Computational and Structural Biotechnology Journal 11 (2014) 1–10

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

Using biological networks to improve our understanding of infectious diseases

Nicola J. Mulder *, Richard O. Akinola, Gaston K. Mazandu, Holifidy Rapanoel

Computational Biology Group, Department of Clinical Laboratory Sciences, IDM, University of Cape Town Faculty of Health Sciences, Anzio Road, Observatory, Cape Town, South Africa

ARTICLE INFO

Available online 27 August 2014

Keywords:
Biological networks
Tuberculosis
Pathogen
Evolution
Protein–protein interaction

ABSTRACT

Infectious diseases are the leading cause of death, particularly in developing countries. Although many drugs are available for treating the most common infectious diseases, in many cases the mechanism of action of these drugs or even their targets in the pathogen remain unknown. In addition, the key factors or processes in pathogens that facilitate infection and disease progression are often not well understood. Since proteins do not work in isolation, understanding biological systems requires a better understanding of the interconnectivity between proteins in different pathways and processes, which includes both physical and other functional interactions. Such biological networks can be generated within organisms or between organisms sharing a common environment using experimental data and computational predictions. Though different data sources provide different levels of accuracy, confidence in interactions can be measured using interaction scores. Connections between interacting proteins in biological networks can be represented as graphs and edges, and thus studied using existing algorithms and tools from graph theory. There are many different applications of biological networks, and here we discuss three such applications, specifically applied to the infectious disease tuberculosis, with its causative agent *Mycobacterium tuberculosis* and host, *Homo sapiens*. The applications include the use of the networks for function prediction, comparison of networks for evolutionary studies, and the generation and use of host–pathogen interaction networks.

Contents

1. Introduction

The biology of organisms is complex and involves the interplay between numerous factors, including proteins, nucleic acids and small

* Corresponding author. Tel.: +27 21 4066058.
E-mail addresses: Nicola.mulder@uct.ac.za (N.J. Mulder), roakinola@gmail.com (R.O. Akinola), gmazandu@cbio.uct.ac.za (G.K. Mazandu), holy@aims.ac.za (H. Rapanoel).

http://dx.doi.org/10.1016/j.csbj.2014.08.006

molecules. These, in turn, are influenced by the environment and evolve to enable adaptation to environmental niches. Bacterial pathogens have evolved to infect their hosts through multiple mechanisms, including horizontal gene transfer [1], mutations [2], gene duplications [3] and gene loss [4]. In order to study infectious diseases caused by bacterial pathogens, we need to improve our understanding of the underlying molecular biology of these organisms so that we can determine how they infect, persist and cause disease, as well as better understand the pharmacokinetic and pharmacogenomic actions of anti-bacterial drugs.

The functioning of a biological system is largely driven by proteins, which interact and work together in pathways and processes. Therefore to understand the system, proteins must be studied within the context of their interactions with other proteins, rather than in isolation. Proteins can interact through direct physical binding, or through indirect associations, such as contributing to the same biological process. Protein–protein interaction networks are probably the most used example of biological networks, and can include interactions from both physical protein–protein binding as well as other functional interactions [5]. The vast amount of data generated over the years by different high-throughput biological technologies has raised the need for an integrative approach where datasets from heterogeneous sources are merged into a single network of interacting proteins. In these biological networks, the nodes are proteins and the edges represent functional interactions between proteins which can be derived from a variety of different data sources [6]. These sources include direct physical binding, for which there are a number of protein–protein interaction databases (e.g. IntAct, DIP, BIND), co-expression, functional similarity, text-mining, co-localization and other functional genomics data sources [6].

Biological networks provide the starting point for a number of analyses that aim to improve our understanding of biological systems [7]. Since biological networks are depicted as network graphs, many of these analysis tools draw on concepts and algorithms from graph theory. These allow us to, for example, determine the properties of nodes, such as their degree (number of neighbours), betweenness and centrality, which provide a feeling of how important that node is in facilitating communication between other nodes in the network and in holding connected components of the network together. We can also perform in silico knock-out studies to determine the potential impact of targeting a particular protein. Identifying the essentiality of proteins and the effect of knocking out the protein in the biological network of a pathogen has the potential to enable in silico prediction of potential drug targets when studying infectious diseases. There are many other applications of biological networks, and in this article we review some of these applications in studying human pathogens, using examples from our work on *Mycobacterium tuberculosis* and related mycobacteria. *M. tuberculosis* is the causative agent of tuberculosis (TB), an infectious disease of epidemic proportions in developing countries. First we review the use of protein–protein functional interaction (PPI) networks for protein function prediction (note, functional interactions include all functional connections between proteins, not only physical binding), and then we demonstrate how networks can facilitate evolutionary studies between pathogenic and non-pathogenic strains with differing genome sizes by comparing three different networks. Finally, we review some methods for generating host–pathogen interaction networks to improve our understanding of the interplay between host and pathogen during infection, not only using the *M. tuberculosis*–human interaction network as an example but also providing use cases from other host–pathogen studies.

2. Use of biological networks for function prediction

The completion of several sequencing projects and other high-throughput biological technologies has generated complete genome sequences and functional genomics data for several organisms. The abundance of these diverse biological data from various sources constitutes a rich source of knowledge, providing valuable insights into the dynamics driving collective and specific features of these organisms, and shedding light on the targeted organism's biology. Despite the uncontested successes recorded from comparative and functional genomics in gaining a better understanding of these organisms' biology and evolution, a number of challenges still remain. One of the main challenges is the lack of functional annotations for a relatively high proportion of genes and thus proteins within genomes. From 20 to 50% of genes within a genome are still annotated as 'unknown', 'uncharacterized' or 'hypothetical', and this limits our ability to exploit these data [8], leading to the paradigm of "a world which is data rich yet information poor". *M. tuberculosis* contains a large number of "uncharacterised" or "hypothetical" proteins, which limits our ability both to understand their role in the pathogenesis of TB and to determine their potential as drug targets.

Proteins perform an astonishing range of biological functions in an organism, including roles as structural proteins, as enzymes and for the transportation of materials within and between cells. Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function, implying that knowledge of protein functions and their biological pathways is crucial for understanding an organism's behaviour. Thus, one of the major tasks in the post-genomic era is genome annotation, or assigning functions to gene products in order to capitalize on the knowledge gained through different biological data produced. This requires a systematic description of the attributes of genes and proteins without any ambiguity using a standardized syntax and semantics in a format that is human readable and understandable, as well as interpretable computationally [9]. One of the biggest accomplishments in this area is the creation of the Gene Ontology (GO), which currently serves as the dominant and most popular functional classification scheme for annotation and functional representation of genes and their products [10].

The initial computational approach for assigning functions to an uncharacterized protein uses sequence similarity search tools, such as the Basic Local Alignment Search Tool (BLAST) [11]. This approach is referred to as homology-based annotation transfer, providing a straightforward scheme for suggesting possible functions for uncharacterized proteins. The key assumption driving this approach is that two proteins with significantly similar sequences are evolutionarily 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, and above all, the most significant database hit may perform a different function due to gene duplication events [12,13], domain shuffling events (deletions), or single point mutations [14]. Several approaches that do not rely directly on sequence similarity have also been implemented, which include using information about gene fusions, phylogenetic profiles of protein families, gene adjacency in genomes and expression patterns [15]. Below we describe the concept of and algorithms for function prediction and the use of GO and biological networks to achieve this.

2.1. Protein function and Gene Ontology

From a mathematical point of view, transference of a functional label from a set A to a set B is a rule which associates each object (input) 'x' in A with at most one object (output) 'y' in B. In this case, 'y' represents the realization of 'x', called a function of 'x'. For a function to be well-defined one needs to know the two sets A and B and the rule of associations of objects or realizations of all objects of A. Without loss of generality, a set is a collection of well-defined objects, and if A and B are well described, then a function is completely determined by knowing just the realizations of objects. Similarly, assuming the context and the scope of interest are known, protein function is a concept used to describe all types of realizations or activities to which the protein contributes, which take place within an organism, and which have consequences at the cellular and system levels [16]. Thus, the concept