پیشگویی عملکرد اتّصال پروتئینها به ریبونوکلئیک اسید بر اساس خواص فیزیکوشیمیایی آنها به کمک روش لوژستیک رگرسیون

مهدی پورشیخعلی اصغری و پرویز عبدالمالکی*

تهران، دانشگاه تربیت مدرس، دانشکده علوم زیستی، گروه بیوفیزیک

تاریخ دریافت: ۹۲/۴/۳۰ تاریخ پذیرش: ۹۳/۷/۹

چکیده

بار، ممان دوقطبی و مقادیر ویژه ممان چهارقطبی به عنوان خواص فیزیکوشیمیایی برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید در این تحقیق مورد استفاده قرار گرفتند. از روش ساده و کارآی لوژستیک رگرسیون برای انجام عمل پیشگویی استفاده شد. در معادله لوجیت حاصل، پارامتر بار بیشترین ضریب (۱۹/۱۹) و بیشترین تأثیر را در پیشگویی داشت و بعد از آن پارامتر ممان دوقطبی اهمیت داشت. روش لوژستیک رگرسیون به صورت Jackknife بر روی ۲۶۰۱ پروتئین (۱۶۰ پروتئین متّصل شونده به ریبونوکلئیک اسید و ۲۴۴۱ پروتئین غیر متّصل شونده به ریبونوکلئیک اسید) آموزش داده شد. مقدار پارامتر ارزیابیِ مساحت زیر منحنی ROC (AUC) مدل نهایی، ۸۳ درصد به دست آمد که در قیاس با روش شبکه عصبی به کار رفته برای پیشگویی خیلی بیشتر می باشد. مقادیر پارامترهای دقت، صحت و معیار F نیز به ترتیب ۹۴، ۵۹ و ۴۶ درصد به دست آمدند که همگی بیشتر از مقادیر حاصل از روش شبکه عصبی می باشند. در نتیجه، در این تحقیق نشان داده شد که به کمک روش ساده، سریع و دقیقِ لوژستیک رگرسیون، پروتئینهای متّصل شونده به ریبونوکلئیک اسید می توانند به خوبی از پروتئینهای غیر متّصل شونده به ریبونوکلئیک اسید (به کمک تعداد اندکِ خواص پیشگوی فیزیکوشیمیایی) متمایز شوند.

واژه های کلیدی: پروتئینهای متّصل شونده به ریبونوکلئیک اسید، خواص فیزیکوشیمیایی پروتئینها، لوژستیک رگرسیون، پیشگویی عملکرد پروتئین

* نویسنده مسئول، تلفن : ۰۹۱۲۲۷۰۴۵۰۱، پست الکترونیکی: parviz@modares.ac.ir

مقدمه

میانکنشهای پروتئین- ریبونوکلئیک اسید (RNA) نقشی اساسی را در بیشتر فرآیندهای سلّولی مثلِ رونویسی، رونویسی معکوس، همانندسازی، انتقالِ RNA، پردازشِ پس از رونویسی، ترجمه و تنظیم مقادیر RNA در سلّول ایفاء می کنند (۹). داده ها و شواهد ژنتیکی و پروتئومیکی جدید حاصل از مدلهای حیوانی آشکار می سازند که پروتئینهای متّصل شونده به ریبونوکلئیک اسید یا به اختصار RBPs در بیشتر بیماریهای انسانی، از بیماریهای عصب شناختی گرفته تا سرطان، دخیل هستند (۱۶). لیکن، مکانیسم تشخیصِ RNA توسط این پروتئینها یا بالعکس به میزان کمی شناخته شده است.

بیشتر RBPs ساختارهایی یکتا دارند و از تکرارهای متعدّدی از تعداد کمی دُومینهای اصلی تشکیل می شوند که این دُومینها به طرق مختلف آرایش می یابند تا نیازمندیهای عملکردی متنوع آنها را برآورده سازند. دُومینهای اصلی متّصل شونده به ریبونوکلئیک اسید در برگیرنده موتیفِ تشخیص ریبونوکلئیک اسید ، دُومینِ کا همولوژی، دُومین متّصل شونده به ریبونوکلئیک اسید دو زنجیره ای، دُومین S1، دُومین PIWI، دُومین TRAP،

دُومین SAM، دُومینِ Pumilio و دُومینِ PAZ هستند. این دومینها از نظر نوع توپولوژی (α، β، αβ)، سطح تشخیص دهنده ریبونوکلئیک اسید (صفحه بتا یا مارپیچ آلفا) و نوع میانکنشهایی که با ریبونوکلئیک اسید برقرار می سازند (پیوند الکتروستاتیک، پیوند هیدروژنی، پیوند واندروالس) با یکدیگر تفاوت دارند (۱۷).

روشهای محاسباتی گوناگونی برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید در سالهای اخیر تکامل و توسعه یافته اند. بخش عمدهٔ آنها از روش ماشین بردار پشتیبان برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید به کمکِ اطّلاعات توالی و ساختار استفاده کرده اند (۷، ۸، ۱۰، ۱۱، ۱۴، ۲۱، ۲۲، ۲۴ و ۲۷). بخشی کوچک از این روشهای محاسباتی نیز از تکنیکهای هم ترازی ساختاری سرتاسری و تکنیکِ تشخیص فولد برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید استفاده کرده اند (۲۸،۲۶و۲۹). در این مطالعات، خواص الکتروستاتیک اصلی پروتئینها مثلِ ممانهای دوقطبی و چهارقطبی در ترکیب با سایر پارامترهای حاصل از توالی یا ساختار در نظر گرفته شده اند و لذا نقش انفرادی این خواص فیزیکوشیمیایی در تشخیص RNA مخفی مانده است. روشهای دیگری نیز برای بررسی میانکنشهای پروتئین–ریبونوکلئیک اسید موجود می باشند لیکن هدف آنها بررسی ساختاری می باشد نه پیشگویی عملکرد (۱۳و۱۸).

در میانکنش پروتئین– دِزوکسی ریبونوکلئیک اسید (DNA) الکتروستاتیک نقشی اصلی را بر عهده دارد. دلیلِ آن باردار بودن خیلی منفی DNA می باشد (منفی ۲ بار به ازای هر جفت باز). از طرف دیگر، پروتئینهای متصل شونده به DNA روی سطحی که با DNA مواجه می شوند، اغلب بارهای مثبت زیادی دارند. نقش کلیدی بارِ مثبت پروتئین ایجاد یک جاذبه الکتروستاتیک برای DNA می باشد (۵).

Ahmad and Sarai در سال ۲۰۱۱ نشان دادند که خواص الکتروستاتیک ساده شامل بار، ممان دوقطبی و ممان چهار قطبی حاوی اطلاعات ارزشمند و مفیدی برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید می باشند. آنها تنها از روش شبکه عصبی مصنوعی برای پیشگویی استفاده کردند (۳).

شبکه عصبی مصنوعی یک کلاس از روشهای یادگیری ماشینی می باشد که به خاطر توانایی بالای آن در سرو کار داشتن با سری داده هایی با ابعاد زیاد و پُرخش Noisy و نیز پارامترهای پیچیده و غیر خطی، در بیشتر مسائل بیوانفورماتیکی تا به حال استفاده شده است. در واقع شبکه عصبی مصنوعی رفتار یک شبکه عصبی زیستی را شبیه سازی می کند. در یک شبکه عصبی مصنوعی، واحدهای پردازش ابتدایی (موسوم به نورونها) در لایه های ورودی، مخفی و خروجی سازمان بندی می شوند. فرآیند آموزش در شبکه عصبی مصنوعی بر اساسِ میان اتصالات بینِ نورونهای تشکیل دهنده شبکه می باشد. پس از اتمامِ مرحله آموزش، شبکه عصبی مصنوعی قادر به پیشگویی خوبِ نمونه های مربوط به سری آزمون یا تست می باشد (۱۵).

در مطالعه جاری نقش انفرادیِ خواص الکتروستاتیک پروتئینها شاملِ بار ، ممان دوقطبی و ممان چهار قطبی در پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید بررسی خواهد شد که تا به حال انجام نشده است. سؤال اصلی تحقیق حاضر این می باشد که آیا تنها به کمک این پارامترهای اندکِ فیزیکوشیمیایی می توان با دقّت خوبی عملکرد اتّصال پروتئینها به ریبونوکلئیک اسید را پیشگویی کرد؟

Sziagyi and Skolnick در سال ۲۰۰۶ پژوهشی انجام دادند که در آن از ترکیبِ فراوانیهای آمینواسیدهای ویژه در توالی پروتئین و ممان دوقطبی مولکول برای پیشگویی عملکرد اتّصال پروتئینها به دِزوکسی ریبونوکلئیک اسید

استفاده کردند. آنها از روش سریع و توانمند لوژستیک رگرسیون برای پیشگویی استفاده کردند و مشاهده کردند روش مزبور قابلیت پیشگویی عملکرد اتّصال پروتئینها به دِزوکسی ریبونوکلئیک اسید را با دقّت بالایی دارد (۲۵).

لوژستیک رگرسیون یک مدل آماری خطی تعمیم یافته می باشد که توانایی پیشگویی یک خروجی مجزا را ازیک سری متغیرهای پیوسته، گسسته یا دوحالته دارد. به طور کلی، متغیر پاسخ در لوژستیک رگرسیون دو حالته می باشد مثلِ وجود/عدم وجود یا پیروزی/شکست. لوژستیک رگرسیون می تواند برای برآوردِ احتمال یک اتفاق ویژه (در اینجا ، عملکرد اتّصال پروتئین به ریبونوکلئیک اسید) استفاده شود (۱۲). در این تحقیق با استفاده از خواص الکتروستاتیکی بار، ممانهای دوقطبی و چهار قطبی به کمکِ روش لوژستیک رگرسیون عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید پیشگویی می گردد. در نهایت، نتایج به دست آمده از پیشگویی با روش لوژستیک رگرسیون با تنها روش قبلاً به کار رفته بر روی این خواص فیزیکوشیمیایی (شبکه عصبی مصنوعی) مقایسه خواهد شد .

مواد و روشها

سریهای داده (بانک اطلاعاتی): بانک اطلاعاتی پروتئینهای متّصل شونده به ریبونوکلئیک اسید (RBPs) از ۱۶۰ زنجیره پروتئینی غیر همولوگ تشکیل می شد (۳). این زنجیره های پروتئینی حداقل با یک زنجیره RNA در تماس هستند که حداقل تعداد رزیدوهای دخیل آنها در تماس با RNA، ۳ می باشد و افزونگی که مبنای کاهش ابعاد بانک اطلاعاتی بود به کمک نرم افزار BLASTCLUST (۴) در سطح ۲۵ درصد تعیین گردید. بانک اطلاعاتی پروتئینهای غیر متّصل شونده به ریبونوکلئیک اسید (Non-RBPs) نیز از ۲۴۴۱ زنجیره پروتئینی تشکیل می شد (۳). این زنجیره ها هیچ گونه شباهتی با RBPs نداشتند و از پایگاه PDBselect (۶) با سطحِ افزونگی ۲۵ درصد برداشته شده بودند.

محاسبه پارامترها: در این تحقیق از پارامترهای محاسبه شده به وسیله Ahmad and Sarai در سال ۲۰۱۱ (۳) برای پیشگویی استفاده گردید. لیست کامل پارامترها و مشخصات آنها در جدول ۱ آورده شده است.

جدول ۱ – مشخصات پارامترهای به کار رفته در مدل

پارامتر	مشخصات	واحد
بار کلّی پروتئین	برحسبِ بار اوّلیه الکترون (e)	$1e \approx 1.6\times10^{-19}$ coulombs
ممان دوقطبی پروتئین	بر حسبِ Debye (دبای)	$1Debye \approx 3.33564\times10^{-30}$ coulomb meters
ممان چهار قطبی پروتئین	بر حسبِ $eÅ^2$	$Coulomb\ metre^2$ $1Å(angstrom)=1.0\times10^{-10}meters$

طریقه محاسبه هر یک از پارامترها در آن تحقیق از این قرار است: برای محاسبه تمامی پارامترها، مختصات زنجیره جانبی پروتئینها نادیده گرفته شد و ممانهای الکتریکی براساسِ کنفورماسیون زنجیره اصلی (تعیین شده توسط موقعیت C_α رزیدوها) محاسبه شدند. برای محاسبه بار، تمامی رزیدوهای لیزین و آرژینین دارای یک بار مثبت و تمامی رزیدوهای آسپارتیک اسید و گلوتامیک اسید دارای یک بار منفی در نظر گرفته شدند و تمامی سایر رزیدوها خنثی در نظر گرفته شدند . هیستیدین به علت اثرات ناچیز حالتهای باردارش خنثی در نظر گرفته شد. در این محاسبات، مولکولهای آب، یونهای فلزی و لیگاندها به حساب نیامدند (۳).

اجزای ممانهای دوقطبی با استفاده از عبارت زیر محاسبه شدند (۳):

$$F = \sum(R_i - R_0)q_i \qquad (۱)$$

که در آن R_0 نقطه مرجع می باشد، که به عنوان مرکز هندسی تمامی رزیدوها (موقعیتهای C_α) در ساختار در نظر گرفته می شود، و i نشان دهنده یک اتم در ساختار پروتئین می باشد. ممان دوقطبی کلی پروتئین به وسیله جمع برداری این اجزا به دست می آید ($p = \sqrt{(P_x)^2 + (P_y)^2 + (P_z)^2}$).

ممان چهارقطبی دارای نه جزء می باشد ($M_{xx}, M_{xy}, M_{xz}, M_{yx}, M_{yy}, M_{yz}, M_{zx}, M_{zy}, M_{zz}$) . هر یک از این اجزا توسط معادله زیر محاسبه می شوند (۳) :

$$M_{\alpha\beta} = 1/2\sum(3r_{i\alpha}r_{i\beta} - r_i^2\,\delta_{\alpha\beta})q_i \qquad (۲)$$

که در آن r_i بردار موقعیت نسبی است، i شاخص بار است و عملیات جمع شامل تمامی بارها می باشد. ماتریس ممان چهار قطبی می تواند قطری شود و ۳ مقدار ویژه آن با ترتیب کاهشی می توانند با Q1، Q2 و Q3 نمایش داده شوند. تمامی مقادیر ممانهای الکتریکی با درنظر گرفتن طول توالی پروتئین نرمالیزه شدند (۳).

لوژستیک رگرسیون: لوژستیک رگرسیون دوتایی رابطه بین یک متغیر پاسخ دوحالته (در اینجا ،متّصل شدن یا نشدن یک پروتئین به RNA) و یک سری متغیرهای پیشگو را توصیف می کند. از نظر ریاضی، خروجی مدل لوژستیک مقدار متغیر پاسخ نمی باشد بلکه احتمال به دست آوردن مقدار ۱ (همان متّصل شدن به RNA) را در قیاس با مقدار ۰ (همان متّصل نشدن به RNA) به دست می دهد. چون احتمال مابین ۰ تا ۱ می باشد، رگرسیون خطی نامناسب برای پیشگویی مقدار مستقیم آن می باشد لذا از تبدیل لوژستیکِ احتمال برای پیشگویی استفاده می شود :

$$Logit(p) = \frac{p}{1-p} \qquad (۳)$$

درحالی که احتمال(p) از ۰ تا ۱ تغییر می یابد، لوجیتِ p از منهای بینهایت تا مثبت بینهایت تغییر می یابد و لوجیت عدد نیم، صفر می باشد. فرم اصلی معادله لوژستیک رگرسیون به صورت زیر می باشد: (۴)

$$Logit(p) = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

که در آن $x_1,x_2,\ldots,x_n$ متغیرهای پیشگو یا مستقل می باشند و $b_1,b_2,\ldots,b_n$ ضرایب معادله می باشند (۱۲).

شرایط شبیه سازی با مدل لوژستیک رگرسیون در جدول ۲ آمده است. در این تحقیق از تابع glm نرم افزار R ورژنِ 3.0.1 (بنیاد محاسبه آماری آر، وِین، اتریش) برای انجام پیشگویی با روش لوژستیک رگرسیون استفاده شده (۲۰). روشِ Jackknife برای ارزیابی روش لوژستیک مورد استفاده قرار گرفت. در این روش ، یک پروتئین از کل بانک جدا گشته و عملیات آموزش مدل به کمک سایر پروتئینها انجام می شود. سپس مدل آموزش دیده برای پیشگویی عملکرد اتصال پروتئین جدا گشته به کار می رود. این عملیات به تعداد کل پروتئینهای بانک تکرار می گردد (یعنی هربار عملکرد اتّصال یک پروتئین پیشگویی می شود). در نهایت مدل با توجه به اجرا بر روی تک تک پروتئینها ارزیابی می گردد. برای انجام روش Jackknife از تابعِ CVbinary کتابخانه DAAG نرم افزار R ورژن 3.0.1 استفاده شد (۱۹). واحد دقّت یک صدم اعشار می باشد.

جدول ۲ – شرایط شبیه سازی با مدل لوژستیک رگرسیون

شرایط	مورد استفاده شده
نرم افزار	R ورژنِ 3.0.1
کتابخانه	stat ,DAAG
نوع تابع	glm,CVbinary
نوع تابع اتصال	logit
روش ارزیابی	Jackknife

معیارهای ارزیابی مدل: چهار پارامتر وابسته به آستانه برای ارزیابی اجرای روش لوژستیک رگرسیون استفاده شد. این چهار پارامتر از چهار اندیس زیر مشتق می شوند: اندیسِ تی پی (تعداد RBPs درست پیشگویی شده)، اندیسِ تی اِن (تعداد Non-RBPs درست پیشگویی شده)، اندیسِ اف پی (تعداد Non-RBPs که اشتباهاً به عنوان RBPs پیشگویی شده اند) و اندیسِ اف اِن (تعداد RBPs که اشتباهاً به عنوان Non-RBPs پیشگویی شده اند). فرمولهای محاسبه این پارامترها از اندیسهای ذکر شده بدین ترتیب می باشند:

(۵) صحّت (Precision) $p = \frac{TP}{TP+FP}$

(۶) حسّاسیت $Recall(r) = \frac{TP}{TP+FN}$

(۷) دقّت (Accuracy) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

(۸) معیار F (F-measure) $F - measure = \frac{2pr}{p+r}$

معیار F میانگین هندسی p و r می باشد.

به علاوه، منحنی ROC،که یک نمودار حسّاسیت (همان r) در مقابل نرخ پیشگویی اشتباه می باشد، برای ارزیابی اجرایِ روش لوژستیک رگرسیون رسم شد. مساحت زیر این منحنی یا AUC به عنوان یک معیار ارزیابی مستقل از آستانه، برای سنجشِ کارآیی مدل استفاده گردید. حداکثر مقدار آن یک می باشد و یک روش پیشگویی ضعیف (تصادفی) مقادیرِ AUC حول و حوشِ نیم دارد.

از توابعِ موجود در کتابخانه ROCR نرم افزار R ورژنِ 3.0.1 برای به دست آوردن مقادیر پارامترهای ارزیابی مدل و نیز رسم منحنی ROC استفاده شد (۲۳).

نتایج و بحث

پس از اجرای روش لوژستیک رگرسیون بر روی کل بانک، تنها متغیرهای بار و ممان دوقطبی از نظر آماری معنادار بودند (یعنی P-value < ۰/۰۵ داشتند). معادله لوجیت حاصل پس از خطی سازی شدن به این صورت به دست آمد:

$$Logit(p) = -3.78 + (19.19 \times \text{بار}) + (0.26 \times \text{ممان دوقطبی})$$

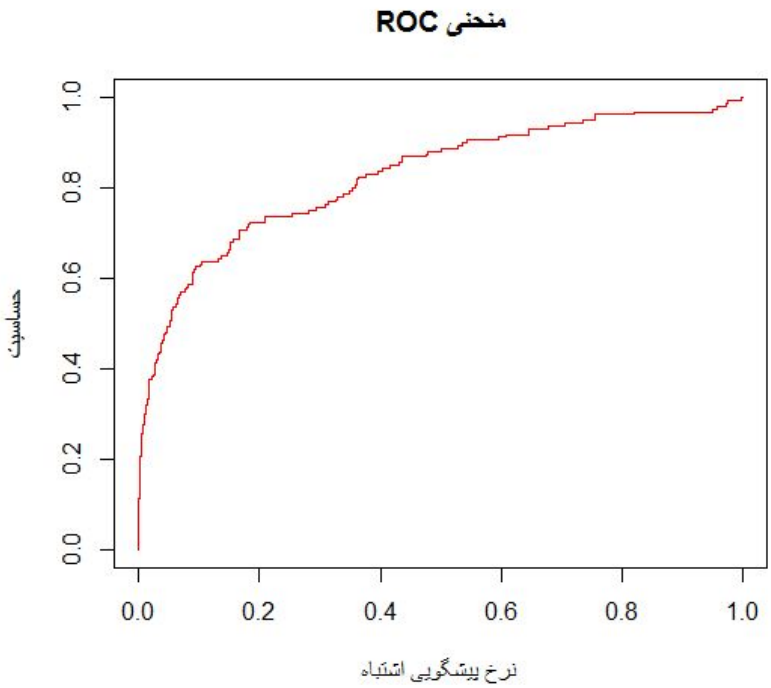
با توجه به معادله بالا مشخص می شود که قدر مطلق ضریب رگرسیون پارامتر بار در قیاس با پارامتر ممان دوقطبی خیلی بیشتر می باشد و لذا تأثیر بار در پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید دارای وزن یا اهمّیت بیشتری است. در مورد تأثیر مقادیر ویژه ممان چهار قطبی نیز می توان گفت که آنها پارامترهای غیر معنادار بوده (P-value > ۰/۰۵) و صرفاً با داشتن مقادیر پارامترهای بار و ممان دوقطبی می توان عمل پیشگویی را به خوبی انجام داد لذا این پارامترها در معادله کاهش یافته وارد نشده اند.

جدول ۳ – ارزیابی اجرای روشهای لوژستیک رگرسیون و شبکه عصبی در پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید در حالتِ Jackknife (تمامی مقادیر بر حسبِ درصد می باشند).

روش	دقّت	صحّت	حسّاسیت	معیار F	AUC
شبکه عصبی (مرجعِ شماره ۳)	91	31	45	37	78
لوژستیک رگرسیون	94	59	38	46	83

پس از اجرای روش Jackknife بر روی بانک (تعداد فولدها =۲۶۰۱)، پارامترهای ارزیابی مدل شاملِ صحّت، حساسیّت، دقّت ،معیار F و مساحت زیر نمودار ROC (AUC) با توجه به بالاترین مقدارِ پارامترِ معیار F محاسبه

شدند. مقادیر این پارامترها در جدول ۳ آورده شده اند. برای مقایسه، نتایج حاصل از پیشگویی با روش شبکه عصبی (۳) نیز در این جدول آورده شده اند. منحنی ROC روش ارائه شده در این پژوهش (به دست آمده در حالتِ Jackknife بر روی بانک)، در شکل ۱ نشان داده شده است.



شکل ۱ – منحنیِ ROC روش لوژستیک رگرسیون

با نگاهی به جدول شماره ۳ می توان دریافت که کارآیی روش لوژستیک رگرسیون در قیاس با روش شبکه عصبی از نظر ۴ معیار ارزیابیِ دقّت، صحّت، معیار F و AUC بهتر است. تنها مقدار معیار ارزیابیِ حساسیّت کمتر از مقدار به دست آمده توسط روش شبکه عصبی می باشد که نشان دهنده بالاتر بودن اندیسِ اف اِن یا به عبارت دیگر پیشگویی تعداد بیشتری از RBPs در دسته Non-RBPs می باشد که این امر به نوبه خود به دلیل نامتعادل بودن بانک اطلاعاتی و بیشتر بودن تعداد Non-RBPs (۲۴۴۱) از RBPs (۱۶۰) قابل توجیه است. متقابلاً، بیشتر بودن پارامتر صحّت دلالت بر پائین تر بودن اندیسِ اف پی دارد که نشان دهنده پیشگویی تعداد کمتری از Non-RBPs در دسته RBPs می باشد و از مزایایِ بارز روش لوژستیک رگرسیون به کار رفته در تحقیق حاضر برای صحیح تر پیشگویی کردن عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید در قیاس با روش شبکه عصبی می باشد. همچنین، مقدار معیار ارزیابی AUC روش لوژستیک رگرسیون ۵ درصد بیشتر از مقدار معادل آن در روش شبکه عصبی می باشد که با توجه به غیر وابسته بودن این معیار به آستانه، دلیل محکم تر دیگری بر کارآیی بهتر این روش نسبت به روش شبکه عصبی به دست می آید.

بهره گیری از ابزارهای آماری (مثلِ لوژستیک رگرسیون) و نیز روشهای هوش مصنوعی (مثل شبکه عصبی) برای پیشگویی ساختار و عملکرد پروتئینها پیشرفتهای قابل توجهی را در علم زیست شناسی به وجود آورده است با علم به این واقعیت که روشهای آزمایشگاهی برای تعیین ساختار و عملکرد زمان بر و هزینه بر می باشند(۱). لیکن، این دو نوع ابزار یا روش با یکدیگر تفاوت دارند. از مزیتهای بارز روش لوژستیک رگرسیون، سرعت بالا و

سادگی اجرای آن می باشد در حالی که روش شبکه عصبی پیچیدگی بالایی دارد و همچنین به صورت جعبه سیاه عمل می کند یعنی اطلاعاتی در مورد وزنِ پارامترهای مختلف دخیل در مدل پیشگویی را نمی دهد. ولی روش لوژستیک رگرسیون این قابلیت را دارد که وزنهای مختلف پارامترهای دخیل در مدل پیشگویی را بدهد. ضرایب مدل لوجیت در معادله بالا همان وزنها می باشند که نشانگر اهمّیت هر پارامتر در تعیین احتمالِ اتّصال پروتئینها به ریبونوکلئیک اسید می باشند.

دیگر مزیّت متدولوژی مطالعه جاری این حقیقت است که می توان به کمک تنها پارامترهای بار و ممان دوقطبی پروتئینهای متّصل شونده به ریبو نوکلئیک اسید را از پروتئینهای غیر متّصل شونده به ریبونوکلئیک اسید با دقّت خوبی تفکیک کرد. در حالی که در روش شبکه عصبی از پارامترهای بار، ممانهای دوقطبی و چهار قطبی برای پیشگویی استفاده شده است و دقّت تفکیک آن روش کمتر از روش تحقیق حال حاضر می باشد. با توجه به نقش پروتئینهای متّصل شونده به ریبونوکلئیک اسید در فرآیند های مختلف زیستی، از تنظیم بیان ژن گرفته تا مقاومت به تنشهای محیطی در گیاهان(۲)، اهمیت تحقیق جاری بیش از بیش آشکار می گردد.

نتیجه گیری

در این مطالعه، از روش آماری لوژستیک رگرسیون برای پیشگویی عملکردِ اتّصال پروتئینها به ریبونوکلئیک اسید به کمک خواص فیزیکوشیمیایی بار، ممانهای دوقطبی و چهار قطبی استفاده گردید. پس از اجرای مدل لوژستیک رگرسیون بر روی بانک اطلاعاتی، مشخّص گردید که پارامترهای بار و ممان دوقطبی برای پیشگویی صحیح و دقیق کافی می باشند. همچنین وزن و اهمّیت هر یک از این پارامترها با به دست آوردن معادله لوجیتِ مربوطه مشخص گردید. نتایج حاصل از مدل به کار رفته در این پژوهش با نتایج حاصل از شبکه عصبی مقایسه گردید و مشاهده شد که مدل به کار رفته در این پژوهش کارآیی خیلی بهتری در قیاس با مدل شبکه عصبی دارد. در این تحقیق نشان داده شد که روش ارائه شده در این پژوهش توانایی تفکیکِ پروتئینهای متّصل شونده به ریبونوکلئیک اسید را با دقّت بالایی (۹۴ درصد) از پروتئینهای غیر متّصل شونده به ریبونوکلئیک اسید دارد. روش ارائه شده در تحقیق حاضر، هم سریع و هم دقیق است و می تواند به آسانی در مطالعات در سطحِ پروتئوم برای پیشگویی پروتئینهای متّصل شونده به ریبونوکلئیک اسید به کار رود.

منابع

۱. اسکندری، و. یخچالی، ب. مینوچهر، ز. ۱۳۸۹. پیشگویی نرم افزاری مناطق مجاز پذیرش پپتیدهای بیگانه در زیر واحد CstH پیلی CS3 اشریشیاکلی. مجله زیست شناسی ایران، جلد ۲۳، شماره ۱، ص ۷۳-۸۴.

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

3. Ahmad S., Sarai A., 2011. Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. BMC Struct. Biol., 11:8.

4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

5. Barbi M., Paillusson F., 2013. Chapter Seven - Protein–DNA Electrostatics: Toward a New Paradigm for Protein Sliding, Pages 253-297, in the *Advances in Protein Chemistry and Structural Biology* Book Series, Volume 92, Pages 1-357, Dynamics of Proteins and Nucleic Acids, Edited by Tatyana Karabencheva-Christova.

6. Berman HM, Henrick K, Nakamura H, 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol., 10 (12), 980.

7. Cai C.Z., Han L.Y., Ji Z.L., Chen X., Chen Y.Z., 2003. SVM-Prot: web-based support vector

machine software for functional classification of a protein from its primary sequence. Nucl. Acids Res., 31, 3692-3697.

8. Cai Y-D., Lin S.L., 2003. Support vector machines for predicting rRNA-, RNA-, and DNA- binding proteins from amino acid sequence. Biochim. Biophys. Acta, 1648, 127-133.
9. Chen Y., Varani G., 2005. Protein families and RNA recognition. FEBS J., 272, 2088-2097.
10. Fujishima K., Komasa M., Kitamura S., Suzuki H., Tomita M., Kanai A., 2007. Proteome-wide prediction of novel DNA/RNA binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon *Pyrucoccus furiosus*. DNA Res., 14, 91-102.
11. Han L.Y., Cai C.Z., Lo S.L., Chung M.C.M., Chen Y.Z., 2004. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA, 10, 355-368.
12. Hosmer D.W. and Lemeshow S., 2000. Applied logistic regression (Eds: Shewh art W. A. and Wilks S. S.). John Wiley & Sons Inc, New York.
13. Huang S.Y. and Zou X. 2013. A nonredundant structure dataset for benchmarking protein-RNA computational docking. J Comput. Chem., 34, 311-318.
14. Kumar M., Gromiha M.M., Raghava G.P.S., 2011. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J. Mol. Recognit., 24, 303-313.
15. Lancashire L.J., Lemetre C. and Ball G.R. 2009. An introduction to artificial neural networks in bioinformatics _ application to complex microarray and mass spectrometry datasets in cancer studies. Brief. Bioinform, 10, 315-29.
16. Lukong K.E., Chang K.W., Khandijan E.W., Richard S., 2008. RNA-binding proteins in human genetic disease. Trends Genet., 24, 416-425.
17. Lunde B.M., Moore C., Varani G., 2007. RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol., 8, 479-490.
18. Mahdavi S., Salehzadeh-Yazdi A., Mohades A., Masoudi-Nejad A. 2013. Computational structure analysis of biomacromolecule complexes by interface geometry. Comput. Biol. Chem., 47, 16-23.
19. Maindonald J. and Braun W.J. 2013. DAAG: Data analysis and graphics data and functions. R package version 1.16. URL http://CRAN.R-project.org/package=DAAG .
20. R-package Development Core Team, 2013. R: a language and environment for statistical computing**.** R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ .
21. Shao X., Tian Y., Wu L., Wang Y., Jing L., Deng N., 2009. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J. Theor. Biol., 258, 289-293.
22. Shazman S., Mandel-Gutfreund Y., 2008. Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput. Biol., 4, e1000146, 1-14.
23. Sing T., Sander O., Beerenwinkel N. and Lengauer T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics , 21(20), 7881.
24. Spriggs R.V., Murakami Y., Nakamura H., Jones S., 2009. Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics , 25, 1492-1497.
25. Szilagyi A., Skolnick J., 2006. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J. Mol. Biol., 358, 922-933.
26. Yang Y., Zhan J., Zhao H., Zhou Y., 2012. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins Struct. Funct. Bioinf., 80, 2080-2088.
27. Yu X., Cao J., Cai Y., Shi T., Li Y., 2006. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J. Theor. Biol., 240, 175-184.
28. Zhao H., Yang Y., Zhou Y., 2011. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucl. Acids Res., 39, 3017-3025.
29. Zhao H., Yang Y., Zhou Y., 2011. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biol., 8, 988-996.

RNA-binding function prediction of proteins based on their physicochemical features using the logistic regression method

Poursheikhali Asghari M. and Abdolmaleki P.

Biophysics Dept., Faculty of Biological Sciences, Tarbiat Modares University, Tehran, I.R. of Iran

Abstract

Charge, dipole moment and quadrupole moment eigenvalues as physicochemical features have been used for the RNA-binding function prediction of proteins, in this study. Simple and efficient logistic regression method was utilized for the prediction process. In the corresponding logit equation, charge feature had the highest coefficient (19.19) and impact on the prediction and dipole moment was the second significant feature. Logistic regression was trained using jackknife procedure on 2601 protein chains (160 RNA-binding proteins and 2441 non RNA-binding proteins). The value for the performance measure of area under the curve of receiver operating characteristics (ROC) was 83% for the final model and is higher than the value obtained by the neural network method for prediction. The values of accuracy, precision and F-measure were 94%, 59% and 46%, respectively, which outperformed the neural network method. In conclusion, we showed that with the help of simple, fast and accurate logistic regression method, RNA-binding proteins can be well distinguished from non RNA-binding proteins using a few number of physicochemical predictor features.

Key words: RNA-binding proteins; Physicochemical features of proteins; Logistic regression; Protein function prediction