CellPress

Cell Reports
2013

Deciphering Signatures of Mutational Processes Operative in Human Cancer

Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge,
Peter J. Campbell, and Michael R. Stratton

Model Definition

- Mutation type

  K-letter alphabet Ξ

- Mutational signature

  discrete probability density function

$$P : \Xi \rightarrow \mathbb{R}_+^K$$

- mutational process P_1

  nonnegative *K*-tuple

$$P_1 = [p_1^1, p_1^2, \ldots, p_1^K]^T \quad \sum_{k=1}^{K} p_1^k = 1$$

Model Definition

- nonnegative mutational signature

$$P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$$

size $K \times N$,

K is the number of mutation types

N is the number of signatures

احتمال رخ دادن هر جهش در هر امضاء

Model Definition

- Exposure $e_g^1 \in \mathbb{N}_0$

  a mutational process P_1 with signature

$$P_1 = [p_1^1, p_1^2, \ldots, p_1^K]^T$$

  in genome g

میزان قرار گرفتن ژنوم در معرض یک فرایند جهش با امضاء مشخص

Model Definition

- nonnegative exposure

$$E=\begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$$

size $N \times G$

N is the number of signatures

G is number of genomes

میزان قرار گرفتن هر ژنوم در معرض هر امضاء

Model Definition

- mutational catalog of a cancer genome g

$$m_g : \Xi \rightarrow \mathbb{N}_0^K$$

- mutational catalog m_1

  nonnegative K-tuple

$$m_1 = [m_1^1, m_1^2, \ldots, m_1^K]^T$$

Model Definition

- nonnegative mutational catalogs

$$M = \begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix}$$

size $K \times G$

K is the number of mutation types

G is number of genomes

میزان هر جهش در هر ژنوم

Framework for Deciphering Signatures of Mutational Processes

- a linear superposition

  of the signatures

  of the mutational processes operative

  in this genome

  and their respective exposures

- *G* genomes and *N* mutational signatures

$$M \approx P \times E$$

Step 1 (Dimension Reduction)

- Reduce the dimensions of *M*

$$\sum_{r \in R} \sum_{g=1}^{G} m_g^r \leq 0.01 \times \sum_{k=1}^{K} \sum_{g=1}^{G} m_g^k$$

new matrix $\dot{M}$

size $\dot{K} \times G$ where $\dot{K} = K - |R|$

Step 2 (Bootstrap)

- Monte Carlo bootstrap resampling

  1. compute $Pr\left(\breve{m}_g^q\right)=\dot{m}_g^q / \sum_{k=1}^{K} \dot{m}_g^k$ for each mutation type *q*

     $=>\left\{\dot{m}_g^q, Pr\left(\breve{m}_g^q\right), q=1,2,\ldots,\dot{K}\right\}$

  2. resample $\dot{K}$ times

     with replacement from weighted set

     such that $\sum_{k=1}^{K} \breve{m}_g^k = \sum_{k=1}^{K} \dot{m}_g^k$

  new matrix $\breve{M}$

Step 3 (NMF)

- the multiplicative update algorithm

$$\min_{P \in M_{R_+}^{(\dot{K},N)}, E \in M_{R_+}^{(N,G)}} \left\| \breve{M} - P \times E \right\|_F^2$$

1. Initialize $p_n^k \geq 0, e_g^n \geq 0, \forall n, g, k$

2. for i= *1,2, . . ., 10000*

   a) If (*P, E*) is stationary, stop

   else

$$e_g^n = e_g^n \frac{\left[P^T \breve{M} \right]_{n,g}}{\left[P^T PE \right]_{n,g}} \quad \forall g, n$$

$$p_n^{\dot{k}} = p_n^{\dot{k}} \frac{\left[\breve{M} E^T \right]_{\dot{k},n}}{\left[PEE^T \right]_{\dot{k},n}} \quad \forall n, \dot{k}$$

Step 4 (Iterate)

- Perform Steps 2 and 3 for I iterations
- I depend on the size and type of the initial matrix M

Step 5 (Cluster)

- $S_P \in M_{R_+}^{(\dot{K},N)}$ set of matrices
  - mutational signatures generated over the I iterations
- $S_E \in M_{R_+}^{(N,G)}$ set of matrices
  - exposures generated over the I iterations
- Cluster the data into N clusters
  - Assign each signature for $\forall P \in S_P$ to exactly one cluster
  - N centroids : average the signatures belonging to each cluster

Step 5 (Cluster)

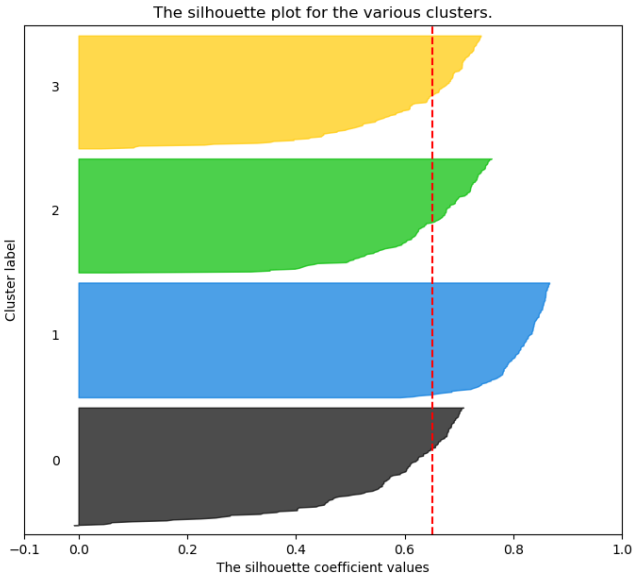
- Similarities between mutational signatures

$$similarity(A,B) = \frac{\sum_{k=1}^{K} A_k B_k}{\sqrt{\sum_{k=1}^{K} (A_k^2)} \sqrt{\sum_{k=1}^{K} (B_k^2)}}$$

- $\overline{P}$ iteration-averaged signature matrix
  - combine the N centroid vectors ordered by their reproducibility

Step 6 (Evaluate)

- Silhouette width → measure reproducibility
  - how similar an object is to its own cluster (cohesion) compared to other clusters (separation)
  - 1.00 : consistently deciphering the same mutational signature
  - 0.00 : lack of reproducibility of the solution



Step 6 (Evaluate)

- Frobenius reconstruction error → measure accuracy
  - $\|M - P \times E\|_F^2$
  - 0.00 : original matrix