Table of Content:

1.) Project Information:

In the first assignment, we computed the rates of substitutions and indels between human chromosome 22 and chimpanzee and mouse genomes. That serves as an initial experiment to study divergence between mammalian genomes. In this assignment, we will study divergence among more species and with more details.

We are given pairwise alignments between human and several other genomes, and an annotation file of human genes. We need to compute the substitution rate, ti /tv ratio, gap rate and length frequencies for different categories of human positions: intergenic regions, intron regions, exons, coding regions, 5' UTRs, 3' UTRs, and promotors (considering 100, 200, 400, and 800 bases before first exon.) You also need to interpret your results.

2.) Introduction:

So far we have learned what is Genome in class. Genome is encoded into four types of DNA bases and we use the first letter of each to represent them. The four Alphabets are A, C, G and T. So, genomes are long strings containing these four alphabets.

We have learnt that all species have experienced a long history of evolution. And couple of these evolutionary events have occurred because of insertion/deletion(indel; as we cannot establish for a fact whether there was an insertion or deletion and hence we call it indel) substitution and few more like duplication, translocation and transposition. But in this project we are just concentrate on substitution.

Substitution is a type where one base is transformed to another (A – G, A – T, A –C, G – A, G – C, G – T, C – A, C – G, C – T, T – A, T – C and T – G). There are two types in substitution:

Transition: When substitution between A – G and C – T occurs.

Transversion: Rest all the types of substitution.

Also we have learnt that an indel is often referred as a gap and it is represented using '-'. A sequence of contiguous '-' characters is counted as one gap. The gap length refers to the number of '-' characters in the gap.

Genes are linearly arranged in the genome, separated by intergenic regions. Each gene has one or more exons, which will be part of a mature RNA product.

Regions	Meaning
Exons region	Between exonStart and exonEnd
Coding region	Exon between cdStart and cdEnd
5' UTRs	Between txStart and cdStart
3' UTRs	Between cdsEnd and txEnd
Promotors	Promoter length before txStart
Intron regions	The region between two exons, i.e., between last txEnd and next txStart.
Intergenic regions	Position between genes

3.) Programming Language Used:

For this project I have used C++ as my programming language. I took around 5 hours to code and around 7 - 8 hours to debug and modify. And around another 4 hours to write the report.

Used the Command g++ -o <execution file name> <program name.cc> to compile the program and ./<execution file name> to execute

4.) System Information Where The Code Was Executed:

The code was executed in Ubuntu 18.04. The System Specification is as follows:

a.) Processor: Intel core i5 7 th Generation, Quad Core Processor
b.) RAM: 8 GB Ram
c.) Hard Disk Size: 1 TB

5.) Findings and Observations:

a.) Comparison between hg19knowngene file and human-chimpanzee gene file.

For base 100:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	12634696	10904008	8678170	11468682	2218966	5894530	99815
Mis-Match Count	229029	152166	135812	157809	29574	100595	1486
Transitions Count	154494	106112	92736	110147	20413	68288	982
Transversions Count	74535	46054	43076	47662	9161	32307	504
Gap count	25119	17924	15086	18211	3895	10904	179
Substitution Rate	0.0178	0.01376	0.01541	0.01357	0.01315	0.01678	0.01467
ti/tv ratio	2.073	2.304	2.153	2.311	2.228	2.114	1.948
Gap Rate	0.001949	0.001619	0.001709	0001564	0.001729	0.001816	0.01764

For base 200:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	12591734	10883155	8646619	11445770	2207374	5876630	195181
Mis-Match Count	228326	151939	135323	157564	29414	100284	2905
Transitions Count	154037	105955	92405	109977	20314	68069	1927
Transversions Count	74289	45984	42918	47587	9100	32215	978
Gap count	25038	17893	15040	18180	3877	10876	337
Substitution Rate	0.01781	0.01377	0.01541	0.01358	0.01315	0.01678	0.01467
ti/tv ratio	2.073	2.304	2.153	2.311	2.232	2.113	1.97
Gap Rate	0.001949	0.001619	0.00171	0.001564	0.00173	0.001816	0.001698

For base 400:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	12509994	10841481	8585173	11399931	2186058	5840665	380041
Mis-Match Count	226932	151293	134434	156877	29144	99706	5834
Transitions Count	153173	105486	91810	109482	20144	67680	3855
Transversions Count	73759	45807	42624	47395	9010	32026	1979
Gap count	28476	17825	14937	18109	3832	10821	670
Substitution Rate	0.01782	0.01376	0.01542	0.01357	0.01316	0.01678	0.01512
ti/tv ratio	2.077	2.303	2.154	2.31	2.235	2.113	1.948
Gap Rate	0.001949	0.001619	0.00171	0.001565	0.001727	0.001818	0.001733

For base 800:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	12357926	10756677	8468470	11306152	2148728	5770267	733616
Mis-Match Count	224227	150105	132642	155600	28661	98486	11519
Transitions Count	151375	104641	90600	108576	19799	66866	7708
Transversions Count	72852	45464	42042	47024	8862	31620	3811
Gap count	24570	17658	14742	17935	3778	10687	1338
Substitution Rate	0.01782	0.01376	0.01542	0.01358	0.01316	0.01678	0.01546
ti/tv ratio	2.078	2.302	2.155	2.309	2.234	2.115	2.023
Gap Rate	0.01949	0.001616	0.001711	0.001562	0.001732	0.001818	0.001792

For base 1000:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	12285424	10714371	8411990	11258882	2131645	5735834	904904
Mis-Match Count	223045	149561	131854	154984	28448	97983	14033
Transitions Count	150588	104269	90051	108163	19647	66510	9416
Transversions Count	72457	45292	41803	46821	8801	31473	4617
Gap count	24412	17583	14646	17859	3746	10624	1667
Substitution Rate	0.01783	0.01377	0.01543	0.01358	0.01317	0.0168	0.01527
ti/tv ratio	2.078	2.302	2.154	2.31	2.232	2.113	2.039
Gap Rate	0.001948	0.001616	0.001711	0.001562	0.001731	0.001818	0.001811

So from the above findings we can safely conclude that as the bases before first exon increases I.e from 100 to 200 to 400 to 800 to 1000, the match count, mismatch count, transition count, transversion count keeps decreasing for all regions(Intergenic, Intron, Exon, Coding, 5'UTR, 3'UTR) whereas it increases only for the promotor region.

Similarly even the Gap count keeps decreasing for all regions except for promotor which keeps increasing. But I did however observe one irregularity where the gap count decreases in the integenic region when considered for base 100 and 200, but increase a bit for base 400 and then again decreases for base 800 and 1000.

When I did check the substitution Rate, ti/tv ration and Gap rate there is minute change in values for different bases, but the difference is very minute.

b.) Comparison between hg19knowngene file and human-mouse gene file:

For base 100:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	2319014	2765730	2364267	3202354	593630	1334013	43043
Mis-Match Count	1105674	1297469	952073	1399828	252201	597513	18203
Transitions Count	652906	774733	566401	837978	150072	353084	10606
Transversions Count	452768	522736	385672	561850	102129	244429	7597
Gap count	122573	152273	103885	156853	31110	68195	2205
Substitution Rate	0.3229	0.3193	0.2871	0.3042	0.2982	0.3093	0.2972
ti/tv ratio	1.442	1.482	1.469	1.491	1.469	1.445	1.396
Gap Rate	0.03455	0.03612	0.03037	0.03296	0.03548	0.0341	0.03475

For base 200:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	2303412	2757277	2351258	3192468	588834	1327233	80107
Mis-Match Count	1098643	1293601	946729	1395634	250144	594552	34446
Transitions Count	648897	772482	563275	835518	148870	351369	19992
Transversions Count	449746	521119	383454	560116	101274	243183	14454
Gap count	121690	151797	103288	156358	30869	67858	4161
Substitution Rate	0.3229	0.3193	0.2871	0.3042	0.2982	0.3094	0.3007
ti/tv ratio	1.443	1.482	1.469	1.492	1.47	1.445	1.383

Gap Rate	0.03453	0.03612	0.03037	0.03296	0.03549	0.03411	0.03505

For base 400:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	2278933	2742339	2326688	3174744	580371	1313912	144094
Mis-Match Count	1087265	1286821	936730	1388173	246413	588965	62603
Transitions Count	642367	768523	557381	831140	146660	348104	36375
Transversions Count	444898	518298	379349	557033	99753	240861	26228
Gap count	120267	150984	102164	155507	30421	67220	7521
Substitution Rate	0.323	0.3194	0.287	0.3042	0.298	0.3095	0.3029
ti/tv ratio	1.444	1.483	1.469	1.492	1.47	1.445	1.387
Gap Rate	0.0345	0.03612	0.03036	0.03296	0.03549	0.03412	0.03511

For base 800:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	2239945	2714465	2281899	3140678	567824	1287862	255745
Mis-Match Count	1068808	1273986	918556	1373740	241005	577797	112069
Transitions Count	631500	760940	546631	822552	143441	341578	65575
Transversions Count	437308	513046	371925	551188	97564	236219	46494
Gap count	117931	149360	100084	153810	29746	65888	13561
Substitution Rate	0.323	0.3194	0.287	0.3043	0.298	0.3097	0.3047

ti/tv ratio	1.444	1.483	1.47	1.492	1.47	1.446	1.41
Gap Rate	0.03442	0.0361	0.03032	0.03295	0.03547	0.03411	0.03556

For base 1000:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	2223400	2702152	2261803	3124479	562614	1276862	304699
Mis-Match Count	1061169	1268112	910633	1366931	238759	573055	133505
Transitions Count	626938	757456	541987	818483	142123	338827	78265
Transversions Count	434231	510656	368646	548448	96626	234228	55240
Gap count	116956	148669	99166	153064	29481	65290	16145
Substitution Rate	0.3231	0.3194	0.287	0.3043	0.2979	0.3098	0.3047
ti/tv ratio	1.444	1.483	1.47	1.492	1.471	1.447	1.417
Gap Rate	0.03438	0.03609	0.03031	0.03296	0.03548	0.03409	0.03553

So, from the above findings we can safely conclude that as the bases before first exon increases I.e from 100 to 200 to 400 to 800 to 1000, the match count, mismatch count, transition count, transversion count keeps decreasing for all regions(Intergenic, Intron, Exon, Coding, 5'UTR, 3'UTR) whereas it increases only for the promotor region.

Similarly even the Gap count keeps decreasing for all regions except for promotor which keeps increasing.

When I did check the substitution Rate, ti/tv ration and Gap rate there is minute change in values for different bases, but the difference is very minute.

c.) Comparison between hg19knowngene and human-dog gene file:

For base 100:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	4233043	4675704	3690735	5122787	929099	2314553	56165
Mis-Match Count	1564294	1635334	1235802	1729751	310876	830509	19735
Transitions Count	951893	1005085	747452	1062566	190042	499929	11500
Transversions Count	612401	630249	488350	667185	120834	330580	8235
Gap count	167702	186268	135910	192371	36960	92847	2354
Substitution Rate	0.2698	0.2591	0.2508	0.2524	0.2507	0.2641	0.26
ti/tv ratio	1.554	1.595	1.531	1.593	1.573	1.512	1.396
Gap Rate	0.02811	0.02867	0.02685	0.02731	0.02894	0.02867	0.03008

For base 200:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	4210825	4664538	3673951	5110095	922821	2305573	106333
Mis-Match Count	1556150	1631306	1229807	1725326	308631	827156	37902
Transitions Count	947198	1002608	743955	1059878	188722	497963	22169
Transversions Count	608952	628698	485852	665448	119909	329193	15733
Gap count	166722	185814	135207	191896	36676	92499	4491
Substitution Rate	0.2698	0.2591	0.2508	0.2524	0.2506	0.264	0.2628
ti/tv ratio	1.555	1.595	1.531	1.593	1.574	1.513	1.409

Gap Rate	0.0281	0.02867	0.02683	0.02731	0.02892	0.02866	0.0302

For base 400:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	4174142	4643191	3640908	5085398	911693	2287008	197406
Mis-Match Count	1542345	1623427	1218791	1716693	304920	820605	70602
Transitions Count	939149	997834	737403	1054687	186451	494099	41544
Transversions Count	603196	625593	481388	662006	118469	326506	29058
Gap count	165087	184920	134006	190961	36246	91719	8221
Substitution Rate	0.2698	0.2591	0.2508	0.2524	0.2506	0.2641	0.2634
ti/tv ratio	1.557	1.595	1.532	1.593	1.574	1.513	1.43
Gap Rate	0.02807	0.02866	0.02683	0.02731	0.02893	0.02867	0.02976

For base 800:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	4112273	4601358	3581997	5036865	893930	2252560	360019
Mis-Match Count	1519107	1608447	1199525	1700085	298975	808912	128086
Transitions Count	925288	988696	725944	1044611	182902	487127	76002
Transversions Count	593819	619751	473581	655474	116073	321785	52084
Gap count	162210	183200	131681	189130	35498	90253	15143
Substitution Rate	0.2698	0.259	0.2509	0.2524	0.2506	0.2642	0.2624

ti/tv ratio	1.558	1.595	1.533	1.594	1.576	1.514	1.459
Gap Rate	0.028	0.02866	0.0268	0.02731	0.0289	0.02864	0.03009

For base 1000:

Regions / Cumulative Count	Intergenic	Intron	Exon	Coding	UTR5	UTR3	Promotor
Match Count	4084009	4581327	3553924	5012883	885734	2236634	436387
Mis-Match Count	1508453	1601608	1190275	1692307	296204	803372	154829
Transitions Count	918885	984458	720369	1039808	181223	483796	92218
Transversions Count	589568	617150	469906	652499	114981	319576	62611
Gap count	160954	182351	130577	188219	35142	89567	18352
Substitution Rate	0.2697	0.259	0.2509	0.2524	0.2506	0.2643	0.2619
ti/tv ratio	1.559	1.595	1.533	1.594	1.576	1.514	1.473
Gap Rate	0.02798	0.02865	0.02679	0.0273	0.02877	0.02862	0.03011

So, from the above findings we can safely conclude that as the bases before first exon increases I.e from 100 to 200 to 400 to 800 to 1000, the match count, mismatch count, transition count, transversion count keeps decreasing for all regions(Intergenic, Intron, Exon, Coding, 5'UTR, 3'UTR) whereas it increases only for the promotor region.

Similarly even the Gap count keeps decreasing for all regions except for promotor which keeps increasing.

When I did check the substitution Rate, ti/tv ration and Gap rate there is minute change in values for different bases, but the difference is very minute.

I have attached the source code file and executable file to Blackboard.