Structural protein descriptors in 1-dimension and their sequence-based predictions

Lukasz Kurgan[1*] and Fatemeh Miri Disfani[1]

[1] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, CANADA

*Corresponding author
email: lkurgan@ece.ualberta.ca; phone (780) 492-5488; fax (780) 492-1811

Abstract
The last few decades observed an increasing interest in development and application of 1-dimensional (1D) descriptors of protein structure. These descriptors project 3D structural features onto 1D strings of residue-wise structural assignments. They cover a wide-range of structural aspects including conformation of the backbone, burying depth/solvent exposure and flexibility of residues, and inter-chain residue-residue contacts. We perform first-of-its-kind comprehensive comparative review of the existing 1D structural descriptors. We define, review and categorize ten structural descriptors and we also describe, summarize and contrast over eighty computational models that are used to predict these descriptors from the protein sequences. We show that the majority of the recent sequence-based predictors utilize machine learning models, with the most popular being neural networks, support vector machines, hidden Markov models, and support vector and linear regressions. These methods provide high-throughput predictions and most of them are accessible to a non-expert user via web servers and/or stand-alone software packages. We empirically evaluate several recent sequence-based predictors of secondary structure, disorder, and solvent accessibility descriptors using a benchmark set based on CASP8 targets. Our analysis shows that the secondary structure can be predicted with over 80% accuracy and segment overlap (SOV), disorder with over 0.9 AUC, 0.6 Matthews Correlation Coefficient (MCC), and 75% SOV, and relative solvent accessibility with PCC of 0.7 and MCC of 0.6 (0.86 when homology is used). We demonstrate that the secondary structure predicted from sequence without the use of homology modeling is as good as the structure extracted from the 3D folds predicted by top-performing template-based methods.

Introduction

Knowledge of protein structure and lack of structure (disorder in the structure) has important implications in characterization, prediction and analysis of protein function and interactions with other molecules. The last few decades observed development of a number of lower-level descriptors of protein structure that provide an alternative and somehow complementary way to describe, analyze, and predict protein structure and function when compared with the structure defined as a set of three-dimensional atomic coordinates. These descriptors quantify certain structural properties of residues, such as packing density, local (with respect to the sequence) structural arrangements, and their position with respect to the protein surface, and they provide interesting insights into sequence-structure-function relationships. Importantly, they already found a wide-range of useful applications, including tertiary structure prediction [1-3] assessment of the quality of structural models [4], sequence to structure alignment [5], characterization of folding dynamics[6], prediction and characterization of binding residues [7] and target selection for structural genomics [8, 9], to name just a few.

We refer to these descriptors as 1-dimensional (1D) descriptors since they project 3D structural features onto 1D strings of residue-wise structural assignments. Examples of 1D structural descriptors include secondary structure, solvent accessibility and annotation of transmembrane helices [10]. Besides the structural descriptors, recent years also observed research into development, analysis and

prediction of 1D functional descriptors. Examples include annotation of residues that participate in protein-protein interactions [11], DNA and RNA binding [12-14], and annotation of catalytic residues [15, 16] and signal peptides [17].

To date, the majority of the structural descriptors were proposed, analyzed and reviewed individually. While some of them, such as secondary structure, are relatively well popularized, others, like residue depth, are less well-known, although we believe that the insights they provide and the applications that they enable merit their introduction to a wider community. Moreover, a number of these descriptors are related with each other and they share certain characteristics that can be exploited for instance in building more effective models for their prediction. We perform first-of-its-kind comprehensive comparative review of existing 1D structural descriptors. We define, review and categorize ten structural descriptors and we also describe, summarize and contrast computational models that are used to predict these descriptors from the protein sequences. Our study extends prior review works that concerned individual descriptors, including secondary structure [10, 18-20], disorder [21-23], solvent accessibility and transmembrane helices [10, 19], as well as a recent review that focuses on the relevant prediction servers [24].

1D structural descriptors

Categorization of descriptors

The 1D descriptors of protein structure can be grouped based on the type of information they provide and the way they encode it. We categorize these descriptors into four following classes that describe similar structural properties:

1. *Backbone conformation* descriptors, which describe relative spatial position of residues along the protein chain. We consider three of these descriptors including torsion angles, secondary structure, and annotation of transmembrane helices. Other examples of the 1D backbone conformation descriptors include annotation of certain secondary and supersecondary conformations such as beta-, gamma-, alpha- and pi-turns [25-29], beta barrels [30, 31], and coiled coils [32, 33]. Information concerning the backbone conformation could be aggregated in a local neighborhood and several studies were devoted to the development and prediction of frequent local backbone structures/motifs which are referred to as structural alphabets [34].
2. Descriptors of *buriedness* that quantify the degree of the exposure of residues to the solvent when they are located on the protein surface and their burying depth when they are positioned in the core of the native fold. We review two of these descriptors, solvent accessibility and residue depth. We note that buriedness can be also described using a recently proposed half sphere exposure descriptor [35, 36].
3. *Inter-chain contacts* descriptors, which represent information concerning connectivity/density of residues in the native fold; they are usually quantified as the number of contacts between residues that are close by in the structure. We introduce three contact-based descriptors including contact number, absolute contact number, and residue-wise contact order. These 1D descriptors are related to the 2D contact map descriptor [37], which is out of scope of this review.
4. Descriptors of *flexibility* that characterize the degree to which the spatial position of a given residue fluctuates. The extreme manifestation of these fluctuations is a lack of stable structure, which is referred to as disorder. These descriptors of the lack/flexibility of protein structure are motivated by the fact that flexible/disordered residues are implicated in various cellular processes such as regulation, recognition, signaling and control [38, 39]. We overview two flexibility descriptors including B-factors and the annotations of the disordered residues.

The possible encodings of the above 1D descriptors are based on nominal, real (floating point), binary, and integer values. The binary encoding denotes presence/absence of a certain structural property for a

given residue in the protein chain, e.g., annotation of residues that form or do not form trans-membrane helices. Nominal values explicitly denote specific states of a given property in case when multiple states are possible, e.g., annotation of the secondary structure states, while usage of integer values additionally signifies a particular ordering of these states, e.g., contact number. Real-numbers are used when the property follows a continuous scale which is not discretized into states, e.g., solvent accessibility, B-factors and torsion angles.

Definitions

Following we review the definitions of the considered descriptors. Although some of the 1D structural descriptors have a unique definition, a number of them have a few, usually slightly different and possibly complementary, definitions.

Descriptors of the backbone conformation

Secondary structure refers to local three-dimensional conformations of amino acid segments in the protein chain which are established through hydrogen bonds between N-H and C=O groups. A number of different systems for the assignment of secondary structure were developed over the last few decades. The first implementation of the secondary structure assignment method was done in late 1970s by Levitt and Greer [40]. This was followed by Kabsch and Sander who developed a method called DSSP [41]. This algorithm is based on the detection of hydrogen bonds defined by an electrostatic criterion. Other assignment methods include DEFINE [42], P-CURVE [43], STRIDE [44], PSEA [45], XTLSSTR [46], SECSTR [47], and KAKSI [48].

Here, we briefly introduce the assignment defined by DSSP since this method is often assumed to be the golden standard and it remains to be the most widely-used program for the secondary structure assignment [48]. Overall, the secondary structures are determined based on the patterns of hydrogen bonds and they are categorized into three major states, helices, sheets, and regions with irregular secondary structure. The DSSP method assigns one of the following eight secondary structure types for each of the structured residues (residues that have three-dimensional coordinates) in the protein sequence:

- G: 3-turn helix (also referred to as 3_{10} helix). In this secondary structure the carboxyl group of a given amino acid forms a hydrogen bond with amid group of the amino acid three positions down in the sequence forming a tight, right-handed helical structure with only three residues per turn.
- H: 4-turn helix (also refereed to as α-helix). This structure is similar to the 3-turn helix, however, the hydrogen bonds are formed between consecutive amino acids that are four positions away in the protein chain. This is the most prevalent helix type.
- I: 5-turn helix (also refereed to as π-helix). In this type of the helix the hydrogen bonding occurs between residues spaced five positions away from each other and which also results in a right-handed helical structure; left-handed π-helices are relatively rare.
- E: extended strand in parallel or anti-parallel sheet conformation. Two or more strands are connected laterally by at least two hydrogen bonds forming a pleated sheet.
- B: residue in an isolated beta-bridge, which is a single residue pair sheet formed based on the hydrogen bond.
- T: hydrogen bonded turn. A turn in the protein chain in which a single hydrogen bond is formed between residues spaced 3, 4, or 5 positions away in the protein chain.
- S: bend, which denotes a fragment of protein chain with high curvature where the angle between the vector from C^{α}_{i} to C^{α}_{i+2} (C^{α} atoms at i^{th} and $i+2^{th}$ positions) and the vector from C^{α}_{i-2} to C^{α}_{i} is$< 70°$; this is the only non-hydrogen bond-based regular secondary structure type.
- – : irregular secondary structure (also referred to as loops and random coils), which corresponds to the remaining conformations.

The above eight types are often mapped into three states as follows

– H: α-helix. This secondary structure state encompasses right or left handed cylindrical/helical conformations that include G, H, and I types.
– E: β-strand. This state corresponds to pleated sheet structures and it includes E and B secondary structure types.
– C: coil. This state represent the remaining types of the local confirmations and it includes S, T, and – types.

Transmembrane helices (TMHs) are helices that are embedded into the lipid bilayer of membranes; they are characteristic to α-helical transmembrane proteins. These proteins constitute about 30% of the proteins encoded in a typical genome and are involved in a wide variety of important processes such as cell signaling, transport of membrane-impermeable molecules and cell recognition [49]. TMHs are typically apolar 12 to 35 residues long helical amino acid segments that are oriented perpendicularly to the surface of the membrane. Transmembrane proteins include several TMHs which are usually approximately parallel to each other and which are packed close to each other in the membrane. TMPDB (Transmembrane Protein Database) [50] provides convenient access to annotation of transmembrane helices.

Torsion angles are the rotational angles that define placement of the backbone atoms in the protein chain. The three rotational angles include ω which is defined about the C–N bond, φ about the C^{α}–N bond and ψ about the C^{α}–C bond. The value of ω is fixed at 180° or 0° and thus the protein backbone is described by the remaining two torsion angles. Different secondary structure states have their characteristic torsion angles that can be visualized using the Ramachandran plot [51]. An improved view of the sparseness of the allowed torsion angles, in particular for multiple consecutive angles, can be obtained using representation described in [52].

Descriptors of the buriedness

Solvent-accessible area of a protein molecule was first defined by Lee and Richards in early 1970s [53] as the area traced out by the center of a virtual probe sphere representing a solvent molecule as it is rolled over the protein surface. In the follow up definition [54], the solvent-accessible area consists of the part of the van der Waals surface of the atoms that are accessible to the probe sphere. The accessible surfaces of atoms are connected to each other by a network of concave and saddle-shaped surfaces that smoothes over the crevices and pits between the atoms. The 1D descriptor of the *solvent accessibility* (also referred to as the *relative solvent accessibility*) is defined as the ratio between the solvent exposed surface area of a given residue observed in a given protein structure (i.e., the corresponding part of the solvent-accessible area of this protein) and the maximum obtainable value of the solvent-exposed surface area for this amino acid [55]. The ratio is used to normalize between different residue types. The values for the accessible surface area are often calculated using the DSSP program [41]. The maximum obtainable values of the solvent exposed surface area correspond to the surface exposed area of a given residue type observed in an extended tripeptide conformation flanked with either glycine [56] or alanine [57] residues. The relative solvent accessibility ranges between 0%, for fully buried residues, and 100%, for fully solvent accessible residues. All residues with 0% relatively solvent accessibility are categorized as fully buried, although their burying depth, with respect to the core of the protein molecule, can be different. This observation motivates the residue depth descriptor.

The *residue depth* is the average atom depth of all atoms, except the hydrogen atoms, that make up a given residue [58]. Several definitions of atom depth have been proposed, including distance [58-60] and volume-based [61]. The depth of an atom could be defined as the distance of this atom from the nearest surface water molecule; the corresponding calculations use Monte Carlo simulations of water

molecules surrounding the protein [58]. The DPX algorithm [60] defines the depth as the distance of a given atom from the closest solvent accessible atom. The DPX-based depth equals zero for solvent accessible atoms and is greater that zero for atoms buried in the protein interior. The volume-based atom depth is defined as

$$D_{ir} = 2V_{ir}/V_{0r}$$

where i is the atom index, V_{ir} is the solvent exposed volume of a sphere with radius r centered on atom i, and V_{0r} is the volume of the same sphere centered on an isolated atom. In contrast to the above definitions that compute an average over all atoms, following Verazzo et al. [61] the residue depth is defined using only the C^{α} atoms and r = 9Å. We note that distance- and volume-based depth values are negatively correlated.

The depth values are usually normalized using mean and standard deviation of the depth values in a large, pre-defined protein dataset as follows [62, 63]

normalized_depth = (depth – mean_depth) / standard_deviation_of_depth

Details concerning calculation of various residue depth definitions are given in [63].

Descriptors of the inter-chain contacts

The *contact number* (also referred to as the coordination number or the Ooi number) is the number of residues making a "contact" with a given residue in a native protein fold. More specifically, this descriptor is defined as the number of C^{α} atoms within the sphere of a predefined radius r centered on the C^{α} atom of a given residue. A few variants of this definition were proposed over the last 30 years:

- The contact number is defined as the number of C^{α} atoms, excluding the two adjacent residues in the sequence, within a sphere with r = 8Å centered at the C^{α} atom of a given residue [64].
- The same as above but with a larger r = 14Å [65]
- Recently, Pollastri et al. proposed coordination number which is a binarized contact number [66]. A given residue is assigned value of 0 if the contact number computed using the sphere with radius r is lower than the average contact number for a given amino acid type; otherwise the residue is assigned 1. The authors used different values of r = 6, 8, 10, and 12Å and they counted all C^{α} atoms in the sphere, including the two residues adjacent in the sequence.

The *absolute contact number*, which is a variant of the contact number, uses C^{β} atoms, except for glycine where C^{α} atoms are used, and the boundary of the sphere is smoothed using a sigmoid function [67]. The absolute contact number of the i^{th} residue in a protein chain is defined as

$$\sum_{i:|j-i|>2} \sigma(r_{ij})$$

where r_{ij} is the distance between C^{β} atoms of i^{th} and j^{th} residue, the two residues adjacent to the i^{th} position are disregarded in the sum, $\sigma(r_{ij}) = 1/(1+e^{w[r_{ij}-d_c]})$, w determines the smoothness of the boundary (by default w = 3), and d_c = 12 is a free parameter which is used as a cutoff to find contacting residues. The sigmoid function $\sigma(r_{ij})$ is approximately equal 1 when $r_{ij} < d_c$ and it is close to 0 when $r_{ij} > d_c$. The absolute contact number is a continuous (floating-point) extension of the discrete (integer) contact number.

The *residue-wise contact order* of the i^{th} residue, which is expressed as the sum of linear distances in the protein sequence between all pairs of contacting residues, excluding the two neighbors on each side of the i^{th} residue, is defined as [68]:

$$\sum_{j=1,|j-i|<3}^{N} S_{ij}$$

Two residues are assumed to be in contact if the distance between their C^{β} atoms (C^{α} atoms for glycine) is < 12 Å.

A related sequence level descriptor called *relative contact order*, which possibly motivated the development of the abovementioned residue level contact descriptors, is defined as the average linear distance in the protein sequence between all pairs of contacting residues normalized by the sequence length [69].

Descriptors of flexibility

The B-factor (also called temperature-factor or Debye-Waller factor) describes the degree to which the electron density of a given atom (or a group of atoms) in the X-ray scattering of the crystal structure of a protein is spread out. The B-factor values quantify mobility of an atom and they are computed as

$$8\pi^2 U_i^2$$

where U_i^2 is the mean square displacement of the i^{th} atom which is averaged over the lattice. Since B-factors depend on several characteristics of the structure determination protocol, such as experimental resolution, crystal contacts, and refinement procedures, they should be normalized to allow comparisons between different structures. Following [70-73] the B-factors of given residues are expressed using the B-factors of C^{α} atoms that are normalized using average and standard deviation of the B-factors in a given chain as follows

$$\text{normalized\_}B_{factor} = (B_{factor} - \text{mean\_}B_{factor}) \,/\, \text{standard\_deviation\_of\_}B_{factor}$$

While the abovementioned descriptors have relatively well-defined and consistent definitions, the annotations of the *disordered residues* (also referred to as intrinsically disordered, intrinsically unstructured, natively unfolded, natively disordered, and highly flexible) that are performed using different experimental methods is not always consistent [74]. Protein disorder is indirectly observed using a diverse set of experimental methods including spectroscopic and NMR-based approaches [75]. To date, there is no golden standard for the assignment of the disordered regions, i.e., segments of disordered residues in the protein chain. In the past CASP (Critical Assessment of techniques for protein Structure Prediction) experiments the disordered regions were defined as residues that lack coordinates in structures solved by X-ray crystallography and as residues that exhibit high variability within structure ensemble or are annotated as disordered in REMARK 465 by experimentalists for the structures solved by NMR [76, 77]. Another source of high-quality disorder annotation is a manually curated DisProt database (Database of Disordered Proteins) [78], which is the main and centralized source of the experimentally validated disordered regions.

Summary

The considered ten 1D structural descriptors are summarized in Table 1. Some of them were originally proposed over half a century ago, including B-factor which was investigated by Debye and Waller in early 1900s [79] and secondary structure which was first proposed by Pauling and Corey in early 1950s [80, 81] and later formalized in 1970s [40]. Although structural properties described by most of these descriptors characterize individual residues, in a few cases, such as the secondary structure and disorder, these properties can be extended over segments of consecutive residues in the protein chain. For instance, helices are formed by stretches of at least 3 consecutive residues. As mentioned above, the 1D descriptors are encoded using several formats including nominal values and binary, integer, and real numbers. The original encoding is sometimes converted into a reduced representation. This includes the binarized version of the integer-valued contact number [66], and the real-valued relative solvent accessibility which is often converted into a binary descriptor by setting a threshold to differentiate between exposed and buried residues. In this case a given residue is defined as solvent exposed when it's relative solvent accessibility > 0.25, and otherwise it is assumed to be buried [73, 82-85]. The computation of the descriptor values most often includes processing of the tertiary structures that are deposited in the Protein Data Bank (PDB) [86] in some cases other databases such as DisProt [78] and TMPDB [50] are also used.

Sequence-based prediction of 1D structural descriptors

The values of the 1D structural descriptors can be either computed from the known protein structures (or experimental annotations of disorder) or predicted from the knowledge of the input protein sequence. The latter is based on the (mostly true) premise that sequence determines native structure for small globular protein [87]. We overview the existing sequence-based predictors of the 1D structural descriptors and present an empirical study that compares selected secondary structure, disorder, and solvent accessibility predictors. The novel aspects of our analysis include the comparison of the quality of the secondary structure predictions with the quality of the secondary structure in the predicted tertiary structure, inclusion of a wide-range of evaluation measures, and per-segment (explained below) evaluation of the disorder predictions.

Overview of existing predictors

The available methods for the sequence-based prediction of the considered descriptors are summarized in Table 2. We provide information concerning publication that introduces a given predictor and the subsequent publications that describe extensions and improvements to this method, details of the prediction model including its inputs and algorithms used, and we comment on the availability of the web server and/or a standalone program that implement this method. As one of the potential measures of the popularity of these predictors we provide the citation counts, both total and per year since publication, which were collected in July 2010 using the ISI Web of Science database. For methods with more than one publication we provide the average per year across all publications. We include the secondary structure predictors that were developed after 1999 and we refer the reader to other recent reviews [10, 18] to learn about older methods. Similarly, for the transmembrane helix prediction methods we primarily concentrate on methods published after mid 1990s; older methods are reviewed in [88, 89]. We exclude the predictors of disorder since they were recently and comprehensively reviewed in several works including [21, 22, 90, 91].

Most of the existing predictors use the sequence profiles or multiple alignments computed using PSI-BLAST [92] as their inputs. Other popular inputs to predict the 1D descriptors include secondary structure and solvent accessibility that are predicted from the sequence. The selection of the predictive method/algorithm mostly depends on the encoding of a given descriptor. We observe that most of the recent predictors are based on so called machine learning methods. They include neural networks which were used for all considered descriptors, hidden Markov models and support vector machines which were used for nominal and binary descriptors, and support vector regression and multiple linear regression that were applied to predict real-value and integer-based descriptors. We note that many predictors for binary and nominal descriptors output probabilities associated with each prediction and often use these probabilities to predict the descriptor values by setting a threshold. The advantage of predicting probabilities instead of directly predicting the descriptor values is that these probabilities can be used to indicate confidence in the predicted value.

Some predictors utilize homology modeling where they find similar sequences with known structure and use them to perform predictions. There are several methods that are based on a consensus, in which case outputs of multiple predictors of a given descriptor are combined together; some of the consensus-based methods also utilize homology modeling. A number of the secondary structure, transmembrane helix, and solvent accessibility predictors received several hundred citations, which suggests is a relatively substantial interest in research and applications of these methods. Finally, we observe that most of these predictors are provided as either standalone software or web servers (in same cases both) to a wider structural biology community. These software are usually publicly accessible and can be used free of charge for non-commercial purposes. The availability of these programs allows for a relatively easy, without any coding and usually user-friendly access for a non-

expert to the predictions of these 1D descriptors from the protein sequence. This, in turn, is one of the major reasons for the high citations / popularity of these prediction programs.

Table 1. Summary of the considered ten 1D structural descriptors. The first two columns provide the types and names of the descriptors. The third column lists the year (or approximate time in case of the transmembrane helices and disorder) when the descriptor was proposed. The next column identifies descriptors that have multiple definitions. The following set of four columns provides details about descriptors including brief descriptions, information whether they are defined for individual residues or for segments of adjacent residues, data type used to encode their values, and information which atoms are used to define their values. The rightmost column briefly explains how the descriptor values are computed.

Type of descriptor	Name of descriptor	Year (approximate time) first introduced	Multiple definitions	Characteristics of the descriptors				Computation of descriptor values
				Brief description	Defined for residues vs. sequence segments	Data type	Atoms used in the definition	
Backbone conformation	Secondary structure	1951 / 1977 [40, 80, 81]	Yes	Local, in the sequence, patterns of hydrogen bonding; DSSP defines 8 secondary structure types that are combined into 3 states, helix, strand and coil.	Segment	Nominal	Backbone	Computed from PDB structures using DSSP
	Transmembrane helices	1970s		Helices embedded into the lipid bilayer of membranes	Segment	Binary	Backbone	Computed from PDB structures or annotated in TMPDB
	Torsion angles	1963 [51]		Rotational angle about the bonds between atoms in the peptide backbone	Residue	Real	Backbone	Computed from PDB structures
Buriedness	Solvent accessibility	1971 [53]	Yes	Amount of residue surface that is exposed to the solvent.	Residue	Real	All atoms	Computed from PDB structures using DSSP
	Residue depth	1999 [58]	Yes	Distance to the protein surface.	Residue	Real	All atoms or C^{α}	Computed from PDB structures using MSMS, NACCESS, or Amber 4.1
Inter-chain contacts	Contact number	1980 [64]	Yes	Number of contacting residues.	Residue	Integer	C^{α}	Computed from PDB structures
	Absolute contact number	2005 [67]		Smoothed (using continuous function) number of contacting residues	Residue	Real	C^{β}	Computed from PDB structures
	Residue-wise contact order	2005 [68]		Sum of linear distances in the protein sequence between all pairs of contacting residues	Residue	Integer	C^{β}	Computed from PDB structures
Flexibility	Disorder	1970s	Yes	Lack of well-defined tertiary structure	Segment	Binary	Not applicable	Computed from PDB structures or annotated in DisProt
	B-factor	1913 [79]		Mobility of an atom in a crystal structure	Residue	Real	C^{α}	Annotated in PDB structures

Table 2. Summary of the sequence-based predictors of 1D structural descriptors. The "name" column gives the name of the method, if available, and the corresponding reference(s). The "inputs" column gives inputs including sequence (Seq), sequence profile (SeqProf), sequence alignment (SeqAlign), predicted secondary structure (PSS), predicted solvent accessibility (PSA), and homology modeling (HomMod); The "algorithm" column lists the prediction algorithms including Neural Network (NN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Support Vector Regression (SVR), Multiple Linear Regression (MLR), consensus (Con), and homology modeling (HomMod); the latter denotes configurations where templates are merged with the predictions rather than being used as inputs to the predictor. The "citations" columns provide the total and the average per year number of citations for papers listed in the "name" column. We divide the total by the sum of the number of years passed for the papers listed in the "name" column to compute the per year average. The citation counts were extracted from ISI Web of Science database on July 2010. The last column identifies whether a standalone program (SP) and/or a web server (WS) is available. The methods are sorted by the year of their first publication in the ascending order.

Descriptor	Prediction method			Citations		standalone program (SP) / web server (WS)
	Name (references)	Inputs	Algorithm	Total	Per year	
Secondary structure	PHD [93, 94]	SeqProf, Seq	NN	1347	61.2	WS
	PSIPRED [95, 96]	SeqProf, Seq	NN	1728	96	WS + SP
	Jpred [84, 97]	SeqProf, Seq	NN	504	36	WS
	PROF-King [98]	SeqProf, Seq	NN, quadratic/linear discrimination	189	15.7	WS + SP
	SSpro [99, 100]	SeqProf, Seq + HomMod	NN	384	25.6	WS + SP
	GeneSilico [101]	PSS	Con	134	16.7	WS
	SAM-T [102-104]	SeqProf, Seq	NN	173	10.8	WS
	SABLE [105]	SeqProf, Seq, PSA	NN	54	9	WS + SP
	YASPIN [106]	SeqProf, Seq	NN, HMM	63	10.5	WS
	PORTER [107]	SeqProf, Seq	NN	93	15.5	WS + SP
	PSIMLR [108]	SeqProf, Seq	Multiple linear regression	9	1.5	WS
	YASSPP [109]	SeqProf, Seq	SVM	16	3.2	WS
	OSSHMM [110]	SeqProf, Seq	HMM	7	1.4	SP
	PROTEUS [111, 112]	PSS	Con, HMM + HomMod	33	4.1	WS
	PORTER_H [113]	SeqProf, Seq + HomMod	NN	14	3.5	WS + SP
	SPINE [114]	SeqProf, Seq	NN	19	4.7	WS + SP
	P.S.HMM [115]	SeqAlign, Seq	Genetic algorithm, NN, HMM	4	1	WS
	DISSPred [116]	SeqProfs, Seq, predicted torsion angles	SVM, clustering	0	0	WS
	SPINEX [3]	SeqProf, Seq, PSA, PSS	NN	2	1	WS
	PCI-SS [117]	SeqProf, Seq	Parallel cascade identification	1	0.5	WS
	[118]	SeqProf, Seq	Fragment matching	0	0	SP
Trans-membrane helix	TopPred [119, 120]	Seq	Rule-based	1662	46.1	WS
	MEMSAT [121-123]	SeqProf, Seq	NN	680	20	WS
	PHDhtm [124]	SeqProf	NN	418	27.8	WS
	HMMTOP [125, 126]	SeqAlign, Seq	HMM	1150	50	WS + SP
	TMHMM [127, 128]	Seq	HMM	2812	122.2	WS
	GeneSilico [101]	Predicted transmembrane helices	Con	134	16.7	WS
	Phobius [129, 130]	Seq	HMM	356	32.3	WS
	SVMtm [131]	Seq	SVM	28	4	WS
	MINNOU [132]	PSA, PSS, Seq	NN	18	3.6	WS + SP
	TMpro [133, 134]	Seq	NN, active learning	4	1	WS
	MEMSAT-SVM [135]	SeqProf, Seq	SVM	2	1	WS + SP
	TOPTMH [136]	SeqProf, Seq	SVM, HMM	Unavailable		Unavailable
Torsion angles	PrISM [137]	PSS, SeqProf, Seq	NN, SVM	28	4	Unavailable
	LOCUSTRA [34]	SeqProf, Seq	SVM	3	1	WS
	REAL SPINE [3, 82, 83, 138]	SeqProfs, Seq, PSS	NN	31	3.4	WS + SP
	ANGLOR [2]	SeqProf, Seq, PSA, PSS	NN, SVM	2	.6	WS + SP
	DISSPred [116]	SeqProfs, Seq, PSS	SVM, clustering	0	0	WS
	SPINE XI [82]	SeqProf, Seq, PSA, PSS	NN, conditional random field	2	1	WS

Descriptor	Prediction method			Citations		standalone program (SP) / web server (WS)
	Name (references)	Inputs	Algorithm	Total	Per year	
Solvent accessibility	[139]	Seq	NN	71	3.3	Unavailable
	PHDacc [85]	SeqAlign, Seq	NN	300	17.6	WS
	[140]	SeqAlign, Seq	Bayesian statistics	53	3.5	Unavailable
	[141]	SeqAlign, Seq	Probabilistic scoring function	20	1.5	SP
	[142]	Seq	Lookup table	23	1.9	Unavailable
	[143]	Seq	Lookup table	25	2.2	Unavailable
	JNET [84]	SeqProf, Seq	NN	372	33.8	WS + SP
	NETASA [57, 144]	Seq	NN	117	6.8	WS + SP
	ACCpro [66]	SeqProf, Seq	NN	99	11	WS + SP
	SABLE [55]	SeqProf, Seq	NN	84	12	WS + SP
	SVMpsi [145]	SeqProf, Seq	SVM	48	6.8	Unavailable
	[146]	Seq	SVR	35	5	Unavailable
	SARpred [147]	SeqProf, Seq, PSS	NN	31	5.1	WS
	PSIMLR [108]	SeqProf, Seq	MLR	9	1.5	WS
	[148]	SeqProf, Seq	MLR	16	2.6	Unavailable
	[149]	SeqProf, Seq	SVR	26	5.2	WS
	Real-SPINE [82, 83]	SeqProfs, Seq, PSS	NN	24	4	WS + SP
	PaleAle and PaleAle_H [113]	SeqProf, Seq + HomMod	NN	14	3.5	WS + SP
	SVM-Cabins [150]	SeqProf, Seq	SVM	13	3.2	Unavailable
	[151]	SeqProfs, Seq, PSS	SVR, MLR	Unavailable		Unavailable
	NetSurfP [152]	SeqProf, Seq, PSS	NN	3	1.5	WS
Residue depth	[62]	SeqProf, Seq	SVR	5	1.6	Unavailable
	[63]	SeqProfs, Seq, PSS	SVR	4	1.3	Unavailable
	[153]	SeqProfs, Seq, PSS, PSA, predicted disordered regions	SVR	1	0.5	WS
Contact number	[64]	Seq	MLR	66	2.1	Unavailable
	[154]	SeqProf, Seq	NN	10	1	WS
	[155]	SeqProf, Seq	NN	28	2.8	WS
	CONpro [66]	SeqProf, Seq	NN	99	11	WS + SP
	[156]	SeqProf, Seq	SVR	7	1.4	Unavailable
Absolute contact number	[67]	SeqProf, Seq	MLR	51	8.5	Unavailable
	[156]	SeqProf, Seq	SVR	24	4	Unavailable
	CRNPRED [157]	SeqProf, Seq	Critical random network	7	0.6	WS + SP
Residue-wise contact order	CRNPRED [67, 157]	SeqProf, Seq	Critical random network	7	0.6	WS + SP
	[159]	SeqProf, Seq, PSS	SVR	18	3.6	Unavailable
	[160]	SeqProf, Seq, PSS	SVR	0	0	SP
B-factor	[161]	SeqProf, Seq	Logistic regression	90	12.8	Unavailable
	[162]	SeqProf, Seq	SVR	41	6.8	Unavailable
	PROFbval [163, 164]	SeqProf, Seq, PSS, PSA	NN	60	5.4	WS + SP
	[165]	SeqAlign, SeqProf, Seq	SVM	4	1	Unavailable
	[73]	PSA	MLR	3	1.5	Unavailable
	[166]	SeqProf, Seq, PSS	SVR	1	0.5	WS

Table 3.Comparison of the quality of the secondary structure predictions by seven recent sequence-based predictors which were introduced in Table 2 and by using tertiary structure predicted by the top-three performing template-based methods in the CASP8 competition (the secondary structure was acquired from the predicted folds using DSSP). The methods were tested on 111 CASP8 targets. The table lists per-residue measures including Q and MCC values, as well as $Q_{HEerror}$. The last four columns provide per-segment SOV measures where the quality was first computed per sequence and than averaged over all sequences; we provide both the average and the corresponding standard deviation (stdev). Methods for each predictor type (secondary vs. tertiary structure) are sorted in descending order by their Q_3 values.

Predictor		**Per-residue evaluation**											**Per-segment evaluation** (average ±stdev)			
Type	**Name**	**Q_3**	**Q_{Hobs}**	**Q_{Eobs}**	**Q_{Cobs}**	**Q_{Hpred}**	**Q_{Epred}**	**Q_{Cpred}**	**MCC_H**	**MCC_E**	**MCC_C**	**$Q_{HEerror}$**	**SOV_3**	**SOV_H**	**SOV_E**	**SOV_C**
Secondary structure	PORTER_H	83.4	87.4	77.5	83.3	87.4	85.7	78.7	0.8	0.76	0.68	1.28	81.1 ±12.8	82.6 ±12.4	78.7 ±24.5	76.9 ±13.6
	SSpro	80.4	86.3	70.8	80.6	85.2	82.6	75	0.77	0.7	0.63	1.59	78.3 ±12.4	82.3±21.3	74.3 ±23.7	73.6 ±14.2
	SPINE	79.1	84.1	71.4	78.9	83.9	80.1	74.1	0.74	0.69	0.61	2.02	76.8 ±12.5	79.4 ±22.3	74.1 ±24.4	72.3 ±12.8
	YASPIN	78.8	85.9	79.2	71.8	84	73.2	77.2	0.76	0.68	0.59	1.98	77.6 ±11.1	81.6 ±20.9	78.3 ±21.6	70.0 ±12.8
	PSIPRED	78.2	84.7	72.9	75.4	82.6	77.2	74.7	0.74	0.67	0.59	2.23	76.8 ±11.7	79.1±22.7	72.8 ±25.8	71.6 ±13.5
	SABLE	77.7	80.6	67.2	81.3	86	79.8	70.3	0.74	0.66	0.58	1.68	75.6 ±12.1	78.0 ±22.4	72.4 ±24.1	71.4 ±12.9
	PROTEUS	76.6	82	75.3	72.1	83.8	72.3	72.5	0.73	0.65	0.55	1.93	73.6 ±11.0	73.3 ±23.9	72.5 ±25.8	67.3 ±12.5
Tertiary Structure	ZHANG-server	80.5	87.4	68.1	81.6	84.3	90.3	73.1	0.77	0.73	0.61	0.61	77.2 ±14.4	82.5 ±21.7	72.8 ±27.2	71.9 ±14.8
	RAPTOR	80.4	80.7	76.2	82.7	86.4	87.7	72.4	0.74	0.77	0.61	0.57	76.9 ±15.8	77.2 ±27.0	77.9 ±25.4	72.4 ±16.4
	TASSER	76.4	78	58.9	85.6	86.4	89.4	65.8	0.72	0.66	0.56	0.61	71.6 ±17.0	78.4 ±21.0	65.3 ±27.6	68.6 ±17.4

Table 4. Summary of statistical significance tests for Q_3 (shown in the upper triangle) and SOV_3 (shown in the lower triangle) for secondary structure predictions from Table 3. We performed two types of tests, per residue (denoted as "res") and per sequence (denoted as "seq"). In case of the per residue evaluation, the test was done by bootstrapping 1000 repetitions of 80% of the chains; the same datasets were used for all methods and the results on the corresponding 1000 datasets were compared for a given pair of methods. We note that SOV_3 is a per sequence measure and it cannot be calculated per residue. Hence, we averaged the SOV_3 values computed for each sequence in the boostrapped dataset. In the per sequence case, we compared the prediction quality without bootstrapping computed for each of the 111 sequence. The differences between each pair of methods were tested using paired t-test (if samples are normal) or Wilcoxon rank sum test (if they are not normal); normality was validated using Shapiro-Wilk test. For each pair of methods, "+"/"–" mean that a method in the corresponding row performs better/worse, respectively, than the method in the corresponding column with *p*-value < 0.05; "=" denotes that the results are not significantly different. Methods for each predictor type (secondary vs. tertiary structure) are sorted in descending order by their Q_3 values in Table 3.

	Secondary structure														**Tertiary structure**					
	PORTER_H		**SSpro**		**SPINE**		**YASPIN**		**PSIPRED**		**SABLE**		**PROTEUS**		**ZHANG-server**		**RAPTOR**		**TASSER**	
Q_3 / SOV_3	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**	**res**	**seq**
PORTER_H			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
SSpro	–	–			+	=	+	=	+	+	+	+	+	+	–	=	–	=	+	+
SPINE	–	–	–	=			+	=	+	+	+	+	+	+	–	=	–	=	+	+
YASPIN	–	–	–	=	+	=			+	=	+	+	+	+	–	=	–	=	+	+
PSIPRED	–	–	–	–	–	=	–	=			+	=	+	+	–	–	–	=	+	=
SABLE	–	–	–	–	–	–	–	–	–	=			+	+	–	–	–	–	+	=
PROTEUS	–	–	–	–	–	–	–	–	–	–	–	–			–	–	–	–	+	=
ZHANG-server	–	–	–	–	+	=	–	=	+	=	+	=	+	+			+	=	+	+
RAPTOR	–	–	–	–	+	=	–	=	+	=	+	=	+	+	–	=			+	+
TASSER	–	–	–	–	–	–	–	–	–	–	–	–	–	=	–	–	–	–		

Table 5.Comparison of the quality of the disorder predictions by eight recent sequence-based predictors that provide publicly available web server/standalone program and the four best-performing disorder predictors in the CASP8 competition (which are identified by the corresponding group number in the curly brackets). The methods were tested on 121 CASP8 targets. The table lists per-residue measures including AUC (which evaluates real-value predictions of probability of disorder), S_W, MCC, accuracy, ACC (average of sensitivity and specificity), and Q values. The last five columns provide per-segment evaluations including SOV_D and the number of missing disordered segments with minimal length >3 and >10; a native disordered segment is deemed missing if none of its residues are predicted (MSeg) or ≤50% of its residues are predicted (MSeg50%). Methods are sorted in descending order based by their S_W values.

Predictor	Per-residue evaluation									Per-segment evaluation				
	Real-value predictions	Binary predictions								Binary predictions				
	AUC	S_W	MCC_D	Accuracy	ACC	Q_{Oobs}	Q_{Opred}	Q_{Dobs}	Q_{Dpred}	SOV_D	$MSeg_{>3}$	$MSeg_{>10}$	$MSeg50\%_{>3}$	$MSeg50\%_{>10}$
GS-MetaServer2 {153}	0.91	0.67	0.55	89.0	83.5	90.4	97.0	76.6	49.0	75.2	0.22	0.17	0.31	0.29
GeneSilicoMetaServer {297}	0.90	0.66	0.57	90.2	83.4	92.0	96.8	74.8	52.9	74.8	0.25	0.06	0.32	0.31
MULTICOM-CMFR {69}	0.89	0.66	0.50	85.7	83.3	86.4	97.3	80.2	41.3	76.7	0.15	0.17	0.25	0.27
MFDp	0.89	0.63	0.61	92.3	81.7	95.2	96.2	68.2	62.9	66.4	0.38	0.24	0.47	0.41
MULTICOM {453}	0.92	0.62	0.67	94.2	81.2	97.8	95.8	64.6	78.0	69.4	0.31	0.41	0.43	0.45
DISOPRED2	0.88	0.61	0.59	92.0	80.5	95.1	95.9	65.9	61.9	66.8	0.27	0.15	0.45	0.45
DISOclust	0.86	0.59	0.42	81.2	79.7	81.5	96.9	77.9	33.6	72.4	0.14	0.16	0.28	0.30
MD	0.85	0.56	0.42	83.5	78.1	84.9	96.2	71.2	35.8	66.1	0.23	0.08	0.37	0.39
IUPredS	0.85	0.52	0.54	91.5	76.2	95.7	94.9	56.7	61.1	65.3	0.32	0.26	0.51	0.56
IUPredL	0.81	0.46	0.54	92.3	73.1	97.5	94.1	48.7	69.9	49.3	0.60	0.23	0.78	0.70
NORSnet	0.79	0.38	0.49	91.9	69.3	98.0	93.3	40.5	70.5	39.1	0.74	0.59	0.91	0.87
Ucon	0.75	0.35	0.29	83.0	67.6	87.1	93.4	48.0	30.6	50.1	0.40	0.27	0.67	0.72
PROFbval	0.78	0.31	0.19	49.1	65.5	44.7	96.5	86.3	15.5	70.6	0.02	0.00	0.19	0.31

Table 6. Summary of statistical significance tests for AUC (shown in the upper triangle) and MCC_D (shown in the lower triangle) for disorder predictions from Table 4. We performed two types of tests, per residue (denoted as "res") and per sequence (denoted as "seq"). In case of the per residue evaluation, the test was done by bootstrapping 1000 repetitions of 80% of the chains; the same datasets were used for all methods and the results on the corresponding 1000 datasets were compared for a given pair of methods. In the per sequence case, we compared the prediction quality without bootstrapping computed for each of the 121-17=104 sequence; we excluded 17 fully ordered/disordered chains since in these cases per chain quality measures could not be computed The differences between each pair of methods were tested using paired t-test (if samples are normal) or Wilcoxon rank sum test (if they are not normal); normality was validated using Shapiro-Wilk test. For each pair of methods, "+"/"–" mean that a method in the corresponding row performs better/worse, respectively, than the method in the corresponding column with *p*-value < 0.05; "=" denotes that the results are not significantly different. Methods are sorted in the same order as in Table 4.

AUC / MCC_D	{153}		{297}		{69}		Mfdp		{453}		DISOPRED2		DISOclust		MD		IUPredS		IUPredL		NORSnet		Ucon		PROFbval	
	res	seq	res	Seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq
GS-MetaServer2 {153}			+	=	+	+	+	+	–	–	+	=	+	+	+	+	+	=	+	+	+	+	+	+	+	+
GeneSilicoMetaServer {297}	+	+			–	–	+	+	–	–	+	=	+	+	+	+	+	=	+	+	+	+	+	+	+	+
MULTICOM-CMFR {69}	–	=	–	=			+	=	–	–	+	=	+	+	+	+	+	=	+	+	+	+	+	+	+	+
MFDp	+	–	+	–	+	–			–	–	+	+	+	+	+	–	+	=	+	+	+	+	+	+	+	+
MULTICOM {453}	+	=	+	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
DISOPRED2	+	–	+	–	+	+	–	=	–	–			+	+	+	–	+	=	+	+	+	+	+	+	+	+
DISOclust	–	–	–	–	–	–	–	=	–	–	–	=			+	=	+	+	+	+	+	+	+	+	+	+
MD	–	–	–	–	–	–	–	=	–	–	–	=	–	=			+	+	+	+	+	+	+	+	+	+
IUPredS	–	–	–	–	+	+	–	=	–	–	–	=	+	=	+	+			+	+	+	+	+	+	+	+
IUPredL	–	–	–	–	+	–	–	–	–	–	–	–	+	–	+	–	–	–			+	–	+	+	+	–
NORSnet	–	–	–	–	–	–	–	–	–	–	–	–	+	–	+	–	–	–	–	=			+	=	–	–
Ucon	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	=	–	+			–	=
PROFbval	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	+		

Table 7.Comparison of the quality of the solvent accessibility predictions by seven recent sequence-based predictors which were introduced in Table 2 and which provide publicly available web server/standalone program. The methods were tested on 113 CASP8 targets. The table lists per-residue measures including MCC, accuracy, PCC (Pearson's correlation coefficient) and MAE (Mean Absolute Error). The first two measures are specific for binary predictions, while the last two are used to measure performance of real valued predictions. The PCC and MAE measures could not be computed for PaleAle, JNET and ACCpro 4 since these methods predict only binary values. A cut-off of 0.25 was used to convert real values to binary predictions. Methods are sorted in descending order based on their MCC values.

Predictor	Binary predictions		Real-value predictions	
	MCC	Accuracy	PCC	MAE
PaleAle	0.86	0.93	N/A	N/A
NetSurfP	0.59	0.79	0.68	0.14
JNET	0.56	0.78	N/A	N/A
ACCpro 4.03	0.56	0.78	N/A	N/A
SABLE	0.55	0.78	0.65	0.15
SARpred	0.50	0.75	0.58	0.17
Real-SPINE3	0.48	0.75	0.66	0.12

Table 8. Summary of statistical significance tests for MCC (shown in the upper triangle) and PCC (shown in the lower triangle) for solvent accessibility predictions from Table 5. The PCC values for NEtSurfP, JNET, and ACCpro could not be computed (they are denoted as N/A) since these methods provide only binary predictions. We performed two types of tests, per residue (denoted as "res") and per sequence (denoted as "seq"). In case of the per residue evaluation, the test was done by bootstrapping 1000 repetitions of 80% of the chains; the same datasets were used for all methods and the results on the corresponding 1000 datasets were compared for a given pair of methods. We note that PCC is a per sequence measure and it cannot be calculated per residue. Hence, we averaged the PCC values computed for each sequence in the boostrapped dataset. In the per sequence case, we compared the prediction quality without bootstrapping computed for each of the 113 sequence. The differences between each pair of methods were tested using paired t-test (if samples are normal) or Wilcoxon rank sum test (if they are not normal); normality was validated using Shapiro-Wilk test. For each pair of methods, "+"/"–" mean that a method in the corresponding row performs better/worse, respectively, than the method in the corresponding column with *p*-value < 0.05; "=" denotes that the results are not significantly different. Methods are sorted in descending order by their MCC values in Table 5.

	PaleAle		NetSurfP		JNET		ACCpro 4.03		SABLE		SARpred		Real-SPINE3	
MCC / PCC	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq	res	seq
PaleAle			+	+	+	+	+	+	+	+	+	+	+	+
NetSurfP	N/A	N/A			+	+	+	+	+	+	+	+	+	+
JNET	N/A	N/A	N/A	N/A			+	=	+	+	+	+	+	+
ACCpro 4.03	N/A	N/A	N/A	N/A	N/A	N/A			+	=	+	+	+	+
SABLE	N/A	N/A	–	–	N/A	N/A	N/A	N/A			+	+	+	+
SARpred	N/A	N/A	–	–	N/A	N/A	N/A	N/A	–	–			+	+
Real-SPINE3	N/A	N/A	–	–	N/A	N/A	N/A	N/A	+	=	+	+		

Empirical comparison of secondary structure predictors

We compare seven recent secondary structure predictors, including PORTER_H [113], SSpro [99, 100], SPINE [114], YASPIN [106], PSIPRED [95, 96], SABLE [105], and PROTEUS [111, 112]. These predictors offer publicly available standalone software and/or web servers that were used to generate predictions. We used standalone versions of SSpro and PROTEUS without homology modeling, and the web server version of PORTER_H that applies homology modeling. The evaluation is performed on the targets from the most recent completed CASP8 competition. We use the secondary structure derived from the native folds using DSSP as the native descriptor values. The results of the secondary structure predictors are also compared against the secondary structures extracted from the tertiary structure predicted by the top-three performing template-based methods in the CASP8 competition [167], the ZHANG-server, RAPTOR and TASSER.

We perform evaluations at the residue level (predictions are assessed for individual residues and the results are aggregated over the entire dataset) and at the segment level (quality of the predictions is quantified for each sequence and averaged over all chains in the dataset). In the latter case we provide the average value and the corresponding standard deviation. We use the performance measurements used in the EVA platform [168, 169] to quantify the quality. The residue-level measurements include $Q_{Hpred,}$ $Q_{Epred\ and}$ Q_{Cpred} which quantify the fraction of correctly predicted secondary structures among all predicted secondary structures of the same type, and Q_{Hobs}, Q_{Eobs}, and Q_{Cobs} that correspond to the fraction of correct secondary structure predictions among all native (observed) secondary structures of the same type.

$$Q_{Hpred} = \frac{N_{HH}}{\sum_{i\in\{H,E,C\}} N_{iH}} \times 100,\ Q_{Epred} = \frac{N_{EE}}{\sum_{i\in\{H,E,C\}} N_{iE}} \times 100,\ Q_{Cpred} = \frac{N_{CC}}{\sum_{i\in\{H,E,C\}} N_{iC}} \times 100$$

$$Q_{Hobs} = \frac{N_{HH}}{\sum_{j\in\{H,E,C\}} N_{Hj}} \times 100,\ Q_{Eobs} = \frac{N_{EE}}{\sum_{j\in\{H,E,C\}} N_{Ej}} \times 100,\ Q_{Cobs} = \frac{N_{CC}}{\sum_{j\in\{H,E,C\}} N_{Cj}} \times 100$$

where $i, j \in$ {helix H, strand E, coil C}, and N_{ij} stands for the number of residues in the native state of i which has been predicted as state j. We include Q_3 that examines the total number of correct predictions for the three secondary structure states and $Q_{HEerror}$ that is defined as the number of strand residues predicted as helices and vice versa dived by total number of residues in the dataset.

$$Q_3 = \frac{\sum_{i\in\{H,E,C\}} N_{ii}}{N} \times 100,\ Q_{HEerror} = \frac{N_{HE} + N_{EH}}{N} \times 100$$

where N denotes the number of all residues.

We also compute the residue level Matthews Correlation Coefficient (MCC)

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}$$

where TP stands for true positive which is the number of correctly predicted positives (say, helix residues when computing MCC_H), FP stands for false positive which is the number of native positives which are predicted as negatives (helix residues predicted as coil or strand residues), TN stands for true negative which is the number of correctly predicted negative residues, and FN stands for false negative and denotes the number of native negatives predicted as positives. The MCC values are computed for prediction of helix (MCC_H), strand (MCC_E), and coil (MCC_C), and they range between -1 and 1 with higher values corresponding to better prediction performance.

Finally, we include the segment-level segment overlap (SOV) measures [170]. The SOV quantifies the amount of overlap between the native and the predicted segments; we compute it for both secondary structure and disorder predictions. We measure SOV for helix segments (SOV_H), strand segments

(SOV_E), coil segments (SOV_C), and SOV_3 which quantifies the overall segment overlap for all three secondary structure states.

The results for the secondary structure prediction are summarized in Table 3. The best-performing PORTER_H achieves 83.4% Q_3 and 81.1 SOV_3. The second best SSpro, which in contrast to PORTER_H does not utilize homology modeling, obtains 80.4% Q_3 and 78.3 SOV_3. The Q_3 values of the considered secondary structure predictors range between 77% and 83% compared to the 76% to 81% range obtained by the best performing fold predictors. The SOV_3 values of three secondary structure predictors, PORTER_H, SSpro, and YASPIN are higher than the segment overlaps computed for the three tertiary structure predictors. The predictions of helix residues and segments are characterized (as expected) by higher quality than for the strands and coils. The SOV_H and MCC_H values of two secondary structure predictors are at above 82 and above 0.77, respectively; these results match the quality of the predictions generated by the ZHANG-server. The YASPIN and PORTER_H provide high-quality predictions of the strand segments with SOV_E at over 78 which equals to the performance of the RAPTOR. We observe that YASPIN slightly over-predicts strand residues, i.e., it has high Q_{Eobs} and low Q_{Epred}, when compared with the other methods. Interestingly, between 1.3% and 2.2% of all secondary structure predictions include helical residues confused for strand residues and vice versa. These mistakes are less prevalent in the tertiary structure predictions where the corresponding rates are at about 0.6%. Overall, we note that there is no clear-cut winner, i.e., none of the methods obtains favorable prediction quality on all measures; a similar conclusion was drawn in a recent evaluation of standalone secondary structure predictors [171]. Although PORTER_H obtains the highest overall Q_3 and SOV_3, the runner-up SSpro, SPINE and YASPIN provide relatively high-quality predictions for helices and coils, coils, and strands, respectively. We also conclude that the secondary structure in the protein folds predicted by the best-performing template-based methods is of comparable quality to the secondary structure predicted by the modern secondary structure predictors that do not utilize templates.

We also evaluate statistical significance of the differences between each pair of the considered methods. We perform the evaluation at the residue and the sequence level. In the per residue case, we follow the procedure from the evaluation of the disorder predictions in the CASP8 experiment [76]. We compare 1000 paired results (two methods compared on the same datasets) obtained using the bootstrapping with 80% of the targets from the CASP8 dataset, i.e., we compare the per-residue evaluation on 1000 subsets of the entire benchmark set. For SOV and PCC measures that are calculated per sequence, we compute their average over all sequence in a given subset. We also perform the per sequence significance test in which case we compare paired results for each sequence over the entire benchmark dataset. We verify whether the input measurements follow normal distribution, as tested using Shapiro-Wilk test at the 0.05 significance. If both measurements are normal then we use paired-test, otherwise we use Wilcoxon rank sum test. We assume that a given pair of methods is significantly different if the corresponding p-value < 0.05.

Table 4 summarizes the result of the significance tests for the predictions of the secondary structure that are evaluated using the Q_3 (in the upper triangle) and SOV_3 (in the lower triangle) measures. The "+"/"–" denote that the method in a given row performs better/worse, respectively, than the method in the corresponding column with p-value < 0.05; "=" means that there is no significant difference in the performance for a given pair of methods. The best performing PORTER_H predictor significantly outperforms the other considered secondary structure prediction methods for both quality measures, which is likely a direct effect of the fact that this method utilizes homology modeling. The second best SSpro, which does not use homology modeling, similarly significantly improves over the remaining 5 secondary structure predictors for Q_3 and SOV_3, except for SPINE and YASPIN when considering the per sequence comparison. We note that the per residue and the per sequence results are relatively

consistent for the considered pairs of the secondary structure prediction methods. This suggests that the prediction quality does not vary much when evaluated for individual sequences and over a dataset of sequences, i.e., there are no methods that do well over a dataset but poorly for some sequences and vice versa. Among the tertiary structure predictors, the Zhang-server significantly outperforms the other two methods, except for RAPTOR which is equivalent to the Zhang-server for the per sequence evaluation. Interestingly, we note that PORTER_H significantly improves over the Zhang-server and the other two tertiary structure predictors when evaluated using both Q_3 and SOV_3. The second best secondary structure predictor SSpro is outperformed by Zhang-server and RAPTOR based on the per residue Q_3 and, but it provides significantly better SOV_3 values.

Empirical comparison of disorder predictors

We contrast eight disorder predictors, including MFDp [172], MD [173], DISOclust [174], NORSnet [175], Ucon [176], PROFbval [164], IUPred that predicts long (IUPredL) and short (IUPredS) disordered regions [177] and DISOPRED2 [178]. Standalone software or web servers of these methods were used to generate predictions. We use the disorder annotations provided by the assessors of the CASP8 as the native descriptor values and we compare the results of the abovementioned disorder predictors with the top-three predictors of the binary disorder annotations from the CASP8 [76]. These three methods are identified by the group number in curly brackets (as registered for the CASP8 meeting) and the group name, and they include GS-MetaServer2 {153}, GeneSilicoMetaServer {297}, MULTICOM-CMFR {69} [179, 180], as well as MULTICOM {453} which is a human (expert-based) meta-predictor. The GS-MetaServer2 and GeneSilicoMetaServer are unpublished but they offer a web server at https://genesilico.pl/meta2/. We use a subset of 111 and 121 CASP8 targets in the case of the secondary structure and disorder evaluations, respectively, for which all considered methods were able to provide predictions.

The same as in the CASP8, we discard the native disordered regions with 3 or fewer residues [76] (private correspondence with authors), i.e., these residues are ignored when computing the quality measures. We use two residue-level measures that are specific to evaluation of disorder predictions that were utilized at the recent CASP experiments [76, 77] including S_w and ACC, which is an average between sensitivity and specificity.

$$ACC = [(TP / N_{disorder}) + (TN / N_{order})] / 2$$

$$S_w = (w_{disorder}*TP - w_{order}*FP + w_{order}*TN - w_{disorder}*FN) / (w_{disorder}* N_{disorder} + w_{order}* N_{order})$$

where TP stands for true positive which is the number of correctly predicted disordered residues, FP stands for false positive which is the number of natively ordered residues which are predicted as disorder, TN stands for true negative which is the number of correctly predicted ordered residues, and FN stands for false negative and denotes the number of natively disordered residues predicted as ordered, $w_{disorder}$ and w_{order} are the fractions of ordered and disordered residues, respectively, and N_{order} and $N_{disorder}$ are the total number of disordered and ordered residues, respectively. Similarly as for the secondary structure, we measure Q_{Oobs} , Q_{Opred}, Q_{Dobs}, and Q_{Dpred} to evaluate the residue level predictions and SOV_D for the disordered segments. The former four measures quantify the fraction of the correct predictions of ordered (O) and disordered (D) residues among all predicted and all native ordered and disordered residues, respectively. The SOV_D denotes the segment overlap between the predicted and the native disordered segments. We also compute residue level MCC_D and accuracy

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

We measure the area under the ROC (AUC) to evaluate the quality of the predicted real-value probabilities of disorder that are outputted by all considered predictors. For each value of probability p predicted by a given method (between 0 and 1), all the residues with probability equal or greater than p are set as disordered, and all other residues are set as ordered. Next, the TP-rate = TP / (TP + FN) and the FP-rate = FP / (FP + TN) are calculated and we use the area under the corresponding curve to quantify the predictive quality.

Finally, we introduce and quantify four segment-level measures for the disorder predictions. They quantify the fraction of the disordered segments that were missing (not predicted) among all segments of a given size range, which include segment with >3 residues and with >10 residues. Our benchmark set includes 170 native segments of 3 or more disordered residues and 86 segments with 10 or more disordered residues. We consider two cut-offs to define a given segment as missing, when none of its residues are predicted as disordered ($MSeg_{>3}$ and $MSeg_{>10}$ measures) and when ≤50% of its residues are predicted as disordered ($MSeg50\%_{>3}$ and $MSeg50\%_{>10}$ measures).

The quality of the disorder predictions is analyzed in Table 5. Seven predictors, including the four CASP8 participants, MFDp, DISOPRED2, and DISOclust achieve AUC > 0.85 and S_w of about 0.6 or higher. The ROCs for these predictors are shown in Figure 1. We constrain the FP-rate range to 0-0.2 since the disordered residues constitute only 11% of all residues in our dataset, i.e., higher FP-rates would lead to a substantial over-prediction of the disordered residues. We observe that MULTICOM outperforms all other considered methods, and among the published predictors, MD works well for FP-rates < 0.085 and MFDp provides favorable TP-rates for FP-rates > 0.1.We note that some of the prediction methods tend to over-predict the disordered residues as their Q_{Dpred} are relatively low and Q_{Dobs} are relatively high, which was also observed in [181]. For instance, GS-MetaServer2, GeneSilicoMetaServer, MULTICOM-CMFR, DISOclust, MD, and PROFbval have Q_{Dpred} at about 50 or below (with Q_{Dobs}> 71), which means that most of the disordered residues that they predict are in fact annotated as structured. These methods can still obtain high (including the highest) S_w values; this measure favors over-prediction of the disorder considering that there are only 11% of disordered residues in our dataset (i.e., w_{order} = 0.11 and $w_{disorder}$ = 0.89). The two highest MCC_D values are achieved by MULTICOM-CMFR and MFDp and these methods also have one of the highest accuracy and AUC values. The segment-level evaluations reveal that when excluding the methods that over-predict disorder (GS-MetaServer2, GeneSilicoMetaServer, MULTICOM-CMFR, DISOclust, MD, and PROFbval) between 27% (for DISOPRED2) and 74% (for NORSnet) of disordered segments are completely missed (none of their residues are predicted), and less than a half of the disordered residues are predicted for between 43% (for MULTICOM) and 91% (for NORSnet) segments. Even when considering only longer disordered segments with at least 10 residues, these rates are relatively high, i.e., 15% to 59% are completely missed and 41% to 87% are predicted with less than a half residues. The segment overlap values for the disordered segments range between 39% and 77%. To compare, the highest SOV values for coils, strand and helix segments, which were achieved by PORTER_H, are slightly higher and they equal 78%, 79%, and 83%, respectively.

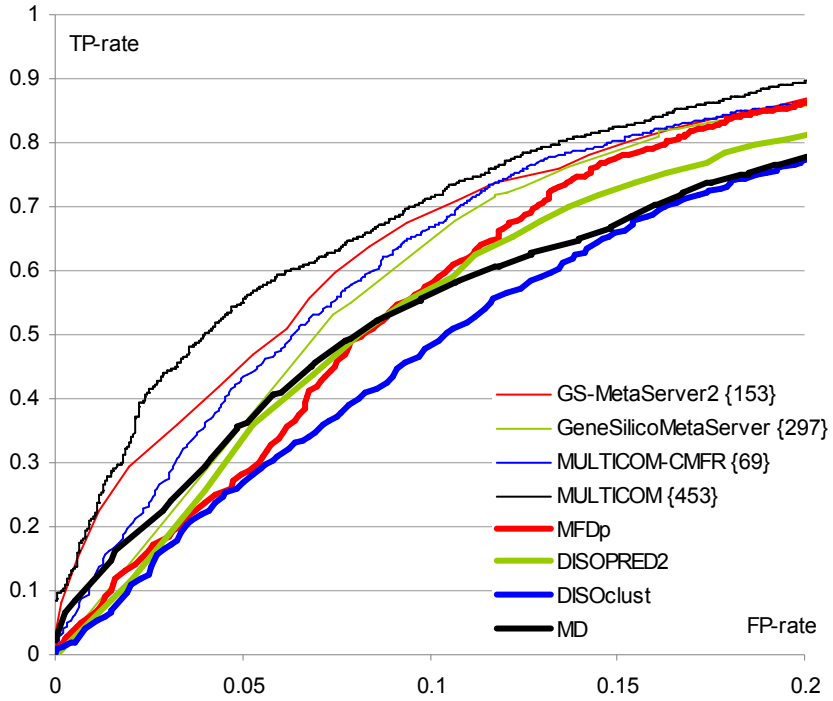

Figure 1.The ROCs of the eight best considered disorder prediction methods, including four CASP4 participants (shown using thin lines) and four public web servers (shown using thick lines). The *x*-axis is the TP-rate, which is constrained to 0-0.2 range, and *y*-axis is the FP-rate.

Table 6 shows the results of the evaluation of the statistical significance of the differences for all pairs of the considered disorder predictors; the setup follows the significance tests for the secondary structure and the methods are sorted by their S_w values from Table 5. Table 6 presents the results for AUC which is used to validate real-value predictions (in the upper triangle) and MCC_D for the binary predictions (in the lower triangle). The best performing MULTICOM significantly outperforms all the other methods in terms of AUC and MCC_D, except only for the GS-MetaServer2 which provides equivalent per sequence MCC_D. The best predictor among the remaining methods with respect to MCC_D is MFDp, see Table 5, and this method significantly improves over the other 11 methods for the per residue MCC_D. In case of the AUC based evaluation, the second best GS-MetaServer2 also outperforms the other 11 predictors for the per residues AUC. We note several results where the per residue and the per sequence tests lead to opposing conclusions, i.e., where a given method that performs significantly better than the other method in the per residue case performs significantly worse in the per sequence case, and vice versa. Examples of such methods are IUPredL, NORSnet, Ucon, PROFbval, DISOPRED2, and MFDp when considering the evaluations based on the MCC_D. We observe that IUPredL, NORSnet, Ucon, and PROFbval are characterized by a substantial difference in magnitude of MCC_D values between the per residue and per sequence evaluations. For example, the IUPredL (MCC_D = 0.54) performs better than both MD (MCC_D = 0.42) and DISOclust (MCC_D = 0.42) in the per residue test, while for the per sequence test the average (over the benchmark dataset) MCC_D of IUPredL drops to 0.20, when compared with MCC_D of MD and DISOclust that equal 0.31 and 0.32, respectively. This suggests that while these four methods (IUPredL, NORSnet, Ucon, and PROFbval) are characterized by good quality predictions on the entire benchmark dataset, they perform relatively poorly on some sequences. We note that although differences in magnitude of MCC_D between the per residue and the per sequence tests for the DISOPRED2 and MFDp are relatively small, these two methods still drop significantly below the GS-MetaServer2 and GeneSilicoMetaServer in the per sequence evaluation. We also observe several differences between the per residue and the per sequence comparison for the AUC-based significance tests. For example, comparing MFDp and DISOPRED2 with MD, we note that the former two methods are significantly better in the per residue test, but they are significantly worse for the per

sequence test. This is once again caused by the lower per sequence AUC of the MFDp and DISOPRED when compared with the per residue AUC; our analysis shows that these differences have relatively small magnitude.

Empirical comparison of solvent accessibility predictors

We empirically contrast seven solvent accessibility predictors, PaleAle [113], NetSurfP [152], JNET [84], ACCpro version 4.03 [100], SABLE [55], SARpred [147], and Real-SPINE3 [82, 83]. These methods provide publicly available standalone software and/or web servers that were used to generate predictions. We used standalone versions of Real-SPINE3 and ACCpro 4.03 without homology modeling. The evaluation utilizes a subset of 113 targets from the CASP8 experiment for which all seven methods generated predictions. The above methods predict relative solvent accessibility (RSA). We use the absolute solvent accessibility (ASA) values derived from the native folds using DSSP to compute the native relative solvent accessibility (RSA) as follows:

$$RSA_i = ASA_i / MSA_i$$

where i is a residue index and MSA_i is the maximum obtainable solvent accessibility for the corresponding amino acid type. The MSA_i is used to normalize the corresponding ASA_i value with respect to the overall size of a given amino acid type. There are several ways to quantify the MSA values, with two most prevalent that are based on the solvent accessibility of residue X in an extended Ala-X-Ala [82] and Gly-X-Gly [56] tripeptides. We normalize the ASA values generated by the DSSP using the same MSA values as were applied in the assessed predictors. Specifically, we used normalization factors from [82] for NetSurfP and SARpred, from [56] for SABLE, from [182] for JNET, from [57] for REAL-Spine3, and the values provided by Dr. Pollastri (personal communication) for PaleAle and ACCpro. Similarly as in [73, 82-85, 108, 113, 152] we converted the RSA values to binary classes using threshold of 0.25, i.e., residues with RSA >0.25 were considered as exposed, and with RSA ≤ 0.25 as buried.

We compute four performance measures including MCC and accuracy for the binary predictions, and Pearson's correlation coefficient (PCC) and mean absolute error (MAE) to evaluate the predicted RSA values. Only the NetSurfP, SABLE, SARpred, and Real-SPINE3 generate RSA values (and consequently binary predictions that are obtained from the RSA values), while the other methods generate only the binary predictions.

The results are summarized in Table 7. The PaleAle achieves MCC equal 0.86 and outperforms the other methods by a wide margin. The second best NetSurfP obtains MCC = 0.59 and three other methods have MCC around 0.56. The reason for this gap is the fact that PaleAle uses homology modeling with a template library that likely includes some of the considered here CASP8 targets. Interestingly, we observe that high quality of the RSA predictions is not necessarily coupled with high quality of the binary predictions, and vice versa. For instance, the NetSurfP which obtains the highest MCC when compared with the other three methods that generate RSA values has higher (worse) MAE value than the Real-SPINE3 which obtains the lowest MCC of 0.12. A potential reason is that some of the considered methods could be optimized to maximize the predictive performance using a different cut-off to define the binary classes.

Table 8 summarizes the results of statistical significance test for the solvent accessibility and it considers the evaluations based on the MCC (in the upper triangle) and PCC (in the lower triangle) measures; the setup follows the significance tests for the secondary structure predictions. The PCC could not be computed for PaleAle, JNET, ACCpro since these methods do not output the RSA; the corresponding cells in the Table 8 are denoted as "not applicable" (N/A). The best performing with respects to MCC PaleAle significantly outperforms all other methods for this quality measure. The second best NetSurfP also provides significantly higher MCC when compared with the remaining five methods. Similarly, NetSurfP that obtains the highest PCC (see Table 7) significantly outperforms

SABLE, SARpred, and Real-SPINE3 when considering correlation between the predicted and the native RSA. The second best Real-SPINE provides significantly improved PCC when compared with the SABLE and SAPpred. We observe that the per residue and the per sequence results are relatively consistent for all pairs of methods.

Conclusions

The 1D descriptors of protein structure cover a wide-range of structural aspects including conformation of the backbone, burying depth/solvent exposure and flexibility of residues, and inter-chain residue-residue contacts. These descriptors are widely used to characterize and analyze protein folds and to predict various structural and functional characteristics of proteins. They can be either computed from the know structure or predicted from the primary sequence. The last two decades observed substantial efforts in the development of accurate sequence-based predictors. Our overview of a several dozens of the most recent predictors shows that they are based on a common architecture in which the protein sequence is represented as a feature vector using evolutionary profiles and, in some cases, predictions of a few related 1D descriptors, which is fed into a machine learning-based prediction model. The most popular models include neural networks, support vector machines, hidden Markov models, and support vector and multiple linear regressions. Our empirical evaluation of the quality of the sequence-based prediction of secondary structure, disorder, and solvent accessibility descriptors shows that these machine learning models generate high-quality results. For instance, the secondary structure can be predicted with over 80% accuracy and segment overlap, the disorder with over 0.9 AUC, 0.6 MCC, and 75% SOV, and relative solvent accessibility with PCC of 0.7 and MCC of 0.6 (0.86 when homology is used). We caution the reader that these results are based on a relatively small, although representative (according to the CASP8 organizers), dataset and thus our conclusions may not generalize to other, larger or more specialized, e.g. constrained to specific types of protein folds, protein sets. The utility of the considered prediction methods is further strengthened by the fact that most of them are accessible to a non-expert user via web servers or standalone software packages. We anticipate that the 1D structural protein descriptors will play a significant role in various related fields such as high-throughput protein structure and function annotation, characterization and prediction of protein-ligand and protein-protein interactions, and rational drug design, to name a few.

References

[1] Kinjo, A.; Nishikawa, K. Recoverable one-dimensional encoding of three-dimensional protein structures. *Bioinformatics,* **2005**, *21*(10), 2167-2170.

[2] Wu, S.; Zhang, Y. ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS One* **2008**, *3*(10), e3400.

[3] Faraggi, E.; Yang, E. ,Zhang, S.; Zhou, Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction *Structure,* **2009**, *17*(11), 1515-1527.

[4] Benkert, P.; Tosatto, S.C.; Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins,* **2008**, *71*(1), 261-277.

[5] Yu, C.; Joachims, T.; Elber, R.; Pillardy, J. Support vector training of protein alignment models. *J. Comput .Biol.,* **2008**, *15*(7), 867-880.

[6] Gao, J.; Zhang, T.; Zhang, H.; Shen, S.; Ruan, J.; Kurgan, L. Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. *Proteins,* **2010**, *78*(9), 2114-2130.

[7] Fischer, J.D.; Mayer, C.E.; Söding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics,* **2008**, *24*(5), 613-620.

[8] Kurgan, L.; Mizianty, M. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Nat. Sci.,* **2009**, *1*(2), 93-106.

[9] Slabinski, L.; Jaroszewski, L.; Rodrigues, A.P.; Rychlewski, L.; Wilson, I.A.; Lesley, S.A.; Godzik, A. The challenge of protein structure determination--lessons from structural genomics. *Protein Sci.,* **2007**, *16*(11), 2472-2482.
[10] Rost, B. Prediction of protein structure in 1D - secondary structure, membrane regions, and solvent accessibility. In *Structural Bioinformatics,* 2nd ed.; Bourne, P.E.; Ed.; Wiley, **2009**; pp 679-714.
[11] Xia, J.; Wang, S.L.; Lei, Y.K. Computational methods for the prediction of protein-protein interactions. *Protein Pept. Lett.,* **2010**, *17*(9), 1069-1078.
[12] Ding, X.; Pan, X.Y.; Xu, C.; Shen, H.B. Computational prediction of DNA-protein interactions: a review. *Curr. Comput. Aided Drug Des.,* **2010**, *6*(3), 197-206.
[13] Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.,* **2010**, *4*(Suppl 1), S3.
[14] Zhang, T.; Zhang, H.; Chen, K,. Ruan, J.; Shen, S.; Kurgan, L. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.,* **2010**, *11*(7), 609-628.
[15] Sankararaman, S.; Sha, F.; Kirsch, J.F.; Jordan, M.I.; Sjölander, K. Active site prediction using evolutionary and structural information. *Bioinformatics,* **2010**, *26*(5), 617-624.
[16] Zhang, T.; Zhang, H.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Accurate sequence-based prediction of catalytic residues. *Bioinformatics,* **2008**, *24*(20), 2329-2338.
[17] Hiss, J.; Schneider, G. Architecture, function and prediction of long signal peptides. *Brief Bioinform.,* **2009**, *10*(5), 569-578.
[18] Pirovano, W.; Heringa, J. Protein secondary structure prediction. *Methods Mol. Biol.,* **2010**, *609*, 327-348.
[19] Rost, B. Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem. Anal.,* **2003**, *44*, 559-587.
[20] Rost, B. Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.,* **2001**, *134*(2-3), 204-218.
[21] Peng, Z.-L.; Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci*., **2011**. [In press].
[22] He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V.N.; Dunker, A.K. Predicting intrinsic disorder in proteins: an overview. *Cell Res.,* **2009**, *19*(8), 929-949.
[23] Kuznetsov, I. Simplified computational methods for the analysis of protein flexibility. *Curr. Protein Pept. Sci.,* **2009**, *10*(6), 607-613.
[24] Chou, K.C.; Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.,* **2009**, *2*, 63-69.
[25] Koch, O.; Klebe, G. Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins,* **2009**, *74*(2), 353-367.
[26] Meissner, M.; Koch, O.; Klebe, G.; Schneider, G. Prediction of turn types in protein structure by machine-learning classifiers. *Proteins,* **2009**, *74*(2), 344-352.
[27] Zheng, C.; Kurgan, L. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. *BMC Bioinform.,* **2008**, *9*, 430.
[28] Hu, X.; Li, Q. Using support vector machine to predict beta- and gamma-turns in proteins. *J. Comput. Chem.,* **2008**, *29*(12), 1867-1875.
[29] Wang, Y.; Xue, Z.D.; Shi, X.H.; Xu, J. Prediction of pi-turns in Proteins using PSI-BLAST profiles and secondary structure information. *Biochem. Biophys. Res. Commun.,* **2006**, *347*(3), 574-580.
[30] Freeman, T.J.; Wimley, W.C. A highly accurate statistical approach for the prediction of transmembrane {beta}-barrels. *Bioinformatics,* **2010**, *26*(16), 1965-1974.

[31] Chou, K.; Carlacci, L.; Maggiora, GM. Conformational and geometrical properties of idealized beta-barrels in proteins. *J. Mol. Biol.,* **1990**, *213*(2), 315-326.
[32] Gruber, M.; Söding, J.; Lupas, .AN. Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.,* **2006**, *155*(2), 140-145.
[33] Lupas, A. Coiled coils: new structures and new functions. *Trends Biochem. Sci.,* **1996**, *21*(10), 375-382.
[34] Zimmermann, O.; Hansmann, U.H. LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J. Chem. Inf. Model.,* **2008**, *48*(9), 1903-1908.
[35] Hamelryck, T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins,* **2005**, *59*(1), 38-48.
[36] Song, J.; Tan, H.; Takemoto, K.; Akutsu, T. HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics,* **2008**, *24*(13), 1489-1497.
[37] Bartoli, L.; Capriotti, E.; Fariselli, P.; Martelli, P.L.; Casadio, R. The pros and cons of predicting protein contact maps. *Methods Mol. Biol.,* 2008, *413*, 199-217.
[38] Dosztányi, Z.; Meszaros, B.; Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform.,* **2010**, *11*(2), 225-243.
[39] Midic, U.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome. *Protein Pept. Lett.,* **2009**, *16*(12), 1533-1547.
[40] Levitt, M.; Greer, J. Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.,* **1977**, *114*(2), 181-239.
[41] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **1983**, *22*(12), 2577-2637.
[42] Richards, F.; Kundrot, C.E. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. . *Proteins,* **1988**, *3*(2), 71-84.
[43] Sklenar, H.; Etchebest, C.; Lavery, R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins,* 1989, 6:46-60. *Proteins,* **1989**, *6*(1), 46-60.
[44] Frishman, D.; Argos, P.; Knowledge-based protein secondary structure assignment. *Proteins,* **1995**, 23(4), 566-579.
[45] Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J.P. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput. Appl. Biosci.,* **1997**, *13*(3), 291-295.
[46] King, S.; Johnson, W.C. Assigning secondary structure from protein coordinate data. *Proteins,* **1999**, *3*(35), 313-320.
[47] Fodje, M.; Al-Karadaghi, S. Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng.,* **2002**, *15*(5), 353-358.
[48] Martin, J.; Letellier, G.; Marin, A.; Taly, J.F.; de Brevern, A.G.; Gibrat, J.F. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.,* **2005**, *5*, 17.
[49] Nugent, T.; Jones, D.T. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput. Biol.,* **2010**, *6*(3), e1000714.
[50] Ikeda, M.; Arai, M.; Okuno, T.;Shimizu, T. TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.,* **2003**, *31*(1), 406-409.
[51] Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*(1), 95-99.
[52] Sims, G.E.; Choi, I.G.; Kim, S.H. Protein conformational space in higher order phi-Psi maps. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*(3), 618-21.

[53] Lee, B.; Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.,* **1971**, *55*(3), 379-400.
[54] Richards, F. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.,* **1977**, *6*, 151-176.
[55] Adamczak, R.; Porollo, A.; Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins,* **2004**, *56*(4), 753-767.
[56] Chothia, C . The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.,* **1976**, *105*,(1), 1-12.
[57] Ahmad, S.; Gromiha, M.M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins,* **2003**, *50*(4), 629-635.
[58] Chakravarty, S.; Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure,* **1999**, *7*(7), 723-732.
[59] Pintar, A.; Carugo, O.; Pongor, S. Atom depth as a descriptor of the protein interior. *Biophys. J.,* **2003**, *84*(4), 2553-2561.
[60] Pintar, A.; Carugo, O.; Pongor, S. DPX: for the analysis of the protein core. *Bioinform.,* **2003**, *19*(2), 313-314.
[61] Varrazzo, D.; Bernini, A.; Spiga, O.; Ciutti, A.; Chiellini, S.V.; Bracci, L.; Niccolai, N. Three-dimensional computation of atom depth in complex molecular structures. *Bioinformatics,* **2005**, *21*(12), 2856-2860.
[62] Yuan, Z.; Wang, Z.X. Quantifying the relationship of protein burying depth and sequence. *Proteins,* **2008**, *70*(2), 509-516.
[63] Zhang, H.; Zhang, T.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. *BMC Bioinform.,* **2008**, *9*, 388.
[64] Nishikawa, K.; Ooi, T. Prediction of the surface-interior diagram of globular *Proteins,* by an empirical method. *Int. J. Pept. Protein Res.,* **1980**, *16*(1), 19-32.
[65] Nishikawa, K.; Ooi, T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem.,* **1986**, *100*(4), 1043-1047.
[66] Pollastri, G.; Baldi, P.; Fariselli, P.; Casadio, R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins,* **2002**, *47*(2), 142-153.
[67] Kinjo, A.; Horimoto, K.; Nishikawa, K. Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins,* **2005**, *58*(1), 158-165.
[68] Kinjo, A.; Nishikawa, K. Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structure from amino acid sequence using critical random network. *Biophysics,* **2005**, *1*, 67-74.
[69] Plaxco, K.; Simons, K.T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*(4), 985-994.
[70] Parthasarathy, S.; Murthy, M.R.N. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.,* **1997**, *6*(12), 2561-2567.
[71] Carugo, O.; Argos, P. Correlation between side chain mobility and conformation in protein structures. *Protein Eng.,* **1997**, *10*(7), 777-787.
[72] Schlessinger, A.; Rost, B. Protein flexibility and rigidity predicted from sequence. *Proteins,* **2005**, *61*(1), 115-126.
[73] Zhang, H.; Zhang, T.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins,* **2009**, *76*(3), 617-636.
[74] Vucetic, S.; Brown, CJ.; Dunker, AK.; Obradovic, Z. Flavors of protein disorder. *Proteins,* **2003**, *52*(4), 573-584.
[75] Bracken, C.; Iakoucheva, L.M.; Romero, P.R.; Dunker, A.K. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.,* **2004**, *14*(5), 570-576.

[76] Noivirt-Brik, O.; Prilusky, J.; Sussman, J.L. Assessment of disorder predictions in CASP8. *Proteins,* **2009**, 77, 210-216.
[77] Bordoli, L.; Kiefer, F.; Schwede, T. Assessment of disorder predictions in CASP7. *Proteins,* **2007**, *69*(Suppl 8), 129-136.
[78] Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; Obradovic, Z.; Dunker, A.K. DisProt: the database of disordered proteins. *Nucleic Acids Res.,* **2007**, *35*(Database issue), D786-D793.
[79] Debye, P. Interferenz von Röntgenstrahlen und Wärmebewegung. *Ann. Phys.,* **1913**, *348*(1), 49-92.
[80] Pauling, L.; Corey, R.B.; Branson, H.R.;The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Nat. Acad. Sci. USA,* **1951**, *37*(4), 205-211.
[81] Pauling, L.; Corey, R.B.; The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl. Acad. Sci. USA,* **1951**, *37*(5), 251-256.
[82] Faraggi, E.; Xue, B.; Zhou, Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins*,* by guided-learning through a two-layer neural network. *Proteins,* **2009**, *74*(4), 847-856.
[83] Dor, O.; Zhou, Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins,* **2007**, *68*(1), 76-81.
[84] Cuff, J.; Barton, G.J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins,* **2000**, *40*(3), 502-511.
[85] Rost, B.; Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins,* **1994**, *20*(3), 216-226.
[86] Berman, H.; Henrick, K.; Nakamura, H.; Markley, J.L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.,* **2007**, *35*(Database issue), D301-3.
[87] Anfinsen, C. Principles that govern the folding of protein chains. *Science,* **1973**, *181*(96), 223-230.
[88] Chen, C.; Kernytsky, A.; Rost, B. Transmembrane helix predictions revisited. *Protein Sci.,* **2002**, *11*(12), 2774-2791.
[89] Fleishman, S.; Ben-Tal, N. Progress in structure prediction of -helical membrane proteins. *Curr.Opin. Struct. Biol.,* **2006**, *16*, 496-504
[90] Dosztányi, Z.; Sándor, M.; Tompa, P.; Simon, I. Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci.,* **2007**, *8*(2), 161-171.
[91] Bourhis, J.; Canard, B.; Longhi, S. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr. Protein Pept. Sci.,* **2007**, *8*(2), 135-149.
[92] Altschul, S.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.,* **1990**, *215*(3), 403-410.
[93] Rost, B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.,* **1996**, *266*, 525-539.
[94] Rost, B.; Yachdav, G.; Liu, J. The PredictProtein Server. *Nucleic Acids Res.,* **2004**, *32*(Web Server issue), W321-W326.
[95] Jones, D. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.,* **1999**, *292*(2), 195-202.
[96] Bryson, K.; McGuffin, L.J.; Marsden, R.L.; Ward, J.J.; Sodhi, J.S.; Jones, DT. Protein structure prediction servers at University College London. *Nucleic Acids Res.,* **2005**, *33*(Web Server issue), W36-W38.
[97] Cole, C.; Barber, J.D.; Barton, G.J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.,* **2008**, *36*(Web Server issue), W197-W201.
[98] Ouali, M.; King, R.D. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.,* **2000**, *9*(6), 1162-1176.

[99] Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins,* **2002**, *47*(2), 228-235.
[100] Cheng, J.; Randall, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.,* **2005**, *33*(Web Server issue), W72-W76.
[101] Kurowski, M.; Bujnicki, J.M. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.,* **2003**, *31*(13), 3305-3307.
[102] Karplus, K.; Karchin, R.; Draper, J.; Casper, J.; Mandel-Gutfreund, Y.; Diekhans, M.; Hughey, R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins,* **2003**, *53*(6), 491-496.
[103] Karplus, K.; Katzman, S.; Shackleford, G.; Koeva, M.; Draper, J.; Barnes, B.; Soriano, M.; Hughey, R. SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins,* **2005**, *61*(Suppl 7), 135-142.
[104] Karplus, K.; SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.,* **2009**, *37*(Web Server issue), W492-W497.
[105] Adamczak, R.; Porollo, A.; Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins,* **2005**, *59*(3), 467-475.
[106] Lin, K.; Simossis, V.A.; Taylor, W.R.; Heringa, J. A simple and fast secondary structure prediction algorithm using hidden neural networks. *Bioinformatics,* **2005**, *21*(2), 152-159.
[107] Pollastri, G.; McLysaght, A. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics,* **2005**, *21*(8), 1719-1720.
[108] Qin, S.; He, Y.; Pan, X.M. Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. *Proteins,* **2005**, *61*(3), 473-480.
[109] Karypis, G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins,* **2006**, *64*(3), 575-586.
[110] Martin, J.; Gibrat, J.F.; Rodolphe, F. Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct. Biol*., **2006**, *6*, 25.
[111] Montgomerie, S.; Sundararaj, S.; Gallin, W.J.; Wishart, D.S. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinform.,* **2006**, *7*, 301.
[112] Montgomerie, S.; Cruz, J.A.; Shrivastava, S.; Arndt, D.; Berjanskii, M.; Wishart, D.S. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res.,* **2008**, *36*(Web Server issue), W202-W209.
[113] Pollastri, G.; Martin, A.J.M.; Mooney, C.; Vullo, A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinform.,* **2007**, *8*, 201.
[114] Dor, O.; Zhou, Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins,* **2007**, *66*(4), 838-845.
[115] Won, K.; Hamelryck, T.; Prügel-Bennett, A.; Krogh, A. An evolutionary method for learning HMM structure: prediction of protein secondary structure. *BMC Bioinform.,* **2007**, *8*, 357.
[116] Kountouris, P.; Hirst, J.D. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinform.,* **2009**, *10*, 437.
[117] Green, J.; Korenberg, M.J.; Aboul-Magd, M.O. PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinform.,* **2009**, *10*, 222.
[118] Zhou, T.; Shu, N.; Hovmöller, S. A novel method for accurate one-dimensional protein structure prediction based on fragment matching. *Bioinformatics,* **2010**, *26*(4), 470-477.
[119] von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **1992**, *255*(2), 487-494.
[120] Claros, M.G.; von Heijnea, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **1994**, *10*(6), 685-686.

[121] Jones, D.; Taylor, W.R. Thornton, JM.; A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* **1994**, *33*(10), 3038-3049.
[122] Jones, D. Do transmembrane protein superfolds exist? *FEBS Lett.,* **1998**, *423*(3), 281-285.
[123] Jones, D. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics,* **2007**, *23*(5), 538-544.
[124] Rost, B.; Fariselli, P.; Casadio, R. Topology prediction for helical transmembrane *Proteins,* at 86% accuracy. *Prot. Sci.,* **1996**, *4*(8), 521-533.
[125] Tusnády, G.; Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.,* **1998**, *283*(2), 489-506.
[126] Tusnády, G.; Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics,* **2001**, *17*(9), 849-850.
[127] Sonnhammer, E.L.; von Heijne, G.; Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Intell. Syst. Mol. Biol.,* **1998**, *6*, 175-182.
[128] Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.,* **2001**, *305*(3), 567-580.
[129] Kall, L.; Krogh, A.; Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.,* **2004**, *338*(5), 1027-1036.
[130] Käll, L.; Krogh, A.; Sonnhammer, E.L. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res.,* **2007**, *35*(Web Server issue), W429-W432.
[131] Yuan, Z.; Mattick, J.S.; Teasdale, R.D. SVMtm: support vector machines to predict transmembrane segments. *J. Comput. Chem.,* **2004**, *25*(5), 632-636.
[132] Cao, B.; Porollo, A.; Adamczak, R.; Jarrell, M.; Meller, J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. *Bioinformatics,* **2006**, *22*(3), 303-309.
[133] Ganapathiraju, M.; Jursa, C.J.; Karimi, H.A.; Klein-Seetharaman, J. TMpro web server and web service: transmembrane helix prediction through amino acid property analysis. *Bioinformatics,* **2007**, *23*(20), 2795-2796.
[134] Osmanbeyoglu, H.; Wehner, J.A.; Carbonell, J.G.; Ganapathiraju, M.K. Active machine learning for transmembrane helix prediction. *BMC Bioinform.,* **2010**, *11*(Suppl 1), S58.
[135] Nugent, T.; Jones, D.T. Transmembrane protein topology prediction using support vector machines. *BMC Bioinform.,* **2009**, *10*, 159.
[136] Ahmed, R.; Rangwala, H.; Karypis, G. TOPTMH: topology predictor for transmembrane alpha-helices. *J. Bioinform. Comput. Biol.,* **2010**, *8*(1), 39-57.
[137] Kuang, R, Leslie, C.S.; Yang, A.-S. Protein backbone angle prediction with machine learning approaches. *Bioinformatics,* **2004**, *20*(10), 1612-1621.
[138] Xue, B.; Dor, O.; Faraggi, E.; Zhou, Y. Real-value prediction of backbone torsion angles. *Proteins,* **2008**, *72*(1), 427-433.
[139] Holbrook, S.R.; Muskal, S.M.; Kim, S.-H. Predicting surface exposure of amino acids from protein sequence. *Protein Eng.,* **1990**, *3*(8), 659-665.
[140] Thompson, M.; Goldstein, R.A. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins,* **1996**, *25*(1), 38-47.
[141] Pascarella, S.; De Persio, R.; Bossa, F.; Argos, P. Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins,* **1998**, *32*(2), 190-199.
[142] Richardson, C.; Barlow, D.J. The bottom line for prediction of residue solvent accessibility. *Protein Eng.,* **1999**, *12*(12), 1051-1054.
[143] Carugo, O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng.,* **2000**, *13*(9), 607-609.

[144] Ahmad, S.; Gromiha, M.M. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics,* **2002**, *18*(6), 819-824.
[145] Kim, H.; Park, H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins,* **2004**, *54*(3), 557-562.
[146] Yuan, Z.; Huang, B. Prediction of protein accessible surface areas by support vector regression. *Proteins,* **2004**, *57*(3), 558-564.
[147] Garg, A.; Kaur, H.; Raghava, G.P. Real value prediction of solvent accessibility in *Proteins,* using multiple sequence alignment and secondary structure. *Proteins,* **2005**, *61*(2), 318-324.
[148] Wang, J.-Y.; Lee, H.-M.; Ahmad, S. Prediction and evolutionary information analysis of *Proteins,* solvent accessibility using multiple linear regression. *Proteins,* **2005**, *61*, 481-491.
[149] Nguyen, M.; Rajapakse, J.C. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins,* **2006**, *63*(3), 542-550.
[150] Wang, J.; Lee, H.M.; Ahmad, S. SVM-Cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. *Proteins,* **2007**, *68*(1), 82-91.
[151] Chen, K.; Kurgan, M.; Kurgan, L. Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. *J. Biom. Sc. Eng.,* **2008**, *1*(1), 1-9.
[152] Petersen, B.; Petersen, T.N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.,* **2009**, *9*, 51.
[153] Song, J.; Tan, H.; Mahmood, K.; Law, RH.; Buckle, AM.; Webb, GI.; Akutsu, T.; Whisstock, JC. Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS One* **2009**, *4*(9), e7072.
[154] Fariselli, P.; Casadio, R.; RCNPRED: prediction of the residue co-ordination numbers in proteins. *Bioinformatics,* **2001**, *17*(2), 202-204.
[155] Pollastri, G.; Baldi, P.; Fariselli, P.; Casadio, R. Improved prediction of the number of residue contacts in Proteins, by recurrent neural networks. *Bioinformatics,* **2001**, *17*(Suppl 1), S234-242.
[156] Ishida, T.; Nakamura, S.; Shimizu, K. Potential for assessing quality of protein structure based on contact number prediction. *Proteins,* **2006**, *64*(4), 940-947.
[157] Yuan, Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinform.,* **2005**, *6*, 248.
[158] Kinjo, A.; Nishikawa, K. CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinform.,* **2006**, *7,* 401.
[159] Song, J.; Burrage, K. Predicting residue-wise contact orders in *Proteins,* by support vector regression. *BMC Bioinform.,* **2006**, *7*, 425.
[160] Rangwala, H.; Kauffman, C.; Karypis, G. svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinform.,* **2009**, *10*, 439.
[161] Radivojac, P.; Obradovic, Z.; Smith, DK.; Zhu, G.; Vucetic, S.; Brown, CJ.; Lawson, JD.; Dunker, AK. Protein flexibility and intrinsic disorder. *Protein Sci.,* **2004**, *13*(1), 71-80.
[162] Yuan, Z.; Bailey, T.L.; Teasdale, R.D. Prediction of protein B-factor profiles. *Proteins*, **2005**, *4*, 905-912.
[163] Schlessinger, A,; Rost, B.; Protein flexibility and rigidity predicted from sequence. *Proteins,* **2005**, *61*(1):115-26.
[164] Schlessinger, A.; Yachdav, G.; Rost, B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics,* **2006**, *22*(7), 891-893.
[165] Chen, P.; Wang, B.; Wong, H.S.; Huang, D.S. Prediction of protein B-factors using multi-class bounded SVM. *Protein Pept. Lett.,* **2007**, *14*(2), 185-190.
[166] Pan, X.; Shen, H.B. Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept. Lett.,* **2009**, *16*(12), 1447-1454.

[167] Cozzetto, D.; Kryshtafovych, A.; Fidelis, K.; Moult, J.; Rost, B.; Tramontano, A. Evaluation of template-based models in CASP8 with standard measures. *Proteins,* **2009**, *77*(Suppl 9), 18-28.
[168] Koh, I.; Eyrich, V.A.; Marti-Renom, M.A.; Przybylski, D.; Madhusudhan, M.S.; Eswar, N.; Graña, O.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.,* **2003**, *31*(13), 3311-3315.
[169] Eyrich, V.; Marti-Renom, M.A.; Przybylski, D.; Madhusudhan, MS.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics,* **2001**, *17*(12), 1242-1243.
[170] Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins,* **1999**, *34*(2), 220-223.
[171] Zhang, H.; Zhang, T.; Chen, K.; Kedarisetti, K.D.; Mizianty, M.J.; Bao, Q.; Stach, W.; Kurgan, L. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Brief Bioinform*. **2011**, doi: 10.1093/bib/bbq088 [Epub ahead of print].
[172] Mizianty, M.; Stach, W.; Chen, K.; Kedarisetti, K.D.; Miri Disfani, F.; Kurgan, L. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics,* **2010**, *26*(18), i489-96.
[173] Schlessinger, A.; Punta, M.; Yachdav, G.; Kajan, L.; Rost, B. Improved disorder prediction by combination of orthogonal approaches. *PLoS One* **2009**, *4*(2), e4433.
[174] McGuffin, L.J. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics,* **2008**, 24(16), 1798-1804.
[175] Schlessinger, A.; Liu, J.F.; Rost, B. Natively unstructured loops differ from other loops. *PLoS Comput. Biol.,* **2007**, *3*, 1335-1346.
[176] Schlessinger, A.; Punta, M.; Rost, B. Natively unstructured regions in proteins, identified from contact predictions. *Bioinformatics,* **2007**, *23*, 2376-2384.
[177] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: web server for the prediction of intrinsically un-structured regions of *Proteins,* based on estimated energy content. *Bioinformatics,* **2005**, *21*(16), 3433-3434.
[178] Ward, J.; McGuffin, L.J.; Bryson, K.; Buxton, B.F.; Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics,* **2004**, *20*, 2138-2139.
[179] Cheng, J.; Sweredoski, M.; Baldi, P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowl. Disc.* **2005**, *11*, 213-222.
[180] Hecker, J.; Yang, J.; Cheng, J. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics,* **2008**, *9*(S1), S9.
[181] Mizianty, M.; Zhang, T.; Xue, B.; Zhou, Y.; Dunker, A.K.; Uversky, V.N.; Kurgan, L. In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinform.,* **2011**, *12*, 245.
[182] Rose, G.D.; Dworkin, J.E. The hydrophobicity profile. Fasman, G.D., Ed.; Prediction of protein structure and the principles of protein conformation. Springer, **1989**, pp. 625-634