Non-negative (tri) matrix factorization and its application to study human disease

Nucleic Acids Research, 2012, ***1–16***
doi:10.1093/nar/gks615

Co-clustering phenome–genome for phenotype classification and disease gene discovery

TaeHyun Hwang[1], Gowtham Atluri[2], MaoQiang Xie[3], Sanjoy Dey[2], Changjin Hong[4], Vipin Kumar[2] and Rui Kuang[2,*]

Outline

1. Non-negative (tri) matrix factorization

2. Background

3. Our approach

   - Regularized non-negative tri matrix factorixation

4. Experiments

5. Discussion

Algorithms for Non-negative Matrix Factorization

Daniel D. Lee*
*Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974

H. Sebastian Seung*†
†Dept. of Brain and Cog. Sci.
Massachusetts Institute of Technology
Cambridge, MA 02138

Problem formulation

Non-negative matrix factorization (NMF) Given a non-negative matrix V, find non-negative matrix factors W and H such that:

$$V \approx WH \tag{1}$$

Why NMF?

Non-negative matrix factorization (NMF) Given a non-negative matrix V, find non-negative matrix factors W and H such that:

$$V \approx WH \tag{1}$$

SVD $\quad \mathbf{A}_k = \mathbf{U}_k \; \Sigma_k \; \mathbf{V}_k^T$

$mixed \quad nonneg \quad mixed$

NMF $\quad \mathbf{A}_k = \mathbf{W}_k \; \mathbf{H}_k$

$nonneg \quad nonneg \quad nonneg$

Multiplicative update rules

Theorem 1 *The Euclidean distance* $||V - WH||$ *is nonincreasing under the update rules*

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}} \tag{4}$$

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu} \left[(W^T V)_{a\mu} - (W^T W H)_{a\mu} \right]. \tag{6}$$

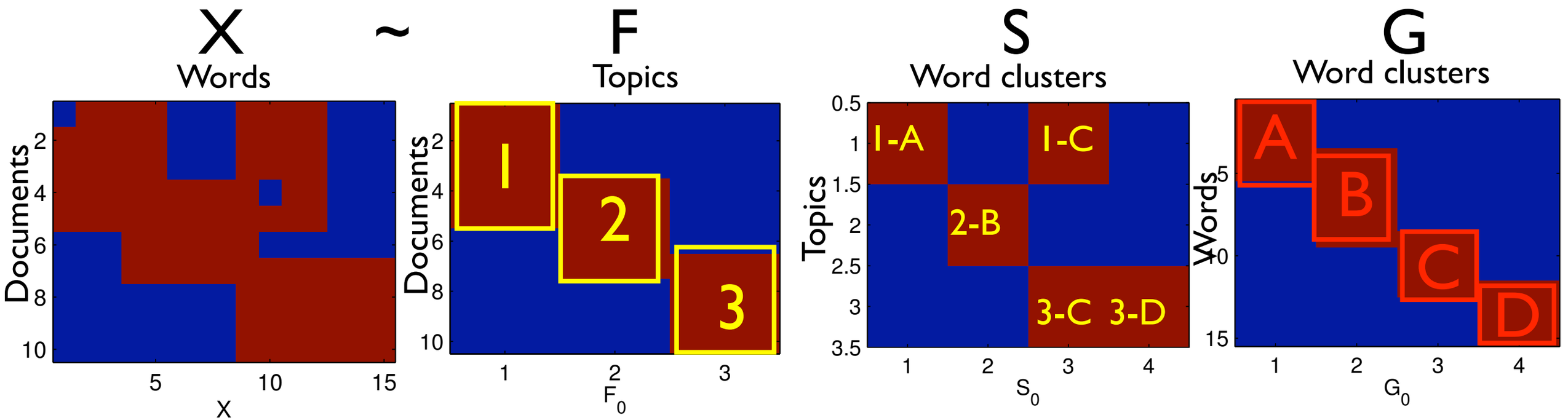
If $\eta_{a\mu}$ are all set equal to some small positive number, this is equivalent to conventional gradient descent. As long as this number is sufficiently small, the update should reduce $||V - WH||$.

Now if we diagonally rescale the variables and set

$$\eta_{a\mu} = \frac{H_{a\mu}}{(W^T W H)_{a\mu}}, \tag{7}$$

then we obtain the update rule for H that is given in Theorem 1.

Non-negative tri matrix factorization



Document-word

Topics

Topic-word cluster association

Word clusters

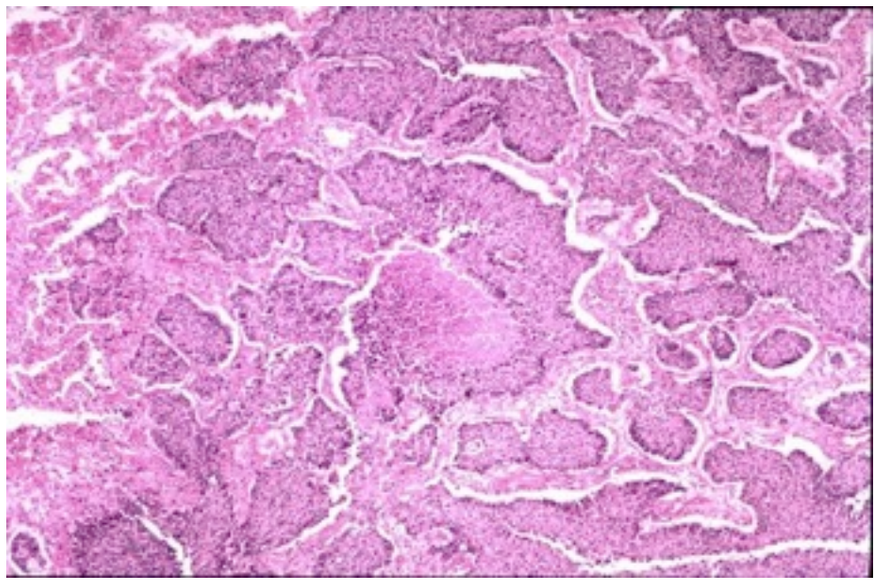
$$\min_{F,S,G} \|X - FSG^T\|_F^2$$

Background

✓ Phenotype

- the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment

- phenotypes could be either disease phenotypes or any other observable characteristics



Cancer

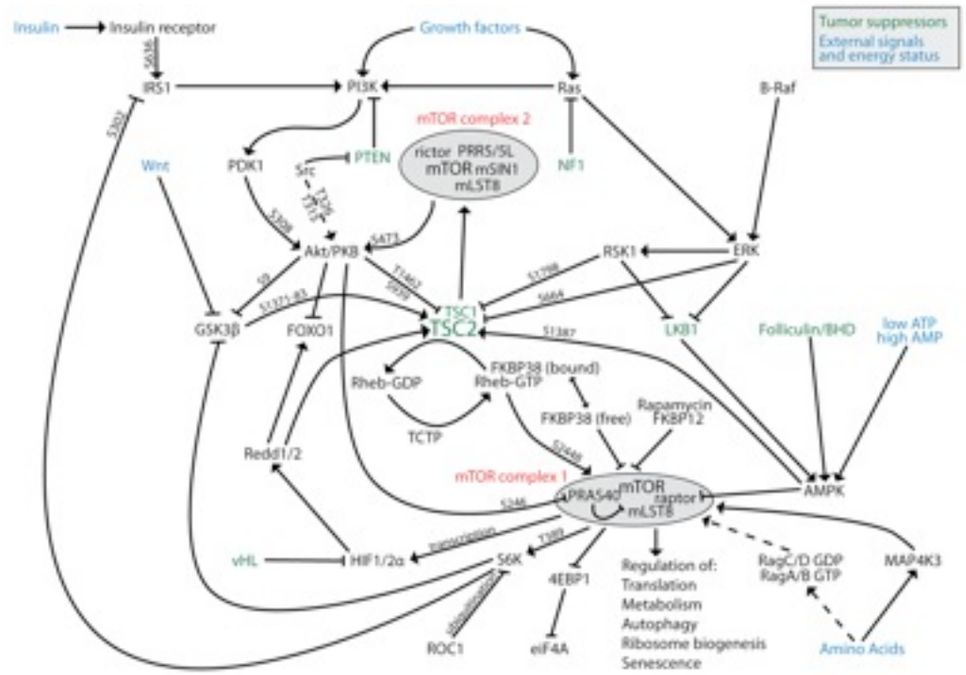
Blond hair

Eye color

Background

✓ Pathway

- a series of actions among molecules (genes) in a cell that leads to a certain end point or cell function



✓ Human disease phenotype classification

- classify a given disease phenotype into 20 disease classes (i.e., cancer, bone, dermatological, neurological, psychiatric, and etc.)

Why human disease classification and disease gene discovery

Gambling is a genetic disorder?

606349

SNOMEDCT: 18085000 ICD10CM: F63.0 ICD9CM: 312.31

GAMBLING, PATHOLOGIC

TEXT

Description

Pathologic gambling is defined as a chronic and progressive failure to resist impulses to gamble accompanied by gambling behavior that compromises or damages personal, family, or vocational pursuits. The prevalence of pathologic gambling in the adult American population is estimated to be between 1 and 3% (review by Eisen et al., 1998).

Comings et al. (2001) noted that some form of gambling is legal in all but 2 states in the U.S., and gambling on the Internet is available to anyone with a computer regardless of the local laws. They stated that as access to gambling has increased, there has been a corresponding increase in the frequency of addiction to gambling, known as pathologic gambling.

Gambling (cont')

Inheritance

{3,2:Eisen et al. (1997,1998)} studied 3,359 mono- and dizygotic U.S. male twin pairs who had served in the military and found that inherited factors explained between 37 and 55% of the prevalence of 5 individual symptoms of pathologic gambling (attempts to wind back losses at same place, gambling larger amounts than intended, repeated efforts to reduce or stop gambling, frequent preoccupation with gambling, and increase betting to maintain interest) from DSM-III-R for which data was informative. In addition, inherited factors plus shared environmental experiences explained 56% of the report of 3 or more symptoms of pathologic gambling and 62% of the diagnosis of pathologic gambling disorder (4 or more symptoms).

Pathogenesis

A number of different neurotransmitters have been thought to be implicated in pathologic gambling. Comings et al. (2001) studied polymorphisms at 31 different genes involved in dopamine, serotonin, norepinephrine, GABA, and other types of metabolism in 139 pathologic gamblers and 139 age, race, and sex-matched controls. Fifteen genes were included in the multivariate regression analysis. The most significant were DRD2 (126450), DRD4 (126452), DAT/DAT1 (SLC6A3; 126455), TPH (191060), ADRA2C (104250), NMDAR1 (138249), and PS1 (104311) genes. Dopamine, serotonin, and norepinephrine genes contributed approximately equally to the risk of pathologic gambling. The results indicated that genes influencing a range of brain functions play an additive role as risk factors for pathologic gambling. Comings et al. (2001) suggested that multigene profiles in specific individuals may help in choosing appropriate treatment

Why human disease classification and disease gene discovery

- Traditional approaches heavily rely on clinical syndromes, and do ***not*** consider molecular basis of disease phenotype

- There is no existing computational method to integrate clinical information and genomic information for disease classification

Motivation

✓ Classify disease phenotypes based on phenotypic and genetic data integration to improve diagnosis, and prognosis accuracy

- only less than ~1000 out of 6543 disease phenotypes are manually annotated in phenotype classes

✓ Identify disease-related pathways/genes to develop efficient targeted therapies, as well as better understanding of molecular mechanisms underlying complex human disease

- which pathways/genes should be effective target to cure disease X?

- which existing drugs targeting specific pathways/genes could be used to treat which diseases?

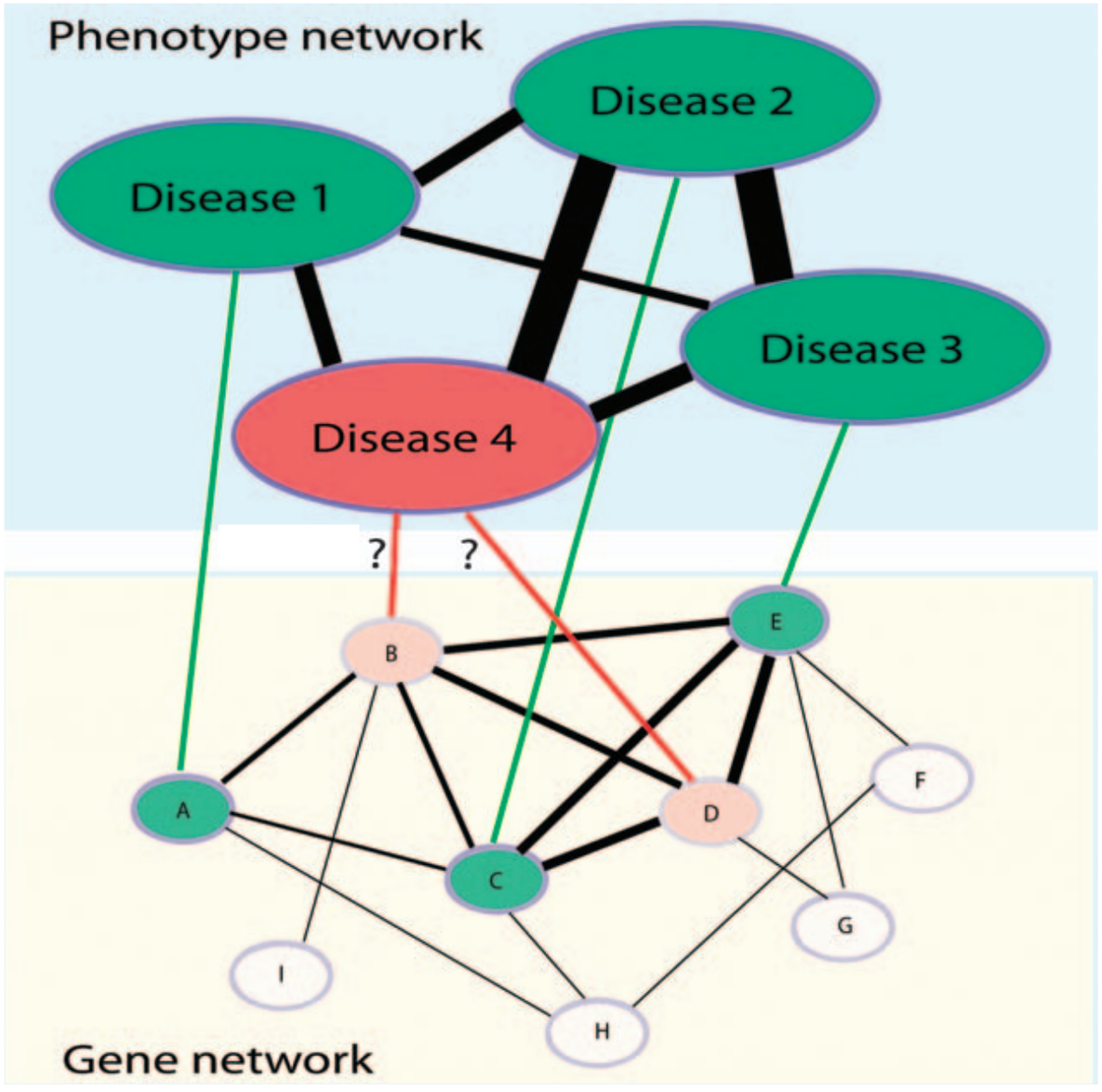
An integrated network view of disease-gene association

- ✓ Phenotypically similar diseases are caused by functionally related disease genes

A)

B)

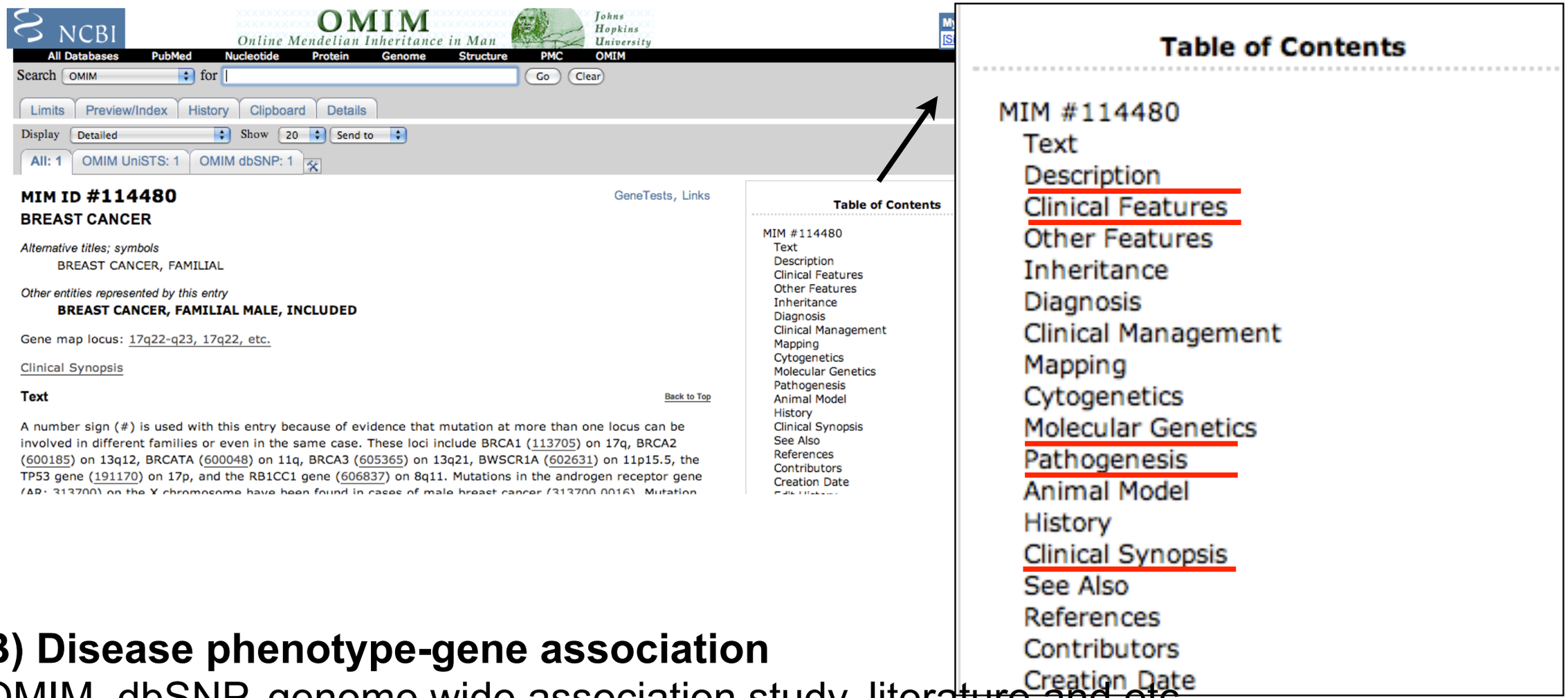
C)



Oti M, Brunner HG, The modular nature of genetic disease, Clinical Genetics 2007

Public database

A) Disease phenotype database (Phenotype similarity network)
Online Mendelian Inheritance in Man (OMIM)

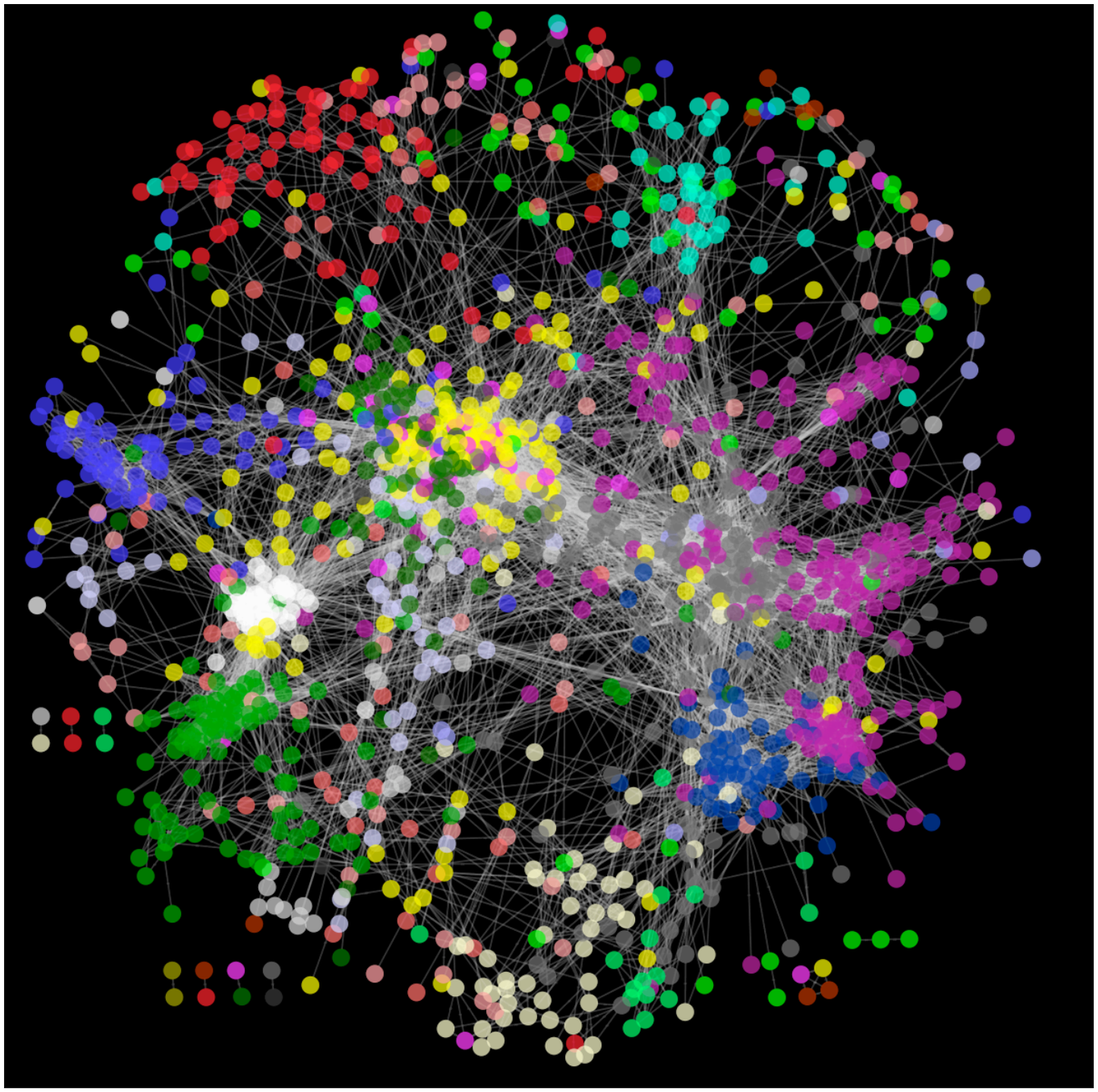


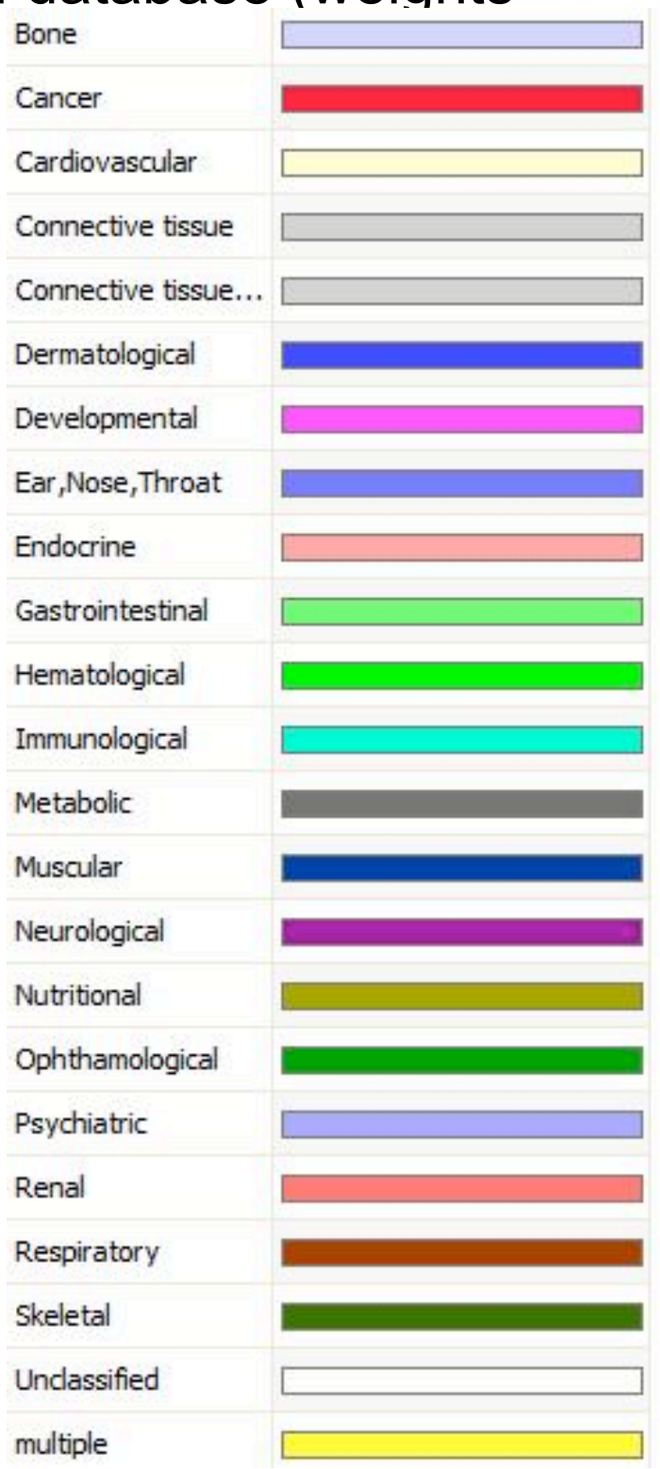
B) Disease phenotype-gene association
OMIM, dbSNP, genome wide association study, literature and etc.

C) Gene-gene interaction database
Protein interaction network (Human Protein Reference Database)
Genetic interaction network, co-expression from microarray gene data
Pathway (KEGG, Biocarta, MSigDB, and etc)

Phenotype similarity network

- Node: disease phenotype in OMIM
- Edge: phenotypical similarity calculated by text mining* with OMIM database (weights > 0.4)





Disease class annotation from Goh et. al, PNAS 2007

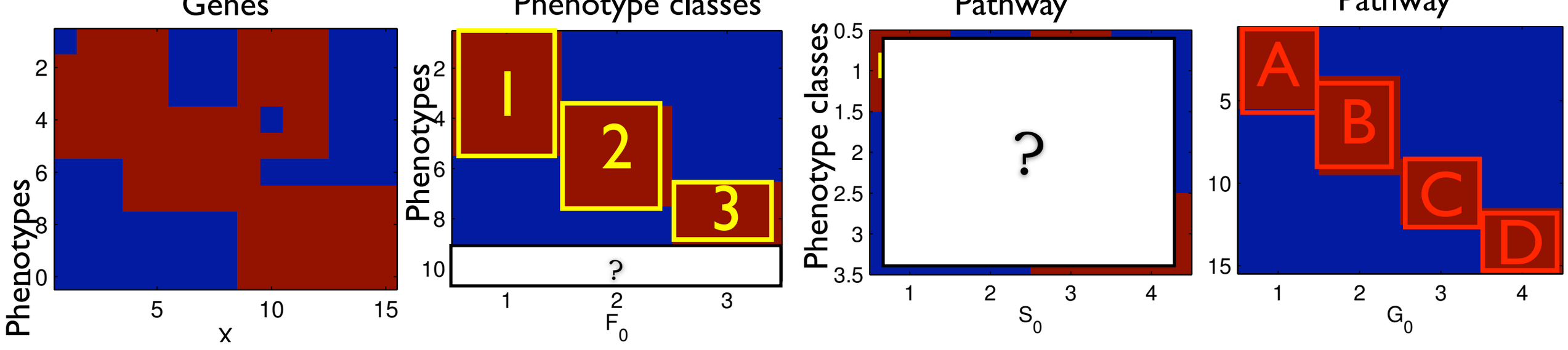
*Marc Driel, et al. "A text-mining analysis of the human phenome", European Journal of Human Genetics 2006

Problem formulation

- Given: phenotype-gene association (X), phenotype class (F), pathway (G), phenotype similarity network (M), gene-gene interaction network (N)
- Task: classify phenotype into 20 phenotype classes, and identify disease-related pathways/genes

Q: Predict the phenotype class of a query disease phenotype

X ~ F S G



Phenotype-gene association

Phenotype classes

Phenotype class-pathway association

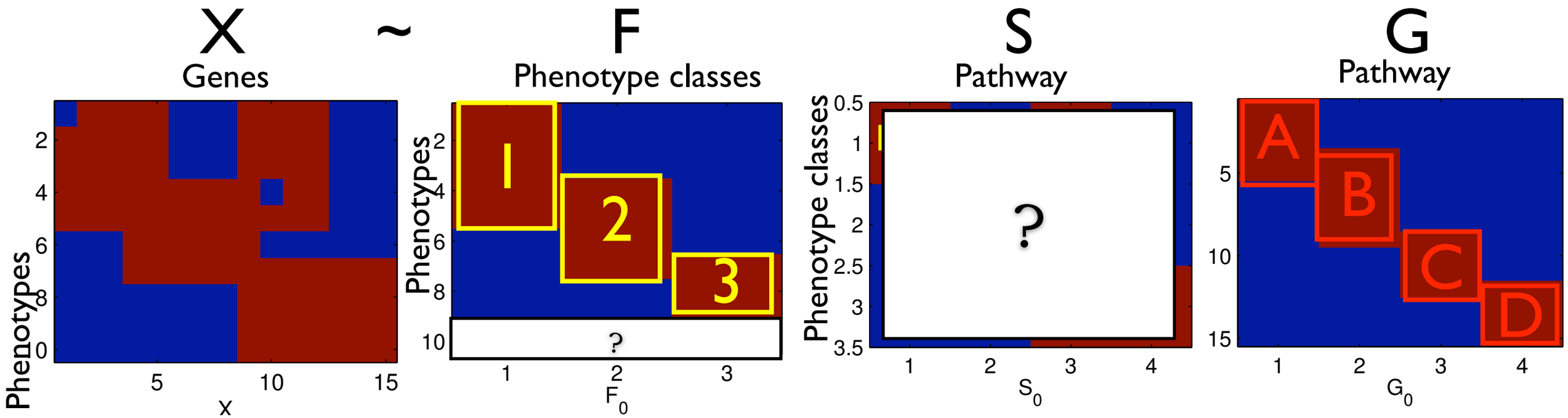
Pathway

$$\min_{F,S,G} \|X - FSG^T\|_F^2 + \alpha\|F - F^0\|_F^2 + \beta\|G - G^0\|_F^2$$

Problem formulation

- Given: phenotype-gene association (X), phenotype class (F), pathway (G), phenotype similarity network (M), gene-gene interaction network (N)
- Task: classify phenotype into 20 phenotype classes, and identify disease-related pathways/genes

Q: Predict the phenotype class of a query disease phenotype



Phenotype-gene association

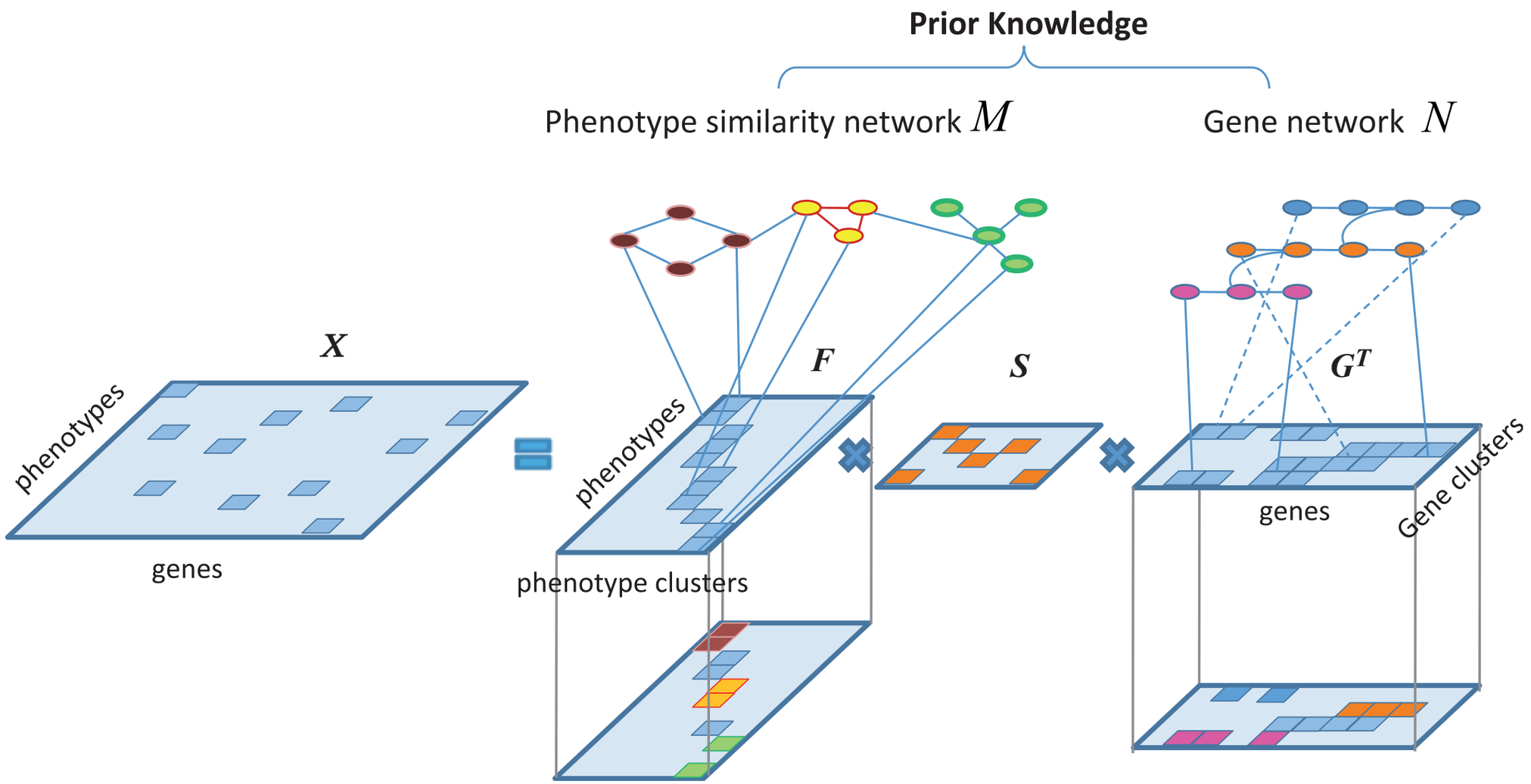
Phenotype classes

Phenotype class-pathway association

Pathway

This model does not incorporate clinical information of phenotypes (M) for classification

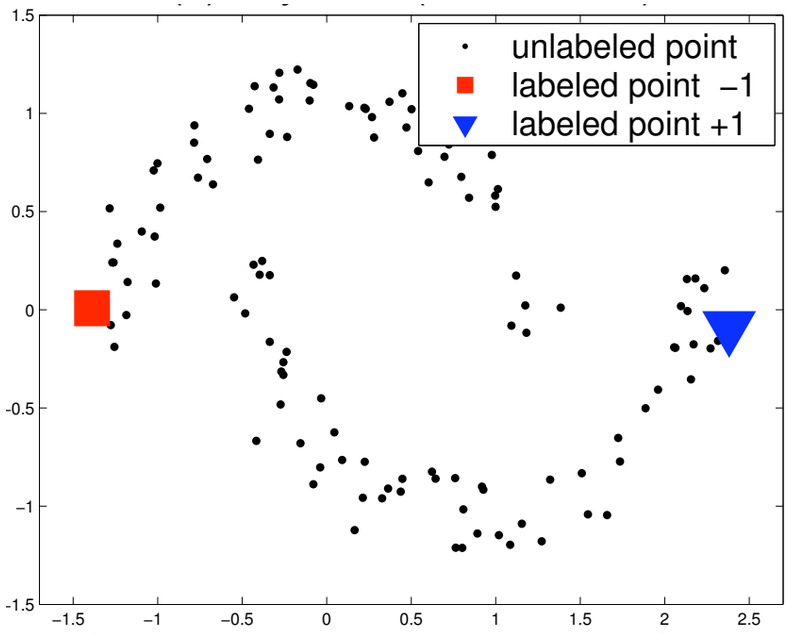
Our proposed method



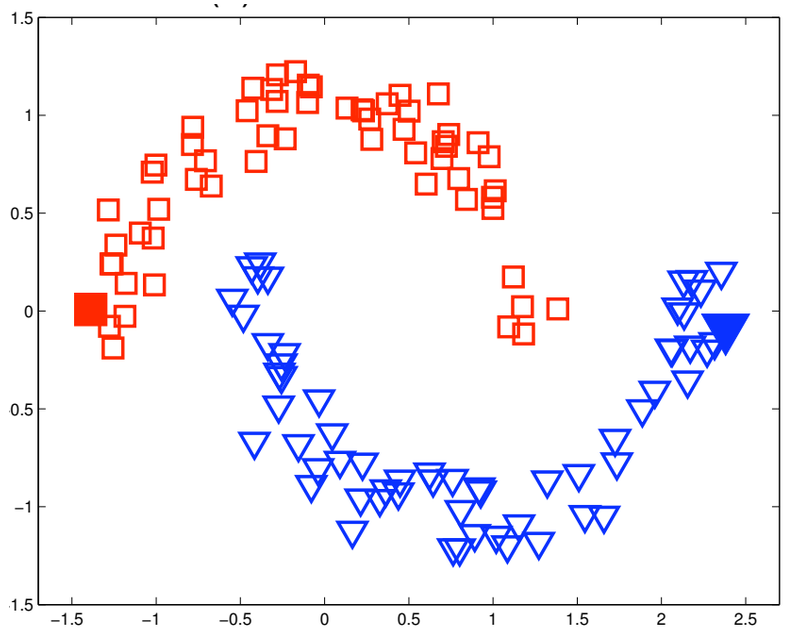
$$\min_{F,S,G} \|X - FSG^T\|_F^2 + \alpha\|F - F^0\|_F^2 + \beta\|G - G^0\|_F^2 + \gamma \mathrm{tr}(F^T(D_M - M)F) + \lambda \mathrm{tr}(G^T(D_N - N)G)$$

✓Integrating phenotype similarity network enables to use clinical information of phenotypes

Graph laplacian



Two moon example



Ideal classification

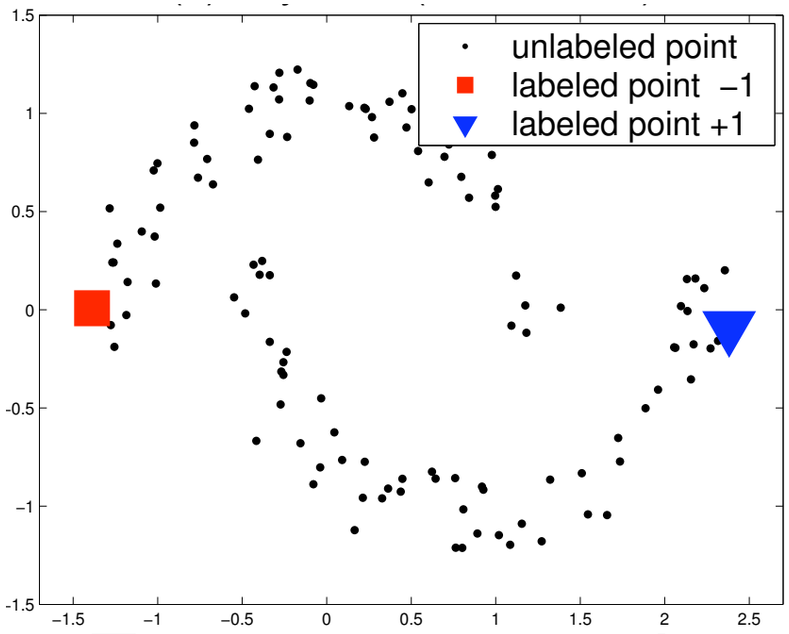
- $n \times n$ weight matrix W
  - symmetric, non-negative
- Diagonal degree matrix D: $D_{ii} = \sum_{j=1}^{n} W_{ij}$
- Graph Laplacian matrix Δ

$$\Delta = D - W$$

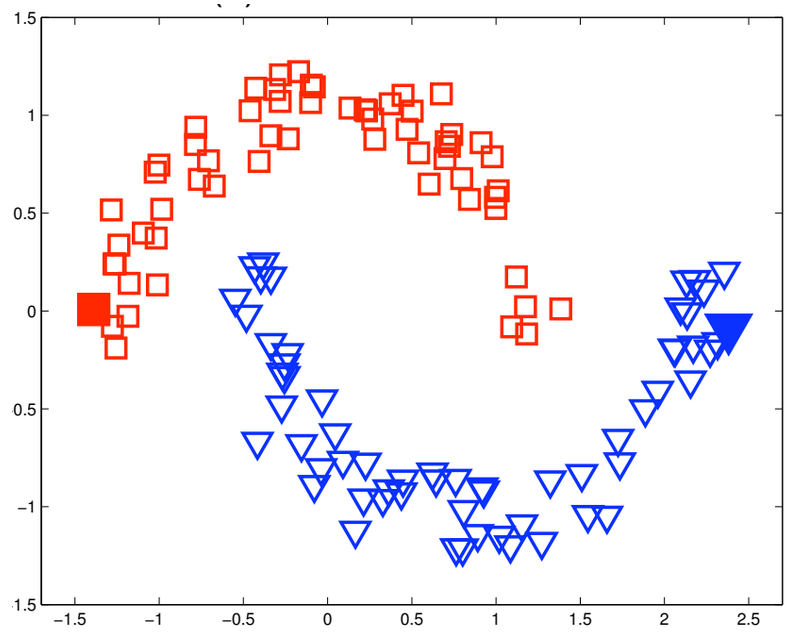
- The energy can be rewritten as

$$\sum_{i \sim j} w_{ij} (f(x_i) - f(x_j))^2 = f^\top \Delta f$$

Graph laplacian



Two moon example



Ideal classification

Key idea: Propagate label information by exploring graph structures

$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 + \mu \sum_{i=1}^{n} \left\| F_i - Y_i \right\|^2 \right)$$

Local consistency — Fitting term

F: label assignment

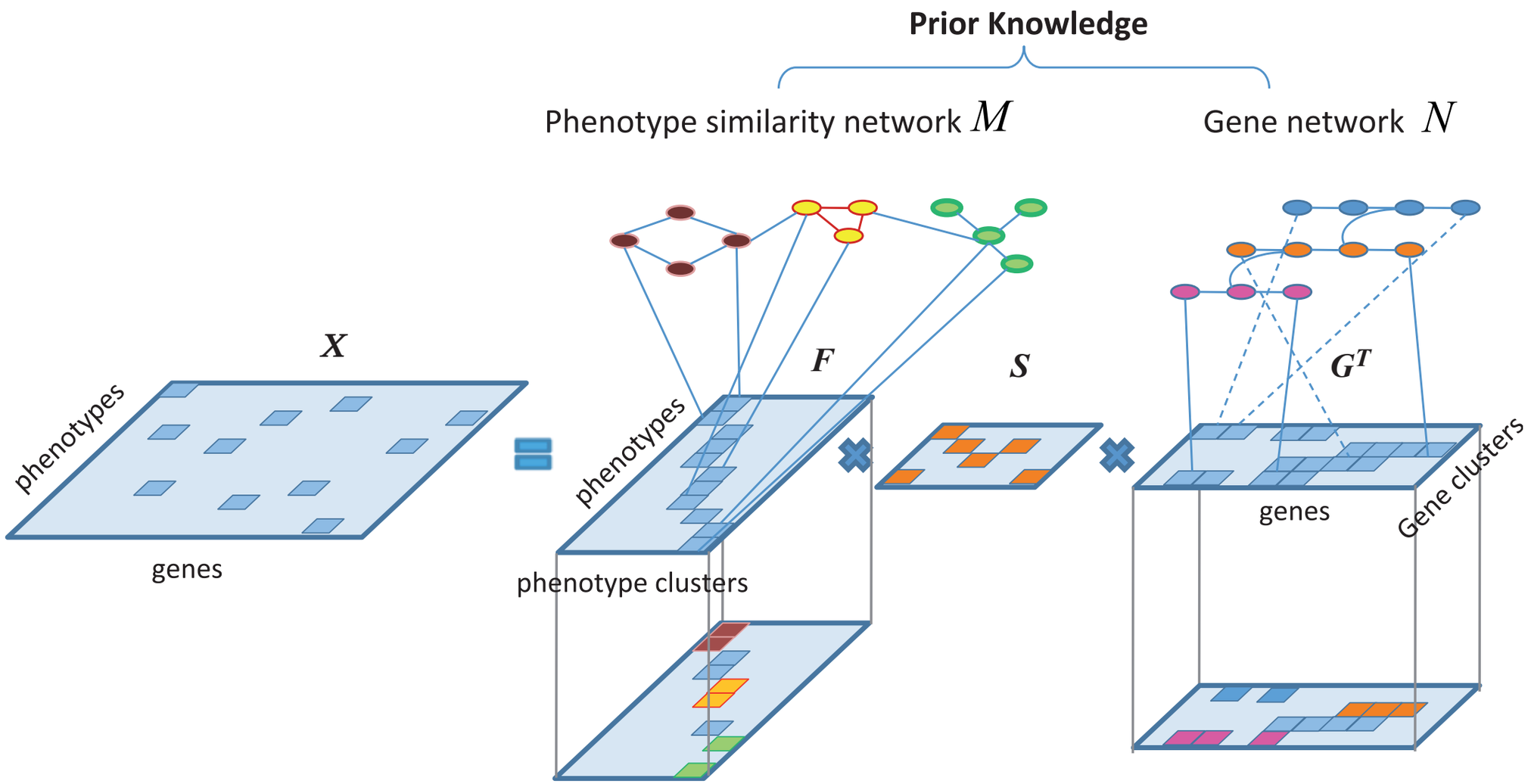
Y: label information

W: affinity matrix

D: diagonal matrix: sum of i-th row of W

D. Zhou et al, "Learning with local and global consistency", NIPS 2004

Our proposed method



$$\min_{F,S,G} \|X - FSG^T\|_F^2 + \alpha\|F - F^0\|_F^2 + \beta\|G - G^0\|_F^2 + \gamma \mathrm{tr}(F^T(D_M - M)F) + \lambda \mathrm{tr}(G^T(D_N - N)G)$$

✓Integrating phenotype similarity network enables to use clinical information of phenotypes

Algorithm

Algorithm 1

Regularized Non-negative Matrix Tri-factorization
INPUT: X, F^0, G^0, L_M, L_N, parameters α, β, γ, and λ, maximum interation T
OUTPUT: F, G, S
while not converged and $t \leq T$ do

(1) Update $F_{ij} \leftarrow F_{ij}\sqrt{\frac{(XGS^T+\alpha F^0+\gamma MF)_{ij}}{(FSG^TGS^T+\alpha F+\gamma D_M F)_{ij}}}$.

(2) Normalize $F_{i.} \leftarrow \frac{F_{i.}}{\sum_{j=1}^{k_1} F_{ij}}$

(3) Update $G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^TFS+\beta G^0+\lambda NG)_{ij}}{(GS^TF^TFS^S+\beta G+\lambda D_N G)_{ij}}}$.

(4) Normalize $G_{i.} \leftarrow \frac{G_{i.}}{\sum_{j=1}^{k_2} G_{ij}}$.

(5) Compute $S_{ij} \leftarrow S_{ij}\sqrt{\frac{(F^TXG)_{ij}}{(F^TFSG^TG)_{ij}}}$.

end while

Multiplicative update rules

Computation of F

If we fix variables S and G, solving equation (2) with respect to F is equivalent to minimizing the following function:

$$L(F) = \|X - FSG^T\|_F^2 + \alpha\|F - F^0\|_F^2 + \gamma \mathrm{tr}(F^T L_M F)$$

subject to $\sum_{j=1}^{k_1} F_{i,j} = 1$, where L_M is $D_M - M$.

The differentiation of L with respect to F is

$$\frac{\partial L(F)}{\partial F} = -2XGS^T + 2FSG^TGS^T + 2\alpha(F - F^0) + 2\gamma L_M F.$$

The multiplicative update rule is

$$F_{ij} \leftarrow F_{ij}\sqrt{\frac{(XGS^T + \alpha F^0 + \gamma MF)_{ij}}{(FSG^TGS^T + \alpha F + \gamma D_M F)_{ij}}}.$$

To satisfy the equality constrain, we normalize F as

$$F_{i.} \leftarrow \frac{F_{i.}}{\sum_{j=1}^{k_1} F_{ij}}.$$

Multiplicative update rules

Computation of G

If we fix variables S and F, solving equation (2) with respect to G is equivalent to minimizing the function,

$$L(G) = \|X - FSG^T\|_F^2 + \alpha\|G - G^0\|_F^2 + \gamma \mathrm{tr}(G^T L_N G)$$

subject to $\sum_{j=1}^{k_2} G_{ij} = 1$, where L_N is $D_N - N$.

The differentiation of L with respect to G is

$$\frac{\partial L(G)}{\partial G} = -2X^T FS + 2GS^T F^T FS^S + 2\beta(G - G^0) + 2\lambda L_N G.$$

The multiplicative update rule is

$$G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^T FS + \beta G^0 + \lambda NG)_{ij}}{(GS^T F^T FS^S + \beta G + \lambda D_N G)_{ij}}}.$$

To satisfy the equality constrain, we normalize G as

$$G_{i.} \leftarrow \frac{G_{i.}}{\sum_{j=1}^{k_2} G_{ij}}.$$

Multiplicative update rules

Computation of S

After F and G are computed, solving equation (2) with respect to S is equivalent to minimizing the following function:

$$L(S) = \|X - FSG^T\|_F^2.$$

The differentiation of L with respect to S is

$$\frac{\partial L(S)}{\partial S} = -2F^T XG + 2F^T FSG^T G.$$

The multiplicative update rule is

$$S_{ij} =\leftarrow S_{ij}\sqrt{\frac{(F^T XG)_{ij}}{(F^T FSG^T G)_{ij}}}.$$

Data preparation

1. Disease phenotype similarity network

- 1325 disease phenotypes
  - 501 labeled disease phenotypes
- Edges are weighted by pairwise disease similarities among 1325 disease phenotypes calculated by text mining techniques [Marc Driel, et al., European Journal of Human Genetics 2006]

2. Disease-gene association network [OMIM database., May 2007]

- an undirected bi-partite graph with disease and gene vertices
- 1126 disease-gene associations

3. Protein interaction networks [HPRD database., May 2007]

- 7997 proteins are mapped to human genes
- 34364 binary-valued undirected interactions between 7997 proteins
- Self-interactions are removed

4. KEGG pathway

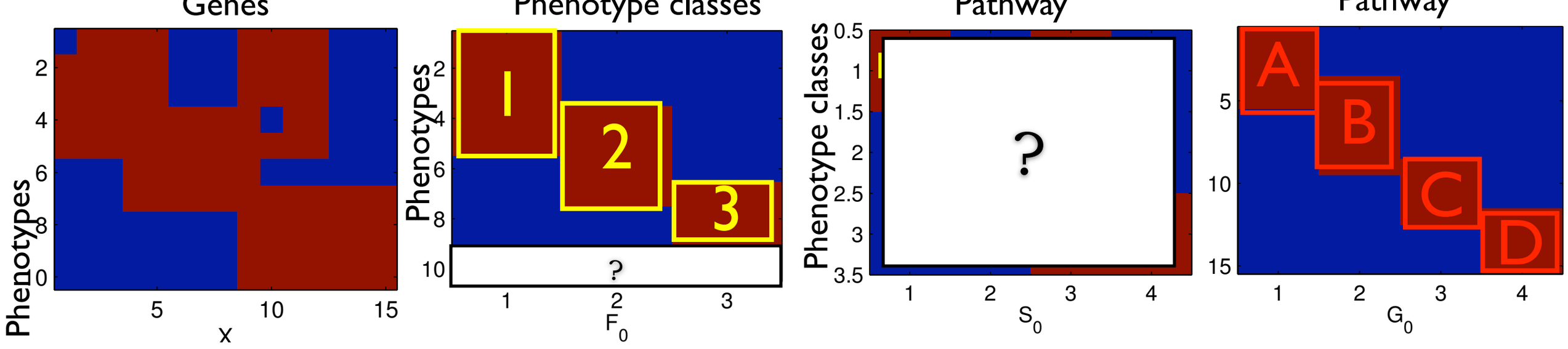
- 4128 genes in 200 pathways

Problem formulation

- Given: phenotype-gene association, phenotype class, pathway, phenotype similarity network, gene-gene interaction network
- Task: classify phenotype into 20 phenotype classes, and identify disease-related pathways/genes

Q: Predict the phenotype class of a query disease phenotype

X ~ F S G



Phenotype-gene association

Phenotype classes

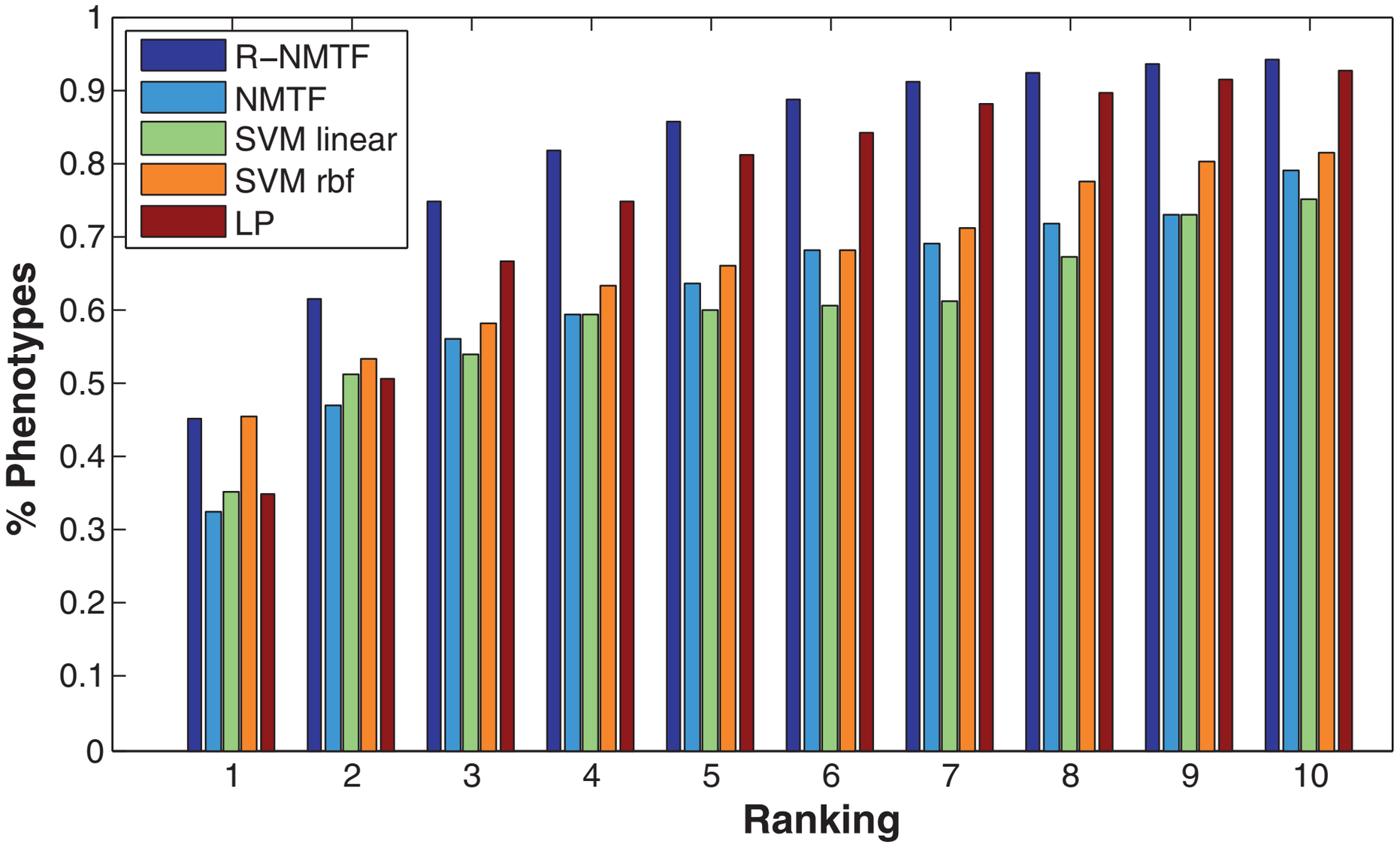
Phenotype class-pathway association

Pathway

Leave-one-out cross validation

Table 2. Performance of phenotype classification in leave-one-out cross-validation

Compared methods	Avg. rank	win/draw/loss (*P*-value)
R-NMTF versus NMTF	**3.124** versus 5.590	300/154/136 (4.617e−13)
versus SVM-linear	versus 6.103	308/154/128 (3.693e−12)
versus SVM-rbf	versus 5.037	268/213/109 (1.497e−4)
versus LP	versus 3.700	161/388/41 (9.145e−05)



Leave-one-out cross validation

Table 3. Disease phenotype classification results by disease classes

Disease classes (No)	Avg. rank				
	R-NMTF	NMTF	SVM-linear	SVM-rbf	LP
Bone (23)	**3.3**	8.5	4.7	7.6	4.7
Cancer (53)	**1.6**	5.0	4.2	2.0	1.9
Cardiovascular (28)	**3.8**	10.1	10.0	6.0	4.3
Connective tissue (16)	**8.5**	8.9	10.6	11.4	11.1
Dermatological (32)	**2.0**	4.4	3.0	4.0	2.5
Developmental (28)	5.7	**2.5**	9.6	9.2	6.5
Ear,Nose,Throat (3)	20.0	20.0	**14.7**	15.0	16.7
Endocrine (30)	**4.2**	5.4	13.4	5.4	4.9
Gastrointestinal (12)	9.7	7.8	**7.8**	9.7	11.7
Hematological (30)	3.5	9.5	**2.3**	6.9	3.8
Immunological (31)	**2.6**	10.0	8.1	5.2	2.8
Metabolic (84)	**1.0**	2.2	4.1	2.2	**1.0**
Muscular (18)	5.7	**5.3**	12.2	9.1	7. 3
Neurological (80)	**1.4**	6.2	5.8	2.7	1.4
Nutritional (2)	16.0	3.0	19.0	**2.0**	20
Ophthamological (35)	**1.9**	4.2	2.5	2.9	2.5
Psychiatric (9)	7.9	**6.1**	8.0	11.4	14.8
Renal (23)	4.1	**3.5**	4.4	6.8	4.9
Respiratory (7)	15.4	**10.4**	**10.4**	14.1	15.7
Skeletal (46)	**1.5**	3.3	4.8	5.2	1.8

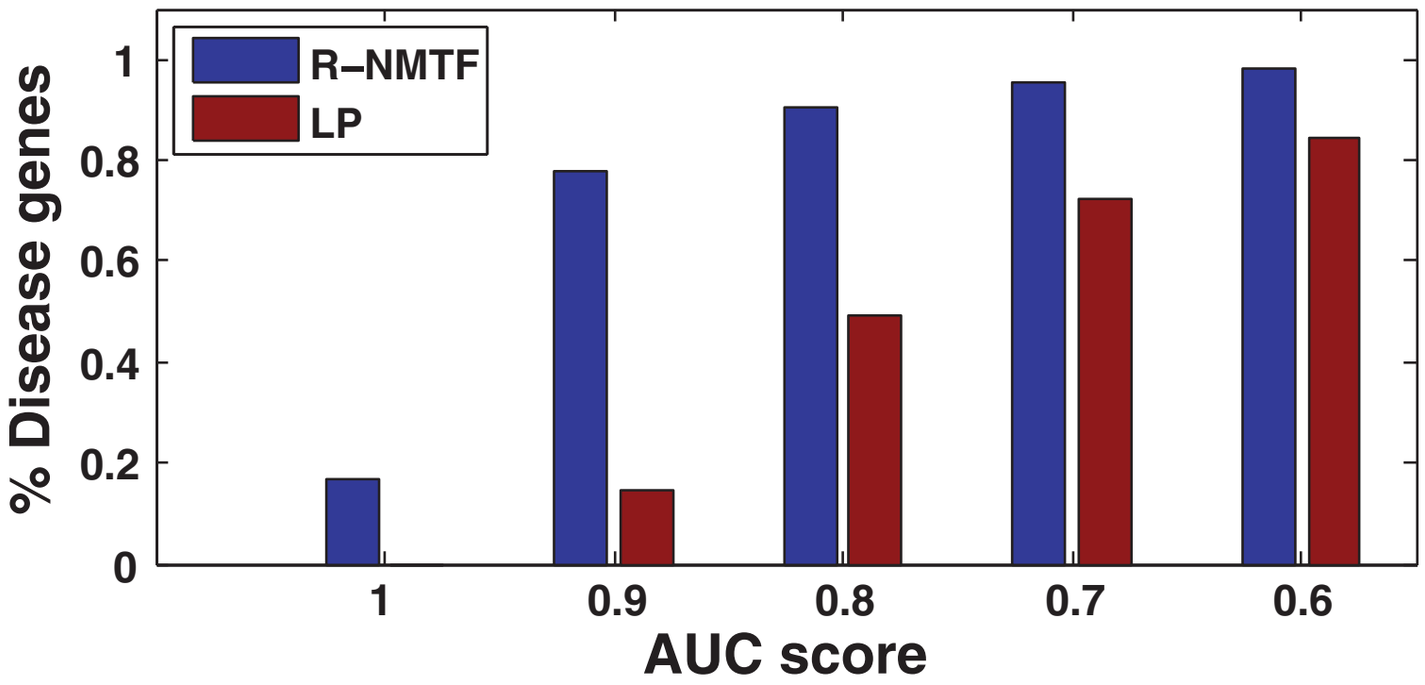
Leave-one-out cross validation (Gene)

- 590 genes are annotated in 27 KEGG disease pathways such as Alzheimer, diabetes or cancer-related pathways
- In leave-one-out cross validation, we remove annotations of one of 590 genes, and classify into 200 KEGG pathways

Table 4. Performance of disease gene discovery in leave-one-out cross-validation

Compared methods	Avg. AUC	win/draw/loss (*P*-value)
R-NMTF versus LP	**0.930** versus 0.730	526/1/63 (5.4482e−113)

This table reports the average AUC for disease gene classification, and the pairwise 'win/draw/loss' comparisons of each leave-one-out case between R-NMTF and LP. The last column reports the statistical significance of ranking results using Wilcoxon rank sum test.



Newly predicted disease phenotypes

Table 5. New disease phenotypes in 20 disease classes

Disease classes	New disease phenotypes				
Bone	Achondrogenesis, Type III (44)	Canine Teeth (Omim:114600)	Dens Evaginatus (45)	Dental Noneruption (46)	Dentin Dysplasia, Type I(47)
Cancer	Fanconi Anemia (48)	Juvenile Myelomonocytic Leukemia	Breast Cancer	Proteus Syndrome (49,50,51,52)	Bannayan-Riley-Ruvalcaba Syndrome (53,54)
Cardiovascular	Cardiomyopathy (Omim:192600)	Atrial Standstill (55)	Cardiomyopathy, Dilated, 1E	Long Qt Syndrome 3 (56,57)	Sudden Infant Death Syndrome (58)
Connective tissue	Arthritis, Sacroiliac (59)	Spondyloarthropathy (Omim:183840)	Slipped Femoral Capital Epiphyses (60)	Facial Asymmetry (61)	Cervical Rib
Dermatological	Deafness; Dfna3 (62)	Epidermolysis Bullosa (Omim:131800)	Pachyonychia Congenita, Type 1 (63)	Epidermolysis Bullosa Herpetiformis (64)	Epidermolysis Bullosa Simplex, Koebner Type (64)
Developmental	Leucine Transport, High	Uterine Anomalies (65)	Testes, Rudimentary (66)	Oligosynaptic Infertility	Hypospadias, Autosomal (67)
Ear,Nose,Throat	Otosclerosis 3 (68)	Otosclerosis 2 (68)	Otosclerosis 5 (68)	Periodontitis, Aggressive, 2	Red Cell Permeability Defect
Endocrine	Diabetes Mellitus	Hypoglycemia (Omim:601820) (69)	Polycystic Ovary Syndrome 1 (70)	Diabetes Mellitus, Transient Neonatal	Goiter, Multinodular 2 (71)
Gastrointestinal	Cholestasis2 (Omim:605479) (72)	Bile Acid, Synthetic Defect Of	Cholestasis; Pfic2 (Omim:601847) (72)	Cholestasis; Pfic3 (Omim:602347) (72)	Pancreatitis, Hereditary (73)
Hematological	Anemia (74)	Hyperheparinemia	Sideroblastic Anemia, Autosomal (75)	Platelet Groups–ko System	Anemia, Familial Pyridoxine-Responsive (76)
Immunological	Herpesvirus Sensitivity (77)	Interleukin (Omim:243110) (78)	Panbronchiolitis, Diffuse (79)	Immune Deficiency Disease	Allergic Bronchopulmonary Aspergillosis (80)
Metabolic	Immunoglobulin D Level In Plasma	Magnesium, Elevated Red Cell	Flood Factor Deficiency	Citrulline Transport Defect	Amobarbital, Deficient N-Hydroxylation of
Muscular	Palmomental Reflex	Myopathy (Omim:255100)	Muscular Hypoplasia	Pleoconial Myopathy With Salt Craving	Myopathy, Congenital
Neurological	Amyotrophic Lateral Sclerosis 1	Amyotrophic Lateral Sclerosis 2	Alzheimer Disease 2	Prion Disease (Omim:603218)	Frontotemporal Dementia (Omim:607485)
Nutritional	Bulimia Nervosa	Red Cell Permeability Defect	Labia Minora (Omim:149600) (81)	Schizophrenia 9 (82)	Amyotrophic Lateral Sclerosis 6 (83)
Ophthamological	Cone Dystrophy 3	Cone-Rod Dystrophy 3	Leber Congenital Amaurosis	Cone-Rod Dystrophy 6	Retinitis Pigmentosa 19
Psychiatric	Fg Syndrome 2 (86)	Fg Syndrome 3 (84)	Schizophrenia 5	Cerebral Angiopathy, Dysphoric (85,86)	Gambling, Pathologic
Renal	Nephrotic Syndrome, Type 2 (87,88)	Hypertensive Nephropathy (89)	Enuresis, Nocturnal, 2 (90)	Enuresis, Nocturnal, 1 (90)	Blue Diaper Syndrome
Respiratory	Hemangiomatosis	Respiratory Underresponsiveness	Emphysema (Omim:130700)	Asthma, Short Stature, and Elevated Iga	Asthma-Related Traits, Susceptibility To, 1
Skeletal	Brachydactyly, Mononen Type	Tibial Hemimelia (91)	Acropectoral Syndrome	Syndactyly, Type IV	Spondyloepimetaphyseal Dysplasia, Irapa Type

The 5 most confident predictions of phenotypes in each disease class are reported.

Newly predicted disease genes

Table 6. New member genes of KEGG disease pathways

Kegg disease pathways	New member genes									
Hsa04930: Type II Diabetes Mellitus	KCNJ8 (92)	EFHC1	ADIPOR2 (93)	ABCC9	LDHA	CDH13 (94)	ENSA	CRYBB1	CASR	KCNJ2
Hsa04940: Type I Diabetes Mellitus	CKAP5	SPTBN4	PTPRT	SNX19	CD74	LILRB1 (95)	LILRB2	GAST	LRRC23	CTLA4 (96)
Hsa04950: Maturity Onset Diabetes of the Young	OLIG2	EN2	PCSK1	PNRC1	PCSK2	GATA5	GATA6	PNRC2	OTX2	RAMP2
Hsa05010: Alzheimers Disease	TMED10 (97)	BRI3	PTX3	APH1B (98)	TFCP2 (99)	HRG	C1R	FKBP2	KHSRP	NEDD8 (100)
Hsa05020: Parkinsons Disease	ARIH1 (101)	AMFR	AGXT	TRIM25	CCNB1IP1	GAN	TMCC2	STUB1	SH2D3C	SLC6A1
Hsa05030: ALS	SSR3	JUB	ALS2CL (102)	APBA1	MTMR2	ABL2	HOXB2	RAB37	PKN1 (103)	CHML
Hsa05040: Huntingtons Disease	HIP1R (104)	SNX5	IFT20	PICALM	RPS10	PQBP1	NECAP1	ARF1	KPNA4	MBTPS1
Hsa05050: Dentatorubropallidoluysian Atrophy	ALG13	TRIM22	CLCN5	ECM1	MYST3	NET1	SYNPO	EFEMP1	CPSF6	NDFIP2
Hsa05060: Prion Disease	PRND (106)	CHD6	LAMA2	RPS21	EIF2AK3	KEAP1	ADAM23	DPP6	MOG	OPCML
Hsa05110: Cholera Infection	SERP1	SEC63	ARFIP2	APOB	ARFIP1	PIP5K1A	FLAD1	TRAM1	ETHE1	AP1B1
Hsa05120: Epithelial Cell Signaling in Helicobacter Pylori Infection	GRLF1	ETHE1	HBA1	EFNA2	TOMM34	DARC	ADD2	SH3D19	PFKM	ANG
Hsa05130: Pathogenic Escherichia Coli Infection Ehec	ARPC4	GRM7	HS1BP3	CGN	PLA2G7	KIAA1543	LAPTM4A	NOX4	ACTR2	SSB
Hsa05210: Colorectal Cancer	EXO1 (106,107)	ADIPOR1 (108,109)	MUTYH (111)	PMS2	CDCA8	ROR2	PMS1	MAZ	WNT5A	WNT7A
Hsa05211: Renal Cell Carcinoma	HIF3A	OS9	EGLN2	ING4	ARNTL2	SIM1	ASB8	LRRC41	SENP6	SIM2
Hsa05212: Pancreatic Cancer	REPS1	REPS2	PLCD1	SHFM1	EXOC1	RAD51AP1	RAD54L	RALGPS1	EXOC5	EXOC3
Hsa05213: Endometrial Cancer	MSR1	BRCA2 (112)	NF1	MXI1 (113)	RNASEL	FH	MSH2	ELAC2	MAD1L1	CHEK2
Hsa05214: Glioma	PDAP1	KIAA1683	RHBDF1	RPS18	ART1	BRD2	NKD2	MYO10	TFDP2	SETD8
Hsa05215: Prostate Cancer	KRT27	MTTP	ATF6 (114)	PTHLH (115)	SEMG1	ATF2 (116)	G6PC	NFIL3 (117)	ASGR1	MALL
Hsa05216: Thyroid Cancer	TSSK2	TMOD2	RNF14	TRIM25	PPP4C	IFI16	CNN1	TMOD1	S100A2	NUP98
Hsa05217: Basal Cell Carcinoma	IHH	DHH	ZIC1	ZIC2	PORCN	SFRP1	ROR2	FRMPD4	GPC3	GAS1
Hsa05218: Melanoma	FGFR4 (118)	FGFR2 (119)	PHEX	FGFR3 (120,121)	SCN8A	EBNA1BP2	RPS2	MAPK8IP2	TFEB	PDAP1
Hsa05219: Bladder Cancer	MLC1	UNC5B (122)	UNC5A	PAWR	AATF	TNXB	CAMK2A	RECK	HIST3H2A	ATF4 (123)
Hsa05220: Chronic Myeloid Leukemia	APBA3	MAP4K5 (124)	BAZ2B	KLF3	TDGF1	MAPK4	FMOD	RAI2	ELF2	SPRY2 (125)
Hsa05221 Acute Myeloid Leukemia	RPL21	NDUFB8 (126)	FBXO18	GATA2 (127)	CEBPD (128)	GFI1 (129)	TAF9B	MYST3 (130)	CBFA2T3	NFATC1
Hsa05222: Small Cell Lung Cancer	CKS2	BCKDK	TBC1D8	TNFRSF19	DUSP1	TNFRSF4	TNFRSF12A	NGFRAP1	LTBR	MAP6
Hsa05223: Non Small Cell Lung Cancer	FDXR (131)	LATS1 (132)	MAP6	NR1H2 (133)	PRKRIR	CSN1S1	NR1H3	CNKSR1	FOXG1 (134)	PNRC1

The 10 most confident predictions of member genes in KEGG disease pathways are reported.

Newly predicted disease genes

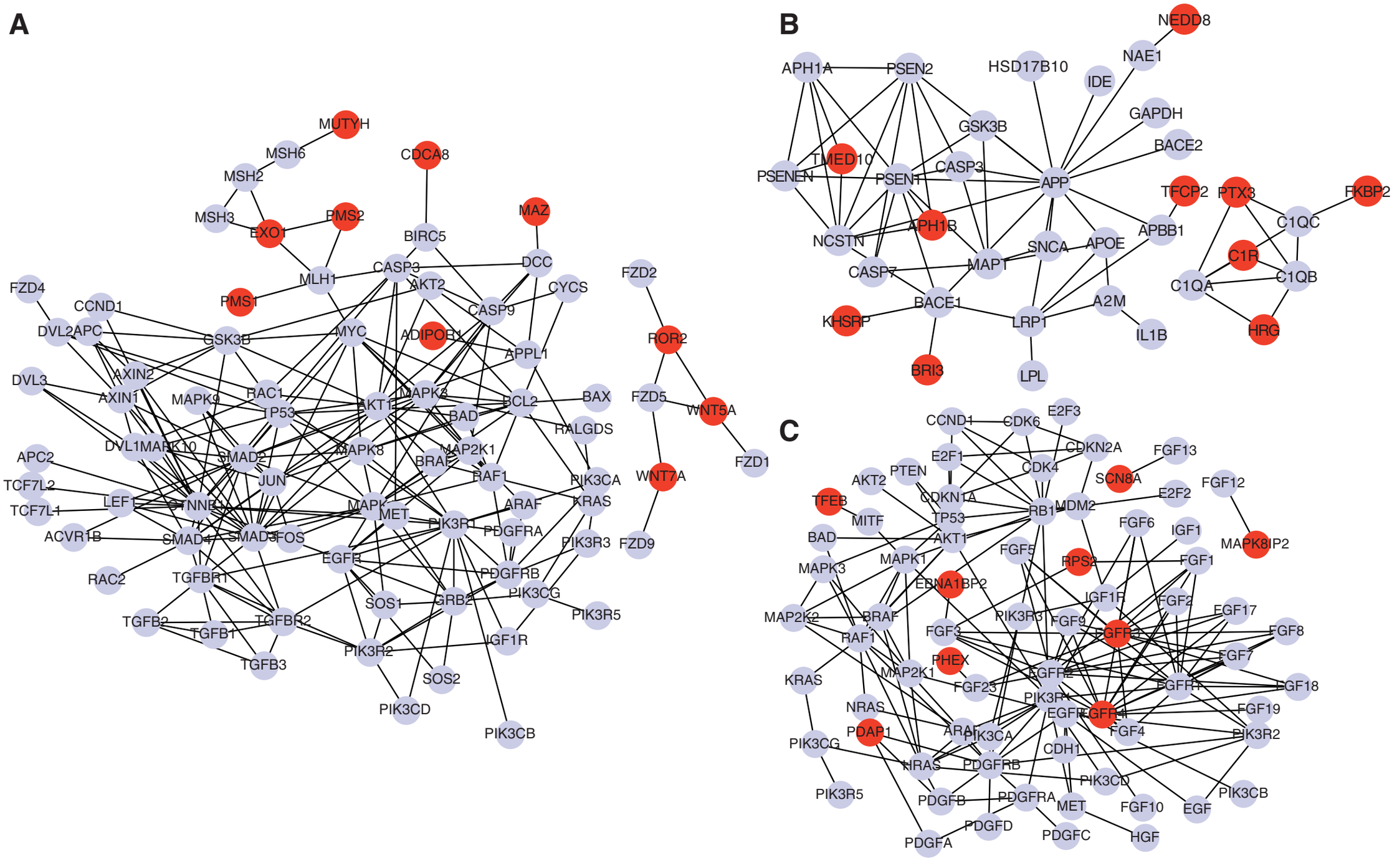
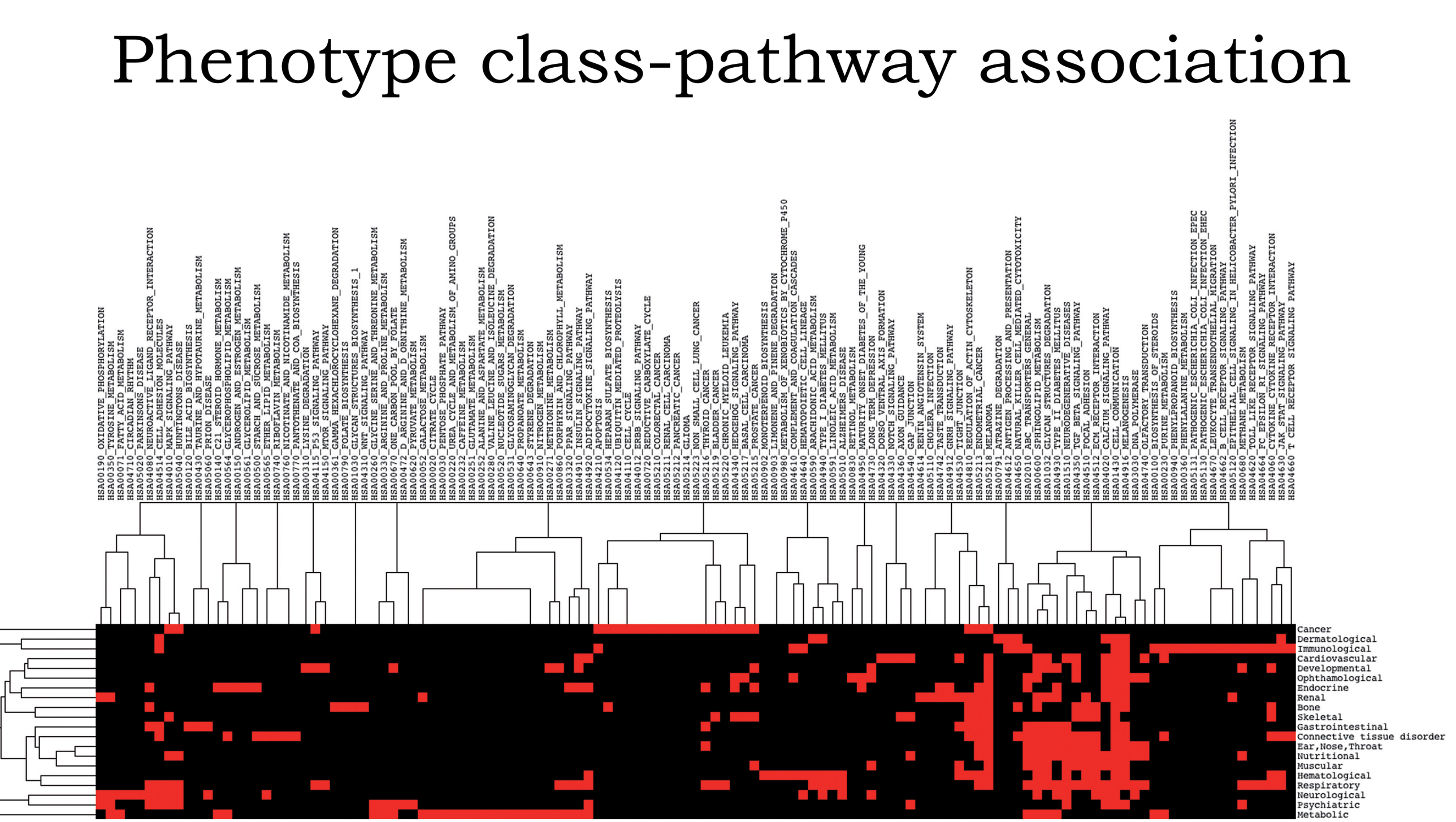
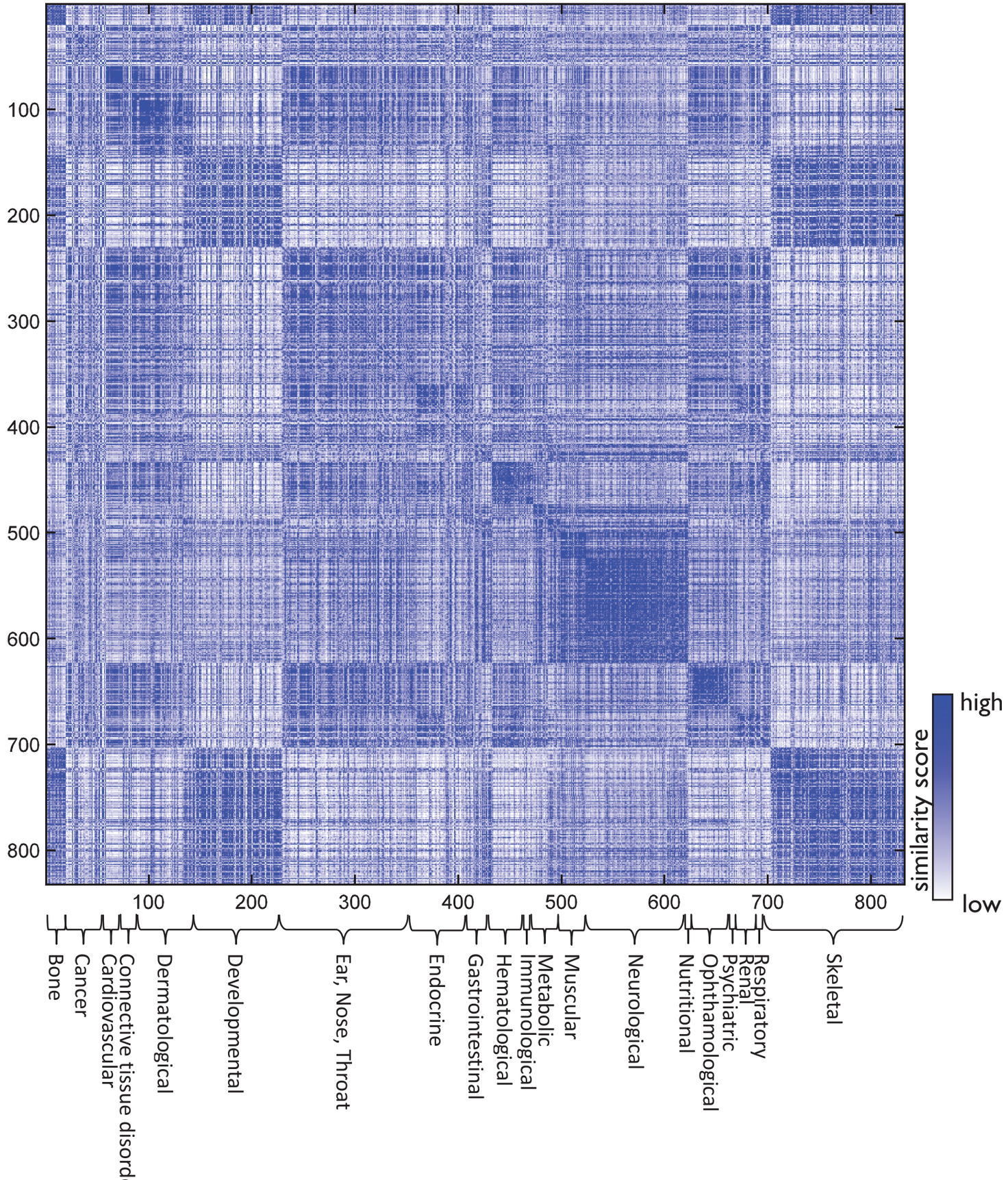


Figure 5. PPI subnetworks of the extended disease pathways. In each pathway, gray nodes are known member genes in the disease pathways and red nodes are newly predicted member genes. Edges represent PPI between two genes. Note that if a known or a newly predicted member gene is not interacting with any other member genes in the pathway, the gene is not included. (**A**) Colorectal cancer pathway. The predicted colorectal cancer genes EXO1 and ADIPOR1 are interacting with many other genes in the colorectal cancer pathway. (**B**) Alzheimer pathway. Over-expression of C1R is known for involving alzheimer disease. (**C**) Melanoma pathway. Mutation and copy number changes in new member gene FGFR3 were recently discovered in melanoma.

Phenotype class-pathway association



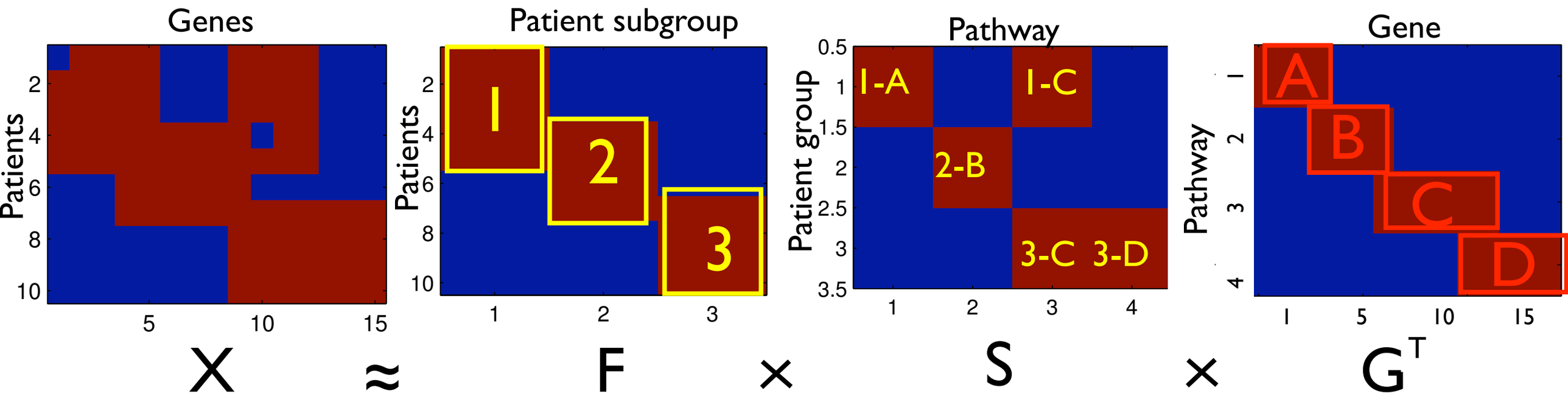
Human Phenotype Ontology



SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., accepted in, Nucleic Acids Research
T. Hwang, In preparation

Matrix tri-factorization

- Given: Gene expression and pathway data
- Task : Identify patient subgroups and pathway activities related with patient subgroups



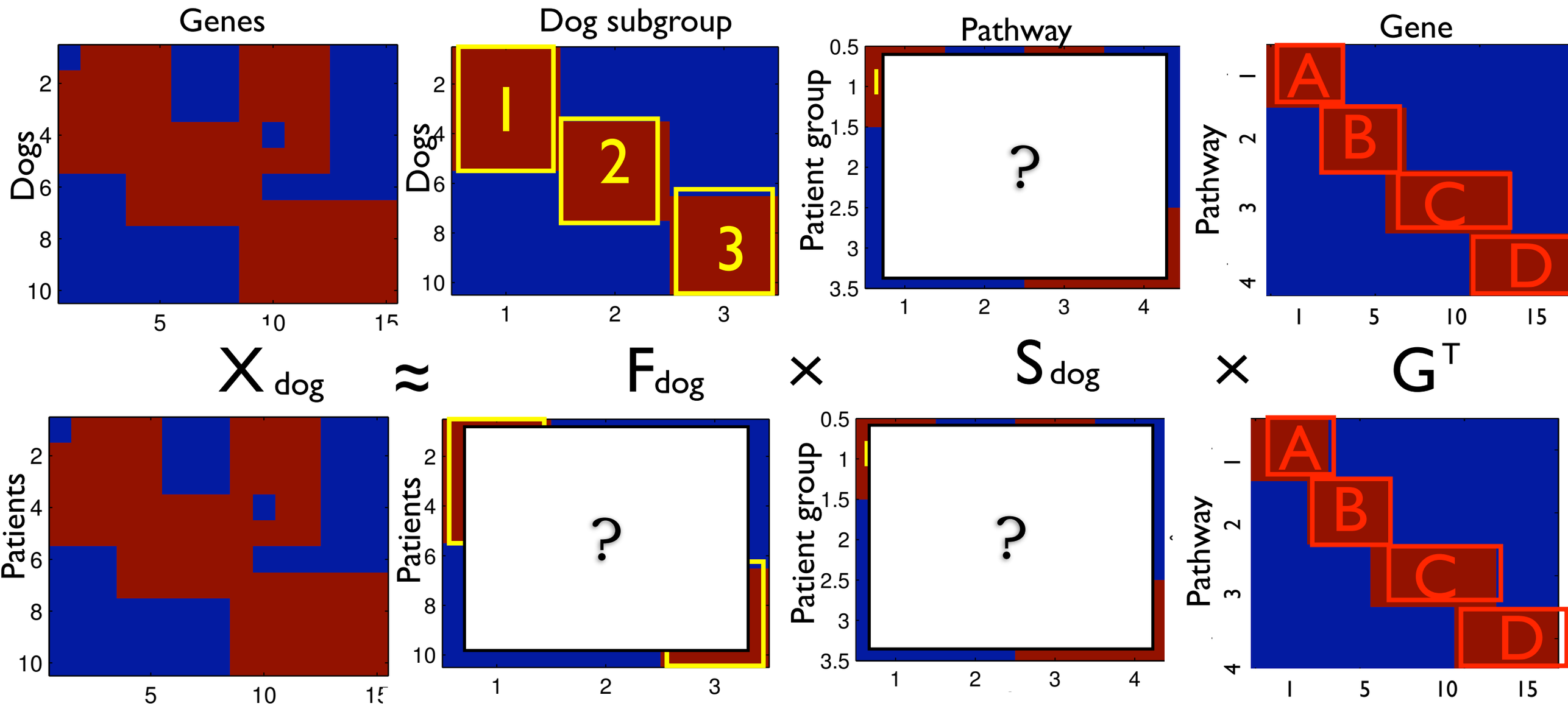
$$\min_{\mathbf{F}\geq 0,\mathbf{S}\geq 0,\mathbf{G}\geq 0} \frac{1}{2}\|\mathbf{X}-\mathbf{FSG}^T\|_F^2 + \lambda_F\|\mathbf{F}\|_1^2 + \lambda_S\|\mathbf{S}\|_1^2 + \lambda_G\|\mathbf{G}\|_1^2$$

X: gene expression data
F: patient subgroups
S: patient subgroup-pathway associati
G: pathway

SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., accepted in, Nucleic Acids Research
T. Hwang, In preparation

Cross-species Matrix tri-factorization

- Given: Dog and human gene expression, pathway data, and dog sugbgroup
- Task : Identify patient subgroups and pathway activities related with patient subgroups in human

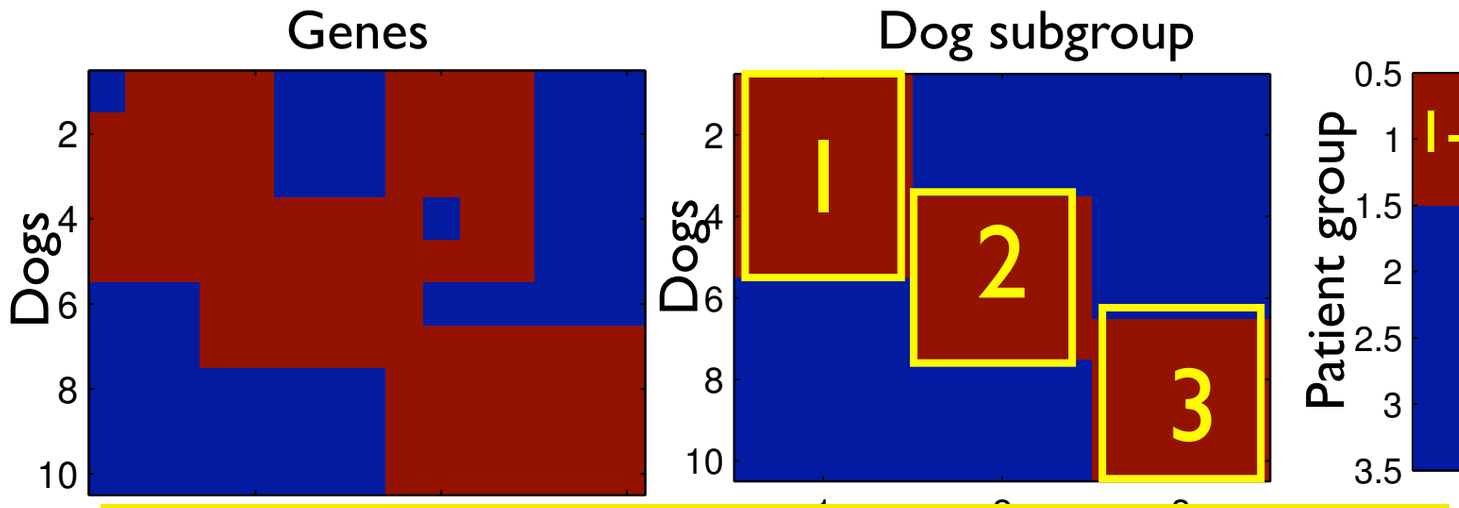


$$\min_{F_{human},\, S_{dog}} \frac{1}{2}\left(\|\mathbf{X}_{dog} - \mathbf{F}_{dog}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2 + \|\mathbf{X}_{human} - \mathbf{F}_{human}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2\right)$$

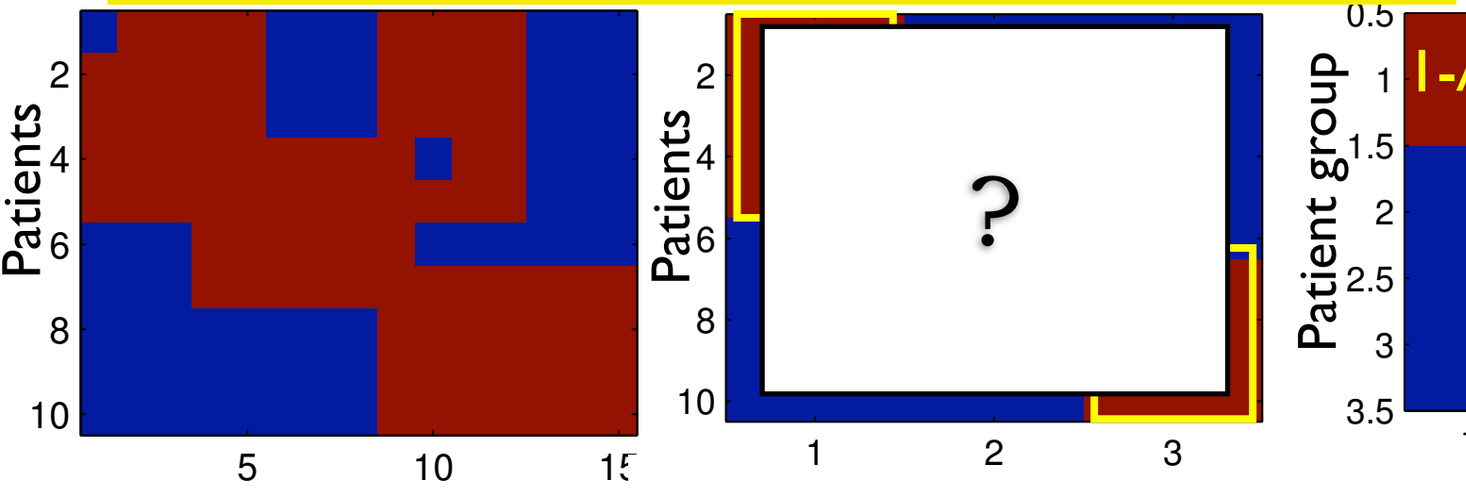
SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., accepted in, Nucleic Acids Research
T. Hwang, In preparation

Cross-species Matrix tri-factorization

- Given: Dog and human gene expression, pathway data, and dog sugbgroup
- Task : Identify patient subgroups and pathway activities related with patient subgroups in human



Use inferred pathway activities from dog to human cancer



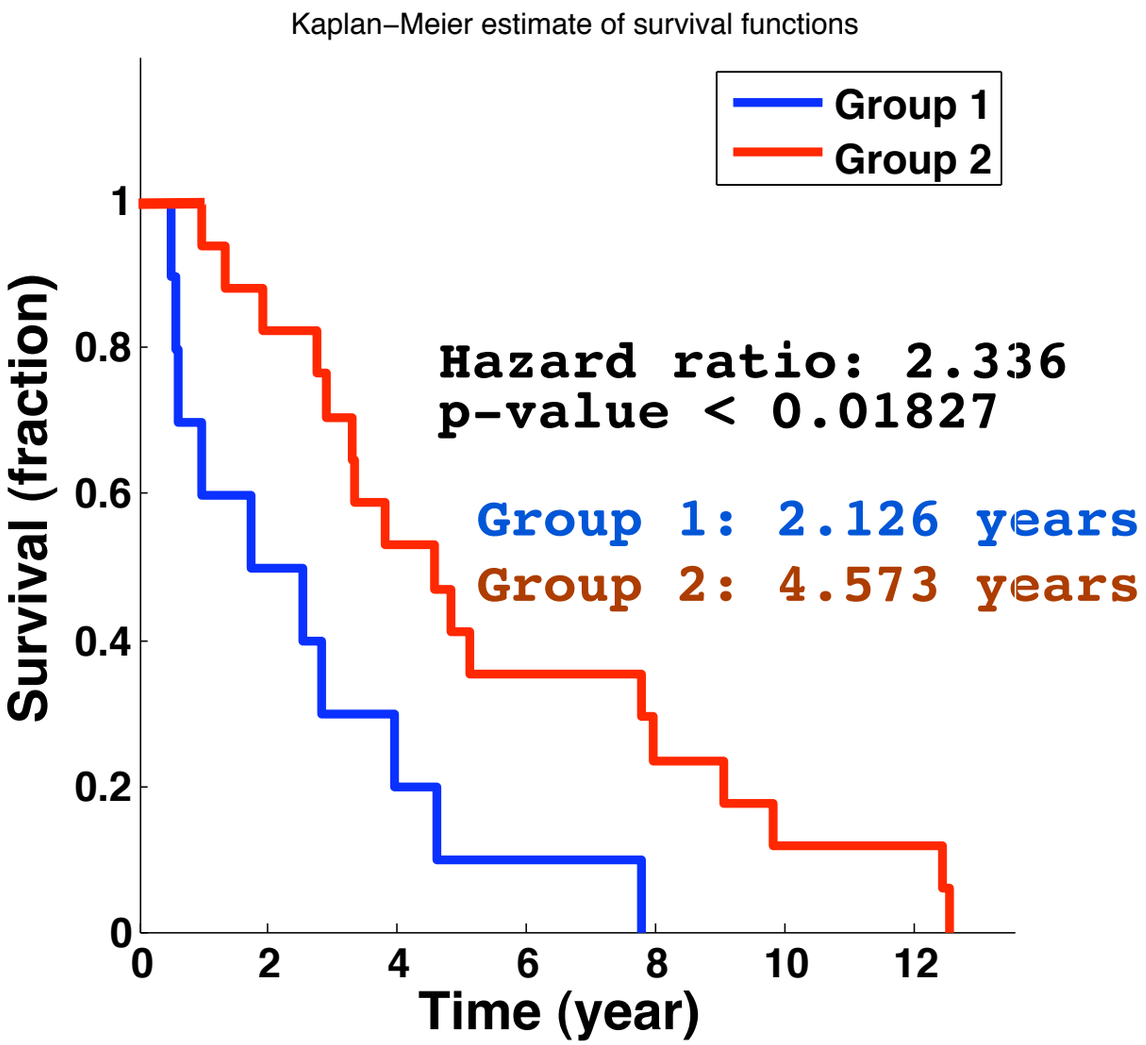
$$\mathbf{X}_{human} \approx \mathbf{F}_{human} \times \mathbf{S}_{dog} \times \mathbf{G}^T$$

$$\min_{\mathbf{F}_{human},\, \mathbf{S}_{dog}} \frac{1}{2}\left(\|\mathbf{X}_{dog} - \mathbf{F}_{dog}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2 + \|\mathbf{X}_{human} - \mathbf{F}_{human}\mathbf{S}_{dog}\mathbf{G}^T\|_F^2\right)$$

SJ Kim, **T. Hwang,** G.B. Giannakis, **CIP** 2012
T. Hwang and et. al., accepted in, Nucleic Acids Research
T. Hwang, In preparation

Experiments (Osteosarcoma)

- Osteosarcoma: 34 dogs (GSE27217) and 34 patients (GSE16091) with clinical data
- Pathway: Reactome pathway (430 pathways)
- 5 (short) vs 12 (long) months for dog subgroup



Ranking	Pathway
1	INFLUENZA LIFE CYCLE
2	**CELL CYCLE CHECKPOINTS**
3	**STABILIZATION OF P53**
4	S PHASE
5	DNA STRAND ELONGATION
6	SCF SKP2 MEDIATED DEGRADATION OF P27 P21
7	CYCLIN E ASSOCIATED EVENTS DURING G1 S TRANSITION
8	**SIGNALING BY NGF**
9	REGULATION OF INSULIN SECRETION BY GLUCAGON LIKE PEPTIDE 1
10	NEURORANSMITTER RECEPTOR BINDING
11	SYNTHESIS OF DNA
12	OPIOID SIGNALLING
13	**SIGNALING BY WNT**
14	ACTIVATION OF NMDA RECEPTOR UPON GLUTAMATE BINDING
15	VIF MEDIATED DEGRADATION OF APOBEC3G