مجله پژوهشهای سلولی و مولکولی (مجله زیست شناسی ایران) جلد ۳۷، شماره ۱، ۱۴۰۳
DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

پیش‌بینی تعاملات بین RNA و پروتئین با استفاده از ترنسفورمرها و شبکه‌های عصبی دوقلو ناهمسان

نیکتا گوهری صدر، آرمین بهجتی و فاطمه زارع میرک آباد*

ایران، تهران، دانشگاه صنعتی امیرکبیر، دانشکده ریاضی و علوم کامپیوتر

تاریخ دریافت: ۱۴۰۰/۱۱/۰۲ تاریخ پذیرش: ۱۴۰۱/۰۷/۱۲

چکیده

تعاملات RNA و پروتئین نقش مهمی در فرآیندهای سلولی بنیادی موثر در بیماری‌های انسان، حیوانات، گیاهان و همچنین تنظیمات بیان ژن دارند. با این حال، الگو و نحوه انتخاب این تعاملات به خوبی درک نشده‌اند. همچنین به دلیل هزینه‌بر و زمان‌بر بودن روش‌های آزمایشگاهی، نیاز به توسعه روش‌های محاسباتی معتبر وجود دارد. پیش‌بینی این تعاملات، نیازمند بررسی اطلاعات ساختاری مولکول‌ها می‌باشد، در حالی که این اطلاعات همیشه در دسترس نیست. از طرفی، نتیجه تحقیقات روی مدل‌های ترنسفورمر نشان می‌دهد که آن‌ها می‌توانند به خوبی از توالی‌های RNA و پروتئین اطلاعات بیوشیمیایی، بیوفیزیکی و ساختاری مهمی را استخراج کنند. در این تحقیق، از دو ترنسفورمر ProtAlbert و DNABERT استفاده شده تا نمایش مناسبی از ویژگی توالی‌های RNA و پروتئین ساخته شود. بردارهای ویژگی استخراج شده به یک مدل یادگیری عمیق دوقلو ناهمسان داده شد تا تعاملات بین این دو مولکول را پیشگویی کند. نتایج بدست آمده نشان داد که روش پیشنهادی این تحقیق با داشتن میانگین دقت ۹۲,۳ درصد و میانگین مساحت زیر منحنی ۹۶,۶ درصد در مقایسه با روش‌های موجود بهتر عمل می‌کند.

واژه‌های کلیدی: DNABERT، یادگیری عمیق، ProtAlbert

* نویسنده مسئول، تلفن: ۰۲۱۶۶۴۶۰۹۴۸، پست الکترونیکی: f.zare@aut.ac.ir

مقدمه

تعاملات بین پروتئین‌ها و RNAها تاثیر **مستقیم بر** فعالیت‌های ابتدایی موجودات زنده دارند [۱۵]. این تعاملات می‌توانند در فرآیندهای بنیادی سلولی مانند همانندسازی کروموزوم، انتقال مواد، رونویسی و ترجمه نقش داشته باشند [۲]. در ضمن بصورت خاص دیده شده که تعامل RNA و پروتئین باعث ایجاد مقاومت گیاهان به تنش‌های محیطی مانند شوری یا سرما می‌شود [۱]. بنابراین پیشگویی و درک این تعاملات می‌توانند تاثیر زیادی بر تحقیقات آسیب‌شناسی و طراحی دارو داشته باشند. روش‌های آزمایشگاهی به دلیل وقت‌گیر و هزینه‌بر بودن نتوانسته‌اند بررسی همه جانبه‌ای در این زمینه داشته باشند. به همین دلیل محققان به روش‌های محاسباتی برای پیش‌بینی تعامل بین پروتئین و RNA علاقه‌مند شده‌اند. از جمله این روش‌های محاسباتی که امروزه بیشتر مورد استفاده قرار می‌گیرد می‌توان به الگوریتم‌های یادگیری ماشین و یادگیری عمیق اشاره نمود. این روش‌ها را بطور کلی می‌توان به دو دسته مبتنی بر توالی [۴، ۶، ۱۲، ۱۳، ۲۰] و مبتنی بر ترکیب توالی و ساختار [۹، ۱۴، ۱۹] تقسیم کرد.

از روش‌های مبتنی بر توالی می‌توان به مدل RPISeq که در سال ۲۰۱۱ پیشنهاد گردید، اشاره نمود [۱۲]. در این روش از طبقه‌بندهای RF (Random forest) و SVM (Support vector machine) برای انجام پیش‌بینی تعامل بین RNA و پروتئین استفاده می‌شود. در سال ۲۰۱۶، پروژه IPMiner با

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

استفاده از روش فراوانی ۳تایی و ۴تایی (3-mer and 4-mer frequency) توالی‌ها را رمزگذاری کرده و در نهایت یک مدل تعمیم پشته‌ای (Stacked ensemble) می‌سازد [۱۳]. پس از آن در سال ۲۰۱۹، الگوریتم CFPR معرفی گردید که با استفاده از انتقال غیرخطی بر روی فراوانی kتایی‌ها می‌تواند ویژگی‌های پیچیده‌تری از توالی استخراج کند. در نهایت، با استفاده از طبقه‌بند RF ابعاد این ویژگی‌ها کاهش می‌یابد تا به‌عنوان خروجی، تعاملات پیش‌بینی گردد [۶]. چنگ و همکارانش نیز در سال ۲۰۱۹ با فراوانی ۳تایی و ۴تایی بردار ویژگی از توالی‌های پروتئین و RNA می‌سازند و تعامل بین آن دو را با SVM، RF و CNN (Convolutional neural network) پیش‌بینی می‌کنند [۴]. در سال ۲۰۲۰ وانگ و همکارانش برای استخراج ویژگی‌ها از شبکه عصبی CNN استفاده کرده و سپس به یک شبکه یادگیری ماشین شدید (Extreme learning machine) می‌دهد [۲۰]. گرچه این روش‌ها فقط به توالی مولکول‌ها برای پیشگویی تعامل نیاز دارند و به سادگی قابل اجرا هستند ولی در تشخیص روابط بین نوکلئوتیدها و اسیدآمینه‌ها که نشان‌دهنده مفاهیم ساختاری مولکول هستند، ضعیف عمل می‌کنند. این ضعف تاثیر جدی در کاهش صحت پیشگویی دارد زیرا پیش‌بینی اینکه یک جفت RNA و پروتئین با یکدیگر تعامل دارند یا خیر وابسته به داشتن ساختار مولکول‌ها می‌باشد.

از روش‌های مبتنی بر توالی و ساختار می‌توان به RPITER در سال ۲۰۱۹ اشاره کرد [۱۴]. این تحقیق، یک معماری سلسله مراتبی یادگیری عمیق طراحی کرده که توالی و ساختار را بعنوان ورودی گرفته و روش CTFE (Conjoint triad feature encoding) را اعمال می‌کند. همچنین فن و همکارانش در سال ۲۰۱۹ ترکیب شبه نوکلئوتید و اسید آمینه را در نظر گرفته و از رگرسیون خطی برای پیش‌بینی تعاملات بین این دو مولکول استفاده می‌کند [۹]. در سال ۲۰۲۱، الگوریتم دیگری به نام EDLMFC معرفی گردید که اطلاعات توالی، ساختار دوم و سوم RNA و پروتئین را دریافت کرده و سپس تعامل بین RNA و پروتئین توسط شبکه‌های CNN و BLSTM (Bidirectional long short-term memory)، پیش‌بینی می‌گردد [۱۹]. هرچند این روش‌ها دقت بالاتری در پیشگویی تعامل بین RNA و پروتئین دارند ولی چالش جدی آن‌ها در دسترس نبودن همیشگی اطلاعات ساختارهای دوم یا سوم مولکول است.

همان‌طور که در بالا اشاره شد، روش‌های مبتنی بر توالی توانمند هستند که تعامل بین دو مولکول را تنها با در دسترس بودن توالی پیشگویی نمایند. هرچند با در نظر نگرفتن اطلاعات ساختاری، این روش‌ها بطور معمول عملکرد ضعیفی دارند. مزیت روش‌های مبتنی بر توالی و ساختار این است که دقت بالایی در پیشگویی تعاملات دارند ولی با این چالش مواجه هستند که در صورت در دسترس نبودن ساختار مولکول‌ها قابل استفاده نیستند. هدف این تحقیق ارائه روشی است که بتواند تنها از توالی دو مولکول برای پیش‌بینی تعامل استفاده نماید و در ضمن ویژگی‌های مورد نیاز ساختاری برای پیشگویی را بدون دریافت مستقیم از ورودی استخراج کند. همچنین در این تحقیق به کم شدن زمان آموزش مدل و عدم نیاز به سخت‌افزاری با هزینه بالا نیز توجه شده است. بنابراین، ارائه مدلی با این مشخصات به استفاده از مزیت هر دو روش‌های مبتنی بر توالی و مبتنی بر ترکیب توالی و ساختار در حل مسئله تعامل دو مولکول RNA و پروتئین کمک می‌کند.

برای رسیدن به این هدف، الگوریتمی به نام TIRP (Transformers for interaction prediction between RNA and protein) (شکل ۱) ارائه شده که اگرچه مبتنی بر توالی است اما می‌تواند ویژگی‌های ساختاری را نیز استخراج کند و از آن اطلاعات در پیشگویی تعامل دو مولکول استفاده نماید. برای انجام این هدف، در الگوریتم TIRP از ترنسفورمرها که در پردازش زبان‌های طبیعی بعنوان ابزارهای قوی برای درک ساختار متن شناخته شده،

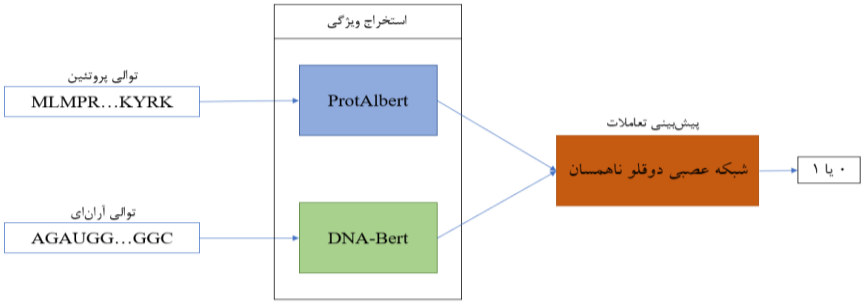
DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

استفاده گردیده است تا برداری از توالی‌ها تولید شود. این بردارها ویژگی‌های ساختاری مولکول را بصورت نهفته در خود دارند که در ادامه به طبقه‌بندی به نام شبکه عصبی دو قلوی ناهمسان (Asymmetric Siamese Neural Network) جهت پیشگویی تعاملات داده می‌شوند. با توجه به این که در این تحقیق از ترنسفورمرهای پیش‌آموزش داده شده استفاده می‌شود، برای آموزش دادن الگوریتم TIRP نیاز به سخت افزاری پیچیده‌ای نیست.

در پروژه‌ی ProtTrans چندین مدل مبتنی بر ترنسفورمر بر روی توالی‌های پروتئین منتشر شده که شامل دو مدل خود همبسته (Auto-regressive) به نام‌های XLNet و Transofrmer-XL و چهار مدل خود رمزگذار (Autoencoder) به نام‌های BERT (Bidirectional encoder representations from transformers)، Albert، Electra و T5 می‌شوند [۸]. با توجه به ماهیت این تحقیق، مدل‌های خود همبسته کمکی به ما نمی‌کنند و در میان مدل‌های خود رمزگذار، به علت بهینه بودن و عدم نیاز به سخت افزار پیچیده و در عین حال نتایج مشابه با سایر مدل‌ها، ترنسفورمر Albert انتخاب شد [۱۱]. بنابراین، در الگوریتمTIRP، ابتدا برای تبدیل یک توالی پروتئین به بردار عددی از ترنسفورمر ProtAlbert [۸] استفاده شده است. این ترنسفورمر مبتنی بر BERT [۷] بوده و بعنوان یکی از بهترین ترنسفورمرهای پیش‌آموزش (Pre-train) داده شده روی توالی‌های پروتئین، شناخته می‌شود. در سال۲۰۲۰ ویگ و همکارانش [۱۸] نشان دادند که ترنسفورمرهای بر پایه BERT می‌توانند زبان توالی‌های مولکولی را درک کنند و اطلاعات ساختاری و دیگر ویژگی‌های زیستی موثر را به خوبی استخراج نمایند. بنابراین بدون در دسترس داشتن ساختار، می‌توان یک نمایش عددی از ویژگی‌های بیوشیمایی، بیوفیزیکی و ساختاری پروتئین تولید نمود. در گام بعدی، برای تبدیل یک توالی RNA به بردار عددی از ترنسفورمر DNABERT [۱۰] استفاده شده است. جی و همکارانش [۱۰] در سال ۲۰۲۱ نشان دادند که این ترنسفورمر به خوبی ویژگی‌های زیستی مولکول RNA را می‌تواند از توالی استخراج کند. در نهایت برای پیش‌بینی تعاملات بین RNA و پروتئین، خروجی این دو ترنسفورمر به یک معماری از نوع شبکه عصبی دوقلو داده شده است. این شبکه عصبی دو بردار که شامل ویژگی‌های نهفته زیستی است را از دو فضای متفاوت به یک فضا منتقل می‌کند. سپس در صورت وجود تعامل بین دو مولکول، بردار ویژگی‌ها در فضای انتقال داده شده به هم نزدیک و در صورت عدم تعامل در فضا از یکدیگر دور می‌شوند.

الگوریتم TIRP روی سه پایگاه داده‌های RPI488 [۱۳]، NPInter v2.0 [۲۱] و RPI1807 [۱۶] اجرا شده است. ارزیابی این الگوریتم در دو مرحله انجام شد. در مرحله اول بررسی گردید که طبقه‌بند شبکه عصبی دوقلوی ناهمسان در ساختار TIRP از طبقه‌بندهای کلاسیک مانند RF، SVM، NN (Neural network) برای پیشگویی تعامل بین دو مولکول مناسب‌تر است. برای انجام این تحلیل بردارهای استخراج شده از ترنسفورمرها به مدل‌های کلاسیک داده شد. مقایسه نتایج آن‌ها با TIRP نشان داد که شبکه عصبی دوقلوی ناهمسان بهتر از طبقه‌بندهای دیگر در پیشگویی تعامل بین دو مولکول عمل می‌کند. سپس الگوریتم TIRP با تعدادی از مدل‌های مبتنی بر توالی و مبتنی بر ترکیب توالی و ساختار مقایسه شد. نتایج مقایسه‌ی دقت الگوریتم‌ها، نشان داد که اگرچه معماری TIRP ساده است و نیاز به سخت افزار پرهزینه‌ای برای اجرا ندارد، میزان دقت بالاتری نسبت به روش‌های مبتنی بر توالی دارد و در ضمن قابل رقابت با روش‌های مبتنی بر ترکیب توالی و ساختار است.

مقاله پژوهشی، صفحات ۹۱-۷۴
DOI: 10.22034/CMR.2022.2200



شکل ۱- نمای کلی از الگوریتم TIRP

مواد و روشها

در این بخش ابتدا مسئله تعامل RNA و پروتئین (RPI= RNA Protein Interaction) و تعاریف اولیه مورد نیاز ارائه می‌گردد، سپس روش پیشنهادی (TIRP) برای حل مسئله RPI و جزئیات آن شرح داده می‌شود. با توجه به این که در بخش نتایج روش پیشنهادی با مدل‌های کلاسیک مانند RF، SVM و NN مقایسه می‌گردد، در این بخش توضیح مختصری هم درباره طبقه‌بندهای کلاسیک داده می‌شود. در ادامه پایگاه دادهای مورد نیاز و معیاری‌های ارزیابی معرفی می‌گردد.

مسئله تعامل RNA و پروتئین: هر توالی RNA مانند R با طول m بصورت

$$R = r_1 ... r_m, \forall r_i \in N,$$

نمایش داده می‌شود بطوری که مجموعه N نشان‌دهنده چهار نوع نوکلئوتید است.

هر توالی پروتئین P به طول n بصورت

$$P = p_1 ... p_n, \forall p_i \in A,$$

نمایش داده می‌شود بطوری که مجموعه A نشان‌دهنده بیست نوع اسید آمینه است.

براساس دو توالی داده شد RNA و پروتئین، مسئله RPI بصورت زیر تعریف می‌گردد:

- ورودی: دو توالی R و P
- خروجی: در صورت وجود تعامل بین دو مولکول، یک و در غیر این صورت صفر تولید می‌گردد.

ترنسفورمرها: ترنسفورمرها نوعی از مدل‌های یادگیری عمیقی با معماری رمزگذار(Encoder) و رمزگشا (Decoder) هستند که با مکانیزم توجه (Attention mechanism) می‌توانند وابستگی‌های متنی را به خوبی تشخیص دهند. یکی از بهترین این ترنسفورمرها برای شناسایی روابط اجزایی متن، معماری BERT [۷] می‌باشد که الگوریتمی دوطرفه (Bidirectional) و بدون ناظر است. یکی از بزرگترین مزیت‌های معماری BERT توانایی درک جملاتی با طول های مختلف و به خاطر سپردن جملات بسیار طولانی می‌باشد. با اینکه توالی‌های زیستی نیز می‌توانند بعنوان زبان دیده شوند اما استفاده مستقیم BERT برای حل مسائل زیستی، منجر به نتایجی خوبی نخواهد شد. در نتیجه، مدل‌های پیش‌آموزش داده شده‌ای از این معماری مانند دو ترنسفورمر DNABERT [۱۰] و ProtAlbert [۸] ساخته شده‌اند که بترتیب توالی مولکول‌های نوکلئوتیدی و پروتئینی را بعنوان ورودی دریافت کرده و از آن‌ها ویژگی استخراج می‌کنند.

ترنسفورمر DNABERT: ترنسفورمرDNABERT [۱۰] روی ژنوم انسان و براساس معماری BERT پیش‌آموزش داده شده است که دارای ۱۲ لایه (Layer) با ۷۶۸ نورون پنهان و ۱۲ هد توجه (Attention head) در هر لایه می‌باشد. این ترنسفورمر روی توالی‌های DNA آموزش داده شده است. با تبدیل باز یورسیل (Uracil) به تیمین (Thymine) در RNA، می‌توان از این ترنسفورمر بمنظور استخراج ویژگی برای توالی‌های RNA استفاده نمود [۱۰].

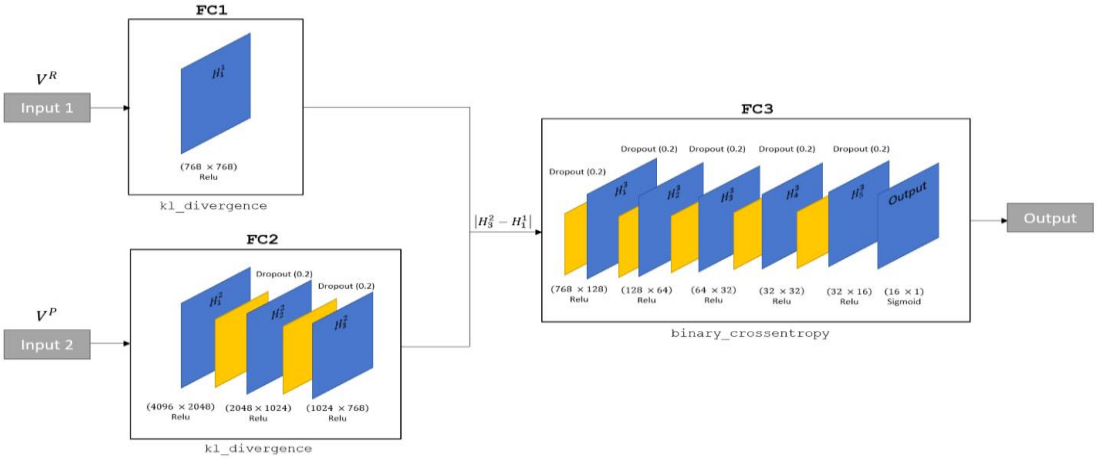
مقاله پژوهشی، صفحات ۹۱-۷۴
DOI: 10.22034/CMR.2022.2200

این ترنسفورمر توالی‌های با طول حداکثر ۵۱۲ دریافت کرده و برداری به طول ۷۶۸ را بعنوان خروجی می‌سازد.

ترنسفورمر ProtAlbert: این ترنسفورمر از معماری Albert که نسخه توسعه یافته BERT است، استفاده می‌کند. ترنسفورمر Albert با کاهش حجم محاسبات و توانایی اجرا روی توالی‌های بلندتر قابل رقابت با ترنسفورمر BERT است. بنابراین ما در این تحقیق برای کد کردن پروتئین از نسخه ProtAlbert [۸] که پیش‌آموزش داده شده Albert روی ۲۱۶ میلیون توالی پروتئین پایگاه داده Uniref100 [۱۷] است، استفاده می‌کنیم. معماری این نسخه شامل ۱۲ لایه و ۶۴ هد توجه است. در ضمن بهجتی و همکارانش [۳] نشان دادند که این نسخه از ترنسفورمر می‌تواند پنج ویژگی پروتئین شامل نزدیک‌ترین تعامل با همسایه، نوع اسیدآمینه، اطلاعات بیوشیمی و بیوفیزیکی اسیدآمینه‌ها و اطلاعات ساختار دوم و سوم را تنها براساس توالی پروتئین تشخیص دهد که شناسایی این ویژگی‌ها می‌تواند تاثیر زیادی در پیش‌بینی تعاملات بین RNA و پروتئین داشته باشد. این ترنسفورمر به ازای هر توالی پروتئین برداری به طول ۴۰۹۶ بعنوان خروجی تولید می‌کند.

شبکه عصبی دوقلو ناهمسان: در مسئله‌های زیستی روش‌های متفاوت خطی مانند فاصله اقلیدسی برای محاسبه فاصله بردارها استفاده می‌شود. اخیرا روش‌های جدیدی در شبکه های عصبی به نام شبکه‌های عصبی دوقلو معرفی شده که می‌توانند از یک شبکه عصبی با وزن مشترک برای مقایسه دو بردار استفاده نمایند. اگر دو ورودی متعلق به یک گونه باشند، ابعاد بردارها را در فضا بصورتی تغییر می‌دهد که اختلاف آن‌ها کم و نزدیک به صفر و در غیر این صورت اختلاف دو ورودی نزدیک به یک شود. این شبکه همچنین توانایی یادگیری بهتر با تعداد داده کم را دارد. شبکه عصبی دوقلوی ناهمسان از دو شبکه عصبی با معماری متفاوت تشکیل شده است که اطلاعات پنهان در دو بردار ورودی را تشخیص دهد. هر دو شبکه عصبی پیشخور (Feedforward) بوده و از پس انتشار خطا (Backpropagation) در طول یادگیری استفاده می‌کنند تا فاصله بین دو بردار محاسبه نماید [۵].

در این تحقیق از معماری شبکه‌های عصبی دوقلو ناهمسان استفاده شده تا بتوان میزان شباهت دو بردار عددی پروتئین و RNA که از ترنسفورمرها استخراج شده و شامل ویژگی‌های پنهان زیستی درتوالی‌ها است را محاسبه نمود. این مدل هر دو بردار را به فضای یکسانی منتقل می‌کند بطوری که در صورت عدم تعامل دو مولکول، فاصله آن‌ها در این فضایی جدید زیاد و در صورت وجود تعامل فاصله آن‌ها در این فضا کم شود. معماری این شبکه در شکل ۲ قابل مشاهده است.



شکل۲- معماری شبکه

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

مدل پیشنهادی برای حل مسئله RPI: در این تحقیق، روشی مبتنی بر ترکیب ترنسفورمر و شبکه‌های عصبی دوقلوی ناهمسان به نام TIRP ارائه گردیده است. مراحل کلی آن بشرح زیر است (شکل ۱):

۱. ورودی: توالی RNA مانند R و توالی پروتئینی P

۲. استخراج ویژگی از توالی‌های داده شده:

أ. استخراج بردار ویژگی عددی به طول ۴۰۹۶ به نام $V^P=[v_1^P...v_{4096}^P]$ از توالی پروتئینی P براساس ترنسفورمر پیش‌آموزش داده شده ProtAlbert.

ب. استخراج بردار ویژگی عددی به طول ۷۶۸ به نام $V^R=[v_1^R...v_{768}^R]$ از توالی RNA مانند R براساس ترنسفورمر پیش‌آموزش داده شده DNABERT.

۳. استفاده از شبکه عصبی دوقلوی ناهمسان برای پیشگویی تعامل دو مولکول (شکل ۲):

أ. ورودی شبکه عصبی دو بردار V^R و V^P می‌باشد.

ب. بردار V^R به یک شبکه تمام همبند به نام FC1 داده می‌شود که پارامترهای آن درجدول ۱ نشان داده شده است.

ج. بردار V^P به یک شبکه تمام همبند به نام FC2 داده می‌شود که پارامترهای آن در جدول ۲ نشان داده شده است.

د. تفاضل خروجی لایه اول FC1 (H_1^1) و لایه سوم FC2 (H_3^2) بعنوان ورودی به شبکه تمام همبند FC3 ($|H_3^2-H_1^1|$) داده می‌شود که پارامترهای آن در جدول ۳ نشان داده شده است.

ه. تولید کردن صفر یا یک در لایه خروجی شبکه عصبی دوقلوی ناهمسان بترتیب نشان دهنده تعامل نداشتن یا تعامل داشتن یک جفت RNA و پروتئین است.

جدول ۱- جزئیات پیاده سازی شبکه FC1.

نام لایه‌ها	H_1^1
تعداد نورون‌ها	۷۶۸
تابع فعال ساز	Relu
بهینه ساز (Optimizer)	Adam (نرخ یادگیری = ۰,۰۰۰۱)
تابع زیان (Loss function)	K1 divergence

جدول ۲- جزئیات پیاده سازی شبکه FC2

نام لایه‌ها	H_1^2	H_2^2	H_3^2
تعداد نورون‌ها	۲۰۴۸	۱۰۲۴	۷۶۸
تابع فعال ساز	Relu	Relu	Relu
دراپ اوت	۰,۲	۰,۲	———
بهینه ساز	Adam (نرخ یادگیری = ۰,۰۰۰۱)		
تابع زیان	K1 divergence		

جدول ۳- جزئیات پیاده سازی شبکه FC3.

نام لایه‌ها	H_1^3	H_2^3	H_3^3	H_4^3	H_5^3	*Output*
تعداد نورونها	۱۲۸	۶۴	۳۲	۳۲	۱۶	۱
تابع فعال ساز	Relu	Relu	Relu	Relu	Relu	Sigmoid
دراپ اوت	۲.۰	۲.۰	۲.۰	۲.۰	۲.۰	
بهینه ساز	Adam (نرخ یادگیری = ۰,۰۰۰۱)					
تابع زیان	Binary Cross Entropy					

با توجه به اینکه در ادامه می‌خواهیم طبقه‌بند شبکه عصبی دوقلوی ناهمسان در الگوریتم TIRP را با طبقه‌بندهای کلاسیک مقایسه نماییم، در این زیر بخش سه نسخه از الگوریتم TIRP تعریف می‌کنیم که شامل طبقه‌بندهای کلاسیک RF، SVM و NN برای پیشگویی تعامل RNA است. این نسخه‌ها براساس نوع طبقه‌بند $TIRP^{RF}$، $TIRP^{SVM}$ و $TIRP^{NN}$ بترتیب نامگذاری شده‌اند. به هر سه نسخه بردار الحاق شده V^P و V^R که استخراج شده از ترنسفورمرها است، بعنوان ورودی داده شد. برای هریک از مدل‌ها پارامترهای متفاوتی بررسی گردید و سپس بهترین آن‌ها برای مقایسه با TIRP با طبقه‌بند شبکه عصبی

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

دوقلو ناهمسان انتخاب شد. در الگوریتم RF تعداد درخت‌های ۵۰، ۱۰۰، ۲۰۰ و ۳۰۰ مورد بررسی قرار گرفت. در مدل SVM توابع کرنل خطی، چندجمله‌ای از درجه ۴ و ۸، سیگموید و RBF (Radial Basis Function) ارزیابی شدند. مدل NN از یک لایه تا ۶ لایه همراه با دراپ اوت ۰,۲، تابع فعال ساز Relu در لایه‌های مخفی و تابع سیگموید در لایه خروجی برای پیشگویی تعامل RNA و پروتئین مورد تحلیل قرار گرفت.

پایگاه داده: برای آموزش و ارزیابی این تحقیق از پایگاه داده‌های RPI488 [۱۳]، NPInter v2.0 [۲۱] و RPI1807 [۱۶] استفاده شده است. پایگاه داده RPI1807، داده‌های خود را از PRIDB (Nucleic acid database) و NDB (Protein-RNA interface database) استخراج کرده که شامل ۱۰۷۸ توالی RNA و ۳۱۳۱ پروتئین می‌باشد. این پایگاه داده در مجموع دارای ۱۸۰۷ جفت تعامل (مثبت) و ۱۴۳۶ جفت عدم تعامل (منفی) است. مجموعه داده NPInter v2.0 از پایگاه داده NPInter گرفته شده که شامل تعاملات فیزیکی بین RNA و پروتئین می‌باشد. این پایگاه داده از ۴۶۳۶ نوع RNA، ۴۴۹ نوع پروتئین و در مجموع ۱۰۴۱۲ جفت مثبت تشکیل شده است. مجموعه داده RPI488 نیز با داشتن ۲۵ نوع RNA، ۲۴۷ نوع پروتئین شامل ۲۴۳ جفت مثبت و ۲۴۵ جفت منفی می‌باشد. باتوجه به اینکه پایگاه داده NPInter فاقد داده عدم تعامل است، ما داده‌های خود را از مقاله [۱۹] (https://github.com/JingjingWang-87/EDLMFC) استخراج کردیم که برای این پایگاه داده نیز داده منفی تولید کرده است. در این تحقیق، بدلیل محدودیت ترنسفورمر DNABERT، RNA های با طول حداکثر ۵۱۲ و بدلیل محدودیت‌های سخت افزاری، پروتئین‌های با طول حداکثر ۱۰۰۰ برای آموزش و ارزیابی مدل استفاده شده است. تعداد جفت‌هایی که باهم در تعامل هستند (جفت‌های مثبت) و جفت‌هایی که در تعامل نیستند (جفت‌های منفی) در جدول ۴ نشان داده شده است.

جدول۴- تعداد جفت‌های مثبت و منفی.

جفت‌های منفی	جفت‌های مثبت	پایگاه داده
۲۸۶	۵۵۴	RPI1807
۱۶۸۵	۱۷۹۳	NPInter v2.0
۲۳۸	۲۱۰	RPI488

معیارهای ارزیابی: در این تحقیق، ۷ معیار ارزیابی که در مقاله‌ها [۶، ۱۳، ۱۴، ۱۹] عموما برای بررسی عملکرد مدل در پیشگویی تعامل RNA و پروتئین استفاده می‌شود، معرفی می‌گردد. این معیارها شامل دقت (ACC= Accuracy) ، حساسیت (TPR= True positive rate)، تشخیص (TNR= True negative rate)، (PPV= Positive predictive value)، (F1= F1 score)، (MCC= Matthews correlation coefficient) و مساحت زیر منحنی (AUC= Area under curve) می‌باشد. رابطه هرکدام از معیارها در ادامه شرح داده می‌شود.

- معیار ACC، توانایی طبقه‌بندی مدل روی تمام داده‌ها را نشان می‌دهد که بصورت زیر است:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

در این رابطه، TP (True positive) به معنی تعداد جفت‌های RNA و پروتئینی است که باهم در تعامل بوده و بدرستی پیش‌بینی می‌شوند. مقدار TN (True negative) نشان‌دهنده تعداد جفت‌هایی است که باهم در تعامل نیستند و بدرستی پیشگویی می‌گردند. تعداد جفت‌هایی که باهم درتعامل نبوده اما مدل آن‌ها را اشتباه پیش‌بینی می‌نماید با FP (False positive) نمایش داده می‌شود. درنهایت، FN (False negative) تعداد جفت‌هایی که باهم در تعامل بوده و توسط مدل به اشتباه پیش‌بینی شده اند را نشان می‌دهد.

- معیار TPR، توانایی مدل در تشخیص جفت‌های مثبت را نشان می‌دهد که:

$$TPR = \frac{TP}{TP + FN}.$$

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

- معیار TNR، توانایی مدل روی جفت‌های منفی را نشان می‌دهد که:

$$TNR = \frac{TN}{TN + FP}.$$

- معیار PPV، توانایی مدل در تشخیص صحیح جفت‌های مثبت نسبت به کل داده‌ای که مثبت پیشگویی می‌شود را نمایش می‌دهد که:

$$PPV = \frac{TP}{TP + FP}.$$

- معیار MCC، عملکرد مدل هنگامی که تعداد جفت‌های مثبت و منفی در تعادل نیستند را نشان می‌دهد که:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

- معیار F1، معیار جامعی است که با در نظر گرفتن TPR و PPV توانایی مدل را می‌سنجد که:

$$F1 = \frac{2 \times TPR \times PPV}{TPR + PPV}.$$

نتایج و بحث

در این بخش، ما به ارزیابی الگوریتم TIRP در دو گام می‌پردازیم. ابتدا نشان داده می‌شود که انتخاب طبقه‌بند شبکه عصبی دوقلوی ناهمسان در TIRP مناسب‌تر از طبقه‌بندهای کلاسیک مانند RF، SVM و NN است. سپس عملکرد الگوریتم پیشنهادی با روش‌های موجود که برای پیشگویی تعامل RNA و پروتئین اخیرا ارائه شده، مقایسه می‌گردد.

مقایسه با روش‌های کلاسیک طبقه‌بندی: در این زیربخش نسخه اصلی TIRP که دارای طبقه‌بند شبکه دوقلوی ناهمسان است برای ارزیابی با سه نسخه $TIRP^{RF}$، $TIRP^{SVM}$ و $TIRP^{NN}$ که مبتنی بر طبقه‌بندهای کلاسیک RF ، SVM و NN هستند، مقایسه می‌گردد. ابتدا هر یک از نسخه‌ها براساس پارامترهای متفاوت اموزش داده شده و به روش اعتبارسنجی متقابل ۵ تایی (5-fold cross validation) تست شده‌اند. سپس مقدار میانگین AUC به ازای هر مدل با پارامترهای متفاوت محاسبه گردیده و بهترین پارامترها برای هر طبقه‌بند در جدول ۵ مشخص شده است. مقایسه نتایج مقدار AUC در نسخه‌های TIRP به ازای طبقه‌بندهای متفاوت در جدول ۶ قابل مشاهده است. این جدول نشان می‌دهد که طبقه‌بند شبکه عصبی دوقلوی ناهمسان از دو طبقه‌بند RF و SVM بطور محسوسی بهتر عمل می‌کند. در طبقه‌بند NN نیز با وجود این‌که عمق لایه‌ها مطابق با شبکه دوقلو ناهمسان در نظر گرفته شده، اما همچنان TIRP میزان AUC بیشتری را نشان می‌دهد.

جدول۵- بهترین پارامترها برای مدل‌های کلاسیک طبقه‌بندها.

مقدار پارامتر	نوع پارامتر	طبقه‌بند
۵۰	تعداد درخت	RF
خطی	تابع کرنل	SVM
۶	تعداد لایه	NN

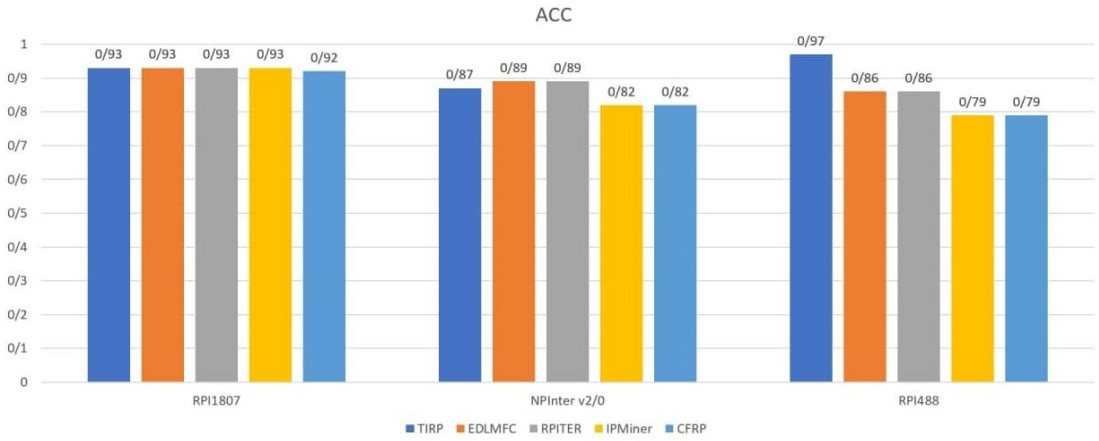
جدول ۶- مقایسه الگوریتم TIRP با روش‌های کلاسیک طبقه‌بندی براساس معیار AUC.

روش‌ها	RPI1807	NPInter v2.0	RPI488
TIRP	۰,۹۸	۰,۹۳	۰,۹۹
$TIRP^{RF}$	۰,۹۰	۰,۸۵	۰,۹۵
$TIRP^{SVM}$	۰,۹۰	۰,۷۷	۰,۹۱
$TIRP^{NN}$	۰,۹۸	۰,۹۲	۰,۹۸

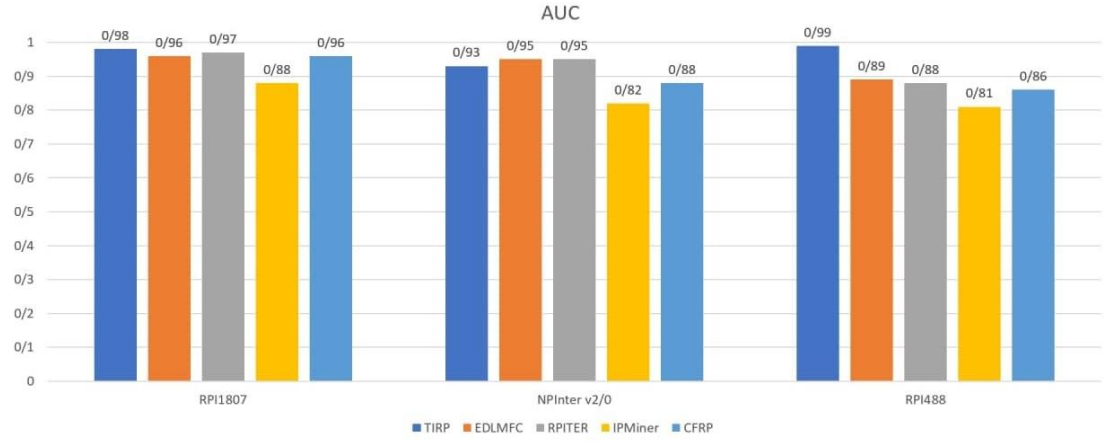
مقایسه الگوریتم TIRP با روش‌های نوین موجود: برای بررسی بهتر عملکرد TIRP، این الگوریتم را با روش‌های نوین موجود مانند EDLMFC [۱۹]، RPITER [۱۴]، IPMiner [۱۳] و CFRP [۶] مقایسه نمودیم (نمودارهای ۱و۲). دو روش EDLMFC و RPITER مبتنی بر ترکیب توالی و ساختار هستند و دو روش IPMiner و CFRP مبتنی بر توالی می‌باشند. مقادیر معیارهای ارزیابی این

مقاله پژوهشی، صفحات ۹۱-۷۴
DOI: 10.22034/CMR.2022.2200

روش‌هایی که در نمودارهای ۱ و ۲ و جدول ۷ نشان داده‌شده، از مقاله [۱۹] استخراج شده است. الگوریتم TIRP نیز از پایگاه داده مقاله [۱۹] استفاده کرده و مشابه روش‌های دیگر با روش اعتبارسنجی متقابل ۵ تایی تست شده است.



نمودار۱- مقایسه روش‌های نوین موجود با TIRP براساس معیار ACC.



نمودار۲- مقایسه روش‌های نوین موجود با TIRP براساس معیار AUC.

نمودار ۱ میزان دقت (ACC) هر یک از الگوریتم‌های TIRP، EDLMFC، RPITER، IPMiner و CFRP را روی سه پایگاه داده نشان می‌دهد. تقریبا همه الگوریتم‌ها روی پایگاه داده RPI1807 در حدود ۹۳ درصد دقت دارند. در پایگاه داده NPInter v.20، روش TIRP از نظر دقت قابل رقابت با EDLMFC و RPITER است. برتری روش TIRP به این دو روش در این است که بدون اطلاعات ساختاری به دقت آن‌ها نزدیک است. دقت روش TIRP برروی پایگاه داده RPI488 بیشتر از همه روش‌ها است. می توان ادعا کرد دلیل این موضوع به علت عملکرد مناسب شبکه‌های دوقلو روی داده‌های کم است زیرا مجموعه داده‌های این پایگاه داده کمتر از بقیه پایگاه داده‌ها است.

نمودار ۲ مقدار AUC هر یک الگوریتم‌ها را روی سه پایگاه داده نشان می‌دهد. به ازای این معیار روی پایگاه داده RPI1807، تقریبا سه روش TIRP، EDLMFC، RPITER مشابه عمل می‌کنند که نشان می‌دهد گرچه به مدل پیشنهادی در این تحقیق اطلاعات ساختاری مولکول‌ها داده نشده ولی ترنسفورمرها بصورت قابل قبولی اطلاعات ساختاری را استخراج می‌کنند. مقدار AUC روی پایگاه

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

داده NPInter v.20 بر روی این نکته تایید می‌کند که مدل ما قابل رقابت با مدل‌های مبتنی بر توالی و ساختار است. در ضمن برروی پایگاه داده RPI488، روش TIRP بصورت چشمگیری مقدار AUC را افزایش می‌دهد که این نیز تایید بر وجود طبقه‌بند مناسب یعنی شبکه‌های دوقلو ناهمسان است که خیلی کارا روی داده‌های با حجم کم هستند.

در جدول ۷ می‌توان نتایج معیارهای TPR، TNR، PPV، F1 و MCC را روی پایگاه داده‌های متفاوت به ازای الگوریتم‌های گوناگون مقایسه نمود. افزایش یا کاهش میزان معیارهای TPR نشان می‌دهد که مدل میزان پیشگویی در تعامل بودن را خوب یا بد تشخیص داده است. اصولا افزایش یا کاهش این پارامتر رابطه معکوس با TNR دارد. این پارامتر نشان دهنده میزان پیشگویی عدم تعامل است. در الگوریتم TIRP، تفاضل این دو معیار تقریبا در همه پایگاه داده‌ها ۴ درصد است. این نشان می‌دهد این روش برخلاف روش‌های دیگر فقط روی یکی از حالت‌های کلاس تشخیص تعامل یا عدم تعامل بایاس نمی‌شود. همچنین با توجه به اینکه معیارهای F1 و MCC جفت‌های در تعامل یا عدم تعامل پیشگویی شده را بصورت همزمان بررسی می‌کند، می‌توانیم ادعا کنیم که در این معیارها، الگوریتم TIRP با روش‌های موجود مقایسه شده، قابل رقابت است.

جدول ۷– مقایسه براساس دیگر معیارهای ارزیابی

پایگاه داده	روش	TPR	TNR	PPV	F1	MCC
RPI1807	TIRP	۰,۹۶	۰,۹۰	۰,۹۷	۰,۹۷	۰,۸۷
	EDLMFC	۰,۹۶	۰,۸۴	۰,۹۴	۰,۹۵	۰,۸۳
	RPITER	۰,۹۷	۰,۸۲	۰,۹۴	۰,۹۵	۰,۸۲
	IPMiner	۰,۹۹	۰,۷۶	۰,۹۲	۰,۹۵	۰,۸۲
	CFRP	۰,۹۷	۰,۷۷	۰,۹۲	۰,۹۵	۰,۷۹
NPInter v2.0	TIRP	۰,۹۱	۰,۸۷	۰,۸۸	۰,۸۹	۰,۷۹
	EDLMFC	۰,۹۱	۰,۸۷	۰,۸۸	۰,۸۹	۰,۷۹
	RPITER	۰,۹۱	۰,۸۶	۰,۸۷	۰,۸۹	۰,۷۸
	IPMiner	۰,۸۴	۰,۸۱	۰,۸۱	۰,۸۳	۰,۶۵
	CFRP	۰,۷۷	۰,۸۶	۰,۸۵	۰,۸۱	۰,۶۴
RPI488	TIRP	۱	۰,۸۵	۰,۸۵	۰,۹۲	۰,۸۵
	EDLMC	۰,۷۴	۰,۹۶	۰,۹۶	۰,۸۲	۰,۷۴
	RPITER	۰,۷۵	۰,۹۵	۰,۹۵	۰,۸۲	۰,۷۴
	IPMiner	۰,۸۴	۰,۷۸	۰,۷۹	۰,۷۹	۰,۶۳
	CFRP	۰,۷۵	۰,۸۵	۰,۸۲	۰,۷۷	۰,۶۰

نتیجه‌گیری

در این مقاله، روش جدیدی برای پیش‌بینی تعاملات بین RNA و پروتئین پیشنهاد شده است. این روش ابتدا از توالی‌های RNA توسط ترنسفورمرهای ProtAlbet و DNABERT ویژگی استخراج کرده و سپس تعامل داشتن یا نداشتن آن دو را با شبکه عصبی دوقلوی ناهمسان

DOI: 10.22034/CMR.2022.2200 مقاله پژوهشی، صفحات ۹۱-۷۴

پیشگویی می‌کند. ارزیابی‌ها نشان داد که این مدل با داشتن میانگین دقت ۹۲,۳ درصد و میانگین مساحت زیر منحنی ۹۶,۶ درصد دقت از روش‌های نوین موجود عملکرد بهتری دارد. در آینده، این روش را می‌توان با تنظیم دقیق (**Fine tuning**) ترنسفورمر DNABERT، روی توالی‌های RNA بهبود بخشید. در ضمن نوع ویژگی از تعاملات در لایه‌های ترنسفورمرها را می‌توان بررسی کرد.

منابع

۱- اقدسی، م. ۱۳۹۲. بررسی پروتئومیکی گیاهان تراریخت شده با RBP2-GR در مقایسه با گیاهان وحشی. مجله پژوهشهای سلولی و مولکولی (مجله زیست شناسی ایران)، جلد ۲۶ ،شماره ۲ ص، ۱۵۴-۱۶۳.

۲- پورشیخعلی اصغری، م. و عبدالمالکی، پ. ۲۰۱۵. پیشگویی عملکرد اتّصال پروتئینها به ریبونوکلئیک اسید بر اساس خواص فیزیکوشیمیایی آنها به کمک روش لوژستیک رگرسیون. مجله پژوهش‌های سلولی و مولکولی (مجله زیست شناسی ایران)، جلد ۲۸، شماره ۴۵-۵۳ ص، ۱

3- Behjati A., Zare-Mirakabad F., Arab S. S., and Nowzari-Dalini A., Jan. 2021, “Protein sequence profile prediction using ProtAlbert transformer”. *bioRxiv*, p. 2021.09.23.461475, doi: 10.1101/2021.09.23.461475.

4- Cheng S., Zhang L., Tan J., Gong W., Li C., and Zhang X., 2019, “DM-RPIs: Predicting ncRNA-protein interactions using stacked ensembling strategy”. *Computational biology and chemistry*, vol. 83, p. 107088.

5- Chicco D., 2021, “Siamese neural networks: An overview”. *Artificial Neural Networks*, pp. 73–94.

6- Dai Q., Guo M., Duan X., Teng Z., and Fu Y., 2019, “Construction of complex features for computational predicting ncRNA-protein interaction”. *Frontiers in genetics*, vol. 10, p. 18.

7- Devlin J., Chang M.-W., Lee K., and Toutanova K., 2018, “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv preprint arXiv:1810.04805*.

8- Elnaggar A. *et al.*, 2020, “ProtTrans: towards cracking the language of Life’s code through self-supervised deep learning and high performance computing”. *arXiv preprint arXiv:2007.06225*.

9- Fan X.-N. and Zhang S.-W., 2019, “LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier”. *Neurocomputing*, vol. 370, pp. 88–93.

10- Ji Y., Zhou Z., Liu H., and Davuluri R., 2021, “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome”. *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120.

11- Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., and Soricut R., 2019, “Albert: A lite bert for self-supervised learning of language representations”. *arXiv preprint arXiv:1909.11942*.

12- Muppirala U. K., Honavar V. G., and Dobbs D., 2011, “Predicting RNA-protein interactions using only sequence information”. *BMC bioinformatics*, vol. 12, no. 1, pp. 1–11.

13- Pan X., Fan Y.-X., Yan J, and Shen H.-B., 2016, “IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction”. *BMC genomics*, vol. 17, no. 1, pp. 1–14.

14- Peng C., Han S., Zhang H., and Li Y., 2019, “RPITER: a hierarchical deep learning framework for ncRNA–protein interaction prediction”. *International journal of molecular sciences*, vol. 20, no. 5, p. 1070.

15- Rinn J. L. and Ule J., “’Oming in on RNA–protein interactions”. *Genome biology*, vol. 15, no. 1. Springer, pp. 1–3, 2014.

16- Suresh V., Liu L., Adjeroh D., and Zhou X., Feb. 2015, “RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information”. *Nucleic Acids Research*, vol. 43, no. 3, pp. 1370–1379, doi: 10.1093/nar/gkv020.

17- Suzek B. E., Wang Y., Huang H., McGarvey P. B., Wu C. H., and Consortium U., 2015, “UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches”. *Bioinformatics*, vol. 31, no. 6, pp. 926–932.

مقاله پژوهشی، صفحات ۹۱-۷۴ DOI: 10.22034/CMR.2022.2200

18- Vig J., Madani A., Varshney L. R., Xiong C., Socher R., and Rajani N. F., 2020, "Bertology meets biology: Interpreting attention in protein language models". *arXiv preprint arXiv:2006.15222*.

19- Wang J. *et al.*, 2021, "EDLMFC: an ensemble deep learning framework with multi-scale features combination for ncRNA–protein interaction prediction". *BMC bioinformatics*, vol. 22, no. 1, pp. 1–19.

20- Wang L., You Z.-H., Huang D.-S., and Zhou F., 2018, "Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions". *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 3, pp. 972–980.

21- Yuan J., Wu W., Xie C., Zhao G., Zhao Y., and Chen R., Jan. 2014, "NPInter v2.0: an updated database of ncRNA interactions". *Nucleic Acids Research*, vol. 42, no. D1, pp. D104–D108, doi: 10.1093/nar/gkt1057.

RNA-Protein interaction prediction using transformers and asymmetric Siamese neural network

Gohari sadr N., Behjati A. and Zare-Mirakabad F.

Dept. of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, I.R. of Iran

Abstarct

RNA-protein interactions play essential roles in many biological processes, such as gene regulation and fundamental cellular processes related to human, animal, and plant diseases. However, the patterns of these interactions are not fully understood. The experimental methods to solve this problem are expensive and time-consuming. Therefore, there is a compelling need for developing reliable computational methods. Predicting these interactions requires structural information about RNA and protein, which is not always available. On the other hand, results of the research on transformers show that they can efficiently extract biochemical, biophysical, and structural features from molecule sequences. In this experiment, we use ProtAlbert and DNABERT transformers to provide a good representation for RNA and protein sequences. Then we feed the feature vectors to an asymmetric Siamese network to predict whether they interact with each other or not. The experimental results indicate that our method achieves superior performance with an average accuracy of 92.3% and an average area under the curve of 96.6%.

Keywords: DNABERT, Deep Learning, ProtAlbert