Matrix decomposition in DNA sequence analysis

Peter J Park
Department of Biomedical Informatics
Harvard Medical School
November 11, 2023
Calvin Cunningham
6 hours ago
If Ted Pick was a real math whiz, he would stop trying to stuff his pockets and attempt to solve the Navier-Stokes Equation.

Reply • 👍 1 • Share

Ted Pick Is a Math Whiz Among Math Whizzes. He’s the New Morgan Stanley CEO
The company lifer will have to keep the wealth-management unit happy, while doing the same for investment banking and trading
Trade-offs?

beautiful & deep

messy but useful
This fall: $350!
• “Reading” the 3-billion nucleotide sequences in a person’s genome
• Four nucleotides: A, C, T, G
• One “whole-genome”: ~100 GB
• Latest machine: ~$1M, 128 genomes in 2 days
Sequencing the human genome and interpreting

War and Peace: ~3 million characters
Human genome: ~3 billion characters
that's the height of nothing. And that's the
is that we know
we know nothing. And
we know nothing. And
to human wisdom.
all we can know
know is that we
the height of human
we can know is
All we can know is that we know nothing. And that's the height of human wisdom. And that's the height of human wisdom.
All we can know is that we know nothing. And that's the height of human wisdom.

We can know is that we know nothing. And that's the height of human wisdom.

We can know is that we know nothing. And that's the height of human wisdom.

We can know is that we know nothing. And that's the height of human wisdom.
All we can know is that we know nothing. And that's the height of human wisdom.

-- Leo Tolstoy
Genome sequencing data

- **DNA fragment**
- **"read"**
- **"read length"** ~150bp
- **"coverage"** 30-100X
- colored dots - **"mismatch"** to the reference genome
Mutational processes in cancer and normal cells

- How is your genome mutated when you have cancer?
- What are the mechanisms generating the mutations?
- With whole-genome sequencing, many types of genomic alterations can be detected
Hubble Telescope
50 Terabytes in 20 years

Large Hadron Collider
15 Petabytes in 1 year

My lab’s data: 2.7 PB
Different mutagenic mechanisms generate different errors on the DNA

- Smoking
- Sunlight

Lung cancer
- Melanoma, skin
- Glioma, brain

Very similar in 6 dimensions although the source is definitely different
Neighboring nucleotides are informative

– Spontaneous deamination of 5-methylcytosine

– C>T at CCN and TCN

– UV radiation is known to cause CC > TT
More than one mutational process in a single tumor

Process A

Process B

Process C
Signal decomposition into multiple processes

Underlying signatures

Observations

Process A

Process B

Process C
Signal decomposition into multiple processes

**Observations**

A large dataset

**Underlying signatures**

Process A

Process B

Process C

When enough observations are present pattern recognition algorithms can be used to discover the underlying signatures.
Non-negative matrix factorization

Mutation count matrix
row: mutation type
column: sample

$$X \in \mathbb{R}^{V \times D}$$

Non-negativity constraint (element-wise):

$$W, H \geq 0$$

NP-hard, requires an iterative algorithm for finding local minima

A probabilistic interpretation ->
maximum likelihood approach

Mutation matrix
Signature matrix
Exposure matrix

$$X \approx WH$$

$$W \in \mathbb{R}^{V \times K}$$
$$H \in \mathbb{R}^{K \times D}$$

$$K$$ mutational signatures $$w_1, \cdots, w_K$$
Decomposed using a dataset of >2400 images; 49 basis; black-positive, red-negative;

\[ V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \]

\[
W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \\
W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}}
\]

\[
H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}
\]

Nature, 1999; cited 14000 times
Non-uniqueness of NMF solutions

Many possible solutions because of non-uniqueness
Minimum-volume NMF (mvNMF)

mvNMF penalizes the volume spanned by the signatures and induces a unique solution.

But the signatures are correlated...

\[
\mathcal{L}(L, U, W, \alpha, \sigma^2) = -D_{KL}(X \| WH) - \frac{m}{2}(K + D) \log(2\pi\sigma^2)
- \frac{1}{2\sigma^2} \left( \sum_k \| \ell_k \|^2 + \sum_d \| u_d \|^2 \right),
\]

mvNMF:
Craig et al., IEEE Transactions on Geoscience and Remote Sensing, 1994
Leplat et al., IEEE Transactions on Signal Processing, 2020
Catalog of mutational signatures

Mutational signatures - examples

Smoking

Homologous recombination deficiency

If your tumor genome shows SBS3, you should be considered for PARP inhibitor treatment.
Catalog of mutational signatures

- How many signatures are there?
- More data -> more signatures?
- What is the mechanism behind each signature?
- What is the best way to determine whether a given patient has a specific signature?
- Are there signatures for other types of mutations?
- Can we identify signatures from blood DNA?

Mutational signature analysis methods

**SigMA**
Signature Multivariate Analysis
Gulhan et al, *Nature Genetics*, 2019

**MuSiCal:**
Mutational Signature Calculator
Hu et al, *Nature Genetics*, in press

Can we find patients who should receive PARP inhibitor?

Can we find signatures more accurately?
Studying mutations in single cells and in the brain

Somatic mutation in single human neurons tracks developmental and transcriptional history

Evrony et al., Neuron, 2015
Lodato et al, Science, 2015
Lodato et al, Science, 2018
Bohrson et al, Nature Genetics, 2019
Dou et al, Nature Biotechnology, 2020
Rodin et al, Nature Neuroscience, 2021
Bizzotto et al, Science, 2021
Sherman et al, Nature Neuroscience, 2021
Luquette et al, Nature Genetics, 2022
Acknowledgement

• Clara Bakker
• Michele Berselli
• Joseph Brew
• Craig Bohrson
• Greg Brunette
• Elizabeth Chun
• Shannon Ehmsen
• Niklas Engel
• William Feng
• Cesar Ferreyra-Mansilla
• Beverly Fu
• Teng Gao
• Benedikt Geiger
• Dominik Glodzik
• Doga Gulhan
• Hu Jin
• Clara Kim
• Yoo-Na Kim
• Joe Luquette
• Julia Markowski
• Dominika Maziec
• Bianca Morris
• Rahi Navelkar
• Cassidy Perry
• Kent Pitman
• Doug Rioux
• Will Ronchetti
• Andy Schroeder
• Alex Veit
• Vinay Viswanadham
• Yifan Zhao
• Yuwei Zhang

Collaborators

• Christopher Walsh (BCH)
• Steve Elledge (BWH)
• David Ting (MGH)
• Kevin Haigis (DFCI)
• Charles Roberts (St Jude)
• Mark Johnson (UMass)
• Peter Kharchenko (BCH)
• Mitzi Kuroda (BWH)
• Fred Winston (HMS)
• Colleagues from TCGA, ENCODE, ICGC, 4D Nucleome, Brain Somatic Mosaicism Network

Recent alumni:

• Katerina Chatzipli
• Simon Chu
• Josh Cook
• Andrea Cosolo
• Jake Lee
• Viktor Ljungstrom
• Catherine Song
• Antuan Tran
• Dana Vuzman
I enjoy learning new things. When you start in a new field you have to ask dumb questions. I often say I’m paid for my ability to tolerate feeling stupid.

- Persi Diaconis