Sequence Complexity of Disordered Protein

Pedro Romero,[1#] Zoran Obradovic,[1¥] Xiaohong Li,[1‡] Ethan C. Garner,[2†] Celeste J. Brown,[2] and A. Keith Dunker[2*]
[1]*School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington*
[2]*School of Molecular Biosciences, Washington State University, Pullman, Washington*

ABSTRACT ***Intrinsic disorder*** **refers to segments or to whole proteins that fail to self-fold into fixed 3D structure, with such disorder sometimes existing in the native state. Here we report data on the relationships among intrinsic disorder, sequence complexity as measured by Shannon's entropy, and amino acid composition. Intrinsic disorder identified in protein crystal structures, and by nuclear magnetic resonance, circular dichroism, and prediction from amino acid sequence, all exhibit similar complexity distributions that are shifted to lower values compared to, but significantly overlapping with, the distribution for ordered proteins. Compared to sequences from ordered proteins, these variously characterized intrinsically disordered segments and proteins, and also a collection of low-complexity sequences, typically have obviously higher levels of protein-specific subsets of the following amino acids: R, K, E, P, and S, and lower levels of subsets of the following: C, W, Y, I, and V. The Swiss Protein database of sequences exhibits significantly higher amounts of both low-complexity and predicted-to-be-disordered segments as compared to a non-redundant set of sequences from the Protein Data Bank, providing additional data that nature is richer in disordered and low-complexity segments compared to the commonness of these features in the set of structurally characterized proteins. Proteins 2001;42:38–48.** © 2000 Wiley-Liss, Inc.

Key words: protein disorder; sequence complexity; neural network predictors

INTRODUCTION

Amino acid sequence determines protein 3D structure[1] with the oft-stated corollary that structure is *prerequisite* to function[2–5] by mechanisms such as lock and key[6] or induced fit.[7] However, a number of proteins remain as flexible ensembles under physiological conditions and yet exhibit function when assayed.[8–11] Such proteins have been called "natively denatured,"[12] "natively unfolded,"[13] and "intrinsically unstructured."[14] Many other proteins are not intrinsically disordered throughout, but rather have functionally significant local regions of disorder.[15–19]

Intrinsic protein disorder has been identified by a variety of methods, including (1) protease digestion, with disorder indicated by sites of hypersensitivity[13,20–22]; (2) X-ray diffraction, with disorder indicated by residues missing from electron density maps[15,17,18,23]; (3) NMR spectroscopy, with disorder indicated by sharp peaks, by the absence of NOEs characteristic of secondary structure or by negative values for ^{1}H-^{15}N heteronuclear NOEs[8,10,20,23–29]; (4) circular dichroism, with disorder indicated by low intensity from ~ 210 to ~ 240 nm[9,13,30–32]; and (5) determination of hydrodynamic values, where an atypically large Stoke's radius for a given molecular weight indicates unfolded protein.[9,12,13,20,30]

We determined that intrinsically disordered regions could be predicted from their amino acid sequences[33–37] and identified long disordered regions (LDRs) having ≥ 40 residues characterized by especially strong predictions of disorder.[38] For our predictors with outputs, q, between 0 to 1.0 where $q > 0.5$ indicates disorder, about 1,000 putative LDRs were identified with $q > 0.85$; here these are called *extreme LDRs*. The Top 20 of these ranged in length from 120 to 576 residues with average predictor output values from 0.94 to 0.99.[38] The extreme LDRs and Top 20 were intended to provide a target list for experimental tests on the predictor.

Although no experimentalist has yet contacted us to confirm or refute any of the extreme LDR or Top 20 predictions, three bioinformaticists (Blackwell, States, and Frishman) told us that the Top 20 have low sequence complexity as defined by Wooten and Federhen.[39,40] This suggested that the predictor could be detecting non-globularity through low complexity rather than through sequence features specifically associated with disorder.

Abbreviations: CD, circular dichroism; LDR, long disordered region; NMR, nuclear magnetic resonance; NRL, Naval Research Laboratory; PDB, Protein Data Bank; PIR, protein identification resourse; PONDR, predictor of natural disordered regions; UV, ultraviolet; SW, Swiss Protein database.

Grant sponsor: NSF; Grant number: CSE-IIS-9711532; Grant sponsor: NIH; Grant number: 1R01 LM06916.

[#]Present Address: Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025. E-mail: promero@ai.sri.com

[¥]Present Address: Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 N. Broad St., Philadelphia, PA 19122. E-mail: zoran@joda.cis.temple.edu

[‡]Present Address: Advance Biometrics, Inc., 2722 East Main Ave., Puyallup, WA 98372. E-mail: xiaohong@livegrip.com

[†]Present Address: Department of Biochemsitry and Biophysics, Unversity of California San Francisco, San Francisco, CA 94143-0448. E-mail: egarner@itsa.ucsf.edu

*Correspondence to: A. Keith Dunker, Department of Biochemistry and Biophysics, Washington State University, Pullman, WA 99164-4660. E-mail: dunker@mail.wsu.edu

Received 3 December 1999; Accepted 24 August 2000

The uncertainties raised by the low complexity of the Top 20 LDRs motivated this study. The results show overlapping distributions for the complexity values of ordered and disordered sequences, with only the topmost being exclusively low complexity due to a dearth of some amino acids, herein called order-promoting, and an abundance of other amino acids, herein called disorder-promoting. Overall, concerted use of sequence complexity and disorder prediction appears to provide a useful tool for the analysis of protein sequences.

MATERIALS AND METHODS

Sequences and Databases

The following databases were used: (1) Protein Data Bank (PDB)[41]; (2) PDB_Select_25[42]; (3) the Naval Research Laboratory 3D (NRL-3D) sequence database, which is maintained and distributed by the Protein Identification Resource (PIR)[43]; and (4) Swiss Protein (SW.)[44]

The NRL-3D database contains amino acid sequences generated from the ATOMS list in PDB files; so, with exceptions in which coordinates are from models rather than from data, the sequences in NRL-3D comprise the ordered subset of PDB.[43] Modeled regions are identified in the PDB files as having zero occupancy. These are rare and usually short.

An all-globular version of NRL-3D (Globular-3D) was constructed by removing all fibrous sequences (coiled coils, collagen, and silk fibroins) and a few additional sequences that were either non-globular or that were classified as low complexity due to an abundance of ambiguous amino acids.

Since NRL-3D contains ordered residues, disorder prediction on this database gives a false-positive error rate.[35,38] However, NRL-3D is highly redundant. In order to remove biases in the evaluation of predictor error rates, we developed a non-redundant set of ordered protein sequences. Starting with the August 3, 1999, version of PDB_Select_25,[42] which contains just 1 representative from each group of related proteins in PDB, a non-redundant set of ordered protein sequences, called O_PDB _Select_25, was constructed by extracting the ordered regions.

Sets of helical coiled-coils, silks, and collagens were collected by key word searches on SW. Sequence regions associated with globular domains in these proteins were deleted.

Databases of disordered regions characterized by X-ray diffraction, NMR, or CD were constructed. The segments characterized as disordered by X-ray were identified as residues having backbone and side chain atoms that were absent from the ATOMS lists in the PDB_Select_25 files, yielding the disordered subset called D_PDB_Select_25. NMR- and CD-characterized segments of disorder were identified from their descriptions in publications found by key-word searches on PubMed.

Sequence Complexity Measure

Entropy as defined in Shannon's information theory[45] was previously applied to amino acid sequences by Wootton.[39] Shannon's entropy, herein called K_2, is given by the following equation:

$$K_2 = -\sum_{i=1}^{N} \frac{n_i}{L}\left(\log_2 \frac{n_i}{L}\right) = -\sum_{i=1}^{N} f_i \log_2 f_i \qquad (1)$$

where N represents the number of letters in the alphabet (20 amino acids in this case) and n_i is the number of times the letter i appears in the window of length L, so f_i corresponds to the fraction of amino acid i over the window. For a window of $L \geq 20$ and an alphabet of 20 letters (one for each amino acid), $0 \leq K_2 \leq \log_2(20) \approx 4.32$ bits.

Statistics

The mole fractions for the amino acids in a database were calculated as :

$$P_j = \sum (n_i P_{ji}) / \sum n_i, \qquad (2)$$

where P_{ji} is the frequency of amino acid j in sequence i of length n_i and the summation is over all sequences in a given database. The variances of the amino acids in the database were calculated as:

$$\mathrm{Var}(P_j) = \{\sum n_i^2 \mathrm{Var}(P_{ji})\} / (\sum n_i)^2, \qquad (3)$$

where $Var(P_{ji}) = P_{ji}(1 - P_{ji})/n_i$.

The fractional difference in composition between two sets a and b is $(P_j^a - P_j^b) / P_j^b$. The variances for these ratios are:

$$\mathrm{Var}(P_j^a - P_j^b)/P_j^b = (P_j^a/P_j^b)^2\{\mathrm{Var}(P_j^a)/(P_j^a)^2 + \mathrm{Var}(P_j^b)/(P_j^b)^2\}, \qquad (4)$$

where P_j^a is the mole fraction of amino acid j for database a, and $Var(P_j^a)$ is the variance of amino acid j for database a.[46]

Neural Network Predictors for Long Disordered Regions

We have developed several neural network predictors of natural protein disordered regions (PONDRs). In the initial studies, disorder was partitioned according to length, with the development of different predictors for short, medium, and long disordered regions.[35] The predictor for long disordered regions (LDRs) is herein renamed PONDR XL1. This predictor used 10 inputs. Later, disorder was partitioned according to position, with the development of different predictors for N-terminal, internal, and C-terminal regions.[34] These predictors used 8 inputs.

A new predictor for internal regions, called PONDR VL1, was developed as briefly described herein. A training set of 15 disordered regions having a total of 1,149 residues was compiled and balanced by an equal number of ordered residues taken randomly from NRL_3D. From an initial pool of 31 attributes, a branch and bound search[47] was used to select 10 attributes that gave the best collective discrimination between the order and disorder in the training set using a Mahalanobis distance criterion. The back-propagation learning algorithm[48] was used to train a

feedforward neural network having the ten selected attributes as inputs, a fully connected hidden layer of ten neurons and a single output. To estimate errors, the training was repeated on 5 disjoint subsets each having 80% of the data with 3 different initializations, so neural network training was repeated $5 \times 3 = 15$ times. Once the accuracy was established by this 5-cross validation procedure, a new neural network was trained to the same accuracy using all the data.

Of the 15 disordered regions in the training set, 8 were characterized by X-ray diffraction (PDB IDs: 2tbv, 2ts1, 1aui, 1bgw, 1elo, 1af3, 1ati, and 1lbh) and 7 by NMR (SW IDs: prio_mouse, h5_chick, flgm_salty, regn_lambd, hsf_klula, and hmgi_human, and PIR accession: S50866). The 31 attributes in the initial pool included the 20 amino acid compositions, two different hydropathy scales,[49,50] flexibility index,[51] α-moment,[52] β-moment,[53] net charge (K + R − D − E),[54] aromatic composition (W + F + Y),[54] coordination number,[55] codon number,[56] alphabet size,[57] and side chain volumes.[58]

To enable prediction from the first to the last residue in a protein, the PONDR VL1 and the predictors for the N- and C-terminal regions were integrated. This integration was carried out in 3 steps. First, predictions were made by the three predictors over their respective domains, with overlapping predictions for positions 11–14 by the N-terminal and VL1 predictors, and, for a protein of length M, with overlapping predictions from M - 14 to M - 11 by the C-terminal and VL1 predictors. Second, the values for each of the 4 pairs of overlapping prediction were averaged. Third, the now integrated prediction outputs were smoothed by averaging over sliding windows of 9 amino acids, with the first and last 4 sequence positions being assigned the unsmoothed prediction output values from the N- and C-terminal predictors, respectively. This integrated predictor is herein called PONDR VL-XT.

RESULTS

Databases of Characterized Order and Disorder

The first step in this study was to collect ordered and disordered sequences and organize them into databases as outlined in Materials and Methods (Table I). The protein identities for Table I are given on our website: http://disorder.chem.wsu.edu

TABLE I. Data Summary

Group	Number of segments	Number of residues
Fibrous sequences		
Coiled coils	28	10,391
Collagen	27	20,109
Silk repeats	14	10,329
Order databases		
Globular-3D	14,540	2,610,197
O_PDB_Select_25	1,111	220,668
XL1 training	7	1,561
Disorder databases		
XL1 training	7	508
VL1 training	15	1,376
XT training	199	2,894
X-ray	56	2,844
NMR	41	4,019
CD	53	10,554
ALL	150	17,417

False-Positive Prediction of Disorder

An estimate of the false-positive error rate is needed in order to determine the extent to which predicted LDRs are contaminated with ordered protein. False-negative predictions of order on actual disordered regions are less relevant here because such miss-classification of LDRs as ordered segments would not significantly affect the subsequent analysis.

Table II shows the false-positive error rates for PONDRs XL1 and VL-XT, using two thresholds for disorder, namely $q > 0.5$ and $q > 0.85$ for predicted and extreme LDRs, respectively. For obvious reasons,[38] the false-positive error rate drops with increasing length or threshold. VL-XT has two advantages compared to VL1: a lower error rate and predictions to the termini. The latter is especially important because the ends of proteins are often disordered.

The contamination of the predicted LDRs with ordered segments was estimated as follows. Using O_PDB_Select_25, the false-positive prediction rates as a function of length were determined. These rates were then used as the expected frequency of false-positive LDRs in SW as a function of length, assuming that all of SW is ordered and that the ordered sequences in O_PDB_Select_25 are representative of those in SW. This estimated (false-positive) frequency was compared with the actual prediction frequency as a function of length (Fig. 1). A lower bound for the contamination was then estimated as the relative areas under the two curves, yielding an estimate of 1.7%. Repeating the simulation for the extreme LDRs gives an estimate of less than 0.08% contamination. Carrying out these simulations with XL1 rather than VL-XT resulted in about 10-fold higher estimates of contamination, providing a strong reason for using the VL-XT predictor for further studies.

Databases of Predicted Order and Disorder

An efficient way to increase the size of the disordered database would be to use prediction.[35,38] Previously we used PONDR XL1 to generate the LDRs, the extreme LDRs, and the Top 20 (reference 38); here we used VL-XT for the LDRs and extreme LDRs, but kept the original Top 20 to test the personal communications indicating that the previous Top 20 segments are low complexity. Protein identities for this database are provided at our website: http://disorder.chem.wsu.edu/. The predictions of disorder were compiled into a database (Table III). In addition, low-complexity segments, defined as $K_2 < 2.9$, were also included.

The attributes used to train XL1 and the three parts of VL-XT are given in Table IV. Even though alphabet size, which is a measure of sequence complexity,[57] was in the attribute pool for the VL-XT, it was not selected. Thus, neither XL1 nor VL-XT use sequence complexity.

TABLE II. False-Positive Disorder Prediction Rates

PONDR	Training error (%)[a]	Threshold	O_PDB_Select_25 (%) Per residue	10 or longer	20 or longer	40 or longer
XL1	26 ± 4	0.5	34	17	7	1.30
VL-XT	17 ± 3	0.5	22	9	3	0.40
XL1	—	0.85	4.4	0.7	0.1	0.00
VL-XT	—	0.85	3.8	0.5	0.1	0.01

[a]5-cross validation was used for predictor training at the 0.5 threshold. The training error was from 15 experiments: 3 neural networks were trained using different initializations for each of 5 disjoint subsets containing 4/5 of the data. Error rates were then estimated by application of the resulting 5 sets of 3 neural nets to the data not used for training for each of the 5 sets.

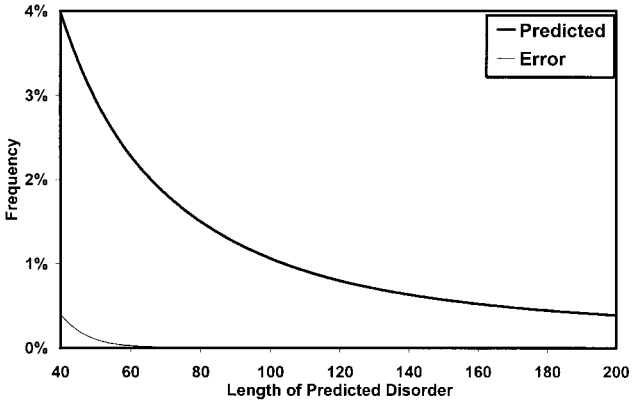

Fig. 1. Estimated contamination of predicted LDRs with ordered segments. PONDR VL-XT was applied to SW and to O_PDB_Select_25. Segments of length ≥ 40 with $q > 0.5$ for every position were collected for both databases. The relative frequencies of disorder prediction on these two databases were then determined and compared: the upper curve is from SW, the lower is from O_PDB_Select_25 and provides an estimate of the error rate vs. length.

TABLE III. Predicted Disorder and Low Complexity

	PONDR used	Number of segments[a]	Number of residues
Swiss-Protein		81,005	30,377,255
LDR	VL-XT	40,764	2,571,711
Extreme LDR	VL-XT	9,393	610,119
Top 20 LDR	XL1	20	4,342
Low complexity	$K_2 < 2.9$	8,300	654,307

[a]Proteins of length 45 and shorter were excluded from this analysis.

Order, Disorder, and Sequence Complexity

The sequence complexities of ordered and disordered proteins were compared by K_2 distributions calculated over sliding windows of 45 residues (Fig. 2). The distribution for Globular-3D is repeated in each panel to provide a common reference for each set of data.

The distributions in Figure 2A show that the fibrous proteins have mostly lower complexity sequences compared to those in Globular-3D. These data are consistent with previous work.[40]

Figure 2B compares the K_2 histograms of both the ordered and the disordered parts of the training set used to develop PONDR XL1 with that of Globular-3D. These data show that this predictor utilized data for which the disordered and ordered fragments had similar complexities.

Figure 2C compares the K_2 distributions of ordered proteins with those from 3 different sets of disordered regions as characterized by X-ray, NMR, and CD. The three differently characterized sets of disordered proteins yield remarkably similar complexity distributions and so were combined into a single database called ALL-Disorder.

In Figure 2D the K_2 distributions of predicted, extreme, and Top 20 LDRs[38] are compared with those of ALL-Disorder and Globular-3D. A progressive shift from high to low complexity is observed with Globular-3D > ALL-Disorder > predicted LDRs > extreme LDRs > Top 20.

Amino Acid Compositions

To gain insight into the relationships between sequence and disorder, we compared the amino acid compositions of the ordered, disordered, fibrous, and low-complexity sets in this study (see http://disorder.chem.wsu.edu for the raw composition data). To visualize differences, the amino acid mole fractions, P, of each amino acid, j, for pairs of protein sets, a and b, are displayed as $(P_j^a - P_j^b) / P_{j^b}$, where set a varies and set b is Globular-3D (Fig. 3). For these comparisons, the lower bound is -1 for cases for which $P_j^a = 0$ and the upper bound for amino acid j is equal to $(100 - P_j^b) / P_j^b$. The amino acids in Figure 3 are arranged from the most rigid to the most flexible according to the scale of Vihinen et al.[51] This scale is based on the averaged B-factor values for the backbone atoms of each residue type as estimated from 92 proteins.

The X-ray-, NMR-, and CD-characterized segments of disorder have amino acid compositions that are similar to each other and different from those of ordered segments (Fig. 3A). Specifically, the disordered segments are ~ 20 to ~ 50% depleted in the amino acids to the left (e.g., the negative peaks for W to L in Fig. 3A) and ~ 20 to ~ 50% enriched in the amino acids to the right (e.g., the positive peaks for M, A, R, Q, S, P, E, and K) with only a few exceptions (especially G, N, and D). Given these results, the amino acids to the left are herein called *order-promoting* and those to the right *disorder-promoting*.

The fibrous proteins are also very significantly depleted in order-promoting amino acids and enriched in disorder-promoting amino acids (Fig. 3B). However, the enrichment patterns of the fibrous proteins are very distinct from the

TABLE IV. PONDR Inputs

PONDR	Attributes[a]									
XL1	Flexibility	Hydropathy	C	W	Y	H	D	E	K	S
VL1	Coordination number	Net charge	WFY	W	Y	F	D	E	K	R
XN	Coordination number	V	VIYFW	M	N	H	D	PEVK		
XC	Coordination number	Hydropathy	VIYFW	M	T	H		PEVK		R

[a]These attributes were calculated as normalized values of the indicated feature over sliding windows. For example, the value for C is simply the number of times C appears in a given window divided by the window length, and the value for WFY is the sum of W + F + Y divided by the window length. The normalization for a given value, Observed, is simply (Observed − Min)/(Max − Min), where Max and Min were determined from the entire dataset. VL1 used windows of 21 residues. XL1 used windows that varied in length according to attribute type (as described previously[35]). XN and XC used windows that varied in length and the position of prediction assignment according to location relative to the end of the chain as described in detail elsewhere.[34]

patterns of the disordered segments (compare Fig. 3A and B).

The segments predicted-to-be-ordered show compositions quite similar to those of segments structurally characterized as ordered, with differences much smaller than those for the other comparisons studied to date (Fig. 3C). For 15 amino acids (W, Y, V, H, M, A. T, R, Q, S, N, P, E, D, and K) the compositional differences are small (12% or less), with significant differences (> 20%) for just 3 amino acids (C, F, and L).

Likewise, the depletions and enrichments of the predicted LDRs and structurally characterized intrinsically disordered segments are very similar to each other. For 11 of the 20 amino acids (W, I, V, L, M, A, T, G, Q, N, and E) both predicted and characterized disorder are very similarly depleted or enriched; for another 6 (C, F, Y, R, S, and P) the changes from structured protein are similar. The patterns of enrichment or depletion are dissimilar for only 3 amino acids (H, D, and K).

Figure 3D compares the amino acid compositions of low-complexity sequences from SW, the extreme LDRs, and the Top 20 predictions of disorder. Like the fibrous proteins, the low-complexity segments are depleted in the order-promoting amino acids and enriched in the disorder-promoting ones, but the patterns of depletion and enrichment are very different (compare Fig. 3D and B). On the other hand, low-complexity and extreme LDRs show the same trends for 17 of the 20 amino acids, with the low-complexity segments typically exhibiting depletions of the order-promoting amino acids and enrichments of the disorder-promoting ones.

The Top 20 LDRs (Fig. 3D) are depleted in the 12 left-most amino acids (W to R) and also in Q. Enrichments are observed for just 7 of the more disorder-promoting amino acids, with very substantial enrichments in S and E.

Low Complexity Segments and Predicted LDRs in SW and PDB_Select_25

To understand relationships between low-complexity and predicted LDRs, the protein chains in SW and PDB_Select_25 having at least one segment with one characteristic and not the other and segments with at least one of each characteristic were determined (Table V). The sequences in SW having at least one low-complexity segment are 10 times richer than the corresponding sequences in PDB_Select_25, that is 7.1% for SW compared to 0.7% for PDB_Select_25, ~ 3 times richer in predicted LDRs, that is 29.1% for SW compared to 11.0% for PDB_Select_25, ~ 15 times richer in extreme LDRs, that is 7.6% for SW compared to 0.5% for PDB_Select_25. Finally, most of the low complexity segments are also predicted to be LDRs. Just 881 out of 5,748, or 15% of the chains in SW with low-complexity segments do not correspond to predicted LDRs and just 2 of 6 chains in PDB_Select_25 with low-complexity segments fall into this category.

DISCUSSION

Identification and Miss-Identification of Intrinsic Order and Disorder

Disordered segments characterized by X-ray diffraction, NMR, and CD have been compiled (Table I). An X-ray-characterized LDR could be identified as disordered due to a wobbly, ordered domain and so could be miss-classified. NMR reveals regions of high local motion and so provides unambiguous indication of disorder; however, NMR-characterized disorder is biased towards random-coils compared to molten globules due to exchange line broadening for the latter. Furthermore, NMR analysis is typically restricted to smaller proteins. Near and far UV CD in combination can distinguish among ordered structure, molten globules, and random coils, but CD is semi-quantitative and lacks position-specific information. Thus, the LDRs in Table I surely contain significant amounts of ordered structure that are miss-classified as disordered.

The ordered protein data likely contains far fewer miss-classified segments than the disordered data, yet Globular-3D should not be considered to be devoid of miss-classification. For example, from the more than 12,000 segments in Globular-3D, we sampled 50 that were predicted to be mostly disordered. Of this sample, 49 were involved in complexes with DNA, proteins, or co-factors, each with no record of crystallization when uncomplexed. These proteins likely don't self-fold but probably undergo disorder-to-order transitions upon complex formation.[18,19,59] Thus, the estimated false-positive error rates

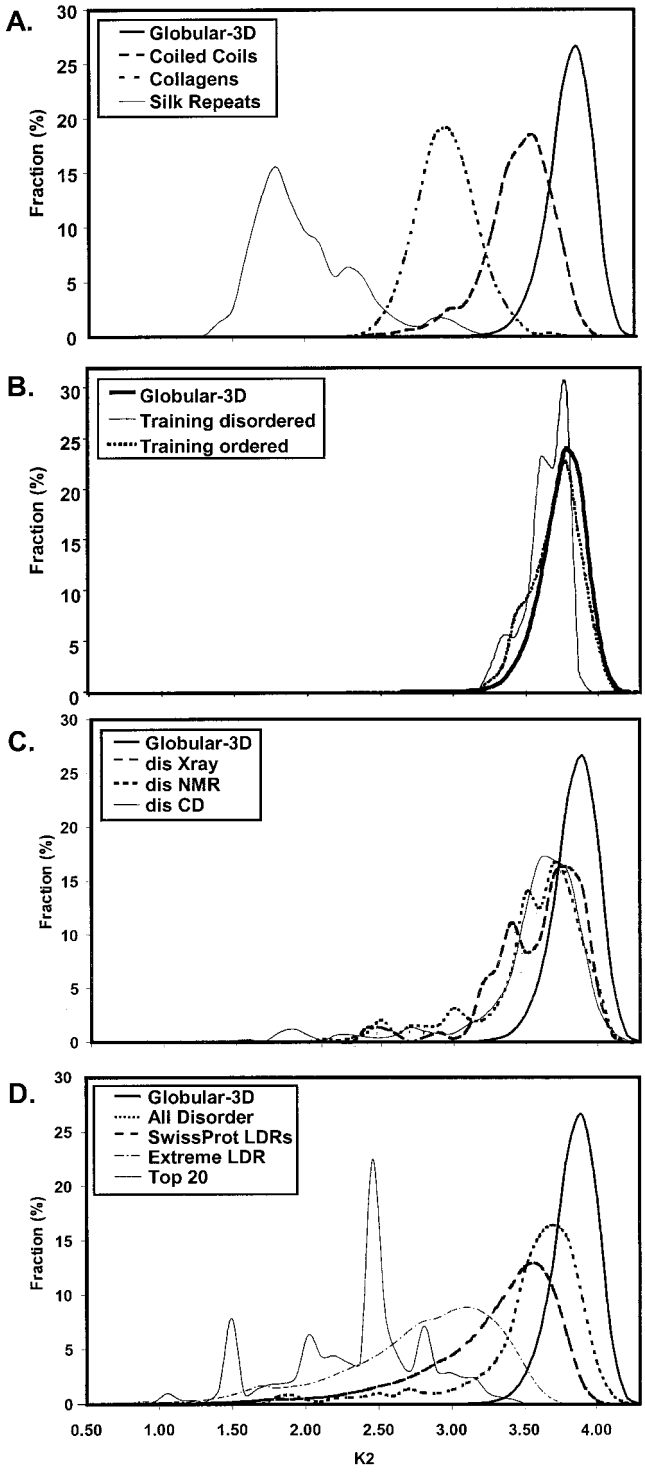

Fig. 2. Complexity distributions. The complexities K_2 were calculated by equation (1) for sliding windows of 45 residues and converted to histograms for the Globular-3D database and for other selected databases from Tables I and III. **A:** Fibrous proteins (silk, collagen, and coiled-coils) are compared with ordered protein (Globular-3D). **B:** Ordered and disordered segments used to train PONDR XL1 are compared with Globular-3D. **C:** The three sets of disordered segments as characterized by X-ray, NMR, or CD are compared with Globular-3D. **D:** ALL-Disorder, predicted LDRs, extreme LDRs, and the Top 20 disorder predictions are compared with Globular-3D.

are probably too high, but the excess error would be difficult to determine at this time.

Predicted Order and Disorder

Because of attempts to use only correctly classified disorder, training sets have been very small: 508 disordered residues for XL1 and 1,376 for the internal regions part of VL-XT (Table I). Nevertheless, both XL1 and VL-XT generalize well for the prediction of order on O_PDB_Select_25. This database contains over 220,000 residues and samples the structured parts of nearly all crystallized protein families. The apparent error rates determined from O_PBD_Select_25 are within two standard deviations of the training values for both XL1 and VL-XT (Table II). Also, the amino acid compositions of the predicted and observed order match quite well (Fig. 3C).

The predicted LDRs and extreme LDRs were estimated to be contaminated with 1.7% (Fig. 1) and 0.08% ordered residues, respectively. However, a 40 or longer residue segment with both order and disorder would be more likely to be predicted as completely disordered as compared to a fully ordered segment of the same length. Since segments with both order and disorder were not considered in the estimation of contamination, the actual contamination with order must be higher than the estimates. On the other hand, the contamination of the structurally characterized disorder with order, especially for the X-ray and CD disorder data, is likely to be quite high, so the predicted LDRs, and almost certainly the extreme LDRs, could be less contaminated with order than are the structurally characterized LDRs.

Despite the uncertain contamination of the predicted LDRs with order, these predictions are still useful. Prediction increased the disorder data from not quite 20,000 residues (Table I) to over 600,000 for the extreme LDRs and to over 2.5 million for the LDRs (Table III). A further complication is that predicted LDRs are biased towards disorder that resembles the training set. Disorder apparently comes in different flavors, with those distinct from the training set examples being miss-predicted as ordered.[37] Thus, using both predicted and observed LDRs should give a better overall understanding of the sequence characteristics of disordered regions in proteins.

Sequence Complexity and Ordered Protein Structure

Figure 2A shows that fibrous sequences have lower complexities and globular proteins have higher complexities, with nearly all of the overlap arising from coiled-coils. These data confirm Wootton's original work[39,40] with the additional insight that complexity shifts to lower values in the order globular protein > coiled coil > collagen > silk.

The Globular-3D K_2 distribution appears to approach the X-axis smoothly (Fig. 2). However, scale expansion reveals a rather abrupt increase, changing from one to thousands of examples over .08 units of K_2. Not one of these 45-residue segments has a sequence complexity below $K_2 \approx 2.9$ (nor fewer than 10 amino acids), suggesting a possible lower bound.[57]

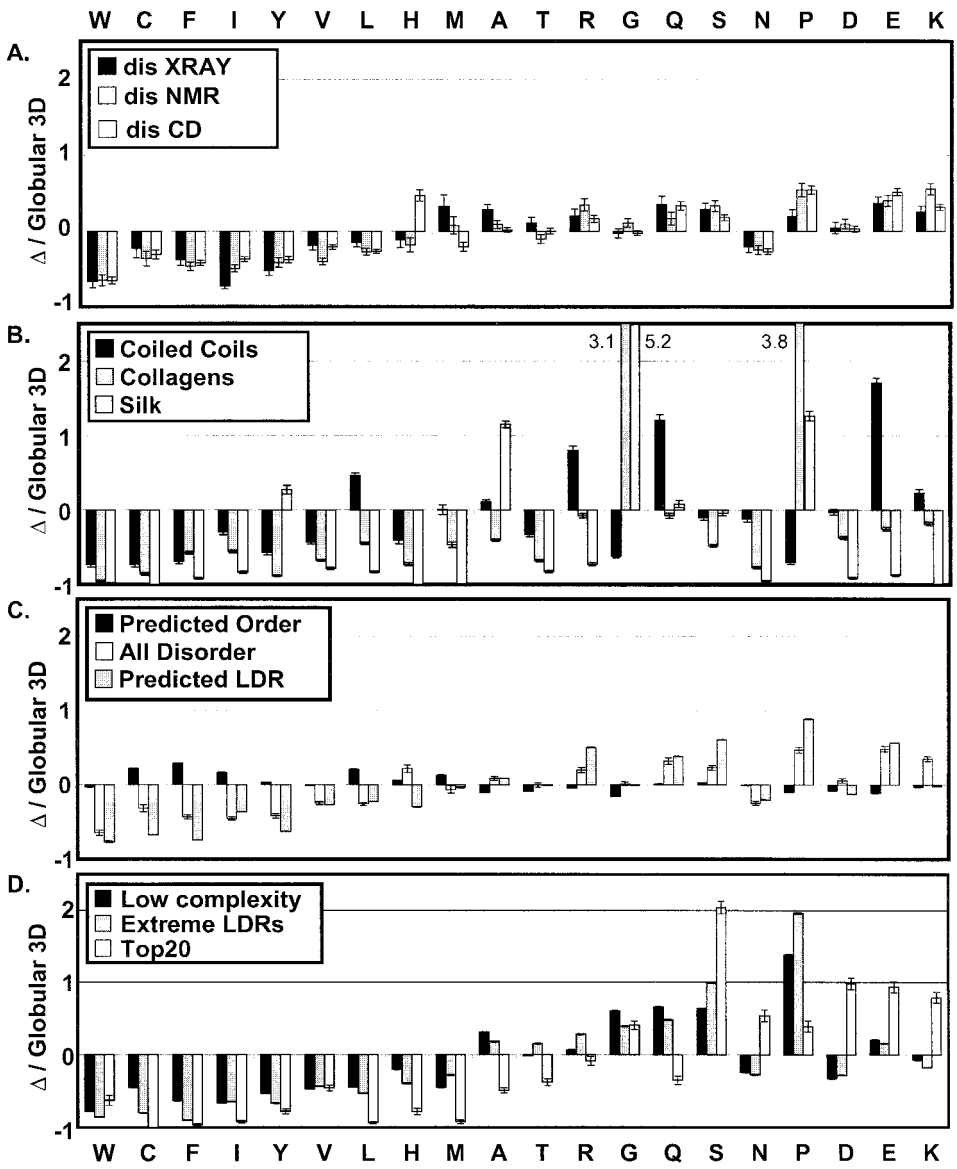

Fig. 3. Comparisons of amino acid compositions for different sets of proteins. For two sets of proteins, *a* and *b*, the ordinates are given by $(P_i^a - P_i^b) / P_i^b = D$ / *Globular-3D*, with the error bars representing one standard deviation, and with the terms and calculations as described in Statistics. The three differently characterized sets of disorder are compared with Globular-3D in **A**, the three types of fibrous proteins in **B**, predicted order, ALL-disorder, and predicted disorder in **C**, and low-complexity sequences, extreme LDRs, and the Top 20 in **D**.

Globular, ordered proteins need to have polar and non-polar amino acids to define the outside surfaces and the core regions. On the outside surface and on the surfaces of crevices and pockets, various polar amino acids are needed to meet solubility and functional requirements. Within the core regions, variously sized and shaped non-polar amino acids are needed to fill the nooks and crannies to yield the tight packing observed in these proteins. In addition, some polar amino acids and water molecules are commonly found within protein cores, perhaps to facilitate conformational changes associated with function. Thus, a complexity lower bound corresponding to about 10 amino acids seems to be reasonable for ordered, globular structure.

We used the observed lower bound of ordered protein structure to define low-complexity sequences. Wootton and co-workers defined low complexity differently, with a "trigger" value and an "extension" value, both of which have been incorporated into a program called SEG.[60] Limited experimentation shows our definition of low complexity to be more stringent than is Wootton's SEG program when used with its default parameters.

Sequence Complexity And Disordered Structure

In contrast to structured proteins, intrinsic disorder exhibits a significant fraction of low complexity sequences (Fig. 2C and D). Of the almost 20,000 disordered amino

TABLE V. Complexity and Disorder in a Structure Database and in Swiss Protein

Category[a]	PDB_Select_25 no. of chains	PDB_Select_25 % of chains	SwissProt no. of chains	SwissProt % of chains
Proteins >45 aa	920		81,005	
Low K2	6	0.7	5,748	7.1
Predicted LDRs	101	11.0	23,570	29.1
Extreme LDRs	5	0.5	6,291	7.8
Low K2, no LDR	2	0.2	881	1.1
LDR, no Low K2	97	10.5	18,703	23.1
Low K2 or LDR	103	11.2	24,451	30.2

[a]Chains having at least 1 of the indicated category.

acids in the ALL-Disorder database, about 1,587 residues or 8%, fall in the region with $K_2 < 2.9$ as compared to 0% out of more than 2.5 million in Globular-3D. Without structural constraints, disordered regions are free to contain just a few amino acids, or even just one.

Low-complexity sequences can also form ordered structure under some circumstances. For example, the fibrous proteins have low complexity and are predicted to be disordered by both XL1 and VL-XT. Yet these sequences form ordered (fibrous) structures upon self-association. However, this folding requires formation of a complex with a binding partner. We have yet to find a protein or region with $K_2 < 2.9$ that self-folds into ordered protein without a partner.

On the other hand, even very high complexity sequences can fail to form ordered structure. For example, two large fragments from thioredoxin are disordered[61] even though their K_2 distributions lie within the domain of 3.6 to 4.1. Such failure to form order despite high complexity may be due to lack of suitable long-range interactions or some other feature and suggests that there might be no upper limit on the sequence complexity of intrinsic disorder. In agreement with this suggestion, ALL-Disorder and Globular-3D are observed to extrapolate to the same upper limit (Fig. 2D); the extrapolation to apparently the same value is conserved when a 100-fold scale expansion is used (not shown).

High-complexity disordered sequences that fail to fold due to the absence of suitable long-range interactions or other features might be identified by PONDR as ordered. Like the fragments in thioredoxin, such sequences could have evolved to associate with partners. Indeed, several of the sequences in the ALL-Disorder database were characterized as disordered in the absence of known partners such as RNA or DNA. Thus, the terms *intrinsically unstructured*[14] or *intrinsically disordered* do not necessarily mean that the proteins are incompletely folded in the cell, but only that the proteins don't self-fold and may be involved in complexes. A more interesting possibility is that some proteins undergo order/disorder transitions upon association/dissociation with their partners as important steps in their biological functions.[18,19,28]

Sequence Complexity and Predicted LDRs

The Top 20 LDRs determined by PONDR XL1 are low-complexity sequences (Fig. 3C), which raised the possibility that disorder was being detected by low complexity rather than by sequence attributes that correlate with disorder. However, the ordered and disordered segments used for training XL1 show a minor difference in complexity (Fig. 2B), thus probably ruling out the possibility that low complexity was an unknown characteristic of the training set used to develop the predictor that identified the Top 20.

Even though sequence complexity has not been an explicit attribute used for the predictors (Table II), complexity is reduced for predicted as compared to actual disorder and complexity decreases still further for segments with higher prediction scores. That is, the modes of the distributions for Globular-3D, ALL-Disorder, predicted LDRs, extreme LDRs, and the Top 20 are $K_2 = 3.92$, 3.75, 3.60, 3.10, and 2.49, respectively, and the percentages of these same distributions with $K_2 \leq 2.9$ are 0, 8, 15, 47, and 65, respectively.

The disorder prediction score depends on attributes (Table IV) associated with depletion of most of the order-promoting amino acids and enrichment of some of the disorder-promoting ones (Fig. 3). The disorder prediction scores increase as the depletions and enrichments increase. Increased depletions and enrichments lead to decreased complexity. Thus, as prediction scores go up, sequence complexity goes down.

Sequence Complexity and Amino Acid Composition

One suggested advantage of the sequence complexity measure is its independence of amino acid type.[40] A separate issue is whether low-complexity sequences contain a random or nonrandom sampling of the amino acids. If low-complexity sequences were a random sampling of the amino acids of ordered protein structure, then their amino acid compositional differences as given in Figure 3D would all be near 0; in contrast, almost none of the amino acids exhibits differences from order near 0. Thus, the low-complexity sequences in SW exhibit compositional biases, not random sampling.

The fibrous and general low-complexity sets both show large depletions of the order-promoting amino acids. The single exception is the atypical enrichment in L for the coiled-coils. This enrichment is perhaps due to an abundance of leucine zippers in SW.

The fibrous and general low-complexity sets both also show substantial enrichments of some disorder-promoting

residues, but with differences in the patterns of enrichment. The fibrous proteins are enriched in the specific amino acids that correspond to their repeating motifs, namely A, G, and P (silk), G and P (collagen), and R, Q, and E (coiled-coils). The non-fibrous, low-complexity segments are also enriched in A, G, Q, P, and E as are one or more of the fibrous proteins, but the amount of enrichment is different. Also, unlike any of the fibrous protein motifs, the general low-complexity sequences are also substantially enriched in S.

The low-complexity sequences from SW and the extreme LDRs from this same database show similar trends of depletion and enrichment for 17 of the 20 amino acids (Fig. 3D). Given the apparent inability of low-complexity sequences to form ordered, globular structure, one possible explanation is that the non-fibrous, low-complexity segments were selected over evolutionary time for the specific purpose of being intrinsically disordered.

Commonness of Predicted LDRs and Low-Complexity Segments

The SW sequence database is far richer in low complexity segments than is PDB_Select_25 (Table V), but the fraction estimated to be low complexity here, about ~ 7%, is far below the ~ 25% estimated previously by the SEG program using default parameters.[60] This comparison points out the greater stringency of our definition of low complexity.

The percentage of protein chains with either low-complexity segments or predicted LDRs is ~ 3 times higher in SW as compared to PDB_Select_25, e.g., about 30.2% as compared to 11.2%. The probable explanation of these results is that the requirement for crystallization biases the latter database against proteins with significant amounts of disordered or fibrous structure.[35,60]

Implications for Deducing Function from Sequence

The amount of amino acid sequence information from the various genome projects is growing rapidly. A companion structural genomics project[62,63] has started recently. The goal of structural genomics is to characterize at least one representative of every protein fold. These representative structures will provide the basis for constructing useful 3D models by means of sequence similarity information, e.g., by homology modeling. Of course the goal of this approach is to identify sequence/function correlations using structural models as intermediates.

Determining the 3D structures of a representative set of folds is and ought to be a top priority. However, many proteins have low-complexity segments or intrinsically disordered regions that are involved in function.[9,10,13,15,17–21,23,26–28, 59, 64–81] As pointed out above, the percentage of protein chains having low complexity segments or putative disordered regions is not small. Also, from the content of the current databases, it appears to be very likely that many of the proteins with disorder or low complexity won't crystallize. Thus, unless intrinsic disorder and low complexity are taken into account, the structural genomics project and other efforts to deduce function from sequence will fall short.

ACKNOWLEDGMENTS

The authors thank Tom Blackwell, David States, and Dmitrij Frishman for pointing out to us that our top 20 list of the most strongly predicted disordered regions are mostly low complexity. Support from NSF-CSE-IIS-9711532 to Z. O. and A.K.D. and from N.I.H. 1R01 LM06916 to A.K.D and Z.O. is gratefully acknowledged.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Mirsky AE, Pauling L. On the structure of native, denatured and coagulated proteins. Proc Natl Acad Sci USA 1936;22:439–447.
3. May AC, Johnson MS, Rufino SD, Wako H, Zhu ZY, Sowdhamini R, Srinivasan N, Rodionov MA, Blundell TL. The recognition of protein structure and function from sequence: adding value to genome data. Phil Trans R Soc Lond B Biol Sci 1994;344:373–381.
4. Koonin EV, Tatusov RL, Galperin MY. Beyond complete genomes: from sequence to structure and function. Curr Opin Struct Biol 1998;8:355–363.
5. Orengo CA, Todd AE. From protein structure to function. Curr Opin Struct Biol 1999;3:374–382.
6. Fischer E. Einfluss der configuration auf die wirkung derenzyme. Ber Dt Chem Ges 1894;27:2985–2993.
7. Koshland Jr. DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–104.
8. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. Nat Struct Biol 1997;4:285–291.
9. Kriwacki RW, Wu J, Tennant L, Wright PE, Siuzdak G. Probing protein structure using biochemical and biophysical methods. Proteolysis, matrix-assisted laser desorption/ionization mass spectrometry, high-performance liquid chromatography and size-exclusion chromatography of p21Waf1/Cip1/Sdi1. J Chromatogr A 1997;777:23–30.
10. Fletcher CM, McGuire AM, Gingras AC, Li H, Matsuo H, Sonenberg N, Wagner G. 4E binding proteins inhibit the translation factor eIF4E without folded structure. Biochemistry 1998;37:9–15.
11. Plaxco KW, Gross M. The importance of being unfolded. Nature 1997;386:657, 659.
12. Schweers O, Schonbrunn-Hanebeck E, Marx A, Mandelkow E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. J Biol Chem 1994;269:24290–24297.
13. Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT, Jr. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry 1996;35:13709–13715.
14. Wright PE, Dyson HJ. Intrinsically unstructured proteins: Reassessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
15. Alber T, Gilbert WA, Ponzi DR, Petsko GA. The role of mobility in the substrate binding and catalytic machinery of enzymes. Ciba Found Symp 1982;93:4–24.
16. Champness JN, Bloomer AC, Bricogne G, Butler PG, Klug A. The structure of the protein disk of tobacco mosaic virus to 5 Å resolution. Nature 1976;259:20–24.
17. Huber R. Conformational flexibility in protein molecules. Nature 1979;280:538–539.
18. Schulz GE. Nucleotide binding proteins. In: Balaban M, editor. Molecular mechanism of biological recognition. New York: Elsevier/North-Holland Biomedical Press; 1979. p 79–94.
19. Spolar RS, Record II MT. Coupling of local folding to site-specific binding of proteins to DNA. Science 1994;263:777–784.
20. Aviles FJ, Chapman GE, Kneale GG, Crane-Robinson C, Bradbury EM. The conformation of histone H5. Isolation and characterisation of the globular segment. Eur J Biochem 1978;88:363–371.
21. Manalan AS, Klee CB. Activation of calcineurin by limited proteolysis. Proc Natl Acad Sci USA 1983;80:4291–4295.
22. Shaiu WL, Hu T, Hsieh TS. The hydrophobic, protease-sensitive terminal domains of eukaryotic DNA topoisomerases have essential function. Pacific Symp Biocomput 1999;4:578–589.

23. Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, Nettesheim D, Chang BS, Thompson CB, Wong SL, Ng SL, Fesik SW. X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. Nature 1996;381:335–341.
24. Aaron BM, Oikawa K, Reithmeier RA, Sykes BD. Characterization of skeletal muscle calsequestrin by 1H NMR spectroscopy. J Biol Chem 1984;259:11876–11881.
25. Billeter M, Riek R, Wider G, Hornemann S, Glockshuber R, Wuthrich K. Prion protein NMR structure and species barrier for prion diseases. Proc Natl Acad Sci USA 1997;94:7281–7285.
26. Cho HS, Liu CW, Damberger FF, Pelton JG, Nelson HC, Wemmer DE. Yeast heat shock transcription factor N-terminal activation domains are unstructured as probed by heteronuclear NMR spectroscopy. Protein Sci 1996;5:262–269.
27. Huth JR, Bewley CA, Nissen MS, Evans JN, Reeves R, Gronenborn AM, Clore GM. The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat Struct Biol 1997;4:657–665.
28. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci USA 1996;93:11504–11509.
29. Newman M, Strzelecka T, Dorner LF, Schildkraut I, Aggarwal AK. Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. Science 1995;269:656–663.
30. Hernandez MA, Avila J, Andreu JM. Physicochemical characterization of the heat-stable microtubule- associated protein MAP2. Eur J Biochem 1986;154:41–48.
31. Loomis RE, Bergey EJ, Levine MJ, Tabak LA. Circular dichroism and fluorescence spectroscopic analyses of a proline-rich glycoprotein from human parotid saliva. Int J Pept Protein Res 1985;26:621–629.
32. Warrant RW, Kim SH. alpha-Helix-double helix interaction shown in the structure of a protamine-transfer RNA complex and a nucleoprotamine model. Nature 1978;271:130–135.
33. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. Genome Informat 1998;9:201–213.
34. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. Genome Informat 1999;10:30–40.
35. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. Proc. I.E.E.E. Int Conf Neural Networks 1997;1:90–95.
36. Romero P, Obradovic Z, Dunker AK. Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Informat 1997;8:110–124.
37. Romero PZ, Obradovic C, Dunker AK. Intelligent data analysis for protein disorder prediction. Artific Intell Rev. In press.
38. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Guilliot S, Garner E, Dunker AK. Thousands of proteins likely to have long disordered regions. Pacific Symp Biocomput 1998;3:437–448.
39. Wootton JC. Statistic of local complexity in amino acid sequences and sequence databases. Comput Chem 1993;17:149–163.
40. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
42. Hobohm U, Sander C. Enlarged representative set of protein structures**.** Protein Sci 1994;3:522–524.
43. Pattabiraman N, Namboodiri K, Lowrey A, Gaber BP. NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. Protein Seq Data Anal 1990;3:387–405.
44. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res 1996;24:21–25.
45. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948:379–423, 623–656.
46. Kendall M, Stuart A. The advanced theory of statistics. C. Griffin & Company Limited; 1977. 455 p.
47. Ripley BD. Pattern recognition and neural networks. Cambridge, UK: Cambridge University Press; 1996. 403 p.
48. Werbos P. Beyond regression: New tools for predicting and analysis in the behavioral sciences. PhD Thesis, 1974, Harvard University, Cambridge, MA. 453 p.
49. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
50. Rose GD. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 1978;272:586–590.
51. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
52. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature 1982;299:371–374.
53. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci USA 1984;81:140–144.
54. Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK. The sequence attribute method for determining relationships between sequence and protein disorder. Genome Informat 1998;9:193–200.
55. Galaktionov SG, Marshall GR. Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D. St. Louis, MO: Washington University Institute for Biomedical Computing; 1996. 42 p.
56. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F, O'Neal C. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc Natl Acad Sci USA 1965;53:1161–1168.
57. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 1999;462:363–367.
58. Janin J. Surface and inside volumes in globular proteins. Nature 1979;277:491–492.
59. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 1996;271:1247–1254.
60. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
61. Yang X-M, Georgescu RE, Li J-H, Yu W-F, Haierhan, Tasayco ML. Recognition between disordered polypeptide chains from cleavage of an a/b domain: self- versus non-self-association. Pacific Symp Biocomput 1999;4:590–600.
62. Gaasterland T. Structural genomics: bioinformatics in the driver's seat. Nat Biotechnol 1998;16:625–627.
63. Shapiro L, Lima CD. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. Structure 1998;6:265–267.
64. Ayala YM, Vindigni A, Nayal M, Spolar RS, Record MT, Jr., Di Cera E. Thermodynamic investigation of hirudin binding to the slow and fast forms of thrombin: evidence for folding transitions in the inhibitor and protease coupled to binding. J Mol Biol 1995;253:787–798.
65. Bloomer AC, Champness JN, Bricogne G, Staden R, Klug A. Protein disk of tobacco mosiac virus at 2.8 Å resolution showing the interactions within and between subunits. Nature 1978;276:362–368.
66. Bode W, Schwager P, Huber R. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution. J Mol Biol 1978;118:99–112.
67. Chambers JL, Stroud RM. Difference Fourier refinment of the structure of DIP-trypsin at 1.8 A with a microcomputer technique. Acta Cryst 1977;B33:1824.
68. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pacific Symp Biocomput 1998;3:473–484.
69. Fehlhammer H, Bode W. The refined crystal structure of bovine beta-trypsin at 1.8 Å resolution. I. Crystallization, data collection and application of patterson search technique. J Mol Biol 1975;98:683–692.
70. Flick KE, Gonzalez L, Jr., Harrison CJ, Nelson HC. Yeast heat shock transcription factor contains a flexible linker between the DNA-binding and trimerization domains. Implications for DNA

binding by trimeric proteins. J Biol Chem 1994;269:12475–12481.
71. Gray CW, Brown RS, Marvin DA. Adsorption complex of filamentous fd virus. J Mol Biol 1981;146:621–627.
72. Gubser CC, Varani G. Structure of the polyadenylation regulatory element of the human U1A pre-mRNA 3′-untranslated region and interaction with the U1A protein. Biochemistry 1996;35:2253–2267.
73. Jaffray E, Wood KM, Hay RT. Domain organization of I kappa B alpha and sites of interaction with NF- kappa B p65. Mol Cell Biol 1995;15:2166–2172.
74. Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, Gastirel LN, Habuka N, Chen X, Maldanado F, Barker JE, Bacquet R, Villafranca, JE. Crystal structures of human calcineurin and the human FKBP12-FK506- calcineurin complex. Nature 1995;378:641–644.
75. Livnah O, Bayer EA, Wilchek M, Sussman JL. Three-dimensional structures of avidin and the avidin-biotin complex. Proc Natl Acad Sci USA 1993;90:5076–5080.
76. Matthews JR, Nicholson J, Jaffray E, Kelly SM, Price NC, Hay RT. Conformational changes induced by DNA binding of NF-kappa B. Nucleic Acids Res 1995;23:3393–3402.
77. Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, Luisi BF, Sigler PB. Crystal structure of trp repressor/operator complex at atomic resolution. Nature 1988;335:321–329.
78. Rayment I, Rypniewski WR, Schmidt-Base K, Smith R, Tomchick DR, Benning MM, Winkelmann DA, Wesenberg G, Holden HM. Three-dimensional structure of myosin subfragment-1: a molecular motor. Science 1993;261:50–58.
79. Riek R, Hornemann S, Wider G, Glockshuber R, Wuthrich K. NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231). FEBS Lett 1997;413:282–288.
80. Sturm RA, Herr W. The POU domain is a bipartite DNA-binding structure. Nature 1988;336:601–604.
81. Thomas DD, Ramachandran S, Roopnarine O, Hayden DW, Ostap EM. The mechanism of force generation in myosin: a disorder-to-order transition, coupled to internal structural changes. Biophys J 1995;68:135S–141S.