

Chapter 2

The GOR Method of Protein Secondary Structure Prediction and Its Application as a Protein Aggregation Prediction Tool

Maksim Kouza, Eshel Faraggi, Andrzej Kolinski,
and Andrzej Kloczkowski

Abstract

The GOR method of protein secondary structure prediction is described. The original method was published by Garnier, Osguthorpe, and Robson in 1978 and was one of the first successful methods to predict protein secondary structure from amino acid sequence. The method is based on information theory, and an assumption that information function of a protein chain can be approximated by a sum of information from single residues and pairs of residues. The analysis of frequencies of occurrence of secondary structure for singlets and doublets of residues in a protein database enables prediction of secondary structure for new amino acid sequences. Because of these simple physical assumptions the GOR method has a conceptual advantage over other later developed methods such as PHD, PSIPRED, and others that are based on Machine Learning methods (like Neural Networks), give slightly better predictions, but have a “black box” nature. The GOR method has been continuously improved and modified for 30 years with the last GOR V version published in 2002, and the GOR V server developed in 2005. We discuss here the original GOR method and the GOR V program and the web server. Additionally we discuss new highly interesting and important applications of the GOR method to chameleon sequences in protein folding simulations, and for prediction of protein aggregation propensities. Our preliminary studies show that the GOR method is a promising and efficient alternative to other protein aggregation predicting tools. This shows that the GOR method despite being almost 40 years old is still important and has significant potential in application to new scientific problems.

Key words Secondary structure prediction, GOR, Information theory, Protein aggregation

1 Introduction

The prediction of protein structure from amino acid sequence is one of the most important problems in molecular biology. With large-scale genome sequencing of various organisms and individuals for personalized (precision) medicine that produces an enormous amount of amino acid sequence data, the problem became even more important. Although prediction of tertiary structure is one of the

ultimate goals of protein science, the prediction of secondary structure from sequence is still a more feasible intermediate step in this direction. Furthermore, some knowledge of secondary structure can serve as an input for prediction. Instead of predicting the full three-dimensional structure, it is much easier to predict simplified aspects of structure, namely the key structural elements of the protein and the location of these elements not in the three-dimensional space but along the protein amino acid sequence. This reduces the complex three-dimensional problem to a much simpler one-dimensional problem. The fundamental elements of the secondary structure of proteins are alpha-helices, beta-sheets, coils, and turns. In 1983, Kabsch and Sander developed the classification of elements of secondary structure based mainly on hydrogen bonds between the backbone carbonyl and NH groups [1]. Their dictionary of secondary structure assignment Database of Secondary Structure in Proteins (DSSP) is widely used in protein science, although there are other alternative assignment methods, such as STRIDE [2]. According to the DSSP classification, there are eight elements of secondary structure assignment denoted by letters: H (alpha-helix), E (extended beta-strand), G (3_{10} helix), I (π -helix), B (bridge, a single residue beta-strand), T (beta-turn), S (bend), and C (coil). The eight-letter DSSP alphabet requires translation into the three-letter code. For instance, for the CASP (Critical Assessment of Structure Prediction) experiments, helices (H, G, and I) in the DSSP code are assigned the letter H in the three-letter secondary structure code, whereas strands (E) and bridges (B) in the DSSP code are translated into sheets (E) in the three-letter code. Other elements of the DSSP structure (T, S, C) are treated as coil (C). There are, however, other alternative ways to make these assignments.

1.1 The Original GOR Method

The GOR program is one of the first major methods proposed for protein secondary structure prediction from sequence. The original article (GOR I) was published by Garnier, Osguthorpe, and Robson in 1978, with the first letters of the authors' names forming the name of the method [3]. The method has been continuously improved and modified during the next 30 years. The first version (GOR I) used a small database of 26 proteins with about 4500 residues. The next version (GOR II) [4] used the enlarged database of 75 proteins containing 12,757 residues. Both versions predicted four conformations (H, E, C, and turns T) and were using singlet frequency. Starting with GOR III [5] the number of predicted conformations was reduced to three (H, E, and C). The GOR III method started to additionally use information about the frequencies of pairs (doublets) of residues within the window, based on the same database as the earlier version. The next version was named GOR IV [6] and it used 267 protein chains containing 63,566 residues and is still available as a web server at https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.

[pl?page=npsa_gor4.html](http://npsa_gor4.html). The latest version GOR V is using several improvements, including multiple sequence alignments and discussed in the next section [7].

Because of its simple assumptions, the GOR method has conceptual advantage over other later developed methods such as PHD [8], PSIPRED [5], SPINE-X [9], and others. While these secondary structure prediction tools rely on machine learning and typically are black boxes in terms of the principles leading to their predictions, as we briefly review below, the GOR method's reasoning for arriving at a particular prediction is clearly evident to the user. In some cases this clarity may be more significant than the slight loss in accuracy of the GOR algorithm.

The GOR algorithm is based on information theory combined with Bayesian statistics. One of the basic mathematical tools of information theory is the information function $I(S;R)$:

$$I(S;R) = \log[P(S|R)/P(S)] \quad (1)$$

For the problem of protein secondary structure prediction, the information function is defined as the logarithm of the ratio of the conditional probability $P(S|R)$ of observing conformation S , [where S is one of the three states: helix (H), extended (E), or coil (C)] for residue R (where R is one of the 20 possible amino acids) and the probability $P(S)$ of the occurrence of conformation S . The information function $I(S;R)$ is computed from a database of proteins used in the program (267 proteins for GOR IV).

The conformational state of a given residue in the sequence depends not only on the type of the amino acid R but also on the neighboring residues along the chain within the sliding window. GOR IV used a window of 17 residues, that is, for a given residue, eight nearest neighboring residues on each side were analyzed.

According to information theory, the information function of a complex event can be decomposed into the sum of information of simpler events, generally:

$$I(\Delta S; R_1, R_2, \dots, R_n) = I(\Delta S; R_1) + I(\Delta S; R_2 | R_1) + \dots + I(\Delta S; R_n | R_1, \dots, R_{n-1}) \quad (2)$$

where the information difference is defined as:

$$I(\Delta S; R_1, R_2, \dots, R_n) = I(S; R_1, R_2, \dots, R_n) - I(n - S; R_1, R_2, \dots, R_n) \quad (3)$$

Here, $n - S$ denotes all conformations different than S . The GOR IV method assumed also that the information function is a sum of information from single residues (singlets) and pairs of residues (doublets) within the window of width $2d + 1$ (i.e., $d = 8$, for the window of 17 residues):

$$\log \frac{P(S_j, \text{LocSeq})}{P(n-S_j, \text{LocSeq})} = \frac{1-2d}{2d+1} \sum_{m=-d}^d \log \frac{P(S_j; R_{j+m})}{P(n-S_j; R_{j+m})} + \frac{2}{2d+1} \sum_{n,m=-d}^d \log \frac{P(S_j; R_{j+m}; R_{j+n})}{P(n-S_j; R_{j+m}, R_{j+n})} \quad (4)$$

Here the first summation is over singlets and the second summation is over doublets within the window centered around the j -th residue. The pair frequencies of residues R_j and R_{j+m} with R_j occurring in conformations S_j and $n-S_j$ are calculated from the database. All 267 proteins in the GOR IV database have well-determined structures (with crystallographic resolution at least 2.5 Å). Using the frequencies calculated from the databases, the program could predict probabilities of conformational states for a new sequence. The accuracy of the prediction with the GOR IV program based on single sequences (without multiple alignments) tested on the database of 267 sequences with the rigorous jack-knife methodology was 64.4 %.

The advantage of the GOR method over other methods is that it clearly identifies all factors that are included in the analysis and calculates probabilities of all three conformational states. Because the GOR IV algorithm is computationally fast, it is possible to perform the full jack-knife procedure: each time when the prediction for the given sequence (of 267 sequences) is done, the sequence is removed from the database and the spectrum of frequencies used for the prediction is recalculated without including the information about the query sequence.

1.2 The GOR V Method

Several changes to the GOR IV program to improve the accuracy of the secondary structure prediction were made in GOR V. The GOR V version of the program is available from <http://gor.bb.iastate.edu/>

Modifications and improvements incorporated into the GOR V version are listed below:

1. Enlarged database of sequences with known secondary structure was used. The GOR IV database of 267 sequences was replaced by a new database of 513 nonredundant domains containing 84,107 residues proposed by Cuff and Barton [10, 11].
2. Some parameters in the GOR algorithm were optimized to increase the accuracy of the prediction. The most important modification was the introduction of the decision constants in the final prediction of the conformational state. The GOR IV program had a tendency to overpredict the coil state (C) at the cost of the helical conformation (H), and to an even greater extent at the cost of beta-strands (E). Decision parameters were therefore introduced to improve predictions.

The predicted probability of the coil (C) conformation must be greater by some critical margins than probability of either the (H) or (E) states to accept C as the winning conformation. The margin for the beta-strands is greater than for helices. The introduction of the decision constants significantly improves the predicted results by about 1.6 %.

3. The GOR algorithm was modified to include the triplet statistics within the window. The previous versions of the program used only single residue statistics (GOR I–II) or the combination of the single residue and pair residue statistics within the window (GOR III–IV). Now the GOR algorithm calculates statistics of singlets, pairs, and triplets for the secondary structure prediction. The addition of the triplets improved the accuracy of the prediction by only 0.3 %.
4. A resizable window was applied in the GOR program. The previous version of the program (GOR IV) was using the window having a fixed width of 17 residues, that is, with eight residues on both sides of the central one. The accuracy of the prediction is slightly better for the smaller window of the width of 13 residues. The Cuff and Barton database on nonredundant sequences of protein domains includes a significant number of short sequences, with many of them as short as 20–30 residues. The prediction of the secondary structure for such short sequences is very inaccurate, because of the artificial end effect of the window. Residues at the beginning or at the end of the sequence have neighbors only on one side of the window. To overcome this problem smaller windows are used for the prediction of the secondary structure of short sequences. For sequences up to 25 residues, the window size is seven residues; for sequences from 26 to 50 residues, the window size is nine residues; for sequences 51–100 residues, the window is 11 residues; and for all sequences longer than 100 residues, the window size is 13. The introduction of the resizable window allows to include all 513 nonredundant sequences in the prediction procedure.
5. Multiple sequence alignments were used for the secondary structure prediction. Multiple sequence alignments from the PSI-BLAST [12] program for each of the 513 nonredundant sequences from the database were used. The nr database which contains all known databases: all nonredundant GeneBank CDS translations + PDB + SwissProt + PIR + PRF was used, with the maximum number of five iterations in the BLAST computations. The number of alignments varied considerably depending on the sequence. For some sequences, the BLAST program produced more than 2000 alignments, whereas for some other sequences, only a few alignments. A small improvement in the prediction is obtained by removing the alignments

that are too similar to the query sequence. The best results are obtained by skipping all alignments that have identity greater than 97% to the query sequence. Besides the identity threshold, various methods of weighting of the alignments in the calculation of the accuracy of the prediction were used. The methodological procedure was based on the calculation of the matrices of the probabilities of various (H, E, and C) secondary structure elements $P_H(i, j)$, $P_E(i, j)$, and $P_C(i, j)$ for each j -th residue in the i -th alignment (with the inclusion of alignment gaps). The averages over alignments $\langle P_H(j) \rangle$, $\langle P_E(j) \rangle$, and $\langle P_C(j) \rangle$ at the j -th position in the alignment were computed and used for the prediction of the secondary structure conformation for the j -th residue. The simplest method is to use the largest probability value $\max \{ \langle P_H(j) \rangle, \langle P_E(j) \rangle, \langle P_C(j) \rangle \}$. We have modified this assignment procedure by introducing decision constants. The coil state is assigned only if the calculated probability of the coil conformation is greater than the probability of the other states (H, E) plus the imposed thresholds (0.15 for E and 0.075 for H). The value of the threshold for the beta-sheets is larger than for alpha-helices, because strands were more often erroneously predicted as coils.

All calculations for the translation of the eight-state DSSP assignments into the three secondary structure states H, E, and C are the same as these used by Frishman and Argos. This means that DSSP states H and E were translated to H and E in a three-state code, and all other letters of the DSSP code were translated to coil (C). Additionally, similar to Frishman and Argos [13], we treated helices shorter than five residues (HHHH or less) and sheets shorter than three residues (EE or E) like coils, assuming that they are most likely prediction errors. The Frishman and Argos assignment scheme is therefore highly compatible with the GOR program performance.

1.3 GOR V Web Server

We have created the GOR V web server for protein secondary structure prediction [14]. The GOR V algorithm combines information theory, Bayesian statistics, and evolutionary information. In its fifth version, the GOR method reached (with the full jackknife procedure) an accuracy of prediction Q_3 of 73.5%.

The GOR V server is based on the database of Cuff and Barton [11] of 513 sequentially nonredundant domains, which contains 84,107 residues. To ensure that such a set was representative of available proteins, nonredundancy was defined with stringent tests. The address of the GOR V web server is <http://gor.bb.iastate.edu>.

The GORV server works in the following manner. When the user provides the input sequence, the GORV server that was trained on 513 proteins calculates the helix, sheet, and coil probabilities at each residue position and makes an initial prediction based on the structural states having highest probabilities. After this initial prediction, heuristic rules are applied. These rules

include converting helices shorter than five residues and sheets shorter than two residues to coil and using decision parameters.

1.4 Input Data

The required input data includes (Fig. 1):

- Sequence name (optional). Name of the sequence or protein will appear in the Result page.
- User's e-mail address. An e-mail address to send the predicted information and notify of job completion.
- Protein sequence. The server accepts 20 single letter codes for standard amino acids, maximum 1000 amino acids in length.

For A domain of protein G (GA), the example protein shown in Fig. 1, the sequence can be accessed from the Protein Data Bank database (<http://www.rcsb.org>) or Uniprot database (<http://www.uniprot.org/>). The sequence of GA protein is deposited in PDB and Uniprot databases under identifiers 2FS1 and Q51918, respectively.

1.5 Output Data

As an output, the user receives the secondary structure prediction for the input sequence and the probabilities for each secondary state element at each position. The prediction results are shown in the web browser, which should stay open during the run, and are also sent to the e-mail address previously provided by the user. Any run-time error message will appear in the web browser, and if any problem arises, the user can contact the system administrator via the e-mail provided on the web page.

GOR V web server output data is shown in Fig. 2. The server provides the following: at the top of the output page a user might get either messages that the submission has been carried out successfully ("The e-mail address is accepted. The sequence is accepted. BLAST run is completed. GOR run is completed.") or error messages. If no submission errors have occurred, the secondary structure prediction is provided in a three-letter code: E-beta structure, H-helix, C-coil or loop. For the GA protein amino acid sequence, MEAV DANSLA¹⁰QAKEA AIKEL²⁰KQYG IGDYI³⁰KLIN NAKTVE⁴⁰GVESL KNEIL⁵⁰KALPTE, we have obtained the following secondary structure prediction: CCCCHH HHHH¹⁰HHHHHH HHHH²⁰HCCCC CHHH³⁰HHCCCCHHHH⁴⁰HHHHHHHHHHC⁵⁰CCCCC. Thus, GOR V web server predicted 3 alpha-helices, H1 [1, 5–18], H2 [25–29], and H3 [34–46] for GA protein which is in excellent agreement with the results of the DSSP algorithm [1] applied to the crystal structure (pdb code 2FS1). The structure predicted by DSSP is shown in Fig. 3. There are 3 alpha-helices: H1 [1, 5–20], H2 [24–31], H3 [36–48]. It is worth noting that more popular PSIPRED server fails to describe the entire range of complexity observed in GA folded structure. Namely, its prediction CHHH HHHHHH¹⁰HHHHH HHHHH²⁰HHCCCH HHHHH³⁰HHHHHHHHHHH⁴⁰HHHH HHHHHH⁵⁰HHCCCC suggests the presence of two helices, first

IOWA STATE UNIVERSITY

PLANT SCIENCES INSTITUTE



Laurence H. Baker Center for Bioinformatics and Biological Statistics

!!! Please use our improved CDM Secondary Structure Prediction Server that also includes GOR V results !!!

GOR V PROTEIN SECONDARY STRUCTURE PREDICTION SERVER

The GOR (Garnier-Osguthorpe-Robson) method uses both information theory and Bayesian statistics for predicting the secondary structure of proteins. Over the years, the method has been improved by including larger databases and more detailed statistics, which account not only for amino acid composition, but also for amino acid pairs and triplets. The most crucial change in the algorithm was the inclusion of evolutionary information using PSI-BLAST (Altschul *et al.* Nucl. Acids Res. 25, 3389, 1997) to increase the information content for improved discrimination among secondary structures. In GOR V, the prediction accuracy Q_3 using full-jackknifing reached 73.5%.

Sequence name (optional):

Your e-mail address :

Paste a protein sequence below:

(Please use one-letter amino acid codes with no comment line--the submission is limited to 1000 residues)

MEAYDANSLAQAKFAAKELKQYGGDYIKLINNAKTYEGVESLKNEIKALPTE

References for GOR V:

* Kloczkowski, A., Ting, K.-L., Jernigan, R.L., Garnier, J., "Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence", *Proteins*, 49, 154-166, 2002. [pdf](#)

* Sen, T.Z., Jernigan, R.L., Garnier, J., Kloczkowski, A., "GOR V server for protein secondary structure prediction", *Bioinformatics*, 21(11), 2787-2788, 2005. [pdf](#)

General References for GOR:

* Garnier, J., Osguthorpe, D.J., Robson, B. *J. Mol. Bio.*, 130, 97-120, 1978.

* Gibrat, J.F., Garnier, J., Robson, B. *J. Mol. Bio.*, 198, 425-443, 1987.

* Garnier, J., Gibrat, J.F., Robson, B. *Methods Enzymol.*, 266, 540-553, 1996.

* Kloczkowski, A., Ting, K.-L., Jernigan, R.L., Garnier, J. *Polymer*, 43, 441-449, 2002.

Please refer your questions/comments about this server to [Taner Z. Sen](#) associated with [the Jernigan group](#).

Fig. 1 GOR V web server screenshot. Example input interface is presented for GA protein sequence

GOR is running, please wait...

GOR run is completed.

This is the secondary structure prediction:

CCCCHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHCCCCHHHHHHHHHHHHHHHHHHCCCCCCCC

Column information:

- 1) Sequence index
- 2) Amino acid type
- 3) Helix probability
- 4) Sheet probability
- 5) Coil probability
- 6) GOR V prediction

```

1 M 0.063 0.126 0.811 C
2 E 0.069 0.147 0.784 C
3 A 0.126 0.210 0.664 C
4 V 0.253 0.280 0.467 C
5 D 0.410 0.291 0.299 H
6 A 0.635 0.200 0.165 H
7 N 0.625 0.186 0.189 H
8 S 0.681 0.163 0.156 H
9 L 0.715 0.150 0.136 H
10 A 0.724 0.137 0.139 H
11 Q 0.713 0.136 0.151 H
12 A 0.745 0.139 0.115 H
13 K 0.743 0.157 0.100 H
14 E 0.749 0.156 0.094 H
15 A 0.710 0.168 0.122 H
16 A 0.717 0.154 0.129 H
17 I 0.685 0.161 0.154 H
18 K 0.647 0.170 0.184 H
19 E 0.579 0.166 0.254 H
20 L 0.462 0.190 0.348 H
21 K 0.408 0.249 0.343 H
22 Q 0.265 0.263 0.472 C
23 Y 0.145 0.306 0.549 C
24 G 0.155 0.325 0.520 C
25 I 0.165 0.418 0.417 C
26 G 0.259 0.316 0.425 C
27 D 0.337 0.254 0.409 C
28 Y 0.444 0.289 0.267 H
29 Y 0.513 0.280 0.207 H
30 I 0.509 0.290 0.201 H
31 K 0.608 0.246 0.147 H
32 L 0.470 0.272 0.258 H
33 I 0.331 0.222 0.447 C
34 N 0.289 0.174 0.537 C
35 N 0.340 0.218 0.442 C
36 A 0.301 0.294 0.405 C
37 K 0.306 0.392 0.302 H
38 T 0.405 0.410 0.185 H
39 V 0.472 0.300 0.228 H
40 E 0.456 0.235 0.309 H
41 G 0.506 0.281 0.213 H
42 V 0.564 0.266 0.170 H
43 E 0.553 0.231 0.216 H
44 S 0.626 0.188 0.186 H
45 L 0.664 0.188 0.148 H
46 K 0.681 0.191 0.128 H
47 N 0.629 0.247 0.124 H
48 E 0.545 0.299 0.156 H
49 I 0.449 0.241 0.310 H
50 L 0.254 0.210 0.535 C
51 K 0.139 0.158 0.704 C
52 A 0.093 0.148 0.759 C
53 L 0.068 0.132 0.800 C
54 P 0.067 0.127 0.807 C
55 T 0.063 0.124 0.813 C
56 E 0.062 0.123 0.814 C

```

The prediction information is sent to your e-mail address. Thank you for using our service.**Fig. 2** GOR V web server screenshot. Example output interface is presented for GA protein

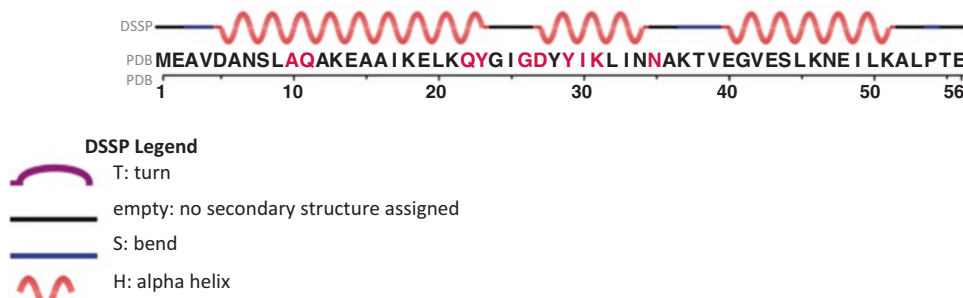


Fig. 3 Sequence chain view for GA protein (pdb code 2FSI) taken from Protein Data Bank (<http://www.rcsb.org>)

helix located from residue 2 to 22, H1 [1–19], and the second one located between residues 25 and 52, H2 [22–49]. Thus, for protein GA, the PSIPRED prediction is less accurate than the results obtained by the GOR server.

1.6 Hardware

Currently, the server is running Linux with 4.5GB RAM and 140GB memory. The program code is compiled using the Intel Fortran Compiler 8.0.034, and the web interface is established with a CGI script written using HTML and PERL. In order to use the GOR web server a user needs a personal computer workstation connected to the Internet. The GOR web server is compatible with most of popular web browsers like Google Chrome, Mozilla, or Safari.

1.7 Availability

The GOR web server is freely available at <http://gor.bb.iastate.edu>. Prediction time depends on the number of amino acids of the input sequence. For a sequence of ~100 amino acids, the secondary structure prediction takes ~30 s. The most time-consuming steps are PSI-BLAST alignments that in some cases, e.g., for many hits or slowly converging iterations, may take considerable time. Note, that the older version of the GOR server [6], GOR secondary structure prediction method version IV, is also available online at https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html.

2 Methods

2.1 Secondary Structure Prediction with the GOR Model – Features and Applications

2.1.1 Combining Secondary Structure Prediction with Structure Prediction Servers to Reduce the Search Space

Secondary structure predictions of user submitted sequence obtained by GOR, PSIPRED, or similar algorithms might be used as additional input for structure prediction servers [15, 16] or as well as for the protein-peptide docking servers [17, 18]. Apart from protein sequence, those servers typically use some information about the predicted secondary structure of a protein or a peptide. For example, the CABS-fold web server provides tools for protein structure prediction not from sequence only (de novo modeling), but also using alternative templates [15]. If you get the secondary structure predicted for some residues of the sequence of interest, you can specify the secondary structure for structure

prediction web servers. By default, CABS-fold uses the PSIPRED method [19] prediction, but alternatively predictions by the GOR server can be used.

Although PSIPRED is probably the most commonly used secondary structure prediction method and, in general, it has been reported to be around 5% more accurate than the GOR method, we have shown above that the GOR-V secondary structure prediction for GA protein is in better agreement with DSSP. An interesting question is whether this observation is also valid for other proteins? To check this, we choose conceptually different protein to GA protein: the B domain of protein G (GB) that consists of one alpha-helix and four beta-strands. Moreover, if we compare predictions by PSIPRED and GOR methods for another well-studied protein, B domain of protein G (GB), we see slightly better performance of the GOR over PSIPRED method. Namely, for GB protein amino acid sequence, MTKLILNGKTLKGETTTEAVDAATAAEKVFQYANDNGVDGEWYDDATKTFTVTE, the following secondary structure predictions are obtained: CCCEEEEC¹⁰CCCCCCHHH²⁰HHHHHHHHHH³⁰HHHHCCCCC⁴⁰CEEECCCCC⁵⁰EEEECC and CEEEE EEECE¹⁰ ECCCCHH HHH²⁰HCHHH HHHHH³⁰HHHHHH CCCC⁴⁰CEEECCCCCE⁵⁰EEEEEC by using GOR and PSIPRED methods, respectively. Figure 4 shows the secondary structure for GB protein assigned by DSSP program. There are four beta-strands and one alpha-helix, S1 [2–8], S2 [1, 6, 11, 13–16], H1 [20–33], S3 [39–43], and S4 [49–52]. Both methods detect S1, S3, and S4 strands as well as alpha-helix observed in the native conformation of protein GB. It should be noted that both methods have shortcomings, e.g., H1 is predicted by GOR to start from the residue 18 and ends at the residue 34, while PSIPRED produces wrong prediction of beta-strand located from residue 9 to 10 and alpha-helix located between residues 16 and 21. As in protein the GA case, structure prediction for protein GB by the GOR-V server appears to be more accurate than by PSIPRED.

2.1.2 Modeling Chameleon Sequences of Proteins with the Help of GOR Method

An advantage of the GOR server is that it not only provides the secondary structure prediction for the input sequence but also offers the probabilities for each secondary state element at each position. This feature can be extremely useful for studying folding process of a protein by considering not only the single native state but also the complement structure(s), which might be observed with lower probabilities. A good example of such scenario is the mutants GA98 and GB98 with sequence identity of 98%, which were obtained by performing a set of mutation experiments starting from wild-type forms of proteins GA and GB [20, 21]. Note that the wild-type proteins GA and GB have no significant sequence homology and have different folds. The mutants GA98 and GB98

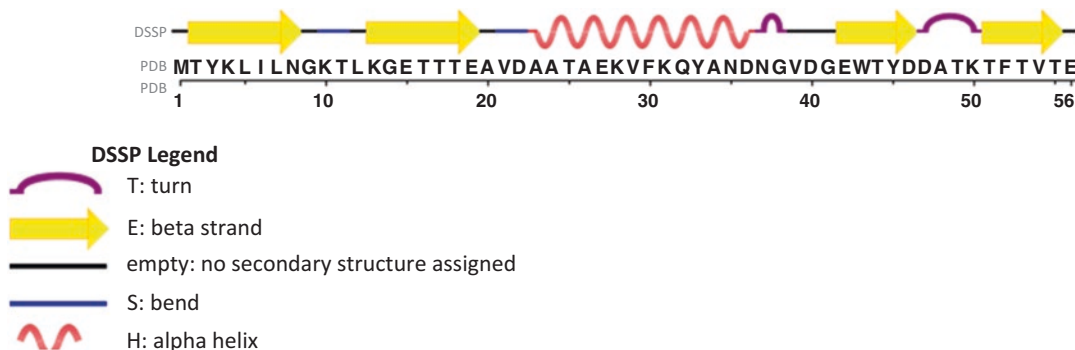


Fig. 4 Sequence chain view for the GB protein (pdb code 1PGB) taken from the Protein Data Bank (<http://www.rcsb.org>)

differ in one amino acid, but fold into different three-dimensional structures and perform different functions. GA98 shares the 3-alpha-helices fold of the parent protein GA, while GB98 shares the mixed alpha/beta fold of the parent protein GB. Interestingly, in the case of GA98, the competing structure (resembling the B domain of protein G instead of the A domain) is also observed experimentally with a certain, but low, probability [22, 23].

Figure 5 shows the probabilities to form alpha-helix and beta-strand as a function of residue position for GA (Fig. 5a), mutants GA30 and GA98 (Fig. 5b, c), GB (Fig. 5d), mutants GB30 and GB98 (Fig. 5e, f). In case of protein GB and its mutants, probabilities remain almost unchanged, while in case of protein GA and its mutants, we observe a switch between alpha fold of GA (and GA30) and alpha/beta fold of GA98. Experimentally in the case of GA98 protein, both alpha and alpha/beta folds were reported, whereas for GB98 only an alpha/beta fold of a parent GB protein [22].

Another example is the chameleon behavior of certain segments of protein sequence, which do not have a high preference for a particular conformation. The N-terminal fragment of the 49-residue protein CFr has been shown to fold into a helical structure, then unfold and finally refold into an extended beta-strand conformation [24, 25]. Incorporation of the GOR server predictions of both (alpha-helix and beta-strand) probabilities instead of only the most probable one might help to detect not only the native state, but also those observed with lower probabilities. We are now using a combination of the structure-based model [26, 27] with CABS software [28, 29] to test the effectiveness of this idea in ongoing simulations.

2.1.3 GOR Server as a Protein Aggregation Prediction Tool

The protein sequence determines its ability not only to fold but also to misfold and aggregate. Fibril formation resulting from protein misfolding and aggregation is a hallmark of several well-known neurodegenerative diseases such as Alzheimer's, type 2 diabetes, or Parkinson's diseases [30, 31]. The list of disorders linked to

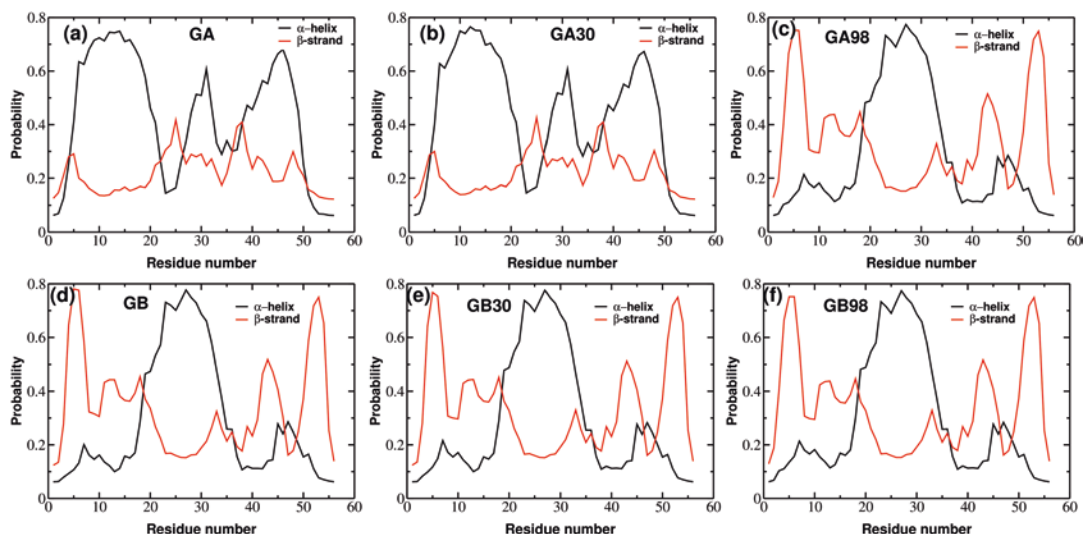


Fig. 5 Amino acid propensities for alpha-helices and beta-sheets as obtained by GOR method for protein GA (a) and GB (d) as well as their mutants GA30 (b), GA98 (c), GB30 (e), GB98 (f). In the case of protein GB and its mutants, probabilities remain almost unchanged (d–f), while in the case of protein GA and its mutants (a–c), we observe a switch from two alpha-helices (a, b) to three beta-strands (c)

protein misfolding continues to grow. For example, very recently the preeclampsia [32], a pregnancy-specific disorder, has been added to the list as it was shown to share pathophysiological features with recognized protein misfolding disorders. Although amyloid forming proteins and peptides exhibit no obvious sequence or structure homology, in many cases protein aggregates take the form of amyloid fibrils with high beta-sheet content. This suggests that the protein ability to misfold and aggregate can be described by general principles.

Recent theoretical and experimental results have indicated that protein aggregation rates depend on a number of factors such as the hydrophobicity of side chains [33, 34], preformed template fluctuations [35, 36], net charge [37], patterns of polar and non-polar residues [38], diverse secondary structure elements [39], high β -content [40], aromatic interactions [41], and the population of the fibril-prone conformation in the monomeric state [42, 43]. Many of those factors are used by bioinformatics predictive tools [44–48] to detect aggregation-prone fragments of proteins. Most of the predictive tools including TANGO [44], Aggrescan [45], Fold-amyloid [46], and Zyggregator [47] rely on a polypeptide chain sequence. For each sequence, those servers calculate own score and report aggregation-prone fragments of polypeptide sequence. Zyggregator predicts the aggregation-prone regions of polypeptide sequence based on a number of factors like hydrophobicity, charge, local stability, and the propensity to adopt alpha-helical or beta-sheet structures [47]. Aggrescan predictions are

based on an aggregation-propensity scale for natural amino acids derived from *in vivo* experiments and on the assumption that short and specific sequence stretches modulate protein aggregation [45].

In this work we test the GOR approach for the prediction of aggregation properties of protein structures. High beta-content in a monomeric state is one of the factors governing the fibril formation time [40]. Based on this, it is reasonable to assume that a high probability of amino acid to form beta-strand might correspond to its high aggregation propensity.

The accumulation of the beta-amyloid peptides found in two forms either 40 (Abeta₁₋₄₀) or 42 (Abeta₁₋₄₂) amino acids in human brains has been linked to Alzheimer's disease (AD) [31, 49]. Parkinson's disease, type 2 diabetes, and a disease known as dialysis-related amyloidosis are associated with the accumulation of amyloidogenic proteins: human alpha-synuclein, amylin, and beta-2 microglobulin respectively [50–52]. We choose these four proteins to investigate the effectiveness of the GOR method to predict aggregation-prone fragments of polypeptide sequence.

Figure 6 shows the propensity to form beta-strand as a function of residue index for typical amyloidogenic proteins, Abeta₁₋₄₀, human alpha-synuclein, amylin, and beta-2 microglobulin. The beta-sheet propensity profile of Abeta₁₋₄₀ has four peaks located at residue positions 4, 12, 17, and 32 (Fig. 6a). Note that two major peaks involving residues 17 and 32 correspond to the regions of high aggregation propensity determined experimentally (illustrated by red boundaries of squares). Namely, the central core of beta-amyloid involving residues 16–21 and the C-terminal fragment (residues 30–40).

The aggregation-prone interval from beta-sheet propensity profile can be identified as, $\Delta Ri = Ri_1 - Ri_2$, where Ri_1 and Ri_2 are the closest points to the half-maximum of the peak, Ri . We defined a fibril-prone conformation as one if beta-sheet probability exceeds the 40%. Using the above definitions we detected aggregation-prone regions for four typical amyloidogenic proteins, abeta₁₋₄₀, human alpha-synuclein, amylin, and beta-2 microglobulin (Table 1). Experimental results are also provided (Table 1 and Fig. 6) for comparison of GOR capability to predict aggregation-prone fragments of protein sequence.

In general, a peak in sheet propensity profiles can be interpreted as a sign of aggregation-prone fragment of protein sequence (except for the second peak for Abeta₁₋₄₀ and the first one for amylin). We believe that given the simplicity of the approach this agreement can be considered satisfactory and validates the use of the GOR method as an efficient alternative to other protein aggregation prediction tools.

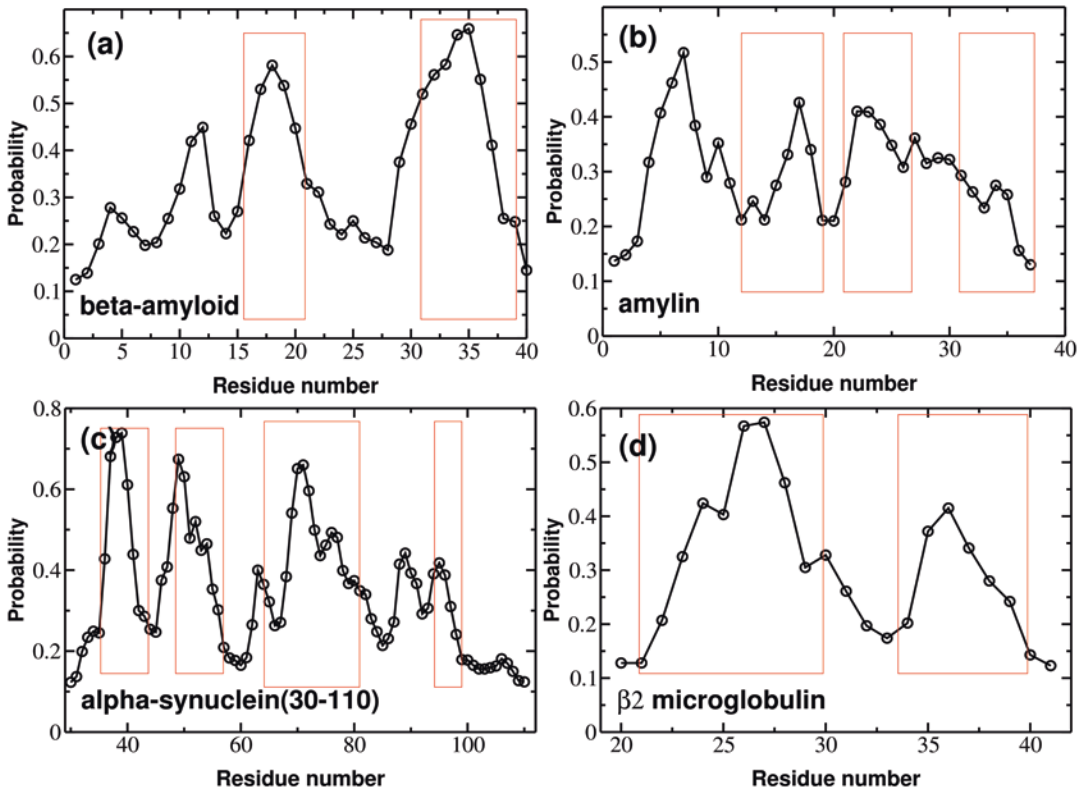


Fig. 6 Beta-sheet propensity profiles for Abeta₁₋₄₀ (a), amylin (b), alpha-synuclein (30–110) (c), beta-2 microglobulin [18–37] (d). Red boundaries of squares illustrate beta-strands of amyloid fibrils as determined by NMR experiment

Table 1
Comparison of GOR predictions with the regions adopting beta-strand configurations of different amyloid fibrils observed by NMR measurements

Protein	Experimentally known beta-strands of amyloid fibrils	GOR prediction
Abeta ₁₋₄₀	–	10–13
	β ₁ : 16–21	17–21
	β ₂ : 30–40	32–38
Alpha-synuclein (30–110)	β ₁ : 38–44	36–41
	β ₂ : 49–58	47–55
	β ₃ : 63–80	68–78
	–	86–90
	β ₄ : 92–96	92–96
Amylin	–	4–8
	β ₁ : 12–17	14–17
	β ₂ : 22–27	21–31
	β ₃ : 31–37	–
Beta-2 microglobulin [17–38]	β ₁ : 21–30	23–29
	β ₂ : 33–40	33–39

Acknowledgments

A. Kloczkowski would like to acknowledge support provided by start-up funds from The Research Institute of Nationwide Children's Hospital. This work was also supported by the Polish Ministry of Science and Higher Education Grant No. IP2012 016872 and "Mobilnosc Plus" No. DN/MOB/069/IV/2015; the National Science Center grant [MAESTRO 2014/14/A/ST6/00088].

References

1. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
2. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579
3. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120:97–120
4. Creighton TE (1990) Prediction of protein structure and the principles of protein conformation. Gerald D. Fasman, Ed. Plenum, New York, 1989. xiv, 798 pp., illus. \$95, *Science* 247:1351–1352
5. Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. *J Mol Biol* 198:425–443
6. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Meth Enzymol* 266:540–553
7. Kloczkowski A, Ting KL, Jernigan RL, Garnier J (2002) Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* 49:154–166
8. Rost B, Sander C, Schneider R (1994) Phd—an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* 10:53–60
9. Faraggi E, Zhang T, Yang YD, Kurgan L, Zhou YQ (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* 33:259–267
10. Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34:508–519
11. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40:502–511
12. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
13. Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27:329–335
14. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A (2005) GOR V server for protein secondary structure prediction. *Bioinformatics* 21:2787–2788
15. Blaszczyk M, Jamroz M, Kmiecik S, Kolinski A (2013) CABS-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res* 41:W406–W411
16. Yang JY, Yan RX, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. *Nat Methods* 12:7–8
17. Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S (2015) CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res* 43:W419–W424
18. Blaszczyk M, Kurcinski M, Kouza M, Wieteska L, Debinski A, Kolinski A, Kmiecik S (2016) Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* 93:72–83
19. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202
20. Alexander PA, He YA, Chen YH, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A* 106:21149–21154
21. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of

- two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A* 104:11963–11968
22. Bryan PN, Orban J (2010) Proteins that switch folds. *Curr Opin Struct Biol* 20:482–488
 23. Kouza M, Hansmann UHE (2012) Folding simulations of the A and B domains of protein G. *J Phys Chem B* 116:6645–6653
 24. Mohanty S, Meinke JH, Zimmermann O, Hansmann UHE (2008) Simulation of Top7-CFR: a transient helix extension guides folding. *Proc Natl Acad Sci U S A* 105:8004–8007
 25. Gaye ML, Hardwick C, Kouza M, Hansmann UHE (2012) Chameleonicity and folding of the C-fragment of TOP7. *Epl-Europhys Lett* 97:68003
 26. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN (2009) An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins* 75:430–441
 27. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298:937–953
 28. Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* 51:349–371
 29. Wabik J, Kmiecik S, Gront D, Kouza M, Kolinski A (2013) Combining coarse-grained protein models with replica-exchange all-atom molecular dynamics. *Int J Mol Sci* 14:9893–9905
 30. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 75:333–366
 31. Nasica-Labouze J, Nguyen PH, Sterpone F, Berthoumieu O, Buchete NV, Cote S, De Simone A, Doig AJ, Faller P, Garcia A, Laio A, Li MS, Melchionna S, Mousseau N, Mu YG, Paravastu A, Pasquali S, Rosenman DJ, Strodel B, Tarus B, Viles JH, Zhang T, Wang CY, Derreumaux P (2015) Amyloid beta protein and Alzheimer’s disease: when computer simulations complement experimental studies. *Chem Rev* 115:3518–3563
 32. Buhimschi IA, Nayeri UA, Zhao G, Shook LL, Pensalfini A, Funai EF, Bernstein IM, Glabe CG, Buhimschi CS (2014) Protein misfolding, congophilia, oligomerization, and defective amyloid processing in preeclampsia. *Sci Transl Med* 6:245ra292
 33. Berhanu WM, Hansmann UHE (2012) Side-chain hydrophobicity and the stability of A beta(16–22) aggregates. *Protein Sci* 21:1837–1848
 34. Otzen DE, Kristensen O, Oliveberg M (2000) Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci U S A* 97:9907–9912
 35. Nguyen PH, Li MS, Stock G, Straub JE, Thirumalai D (2007) Monomer adds to preformed structured oligomers of A beta-peptides by a two-stage dock-lock mechanism. *Proc Natl Acad Sci U S A* 104:111–116
 36. Kouza M, Co NT, Nguyen PH, Kolinski A, Li MS (2015) Preformed template fluctuations promote fibril formation: insights from lattice and all-atom models. *J Chem Phys* 142:145104
 37. Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, Dobson CM (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc Natl Acad Sci U S A* 99:16419–16426
 38. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH (1999) De novo amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci U S A* 96:11211–11216
 39. Kallberg Y, Gustafsson M, Persson B, Thyberg J, Johansson J (2001) Prediction of amyloid fibril-forming proteins. *J Biol Chem* 276:12945–12950
 40. Sgourakis NG, Yan YL, McCallum SA, Wang CY, Garcia AE (2007) The Alzheimer’s peptides A beta 40 and 42 adopt distinct conformations in water: a combined MD/NMR study. *J Mol Biol* 368:1448–1457
 41. Gazit E (2002) A possible role for pi-stacking in the self-assembly of amyloid fibrils. *FASEB J* 16:77–83
 42. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424:805–808
 43. Nam HB, Kouza M, Hoang Z, Li MS (2010) Relationship between population of the fibril-prone conformation in the monomeric state and oligomer formation times of peptides: insights from all-atom simulations. *J Chem Phys* 132:165104
 44. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302–1306
 45. Castillo V, Grana-Montes R, Sabate R, Ventura S (2011) Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J* 6:674–685

46. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 26:326–332
47. Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev* 37:1395–1401
48. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res* 43:W306–W313
49. Selkoe DJ (2004) Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases. *Nat Cell Biol* 6:1054–1061
50. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R, Stenroos ES, Chandrasekharappa S, Athanassiadou A, Papapetropoulos T, Johnson WG, Lazzarini AM, Duvoisin RC, DiIorio G, Golbe LI, Nussbaum RL (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 276:2045–2047
51. Kodali R, Wetzel R (2007) Polymorphism in the intermediates and products of amyloid assembly. *Curr Opin Struct Biol* 17:48–57
52. Floege J, Ketteler M (2001) beta(2)-microglobulin-derived amyloidosis: an update. *Kidney Int* 59:S164–S171

Prediction of Protein Secondary Structure

Zhou, Y.; Kloczkowski, A.; Faraggi, E.; Yang, Y. (Eds.)

2017, XI, 313 p. 67 illus., 56 illus. in color., Hardcover

ISBN: 978-1-4939-6404-8

A product of Humana Press