A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer

Binghui Liu, Chong Wu, Xiaotong Shen, Wei Pan

University of Minnesota, Minneapolis, MN 55455

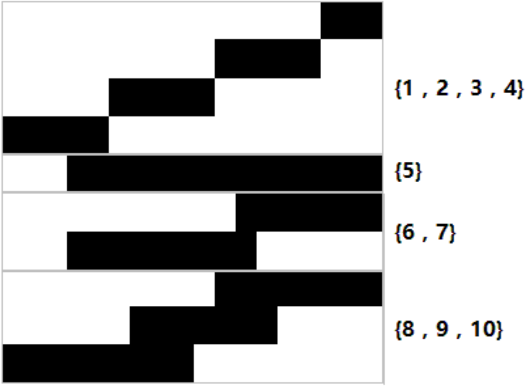
Nov 2017

Introduction

- Given: an $n \times p$ mutation matrix $\mathbf{A}$ with entry $A_{ij} = 1$ if gene j is mutated in patient i, and $A_{ij} = 0$ otherwise.
- Goal: to identify a subset B of genes as **driver** gene.
- Vandin et al (2012, *Genome Res*) proposed two criteria:
    - Coverage: many patients with mutations in B, maximize $|\Gamma(B)|$ with $\Gamma(B) = \bigcup_{j \in B} \Gamma(j)$ and $\Gamma(j) = \{i : A_{ij} = 1\}$.
    - Exclusitivity: mutations in B do not occur simultaneously on any patient, minimize $\omega(B) = \sum_{j \in B} |\Gamma(j)| - |\Gamma(B)|$.
- Overall, minimize

$$f(B) = \frac{\omega(B)}{n} - \frac{|\Gamma(B)|}{n} = \frac{1}{n} \sum_{j \in B} |\Gamma(j)| - \frac{2}{n} |\Gamma(B)|. \qquad (1)$$

- Challenge: a combinatorial (i.e. NP-hard) problem!
  not feasible to have an exact solution.
  use approximate solutions, e.g. Monte Carlo methods...



B_0	$f(B_0)$
{1,2,3,4}	−1.00
{5}	−0.85
{6,7}	−0.80
{8,9,10}	−0.70

A toy example, where each column represents a patient, each row represents a gene and each black entry represents the mutation.

Existing approaches

- Dendrix-MCMC: Vandin et al (2012, *Genome Res*);
- Multi-dendrix-MCMC: Leiserson et al (2013, *PLOS Comp Biol*);
- Binary linear programming (BLP) and genetic algorithm (GA): Zhao et al (2012, *Bioinformatics*);

New formulation

- Define: $B = B(\boldsymbol{\beta}) = \{j \in V : |\beta_j| \neq 0\}$.
- Rewrite (1) as

$$f(B(\boldsymbol{\beta})) = \frac{1}{n}\sum_{j=1}^{p} I(|\beta_j| \neq 0)A_{\cdot,j} - \frac{2}{n}\sum_{i=1}^{n} I\Big(\sum_{j=1}^{p} A_{ij} I(|\beta_j| \neq 0) \neq 0\Big). \quad ($$

- Challenge: discontinuous indicator function $I(\beta_j \neq 0)$.
- Our solution: use a TLP to approximate $I(.)$:

$$\min(|\beta_j|/\tau_1, 1) \to I(|\beta_j| \neq 0) \qquad \text{as } \tau_1 \to 0^+.$$

- Note: TLP as a non-convex penalty better than the popular Lasso in regularization, e.g. regression/classification, Gaussian graphical models (GGMs) (Shen et al 2012, *JASA*); fusion: network-based regression (Kim et al 2013, *Biometrics*; Zhu et al 2013, *JASA*), multiple GGMs (Gao et al 2016, *EJS*), ...

$$\tau_1 \min(|\beta_j|/\tau_1, 1) = \min(|\beta_j|, \tau_1).$$

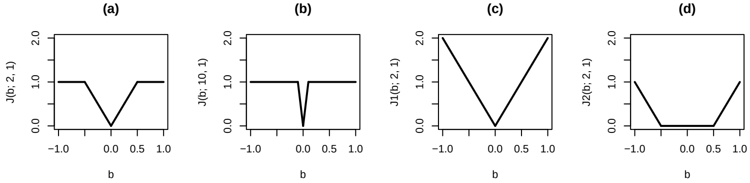


Figure: TLP $J_T(b;\tau)$ (a) with $\tau = 0.5$, or (b) with $\tau = 0.1$; J_T in (a) is decomposed into a difference of two convex functions J_1 in (c) and J_2 in (d).

Computation: DC algorithm

- New target:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{j=1}^{p}\min\left(\beta_j/\tau_1,1\right)A_{\cdot,j} - \frac{2}{n}\sum_{i=1}^{n}\min\left(\sum_{j=1}^{p}A_{ij}\beta_j/\tau_1,1\right) \\ &\quad +\lambda\sum_{j=1}^{p}\min\left(\beta_j/\tau_2,1\right)+\frac{\alpha}{n}\sum_{j=1}^{p}\beta_j^2, \qquad (3) \end{aligned}$$

  with respect to $\boldsymbol{\beta}=(\beta_1,...,\beta_p)'\in[0,+\infty)^p$.

- Note: The TLP and the ridge penalty ensure sparse and proper solutions.
- DC decomposition of TLP:

$$\min(\frac{|z|}{\tau},1)=\frac{|z|}{\tau}-\max\left(\frac{|z|}{\tau}-1,0\right).$$

- At iteration m with the current estimate,

$$
\begin{aligned}
S^{(m)}(\boldsymbol{\beta}) = \quad & \boldsymbol{\beta}' \left(\mathrm{diag}(\mathbf{A}_{\cdot}) I(\hat{\boldsymbol{\beta}}^{(m-1)} \leq \tau_1)/n\tau_1 + \lambda I(\hat{\boldsymbol{\beta}}^{(m-1)} \leq \tau_2)/\tau_2 - 2\mathbf{A}_{\cdot}/n\tau_1 \right) + \\
& \frac{2}{n} \sum_{i=1}^{n} \max\Big(\sum_{j=1}^{p} A_{ij}\beta_j/\tau_1 - 1, 0\Big) + \frac{\alpha}{n}\boldsymbol{\beta}'\boldsymbol{\beta}, \qquad (4)
\end{aligned}
$$

- $S^{(m)}(\boldsymbol{\beta})$ is convex; can apply Matlab CVX or subgradient descent.
- Theorem: The above DC algorithm converges to a local minimizer montonically in finite steps.
- Tuning parameter selection: $\alpha = 0.001$, $\tau_1 = 1$; others by CV.

Table: Applied to the mutation data of glioblastoma multiforme (TCGA 2008),the new method MCSS identified multiple sets of low-cost mutated genes, grouped in terms of associated pathways.

Pathway	Core mutations	$\hat{B}$	$f(\hat{B})$
p53 signalling	*CDKN2A, MDM2, MDM4, TP53*	(*CDKN2A, MDM2, MDM4, TP53*)	-55
		(*CDKN2A, DTX3, TP53*)	-57
		(*CDKN2A, TP53*)	-53
		(*CDKN2B, TP53*)	-53
RB signalling	*CDKN2A/B, CDK4, RB1*	(*CDKN2B, CYP27B1, RB1*)	-62
		(*CDKN2B, ERBB2, RB1, TSPAN31*)	-64
		(*CDKN2A, CYP27B1, RB1*)	-56
		(*CDKN2B, CYP27B1, NF1*)	-56
		(*CDKN2A, CYP27B1, NF1*)	-54
		(*CDKN2B, CYP27B1*)	-54
RAS signalling	*EGFR, NF1*	(*EGFR, KDR, NF1*)	-52
		(*MTAP, TP53, TSFM*)	-56
		(*CYP27B1, MTAP, PTEN*)	-55
		(*CDK4, MTAP, PTEN*)	-55
		(*EGFR, TP53*)	-52

i

Table: Results in Simulation I based on 100 simulation replications with $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$. The sample means (SD in parentheses) of correct (C) or incorrect (IC) numbers of non-zero estimates, average differences of the cost (ADC) between the true gene subset $B_0 = \{1, 2, 3, 4\}$ and the estimated subset $\hat{B}$, that is, $\frac{f(B_0)-f(\hat{B})}{n}$, and the running time (RT) (in minutes) of the algorithms.

n	p	Method	C	IC	ADC	$\hat{c}_1$ $[c_1]$	$\hat{c}_2$ $[c_2]$	R
50	1000	MCSS	4 (0)	0 (0)	0 (0)	.95 [.95]	.01 [.00]	.22 (
		Dendrix	3.80 (.41)	.50 (.94)	-.02 (.04)	.94 [.95]	.01 [.00]	16.89 (
		Mdendrix	3.90 (.30)	.15 (.36)	-.01 (.03)	.95 [.95]	.01 [.00]	.81 (
		BLP	3.39 (.86)	3.82 (2.59)	.05 (.03)	.99 [.95]	.00 [.00]	.01 (
		GA	3.90 (.38)	2.13 (1.69)	.04 (.03)	.98 [.95]	.01 [.00]	2.97 (
50	10000	MCSS	4 (0)	0 (0)	0 (0)	.95 [.95]	.01 [.01]	1.67 (
		Dendrix	1.25 (1.02)	5.96 (3.62)	-.24 (.04)	.83 [.95]	.02 [.01]	67.06 (
		Mdendrix	1.45 (1.23)	5.25 (4.02)	-.24 (.03)	.83 [.95]	.03 [.00]	1.88 (
		BLP	3.42 (.91)	2.72 (2.06)	.05 (.03)	.99 [.95]	.01 [.00]	.27 (
		GA	3.93 (.25)	1.50 (.92)	.04 (.02)	.98 [.95]	.01 [.00]	284.92 (

Discussion

- A general method to approximate the indicator function; more applications (and modifications) may be worthwhile.
- Considered integrative analysis of mutation and gene expression data.
  improved performance with GE data;
  non-linear; binary linear programming (BLP) not applicable.
- TLP is a general non-convex penalty applied to many high-dimensional problems.
  can be regarded as a refined version of Lasso; TLP=Lasso with larger τ;
  better empirical and theoretical performance;
  computationally more demanding (with an extra tuning parameter τ).

References

1. Vandin, F., Upfal, E. and Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Research* **22**, 375–385.
2. Leiserson, M. D. M., Blokh, D., Sharan, R., Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* **9**, e1003054.
3. Zhao, J., Zhang, S., Wu, L., Zhang, X. (2012). Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28**, 2940–2947.
4. Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Am. Statist. Assoc.* **107**, 223–232.
5. Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075.
6. The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature* **455**, 1061–1068.

Acknowledgement

- This research was supported by NIH.

- **Thank you!**