The Annals of Applied Statistics
2021, Vol. 15, No. 2, 902–924
https://doi.org/10.1214/20-AOAS1431

LARGE-SCALE MULTIPLE INFERENCE OF COLLECTIVE DEPENDENCE WITH APPLICATIONS TO PROTEIN FUNCTION

BY ROBERT JERNIGAN[1,*], KEJUE JIA[1,†], ZHAO REN[2] AND WEN ZHOU[3]

[1]*Department of Biochemistry, Biophysics, and Molecular Biology, Program of Bioinformatics and Computational Biology, Iowa State University,* [*]*jernigan@iastate.edu;* [†]*kjia@iastate.edu*

[2]*Department of Statistics, University of Pittsburgh, zren@pitt.edu*

[3]*Department of Statistics, Colorado State University, riczw@stat.colostate.edu*

Measuring the dependence of $k \geq 3$ random variables and drawing inference from such higher-order dependences are scientifically important yet challenging. Motivated here by protein coevolution with multivariate categorical features, we consider an information theoretic measure of higher-order dependence. The proposed collective dependence is a symmetrization of differential interaction information which generalizes the mutual information of a pair of random variables. We show that the collective dependence can be easily estimated and facilitates a test on the dependence of $k \geq 3$ random variables. Upon carefully exploring the null space of collective dependence, we devise a **C**lassification-**A**ssisted **L**arge sca**L**e inference procedure to **DE**tect significant k-**CO**llective **DE**pendence among $d \geq k$ random variables, with the false discovery rate controlled. Finite sample performance of our method is examined via simulations. We apply this method to the multiple protein sequence alignment data to study the residue or position coevolution for two protein families, the elongation factor P family and the zinc knuckle family. We identify novel functional triplets of amino acid residues, whose contributions to the protein function are further investigated. These confirm that the collective dependence does yield additional information important for understanding the protein coevolution compared to the pairwise measures.

1. Introduction. Ever since the pioneering observation on evolutionary interactions between flowering plants and insects by Charles Darwin (Darwin (1859)), coevolution has been an important topic of evolutionary theory to understand the interdependences among living systems (Thompson (1994)). Defined as the coordinated changes taking place among organisms or biomolecules, coevolution typically stabilizes or improves functional interactions among the various parts. Rapid recent advances in genome sequencing has led to growing numbers of studies of coevolution at the molecular level (Fraser et al. (2004), Cusick et al. (2005), Chao et al. (2008), Ochoa and Pazos (2014)). In computational biology, multiple sequence alignments (MSA) for protein families have become accessible for large numbers of species, and the identification of coevolution signals from an MSA has proven to be highly effective for the analysis and prediction of protein structure and function (Afonnikov and Kolchanov (2004), Wells and McClendon (2007), Burger and van Nimwegen (2010), Morcos et al. (2011)). Amino acid mutations within or among protein sequences is considered to be a key factor for understanding protein function. When multiple residues of a protein or even multiple interacting proteins have a change at one position in a functional protein, one then normally expects to find compensating changes at other positions to maintain its normal function. That is, the residue level coevolution can be inferred from the simultaneous mutation among multiple positions in the MSA.

Received December 2019; revised November 2020.

Key words and phrases. Collective dependence, false discovery rate, information theoretic measure, multiple testing, protein coevolution, structural biology.

While the experimental identification of coevolving residues in a protein sequence would be arduous and/or serendipitous, computational methods are simple and have shown success in a wide range of applications (de Juan, Pazos and Valencia (2015)). For example, mutual information has been exploited to predict coevolved amino acids and applied to allosteric pathway identification, functional site detection and residue contact prediction (Korber et al. (1993), Martin et al. (2005), Dunn, Wahl and Gloor (2008)). A mutual information based network approach was proposed to identify catalytic residues by Buslje et al. (2010). Adjusted for the indirect or transitive effects, methods focusing on residue contact prediction have been introduced, such as the direct coupling analysis (Weigt et al. (2009), Morcos et al. (2011)). Utilizing the correspondence between the graph structure and the precision matrix, the protein sparse inverse covariance estimation was proposed by Jones et al. (2011), with similar approaches being commonly exploited in genetic network analysis (Yin and Li (2011), Zhi et al. (2013), Cai et al. (2013), Wang et al. (2015), Xia, Cai and Cai (2018)). By using the so-called statistical potential energy, approaches, such as the statistical coupling analysis, explicitly search for groups of coevolving residues to identify the allosteric pathways within proteins (Lockless and Ranganathan (1999), Halabi et al. (2009), Reynolds, McLaughlin and Ranganathan (2011)). Coevolution analysis has also been applied to improve protein sequence matching (Jia and Jernigna (2018)), understand protein dynamics (Mishra and Jernigna (2018)) and detect deleterious mutations (Hopf et al. (2017)).

Existing measures to extract coevolution signals from the MSA primarily focus on pairwise coherence. Despite their popularity, any extension to higher-order scenarios, where the coevolution would involve $k \geq 3$ residues simultaneously, has not often been carried out even though in practice the intrinsic nature of protein structures is to have such multiple interactions, and these should be considered for predicting the fitness of amino acid substitutions in proteins or evolutionarily conserved protein-protein interactions. In other applications, certain higher-order dependences have been employed to detect functional modules in oncogenic networks (Ciriello et al. (2012), Bueno and Mar (2017)) as well as to define higher-order linkage disequilibrium among three or more alleles in genetics (Slatkin (2008)). Most approaches employ pairwise measures to identify the higher-order dependences, such as the three-residue correlation (Figliuzzi, Barrat-Charlaix and Weigt (2018)). This, however, presents some fundamental challenges. Physically, it is not always realistic to assume an arbitrary network to be represented by the collection of many two-body interactions, even for some ideal models such as the Potts model. Moreover, it is well known in probability that the pairwise dependence, and therefore bivariate measures such as the correlation or the distance covariance, cannot be used directly to detect the joint dependence. A classical example is the Bernstein's coin, as detailed in Section A.1 of the Supplementary Material (Jernigan et al. (2020)).

To directly model the higher-order dependence, various statistical approaches have been identified because of their flexibility with respect to the data type and capability for drawing inferences. For example, by modeling the higher-order dependence as edges containing more than two vertices, the random hypergraphs, such as those from the generalized Erdös–Rényi model, have been employed to study complex community structures beyond pairwise associations (Klamt, Haus and Theis (2009), Yuan et al. (2018)). Using compound Poisson processes, Staude, Rotter and Grün (2010) and Staude, Grün and Rotter (2010) proposed testing procedures to detect higher-order correlations and cumulants in neural spike counts. A similar but less parametric approach has been developed based on the maximum entropy distribution (Onken, Dragoi and Obermayer (2012)). The aforementioned methods either rely on the correct model specification or continuous measurements and, therefore, are not directly applicable to the general discrete and categorical data, such as the MSA data. In addition, the higher-order dependence may encounter certain degeneracy that permits the full knowledge of all variables, given only one in the group, which is referred to as the full dependence (Galas et al. (2010, 2014)). That is, each random variable is deterministically specified

by another one in the group, and, therefore, the joint distribution is degenerate. On the one hand, such full dependence should have already been detected in the pairwise dependence relationship, and, ideally, it should be avoided in the higher-order dependence detection. On the other hand, from the information theoretic point of view, such a degeneracy does not incur for authentic higher-order dependences which require full knowledge of all others collectively to predict one variable. In fact, no information will lose by removing all but one variables from the group (Fleuret (2004), Galas et al. (2010)). In other words, the group is degenerate in the sense that its total entropy equals to the entropy of any individual variable in the group. Finally, inference on the higher-order dependence is by nature of high dimension and requires careful control on the false discover rate (FDR, Benjamini and Hochberg (1995), Dudoit and van der Laan (2008), Sun, Zhang and Owen (2012), Cai and Liu (2016)). However, the large-scale multiple inference of existing measures of higher-order dependence has not been fully explored.

To circumvent these challenges and detect nonfull-dependence higher-order dependences among residues or positions on the MSA, we consider an information theoretic measure, the differential interaction information (DII) and its symmetrization, collective dependence (Galas et al. (2014)). DII and the collective dependence are built upon the interaction information, a multivariate generalization of mutual information. Compared to existing measures, collective dependence vanishes for two extreme cases, full dependence and full independence. In addition, collective dependence accommodates discrete and categorical data more naturally as its building block; the information entropy originally was for an arbitrary discrete distribution and can be effectively estimated based on independent and identically distributed (i.i.d.) observations (Paninski (2003)). Specifically, take the trivariate $\{X_1, X_2, X_3\}$ as an example; DII of X_1 with respect to others is defined by $\mathsf{H}(X_1) - \mathsf{H}(X_1, X_2) - \mathsf{H}(X_1, X_3) + \mathsf{H}(X_1, X_2, X_3)$, where $\mathsf{H}(\cdot)$ denotes the marginal entropy of the corresponding random variables. The collective dependence is then defined as the product of DIIs of each random variable with respect to others (see Section 2 for further details). It can be shown that collective dependence vanishes if a subset of $\{X_1, X_2, X_3\}$ is independent of the remaining variables. In addition, it is also zero for the other extreme case, full dependence. On the other hand, it is nonzero if the trivariate is dependent but not degenerate due to the full dependence. Hence, the collective dependence provides a model-free measure to identify higher-order dependence in multivariate discrete and categorical distributions.

As noted, identification of all k (≥ 3) variables with significant collective dependence from a d-dimensional random vector is a large-scale multiple inference problem with $\binom{d}{k}$ hypotheses, for which FDR controlling procedures can be employed. However, as the collective dependence is the product of individual DIIs, the null space of each test, which corresponds to the lack of higher-order collective dependence, is composite in nature. A naive testing procedure common for all possible distributions under the true null will compromise the power. Whenever additional information on the potential nonzero DIIs is provided, one can improve the power of each test by constructing specific null distribution of the test statistic. To that end, we propose a **C**lassification-**A**ssisted **L**arge sca**L**e inference to **DE**tect **CO**llective **DE**pendence (CALL-DECODE) in this paper. We first estimate k DIIs for each of the $\binom{d}{k}$ groups of random variables using certain debiased entropy estimator (Paninski (2003)). Then, we classify all these $\binom{d}{k}$ groups to mutually exclusive subsets by detecting potential nonzero DIIs in comparison to some large deviation bounds. Leveraging such a classification, the empirical false discovery proportion (FDP) is computed using subset-specific cutoffs and, therefore, with respect to the nominal FDR, defines the threshold for identification of random variables with significant collective dependence.

Collective dependence and our method, CALL-DECODE, can be easily implemented and used to identify coevolution signals for $k \geq 3$ residues or positions from the MSA. For example, in our study on protein coevolution of the elongation factor P family and the zinc knuckle

family, we focus on $k = 3$ and show that our method is capable of identifying the important residues connecting triplets with significant third order collective dependence. Those novel hub residues, such as Lys14, Phe16, Asn17, Ala25 and Lys26 from the zinc knuckle family, are mostly missed by methods based on mutual information. They are critical for the protein's function such as interacting with RNA/DNA and are consistent with recent experimental results and discussions in literature.

1.1. *Outline.* In the next section we introduce the collective dependence to quantify associations of $k \geq 3$ variables from a d-dimensional random vector. Section 3 provides details of the CALL-DECODE procedure. For simplicity of presentation, we focus on $k = 3$ corresponding to triplets of random variables. We report numerical studies in Section 4 to demonstrate the finite sample performance of our method. In Section 5 we present the results of two applications of collective dependence and CALL-DECODE to the motivating protein coevolution problems. The paper is concluded in Section 6 with discussions. Technical discussions, additional numerical experiments and extra results for the real data analysis are deferred to the Supplementary Material (Jernigan et al. (2020)).

2. Collective dependence. In this section we briefly review the differential interaction information (DII) and introduce the collective dependence (Galas et al. (2014)) as a measure to quantify the higher-order dependence for $k \geq 3$ random variables.

Assume that independent d-dimensional observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. copies of a discrete random vector $\mathbf{X} = (X_1, \ldots, X_d)$. The marginal probability function of $\mathbf{X}$ is defined by $\mathbb{P}(X_b = l) = p_{b,l}$ for each $b = 1, \ldots, d$, where $l \in \mathcal{X}_b = \{0, 1, \ldots, L_b\}$ with known $L_b \geq 1$, and $p_{b,l} \in [0, 1]$, $\sum_{l=0}^{L_b} p_{b,l} = 1$. Motivated by our application to protein coevolution, we assume each L_b remains fixed while the sample size n and dimension d can diverge. Besides discrete type distribution with finite possible outcomes for each coordinate, there is no other distributional assumption on the joint probability function. We also call it distribution-free model.

To formally introduce DII, we recall that the mutual information is a fundamental information-theoretic quantity to measure the codependence between two random variables, which is $\mathsf{I}(X_1, X_2) = \mathsf{H}(X_1) + \mathsf{H}(X_2) - \mathsf{H}(X_1, X_2)$, where $\mathsf{H}(\nu)$ is the joint entropy of a general collection of random variables ν. One popular way of generalizing of the mutual information to the multivariate case is the so-called interaction information (McGill (1954)). This quantity has been studied from different view points in literature (McGill (1954), Yeung (1991), Galas et al. (2010)). In the case of three random variables with $k = 3$, the interaction information can be written as the gain (or loss) in mutual information between any two of the variables, due to additional knowledge of the third random variable. That is, $\mathsf{I}(X_1, X_2, X_3) = \mathsf{I}(X_1, X_2) - \mathsf{I}(X_1, X_2|X_3) = \mathsf{H}(X_1) + \mathsf{H}(X_2) + \mathsf{H}(X_2) - \mathsf{H}(X_1, X_2) - \mathsf{H}(X_1, X_3) - \mathsf{H}(X_2, X_3) + \mathsf{H}(X_1, X_2, X_3)$. Intuitively, this is the amount of information shared by all three variables together. It enjoys many desired properties such as permutation symmetric and that it vanishes if any one variable is independent of the other two. The general interaction information of a collection of random variables ν can be defined using the inclusion-exclusion formula (Bell (2003)) as follows:

$$\mathsf{I}(\nu) = \sum_{\{\tau\}} (-1)^{|\tau|+1} \mathsf{H}(\tau),$$

where $\{\tau\}$ is the collection of all subsets of ν and $|\tau|$ denotes the cardinality of τ. This general interaction information is the amount of information shared by all the variables together and has similar good properties as the case $k = 3$ does.

However, the downside of using interaction information as our measure is that it does not vanish for the full dependence scenario but this extreme degeneracy should be avoided in

higher-order interaction detection (Fleuret (2004), Galas et al. (2010)). While there are more than one way to get zeros at this extreme, Galas et al. (2014) suggest a natural choice called the differential interaction information (DII). Specifically, for a target variable $X \in \nu$, the DII of X with respect to others, $\Delta_{\nu,X}$ is defined by the change in the interaction information between sets that differ only by the addition of X, that is, $\Delta_{\nu,X} = \mathsf{I}(\nu) - \mathsf{I}(\nu \setminus \{X\})$. It also can be written in terms of the marginal entropies by the inclusion-exclusion formula of $\mathsf{I}(\nu)$ as follows:

$$\Delta_{\nu,X} = \sum_{\{\tau_X\}} (-1)^{|\tau_X|+1} \mathsf{H}(\tau_X), \tag{2.1}$$

where $\{\tau_X\}$ is the collection of all subsets of ν that contain X. It is interesting to notice that DII is equivalent to the conditional interaction information $\Delta_{\nu,X} = -\mathsf{I}(\nu \setminus \{X\}|X)$ which is equivalent to the conditional mutual information for $k = 3$ (Fleuret (2004), Vejmelka and Paluš (2008)). This definition successfully tackles the challenge due to full dependence since for this case any variable, say Y, fully determines all other variables which further implies $\Delta_{\nu,X} = \mathsf{I}(\nu) - \mathsf{I}(\nu \setminus \{X\}) = \mathsf{I}(Y) - \mathsf{I}(Y) = 0$. By the equivalent definition in (2.1), one can easily see that $\Delta_{\nu,X} = 0$ if any variable in $\nu \setminus \{X\}$ is independent of all remaining variables. However, $\Delta_{\nu,X}$ is not permutation symmetric, as it specifies a target variable X. In addition, one can see that $\Delta_{\nu,X}$ may not vanish if X is independent of all remaining variables. For example, when $k = 3$ and $\nu = \{X_1, X_2, X_3\}$ with X_3 independent from $X_1 = X_2$, it is easy to obtain that $\Delta_{\nu,X_3} = -\mathsf{H}(X_1) = -\mathsf{H}(X_2)$. Finally, to solve these two issues while keeping the property that $\Delta_{\nu,X} = 0$ for the full dependence scenario, Galas et al. (2014) define the collective dependence Δ_ν of a collection of random variables ν by the symmetrization of DII for all variables within ν as follows:

$$\Delta_\nu = (-1)^{|\nu|} \prod_{X \in \nu} \Delta_{\nu,X}. \tag{2.2}$$

We list the desired properties of the collective dependence in the following proposition which is from Galas et al. (2014). We refer readers to Galas et al. (2014) for further details and explanation of the collective dependence, including its connection to the so-called set complexity in biological system.

PROPOSITION 1 (Galas et al. (2014)). *Given random variables* $\nu = \{X_1, \ldots, X_d\}$, *the collective dependence defined in* (2.2) *vanishes if*: (i) *all variables are independent*, *or* (ii) *all the variables are fully dependent or* (iii) *there exists* X_j *independent from the remaining variables for some* $j \in \{1, \ldots, d\}$.

For each $k \geq 3$, we introduce the collective dependence for all k-groups of d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$. Denote $\nu_k \subset \{X_1, \ldots, X_d\}$, where $|\nu_k| = k$, that is $\nu_k = \{X_{h_1}, \ldots, X_{h_k}\}$ with $\{h_1, \ldots, h_k\} \subset \{1, \ldots, d\}$. DII of variable X_{h_j} with respect to a group of k variables, Δ_{ν_k,h_j}, is defined as follows, where $\{\tau_j\}$ is the collection of all subsets of ν_k that contain X_{h_j}:

$$\Delta_{\nu_k,h_j} = \sum_{\{\tau_j\}} (-1)^{|\tau_j|+1} \mathsf{H}(\tau_j). \tag{2.3}$$

Equipped with DII Δ_{ν_k,h_j}, the collective dependence among k variables specified by ν_k, Δ_{ν_k} is then defined as follows:

$$\Delta_{\nu_k} = (-1)^k \prod_{j=1}^{k} \Delta_{\nu_k,h_j}.$$

For a particular k-group ν_k, Δ_{ν_k}, therefore, measures the higher-order dependence among variables specified by ν_k. As demonstrated by Proposition 1, $\Delta_{\nu_k} = 0$ whenever the k variables in ν_k are independent, or fully dependent or any individual variable in ν_k is independent of the others. That is, $\Delta_{\nu_k} = 0$ implies the lack of collective dependence for variables in a particular ν_k.

We close this section by proposing a simple estimator $\widehat{\Delta}_{\nu_k,h_j}$ of each DII Δ_{ν_k,h_j}. Specifically, for small k, Δ_{ν_k} can be estimated well through the estimation of each entropy. Given the distribution-free model discussed before, one can estimate $\mathsf{H}(\tau_j)$ by the maximum likelihood estimator (MLE), that is, $\widehat{\mathsf{H}}_{\mathrm{MLE}}(\tau_j) = -\sum_{\mathbf{c}_{\tau_j}} \widehat{f}_{\tau_j}(\mathbf{c}_{\tau_j}) \log(\widehat{f}_{\tau_j}(\mathbf{c}_{\tau_j}))$, where $\widehat{f}_{\tau_j}(\mathbf{c}_{\tau_j}) = n^{-1}\sum_{i=1}^{n} \mathbb{I}(\mathbf{X}_{i,\tau_j} = \mathbf{c}_{\tau_j})$ with $\mathbf{c}_{\tau_j} \in \mathcal{X}_{\tau_j}$, the set of all possible values of $\mathbf{X}_{\tau_j}$. For fixed k, this simple "plug-in" estimator $\widehat{\mathsf{H}}_{\mathrm{MLE}}(\tau_j)$ is asymptotically minimax with both bias and variance decreasing at the rate of $1/n$ (Basharin (1959), Paninski (2003)). In addition, one can easily provide a sub-Gaussian type large deviation bound using the bounded difference inequality. To further correct the finite sample bias, in this paper we adopt the following Miller–Madow bias-corrected estimator (Miller (1955)) based on MLE, hereafter,

$$\widehat{\mathsf{H}}(\tau_j) = \widehat{\mathsf{H}}_{\mathrm{MLE}}(\tau_j) + \frac{|\mathcal{X}_{\tau_j}| - 1}{2n} = -\sum_{\mathbf{c}_{\tau_j}} \widehat{f}_{\tau_j}(\mathbf{c}_{\tau_j}) \log\big(\widehat{f}_{\tau_j}(\mathbf{c}_{\tau_j})\big) + \frac{|\mathcal{X}_{\tau_j}| - 1}{2n}.$$

By plugging each bias-corrected entropy estimator into (2.3), we, therefore, obtain the estimate of each DII, $\widehat{\Delta}_{\nu_k,h_j} = \sum_{\{\tau_j\}} (-1)^{|\tau_j|+1} \widehat{\mathsf{H}}(\tau_j)$. As a by-product, the collective dependence among k variables specified by ν_k can be easily estimated by $\widehat{\Delta}_{\nu_k} = (-1)^k \prod_{j=1}^{k} \widehat{\Delta}_{\nu_k,h_j}$.

REMARK 2.1. It is worthwhile to mention that, when $k = 3$, the population DII is always nonpositive. This is due to the following fact: without loss of generality, assuming $\nu = \{X_1, X_2, X_3\}$, we have $\Delta_{\nu,3} = \mathsf{H}(X_3) - \mathsf{H}(X_1, X_3) - \mathsf{H}(X_2, X_3) + \mathsf{H}(X_1, X_2, X_3) = -\mathsf{I}(X_1, X_2|X_3)$, where $\mathsf{I}(X_1, X_2|X_3)$ is the conditional mutual information of X_1, X_2 given X_3. For a general $k > 3$, the corresponding DIIs can be either positive or negative. While the targeted true DII is nonpositive, its Miller–Madow estimator due to the bias-correction can be positive. We demonstrate the histograms of DII estimators in the Supplementary Material (Jernigan et al. (2020)).

3. Large scale inference on the collective dependence. As discussed above, vanishing Δ_{ν_k} implies the lack of higher-order dependence. The goal of this paper is to construct a multiple testing procedure for testing the collective dependence of all k-groups. Hereafter, without loss of generality, $k \geq 3$ is prespecified; we index all the possible ν_k by ℓ with $\ell = 1, \ldots, p_k := \binom{d}{k}$, and we will substitute all indices involving ν_k by ℓ. Besides, we still use generic $h_1, \ldots, h_k$ to denote the indices of variables in the ℓth k-group.

Denote the set of true null hypotheses by $\mathcal{H}_0 = \{\ell : 1 \leq \ell \leq p_k, \Delta_\ell = 0\}$. However, for each $1 \leq \ell \leq p_k$, the null hypothesis $H_0^\ell : \Delta_\ell = 0$ is composite. Indeed,

$$\Delta_\ell = 0 \quad \Longleftrightarrow \quad \text{at least one } \Delta_{\ell,h_j} = 0, \quad 1 \leq j \leq k. \tag{3.1}$$

In what follows, we first briefly discuss a natural and simple testing procedure for each H_0^ℓ. To control the type one error for the composite null, this procedure might be conservative. To boost the power for testing each H_0^ℓ, we then formally introduce a novel classification-assisted multiple testing procedure CALL-DECODE. First, we classify each null into one of the $(k+1)$ categories $\mathcal{H}_{0,j}$, $j = 0, \ldots, k$ via a thresholding procedure. The test statistic for each $\mathcal{H}_{0,j}$ distinguishes from each other. Next, we calculate the p-value for each of the p_k null hypotheses according to its category and apply a novel FDR-controlling procedure to test H_0^ℓ simultaneously for all $1 \leq \ell \leq p_k$.

3.1. *A simple testing procedure.* For each fixed $1 \le \ell \le p_k$, consider testing $\Delta_\ell = 0$. One simple idea is to test each sub null hypothesis $H_0^{\ell,j} : \Delta_{\ell,h_j} = 0$ for all $1 \le j \le k$. If there exists at least one j such that $H_0^{\ell,j}$ is not rejected, we fail to reject the null $H_0^\ell : \Delta_\ell = 0$.

To illustrate, we assume that, after some standardization, each scaled estimator of DII $\breve{\Delta}_{\ell,h_j} = \widehat{\Delta}_{\ell,h_j}(\widehat{\sigma}^*_{\ell,jj})^{-1/2}$ follows a standard normal distribution asymptotically under the null $H_0^{\ell,j}$, where $\widehat{\sigma}^*_{\ell,jj}$ is some accurate variance estimator for $\widehat{\Delta}_{\ell,h_j}$. Therefore, it is natural to use $\breve{\Delta}_{\ell,h_j}$ as the test statistic for each $H_0^{\ell,j}$ such that $H_0^{\ell,j}$ is rejected if $|\breve{\Delta}_{\ell,h_j}| > t$ for some critical value $t > 0$. Notice that our goal is to test H_0^ℓ rather than each $H_0^{\ell,j}$. To control the type one error with respect to a prespecified level $\alpha > 0$ for H_0^ℓ, one need to pick t carefully for each $H_0^{\ell,j}$. One can always pick a conservative $t = \Phi^{-1}(\alpha/(2k))$ based on the Bonferroni correction. Nevertheless, this does not reflect the relationship highlighted in (3.1). Indeed, one would not reject H_0^ℓ if and only if at least one of the k subhypotheses fails to be rejected. One extreme under the null H_0^ℓ is that only one sub null is true while all others are false. Without loss of generality, we assume that $\Delta_{\ell,h_1} = 0$ and $\Delta_{\ell,h_j} \neq 0$ for $j \ge 2$. Therefore, asymptotically, we have to pick $t = \Phi^{-1}(\alpha/2)$, since the probability of rejecting any $H_0^{\ell,j}$ is tiny for $j \ge 2$ as n diverges. One can easily see that this choice of $t = \Phi^{-1}(\alpha/2)$ enables us to control the type one error when the number of true sub nulls are greater than one as well. In other words, a simple testing procedure is to reject H_0^ℓ, whenever $\min_{j=1,\ldots,k} |\breve{\Delta}_{\ell,h_j}| > \Phi^{-1}(\alpha/2)$.

While the above simple testing procedure for each H_0^ℓ is expected to be asymptotically valid, it may potentially lack of power. In fact, if the number of zero coordinates (sub nulls) in the vector $(\Delta_{\ell,h_1}, \ldots, \Delta_{\ell,h_k})$ is larger than one, then the choice of t for $\min_{j=1,\ldots,k} |\breve{\Delta}_{\ell,h_j}|$ is conservative, that is, the actual type one error is much smaller than α. To boost the power, we propose a classification-assisted method. The underlying idea is that if we could obtain the information that how many sub nulls are potentially true, then we are able to calibrate the distribution of our test statistic better. To be specific, each null H_0^ℓ can be categorized into k different cases. Set the number of zero coordinates in the vector $(\Delta_{\ell,h_1}, \ldots, \Delta_{\ell,h_k})$ by z_ℓ for $1 \le \ell \le p_k$. Then, define $\mathcal{H}_{0,j} = \{\ell : 1 \le \ell \le p_k, z_\ell = j\}$ as the collection of those k-groups with j zero-valued DII. Clearly, $\mathcal{H}_{0,j} \cap \mathcal{H}_{0,i} = \varnothing$ whenever $i \neq j$. Besides, $\mathcal{H}_0 = \bigcup_{j=1}^k \mathcal{H}_{0,j}$. For notational simplicity, we set alternative hypotheses as $\mathcal{H}_1 = \{\ell : 1 \le \ell \le p_k, \Delta_\ell \neq 0\} = \mathcal{H}_{0,0}$. We formally introduce our procedure in two steps below, respectively, leaving a specific algorithm illustration for the third order collection dependence to conclude Section 3.

3.2. *Classification of differential interaction information.* In this step we apply a simple thresholding procedure to identify Δ_{ℓ,h_j}'s with strong signals. The purpose is to provide a classification rule so that we can classify each k-group into one of the $(k+1)$ cases $\mathcal{H}_{0,j}$, $j = 0, 1, \ldots, k$. That is, we can obtain the number of potentially true sub nulls for each H_0^ℓ, $1 \le \ell \le p_k$.

To this end, we apply a bootstrap procedure to facilitate our classification. Let $\{\mathbf{X}_i^*, 1 \le i \le n\}$ be samples drawn randomly with replacement from the original data $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, where $\mathbf{X}_i^* = (X_{i,1}^*, \ldots, X_{i,d}^*)$ for $1 \le i \le n$. Compute $(\widehat{\Delta}^*_{\ell,h_1}, \ldots, \widehat{\Delta}^*_{\ell,h_k})$ in a similar way, as discussed before, based on $\{\mathbf{X}_i^*, 1 \le i \le n\}$, that is, $\widehat{\Delta}^*_{\ell,h_j} = \sum_{\tau_j^*} (-1)^{|\tau_j^*|+1} \widehat{\mathsf{H}}(\tau_j^*)$, where the sum is taken over all subset τ_j^* of the ℓth k-group that contains X_{h_j}. For some given large positive integer B, we replicate the above procedure B times independently and obtain $\widehat{\Delta}^{*,1}_{\ell,h_j}, \ldots, \widehat{\Delta}^{*,B}_{\ell,h_j}$. Define the sample covariance matrix of $(\widehat{\Delta}^*_{\ell,h_1}, \ldots, \widehat{\Delta}^*_{\ell,h_k})$ for each $1 \le \ell \le p_k$ as $\widehat{\mathbf{\Sigma}}^*_\ell = (\widehat{\sigma}^*_{\ell,ij})_{k\times k}$, where $\widehat{\sigma}^*_{\ell,ij} = B^{-1} \sum_{t=1}^B (\widehat{\Delta}^{*,t}_{\ell,h_j} - \widehat{\Delta}_{\ell,h_j})(\widehat{\Delta}^{*,t}_{\ell,h_i} - \widehat{\Delta}_{\ell,h_i})$.

Finally, we define the scaled DII based on this bootstrap procedure as $\breve{\Delta}_{\ell,h_j} = \widehat{\Delta}_{\ell,h_j}(\widehat{\sigma}^*_{\ell,jj})^{-1/2}$. Then, we apply the thresholding procedure as

$$\breve{\Delta}^{th}_{\ell,h_j} = \mathbb{I}\{|\breve{\Delta}_{\ell,h_j}| > \sqrt{2\log(kp_k)}\}.$$

We set the threshold $\sqrt{2\log(kp_k)}$ as under the null, the distribution of each $\breve{\Delta}_{\ell,h_j}$, $1 \le \ell \le p_k$, $1 \le j \le k$ is expected to be close to a standard normal such that the largest magnitude of those kp_k statistics is no larger than $\sqrt{2\log(kp_k)}$ with probability approaching to 1. After taking this thresholding procedure, those DIIs with strong signals are able to survive. In the end we can define the number of zero coordinates of $(\breve{\Delta}^{th}_{\ell,h_1}, \ldots, \breve{\Delta}^{th}_{\ell,h_k})$ by $\breve{z}_\ell$ for $1 \le \ell \le p_k$ and set our classification rule as

$$\breve{\mathcal{H}}_{0,j} = \{\ell : 1 \le \ell \le p_k, \breve{z}_\ell = j\} \quad \text{for } j = 0, 1, \ldots, k.$$

3.3. *Classification-assisted large scale inference to detect collective dependence.* We are ready to introduce our new multiple testing procedure to detect the significant collective dependence, namely, the CALL-DECODE. For each of the k possible categories under the null $\breve{\mathcal{H}}_{0,j}$, $j = 1, \ldots, k$, we consider a different test statistic.

For each $j = 1, \ldots, k$, we consider those $\ell \in \breve{\mathcal{H}}_{0,j}$. Without loss of generality, we assume that the first j coordinates of $(\breve{\Delta}^{th}_{\ell,h_1}, \ldots, \breve{\Delta}^{th}_{\ell,h_k})$ are zeros. Then, on the event that our classification is accurate, the specific null H_0^ℓ for this k-group becomes that $\Delta_{\ell,h_1} = \Delta_{\ell,h_2} = \cdots = \Delta_{\ell,h_j} = 0$. The validity of our classification result can be justified under certain minimum signal strength condition on each nonzero DII. To test this simplified null, we further scaled these first j coordinates such that they are asymptotically normal with identity covariance, that is, $(\bar{\Delta}_{\ell,h_1}, \ldots, \bar{\Delta}_{\ell,h_j}) = (\widehat{\Delta}_{\ell,h_1}, \ldots, \widehat{\Delta}_{\ell,h_j})(\widehat{\mathbf{\Sigma}}^*_{\ell,[j][j]})^{-1/2}$, where $\widehat{\mathbf{\Sigma}}^*_{\ell,[j][j]}$ is the upper-left j by j submatrix of $\widehat{\mathbf{\Sigma}}^*_\ell$. Motivated by the alternative hypothesis $\ell \in \mathcal{H}_1 = \mathcal{H}_{0,0}$ in which all coordinates of $(\Delta_{\ell,h_1}, \ldots, \Delta_{\ell,h_k})$ are nonzero, we use the smallest magnitude among $(\bar{\Delta}_{\ell,h_1}, \ldots, \bar{\Delta}_{\ell,h_j})$, that is,

$$T_\ell = \min_{t=1,\ldots,j} |\bar{\Delta}_{\ell,h_t}|, \tag{3.2}$$

as the test statistic for this k-group. The cumulative distribution function (c.d.f.) of this variable under the specific null mentioned above can be written asymptotically as $F_j(x) = 1 - [2\{1 - \Phi(x)\}]^j$. We further define the βth quantile of $F_j(x)$ as $F_j^{-1}(\beta) = \inf\{x : F_j(x) \ge \beta\}$. We apply the above procedure for each $\ell \in \bigcup_{j=1}^k \breve{\mathcal{H}}_{0,j}$. Clearly, there are k different $F_j(x)$ corresponding to the k categories.

We now develop the multiple testing procedure motivated by the protocol pioneered by Cai and Liu (2016) for all Δ_ℓ's,

$$H_0^\ell : \Delta_\ell = 0 \quad \text{vs.} \quad H_1^\ell : \Delta_\ell \ne 0 \quad (1 \le \ell \le p_k).$$

Given the category of each k-group with the aid of classification rule proposed in last subsection, we are able to calculate the p-value of each test statistic T_ℓ, $1 \le \ell \le p_k$. It is worthwhile to point out that as a byproduct of the second step, we have already identified $\breve{\mathcal{H}}_{0,0}$ as k-groups with significant collective dependence. Let β be the p-value threshold level such the null hypotheses Δ_ℓ is rejected whenever $T_\ell \ge F_j^{-1}(\beta)$. Then, the FDP of the procedure is defined as

$$\frac{\sum_{\ell \in \mathcal{H}_0} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}}{\max(|\breve{\mathcal{H}}_{0,0}| + \sum_{j=1}^k \sum_{\ell \in \breve{\mathcal{H}}_{0,j}} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}, 1)},$$

where $|\check{\mathcal{H}}_{0,0}|$ is the cardinality of the set $\check{\mathcal{H}}_{0,0}$. An oracle threshold level for controlling the false discovery proportion at a prespecified level $0 < \alpha < 1$ is then

$$\beta^{\text{ore}} = \inf\left\{0 \le \beta \le 1 : \frac{\sum_{\ell \in \mathcal{H}_0} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}}{\max(|\check{\mathcal{H}}_{0,0}| + \sum_{j=1}^{k} \sum_{\ell \in \check{\mathcal{H}}_{0,j}} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}, 1)} \le \alpha\right\}.$$

However, this oracle threshold β^{ore} is unknown and need to be estimated, as it depends on the knowledge of the set of null hypotheses $\mathcal{H}_0$. Motivated from Cai and Liu (2016), with the assumption that majority null hypotheses are true (i.e., $|\mathcal{H}_0|/(p_k - |\check{\mathcal{H}}_{0,0}|) \to 1$), we estimate the numerator, the number of the true nulls, which are falsely rejected by the procedure at the threshold level β, as $(p_k - |\check{\mathcal{H}}_{0,0}|)(1-\beta)$, which leads to the final FDR procedure,

$$\tilde{\beta} = \inf\left\{0 \le \beta \le C_\beta : \frac{(p_k - |\check{\mathcal{H}}_{0,0}|)(1-\beta)}{\max(|\check{\mathcal{H}}_{0,0}| + \sum_{j=1}^{k} \sum_{\ell \in \check{\mathcal{H}}_{0,j}} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}, 1)} \le \alpha\right\}. \tag{3.3}$$

Whenever $\tilde{\beta}$ does not exist, as suggested by Cai and Liu (2016), we simply let $\tilde{\beta} = C_\beta := 1 - (p_k - |\check{\mathcal{H}}_{0,0}|)^{-1}$. We reject H_0^ℓ, whenever either $\ell \in \check{\mathcal{H}}_{0,0}$ or $T_\ell \ge F_j^{-1}(\tilde{\beta})$, where $\ell \in \check{\mathcal{H}}_{0,j}$, $1 \le j \le k$. We emphasize that the above procedure is based on an approximation of $\sum_{\ell \in \mathcal{H}_0} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}/|\mathcal{H}_0|$ by $(1-\beta)$. It is important to impose the constraint $0 \le \beta \le C_\beta$ in (3.3) since this approximation would not be accurate when β is very close to 1. In addition, the dependence among all T_ℓ would not influence this approximation too much when $|\mathcal{H}_0|$ is reasonably large.

We summarize the proposed inference procedure for third order collection dependence as a demonstration below before concluding this section with a few important remarks.

Algorithm: CALL-DECODE for Significant Triplets

Input. Data matrix with n rows and d columns, number of bootstrap iterations B, nominal FDR α, p^{thr} and ϵ if screening.

Step 0. Initial screening as discussed in Section A.2.2.

Step 1. Classification. For each triplet $\ell = 1, \cdots, p_3 := \binom{d}{3}$, do

Step 1.1. estimate differential interaction information $(\widehat{\Delta}_{\ell,h_1}, \widehat{\Delta}_{\ell,h_2}, \widehat{\Delta}_{\ell,h_3})$, where $\widehat{\Delta}_{\ell,h_j} = \sum_{\{\tau_j\}} (-1)^{|\tau_j|+1} \widehat{\mathsf{H}}(\tau_j)$, as defined in Section 2;

Step 1.2. estimate the variance of $\widehat{\Delta}_{\ell,h_j}$, $\widehat{\sigma}^*_{\ell,jj}$ using B bootstrap iterations and obtain scaled differential interaction information $\check{\Delta}_{\ell,h_j} = \widehat{\Delta}_{\ell,h_j}(\widehat{\sigma}^*_{\ell,jj})^{-1/2}$;

Step 1.3. threshold differential interaction information $(\check{\Delta}^{th}_{\ell,h_1}, \check{\Delta}^{th}_{\ell,h_2}, \check{\Delta}^{th}_{\ell,h_3})$ with $\check{\Delta}^{th}_{\ell,h_j} = \mathbb{I}\{|\check{\Delta}_{\ell,h_j}| > \sqrt{2\log(3p_3)}\}$;

Step 1.4. set $\check{\mathcal{H}}_{0,j} = \{\ell : 1 \le \ell \le p_3, \check{z}_\ell = j\}$ for $j = 0, 1, 2, 3$ as the classification rule, where $\check{z}_\ell = \sum_{k=1}^{3} \mathbb{I}(\check{\Delta}^{th}_{\ell,h_k} = 0)$.

Step 2. Compute test statistic for each triplet.

Step 2.1. transform differential interaction information to $(\bar{\Delta}_{\ell,h_1}, \ldots, \bar{\Delta}_{\ell,h_j})$ using estimated covariance from B bootstrap iterations;

Step 2.2. calculate test statistic $T_\ell = \min_{t=1,\ldots,j} |\bar{\Delta}_{\ell,h_t}|$ for the ℓth triplet. The distribution under null $\Delta_\ell = 0$ relies on classification: $F_j(x) = 1 - [2\{1 - \Phi(x)\}]^j$ for class j.

Step 3. Identification. For level α, we compute $\tilde{\beta}$ using (3.3) and reject $\Delta_\ell = 0$ for the ℓth triplet whenever either $\ell \in \check{\mathcal{H}}_{0,0}$ or $T_\ell \ge F_j^{-1}(\tilde{\beta})$.

Output. 3-dimensional index vectors for identified triplets.

REMARK 3.1. Though motivated from Cai and Liu (2016), our procedure has unique novelty due to the classification step. One major difference from Cai and Liu (2016) is that we take into account those k-groups $\breve{\mathcal{H}}_{0,0}$ with strong collective dependence. As a result, our finite sample approximation of the numerator potentially is more accurate. After all, intuitively, we only need $|\mathcal{H}_0|/(p_k - |\breve{\mathcal{H}}_{0,0}|) \to 1$, while routine application of Cai and Liu (2016) requires $|\mathcal{H}_0|/p_k \to 1$.

REMARK 3.2. After the classification step, our test statistic T_ℓ in (3.2) is based on the smallest magnitude. Alternatively, one can also consider chi-square type approach for the simplified null $\Delta_{\ell,h_1} = \Delta_{\ell,h_2} = \cdots = \Delta_{\ell,h_j} = 0$. However, if only some of these quantities are zero while the null H_0^ℓ is still true, the chi-square based test is more likely to reject the null while T_ℓ is relatively insensitive to this pathological case. In practice, to reduce the effect of rare cases that some of the classification were not accurate in the first step, we advocate using T_ℓ over the chi-square type statistic. After all, it is more intuitive to use the weakest signal among all coordinates of $(\Delta_{\ell,h_1}, \ldots, \Delta_{\ell,h_k})$ if the goal is to test whether at least one of them is zero.

REMARK 3.3. In practice, one can always apply a variable screening step before CALL-DECODE to remove those (nearly) deterministic variables and reduce the computational burden. This screening procedure is motivated by the fact that in many applications, including the protein coevolution study, some variables demonstrate little variation and hence are nearly deterministic. Intuitively, those variables are not important in interpreting the dependence relationship among a group of variables. Indeed, this is the case in our setting: one can easily show that there is no (very weak) collective dependence among any k-group including at least one such a (nearly) deterministic variable; see Section A.2 of the Supplementary Material (Jernigan et al. (2020)) for further details on this claim, including a detailed description of the proposed screening procedure with theoretical justification in Section A.2.2.

4. Monte Carlo evidence. In this section we report results from Monte Carol experiments to evaluate the finite-sample performance of our method on detecting the significant higher-order collective dependence of d-dimensional random vectors. For demonstrations we focus on $k = 3$ to identify dependent triplet.

We generate multivariate discrete or categorical data from two settings, the truncated continuous distribution (TCD) and the multinomial distribution (MD). The empirical FDR and power are computed based on 100 replications. In simulations we let the nominal FDR equal to 0.2, $d = 20$ or 26, and $B = 1000$ for the classification step in our method. Notice that, for the third order collective dependence, $d = 20$ or 26 variables have 1140 or 2600 candidate triplets and indeed lead to a large scale inference problem. On the other hand, for the motivating protein coevolution problem many important protein families do have a small number of positions on their sequences, such as the zinc knuckle family with only $d = 16$ residues.

4.1. *TCD setting*. For the TCD setting, first generate random vectors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})$, $i = 1, \ldots, n$, with i.i.d. coordinates $Z_{ib} \sim F$, where F is either the standard normal (C1) or the standardized exponential distribution with mean 5 (C2). Using covariance matrix $\mathbf{\Sigma}$, $\mathbf{Z}_i$'s are further transformed as $\mathbf{Z}_i \mathbf{\Sigma}^{1/2}$ to possess nondegenerate higher-order dependence. We consider two block diagonal settings for $\mathbf{\Sigma}$ that the first design has one large block of size 15 and *five* remaining nonzero diagonals (S1) and the second has three blocks of size 6 and *two* remaining nonzero diagonals (S2). Settings S1 and S2 have 455 and 60 triplets with nontrivial collective dependence, respectively. For both settings we let the diagonal entries belonging to blocks equal to 0.5, the nonzero off-diagonal entries equal to 0.2

TABLE 1

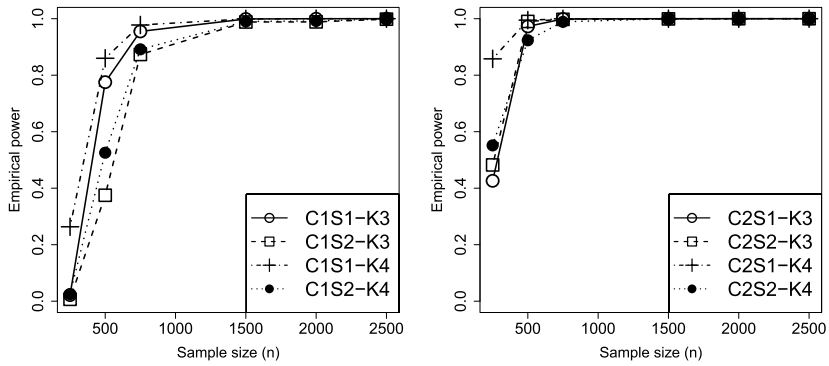
The empirical FDRs for detecting triplets with significant third order collective dependence by the proposed algorithm under the TCD setting for models C1 (standard normal) and C2 (standardized exponential), covariances S1 and S2, and T_K with $K = 3, 4$. The nominal FDR is 0.2, and results are based on 100 repetitions

Model	K/n	250	500	750	1500	2000	2500
C1S1	3	0.068	0.058	0.064	0.025	0.012	0.013
	4	0.223	0.125	0.106	0.114	0.117	0.135
C1S2	3	0.197	0.092	0.082	0.030	0.026	0.022
	4	0.225	0.198	0.149	0.122	0.124	0.106
C2S1	3	0.084	0.074	0.076	0.013	0.014	0.022
	4	0.203	0.167	0.104	0.071	0.050	0.045
C2S2	3	0.151	0.052	0.050	0.023	0.018	0.023
	4	0.188	0.204	0.201	0.181	0.136	0.181

and the remaining diagonals equal to 1.3. From $\mathbf{Z}_i\boldsymbol{\Sigma}^{1/2}$, we obtain data $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ with $X_{ib} = T_K([\mathbf{Z}_i\boldsymbol{\Sigma}^{1/2}]_b)$ for $1 \le b \le d$, where $T_K(x) = \sum_{\iota=0}^{K-1} \iota \mathbb{I}(x \in I_\iota^K)$ is a truncation operator to generate K levels from x. For $K = 3$, we let $I_0^3 = (-\infty, -1]$, $I_1^3 = (-1, 1]$, $I_2^3 = (1, \infty)$ for model C1 and $I_0^3 = (0, 1]$, $I_1^3 = (1, 3]$, $I_2^3 = (3, \infty)$ for model C2. Similarly, when $K = 4$, $I_0^4 = (-\infty, -1]$, $I_1^4 = (-1, 1]$, $I_2^4 = (1, 2]$, $I_3^4 = (2, \infty)$ for model C1 and $I_0^4 = (0, 1]$, $I_1^4 = (1, 2]$, $I_2^4 = (2, 3]$, $I_3^4 = (3, \infty)$ for model C2. In summary, for the TCD setting we have eight different scenarios based on the latent distribution F, the covariance $\boldsymbol{\Sigma}$ and the truncation operator. Though the rigorous characterization of sparsity for collective dependence is nontrivial under the general setting, it is easy to see that the third order collective dependence is relatively dense for setting S1 given $d = 20$. For the TCD setting we let $n = 250, 500, 750, 1500, 2000$ and 2500.

From Table 1 the empirical FDR is under control at the nominal level as n increases for in general. Even for $n = 250$, the empirical FDR is reasonably controlled for $K = 3$ under all four settings while it is slightly inflated when $K = 4$. The slight inflation of empirical FDR for $K = 4$ and $n = 250$ is more or less due to the simultaneous estimation of entropy for a large number of discrete distributions with many unique values. The larger block size in $\boldsymbol{\Sigma}$ (setting S1) leads to more dependent triplets. The seemingly conservative FDR control for large n in Table 1 can be interpreted as a result due to the dependence across triplets and the corresponding hypotheses which usually influences the FDR control in practice (Cai and Liu (2016)). The empirical power, as displayed in Figure 1, approaches toward one as n increases and provides numerical evidence for our procedure's consistency. Our method is more powerful for the setting of covariance with larger blocks. Settings with $K = 4$ are more powerful than those with $K = 3$, which is not surprising as more information will be available, given a larger K (Cover and Thomas (2006)). The difference in power between settings of the standard normal (C1) and the standardized exponential (C2) is mainly caused by the difference in $\mathbb{P}([\mathbf{Z}_i\boldsymbol{\Sigma}^{1/2}]_b \in I_0^3)$ between $Z_{i1} \sim N(0, 1)$ and $Z_{i1} \sim \exp(5)/5$ which leads to different entropy.

4.2. *MD setting*. Under the MD setting we generate data $\mathbf{X}_i$'s using multinomial distribution Multi$(M, \mathbf{r})$ with M trials and probability vector $\mathbf{r} = (r_1, \ldots, r_Q)$, where $\sum_{q=1}^{Q} r_q = 1$. Specifically, we consider two designs. For design I (D1), $\mathbf{X}_i = (\mathbf{X}_{i,(1)}, \mathbf{X}_{i,(2)})$, where the 9-dimensional vectors $\mathbf{X}_{i,(1)} \sim$ Multi$(M, \mathbf{r})$ and independent from $\mathbf{X}_{i,(1)}$, the 11-dimensional vector $\mathbf{X}_{i,(2)}$ has independent components uniformly sampled from $\{0, 1, 2, 3, 4\}$. Design II (D2) is similar to D1 except that $\mathbf{X}_{i,(1)}$ is a 15-dimensional multinomial random vector. That



(a) Model C1 ($Z_{ij} \sim N(0,1)$) (b) Model C2 ($Z_{ij} \sim \exp(5)/5$)

FIG. 1. *The empirical powers for detecting triplets with significant third order collective dependence by the proposed method under the TCD setting for models C*1 (*standard normal*) *and C*2 (*standardized exponential*), *covariances S*1 *and S*2 *and* T_K *with* $K = 3, 4$. *The nominal FDR is* 0.2, *and results are based on* 100 *repetitions*.

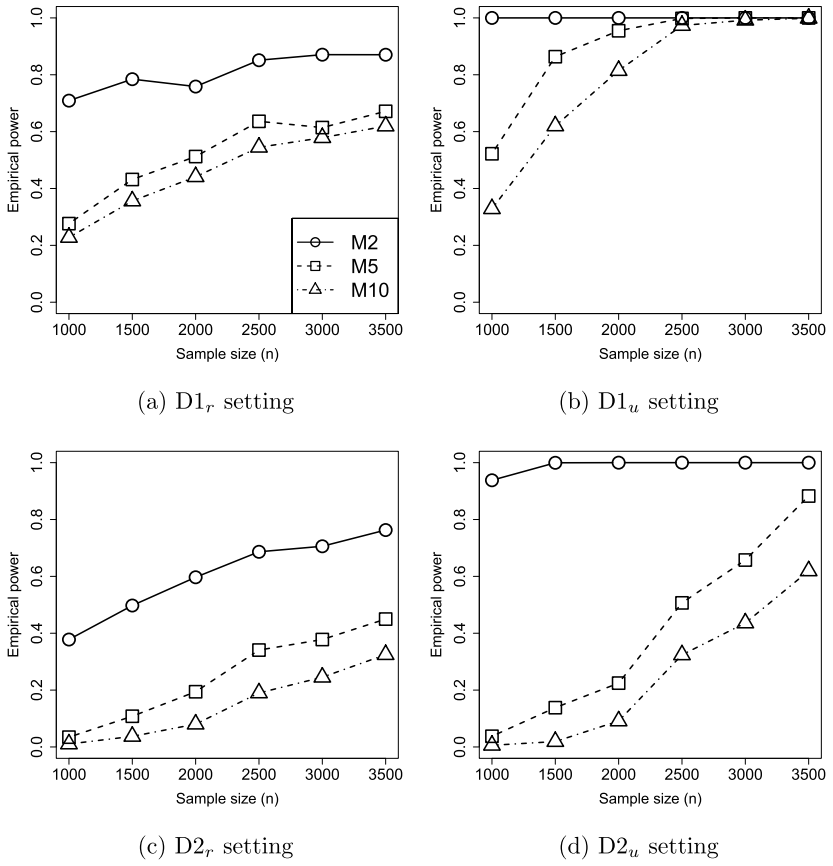
is, $d = 20$ for D1 and $d = 26$ for D2. Designs I and II have 84 and 455 triplets with nontrivial collective dependence, respectively. For all settings, $M \in \{2, 5, 10\}$. The Q-dimensional $\mathbf{r}$ is either fixed with a common value Q^{-1} for all components or random whose entries are determined via dividing Q i.i.d. $U(0, 1)$ random numbers by their sum. In summary, for the MD setting we have four different cases based on the design and the randomness of $\mathbf{r}$. Compared to the TCD setting, the number of dependent triplets for the MD setting is similar. Here, we let $n = 1000, 1500, 2000, 2500, 3000$, and 3500.

In Table 2, the empirical FDRs are satisfactorily controlled for the MD setting. Similar to the TCD setting, the seemingly conservative empirical FDR can be interpreted by the strong dependence among hypotheses associated to triplets. As n increases, numerical evidences in Figure 2 show that our procedure remains consistent. In comparison to the TCD setting, the multinomial distribution has negative covariance between components whose magnitude increases in M. Therefore, larger M will incur stronger dependence among test statistics T_ℓ's and the corresponding hypotheses which compromises the power. For $M = 2$, the data is

TABLE 2

The empirical FDRs for detecting triplets with significant third order collective dependence by the proposed method under the MD setting for Designs I and II with random and fixed $\mathbf{r}$ ($D1_r$ *and* $D1_u$; $D2_r$ *and* $D2_u$, *respectively*), *where* $M = 2, 5, 10$. *The nominal FDR is* 0.2, *and results are based on* 100 *repetitions*

Model	M/n	1000	1500	2000	2500	3000	3500
$D1_r$	2	0.020	0.016	0.013	0.012	0.013	0.010
	5	0.019	0.010	0.015	0.013	0.010	0.014
	10	0.024	0.029	0.026	0.021	0.022	0.028
$D1_u$	2	0.045	0.051	0.022	0.000	0.001	0.000
	5	0.014	0.021	0.031	0.038	0.031	0.020
	10	0.028	0.033	0.024	0.036	0.046	0.038
$D2_r$	2	0.036	0.044	0.043	0.045	0.039	0.038
	5	0.008	0.015	0.018	0.022	0.021	0.025
	10	0.007	0.020	0.016	0.020	0.022	0.025
$D2_u$	2	0.057	0.071	0.072	0.067	0.066	0.066
	5	0.013	0.016	0.016	0.027	0.026	0.031
	10	0.022	0.004	0.022	0.024	0.028	0.029



(a) $D1_r$ setting (b) $D1_u$ setting

(c) $D2_r$ setting (d) $D2_u$ setting

FIG. 2. *The empirical powers for detecting triplets with significant third order collective dependence by the proposed method under the MD setting for Designs I and II with random and fixed* $\mathbf{r}$ ($D1_r$ *and* $D1_u$; $D2_r$ *and* $D2_u$, *resp.*), *where* $M = 2, 5, 10$. *The nominal FDR is* 0.2, *and results are based on* 100 *repetitions*.

sparse by structure, and the estimation of entropy is fast and accurate which leads to substantially higher power compared to other scenarios. With more variability it is not surprising that the random $\mathbf{r}$ leads to a loss in power. Finally, it is interesting that D1 provides a more powerful result than D2 for both random and uniform probability vectors $\mathbf{r}$. Though the number of dependent triplets under D2 is more than that under D1, the total number of candidate triplets under D2, which is 2600, is more than double that under D1, which is 1140. That is, the number of potentially dependent hypotheses to be multiple tested in D2 is much larger than that in D1. In addition, a larger d leads to more marginal entropy to be estimated and may, therefore, undermine the overall accuracy to estimate the collective dependence, particularly when M is large and n is small. Hence, these factors together compromise the empirical power when the sample size is small.

5. An application to protein coevolution. Conservation of function within a protein family dictates the pattern of sequence variation. For example, in order to maintain a functional interaction, all residues involved usually mutate together. Those jointly dependent residues and corresponding positions on the sequence are the key components to understand protein coevolution. In this section we apply our method to the protein coevolution problem. Particularly, we employ the collective dependence as a measure and apply our detection procedure, CALL-DECODE, to identify functional triplet of amino acids residues from the protein MSA data.

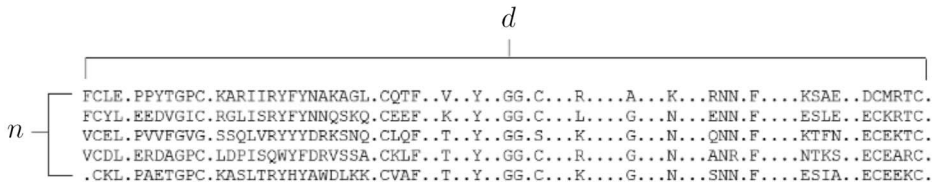


FIG. 3. *A snapshot of the* (*raw*) *MSA data for analysis.*

5.1. *Data description and processing.* As discussed in Section 1, the MSA data is typical for computational approaches to study protein coevolution. Generally speaking, the MSA is the sequence alignment of three or more biological sequences, such as protein or nucleic acid, of same length. During the alignment process the input sequences are considered to have an evolutionary relationship by which they share an evolutionary linkage and a common ancestor. In this study an MSA consists of a group of homologous protein sequences aligned using hidden markov models. The MSA data for each protein family is retrieved from the Pfam database (Finn et al. (2008)), and the corresponding structure is obtained from the Protein Data Bank (PDB, Berman et al. (2000)). Each family describes a protein motif which refers to a structural domain with certain biological functions. As shown in Figure 3, an aligned MSA data is an $n \times d$ matrix with n independent protein sequences from a common family and d positions encoding the type of amino acid. Here, n is the size of a protein family and is usually fairly large for coevolution study. The raw value at each of the d position is from 21 unique alphabets representing the 20 amino acid types plus one gap symbol.

In our analysis, the raw MSA data is further processed as follows. First, we remove the redundant sequences using a similarity threshold of 90% and eliminate the columns that have more than 10% gaps. Then, we remove all the conserved positions with low information entropy. Finally, we categorize all amino acid types and the gap symbol into *six* classes, based on the size and the chemical attributes. That is, the data for analysis contains six unique alphabets, as summarized in Table 3. Extracting coevolution signals of functional triplets of residues is then fulfilled through detecting triplets with significant collective dependence among d discrete random variables taking values on $\{0, 1, 2, 3, 4, 5\}$.

5.2. *Results and findings.* We study two protein families, the elongation factor P family (Pfam ID: PF09285) and the zinc knuckle family (Pfam ID: PF00098) for demonstration purposes. By identifying triplets having significant collective dependences, our method reveals novel amino acid residues that are critical for the protein's function, which are missed by the widely used methods based on mutual information (denoted as MI, Martin et al. (2005)), whose details and additional comparisons are deferred to Section C in the Supplementary Material (Jernigan et al. (2020)).

TABLE 3
Code of the six alphabets in the data for analysis; *see Nelson and Cox* (2005) *for coding of amino acid types*

Alphabet	Amino acid types or gap symbols
0	B,J,O,Z,U,X, '–', '·'
1	A,G
2	C,S,T,N,Q
3	D,R,E,K,H
4	F,Y,W
5	P,V,I,L,M

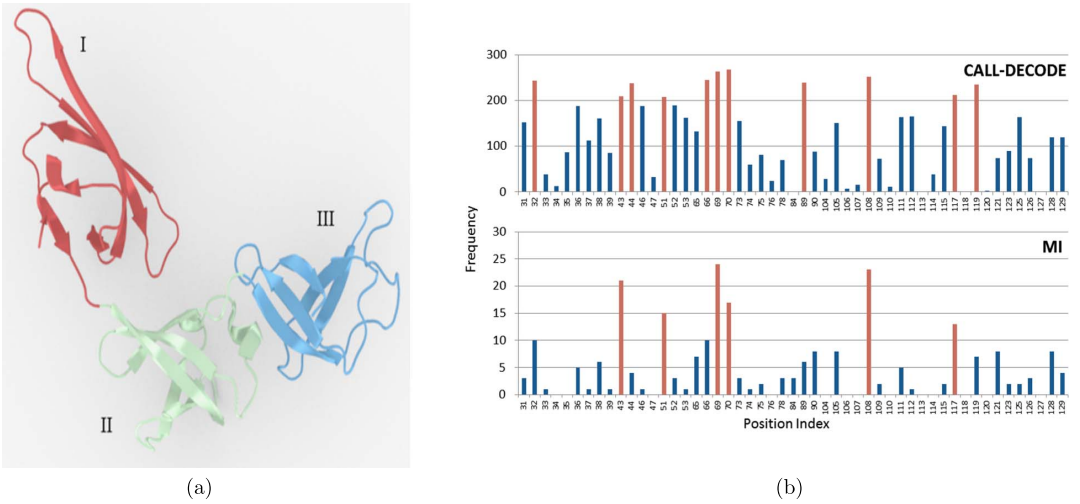


FIG. 4. (*a*) *The structure of EF-P (from Thermus thermophilus). Three domains are colored separately.* (*b*) *The coverage of coevolved positions on the MSA identified as significant by our method and MI. The x-axis is the position index of the MSA, and the y-axis is its appearance frequency based on the corresponding method. A position is included if it has been identified as significant by either method. Hubs are in red.*

5.2.1. *Elongation factor P family.* Elongation factor P (EF-P) is a prokaryotic protein translation factor required for peptide bond synthesis on 70S ribosomes. It stimulates ribosomal peptidyl transferase activity. Although the exact biological mechanism remains unknown, EF-P has been hypothesized to alter the affinity of the ribosome for tRNA and thus increases their reactivity as acceptors for peptidyl transferase. As shown in panel (a) in Figure 4, EF-P consists of three β-barrel domains. Domain I is topologically the same as the N domain of an EF-P homolog of Eukarya and Archaea, eIF-5A. EF-P domains II and III share the same topology as that of the eIF-5A C domain.

In this analysis we focus on the C-terminal domain in EF-P, the domain labeled III in panel (a) in Figure 4. One side of the surface of domain III has a patch of conserved basic residues which are positively charged and favorable for nucleic acid binding (Hanawa-Suetsugu et al. (2004)). Residues on the other side of domain III are mostly negatively charged and conserved. A conserved residue indicates small or no variations at the corresponding position on the MSA. For the coevolution analysis those conserved positions are ignored.

The raw MSA data consists $n = 4886$ sequences. The number of nonconserved positions involved in the analysis is $d = 55$. As a result, 2022 out of 26,235 candidate triplets are identified with significant third order collective dependence by our method with FDR controlled at 0.05. This covers 50 out of the total 55 positions on the MSA. As a comparison, MI is applied to the same data. MI selects 122 out of 1485 candidate pairs which covers 38 positions. Panel (b) in Figure 4 displays the frequencies of each position belonging to a significant triplet or pair based on the corresponding method. From the detection results we also extract the hub positions and corresponding amino acid residues. Hubs are considered to be important for biological functions and structural stability (Buslje et al. (2010), Zhang, Ren and Chen (2018)). Positions on the MSA with the top one third highest appearance frequency are considered to be hubs.

Some positions are commonly identified as hubs by our method and MI, such as positions 43, 51, 69, 70, 108 and 117 on the MSA (Panel (b) in Figure 4), corresponding to residues Gly138, Asp142, Gly147, Ser148, Leu163 and Lys172 (Panel (a) in Figure 5). In contrast, residues Leu130, Val139, Gly146, Ala157 and Asp174 (at positions 32, 44, 66, 89, and 119) are selected uniquely as hubs based on the collective dependence. It is interesting that the

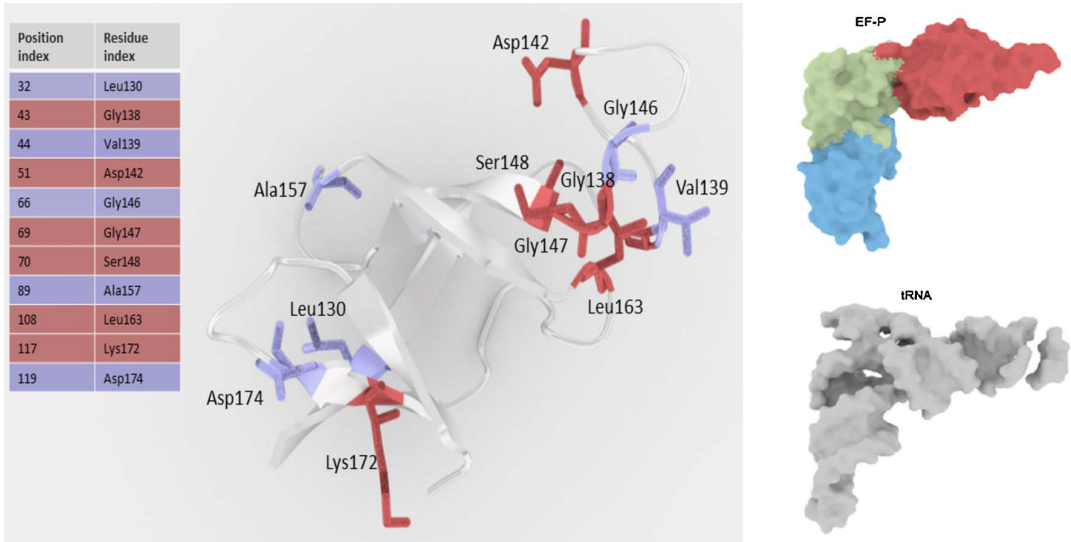


FIG. 5. (*a*) *The hub residues of EF-P C-terminal Domain. Residues in red are commonly identified by both CALL-DECODE and MI. Residues uniquely selected by CALL-DECODE are in purple. The mapping between residue index and position index is displayed in the legend* (*generated from PfamScan*). *The position index refers to the column number in the MSA. The residue index is the unique identifier* (*residue name plus residue number*) *for a residue on the protein structure defined in the PDB file.* (*b*) *The structures of EF-P monomer and tRNA in the surface representation. Domain I, II, III in EF-P are in red, green and blue, respectively.*

hubs loosely fall within two clusters. Within each cluster the spatial proximity of hub residues suggests collective mutations of residues.

The similarlity between the shape of EF-P monomer and the approximaitely 95 degree-L shape of tRNA molecule is well known (Hanawa-Suetsugu et al. (2004), panel (b) in Figure 5). It has been suggested that the similarity of EF-P monomer in shape and size to tRNA may help proteins to pass through the entrance of the ribosome (Vestergaard et al. (2001)). Both the I-II and II-III domain interfaces are formed by hydrophobic side chains with high surface complementarities. The hub residues uniquely discovered by our method, Ala157 and Asp174, are in the loop region connecting domains II and III (Panel (a) in Figure 5). This loop region resembles a joint interface between the two domains. Therefore, this reflects the functional importance of Ala157 and Asp174 for stabilizing the overall structure shape which is involved in RNA (or DNA) binding interactions.

5.2.2. *Zinc knuckle family.* The zinc finger domain is a transcription factor mostly interacting with DNA or RNA. In a zinc finger domain, the zinc ion helps stabilize and hold the domain in a specific shape while all the side chains of functional amino acids reach out to interact with nucleotides. The functional amino acids determine the binding location of protein to DNA or RNA. Zinc finger domains are stable and rarely undergo conformational changes beyond target binding. Based on the overall shape, zinc finger domains can be classified into different types. For example, the zinc knuckle family (ZnkF) is a subtype with a gag-knuckle fold which consists of two β-strands connected by a CCHC-type zinc knuckle followed by a short helix or loop.

In addition to the relatively short sequence, the high variability in nonconserved positions of ZnkF makes it difficult to determine the group of positions with functional or structural specificity using methods that are based on pairwise coherence measures. For example, a traditional approach to identify a group of coevolved positions from pairs is to combine MI-selected pairs by referring to the transitive relationship. However, the high background noise,

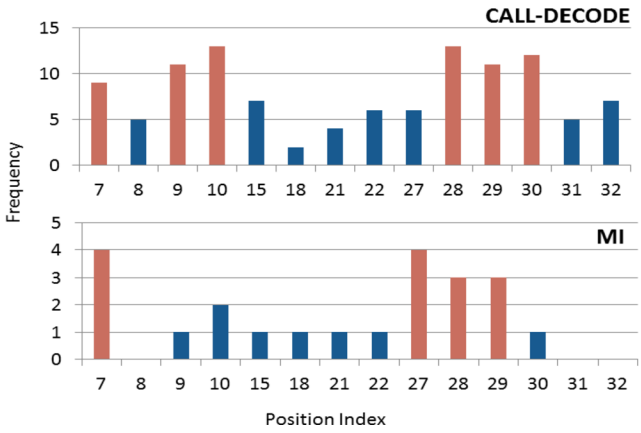


FIG. 6. *The coverage of coevolved positions on the MSA identified as significant by our method and MI. The x-axis is the position index of the MSA, and the y-axis is its appearance frequency based on the corresponding method. A position is included if it has been identified as significant by either method. Hubs are in red.*

which accounts for the intrinsic random variability in data (Dunn, Wahl and Gloor (2008)), makes it unreliable to employ the transitive relationships for identifying higher-order groups of coevolving positions from coevolved pairs.

In the Pfam database, ZnkF includes 808 species with 22,448 sequences and the average length of domain across species is 17.8 amino acids. We focus on a species with $n = 9029$ sequences and $d = 16$ amino acids in this analysis. With nominal FDR controlled at 0.05, out of 560 candidates, our method identifies 37 triplets with significant collective dependence which covers 14 out of the total 16 positions on the MSA. As a comparison, MI selects 11 pairs out of 120 candidates and covers 11 positions. Again, the advantage of collective dependence is demonstrated in the wider coverage of positions than MI. Displayed in Figure 6, MI fails to identify positions 8, 31 and 32 on the MSA and neither does it select hubs Phe16, Asn17 and Asn27 at positions 9, 10 and 30, respectively.

Compared with results from MI, our method identifies dependent triplets and reveals more hub residues critical for protein's function. In fact, hubs Lys14, Phe16, Asn17, Ala25 and Lys26, corresponding to positions 7, 9, 10, 28 and 29 on the MSA in Figure 6, are functional residues interacting with the nucleotides. They are keys to understand the mechanism of HIV-1 (De Guzman et al. (1998)). Panel (a) in Figure 7 displays the protein structure of the HIV-1 nucleocapsid (NC) protein, which posses the zinc knuckle domain, bound to the SL3 Ψ-RNA recognition element (PDB ID is 1a1t). This structure reveals the functions of our identified hubs: Lys14, Phe16, Asn17, Ala25 and Lys26. The 20-nucleotide RNA segment in panel (a) in Figure 7 is part of what is known as Ψ-site, which contains four stem-loop structures, denoted as SL1 through SL4. The 20-nucleotide RNA segment assembles SL3, a highly conserved structure among different strains of HIV-1. In panel (a) in Figure 7, Ala25 interacts with nucleotides A8 and G9 on the 20-nucleotide RNA segment; Lys26 interacts with A8; the backbones of Phe16 and Ala25 for hydrogen bonds with G9; and the Lys14 backbone forms a hydrogen bond with G9. The base of the tetraloop nucleotide, A8, makes hydrophobic contacts with Ala25, Phe16 and Asn17.

Though the function of identified Asn27 at position 30 on the MSA remains uncovered, we notice that the side chains of Phe6, Gln9 and Asn27 point to the same center and are spatially close to each other (Panel (b) in Figure 7). This indicates a noncharged side chain interaction. Chakrabarti and Bhattacharyya (2007) showed that there is a high propensity of Asn-Gln spatial proximity in general due to their stability. In addition, for 1a1t, Phe6 is the amino acid next to Asn27-Gln9 with a high preference for aromatic side chains. In fact, the side chains

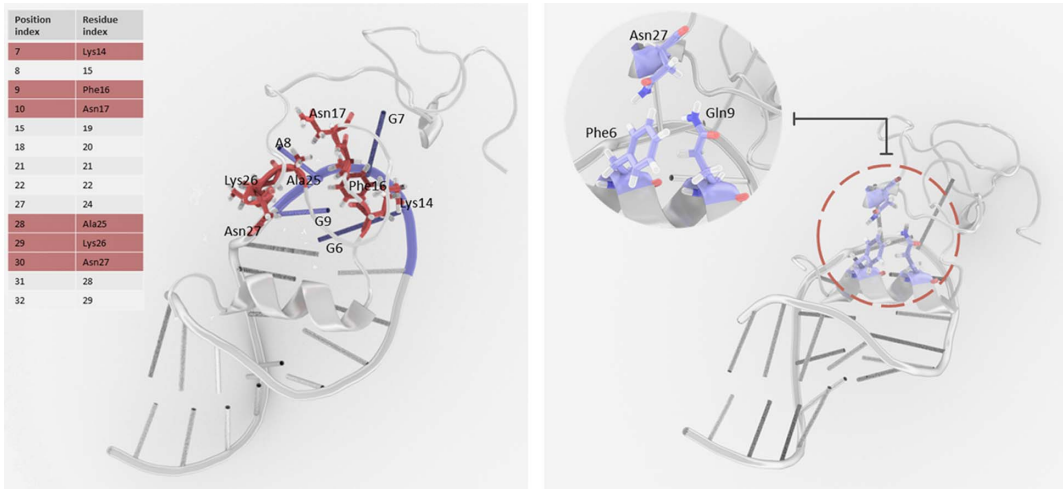


FIG. 7. (*a*) *The interaction between HIV-*1 *NC protein and RNA. Hub residues identified by collective dependence are in red, while the interacting nucleotides are in purple. The mapping between residue index and position index is displayed in the legend* (*generated from PfamScan*). (*b*) *The triplet residue contact formed by Asn*27, *Gln*9 *and Phe*6 *on the HIV-*1 *NC protein* (*in purple*).

of these three amino acids point to the geometric center of this triplet neighborhood. We hypothesize that the triplet residue contacts made by Asn27-Gln9-Ph6 support the structural stability and may further aid the structure positioning to form the protein-RNA interaction.

6. Conclusions with discussions. Measuring the dependence of $k \geq 3$ random variables is fundamentally different from measuring their pairwise dependence. Statistically, bivariate measures, such as mutual information, distance covariance and partial correlation, cannot be directly employed to detect a higher-order dependence. In practice, the data are commonly discrete or categorical such as the protein coevolution study and, therefore, demand measures with fewer parametric assumptions. Efforts on generalizing bivariate measures to $k \geq 3$ variables scattered in literature are often model specific and thus cannot be easily adopted for our purpose. More importantly, the higher-order dependence may encounter certain degeneracy due to the full dependence relationship which should be avoided in higher-order dependence detection. In response to these challenges, we introduce an information theoretic measure, collective dependence (Galas et al. (2014)), to quantify the dependence among k variables. Using a bias-correction estimator on information entropy, we can estimate the collective dependence effectively.

Comparison with other FDR procedures. Identifying k sub-vectors with significant collective dependence from a d-dimensional random vector is intrinsically a large scale multiple testing problem. Furthermore, the symmetry of collective dependence with respect to the permutation of variables leads to composite nulls which compromises the power of naive multiple testing procedures. To that end, we propose a classification-assisted multiple testing procedure, CALL-DECODE, to detect the kth order collective dependence. Our method provides satisfactory powers in practice while controls FDR. We discuss the novelty of CALL-DECODE in the context of multiple testing problems. First, our two-step multiple testing procedure is an extension of the method by Cai and Liu (2016) from the pairwise dependence to the collective dependence with a general order. Similar to Cai and Liu (2016), the major difference from the standard Benjamini–Hochberg (BH) procedure lies in that by im-

posing constraint C_β in (3.3), the dependence among test statistics are expected to be negligible when the nulls are dominant. Second, CALL-DECODE can be viewed as one of the structure-adaptive algorithms which have drawn significant attention recently (Li and Barber (2019)). Without our classification step, one can always follow the simple procedure in Section 3.1 to obtain a conservative p-value for each k-group, and then apply the multiple testing procedures, such as the standard BH method or the method by Cai and Liu (2016), to control the FDR. In particular, for each k-group, one can use $T_\ell = \min_{t=1,\ldots,k} |\breve{\Delta}_{\ell,h_t}|$ with the null c.d.f. $F_1(x) = 1 - [2\{1 - \Phi(x)\}]^1$. In contrast, in a data-adaptive fashion, each p-value is "reweighted" according to an updated c.d.f. $F_j(x) = 1 - [2\{1 - \Phi(x)\}]^j$ whenever the corresponding k-group is classified into the jth category during the classification step of CALL-DECODE, that is, $\ell \in \breve{\mathcal{H}}_{0,j}$. Then, the method by Cai and Liu (2016) is applied to the reweighted p-values to control the FDR in our procedure. More comparisons between our method and other multiple testing procedures that integrate prior structure of hypotheses are detailed in the Supplementary Material (Jernigan et al. (2020)).

Possible failure with classification errors. The minimum signal strength on each nonzero DII to be classified correctly in our CALL-DECODE is at the order of $\sqrt{\log(kp_k)/n}$ because its target is to bound the supnorm of all DII estimators, that is, $\sup_{\ell,j} |\breve{\Delta}_{\ell,h_j}|$. Since this minimum signal strength condition may not be satisfied, one natural question is what would happen if the classification step is not accurate? After all, if all k-groups can be classified perfectly, then it is no bother to conduct multiple testing procures. To this end, we would like to first provide some insights on the potential errors and their consequences on our multiple testing procedure.

Due to the conservative threshold, it is well expected that all DII estimators with a true null would be thresholded to zero in the classification step, that is, $\mathcal{H}_{0,j} \subset \bigcup_{t=j}^{k} \breve{\mathcal{H}}_{0,t}$ for each $0 \le j \le k$. In particular, we expect those k-groups in the category $\mathcal{H}_{0,k}$ would be classified correctly. All potential errors, due to some DII estimators with nonnull, are thresholded to be zero incorrectly and can be categorized into two types. The first one corresponds to those k-groups with nonzero (alternative) collective dependence, that is, $\mathcal{H}_{0,0}$. However, this type of error is less worrisome since a valid test would be conducted in Steps 2 and 3 of our CALL-DECODE with true alternative. The second type of classification errors corresponds to those k-groups with zero (null) collective dependence, but some of their nonzero DIIs are thresholded incorrectly. For example, some $\ell \in \mathcal{H}_{0,j}$, $j > 0$ can be classified into one of those $\breve{\mathcal{H}}_{0,l}$, $j+1 \le l \le k$. For this type of error, the Steps 2 and 3 of CALL-DECODE are no longer accurate because when comparing the minimum test statistic $T_\ell = \min_{t=1,\ldots,l} |\bar{\Delta}_{\ell,h_t}|$ with the quantile $F_l^{-1}(\beta)$, we hope to accept the null (note this k-group ℓ has a true null), but it turns out the alternative of Step 2.2 is true. Since $l > j$, we would expect the calculated p-value based on an inappropriate quantile $F_l^{-1}(\beta)$ is smaller than it should be based on the correct quantile $F_j^{-1}(\beta)$. This potentially would lead to an inflation of the FDR level. However, as we briefly explained in Remark 3.2, we adopted the minimum test statistic T_ℓ over other options (such as chi-square based test) to alleviate the inflation of FDR level due to insensitivity of T_ℓ on j. In Section B.1.2 of the Supplementary Material (Jernigan et al. (2020)), we confirm this intuition with additional simulation experiments, from which we observe that errors in the classification step, especially the second type of classification errors that some $\ell \in \mathcal{H}_{0,j}$ is misclassified into one of those $\breve{\mathcal{H}}_{0,l}$ for $j+1 \le l \le k$, become substantial when the sample size n is small, the number of unique values L_b taken by X_b is large, or the signal strength is weak. Correspondingly, the FDR control is slightly compromised due to the errors in the classification step, particularly for small n and weak signals. On the other hand, as the signal strength or the sample size increasing, the empirical FDR quickly becomes under control with respect to the nominal level, and, meanwhile, the errors of classification vanish rapidly.

Discussion on future theoretical studies. Methodologically, our procedure CALL-DECODE is motivated from the pioneer works on multiple testing entries within the large covariance and precision matrices (Liu (2013), Cai and Liu (2016)). Consider our test statistics $\{T_\ell\}_{\ell=1}^{p_k}$ with corresponding null reference c.d.f. $\{F_\ell(t)\}_{\ell=1}^{p_k}$, where $F_\ell = F_j$ when the ℓth group is in the category j. The desired FDR control of our method can be established if $\sup_{0\le\beta\le 1-p_k^{-1}} |\sum_{\ell\in\mathcal{H}_0} \mathbb{I}\{T_\ell \ge F_j^{-1}(\beta)\}/\{|\mathcal{H}_0|(1-\beta)\} - 1|$ converges to zero in probability. This uniform convergence can be achieved by bounding the pairwise covariances among T_ℓ's at each value of β (see Lemma 3 of Cai and Liu (2016) or Lemma 6.3 of Liu (2013)). In particular, it is sufficient to show that the majority of pairwise covariances among T_ℓ's are small (see Lemma 4 of Cai and Liu (2016), for instance). With additional sparsity conditions that the majority of null hypotheses are true, the last step is potentially manageable once one can show the test statistic T_ℓ is asymptotically linear. Hence, as a future theoretical study, a refined analysis is in need for the linear representation of each T_ℓ as well as for bounding each pairwise covariance among T_ℓ's by certain polynomials of DIIs, under some regularity conditions.

Extensions and potential challenges. Though we focus on $k = 3$ for the purpose of simple presentation, CALL-DECODE can be easily extended to $k \ge 4$ for collective dependence of higher orders. This, however, will encounter two major challenges. First, large k will lead to more delicate composite nulls and, therefore, increases the complexity of the classification step in CALL-DECODE. Meanwhile, investigating the collective dependence with potentially divergent k (and/or for continuous random variables) is of information theoretic interesting by itself. The second challenge is the increasing computational demands due to the exponentially growing number of quadruplets, quintuplets and so on even for small d. The computational cost due to the bootstrap can be resolved by deriving the asymptotic behavior of empirical estimator of collective dependence. In principal, our procedure can also be easily extended to continuous distributions as long as asymptotic normality of individual marginal entropy estimators $\mathsf{H}(\tau_j)$ can be achieved. For instance, under certain smoothness condition on the joint density function, the Kozachenko–Leonenko estimator Kozachenko and Leonenko (1987) based on the nearest neighbor distances is asymptotically normal when $k = 3$, while certain weighted Kozachenko–Leonenko estimator (Berrett, Samworth and Yuan (2019)) is preferred for general k. Both directions, more or less, can be reduced to investigating the asymptotics and large deviations of certain refined estimator to the entropy, such as the weighted Kozachenko–Leonenko estimator. We will pursue this in a future work.

To demonstrate our method for the protein coevolution study, we considered two protein families, the elongation factor P family and the zinc knuckle family. The identified triplets and hub residues from both families suggest interesting mechanisms of protein interactions with RNA or DNA *in vivo*, and some of these have been validated in literature. In addition to the promising application to protein coevolution, the collective dependence and our procedure are of wide applicability and can be easily applied to other areas in which the data are also categorical or counts with finite unique values. For example, the single-nucleotide polymorphism (SNP) data, whose measurements take values at $\{0, 1, 2\}$ to represent homozygous reference allele, heterozygote and homozygous for alternative allele, or the binary DNA methylation data, where the methylation status of a CpG site is encoded as a binary variable to represent it, is either methylated or unmethylated. Here, our method can be applied to reveal the potential higher order dependence among positions in the genome or CpG sites on some regions of DNA. With those information recovered, we can not only identify critical hubs, as did for the protein coevolution, but also we can further investigate how does the higher order dependence vary across different biological conditions. Furthermore, for both the SNP and binary DNA methylation data, the spatial structure among positions or CpG sites are defined

much clearer than that for the MSA data in protein coevolution. Hence, we can learn how does the higher order dependence align or disagree with the spatial structure of positions in the genome or CpG sites on DNA and uncover more sophisticated mechanisms in the genetic variation or DNA methylation. Exploring our method within other areas will be our future efforts.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. The first and second authors were supported, in part, by NIH Grant R01-GM127701. The third author was supported, in part, by NSF Grant DMS-1812030, an AMS Simons Travel Grant and the Central Research Development Fund at the University of Pittsburgh. The fourth author was supported, in part, by DOE Grant DE-SC0018344 and NSF Grants IIS-1545994 and IOS-1922701.

SUPPLEMENTARY MATERIAL

Supplement to "Large scale multiple inference of collective dependence with applications to protein function" (DOI: 10.1214/20-AOAS1431SUPP; .pdf). In this supplement, we provide technical discussions, additional numerical experiments, and extra results for the real data analysis.

REFERENCES

AFONNIKOV, D. A. and KOLCHANOV, N. A. (2004). CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res*. **32** W64–W68.

BASHARIN, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl*. **4** 333–336. MR0127457 https://doi.org/10.1137/1104033

BELL, A. J. (2003). The co-information lattices. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*: *IC* **2003**.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERMAN, H., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T., WEISSIG, H., SHINDYALOV, I. and BOURNE, P. (2000). The protein data bank. *Nucleic Acids Res*. **28** 235–242.

BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist*. **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

BUENO, R. and MAR, J. C. (2017). Changes in gene expression variability reveal a stable synthetic lethal interaction network in BRCA2-ovarian cancers. *Methods* **131** 74–82. https://doi.org/10.1016/j.ymeth.2017.07.021

BURGER, L. and VAN NIMWEGEN, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol*. **6** e1000633, 18. MR2601389 https://doi.org/10.1371/journal.pcbi.1000633

BUSLJE, C. M., TEPPA, E., DI DOMÉNICO, T., DELFINO, J. M. and NIELSEN, M. (2010). Networks of high mutual information define the structural proximity of catalytic sites: Implications for catalytic residue identification. *PLoS Comput. Biol*. **6** e1000978.

CAI, T. T. and LIU, W. (2016). Large-scale multiple testing of correlations. *J. Amer. Statist. Assoc*. **111** 229–240. MR3494655 https://doi.org/10.1080/01621459.2014.999157

CAI, T. T., LI, H., LIU, W. and XIE, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100** 139–156. MR3034329 https://doi.org/10.1093/biomet/ass058

CHAKRABARTI, P. and BHATTACHARYYA, R. (2007). Geometry of nonbonded interactions involving planar groups in proteins. *Prog. Biophys. Mol. Biol*. **95** 83–137. https://doi.org/10.1016/j.pbiomolbio.2007.03.016

CHAO, J. A., PATSKOVSKY, Y., ALMO, S. C. and SINGER, R. H. (2008). Structural basis for the coevolution of a viral RNA-protein complex. *Nat. Struct. Mol. Biol*. **15** 103–105.

CIRIELLO, G., CERAMI, E., SANDER, C. and SCHULTZ, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*. **22** 398–406. https://doi.org/10.1101/gr.125567.111

COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. MR2239987

CUSICK, M. E., KLITGORD, N., VIDAL, M. and HILL, D. E. (2005). Interactome: Gateway into system biology. *Hum. Mol. Genet.* **14** R171–R181.

DARWIN, C. (1859). *The Origin of Species by Means of Natural Selection*. Modern Lib.

DE GUZMAN, R. N., WU, Z. R., STALLING, C. C., PAPPALARDO, L., BORER, P. N. and SUMMERS, M. F. (1998). Structure of the HIV-1 nucleocapsid protein bound to the SL3 Ψ-RNA recognition element. *Science* **279** 384–388.

DE JUAN, D., PAZOS, F. and VALENCIA, A. (2015). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14** 249–261.

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. *Springer Series in Statistics*. Springer, New York. MR2373771 https://doi.org/10.1007/978-0-387-49317-6

DUNN, S. D., WAHL, L. M. and GLOOR, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24** 333–340.

FIGLIUZZI, M., BARRAT-CHARLAIX, P. and WEIGT, M. (2018). How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35** 1018–1027. https://doi.org/10.1093/molbev/msy007

FINN, R. D., TATE, J., MISTRY, J., COGGILL, P. C., SAMMUT, S. J., HOTZ, H. R., CERIC, G., FORSLUND, K., EDDY, S. R. et al. (2008). The Pfam protein families database. *Nucleic Acids Res.* **36** 281–288.

FLEURET, F. (2004). Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **5** 1531–1555. MR2248026

FRASER, H. B., HIRSH, A. E., WALL, D. P. and EISEN, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* **101** 9033–9038.

GALAS, D. J., NYKTER, M., CARTER, G. W., PRICE, N. D. and SHMULEVICH, I. (2010). Biological information as set-based complexity. *IEEE Trans. Inf. Theory* **56** 667–677. MR2724399 https://doi.org/10.1109/TIT.2009.2037046

GALAS, D. J., SAKHANENKO, N. A., SKUPIN, A. and IGNAC, T. (2014). Describing the complexity of systems: Multivariable "set complexity" and the information basis of systems biology. *J. Comput. Biol.* **21** 118–140. MR3164643 https://doi.org/10.1089/cmb.2013.0039

HALABI, N., RIVOIRE, O., LEIBLER, S. and RANGANATHAN, R. (2009). Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138** 774–786.

HANAWA-SUETSUGU, K., SEKINE, S., SAKAI, H., HORI-TAKEMOTO, C., TERADA, T., UNZAI, S., TAME, J. R., KURAMITSU, S., SHIROUZU, M. et al. (2004). Crystal structure of elongation factor P from Thermus thermophilus HB8. *Proc. Natl. Acad. Sci. USA* **101** 9595–9600.

HOPF, T. A., INGRAHAM, J. B., POELWIJK, F. J., SCHÄRFE, C. P. I., SPRINGER, M., SANDER, C. and MARKS, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35** 128–135. https://doi.org/10.1038/nbt.3769

JERNIGAN, R., JIA, K., REN, Z. and ZHOU, W. (2021). Supplement to "Large-scale multiple inference of collective dependence with applications to protein function." https://doi.org/10.1214/20-AOAS1431SUPP

JIA, K. and JERNIGNA, R. L. (2018). SeqStruct: A new amino acid similarity matrix based on sequence correlations and structural contacts yields sequence-structure congruence. *BioRxiv*. https://doi.org/10.1101/268904

JONES, D. J., BUCHAN, D. W., COZZETTO, D. and PONTIL, M. (2011). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28** 184–190.

KLAMT, S., HAUS, U.-U. and THEIS, F. (2009). Hypergraphs and cellular networks. *PLoS Comput. Biol.* **5** e1000385, 6. MR2516078 https://doi.org/10.1371/journal.pcbi.1000385

KORBER, B. T., FARBER, R. M., WOLPERT, D. H. and LAPEDES, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc. Natl. Acad. Sci. USA* **90** 7176–7180.

KOZACHENKO, L. F. and LEONENKO, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16.

LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453 https://doi.org/10.1214/13-AOS1169

LOCKLESS, S. W. and RANGANATHAN, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286** 295–299.

MARTIN, L., GLOOR, G. B., DUNN, S. and WAHL, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21** 4116–4124.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika* **19** 97–116.
Miller, G. (1955). Note on the bias of information estimates. *Inf. Theory Psychol. Probl. Methods* **II-B** 95–100.
Mishra, S. K. and Jernigna, R. L. (2018). Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics. *PLoS ONE* **13** e0199225.
Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. et al. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108** 1293–1301.
Nelson, D. L. and Cox, M. M. (2005). *Principles of Biochemistry*, 4th ed. W.H. Freeman, New York.
Ochoa, D. and Pazos, F. (2014). Practical aspects of protein coevolution. *Front. Cell Dev. Biol.* **2** 14. https://doi.org/10.3389/fcell.2014.00014
Onken, A., Dragoi, V. and Obermayer, K. (2012). A maximum entropy test for evaluating higher-order correlations in spike counts. *PLoS Comput. Biol.* **8** e1002539, 12. MR2958407 https://doi.org/10.1371/journal.pcbi.1002539
Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.
Reynolds, K. A., McLaughlin, R. N. and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell* **147** 1564–1575.
Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9** 477–485.
Staude, B., Grün, S. and Rotter, S. (2010). Higher-order correlations in non-stationary parallel spike trains: Statistical modeling and inference. *Front. Comput. Neurosci.* **4** 16. https://doi.org/10.3389/fncom.2010.00016
Staude, B., Rotter, S. and Grün, S. (2010). CuBIC: Cumulant based inference of higher-order correlations in massively parallel spike trains. *J. Comput. Neurosci.* **29** 327–350. MR2721349 https://doi.org/10.1007/s10827-009-0195-x
Sun, Y., Zhang, N. R. and Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679 https://doi.org/10.1214/12-AOAS561
Thompson, J. N. (1994). *The Coevolutionary Process*. Univ. Chicago Press, Chicago, IL.
Vejmelka, M. and Paluš, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E* (3) **77** 026214, 12. MR2453293 https://doi.org/10.1103/PhysRevE.77.026214
Vestergaard, B., Van, L. B., Andersen, G. R., Nyborg, J., Buckingham, R. H. and Kjeldgaard, M. (2001). Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. *Mol. Cell* **8** 1375–1382.
Wang, Y. X. R., Jiang, K., Feldman, L. J., Bickel, P. J. and Huang, H. (2015). Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis. *Ann. Appl. Stat.* **9** 300–323. MR3341117 https://doi.org/10.1214/14-AOAS792
Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **106** 67–72.
Wells, J. A. and McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450** 1001–1009.
Xia, Y., Cai, T. and Cai, T. T. (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *J. Amer. Statist. Assoc.* **113** 328–339. MR3803468 https://doi.org/10.1080/01621459.2016.1251930
Yeung, R. W. (1991). A new outlook on Shannon's information measures. *IEEE Trans. Inf. Theory* **37** 466–474. MR1145812 https://doi.org/10.1109/18.79902
Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129 https://doi.org/10.1214/11-AOAS494
Yuan, M., Liu, R., Feng, Y. and Shang, Z. (2018). Testing community structures for hypergraphs. Preprint. Available at arXiv:1810.04617.
Zhang, R., Ren, Z. and Chen, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Comput. Biol.* **14** e1006369.
Zhi, W., Minturn, J., Rappaport, E., Brodeur, G. and Li, H. (2013). Network-based analysis of multivariate gene expression data. *Methods Mol. Biol.* **972** 121–139.