Co-expression Network 제작 실습

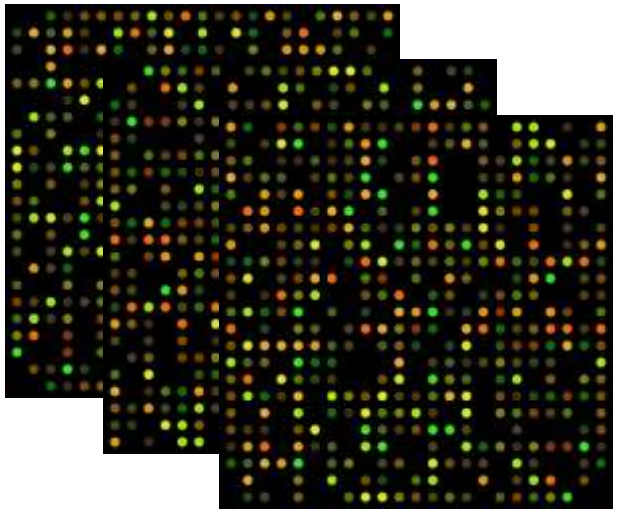
김이루
Network Biology Lab
2016 KSBSB 동계워크샵

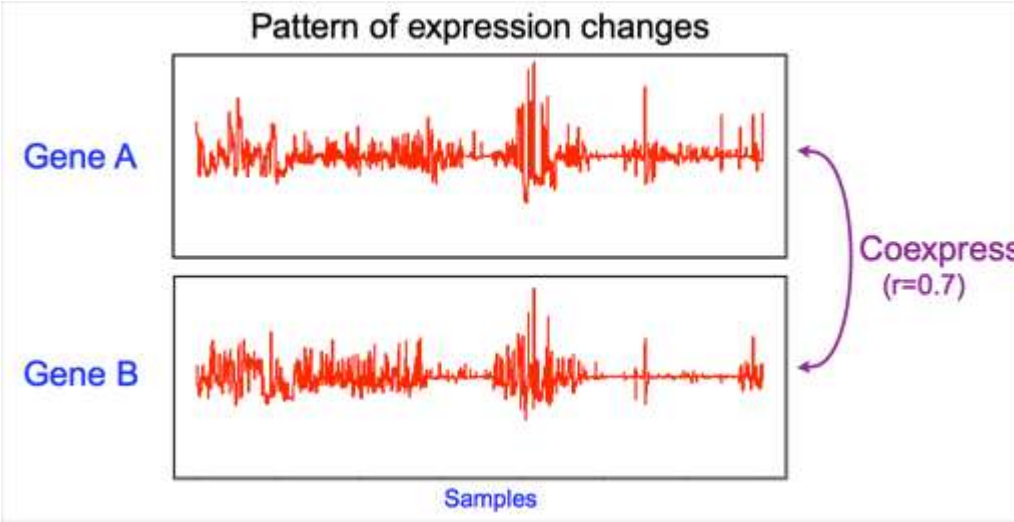
1. 실습 목표

- 네트워크를 만드는 과정 이해
- Public microarray data 사용하는 방법 실습
- Public expression data를 사용하여 유전자 네트워크 제작 방법 실습
- 네트워크를 Benchmark하는 방법 실습
- 네트워크들을 Integration 하는 방법 실습

2. 이론적 배경

- Co-expression pattern
  - Expression의 변화 양상의 일치 정도를 사용하여 기능적 상관관계를 유추하는 방법
  - 같은 Transcription facto에 조절되거나 같은 Pathway에 작용하는 두 유전자는 발현 패턴이 비슷할 것이다.
  - 이에 여러 개의 RNA expression data에 걸쳐서 발현의 패턴이 비슷한 두 유전자는 기능적으로 상관관계가 높을 것이라는 이론에 기반한다.
  - 보통 발현 패턴 비교는 Pearson correlation coefficient로 측정을 한다.





http://atted.jp/overview.shtml

Co-expression 계산 예제

	Sample 1	Sample 2	Sample 3	Control 1	Control 2
Gene A	5423	8342	6463	1043	324
Gene B	6531	10432	5813	1232	300
Gene C	143	424	6419	261	153

Divide by average of controls, and convert to log value

	Sample 1	Sample 2	Sample 3	Control 1	Control 2
Gene A	12.40487546	13.0261776	12.65798828	10.02652344	8.339850003
Gene B	12.67308819	13.34872815	12.50506719	10.26678654	8.22881869
Gene C	7.159871337	8.727920455	12.64813285	8.027905997	7.257387843

PCC calculation

Gene A - Gene B	0.995051
Gene A - Gene C	0.481868
Gene B - Gene C	0.406111

Log likelihood score

- 여러 Biological data는 서로 다른 score 분포와 신뢰성을 가지고 있기 때문에 하나의 기준으로 score를 통일 하는 것이 필요하다
- 이런 score 통일과 일관성 있는 데이터 신뢰성 평가를 위하여 네트워크 link benchmark를 할 때 Log likelihood score (LLS)를 사용한다.
- LLS를 구하기 위해서는 신뢰성 있는 Gold standard set이 필요하다.
- LLS 공식은 다음과 같다.

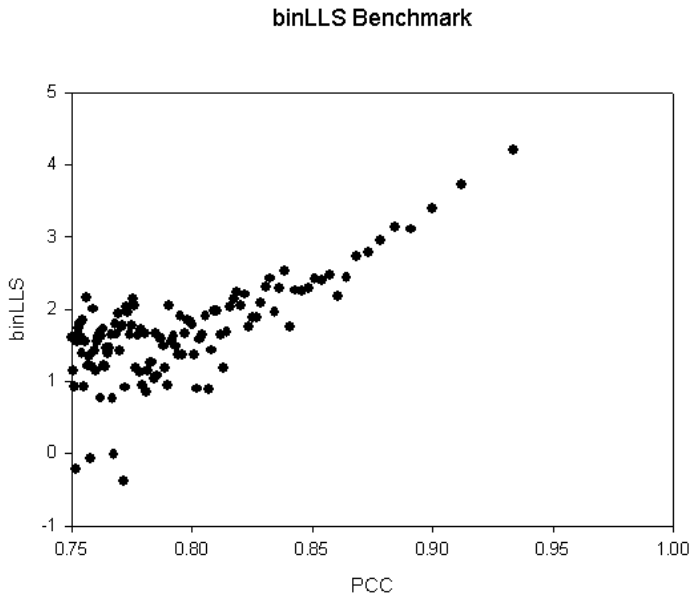
$$LLS = \ln\left(\frac{P(L|E)/\sim P(L|E)}{P(L)/\sim P(L)}\right)$$

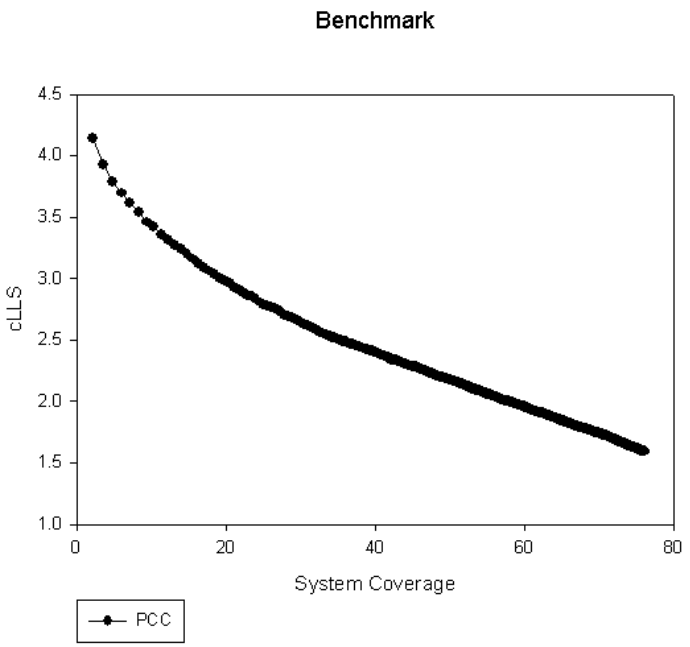
- P(L): 전체 pairs중에서 gold standard positive pair를 **뽑을** 확률이다.
- P(L|E): 특정 pair set에서 gold standard positive pair을 **뽑은** 확률이다.
- 즉 LLS는 특정 pair set에서 확률적으로 random에 비해서 positive pair를 얼마나 많이 뽑았는지를 나타낸다.

Insuk Lee and Ben Lehner et al., *Nat Genet*. 2008

Benchmark

- 네트워크 performance를 측정 및 Log-likelihood score regression을 위한 선행 단계
- Gold standard를 사용하여 network의 performance를 측정한다.
- 성능을 측정할 때 보통 10,00개의 link를 하나의 bin으로 묶어 측정한다
- 2가지 Benchmark를 활용
  - Bin 하나당 PCC 값 대비 LLS를 측정하는 방법 (binLLS)
  - Bin별로 누적 유전자 수 대비 누적 LLS를 측정하는 방법 (cLLS)



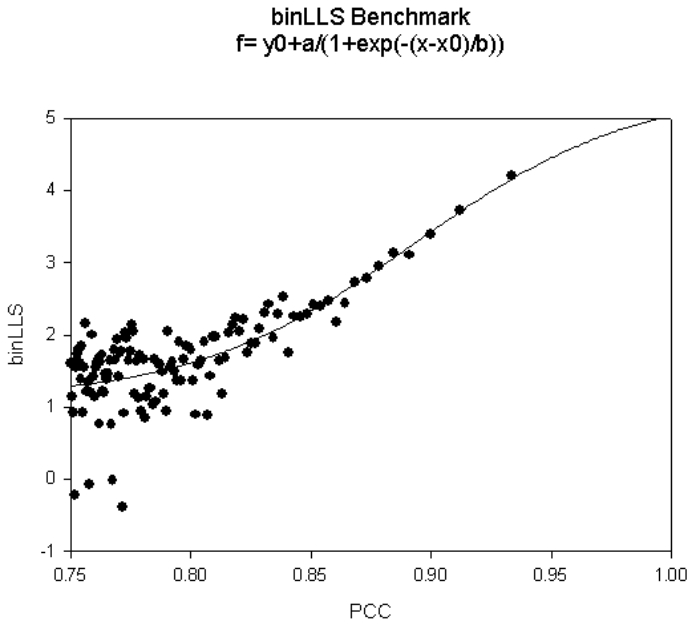


Regression

- Benchmark에서 측정된 LLS를 기반으로 regression model을 만들어 각 functional link별로 LLS를 맵핑하는 과정
- 보통 Sigmoidal function을 사용하여 regression model을 제작한다
- 낮은 PCC 쪽의 노이즈 부분을 적절하게 제거하여 가장 best fit된 model을 사용한다

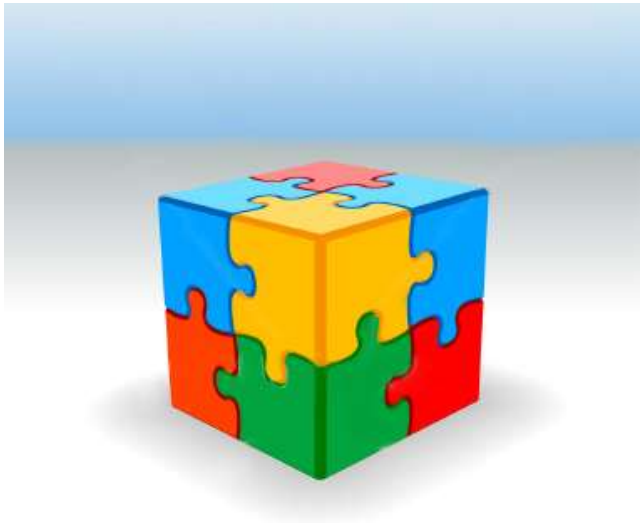
$$f(x) = y_0 + \frac{a}{1 + e^{-\left(\frac{x - x_0}{b}\right)}}$$

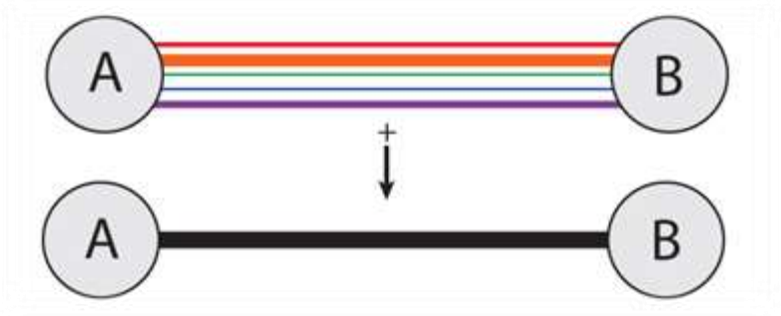
Sigmoidal 4 parameter function



Integration

- Weighted sum 공식
  - $WS = L_0 + \sum_{i=1}^{n} \frac{L_0}{D \cdot i}, for\ all\ L \geq T$
  - N: LLS threshold를 넘는 Evidences 개수
  - L_0: Evidences중 최대값
  - D: Weighting factor
  - T: LLS threshold
- Optimization
  - D값과 T값을 조절 하여 네트워크 성능의 최적화를 시키는 과정





3. 준비 사항

- Linux가 설치된 PC 또는 Virtual machine
- R
  - Affy Package
  - Biocondutor Package
- Co-expression 네트워크 제작 프로그램
  - Expression matrix 제작 스크립트
  - Pearson correlation coefficient 계산 프로그램
  - Benchmark 프로그램
  - Sigmoidal function regression 프로그램
  - Integration 프로그램

4. 실습 내용

- Public에서 Microarray 데이터 받아서 Expression profile 만들기
  1. Gene Expression Omnibus 접속 (www.ncbi.nlm.nih.gov/**geo**/)
  2. 원하는 Expression data 검색 (e.g. GSE5657)
  3. Supplementary file download (TAR of CEL)

     - 또는 다음과 같은 방법으로 다운로드가 가능

     $ wget ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/supplementary/series/GSE5657/GSE5657_RAW.tar
  4. Microarray annotation packages 실행

     $ install_annotation_package.R
  5. Cel폴더 만들고 데이터 이동

     $ mkdir cel

     $ mv GSE5657_RAW.tar cel
  6. TAR파일 압축 풀기

     $ tar -xvf GSE5657_RAW.tar –C cel/
  7. Expression matrix 제작 code 실행

     $ ./make_expression_matrix.R GSE5657

- Output 형식

Expression matrix

geneIDs	GSM252064.CEL	GSM252065.CEL	GSM252066.CEL	GSM252067.CEL	GSM252068.CEL	GSM252069.CEL	GSM252070.CEL	GSM252071.CEL	GSM252072.CEL	GSM252073.CEL
100009600	36.42	10.34	22.82	199.10	28.58	42.45	95.77	17.49	81.44	119.99
100012	16.34	7.23	36.81	12.84	4.61	5.99	5.90	3.07	13.88	3.22
100017	173.03	174.35	156.05	233.99	255.60	238.81	154.65	135.02	147.68	154.31
100019	113.20	141.18	113.05	357.82	268.04	297.34	82.89	120.24	87.04	122.80
100034360	4.70	5.76	41.38	4.15	7.04	1.99	51.37	1.93	8.63	2.47
100034675	10.48	51.83	37.80	60.94	75.87	63.93	49.42	51.52	18.89	58.41
100034733	16.34	69.03	22.09	273.74	226.43	308.88	9.91	26.36	12.88	42.36
100034748	61.32	57.23	119.74	59.64	177.55	98.51	15.52	108.98	20.84	73.39
100036520	45.75	73.14	13.69	15.83	6.13	4.88	6.18	27.62	50.42	15.98
100036521	310.35	354.68	341.59	450.34	460.29	425.82	523.26	460.31	524.48	453.24
100036523	41.79	35.49	120.55	38.80	72.74	58.67	21.22	63.25	16.33	3.62
100036528	67.83	39.32	76.61	3.28	4.85	13.17	12.28	8.94	3.86	14.46
100036537	123.73	126.10	6.24	8.07	9.02	111.83	49.64	35.01	6.90	42.52
100036538	59.89	4.93	36.62	28.61	7.13	16.47	6.94	15.94	20.00	3.55
100036539	88.96	4.46	3.88	15.07	7.63	7.45	6.73	8.75	35.09	31.22
100036541	9.69	85.15	12.94	16.64	10.11	13.31	16.71	15.10	10.23	15.15
100036543	846.73	745.63	955.15	451.50	431.89	337.55	556.60	567.48	492.50	559.11
100036768	6.58	4.74	50.25	105.53	75.51	103.62	11.04	5.87	7.59	29.19
100037258	975.19	1033.00	882.99	1562.16	1750.83	1805.78	919.91	1086.31	886.41	900.41
100037262	5.81	15.83	19.43	31.35	12.73	24.92	10.30	4.45	49.02	12.24
100037278	191.73	214.09	139.94	1304.20	663.73	762.83	82.56	69.42	51.81	214.07
100037283	253.28	230.69	190.57	210.47	229.74	252.62	163.37	237.51	179.18	186.49
100037297	61.21	103.80	15.82	56.46	150.49	86.64	54.91	74.26	78.79	73.06
100038347	199.62	105.62	82.71	354.22	382.36	202.67	69.28	164.40	101.17	192.71
100038358	12.09	22.69	75.47	39.93	3.95	11.95	63.15	33.52	1.99	32.95
100038371	84.98	3.43	1.94	3.83	7.02	4.17	54.54	16.20	24.85	22.28
100038392	88.54	92.11	14.28	77.26	53.68	137.55	51.13	87.73	37.22	52.44
100038394	103.90	48.64	75.24	104.68	73.98	111.79	44.04	122.76	125.52	135.57
100038398	45.98	23.02	88.76	109.37	9.51	94.10	37.17	58.05	17.89	58.54
100038402	12.88	68.78	97.89	33.81	19.36	9.63	9.77	8.71	6.22	9.52
100038416	152.59	123.08	105.62	102.60	45.23	78.35	93.61	150.61	42.33	4.72
100038419	41.12	18.98	7.11	14.21	7.91	12.13	4.46	43.52	4.49	25.96
100038480	4.53	27.60	5.43	5.12	32.78	31.92	1.54	2.15	17.45	17.64
100038489	17.40	43.00	33.50	23.43	9.04	6.18	7.23	14.25	7.84	25.96

- Co-expression Network 만들기

  - 필요 파일: Expression profile matrix, control 정보, co-expression network 제작 프로그램

  1. 실험 Control column 설정

     $ ./make_control GSE5657

  2. Log expression ratio 계산 및 Gene들간 Pearson correlation coefficient 계산

     $ ./Ratio_PCC_Each_Control.sh GSE5657

     - 내부 실행 파일

     - Probes2GenesByMeanValue.pl : 유전자 발현 평균 값 계산

     - Create_log2Ratio_Each_Control.pl : 발현 값에 log를 취함

     - coexpression_S_CC_log2 : PCC 계산 코드

- Output 형식

Log ratio matrix

geneIDs	GSM252064.CEL	GSM252065.CEL	GSM252066.CEL	GSM252067.CEL
100009600	-0.943	-2.759	-1.617	1.508
100012	-1.129	-2.304	0.043	-1.477
100017	-0.721	-0.710	-0.870	-0.286
100019	-0.586	-0.268	-0.588	1.074
100034360	-4.295	-4.002	-1.157	-4.475
100034675	-2.348	-0.042	-0.497	0.192
100034733	-3.384	-1.305	-2.949	0.682
100034748	-0.899	-0.998	0.067	-0.939
100036520	0.293	0.970	-1.448	-1.238
100036521	-0.436	-0.244	-0.298	0.101
100036523	-0.436	-0.672	1.093	-0.543
100036528	1.702	0.915	1.877	-2.666
100036537	0.878	0.905	-3.431	-3.061
100036538	1.276	-2.327	0.566	0.210
100036539	1.311	-3.006	-3.208	-1.251
100036541	-1.898	1.237	-1.481	-1.118
100036543	1.014	0.830	1.187	0.106
100036768	-2.207	-2.680	0.726	1.797
100037258	-1.057	-0.974	-1.200	-0.377
100037262	-2.894	-1.447	-1.151	-0.461
100037278	-0.447	-0.288	-0.901	2.319
100037283	0.761	0.626	0.350	0.494
100037297	-0.160	0.602	-2.111	-0.276
100038347	-2.398	-3.317	-3.669	-1.571
100038358	-1.207	-0.299	1.434	0.516
100038371	0.927	-3.704	-4.523	-3.546
100038392	0.516	0.573	-2.117	0.319

PCC of gene pairs

Gene A	Gene B	PCC
14969	16149	0.985473
12259	12262	0.975932
14961	14969	0.97525
66108	72900	0.973375
14961	16149	0.972895
21973	68612	0.972211
12826	12827	0.970992
107995	52276	0.970453
14960	14961	0.970346
65963	66058	0.970163
14960	16149	0.968855
12260	12262	0.968037
17219	67177	0.967287
68202	68342	0.966784
17218	17219	0.966399
19988	67115	0.966236
66046	66414	0.965862
14960	14969	0.965475
227197	68202	0.964901
27425	66495	0.964883
11950	57423	0.964258
17105	17110	0.963989
12862	21393	0.963506
11946	66445	0.962427
66108	67680	0.962269
19981	54127	0.962156
14127	22177	0.962136
12259	12260	0.961688

- 만들어진 Network를 Benchmark하여 Log-likelihood score를 구하기

  - 필요 파일 : Gold standard pair file, Benchmark code,

  1. 각 Bin들의 Log-likelihood score와 누적 Log-likelihood score 구하기

     Format: ./benchmark_binLLS_cLLS [binsize] [genomesize] [gold standard] [network] [report] [linkcnt limit]

     $ ./Benchmark_BinLLS_0.632BST.py 1000 20058 mouse_trainingset GSE5657/log2ratio.pcc.sort.dat GSE5657/log2ratio.pcc.sort.dat.report 30000

  2. LLS Benchmark Plot을 그려보고 sigmoidal function으로 regression model 구하기

     $ ./draw_benchmark_binLLS.py -i GSE5657/log2ratio.pcc.sort.dat.report -o GSE5657 -a

  3. Network를 sigmoidal 4 parameter function으로 regression 하기

     Format: ./regression_network [network file] [x threshold] [a] [b] [x0] [y0] > output

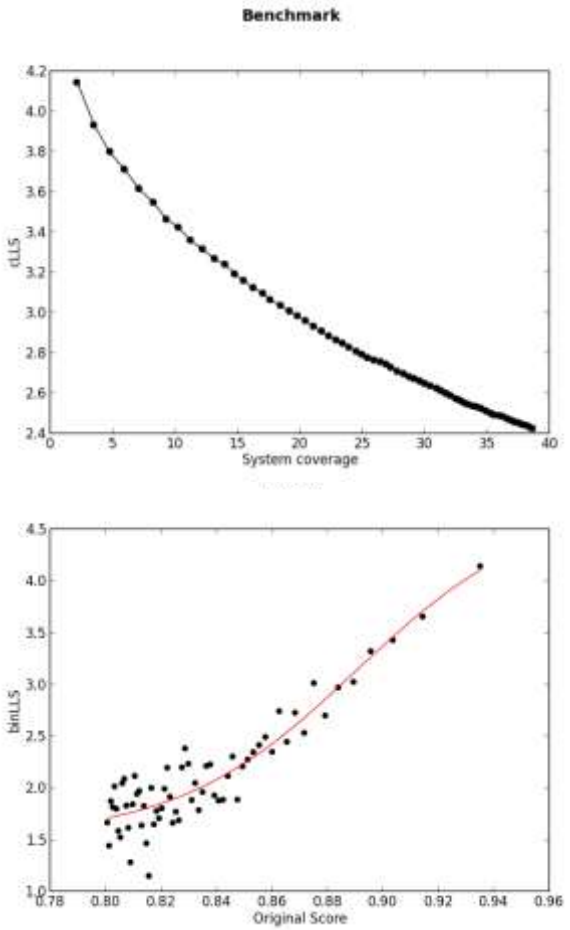
     $ ./regression_sigmoid4.pl GSE5657/log2ratio.pcc.sort.dat 0.93 -46395.00 -0.01239 0.80 2.19 > CoexpressionNetwork_GSE5657

- Output 형식

Benchmark file

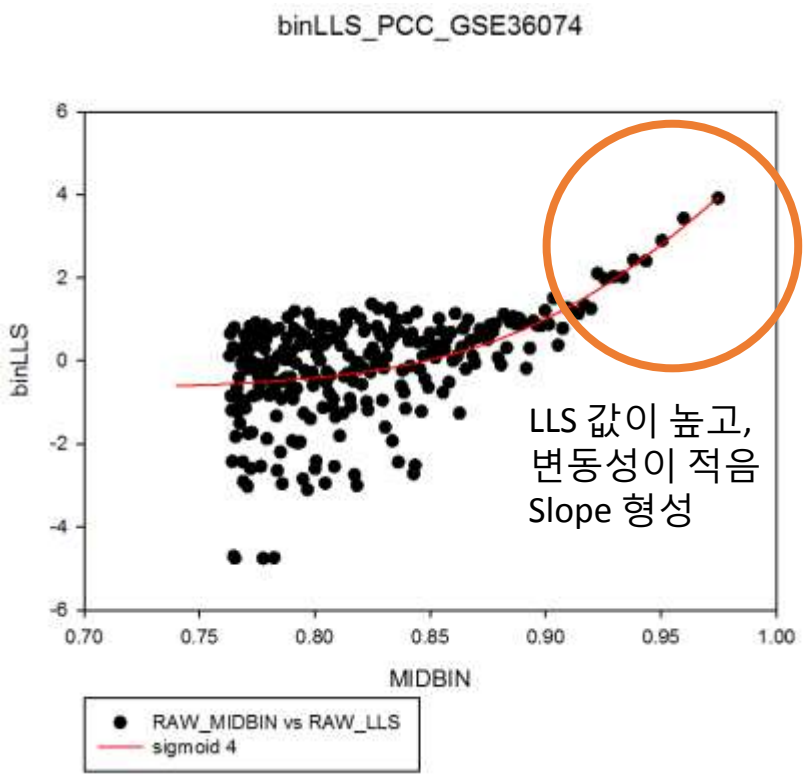
mid_bin	System coverage	cLLS	PCC	binLLS
500	2.153754	4.144708	0.934949	4.144708
1500	3.509822	3.933153	0.914221	3.658233
2500	4.746236	3.798278	0.903452	3.431195
3500	5.967694	3.711857	0.895454	3.322761
4500	7.109383	3.616394	0.889264	3.027053
5500	8.251072	3.547254	0.883775	2.973863
6500	9.317978	3.464404	0.879058	2.702445
7500	10.28517	3.422485	0.874964	3.015216
8500	11.24738	3.359104	0.871512	2.534332
9500	12.19464	3.315568	0.868207	2.729588
10500	13.16183	3.267433	0.865249	2.447238
11500	13.97946	3.240102	0.862482	2.743632
12500	14.77715	3.193578	0.859918	2.351368
13500	15.50504	3.158367	0.857521	2.494666
14500	16.24788	3.125918	0.855225	2.416518
15500	17.04557	3.096223	0.853138	2.347982
16500	17.62888	3.06411	0.851105	2.276943
17500	18.44152	3.037317	0.849237	2.210729
18500	19.13451	3.007182	0.847417	1.889257
19500	19.78263	2.984636	0.845641	2.30575
20500	20.45568	2.958791	0.843962	2.117896
21500	21.0689	2.931797	0.84224	1.888479
22500	21.71702	2.90703	0.840619	1.879228
23500	22.32526	2.882527	0.8391	1.929993
24500	22.87865	2.86481	0.837593	2.229127
25500	23.4121	2.848771	0.836165	2.214212

Benchmark Plot

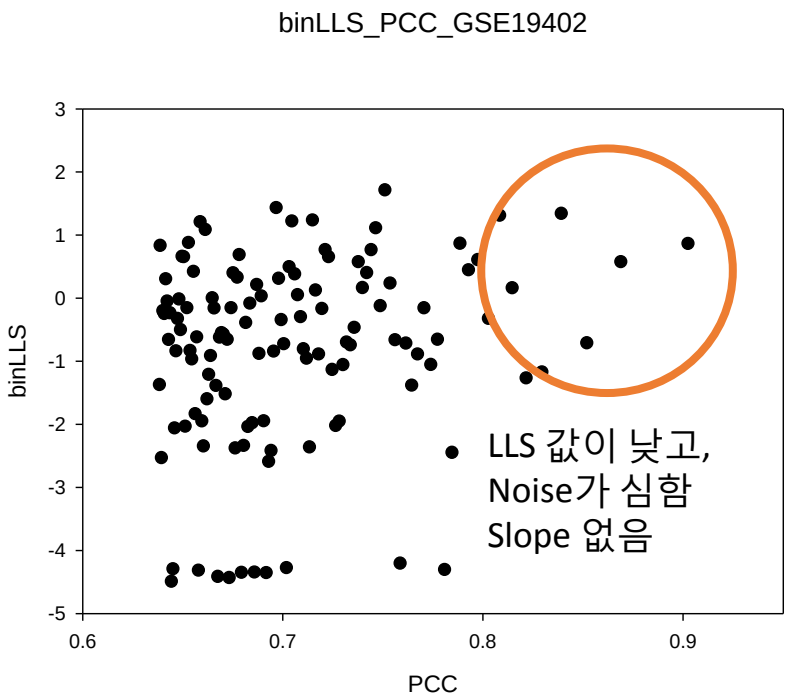


Bin(Dot) 하나당 1,000개의 Network links

- 잘 맞는 모델, 잘 맞지 않는 모델

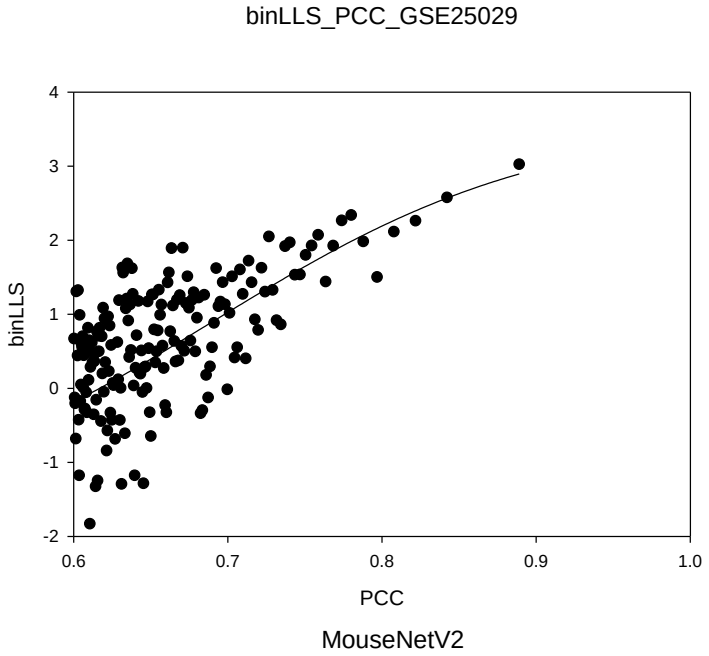


좋은 Network

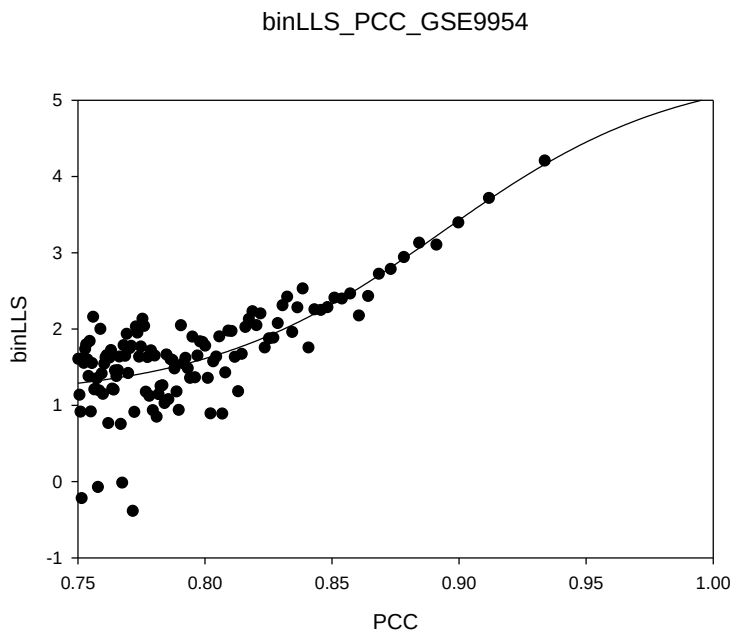


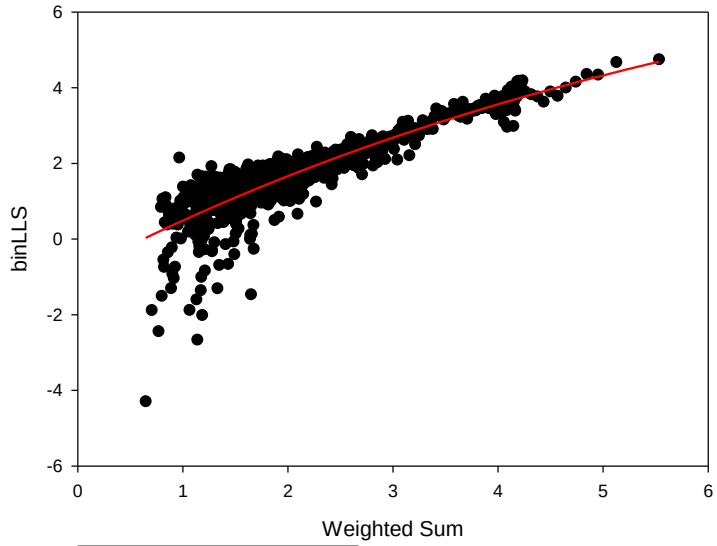
안좋은 Network

- 네트워크의 X threshold를 정하는 방법
  - Noise가 급격히 증가하는 부분
  - 보통 Bin의 LLS가 0 이하로 등장하는 부분에서 자른다



MouseNetV2





- Network Integration 하기

  - 필요 파일: Integration 프로그램, Benchmark 프로그램, Gold standard, Regression 프로그램

  1. 하나의 파일로 Network 합치기

     Format: ./join_pair_scores [output file] [# of networks] [network 1] [network 2] [network 3] …..

     $ ./join_pair_scores.pl join_network 3 GSE11494_network GSE3414_network GSE8836_network

	Study Title (Innate immune response related studies)
GSE11494	Innate immune response of murine nasal-associated lymphoid tissue (NALT) to Streptococcus pyogenes infection.
GSE3414	Immune Response to Nippostrongylus brasiliensis in the mouse lung
GSE8836	CLL in Em-TCL1 mice provides a biologically relevant model to unravel and reverse immune deficiency in human cancer.

- Network Integration 하기

  - 필요 파일: Integration 프로그램, Benchmark 프로그램, Gold standard, Regression 프로그램

  2. 여러 가지 weighting factor 적용 후 Benchmark 해보기

     Format: ./concensus_byWS [# of networks] [join file] [weighting factor] [LLS cutoff] > output file

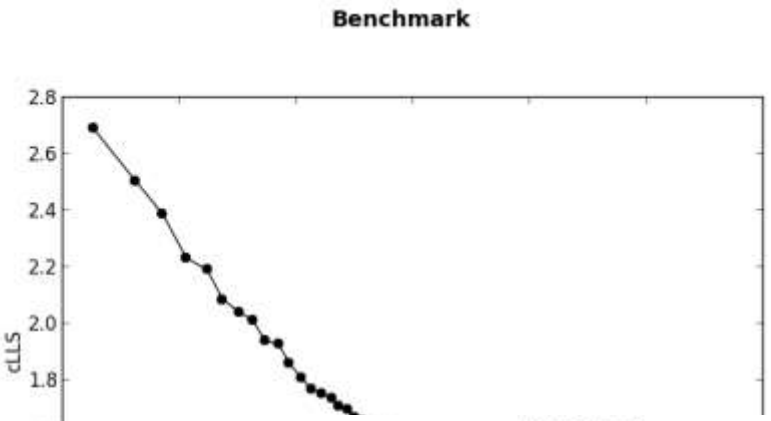
     $ ./concensus_byWS.pl 3 join_network 2 0 > weightedsum_network_D2

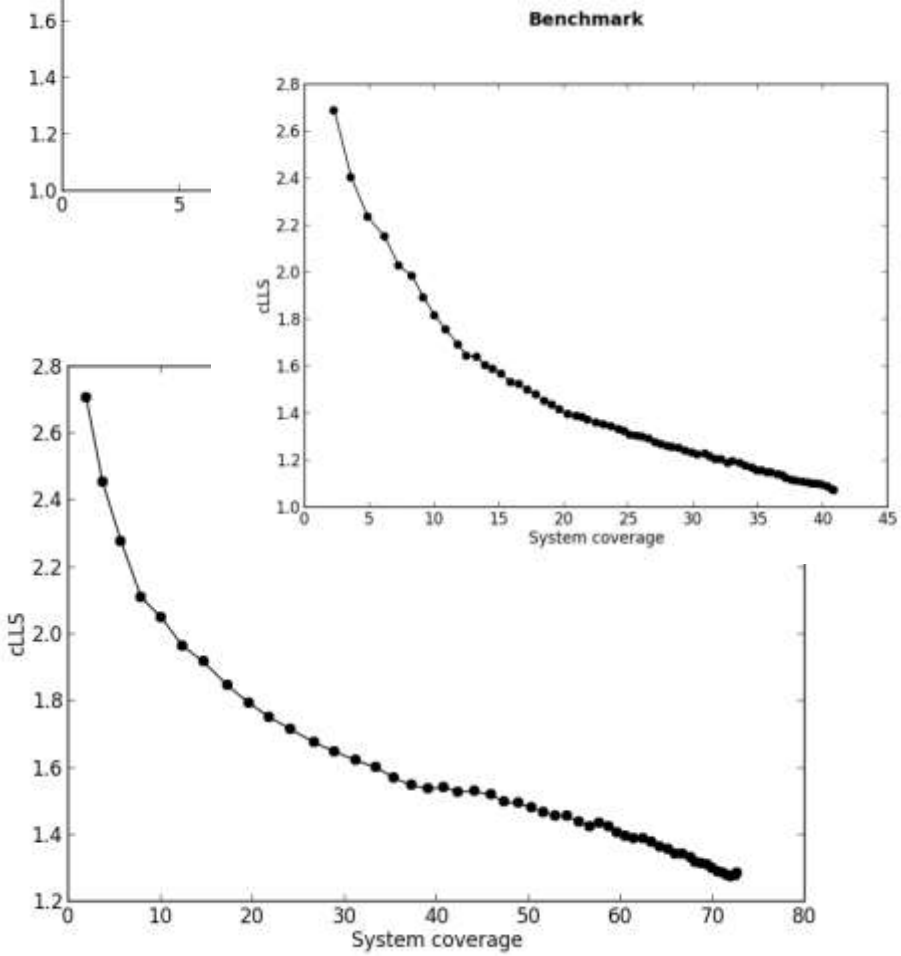
     $ ./cut_sorted_benchmark_regression.sh weightedsum_network_D2

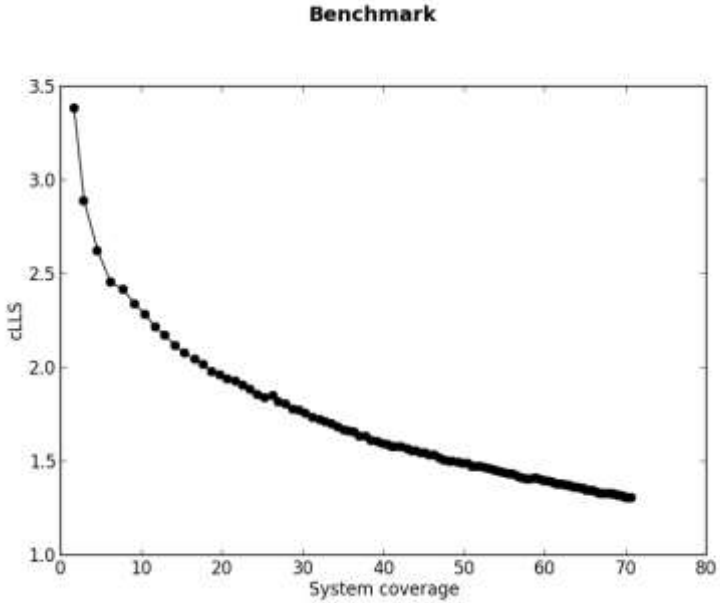
  3. 적절한 weighting factor 결정하기

  4. Network에 regression model 적용하기

     $ ./regression_sigmoid4.pl weightedsum_network_D2.cut.sorted 0.5 12.39 2.41 4.54 -1.56 > integrated_network







Integrated network

5. 참고문헌

- Joshua M. Stuart and Eran Segal et al., Science, 2003
- Insuk Lee and et al., *Science*, 2004
- Insuk Lee and Ben Lehner et al., *Nat Genet*., 2008
- Tak Lee et al., *Nucleic Acids Res.,* 2015
- Eiru Kim et al., *Nucleic Acids Res.,* 2016
- https://www.bioconductor.org/
- http://atted.jp/overview.shtml