SubFeat: Feature Subspacing Ensemble Classifier for Function Prediction of DNA, RNA, and Protein/Peptide Sequences

H. M. Fazlul Haque[a], Fariha Arifin[a], Rafsanjani Muhammod[b], and Swakkhar Shatabda[a]

[a]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

[b]Bioinformatics Research Lab, United International University, Dhaka, Bangladesh

Supplementary Material

SubFeat Version 1.0

Contents

1 Feature Description

We have taken two DNA's FASTA sequences as example. One is for positive (>Positive Sequence), and another is for negative (>Negative Sequence) example respectively.

Box 1: Sample FASTA file (File name: demoFASTAs.txt or demoFASTAs.fa)

```
>Positive Sequence
TCAGGGAGATGTGAGCCAGCTCACCATAAAAAAGCCG
>Negative Sequence
ATTGCGCGGTACAACTAAAAAACGCTGTTCCGATGGA
```

Box 2: Sample label file (File name: demoLabels.txt)

```
1
0
```

Important Definitions:

$$\mathbf{X} = \begin{cases} \{\text{A,C,G,T}\}, & \text{if the problem involves DNA sequences} \\ \{\text{A,C,G,U}\}, & \text{if the problem involves RNA sequences} \\ \{\text{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}\}, & \text{if the problem involves protein sequences} \end{cases}$$

$x_i \in \mathbf{X}$ where i specifies the position of x in some subsequence. Counts of such subsequences of varying lengths is regarded as features in our method.

$j \in \{1, 2, 3, ..., k\}$ where j specifies the number of gaps (don't care) in a subsequence.

1.1 Feature Subspace–1 (F_1)

When n=k, then the $\sum_{i=1}^{n} 4^i$ features will exist for DNA and RNA sequence; but $\sum_{i=1}^{n} 20^i$ features will exist for protein/peptide sequence.

When k=1, feature structure will be **X**.

When k=2, feature structure will be **X**, and **XX**.

When k=3, feature structure will be **X**, **XX**, and **XXX**.

For the **MonoMer Composition**, feature structure will be **X**.

For the **DiMer Composition**, feature structure will be **XX**.

For the **TriMer Composition**, feature structure will be **XXX**.

Described with appropriate examples:

When k=1 then only four (4) features will exist for DNA and RNA, but twenty (20) features will exist for protein. Features will be numbers of A, C, G and T/U of the whole sequence of DNA and RNA respectively.

When k=2 then only twenty (20) features will exist for DNA and RNA, but four hundred and twenty (420) features will exist for protein. Features will be numbers of A, C, G, T, AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT of the whole sequence of DNA respectively.

When k=3 then only eighty four (84) features will exist for DNA and RNA, but eight thousand four hundred and twenty (8,420) features will exist for protein. Features will be numbers of A, C, G, T, AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT, AAA, AAC, AAG, AAT, ACA, ACC, ACG, ACT, AGA, AGC, AGG, AGT, ATA, ATC, ATG, ATT, CAA, CAC, CAG, CAT, CCA, CCC, CCG, CCT, CGA, CGC, CGG, CGT, CTA, CTC, CTG, CTT, GAA, GAC, GAG, GAT, GCA, GCC, GCG, GCT, GGA, GGC, GGG, GGT, GTA, GTC, GTG, GTT, TAA, TAC, TAG, TAT, TCA, TCC, TCG, TCT, TGA, TGC, TGG, TGT, TTA, TTC, TTG, and TTT of the whole sequence of DNA respectively.

1.1.1 Generate dataset using F_1

For '>Positive Sequence': $\sum A$=13, $\sum C$=9, $\sum G$=10, $\sum T$=5, $\sum AA$=3, $\sum AC$=1, $\sum AG$=5, $\sum AT$=2 and so on upto three combination of ACGT; and '>Negative Sequence': $\sum A$=12, $\sum C$=8, $\sum G$=9, $\sum T$=8, $\sum AA$=4, $\sum AC$=3, $\sum AG$=0, $\sum AT$=2 and so on upto three combination of ACGT.

Box 3: Sample dataset using pseudoKNC (File name: fullDataset.csv)

13, 9, 10, 5, 3, 1, 5, 2, ..., 0, 0, 0, 0 1
12, 8, 9, 8, 4, 3, 0, 2, ..., 0, 1, 1, 0, 0

1.2 Feature Subspace–2 (F_2)

1.2.1 1-Gapped Di-Mono Composition

The number of [(4×4)×4] features will exist for DNA and RNA sequence; but [(20×20)×(20)] features will exist for protein; and feature structure will be **XX_X**.

1.2.2 Generate dataset using F_2

For '>Positive Sequence': $\sum AA\_A$=3, $\sum AA\_C$=1, $\sum AA\_G$=1, $\sum AA\_T$=0, and so on; and '>Negative Sequence': $\sum AA\_A$=3, $\sum AA\_C$=1, $\sum AA\_G$=1, $\sum AA\_T$=1, and so on.

Box 4: Sample dataset using diMonoKGap (File name: fullDataset.csv)

3, 1, 1, 0, 1, 0, 0, 0, ..., 1
3, 1, 1, 1, 2, 1, 0, 0, ..., 0

1.3 Feature Subspace–3 (F_3)

1.3.1 1-Gapped Mono-Di Composition

The number of [4×(4×4)] features will exist for DNA and RNA sequence; but [20×(20×20)] features will exist for protein; and feature structure will be **X_XX**.

Described with appropriate examples:

The only sixty four (64) features will exist for DNA and RNA, but eight thousand (8,000) features will exist for protein. Features will be numbers of A_AA, A_AC, A_AG, A_AT, A_CA, A_CC, A_CG, A_CT, A_GA, A_GC, A_GG, A_GT, A_TA, A_TC, A_TG, A_TT, C_AA, C_AC, C_AG, C_AT, C_CA, C_CC, C_CG, C_CT, C_GA, C_GC, C_GG, C_GT, C_TA, C_TC, C_TG, C_TT, G_AA, G_AC, G_AG, G_AT, G_CA, G_CC, G_CG, G_CT, G_GA, G_GC, G_GG, G_GT, G_TA, G_TC, G_TG, G_TT, T_AA, T_AC, T_AG, T_AT, T_CA, T_CC, T_CG, T_CT, T_GA, T_GC, T_GG, T_GT, T_TA, T_TC, T_TG, and T_TT of the whole sequence of DNA respectively.

1.3.2 Generate dataset using F_3

For '>Positive Sequence': $\sum A\_AA$=4, $\sum A\_AC$=0, $\sum A\_AG$=1, $\sum A\_AT$=1, and so on; and '>Negative Sequence': $\sum A\_AA$=4, $\sum A\_AC$=1, $\sum A\_AG$=0, $\sum A\_AT$=0, and so on.

Box 5: Sample dataset using '1-Gapped Mono-Di Composition'

4, 0, 1, 1, 1, 2, 0, 1, ..., 1
4, 1, 0, 0, 0, 0, 1, 1, ..., 0

2 Feature Calculation

Since we have used different types of datasets in this work, we extrated features which are sequence based. For protein dataset the number of fullspace feature is 24,420. For DNA and RNA datasets the number of fullspace feature is 212. A summary of feature subspace of protein, DNA and RNA data is given in Table 6 and Table 7.

Feature Subspace	Feature Type	No. of features
F_1	MonoMer Composition	20
	DiMer Composition	400
	TriMer Composition	8000
F_2	1-Gapped Di-Mono Composition	8000
F_3	1-Gapped Mono-Di Composition	8000

Table 6: Details of feature subspacing for protein/peptide sequence.

Feature Subspace	Feature Type	No. of features
F_1	MonoMer Composition	4
	DiMer Composition	16
	TriMer Composition	64
F_2	1-Gapped Di-Mono Composition	64
F_3	1-Gapped Mono-Di Composition	64

Table 7: Details of feature subspacing for DNA and RNA sequence.

References

[1] Md Siddiqur Rahman, Usma Aktar, Md Rafsan Jani, and Swakkhar Shatabda. ipromoter-fsen: Identification of bacterial σ70 promoter sequences using feature subspace based ensemble classifier. *Genomics*, 2018.

[2] Md Siddiqur Rahman, Usma Aktar, Md Rafsan Jani, and Swakkhar Shatabda. ipro70-fmwin: identifying sigma70 promoters using multiple windowing and minimal features. *Molecular Genetics and Genomics*, 294(1):69–84, 2019.

[3] Rafsanjani Muhammod, Sajid Ahmed, Dewan Md Farid, Swakkhar Shatabda, Alok Sharma, and Abdollah Dehzangi. Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences. *Bioinformatics*, 35(19):3831–3833, 2019.

[4] Md Rafsan Jani, Md Toha Khan Mozlish, Sajid Ahmed, Niger Sultana Tahniat, Dewan Md Farid, and Swakkhar Shatabda. irecspot-ef: Effective sequence based features for recombination hotspot prediction. *Computers in biology and medicine*, 103:17–23, 2018.

[5] Shahana Yasmin Chowdhury, Swakkhar Shatabda, and Abdollah Dehzangi. Idnaprotes: Identification of dna-binding proteins using evolutionary and structural features. *Scientific Reports*, 7(1):14938, 2017.

[6] Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.

[7] Bin Liu, Jinghao Xu, Xun Lan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, and Kuo-Chen Chou. idna-prot— dis: identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PloS one*, 9(9):e106691, 2014.

[8] Wei Chen, Pengmian Feng, Hui Ding, and Hao Lin. Pai: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Scientific reports*, 6:35123, 2016.

[9] Hao Lin, Zhi-Yong Liang, Hua Tang, and Wei Chen. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.

[10] Farshid Rayhan, Sajid Ahmed, Dewan Md Farid, Abdollah Dehzangi, and Swakkhar Shatabda. Cfsboost: Cumulative feature subspace boosting for drug-target interaction prediction. *Journal of Theoretical Biology*, 2018.

[11] Wei Chen, Hui Ding, Pengmian Feng, Hao Lin, and Kuo-Chen Chou. iacp: a sequence-based tool for identifying anticancer peptides. *Oncotarget*, 7(13):16895, 2016.

[12] Atul Tyagi, Pallavi Kapoor, Rahul Kumar, Kumardeep Chaudhary, Ankur Gautam, and GPS Raghava. In silico models for designing and discovering novel anticancer peptides. *Scientific reports*, 3:2984, 2013.

[13] Zohre Hajisharifi, Moien Piryaiee, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. Predicting anticancer peptides with chou s pseudo amino acid composition and investigating their mutagenicity via ames test. *Journal of Theoretical Biology*, 341:34–40, 2014.

[14] M Saifur Rahman, Swakkhar Shatabda, Sanjay Saha, M Kaykobad, and M Sohel Rahman. Dpp-pseaac: A dna-binding protein prediction model using chous general pseaac. *Journal of theoretical biology*, 452:22–34, 2018.

[15] Saravanan Vijayakumar and PTV Lakshmi. Acpp: a web server for prediction and design of anti-cancer peptides. *International Journal of Peptide Research and Therapeutics*, 21(1):99–106, 2015.

[16] Leyi Wei, Chen Zhou, Huangrong Chen, Jiangning Song, and Ran Su. Acpred-fl: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, 2018.

[17] Georges St Laurent, Michael R Tackett, Sergey Nechkin, Dmitry Shtokalo, Denis Antonets, Yiannis A Savva, Rachel Maloney, Philipp Kapranov, Charles E Lawrence, and

Robert A Reenan. Genome-wide analysis of a-to-i rna editing by single-molecule sequencing in drosophila. *Nature structural & molecular biology*, 20(11):1333, 2013.