bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evolutionary Distance of Gene-Gene Interactions:

Estimation under Statistical Uncertainty

Xun Gu[1*]

[1] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA.

*Corresponding author. E-mail: xgu@iastate.edu

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Consider the functional interaction of gene *A* to an interaction subject *X*; for instance, it is the gene-gene interaction if X represents for a gene, or gene-tissue interaction (expression status) if X for a tissue. In the simplest case, the status of this *A*-*X* interaction is r=1 if they are interacted, or r=0 otherwise. A fundamental problem in molecular evolution is, given two homologous (orthologous or paralogous) genes *A* and *B*, to what extent their functional overlapping could be by the means of interaction networks. Given a set of interaction subjects ($X_1, \ldots X_N$), it is straightforward to calculate the interaction distance (I_{AB}) between genes *A* and *B*, by a Markov-chain model. However, since the high throughput interaction data always involve a high level of noises, reliable inference of r=1 or r=0 for each gene remains a big challenge. Consequently, the estimated interaction distance (I_{AB}) is highly sensitive to the cutoff of interaction inference which is subject to some arbitrary. In this paper we will address this issue by developing a statistical method for estimating I_{AB} based on the *p*-values (significant levels). Computer simulations are carried out to evaluate the performance of different *p*-value transformations against the uncertainty of interaction networks.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Thanks to the growing amount and quality of omics data, the assembly of biological networks has played an important role to unveil the underlying cellular processes and evolution. In this scenario, Protein-Protein Interactions (PPIs) (Mosca et al. 2013; Hao et al .2016) and Gene Regulation Networks (GRN) (Halton 2017) are among the most important and widely studied networks. A PPI networks is described in terms of proteins (nodes) and their physical/functional interactions (edges) (Nibbe et al. 2010; Sharan et al. 2007; Procaccini et al. 2016; Gustafsson et al. 2014), while a GRN network described in terms of gene-protein interactions. Note that transcriptome (RNA-seq) can be also viewed as the gene-tissue interactions. Both PPI and GRN networks are analyzed through the identification of subnetworks, or modules, showing specific topological and/or functional characteristics (Barabási et al. 2011; Vella et al. 2017; Hartwell et al 1999; Gursoy et al. 2008; Fraser 2005). For instance, a PPI module represents a group of proteins taking part in specific, separable functions such as protein complexes, metabolic pathways or signal transduction systems.

Here we focus on the evolution of interaction network. Initiated by an influential, but controversial study (Fraser et al. 2002) on the effect of protein-protein interactions on protein sequence evolution, interactivity at the DNA, protein and genetic levels has been the major topic in the study of systems biology and evolution (von Mering et al. 2002; Wagner 2001; 2003; Jordan et al. 2003; Kim WK, Marcotte 2008; Sun and Kim 2011). Given the network sizes, typically involving thousands of elements, it often requires in-silico automated methods (Ma'ayan 2008; Grindrod, P. & Kibble 2004). However, it has been shown that the interaction inferences based on large-scale omics data are subject to a high level of statistical noises (Benjamini and Hochberg 1995; Shaffer 1995; Benjamini and Yekutieli 2001; Efron et al. 2001; Efron 2004; Smyth 2004). Consider the functional interaction of gene *A* to any other gene *X*. In the simplest case, the status of this *A*-*X* interaction is r=1 if these two genes are connected, or r=0 otherwise. For two duplicate genes *A* and *B*, an important measure for their functional overlapping is the number (n_{11}) of other (*X*) genes that interact with both duplicate genes *A* and *B* (*A*-*X* and *B*-*X*). One may utilize a simple Poisson model to estimate the interaction distance (D_I) between duplicates, i.e., the average number of interaction changes (losses and gains). For a set (n)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of genes, each of which interacts with *A*, or *B* or both, one can calculate the proportion of interaction divergence $q=1-n_{11}/n$, and estimate the interaction distance by $D_I=-n \ln(1-q)$. However, since the high throughput functional genomic data involve high level noises, statistical evaluation of inferred interactions becomes a big challenge in the genomic analysis. We shall address this issue, i.e. how to take the statistical uncertainty in the evolutionary analysis.

New Methods

Statistical representation for gene-gene interactions

p-Value presentation for single gene-query interaction. Many genomic studies have adopted a *p*-value to characterize the statistical significance of any gene-query interaction. That is, instead of a binary (r=1 or 0) status of any gene-query (*A*-*X*) interaction, a *p*-value is assigned for the interaction between a gene and a query; a small *p*-value, e.g., p=0.001, means that the interaction is highly statistically significant, and *vice versa*.

Multiple-test problem. Usually *p*-values of high throughput gene-gene interactions are subject to the sophisticated multiple-test problem (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). Intuitively speaking, it depends on a pre-specified cut-off for the *p*-value, that is, the interaction status is assigned as '1' if the *p*-value is less than the cutoff; otherwise the interaction is null. How to determine the cut-off is controlled by the false positive rate of discovery, usually called the *q*-value for a given set of interactions. A *q*-value, say 0.1, means that for all gene-gene interactions predicted as 'active', the proportion of false-positive cases has been set to be 10%. To calculate the *q*-value under a given p-value cut-off, we have to model the *p*-value distributions under the null and the alternative hypotheses, respectively (Efron 2004; Erron et al. 2001).

Challenge for evolutionary analysis of interactions. While it is widely accepted that the evolutionary distance of interactions between homologous genes is important for understanding the pattern of genome evolution, the estimation problem has not been well addressed. Suppose that we have two genes (*A* and *B*) that have *p*-values to *N* interaction

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

queries (X_1,.., X_N), respectively. Apparently, the distance estimation becomes straightforward if all states of *A*-X_i and *B*-X_i (i=1,…, N) are unambiguously observed. As discussed above, statistical inference of interaction status for a set of *p*-values is controlled by the false-positive rate (the *q*-value). As a result, the estimated interaction distance is highly sensitive to the *q*-value cutoff selected.

A general *p*-value framework for the interaction evolution

Our goal is to develop a practically feasible approach to overcome this problem by treating these *p*-values as observations. We use the statistical approach for modeling *p*-values similar to that used in the analysis of multi-test problem. Instead of determining the false positive rate for a given set of predictions, we combine the *p*-value model and the underlying evolutionary model to develop a cutoff-free method for estimating the interaction distance.

Expression distance defined by transformed p-values

For *p*-value presentation of gene-query interactions, we have to develop an explicit model for the interaction evolution (Fig.1). Consider two homologous genes *A* and *B* with an interaction query *X*, there are four combined patterns (r_A, r_B): (1, 1), (1, 0), (0, 1) and (0, 0), respectively. For instance, (1, 1) means that both genes have the interaction with the same query *X*; (1, 0) means gene *A* has the interaction but gene *B* does not, and so forth. The probability of each pattern is denoted by $P(r_A, r_B)$, where r_A, r_B=1 or 0, respectively.

Let p_A and p_B be the *p*-values for interactions *A*-*X* and *B*-*X*, respectively. Let y_A and y_B be any given transformations of p_A and p_B, respectively. Two simple forms that are practically useful are y=-ln (p+p_0) and y=p, respectively (p_0 is the small number to avoid y tends to be infinite when p approaches 0, usually one may choose p_0=1/M; where M is the total number of interactions under study). Next, we define the expectation of squared *y*-score differences between genes *A* and *B* as follows

$$\delta_{AB}^2 = E[(y_A - y_B)^2] \quad (1)$$

where *E* is for taking expectation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While it is straightforward to estimate δ^2_{AB} from the interaction data as long as the transformation formula is specified, the challenge is how to connect δ^2_{AB} to the evolutionary model of interactions. Our idea is to expand the right hand of Eq.(1) by the means of conditional expectations with respect to interaction patterns (r_A, r_B), which can be concisely written by

$$\gamma_{ij}=E[(y_A-y_B)^2|r_A=i, r_B=j] \qquad (2)$$

where i, j=0 or 1. Let $P(r_A, r_B)$ be the probability of an interaction pattern (r_A, r_B). According to the probability theory, we have

$$\begin{aligned} \delta^2_{AB} &= \sum_{r_A,r_B=0,1} E[(y_A-y_B)^2|r_A,r_B]P(r_A,r_B) \\ &= \gamma_{11}P(1,1)+\gamma_{10}P(1,0)+\gamma_{01}P(0,1)+\gamma_{00}P(0,0) \end{aligned} \qquad (3)$$

In short, according to Eq.(3), δ^2_{AB} can be decomposed into two components: the first component is the interaction uncertainty measured by γ_{ij}, and the second component is the evolution of interactions (gain or loss) measured by the probabilities $P(r_A, r_B)$.

Evolutionary model to determine $P(r_A, r_B)$

Suppose that gain and loss of interactions are the major mechanisms for the interaction evolution of a gene, which are independent of other genes. We then develop a simple Markov-chain model as follows. Let π_1 be the probability of active interaction (r=1), and π_0=1-π_1 be that of inactive interaction (r=0). Let λ be the evolutionary rate of interaction. Under the Markov-chain model, the gain rate and the loss rate of an interaction are given by $\pi_1\lambda$ and $\pi_0\lambda$, respectively. It follows that the transition probabilities from state i to state j after the t time units, $P_{ij}(t)$, i, j=0, or 1, are given by $P_{11}(t)= \pi_1+\pi_0 e^{-\lambda t}$, and $P_{10}(t)= \pi_0(1-e^{-\lambda t})$; similarly, we have $P_{00}(t)=\pi_0+\pi_1 e^{-\lambda t}$, and $P_{01}(t)= \pi_1(1-e^{-\lambda t})$. In the case of two homologous genes, the probability for any observed pattern $P(r_A, r_B)$ can be calculated as follows

$$\begin{aligned} P(1,1) &= \pi_1^2+\pi_1\pi_0 e^{-2\lambda t} \\ P(1,0) &= \pi_1\pi_0\left(1-e^{-2\lambda t}\right) \\ P(0,1) &= \pi_1\pi_0\left(1-e^{-2\lambda t}\right) \\ P(0,0) &= \pi_0^2+\pi_1\pi_0 e^{-2\lambda t} \end{aligned} \qquad (4)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A general formula for the evolutionary distance of interactions

When the probability of an interaction pattern (r_A, r_B), $P(r_A, r_B)$, is determined by Eq.(4), we can derive a general formula for the evolutionary distance of interactions by combining Eq.(3) by Eq.(4), resulting in

$$\delta_{AB}^2 = \delta_\infty^2 - (\delta_\infty^2 - \delta_0^2)e^{-2\alpha t} \tag{5}$$

where δ^2_∞ and δ^2_0 are given by

$$\begin{aligned} \delta_\infty^2 &= \gamma_{11}\pi_1^2 + (\gamma_{10} + \gamma_{01})\pi_1\pi_0 + \gamma_{00}\pi_0^2 \\ \delta_0^2 &= \gamma_{11}\pi_1 + \gamma_{00}\pi_0 \end{aligned} \tag{6}$$

Eq.(5) shows that, when t=0, $\delta^2_{AB}=\delta^2_0$, and δ^2_{AB} increases with t and ultimately reaches δ^2_∞ as $t\to\infty$. One may further define the effective proportion of different interactions between genes A and B

$$q_e = \frac{\delta_{AB}^2 - \delta_0^2}{\delta_\infty^2 - \delta_0^2} \tag{7}$$

such that q_e satisfies

$$q_e = 1 - e^{-2\lambda t} \tag{8}$$

Then, given the number (N) of interaction queries, the interaction distance defined by $I_{AB}=2\pi_1\pi_0\lambda t$, is given by

$$I_{AB} = -2\pi_1\pi_0 \ln(1 - q_e) \tag{9}$$

Results and Discussion

Based on the theoretical framework formulated above, we develop a computational procedure that is suitable to a variety of OMICS data types.

The beta-uniform mixture (BUM) distribution of *p*-values

While *p*-values arising from the null hypothesis are distributed uniformly on the

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interval (0, 1), those arising from the alternative hypothesis follows a distribution denoted by φ(*p*), which is generally modeled as a beta distribution. Therefore, the distribution of the set of *p*-values can be written as a beta-uniform mixture (BUM) consisting of a uniform (0, 1) component for the null hypothesis and the beta component for the alternative hypothesis, with the pdf given by

$$\Phi(p) = u_0 + (1 - u_0)\phi(p) \qquad (10)$$

where π_0 is the proportion of null hypothesis (r=0); and the second term for the Beta component. Here we implement a special form of beta distribution by setting β_0=1, resulting in

$$\varphi(p)=\beta p^{\beta-1} \qquad (11)$$

for $0<p\leq 1$, $0<\pi_0<1$, and $0<\beta<1$. As shown in Fig.1(c), Eq.(11) provides a reasonable model for the distribution of *p*-values arising from high throughput genomics (Storey and Tibshirani 2003). It is a curve that asymptotes at x=0 and monotonically decreases to its minimum of $\pi_0+(1-\pi_0)\beta$ at x=1. This curve approximates the anticipated distribution of the *p*-values arising from a genomic experiment. Under the null hypothesis, the *p*-values will have a uniform density corresponding to a flat horizontal line. Under the alternative hypothesis, the *p*-values will have a distribution that has high density for small *p*-values and the density will decrease as the *p*-values increase. The overall distribution will be a mixture of *p*-values arising from the two hypotheses.

Parameters of interaction uncertainty under the BUM model

Suppose that y_A and y_B are independent, which follow the same distribution conditional of r=0 or 1. Then $\boldsymbol{\gamma}_{rA,\,rB}$ defined by Eq.(2) can be further simplified as follows

$$\begin{aligned}\gamma_{r_A,r_B} &= E[y_A^2|r_A] + E[y_B^2|r_B] - 2E[y_A|r_A]E[y_B|r_B] \\ &= E[y^2|r_A] + E[y^2|r_B] - 2E[y|r_A]E[y|r_B]\end{aligned} \qquad (12)$$

To calculate $\boldsymbol{\gamma}_{rA,\,rB}$, we have to consider the null hypothesis of r=0 (no interaction) and the alternative separately. Under the null hypothesis, the *p*-value follows a uniform distribution. For a given *p*-value transformation, $y=f(p)$, let $\bar{y}_0$ and σ_0^2 be the mean and variance of y under the null r=0, respectively, which are given by

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned} \bar{y}_0 &= \int_0^1 f(p)dp \\ \sigma_0^2 &= \int_0^1 f^2(p)dp - (\bar{y}_0)^2 \end{aligned} \tag{13}$$

One may easily show that $E[y|r=0]= \bar{y}_0$ and $E[y^2|r=0]=(\bar{y}_0)^2+\sigma_0^2$. Under the alternative hypothesis of r=1, the distribution of p, denoted by $\varphi(p)$, is usually modeled by a beta distribution given by Eq.(11). Similar to above, let $\bar{y}_1$ and σ_1^2 be the mean and variance of y under the alternative of r=1, respectively, as given by

$$\begin{aligned} \bar{y}_1 &= \int_0^1 f(p)\phi(p)dp \\ \sigma_1^2 &= \int_0^1 f^2\phi(p)dp - (\bar{y}_1)^2 \end{aligned} \tag{14}$$

Obviously we have $E[y|r=1]= \bar{y}_1$ and $E[y^2|r=1]=(\bar{y}_1)^2+\sigma_1^2$. Putting together, those coefficients of interaction uncertainty (γ_{11}, γ_{10}, γ_{01} and γ_{00}) defined by Eq.(12) are given by

$$\begin{aligned} \gamma_{11} &= 2\sigma_1^2 \\ \gamma_{10} &= \sigma_1^2 + \sigma_0^2 + (\bar{y}_1 - \bar{y}_0)^2 \\ \gamma_{01} &= \sigma_1^2 + \sigma_0^2 + (\bar{y}_1 - \bar{y}_0)^2 \\ \gamma_{00} &= 2\sigma_0^2 \end{aligned} \tag{15}$$

respectively.

Types of *p*-value transformation

We shall implement a statistical procedure to estimate those parameters on the right hands of Eq.(15), as long as the p-value transformation is specified. In this study we consider two p-value transformations: the p-based, and the $-\ln(p+1/M)$-based, where M is the total number of interactions.

The p-based method. The simplest one is to use the p-value directly, i.e., $y=p$. Note that under the null of no interaction (r=0), p follows a uniform distribution in [0,1] with the mean 1/2 and the variance 1/12. In this case, Eq.(15) can be simplified as follows

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned} \gamma_{11} &= 2\sigma_p^2 \\ \gamma_{10} &= \sigma_p^2 + 1/12 + (\bar{p} - 1/2)^2 \\ \gamma_{00} &= 1/6 \end{aligned} \tag{16}$$

where $\bar{p} = \beta/(1+\beta)$ and $\sigma_p^2 = \bar{p}/(2+\beta)$ are the mean and variance of *p*-values under the beta distribution of Eq.(11).

The -ln(p+1/M)-based method

In spite of simplicity, we have realized that $y=p$ transformation may be subject to some sampling problems. For instance, an interaction associated with a *p*-value of 0.001 is statistically sound, while that with more than 0.05 is usually considered as non-significance. Consider a hypothetical example that for two duplicate genes *A* and *B*. Assume p_A=0.001 and p_B=0.5 for the functional interactions A-X and B-X, respectively, so that $(p_A-p_B)^2=0.4999^2$. Secondly, for another interaction pair A-X' and B-X', we assume that the *p*-values are p'_A=0.30 and p'_B=0.8 so that $(p'_A-p'_B)^2=0.5^2$. The virtually same score between two cases is apparently counter-intuitive, because one may statistically infer that gene *A* interacts with *X* but not for gene *B*, whereas both *A* and *B* are unlikely to interact with gene *X'*.

We try to use a (negative) log-transformation score (y) for the *p*-value of an interaction, i.e., $y=-\ln p$, to avoid this difficulty. In the above case, we observed that $(y_A-y_B)^2=6.22^2$ (for A-X and B-X) is much higher than $(y'_A-y'_B)^2=0.98^2$ (A-X' and B-X'), which makes much more sense. While this log-transformation can effectively reduce the random effects, one problem is that $y=-\ln p$ does not converge when p approaches to zero, which may cause a considerable (upward) sampling bias. We thus recommend a simple correction by adding a pseudo-count, that is, $y=-\ln(p+1/M)$, where M is the total number interactions under study.

Since M is usually large, one can show that under the null hypothesis y approximately follows an exponential distribution with the mean $\bar{y}_0$ and variance σ_0^2. Together, we have

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned} \gamma_{11} &= 2\sigma^2 \\ \gamma_{10} &= \sigma^2 + 1 + (\bar{y} - 1)^2 \\ \gamma_{00} &= 2 \end{aligned} \tag{17}$$

where $\bar{p}$ and σ^2 are the mean and variation of y-values under the alternative hypothesis of r=1, respectively.

Computational procedure and implementation

Under the *R*-environment, we implement the following computational procedure to estimate the evolutionary distance of interactions based on p-values.

(i) Given high throughput p-values, we use several statistical methods to estimate π_0, the proportion of null hypotheses of interactions. First, the LOESS-based method (Pounds and Cheng 2004) applies LOESS to the p-value spacing to obtain an estimate of PDF, and takes the minimum value of the estimated PDF as an estimate of π_0. Second, the CDF-based method (Cheng et al. 2004) uses an estimator for the CDF (Cumulative Distribution Function) of p-values in the form of a B-spline series with strategically designed knot sequence to achieve a desirable shape for the p-value cumulative distribution. The PDF is simply the first derivative of the CDF and the minimum of PDF is taken as an estimate of π_0. Third, exploiting the fact that p-values from true null hypotheses are uniformly distributed, the asymptotic uniform method (Storey and Tibshirani method 2003) uses a tunable estimate, that is, $\pi_0(c)= \#\{p_i>c; i=1,...M\}/M(1-c)$, with c as the tuning parameter. It then fits a natural cubic spline with 3 degrees of freedom to the data of $\pi_0(c)$ on a series of values of c such as c=0, 0.05,…, 0.90, and finally the value of the fitted spline line at the end point c=1 is taken as the estimate of π_0. The technical details of these methods can be found in the original publications.

(ii) Treating the estimate of π_0 as known, the quasi-maximum likelihood estimate of parameter β can be obtained by fitting the beta-uniform mixture (BUM) model in Eq.(11) to the observed p-values.

(iii) Calculation of γ_{11}, γ_{10}, γ_{01} and γ_{00} numerically for $y=p$ and $y=\ln(p+1/M)$,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively, leading to the estimation of δ^2_∞ and δ^2_0 by Eq.(6).

(iv) Given the sample size (N) of interaction queries, δ^2_{AB} can be calculated as follows

$$\delta^2_{AB} = \sum_{k=1}^{N} (y_{A,k} - y_{B,k})^2 / N \quad (18)$$

The effective proportion of different interactions between genes (q_e) can be estimated according to Eq.(7). Then the evolutionary distance of interactions distance can be estimated by Eq.(9). In short, estimation of I_{AB} turns out to estimate the effective proportion of different interactions between genes, which can be achieved when the transformed p-value (the y-score) is specified.

Simulation study

Statistical uncertainty of interactions

Experimental noises in high throughput genomics data could make any biological analysis unreliable. Hence we wish to investigate this effect on the estimation of interaction distance. Since the value of parameter β is inversely related to the level of experimental noise, we design the following simulation study pipeline. We choose β= =0.01, 0.05, 0.1, 0.3, 0.5, and 0.8. Intuitively, a low β value indicates a high number of biological replicates, and *vice versa*. According to Eq.(11), these values correspond to type-II error at the 0.05 significance level of 0.03, 0.14, 0.26, 0.59, 0.78, and 0.85, respectively, as calculated by $1-\int_0^{0.05} \varphi(p)dp$.

Simulation theme of interaction evolution

Under the BUM model of interaction uncertainty, the simulation theme of interaction evolution is designed as follows. (*i*) Given the evolutionary scenario of two species with N interaction queries, simulate the interaction pattern according to the Markov-chain model. We set the interaction distance I_{AB}=0.1, 0.3, 0.5, 0.8, or 1.0, respectively; the number of interaction queries is N=100, 200 or 500, respectively; and the proportion of active interaction is set to be π_1=0.1, 0.3, 0.5, 0.7 or 0.8, respectively. (*ii*) In each case, we estimate those interaction uncertainty parameters for two p-value transformations, $y=p$, and $y=-\ln(p+1/M)$, respectively, as well as the evolutionary distance. And (*iii*) carry out 1000 simulation replicates and analyze the statistical

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

properties in each case. The goal of our simulation study is to examine the systematic bias and sampling variance of I_{AB} under various conditions (Table 1).

Effects of p-value transformation

When the simplest $y=p$ transformation is used, our simulation shows a large sampling variance and severe underestimation bias, especially in those cases when the parameter β is set be large (>0.5). Indeed, a small β value (<0.05) can effectively reduce both sampling variance and bias. Note that a low β value indicates a high number of biological replicates, our observation suggests that the performance of $y=p$ transformation for the estimation of I_{AB} is reasonable acceptable only when the number of biological replicates is sufficient. Moreover, for closely related species when the number of alternative hypotheses is small, estimation of I_{AB} becomes biologically meaningless when the network uncertainty is overwhelming (a large β value) (Table 1). These results, together, suggest that, using the p-value directly without any transformation is not recommended, because the estimated evolutionary distance of interactions could be highly sensitive to the network uncertainty.

By contrast, the $-\ln(p+1/M)$-distance shows some nice statistical properties in the case of low biological replicates; both sampling variance and estimation bias are generally acceptable. We have examined the effect of the pseudo-count ($1/M$). Indeed, the $-\ln(p)$-distance becomes statistically unreliable in the case of small β value (<0.05), especially when the number of query genes N is small. This observation can be explained as some genes may receive very low p-values (very high significance levels), resulting in some extremely high $y=-\ln(p)$ values. Fortunately, this defect can be effectively corrected by the $-\ln(p+1/M)$ distance. While the $-\ln(p+1/M)$ distance overall performs satisfactory, it tends to underestimate the distance when the number of query genes is small, such as less than 200.

Effects of evolutionary parameters

First, the interaction distance estimation I_{AB}, as expected, is asymptotically unbiased, i.e., the estimate I_{AB} is statistically unbiased for sufficiently large sampling size of interaction queries (N). Similarly, the sampling variance of I_{AB} decreases with the

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sampling size of interaction queries (N). Second, the asymptotic rate is dramatically affected by the parameter β in the BUM model. It becomes very slow when β is large, suggesting that an accurate estimation of interaction distance becomes difficult when the experimental noise is high. Third, the proportion of NA (not applicable) cases increases with the increase of β, which makes the distance estimation practically not useful particular when the sampling size (N) is small. Forth, as a general tendency, the distance estimation usually has nice statistical properties in the cases around $\pi_1=\pi_0=0.5$.

Outlook for further study

Our further study will be focused on how to improve the efficiency and applicability of the proposed method. Several research lines are considered. For instance, we try to find the *p*-value transformation such that it not only can optimize statistical properties but also biologically interpretable. In the case of multiple genes that may represent a gene family evolution or species evolution, one may develop a distance-based approach to investigating the evolutionary pattern of interactions. Yet, it remains a challenge to develop a generalized likelihood function of interactions under a phylogeny when the interaction uncertainty is taken into consideration.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. A summary of simulation studies to evaluate the statistical performance of the interaction distance.

	Distance (I_{AB})		
y	*p*	*-ln p*	*-ln (p+1/M)*
β=0.8			
I_{true}= 0.1	0.033±0.503	0.105±0.043	0.106±0.024
I_{true}= 0.5	0.409±0.204	0.477±0.050	0.475±0.053
I_{true}= 1.0	0.845±0.308	0.945±0.102	0.986±0.103
β=0.5			
I_{true}= 0.1	0.087±0.107	0.107±0.023	0.103±0.022
I_{true}= 0.5	0.468±0.103	0.568±0.043	0.511±0.040
I_{true}= 1.0	0.944±0.308	1.544±0.603	0.974±0.407
β=0.10			
I_{true}= 0.1	0.092±0.036	0.122±0.036	0.102±0.017
I_{true}= 0.5	0.485±0.132	0.582±0.315	0.512±0.036
I_{true}= 1.0	0.968±0.167	1.868±0.804	0.968±0.042

Note: In each case, the number of query genes (N) is set to be 500 and π_1=π_0=0.5. Sampling variances are calculated based on 1000 simulation replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. Nature reviews. Genetics 12, 56 (2011).

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

Benjamini, Y. and D. Yekutieli (2001): "The control of the false discovery rate in multiple testing under dependency," Ann Stat, 29, 1165–88.

Cheng C, Pounds SB, Boyett JM, Pei D, Kuo ML, et al. (2004) Statistical Significance Threshold Criteria For Analysis of Microarray Gene Expression Data. Statistical Applications in Genetics and Molecular Biology 3: Article 36.

Efron, B. (2004): "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," J Am Stat Assoc, 99, 96–104.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): "Empirical bayes analysis of a microarray experiment," Journal of Computational Biology, 96, 1151–60.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: Evolutionary rate in the protein interaction network. *Science* 2002, 296:750-752.

Fraser, H. B. Modularity and evolutionary constraint on proteins. Nature genetics 37, 351 (2005).

Grindrod, P. & Kibble, M. Review of uses of network and graph theory concepts within proteomics. Expert review of proteomics 1, 229–238 (2004).

Gursoy, A., Keskin, O. & Nussinov, R. Topological properties of protein interaction networks from a structural perspective (2008).

Gustafsson, M. et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome medicine 6, 82 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Halton, M.S. 2017. Perspective on gene regulation evolution. Trends in Genetics 33:436-447.

Hao, T., Peng, W., Wang, Q., Wang, B. & Sun, J. Reconstruction and application of protein–protein interaction network. International journal of molecular sciences 17, 907 (2016). 3.

Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. Nature 402, C47 (1999).

Jordan IK, Wolf YI, Koonin EV: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol Biol* 2003, 3:1.

Kim WK, Marcotte EM: Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Comput Biol* 2008, 4:e1000232.

Ma'ayan, A. Network integration and graph analysis in mammalian molecular systems biology. IET systems biology 2, 206–221 (2008).

Mosca, R., Pons, T., Céol, A., Valencia, A. & Aloy, P. Towards a detailed atlas of protein–protein interactions. Current opinion in structural biology 23, 929–940 (2013).

Nibbe, R. K., Koyutürk, M. & Chance, M. R. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. PLoS computational biology 6, e1000639 (2010).

Pounds, S. and S. W. Morris (2003): "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values," Bioinformatics, 19, 1236–1242.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Procaccini, C. et al. The proteomic landscape of human ex vivo regulatory and conventional t cells reveals specific metabolic requirements. Immunity 44, 406–421 (2016).

Shaffer, J. P. (1995): "Multiple hypothesis testing," Annu. Rev. Psychol., 46, 561–84.

Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. Molecular systems biology 3, 88 (2007).

Smyth, G. K. (2004): "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Statistical Applications in Genetics and Molecular Biology, 1, 3.

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

Sun M, Kim P (20111) Evolution of biological interaction networks: from models to real data. *Genome Biology* 2011, **12**:235

Vella, D., Zoppis, I., Mauri, G., Mauri, P. & Di Silvestre, D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. EURASIP Journal on Bioinformatics and Systems Biology 2017, 6 (2017).

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002, 417:399-403.

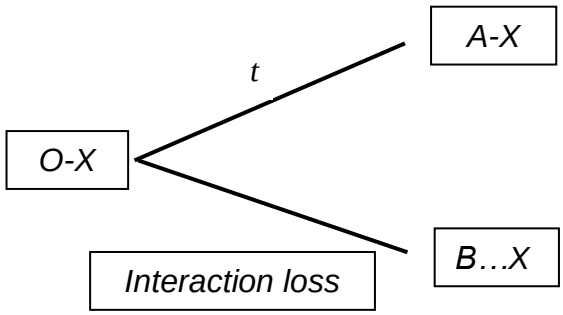
Wagner A: The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 2001, 18:1283-1292.

Wagner A: How the global structure of protein interaction networks evolves. *Proc Biol Sci* 2003, 270:457-466.

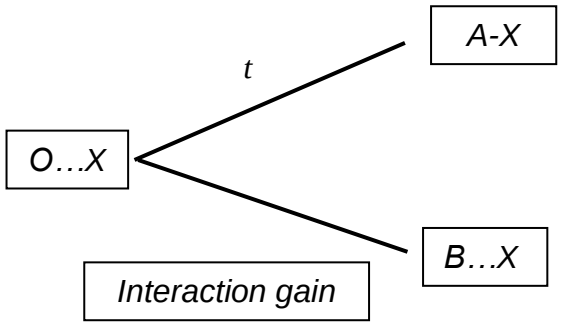
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 1. Schematic illustration of interaction evolution. Suppose two homologous genes (*A* and *B*) diverged *t* time units ago, via either speciation or gene duplication. For a given interaction query gene (X), gene A has an active interaction with X (solid line) whereas gene B has no interaction with X (dashed line). There are two possibilities for their common ancestor (gene O): in the case of active interaction between O and X, there is an interaction loss in the B-lineage (panel a), otherwise there is an interaction gain in the A-lineage (panel b). (c) Schematic illustration of the BUM distribution for *p*-values. Region A corresponds to the occurrence of true positives; region B corresponds to the occurrence of false negatives; Region C corresponds to the occurrence of false positives and region D corresponds to the occurrence of true negatives.

(a)



(b)



bioRxiv preprint doi: https://doi.org/10.1101/2020.03.08.982710; this version posted March 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(c)