Soft computing methods for the prediction of protein tertiary structures: A survey

Alfonso E. Márquez-Chamorro [a,*], Gualberto Asencio-Cortés [a],
Cosme E. Santiesteban-Toca [b], Jesús S. Aguilar-Ruiz [a]

[a] *School of Engineering, Pablo de Olavide University, Seville, Spain*
[b] *Centro de Bioplantas, University of Ciego de Ávila, Cuba*

A B S T R A C T

Keywords:

Protein structure prediction
Soft computing
Protein contact map
Support vector machines
Neural networks
Evolutionary algorithms

The problem of protein structure prediction (PSP) represents one of the most important challenges in computational biology. Determining the three dimensional structure of proteins is necessary to understand their functions at molecular level. The most representative soft computing approaches for solving the protein tertiary structure prediction problem are summarized in this paper. These approaches have been categorized following the type of methodology. A total of 90 relevant works published in last 15 years in the field of protein structure prediction have been reported, including the best competitors in last CASP editions. However, despite large research effort in last decades, a considerable scope for further improvement still remains in this area.

Contents

[*] Corresponding author. Tel.: +34 954062449.
E-mail address: amarcha@upo.es (A.E. Márquez-Chamorro).

1. Introduction: background and purpose

Proteins are an important class of macromolecules present in all biological organisms. They form the basis of cellular and molecular life and significantly affect the structural and functional characteristics of cells and genes [13]. Numerous functions, as structural support, mobility, protection, regulation or transport, are developed by proteins in the cells. Proteins are formed by union of simpler substances called amino acids. Amino acid sequence determines the structure of proteins and is the link between the genetic message in DNA and the three-dimensional structure which is associated to a biological function. Therefore, the knowledge of the sequence is essential to discover the protein functionality [8].

A protein can be seen on four different levels depending on which structures of the protein are considered. Essentially, primary structure of proteins consist of linear sequences of twenty natural amino acids joined together by peptide bonds. The secondary structure of a protein refers to the interactions due to a regular arrangement of hydrogen bonds between CO and NH groups (carboxyl and amino) of its amino acids, forming different motifs (α-helix, β-sheet, loops and turns). The tertiary structure is a description of the complex and irregular folding of the polypeptide chain in three dimensions. These complex structures are held together by a combination of several molecular interactions (e.g. ionic, hydrophobic or hydrogen bonds) that involve the amino acids of the chain. The quaternary structure is the final dimensional structure formed by all the polypeptide chains making up a protein [13].

A protein spontaneously folds into a 3-dimensional structure after having been manufactured in the ribosomes. A specific protein will fold in the same way and will end up with the same 3D structure. This phenomenon is called the native state of the protein [8]. Protein folding represents the process whereby higher structures are formed from the primary structure. A folded protein can have more than one stable folded state or conformation. Each conformation has its own biological activity. Anfinsen's experiment discovered that the amino acid sequence determines the native structure of a protein [1].

Sometimes, a protein can fold into a wrong shape. A single missing or incorrect amino acid could cause such a misfold. As already stated, protein function is determined by its structure, which can be inferred from the sequence of amino acids, therefore a misfold implies that a protein can not fulfill its function correctly. Alzheimer's disease, Cystic fibrosis and other neurodegenerative diseases are now attributed to protein misfolding. The knowledge of the misfolding factors and understanding the protein folding process, would help in developing cures for these diseases. Therefore, the knowledge of the structure of the protein provides a great advantage for the development of new drugs and the design of new proteins.

Despite of having a huge number of protein sequences as result of the genome sequence projects [86,101], the number of known protein structures is significantly lower than the proportion of known sequences. The difficult determination of these structures, using experimental methods such as X-ray or nuclear magnetic resonance (NMR), contributes to increase this gap between sequence and protein structures. Therefore, it is necessary the use of computational methods which predict 3D protein structures in a cheaper and faster way.

Different soft computing approaches have been developed to deal with the PSP problem. The main soft computing paradigms for the application of protein structure prediction are artificial neural networks (ANNs), evolutionary computation (EC) and support vector machines (SVMs). Furthermore, protein structure prediction methods can be further classified according to a biological approximation: homology-based methods, threading methods and *ab initio* methods.

Homology methods are based on the comparative of protein sequences with known structures. These methods are based on the hypothesis that similar protein sequences determine similar 3D structures.

Threading methods, also called sequence-structure alignment or fold recognition methods, try to align a protein sequence to a 3D structure. Threading methods are based on the idea that evolution conserves the structure rather than the sequence. On the other hand, *Ab initio* methods try to find a 3D model of the protein exclusively using the amino acid sequence, according to the laws of physics and chemistry.

In this paper, we present a survey on relevant methods of protein tertiary structure prediction based on soft computing techniques (neural networks, support vector machines and evolutionary computation). The remaining of this paper is structured as follows. Section 2 summarizes some basic concepts of PSP. The following sections describe all the different employed techniques (neural network, support vector machines, evolutionary algorithm, statistical methods and other predictive methods). Finally, last section summarizes the main conclusions of the study.

2. Preliminary concepts

In this section, several basic concepts related to PSP are briefly explained. First, protein structure prediction workflow is described. In the second place, the most relevant data structures used to represent the tertiary structure of a protein, such as the 3D models, *e.g.* torsion angle and lattice models, distance maps (DM) and contact maps (CM), are identified. Additionally, we specify the performance metrics for the validation of the prediction algorithms in this area. This is a particularly important issue, given the variety of methods and different measures employed, for the comparative analysis among algorithms.

2.1. Protein structure prediction workflow

The general steps for the PSP methodology are summarized in this section. Given a new protein sequence with an unknown structure, homology modeling can be considered as first step. This template-based methodology is based on the assumption that similar sequences encode similar 3D structures. If the target sequence does not have homologous proteins with known structure, fold recognition methods can be required. This type of protein modeling methods are based on the threading alignment. Structure templates are aligned with the target sequence by optimizing a scoring function based on statistical knowledge. If these two approaches turn out to be insufficient, we finally need *ab-initio* methods, which exclusively use the sequence information to predict the structure. First, a data input selection is required. Data input can consist of different features obtained from amino acid sequence. For instance, the frequency of appearance of amino acids in the sequence, physico-chemical properties of the residues, evolutionary information extracted from the sequence, such as position specific scoring matrices or correlated mutations, or 1D features prediction as secondary structures (SS) or solvent accessibility (SA), are usually employed as input data. The following step considers the selection of a data output model. Different representation models for the predicted structure, such as contact map or torsion angle model, are classified in the following section. The selection of an algorithmic technique for the prediction is also required. This survey is focused on the classification of the different algorithmic methodologies (statistical and soft computing approaches) for PSP. The quality assessment of the generated model is also crucial to understand the validity of the methods. Performance metrics section specify which are the most common assessment measures used in this area for the validation of models. Finally, the 3D model of the protein can be reconstructed using specific applications [62]. The majority of the works which are summarized in this survey are categorized as *ab-initio* methodologies.

2.2. Input data features

Different features of training dataset are used for the soft computing techniques reviewed. In the following, we explain some of the properties used for the algorithms.

Evolutionary information. Sequence alignment is a standard technique in bioinformatics for visualizing the relationships between residues in a collection of evolutionary or structurally related protein. Existing protein structure prediction algorithms in the literature have used position-specific scoring matrices (PSSM) and correlated mutations as input encoding.

The tendency of residue positions in proteins to mutate coordinately is called correlated mutation [40]. For each residue, its frequency of being correlatively mutated with respect to all other residues present in the same chain is calculated. This is computed by counting the number of times that two residues are either present or absent together and dividing it by the total number of counts.

On the other hand, PSSMs are also obtained from sequence alignments. PSSMs determine the substitution scores between amino acids according to their positions in the alignment. Each cell of the matrix is calculated as the log_2 of the observed substitution frequency at a given position divided by the expected substitution frequency at that position. Thus, a positive score (ratio >1) indicates that the observed frequency exceeds the expected frequency, suggesting that this substitution is surprisingly favored. A negative score (ratio <1) indicates the opposite: the observed substitution frequency is lower than the expected frequency, suggesting that the substitution is not favored.

Physico-chemical properties. The most direct information we can extract from the primary sequence of a protein are physico-chemical characteristics of its residues. With this information, we can generate representations of, for example, how the hydrophobicity varies along the sequence of the protein and obtain information about hydrophobic areas, which may help the prediction of structural characteristics. Properties used in the literature are hydrophobicity, polarity, volume of residues, graph shape index and isoelectric point, among others.

Secondary structures. Secondary structure (SS) prediction consists of predicting the location of α-helices, β-sheets and turns from a sequence of amino acids. The location of these motifs could be used by approximation algorithms to obtain the tertiary structure of the protein.

Sequence separation distance. The separation distance between two residues is defined as $||i-j||$ where i and j are the sequence indices of two residues. According to the sequential distance, we can estimate which pair of residues is bonded. The higher the distance (>100) between two residues, the lower the probability of being bonded.

Protein length and protein molecular weight. Protein length indicates the number of amino acids of each sequence. Molecular weight of a protein is the mass of this molecule. It can be calculated as the sum of the individual isotopic masses of all the atoms in the molecule. These features correspond to the representation of global information of a protein sequence.

2.3. Output data models

Output data models for the representation of tertiary structure of proteins are summarized in this section.

- Torsion angle model

  Torsion or dihedral angles (Φ, Ψ) represent the position of the atoms of an amino acid chain where Φ involves the N—C_α bond and Ψ involves the C_α—C bond. These rotations determine each protein structure. A possible representation of the amino acids of a protein is $[(\Phi_1, \Psi_1)\ldots(\Phi_n, \Psi_n)]$ where n reflects the residue length. The Ramachandran plot, described in [91], allows to avoid the possible collisions among atoms.
- Lattice model

  Other representation of the protein structure can be the lattice models. Each amino acid can be represented by a pair (x, y) where x and y are the coordinates of a 2-dimensional lattice. Considering the possible number of movements to the next point, another representation could be direction vectors, $(L_1, L_2 \ldots L_n)$ where $L_i \in \{UP, DOWN, LEFT, RIGHT\}$ are the locations of each amino acid with respect to the previous one.
- Binary contact map

  A protein contact map is a bidimensional square $L \times L$ matrix, where L represents the residue length. The upper triangle of the matrix represents the observed part and the lower triangle reflects the predicted part. An element of the contact map represents a pair of amino acids (i, j) which are in contact (1) or not (0). If the distance between the i and j is less than or equal to a given threshold, a contact is established. To this aim, a commonly used threshold is 8 angstroms (Å), as in [76]. Alpha carbon (C_α) and beta carbon (C_β) are the reference atoms, to measure the distance between amino acids, most commonly used [35]. Given a contact map of a protein, it is possible to reconstruct a 3D model of the protein backbone, solving the Molecular Distance Geometry Problem (MDGP) described by [62]. Contact maps, as protein structure representation, are also useful to compare protein structures, using the maximum contact map overlap described in [29].
- Distance matrix

  A distance map is a bidimensional square $L \times L$ matrix for the representation of the distances between the residues of a protein. This representation is based on the assumption that the main drawback of binary contact maps is the data loss that occurs in the discretization. The calculation of the distances between the residues is determined by the Euclidean distance. Estimated distances are located in the lower triangle and real distances are located in the upper triangle.

2.4. Experimental validation

In order to promote research in this field, the performance of current methods is assessed in the CASP competition (Critical Assessment of Techniques for Protein Structure Prediction) every two years [78].

2.5. Performance metrics

The quality measures used to evaluate the reliability of the 3D models are:

- Root mean square deviation (RMSD), that represents the absolute deviation (in Å) of individual C_α atoms between the model and the known true structure. The formula for RMSD is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|r_i^{model} - r_i^{real}|^2} \tag{1}$$

where r_i^{model} and r_i^{real} are the positions of ith C_α atoms in the model and the real protein, respectively.

- Global distance test-Total score (GDT_TS) is defined in [119] and is used as major assessment criteria in CASP, and describes the percentage of well-modeled residues in the model with respect to the target. The definition of GDT_TS is shown in Eq. 2:

$$GDT\_TS = 100 \times \frac{\sum_{d_i} \frac{GDT_i}{NT}}{4} \tag{2}$$

where GDT_i is the number of α-carbons of a prediction not deviating from more than an established cutoff d_i (in Å) from the α-carbons of the targets, after optimal superimposition. NT is the number of amino acids of the protein and $d_i \in \{1, 2, 4, 8\}$ expressed in Å.
- TM-score or template modeling score, detailed in [125], measures the global structural similarity between the model and template proteins, according to the distances of each pair of residues. TM-score is defined as:

$$TM\_score = \max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \right] \tag{3}$$

where L_T is the number of aligned residues, L_N is the number of residues, and d_i is the aligned residue distance. The range of values of this measure is (0, 1], the better the higher.

For the accuracy assessment of contact maps, other three measures are also employed. Accuracy represents the number of correctly predicted contacts. Coverage reflects the proportion of predicted contacts divided by the real contacts. Therefore, *accuracy* = C/C_p and *coverage* = C/C_t where C_t represents the real contacts of the protein, C is the number of correct predictions (true positive rate) and C_p reflects the total of predicted contacts. X_d represents the distribution accuracy of the predicted contacts. X_d is defined by Eq. 4:

$$X_d = \sum_{i=1}^{15} \frac{P_i - P_a}{i} \tag{4}$$

where P_i reflect the number of estimated pairs whose distance is in the range $(4(i-1), 4i)$ and P_a represents the number of real pairs whose distance is also in the range $(4(i-1), 4i)$.

3. Neural network methods

The basic principle behind an artificial neural network (ANN), is that the ANN can be trained to recognize patterns of known amino acid structures and use them to predict protein tertiary structures or protein contact maps.

As stated before, ANNs are one of the most popular methods for the tertiary structure prediction. Some ANN proposals base the prediction of contact maps (CMs) on chemico-physical properties and structural information of the amino acids. In [12], authors present a feed-forward neural network which is trained with matching sets of amino acid sequences and two different types of structural information: the corresponding secondary structure and the contact map of known proteins. The input of the network consists of a window of 61 residues. The hidden layer has 300 neurons and the output layer 30 neurons used to predict contacts between the amino acid that occupies the central positions in the window and the rest of the residues of the window. Three additional neurons are used to predict the tertiary motifs. The method proposed in [39] is based on two input neurons which use input vectors with information about the residue–residue contact and its environment, the number of amino acids and the separation between amino acids. In addition, several variables are added, such as the hydrophobicity of the environment, as well as evolutionary information. This work was enhanced with the addition of correlated mutations in [40]. Also a filtering procedure is added to the predictor, to avoid contact overprediction, taking into account the residue–residue interactions that can establish. The technique proposed in [43], adopts two layer feed-forward ANN. The ANN is trained using the results of a study about correlation between sequence separations and distances of each amino acid pair. The method described in [121], improves a method based on radial basis function neural network including a binary encoding scheme for learning the inter-residue contact patterns.

Another popular type of ANNs for contact map prediction is represented by recurrent neural networks (RNN). In [106], authors introduce a predictor based on ensembles of two-layered BRNNs (bidirectional recurrent neural networks). The method classifies the components of the principal eigenvector (PE) and uses predicted secondary structure information and hydrophobicity interaction scales.

Evolutionary, as well as structural information are employed in various ANNs. SPINE-2D [116] consists of two neural networks using one and two layers, respectively. These networks use 34 features as input, including position specific scoring matrices (PSSMs) generated by PSIBLAST ([4]), seven physico-chemical properties of amino acids, including hydrophobicity, volume, polarizability and secondary structure from the DSSP assignment program (Dictionary of Secondary Structure of Proteins) ([52]). In [65], authors propose a novel hybrid architecture based on neural and Markov logic networks with grounding-specific weights, in order to predict beta contacts. Multiple sequence alignments (MSA), SS and two states of SA are used as input data. The method described in [98], predicts contacts using correlated mutations and applies a random sampling of the majority class. In [88], author propose a method that combines different sources of information as protein properties, biophysical features, evolutionary profiles, secondary structure prediction and alignment information. This method, called PROFcon, achieves good accuracy levels for long-range contacts (inter-residue separation of 24***). A machine learning approach for contact map prediction, named CMAPpro, is developed in [30]. This approach consists of three steps. First, two neural networks predict contacts and secondary structure elements. Second, an energy-based method predicts contact probabilities between residues in secondary structure elements. Finally, a deep neural network architecture organizes and refines the prediction of contacts. A hybrid method, described in [37], combines an ANN and a boosting technique. This method employs GPU resources due to the use of extensive profiles of features. These profiles are classified according to the contact range: short, medium or large. In [32], authors present a contact map predictor based on subnetworks which run in parallel. The encoding is performed using conservation weights and SS prediction from HSSP files (homology derived secondary structures of proteins), number of residues and separation between them and the length of the subsequences. Two content windows and a segment window incorporate local information of the contact.

Two networks approaches are also used for the torsion angle predictions. The method proposed in [7], employs PSSM for the torsion angles prediction. Torsion angles are discretized in 5 intervals. Output data is represented as a vector of discretized values for each amino acid of the target sequence. The method described in [34], performs a 3D model from a prediction of the torsion angles. This method uses a neural network with several intermediate layers and the input data is constituted by subchains of characters. These chains are sequence fragments of known structures with values of SS predictions and real torsion angles.

Some approaches predicts inter-residue protein distance maps. In [124], authors present a method for the protein distance matrix prediction. This method employs a neural network whose

Table 1
Summary of neural network methods for tertiary structure prediction.

Method	Ref.	Year	Acc.(%)	Dataset size	Description
	[39]	1999	16.0	408	Evolutionary information, AA
distanceP	[43]	1999	70.3	9	Correlation between distances
	[40]	2001	21.0	173	Correlated mutations
	[124]	2004	80.0 (cov)	40	Distance map prediction
	[121]	2005	32.0	173	RBFNN, binary encoding scheme
PROFcon	[88]	2005	<20.0	748	SS pred., evolutionary profiles
	[106]	2006	36.5	327	Ensembles of 2-layered BRNN
	[98]	2007	58.0	93	Correlated mutations
NNCON	[103]	2009	31.0	48	2D-RNN, β-sheet conformation
	[65]	2009	47.3	80	Markov logic networks
SPINE-2D	[116]	2009	23.0–26.0	500,CASP7	PSSM, AA properties
A3N	[34]	2010	1.92 (RMSD)	PDB, 5	Torsion angles
	[112]	2010	6.4 (RMSD)	CASP8, 24	Threading, eigenvectors
	[7]	2012	85.0	CULLPDB	Torsion angles, PSSM
CMAPpro	[30]	2012	30.0	CASP8,CASP9	SS
	[37]	2012	29.0	CASP9	Boosting
CNNcon	[32]	2013	58.8	1082	Cascaded NN
	[61]	2014	0.5 (RMSD)	CASP7, CASP9	Distance map, 2D-RNN

parameters are previously optimized by a genetic algorithm. Authors develops an *ab initio* approach based on a 2D Recursive NN, described in [61]. As input data the consider evolutionary information, SS, SA and contact density information. The output of the neural network is the distance map. Furthermore, this approach also generates a template-based model (TBM).

2D-Recursive NN are employed in two contact map predictors. The method described in [103], was ranked as one of the most accurate methods from CASP8 [104]. This method performs two steps. First, a 2D-Recursive NN predicts a residue–residue contact map. After that, an ANN predicts the special β-sheet conformation. Distill_roll, described in [75], is a neural network based method that works in two phases. The first one is a fold recognition stage dependent on sets of protein features predicted by machine learning techniques. The second phase involves an optimization algorithm that searches the space of protein backbones under the guidance of a potential based on templates found in the first stage. The structures are fitted directly to distance maps of templates rather than to predict contact maps. Distill_roll uses PSI-BLAST and PSSM to search the PDB for templates. The resulting 1D predictions are used to find remote homologues in the PDB. The inputs for the neural network are the sequences, the MSA and the templates found in the first stage.

Finally, the method described in [112] combines an *ab-initio* approach with classical threading procedures. This approach generates 3D models from known structures. In order to find the best structural templates, the algorithm calculates eigenvalues and eigenvectors from the protein contact matrix.

Neural networks provide a high degree of flexibility. Besides encoded input vectors of pair of amino acids, we may include neurons with additional information, *e.g.* sequence length, hydrophobicity values of the environment or evolutionary information. On the other hand, neural networks have certain limitations, *e.g.*, restriction in the encoding of input data, the use of appropriate parameters of the ANN and overfitting.

All the cited ANN methods for tertiary structure prediction and their achieved results are summarize, in chronological order, in Table 1.

4. Support vector machines

SVM techniques performs classification tasks building a hyperplane in a multidimensional space, trying to maximize the margin between each different classes. The function that performs the transformation of the space is called kernel function. SVMs are used as a machine learning tool to predict tertiary structure from the primary sequence.

Several contact map predictors are based on SVM approaches. In [128], authors use as data input MSA, correlated mutations and several amino acid physico-chemical properties and SS. Authors analyze the reliability of their method according to protein structural classes. The method described in [22] employs as input features SS, SA, pairwise information, contact potentials or local window feature. In [113] and [68], authors use evolutionary information for the contact map prediction. In [113], authors develop a composite set of nine SVM-based contact predictors that are used in [93] in combination with sparse template contact restraints. They use the original energy function of I-TASSER and contact predictions generated by extended versions of SVMSEQ ([114]). Authors propose a hierarchical scheme for contact prediction in [68]. This method uses contact propensities combined with MSA, SA and helical features.

On the other hand, [46] develops a fold recognition method based on SVM and PSIBLAST. The transformation of the MSA vector is used as input of the SVM obtaining a probability of the relation between the template and a query sequence.

Two methods have also been developed to predict disulfide bonds (contacts between cysteines). In [95], authors present a SVM approach combined with HMM for the protein disulfide contacts in eukaryote cells. As input data, the use external subcellular localization of the proteins and evolutionary profiles. The method showed in [96], employs MSA to detect correlated mutations. Sparse inverse covariance is calculated to avoid the possible false positives. They develop a support vector regression to detect the disulfide contacts using global and local characteristics of the proteins, as well as PSSM, which constitute the input profiles.

Finally, in [130], authors develop a support vector regression technique which detects the native structures of a protein among decoy sets. This technique uses contact energy based score, amino acid network score and the fast Fourier transform.

As limitations of these type of methods, we can cite that the kernel models overfits the model selection criterion, the difficulty in the selection of the optimal kernel function parameters and the algorithmic complexity and extensive memory requirements in large-scale tasks.

SVM methods for tertiary structure prediction are summarized in Table 2.

5. Evolutionary computation

Evolutionary computation (EC) is a optimization technique based on the Darwinian concepts of evolution. Evolutionary algorithms (EAs) and Genetic algorithms (GAs) represent two subtypes of EC.

Table 2
Summary of SVM methods for tertiary structure prediction.

Method	Ref.	Year	Acc.(%)	Dataset size	Description
	[128]	2002	22.4	177	Sequence profiles, correlated mutations, SS
	[46]	2005	46.0	16	Estimating the significance of the alignments
SVMcon	[22]	2007	21.0	48	SA, SS, pairwise information
	[68]	2009	56.0	52	Propensities, evolutionary profile, SA
	[113]	2011	31.0	273	I-TASSER, sparse template contact restraints
	[95]	2011	66.0	1797	HMM, subcellular localization
	[96]	2013	51.0	1797	SVR, inverse covariance, PSSM

Methods based on EC may use various possible representations of a protein structure: dihedral or torsion angles and lattice models (direction vector representation and hyprophobic–polar model). Torsion angles and direction vector were described in Section 2. HP (hyprophobic–polar) models are detailed in [31]. In this model, a sequence is represented as a string $s \in (H, P)^+$, where H represents a hydrophobic amino acid and P a polar or hydrophilic one.

We grouped the different evolutionary approaches according to the representation models described. We start with methods that use dihedral or torsion angles. In [26], authors develop a 3-torsion angles representation (Φ, Ψ, ω). No mutation operator is used. The fitness function consists of hydrophobic interactions and van der Waals contact measures. [97] develops a GA for a force field model, that represents chemical reactions and physical forces that occur in a protein. In [54], authors use a fitness function based on the Chemistry at HARvard Macromolecular Mechanics (CHARMM) force fields ([16]), to evaluate the potential energy values, and a scatter search algorithm. For the representation, authors use some amino acid features,*e.g.*, partial charge and van der Waals bond. In [82], a global free energy function of an unfolded conformation is calculated as the sum of threes types of energy: local backbone electrostatic energy, which represents the sum of the interactions between N—H and C═O groups among amino acids, intra-molecular electrostatic energy, that reflects the rest of intra-molecular interactions, and solvation free energy, that takes into account different interactions as hydrogen bonding, ion–dipole, and dipole–dipole attractions or van der Waals forces. Finally, in [127], author develops a tabu search algorithm for the torsion angles prediction. This tabu search, included in the mutation operator, improves the capacity of the local search of the algorithm. An off-latice model, which includes an energy function for the sequence of monomers, is incorporated to represent the amino acids.

Direction vector representation is employed in [15]. This approach uses a three-dimensional protein representation with 32 possible movements for each protein residue. The fitness function analyzes some protein characteristics like hydrophobicity, charge, and side-chain size.

HP models are used in several evolutionary proposals. For instance, a method that adopts a two-dimensional square lattice based on HP model is proposed in [105]. In [64], authors use a hybrid algorithm consists of Monte Carlo optimization and an HP square model. In [25], authors develop a HP model with cubic lattice implementation. This method adopts a fitness function based on the Kronecker-delta function,[1] distance between target residues, overlap involving the residues and free contact energy between target residues. A coefficient evaluates possible penalties due to violation movements.

Various contact and distance map predictors are also based on EAs. A distance matrix representation is presented in [87]. The method analyzes possible distances between each pair of amino acids for each protein. Fitness function is calculated using three terms: two penalty constraint factors and a hydrophobicity interaction term. This method also describes a repair algorithm and a penalization strategy for distance map unfeasible solutions. The method proposed in [44] starts with an initial random contact map population for a given amino acid sequence. A neural network and four physical protein properties (charge, sequence distance, neighborhood hydrophobicity and degree of vertices) are used in the fitness function. The most accurate contact map is selected after last generation. Later, this contact map is compared with a contact map template for each fold using graph theory. The maximum scoring template determines the fold of the protein. In [122], authors propose an EA that employs a 19-bit representation for a protein, where bits 0–8 represent each possible pairwise between amino acids, bits 9–12 represent a residue classification (polar, non-polar, acid or base), bits 13–15 represent which possible secondary structure a residue is among helix, sheet and coil. Bits 16–17 represent the sequence length and bits 18–19 represent the sequence separation. A GA is used to improve a radial basis function neural network. The method proposed in [21] is based on genetic algorithm classifiers (GaCs) for long range contacts prediction. These contacts have a sequence separation between amino acids of more than 24 residues. This method incorporates the sequence profile centers (SPCs). The method described in [69], uses a Self-organizing map described in [56] and Genetic Programming (GP) approach to predict protein contacts. Finally, [9] presents a system based on the prediction of some structural features of protein residues such as SS, SA, Recursive Convex Hull (RCH) and coordination number (CN) and a learning system based on genetic algorithms, called BioHEL. The generated rules by this system contain human-readable explanations and insight information about the residue contact predictions.

An EA that incorporates a local search phase is called a Memetic algorithm. Multimeme algorithms (a type of Memetic algorithms) use several kinds of local searches. This paradigm is adopted in the following four methods. In [60], author introduces a Multimeme algorithm with using a Functional Model Protein ([11]) and a HP model in 2D or 3D. This method analyzes the compatibility of a new offspring using a strategy based on the memory of compatible contacts. The author presents a combination of fuzzy logic and Multimeme algorithms in HP models in [83]. Fuzzy logic is used as a modifier of the memepool local searchers, evaluating possible solutions. Another Memetic algorithm was proposed in [47]. This method uses a HP model and calculates the fitness function, with two new parameters called *H-compliance* and *P-compliance* which measure the situation of the residue according to the hydrophobic core.

In [21], authors implement an ensemble of GA classifiers to predict long-range contacts. The individuals of the GA represent three amino acid windows and 20 properties obtained from the HSSP database of protein structure–sequence alignments ([33]) for each residue in such windows. The method also uses the sequence profile centers (SPCs).

PSP problem can be seen as a multi-objective optimization problem. Multi-objective optimization evolutionary algorithms (MOEA) are described in the following five methods.

[1] The function is 1 if the variables are equal, and 0 otherwise.

Table 3
Summary of evolutionary computation methods for tertiary structure prediction.

Method	Ref.	Year	Acc.(%)	Dataset size	Description
	[82]	1997	3.1 (RMSD)	28	Global free energy function
	[87]	1998	–	3	Distance matrix representation
	[26]	1998	1.48–4.48 (RMSD)	5	Torsion angles, fitness interaction
	[97]	2000	1.08 (RMSD)	3	Force field model
	[64]	2001	–	8	Monte Carlo optim., HP model
	[15]	2002	–	2	3-D protein representation
	[60]	2002	–	200	Multimeme algorithm, HP model
	[25]	2003	–	8	HP model with cubic lattice
	[69]	2004	21.4	CASP5	Self-organizing map, GP
MOFASA	[100]	2004	53.0	27	Folding recognition
	[44]	2005	69.0–88.0	24	Contact map representation
FANS	[83]	2005	–	4	Fuzzy logic and multimeme alg.
	[27]	2006	3.6 (RMSD)	5	I-PAES, torsion angles, CHARMM
	[122]	2007	–	61	Residue properties representation
	[54]	2008	9.43 (RMSD)	2	Fitness function CHARMM based
	[18]	2009	1.8 (RMSD)	2	CHARM
	[47]	2009	–	9	Memetic algorithm, HP model
MI-PAES	[51]	2009	4.23 (RMSD)	4	Torsion angle model
	[21]	2010	21.5	480	Sequence profile centers (SPCs)
	[127]	2010	–	4	Local tabu search
Pitagoras-PSP	[19]	2011	9.15 (RMSD)	CASP8	PAES, torsion angles
BioHEL	[9]	2012	25.7	CASP9	GA, contact rules
MECoMaP	[72]	2012	54.0	173	MOEA, decision contact rules, AA

A method described in [18], develops a parallel MO algorithm using NSGA-II. A torsion angle model is generated and a CHARMM energy function is employed by this algorithm. MOFASA [100] attempt to solve the protein folding problem, incorporating a MOEA and a feature selection method. [27] presents I-PAES algorithm which is used to explore the conformational space searching for the minimal energy. The method uses a torsion angles model representation and CHARMM equation as fitness function. In [51], authors develop a MOEA, which represents protein structures by torsion angles. They modified the classical algorithm PAES, introducing two immune inspired operators. A MOEA is also presented in [19], called Pitagoras-PSP. This algorithm uses an evolutionary *ab initio* approach based on PAES. The algorithm predicts protein torsion angles and uses an energy function as fitness function. Mutation operators maintain values of torsion angles in feasible ranges according to secondary structure of residues. Finally, in [72], authors perform a MOEA for the residue–residue contact prediction based on physico-chemical properties (H,P and net charge) of amino acids and structural features of the proteins (SS and SA). The algorithm generates rules which specify the residue–residue contact criteria, according to the cited features.

The limitations of these type of methods are the difficulty to find a stop criteria, and the possibility to converge to a local maximum as result of an adverse parameter configuration. A correct choice of the representation of the problem, the fitness function, the size of the population and the rate of the genetic operators should be taken into account. For instance, a small size of the population can cause that the EA cannot explore the sufficient space to find a correct solution.

A summary of evolutionary computation methods for tertiary structure prediction is shown in Table 3. In case the value of Accuracy (third column) is not provided by the authors, it is marked by dash.

6. Statistical approaches

Typically, the statistical approaches for PSP are based on homology models. Many techniques based on the comparison of 3D structures have been proposed.

In [126], author develops a prediction server called I-TASSER which is a homology-based protein structures and function predictor. This method was ranked as the No 1 server in recent CASP competitions. Profile–Profile threading Alignment (PPA) and the Threading ASSEmbly Refinement program are the main components of this server application. HMM (Hidden Markov model) and Monte Carlo simulation are also used during the prediction stage.

In [53], a method named SAM-T08, uses HMMs and provides in addition to the 3D model, other information such as MSAs or predictions of protein contacts.

Two protein folding recognition methods are described in [81,92]. The first method introduces a statistical approach for folding recognition. This method is based on the use of sequence conservation[2] and sequence correlations.[3] These properties are extracted from the multiple sequence alignments. It is shown that sequence conservation and correlation provide enough information to detect incorrectly folded proteins. The second method is based on Bayesian networks. This approach is focused on protein fold and superfamily recognition. This Bayesian network also includes HMM's.

In [129], authors propose several approaches to predict contact order from the amino acid sequence only. A first approach is based on a weighted linear combination of predicted SS content and amino acid composition. A second approach is based on sequence similarity to known three-dimensional structures.

Two recent methods combine two different methodologies: *ab initio* and homology modeling. In [115], authors implement an *ab initio* approach, called QUARK, based on NN (nearest neighbors), Monte Carlo simulation and template-based matching algorithm. NN predicts structural features, a global fold is generated by Monte Carlo simulation and threading alignments are carried out. The size of the fragments are not fixed, and varies from 1 to 20 residues. A semi-reduced model, consists of full backbone atoms and residues center of mass (SC), is employed for the representation which is based in both Cartesian and torsion angle models. A force field model based on 11 terms is used as search engine. The method presented in [55], called GalaxyWEB, combines homology and *ab initio* modeling. In the initial phase, templates are selected according

[2] Protein sequence conservation represents the appearing of the same residues at analogous parts of different proteins due to the presence of equivalent functionalities.

[3] Sequence correlations are based on the idea that certain residue substitutions commonly occur in homologous proteins.

Table 4
Summary of statistical methods for tertiary structure prediction.

Method	Ref.	Year	Acc.(%)	Dataset size	Description
	[81]	1999	76.2	71	MSA, sequence conservation
	[92]	2002	77.0	25	Superfamily recognition
	[111]	2008	5.8 (RMSD)	1507	Monte Carlo simulation
	[129]	2008	74.2	499	SS weighted linear combination
I-TASSER	[126]	2009	34.0	CASP8	Protein structure alignment
SAM-T08	[53]	2009	61.4 (GDT-TS)	CASP8	HMMs and MSAs
GalaxyWEB	[55]	2012	68.5 (GDT-TS)	68,CASP9	HHSearch, TBM
QUARK	[115]	2012	69.1 (TM-SCORE)	51,94,CASP9	NN, Monte Carlo simulation
	[17]	2013	61.0	916	β-sheet contacts, max. entropy
	[38]	2013	78.0	17	Max. likelihood, Potts model
	[74]	2013	52.0	15	Substitution probabilities

to their scores using a HMM's. Then, the method refines the terminus regions and loops using a conformation space annealing (CSA) global optimization to avoid the inconsistencies.

The three following works perform their prediction with the aid of MSA's. The method proposed in [38], bases its contact prediction on the correlated mutations from protein MSA. This method infers a Potts model solving an optimization searching problem of pseudo-likelihood maximization. Finally, the method described in [17], is focused on the prediction of parallel and anti-parallel β-sheets contacts. To this aim, a measure of correlation between the distributions of amino acids from the protein MSA is calculated. It also includes an estimation of the maximum entropy values of these distributions to avoid the false positives. ZHOU-SPARKS-X, described in [118], uses both sequence profiles from MSA, and structure profiles, including SS, SA surface area and main-chain torsion angles. The method employs statistical error potentials to estimate the agreement between the native template structure and improved predicted structural properties of the query sequence. The query sequence was aligned with a pre-compiled structural library. The template with the highest alignment scores is selected for model building. A refinement program is used to link the models of different parts of the query sequence and remove clashes by using a potential function. The predictor is trained using the support vector machine library libsvm.

Two threading strategies are described in the following two proposals. In [111], score function is based on the effective connectivity profiles (ECP). The stochastic Monte Carlo algorithm is used for the search of the best template. This search tries to minimize the scoring function previously cited. Residue–residue contacts between the target objective and the templates are considered to determine the best fit. RaptorX-Roll, described in [84], is a single-template protein threading method with a probabilistic graphical model. The method is based on a function that estimates the log-likelihood of alignment state transition between consecutive amino acids based upon protein features. RaptorX-Roll uses neural networks to construct the mentioned function. The model parameter vector consists of all the parameters of 9 neural networks, which are trained by a set of non-redundant sequence-template pairs. The reference alignments used to train the parameter are generated by a protein structure alignment tool named DeepAlign developed by the authors of RaptorX-Roll.

The method proposed in [74], detects co-evolving site pairs by using substitution probabilities as well as physico-chemical properties changes, such as β-turn propensities, hydrogen-bonds and aromatic propensities. This method is specialized in transmembrane proteins.

A summary, in chronological order, of the methods described in this section is shown in Table 4. The first and second columns indicate the name of the method and its reference respectively. The third column represents the available accuracy achieved by the method. The forth column indicates the size of the data set of proteins, while the fifth column shows main characteristics of the algorithms.

7. Other predictive methods

In addition to the cited proposals to PSP, there are other important approaches, such as random forest algorithm, integer linear optimization and sparse inverse covariance. In this section, we will cover some of these strategies.

7.1. Mathematical models

Various contact map predictors are based on mathematical models. The six following methods belong to this category. [41] describes a consensus contact prediction method based on an integer linear programming (ILP) model. This method determines a weight for each prediction server according to an ILP and allows to discern between real and false contacts. In [89,90], author propose an integer linear optimization approach which calculates force field to predict residue–residue contacts in alpha-proteins minimizing the contact energies. This method allows to establish a set of constraints based on distances of elements of SS. The method described in [110], improves a mathematical optimization model to predict the contacts in transmembrane alpha proteins. Physical constraints were also incorporated in the mathematical model. The method presented in [6], employs multiple evolutionary templates extracted from MSA to enhance the protein contact map prediction.

7.2. Correlated mutations

Correlated mutations are employed in the following two methods for the PSP prediction. In [49], author develops PSICOV, which introduces an estimation of the inverse covariance applied to the contact map problem. Data input is obtained using correlated mutations from MSA. This methodology was combined with a fragment assembly method called FRAGFOLD in [57]. This method predicts protein contacts using statistical potentials. The method developed in [80], is focused on the prediction of 3D models of transmembrane proteins using SS and correlated mutations. Sparse inverse covariance is employed for detecting false positives in the correlations. Once the contacts are predicted and the assembly-fragment is carried out, the protein is reconstructed and its 3D model is obtained.

7.3. Evolutionary information

Evolutionary information is employed in the following seven methods of the literature. In [10], authors present a novel HMM method for contact map prediction which employ MSA and SS predictions as input data. The method proposed in [107], describes a multi-level combination approach to improve the various steps in PSP combining complementary and alternative templates, alignments and models. This approach, called MULTICOM, incorporates five automated PSP servers and one human predictor. The method presented in [70], is based on MSA statistics to determine evolutionary constraints from a set of homologous protein sequences. The inferred residue pair couplings constitutes enough information to define an accurate 3D protein fold model. The method described in [102], determines the protein contacts from MSA. SS predictions are included in a structure-based model. This model has two components, for local and non-local interactions which contributes to the reconstruction of the target protein. The method presented in [77], obtains the correlated mutations from MSA for the contact map prediction and use the problem of the maximization of the entropy to avoid false contacts.

In [94], authors propose a homology method for the protein contacts prediction. This clustering approach (*K*-means) generates a database of cluster of sequences with similar structures using a similarity structure measure. This method employs physico-chemical properties and evolutionary information to perform a cross-validation using three machine learnings approaches: DT, SVM and RF. Recently, [23] proposes a Hopfield–Potts model to analyze the residue coevolution from the detected correlated mutations from MSA. This method, which calculates eigenvalues of correlation matrix, improves the performance and computational time with respect to other similar approaches as PCA (principal component analysis) or DCA (direct coupling analysis).

7.4. Random forest

The three following methods are based on Random forest (RF). In [63], authors develop ProC_S3, which uses a set of RF algorithm-based models. Some characteristics of the algorithm are the use of a propensity matrix between residues and a classification of amino acids according to probabilities. The method described in [108], predicts global contacts between alpha-helices of trans-membrane proteins. As input data, large profiles are built using evolutionary information. An attribute selection with CFS and BestFirst is also performed by the algorithm. Finally, in [109], authors present PhyCMAP, a random forest and integer linear programming method which takes into account evolutionary information (correlated mutations) and physical constraints for the contact map predictions. First, the algorithm calculates the contact probabilities between each pair of residues, according to their evolutionary information. Then, they rank these probabilities and establish the physical constraints to reduce the conformational space possibilities.

7.5. Other approaches

In [36], authors perform a conformation ensemble approach. The method collects various models (SVMCon, TASSER and ROSSETA) and complementary information from a variety of methods to enhance the predictions.

A sequence-based protein contact map prediction method, named LRcon, based on logistic regression is detailed in [117]. A feature vector is fed into the logistic regression-based algorithm to make a consensus prediction for each residue pair.

JUSTcon is described in [2]. This contact map predictor combines KNN with a neuro-fuzzy inference system. The features of the input vectors include a PSIBLAST profile and SS and SA predictions. The window size is selected by a simple expert system. The adaptive neuro-fuzzy inference system predicts contacts from the given samples using an affinity score matrix.

A contact-assisted model is proposed in [3]. The prediction of a small number of residue contacts can improve the accuracy of the structural predictions. A contact prediction-assisted category was incorporated in CASP10.

In [48], authors develop an *ab initio*-homology hybrid method, called Bagheerath. This method is based on a template based modeling and energy based function which generates several structures according to a systematic sampling of the conformational space of loop dihedrals. The authors conclude that *ab initio* methods are good for small proteins (<100 residues).

The method described in [99], is based solely in energy models. These models apply energy force fields and physical laws to the

Table 5
Summary of other methods for tertiary structure prediction.

Method	Ref.	Year	Acc. (%)	Dataset size	Description
	[94]	2006	1.19 (RMSD)	27	HMM
FragHMMent	[10]	2009	22.8	151	HMM
	[41]	2009	37.0	CASP7	Hybrid method
	[99]	2009	5.9 (RMSD)	6, CASP7	Energy force fields
	[89]	2009	66.0	48	Integer linear optimization
MULTICOM	[107]	2010	63.0 (GDT-TS)	120	Template-based approach
	[5]	2011	51	5130	Distance map prediction, AA
JUSTcon	[2]	2011	45.2	450	Nearest neighbor algorithm
WMC	[6]	2011	23.6 (PCC)	CASP8	Template-based approaches
	[36]	2011	30.0	CASP9	Hybrid method
ProC_S3	[63]	2011	26.9	1490	RF algorithm based model
EVfold	[70]	2011	2.7-4.8 (RMSD)	15	Corr. mutations
	[77]	2011	55.0 (cov.)	131	Max. entropy model, corr. mut.
	[110]	2011	56.0	5	Integer linear optimization
	[108]	2011	49.0	62	RF, attribute selection
LRcon	[117]	2011	41.5	846	Hybrid method
	[80]	2012	8.5 (RMSD)	28	Correlated mutations
PSICOV	[49]	2012	>50.0	118	Sparse inverse covariance
Bagheerath	[48]	2012	7.0 (RMSD)	80	Ab initio-homology
	[102]	2012	5.37 (RMSD)	15	MSA, structure-based model
	[3]	2013	34.9 (GDT-TS)	CASP10	Contact-assisted model
PhyCMAP	[109]	2013	36.3	36,CASP10	RF, ILP, physical constraints
	[23]	2013	70.0	15	Residue coevolution, MSA
TOPOL	[45]	2014	-	1	Distance matrix
FRAGFOLD	[57]	2014	0.54 (TM-score)	150	Ensemble methods
SSThread	[71]	2014	3.8 (RMSD)	74,21	Knowledge-based potential

Table 6
Summary of latest CASP competitions.

Edition	Year	# proteins	Method	Reference	Acc.(%)
CASP11	2014	208	Zhang-Server	Zhang [126]	61.10
CASP10	2012	145	Zhang-Server	Zhang [126]	56.48
CASP9	2010	144	QUARK	Xu and Zhang [115]	56.13
CASP8	2008	172	Zhang-Server	Zhang [126]	64.83
CASP7	2006	124	Zhang	Zhang [126]	78.03
CASP6	2004	87	TASSER-3D	Zhang [126]	56.23
CASP5	2002	67	Bujnicki	Kosinski et al. [58]	47.17

Table 7
Prediction accuracy values (%) for five recent methods using a 291 protein dataset and the 123 CASP10 targets, called Set600 and described in [109], and <90% sequence identify.

Method	Year	Reference	Top 5	$L/10$	$L/5$
291 DS					
PhyCMAP	2013	[109]	58.2	52.8	46.1
PSICOV	2012	[49]	48.3	41.8	35.8
NNCON	2007	[103]	22.4	18.2	15.2
CMAPPro	2012	[30]	52.7	46.9	41.6
EvFold	2011	[70]	44.2	38.9	34.2
CASP10					
PhyCMAP	2013	[109]	42.1	36.3	32.0
PSICOV	2012	[49]	35.0	27.7	22.6
NNCON	2007	[103]	28.6	23.9	18.8
CMAPPro	2012	[30]	38.0	33.6	29.7
EvFold	2011	[70]	32.8	25.7	22.5

atoms of the molecule. This method also uses dynamic molecular simulations.

Recently, a method for the distance matrix prediction is presented in [45]. It combines two different procedures. The first approach, called TOPOL, is based on the geometrical determination of the arc length between residues on the surface of the protein. The second method consists in a modification of the Dijkstra algorithm.

A knowledge-based potential described in [71], constitutes the base of the contact prediction of this method. These predicted contacts are located in alpha-helices or beta-strands. After the SS prediction, the algorithm performs the cluster prediction among the possible templates according to the Z-score. Last step involves the loop prediction for the finally reconstruction of the protein.

Finally, in [5], authors propose a novel nearest neighbor approach for distance matrix prediction based on physicochemical properties of amino acids. First, they construct sets of subsequence profiles based on physicochemical properties from known protein structures and classify them according to the first and last amino acids of fragments. Second, a search for most similar profiles is performed, producing predicted distances between amino acids. All these methods are summarized in Table 5.

Finally, Table 6 summarizes the main characteristics of the latest CASP winners which were described previously. First column indicates the edition of the competition. The second column shows the number of target proteins. The third and forth column present the name of the best method of each edition for PSP and its reference. The fifth column indicates the achieved accuracy (GDT-TS) of each method.

8. Comparative studies

In order to provide a clear comparison among the different described methods, we have performed two comparative analysis. The first comparative, showed in Table 7, presents the prediction accuracy of PhyCMAP [109], PSICOV [49], NNCON [103], CMAPpro and Evfold [70] for long-range C_α contacts (sequence distance >24), tested on identical conditions with a typical dataset (Set600), published in [109], and the 123 protein targets used in CASP10. CASP evaluators consider a list of predicted contacts sorted by their accuracy value. A number of predicted contact pairs proportional to the protein length is selected for the evaluation of a method. In this case, we have shown for the evaluation a number of pairs corresponding to: Top 5, $L/10$ and $L/5$ where L is the protein length. These recent five contact predictor methods, belongs to the different soft computing approaches defined in this survey. As we can see, maximum prediction accuracy results barely reach the 50%. None of the methods clearly outperform others, however for this dataset, we can highlight a better performance for the random forest [109] and deep neural network approaches [30].

Table 8 shows a comparison among the methods which produce the best results in the last finished CASP competition (CASP 10, in 2012). The table presents the results in domain classification for the Residue–Residue (RR) category, specifically for the subcategories free-modelling and hybrid template-based modelling/free-modelling. The methods were ranked according to sum of average Z-scores for measures Acc and Xd. The results were produced predicting long range contacts (separation >=24) and top $L/5$ predicted contacts (where L is the domain length). The column Dms specifies the number of domains predicted by the method. The columns Acc and Xd indicate the average accuracy and Xd measures for all predicted domains by each method. The Zsc_Acc and Zsc_Xd columns are added and averaged over the number of domains attempted by a method. Note that despite the low values of accuracy (the maximum values was 19.92 for MULTICOM-REFINE), Xd values are promising (up to 12.58 for MULTICOM-CONSTRUCT). Regarding these results, one of the most currently reliable methods for protein structure prediction is MULTICOM and its variants. According to the algorithmic kernel of best methods analyzed in CASP competition, hybrid approaches were the most frequent and reliable.

Table 8
Evaluation of predictions in RR category for CASP10 competitors.

GR Name	Dms	Acc	Zsc_Acc	Xd	Zsc_Xd	Zsc_Acc+Zsc_Xd	Ref.
MULTICOM-CONST	15	19.15	0.54	12.58	0.75	1.29	[107]
RaptorX-Roll	9	14.91	0.63	10.35	0.63	1.26	[84]
IGBteam	16	18.02	0.68	9.68	0.54	1.22	[30]
ZHOU-SPARKS-X	12	12.26	0.62	8.26	0.59	1.21	[118]
MULTICOM-NOVEL	15	19.12	0.46	9.93	0.68	1.14	[107]
MULTICOM-REFINE	15	19.92	0.48	9.88	0.65	1.13	[107]
SAM-T08-server	12	15.33	0.66	10.19	0.45	1.11	[53]
Distill_roll	16	16.42	0.67	9.75	0.42	1.09	[75]
MULTICOM	15	16.94	0.58	9.88	0.51	1.09	[107]
ProC_S4	15	16.71	0.55	9.93	0.43	0.98	[63]
ICOS	15	16.16	0.37	10.10	0.36	0.73	[9]

9. Conclusions

In the last 40 years, the problem of PSP has been tackled by multiple approaches. Nevertheless, a definitive solution has not been found. In these years, several advances in the field must be taken into account, as the development of new prediction algorithms using more powerful computers, the development of communal scientific competitions, such as CASP, the creation of PDB, a database of now more than 90,000 protein structures and the study of the called folding diseases (Alzheimer's, Parkinson's, etc.). Statistical and soft computing methods offer a wide variety of different approaches to address the PSP problem, obtaining satisfactory results in some cases. However, comparing the performance of the different approaches, we cannot draw clear conclusions to determine which is the best methodology. It depends on, the data set used, the input features of the machine learning algorithm, among other issues. Specifically, those methods which use evolutionary information from sequence alignments obtain better results than others.

As it has been shown, despite the relative maturity of this research discipline, progress has not been as satisfactory as might be desired, and the margin of improvement is still significant. The current trend in PSP methods is related to the correlated mutations and direct coupling analysis [70]. Latest methods combine co-evolutionary analysis as input data to enhance the predictions [50]. Furthermore, the methods are increasingly focused on the prediction of specific parts of the PSP problem rather than be a general approach, because of the complexity of the problem. For example, the β-sheets prediction methods [17], inter α-helices prediction methods [108,110], cysteine–cysteine prediction methods [95,96] or specific methods to some types of proteins, such as transmembrane ones [80].

Acknowledgements

This research was funded by the Junta de Andalucia under project P07-TIC-02611 "Intelligent systems for discovering patterns of behavior. Application to biological database" and by the Spanish MEC under project TIN2011-28956-C02-01.

References

[1] C. Anfinsen, The formation and stabilization of protein structure, Biochem. J. 128 (1972) 737–749.

[2] A.A. Abu-Doleh, O.M. Al-Jarrah, A. Alkhateeb, Protein contact map prediction using multi-stage hybrid intelligence inference systems, J. Biomed. Inform. 45 (1) (2011) 1.

[3] B. Adhikari, X. Deng, J. Li, D. Bhattacharya, J.A. Cheng, Contact-assisted approach to protein structure prediction and its assessment in CASP10, in: Proceedings on AAAI Conference on Artificial Intelligence, 2013, pp. 2–7.

[4] S.F. Altschul, T.L. Madden, A.A. Schffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 338.

[5] G. Asencio-Cortés, J.S. Aguilar-Ruiz, Predicting protein distance maps according to physicochemical properties, J. Integr. Bioinform. 8 (3) (2011) 181.

[6] H. Ashkenazy, R. Unger, Y. Kliger, Hidden conformations in protein structures, Bioinformatics 27 (14) (2011) 1941–1947.

[7] Z. Aydin, J. Thompson, J. Bilmes, D. Baker, W.S. Noble, Protein torsion angle class prediction by a hybrid architecture of Bayesian and neural networks, in: Conference on Bioinformatics and Computational Biology, 2012, pp. 2012–2018.

[8] J.M. Berg, J.L. Tymoczko, L. Stryer, Biochemistry, 2002 (W.H. Freeman).

[9] J. Bacardit, P. Widera, A.E. Márquez-Chamorro, F. Divina, J.S. Aguilar-Ruiz, N. Krasnogor, Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features, Bioinformatics 28 (19) (2012) 2441–2448.

[10] P. Bjrkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson, T.R. Hvidsten, Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contacts, Bioinformatics 25 (10) (2009) 1264–1270.

[11] B.P. Blackburne, J.D. Hirst, Evolution of functional model proteins, J. Chem. Phys. 115 (4) (2001) 1935–1942.

[12] H. Bohr, J. Bohr, S. Brunak, R.M.J. Cotterill, H. Fredholm, B. Lautrupt, S.B. Petersen, A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks, FEBS Lett. 261 (1) (1990) 43–46.

[13] P.E. Bourne, J. Gu, Structural Bioinformatics (Methods of Biochemical Analysis), Wiley-Blackwell, 2003.

[15] K.A. Braden, Simple approach to protein structure prediction using genetic algorithms, Stanford Univ. 426 (2002) 36–44.

[16] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, J. Comput. Chem. 4 (1983) 187–217.

[17] N.S. Burkoff, C. Varnai, D.L. Wild, Predicting protein beta-sheet contacts using a maximum entropy-based correlated mutation measure, Bioinformatics 29 (2013) 580–587.

[18] J.C. Calvo, J. Ortega, Parallel protein structure prediction by multiobjective optimization, Parallel Distrib. Netw. Based Process. 12 (4) (2009) 407–413.

[19] J.C. Calvo, J. Ortega, M. Anguita, PITAGORAS-PSP: including domain knowledge in a multi-objective approach for protein structure prediction, Neurocomputing 74 (16) (2011) 2675–2682.

[21] P. Chen, Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers, BMC Struct. Biol. 10 (1) (2010).

[22] J. Cheng, P. Baldi, Improved residue contact prediction using support vector machines and a large feature set, BMC Bioinform. 8 (2007) 113.

[23] S. Cocco, R. Monasson, M. Weigt, From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction, PLOS ONE 9 (8) (2013) e1003176.

[25] C. Cotta, Protein structure prediction using evolutionary algorithms hybridized with backtracking, Lecture Notes Comput. Sci. 2687 (2003) 321–328.

[26] Y. Cui, R.S. Chen, W. Hung, Protein folding simulation with genetic algorithm and supersecondary structure constraints, Proteins 31 (1998) 247–257.

[27] V. Cutello, G. Narzisi, G. Nicosia, A multi-objective evolutionary approach to the protein structure prediction problem, J. R. Soc. Interface 3 (2006) 139–151.

[29] P. Di Lena, P. Fariselli, L. Margara, M. Vassura, R. Casadio, Fast overlapping of protein contact maps by alignment of eigenvectors, Bioinformatics 26 (18) (2010) 2250–2258.

[30] P. Di Lena, K. Nagata, P. Baldi, Deep architectures for protein contact map prediction, Bioinformatics 28 (19) (2012) 2449–2457.

[31] K.A. Dill, Dominant forces in protein folding, Biochemistry 24 (1985) 1501.

[32] W. Ding, J. Xie, D. Dai, H. Zhang, H. Xie, W. Zhang, CNNcon: improved protein contact maps prediction using cascaded neural networks, PLOS ONE 8 (4) (2013) 1–7.

[33] C. Dodge, R. Schneider, C. Sander, The HSSP database of protein structure–sequence alignments and family profiles, Nucleic Acids Res. 26 (1) (1998) 313–315.

[34] M. Dorn, N. Souza, A3N: an artificial neural network n-gram-based method to approximate 3-D polypeptides structure prediction, Expert Syst. Appl. 37 (2010) 7497–7508.

[35] J.M. Duarte, R. Sathyapriya, H. Stehr, I. Filippis, M. Lappe, Optimal contact definition for reconstruction of contact maps, BMC Bioinform. 11 (2010) 283.

[36] J. Eickholt, Z. Wang, J. Cheng, A conformation ensemble approach to protein residue–residue contact, BMC Struct. Biol. 11 (2011) 38.

[37] J. Eickholt, J. Cheng, Predicting protein residue–residue contacts using deep networks and boosting, Bioinformatics 28 (2012) 3066–3072.

[38] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models, Phys. Rev. E 87 (2013) 012707.

[39] P. Fariselli, R. Casadio, A neural network based predictor of residue contacts in proteins, Protein Eng. 12 (1999) 15–21.

[40] P. Fariselli, O. Olmea, A. Valencia, R. Casadio, Prediction of contact map with neural networks and correlated mutations, Protein Eng. 14 (2001) 133–154.

[41] X. Gao, D. Bu, J. Xu, M. Li, Improving consensus contact prediction via server correlation reduction, BMC Struct. Biol. 9 (2009) 28.

[43] J. Gorodkin, O. Lund, C.A. Andersen, S. Brunak, Using sequence motifs for enhanced neural network prediction of protein distance constraints, ISMB 99 (1999) 95–105.

[44] N. Gupta, N. Mangal, S. Biswas, Evolution and similarity evaluation of protein structures in contact map space, Proteins 59 (2005) 196–204.

[45] D. Hall, S. Li, K. Yamashita, R. Azuma, J.A. Carver, D.M. Standley, A novel protein distance matrix based on the minimum arc-length between two amino-acid residues on the surface of a globular protein, Biophys. Chem. (2014), http://dx.doi.org/10.1016/j.bpc.2014.01.005

[46] S. Han, B. Lee, S.T. Yu, C. Jeong, S. Lee, D. Kim, Fold recognition by combining profile–profile alignment and support vector machine, Bioinformatics 21 (2005) 2667–2673.

[47] K. Islam, M. Chetty, Novel memetic algorithm for protein structure prediction, Lecture Notes Artif. Intell. 5866 (2009) 412–421.

[48] B. Jayaran, P. Dhingra, B. Lakhani, S. Shekhard, Bhageerath – targeting the near impossible: pushing the frontiers of atomic models for protein tertiary structure prediction, J. Chem. Sci. 124 (1) (2012) 83–91.

[49] D.T. Jones, D.W.A. Buchan, D. Cozzetto, M. Pontil, PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments, Bioinformatics 28 (2) (2012) 184–190.

[50] D.T. Jones, T. Singh, T. Kosciolek, S. Tetchner, MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range

hydrogen bonding in proteins, Bioinformatics (2014), http://dx.doi.org/10.1093/bioinformatics/btu791

[51] M.V. Judy, K.S. Ravichandran, K. Murugesan, A multi-objective evolutionary algorithm for protein structure prediction with immune operators, Comput. Methods Biomech. Biomed. Eng. 12 (4) (2009) 407–413.

[52] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.

[53] K. Karplus, SAM-T08: HMM-based protein structure prediction, Nucleic Acids Res. 37 (2) (2009) 492–497.

[54] C. Kehyayan, N. Mansour, Evolutionary algorithm for protein structure prediction, in: International Conference on Advanced Computer Theory and Engineering, vol. 199, 2008, pp. 133–154.

[55] J. Ko, H. Park, L. Heo, C. Seok, GalaxyWEB server for protein structure prediction and refinement, Nucleic Acids Res. 40 (2012) 294–297.

[56] T. Kohonen, K. Makisara, The self-organizing feature maps, Phys. Scr. 39 (1989) 168–172.

[57] T. Kosciolek, D.T. Jones, De novo structure prediction of globular proteins aided by sequence variation-derived contacts, PLOS ONE 9 (3) (2014) e92197.

[58] J. Kosinski, I.A. Cymerman, M. Feder, M.A. Kurowski, J.M. Sasin, J.M. Bujnicki, A Frankensteins monster approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation, Proteins 53 (Suppl 6) (2003) 369–379.

[60] N. Krasnogor, B.P. Blackbourne, E.K. Burke, J.D. Hirst, Multimeme algorithms for protein structure prediction, Lecture Notes Comput. Sci. 2439 (2002) 769–778.

[61] P. Kukic, C. Mirabello, G. Tradigo, I. Walsh, P. Veltri, G. Pollastri, Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks, BMC Bioinform. 15 (2014) 6.

[62] C. Lavor, L. Liberti, N. Maculan, A. Mucherino, Recent advances on the discretizable molecular distance geometry problem, Eur. J. Oper. Res. (2011).

[63] Y. Li, Y. Fang, J. Fang, Predicting residue–residue contacts using random forest models, Bioinformatics 27 (24) (2011) 3379–3384.

[64] F. Liang, W.H. Wonh, Evolutionary Monte Carlo for protein folding simulations, J. Chem. Phys. 115 (7) (2001) 3374–3380.

[65] M. Lippi, P. Frasconi, Prediction of protein beta-residue contacts by Markov logic networks with grounding-specific weights, Bioinformatics 25 (18) (2009) 2326–2333.

[68] A. Lo, Y.Y. Chiu, E.A. Rødland, P.C. Lyu, T.Y. Sung, W.L. Hsu, Predicting helix–helix interactions from residue contacts in membrane proteins, Bioinformatics 25 (8) (2009) 996–1003.

[69] R.M. MacCallum, Striped sheets and protein contact prediction, Bioinformatics 20 (2004) 224–231.

[70] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE 6 (12) (2011) 766.

[71] K.J. Maurice, SSThread: template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs, J. Comput. Chem. 35 (2014) 644–656.

[72] A.E. Márquez-Chamorro, G. Asencio-Cortés, F. Divina, J.S. Aguilar-Ruiz, Evolutionary decision rules for predicting protein contact maps, in: Pattern Analysis and Applications, PAAA, Springer, 2012, September.

[74] S. Miyazawa, Prediction of contact residue pairs based on co-substitution between sites in protein structures, PLOS ONE 8 (1) (2013) e54252.

[75] C. Mooney, G. Pollastri, Beyond the twilight zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information, Proteins 77 (1) (2009) 181–190.

[76] B. Monastyrskyy, K. Fidelis, A. Tramontano, A. Kryshtafovych, Evaluation of residue–residue contact predictions in CASP9, Proteins 79 (S10) (2011) 119–125.

[77] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) E1293–E1301.

[78] J. Moult, A large-scale experiment to asses protein structure prediction methods, Proteins 23 (3) (1995) 2–4.

[80] T. Nugent, D.T. Jones, Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) E1540–E1547.

[81] O. Olmea, B. Rost, A. Valencia, Effective use of sequence correlation and conservation in fold recognition, J. Mol. Biol. 295 (1999) 1221–1239.

[82] J.T. Pedersen, J. Moult, Protein folding simulations with genetic algorithms and a detailed molecular description, J. Mol. Biol. 269 (1997) 240–259.

[83] D. Pelta, N. Krasnogor, Multimeme algorithms using fuzzy logic based memes for protein structure prediction, Stud. Fuzziness Soft Comput. 166 (2005) 49–64.

[84] J. Peng, J. Xu, RaptorX: exploiting structure information for protein alignment by statistical inference, Proteins 79 (10) (2011) 161–171.

[86] J. Pevsner, Bioinformatics and Functional Genomics, Wiley-Blackwell, 2005.

[87] A. Piccolboni, G. Mauri, Application of evolutionary algorithms to protein folding prediction, Lecture Notes Comput. Sci. 1363 (1998) 123–135.

[88] M. Punta, B. Rost, PROFcon: novel prediction of long-range contacts, Bioinformatics 21 (2005) 2960–2968.

[89] R. Rajgaria, S.R. McAllister, C.A. Floudas, Towards accurate residue–residue hydrophobic contact prediction for alpha helical proteins via integer linear optimization, Proteins 74 (4) (2009) 929–947.

[90] R. Rajgaria, Y. Wei, C.A. Floudas, Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method astro-fold, Proteins 78 (8) (2010) 1825–1846.

[91] C. Ramakrishnan, G.N. Ramachandran, Stereochemical criteria for polypeptide and protein chain conformation, Biophys. J. 5 (1965) 909–933.

[92] A. Raval, Z. Ghahramani, D.L. Wild, Bayesian network model for protein fold and remote homologue recognition, Bioinformatics 8 (2002) 788–801.

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

[94] O. Sander, I. Sommer, T. Lengauer, Local protein structure prediction using discriminative models, BMC Bioinform. 7 (2006) 14.

[95] C. Savojardo, P. Fariselli, M. Alhamdoosh, P.L. Martelli, A. Pierleoni, R. Casadio, Improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization, Bioinformatics 27 (2011) 2224–2230.

[96] C. Savojardo, P. Fariselli, P.L. Martelli, R. Casadio, Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations, BMC Bioinform. 14 (2013) S10.

[97] S. Schulze-Kremer, Genetic algorithms and protein folding, Protein Struct. Predict. 9 (2000) 175–222.

[98] G. Shackelford, K. Karplus, Contact prediction using mutual information and neural nets, Proteins 69 (2007) 159–164.

[99] M.S. Shell, S.B. Ozkan, V. Voelz, G.A. Wu, K.A. Dill, Blind test of physics-based prediction of protein structures, Biophys. J. 96 (2009) 917–924.

[100] S.Y.M. Shi, N. Suganthan, Multi-Class Protein Fold Recognition Using Multi-Objective Evolutionary Algorithms. KanGAL Report, 2004, pp. 1–7.

[101] L. Stein, Genome annotation: from sequence to biology, Nat. Rev. Genet. 2 (7) (2001) 493–503.

[102] J.I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, J.N. Onuchic, Genomics-aided structure prediction, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 10340–10345.

[103] A.N. Tegge, Z. Wang, J. Eickholt, J. Cheng, NNcon: improved protein contact map prediction using 2D-recursive neural networks, Nucleic Acids Res. 37 (2) (2009) 515–518.

[104] M. Tress, Target domain definition and classification in CASP8, Proteins 77 (Suppl 9) (2009) 10–17.

[105] R. Unger, The genetic algorithm approach to protein structure prediction, Struct. Bond. 110 (2004) 153–175.

[106] A. Vullo, I. Walsh, G. Pollastri, A two-stage approach for improved prediction of residue contact maps, BMC Bioinform. 7 (180) (2006) 1–12.

[107] Z. Wang, J. Eickholt, J. Cheng, Multicom: a multi-level combination approach to protein structure prediction and its assessments in CASP8, Bioinformatics 26 (7) (2010) 882–888.

[108] X.F. Wang, Z. Chen, C. Wang, R.X. Yan, Z. Zhang, J. Song, Predicting residue–residue contacts and helix–helix interactions in transmembrane proteins using an integrative feature-based random forest approach, PLoS ONE 6 (2011) e2676.

[109] Z. Wang, J. Xu, Predicting protein contact map using evolutionary and physical constraints by integer programming, Bioinformatics 29 (2013) 266–273.

[110] Y. Wei, C.A. Floudas, Enhanced inter-helical residue contact prediction in transmembrane proteins, Chem. Eng. Sci. 66 (19) (2011) 4356–4369.

[111] K. Wolff, M. Vendruscolo, M. Porto, Stochastic reconstruction of protein structures from effective connectivity profiles, BMC Biophys. 1 (2008) 5.

[112] K. Wolff, M. Vendruscolo, M. Porto, Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles, Proteins 78 (2010) 249–258.

[113] S. Wu, A. Szilagyi, Y. Zhang, Improving protein structure prediction using multiple sequence-based contact predictions, Structure 19 (8) (2011) 1182–1191.

[114] S. Wu, Y. Zhang, A comprehensive assessment of sequence-based and template-based methods for protein contact prediction, Bioinformatics 24 (7) (2008) 924–931.

[115] D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, Proteins 80 (2012) 1715–1735.

[116] B. Xue, E. Faraggi, Y. Zhou, Predicting residue–residue contact maps by a two-layer: integrated neural-network method, Proteins 76 (1) (2009) 176–183.

[117] J.Y. Yang, X. Chen, A consensus approach to predicting protein contact map via logistic regression, in: Bioinformatics Research and Applications – 7th International Symposium, ISBRA 2011, Changsha, China, May 27–29, 2011, 6674: 136–147.

[118] Y. Yang, E. Faraggi, H. Zhao, Y. Zhou, 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) (2011) 2076–2082.

[119] A. Zemla, LGA: a method for finding 3D similarities in protein structures, Nucleic Acids Res. 31 (13) (2003) 3370–3374.

[121] G. Zhang, D. Huang, Z. Quan, Combining a binary input encoding scheme with RBFNN for globulin protein inter-residue contact map prediction, Pattern Recogn. Lett. 16 (10) (2005) 1543–1553.

[122] G. Zhang, K. Han, Hepatitis C virus contact map prediction based on binary strategy, Comput. Biol. Chem. 31 (2007) 233–238.

[124] G.Z. Zhang, D.S. Huang, Combing genetic algorithm with neural network technique for protein inter-residue spatial distance prediction, Neural Netw. 3 (2004) 1687–1691.

[125] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, Proteins 57 (2004) 702–710.

[126] Y. Zhang, I-TASSER: fully automated protein structure prediction in CASP8, Proteins 77 (2009) 100–113.

[127] X. Zhang, T. Wang, H. Luo, J. Yang, Y. Deng, J. Tang, M. Yang, 3D Protein structure prediction with genetic tabu search algorithm, BMC Syst. Biol. 4 (2010) S6.

[128] Y. Zhao, G. Karypis, Prediction of Contact Maps Using Support Vector Machines, Department of Computer Science, University of Minnesota, Minneapolis, 2002.

[129] J. Zhou, D. Arndt, D.S. Wishart, G. Lin, Y. Shi, J. Zhou, D. Arndt, D.S. Wishart, G. Lin, Protein contact order prediction from primary sequences, BMC Bioinform. 9 (255) (2008) 1–21.

[130] J. Zhou, W. Yan, G. Hu, B. Shen, SVR CAF: an integrated score function for detecting native protein structures among decoys, Proteins 82 (4) (2013) 556–564.