Construction of Suffix Trees for String Matching

Ilya Levner

March 16, 2004

Abstract

In this assignment suffix trees are used to match substrings. The Y chromosomes of human and mouse DNA are compared for string matches.

Keywords: BioInformatics, Suffix Trees, String Matching.

Contents

1 Data Gathering

DNA sequences from human and mouse Y chromosomes were obtained from the NCBI database. Respectively, for human and mouse, the ftp sites were `ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_Y/` and `ftp://ftp.ncbi.nih.gov/genbank/genomes/M_musculus/CHR_Y/` and the extracted files were *hs_chrY.gbk.gz* and *mm_chrY.gbk.gz*. The raw DNA data was first extracted from the files and all instances of known genes removed. The removed genes can be found in the appendix, and totalled 237 for the human DNA and 126 for the mouse DNA. The raw DNA sequences respectively contained 25,142,184 and 19096172 bases for human and mouse. After gene sequence deletion to processed DNA sequences contained 20,015,977 and 13,409,870 bases in 237 and 126 intron files.

A copy of this report and all relevant files can be found at `http://www.cs.ualberta.ca/~ilya/courses/c606-BioInformatics/a2/` Files in the folders `code/human` and `code/mouse` in the form out**i**.introns_**j**.txt contain introns extracted form the NCBI files (indices i,j are used for internal processing). `code/SuffixTree[Main].[h,cc]` and the `code/Makefile` are the key components build and used for the construction of the suffix trees.

2 Human Y Chromosome

The suffix tree for the human y chromosome was build out of 237 individual sequences. The number of within species sequences greater than 21 characters was 282990. The longest common sequence was 1782 characters. In total 95,688 unique sequences were found and are stored in the file named `human_common_subsequences.txt`. The file contains entries in the following format:

```
PathLength: 21 "AAAAAAAAAAAAAAAAAAAAAA" (10,93881)(68,34705)
```

where `PathLength` specifies the length of the string. The string it self is printed in quotes next. The pairs following the common substring denote (*stringindex*, *startposition*) for each intron sequence that this substring occurred in. String index refers to the order in which the intron files were processed, while starting position refers to the start of the common substring within the intron file. By cross-referencing the `human_suffixtree_results.txt` file, the corresponding file index can be found. Figure 1 shows the frequency distribution of common subsequences lengths (for sequences over 21 bases in length). Figure 2 shows arity frequency distribution.

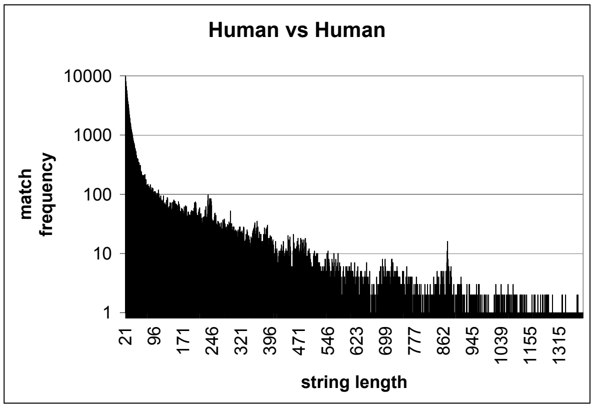

Figure 1: Match Frequency (log scale) versus string length for human introns on the y chromosome.

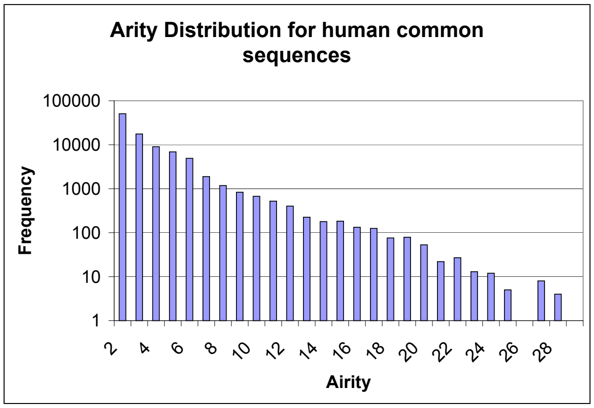

Figure 2: Frequency (log scale) versus string match arity (i.e., the number of introns **all** having a common subsequence) for human introns on the y chromosome.

3 Mouse Y Chromosome

The suffix tree for the mouse y chromosome was build out of 126 individual sequences. In contrast to the human y chromosome, number of within species sequences greater than 21 characters was over 1.4 million. The longest common sequence was 14221 characters. In total 205,290 unique sequences were found and are stored in the file named `mouse_common_subsequences.txt`. Figure 3 shows the frequency distribution of common subsequences lengths (for sequences over 21 bases in length). Figure 4 shows arity frequency distribution. Clearly, when Figures 2 and 4 are compared we see that mouse DNA has primarily binary sequence commonality, whereas human DNA has a large arity meaning that a common strand of DNA within human introns occurs much more frequently.

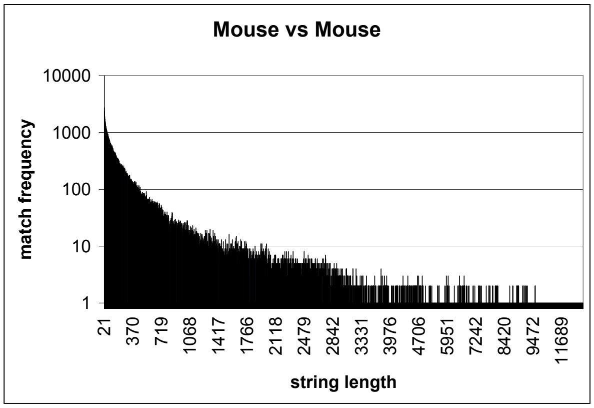

Figure 3: Match Frequency (log scale) versus string length for mouse introns on the y chromosome.

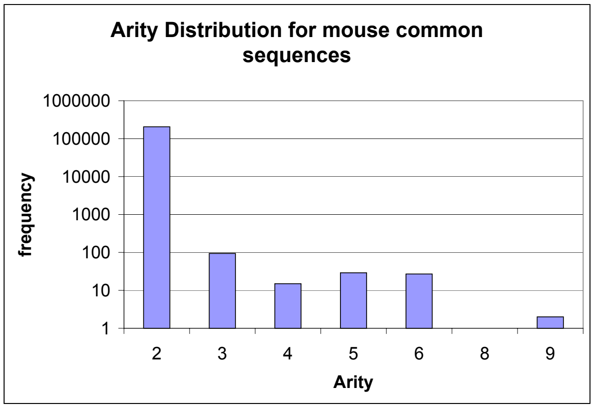

Figure 4: Frequency (log scale) versus string match arity (i.e., the number of introns **all** having a common subsequence) for mouse introns on the y chromosome.

4 Common Sequences

To see genetic commonalities between human and mouse species, pairwise general suffix trees ware created from both human and mouse intron sequences. Pairwise frequency distributions were attained by running 29,862 pairwise comparisons and selecting sequences common to both human and mouse introns. This was done due to the lack of memory needed to create a general suffix tree for all $237 + 126 = 363$ sequences, totaling $20,015,977 + 13,409,870 = 33,425,847 = 33MB$ characters. To see why, recall that each suffix has a leaf node, regardless of whether the characters are in a single sequence or are split across a number of files. This is due to the need to append the sequence with a *unique* terminal character. Each leaf node needs at least 10 bytes requiring 330 Mb just for the leaf nodes, the whole tree requires about 3 times that and thus needs about 1Gb. On top of that to collect statistics another 1Gb of memory is needed to store the C(v) numbers (see [2] for more information). Unfortunately, in order to retrieve the statistics, a DFS needs adequate stack depth to traverse the whole tree. Thus, while the tree can be build (needing only about 5 minutes, there was not enough memory to recursively traverse the tree and print out the common subsequences.

As a result, **pair wise** suffix trees were created.

A Human Genes

The following coding regions were used as a guide for extracting the remaining intron sequences from the human y chromosome.

```
complement(103052..104381) complement(<244436..>245964)
<279336..>310706 <566873..>574252 <643738..>645414
<654414..>669675 <676191..>700633 complement(<720024..>734934)
<854425..>856587 complement(<931734..>947149) 975114..1004969
1013976..1172911 complement(1113999..1116728) 1217593..1233206
complement(<1242581..>1260908) <1407605..>1410451
complement(1561079..1793366) <1599272..>1600250 2016263..2018720
complement(<2030761..>2031533) <2181579..>2181762
complement(2298472..2299208) 2837225..3046239 <2951769..>2954009
complement(<3189910..>3190970) complement(<3597890..>3705307)
<3800550..>3801324 5433707..5442968 complement(<5765917..>5768552)
complement(5824351..5829480) <5877734..>5878913 5970819..5972317
complement(<6005518..>6007419) complement(6018201..6019781)
6051884..6068309 complement(6126631..6129439)
complement(6131319..6133255) 6200589..6202685 6275008..6276967
complement(<6329937..>6345932) complement(6432138..6433936)
6574969..6580102 <6635902..>6638537 complement(<6702371..>6718480)
6740987..6750239 6795892..6797812 complement(<6831284..>6878300)
complement(6904735..6907263) <6923048..>6987561 6979103..6988418
complement(6980661..7022936) complement(<7060119..>7061715)
complement(7068609..7077857) <7097483..>7116502
complement(7121722..7127449) complement(7239869..7241986)
complement(7300647..7326637) <7576719..>7577985
complement(7704652..7752374) complement(<7730251..>7734083)
7816579..7839643 7846526..7855033 complement(7954641..7994159)
complement(8664889..8718449) complement(<8768911..>8769785)
8774946..8792375 <8809334..>8811282 <8926475..>8942121
8955998..8980253 <9060600..>9061762 complement(<9101062..>9106830)
complement(<9420571..>9421002) <9510482..>9518556
complement(9594369..9600783) complement(<9629645..>9636468)
<9667450..>9670060 <9692740..>9701458 9734134..9748539
complement(9782821..9793887) complement(9835922..9838589)
9895466..9898294 complement(37877..40681)
complement(189715..195528) complement(208776..209796)
250730..253524 complement(269733..270753) 288387..293680
complement(303459..309061) complement(309963..311413)
331594..334181 complement(394514..398946)
complement(424009..435030) 474893..477492
complement(490617..499077) complement(501446..505089)
524377..542269 complement(536949..558082) 584946..597038
<614948..>619570 complement(650413..652888) 653207..653965
complement(<689246..>691978) 694591..695755 <735339..>743635
complement(<806009..>819083) 820157..822040
complement(<850187..>852989) <939976..>940237
complement(974232..976091) <981049..>982896 complement(5376..6272)
60134..85472 154026..200371 512329..513320 <826159..>826931
855432..856388 <958947..>961189 complement(<1140963..>1143064)
complement(<2077357..>2079276) 2276573..3018571
complement(<2482714..>2484483) complement(<2613952..>2615327)
<2849443..>2850835 complement(<3069600..>3070124)
<3163857..>3165087 <3435144..>3436179 3522570..3525359
complement(3534680..3541867) <3542894..>3545626
complement(3580913..3582799) <3604535..>3619930 3725815..3734253
complement(4142265..4150374) <4177480..>4177814 4187033..4368030
<4226772..>4226962 complement(<4376745..>4382852) 4550320..4657895
complement(<4947815..>4954325) complement(<4963149..>4965940)
complement(<4989332..>5053052) 5081271..5087030
complement(5189766..5190844) complement(5267614..5268237)
complement(5346058..5396792) <5403470..>5420976
complement(<5454173..>5455514) complement(<5540523..>5555493)
5627431..5628283 complement(5790851..5794145)
complement(6059660..6093729) <6180564..>6192690
complement(6207269..6211217) <6310457..>6316335
complement(98615..112786) complement(<121934..>133858)
<167030..>200303 312966..341550 complement(363261..377661)
complement(<393141..>399471) 404054..407460
complement(491517..493595) 503390..509484
complement(<579260..>582715) complement(634312..636178)
671145..680311 685116..709356 complement(715407..716415)
complement(<775526..>808802) <841975..>849937 920827..923873
complement(<957143..>964509) <1244856..>1245789
complement(1324078..1393729) 1414254..1486071 1414269..1484720
1414275..1484431 <1574203..>1587418 complement(<1731723..>1743652)
complement(<1910508..>1912106) <1965644..>1992871
complement(2049092..2052133) complement(2056291..2057421)
complement(<2111517..>2113115) complement(2131448..2136343)
<2178021..>2189802 complement(2239950..2242735)
complement(2244598..2246534) <2369367..>2372083
complement(2378155..2381952) complement(<2400285..>2400690)
complement(2404688..2409552) complement(<2443299..>2444481)
complement(<2472147..>2475439) 2477675..2486555
complement(<2556757..>2557099) complement(<2590898..>2598254)
2764923..2801746 2878004..2879609 complement(2957796..3008114)
complement(3190781..3192386) complement(3224107..3470374)
<3472149..>3479514 <3513329..>3513671 <3594998..>3598291
3629136..3662509 3662544..3665515 3669566..3679283
3680221..3692358 complement(<3698343..>3701068)
complement(3768576..3771823) 3821473..3825846
complement(<3880664..>3902895) 4018588..4019927
complement(<4077506..>4104737) 4158265..4159862
complement(4267224..4268231) complement(<4342198..>4361028)
<4394218..>4406147 complement(<4550456..>4621111)
```

```
<4673487..>4673763 <4821948..>4822880
complement(<4867216..>4907188) <33734..>34661
complement(42255..53875) 48168..75199 complement(76605..86063)
complement(13592..66233) complement(77793..96744) 168754..169044
303695..338762 669572..673700 11080..57301
complement(60764..66713) 76655..90033 complement(77856..127452)
complement(137188..211719) 266235..277114 270067..317693
complement(416753..439858) complement(86164..100695)
206545..238056 290915..341037 67162..81828 180724..258346
complement(284726..296851) 296957..310193
```

B Mouse Genes

The following coding regions were used as a guide for extracting the remaining intron sequences from the mouse y chromosome.

```
408..25412 41347..184257 complement(15544..26252)
complement(<54788..>55348) 143135..143470
complement(174794..176222) 194612..195799 347234..347569
487688..513840 complement(768544..800521) 812307..824669
complement(881543..893084) complement(1021472..1045491)
1086896..1096526 complement(1442018..1468111) 1497008..1517461
complement(305828..306511) 465415..467781
complement(514614..515297) complement(591768..623105)
complement(727659..730006) complement(812190..817653)
900425..901116 complement(247281..249597)
complement(321947..324268) complement(351522..352310)
478648..478947 complement(561430..569303) 691472..692170
complement(894643..896963) 1027947..1054136 1159606..1161928
1235941..1236624 1427965..1428663 1574339..1575022
complement(1726869..1729049) complement(1840497..2078477)
complement(2113415..2114113) complement(2259156..2259839)
complement(2400170..2400868) complement(2548852..2592948)
2697271..2699846 complement(2905839..2931997) 3035275..3035973
3217859..3218557 complement(3356790..3567662) 4031087..4033403
4179959..4180658 complement(4206190..4206888)
complement(4358760..4360919) complement(4490448..4493332)
complement(4565578..4567930) 4696527..4698012
complement(4799515..4801811) complement(4873556..4875823)
4953020..4953703 5128389..5129087 complement(5164860..5214727)
5320098..5320795 5432338..5499305 complement(5517873..5518571)
5641652..5641951 complement(5743008..5743706) 5998630..5999328
6046201..6046884 complement(15618..15923) complement(75848..79229)
105473..143211 167506..193731 complement(254140..254838)
complement(368682..369380) complement(617974..618672)
```

```
691570..693891 887714..890297 954560..956886 1030200..1030883
complement(65701..68286) 108103..108787 complement(252433..253132)
complement(232281..232964) complement(330781..333123)
complement(408335..409008) complement(476980..477678)
complement(14205..33535) complement(73611..78399) 138421..194077
complement(523475..571593) complement(781558..784863)
complement(989606..990304) 1305434..1307696
complement(1333735..1365270) complement(1397368..1423640)
complement(1437171..1462675) 1595903..1596691 81478..82176
146475..148806 190188..221200 complement(243495..246875)
272772..311619 339366..340064 complement(405155..431319)
493077..493775 complement(611854..614174)
complement(677513..678211) 804231..830458 1084035..1084733
1151225..1153533 1230111..1230809 1334325..1335008
complement(1508036..1508718) complement(30564..86037)
complement(151325..154627) complement(220642..223228)
complement(10946..11629) complement(4893..9822) 70005..102143
<357305..>358526 complement(378117..392066) 25199..27557
complement(269896..342435) 351356..378497 658371..694463
complement(886451..925048) complement(1084767..1116984)
1124126..1186004 complement(1305075..1308399)
complement(58382..130581) complement(58468..130354) 151885..177396
<206797..>207480 231122..261558 101203..127643
complement(165287..168582) complement(232449..235032) 60448..61131
197923..198606 complement(464602..515249) 50753..55170
12543..38054 <67457..>68140 91..16192 82994..100974 Y;
complement(129052..133469)
```