BIMM 185 Lab Report Week 8

5/20/2017
Lihui Lu A99108553

Introduction

Last week we went through some example topics in comparative genomics and started with one example project of finding conserved gene paired across genomes to predict protein-protein interactions. We hypothesize that highly interacted proteins are likely to interact physically and close to each other. We again build an inference model to examine our hypothesis. To identify protein-protein interactions, different methods like Genes within the same operon units, gene neighbors, Protein fusions and Phylogenetic Profile are used. For the in-class exercise, we practiced implementing the gene neighbor method to detect conserved genes pairs A. tumefaciens.

Methods and Codes

Our goal is to find all the conserved gene pairs across E.coli and A.tumefaciens that are less than 5 genes away from each other. In order to do that, we first constructed the orthologs table using the blast results that we obtained weeks ago. A SQL view is constructed with 3 attributes, qid(gene_id of genes in E.coli), sid(gene_id of genes in A.tumefaciens), and rid(replicon id of A.tumefaciens replicons) in order to provide convenience for later query by replicon.

The sql command is below:

```
create view orthologs as select g1.gene_id as qid, g2.gene_id as sid, g2.replic
on_id as rid from (select b1max.qseqid as qseqid, b1max.sseqid as sseqid from
(select b.qseqid, b.sseqid, b.bitscore, b.qcovs, b.scov from blast_1 b,
    (select b1.qseqid,max(bitscore) as bitscore from blast_1 b1 group by b1.qse
qid) maxtable
    where b.qseqid = maxtable.qseqid and b.bitscore = maxtable.bitscore) b1max,

(select b.qseqid, b.sseqid, b.bitscore, b.qcovs, b.scov from blast_2 b,
    (select b2.qseqid,max(bitscore) as bitscore from blast_2 b2 group by b2.qse
```

```
qid) maxtable
    where b.qseqid = maxtable.qseqid and b.bitscore = maxtable.bitscore) b2max
where b1max.qseqid = b2max.sseqid and b1max.sseqid = b2max.qseqid order by b1ma
x.bitscore DESC) o1 inner join genes g1 on o1.qseqid = g1.accession inner join
genes g2 on o1.sseqid = g2.accession;
```

Then for each replicon in A.tumefaciens, directons on that replicon were queried out and stored in a list so that each gene has an index as its position on the replicon.

```
'''
This function gets all the directons of the given replicon
'''
def query_directons(conn, replicon):
    cur = conn.cursor()
    sql_statement = ("select genes.gene_id from genes inner join(select gene_id
, min(left_position) as left_position, max(right_position) as right_position fr
om exons group by gene_id) position on position.gene_id = genes.gene_id where g
enes.replicon_id = {replicon} order by left_position;".format(replicon=replicon
))
    cur.execute(sql_statement)
    result = cur.fetchall()
    #print(result)
    return list(result)
```

To gain the conserved genes(orthologs) between E.coli and the replicon, the previously constructed 'orthologs' view was used.

```
'''
This functions gets all the ortholog gene pairs in E.coli and the given replico
n
'''
def query_orthologs(conn, replicon):
    cur = conn.cursor()
    sql_statement = ("select * from orthologs where rid = {replicon}".format(re
plicon=replicon))
    cur.execute(sql_statement)
    result = cur.fetchall()
    return list(result)
```

After gaining all the data that we need, we iterate through all the genes in E.coli that are also conserved in A.tumefacians replicon, find its location, then iterate through all the genes that are

within 5 genes away from that gene, see if the paired gene is also conserved in A.tumefacians. Then find the position of the 2 conserved orthologs on A.tumefacians and calculate the distance. Output every gene pairs that satisfy our requirement in table with the following columns:
1. gene id of gene1 in E.coli
2. gene id of gene2 in E.coli
3. distance in E.coli
4. distance of orthologs in A. tumefaciens

Results

After went through all 4 replicons of A.tumefaciens, a total number of 745 conserved gene pairs were found.
Among the gene pairs, 605 gene pairs were found between E.coli and replicon 2; 126 were found between E.coli and replicon 3; 13 gene pairs were found between E.coli and replicon 4; and 1 gene pair was found between E.coli and replicon 5.

The full list of the conserved gene pairs will be put at the end of this lab report in the appendix section.

Discussion

Last week we learned the whole process from initializing a hypothesis to building a model to examine the hypothesis and use the hypothesis to predict results on unknown testing data. The inference model that we built last week was more complicated compare to the inference model that we built in order to predict operon units due to the collection of the True positive and True negative training sets. The model is more complicated also in the sense of incorperating multiple methods and intergrating scores resulted from different methods into one uniform scale.

Appendix

Resulting conserved gene pairs between E.coli and A.tumefaciens

```
13  14  1   1
24  25  1   1
50  51  1   1
```

```
50  53  3   3
51  53  2   2
64  65  1   1
64  66  2   2
65  66  1   1
66  71  5   5
70  72  2   1
76  77  1   1
79  80  1   1
79  82  3   3
79  83  4   4
79  84  5   5
80  82  2   2
80  83  3   3
80  84  4   4
80  85  5   5
82  83  1   1
82  84  2   2
82  85  3   3
82  86  4   4
82  87  5   5
83  84  1   1
83  85  2   2
83  86  3   3
83  87  4   4
83  88  5   5
84  85  1   1
84  86  2   2
84  87  3   3
84  88  4   4
84  89  5   5
85  86  1   1
85  87  2   2
85  88  3   3
85  89  4   4
86  87  1   1
86  88  2   2
86  89  3   3
87  88  1   1
87  89  2   2
88  89  1   1
88  90  2   5
89  90  1   4
89  91  2   5
90  91  1   1
```

```
90  92  2   2
90  93  3   3
90  94  4   4
91  92  1   1
91  93  2   2
91  94  3   3
92  93  1   1
92  94  2   2
93  94  1   1
123 124 1   1
138 141 3   5
152 156 4   1
164 165 1   1
164 166 2   2
164 167 3   3
164 169 5   4
165 166 1   1
165 167 2   2
165 169 4   3
165 170 5   4
166 167 1   1
166 169 3   2
166 170 4   3
166 171 5   4
167 169 2   1
167 170 3   2
167 171 4   3
167 172 5   4
169 170 1   1
169 171 2   2
169 172 3   3
169 174 5   4
170 171 1   1
170 172 2   2
170 174 4   3
170 175 5   4
171 172 1   1
171 174 3   2
171 175 4   3
171 176 5   4
172 174 2   1
172 175 3   2
172 176 4   3
172 177 5   5
174 175 1   1
```

```
174 176 2   2
174 177 3   4
175 176 1   1
175 177 2   3
176 177 1   2
233 234 1   1
310 311 1   1
398 399 1   1
400 401 1   1
406 407 1   1
406 408 2   4
406 409 3   5
407 408 1   3
407 409 2   4
408 409 1   1
413 415 2   2
422 423 1   1
422 424 2   2
422 425 3   3
423 424 1   1
423 425 2   2
424 425 1   1
430 431 1   1
430 432 2   2
430 433 3   3
431 432 1   1
431 433 2   2
432 433 1   1
443 444 1   1
443 445 2   1
444 445 1   2
454 455 1   1
462 463 1   1
462 464 2   3
463 464 1   2
486 487 1   1
486 488 2   2
487 488 1   1
633 634 1   1
633 636 3   2
634 636 2   1
654 655 1   1
654 656 2   2
654 657 3   3
654 658 4   4
```

```
655 656 1   1
655 657 2   2
655 658 3   3
656 657 1   1
656 658 2   2
657 658 1   1
710 712 2   2
710 713 3   3
712 713 1   1
712 716 4   5
713 716 3   4
713 717 4   5
714 715 1   1
714 716 2   2
714 717 3   1
715 716 1   3
715 717 2   2
716 717 1   1
743 745 2   3
743 746 3   2
743 747 4   1
745 746 1   1
745 747 2   2
746 747 1   1
750 751 1   1
766 767 1   1
836 837 1   1
836 838 2   2
836 839 3   3
837 838 1   1
837 839 2   2
838 839 1   1
863 864 1   1
990 991 1   1
990 992 2   2
990 993 3   3
990 994 4   4
990 995 5   5
991 992 1   1
991 993 2   2
991 994 3   3
991 995 4   4
992 993 1   1
992 994 2   2
992 995 3   3
```

```
993 994 1   1
993 995 2   2
994 995 1   1
1034    1035    1   4
1051    1052    1   1
1051    1056    5   3
1052    1056    4   2
1053    1054    1   5
1056    1057    1   4
1056    1058    2   2
1057    1058    1   2
1068    1069    1   1
1070    1071    1   1
1070    1072    2   2
1070    1073    3   3
1070    1075    5   4
1071    1072    1   1
1071    1073    2   2
1071    1075    4   3
1072    1073    1   1
1072    1075    3   2
1073    1075    2   1
1076    1078    2   3
1095    1096    1   1
1184    1185    1   3
1241    1242    1   1
1243    1244    1   1
1249    1253    4   5
1282    1286    4   1
1306    1307    1   1
1591    1592    1   1
1627    1630    3   1
1669    1670    1   1
1669    1671    2   2
1669    1672    3   3
1669    1673    4   2
1670    1671    1   1
1670    1672    2   2
1670    1673    3   3
1671    1672    1   1
1671    1673    2   4
1672    1673    1   5
1702    1703    1   1
1702    1705    3   3
1702    1706    4   4
```

```
1702    1707    5   5
1703    1705    2   2
1703    1706    3   3
1703    1707    4   4
1705    1706    1   1
1705    1707    2   2
1706    1707    1   1
1843    1844    1   2
1849    1850    1   1
1849    1851    2   2
1850    1851    1   1
1875    1877    2   2
1877    1881    4   1
1902    1903    1   1
1929    1933    4   4
1929    1934    5   5
1933    1934    1   1
1937    1938    1   2
1956    1957    1   1
1970    1971    1   1
2007    2008    1   2
2007    2009    2   3
2007    2010    3   4
2008    2009    1   1
2008    2010    2   2
2009    2010    1   1
2113    2114    1   1
2113    2115    2   2
2113    2116    3   3
2114    2115    1   1
2114    2116    2   2
2115    2116    1   1
2142    2143    1   1
2150    2151    1   1
2163    2164    1   1
2163    2165    2   2
2163    2166    3   3
2164    2165    1   1
2164    2166    2   2
2165    2166    1   1
2178    2180    2   1
2178    2181    3   2
2179    2183    4   2
2179    2184    5   3
2180    2181    1   1
```

```
2183    2184    1   1
2183    2185    2   2
2184    2185    1   1
2261    2262    1   1
2261    2263    2   2
2261    2264    3   3
2261    2265    4   4
2261    2266    5   5
2262    2263    1   1
2262    2264    2   2
2262    2265    3   3
2262    2266    4   4
2262    2267    5   5
2263    2264    1   1
2263    2265    2   2
2263    2266    3   3
2263    2267    4   4
2263    2268    5   5
2264    2265    1   1
2264    2266    2   2
2264    2267    3   3
2264    2268    4   4
2264    2269    5   5
2265    2266    1   1
2265    2267    2   2
2265    2268    3   3
2265    2269    4   4
2266    2267    1   1
2266    2268    2   2
2266    2269    3   3
2267    2268    1   1
2267    2269    2   2
2267    2271    4   5
2268    2269    1   1
2268    2271    3   4
2269    2271    2   3
2271    2272    1   3
2271    2273    2   4
2272    2273    1   1
2297    2298    1   1
2300    2301    1   1
2481    2482    1   1
2481    2483    2   3
2482    2483    1   4
2512    2516    4   1
```

```
2547    2550    3   3
2547    2551    4   2
2547    2552    5   1
2549    2550    1   3
2549    2551    2   4
2549    2552    3   5
2550    2551    1   1
2550    2552    2   2
2551    2552    1   1
2585    2586    1   2
2585    2587    2   3
2585    2588    3   4
2586    2587    1   1
2586    2588    2   2
2586    2589    3   4
2587    2588    1   1
2587    2589    2   3
2588    2589    1   2
2642    2643    1   1
2642    2644    2   2
2642    2645    3   3
2643    2644    1   1
2643    2645    2   2
2644    2645    1   1
2660    2662    2   1
2706    2707    1   1
2714    2715    1   1
2860    2861    1   1
2860    2862    2   2
2861    2862    1   1
2907    2908    1   1
2909    2910    1   1
2931    2932    1   1
2931    2933    2   2
2932    2933    1   1
2957    2958    1   1
3017    3018    1   1
3092    3094    2   1
3111    3112    1   1
3111    3113    2   2
3111    3115    4   4
3112    3113    1   1
3112    3115    3   3
3112    3116    4   5
3113    3115    2   2
```

```
3113    3116    3   4
3113    3117    4   5
3115    3116    1   2
3115    3117    2   3
3116    3117    1   1
3128    3130    2   3
3128    3131    3   4
3130    3131    1   1
3139    3140    1   1
3145    3146    1   1
3145    3147    2   2
3145    3148    3   3
3145    3149    4   4
3146    3147    1   1
3146    3148    2   2
3146    3149    3   3
3147    3148    1   1
3147    3149    2   2
3148    3149    1   1
3175    3176    1   1
3177    3181    4   1
3200    3201    1   1
3214    3215    1   1
3214    3216    2   2
3215    3216    1   1
3224    3225    1   1
3232    3233    1   1
3232    3235    3   2
3232    3236    4   3
3233    3235    2   1
3233    3236    3   2
3233    3238    5   4
3235    3236    1   1
3235    3238    3   3
3235    3239    4   4
3235    3240    5   5
3236    3238    2   2
3236    3239    3   3
3236    3240    4   4
3236    3241    5   5
3238    3239    1   1
3238    3240    2   2
3238    3241    3   3
3238    3242    4   4
3238    3243    5   5
```

```
3239    3240    1   1
3239    3241    2   2
3239    3242    3   3
3239    3243    4   4
3239    3244    5   5
3240    3241    1   1
3240    3242    2   2
3240    3243    3   3
3240    3244    4   4
3240    3245    5   5
3241    3242    1   1
3241    3243    2   2
3241    3244    3   3
3241    3245    4   4
3241    3246    5   5
3242    3243    1   1
3242    3244    2   2
3242    3245    3   3
3242    3246    4   4
3242    3247    5   5
3243    3244    1   1
3243    3245    2   2
3243    3246    3   3
3243    3247    4   4
3243    3248    5   5
3244    3245    1   1
3244    3246    2   2
3244    3247    3   3
3244    3248    4   4
3244    3249    5   5
3245    3246    1   1
3245    3247    2   2
3245    3248    3   3
3245    3249    4   4
3245    3250    5   5
3246    3247    1   1
3246    3248    2   2
3246    3249    3   3
3246    3250    4   4
3246    3251    5   5
3247    3248    1   1
3247    3249    2   2
3247    3250    3   3
3247    3251    4   4
3247    3252    5   5
```

```
3248    3249    1   1
3248    3250    2   2
3248    3251    3   3
3248    3252    4   4
3248    3253    5   5
3249    3250    1   1
3249    3251    2   2
3249    3252    3   3
3249    3253    4   4
3249    3254    5   5
3250    3251    1   1
3250    3252    2   2
3250    3253    3   3
3250    3254    4   4
3250    3255    5   5
3251    3252    1   1
3251    3253    2   2
3251    3254    3   3
3251    3255    4   4
3251    3256    5   5
3252    3253    1   1
3252    3254    2   2
3252    3255    3   3
3252    3256    4   4
3252    3257    5   5
3253    3254    1   1
3253    3255    2   2
3253    3256    3   3
3253    3257    4   4
3253    3258    5   5
3254    3255    1   1
3254    3256    2   2
3254    3257    3   3
3254    3258    4   4
3254    3259    5   5
3255    3256    1   1
3255    3257    2   2
3255    3258    3   3
3255    3259    4   4
3256    3257    1   1
3256    3258    2   2
3256    3259    3   3
3257    3258    1   1
3257    3259    2   2
3258    3259    1   1
```

```
3277    3278    1   1
3277    3279    2   2
3277    3280    3   3
3278    3279    1   1
3278    3280    2   2
3279    3280    1   1
3390    3391    1   1
3390    3392    2   2
3390    3393    3   3
3391    3392    1   1
3391    3393    2   2
3392    3393    1   1
3498    3499    1   4
3507    3508    1   1
3517    3518    1   1
3582    3586    4   5
3586    3587    1   2
3591    3595    4   1
3596    3597    1   1
3641    3642    1   1
3651    3652    1   1
3651    3654    3   2
3652    3654    2   1
3673    3674    1   1
3673    3675    2   2
3673    3676    3   3
3674    3675    1   1
3674    3676    2   2
3675    3676    1   1
3678    3679    1   1
3680    3681    1   1
3680    3682    2   2
3680    3683    3   3
3681    3682    1   1
3681    3683    2   2
3682    3683    1   1
3689    3690    1   1
3769    3771    2   1
3772    3773    1   1
3772    3774    2   2
3773    3774    1   1
3785    3786    1   1
3797    3798    1   1
3859    3860    1   1
3898    3899    1   2
```

```
3898    3900    2   3
3898    3901    3   4
3898    3902    4   5
3899    3900    1   1
3899    3901    2   2
3899    3902    3   3
3899    3903    4   4
3899    3904    5   5
3900    3901    1   1
3900    3902    2   2
3900    3903    3   3
3900    3904    4   4
3900    3905    5   5
3901    3902    1   1
3901    3903    2   2
3901    3904    3   3
3901    3905    4   4
3901    3906    5   5
3902    3903    1   1
3902    3904    2   2
3902    3905    3   3
3902    3906    4   4
3903    3904    1   1
3903    3905    2   2
3903    3906    3   3
3904    3905    1   1
3904    3906    2   2
3905    3906    1   1
3909    3913    4   3
3962    3965    3   3
3975    3976    1   1
4012    4013    1   1
4014    4015    1   1
4014    4016    2   2
4014    4017    3   3
4014    4018    4   4
4014    4019    5   5
4015    4016    1   1
4015    4017    2   2
4015    4018    3   3
4015    4019    4   4
4016    4017    1   1
4016    4018    2   2
4016    4019    3   3
4017    4018    1   1
```

```
4017    4019    2   2
4018    4019    1   1
4022    4023    1   1
4061    4062    1   1
4065    4066    1   2
4090    4094    4   5
4091    4092    1   1
4093    4094    1   1
4119    4121    2   1
4119    4122    3   3
4121    4122    1   2
4130    4131    1   1
4139    4140    1   1
4176    4178    2   2
4297    4298    1   1
4297    4299    2   5
4298    4299    1   4
40  41  1   1
129 130 1   1
147 148 1   2
193 194 1   1
193 195 2   2
194 195 1   1
280 281 1   1
280 282 2   2
280 283 3   3
281 282 1   1
281 283 2   2
282 283 1   1
467 472 5   1
480 481 1   1
513 514 1   1
580 584 4   5
591 592 1   1
591 593 2   2
591 594 3   3
592 593 1   1
592 594 2   2
593 594 1   1
682 683 1   1
682 684 2   2
682 685 3   3
682 686 4   4
683 684 1   1
683 685 2   2
```

```
683 686 3   3
684 685 1   1
684 686 2   2
685 686 1   1
700 701 1   1
721 722 1   1
721 723 2   2
722 723 1   1
725 726 1   1
725 729 4   4
725 730 5   5
726 729 3   3
726 730 4   4
726 731 5   5
729 730 1   1
729 731 2   2
730 731 1   1
756 757 1   3
756 758 2   2
756 760 4   1
757 758 1   1
757 760 3   2
758 760 2   1
860 861 1   1
867 868 1   1
891 892 1   1
914 915 1   1
1020    1021    1   2
1155    1156    1   1
1155    1157    2   2
1156    1157    1   1
1192    1193    1   1
1224    1225    1   1
1226    1227    1   1
1314    1315    1   1
1639    1640    1   5
1838    1842    4   1
1852    1853    1   1
1852    1856    4   2
1852    1857    5   5
1853    1856    3   1
1853    1857    4   4
1856    1857    1   3
1947    1948    1   1
2023    2024    1   3
```

```
2023    2025    2   2
2024    2025    1   1
2035    2039    4   3
2037    2038    1   1
2088    2089    1   2
2134    2135    1   1
2205    2206    1   1
2403    2404    1   1
2403    2405    2   2
2404    2405    1   1
2501    2502    1   1
2529    2531    2   2
2802    2803    1   1
2883    2884    1   4
3121    3123    2   1
3157    3158    1   1
3304    3307    3   3
3327    3328    1   1
3365    3366    1   3
3365    3367    2   2
3365    3368    3   5
3365    3369    4   1
3366    3367    1   1
3366    3368    2   2
3366    3369    3   2
3367    3368    1   3
3367    3369    2   1
3368    3369    1   4
3387    3388    1   1
3387    3389    2   2
3388    3389    1   1
3422    3424    2   2
3468    3470    2   2
3478    3479    1   1
3478    3480    2   2
3478    3481    3   3
3478    3482    4   4
3479    3480    1   1
3479    3481    2   2
3479    3482    3   3
3480    3481    1   1
3480    3482    2   2
3481    3482    1   1
3484    3487    3   1
3504    3505    1   1
```

```
4145    4146    1   1
4145    4147    2   2
4145    4148    3   3
4146    4147    1   1
4146    4148    2   2
4147    4148    1   1
4166    4167    1   1
4182    4183    1   1
357 358 1   1
586 587 1   1
586 588 2   2
586 590 4   4
587 588 1   1
587 590 3   3
588 590 2   2
1391    1392    1   1
1978    1979    1   1
2819    2820    1   1
2994    2995    1   1
3032    3033    1   1
4187    4188    1   1
3310    3312    2   1
```