Counting and Sampling Problems in Computational Biology

Mohammed El-Kebir[1], Jackie Oh[1], Yuanyuan Qi[1], and Palash Sashittal[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Introduction

Combinatorial optimization problems are ubiquitous in the field of computational biology — from cancer phylogenetics to molecular epidemiology. Typically, these problems are NP-hard. As a consequence, linear programming and integer programming solvers have become part of the standard toolkit for computational biologists. However, these tools are not as effective in exploring uncertainty and degeneracy (non-uniqueness) of optimal solutions to the computational problems. Here, we show three such applications in cancer phylogenetics and molecular epidemiology that can be solved efficiently by recent advances in approximate counting and almost uniform sampling in SATISFIABILITY (SAT).

Transmission Network Inference

Motivation Recent developments in sequencing technologies have made molecular epidemiology an indispensable tool for real-time outbreak management, public health policies and devising disease control strategies during disease outbreak. This is evident from the fact that since the first reported case of COVID-19 on 24$^{\text{th}}$ December 2019, 33,656 consensus sequences [1] and 7,810 bulk sequencing samples [2] have been published by the research community.

Problem statement Given epidemiological and genomic data, the key challenge is to infer the transmission history of the outbreak. We consider a given timed phylogeny of the pathogen samples and a contact map for the infected hosts indicating putative transmissions. Each leaf of the given phylogeny is labeled by the infected host from which the sample was collected (see Figure 1(a)). Our goal is to find a host labeling for the internal vertices of the phylogeny that yields the most parsimonious transmission history that agrees with the given contact map. Parsimony can be defined either in terms of the number of transmitted pathogen strains or the number of transmission events in the outbreak. As many parsimonious solutions exist, we seek to sample uniformly from the solution space.

SAT formulation We introduce Boolean variables to encode the labelings of the internal vertices of the phylogeny and transmission edges between infected hosts. The constraints of the optimization problem can be easily described in CNF form.

Results We applied this approach to find the transmission history of the 2014 Ebola outbreak between different chiefdoms in Sierra-Leone [3] and the transmission chain in an HIV-1 outbreak involving 11 patients from 1988 to 2006 [4].

Phylogenetic Deconvolution of Bulk DNA samples

Motivation The evolutionary process of individual cancers is appropriately modeled by a phylogenetic tree. Tumor phylogenies enable important applications that aim to improve our understanding of tumorigenesis and guide personalized response to treatment. However, multiple phylogenies with distinct topologies may be inferred from the same bulk DNA sequencing data of a tumor. To overcome this challenge, current methods aim to sample trees with maximum likelihood, but empirical and theoretical evidence shows that such methods are not sampling the solution space uniformly [5].

Problem statement Given bulk DNA sequencing data of m samples and n mutations of a tumor in the form of frequency matrix $F \in [0,1]^{m \times n}$. The frequencies F correspond to unknown mixtures $U \in [0,1]^{m \times n}$ of nodes of an unknown phylogenetic tree T, whose nodes are mutations. Our goal is to sample from all possible *complete perfect phylogenetic trees* such that the frequency of each mutation is greater or equal to the sum of the frequencies of its children for all samples. This restriction is also known as the *sum condition* (SC).

SAT formulation We use the ancestry graph G_F h to help us formulate the problem, which is a directed acyclic graph whose vertices correspond to the n mutations and whose edges (i, j) indicate that the frequency of i is greater or equal to the frequency of j for all samples. Our goal is to sample spanning arborescences in G_F that satisfy (SC) (see Fig. 1b). We introduce Boolean variables to indicate the root vertex, presence/absence of each edge, and pairwise reachability. The constraint of T being a spanning tree of G_F is formulated using clause containing these variables. Moreover, we discretize the frequencies by multiplying a large integer and using a bit vector representation. We model (SC) using bit vector arithmetic.

Results We applied this method to a simulated dataset and an maximum likelihood extension of the method to a non-small cell lung cancer cohort (tracerX) [6]. We showed that our method outperforms PhyloWGS [7], an existing MCMC-based method, in terms of sampling uniformity and running time, and report trees with high quality.

Sampling 1-Dollo Phylogenies

Motivation While bulk sequencing technology yields sequence reads from a large collection of cells, single cell sequencing (SCS) technology allows us to determine the presence of single-nucleotide variants (SNVs) in individual cells of a tumor. Given SCS data of a patient's tumor, we seek a phylogenetic tree that models the evolutionary history of cells within the tumor. This can help one better understand the mechanisms behind intra-tumor heterogeneity and identify targets for treatment.

Problem statement Our goal is to use SCS data to find a phylogenetic tree that explains the evolutionary history of cells within a tumor. The leaves of this tree are labelled by cells observed at the time of sequencing, the internal vertices are labelled by ancestral clones of cells within the tumor, and each edge is labeled by a mutation gain or a mutation loss. Under the 1-Dollo parsimony model, each mutation is gained exactly once and can be lost at most once due to a copy number aberration. When we allow for losses and account for possible errors in the sequencing data, we find that there are many possible 1-Dollo Phylogenies for one patient's SCS data. We also find that after correcting for errors, cells generally originate from a small number of clones and mutations cluster together on branches of the tree. So our goal is to uniformly sample phylogenies for SCS data that have a feasible number of false positives and false negatives and contain a pre-determined number of cell and mutation clusters.

SAT formulation We introduce Boolean variables that represent false positives, false negatives, losses, and cell and mutation clustering. A 1-Dollo Phylogeny can be represented as a matrix that does not contain any of 25 forbidden submatrices [8], so we add clauses to enforce the absence of each of the forbidden submatrices in a potential solution. We also use a half/full adder approach to constrain the number of false positives and false negatives and enforce the number of cell and mutation clusters for each solution.

Results We applied this method to a simulated datasets with up to 15 cells and 15 mutations. We are currently looking into a cutting planes approach to make this method more feasible on real data.

References

[1] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.

[2] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[3] Palash Sashittal and Mohammed El-Kebir. SharpTNI: Counting and sampling parsimonious transmission networks under a weak bottleneck. *BMC Medical Genomics*, 2019. In Print. Journal version of RECOMB-CG 2019 paper.

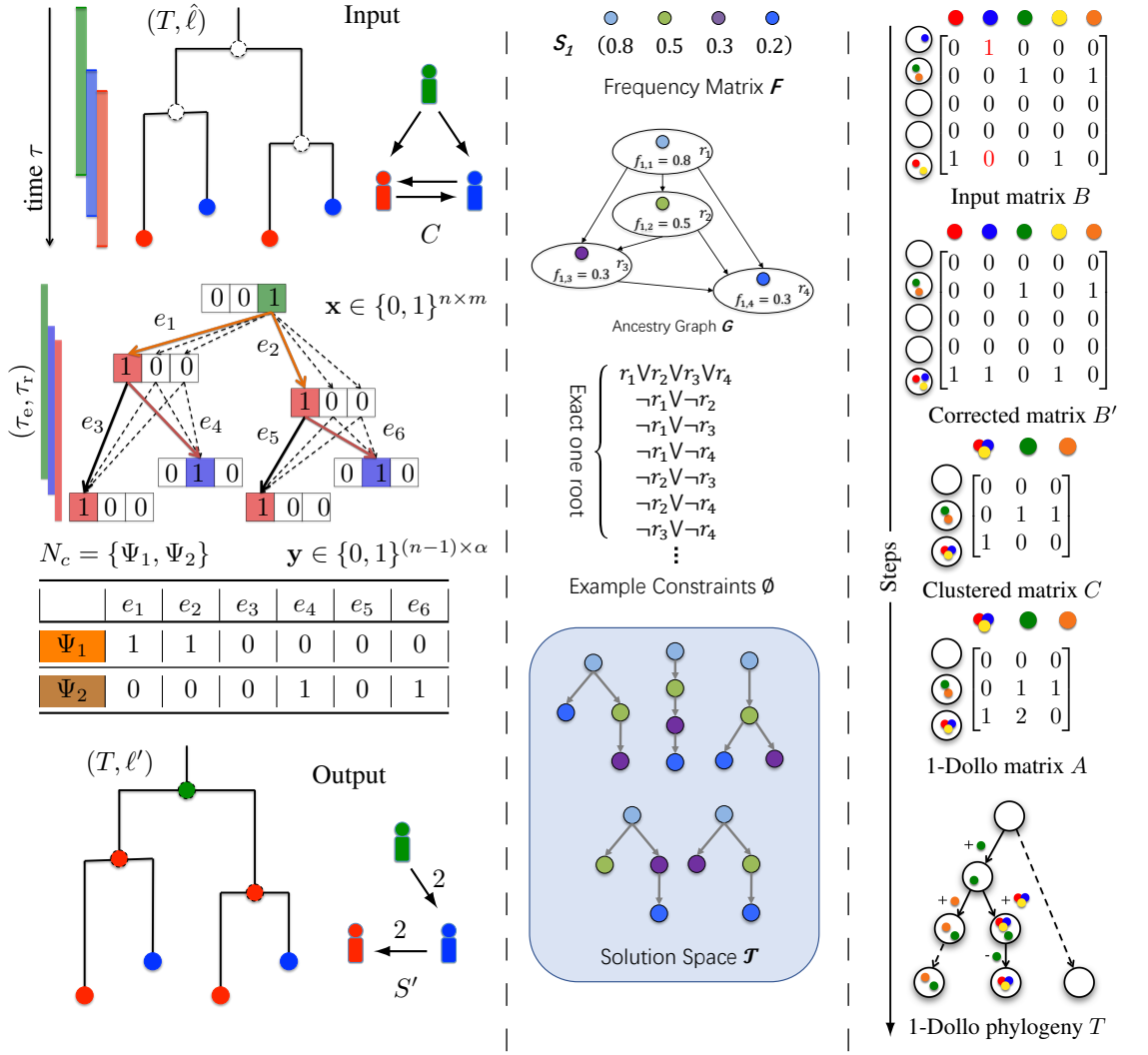
[4] Palash Sashittal and Mohammed El-Kebir. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics/ISMB 2020*.

[5] Yuanyuan Qi, Dikshant Pradhan, and Mohammed El-Kebir. Implications of non-uniqueness in phylogenetic deconvolution of bulk dna samples of tumors. *Algorithms for Molecular Biology*, 14(1):19, 2019.

[6] Mariam Jamal-Hanjani et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[7] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun H Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, February 2015.

[8] Mohammed El-Kebir. SPhyR: Tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics (Oxford, England)*, 34(17):i671–i679, 09 2018.



	e_1	e_2	e_3	e_4	e_5	e_6
Ψ_1	1	1	0	0	0	0
Ψ_2	0	0	0	1	0	1

(a) **Transmission Network Inference** (b) **Phlyogenetic Deconvolution** (c) **1-Dollo Phylogeny Sampling**

Figure 1: (a) **Transmission Network Inference instance –** The input consists of a timed phylogeny T with a leaf labeling ℓ and a contact map C. Each leaf of T corresponds to a pathogen sample collected from an infected host. The output is the internal vertex labeling ℓ' of the timed phylogeny and the transmission network S'. (b) **Phylogenetic Deconvolution instance –** An example of input frequency matrix F with $m = 1$ bulk samples and $n = 4$ mutations that has a set $\mathcal{T}(F)$ of five solutions. Clauses that enforce the presence of a single root vertex are shown. (c) **1-Dollo problem instance –** The input is in the form of a binary matrix, $B \in \{0,1\}^{m \times n}$. After correcting for errors (marked in red) and clustering, we aim to find a 1-Dollo completion matrix $A \in \{0,1,2\}^{s \times t}$. Entries that are 2 in A represent mutation clusters that were lost due to a copy number aberration. The rows of a 1-Dollo completion matrix are the leaves of a 1-Dollo phylogeny T.