Exploiting Heterogeneous Sequence Properties Improves Prediction of Protein Disorder

Zoran Obradovic,[1]* Kang Peng,[1] Slobodan Vucetic,[1] Predrag Radivojac,[2,3] and A. Keith Dunker[3]
[1]*Center for Information Science and Technology, Temple University, Philadelphia, Pennsylvania*
[2]*School of Informatics, Indiana University, Bloomington, Indiana*
[3]*Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana*

ABSTRACT **During the past few years we have investigated methods to improve predictors of intrinsically disordered regions longer than 30 consecutive residues. Experimental evidence, however, showed that these predictors were less successful on short disordered regions, as observed two years ago during the fifth Critical Assessment of Techniques for Protein Structure Prediction (CASP5). To address this shortcoming, we developed a two-level model called VSL1 (CASP6 id: 193-1). At the first level, VSL1 consists of two specialized predictors, one of which was optimized for long disordered regions (>30 residues) and the other for short disordered regions (≤30 residues). At the second level, a meta-predictor was built to assign weights for combining the two first-level predictors. As the results of the CASP6 experiment showed, this new predictor has achieved the highest accuracy yet and significantly improved performance on short disordered regions, while maintaining high performance on long disordered regions. Proteins 2005; Suppl 7:176–182.** © **2005 Wiley-Liss, Inc.**

Key words: disorder prediction; intrinsically disordered; length dependent predictors

INTRODUCTION

Intrinsically disordered proteins or protein regions are characterized by their highly dynamic conformations under putatively physiological conditions in which the atom coordinates and backbone Ramachandran angles vary significantly over time with no specific equilibrium values.[1–5] Instead of folding into a fixed 3D structure, a disordered protein or region exists as an ensemble of noncooperatively interchanging structures. Several operational definitions for protein disorder exist, including random coils,[6,7] high C_α atom B-factor regions,[6,7] dynamically flexible ensembles,[8] absence of regular secondary structure (NORS),[9] and missing coordinates for backbone atoms.[6–8,10] These different definitions provide insight into properties of flexible regions and their relationships to disorder, and are consistent with the view that there are several flavors of intrinsic protein disorder.

Although lacking specific 3D structures, intrinsically disordered proteins or regions carry out, and are required for, essential biological functions. These functions were broadly categorized by Dunker et al.[11] into molecular recognition, molecular assembly, protein modification, and entropic chain activities, while Tompa added scavenger[4] and chaperone[12] functions among others. Disordered proteins or regions can be characterized by various experimental methods,[13,14] but current technologies are still prohibitive for genome-scale analyses. Thus, the ability to reliably predict disordered regions from amino acid sequence is important, and could have significant impact on a wide range of biomedical research, for example, from the design of protein structure–function experiments[1,15] to the understanding of the roles of disorder in cell-signaling and cancer-related proteins[16] and also to the ongoing structural genomics projects.[17–19] As of 2002, the prediction of disordered regions has been externally evaluated as part of CASP.[10] Since the involvement of CASP, disorder prediction has been attracting increased interest.[6–8,20–22]

In the CASP5 experiment we assessed six predictors of intrinsically disordered regions[23] and all of them achieved greater than 70% overall accuracies. These predictors, along with those developed by other groups, strongly support the hypothesis that intrinsic disorder is encoded by amino acid sequence.[6–8,24–29] However, the CASP5 results also revealed that our predictors were significantly less accurate on short disordered regions (≤30 residues) compared to long disordered regions (>30 residues), with accuracies of 25–66% versus 75–95%, for short versus long regions of disorder, respectively. One possible reason for such a discrepancy was the use of large windows for attribute construction (e.g., 41 residues) or output smoothing (e.g., 61 residues), which improved prediction on long disordered and ordered regions, but also filtered out many predicted short regions. Another potential contributing factor is the possibility of heterogeneous amino acid compositions between the short and long disordered regions. Because all attributes were derived from the amino acid sequence, a predictor trained exclusively on long disor-

Grant sponsor: NIH; Grant number: R01 LM007688-01A1 (to A.K.D. and Z.O.).

*Correspondence to: Zoran Obradovic, Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 North Broad Street, Philadelphia, PA 19122. E-mail: zoran@ist.temple.edu

Received 12 April 2005; Accepted 26 May 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20735

The article was originally published online as an accepted preprint. The “Published Online” date corresponds to the preprint version.

dered regions was unlikely to perform well on short disordered regions, and *vice versa*.

In our initial study on the prediction of intrinsic disorder,[28] four predictors were built on disordered regions: three based on different length groups, that is, short (7–21 residues), medium (22–44 residues), and long (45 residues or longer), respectively, and one based on all of the disordered regions combined. All four predictors used the same attributes based on amino acid compositions. All three length-specific predictors outperformed the all-length predictor, with 9–14% accuracy improvements, and when any one of the three predictors was applied to disordered regions from the other two groups, the prediction accuracy dropped significantly, by 6–14%. These experiments were the first indication of a length-dependency of the amino acid compositions of disordered regions. In a more recent study,[30] a set of short disordered regions of ≤10 residues was extracted from the Protein Data Bank (PDB)[31] and then compared to regions of long disorder, high B-factor order and low B-factor order. This study showed that short disordered regions exhibited significantly different amino acid compositions compared to long disordered regions and appeared to be more similar to the high B-factor ordered regions.

Based on these findings, we developed a composite predictor called VSL1[32] (Various Short Long, version 1), which is applicable to disordered regions of arbitrary length (The predictor naming convention can be found in Obradovic et al.;[23] in brief, "Various" indicates training data characterized by different methods, while "Short" and "Long" indicate the lengths of the disordered segments). This predictor exploits the length dependent (heterogeneous) amino acid compositions and sequence properties of disordered regions by using a two-level structure to integrate two specialized predictors optimized for short (≤30 residues) and long (>30 residues) disordered regions, respectively. As results in the latest CASP6 experiment showed, the VSL1 predictor significantly improved the prediction on short disordered regions, while retaining high accuracy on long disordered regions that characterized our previously developed predictors. In this report of our results for the CASP6 experiment, we analyze the performance of the new VSL1 predictor as well as four other predictors of intrinsically disordered regions developed previously.

MATERIALS AND METHODS

Training Data

The dataset for training the VSL1 predictor contained 1,335 nonredundant protein sequences with maximal pairwise sequence identity limited to 25%. These proteins were assembled from four other datasets: 153 sequences from DisProt[33] v1.2 with long disordered regions characterized by various methods,[23,27] 511 PDB chains with short disordered regions identified as missing backbone atom coordinates,[22] 290 completely folded PDB chains,[23,27,30,34] and 381 recent PDB chains with short disordered regions. In total, the training proteins contain 230 long disordered regions with 25,958 residues, 983 short disordered regions with 9,632 residues, and 354,169 ordered residues.

VSL1 Predictor

The VSL1 predictor consists of three component predictors, each as an ensemble of logistic regression models, in a two-level architecture. At the first level are two specialized predictors: a long disorder predictor, VSL1-L, for disordered regions of >30 residues, and a short disorder predictor, VSL1-S, for disordered regions of ≤30 residues. At the second level is a meta-predictor, VSL1-M, whose output can be interpreted as the likelihood that a 61-residue subsequence centered at current sequence position contains or overlaps a *long* disordered region. Thus, weights are derived from the output of VSL1-M for combining the two first-level predictors. For all three predictors, attributes are constructed for each sequence position based on an input window of length W_{in} (odd number) centered at that position. The window is extended outside the N- and C termini by concatenating $(W_{\text{in}}-1)/2$ special "spacer" characters at each terminus. The attributes calculated include amino acid frequencies, the "spacer" frequency, K_2-entropy,[35] charge-hydrophobicity ratio,[36] averaged flexibility index,[37] averaged PSI-BLAST[38] profiles, averaged PHD[39] and PSIPred[40] secondary structure predictions. Attribute selection and window length optimization were performed independently for individual predictors. Further details of the VSL1 predictor will be described elsewhere.[32]

Previous Long Disorder Predictors

In the CASP6 experiment we also tested four previously developed predictors of long disorder: VL-XT[25] (id: 633-1), VL3-BA[23] (id: 633-2), VL2[26] (id: 193-3), and VL3-E[27] (id: 193-2), which enabled us to estimate the progress over time. The VL-XT predictor is a combination of three feedforward neural network predictors specialized for N-terminal, C-terminal and internal regions, respectively. VL2 is a linear predictor built using ordinary least-squares regression.[41] As in CASP5, VL2 predictions were not smoothed. VL3-BA uses a disorder/order boundary predictor to correct the putative boundaries between predicted disordered/ordered regions by the VL3[23,27] predictor. The VL3-E predictor combines two neural network ensemble predictors VL3-H and VL3-P,[23,27] and is currently our best predictor of long disordered regions with overall accuracy higher than 86%.

Evaluation Criteria

The predictors were evaluated on the 63 CASP-curated target structures released on November 18, 2004. These targets have the following CASP identifications: T0196–T0206, T0208, T0209, T0211–T0216, T0222–T0224, T0226–T0235, T0238–T0244, T0246–T0249, T0251, T0262–T0269, T0271–T0277, and T0279–T0282. Among these targets, T0222 (PDB id: 1VLI) is a theoretical model, T0213–T0215, T0224, and T0230 are NMR structures, while the remaining examples are X-ray structures.

TABLE I. Summary of the 63 CASP6 Targets

	Length range	Number of regions	Number of residues
Disordered regions	1–3	35	58
	4–15	41	304
	16–30	9	201
	31–100	4	266
	>100	1	102
	Total	90	931
Ordered regions		90	12,520

Two criteria were used for predictor evaluation. The overall accuracy (ACC) was calculated as the average of *true positive rate* (*sensitivity*), or percentage of disordered residues correctly predicted, and *true negative rate* (*specificity*), or percentage of ordered residues correctly predicted, using a decision threshold of 0.5. A random predictor or a trivial predictor that assigns all examples to one class will have an overall accuracy of 50%. The receiver operating characteristic (ROC) curve is a plot of *true positive rate* against *false positive rate* (or 1 − *specificity*), usually calculated at different decision thresholds. The area under the ROC curve (AUC) is known to be a useful measure of overall predictor quality, with a value of 100 for a perfect predictor and 50 for a random predictor.[42]

RESULTS

CASP6 Targets

A total of 90 disordered regions with 931 residues were identified as missing backbone atom coordinates from 48 of the 63 targets used in our evaluation (Table I). None of the 48 targets was a wholly disordered protein, while of the 90 disordered regions, only 5 were longer than 30 residues and came from T0206, T0235, T0238, T0249, and T0262. The longest disordered region contained the 102 N-terminal residues of T0235. In total, the long disordered regions contained about 40% (368) of all disordered residues. Of the 85 disordered regions of ≤30 residues, 33/25 were at N-/C-termini containing 205/125 residues in total. Finally, the 63 targets also contained 90 ordered regions with 12,520 residues.

Prediction Accuracies

Figure 1 shows the ROC curves plotted by varying the decision threshold from 0 to 1 in increments of 0.005. Table II compares the AUC approximated using the *trapezoid rule*, and the *per-residue* prediction accuracies calculated with the default threshold of 0.5. The standard errors were calculated with 1000 bootstrap samples using a procedure described in Obradovic et al.[23] Among the five predictors tested in CASP6, VSL1 achieved the highest overall accuracy (ACC = 79.4%) and the greatest area under the ROC curve (AUC = 88.3). In addition, VSL1 was clearly better than either one of its two component predictors alone. This indicates that the meta-predictor VSL1-M was effective in combining the two specialized predictors.

As illustrated in Table II, VSL1 was significantly better than all other predictors on short disordered regions (≤30 residues), but considerably less accurate than VL3-E on long disordered regions (>30 residues). To better characterize the differences, we further divided the disordered regions into several length groups and examined the accuracy on each group separately. As evident in Figure 2, the improvement of VSL1 mainly came from the short disordered regions of 1–3 and 4–15 residues, but it was just slightly better than VL2 on short disordered regions of 16–30 residues. On the four long disordered regions of 31–100 residues, VL3-E was nearly perfect (98.5%) while VSL1 also achieved a high accuracy of 82.7%. However, all predictors were less successful on the longest disordered region (102 residues) from T0235, with VL3-E being the most accurate.

Predictions on Individual Targets

Representative predictions by VL-XT (dashed), VL3-E (dashed-dotted), and VSL1 (solid) for targets T0233 and T0201 are shown in Figure 3. Although VSL1 successfully identified all four disordered regions (thick line segments) from T0233 except residue 106, VL-XT predicted part of the second region and completely missed the third one, and VL3-E detected only the first region. In addition, VSL1 had significantly fewer false positives than VL-XT on ordered regions.

For the fully folded T0201 (PDB id: 1S12, 4 chains), VL-XT achieved the highest accuracy of 80.8%, while both VSL1 and VL3-E predicted it to be mostly disordered from residue 38 to 94. The significant enrichment of disorder-promoting residues[24] in this region, namely 11 Lysines, 9 Glutamic Acids, 1 Proline, 4 Serines, 2 Glutamines, 2 Arginines, and 4 Alanines, might have contributed to the large error rates of VSL1 and VL3-E. However, it is also possible that this region does have a flexible structure under physiological conditions, but became stabilized due to the high-salt (1.4 M NaCitrate) crystallization conditions used and/or the multimer formation.

Predictions on High B-Factor Regions

After excluding residues from disordered regions, the five NMR target structures, and T0227 with constant B-factors over its entire sequence, a total of 11,980 residues remained with valid B-factors. The B-factors, averaged over the backbone atoms for each residue, were first normalized to zero mean and unit variance, chain by chain, using a procedure by Smith et al.[34] The residues were then assigned to two groups as *high-B-factor* (1,014 residues) and *low-B-factor* (10,966 residues) depending on whether their normalized B-factor values were higher than 2.0.[34] Interestingly, VSL1 had significantly higher false positive error rate on *high-B-factor* than *low-B-factor* regions (39.8% *versus* 14.7%).

Predictions on Terminal Short Disordered Regions

On the 33 N-terminal and 25 C-terminal short disordered regions, the VSL1 predictor achieved high accuracies of 94.4 ± 1.6% and 82.4 ± 3.4%, respectively, but was less successful on the 27 internal short disordered regions with an accuracy of 63.9 ± 3.1%. Although the terminal

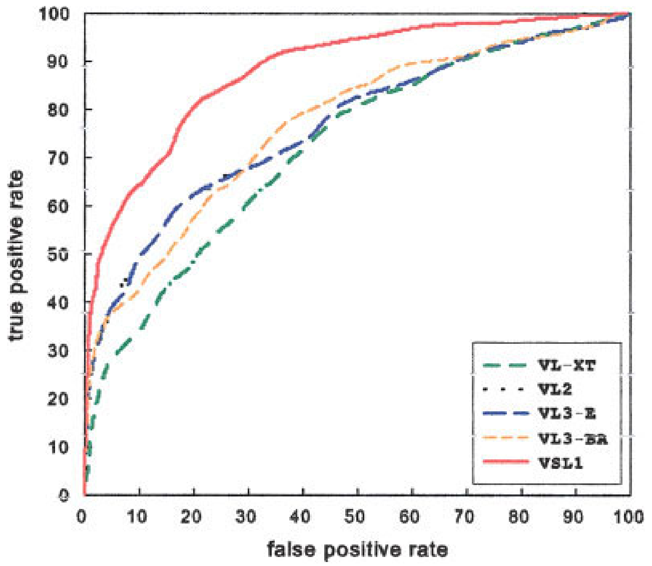

Fig. 1. Comparison of ROC curves for five models tested in CASP6 experiment. ROC curves were plotted by varying threshold from 0 to 1 in increments of 0.005.

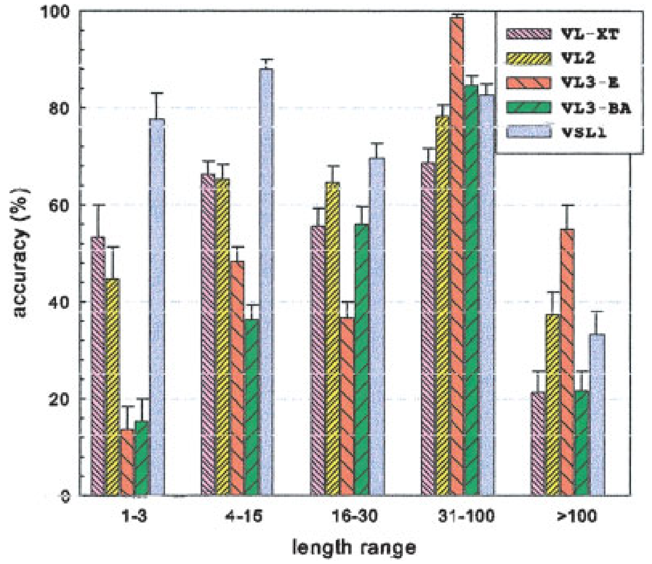

Fig. 2. Length-dependent per-residue prediction accuracies for the five models tested in CASP6 experiment. The standard errors were estimated over 1000 bootstrap samples.

short disordered regions came from 45 targets with a total of 330 residues, VSL1 predicted disordered regions on almost all 63 targets at 59 N- and 57 C-termini with a total of 1,307 residues.

DISCUSSION

VSL1 Predictor

The success of VSL1 model can be attributed to both the enlarged training data and its two-level architecture that exploits the length dependent (heterogeneous) amino acid compositions and sequence properties of disordered regions. Not only was the data size increased substantially, but more importantly, it included a considerable number of short disordered regions that were not used in the training of previous predictors. Under the two-level architecture, the specialized predictors, VSL1-S and VSL1-L, could be optimized separately on more homogeneous data, in terms of attribute selection, model selection, and window optimization. On the other hand, the meta predictor VSL1-M proved to be effective in combining the two specialized predictors and its success further confirmed the previously observed differences between short and long disordered regions.

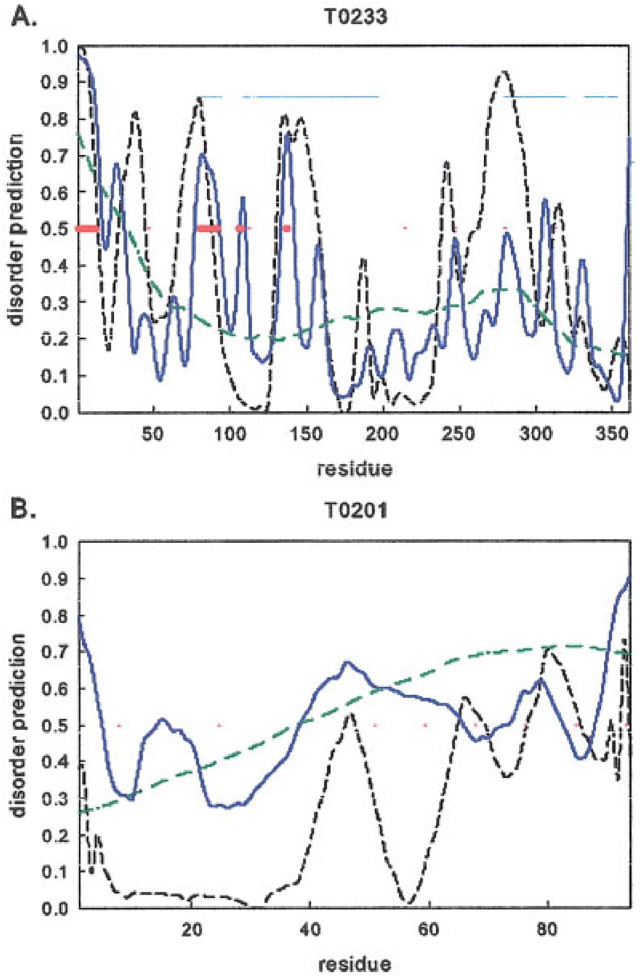

Fig. 3. Representative predictions by VL-XT (dashed), VL3-E (dashed-dotted), and VSL1 (solid) for two targets: (**A**) T0233, with four short disordered regions at residues 1–13, 81–92, 106–108, and 137–138; (**B**) T0201, a completely ordered protein. Disordered regions are marked in thick line segments. The threshold for predicting disorder is 0.5.

The threshold of 30 for partitioning disordered regions into *short* and *long* is artificial[11] and may not necessarily be optimal. A better approach might be to identify different length groups on the basis of maximizing the Kullback-Leibler distance of amino acid compositions between the different groups. It is also possible that certain short disordered regions may share similar properties to long disordered regions and *vice versa*. Therefore, the initial partitioning obtained by the length threshold might be

TABLE II. Prediction Accuracies (%) and Areas under ROC Curves (AUC) on 63 CASP6 Targets

Model	TP[a]	TN[b]	TP$_S$[c]	TP$_L$[d]	ACC[e]	AUC[f]
VL-XT	59.1 ± 1.7	71.3 ± 0.4	61.3 ± 2.1	55.7 ± 2.7	65.2 ± 0.9	72.4 ± 0.85
VL2	64.6 ± 1.6	76.2 ± 0.4	63.1 ± 2.1	66.8 ± 2.4	70.4 ± 0.8	79.2 ± 0.85
VL3-E	58.8 ± 1.6	83.9 ± 0.3	40.7 ± 2.1	86.4 ± 1.7	71.3 ± 0.8	76.7 ± 0.92
VL3-BA	51.3 ± 1.7	84.4 ± 0.3	41.4 ± 2.1	66.6 ± 2.5	67.9 ± 0.9	76.6 ± 0.89
VSL1	75.9 ± 1.4	82.9 ± 0.3	80.5 ± 1.7	69.0 ± 2.3	79.4 ± 0.7	88.3 ± 0.57
VSL1-S	73.3 ± 1.5	84.2 ± 0.3	80.6 ± 1.6	62.0 ± 2.5	78.7 ± 0.7	86.6 ± 0.70
VSL1-L	60.8 ± 1.6	81.8 ± 0.3	54.4 ± 2.2	70.7 ± 2.3	71.3 ± 0.8	78.9 ± 0.82

[a] TP — true positive rate, or, percentage of correctly predicted disordered residues.

[b] TN — true negative rate, or, percentage of correctly predicted ordered residues.

[c] TP$_S$ — percentage of correctly predicted residues from short disordered regions ($\leq$30 residues).

[d] TP$_L$ — percentage of correctly predicted residues from long disordered regions ($>$30 residues).

[e] ACC — overall accuracy, as (TP+TN)/2.

[f] AUC — area under ROC curve, approximated using the *trapezoid rule*.

further improved using a competition procedure developed previously.[26]

False Positives and High B-factors

Quite often residues near a disorder/order boundary have very high B-factors and exhibit a trend of sharp increase toward the disordered region. This suggests that residues within the disordered region might have even higher B-factors if their coordinates could actually be determined. Consistent with this observation, a previous study showed that high-B-factor (flexible) ordered regions share similar amino acid compositions and sequence properties with short disordered regions, and could be predicted fairly accurately from amino acid sequence using attributes similar to those developed for disorder prediction.[30] On the other hand, the independent assessor's report revealed the high correlation (coefficient = 0.92) between averaged VSL1 predictions and experimental B-factors.[43] The overall trend was that the higher the B-factors, the higher the predictions, with values approaching or even exceeding the threshold of 0.5 for declaring disorder. Therefore, it is not surprising to observe that VSL1 had significantly higher percentage of false positives on high-B-factor regions than low-B-factor regions.

Terminal Short Disordered Regions

The significantly higher accuracies on terminal short disordered regions could be attributed to a large proportion (about 60%) of short terminal disordered regions present in the training data. The use of a "spacer" character for extending the input windows might have further led the predictor to memorize "disordered region at terminus" as a rule, which is not necessarily true for most natural proteins. A better approach would likely be to build a predictor optimized for internal short disordered regions and integrate it with VSL1-S using a method similar to the VL-XT predictor.[25]

Missing Sequence Segments

The disordered regions in our evaluation were labeled as missing residues in the atom coordinate files released by the CASP organizers. However, not all of the regions labeled as disordered corresponded to regions shown experimentally to lack organized 3D structure. Included among the experimentally characterized regions of disorder were regions encoded by DNA that was simply omitted from the cloning/expression constructs. The DNA encoding these regions could have been omitted for various regions, such as for convenience, in attempts to characterize an autonomous domain of higher interest, or even to remove a region of predicted nonglobularity (e.g., predicted to be disordered, to be a transmembrane segment, to be a signal sequence, or to be low complexity). Thus, assuming such omitted regions to be disordered could be problematic. As an example, only the first 145 residues of T0234 are present in the corresponding PDB entry 1VL7:A, while as described in the REMARK 999 record, "RESIDUES 146–165 WERE OMITTED FROM THE CONSTRUCT TO ELIMINATE A REGION PREDICTED TO BE DISORDERED." Thus, the missing 20-residue C-terminal region may not be really disordered. On the other hand, only three and five residues were predicted to be disordered in this region by VL-XT and VSL1, respectively.

Another example is the 102-residue N-terminal region of T0235 on which most of the predictors performed poorly. According to the PDB entry (1VJV:A), only a fragment (residues 97–499) of T0235 is present, along with a 12-residue purification tag at its N-terminus, while the target information from the Joint Center for Structural Genomics Web site (http://www1.jcsg.org/cgi-bin/psat/targetinfo.cgi?acc=YFR010W&uid=1314960) indicates the full sequence was included in the construct. The first 18 residues of 1VJV:A, that is, the purification tag and residues 97–102 of T0235, have missing electron densities for their backbone atoms. Furthermore, searching the T0235 sequence against the NCBI Conserved Domain Database (CDD)[44] identified a *globular* ubiquitin-like domain (UBQ, CDD id: cd00196.1) at residues 6–76, with an E-value of 7e-6, and sequence identity of 26.4% and similarity of 47.2% to the domain consensus sequence. Thus, labeling all of the first 102 residues as disordered might be incorrect, and the putative ubiquitin-like domain might account for the low prediction accuracy on this region for several of the predictors.

Multimeric Proteins

Many structures in PDB exist as multimers, that is, complexes of two or more chains or subunits. Even if the chains from the same multimer share identical sequences, they do not necessarily have identical tertiary structures, and can have different regions of missing coordinates, probably due to oligomer interfaces or crystal contacts. As one possible scenario, a sequence region that is intrinsically disordered as a monomer could become ordered upon complexation. Clearly, such protein complexes raise a unique question in data labeling.

For example, only the first residues in 1WDJ:A (target T0273) and 1WDJ:C are disordered, while 1WDJ:B has 35 disordered residues at its N-terminus. On the other hand, all our models predicted most (60–100%) of the first 35 residues to be disordered. Visual inspection reveals that the folded 35-residue N-terminal regions of chains A and C are where the two chains interact, while the missing corresponding part of chain B seems to be pointing outside the molecule if judging from the direction of its residues 36–46. Many examples of protein–protein interactions that very likely involve disorder-to-order transitions upon complex formation have been found and subjected to detailed analysis with quite interesting results.[45]

CONCLUSION

We evaluated our latest predictor VSL1 of intrinsically disordered regions on the 63 CASP6 targets and compared its performance to four predictors VL-XT, VL2, VL3-BA, and VL3-E previously developed by our group. The results suggest that progress is being made, especially in predicting disordered regions of 30 residues or shorter, which could be attributed to (a) the utilization of length-dependency of the statistical properties of the disordered regions, and (b) the substantial increase in the training data.

Additional work is needed to investigate the relationships between short disordered regions and high B-factor ordered regions, oligomer interfaces, and crystal contacts. The treatment of terminal disordered regions needs to be further examined, while techniques of denoising the training data and improved data representation should also be employed. New approaches that would include long-range residue interactions and other types of information will soon become necessary to continue the progress in this area.

REFERENCES

1. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
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, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59.
4. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
5. Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269:2–12.
6. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure (Camb) 2003;11:1453–1459.
7. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 2003;31:3701–3708.
8. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
9. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 2003;31:3833–3835.
10. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53(Suppl 6):561–565.
11. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582.
12. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. FASEB J 2004;18:1169–1175.
13. Rose GD. Unfolded proteins. In: Richards FM, Eisenerg DS, Kuriyan J, editors. Advances in protein chemistry, 62. New York: Academic Press; 2002. p. xv–xxi.
14. Daughdrill GW, Pielak GJ, Uversky VN, Cortese MS, Dunker AK. Natively disordered proteins. In: Buchner J, Kiefhaber T, editors. Handbook of protein folding: Weinheim: Wiley-VCH; 2005. p 271–353.
15. Bracken C, Iakoucheva LM, Romero PR, Dunker AK. Combining prediction, computation and experiment for the characterization of protein disorder. Curr Opin Struct Biol 2004;14:570–576.
16. 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.
17. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44:1989–2000.
18. Oldfield CJ, Ulrich EL, Cheng Y, Dunker AK, Markley JL. Addressing the intrinsic disorder bottleneck in structural proteomics. Proteins 2005;59:444–453.
19. Peti W, Etezady-Esfarjani T, Herrmann T, Klock HE, Lesley SA, Wuthrich K. NMR for structural proteomics of *Thermotoga maritima*: screening and structure determination. J Struct Funct Genomics 2004;5:205–215.
20. Coeytaux K, Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005;21:1891–1900.
21. Thomson R, Esnouf R. Prediction of natively disordered regions in proteins using a bio-basis function neural network. Lecture Notes Comput Sci 2004;3177:108–116.
22. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–839.
23. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53(Suppl 6):566–572.
24. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv Protein Chem 2002;62:25–49.
25. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48.
26. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573–584.
27. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. J Bioinform Comput Biol 2005;3:35–60.
28. 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.
29. 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.
30. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. Protein Sci 2004;13:71–80.
31. 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.

32. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of intrinsic protein disorder. In review.
33. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–140.
34. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12: 1060–1072.
35. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 1993;17: 149–163.
36. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
37. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
38. 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.
39. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266:525–539.
40. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
41. Davidson R, MacKinnon J. Estimation and inference in econometrics. New York: Oxford University Press; 1993.
42. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–843.
43. Jin Y, Dunbrack RL Jr. Assessment of disorder predictions in CASP6. Proteins 2005;Suppl 7:167–175.
44. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for protein classification. Nucleic Acids Res 2005;33:D192–D196.
45. Gunasekaran K, Tsai CJ, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J Mol Biol 2004; 341:1327–1341.