

## Technical Brief

### OPAL+: Length-specific MoRF prediction in intrinsically disordered protein sequences

Ronesh Sharma<sup>1,2, §</sup>, Alok Sharma<sup>1,3,4,5, §</sup>, Gaurav Raicar<sup>1</sup>, Tatsuhiko Tsunoda<sup>3,4,6</sup> and Ashwini Patil<sup>7, \*</sup>

<sup>1</sup>School of Engineering and Physics, The University of the South Pacific, Suva, Fiji

<sup>2</sup>School of Electrical and Electronics Engineering, Fiji National University, Suva, Fiji

<sup>3</sup>Laboratory of Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

<sup>4</sup>Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

<sup>5</sup>Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia

<sup>6</sup>CREST, JST, Tokyo 113-8510, Japan

<sup>7</sup>Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan.

\* To whom correspondence should be addressed.

E-mail: [ashwini@hgc.jp](mailto:ashwini@hgc.jp)

§ Equal contribution

Keywords: Intrinsically disordered proteins; molecular recognition features, support vector machines;

---

## **Abstract**

Intrinsically disordered proteins (IDPs) contain long unstructured regions which play an important role in their function. These intrinsically disordered regions (IDRs) participate in binding events through regions called molecular recognition features (MoRFs). Computational prediction of MoRFs helps identify the potentially functional regions in IDRs. In this study, we present OPAL+, a novel MoRF predictor. OPAL+ uses separate models to predict MoRFs of varying lengths along with incorporating the hidden Markov model (HMM) profiles and physicochemical properties of MoRFs and their flanking regions. Together, these features help OPAL+ achieve a marginal performance improvement of 0.4% - 0.7% over its predecessor for diverse MoRF test sets. This performance improvement comes at the expense of increased run time as a result of the requirement of HMM profiles. OPAL+ is available for download at <https://github.com/roneshsharma/OPAL-plus/wiki/OPAL-plus-Download>.

Intrinsically disordered proteins (IDPs) are proteins having large regions without a stable three-dimensional structure under physiological conditions. These proteins are frequently found and perform important functional roles <sup>[1][2]</sup>. The function of IDPs is often manifested through their binding to other ordered <sup>[3]</sup> and disordered <sup>[4]</sup> proteins. Binding of intrinsically disordered regions (IDRs) to ordered proteins is mediated by molecular recognition features (MoRFs). MoRFs are short regions of 5-25 residues within IDRs that undergo disorder-to-order transition on binding a partner protein <sup>[5]</sup>. Several computational methods are available to predict IDRs and their associated features <sup>[6, 7]</sup>, including MoRFs <sup>[8, 9-14]</sup>. Some of the available MoRF predictors include ANCHOR <sup>[9]</sup>, MoRFpred <sup>[10]</sup>, MoRFchibi <sup>[11]</sup>, MoRFpred-plus <sup>[13]</sup>, MoRFchibi-light <sup>[12]</sup>, MoRFchibi-web <sup>[12, 15]</sup> and OPAL <sup>[14]</sup>. ANCHOR uses the properties of binding regions located in disordered protein sequences to identify segments of protein regions that do not form sufficient interactions to fold on their own, but are likely to gain energy by interacting with globular proteins. On the other hand, MoRFpred uses the features generated from physicochemical properties of disordered regions such as relative solvent accessibility derived from Real-SPINE3 <sup>[16]</sup>, position specific scoring matrices computed by PSI-BLAST <sup>[17]</sup>, flexibility (B-factor) estimated by PROFbval and predictions of five other disorder predictors. MoRFchibi <sup>[11]</sup> employs two SVM models with local physicochemical properties of amino acids as feature vectors to identify MoRFs. These models use amino acid similarity, composition and contrast information between MoRF and non-MoRF regions to predict MoRFs. The scores of MoRFchibi are processed using Bayes rule. MoRFpred-plus <sup>[13]</sup> is trained using hidden Markov model (HMM) profiles and local physicochemical properties of disordered protein sequences. MoRFpred-plus targets the properties of upstream/downstream flank residues of a query residue to predict MoRFs. More accurate predictors like MoRFchibi-web <sup>[12]</sup> and OPAL <sup>[14]</sup> are constructed by combining multiple component predictors. Specifically, MoRFchibi-web uses scores of MoRFchibi, Espirtz <sup>[18]</sup>, a disorder predictor, and conservation information derived from PSI-BLAST <sup>[17]</sup>. On the other hand, OPAL

---

is an ensemble of two predictors, MoRFchibi <sup>[11]</sup> and PROMIS <sup>[14]</sup>. PROMIS incorporates structural information of disordered protein sequences into OPAL <sup>[14]</sup>. Both, MoRFchibi-web and OPAL use a single model to predict MoRFs of varying lengths.

In this study, we present a method to improve the performance of OPAL by developing multiple models specific for MoRFs of different lengths. OPAL+ uses four support vector machine (SVM) models, each trained using MoRFs of different lengths. It also utilizes evolutionary information of the IDRs in the form of HMM profiles obtained from another MoRF predictor, MoRFpred-plus <sup>[13]</sup>. Finally, OPAL+ combines the scores of the length-specific MoRF predictor and MoRFpred-plus with those from MoRFchibi to obtain a final MoRF score for each residue in the query disordered region.

The training and test sets used to develop OPAL+ were obtained from Disfani et al., <sup>[10]</sup> and Malhis et al., <sup>[15]</sup>. These sets are called TRAIN, TEST, TEST464 and EXP53, and are described in Table S1. To assemble TRAIN, TEST and TEST464 sets, sequences were collected and filtered from protein data bank (PDB) depositions made before April 2008. The EXP53 set was assembled by combining sequences from three different studies <sup>[7, 10, 19]</sup> and includes sequences containing experimentally validated MoRFs that are disordered in isolation. These sets were previously used to develop predictors including OPAL <sup>[14]</sup>, MoRFchibi-web <sup>[12]</sup>, MoRFpred-plus <sup>[13]</sup>, MoRFchibi <sup>[11]</sup> and MoRFpred <sup>[10]</sup>. Altogether, there were 938 sequences in training and test sets. The details of the number of MoRF and non-MoRF residues in each set is given in Table S1. The distribution of the MoRF lengths within the training and some test sets are shown in Figure S1. The sequences in the TEST and TEST464 sets contain MoRFs from lengths 5 to 25 residues. The TRAIN and TEST sets primarily contain MoRFs of length 7 to 11 residues, while the EXP53 set contains several MoRFs as long as 30 residues or more. Given the non-uniform distribution of MoRFs over different lengths, we divided the EXP53 dataset according to MoRF length, i.e., EXP53SHORT containing MoRFs of length up to

30 residues, EXP53LONG containing MoRFs of length greater than 30 amino acids and EXP53ALL containing all MoRFs.

An overview of the length-specific MoRF prediction scheme is given in Figure 1. Four different models were constructed to predict MoRFs in disordered protein sequences, each trained to target different MoRF lengths. We partitioned MoRFs into 4 groups based on their lengths - from 5 to 9 residues, 10 to 14 residues, 15 to 19 residues and 20 to 24 residues. Table S1 gives the number of MoRFs in the training and test sets for each group. In the training step, features were computed from MoRFs and non-MoRFs. Since the TRAIN set has a single MoRF region and the number of non-MoRF residues is greater than the number of MoRF residues, balanced sampling is required. To enable balanced sampling, we extracted upstream/downstream flanking amino acid residues along with the MoRF region as a positive sample. We then extracted the same size of the negative sample from a non-MoRF region (See Supporting information). For each length-specific model, we computed bigram feature vectors<sup>[20]</sup> from each MoRF and non-MoRF group using the BigramMoRF method described in Sharma *et al.*,<sup>[14]</sup> and the structural attributes predicted<sup>[21]</sup> using Spider2<sup>[21]</sup>. Bigram feature representation is based on profile bigrams<sup>[20]</sup>, where the feature vector is obtained by counting the bigram frequencies from the evolutionary profiles of a protein sequence. However, in this study, instead of using evolutionary profiles we used structural attributes to compute bigram features (See Supporting information). The Spider2 structural attributes include secondary structure (SS), containing the structural description of proteins such as helix, sheet and coil; accessible surface area (ASA), which measures the exposure level of amino acids to solvent in the proteins; backbone angles, which include the dihedral angles of proteins such as Phi, Psi, Theta and Tau; half-sphere exposure (HSE), which gives the number of *C* $\alpha$  atoms in the upper (HSEu) and lower spheres (HSEd) of a residue. Each length-specific model is trained independently as illustrated in Figure S2. During the test phase,

---

all four length-specific models are used for scoring and the output scores are combined by taking the minimum score as the output score (See Supporting information).

We selected SVM classifier with RBFkernel. The C and Gamma values of the kernel were selected as 1000 and 0.0038, respectively <sup>[14]</sup>. To select the structural attributes for each model in Figure S2, we performed successive feature selection scheme in the forward direction <sup>[22]</sup> and observed the AUC performance measure to select the highly ranked attributes for each model. Table 1 shows the attributes selected for each length-specific MoRF prediction model. As the results indicate, different features are informative at different MoRF lengths. While the contact HSE attributes, dihedral angle and theta are important in all models, the information given by the helical or strand state of residues in the MoRFs varies with length. This suggests that MoRFs of different lengths bind to partner proteins in distinct ways.

To further improve the model performance, we combined MoRFpred-plus and MoRFchibi with the proposed model, since they were constructed using complementary features and learning algorithms. MoRFpred-plus targets the properties of disordered regions flanking the MoRF residue to identify MoRFs <sup>[13]</sup> whereas MoRFchibi utilizes the similarity, composition and contrast information of MoRFs and non-MoRFs together with different SVM kernels <sup>[11]</sup> to predict MoRFs. To calculate the scores for each residue, we applied the common averaging principle where all scores are added and divided by the number of models used (Figure S3).

The final score calculation was performed for each residue by taking a window of scores consisting of the residue score itself and the score of its  $z$  flanking residues on either side <sup>[14]</sup> (Figure S4, see Supporting information). Figure S5 shows the AUCs for varying the value of flank size  $z$  from 1 to 30 for the four proposed models presented in Figure S2. Varying the value of residue flank size,  $z$ , and observing the AUC, we selected  $z$  equal to 25 to process the output scores of the four proposed models (model 1, model 2, model 3 and model 4). In addition, to obtain average performance from

the combined scheme, we processed MoRFchibi scores with  $z$  equal to 15, MoRFpred-plus scores with  $z$  equal to 4 and the final combined model scores with  $z$  equal to 8.

The final AUCs are listed in Table 2. As seen, OPAL+ performs well across all the test sets. Compared to the benchmarked predictors, MoRFchibi-web and OPAL, the performance improvement of 2.2% and 0.7% is obtained for the TEST set, 1.5% and 0.4% for the TEST464 set, and 1.1% and 0.2% for EXP53ALL, respectively. To predict long MoRFs (EXP53LONG), OPAL+, OPAL and MoRFchibi-web achieved AUCs of 0.822, 0.822 and 0.758, respectively, while to predict short MoRFs (EXP53SHORT) AUCs of 0.876, 0.870 and 0.886 were observed (Figure S6 to S9). This shows that OPAL+ provides more accurate prediction for long MoRFs. On the other hand, MoRFchibi-web produces better prediction results for short MoRFs, though OPAL+ improves on the performance of OPAL in this case as well. To analyze the performance of OPAL+, we calculate performance measures including precision, F-measure, accuracy and false positive rate (FPR) for different values of TPR as shown in Table S2 to S4. Thus, it is observed that OPAL+ achieves a minor increase in performance measures for certain TPR values, while performing similar to OPAL for other TPR values. Figures S10-S13 show specific examples of MoRF prediction where OPAL+ outperforms OPAL.

To determine the statistical significance of the difference in the prediction performance of OPAL+ and OPAL, we used the paired t-test with 5% significance level. We computed the paired t-test for different output threshold probabilities of the classifier and averaged the results. The statistical significance of the difference between the performance of OPAL+ and OPAL for the test sets TEST464, EXP53ALL and EXP53SHORT are 0.068, 0.082 and 0.050, respectively. Though the performance of OPAL+ shows a statistically significant improvement over OPAL for the EXP53SHORT test set, it is above the significance levels for TEST464 and EXP53ALL. Thus, we conclude that the predic-

---

tion accuracies obtained by OPAL+ using the length-specific scheme for MoRF prediction are very similar to those of OPAL.

To test the efficiency of the state-of-the-art predictors, Table S5 shows the comparison of AUCs, prediction speed in residues/minute (r/m) and multiple sequence alignment information. The results show that predictors like ANCHOR and MoRFchibi are fast in terms of prediction speed, whereas the remaining predictors requiring multiple sequence alignment are comparatively slow. MoRFchibi-web, OPAL and OPAL+ are constructed by combining multiple component predictors, therefore, their architecture is complex and prediction speed is much lower compared to the other predictors.

In summary, we present an improved MoRF predictor, OPAL+, which is unique in its use of separate models to predict MoRFs of varying lengths and evolutionary profiles of disordered regions. It is available for download at <https://github.com/roneshsharma/OPAL-plus/wiki/OPAL-plus-Download>. OPAL+ shows a marginal performance improvement over OPAL. Additional improvement in MoRF prediction will require exploring novel features of protein sequences with model training strategies.

## References

- [1] Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky, L. Kurgan, *Cellular and Molecular Life Sciences* 2015, 72, 137.
- [2] P. Lieutaud, F. Ferron, A. V. Uversky, L. Kurgan, V. N. Uversky, S. Longhi, *Intrinsically Disord Proteins* 2016, 4.
- [3] K. Sugase, H. J. Dyson, P. E. Wright, *Nature* 2007, 447, 1021.
- [4] A. Borgia, M. B. Borgia, K. Bugge, V. M. Kissling, P. O. Heidarsson, C. B. Fernandes, A. Sottini, A. Soranno, K. J. Buholzer, D. Nettels, B. B. Kragelund, R. B. Best, B. Schuler, *Nature* 2018, 555, 61.
- [5] C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, A. K. Dunker, *Biochemistry* 2005, 44, 12454; A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A.

K. Dunker, V. N. Uversky, *Journal of Molecular Biology* 2006, 362, 1043; J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, A. K. Dunker, *Biochemistry* 2006, 45, 6873; V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, A. K. Dunker, *Journal of Proteome Research* 2007, 6, 2351; J. Yan, A. K. Dunker, V. N. Uversky, L. Kurgan, *Mol Biosyst* 2016, 12, 697.

[6] Z. Peng, L. Kurgan, *Nucleic Acids Res* 2015, 43, 24; F. Meng, V. N. Uversky, L. Kurgan, *Cell Mol Life Sci* 2017, 74, 3069; B. Mészáros, G. Erdős, Z. Dosztányi, *Nucleic Acids Research* 2018, 46, W329.

[7] D. T. Jones, D. Cozzetto, *Bioinformatics* 2015, 31, 857.

[8] Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky, A. K. Dunker, *Biochemistry* 2007, 46, 13468; B. Xue, A. K. Dunker, V. N. Uversky, *Int J Mol Sci* 2010, 11, 3725; C. Fang, T. Noguchi, D. Tominaga, H. Yamana, *BMC Bioinformatics* 2013, 14, 1471.

[9] Z. Dosztányi, B. Mészáros, I. Simon, *Bioinformatics* 2009, 25, 2745.

[10] F. M. Disfani, W. L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky, L. Kurgan, *Bioinformatics* 2012, 28, i75.

[11] N. Malhis, J. Gsponer, *Bioinformatics* 2015, 31, 1738.

[12] N. Malhis, M. Jacobson, J. Gsponer, *Nucleic Acids Research* 2016, 44, W488.

[13] R. Sharma, M. Bayarjargal, T. Tsunoda, A. Patil, A. Sharma, *Journal of Theoretical Biology* 2018, 437, 9.

[14] R. Sharma, G. Raicar, T. Tsunoda, A. Patil, A. Sharma, *Bioinformatics* 2018, 34, 1850.

[15] N. Malhis, E. T. C. Wong, R. Nassar, J. Gsponer, *PLoS ONE* 2015, 10, e0141603.

[16] E. Faraggi, B. Xue, Y. Zhou, *Proteins* 2009, 74, 857.

[17] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Research* 1997, 17, 3389.

[18] I. Walsh, A. J. M. Martin, T. D. Domenico, S. C. E. Tosatto, *Bioinformatics* 2012, 28, 503.

[19] B. Meszaros, I. Simon, Z. Dosztanyi, *PLoS Computational Biology* 2009, 5, 1.

[20] A. Sharma, J. Lyons, A. Dehzangi, K. K. Paliwai, *Theoretical Biology* 2013, 320, 41.

---

[21] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, *Methods Mol Biol* 2017, 1484, 55.

[22] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, *BMC Bioinformatics* 2013, 14, 1.

### **Supporting Information**

Supporting Information is available from the Wiley Online Library.

### **Acknowledgements**

This work was supported 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).

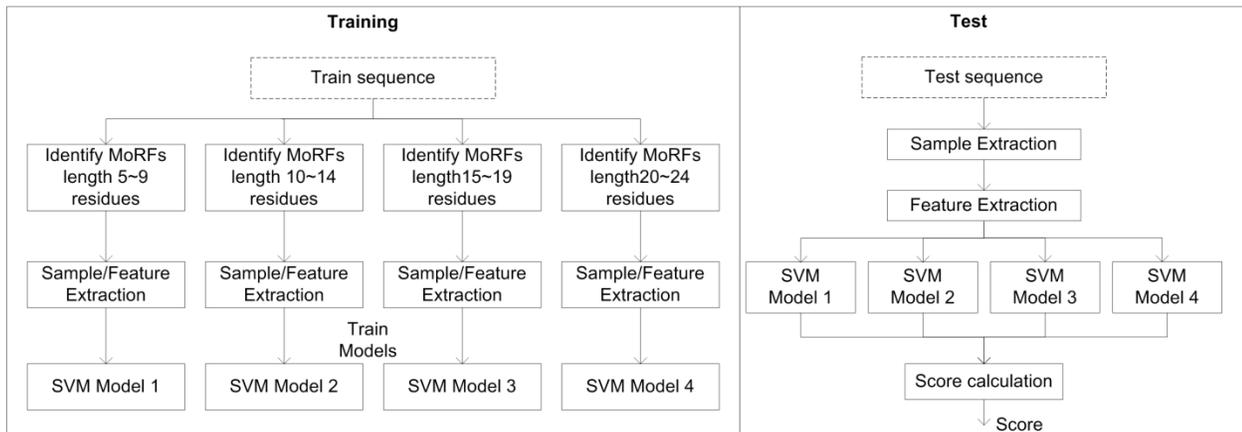
### **Conflict of Interest statement:**

The authors declare no conflict of interest.

### **Authors' contributions**

RS performed the analysis and wrote the manuscript under the guidance of AP and AS. GR helped in algorithm development. TT provided computational resources.

## Figures



**Figure 1:** Overview of the proposed MoRF prediction method.

---

## Tables

**Table 1:** Selected attributes for each of the proposed models

Models for MoRF lengths:	Attributes
5 to 9 residues	CN attribute from HSE $\alpha$ group Theta attribute from backbone dihedral angles Strand state probability attribute from SS group HSEu attribute from HSE $\beta$ group
10 to 14 residues	CN attribute from HSE $\alpha$ group ASA attribute HSEd attribute from HSE $\beta$ group Theta attribute from backbone dihedral angles
15 to 19 residues	CN attribute from HSE $\alpha$ group ASA attribute Helix state probability attribute from SS group HSEd attribute from HSE $\alpha$ group Strand state probability attribute from SS group
20 to 24 residues	CN attribute from HSE $\alpha$ group Theta attribute from backbone dihedral angles

CN: number of contact residues; HSE: half-sphere exposure, HSEu: HSE in the upper sphere, HSEd: HSE in the lower sphere SS: secondary structure and ASA: accessible surface area. For each of the model, attributes are listed from higher to lower rank.

**Table 2:** AUCs of MoRF predictors and models

Predictors/models	TEST	TEST464	EXP53 ALL	EXP53 LONG	EXP53 SHORT
ANCHOR	0.600	0.605	0.615	0.586	0.683
MoRFPred	0.673	0.675	0.620	0.598	0.673
MoRFchibi	0.740	0.743	0.712	0.679	0.790
MoRFPred-plus	0.755	0.724	0.712	0.670	0.821
PROMIS	0.791	0.788	0.818	0.815	0.823
MoRFchibi-light	0.775	0.777	0.799	0.770	0.869
MoRFchibi-web	0.800	0.805	0.797	0.758	0.886
OPAL	0.815	0.816	0.836	0.822	0.870
<b>Length-specific model</b>	<b>0.781</b>	<b>0.777</b>	<b>0.812</b>	<b>0.805</b>	<b>0.830</b>
<b>Length-specific model + MoRFPred-plus</b>	<b>0.813</b>	<b>0.810</b>	<b>0.821</b>	<b>0.807</b>	<b>0.856</b>
<b>OPAL<sup>+</sup> (Length-specific model + MoRFPred-plus + MoRFchibi)</b>	<b>0.822</b>	<b>0.820</b>	<b>0.838</b>	<b>0.822</b>	<b>0.876</b>