Discrete Structural Statistics for Cancer Science: 2016 Progress

離散構造統計学の創出と癌科学への展開

Koji Tsuda (U. Tokyo)
Kenji Kadomatsu (Nagoya U.)
Ichiro Takeuchi (Nagoya Inst. Tech.)
Jun Sese (AIST)
Ryo Yamada (Kyoto U.)

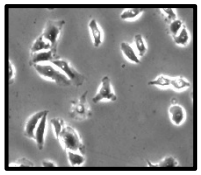
New Statistical Test, Cancer Science

Discrete Mathematics
ZDD, SAT, WISH etc

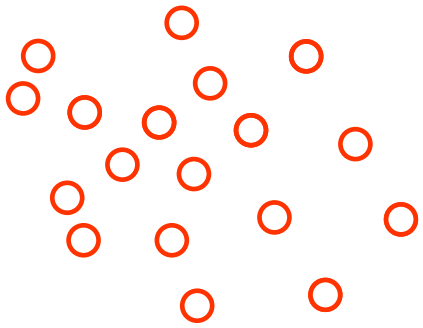
Statistics
FDR,Permutation

Discrete Structural Statistics

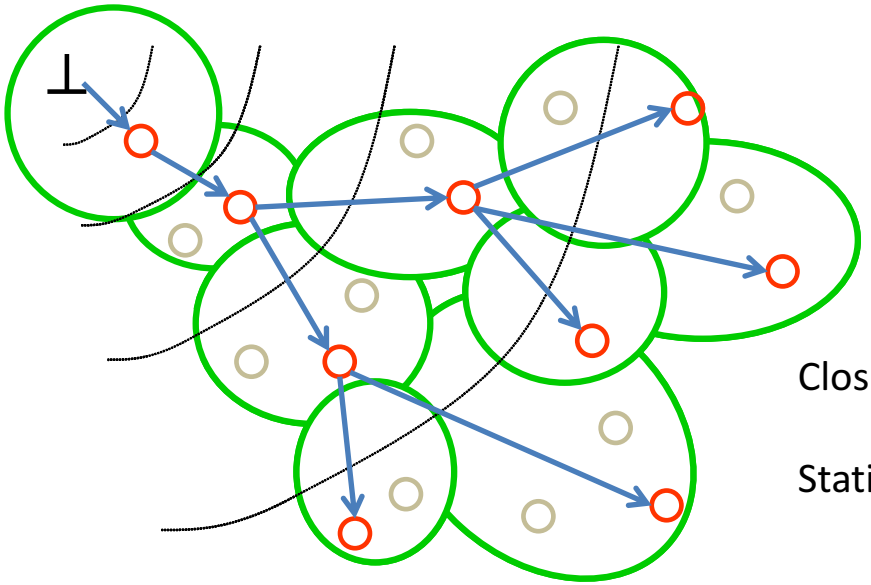
Cancer Science



Hypotheses Set



Structure Introduced



Closure Operator

Statistical Pruning

2016 Highlights

- New statistical principle: Selective inference
  - Suzumura et al., in submission
- Improved efficiency in pattern mining (POSTER)
  - Nakagawa et al., KDD 2016.
- Neuroblastoma Single-Cell RNA-Seq experiments
  - On-going

Testing after selection

- Select features by LASSO and compute p-values
  - Selection bias: p-values are too good!
- Remedy 1: Consider unselected features as well (e.g., LAMP)
- Remedy 2: Condition on the selection event

Selection by LASSO

$$(X, \mathbf{y}) \rightarrow (\mathbf{x}_2, \mathbf{x}_7)$$

Computing p-value

$$\mathrm{corr}(\mathbf{x}_2, \mathbf{y}) = 0.90$$

Compare

$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_1) = 0.10$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_2) = 0.82$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_3) = 0.23$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_4) = 0.30$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_5) = 0.92$$

With permutated outcomes

Selection by LASSO

$$(X, \mathbf{y}) \to (\mathbf{x}_2, \mathbf{x}_7)$$

$$(X, \tilde{\mathbf{y}}_1) \to (\mathbf{x}_1, \mathbf{x}_9)$$
$$(X, \tilde{\mathbf{y}}_2) \to (\mathbf{x}_2, \mathbf{x}_7)$$
$$(X, \tilde{\mathbf{y}}_3) \to (\mathbf{x}_3, \mathbf{x}_4)$$
$$(X, \tilde{\mathbf{y}}_4) \to (\mathbf{x}_8, \mathbf{x}_9)$$
$$(X, \tilde{\mathbf{y}}_5) \to (\mathbf{x}_2, \mathbf{x}_7)$$

Computing p-value

$$\mathrm{corr}(\mathbf{x}_2, \mathbf{y}) = 0.90$$

Compare

$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_1) = 0.10$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_2) = 0.82$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_3) = 0.23$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_4) = 0.30$$
$$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_5) = 0.92$$

With permutated outcomes

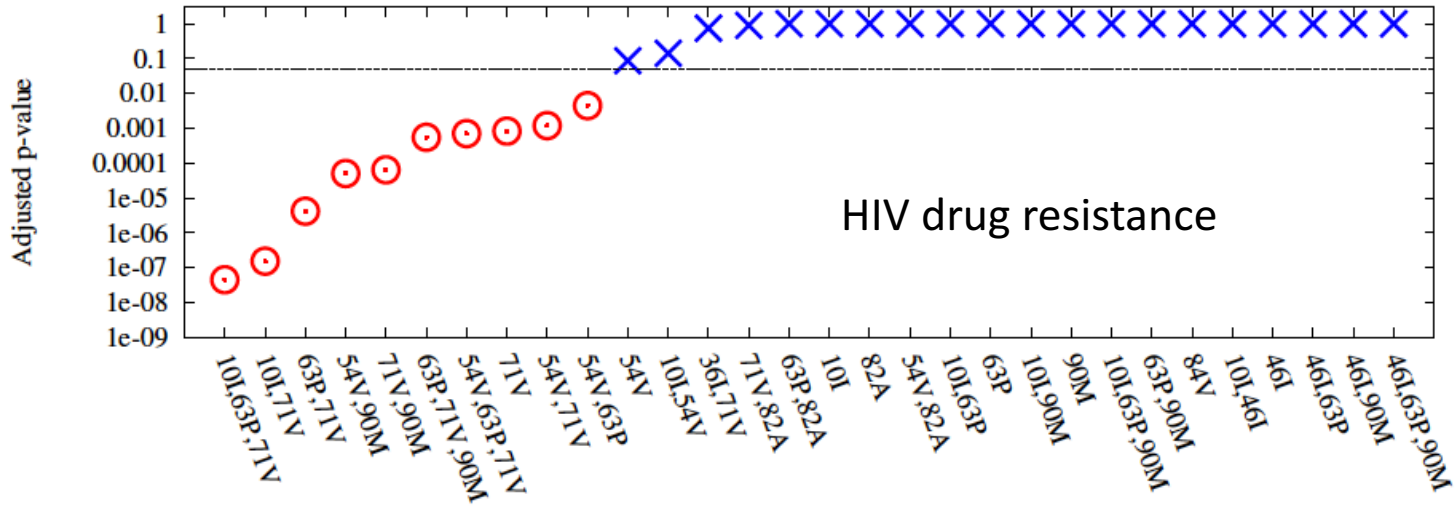
Selection by LASSO	Computing p-value
$(X, \mathbf{y}) \to (\mathbf{x}_2, \mathbf{x}_7)$	$\mathrm{corr}(\mathbf{x}_2, \mathbf{y}) = 0.90$
	Compare
$(X, \tilde{\mathbf{y}}_1) \to (\mathbf{x}_1, \mathbf{x}_9)$	$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_1) = 0.10$
$(X, \tilde{\mathbf{y}}_2) \to (\mathbf{x}_2, \mathbf{x}_7)$	$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_2) = 0.82$
$(X, \tilde{\mathbf{y}}_3) \to (\mathbf{x}_3, \mathbf{x}_4)$	$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_3) = 0.23$
$(X, \tilde{\mathbf{y}}_4) \to (\mathbf{x}_8, \mathbf{x}_9)$	$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_4) = 0.30$
$(X, \tilde{\mathbf{y}}_5) \to (\mathbf{x}_2, \mathbf{x}_7)$	$\mathrm{corr}(\mathbf{x}_2, \tilde{\mathbf{y}}_5) = 0.92$

Use them only !

With permutated outcomes

2016 Highlight 1: Selective inference in pattern mining

- Computation of selective null distribution needs reference to all features
- Impossible in combinatorial cases
- New bound for disregarding large patterns !



2016 Highlight 3:

Single cell RNA-seq for Neuroblastoma

- ～2014: Bulk analysis = Average of many cells
- 2014～: Single cell analysis (10x Genomics)

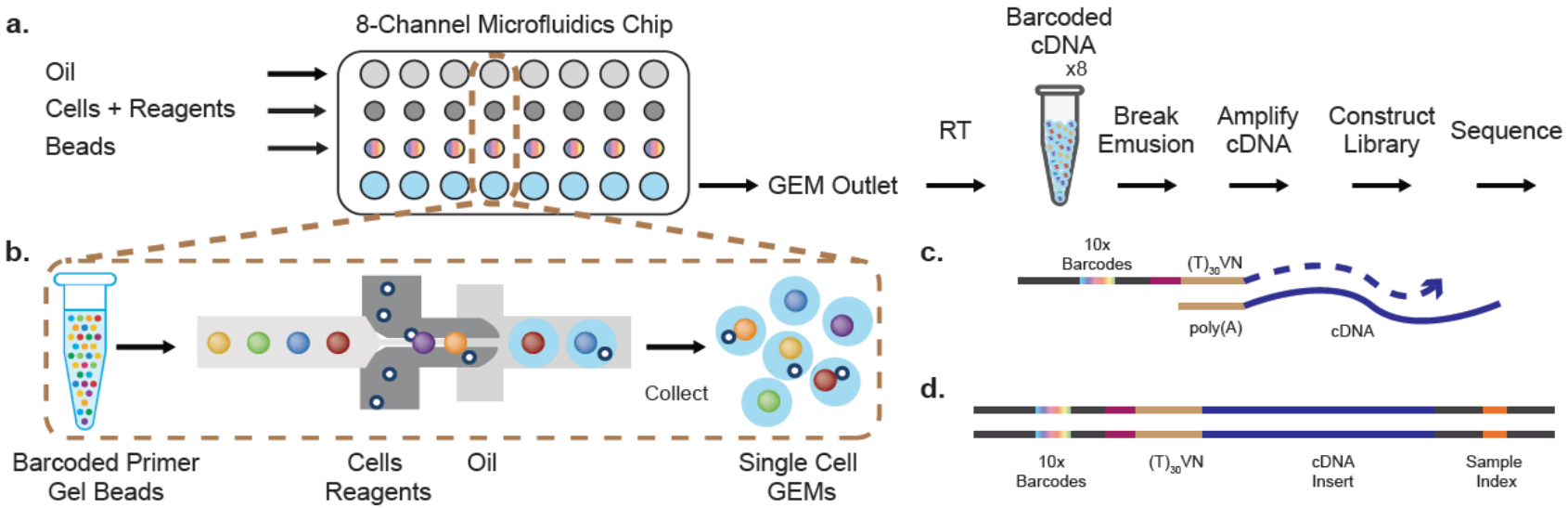
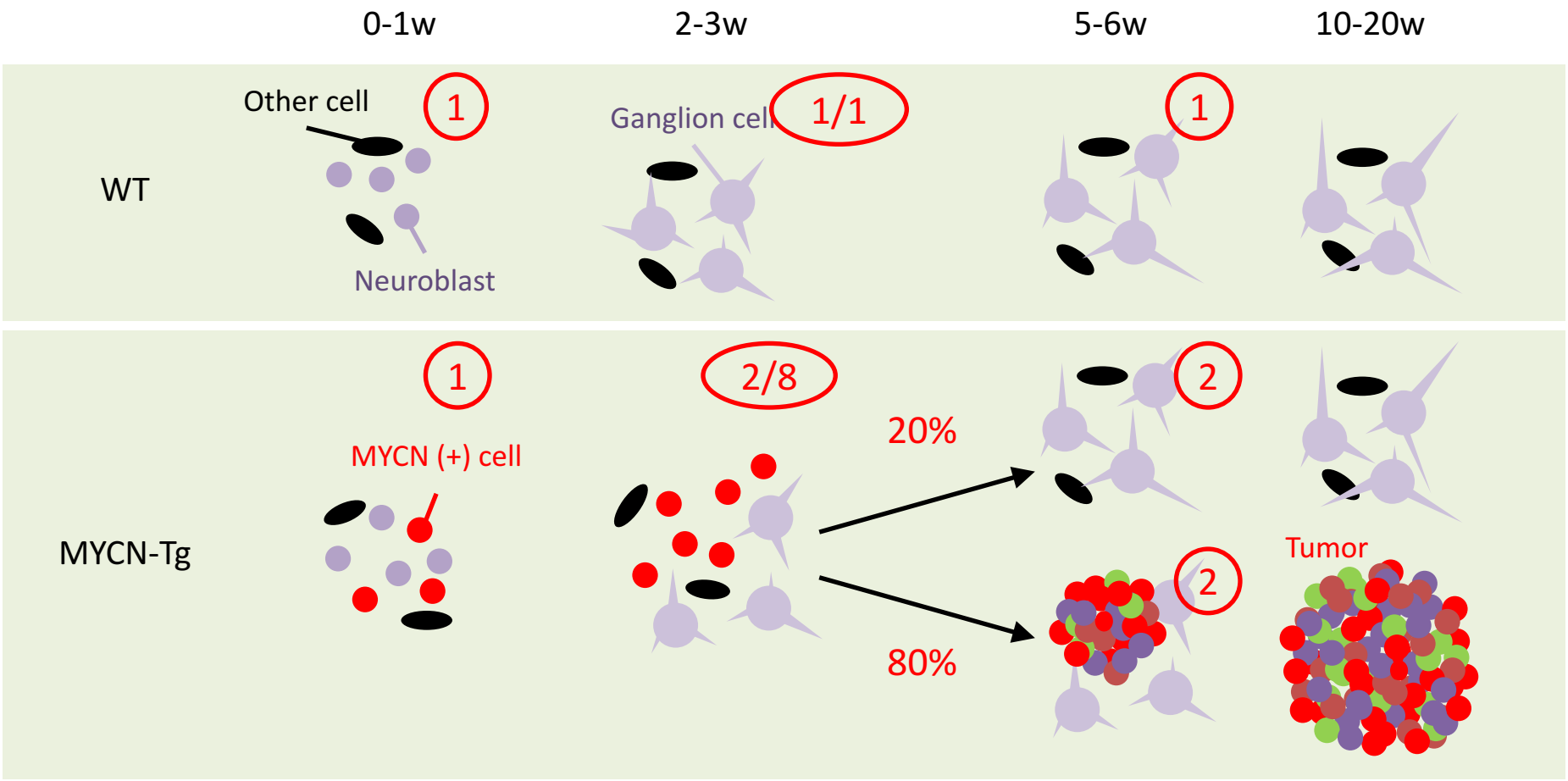
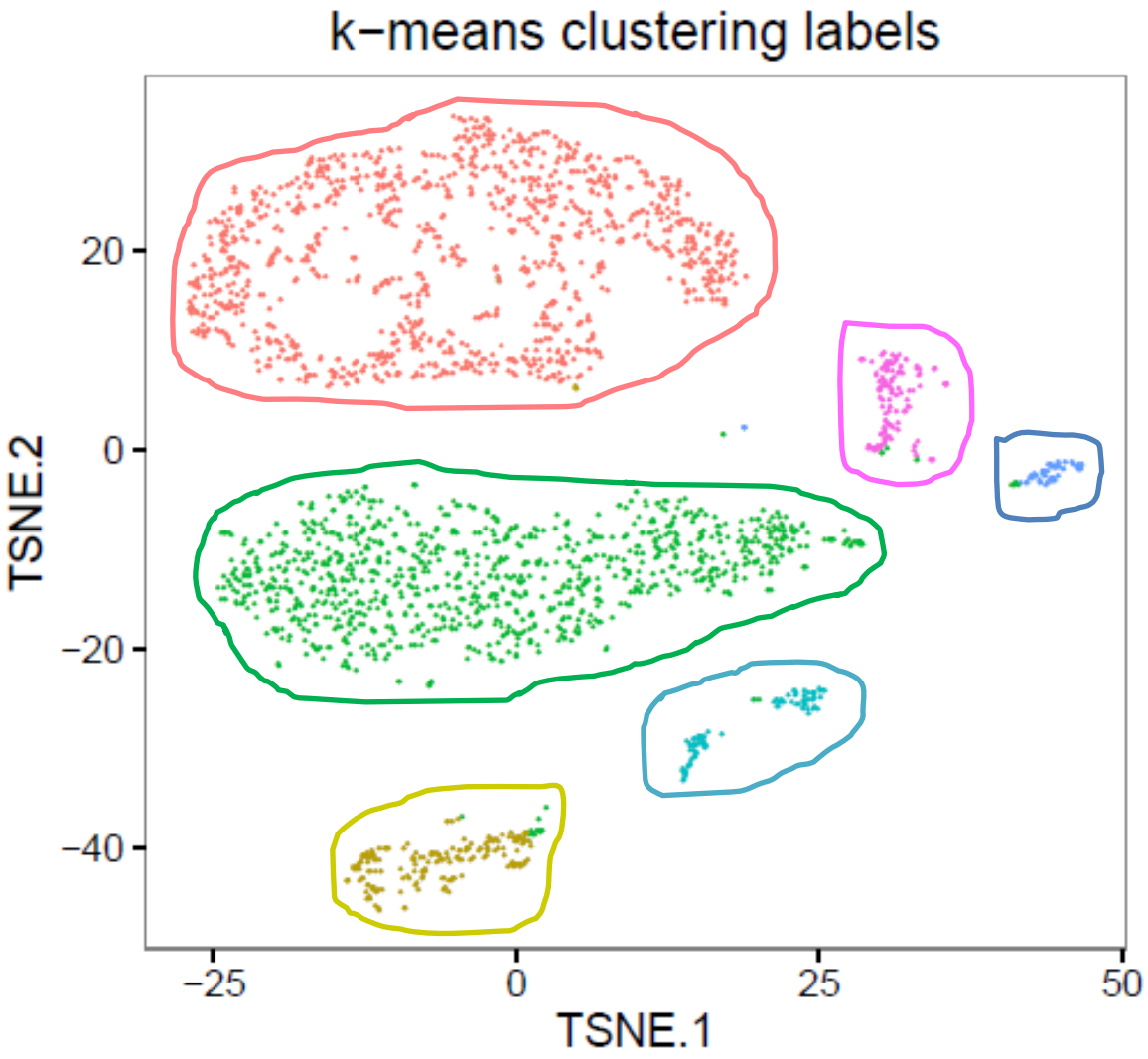


Figure 1. GemCode single cell platform. (a) Formation of GEMs, RT takes place inside each GEM, which is then pooled for cDNA amplification and library construction in bulk. (b) Formation of single-cell GEMs. (c) Barcoded oligonucleotides contained inside GEMs. (d) Final library molecules.

Plan of experiments



Cancer cells detected clearly



Cluster 1: Neuroblastoma cells (MYCN)

Cluster 2: Ganglion cells (Npy, Dbh, Ntrk1)

Cluster 3: Glial cells ? (Dbi, Fabp7, Arpc1b)

Cluster 4: Fibroblasts ? (Dcn, Col3a1, Lum, Igfbp6, Acta2)

Cluster 5: Myeloid cells, probably macrophages (Lyz2, C1qa, C1qb, C1qc, Ftl1)

Cluster 6: Endothelial cells ? (Egfl7, Id3, Plvap, Esam, Cldn5)

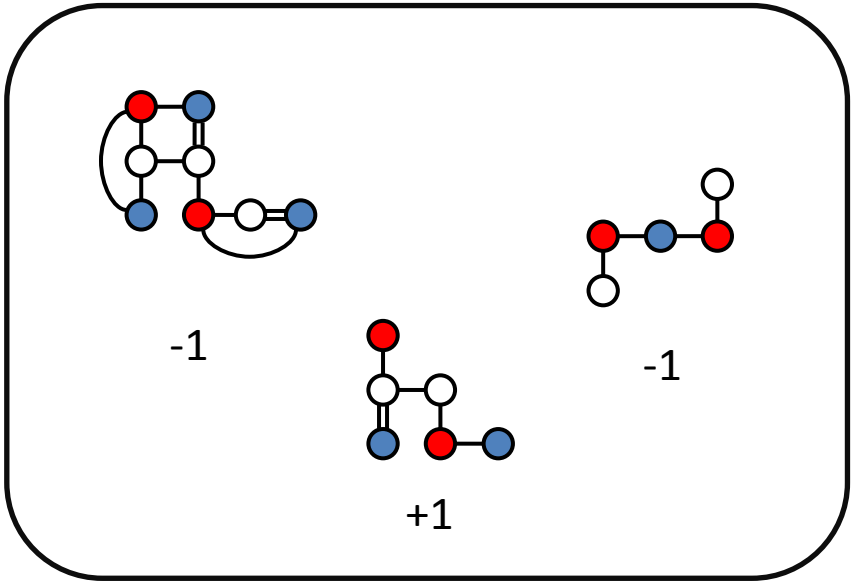
Summary

- New directions emerged in 2016
- Biological experiments going well
- 2017～: New discoveries in cancer science

- Awareness: Invited session in MCP 2017 !

Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining

中川和也(名工大), 鈴村真矢(名工大), 烏山昌幸(名工大), 津田宏治(東大), 竹内一郎(名工大)



データベース

- KDD2016 採択論文
- グラフの性質を特徴づける部分構造を見つけたい
- スパース学習を行う

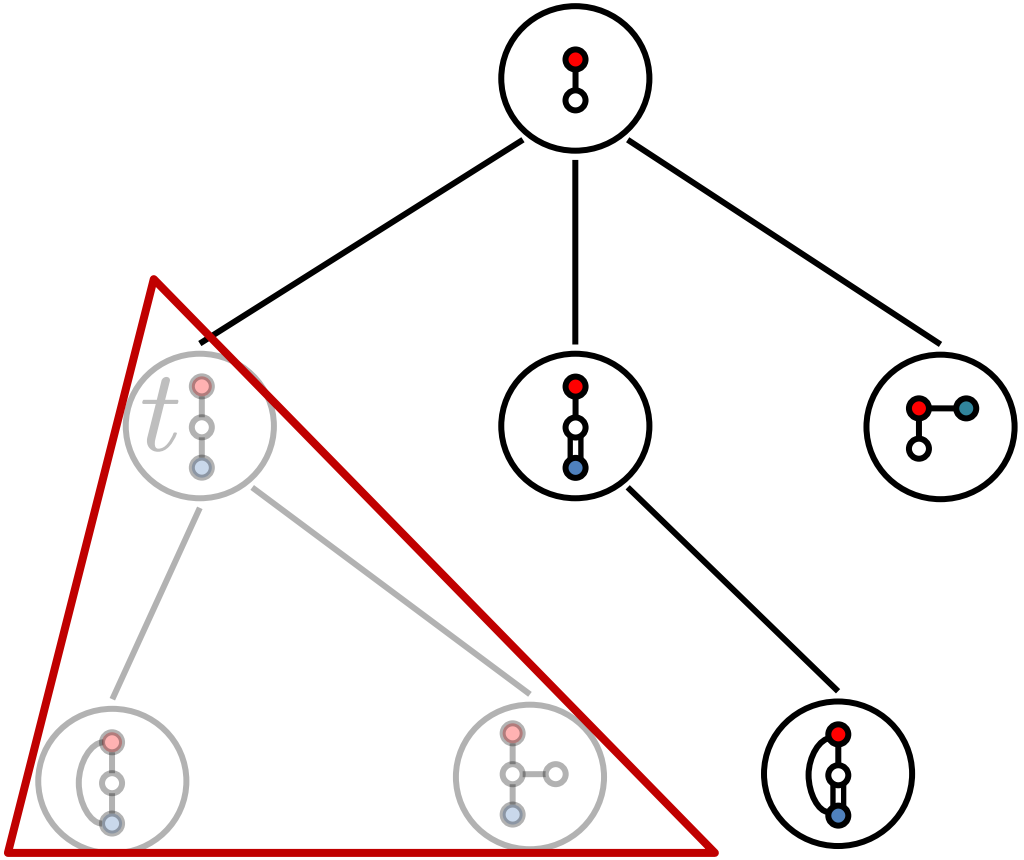
$$f = \; w_1 + \; w_2 + \; w_3 + \; w_4 + \; \cdots$$

Safe Pattern Pruning

$$Score\ (t) < 1$$

$$\Downarrow$$

$$w^*_{t'} = 0$$

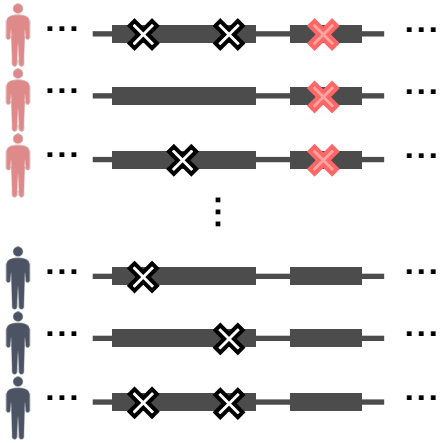


- 最適解においてスパースとなる部分木を同定・プルーニングできる
- 各ノードにおけるscore計算には最適解を必要としない (近似解を使う)

LAMPLINK: detection of statistically significant SNP combinations from GWAS data

発表者: 寺田愛花（JST さきがけ/東京大学）

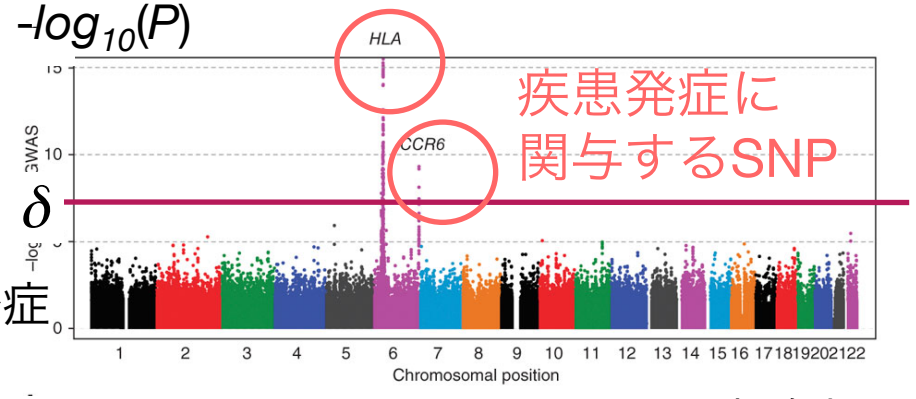
がん患者



健常者

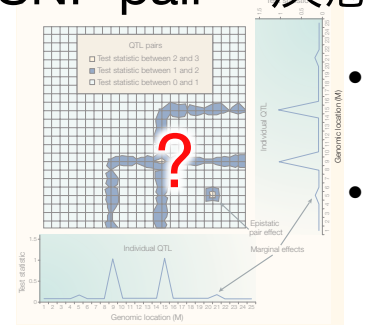
SNP単体 – 疾患発症

SNP: 数万〜数百万カ所



Kochi et al., Nat. Genet., 2010 を一部改変

SNP pair – 疾患発症

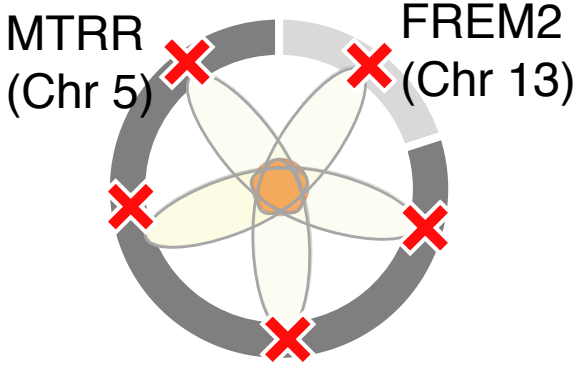


- FastANOVA (Zhang, X., et al. 2008)
- PLINK (Purcell, S., et al. 2007)

Carlborg & Haley, Nat. Rev. Genet. (2004).

LAMP

3個以上のSNP – 疾患発症



Terada, A et al. Bioinformatics 32(22), 3513–3515