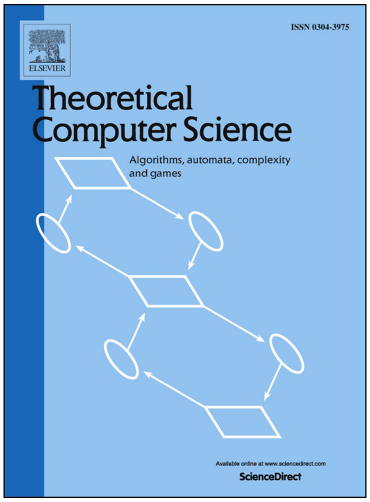
ISSN 0304-3975
ELSEVIER
Theoretical Computer Science
Algorithms, automata, complexity and games
Available online at www.sciencedirect.com
ScienceDirect

A new distributed alignment-free approach to compare whole proteomes

Umberto Ferraro Petrillo, Concettina Guerra, Cinzia Pizzi

PII:	S0304-3975(17)30529-7
DOI:	http://dx.doi.org/10.1016/j.tcs.2017.06.017
Reference:	TCS 11231
To appear in:	*Theoretical Computer Science*
Received date:	22 February 2017
Revised date:	12 June 2017
Accepted date:	14 June 2017

Please cite this article in press as: U. Ferraro Petrillo et al., A new distributed alignment-free approach to compare whole proteomes, *Theoret. Comput. Sci.* (2017), http://dx.doi.org/10.1016/j.tcs.2017.06.017

A new distributed alignment-free approach to compare whole proteomes

Umberto Ferraro Petrillo[a,1], Concettina Guerra[b], Cinzia Pizzi[c,*]

[a] *Dipartimento di Scienze Statistiche, Università di Roma "Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy.*
[b] *College of Computing, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA 30318, USA*
[c] *Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/A, 30131 Padova, Italy*

Abstract

Phylogeny inference has moved in recent years from the analysis of a single or few proteins to that of whole proteomes. However, the reconstruction of evolutionary trees for big number of species poses a significant computational challenge when using complete proteomes, even when relatively fast pairwise sequence comparison algorithms are used. We present a distributed approach that relies on the computation of distance measures based on maximal shared substrings within a bounded Hamming distance. The distributed system we built to implement this approach is flexible in that it supports a variety of design choices. It is based on the Spark framework and covers all the steps required by our approach, starting from the initial indexing of a set of FASTA sequences up to producing a report detailing the distances among these sequences, ranked according to a user-defined measure. Here we apply it to compare all proteins of selected organisms, divide them into groups and perform the comparisons within each group separately. The groups include: the functionally characterized proteins, the ribosomal proteins, and the unannotated proteins. We compute the average distances within the groups and evaluate their relationship and ability to capture the evolutionary closeness of organisms. We run experiments on selected species using a Hadoop computing cluster running Spark. The results show that the system implementing our approach is scalable and accurate.

Keywords: alignment free distances; average common substring; mismatches; distributed systems; bioinformatics.

1. Introduction

Phylogeny and classification of species have generally been based on the comparison of a single molecule (SSU RNA) or a few selected genes or proteins [1]. However, the selection of different genes sometimes produced conflicts in the reconstructed evolutionary trees, suggesting that a more comprehensive comparative analysis of whole genomes/proteomes was an important step towards reliable tree reconstructions [2]. As more genomes become available, this seemed to be finally

[*] Corresponding author

Email addresses: `umberto.ferraro@uniroma1.it` (Umberto Ferraro Petrillo), `guerra@gatech.edu` (Concettina Guerra), `cinzia.pizzi@dei.unipd.it` (Cinzia Pizzi)

[1] Part of this work was done while on a visit to College of Computing, Georgia Institute of Technology