Deciphering Signatures of Mutational Processes Operative in Human Cancer
Tumor Cells Carry Somatic Mutations

Tumor

Sequence

gcttcgctagcgcccccttttaatcgatcccgatcg
cccacgatcggatatgctagatcgactgttttttaatt
agccccacatcactatactccctttttggagacgatc
atgccccggtttcgaatgctaaaatgctaaagttt
cccacgatcggatatgctagatcgactgttttttaatt
cagctactgatcgttttgccggccccgggagat
atgccccggttttgaatgctaaaatgctaaagttt

catalog
1. acgatcg
2. cttcccttt
3. tcggata
4. gactgtttt
5. gcccgg
..... 500
Motivation

• Catalogs have heterogeneity
  – Different mutation types: Substitution, missense, nonsense, indels
  – DNA Repair mechanisms
  – Passenger mutations

• Many different cancer signatures
Aim to create computational framework to bridge the gap between the catalogs and signatures

**Catalog**
1. acgatcg
2. ctcccttt
3. tcggata
4. gactgttt
5. gccccg
..... 500

**Lung Cancer Signature**
1. Gcgta (G:C > T:A)
2. Cttccg Deletion
3. tcggata
Feature of Signatures

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

$P = \text{Mutational Signature}$

$p_{1\ldots k} = \text{probability } P \text{ causes a certain mutation}$

$K = 96 \ (6 \text{ types of substitutions } \times 4 \text{ types of } 5' \text{ bases } \times 4 \text{ types of } 3' \text{ bases})$
Mapping of a Genome

\[ m_g^i \approx \sum_{n=1}^{N} p_n^i e_g^n. \]

P = process/mutation

e = exposure/weight

Diagram showing mutational processes and exposure effects on the genome.
What we end up with

\[ P = \begin{bmatrix}
  p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \\
\end{bmatrix} \begin{bmatrix}
  X \\
\end{bmatrix} = M = \begin{bmatrix}
  m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \\
\end{bmatrix} \begin{bmatrix}
  E \\
\end{bmatrix} \]
Non-Negative Matrix Factorization

• Want to extract “P” and “e” from M

**Step 1 and 2**
Reduce Matrix Dimensions

\[ \sum_{r \in R} \sum_{g=1}^{G} m_{rg}' \leq 0.01 \times \sum_{k=1}^{K} \sum_{g=1}^{G} m_{rg}^k, \]

Use bootstrap resampling
Step 3&4: Non Negative Matrix Factorization

- All inputs must be non-negative
- Aims to recreate P and e from M

Iterate until convergence

\[
\begin{align*}
    e_G^N & \leftarrow e_G^N \left[ P^T \tilde{M} \right]_{N,G} \\
    p_N^K & \leftarrow p_N^K \left[ \tilde{M}E^T \right]_{K,N}
\end{align*}
\]

Minimize

Cost Function

\[
\| \tilde{M} - P \times E \|_F^2
\]

Equivalent to \((K,N)^{th}\) element of matrix
**NMF: Faces**

From Lee and Seung, 1999
NMF: Encyclopedia

Breaks topics into Related words

Uses context to Differentiate

From Lee and Seung, 1999
**Step 5: Clustering**

- Partition-clustering algorithm was applied to cluster data into $N$ clusters
Step 6: Evaluate

• Look at Frobenius reconstruction error to evaluate for accuracy
• Compare mutational signatures:

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

Sim(A,B) = 1 means same signature
Does it work?
Breast Cancer Example
Impact

• Ability to generate cancer signatures from comprehensive ‘omic data
• Opens the door for further work. Eg. Sparsity constraint to use a minimum number of signatures