Archive of SID

ISSN: 2476-6917; Modares Journal of Biotechnology. 2019;10(4):545-555

Inferring Co-Expression Networks from their Associated Attrib-utes by Neural Networks

ARTICLE INFO

Article Type
Original Research

Authors
Mahdevar Gh.*[1] *PhD*

How to cite this article
Mahdevar Gh. Inferring Co-Expression Networks from their Associated Attributes by Neural Networks. Mod-ares Journal of Biotechnology. 2019;10(4):545-555.

ABSTRACT

Gene expression, flow of information from DNA to proteins, is a fundamental biological process. Expression of one gene can be regulated by the product of another gene. These regulatory relationships are usually modeled as a network; genes are modeled as nodes and their relationships are shown as edges. There are many efforts for discovering how genes regulate expression of themselves. This paper presents a new method that employs expression data and ontological data to infer co-expression networks, networks made by connecting genes with similar expression patterns. In brief, the method begins by learning associations between the available ontological information and the provided co-expression data. Later, the method is able to find both known and novel co-expressed pairs of genes. Finally, the method uses a self-organizing map to adjust estimation made by the previous step and to form the GCN for the input genes. The results show that the proposed method works well on the biological data and its predictions are accurate; consequently, co-expression networks generated by the proposed method are very similar to the biological networks or those that constructed with no missing data. The method is written in C++ language and is available upon request from the corresponding author.

Keywords Gene Expression; Neural Networks; Unsupervised Machine Learning

CITATION LINKS

[1] Campbell ... [2] Molecular cell ... [3] Modeling and simulation of genetic regulatory systems: A ... [4] Gene regulatory network inference: Data intgrtion in dynamic models-a ... [5] Combining partial correlation and an information theory ... [6] Advantages and limitations of current network ... [7] Reverse engineering gene regulatory network from microarray data ... [8] Revealing strengths and weaknesses of methods for gene network ... [9] Comparing statistical methods for constructing large scale ... [10] Current approaches to gene regulatory ... [11] Classification of co-expressed genes from DNA regulatory ... [12] Inferring regulatory networks from expression data using ... [13] A general co-expression network-based approach to gene expression ... [14] Construction and analysis of protein-protein interaction ... [15] Boolean dynamics of genetic regulatory networks inferred from ... [16] Predicting gene expression from ... [17] Inferring gene correlation networks from transcription ... [18] Gene ontology: Tool for the unification of ... [19] Use and misuse of the gene ontology ... [20] Quantifying the relationship between co-expression, ... [21] Constructing gene co-expression networks and predicting functions ... [22] Reconstruction of gene co-expression network from microarray ... [23] Correlation between gene expression and GO ... [24] Gene expression correlation and gene ontology-based ... [25] Semantic similarity in a taxonomy: An information-based measure and its ... [26] A new method to measure the semantic similarity of GO ... [27] A gene-coexpression network for global discovery of conserved ... [28] Self-organization and associative ... [29] Understanding machine learning: From theory to ... [30] Introduction to machine ... [31] Data mining: Concepts and ... [32] The elements of statistical ... [33] Beyond regression: New tools for prediction and analysis in the behavioral ... [34] Network biology: Understanding the cell's functional ... [35] Similarities and differences in genome-wide expression data of six ... [36] Reconstruction of metabolic networks from genome ... [37] YeastNet v3: A public database of data-specific and ... [38] Discovering gene association networks by multi-objective evolutionary ... [39] Neural networks: A ... [40] Comparing association network algorithms for reverse ... [41] Towards reconstruction of gene networks from ... [42] Inferring gene regression networks with model ... [43] Inferring genetic regulatory logic from expression ... [44] Inferring adaptive regulation thresholds and association rules from ...

[1]Mathematics Department, Sciences Faculty, University of Isfahan, Isfahan, Iran

***Correspondence**
Address: Mathematics Department, Sciences Faculty, University of Isfahan, Azadi Square, Isfahan, Iran. Postal Code: 8174673441
Phone: +98 (31) 37934611
Fax: -
gh.mahdevar@sci.ui.ac.ir

Article History
Received: September 10, 2018
Accepted: July 15, 2019
ePublished: December 21, 2019

Copyright© 2019, TMU Press. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

www.SID.ir

Archive of SID

استنتاج شبکه هم‌بیانی ژن‌ها از روی ویژگی‌های منتسب‌شده به آنها به‌وسیله شبکه‌های عصبی مصنوعی

قاسم مهدور* PhD

گروه ریاضی، دانشکده علوم، دانشگاه اصفهان، اصفهان، ایران

چکیده

فرآیند شارش اطلاعات از DNA به پروتین‌ها که به بیان ژن موسوم است، یک فرآیند پایه‌ای در زیست‌شناسی است. تنظیم بیان ژن‌ها پاسخ سلول‌ها به محرک‌های فراوانی بوده و برای آنها حیاتی است. ژن‌ها با بیان مشابه در یک سری آزمایش مناسب، ژن‌های هم‌بیان، به‌طور معمول توسط تنظیم‌کننده‌های یکسان مدیریت می‌شوند و باز هم به‌طور معمول تغییر در بیان آنها پاسخ به محرک‌های یکسانی هستند.

در این مقاله ما یک روش جدید ارایه کرده‌ایم که داده‌های مرتبط با بیان و هستی‌شناسی ژن‌ها را به‌کارگرفته و به‌وسیله آنها ژن‌های هم‌بیان را یافته و شبکه هم‌بیانی ژن‌ها را ایجاد می‌کند.

در ابتدای روش ایجادشده یک شبکه عصبی مصنوعی روابط بین خصایص منتسب‌شده به ژن‌ها توسط پروژه هستی‌شناسی ژن‌ها و میزان مشابهتی که در بیان با یکدیگر دارند را فرا می‌گیرد. به‌سادگی، خصایص گردآوری‌شده توسط هستی‌شناسی ژن‌ها شامل عملکرد، فرآیند، و محل فعالیت ژن‌ها هستند. بعد از پایان مرحله یادگیری، شبکه عصبی مصنوعی قادر است ژن‌های هم‌بیان را کشف کند. به‌علاوه، شبکه‌های زیستی از چندین گروه ژنی به‌هم‌پیوسته ساخته شده‌اند، به همین دلیل یافتن این گروه‌ها می‌تواند کیفیت شبکه‌های هم‌بیانی ساخته شده را بالا ببرد. بنابراین، در گام بعدی روش، یک شبکه عصبی مصنوعی دیگر گروه ژن‌ها را از روی همان خصایص هستی‌شناسی پیدا می‌کند. تحلیل‌های ما نشان دادند که نتایج روش ایجادشده شباهت زیادی به نتایج آزمایشگاهی دارد. همچنین، ما نشان دادیم که شبکه‌های هم‌بیانی ساخته‌شده توسط آن مشابه هم‌ارزهای زیستی و حتی مشابه آنهایی است که با داده‌های بدون نقص ساخته شده‌اند. درنهایت، ما از زبان C++ برای نوشتن روش استفاده کرده‌ایم و برنامه آن در دسترس است.

کلیدواژه‌ها: بیان ژن، شبکه عصبی پس انتشار، نگاشت خودسازمان‌دهنده

تاریخ دریافت: ۱۳۹۷/۶/۱۹

تاریخ پذیرش: ۱۳۹۸/۴/۲۴

*نویسنده مسئول: gh.mahdevar@sci.ui.ac.ir

مقدمه

فرآیند تنظیم بیان ژن‌ها توسط پروتئین‌ها، که خود محصول ژن‌ها هستند به سلول‌ها امکان استفاده از یک مجموعه محدود از ژن‌ها برای رشد و گذر از شرایط محیطی مختلف را داده است[1, 2]. سلول‌ها یک دنباله از رویه‌های تنظیمی را روی این فرآیند پراهمیت اعمال می‌کنند. برای مثال، تنظیم موقعیت ژن‌های درون هسته، تنظیم معماری کروماتین، یا تنظیم‌های که توسط عوامل رونویسی هماهنگ می‌شوند[2]. کشف این که چگونه ژن‌ها تنظیم می‌شوند اهمیت زیادی داشته موضوع پژوهش‌های زیادی بوده است[3, 4].

ترسم ژن‌ها به‌صورت گره و کشیدن یال بین ژن‌های هم‌بیان شبکه‌ای را ایجاد می‌کند که به آن *شبکه* هم‌بیانی ژن‌ها گفته می‌شود[5]. این شبکه کاربردهای زیادی دارد که یافتن ژن‌های بیماری‌زا یکی از آنها است[6, 7]. در نتیجه استنتاج یا ساخت این شبکه‌ها از روی یک یا چند منبع زیستی یکی از هدف‌های مهم در زیست‌شناسی محاسباتی است[3, 6-8]. در این مقاله ما نیز به این موضوع پرداخته‌ایم.

روش‌های متفاوت با ایده‌های گوناگون برای استنتاج شبکه ژن‌های داده‌شده ارایه شده است[8-10]. هدف برخی از این روش‌ها یافتن مولکول‌های است که به‌صورت مستقیم بیان ژن‌ها را تنظیم می‌کنند[11]. برخی دیگر، از میان‌کنش مستقیم بین مولکول‌ها صرف نظر کرده و تلاش می‌کنند تا تاثیر یک ژن روی بیان ژن‌ها را بیابند[12, 13].

در اساس، روش‌های موجود اطلاعات داده‌شده درباره ژن‌های ورودی را به کار بسته و شبکه‌ای که ژن‌ها روی آن بنا شده‌اند را با محاسبه میزان تاثیر ژن‌ها روی یکدیگر می‌یابند. چندین روش شبکه ژن‌های داده‌شده را از روی غلظت محصولات آن ژن‌ها می‌سازند[14]. برخی دیگر میزان رونوشت ایجادشده از ژن‌ها را در نظر می‌گیرند[7, 15]. تمام این روش‌ها با پیدایش فناوری‌های با خروجی فراوان رواج بیشتری یافته‌اند. برای نمونه، فناوری میکروواری امکان به‌دست‌آوردن میزان رونوشت موجود از هزاران ژن در یک لحظه دلخواه را به پژوهشگران داده است. به‌علاوه، تعدادی از روش‌ها داده‌های فرازمانی مرتبط با خود ژن‌ها، برای نمونه توالی‌های بالادستی، را به کار گرفته‌اند[11, 16, 17].

یک منبع پرکاربرد دیگر در زیست‌شناسی محاسباتی اطلاعات هستی‌شناسی ژن‌ها است که توسط پروژه‌ای با همین نام و به نشانی www.geneontology.org ارایه می‌شود[18]. این پروژه مجموع دانش موجود درباره چندین گونه است که با کمک سه فرهنگ واژه، فرآیند، عملکرد و محل فعالیت ژن‌ها را تبیین می‌کند[19]. همچنین این پروژه به هر ویژگی یک برچسب منحصربه‌فرد منتسب کرده و امکان فهرست‌کردن ویژگی‌های منتسب‌شده به یک ژن و ژن‌های برچسب علامت زده‌شده توسط یک برچسب را به کاربران می‌دهد. شکل ۱ بخش کوچکی از هستی‌شناسی فرآیندهای زیستی را نشان می‌دهد.

ژن‌های هم‌بیان، ژن‌هایی که سطح بیان آنها در آزمایش‌های به‌عمل‌آمده شباهت زیادی دارند، به‌طور معمول رابطه تنظیمی با یکدیگر دارند، یعنی هم‌بیانی از وجود رابطه تنظیمی حکایت می‌کند[20-22]. همچنین نشان داده شده است که ژن‌های هم‌بیان برچسب‌های مشترک زیادی دارند[23, 24]. به‌علاوه، پژوهشگران معیارهای گوناگونی برای اندازه‌گیری میزان شباهت دو مجموعه از برچسب‌ها به وجود آورده‌اند (برای نمونه، ضریب همبستگی و اطلاعات متقابل) [25, 26].

در این مقاله، ما روشی جدید برای ساخت شبکه هم‌بیانی ژن‌ها از روی اطلاعات هستی‌شناسی و داده‌های مرتبط با بیان آنها بیان می‌کنیم (شکل ۲). به‌صورت خلاصه، در ابتدای این روش رابطه بین ویژگی‌های ژن‌ها و شباهت بیان آنها توسط یک شبکه عصبی فرا

www.SID.ir

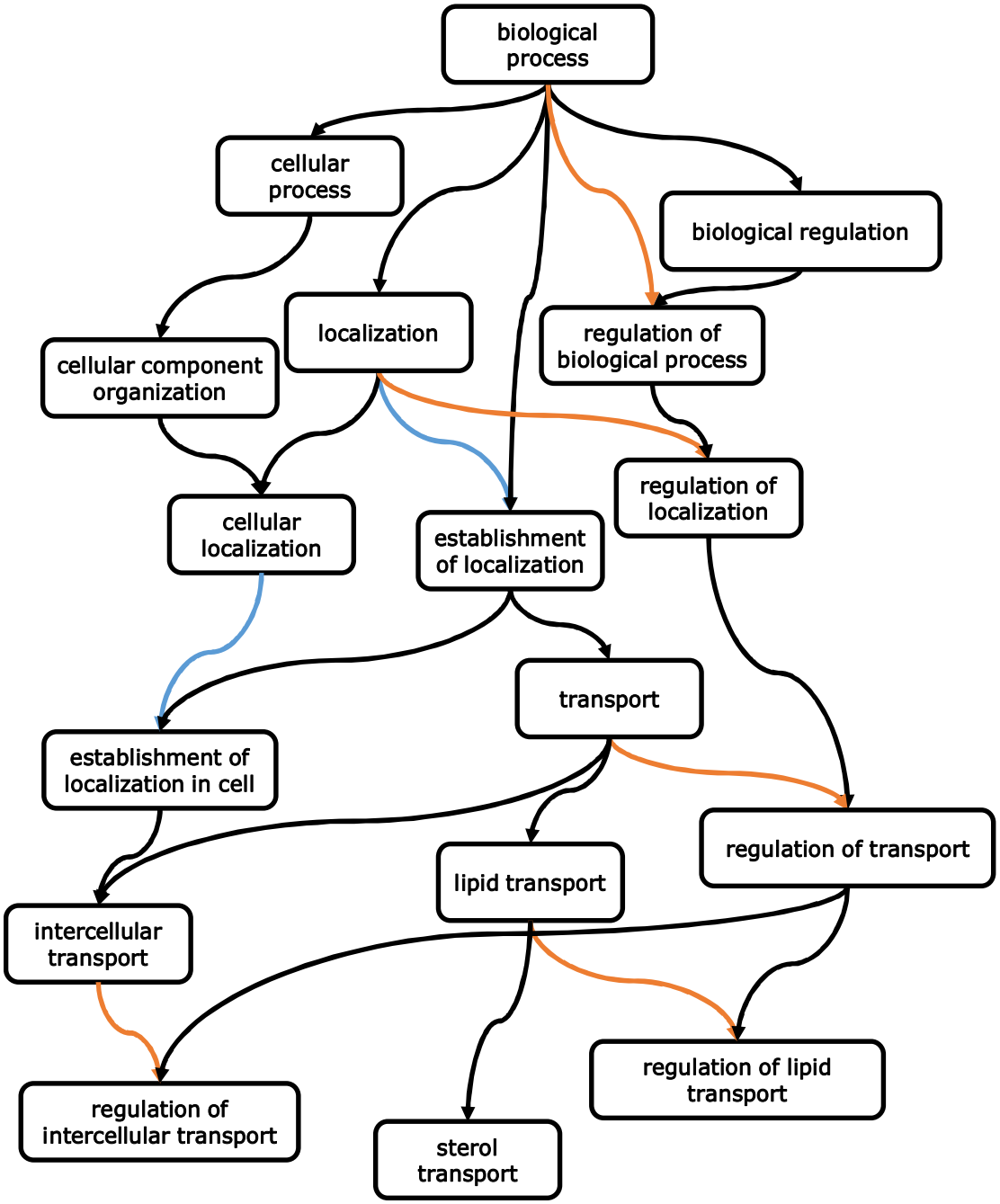
Archive of SID

گرفته می‌شود. در ادامه، با کمک این شبکه می‌توان ژن‌های هم‌بیان را از روی ویژگی‌های آنها تشخیص داد. در انتها ما از یک نگاشت خودسازمان‌دهنده برای یافتن گروه‌های ژنی و بالابردن دقت یافته‌ها استفاده می‌کنیم. جزییات روش در بخش مواد و روش‌ها آمده است. دقت و کارآیی آن نیز در این بخش ارایه خواهد شد.

شکل ۳ بخش کوچکی از ماتریس برچسب‌ها که توسط داده‌های دریافت‌شده از پروژه هستی‌شناسی ژن‌ها ساخته شده است را نشان می‌دهد. خوشبختانه، پایگاه این پروژه به نشانی www.geneontology.org برچسب‌های منتسب‌شده به ژن‌های چندین گونه زیستی را ارایه می‌کند[18].

درایه $c_{i,j}$ ضریب همبستگی مشاهده‌شده بین الگوی بیان ژن‌های iام و jام را نشان می‌دهد. به‌علاوه $c_{i,j} = -$ نشان‌دهنده ضریب همبستگی گزارش‌نشده یا آزموده‌نشده است. شکل ۴ ضریب همبستگی بین بیان ۷ ژن را نشان می‌دهد. ضرایب یا روابط نامشخص با نماد - علامت‌گذاری شده‌اند. پیش‌بینی مقدار این ضرایب موضوع دو زیر بخش آتی است.

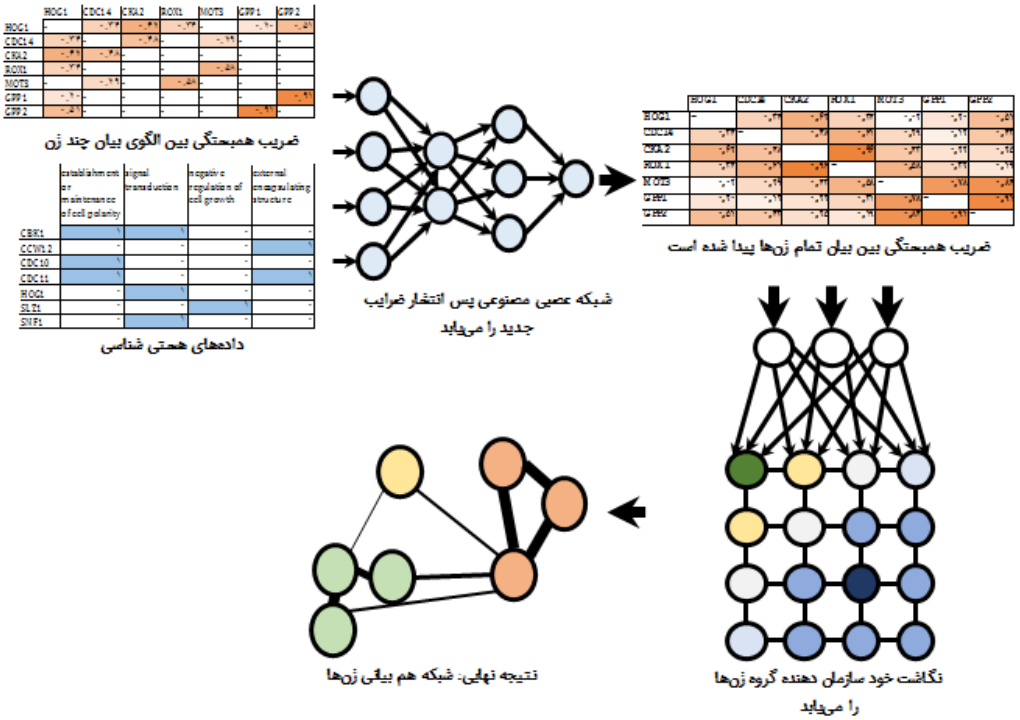
ما از یک شبکه عصبی مصنوعی با پس‌انتشار برای پیش‌بینی همبستگی‌های گزارش‌نشده و از یک نگاشت خودسازمان‌دهنده برای بالابردن کیفیت پیش‌بینی و ساخت شبکه هم‌بیانی استفاده کرده‌ایم. دو زیرقسمت پیش‌رو جزییات این دو شبکه را توصیف می‌کنند.



شکل ۱) بخش کوچکی از هستی‌شناسی فرآیندها

www.SID.ir

Archive of SID



شکل ۲) مراحل محاسباتی روش مطرح‌شده؛ این روش به دو نوع داده احتیاج دارد: ۱) همبستگی بین الگوی بیان چندین جفت ژن و ۲) اطلاعات هستی‌شناسی آنها؛ یک شبکه عصبی مصنوعی تلاش می‌کند تا رابطه بین این دو نوع داده را بیابد و از روی آن همبستگی‌های اعلام‌نشده را تشخیص دهد. یک شبکه عصبی مصنوعی دیگر به هر ژن یک گروه منتسب کرده و کیفیت نتایج شبکه عصبی اول را بالا می‌برد.

	establishment or maintenance of cell polarity	signal transduction	negative regulation of cell growth	external encapsulating structure
CBK1	۱	۱	۰	۰
CCW12	۰	۰	۰	۱
CDC10	۱	۰	۰	۰
CDC11	۱	۰	۰	۱
HOG1	۰	۱	۰	۰
SLZ1	۰	۰	۱	۰
SNF1	۰	۱	۰	۰

شکل ۳) یک ماتریس برچسب الگو با ۴ برچسب و ۷ ژن؛ این ماتریس عملکردها، فرآیندها و محل‌های حضور ژن‌های فهرست‌شده را نشان می‌دهد. برای نمونه، سلول‌ها ژن *CBK1*، *CDC10* و *CDC11* را برای حفظ قطبیت خود بیان می‌کنند. ژن *SLZ1* تاثیر منفی روی رشد دارد.

	HOG1	CDC14	CKA2	ROX1	MOT3	GPP1	GPP2
HOG1	-	۰/۳۴	۰/۶۲	۰/۳۴	-	۰/۲۰	۰/۵۱
CDC14	۰/۳۴	-	۰/۴۸	-	۰/۲۹	-	-
CKA2	۰/۶۲	۰/۴۸	-	-	-	-	-
ROX1	۰/۳۴	-	-	-	۰/۵۸	-	-
MOT3	-	۰/۲۹	-	۰/۵۸	-	-	-
GPP1	۰/۲۰	-	-	-	-	-	۰/۹۱
GPP2	۰/۵۱	-	-	-	-	۰/۹۱	-

شکل ۴) ماتریس همبستگی ۷ ژن؛ روابط تنظیمی ژن‌ها در این ماتریس ذخیره شده است. برای نمونه، *GPP1* و *GPP2* به‌صورت همزمان افزایش یا کاهش می‌یابند، زیرا ضریب همبستگی بین الگوی بیان این دو ژن، ۰/۹۱، بسیار زیاد است. از طرفی دیگر، ضریب همبستگی بین الگوی بیان *GPP1* و *HOG1* ناچیز است، بنابراین محصولات این دو ژن برای پاسخ به شرایط کاملاً متفاوتی ساخته می‌شود[27].

www.SID.ir

Archive of SID

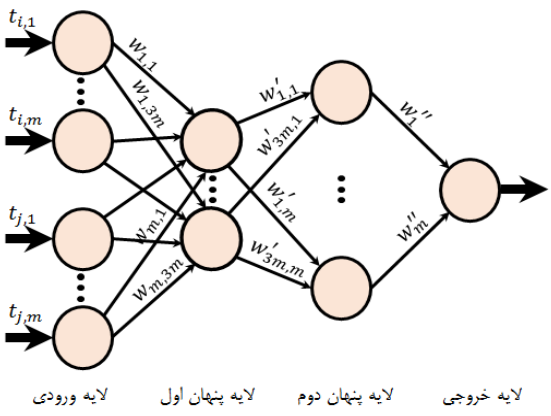
شبکه پس‌انتشار: یک شبکه عصبی مصنوعی یک شبیه‌سازی از شبکه عصبی زیستی ذهن است و از تعداد زیادی واحد موازی با نام سلول عصبی یا نورون ساخته شده است. نورون‌ها را می‌توان به‌وسیله ارتباطات وزن‌دار به یکدیگر متصل کرد. پیش از به‌کاربردن شبکه باید وزن‌ها را با یکی از دو رویه یادگیری با ناظر یا بدون ناظر تنظیم کرد. در رویه یادگیری، وزن‌های شبکه عصبی کم‌کم به‌گونه‌ای تنظیم می‌شوند که میانگین خطای شبکه، فاصله بین خروجی شبکه و مقداری که از آن انتظار می‌رود، کاهش یابد[28, 29].

در رویه یادگیری با ناظر، ورودی‌های آموزشی با پاسخ‌های معلوم متناوباً به شبکه داده می‌شود. در هر تکرار، رویه یادگیری وزن‌ها را به‌گونه‌ای تنظیم می‌کند که میزان خطا کاهش یابد و خروجی‌ها با دقت بالایی از روی ورودی‌ها تولید شوند[28, 30]. این رویه یادگیری برای تخمین توابع مناسب است[31]. ارایه ورودی‌ها به شبکه و تنظیم وزن‌ها تا جایی ادامه می‌یابد که یک شرط ازپیش‌تعیین‌شده محقق شود؛ برای مثال میانگین خطای تخمین خروجی از ۰/۰۰۱ کمتر شود.

هنگامی که دانشی درباره خروجی مطلوب نداریم، یادگیری بدون ناظر کاربردی است. به‌طور معمول طبقه‌بندی یا خوشه‌بندی داده به‌گونه‌ای که ورودی‌های مشابه براساس یک معیار مشخص در یک طبقه یا خوشه قرار بگیرند، هدف این رویه یادگیری است.

یک شبکه عصبی مصنوعی پس‌انتشار شبکه‌ای با چندین لایه از نورون‌ها است که رویه یادگیری آن با ناظر بوده و برای تخمین توابع و کشف روابط بین ورودی‌ها و خروجی‌ها مناسب است[29, 30, 32].

در این گونه از شبکه‌ها یک نورون به تک‌تک نورون‌های لایه بعدی به‌وسیله ارتباطات وزن‌دار متصل است. وزن‌ها در آغاز تصادفی بوده و در حین یادگیری به‌گونه‌ای تنظیم می‌شوند که خطای کل کمتر از یک حد آستانه‌ای شود. شکل ۵ شبکه طراحی‌شده را نمایش می‌دهد. این شبکه برچسب‌های منتسب با دو ژن ورودی، یعنی دو سطر از ماتریس T را دریافت کرده و ضریب همبستگی بین الگوی بیان آن دو ژن را تخمین می‌زند.



شکل ۵) ساختار شبکه عصبی مصنوعی پس انتشار طراحی‌شده

همان‌طور که نشان داده شده است، برای یک آزمایش با m برچسب، $2m$ نورون در لایه ورودی، $3m$ نورون در لایه اول مخفی، m نورون در لایه دوم مخفی، و یک نورون در لایه خروجی وجود دارد (نکته، چندین راهنما برای تعیین تعداد نورون‌های لایه‌های مخفی وجود دارد؛ منتهی، آزمون و خطا روش متداول یافتن آنها است).

آموزش شبکه: هدف از رویه یادگیری تنظیم وزن‌ها است به‌گونه‌ای که خروجی تخمین دقیقی از خروجی مطلوب باشد. در اینجا، بعد از اجرای رویه یادگیری روی داده‌های بیان و هستی‌شناسی موجود، شبکه می‌تواند ژن‌های هم‌بیان را از روی هستی‌شناسی آنها بیابد؛ به‌علاوه، می‌تواند ژن‌های هم‌بیان نوین را نیز بیابد. طبق تعریف رویه یادگیری تکرار این گام‌ها است[33]:

۱) محاسبه خروجی یا پاسخ نورون‌ها به ورودی جاری، لایه‌به‌لایه، از لایه ورودی به لایه خروجی؛ ورودی، یعنی ویژگی‌های منتسب‌شده به دو ژن، مثلاً ژن‌های p و q، که برابر است با بردار $[t_{p,1}, t_{p,2}, \dots, t_{p,m}, t_{q,1}, t_{q,2}, \dots, t_{q,m}]$ به نورون‌های لایه ورودی داده می‌شود. خروجی نورون iام که در لایه‌های پنهان یا خروجی قرار دارد برابر است با $f(\sum_{t=1}^{n} w_{j,i} o_j)$ که در آن o_j خروجی نورون jام لایه قبل و $w_{j,i}$ وزن بین این دو است. نکته اینکه وجود تابع فعالیت $f(x) = \frac{1}{1+e^{-x}}$ باعث می‌شود که شبکه بتواند روابط پیچیده بین ورودی و خروجی را بیابد. براساس نتایج، بدون آن دقت نهایی کاهش می‌یابد.

۲) تعیین میزان خطای تولیدشده، که برابر است با فاصله بین خروجی شبکه و خروجی مورد انتظار: $e = c_{p,q} - o$ که در آن o خروجی نورون لایه آخر و $c_{p,q}$ ضریب همبستگی بین بیان ژن‌های p و q است.

۳) پس‌انتشار خطا: تصحیح وزن‌های متصل به نورون‌ها با توجه به سهم هر یک در میزان خطای محاسبه‌شده، لایه‌به‌لایه، از لایه خروجی به لایه ورودی. میزان تغییرات لازم برای وزن بین نورون jام و نورون iام لایه قبل برابر است با $\Delta w_{i,j} = \alpha o_i \delta_j$ که در آن $\delta_j = f'(o_i) \sum_{j=1}^{n} \delta_j w_{i,j}$ سهم نورون iام لایه قبل در خطای به‌وجودآمده است. متغییر α که به آن نرخ یادگیری نیز گفته می‌شود میزان خطای تصحیح‌شده در هر تکرار را مشخص می‌کند ($\alpha = 0.1$ یک انتخاب مناسب است). پس از محاسبه تغییرات لازم، آنها اعمال می‌شوند: $w_{i,j} = w_{i,j} + \Delta w_{i,j}$.

در شبکه عصبی پس‌انتشار یک نورون لایه ورودی، به‌سادگی، ورودی خود را به لایه بعد می‌دهد. سایر نورون‌ها ورودی وزن‌دار خود را جمع، یک تابع فعالیت روی آن اعمال کرده و نتیجه را به‌عنوان خروجی خود گزارش می‌کنند.

همان‌طور که قبلاً گفته شد، هدف رویه یادگیری کاهش میزان خطا (فاصله بین خروجی تولیدشده و خروجی مورد انتظار) است. تغییر وزن‌ها با توجه به سهمی که در میزان خطای تولیدشده دارند رویه یادگیری را به سمت هدف خود سوق می‌دهد. رویه ارایه‌شده دو ورودی شامل ماتریس برچسب‌ها T و ماتریس همبستگی‌ها C دارد. این رویه متناوباً برچسب‌ها یا اطلاعات هستی‌شناسی ژن‌های p و q، یعنی بردار $[t_{p,1}, t_{p,2}, \dots, t_{p,m}, t_{q,1}, t_{q,2}, \dots, t_{q,m}]$ را به لایه ورودی می‌دهد تا واکنش شبکه نسبت به داده‌های آن دو ژن را

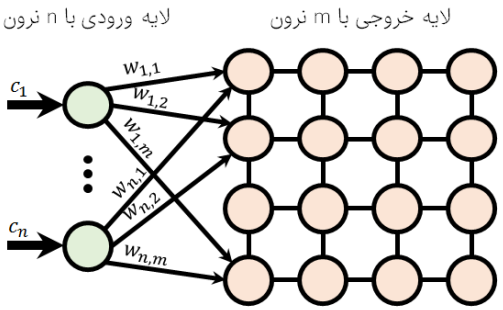
www.SID.ir

Archive of SID

بیابد. پاسخ یا واکنش شبکه در خروجی نورون لایه آخر قرار می‌گیرد. در ادامه رویه خطای به‌وجودآمده را محاسبه می‌کند: $e = c_{p,q} - o$. پس از آن تغییراتی که باعث کاهش خطا می‌شوند به دست آمده و اعمال می‌شوند. پس از پایان رویه یادگیری، شبکه پس‌انتشار می‌تواند روابط داده‌شده، یعنی درایه‌های مشخص ماتریس C، را با دقت بالایی بازتولید کند. بنابراین، پرکردن درایه‌های نامشخص ماتریس C با خروجی این رویه مناسب است.

ساخت شبکه هم‌بیانی ژن‌ها: تا اینجا، ماتریس همبستگی‌ها با مقادیر از قبل مشخص یا با مقادیر تازه کشف‌شده پر شده است. هدف این گام یافتن اعضای E و ساختن گراف $G = (V, E)$ از روی ماتریس پرشده C است. می‌دانیم که در شبکه‌های زیستی ژن‌ها در گروه‌هایی با تعداد زیادی رابطه قرار گرفته‌اند[34-36]. ژن‌های ویژه که تعداد زیادی رابطه دارند این گروه‌ها را به یکدیگر وصل می‌کنند. ویژگی گفته‌شده مشوق این قسمت از پژوهش بوده است. برای یافتن گروه تمامی ژن‌ها ما یک نگاشت خودسازمان‌دهنده را آموزش داده‌ایم.

همان‌طور که در شکل ۶ نشان داده شده است، نگاشت‌های خودسازمان‌دهنده شبکه‌های عصبی مصنوعی با دو لایه به هم متصل ورودی و خروجی هستند[28]. این گونه از شبکه‌های عصبی مصنوعی قابلیت طبقه‌بندی یا خوشه‌بندی داده‌ها براساس ویژگی‌هایشان را دارند. به اختصار، نگاشت طراحی‌شده ضرایب همبستگی محاسبه‌شده برای ژن‌ها، یعنی سطرهای ماتریس C را دریافت کرده و با یافتن سطرهای که رابطه‌شان با سایر سطرها همانند است، گروه‌های ژنی را تشکیل می‌دهد. بعداً، یافتن اعضای E، یعنی یال‌های شبکه همبستگی از روی درایه‌های ماتریس C ساده خواهد بود. جزییات این کار در ادامه بیان شده است.



شکل ۶) ساختار نگاشت خودسازمان‌دهنده طراحی‌شده برای یافتن گروه ژن‌ها؛ ضرایب همبستگی یک ژن، یک سطر از ماتریس C، به آن داده می‌شود. نورون لایه خروجی که کمترین فاصله را با آن ورودی دارد، برنده آن ورودی می‌شود. درنهایت وزن‌های نورون برنده و همسایه‌های آن تصحیح می‌شوند.

نورون‌های لایه ورودی به آنهایی که در لایه خروجی قرار دارند توسط یال‌های وزن‌دار متصل است. رویه یادگیری این وزن‌ها را در چندین تکرار به‌گونه‌ای تغییر می‌دهد که واکنش نورون‌های همسایه در لایه خروجی به ورودی‌های مشابه یکسان باشد.

طبق تعریف رویه یادگیری نگاشت خودسازمان‌دهنده تکرار این گام‌ها است:

۱) در ابتدا نورون برنده، نورونی که وزن‌های متصل به آن به ورودی جاری، یعنی روابط ژن p شباهت بیشتری دارد تعیین می‌شود. نورونی در لایه خروجی برنده است که این رابطه را کمینه کند:

$$\sqrt{\frac{\sum_{i\leftarrow 1}^{n}\left(c_{p,i}-w_{i,v}\right)^2}{n}}$$

۲) پس از یافتن نورون برنده، نورون vام وزن‌های آن نورون و تا حدی همسایه‌های نزدیک آن تصحیح می‌شود. تصحیح لازم برای وزن یال بین نورون iام لایه ورودی و jام لایه خروجی برابر است با $\Delta w_{i,j} = \alpha \times \delta(v,j) \times \left(t_{r,i} - w_{i,j}\right)$ که در آن α نرخ یادگیری بوده و تابع $\delta(v,j)$ فاصله بین نورون برنده، vام، تا نورون j ام را باز می‌گرداند. پس از محاسبه تغییرات لازم، آنها اعمال می‌شوند:

$$w_{i,j} = w_{i,j} + \Delta w_{i,j}.$$

در رویه بالا پس از یافتن نورون با کمترین فاصله از ورودی جاری، نورون برنده، وزن‌های آن نورون و تا حدی همسایه‌های نزدیک آن تصحیح می‌شود. تابع δ فاصله یک نورون در لایه خروجی از نورون برنده را بازمی‌گرداند. از آنجایی که نورون‌های همسایه در لایه خروجی به یک گروه ژنی مرتبط هستند، بنابراین، تعداد آنها m، باید از حداکثر تعداد گروه‌ها، n، بیشتر باشد. آزمایش با ورودی‌های مختلف نشان داد که $m = 50n$ یک انتخاب مناسب برای اکثر شرایط است.

پس از اتمام رویه یادگیری، ارایه ضرایب همبستگی ژن‌های متعلق به یک گروه باعث برنده‌شدن یک نورون خاص می‌شود. به‌علاوه، اکنون می‌توان وزن یال‌های شبکه هم‌بیانی ژن‌ها را با توجه به گروه ژن‌های آن و ضریب همبستگی بین الگوی بیان ژن‌ها تعیین کرد: برای هر $1 \leq i,j \leq n$، اگر ژن iام و jام در یک گروه هستند، به عبارتی دیگر، اگر ارایه سطر iام و jام ماتریس C به لایه ورودی باعث برنده‌شدن یک نورون خاص می‌شود، آنگاه $e_{i,j} \leftarrow c_{i,j}$، در غیر این صورت $e_{i,j} \leftarrow 0.75c_{i,j}$. درنهایت، یال‌های کم‌اهمیت را نیز می‌توان حذف کرد: به‌ازای هر $1 \leq i,j \leq n$، اگر $e_{i,j} < 0.33$ آنگاه $e_{i,j} \leftarrow 0$.

ثابت‌های انتخاب‌شده در اینجا برای تنظیم وزن‌های شبکه هم‌بیانی ژن‌ها، یعنی ۰/۷۵ و ۰/۳۳ باعث ایجاد بهترین نتایج روی داده‌های دریافت‌شده از پایگاه داده‌های YeastNet شده است[37]. این پایگاه داده‌ها ضرایب همبستگی بین بسیاری از ژن‌های مخمر را نگهداری می‌کند، دقیقاً ۳۶۱۹۸۶ ضریب همبستگی. همچنین، داده‌های آن از روی ۲۹۰۰۰ مقاله چاپ‌شده استخراج شده و بنابراین از نظر زیستی

www.SID.ir

Archive of SID

دقیق است. در نتیجه، ثابت‌های انتخاب ده باید برای سایر آزمایش‌ها نیز مناسب باشند.

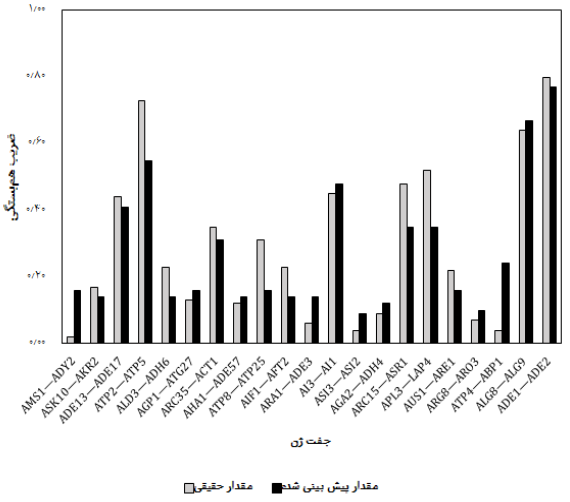
یافته‌ها

ما در این قسمت دقت و کارآیی روش مطرح‌شده را مورد ارزیابی قرار داده و نتایج آن را با نتایج موجود در پایگاه داده‌های زیستی مقایسه می‌کنیم. ارزیابی براساس نتایج به‌دست‌آمده از چندین آزمون است. در ابتدا ما مجموعه داده‌هایی که از آنها در آزمایش‌های مختلف استفاده کرده‌ایم را معرفی می‌کنیم. زیربخش دوم حاوی نتیجه آزمون‌های است که دقت تعیین ضرایب همبستگی را مد نظر قرار داده‌اند. همچنین دقت پیش‌بینی ضرایب همبستگی جدید نیز گزارش شده است. آزمون‌های زیربخش سوم میزان شباهت بین شبکه‌های هم‌بیانی ژنی تولیدشده و واقعی را اندازه‌گیری می‌کنند. درنهایت، نتیجه مقایسه روش مطرح‌شده با یک پنج روش مطرح دیگر در زیر بخش چهارم خواهد آمد.

مجموعه داده‌های استفاده‌شده: در این آزمایش‌ها روابط هم‌بیانی گونه مخمر مورد استفاده قرار گرفته است. پایگاه داده YeastNet ضریب همبستگی بین بیان ۶۴۰۰ ژن مخمر را در ارایه می‌کند. این پایگاه داده با جمع‌آوری نتایج آزمایشگاهی که در ۲۹۰۰۰ مقاله به چاپ رسیده‌اند و انجام برخی محاسبات توانسته است ۳۶۱۹۸۶ ضریب همبستگی با کیفیت را بیابد. حجم بسیار زیاد و پشتوانه زیستی که داده‌های این پایگاه داده دارند باعث پرکاربردشدن آن شده است. تعداد زیادی از پژوهش‌ها برای نشان‌دادن برتری روش خود از داده‌های این پایگاه داده استفاده کرده‌اند. برای نمونه *مارتینز* و همکاران چندین روش مختلف را روی داده‌های مستخرج از این پایگاه داده اجرا کرده و نتایج را گزارش کرده‌اند[38]. ما نیز ضرایب همبستگی بین ژن‌های گونه مخمر را از این پایگاه داده دریافت کرده‌ایم.

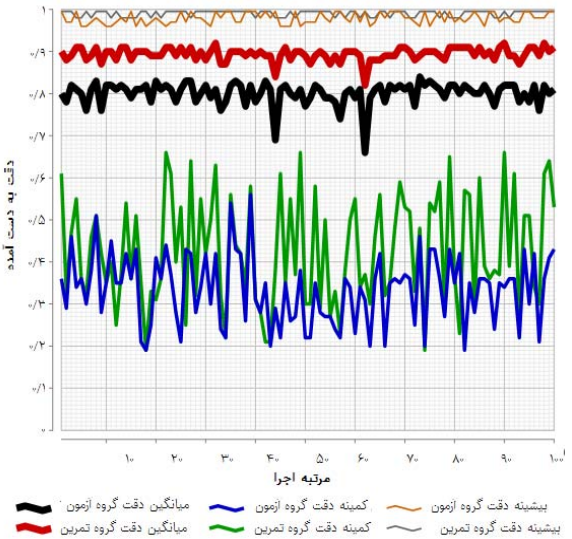
ویژگی‌های منتسب به ژن‌های مخمر از پایگاه داده پروژه هستی‌شناسی ژن‌ها به نشانی www.geneontology.org دریافت شده است. این پروژه که شرکت‌ها و موسسات بزرگی حامی آن هستند گردایه دانش موجود درباره ژن‌ها را به در خود دارد. برای نمونه این پایگاه داده ویژگی‌های که ژن‌های مخمر دارند را از روی نتایج ۵۳۲۴۴ آزمایش به دست آورده است. کیفیت بالای داده‌های ارایه‌شده موجب شده است که این پایگاه داده مرجع اصلی برای دریافت ویژگی‌های ژن‌ها یا اطلاعات هستی‌شناسی آنها باشد.

تعیین ضرایب همبستگی بین الگوی بیان ژن‌ها: در اولین آزمون، ۳۶۱۹۸۶ ضریب همبستگی موجود بین بیان ۶۴۰۰ ژن مخمر از YeastNet استخراج شد. برخی از آنها به‌صورت تصادفی انتخاب و به روش داده شد. این ضرایب همبستگی در کنار هم یک مجموعه را تشکیل می‌دهند که به آن مجموعه آموزشی می‌گوییم. سایر اعضای مجموعه ارزیابی خواهند بود و برای ارزیابی دقت روش از آنها استفاده می‌شود. در اینجا ما دقت را به‌صورت فاصله بین ضریب همبستگی تخمین زده شده و واقعی تعریف می‌کنیم. چند ضریب همبستگی به همراه مقدار تخمین زده‌شده توسط روش برای هر یک از آنها در نمودار ۱ آمده است. البته میانگین دقت تعیین ضرایب همبستگی برای اعضای مجموعه ارزیابی این آزمون ۰/۸۷ بوده است.



نمودار ۱) ضرایب همبستگی بین الگوی بیان ۲۱ جفت ژن؛ این نمونه‌ها از مجموعه ارزیابی برداشته شده‌اند.

برای کاهش وابستگی نتایج به ترکیب دو مجموعه آموزشی و ارزیابی، باید آزمون بالا را چندین بار تکرار و در هر تکرار ضرایب موجود را به‌صورت تصادفی به دو مجموعه یادشده افراز کرد. نمودار ۲ دقت‌های به‌دست‌آمده در ۱۰۰ تکرار را نشان می‌دهد. به‌صورت خلاصه، میانگین دقت تعیین ضرایب همبستگی برای اعضای مجموعه آموزشی و ارزیابی به‌ترتیب ۰/۸۹ و ۰/۸۰ بود. البته برای یک عضو از مجموعه آموزشی، دقت تعیین ضریب می‌تواند تا ۰/۲۱ کاهش یابد و تا ۰/۹۹ بالا رود. برای یک عضو از مجموعه ارزیابی حداقل دقت ۰/۱۹ و حداکثر آن ۰/۹۸ بوده است.



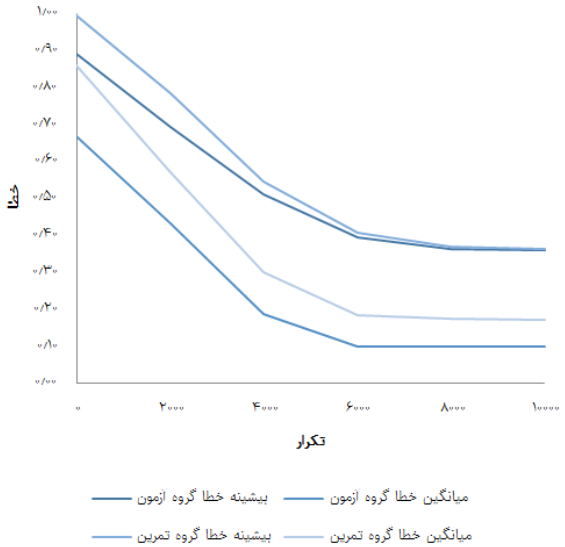
نمودار ۲) دقت‌های به‌دست‌آمده در چندین اجرای آزمون؛ میانگین، بدترین و بهترین نتیجه به‌دست‌آمده در هر اجرا به‌ترتیب با خطوط ضخیم، معمولی و نازک نشان داده شده است.

www.SID.ir

Archive of SID

برای رسیدن به نتایجی که از نظر آماری توجیه داشته باشند، باید رویه بالا را چندین بار تکرار کرد و در هر تکرار مجموعه آموزشی و ارزیابی را به‌صورت تصادفی ایجاد کرد[29, 32]. به همین دلیل در آزمون دوم ورودی ۲۰۰ سری از ضریب همبستگی‌ها است. هر سری بین ۲۰ تا ۵۰ عضو دارد که به‌صورت تصادفی از مجموعه ضرایب پایگاه داده‌های YeastNet انتخاب شده‌اند. این بار مجموعه ارزیابی هم سری فقط یک عضو دارد. یعنی، در هر سری تمام ضرایب به استثنای یکی، عضو مجموعه ارزیابی هستند و برای پیش‌بینی آن یک عضو مورد استفاده روش قرار می‌گیرند. این آزمون که به روش انجام آن اعتبارسنجی متقابل نیز می‌گویند نشان داد که دقت روش مطرح‌شده در تعیین یک عضو مجموعه ارزیابی هر سری ۰/۷۴ است. تعداد ضرایب همبستگی موجود در سومین آزمون نسبت به دو مورد قبلی بیشتر است: ورودی آزمون سوم ۴۱۲۲ ضریب همبستگی است. این ضرایب همبستگی بین الگوی بیان ۳۸۵ ژن گزارش شده است و باز هم به‌صورت تصادفی انتخاب و از پایگاه داده‌های YeastNet دریافت شده‌اند. این ضرایب به دو مجموعه آموزشی با ۳۲۹۸ عضو و ارزیابی با ۸۲۴ عضو افراز شده است.

نمودار ۳ خطای پیش‌بینی اعضای مجموعه آموزشی و ارزیابی در چندین گام از رویه یادگیری را نشان می‌دهد. باید توجه داشت که بعد از چندین تکرار رویه یادگیری، وزن‌ها ثابت می‌شوند و تکرارهای بعدی تقریباً دیگر اثری ندارند؛ البته، باید متذکر شد که با توجه به پژوهش‌های موجود، این رفتار برای شبکه‌های عصبی مصنوعی طبیعی است[29, 30, 32, 39].



نمودار ۳) تکرارهای متوالی رویه یادگیری شبکه عصبی پس انتشار وزن‌های شبکه را به‌گونه‌ای تنظیم می‌کند که اختلاف بین ضرایب همبستگی پیش‌بینی‌شده و واقعی کم شود. به همین دلیل میانگین و حداکثر خطای محاسبه در طول زمان کاهش می‌یابد.

ساخت شبکه هم‌بیانی ژن‌ها: هدف ما در این قسمت مقایسه شبکه‌های هم‌بیانی ساخته‌شده با کمک روش مطرح‌شده با هم ارزهای حقیقی خود به‌وسیله چندین شاخص آماری شناخته شده است. در این آزمون، ۱۵ ژن از مخمر در نظر گرفته شده است. این ژن‌ها را می‌توان به دو گروه با ویژگی‌های مختص به خود تقسیم کرد. با توجه به پایگاه داده هستی‌شناسی:

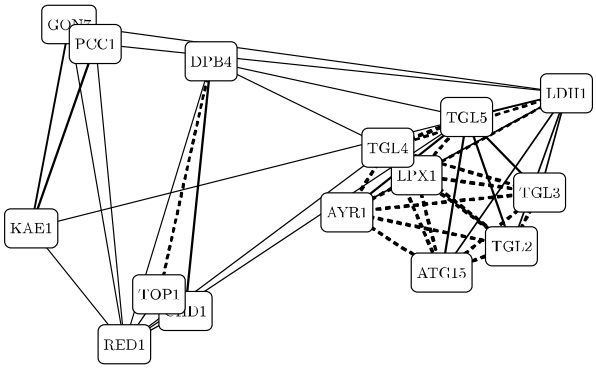
۱) محصولات ژن‌های *ATG15*، *AYR1*، *LDH1*، *LPX1*، *TGL2*، *TGL3*، *TGL4* و *TGL5* همگی به DNA متصل می‌شوند.

۲) محصولات ژن‌های *CHD1*، *DPB4*، *GON7*، *KAE1*، *PCC1*، *RED1* و *TOP1* همگی در شکاندن چربی‌ها نقش دارند.

باز هم دو فعالیتی که این ژن‌ها دارند به تصادف انتخاب شده‌اند. شبکه حاوی این ژن‌ها باید یک شبکه با دو بخش جدا از هم باشد، زیرا، این دو فعالیت کاملاً با هم تفاوت دارند و در نتیجه محصول ژن‌های قیدشده در شرایط کاملاً متفاوتی مورد نیاز هستند[24].

شکل ۷ شبکه ساخته‌شده برای این ژن‌ها را نشان می‌دهد. این شبکه با استفاده از دو نوع داده شامل داده‌های زیستی YeastNet و ضرایب همبستگی استنتاج‌شده توسط روش ما ساخته شده است. یال‌های که توسط هر دو منبع گزارش‌شده با خطوط منقطع و یال‌های نوینی که روش ما یافته با خطوط پیوسته نشان داده شده است.

آزمون‌های قبلی نشان دادند که روش مطرح‌شده می‌تواند با دقت قابل قبولی ضریب همبستگی بین الگوی بیان دو ژن را پیش‌بینی کند. بنابراین، در حالتی که ضریب همبستگی نامشخص است، استفاده از نتایج روش مطرح‌شده جایگزین خوبی است. شایان ذکر است که میانگین اختلاف بین وزن یال‌ها و مقادیر تخمین زده‌شده توسط روش ما ۰/۲۱ است. بنابراین، یال‌های جدید گزارش‌شده توسط روش ما باید ۰/۷۹=۰/۲۱-۱/۰۰ دقت داشته باشند.



شکل ۷) شبکه ۱۵ ژن با فعالیت‌های متفاوت *ATG15*، *AYR1*، *LDH1*، *LPX1*، *TGL2*، *TGL3*، *TGL4* و *TGL5* به DNA متصل می‌شوند؛ از طرفی دیگر *CHD1*، *DPB4*، *GON7*، *KAE1*، *PCC1*، *RED1* و *TOP1* در شکاندن چربی‌ها نقش دارند، بنابراین شبکه باید دوبخشی باشد.

برخی از معیارهای موجود به نتایج گسسته یا دودویی نیاز دارند. برای گسسته‌سازی نتایج، باید یک مقدار آستانه‌ای انتخاب و مقادیر کوچک‌تر و بزرگ‌تر از آن مقدار را به ترتیب با صفر و یک جایگزین کرد[31]. بعد از اجرای این گام می‌توان از دو معیار پرکاربرد که برای ارزیابی کارآیی روش‌ها ابداع شده‌اند، استفاده کرد. این دو معیار حساسیت (SN) و ویژگی (SP) هستند.

www.SID.ir

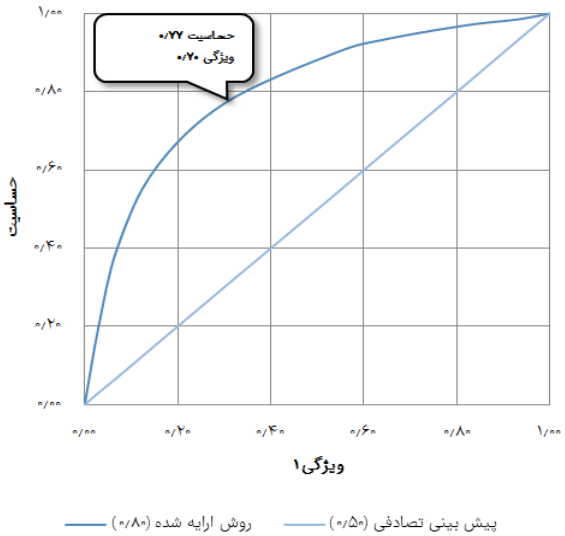
Archive of SID

حساسیت و ویژگی بر مبنای چهار آماره شامل مثبت حقیقی (TP)، مثبت کاذب (TN)، منفی حقیقی (FP)، منفی کاذب (FN) هستند: منفی حقیقی تعداد یال‌های است که در شبکه واقعی و ساخته‌شده وجود ندارند. از طرفی دیگر، مثبت کاذب تعداد یال‌های است که اشتباهاً به شبکه ساخته‌شده افزوده شده‌اند. مثبت حقیقی تعداد یال‌های پیش‌بینی‌شده و منفی کاذب تعداد یال‌های است که نادیده گرفته شده‌اند. اکنون می‌توان دو معیار حساسیت و ویژگی را به‌صورت رسمی معرفی کرد: حساسیت نسبت تعداد یال‌های صحیح است به کل یال‌ها: $SN = \frac{TP}{TP+FN}$؛ در مقابل، ویژگی کیفیت یال‌های رسم نشده‌شده یا نتایج منفی است: $SP = \frac{TN}{TN+FP}$.

بدیهی است که این کسرها و آماره‌ها همگی به آستانه انتخاب‌شده برای گسسته‌سازی بستگی دارند. در حالت کلی، یک مبادله بین حساسیت و ویژگی وجود دارد: افزایش مقدار آستانه‌ای باعث بالارفتن حساسیت و کاهش ویژگی می‌شود؛ کاستن از آن نتیجه عکس خواهد داشت.

تغییر مقدار آستانه‌ای و رسم حساسیت در مقابل ویژگی در حین این کار باعث به‌وجودآمدن منحنی می‌شود که به آن منحنی مشخصه عملکرد سیستم (ROC) می‌گویند و ابزاری پرکاربرد برای مقایسه روش‌های مختلف انجام یک کار است[31]. منحنی روش آرمانی از نقطه بالا سمت چپ نمودار گذر می‌کند. در عوض، خط اریب منحنی بدترین روش ممکن، یعنی روش تصادفی، است. سطح زیر منحنی نیز یک معیار موجز یا خلاصه برای بیان دقت روش‌ها است[32]. بدیهی است که سطح زیر منحنی بدترین روش ۰/۵ و بهترین روش ۱/۰ است.

در این آزمون، مجدداً، از پایگاه داده YeastNet استفاده شده است. این بار مجموعه‌های آموزشی و ارزیابی به‌ترتیب ۱۰۰۰ و ۱۰۰ ضریب همبستگی دارند. روش مطرح‌شده روی داده‌های مجموعه آموزشی اعمال و برای تعیین قدرت پیش‌بینی آن، منحنی مشخصه عملکرد آن تشکیل شد. نمودار ۴ این منحنی را نشان می‌دهد.



نمودار ۴) منحنی مشخصه عملکرد روش مطرح‌شده و روش تصادفی؛ سطح زیر منحنی داخل پرانتز نوشته شده است.

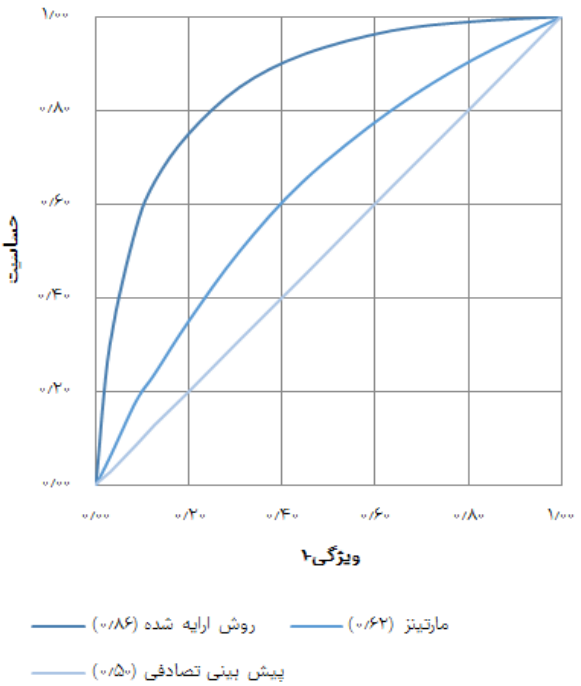
سطح زیر منحنی روش ما ۰/۸۰ است؛ بنابراین منحنی آن شباهت زیادی به منحنی بهترین روش دارد و کیفیت شبکه‌های ساخته‌شده توسط آن قابل قبول است[40]. این نتایج نشان می‌دهند که روش ما قابلیت یافتن روابط تنظیمی بین ژن‌ها را دارد.

مقایسه با سایر روش‌ها: *مارتینز* و همکاران[38] در یک تحقیق دقت پنج روش ساخت شبکه هم‌بیانی ژن‌ها را با هم مقایسه کردند که شامل موارد زیر هستند:

یک روش تکاملی که خودشان ابداع کرده‌اند[38]، یک روش مبتنی بر درخت تصمیم‌گیری[41]، یک روش رگرسیون درختی[42]، یک روش بصری احتمالاتی[43]، و یک الگوریتم بهینه‌سازی ترکیبیاتی[44].

با استفاده از چندین مقایسه، از جمله مقایسه روش‌های متفاوت روی داده‌های زیستی مرتبط با رشد سلول مخمر، نویسندگان به این نتیجه رسیدند که روش آنها بهترین روش برای استنتاج شبکه هم‌بیانی ژن‌ها است و حساسیت و ویژگی روش آنها بیشتر از سایر روش‌ها بوده است[38]. در اینجا ما روش مطرح‌شده را با روش آنها مقایسه می‌کنیم.

همان‌طور که گفته شد، معمولاً از سطح زیر منحنی برای مقایسه دقت روش‌های مختلف استفاده می‌شود و بیشتربودن سطح زیر منحنی به منزله دقت بیشتر است. نمودار ۵ منحنی‌های رسم‌شده براساس نتایج روش طرح‌شده در این مقاله و روش *مارتینز* و همکاران را نشان می‌دهد.



نمودار ۵) منحنی مشخصه عملکرد برای نتایج به‌دست‌آمده به‌وسیله روش مطرح‌شده و روش *مارتینز* و همکاران[38]، سطح زیر منحنی‌ها داخل پرانتز نوشته شده است.

همچنین در نمودار سطح زیر منحنی‌ها و یک نقطه نمونه از هر منحنی نشان داده شده است. همان‌طور که نمودار ۵ نشان می‌دهد،

www.SID.ir

Archive of SID

منحنی روش طرح‌شده در این مقاله به نقطه بالا سمت چپ نزدیک‌تر است، به‌علاوه، سطح زیر آن ۰/۸۶، از سطح زیر منحنی روش *مارتینز* و همکاران ۰/۶۲، بیشتر است. بنابراین، سطح زیر منحنی روش ما از هر پنج روش نیز بیشتر است و می‌توان گفت: شبکه‌های هم‌بیانی ژنی ساخته‌شده توسط روش ما در مقایسه با آنچه پنج روش نام‌برده‌شده می‌توانند ایجاد کنند، دقت بیشتری دارند و بیشتر شبیه شبکه‌های زیستی حقیقی خواهند بود.

از محدودیت‌های مطالعه حاضر وابسته‌بودن آن به یک گونه جانوری است. در مرحله یادگیری داده‌های یک گونه صرفاً مورد استفاده قرار گرفته است. احتمالاً بررسی چندین گونه جانوری که داده‌های آزمایشگاهی دقیق دارند و یادگیری رابطه بین ژن‌های آنها می‌تواند روش ارایه‌شده در این مطالعه را پرقدرت‌تر کند. تقسیم‌کردن داده‌های گردآوری‌شده بین چند شبکه عصبی به‌جای استفاده از یک شبکه عصبی می‌تواند باعث بالارفتن دقت و کارآیی شود. یافتن روشی برای بخش‌کردن داده‌ها به چند جزء برای دادن آنها به چند شبکه عصبی می‌تواند گام بعدی برای این مطالعه باشد.

نتیجه‌گیری

ما در این مقاله یک روش جدید برای استنتاج شبکه‌های هم‌بیانی ژن‌ها با ترکیب دو نوع داده معرفی کردیم. داده‌های مورد نیاز روش شامل موارد زیر است:

۱) اطلاعات هستی‌شناسی ژن‌ها که به‌سادگی یک مجموعه از برچسب‌ها است و خصایص ژن‌ها را نشان می‌دهند.

۲) میزان شباهتی که الگوی بیان ژن‌ها با یکدیگر دارند؛ به‌طور معمول از معیار ضریب همبستگی برای نشان‌دادن میزان شباهت الگوی بیان ژن‌ها استفاده می‌شود.

به‌صورت خلاصه، روش ما در ابتدا یک شبکه عصبی مصنوعی را آموزش می‌دهد تا رابطه بین ویژگی‌های منتسب‌شده به ژن‌ها و میزان شباهتی که الگوی بیان آنها با هم دارد را فرا بگیرد. پس از این مرحله، شبکه عصبی مصنوعی می‌تواند میزان شباهت بین الگوی بیان ژن‌ها را از روی ویژگی‌های منتسب‌شده به آنها تعیین کند. این موضوع یکی از برتری‌های روش ما نسبت به روش‌های موجود است.

در شبکه‌های زیستی ژن‌ها درون گروه‌های کوچک با تعداد زیادی رابطه قرار دارند. به همین دلیل روش ما در ادامه با کمک یک شبکه عصبی مصنوعی دیگر گروه‌های ژنی را یافته و دقت محاسبه روابط بین ژن‌ها را بالا می‌برد.

بر مبنای تحلیل‌های انجام‌شده روی داده‌های زیستی به‌دست‌آمده از پایگاه داده‌های معتبر و مقایسه‌های به‌عمل‌آمده با روش‌های موجود به این نتیجه رسیدیم که روش مطرح‌شده در این مقاله می‌تواند با دقت بالایی شباهت‌های موجود بین الگوی بیان ژن‌ها را بیابد و پیش‌بینی‌های آن برای شباهت‌های جدید هم از دقت بالای برخوردار است. بنابراین، یک برتری روش ما نسبت به سایر روش‌ها این است که افزودن خروجی آن به اطلاعات موجود درباره روابط بین ژن‌ها موجب بالارفتن دقت استنتاج شبکه‌های می‌شود. به همین دلیل شبکه‌های ساخته‌شده به‌وسیله این روش شباهت زیادی به شبکه‌های زیستی دارند. با توجه به نتایج با کیفیت به‌دست‌آمده با احتمال بالای روش ارایه‌شده در این مقاله روی گونه‌های که از نظر زیستی پیچیده‌تر هستند، برای مثال انسان، نیز قابل اعمال است. به همین جهت اعمال این روش روی داده‌های زیستی مرتبط با انسان و یافتن شبکه‌های دخیل در بیماری‌ها هدف بعدی ما است.

تشکر و قدردانی: از همکاری ارزشمند کارشناسان آزمایشگاه‌های دانشگاه اصفهان در پیشبرد این پژوهش تشکر و قدردانی می‌نمایم.

تاییدیه اخلاقی: موردی از سوی نویسنده گزارش نشده است.

تعارض منافع: هیچ گونه تعارضی در منافع وجود ندارد.

منابع مالی: موردی وجود ندارد.

منابع

1- Reece JB, Urry LA, Cain ML, Jackson RB, Wasserman SA, Minorsky PV. Campbell biology. 9th Edition. San Francisco: Benjamin Cummings; 2010.

2- Lodish HF, Berk A, Kaiser CA, Krieger M, Scott MP. Molecular cell biology. 6th Edition. New York: Macmillan Higher Education; 2007.

3- De Jong H. Modeling and simulation of genetic regulatory systems: A literature review. J Comput Biol. 2002;9(1):67-103.

4- Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: Data intgrtion in dynamic models-a review. Biosystems. 2009;96(1):86-103.

5- Reverter A, Chan EKF. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24(21):2491-7.

6- De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 2010;8(10):717-29.

7- Kabir M, Noman N, Iba H. Reverse engineering gene regulatory network from microarray data using linear time-variant model. BMC Bioinform. 2010;11(1):S56.

8- Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci. 2010;107(14):6286-91.

9- Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. PLoS One. 2012;7(1):e29348.

10- Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC Bioinform. 2007;8(6):S9.

11- Pavesi G, Valentini G. Classification of co-expressed genes from DNA regulatory regions. Inf Fusion. 2009;10(3):233-41.

12- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5(9):e12776.

13- Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis:

www.SID.ir

Archive of SID

memory. 3rd Edition. Heidelberg: Springer; 1989.
29- Shalev-Shwartz Sh, Ben-David Sh. Understanding machine learning: From theory to algorithms. 1st Edition. Cambridge: Cambridge University Press; 2014.
30- Alpaydin E. Introduction to machine learning. 2nd Edition. Cambridge: MIT Press; 2010.
31- Han J, Kamber M. Data mining: Concepts and techniques. 2nd Edition. Amsterdam: Elsevier; 2006.
32- Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd Edition. New York: Springer; 2009.
33- Werbos PJ. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Cambridge: Harvard University; 1974.
34- Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101-13.
35- Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organism. PLoS Biol. 2004;2(1):e9.
36- Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics. 2003;19(2):270-7.
37- Kim H, Shin J, Kim E, Kim H, Hwang S, Shim JE, et al. YeastNet v3: A public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. Nucleic Acids Res. 2013;42(D1):D731-6.
38- Martínez-Ballesteros M, Nepomuceno-Chamorro IA, Riquelme JC. Discovering gene association networks by multi-objective evolutionary quantitative association rules. J Comput Syst Sci. 2014;80(1):118-36.
39- Haykin S. Neural networks: A comprehensive foundation. 2nd Edition. New Jersey: Pearson Education Canada; 1998.
40- Soranzo N, Bianconi G, Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: Synthetic versus real data. Bioinformatics. 2007;23(13):1640-7.
41- Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol. 2003;4(1):R6.
42- Nepomuceno-Chamorro IA, Aguilar-Ruiz JS, Riquelme JC. Inferring gene regression networks with model trees. BMC Bioinform. 2010;11:517.
43- Bulashevska S, Eils R. Inferring genetic regulatory logic from expression data. Bioinformatics. 2005;21(11):2706-13.
44- Ponzoni I, Azuaje F, Augusto J, Glass D. Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. IEEE ACM Trans Comput Biol Bioinform. 2007;4(4):624-34.
Comparison and applications. BMC Syst Biol. 2010;4:8.
14- Raman K. Construction and analysis of protein-protein interaction networks. Autom Exp. 2010;2(1):2.
15- Martin Sh, Zhang Z, Martino A, Faulon JL. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics. 2007;23(7):866-74.
16- Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004;117(2):185-98.
17- Mahdevar G, Nowzari-Dalini A, Sadeghi M. Inferring gene correlation networks from transcription factor binding sites. Genes Genet Syst. 2013;88(5):301-9.
18- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nat Genet. 2000;25(1):25-9.
19- Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9(7):509-15.
20- Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinform. 2004;5(1):18.
21- Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinform. 2007;8(1):299.
22- Roy S, Bhattacharyya DK, Kalita JK. Reconstruction of gene co-expression network from microarray data using local expression patterns. BMC Bioinform. 2014;15(Suppl 7):S10.
23- Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, et al. Correlation between gene expression and GO semantic similarity. IEEE ACM Trans Comput Biol Bioinform. 2005;2(4):330-8.
24- Wang H, Azuaje F, Bodenreider O, Dopazo J. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 7-8 October, 2004, La Jolla, California, USA. Piscataway: IEEE; 2004. pp. 25-31.
25- Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res. 1999;11:95-130.
26- Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274-81.
27- Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249-55.
28- Kohonen T. Self-organization and associative

www.SID.ir