Experimental Evaluation of Protein Secondary Structure Predictors

Luca Miceli¹, Luigi Palopoli¹, Simona E. Rombo¹, Giorgio Terracina²,
Giuseppe Tradigo³, and Pierangelo Veltri³

¹ DEIS, Università della Calabria

² Dip. di Mat., Università della Calabria

³ Dip. Med. Sp. e Cl., Univ. di Catanzaro

Abstract. Understanding protein biological function is a key issue in modern biology, which is largely determined by its 3D shape. Protein 3D shape, in its turn, is functionally implied by its amino acid sequence. Since the direct inspection of such 3D structures is rather expensive and time consuming, a number of software techniques have been developed in the last few years that predict a spatial model, either of the secondary or of the tertiary form, for a given target protein starting from its amino acid sequence.

This paper offers a comparison of several available automatic secondary structure prediction tools. The comparison is of the experimental kind, where two relevant sets of proteins, a non-redundant one including 100 elements, and a 180-protein set taken from the CASP 6 contest, were used as test cases. Comparisons have been based on evaluating standard quality measures, such as the Q3 and SOV.

1 Introduction

Proteins are the basic constituents of living beings. They form the basis for structural components of cells as well as for metabolic processes involved in organic life. Understanding protein functions has, therefore, a central role in the analysis of the biological mechanisms underlying life processes. A protein biological function is largely determined by its 3D shape [15], which is functionally implied, in its turn, by the sequence of amino acids that form the protein [3]. The amino acidic sequence of a protein is called its *primary structure*, whereas its 3D shape is encoded in two different representations, namely, its *secondary* and its *tertiary* structure [15]. To illustrate, the tertiary structure of a protein tells, with respect to a given 3D fixed axis origin point, the exact positions of protein constituent atoms. The secondary structure, instead, provides information about the composition of the protein structure in terms of regular substructures. In fact, amino acids tend to dispose themselves within some few substructures, namely, sheets consisting of β -strands (denoted E in the following) laterally connected to form a pleated sheet, and α -helices (denoted H in the following). Moreover, amino acids might contribute to form kinds of irregular structures, which link regular ones to one another, and which are usually referred to as *loops* (denoted L in the sequel). Thus, a secondary protein structure is denoted by a sequence of letters over the alphabet E, H, L , one letter for each of the amino acids occurring in the primary structure of the protein. These letters are called the *conformational states* of the amino acids.

All that given, the relevance of associating secondary and tertiary structure with protein amino acid sequences is immediately understood. To do that, complex lab methods (that are, X-ray crystallography and nucleic magnetic resonance spectroscopy (NMR)) can be employed, which are however expensive and time-consuming. As a result, protein sequence discovering rate is much higher than protein structure identification rate and, therefore, while some millions of protein primary structures are known to date, only a few tens of thousands of secondary and tertiary structures have been discovered [2]. This is the reason why a large interest has been witnessed in the community towards computational methods for *predicting* [10,25,26,28] such secondary and tertiary structures [4,5,13,16,17,21,22,29]. Prediction methods, though fast and cheap to run, do not achieve to date the same accuracy as lab methods in reconstructing protein structures.

This paper is concerned with comparing computational methods for protein *secondary* structure prediction. Therefore, in the following, we will make no further reference to tertiary structures. In this setting, this paper offers an experimental analysis of several available protein secondary structure predictors. The analysis has been conducted by using two relevant data sets (a non-redundant and a CASP protein set), which will be described below and overall including 280 proteins. The experimental evaluation has been carried out by querying a set of available prediction tools over proteins having known secondary structure, in order to evaluate their accuracy according to quality evaluation parameters, namely, the Q3 and the SOV, which are commonly adopted in the literature [23,27]. Both the average behavior of each predictor on the whole data sets and its specific accuracy for each protein are reported and analyzed so as to result in a comparison of the (relative) performances of the considered prediction tools.

It is worth mentioning here that monitoring tools like those published by EVA server [7] can be referred to in order to verify the quality of results delivered by prediction tools made available on-line, providing the evaluations *Q3* and *SOV* for secondary structure prediction. In particular, at for today, the average values for such parameters are 75.9% for *Q3* and 72.7% for *SOV*, according to data reported at http://cubic.bioc.columbia.edu/eva/sec/res_sec.html, where 100% would be the value scored by the *perfect* predictor for both parameters. International challenges, such as the biannual CASP competition [30], have been instituted to encourage studying and designing high-quality synthetic predictors, even if CASP focuses on 3D structure prediction tools.

The work presented in this paper differs in both objectives and contributions from those either resulting from the CASP competition or obtained using the EVA server, for the following reasons. First, we notice that CASP is intended to evaluate the impact of current prediction methods and techniques in helping experts to design 3D structure predictions; in fact, groups competing in CASP are not constrained to exploit automatic tools only; rather, they can refine “by hand” resulting predictions with the guide of current human expertise in the field. As a consequence, CASP is meant to determine a measure of human capabilities in predicting protein structures with the aid of available automatic tools, rather than on evaluating prediction tools “per se”. So, CASP is, in a sense, *expert-oriented*. On the contrary, the purpose of EVA is to determine what confidence a biologist should rely on a given specific tool “as they are” to be a good predictor, by monitoring the quality of its predictions on a wide variety of proteins and a large time range. Therefore, EVA is intended to evaluate the performances of the

single tools, where different protein data sets may be used in evaluating a tool, with no reference whatsoever to those which were used to evaluate *other* tools.

Therefore, neither CASP, nor EVA address direct experimental cross-comparison between prediction tools, which is the focus of this paper, inasmuch as the objective of our work is to provide a thorough comparison of tool performances on protein data sets specifically selected to be both statistically and biologically relevant.

The rest of the paper is organized as follows. The next section presents a very brief overview of main tool categories, whereas Section 3 introduces data sets and prediction tools we exploited for our experimental evaluations, illustrates the experimental results, and discusses about predictor performances as resulting from the experiments is reported. Finally, in Section 4, some conclusions are drawn.

2 Secondary Structure Prediction Methods

As already stated, secondary structure prediction consists in associating a string of characters representing amino acids conformational states to the string representing the primary structure of a protein sequence, whose structure is not determined experimentally. The correctness of the prediction can only be determined by comparing the predicted string with the secondary structure obtained from lab methods (called its *observed* structure), once it has been singled out [26]. It is largely accepted that evaluation of quality of a prediction is done by using *Q3* and *SOV* parameters, where *Q3* punctually measures the percentage of correctly guessed structures for the target protein, i.e., the conformational state of single amino acids, whereas *SOV* is obtained by computing per-segment overlaps [7].

Protein secondary structure prediction methods (as well as tertiary ones) can be classified as either *ab initio* methods or *evolution-based* methods (aka, *homology-based* methods). In the first case protein structure prediction is based on evaluating the minimal free energy [6]. Indeed the three dimensional conformation of a protein is defined by the spatial conformation of amino acids chain that correspond to that structure featuring the lowest free energy determined by the mutual interactions of amino acids. This is the idea underlying the development of *ab initio* methods. However, an exhaustive search of all possible configurations of a polypeptide chain is such a formidable task that *ab initio* methods usually produce satisfactory results just in case of chains with a low number of amino acids, in which cases, it is actually feasible to evaluate the energy configuration of the folding process rather precisely. Simulating algorithms (such as Monte Carlo methods) are useful in the case of proteins with a low number of amino acids, or in the case of prediction processes guided by information used to select (and, thus, reduce) the number of possible configurations to test for good. As an example, the Rosetta [24] predictor uses a Monte Carlo algorithm to reduce the possible combinations of amino acids while predicting single regions.

Counter-wisely, *evolution-based* methods look at the target protein's primary structure, comparing this protein sequence with known ones published in available databases and trying to exploit evolutionary protein relationships (e.g., with protein family identification). Then, comparative modeling algorithms are used to build a spatial configuration for the unknown structure. Given a set of proteins that are evolutionary related

to the target one, comparative based algorithms construct a multi alignment and deduce correspondences between amino acids, possibly identify regions affected by insertion, deletion and modification caused by evolution, and build a structure corresponding to amino acids that did not change. Finally the outside region is designed and the final structure is optimized. Such methods strictly rely on how close is the evolutionary relation among the target protein and a set of known proteins. Sequence alignments may be performed by using algorithms of global alignment, such as the Needleman and Wunsch [19], or by using local alignment algorithms such as BLAST [1]. Multiple alignments can be performed using Clustal and TCOFFEE algorithms [11,31].

In the next section, the main methodologies on which each of the considered tools is based are also specified.

3 Experiments on Protein Prediction

In this section we present the experimental framework we adopted to compare the performance of nine secondary structure prediction tools, that are: Jufo [18], Prof [20], Porter [21], Psipred [17], Nn-predict [14], all exploiting neural network-based approaches; HMMSTR/Rosetta [4], based on an *ab initio* method; SAM [13], based on linear hidden Markov models; Gor IV [8], using frequency analysis of amino acid conformational states; Hnn [9], based on a hierarchical and modular approach.

To begin with, we describe the data sets we have exploited in experiments and, then, we illustrate in detail the experimental tests we carried out using the aforementioned prediction tools.

3.1 Data Sets

In order to carry out a robust analysis of the performances of the analyzed tools, we considered two protein data sets.

The former data set includes 100 proteins from the *PDB25Select* database [12], a set purposely constructed to be non-redundant. Proteins in this data set are in fact characterized by a sequence identity which is less than 25%; this feature eliminates, in practice, the possible bias determined by testing homologues proteins, which could unfairly give an advantage to those techniques better working on that kind of proteins. This clearly makes the performance analysis more statistically relevant. This data set, which we will call PDB25 in the following, is shown in Figure 1(a), where proteins are identified by their PDB id [2].

The second data set we considered includes 180 proteins from the CASP 6 edition [30] and available at <http://predictioncenter.org/>. Structures of proteins selected for CASP represent both a non-trivial and a biologically relevant test-bed for protein structure prediction tools. This data set, called CASP in the following, is reported in Figure 1(b) where, again, each protein is identified by its PDB id.

3.2 Quality Parameters

In order to test the accuracy of analyzed prediction tools, we exploited the well known *SOV* and *Q3* parameters described in [23,27]. We recall that *Q3* represents the

PDB25 Proteins									
1AAF	1ARB	1BBPA	1CAUA	1CPCL	1FHA	1GPS	1HLE	1LTS	1PAZ
1PPFE	1TNFA	1XIMA	2BPA3	2GBIA	2MADL	2POR	3RUBS	4ICD	7APIB
9WGAA	1BBT1	1CAUB	1EGF	1FIAB	1GRCA	1HSBA	1LTSC	1MUP	1PPN
1RHD	1SNC	1YCC	2CBH	2GLSA	2MEV1	2RN2	3CD4	3SC2A	4INSB
7ZNF	2AAK	1ATX	1BBT2	1CBN	2END	1FNB	2HSDA	1LTS	1NIPB
1PDA	1PRCC	1RND	1TABI	1TRB	256BA	2CCYA	2HAD	2MEV4	2SCPA
3CHY	3SC2B	4RCRH	8ABP	1AAPA	1AVHA	1BGC	1FXIA	1MAMH	1OIAA
1PRCM	1RPRA	1TEN	1TREA	2AAA	2CDV	2HHMA	1ABA	1CD8	1D66A
1ERP	1GSSA	1IFC	1MDAA	1RRO	1TFG	1TROA	2ACHA	2CPL	2MHU
3DFR	2SN3	3SICI	4SGBI	8ADH	1BOVA	1CDTA	1DFNA	1EZM	1GLAF

(a)

CASP Proteins									
1STZA	1SUWA	1T70A	1S12A	1SUMB	1TD6A	1T6SA	2BH8A	1VDHA	1VGGA
1WD7A	1WD5A	1WD6A	1WDEA	1WDIA	1WDJA	1WDVA	1WFEX	1WHZA	1WTYA
1WJ9A	1WJGA	1WK4A	1WKCA	1VLCA	1BNKA	1HKA	1XFKA	1WG8A	1GA7A
1G6EA	1B9KA	2A2UA	1L7CA	1VL7A	1VLOA	1J3GA	1VQUA	1XCBA	1VM0A
1VKPA	1ZWJA	2FNOA	2J85A	1E68A	1W33A	1KRHA	1EXSA	1M33A	1TROA
1G7DA	1IZNA	1QGOA	1GGWA	1L7AA	1FW9A	1N2XA	1DW9A	1BVB	2EZM
1DPTA	1J90A	1MZHA	1KOYA	1JWEA	1G6UA	1C8UA	1TZ9A	1FZRA	1E91A
1C18A	1BQV	1E2XA	1I8UA	1UZCA	1J83A	1QFJA	1E4FT	1Y0BA	1XQAA
1Y12A	2CV4A	1VKKA	1CCWA	1MKIA	1B7GO	1NG4A	1EUGA	1BG8A	1NO5A
1IM8A	1IZMA	1YLIA	1INO	1DBUA	1MW5A	1FWKA	1TVGA	1M3SA	1PV1A
1TE7A	1S04A	1FL9A	1MW7A	1YK3A	1N91A	1NYNA	1NKVA	1XQBA	1M2FA
1E3JA	1H7MA	1G8PA	1G29I	1BL0A	1B93A	1EWQA	1Q74A	1VLP	1NTFA
1XG8A	1XV2A	1F35A	1E54A	1RKIA	1MOPA	1GA6A	1QJVA	1YEMA	1XE1A
1BD9A	1JADA	1BYFA	1BHE	1I74A	1NBUA	1NXJA	1R9QA	1W81A	1BK7A
1C94A	1QMHA	1IJXA	1NVT	1UE6A	1E6WA	1B0NA	1B0NB	1D3BA	1YUDA
1TZAA	1H5PA	1JAKA	1IY9A	1FC3A	1VLIA	1X9BA	1EH2	1J6UA	1UWDA
1VL4A	1O12A	2AJRA	1VLAA	1X9AA	1O0WA	1O0XA	1O0YA	2B8NA	1VKWA
1VPQA	1O13A	1U07A	1ON9A	1BKB	1GEQA	1WIWA	1WK2A	1VJVA	1QY6A

(b)

Fig. 1. The exploited protein data sets: (a) The PDB25 data set. (b) The CASP data set.

percentage of amino acids correctly predicted by a prediction tool t for a given input protein p . Conversely, the SOV parameter represents the percentage of *segments* correctly predicted by the prediction tool t for the protein p , where a segment is a portion of a secondary structure made exclusively of the same conformational state (e.g., of α -helices, or of β -strands, or of loops) and satisfies some other structural constraints. The interested reader is referred to the cited references for a more formal definition of $Q3$ and SOV .

$Q3$ and SOV measure two quite different (and often contrasting) characteristics of the predictions. Consequently, neither $Q3$ nor SOV alone are suitable to measure the overall accuracy of a tool. Thus, in order to embed into a single parameter information provided by both $Q3$ and SOV , we considered a further parameter, that we call *Prediction Accuracy*, defined as:

$$PA = \frac{2}{3} \cdot SOV + \frac{1}{3} \cdot Q3$$

which is simply obtained as a weighted mean of SOV and $Q3$; more precisely, since SOV takes into more account structural kind of information which $Q3$ does not consider, it appears sensible to weigh the SOV more than the $Q3$. Such weights have been also tested experimentally.

Tests have been carried out as follows. Each analyzed prediction tool has been required to predict the secondary structure of each considered protein. Then, each prediction has been compared with the real, known, structure to compute $Q3$, SOV and, consequently, the PA .

We next provide a description of the measures recorded in our tests; a discussion of the corresponding results is addressed in the next section.

3.3 Results

Table 1 summarizes the results we obtained by comparing different prediction tools. In particular, it shows the *average PA* scored by each prediction tool as well as the corresponding standard deviation on (i) the PDB25 data set, (ii) the CASP data set, (iii) all proteins included in either data sets.

Table 2 considers $Q3$ and SOV separately; in particular, it shows the average $Q3$ (resp., SOV) scored by each tool on (i) the PDB25 data set, (ii) the CASP data set, (iii) all proteins.

Finally, in Figure 2 some graphs are presented showing the percentage of test cases in which each predictor resulted in the set of the best ones (again, distinguishing between PDB25 and CASP data sets). The set of the best predictors results by simply considering the best performances up to a given tolerance threshold. In particular, since the differences in the PA s of the various tools are often very small we considered different levels of “tolerance” in determining the set of best predictors.

To illustrate, let p_i be an input protein, let t_j be a tool, PA_{ij} the PA of the tool t_j on protein p_i and PA_i^* the best value of PA obtained for p_i . In order to select the set of best predictors on each p_i we considered the following formula:

$$bestTools(p_i) = \{t_j \mid PA_{ij} \geq (1 - \tau)PA_i^*\}$$

where $\tau \in [0, 1]$ is the “tolerance” factor, stating the maximum tolerated deviation from the best PA . We computed the best sets of tools for each protein using the following three levels of tolerance: (i) $\tau = 0.00$, which selects only those predictors scoring exactly the best PA ; the results for this case are shown in Figures 2(a) and 2(b). (ii) $\tau = 0.05$, which considers those predictors scoring a PA within 5% of the best PA ; the results for this case are shown in Figures 2(c) and 2(d). (iii) $\tau = 0.15$, which considers those predictors scoring a PA within 15% of the best PA ; the results for this case are shown in Figures 2(e) and 2(f).

3.4 Discussion

From an overall analysis of the results we obtained from experiments, we note that the PDB25 data set actually turned out to be a more demanding test-bed than the CASP set. Indeed, the average PA ’s measured on the CASP proteins is larger than the PA scored on PDB25 proteins for all tools except Gor IV and HMMSTR/Rosetta. This result is quite interesting because it points out that possible homologies between proteins can simplify the structure prediction process not only for those tools based on comparative modeling, but, actually, for most of them.

Table 1. Average PA and corresponding standard deviation for each prediction tool

Tool	Technique	PDB25 Proteins		CASP Proteins		All Proteins	
		Avg PA	st.dev.	Avg PA	st.dev.	Avg PA	st.dev.
Jufo	neural networks	70.10	13.49	74.39	10.49	72.86	11.83
Prof	neural networks	71.56	14.45	73.14	9.71	72.58	11.65
Porter	neural networks	87.04	10.69	92.10	6.85	90.29	8.77
Pspired	neural networks	77.64	11.95	80.63	9.23	79.56	10.38
Nn-predict	neural networks	55.11	11.26	57.76	10.08	56.82	10.59
HMMSTR/Rosetta	ab initio	73.43	14.07	72.84	14.01	73.05	14.03
Sam	linear hidden Markov models	75.20	13.131	79.06	9.34	77.68	11.00
Gor IV	freq. analysis of amino acid conf. states	61.21	11.96	60.81	10.11	60.95	10.81
Hnn	hierarch. and modular approach	62.46	11.74	64.25	10.68	63.61	11.10

Fig. 2. Percentage of times each prediction tool has been detected among the best ones at various tolerance levels

Table 2. Average $Q3$ and SOV scored by each prediction tool

Tool	PDB25		CASP		All	
	AvgQ3	AvgSOV3	AvgQ3	AvgSOV3	AvgQ3	AvgSOV3
Jufo	74.67	67.64	75.34	73.31	75.16	71.49
Prof	75.88	69.34	75.24	72.06	75.50	71.19
Porter	90.34	85.33	93.19	91.54	92.20	89.40
Psipred	81.21	76.03	82.41	79.71	82.05	78.44
Nn-predict	62.53	51.34	62.33	55.41	62.40	54.00
HMMSTR/Rosetta	79.36	70.56	75.47	71.49	76.82	71.17
Sam	79.25	73.11	80.56	78.29	80.09	76.44
Gor IV	65.87	58.93	62.83	59.77	63.92	59.43
Hnn	66.82	60.45	67.25	62.71	67.08	61.83

Second, the average performance of the analyzed tools appears notable, especially when compared to the complexity of the exploited data sets. To illustrate, while the average PA recorded for the top scoring tools in our tests agrees with the results reported in http://cubic.bioc.columbia.edu/eva/sec/res_sec.html, it turns out that significant improvements can still be achieved in prediction accuracy, as long as the quality attained by the tools on a protein-to-protein basis is still uncertain for almost all of them, which is well mirrored in the values of the standard deviation of the PA we recorded in our tests (see Table 1). This demonstrates that the *guarantee* for a sufficiently high precision of the prediction still remains rather low, in general, which eventually implies that the predictions yielded by those tools, even if quite good on the average, can hardly be looked at as a reliable reference by biologists.

A further interesting observation that can be drawn from Table 1 is that our results do not allow the identification of a *prominent technique* for protein secondary structure prediction, even if recent trends seem to pay much attention on neural network-based approaches. As for neural network-based techniques, in particular, on the one hand, they showed rather varying results in terms of PA (one of them, namely Porter, appears as the best predictor in the analyzed set, whereas another one of them, namely Nn-predict, scored the lowest average PA) while, on the other hand, tools based on different techniques scored competitive PA s. This, again, suggests that there is still space for improvements in protein structure prediction accuracy.

The analysis of Table 2 confirms our former claim that neither $Q3$ nor SOV alone can provide a reliable measure of the tools prediction accuracy. Indeed, the analysis of this table points out that there are cases of pairs tools (e.g., Sam and HMMSTR/Rosetta), scoring, respectively, the best $Q3$ and and the best SOV , making it difficult to figure out which one attains the best quality prediction. In this respect, our choice of defining the PA as a composed quality parameter englobing both the $Q3$ and the SOV , seems to be a sensible one.

Finally, as for the percentage of proteins upon which a predictor can be counted among the best ones, we can observe from Figure 2 that the general trend of the various tools is actually independent of the tolerance degree. Clearly, the higher the tolerance is, the higher the percentage associated with each tool will be. The interesting observation is that the *overall* ranking of the various tools is almost invariant with respect to the tolerance degree. In more detail, these graphs show that, on the employed data sets, Porter was the best performing tool both in terms of average PA and, notably, also in terms of number of successes.

4 Conclusion

In this paper we have illustrated an experimental campaign we have developed in order to compare the quality of predictions returned by protein secondary structure predictors. Analyzed predictors include the systems Jufo [18], Prof [20], Porter [21], Psipred [17], Nn-predict [14], HMMSTR/Rosetta [4], SAM [13], Gor IV [8], Hnn [9].

Two test data sets were selected: one non-redundant set (used to minimize the influence of homology amongst tested proteins on returned predictions) and a second set of proteins taken from those used in CASP, the well-known biannual protein prediction contest.

The experimental results showed that some of the available predictors are able to score quite good average accuracies, but the quest for a truly reliable automatic tool is still open, as the predictions are still characterized by significant standard deviations (which means that, for a given generic protein, the confidence for the returned prediction to have a high accuracy is largely unpredictable).

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17), 3389–3402 (1997)
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. *Nucleic Acids Research* 28, 235–242 (2000)
3. Branden, C.I., Tooze, J.: *Introduction to Protein Structure*. Garland Publishing, Inc. (1998)
4. Bystroff, C., Shao, Y.: Fully automatic ab initio protein structure prediction using I-SITES, HMMSTR and HMMSTR/ROSETTA. *Bioinformatics* 18(suppl. 1), S54–S61 (2002)
5. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., Barton, G.J.: Jpred: a consensus secondary structure prediction server. *Bioinformatics* 14, 892–893 (1998)
6. Dudek, M., Ramnarayan, K., Ponder, J.W.: Protein structure prediction using a combination of sequence homology and global energy minimization. *Journal of Computational Chemistry* 2(19), 548–573 (1998)
7. Eyrich, V.A., Martin-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A., Rost, B.: Eva: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* 17(12), 1242–1243 (2001)
8. Garnier, J., Gibrat, J.F., Robson, B.: GOR secondary structure prediction method version IV. *Methods in Enzymology* 266, 540–553 (1996)
9. Guermeur, Y.: *Combinaison de classifieurs statistiques, Application a la prediction de structure secondaire des proteines.*, PhD Thesis (1997)
10. Guerra, C., Istrail, S.: *Mathematical Methods for Protein Structure Analysis and Design: Advanced Lectures*, August 13, 2003. *Lecture Notes in Computer Science / Lecture Notes in Bioinformatics*. Springer, Heidelberg (2003)
11. Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressively multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680 (1994)
12. Hobohm, U., Sander, C.: Enlarged representative set of protein structures. *Protein Science* 3, 522–524 (1994)

13. Hughey, R., Krogh, A.: SAM: Sequence alignment and modeling software system., Technical Report UCSC-CRL-95-7, University of California, Santa Cruz, CA (1995)
14. Kneller, D.G., Cohen, F.E., Langridge, R.: Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology* 214, 171–182 (1990)
15. Lesk, A.M.: *Introduction to Protein Architecture*. Oxford University Press (2001)
16. Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J.: A Simple and Fast Secondary Structure Prediction Algorithm using Hidden Neural Networks. *Bioinformatics* 21(2), 152–159 (2005)
17. McGuffin, L.J., Bryson, K., Jones, D.T.: The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405 (2000)
18. Meiler, J., Mueller, M., Zeidler, A., Schmaeschke, F.: JUFO: Secondary Structure Prediction for Proteins (2002), <http://www.jens-meiler.de>
19. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 442–453 (1970)
20. Ouali, M., King, R.D.: Cascaded multiple classifiers for secondary structure prediction. *Protein Science* 9, 1162–1176 (2000)
21. Pollastri, G., McLysaght, A.: Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21(8), 1719–1720 (2005)
22. Rost, B., Sander, C.: Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 232, 584–599 (1993)
23. Rost, B., Sander, C., Schneider, R.: Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology* 235, 13–26 (1994)
24. Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D.: Ab Initio Protein Structure Prediction of CASP III targets using Rosetta. *Proteins* (suppl. 3), 171–176 (1999)
25. Tramontano, A.: *The Ten Most Wanted Solutions in Protein Bioinformatics*. CRC Press, Boca Raton (2005)
26. Tramontano, A., Lesk, A.M.: *Protein Structure Prediction: Concepts and Applications*. John Wiley & Sons, Chichester (2006)
27. Venclovas, C., Zemla, A., Fidelis, K., Moult, J.: Some measures of comparative performance in the three CASPs. *Proteins: Structure, Function, and Genetics* 34, 220–223 (1999)
28. Webster, D.M.: *Protein Structure Prediction: Methods and Protocols* (Methods in Molecular Biology). Humana Press, August 15 (2000)
29. White, J.V., Stultz, C.M., Smith, T.F.: Protein Classification by Stochastic Modeling and Optimal Filtering of Amino-Acid Sequences. *Mathematical Biosciences* 119, 35–75 (1994)
30. Critical assessment of techniques for protein structure prediction, <http://predictioncenter.llnl.gov/casp6/Casp6.html>
31. TCOFFEE Multiple Alignments tools. Swiss institute of bioinformatics, <http://www.tcoffee.org/>