Supplementary Text S1 for OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences

Ronesh Sharma2, Gaurav Raicar2, Tatsuhiko Tsunoda1,3,4, Ashwini Patil5,+,\*, and Alok Sharma1,3,4,6,+,\*

1RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

2The University of the South Pacific, Fiji

3Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

4CREST, JST, Tokyo 113-8510, Japan

5Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan

6Griffith University, Australia

\* [ashwini@hgc.jp](mailto:ashwini@hgc.jp), [alok.sharma@griffith.edu.au](mailto:alok.sharma@griffith.edu.au)

+Last authors

In this supplementary text, we describe the detailed process of developing the OPAL model. We first illustrate the training steps of the OPAL model and then we outline the test steps. The output of this model is a score for a query sequence. Moreover, we also provide an example with the help of a toy protein sequence to demonstrate the procedure of scoring a query sequence.

**Training steps**

In the training set, we extracted features from MoRFs and non-MoRFs. Let MoRF (with flanks) be depicted as and non-MoRF (with flanks) be denoted as. Segments and are from a protein sequence of the training set. Assuming a protein sequence is given as

= ( (1)

where is the *j*-th amino acid in the sequence, is the total number of protein sequences in the training set and is the length of protein sequence For a protein sequence, suppose and MoRF is present between and . For this protein , is illustrated in Fig.s1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |

MoRFs

Left flanks of 20 amino acids Right flanks of 20 amino acids

Fig.s1:

In addition, a non-MoRF is illustrated in Fig.s2. Here segment is from amino acids .

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |

non-MoRFs

Left flanks of 20 amino acids Right flanks of 20 amino acids

Fig.s2:

If a MoRF is present at the beginning or end of the protein sequence, zeros are padded to create flanks of length 20. We selected flank length to be 20 because at this length the AUC obtained was highest on test data. In our training set we had 421 protein sequences, this in turn gave 421 representing MoRFs. A MoRF can have a variable length between 5 and 25 residues. Positive samples representing MoRF residues are extracted from and negative samples representing non-MoRF residues are extracted from In the following subsections, we describe the procedure of extracting positive and negative samples from MoRF and non-MoRF regions. We described the procedure for 2 methods in this supplement. They are referred to as StructMoRF and BigramMoRF.

***StructMoRF***

In this method, we use window of size 41 to extract sample to represent each MoRF residue. The center of the window is placed on the MoRF residue with 20 neighboring amino acids as flanks on both sides, upstream and downstream of the residue. For a MoRF residue, sample is defined as

(2)

where is the MoRF residue, i.e. in Fig.s3 (a), =30 and =11.Using equation (2), positive samples extracted from a segment is interpreted as

, (5<<25) (3)

where is the MoRF length. Fig.s3 (a) shows the schematic illustration to extract samples from segment defined in Fig.s1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |

a

*l*

RF

RF

RF

RF

RF

RF

RF

RF

RF

Structural matrix of size 41 by 3

b

, , ,

RF

RF

RF

RFASAAS

*CN , HSEu, HSEd* Structural attributes

RF

RFASAAS

41 amino acids ×3 structural attributes = 123 features

Fig.s3: a) Schematic illustration of extracting samples from segment, b) Illustration of structural attributes used.

The number of positive samples extracted per segment is equal to the number of MoRF residues per segment. To obtain feature vector, structural attributes are considered and here the attribute values are treated as features as depicted in Fig.s3 (b). Suppose 3 structural attributes are considered, the feature vector for a sample can be interpreted as:

(4)

where is the element of structural matrix of size 41 by . Similarly, for a negative sample extracted from segment , feature vector is interpreted as , where is the element of structural matrix of size 41 by .

***BigramMoRF***

In BigramMoRF method, we take the entire MoRF region of length (as depicted in Fig.s3 (a)) to extract bigram features as:

(5)

where is the element of structural matrix of size 40+ by (similar to Fig.s3 (b), however, structural matrix size is 40+ by ). Equation (5) returns a matrix of 9 occurrence frequencies for 9 bigram interactions, thus we define this matrix as the occurrence matrix and it can be interpreted as a 9-element feature vector defined for a sample as:

(6)

Similarly, for a negative sample extracted from segment , feature vector is interpreted as .

**Test steps**

To score each residue in query protein sequence, below we describe the steps involved with StructMoRF and BigramMoRF methods.

***StructMoRF***

To score, in this method we use window of size 41 to extract sample for each query residue. For a query residue, sample is defined as

(7)

where is the query residue in the query sequence and =1,2,.... Using equation (7), samples for a query sequence of length is interpreted as

(8)

Fig.s4 shows the schematic illustration of extracting query samples from a query sequence. Feature vector is extracted from the sample and is used for prediction. For query residues at the beginning or end of the sequences, zeros are padded to make length of flanks to 20.

Query Sequence

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

RF

RF

RF

0,0, ,0,

RF

RF

RF

RF

RF

RF

RF

RF

RF

Fig.s4: Schematic illustration of extracting samples to score query sequence. is the *j*-th amino acid in the query sequence and refers to the length of query protein sequence.

***BigramMoRF***

To score using BigramMoRF method, we use window of size 40+ to extract samples from a query sequences. For a query sequence, sample is defined as

(9)

where is the query residue in the query sequence and =,,3,...,. Using equation (9), samples for a query sequence of length is interpreted as

(10)

where =6,7,….,30. Fig.s6 shows the schematic illustration of extracting samples from a query sequence of length . Equation (11) outlines the samples utilized for scoring each of the residues.

(11)

where =1,2,….,. Bigram feature vector is computed from the sample and is used for scoring. Varying from 6 to 30, we end up with 450 (6+7+ - - - +29+30) samples for each residue, except for the residues at the beginning and end of the query sequence, this in turn gives 450 predicted scores. Thus maximum score is taken for scoring.

Query Sequence

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

jcccccbkoojbjlkbkljbklbklb

jcccccbkoojbjlkbkljbklbklb

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

jcccccbkoojbjlkbkljbklbklb

jcccccbkoojbjlkbkljbklbklb

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

jcccccbkoojbjlkbkljbklbklb

jcccccbkoojbjlkbkljbklbklb

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

jcccccbkoojbjlkbkljbklbklb

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

jcccccbkoojbjlkbkljbklbklb

Fig.s6: Schematic illustration to extract samples from a query sequence of length.

***An illustration to predict query residues using an example of a toy protein problem.***

To illustrate, let us consider a toy query protein sequence as of length 6. Note that this toy protein is made up of three amino acids A, N and R. Suppose number of structural attributes used is 3 (*CN, HSEu and HSEd*), Table s1 shows the structural matrix for this protein.

Table s1: Structural matrix of protein

|  |  |  |  |
| --- | --- | --- | --- |
| Amino acids | CN | HSEu | HSEd |
| A | 0.2 | 0.9 | 0.1 |
| R | 0.1 | 0.7 | 0.1 |
| N | 0.2 | 0.8 | 0.2 |
| R | 0.3 | 0.1 | 0.2 |
| A | 0.2 | 0.9 | 0.1 |
| N | 0.4 | 0.2 | 0.2 |

To score using StructMoRF, let us consider a window of size 3. The samples extracted for the residues in terms of amino acids and structural matrix with feature vectors computed is shown in Table s2.

Table s2: Samples and feature vectors for the residues using StructMoRF method

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Query residue | Samples | | | | | Feature vectors |
|  | Amino acids | CN | HSEu | HSEd |  |
| A |  | 0 | 0 | 0 | 0 |  |
| A | 0.2 | 0.9 | 0.1 | 0, 0.2, 0.1, 0, 0.9, 0.7, 0, 0.1 ,0.1 |
| R | 0.1 | 0.7 | 0.1 |  |
| R |  | A | 0.2 | 0.9 | 0.1 |  |
| R | 0.1 | 0.7 | 0.1 | 0.2, 0.1, 0.2, 0.9, 0.7, 0.8, 0.1, 0.1, 0.2 |
| N | 0.2 | 0.8 | 0.2 |  |
| N |  | R | 0.1 | 0.7 | 0.1 |  |
| N | 0.2 | 0.8 | 0.2 | 0.1, 0.2, 0.3, 0.7, 0.8, 0.1, 0.1 0.2, 0.2 |
| R | 0.3 | 0.1 | 0.2 |  |
| R |  | N | 0.2 | 0.8 | 0.2 |  |
| R | 0.3 | 0.1 | 0.2 | 0.2, 0.3, 0.2, 0.8, 0.1, 0.9, 0.2, 0.2, 0.1 |
| A | 0.2 | 0.9 | 0.1 |  |
| A |  | R | 0.3 | 0.1 | 0.2 |  |
| A | 0.2 | 0.9 | 0.1 | 0.3,0.2, 0.4, 0.1, 0.9, 0.2, 0.2, 0.1, 0.2 |
| N | 0.4 | 0.2 | 0.2 |  |
| N |  | A | 0.2 | 0.9 | 0.1 |  |
| N | 0.4 | 0.2 | 0.2 | 0.2, 0.4, 0, 0.9, 0.2, 0, 0.1, 0.2, 0 |
| 0 | 0 | 0 | 0 |  |

To score using BigramMoRF, let us consider a window size of +2 and suppose =2. The samples used for scoring the query residues in terms of amino acids and structural matrices with feature vectors computed is shown in Table s3 and Table s4.

Table s3: Samples extracted and feature vectors computed using BigramMoRF method

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Samples | | | | | Feature vectors |
| No | Amino acids | CN | HSEu | HSEd |
|  | 0 | 0 | 0 | 0 | 0.010, 0.057, 0.008, 0.055, 0.298, 0.038, 0.010, 0.057, 0.008 |
| A | 0.2 | 0.9 | 0.1 |
| R | 0.1 | 0.7 | 0.1 |
| N | 0.2 | 0.8 | 0.2 |
|  | A | 0.2 | 0.9 | 0.1 | 0.025, 0.118, 0.023, 0.060, 0.318, 0.043, 0.020, 0.098, 0.018 |
| R | 0.1 | 0.7 | 0.1 |
| N | 0.2 | 0.8 | 0.2 |
| R | 0.3 | 0.1 | 0.2 |
|  | R | 0.1 | 0.7 | 0.1 | 0.035, 0.100, 0.030, 0.093, 0.182, 0.070, 0.0230 , 0.078, 0.020 |
| N | 0.2 | 0.8 | 0.2 |
| R | 0.3 | 0.1 | 0.2 |
| A | 0.2 | 0.9 | 0.1 |
|  | N | 0.2 | 0.8 | 0.2 | 0.050, 0.155, 0.035, 0.083, 0.088, 0.055, 0.028, 0.088, 0.020 |
|  | R | 0.3 | 0.1 | 0.2 |
|  | A | 0.2 | 0.9 | 0.1 |
|  | N | 0.4 | 0.2 | 0.2 |
|  | R | 0.3 | 0.1 | 0.2 | 0.035, 0.095, 0.020, 0.078, 0.068, 0.050, 0.018, 0.048, 0.010 |
|  | S | 0.2 | 0.9 | 0.1 |
|  | N | 0.4 | 0.2 | 0.2 |
|  | 0 | 0 | 0 | 0 |

Table s4: Samples used for scoring each residue using BigramMoRF method

|  |  |  |
| --- | --- | --- |
| -th query position | Query residue | Samples |
|  | A |  |
|  | R | , |
|  | N | , |
|  | R | , |
|  | A | , |
|  | N |  |