10-2013

Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure

S. Saraswathi Battelle
The Research Institute at Nationwide Children's Hospital

J .L. Fernández-Martínez
University of Oviedo

A. Koliński
Warsaw University

R. L. Jernigan
Iowa State University, jernigan@iastate.edu

A. Kloczkowski Battelle
The Research Institute at Nationwide Children's Hospital

Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure

Abstract
Exponential growth in the number of available protein sequences is unmatched by the slower growth in the number of structures. As a result, the development of efficient and fast protein secondary structure prediction methods is essential for the broad comprehension of protein structures. Computational methods that can efficiently determine secondary structure can in turn facilitate protein tertiary structure prediction, since most methods rely initially on secondary structure predictions. Recently, we have developed a fast learning optimized prediction methodology (FLOPRED) for predicting protein secondary structure (S. Saraswathi, et al., [1]). Data are generated by using knowledge-based potentials combined with structure information from the CATH database. A neural network-based extreme learning machine (ELM) and advanced particle swarm optimization (PSO) are used with this data to obtain better and faster convergence to more accurate secondary structure predicted results. A five-fold cross-validated testing accuracy of 83.8 % and a segment overlap (SOV) score of 78.3 % are obtained in this study.

Secondary structure predictions and their accuracy are usually presented for three secondary structure elements: α-helix, β-strand and coil but rarely have the results been analyzed with respect to their constituent amino acids. In this paper, we use the results obtained with FLOPRED to provide detailed behaviors for different amino acid types in the secondary structure prediction. We investigate the influence of the composition, physico-chemical properties and position specific occurrence preferences of amino acids within secondary structure elements. In addition, we identify the correlation between these properties and prediction accuracy. The present detailed results suggest several important ways that secondary structure predictions can be improved in the future that might lead to improved protein design and engineering.

Disciplines
Biochemistry | Biophysics | Molecular Biology | Structural Biology

Comments
This is a manuscript of an article published as Saraswathi, Saras, Juan Luis Fernández-Martínez, Andrzej Koliński, Robert L. Jernigan, and Andrzej Kloczkowski. "Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure." Journal of molecular modeling 19, no. 10 (2013): 4337-4348. The final publication is available at Springer via http://dx.doi.org/10.1007/s00894-013-1911-z. Posted with permission.

Published in final edited form as:
J Mol Model. 2013 October ; 19(10): 4337–4348. doi:10.1007/s00894-013-1911-z.

Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure

S. Saraswathi Battelle,
Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH, USA

J. L. Fernández-Martínez,
Department of Mathematics, University of Oviedo, Oviedo, Spain

A. Koliński,
Laboratory of Theory of Biopolymers, Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

R. L. Jernigan, and
Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, USA

A. Kloczkowski Battelle
Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Department of Pediatrics, The Ohio State University College of Medicine, 700 Children's Drive, Columbus, OH, 43205 USA, Tel.: +1-614-722-3880, Fax: +1-614-355-2728,

A. Kloczkowski Battelle: Andrzej.Kloczkowski@NationwideChildrens.org

Abstract

Exponential growth in the number of available protein sequences is unmatched by the slower growth in the number of structures. As a result, the development of efficient and fast protein secondary structure prediction methods is essential for the broad comprehension of protein structures. Computational methods that can efficiently determine secondary structure can in turn facilitate protein tertiary structure prediction, since most methods rely initially on secondary structure predictions. Recently, we have developed a fast learning optimized prediction methodology (FLOPRED) for predicting protein secondary structure (S. Saraswathi, et al., [1]). Data are generated by using knowledge-based potentials combined with structure information from the CATH database. A neural network-based extreme learning machine (ELM) and advanced particle swarm optimization (PSO) are used with this data to obtain better and faster convergence to more accurate secondary structure predicted results. A five-fold cross-validated testing accuracy of 83.8 % and a segment overlap (SOV) score of 78.3 % are obtained in this study.

Secondary structure predictions and their accuracy are usually presented for three secondary structure elements: α-helix, β-strand and coil but rarely have the results been analyzed with respect to their constituent amino acids. In this paper, we use the results obtained with FLOPRED to provide detailed behaviors for different amino acid types in the secondary structure prediction. We investigate the influence of the composition, physico-chemical properties and position specific occurrence preferences of amino acids within secondary structure elements. In addition, we identify the correlation between these properties and prediction accuracy. The present detailed

results suggest several important ways that secondary structure predictions can be improved in the future that might lead to improved protein design and engineering.

Introduction

Due to advances in sequencing, millions of protein sequences are available in the protein data bank [2]. Yet, currently we have only about 80,000 solved protein structures. In principle, this large gap can be filled by protein structure prediction. Expensive and time consuming experimental protein structure determination methods such as X-ray crystallography and nuclear magnetic resonance (NMR) are not possible for large scale applications on the genome scale. Secondary structures can be predicted cheaply and easily using a variety of computational methods including machine learning. Secondary structure prediction is often a prerequisite to 3-D structure prediction. Hence developing faster and more accurate secondary structure prediction methods remains important. In addition to achieving high prediction accuracy, it is important to discover the influence of the content and position specific occurrences of amino acids on prediction accuracy. This knowledge can result in improved secondary structure predictions by ensuring that training models make use of such characteristics to make more successful and more accurate predictions.

Neural networks have been the most successful computational approach used for secondary structure prediction. Other prediction methods that have been successfully applied to secondary structure prediction include statistical methods, nearest neighbor methods, hidden Markov models (HMM) and support vector machines (SVM) [3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23]. On average about 70 % accuracy was obtained when predictions were based on a single amino acid sequence. A 10 % improvement in prediction accuracy was achieved when evolutionary information from multiple sequence alignments (MSA) was included [24,25,26,8,27,28,29,30,31,32,33]. Some accuracy improvements were obtained by including long-range interactions [34,35]. A compound pyramid model (CPM) that used a two-level mixed-modal SVM (MMS) [32] for secondary structure predictions has the highest reported accuracy of 85.6 % [36].

Our method (FLOPRED) uses novel data based on knowledge-based potential information calculated by using the CABS algorithm [37], which captures structural information for predicting probable structures. FLOPRED uses a simple single layer neural network called extreme learning machine (ELM) [38,39,40]. The results from this algorithm are further optimized by using an advanced particle swarm optimization (PSO) algorithm [41,42,43,44,45,46,47]. These features make this algorithm highly efficient, accurate and less expensive to use, compared to other algorithms that apply more complicated algorithms and require larger computational resources for structure prediction. FLOPRED learns from the information encoded in individual protein sequences and predicts the three secondary structure elements: α-helix, β-sheet and coil accurately with an average five-fold cross-validated accuracy of 83.8 % and a segment overlap score (SOV) [48,49] of 78.3 % (see Sect. S1). SOV score is an alternate measure for the evaluation of secondary structure prediction methods that is based on secondary structure segments rather than individual residues. SOV score is considered superior to traditional scores obtained through residue-

based approach. FLOPRED is very simple, requires fewer resources and yields an accuracy comparable to the best found in the literature. A comprehensive discussion of the data, methods and analysis of our previous study using FLOPRED is given in [1] and its supplementary materials. Our results are significantly better (see Tables 1 and 2) than those found in the literature for studies that do not use the evolutionary information contained in multiple sequence alignments, and is comparable to results that include MSA, although it may not be entirely fair to compare results of studies that use different datasets. Our method does not use multiple sequence alignments; however certain evolutionary information is implicitly embedded within the CATH library of folds.

Rarely have analyses been carried out that look in detail at the importance of the specific amino acid types for the predictions, with only a few exceptions [50]. For example, the results from several secondary structure prediction servers were studied and reanalyzed to discern patterns of prediction accuracies with respect to different amino acids [51]. Chou and Fasman used the frequencies of the occurrences of each amino acid as a parameter to predict secondary structure [4]. Those studies found that amino acids have different propensities for the three protein secondary structures, α-helix, β-sheet and coil. Later studies showed that the accuracy of predictions vary widely, depending on the parameters used for prediction [52]. These parameters differ depending on the frequency of occurrence of amino acids and extent of homogeneity of the secondary structures present in the proteins being studied.

In contrast to earlier studies, the emphasis in this paper is on the analysis of secondary structure prediction results with respect to the content and position specific occurrences of amino acids. An in-depth analysis of amino acid accuracies obtained by applying FLOPRED to secondary structure prediction is given. In addition, interesting and intriguing patterns in the classification results obtained are highlighted. In order to discern patterns of prediction that might help to improve secondary structure prediction accuracy, we inspect and analyze our results with respect to the behavior of hydrophobic and hydrophilic amino acids in order to derive the :

- Correlation between individual amino acid content, in variation and secondary structure prediction accuracy.
- Contribution of each secondary structure accuracy to the overall prediction.
- Difference in prediction accuracies for amino acids occurring in the *Middle* and *Ends* of secondary structures.
- Pattern of prediction accuracies for amino acids that occur in the *Middle* of secondary structure elements considered separately from those residues that occur at the *Ends*.

These derivations will then will permit us to draw conclusions to suggest ways to improve secondary structure prediction in future studies. Additionally, our results might be helpful in the development of future protein design and protein engineering applications and in the prediction of the effect of substitutions of individual amino acids on protein structure.

Note: All tables, figures or section numbers starting with the letter 'S' are found in the supplementary materials.

Data and methods

A three-class secondary structure assignment of the eight states in the DSSP alphabet [53] is used in our predictions. The three states that are grouped under Helix (H) include the regular α-helix H, the extended 3_{10} helix G and the compressed π-helix I; E and bridge B are grouped under β-sheet (E); while turns T, bends S, blanks and C are classified as coil. FLOPRED is tested on a subset of proteins culled from the CB513 [26] dataset. Target sequences that had more than 70% sequence identity (according to a global Needleman-Wunsch sequence alignment [54] using BLOSUM62 [55]) were removed. Sequences with structural similarity, according to HSSP [56], the Homology-derived Secondary Structure of Proteins database, were also eliminated from our data set. This reduced dataset, of 387 protein sequences is used for secondary structure predictions. While the number of proteins is relatively small compared to the original set of 513 sequences, nonetheless you do have a total of 63, 383 residues and a significant number of amino acids of each type, and there is always greater uncertainty for the rarer amino acids such as tryptophan and methionine. The FLOPRED algorithm is trained using this data and its efficiency and robustness are tested using five-fold cross validation.

Knowledge-based potentials

Knowledge-based potentials are extracted by using the CABS [37] algorithm to obtain 27 features to represent each amino acid in a sequence. CABS stands for C-α-C-β-Side group protein model where C-α is the α-carbon and C-β is the β-carbon of an amino acid. The associated force field encodes both short-range and long-range interactions in proteins. It uses a lattice model to represent hundreds of possible orientations of the virtual α-carbon- α-carbon virtual bond and uses highly efficient Replica Exchange Monte Carlo for sampling the conformational space. The knowledge-based potentials of the force field include the following information:

- Protein-like conformational biases.
- Statistical potentials for the short-range interactions.
- A representation of main chain hydrogen bonds.
- Statistical potentials describing the side chain interactions.

Please refer to Sect. S4 and Saraswathi et al., [1] and its supplementary materials for a comprehensive description of the data generation and methods used. We have given only an abbreviated account here since we aim to discuss our previous results with specific emphasis on the behavior of the amino acid and its various properties with respect to accuracy. Description of the selection criteria and other details of the data generating algorithm such as energy calculations and creation of profile matrices are described in Sect. S4.1.

FLOPRED- an Extreme Learning Machine classifier

A single layer feed forward-network based classifier called an Extreme Learning Machine (ELM) is used for secondary structure predictions here. Compared to other neural networks, ELM offers the smallest training error resulting in rapid convergence and best generalization performance. The parameters of this model, such as input weights and bias are randomly assigned and optimized using PSO. The number of hidden neurons which are limited to between 5 % and 10 % of the number of training inputs are randomly assigned and optimized by the PSO. A sigmoidal activation function is used for the hidden layer and a linear activation function is used for the output neurons. By assuming the network output (Y) is equal to the coded class label, the output weights (W) are calculated analytically as,

$$W = YY_h^{\dagger},$$

where $Y_h^{\dagger}$ is the Moore-Penrose generalized pseudo-inverse of the hidden layer output matrix Y_h An overview of ELM is given in Sect. S4.2 in the supplementary materials.

Particle Swarm Optimization

Improved and extended versions of the PSO algorithm [41,42,43,57,45,47] have been used to improve secondary structure prediction accuracy by tuning parameters of the ELM, such as weights, bias and the number of hidden neurons. The natural behavior of individuals in groups, such as a swarm of bees or a flock of birds, is mimicked by the PSO global optimization algorithm. For example, such a group may need to collectively solve an optimization problem such as how best to reach their nest or hive. Through intelligent sampling of the prismatic volume in the model space, PSO tries to find the best parameters that are nearest to the global minimum which results in minimum error in classification. A comprehensive description of this algorithm is given in Sect. S.4.3 in the supplementary materials. Our predictions make use of these advanced and efficient PSO algorithms to achieve additional robustness and significantly improved prediction accuracy.

Results and Discussion

Amino acids occur in various quantities and positions in different secondary structures. Secondary structure prediction results that have been averaged over many iterations mask the underlying variation in the predictions for individual amino acids in different secondary structures. We analyzed prediction results in the context of amino acids with high and low hydrophobicity index. According to the hydrophobicity scale determined by Kyte and Doolittle [58], amino acids such as Ile, Val, Leu, Phe, Cys, Met and Ala are considered hydrophobic (group1) with a hydrophobicity index that ranges between 1.8 and 4.5. Gly, Thr, Ser, Trp, Tyr and Pro that have values between −1.6 and −0.4, are less hydrophobic (group 2). His, Asp, Glu, Asn, Gln, Lys and Arg are very hydrophilic (group 3) with hydrophobicity values that range between −4.5 and −3.2. We consider residues in group 1 and group 2 to be hydrophobic and those in group 3 to be hydrophilic. We examined accuracies with respect to amino acid content and position specific propensities of amino acids in each secondary structure, to see whether there is any correlation between these

features and prediction accuracy. This analysis could help to find a better representation of proteins in the training model with the aim of improving prediction accuracy, especially for β-sheet predictions that have traditionally had a lower accuracy than for α-helix and coil structures. A brief discussion of the prediction results obtained by FLOPRED with respect to the three secondary structures: α-helix, β-sheet and coil is given next.

Secondary structure prediction

Prediction accuracies obtained through a 5-fold cross-validation study are highest for α-helix and lowest for β-sheet with an average accuracy of 80.4 % and an overall Q_3 score of 83.8 % for the three secondary structures, as seen in Table 1. The 3.4 % difference in accuracy between overall and average accuracy is due to the variability in the content of amino acids in the three secondary structures. Table 3 and Figs. S1 and S4 show that the variability in prediction accuracies is highest for β-sheet and lowest for α-helix, while coil falls in the middle. Table. 3 and Fig. S3 gives the overall variability in the content of amino acids across the three secondary structures and the corresponding difference between the overall and average accuracy for each amino acid. It can be seen that as the variability increases, the difference in accuracy also increases (see Fig. S4).

SOV scores are given in Table 1 along with training and cross-validation results. The SOV scores for the test data is observed to be, highest for α-helix at 85.8 %, 77.5 % for β-sheet and 71.8 % for coil with an overall SOV score of 78.3 %. FLOPRED results compared favorably with those studies in the literature that use the CB513 dataset for secondary structure prediction and use multiple sequence (evolutionary) information, to develop their models (see Table 2). Note that, in addition, we are using information derived from protein sequences and knowledge-based potentials calculated with the CABS algorithm. A detailed discussion of these results can be found in Saraswathi et al., [1] and in the supplementary materials under Sect. S5. We have included a brief comparison here. Except in one case (CPM method, [36]), our average testing accuracy of 83.8 % is higher than the accuracies found in the literature (see Table 2). FLOPRED achieves gains between 1.7 % and 10.4 % in Q_3 results compared to previous methods and its accuracy is 1.8 % lower than the best method (CPM). The α-helix testing accuracy (91.1 %) is still the highest compared to other studies, while the β-sheet accuracy is lower than the CPM method by 5.8 %. Coil accuracies do not fare so well compared to previous studies. The SOV scores of 78.3 %, is between 1.8 % and 7.5 % higher than the first four studies listed and is less than the CPM and SPINE X [50] studies by 0.7 % and 1.5 % respectively. The higher accuracies of FLOPRED can be attributed to the learning capabilities of the ELM algorithm and the advanced optimization techniques offered by the PSO algorithms [45] that were used to tune the parameters of the neural network. When using neural networks and other machine learning techniques, improved accuracies might be obtained by having a good representation of the test data in the training models. We have investigated whether, the content for all twenty amino acids is adequately represented in the data set (to enable the training model to learn about each amino acid), by analyzing results at the amino acid level.

Amino acid content and secondary structure accuracy

The secondary structure predictions of 63, 383 amino acids that are present in 387 proteins are analyzed here. Amino acids are present in proteins in varying quantities (see Table 3 and Figs. S3), with large and small variations in frequency of occurrence, which in turn lead to variation in prediction accuracies for different amino acids (see Fig. S4). The overall rankings range from the highest accuracy (rank 1) for Alanine to the lowest accuracy (rank 20) for Histidine. It can be observed from Table 3 that there is not much correlation between the content and corresponding secondary structure prediction accuracies and there are not any discernible patterns with regard to the effect of content on prediction accuracy (See Sects. S5.2 and Fig. S2). Although Ala has the largest content (8.4 %) and holds the top rank for accuracy, other residues like Met and Trp with very low content still enjoy higher ranks than some residues with much larger content. We investigated further to examine if a pattern of low content with high variability across the three secondary structures, results in correspondingly lower prediction accuracies.

Correlation between variability, content and accuracy

In this section the overall content, variability, accuracy and ranking of amino acids are discussed. The discussion for each secondary structure: α-helix, β-sheet and coil is given under Sects. S5.3 – S5.6.

Overall amino acid content and variability in secondary structures—Amino acid content is shown in Table 3 and Figs. S3a, S3b and S3c for α-helix, β-sheet and coil. The variation in amino acid content is highest for α-helix and coil while they are comparatively more uniform for β-sheet conformations. Our data shows that 39.9 % of amino acids occur in α-helices and individual amino acid contents vary between 20.2 % and 56.5 % in α-helix. Only 20.1 % of amino acids occur in β-sheet and individual amino acid contents vary between 9.7 % and 39.2 % in β-sheet. The remaining 40 % of amino acids occur in coil and amino acid contents vary between 27.5 % and 69.6 % in coil. It is interesting to note that the variability of 10 out of 12 hydrophobic type residues, are the lowest (between 3.8 % and 12.8 %), while Gly has a high variability at 22.8 %. The variability for the hydrophobic Ala (17.3 %) and other hydrophilic residues are above mid-range between 12.8 % and 18.7 %. Pro has the highest variability at 26 %, which is almost 7 times larger than the lowest variability for Cys (3.8 %).

Overall amino acid accuracies in secondary structures—The correlation between the content and accuracy for α-helix is 0.57, for β-sheet 0.77 and for coil it is 0.58. There is a negative correlation of −0.12 between variability and accuracy. Correlation between average content and overall accuracy, for all three secondary structures, is only 0.14. This might be due to the fact that most of the variances in accuracy for individual secondary structures are being hidden, where the higher accuracy in one secondary structure element is offset by a corresponding lower accuracy for the same amino acid in another secondary structure element.

Table 3 and Figs. S4a, S4b and S4c give the α-helix, β-sheet and coil classification accuracies for the twenty amino acids for all three secondary structures. Here, α-helix

accuracies range between 84 % and 97.2 %, whereas β-sheet accuracies are much lower and range only between 46.1 % and 87.5 %. Coil accuracies are slightly better and range between 64.6 % and 90 %. Average accuracies for α-helix, β-sheet and coil are 91.1 %, 71.9 % and 78.2 % with standard deviation of 4.3 %, 13.2 % and 9.3 % respectively.

The amino acid deviations from mean accuracy differ widely for different secondary structures. When deviations from their average mean for each of the three secondary structures are plotted separately, as shown in Fig 1, some interesting patterns emerged. Fig. 1d plots the deviations from the overall mean accuracy (83.8 %) for all amino acids. The largest negative deviation is for hydrophilic Asp at −4.05 % and the largest positive deviation is for hydrophobic Ala at 4.2 %. The positive deviations for α-helix and β-sheet in Figs 1a and 1b are compensated by the opposite deviation for coil in in Fig 1c. We find that the overall pattern of deviations is quite different from the tendencies for individual secondary structures. There is no noticeable trend in these deviations except that the hydrophilic residues seem to have slightly larger negative deviations compared to hydrophobic residues.

The variations in residue frequencies among the three secondary structures seem to influence the overall accuracy for all twenty amino acids as shown in the accuracy rankings in Figs. 2a, 2b and 2c. There is also a tendency for alpha helix and beta sheet to have enhanced accuracy for the same amino acids for many of the residues (see Fig. 2d). Opposing this is the lower accuracy for the same amino acids in coil. In general, we can see higher accuracies in the form of larger positive deviations when content for an amino acid is higher in one secondary structure compared to its content in the other two secondary structures. The hydrophilic amino acids do well in coil predictions while the hydrophobic residues do well in α-helix and β-sheet. But in general the hydrophilic residues do poorly in the final predictions since their gains in coil predictions are offset by their poor performance in the other two secondary structures. Better overall prediction accuracies are likely, when content in each secondary structure is evenly distributed (or at least have lesser variations) among the three secondary structures, although there are some exceptions to these rules.

Summary of analysis on content, variation and accuracies in secondary structures—The length and amino acid composition of secondary structures and the fold compositions of proteins that are used in the data can greatly impact and influence the final accuracy obtained in secondary structure prediction. Our observations indicate that a greater presence of an amino acid in a particular secondary structure does not assure higher accuracy for that residue in that secondary structure. Amino acids which appear in very small quantities in the data, like Cys, Met and Trp and show low variability still do well, since their content is evenly distributed among all three secondary structures. Some trend is visible for the hydrophilic residues in β-sheet where there is very low content between 10 and 15 %, leads to a very low ranking of between 16 and 20. In other cases even if there is a fairly even distribution of residues among the three secondary structures, the ranking is still low for residues in coil, unless a majority of the residues present are in coil.

The discussions above indicate that higher α-helix content will have positive influence, β-sheet content will generally negatively impact a residue's overall accuracy, which can be

slightly counteracted with a higher coil content which has better accuracies. We observe that the ranking tendencies for α-helix and β-sheet match their overall accuracies more closely than they do with coil. The highest overall ranking is for Ala, followed by Lys and Arg, while the lowest rankings are for Asn, Aspand His. Some residues like Val, Ile and Trp have high rankings for α-helix and β-sheet, but they have low overall ranking due to very low rankings in coil, while some residues like Ala, Lys and Arg have final high ranking despite having only an average ranking for the individual secondary structures. Due to these tendencies the loss or gain in ranking from two of the secondary structures is commonly equalized by the tendencies in coil. Hence the final overall accuracies for each of the twenty amino acids do not reflect the individual characteristics separately for each of the three secondary structures.

The comparative studies for content vs. accuracy leads to a clear way forward to improve secondary structure predictions. Consider predicted probabilities for all positions in a sequence. If each amino acid and each secondary structure type has a further factor based on the accuracy from the rankings in Figs. 2a, 2b and 2c, then it will give a higher probability to the cases that are most accurate in their predictabilities. This would be a straightforward procedure to implement. Next, we look at prediction accuracies for residues that appear in the *Middle* of secondary structures in comparison with those that appear at the *Ends* of secondary structures, to see whether there is any position specific preference with respect to accuracy.

Influence of position specific amino acid occurrences on secondary structure prediction accuracies

Position specific preferences of amino acids result in a variety of content in the *Middle* and *Ends* (see Table S2 and Fig. S7a) of secondary structures. A residue is considered to be an *End* residue, if it is part of the first three or the last three residues of a secondary structure. All other residues are considered to occur *in the middle* of secondary structures. We find that overall prediction accuracies differ considerably for residues that are present in the *Middle* vs. *Ends* of secondary structures (see Table 4).

Comparative analysis of correctly predicted residues in the *Middle* vs. *Ends* of secondary structures, when these two regions are considered collectively—

Average content in the *Middle* of secondary structures is 45.5 % for all amino acids (Table S2). But, the content in the *Middle* regions vary widely when we consider each of the secondary structures individually. For longer α-helix structures, there will be more residues specified as being in *Middle* regions than in *End* regions. Accuracy for *Middle* regions range between 32.8 % and 56 % with an average accuracy of 49 %. The accuracy for *End* regions range between 44 % and 67.2 % with an average accuracy of 51 % (see Table S3). The residues that are correctly predicted seem to be distributed evenly between the two regions, while the residues that are incorrectly predicted seem to occur overwhelmingly at the *END* regions.

Table 5 shows amino acid content variation at the *Ends* regions, for the three secondary structures. In contrast to correct predictions, an overwhelming majority of errors occur at the

Ends of secondary structures. Table 4 and Fig. S6 show errors that occur in the *Middle* vs. *Ends* of secondary structures. Fig. 3 shows that the *Ends* exhibit large variation in their error occurrence.

All prediction errors (100 %) for Ile, Met, Trp and Tyr in α-helix (see Table 4 and Fig. S6-a) are only at the *Ends*, although the content for these residues range only between 25.1 % and 39.3 % for the *End* regions. The lowest error for *End* regions occur for Gln (84.5 %) followed by His at 88.6 % (their contents are 33 % and 41.7 %, respectively). The complement of these error percentages occur for the *Middle* of α-helix and are limited to between 0 % and 10 % for 18 out of the 20 amino acid types.

All prediction errors (100 %) that occur for Met, Phe and Tyr are only at the *Ends* of β-sheet (see Table 4 and Fig. S6-b), although their content is around 70%. The lowest error is for Glu (84.7 %), but this is still very high. Errors in the *Middle* of β-sheet are limited to between 0 % and 10 % for 19 out of 20 amino acid types.

All prediction errors in coil for Gln (100 %) occurs only at the *Ends* (see Table 4 and Fig. S6-c) although *End* content is about 60%. The lowest error occurs for Phe (78.6 %) and Tyr (79.1 %). The errors for coil in the *Middle* of secondary structures are larger compared to the other two secondary structures.

Summary of analysis of predictions in the *Middle* vs. *Ends* of secondary structures—*End* predictions differ widely for different amino acids. It is well known in structure prediction that the determination of boundaries between secondary structures states along the sequence is a common problem. What we are seeing here is a reflection of this type of prediction problem. Some residues have higher predictions accuracies for a particular secondary structure compared to others. For example, predictions for residues such as Met or Ile for α-helix and β-sheet are much more reliable than their predictions for coil. These statistics, gathered for representative sets of proteins, can be applied to improve secondary structure predictions for some of these residues by introducing different reliability factors for the different situations. Although the average overall content for all amino acids is 45.5 %, there are large variations (between *Middle* regions and *Ends*) in the amino acid content for each of the three secondary structures (Table S2).

Looking at the combined content for *Middle* vs. *End* regions does not yield a clear picture of the underlying patterns in predictions, although we found that most of the errors do occur only at the End regions. There appear to be no correlations between higher content and higher accuracy for individual secondary structures for *Middle* vs. End regions. Aggregating results without considering position specific preferences (for *Middle* vs. End regions), masks the large differential behavior of errors for different amino acids.

Next, we look at prediction vs. content patterns for residues that appear in the *Middle* and *End* regions separately, in order to understand the nature of these errors and draw our conclusions.

Comparative analysis of errors that occur in the *Middle* vs. *Ends* of secondary structures, when these two regions are considered separately—If we consider

residues that occur only in the *Middle* regions (separately, from those residues that appear at the *Ends*, the percentage of residues that are correctly predicted (see Table S5) is very high and ranges between 89.7 % and 98.6 %. Mean overall accuracy for hydrophobic residues is 96 % vs. 95 % for hydrophilic residues. Hence the overall errors in prediction in the *Middle* regions are very low and range between 1.4 % and 10.3 % (average of 4.5 %).

Table 5 gives the content and errors in prediction for residues that occur at the *Ends* of secondary structures. Errors in prediction for residues in individual secondary structures (at the *Ends*) differ widely, as seen Fig. S7b. The content in the *Ends* regions show large variability across the three secondary structures, that possibly leads to large errors. Errors occurring at the *Ends* of α-helix are lower than those occurring in the other two secondary structures. Errors occurring in β-sheet are lower than those occurring in coil for some hydrophobic residues while the errors are generally higher for all hydrophilic residues. In contrast to *lpha*-helix and β-sheet, hydrophilic residues at the *Ends* of coil have better accuracies than hydrophobic residues. The variations in errors at the *Ends* of secondary structures are discussed more extensively in the supplement.

Although it appeared during the above discussions that the predictions at the ends of secondary structures are very prone to errors (showing over 90 % errors when analyzed with respect to overall content), we see a brighter picture here with much lower errors (see Fig. S7b and Table 5) when the residues are considered only with respect to content at the *Ends*. Table S4 indicates that an average of 70.6 % of residues that appear at the *Ends* of secondary structures are correctly predicted. The average accuracy for the hydrophobic residues is 72.5 % (with variation of 12.3 %) and 67.8 % (with much higher variation of 19.7 %) for hydrophilic residues.

Contrary to what we saw earlier, the variations (average of 13.6 %) in errors are thrice as much for the *End* predictions when compared to variations for residues in the *middle* (4.5 %). The larger variations possibly lead to a higher error rate (between 25 % and 41.5 %) and less reliability for *End* predictions. The highest errors are for Gly, His, Pro, Ser, Asn, Asp and Gln. Most of these residues have high variations in content except for His and Gln.

Summary of pattern of predictions at the *End* regions of secondary structures—The correlation coefficient between content and errors at the *Ends* of secondary structures are −0.19 for *alpha*-helix, −0.8 for β-sheet and −0.55 for coil. There is a strong negative correlation between content and errors for all three secondary structures, with this relationship being much stronger in β-sheet than it is for the other two secondary structures. The content for correct predictions is dominated by *alpha*-helix structures, while those for errors are dominantly for coil structures. So, it is possible that a paucity of residues contributes to the low accuracies of β-sheet compared to other structures. *alpha*-helices are also much longer than the other two secondary structures on average and hence for *Middle* regions, the number of *alpha*-helix residues is larger than it is for β-sheet which occur on average in shorter lengths and in smaller number of cases compared to both *alpha*-helix and coil. The residues listed for α-helix and β-sheet as having the highest errors at *End* regions do correspond to residues having lower content, but there is no such relationship for residues in coil.

Conclusions

A five-fold cross-validation for predictions yielded 91.1 % accuracy for α-helix, 71.9 % for β-sheet and 78.2 % for coil, with an overall Q_3 score of 83.8 %. These results have been analyzed with respect to their amino acid content in order to investigate possible reasons for lower accuracies of β-sheet and to discover patterns that might be used to improve secondary structure predictions. We found that each amino acid in the three secondary structures has its own prediction pattern. Considering prediction results only at the secondary structure level masks these patterns, which could be gainfully used to improve classification of secondary structures. FLOPRED seems to have very good prediction accuracies for residues, which occur in the *Middle* of secondary structures (over 90 %) and has nearly 70 % accuracy for residues that occur at the *Ends* of structures. Amino acids with lower content or uneven distribution among the three secondary structures may have lower prediction accuracies. We find that there is some correlation between the content of amino acids in each of the secondary structures and their prediction accuracies. Hydrophobic residues tend to have higher prediction accuracies compared to hydrophilic residues. The lower accuracies for β-sheet seem to be related to its poor representation in the data set in comparison with α-helix and coil. A more uniform representation of residues in all the three secondary structures might lead to better secondary structure prediction accuracies. We find that amino acid predictions differ for *Middle* vs. *End* regions in all three secondary structures. Errors in the *Middle* regions occur mostly for a few residues. Errors in predictions occur primarily at the Ends, where only about 70 % of residues are predicted correctly. Some of these statistics can be used to improve prediction accuracies. Those residues with large errors are associated with lower content at the End regions for α-helix and β-sheet but this correspondence does not hold for residues in coil. It might be possible to achieve higher accuracies (if *End* predictions could be improved) by building a training model that has an even representation of all amino acidsin all three secondary structures, with separate consideration of the *Middle* and the *End* regions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

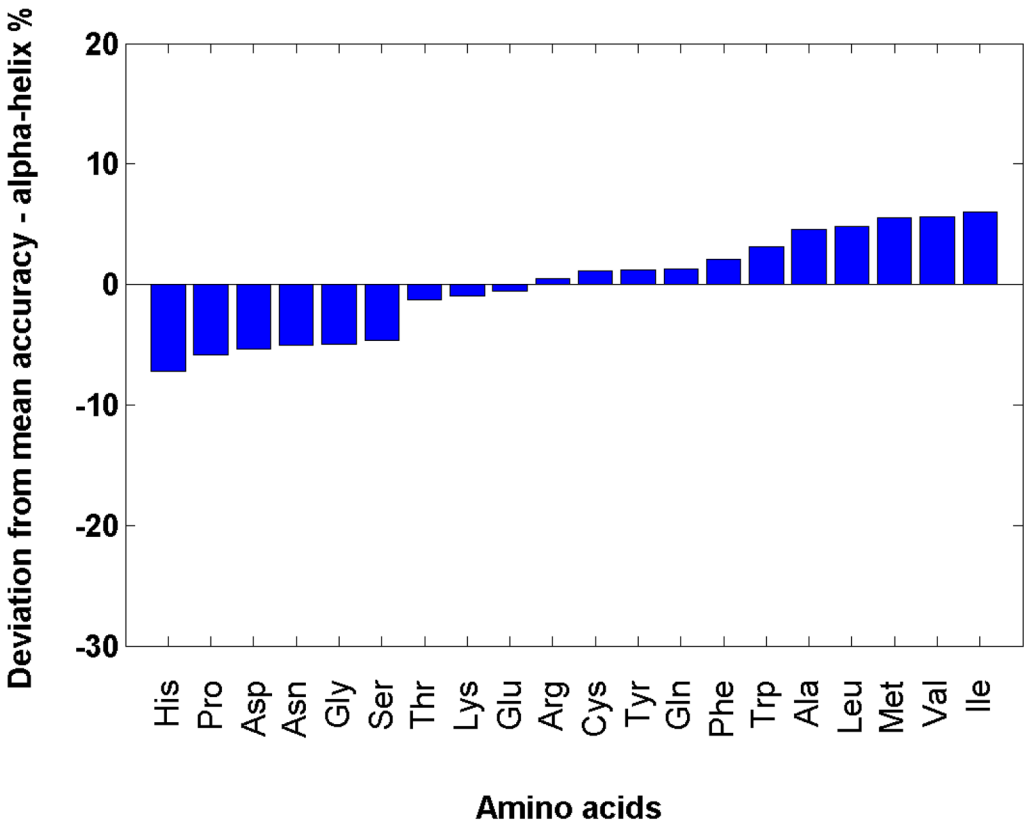
The *algorithm for the knowledge-based potentials data*, was developed by members from the Kolinski [37] lab. This work was supported by the NSF grant IGERT-0504304, NSF grant MSB-1021785 and National Institutes of Health grants R01GM081680 and R01GM072014.

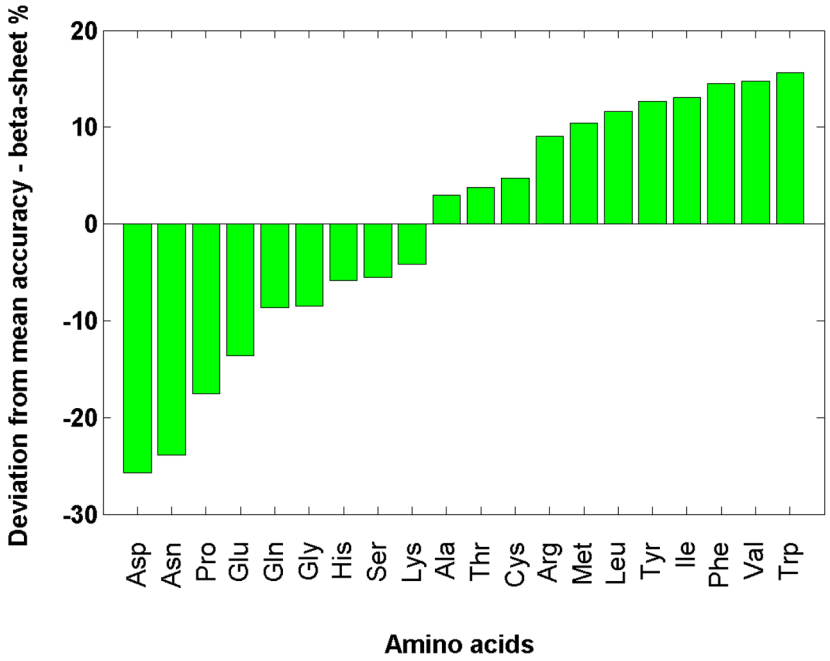
References

1. Saraswathi S, Fernández-Martínez JL, Koliński A, Jernigan RL, Kloczkowski A. JMM. 2012; 18:4275.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. Nucleic Acids Res. 2000; 28:235. [PubMed: 10592235]
3. Qian N, Sejnowski TJ. J Mol Biol. 1988; 202:865. [PubMed: 3172241]
4. Chou PY, Fasman GD. Biochemistry. 1974; 13:222. [PubMed: 4358940]
5. Garnier J, Osguthorpe DJ, Robson B. J Mol Biol. 1978; 1:97. [PubMed: 642007]

6. Garnier J, Gibrat JF, Robson B. Methods Enzymol. 1996; 226:540. [PubMed: 8743705]
7. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. J Mol Biol. 1987; 195:957. [PubMed: 3656439]
8. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Proteins. 2002; 49:154. [PubMed: 12210997]
9. Salzberg S, Cost S. J Mol Biol. 1992; 227:371. [PubMed: 1404357]
10. Yi TM, Lander ES. J Mol Biol. 1993; 232:1117. [PubMed: 8371270]
11. Salamov AA, Solovyev VV. J Mol Biol. 1995; 247:11. [PubMed: 7897654]
12. Salamov VV, Solovyev AA. J Mol Biol. 1997; 268:31. [PubMed: 9149139]
13. Vapnik, VN. The Nature of Statistical Learning Theory (Information Science and Statistics). Springer-Verlag; New York: 2000.
14. Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Bioinformatics. 2003; 19:1650. [PubMed: 12967961]
15. Montgomerie S, Sundaraj S, Gallin W, Wishart D. BMC Bioinformatics. 2006; 301:301. [PubMed: 16774686]
16. Pollastri G, Martin A, Mooney C, Vullo A. BMC Bioinformatics. 2007; 8:201. [PubMed: 17570843]
17. Wang G, Zhao Y, Wang D. Neurocomputing. 2008; 72:262.
18. Malekpour SA, Naghizadeh S, Pezeshk H, Sadeghi M, Eslahchi C. Mathematical Biosciences. 2009; 217:145. [PubMed: 19046975]
19. Palopoli L, Rombo SE, Terracina G, Tradigo G, Veltri P. Information Fusion. 2009; 10:217.
20. Santiago-Gómez MP, Kermasha S, Nicaud JM, Belin JM, Husson F. J Mol Catal B-Enzym. 2010; 65:63.
21. Yang B, Wei H, Zhun Z, Huabin Q. Expert Syst Appl. 2009; 36:9000.
22. Zhou Z, Yang B, Hou W. Expert Syst Appl. 2010; 37:6381.
23. Babaei S, Geranmayeh A, Seyyedsalehi SA. Comput Meth and Prog Bio. 2010; 100:237.
24. Rost B, Sander C. J Mol Biol. 1993; 232:584. [PubMed: 8345525]
25. Rost B. Methods Enzymol. 1996; 266:525. [PubMed: 8743704]
26. Cuff JA, Barton GJ. Proteins. 2000; 40:502. [PubMed: 10861942]
27. Cheng H, Sen TZ, Jernigan RL, Kloczkowski A. Bioinformatics. 2007; 23:2628. [PubMed: 17660202]
28. Cheng H, Sen TZ, Kloczkowski A, Margaritis D, Jernigan R. Polymer. 2005; 46:4314. [PubMed: 19081746]
29. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A. Bioinformatics. 2005; 21:2787.10.1093/bioinformatics/bti408 [PubMed: 15797907]
30. Sen TZ, Cheng H, Kloczkowski A, Jernigan R. Prot Sci. 2006; 15:2499.
31. Rost B, Yachdav G, Liu J. Nucleic Acids Res. 2004; 32:W321. [PubMed: 15215403]
32. Eddy SR. Bioinformatics. 1998; 14:755. [PubMed: 9918945]
33. Jones D. J Mol Biol. 1999; 292:195. [PubMed: 10493868]
34. Kihara D. Protein Science. 2005; 14:1955. [PubMed: 15987894]
35. Madera M, Calmus R, Thiltgen G, Karplus K, Gough J. Bioinformatics. 2010; 26:596. [PubMed: 20130034]
36. Yang B, Wu Q, Ying ZSH. Knowl-Based Syst. 2011; 24:304.
37. Koliński A. ACTA Biochem Pol. 2004; 51:349.
38. Huang GB, Zhu QY, SCK. Neurocomputing. 2006; 70:489.
39. Saraswathi S, Jernigan RL, Koliński A, Kloczkowski AP. IJCCI/ICNC. 2010:370–375.
40. Suresh S, Saraswathi S, Sundararajan N. EAAI. 2010; 23:1149.
41. Kennedy J, Eberhart RC. P ICNN. 1995; 4:1942.
42. Fernández-Martínez JL, García-Gonzalo E. JAEA. 2008; 2008:15.
43. Fernández-Martínez JL, García-Gonzalo E, Fernández-Alvarez JP. IJCIR. 2008; 4:93.
44. García-Gonzalo E, Fernández-Martínez JL. P ICCMS. 2009:1280–1290.

45. Fernández-Martínez JL, García-Gonzalo EP. IJCCI/ICNC. 2010:237–242.
46. Fernández-Martínez JL, García-Gonzalo E. IEEE Trans Evol Comput. 2011; 15:405.
47. Fernández-Martínez JL, García-Gonzalo E, Saraswathi S, Jernigan RL, Kloczkowski A. ASILNCS. 2011; 6728:1.
48. Rost B, Sander C. Proteins. 1994; 20:216. [PubMed: 7892171]
49. Zemla A, Venclovas e, Fidelis K, Rost B. Proteins: Struct, Funct, Bioinf. 1999; 34:220.
50. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. J Comput Chem. 2012; 33(3):259. [PubMed: 22045506]
51. Kazemian M, Moshiri B, Nikbakht H, Lucas C. Computational Biology and Chemistry. 2007; 31:44. [PubMed: 17270498]
52. Costantini S, Colonna G, FAM. BBRC. 2006; 342:441–451. [PubMed: 16487481]
53. Kabsch W, Sander C. Biopolymers. 1983; 22:2577. [PubMed: 6667333]
54. Needleman SB, Wunsch CD. J Mol Biol. 1970; 48:443. [PubMed: 5420325]
55. Henikoff S, Henikoff J. Proc Natl Acad Sci U S A. 1992; 89:10915. [PubMed: 1438297]
56. Sander C, Schneider R. Proteins. 1991; 9:56. [PubMed: 2017436]
57. Fernández-Martínez JL, García-Gonzalo E. Swarm Intell: Spec Publ PSO. 2009; 3:245.
58. Kyte J, Doolittle RF. JMB. 1982; 157:105.





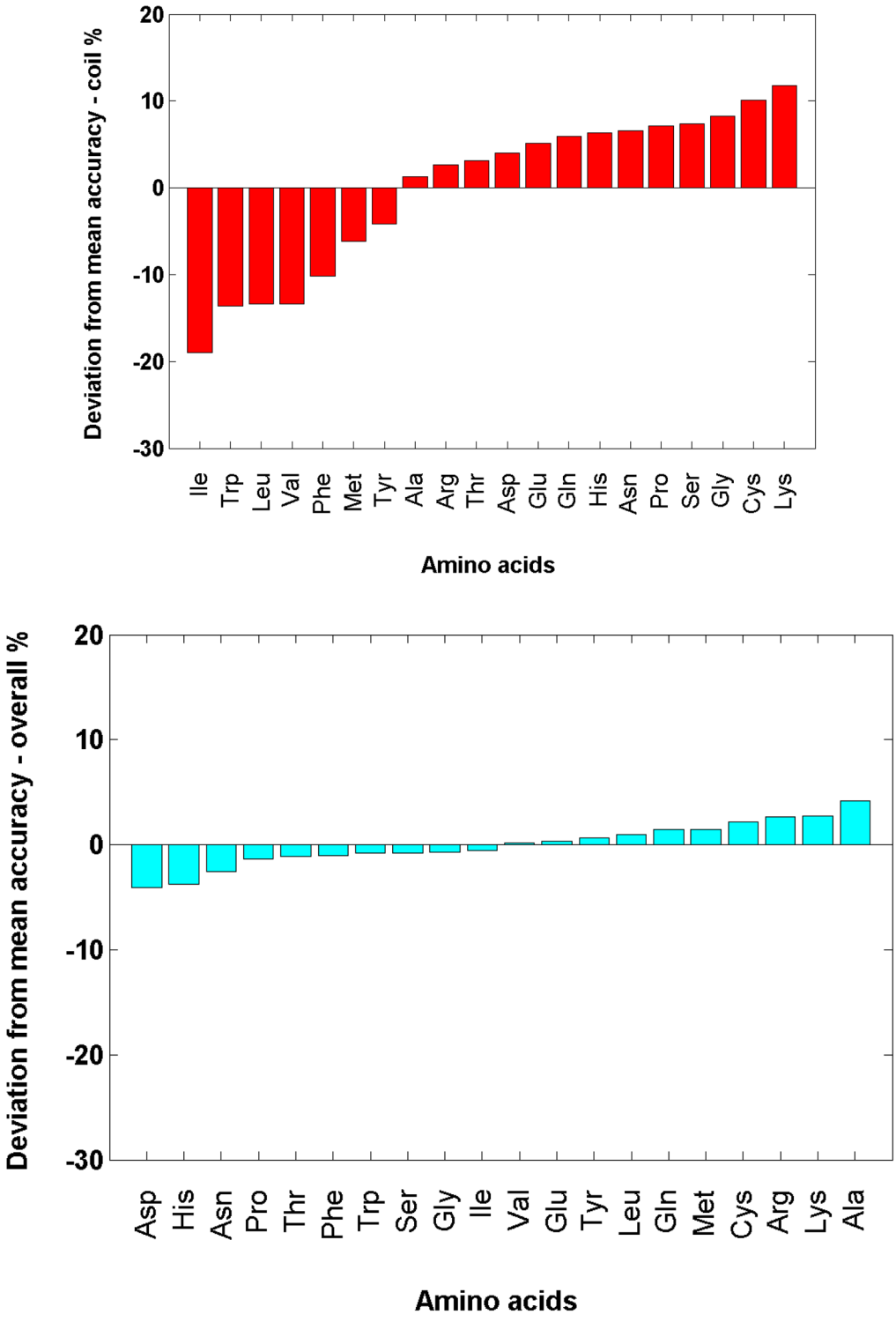
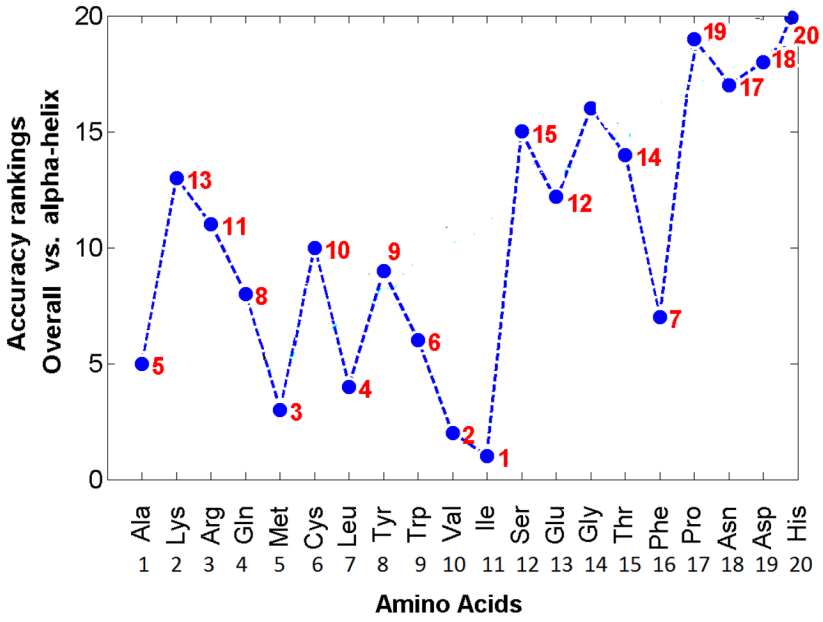
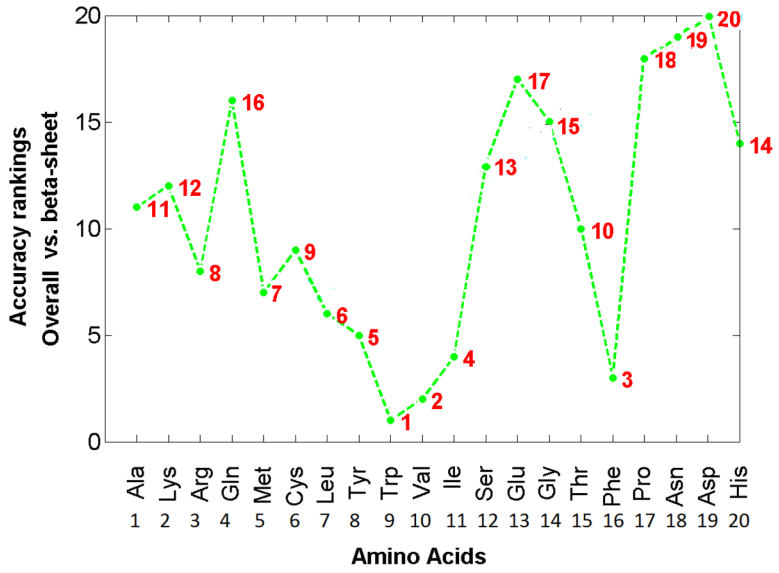


Fig. 1.

Accuracies- as deviations from the mean for overall and secondary structure average accuracies. Amino acids are listed according to their deviation values (from the largest negative deviation to the highest positive deviation), for all figures. There seems to be some

clear differences between hydrophobic and hydrophilic residues in each of the individual secondary structure types shown in Figs. 1a, 1b and 1c. The amino acid deviations from mean accuracy (91.1 % for α-helix, 71.9 % β-sheet and 78.2 % for coil) differ widely for different secondary structures. Fig. 1d plots the overall mean deviations from the mean accuracy of 83.8 % for all amino acids. Hydrophilic residues seem to have larger negative deviations compared to hydrophobic residues. The overall deviations compensate for the complimentary differences in the individual secondary structures and show much lower overall deviations for all residues, which could be misleading. The overall accuracies appear to be higher than they actually are for the individual secondary structures.





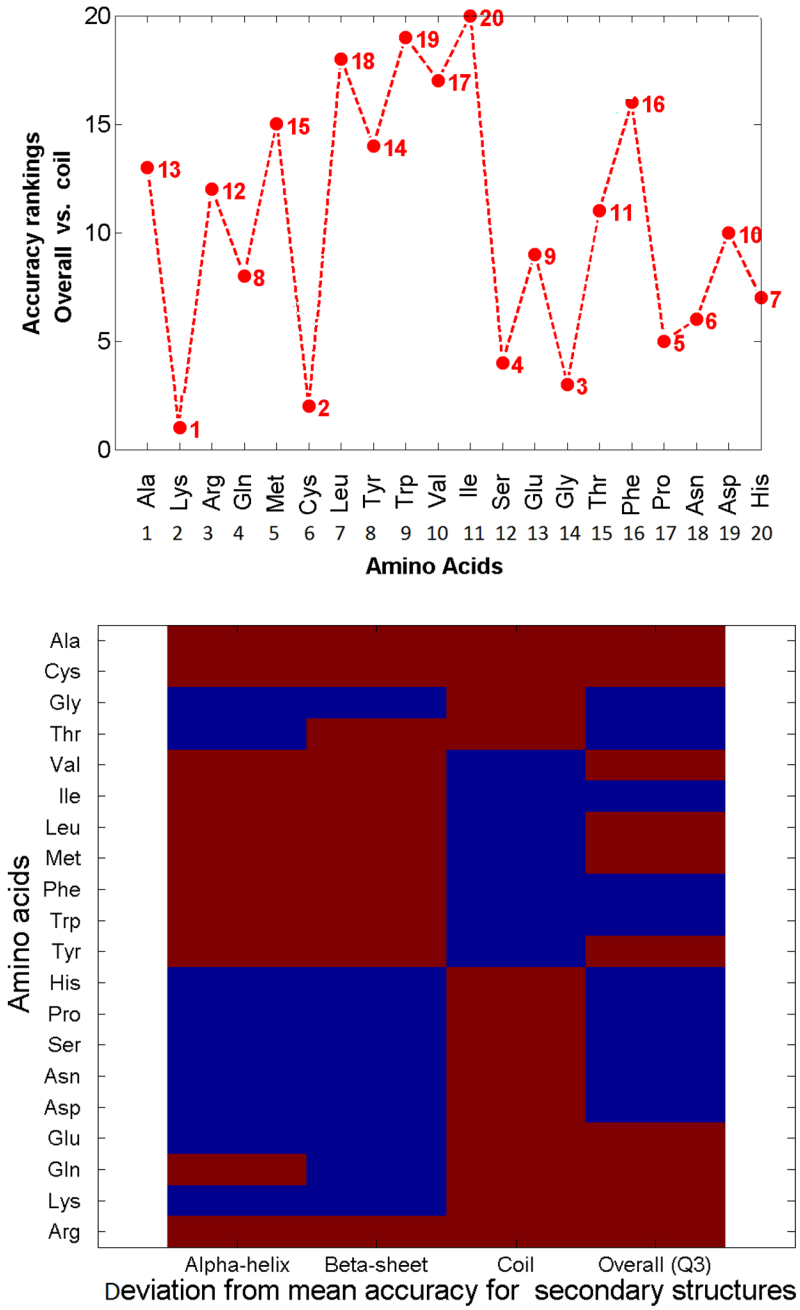


Fig. 2.

Trends in accuracy rankings for secondary structures are given here. Amino acids are listed according to their overall ranking for Figs. 2a, 2b and 2c. The accuracies obtained for each amino acid in α-helix, β-sheet and coil (Figs. 2a, 2b and 2c) are ranked from 1 (highest accuracy) to 20 (lowest accuracy). These rankings are shown as blue (α-helix), green (β-sheet) and red (coil) lines, with the ranks for this particular secondary structure indicated in red to the right of each point marking an amino acid. The overall ranking for each amino acid is given on the x-axis (ranks 1 to 20), which differ a lot from the rankings given inside each figure. Fig. 2d shows the deviations from the overall mean accuracy of 83.8 % for

amino acids; In Fig. 2d, hydrophobic amino acids are listed first followed by hydrophilic residues (see results section for details). Positive deviations are marked in red and the negative deviations are marked in blue. Positive trends (higher accuracies) for α-helix (2a) and β-sheet (2b) are at least partially offset by opposite tendencies for the same residues in coil (as can be seen in Fig. 2d for residue types Val, Ile, Leu, Met, Phe, Trp, Tyr, His, Pro, Ser, Asn, Asp, Glu, Lys and Gly). Consequently, *the final overall accuracies for each of the twenty amino acids do not reflect their individual accuracy characteristics in each of the three secondary structure states*. One way to immediately improve secondary structure predictions would be to introduce amino acid specific weights according to the values seen in Fig. 2d. The amino acids predicted in a particular secondary structure having red values in this figure would receive stronger weights than the ones that are in blue.

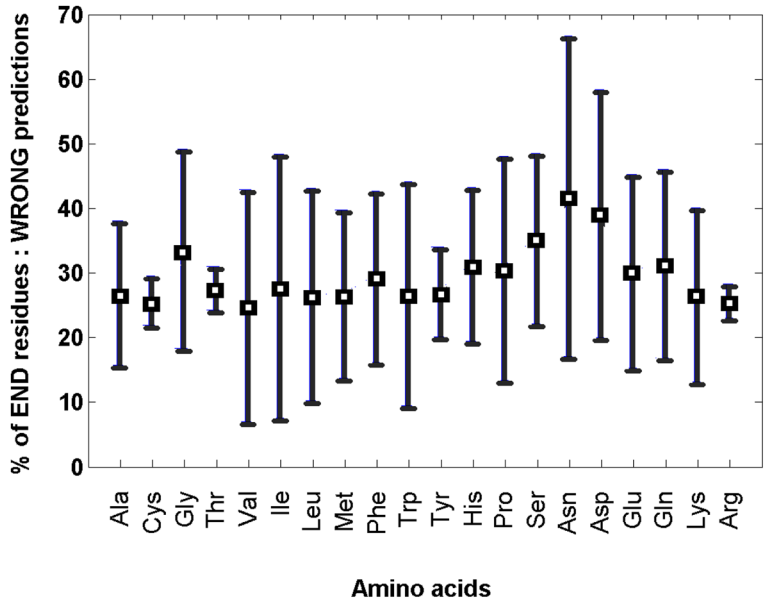


Fig. 3.
Errors at the *Ends* of secondary structures

Table 1

Overall training validation and testing accuration and SOV accuracies and SOV scores for secondary structures on five-fold-cross-validation

Model	α-helix	β-sheet	Coil	Overall
Training-Q_3	91.3	78.3	81.2	84.0
Training-SOV	86.0	77.8	75.2	79.4
Validation-Q_3	90.5	75.0	79.1	81.8
Validation-SOV	83.8	76.8	71.1	76.3
Testing-Q_3	91.1	71.9	78.2	83.8
Testing-SOV	85.8	77.5	71.8	78.3
0.1em.8em.8em				

Table 2

Comparison of FLOPRED predictions against other secondary structure predictions in the literature that have used multiple sequence alignments (MSA) with the CB513 dataset (* with the exception of the PHD method which used the RS126 set consisting of 126 proteins [24] and the SPINE X [50] server which used a dataset of 1833 proteins). FLOPRED makes use of only sequence and knowledge-based potential information. FLOPRED had results much better than the best results from the literature for those studies that did not include MSA (not given) and comparable with the results for those studies which do include MSA (as listed in this table), although it may not be entirely fair to compare results from different datasets

Method	Q_H%	Q_E %	Q_C %	Q_3 %	SOV %
PHD Expert [24]	78.9	73.3	78.8	77.6	75.0*
GOR V [8]	74.0	50.6	82.1	73.4	70.8
J Net [26]	78.4	63.9	80.6	76.4	74.2
PSIPRED [33]	83.5	70.3	83.8	80.0	76.5
SPINE X [50]	87.1	71.8	83.0	82.1	79.0*
CPM[]	87.6	77.7	87.4	85.6	79.8
FLOPRED	91.1	71.9	78.2	83.8	78.3

J Mol Model. Author manuscript; available in PMC 2014 October 01.

Table 3

Prediction accuracies for five-fold cross-validation are given, for each of the twenty amino acids for each secondary structure, along with average prediction accuracy. Hydrophobic amino acids are listed first followed by hydrophilic residues (see results section for details). Overall and average content and their variability is given. The average variability is calculated across the three secondary structures for a particular residue, while the overall variability is calculated with respect to all amino acids in all secondary structures (over the complete data set). The difference in overall and average accuracy shows the contribution of each amino acid to the discrepancy in these two values. Higher variability in amino acid representation leads to increasing difference between average and overall accuracy, as shown in Fig. S4. As the variability increases, the difference in accuracies also increase (see Fig. S5). The last column gives the overall accuracy ranking (where 1 is the highest and 20 is the lowest)

AA	Overall AA content %	α-helix Accuracy % (content %)	β-sheet Accuracy % (content %)	Coil Accuracy % (content) %)	Average Accuracy %) (variability)	Overall Accuracy (variability)	Difference between Accuracy %	Overall Rank
Ala	8.4	95.8 (56.5)	74.8 (15.0)	79.5 (28.5)	83.36 (17.3)	88.02 (0.045)	4.66	1
Cys	1.5	96.8 (33.3)	86.6 (39.2)	64.9 (27.5)	85.76 (4.8)	86.02 (0.002)	0.27	6
Gly	7.6	96.7 (44.7)	82.3 (20.8)	72.0 (34.5)	78.69 (9.8)	83.10 (0.079)	4.42	14
Thr	5.4	90.6 (55.0)	58.3 (13.0)	83.4 (32.0)	82.27 (17.2)	82.69 (0.011)	0.43	15
Val	7.1	91.7 (51.3)	81.0 (15.5)	80.9 (33.2)	82.75 (14.6)	84.02 (0.003)	1.27	10
Ile	5.6	94.3 (44.2)	87.5 (22.4)	64.6 (33.3)	80.47 (8.9)	83.24 (0.006)	2.77	11
Leu	9.3	92.4 (40.5)	84.6 (28.5)	74.0 (31.0)	81.45 (5.2)	84.78 (0.025)	3.33	7
Met	1.9	85.3 (20.7)	54.4 (9.7)	85.4 (69.6)	83.66 (26.0)	85.23 (0.015)	1.57	5
Phe	3.9	86.5 (35.7)	66.3 (15.1)	85.6 (49.2)	82.57 (14.0)	82.81 (0.006)	0.24	16
Trp	1.3	92.4 (52.5)	63.2 (15.9)	84.1 (31.6)	82.12 (15.0)	82.98 (0.014)	0.86	9
Tyr	3.8	90.2 (47.2)	67.7 (16.2)	90.0 (36.6)	83.66 (12.8)	84.47 (0.004)	0.82	8
Pro	4.7	86.1 (34.5)	48.0 (10.9)	84.8 (54.6)	75.02 (17.8)	82.42 (0.103)	7.40	17
Ser	6.3	85.8 (33.7)	46.1 (10.3)	82.2 (56.0)	79.49 (18.7)	83.04 (0.030)	3.56	12
His	2.3	93.2 (40.0)	86.4 (25.5)	68.1 (34.5)	78.20 (6.0)	80.04 (0.011)	1.85	20
Asn	4.7	86.2 (20.2)	63.4 (14.4)	86.5 (65.4)	72.98 (22.8)	81.29 (0.048)	8.31	18
Asp	6.1	89.9 (32.0)	75.6 (23.9)	81.3 (44.1)	71.39 (8.3)	79.76 (0.053)	8.38	19
Glu	6.0	96.0 (51.1)	83.5 (21.5)	64.8 (27.5)	77.42 (12.8)	84.12 (0.045)	6.71	13
Gln	3.9	97.2 (40.8)	85.0 (32.9)	59.3 (26.3)	79.93 (6.0)	85.23 (0.035)	5.30	4
Lys	6.0	84.0 (32.2)	66.1 (23.6)	84.5 (44.2)	82.66 (8.4)	86.52 (0.025)	3.86	2
Arg	4.3	92.3 (31.2)	76.6 (30.1)	88.4 (38.7)	84.52 (3.8)	86.47 (0.033)	1.95	3
Mean	5.0	91.1 (39.9)	71.9 (20.1)	78.2 (40.0)	80.4 (12.5)	83.8 (0.003)	3.4	

AA	Overall AA content %	α-helix Accuracy % (content %)	β-sheet Accuracy % (content %)	Coil Accuracy % (content) %)	Average Accuracy %) (variability)	Overall Accuracy (variability)	Difference between Accuracy %	Overall Rank
Stdev	2.1	4.3 (10.0)	13.2 (8.2)	9.3 (12.9.)	3.9 (6.3)	2.1 (0.003)	2.7	
0.1em.8em.8em								

Table 4

Errors- in the *Middle* and *Ends* of secondary structures are shown below. Hydrophobic amino acids are listed first followed by hydrophilic amino acids (see results section for details). There is a consistently and strikingly higher error rate for predictions at the ends of secondary structure segments for all amino acids in all secondary structure types, while there are very few errors that occur in the *Middle*

AA	% Errors *in the middle'*				% Errors *'at the ends'*			
	α- helix	*β*- sheet	Coil	Q_3	*α*-helix	*β*-sheet	Q_3	Coil
Ala	1.5	1.5	6.9	3.3	98.5	98.5	93.1	96.7
Cys	2.9	8.0	1.5	4.1	97.1	92.0	98.5	95.9
Gly	7.8	8.1	3.4	6.4	92.2	91.9	96.6	93.6
Thr	9.8	4.4	7.6	7.3	90.2	95.6	92.4	92.7
Val	9.2	2.5	17.2	9.6	90.8	97.5	82.8	90.4
Ile	0.0	2.7	11.0	4.6	100.0	97.3	89.0	95.4
Leu	0.2	4.7	13.2	6.1	99.8	95.3	86.8	93.9
Met	0.0	0.0	4.7	1.6	100.0	100.0	95.3	98.4
Phe	2.5	0.0	21.4	8.0	97.5	100.0	78.6	92.0
Trp	0.0	2.8	16.5	6.4	100.0	97.2	83.5	93.6
Tyr	0.0	0.0	20.1	6.7	100.0	100.0	79.9	93.3
Pro	0.0	7.7	16.7	8.1	100.0	92.3	83.3	91.9
Ser	1.7	5.1	6.3	4.4	98.3	94.9	93.7	95.6
His	11.4	0.2	5.3	5.6	88.6	99.8	94.7	94.4
Asn	3.7	3.8	2.7	3.4	96.3	96.2	97.3	96.6
Asp	0.3	4.6	4.7	3.2	99.7	95.4	95.3	96.8
Glu	5.0	15.3	9.0	9.7	95.0	84.7	91.0	90.3
Gln	15.5	4.2	0.0	6.5	84.5	95.8	100.0	93.5
Lys	2.4	5.9	8.6	5.6	97.6	94.1	91.4	94.4
Arg	6.9	7.4	10.4	8.2	93.1	92.6	89.6	91.8
Mean	4.0	4.4	9.4	5.9	96.0	95.6	90.6	94.1
0.1em.8em.8em								

Table 5

Amino acid content at the *Ends* with their standard deviations across the three secondary structures (var) is shown here. Hydrophobic amino acids are listed first followed by hydrophilic amino acids (see results section for details). The content values reflect some well-known characteristics. For example, turns that are considered to be coil are known to be particularly enriched in Gly, Pro, Ser, Asn and Asp, which are seen to have the highest values here. The values inside the brackets show the error % at the Ends of secondary structures and their standard deviations across the three secondary structures in the last column

AA	% Content and Errors in the *End* regions			
	α-helix (error)	*β*-sheet (error)	Coil (error)	Var (error var)
Ala	36.5 (14.2)	22.4 (36.2)	41.2 (29.0)	9.8 (11.2)
Cys	17.1 (25.8)	43.1 (28.8)	39.7 (21.3)	14.1 (3.8)
Gly	12.3 (32.4)	15.3 (49.1)	72.4 (18.4)	33.9 (15.4)
Thr	20.7 (24.0)	30.8 (30.8)	48.5 (26.9)	14.1 (3.4)
Val	17.5 (10.1)	49.9 (18.7)	32.7 (44.7)	16.2 (18.0)
Ile	19.7 (11.2)	44.1 (21.0)	36.2 (50.5)	12.5 (20.5)
Leu	33.5 (11.9)	29.3 (22.5)	37.2 (44.3)	4.0 (16.5)
Met	23.5 (13.1)	28.0 (26.7)	48.6 (39.1)	13.4 (13.0)
Phe	24.9 (22.0)	34.6 (20.7)	40.6 (44.2)	7.9 (13.2)
Trp	28.2 (17.2)	32.2 (15.5)	39.6 (46.3)	5.8 (17.3)
Tyr	26.1 (23.0)	38.0 (22.3)	35.8 (34.6)	6.3 (6.9)
Pro	25.2 (19.0)	12.6 (50.3)	62.2 (21.6)	25.8 (17.4)
Ser	29.1 (31.0)	18.4 (49.6)	52.4 (24.1)	17.4 (13.2)
His	25.2 (27.8)	27.5 (44.0)	47.3 (21.0)	12.2 (11.8)
Asn	23.4 (33.4)	13.2 (69.3)	63.4 (21.6)	26.5 (24.8)
Asp	26.3 (29.5)	14.0 (60.8)	59.7 (26.0)	23.6 (19.2)
Glu	45.1 (19.6)	17.3 (47.1)	37.6 (23.0)	14.4 (15.0)
Gln	35.4 (19.9)	24.2 (47.5)	40.4 (25.7)	8.3 (14.6)
Lys	36.2 (22.9)	22.0 (41.0)	41.8 (14.8)	10.2 (13.4)
Arg	35.1 (22.8)	19.4 (27.9)	45.5 (25.0)	13.1 (2.6)
Mean	27.1 (21.5)	26.8 (36.5)	46.1 (30.1)	14.5 (13.6)
0.1em.8em.8em				