SigLASSO: a LASSO regression approach for mutational signatures identification in cancer genomics

Abstract: Multiple mutational processes fuel carcinogenesis and leave characteristic signatures in cancer genomes. Identifying operative mutational processes by signatures helps understand cancer initiation and development. The task is to break down cancer mutations by nucleotide context into a linear combination of mutational signatures. The solution should be sparse and biologically interpretable. Previously published methods use empirical forward selection or iterate all signature combinations using brutal force. Here, we formulate it as a more mathematically justified LASSO linear regression problem. By parsimoniously assigning signatures to cancer genome mutation profiles, the solution becomes sparse and biologically interpretable. Additionally, LASSO organically integrates biological prior into the solution by fine-tuning penalties on coefficients. Compared with current approach of subseting signatures before fitting, our method leaves leeway for noises and unknown signatures, leading to a more reliable and interpretable signature solution. Last, our method is automatically parameterized based on cross-validation and subsampling. The model complexity is informed by the size and complexity of the data. This objective, robust approach promotes data replicability and fair comparison across samples.

TOO CONGRAT

Introduction

Mutagenesis is the fundamental process for cancer development. Examples include spontaneous deamination of cytosine, ultraviolet light inducing pyrimidine dimer and alkylating agents crosslinking guanines. Multiple endogenous and exogenous mutational processes drive cancer mutagenesis and leave distinct fingerprints. Noticeably, these processes have characteristic mutational nucleotide context biases. Mutation profiling of cancer sample at manifestation finds all mutations accumulate over lifetime, including somatic alterations both before the cancer initiation and during cancer development. In a generative model, over time multiple latent processes generate mutations drawing from their corresponding nucleotide context distributions ("mutation signature"). In cancer sample, mutations from various mutation processes are mixed and observable by sequencing.

Many mutation processes are recognized and linked with known etiologies. Understanding the fundamental underlying processes helps understand cancer initiation and development.

EXPL

Previously published methods use empirical forward selection or iterate all combinations (brutal force). Here, we formulate it as a more mathematically rigorous LASSO linear regression problem. By penalizing the L1 norm of coefficients, the algorithm produces sparse and biologically interpretable solutions. Additionally, LASSO organically integrates biological prior into the solution by fine-tuning penalties on coefficients. Compared with current approach of subseting signatures before fitting, our method leaves leeway for noises and unidentified signatures, leading to a more reliable and interpretable signature solution. Last, our method is automatically parameterized based on cross-validation and subsampling. This objective, robust approach promotes data replicability and fair comparison across datasets.

PGNITA

Material and methods

Signature identification problem

Different mutational processes leave mutations in the genome with distinct nucleotide context. In particular, we consider the mutant nucleotide context and look one nucleotide ahead and behind. This divides mutations into 96 trinucleotide contexts. Each mutational process carries its unique signature, which is represented by a mutational trinucleotide context distribution (Fig 1A). 30 signatures are identified from nonnegative matrix factorization (NMF) and clustering from large-scale pan cancer analysis (REF). Here our object is to leverage on the pan cancer analysis and decompose mutations observed in new samples into a linear combination of signatures. Mathematically, the problem is formulated as the following nonnegative regression problem:

$$\min_{W \in R^+} \|SW - M\|_2$$

The mutation matrix, M, contains mutations of each sample broken down into 96 nucleotide contexts. S is a 96×30 signature matrix, containing the mutation probability in 96 trinucleotide contexts of the 30 signatures. W is the weighting matrix, representing the contribution of 30 signatures in each sample.

SigLASSO workflow

To promote sparsity and interpretability of the solution, SigLASSO uses LASSO regression, adding an L1 norm regularizer on the weights of signatures (coefficients). LASSO is both mathematically justified and computationally efficient.

$$\min_{W \in R^+} (\|SW - M\|_2 + \sum \lambda I \|W\|)$$

λ is parameterized by 10-fold cross validation. We use the smallest λ that gives mean square error within 3 standard deviances (SD) of the minimum.

Mutation count is an important factor for signature identification. I is a vector of indicator functions of whether a signature should be penalized. If we have strong prior belief that a signature should be active, the corresponding I is zero. Mathematically, LASSO is equivalent to a Bayesian linear regression framework with Laplace prior.

To assess the solution stability and account for lower signature ascertainment when less mutations are observed, we perform subsampling. At each subsampling step, we sample 50% mutations, solve the SigLASSO problem and find active (i.e. have nonnegative coefficients) signatures. In the end, we only retain signatures that are active in more than τ fraction of all subsampling trials. τ can be set empirically between 0.6 to 0.9 (REF). In our study, we use 0.6 and set subsampling to 100 times.

A schematic illustration of the SigLASSO workflow is shown here (Fig 1B).

Data simulation and model evaluation

First we downloaded 30 previously identified signatures (http://cancer.sanger.ac.uk/cosmic/signatures, REF). We created simulated dataset by randomly and uniformly drawing signatures (2 to 5 signatures) and corresponding weights (minimum: 0.02). Noise was simulated at various levels with a uniform distribution on 96 trinucleotide contexts. Then we summed up all the signatures and noise to form a mutation distribution. We randomly drew mutations from this distribution with different mutation counts.

We ran deconstructSigs according to the original publication (REF). To evaluate the performances, we compared the inferred signature distribution with the true distribution and calculated mean square error (MSE). We also measured the number of truth positive signatures in the solution as well as the false positive and negative ones.

Testing on real dataset

We realized the real cancer mutational profiles are much noisier than our simulation and exhibit highly nonrandom distribution of signatures. To assess the performance of our method on real world cancer dataset, we use TCGA somatic mutations from various cancer types. VCF files are downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). A detailed list of files used in this study can be found in Appendix X.

The signature composition results are compared with previous pan-cancer signature analysis (http://cancer.sanger.ac.uk/cosmic/signatures, REF). Priors used in SigLASSO are also extracted from this analysis.

SigLASSO software suite

SigLASSO accepts (vcf files or) processed mutational spectrums. It allows the users to specify biological priors, subsampling steps and subsampling cutoff. SigLASSO uses the 30 published signatures by default. Users are given the option to also supply customized signature files. LASSO is computationally efficient. Using default settings, the program could successfully decompose a cancer sample data in a few seconds on a regular laptop (3 GHz i7 CPU, 16 GB DDR3 memory).
SigLASSO is released as an R package (SigLASSO). Updated code is also distributed on GitHub (https://github.com/ShantaoL/SigLASSO).

Results

1. Performance on simulated dataset

LASSO if more computationally efficient than forward selection.

2. Performance on real dataset

2.1 WGS scenario: renal cancer datasets, prior matters

35 Whole-genome sequenced papillary kidney cancer samples. Priors are from previous large-scale Pan-cancer studies.

2.2 WXS scenario: esophageal carcinoma, our method is sensitive to mutation counts

181 Whole-exome sequenced esophageal carcinoma samples[6] with mutation counts > 20. No prior is used.

2.3 Implications in infer signature changes in tumor evolution

Discussion