

## پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک 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 پذیرش: ۹۳/۷/۹

تاریخ دریافت: ۹۲/۴/۳۰

### چکیده

بار، ممان دوقطبی و مقادیر ویژه ممان چهارقطبی به عنوان خواص فیزیکوشیمیایی برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید در این تحقیق مورد استفاده قرار گرفتند. از روش ساده و کارآی لوژستیک رگرسیون برای انجام عمل پیشگویی استفاده شد. در معادله لجیت 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 متصل شونده به ریبونوکلیک اسید و ۲۴۴۱ پروتئین غیر متصل شونده به ریبونوکلیک اسید) آموزش داده شد. مقدار پارامتر ارزیابی 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 دقیق لوژستیک رگرسیون، پروتئینهای متصل شونده به ریبونوکلیک اسید می توانند به خوبی از پروتئینهای غیر متصل شونده به ریبونوکلیک اسید (به کمک تعداد اندک خواص پیشگوی فیزیکوشیمیایی) متمایز شوند.

**واژه های کلیدی:** پروتئینهای متصل شونده به ریبونوکلیک اسید، خواص فیزیکوشیمیایی پروتئینها، لوژستیک رگرسیون، پیشگویی عملکرد پروتئین

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

### مقدمه

مکانیسم تشخیص RNA توسط این پروتئینها یا بالعکس به میزان کمی شناخته شده است.

بیشتر RBPs ساختارهایی یکتا دارند و از تکرارهای متعددی از تعداد کمی ڈومینهای اصلی تشکیل می شوند که این ڈومینها به طرق مختلف آرایش می یابند تا نیازمندیهای عملکردی متنوع آنها را برآورده سازند. ڈومینهای اصلی متصل شونده به ریبونوکلیک اسید در برگیرنده موتیف تشخیص ریبونوکلیک اسید، ڈومین کا همولوژی، ڈومین متصل شونده به ریبونوکلیک اسید دو زنجیره ای، ڈومین S1، ڈومین PIWI، ڈومین TRAP،

میانکنشهای پروتئین-ریبونوکلیک اسید (RNA) نقشی اساسی را در بیشتر فرآیندهای سلولی مثل رونویسی، رونویسی معکوس، همانندسازی، انتقال RNA، پردازش پس از رونویسی، ترجمه و تنظیم مقادیر RNA در سلول ایفاء می کنند (۹). داده ها و شواهد ژنتیکی و پروتئومیکی جدید حاصل از مدل‌های حیوانی آشکار می سازند که پروتئینهای متصل شونده به ریبونوکلیک اسید یا به اختصار RBPs در بیشتر بیماریهای انسانی، از بیماریهای عصب شناختی گرفته تا سرطان، دخیل هستند (۱۶). لیکن،

دومین SAM، دومین Pumilio و دومین PAZ هستند. این دومینها از نظر نوع توپولوژی ( $\alpha$ ،  $\beta$ ،  $\alpha\beta$ )، سطح تشخیص دهنده ریبونوکلیک اسید (صفحه بتا یا ماریچ آلفا) و نوع میانکنشهایی که با ریبونوکلیک اسید برقرار می‌سازند (پیوند الکتروستاتیک، پیوند هیدروژنی، پیوند واندروالس) با یکدیگر تفاوت دارند (۱۷).

روشهای محاسباتی گوناگونی برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید در سالهای اخیر تکامل و توسعه یافته اند. بخش عمده آنها از روش ماشین بردار پشتیبان برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید به کمک اطلاعات 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 تراز ساختاری سرتاسری و تکنیک تشخیص فولد برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید استفاده کرده اند (۲۶، ۲۸ و ۲۹). در این مطالعات، خواص الکتروستاتیک اصلی پروتئینها مثل 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-ریبونوکلیک اسید موجود می‌باشند لیکن هدف آنها بررسی ساختاری می‌باشد نه پیشگویی عملکرد (۱۳ و ۱۸).

روشهای محاسباتی گوناگونی برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید در سالهای اخیر تکامل و توسعه یافته اند. بخش عمده آنها از روش ماشین بردار پشتیبان برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید به کمک اطلاعات 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 تراز ساختاری سرتاسری و تکنیک تشخیص فولد برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید استفاده کرده اند (۲۶، ۲۸ و ۲۹). در این مطالعات، خواص الکتروستاتیک اصلی پروتئینها مثل 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-ریبونوکلیک اسید موجود می‌باشند لیکن هدف آنها بررسی ساختاری می‌باشد نه پیشگویی عملکرد (۱۳ و ۱۸).

در مطالعه جاری نقش انفرادی خواص الکتروستاتیک پروتئینها شامل بار، ممان دوقطبی و ممان چهارقطبی در پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید بررسی خواهد شد که تا به حال انجام نشده است. سؤال اصلی تحقیق حاضر این می‌باشد که آیا تنها به کمک این پارامترهای اندک فیزیکوشیمیایی می‌توان با دقت خوبی عملکرد اتصال پروتئینها به ریبونوکلیک اسید را پیشگویی کرد؟

Sziagyi and Skolnick در سال ۲۰۰۶ پژوهشی انجام دادند که در آن از ترکیب فراوانیهای آمینواسیدهای ویژه در توالی پروتئین و ممان دوقطبی مولکول برای پیشگویی عملکرد اتصال پروتئینها به دزوکسی ریبونوکلیک اسید

در میانکنش پروتئین-دزوکسی ریبونوکلیک اسید (DNA) الکتروستاتیک نقشی اصلی را بر عهده دارد. دلیل آن باردار بودن خیلی منفی DNA می‌باشد (منفی ۲ بار به ازای هر جفت باز). از طرف دیگر، پروتئینهای متصل شونده به DNA روی سطحی که با DNA مواجه می‌شوند، اغلب بارهای مثبت زیادی دارند. نقش کلیدی بار مثبت پروتئین ایجاد یک جاذبه الکتروستاتیک برای DNA می‌باشد (۵).

در میانکنش پروتئین-دزوکسی ریبونوکلیک اسید (DNA) الکتروستاتیک نقشی اصلی را بر عهده دارد. دلیل آن باردار بودن خیلی منفی DNA می‌باشد (منفی ۲ بار به ازای هر جفت باز). از طرف دیگر، پروتئینهای متصل شونده به DNA روی سطحی که با DNA مواجه می‌شوند، اغلب بارهای مثبت زیادی دارند. نقش کلیدی بار مثبت پروتئین ایجاد یک جاذبه الکتروستاتیک برای DNA می‌باشد (۵).

**سریهای داده (بانک اطلاعاتی):** بانک اطلاعاتی پروتئینهای متصل‌شونده به ریبونوکلیک اسید (RBPs) از ۱۶۰ زنجیره پروتئینی غیر همولوگ تشکیل می‌شد (۳). این زنجیره‌های پروتئینی حداقل با یک زنجیره RNA در تماس هستند که حداقل تعداد رزیدوهای دخیل آنها در تماس با RNA، ۳ می‌باشد 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 متصل‌شونده به ریبونوکلیک اسید (Non-RBPs) نیز از ۲۴۴۱ زنجیره پروتئینی تشکیل می‌شد (۳). این زنجیره‌ها هیچ‌گونه شباهتی با RBPs نداشتند و از پایگاه PDBselect (۶) با سطح 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.

استفاده کردند. آنها از روش سریع و توانمند لوژستیک رگرسیون برای پیشگویی استفاده کردند و مشاهده کردند روش مزبور قابلیت پیشگویی عملکرد اتصال پروتئینها به دزوکسی ریبونوکلیک اسید را با دقت بالایی دارد (۲۵).

لوژستیک رگرسیون یک مدل آماری خطی تعمیم یافته می‌باشد که توانایی پیشگویی یک خروجی مجزا را از یک سری متغیرهای پیوسته، گسسته یا دوحالتی دارد. به طور کلی، متغیر پاسخ در لوژستیک رگرسیون دو حالتی می‌باشد مثل وجود/عدم وجود یا پیروزی/شکست. لوژستیک رگرسیون می‌تواند برای برآورد احتمال یک اتفاق ویژه (در اینجا، عملکرد اتصال پروتئین به ریبونوکلیک اسید) استفاده شود (۱۲). در این تحقیق با استفاده از خواص الکتروستاتیکی بار، ممانهای دوقطبی و چهار قطبی به کمک روش لوژستیک رگرسیون عملکرد اتصال پروتئینها به ریبونوکلیک اسید پیشگویی می‌گردد. 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

جدول ۱ - مشخصات پارامترهای به کار رفته در مدل

پارامتر	مشخصات	واحد
بار کلی پروتئین	برحسب بار اولیه الکترون (e)	$1e \approx 1.6 \times 10^{-19} \text{ coulombs}$
ممان دوقطبی پروتئین	بر حسب Debye (دبای)	$1\text{Debye} \approx 3.33564 \times 10^{-30} \text{ coulomb meters}$
ممان چهار قطبی پروتئین	بر حسب $e\text{\AA}^2$	$\text{Coulomb metre}^2$ $1\text{\AA}(\text{angstrom})=1.0 \times 10^{-10} \text{ meters}$

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 حالتی باردارش خنثی در نظر گرفته شد. در این محاسبات، مولکولهای آب، یونهای فلزی و لیگاندها به حساب نیامدند (۳).

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 براساس کنفورمسیون زنجیره اصلی (تعیین شده توسط موقعیت  $C_{\alpha}$  رزیدوها) محاسبه شدند. برای محاسبه بار، تمامی رزیدوهای لیزین و آرژینین دارای یک بار مثبت و

درحالی که احتمال (p) از ۰ تا ۱ تغییر می‌یابد، لوجیت p از منهای بینهایت تا مثبت بینهایت تغییر می‌یابد و لوجیت عدد نیم، صفر می‌باشد. فرم اصلی معادله لوجستیک رگرسیون به صورت زیر می‌باشد: (۴)

$$\text{Logit}(p) = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

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

شرایط شبیه‌سازی با مدل لوجستیک رگرسیون در جدول ۲ آمده است. در این تحقیق از تابع glm نرم افزار R ورژن 3.0.1 (بنیاد محاسبه آماری آر، وین، اتریش) برای انجام پیشگویی با روش لوجستیک رگرسیون استفاده شده (۲۰). روش Jackknife برای ارزیابی روش لوجستیک مورد استفاده قرار گرفت. در این روش، یک پروتئین از کل بانک جدا گشته و عملیات آموزش مدل به کمک سایر پروتئینها انجام می‌شود. سپس مدل آموزش دیده برای پیشگویی عملکرد اتصال پروتئین جدا گشته به کار می‌رود. این عملیات به تعداد کل پروتئینهای بانک تکرار می‌گردد (یعنی هر بار عملکرد اتصال یک پروتئین پیشگویی می‌شود). در نهایت مدل با توجه به اجرا بر روی تک تک پروتئینها ارزیابی می‌گردد. برای انجام روش Jackknife از تابع CVbinary کتابخانه DAAG نرم افزار R ورژن 3.0.1 استفاده شد (۱۹). واحد دقت یک صدم اعشار می‌باشد.

جدول ۲ - شرایط شبیه‌سازی با مدل لوجستیک رگرسیون

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

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

$$F = \sum (R_i - R_0) q_i \quad (1)$$

که در آن  $R_0$  نقطه مرجع می‌باشد، که به عنوان مرکز هندسی تمامی رزیدوها (موقعیتهای  $C_\alpha$ ) در ساختار در نظر گرفته می‌شود، و  $i$  نشان دهنده یک اتم در ساختار پروتئین می‌باشد. ممان دوقطبی کلی پروتئین به وسیله جمع برداری این اجزا به دست می‌آید

$$p = \sqrt{(F_x)^2 + (F_y)^2 + (F_z)^2}$$

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

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

که در آن  $r_i$  بردار موقعیت نسبی است،  $i$  شاخص بار است و عملیات جمع شامل تمامی بارها می‌باشد. ماتریس ممان چهار قطبی می‌تواند قطری شود و ۳ مقدار ویژه آن با ترتیب کاهشی می‌تواند با  $Q_1, Q_2$  و  $Q_3$  نمایش داده شوند. تمامی مقادیر ممانهای الکتریکی با در نظر گرفتن طول توالی پروتئین نرمالیزه شدند (۳).

لوجستیک رگرسیون: لوجستیک رگرسیون دوتایی رابطه بین یک متغیر پاسخ دوحالته (در اینجا، متصل شدن یا نشدن یک پروتئین به RNA) و یک سری متغیرهای پیشگو را توصیف می‌کند. از نظر ریاضی، خروجی مدل لوجستیک مقدار متغیر پاسخ نمی‌باشد بلکه احتمال به دست آوردن مقدار ۱ (همان متصل شدن به RNA) را در قیاس با مقدار ۰ (همان متصل نشدن به RNA) به دست می‌دهد. چون احتمال مابین ۰ تا ۱ می‌باشد، رگرسیون خطی نامناسب برای پیشگویی مقدار مستقیم آن می‌باشد لذا از تبدیل لوجستیک احتمال برای پیشگویی استفاده می‌شود:

$$\text{Logit}(p) = \frac{F}{1-p} \quad (3)$$

از توابع موجود در کتابخانه ROCR نرم افزار R ورژن 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  $P\text{-value} < 0.05$  داشتند). معادله لجیت حاصل پس از خطی سازی شدن به این صورت به دست آمد:

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

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 عملکرد اتصال پروتئینها به ریبونوکلیک اسید دارای وزن یا اهمیت بیشتری است. در مورد تأثیر مقادیر ویژه ممان چهار قطبی نیز می توان گفت که آنها پارامترهای غیر معنادار بوده ( $P\text{-value} > 0.05$ ) 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.

**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: اندیس تی پی (تعداد RBPs درست پیشگویی شده)، اندیس تی ان (تعداد Non-RBPs درست پیشگویی شده)، اندیس اف پی (تعداد Non-RBPs که اشتبهاً به عنوان RBPs پیشگویی شده اند) و اندیس اف ان (تعداد RBPs که اشتبهاً به عنوان Non-RBPs پیشگویی شده اند). فرمولهای محاسبه این پارامترها از اندیسهای ذکر شده بدین ترتیب می باشند:

$$P = \frac{TP}{TP+FP} \quad (5) \text{ صحت (Precision)}$$

$$Recall(r) = \frac{TP}{TP+FN} \quad (6) \text{ حساسیت}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7) \text{ دقت (Accuracy)}$$

$$F\text{-measure} = \frac{2Pr}{p+r} \quad (8) \text{ معیار F (F-measure)}$$

معیار F میانگین هندسی p و r می باشد.

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

جدول ۳- ارزیابی اجرای روشهای لوژستیک رگرسیون و شبکه عصبی در پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید در حالت Jackknife (تمامی مقادیر بر حسب درصد می باشند).

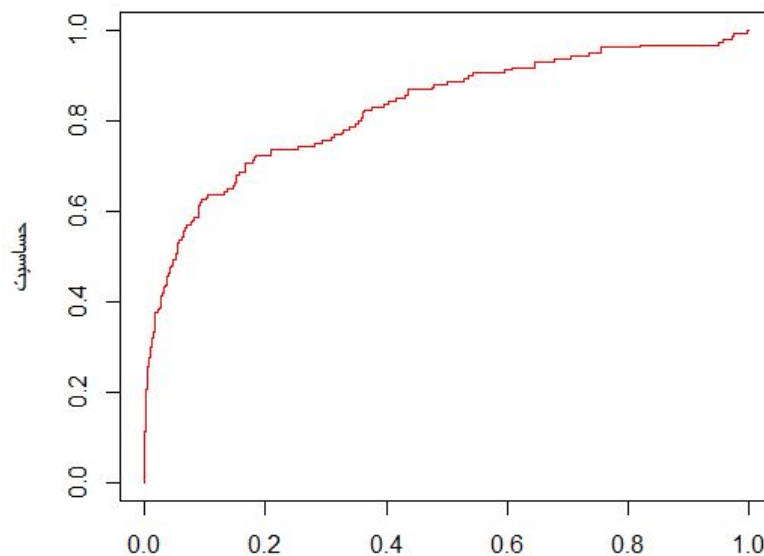
روش	دقت	صحت	حساسیت	معیار F	AUC
شبکه عصبی (مرجع شماره ۳)	91	31	45	37	78
لوژستیک رگرسیون	94	59	38	46	83

حساسیت، دقت، معیار F و مساحت زیر نمودار ROC (AUC) با توجه به بالاترین مقدار پارامتر معیار F محاسبه

پس از اجرای روش Jackknife بر روی بانک (تعداد فولدها = ۲۶۰۱)، پارامترهای ارزیابی مدل شامل صحت،

شدند. مقادیر این پارامترها در جدول ۳ آورده شده‌اند. روش ارائه شده در این پژوهش (به دست آمده در حالت برای مقایسه، نتایج حاصل از پیشگویی با روش شبکه عصبی (۳) نیز در این جدول آورده شده‌اند. منحنی ROC است.

### منحنی ROC



شکل ۱ - منحنی ROC روش لوژیستیک رگرسیون نرخ پیشگویی اشتباه

اسید در قیاس با روش شبکه عصبی می‌باشد. همچنین، مقدار معیار ارزیابی AUC روش لوژیستیک رگرسیون ۵ درصد بیشتر از مقدار معادل آن در روش شبکه عصبی می‌باشد که با توجه به غیر وابسته بودن این معیار به آستانه، دلیل محکم‌تر دیگری بر کارایی بهتر این روش نسبت به روش شبکه عصبی به دست می‌آید.

بهره‌گیری از ابزارهای آماری (مثل لوژیستیک رگرسیون) و نیز روشهای هوش مصنوعی (مثل شبکه عصبی) برای پیشگویی ساختار و عملکرد پروتئینها پیشرفتهای قابل توجهی را در علم زیست‌شناسی به وجود آورده است با علم به این واقعیت که روشهای آزمایشگاهی برای تعیین ساختار و عملکرد زمان بر و هزینه بر می‌باشند (۱). لیکن، این دو نوع ابزار یا روش با یکدیگر تفاوت دارند. از مزایای بارز روش لوژیستیک رگرسیون، سرعت بالا و

با نگاهی به جدول شماره ۳ می‌توان دریافت که کارایی روش لوژیستیک رگرسیون در قیاس با روش شبکه عصبی از نظر ۴ معیار ارزیابی دقت، صحت، معیار F و AUC بهتر است. تنها مقدار معیار ارزیابی حساسیت کمتر از مقدار به دست آمده توسط روش شبکه عصبی می‌باشد که نشان دهنده بالاتر بودن اندیس اف ان یا به عبارت دیگر پیشگویی تعداد بیشتری از RBPs در دسته Non-RBPs می‌باشد که این امر به نوبه خود به دلیل نامتعادل بودن بانک اطلاعاتی و بیشتر بودن تعداد Non-RBPs (۲۴۴۱) از RBPs (۱۶۰) قابل توجه است. متقابلاً، بیشتر بودن پارامتر صحت دلالت بر پائین تر بودن اندیس اف پی دارد که نشان دهنده پیشگویی تعداد کمتری از Non-RBPs در دسته RBPs می‌باشد و از مزایای بارز روش لوژیستیک رگرسیون به کار رفته در تحقیق حاضر برای صحیح تر پیشگویی کردن عملکرد اتصال پروتئینها به ریبونوکلیک

### نتیجه‌گیری

در این مطالعه، از روش آماری لوژستیک رگرسیون برای پیشگویی عملکرد اتصال پروتئینها به ریبونوکلیک اسید به کمک خواص فیزیکوشیمیایی بار، ممانهای دوقطبی و چهار قطبی استفاده گردید. پس از اجرای مدل لوژستیک رگرسیون بر روی بانک اطلاعاتی، مشخص گردید که پارامترهای بار و ممان دوقطبی برای پیشگویی صحیح و دقیق کافی می‌باشند. همچنین وزن و اهمیت هر یک از این پارامترها با به دست آوردن معادله لوجیت 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 کارایی خیلی بهتری در قیاس با مدل شبکه عصبی دارد. در این تحقیق نشان داده شد که روش ارائه شده در این پژوهش توانایی تفکیک پروتئینهای متصل شونده به ریبونوکلیک اسید را با دقت بالایی (۹۴ درصد) از پروتئینهای غیر متصل شونده به ریبونوکلیک اسید دارد. روش ارائه شده در تحقیق حاضر، هم سریع و هم دقیق است و می‌تواند به آسانی در مطالعات در سطح پروتئوم برای پیشگویی پروتئینهای متصل شونده به ریبونوکلیک اسید به کار رود.

سادگی اجرای آن می‌باشد در حالی که روش شبکه عصبی پیچیدگی بالایی دارد و همچنین به صورت جعبه سیاه عمل می‌کند یعنی اطلاعاتی در مورد وزن پارامترهای مختلف دخیل در مدل پیشگویی را نمی‌دهد. 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 وزنهای می‌باشند که نشانگر اهمیت هر پارامتر در تعیین احتمال اتصال پروتئینها به ریبونوکلیک اسید می‌باشند.

دیگر مزیت متدولوژی مطالعه جاری این حقیقت است که می‌توان به کمک تنها پارامترهای بار و ممان دوقطبی پروتئینهای متصل شونده به ریبونوکلیک اسید را از پروتئینهای غیر متصل شونده به ریبونوکلیک اسید با دقت خوبی تفکیک کرد. در حالی که در روش شبکه عصبی از پارامترهای بار، ممانهای دوقطبی و چهار قطبی برای پیشگویی استفاده شده است و دقت تفکیک آن روش کمتر از روش تحقیق حال حاضر می‌باشد. با توجه به نقش پروتئینهای متصل شونده به ریبونوکلیک اسید در فرآیند های مختلف زیستی، از تنظیم بیان ژن گرفته تا مقاومت به تنشهای محیطی در گیاهان (۲)، اهمیت تحقیق جاری بیش از پیش آشکار می‌گردد.

### منابع

- اسکندری، و. یخچالی، ب. مینوچهر، ز. ۱۳۸۹. پیشگویی نرم افزاری مناطق مجاز پذیرش پپتیدهای پیگانه در زیر واحد CstH پیلی CS3 اشریشیاکلی. مجله زیست‌شناسی ایران، جلد ۲۳، شماره ۱، ص ۷۳-۸۴.
- اقدسی، م. ۱۳۹۲. بررسی پروتئومیک گیاهان تراریخت شده با GR-RBP2 در مقایسه با گیاهان وحشی. مجله پژوهشهای سلولی و مولکولی (مجله زیست‌شناسی ایران)، جلد ۲۶، شماره ۲، ص ۱۵۴-۱۶۳.
- Ahmad S., Sarai A., 2011. Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. BMC Struct. Biol., 11:8.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J. Mol. Biol., 215, 403-410.
- 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 Karabancheva-Christova.*
- Berman HM, Henrick K, Nakamura H, 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol., 10 (12), 980.
- 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 *Pyrococcus 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: Shewhart 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