Modares Journal of Biotechnology. 2018;9(2):259-265

Introduction of *GRAP* Gene as Alzheimer's Disease Candidate Gene Using Microarray Data Analysis

ARTICLE INFO

Article Type
Original Research

Authors
Paylakhi S.Z. [1] *MSc*,
Ozgoli S.* *PhD*,
Paylakhi S.H.[2] *PhD*

How to cite this article
Paylakhi S Z, Ozgoli S, Paylakhi S H. Introduction of *GRAP* Gene as Alzheimer's Disease Candidate Gene Using Microarray Data Analysis. Modares Journal of Biotechnology. 2018;9(2):259-265.

ABSTRACT

Aims One of the most important areas in medical research is the identification of disease-causing genes, which helps the identification of mechanisms underlying disease and as a result helps the early diagnosis of disease and the better treatment. In recent years, microarray technology has assisted biologists to gain a better understanding of cellular processes. To this end, the application of efficient methods in microarray data analysis is very important. The aim of this study was the introduction of *GRAP* Gene as Alzheimer's disease candidate gene using microarray data analysis.

Materials & Methods In the present bioinformatic study, which was conducted on an Alzheimer's microarray data set containing 12990 genes, 15 patients, and 16 healthy subjects, by combining Fisher, Significance Analysis of Microarray (SAM), and Particle Swarm Optimization (PSO) methods as well as Classification and Regression Tree (CART), a new method was presented for analyzing microarray gene expression data to identify genes involved in Alzheimer's incidence.

Findings The accuracy level of the proposed method was 90.32% and the interpretation of the results from a biological point of view indicated that the proposed method has worked well; finally, the proposed method introduced 4 genes, of which, until now, 3 genes (75%) have been reported in biological studies as genes that cause Alzheimer's disease.

Conclusion In addition to proposing a new feature selection method for the analysis of microarray data, this study has introduced a new gene (*GRAP*) as a candidate gene related to Alzheimer's disease.

Keywords Microarray; Decision Tree; Machine Learning

CITATION LINKS

[1] Microarray technology [2] A comparative study of different machine learning methods on microarray gene expression data [3] Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses [4] Identification of disease-causing genes using microarray data mining and Gene Ontology [5] Identification of pseudo-periodic gene expression profiles [6] A new optimizer using particle swarm theory [7] On classification and regression trees for multiple responses and its application [8] The introduction and application of classification tree model for determination of risk factor for esophageal cancer in Golestan province [9] Applied data mining [10] Effect of ultrasonic treatment on the recovery ... [11] Deficiency of patched 1-induced Gli1 signal transduction results in astrogenesis in Swedish mutated APP transgenic mice [12] CREB expression mediates amyloid β-induced basal BDNF downregulation [13] Fernandez-de-Cossio J, Bringas R. BisoGenet: A new tool for gene network building, visualization and analysis [14] Role of TGFβ signaling in the pathogenesis of Alzheimer's disease [15] Epigenetic mechanisms underlying cognitive impairment and Alzheimer disease hallmarks in 5XFAD mice [16] A chemical with proven clinical safety rescues Down-syndrome-related phenotypes in through DYRK1A inhibition [17] The mitochondrial-derived peptide humanin activates the ERK1/2, AKT, and STAT3 signaling pathways and has age-dependent signaling differences in the hippocampus [18] RNA interference silencing of the adaptor molecules ShcC and Fe65 differentially affect amyloid precursor protein processing and Abeta generation [19] Distribution of fibroblast growth factor receptor-1 (FGFR-1) and FGFR-3 in the hippocampus of patients with Alzheimer's disease [20] Fibroblast growth factor receptor signaling affects development and function of dopamine neurons - inhibition results in a schizophrenia-like syndrome in transgenic mice [21] A common brain network links development, aging, and vulnerability to disease [22] Gap junctions and neurological disorders of the central nervous system [23] Tumor diagnosis preceding Alzheimer's disease onset: Is there a link between cancer and Alzheimer's disease? [24] Grap is a novel SH3-SH2-SH3 adaptor protein that couples tyrosine kinases to the Ras pathway [25] Elevated expression of p21ras is an early event in Alzheimer's disease and precedes neurofibrillary degeneration [26] Transformation by oncogenic ras-p21 alters the processing and subcellular localization of the lysosomal protease cathepsin D [27] Tissue-based Alzheimer gene exp-ression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets biomarker sets

*Control Department, Electrical & Computer Engineering Faculty, Tarbiat Modares University, Tehran, Iran
[1]Control Department, Electrical & Computer Engineering Faculty, Tarbiat Modares University, Tehran, Iran
[2]Ophthalmology Department, Ophthalmology Faculty, University of California, San Francisco, United States

Correspondence
Address: Tarbiat Modares University, Nasr Bridge, Jalal-Al-Ahmad Highway, Tehran, Iran
Phone: +98 (21) 82883309
Fax: +98 (21) 82884325
ozgoli@modares.ac.ir

Article History
Received: October 25, 2016
Accepted: December 11, 2017
ePublished: June 21, 2018

Copyright© 2018, 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.

معرفی ژن *GRAP* به‌عنوان ژن نامزد عامل آلزایمر با استفاده از تحلیل داده‌های ریزآرایه

سیده‌زهرا پای‌لاخی MSc
گروه کنترل، دانشکده برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
سجاد ازگلی* PhD
گروه کنترل، دانشکده برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
سیدحسن پای‌لاخی PhD
گروه چشم‌پزشکی، دانشکده چشم‌پزشکی، دانشگاه کالیفرنیا، سانفرانسیسکو، ایالات متحده

چکیده
اهداف: شناسایی ژن‌های دخیل در بروز یک بیماری، یکی از حوزه‌های مهم تحقیقات پزشکی است که به تشخیص مکانیزم بیماری و در پی آن تشخیص به‌موقع و درمان بهتر بیماری کمک می‌کند. در سال‌های اخیر فناوری ریزآرایه به دانشمندان علوم زیستی برای فهم فرآیندهای سلولی کمک شایانی کرده است. بدین منظور استفاده از روش‌های کارآمد در تحلیل داده‌های ریزآرایه بسیار کلیدی است. هدف مطالعه حاضر معرفی ژن *GRAP* به‌عنوان ژن نامزد عامل آلزایمر با استفاده از تحلیل داده‌های ریزآرایه بود.
مواد و روش‌ها: در مطالعه بیوانفورماتیکی حاضر که روی یک مجموعه داده ریزآرایه مربوط به آلزایمر شامل ۱۲۹۹۰ ژن، ۱۵ فرد بیمار و ۱۶ فرد سالم صورت گرفت، با ترکیب روش‌های فیشر، تجزیه و تحلیل اهمیت میکروآرایه (SAM) و الگوریتم بهینه‌سازی ازدحام ذرات (PSO) و با استفاده از روش طبقه‌بندی و رگرسیون مبتنی بر درخت تصمیم (CART)، روش جدیدی به‌منظور تحلیل داده‌های بیان ژن ریزآرایه، برای شناسایی ژن‌های دخیل در بروز آلزایمر ارایه شد.
یافته‌ها: سطح دقت مدل به‌دست‌آمده ۹۰/۳۲% بود و ارزیابی نتایج از دیدگاه زیستی نشان داد که روش پیشنهادی موفق عمل کرده و در نهایت ۴ ژن ارایه کرده است که ۳ ژن از این ۴ ژن (۷۵%)، تاکنون به‌عنوان ژن‌های دخیل در آلزایمر در منابع زیستی معتبر گزارش شده‌اند.
نتیجه‌گیری: این مطالعه علاوه بر ارایه یک روش انتخاب ویژگی جدید و تلفیقی برای تحلیل داده‌های ریزآرایه، یک ژن جدید (*GRAP*) به‌عنوان ژن نامزد مرتبط با آلزایمر معرفی کرده است.
کلیدواژه‌ها: ریزآرایه، درخت تصمیم، یادگیری ماشین

تاریخ دریافت: ۱۳۹۵/۰۸/۰۴
تاریخ پذیرش: ۱۳۹۶/۰۹/۲۰
*نویسنده مسئول: ozgoli@modares.ac.ir

مقدمه

از آنجا که در حال حاضر انواع بیماری‌ها با علایم و علل خاص در انسان‌ها بروز پیدا می‌کنند و افراد بسیاری با بیماری‌های مختلف دست‌وپنجه نرم می‌کنند، شناسایی ژن‌های دخیل در بیماری‌های مختلف که منجر به پیشگیری، تشخیص به‌موقع بیماری، درمان بهتر و پیش‌بینی بازگشت مجدد بیماری می‌شود، تبدیل به یکی از مهم‌ترین حوزه‌های تحقیقات پزشکی شده است. راهکارها و روش‌های بسیاری در این زمینه ارایه و توسط محققان بررسی شده است.

نتایج نشان می‌دهد که یکی از بهترین و دقیق‌ترین روش‌ها در این زمینه برای شناسایی این گونه ژن‌ها، پایش مقادیر بیان ژن افراد (نمونه‌ها) مختلف با استفاده از فناوری ریزآرایه است. فناوری ریزآرایه امکان پایش بیان هزاران ژن به‌صورت همزمان را فراهم کرده و این باعث تحول عظیمی در فهم ما از پاتوفیزیولوژی بسیاری از بیماری‌ها شده است[1].

برای بررسی حجم وسیع اطلاعات تولیدشده از فناوری ریزآرایه و استخراج اطلاعات مفید از آنها، در سال‌های اخیر تلاش‌های فراوانی صورت گرفته‌است. تحلیل داده‌های بیان ژن به‌دست‌آمده از ریزآرایه بسیار مهم است و این داده‌ها بدون وجود ابزاری برای تفسیر آنها بدون استفاده خواهند بود.

تحلیل پایه داده‌های بیان ژن شامل یک مرحله پیش‌پردازش و آماده‌سازی مجموعه داده برای مراحل تحلیل سطوح بالاتر مانند خوشه‌بندی یا دسته‌بندی است و انتخاب ویژگی یکی از مراحل مهم در پیش‌پردازش داده‌ها است. انتخاب ویژگی فرآیندی است که در آن، هدف رسیدن به بهترین زیرمجموعه از مجموعه ویژگی‌ها است. این هدف به کمک شناسایی ویژگی‌های حاوی اطلاعات و مرتبط با ویژگی برچسب و همچنین حذف ویژگی‌های زاید و اضافی محقق می‌شود.

در یک تقسیم‌بندی کلی، روش‌های انتخاب ویژگی به ۳ مقوله تقسیم می‌شوند و در هر روش به یک تابع معیار نیاز است تا بتوان مناسب‌بودن انتخاب ویژگی‌ها را ارزیابی نمود. اگر تابع معیار مبتنی بر دسته‌بندی‌کننده باشد (مثلاً دقت دسته‌بندی)، روش دسته‌بندی (Wrapper) و اگر مستقل از دسته‌بندی‌کننده باشد، روش فیلتر نامیده می‌شود و در نهایت اگر از معیارهای ارزیابی مختلف در مراحل متفاوت جستجو استفاده شود، روش تعبیه‌شده (embedded) نامیده می‌شود[2].

تاکنون روش‌های بسیاری برای انتخاب ویژگی مورد بررسی قرار گرفته‌اند. اگرچه نقاط قوت روش‌های پیشنهادی در تحقیقات مختلف ثابت شده است، اما محققان معتقدند تاکنون هیچ روشی به‌عنوان بهترین روش شناخته نشده است، به عبارتی تمام روش‌های ارایه‌شده در کنار نقاط قوت خود، نقطه ضعف‌هایی را نیز شامل می‌شوند که باعث ادامه تلاش محققان برای یافتن روش‌های جدیدی شده است که به‌طور عمده ترکیبی از روش‌های پیشین هستند. در حقیقت هدف محققان از یافتن روش‌های جدید، ترکیب روش‌های موجود و استفاده از نقاط قوت آنها است.

ایده روش‌های ترکیبی بدین صورت است که در ابتدا برای کاهش سریع ویژگی‌های ورودی از روش‌های فیلتر استفاده می‌کنند و سپس روش‌های Wrapper را به‌منظور انتخاب ویژگی‌های مهم روی خروجی مرحله اول اعمال می‌نمایند.

در این مقاله نیز روشی نوین برای انتخاب ویژگی ارایه شده است که ترکیبی از روش‌های فیشر، تجزیه و تحلیل اهمیت میکروآرایه (SAM؛ نمونه‌ای از روش فیلتر) و الگوریتم بهینه‌سازی ازدحام ذرات (PSO ؛نمونه‌ای از روش Wrapper) است. بدین ترتیب که ابتدا از معیار فیشر برای کاهش اولیه ابعاد مجموعه داده ورودی یا به عبارتی حذف ژن‌های نویزی و نامربوط از مجموعه داده اولیه استفاده می‌شود. سپس روش SAM به‌منظور انتخاب ژن‌های حاوی اطلاعات و مهم، پیاده‌سازی می‌شود. در مرحله بعد PSO روی خروجی مرحله پیشین اعمال شده و در نتیجه ژن‌های مهم و اساسی در بروز بیماری انتخاب می‌شوند. در نهایت نیز با استفاده از روش طبقه‌بندی و رگرسیون مبتنی بر درخت تصمیم (CART) به یادگیری مدل بر مبنای ژن‌های انتخاب‌شده از مرحله قبل پرداخته می‌شود. در واقع از مزیت‌های روش‌های فیلتر و Wrapper به‌طور همزمان استفاده می‌شود.

هدف مطالعه حاضر، معرفی ژن پروتئین آداپتور وابسته به GRB2 یا پروتئین شمارهٔ ۲ متصل به گیرنده فاکتور رشد (*GRAP*) به‌عنوان ژن نامزد عامل آلزایمر با استفاده از تحلیل داده‌های ریزآرایه بود.

مواد و روش‌ها

مطالعه بیوانفورماتیکی حاضر روی یک مجموعه داده ریزآرایه مربوط به آلزایمر شامل ۱۲۹۹۰ ژن، ۱۵ فرد بیمار و ۱۶ فرد سالم

صورت گرفت[3].

معیار فیشر: معیار فیشر که به‌صورت فرمول زیر نمایش داده می‌شود از جمله روش‌های فیلتری است که به‌طور گسترده برای تحلیل داده‌های ریزآرایه مورد استفاده قرار گرفته و نتایج مطلوبی را نیز ارایه کرده است:

$$Fisher(g) = \frac{(m_1(g)-m_2(g))^2}{(s_1^2(g)+s_2^2(g))}$$

که m_1 و m_2 میانگین ژن g و s_1 و s_2 واریانس ژن g در برچسب‌های ۱ و ۲ (سالم و بیمار) را نشان می‌دهند. این روش به‌منظور حذف ژن‌های زاید و نویزی موجود در داده‌های ریزآرایه مورد استفاده قرار می‌گیرد[4].

روش SAM: روش SAM مخصوص تحلیل مجموعه داده‌های ریزآرایه ارایه‌شده است. در واقع به‌منظور رفع مشکلی که بالابودن حجم داده‌های ریزآرایه در تحلیل این سری داده‌ها ایجاد می‌کند، در سال ۲۰۰۱ روش SAM پیشنهاد شد[5]:

$$d(x) = \frac{x_+ - x_-}{s_0 + \sqrt{\frac{1}{n_+}+\frac{1}{n_-}}\sqrt{\frac{(\sum_{i_+}(x_i-x_+)^2+\sum_{i_-}(x_i-x_-)^2}{n_+ + n_- - 2}}}$$

که $i_{+/-}$ شاخص‌ها را نمایش می‌دهد و $x_{+/-}$ نیز بیانگر میانگین تمام نمونه‌های متعلق به برچسب‌های مثبت و منفی است و S_0 نیز یک پارامتر اصلاح‌کننده برای کنترل اثر واریانس است.

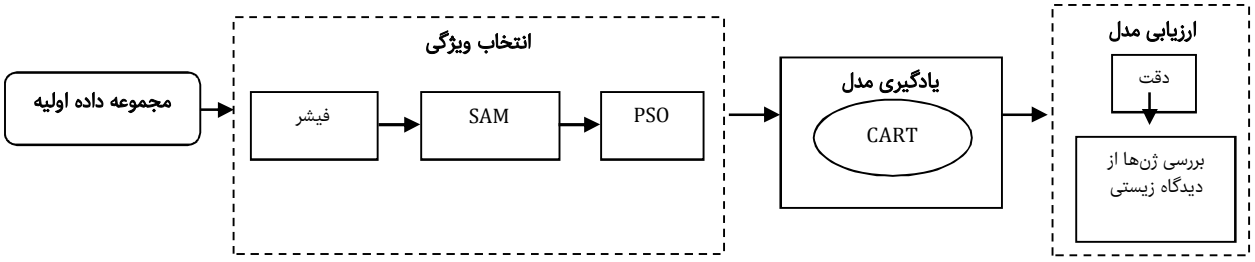
الگوریتم PSO: الگوریتم PSO یکی از مهم‌ترین الگوریتم‌های بهینه‌سازی هوشمند است که در حوزه هوش ازدحامی قرار می‌گیرد. این الگوریتم در سال ۱۹۹۵ توسط *کندی* و *ابرهارت* معرفی شد و با الهام از رفتار اجتماعی حیواناتی چون ماهی‌ها و پرندگان که در گروه‌های کوچک و بزرگ کنار هم زندگی می‌کنند، طراحی شد. الگوریتم PSO برای انواع مسایل پیوسته و گسسته مناسب است و پاسخ‌های بسیار مناسبی برای مسایل بهینه‌سازی مختلف داده است[6].

درخت تصمیم و مدل رده‌بندی درختی: درخت تصمیم یکی از روش‌های ناپارامتری رده‌بندی‌کردن است که با توجه به نوع متغیر وابسته به دو دسته رده‌بندی درختی برای متغیر دسته‌ای و رگرسیون درختی برای متغیر پیوسته تقسیم می‌شود. رده‌بندی درختی در راستای روش‌هایی نظیر تحلیل ممیزی (تابع تشخیص) و رگرسیون لجستیک است. در این روش مجموعه‌ای از شرط‌های منطقی به‌صورت یک الگوریتم با ساختار درختی برای رده‌بندی یا پیش‌بینی یک پیامد به کار می‌رود[7]. با توجه به این که شاخص‌های متنوع و روش‌های گوناگونی برای تعیین درخت تصمیم معرفی شده است، الگوریتم‌های متنوع و گوناگونی نیز ارایه شده است که یکی از مهم‌ترین و شناخته‌شده‌ترین آنها، الگوریتم CART است.

الگوریتمCART: الگوریتم CART که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌شود، توسط *بهنام‌پور* و همکاران در سال ۱۹۸۴ به‌طور کامل معرفی شد. این روش برای متغیرهای کمی طراحی شده ولی برای هر نوع متغیری قابل استفاده است. در این روش و برای متغیر پاسخ کیفی، شاخص جینی به‌عنوان معیاری برای انتخاب متغیرهای مناسب معرفی شده‌است.

مدل CART را می‌توان به‌عنوان یکی از شناخته‌شده‌ترین الگوهای رده‌بندی به‌منظور تشخیص و پیش‌گویی در علوم پزشکی برشمرد[8].

روش پیشنهادی این مطالعه: شماتیک ساده‌ای از مراحل پیاده‌سازی روش پیشنهادی نمایش داده شد (شکل ۱). روند کار بدین صورت بود که ابتدا روش انتخاب ویژگی با ترکیب سه روش فیشر، SAM و PSO روی مجموعه داده ورودی پیاده‌سازی شد، سپس یادگیری مدل با استفاده از روش دسته‌بندی‌کننده CART براساس خروجی مرحله اول صورت پذیرفت و در نهایت به ارزیابی مدل پرداخته شد.



شکل ۱) شماتیک مراحل پیاده‌سازی روش پیشنهادی

دقت دسته‌بندی با استفاده از فرمول زیر به‌دست می‌آید:

$$CA = \frac{TP + TN}{TP + FP + TN + FN}$$

TN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی منفی تشخیص داده است.

TP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی مثبت تشخیص داده است.

FP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه مثبت تشخیص داده است.

FN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه منفی تشخیص داده است[9].

برای پیاده‌سازی روش پیشنهادی از نرم‌افزار Rapid minder 5®، برای رسم شبکه برهم‌کنش‌های پروتئین- پروتئین (PPI) مرتبط با ۴ ژن از نرم‌افزار BisoGenet 3.4.0، برای بررسی PPI بین ژن‌های انتخاب‌شده از نرم‌افزار DAVID 6.8® و برای دسته‌بندی داده‌ها با انتخاب روش خواسته‌شده، از نرم‌افزار Rapid minder 5® استفاده شد. از آنجا که یک مرحله مهم در الگوریتم‌های داده‌کاوی مرحله آموزش است، بنابراین نرم‌افزار، داده‌ها را به دو دسته داده‌های آموزش و آزمون تقسیم می‌کند.

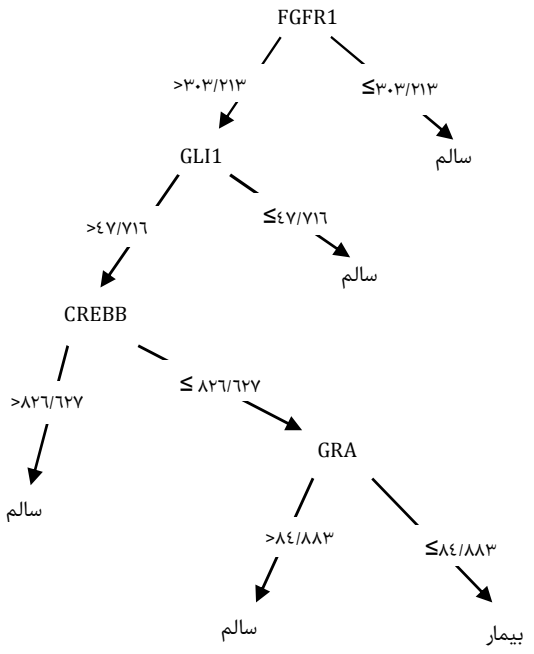
چون در هر بار اجرا در نرم‌افزار Rapid minder 5® نتایج متفاوتی به‌دست می‌آید، می‌توان از روش 10F-CV استفاده کرد. در این روش کل مجموعه داده به ۱۰ قسمت مساوی تقسیم شده و الگوریتم ۱۰ بار اجرا می‌شود. در هر بار ۹ قسمت از این داده‌ها را برای آموزش استفاده می‌کند و یک قسمت باقی‌مانده را به‌عنوان داده آزمون برای ارزیابی‌کردن الگوریتم در نظر می‌گیرد. میانگین نتایجی که به دست می‌آید به‌عنوان نتیجه نهایی گزارش می‌شود[9]. مهم‌ترین معیار برای تعین کارآیی یک الگوریتم دسته‌بندی، دقت یا نرخ دسته‌بندی (CA) است که نشان می‌دهد مدل طراحی‌شده چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است. دقت مدل یا نرخ دسته‌بندی با توجه به ماتریس درهم‌ریختگی (CM) محاسبه شد (جدول ۱).

جدول ۱) ماتریس درهم‌ریختگی

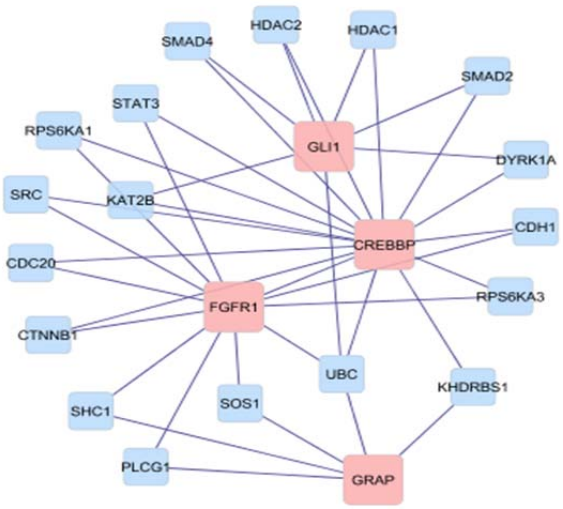
	رکوردهای پیش‌بینی‌شده	
رکوردهای واقعی	دسته −	دسته +
دسته +	FN	TP
دسته −	TN	FP

یافته‌ها

مدل رده‌بندی درختی به‌دست‌آمده با استفاده از روش پیشنهادی نمایش داده شد (شکل ۲) و در ماتریس درهم‌ریختگی، دقت این مدل ۹۰/۳۲% محاسبه شد (جدول ۲). هریک از الگوریتم‌های انتخاب ویژگی باعث بهبود سطح دقت دسته‌بندی‌کننده شدند (جدول ۳) و شبکه PPI مرتبط با ۴ ژن مذکور رسم شد (شکل ۳). سطح معنی‌داری مسیرهای بیولوژیکی (موجود در پایگاه داده KEGG) مرتبط با گیرنده فاکتور رشد فیبروبلاست (*FGFR1*)، خانواده انگشت روی (*GLI1*)، پروتئین متصل‌شونده CREB (*CREBBP*) و *GRAP* ارایه شد (جدول ۴). همچنین نتایج روش پیشنهادی این مطالعه با دیگر روش‌های استفاده‌شده مقایسه شدند (جدول ۵).



شکل ۲) مدل رده‌بندی درختی به‌دست‌آمده با استفاده از روش پیشنهادی



شکل ۳) شبکه برهم‌کنش‌های پروتئین- پروتئین مرتبط با ۴ ژن *FGFR1*، *GLI1*، *CREBBP* و *GRAP*

جدول ۱) ماتریس درهم‌ریختگی مرتبط با مدل به‌دست‌آمده

−	**بیمار** (واقعی)	**سالم** (واقعی)	**دقت کلاس** (%)
بیمار (پیش‌بینی)	۱۳	۱	۹۲/۸۶
سالم (پیش‌بینی)	۲	۱۵	۸۸/۲۴
یادآوری کلاس (%)	۸۶/۶۷	۹۳/۷۵	−

جدول ۳) مقایسه نتایج پیاده‌سازی روش‌های مختلف انتخاب ژن

روش‌ها	**دقت** (%)
دسته‌بندی بدون پیاده‌سازی الگوریتم انتخاب ویژگی	۴۳/۸۹
دسته‌بندی پس از پیاده‌سازی معیار فیشر	۶۱/۵
SAMدسته‌بندی پس از پیاده‌سازی معیار فیشر +	۸۱/۹۸
PSO + SAMدسته‌بندی پس از پیاده‌سازی فیشر +	۹۰/۳۲

جدول ۴) مسیرهای بیولوژیک مرتبط با ۴ ژن *FGFR1*، *GLI1*، *CREBBP* و *GRA*

مسیرها	**تعداد ژن‌ها**	**اسامی ژن‌ها**	**سطح معنی‌داری**
Pathways in cancer (bta05200)	۳	*FGFR1, CREBB , GLI*	۰/۰۰۳۸
Adherens junction (bta04520)	۲	*FGFR1, CREBBP*	۰/۰۲۶۹
Prostate cancer (bta05215)	۲	*FGFR1, CREBBP*	۰/۰۳۳۷۲۲

جدول ۵) مقایسه نتایج روش پیشنهادی اول و روش‌های استفاده‌شده دیگر

روش	**دقت** (%)	**تعداد ژن‌های انتخاب‌شده**	**نسبت تعداد ژن‌های شناسایی‌شده /تعداد ژن‌های انتخاب‌شده** (%)
GA + SVM	۹۱	۲۰	۱۵
IG + SVM	۸۷	۲۰	۱۵
RF + SVM	۸۷	۲۰	۱۵
روش پیشنهادی	۹۰/۳۲	۴	۷۵

الگوریتم ژنتیک (GA)، بهره اطلاعات (IG)، جنگل‌های تصمیم تصادفی (RF)، ماشین بردار پشتیبان (SVM)

بحث

شناسایی ژن‌های عامل بیماری یکی از مهم‌ترین حوزه‌های تحقیقات پزشکی است که با استفاده از تحلیل داده‌های ریزآرایه می‌توان این هدف را تحقق بخشید. تعداد زیاد ژن‌ها و در مقابل تعداد کم نمونه‌ها در داده‌های ریزآرایه، تحلیل این سری از داده‌ها

را با مشکل همراه می‌سازد. در این مقاله برای کاهش این مشکل روشی ترکیبی شامل هر دو روش فیلتر و Wrapper ارایه شد. در حقیقت این روش از ترکیب معیار فیشر و روش SAM و PSO به‌منظور انتخاب ژن‌های حاوی اطلاعات استفاده کرد. بدین ترتیب که ابتدا با استفاده از معیار فیشر ژن‌های نویزی و زاید از مجموعه داده حذف شده و سپس با استفاده از روش SAM و PSO ژن‌های مرتبط و باارزش انتخاب شدند. در نهایت نیز با استفاده از الگوریتم CART به یادگیری مدل بر مبنای ژن‌های انتخاب‌شده از مرحله قبل پرداخته شد.

در این مدل ۴ ژن شامل گیرنده فاکتور رشد فیبروبلاست (*FGFR1*)، خانواده انگشت روی (*GLI1*)، پروتئین متصل‌شونده CREB (*CREBBP*) و *GRAP* به‌عنوان گره‌های درخت تصمیم مشخص شده‌اند و ژن‌های انتخاب‌شده نهایی با آلزایمر مرتبط هستند.

سه ژن (*FGFR1*، *GLI1* و *CREBBP*) از این ۴ ژن تاکنون به‌عنوان ژن‌های دخیل در آلزایمر معرفی و در مقالات زیستی گزارش شده‌اند[10-12] و در واقع ژن چهارم به‌عنوان یک ژن نامزد جدید مرتبط با این بیماری ارایه شده است. به‌منظور بررسی بیشتر این ژن‌ها، با استفاده از نرم‌افزار BisoGenet 3.4.0 [13]، شبکه PPI مرتبط با ۴ ژن مذکور رسم و در شکل ۳ نمایش داده شده است. این شبکه شامل ۲۲ گره (ژن) و ۳۹ یال است که با بررسی ژن‌های موجود در این شبکه بیش از ۱۴ ژن دیده می‌شود که به‌عنوان ژن‌های دخیل در آلزایمر در علم پزشکی شناسایی و در مقالات زیستی گزارش شده‌اند. در ژن‌های *SMAD2*، *SMAD4*، *HDAC2*، *DYRK1A*، *STAT3*، *CTNNB1*، *SOS1* و *SHC1*[14-18] با توجه به شبکه PPI به‌ترتیب ۴ ژن اول با *ژن GLI1* ، ۵ ژن اول با ژن *CREBBP* و در نهایت ۳ ژن آخر با ژن *FGFR1* در ارتباط هستند. این ارتباطات بیانگر وجود مسیرهای سلولی و ارتباطات مولکولی مهم مرتبط با آلزایمر هستند که ۳ ژن *FGFR1*، *GLI1*، و *CREBBP* در آنها ایفای نقش می‌کنند. از طرفی ریشه درخت رده‌بندی براساس بیان ژن *FGFR1* تشکیل شده است، به‌طوری که براساس این مدل خطر ابتلا به آلزایمر در افرادی که دارای بیان ژن *FGFR1* کمتر از ۳۰۳/۲۱۳ هستند تقریباً صفر گزارش شده است و این به‌معنی ایفای نقش کلیدی این ژن در بروز بیماری آلزایمر است که با مشاهدات بالا و همچنین مطالعات زیستی معتبری که تاکنون در این رابطه انجام شده است، مطابقت دارد.

به‌عنوان نمونه‌ای از این مطالعات می‌توان به مطالعات *مو* و *یانگ*[10] و *فرر* و *مارتی*[19] اشاره کرد که در آنها به بررسی نقش ژن *FGFR1* در بروز آلزایمر پرداخته و نقش اساسی این ژن در بروز بیماری مذکور را تایید می‌کنند. همچنین در مطالعه‌ای دیگر *کلجبور* و همکاران به نقش ژن *FGFR1* در بیماری اسکیزوفرنی پرداخته‌اند[20]. آنها در این مطالعه با صراحت بیان می‌کنند که ژن *FGFR1* مانند یک رهبر ارکستر عمل می‌کند و به‌طور فیزیکی با همه ژن‌هایی که روی اسکیزوفرنی اثر می‌کنند، در ارتباط است. از طرفی در تحقیقات دیگر به ارتباط بین آلزایمر و اسکیزوفرنی اشاره شده است[21]. در واقع محققان انگلیسی با بررسی اسکن‌های بیماران مبتلا به آلزایمر و اسکیزوفرنی، متوجه شدند نواحی یکسانی از مغز تحت تاثیر این شرایط قرار گرفته بودند. این نکته که اسکیزوفرنی و آلزایمر با یکدیگر مرتبط هستند با نتایج متخصصان دیگر مطابقت دارد. در نتیجه با توجه به این مطالعات و مشاهدات می‌توان کلیدی‌بودن نقش ژن *FGFR1* را در بروز آلزایمر تایید کرد.

در ۳ مسیر KEGG به‌دست‌آمده برای ۴ ژن مذکور، ۲ مورد از آنها (Pathways in cancer و Adherens junction) تاکنون به‌عنوان مسیرهایی که در ارتباط با بروز آلزایمر هستند، شناسایی شده‌اند[22, 23]. به ترتیب در مسیر اول (Pathways in cancer) ژن‌های *FGFR1*، *GLI1*، و *CREBBP* و در مسیر دوم (Adherens junction) ژن‌های *FGFR1* و *CREBBP* ایفای نقش می‌کنند و این در واقع به‌معنای ارتباط و اهمیت ۴ ژن انتخاب‌شده نهایی در بروز آلزایمر است.

تمامی این یافته‌ها مهر تاییدی بر اعتبار مدل به‌دست‌آمده و کارآیی روش پیشنهادی برای شناسایی ژن‌های عامل بیماری است. در واقع در این مطالعه با استفاده از روش پیشنهادی با دقت مطلوبی، ژن *GRAP* به‌عنوان یک ژن نامزد حاوی اطلاعات مرتبط با آزایمر ارایه شد که با بررسی شبکه PPI نیز می‌توان این ادعا را تایید کرد. زیرا همان طور که در شکل ۳ مشخص است این ژن با ژن‌های *SOS1* و *SHC1* به‌صورت مستقیم و با ژن‌های *GLI1*، *CREBBP* و*FGFR1* به‌صورت غیر مستقیم در ارتباط است و پیش‌تر نیز بیان شد که تمام این ۵ ژن تاکنون در حوزه پزشکی به‌عنوان ژن‌های عامل آلزایمر شناسایی شده‌اند. به بیان دیگر می‌توان گفت که ژن *GRAP* در ارتباط مولکولی با ژن‌هایی است که در بروز آلزایمر دخیل هستند و در واقع ژن مذکور در مسیرهای سلولی مشخص به‌همراه این ژن‌ها در بروز آلزایمر ایفای نقش می‌کند.

هر چند تاکنون یافته‌ای دال بر نقش ژن *GRAP* در آلزایمر گزارش نشده است، اما بررسی مطالعات مرتبط سرنخ‌هایی را آشکار می‌سازد که تاییدکننده نقش احتمالی این ژن در آلزایمر و اهمیت روش ما در شناسایی ژن‌های جدید عامل بیماری است. از یک سو، ژن *GRAP* از طریق دامنه SH3-SH2-SH3 خود مسیر Ras را تنظیم می‌کند[24] و از سوی دیگر تحقیقات نشان داده که بیان Ras در کورتکس فرانتال مغز بیماران آلزایمری افزایش می‌یابد[25]. همچنین شواهدی وجود دارد که نشان می‌دهد در مغز بیماران مبتلا به آلزایمر، فعالیت گلیکولازهای لیزوزونی افزایش می‌یابد و Ras در تنظیم فعالیت‌های لیزوزومی نقش فعال دارد[26].

در مطالعه *اسکوبرت* و همکاران عملکرد سه روش الگوریتم ژنتیک (GA)، بهره اطلاعات (IG) و جنگل‌های تصمیم تصادفی (RF) در پیاده‌سازی روی مجموعه داده‌های آلزایمر، با یکدیگر مقایسه شد[27]. در واقع این مقاله به بررسی تاثیر پیاده‌سازی هریک از سه روش مذکور قبل از مدل‌سازی با استفاده از روش دسته‌بندی‌کننده ماشین بردار پشتیبان (SVM)، بر سطح دقت مدل به‌دست‌آمده پرداخته است.

نتیجه‌گیری

پیاده‌سازی روش پیشنهادی روی داده‌های ریزآرایه آلزایمر و سپس مدل‌سازی با استفاده از الگوریتم رده‌بندی درختیCART منجر به شناسایی ۴ ژن مرتبط با آلزایمر شد که ۳ مورد از این ۴ ژن تاکنون به‌عنوان ژن‌های دخیل در بروز بیماری مذکور در مقالات زیستی معرفی شده‌اند. در واقع روش پیشنهادی، موفق به ارایه یک ژن نامزد (*GRAP*) به‌عنوان ژن دخیل در بروز آلزایمر شده است و احتمال مرتبط‌بودن این ژن با بیماری مذکور بسیار بالا است.

تشکر و قدردانی: موردی از سوی نویسندگان گزارش نشده است.

تاییدیه اخلاقی: اینجانب سجاد ازگلی نویسنده مسئول مقاله حاضر گواهی می نمایم که این مقاله قبلاً در هیچ نشریه‌ای اعم از داخلی یا خارجی چاپ نشده است، این مقاله صرفاً به‌منظور بررسی و چاپ به مجله زیست‌فناوری تربیت مدرس ارسال شده و تا هنگام پایان بررسی، داوری مقاله و اعلام نظر نهایی فصل‌نامه، به مجله دیگری ارسال نخواهد شد، این مقاله در نتیجه فعالیت‌های تحقیقاتی اینجانب و همکاران (سیده زهرا پای‌لاخی و سید حسن پای‌لاخی) تهیه و تحریر شده و حقوق کلیه افرادی که به نحوی در اجرای این تحقیق مشارکت و همکاری داشته‌اند، رعایت شده است. در جریان اجرای این تحقیق و تهیه مقاله کلیه اصول اخلاق حرفه‌ای مرتبط با موضوع تحقیق از جمله رعایت حقوق آزمودنی‌ها، سازمان‌ها و نهادها و نیز مولفین و مصنفین رعایت شده است.

تعارض منافع: موردی از سوی نویسندگان گزارش نشده است.

سهم نویسندگان: سیده زهرا پای‌لاخی (نویسنده اول)، نگارنده مقدمه/روش‌شناس/پژوهشگر اصلی/تحلیلگر آماری/نگارنده بحث (۷۰%)؛ سجاد ازگلی (نویسنده دوم)، پژوهشگر کمکی/نگارنده بحث (۱۵%)؛ سیدحسن پای‌لاخی (نویسنده سوم)، پژوهشگر کمکی/نگارنده بحث (۱۵%).

منابع مالی: موردی از سوی نویسندگان گزارش نشده است.

منابع

1- Fakhr E, Motamed N, Habibi Rezaei M. Microarray technology. Genet 3rd Millenn. 2011;9(3):2481-8. [Persian]

2- Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9 Suppl 1:S13.

3- Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A. 2004;101(7):2173-8.

4- Mohammadi A, Saraee MH, Salehi, M. Identification of disease-causing genes using microarray data mining and Gene Ontology. BMC Med Genom. 2011;4:12.

5- Tian LP, Liu LZ, Wu FX. Identification of pseudo-periodic gene expression profiles. Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP). Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2011.

6- Eberhart R, Kennedy J. A new optimizer using particle swarm theory. MHS'95, Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 4-6 Oct, 1995, Nagoya, Japan. Piscataway: IEEE; 1995.

7- Lee SK. On classification and regression trees for multiple responses and its application. J Classif. 2006;23(1):123-41.

8- Behnampour N, Hajizadeh E, Semnani S, Zayeri F. The introduction and application of classification tree model for determination of risk factor for esophageal cancer in Golestan province. Jorjani Biomed J. 2013;1(2):38-46. [Persian]

9- Saniee Abadeh M, Mahmoodi S, Taherpour M. Applied data mining. Tehran: Niyaz Danesh; 2012.

10- Kong LN, Zuo PP, Mu L, Liu YY, Yang N. Gene expression profile of amyloid beta protein-injected mouse model for Alzheimer disease. Acta Pharmacol Sin. 2005;26(6):666-72.

11- He P, Staufenbiel M, Li R, Shen Y. Deficiency of patched 1-induced Gli1 signal transduction results in astrogenesis in Swedish mutated APP transgenic mice. Hum Mol Genet. 2014;23(24):6512-27.

12- Rosa E, Fahnestock M. CREB expression mediates amyloid β-induced basal BDNF downregulation. Neurobiol Aging. 2015;36(8):2406-13.

13- Martin A, Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de-Cossio J, Bringas R. BisoGenet: A new tool for gene network building, visualization and analysis. BMC Bioinformatics. 2010;11:91.

14- Von Bernhardi R, Cornejo F, Parada GE, Eugenín J. Role of TGFβ signaling in the pathogenesis of Alzheimer's disease. Front Cell Neurosci. 2015;9:426.

15- Griñán-Ferré C, Sarroca S, Ivanova A, Puigoriol-Illamola D, Aguado F, Camins A, et al. Epigenetic mechanisms underlying cognitive impairment and Alzheimer disease hallmarks in 5XFAD mice. Aging (Albany NY). 2016;8(4):664-84.

16- Kim H, Lee KS, Kim AK, Choi M, Choi K, Kang M, et al. A chemical with proven clinical safety rescues Down-syndrome-related phenotypes in through DYRK1A inhibition. Dis Model Mech. 2016;9(8):839-48.

17- Kim SJ, Guerrero N, Wassef G, Xiao J, Mehta HH, Cohen P, et al. The mitochondrial-derived peptide humanin activates the ERK1/2, AKT, and STAT3 signaling pathways and has age-dependent signaling differences in the hippocampus. Oncotarget. 2016;7(30):46899-912.

18- Xie Z, Dong Y, Maeda U, Xia W, Tanzi RE. RNA interference silencing of the adaptor molecules ShcC and Fe65 differentially affect amyloid precursor protein processing and Abeta generation. J Biol Chem. 2007;282(7):4318-25.

19- Ferrer I, Martí E. Distribution of fibroblast growth factor receptor-1 (FGFR-1) and FGFR-3 in the hippocampus of patients with Alzheimer's disease. Neurosci Lett. 1998;240(3):139-42.

20- Klejbor I, Myers JM, Hausknecht K, Corso TD, Gambino AS, Morys J, et al. Fibroblast growth factor receptor signaling affects development and function of dopamine neurons - inhibition results in a schizophrenia-like syndrome in transgenic mice. J Neurochem. 2006;97(5):1243-58.

21- Douaud G, Groves AR, Tamnes CK, Westlye LT, Duff EP, Engvig A, et al. A common brain network links development, aging, and vulnerability to disease. PNAS. 2014;111(49):17648-53.

22- Nakase T, Naus CC. Gap junctions and neurological disorders of the central nervous system. Biochim Biophys Acta. 2004;1662(1-2):149-58.

23- Realmuto S, Cinturino A, Arnao V, Mazzola MA, Cupidi C, Aridon P, et al. Tumor diagnosis preceding Alzheimer's disease onset: Is there a link between cancer and Alzheimer's disease?. J Alzheimers Dis. 2012;31(1):177-82.

24- Feng GS, Ouyang YB, Hu DP, Shi ZQ, Gentz R, Ni J. Grap is a novel SH3-SH2-SH3 adaptor protein that couples tyrosine kinases to the Ras pathway. J Biol Chem. 1996;271(21):12129-32.

25- Gärtner U, Holzer M, Arendt T. Elevated expression of p21ras is an early event in Alzheimer's disease and precedes neurofibrillary degeneration. Neuroscience. 1999;91(1):1-5.

26- Démoz M, Castino R, Dragonetti A, Raiteri E, Baccino FM, Isidoro C. Transformation by oncogenic ras-p21 alters the processing and subcellular localization of the lysosomal protease cathepsin D. J Cell Biochem. 1999;73(3):370-8.

27- Scheubert L, Luštrek M, Schmidt R, Repsilber D, Fuellen G. Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. BMC Bioinformatics. 2012;13:266.