Supplemental Material: Incorporating Intergenic Regions into Reversal and Transposition Distances with Indels

Alexsandro Oliveira Alexandrino*‡, Andre Rodrigues Oliveira‡, Ulisses Dias§, and Zanoni Dias‡

‡*Institute of Computing, University of Campinas*
1251 Albert Einstein Ave., 13083-852 Campinas, São Paulo, Brazil
{*alexsandro,andrero,zanoni*}*@ic.unicamp.br*

§*School of Technology, University of Campinas*
1888 Paschoal Marmo St., 13484-332 Limeira, São Paulo, Brazil
ulisses@ft.unicamp.br

1. Examples of Genome Representation, Rearrangements, and Breakpoints

In the following, we exemplify important concepts defined throughout the paper. Figure 1 presents two genomes and maps them into an instance for the rearrangement distance problems. Figures 2, 3, 4, and 5 present how a reversal, transposition, insertion, and deletion, respectively, affect a genome. At last, Example 1 shows the intergenic breakpoints and strips for an instance of the Reversals and Indels Distance problem.

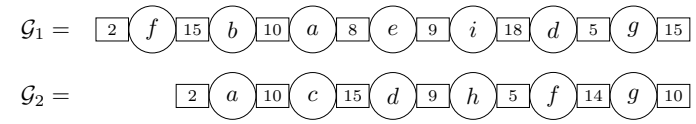

Fig. 1. Genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ have their genes and intergenic regions sizes represented, respectively, by letters inside circles and numbers inside rectangles. We represent $\mathcal{G}_2 = (\iota^n, \breve{\iota}^n)$, where $n = 6$ and $\breve{\iota}^n = (2, 10, 15, 9, 5, 14, 10)$, and $\mathcal{G}_1 = (A, \breve{A})$, where $A = (5\ \alpha\ 1\ \alpha\ 3\ 6)$ and $\breve{A} = (2, 15, 10, 8, 18, 5, 15)$. Observe that, in this example, genes e and i exist in $\mathcal{G}_1$ but not in $\mathcal{G}_2$, so the segment from $\mathcal{G}_1$ that goes from e to i along with the intergenic region of size 9 are all represented by only one element α in A.

*Corresponding author.

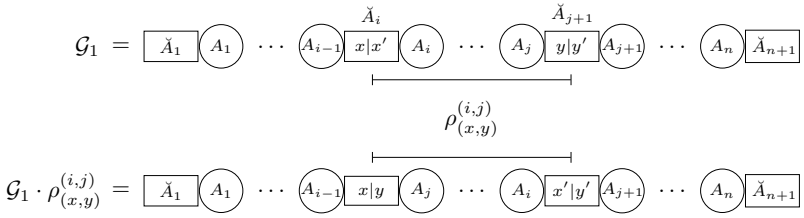

Fig. 2. Example of how an intergenic reversal $\rho^{(i,j)}_{(x,y)}$ affects $\mathcal{G}_1 = (A, \breve{A})$.

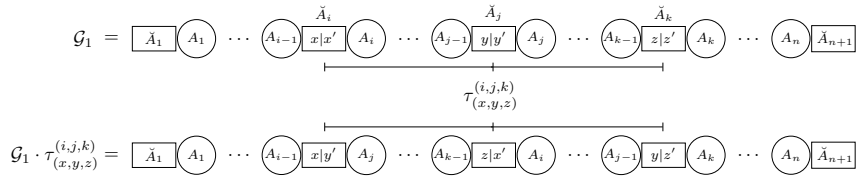

Fig. 3. Example of how an intergenic transposition $\tau^{(i,j,k)}_{(x,y,z)}$ affects $\mathcal{G}_1 = (A, \breve{A})$.

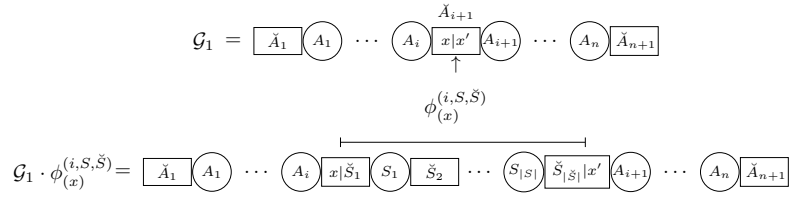

Fig. 4. Example of how an insertion $\phi^{(i,S,\breve{S})}_{(x)}$ affects $\mathcal{G}_1 = (A, \breve{A})$.

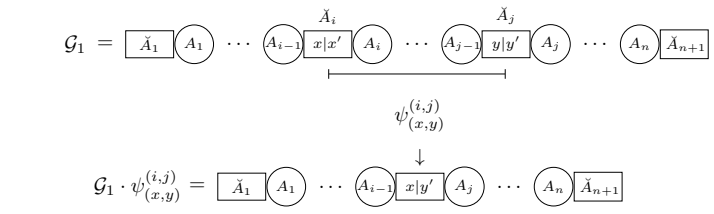

Fig. 5. Example of how a deletion $\psi^{(i,j)}_{(x,y)}$ affects $\mathcal{G}_1 = (A, \breve{A})$.

Example 1. Consider $\mathcal{G}_1 = (A, \breve{A})$ and $\mathcal{G}_2 = (\iota^n, \breve{\iota}^n)$, where $n = 6$, A is the extended string (0 5 α 3 2 1 6 7), $\breve{A} = (2, 15, 10, 8, 10, 18, 10)$, and $\breve{\iota}^n = (2, 10, 15, 9, 5, 14, 10)$. For the model $\mathcal{M}_r$, we have the following intergenic breakpoints: $(0, 5)$, $(5, \alpha)$, $(\alpha, 3)$, $(3, 2)$, and $(1, 6)$. The intergenic breakpoint $(3, 2)$ is undercharged. For the model $\mathcal{M}_r$, the string A has the following strips: $(0), (5)$, (α), (3), (2 1), and (6 7).

2. Experimental Results on Simulated Data

We performed experiments on simulated data to evaluate the algorithms in practical scenarios.

Given three parameters n, k, and $\mathcal{M}$, where n is the size of the target string ι^n, k is the number of rearrangements applied of each type (reversals or transpositions, insertions, and deletions), and $\mathcal{M}$ is the rearrangement model, we create a random instance as follows:

- We create the target genome $\mathcal{G}_2 = (\iota^n, \breve{\iota}^n)$, where $\breve{\iota}^n$ is a list of $n+1$ numbers chosen from a random uniform distribution in the interval $[0, 100]$.
- Build the genome $\mathcal{G}_1$ by applying on $\mathcal{G}_2$:
  - k reversals, if $\mathcal{M}$ contains reversals and does not contain transpositions (i.e., $\mathcal{M} = \mathcal{M}_r$);
  - k transposition, if $\mathcal{M}$ contains transpositions and does not contain reversals (i.e., $\mathcal{M} = \mathcal{M}_t$);
  - k reversals or transpositions, if $\mathcal{M}$ contains transpositions and reversals (i.e., $\mathcal{M} = \mathcal{M}_{rt}$). In this case, each operation has a 50% chance of being a reversal or a transposition.
- Apply k deletions on $\mathcal{G}_1$;
- Apply k insertion on $\mathcal{G}_1$.

Each rearrangement applied has its parameters chosen from a random uniform distribution in the set of all valid values for the parameters.

Our dataset is divided into sets, grouped by the parameters n, k, and $\mathcal{M}$. Each set $DS^{\mathcal{M}}_{n,k}$ has 1000 random instances created using the parameters n, k, and $\mathcal{M}$, for $n \in \{50, 100, \ldots, 500\}$, $k \in \{0.1n, 0.5n, n\}$, and $\mathcal{M} \in \{\mathcal{M}_r, \mathcal{M}_t, \mathcal{M}_{rt}\}$. The code for the algorithms implemented and the datasets used in this study are available in the following public repository: https://github.com/compbiogroup/unsigned-rearrangements-indels-distance-with-intergenic-regions.

Tables 1, 2, and 3 show the results obtained considering the 4-approximation algorithm for Reversals and Indels Distance, the 4.5-approximation algorithm for Transpositions and Indels Distance, and the 6-approximation algorithm for Reversals, Transpositions, and Indels Distance, respectively. We used the appropriate sets of random instances for each problem, considering the respective rearrangement model.

In all tables, the first and second columns (n and k) are the parameters used to create the set of random instances $DS^{\mathcal{M}}_{n,k}$, where $\mathcal{M}$ is the model corresponding to each algorithm. Each line of these tables has data for the set $DS^{\mathcal{M}}_{n,k}$. The third to fifth columns present the minimum, average, and maximum number of operations used in the solutions, respectively. The sixth to eighth columns present the minimum, average, and maximum values for the practical approximation factor, respectively. The practical approximation factor for an instance is the number of operations used by the algorithm divided by the corresponding lower bound for that instance. Note that each problem has its specific lower bound.

Overall, the three algorithms presented a practical approximation factor considerably lower than the theoretical approximation factor. For a fixed value of n, we observe that the average approximation factor increases as the value of k increases. As expected, the number of operations applied by the algorithms increases as the value of k increases.

Considering both the Reversals and Indels Distance and the Reversals, Transpositions, and Indels Distance, the average approximation factor increases as the size of the target genome (parameter n) increases, until n reaches 150, and it slightly changes for the values of n in the interval $[150, 500]$.

For the Reversals and Indels Distance, the average practical approximation factor was between 1.66 and 2.01. For the Reversals, Transpositions, and Indels Distance, the average practical approximation factor was between 2.42 and 3.03.

Considering the Transpositions and Indels Distance, the average approximation factor increases as the size of the target genome (parameter n) increases. The average practical approximation factor was between 2.43 and 3.27.

Table 1. Experimental results considering the 4-approximation algorithm presented for the Reversals and Indels Distance problem. Each row presents results for this algorithm considering the set $DS_{n,k}^{\mathcal{M}_r}$ as input.

n	k	**Operations**			**Approximation Factor**		
		Minimum	**Average**	**Maximum**	**Minimum**	**Average**	**Maximum**
50	5	3	4.91	6	1.33	1.66	2.00
50	25	25	33.95	43	1.67	1.92	2.18
50	50	41	52.32	62	1.86	1.99	2.15
100	10	10	14.71	19	1.43	1.70	2.25
100	50	55	67.29	78	1.78	1.93	2.12
100	100	91	105.81	120	1.92	2.00	2.10
150	15	20	24.52	30	1.53	1.73	2.07
150	75	88	104.01	121	1.80	1.95	2.09
150	150	142	158.82	173	1.95	2.01	2.07
200	20	22	29.55	35	1.53	1.73	2.06
200	100	119	137.60	152	1.81	1.95	2.07
200	200	191	210.81	232	1.94	2.01	2.06
250	25	33	39.31	46	1.57	1.74	2.00
250	125	152	171.23	194	1.85	1.95	2.09
250	250	245	263.88	286	1.95	2.01	2.07
300	30	42	49.08	55	1.59	1.75	1.96
300	150	189	208.04	232	1.84	1.95	2.04
300	300	290	317.18	338	1.96	2.01	2.06
350	35	47	54.05	62	1.61	1.74	2.00
350	175	223	241.62	267	1.87	1.95	2.04
350	350	346	369.08	396	1.96	2.01	2.06
400	40	56	63.88	73	1.61	1.75	1.94
400	200	253	275.30	298	1.85	1.95	2.03
400	400	398	422.68	448	1.97	2.01	2.05
450	45	64	73.54	83	1.62	1.75	2.00
450	225	285	311.99	335	1.86	1.95	2.02
450	450	446	476.64	508	1.97	2.01	2.05
500	50	69	78.47	87	1.62	1.74	1.91
500	250	311	345.38	373	1.88	1.95	2.03
500	500	498	527.72	558	1.97	2.01	2.04

Table 2. Experimental results considering the 4.5-approximation algorithm for the Transpositions and Indels Distance problem. Each row presents results for this algorithm considering the set $DS_{n,k}^{\mathcal{M}_t}$ as input.

n	k	**Operations**			**Approximation Factor**		
		Minimum	**Average**	**Maximum**	**Minimum**	**Average**	**Maximum**
50	5	3	6.70	7	1.33	2.43	3.50
50	25	29	39.18	48	2.62	3.04	3.42
50	50	47	56.73	66	2.82	3.09	3.41
100	10	14	19.66	23	2.43	2.90	3.29
100	50	66	79.27	95	2.92	3.14	3.36
100	100	100	116.52	129	2.97	3.16	3.38
150	15	26	32.69	38	2.70	2.98	3.40
150	75	102	123.16	138	3.00	3.19	3.37
150	150	159	176.08	192	3.05	3.19	3.36
200	20	31	39.74	46	2.69	3.01	3.36
200	100	143	164.29	185	3.06	3.21	3.37
200	200	214	235.18	256	3.09	3.21	3.34
250	25	43	53.12	60	2.78	3.05	3.35
250	125	181	205.14	227	3.10	3.23	3.36
250	250	273	295.39	327	3.10	3.22	3.33
300	30	57	66.51	74	2.86	3.08	3.38
300	150	228	249.65	271	3.12	3.24	3.36
300	300	326	355.81	382	3.13	3.23	3.33
350	35	64	73.84	82	2.91	3.10	3.35
350	175	266	291.21	316	3.15	3.25	3.36
350	350	391	414.90	443	3.14	3.24	3.34
400	40	73	87.38	96	2.83	3.13	3.32
400	200	303	331.99	369	3.16	3.26	3.36
400	400	451	475.34	505	3.14	3.24	3.34
450	45	88	101.27	113	2.97	3.14	3.33
450	225	352	377.03	411	3.14	3.26	3.35
450	450	508	535.65	566	3.15	3.24	3.33
500	50	97	108.62	119	2.97	3.15	3.32
500	250	385	418.31	448	3.17	3.27	3.38
500	500	562	595.30	627	3.17	3.25	3.32

Table 3. Experimental results considering the 6-approximation algorithm for the Reversals, Transpositions, and Indels Distance. Each row presents results for this algorithm considering the set $DS_{n,k}^{\mathcal{M}_{rt}}$ as input.

n	k	**Operations**			**Approximation Factor**		
		Minimum	**Average**	**Maximum**	**Minimum**	**Average**	**Maximum**
50	5	3	5.78	8	1.50	2.42	3.50
50	25	26	36.26	44	2.50	2.92	3.36
50	50	44	53.87	62	2.78	2.98	3.28
100	10	11	16.93	23	2.00	2.65	3.50
100	50	57	72.15	85	2.71	2.96	3.22
100	100	96	108.67	122	2.89	3.00	3.15
150	15	22	28.09	36	2.27	2.72	3.30
150	75	94	110.99	128	2.79	2.97	3.16
150	150	149	163.50	179	2.92	3.01	3.13
200	20	25	33.50	43	2.33	2.71	3.31
200	100	126	146.86	165	2.82	2.98	3.14
200	200	191	216.42	234	2.93	3.02	3.12
250	25	37	44.55	55	2.38	2.73	3.13
250	125	166	183.01	207	2.87	2.98	3.15
250	250	251	271.58	291	2.93	3.02	3.12
300	30	45	55.46	66	2.40	2.73	3.11
300	150	197	221.84	243	2.86	2.98	3.12
300	300	304	326.24	349	2.97	3.02	3.10
350	35	49	61.01	70	2.45	2.73	3.09
350	175	236	257.60	286	2.88	2.99	3.09
350	350	358	380.11	409	2.98	3.02	3.11
400	40	61	71.89	84	2.48	2.73	3.08
400	200	268	293.89	319	2.88	2.99	3.12
400	400	406	434.65	461	2.97	3.02	3.08
450	45	72	82.62	94	2.50	2.73	3.07
450	225	299	332.71	359	2.88	2.99	3.08
450	450	464	490.05	515	2.97	3.03	3.09
500	50	75	88.20	99	2.48	2.73	3.00
500	250	341	368.73	397	2.89	2.99	3.10
500	500	514	543.52	574	2.98	3.03	3.08