Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2011 ; 8(5): 1208–1222. doi:10.1109/TCBB.2010.95.

Continuous Cotemporal Probabilistic Modeling of Systems Biology Networks from Sparse Data

David J. John,
Department of Computer Science, Wake Forest University. djj@wfu.edu

Jacquelyn S. Fetrow, and
Dean of Wake Forest College and is with the Departments of Computer Science and Physics, Wake Forest University. fetrowjs@wfu.edu

James L. Norris
Department of Mathematics, Wake Forest University. norris@wfu.edu

Abstract

Modeling of biological networks is a difficult endeavour, but exploration of this problem is essential for understanding the systems behaviour of biological processes. In this contribution, developed for sparse data, we present a new continuous Bayesian graphical learning algorithm to cotemporally model proteins in signaling networks and genes in transcriptional regulatory networks. In this continuous Bayesian algorithm the correlation matrix is singular because the number of time points is less than the number of biological entities (genes or proteins). A suitable restriction on the degree of the graph's vertices is applied and a Metropolis-Hastings algorithm is guided by a BIC-based posterior probability score. Ten independent and diverse runs of the algorithm are conducted, so that the probability space is properly well-explored. Diagnostics to test the applicability of the algorithm to the specific data sets are developed; this is a major benefit of the methodology. This novel algorithm is applied to two time course experimental data sets: 1) protein modification data identifying a potential signaling network in chondrocytes; and 2) gene expression data identifying the transcriptional regulatory network underlying dendritic cell maturation. This method gives high estimated posterior probabilities to many of the proteins' directed edges that are predicted by the literature; for the gene study, the method gives high posterior probabilities to many of the literature-predicted sibling edges. In simulations, the method gives substantially higher estimated posterior probabilities for true edges and true subnetworks than for their false counterparts.

Keywords

Biological system modeling; statistical computing; multivariate statistics; correlation and regression analysis; signal transduction networks; transcriptional regulatory networks; biological network modeling

1 Introduction

Protein modification and gene expression studies are often conducted to study the temporal relationships between different entities, proteins or genes, after some perturbation of their environment. These studies have the potential to yield important information about the entities' interactions, pathways and associations, information that might be identified by appropriate modeling methods. Often, the very high cost of obtaining and processing the samples forces the number of sampled time points to be small; however, many different

entities are often examined. This causes the time series data of the various entities to be sparse, thus presenting modeling challenges.

In contrast to this sparse situation, when the number of time points *t* exceeds the number of entities *k* then direct use of methods such as partial correlations [1] and testing for non-zeros (and thus associations) in marginal covariance matrices [2] are potential approaches for inferring connections between the entities as graph edges.

In order to use partial correlation or zero covariance methods in the sparse data situation, one might be tempted to restrict the number of *utilized entities* or *modules* to be less than the number of time points by either selecting a representative or average of each type (e.g. a type might be a cluster of entities with a similar response pattern) [3] or by simultaneously predicting *modules* of commonly acting entities while predicting the network [4]. This can be a benefit, but when *t* is quite small, say around 6, only a very small number of entities or modules could then be used, which would yield little biological information.

Because (full) partial correlations require non-singular covariance matrices which do not exist in the sparse, $k > t$, setting, some researchers use shrinkage, lasso, and other regularization techniques to estimate these partial correlations, e.g., [5], [6], [7], [8] and [9]. Others compute low-order partial correlations which only adjust for a small number of entities, e.g. [10], [11] and [12].

In developing models, some researchers [13], [14], discretize each entity's values into a small number of bins, while others [12] use continuous data. In this paper, we retain the original continuous data, develop rigorous continuous modeling techniques based on multivariate log-normal theory, construct independent robust probability-based movements through the association space, and produce diagnostics to test the suitability of our method for specific data sets.

One can develop network models based on two different time paradigms. In a previous paper [15], the *next state* time paradigm was considered. In next state, it is assumed that a *predicting* entity's level at one sampled time point influences a *response* entity's level at the next time point in accordance with a stationary Markov process. For a data set which fits this model well, based on the next-state diagnostic tests, network models can be developed. However, in some cases, these diagnostics suggest that the next-state model is not appropriate; in particular, when modeling biological processes the next-state model can be very sensitive to the sample times that are chosen.

In this paper, we focus exclusively on the *cotemporal* time paradigm. The resulting models represent *associations* between entities' measurements at the time points, cotemporally, rather than from one sampled time point to the next. For this cotemporal setting we assume that the *k* entities' values at the *t* sampled time points give us approximations to *t* independent samples of the associations between the entities. In Section 4, we present diagnostics to test this and other model assumptions for particular data sets.

Just as with general observational studies that are cotemporal in nature, when searching for a potential causal association between two particular variables (entities) it is necessary to adjust for the levels of the other variables [16], [17]; otherwise they confound searches. When the number of time points, *t*, does not exceed the number of entities, *k*, despite not being able to compute the full partial correlations (and thus linearly adjust for all other entities), the method searches for the best small sets of predictors and estimates probabilities for other restricted sized sets. Entity *i* is in a highly predicting set for entity *j* if it out competes (in a likelihood sense) all but a small set of other entities, and it provides additional predicting power for *j* beyond that of the other entities in the small set. Restricting

the number of predictors (here, parents in the directed acyclic graph) has the added benefit of only claiming associations that are most profound and thus greatly simplifying interpretations and increasing algorithmic speed [18].

This cotemporal model is an example of a dependency network [19] which is based on low-order (small number of predictors) regression. This limitation on the number of parents has similarity to sparse networks that are developed under low-order conditional independence [12], [20]. For both, a particular directed association (edge) into an entity is only claimed if its presence substantially improves a model which already has a small number of potentially strong (predicting) edges. As in dependency networks, the graphical modeling adjusts for potentially confounding entities before claiming an association, represented by an edge, between two particular entities.

In this paper, we conduct rigorous cotemporal modeling after extensive searching over the hilly graphical network space, and we estimate probabilities for graphs and edges. Just as importantly, we construct diagnostics to evaluate of the utility of the models for a given data set. In Section 2, the details of the modeling are presented, beginning with the assumption that the vector of the k entities' values at a given time point follows a multivariate log-normal distribution. The number of possible networks is often extremely large. Section 3 describes an algorithmic search for the *best* networks and for estimating probabilities of edges and networks. In Section 4, the testing diagnostics, used to evaluation the suitability of our method for particular data sets, are presented. Section 5 describes the two sets of biological data that are used in this paper. Sections 6 and 7 apply the cotemporal modeling process to experimentally derived protein modification and gene expression data sets, respectively. Simulations involving our method and other methods are presented in Section 8. Section 9 summarizes the methodology and results.

2 Multivariate Log-Normal Distributions and Directed Acyclic Graphs

The application of a log transform to the original data changes the common multiplicative (percentage change) chance errors to additive ones, which converts rightskewed distributions for the entities' levels into more normal ones [21]. In order for each of the k (protein or gene) entities to be calibrated on the same scale, we standardize the list of logarithmic values for each respective variable; specifically for each value, we subtract the list's mean over the t time points and divide by its standard deviation. The resulting data is then modeled based on multivariate normal distributions. Use of the above normal (Gaussian) assumption is very popular in graphical modeling. For example, this distribution is used to model the *Arabidopsis thaliana* isoprenoid gene network [12], the *Arabidopsis thaliana* transcriptome gene network [22], and human cortical networks [23]. Due to standardization, mean vectors will consist of zeros while the covariance matrix will be equivalent to the correlation matrix.

For this paper, t, the number of time points, does not exceed k, the number of entities; as a consequence, our sample covariance matrix is singular [24] hence maximum likelihood estimation over all entries is not feasible. Therefore, we use directed acyclic graphs (DAGs) to iteratively separate the likelihood into smaller parts which can be evaluated and then combined to maximize directed likelihoods. These likelihoods are then utilized to estimate information scores which allows for comparisons of directed graphs.

In the formation of a particular DAG, the entities are ordered and, as a consequence, a potential entity's parents can only come from those entities above it in the ordering. The DAG achieves separation of the full log likelihood by expressing it as the sum of the conditional log likelihoods of each entity given its respective parents. Also, each conditional

distribution can be estimated independently of the others. In [25] these directed acyclic graph properties are proved when the number of parents is not limited, while [26] and [12] reason that these properties should approximately hold when there are restrictions on the number of parents. It should be emphasized that a given DAG has the same likelihood no matter which allowable ordering is used in forming the DAG; this common likelihood is the likelihood of the DAG. For each parent of a entity, a directed edge exists from the parent to the entity in the DAG.

A non-singular sample covariance matrix for an entity and its m parents will exist if and only if $m \leq t - 2$ [24]. Since for undirected graphs, entity i could be a parent of j ($i \rightarrow j$) or j could be a parent of i ($j \rightarrow i$), we restrict the total number of edges into or out from any specific entity to not exceed $t - 2$.

The Bayesian Information Criterion (BIC) is utilized as an inverse measure of the posterior probability of a DAG. For a particular DAG G,

$$BIC_G = -2*\max(\ln(\text{likelihood})) + \ln(t)*(\text{number of parameters}).$$

where the maximum is over the mean and covariance parameter set. Under non-informative priors the posterior probability of a given DAG G is asymptotically proportional to e^{-05BIC_G}, which is larger for smaller values of BIC_G [27]. Thus, a DAG's predicted posterior probability is higher if the graph has better fit to the data, i.e., higher likelihood; while it is lower if the graph has more complexity, i.e., more parameters; thus, parsimony (simplicity) is appropriately rewarded for this model selection criterion. Using BIC in graphical models is common, e.g. backward selection use in *MIM version 3.2* (MIM 3.2 ©David Edwards, 2004) [28], and a non-linear use in [29]. In our setting, the sample size n is often not large, say $n = 6$, so BIC may not closely (inversely) relate to the actual posterior probability. However, BIC is a Laplace (Taylor series based) approximation to the posterior probability. Under our normal setting, BIC is an order $n^{-½}$ [$\Theta(n^{-½})$] approximation [30]. Also, BIC is a model selection criteria which directly balances model fit and model complexity. In Section 8 we further examine BIC approximation by comparing it to a specific *non-informative prior* posterior.

Note that for two different "non-informative" prior distributions, a given DAG's exact posterior distributions differ but the BIC scores and their estimated probabilities will not be altered [31]. Under our normal "linear" setting, for a specific prior distribution, an exact posterior distribution on a DAG can be computed based on the results of [30]. In Section 8, we perform simulations to compare our BIC-based results with these prior-specific results.

For any restricted subset of the k total entities which contain both entities i and j, the population covariance between i and j has as its maximum likelihood estimate the (same) sample covariance between i and j (which is equivalent to the Pearson correlation, r_{ij}, since the data are standardized).

Let R denote the full $k \times k$ sample correlation matrix. The number of parameters of a given DAG is the sum of $2k$ and the number of $\{i,j\}$ sets that are a component of at least one of the conditional likelihoods. The number of parameters may be considerably less than the sum of the numbers of parents over all entities. A DAG's maximum log likelihood is the sum of the maximum of the logs of the conditional likelihood's over the k entities; each of these is easily estimated using properties of conditional multivariate normal distributions [24]. Specifically, if for a DAG G, entity j does not have any parents, then entity j's estimated (maximum) log conditional likelihood is $c - \frac{t}{2}\log(r_{jj}) = c - \frac{t}{2}\log(1) = c$. The constant c in

the above log conditional likelihoods has the value of $-\frac{t}{2}\log(2\pi)$ which does not depend on the graph *G*. Otherwise, if we let *j* denote entity *j*'s parent set for *G*, then entity *j*'s log conditional likelihood is estimated by $c-\frac{t}{2}\log(r_{jj}-R_{j\tilde{j}}R_{\tilde{j}\tilde{j}}R_{\tilde{j}j})$ where the two subscripts on *R* specify the respective row and column(s) determining the submatrix of *R*. Thus the BIC_G score and estimated relative posterior probability can be determined for any DAG. The posterior probability for an undirected graph is the sum of the posterior probabilities for its associated DAGs.

Algorithm 1

The decision process of the Metropolis-Hastings algorithm, searching for best models and high probability edges using Bayesian Information Criterion. The function *random*() returns a random value between 0 and 1, uniformly. For a more liberal *burn-in* stage the −½ on line 4 is replaced with a larger value, −0.1.

1: Generate *New* from the immediate neighbors of *Current*
{If *New* is an improvement over *Current* then unconditionally accept *New*, else probabilistically accept *New*}
2: **if** $\text{BIC}_{New} < \text{BIC}_{Current}$ **then**
3: *Current* ← *New*
4: **else if** random() < $e^{-½(\text{BIC}_{New} - \text{BIC}_{Current})}$ **then**
5: *Current* ← *New*
6: **end if**

When there are no restrictions on the number of parents, then there are Markov likelihood equivalence classes for graphs where two DAGs are in the same class if their associated undirected graphs are the same and if the collider directed edges are the same [32]. (Two directed edges are *collider* directed edges when they terminate in the same node.) It has been hypothesized that when strong directed associations are observed then a cause-effect relation may hold [33]. Our setting is an approximation to the above since we have restrictions on the number of parents. In the cotemporal paradigm, one can observe edges between *siblings* (both sharing a common parent). However, since our procedure assigns high probability to an edge only if it substantially improves predictability (via likelihood) above that of other edges, the chances of sibling edges is lessened. Specifically, the parent/sibling edges might be valued higher than the sibling/sibling edges, and the siblings may add little additional predictability.

Given our restriction on the number of parents, we chose to conduct our algorithm over the commonly used space of directed acyclic graphs. In other studies, some researchers have conducted searches over the space of entity orderings [34], or over Markov equivalence classes [35].

3 Metropolis-Hastings Algorithm

3.1 A Particular Run of the Algorithm

Under our restriction of no more than $t-2$ undirected edges into *k* vertices, there is a total of

$$M_U = \left(\sum_{i=0}^{t-2}\binom{k-1}{i}\right)^k \quad (1)$$

undirected graphs. This number is quite large for even moderate k and t. For example, when $k = 12$ and $t = 6$ then the number of undirected graphs, M_U, exceeds 10^{32}. Similarly, if we let M_D be the number of directed graphs with all vertex degrees bounded by $t - 2$, then $M_D > M_U$. It is not computationally feasible to perform a brute force search of the spaces of directed or undirected graphs.

In order to search for the DAGs and the undirected graphs of highest estimated (BIC based) posterior probability as well as to estimate posterior probabilities for entities' pairwise associations and for graphs, we developed a version of the Metropolis-Hastings algorithm [36], [37]. Our version is for the cotemporal setting with a vertex degree restriction not to exceed $t - 2$. By the theory of aperiodic irreducible Markov chains [36], the longrun fraction of the time that our Metropolis-Hastings algorithm visits a particular DAG should approximate its (BIC based) posterior probability.

Our algorithm begins with an initial DAG, *Current*. Fundamentally, as shown in Algorithm 1, a single replicate in the algorithm makes a single move from the *Current* DAG to a potential replacement DAG, *New*. A single move within the space consists of one edge added or removed in the current DAG's underlying undirected graph as well as a re-ordering of the entities, yielding a re-orientation of the edges. A valid single move retains the restriction of no more than $t - 2$ edges associated with any entity. If the potential replacement DAG, *New*, has higher proportional posterior probability than that of the current DAG *Current*, $e^{-0.5BIC_{New}} > e^{-0.5BIC_{Current}}$, then *New* is accepted as the current DAG, *Current*. Otherwise, the potential replacement DAG, *New*, is accepted as the current DAG with probability $e^{-0.5(BIC_{New}-BIC_{Current})}$. This possible acceptance of a non-improving model makes the Metropolis-Hastings algorithm a robust search process. If the potential DAG, *New*, is not accepted, then the current DAG remains as *Current*.

The algorithm continues for a very large number of replications. For a single run, we use 5 million *burnin* iterations and then 50 million regular iterations. Run data collected include the sorted list of the top 200 observed DAGs based on estimated relative posterior probability, $e^{-.5*BIC_{DAG}}$, over all 50+5 million iterations. If a DAG with high estimated posterior probability is only viewed in the *burn-in*, it still has high estimated posterior probability. The relative frequencies of each of the entity-to-entity directed and undirected edges are also computed over the 50 million regular replicates.

3.2 Multiple Runs of the Algorithm

As with any global searching algorithm, we must be concerned with the effectiveness of the Metropolis-Hastings guided search over the space of DAGs. Even though we are using continuous data, the small number of time points means that the landscape of the space has the potential for many models corresponding to local probability maxima. To better insure the success of the algorithm in visiting virtually all maxima of large probability, ten independent runs are used. These are combined as discussed in Section 3.3.

The 10 separate runs of our algorithm are based upon 5 different strategies for determining the initial undirected graphs, each coupled to 2 different *burn-in* acceptance criteria. Specifically, the 5 different strategies for the initial undirected graphs are: (a) initial selected undirected edges corresponding to the highest absolute Pearson correlations (r's) between the entities, (b) same as *a* but replacing up to 10% of the original selected edges with edges corresponding to lower absolute r, (c) same as *a* but randomly replacing up to 20% of the original selected edges, (d) same as *a* but randomly replacing up to 50% of the original selected edges and (e) initial selection of edges completely at random. These give a diverse array of starting points. The 2 different acceptance criteria for the *burn-in* period are: (1) the regular acceptance criterion (as discussed in Section 3.1) and (2) a more liberal acceptance

criterion, based on a larger exponent, −0.1, than the −½ shown on line 4 in Algorithm 1, this exponent is reset to −½ at the end of the *burn-in* period. The more liberal criterion gives more wide-spread movement during the *burn-in* period.

3.3 Combining the Runs and Probability Estimates

The 10 runs are combined, and we obtain relative frequencies of each of the entity-to-entity directed and undirected edges over the 500 million regular replications. Specifically the combined relative frequency for a particular entity-to entity direct edge is simply the sum of the frequencies of those directed graphs containing this directed edge, divided by 500 million. An analogous method produces the combined relative frequencies for an entity-to-entity undirected edge.

Many of the top 200 DAGs from one run will also be in the top 200 DAGs of another run. Ideally, the combined list of distinct top DAGs over the 10 runs will have far less than 10 * 200 = 2000 DAGs. This combined list, *TopD*, is the amalgamation of the ten runs' top 200 DAG lists. The estimated posterior probability of a particular DAG *j* is proportional to $e^{-.5BIC_{DG_j}}$. If the combined top directed graphs list possesses nearly all of the total probability (see Section 3.4) we are able to estimate the probability of a DAG DG_j in the list by

$$\widehat{p}(DG_j) \equiv \frac{e^{0.5BIC_{DG_j}}}{\sum_{DG \in TopD} e^{-0.5BIC_{DG}}} \quad (2)$$

Any undirected graph can have either direction for each of its edges; so, any undirected graph is the disjoint union of its corresponding DAGs. Thus, the probability of any undirected graph is the sum of the probabilities of its corresponding DAGs. Therefore, we form a combined undirected list of top undirected graphs, *TopU*, which consists of those undirected graphs which have at least one corresponding DAG in the combined top directed list. Finally, we initially estimate the probability of any undirected graph UG_k in the top undirected list by

$$\widehat{p}(UG_k) \equiv \frac{\sum_{DG \in X_k} e^{-0.5BIC_{DG}}}{\sum_{DG \in TopD} e^{-0.5BIC_{DG}}} \quad (3)$$

where $DG \in X_k$ if and only if $DG \in TopD$ and the undirected graph induced by *DG* is UG_k.

The probability of a given entity-to-entity directed edge is the sum of the probabilities of all DAGs that have the particular directed edge. The probability of the entity *m* to entity *n* directed edge is initially estimated by

$$\widehat{p}(m \to n) \equiv \frac{\sum_{DG \in Y_{m,n}} e^{-0.5BIC_{DG}}}{\sum_{DG \in TopD} e^{-0.5BIC_{DG}}} \quad (4)$$

where $DG \in Y_{m,n}$ if and only if $DG \in TopD$ and $m \to n$ belongs to *DG*. Similarly the probability of the undirected edge between entity *m* and entity *n* is initially estimated by

$$\hat{p}(m \to n) \equiv \frac{\sum_{DG \in Y_{m,n} \cup Y_{n,m}} e^{-0.5BIC_{DG}}}{\sum_{DG \in TopD} e^{-0.5BIC_{DG}}} = \sum_{UG \in Z_{m,n}} \hat{p}(UG) \quad (5)$$

where $UG \in Z_{m,n}$ if and only if $UG \in TopU$ and $m \leftrightarrow n$ is an edge in UG.

3.4 Estimated fraction of the probability space

The estimated probabilities of the top directed graphs tend to diminish rapidly, often in an approximately exponential fashion, as one goes down the top directed list. The lower ranked directed graphs in the list often have trivially small estimated proportional probability. It appears that the set of directed graphs that are not in the top directed list might have very small probability and thus the fraction of the total probability that is accounted for by the top directed list would be close to 1.0.

A rough estimate of this fraction is obtained by first examining the natural logarithm of the initially estimated probability of a top DAG versus its index in the *TopD* list, listed in decreasing order of their probabilities. (Fig. 5(c) and 12(c) are examples of this.) If the logarithm is asymptotically linear over the entire domain then a linear regression of the logarithm of the initially estimated probability versus index over the nearly linear region is utilized in order to forecast the proportional probability for the set of DAGs that are not in the top list. The total area under the curve over all M_D directed models, is then roughly approximately by

$$\sum_{i=1}^{N_D} e^{\log(\hat{p}(DG_i))} + \sum_{i=N_D+1}^{M_D} e^{\beta_0 + \beta_1 i}$$

where N_D is the number of DAGs in *TopD*. The adjusted posterior probability $\hat{f_i}$ of DG_i is

$$\frac{\hat{p}(DG_i)}{\sum_{i=1}^{N_D} e^{\log(\hat{p}(DG_i))} + \sum_{i=N_D+1}^{M_D} e^{\hat{\beta}_0 + \hat{\beta}_1 i}} = \frac{\hat{p}(DG_i)}{1 + \sum_{i=N_D+1}^{M_D} e^{\hat{\beta}_0 + \hat{\beta}_1 i}}$$

Since M_D is quite large, the denominator of this probability can not be reliably computed. However,

$$\sum_{i=N_D+1}^{M_D} e^{\hat{\beta}_0 + \hat{\beta}_1 i} \approx \int_{N_D+1}^{M_D+1} e^{\hat{\beta}_0 + \hat{\beta}_1 x} dx = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 (N_D+1)}}{\hat{\beta}_1} (e^{\hat{\beta}_1 (M_D - N_D)} - 1)$$

Since $M_D >> N_D$ and $\hat{\beta_1} < 0$, $e^{\hat{\beta 1}(M_D - N_D)}$ is essentially 0. Now the above expression further reduces to

$$-\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 (N_D+1)}}{\hat{\beta}_1}$$

Hence the adjusted probability for DG_i, $\hat{f_i}$, is approximated as

$$\hat{f}_i \approx \frac{\hat{\beta}_1 \hat{p}(DG_i)}{\hat{\beta}_1 - e^{\hat{\beta}_0 + \hat{\beta}_1 (N_D + 1)}}$$

It follows that $\hat{f}_D \approx \sum_{i=1}^{N_D} \hat{f}_i = \frac{\hat{\beta}_1}{\hat{\beta}_1 - e^{\hat{\beta}_0 + \hat{\beta}_1 (N_D + 1)}}$ estimates the fraction of the total probability which is accounted for by the *combined* top directed list, *TopD*. However, it should be emphasized that even if there is a strong log-linear relationship over indices and $\hat{f}$ is close to 1.0 then there could still be non-trivial probability that was not discovered during the search, especially if a well fitting log curvi-linear trend over indices would suggest it. Even if such non-trivial probability exists, then our restriction of model averaging to the *TopD* list (using the Eq. 2–5) is in the spirit of Occam's window restriction to a top group of models, [31] and [38].

In the biological studies, as will be discussed in Sections 6 and 7, $\hat{f}_D$ is extremely close to 1.0 so no adjustments to the probabilities, as computed in Section 3.2, are recommended. If the fraction is non-trivially below 1.0, then we could adjust each of the estimated probabilities in Section 3.3 by multiplying each by $\hat{f}_D$ in order to obtain adjusted probability estimates.

As another look to see if the probability space has been well explored, we roughly estimate the *fraction* of the total probability that is accounted for by the the top undirected list, $\hat{f}_U$. The estimated probabilities of the top undirected graphs also diminish as the index of list increases. The derivation of this estimate is analogous to the one for the directed list except now N_U is the number of observed top undirected graphs, and the logarithm of the initially estimated probability of an undirected graph is used instead of the one for a directed graph (e.g. Fig. 5(d) and 12(d). In our biological studies, $\hat{f}_U$ like $\hat{f}_D$, is extremely close to 1.0; this provides some complementary evidence that these probability spaces have been well explored by the algorithm.

4 Statistical Diagnostics

One major advantage of our method is the ability to use diagnostics to examine the suitability of our method for a given data set. These diagnostics include evaluation of normality, non-existence of outliers, independent time point contribution, and homogeneity of variance.

First, we examine the original multivariate log normal assumptions by examining q-score plots [24] for each entity's list of $\boldsymbol{t}$ log-transformed values. A *q*-score plot on standardized data shows the expected standard normal values, the *q-scores*, versus the standardized data values. Therefore, these points should fall close to the line $y = x$, and thus have high Pearson correlation *r*, if the (logged) data values follow a normal distribution.

Second, we examine the residuals from the predictions with our *best*, highest posterior probability, DAG. Separately for each entity, we examine the residual values over time after adjusting for associated entities' values. Each entity is checked for compliance with the following: a.) that the magnitudes of the standardized residuals are not extreme, e.g. do not exceed 3.0 in absolute value, indicating that outliers do not exist, b.) that the plot of the residuals resemble white noise (random independent scatter) over time, with non-extreme first-order autocorrelations [40], in order to support independent time point contributions,

and c.) that the q-score plots of the standardized residuals reasonably comply with those of a standard normal distribution.

Third, we produce plots of the above standardized residuals versus the entities' predicted values. We look for close to white noise over the predicted values to check for homogeneity of variances within entities. The lack of increasing or decreasing spread about zero would suggest such homogeneity. If all of these above diagnostics are met, then the method has strong utility for the data set.

5 Biological Data

Two different sets of time course data are used to illustrate this algorithm and to evaluate the effectiveness of the algorithm across different modalities. These data sets were chosen to represent diversity of data types, protein modification (representing a signal transduction network) and gene expression (representing a transcriptional regulatory network), and a diversity of time frames, from minutes to twenty-four hours. Both of these data sets have a small number of time points in relationship to the number of proteins or genes.

The signal transduction data was originally collected in the Loeser laboratory [41]. These data were collected in chondrocytes, using Western blots to monitor the phosphorylation of proteins following stimulation of the chondrocytes with *IGF-1*, insulin-like growth factor. Phosphorylation of isoforms and, in several cases, multiple phosphorylation sites was observed. The data set of 11 protein measurements was sampled at 6 times points. The protein names, isoforms and phosphorylation sites are: *Src Homology Domain* isoforms: *Shc p46 Y317*, *Shc p52 Y317*, and *Shc p66 Y317*; *Akt* phosphorylation sites: *Akt T308* and *Akt S473; extracellular signal-regulated kinase* isoforms: *Erk p44 Y202/Y204* and *Erk p42 Y202/Y204*; *p70 S6 Kinase:* phosphorylation sites (*p70 S6K T389* and *p70 S6K T421/S424*); *Forkhead*: *Fkhd S256*; and, *glycogen synthase kinase 3: GSK 3b S9*. Measurements were taken at 0, 5, 10, 15, 30 and 60 minutes.

Based on the work of [41]–[43] some of the components of the signaling network from IGF-1 stimulation of chondrocytes are known. In particular, intensive studies of the *MAPK/ERK* and PI3 kinase pathways yield strong suggestions for the chondrocyte signaling pathway network. The literature-based model shown in Fig. 1 was established by searching the Signal transduction Knowledge Environment Connection Maps (STKE, http://stke.sciencemag.org) [44]. There were four relevant pathways, namely insulin signaling [45], integrin signaling [46], fibroblast growth factor receptor [47], and epidermal growth factor receptor [48]. A composite signaling network extracted from these literature sources is shown in Fig. 1.

Gene expression in dendritic cells was observed using microarrays in the Hiltbold laboratory (Hiltbold et al., submitted). Dendritic cells are a key regulator of the human immune response. Gene expression was observed during the maturation of dendritic cells following stimulation with *poly I:C*, polyinosinic-polycytidic acid sodium salt, a mimic of viral infection. The expression of **12** genes was extracted from the microarrays at **5** time points. Expression of the following genes was observed: *Interferon beta (Ifn-beta), Interferon alphas (Ifn-alphas)* (represented by four genes α2, α 4, α5, and α7), *2'5'OAS, IFI35, IP10, IRF-1, IRF-7, IRF-8, PSMB8*. Dendritic cells were harvested and gene expression measured on microarrays at 0, 1, 3, 16, 12 and 24 hours. Gene expression measurements were normalized and analyzed in a standard fashion using the Affymetrix QC Toolbox (www.affymetrix.com). Data reported are the log ratio compared to time 0, yielding 5 measurements that were input into the algorithm.

The literature-based transcriptional regulatory network underlying dendritic cell maturation is shown in Fig. 2. This model was compiled from several sources: two literature reviews [49] and [39] and from the pathways generated through the use of Ingenuity Pathways Analysis (Ingenuity®Systems, www.ingenuity.com). Review of the literature indicates that dendritic cell maturation is divided into early, middle and late stages, represented in Fig. 2. Sibling relationships should exist between genes expressed in the early stage; likewise, genes expressed in the middle stage (or the late stage) should be related as siblings. Genes for interferons alpha and beta are expressed in the early and late stages. All other genes are expressed in the middle stage of dendritic cell maturation.

6 Chondrocyte Signaling Models

Recall that the chondrocyte data set consists of 11 protein modifications measured at 6 time points. For easier comparison to the dendritic cell gene study with 5 utilized time points, we restrict the number of edges into any of the proteins to be at most $5 – 2 = 3$. As discussed in Section 3.2, we conduct 10 independent analyses of these data; each run consisted of 5,000,000 steps during initialization, and 50,000,000 regular steps. The burn-in and regular stages for the 10 runs were discussed earlier in Section 3.2.

Relative edge frequencies, the fractions of times the individual edges occurred in models, are computed across the combined 500,000,000 regular examinations of the search space. These directed and undirected edge frequencies are shown in Fig. 3(a) and 3(b). These edge frequency graphs are not subject to the vertex degree restriction since the edges here reflect edge frequencies over all 500,000,000 steps.

Combining the 10 sets of top 200 directed acyclic models from the chondrocyte protein data set yields a composite set of 269 DAG models, considerably less than the maximum possible 2,000. Fig. 5(a) and Fig. 5(c) show the relative posterior probabilities of these top directed acyclic models, and the logarithm of this distribution, respectively. From these 269 directed models, 57 undirected models were found. Fig. 5(b) and Fig. 5(d) show the relative posterior probabilities of these top undirected graphs, and the logarithm of this distribution, respectively. Using methods discussed in Section 3.4, our estimates of the total probability visited are $\hat{f_d} = 0.999849$ and $\hat{f_u} = 0.999808$, with left-hand endpoints of 100 and 5, respectively. Thus, although a small fraction of the model space was examined, we roughly estimate that nearly all of the probability associated with that space has been examined.

From the 269 top directed models and 57 top undirected models and using Equations (4) and (5), directed and undirected edge probabilities are estimated and shown in Fig. 3(c) and 3(d), respectively. These edge probabilities support many of the relationships in the literature model shown in Fig. 1. The high probability red and green directed edges from the *Shc* isoforms to various other downstream proteins, coincide with the literature model. Also, *Akt* (*T308*) has a high posterior probability of being a predictor (parent) of *Fkhd (S256)*. Also, there was high posterior probability of associations for both the sibling relation of *GSK 3b (S9)* with *Fkhd (S256)* and for the common descendant relation of *p70S6k (T389)* with *GSK 3b (S9)*. Figure 3(c) suggests there may be some cyclic association between *Shc p66 (Y317)* and *Akt (S473)* even though the model is based on acyclic graphs.

Theoretically, in the long run the directed edge probabilities seen in Fig. 3(c) and 3(d) should agree with the directed frequencies as shown in Fig. 3(a) and 3(b). However, the difference in the edge frequencies and probabilities underscores the difficulty of working with the sparse data, and demonstrate that the edge frequencies can poorly predict probabilities even after a very large number of replications.

The top overall directed acyclic and undirected models are shown in Fig. 4(a) and (b), respectively. Their respective probabilities, *prob(DG)* = 0.068781 and *prob(UG)* = 0.302922, have been computed using Equations (2) and (3). From Equation (1), the *a priori* probability of any undirected model is on the order of 10^{-28}, thus the algorithm has non-trivially identified the Fig. 4(b) model as the most likely undirected model. As seen in Fig. 4(a), proteins *Shc p46 (Y317)* and *Shc p42 (Y317)* have no parents; as a consequence, these proteins will not contribute to all parts of the upcoming diagnostics.

The q-score plot for the standardized, logged data of all the proteins is seen in Fig. 6. The plot has a correlation of 0.9561 over all data values. The figure's caption also contains the correlation values for the individual proteins. Eight of the eleven proteins' original (logged) data were very good matches to normal distributions. The other three proteins appear to have small deviations from normality, but they are not large and our procedures should be robust to small deviations from normality, as are the *t*-test and linear regression test [52].

In Fig. 7, the residuals from the best DAG of the chondrocyte data over time are plotted. The residuals are the differences between the data and the predicted values from the best directed graph. Each of the proteins' residual curves do not deviate strongly from random variation about zero, with non-significant first-order autocorrelations, suggesting that the residuals are close to white noise over time; thus the independent time point assumption is reasonable for this data set. Also note that *GSK 3b (S9)* and *p70S6K (T421/S424)* are predicted best by the graph since their residuals were the smallest.

Fig. 8 is a plot of the standardized residuals from the best DAG versus the predicted values. Again, it is close to white noise over the *x*-axis now suggesting homogeneous residual variances over parents' (predictors') levels. Also, there are no outlying observations since all the standardized residuals easily fall in the interval (−2.0, +2.0).

The q-score plot for the standardized residuals from the best DAG is shown in Fig. 9. The near linearity of the plot and the high correlations presented in the figure's caption, suggest that the normality assumption is reasonable. Ten of the eleven proteins' correlations were not significantly different from 1.0.

Based on these diagnostics, our methodology is applicable to the above chondrocyte data set. Approximate normality is very reasonable for most proteins' (logged) data as it is for residuals from the best overall directed model. Also, the residuals resemble random independent scatter over both time and predicted values, suggesting valid regression models. Finally, Fig. 3(c) shows high probabilities for several protein-to-protein directed edges which appear in Fig. 1, the composite chondrocyte literature model.

One major additional benefit of our method is the use of the *TopD* and *TopU* lists and their associated graph's estimated probabilities to approximate the probability of subgraphs. For example, suppose one desired to examine the probability of connectivity between *Akt (S473)*, *Akt (T308)*, and *Fkhd (S256)*. This probability can be directly estimated as a sum of the probabilities of all undirected graphs in *TopU* in which there are at least two edges between the three proteins. For the chondrocyte data, the estimate of this probability is 0.99. Using the proteins *Akt (S473)*, *GSK 3b (S9)* and *Shc p46 (Y317)*, the estimated probability is 0.00. Both of these estimates are consistent with Fig. 4. This idea will be discussed further in Section 8.

The five methods from [5] and [53], based on estimating partial correlation, were also applied to the chondrocyte data set. Of the five methods, only *pls*, based on partial least squares regression, discovered significant edges. Of the 7 significant edges, 5 are present in our models shown in Fig. 3(d) and 4(b). All of these five significant edges are colored red

and blue in Fig. 3(d). The other two edges had the least significant partial correlation values of the 7.

7 Dendritic Cell Maturation Models

The gene expression data set consists of 12 genes at 5 time points during dendritic cell maturation. We again restrict the number of edges into or out from a given gene to be at most 3. As described in Section 3.2, there were 10 independent runs. As before the top 200 graphs of each of these runs were amalgamated.

Relative edge frequencies are computed across the combined 500,000, 000 regular probes conducted by the Metropolis-Hastings algorithm. These frequencies are shown in Fig. 10(a) and 10(b).

Combining the 10 sets of top 200 directed acyclic models from the dendritic set yields a composite set of 211 DAGs. From these 211 directed acyclic models, 21 undirected models are found. Fig. 12 shows plots of these distributions and of the log of the these distributions. The values $f_u = 0.992799$ and $f_{\hat{d}} = 0.999491$, with both left-hand endpoints 1, roughly suggest that the 10 independent runs of the algorithm successfully searched the probability space.

From the 211 top DAG models and 21 top undirected graph models, edge posterior probabilities are estimated and shown in Fig. 10(c) and 10(d). The edge probability information shown in Fig. 10(d) suggests *IRF1* is an important predictor of *IP10*, while *IRF8* is an important predictor of *PSM8*. Fig. 10(d) additionally suggests the existence of strong associations between *IFNβ* and *IFNα4*, between *IFNα2* and *IFNα5*, and between *2'5'OAS* and *IFI35*. All of the above relationships are thought to be sibling relationships in the dendritic cell literature model, Fig. 2. The literature model's parent/child relationships of *IFNα7/IFI35*, and *IFNβ/IP10* also receive high estimated posterior probabilities by our method.

Six models tied for the top DAG, a representative one is shown in Fig. 11(a). It (and each of the other five) has a probability of 0.012621, computed using Equation (2). The top undirected graph is shown in Fig. 11(b); its probability is 0.140293, from Equation (3).

Fig. 13 has the overall q-score plot for the original standardized, logged data. There is some lack of linearity suggesting slightly less tail probability than that of a normal distribution. However, this should not bias the results, especially since all genes' r values do not significantly differ from 1.0.

The residuals from the (representative) best directed model for the dendritic cell data are plotted over time, Fig. 14. There are slightly more trends in this plot than were in the corresponding plot of the chondrocyte data; however, the independent time point assumption approximately holds. A few of the genes, *IFNβ, IFI35* and *2'5'OAS*, were especially well predicted by their *parents* (predictors) as demonstrated by their small absolute residuals. The residuals are the differences between the data and the extracted signals (predicted values). Here, as with the chondrocyte data, the residuals are similar to white noise, as they should be.

Fig. 15 is a plot of the standardized residuals from the best DAG versus predicted values. There is no major deviation from white noise, and all of these residuals are between −2.0 and +2.0 which suggests homogeneity of variance. Also, there are no outlying values that bias the estimates.

The q-score plot for the best directed graph's standardized residuals over all the genes is shown in Fig. 16. The q-scores *r*'s for the genes' residuals are large and not significantly different from 1.0, suggesting normality.

Our diagnostics demonstrate that our methodology is applicable to the dendritic cell data set. All the genes' (logged) data and residuals suggest approximate normality. Except for a high first-order autocorrelation for one gene, *IRF7*, all residuals show random independent scatter over both time and predicted values. The single high autocorrelation might make edges involving *IRF7* to be less likely than they would be otherwise, but it will not have dramatic effects on the overall picture of predicted associations between the genes. Finally, our model gave high probabilities for several (gene-to-gene) sibling edges and for two directed edges that are claimed in the composite literature model (as presented in the discussion of Fig. 10).

The estimated probability of connectivity between *IRF1, IP10* and *PSMB8* is estimated from the *TopU* list and the associated graph probabilities; it is 0.93. Similarly, the estimate of the probability of a connection between *IFNa5, IRF8* and *2'5'OAS* is 0.00. These computed estimates are consistent with Fig. 11. This idea will be discussed further in the next section.

The five methods from [5] and [53] were also applied to the dendritic cell data set. *pls* discovered 8 significant edges, and *ridge* discovered 1 significant edge. Of the 8 significant edges discovered by *pls*, 6 are present in Fig. 10(d), 5 are present in Fig. 11(b), and 2 are in neither of these two figures.

8 Simulations

Several randomly generated data sets for each of two settings and one data set for each of many other instructive settings were utilized to examine the performance of our method. For comparison, several other estimation procedures taken from the literature were applied to the same generated data sets.

The methods discussed in this paper assume that the standardized logged data follow a multivariate *k*-variate (entity) normal distribution with mean, 0, and covari-ance/correlation matrix $\Sigma = \rho$. The data sets used for the statistical simulations involve randomly generating *t* (the number of time points) vectors from such a distribution with a specified ρ. Another important parameter for our estimation method is the maximum vertex degree *m*.

Nine data sets were generated and analyzed from the following settings: $k = 9$, $t = 5$, $m = 2$, ρ a 9×9 block diagonal matrix of three 3×3 blocks each of which is

$$\begin{bmatrix} 1 & 0.9 & 0.81 \\ 0.9 & 1 & 0.9 \\ 0.81 & 0.9 & 1 \end{bmatrix}$$

Another nine data sets were generated and analyzed from $k = 15$, $t = 10$, $m = 3$, and ρ in a similar block diagonal structure with now five of the above 3×3 blocks.

Also, single data sets were generated and analyzed using an analogous form for ρ and the following different combinations of (*k, t,m*): (9, 5, 3), (12, 7, 2), (12, 7, 3), (12, 7, 5), (30, 20, 2), (30, 20, 3), (30, 20, 18), (15, 10, 2), and (15, 10, 8).

For the 23 simulations when *m* was low ($m = 2, 3$) the method's edge probability estimates show consistency with the generating model. In these simulations, the generating correlation matrix had some entries with high (off diagonal) correlations (0.81,0.9), while the rest had

zero correlations. For those entries with high generators, the method's undirected edge probability estimates were consistently higher than those corresponding to the zero generators. For nearly all of the simulations, the simulation's average of the estimates over the high generators' entries was more than three times the average of our estimates from the zero generators. The average of the high generators' undirected probability estimates over the 23 simulations is 0.39, while the corresponding average over the zero generators is 0.08.

However, for simulations when m was large ($m = 5, 8, 18$), the space of allowed edges was expanded to the point where the *TopD* lists from different runs were often extremely different. The total number of DAGs in the overall *TopD* list often exceeded 1,000, and the resulting edge probability matrices were not suggestive of the generated setting.

For all of the above simulated data sets, several published regularized estimates of the partial correlation matrix were also computed. Specifically, the shrinkage, partial least squares, ridge regression, lasso and adaptive lasso methods with their associated selection choices were examined as in [5] and [53]. For nearly all of the simulations, the partial correlation estimates were close to the partial correlations of their respective generators.

As previously mentioned at the end of Sections 6 and 7, one benefit of the method is the use of the amalgamated *TopD* and *TopU* lists, along with the probabilities of the associated graphs, to estimate the probabilities of subgraphs. Recall that in these simulations, the correlation matrices consisted of a block diagonal structure of 3×3 blocks of high ($\rho \geq 0.81$) positive association within blocks, and no ($\rho = 0$) association between blocks. For the 23 simulations where m was kept low ($m = 2,3$), the method did quite well in suggesting 3-way connectivity when it exists and not suggesting it when it does not. Specifically, for each of the simulations, three true triples and seven false triples were randomly chosen, and the probability estimates of a 3-way connection were computed for each. The average of the $3 \times 23 = 96$ probabilities of 3-way connectivity estimates over the true triples was 0.78, while the average of the $7 \times 23 = 161$ probability of 3-way connectivity estimates over the false triples was 0.08. Separately, for each of the 23 individual simulations, the true triple probability average was at least twice that of the false triple average; for most of the simulations, the true triple probability average was more than 20 times larger than the false triple probability average.

The BIC-based posterior probabilities were compared with Raftery et al.'s [54] specific *non-informative prior* posterior likelihoods by utilizing their Bayes factor for a response (entity node) under normal linear regression and their specific set of hyperparameters. Since the Raftery et al.'s likelihood equation is only applicable when there is at least one predictor (parent), it was necessary to derive (using ideas from [55, Theorem 7.6.1, pages 417–419]) the corresponding Bayes factor for a node when there is not a predictor. The product of the Bayes factors over all nodes (with or without parents) of a DAG yields the Bayes factor for the DAG. For both the chondrocyte and dendritic cell data sets and for many of the simulated data sets, the *TopD* DAG's BIC-based probability ratios were compared to their respective specific noninformative prior-based probability ratios. Each DAG's ratio was relative to the corresponding performance of the DAG with the best BIC score. Due to outlying ratio pairs, Spearman's correlation was used as a measure of the ratio's relationship; this correlation was found usually to be at a low positive level, nearly always in the −0.1 to 0.3 range. The BIC-based posterior probability estimates are not greatly aligned with the specific *non-informative prior* posterior probability estimates.

9 Conclusions

The method focuses on the all too common biological modeling situation where the number of time points does not exceed the number of entities. In this situation, the data is quite sparse and the covariance matrix is singular. Multivariate normal theory was employed to find the maximum likelihoods as well as to accurately compute the numbers of parameters for restricted networks. The Bayesian Information Criterion (BIC) scores were used in order to estimate approximate posterior probabilities.

Diagnostics which test the suitability of this paper's method for a particular data set are developed. These diagnostics test the three main assumptions of the method: (log) normality, independent time point contributions, and homogeneity of residual variances. For both of this paper's biological data sets, the diagnostic tests suggested reasonable adherence to the assumptions for nearly all of the entities.

In order to cover the probability space well, our method conducts and amalgamates 10 independent diverse runs of a robust search algorithm; a rough estimate is computed for the method's coverage of the probability space. The paper's results include not only probability estimates for full networks, but also probability estimates for both directed and undirected edges between entities as well as for any subgraphs of interest. These edge probabilities are often considerably different from their frequencies of occurrences in the algorithm.

For both the chondrocyte and dendritic cell data sets, diagnostics suggest that the method gives insight into the network. Also, the method's results match several of the literature based theorized results (see discussions of Fig. 3 and Fig. 10). Importantly, the diagnostics test the applicability of the process for other data sets.

Several simulations examined the utility of this paper's method. When the maximum vertex degree m was kept low the method's probability estimates were often substantially larger for true edges than for false edges. Also, the probability estimates for true 3-way connectivity was much greater than for false 3-way connectivity. The simulations suggest that there is often low positive association between our BIC-based method and a specific non-informative prior-based method. However, the BIC-based method, which rewards for higher likelihood and penalizes higher complexity, helps in detecting associations.

In future work, further examination of the use of specific prior settings is planned. As well, the extension of the core idea of robust search of the DAG probability space in order to estimate probabilities for edges, subnetworks and networks will be applied to more complex experimental designs.

Acknowledgments

The authors thank the reviewers for valuable suggestions that significantly improved this paper.

The work of the authors is supported by NSF-NIGMS Program in Mathematical Biology through a grant, NIH R01-GM075304.

References

1. Lauritzen, SL. Graphical Models. Oxford Clarendon Press; 1996.
2. Chaudhuri S, Drton M, Richardson T. Estimation of a covariance matrix with zeros. Biometrika. Jan; 2007 vol. 94(no. 1):199–216.
3. Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics. 2002; vol. 6(no. 2):287–297. [PubMed: 11847076]

4. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics. Jun; 2003 vol. 34(no. 2):166–276. [PubMed: 12740579]
5. Krämer N, Schäfer J, Boulesteix A-L. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics. 2009; vol. 10(no. 384)
6. Dobra A, Hans C, Jones B, Nevins JR, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; vol. 90:196–212.
7. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology. 2005; vol. 4:32.
8. Schäfer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005; vol. 21:754–764. [PubMed: 15479708]
9. Li H, Gai J. Gradient directed regularization for sparse Gaussian, concentration graphs with applications to inference of genetic networks. Biostatistics. 2008; vol. 7(no. 2):302–317.
10. de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using paritial correlation coefficients. Bioinformatics. 2004; vol. 20(no. 18):3565–3574. [PubMed: 15284096]
11. Magwene PM, Kim J. Estimating genome expression networks using first-order conditional independence. Genome Biology. 2004; vol. 5(no. 12)
12. Wille A, Zimmermann P, Vranova E, Fü rholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, Zit-zler E, Gruissem W, Bühlmann P. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. Genome Biology. 2004; vol. 5
13. Allen EE, Fetrow JS, Daniel LW, Thomas SJ, John DJ. Algebraic dependency models of protein signal transduction networks from time-series data. Journal of Theoretical Biology. Jan; 2006 vol. 238(no. 2):317–330. [PubMed: 16002094]
14. Allen, EE.; Pecorella, A.; Fetrow, J.; John, DJ.; Turkett, W. Reconstructing networks using co-temporal functions. In: Silaghi, M., editor. Proceedings of the 44th Annual Association for Computing Machinery Southeast Conference; ACM; Melbourne, Florida. Mar. 2006 p. 417-422.
15. John, DJ.; Fetrow, JS.; Norris, JL. Metropolis-Hastings algorithm and continuous regression for finding next-state models of protein modification using information scores. In: Yang, JY.; Yang, MQ.; Zhu, MM.; Zhang, Y.; Arabnia, HR.; Deng, Y.; Bourbakis, N., editors. Proceedings of the 7th International Symposium on BioInformatics and BioEngineering; IEEE; Oct. 2007 p. 35-41.
16. Freedman, D.; Pisani, R.; Purves, R. Statistics. 4th ed.. W. W. Norton; 2007.
17. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology. 2000; vol. 7(no. 3):601–620. [PubMed: 11108481]
18. Friedman N, Nachman I, Pe'er D. Learning Bayesian network structures from massive datasets: The sparse candidate algorithm. Proceedings of Uncertainly in Artificial Intelligence. 1999
19. Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C. Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research. Oct.2000 vol. 1:49–75.
20. Wille A, Bühlmann P. Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology. 2006; vol. 5
21. Neter, J.; Wasserman, W.; Kutner, MH. Applied Linear Statistical Models. 2nd ed.. Irwin; 1985.
22. Ma S, Gong Q, Bohnert HJ. An *Arabidopsis* gene network based on the graphical Gaussian model. Genome Research. 2007; vol. 17:1614–1625. [PubMed: 17921353]
23. Schmitt JE, Lenrat RK, Wallace GL, Ordez S, Taylor KN, Kabani N, Greenstein D, Lerch JP, Kendler KS, Neabes MC, Gredd IN. Identification of genetically medrated cortical networks: a multivariate study of pediatric twins and siblings. Cerebral Cortex. 2008; vol. 18(no. 8):1737–1747. [PubMed: 18234689]
24. Johnson, RA.; Wichern, DW. Applied multivariate statistical analysis. Prentice-Hall; 1982.
25. Pearl, J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann; 1988.
26. Segal E, Pe'er D, Regev A, Koller D, Friedman N. Learning module networks. Journal of Machine Learning Research. 2005; vol. 6:557–588.

27. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; vol. 6:461–464.
28. Edwards, D. Introduction to graphical modelling. 2nd ed.. New York: Springer-Verlag; 2000.
29. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 2004; vol. 20:1248–1256.
30. Raftery, AE. Bayesian model selection in social research. In: Marsden, PV., editor. Sociological Methodology. Blackwell: Cambridge, Massachusetts; 1995. p. 111-195.
31. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical Science. 1999; vol. 14(no. 4)
32. Pearl, J.; Verma, TS. A theory of inferred causation; Principles of Knowledge Representation and Reasoning: Proc. Second International Conference; 1991.
33. Spirtes, P.; Glymour, C.; Scheines, R. Causation, prediction, and search. Springer-Verlag; 1993.
34. Friedman N, Koller D. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. Machine Learning. 2003; vol. 50:95–126.
35. Chickering, DM. Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann; 1996. Learning equivalence classes of Bayesian network structures; p. 150-157.
36. Hastings WK. Monte Carlo sampling methods using Markov chains and their applicatons. Biometrika. 1970; vol. 57:97–109.
37. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. The American Statistician. Nov; 1995 vol. 49(no. 4):327–335.
38. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association. 1994; vol. 89:1535–1546.
39. Bonjardim CA. Interferons (Ifns) are key cytokines in both innate and adaptive antiviral immune responses–and viruses counteract Ifn action. Microbes Infect. 2005; vol. 7(no. 3):569–578. [PubMed: 15792636]
40. Mendenhall, W.; Sincich, T. A second course in statistics: regression analysis. Prentice-Hall; 1996.
41. Starkman BG, Cravero JD, Delcarlo M, Loeser RF. IGF-1 stimulation of proteoglycan synthesis by chondrocytes requires activation of the PI 3-kinase pathway but not ERK MAPK. Biochemical Journal. Aug; 2005 vol. 389(no. 3):723–729. [PubMed: 15801908]
42. Butler AA, Yakar S, Gewolb IH, Karas M, Okubo Y, LeRoith D. Insulin-like growth factor-1 receptor signal trans-duction: at the interface between physiology and cell biology. Comparative Biochemistry Physiology-Part B: Biochemistry & Molecular Biology. Sep; 1998 vol. 121(no. 1): 19–26.
43. Baserga R, Hongo A, Rubini M, Prisco M, Valentinis B. The IGF-1 receptor in cell growth. Biochimica et Biophysica Acta. Jun; 1997 vol. 1332(no. 3):105–125.
44. Gough NR. Science's signal transduction knowledge environment: the connections map database. Ann N Y Acad Sci. 2002; vol. 971
45. White MF. Insulin signaling pathway. Sci. STKE (Connections Map). 2007 cmp_12069.
46. Martin KH, Slack JK, Boerner SA, Martin CC, Parsons JT. Integrin signaling pathway. Sci. STKE (Connections Map). 2002 cmp_12069.
47. Schlessinger J. Fibroblast growth factor receptor pathway. Sci. STKE (Connections Map). 2008 cmp_15049.
48. Schlessinger J. Epidermal growth factor receptor pathway. Sci. STKE (Connections Map). 2008 cmp_14987.
49. Takaoka A, Yanai H. Interferon signalling network in innate defence. Cell Microbiol. 2006; vol. 8(no. 6):907–922. [PubMed: 16681834]
50. Filliben JJ. The probability plot correlation coefficient test for normality. Technometrics. Feb; 1975 vol. 17(no. 1):111–117.
51. Snedecor, GW.; Cochran, WG. Statistical Methods. 6th ed.. The Iowa State University Press; 1967.
52. Heeren T, D´ıAgostino R. Robustness of the two independent samples *t*-test when applied to ordinal scaled data. Statistics in Medicine. 2006; vol. 6(no. 1):79–90.

53. Mevik B-H, Wehrens R. The pls package: Principal component and partial least squares regression in R. Journal of Statistical Software. 2007; vol. 18(no. 2):1–24. [Online]. Available: http://www.jstatsoft.org/v18/i02.

54. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association. 1997; vol. 92(no. 437):179–191.

55. DeGroot, MH.; Schervish, MJ. Probability and Statistics. Addison-Wesley; 2002.

Biographies

David J. John Dr. David John received his PhD in mathematics and his MS in computer science from Emory University in 1978. He has been on the faculty at Wake Forest University since 1982, and is currently Professor of Computer Science. For the last six years, Dr. John has actively participated in the Wake Forest University BioNet interdisciplinary research group; the focus of the group is to algorithmically infer interactions between genes/proteins from sparse time series laboratory measurements. His contribution to this work has been in the development and analysis of interaction modeling algorithms.

Jacquelyn S. Fetrow Dr. Jacquelyn S. Fetrow received her PhD in biological chemistry from Pennsylvania State University Medical School in 1986. Since 2003, she has held the position of Reynolds Professor of Computational Biophysics at Wake Forest University, with joint appointments in both the departments of physics and computer science. In 2008, she was appointed to the position of Dean of Wake Forest College. Prior to these appointments, she was on the faculty of the Biology Departments at SUNY Albany, and served as Chief Scientific Officer and Director of GeneFormatics, a biotechnology software company. Dr. Fetrow's research program focuses on understanding the relationship between protein structure, function, and dynamics and biological networks with a long-range goal of understanding disease mechanisms and improving the structure-based drug discovery process.

James L. Norris Dr. James Norris received the PhD degree in statistics from The Florida State University in 1990. since 1989, he has been a member of the Department of Mathematics at Wake Forest University, U.S.A. where he is currently a full professor.

Throughout, his research has focused on developing statistical methods for biological settings. The major emphasis of his recent work has involved modeling biological signaling and gene interaction networks.

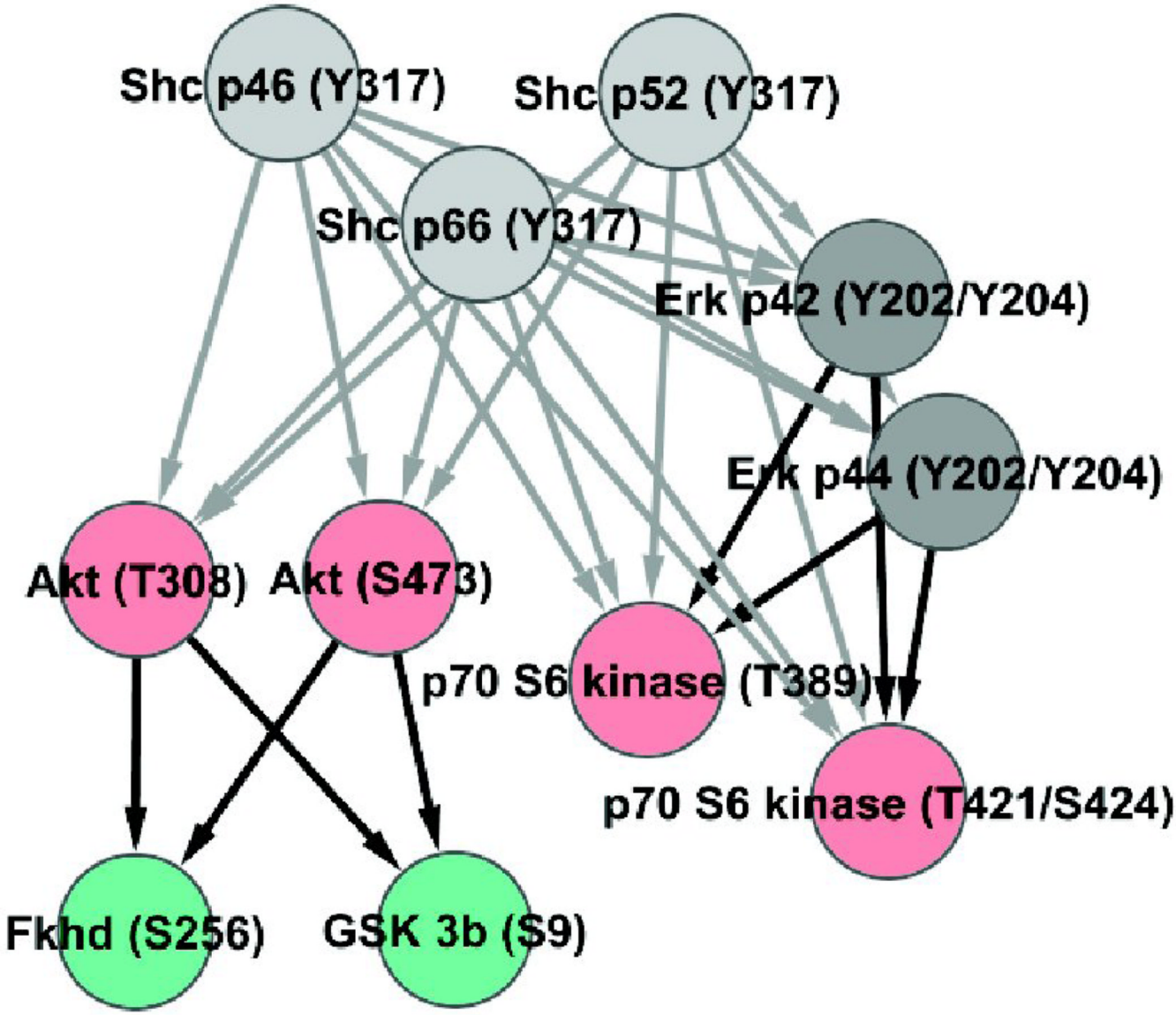

Fig. 1.

Composite signal transduction network extracted from the literature and expected to be observed following IGF-1 stimulation. Phosphorylation in the following proteins was measured: *Shc, Erk, Akt, p70 S6 kinase, Forkhead (Fkhd)*, and *glycogen synthase kinase 3b (GSK 3b). Shc* and *Erk* exist as multiple isoforms and these were observed independently in the experiments: *Shc p46, p52*, and *p66* and *Erk p42* and *p44* (indicated by light and dark gray nodes, respectively). The relationship between these isoforms within the signaling network is not known from the literature. Phosphorylation of *Akt* and *p70 S6 kinase* was measured at two sites each: *Akt T308 and S473; p70 S6 kinase T389* and *T421/S424* (indicated by pink nodes). *Fkhd* and *GSK 3b* were each measured at one phosphorylation site *(S256* and *S9*, respectively). The dark *(direct)* directed edges from the *Erk* phosphorylation sites to the *p70 S6 kinase* states that phosphorylation of *Erk* directly influences phosphorylation of this kinase. Similarly the dark directed edges from the *Akt* phosphorylation sites to *Fkhd (S256)* and *GSK3b (S9)* suggest that phosphorylation of *Akt* directly influences *Fkhd (S256)* and *GSK3b (S9)* phosphorylation. The light *(non-direct)*

directed edges going from the *Shc* isoforms indicates that phosphorylation of *Shc* directly influences phosphorylation of non-measured proteins which influence activation of the measured *Erk*, *p70 S6 kinase*, and *Akt*.

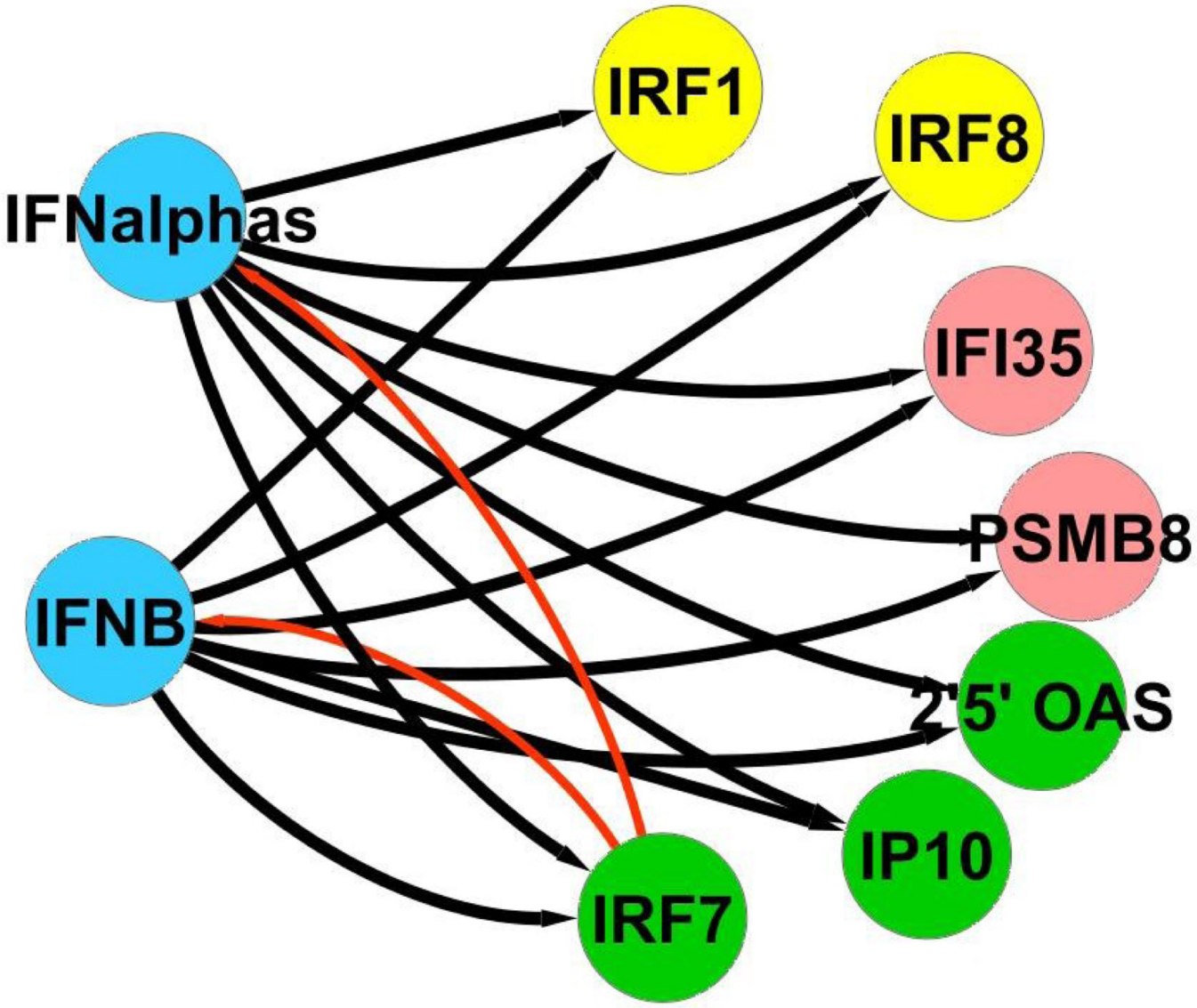

Fig. 2.

Composite literature model of the genes expressed during the early (blue nodes), middle (yellow, pink, and green nodes), and late (blue nodes) stages of dendritic cell maturation. Different colors for the genes in the middle stage represent groups of genes under control of a single transcription factor. Directed edges show gene expressions that should follow in a time-based fashion. *Irf-7*, expressed in the middle stage of maturation, is thought to cause the re-expression of the interferons in the late stage [39]; this is represented by the (red and black) directed edges between the interferons and *Irf-7* that extend in both directions. The four interferon alpha genes *IFNα2*, *IFNα4*, *IFNα5*, and *IFNα7* are represented in the literature model by the single node *IFNαs;* no literature distinction is currently available between the four.

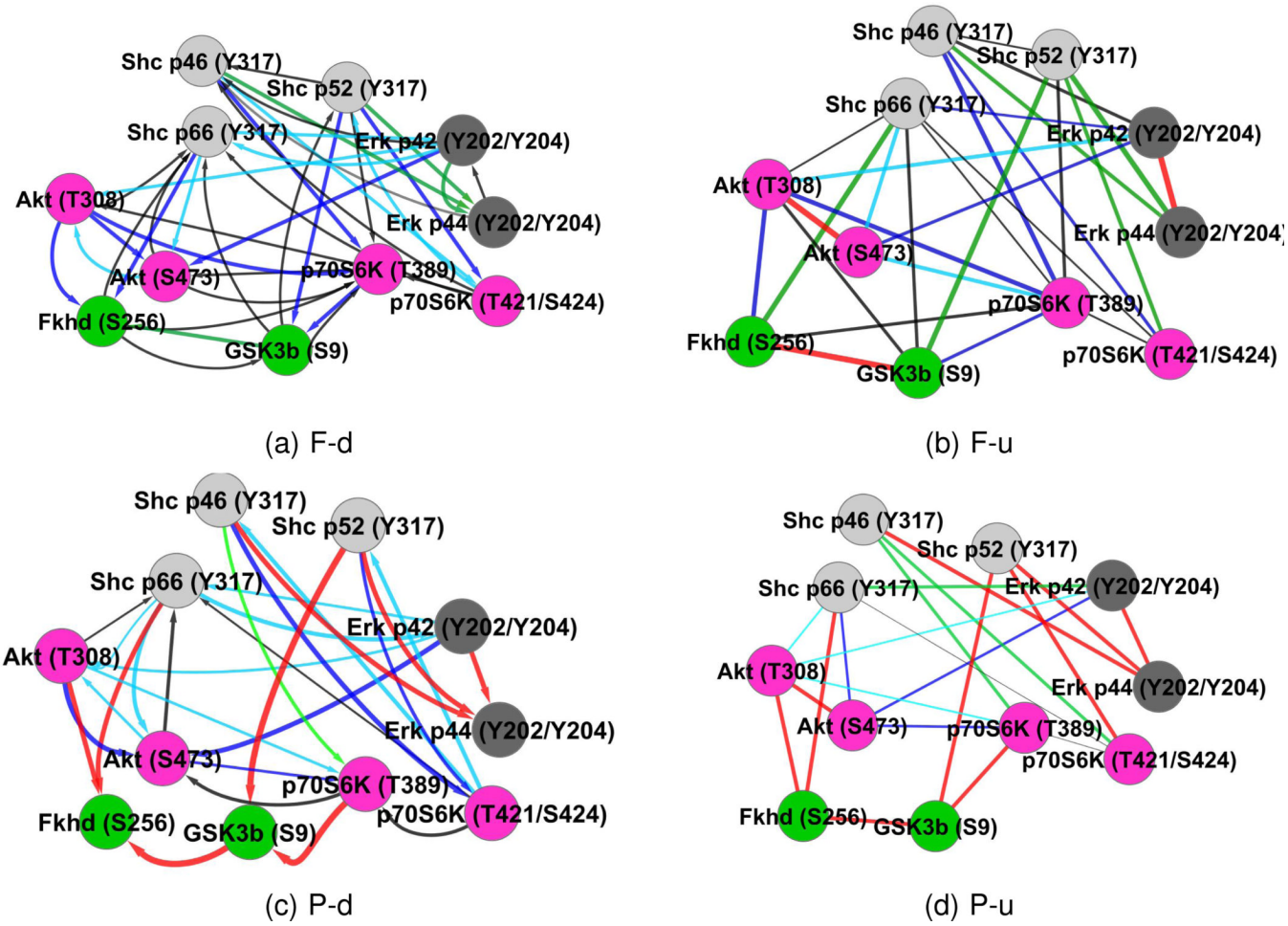

Fig. 3.

Edge frequencies (F) and edge probabilities (P) in directed (d) graphs and undirected (u) graphs for the chondrocyte data. The frequencies (F) are the fractions of time that the individual protein-to-protein directed (d) and undirected (u) edges occurred in models obtained from the combined 500, 000, 000 regular M-H examinations. The probabilities (P) are estimated posterior probabilities for directed edges and undirected edges using Equations 4 and 5, respectively, across the 550, 000, 000 regular and burn-in M-H examinations. The edge frequencies are often considerably different from their corresponding edge probabilities. This difference demonstrates that frequencies can poorly predict probabilities even with a very large number of M-H examinations of the search space. The edge frequency/probability color legend, used here and in following figures is: [0.95,1.0], red; [0.8, 0.95), green; [0.5, 0.8), blue; [0.2, 0.5), light blue; [0.5, 0.2), black; [0.0, 0.05), clear.

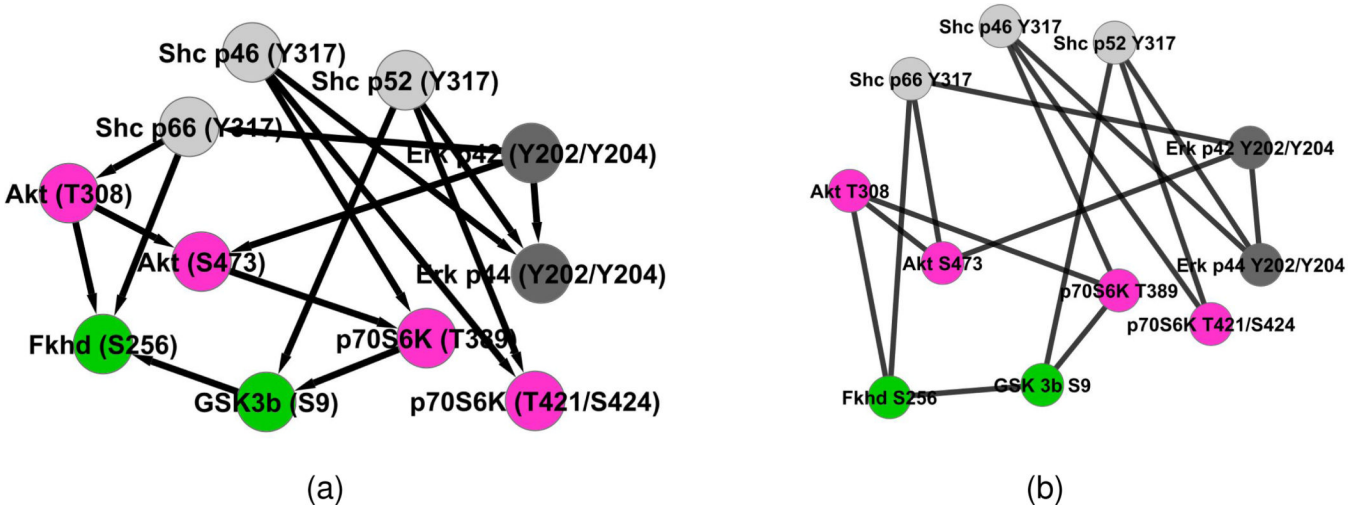

Fig. 4.

Highest posterior probability overall directed and undirected networks for the chondrocyte data. The respective probabilities, $prob(DG_d)$ = 0.068781 and $prob(UG_u)$ = 0.302922, were computed using Equations (2) and (3), respectively, across all 550,000,000 regular and burn-in M-H examinations. The algorithm produced these as the most likely predictions for the overall directed and undirected chondrocyte network models.

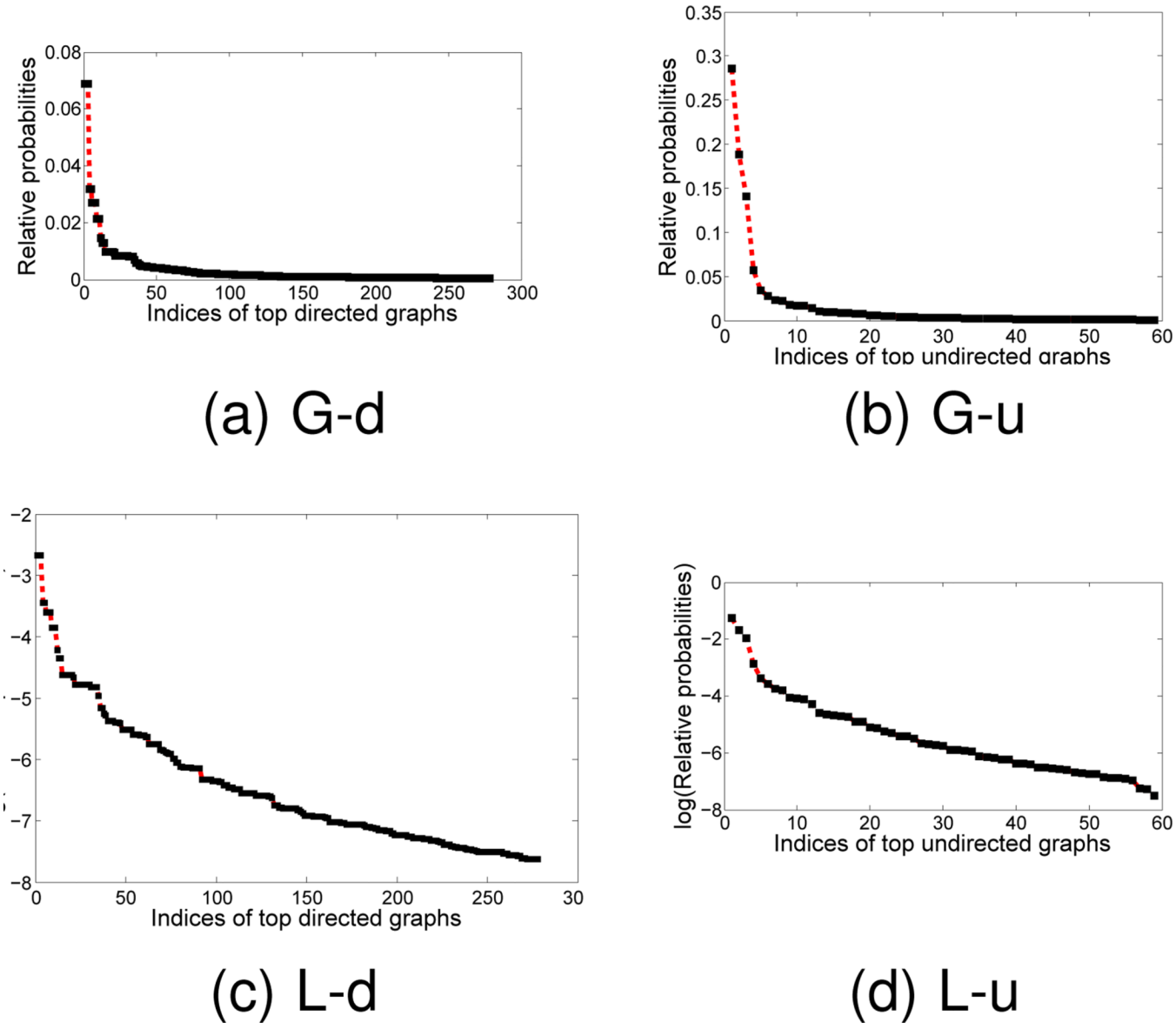

Fig. 5.

Relative graph probabilities (G) and their natural logarithms (L) versus Index for the top directed (d) graphs and top undirected (u) graphs across all 550,000,000 regular and burn-in M-H examinations for the chondrocyte data. The values of $\hat{f_d}$ = 0.999849 and $\hat{f_u}$ = 0.999808 were computed using starting index values of 100 and 5, respectively.

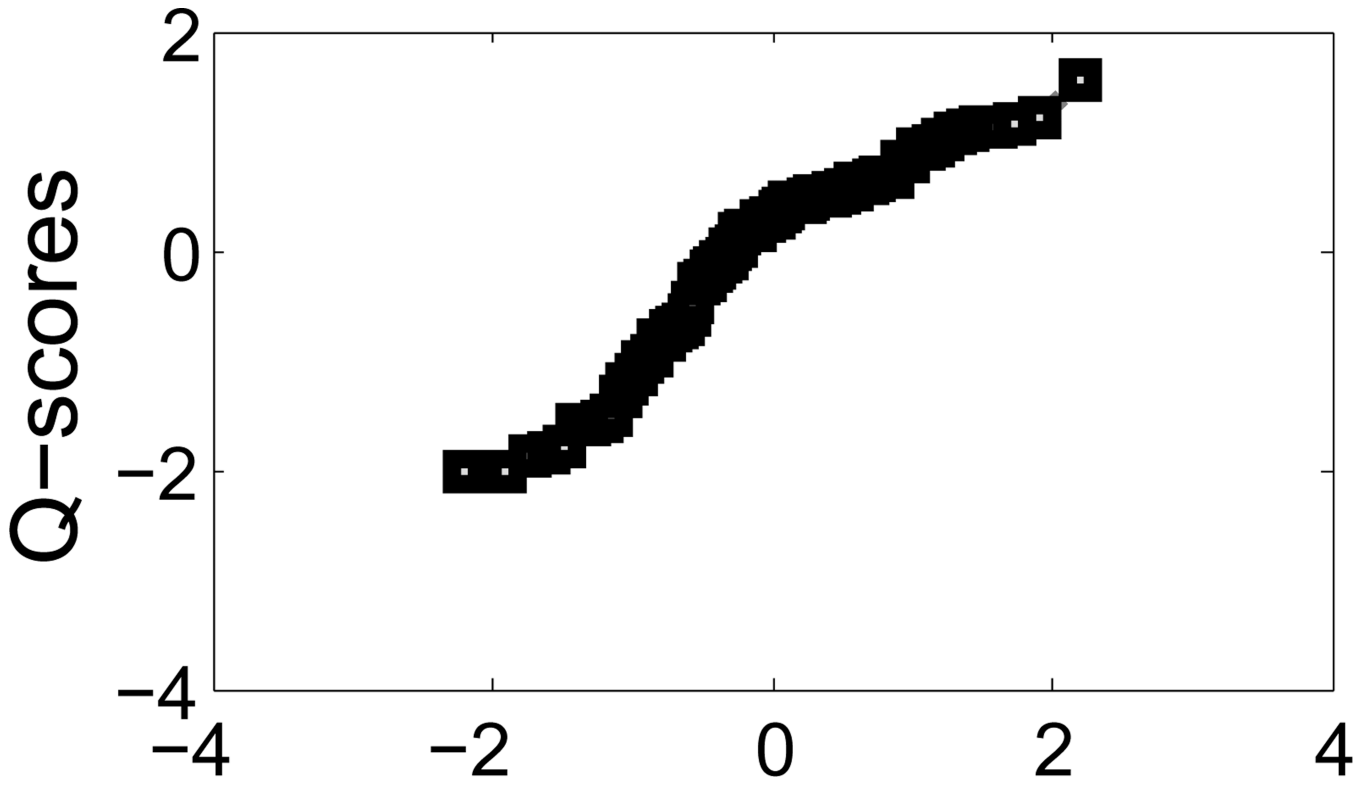

Fig. 6.

Q-score plot for all of the standardized logged chondrocyte data. The individual Pearson correlation, *r*, values for the 11 sets of proteins are (in the order of Fig. 7 legend): 0.7795*, 0.7590*, 0.9875, 0.9648, 0.9522, 0.9556, 0.9058, 0, 8681*, 0.9174, 0.9927, and 0.9762. The *r* value for the entire set is 0.9561. The three starred values are significantly different from 1.0 suggesting some deviation from normality for the original logged data of these three proteins; the other 8 proteins' *r* values are not significantly different from 1.0, suggesting their normality [50].

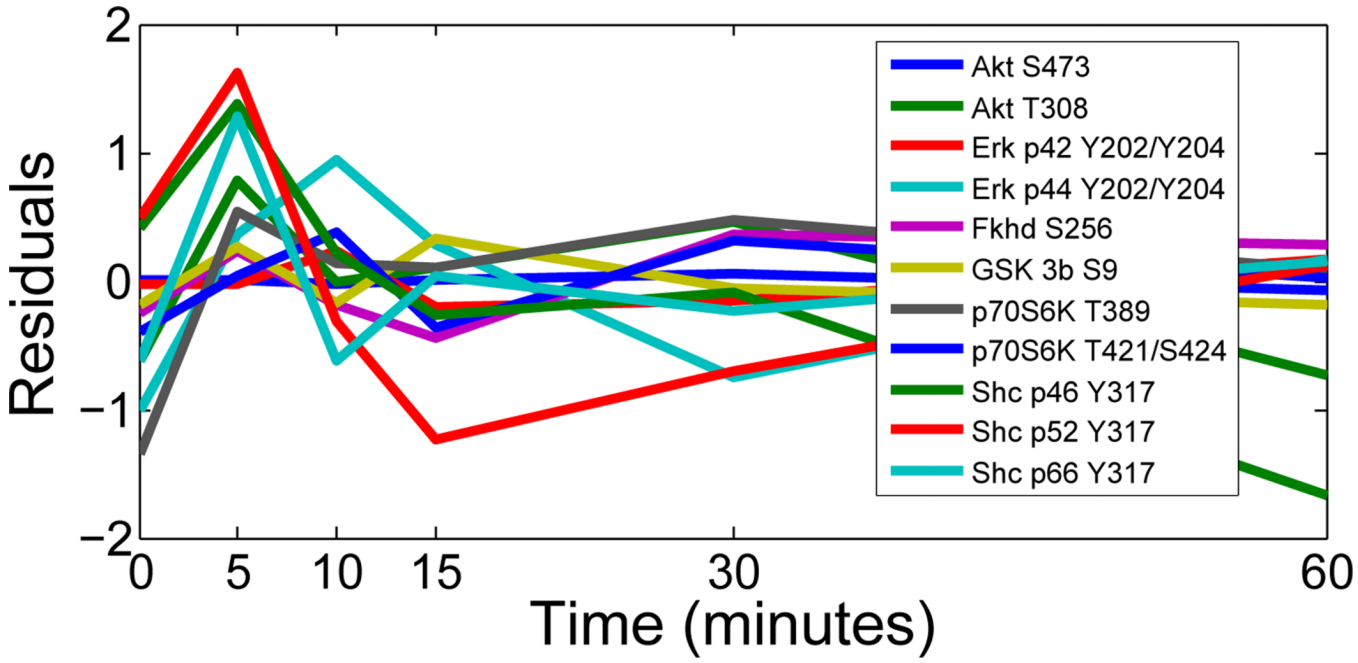

Fig. 7.

The residuals exhibit the differences of the data and the best directed model's predicted values for the chondrocyte data. Each protein's residuals over time do not substantially deviate from white noise with frequent short runs of both positive and negative residuals over time. First-order autocorrelations in order of residuals of the above-listed chondrocyte proteins are: −0.8211, 0.6451, 0.1702, −0.3169, −0.7028, −0.7905, −0.1616, −0.7947, 0.4377, 0.3212, and −0.5981. None of these are significantly different from 0.0 [51], the assumption of random uncorrelated residuals is supported.

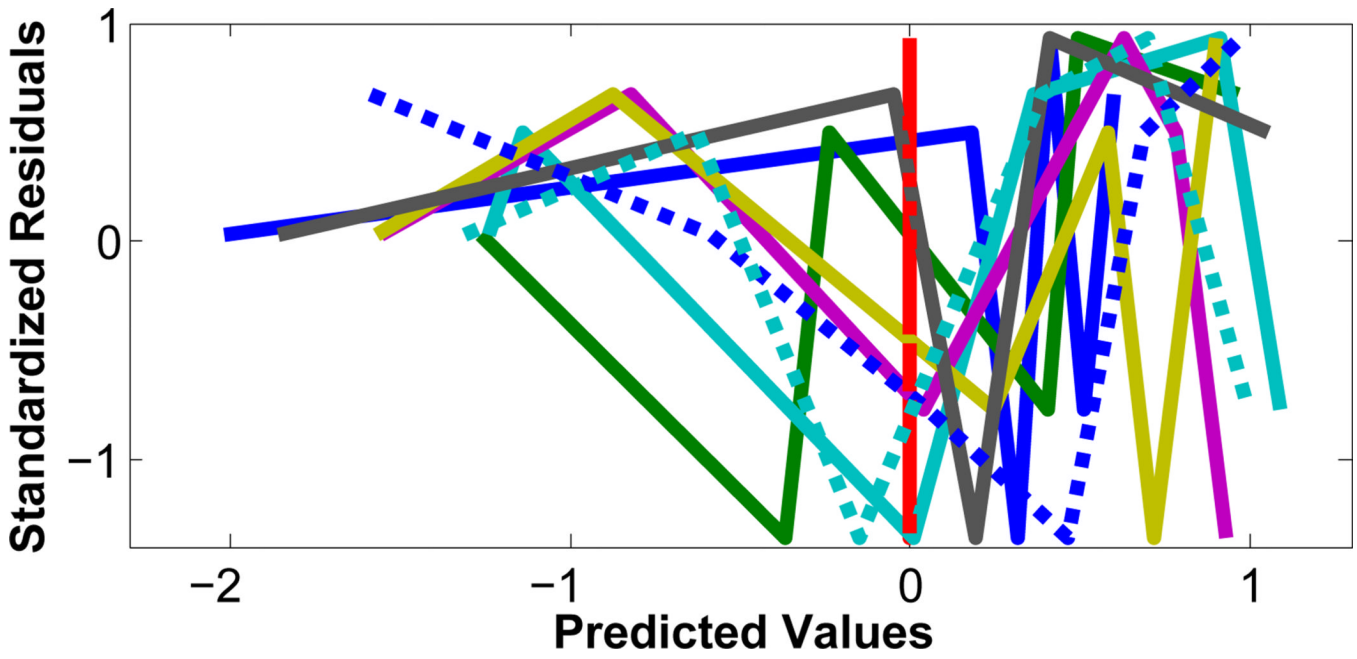

Fig. 8.

The standardized residuals versus the predicted values for the best directed model representing the chondrocyte data are shown (line legend shown in Fig. 7). The vertical lines for *Erk p42 (Y202/Y204)*, *Shc p46 (Y317)* and *Shc p42 (Y317)* reflect the absence of parents, as seen in Fig. 4(a). The lack of any strong trends in the standardized residuals over the predicted levels suggests homogeneous variances. Also, since all standardized residuals are in the interval [−2.0, +2.0], there are no apparent outliers.

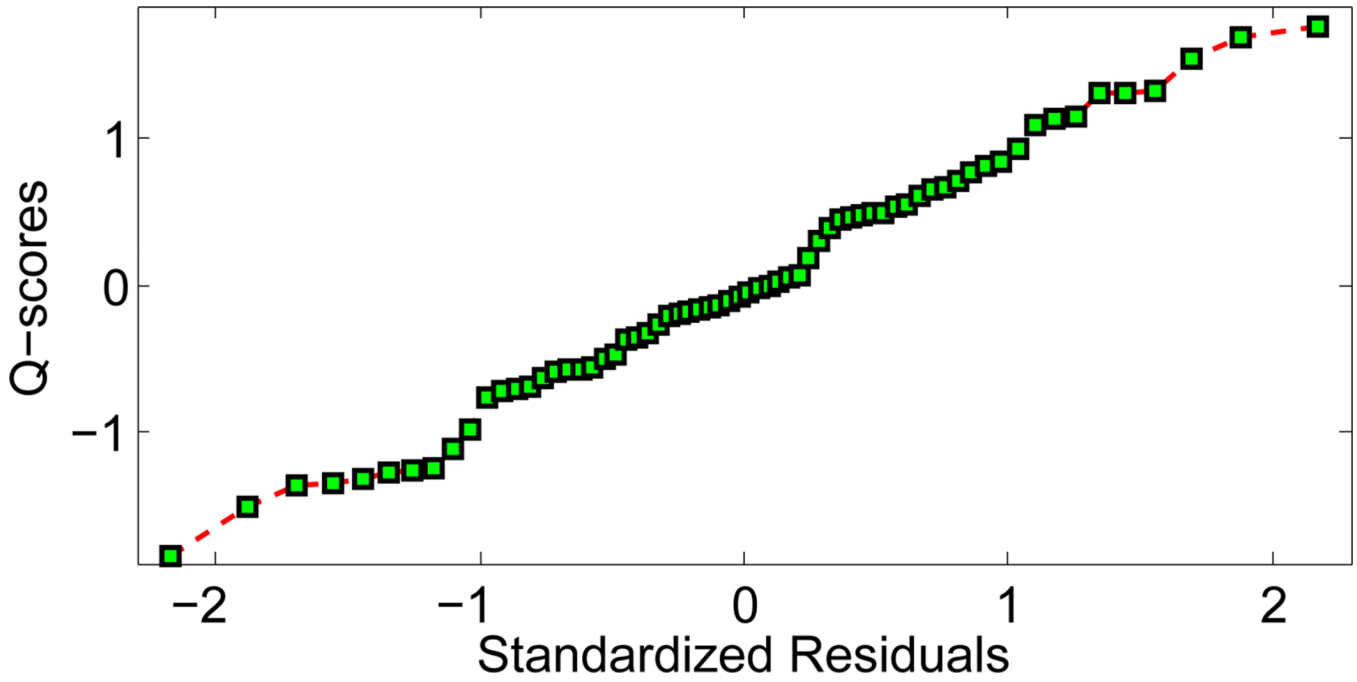

Fig. 9.

The q-score plot of the standardized residuals for all proteins from the best directed chondrocyte graph. The Pearson correlation r over all the proteins is 0.9961. The individual *r* values for the 11 proteins are: 0.9679, 0.9627, 0.9875, 0.9342, 0.9642, 0.9758, 0.8414*, 0.9880, 0.9174, 0.9927, and 0.9461. The one starred value is statistically significantly different from 1.0 (but not highly so) suggesting some deviation from normality for the residual of *p70S6K (T389)* from the best directed model; however, the normality assumption is reasonable for the other 10 proteins. Overall, the residuals are close to being normally distributed.

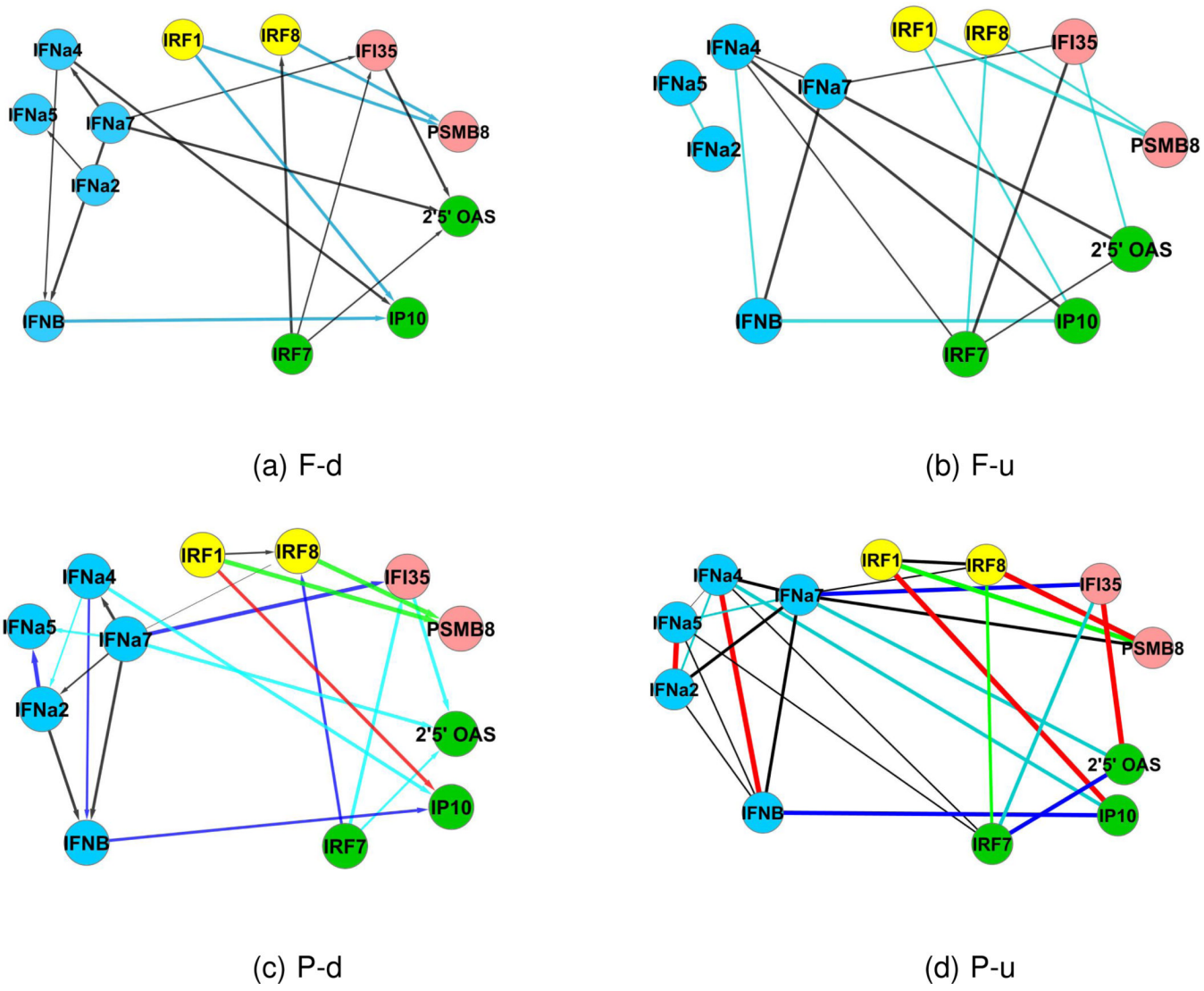

Fig. 10.

Edge frequencies (F) and edge probabilities (P) in directed (d) graphs and undirected (u) graphs for the dendritic model space. The frequencies (F) are the fractions of time that the individual protein-to-protein directed (d) and undirected (u) edges occurred in dendritic models obtained from the combined 500, 000, 000 regular M-H examinations. The probabilities (P) are estimated posterior probabilities for directed edges and undirected dendritic edges computed using Equations 2 and 3, respectively, across the 550,000,000 regular and burn-in M-H examinations. The edge colorings are as in Fig. 3. The edge frequencies are often considerably different from their corresponding edge probabilities. This difference demonstrates that frequencies can poorly predict probability even with a very large number of M-H examinations of the search space.

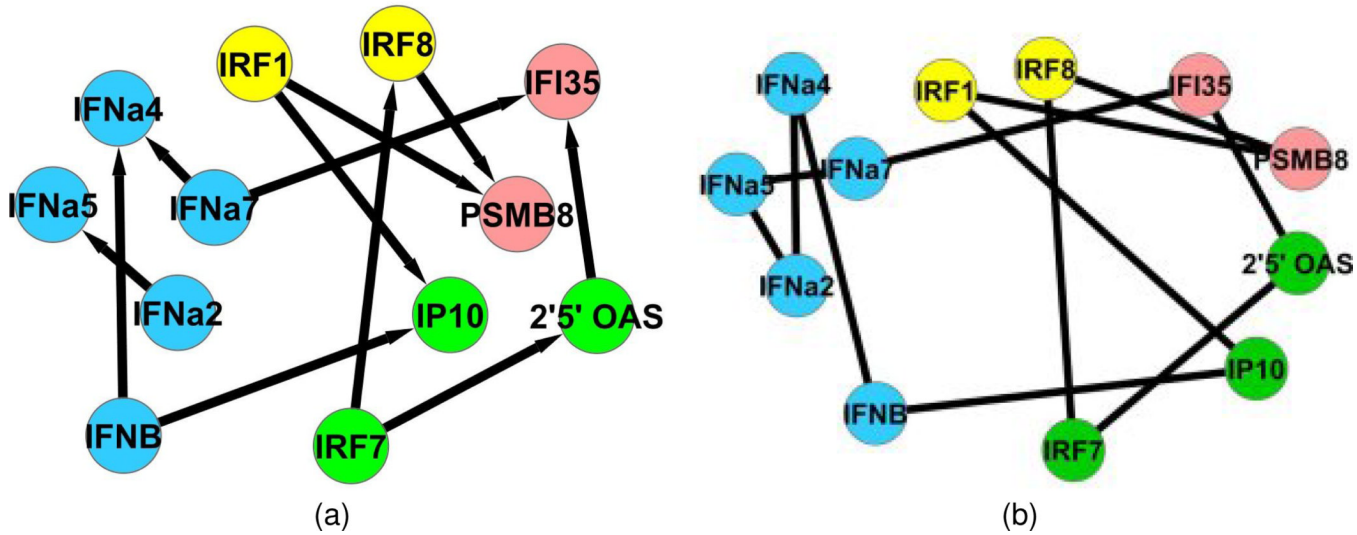

Fig. 11.

The top overall network undirected graph (b) and a representative top directed graph (a) for the dendritic cell data after 10 runs searched through 500 million graphs. The estimated probabilities of these top models are $prob(DG_d)$ = 0.012621 and $prob(UG_u)$ = 0.140293. These predict overall directed and undirected network models.

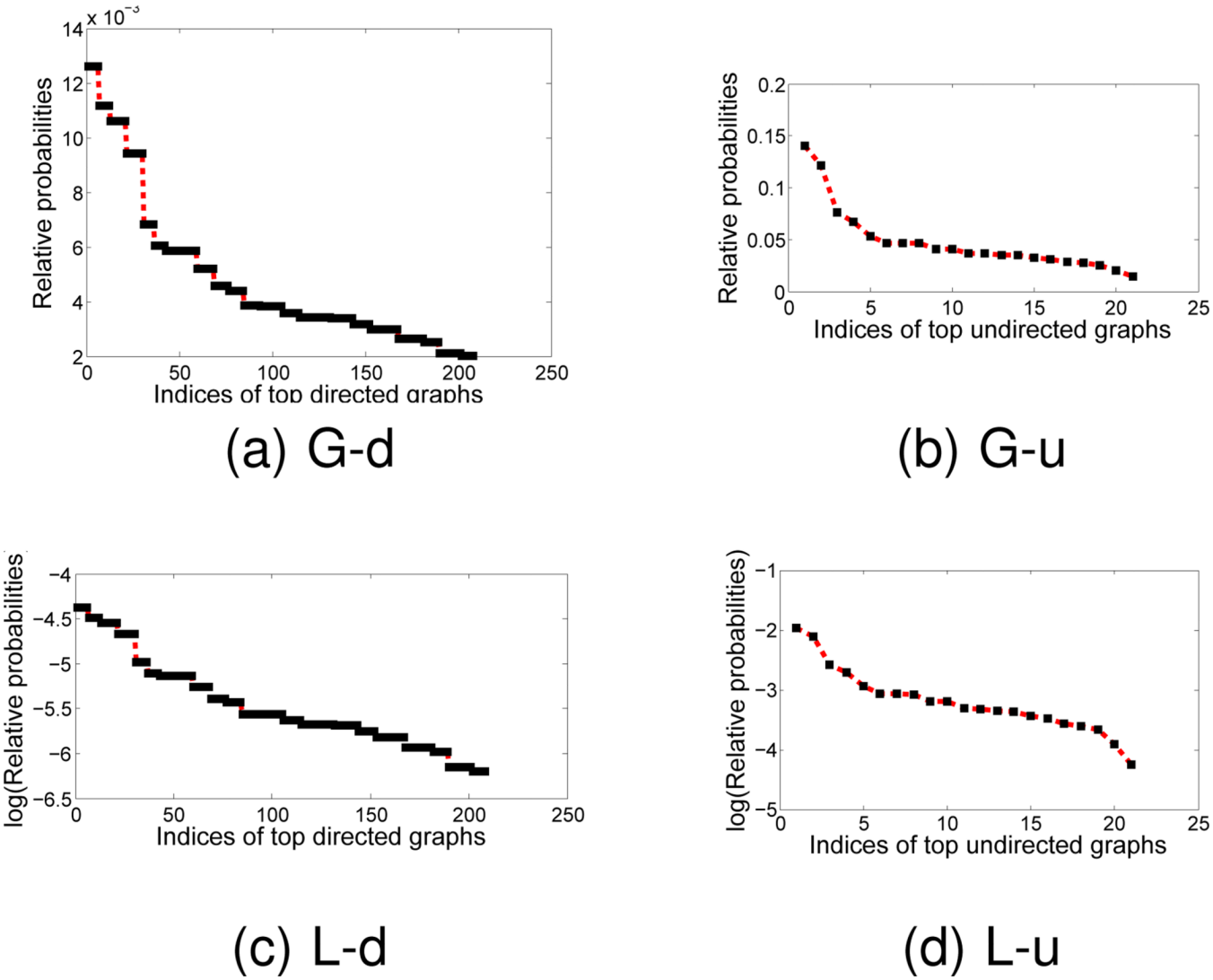

Fig. 12.

Relative graph probabilities (G) and their natural logarithms (L) versus Index for the top directed (d) graphs and top undirected (u) graphs across all 550,000,000 regular and burn-in M-H examinations for the dendritic cell data. The values of $\hat{f_d}$ = 0.999491 and $\hat{f_u}$ = 0.992799 were computed using starting index values of 1 and 1, respectively. The linearity of L-u is not optimal for high indices but if we adjust for their concave nature the estimate for $\hat{f_u}$ would be even closer to 1.000.

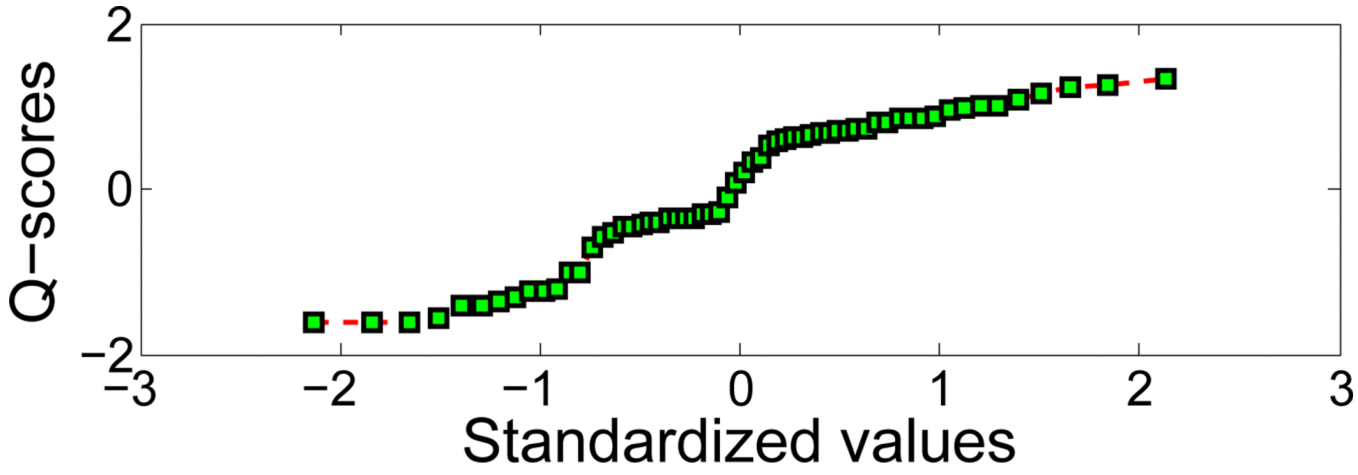

Fig. 13.

Q-score plot for all of the original standardized logged dendritic cell data. The Pearson correlation *r* value for the entire set is 0.9622. The corresponding individual r values for the 12 sets of genes (in the order of the gene legend in Fig. 14) are: 0.9369, 0.9801, 0.9942, 0.9864, 0.9450, 0.9493, 0.8913, 0.9455, 0.8924, 0.9690, 0.9329, and 0.9573. None of these *r* values are significantly different from 1.0, thus suggesting normality for the original logged data [50].

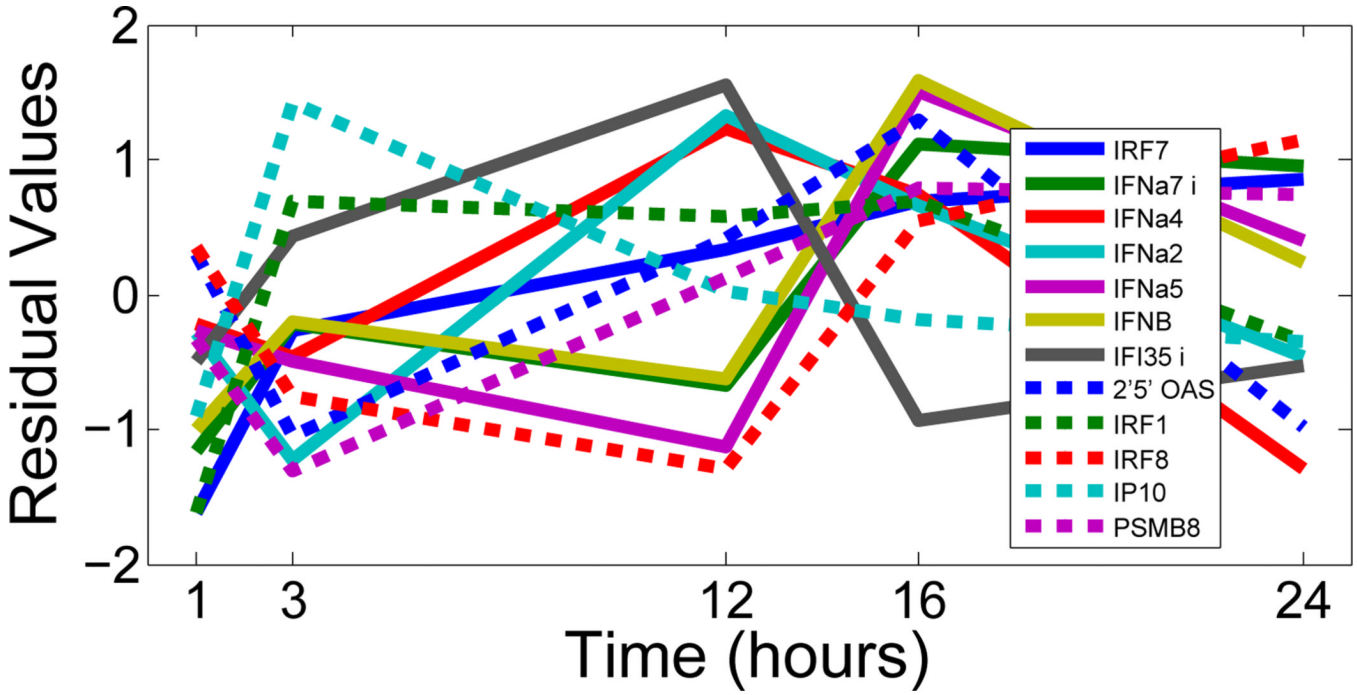

Fig. 14.

Dendritic cell residuals from the best directed model, Fig. 11(b), versus time, separately for each gene. First-order autocorrelations in order of the above listed genes are 0.9984*, 0.3826, −0.2126, −0.1950, −0.0997, −0.0734, −0.1563, −0.4261, −0.4078, 0.1780, −0.5361, and 0.3993; only the *IRF7* gene's first-order autocorrelation is significantly different from 0.0, and white noise seems reasonable for all genes except possibly *IRF7* [51].

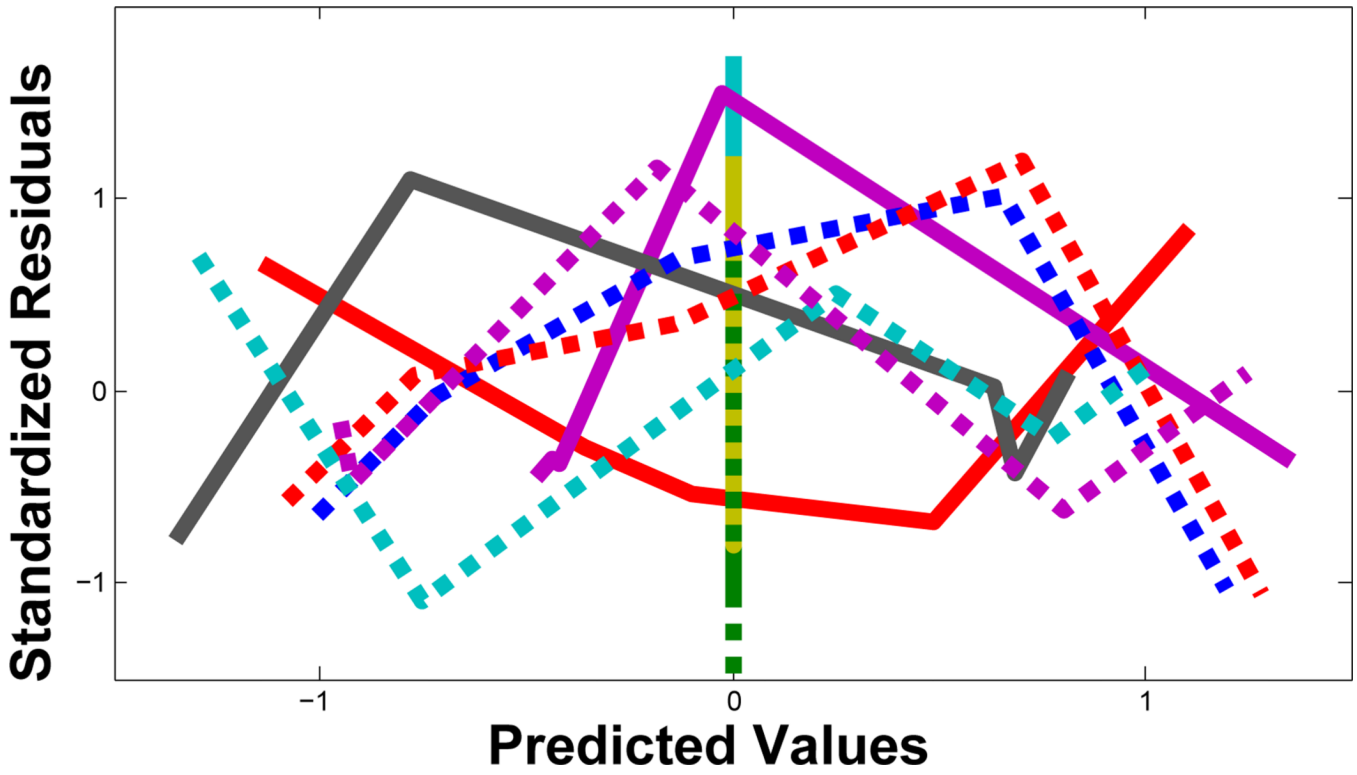

Fig. 15.

The dendritic cell standardized residuals versus the predicted values for the best directed model (line legend shown in Fig. 14). The vertical lines for *IFNα2*, *IFNα7*, *IFNβ*, *IRF7* and *IRF8* reflect the absence of parents (predictors), as seen in Fig. 11(a). The lack of any strong trends in the standardized residuals over the predicted levels suggests homogeneous variances. Also, since all standardized residuals are in the interval [−2.0, +2.0], there are no apparent outliers.

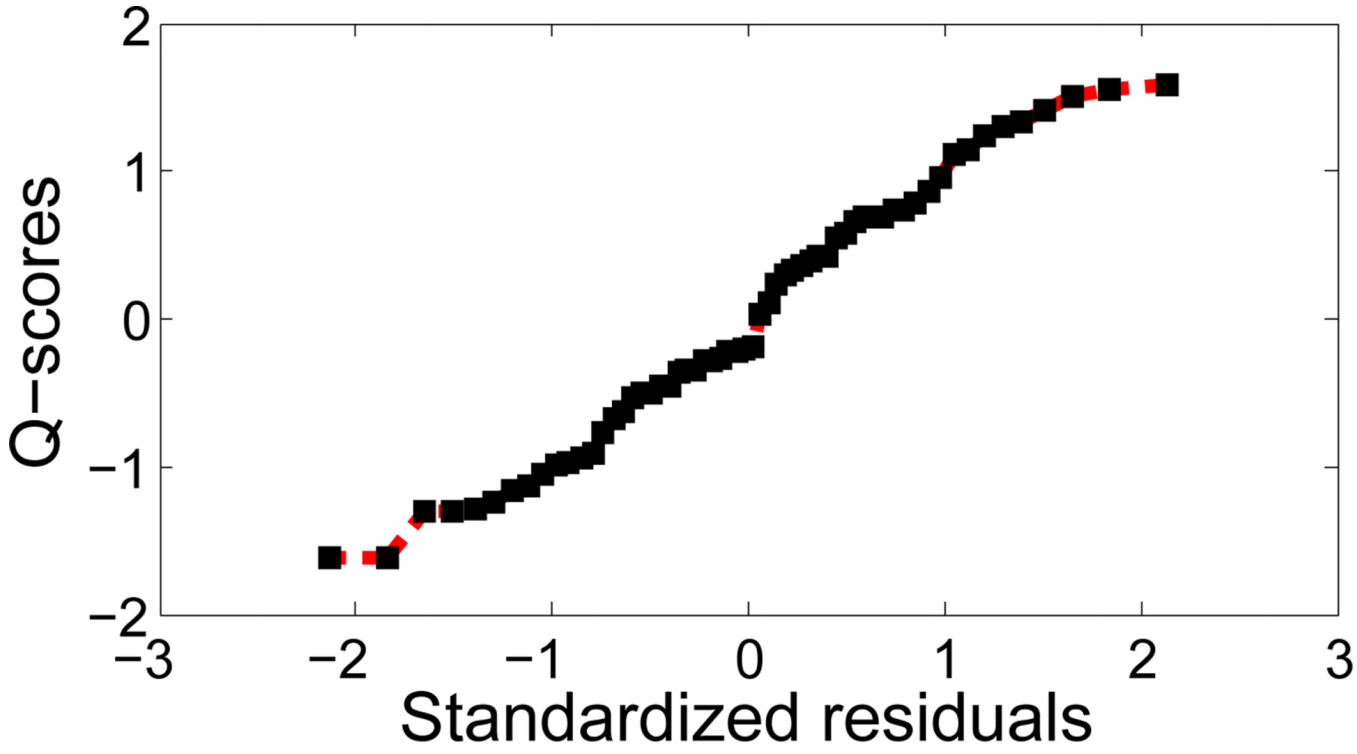

Fig. 16.

The q-score plot of the standardized residuals for all genes from the best directed dendritic model. The Pearson correlation *r* over all the genes is 0.99026. The individual *r* values for the 12 sets of genes are 0.93692, 0.96887, 0.98886, 0.98644, 0.97715, 0.95892, 0.94196, 0.95744, 0.89238, 0.97731, 0.93216, and 0.9599. None of these *r* values are significantly different from 1.0, thus suggesting normality for all genes residuals from the best directed model.