Protein J (2008) 27:234–239
DOI 10.1007/s10930-008-9129-0

On the Relation Between the Predicted Secondary Structure and the Protein Size

Lukasz Kurgan

Published online: 26 February 2008

Abstract Accurately predicted protein secondary structure provides useful information for target selection, to analyze protein function and to predict higher dimensional structure. Existing research shows that *more data + refined search = better prediction*. We analyze relation between the prediction accuracy and another crucial factor, the protein size. Empirical tests performed with two secondary structure predictors on a large set of high-resolution, non-redundant proteins show that the average accuracies for small proteins (<100 residues) equal 73% and 54% for α-helices and β-strands, respectively. The α-helix/β-strand accuracies for very large proteins (>300 residues) equal 77%/68%, respectively. Similarly, the tests with three secondary structure content predictors show that the prediction errors for the small/very large proteins equal 0.13/0.09 and 0.09/0.06 for α-helix and β-strand content, respectively. Our tests confirm that the secondary structure/content predictions for the very large proteins are characterized statistically significantly better quality than prediction for the small proteins. This is in contrast with the tertiary structure predictions in which higher accuracy is obtained for smaller proteins.

Keywords Secondary protein structure · Secondary protein structure content · Protein size · PSI-PRED · PSSC-core

L. Kurgan (✉)
Department of Electrical and Computer Engineering, University of Alberta, 2nd floor, ECERF (9107 116 Street), Edmonton, AB, Canada T6G 2V4
e-mail: lkurgan@ece.ualberta.ca

1 Introduction

The secondary protein structure was postulated over 50 years ago by Pauling and Corey, who predicted the existence of two local periodic motifs: the α-helix [33] and the β-sheet [32]. The secondary structure is widely used in a number of structural biology applications, such as structure comparison [11], classification [27, 30], and visualization [17, 42]. It can also be used to successfully identify family, superfamily, and tertiary fold of the underlying protein [12]. Experimental methods for determination of the secondary structure depend on the experimentally derived tertiary structure. The most popular secondary structure assignment method, Dictionary of Protein Secondary Structure (DSSP[1]), was developed in early 1980's [19]. It defines eight types of secondary structures that are combined into three basic secondary structure states: α-helix, β-strand, and coil. Protein structures deposited in the Protein Data Bank (PDB) [2] contain secondary structure description that is either provided by the depositor (optional) or, in most cases, generated by DSSP. We note that assignment of the secondary structures performed by DSSP and other assignment methods [26] is based on the atomic coordinates, i.e., the tertiary structure.

In contrast to the experimental methods, computational methods for prediction of the secondary structure use only the protein sequence. Availability of accurately predicted secondary structure is crucial for target selection in structural genomics to obtain clues about protein function and for predictions of higher dimensional aspects of protein structure [39]. More specifically, the predicted secondary structure and secondary structure content, which is defined

[1] List of abbreviations: Dictionary of Protein Secondary Structure (DSSP), amino acid (AA), Protein Data Bank (PDB).

as the percentage amount of α-helices and β-strands in a sequence, are used in prediction of tertiary structure [4, 5, 7, 25, 31, 40, 41], protein fold [10, 43], pi-turns [44], protein topology [13], and structural class [21], as well as in reduction of the complexity of molecular dynamics simulations [6], characterization of protein domains [28], structural analysis of individual proteins [29], and identification of putative active sites [1].

Current secondary structure prediction algorithms provide accuracy of about 80% for the three state predictions [18, 34–36]. In 2001, Rost proposed the following formula: *more data + refined search = better prediction* [38]. To be more specific, research shows that accuracy is affected by the following factors:

- the size of the dataset used to derive the prediction model [8, 37, 38]
- quality of the underlying sequence alignment method [9, 14, 37]
- protein family size [37],
- the quality of the used classification algorithm [38, 39]

We perform a systematic analysis of another, currently unexplored and important factor that impacts quality of the secondary structure prediction, namely the size of the protein. We show that the quality of the experimentally derived tertiary structure (and consequently the secondary structure) does not depend on the protein size, while the quality of the predicted secondary structure and secondary structure content strongly depends on the protein size. This work is unique in two aspects: (1) we consider both secondary structure and secondary structure content predictions, and (2) we show that the discovered relation holds true for several representative prediction methods.

2 Materials and Methods

2.1 Experimentally Derived Structure

We extracted 48,323 protein sequences (including multiple chains) that were deposited in PDB as of April 12, 2007. Among them, 39,884 protein chains for which the sequence includes at least 30 AAs and the experimental resolution of the structure is known, were kept. Figure 1 shows relation between the size of a protein (expressed as the sequence length) and the average resolution of the corresponding structures measured in Å. The Figure shows that the resolution (and consequently the quality) of the experimental structures does not depend on the size of the protein, i.e., the average resolution ranges between 2.1 and 2.4 Å over different sizes of proteins. This implies that the secondary structure assigned by DSSP based on these tertiary

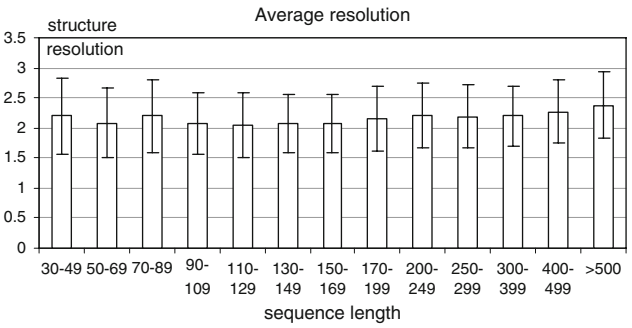

Fig. 1 The relation between the average resolution and sequence length for protein structures deposited in PDB. The error bars denote ±standard deviation of the resolutions for the corresponding intervals of the sequence length

structures is also characterized by a relatively similar, with respect to the sequence length, quality.

2.2 Computationally Predicted Structure

The relation between the quality of the computationally predicted secondary structure and the underling sequence length (protein size) is studied based on two mainstream types of the related prediction methods:

1. *Secondary structure* predictors, which predict the secondary structure from a protein sequence. The two representative methods which use different underlying prediction architectures are PSI-PRED [18] and YASPIN [24]. The PSI-PRED method was recently found to provide superior prediction quality [3] and is used for tertiary structure prediction performed by the ROSETTA method [5], while YASPIN method was shown to provide high quality predictions for β-strands [24].
2. *Secondary structure content* predictors, which predict the amount of α-helices and β-strands from a protein sequence and without providing information about the location of these secondary structures. The three representative methods include method by Zhang and colleagues [45] (referred to as ZSZ01), by Lin and Pan [23] (referred to as LP01), and PSSC-core [16]. The main motivation to select these three methods is that they provide superior prediction quality when compared with other content prediction methods [16].

2.3 Dataset and Experimental Setup

The predictions were performed with a large benchmark dataset of high-resolution, low homology proteins published in [22]. The dataset, named *25PDB*, was originally built based on the 25% PDBSELECT list [15], and includes 1,673 proteins scanned with at least 3 Å

resolution, characterized by average identity of 25%, and which include at least 30 residues. These assumptions allow removing bias due to sequence identity and assuring that the DSSP structures derived for the included proteins, which are used to validate the predicted secondary structure, are of high quality. The prediction quality was evaluated using standard measures. In case of the secondary structure prediction we computed accuracy of prediction for α-helix (Q_H), β-strand (Q_E), and coil (Q_C) structures. The accuracy is defined as the numbers of residues that were correctly predicted to form a given secondary structure divided by the total number of residues in that secondary structure. The corresponding Q_H, Q_E, and Q_C values were first computed for each sequence, and next these values were used to compute average accuracies for proteins grouped by their size. To assure that the results for the secondary structure prediction are fair, the accuracy of zero was assumed only when a given secondary structure was present in the sequence but it was not correctly predicted; in case when a given secondary structure was not present in the sequence, the corresponding accuracy was ignored. For the content prediction we computed mean absolute error between the predicted and the actual amount of α-helix (e_H) and β-strand (e_E) structures, respectively. Similarly as for the secondary structure, the content prediction errors were computed for each sequence, and next these values were averaged for proteins grouped by their size.

To facilitate our analysis of the impact of the protein size on the accuracy (error rate) of the predicted secondary structure (content), we divided the dataset into four subsets: *very small proteins* with 50 or less AAs, *small proteins* with 100 or less AAs, *large proteins* with chains that contain between 100 and 300 residues, and *very large proteins* that include 300 or more AAs.

3 Results

3.1 Secondary Structure Prediction

Table 1 shows average accuracy for helix (Q_H), strand (Q_E) and coil (Q_C) prediction for the PSI-PRED and YASPIN secondary structure prediction methods grouped by protein sizes, which are expressed based on the corresponding sequence length. The results are consistent with the overall quality of the two prediction methods, i.e., PSI-PRED is characterized by higher accuracy for α-helix and coil predictions, i.e., 77 and 75.4%, while YASPIN gives better predictions for β-strands, i.e., 67.5%. Most importantly, the accuracies of α-helix and β-strands predictions vary with the protein sizes. In both cases the predictions for small proteins suffer lower accuracies while predictions for large and very large proteins are characterized by higher accuracies. Overall for both PSI-PRED and YASPIN, the average accuracy for the small proteins equals 73.2% and 54.3% for the α-helix and β-strand predictions, respectively. Results for the very small proteins are characterized by even smaller average accuracies, i.e., 63.5% and 38.0% for the α-helix and β-strands predictions, respectively. In contrast, for the large proteins the average accuracies equal 75.3% and 69.9% for the α-helix and β-strand predictions, respectively, while for the very large proteins the average accuracies equal 76.9% and 68.5%, respectively. Finally, the prediction accuracies for the coils do not depend on the underlying sequence length, see Table 1.

Table 1 Average, over proteins in a given size range, prediction accuracy for the PSI-PRED and YASPIN secondary structure prediction methods in the function of protein size

Protein size	Accuracy for α-helix (Q_H)		Accuracy for β-strand (Q_E)		Accuracy for coil (Q_C)	
	PSI-PRED	YASPIN	PSI-PRED	YASPIN	PSI-PRED	YASPIN
[30–49]	66.2	60.8	32.0	43.9	75.9	68.5
[50–69]	74.2	73.6	47.3	57.8	77.3	71.1
[70–89]	79.1	75.8	55.9	61.4	74.7	68.9
[90–109]	79.4	76.5	66.1	67.8	75.4	68.7
[110–129]	73.9	71.4	68.2	72.1	76.0	70.0
[130–149]	79.2	74.3	67.2	71.8	75.5	70.2
[150–169]	77.9	71.5	65.2	70.7	73.5	69.8
[170–199]	75.8	71.9	66.2	72.0	74.9	70.1
[200–249]	77.7	73.0	71.1	74.9	73.7	68.7
[250–299]	81.0	74.6	64.8	73.0	76.4	69.8
[300–399]	77.6	75.6	66.8	74.2	77.0	70.0
[400–499]	78.9	74.9	64.3	72.4	75.0	69.7
>500	80.2	75.6	57.6	65.2	75.3	69.5
Average	77.0	73.0	61.0	67.5	75.4	69.6

Table 2 Statistical significance of the differences in accuracy of the secondary structure prediction between small, large, and very large proteins

Protein sizes	Prediction methods	Prediction of β-strands		Prediction of α-helices	
		t-value	Predictions for larger protein sizes are significantly better than for smaller sizes	t-value	Predictions for larger protein sizes are significantly better than for smaller sizes
Small $\leq$100 AAs versus large <100, 300>	PSI-PRED	10.97	Yes	1.73	No
	YASPIN	10.74	Yes	1.14	No
Small $\leq$100 AAs versus very large $\geq$300 AAs	PSI-PRED	6.22	Yes	1.99	Yes
	YASPIN	7.12	Yes	1.98	Yes
Large <100, 300> versus very large $\geq$300 AAs	PSI-PRED	−1.77	No	1.07	No
	YASPIN	−0.04	No	1.47	No

Yes (no) denotes that the accuracy for larger protein sizes is (is not) statistically significantly higher at 95% confidence level than the accuracy for smaller protein sizes

Positive (negative) t-values denote that the accuracies for larger protein sizes are higher (lower) than the accuracies for smaller protein sizes

The differences in the prediction quality for different protein sizes were evaluated with respect to their statistical significance. We performed independent group t-test for the predictions of the two secondary structure prediction methods to contrast the prediction quality between the small ($\leq$100 AAs), the large (100–300 AAs), and the very large ($\geq$300 AAs) proteins, see Table 2. The results show a statistically significant relation between the prediction quality and the protein size. Predictions for the very large proteins are shown to be statistically significantly better than prediction for the small proteins. Additionally, predictions for the large proteins are shown to be significantly better than prediction for the small proteins in case of the β-strand structure prediction. Positive t-values show that prediction accuracies for larger proteins are on average higher than accuracies for smaller proteins. The only exception is the difference in prediction of β-strands for very large and large proteins, in which case the accuracies are similar. These conclusions are shown to be consistent for both considered prediction methods.

3.2 Secondary Structure Content Prediction

Secondary structure content predictions are summarized in Table 3. Predictions of α-helix and β-strand content that were generated by three different state-of-the-art methods show consistent relation between the protein size and the prediction error. Namely, predictions for the very large proteins are characterized by a smaller error than prediction for smaller size proteins. In case of α-helix content prediction the relation between the error and size is proportional, i.e., the larger the proteins the smaller the error. In case of β-strand content predictions, the error for the very small, the small, and the large proteins are comparable, but errors for the very large proteins are substantially smaller. More specifically, average errors for

Table 3 Average, over proteins in a given size range, prediction error for the LP01, ZSZ01 and PSSC-core secondary structure content prediction methods in the function of protein size

Protein size	Error for α-helix content (e_H)			Error for β-strand content (e_E)		
	LP01	ZSZ01	PSSS-core	LP01	ZSZ01	PSSS-core
[30–49]	0.16	0.15	0.16	0.10	0.10	0.08
[50–69]	0.15	0.15	0.14	0.10	0.10	0.09
[70–89]	0.13	0.13	0.12	0.09	0.09	0.08
[90–109]	0.11	0.11	0.11	0.09	0.09	0.09
[110–129]	0.11	0.11	0.10	0.09	0.09	0.09
[130–149]	0.11	0.11	0.11	0.10	0.09	0.09
[150–169]	0.12	0.11	0.11	0.09	0.09	0.08
[170–199]	0.10	0.09	0.09	0.10	0.10	0.09
[200–249]	0.10	0.10	0.09	0.08	0.08	0.08
[250–299]	0.09	0.08	0.08	0.07	0.07	0.07
[300–399]	0.10	0.10	0.09	0.07	0.07	0.06
[400–499]	0.09	0.09	0.07	0.06	0.06	0.05
>500	0.09	0.09	0.09	0.06	0.06	0.05
Average	0.11	0.11	0.10	0.09	0.08	0.08

Table 4 Statistical significance of the differences in error of the secondary structure content prediction between small, large, and very large proteins

Protein sizes	Prediction methods	Prediction of β-strand content		Prediction of α-helix content	
		t-value	Predictions for larger protein sizes are significantly better than for smaller sizes	*t*-value	Predictions for larger protein sizes are significantly better than for smaller sizes
Small $\leq$100 AAs versus large <100, 300>	PSSC-core	0.01	No	8.03	Yes
	LP01	2.65	Yes	7.37	Yes
	ZSZ01	2.23	Yes	7.40	Yes
Small $\leq$100 AAs versus very large $\geq$300 AAs	PSSC-core	5.62	Yes	6.73	Yes
	LP01	6.37	Yes	5.99	Yes
	ZSZ01	5.99	Yes	6.06	Yes
Large <100, 300> versus very large $\geq$300 AAs	PSSC-core	6.25	Yes	2.81	Yes
	LP01	5.23	Yes	2.20	Yes
	ZSZ01	5.21	Yes	2.22	Yes

Yes (no) denotes that the error for larger protein sizes is (is not) statistically significantly smaller at 95% confidence level than the error for smaller protein sizes

Positive *t*-values denote that the errors for larger protein sizes are smaller than the errors for smaller protein sizes

the small proteins over the three prediction methods equal 0.13 and 0.09 for the α-helix and the β-strand content predictions, respectively. The same errors for the large proteins equal 0.10 and 0.08, respectively, and for the very large proteins they equal 0.09 and 0.06, respectively.

Similarly as for the secondary structure prediction, the significance of the differences in the content prediction error for different protein sizes was evaluated using independent group *t*-test. The quality of predictions of the three secondary structure content prediction methods were contrasted between the small ($\leq$100 AAs), the large (100–300 AAs), and the very large ($\geq$300 AAs) proteins, see Table 4. The results show that in virtually all the cases, i.e., for all the three prediction methods, predictions of both α-helix and β-strand content, and all combinations of sizes, errors for larger proteins are statistically significantly smaller than for smaller proteins. The only exception is prediction of β-strand content by PSSC-core where results for small and large proteins are comparable. The positive *t*-values confirm the strong trend that content predictions for larger proteins is characterized by better quality than for smaller proteins.

4 Discussion

The quality of the predicted secondary structure is affected by several known factors that include the size of the protein database, and the quality of the applied sequence alignment and classification algorithms. Our empirical results show that the accuracy also depends on the size of the predicted protein. Predictions for the very large proteins are characterized by statistically significantly better quality when compared with prediction for the small proteins. We believe that the better predictions are the results of availability of a larger amount of information, i.e., predictions with very large proteins are based on long sequences, and thus they use more reliable statistical/evolutionary information. The reported relation suggests that the predicted secondary structure for the very large proteins constitutes a more reliable source of information for the structural biologists when compared with the same predictions for the small proteins. In contrast, in the case of tertiary structure prediction accuracy increases with the decreasing protein size [20], which further highlights importance of our result.

Acknowledgments The author would like to thank Mr. Ke Chen for help with the preparation of the data. This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Advani S, Roy KB (2000) Biochem Biophys Res Commun 279(1):11–16
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
3. Birzele F, Kramer S (2006) Bioinformatics 22:2628–2634
4. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief W, Wedemeyer W, Schueler-Furman O, Murphy P, Schonbrun J, Strauss C, Baker D (2003) Proteins 53(S6):457–468
5. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D (2005) Proteins 54:282–288
6. Chen CC, Singh JP, Altman RB (1999) Bioinformatics 15:53–65
7. Cheng J, Baldi P (2006) Bioinformatics 22(12):1456–1463
8. Cuff JA, Barton GJ (1999) Proteins 34(4):508–519
9. Cuff JA, Barton GJ (2000) Proteins 40:502–511
10. Ding CH, Dubchak I (2001) Bioinformatics 17:349–358

11. Gibrat JF, Madej T, Bryant SH (1996) Curr Opin Struct Biol 6:377–385
12. Gong H, Rose GD (2005) Proteins 61:338–343
13. Gubbi J, Shilton A, Parker M, Palaniswami M (2006) Genome Inform 17(2):259–269
14. Heringa J (2000) Curr Protein Pept Sci 1(3):273–301
15. Hobohm U, Sander C (1994) Protein Sci 3:522–524
16. Homaeian L, Kurgan LA, Cios KJ, Ruan J, Chen K (2007) Proteins 69(3):486–498
17. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33–38, 27–28
18. Jones DT (1999) J Mol Biol 292:195–202
19. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
20. Kolinski A, Skolnick J (1998) Proteins 32(4):475–494
21. Kurgan LA, Chen K (2007) Biochem Biophys Res Commun 357(2):453–460
22. Kurgan LA, Homaeian L (2006) Pattern Recognit 39:2323–2343
23. Lin Z, Pan X (2001) J Protein Chem 20:217–220
24. Lin K, Simossis VA, Taylor WR, Heringa J (2005) Bioinformatics 21:152–159
25. Lomize AL, Pogozheva ID, Mosberg HI (1999) Proteins S3:199–203
26. Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) BMC Struct Biol 5:17
27. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) J Mol Biol 247:536–540
28. Nagy I, Trexler M, Patthy L (2003) Biochem Biophys Res Commun 302(3):554–561
29. Neves-Ferreira A, de Andrade CM, Vannier-Santos MA, Perales J, Nascimento HJ, da Silva Jr JG (2004) Protein J 23(1):71–77
30. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5:1093–1108
31. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Proteins S3:177–185
32. Pauling L, Corey RB (1951b) Proc Natl Acad Sci USA 37:251–256
33. Pauling L, Corey RB, Branson HR (1951a) Proc Natl Acad Sci USA 37:205–211
34. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O (2000) Proteins 41:17–20
35. Pollastri G, McLysaght A (2005) Bioinformatics 21:1719–1720
36. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Proteins 47:228–235
37. Przybylski D, Rost B (2002) Proteins 46:197–205
38. Rost B (2001) J Struct Biol 134:204–218
39. Rost B (2003) In: Bourne P, Weissig H (eds) Structural Bioinformatics. John Wiley & Sons, New Jersey, pp 559–587
40. Sali A, Blundell TL (1993) J Mol Biol 234:779–815
41. Samudrala R, Huang ES, Koehl P, Levitt M (2000) Protein Eng 13:453–457
42. Sayle RA, Milner-White EJ (1995) Trends Biochem Sci 20:374
43. Shen HB, Chou KC (2006) Bioinformatics 22:1717–1722
44. Wang Y, Xue Z-D, Shi X-H, Xu J (2006) Biochem Biophys Res Commun 347(3):574–580
45. Zhang ZD, Sun ZR, Zhang CT (2001) J Theor Biol 208:65–78