Detecting presence of mutational signatures in cancer with confidence

Supplementary materials

Xiaoqing Huang, Damian Wojtowicz, and Teresa M. Przytycka

National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda MD 20894, USA

Quadratic Programming (QP) Taking advantage of the COSMIC mutational signature matrix $\mathbf{P}$, as columns of $\mathbf{P}$ are linearly independent so $\mathbf{P}^T\mathbf{P}$ is positive definite, our minimization problem is equivalent to optimization of a strictly convex quadratic problem. Dual method [1] is an excellent technique to get efficient and numerically stable solutions to this kinds of quadratic problems by utilizing the Cholesky and QR factorizations and updating procedures. Our objective function can be re-written as:

$$\min_{\mathbf{E}} (\mathbf{M}\text{-}\mathbf{PE})^T(\mathbf{M}\text{-}\mathbf{PE}) \qquad \text{s.t.: } \mathbf{C}^T\mathbf{E} - \mathbf{b} \geq 0$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & \cdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}_{N \times N+1} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{N+1 \times 1}$$

and the first constraint here is an equality constraint, and all further are inequality constraints. And further we can re-write the problem in the form of a problem solved by the dual method:

$$\min_{\mathbf{E}} \mathbf{a}^T\mathbf{E} + \frac{1}{2}\mathbf{E}^T\mathbf{G}\mathbf{E} \qquad \text{s.t.: } \mathbf{C}^T\mathbf{E} - \mathbf{b} \geq 0$$

where $\mathbf{a}^T = -\mathbf{M}^T\mathbf{P}$, $\mathbf{G} = \mathbf{P}^T\mathbf{P}$. Note that $\mathbf{G}$ is positive definite for the COSMIC matrix $\mathbf{P}$.

By calling function `solve.QP` from the R package *quadprog*, our optimization problem is solved very fast in less than 30 iterations on average. A similar approach based on QP was also suggested by Lynch [2].

Simulated annealing (SA) One of the generically popular optimization algorithms is simulated annealing, a method that simulates the annealing process in metallurgy to find a global minimum of the objective function. We used the generalized simulated annealing (GSA) approach [3], implemented in the R package *GenSA* [4], as it has been widely used as a global optimization tool for multidimensional functions. The GSA method propose new schemas for visiting distribution, accepting rule, and cooling schedule. The main improvement of the Tsallis statistics [3] lies in the use of a distorted Cauchy-Lorentz distribution, instead of the Gaussian distribution as in the classical SA [5], to randomly generate a neighboring state to visit. Such visiting distribution allows frequently local jumps, but occasionally the jumps can be quite long. GSA is able to accurately locate the absolute minimum of a given function, and convergence is reached much more rapidly than in the classical SA.

Synthetic datasets For each synthetic dataset, we randomly picked operating signatures, selected their contributions and computed a corresponding synthetic mutational profile. Such profile was decomposed into predefined signatures using each of three methods (dS, SA and QP). The inferred signature contribution were compared with the known ground truth contributions of synthetic dataset using the cosine distance. We tested the methods' performance on 1000 synthetic datasets generated based on breast cancer signatures.

References

[1] D. Goldfarb and A. Idnani. "A numerically stable dual method for solving strictly convex quadratic programs". In: *Mathematical Programming* 27.1 (1983), pp. 1–33.

[2] A. G. Lynch. *Decomposition of mutational context signatures using quadratic programming methods [version 1; referees: 1 approved, 1 approved with reservations]*. Vol. 5. 2016.

[3] C. Tsallis and D. A. Stariolo. "Generalized simulated annealing". In: *Physica A,* 233 (1996), pp. 395–406.

[4] Y. Xiang, S. Gubian, B. Suomela, and J. Hoeng. "Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R." In: *The R Journal Volume 5/1, June 2013* (2013).

[5] C. D. Gelatt S. Kirkpatrick and M. P. Vecchi. "Optimization by simulated annealing". In: *Science* 220 (1983), pp. 671–680.

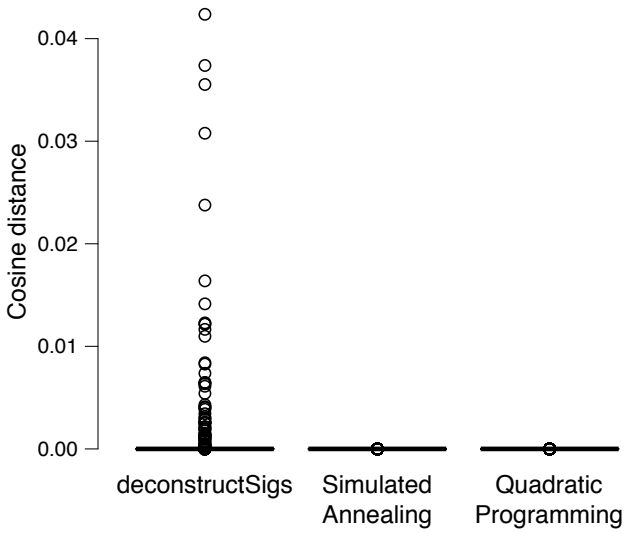

Fig. S1. Comparison of three decomposition methods on 1000 synthetic datasets. For each method, distribution of cosine distance between inferred signature contributions and the ground truth contributions of synthetic datasets is shown.

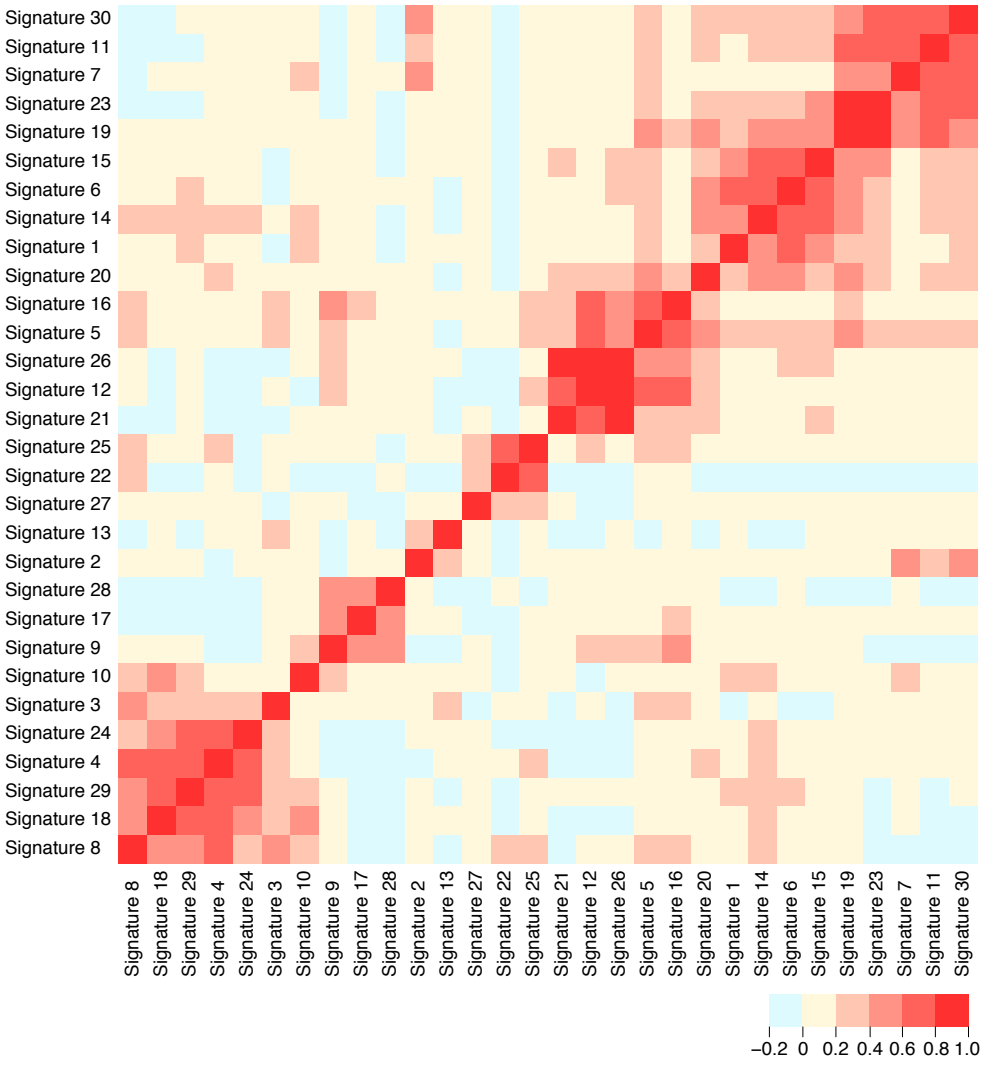

Fig. S2. Pearson's correlation between COSMIC mutational signatures.