Bioinformatics, 2018, 1–8
doi: 10.1093/bioinformatics/bty691
Advance Access Publication Date: 9 August 2018
Original paper

OXFORD

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

Systems biology

On entropy and information in gene interaction networks

Z. S. Wallace[1], S. B. Rosenthal[2], K. M. Fisch [2], T. Ideker[3] and R. Sasik [2,*]

[1]Department of Mathematics, Tufts University School of Arts and Sciences, Medford, MA 02155, USA, [2]Department of Medicine, Center for Computational Biology and Bioinformatics, University of California San Diego, La Jolla, CA 92093, USA and [3]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 8, 2018; revised on June 14, 2018; editorial decision on August 6, 2018; accepted on August 8, 2018

Abstract

Motivation: Modern biological experiments often produce candidate lists of genes presumably related to the studied phenotype. One can ask if the gene list as a whole makes sense in the context of existing knowledge: Are the genes in the list reasonably related to each other or do they look like a random assembly? There are also situations when one wants to know if two or more gene sets are closely related. Gene enrichment tests based on counting the number of genes two sets have in common are adequate if we presume that two genes are related only when they are in fact identical. If by related we mean well connected in the interaction network space, we need a new measure of relatedness for gene sets.

Results: We derive entropy, interaction information and mutual information for gene sets on interaction networks, starting from a simple phenomenological model of a living cell. Formally, the model describes a set of interacting linear harmonic oscillators in thermal equilibrium. Because the energy function is a quadratic form of the degrees of freedom, entropy and all other derived information quantities can be calculated exactly. We apply these concepts to estimate the probability that genes from several independent genome-wide association studies are not mutually informative; to estimate the probability that two disjoint canonical metabolic pathways are not mutually informative; and to infer relationships among human diseases based on their gene signatures. We show that the present approach is able to predict observationally validated relationships not detectable by gene enrichment methods. The converse is also true; the two methods are therefore complementary.

Availability and implementation: The functions defined in this paper are available in an R package, *gsia*, available for download at https://github.com/ucsd-ccbb/gsia.

Contact: rsasik@ucsd.edu

1 Introduction

Gene expression studies based on microarrays (Schena *et al.*, 1995) and RNA sequencing technologies (Morin *et al.*, 2008) typically produce lists of genes related to the studied biological phenotype. Other genomic-scale experiments, even though not necessarily gene centric, such as genome-wide association studies (Klein *et al.*, 2005), chromatin immunoprecipitation studies (Robertson *et al.*, 2007) or DNA methylation studies (Cokus *et al.*, 2008), also frequently summarize their findings as gene lists. The gene list usually has a measure of statistical significance assigned to it, such as the false discovery rate *fdr* (Benjamini and Hochberg, 1995). Each gene on the list may also have an individual measure of statistical

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

significance, either a *P*-value, a *q*-value (Storey and Tibshirani, 2003) or local false discovery rate *lfdr* (Efron, 2008). Gene lists can be interrogated biologically by performing gene set enrichment analyses against curated gene sets, such as KEGG (Kanehisa *et al.*, 2004), Gene Ontology (Harris *et al.*, 2004) and others, all found conveniently under one roof in MSigDB (Subramanian *et al.*, 2005). Gene set enrichment methods depend on exact gene matching; i.e. two gene sets are considered significantly related when they have enough genes in common. A logical extension is that two genes are related only when they are in fact identical.

A different kind of relationship can be defined in the context of a gene interaction network: two genes, represented by nodes of the network, are related when they are closely connected by edges of the network, preferably along multiple paths. This relationship can be quantified using physical models of the network and their properties. Two examples are the heat diffusion model (Köhler *et al.*, 2008) and the electrical conduction model (Klein and Randić, 1993). In the heat diffusion model, heat propagates along the edges of the network. Two nodes are considered related if enough heat reaches one node from the other. A very popular variant of this model is the network propagation model [for review see Cowen *et al.* (2017)], which shares some qualitative properties with heat diffusion but does not conserve energy. In the electrical conduction model, the edges of the network are modeled as electrical resistors. Two genes are related when the effective resistance between the corresponding nodes is small. Electrical resistance between two nodes is also a true distance, which is a very attractive property.

There is a deep connection between resistance distance on a network and statistical mechanics of a set of interacting one-dimensional harmonic oscillators (Estrada and Hatano, 2010). The energy function of the oscillator system has translational symmetry (does not depend on the position of the center of mass). This model is directly applicable to vibrations of complex organic molecules moving freely in a solution, in which case the molecule's atoms themselves are the oscillators. Electrical resistance between two nodes in a corresponding resistor network can be written as the thermal average of the relative square displacement of the two corresponding oscillators. However, the resistor model is pathological in the sense that the entropy of the system diverges. This fact is intimately tied to the translational symmetry of the model.

Here we show, starting from some very basic assumptions, that gene expression levels in a cell can also be represented by a formally similar set of interacting one-dimensional harmonic oscillators. In this case, however, translational symmetry is broken and individual fluctuations remain finite as demanded by everyday experience. As a consequence, entropy of any set of these oscillators is finite. We use this entropy to derive interaction information for a single gene set, as well as mutual information between two gene sets. The utility of mutual information as a network-based similarity measure between gene sets has been recognized earlier by Chuang *et al.* (2007) These authors define mutual information heuristically; here it follows naturally from the statistical properties of the associated harmonic oscillator model. We say that genes or gene sets are related in the network sense when they are mutually informative.

To formalize our task, we will give quantitative answers to these questions:

1. [Interaction information] Are genes in set $\mathcal{A}$ significantly mutually informative? Could genes in set $\mathcal{A}$ have been generated by random selection?
2. [Mutual information] Given two gene sets $\mathcal{A}$ and $\mathcal{B}$, is gene set $\mathcal{B}$ significantly informative of $\mathcal{A}$? Could genes in set $\mathcal{B}$ have been generated randomly and independently from genes in set $\mathcal{A}$?

2 Materials and methods

We begin with a simplified physical model of a cell in its microenvironment. The state of the cell is simplistically described by the set expression levels of all biomolecules (mRNA, proteins, lipids, metabolic intermediaries, etc.) $\{e_i, i \in \mathcal{N}\}$, where $\mathcal{N}$ is the set of all expressed molecular species. Let us denote these collectively by a vector $\mathbf{e}$. The microenvironment is described by a set of external fields, such as temperature, pH, mechanical stress, concentration of nutrients and other molecules such as electrolytes, receptor ligands or man-made molecules (drugs). These will be referred to collectively as $\mathbf{h}$. We assume that $\mathbf{h}$ is such that homeostasis is possible and therefore a stable steady state $\mathbf{e}^0$ exists. It is a dynamical equilibrium state, possible only for as long as the flow of energy, nutrients and other molecules is maintained. Consequently, a small transient deviation of $\mathbf{e}$ from equilibrium is followed by restoration of equilibrium. We postulate therefore that the equilibrium is a local minimum of some multivariate 'potential energy' function $V(\mathbf{e}, \mathbf{h})$. Since different cell types express a different set of biomolecules, V must be cell type specific. Close to this minimum, with $\mathbf{h}$ fixed, we can write

$$V(\mathbf{e}) = V(\mathbf{e}^0) + \frac{1}{2}\sum_{i,j}\frac{\partial^2 V}{\partial e_i \partial e_j}\Big|_{\mathbf{e}^0}(e_i - e_i^0)(e_j - e_j^0) + \cdots \quad (1)$$

Defining normalized expression levels as $y_i \equiv e_i/e_i^0$, dropping the constant (it is irrelevant for our purposes) and neglecting the higher-order terms, Equation (1) becomes

$$V(\mathbf{y}) = \frac{1}{2}\sum_i A_{ii}(y_i - 1)^2 + \sum_{i<j} A_{ij}(y_i - 1)(y_j - 1) \quad (2)$$

The off-diagonal elements A_{ij} are non-zero only when genes i and j interact, and are positive when i inhibits expression of j and negative when i induces expression of j. We only consider $A_{ij} < 0$ in the following. We anticipate $\mathbf{A}$ to be a sparse matrix with structure corresponding to the topology of the gene-gene interaction network. In the absence of quantitative knowledge, we set $A_{ij} \equiv -A_1 < 0$ (for i, j interacting, 0 otherwise). Equation 2 becomes

$$V(\mathbf{x}) = \frac{1}{2}\sum_i (A_{ii} - A_1 d_i)x_i^2 + \frac{A_1}{2}\sum_{i<j}(x_i - x_j)^2. \quad (3)$$

where $\mathbf{x} \equiv \mathbf{y} - 1$ are deviations from equilibrium and d_i is the degree of node i in the network. Since the expansion (1) was away from equilibrium, the quadratic form $V(\mathbf{x})$ must be positive definite. The form (3) makes it explicit that it will be so if $A_{ii} - A_1 d_i > 0$ for all i. In the absence of specific knowledge, we adopt a simple prescription $A_{ii} - A_1 d_i = A_0 > 0$. We now write

$$V(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}, \quad (4)$$

where

$$\mathbf{A} = A_0\mathbf{I} + A_1\mathbf{L}, \quad (5)$$

with $\mathbf{I}$ the identity matrix and $\mathbf{L}$ the network Laplacian. Equation (4) represents the system of interacting one-dimensional harmonic oscillators, in which the oscillator coordinates are deviations of normalized gene expression levels from their steady-state values.

We would like to know which genes respond to perturbations in a coordinated manner. These will likely be genes who are directly

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

interacting, but also genes who are not directly interacting, as long as there are paths of physical interactions that connect them in the network. Note that (4) formally represents a system of interacting harmonic oscillators. We can investigate all possible perturbations within the framework of statistical mechanics, by immersing the system in a heat bath of temperature $1/\beta$ (this temperature is a theoretical device and is not related to the actual temperature of the cell). The probability of finding the system in a state $\mathbf{x}$ is given by the Boltzmann factor $P(\mathbf{x}) = \frac{1}{Z}\exp[-\beta V(\mathbf{x})]$, where Z is a normalization constant (partition sum). As the energy is a quadratic form of the degrees of freedom, $P(\mathbf{x})$ is a multivariate Gaussian distribution and the model is exactly soluble. All thermodynamic expectation values (such as covariance) can be expressed in terms of matrix $\mathbf{A}$. Our goal is to calculate the entropy of a subset of oscillators, from which we can derive other quantities such as mutual information and variation of information between sets of oscillators. To that end, we take advantage of the fact that in a multivariate Gaussian ensemble with probability density of the form $P(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)$, the covariance matrix of variables $\mathbf{x}$ is simply $\mathbf{\Sigma}$. Explicitly,

$$\langle x_i x_j\rangle \equiv \frac{\int x_i x_j \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)d\mathbf{x}}{\int \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)d\mathbf{x}} = \Sigma_{ij}. \tag{6}$$

Another useful property of the multivariate Gaussian ensemble is that the marginal covariance matrix of a *subset* of variables $\mathcal{A}$, $\mathbf{\Sigma}_\mathcal{A}$, is obtained from $\mathbf{\Sigma}$ simply by removing all rows and columns corresponding to the marginalized variables (i.e. coordinates not in $\mathcal{A}$). The entropy of any subset of oscillators can then be written in terms of their marginal covariance matrix as

$$H(\mathcal{A}) = \frac{1}{2}k_B \ln\det(2\pi e\mathbf{\Sigma}_\mathcal{A}). \tag{7}$$

Entropy is a thermodynamic quantity, which does not depend on the expression values, only on temperature via matrix $\mathbf{\Sigma}$. From now on, we will measure entropy in multiples of k_B. We also set $A_1 = A_0$ for simplicity. The covariance matrix becomes

$$\mathbf{\Sigma} = \frac{1}{\beta A_0}(\mathbf{I}+\mathbf{L})^{-1}. \tag{8}$$

Matrix $\mathbf{\Sigma}$ exists because $\mathbf{A}$ is positive definite. We know from experience that for as long as the interaction network is connected, $\mathbf{\Sigma}$ is a full matrix, and by Eq. (6) this means that pairwise correlations exists between all pairs of oscillators, not just those who interact directly. This reinforces the old adage 'correlation does not imply causation.' In this way, physical interactions among some oscillators generate correlations among all (of varying magnitude). We note that if the covariance matrix of all genes were known, one could in principle find the raw interaction matrix $\mathbf{A}$ by inversion.

It is convenient to define the matrix $\mathbf{G} \equiv (\mathbf{I}+\mathbf{L})^{-1}$, which depends only on the topological properties of the network. In terms of $\mathbf{G}$,

$$H(\mathcal{A}) = \frac{|\mathcal{A}|}{2}\ln\frac{2\pi e}{\beta A_0} + \frac{1}{2}\ln\det\mathbf{G}_\mathcal{A}, \tag{9}$$

Entropy of a single gene i is

$$H_i = \frac{1}{2}\ln\frac{2\pi e}{\beta A_0} + \frac{1}{2}\ln G_{ii}, \tag{10}$$

where we use the short notation $H_i \equiv H(\{i\})$. Single-gene entropy is not uniform, because interactions have renormalized the fluctuations. Small diagonal elements G_{ii} imply low entropy, which is a measure of high information centrality characteristic of highly connected genes in the network.

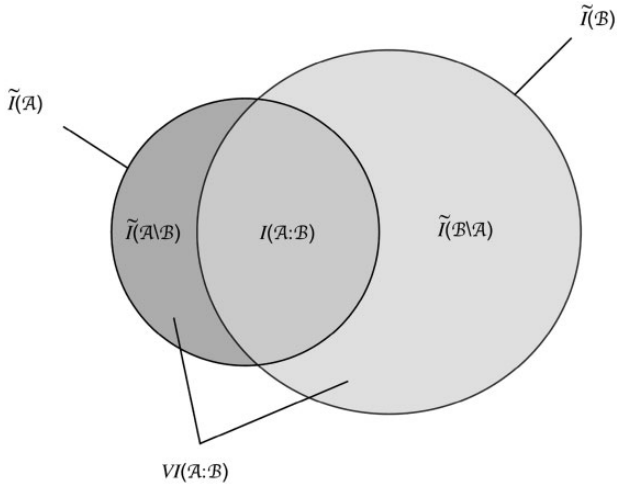

Fig. 1. Definition of mutual information I and variation of information VI in terms of interaction information of sets of harmonic oscillators $\mathcal{A}$ and $\mathcal{B}$

Let us now define a diagonal matrix formed from the diagonal entries of $\mathbf{G}$, $\mathbf{G}_0 \equiv \mathrm{diag}(\mathbf{G})$. The system whose covariance matrix is $\mathbf{G}_0$ is special. In this diagonal system, individual genes have the same entropy as in the fully interacting system, yet their oscillations are statistically independent. This means that physically distinct harmonic oscillators share no information. In this sense, the diagonal system is a most suitable reference. Let us denote the entropy of a gene set in the diagonal model as $H_0(\mathcal{A})$.

We define *interaction information* as

$$\tilde{I}(\mathcal{A}) \equiv H_0(\mathcal{A}) - H(\mathcal{A}), \tag{11}$$

i.e. as the reduction of entropy of set $\mathcal{A}$ due to interactions. This quantity is always non-negative as interactions decrease the entropy, and *does not depend on temperature*. By construction, $\tilde{I}_i = 0$. In terms of $\mathbf{G}$,

$$\tilde{I}(\mathcal{A}) = \frac{1}{2}\sum_{i\in\mathcal{A}}\ln G_{ii} - \frac{1}{2}\ln\det\mathbf{G}_\mathcal{A}. \tag{12}$$

Mathematically, this is equal to interaction information of a system described by the 'normalized' covariance matrix $\tilde{\mathbf{\Sigma}} = \frac{1}{\beta A_0}\tilde{\mathbf{G}}$, where

$$\tilde{\mathbf{G}} = \mathbf{G}_0^{-1/2}\mathbf{G}\mathbf{G}_0^{-1/2}. \tag{13}$$

Formally,

$$\tilde{I}(\mathcal{A}) = -\frac{1}{2}\ln\det\tilde{\mathbf{G}}_\mathcal{A}. \tag{14}$$

This is our central result. $\tilde{\mathbf{G}}$ has a straightforward interpretation as the Pearson correlation matrix,

$$\tilde{G}_{ij} = \frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}} = \varrho_{ij}. \tag{15}$$

We now define mutual information between gene sets $\mathcal{A}$ and $\mathcal{B}$ as $I(\mathcal{A}:\mathcal{B}) = \tilde{I}(\mathcal{A}\cup\mathcal{B}) - \tilde{I}(\mathcal{A}\backslash\mathcal{B}) - \tilde{I}(\mathcal{B}\backslash\mathcal{A})$, and variation of information as $VI(\mathcal{A}:\mathcal{B}) = \tilde{I}(\mathcal{A}\backslash\mathcal{B}) + \tilde{I}(\mathcal{B}\backslash\mathcal{A})$; confer Figure 1. It is understood that $\tilde{I}(\varnothing) = 0$. Naturally, $I(\mathcal{A}:\mathcal{A}) = \tilde{I}(\mathcal{A})$. Mutual information is the amount of information one can learn

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

about set $\mathcal{A}$ from only knowing set $\mathcal{B}$ and vice versa. Variation of information is the amount of extra information contained within $\mathcal{A} \cup \mathcal{B}$ that is not shared by $\mathcal{A}$ and $\mathcal{B}$. This extra information can be considered unwanted in some contexts. When sets $\mathcal{A}$ and $\mathcal{B}$ contain single genes i and $j \neq i$, we obtain

$$I_{ij} = -\frac{1}{2}\ln(1 - \varrho_{ij}{}^2), \tag{16}$$

where we used the shorthand notation $I_{ij} \equiv I(\{i\} : \{j\})$. This is the well-known Gel'fand-Yaglom formula (Gel'fand and Yaglom, 1957) for mutual information of two harmonic oscillators.

It is possible to obtain approximate formulas in what we call the two-body approximation, which is appropriate when no three genes in the set are very correlated, i.e. $\varrho_{ij} \ll 1$ for at least two of the three possible pairs made from of any triple of genes from the set. This approximation holds well for random sets of small size ($|\mathcal{A}| \ll N$), where triple collisions are rare. In this approximation, interaction information becomes

$$\tilde{I}(\mathcal{A}) = \sum_{\substack{i,j \in \mathcal{A} \\ i<j}} I_{ij}. \tag{17}$$

This formula is of course exact when $|\mathcal{A}| = 2$. This approximation can be derived using Laplace's expansion of the determinant in Eq. (14), using the stated assumption and the fact that $\tilde{G}_{ii} = 1$. Mutual information between disjoint sets $\mathcal{A}$ and $\mathcal{B}$ then becomes

$$I(\mathcal{A} : \mathcal{B}) = \sum_{\substack{i \in \mathcal{A} \\ j \in \mathcal{B}}} I_{ij}, \tag{18}$$

which makes sense intuitively. This equation makes it explicit that mutual information has a size bias: it is reasonable to expect that $I(\mathcal{A} : \mathcal{B}) \propto |\mathcal{A}| \cdot |\mathcal{B}|$ as there are $|\mathcal{A}| \cdot |\mathcal{B}|$ terms in the sum. This is not a concern when we are comparing only two sets of fixed sizes, or when all sets to be compared have the same size. When one needs to compare gene sets of varying sizes and be free of the size bias, a useful quantity that largely removes this bias is information quality ratio (Wijaya *et al.*, 2017):

$$IQR(\mathcal{A} : \mathcal{B}) = \frac{I(\mathcal{A} : \mathcal{B})}{\tilde{I}(\mathcal{A} \cup \mathcal{B})}. \tag{19}$$

It is the fraction of the total interaction information within two sets that is mutual. It is always between 0 and 1.

Functions that calculate matrix $\tilde{\mathbf{G}}$ for an arbitrary undirected graph as well as functions that calculate $\tilde{I}(\mathcal{A})$, $I(\mathcal{A} : \mathcal{B})$, $VI(\mathcal{A} : \mathcal{B})$ and $IQR(\mathcal{A} : \mathcal{B})$, and also perform rudimentary statistical testing can be found in the *R* package *gsia* at https://github.com/ucsd-ccbb/gsia.

3 Results and discussion

We now return to interaction information and mutual information problems defined in the Introduction.

The first question is answered by comparing interaction information of a given set, $\tilde{I}(\mathcal{A})$, with that of a random set $\mathcal{R}$ of the same size as $\mathcal{A}$ but drawn from the null distribution. The null distribution is highly context specific and should reflect the process that generated set $\mathcal{A}$. If for example the null experiment can produce any expressed gene with equal probability, then the null distribution contains all gene sets of size $|\mathcal{A}|$ drawn uniformly randomly from the expressed gene set. We define *P*-value of $\tilde{I}(\mathcal{A})$ as the probability $P[\tilde{I}(\mathcal{R}) \geq \tilde{I}(\mathcal{A})]$ in the context of the corresponding null model.

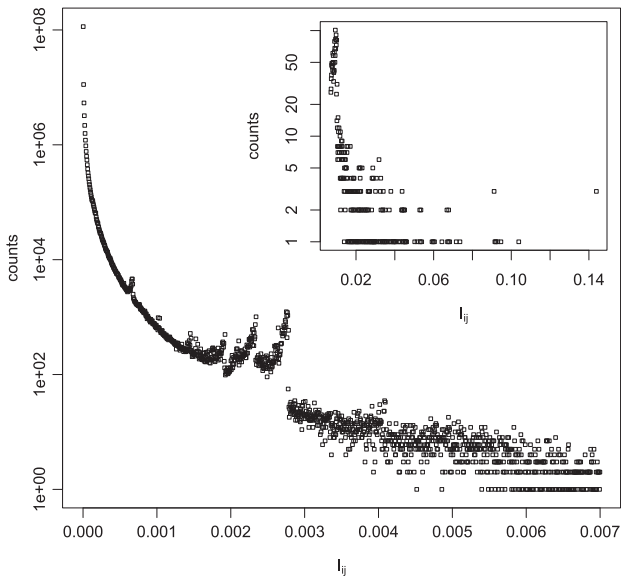

Fig. 2. Histogram of I_{ij} for InBioMap interaction network. Inset: continuation for $I_{ij} > 0.007$. Large values are extremely rare ($I_{ij} \geq 0.02$ for only 0.00017% of gene pairs), which no doubt stems from the sparsity of the raw interaction matrix **A**

This probability can be estimated numerically, either exactly using Eq. f(14) or approximately using Eq. (17).

The second question is answered by comparing mutual information of the two gene sets, $I(\mathcal{A} : \mathcal{B})$, with that of sets $\mathcal{A}$ and a random set $\mathcal{R}$ of the same size as $\mathcal{B}$ but drawn from the null distribution. We define the *P*-value as the probability $P[I(\mathcal{A} : \mathcal{R}) \geq I(\mathcal{A} : \mathcal{B})]$. Note that this probability may be significant even when $\mathcal{A}$ and $\mathcal{B}$ are disjoint. Again, the null model is very context dependent and should reflect the process that generated set $\mathcal{B}$. For instance, sets $\mathcal{R}$ and $\mathcal{A}$ may be allowed to overlap if $\mathcal{B}$ and $\mathcal{A}$ can overlap in principle. This definition of significance is explicitly asymmetric as $\mathcal{A}$ has a position of the reference set. It is therefore possible for a set $\mathcal{B}$ to be significantly related to $\mathcal{A}$, while $\mathcal{A}$ is not significantly related to $\mathcal{B}$. A symmetric definition follows readily.

To illustrate these concepts, we consider three different bioinformatic applications: genome-wide association studies (GWAS), identification of related metabolic pathways, and finding related human diseases based on disease gene signatures. As for the interaction network, we use the InBioMap protein-protein interaction network (Li *et al.*, 2017), which is an integrative database with more than 17000 proteins and $\sim 6 \times 10^5$ interactions, both direct (physical) and indirect (functional). This database does not contain 'interactions' inferred from co-expression nor co-citation evidence and is therefore ideally suited for our approach. Gene expression correlations emerge naturally in our model between any two genes that are connected via the network even when they do not interact directly. In fact, all gene pairs within a connected component of the graph are correlated to various degree, as shown in Figure 2. InBioMap performed very well in a recent series of gene set recovery tests done by Huang *et al.* (2018).

The InBioMap network is not tissue specific but in the context of GWAS or diseases this is appropriate. After conversion of identifiers from UNIPROT to Entrez gene and retaining the largest connected component of the network, we obtained a network of $N = 17171$ genes with 588 897 edges. This is a sparse network with just 0.40% of possible interactions. The distribution of the pairwise mutual

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

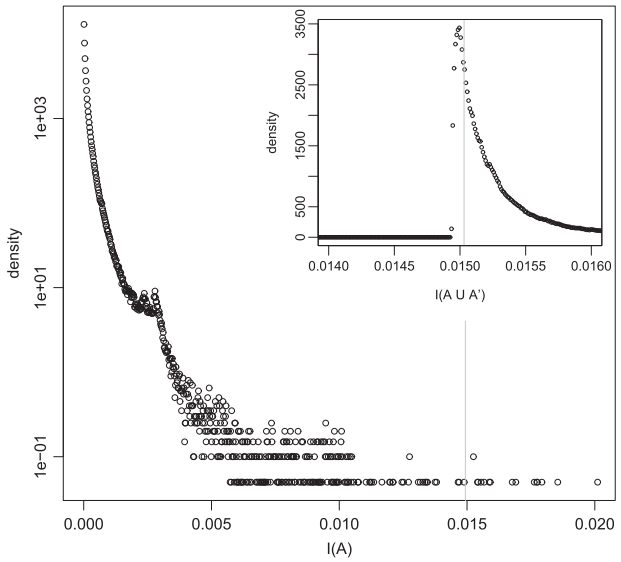

Fig. 3. Null distribution of the interaction information $\tilde{I}(\mathcal{A})$ for a set of six uniformly randomly selected genes from the network. The vertical line marks the value of $\tilde{I}(\mathcal{A})$ for the set of six candidate genes associated with alcohol dependency. $p = 3.4 \times 10^{-5}$. Inset: Null distribution of interaction information for a set consisting of the six candidate genes and three uniformly randomly selected genes from the network. The vertical line marks the value of $\tilde{I}(\mathcal{A} \cup \mathcal{A}')$ when $\mathcal{A}'$ is the set of three 'best' non-significant genes suggested by Zuo *et al.* (2015). $P = 0.72$

information matrix I_{ij} is drawn in Figure 2. Large values are relatively rare. The sharp local maxima in the I_{ij} distribution are caused by frequent repetition of local connectivity patterns or graph motifs throughout the network.

3.1 GWAS studies of alcohol dependency

Alcohol dependency is a complex psychiatric diagnosis for a person who is either physically or psychologically dependent on alcohol. It has now been re-classified as alcohol use disorder (AUD) [American Psychiatric Association (2013)]. To our knowledge, there have been six successful GWAS studies of AUD since 2009; by success we mean detection of at least one single nucleotide polymorphism (SNP) at the level $P < 5 \times 10^{-8}$ (Frank *et al.*, 2012; Gelernter *et al.*, 2014; Park *et al.*, 2013; Quillen *et al.*, 2014; Treutlein *et al.*, 2009; Zuo *et al.*, 2015). After association of SNP's to genes, these studies collectively produced seven candidate genes: PECR, ADH1C, ADH7, ALDH2, LOC100507053, ADH1B and SERINC2. Only two studies reproduced the same genetic locus albeit with different SNP's. The absence of agreement between studies can be explained by small power and/or by true genetic differences among the cohorts. Regardless, if there is a common molecular pathway that is contributing to the risk of AUD, then we would expect these genes to be highly mutually informative. Quantitatively speaking, their interaction information should be unusually large. We test this hypothesis by calculating the probability that interaction information of a null gene set will be no smaller than the observed value. Since GWAS studies can in principle identify any genomic locus, we define the null set as a set of six genes drawn uniformly randomly from the network (one gene, LOC100507053, is not found in InBioMap so it will be ignored; the remaining six are part of the candidate gene set $\mathcal{A}$). We sampled 10^6 null sets, with the result $p[\tilde{I}(\mathcal{R}) \geq \tilde{I}(\mathcal{A})] = 4.1 \times 10^{-5}$, which is highly significant. We conclude therefore that the candidate genes are significantly mutually informative and the likelihood that they were selected uniformly randomly is very small. The null distribution and the observed value for the set of six candidate genes is in Figure 3. The distribution of $\tilde{I}(\mathcal{A})$ resembles that of I_{ij} in Figure 2. This is not a coincidence: set $\mathcal{A}$ is small, and Eq. 17 is a good approximation for a typical null set. Chances are that a single one of the $\binom{6}{2} = 15$ terms I_{ij} makes a dominant contribution, which explains why the tail of the distribution in Figure 3 resembles that of Figure 2.

Even though the GWAS studies produced disjoint SNP's, the researchers were confident that the candidate genes were mostly valid because several of them are involved in alcohol metabolism. This is precisely the point of our work: just as these researchers were able to use prior knowledge, which increased their confidence in their results, our method puts a quantitative *P*-value on that confidence without the need to know anything about alcohol metabolism specifically, as our method leverages the knowledge of the entire interaction network. This way, we can in principle uncover relationships between genes even when the underlying pathway is unknown or unnamed, or lies at the boundary of two canonical pathways, or is a fraction of a canonical pathway, provided the interaction network is reasonably complete.

An interesting wrinkle to this story are the genes that did not achieve the significance threshold of $P < 5 \times 10^{-8}$ (Zuo *et al.*, 2015), yet the authors felt they were 'significantly or suggestively associated' with AUD, and 'most appropriate for follow-up' as contributors to risk for AUD. In the European ancestry combined cohort, these genes were STK40, KIAA0040 and IPO11. Can we support their assertion? Let us denote the set of these three added genes by $\mathcal{A}'$. If these genes were part of the same network neighborhood as the six candidate genes, we would expect $\tilde{I}(\mathcal{A} \cup \mathcal{A}')$ to be unusually large, given $\mathcal{A}$. The conditional probability that three random genes $\mathcal{R}'$ when added to the six candidate genes $\mathcal{A}$ will have interaction information no less than the set $\mathcal{A} \cup \mathcal{A}'$ is $p[\tilde{I}(\mathcal{A} \cup \mathcal{R}') \geq \tilde{I}(\mathcal{A} \cup \mathcal{A}'); |\mathcal{R}'| = |\mathcal{A}'|] = 0.72$. This means that the three added genes are not significantly informative of the six already established candidate genes and we cannot support the authors' assertion that they are suggestively associated with AUD. The null distribution and the observed value of interaction information for the set of six candidate genes with three added genes is in Figure 3 (inset).

3.2 Metabolic pathways

We now turn to canonical pathways, specifically A) Glyoxylate and dicarboxylate metabolism, and B) Nitrogen metabolism [KEGG pathways 00630 and 00910 (Kanehisa *et al.*, 2004)]. These are metabolic pathways in plants, which are coupled via a metabolite formate, but otherwise share no gene products (enzymes). It is not rare for canonical pathways to be disjoint; in fact, 79% of all KEGG pathway pairs are. We think this is more a reflection of a human tendency to compartmentalize knowledge rather than the nature's way of functioning. Let us denote the set of genes involved in pathway A by $\mathcal{A}$, and genes of pathway B by $\mathcal{B}$. We have $|\mathcal{A}| = 16$, $|\mathcal{B}| = 23$, and $\mathcal{A} \cap \mathcal{B} = \varnothing$. Activity of these pathways depends on the level of their common metabolite formate, which suggests that they are coordinated as parts of a greater metabolic network. In this context therefore, gene sets $\mathcal{A}$ and $\mathcal{B}$ are functionally related and should be mutually informative. Yet, neither pathway is significantly enriched by the other in terms of genes, because their gene sets are disjoint. Gene set enrichment methods fail to find a relationship between these two pathways. Can we detect a relationship using the present

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

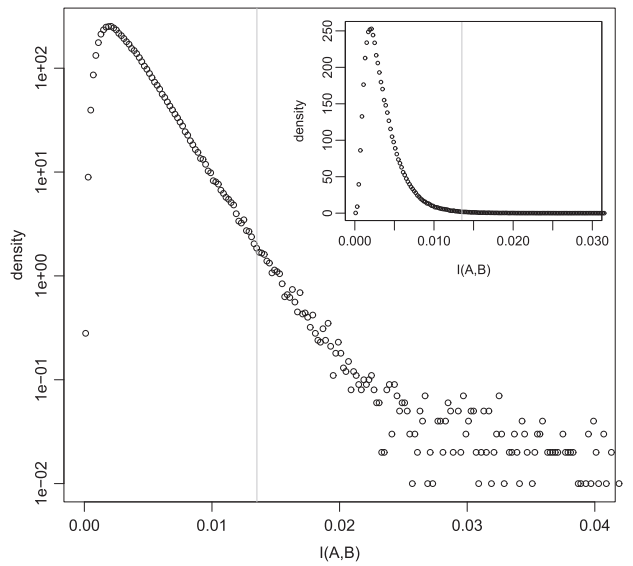

Fig. 4. Null distribution of mutual information $I(\mathcal{A},\mathcal{B})$ for $|\mathcal{A}| = 16$, $|\mathcal{B}| = 23$ and $\mathcal{A} \cup \mathcal{B} = \varnothing$, calculated numerically with 5×10^5 random sets. The observed value for two canonical pathways of this size, Glyoxylate and dicarboxylate metabolism and Nitrogen metabolism, is marked by the vertical line ($P = 0.0055$). Insert: same as main panel, in linear scale

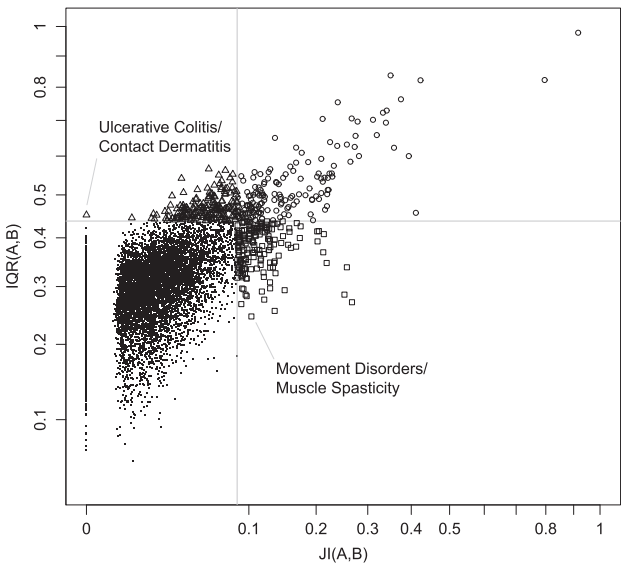

Fig. 5. Information quality ratio versus Jaccard index for all pairs of diseases. Pairs found significant by both statistics (◯), only by the information statistic *IQR* (△) and only by the enrichment statistic *JI* (□). The horizontal line separates the data points at the level *lfdr* = 0.3; the vertical line into two parts of the same size as the horizontal line. The annotated disease pairs were chosen as extreme representatives of what is considered significant by either method alone

method? Indeed, the probability that two uniformly randomly selected gene sets $\mathcal{R}$ and $\mathcal{R}'$ such that $|\mathcal{R}| = |\mathcal{A}|$, $|\mathcal{R}'| = |\mathcal{B}|$ and $\mathcal{R} \cap \mathcal{R}' = \varnothing$ will have mutual information no less than the observed $I(\mathcal{A},\mathcal{B})$ is 0.0055, which is highly significant. The null distribution of mutual information is in Figure 4. We conclude that Glyoxylate and dicarboxylate metabolism and Nitrogen metabolism are significantly mutually informative pathways.

3.3 Human disease set signatures

Human diseases are a set of pathologies that manifest themselves in distinct ways. In the genomics era, diseases came to be associated with genes whose altered expression or altered sequence can either cause or increase the likelihood or severity of the symptoms. These gene set signatures are catalogued in the DisGeNet database (Piñero *et al.*, 2017). There are 130 821 diseases in DisGeNet, but we considered only diseases with distinct gene signatures and sizes between 50 and 200 genes. From the resulting 357 diseases, we manually removed predominantly congenital disorders and abnormalities like retrognathia and flat face. The final set contains 130 predominantly metabolic, degenerative, inflammatory and neoplastic diseases.

To illustrate mutual information, we will organize the diseases based on mutual information of gene signatures and build a network of significant disease–disease relationships. The central quantity is mutual information; however, because of the size bias mentioned above, information quality ratio (19) is more appropriate. We calculate *IQR* for every pair of diseases. In order to find the appropriate significance cutoff, we need to define the null model first. The appropriate null 'disease' gene set should have all the characteristics of actual disease gene sets such as size distribution, perhaps also the interaction information distribution, yet it should be statistically independent from any other null set. A naive uniformly random null model used in the above two cases fails, because real disease gene sets show highly non-uniform gene frequency distribution. This also means that the observed disease–disease gene overlap is typically much larger than expected in the uniform random model, and so is mutual information. After much consideration we decided it would be best to learn what it means for two diseases to be significantly related from the diseases themselves. In essence, we assume that most disease gene sets should be called 'not significantly related' on the basis of them being distinct clinical diagnoses. This means that the observed distribution of $IQR(\mathcal{A},\mathcal{B})$ can be viewed as a mixture of the actual null distribution and a small fraction of significant values. The significant values can be found with the empirical Bayesian approach of Efron (2008) using the *R* function *locfdr* (https://cran.r-project.org/package=locfdr). We call disease pairs significantly related if their *lfdr* < 0.3 (posterior probability better than 70%). There are 331 significant relationships that define a network of disease–disease relationships.

To contrast the present method with the established gene enrichment methodology, we also construct the network using a gene enrichment statistic. The quantity analogous to $IQR(\mathcal{A},\mathcal{B})$ is the Jaccard index

$$JI(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}. \tag{20}$$

Like *IQR*, *JI* is also normalized to interval [0, 1]. Unfortunately, the empirical distribution of *JI* among the 130 diseases is multimodal with an isolated peak at $JI(\mathcal{A},\mathcal{B}) = 0$ (14.7% of disease–disease gene set pairs have empty intersection, cf. Fig. 5) and does not lend itself to straightforward empirical Bayes analysis, so it is not obvious where to draw the significance threshold for *JI* based on *JI* values alone. For purposes of comparison with the present method, we draw the threshold at a value that yields the same number of significant disease pairs as the information method. This will suffice for a fair comparison of the two methods.

The entire disease–disease network of 130 diseases can be viewed interactively at the Network Data Exchange (NDEx) (Pratt *et al.*, 2015), UUID: 4a7f0c68-69b4-11e8-a4bf-0ac135e8bacf].

Of the 506 disease pairs found significant by at least one method (6.0% of all disease pairs), 156 were identified by both methods,

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

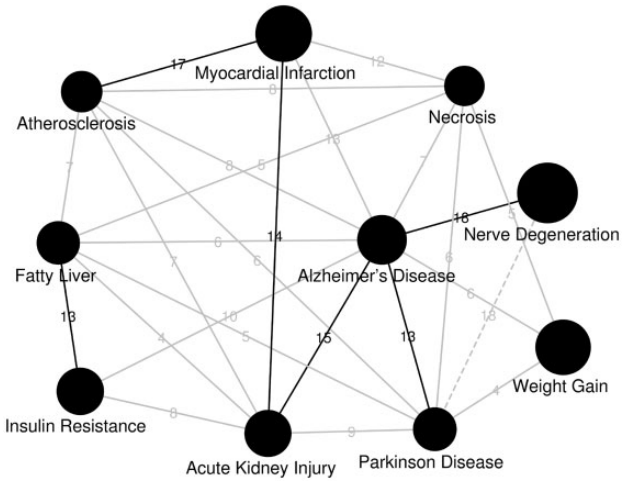

Fig. 6. Network of diseases in the nearest neighborhood of Alzheimer's disease. Node area is proportional to the size of the corresponding disease gene set and ranges from 52 (Necrosis) to 119 (Nerve Degeneration). The numbers along the edges are numbers of genes shared by the two connected diseases. Black edges were found significant by both the information (*IQR*) and the enrichment (*JI*) statistics; solid grey edges were detected only by the information statistic, and the dashed grey edges were detected only by the enrichment statistic

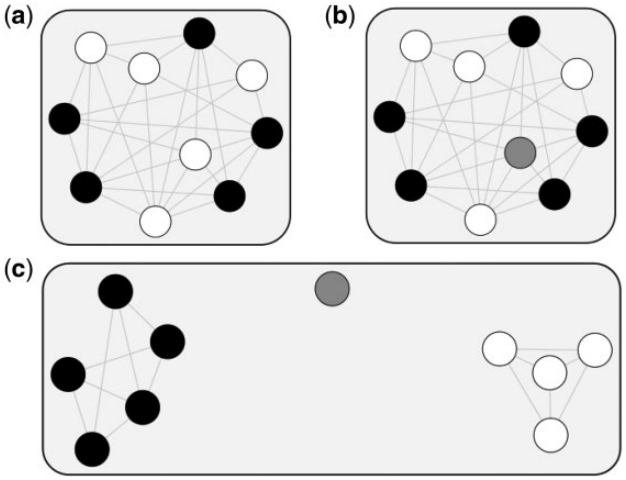

Fig. 7. Illustration of three cases relevant to the comparison of the information method and gene set enrichment methods: (**a**) Two gene sets (black and white) have no genes in common but they are proximal in the network sense (are highly mutually informative), (**b**) Two gene sets are highly mutually informative and at the same time have a significant number of genes in common (grey) and (**c**) two gene sets are far removed from each other in the network sense (are mutually uninformative) but have a significant number of genes in common (grey nodes)

175 pairs were detected only by the information method, and 175 only by the gene enrichment method. The overlap is highly significant, as 47.1% of disease pairs detected by one method are also detected by the other. However, the complement of 52.9% disease pairs detected uniquely by either method deserves scrutiny. We would like to know if one method is clearly 'better' at detecting relationship or are the two methods complementary. To this end we explore a small part of the disease–disease network centered around Alzheimer's Disease (AD) in more detail (Fig. 6). This subnetwork is also accessible at NDEx, UUID 839b1acb-69b4-11e8-a4bf-0ac135e8bacf.

Dark edges of this graph represent relationships detected by both methods. From the perspective of AD, *Nerve Degeneration* and *Parkinson Disease* are certainly relevant relationships. A less obvious one is *Acute Kidney Injury*. However, there is epidemiological evidence that people who had been hospitalized with acute kidney injury (and recovered) have a significantly increased risk of developing dementia later in life (Tsai *et al.*, 2017). Hence we conclude that these three relationships detected by both methods are true. How about relationships detected by the information statistic alone? Let us take them in the counter-clockwise order according to Figure 6: *Weight Gain* is a major risk factor for developing AD and is probably causative (Kivipelto *et al.*, 2005). *Necrosis* is part of the spectrum of neuronal cell death seen in neurodegenerative diseases (Gorman, 2008). *Myocardial Infarction* seems far removed from AD, but they share the same pathogenic mechanisms, namely cholesterol metabolism and inflammation (Licastro *et al.*, 2011). By the same token, *Atherosclerosis* is highly relevant to AD and the two have even been postulated to be one disease with different presentations (Lathe *et al.*, 2014; Roher *et al.*, 2004). *Fatty Liver* (non-alcoholic) induces symptoms of AD (Kim *et al.*, 2016). Finally, *Insulin Resistance* is emerging as an important feature of AD and the two are possibly synergistic rather than coincidental diseases (Arnold *et al.*, 2018). Now what about relationships detected only by the enrichment statistic? In the nearest network neighborhood of AD, there is only one, between *Nerve Degeneration* and *Parkinson Disease*. This is certainly correct. It appears that the present method and the gene enrichment method are complementary rather than exclusive.

Let us now look at the most extreme disease pairs called significant by either statistic alone (see Fig. 5). *Ulcerative Colitis* and *Contact Dermatitis* seem unrelated at first sight. We note however that ulcerative colitis can affect any organ system in the body, and skin is the most commonly affected one, with dermatitis ulcerosa (pyoderma gangrenosum) being the most common type of dermatitis (Huang *et al.*, 2012). While this is not the same as contact dermatitis, we do not rule this connection to be false positive. *Ulcerative Colitis* and *Contact Dermatitis* is a special disease pair in the sense that it was the only one detected by the information method in the complete absence of gene intersection ($\mathcal{A} \cap \mathcal{B} = \varnothing$, $JI(\mathcal{A}, \mathcal{B}) = 0$). On the other side of the spectrum, the relationship between *Movement Disorders* and *Muscle Spasticity* is certainly real as spasticity is a type of movement disorder. Again, it appears that the information statistic and enrichment statistic are complementary.

We now describe, in qualitative terms, the kinds of gene set pairs whose relationship can be detected by only one or by both methods. Figure 7 has cartoon representations of them. Panel a) contains two disjoint gene sets that are highly interconnected in the network. In the harmonic oscillator representation, the oscillators are strongly coupled, which means that they will be highly correlated. From the information point of view, these gene sets are highly mutually informative, therefore they can be detected by the information method. Enrichment methods will not detect a relationship. Panel b) has two highly interconnected gene sets with a significant number of genes in common (probability that two gene sets of size 5 and 6 chosen independently and uniformly randomly from a set of 10^4 genes will have at least one gene in common is 0.0030). These gene sets can be recognized by both information and enrichment methods. Finally, panel c) has two sets who are mutually uninformative except for a small yet significant number of genes in common. Gene sets like these can be detected only by enrichment methods.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018

4 Conclusions

We derived mathematically simple information measures for gene sets in the context of an underlying gene interaction network. These information measures follow from very few biological assumptions, chief among them being the assumed existence of a stable steady state, and from established methods of statistical physics. We applied these information concepts to the problems of interaction information and mutual information for disease gene signatures. Our results indicate that these information measures are able to recapitulate known disease relationships; in some cases even when established gene enrichment methods fail. We find that the information methods presented here and the established gene enrichment methods are complementary rather than exclusive. The utility of each depends on the kind of gene set similarity a researcher is looking for. We expect that the present methodology will be used in other contexts, such as drug repurposing based on drug and disease signature gene sets, and to relate experimentally relevant gene sets to curated pathways.

Funding

This work has been supported by the UC San Diego Clinical and Translational Research Institute Grant UL1TR001442. TI acknowledges funding from the National Resource for Network Biology Grant P41GM103504.

Conflict of Interest: none declared.

References

American Psychiatric Association. (2013) *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Publishing, Arlington, VA.

Arnold,S.E. *et al*. (2018) Brain insulin resistance in type 2 diabetes and Alzheimer disease: concepts and conundrums. *Nat. Rev. Neurol*., **14**, 168–181.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol*., **3**, 140.

Cokus,S.J. *et al*. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Cowen,L. *et al*. (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet*., **18**, 551–562.

Efron,B. (2008) Microarrays, empirical Bayes and the two-groups model. *Stat. Sci*., **23**, 1–22.

Estrada,E. and Hatano,N. (2010) Topological atomic displacements, Kirchhoff and Wiener Indices of Molecules. *Chem. Phys. Lett*., **486**, 166–170.

Frank,J. *et al*. (2012) Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addict. Biol*., **17**, 171–180.

Gelernter,J. *et al*. (2014) Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Mol. Psychiatry*, **19**, 41–49.

Gel'fand,I.M. and Yaglom,A.M. (1957) Calculation of amount of information about a random function contained in another such function. *Am. Math. Soc. Transl. Ser. 2*. **12**, 199–246 (English Translation of Original in Uspekhi Matematicheskikh Nauk, 12, 3–52).

Gorman,A.M. (2008) Neuronal cell death in neurodegenerative diseases: recurring themes around protein handling. *J. Cell Mol. Med*., **12**, 2263–2280.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.

Huang,B.L. *et al*. (2012) Skin manifestations of inflammatory bowel disease. *Front. Physiol*., **3**, 13.

Huang,J.K. *et al*. (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst*., **6**, 484–495.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.

Kim,D.-G. *et al*. (2016) Non-alcoholic fatty liver disease induces signs of Alzheimer's disease (AD) in wild-type mice and accelerates pathological signs of AD in an AD model. *J. Neuroinflam*., **13**,

Kivipelto,M. *et al*. (2005) Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. *Arch. Neurol*., **62**, 1556–1560.

Klein,D.J. and Randić,M. (1993) Resistance Distance. *J. Math. Chem*., **12**, 81.

Klein,R.J. *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Köhler,S. *et al*. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet*., **82**, 949–958.

Lathe,R. *et al*. (2014) Atherosclerosis and Alzheimer - diseases with a common cause? Inflammation, oxysterols, vasculature. *BMC Geriatr*., **14**, 36.

Li,T. *et al*. (2017) A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods*, **14**, 61–64.

Licastro,F. *et al*. (2011) Sharing pathogenetic mechanisms between acute myocardial infarction and Alzheimer's disease as shown by partially overlapping of gene variant profiles. *J. Alzheimers Dis*., **23**, 421–431.,

Morin,R. *et al*. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, **45**, 81–94.

Park,B.L. *et al*. (2013) Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication. *Hum. Genet*., **132**, 657–668.

Piñero,J. *et al*. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Pratt,D. *et al*. (2015) NDEx, the Network Data Exchange. *Cell Syst*., **1**, 302–305.

Quillen,E.E. *et al*. (2014) *ALDH2* is associated to alcohol dependence and is the major genetic determinant of "daily maximum drinks" in a GWAS study of an isolated rural Chinese sample. *Am. J. Med. Genet. B Neuropsychiatr. Genet*., **165**, 103–110.

Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Roher,A.E. *et al*. (2004) Atherosclerosis of cerebral arteries in Alzheimer disease. *Stroke*, **35**, 2623–2627.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Treutlein,J. *et al*. (2009) Genome-wide association study of alcohol dependence. *Arch. Gen. Psychiatry*, **66**, 773–784.

Tsai,H.-H. *et al*. (2017) Increased risk of dementia in patients hospitalized with acute kidney injury: a nationwide population-based cohort study. *PLoS One*, **12**, e0171671.

Wijaya,D.R. *et al*. (2017) Information Quality Ratio as a novel metric for mother wavelet selection. *Chemometr. Intell. Lab. Syst*., **160**, 59–71.

Zuo,L. *et al*. (2015) A new genome-wide association meta-analysis of alcohol dependence. *Alcohol Clin. Exp. Res*., **39**, 1388–1395.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty691/5068595 by University of California, San Diego user on 07 December 2018