De novo Discovery of Mutated Driver Pathways in Cancer

Fabio Vandin[1,2,⋆,⋆⋆], Eli Upfal[1,⋆⋆], and Benjamin J. Raphael[1,2,⋆⋆]

[1] Department of Computer Science, Brown University, Providence, RI
[2] Center for Computational Molecular Biology, Brown University, Providence, RI
{vandinfa,eli,braphael}@cs.brown.edu

Motivation. Next-generation DNA sequencing technologies are enabling genome-wide measurements of somatic mutations in large numbers of cancer patients. A major challenge in interpretation of this data is to distingush functional *driver* mutations that are important for cancer development from random, *passenger* mutations. A common approach to identify driver mutations is to find genes that are mutated at significant frequency in a large cohort of cancer genomes. This approach is confounded by the observation that driver mutations target multiple cellular signaling and regulatory pathways. Thus, each cancer patient may exhibit a different combination of mutations that are sufficient to perturb the necessary pathways. However, the current understanding of the somatic mutational process of cancer [3, 5, 6] places two additional constraints on the expected patterns of somatic mutations in a cancer pathway. First, an important cancer pathway should be perturbed in a large number of patients. Thus we expect that with genome-wide measurements of somatic mutations a driver pathway will exhibit high *coverage*, where most patients will have a mutation in some gene in the pathway. Second, since driver mutations are relatively rare and typically a single driver mutation is sufficient to perturb a pathway, a reasonable assumption is that most patients have a single driver mutation in a pathway. Thus, the genes in a driver pathway exhibit a pattern of *mutually exclusive* driver mutations, where driver mutations are observed in exactly one gene in the pathway in each patient. There are numerous examples of sets of mutually exclusive mutations [5, 6].

Methods. Motivated by these biological observations and the availability of somatic mutation data on large sets of patients, we introduce the problem of finding sets of genes with the following properties: (i) *coverage*: most patients have at least one mutation in the set; (ii) *exclusivity*: nearly all patients have no more than one mutation in the set. We define a measure on sets of genes that quantifies how much sets exhibit both properties and show that finding sets of genes that optimize this measure is NP-hard. In contrast, we prove that a straightforward

⋆ Corresponding Author.

⋆⋆ This work is supported by NSF grant IIS-1016648, the Department of Defense Breast Cancer Research Program, the Alfred P. Sloan Foundation, and the Susan G. Komen Foundation. BJR is also supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

greedy algorithm produces an optimal solution with high probability when given a sufficiently large sets of patients and subject to some statistical assumptions on the distribution of the mutations.

Since the number of patients in currently available datasets is only in the hundreds and the statistical assumptions for the greedy algorithm may be too restrictive (e.g. they are not satisfied by copy number aberrations), we also introduce a second approach. We use a Markov Chain Monte Carlo (MCMC) algorithm to sample from sets of genes with a distribution that gives significantly higher probability to sets of genes with high coverage and exclusivity. Our MCMC algorithm is based on the Metropolis-Hastings method. Although the Metropolis-Hastings method defines a chain that is guaranteed to converge to the desired stationary distribution, there is in general no guarantee how rapidly the chain will converge. While there has been significant progress in recent years in developing mathematical tools for analyzing the convergence time [4], our ability to analyze useful chains is still limited, and in practice most MCMC algorithms rely on simulations to provide evidence of convergence to stationarity [2]. Nevertheless, we prove that our MCMC algorithm converges rapidly to the equilibrium distribution.

Results. We apply the MCMC algorithm to analyze sequencing data from 623 genes in 188 lung adenocarcinoma patients and 601 genes in 84 glioblastoma patients. In both datasets, we find sets of 2-3 genes that are mutated in large subsets of patients and are largely exclusive. These sets include genes in the Rb, p53, and mTor signaling pathways, all pathways known to be important in cancer. In glioblastoma, the set of three genes that we identify was shown to be associated with shorter survival [1]. Finally, we show that the MCMC algorithm efficiently identifies sets of six genes with high coverage and exclusivity in simulated mutation data with thousands of genes and patients.

References

1. L. M. Backlund, B. R. Nilsson, H. M. Goike, E. E. Schmidt, L. Liu, K. Ichimura, and V. P. Collins. Short postoperative survival for glioblastoma patients with a dysfunctional Rb1 pathway in combination with no wild-type PTEN. *Clin. Cancer Res.*, 9:4151–4158, Sep 2003.
2. W. Gilks. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London, 1998.
3. F. McCormick. Signalling networks that cause cancer. *Trends Cell Biol.*, 9:M53–56, 1999.
4. Dana Randall. Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering*, 8(2):30–41, 2006.
5. B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat. Med.*, 10:789–799, Aug 2004.
6. C.H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, 2008.