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

Supporting Information (#2):

List of features selected by using the AdaBoost feature selection technique. AdaBoost has selected features based on relation probability. Among 40,679 features only 425 features are essential for this dataset. However, The number 425 is not fixed, it depends on the dataset.

The table below contains feature ranking, feature number, feature structure (looking view), and feature importance (probability) it comes from.

Table 1: Features observation

Feature Ranking	**Feature Number**	**Feature Structure**	**Feature Importance**
1	9954	z-axis = (A+T)-(C+G)	0.008
2	30769	AGG_____CT	0.008
3	30710	AGA_____GC	0.008
4	12413	ATA__TC	0.006
5	13005	CTT__CC	0.006
6	40130	GTG_____TT	0.006
7	35526	CTT_____AC	0.006
8	4949	CGG_____G	0.006
9	1610	GC_____C	0.006
10	18664	CAA_____AG	0.006
11	20716	CAA_____CA	0.004
12	813	C_____CG	0.004
13	33333	CGG_____GC	0.004
14	38464	TGG_____TG	0.004
15	22625	AGT_____TT	0.004
16	20755	CAC_____GA	0.004
17	40214	TAC_____GC	0.004
18	35206	CCC_____AC	0.004
19	7711	TAT_____T	0.004
20	32559	AAG_____CG	0.004
21	2535	GTA_T	0.004
22	38322	TCG_____CT	0.004
23	9325	TTC_____G	0.004
24	21984	TCT_____TG	0.004
25	2985	GCC__A	0.004
26	7903	AGT_____T	0.004
27	6863	AGC_____T	0.004

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
28	32069	TAT_____AC	0.004
29	32588	AAT_____CA	0.004
30	18839	CCC____GG	0.004
31	30459	AAA_____TA	0.004
32	5848	A_____GTA	0.004
33	8421	ATA_____G	0.004
34	13508	GTT__AA	0.004
35	37409	CGC_____TT	0.004
36	34394	TGT_____GT	0.004
37	24020	TCT_____GA	0.004
38	33987	GTT_____AA	0.004
39	27838	GTG_____TC	0.004
40	11912	TTC_AG	0.004
41	37378	CGA_____TT	0.004
42	1486	GC_____T	0.004
43	21097	CTA_____AT	0.004
44	12169	ACC__AT	0.004
45	1993	G_____CA	0.004
46	31502	GAC_____CC	0.004
47	32006	TAC_____AC	0.004
48	32638	ACA_____TC	0.004
49	4697	TGT____A	0.004
50	23060	CGC_____GA	0.004
51	23479	GCG_____GG	0.004
52	20219	AAA_____TA	0.004
53	20483	AGC_____AA	0.004
54	11249	GGA_CT	0.004
55	24933	CCA_____AC	0.004
56	22411	ACC_____CA	0.004
57	37332	CCT_____GA	0.004
58	4021	GCG____G	0.004
59	18452	AGC____GA	0.004
60	20211	AAA_____GA	0.004
61	6387	ATG_____C	0.004
62	25548	GCT_____CA	0.004
63	13650	TAT__CT	0.002
64	7367	AGA_____T	0.002
65	11380	GTA_GA	0.002
66	16047	TTG___CG	0.002
67	18128	TTT____CG	0.002
68	13069	GAC__CC	0.002
69	23433	GCC_____AT	0.002
70	29403	CTT_____TA	0.002
71	40324	TCC_____AA	0.002
72	10603	CCA_CA	0.002
73	9261	TCC_____G	0.002
74	14066	AAA___CT	0.002
75	7131	GGT_____C	0.002
76	1712	CG_____A	0.002
77	17305	GCC____GT	0.002
78	35091	CAC_____GA	0.002
79	31996	TAA_____TA	0.002
80	7139	GTA_____C	0.002
81	3231	AAT___T	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
82	13060	GAC__AA	0.002
83	2414	C_TCT	0.002
84	3659	TGC___C	0.002
85	25992	TCC_____AG	0.002
86	1701	C_____AG	0.002
87	17295	GCC____CG	0.002
88	2405	CTA_G	0.002
89	8612	G_____CAG	0.002
90	23847	TAG_____AG	0.002
91	18410	AGA_____AT	0.002
92	22611	AGT_____GA	0.002
93	33263	CGA_____CG	0.002
94	29993	TAG_____AT	0.002
95	15424	GGG___TG	0.002
96	36977	ATA_____CT	0.002
97	25151	CGG_____TG	0.002
98	19953	TGA_____CT	0.002
99	7868	A_____CTG	0.002
100	10503	CAC_AG	0.002
101	3273	AGC___A	0.002
102	31944	GTT_____AG	0.002
103	27362	CTT_____TT	0.002
104	24247	TTG_____GG	0.002
105	1266	TG_____C	0.002
106	4053	GGG____G	0.002
107	24520	ACT_____AG	0.002
108	24757	ATG_____GC	0.002
109	10901	CTC_GC	0.002
110	38219	TAT_____CA	0.002
111	9012	C_____CGG	0.002
112	15383	GGC___GG	0.002
113	37173	CAG_____GC	0.002
114	34539	AAA_____CA	0.002
115	12218	ACG__GT	0.002
116	32396	TTC_____CA	0.002
117	32760	AGA_____GG	0.002
118	8414	A_____GTT	0.002
119	4042	G____GCC	0.002
120	2498	G_GAC	0.002
121	29431	GAA_____GG	0.002
122	2477	GCC_G	0.002
123	36947	AGT_____GA	0.002
124	36948	AGT_____GA	0.002
125	12550	CAC__AC	0.002
126	1228	GC_____G	0.002
127	15439	GGT___CG	0.002
128	32703	ACG_____TG	0.002
129	19031	CGT_____GG	0.002
130	14059	AAA___CA	0.002
131	27281	CTC_____CT	0.002
132	27279	CTC_____CG	0.002
133	32711	ACT_____AG	0.002
134	33722	GCG_____GT	0.002
135	5586	G_____GGC	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
136	27544	GCC_____GG	0.002
137	40354	TCC_____TT	0.002
138	6337	AGA_____A	0.002
139	14282	ACT___AT	0.002
140	21480	GGA_____AG	0.002
141	14527	ATG___TG	0.002
142	26020	TCG_____AA	0.002
143	1252	TA_____G	0.002
144	29682	GGA_____CT	0.002
145	11792	TGC_CG	0.002
146	26262	TTC_____GC	0.002
147	21203	CTT_____GA	0.002
148	30633	ACG_____AT	0.002
149	31328	CGT_____TG	0.002
150	2360	C_CTA	0.002
151	21414	GCG_____AC	0.002
152	14606	CAC___CC	0.002
153	9672	G_____GCA	0.002
154	38379	TGA_____CA	0.002
155	28653	AGA_____CC	0.002
156	9671	GGA_____T	0.002
157	19731	TAC_____GA	0.002
158	8484	C_____CAG	0.002
159	11910	TTC_AC	0.002
160	22009	TGA_____GT	0.002
161	9223	TAA_____T	0.002
162	9666	G_____GAC	0.002
163	3123	TCG__C	0.002
164	13154	GAT__TT	0.002
165	16136	AAC____AG	0.002
166	22442	ACG_____AT	0.002
167	6219	TGC_____C	0.002
168	40070	GTC_____AC	0.002
169	40167	TAA_____AG	0.002
170	13547	TAA__CA	0.002
171	25846	TAA_____GC	0.002
172	5673	TCC_____A	0.002
173	14700	CCA___CA	0.002
174	26959	CAT_____CG	0.002
175	31432	CTT_____AG	0.002
176	1039	A_____CT	0.002
177	9767	TCA_____T	0.002
178	38867	ACT_____GA	0.002
179	12362	AGT__AT	0.002
180	6199	TCG_____T	0.002
181	37097	CAA_____AT	0.002
182	28228	TGT_____AA	0.002
183	33411	CTC_____AA	0.002
184	1813	A_____GG	0.002
185	14636	CAG___CA	0.002
186	33791	GGA_____TG	0.002
187	1069	C_____CG	0.002
188	4239	AAC____T	0.002
189	24897	CAG_____TT	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
190	1072	CG_____A	0.002
191	24900	CAT_____AA	0.002
192	4696	T____GTA	0.002
193	8076	G_____ACG	0.002
194	10732	CGA_CA	0.002
195	12349	AGG__TC	0.002
196	1076	CG_____G	0.002
197	10745	CGA_GT	0.002
198	23764	GTT_____GA	0.002
199	18025	TTA____AT	0.002
200	33785	GGA_____GT	0.002
201	5741	TTC_____G	0.002
202	18473	AGG____AT	0.002
203	38859	ACT_____CA	0.002
204	26343	AAA_____AG	0.002
205	4733	TTT____G	0.002
206	22528	AGA_____TG	0.002
207	23798	TAA_____GC	0.002
208	29736	GGG_____AG	0.002
209	10802	CGG_CT	0.002
210	40233	TAG_____AT	0.002
211	8566	C_____TGT	0.002
212	3067	GTT__C	0.002
213	30767	AGG_____CG	0.002
214	23420	GCA_____TA	0.002
215	19098	CTC____GT	0.002
216	13095	GAG__AG	0.002
217	11039	GAC_TG	0.002
218	3214	A___ACT	0.002
219	31478	GAA_____GC	0.002
220	10630	CCC_AC	0.002
221	2930	C__TGC	0.002
222	14582	CAA___GC	0.002
223	20657	ATG_____CT	0.002
224	6270	T_____TTT	0.002
225	33860	GGT_____AA	0.002
226	34838	AGC_____GC	0.002
227	31486	GAA_____TC	0.002
228	20649	ATG_____AT	0.002
229	12280	AGA__GG	0.002
230	34833	AGC_____CT	0.002
231	26443	AAT_____CA	0.002
232	7964	C_____ATG	0.002
233	32645	ACC_____AC	0.002
234	8584	G_____ACA	0.002
235	20680	ATT_____AG	0.002
236	6255	TTC_____T	0.002
237	9125	GCA_____G	0.002
238	24848	CAC_____CG	0.002
239	36377	TGC_____GT	0.002
240	18435	AGC____AA	0.002
241	4788	A_____CGG	0.002
242	38338	TCG_____TT	0.002
243	14719	CCA___TG	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
244	20709	CAA_____AC	0.002
245	19777	TAG____TT	0.002
246	16163	AAG____AA	0.002
247	34985	ATG_____AT	0.002
248	11207	GCT_AG	0.002
249	25819	GTT_____TA	0.002
250	19122	CTG____CT	0.002
251	11966	TTG_TC	0.002
252	9138	G_____CGC	0.002
253	25945	TAT_____GT	0.002
254	452	GA__G	0.002
255	4166	T____GAT	0.002
256	18428	AGA____TA	0.002
257	8557	CTC_____G	0.002
258	32135	TCC_____AG	0.002
259	14235	ACC___TA	0.002
260	7998	C_____CTT	0.002
261	14130	AAG___CT	0.002
262	19628	GTG____CA	0.002
263	14508	ATG___CA	0.002
264	4704	T____TAA	0.002
265	25365	GAC_____GC	0.002
266	18690	CAA____TT	0.002
267	5269	AAG_____G	0.002
268	2584	T_ATA	0.002
269	2585	TAT_A	0.002
270	12905	CTA__AT	0.002
271	37277	CCC_____TC	0.002
272	19426	GCT____TT	0.002
273	13955	TTC__AA	0.002
274	31089	CCA_____CT	0.002
275	10184	ACT_AG	0.002
276	24607	AGC_____TG	0.002
277	25497	GCC_____GT	0.002
278	28480	AAG_____TG	0.002
279	15042	CTG___TT	0.002
280	20403	ACG_____GA	0.002
281	12652	CCA__CA	0.002
282	16393	AGC____AT	0.002
283	6051	GCA_____C	0.002
284	3425	CTA___A	0.002
285	19366	GCG____AC	0.002
286	35374	CGG_____CC	0.002
287	26674	AGG_____CT	0.002
288	22175	TTC_____TG	0.002
289	28421	AAC_____AC	0.002
290	27090	CCT_____CT	0.002
291	7664	G_____TGA	0.002
292	5134	T_____ACT	0.002
293	22785	CAA_____TT	0.002
294	22155	TTC_____CA	0.002
295	24710	ATC_____AC	0.002
296	27721	GGT_____AT	0.002
297	31719	GGA_____AG	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
298	23959	TCC_____GG	0.002
299	23076	CGG_____AA	0.002
300	11498	TAA_AT	0.002
301	6512	C_____TGA	0.002
302	30304	TGT_____TG	0.002
303	30386	TTG_____CT	0.002
304	21575	GGT_____AG	0.002
305	13340	GGC__TA	0.002
306	2022	TA_____T	0.002
307	17703	TAG____AG	0.002
308	18213	AAG_____AC	0.002
309	5104	G_____TGA	0.002
310	16504	ATA____GG	0.002
311	6581	GCG_____G	0.002
312	14332	AGA___TA	0.002
313	30407	TTT_____AG	0.002
314	30597	ACC_____AC	0.002
315	778	AC_____C	0.002
316	22177	TTC_____TT	0.002
317	31679	GCG_____TG	0.002
318	32236	TGA_____CA	0.002
319	36690	AAT_____CT	0.002
320	35516	CTG_____TA	0.002
321	15191	GAT___GG	0.002
322	11694	TCG_CC	0.002
323	20308	AAT_____GA	0.002
324	5331	AGG_____C	0.002
325	4385	CCA_____A	0.002
326	4490	G_____ACC	0.002
327	18897	CCT_____CT	0.002
328	16372	AGA____GA	0.002
329	24162	TGT_____TT	0.002
330	27904	TAA_____TG	0.002
331	39092	ATG_____GA	0.002
332	20296	AAT_____AG	0.002
333	17698	TAC____TT	0.002
334	10090	ACA_AT	0.002
335	5177	TCT_____A	0.002
336	12661	CCA__GC	0.002
337	28442	AAC_____GT	0.002
338	35473	CTC_____CT	0.002
339	30487	AAC_____GG	0.002
340	1568	CA_____A	0.002
341	21795	TAG_____AA	0.002
342	15966	TGT___TC	0.002
343	31202	CCT_____TT	0.002
344	27671	GGC_____GG	0.002
345	32258	TGA_____TT	0.002
346	7786	T_____TCC	0.002
347	8708	T_____AAG	0.002
348	7487	CCT_____T	0.002
349	12196	ACG__AA	0.002
350	5442	C_____GAC	0.002
351	22213	TTT_____AC	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
352	40435	TGA_____GA	0.002
353	14301	ACT___TC	0.002
354	30121	TCG_____AT	0.002
355	8681	GTC_____A	0.002
356	14930	CGT___CT	0.002
357	7701	TAG_____G	0.002
358	16313	ACG____GT	0.002
359	3727	AAC____T	0.002
360	6997	CGG_____G	0.002
361	4427	CGC_____C	0.002
362	23555	GGC_____AA	0.002
363	33156	CCC_____AA	0.002
364	40660	TTT_____GA	0.002
365	34745	ACG_____GT	0.002
366	3298	A___TAC	0.002
367	22737	ATT_____CT	0.002
368	23669	GTA_____GC	0.002
369	20229	AAC_____AC	0.002
370	6675	TAG_____C	0.002
371	7437	CAC_____G	0.002
372	11622	TCA_AC	0.002
373	13406	GGT__TC	0.002
374	36304	TCT_____CG	0.002
375	36607	AAA_____TG	0.002
376	9313	TTA_____A	0.002
377	3924	C____GGG	0.002
378	3856	C____AGA	0.002
379	7831	AAG_____T	0.002
380	27979	TAT_____CA	0.002
381	8886	A_____CGT	0.002
382	22730	ATT_____AT	0.002
383	26695	AGT_____AG	0.002
384	14005	TTG__GC	0.002
385	21851	TAT_____TA	0.002
386	32781	AGC_____CC	0.002
387	32969	ATT_____AT	0.002
388	6552	G_____ATA	0.002
389	20976	CGA_____CG	0.002
390	31159	CCG_____GG	0.002
391	8769	TGA_____A	0.002
392	2010	GT_____C	0.002
393	32955	ATG_____TA	0.002
394	23097	CGG_____GT	0.002
395	36106	TAC_____AT	0.002
396	24003	TCT_____AA	0.002
397	4533	GCG_____G	0.002
398	7796	T_____TGG	0.002
399	11050	GAG_AT	0.002
400	33012	CAA_____GA	0.002
401	6648	G_____TTA	0.002
402	21559	GGG_____GG	0.002
403	22209	TTG_____TT	0.002
404	38571	TTG_____CA	0.002
405	24011	TCT_____CA	0.002

Continued on next page

Table 1 Continued from previous page

Feature Ranking	Feature Number	Feature Structure	Feature Importance
406	3888	C___CGA	0.002
407	6596	G_____GAG	0.002
408	31631	GCC_____CG	0.002
409	20325	ACA_____AC	0.002
410	8233	TCC_____A	0.002
411	14395	AGG___TA	0.002
412	3849	CAC____A	0.002
413	18661	CAA_____AC	0.002
414	2543	GTC_T	0.002
415	39207	CAG_____AG	0.002
416	5210	T_____GTC	0.002
417	16933	CGG____AC	0.002
418	39209	CAG_____AT	0.002
419	27834	GTG_____GT	0.002
420	31828	GGT_____GA	0.002
421	1451	C_____CC	0.002
422	33693	GCC_____TC	0.002
423	25587	GGA_____GA	0.002
424	12616	CAT__AG	0.002
425	34628	AAT_____AA	0.002

Table 1 shows, feature ranking, feature number/index (0-base indexing), feature structure, and feature importance. From the above table '_' represents 1-gap, '__' represents 2-gap, '___' represents 3-gap and so on.

Table 2: Probability calculation among features

Ranks of feature	Probability per feature	Group Probability
1-3	0.008	$0.008 \times 3 = 0.024$
4-10	0.006	$0.006 \times 7 = 0.042$
11-62	0.004	$0.004 \times 52 = 0.208$
63-425	0.002	$0.002 \times 363 = 0.726$

Table 2 shows, the important number of features according to their probability. The overall probability is 1.00.