Supplementary Material: Evolutionary analysis of base-pairing interactions in DNA and RNA secondary structures

1 Supplementary methods

1.1 Stationarity and time-reversibility

1.2 Modelling site-to-site rate variation

In the M95 and extended M95 models, the substitution rate was assumed to be the same for each of the two nucleotide positions within a pair, as well as across all possible site pairs. However, it is well-established that the rate of substitution can vary across nucleotide positions and that failing to account for rate variability can lead to biased parameter estimates (Yang, 1996). Additionally, many of the datasets analysed in this thesis have coding regions, where it is expected that the third nucleotide position in each codon (the so-called 'codon wobble position') will have relatively higher substitution rates associated with it, due to there being a lower chance of mutations modifying the encoded amino acid.

To model variable substitution rates across sites, the gamma distributed sites rate approach of (Yang, 1993, 1994) was implemented. The GTR model with gamma distributed sites, denoted GTR + Γ, is well-understood for single site substitution models, however, an extension to paired site models had not been previously described. The modified M95 paired site model with gamma distributed site rates, denoted M95 + Γ, is given below:

Table S1: Consensus structure ranking. 118 non-overlapping HIV consensus substructures ranked from highest to lowest z-score based on their degrees of coevolution within an alignment of HIV-1 subtype B sequences. Where the canonical structure was treated as unknown and a consensus structure predicted

Rank	Alignment position	Mapped position	Length	Name and reference	Median	z-score
1	8240 - 8577	7256 - 7590	338	Rev Response element (RRE, Heaphy *et al.* (1990); Mandal and Breaker (2004))	5.64	6.53
2	2202 - 2229	1645 - 1672	28	Gag-pol frameshift (Chamorro *et al.*, 1992)	8.17	4.56
3	1710 - 1845	1177 - 1312	136		6.44	4.50
4	4751 - 4833	4134 - 4216	83		6.47	3.97
5	4505 - 4709	3888 - 4092	205		5.22	3.21
6	591 - 939	108 - 445	349	5' Untranslated region (5'UTR, Siegfried *et al.* (2014))	5.38	3.16
7	133 - 151	NA	19		6.85	2.94
8	2564 - 2890	1947 - 2273	327	Longest continuous helix (Siegfried *et al.*, 2014)	4.44	2.62
9	9782 - 9800	8645 - 8663	19		6.92	2.55
10	3612 - 3623	2995 - 3006	12		6.74	2.50
11	5690 - 5720	5070 - 5100	31		5.76	2.43
12	9873 - 9881	8736 - 8744	9		10.71	2.31
13	5622 - 5642	5005 - 5025	21		7.22	2.26
14	9733 - 9781	8597 - 8644	49		5.20	2.22
15	10008 - 10020	8871 - 8883	13		7.39	2.00
16	10061 - 10390	NA	330	3' Untranslated region (3'UTR, Siegfried *et al.* (2014))	4.54	1.82
17	7410 - 7425	6537 - 6552	16		5.65	1.78
18	1185 - 1211	NA	27		5.63	1.76
19	3538 - 3552	2921 - 2935	15		5.97	1.50
20	8057 - 8214	7088 - 7230	158		4.77	1.49
21	1080 - 1146	586 - 652	67	5' Untranslated region (5'UTR, Siegfried *et al.* (2014))	6.49	1.45
22	6627 - 6643	5966 - 5982	17	SP stem (Siegfried *et al.* (2014))	4.98	1.44
23	3261 - 3268	2644 - 2651	8		5.79	1.28
24	3577 - 3586	2960 - 2969	10		5.64	1.09
25	7434 - 7468	NA	35		4.99	1.00
26	105 - 118	NA	14		4.82	0.86
27	9443 - 9464	NA	22		5.35	0.80
28	7536 - 7648	6660 - 6766	113		4.63	0.77
29	1866 - 1882	1333 - 1349	17		4.98	0.77
30	4161 - 4187	3544 - 3570	27		4.41	0.66
31	8005 - 8054	7036 - 7085	50		4.39	0.52
32	3342 - 3393	2725 - 2776	52		3.98	0.52
33	3270 - 3299	2653 - 2682	30		4.93	0.48
34	1908 - 1920	1375 - 1387	13		4.28	0.28
35	4306 - 4320	3689 - 3703	15		4.02	0.26
36	953 - 961	459 - 467	9	5' Untranslated region (5'UTR, Siegfried *et al.* (2014))	4.56	0.19
37	1422 - 1707	889 - 1174	286		4.23	0.14

Rank	Alignment position	Mapped position	Length	Name and reference	Median	z-score
38	419 - 476	NA	58		3.98	0.11
39	7918 - 7923	6982 - 6987	6		4.48	0.06
40	9895 - 9932	8758 - 8795	38		4.16	0.04
41	381 - 397	NA	17		4.51	-0.02
42	3985 - 4015	3368 - 3398	31		4.01	-0.04
43	9135 - 9218	8127 - 8210	84		4.26	-0.04
44	8955 - 9069	7947 - 8061	115		4.04	-0.04
45	5835 - 6054	5215 - 5426	220		3.97	-0.11
46	1285 - 1308	752 - 775	24		4.05	-0.27
47	478 - 546	NA	69	5' Trans-activation response element (5' TAR, Roy *et al.* (1990))	4.16	-0.50
48	6204 - 6212	5573 - 5581	9		2.64	-0.55
49	9258 - 9266	8250 - 8258	9		3.12	-0.55
50	4031 - 4058	3414 - 3441	28		4.41	-0.62
51	5761 - 5769	5141 - 5149	9		3.22	-0.70
52	3588 - 3596	2971 - 2979	9		3.20	-0.73
53	9661 - 9672	NA	12		3.72	-0.74
54	3646 - 3655	3029 - 3038	10		3.40	-0.76
55	10021 - 10030	8884 - 8893	10		3.14	-0.76
56	3685 - 3817	3068 - 3200	133		3.29	-0.77
57	4862 - 4875	4245 - 4258	14	Central polypurine tract (CPPT, Siegfried *et al.* (2014))	3.65	-0.78
58	6092 - 6192	5464 - 5561	101		3.81	-0.80
59	1994 - 2002	NA	9		2.67	-0.83
60	7856 - 7875	NA	20		2.93	-0.86
61	4134 - 4145	3517 - 3528	12		3.23	-0.86
62	7987 - 8000	7018 - 7031	14		3.26	-0.87
63	9940 - 9986	8803 - 8849	47		3.09	-0.87
64	7687 - 7709	6805 - 6821	23		3.24	-0.91
65	3396 - 3536	2779 - 2919	141		3.60	-0.98
66	9080 - 9096	8072 - 8088	17		2.82	-1.05
67	5799 - 5834	5179 - 5214	36		4.06	-1.06
68	5566 - 5602	4949 - 4985	37		3.23	-1.14
69	6727 - 6875	NA	149		3.03	-1.18
70	2234 - 2563	1677 - 1946	330	Gag-pol frameshift (Chamorro *et al.*, 1992)	3.68	-1.21
71	6068 - 6078	5440 - 5450	11		2.03	-1.22
72	6646 - 6659	5985 - 5998	14	SP stem (Siegfried *et al.* (2014))	3.13	-1.23
73	1895 - 1901	1362 - 1368	7		2.00	-1.23
74	6403 - 6443	5757 - 5788	41		3.05	-1.24
75	3867 - 3878	3250 - 3261	12		3.32	-1.25
76	7715 - 7729	NA	15		2.57	-1.27
77	6547 - 6583	5886 - 5922	37	SP stem (Siegfried *et al.* (2014))	3.26	-1.27
78	4440 - 4463	3823 - 3846	24		3.10	-1.29

Rank	Alignment position	Mapped position	Length	Name and reference	Median	z-score
79	1231 - 1263	704 - 730	33		2.93	-1.29
80	4084 - 4092	3467 - 3475	9		3.18	-1.31
81	1954 - 1976	1421 - 1443	23		3.42	-1.34
82	995 - 1026	501 - 532	32	5' Untranslated region (5'UTR, Siegfried *et al.* (2014))	3.62	-1.35
83	7290 - 7298	6417 - 6425	9		1.74	-1.36
84	3166 - 3214	2549 - 2597	49		3.36	-1.38
85	3624 - 3641	3007 - 3024	18		2.44	-1.38
86	1314 - 1394	781 - 861	81		3.31	-1.42
87	7395 - 7409	6522 - 6536	15		1.32	-1.52
88	4390 - 4403	3773 - 3786	14		3.00	-1.53
89	5784 - 5792	5164 - 5172	9		2.19	-1.59
90	2030 - 2052	1482 - 1498	23		2.48	-1.61
91	2110 - 2197	1556 - 1640	88	Gag-pol frameshift (Chamorro *et al.*, 1992)	3.73	-1.69
92	4466 - 4499	3849 - 3882	34		2.49	-1.75
93	4938 - 4962	4321 - 4345	25	Central polypurine tract (CPPT, Siegfried *et al.* (2014))	1.90	-1.81
94	7742 - 7759	6848 - 6862	18		2.70	-1.84
95	6590 - 6600	5929 - 5939	11	SP stem (Siegfried *et al.* (2014))	1.84	-1.86
96	9605 - 9637	8475 - 8507	33		2.21	-1.89
97	8805 - 8869	7800 - 7864	65		3.22	-1.91
98	1933 - 1953	1400 - 1420	21		1.99	-2.02
99	9098 - 9117	8090 - 8109	20		2.70	-2.03
100	8925 - 8942	7920 - 7937	18		1.97	-2.07
101	6461 - 6472	5806 - 5817	12		1.46	-2.08
102	2986 - 3113	2369 - 2496	128		3.09	-2.11
103	6504 - 6522	5846 - 5861	19	SP stem (Siegfried *et al.* (2014))	2.17	-2.20
104	9273 - 9354	NA	82		3.22	-2.21
105	3883 - 3922	3266 - 3305	40		2.65	-2.28
106	9564 - 9573	8434 - 8443	10		1.50	-2.29
107	7880 - 7898	6956 - 6962	19		2.26	-2.32
108	7028 - 7258	6245 - 6385	231		3.02	-2.32
109	9364 - 9398	NA	35		2.71	-2.33
110	5192 - 5541	4575 - 4924	350		3.69	-2.34
111	6284 - 6391	5638 - 5745	108		2.38	-2.36
112	7783 - 7822	6886 - 6925	40		2.53	-2.42
113	5114 - 5133	4497 - 4516	20		1.91	-2.52
114	9221 - 9238	8213 - 8230	18		2.39	-2.60
115	6885 - 6922	NA	38		1.99	-2.67
116	8580 - 8606	7593 - 7607	27		1.32	-2.87
117	8622 - 8762	7623 - 7757	141		2.61	-2.92
118	4991 - 5072	4374 - 4455	82		2.65	-3.25

References

Chamorro, M., Parkin, N., and Varmus, H. E. 1992. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proceedings of the National Academy of Sciences*, 89(2): 713–717.

Heaphy, S., Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Kern, J., Lowe, A. D., Singh, M., and Skinner, M. A. 1990. HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell*, 60(4): 685–693.

Mandal, M. and Breaker, R. R. 2004. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5(6): 451–463.

Roy, S., Delling, U., Chen, C., Rosen, C., and Sonenberg, N. 1990. A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation. *Genes & development*, 4(8): 1365–1373.

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A., and Weeks, K. M. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods*, 11(9): 959–965.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6): 1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3): 306–314.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9): 367–372.