International Journal of Genomics and Data Mining

Fu LM. Int J Genom Data Min 01: 112.

Research Article

DOI:10.29011/2577-0616.000112

A New Model for Analysis of Upstream Mechanisms in a Biological Network Based on Genomics Data

Li M Fu*

Department of Biomedical Engineering, AHMC Healthcare, Los Angeles, CA, USA

***Corresponding author:** Li M Fu, Department of Biomedical Engineering, AHMC Healthcare, Los Angeles, CA 91801, USA. Tel: +19493314196, Email: lifu.usa@gmail.com

Citation: Fu LM (2017) A New Model for Analysis of Upstream Mechanisms in a Biological Network Based on Genomics Data. Int J Genom Data Min 01: 112. DOI: 10.29011/2577-0616.000112

Received Date: 24 October, 2017; **Accepted Date:** 07 November, 2017; **Published Date:** 14 November, 2017

Abstract

It is a common to describe a biological system as a network where a node represents a biological entity, such as a concept, event, or process, and an edge represents a relation. A biological network can model and simulate a biological phenomenon in a practical problem domain. To capture a dynamic behavior, it is important to understand which part of the network serves to control or activate the rest of the network. Here the subnetwork assuming such a role is called the upstream network as distinguished from downstream networks. It should be clear that the upstream mechanisms must occur in time before the downstream mechanisms, hence a specific relationship known as temporal precedence. In cross-sectional designs, it would be difficult to determine temporal precedence between two variables because of the lack of time-controlled experiments. To cope with this problem, we developed a new model with mathematical analysis for determining upstream mechanisms in a biological network based on genomics data. This model has been validated against the gold standard based on time-series data.

Keywords: Biology; Gene Expression; Genomics; Upstream Mechanism

Introduction

A biological system is represented by a network referred to as a biological network, such as a metabolic network, a gene-regulatory network, a neural network, and so on. In general, a network consists of nodes and edges, with the former representing biological variables and the latter for inter-variable relations. When a biological network describes a dynamic behavior, an important question is which part of the network plays a critical upstream role in controlling or activating the rest of the network. For instance, in a pathophysiological network of a disease, the upstream mechanisms are suggestive of the disease pathogenesis, which is vital for disease management. An upstream biological process must temporally precede a downstream process, a relationship known as temporal precedence. In a time-controlled experiment, upstream processes can be defined as those occurring in the beginning stage of the system under study. However, in a cross-sectional design where the data are collected without reference to time, it would be difficult to determine temporal precedence between variables when their order can go either direction. In this work, we present a novel model for the analysis of upstream mechanisms in a biological network without resorting to temporal information. Our approach is different from traditional causal analysis that establishes a cause-effect relationship by three criteria: association/correlation, temporal precedence, and non-spuriousness. While the identification of upstream mechanisms is an important research problem, no solution has been found in the literature to address this problem for cross-sectional (non-temporal) designs. Hence, the model described in the present work is significant.

Recent technology in genomics and bioinformatics has made possible a high-throughput approach based on microarrays or next-generation sequencing technology for exploring up- and down-regulated biological pathways in diseases such as cancer, resulting in highly complex biological networks, in particular, gene regulatory networks that can project biological concepts from the molecular to system levels. In the past, the gene network based on gene expression data was implemented in various forms, such as the Boolean network [1], the differential equation, cluster analysis [2], the Bayesian network [3], regression [4] and graphical. Gaussian modeling [5]. The system dynamics of the gene network can be better modeled using time-series data through methods such as dynamic Bayesian networks and differential equations [6]. In the

present research, the gene-expression data were modeled based on biological-theme networks.

The theme-based approach abstracts enriched biological themes from a large list of genes derived from genomic experiments [7-9]. This approach has the main advantage of yielding similar results despite different gene lists obtained with different methods based on the same experimental data. In the genomics community, the Gene Ontology (GO) has been developed to maintain a unified controlled vocabulary as well as annotations for genes and gene products [10]. We applied the theme analysis software called EASE [7] to identify which GO categories are over-represented on the gene sets of interest. In this way, we construct a network of biological themes from a genome-wide gene-expression data set for a particular application. Our model is devoted to determining the upstream subnetwork for the application. In the remaining sections, we present the model, mathematical analysis, validation, and application results.

Methods

A model for upstream analysis in a biological network

Our method is based on the idea that the root node in a network starts to transmit a biological signal and then activates subsequent nodes. A downstream node receives the original signal modified by some unknown factors, acting like noise. Thus, in a network of information propagation top-downward, the information associated with an upstream node will have less noise than a downstream node. Since there is no correlation between the signal propagated and the random noise, a downstream node will generally be less correlated with another distinct node in the network, compared with an upstream node. As a result, an upstream node has a greater inter-nodal correlation than a downstream node on average.

We defined the upstream-index for a node in the network as the average value of its correlation (the Pearson correlation coefficient) with each of all the other nodes in the network based on the experimental sample. The upstream-index is the key system parameter to assess the degree of the upstream role a node plays. The above arguments imply that an upstream node has a higher upstream-index value than a downstream node. The average of correlation values along with the *p*-value can be calculated using Fisher's *z*-transformation [11] below: the correlation coefficients are transformed into z-values and then the average of the *z*-values is converted back to the average correlation value. Upstream nodes are determined by the upstream-index. One limitation is that the index is only approximately precise for imposing a temporal order among all nodes because of statistical uncertainty. Additionally, equivalent nodes in the network should be removed to avoid virtual correlation.

The number of upstream nodes (one or more) depends on the complexity of the system. There is no definite minimum upstream-index threshold to determine upstream nodes, and the threshold may vary from system to system. Since this is an unsupervised learning problem, we resort to cluster analysis. All the nodes in the system are grouped into clusters in the decreasing order of the associated upstream index so that the nodes in the first cluster are called upstream and the rest of the nodes are downstream.

For cluster analysis, we use the *k*-means clustering algorithm. Recall that this clustering method clusters n objects into *k* clusters according to a similarity measure (here the upstream-index) among these objects. The optimal choice of *k* depends on the distribution of the data points. Several information theoretical measures or statistical criteria have been developed to strike a balance between a single big cluster and multiple single-object clusters. By default, a rule of thumb adopted is to let $k = \lceil\sqrt{n/2}\rceil k = \lceil\sqrt{n/2}\rceil$ (where $\lceil x\rceil\lceil x\rceil$ is the ceiling function) [12]. The *k*-means clustering is applied to the *n* nodes in the system to determine the upstream cluster.

Mathematical Analysis

The central idea is that up-stream nodes generally have a higher upstream-index (i.e., average inter-nodal correlation) than downstream nodes in the system. In our approach, such a system is conceived as top-down source-signal propagation complicated with noise. In a signal-noise model, $X_u=S_U+N_U$ where X is the measured or observed variable, S is the signal component, and N is the random noise with zero mean and magnitude given by the variance $\sigma_N^2\sigma_N^2$. The signal and noise are uncorrelated.

In the system framework thus defined, the original signal and all measurement variables are normalized in the same range. As the given signal propagates along a chain, the signal strength is decreasing or non-increasing (because some energy carried by the signal may be lost) and the noise level is increasing due to added environmental noise. The information held at a downstream node has a larger variance than an upstream node. This claim can be proven as follows: Consider the original signal S is propagated from an upstream node *U* to a downstream node DD. Suppose we consider the case of up-regulated gene expression so that the signal component is positive.

We have: $X_u = S_u + N_u X_u = S_u + N_u$ and $X_d = S_d + N_d X_d = S_d + N_d$ where $S_u \geq S_d S_u \geq S_d$ and $\sigma_{N_d}^2 > \sigma_{N_u}^2 \sigma_{N_d}^2 > \sigma_{N_u}^2$ (from model assumptions).

Treat the signal component as a constant so that $\sigma_{X_u}^2 = \sigma_{N_u}^2 \sigma_{X_u}^2 = \sigma_{N_u}^2$ and $\sigma_{X_d}^2 = \sigma_{N_d}^2 \sigma_{X_d}^2 = \sigma_{N_d}^2$.

Hence $\sigma_{X_d}^2 > \sigma_{X_u}^2 \sigma_{X_d}^2 > \sigma_{X_u}^2$

Let $S_d = \lambda_{u,d} S_u S_d = \lambda_{u,d} S_u$ such that $\lambda_{u,d} = \frac{S_d}{S_u} \leq 1 \lambda_{u,d} = \frac{S_d}{S_u} \leq 1$. Rewrite $X_d X_d$ in terms of $X_u X_u$:

$X_d = S_d + N_d = \lambda_{u,d} S_u + N_d = \lambda_{u,d}(X_u - N_u) + N_d = \lambda_{u,d} X_u + (N_d - \lambda_{u,d} N_u)$

Let $N_{u,d} = N_d - \lambda_{u,d} N_u N_{u,d} = N_d - \lambda_{u,d} N_u$. We derive: $X_d = \lambda_{u,d} X_u + N_{u,d} X_d = \lambda_{u,d} X_u + N_{u,d}$.

A Corollary can be derived as follows:

$X_d = \lambda_{u,d} X_u + N_{u,d} X_d = \lambda_{u,d} X_u + N_{u,d}$ where $\lambda_{u,d} = \frac{S_d}{S_u} \leq 1 \lambda_{u,d} = \frac{S_d}{S_u} \leq 1$ and $N_{u,d} = N_d - \lambda_{u,d} N_u N_{u,d} = N_d - \lambda_{u,d} N_u$.

Given an upstream node *U*, a downstream node *D*, and another distinct node *A* in the system where the noise associated with each node is uncorrelated with the noise at other nodes. The correlation between node *D* and node *A* is smaller than the correlation between node *U* and node *A*. This claim can be proven as follows:

Let variables X_u, X_d, X_a X_u, X_d, X_a denote the values measured at the upstream node, downstream node, and another distinct node in the system, respectively. Consider the case of up-regulated gene expression. Based on Corollary 1, $X_d X_d$ is expressed in terms of $X_u X_u$: $X_d = \lambda_{u,d} X_u + N_{u,d} X_d = \lambda_{u,d} X_u + N_{u,d}$ where $\lambda_{u,d} \lambda_{u,d}$ ($\lambda_{u,d} \leq 1 \lambda_{u,d} \leq 1$) is the parameter of signal propagation from node *U* to node *D*, and $N_{u,d} N_{u,d}$ ($N_{u,d} = N_d - \lambda_{u,d} N_u N_{u,d} = N_d - \lambda_{u,d} N_u$) is the additive noise. Calculate the (Pearson) correlation coefficient (μ: mean, σ: standard deviation):

Correlation between $X_u X_u$ and $X_a X_a$: $\rho(X_u, X_a) = \frac{E[X_u X_a] - E[X_u] E[X_a]}{\sigma_{X_u} \sigma_{X_a}} = \frac{\mu_{X_u X_a} - \mu_{X_u} \mu_{X_a}}{\sigma_{X_u} \sigma_{X_a}} \rho(X_u, X_a) = \frac{E[X_u X_a] - E[X_u] E[X_a]}{\sigma_{X_u} \sigma_{X_a}} = \frac{\mu_{X_u X_a} - \mu_{X_u} \mu_{X_a}}{\sigma_{X_u} \sigma_{X_a}}$

Correlation between $X_d X_d$ and $X_a X_a$: $\rho(X_d, X_a) = \frac{E[X_d X_a] - E[X_d] E[X_a]}{\sigma_{X_d} \sigma_{X_a}} = \frac{\mu_{X_d X_a} - \mu_{X_d} \mu_{X_a}}{\sigma_{X_d} \sigma_{X_a}} \rho(X_d, X_a) = \frac{E[X_d X_a] - E[X_d] E[X_a]}{\sigma_{X_d} \sigma_{X_a}} = \frac{\mu_{X_d X_a} - \mu_{X_d} \mu_{X_a}}{\sigma_{X_d} \sigma_{X_a}}$

Rearrange $\rho(X_d, X_a) \rho(X_d, X_a)$ as follows:

$E[X_d X_a] = E[(\lambda_{u,d} X_u + N_{u,d}) X_a] = E[\lambda_{u,d} X_u X_a] + E[N_{u,d} X_a] E[X_d X_a] = E[(\lambda_{u,d} X_u + N_{u,d}) X_a] = E[\lambda_{u,d} X_u X_a] + E[N_{u,d} X_a]$ ($N_{u,d} N_{u,d}$ and $X_a X_a$ are independent.)

$= \lambda_{u,d} E[X_u X_a] + E[N_{u,d}] E[X_a] = \lambda_{u,d} E[X_u X_a] + 0 = \lambda_{u,d} \mu_{X_u X_a}$

$= \lambda_{u,d} E[X_u X_a] + E[N_{u,d}] E[X_a] = \lambda_{u,d} E[X_u X_a] + 0 = \lambda_{u,d} \mu_{X_u X_a}$ ($N_{u,d} N_{u,d}$ has zero mean.)

$E[X_d] = E[\lambda_{u,d} X_u + N_{u,d}] = E[\lambda_{u,d} X_u] + E[N_{u,d}] = \lambda_{u,d} E[X_u] + 0 = \lambda_{u,d} \mu_{X_u}$

$E[X_d] = E[\lambda_{u,d} X_u + N_{u,d}] = E[\lambda_{u,d} X_u] + E[N_{u,d}] = \lambda_{u,d} E[X_u] + 0 = \lambda_{u,d} \mu_{X_u}$

Hence, $\rho(X_d, X_a) = \frac{E[X_d X_a] - E[X_d] E[X_a]}{\sigma_{X_d} \sigma_{X_a}} = \frac{\lambda_{u,d} \mu_{X_u X_a} - \lambda_{u,d} \mu_{X_u} \mu_{X_a}}{\sigma_{X_d} \sigma_{X_a}} = \frac{\lambda_{u,d} (\mu_{X_u X_a} - \mu_{X_u} \mu_{X_a})}{\sigma_{X_d} \sigma_{X_a}}$

$\rho(X_d, X_a) = \frac{E[X_d X_a] - E[X_d] E[X_a]}{\sigma_{X_d} \sigma_{X_a}} = \frac{\lambda_{u,d} \mu_{X_u X_a} - \lambda_{u,d} \mu_{X_u} \mu_{X_a}}{\sigma_{X_d} \sigma_{X_a}} = \frac{\lambda_{u,d} (\mu_{X_u X_a} - \mu_{X_u} \mu_{X_a})}{\sigma_{X_d} \sigma_{X_a}}$

Compare $\rho(X_d, X_a)\rho(X_d, X_a)$ with $\rho(X_u, X_a)\rho(X_u, X_a)$:

The numerator: $\lambda_{u,d}\left(\mu_{X_u X_a} - \mu_{X_u}\mu_{X_a}\right) \leq \left(\mu_{X_u X_a} - \mu_{X_u}\mu_{X_a}\right)$ $\lambda_{u,d}\left(\mu_{X_u X_a} - \mu_{X_u}\mu_{X_a}\right) \leq \left(\mu_{X_u X_a} - \mu_{X_u}\mu_{X_a}\right)$ (because $\lambda_{u,d} \leq 1 \lambda_{u,d} \leq 1$)

The denominator: $\sigma_{X_d}\sigma_{X_a} > \sigma_{X_u}\sigma_{X_a}\sigma_{X_d}\sigma_{X_a} > \sigma_{X_u}\sigma_{X_a}$ (From Theorem 1, $\sigma_{X_d}^2 > \sigma_{X_u}^2\sigma_{X_d}^2 > \sigma_{X_u}^2$)

Hence, $\rho(X_d, X_a) < \rho(X_u, X_a)\rho(X_d, X_a) < \rho(X_u, X_a)$.

Construction of Biological Themes

Given a functional genomics data set, significantly expressed genes are determined based on the comparison of the test (experimental) against control samples, using a statistical analysis program called GEO2R accessible at NCBI GEO website [13]. GEO2R compares two or more groups of samples to identify genes differentially regulated across experimental conditions. GEO2R returns several statistical results, in particular, the t-statistic and *p*-values, including both the raw and adjusted *P*-values, and the gene selection criteria were based on the adjusted *P*-values.

Then up-regulated and down-regulated genes were analyzed separately, using a software tool (http://david.abcc.ncifcrf.gov/ease/ease.jsp) known as EASE (the Expression Analysis Systematic Explorer) [7], resulting in up-regulated and down-regulated biological themes, respectively. EASE discovers themes from a given list of genes by performing an over-representation analysis with respect to selected biological system categories with the Gene Ontology (GO) [10] ("Biological process" as the default). Over-representation means that a disproportionately higher number of genes belong to a particular system category in the given gene list than in the gene population (e.g., the genome).

A biological network can be constructed where each node represents a biological theme. The identified biological theme output by the EASE program is referred to the "Biological process". Each biological process *bp* consists of a group of differentially expressed genes sharing a specific functional feature: $bp = \langle g_1, g_2, \ldots g_k \rangle$ where *g* denotes a gene. The expression activity of a process is defined as the average expression level of all genes associated with the process. The inter-process correlation between processes A and B refers to the correlation between the process expression-activity level of A and that of B in the given experimental sample. The biological processes so-obtained are collected to form a fully-connected biological network, to which the upstream-analysis model can be applied. It should be noted that the upstream analysis model is applicable to any biological network where each node represents a meaningful variable in the system; it does not matter whether the variable corresponds to a biological theme or not.

The genomics (genome-wide gene expression) datasets used in this research were obtained from the NCBI GEO database in the public domain (https://www.ncbi.nlm.nih.gov/gds), including human Alzheimer's disease (accession number GSE1297), human Parkinson's disease (GSE8397), murine prion diseases (GSE63930), murine traumatic brain injury (GSE24047), murine mucosa wound healing (GSE23006), and murine immune response to bacterial infection (GSE23014).

Results and Discussions

Validation Against the Gold Standard Based on Time-Series Data

In a time-series experiment, the gene expression of the sample is measured at a finite number of time points along the time axis, starting at $t = 0$. The gene expression patterns observed at early time intervals tend to suggest upstream patterns. It seems that the biological processes identified in the first-time period will be upstream or approximately upstream.

Suppose the processes identified for each period are mixed up in a random order. The upstream processes identified by upstream analysis were compared with the upstream processes according to the time parameter in an effort to test the validity of our model, which used the defined upstream-index rather than expression time as the basis for the inference of upstream processes. By so doing, it should be clear that the time-series data were used as the gold standard for validation.

We analyzed multiple experimental datasets each with the time axis included: (i) traumatic brain injury (GSE24047) (ii) mucosa wound healing (GSE23006) (iii) immune response to bacterial infection (GSE23014). These datasets were selected independently and unbiasedly, satisfying the constraint that the gene-expression pattern in each defined interval on the time axis comprised some significantly expressed genes (adjusted *p* value < 0.05 or less), as needed for upstream analysis. In each domain, the gene expression changes were identified by comparing the test sample against the control model at each predefined time point. For each dataset, the biologi-

cal processes corresponding to the time interval-based gene expression changes were analyzed for upstream processes based on the upstream-index, irrespective of the expression time. It turned out that our method successfully identified upstream processes consistent with their early expression time and temporal order for all three datasets under testing (Table 1). Time-series experiments are a useful lab approach in experimental biology. Since, however, it would be cumbersome and time consuming to conduct a time-series study for a clinical disease with complex etiology, there is strong need for a computational tool that performs upstream analysis in a cross-sectional study.

Traumatic Brain Injury			Wound Healing			Immune Response		
Time	**Ups-index**	**Role**	**Time**	**Ups-index**	**Role**	**Time**	**Ups-index**	**Role**
3h	0.528	Up-stream	6h	0.891	Up-stream	12d	0.857	Up-stream
6h	0.670	Up-stream	12h	0.910	Up-stream	15d	0.876	Up-stream
12h	0.355	Down-stream	24h	0.866	Up-stream	21d	0.801	Down-stream
48h	0.413	Down-stream	3d	0.588	Down-stream			
			5d	0.813	Down-stream			
			7d	0.459	Down-stream			
			14d	Removed (~ 7d)				

Table 1: The results of our analysis on three experiment datasets concerning molecular biology based on gene expression. The datasets recorded the temporal events in response to a specific insult. (a) The murine brain response following traumatic injury evaluated at 4-time points, using a sham brain as the control model. (b) The murine wound healing process following tongue mucosa injury observed at 7-time points, using unwounded tongue as the control model. (c) The murine lung immune response following an infection by tuberculosis bacteria measured at 3-time points, using day 0 as the control model. The biological process in each time interval was generally represented by the 10 most significantly induced or repressed genes that were identified by comparing the test sample collected at the end point against the control sample. The upstream-indexes of the biological processes on each dataset were computed without referencing any temporal information. The biological processes were then classified as upstream or downstream on the basis of the upstream-index. The results show that the biological processes in early post-insult time intervals were classified as upstream consistently in all domains under investigation, demonstrating the validity and generality of the upstream analysis model.

Robustness of the upstream-index

The identified themes are ranked by their statistical significance according to the associated EASE score (the Fisher's exact probability modified in favor of themes supported by more genes) [7]. EASE themes have the advantage that they are consistent despite the variability in gene lists derived by different gene selection criteria. This approach facilitates the interpretation of the results of microarray or other high-throughput experiments. However, theme analysis is limited in that it is neither causal analysis nor upstream analysis.

Because the success of our model hinges on the accuracy of the upstream-index for upstream analysis, it is important to validate the upstream-index per se. If the upstream-index calculated based on the given data is consistent with the true temporal order, the index is considered valid. Since, however, it is impractical to obtain all possible data, we employed an idea in statistical learning for method validation so-called cross-validation, a technique for validating a supervised learning system when response values are given. Although the problem addressed in the present study is an unsupervised-learning problem without known responses given in the data, we designed a similar cross-validation scheme as follows. First, the given data were randomly divided into two sets of about equal size. Then the upstream-index was calculated for each variable based on each set of the data separately. Finally, the Spearman rank correlation coefficient between the results based on each data set was computed. Then a statistically significant correlation between the two results would suggest that our method can arrive at a consistent temporal order based on the upstream-index, independent of a particular data set. Here, the Spearman correlation is a better choice because the Pearson correlation coefficient evaluates the linear relationship between two variables whereas the Spearman correlation evaluates the monotonic relationship. This validation experiment was conducted in the domain of Parkinson's Disease (PD). The dataset (GSE8397) consists of 29 PD cases, which were randomly divided into two sets, 15 and 14 cases, respectively. For up-regulated biological processes, the Spearman correlation is 1 and the two-tailed p-value is 0. For down-regulated biological processes, the Spearman correlation is 0.565 and the two-tailed p-value is 0.023. By normal standards, both results would be considered statistically significant, and hence the robustness of the upstream-index was demonstrated.

Test on real-world data

Our method was applied to three clinical genomics data sets concerning the three neurodegenerative diseases, including Alzheimer's disease (AD), Parkinson's disease (PD), and prion diseases, respectively. Our analysis on the AD genomics data selected

1977 up-regulated and 1436 down-regulated differentially expressed genes, and identified 29 up-regulated and 92 down-regulated over-represented biological processes (or called pathological processes). For PD data, we selected 127 up-regulated and 429 down-regulated differentially expressed genes, and identified 10 up-regulated and 53 down-regulated over-represented biological processes. For prion-disease data, we selected 859 up-regulated and 174 down-regulated differentially expressed genes, and identified 30 up-regulated and 7 down-regulated over-represented biological processes. The biological processes of each disease were clustered by their upstream-index. The pathogenesis processes corresponded to the cluster of the biological processes associated with the highest upstream-index. The up-regulated and down-regulated processes were handled separately. The final pathogenesis processes identified in this analysis along with their upstream-indexes and statistical significance scores (EASE scores) are displayed in Table 2 and Table 3. The agreement between the pathogenesis (i.e., upstream biological mechanisms) identified by our method and the main pathogenesis theories described in the literature (Table 2-4) is evidence of our model validity.

Alzheimer's Disease		**Parkinson's Disease**		**Prion Disease**	
Process	UI/ES	Process	UI/ES	Process	UI/ES
Regulation of transcription (DNA-dependent)	0.919/0.000	Transcription, DNA-dependent	0.790/0.004	Physiological process	0.985/0.044
Development	0.911/0.002	RNA metabolism	0.773/0.007	Protein metabolism	0.985/0.007
Protein amino acid phosphorylation	0.906/0.046	Regulation of biological process	0.693/0.025	Cell death	0.983/0.046
Transcription from Pol II promoter	0.905/0.033			Catabolism	0.982/0.026
Cell cycle	0.904/0.005			Cell proliferation	0.979/0.000
				Defense response	0.977/0.000

Table 2: The up-regulated pathogenesis derived from upstream biological mechanisms for the neurodegenerative diseases. UI: Upstream index. ES: EASE score.

Alzheimer's Disease		**Parkinson's Disease**		**Prion Disease**	
Process	UI/ES	Process	UI/ES	Process	UI/ES
Transport	0.949/0.000	Transport	0.926/0.000	Lipid biosynthesis	0.720/0.005
Cellular process	0.944/0.026	Intracellular transport	0.921/0.002	Potassium ion transport	0.720/0.007
Protein metabolism	0.943/0.018	Cation transport	0.912/0.006		
Catabolism	0.942/0.035	Microtubule based process	0.887/0.001		
Intracellular transport	0.941/0.000	Post-Golgi transport	0.886/0.046		
Phosphate Metabolism	0.940/0.044	Exocytosis	0.873/0.003		
Energy pathways	0.940/0.000				
Biosynthesis	0.939/0.013				

Table 3: The down-regulated pathogenesis derived from upstream biological mechanisms for the neurodegenerative diseases. UI: Upstream index. ES: EASE score.

Disease	**Pathogenesis based on upstream model**	**Pathogenesis found in Literature**
Alzheimer's disease	Protein amino acid phosphorylation	Amyloid plaques and intraneuronal neurofibrillary tangles [14]

Parkinson's disease	Transcription, RNA metabolism	α-synuclein and misfolded protein [15,16]
Prion disease	Protein metabolism, defense response	Infectious misfolded prions [17]

Table 4: The comparison between the pathogenesis based on the upstream-analysis model and the pathogenesis documented in the literature for the neurodegenerative diseases.

Our analysis produced different pathogenesis patterns for AD and PD. Both diseases activated transcriptional pathways, but they were different. In particular, AD induced the tumor suppressor response [18], whereas PD exhibited the unfolded protein response. Additionally, the processes of protein and amino acid phosphorylation as well as the cell cycle were implicated in AD, while the process of RNA metabolism was implicated in PD. In the literature, the potential role of RNA metabolism, notably mRNA translation, was hypothesized in PD pathogenesis [19]. Both diseases are characterized by accumulation of certain pathological proteins, but these proteins are of different nature in response to different pathogenesis factors.

The transport mechanism was identified as the pathogenesis mechanism in down-regulated genomics activities for both AD and PD. This result is consistent with our knowledge that the transportation of the neurotransmitters across neurons is fundamental to the nervous systems, including the brain. Furthermore, endoplasmic reticulum dysfunction is proposed as a potential cause for loss of dopaminergic neurons [15]. In another aspect, protein metabolism was identified as a pathogenesis mechanism in the down-regulated genomics of AD but not PD, suggesting a different pathogenesis mechanism between AD and PD.

Prion diseases are fatal neurodegenerative disorders in mammals that are caused by a transmissible pathogenic isoform of the prion protein [20]. The most common human prion disease is Creutzfeldt-Jakob Disease (CJD). The mechanism of defense and immune responses was the most significant biological process in the prion disease in terms of the EASE score, but the pattern of pathogenesis produced by our analysis also included several other biological processes, such as protein metabolism and cell death, which are known to play important roles in this disease (Table 2). The original study that created the data set of prion diseases, GSE63930, performed a pathway analysis using the DAVID database and also found that the gene expression profile of the prion disease was characterized by immune response and inflammation [17]. The prion disease is currently viewed as an infectious neurodegenerative disease carrying misfolded prion proteins that can infect other proteins. Because of its infectious nature, the disease elicits immune and inflammatory responses.

Some research indicates that prions can cause neurodegeneration, but there is no evidence of transmissibility of AD and PD between humans [21]. In other words, AD and PD may not be a prion disease from the clinical standpoint. The pathogenesis patterns identified for AD and PD in our analysis failed to show any infection and immune response, as typical of prion diseases, although It had been hypothesized that some pathological proteins in AD and PD may behave like prions [22]. The elucidation of genomics-based findings demands further biological or neuropathological experiments in the future.

Conclusions

Recent genomics technology has enabled the construction of a complex network of biological pathways based on gene expressions measured using the microarray or next-generation sequencing platforms. However, many public-domain gene-expression data sets concerning complex diseases such as neurodegenerative diseases are not time-series data, which imposes difficulty to determine the upstream biological mechanisms of the disease under study. To cope with this issue, we have developed an effective new model for analysis of upstream mechanisms in a biological network. Our model is able to reconstruct an approximate temporal order from non-temporal data. This model is significant since the problem of upstream analysis in cross-sectional (non-temporal) designs has not been addressed in other's research work. This model has been validated against the gold standard based on several time-series data sets. The application of this model to real diseases also generated important findings consistent with the literature, suggesting its potential applications to new disease research.

Acknowledgments

The genomics datasets used in this study were obtained from the NCBI GEO database in the public domain.

References

1. S Liang, S Fuhrman, R.Somogyi (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures, Pac Symp Biocomput 18-29.
2. S Tavazoie, JD Hughes, MJ Campbell, RJ Cho, GM Church (1999) Systematic determination of genetic network architecture. Nat Genet 22: 281-285.
3. N Friedman, M Linial, I Nachman, D Pe'er (2000) Using Bayesian networks to analyze expression data. Journal of computational biology : a journal of computational molecular cell biology 7: 601-620.
4. A DobraDobra, C Hans, B Jones, JR Nevins, G Yao, et al. (2004) Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis 90: 196-212.
5. H Toh, K Horimot (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 18: 287-297.

6. C Sima, J Hua, S Jung (2009) Inference of gene regulatory networks using time-series data: a survey. Current genomics 10: 416-429.

7. DA Hosack, G Dennis, BT Sherman, HC Lane, RA Lempicki (2003) Identifying biological themes within lists of genes with EASE, Genome Biol, 4 R70.

8. W Huang da, BT Sherman, RA Lempicki (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

9. W Huang da, BT Sherman, Q Tan, JR Collins, WG Alvord, et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, Genome Biol, 8: R183.

10. M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

11. DM Corey, WP Dunlap, MJ Burke (1998) Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations. The Journal of General Psychology 125: 245-261.

12. KV Mardia, JT Kent, JM Bibby (1979) Multivariate Analysis, Academic Press, London.

13. T Barrett, SE Wilhite, P Ledoux, C Evangelista, IF Kim, et al. (2013) NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 41: D991-995.

14. P Tiraboschi, LA Hansen, LJ Thal, J Corey-Bloom (2004) The importance of neuritic plaques and tangles to the development and evolution of AD. Neurology 62: 1984-1989.

15. LB Moran, DC Duke, M Deprez, DT Dexter, RK Pearce, et al. (2006) Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics 7: 1-11.

16. LM Fu, KA Fu (2015) Analysis of Parkinson's disease pathophysiology using an integrated genomics-bioinformatics approach, Pathophysiology : the official. journal of the International Society for Pathophysiology / ISP 22: 15-29.

17. A Herbst, A Ness, CJ Johnson, D McKenzie, JM Aiken (2015) Transcriptomic responses to prion disease in rats. BMC genomics 16: 682.

18. EM Blalock, JW Geddes, KC Chen, NM Porter, WR Markesbery, et al. (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A 101: 2173-2178.

19. B Lu, S Gehrke, Z Wu (2014) RNA metabolism in the pathogenesis of Parkinsons disease, Brain research.

20. RT Johnson (2005) Prion diseases, The Lancet. Neurology 4: 635-642.

21. M Beekes, A Thomzig, WJ Schulz-Schaeffer, R Burger (2014) Is there a risk of prion-like disease transmission by Alzheimer- or Parkinson-associated protein particles?. Acta neuropathologica 128: 463-476.

22. M Goedert (2015) NEURODEGENERATION. Alzheimer's and Parkinson's diseases: The prion concept in relation to assembled Abeta, tau, and alpha-synuclein, Science 349: 1255555.