Predicting Intrinsic Disorder From Amino Acid Sequence

Zoran Obradovic,[1]* Kang Peng,[1] Slobodan Vucetic,[1] Predrag Radivojac,[1,2] Celeste J. Brown,[3] and A. Keith Dunker[3]*
[1]*Center for Information Science and Technology, Temple University, Philadelphia, Pennsylvania*
[2]*Molecular Kinetics, Pullman, Washington*
[3]*School of Molecular Biosciences, Washington State University, Pullman, Washington*

ABSTRACT **Blind predictions of intrinsic order and disorder were made on 42 proteins subsequently revealed to contain 9,044 ordered residues, 284 disordered residues in 26 segments of length 30 residues or less, and 281 disordered residues in 2 disordered segments of length greater than 30 residues. The accuracies of the six predictors used in this experiment ranged from 77% to 91% for the ordered regions and from 56% to 78% for the disordered segments. The average of the order and disorder predictions ranged from 73% to 77%. The prediction of disorder in the shorter segments was poor, from 25% to 66% correct, while the prediction of disorder in the longer segments was better, from 75% to 95% correct. Four of the predictors were composed of ensembles of neural networks. This enabled them to deal more efficiently with the large asymmetry in the training data through diversified sampling from the significantly larger ordered set and achieve better accuracy on ordered and long disordered regions. The exclusive use of long disordered regions for predictor training likely contributed to the disparity of the predictions on long versus short disordered regions, while averaging the output values over 61-residue windows to eliminate short predictions of order or disorder probably contributed to the even greater disparity for three of the predictors. This experiment supports the predictability of intrinsic disorder from amino acid sequence. Proteins 2003;53:566–572.**
© 2003 Wiley-Liss, Inc.

Key words: natively unfolded; intrinsically disordered; neural networks; ordinary least squares regression; machine learning

INTRODUCTION

Proteins or local protein regions that fail to form specific 3-D structure under physiological conditions have been called natively unfolded,[1] intrinsically unfolded,[2] and intrinsically disordered.[3] Such segments of intrinsic disorder have been identified by X-ray diffraction, hypersensitivity to protease digestion, and NMR spectroscopy, while wholly disordered proteins have been identified by NMR and CD spectroscopy and by hydrodynamic measurements. Since each of these methods has limitations, and since disorder can exhibit differences, from extended chains to collapsed globules, characterization of intrinsic disorder should ideally involve several of the methods just mentioned (reviewed in ref. 3)

For the fifth Critical Assessment of Structure Prediction (CASP5) experiment, local regions of disorder were identified by missing coordinates in their X-ray structures. A region of missing electron density in an X-ray-determined structure can result from causes other than intrinsic disorder, such as crystal packing irregularities or the rigid body wobble of an ordered domain,[4] so this simple measure sometimes gives incorrect disorder assignments. In addition, a wholly disordered protein was included in the target set. The absence of ordered 3-D structure for this target was indicated both by NMR and by CD spectroscopy.[5]

Our interest in intrinsically disordered protein was sparked in the early 1990s by discoveries on the filamentous phage coat protein. Others showed that a morphological change in the phage capsid activated the coat protein for spontaneous association with fluid lipid bilayers,[6,7] and we extended those findings with experiments indicating that the morphological transition was accomplished by a change from tight to loose side chain packing,[8,9] corresponding to a molten globule-like form.[10,11] These data suggested that a disordered, molten globule-like form, but not the ordered state, provided the basis for the important function of membrane penetration for this phage capsid protein. More recently we mined the literature to assemble a catalogue of functions for about 110 regions of intrinsic disorder: 28 distinct functions were found. These functions could be grouped into four broad categories: molecular recognition, molecular assembly/disassembly, protein modification, and entropic chain activities.[12,13] Especially important appears to be the use of intrinsic disorder for signaling and regulation,[14] with ~80% of cancer-associated

Grant sponsor: National Institutes of Health; Grant number: 1R01 LM06916; Grant sponsor: National Science Foundation; Grant numbers: CSE-IIS-971153 and CSE-IIS-0196237.

C.J. Brown's present address is IBEST, Department of Biological Sciences, University of Idaho, 83844-3051.

*Correspondence to either: Zoran Obradovic, Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-254), 1805 N. Broad St., Philadelphia, PA 19122. E-mail: zoran@ist.temple.edu; or else send correspondence to A. Keith Dunker's present address: Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202. E-mail: kedunker@iupui.edu

Received 17 February 2003; Accepted 17 June 2003

proteins predicted to have large regions of disorder.[15] Thus, predictions of intrinsic disorder may be useful for helping to identify regulatory domains, to identify domain boundaries, and to provide other important insights into protein structure-function relationships.

We initially attempted the predictions of order and disorder as a means to test the hypothesis that absence of 3-D structure is encoded by the sequence.[16] Later work used larger datasets and compared different prediction strategies (reviewed in ref. 13). Unknown to us but almost 20 years before our first published prediction of disorder, R. J. P. Williams showed that a low ratio of hydrophobic to charged residues could distinguish between a set of globular proteins and two proteins that were random coil-like in solution.[17] Using a much larger set of proteins, Uversky et al.[18] again applied the charge-hydropathy combination to distinguish natively unfolded from folded proteins.

Three years ago we contacted the organizers about including disorder predictions in the CASP experiment and worked informally with the organizers during CASP4. Disorder prediction was a new addition to CASP5. In the CASP5 experiment, we officially participated in two groups, the Obradovic group (predictors with id 454; first three authors) and the Dunker group (predictors with id 355; last three authors). Although all presented work is a result of our close collaboration over the last several years, we operated independently for CASP5. Here we report our joint contribution to the CASP5 experiment on the prediction of intrinsic disorder.

MATERIALS AND METHODS

Data Sets

The character of the datasets used in model training provided the bases for naming the predictors. The letter V in the name stands for variously characterized disordered regions, and X stands for X-ray characterized. The L designates that long regions of disorder (longer than 30 residues) were used for training while the T designates that the training sequences were from the termini.

VL1 data set[19]

The 15 internal long (L) disordered regions used in model training were characterized by various (V) methods including X-ray diffraction, NMR spectroscopy, circular dichroism and limited proteolysis. The order training set included random samples from NRL-3D.[20]

XT data set[21]

This disorder training set consisted of X-ray characterized (X) regions, 5-14 amino acids long, from the N- and C-termini (T), while the order set included the terminal regions of 130 completely ordered proteins chosen from a non-redundant set of proteins from the Protein Data Bank (PDB) called PDB-Select-25,[22] abbreviated here as PDBS25.

VL2 data set[23]

The VL1 dataset was expanded to 145 variously characterized, long regions of disorder. The set of 130 ordered proteins used for the XT dataset was also used here.

VL3 data set[24]

A set of 152 long regions of disorder characterized by various methods was derived from the VL2 dataset by removing some incorrectly identified sequences and by adding proteins (e.g. titin, neurofilament H) that were characterized by other physical means. The set of ordered proteins consisted of 290 PDB-Select-25 chains having no disordered residues.[25] These datasets can be found on our website: http://www.ist.temple.edu/DisProt.

Attribute Construction

An attribute or feature vector is constructed for each residue in a protein. We used amino acids within a symmetric input window of size W_{in}, since spatial conformation is largely influenced by neighboring amino acids. The input window extends/collapses at the N-/C-terminus. The first twenty attributes were the relative frequencies of each amino acid within the input window. Several other attributes were constructed for each position: the flexibility index,[26] hydropathy,[27] net charge and coordination number,[28] which were also averaged over the window W_{in}, and entropy, a measure of sequence complexity.[29] Feature selection was performed independently for each model such that accuracy was maximized.

Predictors

Naming of predictors follows the names of datasets used in training, while the designations after the names of the VL3 group of predictors specifies different techniques used in model training.

VLXT (CASP5 id: 355–1)

The VLXT predictor integrates three feed-forward neural networks: the VL1 predictor from Romero et al.[19] and the N- and C-terminal predictors (XT) from Li et al.[21] The attributes used by these predictors are coordination number, net charge and the relative frequencies of various combinations of W, F, Y, K, R, D, and E. Output for the VL1 predictor starts and ends 11 amino acids from the termini. The XT predictors output predictions up to 14 amino acids from their respective ends. A simple average is taken for the overlapping predictions; and a sliding window of 9 amino acids is used to smooth the prediction values along the length of the sequence. Unsmoothed prediction values from the XT predictors are used for the first and last 4 sequence positions.

VL2 (id: 355-2)

The VL2 predictor is a linear predictor[23] built using ordinary least-squares regression. It is based on 20 attributes (18 amino acid frequencies, average flexibility and sequence complexity) in an input window of length 41. For CASP5, the raw predictions were not averaged over an output window.

VL3 (id: 454-1)

The VL3 predictor is based on an ensemble of feed-forward neural networks and uses the same 20 attributes as VL2 (18 amino acid frequencies, average flexibility and sequence complexity) in an input window of length 41. The raw predictions are averaged over an output window of length 61 to obtain the final prediction for a given position.

VL3-BA (boundary augmented; id 355-4)

After prediction of disorder by VL3, the putative boundaries between order and disorder were corrected using the order/disorder boundary predictor described in Radivojac et al.[24] The closest maximum prediction from the boundary predictor (above 0.5) became the new boundary between the ordered and disordered regions. Input and output windows (W_{in} = 41, W_{out} = 31) were set to maximize the total sensitivity of prediction of long disordered regions.

VL3-H (homology; id: 454-2)

Using PSI-BLAST[30], a set of hypothetical disordered regions was found for each experimentally determined disordered region, thus enlarging the training set. Only homologous sequences from the range $10^{-20} <$ E-value $< 10^{-5}$ were retained and subsequently used in model training. Also, no two homologous sequences were allowed to have sequence identity above 90%. This predictor uses the same set of attributes and averages the input values and raw predictions in the same manner as VL3 (W_{in} = 41, W_{out} = 61).

VL3-P (profile; id 454-3)

The VL3-P predictor uses the evolutionary information both at the stage of training and prediction. In addition to the attributes exploited in other VL3 models, the sequence profile, generated by PSI-BLAST was used to derive 20 additional attributes for each example in the prediction process. Input and output windows were set to W_{in} = 41, W_{out} = 61.

Evaluation Criteria

The results presented here are for the 42 proteins available on Nov. 29, 2002: T0129, T0130, T0132, T0133, T0134, T0135, T0137, T0138, T0139, T0141, T0142, T0145, T0146, T0147, T0148, T0149, T0150, T0153, T0157, T0159, T0160, T0165, T0167, T0168, T0169, T0170, T0172, T0173, T0174, T0182, T0183, T0184, T0185, T0186, T0187, T0188, T0189, T0190, T0191, T0192, T0193, T0195.

The frequency of disordered residues is typically small in proteins characterized by X-ray diffraction. Always predicting the majority class in such an unbalanced dataset yields a predictor of very high accuracy that completely misses the event in which we are interested, in this case, disorder. A common practice in class-imbalance problems is to separately estimate accuracies on the minority class (SN) and the majority class (SP) and take their average to give the overall accuracy (A = (SN + SP)/2). In such a case, a random predictor or a predictor outputting only one class will have an accuracy of 50%.

Large variability in lengths of disordered segments, however, may cause a few proteins with very long disordered regions to completely outweigh many other proteins with significantly shorter segments. For example, the 2,174 residue segment of elastic titin (gi:1017427) and the 1,827 residue segment of microtubule-associated protein 2, isoform 1 (gi:14195624) can dominate many proteins with significantly shorter disordered segments. Since this is not the case for this data set, we present the A measure.

TABLE I. Number of CASP5 Targets With Various Percentages of Disordered Residues

Disordered residues (%)	No. of CASP targets
0	14
1–5	15
6–10	6
11–15	4
16	1
25	1
100	1

TABLE II. Comparison of Disordered Residues in the CASP5 Targets and in PDB-Select-25

	CASP5	PDBS25
No. of chains	42	1223
No. of residues	9,609	239,527
No. of chains with no disorder	14	527
Chains with no disorder (%)	33	32
No. of disordered regions ≤ 30 (residues)	24 (284)	1,100 (8,428)
Disordered residues in regions ≤ 30 (%)	3.0	3.5
No. of disordered regions > 30 (residues)	2 (281)	68 (3,710)
Disordered residues in regions > 30 (%)	2.9	1.5

RESULTS

Disordered Residues in CASP5 and in PDB-Select-25

Of the 42 targets used in our evaluation, 14 had no disordered residues, while 26 had at least one disordered region (Table I). Most of these 26 proteins had only short regions of disorder, while almost half of the disordered residues were in two targets (281 residues total), and the rest of the residues (284) were distributed among 24 targets (Table II).

The ordered and disordered data from the CASP5 targets were compared with the data from PDBS25 as of Oct. 31, 2001 (Table II). While the CASP5 set consisted of 42 proteins with 9,609 total residues, the PDBS25 contained 1,223 chains with 239,527 residues. The two datasets have quite similar percentages of proteins with no disorder, 33% and 32% for CASP5 and PDBS25, respectively, and also quite similar percentages of residues in disordered regions of length $\leq$30 residues, 3.0% and 3.5% for CASP5 and PDBS25, respectively. The CASP5 set had a higher fraction of residues in disordered regions of length $>$30 residues due to the inclusion of one entirely disordered protein.

TABLE III. Prediction Accuracies and Standard Errors (%) Estimated on CASP5 Targets[a]

Model	SN[b]	SP[c]	*A*[d]	SN (≤30)[e]	SN (>30)[f]
VLXT	68.5 ± 1.9	77.1 ± 0.4	72.8 ± 1.0	58.0 ± 3.0	79.0 ± 2.5
VL2	78.1 ± 1.7	76.6 ± 0.5	77.4 ± 0.9	65.9 ± 2.8	90.1 ± 1.8
VL3	56.3 ± 2.1	90.9 ± 0.3	73.6 ± 1.0	25.0 ± 2.6	86.2 ± 2.1
VL3-BA	61.6 ± 2.0	89.3 ± 0.3	75.5 ± 1.0	49.2 ± 3.0	75.2 ± 2.7
VL3-H	58.7 ± 2.0	89.4 ± 0.3	74.1 ± 1.0	35.9 ± 2.9	86.1 ± 2.1
VL3-P	62.8 ± 2.0	89.8 ± 0.3	76.3 ± 1.0	31.0 ± 2.8	95.0 ± 1.4

[a] Estimated on 42 target structures released on Nov. 29, 2002 (including T0145).

[b] SN, % correct predictions of disorder.

[c] SP, % correct predictions of order.

[d] *A*, (SN + SP)/2.

[e] SN (≤30), % correct predictions of disordered regions 1–30 residues in length ($N = 284$).

[f] SN (>30), % correct predictions of disordered regions greater than 30 residues in length ($N = 281$).

Prediction Summary for the 42 Target Proteins

The prediction accuracies and bootstrap estimated standard errors[31] for the six predictors are given in Table III. For each model all target proteins were connected into a single long sequence of length 9,609 from which 1,000 bootstrap samples (sequences of length 9,609) were generated by randomly sampling residues with replacement. For each bootstrap replicated sample, prediction accuracy was determined and stored. Finally, the overall accuracy and its standard error were estimated from each stored array. The sensitivities were calculated separately for disordered regions 30 residues or shorter and longer than 30 residues because all of the predictors were trained and optimized on long regions of disorder.

All of the predictors have greater than 70% average accuracy on the CASP5 targets, although there is considerable variability in sensitivity and specificity. While the first 2 predictors exhibited about 77% accuracy in the prediction of ordered residues, the last 4 predictions exhibited about 90% accuracy. This >10% increase in accuracy correlated with the use of ensembles of predictors for the last 4 models.

Dividing the disordered regions into short (≤30 residues) and long (>30 residues) categories revealed additional information about the various predictors. All 6 predictors had much higher accuracy on long as compared to short regions of disorder. Three predictors averaged their output values over 61 residues in order to decrease short false positive predictions of order and disorder; these predictors showed an even larger difference in accuracy between short versus long disordered segments.

VLXT and VL2 Predictions on Individual Proteins

The charge-hydropathy plot developed by Uversky et al.[18] was applied to the CASP5 targets, with the result that 2 proteins, T0145 and T0170 were predicted to be entirely unfolded, and 2 other proteins, T0129 and T0174, were predicted to be ordered but were very close to the order-disorder dividing line (data not shown). Representative per-residue predictions using VLXT (green) and VL2 (black) are shown for these 4 proteins in Figure 1. T0145 was revealed to be entirely disordered and T0170 to be entirely ordered, while T0129 and T0174 were mostly ordered with various disordered segments as indicated. Overall, the charge-hydropathy plot, VLXT and VL2 all concurred that T0145 is a natively unfolded protein and that T0129 and T0174 are (mostly) ordered proteins. The various methods gave inconsistent results for T0170, with the charge-hydropathy plot and VL2 indicating a natively unfolded protein, but with VLXT indicating an ordered protein.

DISCUSSION

Order Predictions

The VL3 predictors exhibited a >10% improvement in the prediction of order as compared to the first 2, from about 90% as compared to about 77%, respectively. These four predictors averaged the results from ensembles of predictors, where the individual predictors used the same set of disordered residues but different random draws from the much larger set of ordered residues. This would be expected to give better coverage of ordered sequence space and thus could be contributing to the improved prediction of order. A second factor is that the outputs were averaged over a larger number of residues. This would improve the order prediction accuracy by reducing short, false positive predictions of disorder. The relative importance of these two factors has not been determined.

Disorder Predictions on Short Regions of Disorder

All of the predictors discussed here were trained on long regions of disorder, with one result being that VLXT gives a high error rate predicting short regions of disorder. Thus, in our submission to CASP5 we included in the REMARKS section the caveat that internal regions of predicted disorder less than 15 amino acids in length were too short to be considered significant. Similarly, the VL3 predictors were optimized to remove short regions of both predicted order and disorder thus sacrificing prediction accuracy on short segments. Since over half of the disordered residues in the CASP5 targets are in short regions, it is not surprising that these predictors had fairly low sensitivities. VL2 was also originally optimized on an output window size of 41,[23] however, for CASP5 the output

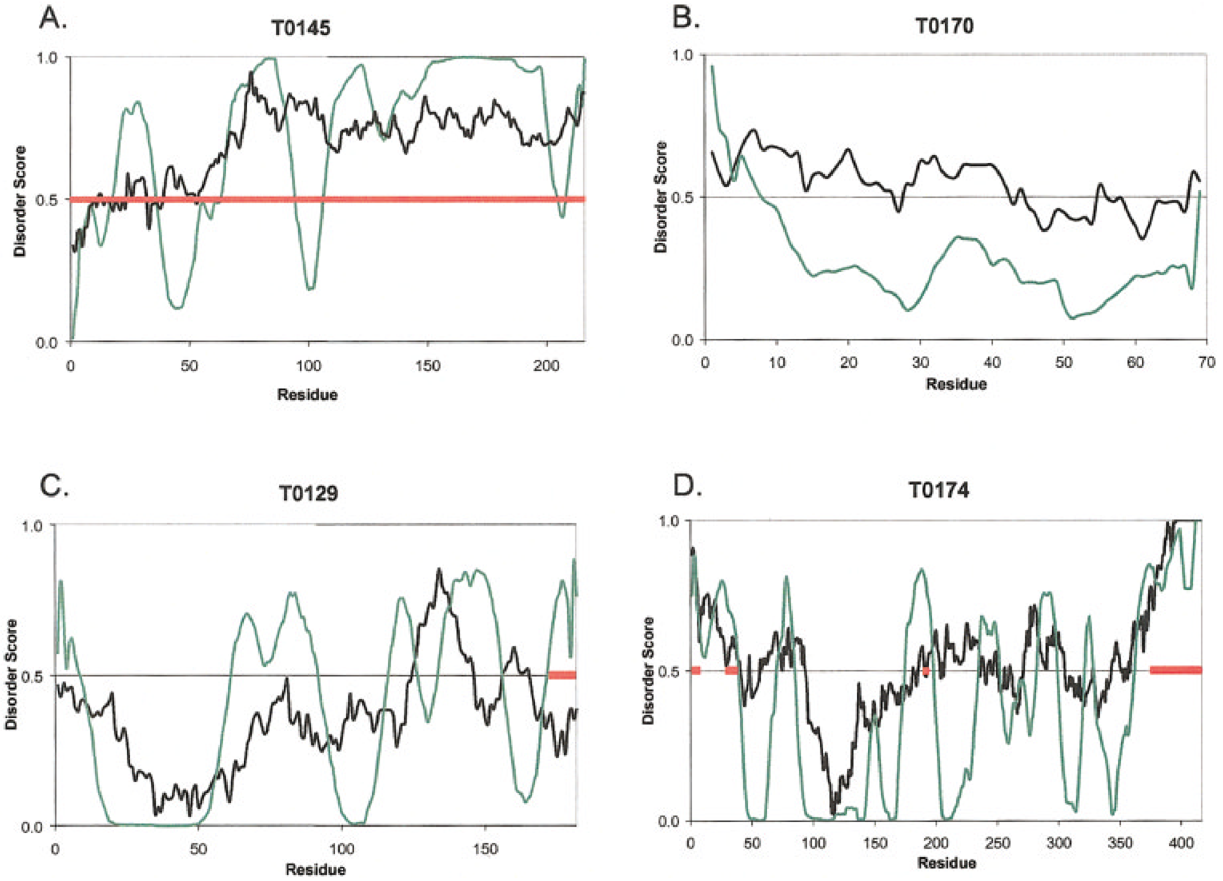

Fig. 1. VLXT (green) and VL2 (black) disorder scores for targets: A. T0145, an entirely disorderd protein; B. T0170, an entirely ordered protein (a 3-helix bundle); C. T0129, contains disordered region: 172-182; D. T0174, contains disordered regions: 1-7, 29-38, 190-194, 375-417. Disordered regions are marked in red. A prediction score above or equal to 0.5 is considered disordered and below 0.5 is considered ordered.

window size was changed to 1. This change resulted in increased sensitivity as compared to the VL3 models.

Disorder Predictions on Long Regions of Disorder

All of the predictors had better sensitivity on long regions of disorder than on short regions. The VL3-P predictor had the highest sensitivity on long disordered regions of any of the remaining five predictors. To train this predictor, sequence profiles for disordered proteins were developed based upon the alignment of homologous disordered regions. The profiles were then used to train an ensemble of neural networks exploiting an inherent ability of ensembles to sample from a larger pool of ordered regions for each network, improving the overall specificity. The fundamental assumption made in this method is that regions of proteins that are homologous to a disordered region of a protein are also disordered. The high success rate of this predictor indicates that this assumption is likely to be correct, and that disordered structure is conserved even when disordered sequence is poorly conserved.[32]

Prediction Comparisons

VLXT had the lowest overall accuracy of the six predictors and VL2 had the highest (Table I). The training set differences between VLXT and VL2 may explain these results. VL2 was trained on a much larger set of disordered proteins, and it was trained on a non-redundant set of completely ordered proteins. Another difference is that VL2 was a linear predictor, while VLXT was an integration of 3 neural networks. We do not believe that the use of a linear predictor rather than a neural network predictor accounts for the improved accuracy of VL2, however, because neural network predictors developed on the same data sets had slightly higher training accuracy than the linear predictor.

The apparent prediction accuracy on short disordered regions is reduced by averaging over large windows such as for VL3, VL3-H, and VL3-P, while the prediction accuracy for long disordered regions is improved. This leads to a greater disparity in the prediction accuracy of short versus long regions of disorder (Table III).

Since all of the predictors were trained on long regions for both order and disorder, it is reasonable to compare their performance using just these data. By this measure, VL3-P appears to be the best of the 6 predictors in this experiment, with about 90% correct prediction of order and about 95% correct prediction of long regions of disorder.

Comparing the predictors for several specific proteins illustrates their successes and failures. Figure 1 illustrates the predictions for four targets, the wholly disordered T0145, the ordered T0170, and two mostly ordered proteins, T0129 and T0174. As indicated above, T0145 and T0170 were indicated to be wholly disordered by the charge-hydropathy plot,[18] while T0129 and T0174 were indicated to be ordered, but were close to the order-disorder dividing line in this plot (data not shown).

All 6 predictors and the charge-hydropathy plot all correctly indicated that T0145 is natively unfolded. Interestingly, our 6 predictors concurred in indicating a short region of order at the N-terminus as shown in Figure 1 for two of the predictors. Perhaps these results suggest latent 3-D structure, such as a short α-helix, in this region. Indeed, VLXT has successfully detected several binding sites that undergo disorder-to-order transitions upon binding with a partner by predicting short regions of order within longer stretches of predicted disorder,[33] and this short prediction of order might correspond to just such a region.

Although T0170 is ordered, both the charge-hydropathy plot and 4 of the 6 predictors indicated this protein to be completely or at least mostly unfolded. This protein is a 3-helix bundle, which indicates a rather large surface-to-volume ratio. In addition, this protein binds to DNA and has a very high positive charge. The high net charge and large-surface to volume ratio accounts for the prediction error by the charge-hydropathy plot and certainly contributes to the large errors made by 4 of the 6 predictors. VLXT predicted 88% of this protein to be ordered, but this apparent success may relate to the small training set used to develop this predictor. Interestingly, VL3 predicted just 28% of the residues to be ordered, while VL3-BA predicted 78% to be ordered, suggesting perhaps that the boundary augmentation improved the prediction of this protein. As indicated in Table III, boundary augmentation led to a general improvement in the VL3 predictor on short regions of disorder, but led to degradation of the accuracy for long regions of disorder. Since the long regions of disorder in this study comprised just 2 segments, further study on a larger sample is needed to further test the effects of boundary augmentation.

The last 11 residues of T0129 (Fig. 1C) are disordered, and only one of our six predictors, VLXT, correctly predicted this region. All 6 of the predictors mistakenly indicated that a highly charged coil from residues 136-146 is disordered. The atoms in this coil have very high B-factors. It would be interesting to know whether this region is involved in crystal contacts.

T0174 contained the second largest amount of disorder of the CASP5 targets, one segment of 43 residues, as well as 3 short regions of disorder. All of our predictors accurately predicted the disordered termini, however, their disorder predictions extended into the ordered region by varying lengths; VL3-H incorrectly predicted an additional 178 residues to be disordered in the C-terminus! With regard to disorder prediction for the N-terminus, the presence of a second disordered region from residues 29-38 probably contributed to the over-prediction of disorder in this region. Furthermore, the false positive disorder prediction errors are not consistent across different predictors (Figure 1D), suggesting that development of combined predictors is worth exploring.

Future Directions

The 6 predictors included 2 that were described in previous publications, VLXT and VL2, and 4 new ones. The evident improvement of the new ones on long regions of disorder suggest that progress is being made, with the VL3-P accuracy exceeding 90% for long regions order and disorder. The poor performance on short disordered regions calls for additional work on these. Such short disordered regions are likely to be more context dependent than long disordered regions. Furthermore, the same boundary error contributes a greater percentage to a short as compared to a long region of disorder. Thus, predicting disorder in short regions is likely to be more difficult than predicting disorder in long regions, but, compared to predictions on long regions of disorder, predictions on short regions have much greater room for improvement.

ACKNOWLEDGMENTS

While working at Molecular Kinetics, P.R. was supported by funds provided by Amgen, Inc. of Thousand Oaks, CA.

REFERENCES

1. 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.
2. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 1999;293:321–331.
3. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW and others. Intrinsically disordered protein. J. Mol. Graph. Model. 2001;19:26–59.
4. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. Genome Inform. Ser. Workshop Genome Inform. 1998;9:201–213.
5. Melamud, E. and Moult, J. (2003) Evaluation of disorder predictions in CASP5. Proteins. 2003;6:562–566.
6. Manning M, Chrysogelos S, Griffith J. Insertion of bacteriophage M13 coat protein into membranes. Biophys. J. 1982;37:28–30.
7. Manning M, Griffith J. Association of M13 I-forms and spheroids with lipid vesicles. Arch. Biochem. Biophys. 1985;236:297–303.
8. Dunker AK, Ensign LD, Arnold GE, Roberts LM. Proposed molten globule intermediates in fd phage penetration and assembly. FEBS Lett. 1991;292:275–278.
9. Dunker AK, Ensign LD, Arnold GE, Roberts LM. A model for fd phage penetration and assembly. FEBS Lett. 1991;292:271–274.
10. Dolgikh DA, Gilmanshin RI, Brazhnikov EV, Bychkova VE, Semisotnov GV, Venyaminov S, Ptitsyn OB. Alpha-Lactalbumin: compact state with fluctuating tertiary structure? FEBS Lett. 1981;136:311–315.
11. Ohgushi M, Wada A. 'Molten-globule state': a compact form of globular proteins with mobile side-chains. FEBS Lett. 1983;164: 21–24.
12. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41: 6573–6582.
13. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv. Protein Chem. 2002;62:25–49.
14. Dunker AK, Obradovic Z. The protein trinity - linking function and disorder. Nat. Biotechnol. 2001;19:805–806.
15. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK.

Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 2002;323:573–584.

16. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. IEEE Int. Conf. Neural Netw. 1997;1:90–95.
17. Williams RJP. The conformational mobility of proteins and its functional significance. Biochem. Soc. Trans. 1978;6:1123–1126.
18. Uversky V, Gillespie J, Fink A. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
19. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48.
20. 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.
21. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. Genome Inform Ser Workshop Genome Inform 1999;10:30–40.
22. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci. 1994;3:522–524.
23. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573–584.
24. Radivojac P, Obradovic Z, Brown CJ, Dunker AK. Prediction of boundaries between intrinsically ordered and disordered protein regions. Pac. Symp. Biocomput. 2003;8:216–227.
25. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Science 2003; 12: 1060–1072.
26. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
27. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 1982;157:105–132.
28. Galaktionov SG, Marshall GR. Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D. In: States DJ, Agarwal P, Gaasterland T, Hunter L, Smith RF, editors. Proc. Int. Conf. Intell. Syst. Mol. Biol.; 1996 June 12–15, 1996; Washington University Institute for Biomedical Computing, Center for Molecular Design, St. Louis, MO. AAAI Press. p 42.
29. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.. 1997;25: 3389–3402.
31. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
32. Brown CJ, Takayama S, Campen AM, Vise P, Marshall T, Oldfield CJ, Williams CJ, Dunker AK. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 2002;55:104–110.
33. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting binding regions within disordered proteins. Genome Inform. Ser. Workshop Genome Inform. 1999;10:41–50.