Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences

Ronesh Sharma[2,3], Gaurav Raicar[2], Tatsuhiko Tsunoda[1,4,5], Ashwini Patil[6,§,*] and Alok Sharma[1,4,5,7,§,*]

[1]RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan, [2]The University of the South Pacific, Fiji, [3]Fiji National University, Fiji, [4]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan, [5]CREST, JST, Tokyo 113-8510, Japan, [6]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan, [7]Griffith University, Australia.

* To whom correspondence should be addressed.

§Last authors

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Intrinsically disordered proteins lack stable 3-dimensional structure and play a crucial role in performing various biological functions. Key to their biological function are the molecular recognition features (MoRFs) located within long disordered protein sequences. Computationally identifying these MoRFs is a challenging task. In this study, we present a new MoRF predictor, OPAL, to identify MoRFs in disordered protein sequences.
Method: OPAL utilizes two independent sources of information computed using different component predictors whose scores are processed and combined using common averaging method. The first score is predicted using a component MoRF predictor which utilizes composition and sequence similarity of MoRF and non-MoRF regions to detect MoRFs. The second score is predicted using half-sphere exposure (HSE), solvent accessible surface area (ASA) and backbone angle information of the disordered protein sequence. The second score mainly targets the amino acid properties of flanking regions surrounding the MoRFs to distinguish MoRF and non-MoRF residues.
Results. OPAL is evaluated using multiple test sets that have been previously used to evaluate MoRF predictors. The results demonstrate that OPAL outperforms all the available MoRF predictors and is the most accurate predictor available for MoRF prediction.
Availability: http://www.alok-ai-lab.com/tools/opal/
Contact: ashwini@hgc.jp, alok.sharma@griffith.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent progress in computational and experimental methods have revealed many protein regions lacking stable 3-dimensional structure (Dyson and Wright, 2005; Lee, et al., 2014; Uversky, 2014; Wright and Dyson, 2015). These protein regions perform various biological functions such as cell regulation and signal transduction (Lee, et al., 2014; Uversky, 2014) . Proteins with such regions are known as intrinsically disordered proteins (IDPs) (Dyson and Wright, 2005; Tompa, 2011). IDPs often execute their function with loosely structured short protein regions that bind to a structured partner and undergo a disorder-to-order transition to adopt a well-defined conformation (Lee, et al., 2014; Mohan, et al., 2006; Vacic, et al., 2007). These short regions are known as short linear motifs (SLiMs) and molecular recognition features (MoRFs). SLiMs are short linear sequence motifs that vary in size from 3 to 10 amino acids and are enriched in intrinsically disordered regions (IDRs) (Edwards, et al., 2007). On the other hand, MoRFs are long disordered regions that fold upon binding to their partner protein and are up to 70 amino acids in length (Mohan, et al., 2006). MoRFs were first introduced as Molecular Recognition Elements (MoREs) (Oldfield, et al., 2005) and their role in protein-protein interactions was elucidated (Liu, et al., 2006; Mohan, et al., 2006; Vacic, et al., 2007).

The functional importance of MoRFs has led to the development of several computational methods and predictors including the very early ones (Cheng, et al., 2007; Oldfield, et al., 2005), and more recent efforts such as ANCHOR (Dosztányi, et al., 2009), MoRFpred (Disfani, et al., 2012), MoRFchibi (Malhis and Gsponer, 2015), MoRFchibi-light (Malhis, et al., 2016) and MoRFchibi-web (Malhis, et al., 2016; Malhis, et al., 2015) and our previous work (Sharma, et al., 2016).

In this study, we present OPAL, to predict MoRFs of sizes 5 to 25 residues located within long disordered protein sequences. OPAL is an ensemble of two predictors: MoRFchibi and Prediction of MoRFs Incorporating Structure (PROMIS), which is also described in this work. OPAL combines MoRF scores at multiple stages using common averaging method. The first score is calculated using a component MoRF predictor, MoRFchibi. The score is processed and is combined with the score of PROMIS. PROMIS is constructed using half-sphere exposure (HSE) (Hamelryck, 2005), solvent accessible surface area (ASA) and backbone angle information of disordered protein sequences to predict MoRFs. The development of PROMIS offered a significant improvement in prediction accuracies when compared with MoRFchibi, MoRFpred and ANCHOR predictors. Overall, the integration of PROMIS with MoRFchibi provided better prediction quality for OPAL compared with predictors of similar approach e.g. MoRFchibi-light and MoRFchibi-web. OPAL is available as an online server at http://www.alok-ai-lab.com/tools/opal.

2 Method

2.1 Benchmark dataset

We used training and test sets that were previously introduced by Disfani *et. al*. (Disfani, et al., 2012) to develop MoRF predictors. These sets were recently used to train and benchmark predictors MoRFchibi, MoRFchibi-light and MoRFchibi-web. The training set contains 421 protein sequences with 245,984 residues, of which 5,396 are MoRF residues. The test set contains 419 protein sequences with 258,829 residues, of which 5,153 are MoRF residues. A second test set, named NEW in Malhis *et. al*. (Malhis and Gsponer, 2015) contains 45 sequences with 37,533 residues, of which 626 are MoRF residues. These sets were collected from Protein Data Bank (PDB) (Disfani, et al., 2012) and peptide regions of 5 to 25 residues were identified as MoRF regions. All the test sequences share less than 30% sequence identity to sequences in the training set (Disfani, et al., 2012). We use the training set to train OPAL and the first test set to evaluate it. We further combine first and the second test sets into single set, TEST464 as in Malhis *et. al*. (Malhis, et al., 2016) and use it to compare the proposed predictor with the state-of-the-art MoRF predictors. Although TEST464 contains sequences that were used to test previous MoRF predictors (Disfani, et al., 2012; Malhis and Gsponer, 2015; Malhis, et al., 2016), we found that 42% of the sequences in TEST464 have sequence identity 30% or more with one other sequence in the set. To address this, we used cd-hit (Li and Godzik, 2006) to remove the sequences from TEST464 which share 30% or more sequence identity. This resulted in 266 sequences and we called this filtered set TEST266. We used an additional test set, EXP53, which was collected and assembled by Malhis *et. al*. (Malhis, et al., 2015). There are 53 non-redundant protein sequences in this set containing MoRF regions that are experimentally verified to be disordered in isolation. EXP53 set was filtered to have sequences with less than 30% sequence identity to those in the training set. EXP53 sequences also share less than 30% sequence identity with each other (Malhis, et al., 2015). EXP53 set contains 25,186 residues, of which 2,432 are MoRF residues. Since protein sequences in EXP53 set contain MoRFs with length greater than 30 residues, MoRFs are further divided into short MoRFs (up to 30 residues) and long MoRFs (longer than 30 residues). For the rest of the paper, we refer to short MoRFs as EXP53short, long MoRFs as EXP53long and all MoRFs as EXP53all. We used TEST266 and EXP53 sets to compare and validate the proposed predictor.

2.2 The PROMIS model

In order to distinguish between MoRF and non-MoRF residues, the proposed PROMIS model uses structural information of disordered regions to compute amino acid properties of flanking regions surrounding the MoRFs. The structural information includes attributes such as HSE (Hamelryck, 2005), ASA and backbone angles of amino acids in disordered regions predicted via Spider2 (Heffernan, et al., 2015; Yang, et al., 2017), a sequence predictor of local and non-local structural features of protein sequences. Two different methods of feature extraction are employed to retrieve meaningful features from structural attributes. The first method is based on profile bigram (Sharma, et al., 2013), where the feature vector is obtained by counting the bigram frequencies from the position specific scoring matrix (PSSM) representing a protein region. However, in this paper we do not apply PSSM to compute profile bigram, instead we used structural attributes to evaluate bigram features. The second method is based on the properties of flanks surrounding the MoRF residue. In this method, feature vector is extracted from structural attributes to encode the properties of flanks surrounding the query residue. More details on the above two methods are given later. For the rest of the paper, we refer to the above two methods as BigramMoRF and StructMoRF, respectively (please see Supplementary Text S1). The feature vectors generated using the above two methods are sent to an SVM model for prediction. The architecture of the proposed PROMIS predictor is shown in Figure 1. The prediction scores of the SVM model obtained from each of the methods described above are combined using the common averaging strategy to produce propensity scores of PROMIS. In common averaging, the score of all SVM models is added and further divided by the number of models used.

To construct PROMIS, we took a similar approach as we took in our previous study (Sharma, et al., 2016) to divide each training sequence into two segments. Using the first segment, we extract positive samples representing MoRFs and using the second segment, we extract negative samples representing non-MoRFs. Our previous study used a fixed flank length of 12 amino acids surrounding the MoRF region to create segments. However, in this study we varied the flank length from 12 to 25 amino acids to identify the length that best discriminates MoRF residues from non-MoRF residues. Using AUC performance measure on the test data (please see Supplementary Table S1), we selected the optimal flank length to be 20. The following subsections outline the structural attributes, feature extraction methods, and training and test of the SVM model.

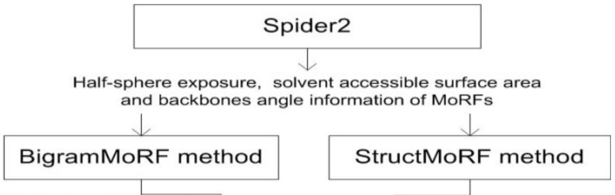

Spider2
Half-sphere exposure, solvent accessible surface area
and backbones angle information of MoRFs
BigramMoRF method
StructMoRF method

Fig.1: Architecture of the PROMIS predictor. PROMIS is constructed using structural attributes of disordered regions. The scores are processed using common averaging. In common averaging, the score of all SVM models is added and further divided by the number of models used.

2.2.1 Structural attributes

Spider2 (Yang, et al., 2017) output is used as a source of feature extraction. It predicts structural information about the protein sequences which includes:

- Secondary structure (SS): contains structural description of each amino acid residue in a number of discrete states, such as helix, sheet and coil.
- Accessible surface area (ASA): measures the exposure level of amino acid residue to solvent in a protein region and is a one–dimensional structural property.
- Backbone angles: includes backbone dihedral angles of amino acids in protein region. We consider the Phi, Psi, Theta (θ) and Tau (τ) angles. θ is the angle between $C\alpha$ atoms ($C\alpha_{i-1}$ – $C\alpha_i$ – $C\alpha_{i+1}$) and τ is the dihedral angle rotated about the $C\alpha_i - C\alpha_{i+1}$ bond .
- Half-sphere exposure (HSE): is an alternative measure of the solvent exposure of a residue and has been shown to perform better than ASA (Heffernan, et al., 2016). It gives the number of C alpha atoms in the upper and lower spheres defined for each residue. We use two measures specifying the HSE alpha and beta (HSEu and HSEd) along with the contact number for each residue.

2.2.2 Feature extraction

To extract feature vectors from structural attributes, the following feature extraction methods are considered:

- BigramMoRF: in this method, we computed bigram features by utilizing structural attributes. The bigram features from *k*-th attribute to *l*-th attribute (in a protein sequence) is computed as follows:

  $$B_{k,l} = \frac{1}{L}\sum_{i=1}^{L-1} M_{i,k}\ M_{i+1,l}\ (1 \le k \le n \text{ and } 1 \le l \le n) \qquad (1)$$

  where $M_{i,k}$ is the element of structural matrix M of size L by n, L is the length of a protein region and n is the number of structural attributes. Computing the bigram frequencies $B_{k,l}$ for $k = 1,2, \dots n$ and $l = 1,2, \dots, n$ would give a bigram matrix B of size $n \times n$. This matrix B can be represented as a vector form F_b by reshaping the $n \times n$ matrix into a vector of length n^2. F_b is a bigram of attributes and not basis expansion of feature vectors. The use of bigram features has shown promising results in protein fold recognition, protein subcellular localization, structural class prediction, functional analysis, drug-interaction and other related problems (Kavianpour and Vasighi, 2017; Lyons, et al., 2015; Mousavian, et al., 2016; Peng, et al., 2017; Sharma, et al., 2013; Sharma, et al., 2015; Xia, et al., 2017).
- StructMoRF: to represent a protein region, in this method the attribute values are treated as features. i.e., the feature vector for a sample can be interpreted as $F_s = [M_{1,1}, M_{2,1}, \dots, M_{i,j}, \dots, M_{L,n}]$, where $M_{i,j}$ is an element of structural matrix M of size L by n. As before L is the length of a protein region and n is the number of attributes. F_s is a tensor sum of attributes.

2.2.3 SVM model

An SVM classifier with radial basis function (RBF) is used to evaluate the features generated. Performing a grid search, SVM kernel parameters *C* and gamma were selected as 1000 and 0.0038, respectively (please see Supplementary Text S2).

2.2.4 Training

For training, since there are more non-MoRF residues compared to MoRF residues in training data, balanced sampling is done by extracting equal number of positive and negative samples. This ratio is further increased to 1:2, i.e. two non-MoRF samples for each MoRF sample, to obtain higher AUCs in detecting MoRF residues (please see Supplementary Table S2 for details). For BigramMoRF method, samples are chosen to represent a region of MoRF residues with a flank of 20 amino acids upstream and downstream of the selected region. On the other hand, for StructMoRF method, samples are chosen to represent a MoRF residue with a flank of 20 amino acids upstream and downstream of the selected residue. The feature vector is computed from the sample and is used for training the model. The detailed procedure of extracting positive and negative samples from training data is illustrated in supplementary information (please see Supplementary Text S1).

2.2.5 Testing

To score a query sequence, we take a query sample from the query sequence to represent each residue. The feature vector is extracted from the sample and is used for scoring. The detailed explanation and illustration of scoring a query sequence is demonstrated in supplementary information (please see Supplementary Text S1).

2.3 Score calculation

In order to validate that the predicted residue is a part of a binding region, we process each predicted score using its neighboring residue scores. Taking the score of query residue at *i*-th location and its neighboring residue scores of size *z* on both sides, we compute the processed propensity score for each residue as follows:

$$\text{Processed score}_i = (\max(\text{scores}_{\mathrm{x}}) + \text{median}(\text{scores}_x))/2 \qquad (2)$$

where *i*=1,2,…, *L*, *L* is the length of query protein sequence and *x* varies from *i-z* to *i+z*.

2.4 Combined model (OPAL)

We applied common averaging technique to combine the proposed PROMIS predictor with component MoRF predictor, MoRFchibi into a single model called OPAL. To predict MoRFs, MoRFchibi targets similarity, composition and contrast information of MoRF and non-MoRF regions using physicochemical properties of amino acids. MoRFchibi utilizes two SVM models with two different kernel functions (sigmoid and RBF). The first kernel (sigmoid) is used to distinguish sequence similarity of query regions to that of training regions and the second kernel (RBF) is used to extract composition and contrast information between the MoRF region and its surrounding regions. In this framework, MoRFchibi is used as one of the components of OPAL and its propensity scores are processed and combined with PROMIS in this study. The details of score processing is described in subsection 2.3. The overview of the combined predictor, OPAL is shown in Figure 2.

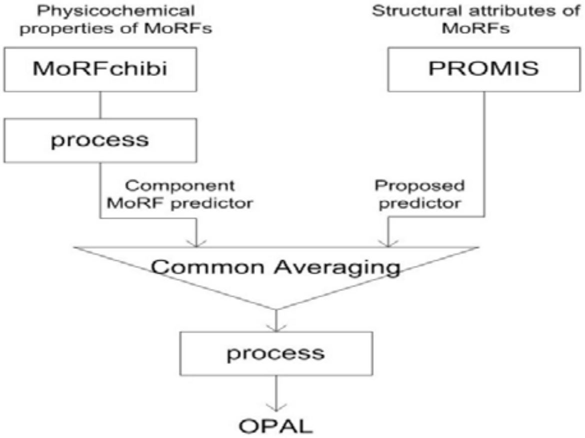


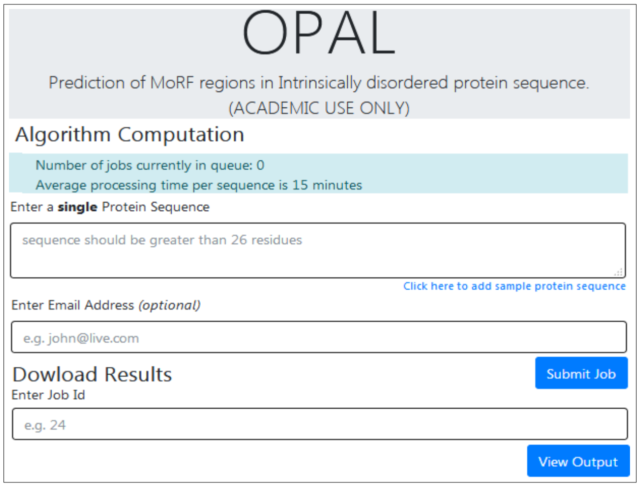
Fig.2: Overview of OPAL predictor. OPAL is constructed using processed MoRFchibi component predictor and PROMIS predictor proposed in this study. The scores are processed using common averaging. To use MoRFchibi as a component predictor, we downloaded MoRFchibi and interfaced it with PROMIS.

2.5 Performance measure

To evaluate OPAL, we used AUC performance measure. AUC is the area under the receiver operating characteristics (ROC) curve and is commonly used to evaluate a predictor to see how well it separates two classes of information, i.e. MoRF and non-MoRF residues. We also report precision, F-measure and false positive rate (FPR) for different values of true positive rate (TPR), since we are interested in predicting MoRFs at a high threshold probability which is near the lower left corner of the AUC curve. TPR is defined as TP/N_{MoRF} and FPR is defined as $FP/N_{non-MoRF}$, where TP is the number of correctly classified MoRF residues, FP is the number of incorrectly predicted MoRF residues, N_{MoRF} is the total number of MoRF residues and $N_{non-MoRF}$ is the total number of non-MoRF residues. To report the processing speed of the predictor, we noted the processing time of the predictor to score a protein sequence and used it to compute the number of residues it predicts in one minute i.e., residues/minute (r/m).

2.6 OPAL Online Server

OPAL is available as an online server at http://www.alok-ai-lab.com/tools/opal/. The details of using OPAL online server are as follows: Opal accepts input as a single protein sequence of length greater than 26 amino acids. A screenshot of the top page of OPAL online server is shown in Figure 3. To use OPAL, users need to enter a protein sequence and email address (optional) before submitting a job to OPAL online server. Once the job is processed, the result can be downloaded using the job ID assigned to the submission. It takes an average processing time of 15 to 20 minutes to process a job. If the user provides an email address with the job submission, then notification is sent to the email once the job is processed. A screenshot of the output is shown in supplementary Figure (please see Supplementary Figure.S1).



OPAL

Prediction of MoRF regions in Intrinsically disordered protein sequence.
(ACADEMIC USE ONLY)

Algorithm Computation

Number of jobs currently in queue: 0
Average processing time per sequence is 15 minutes

Enter a **single** Protein Sequence

sequence should be greater than 26 residues

Click here to add sample protein sequence

Enter Email Address *(optional)*

e.g. john@live.com

Dowload Results

Submit Job

Enter Job Id

e.g. 24

View Output

Fig.3: OPAL online server homepage. A screenshot to show the top page of OPAL online server. Its website address is http://www.alok-ai-lab.com/tools/opal/.

3 Results

OPAL is trained and tested using the same data that was used to train and evaluate the predictors, MoRFpred, MoRFchibi and MoRFchibi-web. These datasets are described in detail in Disfani *et. al*. (Disfani, et al., 2012) and Malhis *et. al.* (Malhis, et al., 2015). We use test sets to evaluate the proposed predictor and the set EXP53 to validate that the performance improvement is not the result of over fitting. In addition, since all the mentioned and the proposed predictors are trained to predict MoRFs of sizes 5 to 25 residues while EXP53 set contains sequences with MoRFs of length greater than 30 residues, we show performance of EXP53 as EXP53all, EXP53long and EXP53short, where EXP53all contains all the MoRFs from 53 sequences, EXP53long contains MoRFs that are greater than 30 residues in size and EXP53short contains MoRFs that are up to 30 residues in size.

3.1 Attribute and model selection

Evaluating the test set, we select important structural attributes and models to identify MoRFs in disordered protein sequence. We use successive feature selection scheme in the forward direction (Sharma, et al., 2013) to rank each structural attribute according to its contribution towards successfully predicting MoRFs. For BigramMoRF method, HSEα attributes were ranked highest amongst structural attributes, θ attribute was ranked highest amongst the dihedral angles and was ranked second overall. Moreover, ASA attribute performed average and τ angle attribute was ranked lowest. For StructMoRF method, HSEu attribute from HSEα group was ranked highest and gave good prediction accuracies in first stage of selection, however, in the second stage its combination with other attributes deteriorated the accuracies. Thus, to obtain average performance, we construct three SVM models, MoRFbi-1, MoRFbi-2 and MoRFwin as shown in Figure 1. MoRFbi-1 and MoRFbi-2 are constructed using BigramMoRF method and MoRFwin is constructed using StructMoRF method. For feature extraction MoRFbi-1 uses attributes HSEα and ASA; MoRFbi-2 uses attribute θ from dihedral angles; and, MoRFwin uses attribute HSEu. Furthermore, for scoring a query sequence, the window size for extracting a sample was set as 70 and 41 for BigramMoRF and StructMoRF methods, respectively. We selected these sizes, because AUCs computed were highest with these sizes compared with other window sizes for each method.

Table 1 shows the AUCs for model trained with training sampling ratio 1:1 and 1:2. First model performed well with sampling ratio of 1:2, whereas second and third models gave good AUCs with sampling ratio 1:1 (for more details on model selection please see Supplementary Table S2). Thus, we select best performing models and combine their scores using common averaging method. The selected and combined models are shown in Table 2.

Table 1: AUCs for models with sampling ratio 1:1 and 1:2 using test set.

		Sampling ratio 1:1	Sampling ratio 1:2
	Models	AUC	AUC
1	MoRFbi-1	0.734	**0.760**
2	MoRFbi-2	**0.689**	0.652
3	MoRFwin	**0.769**	0.738

Bold numbers indicate best performance measure.

Table 2: AUCs for selected and combined models using test set.

	Models	Sampling ratio	Test AUC
1	MoRFbi-1	1:2	0.760
2	MoRFbi-2	1:1	0.689
3	MoRFwin	1:1	0.769
Combined	PROMIS		**0.791**

Combined PROMIS model with significant improvement in AUCs compared to individual selected models.

To validate that the predicted scores as MoRFs form part of the binding region, we use equation (2) to process each predicted score by varying parameter *z* (parameter *z* refers to the size of neighboring residue scores). Figure 4 shows the AUCs for different values of *z* for each of the model. It is observed that models MoRFbi-1, and MoRFbi-2 obtain optimal results at *z* =20, whereas MoRFwin obtain optimal result at *z* =12. Furthermore, since MoRFchibi is used in our proposed combined model OPAL, we also process MoRFchibi scores and found that it obtains optimal result at *z*=4 as observed in Figure 4. Thus, to develop our final predictors PROMIS and OPAL, we process the mentioned model scores at specified *z* parameters giving the best results.

3.2 Comparison with state-of-the-art predictors

To compare the performance of the proposed predictor with available state-of-the-art MoRF predictors, we use datasets TEST464, TEST266 and EXP53 to report the AUCs. We show performance of PROMIS and OPAL, in Table 3 and Table 4, respectively. In Table 3 PROMIS is compared with predictors ANCHOR, MoRFpred and MoRFchibi. These predictors are developed using similar approaches, whereas in Table 4 we compare MoRF predictors which are constructed using many other component predictors and their scores are combined at several stages to produce the final MoRF propensity scores. These predictors are MoRFchibi-light, MoRFchibi-web and OPAL.

PROMIS achieves significant improvements in predicting MoRFs. Compared with MoRFchibi, it provided 4.7% increase in AUCs for TEST464 dataset, 10.6% increase in AUCs for EXP53all, 13.6% increase in AUCs for EXP53long and 3.3% increase in AUCs for EXP53short datasets, respectively.

Table 3: AUCs for predictors of similar approach.

Predictor /methods	TEST464	TEST266	EXP53all	EXP53long	EXP53short
ANCHOR	0.605	0.599	0.615	0.586	0.683
MoRFpred	0.675	0651	0.620	0.598	0.673
MoRFchibi	0.743	0.709	0.712	0.679	0.790
PROMIS	**0.790**	**0.770**	**0.818**	**0.815**	**0.823**

In bold, PROMIS shows significant improvement in prediction accuracies, compared to MoRFchibi. MoRFpred and ANCHOR.

Table 4: AUCs for combined component MoRF predictors.

Predictor /methods	TEST464	TEST266	EXP53all	EXP53long	EXP53short
MoRFchibi-light	0.777	0.762	0.799	0.770	0.869
MoRFchibi-web	0.805	0.785	0.797	0.758	0.886
OPAL	**0.816**	**0.795**	**0.836**	**0.823**	0.870

In bold, OPAL shows overall improvement in prediction accuracies, compared to MoRFchibi-light and MoRFchibi-web.

Incorporating a number of component predictors is thought to increase the performance; this is observed in Table 4. All the combined predictors perform well in comparison with individual predictors observed in Table 3. Compared with MoRFchibi-light and MoRFchibi-web, OPAL obtained 3.9% and 1.1% increase in AUCs for TEST464 dataset, 3.9% and 3.7% increase in AUCs for EXP53all, 6.5% and 5.3% increase in AUCs for EXP53long and performed very similar to MoRFchibi-light and MoRFchibi-web for EXP53short dataset, respectively. The AUC curves generated using each of the dataset is shown in Figure 5. Moreover, it is observed that the proposed predictor achieves much lower false positive rate (FPR) at any given true positive rate (TPR) (please see Supplementary Table.S3). All the mentioned MoRF predictors in this study were designed to predict MoRFs up to size of 25 residues, therefore, it was important to know their performance in scoring longer MoRFs. Thus, PROMIS and OPAL have shown significant increase in accuracies for predicting these MoRFs. For comparison, we also computed precision and F-measure for different values of TPR as observed in Table 5.

3.3 Processing speed

MoRF predictors are used to score large sets of proteins; therefore, it is necessary to test its efficiency. We compare and report the prediction speed for each of the predictor. For MoRFchibi-light, MoRFchibi and ANCHOR, we tested these predictors using the entire TEST set using i5, 3.5GHz computer, since these predictors do not require multiple sequence alignments (MSA). On the other hand, MoRFchibi-web and OPAL required MSA, therefore, we test both these predictors using a single sequence from test set (Uniprot:Q38087) with 903 residues. Predictor MoRFpred is not downloadable, so it was tested on its prediction server with single sequence (Uniprot:Q38087). The processing speed of each predictor with its AUCs are summarized in Table 5. Prediction speed for ANCHOR, MoRFchibi and MoRFchibi-light do not require generation of evolutionary profiles, therefore were fastest with speeds of 3.9×10e+6 r/m, 10.5×10e+3 r/m and 9.9×10e+3 r/m, respectively. The prediction speed of OPAL came third with 215 r/m, whereas MoRFchibi-web provided speed of 80 r/m and MoRFpred came slowest with 48 r/m. Additionally, processing single sequence using MoRFchibi-web on its prediction server provided its speed of 588 r/m, however, the server hardware processor is unknown. The comparison might not be entirely fair, since the prediction server processor for some predictors are unknown and some predictors required the generation of evolutionary profiles.

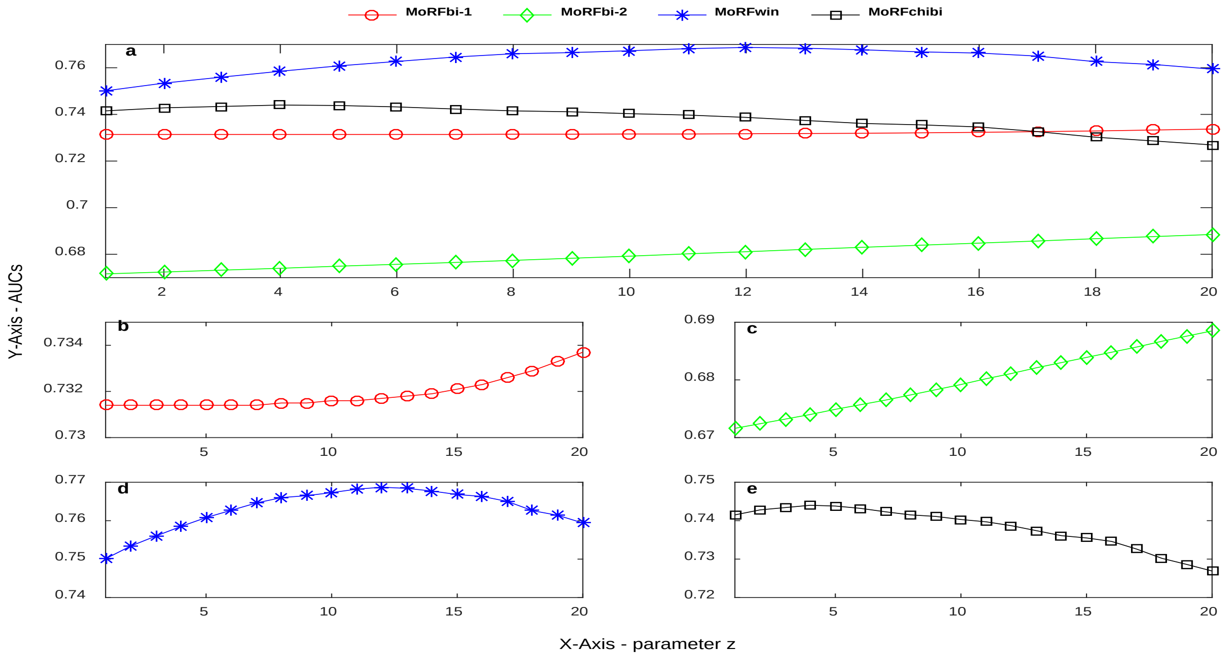


Fig.4: Processed AUCs for each model. The size of parameter *z* in equation (2) is varied from 1 to 20 and suitable size is selected for each model. a) all models AUCs are shown. b) MoRFbi-1 AUCs. c) MoRFbi-2 AUCs. d) MoRFwin AUCs. e) MoRFchibi AUCs. MoRFbi-1 and MoRFbi-2 performed well at *z* =20, MoRFwin performed well at *z* =12, and MoRFchibi performed well at *z* =4.

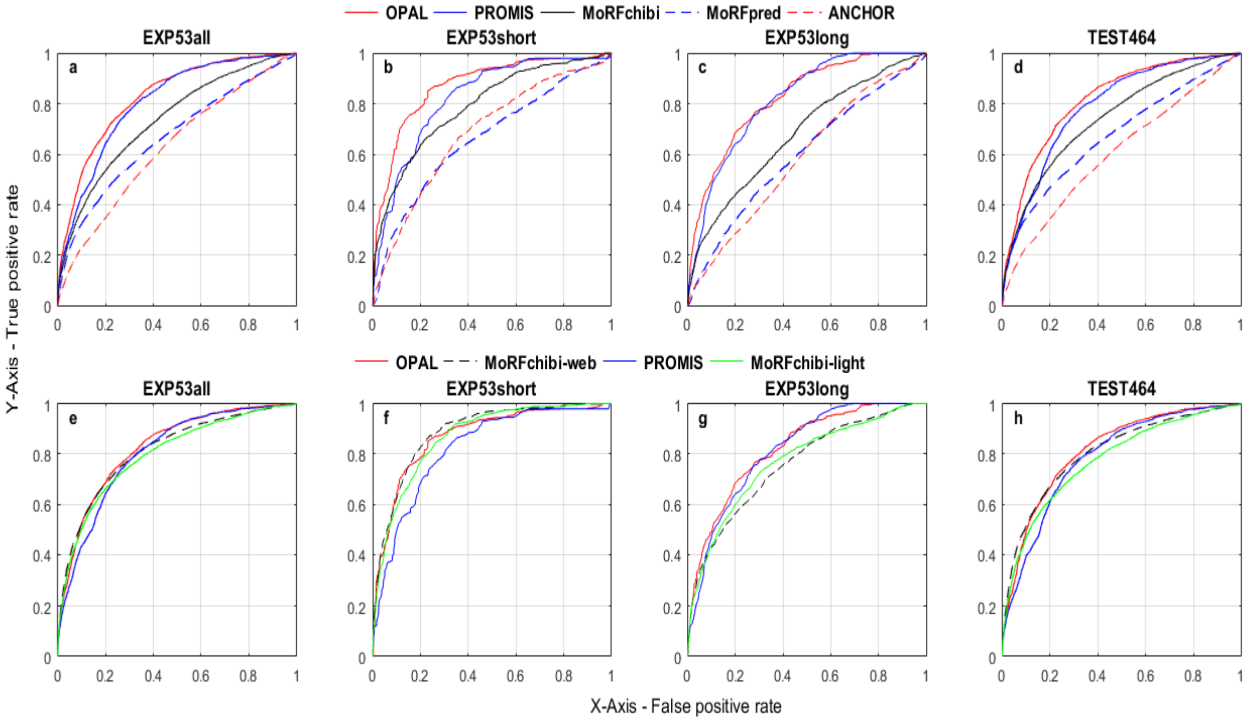


Fig.5: AUC curves generated for each of the datasets, EXP53all, EXP53long, EXP53short and TEST464. Curves a, b, c, and d show OPAL and PROMIS compared with MoRFchibi, MoRFpred and ANCHOR, respectively. Curves e, f, g, and h show OPAL and PROMIS compared with MoRFchibi-web and MoRFchibi-light, respectively.

Table 5: Overall comparison of results

Predictors	Precision	F-measure	AUC	i5 4 core 3.50GHz desktop	Server	Multiple sequence alignments	Combined component predictors
ANCHOR	0.156, 0.134	0.201, 0.212	0.605, 0.615	3.9*10^6	-	×	×
MoRFchibi	0.334, 0.210	0.316, 0.296	0.743, 0.712	10.5*10^3	-	×	×
MoRFpred	0.181, 0.147	0.226, 0.228	0.675, 0.620	-	48	√	×
PROMIS	0.363, 0.332	0.329,0.400	0.790, 0.818	220	-	√	×
MORFchibi light	0.431, 0324	0.354, 0.392	0.777, 0.799	9.9*10^3	-	×	√
MoRFchibi-web	0.495, 0.332	0.373, 0.399	0.805, 0.797	80	588	√	√
OPAL	0.530, 0.386	0.384, 0.436	0.816, 0.836	215	-	√	√

Precision and F-measure is given for TPR values of 0.3 and 0.5, respectively, for EXP53all set and AUC is given for TEST464 and EXP53all sets, respectively.

4 Discussion

We present OPAL, a new sequence based predictor for MoRFs in IDRs. OPAL is developed using processed scores of component MoRFchibi predictor and the scores of proposed PROMIS predictor. We compared its performance with predictors ANCHOR, MoRFpred, MoRFchibi, MoRFchibi-light and MoRFchibi-web. The predictors like MoRFchibi-light and MoRFchibi-web are recently published and they are constructed by combining several other disorder and MoRF component predictors. On the other hand, predictors like ANCHOR, MoRFpred and MoRFchibi are similar to PROMIS. Therefore, we first compare PROMIS with ANCHOR, MoRFpred, and MoRFchibi and then we compared OPAL with MoRFchibi-light and MoRFchibi-web. Using test sets TEST464, TEST266 and EXP53, the results demonstrate that PROMIS outperforms ANCHOR, MoRFpred and MoRFchibi, by observing significant improvement in AUCs. Furthermore, combining component MoRF predictors (such as MoRFchibi-light and MoRFchibi-web), OPAL demonstrated improvement in performance and outperformed the benchmarked MoRFchibi-web predictor.

Achieving higher prediction accuracy for OPAL is the result of combining predictors, PROMIS and MoRFchibi. PROMIS uses structural information of disordered regions for prediction. In the result, it was observed that PROMIS alone provided AUC of 0.790 for TEST464 dataset, whereas MoRFchibi provided AUC of 0.743 only. Using structural features for predicting MoRFs provided enough discrimination information to differentiate MoRFs from its surrounding regions. Compared with physicochemical features of MoRFchibi, they perform well along with the solvent exposure level of amino acids contained in disordered regions. Furthermore, combining PROMIS with MoRFchibi, outperformed all the predictors across the test sets. These predictors use different source of features with different learning algorithms, thus, combining them utilizes the complementary information provided by each, which results in performance improvement. Using validation data set EXP53all with 53 protein sequences, where MoRF regions are experimentally verified to be disordered in isolation, OPAL showed 3.9% performance improvement over MoRFchibi-web and provided lower FPR at any given TPR as shown in Table 6.

The additional improvement for the proposed predictor is the outcome of processing the propensity scores at each stage. By varying and selecting the best size of parameter *z* in equation (2) showed improvement of 0.3% in MoRFbi-1 model, 1.8% in MoRFbi-2 model, 2.6% in MoRFwin model, and 0.4% in MoRFchibi model.

Table 6: FPR as a function of TPR for validating OPAL using EXP53all set

TPR	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MoRFchibi-web	0.016	0.033	0.061	0.107	0.166	0.261	0.382	0.539
OPAL	**0.015**	**0.029**	**0.056**	**0.085**	**0.128**	**0.193**	**0.286**	**0.437**

FPR for TPR values of 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 for predictor OPAL compared with MoRFchibi-web. Note: OPAL beats MoRFchibi-web at any given TPR.

To predict residues in protein sequences as MoRFs or non-MoRFs, predictors are supposed to be consistent over the entire length of the protein sequence. However, if the query samples taken for the regions are very similar to that of training samples, the predictor will over score and produce biased prediction. To overcome this bias, OPAL implements and combines several approaches such as, using two independent sources of information, two different feature extraction methods, selecting best sampling ratios between MoRF and non-MoRF samples during training, and excluding non-MoRF residues neighboring MoRF regions as negative samples. Moreover, using common averaging to combine different models and component predictors with different machine leaning approach makes OPAL less likely to produce biased scores. To show the importance of combining PROMIS with MoRFchibi, we plotted propensity scores of protein P42768 from EXP53 set. This protein contains two verified MoRF regions. Figure 6 shows the propensity scores for models OPAL, PROMIS and MoRFchibi. It is noted that MoRFchibi obtains high scores at the verified MoRF regions, however, it also gives high scores where MoRFs do not exist, i.e., high scores between residues 75 to 140. On the other hand, PROMIS keeps the scores less between residues 75 to 140 and provides above average scores at verified MoRF regions, therefore combining PROMIS with MoRFchibi suppresses the scores to produce higher scores where MoRFs exists

In summary, we have proposed a new MoRF predictor named OPAL using structural information of disordered regions and physicochemical properties of amino acids. Overall, OPAL is the most accurate MoRF predictor available today and it has outclassed the state-of-the-art predictors ANCHOR, MoRFpred, MoRFchibi and MoRFchibi-web.

Acknowledgements

We would like to acknowledge the authors of MoRFpred predictor Disfani *et al.* (2012) and MoRFchibi predictor Malhis *et al* (2015) for making the predictors, and train and test sequence data for MoRF prediction publicly available.

Funding

This work was support by CREST, JST, Yokohama 230–0045, Japan; RIKEN, Center for Integrative Medical Sciences, Japan and Japan Agency for Medical Research and Development (Grant number: 16cm0106320h0001).

Conflict of Interest: None.

Author contributions statement

RS, AP and AS conceived the project. RS performed the analysis and wrote the manuscript under the guidance of AP and AS. GR developed the web-server and TT provided computational resources.

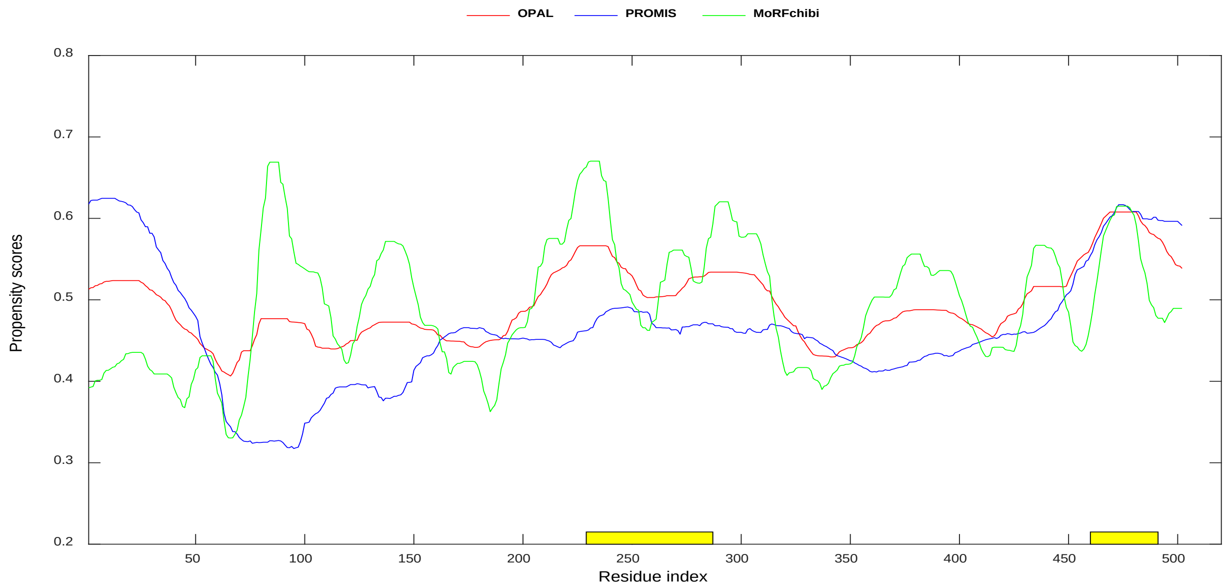


Fig.6: **Propensity scores of protein P42768**. Propensity scores are given by OPAL in red, PROMIS in blue and MoRFchibi in green. The two verified sections of MoRFs are marked in yellow color.

References

Cheng, Y., *et al.* (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments, *Biochemistry*, **46**, 13468–13477.

Disfani, F.M., *et al.* (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, *Bioinformatics*, **28**, i75–i83.

Dosztányi, Z., Mészáros, B. and Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins, *Bioinformatics*, **25**, 2745-2746.

Dyson, H.J. and Wright, E.P. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol*, **6**, 197-208.

Edwards, R.J., Davey, N.E. and Shields, D.C. (2007) SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins, *PLoS ONE*, **2**, e967.

Hamelryck, T. (2005) An amino acid has two sides: A new 2D measure provides a different view of solvent exposure, *Proteins: Structure, Function, and Bioinformatics*, **59**, 38-48.

Heffernan, R., *et al.* (2016) Highly Accurate Sequence-based Prediction of Half-Sphere Exposures of Amino Acid Residues in Proteins, *Bioinformatics*, **32**, 843-849

Heffernan, R., *et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, *Scientific Report*, **5**, 11476.

Kavianpour, H. and Vasighi, M. (2017) Structural classification of proteins using texture descriptors extracted from the cellular automata image, *Amino Acids*, **49**, 261-271.

Lee, R.V.D., *et al.* (2014) Classification of Intrinsically Disordered Regions and Proteins, *Chemical Reviews*, **114**, 6589-6631.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.

Liu, J., *et al.* (2006) Intrinsic Disorder in Transcription Factors, *Biochemistry*, **45**, 6873-6888.

Lyons, J., *et al.* (2015) Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models, *IEEE Transaction on Nanabioscience*, **14**, 761-772.

Malhis, N. and Gsponer, J. (2015) Computational identification of MoRFs in protein sequences, *Bioinformatics*, **31**, 1738–1744.

Malhis, N., Jacobson, M. and Gsponer, J. (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences, *Nucleic Acids Research*, **44**, W488–W493.

Malhis, N., *et al.* (2015) Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule, *PLoS ONE*, **10**, e0141603.

Mohan, A., *et al.* (2006) Analysis of Molecular Recognition Features (MoRFs), *Journal of Molecular Biology*, **362**, 1043-1059.

Mousavian, Z., *et al.* (2016) Drug–target interaction prediction from PSSM based evolutionary information, *Journal of Pharmacological and Toxicological Methods*, **78**, 42-51.

Oldfield, C.J., *et al.* (2005) Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements, *Biochemistry*, **44**, 12454-12470.

Peng, L., *et al.* (2017) Screening drug-target interactions with positive-unlabeled learning, *Scientific Reports*, **7**, 8087.

Sharma, A., *et al.* (2013) A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, *Theoretical Biology*, **320**, 41-46.

Sharma, A., *et al.* (2013) A Strategy to Select Suitable Physicochemical Attributes of Amino Acids for Protein Fold Recognition, *BMC Bioinformatics*, **14**, 1-11.

Sharma, R., *et al.* (2015) Predict Gram-positive and Gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into Chou's general PseAAC, *IEEE transactions on nanobioscience*, **14**, 915-926.

Sharma, R., *et al.* (2016) Predicting MoRFs in Protein Sequences using HMM Profiles, *BMC Bioinformatics*, **17 Suppl X, S14**.

Tompa, T. (2011) Unstructural biology coming of age, *Curr. Opin. Struct. Biol*, **2011**, 419–425.

Uversky, V. (2014) Introduction to Intrinsically Disordered Proteins (IDPs), *Chemical Reviews*, **114**, 6557-6560.

Vacic, V., *et al.* (2007) Characterization of Molecular Recognition Features, MoRFs, and Their Binding Partners, *Journal of Proteome Research*, **6**, 2351-2366.

Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation, *Nature Reviews:molecular cell biology*, **16**, 18-29.

Xia, J., *et al.* (2017) An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier, *Bioinformatics*, **33**, 863-870.

Yang, Y., *et al.* (2017) SPIDER2: A package to predict sccondary structure, accessible surface area and main-chain torsional angles by deep neural networks, *Methods Mol Biol*, **1484**, 55-63.