Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein

Edward A. Weathers[a,b], Michael E. Paulaitis[a,c], Thomas B. Woolf[b,d], Jan H. Hoh[a,b,*]

[a]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 221 Maryland Hall, 3400 North Charles Street, Baltimore, MD 21218, USA
[b]Department of Physiology, Johns Hopkins School of Medicine, 217 Hunterian Building, 725 North Wolfe Street, Baltimore, MD 21205, USA
[c]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21205, USA
[d]Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

Received 4 June 2004; revised 14 September 2004; accepted 14 September 2004

Available online 25 September 2004

Edited by Takashi Gojobori

Abstract Intrinsically disordered proteins are an important class of proteins with unique functions and properties. Here, we have applied a support vector machine (SVM) trained on naturally occurring disordered and ordered proteins to examine the contribution of various parameters (vectors) to recognizing proteins that contain disordered regions. We find that a SVM that incorporates only amino acid composition has a recognition accuracy of 87 ± 2%. This result suggests that composition alone is sufficient to accurately recognize disorder. Interestingly, SVMs using reduced sets of amino acids based on chemical similarity preserve high recognition accuracy. A set as small as four retains an accuracy of 84 ± 2%; this suggests that general physicochemical properties rather than specific amino acids are important factors contributing to protein disorder.

© 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

Keywords: Unstructured protein; Support vector machine; Amino acid composition; Protein classification; Sequence complexity

1. Introduction

It is becoming increasingly clear that proteins or segments of proteins that lack a stable and well-defined three-dimensional structure, often referred to as intrinsically disordered proteins, have a range of important properties and functions that depend on or derive from being disordered [1–4]. For example, Wright and colleagues have proposed that intrinsically disordered protein segments confer conformational flexibility to some proteins allowing a functionally important promiscuity in binding [5]. Other functions such as regulators of nuclear port transport and entropic clocks for ion channel gating have also been proposed [5,6]. Our interest in this problem derives from a proposal that certain cytoskeletal proteins have intrinsically disordered protein segments [7]. In particular, the side-arms of neurofilament proteins NF-M and NF-H and the projection domain of MAP2 are highly unstructured and as a consequence exert long range repulsive forces that are largely entropic in origin; these forces are critical to organizing the neuronal cytoskeleton [8–10].

The emerging importance of disordered proteins has led to the development of tools and approaches for recognizing and predicting the propensity for any given protein sequence to be disordered. Developing and testing these tools requires defining libraries of disordered protein sequences; however, there are no rigorous experimental criteria for defining disorder. Criteria used for identifying collections of disordered regions include considering data from X-ray crystallography, NMR, circular dichroism, and protease sensitivity [11–14]. One notable feature of disordered proteins identified in such collections is a strong bias towards charged and polar amino acids and against hydrophobic amino acids [12]. While there is no unambiguous test of these groupings, it is reasonable to assume that they are at least strongly biased in their relative composition of disordered versus ordered proteins. With that caveat in mind, Dunker and co-workers developed PONDR, a neural net-based predictor [15]. There are now a variety of implementations of PONDR with prediction accuracies as high as 87% [16]. Linding and coworkers also developed a neural net predictor for disorder, DisEMBL, which uses three data sets based on different definitions of disorder [13]. These sets are based on an analysis of proteins with known three-dimensional structure. Consistent with the previous work, the propensities of these sets show a bias for charged and polar amino acids and against hydrophobic amino acids, although there are significant differences in the relative compositions.

A significant limitation of these neural net-based approaches is that it is difficult to interrogate the relative contribution of individual parameters to recognizing or predicting disorder. Here, we have trained a support vector machine (SVM) to recognize intrinsically disordered proteins. SVMs are learning machines based on the development of statistical learning theory by Vapnik and colleagues [17]. An important feature of SVMs is that the results of the learning process can be quantified; thus, the relative influence of different parameters on the ability of the SVM to recognize disordered proteins can be measured. SVMs operate in two stages: data sets from two different classes are first mapped into a higher dimensional space based on vectors that represent some particular parameter, then the hyperplane that optimally separates the two classes is calculated. SVMs are designed to provide a globally optimized solution that ensures the highest level of recognition

[*] Corresponding author. Fax: +1-410-614-3797.
E-mail address: jhoh@jhmi.edu (J.H. Hoh).

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.09.036

accuracy. SVMs have been successfully applied to many pattern classification and recognition problems; applications to biology include predictions of secondary structure, subcellular location, and solvent accessibility [18–20]. Jones and colleagues [21] have recently shown that SVMs are effective tools for predicting disordered proteins. Here, we use an SVM based approach to gain further insight into the physicochemical principles important for recognition of disordered proteins.

2. Materials and methods

2.1. Protein data

The training set was that compiled by Dunker and colleagues [15]. This set contains 718 segments classified as disordered and 1190 sequence classified as structured.

2.2. Support vector machine

We used the mySVM implementation of support vector machine theory by Rüping (http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/). The initial stage of mapping data sets into higher dimensional spaces was accomplished using a kernel function, $K(s_i, x)$, where s_i is a support vector and x is the input sequence. For our analysis, we chose a dot kernel function where $K(s_i, x) = s_i \cdot x$. This kernel function provides high accuracy while avoiding the long training and testing times associated with higher order kernel functions. The results of the mapping process are represented as a set of vectors, $x_i, i = 1, \ldots, N$, and a label vector y_i, which equals 1 for one class and -1 for the alternate class. The optimally separating hyperplane (OSH) is represented by $w^T x_i + b = 0$, where w is the set of vector weights and b is the bias. The vector weight w represents the relative importance of each contributing factor to classification. For ideal data sets, OSH is found by minimizing $1/2w^T w$ subject to the constraint $y_i(w^T x_i + b) \geqslant 1$. For non-ideal data sets, the individual vectors may not be linearly separable. Thus, parameters are introduced to allow for non-linear separation while limiting training error. For this case, the OSH is found by minimizing $1/2w^T w + C \sum \varepsilon_i$ subject to the constraint that $y_i(w^T x_i + b) \geqslant 1 - \varepsilon_i$, where $\varepsilon_i \geqslant 0$. ε_i are slack variables that represent the deviation from ideal separation; these values are minimized in the training process. C is a regularization parameter that balances the trade-off between complexity and error. For our analysis, a range of values for C were tested (data not shown) and C was set at 0.07. Software and data sets used in this analysis are available upon request.

2.3. Measurement of prediction accuracy

Prediction accuracy was determined using 5-fold cross validation (Fig. 1). The ordered and disordered datasets were combined, and 80% of this data set was randomly chosen and used to train the SVM. The prediction accuracy was then measured by testing the SVM on the remaining 20% of the original dataset. The overall prediction accuracy is the average of 10 rounds of testing; 50% reflects random classification.

3. Results and discussion

Each protein in the data set of ordered and disordered proteins was translated into a vector representation. Our initial vector set was based on sequence composition information for each amino acid; proteins were represented with one vector for each amino acid (20-AA SVM). The SVM was trained on a randomly chosen selection of sequences comprising 80% of the total set. The prediction accuracy was calculated by testing the ability of the SVM to correctly categorize proteins in the remaining 20% of the data set. Using this approach, the 20-AA SVM has an accuracy of 87 ± 2%, demonstrating that amino acid composition alone is sufficient to accurately recognize disordered proteins. The vector weights for the 20 amino acids indicate a strong bias against hydrophobic groups and a weaker bias toward charged or polar groups (Fig. 2).

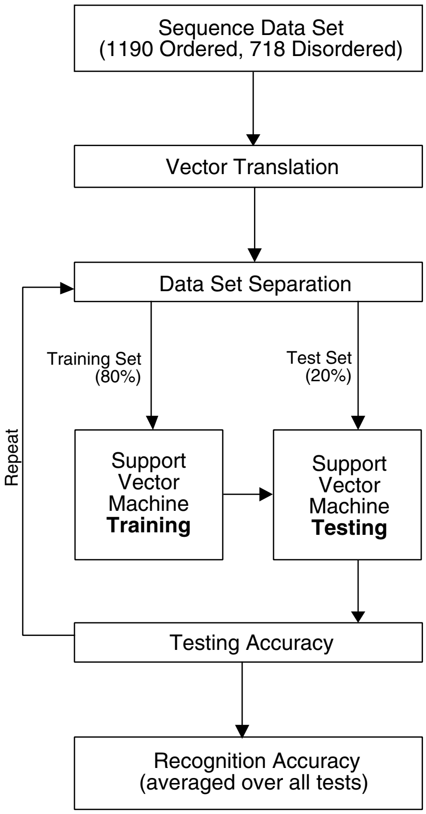

Fig. 1. Schematic of development and testing of the SVM for recognizing intrinsically disordered proteins.

A number of additional parameters that have been associated with disordered proteins were also examined, including Wootton sequence complexity [22], phosphorylation content [23], and net charge. The Wootton complexity is related to the complexity of the numerical state of a sequence and effectively is a measure of the number of distinct ways in which a given sequence can be rearranged. The phosphorylation content is based on the frequency of consensus motifs cAMP-dependent protein kinase, protein kinase C, casein kinase II, and tyrosine kinase obtained from Prosite (http://us.expasy.org/prosite/). The charge vector reflects net charge, where K and R are positively charged and D and E are negatively charged. Used

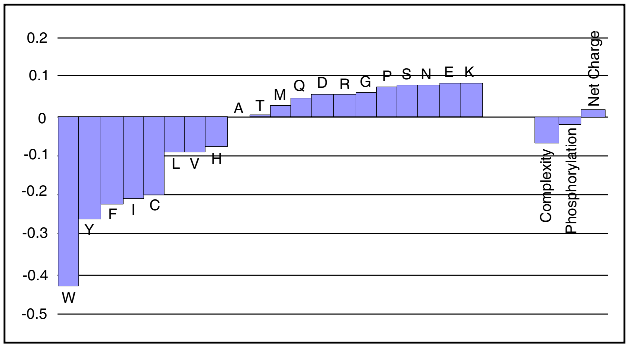

Fig. 2. SVM vector weights for the 20 amino acid SVM predictor and three additional parameters. Positive values indicate residues that are associated with disorder, while residues with negative values are associated ordered regions.

together these three vectors have a recognition accuracy of 71%, poorly compared to the 20-AA SVM. Adding the three vectors to the 20 individual amino acid vectors resulted in no change in the accuracy, and the weights of the new vectors were small, suggesting that they add little new information over sequence composition (Fig. 2). The role of higher order parameters was further investigated by using vector sets based on increased block size. Vector sets were developed for all possible amino acid dimers (400 vectors) and trimers (8000 vectors). Recognition accuracy for the dimers was identical to the single amino acids, while using the trimers increased accuracy slightly to 90 ± 1%.

To investigate how a particular class or property of amino acids affects recognition accuracy and to determine the minimal amount of information needed for recognition, a number of reduced amino acid sets were studied. Reduced sets developed by Andorf and colleagues based on the BLOSUM50 substitution matrix were used to decrease the number of vectors needed to represent protein sequences [24,25]. Sets of 15, 10 and 8 vectors each had 85 ± 2% recognition, and a reduced set of 4 retained 84 ± 1% recognition accuracy (Table 1). Additional reduced sets of amino acids were created based on chemical properties. Sets based on charge had relatively poor recognition (62 ± 3%), while sets based on hydrophobicity performed well (82 ± 1%). The vector weights for these reduced sets also showed a similar strong bias against hydrophobic amino acids and weaker bias for charged or polar groups (Fig. 3). Random groupings of amino acids into four categories produced recognition accuracies near random.

A central finding from our SVM analysis is that a small number of vectors based on general chemical properties of amino acids is sufficient to recognize disordered protein. Using a full 20-amino acid representation of protein sequence can achieve a recognition accuracy of 87%, while a reduced set as small as 4 preserves an 84% recognition accuracy. In the 4 vector set, two vectors with amino acids of a more hydrophilic character show a positive relationship with disorder (disorder-associated), while the two vectors representing more hydrophobic amino acids show a negative relationship (order-associated) [11]. For all the amino sets, the negative vectors are stronger than the positive vectors, suggesting that a high ratio of hydrophilic to hydrophobic amino acids is characteristic of disordered proteins. There are a number of ways to interpret these results. It has been suggested that functionally important properties of disordered proteins may be less sensitive to specific amino acid content than well-folded proteins [26]. This line of thinking is based on analytical treatments of polymers of the type developed by Flory [27] and de Gennes [28], where the polymers are highly unstructured. In these models, relatively simple bead-spring representations of polymers, often with only attractive or repulsive interactions, are remarkably powerful in capturing measurable properties. The general conclusion is that for polymers (proteins) in this regime, atomic details of the monomers are much less important than general characters such as hydrophilicity and hydrophobicity. This is consistent with the findings here, which imply that disorder is related to general chemical properties rather than interactions between specific amino acids. We also note that it is well established that the hydrophobic amino acids play a central role in stabilizing folded proteins [29]. This fact has been exploited to recognize native folds [30] and predict protein globularity [31,32]. In one such approach globularity prediction is based on the ratio of surface accessible to buried amino acids; given the close relationship between surface accessibility and hydrophobicity/hydrophilicity, this means that the general character of amino acid composition provides information about how well a protein will fold [31]. The corollary to this finding would be, as we find here, that a significant under-representation of hydrophobic amino acids would tend to produce less globular and less well-folded proteins.

In general, higher-order correlations seem to play little role in the recognition of disorder. A slight improvement in prediction accuracy was observed for amino-acid blocks of three. However, this difference is at the border of statistical significance; also, when using block sizes larger than one, a potential drawback is overestimation of the recognition accuracy. This can occur when the dataset used in SVM training contains homologous proteins; when large block sizes are used the SVM can overpredict as a result of this homology. Additionally, the lower frequency of appearance of some dimers and trimers in the dataset creates difficulties for statistically accurate predictions. Another issue related to higher-order correlations is the effect of different sequence arrangements on disorder prediction. A protein with a hydrophobic region followed by a hydrophilic region could produce the same SVM score as a protein with alternating hydrophobic and hydrophilic residues, even though these arrangements would not be expected to behave in the same way. However, naturally occurring proteins tend not to be arranged in blocks of amino acids and thus this is not a problem with distinguishing between such proteins.

Table 1
Summary of the SVM recognition accuracy for all vector sets

Classification property	Vector size	Prediction accuracy
20-AA SVM	20	87 ± 2%
Others (charge, phosphorylation and complexity)	3	71 ± 2%
20-AA SVM + others	23	87 ± 2%
Amino acid dimers	400	87 ± 2%
Amino acid trimers	8000	90 ± 1%
Reduced 15 (sub. matrix)	15 (FY,ILMV,KR)	85 ± 2%
Reduced 10 (sub. matrix)	10 (FWY,ILMV,ST,EDNQ,KR)	85 ± 1%
Reduced 8 (sub. matrix)	8 (FWY,CILMV,AG,ST,EDNQ,KR)	85 ± 2%
Reduced 4 (sub. matrix)	4 (FWY,CILMV,AGPST,DEHKNQR)	84 ± 1%
Hydrophobicity	4 (FILVWY,ACGMP,DEHNR,KQST)	82 ± 1%
Charge	3 (KR,DE,ACFGHILMNPQSTVWY)	62 ± 3%

Amino acids in parentheses denote the grouping of residues in the reduced alphabets.

Previous work on disordered proteins has demonstrated a very clear propensity for such proteins to be over-represented in polar and charged amino acids [11–14]. However, the propensity itself, based on a composition profile, does not allow one to evaluate the importance of a given amino acid (or other parameter) to recognizing or predicting disorder. One significant contribution that the SVM approach can make in this context is that it allows quantitative weights to be assigned to individual parameters; these weights are objectively tied to the recognition performance of the SVM. Vector weights for our 20-AA SVM show significant deviations from the overall amino acid composition profiles of the input data (Fig. 4) [11]. The composition profiles indicate the same hydrophilic/hydrophobic separation between order-associated and disorder-associated amino acids. However, our weight vectors show deviations from these propensities, most significantly for tryptophan. The composition profile also indicates that asparagine and aspartic acid are associated with order, while the weight vectors suggest that both are significantly associated with disorder. This suggests that while asparagines/aspartic acid content is relatively low in the overall disordered dataset, high asparagine/aspartic acid content in an individual protein sequence is an indicator of disorder. This conclusion is in agreement with the propensity scales developed by Linding and colleagues: two of the three scales indicate a high propensity for asparagine and aspartic acid to be disordered [13]. These propensity scales again show similar trends for the vector weights although with some minor differences. While the vector weights indicate that charged residues are associated with disorder, the propensity values for some charged amino acids show a bias towards order for one propensity scale. This difference may be a result of the particular scale's derivation from known loop regions, which include both ordered and disordered segments. The SVM vector weights agree best with the values for the "hot loop" propensity scales, which are taken from loop regions with high B factors.

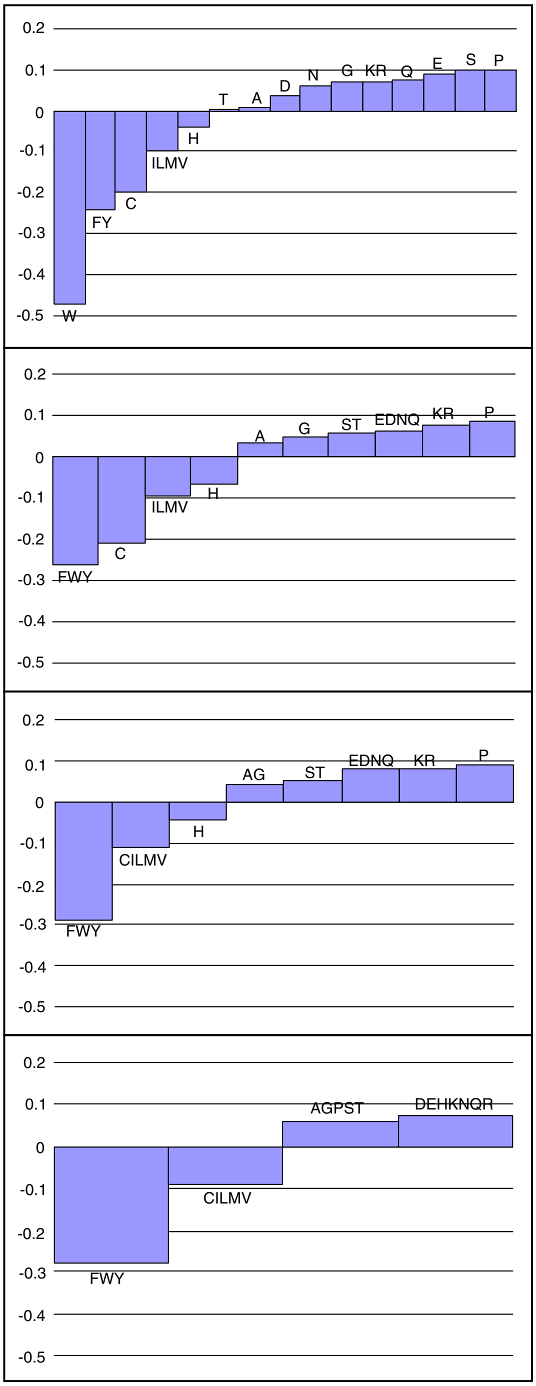

Fig. 3. SVM vector weights for reduced amino acid sets based on the BLOSUM50 substitution matrix. Set of (a) 15, (b) 10, (c) 8 and (d) 4.

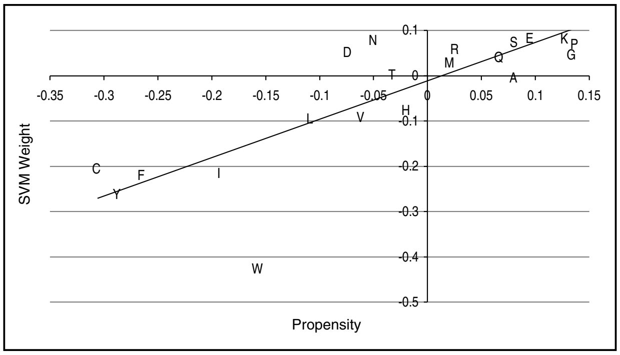

Fig. 4. Comparison of amino acid propensity versus SVM vector weights. Propensities are calculated by taking the log difference of each amino acid's percent composition in the ordered and disordered datasets. Positive propensities denote amino acids overrepresented in disordered proteins.

The SVM used in our analysis is a binary classifier that assumes that proteins will fall into one of two predefined classes: they have a disordered segment of >40 amino acids or they do not. However, naturally occurring proteins can contain both ordered and disordered segments. This suggests that an analysis of proteins in nature should use local (along the chain), rather than overall, amino acid composition as the metric for identifying regions of disorder. Disordered segments can also vary in extent and type; it is likely that there are qualitatively different functions for disordered proteins and it is likely that the nature of the disorder in these cases will be different. Identifying the different classes of disordered proteins and their associated functions will become increasingly important; the SVM based approach used here may prove useful in that endeavor.

Acknowledgements: This work was supported in part by the Burroughs Wellcome Fund (EW).

References

[1] Tompa, P. (2002) Trends Biochem. Sci. 27, 527–533.

[2] Wright, P.E. and Dyson, H.J. (1999) J. Mol. Biol. 293, 321–331.

[3] Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. (2002) Biochemistry 41, 6573–6582.

[4] Uversky, V.N. (2002) Protein Sci. 11, 739–756.

[5] Dyson, H.J. and Wright, P.E. (2002) Curr. Opin. Struc. Biol. 12, 54–60.

[6] Rout, M.P., Aitchison, J.D., Magnasco, M.O. and Chait, B.T. (2003) Trends Cell Biol. 13, 622–628.

[7] Hoh, J.H. (1998) Proteins 32, 223–228.

[8] Brown, H.G. and Hoh, J.H. (1997) Biochemistry 36, 15035–15040.

[9] Kumar, S., Yin, X., Trapp, B.D., Hoh, J.H. and Paulaitis, M.E. (2002) Biophys. J. 82, 2360–2372.

[10] Mukhopadhyay, R. and Hoh, J.H. (2001) FEBS Lett. 505, 374–378.

[11] Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.R., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C.H., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, M., Garner, E.C. and Obradovic, Z. (2001) J. Mol. Graph. Model. 19, 26–59.

[12] Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000) Proteins 41, 415–427.

[13] Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Structure (Camb.) 11, 1453–1459.

[14] Liu, J., Tan, H. and Rost, B. (2002) J. Mol. Biol. 322, 53–64.

[15] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S. and Dunker, A.K. (1998) Pac. Symp. Biocomput., 437–448.

[16] Vucetic, S., Brown, C.J., Dunker, A.K. and Obradovic, Z. (2003) Proteins 52, 573–584.

[17] Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, Berlin.

[18] Hua, S. and Sun, Z. (2001) J. Mol. Biol. 308, 397–407.

[19] Cai, Y., Liu, X., Xu, X. and Chou, K. (2002) J. Cell. Biochem. 84, 343–348.

[20] Yuan, Z., Burrage, K. and Mattick, J.S. (2002) Proteins 48, 566–570.

[21] Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) J. Mol. Biol. 337, 635–645.

[22] Wootton, J.C. and Federhen, S. (1993) Computers Chem. 17 (2), 149–163.

[23] Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. and Dunker, A.K. (2004) Nucleic Acids Res. 32, 1037–1049.

[24] Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad. Sci. USA 89, 10915–10919.

[25] Andorf, C.M., Dobbs, D.L. and Honavar, V.G. (2003) Inform. Sciences, in press.

[26] Bright, J.N., Woolf, T.B. and Hoh, J.H. (2001) Prog. Biophys. Mol. Biol. 76, 131–173.

[27] Flory, P.J. (1953) Principles of Polymer Chemistry. Cornell University Press, Ithaca.

[28] de Gennes, P.G. (1979) Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca.

[29] Dill, K.A. (1990) Biochemistry 29, 7133–7155.

[30] Huang, E.S., Subbiah, S. and Levitt, M. (1995) J. Mol. Biol. 252 (5), 709–720.

[31] Rost, B. and Liu, J. (2003) Nucleic Acids Res. 31, 3300–3304.

[32] Linding, R., Russell, R.B., Neduva, A. and Gibson, T.J. (2003) Nucleic Acids Res. 31, 3701–3708.