

Length-Dependent Amino Acid Compositions of Secondary Structure Segments and Disordered Regions

Jack Y. Yang¹, YiZhi Zhang¹, Mary Qu Yang², Marc Cortese¹,
A. Keith Dunker¹

Keywords: protein disorder, secondary structure, amino acid composition

1 Introduction.

Many protein regions and some entire proteins lack specific three-dimensional structures, existing instead as dynamic, disordered ensembles under physiological conditions. These proteins and regions have been variously called natively unfolded [1], intrinsically unstructured [2], and natively or intrinsically disordered [3]. We have developed predictors [4] that use amino acid sequences as inputs and that give order or disorder assignments as outputs (reviewed in [5]). Prediction of disorder has been included in the two most recent meetings for the Critical Assessment of Protein Structure Prediction (CASP5 and CASP6; see <http://predictioncenter.llnl.gov/>), with the result that successful blind predictions of disorder by several groups [6] provide additional support for a relationship between amino acid sequence and intrinsic protein disorder.

Intrinsically disordered regions and disordered proteins exhibit significantly different amino acid compositions as compared to proteins that fold into three-dimensional structures [3], [5]. In two studies, intrinsically disordered regions were observed to exhibit amino acid compositions that were length-dependent, with different amino acid compositions for short and long regions of disorder [4], [7]. In the CASP6 experiment, our length-dependent predictors of disorder based on simple linear models out-performed predictors based on more complex learning models such as neural networks and support vector machines, suggesting that length-dependent compositions might be important for the prediction of disorder. More study on this problem is needed.

Here we report systematic studies of the amino acid compositions of intrinsically disordered protein regions of different length. For comparative purposes, we also determined the length-dependent compositions of regions of secondary structure with the usual grouping into helix, sheet and irregular (or coil). We anticipate that these data will be useful for improving predictors of disorder, and perhaps also for new secondary structure predictors based on length-dependent strategies.

2 Materials and Methods

Ordered and disordered segments from length 6 to 40 were extracted from a non-redundant set of sequences taken from the Protein Data Bank (<http://www.rcsb.org/pdb>) and from the Database of Protein Disorder (www.disprot.org). Disordered regions were identified as regions of missing

¹ Indiana University School of Medicine, Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biology, Indiana University Purdue University Indianapolis, Indianapolis, Indiana 46202 USA. E-mail: jayyang@iupui.edu

² Purdue University, College of Engineering, School of Electrical and Computer Engineering, Division of Computer Engineering, West Lafayette, Indiana, 47907 USA. E-mail: purduexy@purdue.edu

coordinates. DSSP was used to assign secondary structure [8]. The resulting data set contained the following numbers of fragments and residues: coil - 16,318 and 179,388; helix - 17,794 and 219,788; sheet - 10,216 and 82,070; and disordered – 5,216 and 74,724.

To determine the degree of difference between a pair of amino acid compositions p_1 and p_2 , the Kullback - Leibler (KL) distance was determined as indicated previously [7]. The KL distance was also used as a test statistic to evaluate the significance of the differences between pairs of sample distributions by means of bootstrapping (5000 iterations).

3 Results and Discussion

Sheet structure showed the largest length-dependent amino acid compositional changes. For example, I and V drop with increasing length over the range from 6 to 18 while T, S, E and K increase with length over this range. For helices, significant decreases in P, D, and E with increasing length are observed, while H, M, and G increase with length. These changes with length likely represent end effects, with certain amino acids being more likely to be at the ends of helices and sheets with others being more likely to be excluded from the ends.

KL distance comparisons indicate amino acid compositions of disorder to be most different from compositions of sheets, with a very strong decrease in the KL distance with length. On the other hand, the KL distance comparisons indicate that the amino acid compositions are most similar for regions of disorder and regions of coil. The next most similar pair is helix and sheet.

Experiments are in progress to use these findings as the basis for developing improved predictors of disorder and of secondary structure by using training and testing sets of different length classes.

References.

- [5] Dunker, A.K., Brown, C.J., and Obradovic, Z. 2002. Identification and functions of usefully disordered proteins. *Advances in Protein Chemistry* 62:25-49.
- [3] Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J., Oldfield, C.J., Campen, A. M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C.H., Kissinger, C.R., Bailey, R. W., Griswold, M.D., Chiu, W., Garner, E.C., and Obradovic, Z. 2001. Intrinsically disordered protein. *J. Mol. Graph. and Mod.* 19: 26-59.
- [8] Kabasch, W. and Sander, C. 1983. Dictionary of protein secondary structure: patterns recognition of hydrogen-bonded and geometric features. *Biopolymers* 22:2577-2637.
- [6] Melamud, E., and Moulton, J. 2003. Evaluation of disorder predictions in CASP5. *Proteins*. 53 Suppl 6:561-5.
- [7] Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., and Dunker, A. K. 2004. Protein flexibility and intrinsic disorder. *Protein Science* 13: 71-80.
- [4] Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., and Dunker, A. K., 1997. Identifying disordered regions in proteins from amino acid sequence. Proceedings of the *International Conference on Neural Networks*. 1: 90-95.
- [1] Weinreb P.H., Zhen W., Poon, A.W., Conway, K.A., Lansbury, P.T., Jr. 1996. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*. 35:13709-13715.
- [2] Wright, P.E, Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293:321-331.