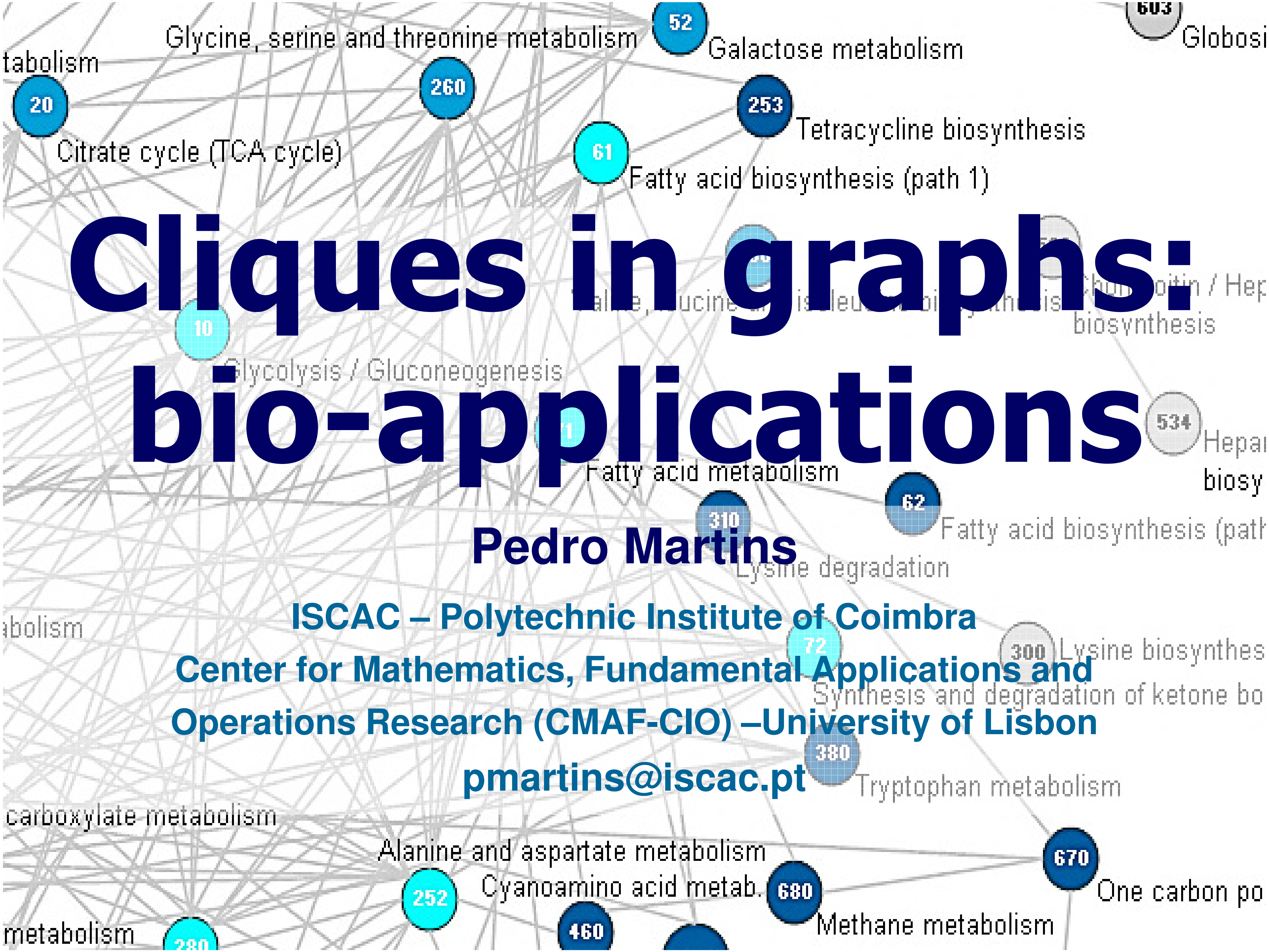


Cliques in graphs: bio-applications

Pedro Martins

ISCAC – Polytechnic Institute of Coimbra

Center for Mathematics, Fundamental Applications and Operations Research (CMAF-CIO) –University of Lisbon

pmartins@iscac.pt

Maximum Clique Problem

Given a simple and undirected graph $G = (V, E)$,
with $V = \{1, \ldots, n\}$ the set of nodes and $E \subseteq V \times V$ the set of edges

A clique of G is a subset $C \subseteq V$ whose elements are pairwise adjacent

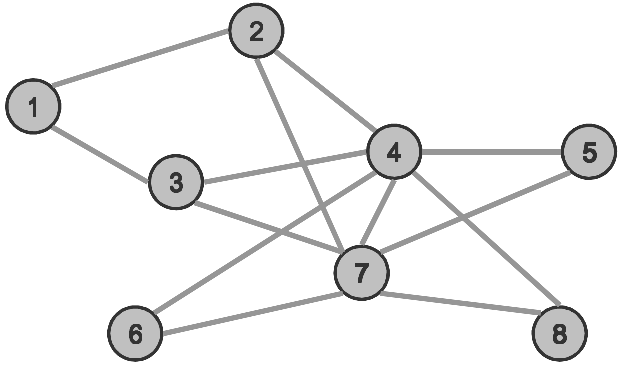
Luce and Perry (1949)

$C \subseteq V$ is a **clique** if $(i,j) \in E$, for all $i,j \in C$

$\omega(G) = \max\{ |C| : C \text{ is a clique of } G \}$

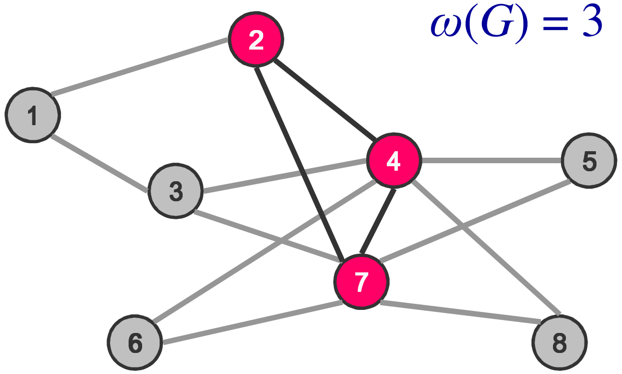
A clique C is maximum if it is the largest clique in G

$G = (V, E)$

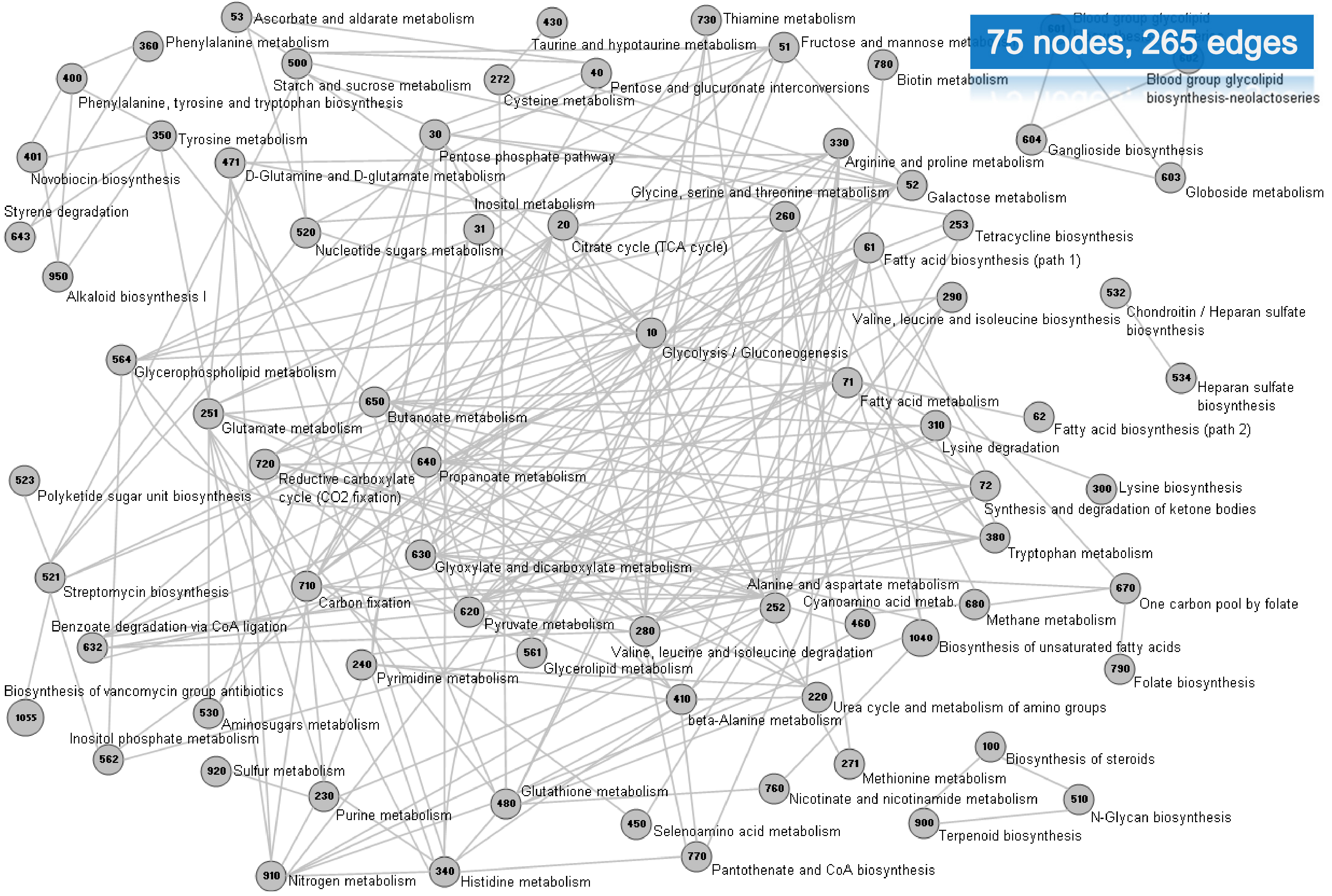


maximum clique

$\omega(G) = 3$

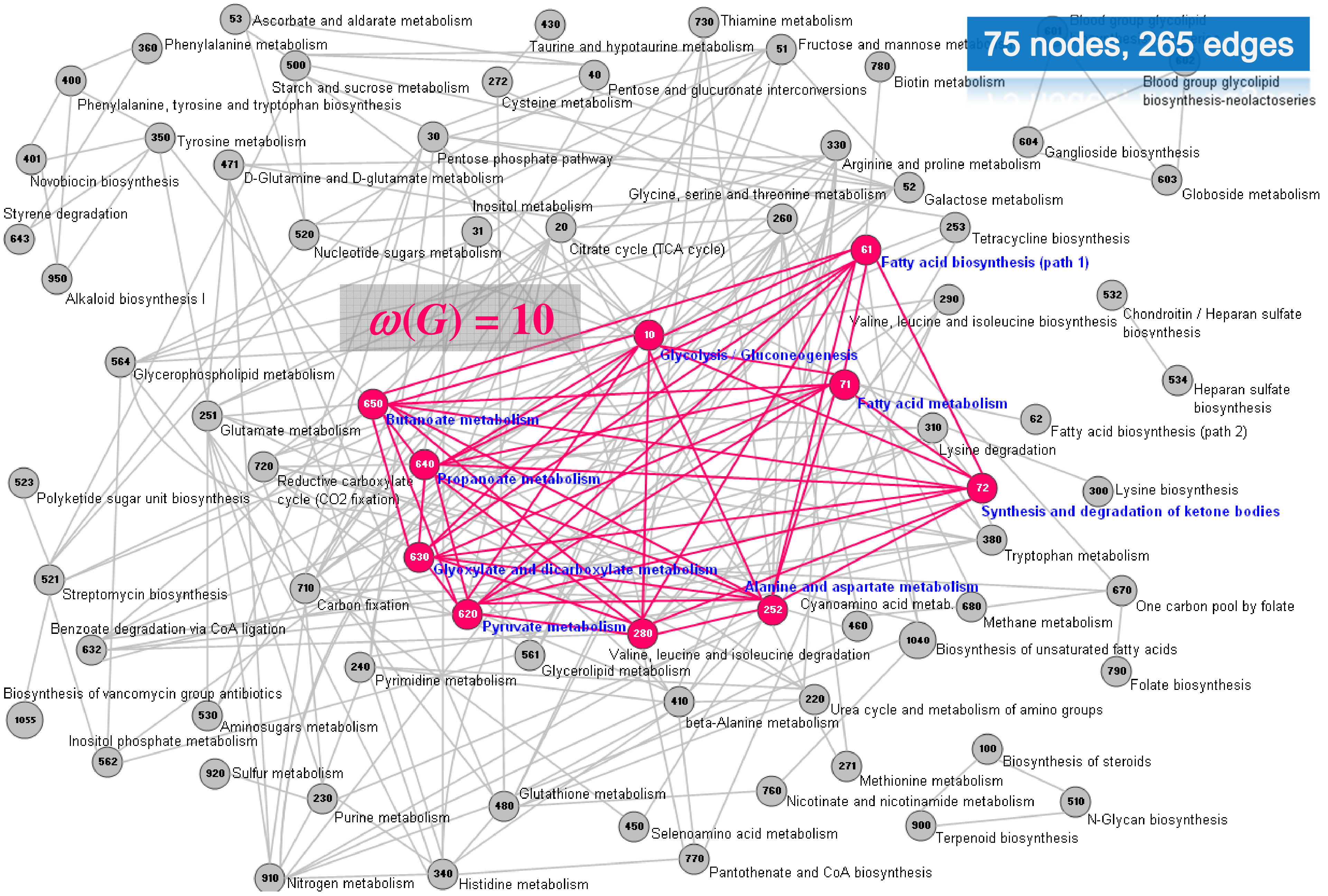


Homo sapiens – Network of Interacting Pathways (NIP)


53 Ascorbate and aldarate metabolism
430
730 Thiamine metabolism
601
Blood group glycolipid
602
360 Phenylalanine metabolism
Taurine and hypotaurine metabolism
51 Fructose and mannose metabolism
500
780
400
272
40
Starch and sucrose metabolism
Biotin metabolism
Blood group glycolipid
biosynthesis-neolactoseries
Phenylalanine, tyrosine and tryptophan biosynthesis
Cysteine metabolism
Pentose and glucuronate interconversions
350 Tyrosine metabolism
30
330
604 Ganglioside biosynthesis
401
471
Pentose phosphate pathway
Arginine and proline metabolism
D-Glutamine and D-glutamate metabolism
52
603
Novobiocin biosynthesis
Glycine, serine and threonine metabolism
Galactose metabolism
Globoside metabolism
Inositol metabolism
Styrene degradation
260
520
31
20
253 Tetracycline biosynthesis
643
Citrate cycle (TCA cycle)
61
Nucleotide sugars metabolism
Fatty acid biosynthesis (path 1)
950
290
532
Alkaloid biosynthesis I
Chondroitin / Heparan sulfate
biosynthesis
Valine, leucine and isoleucine biosynthesis
10
564
Glycolysis / Gluconeogenesis
Glycerophospholipid metabolism
534 Heparan sulfate
biosynthesis
71
650
Fatty acid metabolism
251
62
Butanoate metabolism
Fatty acid biosynthesis (path 2)
310
Glutamate metabolism
Lysine degradation
720
640
Reductive carboxylate
cycle (CO2 fixation)
Propanoate metabolism
523
72
300 Lysine biosynthesis
Polyketide sugar unit biosynthesis
Synthesis and degradation of ketone bodies
380
630
Tryptophan metabolism
521
Glyoxylate and dicarboxylate metabolism
710
Alanine and aspartate metabolism
670
Streptomycin biosynthesis
Carbon fixation
252
Cyanoamino acid metab.
680
One carbon pool by folate
620
Methane metabolism
Benzoate degradation via CoA ligation
Pyruvate metabolism
280
460
1040
Biosynthesis of unsaturated fatty acids
632
561
Valine, leucine and isoleucine degradation
240
Glycerolipid metabolism
790
Pyrimidine metabolism
Folate biosynthesis
Biosynthesis of vancomycin group antibiotics
410
220
1055
530
Urea cycle and metabolism of amino groups
beta-Alanine metabolism
Aminosugars metabolism
Inositol phosphate metabolism
100
562
Biosynthesis of steroids
271
920 Sulfur metabolism
Methionine metabolism
230
760
510
480
Glutathione metabolism
Nicotinate and nicotinamide metabolism
Purine metabolism
N-Glycan biosynthesis
450
900
Selenoamino acid metabolism
Terpenoid biosynthesis
770
910
340
Pantothenate and CoA biosynthesis
Nitrogen metabolism
Histidine metabolism

75 nodes, 265 edges

Homo sapiens NIP – Maximum Clique


75 nodes, 265 edges

$\omega(G) = 10$

53 Ascorbate and aldarate metabolism
430
730 Thiamine metabolism
360 Phenylalanine metabolism
500
Starch and sucrose metabolism
Taurine and hypotaurine metabolism
51 Fructose and mannose met
780
Biotin metabolism
400
272
40
Cysteine metabolism
Pentose and glucuronate interconversions
Blood group glycolipid
biosynthesis-neolactoseries
Phenylalanine, tyrosine and tryptophan biosynthesis
350 Tyrosine metabolism
30
Pentose phosphate pathway
330
Arginine and proline metabolism
604 Ganglioside biosynthesis
401
471 D-Glutamine and D-glutamate metabolism
603 Globoside metabolism
Novobiocin biosynthesis
52 Galactose metabolism
Glycine, serine and threonine metabolism
Styrene degradation
Inositol metabolism
643
520
31
20
260
253 Tetracycline biosynthesis
Nucleotide sugars metabolism
Citrate cycle (TCA cycle)
61 Fatty acid biosynthesis (path 1)
950
Alkaloid biosynthesis I
290
532
Valine, leucine and isoleucine biosynthesis
Chondroitin / Heparan sulfate
biosynthesis
10
Glycolysis / Gluconeogenesis
564 Glycerophospholipid metabolism
71
534 Heparan sulfate
biosynthesis
650
Fatty acid metabolism
251
Butanoate metabolism
310
62 Fatty acid biosynthesis (path 2)
Glutamate metabolism
Lysine degradation
720
640
Propanoate metabolism
Reductive carboxylate
cycle (CO2 fixation)
523
72
300 Lysine biosynthesis
Polyketide sugar unit biosynthesis
Synthesis and degradation of ketone bodies
380
630
Tryptophan metabolism
Glyoxylate and dicarboxylate metabolism
521
710
Alanine and aspartate metabolism
670
Streptomycin biosynthesis
Carbon fixation
252 Cyanoamino acid metab.
680
One carbon pool by folate
620
Methane metabolism
Pyruvate metabolism
280
460
Benzoate degradation via CoA ligation
1040 Biosynthesis of unsaturated fatty acids
632
561
Valine, leucine and isoleucine degradation
240
Glycerolipid metabolism
790
Pyrimidine metabolism
Folate biosynthesis
Biosynthesis of vancomycin group antibiotics
410
220
Urea cycle and metabolism of amino groups
1055
530
beta-Alanine metabolism
Aminosugars metabolism
Inositol phosphate metabolism
100
Biosynthesis of steroids
562
271
920 Sulfur metabolism
Methionine metabolism
760
510
230
480
Glutathione metabolism
Nicotinate and nicotinamide metabolism
N-Glycan biosynthesis
Purine metabolism
900
450 Selenoamino acid metabolism
Terpenoid biosynthesis
770
Pantothenate and CoA biosynthesis
910
340
Nitrogen metabolism
Histidine metabolism

Maximum Independent Set

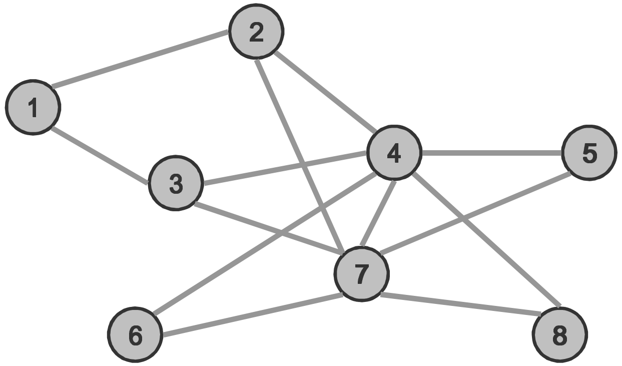
An independent set is a subset $S \subseteq V$ where all its nodes are pairwise nonadjacent

Independent Set

$S \subseteq V$ is an **independent set** if, for all $i,j \in S$, $(i,j) \notin E$

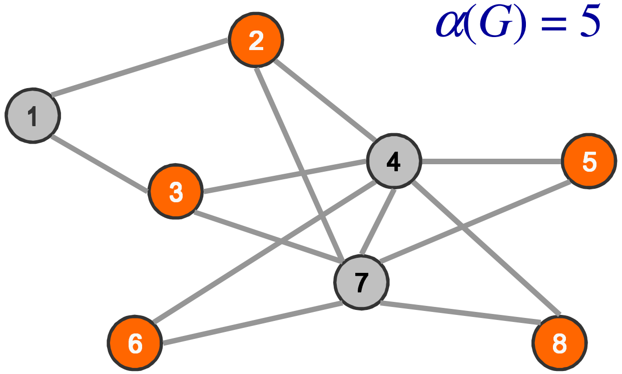
$$\alpha(G) = \max\{\ |S| : S \text{ is an independent set in } G\ \}$$

G = (*V*, *E*)

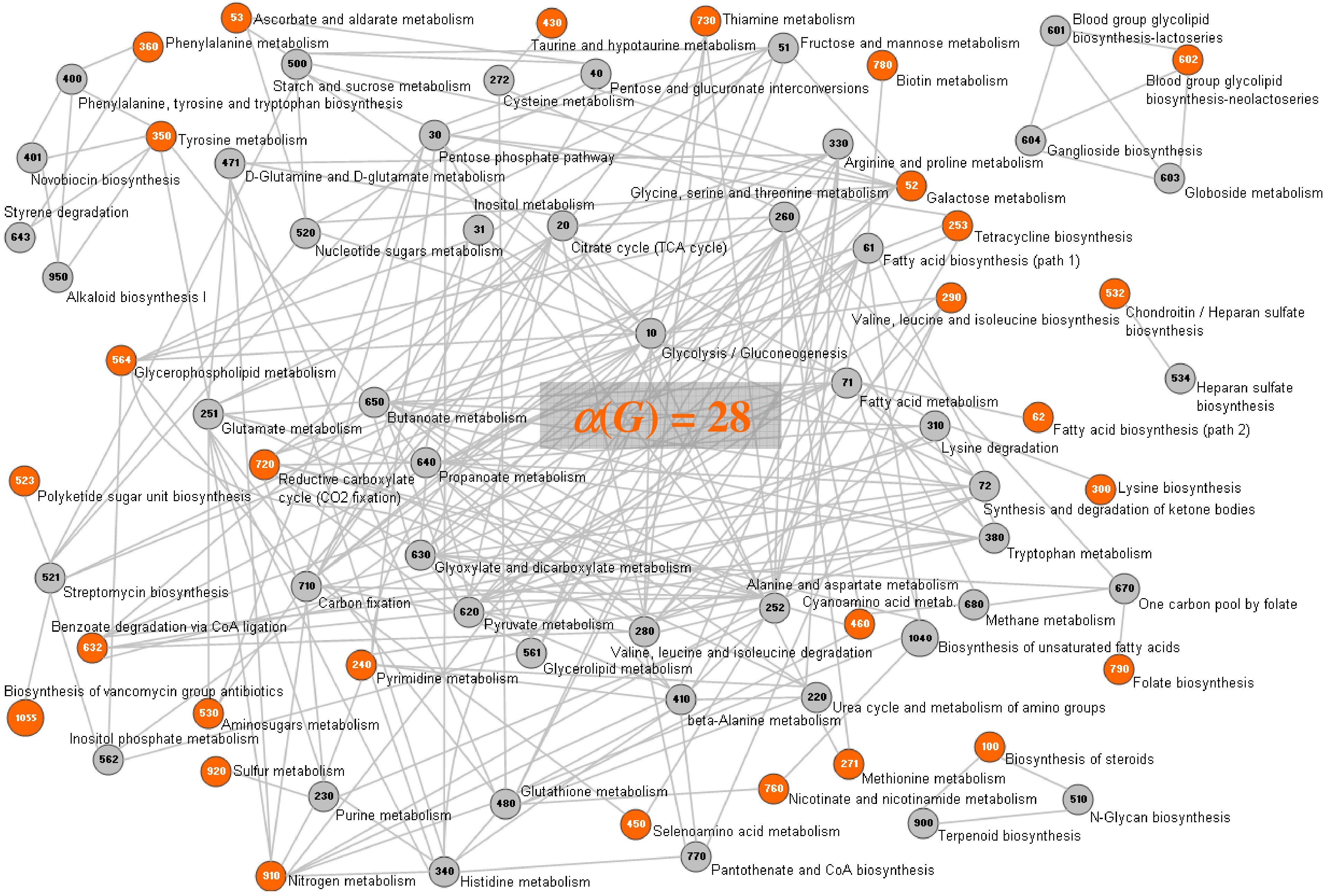


maximum independent set

$\alpha(G) = 5$



Homo sapiens NIP – Maximum Independent Set



53 Ascorbate and aldarate metabolism
360 Phenylalanine metabolism
430 Taurine and hypotaurine metabolism
730 Thiamine metabolism
601 Blood group glycolipid biosynthesis-lactoseries
51 Fructose and mannose metabolism
500 Starch and sucrose metabolism
272 Cysteine metabolism
40 Pentose and glucuronate interconversions
780 Biotin metabolism
602 Blood group glycolipid biosynthesis-neolactoseries
400 Phenylalanine, tyrosine and tryptophan biosynthesis
350 Tyrosine metabolism
30 Pentose phosphate pathway
330 Arginine and proline metabolism
604 Ganglioside biosynthesis
401 Novobiocin biosynthesis
471 D-Glutamine and D-glutamate metabolism
52 Galactose metabolism
603 Globoside metabolism
Styrene degradation
643
Inositol metabolism
Glycine, serine and threonine metabolism
520 Nucleotide sugars metabolism
31
20 Citrate cycle (TCA cycle)
260
253 Tetracycline biosynthesis
61 Fatty acid biosynthesis (path 1)
950 Alkaloid biosynthesis I
290 Valine, leucine and isoleucine biosynthesis
532 Chondroitin / Heparan sulfate biosynthesis
10 Glycolysis / Gluconeogenesis
564 Glycerophospholipid metabolism
534 Heparan sulfate biosynthesis
71 Fatty acid metabolism
650 Butanoate metabolism
251 Glutamate metabolism
α(G) = 28
62 Fatty acid biosynthesis (path 2)
310 Lysine degradation
720 Reductive carboxylate cycle (CO2 fixation)
640 Propanoate metabolism
523 Polyketide sugar unit biosynthesis
72 Synthesis and degradation of ketone bodies
300 Lysine biosynthesis
380 Tryptophan metabolism
630 Glyoxylate and dicarboxylate metabolism
521 Streptomycin biosynthesis
710 Carbon fixation
Alanine and aspartate metabolism
670 One carbon pool by folate
252 Cyanoamino acid metab.
680 Methane metabolism
620 Pyruvate metabolism
280 Valine, leucine and isoleucine degradation
460
Benzoate degradation via CoA ligation
632
561 Glycerolipid metabolism
1040 Biosynthesis of unsaturated fatty acids
240 Pyrimidine metabolism
790 Folate biosynthesis
Biosynthesis of vancomycin group antibiotics
1055
530 Aminosugars metabolism
410 beta-Alanine metabolism
220 Urea cycle and metabolism of amino groups
Inositol phosphate metabolism
562
100 Biosynthesis of steroids
271 Methionine metabolism
920 Sulfur metabolism
230 Purine metabolism
480 Glutathione metabolism
760 Nicotinate and nicotinamide metabolism
510 N-Glycan biosynthesis
450 Selenoamino acid metabolism
900 Terpenoid biosynthesis
770 Pantothenate and CoA biosynthesis
910 Nitrogen metabolism
340 Histidine metabolism

Maximum Clique Problem – Complexity

The maximum clique problem is *NP*-hard (Karp, 1972)

There is no polynomial-time approximation algorithm to solve the maximum clique problem, unless $P = NP$ (Panconesi and Ranjan, 1990, 1993)

Even if $\omega(G)$ is known in advance, we would still have a difficult problem to solve, if our goal is to find a ω-sized clique C in G (belongs to the W[1]-hard class)

Just for fun:

Given a graph G with 500 nodes. Lets assume that we know $\omega(G) = 40$

If we try to ***enumerate*** (*!!!*) all candidate combinations, we would obtain

$$\binom{500}{40} = \frac{500!}{40!\ 460!} > 2\times 10^{59}$$ different sets

Even if we could eliminate 75% of the nodes, we would still have $\binom{125}{40} = \frac{125!}{40!\ 85!} > 8\times 10^{32}$ candidate sets

Cliques – Applications

Computing, information systems, telecommunications and robotics

- Signal processing and image processing (Balas and Yu, 1986), (Hotta et al., 2003)
- Computer vision and pattern recognition (Bolles and Horaud, 1986), (Ogawa, 1986), (Pla and Marchand, 1997), (Pellilo, Siddiqi and Zucker, 1998), (Busiygin, Butenko and Pardalos, 2002)
- Fault diagnosis and detection (Berman and Pelc, 1990)
- Mobile telecommunication networks and wireless networks (Krishna et al., 1997)
- Satellite communication network of the GPS III system (Brinkmann, Crevals and Frye, 2012)
- Hypertext and the World Wide Web (Gibson, Kleinberg and Raghavan, 1998)
- Data mining (Cook and Holder, 2000), (Washio and Motoda, 2003)

Other

- Financial markets, Marketing, Criminal networks, Transmitted disease networks, …

Cliques – Applications

Biology, Genetics and Biochemistry

- Spot matching for 2D gel electrophoresis images (Bahadur et al., 2002)
- Protein structure alignment (Bahadur et al., 2002), (Caprara and Lancia, 2002), (Strickland, Barnes and Sokol, 2005)
- Nonoverlapping local alignments (Butenko and Wilhelm, 2005)
- Matching 3D molecular structures (Butenko and Wilhelm, 2005)
- Integration of genome mapping data (Butenko and Wilhelm, 2005)
- Data mining in molecular structures (Fischer and Meinl, 2004)

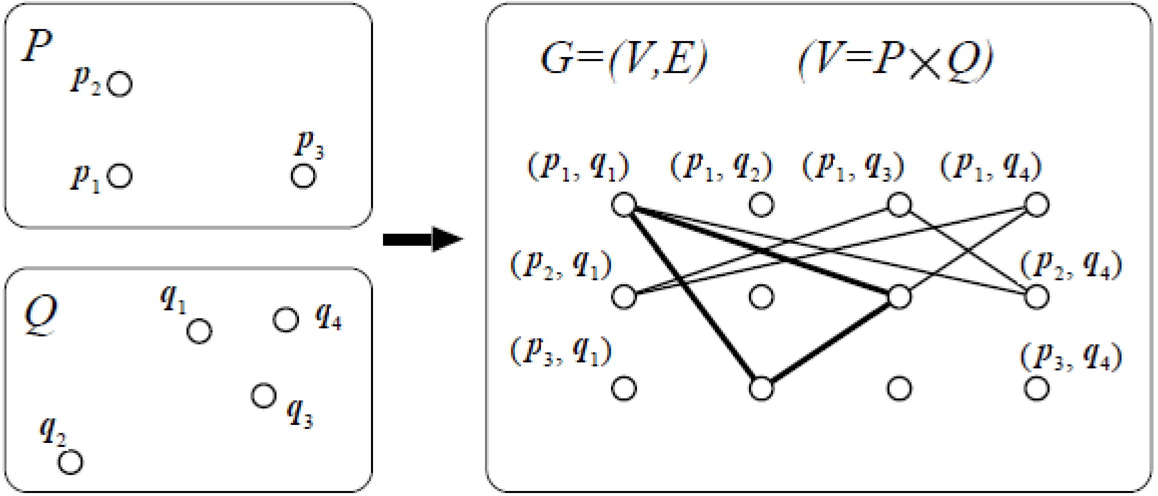
Cliques' drawbacks

- the structure is too rigid

- some times we are simply looking for a dense region in the graph

Cliques – Applications – point matching

Point matching - reduction from point matching to maximum clique



Vertices $(\boldsymbol{p}_i, \boldsymbol{q}_j), (\boldsymbol{p}_k, \boldsymbol{q}_h) \in V$ are connected by an edge if and only if distances $|\boldsymbol{p}_k - \boldsymbol{p}_i|$ and $|\boldsymbol{q}_h - \boldsymbol{q}_j|$ are similar to each other.

The maximum clique $\{(\boldsymbol{p}_1, \boldsymbol{q}_1), (\boldsymbol{p}_2, \boldsymbol{q}_3), (\boldsymbol{p}_3, \boldsymbol{q}_2)\}$ of G corresponds to the maximum match between P and Q.

(in, **K.C. Dukka Bahadur, T. Akutsu, E. Tomita, T. Seki and A. Fujiyama**, "Point Matching Under Non-Uniform Distortions and Protein Side Chain Packing Based on an Efficient Maximum Clique Algorithm", *Genome Informatics*, 13: 143-152, 2002)

Cliques – Applications – protein structure alignment

Protein structure alignment

(in, D.M. Strickland, E. Barnes, and J.S. Sokol, “Optimal protein structure alignment using maximum cliques”, *Operations Research*, 53(3): 389-402, 2005)

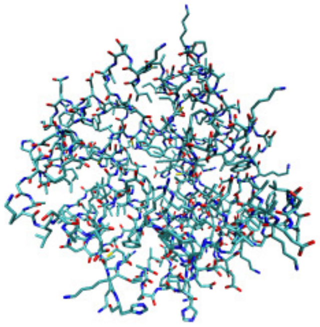
Proteins with very similar tertiary structure usually have analogous functions

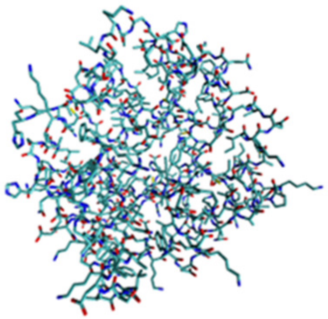
important for protein based clinical treatments

One of the known methods: Maximum Contact Map Overlap (CMO)

it resorts to the maximum clique problem on an appropriate graph

It assesses the similarity between the tertiary structure of two proteins, comparing the proximity among non consecutive amino acids





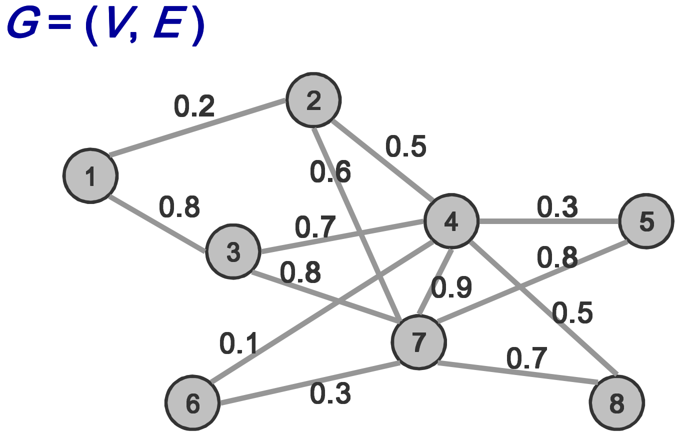
Maximum Edge-Weight Clique (MEWC) Problem

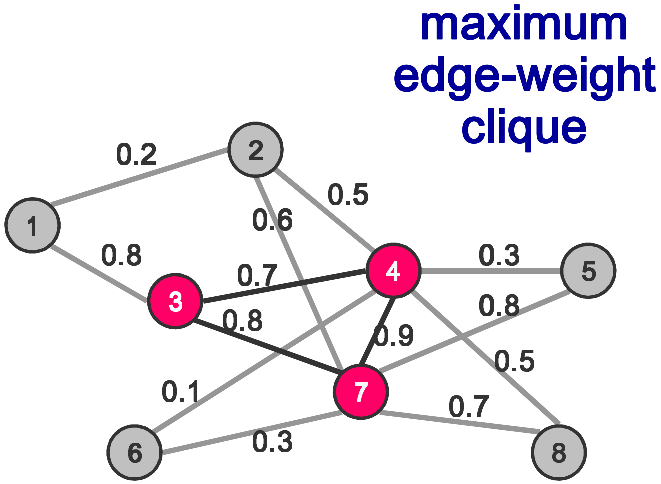
If we assign weights a_{ij} to each edge $(i,j)\in E$

maximum edge-weight clique

Let C be a clique and $A(C)=\sum_{i,j\in C} a_{ij}$

we want to find a clique C with maximum $A(C)$





MEWC – Applications – Network of Interacting Metabolites

Saccharomyces cerevisiae metabolic networks

The data involves 1394 metabolic reactions that use 991 metabolites. Each metabolic reaction is a chemical pathway that uses reactants to generate products. Both reactants and products are metabolites, being shared among reactions.

$$\mathbf{A} + \mathbf{B} \rightarrow \mathbf{C} + \mathbf{D}$$

Network of interacting metabolites:

the set of nodes *V* represents all metabolites ($|V| = 991$)

the set of edges *E* represents pairs of metabolites sharing at least one reaction ($|E| = 4161$) (density: 0.00848)

weights on the edges: $a_{ij} \equiv$ number of reactions sharing metabolites *i* and *j*

Taken from: Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J., 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Research (13), 244-253

MEWC – Applications – Network of Interacting Metabolites

Saccharomyces cerevisiae metabolic networks

The data involves 1394 metabolic reactions that use 991 metabolites. Each metabolic reaction is a chemical pathway that uses reactants to generate products. Both reactants and products are metabolites, being shared among reactions.

List of metabolites:

```
10=Formyltetrahydrofolate
10=FormyltetrahydrofolateM
1=(2=Carboxyphenylamino)=1=deoxy=D=ribulose_5=phosphate
1,3=beta=D=Glucan
1,3=Diaminopropane
1=(5=Phospho=D=ribosyl)=5=amino=4=imidazolecarboxylate
1=(5'=Phosphoribosyl)=5=amino=4=imidazolecarboxamide
1=(5'=Phosphoribosyl)=5=amino=4=(N=succinocarboxamide)=imidazole
1=(5'=Phosphoribosyl)=5=formamido=4=imidazolecarboxamide
1=alpha=D=Galactosyl=myo=inositol
```

• • •

Taken from: Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J., 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Research (13), 244-253

MEWC – Applications – Network of Interacting Metabolites

Saccharomyces cerevisiae metabolic networks

The data involves 1394 metabolic reactions that use 991 metabolites. Each metabolic reaction is a chemical pathway that uses reactants to generate products. Both reactants and products are metabolites, being shared among reactions.

List of reactions:

```
"ADP" + "ATPM" + "Orthophosphate" -> "ADPM" + "ATP" + "H+M" + "OrthophosphateM"
"Adenosine" -> "Inosine" + "NH3"
"Deoxyadenosine" -> "Deoxyinosine" + "NH3"
"Adenine" -> "HYXN" + "NH3"
"GlutamateM" + "OxaloacetateM" -> "2=OxoglutarateM" + "L=AspartateM"
"2=OxoglutarateM" + "L=AspartateM" -> "GlutamateM" + "OxaloacetateM"
"3=(4=Hydroxyphenyl)pyruvate" + "L=Glutamate" -> "2=Oxoglutarate" + "L=Tyrosine"
"2=Oxoglutarate" + "L=Tyrosine" -> "3=(4=Hydroxyphenyl)pyruvate" + "L=Glutamate"
"L=Glutamate" + "Oxaloacetate" -> "2=Oxoglutarate" + "L=Aspartate"
"2=Oxoglutarate" + "L=Aspartate" -> "L=Glutamate" + "Oxaloacetate“
"L=Glutamate" + "Oxaloacetate" -> "2=Oxoglutarate" + "L=Aspartate"
"2=Oxoglutarate" + "L=Aspartate" -> "L=Glutamate" + "Oxaloacetate"
"Chorismate" + "L=Glutamine" -> "4=amino=4=deoxychorismate" + "L=Glutamate"
"Acetyl=CoA" + "ATP" + "CO2" -> "ADP" + "Malonyl=CoA" + "Orthophosphate"
```

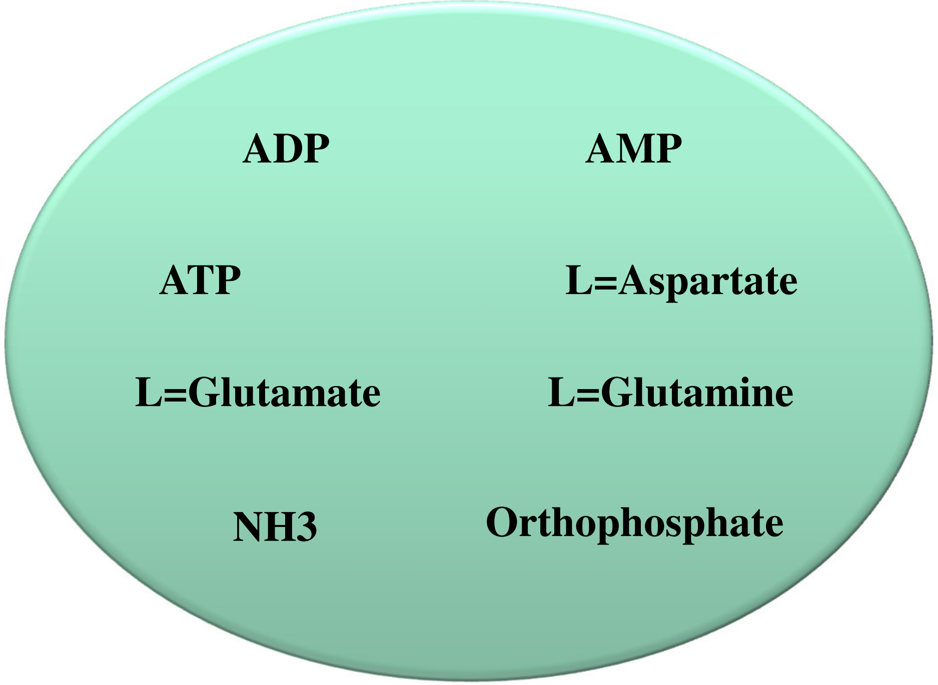
• • •

Taken from: Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J., 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Research (13), 244-253

MEWC – Applications – Network of Interacting Metabolites

Network of interacting metabolites: 991 nodes and 4161 edges (*d* = 0.0085)

maximum edge-weight clique



ADP

AMP

ATP

L=Aspartate

L=Glutamate

L=Glutamine

NH3

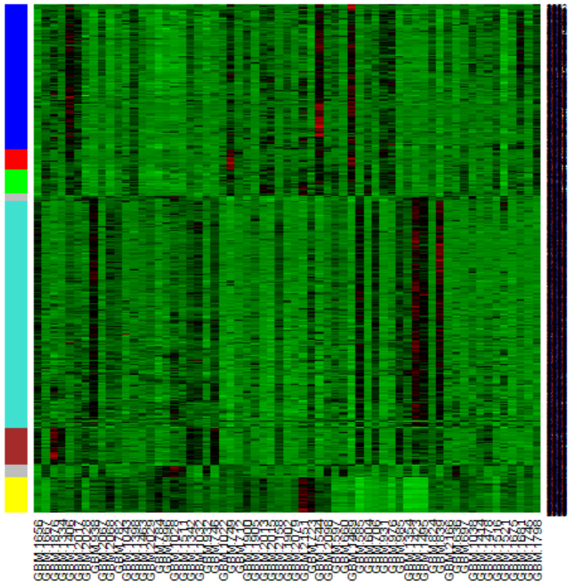
Orthophosphate

these pairs share
343 reactions

191 of which are
distinct reactions

**0.81% of the entire set of metabolites
being shared among
13.7% of the entire set of chemical reactions**

MEWC – Applications – Genes co-expression networks



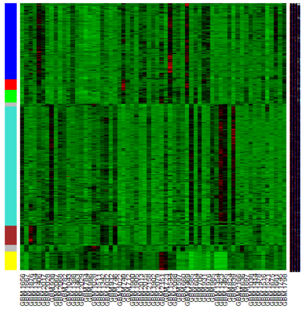
Microarray data from the Human Huntington’s brain disease (HD)

(in, A. Hodges *et al*, “Regional and cellular gene expression changes in human Huntington’s disease brain”, *Hum Mol Genet*, 15(6): 965-977, 2006)

HD causes extensive neurodegeneration in the caudate nucleus, where medium spiny neurons are preferentially destroyed in early stages of the disease. Comparatively, the other analyzed brain regions are relatively spared.

(in, M.C. Oldham, P. Langfelder and S. Horvath, “Network methods for describing sample relationship in genomic datasets: application to Huntington’s disease”, *BMC Systems Biology*, 6(63): 1-26, 2012)

MEWC – Applications – Genes co-expression networks



Microarray data from the Human Huntington's brain disease (HD)

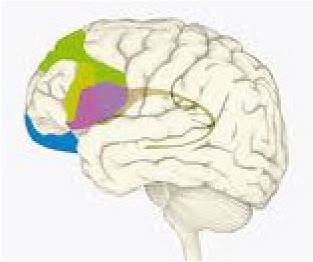
(in, A. Hodges *et al*, "Regional and cellular gene expression changes in human Huntington's disease brain", *Hum Mol Genet*, 15(6): 965-977, 2006)

Data from brain samples of patients with HD (44 indiv.) and unaffected controls (36 indv.)

five grades of HD severity: 0 (least) to 4 (most) using Vonsattel's neuropath logical criteria

Affymetrix U133A microarrays to survey gene expression in:

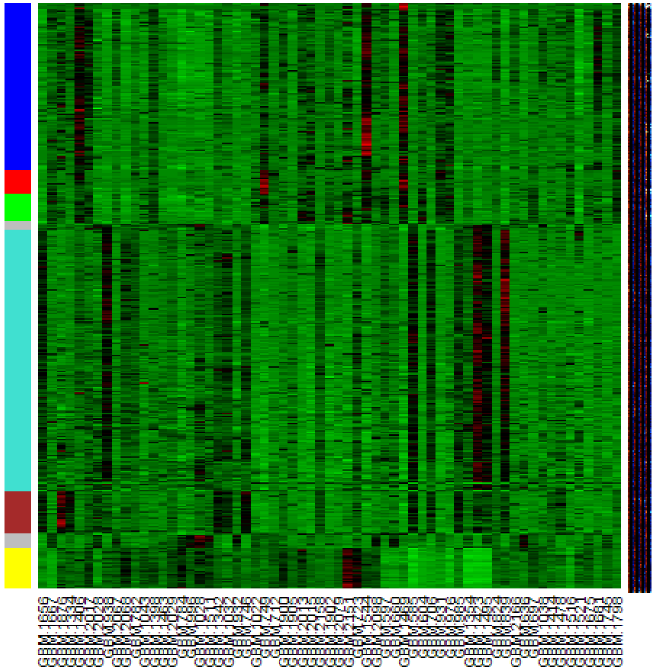
- caudate nucleus (CN)
- cerebellum (CB)
- primary motor cortex (Brodmann's area 4, BA4)
- prefrontal cortex (Brodmann's area 9, BA9)



Samples were processed in the United States (US) and New Zealand (NZ)

(in, M.C. Oldham, P. Langfelder and S. Horvath, "Network methods for describing sample relationship in genomic datasets: application to Huntington's disease", *BMC Systems Biology*, 6(63): 1-26, 2012)

MEWC – Applications – Genes co-expression networks

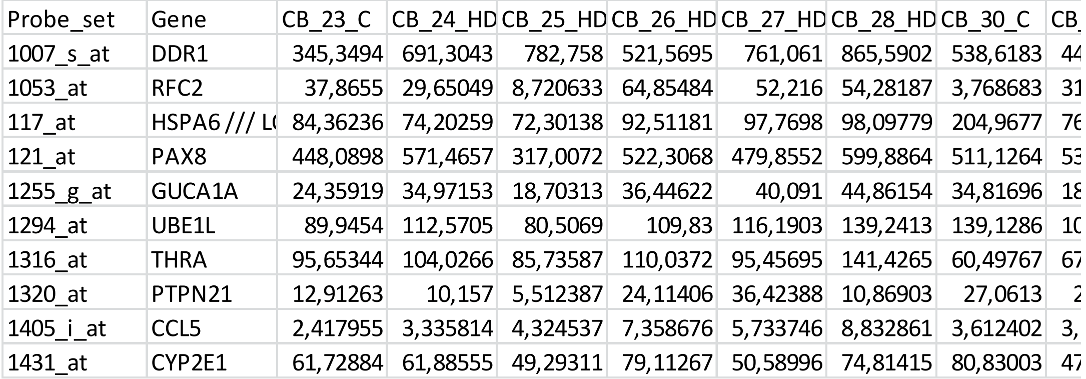


(in, M.C. Oldham, P. Langfelder and S. Horvath, “Network methods for describing sample relationship in genomic datasets: application to Huntington’s disease”, *BMC Systems Biology*, 6(63): 1-26, 2012)

Microarray data

samples

probes/genes



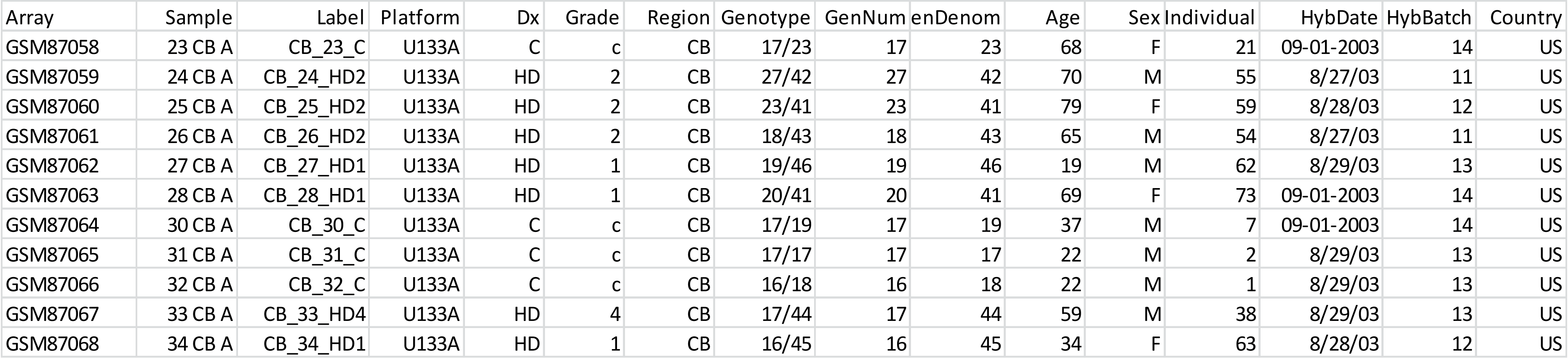
Probe_set	Gene	CB_23_C	CB_24_HD	CB_25_HD	CB_26_HD	CB_27_HD	CB_28_HD	CB_30_C	CB
1007_s_at	DDR1	345,3494	691,3043	782,758	521,5695	761,061	865,5902	538,6183	44
1053_at	RFC2	37,8655	29,65049	8,720633	64,85484	52,216	54,28187	3,768683	31
117_at	HSPA6 /// L(	84,36236	74,20259	72,30138	92,51181	97,7698	98,09779	204,9677	76
121_at	PAX8	448,0898	571,4657	317,0072	522,3068	479,8552	599,8864	511,1264	53
1255_g_at	GUCA1A	24,35919	34,97153	18,70313	36,44622	40,091	44,86154	34,81696	18
1294_at	UBE1L	89,9454	112,5705	80,5069	109,83	116,1903	139,2413	139,1286	10
1316_at	THRA	95,65344	104,0266	85,73587	110,0372	95,45695	141,4265	60,49767	67
1320_at	PTPN21	12,91263	10,157	5,512387	24,11406	36,42388	10,86903	27,0613	2
1405_i_at	CCL5	2,417955	3,335814	4,324537	7,358676	5,733746	8,832861	3,612402	3,
1431_at	CYP2E1	61,72884	61,88555	49,29311	79,11267	50,58996	74,81415	80,83003	47

• • •

• • •

there are 18631 probes/genes and 201 samples

Samples information



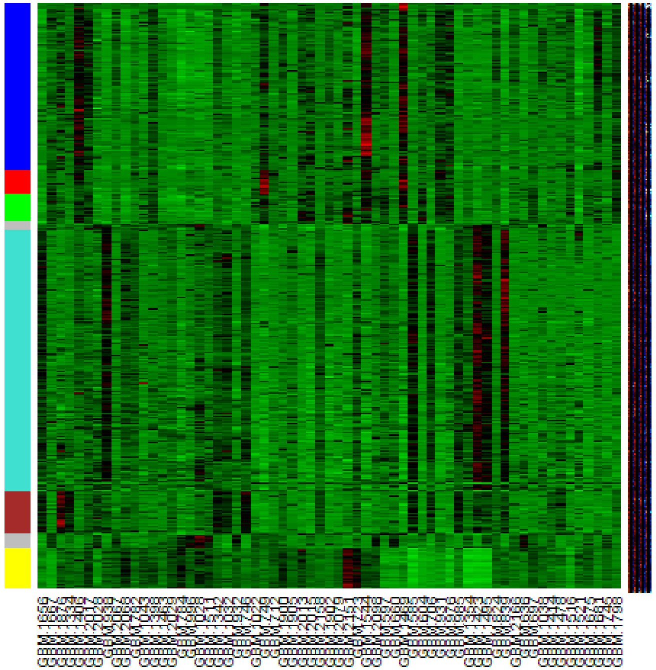
Array	Sample	Label	Platform	Dx	Grade	Region	Genotype	GenNum	enDenom	Age	Sex	Individual	HybDate	HybBatch	Country
GSM87058	23 CB A	CB_23_C	U133A	C	c	CB	17/23	17	23	68	F	21	09-01-2003	14	US
GSM87059	24 CB A	CB_24_HD2	U133A	HD	2	CB	27/42	27	42	70	M	55	8/27/03	11	US
GSM87060	25 CB A	CB_25_HD2	U133A	HD	2	CB	23/41	23	41	79	F	59	8/28/03	12	US
GSM87061	26 CB A	CB_26_HD2	U133A	HD	2	CB	18/43	18	43	65	M	54	8/27/03	11	US
GSM87062	27 CB A	CB_27_HD1	U133A	HD	1	CB	19/46	19	46	19	M	62	8/29/03	13	US
GSM87063	28 CB A	CB_28_HD1	U133A	HD	1	CB	20/41	20	41	69	F	73	09-01-2003	14	US
GSM87064	30 CB A	CB_30_C	U133A	C	c	CB	17/19	17	19	37	M	7	09-01-2003	14	US
GSM87065	31 CB A	CB_31_C	U133A	C	c	CB	17/17	17	17	22	M	2	8/29/03	13	US
GSM87066	32 CB A	CB_32_C	U133A	C	c	CB	16/18	16	18	22	M	1	8/29/03	13	US
GSM87067	33 CB A	CB_33_HD4	U133A	HD	4	CB	17/44	17	44	59	M	38	8/29/03	13	US
GSM87068	34 CB A	CB_34_HD1	U133A	HD	1	CB	16/45	16	45	34	F	63	8/28/03	12	US

• • •

MEWC – Applications – Genes co-expression networks

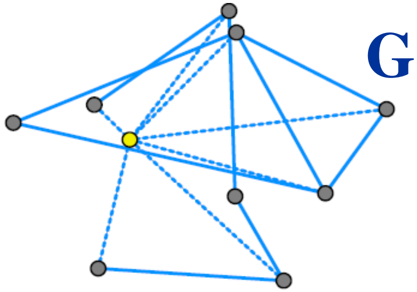
Microarray genes expression data

tissue samples



genes

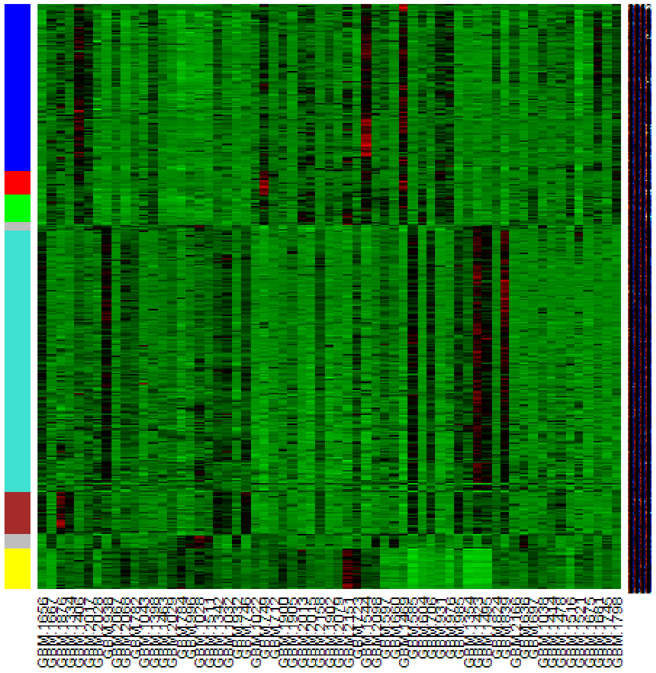
color band indicates module membership



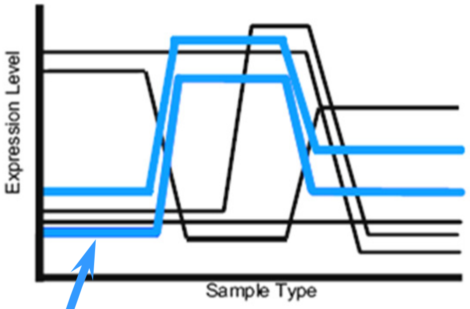
Genes co-expression network

(in, M. Carlson, B. Zhang, Z. Fang, P.S. Mischel, S. Horvath, and S.F. Nelson, "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks", *BMC Genomics*, 7(40): 1-15, 2006)

MEWC – Applications – Genes co-expression networks



Microarray data



strongly correlated genes

Genes co-expression networks

(in, M. Carlson, B. Zhang, Z. Fang, P.S. Mischel, S. Horvath, and S.F. Nelson, “Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks”, *BMC Genomics*, 7(40): 1-15, 2006)

Similarity matrix (correlation)

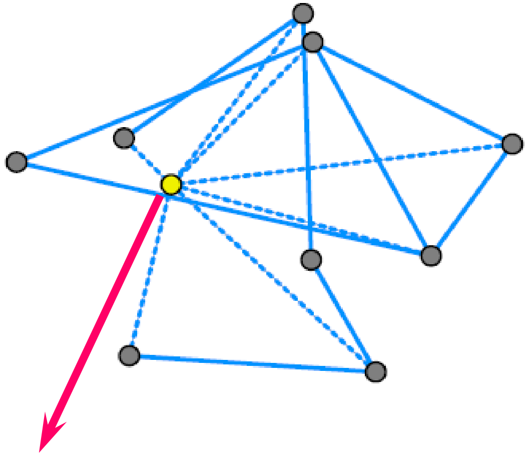
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	1	0.6	0.2	0.8	0.9	0.6	0.9	0.1	0.5	0.3
G2	0.6	1	0.9	0.1	0.2	0.6	1.0	0.1	0.3	0.4
G3	0.2	0.9	1	0.2	0.3	0.4	0.8	0.2	0.3	0.9
G4	0.8	0.1	0.2	1	0.9	0.9	0.8	0.3	0.6	0.0
G5	0.9	0.2	0.3	0.9	1	0.9	0.9	0.6	0.1	0.5
G6	0.6	0.6	0.4	0.9	0.9	1	0.6	0.2	0.7	0.1
G7	0.9	1.0	0.8	0.8	0.9	0.6	1	0.8	0.9	0.2
G8	0.1	0.1	0.2	0.3	0.6	0.2	0.8	1	0.9	0.2
G9	0.5	0.3	0.3	0.6	0.1	0.7	0.9	0.9	1	0.9
G10	0.3	0.4	0.9	0.0	0.5	0.1	0.2	0.2	0.9	1

Adjacency matrix

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	NA	O	O	E	E	O	E	O	O	O
G2	O	NA	E	O	O	O	E	O	O	O
G3	O	E	NA	O	O	O	E	O	O	E
G4	E	O	O	NA	E	E	E	O	O	O
G5	E	O	O	E	NA	E	E	O	O	O
G6	O	O	O	E	E	NA	O	O	O	O
G7	E	E	E	E	E	O	NA	E	E	O
G8	O	O	O	O	O	O	E	NA	E	O
G9	O	O	O	O	O	O	E	E	NA	E
G10	O	O	E	O	O	O	O	O	E	NA

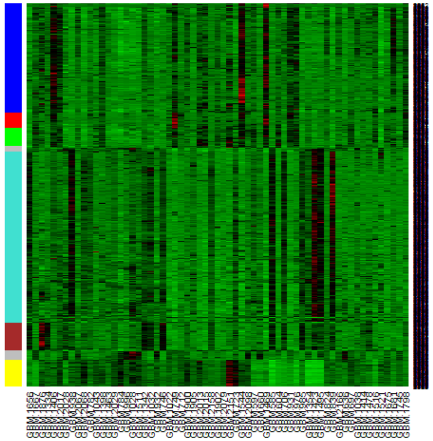
However, some authors suggest that we should **focus on modules instead of individual genes**

Genes co-expression network



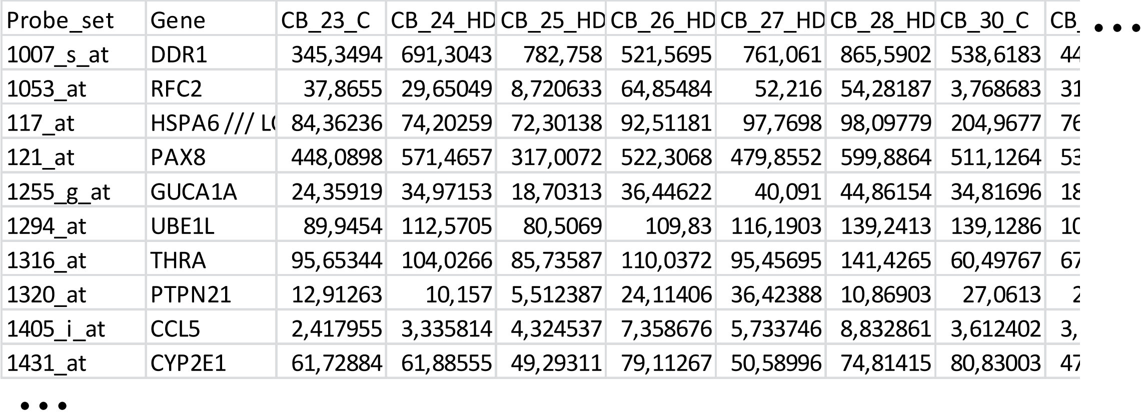
strong chance for being an essential gene

MEWC – App – Genes co-expression network / Samples network



samples

probes/genes



Probe_set	Gene	CB_23_C	CB_24_HD	CB_25_HD	CB_26_HD	CB_27_HD	CB_28_HD	CB_30_C	CB_
1007_s_at	DDR1	345,3494	691,3043	782,758	521,5695	761,061	865,5902	538,6183	44
1053_at	RFC2	37,8655	29,65049	8,720633	64,85484	52,216	54,28187	3,768683	31
117_at	HSPA6 /// L(	84,36236	74,20259	72,30138	92,51181	97,7698	98,09779	204,9677	76
121_at	PAX8	448,0898	571,4657	317,0072	522,3068	479,8552	599,8864	511,1264	53
1255_g_at	GUCA1A	24,35919	34,97153	18,70313	36,44622	40,091	44,86154	34,81696	18
1294_at	UBE1L	89,9454	112,5705	80,5069	109,83	116,1903	139,2413	139,1286	10
1316_at	THRA	95,65344	104,0266	85,73587	110,0372	95,45695	141,4265	60,49767	67
1320_at	PTPN21	12,91263	10,157	5,512387	24,11406	36,42388	10,86903	27,0613	2
1405_i_at	CCL5	2,417955	3,335814	4,324537	7,358676	5,733746	8,832861	3,612402	3,
1431_at	CYP2E1	61,72884	61,88555	49,29311	79,11267	50,58996	74,81415	80,83003	47

• • •

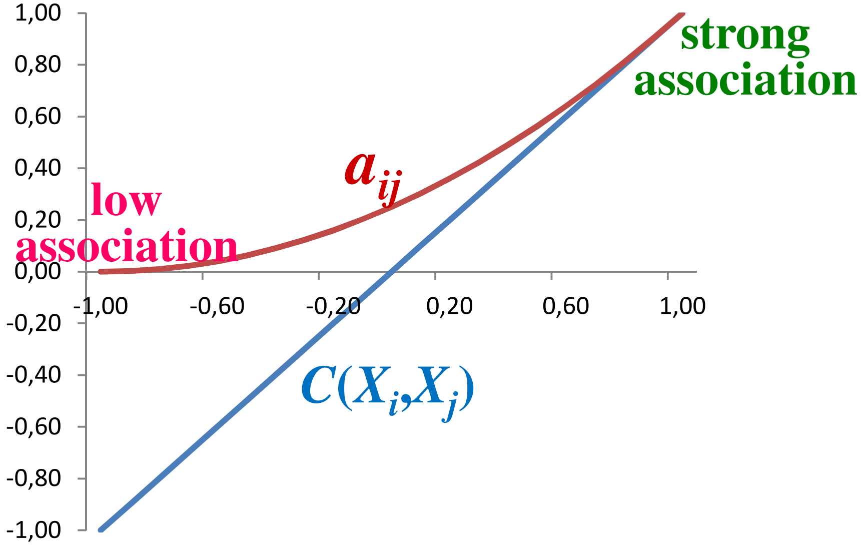
• • •

Microarray data – correlation matrix among a given subset of probes/genes

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

for all pairs of (genes or samples) (i, j) in the selected subset

with $C(X_i, X_j)$ the Pearson correlation among (genes or samples) i and j



(in, M.C. Oldham, P. Langfelder and S. Horvath, "Network methods for describing sample relationship in genomic datasets: application to Huntington's disease", *BMC Systems Biology*, 6(63): 1-26, 2012)

Cliques – Applications – Genes co-expression network

correlation matrix (A) among probes/genes $$a_{ij}=\left(\frac{C(X_i,X_j)+1}{2}\right)^2\in[0,1]$$

Instance with the first 1000 probes/genes threshold $\tau = 0.85 \rightarrow C(X_i,X_j)\approx 0.844$

instance information

nodes (excluding singletons)	edges	density	min$\{a_{ij}\}$	avg$\{a_{ij}\}$	max$\{a_{ij}\}$
664	8300	3.77%	0.85	0.87	0.99

max weight clique solution

total weight = 314.91

clique size = 27

exec time: 1843.82 sec

probes/genes			
200027_a NARS	200749_a RAN	200987_x PSME3	201192_s PITPNA
200030_s SLC25A3	200750_s RAN	201000_a AARS	201198_s PSMD1
200078_s ATP6V0B	200802_a SARS	201001_s UBE2V1	201241_a DDX1
200093_s HINT1	200818_a ATP5O	201002_s UBE2V1	201245_s OTUB1
200614_a CLTC	200870_a STRAP	201022_s DSTN	201472_a VBP1
200638_s YWHAZ	200883_a UQCRC2	201077_s NHP2L1	201523_x UBE2N
200738_s PGK1	200950_a ARPC1A	201191_a PITPNA	

Cliques – Applications – Genes co-expression network

correlation matrix (A) among probes/genes

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

Instance with the first 1000 probes/genes

threshold $\tau = 0.9$ → $C(X_i,X_j) \approx 0.897$

instance information

nodes (excluding singletons)	edges	density	$\min\{a_{ij}\}$	$\operatorname{avg}\{a_{ij}\}$	$\max\{a_{ij}\}$
321	982	1.91%	0.9	0.91	0.99

max weight clique solution

probes/genes		
200030_s SLC25A3	200749_a RAN	201077_s NHP2L1
200078_s ATP6V0B	200870_a STRAP	201198_s PSMD1
200093_s HINT1	201002_s UBE2V1	201472_a VBP1
200614_a CLTC	201022_s DSTN	

total weight = 50.64

clique size = 11

exec time:
0.16 sec

Cliques – Applications – Genes co-expression network

correlation matrix (A) among probes/genes

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

Instance with the first 3000 probes/genes

threshold τ = 0.9 → $C(X_i,X_j) \approx 0.897$

instance information	nodes (excluding singletons)	edges	density	min$\{a_{ij}\}$	avg$\{a_{ij}\}$	max$\{a_{ij}\}$
	788	3782	1.22%	0.9	0.91	0.99

max weight clique solution

probes/genes			
200041_s BAT1	203486_s ARMC8	202181_a KIAA0247	201244_s RAF1
203288_a KIAA0355	203616_a POLB	202220_a KIAA0907	202761_s SYNE2
201697_s DNMT1	202392_s PISD	203073_a COG2	201906_s CTDSPL
202743_a PIK3R3	203487_s ARMC8	200965_s ABLIM1	202328_s PKD1
202360_a MAML1	203345_s MTF2	203298_s JARID2	

total weight = 158.82

clique size = 19

exec time:
1.54 sec

Cliques – Applications – Genes co-expression network

correlation matrix (A) among probes/genes

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2}\right)^2 \in [0,1]$$

Instance with the first 7000 probes/genes

threshold $\tau = 0.9$ → $C(X_i,X_j) \approx 0.897$

instance information	nodes (excluding singletons)	edges	density	min{a_{ij}}	avg{a_{ij}}	max{a_{ij}}
	1378	11449	1.21%	0.9	0.92	0.99

max weight clique solution

total weight = 842.72

clique size = 43

exec time: 3250.80 sec

probes/genes

200920_s BTG1
200965_s ABLIM1
202181_a KIAA0247
202328_s PKD1
202392_s PISD
202761_s SYNE2
202743_a PIK3R3
203298_s JARID2
203486_s ARMC8
203487_s ARMC8

203616_a POLB
203895_a PLCB4
203910_a ARHGAP29
204069_a MEIS1
204431_a TLE2
204520_x BRD1
204791_a NR2C1
205022_s CHES1
205029_s FABP7
205030_a FABP7

205070_a ING3
205390_s ANK1
205391_x ANK1
205528_s RUNX1T1
205529_s RUNX1T1
205646_s PAX6
205728_a ---
205730_s ABLIM3
205795_a NRXN3
205923_a RELN

206163_a MAB21L1
206230_a LHX1
206282_a NEUROD1
206328_a CDH15
206373_a ZIC1
206914_a CRTAM
207060_a EN2
207087_x ANK1
207103_a KCND2
207182_a GABRA6

207197_a ZIC3
207637_a PRKAR2B
208072_s DGKD

Cliques – Applications – Genes co-expression network

correlation matrix (A) among probes/genes

$$a_{ij} = \left(\frac{C(X_i, X_j)+1}{2}\right)^2 \in [0,1]$$

Instance with the first 7000 probes/genes

threshold $\tau = 0.9$ → $C(X_i,X_j) \approx 0.897$

instance information

nodes (excluding singletons)	edges	density	min$\{a_{ij}\}$	avg$\{a_{ij}\}$	max$\{a_{ij}\}$
1378	11449	1.21%	0.9	0.92	0.99

max clique solution

clique size = 43

exec time:
40.95 sec
(DF)

neighborhood:
2381 edges
167 nodes

probes/genes

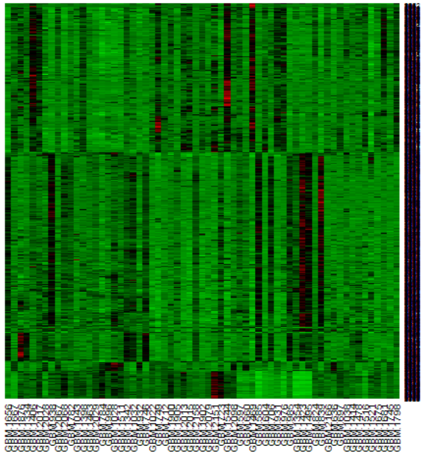
200920_s BTG1
200965_s ABLIM1
202181_a KIAA0247
202328_s PKD1
202392_s PISD
202743_a PIK3R3
202761_s SYNE2
203298_s JARID2
203486_s ARMC8
203487_s ARMC8
203616_a POLB
203895_a PLCB4
203910_a ARHGAP29
204069_a MEIS1
204431_a TLE2
204520_x BRD1
204791_a NR2C1
205022_s CHES1
205029_s FABP7
205030_a FABP7
205070_a ING3
205390_s ANK1
205391_x ANK1
205528_s RUNX1T1
205529_s RUNX1T1
205646_s PAX6
205728_a ---
205730_s ABLIM3
205795_a NRXN3
205923_a RELN
206163_a MAB21L1
206230_a LHX1
206282_a NEUROD1
206328_a CDH15
206373_a ZIC1
206914_a CRTAM
207060_a EN2
207087_x ANK1
207103_a KCND2
207182_a GABRA6
207197_a ZIC3
207637_a PRKAR2B
208072_s DGKD

Technique: *k*-means (non-hierarchical clustering)

Metric: Euclidean distance
Technique: *k*-means

genes co-expression network

$k = 8$



elements: the 201 samples
attributes: the 18631 probes/genes

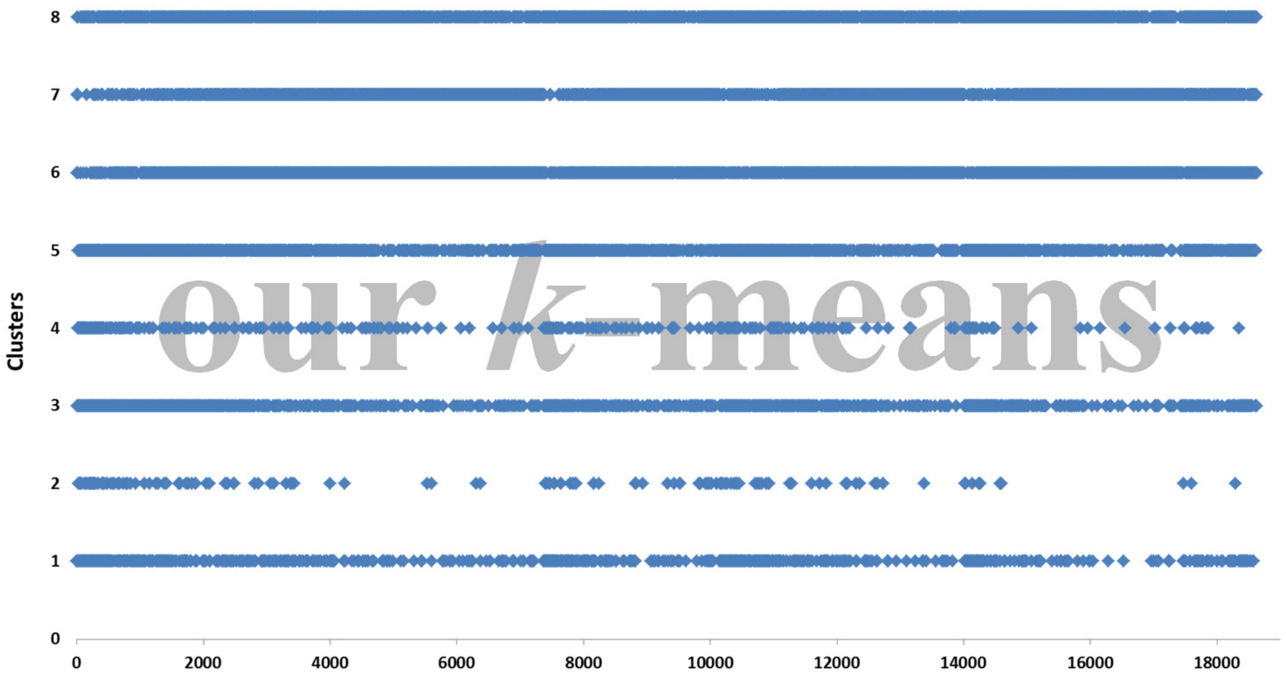
Our *k*-means Best sol. value: 30,401,540.00

R-square test (PROC CLUSTER of SAS): *R-square* = 0.808

SPSS 24.0 solution Best sol. value: 44,333,323.88 *R-square* = 0.808

number of genes in each cluster

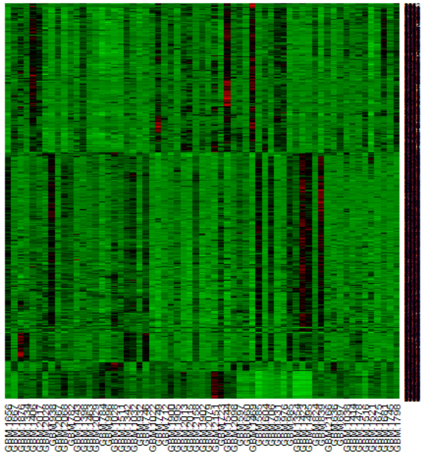
cluster	our alg	SPSS
1	811	2272
2	158	62
3	1598	34
4	371	88
5	2468	441
6	4541	3
7	5294	75
8	3390	15565



Technique: *k*-means (non-hierarchical clustering)

Metric: Euclidean distance
Technique: *k*-means

genes co-expression network



***k* = 8**

elements: the 201 samples
attributes: the 18631 probes/genes

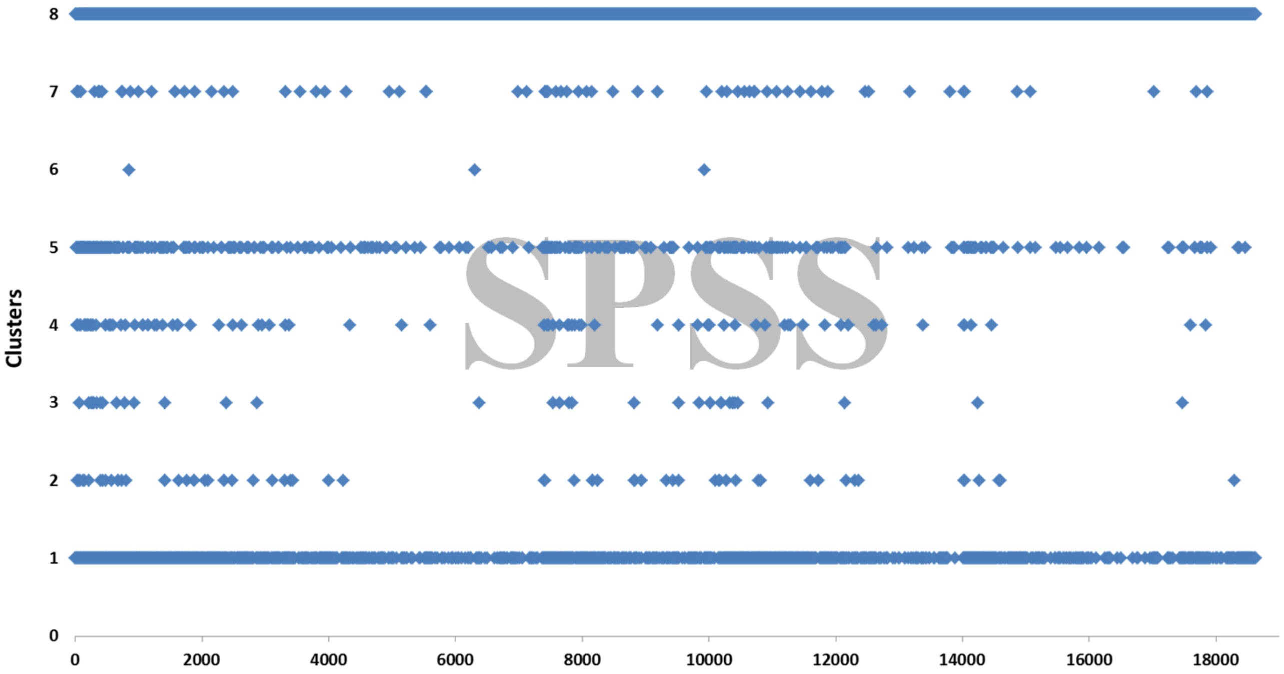
Our *k*-means Best sol. value: 30,401,540.00

R-square test (PROC CLUSTER of SAS): *R-square* = 0.808

SPSS 24.0 solution Best sol. value: 44,333,323.88 *R-square* = 0.808

number of genes in each cluster

cluster	our alg	SPSS
1	811	2272
2	158	62
3	1598	34
4	371	88
5	2468	441
6	4541	3
7	5294	75
8	3390	15565



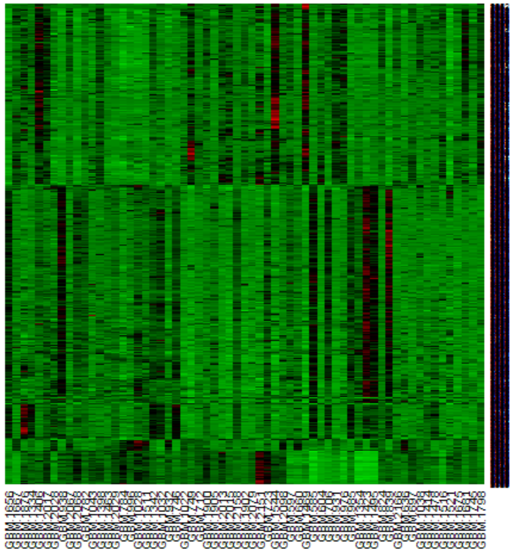
MEWC – Applications – Samples network

correlation matrix (A) among samples

$$a_{ij} = \left(\frac{C(X_i, X_j)+1}{2}\right)^2 \in [0,1]$$

this time X_i and X_j are samples

instances with all 18631 probes/genes and all 201 samples

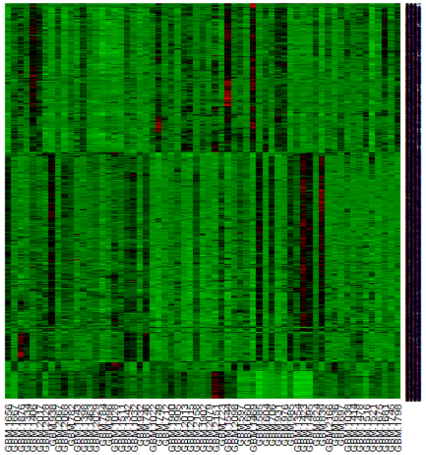


threshold (τ)	nodes (excluding singletons)	edges	density	min{a_{ij}}	avg{a_{ij}}	max{a_{ij}}	optimum	clique size
0.75	201	19053	94.79%	0.75	0.87	0.99	12864.75	172
0.85	201	10521	52.34%	0.85	0.92	0.99	3067.21	82
0.9	201	6180	30.75%	0.90	0.95	0.99	1805.51	62
0.95	196	3205	16.77%	0.95	0.96	0.99	678.90	38
0.98	80	144	4.56%	0.98	0.98	0.99	9.80	5

MEWC – Applications – Samples network

correlation matrix (A)
among samples

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$



Max weight cliques:

$\tau = 0.98$
avg weight = 0.98

BA4_H115_C
BA4_H117_C
BA4_HC80_HD2
BA4_HC86_HD1
BA9_118_C

2.98%

$\tau = 0.95$
avg weight = 0.966

CB_23_C
CB_24_HD2
CB_27_HD1
CB_28_HD1
CB_30_C
CB_31_C
CB_32_C
CB_33_HD4
CB_35_HD1
CB_38_HD4
CB_39_C
CB_40_C
CB_41_C

CB_66_C
CB_69_HD0
CB_71_HD3
CB_74_HD3
CB_76_HD4
CB_81_C
CB_82_C
CB_H104_C
CB_H111_C
CB_H115_C
CB_H117_C
CB_H118_C
CB_H120_C

CB_H121_C
CB_H126_C
CB_H129_C
CB_H132_C
CB_H137_C
CB_HC55_HD1
CB_HC57_HD2
CB_HC65_HD2
CB_HC68_HD1
CB_HC69_HD2
CB_HC81_HD1
CB_HC86_HD1

18.91%

$\tau = 0.90$
avg weight = 0.955

CB_26_HD2
CB_34_HD1
CB_67_C
CB_69_HD0
CB_70_HD3
CB_72_HD2
CB_75_HD2
CB_79_HD2
CB_80_C
CB_H110_C
CB_H123_C
CB_H124_C
CB_H131_C

CB_HC102_HD3
CB_HC103_HD1
CB_HC105_HD1
CB_HC51_HD1
CB_HC53_HD1
CB_HC61_HD2
CB_HC62_HD2
CB_HC71_HD0
CB_HC72_HD2
CB_HC74_HD1
CB_HC80_HD2
CB_HC82_HD2

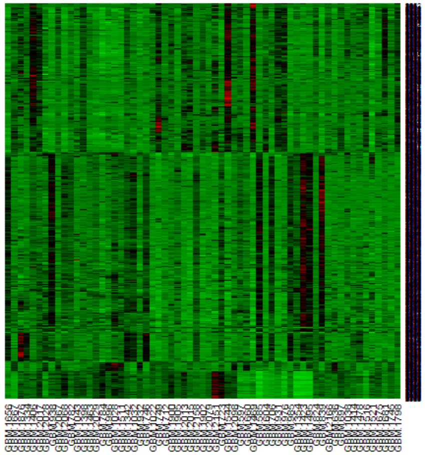
31.34%

mixed samples in each layer ! ! !

MEWC – Applications – Samples network

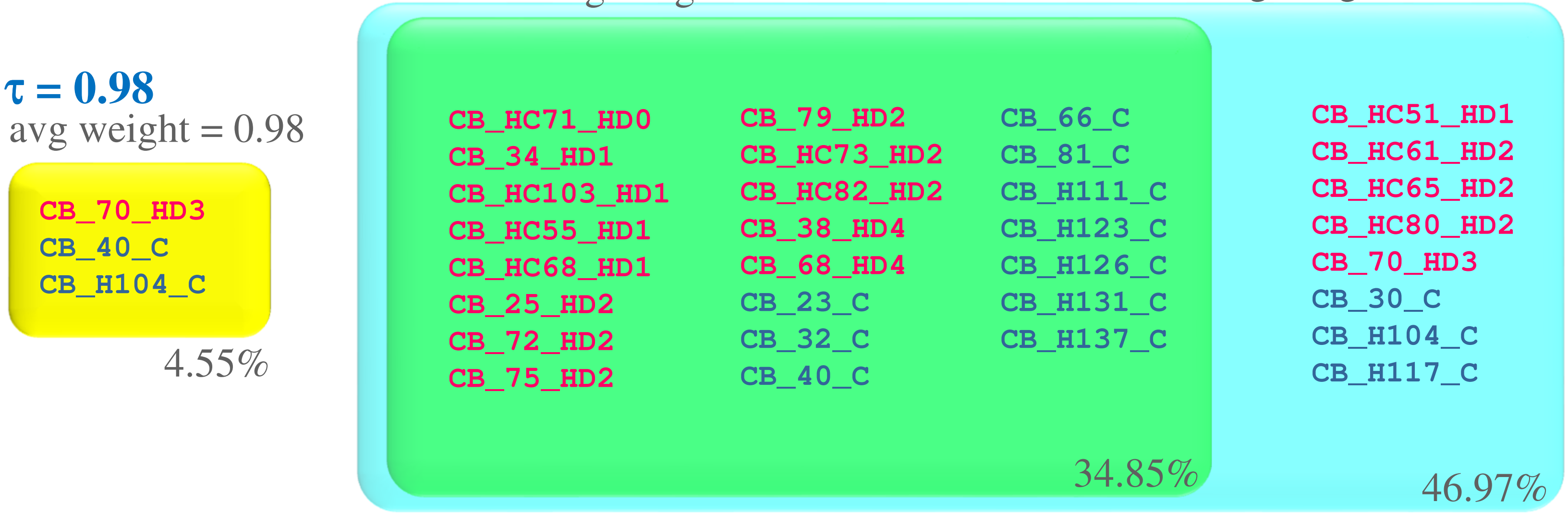
correlation matrix (A)
among **samples**

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$



just among the **cerebellum (CB)** samples (66 samples)

Max weight cliques:



τ = 0.98
avg weight = 0.98

CB_70_HD3
CB_40_C
CB_H104_C

4.55%

τ = 0.95
avg weight = 0.965

CB_HC71_HD0
CB_34_HD1
CB_HC103_HD1
CB_HC55_HD1
CB_HC68_HD1
CB_25_HD2
CB_72_HD2
CB_75_HD2

CB_79_HD2
CB_HC73_HD2
CB_HC82_HD2
CB_38_HD4
CB_68_HD4
CB_23_C
CB_32_C
CB_40_C

CB_66_C
CB_81_C
CB_H111_C
CB_H123_C
CB_H126_C
CB_H131_C
CB_H137_C

34.85%

τ = 0.90
avg weight = 0.958

CB_HC51_HD1
CB_HC61_HD2
CB_HC65_HD2
CB_HC80_HD2
CB_70_HD3
CB_30_C
CB_H104_C
CB_H117_C

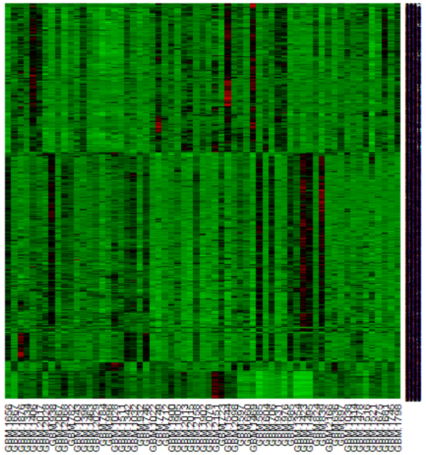
46.97%

mixed samples in each layer *! ! !*

MEWC – Applications – Samples network

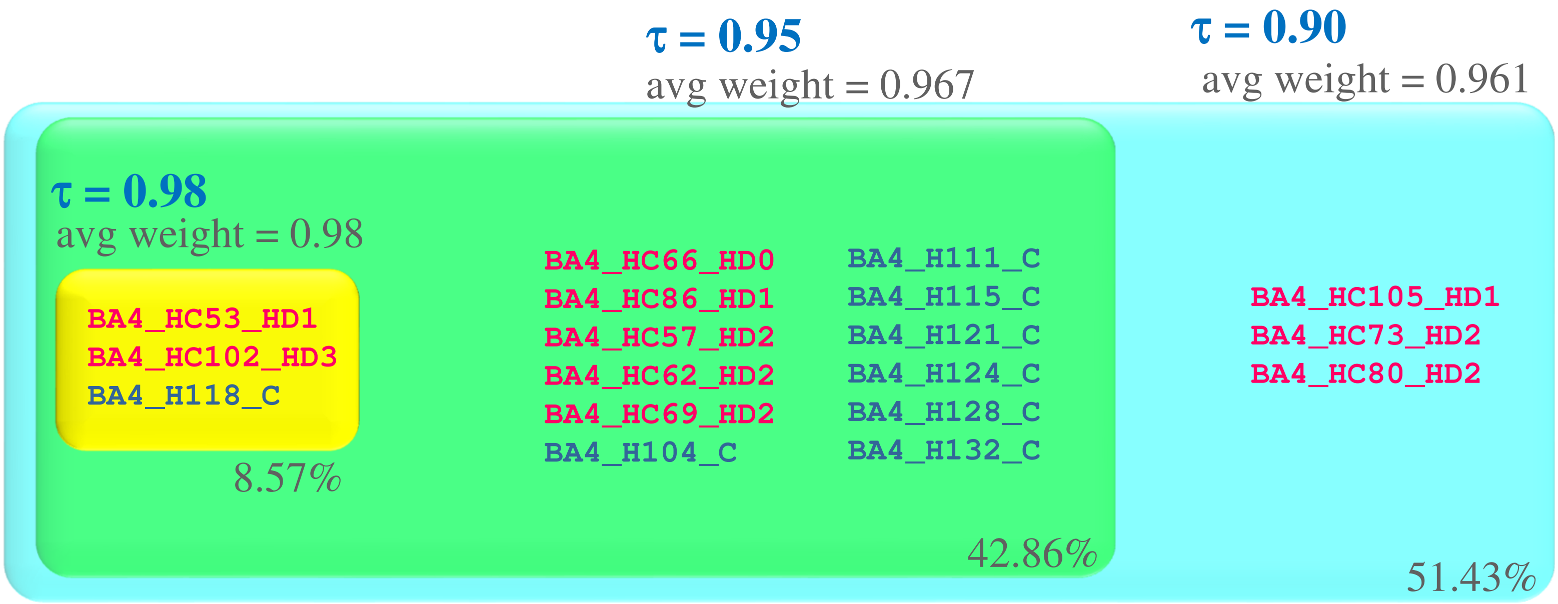
correlation matrix (A)
among samples

$$a_{ij} = \left(\frac{C(X_i, X_j)+1}{2}\right)^2 \in [0,1]$$



just among the **primary motor cortex**
(Brodmann's area 4 – BA4) samples (35 samples)

Max weight cliques:



τ = 0.95
avg weight = 0.967

τ = 0.90
avg weight = 0.961

τ = 0.98
avg weight = 0.98

BA4_HC53_HD1
BA4_HC102_HD3
BA4_H118_C

8.57%

BA4_HC66_HD0
BA4_HC86_HD1
BA4_HC57_HD2
BA4_HC62_HD2
BA4_HC69_HD2
BA4_H104_C

BA4_H111_C
BA4_H115_C
BA4_H121_C
BA4_H124_C
BA4_H128_C
BA4_H132_C

42.86%

BA4_HC105_HD1
BA4_HC73_HD2
BA4_HC80_HD2

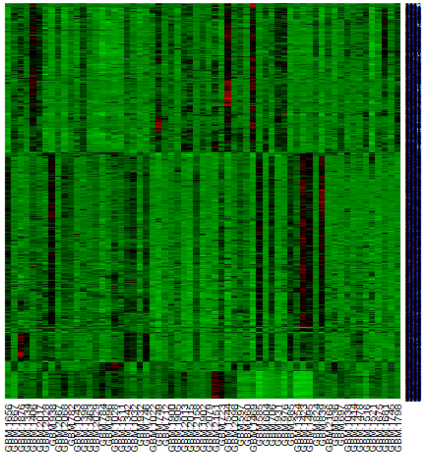
51.43%

mixed samples in each layer ***! ! !***

MEWC – Applications – Samples network

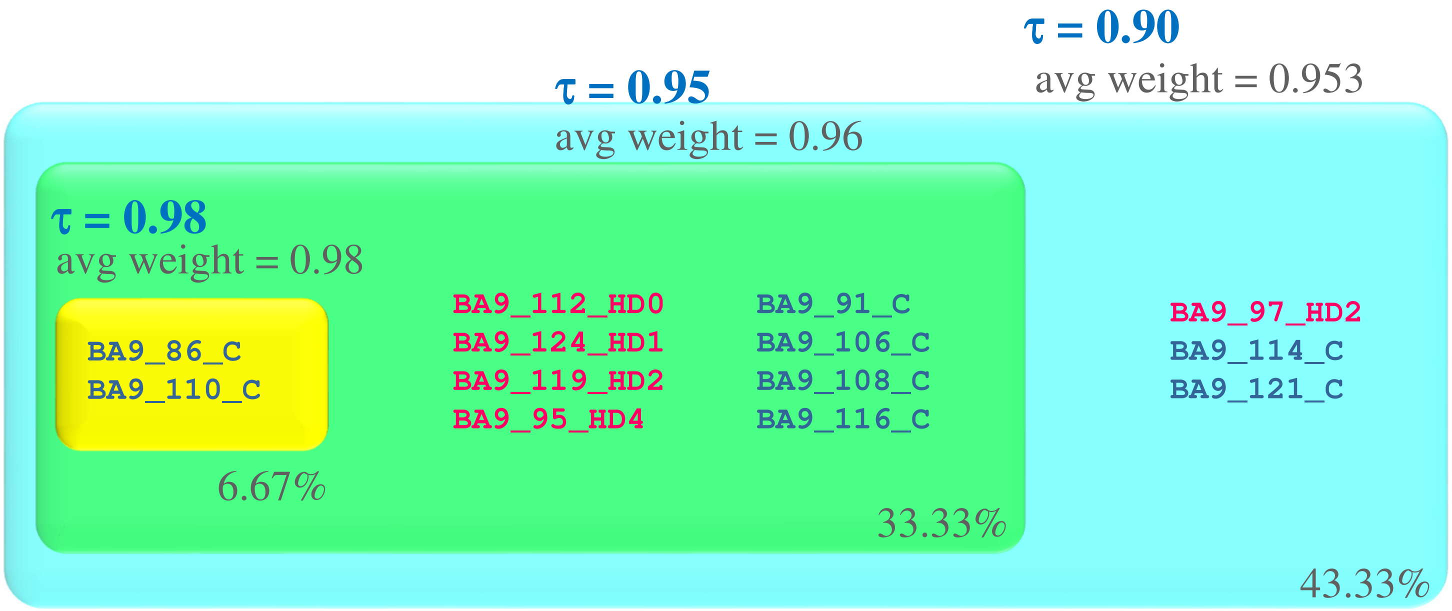
correlation matrix (A)
among **samples**

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$



just among the **prefrontal cortex**
(Brodmann's area 9 – BA9) samples (30 samples)

Max weight cliques:



$\tau = 0.90$
avg weight = 0.953

$\tau = 0.95$
avg weight = 0.96

$\tau = 0.98$
avg weight = 0.98

BA9_86_C
BA9_110_C

6.67%

BA9_112_HD0
BA9_124_HD1
BA9_119_HD2
BA9_95_HD4

BA9_91_C
BA9_106_C
BA9_108_C
BA9_116_C

33.33%

BA9_97_HD2
BA9_114_C
BA9_121_C

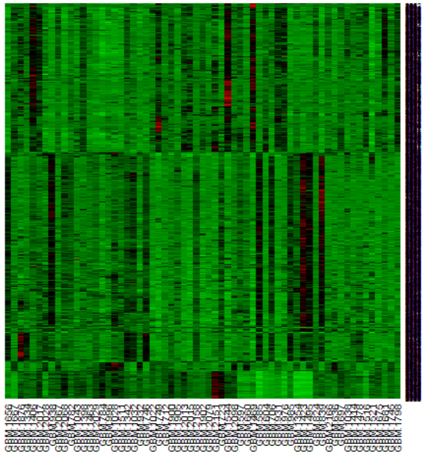
43.33%

mixed samples in each layer *! ! !*

MEWC – Applications – Samples network

correlation matrix (*A*)
among samples

$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

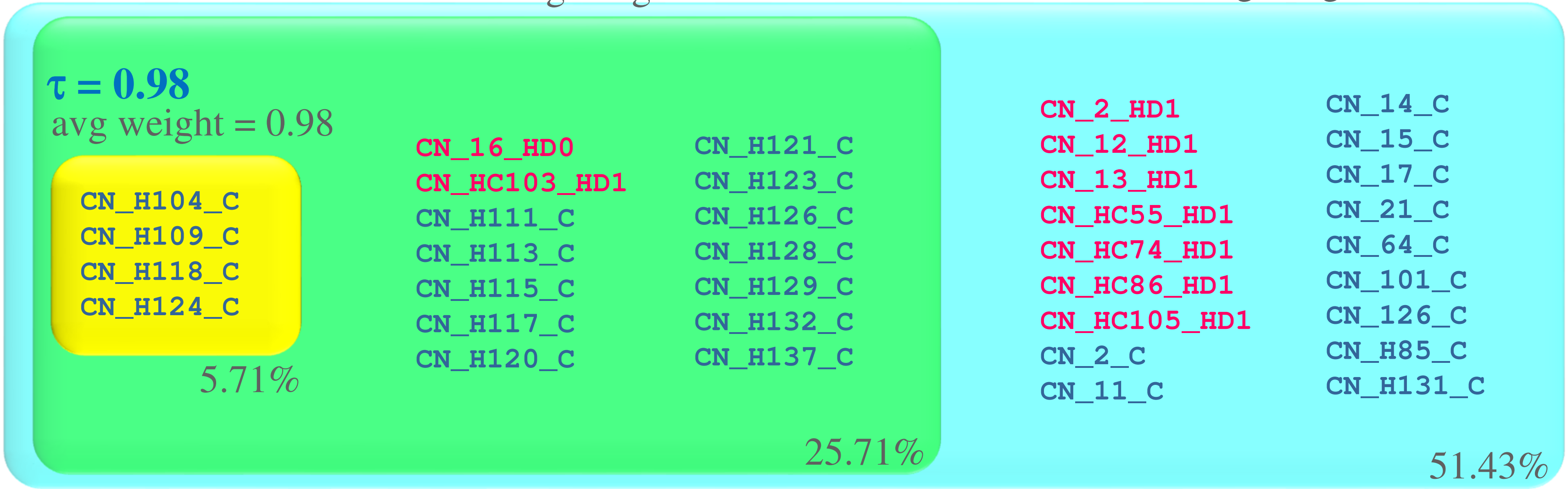


just among the **caudate nucleus** (**CN**) samples (70 samples)

Max weight cliques:

τ = 0.95
avg weight = 0.964

τ = 0.90
avg weight = 0.938



τ = 0.98
avg weight = 0.98

CN_H104_C
CN_H109_C
CN_H118_C
CN_H124_C

5.71%

CN_16_HD0
CN_HC103_HD1
CN_H111_C
CN_H113_C
CN_H115_C
CN_H117_C
CN_H120_C

CN_H121_C
CN_H123_C
CN_H126_C
CN_H128_C
CN_H129_C
CN_H132_C
CN_H137_C

25.71%

CN_2_HD1
CN_12_HD1
CN_13_HD1
CN_HC55_HD1
CN_HC74_HD1
CN_HC86_HD1
CN_HC105_HD1
CN_2_C
CN_11_C

CN_14_C
CN_15_C
CN_17_C
CN_21_C
CN_64_C
CN_101_C
CN_126_C
CN_H85_C
CN_H131_C

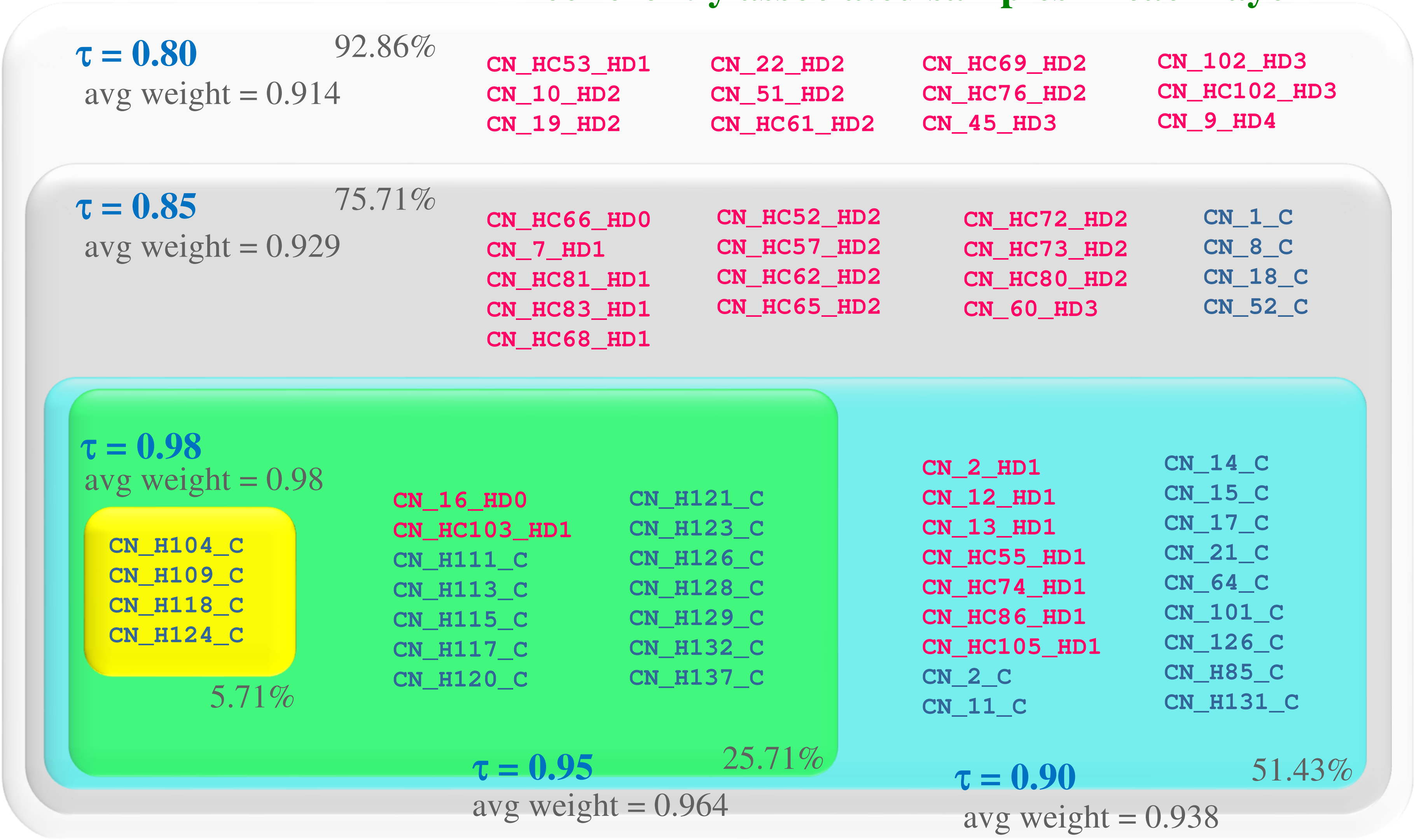
51.43%

coherently associated samples in each layer ✓

MEWC – Applications – Samples network

correlation matrix (*A*) among samples **caudate nucleus (CN)** (70 samples)

Max weight cliques: **coherently associated samples in each layer** ✓



$\tau = 0.80$ 92.86%
avg weight = 0.914

CN_HC53_HD1
CN_10_HD2
CN_19_HD2

CN_22_HD2
CN_51_HD2
CN_HC61_HD2

CN_HC69_HD2
CN_HC76_HD2
CN_45_HD3

CN_102_HD3
CN_HC102_HD3
CN_9_HD4

$\tau = 0.85$ 75.71%
avg weight = 0.929

CN_HC66_HD0
CN_7_HD1
CN_HC81_HD1
CN_HC83_HD1
CN_HC68_HD1

CN_HC52_HD2
CN_HC57_HD2
CN_HC62_HD2
CN_HC65_HD2

CN_HC72_HD2
CN_HC73_HD2
CN_HC80_HD2
CN_60_HD3

CN_1_C
CN_8_C
CN_18_C
CN_52_C

$\tau = 0.98$
avg weight = 0.98

CN_H104_C
CN_H109_C
CN_H118_C
CN_H124_C

5.71%

CN_16_HD0
CN_HC103_HD1
CN_H111_C
CN_H113_C
CN_H115_C
CN_H117_C
CN_H120_C

CN_H121_C
CN_H123_C
CN_H126_C
CN_H128_C
CN_H129_C
CN_H132_C
CN_H137_C

$\tau = 0.95$ 25.71%
avg weight = 0.964

CN_2_HD1
CN_12_HD1
CN_13_HD1
CN_HC55_HD1
CN_HC74_HD1
CN_HC86_HD1
CN_HC105_HD1
CN_2_C
CN_11_C

CN_14_C
CN_15_C
CN_17_C
CN_21_C
CN_64_C
CN_101_C
CN_126_C
CN_H85_C
CN_H131_C

$\tau = 0.90$ 51.43%
avg weight = 0.938

Technique: *k*-means (non-hierarchical clustering)

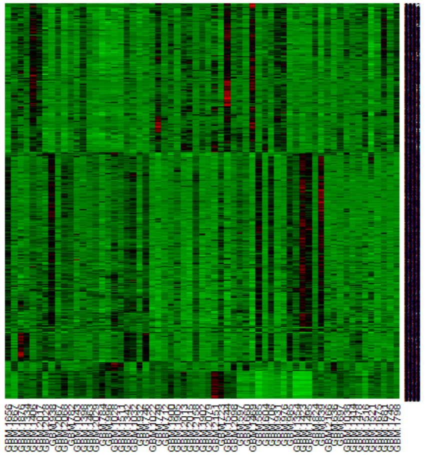
Metric: Euclidean distance
Technique: *k*-means

elements: the 70 CN samples
attributes: the 18631 probes/genes

genes co-expression network

just caudate nucleus (CN)

***k* = 6**



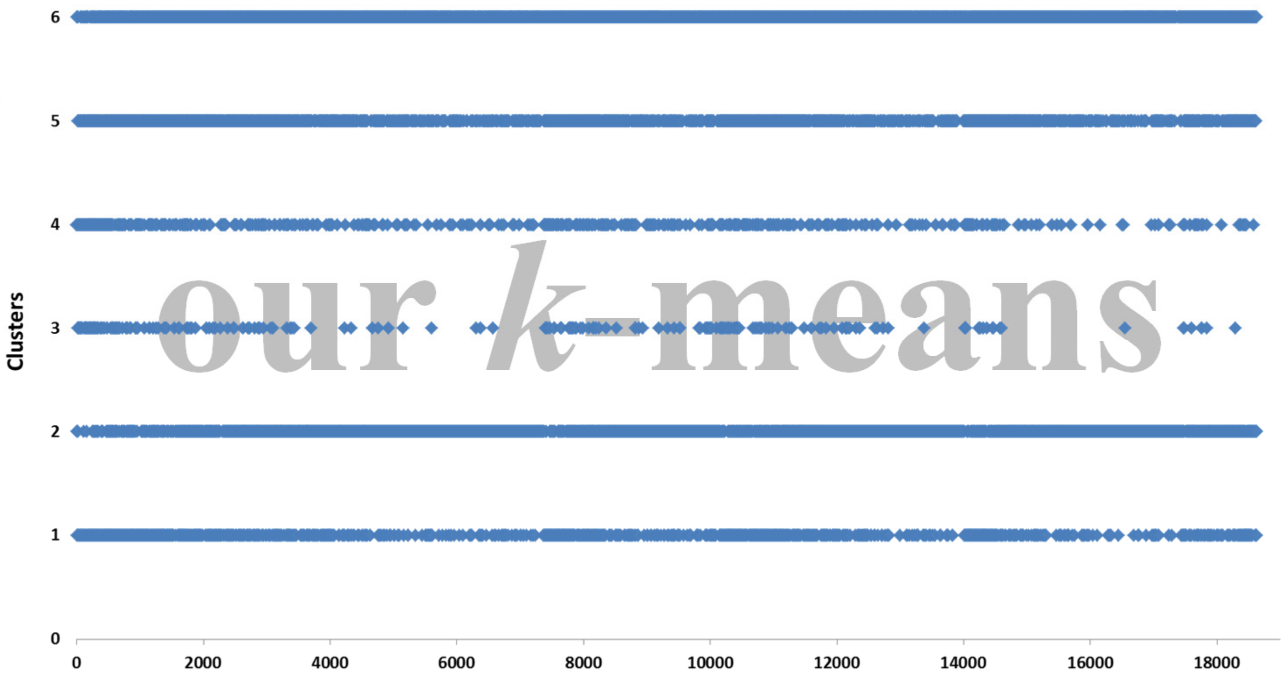
Our *k*-means Best sol. value: 14,844,600.00

R-square test (PROC CLUSTER of SAS): *R-square* = 0.802

SPSS 24.0 solution Best sol. value: 23,036,406.77 *R-square* = 0.816

number of genes in each cluster

cluster	our alg	SPSS
1	1564	9
2	7361	1748
3	245	3
4	706	295
5	3287	16506
6	5468	70



Technique: *k*-means (non-hierarchical clustering)

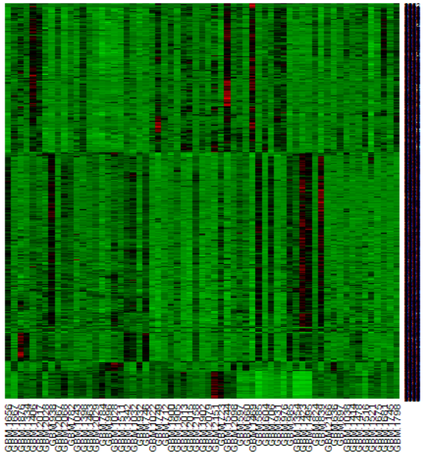
Metric: Euclidean distance
Technique: *k*-means

elements: the 70 CN samples
attributes: the 18631 probes/genes

genes co-expression network

just caudate nucleus (CN)

k = 6



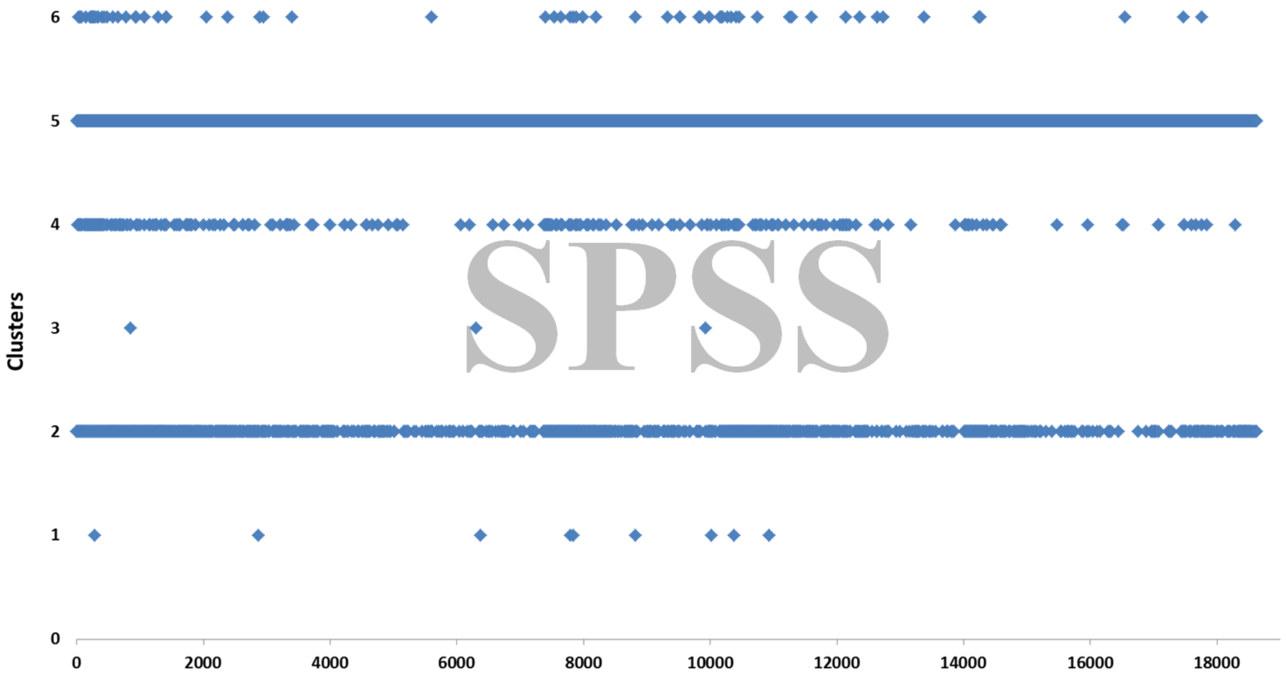
Our *k*-means Best sol. value: 14,844,600.00

R-square test (PROC CLUSTER of SAS): *R-square* = 0.802

SPSS 24.0 solution Best sol. value: 23,036,406.77 *R-square* = 0.816

number of genes in each cluster

cluster	our alg	SPSS
1	1564	9
2	7361	1748
3	245	3
4	706	295
5	3287	16506
6	5468	70



Maximum Weight Cliques Partitioning (MWCP) Problem

Given the weighted graph $G = (V, E, a)$, with a_{ij} the weight of edge $(i,j) \in E$

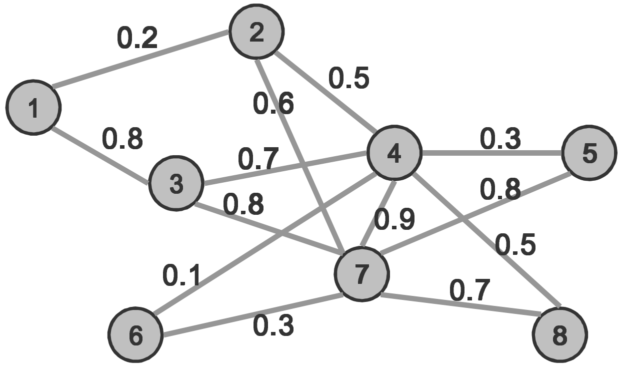
maximum cliques partitioning

Let $C^1 \cup C^2 \cup \ldots \cup C^k$ be a partition of V into cliques and

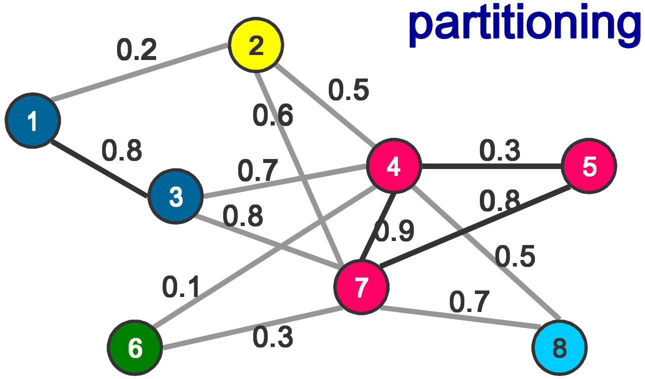
$$A(C^1 \cup C^2 \cup \ldots \cup C^k) = \sum_{i,j \in C^1} a_{ij} + \ldots + \sum_{i,j \in C^k} a_{ij}$$ its total weight.

We want to find a partition $C^1 \cup C^2 \cup \ldots \cup C^k$ with maximum total weight

$G = (V, E)$



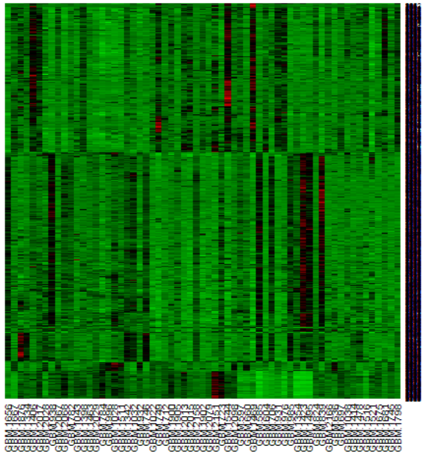
maximum cliques partitioning



MWCP – Applications – Samples network

correlation matrix (A)
among **samples**

$$a_{ij} = \left(\frac{C(X_i, X_j)+1}{2}\right)^2 \in [0,1]$$



τ = 0.90

just among the **caudate nucleus** (**CN**) samples (70 samples)

max weight cliques partitioning: max total weight = 770.37

CN_16_HD0
CN_2_HD1
CN_12_HD1
CN_13_HD1
CN_HC55_HD1
CN_HC74_HD1
CN_HC86_HD1
CN_HC103_HD1
CN_HC105_HD1
CN_2_C
CN_11_C
CN_14_C

CN_15_C
CN_17_C
CN_21_C
CN_64_C
CN_101_C
CN_126_C
CN_H85_C
CN_H104_C
CN_H109_C
CN_H111_C
CN_H113_C
CN_H115_C

CN_H117_C
CN_H118_C
CN_H120_C
CN_H121_C
CN_H123_C
CN_H124_C
CN_H126_C
CN_H128_C
CN_H129_C
CN_H131_C
CN_H132_C
CN_H137_C

weight = 596.71 51.43%

CN_HC66_HD0
CN_HC81_HD1
CN_HC83_HD1
CN_10_HD2
CN_19_HD2
CN_22_HD2
CN_51_HD2
CN_HC52_HD2
CN_HC61_HD2

CN_HC65_HD2
CN_HC69_HD2
CN_45_HD3
CN_60_HD3
CN_62_HD3
CN_HC102_HD3
CN_9_HD4
CN_1_C
CN_18_C

weight = 143.07 25.71%

CN_HC57_HD2
CN_HC62_HD2
CN_HC72_HD2
CN_HC73_HD2
CN_HC76_HD2
CN_HC80_HD2
CN_HC82_HD2
CN_102_HD3

25.99 11.43%

CN_20_C

0.00 1.43%

CN_HC71_HD0
CN_59_HD2

0.90 2.86%

CN_7_HD1
CN_HC68_HD1

0.90 2.86%

CN_HC53_HD1
CN_8_C
CN_H52_C

2.80 4.29%

Maximum p-Median Problem (p medoid)

Given the weighted graph $G = (V, E, a)$, with a_{ij} the weight of edge $(i,j) \in E$
and an integer p

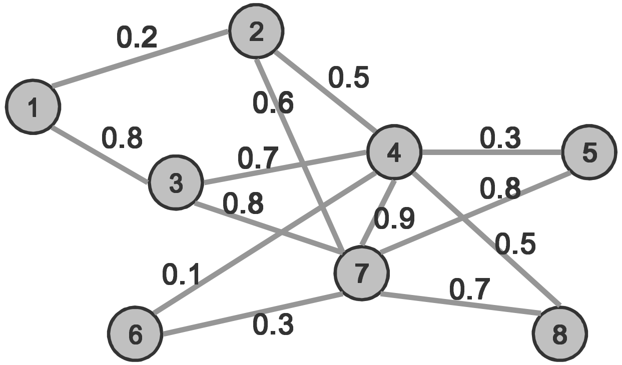
maximum p-median (usually observed as minimum p-median)

Let $V' = \{i^1, \ldots, i^p\} \in V$ and $V^r \in V \backslash V'$ the subset of nodes furthest to i^r in G, for $r = 1, \ldots, p$

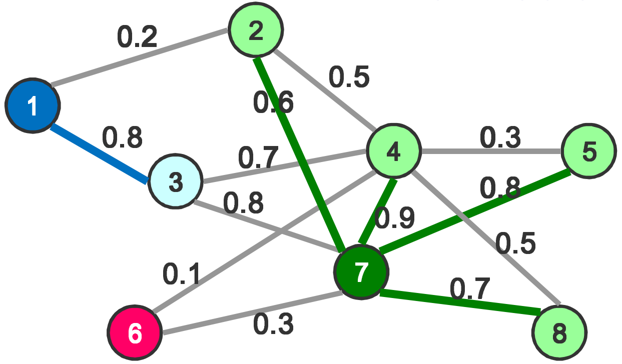
$$A(V') = \sum_{j \in V^1} a_{ji_1} + \ldots + \sum_{j \in V^p} a_{ji_p} \quad \text{its total weight.}$$

We want to find the nodes in V' such that $A(V')$ is the maximum

$G = (V, E)$



maximum
3-median

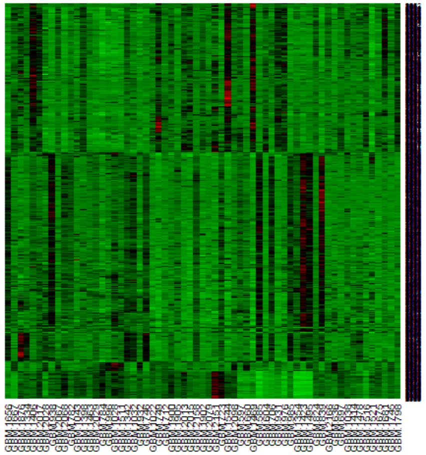


p-median – Applications – Samples network

correlation matrix (*A*)
among samples

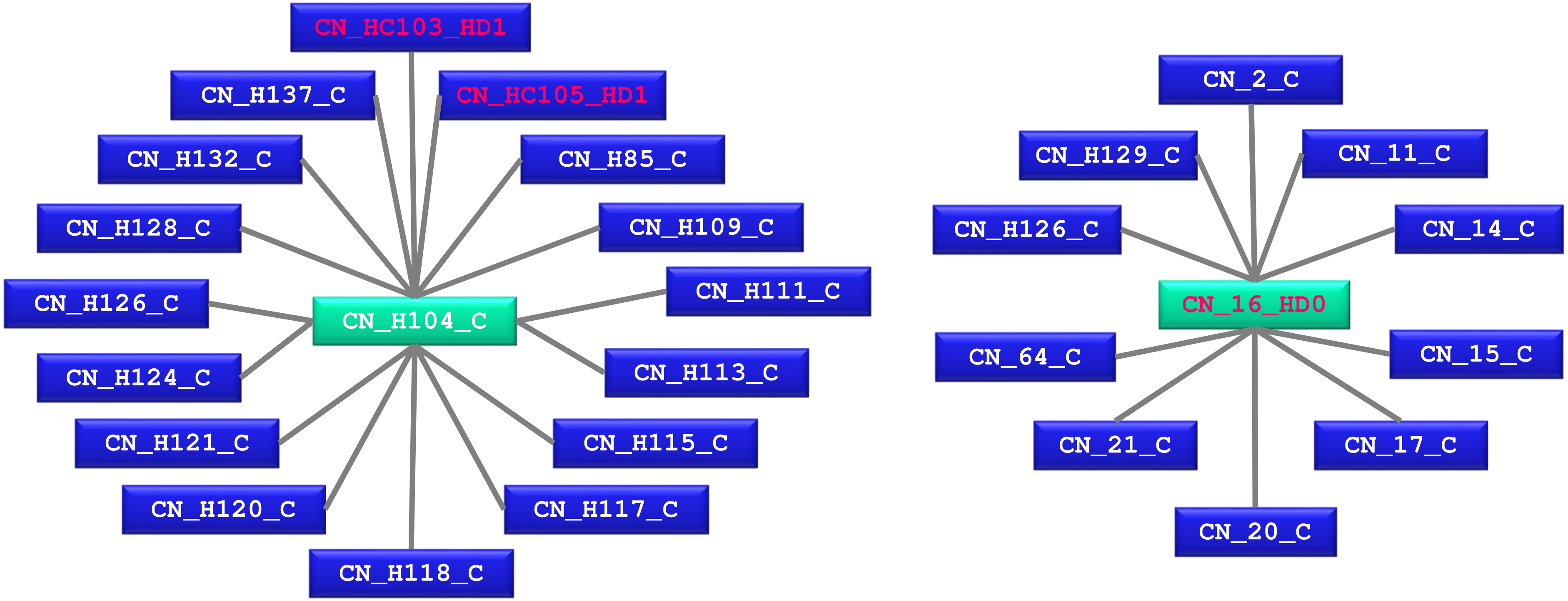
$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

$\tau = 0.90$



just among the **caudate nucleus (CN)** samples (70 samples)

max *p*-median: max total weight = 60.51

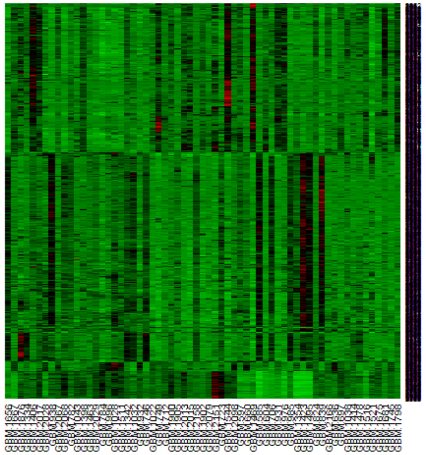


p-median – Applications – Samples network

correlation matrix (*A*)
among samples

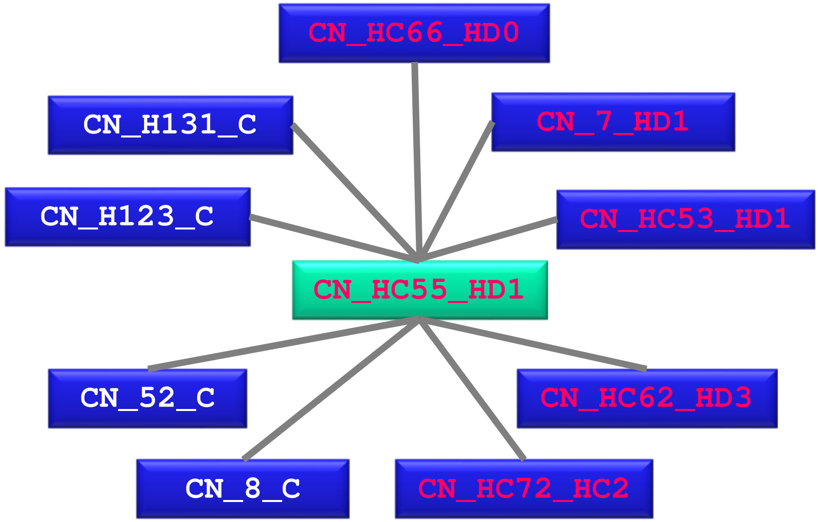
$$a_{ij} = \left(\frac{C(X_i, X_j)+1}{2}\right)^2 \in [0,1]$$

$\tau = 0.90$

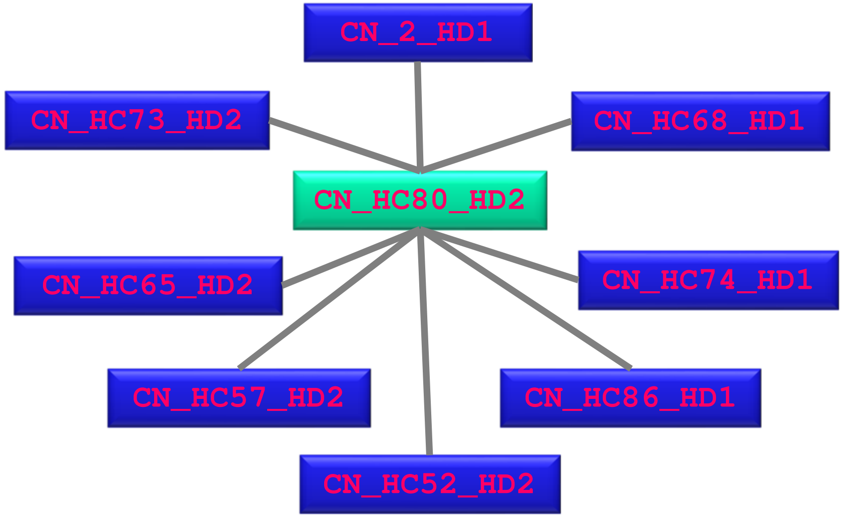


just among the **caudate nucleus (CN)** samples (70 samples)

max *p*-median: max total weight = 60.51



CN_HC66_HD0
CN_H131_C
CN_7_HD1
CN_H123_C
CN_HC53_HD1
CN_HC55_HD1
CN_52_C
CN_HC62_HD3
CN_8_C
CN_HC72_HC2



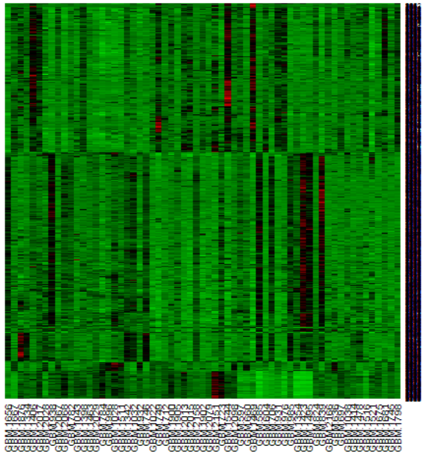
CN_2_HD1
CN_HC73_HD2
CN_HC68_HD1
CN_HC80_HD2
CN_HC65_HD2
CN_HC74_HD1
CN_HC57_HD2
CN_HC86_HD1
CN_HC52_HD2

p-median – Applications – Samples network

correlation matrix (*A*)
among samples

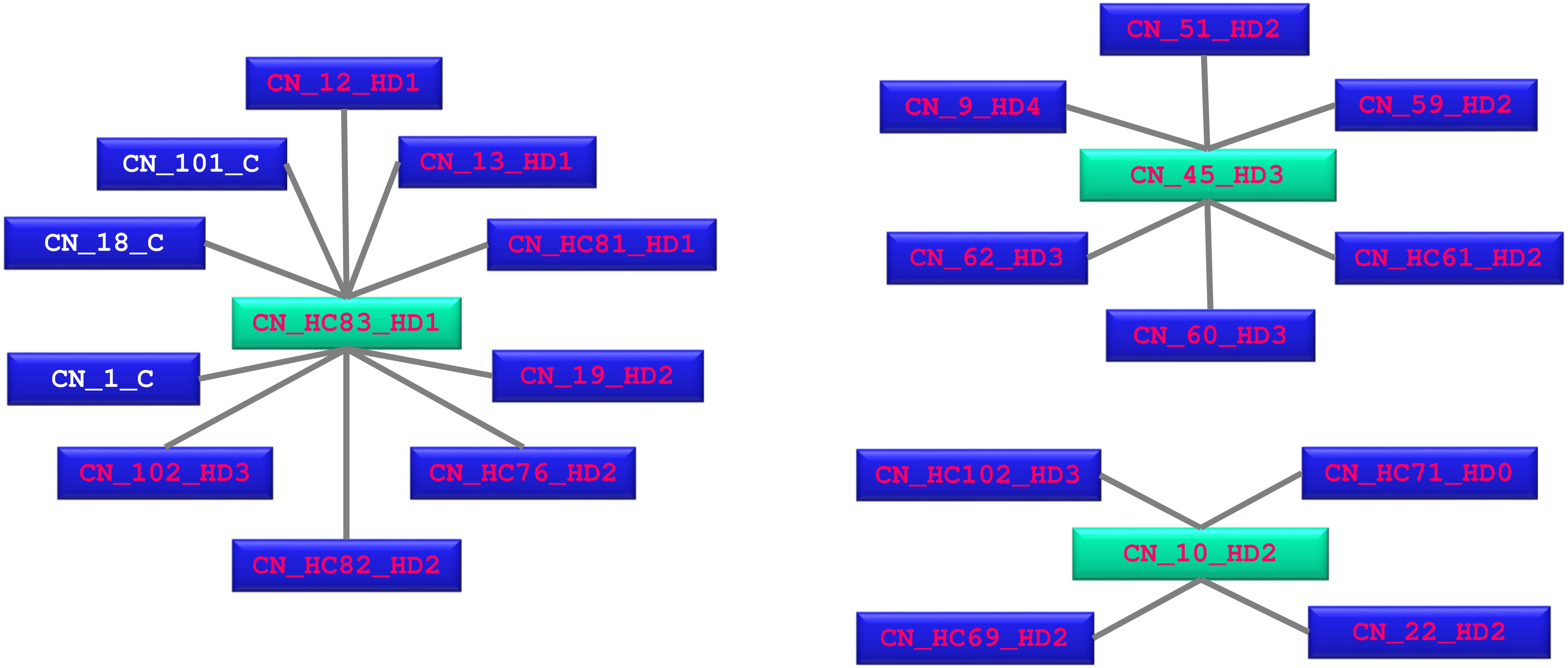
$$a_{ij} = \left(\frac{C(X_i, X_j) + 1}{2} \right)^2 \in [0,1]$$

$\tau = 0.90$



just among the **caudate nucleus** (**CN**) samples (70 samples)

max *p*-median: max total weight = 60.51



CN_HC83_HD1
CN_12_HD1
CN_101_C
CN_13_HD1
CN_18_C
CN_HC81_HD1
CN_1_C
CN_19_HD2
CN_102_HD3
CN_HC76_HD2
CN_HC82_HD2

CN_45_HD3
CN_51_HD2
CN_9_HD4
CN_59_HD2
CN_62_HD3
CN_HC61_HD2
CN_60_HD3

CN_10_HD2
CN_HC102_HD3
CN_HC71_HD0
CN_HC69_HD2
CN_22_HD2