Computational Biology and Chemistry 59 (2015) 139–149

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Principles for the organization of gene-sets

Wentian Li*, Jan Freudenberg[1], Michaela Oswald

The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY, USA

ARTICLE INFO

Article history:
Received 24 February 2015
Accepted 8 April 2015
Available online 10 June 2015

Keywords:
Gene-sets
Gene families
Co-localization
Operon
Protein complex
Protein domains
Protein–protein interaction
Transcription factor target
Gene Ontology (GO)
Biological pathways
Co-expression
Co-differential expression
Essential genes
Housekeeping genes
Tissue-specific genes
Disease genes
Modules

ABSTRACT

A gene-set, an important concept in microarray expression analysis and systems biology, is a collection of genes and/or their products (i.e. proteins) that have some features in common. There are many different ways to construct gene-sets, but a systematic organization of these ways is lacking. Gene-sets are mainly organized ad hoc in current public-domain databases, with group header names often determined by practical reasons (such as the types of technology in obtaining the gene-sets or a balanced number of gene-sets under a header). Here we aim at providing a gene-set organization principle according to the level at which genes are connected: homology, physical map proximity, chemical interaction, biological, and phenotypic-medical levels. We also distinguish two types of connections between genes: actual connection versus sharing of a label. Actual connections denote direct biological interactions, whereas shared label connection denotes shared membership in a group. Some extensions of the framework are also addressed such as overlapping of gene-sets, modules, and the incorporation of other non-protein-coding entities such as microRNAs.

1. Introduction

The reductionist approach to biology (Crick, 1966) attempts to explain a phenotype by its lowest possible entities, usually on the level of genes, or single-nucleotide-base mutations (point mutation, single nucleotide polymorphism). Living organisms, however, are multi-level systems. Many phenotypes and medical conditions can be caused by alterations in multiple genes that are related to each other on a higher level. This is the reason for the interest in gene-sets from a biomedical research perspective.

Many terms describing the concept of gene-sets exist in the literature. In microarray data analysis, a gene *signature* (van de Vijver et al., 2002; Liu et al., 2007; Sortiriou and Piccart, 2007; Schramm et al., 2012; Sanz-Pamplona et al., 2012; Kuner, 2013; Chibon, 2013) refers to a group of genes that are differentially expressed in a patient group as compared with the control group. A *meta-signature* is a generalization of expression signatures across multiple datasets (Rhodes et al., 2004; Shen et al., 2004; Mistry and Pavlidis, 2010; Verweij and Vosslamber, 2011). *Co-differential-expressed* gene-sets is another name for signatures, and the name *multivariate differential expression* also appears in the literature (Szabo et al., 2003; Xiao et al., 2004; Lu et al., 2005; Nilsson et al., 2007). In control samples or in time-course expression data from model organisms, *co-expressed* genes refer to groups of genes whose expression move up and down in synchronized manner (or are positively correlated), typically used in the study of transcription regulation (Boutanaev et al., 2002) with the implication of co-regulation (D'haeseleer et al., 2000; Yeung et al., 2004). Genes that cluster together in patient-only samples are co-expressed genes in the sense of being co-expressed in multiple samples. Biclustering is an extension of clusters in gene space to clusters in both gene and sample spaces (Cheng and Church, 2000; Tanay et al., 2002; Prelić et al., 2006; Eren et al., 2013).

* Corresponding author. Tel.: +1 516 562 1076.
E-mail addresses: wli@nshs.edu, wtli2012@gmail.com (W. Li).
[1] Current address: Regeneron Genetics Center, Regeneron Pharmaceuticals, Inc., Tarrytown, NY, USA.

http://dx.doi.org/10.1016/j.compbiolchem.2015.04.005
1476-9271/

Besides data-based gene-sets, there are also biological knowledge-based gene-sets. Among them, two are broadly used: Gene Ontology (GO) (The Gene Ontology Consortium, 2000; Rhee et al., 2008; du Plessis et al., 2011; Huntley et al., 2014) and biological pathways (Ogata et al., 1999; Khatri et al., 2012). GO concerns the what, how, and where questions of gene products: molecular function (MF), biological process (BP), and cellular component (CC). A biological pathway is comparable to the biological process part of GO category, and a pathway can be typically in one of these three types: metabolic, signaling, and regulatory. However, GO and pathway are two conceptually different approaches in grouping genes. GO puts a label to a gene, but does not necessarily place the gene in one specific process.

To make things even more complicated, various types of pathways can be coupled to form networks or a cascade of processes. For example, a signaling pathway can be activated by a signal from outside the cell membrane, which in turn activates the transcription of certain genes, which in turn produces proteins to participate in a metabolic process. Pathways represented by a graph provide more detailed information than the biological process GO category, as it specifies the relative order of genes along a path, as well as the reaction direction.

A large collection of gene-sets (more than 10,000) reside in the *Molecular signatures database* (MSigDB, release 4.0, http://www.broadinstitute.org/gsea/msigdb/) (Subramanian et al., 2005; Liberzon et al., 2011). These gene-sets are organized in seven catalogs: C1 for positional gene-sets, C2 for curated gene-sets such as pathway, C3 for conserved regulation motifs, C4 for microarray-based gene clusters, C5 for gene-ontology annotated gene-sets, C6 and C7 are more specific microarray-based signature, one for cancer and another for immunology. Another comprehensive resource is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2013), where a KEGG MODULE contains manually maintained biologically functional units. These modules are grouped into four types: pathway modules, structural complexes, functional sets, and signature modules.

Despite the putative comprehensiveness of these databases, group header names are often given by operational convenience, not according to some organizational principle. Take GO annotations for example, Gene Ontology actually consists of three ontologies, based on very different principles. The cellular location of a gene product groups them by their physical proximity, whereas the molecular function of a gene product is highly correlated with the biochemical properties of the protein. Gene-sets obtained from mRNA profiling by microarray provide another example for multiple organizing principles. Many expression profiling experiments in model organisms target a specific biological process, whereas those in human disease studies encompass consequences of many biological processes, some of which could have taken place years ago in patients' history.

The conceptual label in GO versus the actual interaction in pathways illustrates one important aspect in gene-set organization. Genes can be grouped based on actual connections or by sharing a common label as shown in Fig. 1. In Fig. 1(A), a line linking two genes represents an actual pairwise relationship. When these pairwise relationships are extended to include multiple genes (either fully connected, as in set-1, or "spanning", as in set-2), gene-sets can be constructed. In Fig. 1(B), genes sharing a same label (e.g. solid circles for set-1, solid triangles for set-2) are grouped together.

Another important aspect in gene-set organization concerns the level at which two genes are connected. We will explore the following levels: the similarity/homology level, physical level, biochemical level, biological level, and phenotypic/medical level. In some text mining programs, genes are considered to be connected when they are co-cited in the literature. The level at which the co-citation occurs has been used as group/category headers: co-cited

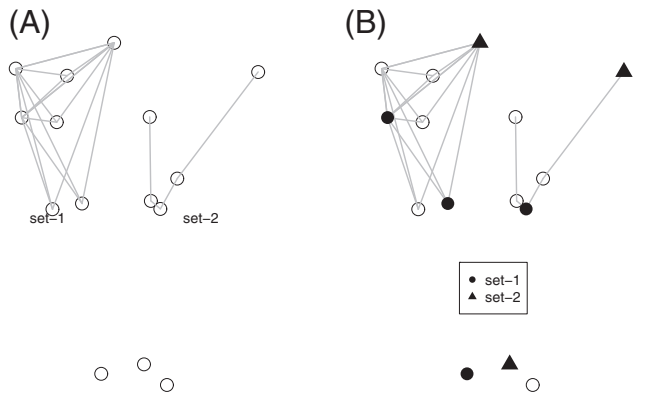

Fig. 1. Illustration of some relationship between genes (dots). (A) Two genes may be related at a specific level (e.g. chemical interaction) and that relationship is represented by a line. A group of genes with high degree of inter-connectivity can be the basis of a gene-set (set-1). For other type of gene-set, only one connection to another gene is enough, with the overall degree of inter-connectivity being low (e.g. pathway) (set-2). (B) Another type of gene-sets consists of genes with the same label: black dots (set-1) and black triangles (set-2).

in the same abstract, same sentence, same function word list, plus the "gene-functionword-gene" level and the expert level (Epple and Scherf, 2009). These multiple level groupings provide an interesting comparison to what we plan to discuss in this paper.

In this paper, we will argue that when the above two key points are considered in our organization principle, i.e. actual link versus shared concept, and level of the connection, we can organize most, if not all, gene-sets discussed in the literature. This paper is partly a review, and partly a perspective. Though we will use many public-domain databases to illustrate various types of gene-sets, a comprehensive listing of all databases is not an intended goal here. Knowledge-based gene-sets are usually taken for granted, and they are often used as being definitive. In reality, pathways can still be fuzzy and not as definitive as we would believe. Through the discussion in this paper, we hope that more information should be provided in any given group of linked genes, such as at what level a gene-set is constructed, in what sense genes in a set are associated, and how to delineate a gene-set with what inclusion/exclusion criteria.

2. Gene-set organization principle at different levels

2.1. Homology level

On the most basic level, two genes are related if their DNA sequences are similar. The widespread existence of similar DNA sequences in the genome is the consequence of duplication (Ohno, 1970; Taylor and Raes, 2004; Li et al., 2014). Genes duplicates originate from a common ancestor and can evolve to perform different functions by mutations and become a multigene family (Ohta, 1980; Walsh and Stephan, 2001). Within a genome, all genes that derived from duplication events are called paralogs (Koonin, 2005), whereas the general term homologs refers to those derived from the common ancestor. The distinction between different forms of homology was made in Fitch (1970). Detection of genes belonging to a multigene family is typically by sequence alignment. The paralog genes in *Ensembl*, for example, are labeled by *within_species_paralog* (Vilella et al., 2009). If duplication is followed by a degradation and one of the duplicated genes is not maintained as a functional copy, it is a pseudogene. Besides duplication, retrotransposition is another mechanism for producing (processed) pseudogenes (Vanin, 1985; Pei et al., 2012).

Using the one-to-four rule (single copy of a gene in invertebrates but four copies in mammals), duplicated genes resulted from the