پیشگویی کلاسهای ساختاری پروتئینها در دو وضعیت با استفاده از مدل ترکیبی عصبی-لوجستیک

صمد جهاندیده[۱،۲]، پرویز عبدالمالکی[۱،*]، محمد مهدی موحدی[۲]
[۱]گروه بیو فیزیک، دانشکده علوم پایه، دانشگاه تربیت مدرس، تهران، ایران
[۲]گروه فیزیک پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی شیراز، شیراز، ایران
* مسئول مکاتبات- آدرس الکترونیکی: *parviz@modares.ac.ir*
(دریافت: ۸۴/۷/۲۰ ؛ پذیرش:۸۵/۵/۹)

چکیده

هدف مطالعه پیشنهادی پیشگویی کلاسهای ساختاری پروتئین ها در دو وضعیت (تمام آلفا و تمام بتا) با استفاده از روش مدل سازی ترکیبی دو مرحله‌ای با شبکه عصبی مصنوعی (**ANN**) و مدل لوجستیک رگرسیون (**LRM**) است. ساختار شبکه و فرایند آموزش طولانی، کاربرد آن را در تمام زمینه‌ها محدود کرده است. پیشگویی کلاسهای ساختاری پروتئین‌ها روی یک سری داده (**n=۱۰۴**) از پروتئین‌های غیر همولوگ تک‌دومین کروی اجرا شد. پارامترهای معنادار آماری در مدل لوجستیک رگرسیون فراوانی تکی اسیدآمینه‌های والین و گلیسین همچنین فراوانی ترکیبات دوتایی لیزین-پرولین، ایزولوسین-سرین، گلوتامین- سرین و گلوتامین- پرولین بودند. در بین ترکیبات سه تایی نیز تنها فراوانی ترکیب آسپاراژین- لوسین- آسپارتیک اسید معنادار بود. پیشگویی کلاسهای ساختاری در دو وضعیت (تمام آلفا و تمام بتا) تنها بر اساس هفت شاخص ساختاری معنی‌دار به عنوان متغیرهای ورودی از بین ۶۴۲ متغیر ساختاری ۸۸٪ کارایی نشان داد. در این مطالعه، هر دو شاخص وابسته به آستانه و غیر وابسته (**ROC**) برای ارزیابی مدل استفاده شده است.

واژه‌های کلیدی: مدل لوجستیک رگرسیون، شبکه عصبی مصنوعی، پیشگویی کلاس ساختاری پروتئین.

مقدمه

پیشگویی ساختار سه‌بعدی یک پروتئین از توالی اسیدآمینه‌اش هنوز جای تأمل دارد. ساختار سه‌بعدی یک پروتئین خواص عملکردی پروتئین را تعیین می‌کند، از طرفی خواص عملکردی پروتئین نیز به توالی اسیدآمینه پروتئین وابسته است. در واقع هر رشته پلی‌پپتید می‌تواند به شکل دمین فشرده با ساختار ویژه سه‌بعدی چین بخورد (Anfinsen 1973). به‌منظور دریافت عمل پروتئین دانشمندان سعی در پیشگویی ساختار سه بعدی پروتئین از توالی اسید آمینه دارند (Jari & Jarkko 1998). دانش ساختار پروتئین نقش اساسی در دریافت عملشان بازی می‌کند. ارائه اصولی که ارتباط توالی اسیدآمینه را به ساختار سه‌بعدی پروتئین مرتبط می‌کند یکی از اهداف اساسی علم زیست شناسی در عصر معاصر می‌باشد. کلاس ساختاری به طریقی نشان‌دهنده شرح کلی از ساختار سوم پروتئین می‌باشد. با این وجود، ضعف کنونی روشهای تئوریک رایج، کارایی این روشها را در پیشگویی دقیق ساختار سوم در کاربردهای تجربی محدود می‌کند (Chou & Zhang 1995).

بیش از ۲۰ سال پیش کلاسهای ساختاری پروتئین به منظور طبقه بندی چین‌خوردگی پروتئین‌ها که بیان‌کننده محتوای عناصر ساختار دوم و چگونگی نظم آنهاست تعریف شد (Levitt & Chothia 1976). ساختارهای شناخته‌شده پروتئین در چهار کلاس اصلی، تمام α، تمام β، $\alpha+\beta$ و α/β طبقه‌بندی شده است.

اخیراً، کریستالوگرافی پرتو X و اسپکتروسکوپی NMR چندبعدی تنها روشهای استاندارد هستند که قادر به تعیین دقیق ساختار پروتئین با قدرت تفکیک بالا می‌باشند. با این وجود، این تکنیکها ضعف‌هایی دارند که شامل ۱) نیاز مندی به کریستالهای با کیفیت بالا در کریستالوگرافی پرتو X که همیشه قابل تهیه نمی‌باشد. ۲) این احتمال وجود دارد که طبیعت پایدار کریستال پروتئین، طبیعت دینامیک پروتئین در محلول را نشان ندهد. علاوه بر این، کل فرایند تعیین ساختار یک پروتیین با تکنیکهای ذکر شده بخاطر نیاز به خالص‌سازی در اسپکتروسکوپی NMR چندبعدی و کریستاله کردن پروتئین در کریستالوگرافی اشعه X فرایندی بسیار کند می‌باشد. اسپکتروسکوپی NMR این مزیت را دارد که می‌توان از آن در تعیین ساختار پروتئین محلول استفاده کرد اما تفسیر منحنی‌های NMR در مورد پروتئینهای بزرگ خیلی پیچیده است و تکنیک محدود به پروتئینهای کوچک می‌باشد. در نتیجه، فضای خالی بزرگی بین تعداد توالی‌های پروتئینی شناخته شده و تعداد توالی‌های پروتئینی سه‌بعدی شناخته شده بوجود

Archive of SID

www.SID.ir

آمده است. این محدودیت های تکنیکی پیشرفت روشهای تئوریک را در پیشگویی ساختار بر اساس اطلاعات موجود تحریک کرده است. پیشگویی ساختار پروتئین با استفاده از روشهای آماری و روشهای هوش مصنوعی از قبیل شبکه عصبی مصنوعی در کاهش فضای خالی کمک خواهد کرد.

بعد از بررسی دقیق متون زیستی، به این نتیجه می‌رسیم که مدل آماری لوجستیک رگرسیون به عنوان تکنیک آماری مفید در پیشگویی ساختار پروتئین استفاده نشده است. اساساً، مدل لوجستیک رگرسیون به عنوان تکنیکی در پیشگویی حالتهای دو وضعیته شناخته شده است.

شبکه عصبی مصنوعی راهکاری دیگر در پیشگویی ساختار پروتئین فراهم می‌آورد، مخصوصاً در موقعیت‌هایی که متغیرهای مستقل (در این مورد پارامترهای استخراج شده از توالی) با متغیر وابسته (کلاس ساختاری پروتئین) ارتباط غیرخطی پیچیده‌ای نشان می‌دهند.

با این وجود، این عیب بر شبکه عصبی وارد است که فرایند آموزش شبکه وقت گیر و بهینه سازی ساختار شبکه وقت گیر است. علاوه بر این، شناسایی اهمیت متغیرهای ورودی شبکه آسان نیست. این دلایل کاربرد آن را در طبقه بندی‌های معمول محدود می‌کند. هدف این پژوهش بررسی کارایی پیشگویی ساختار پروتئین با روش ترکیبی مدل لوجستیک رگرسیون و شبکه عصبی مصنوعی است. در این پژوهش ابتدا مدل پیشگوی لوجستیک رگرسیون خود را ایجاد کردیم، سپس پارامترهای معنی‌دار (فراوانی اسیدهای آمینه و ترکیب اسیدهای آمینه) را به عنوان متغیرهای ورودی مدل شبکه عصبی استفاده کردیم. مدل لوجستیک رگرسیون به عنوان مدل پذیرفته شده معمول برای شناسایی متغیرهای مهم، زمانی که تعداد زیادی متغیر مستقل مورد بررسی قرار می‌گیرد قابل استفاده است. سرانجام، به عنوان فرایند مدل سازی دو مرحله‌ای می‌توانیم از متغیرهای معنی‌دار بدست آمده از مدل لوجستیک رگرسیون به‌عنوان گره‌های لایه ورودی شبکه عصبی استفاده کنیم. از این رو این مدل می تواند تعداد گره‌ها در ورودی را کاهش دهد و ساختار شبکه را ساده کند و زمان آماده سازی مدل بهینه‌شده را کوتاه کند. برای نشان دادن توانایی مدل پیشگوی ساختار پروتئین، پیشگویی ساختار پروتئین روی بانکی از پروتئین شامل ۱۰۴ پروتئین غیر مشابه تک دمین انجام شد.

در این پژوهش ما پیشگویی کلاسهای ساختاری پروتئین‌ها را در دو کلاس (تمام آلفا و تمام بتا) با استفاده از مدل ترکیبی دو مرحله‌ای شبکه‌عصبی (ANN) و مدل لوجستیک رگرسیون (LRM) انجام دادیم. مدل لوجستیک رگرسیون در پیشگویی عضویت کلاس ساختاری پروتئین با استفاده از پارامترهای استخراج شده از ساختار اول، تعیین‌گر اهمیت پارامترهای ساختاری در طبقه‌بندی دقیق کلاس ساختاری پروتئین (تمام آلفا و تمام بتا) می‌باشد. در واقع ما از مدل لوجستیک رگرسیون به عنوان پیش پردازشگر در فرایند انتخاب پارامترهای معنی‌دار (فراوانی اسیدهای آمینه و ترکیب اسیدهای آمینه) استفاده کردیم.

مواد و روشها

بانک پروتئینهای کروی غیر مشابه تک دمین

ما توالی ۱۰۴ پروتئین تک‌دمین غیرمشابه (شامل ۱۸۰۳۶ اسید آمینه) را از بانک اطلاعات پروتئین (PDB) استخراج کردیم. این پروتئین‌ها متعلق به دو کلاس ساختاری (جدول شماره ۱) بر اساس طبقه بندی بانک SCOP بودند. این بانک طبقه‌بندی کلاسهای ساختاری برای دمین‌های پروتئین را بر اساس ارتباط تکاملی و قوانینی که حاکم بر ساختار سه‌بعدی پروتئین‌هاست انجام می‌دهد. بنابر این، آستانه ما برای تعریف کلاس قابل اعتمادتر می‌باشد. ساختار تمام رشته‌ها با حداقل قدرت تفکیک ۲/۵ آنگستروم و تنها توسط تکنیک کریستالوگرافی اشعه X انجام شده بود.

جدول ۱- کد چهار حرفی پروتئینهای استخراجی از بانک PDB که در تحقیق استفاده شده است (n=۱۰۴)، این بانک شامل ۵۶ پروتئین تمام آلفا و ۴۸ پروتئین تمام بتا می‌باشد.

پروتئین های تمام آلفا:
1PSJ, 1PWT, 1FS8, 1FT5, 2YGS, 3WRP, 1XNB, 1IWK, 1JFB, 1JL7, 1ECA, 1LIS, 1ROP, 2HBG, 3C2C, 2A0B, 1CTJ, 1ALU, 2ABK, 1BGF, 1HYP, 1MKS, 1HZ4, 1HUW, 1BEA, 1B7V, 1AYX, 1AA2, 1EA8, 1A6G, 2PSR, 1XSM, 1JKO, 1AX8, 1CZJ, 1RZL, 1FS6, 1HZF, 1HXA, 1HU3, 1IQV, 1BFA, 1B5L, 2ASR, 1K40, 1GSQ, 1LE4, 1ENI, 1UTG, 3ICB, 1CC5, 2HTS, 5PAL, 1PPA, ILKI, 1CGO.
پروتئین های تمام بتا:
1A1X, 1AAJ, 1AT0, 1AX0, 1DUC, 1EUS, 1IKO, 1IW2, 1IWL, 1JAB, 1JHC, 1JPC, 1JPE, 1K12, 1KNB, 1MJC, 1JNW, 2SNS, 2SIL, 2RHE, 1A3K, 8PRN, 1BFS, 8I1B, 1BMG, 1GEN, 2STV, 1INO, 1IWN, 1JK4, 1INW, 1LIB, 2IZB, 2CPL, 2BDO, 2AYH, 1WWC, 1WHO, 1VIE, 1TUC, 1TSP, 1THW, 1TEN, 1RSY, 1PXF, 1MKH, 2MCM, 1UPJ.

ما پارامترهای خود را که شامل فراوانی تکی اسیدهای‌آمینه، فراوانی همه ترکیبات دوتایی و فراوانی چندین ترکیب سه، چهار، پنج و شش تایی است را با استفاده از برنامه‌ای در محیط MATLAB استخراج کردیم. به منظور بررسی صحت برنامه نوشته شده، نتایج با خروجی برنامه COMPSEQ به آدرس شبکه جهانی اینترنت http://bioweb.pasteur.fr/seqanal/interfaces/compseq.html روی توالی‌های مشابه بررسی شد.

Archive of SID

www.SID.ir

مدل لوجستیک رگرسیون

لوجستیک رگرسیون یا آنالیز لاجیت یک مدل آماری معمول با کارایی بالا در زمینه‌های مختلف علمی می‌باشد (Condous *et al.* 2004, Tau *et al.* 2005, Rau *et al.* 2004). این مدل احتمال خروجی‌های دو وضعیته را مورد بررسی قرار می‌دهد که خروجی به یک سری از متغیرها به شکل زیر وابسته است و در آن P احتمال خروجی، α عرض از مبدا، $\beta_1,...,\beta_n$ ضرایب اهمیت پارامترهای $x_1,...,x_n$ می‌باشد. متغیر وابسته مقدار $\{Ln(P/1-P)\}$ می‌باشد که این مقدار به طور خطی به متغیرهای مستقل وابسته می‌باشد (Hosmer & lemeshow 1989). در این رابطه $x_1,...,x_n$ فراوانی اسیدهای آمینه و ترکیبات مختلف آنهاست.

$$Ln\left(P/_{1-P}\right) = Logit\,(p) = \alpha + \beta_1 X_1 + ... + \beta_n X_n$$

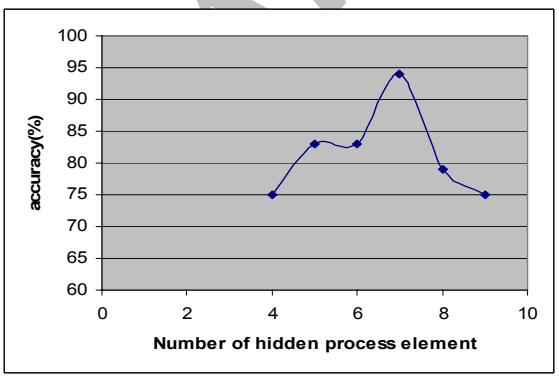
ما آموزش و آزمایش مدل خود را با استفاده از تکنیک jackknife انجام دادیم. در این روش، همه به جز یکی از نمونه‌ها برای آموزش مدل لوجستیک رگرسیون استفاده شده و تک نمونه کنار گذاشته شده به منظور آزمایش مدل استفاده می‌شود. در واقع با آموزش مدل روی تمام نمونه‌ها بجز تک نمونه کنار گذاشته شده ضرایب اهمیت پارامترها ($\beta_1,...,\beta_n$) بدست آمده که در مرحله‌ی آزمایش روی تک نمونه‌ی کنار گذاشته شده اعمال می‌شود. این فرایند تا آنجا تکرار می‌شود که همه نمونه‌های موجود در بانک یک بار به عنوان نمونه آزمایش استفاده شوند. متغیر وابسته دو وضعیته که نتیجه کریستالوگرافی اشعه X می‌باشد به شکل کلاس تمام آلفا (کلاس۱) و کلاس تمام بتا (کلاس) تعریف می‌شود. سرانجام، مدل لوجستیک رگرسیون را با روش forward در محیط نرم افزارآماری (Chicago, USA, SPSS) SPSS۱۱/۵ انجام دادیم.

شبکه عصبی مصنوعی و ساختار مدل ترکیبی

شبکه عصبی مصنوعی مدلی است که از شبکه عصبی بیولوژیک الگو گرفته شده است. شبکه عصبی مصنوعی از چند لایه تشکیل شده که لایه اول لایه ورودی است و بردار ورودی را دریافت می‌کند. هر نرون در لایه مخفی یا لایه میانی و لایه خروجی، بردار خروجی از لایه قبل را دریافت می‌کند سپس مجموع حاصل ضرب بردارهای ورودی در وزنهای خاص آنها که تحت تاثیر تابع فعالسازی قرار می‌گیرد بردار خروجی را ایجاد می‌کند (Patterson & Dan 1996). ما در این تحقیق از یک شبکه عصبی پیش خور سه لایه با الگوریتم یادگیری پس انتشار خطا استفاده کردیم(Patterson & Dan 1996) . شبکه برای شناسائی دو کلاس تمام آلفا و تمام بتا بر پایه پارامترهای استخراج شده از توالی‌های اسید آمینه‌ای طراحی شده است. بانک پروتئین ما شامل ۱۰۴ پروتئین (۵۶ تمام آلفا و ۴۸ تمام بتا) بود.

مدل ترکیبی ما برپایه خروجی مدل لوجستیک بدین شکل طراحی شد که هفت پارامتر معنی‌دار آماری از خروجی مدل لوجستیک (فراوانی هفت ترکیب اسیدآمینه‌ای ذکر شده در قانون ممیزی) به عنوان گره‌های ورودی به شبکه عصبی استفاده شد. لایه دوم یا لایه مخفی، هفت گره داشت و لایه خروجی یک گره که خروجی برابر با یک نشان دهنده کلاس تمام آلفا و خروجی برابر با صفر نشان دهنده کلاس تمام بتا می‌باشد.

جهت تعیین تعداد نرونهای لایه میانی بین تعداد نرونهای ۴ الی ۹، مناسبترین مقدار را با آموزش شبکه روی ۸۰ نمونه مشابه و آزمایش شبکه با استفاده از مابقی ۲۴ نمونه و ثابت نگه داشتن پارامترهای دیگر جستجو کردیم. با استفاده از این روش و دقت بدست آمده از پیشگویی نهایی روی پروتئینها در هر مرحله همانطور که در شکل (۱) مشاهده می‌گردد تعداد نرونهای لایه میانی ۷ عدد انتخاب شد. تابع سیگموئید به عنوان تابع فعالیت شبکه انتخاب شد. الگوریتم یادگیری شبکه پس انتشار خطا بود. این الگوریتم بدین شکل عمل می‌کند که فرایند کاهش خطا را در دو مرحله پیشرو و پسرو انجام می‌دهد. در مرحله پیشرو مقادیر وزنهای پارامترها به طور اتفاقی از فضایی که خود تعیین کرده‌ایم انتخاب شده و در کل مسیر پیشرو ثابت باقی می‌ماند. در نهایت اختلاف خروجی شبکه با هدف محاسبه شده و در مرحله پسرو، وزنها از آخرین لایه به سمت لایه‌های داخلی تغییر می‌یابند تا به لایه ورودی برسیم و دوباره مرحله پیشرو انجام می‌شود. این فرایند بارها و بارها انجام می شود تا به سطح خطای مورد انتظار برسیم. در این حالت وزنهای حاصل شده روی نمونه‌های آزمایش اعمال شده و دقت مدل حاصل می شود. شبکه ما با ۱۵۰۰۰۰۰ تکرار آموزش دید، مشخصات ساختار شبکه بهینه در جدول (۲) آورده شده است. نرم افزار استفاده شده در رابطه با ساختار شبکه در محیط برنامه نویسی MATLAB (Mathworks, Natick, MA, USA) نوشته شده بود.



شکل ۱- دقت شبکه عصبی با توجه به تغییر تعداد گرههای لایه مخفی در بهینه سازی شبکه.

Archive of SID

www.SID.ir

جدول ۲- مشخصات شبکه عصبی بهینه ی بکار رفته در این پژوهش.

مشخصـــات شبکـــه عصبی	
تعداد نرون‌های لایه ورودی	۷ عدد
تعداد لایه‌های پنهان	۱عدد
تعداد نرون‌های لایه پنهان	۷ عدد
تعداد نرون‌های لایه خروجی	۱ عدد
نوع تابع فعالیت شبکه	سیگموئید
نوع شبکه	پس انتشار خطا
مجموع مربع خطا	۰/۰۲
سرعت آموزش	۰/۲
تعداد تکرار	۱۵۰۰۰۰۰

مقیاسهای کارایی مدل

ما در این تحقیق از دو دسته مقیاسهای وابسته به آستانه و مستقل از آستانه به منظور ارزیابی کارایی مدل لوجستیک رگرسیون و مدل ترکیبی استفاده کردیم.

مقیاسهای وابسته به آستانه

ما از شش مقیاس متفاوت برای ارزیابی کارایی مدل استفاده کردیم. این شش مقیاس از چهار شاخص زیر استخراج شده اند.

۱- True positives (TP): تعداد پروتئین‌هایی از کلاس تمام آلفا که صحیح طبقه بندی شده اند.

۲- True negatives (TN): تعداد پروتئین‌هایی از کلاس تمام بتا که صحیح طبقه بندی شده اند.

۳- False positives (FP): تعداد پروتئین‌هایی از کلاس تمام بتا که اشتباهاً در طبقه تمام آلفا قرار گرفته اند.

۴-False negative (FN): تعداد پروتئین‌هایی از کلاس تمام آلفا که اشتباهاً در طبقه تمام بتا قرار گرفته اند.

با استفاده از فرمولهایی که قبلا در گزارشات علمی چاپ شده است دقت پیشگویی، حساسیت، ویژگی و فاکتور همبستگی متیو برای خروجی مدل چنین محاسبه می‌شود (Kaur & Raghava 2004).

۱-دقت پیشگویی: درصدی از پروتئین‌ها که طبقه ساختاری آنها بدرستی پیشگویی شده است.

۲-حساسیت: درصدی از پروتئین‌های کلاس تمام آلفا که بطور صحیح طبقه بندی شده اند.

۳-ویژگی: درصدی از پروتئین‌های کلاس تمام بتا که بطور صحیح طبقه بندی شده اند.

۴-فاکتور همبستگی متیو (MCF): ما از MCF به عنوان یک مقیاس کاراتر (در مقایسه با دیگر مقیاسها) به منظور ارزیابی قابلیت اعتماد پذیری مدل استفاده کردیم. این مقیاس با فرمول زیر بیان می‌شود.

۵-

$$MCF = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCF مقداری بین ۱و۱- می‌باشد. اگر ارتباطی بین مقادیر پیشگویی شده و مقادیر واقعی وجود نداشته باشد باید صفر یا مقدار خیلی کمی باشد. در صورتیکه ارتباط بین مقادیر پیشگویی شده و مقادیر واقعی افزایش یابد مقدار MCF نیز زیاد خواهد شد. واضح است که در پیشگویی کاملاً صحیح مقدار MCF برابر با یک خواهد بود. در واقع این مقیاس میزان اعتماد پذیری مدل را نشان می‌دهد.

مقیاس مستقل از آستانه

مقیاس هایی که تا اینجا معرفی شدند وابسته به مقدار آستانه هستند و شدیداً وابسته به بانک گردآوری شده در تحقیق هستند. مهمترین مشکل این مقیاس ها هم این است که همه وابسته به آستانه‌ی داده شده هستند. ROC مقیاسی مستقل از آستانه است که در آن بازدهی با استفاده از دو شاخص کسر مثبت درست (TP) و نسبت مثبت نادرست (FP) گزارش می‌شود، بدین ترتیب که خروجی مدل روی نمونه‌های آزمایش آنالیز شده و نهایتاً برای رسم منحنی استفاده می‌شود. از سطح زیر منحنی برای ارزیابی بازدهی مدل استفاده می‌شود.

منحنی همیشه از دو نقطه (۰ ,۰) و (۱ ,۱) می گذرد. نقطه اول (۰ ,۰) جایی است که هیچ نمونه‌ای صحیح طبقه بندی نمی‌شود و نقطه‌ی دوم (۱ ,۱) که همه نمونه‌ها صحیح طبقه‌بندی می‌شوند. مدلی که اتفاقی نمونه‌ها را طبقه‌بندی می‌کند مقدارسطح زیر منحنی ROC آن برابر با ۰/۵ خواهد بود. این مقدار برابر سطح زیر خط مستقیمی است که دو نقطه (۰ ,۰) و (۱ ,۱) را به هم وصل می‌کند. سطح زیر منحنی برای یک مدل ایده آل باید برابر با مقدار یک باشد که به سادگی کارایی کامل مدل را نشان می‌دهد (Deleo 1993).

ما از نسخه ۱۱/۵ نرم افزار SPSS برای ایجاد مدل و رسم منحنی استفاده کردیم.

نتایج

نتایج آنالیز لوجستیک رگرسیون

مدل لوجستیک رگرسیون با استفاده از روش forward stepwise اجرا گردید. در این روش متغیرها (n=۶۴۲) یکی پس از دیگری قبل از

Archive of SID

www.SID.ir

پایان آموزش به مدل وارد می‌شوند. قانون ممیزی برای پارامترهای معنادار چنین بدست آمد.

$$Logit(p) = 38 - 24\times(N-L-D) - 1190\times(S-Q) - 1377\times (Q-P) - 757\times(K-P) + 369\times(I-S) - 2.7(V) - 1.5\times(G)$$

در این تساوی G وV به ترتیب فراوانی والین و گلیسین، همچنین K-P، Q-P، I-P و S-Q نیز بترتیب فراوانی ترکیبهای دوتایی لیزین - پرولین، گلوتامین- پرولین، ایزولوسین- سرین و سرین- گلوتامین است. علاوه بر این N-L-D فراوانی ترکیب سه تایی آسپاراژین- لوسین- آسپارتیک اسید می‌باشد.

یکی از مهمترین مزایای این مدل روشن کردن اهمیت پارامترهای ساختاری در طبقه‌بندی دقیق طبقه ساختاری پروتئین هاست. بدین صورت که پارامترهایی که قدر مطلق ضریبشان بیشتر است اهمیت بیشتری در طبقه‌بندی ساختار پروتئین‌ها دارند. بین پارامترهای ارزیابی‌شده فراوانی ترکیب دوتایی Q-P با بیشترین ضریب (۱۳۷۷ -=β) و بعد از آن فراوانی ترکیب دوتایی S-Q و فراوانی ترکیب دوتایی K-P به ترتیب دارای اهمیت بودند. همچنین فراوانی اسید آمینه گلیسین به عنوان ضعیف ترین پارامتر معنادار قابل توجه بود. علاوه بر این از نتایج می توان چنین نتیجه گرفت که توالی‌های طولانی تر اهمیت بیشتری در پیشگویی ساختار پروتئین دارند.

خروجی مدل به ترتیب دقت، حساسیت و ویژگی برابر با ۹۴٪، ۹۵٪ و ۹۴٪ نشان می دهد، همچنین MCF برابر با ۰/۳۱ حاصل شد. این نتایج وابسته به آستانه هستند به این معنا که با تغییر آستانه مقادیر این شاخص ها هم برای یک روش ثابت تغییر می‌کنند.

شاخص ارزیابی بهتر برای مدل استفاده از شاخص یگانه مستقل از آستانه ROC می باشد. سطح زیر منحنی ROC برای این مدل ۰/۹۴۲ بود. مقدار حاصل نتایج حاصل از شاخص‌های وابسته به آستانه را تایید می‌کند.

نتایج مدل ترکیبی

در واقع ما با استفاده از توان مدل لوجستیک رگرسیون در شناسایی پارامترهای معنادار آماری مدلی ترکیبی از دو تکنیک لوجستیک رگرسیون و شبکه‌عصبی ارائه دادیم. در این راستا، بجای تمام پارامترهای استخراج شده از پروتئین‌ها (n=۶۴۲) تنها هفت پارامتر معنادار از خروجی لوجستیک رگرسیون که در قانون ممیزی آورده شد به عنوان ورودی شبکه عصبی مورد استفاده قرار گرفت. همانطور که قبلا گفته شد تعداد ۸۰ پروتئین به عنوان سری آموزش انتخاب شد و شبکه روی ۲۴ پروتئین باقی مانده آزمایش شد. در پایان با بررسی پاسخ‌های خروجی شبکه عصبی مصنوعی، مشخص گردید که مدل ایجاد شده قادر است از میان ۱۶ نمونه موجود ۱۵ نمونه را به درستی پروتئین‌های تمام آلفا گزارش نماید که نشان دهنده حساسیت تشخیص معادل ۹۴٪ می‌باشد. همچنین از میان ۸ نمونه از کلاس تمام بتا شبکه قادر به تشخیص ۶ مورد از نمونه‌های موجود در بانک اطلاعاتی گردید که بیانگر ویژگی تشخیص معادل ۷۵٪ می‌باشد. در مجموع شبکه ایجادشده با تشخیص صحیح ۲۱ نمونه از موارد مورد مطالعه از میان ۲۴ پروتئین موجود توانست به دقت تشخیصی نسبتاً قابل ملاحظه ۸۷/۵٪ برسد که اگر چه در مقایسه با مدل لوجستیک رگرسیون (۹۴٪) دقت کمتری نشان می‌دهد ولی این در حالی است که تنها هفت پارامتر از بین ۶۴۲ پارامتر اولیه به شبکه عصبی به عنوان ورودی ارائه شده است. پارامترهای آماری مربوط به مدل لوجستیک رگرسیون و مدل ترکیبی در جدول ۳ آورده شده است. بیشترین تفاوت بین پارامتر آماری MCF بین این دو مدل دیده می‌شود و مقدار آن در رابطه با مدل ترکیبی برابر با ۰/۷۱ می‌باشد که نشان دهنده اعتماد پذیری بالاتر خروجی مدل ترکیبی در مقایسه با مدل لوجستیک رگرسیون است. شکل ۲ مقایسه‌ای بین تعدادی از پارامترهای آماری مربوط به مدل لوجستیک رگرسیون و مدل ترکیبی را نشان می‌دهد.

جدول ۳- مقادیر محاسبه شده ی شاخصهای کارایی مربوط به دو مدل لوجستیک رگرسیون و ترکیبی.

شاخص کارایی	مدل لوجستیک رگرسیون	مدل ترکیبی
دقت	٪۹۴	٪۸۷/۵
حساسیت	٪۹۵	٪۹۴
ویژگی	٪۹۴	٪۷۵
فاکتور همبستگی متیو	۳۱/۰	۷۱/۰
ROC	۹۴/۰	۸۴/۰



شکل ۲- نمودار ستونی، مقایسه‌ای بین تعدادی از شاخصهای کارایی مربوط به مدل لوجستیک رگرسیون (ستون روشن) و مدل ترکیبی (ستون تیره) را نشان می دهد.

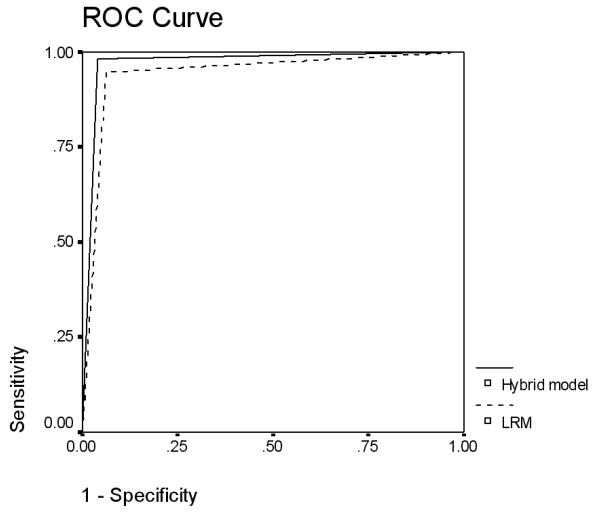
به منظور مقایسه نتایج بدست آمده با کارهای گذشته به مقاله‌ی Munson و همکارانش که با مدلی مشابه مدل آماری ما انجام شده می‌پردازیم (Munson *et al.* 1994). در مقاله با استفاده ار ۴۰۰ پارامتر ساختاری پیشگویی ساختاری را انجام داده‌اند که دقت مدل در بهترین حالت ۶۵/۹ درصد گزارش شده که اگر چه ما در این تحقیق از تعداد

Archive of SID

www.SID.ir

پارامتر بیشتری استفاده کردیم (N=۶۴۲) ولی دقت مدل پیشنهادی به میزان قابل توجهی افزایش نشان داد (۹۴٪). علاوه بر این، مدل ترکیبی پیشنهادی ما در این تحقیق نیز تنها با استفاده از ۷ پارامتر معنادار بدست آمده از مدل لوجستیک رگرسیون موفق به پیشگویی کلاسهای ساختاری با دقت ۸۸٪ شد، البته لازم به تذکر است که ما پیشگویی کلاسهای ساختاری را بخاطر ماهیت ذاتی مدل لوجستیک رگرسیون را تنها در دو کلاس ساختاری انجام دادیم.

علاوه بر این منحنی ROC برای هر دو مدل با استفاده از اطلاعات موجود ترسیم گردید. مقدار سطح زیر منحنی بعنوان شاخص مناسب‌تر در ارزیابی کارائی برای مدل ترکیبی، معادل ۰/۸۴ استخراج گردید که این مقدار در مقایسه با سطح زیر منحنی مربوط به مدل لوجستیک رگرسیون مقدار قابل ملاحظه‌ای می‌باشد (شکل۳).



شکل ۳- منحنی ROC مربوط به مدل ترکیبی (نقطه چین) و لوجستیک رگرسیون (خط پیوسته)، سطح زیر این منحنی شاخص مناسبتری از کارایی مدل در مقایسه با بقیه شاخص هاست.

بحث

نقش حیاتی پروتئین‌ها نیاز به توضیح ندارد. علی‌رغم یافته‌های زیاد در تحقیقات بر روی پروتئین ها تعیین ساختار سه بعدی پروتئین هنوز کار مشکلی است. در واقع تکنیکهای تجربی موجود مثل NMR و X-ray کریستالوگرافی وقت‌گیر و گران هستند. در نتیجه شکاف بزرگی بین پروتئین‌های تعیین توالی شده و پروتئین‌های تعیین ساختار شده ایجاد شده است. پیشگویی‌های محاسباتی ساختارها از توالی اسیدهای آمینه نقش کلیدی در کاهش آن شکاف بازی می‌کنند. گزارشات گذشته نشان داده است که این روشهای محاسباتی اطلاعات مفیدی را برای مجامع تحقیقاتی زیست شناسی فراهم می‌آورد.

در این تلاش، از مدل لوجستیک رگرسیون و یک مدل ترکیبی به منظور طبقه‌بندی پروتئین‌ها با استفاده از اطلاعات ساختار اول استفاده شد. فرض اولیه بر این بود که پارامترهای استخراج شده از توالی رشته‌های پروتئین و آنالیز آنها با مدل لوجستیک رگرسیون می‌تواند در وهله نخست پارامترهای مهم و اهمیت پارامترها را تعیین کند و سپس شبکه عصبی با استفاده از خروجی مدل لوجستیک رگرسیون قادر به پیشگویی کلاس ساختاری پروتئین‌ها با دقت مناسب باشد.

نتایج بدست آمده قابل مقایسه با نتایج روشهایی چون تئوری اطلاعات (Garnier *et al.* 1978)، نزدیکترین همسایگی (Lander Yi & 1993)، شبکه عصبی (Qian & Sejnowski 1988 Kneller *et al.* 1990) و... می‌باشد. همه روشها پروتئین‌های طبقه تمام آلفا را بهتر از طبقه تمام بتا پیشگویی می‌کنند. دلیل منطقی آن نیز نقش برهمکنشهای با دامنه کوتاه و متوسط در پروتئین‌های طبقه تمام آلفاست. به همین شکل دقت کمتر در پیشگویی طبقه‌های ساختاری دیگر بدلیل نقش غالب میانکنشهای با دامنه‌ی بلند است.

در انتخاب پارامترها، اسیدآمینه‌های مختلف تمایل قوی به ساختارهای دوم ویژه دارند. برای مثال آلانین و گلوتامیک اسید غالباً در مارپیچها در حالی که والین، سیستئین و ایزولوسین غالباً در صفحات چین‌دار بتا یافت می‌شوند. دلیل اصلی برای این تمایل به ساختار خاص اختلاف در زنجیره های جانبی اسیدهای‌آمینه است که هرکدام به نوعی حاوی اطلاعات ساختاری ویژه هستند. حال اگر تعدادی از اسیدهای‌آمینه که تمایل به تشکیل یک ساختار ویژه دارند بصورت متوالی در یک زنجیره قرار گیرند آن بخش از زنجیره شانس بالایی برای ایجاد آن ساختار خاص را خواهد داشت.

یکی از مهمترین مزایای مدل لوجستیک رگرسیون در مقایسه با شبکه‌عصبی روشن کردن اهمیت پارامترهای ساختاری در طبقه‌بندی دقیق طبقه ساختاری پروتئین‌هاست در صورتی که شبکه‌عصبی مصنوعی مانند جعبه‌ای سیاه عمل می‌کند و اطلاعاتی در رابطه با اهمیت پارامترها نمی‌دهد.

مدلهای آماری بر پایه قوانین محکم و منطقی ریاضی بنا نهاده شده اند. بنابراین با توجه به این که داده‌های حاصل از پارامترهای ساختاری ارتباط نزدیکی با کلاس ساختاری پروتئین دارند و اینکه مدل لوجستیک به طور اختصاصی برای حالتهای دو وضعیتی طراحی شده توانایی بسیار زیاد در پیشگویی ها غیر منتظره نبود.

در نتیجه ما مدل ترکیبی خود که قادر به پیشگویی طبقه ساختاری پروتئین با استفاده از پارامترهای استخراج شده از ساختار اول است را ایجاد کردیم که این مدل اهمیت پارامترهای ورودی را نیز تعیین می‌کند و به ما کمک می‌کند که شبکه عصبی را با ساختمانی ساده تر طراحی کنیم که کار کردن با آن ساده تر است و علاوه بر این بهینه سازی شبکه و فرایند آموزش آن راحت تر صورت می‌گیرد.

Archive of SID

www.SID.ir

منابع:

Anfinsen C.B. 1973: Principles that govern the folding of protein chains. *Science* **181**: 223–230.

Chou K.C., Zhang C.T. 1995: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**: 275–349.

Condous G., Okaro E., Khalid A., Timmerman D., Lu C., Zhou Y., Van Huffel S., Bourne T. 2004: The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Hum. Reprod.* **19**: 1900-10.

Deleo J.M. 1993: Receiver operating characteristic laboratory (ROCLAB): Software fordeveloping decision strategies that account for uncertainty, In proceeding of the second international symposium on Uncertainty Modelling and Analysis, MD.

Garnier, J., Osguthorpe D., Robson B. 1978: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J.Mol.Biol.* **120**: 97-120.

Hosmer D.W., Lemeshow S. 1989: Applied Logistic Regression. Wiley, New York.

Jari K., Jarkko V. 1998: Unreliability of the Chou-Fasman parameters in predicting protein secondary structure. *Protein. Eng.* **5**: 345-348.

Kaur H., Raghava G.P.S. 2004: Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins. *FEBS Letters* **564**: 47-57.

Kneller D.G., Cohen F.E. Langridge R. 1990: Improvements in secondary structure prediction by enhanced neural network. *J. Mol. Biol.* **214**: 171–182.

Levitt M., Chothia C. 1976: Structural patterns in globular proteins. *Nature* **261**: 552–557.

Munson P.J., Francesco V.D., Porrelli R. 1994: protein secondary structure prediction using periodic quadratic-logistic. *System Sciences* **5**: 375-384.

Rau R.H., Li Y.C., Cheng J.K., Chen C.C., Ko Y.P., Huang C.J. 2004: Predicting blood pressure change caused by rapid injection of propofol during anesthesia induction with a logistic regression model. *Acta Anaesthesiol Taiwan* **42**: 81-6.

Qian N., Sejnowski T.J. 1988: Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**: 865-884.

Patterson A., Dan W. 1996: Artificial Neural Networks, Singapore: Prentice Hall.

Tau Q., Christiansen L., Bathum L., Zhao J.H., Vach W., Vaupel J.W., Christensen K., Kruse T.A. 2005: Haplotype effects on human survival: logistic regression models applied to unphased genotype data. *Ann. Hum. Genet.* **69**: 168-75.

Yi T.M., Lander E.S. 1993: Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* **232**: 1117–1129.

Archive of SID

www.SID.ir