Journal of Theoretical Biology 362 (2014) 9–16

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A survey on computational approaches to identifying disease biomarkers based on molecular networks

Guimin Qin [a,b], Xing-Ming Zhao [a,*]

[a] School of Electronics and Information Engineering, Tongji University, 4800 Caoan Highway, Shanghai 201804, China
[b] School of Software, Xidian University, No. 2 South Taibai Road, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form
3 June 2014
Accepted 4 June 2014
Available online 12 June 2014

Keywords:
Diseasome
Dysfunctional module
Dysregulated pathway
Interactome
Regulatome

ABSTRACT

The disease biomarkers can help make accurate diagnosis and therefore give appropriate interventions. In the past years, the accumulation of various kinds of 'omics' data, e.g. genomics and transcriptomics, makes it possible to identify disease biomarkers in a more efficient way. In particular, the molecular networks that describe the functional relationships among molecules enable the identification of disease biomarkers from a systematic perspective. In this paper, we surveyed the recent progress on the computational approaches that have been developed to identify disease biomarkers based on molecular networks. In addition, we introduced the popular resources about human interactomes and regulatomes as well as human diseasomes, whose availability makes it possible to predict the disease biomarkers with the utility of networks.

1. Introduction

Disease biomarkers are biological characteristics of the pathogenic processes, thereby providing an alternative way for accurate diagnosis and prognosis of diseases, where the early diagnosis of disease risks can help prevent the development of diseases while the precise prognosis of disease states can help make proper interventions. Furthermore, the biomarkers can help identify subtypes of heterogeneous diseases (e.g. breast cancer) so that appropriate treatments can be decided. Therefore, with the human genome available, the identification of molecular biomarkers is getting more and more attention with the promise of 'personalized medicine'. However, it is a very challenging task to find reliable and useful biomarkers for diseases considering more than 20,000 genes encoding about 30,000 proteins within the human genome, where complex interactions can be found among proteins. Recently, with the rapid progress in biotechnologies, especially in high-throughput technologies, the accumulation of various kinds of 'omics' (e.g. genomics, transcriptomics and proteomics) data makes it possible to identify molecular biomarkers that can predict disease risks more efficiently. For example, the genome-wide association study (GWAS) is able to identify genetic variants associated with common diseases (Wellcome Trust Case Control, 2007), and the transcriptome profiles monitoring the expression of tens of thousands of genes can successfully discriminate myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) that is otherwise a difficult problem in clinic (Golub et al., 1999).

Despite the success achieved by various kinds of omics data, the gene biomarkers they screened are generally not reliable, where the gene biomarkers identified based on one dataset sometimes fail to work in another dataset for the same disease. This phenomenon arises due to the fact that most diseases happen because of the dysregulation of sets of functionally interacting genes instead of the mutation of a single gene. Therefore, it is necessary to systematically take into account the interactions among genes when predicting gene biomarkers. The molecular networks, e.g. protein–protein interaction networks and gene regulation networks, can describe the interactions among genes/proteins in a natural way (Barabási and Oltvai, 2004), thereby providing an alternative way to predict gene biomarkers at the systematic level. It was found that the biomarkers predicted based on molecular networks are more accurate and robust (Barabási et al., 2011; Chuang et al., 2007; Ideker and Sharan, 2008; Taylor et al., 2009). For example, the intermodular hub proteins found from protein–protein interaction networks are found to be associated with oncogenesis and the altered modularity can be used for prognosis of breast cancer (Taylor et al., 2009). Compared with those single genes, the biomarkers identified from the molecular networks can provide insights into the molecular underpinnings of diseases, and help develop efficient therapeutic strategies (Barabási et al., 2011; Carter et al., 2013; Moreau and Tranchevent, 2012; Navlakha and Kingsford, 2010; Wang et al., 2011).

In this review paper, we surveyed the recent progress in the bioinformatics field about the computational approaches that have

* Corresponding author. Tel.: +86 2169583959; fax: +86 2169583959.
E-mail address: xm_zhao@tongji.edu.cn (X.-M. Zhao).

http://dx.doi.org/10.1016/j.jtbi.2014.06.007
0022-5193/

been developed to identify disease biomarkers based on molecular networks. We also introduced the popular resources about human interactomes and regulatomes as well as human diseasomes, which facilitate the prediction of the disease biomarkers by exploring molecular networks. Note that this survey aims to summarize the recent progress on biomarker identification based on molecular networks; however, it is by no means comprehensive due to the rapid evolvement of the field.

2. Human interactome and regulatome

With the advent of high-throughput technologies, more and more molecular interactions are being accelerated in various databases that can be publicly accessible. Accordingly, a large amount of human-specific molecular interactions are also available in recent years. These molecular interactions can be categorized into interactions with or without directions based on their biological nature. The undirected interactions are mainly protein–protein interactions, while the directed interactions are various kinds of regulations from regulators to their targets. Hereafter, we refer the human interactome to the human protein–protein interactions while the regulatome to those directed regulations.

2.1. Protein–protein interactome

Table 1 shows the popular public databases for human interactome, including protein binding, complex relationships, and pathway information. Based on these interactome data, the biological system can be described as a protein–protein interaction (PPI) network, where nodes are proteins while the edges represent the interactions among proteins (Rual et al., 2005; Stelzl et al., 2005). The protein–protein interactions can be determined in lab with high-throughput technologies, e.g. yeast two-hybrid and mass spectrometry (MS). However, it is expensive to detect those interactions with experimental technologies. Furthermore, it is found that the interactions detected by different labs for the same species have low consistency (Yu et al., 2008). Under these circumstances, the human-specific interactome maps remain incomplete and noisy. Therefore, some computational approaches have been proposed to predict molecular interactions and make the interactome maps more comprehensive (McDowall et al., 2009; Mohamed et al., 2010; Zhao et al., 2010; Zhao et al., 2009a).

With these interactome maps available, the functional roles of genes/proteins can be investigated. For example, the protein–protein interaction networks are found to be scale-free networks, and the hub proteins have more important roles than other proteins (Han et al., 2004). Another interesting finding is that disease genes have no tendency to encode hub proteins and interacting proteins tend to be associated with same/similar diseases (Goh et al., 2007). In addition, the interactome maps enable the identification of condition-specific pathways and detection of aberrant pathways (Liu et al., 2012a, 2012b; Zhao et al., 2008; Zhao et al., 2009b), thereby providing insights into the molecular mechanisms underlying diseases.

2.2. Human regulatome

Except for the undirected interactomes mentioned above, there are some kinds of regulatomes that describe the interactions between various regulators and their targets, and these regulatomes can be described as directed graphs. For example, in the gene regulation network, the nodes denote either transcription factors (TFs) or genes and the links between nodes denote the regulations of transcription factors on corresponding target genes. Note that each TF may regulate multiple genes while each gene may be regulated by more than one TF. In the post-transcriptional regulation network, the nodes denote either microRNAs (miRNAs) or genes while the edges denote the post-transcriptional regulations of miRNAs on their corresponding target genes. The metabolic network is composed of metabolites with edges denote either the reactions that convert one metabolite into another or the enzymes that catalyze these reactions (Duarte et al., 2007; Jeong et al., 2000). Table 2 lists the popular regulatomes deposited in publically accessible databases.

Similar to the interactomes, the regulatomes are far from complete right now. Therefore, a lot of bioinformatics approaches have been developed to predict the regulatomes. For example, we have proposed a novel algorithm, namely NARROMI, to predict the gene regulations (Zhang et al., 2013). Some approaches have been proposed to predict the translational modification sites as well as the substrates of kinases (Linding et al., 2007; Song et al., 2012; Sun et al., 2013), while some are developed to construct the

Table 1
Popular human protein–protein interaction databases.

Database	Website	Description
BioGRID	http://thebiogrid.org	Collections of physical and genetic interactions for major model organisms
DIP	http://dip.doe-mbi.ucla.edu	The database contains experimentally determined interactions between proteins
HPRD	http://www.hprd.org	A database about human proteins and their interactions
IntAct	http://www.ebi.ac.uk/intact	A database of freely available molecular interactions
IRView	http://ir.hgc.jp/	A database about interacting regions in protein sequences
MINT	http://mint.bio.uniroma2.it/mint	Collection of molecular interactions extracted from literature
MIPS-MPPI	http://mips.helmholtz-muenchen.de/proj/ppi/	A collection of manually curated high-quality mammalian PPI data collected from the scientific literature by expert curators
NetPath	http://www.netpath.org/	A public resource of curated signal transduction pathways
ConsensusPathDB	http://cpdb.molgen.mpg.de/	A database consists of various interactions for human, e.g. protein–protein interactions, metabolic interactions, and drug–protein interactions
STRING	http://string.embl.de/	A database about functional interactions for various organisms
UniHI	http://www.unihi.org/	A comprehensive database of computationally predicted and experimentally determined human protein interactions
BioCyc	http://biocyc.org	A collection of 2988 pathway/genome databases
KEGG	http://www.genome.jp/kegg	A resource about high-level functional information for distinct organisms
SMPDB	http://www.smpdb.ca/	An interactive, visual database containing small molecule pathways found in humans
NCI Nature –Pathway Interaction Database	http://pid.nci.nih.gov	A database about expert-reviewed pathways curated by Nature Publishing Group
IPAVS	http://ipavs.cidms.org	A collection of manually curated human pathway data
Reactome	http://reactome.org	A free, open-source, curated and peer reviewed pathway database