ABSTRACT

Motivation: Order and Disorder prediction using Conditional Random Fields (OnD-CRF) is a new method for accurately predicting the transition between structured and mobile or disordered regions in proteins. OnD-CRF applies CRFs relying on features which are generated from the amino acids sequence and from secondary structure prediction. Benchmarking results based on CASP7 targets, and evaluation with respect to several CASP criteria, rank the OnD-CRF model highest among the fully automatic server group.

Availability: http://babel.ucmp.umu.se/ond-crf/
Contact: Uwe.Sauer@ucmp.umu.se

1 INTRODUCTION

Many proteins carry out important biological functions by means of intrinsically unstructured sequence intervals (Dunker et al., 2002; Romero et al., 1999). Identification of disordered regions in protein sequences can help to reduce bias in sequence similarity analyses and delineate boundaries of protein domains to guide structural and functional studies (Ferron et al., 2006).

Several state-of-the-art approaches have been proposed for prediction of ordered and disordered residues, such as neural networks (NNs) and support vector machines (SVMs). Similar to NNs and SVMs, the conditional random fields (CRFs) (Lafferty et al., 2001) are discriminative supervised machine learning methods. CRFs need training with labeled empirical data in order to learn the classification. However, compared to NNs and SVMs, CRFs are able to take into account interrelation information between two labels of neighboring residues.

Here, we describe how to apply CRFs to build an order and disorder prediction model. We then compare the OnD-CRF method to prediction methods that successfully participated in CASP7 (Bordoli et al., 2007).

2 METHODS

The OnD-CRF training dataset is derived from high-resolution crystal structures. It contains 215 612 residues, of which 13 909 are defined as disordered (Cheng et al., 2005a) since they are part of a crystallized protein but lack a coordinate entry in the PDB file. The training dataset does not contain any of the CASP7 targets.

Performance is assessed with respect to the area under the ROC curve (AUC), the average of sensitivity and specificity (ACC) and a weighted score, $S_w$, that considers the rates of ordered and disordered residues in the dataset (Jin and Dunbrack, 2005). The AUC is a measure of the overall predictor quality, with a value of 1.0 for a perfect predictor and 0.5 for a random predictor. The weighted score $S_w$ and the ACC, introduced in CASP6 and CASP7, are used to evaluate the overall prediction accuracy based on an imbalanced dataset.

Throughout, we use CRF++ 0.49 (http://crfpp.sourceforge.net/) to generate the OnD-CRF. The template file contains the rules for selecting the features that we use for training the OnD-CRF model. The features are extracted only from the amino acid sequence and, using SSpro (Cheng et al., 2005b), from the predicted secondary structure.

After 10-fold cross validation, we find that a sliding window size of nine amino acids yields the best template file. The set of parameters which give rise to an AUC value of 0.864 for the OnD-CRF build on the training dataset are: 1.018 for the hyper-parameter ‘C’, which trades the balance between overfitting and underfitting and 5 for the parameter ‘I’, which sets the cut-off threshold for the features. For all other parameters we use the default CRF++ 0.49 values. As a result of 10-fold cross validation, we find an optimal $P$-value cut-off of $P < 0.05$ for ordered and $P \geq 0.05$ for disordered amino acids. Using this cut-off the OnD-CRF model achieves an ACC of 0.790 and a weighted score $S_w$ of 0.580, based on the training dataset.

3 BENCHMARKING RESULTS

For benchmarking, we use all 96 targets available during CASP7 and compare the results obtained with OnD-CRF to the results of the 15 methods that predicted 93 or more targets. Evaluation is done with respect to the AUC, the sensitivity, $S_{\text{sens}}$, the specificity, $S_{\text{spec}}$, their product, $S_{\text{prod}}$, the ACC and $S_w$. The sensitivity and specificity can be interpreted as the fraction of correctly identified disordered and ordered residues, respectively.

The benchmarking results for all 16 disorder prediction methods, subdivided into the fully automated server group and the human expert group, are listed in Table 1. Within the automatic server group, the OnD-CRF method reaches the best overall performance with highest scores for AUC, $S_{\text{sens}}$, $S_{\text{prod}}$, ACC and $S_w$. The performance of OnD-CRF method is comparable to the best human expert methods such as ISTZORAN and fais. The results show, that OnD-CRF is an

*To whom correspondence should be addressed.
Table 1. Comparing OnD-CRF with prediction methods that participated in CASP7

| Method                  | AUC  | $S_{\text{sens}}$ | $S_{\text{spec}}$ | $S_{\text{prod}}$ | ACC  | $S_w$ |
|-------------------------|------|------------------|-------------------|-------------------|------|------|
| CASP7 Automatic Server Group |      |                  |                   |                   |      |      |
| OnD-CRF$^a$             | 0.839 | 0.688            | 0.813             | 0.560             | 0.750 | 0.501 |
| DISpro$^c$              | 0.822 | 0.597            | 0.854             | 0.510             | 0.726 | 0.451 |
| GeneSilicoMetaServer$^d$| 0.804 | 0.527            | 0.912             | 0.481             | 0.720 | 0.440 |
| BIME@NTU$_{\text{serv}}^a$ | 0.798 | 0.591            | 0.839             | 0.496             | 0.715 | 0.430 |
| DISOPRED$^e$            | 0.837 | 0.425            | 0.953             | 0.405             | 0.689 | 0.378 |
| Distill$^a$             | 0.724 | 0.558            | 0.788             | 0.440             | 0.673 | 0.346 |
| MBI-NTU$_{\text{serv}}^a$ | 0.796 | 0.327            | 0.971             | 0.318             | 0.649 | 0.298 |
| DRIPPRED$^b$            | 0.758 | 0.383            | 0.908             | 0.348             | 0.646 | 0.291 |
| CASP7 Human Expert Group |      |                  |                   |                   |      |      |
| ISTZORAN$^b$            | 0.860 | 0.725            | 0.837             | 0.607             | 0.781 | 0.562 |
| fas$^a$                 | 0.844 | 0.556            | 0.924             | 0.514             | 0.740 | 0.481 |
| CBRC-DR$^a$             | 0.850 | 0.454            | 0.966             | 0.439             | 0.710 | 0.420 |
| BIME@NTU$_{\text{spec}}^f$ | 0.804 | 0.536            | 0.883             | 0.473             | 0.710 | 0.419 |
| IUPred$^b$              | 0.777 | 0.396            | 0.947             | 0.375             | 0.672 | 0.343 |
| CBRC-DRP$_{\text{DE spec}}^a$ | 0.704 | 0.338            | 0.971             | 0.328             | 0.655 | 0.309 |
| Oka$^b$                 | 0.609 | 0.280            | 0.937             | 0.262             | 0.609 | 0.218 |
| Softberry$^a$           | 0.704 | 0.201            | 0.971             | 0.195             | 0.586 | 0.172 |

The entries are sorted with respect to the weighted score $S_w$. Number of predicted targets: $a96$; $b95$; $c94$; $d93$; AUC: Area Under ROC Curve (Bordoli et al., 2007); $S_{\text{sens}} = TP/(TP + FN)$; $S_{\text{spec}} = TN/(FN + TP)$; $S_{\text{prod}} = S_{\text{sens}} + S_{\text{spec}}$; ACC = $(S_{\text{sens}} + S_{\text{spec}})/2$; $S_w = (W_{\text{order}}N_{\text{order}} + W_{\text{disorder}}N_{\text{disorder}})/(W_{\text{disorder}}N_{\text{disorder}} + W_{\text{order}}N_{\text{order}})$. OnD-CRF predictions can be used to determine domain boundaries for 3D structure analysis.

ACKNOWLEDGEMENT

Funding: This work is supported by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at Karolinska Institute, Strategy and Development Office as well as by BIOVITRUM AB and Umetrics AB.

Conflict of Interest: none declared.

REFERENCES

Bordoli, L., et al. (2007) Assessment of disorder predictions in CASP7. Proteins, 69 (Suppl. 8), 129–136.
Cheng, J., et al. (2005a) Accurate prediction of protein disordered regions by mining protein structure data. Data Min. Knowl. Disc., 11, 213–222.
Cheng, J., et al. (2005b) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res., 33, W72-W76.
Donaldson, L.W., et al. (2002) Structure of a regulatory complex involving the Abl SH3 domain, the Crk SH2 domain, and a Crk-derived phosphopeptide. Proc. Natl Acad. Sci. USA, 99, 14053-14058.
Dunker, A.K., et al. (2002) Intrinsic disorder and protein function. Biochemistry, 41, 6573–6582.
Ferron, F., et al. (2006) A practical overview of protein disorder prediction methods. Proteins, 65, 1-14.
Jin, Y. and Dunbrack, R.L., Jr. (2005) Assessment of disorder predictions in CASP6. Proteins, 61 (Suppl. 7), 167–175.
Kobashigawa, Y., et al. (2007) Structural basis for the transforming activity of human cancer-related signaling adaptor protein CRK. Nat. Struct. Mol. Biol., 14, 503–510.
Lafttery, J., et al. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In 18th International Conference on Machine Learning (ICML). pp. 282–289.
Romero, P., et al. (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett., 462, 363-367.