Kevin Dang Nguyen

ECS 124

**ECS 124A: Theory and Practice of Bioinformatics**

**Lab Assignment 4**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 1: Network Inference and GO analysis**

**Question 1.1:**

from.gene to.gene im

1 ycfP\_b1108\_at malF\_b4033\_at 0.008649257

2 yjjG\_b4374\_at pfs\_b0159\_at 0.007627152

3 gltW\_b2590\_s\_at gltT\_b3969\_s\_at 0.007448565

4 sulA\_b0958\_at yhdM\_b3292\_at 0.007423980

5 yijI\_b3948\_at trpB\_b1261\_at 0.007154548

6 gdhA\_b1761\_at metT\_b0673\_s\_at 0.007140683

7 yhjV\_b3539\_at gltW\_b2590\_s\_at 0.006924559

8 yrfE\_b3397\_at atoD\_b2221\_at 0.006841340

9 ileU\_b3277\_s\_at trpE\_b1264\_at 0.006815639

10 cyoE\_b0428\_at fhuA\_b0150\_at 0.006797588

11 glnB\_b2553\_at yheO\_b3346\_at 0.006783172

12 ybgA\_b0707\_at ygcF\_b2777\_at 0.006764527

13 pitA\_b3493\_at sfcA\_b1479\_at 0.006763321

14 ybiA\_b0798\_at rpoZ\_b3649\_at 0.006735039

15 yrbF\_b3195\_at yfjB\_b2615\_at 0.006732101

16 pstB\_b3725\_at FPLMCG\_Protein-E\_w\_at 0.006708946

17 yadM\_b0138\_at INM13X\_II\_3\_at 0.006664733

18 rsuA\_b2183\_at uxaC\_b3092\_at 0.006659602

19 gapC\_2\_b1416\_at napH\_b2204\_at 0.006644846

20 htpX\_b1829\_at LAMCG\_Fii\_w\_at 0.006635637

21 dnaN\_b3701\_at hyaF\_b0977\_at 0.006622997

22 yghQ\_b2983\_at aspU\_b0206\_s\_at 0.006622373

23 htrL\_b3618\_at pheV\_b2967\_s\_at 0.006621409

24 sohB\_b1272\_at ybhE\_b0767\_at 0.006598188

25 b2670\_at ppc\_b3956\_at 0.006590540

26 lrhA\_b2289\_at ybhE\_b0767\_at 0.006575160

27 rpoD\_b3067\_at folD\_b0529\_at 0.006568140

28 pnuC\_b0751\_at wecG\_b3794\_at 0.006510955

29 yecI\_b1902\_at IG\_2863\_4606893\_4606982\_fwd\_at 0.006480678

30 yidR\_b3689\_at IG\_486\_765099\_783104\_rev\_at 0.006401293

31 gpmB\_b4395\_at malM\_b4037\_at 0.006367093

32 seqA\_b0687\_f\_at yecP\_b1871\_at 0.006348263

33 folE\_b2153\_at appB\_b0979\_at 0.006319165

34 yjeF\_b4167\_at secG\_b3175\_at 0.006310741

35 hypD\_b2729\_at INM13X\_IV\_at 0.006307464

36 metG\_b2114\_at yheH\_b3330\_at 0.006306341

37 ugpQ\_b3449\_at aroA\_b0908\_at 0.006273674

38 napG\_b2205\_at yaaF\_b0030\_at 0.006255732

39 yigN\_b3832\_at hupB\_b0440\_at 0.006255724

40 yigP\_b3834\_at miaA\_b4171\_at 0.006244329

41 b1686\_at ycbB\_b0925\_at 0.006239926

42 yagU\_b0287\_at yjiL\_b4334\_at 0.006200768

43 mdlB\_b0449\_at ycjX\_b1321\_at 0.006188629

44 yggG\_b2936\_at map\_b0168\_at 0.006184506

45 yiiX\_b3937\_at rrlC\_b3758\_f\_at 0.006155406

46 vacJ\_b2346\_at yjhT\_b4310\_at 0.006149814

47 IG\_1427\_2261516\_2275912\_rev\_at IG\_196\_310561\_311737\_rev\_at 0.006119229

48 yfiQ\_b2584\_at yncB\_b1449\_at 0.006110619

49 smpA\_b2617\_at yqjB\_b3096\_at 0.006094159

50 dsbC\_b2893\_at lpdA\_b0116\_at 0.006092243

51 aroF\_b2601\_at exuT\_b3093\_at 0.006083861

52 b2989\_at PP1REP\_parA\_at 0.006083427

53 ybeN\_b0639\_at phnO\_b4093\_at 0.006067737

54 yagJ\_b0276\_at wzzE\_b3785\_at 0.006047745

55 ydaR\_b1356\_at b1840\_at 0.006037191

56 tyrR\_b1323\_at yhjT\_b3537\_at 0.006032022

57 hyfE\_b2485\_at aas\_b2836\_at 0.006019013

58 rpsQ\_b3311\_at hyaF\_b0977\_at 0.006018146

59 fabZ\_b0180\_at yfjD\_b2613\_at 0.006009509

60 valY\_b2403\_s\_at atoC\_b2220\_at 0.005996961

61 mltA\_b2813\_at ycbW\_b0946\_at 0.005995548

62 ogrK\_b2082\_at ykgG\_b0308\_at 0.005987910

63 yhhY\_b3441\_at lit\_b1139\_at 0.005986249

64 yggA\_b2923\_at ycdZ\_b1036\_at 0.005985990

65 SYNPBR322\_tet\_w\_at malG\_b4032\_at 0.005984899

66 dadX\_b1190\_at b2294\_at 0.005966532

67 pgi\_b4025\_at otsB\_b1897\_at 0.005962298

68 rpsE\_b3303\_at nadR\_b4390\_at 0.005960226

69 iclR\_b4018\_at xerC\_b3811\_at 0.005958288

70 hcaT\_b2536\_at uvrD\_b3813\_at 0.005957137

71 thrU\_b3976\_at b1645\_at 0.005953883

72 rfaY\_b3625\_at uraA\_b2497\_at 0.005948799

73 holE\_b1842\_at yejA\_b2177\_at 0.005908383

74 INM13X\_IV\_at b1624\_at 0.005905029

75 yhjG\_b3524\_at LAMCG\_ea31\_c\_at 0.005899384

76 bax\_b3570\_f\_at caiF\_b0034\_at 0.005884446

77 ybcW\_b0559\_at ymfJ\_b1144\_at 0.005872644

78 fixX\_b0044\_at b1213\_at 0.005860208

79 folE\_b2153\_at b1674\_at 0.005854227

80 amiA\_b2435\_at pstB\_b3725\_at 0.005851208

81 yjfO\_b4189\_at ccmH\_b2194\_at 0.005847709

82 fumB\_b4122\_at glyU\_b2864\_at 0.005842815

83 gltU\_b3757\_s\_at yahN\_b0328\_at 0.005838560

84 yjbN\_b4049\_at leuT\_b3798\_s\_at 0.005837956

85 ndk\_b2518\_at phoQ\_b1129\_at 0.005836555

86 b2636\_at ygjU\_b3089\_at 0.005832033

87 b2475\_at putA\_b1014\_at 0.005816608

88 dsbG\_b0604\_at b1441\_at 0.005815875

89 dedA\_b2317\_at grpE\_b2614\_at 0.005808274

90 kefC\_b0047\_at metJ\_b3938\_at 0.005806788

91 ydcG\_b1424\_at modE\_b0761\_at 0.005801352

92 b1192\_at trpB\_b1261\_at 0.005788851

93 ynhC\_b1681\_at ubiH\_b2907\_at 0.005786866

94 trpS\_b3384\_at tnaB\_b3709\_at 0.005781448

95 fucI\_b2802\_at galK\_b0757\_at 0.005772254

96 ydbK\_b1378\_at yhjG\_b3524\_at 0.005768364

97 yphA\_b2543\_at recB\_b2820\_at 0.005764110

98 katE\_b1732\_at wrbA\_b1004\_at 0.005761748

99 IG\_1721\_2765013\_2765055\_fwd\_at INM13X\_II\_3\_at 0.005761484

100 tag\_b3549\_at INM13X\_IV\_at 0.005760457

**Question 1.2:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**from.gene:**

|  |  |  |  |
| --- | --- | --- | --- |
| **From** | **To** | **Species** | **David Gene Name** |
| YPHA\_B2543\_AT | yphA | Escherichia coli | Inner membrane protein yphA |
| LRHA\_B2289\_AT | lrhA | Escherichia coli | Probable HTH-type transcriptional regulator lrhA |
| HOLE\_B1842\_AT | holE | Escherichia coli | DNA polymerase III subunit theta |
| YRBF\_B3195\_AT | yrbF | Escherichia coli | Uncharacterized ABC transporter ATP-binding protein yrbF |
| YRBF\_B3195\_AT | mlaF | Escherichia coli | Uncharacterized ABC transporter ATP-binding protein yrbF |
| PITA\_B3493\_AT | pitA | Escherichia coli | Low-affinity inorganic phosphate transporter 1 |
| DSBG\_B0604\_AT | dsbG | Escherichia coli | Thiol:disulfide interchange protein dsbG |
| YIDR\_B3689\_AT | ECs4629 | Escherichia coli | ECs4629 |
| RFAY\_B3625\_AT | rfaY | Escherichia coli | Lipopolysaccharide core biosynthesis protein rfaY |
| B2475\_AT | ypfJ | Escherichia coli | Uncharacterized protein ypfJ |
| YJBN\_B4049\_AT | ECs5031 | Escherichia coli | tRNA-dihydrouridine synthase A |
| YDAR\_B1356\_AT | racR | Escherichia coli | Rac prophage repressor |
| YNHC\_B1681\_AT | ynhC | Escherichia coli | Z2709; c2076 |
| KATE\_B1732\_AT | katE | Escherichia coli | Catalase HPII; Catalase |
| YBEN\_B0639\_AT | nadD | Escherichia coli | Nicotinate-nucleotide adenylyltransferase; Probable nicotinate-nucleotide adenylyltransferase |
| NDK\_B2518\_AT | ndk | Escherichia coli | Nucleoside diphosphate kinase |
| YGHQ\_B2983\_AT | yghQ | Escherichia coli | Inner membrane protein yghQ |
| YHJV\_B3539\_AT | ECs4419 | Escherichia coli | ECs4419 |
| YECI\_B1902\_AT | yecI | Escherichia coli | Ferritin-like protein 2 |
| YAGJ\_B0276\_AT | yagJ | Escherichia coli | Uncharacterized protein yagJ |
| B2636\_AT | yfjS | Escherichia coli | Uncharacterized lipoprotein yfjS |
| SULA\_B0958\_AT | sulA | Escherichia coli | Cell division inhibitor sulA |
| YIGN\_B3832\_AT | yigN | Escherichia coli | DNA recombination protein rmuC |
| DSBC\_B2893\_AT | dsbC | Escherichia coli | Thiol:disulfide interchange protein dsbC |
| YHJG\_B3524\_AT | yhjG | Escherichia coli | Uncharacterized protein yhjG |
| RPOD\_B3067\_AT | rpoD | Escherichia coli | RNA polymerase sigma factor rpoD; RNA polymerase sigma factor |
| TYRR\_B1323\_AT | tyrR | Escherichia coli | Transcriptional regulatory protein tyrR |
| FUMB\_B4122\_AT | ECs5104 | Escherichia coli | ECs5104 |
| FABZ\_B0180\_AT | fabZ | Escherichia coli | (3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase |
| YADM\_B0138\_AT | yadM | Escherichia coli | Uncharacterized protein yadM |
| GAPC\_2\_B1416\_AT | c1843 | Escherichia coli | c1843 |
| KEFC\_B0047\_AT | kefC | Escherichia coli | Glutathione-regulated potassium-efflux system protein kefC |
| B1686\_AT | ydiI | Escherichia coli | Esterase ydiI |
| RPSQ\_B3311\_AT | rpsQ | Escherichia coli | 30S ribosomal protein S17 |
| AMIA\_B2435\_AT | ECs3306 | Escherichia coli | ECs3306 |
| YIIX\_B3937\_AT | yiiX | Escherichia coli | Uncharacterized protein yiiX |
| HTRL\_B3618\_AT | ECs4496 | Escherichia coli | ECs4496 |
| BAX\_B3570\_F\_AT | BAX | Escherichia coli | Protein bax |
| HCAT\_B2536\_AT | hcaT | Escherichia coli | Probable 3-phenylpropionic acid transporter |
| IG\_1721\_2765013\_2765055\_FWD\_AT | yfjO | Escherichia coli | Uncharacterized protein yfjO |
| IG\_1427\_2261516\_2275912\_REV\_AT | c2707 | Escherichia coli | UPF0153 protein yeiW |
| METG\_B2114\_AT | metG | Escherichia coli | Methionyl-tRNA synthetase |
| YFIQ\_B2584\_AT | yfiQ | Escherichia coli | Uncharacterized protein yfiQ |
| TAG\_B3549\_AT | ECs4434 | Escherichia coli | ECs4434 |
| B2670\_AT | ygaW | Escherichia coli | Uncharacterized protein ygaW |
| AROF\_B2601\_AT | aroF | Escherichia coli | Phospho-2-dehydro-3-deoxyheptonate aldolase, Tyr-sensitive |
| VACJ\_B2346\_AT | ECs3229 | Escherichia coli | ECs3229 |
| B2989\_AT | yghU | Escherichia coli | Uncharacterized GST-like protein yghU |
| YGGG\_B2936\_AT | ECs3811 | Escherichia coli | ECs3811 |
| YBIA\_B0798\_AT | ybiA | Escherichia coli | Swarming motility protein ybiA |
| YJJG\_B4374\_AT | yjjG | Escherichia coli | 5'-nucleotidase yjjG |
| HYPD\_B2729\_AT | hypD | Escherichia coli | Hydrogenase isoenzymes formation protein hypD |
| YIJI\_B3948\_AT | ECs4876 | Escherichia coli | ECs4876 |
| SMPA\_B2617\_AT | ECs3479 | Escherichia coli | ECs3479 |
| YHHY\_B3441\_AT | yhhY | Escherichia coli | Uncharacterized N-acetyltransferase yhhY |
| DEDA\_B2317\_AT | ECs3201 | Escherichia coli | Protein dedA |
| SOHB\_B1272\_AT | sohB | Escherichia coli | Probable protease sohB |
| YDBK\_B1378\_AT | ECs2000 | Escherichia coli | ECs2000 |
| HYFE\_B2485\_AT | hyfE | Escherichia coli | Hydrogenase-4 component E |
| GPMB\_B4395\_AT | gpmB | Escherichia coli | Probable phosphoglycerate mutase gpmB |
| DNAN\_B3701\_AT | dnaN | Escherichia coli | DNA polymerase III subunit beta |
| ICLR\_B4018\_AT | ECs4936 | Escherichia coli | ECs4936 |
| GLNB\_B2553\_AT | ECs3419 | Escherichia coli | Nitrogen regulatory protein P-II 1 |
| YCFP\_B1108\_AT | ycfP | Escherichia coli | UPF0227 protein ycfP |
| INM13X\_IV\_AT | I | Enterobacteria phage M13 | Gene 1 protein |
| PSTB\_B3725\_AT | pstB | Escherichia coli | Phosphate import ATP-binding protein pstB |
| B1192\_AT | ldcA | Escherichia coli | Muramoyltetrapeptide carboxypeptidase |
| TRPS\_B3384\_AT | trpS | Escherichia coli | Tryptophanyl-tRNA synthetase |
| YIGP\_B3834\_AT | yigP | Escherichia coli | Uncharacterized protein yigP |
| NAPG\_B2205\_AT | NAPG | Escherichia coli | Ferredoxin-type protein napG |
| YRFE\_B3397\_AT | nudE | Escherichia coli | ADP compounds hydrolase nudE |
| CYOE\_B0428\_AT | ECs0482 | Escherichia coli | Protoheme IX farnesyltransferase |
| MLTA\_B2813\_AT | mltA | Escherichia coli | Membrane-bound lytic murein transglycosylase A |
| YBCW\_B0559\_AT | ECs1627 | Escherichia coli | Uncharacterized protein ybcW |
| RPSE\_B3303\_AT | rpsE | Escherichia coli | 30S ribosomal protein S5 |
| FOLE\_B2153\_AT | folE | Escherichia coli | GTP cyclohydrolase 1 |
| RSUA\_B2183\_AT | ECs3075 | Escherichia coli | Ribosomal small subunit pseudouridine synthase A |
| OGRK\_B2082\_AT | ogrK | Escherichia coli | Prophage P2 OGR protein |
| YAGU\_B0287\_AT | ECs0317 | Escherichia coli | Inner membrane protein yagU |
| HTPX\_B1829\_AT | HtpX | Escherichia coli | Probable protease htpX |
| FUCI\_B2802\_AT | fucI | Escherichia coli | L-fucose isomerase |
| THRU\_B3976\_AT | thrU | Escherichia coli | ECs5571; Z5548; b3976; ECDH10B\_4164; c5587; E2348C\_tRNA79 |
| MDLB\_B0449\_AT | ECs0503 | Escherichia coli | Multidrug resistance-like ATP-binding protein mdlB |
| PNUC\_B0751\_AT | pnuC | Escherichia coli | Nicotinamide riboside transporter pnuC |
| GDHA\_B1761\_AT | ECs2467 | Escherichia coli | ECs2467 |
| DADX\_B1190\_AT | dadX | Escherichia coli | Alanine racemase, catabolic; Alanine racemase |
| YGGA\_B2923\_AT | argO | Escherichia coli | Arginine exporter protein argO |
| PGI\_B4025\_AT | pgi | Escherichia coli | Glucose-6-phosphate isomerase |
| UGPQ\_B3449\_AT | ugpQ | Escherichia coli | Glycerophosphoryl diester phosphodiesterase |
| YDCG\_B1424\_AT | mdoD | Escherichia coli | Glucans biosynthesis protein D |
| FIXX\_B0044\_AT | fixX | Escherichia coli | Ferredoxin-like protein fixX |
| SEQA\_B0687\_F\_AT | seqA | Escherichia coli | Protein seqA |
| YJEF\_B4167\_AT | yjeF | Escherichia coli | Uncharacterized protein yjeF |
| YBGA\_B0707\_AT | ybgA | Escherichia coli | Uncharacterized protein ybgA |
| YJFO\_B4189\_AT | bsmA | Escherichia coli | Uncharacterized lipoprotein yjfO |
| YJFO\_B4189\_AT | yjfO | Escherichia coli | Uncharacterized lipoprotein yjfO |

|  |  |
| --- | --- |
| |  | | --- | |  | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**to.gene:**

|  |  |  |  |
| --- | --- | --- | --- |
| Top of Form  Bottom of Form | | | |
| **From** | **To** | **Species** | **David Gene Name** |
| RECB\_B2820\_AT | recB | Escherichia coli | Exodeoxyribonuclease V beta chain |
| METJ\_B3938\_AT | metJ | Escherichia coli | Met repressor |
| YCJX\_B1321\_AT | ycjX | Escherichia coli | Uncharacterized protein ycjX |
| LAMCG\_FII\_W\_AT | Rz | Enterobacteria phage lambda | Endopeptidase |
| HYAF\_B0977\_AT | hyaF | Escherichia coli | Hydrogenase-1 operon protein hyaF |
| ATOC\_B2220\_AT | atoC | Escherichia coli | Acetoacetate metabolism regulatory protein atoC |
| YJIL\_B4334\_AT | ECs5297 | Escherichia coli | ECs5297 |
| IG\_196\_310561\_311737\_REV\_AT | rpmJ | Escherichia coli | 50S ribosomal protein L36 2; 50S ribosomal protein L36; 50S ribosomal protein L36 1 |
| YHDM\_B3292\_AT | zntR | Escherichia coli | HTH-type transcriptional regulator zntR |
| NAPH\_B2204\_AT | napH | Escherichia coli | Ferredoxin-type protein napH |
| SFCA\_B1479\_AT | maeA | Escherichia coli | NAD-dependent malic enzyme |
| FHUA\_B0150\_AT | ECs0154 | Escherichia coli | ECs0154 |
| SECG\_B3175\_AT | secG | Escherichia coli | Protein-export membrane protein secG |
| MALG\_B4032\_AT | malG | Escherichia coli | Maltose transport system permease protein malG |
| YHEO\_B3346\_AT | yheO | Escherichia coli | Uncharacterized protein yheO |
| ATOD\_B2221\_AT | atoD | Escherichia coli | Acetate CoA-transferase subunit alpha |
| YFJD\_B2613\_AT | yfjD | Escherichia coli | UPF0053 inner membrane protein yfjD |
| TRPB\_B1261\_AT | trpB | Escherichia coli | Tryptophan synthase beta chain |
| MAP\_B0168\_AT | map | Escherichia coli | Methionine aminopeptidase |
| APPB\_B0979\_AT | appB | Escherichia coli | Cytochrome bd-II oxidase subunit 2 |
| LPDA\_B0116\_AT | lpdA | Escherichia coli | Dihydrolipoyl dehydrogenase |
| UBIH\_B2907\_AT | ECs3778 | Escherichia coli | ECs3778 |
| B1840\_AT | yebZ | Escherichia coli | Inner membrane protein yebZ |
| MALF\_B4033\_AT | malF | Escherichia coli | Maltose transport system permease protein malF |
| MIAA\_B4171\_AT | miaA | Escherichia coli | tRNA Delta(2)-isopentenylpyrophosphate transferase |
| B2294\_AT | ECs3178 | Escherichia coli | UPF0304 protein yfbU |
| YAHN\_B0328\_AT | yahN | Escherichia coli | Uncharacterized membrane protein yahN |
| AAS\_B2836\_AT | aas | Escherichia coli | Bifunctional protein aas |
| B1441\_AT | ECs2045 | Escherichia coli | ECs2045 |
| URAA\_B2497\_AT | ECs3359 | Escherichia coli | Uracil permease |
| YCBB\_B0925\_AT | ycbB | Escherichia coli | Uncharacterized protein ycbB |
| YQJB\_B3096\_AT | ECs3978 | Escherichia coli | ECs3978 |
| YBHE\_B0767\_AT | ybhE | Escherichia coli | 6-phosphogluconolactonase |
| WRBA\_B1004\_AT | wrbA | Escherichia coli | Flavoprotein wrbA; Putative flavoprotein wrbA |
| PPC\_B3956\_AT | ppc | Escherichia coli | Phosphoenolpyruvate carboxylase |
| YGJU\_B3089\_AT | ECs3971 | Escherichia coli | Serine/threonine transporter sstT |
| LAMCG\_EA31\_C\_AT | Rz | Enterobacteria phage lambda | Endopeptidase |
| YCDZ\_B1036\_AT | ycdZ | Escherichia coli | Inner membrane protein ycdZ |
| B1674\_AT | c2068 | Escherichia coli | c2068 |
| PUTA\_B1014\_AT | putA | Escherichia coli | Bifunctional protein putA |
| CAIF\_B0034\_AT | ECs0037 | Escherichia coli | ECs0037 |
| YHJT\_B3537\_AT | yhjT | Escherichia coli | Uncharacterized protein yhjT |
| TRPE\_B1264\_AT | trpE | Escherichia coli | Anthranilate synthase component 1 |
| YECP\_B1871\_AT | cmoB | Escherichia coli | tRNA (mo5U34)-methyltransferase |
| IG\_486\_765099\_783104\_REV\_AT | c0824 | Escherichia coli | c0824 |
| IG\_2863\_4606893\_4606982\_FWD\_AT | c5455 | Escherichia coli | c5455 |
| YGCF\_B2777\_AT | ygcF | Escherichia coli | Uncharacterized protein ygcF |
| MODE\_B0761\_AT | modE | Escherichia coli | Transcriptional regulator modE |
| OTSB\_B1897\_AT | otsB | Escherichia coli | Trehalose-phosphate phosphatase |
| CCMH\_B2194\_AT | ccmH | Escherichia coli | Cytochrome c-type biogenesis protein ccmH |
| RRLC\_B3758\_F\_AT | rrlE | Escherichia coli | b4009; c5613; ECDH10B\_4198; ECs5578; EcE24377A\_5094; E2348C\_rRNA21; Z5589; UTI89\_C3807 |
| FOLD\_B0529\_AT | folD | Escherichia coli | Bifunctional protein folD |
| YJHT\_B4310\_AT | ECs5269 | Escherichia coli | N-acetylneuraminate epimerase |
| YKGG\_B0308\_AT | ykgG | Escherichia coli | UPF0707 protein ykgG |
| YMFJ\_B1144\_AT | ymfJ | Escherichia coli | Uncharacterized protein ymfJ |
| B1213\_AT | c1671 | Escherichia coli | c1671 |
| GALK\_B0757\_AT | galK | Escherichia coli | Galactokinase |
| AROA\_B0908\_AT | ECs0991 | Escherichia coli | 3-phosphoshikimate 1-carboxyvinyltransferase |
| GRPE\_B2614\_AT | ECs3476 | Escherichia coli | Protein grpE |
| TNAB\_B3709\_AT | tnaB | Escherichia coli | Low affinity tryptophan permease |
| LIT\_B1139\_AT | lit | Escherichia coli | Bacteriophage T4 late gene expression-blocking protein |
| YNCB\_B1449\_AT | ECs2053 | Escherichia coli | ECs2053 |
| INM13X\_IV\_AT | I | Enterobacteria phage M13 | Gene 1 protein |
| PSTB\_B3725\_AT | pstB | Escherichia coli | Phosphate import ATP-binding protein pstB |
| YEJA\_B2177\_AT | yejA | Escherichia coli | Uncharacterized protein yejA |
| HUPB\_B0440\_AT | hupB | Escherichia coli | DNA-binding protein HU-beta |
| YFJB\_B2615\_AT | ppnK | Escherichia coli | Probable inorganic polyphosphate/ATP-NAD kinase |
| XERC\_B3811\_AT | xerC | Escherichia coli | Tyrosine recombinase xerC |
| NADR\_B4390\_AT | ECs5348 | Escherichia coli | ECs5348 |
| MALM\_B4037\_AT | malM | Escherichia coli | Maltose operon periplasmic protein |
| WZZE\_B3785\_AT | wzzE | Escherichia coli | Lipopolysaccharide biosynthesis protein wzzE |
| YCBW\_B0946\_AT | ECs1030 | Escherichia coli | ECs1030 |
| PHOQ\_B1129\_AT | phoQ | Escherichia coli | Sensor protein phoQ; Sensor protein |
| YHJG\_B3524\_AT | yhjG | Escherichia coli | Uncharacterized protein yhjG |
| FPLMCG\_PROTEIN-E\_W\_AT | repE | Plasmid F | Replication initiation protein |
| PFS\_B0159\_AT | pfs | Escherichia coli | 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase |
| PHNO\_B4093\_AT | ECs5076 | Escherichia coli | ECs5076 |
| B1624\_AT | ECs2332 | Escherichia coli | ECs2332 |
| YHEH\_B3330\_AT | yheH | Escherichia coli | c4101; UTI89\_C3785 |
| INM13X\_II\_3\_AT | I | Enterobacteria phage M13 | Gene 1 protein |
| WECG\_B3794\_AT | wecG | Escherichia coli | Probable UDP-N-acetyl-D-mannosaminuronic acid transferase |
| EXUT\_B3093\_AT | ECs3975 | Escherichia coli | Hexuronate transporter |
| UXAC\_B3092\_AT | uxaC | Escherichia coli | Uronate isomerase |
| B1645\_AT | ydhK | Escherichia coli | Uncharacterized transporter ydhK |
| YAAF\_B0030\_AT | ECs0033 | Escherichia coli | Non-specific ribonucleoside hydrolase rihC |
| UVRD\_B3813\_AT | uvrD | Escherichia coli | DNA helicase II |
| RPOZ\_B3649\_AT | rpoZ | Escherichia coli | DNA-directed RNA polymerase subunit omega |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 2: Phylogenetic analysis**

**Question 2.1.** Calculate the pair-wise Jukes-Cantor distance between these three sequences that

were taken from three different species:

S1: AAAATCGATCAAATCAT

S2: AATCTCGATCAATTCAT

S3: ATATTCGATAAATTAAT

pt = 6/17

dJC = -3/4ln(1- 4/3pt)

dJC = 0.477

What can tell from these calculations regarding the evolutionary similarity of the corresponding

species?

* The Jukes-Cantor distance gives us the estimation of the genetic distance between 2 or more homologous species (the number of substitutions over time t for a given nucleotide position pair since they diverged from their common ancestor). For this case, they are not very similar.

**Question 2.2.**

What are the assumptions that the Jukes-Cantor Model makes?

* Assuming they have a common ancestor. The JC distance over time t to be equal to 6(apha)t.

What would be the Jukes-Cantor distance between two sequences, where the first has a frame shift mutation (indel) but otherwise is identical? For example assume the following genomic sequences:

S1: ATTCGAAA …. **TCAAAATGCA** …. ATTGDSAAAA

S2: ATTCGAAA …. **CAAAATGCAT** …. ATTGDSAAAA

Only the sequences in bold are given (second sequence identical to the first but shifted one position to the left). Calculate the JC distance.

* Since the two sequences in bold only differ by one position. pt = 1/28
* dJC = 0.0366

**Question 2.3.** Derive the Jukes-Cantor probability P that a nucleotide will remain unchanged

over a period of T time units, which is given by:

P = ¼ (1+ 3e-4aT)

Assume rate of change α for a nucleotide mutating to any of the 3 other nucleotides (i.e. the

mutation probability for any given nucleotide during an infinitesimal period of time is 3αΔt).

You can follow the same reasoning that we used in class and then set the derivative of the

substitution matrix to zero to get the retention/substitution probabilities.

1 – 3αΔt αΔt αΔt αΔt

αΔt 1 – 3αΔt αΔt αΔt

αΔt αΔt 1 – 3αΔt αΔt

αΔt αΔt αΔt 1 – 3αΔt

qt+Δt= qt[(1 – 3αΔt)^2 + 3(αΔt)^2] + (1-qt)[2(1 – 3αΔt)( αΔt) + 2(αΔt)^2]

qt+Δt = qt(1 – 6αΔt) + (1-qt)(2 αΔt)

qt+Δt = qt - 6αΔtqt + 2 αΔt - 2 αΔtqt

(qt+Δt – qt)/Δt = (-8 αΔtqt + 2 αΔt)/Δt

dqt/dt = 2α -8 αqt

∫ dq/2α -8 αqt = ∫ dt

1/(-8qt)(log(2α - 8αqt))’ = t + c1

2α - 8αqt = c1e^-8αt + ¼

Qt=0 = 1

1 = c + ¼

c = ¾

qt = 3/4e^(-4αt) + ¼

**Question 2.4.** Below is a distance matrix between different species.

**HUMAN MOUSE RAT DOG CAT**

**HUMAN** 0 3 5 10 11

**MOUSE** 0 2 7 8

**RAT** 0 5 8

**DOG** 0 2

**CAT** 0

Is this an ultra-metric distance matrix? Why?

* Yes, because all paths from leaves are equidistant to the root. Also, the 3-point rule applies.
* D(x,y) > 0
* D(x,x) = 0
* D(x,y)= D(y,x)
* D(x,y)≤D(x,z)+D(z,y)
* D(x,y)≤ max[D(x,z), D(z,y)] or D(x,y)≤ D(x,z) = D(y,z)

Use UPGMA to create the corresponding tree

Species HUMAN MOUSE RAT DOG CAT

HUMAN 0 3 5 10 11

MOUSE 3 0 2 7 8

RAT 5 2 0 5 8

DOG 10 7 5 0 2

CAT 11 8 8 2 0

+-------------+ 3: RAT {1.000000}

|

+-------------+ 6 {1.000000}

| |

| +-------------+ 2: MOUSE {1.000000}

|

+----------------------------+ 8 {2.083300}

| |

| +--------------------------+ 1: HUMAN {2.000000}

|

+---+ 9

|

| +-------------+ 4: DOG {1.000000}

| |

+-----------------------------------------+ 7 {3.083300}

|

+-------------+ 5: CAT {1.000000}

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 3: Unsupervised learning (bonus)**

**Question 3.1.** Use any of the partitioning methods to cluster the Deluc\_Grapes\_Dataset1. Do

that for number of clusters k=5,10,20. For each cluster only, in each of these cases (i.e. for

k=5,10,20) report (a) the top 3 genes, (b) the cluster center, (c) the number of genes in the

cluster. Visualize the results (any heatmap will do).

install.packages('cluster')

library(cluster)

mydata = read.table("Deluc\_Grapes\_DataSet1.txt", sep = "\t", header = FALSE)

mydata <-na.omit(mydata)

day21 = clara(as.numeric(mydata$V2), 5)

day42 = clara(mydata$V3,5)

day42$clusinfo

plot(day42)

day42 = clara(mydata$V3,10)

day42$clusinfo

plot(day42)

day42 = clara(mydata$V3,20)

day42$clusinfo

plot(day42)

day49 = clara(mydata$V4,5)

day49$clusinfo

plot(day49)

day49 = clara(mydata$V4,10)

day49$clusinfo

plot(day49)

day49 = clara(mydata$V4,20)

day49$clusinfo

plot(day49)

day56 = clara(mydata$V5,5)

day56$clusinfo

plot(day56)

day56 = clara(mydata$V5,10)

day56$clusinfo

plot(day56)

day56 = clara(mydata$V5,20)

day56$clusinfo

plot(day56)

day63 = clara(mydata$V6,5)

day63$clusinfo

plot(day63)

day63 = clara(mydata$V6,10)

day63$clusinfo

plot(day63)

day63 = clara(mydata$V6,20)

day63$clusinfo

plot(day63)

day84 = clara(mydata$V7,5)

day84$clusinfo

plot(day84)

day84 = clara(mydata$V7,10)

day84$clusinfo

plot(day84)

day84 = clara(mydata$V7,20)

day84$clusinfo

plot(day84)

day112 = clara(mydata$V8,5)

day112$clusinfo

plot(day112)

day112 = clara(mydata$V8,10)

day112$clusinfo

plot(day112)

day112 = clara(mydata$V8,20)

day112$clusinfo

plot(day112)