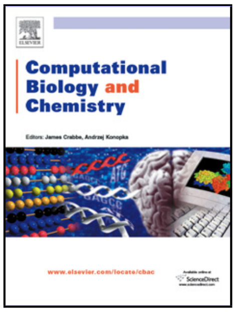

Computational Biology and Chemistry
Editors: James Crabbe, Andrzej Konopka
www.elsevier.com/locate/cbac

Title: Three-Dimensional Protein Structure Prediction: Methods and Computational Strategies

Author: Márcio Dorn Mariel Barbachan e Silva Luciana S. Buriol Luis C. Lamb

PII: S1476-9271(14)00124-8
DOI: http://dx.doi.org/doi:10.1016/j.compbiolchem.2014.10.001
Reference: CBAC 6371

To appear in: *Computational Biology and Chemistry*

Received date: 12-6-2014
Revised date: 3-10-2014
Accepted date: 7-10-2014

Please cite this article as: Márcio Dorn, Mariel Barbachan e Silva, Luciana S. Buriol, Luis C. Lamb, Three-Dimensional Protein Structure Prediction: Methods and Computational Strategies, *Computational Biology and Chemistry* (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.10.001

Highlights

- Predicting the correct three-dimensional structure of a protein molecule is an intricate and arduous task;
- First principle methods without database information;
- First principle methods with database information;
- Fold recognition and threading methods;
- Comparative modeling methods and sequence alignment strategies.

Click here to view linked References

Conformational Space Annealing

Artificial Intelligence

Simulated Annealing

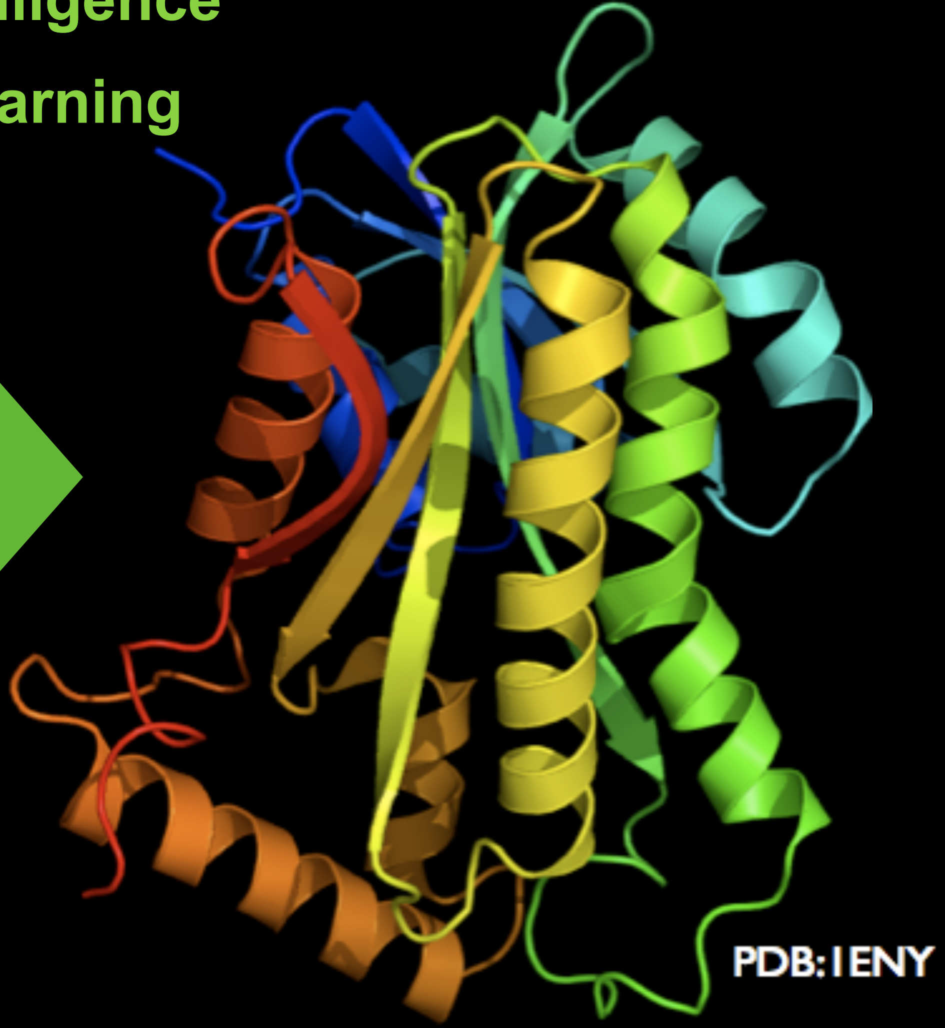
Hidden Markov Models

Machine Learning

Optimization Algorithms

AGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTG
FDRLRLIQRITDRLPAKAPLLELDVQNEEHLASLAGRV
TEAIGAGNKLDGVVHSIGFMPQTGMGINPFFDAPYA
DVSKGIHISAYSYASMAKALLPIMNPGGSIVGMDFDP
SRAMPAYNWMTVAKSALESVNRFVAREAGKYGVRS
NLVAAGPIRTLAMSAIVGGALGEEAGAQIQLLEEGWD
QRAPIGWNMKDATPVAKTVCALLSDWLPATTGDIIYA
DGGAHTQLL

PSP



FUNCTION

Monte Carlo

High Performance Computing

Conjugate Gradient

Data Mining

Genetic Algorithms

Clustering Algorithms

Three-Dimensional Protein Structure Prediction: Methods and Computational Strategies

Márcio Dorn*[a], Mariel Barbachan e Silva[b], Luciana S. Buriol[a], Luis C. Lamb[a]

[a]Federal University of Rio Grande do Sul, Institute of Informatics, Av. Bento Gonçalves 9500, 91501-970, Porto Alegre, RS, Brazil.
[b]Federal University of Rio Grande do Sul, Center of Biotechnology, Av. Bento Gonçalves 9500, 91501-970, Porto Alegre, RS, Brazil.

Abstract

A long standing problem in Structural Bioinformatics is to determine the three-dimensional (3-D) structure of a protein when only a sequence of amino acid residues is given. Many computational methodologies and algorithms have been proposed as a solution to the 3-D Protein Structure Prediction (3D-PSP) problem. These methods can be divided in four main classes: (a) first principle methods without database information; (b) first principle methods with database information; (c) fold recognition and threading methods; and (d) comparative modeling methods and sequence alignment strategies. Deterministic computational techniques, optimization techniques, data mining and machine learning approaches are typically used in the construction of computational solutions for the PSP problem. Our main goal with this work is to review the methods and computational strategies that are currently used in 3-D protein prediction.

Keywords: three-dimensional protein structure prediction, structural bioinformatics, Ab initio methods, knowledge-based methods

1. Introduction

Structural Bioinformatics is one of the key research areas in the field of Computational Biology (Zhang et al., 2005; Altman and Dugan, 2005; Clote and Backofen, 2000; Pevzner, 2000; Liljas et al., 2001; Gopakumar, 2012). Structural Bioinformatics concerns the analysis and prediction of three-dimensional (3-D) structures of biological macromolecules such as Proteins[1], RNA and DNA (Zhang et al., 2005; Altman and Dugan, 2005). This structural information corresponds to 3-D macromolecular structures obtained through different experimental methods such as protein crystallography (X-ray diffraction), electron microscopy or nuclear magnetic resonance (NMR). This information allows one to study folds and local motifs in proteins, molecular folding, evolution and structure/function relationships.

One of the main research problems in Structural Bioinformatics is the prediction of three-dimensional protein structures. Proteins are long sequences formed out of 20 different amino acid residues that in physiological conditions adopt a unique 3-D structure[2] (Anfinsen et al., 1961). Knowledge of the protein structure allows the investigation of biological processes more directly, with higher resolution and finer detail. The sequence-protein-structure paradigm (also known as the ”lock-and-key” hypothesis) says that the protein can achieve its biological function only by folding into a unique, structured state determined by its amino acid sequence (Anfinsen, 1973). Nevertheless, currently it has been recognized that not all protein functions are associated to a folded state (Dunker et al., 2008; Uversky, 2001; Tompa and Csermely, 2004; Tompa, 2002; Wright and Dyson, 1999; Dunker et al., 2001). In some cases proteins must be unfolded or disordered to perform their functions (Gunasekaran et al., 2003). These proteins are called intrinsically disordered proteins (IDP) and represent around 30% of the protein se-

*Corresponding author.
Email addresses: mdorn@inf.ufrgs.br (Márcio Dorn*), mariel.barbachan@ufrgs.br (Mariel Barbachan e Silva), buriol@inf.ufrgs.br (Luciana S. Buriol), lamb@inf.ufrgs.br (Luis C. Lamb)

[1] In this review proteins and polypeptides are treated as synonymous.

[2] Anfinsen hypothesis states that “all the information that dictates the native fold of protein domain is encoded in their amino acid sequence”.

quences. Despite the presence of IDP proteins an important aspect of understanding and interpreting the function of a given protein involves characterizing molecular interactions. These interactions can be intramolecular (ionic bonds, covalent bonds, metallic bonds, etc) or intermolecular (hydrogen bonds and other non-covalent bonds such as van der Waals forces). The knowledge of the 3-D structure of polypeptides gives researchers very important information to infer the function of the protein in the cell (Branden and Tooze, 1998; Laskowiski et al., 2005a,b; Lesk, 2002): structural functions; catalysis in chemical reactions; transport and storage; regulatory functions; gene transcription control; recognition functions. Further details about protein function prediction can be found in Whisstock and Lesk (Whisstock and Lesk, 2003), Rentzsch and Orengo (Rentzsch and Orengo, 2009) and Lee et al. (Lee and Orengo, 2007).

The determination of protein structure is both experimentally expensive (due to the costs associated to crystallography, electron microscopy or NMR), and time consuming (Guntert, 2004). The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the Genome Projects[3] and the number of 3-D structures of proteins which are currently known. Only a tiny portion of protein sequences have experimentally solved three-dimensional structures. These figures not only clearly illustrate the need for, but also motivate further research in Computational Protein Structure Prediction Methods. Over the last 10 years several computational methodologies, systems and algorithms have been proposed as a solution to the Three-Dimensional Protein Structure Prediction (3-D PSP) problem (Bujnicki, 2006; Moult, 2005; Osguthorpe, 2000; Tramontano, 2006). These methods are divided into four classes, that shall be described in detail in this review (Floudas et al., 2006): (1) First principle methods without database information (Osguthorpe, 2000); (2) First principle methods with database information (Rohl et al., 2004; Srinivasan and Rose, 1995); (3) Fold recognition and threading methods (Bowie et al., 1991; Jones et al., 1992; Bryant and Altschul, 1995; Turcotte et al., 1998); and (4) Comparative modeling methods and sequence alignment strategies (Martí-Renom et al., 2000; Sánchez and Sali, 1997). The first group of methods aims at predicting new folds only through (computational) simulation of physicochemical properties of the folding process of the proteins in nature. The other groups represent the methods that are able of performing fast and effective prediction of protein 3-D structures when known template structures and fold libraries are available (Kolinski, 2004).

Predicting the correct 3-D structure of a protein molecule is an intricate and arduous task. The 3-D PSP and Protein Folding (PF) problems[4] are classified in computational complexity theory as NP-complete problems (Crescenzi et al., 1998; Fraenkel, 1993; Hart and Istrail, 1997; Levinthal, 1968; Ngo et al., 1997), i.e, they are among the hardest problems in terms of computational requirements. For a formal definition of NP-completeness see Garey and Johnson (Garey and Johnson, 1979). This complexity is due to the folding process of a protein being highly selective. A long amino acid chain ends up in one out of a huge number of 3-D conformations. In contrast, the conformational preferences of single amino acid residues are weak. Thus, the high selectivity of protein folding is only possible through the interaction of many residues. Therefore, non-local interactions play an important role in protein three-dimensional structure, as local sequence-structure relationships are not absolute (Rackovsky, 2010). *Ab initio* methods (first principle methods without database information) can obtain novel and unknown protein folds. Nevertheless, the complexity and the high dimensionality of the search space (Ngo et al., 1997) even for a small protein molecule makes the problem intractable (Levinthal, 1968). The direct simulation of protein folding in atomic details, as used in Molecular Dynamics (MD)[5], is not tractable (van Gunsteren and Berendsen, 1990) (for large proteins of medical and scientific interest) due to high computational costs, despite the efforts towards the development of distributed computing platforms. On the other hand, homology modelling does not lead to such problems; however, it can only predict structures of protein sequences which are similar or nearly identical to other sequences of known structures. Fold recognition via threading, in turn, is limited to the fold library derived from the Protein Data Bank (PDB) structures (Berman et al., 2000).

In order to tackle the computational complexity of the 3-D PSP problem, current 3-D protein structure prediction methods make use of a wide range of optimiza-

[3] DOE Genomic Science. `http://genomics.energy.gov` (accessed Sep 01, 2014).

[4] Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil.

[5] MD is a simulation method in which the protein system is placed into a random conformation and then the system reacts to force atoms to exert on each other. The model assumes that, as a result of these forces, atoms move in a Newtonian manner. The trajectory of the system should lead to the native conformation.

tion algorithms (Klepeis et al., 2003). Metaheuristics are used to provide near optimal solutions. In addition, considering the limitations of the four classes of protein structure prediction methods, researchers have recently developed hybrid methods which combine principles of the four classes, as can be observed in last CASP editions (Moult et al., 2014, 2011). For example, the accuracy presented by homology modeling methods is combined with the capacity of *Ab initio* methods in predicting novel folds (Dhingra and Jayaram, 2013; Dorn et al., 2008; Fan and Mark, 2004). In order to reduce the complexity and the high dimensionality of the conformational search space inherent to *Ab initio* methods, information about structural motifs found in known protein structures can be used to construct approximate conformations. These approximate conformations are expected to be sufficient to allow later refinement by means of Molecular Mechanics (MM) such as MD simulation (van Gunsteren and Berendsen, 1990). In a refinement step, global interactions between all atoms in the molecule (including e.g. non-bond interactions) are evaluated and deviations in the polypeptide main-chain and side-chain torsion angles can be corrected (Fan and Mark, 2004). These in turn reduce the total time spent by *Ab initio* methods - which usually start from a fully extended conformation of a polypeptide - to fold a sequence of unknown structure (Breda et al., 2007). The first principle methods that make use of database information cover this class. Such methods use previous protein structural information from existing databases in order to construct starting point 3-D protein structures. Machine learning and data mining techniques are also applied in order to extract useful information from known protein 3-D structures.

Our main goal is to review the methods and computational strategies that are currently used in 3-D protein structure prediction. In order to do so, we present the most important results needed to understand the four classes of prediction methods. The main contributions of this review are addressed through the organization and description of the main computational techniques and strategies that are currently used in the development of in silico methods for the 3-D PSP problem. This will contribute toward the development of new computational methods and strategies for the 3-D PSP problem. The review is structured as follows. In Section 2 the fundamental concepts of proteins are briefly described (readers familiar with these fundamental concepts can clearly skip this section). Section 3 describes the four classes in which the 3-D protein structure prediction methods and algorithms are classified. In addition, we present details of the main prediction methods and outline the computational strategies performed. Section 4 concludes the paper and points out directions for further research.

2. On Proteins, Structure and Representation

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acid residues. Each protein is defined by its unique sequence of amino acids. This sequence causes the protein to fold into a particular three-dimensional shape. Predicting the folded structure of a protein only from its amino acid sequence remains a challenging problem in mathematical optimization (Lander and Waterman, 1999). The challenge arises due to the combinatorial explosion of plausible shapes each of which represent a local minimum of an intricate non-convex function of which the global minimum is sought. In nature, proteins typically present 50 to 500 amino acid residues. The books by Lesk (Lesk, 2002) and Tramontano (Tramontano, 2006) present elegant, comprehensive overviews of protein structure.

In nature there are 20 distinct proteinogenic amino acids, each one with its own chemical properties (including size, charge, polarity, hydrophobicity, i.e. the tendency to avoid water packing) (Lodish et al., 1990; Lehninger et al., 2005). Depending on the polarity of the side-chain, amino acids vary in their hydrophilic or hydrophobic character. The importance of the physical properties of the side-chains comes from the influence they have on the amino acid residues interactions in the 3-D structure. The distribution of the hydrophilic and hydrophobic amino acids are important to determine the tertiary structure of the polypeptide. A detailed description of the amino acid properties can be found in Lehninger (Lehninger et al., 2005) and Lodish (Lodish et al., 1990). A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule (H_2O). Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins (Creighton, 1990; Lesk, 2002). The peptide bond (C-N) has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-C_α and C_α-C. These bonds are known as PHI (ϕ) and PSI (ψ) angles, respectively, and are free to rotate (Lesk, 2002; Lodish et al., 1990). This freedom is mostly responsible for the conformation adopted by the polypeptide backbone. However, the rotational freedom around the ϕ (N-C$_\alpha$) and ψ (C$_\alpha$-C) angles is limited by steric hindrance between the side chain of the amino acid residue and the peptide backbone (Branden and Tooze, 1998; Lesk, 2002; Scheef and Fink, 2003). As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties. The peptide bond itself tends to be planar, with two allowed states: trans, $\omega \simeq 180°$(usually) and cis, $\omega \simeq 0°$(rarely) (Branden and Tooze, 1998; Lesk, 2002). The sequence of ϕ, ψ and ω angles of all residues in a protein defines the backbone conformation or fold (Hovmoller and Ohlson, 2002). The angles ϕ and ψ can have any value between $-180°$ and $+180°$. However, some combinations are prohibited by steric interferences between atoms from the main-chain and atoms from the side-chain (two atoms cannot occupy the same space) (Hovmoller and Ohlson, 2002). The allowed and prohibited values for the torsion angles ϕ and ψ are graphically demonstrated by the map of *Sasisekharan-Ramakrishnan-Ramachandran*, or simply Ramachandran map (Ramachandran and Sasisekharan, 1968).

Proteins can be analysed at four levels (Lehninger et al., 2005; Lodish et al., 1990): (a) primary structure, (b) secondary structure, (c) tertiary structure and (d) quaternary structure. This hierarchy facilitates the description and the understanding of proteins. However, it does not aim at describing precisely the physical laws that produce protein structures; it is an abstraction that aims at making protein structure studies more tractable (Scheef and Fink, 2003). The primary structure simply describes the sequence of amino acid residues in a linear order (Branden and Tooze, 1998; Lehninger et al., 2005; Lesk, 2002; Lodish et al., 1990). Each amino acid residue binds to other amino acid residue through a peptide bond. The beginning of the primary structure corresponds to its N-terminal region and the end of its primary structure is the C-terminal region. Proteins are linear polymers that can assume several conformations. The stable arrangement of amino acid residues of the polypeptide forms structural patterns (Lehninger et al., 2005). These structural patterns represent the secondary structure of a polypeptide. The secondary structure is defined by the presence of hydrogen bond patterns between the hydrogen atoms of the amino groups and the oxygen atoms of the carboxyl groups in the polypeptide chain. A regularity in the spatial conformation is maintained through these intermolecular interactions. There are two regular secondary structures: α-helices (Pauling et al., 1951) and β-sheets (Pauling and Corey, 1951). There are also irregular conformations (coil and turns), but the α-helix and β-sheets are the most stable and can be considered as the main elements present in 3-D protein structures.

The tertiary structure of a protein is represented by the distribution of secondary structures in a 3-D space. The three-dimensional shape assumed by a protein is also called native structure of the protein or functional structure. The native structure of a protein is formed by the variation of thermodynamic factors, i.e., covalent interactions, hydrogen bonds, hydrophobic interactions, electrostatic interactions, van der Waals, and repulsive forces (Gibas and Jambeck, 2001; Lehninger et al., 2005; Lesk, 2002; Lodish et al., 1990). In addition, the side-chains play an important role in creating the final structure of the polypeptide (Scheef and Fink, 2003). The tertiary structure of a protein allows the analysis and prediction of the function of the protein in the cell. It is possible to identify the active site, binding sites on a receptor, or a recombination site for the action of another protein (Lehninger et al., 2005). The tertiary structure of a protein is related to its topology (or fold). The topology of a protein is given by the type of succession of secondary structures that are connected to and from the shape in which these structures are organized in a 3-D space. A protein may have different polypeptide chains (or subunits) forming a quaternary structure. The quaternary structure of a protein is the arrangement of various tertiary structures. This structure is maintained by the same forces that determine the secondary and tertiary structures (hydrogen bonds, hydrophobic interactions, hydrophilic interactions) (Lehninger et al., 2005; Lesk, 2002; Lodish et al., 1990).

3. Three-Dimensional Protein Structure Prediction Methods

The prediction of the 3-D structure of polypeptides based only on the amino acid sequence (primary structure) is a problem that has, over the last decades, challenged biochemists, biologists, computer scientists and mathematicians (Baxevanis and Quellette, 1990). The Protein Structure Prediction Problem (Creighton, 1990) is one of the main research problems in Structural Bioinformatics. The main challenge is to understand how the information encoded in the linear sequence of amino acid residues is translated into the 3-D structure, and from this acquired knowledge, to develop computational methodologies that can correctly predict the native structure of a protein molecule. Many methods and algorithms have been proposed, tested and analyzed over the years as a solution to this complex problem. In the literature, one can find several classifications of the

3-D protein structure prediction methods. In this work, we adopt a classification similar to the one described in Floudas (Floudas et al., 2006), which classifies the computational methods for protein structure prediction into four groups:

1. First principle methods without database information - Section 3.1 ;
2. First principle methods with database information - Section 3.2;
3. Fold recognition and threading methods - Section 3.3; and
4. Comparative modeling methods and sequence alignment strategies - Section 3.4.

Regardless of the group, all developed 3-D protein structure prediction methods have to be tested for the ability to predict new protein structures. Every other year since 1994 a worldwide experiment called CASP (Critical Assessment of Structure Prediction) is performed to test protein structure prediction methods. Structural biologists who are about to publish a structure are asked to submit the corresponding sequence for structure prediction. The predictions are then compared with the newly experimentally determined structures (by NMR or X-ray crystallography methods). CASP allows research groups with an opportunity to objectively test their structure prediction methods and provides an independent assessment of the state-of-the art protein structure modeling. The CASP competition involves a large number of research groups using a variety of methods from the four groups listed above. The most significant progress in last CASP was identified by template-based modeling methods (methods that use database information) (Huang et al., 2014; Koop et al., 2007; Cozzetto et al., 2009; Zhang, 2008B; Xu et al., 2011). There was evidence of improved accuracy for targets of mid range difficulty, probably due to improved methods that combine information from multiple templates (Wu and Zhang, 2007a; Cheng, 2008). The major remaining challenge in this class of methods is the development of better methods for template production and identification (Sding, 2005); accurate structure for those regions are not easily derived from an obvious template.

CASP9 an CASP10 did not reveal much progress in Free Modeling methods (first principle methods without database information) (Tai et al., 2014; Jauch et al., 2007; Ben-David et al., 2009; Floudas et al., 2006; Xu et al., 2011). Among the methods that have been tested, I-Tasser presented a significant improvement in its predictions (Xu et al., 2011; Roy et al., 2010). This improvement happened because I-Tasser incorporates two components: REMO (Li and Zhang, 2009) and FG-MD (Li and Zhang, 2011). REMO is a method for atomic structure construction and improvement of hydrogen-bonding network and FG-MD is fragment-guided molecular dynamics based method that uses constrained molecular dynamics simulation to adjust the position of each atom in the protein. Each of the four classes of protein structure prediction methods that will be detailed below have some limitations. The analysis of CASP9 experiments reveals that the best results are achieved by methods which combine principles of the four groups of methods. First principle methods without database information have limitations with respect to the size of the conformational search space (Karplus, 1997; Levinthal, 1968). It is not possible to simulate, in plausible time, all folding process of long sequences of amino acid residues. Methods that use fragments still have two major limitations: the first one is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments; the second refers to the challenge of reducing the potential energy in regions where combination of fragments occur. Despite the high quality predictions, comparative modeling and fold recognition also have some limitations such as the inability to perform prediction of new folds. This is explained by the fact that these methodologies can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures in the PDB. Another limitation is that it is not possible to study the folding process of the protein, i.e., the path that an unfolded protein traverses to the functional state (native state).

3.1. First Principle Methods without Database Information

Ab initio methods, the first principle methods without database information, are founded on thermodynamics and based on the fact that the native structure of a protein corresponds to the global minimum of its free energy (Anfinsen et al., 1961; Anfinsen, 1973; Tramontano, 2006). *Ab initio* structure prediction methods aim at predicting the native conformation of a protein considering only the amino acid sequence (Bonneau and Baker, 2001). Osguthorpe (Osguthorpe, 2000) defines “*Ab initio* folding” as the class of methods that are based on potential energy functions that describe the physics of a current conformational state and where only this potential function is used to search the native structure of the polypeptide. In pure *Ab initio* methods the use of structural templates from a database such as the PDB is not allowed. The structural information from determined structures is only used in the parameterization of empirical all-atoms potentials used in force-fields (potential energy functions) such as AMBER (Cornell et al., 1995), CHARMM (Brooks et al., 1983; MacKerell et al., 1998a), GROMOS (Christen et al., 2005), GROMACS (van der Spoel et al., 2005), OPLS (Jorgensen et al., 1996) and ECEPP/2 (Momany et al., 1975), among others. *Ab initio* protein folding is considered a global optimization problem where the goal is to identify the values of a variable set (torsion angles, position of all atoms or a specific set of atoms in the protein structure) that describe the minimum energy of the polypeptide conformation.

Ab initio methods simulate the protein conformational space using an energy function, which describes the internal energy of the protein and its interactions with the environment in which it is inserted. The goal is to find a global minimum of free energy that corresponds to the native or functional state of the protein (Osguthorpe, 2000; Tramontano, 2006). *Ab initio* methods can predict new folds because they are not limited to templates from the PDB. However, these methods have some limitations with respect to the size of the conformational search space (Karplus, 1997; Levinthal, 1968). This problem is frequently referred to by many authors as the Levinthal's paradox (Zwanzig et al., 1991) following studies carried out by Cyrus Levinthal in 1968 (Levinthal, 1968). In his experiments, Levinthal noted that due to the very large number of degrees of freedom in an unfolded polypeptide chain, a protein molecule has an enormous number of possible conformations (thus rendering a NP-Complete problem) (Crescenzi et al., 1998; Fraenkel, 1993; Hart and Istrail, 1997; Ngo et al., 1997; Levinthal, 1968). In general an *Ab initio* method requires three elements (Chivian et al., 2003; Osguthorpe, 2000): (i) a geometric representation of the protein chain, (ii) a potential function and (iii) an energy surface searching technique. In the sequel, each of these elements are described in further detail.

Geometric Representation: this representation corresponds to the way that computationally we will represent the structure of a protein. The most detailed representations include all atoms of the protein and the surrounding solvent molecules (for example, H_2O). Using all atoms to represent the protein is computationally expensive. Such representations can be simplified in a number of ways (Chivian et al., 2003): the all-atom model of both the protein and the solvent environment (explicit solvent) is usually replaced by employing an united atom model, where the solvent is modeled by potential fields of various descriptions (implicit solvent). In general, the united-atom model is frequently used to reduce the computational cost (Khalili et al., 2005). In this model, explicit hydrogen atoms - with the exception of those that have the capability to participate in hydrogen bonds - are eliminated. Virtual-atoms can also be used to represent one residue and reduce the computational cost (Osguthorpe, 2000). In turn, Rotamers (Dunbrack and Karplus, 2003) can also be used to represent a limited set of conformations that side-chains can adopt in the polypeptide structure. Almost all *Ab initio* folding methods use some form of simplified geometry model, in which single virtual atoms of the model represent a number of atoms in the all-atom model (Osguthorpe, 2000). The geometric representation is one of the most important elements of an *Ab initio* method and is directly related to the reduction or increase of the associated computational complexity. An all-atom model can demand enormous computational effort during a simulation. On the other hand, simplified representation models can preserve the main structure characteristics and reduce the computational time demanded by a protein folding simulation.

Potential Functions: the second element of an *Ab initio* method is a potential energy function. Potential energy functions are used in Molecular Mechanics (MM) simulations (Jorgensen and Tirado-Rives, 2005; Mackerell, 2004), protein design (Li et al., 2013; Gordon et al., 1999; Pokala and Handel, 2000) and protein structure prediction (Lazaridis and Karplus, 2000). There are two categories of potentials: MM potentials and protein structure-derived potential functions (scoring functions) (Zhang and Skolnick, 2004a). The first category aims at modeling the forces that determine protein conformations using physically-based parameterized functional forms from small molecule data or in vacuo quantum mechanics calculations (Chivian et al., 2003). The second category is empirically derived from experimental structures from the PDB (Chivian et al., 2003; Hao and Scheraga, 1999; Koppensteiner and Sippl, 1995; Lazaridis and Karplus, 2000; Mohanty et al., 1999; Sippl et al., 1992; Sippl, 1995; Lu and Skolnick, 2001; Gohlkea et al., 2000). These two classes of potentials represent the forces that determine the macromolecular conformation: solvation[6], electrostatic[7], van der Waals

[6] Solvation is the process of attraction and association of molecules of a solvent with molecules or ions of a solution.

[7] Composed by hydrogen bonds, salt bridges and van der Waals. It provides attractive forces between molecules.

interactions[8], covalent bonds[9], angles, torsions (Boas and Harbury, 2007; Chivian et al., 2003; Park et al., 1997; Pokala and Handel, 2000). The main advantage of using a knowledge-based energy function is that it can model any behavior observed in known protein crystal structures, even when there is no good physical understanding of their behavior (Boas and Harbury, 2007). The disadvantage is that these functions cannot predict new behaviors absent in the training set obtained from the experimental database. A potential energy function incorporates two types of terms: bonded and non-bonded. The bonded terms (*bonds*, *angles* and *torsions*) are covalently linked. The bonded terms constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential (*torsion*) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential includes: ionic bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, and dipole-dipole bonds. There is a great number of potential energy functions used in computational molecular biology. AMBER (Cornell et al., 1995), CHARMM (Brooks et al., 1983; MacKerell et al., 1998a) and ECEPP (Momany et al., 1975) are the most widely used potential energy functions in 3-D PSP and Protein Folding problems. A review of potential energy functions is found in Halgren (Halgren, 1995). Equation 1 presents the CHARMM force field (MacKerell et al., 1998a).

$$
\begin{aligned}
E_{\text{total}} &= \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{UB}} K_{UB}(S-S_0)^2 \\
&+ \sum_{\text{angle}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} K_\chi(1+\cos(\eta-\delta)) \\
&+ \sum_{\text{impropers}} K_{imp}(\varphi-\varphi_0)^2) \\
&+ \sum_{\text{nonbond}} \epsilon\left[\left(\frac{R_{minij}}{r_{ij}}\right)^{12} - \left(\frac{R_{minij}}{r_{ij}}\right)^{6}\right] + \frac{q_iq_j}{\epsilon_1 r_{ij}}
\end{aligned}
\qquad (1)
$$

where K_b, K_{UB}, K_θ, K_χ and K_{imp} are the bond, Urey-Bradley angle (Hagler et al., 1979; Lifson and Warshel, 1968), dihedral angle and improper dihedral angle force constants, respectively; b, S, θ, χ and φ are the bond length, Urey-Bradley 1.3 distance, bond angle, dihedral angle, and improper torsion angle, respectively. The subscript zero represents the equilibrium value for the individual terms. Coulomb and Lennard-Jones 6-12 terms contribute to the external or non-bonded interactions; ϵ is the Lennard-Jones (the depth of the potential well) and R_{min} is the distance at the Lennard-Jones minimum, q_i is the partial atomic charge, ϵ_1 is the effective dielectric constant, and r_{ij} is the distance between atoms i and j.

Energy Surface Search Techniques: methods for *Ab initio* prediction include Molecular Dynamics simulations of proteins and protein-substrate complexes (van Gunsteren and Berendsen, 1990; Rapaport, 2004; Koza, 1992); Monte Carlo simulations that do not use forces but rather compare energies, via the use of Boltzmann probabilities (Simons et al., 1997); Genetic Algorithms which are based on populations of solutions by iterative cycles of operations (Holland, 1993; Pokala and Handel, 2000) and try to improve on the sampling and the convergence of Monte Carlo approaches (Pedersen and Moult, 1997; Tuffery et al., 1991; Bowie and Eisenberg, 1994), and exhaustive and semi-exhaustive lattice-based studies which are based on using a crude/approximate fold representation (such as two residues per lattice point) and then exploring all or large amounts of the conformational space given the crude representation.

There are many computational packages used in *Ab initio* protein structure simulations. These simulation packages are frequently used in the protein folding problem and in other molecular modeling problems such as molecular docking (Lengauer and Rarey, 1996; Kitchen et al., 2004), which predicts the preferred orientation of a molecule with respect to another molecule when bound to each other to form a stable complex (Lengauer and Rarey, 1996). There are also *Ab initio* algorithms developed specifically for the 3-D PSP problem. Section 3.1.1 present an overview of the most common *Ab initio* approaches for the 3-D PSP problem.

3.1.1. Overview

The most common simulation packages used in first principle methods without database information are: AMBER (Assisted Model Building with Energy Refinement) (Case et al., 2005; Pearlman et al., 1995), CHARMM (Chemistry at HARvard Molecular Mechanics) (Best et al., 2012; Brooks et al., 1983; MacKerell et al., 1998b), UNRES (Liwo et al., 1998, 1999a,b, 1997), GROMACS (Groningen MAchine for Chemical Simulation) (Pronk et al., 2013; van der Spoel et al., 2005; Hess et al., 2008) and TINKER (Software Tools for Molecular Design) (Ponder and Richards, 1987; Kundrot et al., 1991). *Ab initio* protein structure prediction methods include: LINUS (Local Independent Nu-

[8] van der Waals are the attractive or repulsive forces between molecules or between parts of the same molecule.

[9] A covalent bond is a form of chemical bonding that is characterized by the sharing of pairs of electrons between atoms, and other covalent bonds.

cleated Units of Structure) (Srinivasan and Rose, 2002, 1995), ASTROFOLD (Subramani et al., 2012; Klepeis and Floudas, 2003) and BHAGEERATH (Shenoy and Jayaram, 2010; Jayaram et al., 2006; Narang et al., 2006).

Each simulation package and protein structure prediction method make use of specific computational strategies in order to search the conformational space and find the native structure of the target polypeptide. Tables 1 and 2 summarizes the main computational strategies implemented and used in the described molecular packages and *Ab initio* protein structure prediction methods. Molecular modeling packages implement many potential energy functions and their application and use depend on the type of the molecular simulation problem and simulation parameters used as solvent, temperature, pressure, etc. Usually, *Ab initio* prediction methods use only one potential energy or scoring functions that analyze specific features. For example, BHAGEERATH uses an empirical energy function that considers the non-bonded energy of a protein, expressed as a sum of three energy terms: electrostatic, van der Waals, and hydrophobic (Arora and Jayaram, 1998); LINUS is based on steric and conformational entropy and the terms used in the scoring functions are the hydrogen bonds and hydrophobic interactions; ASTROFOLD uses the ECEPP/3 force-field. Most used *ab initio* packages and prediction methods are detailed below.

AMBER (Salomon-Ferrer et al., 2013; Case et al., 2005; Pearlman et al., 1995) is an example of *ab initio* package that allows users to carry out and analyze MD simulations for proteins, nucleic acids and carbohydrates. Basically, it is composed of two parts: (1) a set of molecular mechanical force fields for the simulation of biomolecules and (2) a set of molecular simulation programs. The first part covers the set of empirical parameters used in the simulations. The second part is concentrated on the methods used for energy minimization and molecular dynamics. There are three main steps in an AMBER *Ab initio* simulation task: (i) system preparation; (ii) simulation and (iii) trajectory analysis. AMBER provides support for explicit and implicit solvent models (Richards, 1977). In case of explicit solvents it provides support for water models, methanol, chloroform, N-methylacetamide and urea/water mixtures. The implicit solvent model has several advantages over the explicit water representation; the main advantage is related to the fact that implicit models are often computationally less expensive. AMBER implements implicit solvent models with Poisson-Boltzmann (Fogolari et al., 2002) and Generalized Born approach (Still et al., 1990; Onufriev et al., 2002) and explicit solvent models with Particle-Mesh Ewald summation (Darden et al., 2009). A good general overview of the AMBER codes can be found in Case et al. (Case et al., 2005) and Salomon-Ferrer et al. (Salomon-Ferrer et al., 2013).

CHARMM (Brooks et al., 1983; MacKerell et al., 1998b) provides a molecular dynamics simulation and analysis package as well as a widely used set of force fields for molecular dynamics. The package allows generation and analysis of a wide range of molecular simulations: energy minimization and molecular dynamics simulations. There are more advanced features such as free energy perturbation, quasi-harmonic entropy estimation, correlation analysis, and combined quantum and molecular mechanics methods. CHARMM is one of the most used programs for molecular dynamics. It is also considered the oldest program for MD simulations and has accumulated a huge number of features. The CHARMM force fields for proteins include: united-atom model, CHARMM19, all-atom CHARMM22 (MacKerell et al., 1998a) and its dihedral potential corrected variant CHARMM22/CMAP (MacKerell et al., 2004) and CHARMM36 (Best et al., 2012). CHARMM22 is parameterized for the TIP3P explicit water model (Jorgensen et al., 1983) and frequently used with implicit solvents. For DNA, RNA, and lipids, CHARMM27 (MacKerell et al., 2001) is used.CHARMM36 is an improved version of CHARMM22 in several protein force field parameters, such as a new backbone CMAP potential, refined against a range of data for dipeptides as well as experimental data on small peptides such as hairpins and helices, and new side-chain dihedral potentials optimized against quantum mechanical energies from dipeptides and NMR data from unfolded proteins.

GROMACS (Pronk et al., 2013; Hess et al., 2008; van der Spoel et al., 2005)] provides an MD program with source code specially directed towards the simulation of biological macromolecules in aqueous and membrane environments. It does not have a force field of its own, but it is compatible with other force fields such as AMBER (Cornell et al., 1995), OPLS (Jorgensen et al., 1996), GROMOS (Christen et al., 2005) and ENCAD (Levitt, 1983). The package provides micro-canonical Hamiltonian Mechanics (LaValle, 2006), stochastic dynamics and energy minimization algorithms. GROMACS package includes a set of analysis tools that allows trajectory and structural fluctuation analysis. A molecular system is defined by its size and shape, the number of types of molecules it contains, and the coordinates and velocities of each atom. The forces and energies are computed on the basis of three different types of interactions: bonded interactions (between two, three or four particles), non-bonded interactions (between pairs of particles) and special interactions (that can define or impose position, angle or distance constraints on the motion of the system). GROMACS implements quantum mechanics and molecular mechanics approaches that are frequently used to simulate chemical reactions in solution or in enzymes. Both approaches have interfaces to several Quantum Chemistry Packages (MOPAC (Dewar, 1983), GAMESS-UK (Guest et al., 2005), GAUSSIAN[10]). GROMACS uses an united-atom model in order to reduce the complexity of representing the molecular structure and for removing some degrees of freedom. The package has long been used in the protein folding problem (van der Spoel et al., 1996; van der Spoel and Berendsen, 1997; van der Spoel, 1998).

TINKER (Kundrot et al., 1991; Ponder and Richards, 1987)] is a software package used in empirical force field molecular mechanics and Molecular Dynamics calculations. It implements a variety of algorithms including distance geometry with fast metrization and Gaussian trial distances (Huang et al., 1998); Elber's reaction path method (Elber and Karplus, 1987), global optimization via Potential Smoothing and Search algorithms (Pappu et al., 1998), Molecular Dynamics (van Gunsteren and Berendsen, 1990) with simulated annealing and stochastic dynamics (Guarnieri and Sitill, 1994); Particle Mesh Ewald summation (Darden et al., 2009); Monte Carlo minimization; atomic multi-pole treatment of electrostatics with explicit dipole (Williams, 1998); Eisenberg-McLachlan ASP (Eisenberg and McLachlan, 1986; Wesson and Eisenberg, 1992) and GB/SA (Qiu et al., 1997; Still et al., 1990) continuum solvation models and truncated Newton TNCG local energy minimization (Ponder and Richards, 1987; Dembo and Steihaug, 1983; Derreumaux et al., 1994; Eisenstat and Walker, 1996). The routines from TINKER[11] package provide many functions that can be used in the protein folding problem (Ponder, 2010): (1) energy minimization and structural optimization via conjugate gradient, variable metric or truncate Newton method over Cartesian coordinates, torsion angles or rigid bodies; (2) molecular, stochastic and rigid body dynamics with periodic boundaries[12] and control of temperature and pressure; (3) analysis of energy distribution within a structure; (4) simulated annealing with various cooling protocols; (5) normal mode vibrational analysis; (6) conformational search and global optimization; (7) transition state location and conformational pathways; (8) fitting of energy parameters to crystal data; (9) distance geometry with pairwise metrization; (10) molecular volumes and surface areas; (11) free energy changes for structural mutations, and (12) global optimization via energy surface smoothing including Potential Smoothing and Search method.

BHAGEERATH (Shenoy and Jayaram, 2010; Jayaram et al., 2006; Narang et al., 2006) is an *ab initio* protein structure prediction algorithm. It reduces the search space to generate probable candidates for the protein native structure using a set of eight modules. Module one (Generate PDB from FASTA sequence) involves the formation of a 3-D structure from the amino acid sequence with the secondary structure information. Module two (Generate trial structures) involves the generation of a large number of trial structures with a systematic sampling of the conformational space of loop dihedrals. Module three (pad through biophysical filters) has the objective of reducing the number of improbable candidates through the application of a screening procedure based on persistence length[13] and radios of gyration filters[14]. The resultant structures are refined in module four by a Monte Carlo sampling procedure in dihedral space to remove steric clashes between atoms of the main chain and side chains. In Module five the energy of the structures is minimized (step descent and conjugate gradient approaches) to further optimize the side chains. Module six consists of ranking the structures using an all atom energy based empirical scoring function (Narang et al., 2006). Module seven reduces the probable candidates based on the protein regularity index (Thukral et al., 2007). The last module selects the 10 best structures using a topological equivalence criterion and the accessible surface area (Richards, 1977).

UNRES (Liwo et al., 1998, 1999a,b, 1997) is a united-residue force field for energy-based prediction of protein structure. In the UNRES model, a polypeptide chain is presented as a sequence of α-carbon(C_α) atoms linked by virtual bonds with attached united side-chains

[10] Gaussian. http://www.gaussian.com (accessed Sep 01, 2014).

[11] TINKER - Software Tools for Molecular Design. http://dasher.wustl.edu/tinker/ (accessed Sep 01, 2014).

[12] In Molecular Dynamics, periodic bond conditions are usually applied to simulate bulk gasses, liquids, crystals or mixtures.

[13] Persistence length: is the maximum length of the uninterrupted polypeptide chain persisting in a particular direction.

[14] Radius of gyration: describes the overall spread of the molecule and is defined as the root mean square distance of the collection of atoms from their common gravity center.

and united peptide groups. Each united peptide group is located in the middle between two consecutive C_α, with a peptide group being located between a $C_{\alpha i}$ and $C_{\alpha i+1}$. UNRES considers as interaction sites only the united peptide groups and the united side-chains, the C_α are used only to define the geometry of the polypeptide chain. UNRES force-field is widely used in protein folding simulations and protein structure prediction (Oldziej et al., 2005; Liwo et al., 2010; Maisuradze et al., 2010; Czaplewski et al., 2009; Shen et al., 2009; He et al., 2009; Nanias et al., 2009; Shen et al., 2008).
ASTROFOLD (Klepeis and Floudas, 2003) is a combinatorial and global optimization framework for the *Ab initio* prediction of 3-D structures of proteins. It is composed by four main steps: (1) α-helix prediction; (2) β-sheet prediction; (3) loop modelling and (4) tertiary structure prediction. In the first step, the principle of hierarchical folding is used to predict α-helices (Klepeis and Floudas, 2002), where the polypeptide sequence is divided into sub-sequences and optimization techniques are employed in order to find the conformation of a given peptide with the lowest energy (Klepeis et al., 1998). Two algorithms are used to generate low energy assembles: (1) a deterministic branch and bound algorithm (αBB) (Klepeis et al., 1998B; Klepeis and Floudas, 1999) and (2) conformation space annealing (Lee and Scheraga, 1999; Lee et al., 2000, 2001). After predicting the α-helices the remaining residues are analyzed in order to identify the formation of β-sheets. The β-sheet prediction is based mainly on the hydrophobic information and on the prediction of tertiary hydrophobic contacts to identify parallel and anti-parallel structures (Kepleis and Floudas, 2002B). The formulation of hydrophobic interactions between β-sheets residues produces an integer linear programming problem that is solved through an iterative solution and integer cut constrains (Kepleis and Floudas, 2002B). In the loop modeling and restraint step, the structure prediction problem is formulated based on the development of atomic distance and dihedral-angle restraints derived from the α-helix and β-sheet prediction results. The dihedral angles bonds are assigned according to the predicted structure class: α-helix, β-sheet or loop. The loop region has a large structural variability and its prediction is a complex computational task. ASTROFOLD uses a physics-based *ab initio* protein structure approach (Klepeis and Floudas, 2003B) in order to predict the loop segments. After determining the appropriate bounds on dihedral angles and inter-atomic distances, a combination of an αBB global optimization algorithm, stochastic global optimization and MD in torsion angle space (Klepeis and Floudas, 2003B) is used in order to find the polypeptide structure with the lowest internal energy.
LINUS (Srinivasan and Rose, 2002, 1995) Local Independent Nucleated Units of Structure is an implementation of a hierarchical fold model (Rose, 1979; Rose and Wolfenden, 1993) to predict the fold of a protein. It is based on the idea that globular proteins are organized as a structural hierarchy (Grippen, 1978; Rose, 1979) and that a complex fold can be decomposed into secondary structure elements (α-helix, β-sheet, coil, loops) together with their superstructure (Richards and Kundrot, 1988). The LINUS algorithm accumulates favourable structures that are acceptable in a fixed interval of allowed interactions, and repeats this in stages as the size of the interval increases. In each stage the polypeptide chain is allowed to randomly move by under the influence of an energy function. A hierarchy is established in order to recognize favourable conformations at an early stage; they are then constrained in order to persist during the algorithm stages. It considers two types of interactions during the simulation stage: repulsive (two non-bonded atoms can not occupy the same space at the same time), and attractive interactions (hydrogen bonds and the tendency of apolar residues to cluster). The algorithm starts with a target amino acid sequence and sequentially three residues are perturbed simultaneously to generate a new trial conformation. This generated conformation is discarded when one of its amino acid residues overlap. Otherwise, the energy of the conformation is calculated by adding interactions between all residues separated in the sequence by no more than the current interval. The energy is then evaluated; if not rejected, this conformation is defined as the new current conformation. A complete progression from N to C is a cycle. For each interval interaction 6000 cycles are performed (1000 equilibrium steps followed by 5000 trial structures are generated). The trial conformations are retained. Chain segments in the trial ensembles that adopt a persisting conformation in an interval are constrained, and remain in that conformation during subsequent intervals. The energy of a current conformation decreases over the course of each interval and from each interval to the next. At the end, the predicted structure is the conformation with the lowest energy in the last interval. A Monte Carlo procedure is used by the algorithm to escape energy local minimal.

There are several other *ab initio* simulation packages and *Ab initio* algorithms that are also used in the context of the 3-D PSP problem. These packages and algorithms are similar in some aspects to the ones described above. The structure of these packages and algorithms are basically the same. Here, we list the other commonly used *Ab initio* packages and prediction algorithms: ABALONE[15], GROMOS (Scott et al., 1999), MACROMODEL (MacroModel, Schrödinger, LLC, New York, NY - `http://www.schrodinger.com`), MOIL (Elber et al., 1995; Elber, 2005), MOE (The Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada - `http://www.chemcomp.com`), NAB (Nucleic Acid Builder) (Macke and Case, 1998), ADUN (Johnston et al., 2005), ACEMD (Harvey et al., 2009), SPARTAN (Wavefunction, Inc., Irvine, California, USA - `http://www.wavefun.com`), PLOP (Protein Local Optimization Program) (Jacobson et al., 2002, 1968B, 2004), BOSS (Biochemical and Organic Simulation System) (Jorgensen and Tirado-Rives, 2005B), HOOMD (Highly Optimized Object Oriented Molecular Dynamics) (Anderson and Travesset, 2008), LAMMPS (Plimpton, 1995), ITAP (ITAP Molecular Dynamics Program) (Stadler et al., 1997), CPMD (Carr-Parrinello Molecular Dynamics) (Andreoni and Curioni, 2000; Hutter and Curioni, 2005), SMMP (Simple Molecular Mechanics for Proteins) (Eisenmenger et al., 2001, 2006; Meinke et al., 2008), MOLDY (Refson, 2000), MACSIMUS (MACromolecule SIMUlation Software) (J. Kolafa, Prague Institute of Chemical Technology, Czech Republic - `http://www.vscht.cz/fch/software/MACSIMUS`), DL POLY (Smith and Forester, 1996; Smith et al., 2002), ESPRESSO (Extensible Simulation Package for Research on Soft matter) (Limbach et al., 2006), MDYNAMIX (Molecular Dynamics of Mixtures) (Lyubartsev and Laaksonen, 2000), MCPRO (Jorgensen and Tirado-Rives, 2005B), OPENMD (Kuang et al., 2009), ORAC (Marsili et al., 2010; Procacci et al., 1997), PACKMOL (Martínez et al., 2009), PINYMD (Tuckerman et al., 2000), Q (Marelius et al., 1999), SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms) (Soler et al., 2001), VASP (Kresse et al., 2009), SAGEMD (Selezenev et al., 2003), NAMD (Phillips et al., 2005), MOSCITO (Paschek and Geiger, 2003), MCCCS TOWHEE (Martin and Siepmann, 1999). Table 1 lists the simulation packages most widely used in the protein folding and 3-D PSP problems. The main computational methods offered by each package are also listed (Table 2).

3.2. First Principle Methods with Database Information

In first principle methods with database information general rules of protein structures are extracted from protein databases and used to build starting point 3-D protein structures. These methods do not compare a target sequence to a known structure, but they compare fragments, i.e. short amino acid sub-sequences of a target fragment against fragments of known protein structures (Floudas et al., 2006). This arises from the observation that when a new fold is discovered, it is composed of common structural motifs or fragments from super-secondary structures of proteins with known structures (Tramontano, 2006). Thus, if there are protein fragments that fold into similar structures, then this information or these fragments can be used to construct 3-D structural models of proteins. This is the essence of the methods based on fragments. The conformation of a protein is seen as a set of various fragments of amino acid sequences representing various structural motifs that are combined to form a 3-D protein structure. 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 (Simons et al., 1997; Jones, 1997) with the purpose of finding the structure with the lowest potential energy. When finding polypeptide structures with the lowest energy potential, these methods are similar to *ab initio* methods. However, they cannot be classified as *ab initio* methods because they use database information to predict the structure of polypeptides. Fragment-based methods are based on the premise that local interactions can define local structures in proteins. Local structures present in known protein structures are used in order to predict the structure of a target amino acid sequence. When appropriate fragments have been identified, compact structures can be assembled by randomly combining fragments using, for example, a simulated annealing approach (Simons et al., 1997; Rohl et al., 2004).

Similar local sequences do not always present the same 3-D structure. This occurs because in a 3-D structure a large number of physicochemical interactions are present; such interactions contribute not only to the stability of the global structure, but also to the configuration of the secondary structures. Thus, fragment-based methods cannot fragment the target amino acid sequence, search database template fragments, get their information and combine these fragments without any combination criterion. Non-covalent interactions between atoms of different regions of the molecule influence the formation of local structures (Tramontano,

[15] Biomolecular simulations with Abalone. `http://www.biomolecularmodeling.com/Abalone/index.html` (accessed Sep 01, 2014).

Table 1: First principle methods without database information: simulation packages. Main internal computational methods: MD (Molecular Dynamics simulations); EM (Energy Minimization), CG (Conjugate Gradient), SD (Step Descendants), MM (Molecular Mechanics calculations), MC (Monte Carlo) statistical mechanics simulations, semi-empirical Molecular Orbital (MO), Quantum Mechanics (QM), mixed Quantum and Molecular Mechanics (QM/MM), Density Functional Theory (DFT), Truncated Newton (TN) and LBFGS (Limited memory Broyden Fletcher Goldfarb Shanno method).

SIMULATION PACKAGES	MD	EM	SD	CG	L-BFGS	TN	MC	MM	QM/MM	DFT	QM	MO
ACEMD (Harvey et al., 2009)	•	•										
ADUN (Johnston et al., 2005)	•	•									•	
AMBER (Case et al., 2005; Pearlman et al., 1995)	•	•	•	•								
BOOS (Jorgensen and Tirado-Rives, 2005B)							•	•	•			•
CHARMM (Brooks et al., 1983)	•	•					•					
CPMD (Andreoni and Curioni, 2000; Hutter and Curioni, 2005)										•		
DLPOLY (Smith and Forester, 1996; Smith et al., 2002)	•											
ESPRESSO (Limbach et al., 2006)	•						•					
GROMACS (Hess et al., 2008; van der Spoel et al., 2005)	•	•	•	•	•							
HOOMD (Anderson and Travesset, 2008)	•											
ITAP (Stadler et al., 1997)	•											
LAMMPS (Plimpton, 1995)	•											
MACSIMUS	•											
MCCST. (Martin and Siepmann, 1999)							•					
MCPRO (Jorgensen and Tirado-Rives, 2005B)							•					
MDYNAMIX (Lyubartsev and Laaksonen, 2000)	•											
MOIL (Elber et al., 1995; Elber, 2005)	•						•					
MOE	•	•										
MOLDY (Refson, 2000)	•											
MOSCITO (Paschek and Geiger, 2003)	•											
NAB (Macke and Case, 1998)		•										
NAMD (Phillips et al., 2005)	•											
OPENMD (Kuang et al., 2009)	•											
ORAC (Marsili et al., 2010; Procacci et al., 1997)	•											
PINYMD (Tuckerman et al., 2000)	•											
Q (Marelius et al., 1999)	•											
SAGEMD (Selezenev et al., 2003)	•						•			•		
SIESTA (Soler et al., 2001)	•											
SMMP (Eisenmenger et al., 2001, 2006; Meinke et al., 2008)	•	•										
PLOP (Jacobson et al., 2002, 1968B, 2004)	•											
SPARTAN		•					•				•	
TINKER (Kundrot et al., 1991; Ponder and Richards, 1987)	•	•		•		•						
UNRES (Liwo et al., 1998, 1999a,b, 1997)							•	•				
VASP (Kresse et al., 2009)	•										•	

Table 2: First principle methods without database information: computational *Ab initio* protein structure prediction methods. Main internal computational methods: BB (Branch and Bound algorithms), CSA (Conformational Space Annealing), MC (Monte Carlo), Stochastic Tunneling (ST), Parallel Tempering (PT), Swarm-based optimization algorithms (SB), Memetic algorithm (MME), Genetic Algorithms (GA) and Replica Exchange Monte Carlo (REMC).

FIRST PRINCIPLE METHODS	MC	BB	CSA	GA	REMC	ST	PT	SB	MME
Abagyan (Abagyan and Totrov, 1994)	•								
Astrofold (Klepeis and Floudas, 2003)	•	•	•						
Bahamisch et al. (Bahamish et al., 2009)								•	
Bhageerath (Jayaram et al., 2006; Narang et al., 2006)	•								
Dandekar and Argos (Dandekar and Argos, 1992, 1994)				•					
Derreumaux (Derreumaux, 1999)	•								
Fonseca et al. (Fonseca et al., 2010)								•	
Gibbs et al. (Gibbs et al., 2001)	•								
Grand and Merz (Le Grand and Merz, 1993)				•					
Herges et al. (Herges et al., 2003)						•			
Hoque (Hoque et al., 2005, 2006, 2009)				•					
Linus (Srinivasan and Rose, 2002, 1995)	•								
Pedersen and Moult (Pedersen and Moult, 1997)				•					
Pokarowski (Pokarowski et al., 2003)					•				
Schug et al. (Schug et al., 2005)							•		
Smith (Smith, 2005)									•
Sun (Sun, 1995)				•					
Thachuk (Thachuk et al., 2007)					•				
Unger and Moult (Unger and Moult, 1993, 1995B)				•					

2006). Fragment-based methods need to establish a relationship criterion between the fragments so that they can determine the fragments with higher probability of insertion during the prediction of the final structure. In this sense, scoring functions are frequently used. The fitness of a conformation can be assessed with scoring functions (Zhang and Skolnick, 2004a) derived from conformational statistics of known proteins (Floudas et al., 2006). Additional information can be used in order to improve the scoring functions, for example, secondary structure information (Simons et al., 1999B). Figure 1 depicts a generic schematic representation of a fragment-based method.

Usually, given the complete sequence of amino acids in a protein, a fragment-based method is composed of five distinct stages where:

1. It divides the target sequence into fragments;
2. It carries out the search for similar sequences from each fragment, in a database of known structures;
3. It classifies the fragments (scoring);
4. It constructs the three-dimensional structure from the fragment template using a combination technique;
5. Finally, it refines the conformation.

As first principle method without database information, fragment-based methods offer advantages over the other classes of prediction methods. The first advantage refers to the ability of predicting new folds, which cannot be achieved by methods based on homology modelling (Section 3.4). In comparison with *Ab initio* methods, fragment-based methods take advantage of the reduction of the conformational search space. This reduction is due to the fact that in a simple replacement of a fragment in the target protein, this fragment moves from one region of a protein which has a structure with minimum potential energy. However, despite reducing the conformational search space, the methods that use fragments still have two major limitations. The first one is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments. The second one refers to the challenge of reducing the potential energy in regions where combination of fragments occur. Fragment-based methods produced very positive results in the CASP experiments (Moult et al., 2014, 2011).

3.2.1. Overview

The most common first principle methods with database information for the 3-D PSP problem are presented below.

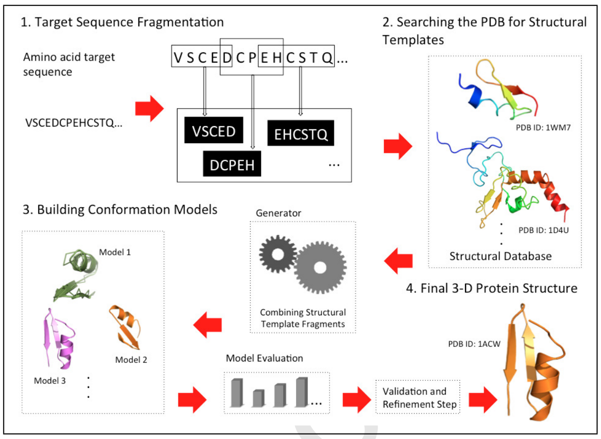

1. Target Sequence Fragmentation
Amino acid target sequence
V S C E D C P E H C S T Q...
VSCEDCPEHCSTQ...
VSCED
EHCSTQ
DCPEH
...
2. Searching the PDB for Structural Templates
PDB ID: 1WM7
PDB ID: 1D4U
Structural Database
3. Building Conformation Models
Generator
Combining Structural Template Fragments
Model 1
Model 2
Model 3
Model Evaluation
Validation and Refinement Step
4. Final 3-D Protein Structure
PDB ID: 1ACW

Figure 1: General schematic representation of a fragment-based method for the 3-D PSP problem: a target sequence is fragmented, molds are obtained from the PDB, the fragments are classified, the conformation is constructed and when appropriate, conformation is refined.

I-TASSER (Roy et al., 2010; Wu and Zhang, 2007b; Zhang, 2007, 2008, 2009) is an interactive implementation of the TASSER method (Zhang and Skolnick, 2004C,B). In the first stage the target sequence is threaded through the PDB to identify appropriate local fragments. Such fragments will incur further structural reassembly. The threading method used in I-TASSER is a simple profile-profile alignment (PPA) approach (von Ohsen et al., 2003). The frequency of amino acid residues obtained with a PDB PSI-BLAST (Altschul et al., 1990) search, the secondary structure prediction from PSIPRED (Jones, 1999B) of the query sequence, and the secondary structure assignment by DSSP (Kabsch and Sander, 1983) are used as terms in a score function. In I-TASSER the protein chain is divided into aligned and unaligned regions based on the PPA results and for a given target sequence the method proceeds as follows: (1) an initial model is built by connecting contiguous secondary structure fragments (I-TASSER considers a contiguous secondary structure a sequence with at least 5 residues) through a random walk of C_α-C_α bond vectors (Wu and Zhang, 2007b); (2) this initial structure is submitted to a parallel Monte Carlo sampling for assembling/refinement (Zhang et al., 2002). I-TASSER uses an energy function that includes predicted secondary structure propensities from PSIPRED, backbone hydrogen bonds (Zhang et al., 2006), a variety of statistical short-range and long-range correlations (Zhang and Skolnick, 2004C) and predicted accessible surface area through an artificial neural network approach (Chen and Zhou, 2005). Secondly, the trajectories obtained by the simulation in the first stage are clustered (Zhang and Skolnick, 2004D); the cluster centroids are obtained and a Monte Carlo simulation is applied beginning with the cluster centroids conformation. Contact restraints are obtained from the combination of centroid structures and PDB structures searched by the structure alignment program TM-align (Zhang and Skolnick, 2005) based on the cluster centroids. The conformation with the lowest energy is selected and the backbone atoms are added by PULCHRA (Feig et al., 2000) and the side-chains are added and optimized by SCWRL (Canutescu et al., 2001).

ROSETTA (Kim et al., 2014; Rohl et al., 2004; Schueler-Furman et al., 2005; Simons et al., 1999B) is a fragment-based method for the 3-D PSP problem that makes use of an assembly strategy to combine native-like structures of fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. The main goal of the ROSETTA scoring function is to search for the most probable structure of a protein given the amino acid sequence and the large number of examples of sequences with known structure in the PDB. A Bayes-based theorem (Simons et al., 1999) gives the probability of a structure depending on the amino acid sequence. The use of this theorem includes some biological information such as radius gyration[16], solvation[17] and residue pair interactions. The 3-D structures are generated by splicing together fragments of known structures with similar local sequences and evaluating them using the scoring function. ROSETTA represents a protein structure using a simplified model consisting of the heavy atoms of the main-chain and the C_β atom of the side-chain and the backbone torsional angles. All bond lengths are held constant. The low scoring conformations with distributions of residues of known proteins are identified through a simulated annealing approach in conjunction with the replacement of the torsion angles of segments in the polypeptide chain. ROSETTA uses both 3 and 9 amino acid residues as the fragment length. There are some variations of the ROSETTA method that implement a Monte Carlo procedure to search the protein conformation with the lowest energy (Schueler-Furman et al., 2005; Bradley et al., 2005).

ROSETTA@home (Das et al., 2007) is a computing network based on the Berkeley Open Infrastructure Network Computing protocol(Anderson, 2004) that implements a ROSETTA-based algorithm on a Grid computing platform. The all-atom energy function and the refinement procedure used by ROSETTA@home is the same used in Schueler-Furman et al. (Schueler-Furman et al., 2005). ROSETTA@home uses three different template-based modeling strategies depending on the sequence size and sequence identity: (1) Loop modeling (targets with sequence identity with the closest template greater than 30% and targets longer than 200 residues with 20-30% sequence identity with the closest template); (2) Loop modeling with constrained all-atom refinement (targets longer than 200 residues with template sequence identity below 20%), and (3) Iterative segment rebuilding and all-atom refinement (targets shorter than 200 residues with template sequence identity below 30%). See (Das et al., 2007) for a complete description of the three protocols.

FRAGFOLD (Jones et al., 2005a; Jones, 2001) is based on the assembly of super-secondary structural fragments obtained from highly resolved protein structures using a simulated annealing approach. This method presents an objective function composed by a set of pairwise potentials of mean force, determined by a statistical analysis of highly resolved X-ray crystallized protein structures and the application of the inverse Boltzmann equation (O'Toole and Dahler, 1960) with a solvation potential and a set of terms that describe the hydrogen network and the steric clashes between atoms of the protein. FRAGFOLD is composed by four basic steps: (1) favorable super-secondary structural fragments at each residue position along the target sequence are selected - the super-secondary structure classification model used by FRAGFOLD is defined as: α-hairpin (consecutive α-helices in a compact arrangement); α-corner (consecutive α - helices in a non-compact arrangement); β-hairpin (hydrogen-bonded consecutive β-strands); β-corner (non-hydrogen-bonded β-strands with intervening α-helix); split β-α-β unit (parallel non-hydrogen-bonded β-strands with intervening α-helix) - the fragment selection involves also the summation of pairs of potential terms and solvation terms for the target sequence onto each super-secondary motif, at each position in the sequence; at the end, a sequence-specific list is generated; (2) a general fragment list is build from all tripeptide, tetrapeptide and pentapeptide fragments from the highly resolved structures; (3) a single folding simulation is executed: (a) a random sequence for the target sequence is generated by selecting fragments entirely randomly, (b) fragments are spliced by superposing the α-carbons and the main chain nitrogen and carboxyl-carbon atoms of the C-terminus of one fragment on the equivalent atoms of the N-terminus of the other fragment, (c) a random confor-

[16] The radius of gyration determines the protein structure compactness.

[17] A solvation layer includes well-ordered sites of hydration around polar and charged protein atom.

mation is generated to each amino acid residue and a steric check is performed, (d) weights for the energy function are calculated, (e) after the weights and the random starting conformation have been determined a simulated annealing approach is used to minimize the energy function; (4) the final structure is selected - for each target sequence, twenty separated simulations using different random number seed values are carried out and the resulting structures are clustered (Kelley et al., 1996). The five most representative populated clusters are assumed with the predicted final structure. FRAGFOLD uses an energy function composed by a pairwise, a solvation, a steric and hydrogen bonding terms.

CABS-Fold (Maciej et al., 2013) is a web server based on CABS (C-Alpha, c-Beta, Side-chain) modelling procedures (Kolinski, 2004). It allows user to choose between performing *ab initio* or template-based modelling. In both cases, CABS-fold sampling is controlled by a Replica Exchange Monte Carlo scheme and the resulting trajectories from the CABS algorithm are clustered using the K-means method. Given a protein sequence, the server workflow that result on a protein model is as follows: CABS modelling, structural clustering, backbone and side-chain reconstruction and model refinement.

SIMFOLD (Chikenjia et al., 2003; Fujitsuka et al., 2006) is a fragment assembly algorithm for protein structure prediction. In the first stage of the basic SIMFOLD algorithm (Chikenjia et al., 2003) fragment candidates are obtained from a given sequence of amino acid residues. These fragments are three-residues long. Contiguous three-residues fragment templates are searched for an exact match in a non-redundant database (SIMFOLD uses the CULLPDB database(Wang and Dunbrack, 2003)) and the resulting fragment templates are stored in a fragment library. The homologous database is also searched to find protein templates for each target fragment (using BLOSUM62 (Henikoff and Henikoff, 1993) scoring over nine residues, which includes an additional three residues in the N-terminal region and in the C-terminal region around the central amino acid fragment). After obtaining the template fragments from the structural database, the algorithm starts a simulation with a random conformation. In this simulation, a move consists of substituting the torsional angles of a randomly chosen candidate in a randomly chosen three-residue fragment for those of the current configuration. Each movement is evaluated using a Metropolis criterion and this procedure is repeated with a decrease in the temperature. SIMFOLD uses a "replicated system" (Chikenjia et al., 2003) to chose the contiguous fragments with high probability. A multicanonical ensemble Monte Carlo (Berg and Neuhaus, 1991) algorithm is used to search the conformational assembly space. SIMFOLD applies an energy function based on physical terms: $V_{tot} = V_{\omega} + V_{\phi} + V_{\psi} + V_{vdw} + V_{HB} + V_{HP} + V_{Rama} + V_{pair}$, where V_{ω}, V_{ϕ} and V_{ψ} are torsion angle potentials, V_{vdw} is the van der Waals interaction, V_{HB} is the Hydrogen bonding term, V_{HP} is the hydrophobic interaction, V_{Rama} represents the secondary structure propensity based on the entropy contribution of the side chain, and V_{pair} denotes pairwise interaction (such as Coulomb interactions)(Fujitsuka et al., 2004). Recent SIMFOLD versions present some optimizations in the energy function where the energetic parameters are optimized using a set of proteins with known X-ray crystal structures (Fujitsuka et al., 2006).

PROFESY (PROFile Enumerating SYstem) (Lee et al., 2004) predicts three-dimensional protein structures that use secondary structure prediction information of the query sequence and the fragment assembly procedure based on global optimization. PROFESY uses the information obtained from the secondary structure prediction method PREDICT (PRofile Enumeration DICtionary) (Joo et al., 2004). For a given sequence of amino acid residues PREDICT, using PSI-BLAST (Altschul et al., 1997), defines patterns for its amino acid residues. Each pattern is a pair of fifteen amino acid residues. Each pattern is compared with those in the PDB and the patterns closest to the query sequence are selected to determine the secondary structure of the query residues. For each amino acid residue in consideration, a fragment library is built and is composed by the backbone dihedral angles of the patterns. The tertiary structure of a given sequence is generated by this library by fragment assembly. The random conformations are built from a N to C-terminal region that selects a random fragment from the fragment library that is related to an amino acid residue from the target sequence. The global energy minimization of the energy function is performed by the Conformational Space Annealing method (CSA) (Lee et al., 1997, 1998). PROFESY energy function includes the number of long-range hydrogen bonds, the radius gyration, the Lennard-Jones, van der Waals interactions of the CHARMM (Brooks et al., 1983; MacKerell et al., 1998a) (available in the TINKER package), force field for avoiding steric clashes, and the accessible surface area solvation energy (Ooi et al., 1987).

CREF(Central-Residue-Fragment-based method)(Dorn and Norberto de Souza, 2008, 2010) is a 3-D PSP method based on short fragments from the PDB. The main goal of CREF is to predict approximate 3-D protein structure that can then be refined through MM techniques. The main characteristic of CREF is that it does not use entire template fragments, but only the ϕ and ψ torsion angles of the main chain of the central amino acid residue of the template fragments obtained from the PDB. Clustering techniques are applied to the template information to identify the Ramachandran plot regions where the central amino acid residues of the template are more concentrated. CREF uses a fixed fragment length equal to five amino acid residues. The clusters identified in the clustering step are labeled with respect to the conformational state indicated by the regions in the Ramachandran plot. The secondary structure of the target sequence is predicted and the approximated conformation is built through a mapping function using the clustering results.

A3N (Artificial Neural Network N-gram-based method) (Dorn and Norberto de Souza, 2010B) is a fragment-based method to predict approximate native-like protein structure from primary sequences of amino acid residues. A3N fragments the target sequence in consecutive amino acid fragments. All fragments with five, seven, nine and eleven amino acid residues are generated. A search procedure in the PDB is performed for each target amino acid fragment. Only the information from the central amino acid residue from the templates is considered for analysis. The structural (torsion angles) information from protein templates is analyzed through a statistical function and the secondary structure of the target sequence is predicted. A clustering algorithm is applied in order to identify similar correlated templates in specific regions of the Ramachandran plot. Each Ramachandran region represents a class of conformational states and torsion angle values. A mapping function is used to create training patterns for each amino acid residue from the target sequence. The training patterns of one amino acid residue are learned using backpropagation in artificial neural networks (d'Avila Garcez et al., 2009; Haykin, 1998; Rumelhart et al., 1986; Garcez et al., 2007; Garcez and Lamb, 2006). The torsion angles ϕ and ψ are predicted for each amino acid residue of the target sequence. The polypeptide structure is then predicted.

QUARK (Xu and Zhang, 2012)[18]focuses on minimizing the two major difficulties regarding to *ab initio* protein structure prediction: Development of a force field and identifying the global energy minimum. To facilitate the force field development and search engine design, QUARK takes a semi-reduced model to represent protein residues by the full backbone atoms (N, Ca, C, O, Cb, H) and the side-chain center of mass. It uses neural networks to predict structural features selected and the global folding is then generated by replica-exchange Monte Carlo simulation by gathering the continuous fragments that were generated using a template library and whose size varies from 1-20 residues. For the first time, a reduced model allows the design of atomic-level force fields such as H-bonding, van der Waals, backbone torsion-angle, and atomic pair-wise interactions. QUARK takes into consideration three levels of packing: the atom, and the residue and the topology level; the latter is particularly essential for the overall folding of proteins in template-free simulations.

There are other prediction methods that use the concept of knowledge-fragments for predicting protein structures: CABS (Kolinski, 2004; Kolinski and Bujinicki, 2005), UNDERTAKER (Karplus et al., 2003), ABLE(Ishida et al., 2003), Fragment-HMM (Li et al., 2008) and ANGLOR (Wu and Zhang, 2008a). Park(Park, 2005) uses a genetic algorithm for fragment assembly to find low-energy conformations. Cutello et al. (Cutello et al., 2006) use a genetic algorithm for solving a multi-objective representation of a protein structure. Table 3 lists the main computational strategies used in the context of this class of methods.

3.3. Fold Recognition and Threading Methods

Fold recognition methods are motivated by the notion that structure is more evolutionary preserved than sequence, i.e., proteins with no apparent sequence similarity could have similar folds (Finkelstein and Ptitsyn, 1987; Levitt and Chothia, 1976; Setubal and Meidanis, 1997; Floudas et al., 2006). Several studies in the last years have indicated that the number of protein structural folds in nature are limited (Richardson, 1981; Li et al., 1996; Wang, 1998). Today, for example, there are approximately ten different folds in fifty percent of the proteins with known structure (Russell and Barton, 1994). The general goal of 3-D protein structure prediction by threading methods is to fit a protein sequence correctly against a structural model. During this procedure the target amino acid sequence is placed, following their sequential order, into structural positions of a template 3-D structure in an optimal way. This involves two basic procedures: (a) selecting a structural model from a library of models and (b) finding the correct replacement between the target sequence against the structural models in the space of possible sequence-structure alignments. Threading methods use structural information such as residue-residue contact patterns,

[18] Quark Online. `http://zhanglab.ccmb.med.umich.edu/QUARK/` (accessed Sep 01, 2014).

Table 3: First principle methods with database information. Main internal computational methods: Clustering Algorithms (CA), Monte Carlo (MC), Simple Profile Alignment Method (PPA), Simulated Annealing (SA), Genetic Algorithms (GA), Multi-canonical Ensemble Monte Carlo (MEMC) with Metropolis criterium, Conformational Space Annealing (CSA), Replica Exchange Monte Carlo (REMC), Hidden Markov Model (HMMS), Artificial Neural Networks (ANN).

FIRST PRINCIPLE METHODS WITH DATABASE	MC	CA	PPA	SA	GA	REMC	CSA	ANN	HMMS	MEMC
Able (Ishida et al., 2003)	•	•		•						
Anglor (Wu and Zhang, 2008a)								•		
A3n (Dorn and Norberto de Souza, 2010B)		•						•		
Cabs (Kolinski, 2004; Kolinski and Bujinicki, 2005)	•	•								
CReF (Dorn and Norberto de Souza, 2008, 2010)		•								
Cutello et al. (Cutello et al., 2006)					•					
Fragment-HMM (Li et al., 2008)	•								•	
Fragfold (Jones, 2001)		•		•	•					
I-Tasser (Wu and Zhang, 2007b; Zhang, 2007, 2008, 2009)	•	•	•							
Park (Park, 2005)					•					
Profesy (Lee et al., 2004)							•			
Rosetta (Rohl et al., 2004; Schueler-Furman et al., 2005)	•			•						
Rosetta@ (Das et al., 2007)	•			•						
Simfod (Chikenjia et al., 2003)										•
Undertaker (Karplus et al., 2003)					•				•	
Quark (Xu and Zhang, 2012)						•		•		
Cabs-Server (Maciej et al., 2013)		•				•				

secondary structure and solvent accessibility, and after identifying the structural similarities, which cannot be detected solely by the similarities between the amino acid sequences, the predicted structural models are constructed.

In threading methods for the 3-D PSP problem it is necessary to solve the problem of sequence-structure replacement, where, given a solved structure $T = t_1, t_2, \ldots, t_n$ and a target sequence $S = s_1, s_2, \ldots, s_m$ the main goal is to find the best match between S and T. Threading methods use known 3-D protein structures as templates for sequences of unknown structures. Threading methods try to identify templates with similar fold with or without direct evolutionary relations (analogue). Homologue proteins are the result of divergent evolution and often share a common function. Analogue proteins do not have a common ancestor and generally do not have a common function. In both cases the proteins share a common three-dimensional structure without a significant sequence similarity (Russell et al., 1998). Comparative modeling usually employs sequence-sequence comparison while threading usually exploits structure information to assist alignment (Zhang, 2009-B). Compared to first principle methods without database information (*Ab initio*), threading methods seek to optimize a potential energy function (an objective or scored function) measuring the fitting quality of a sequence in a particular 3-D configuration. This measure will be assessed using statistical or energetic measurements for the overall likelihood of the target amino acid sequence adopting one of the available structural folds.

In a general form fold recognition methods can be divided in two group (Russell et al., 1998): profile-based (Fischer and Eisenberg, 1996; Rice and Eisenberg, 1997; Rost et al., 1997; Russell et al., 1996) and pair potentials-based (Bryant and Lawrence, 1993; Godzik et al., 1992; Jones et al., 1992; Taylor, 1997). On the first group the information of the structural database containing potential target structures is represented in a linear form or profile. In this case the target protein is matched in turn with this profile. The second group uses pair potentials which score the propensity of two residues being at a certain distance. A threading method typically consists of three components (Smith et al., 1997): (1) construction of a library of potential folds or structural templates; (2) a scoring schema to evaluate any particular placement of a target sequence into each fold; (3) a method to search over the vast space of possible replacements between each sequence and each fold for the best set that gives the best total score. Next, we detail these four components.

(1) Construction of a library of potential folds or structural templates: the library of folds is constructed from known native protein structures derived, for example, from the PDB (Berman et al., 2000). Usually, the 3-D coordinates of a protein structure are reduced to more abstract representations. Structural core elements are defined by the secondary structure elements: β-sheet, α-helix, left handed helix, coil, strands. Frequently, side-chain information is removed. What remains is a backbone template of blank or empty amino acid posi-

tions (Smith et al., 1997).

(2) A scoring schema to evaluate any particular placement of a sequence into each fold: the scoring functions are usually a list of statistical references of each amino acid residue to each structural or fold environment (Smith et al., 1997). These functions describe how favorable a replacement of a query sequence and a template structure are (Jiang et al., 2002). Most threading methods do not use physical full-atom free energy function as used by first principle methods without database information. Most threading objective energy functions are determined empirically by statistical analysis of 3-D data obtained from the PDB. These functions are referred to in general as knowledge-based functions and are used in both profile-based (Fischer and Eisenberg, 1996; Rice and Eisenberg, 1997; Rost et al., 1997; Russell et al., 1996) and pair potentials-based (Bryant and Lawrence, 1993; Godzik et al., 1992; Jones et al., 1992; Taylor, 1997) methods. Different approaches for potential functions can be found: Boltzmann statistics (Sippl, 1995), hydrophobic contact potential (Huang et al., 1996), probability model based on Markov Random Fields (White et al., 1994), logistic regression (Bryant and Lawrence, 1993).

(3) A method to search over the vast space of possible replacements: the use of an algorithm to identify the optimal sequence-structure replacement is essential in a threading method. The main task is to identify the global best score and the optimal fitting/threading. There are at least two main approaches to the sequence-structure replacement: (1) 3-D profile methods (Bowie et al., 1991; Luthy et al., 1992; Alexandrov et al., 1996; Kelley et al., 2000; Shi et al., 2001); and (2) contact potentials (Casari and Sippl, 1992; Bryant and Lawrence, 1993; Sippl et al., 1992; Hendlich et al., 1990). Today most threading methods fall into category 2 above.

3.3.1. Overview

Many threading methods have been developed recently. The most commonly used methods are presented below.

GENTHREADER (Jones, 1999b) performs sequence alignment, followed by the calculation of pair potential and solvation terms, and thus uses an implementation of an artificial neural network in order to evaluate the replacement. The prediction method uses a simplified version of the multi-sequence alignment algorithm (MULTAL (Taylor, 1988)). A sequence profile (Gribskov et al., 1987; Gribskov, 1994) is constructed using a BLOSUM 50 matrix (Henikoff and Henikoff, 1992, 1993). GENTHREADER uses an evaluation function based on a set of pairwise potentials of mean forces (Hendlich et al., 1990) to determine and select the conformation with lowest potential energy. THREADER (Jones et al., 1992) is the first version of GenTHREADER, while the improved GenThereader (McGuffin and Jones, 2003a), mGenTHREADER (Jones et al., 2005b), pGenTHREADER and pDomTHREADER (Lobley et al., 2009a), are later versions. Differently than GENTHREADER, pGenTHREADER and pDomTHREADER use profile-profile alignments and predicted secondary structure as inputs. In pGenTHREADER and pDomTHREADER these features are represented and combined differently, improving the detection of useful templates and the accuracy of sequence-structure alignments. pGenTHREADER is recommend for fold recognition and identification of distant homologues, while that pDomTHREADER is recommended for discriminating super-families.

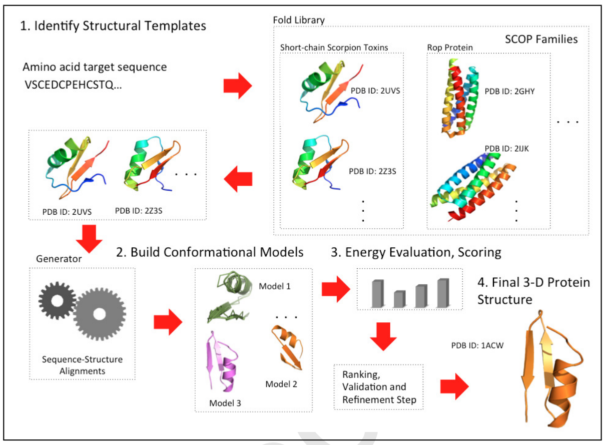

1. Identify Structural Templates
Fold Library
SCOP Families
Short-chain Scorpion Toxins
Rop Protein
Amino acid target sequence
VSCEDCPEHCSTQ...
PDB ID: 2UVS
PDB ID: 2GHY
PDB ID: 2IJK
PDB ID: 2Z3S
PDB ID: 2UVS
PDB ID: 2Z3S
Generator
2. Build Conformational Models
3. Energy Evaluation, Scoring
4. Final 3-D Protein Structure
Model 1
Model 2
Model 3
Sequence-Structure Alignments
Ranking, Validation and Refinement Step
PDB ID: 1ACW

Figure 2: General schematic representation of a threading procedure: template folds are selected from a library of protein structures, models for the target protein are constructed, the potential energy of the structures is calculated and the models are scored, the structures are ranked and validated and, when necessary, the best ranked structure is refined.

123D (Alexandrov et al., 1996) is a threading method that uses a single empirical potential function to map sequences onto structural positions of any of the proposed folds. The empirical scoring function is derived from an analysis of a non-redundant database of known structures by converting relative frequencies into pseudo-energies using a normalization according to the inverse Boltzmann law. After the sequence-structure replacement, the solutions are evaluated and ranked according to their potential and statistical significance. The best replacement is estimated in comparison with the other replacement. 123D uses a fast dynamic programming optimization procedure adapted to CCPs (Contact Capacity Potentials) (Bowie et al., 1991; Ouzounis et al., 1993), mostly for position and secondary structure dependent costs (specially gap costs) to identify similar structures that can be used to model the three-dimensional structure of the target sequence.

ORFEUS (Ginalski et al., 2003b) can be considered a hybrid threading approach. It combines predicted secondary structure information with the information about sequence conservation and variability. The secondary structure information is stored as profile of probabilities (it uses the FFAS strategy (Rychlewski et al., 2000)). The original algorithm uses the PSIPRED algorithm (Jones, 1999B) to predict the secondary structure. However, any other secondary prediction algorithm that produces estimated probabilities for local structures can be used. ORFEUS uses the SCOP classification database to extract protein families and a genetic algorithm implementation to improve the parameters of FFAS.

PROSPECT (Protein Structure Prediction and Evaluation Computer Toolkit) (Kim et al., 2003; Xu and Xu, 2000) uses a scoring function for sequence-structure replacement composed by four terms: (1) mutation term, (2) singleton fitness term, (3) pairwise-contact potential term, and (4) alignment gap penalties. The energy function has the following form: $E_{\text{total}} = \omega_{mutate}E_{mutate} + \omega_{single}E_{single} + \omega_{pair}E_{pair} + \omega_{gap}E_{gap}$. The mutation energy E_{mutate} is the sum of the compatibility measurements $e_{mutate}(a_1, a_2)$ for substituting the template amino acid a_1 by the target amino acid a_2. PROSPECT uses the PAM50 matrix (Gonnet et al., 1992) for calculating E_{mutate}. The singleton energy E_{single} represents the sum of the preferences $E_{single}(a, s, t)$ for aligning amino acid a of the target sequence onto a template position with a structural environment defined by secondary structure s and solvent accessibility[19], or Accessible Surface Area (ASA) t. E_{pair} is the sum of pair-contact potentials $e_{pair}(a_1, a_2)$ between amino acids a_1 and a_2 of the target sequence when they are aligned to template positions that are spatially close. The E_{gap} is the sum of the penalties $e_{gap}(g)$ for an alignment gap of length g (Gonnet et al., 1992). All the ω terms are scaling factors, which are determined by optimizing the threading replacements of the training set against the structure-structure replacements. PROSPECT considers pair contacts only between core residues (α-helix or β-sheet) and alignment gaps only in loop regions. All statistics for estimating the terms in the above equation are collected from FSSP (Holm et al., 1992). The algorithm employs a divide-and-conquer strategy to solve the optimal threading problem. The algorithm solves the entire optimal replacement problem by recursively solving a series of alignment problems between sub-structures and sub-sequences, under various constraints, and then combining these sub-alignments in a consistent and optimal way.

BioShell-Threading (Gniewek et al., 2014) is a versatile tool for the fast and extensive aligning of proteins. The alignment can be based on the two sequences, one sequence and one structure or on the two structures. BioShell-Threading presents a range of special features that approach NP-hard issues, such as 3D structure alignment and threading, using a integrated framework to tackle these problems. It employs Monte Carlo to sample the alignment space so an approximate solution to NP-hard 3D threading problem can be found. In a single run, several secondary structure similarity scores may be used, each of them based on a different secondary structure prediction, the user obtains several alignment, ranked by their score.

FFAS03 server (Xu et al., 2013; Jaroszewski et al., 2011; Rychlewski et al., 2000; Jaroszewski et al., 2005) provides profile-profile alignment using dynamic programming and fold recognition algorithm of fold and function assignment system (FFAS). Released in 2013, FFAS-3D (Xu et al., 2013) brings improvements using a neural network approach for re-ranking templates and adding optimized structural features from both experimental and prediction analyses. These new features have contributed to an 11% increase in the accuracy of the alignment method, as well as a high sucess rate in protein classification in SCOP families.

RaptorX server (Källberg et al., 2012) is a tool for predicting the structure of proteins with no close homologous. It has a profile-entropy scoring method that allows the optimisation of the modelling strategy specifically to the target, it brings a whole new nonlinear threading score function that uses conditional random fields of probabilistic nature to integrate a variety of biological signals. RaptorX uses a multiple-template threading procedure to allow the alignment of the target sequence to any number of existing templates. To sample the alignment space, a dynamic-programming algorithm (revised form of the Smith-Waterman algorithm) is implemented. It predicts the quality of an alignment using a neural network approach to estimate the similarity, measured by TM score (Zhang and Skolnick, 2004b). Furthermore, RaptorX provides a conditional neural fieldbased prediction protocol for determining the secondary structure distribution for each residue in

[19] Solvent Accessibility is the surface area of a biomolecule that is accessible to a solvent.

a target protein. RaptorX also provides domain parsing of long protein and disorder prediction in order to assist the interpretation of the generated results.

Phyre server (Kelley and Sternberg, 2009) gather information from both PDB and SCOP for the construction of a profile and a so called fold library. A profile is also constructed with the given sequence, based on scanning against a non-redundant database. Both close and remote sequence homologous are collected using PSI-BLAST. The profile and the secondary structure which is a consensus of the outputs from Psi-Pred (Buchan et al., 2013), SSPro (Pollastri et al., 2002) and Jnet (Cole et al., 2008) are profile-profile aligned with the fold library and E-value are given and ranked. The top ten E-value results are used to construct 3D structure models of the target sequence. Side-chains are added using a graph-based algorithm and a side-chain rotamer library.

HHpred (Soding et al., 2006a) implements pairwise comparison of profile hidden Markov models (HMM-HMM comparison), being able to improve sensitivity and alignment quality over HMM-sequence alignment comparison. A hidden Markov model (HMM) is a statistical Markov model with hidden states. In a Markov model, the state is visible by the observer, and the state transition probabilities are the parameters. In a hidden Markov model, each state has a probability distribution over the possible outputs. The adjective ”hidden” refers to the state sequence through which the model passes. HMM are being used for alignments and for the detection of homologous. Profile HMM are similar to sequence profiles. However, HMM considers probabilities for insertions and deletions of amino acids along the alignment. It connects to a broad range of databases, such as PDB (Berman et al., 2000), SCOP (Lo Conte et al., 1999), Pfam (Sonnhammer et al., 1998), SMART (Ponting et al., 1999), among others. HHpred is able to generate a sequence-template pairwise alignment, multiple alignment using sundry templates, as well as 3D models calculated by the MODELLER (Eswar et al., 2006) using these alignments.

LOOPP server (Vallat et al., 2009, 2008; Teodorescu et al., 2004) is a homology-based protein structure prediction method. It employs a discriminatory learning decision tree for homology prediction, this tree has four branches, each one implements a different algorithm. The true models are, then, subjected to MODELLER (Eswar et al., 2006; Martí-Renom et al., 2000) for construction and evaluation of atomically detailed structures.

SPARKS-X (Yang et al., 2011) is a fold recognition method that uses a probabilistic approach (Hidden Markov Models) to estimate the probability of a match between predicted and actual 1D structural properties. It uses SPINE-X (Faraggi et al., 2012) to acquire secondary structure information for the models.

Other threading methods can be found in the literature: SEGMER (Wu and Zhang, 2010), THREADER 2 (Jones et al., 1995), ESYPRED3D(Lambert et al., 2002), RAPTOR (RApid Protein Threading predictOR) (Xu et al., 2003,B), LIBRA I (Ota and Nishikawa, 1997), TOPITS (Rost, 1995a,b) and COTH[20]. MUSTER (MUlti-Sources ThreadER) (Wu and Zhang, 2008b) is used to identify template structures from the PDB library. It generates sequence-structure replacements by combining sequence profile-profile alignment with multiple structural information. Turcotte et.al (Turcotte et al., 1998, 2001,B,C) apply Inductive Logic Programing (ILP) to discover rules that govern the three-dimensional topology of protein structure. Xu et. al. (Xu et al., 1998) developed an algorithm that solves the globally optimal threading problem efficiently. Table 4 summarizes the main computational strategies used in the context of the main threading methods.

One of the most recent advancements in the field of 3-D protein structure prediction and threading methods is the idea of meta-strategies or meta-serves (Bujnicki et al., 2001). This idea is related to the concept of consensus-based approach (Lundstrom et al., 2001). As shown in the last 9 editions of CASP (Moult et al., 2014, 2009, 2007, 2005, 2003, 2001, 1997, 1999, 1995) there is no method that is always the best in the predictions. This occurs because the quality of the predictions depends on many factors which are unknown when the prediction is run. In meta-servers all prediction methods are applied to a given sequence; a computational strategy, such as ANNs (Artificial Neural Networks) are used in 3D-Judge (Jaskowski et al., 2007), are then applied in order to choose the most realistic prediction. The meta-servers approach has many advantages: (1) as shown in the CASP experiments, meta-servers produce generally better results than individual servers; (2) 3-D protein structure prediction prediction in meta-serves are more stable than those made when only a single prediction method is used. Meta-server approaches represent one of the most significant advances in the field of protein structure prediction problem. Currently, 3D-Jury (Ginalski et al., 2003a) is one of the most popular meta-servers. It computes structure similarities between models using

[20] COTH: CO-THreader. `http://zhanglab.ccmb.med.umich.edu/COTH/` (accessed Sep 2, 2014).

a MaxSub measure (Siew1 et al., 2000) and chooses the most realistic one as the predicted final result. Other examples of meta-servers can be found in the literature: 3D-Judge (Jaskowski et al., 2007), LOMETS (Wu and Zhang, 2007a), STRUCLA (Sasin et al., 2003), Pcons.net (Wallner et al., 2007), ProCKSi (Barthel et al., 2007) and TASSER (Zhou and Skolnick, 2007, 2009; Zhou et al., 2009). A good review of meta-servers can be found in Fischer's work (Fischer, 2006).

3.4. Comparative Modelling Methods and Sequence Alignment Strategies

In comparative modeling a target sequence of amino acid residues (target protein) is aligned against the amino acid sequence of another protein with known structure (template protein) and stored in the PDB (Berman et al., 2000). If the target sequence is similar to the sequence of the template protein, the structural information obtained from the known structure is used for modeling the target protein (McLachlan, 1992; Bajorath et al., 1994; Blundell et al., 1987; Johnson et al., 1994; Sali, 1995; Sánchez and Sali, 1997; Peitsch, 1996). The main idea of this kind of method is to construct an atomic-resolution model of the target protein from its amino acid sequence and an experimental 3-D structure of a related homologous protein. Comparative modeling can be applied whenever it is possible to detect an evolutionary relationship between the target protein and the template protein of which the 3D structure is known (Martí-Renom et al., 2000). The evolutionary relationship between proteins is a fundamental factor in comparative modeling methods and the target protein can be modeled from homologous proteins with 3-D structures determined experimentally (Sternberg, 1997). The structure of these proteins are similar in the sense that amino acid residues with identical physico-chemical properties occupy the same position in homologous proteins.

Currently, comparative modeling methods can achieve 3-D protein structures with high accuracy and are also used in the field of drug design, virtual screening, and site direct mutagenesis (Koop and Schwede, 2004). The quality of the comparative modeling methods depends on the quality of the sequence alignment methods. The sequence alignment is used to produce a structural model of the target sequence. There are two main classes of methods used as sequence alignment strategies (Martí-Renom et al., 2000):

Sequence-sequence comparison (pairwise): this class includes the methods that compare the target sequence with each candidate sequence in the database independently (Apostolico and Giancarlo, 1998). FASTA (Pearson and Lipman, 1988; Lipman and Pearson, 1985), PSI-BLAST (Altschul et al., 1997) and BLAST (Altschul et al., 1990) are examples of methods used in sequence-sequence comparison.

Multiple sequence comparison: they perform multiple sequence alignments (Notredame, 2002; Thompson et al., 1999; Wallace et al., 2005; Notredame, 2007) to improve the sensitivity of the search (Gribskov, 1994; Krogh et al., 1994; Altschul et al., 1997; Henikoff and Henikoff, 1994). CLUSTALW (Thompson et al., 1994), PSI-BLAST (Altschul et al., 1997) and T-COFFEE(Notredame et al., 2000) are examples of multiple sequence alignment methods. T-COFFEE (Notredame et al., 1998, 2000) provides a simple and flexible means of generating multiple alignments using heterogeneous data sources through a library of pairwise alignments that use a structure similar to the one presented in Notredame et.al (Notredame et al., 1998). T-COFFEE uses a progressive strategy (Feng and Doolittle, 1987; Taylor, 1988; Thompson et al., 1994) (dynamic programming) to find the best multi-alignment. It uses the information in the library of the pairwise alignments to carry out progressive alignment in a way that considers the alignments between all pairwise alignments, while each step of the progressive multi-alignment is executed. In the progressive alignment, pairwise alignments produce a distance matrix between all the sequences, which in turn is used to produce a guide tree using the neighbor-joining method of Saitou and Nei (Saitou and Nei, 1987).

For a given a protein sequence, the comparative modeling procedure requires the identification of homologous sequences with known structure, alignment of the query against the template sequences, construction of the 3-D models and refinement at the final stage. Martí-Renom and Sanchez (Martí-Renom et al., 2000; Sánchez and Sali, 1997) enumerate four basic steps of a comparative modeling procedure: (1) fold assignment and template selection, (2) template target alignment, (3) model building, and (4) model evaluation and refinement. Initially, sequences similar to the target sequence are collected using search engines over a database (fold assignment and template selection). Templates can be found searching in structural databases such as the PDB (Berman et al., 2000). The four basic steps are detailed as follows.

Fold assignment and template selection: the starting point in a Comparative Modeling method is to identify all protein structures with sequences related to the target sequence, then to select templates that will be

Table 4: Threading methods. Main internal computational methods: Linear Programming (LP), Genetic Algorithms (GA), Hidden Markov Models (HMMS), Artificial Neural Networks (ANN), Dynamic Programming (DP), Divide and Conquer (DC), Profile-analysis (PA), Screening Techniques (ST), Sequence-based (SEB), Monte Carlo(MC), Structure-based (STU), Clustering Algorithms (CA), Suport Vector Machines (SVM), Discriminatory Learning (DL) and Inductive Logic Programing (ILP).

Method	ANN	DP	GA	HMMS	LP	DC	PA	ST	SEB	STU	ILP	MC	CA	SVM	DL
123D (Alexandrov et al., 1996)		•													
EsyPred3D (Lambert et al., 2002)								•	•						
Ffas (Rychlewski et al., 2000)							•		•						
Genthareader (Jones, 1999b)	•									•					
3DPssm (Kelley et al., 2000)		•								•					
Libra I (Ota and Nishikawa, 1997)							•			•					
Muster (Wu and Zhang, 2008b)		•													
Orfeus (Ginalski et al., 2003b)			•						•						
Prospect (Xu and Xu, 2000)						•				•					
Raptor (Xu et al., 2003,B)					•					•					
Segmer (Wu and Zhang, 2010)				•											
Threader2 (Jones et al., 1995)	•														
Topits (Rost, 1995a,b)	•	•							•						
Turcotte et. al. (Turcotte et al., 1998, 2001)											•				
Loopp (Vallat et al., 2009, 2008)															•
RaptorX Server (Källberg et al., 2012)													•	•	
Phyre Server (Kelley and Sternberg, 2009)															
SparksX (Yang et al., 2011)				•											
BioShell-Threading (Gniewek et al., 2014)												•			
Ffas03 Server (Xu et al., 2013)	•	•													
HHPred (Soding et al., 2006a)				•											

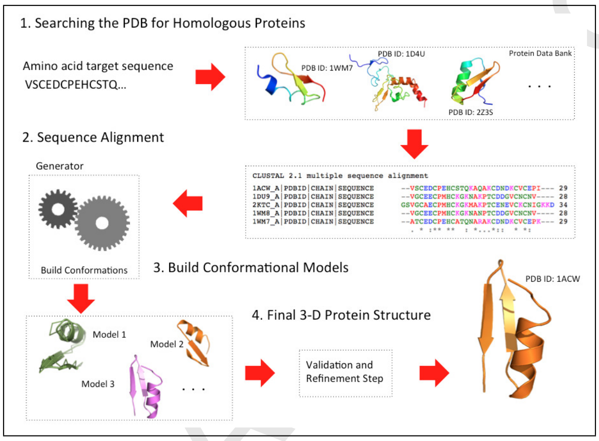

1. Searching the PDB for Homologous Proteins
Amino acid target sequence
VSCEDCPEHCSTQ...
PDB ID: 1WM7
PDB ID: 1D4U
Protein Data Bank
PDB ID: 2Z3S
. . .
2. Sequence Alignment
Generator
Build Conformations
CLUSTAL 2.1 multiple sequence alignment
1ACW_A|PDBID|CHAIN|SEQUENCE --VSCEDCPEHCSTQKAQAKCDNDKCVCEPI--- 29
1DU9_A|PDBID|CHAIN|SEQUENCE --VGCEECPMHCKGKNAKPTCDDGVCNCNV---- 28
2KTC_A|PDBID|CHAIN|SEQUENCE GSVGCAECPMHCKGKMAKPTCENEVCKCNIGKKD 34
1WM8_A|PDBID|CHAIN|SEQUENCE --VGCEECPMHCKGKNANPTCDDGVCNCNV---- 28
1WM7_A|PDBID|CHAIN|SEQUENCE --ATCEDCPEHCATQNARAKCDNDKCVCEPK--- 29
3. Build Conformational Models
PDB ID: 1ACW
4. Final 3-D Protein Structure
Model 1
Model 2
Model 3
. . .
Validation and Refinement Step

Figure 3: Schematic representation of a typical process of comparative modeling by homology. Initially, template proteins are identified. Then the sequence of the target protein is aligned against the sequence of the protein-templates, and then a model is built and validated, obtaining in the end, the 3-D structure of the target protein. If necessary, the final structure may undergo a refinement process.

used as templates. There are numerous protein sequence and structure databases and database scanning software (Altschul et al., 1994; Holm et al., 1992). Templates can be found using the target sequence as a query for searching structure databases such as the PDB (Berman et al., 2000).

Template target alignment: in the alignment step the sequence of the target protein is aligned with sequence(s) of protein(s) with known structure(s). It forms the base model. There are other methods that are usually tuned for detection of remote relationships (Martí-Renom et al., 2000; Baxevanis, 1998; Holm and Sander, 1996; Smith, 1999; Taylor, 1996). In these methods, non-optimal alignments are exploited. Pairwise sequence alignment methods (Apostolico and Giancarlo, 1998) are used to find the best-matching local or global alignments of two sequences. Pairwise alignments can only be used between two sequences at a time. The three primary methods of producing pairwise alignments are dot-matrix methods and dynamic programming. Multiple sequence alignment (Lipman et al., 1989; Hirosawa et al., 1995; Grasso and Lee, 2004; Kim et al., 1994; Edgar, 2004; Brudno et al., 2003; Notredame, 2002; Thompson et al., 1999; Wallace et al., 2005; Notredame, 2007) is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set.

Model building: when building the model of the protein, it is common that, first, all the backbone from the homologous regions is constructed, then the different regions, loop regions and, finally, the side chains (Vásquez, 1996). A variety of methods can be used to construct the 3-D model of the target protein. These methods are divided into three groups: modeling by assembly of rigid bodies (Blundell et al., 1987; Greer, 1990), modeling by segment matching or coordinate reconstruction (Levitt, 1992; Jones and Thornton, 1997B), and modeling by satisfaction of spatial restraints (Sali and Blundell, 1993; Srinivasan et al., 1993; Aszódi and Taylor, 1996).

Model evaluation and refinement: the evaluation of the final model takes into account all available information of the target protein (Tramontano, 2006). According to Baxevanis (Baxevanis and Quellette, 1990), the most critical step in homology modeling is the alignment. A misalignment can have a distorting effect on the other steps, generating a distorted and incorrect final structural model.

The actual comparative modeling process is more complex. As described above, many computational strategies are employed by various prediction servers. Comparative modeling is probably the most used method in protein structure prediction for two main reasons (Tramontano, 2006): (1) the quality of the predicted models - when a reasonable evolutionary relationship is present then the accuracy of the predicted models is greater than those produced with other techniques; (2) the reliability of the model can be estimated a priori and the quality of the predicted structures can be estimated. In the last years considerable progress has been made in *Ab initio* protein structure prediction methods; however, comparative modeling is a very precise and accurate prediction method (Koehl and Levitt, 1999; Martí-Renom et al., 2000). Despite the high quality predictions, comparative modeling by homology has some limitations. The first limitation concerns the inability to perform prediction of new folds. This is explained by the fact that this methodology can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures in the PDB. The second limitation is that it is not possible to study the folding process of the protein, i.e., the path that an unfolded protein traverses to the functional state (native state).

3.4.1. Overview

Many Comparative Modeling methods and sequence alignment strategies were developed over the last years. The main computational methods are listed below.

SWISS-MODEL (Arnold et al., 2006; Kiefer et al., 2009; Biasini et al., 2014) is a web-based integrated service dedicated to protein structure homology modeling. It employs an automated, knowledge-based protein modeling tool: ProMod (Peitsch and Jongeneel, 1993; Peitsch, 1996). SWISS-MODEL presents three types of modeling modules: (1) automated mode, (2) alignment mode and (3) project mode. The first is computed by the SWISS-MODEL server homology pipeline (Schwede et al., 2003). This module is used in cases where the target sequence similarity is sufficiently high to allow for a fully automated mode. In alignment mode, the submitted alignment is matched against the sequence of the template structure extracted from the SWISS-MODEL template library[21]. A rigid fragment assembly modeling and heuristics are used to improve the placement of insertions and deletions based on the structural context. In the project mode, the correct alignment between target and template cannot be clearly determined by sequence-based methods. Further, visual inspection and manual manipulation of the alignment are used (Bates et al., 2001). SWISS-MODEL provides access to a set of increasingly complex and computationally demanding methods for templates searching: BLAST (Altschul et al., 1997), Interactive profile BLAST (Altschul et al., 1997) that uses information from the NR (non-redundant) database, HMM-based template library search that uses a library of Hidden Markov Models, where each model of the library was created from multiple sequence alignment generated by iterative search of NR databases using SAM-T2K (Hughey and Krogh, 1996).

MODELLER (Eswar et al., 2006; Martí-Renom et al., 2000) is a computing system for comparative protein structure modeling that incorporates a large set of functionalities. In the most simple case, MODELLER is used to calculate a model containing all non-hydrogen atoms with only the input of an alignment of a sequence with the template structures and the atomic coordinates of the templates. In other cases, MODELLER can perform fold assignment alignment of two protein sequences (Eswar et al., 2003) or their profiles (Martí-Renom et al., 2004), comparative structure modeling by satisfaction of spatial restraints, multi alignment of sequences and/or structures (Madhusudhan et al., 2006), calculation of phylogenetic trees (Fitch and Margoliash, 1967), and *de novo* modelling of loops in

[21] The template structure database is derived from the PDB.

proteins (Fiser et al., 2000). A 3D model is obtained by optimization of a molecular Probability Density Function (PDF). In order to optimize this function methods of conjugate gradient and Molecular Dynamics with simulated annealing are employed.

ReformAlign (Lyras and Metzler, 2014) is a profile-based meta-alignment approach that requires a initial alignment-for which the previous cited methods can be used. This method works using a non-probabilistic profile constructed based on an existing alignment and then obtaining a reformed alignment using dynamic programming to realign all the sequences against the profile. The employment of ReformAlign may often increase the accuracy of the alignment of the popular aligners.
PyMOD (Bramucci et al., 2012) is a ”easy-to-use” tool that integrate some of the most common tools for both sequence alignment and homology modelling. For database search, it uses PSI-Blast (Altschul et al., 1990), the user can decide between sequence-using MUSCLE (Edgar, 2004) or ClustalW (Larkin et al., 2007; Higgins and Sharp, 1988) or structural alignment-using CE aling (Guda et al., 2004; Prlic et al., 2010), then it performs a structure-based multiple sequence alignment followed by homology modelling using MODELLER (Eswar et al., 2006) and for the visualization of the final model it is integrated with PYMOL (Schrödinger, LLC, 2010). PyMod can input and output sequences and 3D-structures in the FASTA and PDB formats.
TIP-STRUCTFAST (STructure Realization Utilizing Cogent Tips From Aligned Structural Templates) (Debe et al., 2006), MULTALIN (Corpet, 1988), COMPASS (Sadreyev and Grishin, 2003), HHPRED (Sding, 2005; Soding et al., 2005),3DPSSM (Kelley et al., 2000), FUGUE (Shi et al., 2001), SAM-T99 (Karplus et al., 1992), SAM-T02 (Karplus et al., 2001) and CLUSTALW (Larkin et al., 2007; Higgins and Sharp, 1988; Thompson et al., 1994) are examples of other comparative modeling methods found in the literature. Table 5 lists the main computational strategies used in the context of comparative modeling methods for the 3-D PSP problem. These methods are also classified into three groups according to the type of structural information used and the strategy used to build the polypeptide structures: modeling by assembly of rigid body; modeling by segment matching or coordinate reconstruction and modeling by satisfaction of spatial restraints.

Methods such as HHpred (Soding et al., 2006a), HHsenser (Soding et al., 2006b) and HMMER (Finn et al., 2011) implement pairwise comparison of profile hidden Markov models (HMM-HMM comparison), being able to improve sensitivity and alignment quality over HMM-sequence alignment comparison. A hidden Markov model (HMM) is a statistical Markov model with hidden states. In a Markov model, the state is visible by the observer, and the state transition probabilities are the parameters. In a hidden Markov model, each state has a probability distribution over the possible outputs. The adjective ”hidden” refers to the state sequence through which the model passes. HMM are being used for alignments and for the detection of homologous. Profile HMM are similar to sequence profiles. However, HMM considers probabilities for insertions and deletions of amino acids along the alignment.

3.5. Ensemble Protein Structure Prediction Methods

Recently, particularly given the wide range of state-of-the-art tools already implemented, is becoming increasingly common to combine tools that are part of more than one of the above mentioned subdivisions aiming to improve the outcome of structure prediction protein. PSIPRED Protein Analysis Workbench (Buchan et al., 2013) is a server incorporated with a range of protein analysis methods, some are described hereinafter: PSIPRED (Jones, 1999b) for secundary structure prediction; MEMSAT-SVM/MEMSAT3 (Nugent and Jones, 2009) which is a support vector machines based tool for transmembrane topology prediction; GenTHREADER (Jones, 1999a; McGuffin and Jones, 2003b) for fold recognition as well as its more sensitive version, pGenTHREADER (Lobley et al., 2009b); pDomTHREADER (Lobley et al., 2009b) for homologous domain recognition; BioSerf for homology and *ab initio* modelling; HSPred (Lise et al., 2011) is a method for predicting residues responsible for protein-protein interaction; FFPred (Lobley et al., 2008) that uses support vector machines for function prediction and the HADOOP[22] packages, allowing high-throughput analysis. I-TASSER/QUARK (Zhang, 2014) pipeline is able to perform either template-free and template-based modeling. This novel pipeline gathers the outcome of QUARK *ab initio* approach to use as probes for the LOMETS (Local-Meta-Threading-Server) (Wu and Zhang, 2007b) threading method. The assembly refinement is carried out by I-TASSER. Other examples of ensemble protein structure prediction methods are GalaxyTBM (Ko et al., 2012), AMPLE (Bibby et al., 2012).

[22] `http://hadoop.apache.org`, accessed on 02 Sep 2014.

Table 5: Comparative Modeling Methods Summary. Main internal computational methods: Conjugate Gradient (CG), Molecular Dynamics (MD), Simulated Annealing (SA), Hidden Markov Model (HMMS), Clustering Algorithm (CL), Genetic Algorithms (GA), Dynamic Programming (DP), Modeling by Assembly of Rigid Body (RB), Modeling by Segment Matching or Coordinate Reconstruction (SM), Modeling by Satisfaction of Spatial Restraints (SR).

Method	CG	MD	SA	HMMS	CA	DP	RB	SM
SWISS-MODEL (Arnold et al., 2006; Kiefer et al., 2009)				•			•	•
MODELLER (Eswar et al., 2006; Martí-Renom et al., 2000)	•	•	•					•
T-COFFEE (Notredame et al., 1998, 2000)						•		•
CLUSTALW (Larkin et al., 2007; Higgins and Sharp, 1988)						•		
TIP-STRUCTFAST (Debe et al., 2006)						•		•
MULTALIN (Corpet, 1988)					•			•
COMPASS (Sadreyev and Grishin, 2003)						•		•
HHPRED (Sding, 2005; Soding et al., 2005)				•				
SAM-T02 (Karplus et al., 2001)				•				
SAM-T99 (Karplus et al., 1992)				•				
Phyre (Kelley et al., 2000)						•		
FUGUE (Shi et al., 2001)						•		
ReformAlign (Lyras and Metzler, 2014)						•		
PyMOD (Bramucci et al., 2012)						•		

4. Conclusions

The study of protein structure and the prediction of their three-dimensional (3-D) structures is one of the key research problems in Structural Bioinformatics. Predicting the three-dimensional structure of a protein that has no templates in the Protein Data Bank is a very hard and sometimes virtually intractable task. Over the last years, many computational methods, systems and algorithms have been developed with the purpose of solving this complex problem. However, the problem still challenges biologists bioinformaticians, chemists, computer scientists, and mathematicians because of the complexity and high dimensionality of the protein conformational search space.

Experimentally, the generation of a protein sequence is considerably easier than the determination of its 3-D structure. However, the knowledge of the 3-D structure of the polypeptide gives researchers very important information about the function of the protein in the cell. The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the Genome Projects[23] and the number of 3-D structures of proteins which are known nowadays. These figures not only clearly illustrate the need for, but also motivate further research in Computational Protein Structure Prediction Methods.In addition, this analysis demonstrates the importance of the development of accurate computational methods that can compute and predict the 3-D structure of proteins when only their amino acid sequence is known. We have presented several computational techniques that have been widely applied in the context of the 3-D PSP problem. All these techniques present good results in specific case studies. However, there is still a paucity of general methods applicable to all classes of proteins, or to very large amino acid sequences.

The classification of the prediction methods into four classes, (1) first principle methods without database information; (2) first principle methods with database information; (3) fold recognition and threading methods and (4) comparative modeling methods and sequence alignment strategies - gives a more general view about which methods can be used in the prediction, how experimental data can be used in the prediction tasks, and how a protein conformation can be represented in terms of physical and chemical laws (in the protein folding process). Knowledge-based methods are limited to experimental data, e.g., homology modeling can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structure. Fold recognition via threading is limited to the fold library derived from the PDB structure database. *Ab initio* methods can obtain new structures with novel folds. However, the complexity and high dimensionality of the conformational search space even for a small protein molecule still makes the problem intractable.

[23] DOE Genomic Science. `http://genomics.energy.gov` (accessed Sep 01, 2014).

Over the last years, probably the most important results in this field were produced by hybrid methods such as the ones based on first principles with database information. Such hybrid methods combine the accuracy of knowledge-based methods with a more realistic, force field-based, physicochemical description of a protein. The last results presented in the CASP competition corroborate this statement. ROSETTA, FRAGFOLD, I-TASSER and LINUS all belong to this class of methods. ROSETTA and I-TASSER have been the most successful predictors over the last years according to data from the CASP experiments. In the last CASP, the bioinformatics community focused on the problem of predicting the local and global regions of the 3-D model when experimental structural data are not available. Machine learning techniques, statistical potentials, physical energy functions have been applied in order to find accurate structures.

Finally, Protein Structure Prediction is a very difficult problem and further research remains to be done. The development of new strategies, the adaptation and investigation of new methods and the combination of existing and state-of-the-art computational methods and techniques to the 3-D PSP problem are clearly needed. Understanding how experimental data can be better used in combination with *Ab initio* techniques is another open research question. In summary, there are several research opportunities and avenues to be explored in this field, with relevant multidisciplinary applications in computer science, bioinformatics, chemistry, biochemistry, and the medical sciences.

5. ACKNOWLEDGEMENTS

This work was supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9) - Brazil.

References

Abagyan, R., Totrov, M., 1994. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235 (3), 983.

Alexandrov, N., Nussinov, R., Zimmer, R., 1996. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. World Scientific, Singapore.

Altman, R. B., Dugan, J. M., 2005. Defining bioinformatics and structural bioinformatics. John Wiley and Sons, Inc., Hoboken.

Altschul, S., Boguski, M., Gish, W., Wootton, J., 1994. Issues in searching molecular sequence databases. Nat. Genet. 6, 119.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389.

Anderson, D., 2004. BIONC: A system for public-resource computing and storage. IEEE Computer Society, Washington.

Anderson, J., Travesset, A., 2008. Molecular dynamics on graphic processing units: Hoomd to the rescue. Comput. Sci. Eng. 10 (6), 6.

Andreoni, W., Curioni, A., 2000. New advances in chemistry and materials science with cpmd and parallel computing. Parallel Comput. 26, 819.

Anfinsen, C., 1973. Principles that govern the folding of protein chains. Science 181 (96), 223.

Anfinsen, C., Haber, E., Sela, M., White, F. H. J., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 47, 1309.

Apostolico, A., Giancarlo, R., 1998. Sequence alignment in molecular biology. J. Comput. Biol. 5 (2), 173.

Arnold, K., Bordoli, L., Kopp, J., Schwede, T., 2006. The swiss-model workspace: A web-based environment for protein structure homology modeling. Bioinformatics 22 (2), 195.

Arora, N., Jayaram, B., 1998. Energetics of base pairs in b-dna in solution: An appraisal of potential functions and dielectric treatments. J. Phys. Chem. B 102, 6139.

Aszódi, A., Taylor, W. R., 1996. Homology modeling by distance geometry. Folding Des. 1 (5), 325.

Bahamish, H., Abdullah, R., Salam, R., 2009. Protein tertiary structure prediction using artificial bee colony algorithm. Institute of Electrical and Electronics Engineers, New York.

Bajorath, J., Stenkamp, R., Aruffo, A., 1994. Knowledge-based model building of proteins: concepts and examples. Protein Sci. 2 (11), 1797.

Barthel, D., Hirst, J., Blazewicz, J., Burke, E., Krasnogor, N., 2007. Procksi: a decision support system for protein (structure) comparison, knowledge, similarity and information. BMC Bioinf. 8, 416.

Bates, P. A., Kelley, L. A., MacCallum, R. M., Sternberg, M., 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3d-jigsaw and 3d-pssm. Proteins: Struc., Func. Gen. 5, 39.

Baxevanis, A., 1998. Practical aspects of multiple sequence alignment. Methods Biochem. Anal. 39, 172.

Baxevanis, A., Quellette, B., 1990. Bioinformatics: A practical guide to the analysis of genes and proteins, 2nd Edition. John Wiley and Sons, Inc., New York.

Ben-David, M., Noivirt-Brik, O., Prilusky, J., Sussman, J., Levy, Y., 2009. Assessments of casp8 structure predictions for template free targets. Proteins: Struct., Funct., Bioinf. 77 (9), 50.

Berg, B. A., Neuhaus, T., 1991. Multi-canonical algorithms for first order phase transitions. Phys. Lett. B 267 (2), 249.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bath, T., Weissig, H., Shindyalov, I., Bourne, P., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235.

Best, R. B., Zhu, X., Shim, J., Lopes, P., Mittal, J., Feig, M., MacKerell, A., 2012. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. J. Chem. Theory Comput. 8 (9), 3257–3273.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T., Bertoni, M., Bordoli, L., Schwede, T., 2014. Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 12, 252–258.

Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D., Rigden, D. J., 2012. Ample: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models. Acta Crystallogr., Sect. D: Biol. Crystallogr. 68 (12), 1622–1631.

Blundell, T., Sibanda, B., Sternberg, M., Thornton, J., 1987. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326, 347.

Boas, F. E., Harbury, P. B., 2007. Potential energy functions for protein design. Curr. Opin. Struct. Biol. 17 (2), 199.

Bonneau, R., Baker, D., 2001. Ab initio protein structure prediction: progress and prospects. Annu. Rev. Biophys. Biomol. Struct. 30, 173.

Bowie, J. U., Eisenberg, D., 1994. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and empirical guiding fitness function. Proc. Natl. Acad. Sci. U.S.A. 91 (10), 4436.

Bowie, J. U., Luthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253 (5016), 164.

Bradley, P., Misura, K., Baker, D., 2005. Toward high-resolution de novo structure prediction for small proteins. Science 309 (5742), 1868.

Bramucci, E., Paiardini, A., Bossa, F., Pascarella, S., 2012. Pymod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within pymol. BMC Bioinformatics 13 (Suppl 4), S2–S7.

Branden, C., Tooze, J., 1998. Introduction to protein structure, 2nd Edition. Garland Publishing Inc., New York.

Breda, A., Santos, D., Basso, L., Norberto de Souza, O., 2007. Ab initio 3-d structure prediction of an artificially designed three-a-helix bundle via all-atom molecular dynamics simulations. Genet. Mol. Res. 6 (2), 901.

Brooks, R., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., Karplus, M., 1983. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4 (2), 187.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., Morgenstern, B., 2003. Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinf. 4 (66), 1.

Bryant, S., Lawrence, C., 1993. An empirical energy function for threading protein sequence through the folding motif. Proteins: Struc., Func. Gen. 16 (1), 92.

Bryant, S. H., Altschul, S., 1995. Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. 5 (2), 236.

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., Jones, D. T., 2013. Scalable web services for the psipred protein analysis workbench. Nucleic Acids Research 41 (W1), W349–W357.

Bujnicki, J., 2006. Protein structure prediction by recombination of fragments. ChemBioChem 7 (1), 19.

Bujnicki, J., Elofsson, A., Fischer, D., Rychlewski, L., 2001. Structure prediction meta server. Bioinformatics 17, 750.

Canutescu, A., Shelenkov, A., Dunbrack, R. J., 2001. A graph-theory algorithm for rapid protein side chain prediction. Proteins: Struc., Func. Gen. 12 (9), 2001.

Casari, G., Sippl, M. J., 1992. Structure-derived hydrophobic potential. hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. J. Mol. Biol. 224 (3), 725.

Case, D., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz, K., Onufriev, A., 2005. The amber biomolecular simulation program. J. Comput. Chem. 26 (16), 1668.

Chen, H., Zhou, H., 2005. Prediction of solvent accessibility and sites of deleterious mutation from protein sequence. Nucleic Acids Res. 33 (10), 3193.

Cheng, J., 2008. A multi-template combination algorithm for protein comparative modeling. BMC Struct. Biol. 8 (18), 1.

Chikenjia, G., Fujitsukab, Y., Takadac, S., 2003. A reversible fragment assembly method for de novo protein structure prediction. J. Chem. Phys. 119 (13), 6895.

Chivian, D., Robertson, T., Bonneau, R., Baker, D., 2003. Ab initio methods. Methods Biochem. Anal. 44, 547.

Christen, M., Hunenberger, P., Bakowies, D., Baron, R., Burgi, R., Geerke, D., Heinz, T., Kastenholz, M., Krutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., van Gunsteren, W., 2005. The gromos software for biomolecular simulation: Gromos05. J. Comput. Chem. 26 (16), 1719.

Clote, P., Backofen, R., 2000. Computational molecular biology: An introduction, 1st Edition. John Wiley and Sons, Inc, West Sussex.

Cole, C., Barber, J. D., Barton, G. J., 2008. The jpred 3 secondary structure prediction server. Nucleic Acids Res. 36 (suppl 2), W197–W201.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K. J., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., Kollman, P., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117 (19), 5179.

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16 (22), 10881.

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., Tramontano, A., 2009. Evaluation of template-based models in casp8 with standard measures. Proteins: Struct., Funct., Bioinf. 77 (9), 18.

Creighton, T. E., 1990. Protein folding. Biochem. J. 270, 1.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M., 1998. On the complexity of protein folding. J. Comput. Biol. 5 (3), 423.

Cutello, V., Narzisi, G., Nicosia, G., 2006. A multi-objective evolutionary approach to the protein structure prediction problem. J. R. Soc., Interface 3 (6), 139.

Czaplewski, C., Kalinowski, S., Liwo, A., Scheraga, H., 2009. Application of multiplexed replica exchange molecular dynamics to the unres force field: Tests with alpha and alpha+beta proteins. J. Chem. Theory Comput. 5 (3), 627.

Dandekar, T., Argos, P., 1992. Potential of genetic algorithms in protein folding and protein eng. simulations. Protein Eng. 5 (7), 637.

Dandekar, T., Argos, P., 1994. Folding the main chain of small proteins with the genetic algorithm. J. Mol. Biol. 236 (3), 844.

Darden, T., York, D., Pedersen, L., 2009. Particle mesh ewald: An n.log n method for ewald sums in large systems. The J. Chem. Phys. 98 (12), 10089.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M., Bhat, D., Chivian, D., Kim, D., Sheffler, W., Malmstroem, L., Wollacott, A., Wang, C., Andre, I., Baker, D., 2007. Structure prediction for casp7 targets using extensive all-atom refinement with rosetta@home. Proteins: Struc., Func. Gen. 68 (S8), 118.

d'Avila Garcez, A., Lamb, L., Gabbay, D., 2009. Neural-Symbolic Cognitive Reasoning, 1st Edition. Springer, New York.

Debe, D., Danzer, J., Goddard, W., Poleksic, A., 2006. Structfast: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. Proteins: Struc., Func. Gen. 64 (4), 960.

Dembo, R., Steihaug, T., 1983. Truncated-newton algorithms for large-scale unconstrained optimization. Math. Program. 26, 190.

Derreumaux, P., 1999. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. J. Chem. Phys. 111 (5), 2301.

Derreumaux, P., Zhang, G., Schlick, T., Brooks, B., 1994. A truncated newton minimizer adapted for charmm and biomolecular applications. J. Comput. Chem. 15 (5), 532.

Dewar, M., 1983. Development and status of mindo/3 and mndo. J. Mol. Struct. 100, 41.

Dhingra, P., Jayaram, B., 2013. A homology/ab initio hybrid algorithm for sampling near-native protein conformations. J. Comput. Chem. 34 (22).

Dorn, M., Breda, A., Norberto de Souza, O., 2008. A hybrid method for the protein structure prediction problem. Lect. Notes Bioinf. 5167, 47.

Dorn, M., Norberto de Souza, O., 2008. CReF: A central-residue-fragment-based method for predicting approximate 3-D polypeptides structures. ACM, New York.

Dorn, M., Norberto de Souza, O., 2010. Mining the protein data bank with cref to predict approximate 3-d structures of polypeptides. Int. J. Data Min. and Bioin. 4 (3), 281.

Dorn, M., Norberto de Souza, O., 2010B. A3n: an artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction. Expert Syst. Appl. 37 (12), 7497.

Dunbrack, R. J., Karplus, M., 2003. Backbone-dependent rotamer library for proteins: application to side-chain prediction. J. Mol. Biol. 230 (2), 543.

Dunker, A., Lawson, J., Brown, C., Williams, R., Romero, P., Oh, J., Oldfield, C., Campen, A., Ratliff, C., Hipps, K., Ausio, J., Nissen, M., Reeves, R., Kang, C., Kissinger, C., Bailey, R., Griswold, M., Chiu, W., Garner, E., Obradovic, Z., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19 (1), 26.

Dunker, A., Silman, I., Uversky, V., Sussman, J., 2008. Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. 18 (6), 756.

Edgar, R., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (5), 1792.

Eisenberg, D., McLachlan, A., 1986. Solvation energy in protein folding and binding. Nature 319, 199.

Eisenmenger, F., Hansmann, U., Hayryan, S., Hu, C., 2001. Smmp a modern package for simulation of proteins. Comput. Phys. Commun. 138, 192.

Eisenmenger, F., Hansmann, U., Hayryan, S., Hu, C., 2006. An enhanced version of smmp - open-source software package for simulation of proteins. Comput. Phys. Commun. 174, 422.

Eisenstat, S., Walker, H., 1996. Choosing the forcing terms in an inexact newton method. Society for Industrial and Applied Mathematics J. Sci. Comput. 17 (1), 16.

Elber, R., 2005. Computer simulations of protein folding: Classical trajectories by optimization of action. Comput. Phys. Commun. 169 (1-3), 277.

Elber, R., Karplus, M., 1987. A method for determining reaction paths in large molecules: application to myoglobin. Chem. Phys. Lett. 139 (5), 375.

Elber, R., Roitberg, A. S., C. Goldstein, R., Li, H., Verkhivker, G., Keasar, C., Zhang, J., Ulitsky, A., 1995. Moil- a program for simulation of macromolecules. Comput. Phys. Commun. 91 (1-2), 159.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V., Pieper, U., Stuart, A., Martí-Renom, M., Madhusudhan, M., Yerkovich, B., Sali, A., 2003. Tools for comparative protein structure modeling and analysis. Nucleic Acids Res. 31 (13), 3375.

Eswar, N., Martí-Renom, M., Webb, B., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U., Sali, A., 2006. Comparative protein structure modeling with modeller. Curr. Protoc. Bioinf. 15, 561.

Fan, H., Mark, A., 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci. 13 (1), 211.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y., 2012. Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Comput. Chem. 33 (3), 259–267.

Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J., Brooks, C., 2000. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins: Struc., Func. Gen. 41 (1), 86.

Feng, D., Doolittle, R., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25 (4), 351.

Finkelstein, A., Ptitsyn, O., 1987. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50 (3), 171.

Finn, R., Clements, J., Eddy, S., 2011. Hmmer web server: interactive sequence similarity searching. Nucleic Acids Res. 39, 29.

Fischer, D., 2006. Servers for protein structure prediction. Curr. Opin. Struct. Biol. 16, 178.

Fischer, D., Eisenberg, D., 1996. Protein fold recognition using sequence-derived predictions. Protein Sci. 5 (5), 947.

Fiser, A., Do, R., Sali, A., 2000. Modeling of loops in protein structure. Protein Sci. 9 (9), 1753.

Fitch, W., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155 (760), 279.

Floudas, C., Fung, H., McAllister, S., Moennigmann, M., Rajgaria, R., 2006. Advances in protein structure prediction and de novo protein design: A review. Chem. Eng. Sci. 61 (3), 966.

Fogolari, F., Brigo, A., Molinari, H., 2002. The poisson-boltzmann equation for biomolecular electrostatics: a tool for structural biology. J. Mol. Recognit. 15 (6), 377.

Fonseca, R., Paluszewski, M., Winter, P., 2010. Protein structure prediction using bee colony optimization metaheuristic. J. Math. Model. Alg. 9 (2), 181.

Fraenkel, A. S., 1993. Complexity of protein folding. Bull. Math. Biol. 55 (6), 1199.

Fujitsuka, Y., Chikenji, G., Takada, S., 2006. Simfold energy function for de novo protein structure prediction: consensus with rosetta. Proteins: Struc., Func. Gen. 62 (2), 381.

Fujitsuka, Y., Takada, S., Luthey-Schulten, Z., Wolynes, P., 2004. Optimizing physical energy functions for protein folding. Proteins: Struc., Func. Gen. 54 (1), 88.

Garcez, A., Lamb, L., 2006. A connectionist computational model for epistemic and temporal reasoning. Neural Computation 18 (7), 1711.

Garcez, A., Lamb, L., Gabbay, D., 2007. Connectionist modal logic: Representing modalities in neural networks. Theoretical Computer Science 371 (1-2), 34.

Garey, M., Johnson, D., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness, 1st Edition. W.H. Freeman, New York.

Gibas, C., Jambeck, P., 2001. Developing bioinformatics computer skills, 1st Edition. O'Reilly, New York.

Gibbs, N., Clarke, A., Sessions, R., 2001. Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. Proteins: Struc., Func. Gen. 43 (2), 186.

Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L., 2003a. 3d-jury: a simple approach to improve protein structure predictions. Bioinformatics 19 (8), 1015.

Ginalski, K., Pas, J., Wyrwicz, L., Grotthuss, M., Bujnicki, J., Rychlewski, L., 2003b. Orfeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res. 31 (13), 3804.

Gniewek, P., Kolinski, A., Kloczkowski, A., Gront, D., 2014. Bioshell-threading: versatile monte carlo package for protein 3d threading. BMC Bioinformatics 15 (1), 22–29.

Godzik, A., Kolinski, A., Skolnick, J., 1992. A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. 227 (1), 227.

Gohlkea, H., Hendlicha, M., Klebe, G., 2000. Knowledge-based scoring function to predict protein-ligant interactions. J. Mol. Biol. 295, 337.

Gonnet, G., Cohen, M., Benner, S., 1992. Exhaustive matching of the entire protein sequence database. Science 256 (5062), 1443.

Gopakumar, O., 2012. Bioinformatics: Sequence and Structural Analysis. Alpha Science Intl Ltd.

Gordon, D., Marshall, S., Mayo, S., 1999. Energy functions for protein design. Curr. Opin. Struct. Biol. 9 (4), 509.

Grasso, C., Lee, C., 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics 20 (10), 1546.

Greer, J., 1990. Comparative modeling methods: application to the family of the mammalian serine protease. Proteins: Struc., Func. Gen. 7 (4), 317.

Gribskov, M., 1994. Profile analysis. Humana Press, New York.

Gribskov, M., McLachlan, A., Eisenberg, D., 1987. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84 (13), 4355.

Grippen, G., 1978. The tree structural organization of proteins. J. Mol. Biol. 126 (3), 315.

Guarnieri, F., Sitill, W., 1994. A rapidly convergent simulation method: Mixed monte carlo/stochastic dynamics. J. Comput. Chem. 15 (11), 1302.

Guda, C., Lu, S., Scheeff, E., Bourne, P., Shindyalov, I., 2004. Ce-mc: a multiple protein structure alignment server. Nucleic Acids Res. 32 (2), W100–W103.

Guest, M., Bush, I., van Dam, H., Sherwood, P., Thomas, J., van Lenthe, J., Havenith, R., Kendrick, J., 2005. The gamess-uk electronic structure package: algorithms, developments and applications. Mol. Phys. 103 (6), 719.

Gunasekaran, K., Tsai, C., Kumar, S., Zanuy, D., Nussinov, R., 2003. Extended disordered proteins: targeting function with less scaffold. Trends Biochem. Sci. 28 (2), 81.

Guntert, P., 2004. Automated nmr structure calculation with cyana. Methods Mol. Biol. 278, 353.

Hagler, A. T., Stern, P., Lifson, S., Ariel, S., 1979. Urey-bradley force field, valence force field, and ab initio study of intramolecular forces in tri-tert-butylmethane and isobutane. J. Am. Chem. Soc. 101 (4), 813.

Halgren, T. A., 1995. Potential energy functions. Curr. Opin. Struct. Biol. 5 (2), 205.

Hao, M., Scheraga, H., 1999. Designing potential energy functions for protein folding. Curr. Opin. Struct. Biol. 9 (2), 184.

Hart, W., Istrail, S., 1997. Robust proofs of np-hardness for protein folding: general lattices and energy potentials. J. Comput. Biol. 4 (1), 1.

Harvey, M., Giupponi, G., Fabritiis, G. D., 2009. Acemd: Accelerating biomolecular dynamics in the microsecond time scale. J. Chem. Theory Comput. 5 (6), 1632.

Haykin, S., 1998. Neural Networks: A comprehensive foundation, 2nd Edition. Prentice Hall Inc., New York.

He, Y., Xiao, Y., Liwo, A., Scheraga, H., 2009. Exploring the parameter space of the coarse-grained unres force field by random search: selecting a transferable medium-resolution force field. J. Comput. Chem. 30 (13), 2127.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M., 1990. Identification of native protein folds amongst a large number of incorrect models. the calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216 (1), 167.

Henikoff, S., Henikoff, J., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915.

Henikoff, S., Henikoff, J., 1993. Performance evaluation of amino acid substitution matrices. Proteins: Struc., Func. Gen. 17 (1), 49.

Henikoff, S., Henikoff, J., 1994. Protein family classification based on searching a database of blocks. Genomics 19, 97.

Herges, T., Schug, A., Merlitz, H., Wenzel, W., 2003. Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. Nanotechnology 14, 1161.

Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E., 2008. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. 4 (3), 435.

Higgins, D., Sharp, P., 1988. Clustal: a package for performing multiple sequence alignment on a microcomputer. Gene 73 (1), 237.

Hirosawa, M., Totoki, Y., M., H., Ishikawa, M., 1995. Comprehensive study on iterative algorithms of multiple sequence alignment. CABIOS, Comput. Appl. Biosci. 11 (1), 13.

Holland, J., 1993. Adaptation in natural and artificial systems, 1st Edition. The MIT Press, Boston.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., Vriend, G., 1992. A database of protein structure families with common folding motifs. Protein Sci. 1 (12), 1691.

Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273 (5275), 595.

Hoque, M., Chetty, M., Dooley, L., 2005. A new guided genetic algorithm for 2D hydrophobic-hydrophilic model to predict protein folding. IEEE Computer Society Press, New York.

Hoque, M., Chetty, M., Dooley, L., 2006. A guided genetic algorithm for protein folding prediction using 3D hydrophobic-hydrophilic model. IEEE Computer Society Press, New York.

Hoque, M., Chetty, M., Sattar, A., 2009. Genetic algorithm in *ab initio* protein structure prediction using low resolution model: A review. Vol. 224. Springer, Berlin.

Hovmoller, T., Ohlson, T., 2002. Conformation of amino acids in protein. Acta Crystallogr. 58 (5), 768.

Huang, E., Samudrala, R., Ponder, J., 1998. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. Protein Sci. 7 (9), 1998.

Huang, E., Subbiah, S., Tsai, J., Levitt, M., 1996. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J. Mol. Biol. 257 (3), 716.

Huang, Y. J., Mao, B., Aramini, J. M., Montelione, G. T., 2014. Assessment of template-based protein structure predictions in casp10. Proteins: Struct., Funct., Bioinf. 82, 43–56.

Hughey, R., Krogh, A., 1996. Hidden markov models for sequence analysis extension and analysis of the basic method. CABIOS, Comput. Appl. Biosci. 12 (2), 95.

Hutter, J., Curioni, A., 2005. Dual-level parallelism for ab initio molecular dynamics: Reaching teraflop performance with the cpmd code. Parallel Comput. 31 (1), 1.

Ishida, T., Nishimura, T., Nozaki, M., Inoue, T., Terada, T., Nakamura, S., Shimizu, K., 2003. Development of an ab initio protein structure prediction system able. Genome Inf. 14, 228.

Jacobson, M., Friesner, R. X., Honig, B., 2002. On the role of the crystal environment in determining protein side-chain conformations. J. Mol. Biol. 320 (3), 597.

Jacobson, M., Kaminski, G., Friesner, R., Rapp, C., 1968B. Force field validation using protein side-chain prediction. J. Phys. Chem. B 106 (44), 11673.

Jacobson, M., Pincus, D., Rapp, C., Day, T., Honig, B., Shaw, D., Friesner, R., 2004. A hierarchical approach to all-atom loop prediction. Proteins: Struc., Func. Gen. 55, 351.

Jaroszewski, L., Li, Z., hui Cai, X., Weber, C., Godzik, A., 2011. Ffas server: novel features and applications. Nucleic Acids Res. 39 (Web-Server-Issue), 38–44.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., Godzik, A., 2005. Ffas03: a server for profileprofile sequence alignments. Nucleic Acids Res. 33 (suppl 2), W284–W288.

Jaskowski, W., Blazewicz, J., Lukasiak, P., Milostan, M., Krasnogor, N., 2007. 3d-judge - a metaserver approach to protein structure prediction. Found. Comput. Decis. Sci. 31 (1), 3.

Jauch, R., Yeo, H., Kolatkar, P., Clarke, N., 2007. Prediction of protein structures and their docking. Proteins: Struct., Funct., Bioinf. 69 (8), 57.

Jayaram, B., Bhushan, K., Shenoy, S., Narang, P., Bose, S., Agrawal, P., Sahu, D., Pandey, V., 2006. Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res. 34 (21), 6195.

Jiang, T., Xu, Y., Zhang, M., 2002. Protein structure prediction by protein threading and partial experimental data. MIT Press, Boston.

Johnson, M., Srinivasan, N., Sowdhamini, R., Blundell, T., 1994. Knowledge-based protein modeling. Crit. Rev. Biochem. 29 (1), 1.

Johnston, M., Fernández-Galván, I., Villà-Freira, J., 2005. Framework-based design of a new all-purpose molecular simulation application: the adun simulator. J. Comput. Chem. 26 (15), 1647.

Jones, D., 1997. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins: Struc., Func. Gen. S1, 185.

Jones, D., 1999a. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287 (4), 797 – 815.

Jones, D., 1999b. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292 (2), 195 – 202.

Jones, D., 1999B. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292 (2), 195.

Jones, D., 2001. Predicting novel protein folds by using fragfold. Proteins: Struc., Func. Gen. 45 (S5), 127.

Jones, D., Bryson, K., Coleman, A., McGuffin, L., Sadowski, M., Sodhi, J., Ward, J., 2005a. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 61, 143–151.

Jones, D., Miller, R., Thornton, J., 1995. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins: Struc., Func. Gen. 23 (3), 387.

Jones, D., Taylor, W., Thornton, J., 1992. A new approach to protein fold recognition. Nature 358 (6381), 86.

Jones, D. T., Bryson, K., Coleman, A., McGuffin, L. J., Sadowski, M. I., Sodhi, J. S., Ward, J. J., 2005b. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 61 (7), 143.

Jones, S., Thornton, J., 1997B. Prediction of protein-protein interaction sites using patch analysis. J. Mol. Biol. 272 (1), 133.

Joo, K., Lee, J., Kim, S., Kim, I., Lee, J., Lee, S., 2004. Profile-based nearest neighbor method for pattern recognition. J. Korean Phys. Soc. 44 (3), 599.

Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R., Klein, M., 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79, 926.

Jorgensen, W., Maxwell, D., Tirado-Rives, J., 1996. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118 (45), 11225.

Jorgensen, W., Tirado-Rives, J., 2005. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc. Natl. Acad. Sci. U.S.A. 102 (19), 6665.

Jorgensen, W., Tirado-Rives, J., 2005B. Molecular modeling of organic and biomolecular systems using boss and mcpro. J. Comput. Chem. 26 (16), 1689.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., Xu, J., 2012. Template-based protein structure modeling using the raptorx web server. Nature protocols 7 (8), 1511–1522.

Karplus, K., Barrett, C., Hughey, R., 1992. Hidden markov models for detecting remote protein homologies. Bioinformatics 14 (10), 846.

Karplus, K., Karchin, R., Barret, C., Tu, S., Cline, M., Diekhans, M., Grate, L., Gasper, J., Hughey, R., 2001. What is the value added by human intervention in protein structure prediction? Proteins: Struc., Func. Gen. 5, 86.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., Hughey, R., 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins: Struc., Func. Gen. 56 (S6), 491.

Karplus, M., 1997. The levinthal paradox: yesterday and today. Folding Des. 2 (1), S69.

Kelley, L., Gardner, S. P., Stutcliffe, M., 1996. An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally-related subfamilies. Protein Eng. 9, 1063.

Kelley, L., MacCallum, R., Sternberg, M., 2000. Enhanced genome annotation using structural profiles in the program 3d-pssm. J. Mol. Biol. 299, 501.

Kelley, L. A., Sternberg, M., 2009. Protein structure prediction on the web: a case study using the phyre server. Nature protocols 4 (3), 363–371.

Kepleis, J., Floudas, C., 2002B. Prediction of β-sheet topology and disulfide bridges in polypeptides. J. Comput. Chem. 24 (2), 191.

Khalili, M., Liwo, A., Jagielska, A., Scheraga, H., 2005. Molecular dynamics with the united-residue model of polypeptide chains. ii. langevin and berendsen-bath dynamics and tests on model alpha-helical systems. J. Phys. Chem. B 109 (28), 13798.

Kiefer, F., Arnold, K., Kuenzli, M., Bordoli, L., Schwede, T., 2009. The swiss-model repository and associated resources. Nucleic Acids Res. 37, D387.

Kim, D., DiMaio, F., Yu-Ruei Wang, R., Song, Y., Baker, D., 2014. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins 82 (S2), 208–218.

Kim, D., Xu, D., Guo, J., Ellrott, K., Xu, Y., 2003. Prospect ii: protein structure prediction program for genomescale applications. Protein Eng. 16 (9), 641–650.

Kim, J., Pramanik, S., Chung, M., 1994. Multiple sequence alignment using simulated annealing. CABIOS, Comput. Appl. Biosci. 10 (4), 419.

Kitchen, D. B., Decornez, H., Furr, J. R., Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discovery 3 (11), 935.

Klepeis, J., Androulakis, M., Floudas, C., 1998B. Predicting solvated peptide conformations via global minimization of energetic atom-to-atom interactions. Comput. Chem. Eng. 22, 765.

Klepeis, J., Floudas, C., 1999. Free energy calculations for peptides via deterministic global optimization. J. Chem. Phys. 110, 7491.

Klepeis, J., Floudas, C., 2002. Ab initio prediction of helical segments in polypeptides. J. Comput. Chem. 23, 245.

Klepeis, J., Floudas, C., 2003. Astro-fold: a combinatorial and global optimization framework for ab initio prediction of the three-dimensional structures of proteins from the amino acid sequence. Biophys. J. 85, 2119.

Klepeis, J., Floudas, C., 2003B. Ab initio tertiary structure prediction of proteins. J. Global Optim. 25, 113.

Klepeis, J., Ierapetritou, M. G., Floudas, C., 1998. Protein folding and peptide docking: a molecular modeling and global optimization approach. Comput. Chem. Eng. 22, 3.

Klepeis, J., Pieja, M., Floudas, C., 2003. Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. Biophys. J. 84, 869.

Ko, J., Park, H., Seok, C., 2012. Galaxytbm: template-based modeling by building a reliable core and refining unreliable local regions.

BMC bioinformatics 13 (1), 198–207.

Koehl, P., Levitt, M., 1999. A brighter future for proteins structure prediction. Nat. Struct. Mol. Biol. 6, 108.

Kolinski, A., 2004. Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. 51 (2), 349–371.

Kolinski, A., Bujinicki, J., 2005. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins: Struc., Func. Gen. 7, 84.

Koop, J., Schwede, T., 2004. Automated protein structure homology modeling: a progress report. Pharmacogenomics 5 (4), 405.

Koop, S., Bordoli, L., Battey, J., Kiefer, F., Schwede, T., 2007. Assessment of casp7 predictions for template-based modleing targets. Proteins: Struct., Funct., Bioinf. 69 (8), 38.

Koppensteiner, W. A., Sippl, M. J., 1995. Knowledge-based potentials-back to the roots. Biochemistry 63, 247.

Koza, J. R., 1992. Molecular dynamics simulations: Elementary methods, 1st Edition. John Wiley and Sons, Inc., New York.

Kresse, G., Marsman, M., Furthmuller, 2009. Computational Physics, Faculty of Physics, Wien University, Wien.

Krogh, A., Brown, M., Mian, I., Sjolander, K., Haussler, D., 1994. Hidden markov models in computational biology: application to protein modeling. J. Mol. Biol. 235 (5), 1501.

Kuang, S., Li, C., Vardeman II, C., Lin, T., Fennell, C., Sun, X., Daily, K., Zheng, Y., Meineke, M., Gezelter, J., 2009. Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame.

Kundrot, C., Ponder, J., Richards, F., 1991. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. J. Comput. Chem. 12 (3), 402.

Lambert, C., Leonard, N., De Bolle, X., Depiereux, E., 2002. Esypred3d: Prediction of proteins 3d structures. Bioinformatics 18 (9), 1250.

Lander, E., Waterman, M., 1999. The secrets of life: a mathematician's introduction to Molecular Biology. National Academy Press, Washington D. C.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., Higgins, D., 2007. Clustal w and clustal x version 2.0. Bioinformatics 23 (21), 2947.

Laskowiski, R., Watson, J., Thornton, J., 2005a. Profunc: a server for predicting protein functions from 3d structure. Nucleic Acids Res. 33, 89.

Laskowiski, R., Watson, J., Thornton, J., 2005b. Protein function prediction using local 3d templates. J. Mol.Biol. 351, 614.

LaValle, S. M., 2006. Planning Algorithms, 1st Edition. Cambridge University Press, New York.

Lazaridis, T., Karplus, M., 2000. Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol. 10 (2), 139.

Le Grand, S., Merz, K. J., 1993. The application of the genetic algorithm to the minimization of potential energy functions. J. Global Optim. 3 (1), 49.

Lee, J., Kim, S., Joo, K., Kim, I., Lee, J., 2004. Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing. Proteins: Struc., Func. Gen. 56 (4), 704.

Lee, J., Pillardy, J., Czaplewski, C., Arnautova, Y., Ripoll, D., Liwo, A., Gibson, K. D., Wawak, R., Scheraga, H., 2000. Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins, and crystals. Comput. Phys. Commun. 128 (1-2), 399.

Lee, J., Ripoll, D., Czaplewski, C., Pillardy, J., Wedemeyer, W., Scheraga, H., 2001. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. J. Phys. Chem. B 105 (30), 7291.

Lee, J., Scheraga, H., 1999. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. Int. J. Quantum Chem. 75, 255.

Lee, J., Scheraga, H., Rackovsky, S., 1997. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J. Comput. Chem. 18 (9), 1222.

Lee, J., Scheraga, H., Rackovsky, S., 1998. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46 (2), 103.

Lee, D. amd Redfern, O., Orengo, C., 2007. Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. 8, 995.

Lehninger, A., Nelson, D., Cox, M., 2005. Principles of Biochemistry, 4th Edition. W.H. Freeman, New York.

Lengauer, T., Rarey, M., 1996. Computational methods for biomolecular docking. Curr. Opin. Struct. Biol. 6 (3), 402.

Lesk, A. M., 2002. Introduction to Bioinformatics, 1st Edition. Oxford University Press Inc., New York.

Levinthal, C., 1968. Are there pathways for protein folding? J. Chim. Phys. Phys.-Chim. Biol. 65 (1), 44.

Levitt, M., 1983. Molecular dynamics of native protein: Computer simulation of trajectories. J. Mol. Biol. 168 (3), 595.

Levitt, M., 1992. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226, 507.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261 (5561), 552.

Li, H., Helling, R., Tang, C., Wingreen, N., 1996. Emergence of preferred structures in a simple model of protein folding. Science 273 (5275), 666.

Li, S. C., Bu, D., Xu, J., Li, M., 2008. Fragment-hmm: a new approach to protein structure prediction. Proteins: Struct., Funct., Bioinf. 17 (11), 1925.

Li, Y., Zhang, Y., 2009. Remo: a new protocol to refine full atomic protein models from c-α traces by optimizing hydrogen-bonding networks. Proteins: Struct., Funct., Bioinf. 76, 665.

Li, Y., Zhang, Y., 2011. Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling. Structure 19 (12), 1784.

Li, Z., Yang, Y., Zhan, J., Dai, L., Zhou, Y., 2013. Energy functions in de novo protein design: Current challenges and future prospects. Annu. Rev. Biophys. 42 (1), 315–335.

Lifson, S., Warshel, A., 1968. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and nalkane molecules. J. Chem. Phys. 49 (5116), 14.

Liljas, A., Liljas, L., Pskur, J., Lindblom, G. amd Nissen, P., Kjeldgaard, M., 2001. Textbook of structural biology. World Scientific Printers, Singapore.

Limbach, H., Arnold, A., Mann, B., Holm, C., 2006. Espresso: an extensible simulation package for research on soft matter systems. Comput. Phys. Commun. 174 (9), 704.

Lipman, D., Altschul, S., Kececioglu, J., 1989. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 86 (12), 4412.

Lipman, D., Pearson, W., 1985. Rapid and sensitive protein similarity searches. Science 227 (4693), 1435.

Lise, S., Buchan, D., Pontil, M., Jones, D. T., 2011. Predictions of hot spot residues at protein-protein interfaces using support vector machines. PLoS one 6 (2), e16774.

Liwo, A., Czaplewski, C., Kleinerman, D., Blood, P., Scheraga, H., 2010. Implementation of molecular dynamics and its extensions with the coarse-grained unres force field on massively parallel systems; towards millisecond-scale simulations of protein structure, dynamics, and thermodynamics. J. Chem. Theory Comput. 6 (3), 890.

Liwo, A., Kazmierkiwicz, R., Czaplewski, C., Groth, M., Oldziej, S., Wawak, R., Rackovsky, S., Pincus, M., Scheraga, H., 1998.

United-residue force field for off-lattice protein-structure simulations; iii. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. J. Comput. Chem. 19, 259.

Liwo, A., Lee, J., Ripoll, D., Pillardy, J., Scheraga, H., 1999a. Protein structure prediction by global optimization of a potential energy function. Proc. Natl. Acad. Sci. U.S.A. 96, 5482.

Liwo, A., Oldziej, S., Pincus, M., Wawak, R., Rackovsky, S., Scheraga, H., 1999b. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J. Comput. Chem. 18, 849.

Liwo, A., Pincus, M., Wawak, R., Rackovsky, S., Oldziej, S., Scheraga, H., 1997. A united-residue force field for off-lattice protein-structure simulations. h:parameterization of local interactions and determination of the weights of energy terms by z-score optimization. J. Comput. Chem. 18, 874.

Lo Conte, L., Ailey, B., Hubbard, T., Brenner, S., Murzin, A., Chothia, C., 1999. Scop: a structural classification of protein database. Nucleic Acids Res. 28 (1), 257.

Lobley, A., Sadowski, M. I., Jones, D. T., 2009a. pgenthreader and pdomthreader: New methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25, 1761.

Lobley, A., Sadowski, M. I., Jones, D. T., 2009b. pgenthreader and pdomthreader: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25 (14), 1761–1767.

Lobley, A. E., Nugent, T., Orengo, C. A., Jones, D. T., 2008. Ffpred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res. 36 (suppl 2), W297–W302.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M., 1990. Molecular Cell Biology, 5th Edition. Scientific American Books, W.H. Freeman, New York.

Lu, H., Skolnick, J., 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins: Struct., Funct., Bioinf. 44, 223.

Lundstrom, J., Rychlewski, L., Bujnicki, J., Elofsson, A., 2001. Pcons: A neural-network based consensus predictor that improves fold recognition. Protein Sci. 10, 2354.

Luthy, R., Bowie, J., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature 356, 83.

Lyras, D., Metzler, D., 2014. Reformalign: improved multiple sequence alignments using a profile-based meta-alignment approach. BMC Bioinformatics 15 (1), 265–282.

Lyubartsev, A., Laaksonen, A., 2000. M.dynamix - a scalable portable parallel md simulation package for arbitrary molecular mixtures. Comput. Phys. Commun. 128, 565.

Maciej, B., Michal, J., Sebastian, K., Andrzej, K., 2013. Cabs-fold: server for the de novo and consensus-based prediction of protein structure. Nucleic Acids Res. 41, W406–W411.

Macke, T., Case, D., 1998. Modeling unusual nucleic acid structures. American Chemical Society, New York.

Mackerell, A. J., 2004. Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. 25 (13), 1584.

MacKerell, A. J., Banavali, N., Foloppe, N., 2001. Development and current status of the charmm force field for nucleic acids. Biopolymers 56 (4), 257.

MacKerell, A. J., Bashford, D., Bellott, M., Dunbrack, R. J., Evanseck, J., Field, M., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F., Mattos, C., Michnick, S., Ngo, T., Nguyen, D., Prodhom, B., Reiher, W. I., Roux, B. Schlenkrich, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wirkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998a. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. 102 (18), 3586.

MacKerell, A. J., Brooks, B., Brooks, C., Nilsson, L., Roux, B., Won, Y., Karplus, M., 1998b. CHARMM: The energy function and its parameterization with an overview of the program. Vol. 1. Wiley, New York.

MacKerell, A. J., Feig, M., Brooks, C., 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comput. Chem. 25 (11), 1400.

Madhusudhan, M., Martí-Renom, M., Sanchez, R., Sali, A., 2006. Variable gap penalty for protein sequence-structure alignment. Protein Eng. Des. Sel. 19 (3), 129.

Maisuradze, G., Senet, P., Czaplewski, C., Liwo, A., Scheraga, H., 2010. Investigation of protein folding by coarse-grained molecular dynamics with the unres force field. J. Phys. Chem. A 114 (13), 4471.

Marelius, J., Kolmodin, K., Feierberg, I., Avist, J., 1999. Q: An md program for free energy calculations and empirical valence bond simulations in biomolecular systems. J. Mol. Graphics Modell. 16 (4-6), 213.

Marsili, S., Signorini, G., Chelli, R., Marchi, M., Procacci, P., 2010. Orac: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. J. Comput. Chem. 31 (5), 1106.

Martí-Renom, M., Madhusudhan, M., Sali, A., 2004. Alignment of protein sequences by their profiles. Protein Sci. 13 (4), 1071.

Martí-Renom, M., Stuart, A., Fiser, A., Sanchez, A., Mello, F., Sali, A., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29 (16), 291.

Martin, M., Siepmann, J., 1999. Novel configurational-bias monte carlo method for branched molecules. transferable potentials for phase equilibria. 2. united-atom description of branched alkanes. J. Phys. Chem. B 103 (21), 4508.

Martínez, L., Andrade, R., Birgin, E., Martínez, J., 2009. Packmol: A package for building initial configurations for molecular dynamics simulations. J. Comput. Chem. 30 (13), 2157.

McGuffin, L. J., Jones, D. T., 2003a. Improvement of the genthreader method for genomic fold recognition. Bioinformatics 19, 874.

McGuffin, L. J., Jones, D. T., 2003b. Improvement of the genthreader method for genomic fold recognition. Bioinformatics 19 (7), 874–881.

McLachlan, A., 1992. Rapid comparison of protein structures. Acta Crystallogr. A38, 871.

Meinke, J., Mohanty, S., Eisenmenger, F., Hansmann, U., 2008. Smmp v. 3.0 - simulating proteins and protein interactions in python and fortran. Comput. Phys. Commun. 178 (6), 459.

Mohanty, D., Dominy, B., Kolinski, A., Brooks, C., Skolnick, J., 1999. Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the gcn4 leucine zipper. Proteins: Struct., Funct., Bioinf. 35 (4), 447.

Momany, F., McGuire, R., Burgess, A., Scheraga, H., 1975. Energy parameters in polypeptides vii, geometric parameters, partial charges, non-bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally occurring amino acids. J. Phys. Chem. 79 (22), 2361.

Moult, J., Fidelis, K., Hubbard, T., 2003. Critical assessment of methods of protein structure prediction (casp): round v. Proteins: Struc., Func. Gen. 53, 334.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., Tramontano, A., 2007. Critical assessment of methods of protein structure prediction: round vii. Proteins: Struc., Func. Gen. 69, 3.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Tramontano, A., 2009. Critical assessment of methods of protein structure prediction: round viii. Proteins: Struc., Func. Gen. 77, 1.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A., 2014. Critical assessment of methods of protein structure prediction (casp) round x. Proteins: Struct., Funct., Bioinf. 82, 1–6.

Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A., 2011. Critical assessment of methods of protein structure prediction (casp) - round ix. Proteins: Struct., Funct., Bioinf. 79 (S10), 1.

Moult, J., Fidelis, K., Rost, B., Hubbard, T., Tramontano, A., 2005. Critical assessment of methods of protein structure prediction (casp): round vi. Proteins: Struc., Func. Gen. 61, 3.

Moult, J., Fidelis, K., Zemla, A., Hubbard, T., 2001. Critical assessment of methods of protein structure prediction (casp): round iv. Proteins: Struc., Func. Gen. 45, 2.

Moult, J., Hubbard, T., Bryant, S., Fidelis, K., Pedersen, J., 1997. Critical assessment of methods of protein structure prediction (casp): round ii. Proteins: Struc., Func. Gen. 29, 2.

Moult, J., Hubbard, T., Fidelis, K., Pedersen, J., 1999. Critical assessment of methods of protein structure prediction (casp): round iii. Proteins: Struc., Func. Gen. 37, 2.

Moult, J., Pedersen, J., Judson, R., Fidelis, K., 1995. A large-scale experiment to assess protein structure prediction methods. Proteins: Struc., Func. Gen. 23, 2.

Moult, J. A., 2005. Decade of casp: progress, bottlenecks an prognosis in protein structure prediction. Curr. Opin. Struct. Biol. 15 (3), 285.

Nanias, M., Czaplewski, C., Scheraga, H., 2009. Replica exchange and multicanonical algorithms with the coarse-grained unres force field. J. Chem. Theory Comput. 2 (3), 513.

Narang, P., Bhushan, K., Bose, S., Jayaram, B., 2006. Protein structure evaluation using an all-atom energy based empirical scoring function. J. Biomol. Struct. Dyn. 23 (4), 385.

Ngo, J., Marks, J., Karplus, M., 1997. The protein folding problem and tertiary structure prediction. In: Merz Jr, K., Grand, S. (Eds.), Computational complexity, protein structure prediction and the Levinthal Paradox. Birkhauser, Boston, p. 435.

Notredame, C., 2002. Recent progresses in multiple sequence alignment: a survey. Pharmacogenomics 31 (1), 131.

Notredame, C., 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput. Biol. 8 (3), 1405.

Notredame, C., Higgins, D., Heringal, J., 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302 (1), 205.

Notredame, C., Holm, L., Higgins, D., 1998. Coffee: an objective function for multiple sequence alignments. Bioinformatics 14 (5), 407.

Nugent, T., Jones, D., 2009. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 10 (1), 159–169.

Oldziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J., Khalili, M., Arnautova, Y., Jagielska, A., Makowski, M., Schafroth, H., Kazmierkiewicz, R., Ripoll, D., Pillardy, J., Saunders, J., Kang, Y., Gibson, K., Scheraga, H., 2005. Physics-based protein-structure prediction using a hierarchical protocol based on the unres force field: assessment in two blind tests. Proc. Natl. Acad. Sci. U.S.A. 102 (21), 7547.

Onufriev, A., Bashford, D., Case, D., 2002. Effective born radii in the generalized born approximation: The importance of being perfect. J. Comput. Chem. 23 (14), 1297.

Ooi, T., Oobatake, M., Nemethy, G., Scheraga, H., 1987. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc. Natl. Acad. Sci. U.S.A. 84 (10), 3086.

Osguthorpe, D., 2000. Ab initio protein folding. Curr. Opin. Struct. Biol. 10 (2), 146.

Ota, M., Nishikawa, K., 1997. Assessment of pseudo-energy potentials by the best-five test: A new use of the three-dimensional profiles of proteins. Protein Eng. 10 (4), 339.

O'Toole, J., Dahler, J., 1960. Boltzmann equation and inverse collisions. J. Chem. Phys. 33 (5), 1487.

Ouzounis, C., Sander, C., Scharf, M., Schneider, R., 1993. Prediction of protein structure by evaluation of sequence structure fitness aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. 232 (3), 805.

Pappu, R., Hart, R., Ponder, J., 1998. Analysis and application of potential energy smoothing and search methods for global optimization. J. Phys. Chem. B 102 (48), 9725.

Park, B., Huang, E., Levitt, M., 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J. Mol. Biol. 266 (4), 831.

Park, S., 2005. A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. Genome Inf. 16 (2), 104.

Paschek, D., Geiger, A., 2003. Physikalische Chemie, Dortmund University, Dortmund.

Pauling, L., Corey, R., 1951. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. U.S.A. 37 (5), 251.

Pauling, L., Corey, R., Branson, H., 1951. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 37 (4), 205.

Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T. I., DeBolt, S., Ferguson, D., Seibel, G., Kollman, P., 1995. Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput. Phys. Commun. 91 (1-3), 1.

Pearson, W., Lipman, D., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85 (8), 2444.

Pedersen, J., Moult, J., 1997. Protein folding simulations with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269 (2), 240.

Peitsch, M., 1996. Prodmod and swiss-model: Internet-based tools for automated comparative protein modeling. Biochem. Soc. Trans. 24 (1), 274.

Peitsch, M., Jongeneel, C., 1993. A 3-d model for the cd40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. Int. Immunol. 5 (2), 233.

Pevzner, P. A., 2000. Computational Molecular Biology: An Algorithmic Approach, 1st Edition. The MIT Press, Cambridge.

Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., Schulten, K., 2005. Scalable molecular dynamics with namd. J. Comput. Chem. 26 (16), 1781.

Plimpton, S., 1995. Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. 117, 1.

Pokala, N., Handel, T., 2000. Review: protein design - where we were, where we are, where we're going. J. Struct. Biol. 134 (2-3), 269.

Pokarowski, P., Kolinski, A., Skolnickz, J., 2003. A minimal physically realistic protein-like lattice model: Designing an energy landscape that ensures all-or-none folding to a unique native state. Biophys. J. 84 (3), 1518.

Pollastri, G., Przybylski, D., Rost, B., Baldi, P., 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47 (2), 228–235.

Ponder, J., 2010. Jay Ponder Lab, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Saint Louis.

Ponder, J., Richards, F., 1987. An efficient newton-like method for molecular mechanics energy minimization of large molecules. J. Comput. Chem. 8 (7), 1016.

Ponting, C. P., Schultz, J., Milpetz, F., Bork, P., 1999. Smart: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 27 (1), 229–232.

Prlic, A., Bliven, S., Rose, P. W., Bluhm, W., Bizon, C., Godzik, A., Bourne, P., 2010. Precalculated protein structure alignments at the rcsb pdb website. Bioinformatics.

Procacci, P., Paci, E., Darden, T., Marchi, M., 1997. Orac: A molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions. J. Comput. Chem. 18 (15), 1848.

Pronk, S., Pll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., Lindahl, E., 2013. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29 (7), 845–854.

Qiu, D., Shenkin, P., Hollinger, F., Still, W., 1997. The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. J. Phys. Chem. A 101, 3005.

Rackovsky, S., 2010. Global characteristics of protein sequences and their implications. Proc. Natl. Acad. Sci. U.S.A. 107 (19), 8623.

Ramachandran, G., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 238.

Rapaport, D. C., 2004. The art of molecular dynamics simulation, 2nd Edition. Cambridge University Press, Cambridge.

Refson, K., 2000. Moldy: a portable molecular dynamics simulation program for serial and parallel computers. Comput. Phys. Commun. 126 (3), 310.

Rentzsch, R., Orengo, C., 2009. Protein function prediction - the power of multiplicity. Trends Biotechnol. 27 (4), 210.

Rice, D., Eisenberg, D., 1997. A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. 267 (4), 1026.

Richards, F., 1977. Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6, 151.

Richards, F., Kundrot, C., 1988. Identification of structural motifs from protein coordinate data: secondary structure and first level super-secondary structure. Proteins: Struct., Funct., Bioinf. 3 (2), 71.

Richardson, J., 1981. The anatomy and taxonomy of protein structures. Adv. Protein Chem. 34, 167.

Rohl, C., Strauss, C., Misura, K., Baker, D., 2004. Protein structure prediction using rosetta. Methods Enzymol. 383 (2), 66.

Rose, G., 1979. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134 (3), 447.

Rose, G., Wolfenden, R., 1993. Hydrogen bonding, hydrophobicity, packing and protein folding. Annu. Rev. Biophys. Biomol. Struct. 22, 381.

Rost, B., 1995a. Fitting 1-D predictions into 3-D structures. CRC Press, Boca Raton.

Rost, B., 1995b. Topits: Threading one-dimensional predictions into three-dimensional structures. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 314.

Rost, B., Schneider, R., C., S., 1997. Protein fold recognition by prediction-based threading. J. Mol. Biol. 270 (3), 471.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-tasser: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5 (4), 725–738.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning representations by back-propagating errors. Nature 323, 533.

Russell, R., Barton, G., 1994. Structural features can be unconserved in proteins with similar folds. an analysis of side-chain to side-chain contacts secondary structure and accessibility. J. Mol. Biol. 244 (3), 332.

Russell, R., Copley, R., Barton, G., 1996. Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 259 (3), 349.

Russell, R. B., Saqi, M. A., Bates, P. A., Sayle, R. A., Sternberg, M. J., 1998. Protein Eng. 11 (1), 1.

Rychlewski, L., Jaroszewski, L., Li, W., Godzik, A., 2000. Comparison of sequence profiles. strategies for structural predictions using sequence information. Protein Sci. 9 (2), 232.

Sadreyev, R., Grishin, N., 2003. Compass: A tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326 (1), 317.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406.

Sali, A., 1995. Modelling mutations and homologous proteins. Curr. Opin. Biotechnol. 6 (4), 437.

Sali, A., Blundell, T., 1993. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234 (3), 779.

Salomon-Ferrer, R., Case, D. A., Walker, R. C., 2013. An overview of the amber biomolecular simulation package. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 3 (2), 198–210.

Sánchez, R., Sali, A., 1997. Advances in comparative protein-structure modeling. Curr. Opin. Struct. Biol. 7 (2), 206.

Sasin, J., Kurowski, M., Bujnicki, J., 2003. Strucla: a www meta-server for protein structure comparison and evolutionary classification. Bioinformatics 19, 252.

Scheef, E., Fink, J., 2003. Fundamentals of protein structure: Structural Bioinformatics. Ch. 2, p. 15.

Schrödinger, LLC, August 2010. The PyMOL molecular graphics system, version 1.3r1.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., Baker, D., 2005. Progress in modeling of protein structures and interactions. Science 310 (5748), 638.

Schug, A., Herges, T., Verma, A., Wenzel, W., 2005. Investigation of the parallel tempering method for protein folding. J. Phys.: Condens. Matter 17, 1641.

Schwede, T., Kopp, J., Guex, N., Peitsch, M., 2003. Swiss-model: An automated protein homology-modeling server. Nucleic Acids Res. 31 (13), 3381.

Scott, W., Hunenberger, P., Tironi, I., Mark, A., Billeter, S., Fennen, J., Torda, A., Huber, T., Kruger, P., van Gunsteren, W., 1999. The gromos biomolecular simulation program package. J. Phys. Chem. A 103 (19), 3596.

Selezenev, A., Aleynikov, A., Gantchuk, N., Yermakov, P., Labanowski, J., Korkin, A., 2003. Sage md: molecular-dynamic software package to study properties of materials with different models for interatomic interactions. Comput. Mater. Sci. 28 (2), 107.

Setubal, J., Meidanis, J., 1997. Introduction to Computational Molecular Biology, 1st Edition. PWS Publishing Company, Boston.

Shen, H., Czaplewski, C., Liwo, A., Scheraga, H., 2008. Implementation of a serial replica exchange method in a physics-based united-residue (unres) force field. J. Chem. Theory Comput. 4 (8), 1386.

Shen, H., Liwo, A., Scheraga, H., 2009. An improved functional form for the temperature scaling factors of the components of the mesoscopic unres force field for simulations of protein structure and dynamics. J. Phys. Chem. B 113 (25), 8738.

Shenoy, S. R., Jayaram, B., 2010. Proteins: sequence to structure and function-current status. Curr.Protein Pept. Sci. 11 (7), 498–514.

Shi, J., Blundell, T., Mizuguchi, K., 2001. Fugue: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. 310 (1), 243.

Siew1, N., Elofsson, A., Rychlewski, L., Fischer, D., 2000. Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16 (9), 776.

Simons, K., Bonneau, R., Ruczinski, I., Baker, D., 1999B. Ab initio protein structure prediction of casp iii targets using rosetta. Proteins: Struct., Funct., Bioinf. 3 (3), 171.

Simons, K., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated anneling and bayesian score functions. J. Mol. Biol. 268 (1), 209.

Simons, K., Ruczinki, I., Kooperberg, C., Fox, B., Bystroff, C., Baker, D., 1999. Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins: Struct., Funct., Bioinf. 34 (1), 82.

Sippl, M., 1995. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5 (2), 229.

Sippl, M., Hendlich, M., Lackner, P., 1992. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin beta 4. Protein Sci. 1, 625.

Smith, J., 2005. The co-evolution of memetic algorithms for protein structure prediction. Stud. Fuzziness Soft Comput. 166, 105.

Smith, T., 1999. The art of matchmaking: sequence alignment methods and their structural implications. Structure 7 (1), R7.

Smith, T., Lo Conte, L., Bienkowska, J., Gaitatzes, C., Rogers, R. J., Lathrop, R., 1997. Current limitations to protein threading approaches. J. Comput. Biol. 4 (3), 217.

Smith, W., Forester, T., 1996. Dl poly 2.0: A general-purpose parallel molecular dynamics simulation package. J. Mol. Graphics 14 (3), 136.

Smith, W., Yong, C., Rodger, P., 2002. Dl poly: application to molecular simulation. Mol. Simul. 28 (5), 385.

Soding, J., Biegert, A., Lupas, A., 2005. The hhpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33 (Web Server Issue), 244.

Soding, J., Remmert, M., Biegert, A., 2006a. Hhrep: de novo protein repeat detection and the origin of tim barrels. Nucleic Acids Res. 34, 137.

Soding, J., Remmert, M., Biegert, A., A., L., 2006b. Hhsenser: exhaustive transitive profile search using hmmhmm comparison. Nucleic Acids Res. 34, 374.

Soler, J., Artacho, E., Gale, J., García, J., Ordejón, J., Sánchez-Portal, D., 2001. Levinthal's paradox. J. Phys. Condens. Matter 14, 2745.

Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A., Durbin, R., 1998. Pfam: Multiple sequence alignments and hmm-profiles of protein domains. Nucleic Acids Res. 26 (1), 320–322.

Srinivasan, R., Rose, G., 1995. Linus - a hierarchic procedure to predict the fold of a protein. Proteins: Struct., Funct., Bioinf. 22 (2), 81.

Srinivasan, R., Rose, G., 2002. Ab initio prediction of protein structure using linus. Proteins: Struct., Funct., Bioinf. 47 (4), 489.

Srinivasan, S., March, C., Sudarsanam, S., 1993. An automated method for modeling proteins on known templates using distance geometry. Protein Sci. 2 (2), 227.

Stadler, J., Mikulla, R., Trebin, H., 1997. Imd: A software package for molecular dynamics studies on parallel computers. Int. J. Mod. Phys. C 8 (5), 1131.

Sternberg, M., 1997. Protein Structure Prediction: A practical approach, 1st Edition. Oxford University Press, New York.

Still, W., Tempczyk, A., Hawley, R., Hendrickson, T., 1990. Semi-analytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112 (16), 6127.

Subramani, A., Wei, Y., Floudas, C. A., 2012. Astro-fold 2.0: An enhanced framework for protein structure prediction. AIChE Journal 58 (5), 1619–1637.

Sun, S., 1995. A genetic algorithm that seeks native states of peptides and proteins. Biophys. J. 69 (2), 340.

Sding, J., 2005. Protein homology detection by hmmhmm comparison. Bioinformatics 21 (7), 951–960.

Tai, C., Bai, H., Taylor, T. J., Lee, B., 2014. Assessment of template-free modeling in casp10 and roll. Proteins: Struct., Funct., Bioinf. 82, 57–83.

Taylor, W., 1988. A flexible method to align large numbers of biological sequences. J. Mol. Evol. 28 (1-2), 161.

Taylor, W., 1996. Multiple protein sequence alignment: algorithms and gap insertion. Methods Enzymol. 266, 343.

Taylor, W. R., 1997. Multiple sequence threading: an analysis of alignment quality and stability. J. Mol. Biol. 269 (5), 902.

Teodorescu, O., Galor, T., Pillardy, J., Elber, R., 2004. Enriching the sequence substitution matrix by structural information. PROTEINS: Structure, Function, and Bioinformatics 54 (1), 41–48.

Thachuk, C., Shmygelska, A., Hoos, H., 2007. A replica exchange monte carlo algorithm for protein folding in the hp model. BMC Bioinf. 8, 20.

Thompson, J., Higgins, D., Gibson, T., 1994. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673.

Thompson, J., Plewniak, F., Poch, O., 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27 (13), 2682–2690.

Thukral, L., Shenoy, S., Bhushan, K., Jayaram, B., 2007. Proregin: A regularity index for the selection of native-like tertiary structures of proteins. J. Biosci. 32 (1), 71.

Tompa, P., 2002. Intrinsically unstructured proteins. Trends Biochem Sci. 27 (10), 527.

Tompa, P., Csermely, P., 2004. The role of structural disorder in the function of rna and protein chaperones. FASEB J. 18 (11), 1169.

Tramontano, A., 2006. Protein structure prediction, 1st Edition. John Wiley and Sons, Inc., Weinheim.

Tuckerman, M., Yarne, D., Samuelson, S., Hughes, A., Martyna, G., 2000. Exploiting multiple levels of parallelism in molecular dynamics based calculations via modern techniques and software paradigms on distributed memory computers. Comput. Phys. Commun. 128 (1-2), 333.

Tuffery, P., Etchebest, C., Hazout, S., Lavery, R., 1991. A new approach to the rapid determination of protein sidechain conformations. J. Biomol. Struct. Dyn. 8 (6), 1267.

Turcotte, M., Muggleton, S., Sternberg, M., July 1998. Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure. Springer, Madison.

Turcotte, M., Muggleton, S., Sternberg, M., 2001. Automated discovery of structural signatures of protein fold and function. J. Mol. Biol. 306, 591.

Turcotte, M., Muggleton, S., Sternberg, M., 2001B. The effect of relational background knowledge on learning of protein three-dimensional fold signatures. Machine Learning 43 (1-2), 81.

Turcotte, M., Muggleton, S., Sternberg, M., 2001C. Generating protein three-dimensional fold signatures using inductive logic programming. Comput. Chem. 26, 57.

Unger, R., Moult, J., 1993. On the applicability of genetic algorithms to protein folding. IEEE Computer Society Press, New York.

Unger, R., Moult, J., 1995B. Genetic algorithms for protein folding simulations. J. Mol. Biol. 231 (1), 75.

Uversky, V., 2001. What does it mean to be natively unfolded? Eur. J. Biochem. 269 (1), 2.

Vallat, B. K., Pillardy, J., Elber, R., 2008. A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins. Proteins: Structure, Function, and Bioinformatics 72 (3), 910–928.

Vallat, B. K., Pillardy, J., Májek, P., Meller, J., Blom, T., Cao, B., Elber, R., 2009. Building and assessing atomic models of proteins from structural templates: Learning and benchmarks. Proteins: Structure, Function, and Bioinformatics 76 (4), 930–945.

van der Spoel, D., 1998. The solution conformation of amino acids from molecular dynamics simulations of gly-x-gly peptides: comparison with nmr parameters. Biochem. Cell Biol 76 (2-3), 164.

van der Spoel, D., Berendsen, H., 1997. Molecular dynamics simulations of leu-enkephalin in water and dmso. Biophys. J. 72 (5), 2032.

van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A., Berendsen, H., 2005. Gromacs: fast, flexible, and free. J. Comput. Chem. 26 (16), 1701.

van der Spoel, D., van Buuren, A., Tieleman, D., Berendsen, H., 1996. Molecular dynamics simulations of peptides from bpti: A closer look at amide-aromatic interactions. J. Biomol. NMR 8 (3), 229.

van Gunsteren, W., Berendsen, H., 1990. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. Angew. Chem., Int. Ed. Engl. 29 (9), 992.

Vásquez, M., 1996. Modeling side-chain conformation. Curr. Opin. Struct. Biol. 6 (2), 217.

von Ohsen, N., Sommer, I., Zimmer, R., 2003. Profile-Profile alignment: a powerful tool for protein structure prediction. World Scientific Publishing Co Pte Ltd, Singapore.

Wallace, I., Blackshields, G., Higgins, D., 2005. Multiple sequence alignments. Curr. Opin. Struct. Biol. 15 (3), 261.

Wallner, B., Larsson, P., Elofsson, A., 2007. Pcons.net: protein structure prediction meta server. Nucleic Acids Res. 35, 369.

Wang, G., Dunbrack, R. J., 2003. Pisces: a protein sequence culling server. Bioinformatics 19 (12), 1589.

Wang, Z., 1998. A re-estimation for the total numbers of protein folds and super-families. Protein Eng. 11 (8), 621.

Wesson, L., Eisenberg, D., 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci. 1 (2), 227.

Whisstock, J., Lesk, A., 2003. Prediction of protein function from protein sequence and structure. Q. Rev. Biophys 36 (3), 307.

White, J., Muchnik, I., Smith, T., 1994. Modeling protein cores with markov random fields. Math. Biosci. 124 (2), 149.

Williams, D., 1998. Representation of the molecular electrostatic potential by atomic multi-pole and bond dipole models. J. Comput. Chem. 9 (7), 745.

Wright, P., Dyson, H., 1999. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J. Mol. Biol. 293 (2), 321.

Wu, S., Zhang, Y., 2007a. Lomets: A local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35 (10), 3375.

Wu, S., Zhang, Y., 2007b. Lomets: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35 (10), 3375–3382.

Wu, S., Zhang, Y., 2008a. Anglor: A composite machine-learning algorithm for protein backbone torsion angle prediction. Plos One 2, 3400.

Wu, S., Zhang, Y., 2008b. Muster: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Struc., Func. Gen. 72, 547.

Wu, S., Zhang, Y., 2010. Segmer:identifying protein sub-structural similarity by segmental threading. Structure 18, 858.

Xu, D., Jaroszewski, L., Li, Z., Godzik, A., 2013. Ffas-3d: improving fold recognition by including optimized structural features and template re-ranking. Bioinformatics 30 (5), 660–667.

Xu, D., Zhang, J., Roy, A., Zhang, A., 2011. Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding and fg-md-based strcuture refinement. Proteins: Struct., Funct., Bioinf. 79 (10), 147.

Xu, D., Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Struct., Funct., Bioinf. 80 (7), 1715–1735.

Xu, J., Li, M., Kim, D., Xu, Y., 2003B. Raptor: optimal protein threading by linear programming. J. Bioinf. Comput. Biol. 1 (1), 95.

Xu, J., Li, M., Lin, G., Kim, D., Xu, Y., 2003. Protein structure prediction by linear programming. World Scientific, Singapure.

Xu, Y., Xu, D., 2000. Protein threading using prospect: Design and evaluation. Proteins: Struct., Funct., Bioinf. 40 (3), 343.

Xu, Y., Xu, D., Uberbacher, E., 1998. An efficient computational method for globally optimal threading. J. Comput. Biol. 5 (3), 597.

Yang, Y., Faraggi, E., Zhao, H., Zhou, Y., 2011. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27 (15), 2076–2082.

Zhang, Q., Veretnik, S., Bourne, P. E., 2005. Overview of structural bioinformatics. Springer, Heidelberg.

Zhang, Y., 2007. Template-based modeling and free modeling by i-tasser in casp7. Proteins: Struct., Funct., Bioinf. 69 (8), 108.

Zhang, Y., 2008. I-tasser server for protein 3d structure prediction. BMC Bioinf. 9 (40), 1.

Zhang, Y., 2008B. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18, 342.

Zhang, Y., 2009. I-tasser: Fully automated protein structure prediction in casp8. Proteins: Struct., Funct., Bioinf. 77 (S9), 100.

Zhang, Y., 2009-B. Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 19 (2), 145.

Zhang, Y., 2014. Interplay of i-tasser and quark for template-based and ab initio protein structure prediction in casp10. Proteins: Struct., Funct., Bioinf. 82 (S2), 175–187.

Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E., Skolnick, J., 2006. On the origin and completeness of highly likely single domain protein structures. Proc. Natl. Acad. Sci. U.S.A. 103 (8), 2605.

Zhang, Y., Kihara, D., Skolnick, J., 2002. Local energy landscape flattering: Parallel hyperbolic monte carlo sampling of protein folding. Proteins: Struct., Funct., Bioinf. 48, 192.

Zhang, Y., Skolnick, J., 2004a. Scoring function for automated assessment of protein structure template quality. Proteins: Struct., Funct., Bioinf. 57 (4), 702.

Zhang, Y., Skolnick, J., 2004b. Scoring function for automated assessment of protein structure template quality. Proteins: Struct., Funct., Bioinf. 57 (4), 702–710.

Zhang, Y., Skolnick, J., 2004B. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys. J. 87 (4), 2647.

Zhang, Y., Skolnick, J., 2004C. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. U.S.A. 101 (20), 7594.

Zhang, Y., Skolnick, J., 2004D. Spicker: a clustering approach to identify near-native protein folds. J. Comput. Chem. 25 (6), 20.

Zhang, Y., Skolnick, J., 2005. Tm-align: A protein structure alignment algorithm based on tm-score. Nucleic Acids Res. 33, 2302.

Zhou, H., Pandit, S., Skolnick, J., 2009. Performance of the pro-sp3-tasser server in casp8. Proteins: Struc., Func. Gen. 77 (S9), 123.

Zhou, H., Skolnick, J., 2007. Ab initio protein structure prediction using chunk-tasser. Biophys. J. 93, 1510.

Zhou, H., Skolnick, J., 2009. Protein structure prediction by pro-sp3-tasser. Biophys. J. 96 (6), 2119.

Zwanzig, R., Szabo, A., Bagchi, B., 1991. Levinthal's paradox. Proc. Natl. Acad. Sci. U.S.A. 89, 20.