iRecSpot-EF: Effective Sequence Based Features for Recombination Hotspot Prediction

Md Rafsan Jani, Md Toha Khan Mozlish, Sajid Ahmed, Niger Sultana Tahniat, Dewan Md Farid, and Swakkhar Shatadba

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

1 Supplementary Information (Sp3): How does our web service work?

1.1 Step #1:

Access the service by clicking the link at http://irecspot.pythonanywhere.com/. We will see the homepage.

iRecSpot-EF Home Server Read Me Downloads Citation Contributors

iRecSpot-EF: Effective Sequence Based Features for Recombination Hotspot Prediction

Abstract

In genetic evolution, meiotic recombination plays an important role. Recombination introduces genetic variations and is a vital source of biodiversity and appears as a driving force in evolutionary development. Local regions of chromosomes where recombination events tend to be concentrated are known as hotspots and regions with relatively low frequencies of recombination are called coldspots. Predicting hotspots and coldspots can enlighten structure of recombination and genome evolution. In this paper, we proposed a predictor, called iRecSpot-EF to predict recombination and cold spots. iRecSpot-EF uses a novel set of features extracted from the genome sequences. We introduce the frequency of *I*, *k*, *p*-mers in the sequence as features. After the feature extraction, the best features are selected using AdaBoost algorithm. We have selected logistic regression algorithm as the classification algorithm. iRecSpot-EF was tested on a standard benchmark dataset using cross-fold validation. It achieved an accuracy of 95.14% and area under Receiver Operating Characteristic curve (auROC) of 0.985. The performance of iRecSpot-EF is significantly better than the state-of-the-art methods.

Figure 1: Home Page

1.2 Step #2:

Click on the 'Server' button on the top menu or go to this link http://irecspot.pythonanywhere.com/ server. We will see the server page.



Figure 2: Server Page

1.3 Step #3:

Once in the server page, either type or copy and paste the query DNA sequence into the input box. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with the symbol '>' in the first column, followed by lines of sequence data in which nucleotides are represented using single-letter codes. Except for the mandatory symbol '>', all the other characters in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with the symbol '>' appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

1.4 Step #4:

Users are able to set the step size which denotes length of the sub-sequences of interest. The predictor predicts the type (hotspot, coldspot or none) for each sub-sequence according to the selected step size. Example: if sequence length is 519 and step size is 200, then the model will predict the types of three sub-sequences. First one ranging from position 1 to 200 of the input sequence, the second one from position 201 to 400 and the third one from 401 to 519.

at iRecSpot-EF Home Server Read Me Downloads Citation Contributors	
Effective Sequence Based Features for Recombination Hotspot Prediction	
Enter the query sequences in FASTA format : >YDL087C 1.67 ATGTCAACTATGTCAACGCCTGCCGCAGAACAACGCAAACTCGTCGAACAGGCTGATGGGCAGGAC TTCAGTTTCCGCCACACAGGTATTCGCATCAAAAAAGAGACCTCGGACTACACGATCCAAGATC TGCAAGTCATCTGTTGGCGAGTGCCCCCTACGACCAGGGCCAACAGAGCCTGGGA AAATGCCCCGCCAGATGCATCTACCAAGCATAAAATTCAGTACGAGAGAGGGCCAAGCAGGGCCAAA	
ACGTTTCCCGAATTCGAAAGAGAATATCTGGCCATTCTATCTGGGTGTTAATGAGTGTAATGGC CAGATATCCGTAGCACTACAAAATCTAAAACACACCCCCTGAGAACGAATGAAGATTCAGCAGGTT ACCGAAGAACTAGATGTCTTGGACGTGCGGGCCAAGGGCAAGAGAGATTGATCTTTAATT CCTGCAGATGAAGTCAGTATGGGTATGCTACAGCACGCAAAGGACTGATTACTAAAAGG AAAGAAGTTGCAAAACGTGTACGAAACATACAGAAAACGTGGGCCCAGGGCCCAGCACAAGGTA Recomended browser : Google Chrome or Mozilla Firefox Enter Step Size: 200 (Recommended step size > 200) Submit Reset Example	

Figure 3: Enter FASTA and step size

1.5 Step #5:

Click on the 'Submit' button for predictions. Please wait, the system will take some time to generate results.

🛫 iRecSpot-EF	Home	Server	Read Me	Downloads	Citation	Contributors	
---------------	------	--------	---------	-----------	----------	--------------	--

Results

FASTA: >YDL087C 1.67 Result Visualization								
	Position	Predicted Type	Probability	Sequence Details				
1	1 - 200	hotspot	0.957988204916	Sequence Details				
2	201 - 400	None	0.525209029969	Sequence Details				
3	401 - 600	hotspot	0.896245125194	Sequence Details				
4	601 - 773	None	0.664904784768	Sequence Details				
FASTA: >YDR163W 0.82 Result Visualization								
	Position	Predicted Type	Probability	Sequence Details				
1	1 - 200	coldspot	0.772002728263	Sequence Details				
2	201 - 400	None	0.672076476888	Sequence Details				
3	401 - 519	None	0.592098291188	Sequence Details				

Figure 4.1: Predictions for respective sub-sequences

Note that each of the predictions is accompanied by the corresponding probability for the predicted class which can be interpreted as prediction-confidence.

1.5.1 Step #5.1:

Click 'Result visualization' to see the sequence details.

🧈 iRecSpo	t-EF	Home	Server	Read Me	Downloads	Citation	Contributors	
Result visualization of FASTA: >YDL087C 1.67								
Postion	Hots	pot Color:	-	Coldspot Col	or:	None Type Co	lor:	
1 - 200	ATG AGG TCC/ AAC/	ICAACTAT ACTTCAG AGATCTC AGAGCCTC	GTCAACG ITTCCGCC GCAAGTCA GGGAAAAT	CCTGCCGCA CACAACAGGT TACCTTGTT(G	GAACAACG ATTCGCATC GGCGAGTG	CAAACTCGTC CAAAAAAGAGA CCCCTACGAC	GAACAGCTGATGGGC ICCTCGGACTACACGA CTGTTTCAGGGCACC	
201 - 400	CCC ACG TGG AGC	GCAGATG ITTCCCG CCAGATA AGGTTAC	CATCTTAC AATTCGAA TCCGTAGC CGAAGAAC	CAAGCATAA AGAGAATAT CACTACAAAA	AATTCAGTA CTGGCCATT TCTAAAACA	CGAGAGAGAGG CTATCTCGGT CACCGCTGAG	GTCAAGCAGGGCAAA TGTTAATGAGTGTAA GAACGAATGAAGATTC	
401 - 600	TAG/ AGA ^T AAG/ GTT/	ATGTCTTC IGAAGTC/ AGTTGC/ ACAGGTTT	AGACGTGO AGTATGGO AAAACGTG TGCGAGG	CGGATAGGC TATGCTACA TACGAAACA TG	CAATGGGAC GCAGTCAA/ TACAGAAAA	CAAGAGATTGA ATTACAAGAAC CGTGGGCCAC	ATTCTTTAATTCGTGC CTGATTAGTAAAAGGA GAGCGCCCAGCAAAA	
601 - 773	TGTO AGA ACA	GGGGCAT ITCATCTO ACTAACGO	CCTATCG GATATGT CAGCAAGA	CGTTTAGATA CAAGATGAG CAGCTACTA	ACAGACAGA AGAGGATTA CACTACCCG	AGGCTTGCTG ATGATCGGCT AGAAGACGCT	ACCACTTCTTGGGGA AATGAAGAATAACCGG ITGTGTA	

Figure 4.2: Results visualization