algorithms

MDPI

Article

Network Inference from Gene Expression Data with Distance Correlation and Network Topology Centrality

Kuan Liu, Haiyuan Liu *, Dongyan Sun † and Lei Zhang †

Department of Communication Engineering, College of Electronic Information and Optical Engineering, Nankai University, Tianjin 300350, China; liukuan@mail.nankai.edu.cn (K.L.); sundy@nankai.edu.cn (D.S.); 2120180334@mail.nankai.edu.cn (L.Z.)

* Correspondence: liuhaiyuan@nankai.edu.cn

† These authors contributed equally to this work.

Abstract: The reconstruction of gene regulatory networks based on gene expression data can effectively uncover regulatory relationships between genes and provide a deeper understanding of biological control processes. Non-linear dependence is a common problem in the regulatory mechanisms of gene regulatory networks. Various methods based on information theory have been developed to infer networks. However, the methods have introduced many redundant regulatory relationships in the network inference process. A recent measurement method called distance correlation has, in many cases, shown strong and computationally efficient non-linear correlations. In this paper, we propose a novel regulatory network inference method called the distance-correlation and network topology centrality network (DCNTC) method. The method is based on and extends the Local Density Measurement of Network Node Centrality (LDCNET) algorithm, which has the same choice of network centrality ranking as the LDCNET algorithm, but uses a simpler and more efficient distance correlation measure of association between genes. In this work, we integrate distance correlation and network topological centrality into the reasoning about the structure of gene regulatory networks. We will select optimal thresholds based on the characteristics of the distribution of each gene pair in relation to distance correlation. Experiments were carried out on four network datasets and their performance was compared.

Keywords: distance correlation; gene regulatory networks; integrate; network topology centrality

Citation: Liu, K.; Liu, H.; Sun, D.; Zhang, L. Network Inference from Gene Expression Data with Distance Correlation and Network Topology Centrality. *Algorithms* **2021**, *14*, 61. https://doi.org/10.3390/a14020061

Academic Editor: Tatsuya Akutsu

Received: 19 December 2020

Accepted: 11 February 2021

Published: 15 February 2021

1. Introduction

Systems biology is not only an emerging field, but more importantly, it represents a new approach to biological research [1,2]. In the past, the structure of gene regulatory networks (GRNs) was inferred from experimental interventions, but such experiments required considerable time and cost. With the rapid development of high-throughput technologies, a large number of research studies have generated a large amount of gene expression data [3,4], which has made it possible to infer gene regulatory networks from these expression data based on computational methods. In recent years, network inference based on computational methods has become one of the most important goals in the post-genomic era. To this end, the "Dialogue on Reverse Engineering Evaluation and Methods" challenge aims to stimulate researchers to develop new and efficient arithmetics [5].

Much progress has been made in inferring GRNs' structure from gene expression data. In the early days, Boolean networks [6,7] were popular in GRN inferencing, where the states of genes were represented by Boolean variables, and interactions between genes were represented by Boolean functions, which determined the states of genes on top of some other regulatory gene states. At present, information-theoretic approaches are increasingly being used for reconstructing GRNs. Several mutual information (MI)-based methods have been successfully applied to infer GRNs, such as the relevance network (REL), context likelihood of relatedness (CLR), and the ARACNE and minimum redundancy network

(MRNET). The REL algorithm [8] calculates MI values between genes and then infers interactions based on the threshold values. The CLR algorithm [9] extends the REL algorithm, which infers interactions based on scores derived from the background distribution of mutual information. For both the REL and CLR algorithms, it is easy to introduce indirect interactions, which can lead to more false edges. In order to eliminate indirect interactions, Margolin et al. proposed the ARACNE algorithm [10] based on data-processing inequality, which takes into account indirect interactions in interaction triangles. The MRNET algorithm [11] by Meyer is a network inference algorithm using a feature selection strategy, in which an iterative search process is applied to select direct interactions. PCA-CMI [12] measures the dependence between genes through conditional mutual information(CMI), successfully differentiated direct interaction, and indirect association. CMI2NI [13] calculated the mutual information between two genes when given the third gene by calculating the Kullback–Leibler divergence between the hypothetical distributions at the boundary between the two genes. Although MI can characterize nonlinear dependence, MI's calculation usually requires probability or density estimation, which often requires assumptions. However, it is not accurate because the distribution of gene expression data is uncertain.

Recently, a novel and simple statistic of dependency relationships, distance correlation [14], has emerged, which is sensitive to any deviation from independent behaviors, such as nonlinear or non-monotone dependent structures [15]. Better than MI, the distance correlation statistics were calculated very merely without any distribution assumptions. Recently, an approach has been proposed to infer GRNs based on distance correlation. It successfully combined distance correlation with existing CLR algorithms and MRNET algorithms to raise the accuracy of the GRNs [16].

GRNs, or more generally, biochemical networks are sparse, meaning that a gene is regulated by a small number of genes relative to the total number of genes in the network. A range of sparsification-based features have now been proposed to infer GRNs from gene expression data [17,18]. Currently, we can sparse networks by using graph properties in the network [19,20]. In the existing studies, most algorithms only consider the connections in the gene expression data. Still, they do not include the known graph attributes in the reasoning process, which reduces these methods' prediction veracity rates and restricts their availability in practice [21]. A large number of studies have shown a hierarchical, scale-free nature of biological networks [5,22]. This attribute makes most nodes in the network sparsely connected, where a few positively associated genes account for most of the interaction, which are hub nodes. The hub node is a node with high network centrality. The connection rules between nodes reflect the relative position information between nodes to some extent. Recently, the LDCNET algorithm has successfully merged MI and network topology centrality for reconstructing GRNs [23].

In this article, we develop a novel method, namely DCNTC, which incorporate the distance correlation and network topology centrality into GRNs inferring algorithms and test the performance. Our approach adopts a novel estimation of measurement, and the sparse (scale-free) structure of the gene regulatory network is used to calculate the network topology centrality. Compared with the traditional methods, we use the latest distance correlation statistics to measure the nonlinear relationship, and combine the network topology centrality to infer the gene regulation network. At the same time, we select the optimal threshold according to the distribution characteristics of the values calculated by the distance correlation of each gene pair. Real data from the SOS DNA repair network and DREAM-simulated data suggest that the DCNTC algorithm can improve the GRNs' inference accuracy.

2. Methods

In this section, the definitions of distance correlation and network topology centrality will be reviewed, as well as the algorithm of the DCNTC for inferring GRNs.

2.1. Distance Correlation

Distance correlation provides a new approach to the problem of testing the joint independence of random vectors [14,24]. The energy package in R provides the calculation function of distance correlation [25]. Distance correlation [14] was raised as an innovative method to detect the dependence. The key idea is to calculate the difference between the joint eigenfunction and the product of its marginal eigenfunction in a special, weighted L2 space. Specifically, for random variables (X, Y), denote an innovative method of (X, Y) by $f_{(X,Y)}$, and its marginal eigenfunction $f_{(X)}$ and $f_{(Y)}$. The distance covariance between X and Y is defined as the root of the following equation:

$$dcov^2(X,Y) = \int_{R^{(p+q)}} |f_{(X,Y)}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)d_t d_s, \quad (1)$$

where p and q are the dimensions of X and Y, respectively, and $w(t,s)$ is the weight function given by $(C_p C_q |t|_p^{q+1})^{-1}$ with $C_p = \pi^{1+p}/2\Gamma(1+p)/2$ and $C_q = \pi 1 + q/2\Gamma((1+q)/2)$. By standardizing the distance covariance, the distance correlation can be defined as,

$$dcor(X,Y) = \frac{dcov(X,Y)}{\sqrt{dcov(X,X)}\sqrt{dcov(Y,Y)}}. \quad (2)$$

2.2. Network Topology Centrality

Network centrality is a network topology feature used to measure nodes' importance because it can effectively evaluate the network position relative to other nodes in a local scope. Therefore, network topology centrality can be applied to assess the significance of nodes in a network. Commonly used network centralities are closeness centrality, degree centrality, and betweenness centrality. In the network, the node's importance is usually calculated by its degree centrality of the node. Degree centrality was formally defined as the count of links on a node, which is often known as an analytical method of how nodes can be affected by flow into a given network [26]. In an undirected graph, the node's degree is the count of other nodes to which it is connected. In a graph G with n nodes, we commonly use the adjacency matrix A = $[a_{\mu v}]$ to describe the connectivity between nodes. Define the adjacency matrix A, where $a_{\mu v} = 1$, μ is adjacent to v, and $a_{\mu v} = 0$, μ is not adjacent to v or $\mu = v$. Degree centrality is based on the number of connected edges of node v as the importance of node v. The calculation of node centrality is as follows:

$$DC(v) = \sum_{\mu \in V_{NB(v)}} a_{\mu v}, \quad (3)$$

where $a_{\mu v}$ is the element in the adjacency matrix A, and $V_{NB(v)}$ is the adjacency subgraph of v.

2.3. GRNs Inference with DCNTC Algorithm

In this paper, we propose an algorithm DCNTC for inferring the structure of a regulatory network based on network topology centrality and distance correlation, which is based on the network properties and correlation between genes. The algorithm consists of three main parts: (1) Initialisation of the regulatory relationships, (2) calculation of the network topology centrality and optimisation of the ranking, and (3) inference of the regulatory network structure.

2.3.1. Initialization of Regulatory Relationships

The first step in the DCNTC algorithm is to initialize and pre-treat the regulatory relationships between the genes. As distance correlation is an effective way to quantitatively describe non-linear relationships between genes, a distance correlation matrix M is constructed for the input gene expression data based on Equation (2). The greater the value of the elements in this matrix, the greater the likelihood that the gene pair to which

it is addressed has a regulatory relationship. Considering the high noise level of gene expression data and the sparse nature of the regulatory network, elements of the M-matrix need to be pre-processed to eliminate some of the redundant regulatory relationships before the network structure can be inferred. This is usually done by setting a fixed threshold value. When the value of an element in the matrix is greater than the given threshold value, there is a preliminary regulatory relationship between the two genes to which the element corresponds; when the value of an element in the matrix is less than the threshold value, there is no regulatory relationship between the two genes to which the element corresponds, and the matrix element is set to zero.

In the regulatory relationship matrix, we considered that there was no regulatory relationship between pairs of genes with small distance correlations, which can also be described as redundant relationships. In the process of selecting the threshold for initial de-redundancy, we further analysed the distribution of distance correlation values for each gene pair in the different datasets, as shown in Figure 1. In Figure 1, we separated the range of distance correlations (0–1) in steps of 0.1 and counted the ratio of how many values there were in each interval across the different datasets. We found a single peak in the distance correlation distribution plot. To initially remove redundant relationships from the gene regulatory network, we chose the left boundary value of the interval where the peak was located as the threshold value θ for redundancy removal. This allowed us to initially remove redundant relationships from the initially obtained regulatory relationship matrix by threshold θ.

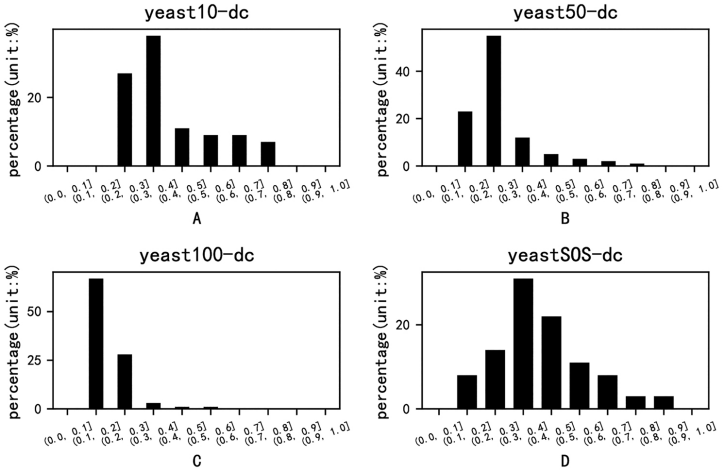

Figure 1. Distribution characteristics of distance correlation statistics.(**A**) Distribution of yeast 10-distance correlation characteristics. (**B**) Distribution of yeast 50-distance correlation characteristics. (**C**) Distribution of yeast 100-distance correlation characteristics. (**D**) Distribution of distance-related features in the sos dataset.

2.3.2. Calculation of the Network Topology Centrality and Optimization of the Ranking

The algorithm then uses the pre-processed gene regulatory relationships and the matrix M to calculate and rank the network topology centrality of each gene. Given the simple and easy-to-implement nature of node centrality, the algorithm uses it as a measure of the neutrality of the network topology of each gene. In this paper, distance correlation is used to measure the regulatory relationships between genes, so that the degree centrality of gene g_v in the network G can be expressed as follows:

$$DC(v) = \sum_{j=1}^{n} \chi((dcor(g_v, g_j)) - d_c) \quad (4)$$

$$\chi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

where d_c represents the given cut-off distance. Essentially, the value of $DC(v)$ is equal to the number of genes for which the value associated with gene g_v distance exceeds the given cut-off distance d_c. From Equation (4), it can be seen that the topological centrality of a gene network is influenced by the value of the cut-off d_c. The magnitude of the value is directly related to the calculation of the topological centrality of the gene network. Specifically, for the updated regulatory relationship matrix M, M_{ij} represents the initial regulatory relationship values for genes i and j. The sequence $M_{ij}(i \leq j)$ is ranked downwards, and the resulting sequence is $M_1 \leq M_2 \leq ... \leq M_T$. We set the d_c value from the αT regulatory relationship value in the sequence, where the parameter α is set to 20% by default. This strategy uses statistical means to obtain the truncation distance values, so the values obtained are more scientifically valid.

Once the value of the cut-off d_c has been determined, the centrality of the node is calculated for each gene in turn, according to Equation (4). From this process, it can be seen that there may be cases where two or more genes have the same node centrality value. In order to effectively distinguish the importance of genes with the same nodal centrality, we consider the re-sequencing of all genes with the same nodal centrality. The ranking process takes into account the nodal centrality of all the genes directly adjacent to the target gene, and measures the importance of the target gene according to the following formula:

$$SDC(\mu) = \sum_{v \in V_{NB(\mu)}} DC(v), \tag{5}$$

where $V_{NB(\mu)}$ represents the set of nodes whose distance from node μ is more correlated than d_c. The larger the value of this formula, the greater the importance of the target gene.

For different genes with the same node centrality, we calculate the corresponding values according to Equation (5), where the gene with the higher score is positioned ahead of the gene with the lower score, and the new standard sequence $\left[q_i'\right] i = 1, 2, ...n$ is obtained.

2.3.3. Inference of the Regulatory Network Structure

In order to construct the complete gene regulatory network, we will select regulatory genes for each gene based on the final sequencing results obtained. According to the scale-free nature of the network, genes with high network centrality will be linked to genes with low network centrality, so that in sequence a, genes with lower order are linked by genes with higher order. Additionally, given the sparsity of biological networks, we limit the number of regulatory genes selected for the target gene to one, meaning that we select only the gene with the greatest distance correlation to the target gene as the regulatory gene. For a network with n genes, this indirectly limits the number of correct edges predicted to, at most, $n - 1$. In summary, the computational process can be implemented by the following function:

$$X_j = arg \max_{X_k} dcor(X_i, X_k), \tag{6}$$

where X_k is the gene whose position precedes gene X_i in the standard sequence $\left[q_i'\right] i = 1, 2, ...n$. This function allows us to select the genes with the greatest distance correlation to gene X_i as regulatory genes for gene X_i from among the genes whose sequence position precedes gene X_i. When the regulatory genes for all the genes are available, we can then construct them into a complete regulatory network structure. The complete algorithm implementation flow is shown in Algorithm 1.

Algorithm 1 DCNTC algorithm

Input: Microarray data $G = g_1, g_2, ...g_n$, the threshold θ
Output: A gene network

```
Initialize Q← ∅
Construct a distance correlation matrix M according to Equation (2)
Adjust matrix M using the threshold θ
for each gene g_c: 1 to n do do
    compute the DC value of g_c according to Equation (4)
Rank the genes g_c ∈ G according to the DC value in descending order and store in Q
for each gene g_c:1 to n do do
    select a regulatory gene of g_c using Equation (6) from Q
return result
```

3. Results

In this section, we describe extensive experiments evaluating the performance of the proposed method. Four regulatory network datasets were used in the experiments. Our proposed method was compared with four network inference algorithms based on information theory: CLR, ARACNE, MRNET, LDCNET, and two methods based on distance correlation: REL-DC and MRNET-DC. For the ARACNE algorithm, we chose the default threshold to discriminate the final regulatory relationship. The CLR and MRNET algorithms were implemented using the optimal threshold selection method from article [27]. The code for all the algorithms was implemented in R and Matlab, respectively. The distance dependence was calculated using the existing dcor function in R. The REL, CLR, ARACNE, and MRNET algorithms were implemented using the existing R package MINET [28]; the LDCNET and DCNTC algorithms were edited and implemented in Matlab.

All of the experiments were performed on four network datasets, including simulated and real data. The datasets can be obtained from previous studies.

The DREAM3-10 gene dataset [5], contains 10 samples for 10 genes. It is from the DREAM ("Dialogue for Reverse Engineering Assessments and Methods") project, and represents a yeast gene network. The true network is composed of 10 nodes and 10 edges.

The DREAM3-50 gene dataset [5], contains 50 samples for 50 genes. It also belongs to the DREAM project and represents a yeast gene network. The true network is composed of 50 nodes and 77 edges.

The DREAM3-100 gene dataset [5], contains 100 samples for 100 genes. It also belongs to the DREAM project and represents a yeast gene network. The true network is composed of 100 nodes and 166 edges.

SOS [29,30], contains nine samples for nine genes. It is an SOS DNA repair network in Escherichia coli. The true network is composed of 9 nodes and 24 edges.

To contrast the reasoning methods objectively, it is necessary to measure their performance quantitatively. The predictive results are defined as follows: false-positive (FP), true-positive (TP), false-positive rate (FPR), true-positive rate (TPR), positive predictive value (PPV), accuracy (ACC), and Matthews coefficient constant (MCC) [31]. Mathematically, they are defined by:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$PPV = \frac{TP}{TP + FP} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (11)$$

where TP, FP, TN, and FN are the elements in the formula. These are accounts of true-positives, false-positives, true-negatives, and false-negatives, respectively.

3.1. Results for DREAM3 Challenge Network

To test the prediction effect of the DCNTC algorithm on simulated data, DREAM challenge data, which were widely used, will be used as test data to reconstruct the GRNs [32]. DREAM3 is one of many DREAM Challenge subprojects that provide users with baseline data and control networks to test and evaluate the regulatory network model's effectiveness. We validated our approach on the DREAM3 dataset, where the sizes of the yeast knockout gene expression data were 10, 50, and 100, respectively.

In the first, we tested the DCNTC algorithm on the yeast gene expression data of network size 10 and a sample size of 10. Figure 2 indicates the network extrapolated from gene expression data in different ways. Figure 2A is the real GRN with 10 nodes and 10 edges chosen from an experimentally validated network in yeast genomes. Figure 2B is the network derived from gene expression data with the LDCNET algorithm. The edges of solid dotted black lines are correctly deduced, and the edges of red dotted lines are incorrect. On the edge of false inferences, G2–G9 is a redundant edge, whereas edge G3–G5 is unfound. Figure 2C shows the network deduced from the gene expression data, which is the network inferred from us. Clearly, in our network, the inexistent regulations of G2–G9 were successfully removed. The performance data for each of the seven algorithms is shown in Table 1.

Table 1. Comparison of the performance of different methods for inferring a 10-gene network in DREAM3.

Method	TP	FP	TN	FN	TPR	FPR	PPV	ACC	MCC
CLR	6	10	25	4	0.600	0.286	0.375	0.689	0.273
ARACNE	6	6	29	4	0.600	0.171	0.500	0.778	0.403
MRNET	6	12	23	4	0.600	0.343	0.333	0.644	0.218
REL-DC	10	4	31	0	1	0.114	0.714	0.911	0.795
MRNET-DC	10	13	22	0	1	0.371	0.435	0.711	0.523
LDCNET	8	1	34	2	0.8	0.029	0.889	0.933	0.802
DCNTC	**9**	**0**	**35**	**1**	**0.9**	**0**	**1**	**0.978**	**0.936**

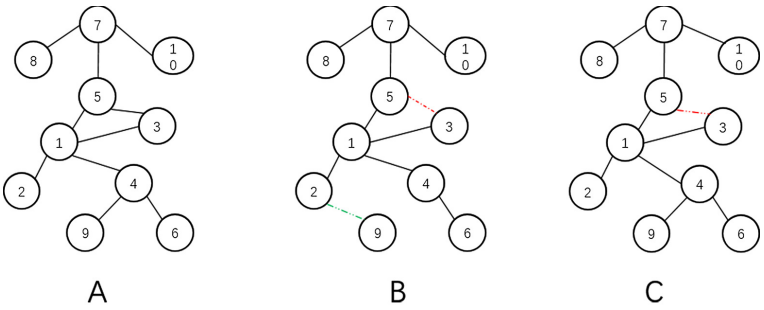

Figure 2. Comparison of a 10-gene network inferred from a DREAM3 dataset. (**A**) The true network with 10 nodes and 10 edges. (**B**) The network inferred by LDCNET from gene expression data. The edge with red dotted lines G3–G5 is false-positive, while the edge G2–G9 is false-negative. (**C**) The network inferred from gene expression data. The false-negative edge G2–G9 in LDCNET was successfully removed by DCNTC.

In the second, we tested the DCNTC algorithm on the yeast gene expression data of network size 50 and a sample size of 50. The real network consists of 50 nodes and 77 edges. In Table 2, we can find that the REL algorithm predicts more accurate edges, but this algorithm also expects a lot of wrong edges. The MRNET algorithm can only predict a tiny number of edges among the four algorithms. In comparison, among the four algorithms, the DCNTC algorithm has higher accuracy (ACC) and lower FPR.

Table 2. Comparison of the performance of different methods for inferring a 50-gene network in DREAM3.

Method	TP	FP	TN	FN	TPR	FPR	PPV	ACC	MCC
CLR	19	165	983	58	0.247	0.144	0.103	0.818	0.070
ARACNE	13	125	1023	64	0.170	0.109	0.094	0.846	0.046
MRNET	21	215	933	56	0.273	0.187	0.089	0.779	0.053
REL-DC	34	49	1099	43	0.442	0.043	0.410	0.925	0.385
MRNET-DC	70	465	683	7	0.909	0.405	0.131	0.615	0.247
LDCNET	23	26	1122	54	0.299	0.023	0.469	0.935	0.342
DCNTC	**29**	**20**	**1128**	**48**	**0.377**	**0.017**	**0.592**	**0.945**	**0.445**

In the third, we tested the DCNTC algorithm on yeast gene expression data with a network size of 100 and a sample size of 100. The whole network consists of 100 nodes and 166 edges. In Table 3, we can find that the MRNET algorithm predicts more accurate boundaries, but this algorithm also expects a lot of unfair advantages. In comparison, among the four algorithms, the DCNTC algorithm has higher accuracy (ACC) and lower FPR.

Table 3. Comparison of the performance of different methods for inferring a 100-gene network in DREAM3.

Method	TP	FP	TN	FN	TPR	FPR	PPV	ACC	MCC
CLR	39	713	4071	127	0.235	0.149	0.052	0.830	0.044
ARACNE	20	417	4367	146	0.121	0.087	0.046	0.886	0.403
MRNET	49	984	3800	117	0.295	0.206	0.047	0.778	0.040
REL-DC	121	386	4398	45	0.729	0.081	0.239	0.913	0.385
MRNET-DC	145	2011	2773	21	0.874	0.421	0.067	0.590	0.165
LDCNET	45	54	4730	121	0.271	0.011	0.455	0.965	0.334
DCNTC	**55**	**42**	**4742**	**111**	**0.331**	**0.009**	**0.567**	**0.969**	**0.419**

On the simulated dataset, the algorithm in this paper performed better in all aspects, mainly because only genes with the greatest distance correlation to the target gene were selected when picking regulatory genes for the target gene. This controls the introduction of false-positive edges to a certain extent. Compared to the other algorithms, the ARACNE algorithm removes the introduction of redundant edges to some extent by removing the loops present in the gene regulatory network; the MRNET algorithm uses a maximum correlation and minimum redundancy strategy to select regulatory genes, which can predict more true-positive edges, but also introduces the most false-positive edges in humans, leading to a decrease in prediction performance. In summary, our algorithm can better control the introduction of redundant edges, thus improving the prediction accuracy.

3.2. Result for SOS Network in E. coil

The SOS network [29] is a signal pathway in the DNA repair system. It is inferred from real gene expression data and is frequently used to test the effectiveness of network inference methods. Here, we test the DCNTC on the network of *E. coli*. Table 4 shows the

results of the seven methods applied to the SOS network in the *E. coli* dataset. The results show that the method outperforms all other methods except MRNET and LDCNET in terms of MCC. Although our method did not identify the most correct edges, it produced the least redundant edges of most methods. Furthermore, the validity of the model on the SOS network was not satisfactory compared to previous experiments on other datasets. The main reason for this finding is related to the characteristics of the SOS network and is due to two main factors: firstly, noise in the real data can be taken into account, making the calculation of node centrality inaccurate and leading to biases in the relative positions between gene regulation, resulting in fewer true-positive edges being predicted; secondly, the SOS network consists of nine nodes and 22 edges. In the final construction of the network, the algorithm selects only the connected edges with the most important relationships to the target gene regulation, which indirectly limits the count of boundaries in the prediction network to eight. Although the experimental results are not ideal, Table 4 shows that our method is no more effective than any other test method other than MRNET and LDCNET.

Table 4. Comparison of different methods on the SOS DNA repair network.

Method	TP	FP	TN	FN	TPR	FPR	PPV	ACC	MCC
CLR	12	5	7	12	0.500	0.417	0.706	0.528	0.079
ARACNE	7	3	9	17	0.292	0.250	0.700	0.444	0.044
MRNET	17	6	6	7	0.708	0.500	0.739	0.639	0.205
REL-DC	6	3	9	18	0.250	0.250	0.667	0.417	0
MRNET-DC	12	9	3	12	0.500	0.750	0.572	0.417	−0.239
LDCNET	6	1	11	18	0.250	0.083	0.857	0.472	0.199
DCNTC	**6**	**2**	**10**	**18**	**0.25**	**0.167**	**0.75**	**0.444**	**0.095**

4. Discussion

In this paper, we fused simple distance correlation and network topological centrality into a structural inference algorithm for GRNs and tested its performance. We selected simulated and real data sets commonly used in gene regulatory network construction algorithms and compared them with existing mutual information-based and distance-correlation-based algorithms. The algorithm can better control the introduction of redundant relationships by ranking the network topological centrality and selecting the maximum distance correlation as the regulatory gene. Therefore, it is conducive to the extension of large-scale network applications, but with many nodes in large-scale networks, there may be more genes with the same node importance when measuring node importance, which may require consideration of higher-order neighbourhood information to measure node importance. As the number of genes increases, it can also lead to inaccuracies and difficulties when inferring the network.

Node centrality is an important metric to characterise the criticality of a node. Many node importance metrics have been proposed, including degree centrality, median centrality, proximity centrality, and eigenvector centrality. The degree centrality is the simplest metric to characterise the importance of a node. Gabrys et al. found that in scale-free or exponential networks, there are only a small number of nodes of large degree, and this are of high importance [33]. Median centrality and proximity centrality are good measures of the importance of nodes in terms of network connectivity, but are computationally complex due to the need to know global information about the network in advance, and are not suitable for large and complex networks. Fabian et al. investigated the extent to which different centrality measures (degree, strength, tightness, mediation, and eigenvectors) recover potential causal interactions in directed acyclic graphs [34,35], whereas feature vector centrality shows powerful effects in measuring the importance of nodes. In summary, the algorithm in the paper has some limitations. In the next study, we will try to reconstruct

the gene regulatory network using different approaches to compute the centrality of the network topology.

The DCNTC algorithm has a clear advantage over the distance-based REL algorithm and the distance-based MRNET algorithm in every performance indicator. The algorithm is similar to the LDCNET algorithm in that both have a central network node location. In combination with the distance-dependent initial network redundancy and network topology centrality, the DCNTC algorithm can predict the structure of the GRN more accurately.

5. Conclusions

In this article, we proposed a novel algorithm DCNTC for inferring GRNs from gene expression data, taking into account the sparse structure of GRNs and the nonlinear dependence. In this method, the nonlinear dependence is represented by distance correlation (without assuming the probability distribution) between this gene pair. The sparse (scale-free) control structure of the gene regulatory network is used to calculate the network topology centrality and test the predictive performance of the algorithm on four data sets, which can effectively predict the structure of the gene regulatory network. However, the algorithm has certain limitations. How could we further confirm the direction of the gene regulatory network structure and its centrality in exploiting the network topology? Further research is still needed in the precise construction of gene regulatory networks.

Author Contributions: Conceptualization, K.L. and L.Z.; Data curation, K.L.; Formal analysis, L.Z.; Investigation, H.L., D.S. and L.Z.; Methodology, K.L.; Project administration, H.L.; Resources, D.S. and L.Z.; Software, K.L.; Supervision, H.L.; Validation, K.L., H.L. and D.S.; Visualization, K.L.; Writing—original draft, K.L. and L.Z.; Writing—review, H.L. and D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found on the Dream Challenges website and in the datasets provided in the published [12] or [16].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; Bernardo, D.D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **2007**, *3*, 78. [CrossRef]
2. Markowetz, F.; Spang, R. Inferring cellular networks—A review. *BMC Bioinform.* **2007**, *8*, 1–17. [CrossRef]
3. Basso, K.; Margolin, A.A.; Stolovitzky, G.; Klein, U.; Dalla-Favera, R.; Califano, A. Reverse Engineering of Regulatory Networks in Human B Cells. *Nat. Genet.* **2005**, *37*, 382–390. [CrossRef]
4. Margolin, A.A.; Wang, K.; Lim, W.K.; Kustagi, M.; Nemenman, I.; Califano, A. Reverse engineering cellular networks. *Nat. Protoc.* **2006**, *1*, 662–671. [CrossRef]
5. Marbach, D.; Prill, R.J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 6286–6291. [CrossRef]
6. Liang, S. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In Proceedings of the Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, Hawaii, HI, USA, 4–9 January 1998.
7. Ruz, G.A.; Goles, E. Learning gene regulatory networks using the bees algorithm. *Neural Comput. Appl.* **2013**, *22*, 63–70. [CrossRef]
8. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In Proceedings of the Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, Hawaii, HI, USA, 4–9 January 2000; Volume 5, pp. 418–429.
9. Faith, J.J.; Hayete, B.; Thaden, J.T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J.J.; Gardner, T.S. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol.* **2007**, *5*, e8. [CrossRef]
10. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.H.; Stolovitzky, G.; Favera, R.D.; Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinform.* **2006**, *7*, 1–15. [CrossRef]
11. Meyer, P.E.; Kontos, K.; Lafitte, F.; Bontempi, G. Information-Theoretic Inference of Large Transcriptional Regulatory Networks. *Eurasip J. Bioinform. Syst. Biol.* **2007**, *2007*, 1–9. [CrossRef] [PubMed]

12. Zhang, X.; Zhao, X.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J.; Liu, Z.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **2012**, *28*, 98–104. [CrossRef]
13. Zhang, X.; Zhao, J.; Hao, J.; Zhao, X.; Chen, L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.* **2015**, *43*, e31. [CrossRef]
14. Szekely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
15. Li, R.; Zhong, W.; Zhu, L. Feature Screening via Distance Correlation Learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef]
16. Guo, X.; Zhang, Y.; Hu, W.; Tan, H.; Wang, X. Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation. *PLoS ONE* **2014**, *9*, e87446. [CrossRef] [PubMed]
17. Li, Y.; Liu, D.; Chu, J.; Zhu, Y.; Cheng, X. A Sparse Bayesian Learning Method for Structural Equation Model-based Gene Regulatory Network Inference. *IEEE Access* **2020**, *8*, 40067–40080. [CrossRef]
18. Chen, C.; Zhang, M.; Zhang, D. Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations. *Statistics* **2015**, *19*, 40–73.
19. Kim, J.; Kim, I.; Han, S.K.; Bowie, J.U.; Kim, S. Network rewiring is an important mechanism of gene essentiality change. *Sci. Rep.* **2012**, *2*, 900. [CrossRef]
20. Estrada, E.; Rodriguezvelazquez, J.A. Subgraph centrality in complex networks. *Phys. Rev. E* **2005**, *71*, 056103. [CrossRef]
21. Ruyssinck, J.; Demeester, P.; Dhaene, T.; Saeys, Y. Netter: Re-ranking gene network inference predictions using structural network properties. *BMC Bioinform.* **2016**, *17*, 76. [CrossRef]
22. Jeong, H.; Mason, S.P. Lethality andcentrality in protein networks. *Nature* **2001**, *411*, 41–42. [CrossRef]
23. Liu, W. Research on the Structure Prediction Algorithm of Gene Regulation Network Based on Information Theory. Ph.D. Thesis, Hunan University, Hunan, China, 2017. (In Chinese)
24. Kosorok, M.R. Discussion of: Brownian distance covariance. *Ann. Appl. Stat.* **2009**, *3*, 1270–1278. [CrossRef]
25. Rizzo ML, S.G. Energy: E-Statistics (Energy Statistics). R Package Version 1.6.2. 2014. Available online: http://CRAN.R-project.org/package=energy (accessed on 1 May 2019).
26. Batool, K.; Niazi, M.A. Correction: Towards a Methodology for Validation of Centrality Measures in Complex Networks. *PLoS ONE* **2014**, *9*, e98379. [CrossRef]
27. Liu, W.; Zhu, W.; Liao, B.; Chen, H.; Ren, S.; Cai, L. Improving gene regulatory network structure using redundancy reduction in the MRNET algorithm. *RSC Adv.* **2017**, *7*, 23222–23233. [CrossRef]
28. Meyer, P.E.; Lafitte, F.; Bontempi, G. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinform.* **2008**, *9*, 1–10. [CrossRef]
29. Ronen, M.; Rosenberg, R.; Shraiman, B.I.; Alon, U. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 10555–10560. [CrossRef]
30. Shenorr, S.S.; Milo, R.; Mangan, S.; Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **2002**, *31*, 64–68. [CrossRef]
31. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451. [CrossRef]
32. Hache, H.; Wierling, C.; Lehrach, H.; Herwig, R. GeNGe: Systematic Generation of Gene Regulatory Networks. *Bioinformatics* **2009**, *25*, 1205–1207. [CrossRef] [PubMed]
33. Gabrys, B. Propagation Phenomena in Real World Networks. *Intell. Syst. Ref. Libr.* **2015**, *85*, 1–24.
34. Dablander, F.; Hinne, M. Node Centrality Measures are a Poor Substitute for Causal Inference. *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef]
35. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [CrossRef]