Background

Cancer is caused by mutations in our DNA. Cancerous cells contain 1000’s of mutations. Mutations can be classified into types based, e.g., on their kind (C→T) and context (ACG→ATG). Mutations may come from different sources, such as smoking, exposure to UV light, failure of DNA damage repairs. These mutational processes generate mutation types according to unique mutational distributions, also called signatures of the mutational process. Different signatures co-occur together in cancer. Active signatures and their exposures (i.e. mixing coefficients) may vary across tumors.

In topic modeling terms:
Tumor samples are documents
Mutations are words
Mutation types are the vocabulary
Signature are topics
Exposures are mixing coefficients

Motivation

Questions:
How much do signature exposures change during tumor evolution?
Do changes correspond to important events in tumor development?

Goals:
Reconstruct signature trajectories over time
Find change points in signature trajectories

Data

Pseudo-time series of mutations from cancer genome sequencing
Mutations are grouped into time points
• Each time point contains 400 mutations divided into 96 classes
• 1633 samples from 40 cancer types

Model

I) Inferring signature mixtures
We fit mixture of multinomials in each time point

\[ p(z_i | \pi) = \pi_i \]

Assignment to signature

Signature mixtures

Probability of each mutation:

\[ p(x_n | z_i, \pi, \mu_1:K) = \prod_{k=1}^{K} \mu_{ik}^{x_{nk}} \]

Fixed signature multinomials

Likelihood of data

\[
\log \mathcal{L}(\pi, \mu, x_{1:N}) = \sum_{n=1}^{N} \log p(x_n | \pi, \mu_1:K) = \sum_{n=1}^{N} \sum_{i=1}^{M} \log p(x_n | z_i, \pi, \mu_1:K) p(z_i | \pi)
\]

E-step: 

\[ p(z_i | x_n, \pi, \mu_1:K) = \pi_i \prod_{k=1}^{K} \mu_{ik}^{x_{nk}} \]

M-step: 

\[ \pi_i = \frac{1}{N} \sum_{n=1}^{N} p(z_i | x_n, \pi, \mu_1:K) \]

II) Finding change points
1) Iterate through all time points
2) Recompute mixtures of multinomials separately before and after change point
3) Point with maximum likelihood is considered a new checkpoint.
4) We use Bayesian Information Criterion to estimate optimal number of change points.

\[ BIC = -2 \ln \mathcal{L} + (#\text{changepoints} + #\text{signatures}) \ln(N) \]

Results

I) Signatures change through tumor development

II) Most changing tumor types

Breast cancer (BRCA) Leukemia (CLL)

46% of samples have maximum change in exposure >20%
31% samples with max change >30%

Most active:
• S3 (failure of DNA double-strand break-repair)
• S5 (aging)
• S8 (typical for breast cancer)

Most active:
• S5 (aging)
• S9 (somatic hypermutation, typical for leukaemias and B-cell lymphomas)

III) Correlation with tumor transitions

IV) Signature changes are not random

Summary

We developed a variant of topic modelling over time on the mutation types to track changes in tumor evolution.
We can: infer active processes in the tumor
detect changes in tumor development
find subclones with greater sensitivity

References