ISSN 0006-2979, Biochemistry (Moscow), 2010, Vol. 75, No. 2, pp. 192-200. © Pleiades Publishing, Ltd., 2010.
Published in Russian in Biokhimiya, 2010, Vol. 75, No. 2, pp. 236-246.
Originally published in Biochemistry (Moscow) On-Line Papers in Press, as Manuscript BM09-196, January 17, 2010.

Statistical Analysis of Unstructured Amino Acid Residues in Protein Structures

M. Yu. Lobanov, S. O. Garbuzynskiy, and O. V. Galzitskaya*

Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia; fax: (495) 632-7871; E-mail: ogalzit@vega.protres.ru

Received July 2, 2009
Revision received August 13, 2009

Abstract—We have performed a statistical analysis of unstructured amino acid residues in protein structures available in the databank of protein structures. Data on the occurrence of disordered regions at the ends and in the middle part of protein chains have been obtained: in the regions near the ends (at distance less than 30 residues from the N- or C-terminus), there are 66% of unstructured residues (38% are near the N-terminus and 28% are near the C-terminus), although these terminal regions include only 23% of the amino acid residues. The frequencies of occurrence of unstructured residues have been calculated for each of 20 types in different positions in the protein chain. It has been shown that relative frequencies of occurrence of unstructured residues of 20 types at the termini of protein chains differ from the ones in the middle part of the protein chain; amino acid residues of the same type have different probabilities to be unstructured in the terminal regions and in the middle part of the protein chain. The obtained frequencies of occurrence of unstructured residues in the middle part of the protein chain have been used as a scale for predicting disordered regions from amino acid sequence using the method (FoldUnfold) previously developed by us. This scale of frequencies of occurrence of unstructured residues correlates with the contact scale (previously developed by us and used for the same purpose) at a level of 95%. Testing the new scale on a database of 427 unstructured proteins and 559 completely structured proteins has shown that this scale can be successfully used for the prediction of disordered regions in protein chains.

DOI: 10.1134/S0006297910020094

Key words: unstructured regions, intrinsically disordered regions, natively unfolded proteins, globular proteins, stability

Prediction of protein structure and function is one of the general directions in structural genomics. Of special interest is prediction of the so-called disordered regions of protein chains (regions having no fixed spatial structure in the native state). Such disordered regions often play an important functional role (see reviews [1, 2]). Disordered regions are structured only when they bind to other molecules (for example, the CREB–CBP complex [3], where CREB is Cyclic AMP Response Element Binding Protein, and CBP is CREB Binding Protein) or under changing conditions of the biochemical medium [4]. Disordered regions of protein chains often cause complexities upon expression, purification, and crystallization of such proteins. It is assumed that the absence of globular structures under physiological conditions is an essential functional advantage for natively unfolded proteins because their large accessible surface for small protein size and their flexibility allow them to more effectively interact with proteins and nuclear acids as compared to globular proteins of the same size with confined conformational flexibility [1, 5].

More than 500 proteins with disordered regions are now known [6]. These proteins and domains are either entirely unstructured in the native state (so-called natively unfolded proteins) or have lengthy disordered regions. It turns out that functionally important protein regions in such proteins are often situated outside globular domains, i.e. just in the disordered regions [4, 6].

Since disordered regions of the protein chain play an important role in protein functioning, much attention is being given to their prediction. At present, special programs such as FoldUnfold, PONDR, RONN, DISOPRED, PreLINK, IUPred, GlobPlot, FoldIndex, and others are available for this purpose. The programs can be separated in two groups according to the principle of their operation. Programs FoldUnfold, PONDR, IUPred, GlobPlot, PreLINK, and FoldIndex predict unstructured regions of the protein chain based on physicochemical

* To whom correspondence should be addressed.

properties of amino acids in proteins. Such properties can include local amino acid composition and hydrophobicity (PONDR) [7, 8], number of expected contacts (FoldUnfold) [9-11], propensity of a chain region to form a hydrophobic cluster (PreLINK) [12], or estimation of the energy interaction between neighboring amino acids (IUPred) [13, 14]. The GlobPlot program estimates the tendency of residues to be present in a regular secondary structure [15]. The FoldIndex program is based on a specially developed charge/hydrophobicity scale for amino acid residues [16]. The other group of programs uses alignments of homologous protein sequences. The RONN program uses a neural network and compares the given sequence with a number of sequences whose structure can be *a priori* determined as ordered, disordered, or a mixture of such structures [17]. DISOPRED is a method using the network trained in such a way as to distinguish regions that are missed in the structure obtained by X-ray analysis [18].

It has been shown that unstructured proteins (or disordered regions in globular proteins) are enriched with the following amino acid residues: Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys [19-22]. Of interest is the scale TOP-IDP, which is a result of consideration of 517 different scales elaborated for predicting disordered regions in protein chains [23].

It is shown herein that the new scale obtained by us based on statistics of unstructured residues in the Protein Data Bank can be used for prediction of disordered regions in a protein chain with help of our earlier method FoldUnfold. It is interesting that the two scales obtained from different statistics (statistics of contacts in globular structures and statistics of unstructured residues in the Protein Data Bank) correlate at a level of 95%.

METHODS OF INVESTIGATION

Creation of database of disordered regions in globular proteins. We have considered all protein structures determined by X-ray analysis with a resolution higher than 3 Å published by December 20, 2008. All 100% homologs were grouped, and the data for them were averaged (all analyzed protein chains with identical amino acid sequences were taken with equal weight). The resulting database includes 95,786 protein chains (among them 28,727 are unique chains, i.e. are not 100% homologs of each other). In total, there are 7,487,366 residues in the unique protein chains, from that 347,872 are unstructured residues, which make 4.65%. We consider residues to be unstructured if they are not resolved by X-ray analysis. To search for such residues, we have compared (for each protein chain) the record SEQRES and the record ATOM in the corresponding PDB file. Residues which are present in the record SEQRES, but their coordinates are absent in the record ATOM (namely, the coordinates of C_α-atom are absent in the record ATOM), are considered as unstructured ones.

Databases of entirely unstructured (natively-unfolded) and entirely structured (folded) proteins. The database including 427 natively unfolded proteins has been compiled using the list of proteins from paper [9]. The database including 129 natively unfolded proteins (http://phys.protres.ru/resources/unfolded_129.html) has been compiled using the list of proteins from work [13]. All amino acid sequences have been taken from the SWISS-PROT database [24].

The database including 559 entirely structured globular proteins (http://phys.protres.ru/resources/folded_559.html) has been created by using PDB codes given in [13].

Observed average number of contacts in globular state at a given distance. The average number of contacts (average environment density) for each of the 20 types of amino acid residues has been taken from our work [10]. The number of contacts (environment density) for amino acid residues for each of the 20 types has been determined from our earlier database of globular proteins [10, 11] with identity not exceeding 80% (the database consisted of 5829 proteins). Amino acid residues are considered as having a contact if at least one pair of their atoms is situated at a distance less than 8 Å. The contacts of adjacent residues in the chain (± 1) are not taken into account because they are covalently connected with each other and therefore have contacts in any conformation of the protein chain. For each amino acid residue, the number of contacts with other residues was calculated. Then the average number of contacts for each of the 20 types of amino acid residues was calculated. These 20 values were used by us as a scale for calculating the expected average number of contacts (close residues) based on the amino acid sequence of the protein [10, 11].

Basic principle of the FoldUnfold program. The FoldUnfold program is accessible at http://calc.protres.ru/ogu/ogu.cgi. The principle of its operation is described elsewhere [9, 10]. Let us remind briefly the main points. Each residue is assigned the expected number of contacts in globular state (the expected number of contacts in globular state is equal to the average number of the observed number of contacts in spatial structures for amino acid residues of a given type) upon prediction of unstructured residues by amino acid sequence. Then averaging of values is done by the region, which is equal to the window width (an adjustable parameter; we used the window size of 41 residues). The resulting average value of expected contacts is ascribed to the central residue in the chosen window. After that, the window is shifted by one residue downstream the amino acid sequence and the procedure is repeated. On the resulting profile of expected contacts, a boundary is marked that separates ordered and disordered regions. A region is considered to be disordered if after averaging all its amino

acid residues have the number of expected contacts less than the chosen threshold value (as in previous works [10, 11] we used 20.4 contacts per residue as a threshold value, which is optimal upon prediction of disordered regions using the protein amino acid sequence) and its size is larger than or equal to the window width of averaging as shown earlier [9-11].

Evaluation of the quality of the prediction of disordered regions. For estimation of the quality of the predictions, standard definitions of sensitivity and specificity have been used [25]:

$$S_n = TP/N_d,$$

$$S_p = TN/N_o.$$

Here S_n is sensitivity, S_p is specificity, TP ("true positives") is the number of correctly predicted unstructured amino acid residues, N_d is the total number of unstructured (disordered) residues, TN ("true negatives") is the number of correctly predicted structured amino acid residues, N_o is the total number of structured (ordered) residues. Thus, sensitivity is a fraction of correctly predicted unstructured residues, and specificity is a fraction of correctly predicted structured residues [25].

Besides, in this work we consider the criterion of evaluation of the quality of prediction that is used in the CASP competition ("Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction" is a competition devoted to the evaluation of the quality of prediction of 3D protein structure) in the category devoted to the evaluation of the quality of prediction of unstructured residues [26, 27] (http://predictioncenter.org/casp8/doc/presentations/CASP8_DR_Sussman.pdf):

$$S_w = \frac{W_1TP - W_2FP + W_2TN - W_1FN}{W_1N_d + W_2N_o},$$

where FP ("false positives") is the number of false positive predictions (the number of residues predicted as unstructured although these residues are in fact structured), FN ("false negatives") is the number of false negative predictions: the number of residues predicted as structured although these residues are in fact unstructured, and W_1 and W_2 are coefficients calculated as follows:

$$W_1 = \frac{N_o}{N} \cdot 100\%, \qquad W_2 = \frac{N_d}{N} \cdot 100\%,$$

($N = N_d + N_o$ is the total number of amino acid residues).

As seen, the equation for calculation of S_w can be rewritten using a smaller number of symbols than that in [26]. Substituting equations instead of W_1 and W_2, we obtain:

$$S_w = \frac{N_o(TP - FN) + N_d(TN - FP)}{2N_dN_o}.$$

Taking into account that $FN = N_d - TP$ and $FP = N_o - TN$, we have:

$$S_w = \frac{N_o(2TP - N_d) + N_d(2TN - N_o)}{2N_dN_o} = \frac{TP}{N_d} + \frac{TN}{N_o} - 1.$$

Or, using the definitions for sensitivity and specificity given above, we obtain:

$$S_w = S_n + S_p - 1.$$

RESULTS AND DISCUSSION

Distribution of unstructured amino acid residues in a protein chain. Fraction of unstructured amino acid residues at different positions of a protein chain. Statistical analysis. To analyze the frequencies of occurrence of unstructured residues in protein structures, we have created a database using all protein structures available in the Protein Data Bank (PDB) by December 20, 2008 (see "Methods of Investigation"). We considered residues to be unstructured if they were not determined by X-ray analysis, or more strictly residues with a non-determined (i.e. having no coordinates) C_α atom. Our database is maximally complete: all protein chains of all structures (determined by X-ray analysis) available in the PDB (by December, 2008) are present in it. This is especially important for the analysis of disordered regions, since different structures of the same protein (or even different protein chains in the same PDB file) can have non-identical disordered regions.

We have analyzed the distribution of unstructured residues in the resulting database. The statistics of occurrence of disordered regions of different length has been calculated. The N-terminal disordered regions and the C-terminal ones, and internal disordered loops (disordered regions at both ends of which there are ordered regions) have been considered separately. The distribution of disordered regions by their lengths is shown in Fig. 1. As seen, the disordered regions one-residue-long occur more frequently at the N- and C-termini of proteins. Disordered regions four residues long occur the most frequently in the middle part of the protein chain.

The statistics of distribution of unstructured residues in protein chains is given in Table 1. It is interesting that 2/3 (66%) of all unstructured amino acid residues are near the termini of protein chains (at a distance less than 30 residues from the N- or C-terminus of the protein chain), although these terminal regions include only 23% of amino acid residues of protein molecules. Therefore, for further study of the occurrence of unstructured residues we

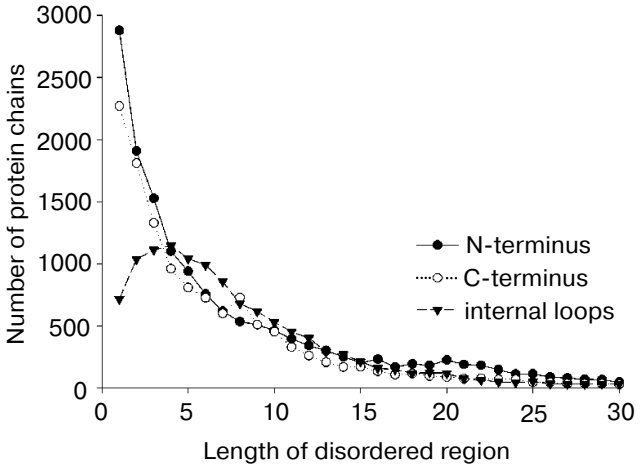

Fig. 1. Length distribution of disordered regions in protein chain (separately for N-terminus, C-terminus, and internal disordered loops).

considered separately the terminal regions (at a distance less than 30 residues from the N- or C-terminus) and the middle part of the protein chain (all the other residues).

We analyzed in more detail the frequency of occurrence of unstructured residues depending on the distance from the N- and C-termini of a protein. The fraction of unstructured residues depending on their position relative to the N- and C-termini of a protein is presented in Fig. 2a. One can see that most of the disordered residues are at the very termini of proteins, the first residue from the N-terminus being unstructured in more than half of cases (the fraction of unstructured residues in this position is 54%), and as the distance from the termini increases the fraction of unstructured residues approaches the value occurring in the middle part of the protein chain – 2.05% (the horizontal dashed line in Fig. 2). At the N- and C-termini, we separated the following regions: the first position from the end (N_1 and C_1, respectively), positions from 2 to 10 ($N_{2\text{-}10}$ and $C_{2\text{-}10}$), positions from 11 to 20 ($N_{11\text{-}20}$ and $C_{11\text{-}20}$), positions from 21 to 30 ($N_{21\text{-}30}$ and $C_{21\text{-}30}$). In such a way, we separated (and considered separately) nine regions (see Fig. 2b). The fraction (the probability of occurrence) of unstructured residues in each of the nine regions $p(region)$ is presented in Fig. 2c.

Then in each region we considered the fraction (the probability of occurrence) of unstructured residues depending on the type $p(type, region)$, and also the normalized fraction of unstructured residues depending on their type:

$$\tilde{p} = \frac{p(type, region)}{p(region)}.$$

Such normalization is necessary to exclude terminal effects to be able to consider the fraction of unstructured residues of a specific type in different regions.

The fractions of unstructured residues (the probabilities of occurrence) for each of the 20 types of amino acid residues in each of the nine mentioned positions are presented in Fig. 3a. As one can see, maximal fractions of unstructured residues are observed for histidine in the four terminal regions (N_1, $N_{2\text{-}10}$, $C_{2\text{-}10}$, C_1), and also for glycine and methionine (for both in region N_1). In the case of histidine, the large bursts in all four regions are connected with poly-histidine tags at the N- and C-termini of protein molecules. If poly-histidine tags are excluded from the database, then all four bursts observed for histidine greatly decrease (the data are not presented). So, histidine residues included in the composition of poly-histidine tags are unstructured more frequently than other histidine residues at the N- and C-termini of proteins.

Then to compare the results of occurrence of unstructured residues at the end and in the middle parts, we normalized the fraction of unstructured residues of each type to the total fraction of unstructured residues for each region (see above). And if the amino acid residue of the given type is unstructured more often than this is usually observed in the given region, then it will have $\tilde{p} > 1$, and *vice versa*. From Fig. 3b, one can see the difference between the relative frequencies of occurrence of unstructured residues of 20 types at the termini of protein chains and in the middle part of protein chains. As before, one can see the bursts connected with poly-histidine tags (His at the N- and C-termini of a protein chain).

Correlation coefficients of the probabilities of occurrence of unstructured residues of 20 types in the middle part of a protein chain and in other regions as well as

Table 1. Distribution of unstructured amino acid residues in protein structures from the Protein Data Bank

		Fraction of all residues, %		Fraction of unstructured residues, %	
Terminal parts	30 residues near the N-terminus	23	11.5	66	38
	30 residues near the C-terminus		11.5		28
Middle part (all other residues)		77		34	

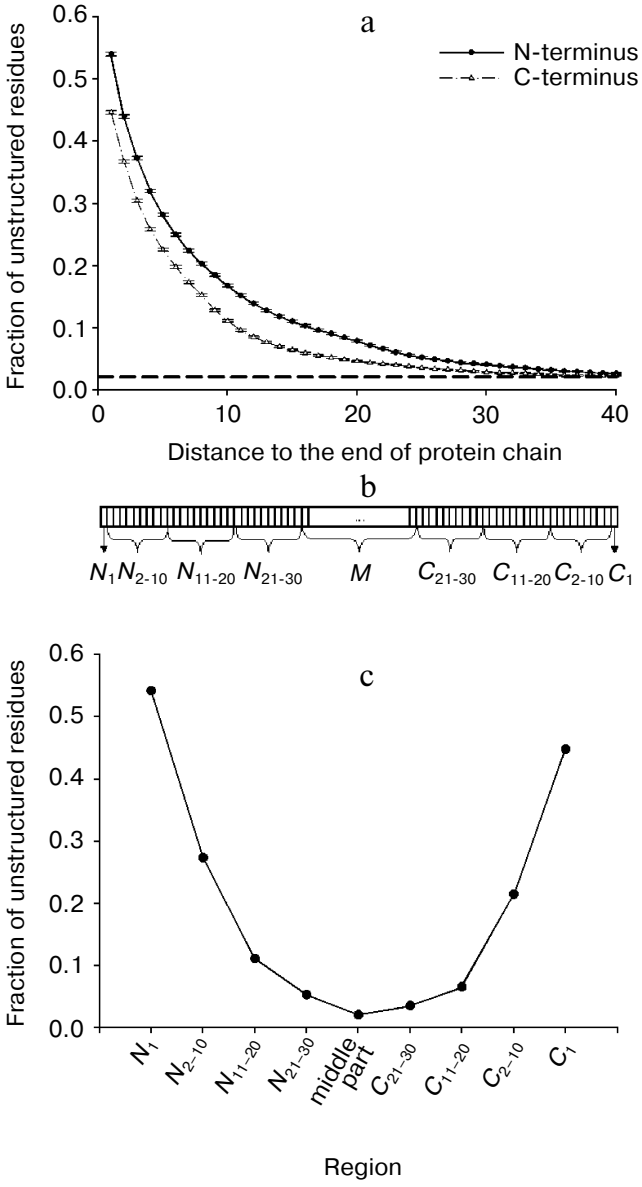

Fig. 2. a) Fraction of unstructured amino acid residues in dependence on the distance to the end of a protein chain. The horizontal line presents the total fraction of unstructured residues in the middle part of the protein chain. b) Scheme illustrating the separation of protein chain into nine regions: N_1, N_{2-10}, N_{11-20}, N_{21-30}; the same for the C-end (C_1, C_{2-10}, C_{11-20}, C_{21-30}); M, the middle part of protein chain. c) Fraction of unstructured amino acid residues in each position mentioned above for nine regions.

through the whole protein are given in Table 2. As one can see, the correlation at the termini is minimal and then (as the center is approached) tends to rise to 100%. The correlation coefficient of the probabilities of occurrence of unstructured residues in the middle part of protein chain and through whole proteins is 69%. It is interesting to note that the general difference in the given case is connected with methionine (if methionine is excluded the correlation coefficient increases to 97%).

The fraction of unstructured amino acid residues for each of the 20 types in the middle part of protein chain is presented in Fig. 4a (region *M* in Fig. 2b). As it is seen from the presented histogram, the fraction of unstructured residues in the middle part of a protein chain varies from 0.01 (for tryptophan) to 0.03 (for serine). As should be expected, the fraction of unstructured amino acid residues is lower for hydrophobic residues and higher for the hydrophilic ones. It is interesting that serine is more often unstructured than any other type of amino acid residues (including glycine and proline which (at least one of them) are usually chosen [9, 15, 23] as the residues with a higher "predisposition" to be in disordered regions). The errors indicated on the histogram shown that the difference is reliable.

In Fig. 4b, the probabilities of occurrence of unstructured residues in the middle part of a protein chain and through whole proteins are presented in the ordinate and in the abscissa, respectively. It can be seen from the figure that all amino acid residues were separated into four groups. The first group (the fraction of unstructured residues in the whole protein is less than 0.03 and in the middle protein part it is less than 0.02) includes mainly hydrophobic amino acid residues (Trp, Phe, Ile, Tyr, Cys, Leu, Val). The second group (the fraction of unstructured residues in the whole protein varies from 0.04 to 0.06) contains hydrophilic and small amino acid residues (besides Ser). The third and fourth groups include one residue (correspondingly Ser and Met). As can be seen from the figure, serine has a high probability to be unstructured both in the middle part of a protein chain and in the whole protein. On the contrary, the probability of methionine to be unstructured in the middle part of

Table 2. Correlation coefficients of probabilities of occurrence of unstructured residues of 20 types in the middle part of a protein chain and in the other regions as well as through the whole protein

Region	Correlation coefficient with the middle part of protein chain, %
N_1	30
N_{2-10}	51
N_{11-20}	65
N_{21-30}	61
Middle part	100
C_{21-30}	97
C_{11-20}	95
C_{2-10}	52
C_1	68
Entire protein	69

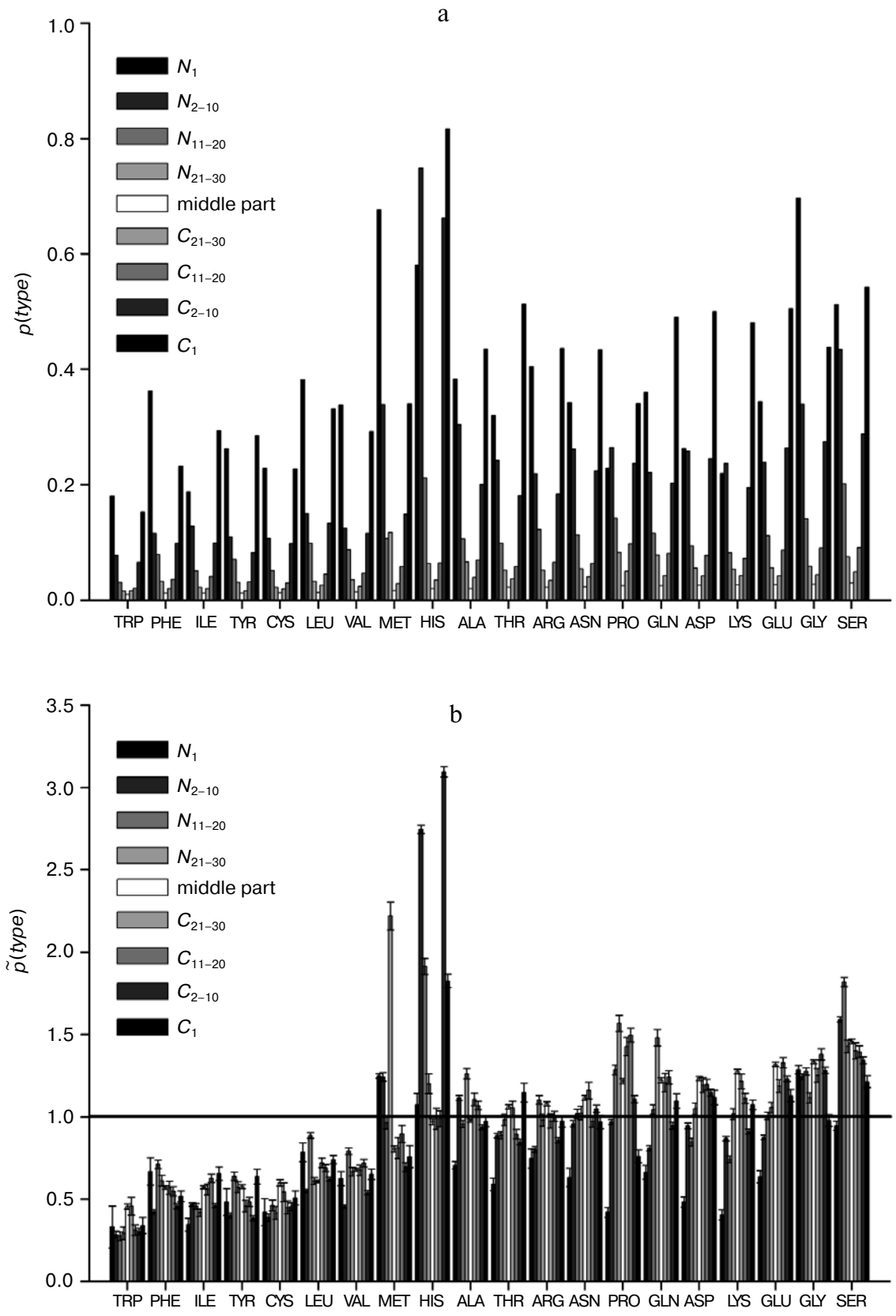

Fig. 3. a) Fraction of unstructured amino acid residues for each of the 20 types in each of nine regions. b) Normalized fraction of unstructured amino acid residues for each of the 20 types in each of nine regions.

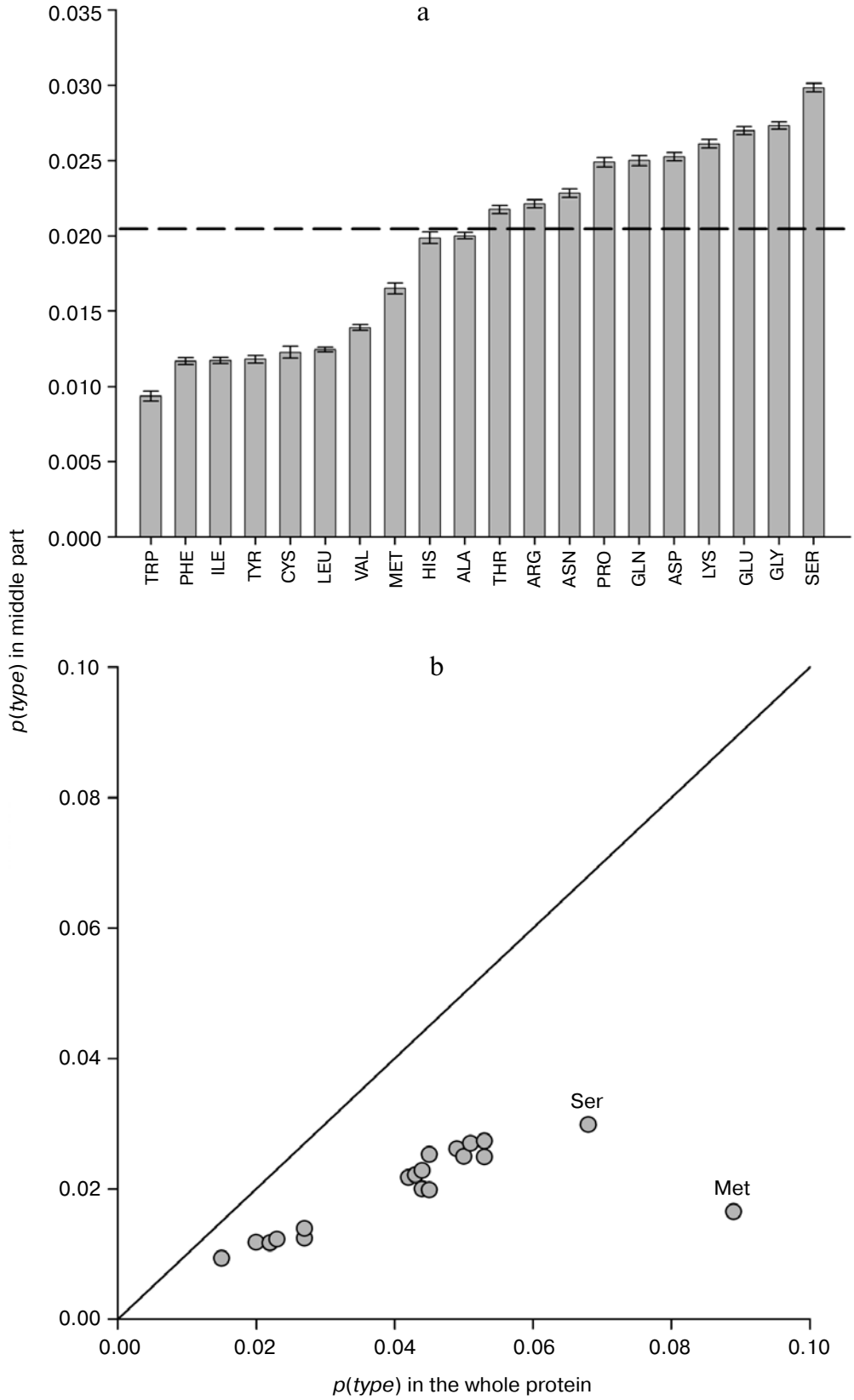

Fig. 4. a) Fraction of unstructured amino acid residues for each of the 20 types in the middle part of a protein chain. b) Fraction of unstructured amino acid residues for each of the 20 types for the whole protein in comparison with the middle part of the protein chain.

a protein chain is only a little higher than that of hydrophobic residues, whereas in the whole protein methionine has the highest probability from the other 20 types to be unstructured (0.089). It is interesting that such a phenomenon is not observed for histidine despite its high probability to be unstructured at the termini of proteins (see Fig. 3) as a result of high disordered poly-histidine tags.

Prediction of disordered regions of a protein chain using the FoldUnfold program. The basic principle of the FoldUnfold program for the prediction of unstructured amino acid residues was based (initial version of the program [9, 10]) on the scale of predicted (expected) contacts, obtained by us [9-11] through an analysis of contacts observed in globular protein structures (see "Methods of Investigation") and used by us for searching regions forming an anomalously small number of contacts. The statistics of occurrence of unstructured residues in the middle part of a protein chain obtained by us could be used also as a scale for prediction of unstructured amino acid residues using the same FoldUnfold program. The comparison of two scales (contact and statistical) has shown that the resulting scale of occurrence of unstructured residues in the middle part of a protein chain correlates with the scale of contacts at about 95%. One can suggest that the new scale will also predict disordered regions rather well if the FoldUnfold program is used. We tested the results of the FoldUnfold program with the new scale on the same databases that had been used earlier [9] for testing the FoldUnfold program with the contact scale: the database of natively unfolded (427 proteins) and globular (559) proteins. Consideration of the two databases allows us to compare the results of our method for the two scales (contact and statistical).

To estimate the fraction of true and false predictions for disordered regions of a protein chain we constructed ROC-curves (Fig. 5) for two scales varying the value of threshold that separates ordered and disordered regions (see "Methods of Investigation"). As seen from the figure, the ROC curve for the statistical scale passes slightly higher than the curve obtained with the contact scale. In other words, the FoldUnfold method works slightly better if the statistical scale is used. The point with the maximal value of S_w has been chosen as the optimal threshold value (see "Methods of Investigation"). The point is indicated by a large symbol in Fig. 5. The maximal value of S_w is 0.74. The position of the border separating ordered and disordered regions is equal to 0.0212; this is slightly higher than the value of $p(region)$ for the middle part of protein chain.

Besides, we have tested the results of the FoldUnfold program with the statistical scale obtained for the whole protein rather than for the middle part of amino acid sequence. In this case, the program works worse (the maximal value of parameter S_w is equal to 0.66).

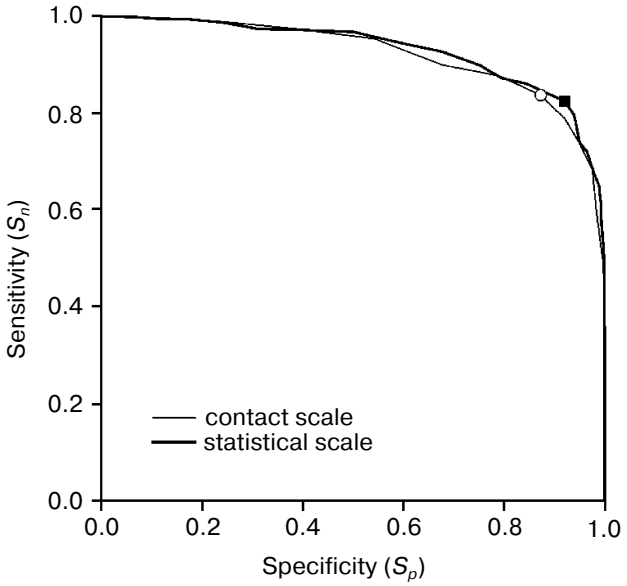

Fig. 5. Sensitivity and specificity for the FoldUnfold method upon varying the value of the threshold separating ordered and disordered regions. The thick curve shows the data obtained using the statistical scale (the middle part of protein chain). The thin curve shows the data obtained using the contact scale. Large symbols indicate the points with the maximal value of S_w.

Comparison of different methods for predicting disordered regions of a protein chain. We compared the predictions of our method with the results of four known methods of predicting disordered regions of a protein chain: GlobProt [15] is a simple approach that estimates the tendency of residues to be involved in a regular secondary structure; PONDR VL3H [28] is a method which utilizes a neuronal network trained to distinguish natively unfolded proteins (whose disordered structure has been verified experimentally) from globular proteins; DISOPRED [18] utilizes a neuronal network trained to identify the regions that have been omitted in the structure obtained by X-ray analysis; IUPred [14] is a method which assigns the structured or unstructured status to a residue on the basis of its capability to form advantageous pairwise contacts. Sensitivity and specificity of predicting disordered regions using protein amino acid sequence for five methods including our FoldUnfold, for two scales (contact and statistical) are presented in Table 3. It is seen from the data presented in the table that our method yields the best results (for the given database of proteins) among the considered methods of prediction of disordered regions using protein amino acid sequence.

Thus, in this work we studied the statistics of unstructured amino acid residues in the Protein Data Bank. It has turned out that 38% of unstructured residues are near the N-terminus of proteins, 28% are near the C-terminus, and the remaining 34% are in the middle part of the protein chain. It has been shown that the relative frequencies of occurrence of unstructured residues at the

Table 3. Comparison of different methods for predicting disordered regions for two databases: database of natively unfolded proteins (129 proteins) and database of globular proteins (559 proteins)

Method	Sensitivity	Specificity	S_w
FoldUnfold (statistical scale)	0.92	0.93	0.85
FoldUnfold (contact scale) [9-11]	0.85	0.95	0.80
IUPred [14]	0.76	0.95	0.71
PONDR VL3H [28]	0.66	0.95	0.61
DISOPRED2 [18]	0.63	0.95	0.58
GlobPlot [15]	0.33	0.82	0.15

termini of protein chains differ from the ones in the middle part of protein chain. It has been shown also that the obtained scale (the fraction of unstructured amino acid residues in the middle part of a protein chain) can be used for the prediction of disordered regions using the FoldUnfold program.

This work was made at a financial support from the Russian Foundation for Basic Research (grant No. 08-04-00561), Russian Academy of Sciences (programs "Molecular and Cell Biology" and "Fundamental Sciences to Medicine"), Howard Hughes Medical Institute (grant No. 55005607), and Russian Science Support Foundation, as well as a grant from the Federal Agency for Science and Innovations (No. 02.740.11.0295).

REFERENCES

1. Tompa, P. (2002) *Trends Biochem. Sci.*, **27**, 527-533.
2. Wright, P. E., and Dyson, H. J. (1999) *J. Mol. Biol.*, **293**, 321-331.
3. Radhakrishnan, I., Perez-Alvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., and Wright, P. E. (1997) *Cell*, **91**, 741-752.
4. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) *Biochemistry*, **41**, 6573-6582.
5. Dyson, H. J., and Wright, P. E. (2002) *Adv. Protein Chem.*, **62**, 311-340.
6. Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., and Dunker, A. K. (2007) *Nucleic Acids Res.*, **35**, D786-793.
7. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, L. E., and Dunker, A. K. (1997) *Proc. IEE Int. Conf. on Neural Networks*, pp. 90-95.
8. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, A. Z. (1999) *Genome Inform.*, **10**, 30-40.
9. Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) *PLoS Comput. Biol.*, **2**, 1639-1648.
10. Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Yu. (2006) *Bioinformatics*, **22**, 2948-2949.
11. Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) *Mol. Biol.* (Moscow), **40**, 341-348.
12. Coeytaux, K., and Poupon, A. (2005) *Bioinformatics*, **21**, 1891-1900.
13. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) *J. Mol. Biol.*, **347**, 827-839.
14. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) *Bioinformatics*, **21**, 3433-3434.
15. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) *Nucleic Acids Res.*, **31**, 3701-3708.
16. Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I., and Sussman, J. L. (2005) *Bioinformatics*, **21**, 3435-3438.
17. Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005) *Bioinformatics*, **21**, 3369-3376.
18. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004) *Bioinformatics*, **20**, 2138-2139.
19. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) *J. Mol. Graph. Model.*, **19**, 26-59.
20. Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., and Dunker, A. K. (2007) *Biophys. J.*, **92**, 1439-1456.
21. Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J., and Dunker, A. K. (2001) *Pac. Symp. Biocomput.*, 89-100.
22. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) *Proteins*, **42**, 38-48.
23. Campen, A., Williams, R. M., Brown, C. J., Meng, J., Uversky, V. N., and Dunker, A. K. (2008) *Protein Pept. Lett.*, **15**, 956-963.
24. Bairoch, A., and Apweiler, R. (2000) *Nucleic Acids Res.*, **28**, 45-48.
25. Melamud, E., and Moult, J. (2003) *Proteins: Structure, Function, and Bioinformatics*, **53** (Suppl. 6), 561-565.
26. Jin, Y., and Dunbrack, R. L., Jr. (2005) *Proteins: Structure, Function, and Bioinformatics*, **61 (**Suppl. **7)**, 167-175.
27. Bordoli, L., Kiefer, F., and Schwede, T. (2007) *Proteins*, **69** (Suppl. 8), 129-136.
28. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003) *Proteins*, **53**, 566-572.