Dynamic Phylogenetic Inference for Sequence-based Typing Data[*]

Alexandre P Francisco
INESC-ID / IST, University of Lisbon, Portugal

Marta Nascimento
INESC-ID / IST, University of Lisbon, Portugal

Cátia Vaz
INESC-ID / ISEL, Polythenic Institute of Lisbon, Portugal

ABSTRACT

Typing methods are widely used in the surveillance of infectious diseases, outbreaks investigation and studies of the natural history of an infection. And their use is becoming standard, in particular with the introduction of High Throughput Sequencing (HTS). On the other hand, the data being generated is massive and many algorithms have been proposed for phylogenetic analysis of typing data, such as the goeBURST algorithm. These algorithms must however be run whenever new data becomes available starting from scratch. We address this issue proposing a dynamic version of goeBURST algorithm. Experimental results show that this new version is efficient on integrating new data and updating inferred evolutionary patterns, improving the update running time by at least one order of magnitude.

CCS CONCEPTS

• **Theory of computation** → *Dynamic graph algorithms*; • **Applied computing** → *Computational biology*; • **Mathematics of computing** → Graph algorithms;

KEYWORDS

phylogenetic inference; phylogenetic trees; dynamic algorithms; sequence-based typing data

ACM Reference format:
Alexandre P Francisco, Marta Nascimento, and Cátia Vaz. 2017. Dynamic Phylogenetic Inference for Sequence-based Typing Data. In *Proceedings of ACM-BCB '17, Boston, MA, USA, August 20-23, 2017,* 1 pages.
https://doi.org/10.1145/3107411.3108214

DYNAMIC GOEBURST

Sequence-based typing methods are fundamental in epidemiological and genetic studies [4]. On the other hand, the introduction of High Throughput Sequencing (HTS) technology, and the decrease in costs of using it, have contributed to large repositories of typing data, creating the need of developing efficient and scalable methods that are suitable for large scale phylogenetic analyses [3, 5]. Since most methods are distance-based, their running time is dominated by the computation of pairwise distances among taxa, leading to at least quadratic running time [5]. We propose here a dynamic version of the goeBURST algorithm [1].

The problem solved by goeBURST can be stated as a graphic matroid, with the solution following a classic greedy approach. Its running time is then $O(\ell \min\{n^2, m \log n\})$, where n is the number of taxa, m is the number of pairs for which the distance is defined, and ℓ is the number of loci under analysis. Usually we consider all-pairs Hamming distance [2], leading to $O(\ell n^2)$ running time.

The dynamic version of goeBURST allows the addition of a new taxa to a previously computed minimum spanning tree (MST) without the need of running goeBURST algorithm from scratch. Although there are well know dynamic algorithms for computing and updating MSTs, they are not directly usable in this context due to distance updating and tie breaking rules. The basic process of adding a new edge to a MST can be done by checking if that edge creates a cycle and, in that case, removing the heaviest edge from that same cycle. Since goeBURST has some specific tie breaking rules regarding the number of locus variants, we have implemented two versions for dynamic updating: dynamic goeBURST algorithm with tie breaking rules and without (relying only on edge weights/distances). The dynamic algorithm with tie breaking rules is more complex because adding a taxon can change the entire tree. This is due to the fact that the overall number of locus variants for all taxa will vary causing a possible change on the tie breaking rule applied before. Adding a new taxon takes linear time on average for real data, being dominated by the time for processing n edges and pairwise comparisons. We achieve this running through the use of efficient data structures, namely for MST representation.

Experimental results were obtained with the *Streptococcus pneumoniae* MLST dataset available at https://pubmlst.org, with the number of taxa varying from $n = 10$ to $n = 1000$ and running times averaged over 500 executions. The update time for adding a new taxon grows linearly with the number of taxa on the tree, while the static goeBURST takes quadratic time for computing the tree, as expected. In particular, static goeBURST takes more than 1500 ms to add 1000 taxa, while both dynamic versions take less than 150 ms to add a new taxon to a previously computed 999 taxa MST.

REFERENCES

[1] Alexandre P Francisco, Miguel Bugalho, Mário Ramirez, and João A Carriço. 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* 10, 1 (2009), 152.

[2] R. W. Hamming. 1950. Error Detecting and Error Correcting Codes. (1950), 147–160 pages.

[3] Marta Nascimento, Adriano Sousa, Mário Ramirez, Alexandre P Francisco, João A Carriço, and Cátia Vaz. 2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics* 33, 1 (2017), 128–129.

[4] D Ashley Robinson, Edward J Feil, and Daniel Falush. 2010. *Bacterial population genetics in infectious disease.* John Wiley & Sons.

[5] Naruya Saitou. 2013. *Introduction to evolutionary genomics.* Springer.

[*] This work was partly supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under projects BacGenTrack (TUBITACK/0004/2014), PRECISE (SAICTPAC/0021/2015) and UID/CEC/500021/2013.

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4722-8/17/08.
https://doi.org/10.1145/3107411.3108214