Bioinformatics, 31(7), 2015, 1133–1135
doi: 10.1093/bioinformatics/btu755
Advance Access Publication Date: 13 November 2014
Applications Note

OXFORD

Phylogenetics

SNV-PPILP: refined SNV calling for tumor data using perfect phylogenies and ILP

Karen E. van Rens[1,2], Veli Mäkinen[1] and Alexandru I. Tomescu[1,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland and [2]HAN University of Applied Sciences, Nijmegen, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on February 6, 2014; revised on November 8, 2014; accepted on November 11, 2014

Abstract

Motivation: Recent studies sequenced tumor samples from the same progenitor at different development stages and showed that by taking into account the phylogeny of this development, single-nucleotide variant (SNV) calling can be improved. Accurate SNV calls can better reveal early-stage tumors, identify mechanisms of cancer progression or help in drug targeting.

Results: We present SNV-PPILP, a fast and easy to use tool for refining GATK's Unified Genotyper SNV calls, for multiple samples assumed to form a phylogeny. We tested SNV-PPILP on simulated data, with a varying number of samples, SNVs, read coverage and violations of the perfect phylogeny assumption. We always match or improve the accuracy of GATK, with a significant improvement on low read coverage.

Availability and implementation: SNV-PPILP, available at cs.helsinki.fi/gsa/snv-ppilp/, is written in Python and requires the free ILP solver lp_solve.

Contact: tomescu@cs.helsinki.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent studies on cancer tissue (Gerlinger *et al.*, 2012; Newburger *et al.*, 2013; Potter *et al.*, 2013) suggest that as the disease progresses, the original tumor progresses into several different sub-clones. These sub-clones contain not only the original mutations but also new mutations added over time; this work focuses on single-nucleotide variants (SNVs). Accurate SNV calls can better reveal early-stage tumors, identify mechanisms of cancer progression or help in drug targeting.

This phylogenetic assumption has been recently taken into account in Salari *et al.* (2013), by proposing a tool for refining the SNV calling of GATK's Unified Genotyper (McKenna *et al.*, 2010), the state-of-the-art SNV multi-sample caller for next-generation sequencing data.

In this note, we present SNV-PPILP (SNV calling with Perfect Phylogenies and Integer Linear Programming), a tool for refining GATK's Unified Genotyper SNV calls for multiple samples. We assume these samples form a character-based phylogeny, the characters being the SNVs reported by GATK. As in Salari *et al.* (2013), we work with the *perfect phylogeny* model; however, we have a new problem formulation for fitting GATK's calls to such a phylogeny, which we solve exactly using integer linear programming (ILP).

SNV-PPILP can be run especially on low-coverage samples, without big violations of the perfect phylogeny model. For a significant improvement in accuracy, six or more samples are needed.

2 Methods

The perfect phylogeny model assumes that (i) once a mutation occurred in a node of the phylogenetic tree, it is passed along to all its descendants and (ii) a mutation does not recur in other nodes apart from these. As argued in Newburger *et al.* (2013) and Salari *et al.* (2013), these assumptions are reasonable in the context of cancer genomics. From the set of SNVs reported by GATK's Unified Genotyper for each sample, we construct a binary matrix M whose

n rows represent the n samples and whose m columns represent the set of all m SNVs found by GATK's Unified Genotyper in at least one sample. For every entry (i, j) where GATK's Unified Genotyper did make a call (either of presence or absence), GATK also provides a *likelihood* (PL field), which we denote here by $L(i, j)$.

It is a standard result, e.g. Estabrook *et al.* (1975), that this matrix corresponds to a perfect phylogeny if and only if every two columns j_1 and j_2 are *compatible*, in the sense that for no three rows i_1, i_2, i_3 of M, *all* of the following conditions hold:

1. $M(i_1, j_1) = 1$ and $M(i_1, j_2) = 1$ and
2. $M(i_2, j_1) = 1$ and $M(i_2, j_2) = 0$ and
3. $M(i_3, j_1) = 0$ and $M(i_3, j_2) = 1$.

Our strategy is, in the first step, to select a set of pairwise compatible columns by a maximum-weight independent set (MWIS) formulation, similar to Salari *et al.* (2013) but which we solve here exactly using ILP. In the second step, we iteratively edit the remaining columns, so that they become compatible with the columns initially selected and with those corrected so far. This is achieved using a new problem formulation, which we also solve by ILP.

In the first step, we collapse identical columns into *mutation groups* and consider the graph in which they are the nodes and where two mutation groups are adjacent if and only if they are not compatible. Using ILP, we find a MWIS in this graph, where, contrary to Salari *et al.* (2013), a mutation group has as weight the sum of the likelihoods of its calls. Denote by A the matrix made up of the columns in the mutations groups of this independent set; A corresponds to a partial perfect phylogeny. Denote by B the matrix made up of the other columns of matrix M.

In the second step, contrary to Salari *et al.* (2013), we iteratively edit each column in B, so that it becomes compatible with all columns of A, remove it from B and add it to A. We consider the columns of B in decreasing order on their average likelihoods. We propose a new editing strategy, for each such column c, as follows. For each row i, if $M(i, c) = 1$, then the weight $w(i, c)$ of correcting the ith row of c is $L(i, c)$. If $M(i, c) = 0$, then we set $w(i, c)$ to be the mean of the likelihoods of the '1' entries in column c minus $\sqrt{2}$ times their standard deviation. Having these weights, we then ask for the correction of the rows of c, such that the sum of the weights of all corrections is minimum. We solve this by another ILP, as follows. We transform all '0' entries into '–1' entries to flip them by a multiplication with –1. For every row i, we have a variable $x_i \in \{0, 1\}$, with the meaning that if $x_i = 0$, then we flip the entry in row i of column c, and if $x_i = 1$ then this entry is not flipped. Thus, each row i of the corrected column becomes $(2x_i - 1)M(i, c)$. We want to find the binary vector $(x_1, \ldots, x_n)$ which maximizes $\sum_{i=1}^{n} x_i w(i, c)$, under the constraint that the edited column is compatible with each column in A. These compatibility constraints are imposed by writing, for every three set of rows $\{i_1, i_2, i_3\}$, such that $M(i_1, c) = M(i_2, c) = 1 \neq M(i_3, c)$, and every column j in A, the following two linear inequalities:

- $(2x_{i_1} - 1)M(i_1, c)M(i_1, j) - (2x_{i_2} - 1)M(i_2, c)M(i_2, j) - (2x_{i_3} - 1)M(i_3, c)M(i_3, j) \leq 2$,
- $-M(i_1, j)(2x_{i_1} - 1)M(i_1, c) + 2M(i_2, j)(2x_{i_2} - 1)M(i_2, c) - M(i_3, j)(2x_{i_3} - 1)M(i_3, c) \leq 2$.

Finally, we add the edited column to A, and remove it from B and proceed by editing the next column in B. At the end of this process, matrix A is the edited matrix of all SNVs reported by our method.

3 Experiments and discussion

We conducted three types of experiments on simulated data, as follows. (i) First, we created a 'perfect' scenario, by varying the number of samples, between 3 and 10, and generating a random tree with that number of leaves. We randomly assigned mutation groups (i.e. sets of known SNVs from GATK resource bundle dbsnp_137.b37) to the edges of these random trees. The size of the mutation groups was varied according to four simulation scenarios, described in Table 1 (right). This resulted into $8 \times 4 = 32$ experiments. We ran the experiments on chromosome 21. In each leaf (i.e. sample) of the simulated phylogenetic tree, we mutate chromosome 21 with the set of SNVs appearing in mutation groups on the path from the root to it. With DWGSIM (Li, 2012), we generate Illumina reads from it, of length 100, coverage 15× and base error rate set to the default 2%. We align them with BWA-MEM (Li *et al.*, 2009), and the alignments are given to GATK's Unified Genotyper. As in Salari *et al.* (2013), we followed GATK's best practices guidelines (see Supplementary Material for details) and ran GATK in multi-sample mode. (ii) Second, we repeated the above 'perfect' scenario but with coverages 30× and 100×. (iii) Third, we considered violations of the 'perfect' scenario, by allowing recurring mutations [violations to property (ii) of the perfect phylogeny model]. For each $r \in \{1, 5, 15\}$, we selected r% of all mutations present in the tree and assigned them to one edge of the tree at random. As above, we then propagated them to all the leaves descending from this edge. This gave $8 \times 4 \times 3 = 96$ other experiments.

For each SNV in the union of *all* mutation groups labeling a tree and for each sample (i.e. leaf of the tree), we checked whether the SNV was reported correctly in the sample (true positives (TP)), if

Table 1. The F measure of GATK's Unified Genotyper's calls (column A), SNV-PPILP's calls (column B) and of the method of Salari *et al.* (2013) (column C)

Samples	Perfect scenario			$r = 1\%$		$r = 5\%$		$r = 15\%$		coverage 30×	
	A	B	C	A	B	A	B	A	B	A	B
3	0.476	0.476	0.318	0.465	0.466	0.547	0.548	0.486	0.486	0.828	0.839
4	0.577	0.601	0.393	0.560	0.580	0.623	0.643	0.641	0.675	0.826	0.828
5	0.605	0.640	0.310	0.595	0.624	0.622	0.646	0.472	0.479	0.837	0.842
6	0.530	0.557	0.293	0.567	0.603	0.626	0.654	0.672	0.726	0.805	0.807
7	0.559	0.603	0.270	0.626	0.688	0.687	0.738	0.569	0.617	0.784	0.787
8	0.651	0.724	0.385	0.568	0.618	0.714	0.777	0.604	0.655	0.861	0.876
9	0.621	0.681	0.305	0.644	0.712	0.642	0.684	0.639	0.713	0.826	0.837
10	0.672	0.752	0.255	0.658	0.730	0.664	0.722	0.605	0.663	0.813	0.823

Mutation group size	M.	S.D.
Low M., Low S.D.	1000	25
Low M., High S.D.	1000	250
High M., Low S.D.	2000	50
High M., High S.D.	2000	500

The values are averages over the four mutation group sizes. These are distributed normally, with mean and standard deviation as shown on the right

it was missing from the report, when it should have been present (false negatives (FN)) or if it was reported even though it should not be present (false positives (FP)). We computed the F measure $= 2 \times \mathrm{TP}/(2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ shown in Table 1 (left).

We experimented also with the method of Salari *et al.* (2013), but its F measure was consistently worse than GATK's original calls. For this reason, we report it only in the perfect scenario.

SNV-PPILP's F measure is always the same or better than GATK's. At 15× coverage, the difference with respect to GATK's F measure is significant. On 30× coverage, we see ∼1% difference in F measure. At coverage 100×, GATK's F measure is very good, at ∼0.98. SNV-PPILP's F measure remained the same up to the third decimal. However, we noticed a slight improvement in the absolute numbers of TPs and FNs (see Supplementary Material). Even though for high coverages the relative improvement is small, it might reveal some critical SNVs in cancer progression studies. Moreover, SNV-PPILP seems resilient to a degree of heterogeneity, as the F measure improved even for a recurring rate $r = 15\%$. However, big violations to the perfect phylogeny model cannot be handled by our model, as this is its core editing principle.

We also ran SNV-PPILP on six whole-genome breast cancer samples from NCBI: PRJNA193652. SNV-PPILP's runtime is only a fraction of GATK Unified Genotyper's overall running time. All running times are provided in the Supplementary Material.

Acknowledgement

We thank the anonymous referees for very helpful comments.

Funding

This work was supported by Academy of Finland [250345] (CoECGR) and [274977 to A.I.T.].

Conflict of Interest: none declared.

References

Estabrook,G.F. *et al.* (1975) An idealized concept of the true cladistic character. *Math. Biosci.,* **23**, 263–272.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.,* **366**, 883–892.

Li,H. (2012) Whole genome simulation: dwgsim, https://sourceforge.net/apps/mediawiki/dnaa/index.php.

Li,H. *et al.* (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* **25**, 1754–1760.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* **20**, 1297–1303.

Newburger,D.E. *et al.* (2013) Genome evolution during progression to breast cancer. *Genome Res.,* **23**, 1097–1108.

Potter,N.E. *et al.* (2013) Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Res.,* **23**, 2115–2125.

Salari,R. *et al.* (2013) Inference of tumor phylogenies with improved somatic mutation discovery. In: Deng,M. *et al.*, *RECOMB 2013, Research in Computational Molecular Biology - 17th Annual International Conference,* volume 7821. Springer, Beijing, China, pp. 249–263.