Computers in Biology and Medicine 55 (2014) 49–52

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Degrees of separation as a statistical tool for evaluating candidate genes

Ronald M. Nelson *, Mats E. Pettersson

Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 22 May 2014
Accepted 1 October 2014

Keywords:
Candidate genes
Network analysis
GWAS
Degree of separation
Gene–gene interaction

ABSTRACT

Selection of candidate genes is an important step in the exploration of complex genetic architecture. The number of gene networks available is increasing and these can provide information to help with candidate gene selection. It is currently common to use the degree of connectedness in gene networks as validation in Genome Wide Association (GWA) and Quantitative Trait Locus (QTL) mapping studies. However, it can cause misleading results if not validated properly. Here we present a method and tool for validating the gene pairs from GWA studies given the context of the network they co-occur in. It ensures that proposed interactions and gene associations are not statistical artefacts inherent to the specific gene network architecture. The **CandidateBacon** package provides an easy and efficient method to calculate the average degree of separation (*DoS*) between pairs of genes to currently available gene networks. We show how these empirical estimates of average connectedness are used to validate candidate gene pairs. Validation of interacting genes by comparing their connectedness with the average connectedness in the gene network will provide support for said interactions by utilising the growing amount of gene network information available.

1. Introduction

Interaction databases offer researchers with the opportunity to validate candidate genes identified from any of a wide range of analytical methods such as genome wide association studies (GWAS), linkage mapping and selective sweep analysis [1]. Once a set of regions of interest has been found using one of the methods above, the genes in the regions are typically screened for known interactions with each other or previously known candidate genes (see [1] and references within). If interactions are found, that is taken as support for the validity of the association-, linkage- or sweep signal [2].

However, the *small world phenomenon* that is inherent to most networks, including those describing gene–gene interactions [3–5], makes this approach problematic, since all genes are connected to each other. Thus, it is not sufficient to only find links between possible candidate genes (i.e. guilt by association). Those links must be evaluated based on the context of the network they are derived from [4,6–10].

We propose that the degree of separation (*DoS*) is a good metric to validate the likelihood of genes being truly associated. For example, two genes interacting with one intermediary may have a measurable biological effect on each other's function. On the other hand, when two genes are connected, via a path of 10 other intermediates, they are less likely to have measurable effects on each other function. The length of the shortest route between two genes, relative to the average path length between any random genes in the network are therefore an estimate of the likelihood that these genes are associated by chance alone.

Evaluating the *DoS* is not currently part of the normal analysis pipeline in GWA studies. This is becoming more important as analyses of expression data is used where eQTL are linked (often with several intermediate steps) to the expression phenotype. Additionally, as the connections in gene networks become denser, it is important to validate the connection between any two genes is not merely sporadic. For this purpose, we present an R package **CandidateBacon** that can, for an arbitrary network of pair wise interactions, provide both a measure of the average connectedness among a set of candidate genes and a network-specific distribution of the connectedness of random gene pairs. This distribution can then be used to assign a *p*-value to the observed connectedness of the candidate set, which is indicative of whether or not the observed links actually serve to validate the original associations findings, or simply reflect the general connectedness of the network.

Since the rapid accumulation of biological interaction data through large-scale experiments means networks of known interactions expand on a regular basis, the degree of connectedness between a

* Correspondence to: Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala 75007, Sweden. Tel.: +46700719244.
E-mail address: ronnie.nelson@slu.se (R.M. Nelson).

http://dx.doi.org/10.1016/j.compbiomed.2014.10.004
0010-4825/

Table 1
Network information from analysed BioGRID data. Empirical estimates of network connectedness (*DoS*) obtained by using CandidateBacon.

Organism	Connections	Node interactions			*DoS*	Theoretical *DoS*
		Maximum	Mean	Unconnected (%)		
Arabidopsis thaliana	5922	438	4.58	10.35	4.85	5.71
Bos taurus	175	12	1.62	92.76	4.03	10.67
Caenorhabditis elegans	3642	526	3.82	9.91	4.35	6.12
Danio rerio	147	12	2.14	91.86	2.78	6.58
Drosophila melanogaster	8157	175	9.21	2.50	4.13	4.06
Gallus gallus	211	110	1.80	69.82	2.04	9.14
Homo Sapiens	15319	9536	10.81	44.73	2.86	4.05
Human Herpesvirus 4	201	152	1.93	41.50	1.99	8.06
Human Immunodeficiency Virus 1	381	265	2.17	0.00	3.03	7.69
Mus musculus	5221	295	3.71	13.84	4.90	6.53
Rattus norvegicus	1488	189	2.32	36.50	5.59	8.70
Saccharomyces cerevisiae	6252	2584	67.17	0.08	2.42	2.08
Saccharomyces cerevisiae (*low*)*	4946	265	13.79	3.98	3.80	3.24
Saccharomyces cerevisiae (*high*)*	5822	2572	63.89	0.00	2.38	2.09
Schizosaccharomyces pombe	2702	304	11.97	3.21	3.55	3.18
Xenopus laevis	387	35	2.22	73.47	5.62	7.45

* Two additional analyses were performed on the high- and low-throughput interactions respectively in the *Saccharomyces cerevisiae* data.

given set of genes will change over time [11]. The **CandidateBacon** package provides an extremely fast efficient way to obtain the average *DoS* (i.e. the average shortest path between any two genes) in the currently known network, and relate it to the properties of that specific network, which is a necessity for validation.

2. Materials and methods

We used organism-wide network data from the BioGRID (Release 3.1.94, compiled on October 2012) repository to evaluate the different levels of connectedness. We included all species with networks having more than 100 interactions (both physical and genetic interactions included) in the analyses (Table 1). Additionally the *Saccharomyces cerevisiae* network was also divided into two datasets separated on high- and low-throughput interactions, respectively. The *degree_of_separation* function was repeated 10,000 times for each network, and the network-specific *DoS* distribution was calculated from the result (Table 1).

We used the package **CandidateBacon** to find the shortest path between two nodes in each network. Using a standard computer (2.66 GHz Intel), it takes, on average, less than 0.3 s to find the shortest path between any two nodes in a large network and complicated (e.g. *S. cerevisiae*). The package was specifically develop to be easy to use for biological networks in the R framework and is freely available at: http://www.computationalgenetics.se/?page_id=7. A full description of the algorithm is available in the package documentation (Supplementary material 1).

In order to show the importance of obtaining an empirical distribution from each network, we compared the observed empirical distributions with theoretical *DoS*. The theoretical *DoS* is calculated using the following equation:

$$DoS = \frac{\ln N}{\ln K}$$

where N is the total number of nodes in the network and K the average number of interactions per node [3]. K is calculated by dividing the total number of interactions (i.e. node pairs) in the network by the total number of nodes in the network. Node hubs and unequal distribution of connections between nodes are thus corrected for in the theoretical *DoS*.

In order to provide a working example of the intended use of the package, we used the list of candidate genes for *S. cerevisiae* reported in Ref. [12] (Supplementary material 2). The proposed interactions were evaluated using network data from BioGrid, as cited above. The average *DoS* for this set of genes is 1.7 nodes per interaction pair. A *t*-test indicated that this is significantly less than random interactions, which have an average of 2.42 (p-value = 0.008). Thus, in this case the detected interactions do lend support that the GWA signals are above the random expectation.

3. Results and discussion

Comparison between the different networks indicates that the *DoS* distribution is different in each network, and thus needs to be evaluated before candidate genes can be validated using that network (Fig. 1 and Table 1). In accordance with that, the theoretical estimates of connectedness within a network are often very different from empirical observations. This is clearly visible in Table 1 where the theoretical estimates of connectedness are often very different from the observed connectedness. Another useful metric provided is the number of unconnected genes within each of the networks. This metric, in addition to the *DoS*, can also be used to support biological interactions if a network with many sub-networks is evaluated. E.g. *Danio rerio*: more than 90% of any two random genes are in separate networks. Given that this is the current information available for this network, the significance of finding several pair of specific genes connected in the same network can be estimated.

It is clear from the data that there is variation between the networks, due to their differing underlying structure. This is also true within a single species, as can be seen when comparing the *S. cerevisiae* high-throughput and low-throughput datasets. The sensitivity in the methods used to score the interactions, have an influence on the resulting network. This further strengthens the point that any detected interactions must be validated in their specific context, if they are to provide support for any detected statistical association.

Evaluation of gene lists with the average connectedness does however provide researchers with a statistical estimate of the validity of candidate genes. For example, the list of candidate genes reported in [12], has a smaller average *DoS* than random gene pairs. It indicates that the genes reported in the analyses are more connected than random and adding this statistical support increase the confidence that true causal genes were discovered. This type of support is important for the follow-up of GWA studies, which often involve substantial investments into the exploration of candidate genes.