New name?

LIBRA: **Li**kelihood-**B**ased mutational signatu**R**es **A**ttribution in cancer genomics

Other keywords: Jointly optimizing, multinomial distribution, sampling error, adaptive?

Abstract (Word count: 249)

Multiple mutational processes fuel carcinogenesis. These processes leave characteristic signatures in cancer genomes. Deciphering the signatures of mutational processes operative in cancer can help elucidate the mechanisms underlying cancer initiation and development. This process involves decomposing cancer mutations by nucleotide context into a linear combination of mutational signatures. Previously published methods use forward selection or iterate all combinations by brute force (REF). Other approaches use linear programming (REF), which is not efficient in optimization. Here, we formulated the task as a likelihood based optimization problem with L1 regularization and developed a software tool, LIBRA. First, by explicitly formulating multinomial sampling into likelihood function and jointly optimizing a multinomial sampling process and signature fitting, LIBRA is aware of the sampling uncertainty. It is especially pivotal in high sampling variance settings, for example, when we only observe low mutation counts in whole exome sequencing (WES). Moreover, LIBRA uses L1 regularization to parsimoniously assign signatures to cancer genome mutation profiles, leading to sparse and more biologically interpretable solutions. Additionally, LIBRA integrates prior biological knowledge harmoniously into the solution by fine-tuning penalties on coefficients. Compared with hard thresholding signatures, our method leaves leeway for noise and rare signatures. Last, the model complexity is informed by the size and complexity of the data through empirical parameterizing based on performance. In sum, LIBRA fits a signature attribution jointly with a multinomial sampling process, while using regularization to promote sparsity and interpretability. Meanwhile, this framework

Shantao 5/6/2018 9:44 PM
Deleted: with an empirically derived stopping criterion

Shantao 5/6/2018 9:48 PM
Deleted: adding a

Shantao 5/6/2018 9:48 PM
Deleted: step

Shantao 5/6/2018 9:44 PM
Deleted: uncertainty

Shantao 5/7/2018 12:10 AM
Deleted: mixture

Shantao 5/6/2018 9:49 PM
Deleted: critical

Shantao 5/6/2018 9:52 PM
Deleted: By

Shantao 5/6/2018 9:52 PM
Deleted: assigning

Shantao 5/6/2018 9:53 PM
Deleted: the solution becomes

Shantao 5/6/2018 9:55 PM
Deleted: sigLASSO

Shantao 5/6/2018 9:54 PM
Deleted: subsetting

Shantao 5/6/2018 9:55 PM
Deleted: before fitting

Shantao 5/6/2018 9:54 PM
Deleted: unknown

Shantao 5/7/2018 7:54 PM
Formatted: Highlight

Shantao 5/6/2018 9:55 PM
Deleted: T

Shantao 5/7/2018 7:54 PM
Deleted: using cross-validation and subsampling

Shantao 5/6/2018 9:55 PM
Deleted: sigLASSO

Shantao 5/6/2018 11:57 PM
Deleted: offers a framework that

empowers researchers to use and integrate their biological knowledge and expertise into the model.

Introduction

Mutagenesis is a fundamental process underlying cancer development. Examples include spontaneous deamination of cytosines, the formation of pyrimidine dimers by ultraviolet (UV) light, and the crosslinking of guanines by alkylating agents [REF]. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints [REF]. Notably, these processes have characteristic mutational nucleotide context biases. Mutation profiling of cancer samples at manifestation has revealed that mutations accumulate over a lifetime; this includes somatic alterations that occur both before cancer initiation and during cancer development. In a generative model, multiple latent processes generate mutations over time, drawing from their corresponding nucleotide context distributions (“mutation signature”). In cancer samples, mutations from various mutational processes are mixed and observable by sequencing.

By applying unsupervised methods such as non-negative matrix factorization (NMF) and clustering to large-scale cancer studies, researchers have identified at least 30 mutational processes [REF]. Many processes have been recognized and linked with known etiologies, such as aging, smoking, or ApoBEC activity. Investigating the fundamental processes underlying mutagenesis can help elucidate cancer initiation and development.

One major task in cancer research is to leverage signature studies on large-scale cancer cohorts and efficiently attribute active signatures to new cancer samples [REF]. Although we do not fully know the latent mutational processes in cancer samples, we can make reasonable and logical assumptions about the solutions of such studies. Here, we aimed to design a computational framework that could meet these expectations. For example, we believe a solution should be sparse as past studies indicate that not all signatures can be active in a single sample or

Shantao 5/7/2018 12:47 AM
Deleted: assign

even a given cancer type. An apparent example is, we should not observe UV-associated signatures in tissues that are not exposed to UV. Likewise, we only expect to observe activation-induced cytidine deaminase (AID) mutational processes, which are biologically involved in antibody diversification, in B cell lymphomas. We also prefer a sparser solution as it explains an observation in a simpler fashion, consistent with Occam's principle.

Previously published methods use forward selection with a post hoc empirical pruning to achieve sparsity or iterate all combinations by brute force (REF) with a pre-fixed, small number of signatures. Other approaches use linear programming (REF), which is not efficient in optimization. None of the approaches explicitly formulates the multinomial sampling process into the model. Here, we formulated the task as a likelihood based optimization problem with L1 regularization. First, by jointly fitting signatures with a multinomial sampling process, LIBRA is aware of the sampling uncertainty. This property is especially critical in high sampling variance settings, for example, when we only observe low mutation counts in whole exome sequencing (WES). Second, LIBRA penalizes the model complexity by regularization. The most straightforward way to do this would be to use the L0 norm (cardinality of active signatures), but this approach cannot be effectively optimized. Conversely, using the L2 norm flattened out at small values leads to many tiny, non-zero coefficients, which are hard to interpret biologically. LIBRA uses L1 norm, which promotes sparsity. Meanwhile, L1 norm is a convex map, thus allows efficient optimization. Additionally, this approach is able to harmoniously integrate prior biological knowledge into the solution by fine-tuning penalties on the coefficients. Compared with the current approach of hardly subsetting signatures before fitting, our soft thresholding method leaves leeway for noise and unidentified signatures. Finally, unlike previous methods, LIBRA is aware of data complexity such as mutational number and patterns in the observation. Our method is automatically parameterized empirically on performance, allowing data complexity to inform

Shantao 5/7/2018 12:22 AM
Deleted: with an empirically derived stopping criterion

Shantao 5/7/2018 12:52 AM
Deleted: explicitly adding a multinomial sampling step into likelihood function and

Shantao 5/7/2018 12:52 AM
Deleted: optimizing

Shantao 5/7/2018 12:52 AM
Deleted: sampling uncertainty and mixture

Shantao 5/7/2018 12:52 AM
Deleted: fitting

Shantao 5/7/2018 12:53 AM
Deleted: It

Shantao 5/6/2018 9:50 PM
Deleted: our sigLASSO approach

Shantao 5/7/2018 12:58 AM
Deleted: Our sigLASSO

Shantao 5/7/2018 12:58 AM
Deleted: sigLASSO

Shantao 5/7/2018 7:54 PM
Deleted: based on cross-validation

model complexity. This approach promotes result reproducibility and fair comparison of datasets.

Material and Methods

Signature identification problem

Mutational processes leave mutations in the genome with distinct nucleotide contexts. Specifically, we considered the mutant nucleotide context and looked one nucleotide ahead and behind. This divides mutations into 96 trinucleotide contexts. Each mutational process carries a unique signature, which is represented by a mutational trinucleotide context distribution (Fig. 1A).

Thirty signatures were identified by NMF (with Frobenius norm penalty) and clustering from large-scale pan-cancer analysis (REF). Here, our objective was to leverage the pan-cancer analysis and decompose mutations from new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following non-negative regression problem. It maintains the original Frobenius norm:

$$W = \underset{W \in Z^+}{\mathrm{argmin}} \|M - SW\|_2^2$$

The mutation matrix, *M*, contains mutations of each sample cataloged into 96 trinucleotide contexts. *S* is a 96×30 signature matrix, containing the mutation probability in 96 trinucleotide contexts of the 30 signatures. *W* is the weights matrix, representing the contributions of 30 signatures in each sample.

Sampling variance

In practice, this problem is optimized on R$^+$ instead of integers for efficiency and simplicity (REF), ignoring the discrete nature of mutation counts. This approach essentially transforms observed mutations into a multinomial probability distribution, making model insensitive to the total mutation count. Yet the total mutation count plays a critical role in inference. Assuming mutations are drawn from an underlying probability distribution (which is the mixture of several

Shantao 5/7/2018 1:01 AM
Deleted: It

Shantao 5/7/2018 7:08 PM
Deleted: to optimization

Shantao 5/7/2018 1:03 AM
Deleted: If we assume

mutational signatures), the mutations follow a multinomial distribution. The total mutation count is the sample size of the distribution, and affects the variance.

For instance, 20 mutations of 96 categories give us very little confidence in inferring the underlying mutation distribution. If we observed 2,000 mutations, we would have much higher confidence. Methods undiscriminating these two scenarios are clearly defective. Here, we aim to use a likelihood-based approach to acknowledge the sampling variance and design a tool sensitive to the total mutation count.

LIBRA model (I still need to fix the notations…want to check if the journal accept LaTeX or not first)

We break data generation process into two parts: first, multiple mutational signatures mix together to form an underlying mutation distribution. Second, we observe a set of categorical data (mutations), which is a realization of the underlying mutation distribution. We use y_i (i = 1…n) to denote the mutation count of the ith category. $\vec{m}$ is the underlying mutation probability distribution with m_i denote the probability of the jth category.

$$L = P(\vec{y}|SW) = P(\vec{y}|\vec{m})P(\vec{m}|SW)$$

To promote sparsity and interpretability of the solution, LIBRA uses adds an L1 norm regularizer on the weights (i.e., coefficients) of the signatures. LASSO is mathematically justified and can be computationally efficiently solved (REF). Now the log-likelihood looks like:

$$1 \propto \sum_{i=1}^{n}\{y_i \log m_i - \frac{\alpha}{2}(m_i - \sum_{k=1}^{K} s_{ik}w_k)^2 - \lambda \sum_{k=1}^{K} c_k w_k\}$$

$$s.t. \forall w_k \geq 0, \forall m_i \geq 0, \sum_{i=1}^{n} m_i = 1$$

Here, $\alpha = 1/\sigma^2$. We will infer α from the residual errors from linear regression. λ is parameterized empirically (see below). $\vec{c}$ is a vector of 30 penalty weights (c_1,

Shantao 5/7/2018 1:06 AM
Deleted: count …ollows ... [1]

Shantao 5/7/2018 1:04 AM
Deleted: likelihood based…ikelihoo ... [2]

Shantao 5/7/2018 7:16 PM
Deleted: y_i \in Y …o denote the ... [3]

Shantao 5/7/2018 7:15 PM
Deleted: M

Shantao 5/7/2018 7:16 PM
Formatted: Subscript

Shantao 5/7/2018 7:16 PM
Deleted: $_j$

Shantao 5/7/2018 7:14 PM
... [4]

Shantao 5/7/2018 7:23 PM
Deleted: sigLASSO …IBRA uses L ... [5]

Shantao 5/7/2018 7:46 PM
Deleted: $1 \propto \sum_{i=1}^{n}\{y_i \log m$

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Shantao 5/7/2018 8:05 PM
Deleted: $s.t. \forall w_k \geq 0, \forall p_i \geq 0,$

Shantao 5/7/2018 7:55 PM
Deleted: by cross-validation… ... [6]

Shantao 5/7/2018 7:21 PM
Deleted: *C*…is a vector of 30 indica ... [7]

c_2, …, c_k), each indicating whether a certain signature should be fully penalized (i.e., 1), partially penalized (e.g., 0.5), or not penalized (i.e., 0). We can also use C to perform adaptive LASSO where C is of the form 1/β^{OLS}. The aim is to get less biased estimator by applying smaller penalties on larger values. This value should be tuned to reflect the level of confidence in prior knowledge.

Optimizing LIBRA

The log likelihood is concave in respect to both $\vec{m}$ and $\vec{w}$. Hence the loss function, the negative log likelihood, is biconvex. We optimize the function by iteratively updating these two variables.

Initialization:

$$\mathrm{m}_i^0 = m_i^{mle} = \frac{y_i}{\sum_{i=1}^{n}(y_i)}$$

$\vec{w}$-step:

$$\vec{w}^{t+1} = \underset{\vec{w}=w_1,w_2,..w_i,...w_n}{\mathrm{argmax}} \sum_{i=1}^{n}\{\tfrac{\alpha}{2}(m_i^{t+1} - \sum_{k=1}^{K} s_{ik}w_k)^2 - \lambda\sum_{k=1}^{K} c_k w_k\}$$

$\vec{m}$ -step:

$$\alpha = \frac{N-\hat{s}}{N}\frac{1}{\sum_{i=1}^{n}(\hat{m}_i - m_i)^2}$$

$$\vec{m}^{t+1} = \underset{\vec{m}=m_1,m_2,..m_i,...m_n}{\mathrm{argmax}} \sum_{i=1}^{n}\{y_i \log m_i - \tfrac{\alpha}{2}(m_i - \sum_{k=1}^{K} s_{ik}w_k)^2\}$$

To begin the iteration, we initialize $\vec{m}$ using its maximum likelihood estimator and start with the $\vec{w}$-step. $\vec{w}$-step is a nonnegative linear LASSO regression that can

Shantao 5/7/2018 7:22 PM
Moved down [1]: This value should be tuned to reflect the level of confidence in prior knowledge.

Shantao 5/7/2018 7:22 PM
Deleted: less …maller penaltiesy ... [8]

Shantao 5/7/2018 7:22 PM
Moved (insertion) [1]

Shantao 5/7/2018 7:28 PM
Formatted: Font:Not Italic

Unknown
Formatted: Font:(Default) Cambria Math, Italic

Shantao 5/7/2018 7:25 PM
Deleted: M is a latent variable, ... [9]

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Shantao 5/7/2018 7:33 PM
Deleted: E-step:

$$\mathrm{E}^{t+1} = \underset{W}{\mathrm{argmax}}\ \tfrac{\alpha}{2}(m_i^{t+}$$

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Shantao 5/7/2018 7:46 PM
Deleted:

Unknown
Formatted: Font:(Default) Arial, Bold

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Shantao 5/7/2018 8:03 PM
Deleted: m_i = y_i/\sum{y_i} …nd ... [10]

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Shantao 5/7/2018 7:53 PM
Deleted: M

be efficiently solved by glmnet (REF). λ is parameterized empirically by repeatedly splitting the nucleotide contexts into training set and testing set. At every step, we split the data set into eight subsets. Each subset contains two of every single nucleotide substitutions. We then hold off one subset as the testing dataset and only fit the signatures on the remaining ones. After circling all eight subsets and repeating the process for ten times, we used the largest λ (which leads to a sparser solution) that gives mean square error (MSE) within one standard deviations (SD) of the minimum.

Then we use the LASSO error variance estimator to estimate \alpha (REF). We solve the $\vec{m}^{t+1}$ with a Lagrange multiplier to maintain the linear summation constrain $\sum_{i=1}^{n} m_i = 1$. The nonnegative constrain of m_i is satisfied in only retain a nonnegative root of the solution (see Appendix?). This process is iterated until convergence.

The key step is the $\vec{m}$ –step. In this step, we try to estimate $\vec{m}$ that optimizes the multinomial likelihood *while* is not too far away from the fitted \hat{m}, as measured by the L2 norm. The trade-off between the multinomial likelihood and the L2 loss reflects the sampling error. The sampling size (sum of m_i), thegoodness of signature fit (as reflected in \alpha) and the overall shapes of Y and \hat{m} all controls this trade-off tension.

Shantao 5/7/2018 7:55 PM
Deleted: by 12-fold (since 12 is a divisor of 96) cross-validation

Shantao 5/7/2018 8:00 PM
Deleted: hold off eight trinucleotide context

Shantao 5/7/2018 8:00 PM
Deleted: s

Shantao 5/7/2018 7:58 PM
Deleted: 88 trinucleotide contexts

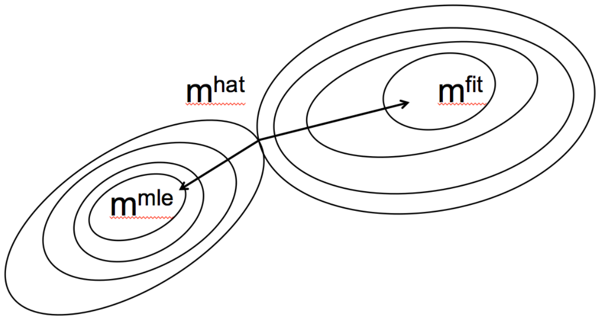
Shantao 5/7/2018 8:01 PM
Deleted: We

Shantao 5/7/2018 7:59 PM
Deleted: three

Shantao 5/7/2018 8:02 PM
Deleted: M^{t+1} using

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial



Unknown

Formatted: Font:(Default) Arial