Unsupervised Integration of Multiple Protein Disorder Predictors

Ping Zhang and Zoran Obradovic
Center for Information Science and Technology, Temple University, Philadelphia, PA 19122
ping@temple.edu, zoran@ist.temple.edu

Abstract—**Studies of intrinsically disordered proteins that lack a stable tertiary structure but still have important biological functions critically rely on computational methods that predict this property based on sequence information. Although a number of fairly successful models for prediction of protein disorder were developed over the last decade, the quality of their predictions is limited by available cases of confirmed disorders. To more reliably estimate protein disorder from protein sequences, an iterative algorithm is proposed that integrates predictions of multiple disorder models without relying on any protein sequences with confirmed disorder annotation. The iterative method alternately provides the maximum a posterior (MAP) estimation of disorder prediction and the maximum-likelihood (ML) estimation of quality of multiple disorder predictors. Experiments on data used at the Critical Assessment of Techniques for Protein Structure Prediction (CASP7 and CASP8) have shown the effectiveness of the proposed algorithm.**

Keywords—protein structure; disorder prediction; meta predictor; unsupervised learning; iterative MAP-ML algorithm

I. Introduction

Identification of regions in proteins that do not have a unique structure, called intrinsic disorders, is addressed computationally by a number of groups that aim to predict this property from sequence information. Contrary to the lock and key paradigm, disordered regions were recently found to be involved in many important functions [1] and in various diseases [2].

Computational characterization of disorder in proteins is appealing due to the difficulties and high cost involved in experimental characterization of disorders. The first predictor of protein disorder was developed by our group in the year 1997 [3]. Based on the importance of predicting this property, in the year 2002, protein disorder prediction was introduced as a category of the CASP contests [4], which promoted the development of new methods for prediction of protein disorder. Consequently, the number of prediction methods available through the Internet has increased rapidly. More than 50 predictors of intrinsic protein disorder have been described in a recent review by He et al [5], enabling researchers to use a meta approach to predict protein disorder by integrating the prediction results of several methods. Recently, four such meta predictors, i.e. metaPrDOS [6], MD [7], PONDR-FIT [8], and MFDp [9], have been developed for the purpose of improving disorder prediction accuracy. They showed significantly improved performance in performed experiments as compared to using individual component predictors.

A limitation of these supervised learning based meta predictors is that they are prone to over-optimization in their integration processes since they are developed relying on disorder/order labeled training datasets that contain a very small number of proteins that have not already been used for development of the component predictors (e.g. sets as small as the Disordered Proteins Database (DisProt) [10] or as specialized as missing coordinates from the Protein Data Bank (PDB) [11]). Therefore, the prediction results of previous meta predictors may not be so good for proteins that have sequence patterns much different from cases used for integration. For example, although it achieved higher prediction accuracy than all predictors participating in CASP7 as stated in its paper [6], metaPrDOS failed to be one of the top predictors in CASP8 [12]. Moreover, one of metaPrDOS' component predictors, i.e. DISOPRED [13], was more accurate than metaPrDOS in CASP8 [12].

To address potential over-optimization problems of meta predictor development by learning from small labeled data, here we introduce a new disorder meta prediction method. By following the idea from Raykar et al. [14], we derived an iterative MAP and ML estimation (MAP-ML) based algorithm for the construction of a meta predictor in a completely unsupervised process using protein sequences without confirmed disorder/order annotations. The new algorithm is presented in Section III after settings and notations are introduced in Section II. Performance evaluation of the new meta method is presented in Section IV by using CASP7 and CASP8 prediction targets as the test sets, which enabled us to compare the prediction results with other methods used in the CASP contests.

II. Problem Statement

Let us define the dataset as $D = \{\boldsymbol{x}_i, y_i^1, ..., y_i^M\}_{i=1}^N$. Here, $\boldsymbol{x}_i$ is an amino acid composition feature vector which is derived from the subsequence covered by a moving window centered at the i-th amino acid within the current protein. $y_i^j \in \{1, 0\}$ (1 represents a disordered state while 0 represents an ordered state) is the prediction label assigned to the instance $\boldsymbol{x}_i$ by the j-th predictor. M is the number of predictors. N is the number of amino acids in the protein.

The first task of our interest is to estimate the sensitivity (i.e., true positive rate) $\boldsymbol{\alpha} = [\alpha^1, ..., \alpha^M]$ and the specificity (i.e., true negative rate) $\boldsymbol{\beta} = [\beta^1, ..., \beta^M]$ of the M predictors. The second task is to get an estimation of the unknown true labels $y_1, ..., y_N$.

III. Method

A. The Proposed MAP-ML Algorithm

To fulfill the two tasks defined in section II, we propose an iterative algorithm that we will call MAP-ML. Given dataset D, we use majority voting to initialize the probabilistic labels μ_i (i.e., the probability when the hidden true label is 1). Then, the algorithm alternately carries out the ML estimation and the MAP estimation described in details in subsections B and C. Given the current estimates of probabilistic labels, the ML estimation measures predictors' performance (i.e., their sensitivity $\boldsymbol{\alpha}$ and specificity $\boldsymbol{\beta}$) and learns a classifier with parameter $\boldsymbol{w}$. Given the estimated sensitivity $\boldsymbol{\alpha}$, specificity $\boldsymbol{\beta}$, and the prior probability which is provided by the learned classifier, the MAP estimation gets the updated probabilistic labels μ_i based on the Bayesian rule. After the two estimations converge, we get the algorithm outputs which include both the probabilistic labels μ_i and the model parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

The proposed iterative MAP-ML algorithm is summarized in Algorithm 1, and the estimations are described in the following subsections B and C.

B. ML Estimation of the Model Parameters

Given the dataset D and the current estimates of μ_i, the algorithm estimates the model parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ by maximizing the conditional likelihood. According to the definitions of sensitivity and specificity, we get

$$\begin{aligned} \alpha^j &= \sum_{i=1}^{N} \mu_i y_i^j \Big/ \sum_{i=1}^{N} \mu_i \\ \beta^j &= \sum_{i=1}^{N} (1-\mu_i)(1-y_i^j) \Big/ \sum_{i=1}^{N} (1-\mu_i) \end{aligned} \tag{1}$$

Given probabilistic labels μ_i, we can learn any classifier using ML estimation. However, in this section for convenience, we will explain it with a logistic regression classifier. By using that classifier, the probability for the positive class is modeled as a sigmoid acting on the linear discriminating function, that is,

$$\Pr[y = 1 \mid \boldsymbol{x}, \boldsymbol{w}] = \sigma(\boldsymbol{w}^T \boldsymbol{x}) \tag{2}$$

where the logistic sigmoid function is defined as $\sigma(z) = 1/(1+e^{-z})$. To estimate the classifier's parameter $\boldsymbol{w}$, we use a gradient descent method, that is, the Newton-Raphson method [15]

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{H}^{-1} \boldsymbol{g} \tag{3}$$

where **g** is the gradient vector, **H** is the Hessian matrix, and η is the step length. The gradient vector is given by

$$\boldsymbol{g}(\boldsymbol{w}) = \sum_{i=1}^{N} [\mu_i - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)] \boldsymbol{x}_i \ ,$$

and the Hessian matrix is given by

$$\boldsymbol{H}(\boldsymbol{w}) = -\sum_{i=1}^{N} [\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)][1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)] \boldsymbol{x}_i \boldsymbol{x}_i^T \ .$$

Algorithm 1. Iterative MAP-ML Algorithm

```
Input: Protein sequences with prediction labels from M predictors.
  Step 1: Convert the protein sequences into amino acid composition feature vectors.
  Step 2: Use majority voting to initialize $\mu_i = \sum_{j=1}^{M} y_i^j \Big/ M$ .
  Step 3: Iterative optimization.
    3.1: ML estimation – Estimate the model parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ based on current probabilistic labels $\mu_i$ using (1) and (3) in subsection B.
    3.2: MAP estimation – Given the model parameters $\theta$, update $\mu_i$ using (8) in subsection C.
  Step 4: If $\theta$ and $\mu_i$ do not change between two successive iterations or the maximum number of iterations is reached, go to the Step 5; otherwise, go back to the Step 3.
  Step 5: Estimate the hidden true label $y_i$ by applying a threshold on $\mu_i$, that is, $y_i$ =1 if $\mu_i > \gamma$ and $y_i$ =0 otherwise. Here use $\gamma$ =0.5 as the threshold.
Output:
1. The estimated sensitivity and specificity of each predictor;
2. The weight parameter of a classifier;
3. The probabilistic labels $\mu_i$ ;
4. The estimation of the hidden true labels $y_i$ .
```

C. MAP Estimation of the Unknown True Labels

Given the dataset D and the model parameters $\theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$, we define probabilistic labels $\mu_i = \Pr[y_i = 1 \mid y_i^1, ..., y_i^M, \boldsymbol{x}_i, \theta]$. Using the Bayesian rule we have

$$\mu_i = \frac{\Pr[y_i^1, ..., y_i^M \mid y_i = 1, \theta] \cdot \Pr[y_i = 1 \mid \boldsymbol{x}_i, \theta]}{\Pr[y_i^1, ..., y_i^M \mid \theta]} \tag{4}$$

which is a MAP estimation problem.

Conditioning on the true label $y_i \in \{1, 0\}$, the denominator of formula (4) is decomposed as

$$\begin{aligned} &\Pr[y_i^1, ..., y_i^M \mid \theta] = \\ &\Pr[y_i^1, ..., y_i^M \mid y_i = 1, \boldsymbol{\alpha}] \Pr[y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{w}] \\ &+ \Pr[y_i^1, ..., y_i^M \mid y_i = 0, \boldsymbol{\beta}] \Pr[y_i = 0 \mid \boldsymbol{x}_i, \boldsymbol{w}] \end{aligned} \tag{5}$$

Given the true label y_i, we assume that $y_i^1, ..., y_i^M$ are independent, that is, the predictors label the instances independently. Hence,

$$\Pr[y_i^1,\dots,y_i^M \mid y_i=1,\boldsymbol{\alpha}] = \prod_{j=1}^{M} \Pr[y_i^j \mid y_i = 1, \alpha^j] \tag{6}$$

$$= \prod_{j=1}^{M} [\alpha^j]^{y_i^j} \, [1-\alpha^j]^{1-y_i^j}$$

Similarly, we have

$$\Pr[y_i^1,\dots,y_i^M \mid y_i=0,\boldsymbol{\beta}] = \prod_{j=1}^{M} [\beta^j]^{1-y_i^j} \, [1-\beta^j]^{y_i^j} \tag{7}$$

From (2), (4), (5), (6), and (7), the posterior probability μ_i which is a soft probabilistic estimate of the hidden true label is computed as

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1-p_i)} \tag{8}$$

where

$$p_i = \Pr[y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{w}] = \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)$$

$$a_i = \prod_{j=1}^{M} [\alpha^j]^{y_i^j} \, [1-\alpha^j]^{1-y_i^j} \quad .$$

$$b_i = \prod_{j=1}^{M} [\beta^j]^{1-y_i^j} \, [1-\beta^j]^{y_i^j}$$

D. *Analysis of the MAP Estimation*

To explain how the MAP estimation model works, we apply logit function to the posterior probability μ_i . From (8), the logit of μ_i is written as

$$\mathrm{logit}(\mu_i) = \ln\frac{\mu_i}{1-\mu_i} = \ln\frac{\Pr[y_i = 1 \mid y_i^1,\dots,y_i^M, \boldsymbol{x}_i, \theta]}{\Pr[y_i = 0 \mid y_i^1,\dots,y_i^M, \boldsymbol{x}_i, \theta]} \tag{9}$$

$$= \boldsymbol{w}^T \boldsymbol{x}_i + \sum_{j=1}^{M} y_i^j [\mathrm{logit}(\alpha^j) + \mathrm{logit}(\beta^j)] + c$$

where $c = \sum_{j=1}^{M} \ln[(1-\alpha^j)/\beta^j]$ is a constant. The first term of (9) $\boldsymbol{w}^T \boldsymbol{x}_i$ is a linear combination (provided by the learned classifier) of current amino acid's composition features. The second term of (9) is a weighted linear combination of the prediction labels from all the predictors. The weight of each predictor is the sum of the logit of the estimated sensitivity and specificity. From (9), we can infer that the estimates of the hidden true labels (in logit form) depend both on protein sequence information and on the prediction labels from all the predictors.

IV. Results and Evaluation

The performance of predictors was evaluated by three measures: the area under the ROC curve (AUC), the average of sensitivity and specificity (ACC), and a weighted score (i.e., S_w) that considers the rates of ordered and disordered residues in the datasets [16].

To assess prediction performance, we used CASP8 data [12] consisting of 121 experimentally characterized protein sequences with 24548 ordered and 2941 disordered residues. To reduce noise due to experimental uncertainty, in the evaluation process we didn't consider disorder segments shorter than four residues. We have also obtained prediction labels with disorder probabilities of all predictors which participated in CASP8 from the contest's official website [4]. We selected 13 predictors developed by different groups assuming that their errors are independent.

In the experiment, as the input of our iterative MAP-ML algorithm we used the sequences of 121 protein targets and the prediction labels from the 13 component predictors. After the algorithm had converged, we used the estimation of the hidden true labels y_i produced by MAP-ML as the binary disorder/order predictions and the probabilistic labels μ_i from MAP-ML outputs as the disorder probability.

Estimated sensitivity α and specificity β of 13 component predictors using our MAP-ML meta predictor without relying on true disorder/order labels are shown in Figure 1. The obtained estimates are sorted according to the average of their estimated sensitivity and specificity and were quite consistent with evaluations reported by CASP8 committee [12] who for their evaluations used labeled data of confirmed disorder/order residues.

A comparison of 13 predictors and our MAP-ML meta predictor on CASP8 labeled data with confirmed disorder/order is shown in Figure 2. On this comparison our iterative MAP-ML algorithm had ACC score of 0.843, S_w score of 0.686, and AUC score of 0.922. These scores were superior to the 13 component predictors in the CASP8 contest. In addition, Figures 1 and 2 could be used to assess similarity of accuracies and rankings of 13 predictors obtained by MAP-ML algorithm without any labeled data versus their evaluation on true labels by CASP8 committee.

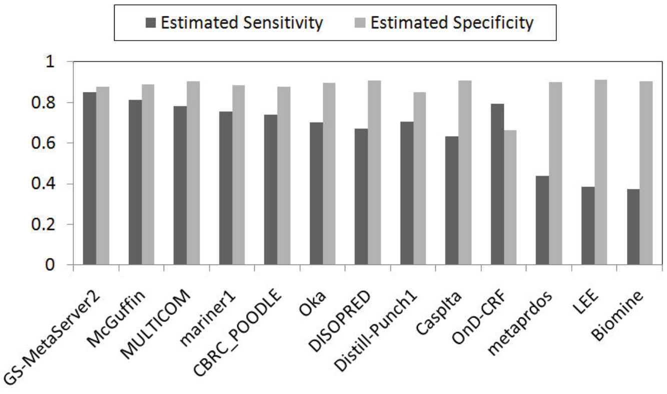

Figure 1. CASP8 accuracy estimates without using labeled data. Estimated sensitivity and specificity of 13 disorder predictors is obtained by MAP-ML algorithm at CASP8 protein sequences without using CASP8 experimentally determined disorder/order labels. The predictors are sorted in descending order of the average of the estimated sensitivity and specificity.

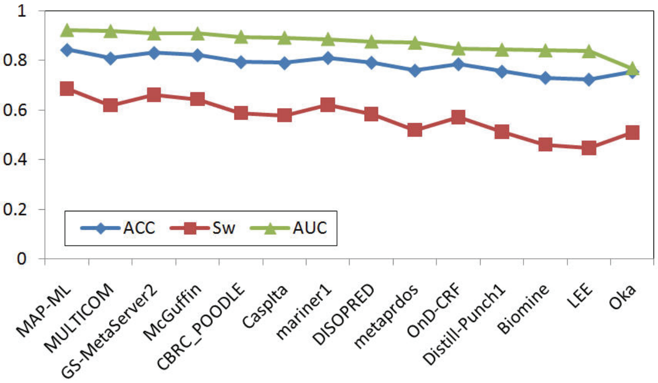

Figure 2. CASP8 comparison on labeled data. Evaluation scores are shown for MAP-ML algorithm and the 13 component predictors at disorder/order labeled CASP 8 protein sequences and the corresponding experimentally determined disorder/order labels. ACC, S_w, and AUC scores are sorted in descending order of the AUC score.

Using the same measures and procedures, we assessed accuracy of 11 CASP7 disorder predictors on CASP7 data [17] without using the corresponding experimentally determined disorder/order labels. Similar to CASP8, for most of predictors ranks obtained by MAP-ML algorithm were quite consistent with their true accuracy on CASP7 data. The scores (ACC=0.798, S_w=0.595, AUC=0.881) of our MAP-ML meta predictor were better than the corresponding scores of 11 component predictors in the CASP7 contest (the details of the experiments are omitted for lack of space).

V. Conclusion

In this study, we proposed an iterative MAP-ML algorithm to predict protein disorder. The algorithm alternately provides the MAP estimation of disorder prediction and the ML estimation of the quality of multiple component disorder predictors. We evaluated the performance of the MAP-ML algorithm versus the performance of other predictors using CASP7 and CASP8 datasets. The results showed that our meta predictor not only outperformed other predictors but also appropriately ranked other predictors without knowing the true labels.

The proposed algorithm assumed that the accuracy of each predictor did not depend on the given protein sequences and that the predictors make their errors independently. Therefore, in our experiments we used the component predictors developed by groups at different institutions. We emphasize that in practice the independence assumptions might not be always true which is the limitation of the proposed algorithm. To relax the independence assumptions and to make even more accurate disorder predictions by the probabilistic meta model, our research in progress includes additional parameters such as disorder flavor and difficulty of a prediction task.

Acknowledgment

This project is funded in part under a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

[1] H. Xie, S. Vucetic, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, V.N. Uversky, and Z. Obradovic, "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions," *Journal of Proteome Research*, 2007, 6(5), pp. 1882-1898.

[2] U. Midic, C.J. Oldfield, A.K. Dunker, Z. Obradovic, and V.N. Uversky, "Protein disorder in the human diseasome: unfoldomics of human genetic diseases," *BMC Genomics*, 2009, 10(Suppl 1), S12.

[3] P. Romero, Z. Obradovic, C. Kissinger, J.E. Villafranca, and A.K. Dunker, "Identifying disordered regions in proteins from amino acid sequence," *Proc. IEEE Int. Conf. on Neural Networks*, Houston, 1997, pp. 90-95.

[4] "CASP Contests Home Page," Internet: http://predictioncenter.org/ [October 26, 2010].

[5] B. He, K. Wang, Y. Liu, B. Xue, V.N. Uversky, and A.K. Dunker, "Predicting intrinsic disorder in proteins: an overview," *Cell Res.*, 2009, 19(8), pp. 929–949.

[6] T. Ishida, and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, 2008, 24(11), pp. 1344–1348.

[7] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved disorder prediction by combination of orthogonal approaches," *PLoS ONE*, 2009, 4(2), e4433.

[8] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, and V.N. Uversky, "PONDR-FIT: a meta-predictor of intrinsically disordered amino acids," *Biochim Biophys Acta*, 2010, 1804(4), pp. 996-1010.

[9] M.J. Mizianty, W. Stach, K. Chen, K.D. Kedarisetti, F.M. Disfani, and L. Kurgan, "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources," *Bioinformatics*, 2010, 26(18), pp. i489-i496.

[10] M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, Z. Obradovic, and A.K. Dunker, "DisProt: the Database of Disordered Proteins," *Nucleic Acids Res.*, 2007, 35(Database issue), pp. 786-793.

[11] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: a computer-based archival file for macromolecular structures," *J. Mol. Biol.*, 1977, 112(3), pp. 535-542.

[12] O. Noivirt-Brik, J. Prilusky, and J.L. Sussman, "Assessment of disorder predictions in CASP8," *Proteins*, 2009, 77(Suppl 9), pp. 210-216.

[13] J.J. Ward, L.J. McGuffin, K. Bryson, B.F. Buxton, and D.T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinformatics*, 2004, 20(13), pp. 2138-2139.

[14] V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, and L. Moy, "Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit," *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, Montreal, 2009, pp. 889-896.

[15] C. Bishop, *Pattern recognition and machine learning*, New York: Springer, 2006, pp. 203-213.

[16] Y. Jin, and R.L. Dunbrack, "Assessment of disorder predictions in CASP6," *Proteins*, 2005, 61(Suppl 7), pp. 167-175.

[17] L. Bordoli, F. Kiefer, and T. Schwede, "Assessment of disorder predictions in CASP7," *Proteins*, 2007, 69(Suppl 8), pp. 129-136.