COATi: statistical pairwise alignment of protein coding sequences

Supplementary Materials

Juan J. Garcia Mesa, Ziqi Zhu, Reed A. Cartwright

Contents

1 Aligner Commands **1**

2 FST Alignment Example **2**

3 Gap Patterns **3**

3.1 Lengths 3

3.2 Phases 4

3.3 Gapiness 5

1 Aligner Commands

We evaluated five different aligners. Below are the commands that we used to run them. We have abbreviated the commands for clarity, stripping out unimportant arguments. Complete workflows can be found in our coati-testing repository on Github.

- COATi: `coati alignpair -m tri-mg ...`; We used COATi two different ways. The primary way used human as the reference sequence. The secondary way, COATi-rev, used gorilla as the reference, via a wrapper script that reversed the order of sequences before alignment and reversed the order back afterwards.
- ClustalΩ v1.2.4: `clustalo --seqtype=Protein ...`; We wrapped ClustalΩ with a script that translates DNA sequences into amino acid sequences (including stops) before alignment. Any codons that were partial or containing ambiguous characters were translated as “X”. The script then aligned these translated sequences with ClustalΩ, and then created a DNA alignment that was consistent with the amino-acid alignment.
- MACSE v2.06: `java -jar macse.jar -prog alignSequences -seq human.fasta -seq_lr gorilla.fasta ...`; We wrapped MACSE with a script that created two temporary fasta files, one containing the human sequence and another containing the gorilla sequence. The human sequence was specified as the reliable sequence and the gorilla sequence was specified as the less-reliable sequence. MACSE uses “!” to mark gaps that result from frameshifts, and the wrapper script replaced these with “-”. Additionally, MACSE sometimes produced columns that only contained gaps, and these columns were removed.
- PRANK v.150803: `prank -codon ...`
- MAFFT v7.520: `mafft --preservecase --globalpair --maxiterate 1000 ...`

COATi can use different alignment models for pairwise alignment. Below are the commands that we used to run different models.

- tri-mg: `coati alignpair -m tri-mg ...`
- tri-ecm: `coati alignpair -m tri-ecm ...`

- mar-mg: `coati alignpair -m mar-mg ...`
- mar-ecm: `coati alignpair -m mar-ecm ...`

2 FST Alignment Example

When using the triplet models (tri-mg and tri-ecm), COATi uses the OpenFST library to generate best alignments by composing the input and output sequences with the COATi FST model. While the COATi FST model is too large to display, we can show the result of a composition.

Fig. S1 shows a graph depicting the FST that results from composing the COATi FST with the input sequence "CTC" and the output sequence "CTG". Every path through this FST represents one possible way to align "CTC" and "CTG", and the sum of all weights along a path is the total weight of the respective alignment. Here, the weights of each arc are in negative-log space. Note that this graph has been optimized, and weight has been pushed towards the initial state. The weight of any specific arc may not be directly mapable to a weight described in the model.

Fig. S2 is the best alignment between "CTC" and "CTG", as determined by the shortest path algorithm. Figs. S1 and S2 were produced by the OpenFST library. A bold circle represents a starting node, and a double circle represents a termination node.

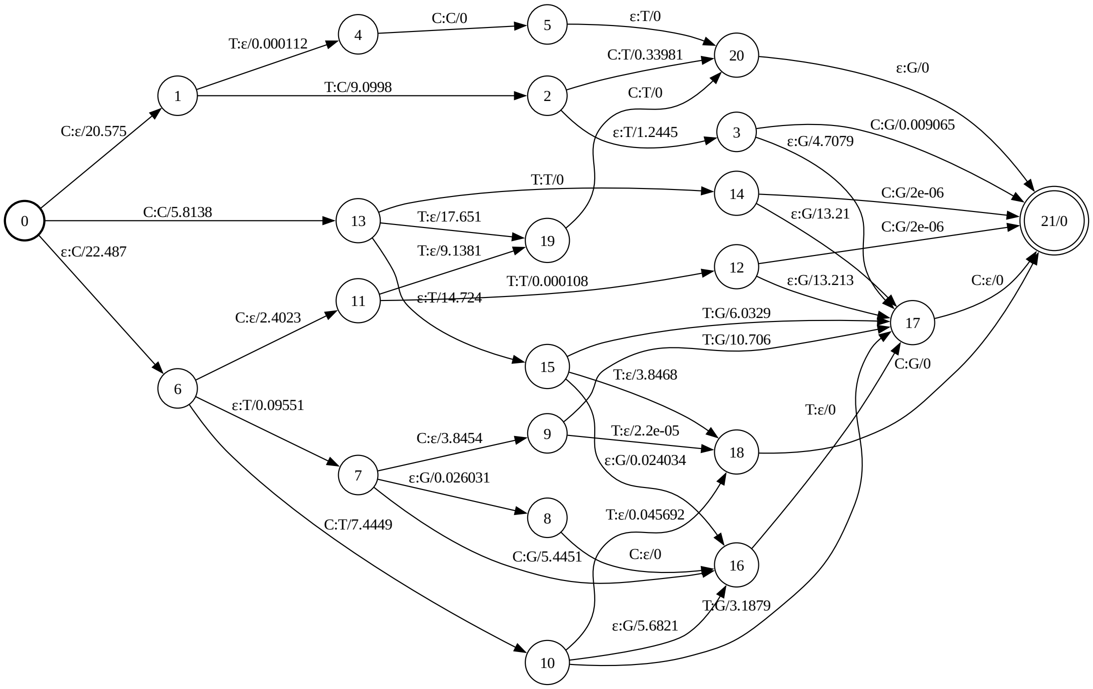

Figure S1: The FST of all possible alignments between 'CTC' and 'CTG'.

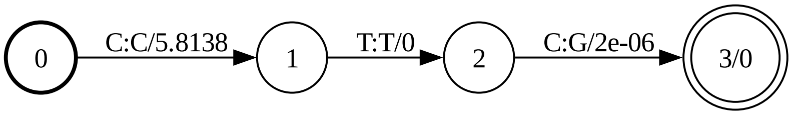

Figure S2: The best alignment of 'CTC' and 'CTG'.

3 Gap Patterns

3.1 Lengths

We quantified the lengths of gaps produced by each method across the empirical dataset of human-gorilla protein-coding-sequence pairs (Tab. S1). Here the gap type is either "D" for gaps introduced into the gorilla sequence or "I" for gaps introduced in the human sequence. "COATi" refers to the full COATi FST model with a MG substitution model (i.e. tri-mg). "COATi-rev" refers to the same model, but using gorilla as the reference and human as the non-reference sequence. Gap lengths of 1–6 nucleotides were binned into their respective columns. Gap lengths longer than 6 nucleotides were binned into columns 7+, 8+, and 9+ depending on whether their lengths were 1, 2, or 3 nucleotides longer than a multiple of three.

We also quantified the lengths of gaps produced by different COATi models (Tab. S2).

Note that ClustalΩ gaps with lengths that are not a multiple of 3 are created by how our wrapper script handles DNA sequences with lengths that are not multiple of 3.

Table S1: Number of gaps introduced by each method separated by length and type.

		Gap Lengths								
Method	Gap Type	1	2	3	4	5	6	7+	8+	9+
COATi	D	106	99	651	94	94	226	1461	1386	3240
COATi	I	87	73	525	80	97	239	1399	1315	3105
COATi-rev	D	105	99	634	89	96	217	1425	1348	3169
COATi-rev	I	82	79	513	80	97	231	1360	1288	3031
ClustalΩ	D	1	0	887	1	1	400	3	7	3910
ClustalΩ	I	0	0	857	0	1	424	3	2	3747
MACSE	D	396	306	1237	5	3	657	71	54	4784
MACSE	I	60	27	1133	5	9	666	78	97	4381
MAFFT	D	204	156	714	147	108	276	544	509	2680
MAFFT	I	187	154	708	112	92	295	454	424	2642
PRANK	D	0	0	552	0	0	167	0	0	4812
PRANK	I	0	0	467	0	0	178	0	0	4686

Table S2: Number of gaps introduced by different COATi models separated by length and type.

		Gap Lengths								
Model	Gap Type	1	2	3	4	5	6	7+	8+	9+
TRI-MG	D	106	99	651	94	94	226	1461	1386	3240
TRI-MG	I	87	73	525	80	97	239	1399	1315	3105
TRI-ECM	D	93	96	665	125	122	257	1438	1441	3203
TRI-ECM	I	86	103	540	105	122	254	1358	1370	3099
MAR-MG	D	102	106	646	95	89	224	1465	1399	3226
MAR-MG	I	84	83	526	80	101	238	1406	1313	3087
MAR-ECM	D	110	110	653	107	89	224	1424	1389	3250
MAR-ECM	I	86	89	523	77	99	249	1408	1310	3128
DNA	D	101	103	646	95	91	224	1446	1409	3206
DNA	I	79	79	520	78	99	239	1369	1316	3107

3.2 Phases

We quantified the phases of gaps produced by different aligners and COATi models (Tab. S3 and S4). Phase 1 gaps begin after the 1st position in a codon in the reference sequence, phase 2 gaps begin after the 2nd position in a codon, and phase 3 begin after the 3rd position in a codon (i.e. between codons). Phase 3 gaps are also known as phase 0 gaps.

Note that ClustalΩ gaps with phases of 1 and 2 are created by how our wrapper script handles DNA sequences with lengths that are not multiple of 3.

Table S3: Number of gaps introduced by each method separated by phase.

	Gap Phases		
Method	1	2	3
COATi	8574	8342	24339
COATi-rev	7760	7714	24704
ClustalΩ	8	6	30986
MACSE	997	1072	37384
MAFFT	4560	5293	22271
PRANK	0	0	31262

Table S4: Number of gaps introduced by different COATi models separated by phase.

	Gap Phases		
Model	1	2	3
TRI-MG	8574	8342	24339
TRI-ECM	7920	9648	24025
MAR-MG	7918	7611	25557
MAR-ECM	7893	7774	25578
DNA	8179	8064	24866

3.3 Gapiness

We quantified the gapiness of different aligners and COATi models (Tab. S5 and S6). Here we define gapiness as the tendency of an aligner to introduce gaps into an alignment, and we measure it by calculating the percentage of nucleotides aligned with a gap. Note that this is different than the percent of columns that contain a gap.

Table S5: Average gapiness of alignments separated by method.

COATi	COATi-rev	ClustalΩ	MACSE	MAFFT	PRANK
3.28%	3.25%	2.52%	2.54%	2.52%	3.52%

Table S6: Average gapiness of alignments separated by model.

TRI-MG	TRI-ECM	MAR-MG	MAR-ECM	DNA
3.28%	3.30%	3.28%	3.26%	3.27%