Original Article

Identification of Heat-Resistant Bacteria Based on Selection of Proper Representation of Protein Sequences Using Deep Learning Approach

***Reza Ahsan*[1], *Mansour Ebrahimi*[2*]**

[1]*School of Engineering, University of Qom, Qom, Iran.*

[2]*Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran.*

[*]***Corresponding Author:*** **Mansour Ebrahimi**; Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran.

Email: mansour@future.edu

Received: 2 Jan, 2020
Accepted: 7 Jun, 2020

Abstract

Background and Objectives: Identification of effective mechanisms in heat-resistance in bacteria is of great importance in some industries, such as food industry, textile manufacturing, and especially in detergent production industries. For this purpose, deep learning tools were used to identify the characteristics of heat-resistant bacteria based on protein properties.

Methods: Some characteristics of heat-resistant and non-heat-resistant proteins, such as the structural properties of amino acids, the number and the frequency of each amino acid, and their physicochemical properties, were calculated. Bacterial classification was performed in three steps: first, attribute weighting methods were used to select the important variables, then those variables, were selected and finally deep learning networks were employed to extract the hierarchy of the features.

Results: The results of 10 weighting methods showed that out of 73 characteristics of the number and frequency of amino acids, only 40 had weights higher than zero. Of these variables, 13 variable gained weight higher than 0.5 and only 10 variables had weight above 0.09. These 10 features were selected as important variables. The frequencies of glutamine and glutamic acid obtained the highest possible weights and were considered as two important features in the classification of heat-resistant and non-heat-resistant bacteria. The highest prediction accuracy of the deep learning networks was 92.42% for the classification of heat resistant bacteria.

Conclusion: The deep neural networks can be effectively used to identify heat-resistant bacteria based on their protein properties.

Keywords: Thermostable; Protein sequence; Classification; Deep learning networks.

DOI: 10.29252/qums.14.3.54

مجله دانشگاه علوم پزشکی قم
دوره چهاردهم، شماره سوم، خرداد ۹۹
صفحه ٥٤ الی ٦٣

مقاله پژوهشی
(Original Article)

شناسایی باکتری‌های مقاوم به گرما براساس انتخاب بازنمایی مناسب از توالی پروتئین با استفاده از رویکرد یادگیری عمیق

رضا احسن[۱]، منصور ابراهیمی[۲*]

[۱]دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران.

[۲]گروه زیست‌شناسی، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

[*]نویسنده مسئول مکاتبات:
منصور ابراهیمی؛ گروه زیست‌شناسی، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

آدرس پست الکترونیکی:
mansour@future.edu

تاریخ دریافت: ۱۳۹۸/۱۰/۱۲
تاریخ پذیرش: ۱۳۹۹/۳/۱۸

چکیده

زمینه و هدف: شناسایی مکانیزم‌های مؤثر در مقاومت باکتری‌ها به گرما به منظور ایجاد سویه‌های مقاوم به گرما در صنایعی از جمله صنایع غذایی، ساخت منسوجات و به ویژه در صنایع تولیدکننده مواد شوینده بسیار حائز اهمیت است. برای این منظور، از ابزارهای یادگیری عمیق برای شناسایی ویژگی‌های باکتری‌های مقاوم به گرما براساس خصوصیات پروتئینی استفاده گردید.

روش بررسی: در این مطالعه برخی از ویژگی‌های پروتئین‌های مقاوم و غیر مقاوم به گرما از قبیل ویژگی‌های ساختاری اسید آمینه‌ها، تعداد و فرکانس هر اسید آمینه و ویژگی‌های فیزیکوشیمیایی آن‌ها محاسبه گردیدند. سپس جهت رده‌بندی باکتری‌ها، ابتدا از مدل‌های وزن‌دهی برای شناسایی متغیرهای مهم استفاده گردید. در ادامه آن‌ها انتخاب شدند و در نهایت با استفاده از شبکه عصبی عمیق نسبت به استخراج سلسله‌مراتب ویژگی‌ها اقدام گردید.

یافته‌ها: نتایج ۱۰ روش وزن‌دهی نشان دادند که از بین ۷۳ ویژگی تعداد و فرکانس اسیدهای آمینه، تنها ۴۰ ویژگی دارای وزن بالاتر از ۰ بودند که از این تعداد، ۱۳ ویژگی وزن بالاتر از ۰/۵ را کسب نموده و تنها ۱۰ ویژگی دارای میانگین وزن بالاتر از ۰/۰۹ بودند. این ۱۰ ویژگی به عنوان متغیرهای مهم انتخاب شدند. ویژگی‌های فرکانس‌های گلوتامین و اسید گلوتامیک بیشترین وزن را اخذ نموده و به عنوان دو ویژگی مهم در رده‌بندی باکتری‌های مقاوم و غیر مقاوم به گرما معرفی شدند. بیشترین دقت پیش‌بینی رده‌بندی باکتری‌های مقاوم به گرما توسط شبکه یادگیری عمیق معادل ۹۲/۴۲ درصد بود.

نتیجه‌گیری: شبکه‌های عصبی عمیق با استخراج سلسله‌مراتب ویژگی می‌توانند به خوبی باکتری‌های مقاوم به‌گرما را براساس ویژگی‌های پروتئینی آن‌ها شناسایی کنند.

کلیدواژه‌ها: مقاوم به گرما؛ توالی پروتئین؛ رده‌بندی؛ یادگیری عمیق.

لطفاً به این مقاله به‌صورت زیر استناد نمایید:

Ahsan R, Ebrahim M. Identification of Heat-Resistant Bacteria Based on Selection of Proper Representation of Protein Sequences Using Deep Learning Approach. [Qom Univ Med Sci J 2020;14(3):54-63. [Full Text in Persian

مقدمه

بیوانفورماتیک دانش استفاده از علوم کامپیوتر، آمار و احتمالات در شاخه زیست‌شناسی مولکولی است. امروزه توالی ژنوم و پروتئین بسیاری از موجودات ساده مانند باکتری‌ها و ویروس‌ها تا موجودات بسیار پیشرفته همچون یوکاریوت‌های پیچیده شناسایی شده است. پیشرفت‌های فناوری در علم ژنتیک و تصویربرداری، انفجاری را در آنالیز حجم زیادی از نمونه‌های مولکولی و سلولی ایجاد کرده است. این افزایش سریع در نرخ استفاده از آنالیز روش‌های متعارف ابعاد داده‌های بیولوژیکی چالش‌برانگیز می‌باشد. روش‌های مدرن یادگیری ماشین از قبیل یادگیری عمیق، نویدی برای قدرت نفوذ به ساختار مخفی بین مجموعه داده‌های بسیار بزرگ و ساخت پیش‌بینی‌های دقیق می‌باشد.

از علم بیوانفورماتیک برای رده‌بندی باکتری‌ها استفاده شده است. این امر به ویژه با توسعه سریع فناوری توالی نسل بعدی با استفاده از پردازش داده‌های ژنومی، نقش مهمی را در شناسایی و رده‌بندی باکتری ایفا نموده است (۱). به دلیل نیاز روزافزون صنایع مختلف از جمله صنایع تولید مواد غذایی و صنایع شوینده به آنزیم‌های مقاوم به گرما به منظور بهینه کردن روندهای تولید محصولات مختلف، در چند دهه اخیر تحقیقات گسترده‌ای در زمینه شناسایی و یافتن دلایل مقاومت به گرما در آنزیم‌ها و به ویژه آنزیم‌هایی که از باکتری‌ها جدا شده‌اند صورت گرفته است. استفاده از شبکه‌های عصبی برای شناسایی باکتری‌های مقاوم به گرما توسط برخی از پژوهشگران مورد توجه قرار گرفته است. روش‌های مصنوعی هوشمند مانند تابع رادیال، الگوریتم ژنتیک، شبکه عصبی مصنوعی و ماشین بردار پشتیبان از پتانسیل کافی در زمینه رده‌بندی و شناسایی باکتری‌ها برخوردار هستند. تجزیه و تحلیل مقایسه‌ای با استفاده از پلت‌فرم داده‌کاوی صورت گرفته شده است که در آن، ماشین بردار پشتیبان، بهترین روش بوده و حداکثر دقت (۹۱ درصد) را فراهم کرده است (۲). شناسایی و تشخیص ویژگی‌های مهم به منظور رده‌بندی باکتری‌ها بر پایه ساختار توالی پروتئین آن‌ها صورت گرفته است؛ به طوری که اسیدهای آمینه مختلف، آب‌گریزی، درصد گوگرد نسبی و تعداد کدون به عنوان پارامترهای مهم شناخته شده‌اند (۳). تجزیه و تحلیل ساختار و توالی چندین پروتئین بیش از حد مقاوم به حرارت از منابع مختلف نشان‌دهنده دو مکانیزم فیزیکی عمده مبتنی بر ساختار و مبتنی بر توالی در آن‌ها می‌باشد (۴).

رسیدن به فضای جستجو شامل تمام زیرمجموعه‌های ممکن برای یافتن مناسب‌ترین ویژگی‌ها و رسیدن به زیرمجموعه بهینه نشان داده است که مسأله حل نشدنی در زمان چند جمله‌ای بر حسب اندازه ورودی مسئله (Non-deterministic Polynomial-time hard) می‌باشد (۵). جستجوی جامع تضمین می‌کند که مناسب‌ترین ویژگی‌ها به دست آیند؛ اما معمولاً این کار به لحاظ محاسباتی حتی برای مجموعه‌های داده‌ای با اندازه متوسط غیر ممکن می‌باشد؛ از این رو در پژوهش حاضر از فرکانس تکرار اسیدهای آمینه در توالی پروتئین به عنوان ویژگی و از روش‌های وزن‌دهی به منظور انتخاب مناسب‌ترین ویژگی‌ها به عنوان مرحله پیش‌پردازش روش پیشنهادی استفاده گردید.

ارزش شبکه عصبی عمیق در این زمینه از دو جنبه است. نخست اینکه شبکه‌های عصبی عمیق می‌توانند به یادگیری ویژگی‌ها از داده‌ها کمک کند و مورد دوم این است که به دلیل توانایی در استخراج سلسله‌مراتب ویژگی‌ها می‌توانند وابستگی‌های غیر خطی در توالی و همچنین اثرات متقابل آن‌ها را استخراج نموده و محدوده توالی گسترده‌تر در مقیاس ژنومی چندگانه را اندازه‌گیری کنند (۶). کاربرد مفید دیگر یادگیری عمیق موفقیت در تقسیم‌بندی، پیش‌بینی نتایج ترتیب کنار هم قرار دادن بخش‌های اطلاعاتی کدکننده ژن یعنی اگزون‌ها (۷،۸) است. در تشخیص ویژگی‌های پروتئین‌های متصل شونده به DNA (Deoxyribonucleic Acid) و RNA (Ribonucleic Acid) (۹)، رمز گشایی کد نظارتی بیان ژن و درک اثرات رونویسی اثرات ژنوم (۱۰،۱۱)، و همچنین در رده‌بندی زیرگروه‌های توالی پروتئین ویروس آنفولانزا (۱۲)، یادگیری عمیق موفقیت آمیز عمل کرده است. با توجه به مطالب بیان شده، مطالعه حاضر در مرحله اول با هدف ارائه بازنمایی مناسب از توالی اسید آمینه پروتئین به منظور رده‌بندی توالی پروتئین و در مرحله دوم با هدف شناسایی باکتری‌های مقاوم به‌گرما که در صنایعی از جمله تولید شربت گلوکز و فروکتوز، ساخت منسوجات و به ویژه پودرهای شستشو مورد استفاده قرار می‌گیرند، انجام شد.

در مرحله سوم شناسایی ویژگی‌های مؤثر در ویژگی‌های بازنمایی شده از توالی اسید آمینه پروتئین از طریق رأی‌گیری از روش‌های وزن‌دهی در تشخیص باکتری‌های مقاوم به گرما انجام شد. در مرحله چهارم نیز به‌کارگیری شبکه عصبی عمیق به منظور استخراج سلسله‌مراتب ویژگی‌ها جهت بهبود دقت رده‌بندی باکتری‌ها براساس مقاومت به گرما در نظر گرفته شد.

روش بررسی

پژوهش نظری- کاربردی حاضر در ارتباط با مجموعه دیتاست ۴۶۳۳ توالی پروتئینی باکتری‌های مقاوم و غیر مقاوم به گرمای استخراج شده از پایگاه داده NCBI (National Center for Biotechnology Information) انجام شد.

جامعه پژوهش عبارت بود از ویژگی‌های استخراج شده از توالی پروتئین که شامل: تعداد تکرار هر اسید آمینه در هر توالی و نیز فرکانس هر اسید آمینه (که نسبت تکرار هر اسید آمینه به طول توالی است) بود. مجموعه داده مورد استفاده در این پژوهش حاوی اطلاعات ۴۶۳۳ نمونه با داده مفقود شده و پرت و ۴۶۱۵ نمونه بدون داده مفقود شده و پرت بود. هر سطر از سطرهای این مجموعه داده، ویژگی‌هایی از یک توالی پروتئین باکتری را نشان می‌دهد. شایان ذکر است که ۱۵۶۵ نمونه مربوط به باکتری مقاوم به گرما و ۳۰۶۸ نمونه مربوط به باکتری غیر مقاوم به گرما بودند. این داده‌ها به ۱۰ بخش تقسیم شدند: ۹۰ درصد به عنوان داده‌های آموزش برای ساخت مدل یادگیری و ۱۰ درصد از داده‌های آزمون جهت سنجش عملکرد رده‌بندی باکتری‌ها براساس مقاومت به گرما در نظر گرفته شدند.

دو راه‌کار عمده برای کاهش ابعاد مجموعه داده‌ای ارائه شده است: استخراج ویژگی و انتخاب ویژگی. در استخراج ویژگی، فضای اولیه ویژگی‌ها به یک فضای دیگر نگاشت می‌شود. در واقع در این راه‌کار با ترکیب ویژگی‌های موجود، تعدادی ویژگی جدید ایجاد می‌شود؛ به طوری که این ویژگی‌ها دارای تمام یا بخش اعظمی از اطلاعات موجود در ویژگی‌های اولیه می‌باشند. در این پژوهش استخراج سلسله‌مراتبی ویژگی‌ها با رویکرد پیشنهادی شبکه عصبی عمیق به صورت خودرمزگذار پشته‌ای انجام شد. از سوی دیگر در انتخاب ویژگی، یک زیرمجموعه از ویژگی‌های اولیه انتخاب گردید. انتخاب ویژگی یک روش مهم و پراستفاده در پیش‌پردازش داده‌ها محسوب می‌شود که موجب افزایش سرعت الگوریتم‌های یادگیری ماشین و بهبود عملکرد طبقه‌بندی‌کننده می‌گردد. برای تحقق این امر از ۱۰ روش وزن‌دهی بهره گرفته شد. جزئیات هر روش وزن‌دهی و معادلات آماری آن‌ها در مقالات قبلی نویسندگان شرح داده شده است. به طور خلاصه می‌توان گفت که در روش قانون، ارتباط یک صفت با محاسبه میزان خطا با در نظر نگرفتن آن صفت در پیش‌بینی نوع باکتری محاسبه می‌گردد. در ماشین بردار پشتیبان نیز از ضرایب بردار نرمال که از ماشین بردار پشتیبان گرفته می‌شود به عنوان وزن‌دهی ویژگی استفاده می‌گردد. همچنین در عدم قطعیت، هر صفتی که مقادیرش برای رسیدن به هدف تصادفی‌تر باشد، انتخاب نمی‌شود. از سوی دیگر در برجستگی، به صورت تصادفی یک نمونه از زیرمجموعه نمونه انتخاب می‌شود و برای هریک از ویژگی‌های این نمونه، ویژگی‌هایی که به خوبی دو کلاس را از یکدیگر تمیز می‌دهند، انتخاب می‌شوند؛ زیرا برای نمونه‌های متعلق به دو کلاس متفاوت، مقادیری نزدیک به یکدیگر را ارائه نمی‌دهند و یک فاصله معنادار را بین مقادیر نمونه‌های یک کلاس در نظر می‌گیرند و مقادیری که به سایر کلاس‌ها می‌دهند وجود دارد. در تحلیل مولفه‌های اصلی، تبدیلی در فضای برداری برای کاهش ابعاد مجموعه داده‌ها مورد استفاده قرار می‌گیرد، به این ترتیب مولفه‌هایی از مجموعه داده را که بیشترین تاثیر در واریانس را دارند، حفظ می‌کند و بقیه حذف می‌شوند. علاوه‌براین، در انحراف معیار از اختلاف داده‌های مربوط به هر کلاس در هر ویژگی با میانگین به منظور تعیین ضریب اطمینان ویژگی استفاده می‌کند. همچنین در معیار آزمون مربع خی‌دو که مشابه با آزمون‌های دیگر آماری است، میزان ارتباط یا وابستگی بین متغیرها را اندازه‌گیری می‌کند. علاوه‌براین می‌تواند برای آزمون وابستگی بین یک یا چند گروه نیز به کار رود که این عمل را از طریق مقایسه تعداد واقعی نمونه‌های مشاهده شده در هر گروه با نمونه‌هایی که مطابق با تئوری یا احتمال، انتظار می‌رود مشاهده شوند، انجام می‌دهد. آزمون مربعی چای بر پایه آزمون آماری خی دو بوده و برای استفاده از این آزمون نیاز است که داده‌ها

به شکل تعداد تکرار بیان شوند. از سوی دیگر، در معیار شاخص جینی هرچه تعیین نوع کلاس برای مقادیر یک ویژگی محتمل‌تر باشد، آن ویژگی وزن بیشتری خواهد داشت. باید خاطرنشان ساخت که در بهره اطلاعاتی از بین ویژگی‌ها آن‌هایی انتخاب می‌شوند که اطلاعات بیشتری را برای انتخاب کلاس هدف ارائه می‌دهند. در بهره اطلاعاتی نسبی از بین ویژگی‌ها، آن که نسبت بهره اطلاعاتی بر آنتروپی آن بزرگتر باشد، وزن بیشتری خواهد داشت. در حقیقت، این معیار بهره اطلاعاتی را نرمال‌سازی می‌کند. معیارهای قبلی به سوی یک ویژگی با مقادیر دامنه بزرگتر گرایش دارند. به عبارت دیگر، این معیارها یک ویژگی با مقدار زیاد را به یک ویژگی با مقدار کم ترجیح می‌دهند؛ به همین دلیل، نرمال‌سازی این معیارها مفید به نظر می‌رسد. می‌توان گفت که بهره اطلاعاتی نسبی در مقایسه با بهره اطلاعاتی، عملکرد بهتری را در دقت و پیچیدگی مدل از خود نشان می‌دهد.

متغیرهای مورد بررسی در این مطالعه شامل ۷۱ ویژگی بودند که متغیر درجه حرارت که مقاومت باکتری به گرما را نشان می‌دهد به عنوان متغیر اصلی یا متغیر هدف لحاظ گردید و ۷۰ ویژگی نیز به عنوان متغیر مستقل جهت رده‌بندی در نظر گرفته شدند. برای تجزیه و تحلیل دادها متناسب با اهداف پژوهش و متغیرهای کمی و کیفی، داده‌های جمع‌آوری شده ابتدا با استفاده از نرم‌افزار رپیدماینر (Rapidminer) وزن‌دهی ویژگی‌ها را انجام دادند و در ادامه، ویژگی‌های مؤثرتر در هریک از روش‌های وزن‌دهی انتخاب شدند. به منظور رده‌بندی مقاومت باکتری به گرما بر مبنای ویژگی‌های انتخاب شده با استفاده از روش‌های مختلف یادگیری ماشین شامل: ماشین بردار پشتیبان (تقسیم خطی داده‌ها؛ سعی می‌کنیم خطی را انتخاب نماییم که حاشیه اطمینان بیشتری داشته باشد)، درخت تصمیم (یک ابزار برای پشتیبانی از تصمیم است که از درخت‌ها برای مدل کردن استفاده می‌کند)، درخت جنگل تصادفی (انجام تصمیم براساس ۱۰۰ درخت تصادفی)، درخت تصادفی (انتخاب شروط تصادفی)، K نزدیکترین همسایه (تعیین کلاس با k همسایه نزدیک)، القای قانون (استخراج قوانین رسمی از مشاهدات)، رگرسیون خطی، رگرسیون منطقی، بیز ساده (روشی برای دسته‌بندی پدیده‌ها بر پایه احتمال وقوع یا عدم وقوع یک پدیده)، بیز کرنل (دسته‌بندی پدیده‌ها با تابع کرنل)، شبکه عصبی با یک لایه (پردازش داده‌ها و اطلاعات به منظور یادگیری و ایجاد دانش با عناصر پردازشی به نام نورون)، شبکه عصبی با لایه‌های خودکار و یک روش یادگیری عمیق (یک شبکه عصبی دو لایه با ۵۰ نورون در هر لایه) استفاده گردید.

در این مطالعه دو رویکرد برای مقایسه عملکرد روش‌های وزن‌دهی مورد بررسی قرار گرفت. در رویکرد اول در ارتباط با ویژگی‌های مؤثرتر در هر روش وزن‌دهی به منظور رده‌بندی باکتری‌های مقاوم به گرما، چهار مدل یادگیری ماشین (بیز ساده، بیز کرنل، شبکه عصبی و شبکه عصبی با تعداد لایه‌های خودکار) و یک روش یادگیری عمیق مورد مقایسه قرار گرفتند. در رویکرد دوم از ویژگی‌هایی برای رده‌بندی استفاده گردید که در ۱۰ روش وزن‌دهی، حداقل یک رأی داشته باشند و میانگین وزن آن‌ها در ۱۰ روش از ۰/۰۹ بیشتر باشد. برای مقایسه عملکرد رویکرد دوم در رده‌بندی باکتری‌های مقاوم به گرما از ۱۲ روش یادگیری ماشین (ماشین بردار پشتیبان، درخت تصمیم، درخت جنگل تصادفی، درخت تصادفی، K نزدیکترین همسایه، القای قانون، رگرسیون خطی، رگرسیون منطقی، بیز ساده، بیز کرنل، شبکه عصبی با یک لایه و شبکه عصبی با لایه‌های خودکار) و یک شبکه عصبی عمیق استفاده شد.

یافته‌ها

در رویکرد اول در هر روش وزن‌دهی، ویژگی‌هایی انتخاب شدند که وزنی بالاتر از ۰/۵ را به خود اختصاص داده بودند. جدول ۲ نتایج دقت چهار مدل یادگیری ماشین و یک روش یادگیری عمیق را نشان می‌دهد. بر مبنای نتایج دقت روش‌های یادگیری ماشین و یادگیری عمیق در ویژگی‌های انتخاب شده با وزن‌دهی شبکه بردار پشتیبان نسبت به سایر روش‌های وزن‌دهی بیشتر شد؛ از این رو این روش وزن‌دهی در این مجموعه از داده‌ها، ویژگی‌های مؤثرتری را شناسایی کرد. از میان ۴۰ ویژگی که وزنی بالاتر از ۰ دریافت کردند، ۱۳ ویژگی در حداقل یک روش وزن‌دهی، وزنی بالاتر از ۰/۵ را به خود اختصاص دادند؛ اما برای اینکه مشخص شود که کدام ویژگی‌ها در مجموعه ۱۰ روش وزن‌دهی مناسب تشخیص داده شده‌اند، از رویکرد دوم استفاده گردید. در این رویکرد براساس رأی، ویژگی‌ها انتخاب می‌شوند.

بر این اساس، ۱۰ ویژگی که حداقل میانگین وزن آن‌ها در ۱۰ روش وزن‌دهی بیشتر از حد آستانه ۰/۰۹ بود، انتخاب گردیدند. همان‌طور که در جدول ۱ مشاهده می‌شود، ویژگی‌های فرکانس‌های گلوتامین و اسید گلوتامیک، بیشترین رأی را در روش‌های وزن‌دهی کسب نمودند.

جدول شماره ۱: نتایج رأی‌گیری ۱۰ روش وزن‌دهی برای استخراج ویژگی‌های مؤثر

متغیر	میانگین وزن اخذ شده در ۱۰ روش وزن‌دهی	تعداد رأی اخذ شده در ۱۰ روش وزن‌دهی با حداقل وزن ۰/۵
فراوانی گلوتامین	۰/۶۷	۷
فراوانی گلوتامیک اسید	۰/۴۷	۷
فراوانی هیدروفوبیک	۰/۴۴	۶
گلوتامین	۰/۳۲	۱
فراوانی آرژنین	۰/۲۱	۱
فراوانی تیروزین	۰/۱۹	۱
نیتروژن	۰/۱۴	۱
فراوانی سیستئین	۰/۱۲	۱
تریپتوفان	۰/۱۱	۱
سیستئین	۰/۰۹	۱

در مقایسه نتایج دو رویکرد پیشنهادی برای انتخاب ویژگی با وزن‌دهی (جداول ۲ و ۳) ملاحظه شد که ویژگی‌های حاصل از رویکرد دوم در بیشتر موارد ویژگی‌های مؤثرتری در جهت رده‌بندی توالی‌های پروتئین مقاوم به گرما هستند.

جدول شماره ۲: نتایج صحت رده‌بندی باکتری براساس مقاومت به گرما با استفاده از روش‌های یادگیری ماشین

روش وزن‌دهی	بیز ساده (درصد)	بیز کرنل (درصد)	یادگیری عمیق (درصد)	شبکه عصبی (درصد)	شبکه عصبی خودکار (درصد)
قانون	۷۳/۳۰	۶۵/۰۳	۸۵/۴۵	۸۵/۴۳	۸۵/۱۳
ماشین بردار پشتیبان	۸۳/۳۸	۷۴/۶۶	۸۷/۳۱	۸۷/۱۱	۸۶/۹۰
عدم قطعیت	۸۴/۲۶	۷۰/۲۳	۷۰/۸۸	۷۰/۶۷	۶۹/۹۱
برجستگی	۸۰/۰۳	۶۶/۲۲	۷۹/۳۲	۷۹/۱۵	۷۹/۲۱
تحلیل مؤلفه‌های اصلی	۳۷/۵۵	۶۵/۶۱	۳۶/۹۵	۳۳/۹۱	۳۴/۰۶
انحراف معیار	۳۷/۵۵	۶۵/۶۱	۳۷/۳۶	۳۶/۸۵	۳۴/۰۸
تست مربعی خی‌دو	۸۴/۲۶	۷۰/۲۳	۸۵/۴۳	۸۵/۱۱	۸۵/۲۸
ایندکس جینی	۵۴/۴۶	۴۳/۰۰	۸۵/۷۳	۸۵/۷۳	۸۵/۶۹
بهره اطلاعاتی	۵۴/۴۶	۴۳/۰۰	۸۵/۸۸	۸۵/۴۷	۸۵/۶۰
بهره اطلاعاتی نسبی	۵۵/۳۲	۴۶/۶۴	۸۶/۳۲	۸۶/۲۵	۸۵/۶۰

جدول ۳ نشان می‌دهد که یادگیری عمیق نسبت به سایر روش‌های یادگیری ماشین در رده‌بندی توالی‌های مقاوم به گرما موفق‌تر بوده است. روش شبکه عصبی عمیق با دقت ۸۷/۷۴ درصد و پس از آن شبکه عصبی با ۸۷/۵۷ درصد، ماشین بردار پشتیبان با ۸۷/۴۸ درصد، شبکه عصبی با تعداد لایه خودکار ۸۷/۴۸ درصد و رگرسیون منطقی با ۸۷/۲۶ درصد روی ۱۰ ویژگی رأی‌گیری شده با استفاده از روش‌های وزن‌دهی نسبت به روش‌های دیگر یادگیری ماشین دقت بهتری در رده‌بندی باکتری‌های مقاوم به گرما نشان دادند. از بین روش‌های یادگیری ماشین، روش بیز ساده و بیز کرنل کمترین دقت را داشتند.

جدول شماره ۳: نتایج دقت روش‌های رده‌بندی در ارتباط با ۱۰ ویژگی مؤثر رأی‌گیری شده با روش‌های وزن‌دهی

روش‌های رده‌بندی	دقت (درصد)
ماشین بردار پشتیبان	۸۷/۴۸
درخت تصمیم	۸۴/۵۲
جنگل تصادفی	۷۹/۹۷
درخت تصادفی	۷۷/۱۴
K نزدیکترین همسایه	۷۷/۲۹
القای قانون	۸۶/۴۲
رگرسیون خطی	۸۶/۶۸
رگرسیون لجستیک	۸۷/۲۶
بیز ساده	۵۴/۰۷
بیز کرنل	۵۹/۹۳
شبکه عصبی	۸۷/۵۷
شبکه عصبی خودکار	۸۷/۴۸
یادگیری عمیق	۸۷/۷۴

در رویکرد سوم، شبکه عصبی عمیق روی تمام ویژگی‌ها به صورت سلسله‌مراتبی استخراج ویژگی را انجام داده است. روش‌های مختلف ارزیابی نتایج روش پیشنهادی شبکه عصبی عمیق در جدول ۴ نشان داده شده است. رویکرد سوم استخراج ویژگی‌ها با استفاده از شبکه عصبی عمیق با دقت ۹۲/۴۲ درصد توالی‌های پروتئین باکتری مقاوم به گرما را رده‌بندی کرد.

در رویکرد سوم، شبکه عصبی عمیق روی تمام ویژگی‌ها به صورت سلسله‌مراتبی استخراج ویژگی را انجام داد؛ در نتیجه نسبت به دیگر روش‌های یادگیری ماشین ذکر شده در جدول ۳، دقت بالاتری را برای رده‌بندی باکتری‌ها براساس مقاومت به گرما نشان داد.

جدول شماره ٤: نتایج صحت و دقت روش پیشنهادی شبکه عصبی عمیق

		واقعاً درست			
	جامعه آماری	واقعاً مثبت	واقعاً منفی	شیوع= $\frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$ ٪ ۳۳/۷۷=	صحت و دقت (ACC)= $\frac{\Sigma TP + \Sigma TN}{\Sigma \text{ total population}}$ = ۹۲/۴۴ درصد
پیش‌بینی	پیش‌بینی مثبت	مثبت درست (TP) ۱۳۸۶	مثبت نادرست (FP) ۱۷۲ خطای نوع اول و دوم	دقت و بازیابی = $\frac{\Sigma TP}{\Sigma \text{prediction positive}}$ = ۸۸/۹۶ درصد	میزان کشف اشتباه $\frac{\Sigma FP}{\Sigma \text{ prediction positive}}$ = ۱۱/۱۴ درصد
	پیش‌بینی منفی	منفی نادرست (FN) ۱۷۹ خطای نوع اول و دوم	منفی درست (TN) ۲۸۹۶	نرخ غفلت نادرست= $\frac{\Sigma FN}{\Sigma \text{prediction negative}}$ = ۵/۸۲ درصد	ارزش پیش‌بینی منفی= $\frac{\Sigma TN}{\Sigma \text{ prediction negative}}$ = ۹۴/۱۸ درصد
		حساسیت و ویژگی، دقت و بازیابی، احتمال تشخیص و توان آماری= $\frac{\Sigma TP}{\Sigma \text{ condition positive}}$ = ۸۸/۵۶ درصد	ارزش مثبت نادرست، بازیابی اطلاعات و احتمال هشدار اشتباه = $\frac{\Sigma FP}{\Sigma \text{ condition negative}}$ = ۵/۶ درصد	نرخ احتمال مثبت = $\frac{TPR}{FPR}$ = ۱۵/۸۱ درصد	نسبت شانس تشخیص= $\frac{LR+}{LR-} = \frac{TP*FN}{FP*TN}$ = ۴۹/۸۱ درصد
		نرخ اشتباه و ارزش منفی نادرست= $\frac{\Sigma FN}{\Sigma \text{ condition positive}}$ = ۱۲/۴۴ درصد	حساسیت و ویژگی و گزینش‌پذیری= $\frac{\Sigma TN}{\Sigma \text{ condition negative}}$ = ۹۴/۴ درصد	نرخ احتمال منفی = $\frac{FNR}{TNR}$ = ۱۱/۸۸ درصد	امتیاز اف ۱ = $2 * \frac{Precision * Recall}{Precision + Recall}$ = ۸۸/۷۶ درصد

بحث

فعالیت آنزیمی با افزایش دما تا دمایی که در آن فعالیت باقی می‌ماند، افزایش می‌یابد (۱۳). آنزیم‌های حرارتی معمولاً به عنوان حفظ فعالیت پس از حرارت دادن در دمای انتخاب شده برای دوره طولانی مدت تعریف می‌شوند. مناسب‌ترین روش برای بیان حرارت‌پذیری، اندازه‌گیری نیمه عمر فعالیت آنزیم در درجه حرارت بالا است (۱۴). آنزیم‌های حرارتی توسط موجودات ترموفیل و مازوفیلی تولید می‌شوند. اگرچه میکروارگانیسم‌های ترموفیل یک منبع بالقوه برای آنزیم‌های حرارتی هستند، اکثر آنزیم‌های ترموسپتیک صنعتی از مواد ارگانیسم مزوفیلی تشکیل شده‌اند (۱۵).

در پژوهش حاضر برای رسیدن به هدف مورد نظر که عبارت بود از شناسایی باکتری‌های مقاوم به گرما به منظور استفاده بیشتر از این باکتری‌ها در صنایعی از جمله تولید شربت گلوکز و فروکتوز، ساخت منسوجات و به‌ویژه پودرهای شستشو، یک فرایند کشف دانش از داده‌های واقعی طراحی و اجرا شد. این فرایند به صورت خلاصه شامل: پیش‌پردازش و آماده‌سازی داده‌ها، یکپارچه‌سازی داده‌ها، کاهش بعد و رده‌بندی باکتری‌ها با استفاده از روش‌های یادگیری ماشین و انتخاب دو روش قانون بیز و شبکه عصبی عمیق به عنوان روش‌های بهتر و مقایسه نتایج صحت این دو روش با مجموعه ویژگی‌هایی که هرکدام از روش‌های وزن‌دهی ارائه داده‌اند، می‌باشد.

در این مطالعه از میان ۷۳ ویژگی، ۱۳ ویژگی مستقل مؤثر در رده‌بندی باکتری براساس مقاومت به گرما بر مبنای رأی‌گیری با استفاده از ۱۰ روش وزن‌دهی استخراج گردید. ویژگی‌های مؤثر انتخاب شده با استفاده از روش‌های وزن‌دهی در روش‌های یادگیری ماشین کاربرد دارند؛ اما در روش‌های یادگیری عمیق، استخراج ویژگی به صورت سلسله‌مراتبی انجام می‌شود؛ بدین‌معنا که در لایه‌های اول، ویژگی‌های اولیه و در لایه‌های بعد از ویژگی‌های لایه قبل به صورت سلسله‌مراتبی، ویژگی‌های خوب برای افزایش صحت رده‌بندی استخراج می‌شوند. روش شبکه عصبی عمیق با استخراج سلسله‌مراتب ویژگی‌ها در لایه‌های مختلف توانست با دقت ۹۲/۴۲ درصد باکتری‌ها را براساس مقاومت به گرما رده‌بندی نماید.

یافتن یا ساختن آنزیم‌های حرارتی به عنوان هدف مهم در تعدادی از صنایع مختلف شناسایی شده است؛ بنابراین درک ویژگی‌های مربوط به ثبات حرارت آنزیمی بسیار اهمیت داشته و از روش‌های مختلفی برای استخراج و یا تولید آنزیم‌های حرارتی پایدار استفاده شده است. ۲۹۴۶ ویژگی که به حرارت پروتئین کمک می‌کنند، پیش‌تر مورد بررسی قرار گرفته‌اند. در این راستا از روش‌های مختلف یادگیری ماشین مانند انتخاب ویژگی، روش‌های خوشه‌بندی و مدل‌های درخت تصمیم‌گیری استفاده شده است (۱۶). علت بالاتر بودن دقت جداسازی باکتری‌های مقاوم به گرما از باکتری‌های غیر مقاوم در این مطالعه آن است که شبکه عصبی عمیق به خوبی می‌تواند با استفاده از لایه‌های عمیق خود به درستی استخراج ویژگی را انجام دهد و نیاز به انتخاب ویژگی ندارد.

نتایج حاصل از پژوهش حاضر می‌توانند در شناسایی باکتری‌های مقاوم به گرما که در صنایعی از جمله تولید شربت گلوکز و فروکتوز، ساخت منسوجات و به ویژه در پودرهای شستشو بسیار حائز اهمیت می‌باشند، مورد استفاده قرار گیرند.

نتیجه‌گیری

مدل‌های پیش‌بینی مبتنی بر آمار، یادگیری ماشین و به ویژه یادگیری عمیق، امکانات جدیدی را در تحلیل، تشخیص و رده‌بندی توالی پروتئین ارائه کرده‌اند. آزمایشات و پیامدهای آن‌ها در این زمینه به شدت در حال رشد هستند. در پژوهش حاضر از ویژگی‌های پروتئین‌های مقاوم و غیر مقاوم به گرما از قبیل ویژگی‌های ساختاری اسید آمینه‌ها، تعداد و فرکانس هر اسید آمینه و ویژگی‌های فیزیکوشیمیایی آن‌ها به منظور بازنمایی توالی پروتئین به ویژگی استفاده شد. مهم‌ترین ویژگی‌های تأثیرگذار بر شناسایی باکتری‌های مقاوم به گرما، فرکانس‌های گلوتامین و اسید گلوتامیک می‌باشند. در این مطالعه مجموعه داده‌ها برای رده‌بندی باکتری‌ها براساس سه رویکرد متفاوت مورد آنالیز قرار گرفت. در رویکرد اول از یک روش وزن‌دهی برای انتخاب ویژگی‌های موثر استفاده شد، در رویکرد دوم از رأی‌گیری روش‌های وزن‌دهی، ویژگی‌های مؤثر انتخاب شدند و در رویکرد سوم از شبکه‌های عصبی عمیق به منظور استخراج سلسله‌مراتب ویژگی‌ها

استفاده گردید. در رویکرد سوم شبکه عصبی عمیق به دلیل اینکه از ساختار سلسله‌مراتبی برای استخراج ویژگی استفاده می‌کند، نیاز به انتخاب ویژگی ندارد و در مجموعه داده‌هایی که نمونه‌های زیادی دارد، عملکرد بهتری را نسبت به روش‌های رده‌بندی یادگیری ماشین از خود نشان می‌دهد. در این مطالعه بیشترین دقت رده‌بندی باکتری‌های مقاوم به گرما در رویکرد اول برابر با ۸۷/۳۱ درصد، در رویکرد دوم معادل ۸۷/۷۴ درصد و در رویکرد سوم برابر با ۹۲/۴۲ به دست آمد. نتایج دقت بالا در رده‌بندی باکتری‌ها به لحاظ مقاومت به گرما، گامی مهم در شناسایی این باکتری‌ها بر مبنای توالی اسید آمینه پروتئین آن‌ها می‌باشد. به منظور دستیابی به این مهم، در مرحله اول توالی اسید آمینه پروتئین به ویژگی‌های ساختاری اسید آمینه‌ها، تعداد و فرکانس هر اسید آمینه و ویژگی‌های فیزیکوشیمیایی بازنمایی شدند. در مرحله دوم از طریق رأی‌گیری ویژگی‌های مؤثر انتخاب شده با روش‌های وزن‌دهی، ویژگی‌های مؤثرتر انتخاب شدند.

در این مطالعه از روش‌های مختلف یادگیری ماشین به منظور رده‌بندی باکتری استفاده شد. در مرحله سوم استخراج ویژگی به صورت سلسله‌مراتب در شبکه عصبی عمیق انجام شد. در این رویکرد، دقت رده‌بندی باکتری مقاوم به‌گرما افزایش یافت. افزایش دقت تشخیص دو کلاس باکتری مقاوم و غیر مقاوم به‌گرما با توجه به پرفایده بودن باکتری‌های مقاوم به گرما به ویژه در تولید شیرین‌کننده‌ها، تولید شربت گلوکز و فروکتوز، ساخت منسوجات و پودرهای شستشو و غیره بر مبنای توالی اسید آمینه پروتئین باکتری اهمیت بالایی دارد. در این راستا به پژوهشگران پیشنهاد می‌شود با تغییر بازنمایی روی توالی اسید آمینه پروتئین، دقت رده‌بندی باکتری را با رویکردهای یادگیری عمیق مورد بررسی قرار دهند؛ به عنوان مثال در پژوهش حاضر موفق شدیم با بازنمایی توالی پروتئین به دو شکل تصویر باینری و سری زمانی با دقت ۱۰۰ درصد، سویه‌های مختلف HA (Hemagglutinin) و NA (Neuraminidase) ویروس آنفولانزا را رده‌بندی نماییم.

References:

1. Zhang C, Zheng G, Xu SF, Xu D. Computational challenges in characterization of bacteria and bacteria-host interactions based on genomic data. J Comput Sci Technol 2012;27(2):225-39. Link

2. Banerjee AK, Ravi V, Murty US, Sengupta N, Karuna B. Application of intelligent techniques for classification of bacteria using protein sequence-derived features. Appl Biochem Biotechnol 2013;170(6):1263-81. PMID: 23657902

3. Berezovsky IN, Shakhnovich EI. Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci U S A 2005;102(36):12742-7. PMID: 16120678

4. Fujita M, Kanehisa M. Comparative analysis of DNA-binding proteins between thermophilic and mesophilic bacteria. Genome Inform 2005;16(1):174-81. PMID: 16362920

5. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol 2016;12(7):878. PMID: 27474269

6. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Advances in neural information processing systems. Vancouver: Neural Information Processing Systems location; 2014. P. 3320-8. Link

7. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 2015;347(6218):1254806. PMID: 25525159

8. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics 2014;30(12):i121-9. PMID: 24931975

9. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature Biotechnol 2015;33(8):831-8. PMID: 26213851

10. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods 2015;12(10):931-4. PMID: 26301843

11. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50(8):1171-1179. PMID:30013180

12. Ahsan R, Ebrahimi M. Image processing techniques represent innovative tools for comparative analysis of proteins. Comput Biol Med 2020;117:103584. PMID: 32072976

13. Paloheimo M, Mäntylä A, Kallio J, Puranen T, Suominen P. Increased production of xylanase by expression of a truncated version of the xyn11A gene from Nonomuraea flexuosa in Trichoderma reesei. Appl Environ Microbiol 2007;73(10):3215-24. PMID: 17384308

14. Yang HM, Yao B, Meng K, Wang YR, Bai YG, Wu NF. Introduction of a disulfide bridge enhances the thermostability of a Streptomyces olivaceoviridis xylanase mutant. J Ind Microbiol Biotechnol 2007;34(3):213-8. PMID: 17139507

15. Yang HM, Yao B, Fan YL. Recent advances in structures and relative enzyme properties of xylanase. Sheng Wu Gong Cheng Xue Bao 2005;21(1):6-11. PMID: 15859321

16. Ebrahimie E, Ebrahimi M. Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms. EXCLI 2009;8:218-33. Link