ALGORITHMIC DATA FUSION METHODS FOR TUBERCULOSIS

By

Cagri Ozcaglar

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

Approved by the
Examining Committee:

_____
Bülent Yener, Thesis Adviser

_____
Kristin P. Bennett, Member

_____
Mohammed Zaki, Member

_____
Chris Bystroff, Member

_____
Qiang Ji, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2012
(For Graduation August 2012)

ABSTRACT

Exponentially-growing genomic data after the advent of gene sequencing technologies shifted the emphasis on to the analysis of many datasets from as many sources as possible. Data from multiple sources in the form of matrices and tensors can be analyzed separately, or they can be coupled and decomposed simultaneously. This data deluge is also observed in patient datasets of tuberculosis (TB), an infectious disease caused by *Mycobacterium tuberculosis* complex (MTBC). Epidemiologists, clinicians, and health care practitioners aim to find transmission routes, detect or rule out possible outbreaks, and control TB. For this purpose, patient isolates are routinely genotyped by multiple biomarkers which include spacer oligonucleotide types (spoligotypes) and Mycobacterial Interspersed Repetitive Units - Variable Number Tandem Repeats (MIRU-VNTR). Now it remains to make inferences from this data congestion. In this thesis, we propose algorithmic data fusion methods for tuberculosis using multiple sources of information from MTBC strains and TB patients.

In the first study, we propose the Tensor Clustering Framework (TCF) on multiple-biomarker tensors (MBT) and subdivide major lineages of MTBC into sublineages via genomic data fusion. The MBT holds data from two biomarkers, spoligotypes and MIRU patterns. We factorize the MBT into its component matrices using multiway models. Based on the component matrix of strain mode, we cluster MTBC strains into sublineages. Our new definition of sublineages based on two biomarkers confirms some of the existing sublineages, and suggests subdividing or merging other sublineages.

In the second study, we propose a new mutation model of spoligotypes based on both spoligotypes themselves and MIRU patterns. The model uses a maximum parsimony method based on three genetic distance measures on these two biomarkers. The resulting putative mutation history of spoligotypes depicted via a spoligoforest shows notable topological attributes. Number of descendant spoligotypes follows a power-law distribution. In addition, number of mutations at each spacer in the

DR region follows a spatially bimodal distribution. Based on this observation, we built two alternative models for mutation length frequency: Starting Point Model (SPM) and Longest Block Model (LBM). Both models plausibly fit mutation length frequency distribution in the spoligoforest.

In the third study, we propose the Unified Biclustering Framework (UBF) for host-pathogen association analysis of tuberculosis patients via genome-phenome data fusion. UBF is flexible in the sense that we can incorporate genetic distance between MTBC strains, spatial distance between TB patients, and time into domain knowledge, and factorize these joint datasets via coupled matrix-matrix and matrix-tensor factorization. We calculate feature pattern similarity matrix of (spoligotype, country) pairs and use it as input to our novel density-invariant biclustering algorithm. Finally, we select statistically significant biclusters using average best-match score. The resulting biclusters verify some of the well-known host-pathogen associations between MTBC strains and geographic distribution of their hosts, as well as suggest new patient-strain relationships.