Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Subject Section

Phylotyper: In silico predictor of molecular subtypes from gene sequences

Matthew D. Whiteside [1,*], Chad R. Laing [1] and Victor P.J. Gannon [1,*]

[1] National Microbiology Laboratory, Public Health Agency of Canada, Lethbridge, AB, Canada, T1J 3Z4

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Whole genome sequencing is being adopted in public health for improved surviellence and outbreak analysis. Molecular subtyping has been used in public health to infer phenotypes and flag high-risk bacterial strain groups. In silico tools that predict molecular subtypes from WGS data are needed to transition historical data and systems to WGS-based protocols.
Results: Phylotyper is a novel solution for in silico molecular subtype prediction from WGS data. It is a general prediction tool that can be applied to most molecular subtype schemes. Phylotyper uses phylogeny to model the evolution of the subtype and infer subtypes for unannotated sequences. The phylogenic framework in Phylotyper, improves accuracy, provides useful contextual feedback and confidence scores, and is more capable at identifying novel subtypes over in silico approaches based on sequence similarity.
Availability: Phylotyper is available for download from: `https://github.com/superphy/insilico-subtyping`.
Contact: matthew.whiteside@phac-aspc.gc.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole-genome sequencing (WGS) is transforming the public health field by providing a highly efficent tool for interogating bacterial populations. The speed, discriminatory power and broad utility of WGS data can improve surviellence and outbreak analysis. Adoption of WGS in the public health, however, requires bridging of historical data and systems with new the technologies and data. One of the workhorse methods used in public health are molecular subtyping (such as serotyping). As a survillence tool, subtypes provide an straight-forward designation that is typically used to distinguish taxonomic groups and infer phenotypes, for example, pathogens from non-pathogens. A WGS-based approach to subtyping would have several benefits over current subtype systems; would be faster, have improved discrimination and would be cheaper and easier to maintain and routinely run. Accordingly, new *in silico* tools have been developed to predict subtypes from WGS data.

Many of the current *in silico* predictors of bacterial subtypes use a similar approach; genes or genome sequences with unknown type are compared against a reference database of sequences with known subtypes. Subtype assignment is based on sequence similarity with the subtype annotation from the top match above a pre-selected sequence similarity threshold being used to assign a subtype to the unknown query. These tools do not examine the phylogenetic context or consider rate of sequence mutation within and between subtypes. Sequence similarity scores are not direct indicators of the level of uncertainty assocated with a subtype prediction.

Phylotyper is a novel in silico predictor of subtypes from sequence data. It builds a phylogenetic tree consisting of reference sequences with known subtype and the unknown query sequences. Using phylogenetic ancestral reconstruction to assign likelihoods of each type to the branch points in the tree and also compute the transition rates between subtypes, Phylotyper assigns an unknown query sequence a subtype based on the extrapolated value from its ancestors in the tree. Subtype information is mainly used as a proxy for evolutionarily-related bacterial strain groups or to infer phenotypes. A phylogenetic framework for predicting subtypes is more consistent with the main uses of subtype information.

2 The Phylotyper Approach

The core of Phylotyper, is an ancestral state reconstruction method that has been adapted for hidden state prediction. In phylogenetics, ancestral state

reconstruction involves the prediction of traits of ancestors from extant descendants. This methodology can be extended to predict properties in existing strains under investigation for which the a properties' state is unknown. Ancestral state reconstruction (ASR) has been successfully applied to hidden state prediction in the field of microbial metagenomics; the tool PICRUSt uses ASR to estimate the gene family content contributed by a bacteria to a metagenomic sample.

In Phylotyper, the `rerootingMethod` function from the phytools R package is used to perform the ASR. It is based on a method originally described in Yang *et al*. (1995) for estimating the marginal likelihood of a discrete set of states for the internal nodes in a phylogenetic tree. The phytools `rerootingMethod` function was selected over alternatives, such as APE ace, because it can handle extant tip nodes with unkown states in the phylogenetic tree and also compute posterior probability for those nodes, hence, making applicable for hidden state prediction. The maximum marginal likelihood (also called emperical Bayesian posterior probability) calculated for unknown tip nodes reflects the most likely state for the node given the emperically estimated evolution model and phylogeny. It captures the uncertainty associated with phylogenetic branch lengths and rates of change associated with the states. In the context of Phylotyper, the posterior probabilities provides a confidence value associated with a predicted subtype. The evolutionary model used in `rerootingMethod` is the M_k model for discrete states (a continuous-time Markov process model)). It is estimated from the data, so in Phylotyper, the reference set of genes used to build a phylogenetic tree and run the ASR, is key to the performance of the predictor. For subtype schemes packaged in Phylotyper, the annotated genes will be maintained and routinely updated to provide a robust and comprehensive reference set. The other key assumption is that gene phylogeny is correlated with the evolution of subtype state. Because of these assumptions are central to the method, all new subtype schemes added into Phylotyper are evaluated for their predictive performance.

Phylotyper is developed in python and R. Outlined below are the steps and tools used in the Phylotyper pipeline:

1. Identify subtype gene loci in input genomes using blastn or blastx
2. Align input genes with unknown sunbtype against a pre-aligned set of reference genes using the tool MAFFT's `--add` feature (Katoh and Standley, 2013).
3. If multiple loci are involved, concatenate individual alignments into superalignment.
4. Generate phylogenetic tree of aligned genes with FastTree.
5. Run phytools `rerootingMethod` using the phylogenetic tree and assigned subtypes. Genes with unknown subtype are assigned a flat prior.
6. Identify the subtype with maximum marginal likelihood for the unknown genes and report to user. Users are also provided with a image of the phylogenetic tree that shows the position of the unknown genes. An example is shown in Figure 1.

3 Functionality in the Phylotyper Tool

Phylotyper was designed to be incorporated into a whole genome sequencing public health workflow. The main input into Phylotyper is assembled (but not necessarily closed) genome sequences. Putative loci needed for the selected subtype scheme are identified in the input genomes using BLAST. The discovered loci are then sent to the Phylotyper subtype prediction module. It is possible to bypass the loci search step and provide genes directly to Phylotyper using the subtype run mode.

The Phylotyper approach can be used to predict any biological property that can be inferred from the phylogenetic distribution of a nominal set of genomic loci. Currently, subtype schemes for *Escherichia coli* Shiga-toxin 1 and 2 subtypes, Serotype O and H-types and Intimin subtypes are available in the Phylotyper package. Details about the current schemes available in the Phylotyper package is provided in table XXX. However, included in the Phylotyper software, is the capability to add new subtype schemes. Creating a new subtype scheme will save the required reference files in a data directory, allowing the newly added schemes to be easily run from Phylotyper. We also encourage users to contribute their subtype schemes to the main software repository (please send us subtype schemes you have developed and wish to share).

Checks are built-in new subtype scheme pipeline to ensure the main assumptions of the Phylotyper approach are not violated. A scan of the reference phylogenetic tree is conducted to search for tightly clustered subclades in the tree that have divergent subtypes. We first estimate the distribution of the inter-patristic distances of genes with the same subtype. The distribution is used to dynamically select a distance threshold, which we employ to flag subclades wherein the inter-patristic distance is less than the threshold but contains multiple distinct subtypes. We also conduct a leave-one-out cross validation for each gene in the new subtype scheme to assess the scheme's predictive capability. Users are alerted if their subtype schemes that have a F1-score below 0.9 from the cross-validation analysis.

Phylotyper also offers flexibility in the parameterization of evolution model used in the ASR step. A component of the underlying ASR framework, is an Mk or markov model of subtype evolution, for which an emperical transition rate matrix is estimated from the data. The transition matrix is used to calculate the expected number of subtype state changes given a distance in the phylogenetic tree. Different model parameterizations can be defined for the transition rate matrix. The simplest parameterization available in Phylotyper is the equal rates model; all subtypes have the same forward and reverse rate. The most complex parameterization available im Phylotyper is the symmetric model, wherein each forward and reverse rate for a given pair of subtypes are assigned a separate parameter. Frequently, the number of subtypes makes the symmetric model too computationally prohibitive. To offer more flexible models in these situations with reduced numbers of free parameters, two custom parameterization approaches were developed. The custom approaches both use a binning strategy that attempts to identify sets of subtypes that would have similar rates and assign them a single parameter as a set. These approaches are described in detail in the Supplementary Information XXX. Each of these model parameterizations; equal, symmetric and the two custom models are tested and evaluated in new subtype pipeline. The parameterization that has highest accuracy (based on a LOOCV analysis) is selected. In the case of ties, the model with the fewest parameters is given precidence (the symmetric model is not tested when the number of subtypes is over 10)

In testing the predictive capability of Phylotyper, we found that in some cases, jointly using multiple loci increased the subtype prediction accuracy. And so, in Phylotyper, multiple loci can be used in creating subtype schemes. The loci will be independently identified in the genome using BLAST and then independently aligned. The individual loci alignments are then concatenated to form a single superalignment that is used to build the predictive phylogenetic tree. The Stx1 and Stx2 schemes are examples of a multi-loci subtype schemes. The A and B subunit genes that make up the holo-toxin in Shiga-toxin are the inputs in these subtype schemes.

4 Comparison to a Sequence Similarity Approach

To assess the performance of the phylogenetic-based method used in Phylotyper and compare it to a direct sequence similarity-based approach, we ran a leave-one-out cross validation analysis. Each reference gene with experimentally validated subtype was removed from the training

Table 1. Subtype schemes available in Phyloyper

Name	Description	Species	Loci
stx1	Shiga-toxin 1 subtype	*E. coli*	2
stx2	Shiga-toxin 1 subtype	*E. coli*	2
eae	Intimin subtype	*E. coli*	1
flic	H-serotype based on flagellin gene	*E. coli*	1
wz	O-serotype based on wzx and wzy genes	*E. coli*	2

dataset and used as a test input into the Phylotyper program. Similarily, using a mock sequence-similarity based tool we developed for assessment purposes, each gene was witheld from the BLAST reference database and used as input (Source code is also available in the phyloyper git repository). The sequence-similarity tool assigns putative subtypes using a BLAST sequence search. The predicted subtype annotation for the test input is assigned from top BLAST hit, provided the top hit passes a percent identity, e-value and alignment coverage minimum thresholds. The threshold values are listed in TODO. This approach of transfering annotations from genes with highest sequence similarity is the core of most current subtying methods.

5 Discussion

From assembled WGS data, Phylotyper can assign unclassifed strains a molecular subtype. Currently the Phylotyper software offers several subtyping schemes for *E. coli*. It can be applied to most molecular subtype schemes, however. To add a new scheme, it requires a reference set of loci sequences that describe the subtype scheme profile. The Phylotyper suite includes functionality to build and run new schemes by users. Performance testing showed that the Phylotyper method is highly accurate.

Phylotyper computes an empirical model of subtype evolution to predict subtypes for unclassified sequences. The model can accomodate subtypes that evolve at different rates; a feature that would be difficult in a sequence similarity top-hit approach. By estimating the phylogenetic distribution of each subtype in a particular scheme, Phylotyper more easily identify a novel subtype or accurately classify a novel sequence allele. The performance of sequence similarity approaches is more dependent on the reference database containing the exact allele. They do not provide straightforward mechanism to interpret novel alleles that are not in the database. The phylogenetic framework and empirical evolution model provides a statistical confidence value for interpreting novel alleles.

To exand the Phylotyper program suite, we are working on developing functionality that will automatically identify loci in a whole genome sequence that is highly evolutionarily correlated with a target subtype.

6 Conclusion

Text Text text

Acknowledgements

text text text

Funding

This work has been supported by the...

References

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4), 772–780.