A Bi-ordering Approach to Linking Gene Expressions with Clinical Annotations in Cancer

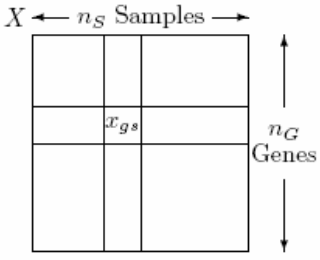
Fan Shi[1], Geoff MacIntyre[1], Christopher Leckie[1],
Izhak Haviv[2], Alex Boussioutas[3], Adam Kowalczyk[1]

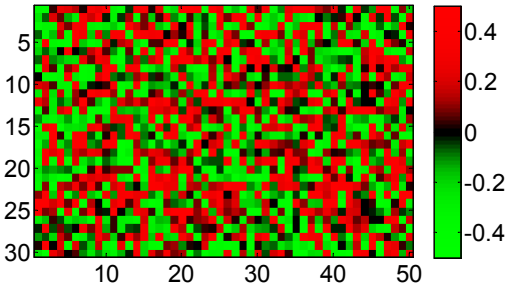
1 National ICT Australia, VRL & The University of Melbourne
2 Baker IDI Australia
3 Peter MacCallum Cancer Centre, Australia

Background: Gene expressions

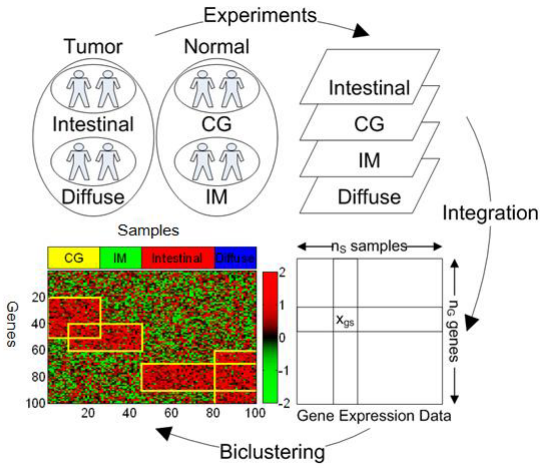
Gene expression microarrays

The associations between genes and samples discovered in gene expression microarrays can be used to help the diagnosis of cancers





Background: Motivation



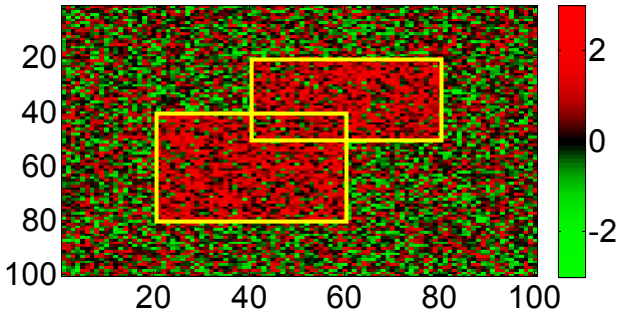
Introduction to bi-clustering

Bi-clustering

An unsupervised method that clusters genes and samples simultaneously

Bicluster —— a subset of genes co-expressed across a subset of samples

Two overlapping biclusters



Bi-ordering Analysis: Overview

Our approach is based on a protocol for exploring biclusters that exhibit statistical, biological and clinical significance

Given an input gene expression data matrix

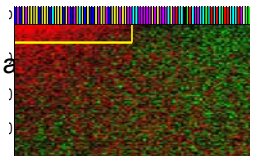
1. Generate biclusters based on Bi-Ordering Approach (*BOA*)
2. Merge similar biclusters into "super-biclusters" to identify robust modules
3. Evaluation: three statistics as measurements
   - Over-representation of histological categories in biclusters
   - Gene Ontology (GO) annotations
   - Concordance of sample order with various phenotype gradients
4. Biological Interpretation

Bi-ordering Analysis: Algorithm

Input Gene expression matrix $[x_{gs}]$ with n_G genes and n_S samples
Standardize data and select **pre-defined thresholds** θ_G and θ_S
Randomly select a submatrix (G, S) as initial bicluster
Repeat

1. Update gene scores $f[g] \leftarrow \langle x_{gs} \rangle_{s \in S}$, for $g = 1, ..., n_G$
2. Select a subset of genes $G \leftarrow \left\{ g \; ; \; f[g] - \langle f[g] \rangle_{g=1,...,n_G} > \theta_G \Big/ \sqrt{|S|} \right\}$
3. Update sample scores $h[s] \leftarrow \langle x_{gs} \rangle_{g \in G}$, for $s = 1, ..., n_S$
4. Select a subset of samples $S \leftarrow \left\{ s \; ; \; h[s] - \langle h[s] \rangle_{s=1,...,n_S} > \theta_S \Big/ \sqrt{|G|} \right\}$

Until G and S are stable
(G, S) is a bicluster with ordering $f(g)$ $h(s)$ a
*Note: $\langle x_{gs} \rangle_{s \in S}$ denotes the mean of $\{x_{gs} \; ; \; s \in S\}$



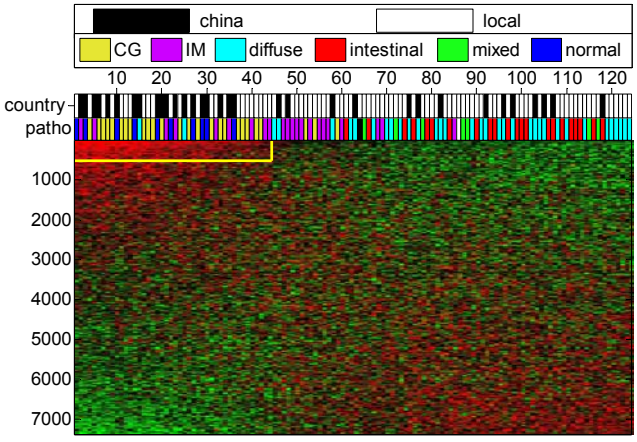
Bi-ordering Analysis: Super-biclustering

- Multiple biclusters
  - Distinct initializations in BOA result in different biclusters in general
  - Some of the biclusters differ slightly due to local optima
- Super-biclustering
  A hierarchical clustering is applied on 'biclusters' to obtain a group of super-biclusters (SBC)
  - Objects: biclusters
  - Distance metric: Jaccard coefficient on genes
  - Prototype selection: a single bicluster is selected as the prototype of SBC

Finally, we obtain a small number of distinct biclusters as output

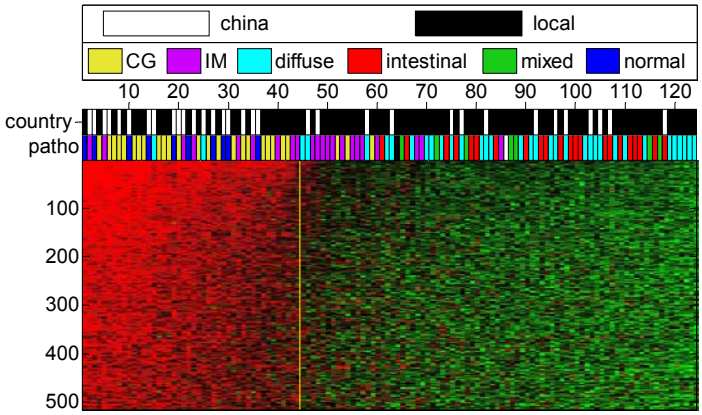
Bi-ordering Analysis

Heat map of a typical bicluster in gastric cancer



Bi-ordering Analysis

Heat map of a typical bicluster in gastric cancer



Experiments: Gastric cancer

Gastric Cancer dataset

- 7383 genes, 124 samples
- 6 pathological categories, and other clinical annotations

Phenotype	Subtype	Malignant score *
Pre-malignant	Normal	1
	Chronic Gastritis (CG)	2
	Intestinal Metaplasia (IM)	3
Malignant	Diffused	4
	Intestinal	4
	Mixed	4

* Malignant score reflects the biological progression of gastric cancer
It is defined by oncologist

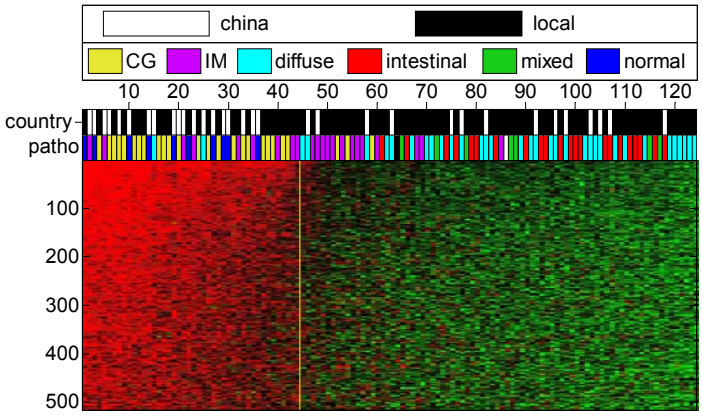
Experiments: Evaluation metrics

Three evaluation metrics

- Saturation metric on samples
  —— Homogeneity of samples in terms of clinical annotations
- Trend statistics
  —— Associations between the sample orders and clinical annotations
- Gene Ontology (GO)
  —— The abundance of genes for particular pathways

Experiments: Results

Heat map of SBC7



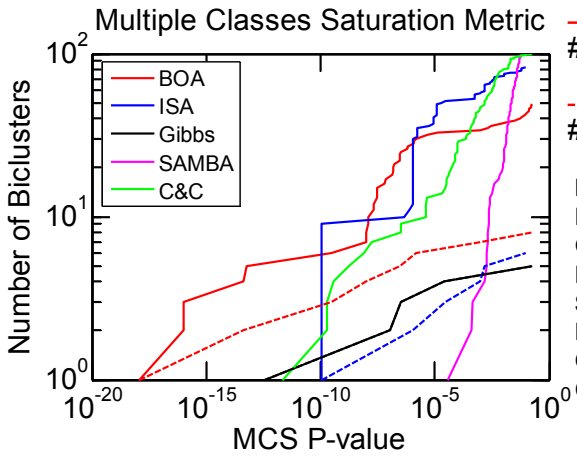
Experiments: Results

The Significance Table of Super-biclusters

		p-value			Most significant annotation
SBC	Converge	MCS	Malignancy Score	GO	GO
SBC1	11	9.4E-04	1.8E-13	5.1E-09	epidermis development
SBC2	188	1.0E-08		7.1E-07	lipid metabolic process
SBC3	2	1.5E-06	5.5E-08	3.2E-32	immune system process
SBC4	96	1.8E-01		2.0E-53	immune system process
SBC5	15	1.1E-18	7.7E-21	1.8E-14	cell cycle process
SBC6	328	3.0E-07	4.9E-08	1.8E-20	multicellular organismal process
SBC7	359	4.0E-14	-5.4E-22	3.2E-22	gen. of precursor metab. & energy
SBC8	1	3.0E-10	-5.2E-08	2.2E-02	lipid metabolic process

Experiments: Saturation metric

Comparison with existing algorithms in terms of saturation metric



ISA: J. Ihmels et al., Bioinformatics, 2004
Gibbs: Q. Sheng, et al., Bioinformatics, 2003
SAMBA: A. Taney et al., Bioinformatics, 2002
C&C: Y. Cheng, G. M. Church, ISMB, 2000

Experiments: Biological interpretation

Compare with “Distinctive patterns of gene expression of premalignant gastric mucosa and gastric cancer”, Alex Boussioutas et al., 2003, Cancer Research.

Region in [Bou03]			SBC_1	SBC_2	SBC_3	SBC_4	SBC_5	SBC_6	SBC_7	SBC_8
Symbol	Annotation	No.Genes	41	217	194	158	227	409	515	146
B	Mitochondrial	665	0	0	0	0	0	1	**416**	9
D1-D3	Proliferation	201	0	0	0	0	**76**	0	0	0
E	Intestinal	294	1	**81**	0	0	0	0	1	44
F	Intestinal	157	0	**112**	0	0	7	1	0	27
G	Squamous	37	**25**	0	0	0	0	0	0	0
H	Inflamation	330	7	0	**117**	**135**	9	7	0	30
K	Extracellular	877	3	0	**67**	0	**74**	**392**	1	0

B: Encoding mitochondrial proteins (CG)

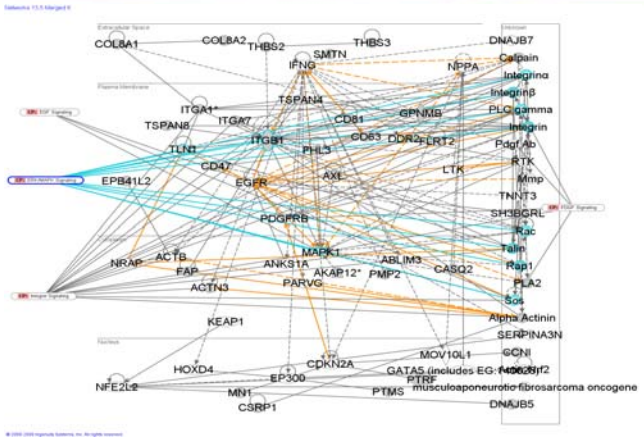
D1-3: Cell proliferation (Intestinal GC)

E&F: Intestinal genes (IM)

H: Inflammation

K: Extra cellular matrix (Diffuse GC)

Experiments: Biological interpretation



Signal transduction pathways predicted to operate in SBC_6 signature-expressing cells (via

analysis in Ingenuity Pathway Analysis®). Notice the genes PDGFRB, EGFR, DDR2, AXL and LTK

are all tyrosine kinase receptors. Also noteworthy is ITGB1. The integrin and tyrosine kinase

Conclusion

- **Developed Bi-Ordering Analysis for gene expression data**
  - An effective method to find significant patterns and orderings
  - Scalability: very efficient on large scale gene expression data
- **Evaluation of statistical significance**
  - SCS and MCS for validating the homogeneity of samples
  - Gene Ontology for validating the significance of gene modules
  - Trend statistics for validating the concordance between sample ordering and clinical annotations
- **Biological findings**
  - The results of gastric cancer are in concordance with previous studies
  - Some novel findings deserve further investigation

Future Work

- Super-biclustering
  How to choose an appropriate distance metric in super-biclustering to detect a robust set of biclusters is an important issue in computational point of view

- Data dependency
  Verify our approach in more synthetic and real datasets are critical to solve the problem of various performance dependent on datasets


NICTA
Thanks
Questions?
19