Infection, Genetics and Evolution 12 (2012) 922–932

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Using the underlying biological organization of the *Mycobacterium tuberculosis* functional network for protein function prediction

Gaston K. Mazandu, Nicola J. Mulder *

Computational Biology Group, Department of Clinical Laboratory Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Medical School, 7925 Observatory, Cape Town, South Africa

ARTICLE INFO

Article history:
Available online 7 November 2011

Keywords:
Tuberculosis
Functional network
Data integration
Function prediction

ABSTRACT

Despite ever-increasing amounts of sequence and functional genomics data, there is still a deficiency of functional annotation for many newly sequenced proteins. For *Mycobacterium tuberculosis* (MTB), more than half of its genome is still uncharacterized, which hampers the search for new drug targets within the bacterial pathogen and limits our understanding of its pathogenicity. As for many other genomes, the annotations of proteins in the MTB proteome were generally inferred from sequence homology, which is effective but its applicability has limitations. We have carried out large-scale biological data integration to produce an MTB protein functional interaction network. Protein functional relationships were extracted from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database, and additional functional interactions from microarray, sequence and protein signature data. The confidence level of protein relationships in the additional functional interaction data was evaluated using a dynamic data-driven scoring system. This functional network has been used to predict functions of uncharacterized proteins using Gene Ontology (GO) terms, and the semantic similarity between these terms measured using a state-of-the-art GO similarity metric. To achieve better trade-off between improvement of quality, genomic coverage and scalability, this prediction is done by observing the key principles driving the biological organization of the functional network. This study yields a new functionally characterized MTB strain CDC1551 proteome, consisting of 3804 and 3698 proteins out of 4195 with annotations in terms of the biological process and molecular function ontologies, respectively. These data can contribute to research into the development of effective anti-tubercular drugs with novel biological mechanisms of action.

1. Introduction

In 1882, a German bacteriologist Robert Koch established the aetiology of TB, and stated (Koch, 1882): "If the importance of a disease for mankind is measured by the number of fatalities it causes, then tuberculosis must be considered much more important than other most feared infectious diseases, plaque, cholera and the like. If one only considers the productive middle-age groups, tuberculosis carries away one-third, and often more". This statement is still true after more than one and a quarter centuries. Indeed, according to the World Health Organization (WHO) (World Health Organization (WHO) Report, 2008, 2009), tuberculosis remains a potent infectious killer, as more than two billion people, corresponding to approximately one-third of the world's population, are infected with MTB, of which, one in ten develops active tuberculosis, affecting mostly young adults in their most productive years. The objective is to reduce the incidence of TB by half by 2015, thus a lot of work is required if we are to achieve the Millennium Development Goal of halting and beginning to reverse the spread of TB as one of the world's major diseases (Annan, 2005).

A number of drugs have been developed against TB (Global Tuberculosis Institute), including streptomycin (1943), p-aminosalicylic acid (1949), isoniazid (1952), pyrazinamide (1954), cycloserine (1955), ethambutol (1962), and rifampin (1963). This rapid succession of drugs with anti-TB activity was aimed at accelerating the exit of TB as a public health challenge. These drugs are of immense value for controlling the disease but have several shortcomings. The widespread emergence of drug resistant strains constitutes a major impediment to these global TB eradication programmes. Furthermore, the interactions between TB and Human Immunodeficiency Virus (HIV) or Acquired Immunodeficiency Syndrome (AIDS) have led to further challenges for anti-tubercular drug discovery. These require the effectiveness of coordinated strategies towards new anti-TB compounds with novel mechanisms of action.

The biggest step towards understanding MTB virulence and its specific abilities for invasion and division inside host macrophages

* Corresponding author. Tel.: +27 21 406 6058; fax: +27 21 406 6068.
E-mail address: nicola.mulder@uct.ac.za (N.J. Mulder).

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.10.027

and defeating the antibacterial mechanisms of these cells, was realized when the MTB genome sequence of H37Rv was published in 1999 (Cole et al., 1998). The global analysis of bacterial proteomes (Prentice, 2004; Ferretti et al., 2004) has yielded insights into functional features of many uncharacterized proteins. This has led to the elucidation of biological properties unique to MTB, such as its virulence, its slow growth and persistence, and the complexity of its cell wall, thus answering certain pressing and interesting questions about the pathogenicity of the tubercle bacillus. One isolated clinical strain of MTB, CDC1551, which is seen as highly transmissible and virulent for humans, was sequenced at the institute of Genomic Research (TIGR) (http://cmr.jcvi.org/tigr-scripts/CMR/GenomePage.cgi?org=gmt) in 2002 (Fleischmann et al., 2002). This offered the opportunity to compare the two genomes and several characteristics of these genomes that were previously unknown have been revealed (Marri et al., 2006). Unfortunately, about half of the identified proteins are labeled "uncharacterized" or "unknown" or "hypothetical" proteins, limiting the ability to exploit these data. In addition, this high proportion of proteins of unknown function in the MTB genome hampers the search for new drug targets and the advancement of research on this pathogenic organism. Therefore, there is an urgent need for predicted functional annotations for this large number of uncharacterized proteins. This has the downstream potential to enable researchers to apply these data to the search for new drug targets and thus for the development of novel and effective drugs with new biological mechanisms of action.

Finding functions of these uncharacterized proteins experimentally is likely to be difficult for several reasons. These include (Baldi and Brunak, 2001): (1) possible relationship of the function to the native environment in which a particular organism lives, (2) inclusion of many genes in the genome for securing its survival in a particular environment, which may have no use in the laboratory environment, and (3) it may even, in many cases, be almost impossible to imitate the natural host, with its myriad of other microorganisms, and thereby determine the exact function of a gene or gene product by experiment alone. The only effective route toward the elucidation of the function of uncharacterized proteins may be a combination of experimental approaches and predictions through computational analysis. To this end, sequence similarity search tools, such as Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990), have been extensively used for predicting functions of uncharacterized proteins. This approach is referred to as homology-based annotation transfer, providing a straightforward scheme of suggesting possible functions for uncharacterized proteins. The key assumption driving this approach is that two proteins with significantly similar sequences are evolutionary linked and might thus share common functions. However, some factors limit its applicability, for example, no known sequence may be similar to the novel protein sequence in the database. In this work, we are following a data integration approach where different protein functional interaction datasets are merged into a single functional interaction network. This network is used to characterize proteins of unknown function. Our system framework is described in Fig. 1, and follows these steps: (1) Generate Functional Interaction Networks enhanced by integrating data from heterogeneous sources (Homology-based, Genomic context and High throughput data); (2) use Gene Ontology (GO) (Ashburner et al., 2000; GO-Consortium, 2006; Dimmer et al., 2008; GO-Consortium, 2009) and prediction algorithms to predict functions of uncharacterized proteins based on the functional interaction networks. These predictions can provide a first hint about functionality that can later be subjected to experimental verification.

There are several goals that a function prediction algorithm needs to meet in the current genomic era. These include improvement of annotation quality and genomic coverage, i.e., to increase the proportion of genes or gene products in a genome which are annotated (Friedberg, 2006). Despite the high degree of noise that interaction data from high throughput experiments contains, making them potentially unreliable, uncontested successes have been recorded from the use of computational approaches to predict functions of uncharacterized proteins using these data. Several approaches have been proposed for predicting protein functions from functional networks and are mainly classified into two categories, namely global network topology and local neighborhood based approaches. Global network topology based approaches use global optimization (Vazquez et al., 2003; Tsuda et al., 2005; Nabieva et al., 2005), probabilistic methods (Troyanskaya et al., 2003; Deng et al., 2004; Letovsky and Kasif, 2003; Cho et al., 2008) or machine learning (Lanckriet et al., 2004; Chen and Xu, 2004; Xiong et al., 2006) to improve the prediction accuracy using the global structure of the network under consideration. In the case of local neighborhood based approaches, known as 'Guilt-by-Association', 'Majority Voting' or 'Neighbor Counting' (Schwikowski et al., 2000), direct interacting neighbors of proteins are used to predict protein functions.

The dualism of "Guilt-by-Association" and "Global" prediction approaches for characterizing a protein has raised a debate separating the Bioinformatics community into divergent groups with differing views. On one hand, there are proponents of the "Guilt-by-Association" strategy, stating that a gene or gene product shares the function of the most closely related genes of known functions, thus predicting protein functions by observing the patterns of each protein's neighborhood. This fraction highlights the

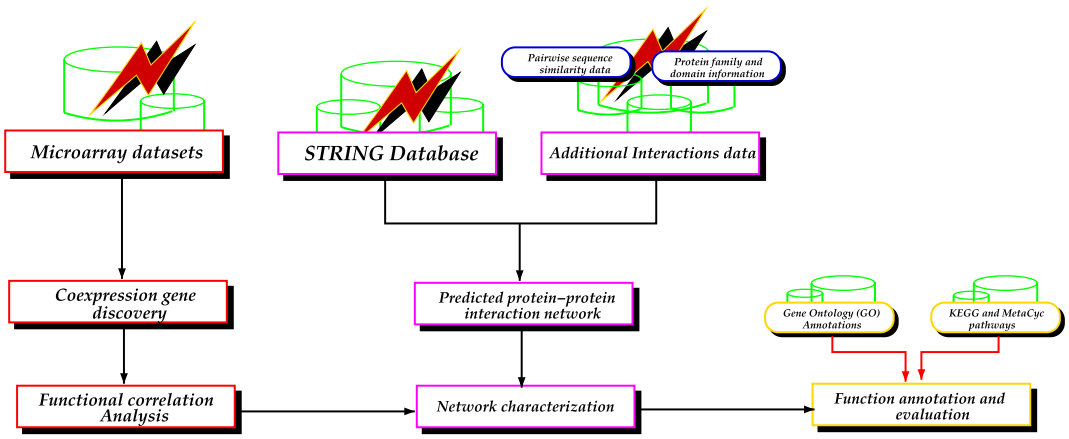

Fig. 1. System framework for protein function prediction.