
بسم الله

دانشگاه تهران

دانشکده ریاضی، آمار و علوم کامپیوتر

مدل های مارکفی پنهان در پیش بینی ساختار دوم پروتئین ها

نگارش: سید امیر ملک پور نرگسی

استاد راهنما: دکتر حمید پزشک

استاد مشاور: دکتر مهدی صادقی

پایان نامه برای دریافت درجه کارشناسی ارشد
در
آمار ریاضی

شهریور ۱۳۸۷

تقدیم به

مادر و پدر مهربانم

مدل های مارکفی پنهان در پیش بینی
ساختار دوم پروتئین ها

چکیده

پروتئین ها از توالی ۲۰ اسید آمینه ساخته شده اند و برای هر یک از اسید های آمینه یکی از سه ساختار $Helix$ ، $Strand$ یا $Coil$ در نظر گرفته می شود که ساختار دوم پروتئین ها نامیده می‌شوند. شناسایی ساختار دوم پروتئین ها یکی از مراحل کلیدی در تعیین ساختمان سه بعدی و عملکرد پروتئین‌ها به شمار می رود. یافتن ساختار دوم پروتئین در درمان بیماری ها و ساخت دارو کمک بسیاری می کند. اگرچه تعیین توالی اسید های آمینه آسان است، تعیین ساختار دوم گران و وقت گیر است. به همین دلیل از روش های محاسباتی برای پیش بینی این ساختاراستفاده می شود.

در این تحقیق کاربرد مدل های مارکفی پنهان در پیش بینی ساختار دوم مورد مطالعه قرار می‌گیرد و حالت خاصی از این مدل ها که مدل نیمه مارکف قطاعی $(SSMM)$ نامیده می شود مرور می‌شود. همچنین امکان استفاده از این مدل ها در پیش بینی $Contact\ \ Map$ ها که یکی از مزیت های مدل‌های احتمالی است و در تعیین ساختار سه بعدی پروتئین ها کمک بسیاری می کند مورد مطالعه قرار می گیرد. این کار معمولا به کمک روش های $MCMC$ انجام می شود. از قابلیت های دیگر مدل‌های مارکفی پنهان توانایی این مدل ها برای در نظر گرفتن $Helical\ \ Capping\ \ Signal$ ها است. همچنین امکان استفاده از این مدل ها را برای در نظر گرفتن اطلاعات تکاملی موجود در توالی های $Homolog$ مرور می کنیم و از آزمون های آماری χ^2 برای یافتن همبستگی های معنی دار بین

اسیدهای آمینه استفاده خواهد شد.

مدل های $SSMM$ تنها از اطلاعات و وابستگی های موجود در سمت چپ هر اسید آمینه استفاده می‌کنند. در اینجا روشی برای در نظر گرفتن وابستگی های چپ به راست و راست به چپ در درون توالی پروتئینی معرفی می شود. همچنین از شبکه های عصبی به منظور کاهش حجم داده‌ها و به کار گیری اطلاعات موجود در توالی های $Homolog$ پروتئینی استفاده خواهد شد، که در آن از مدل های $SSMM$ برای انجام پیش بینی نهایی استفاده می شود.

واژه‌های کلیدی : مدل های مارکفی پنهان، مدل های نیمه مارکف پنهان، مدل های ممیزی، نمونه‌گیری به کمک زنجیر مارکف مونت کارلو $(MCMC)$ ، آزمون های χ^2، روش های بیزی، ساختار دوم پروتئین ها.

قدردانی

با سپاس و ستایش از معبود یگانه که پرتو الطاف بی‌شمارش بر لحظه لحظهٔ زندگی‌ام ساطع و آشکار است. تحقیق حاصل نتیجه همکاری و راهنمایی اساتید و دوستان بزرگواری است که از محبت آنها نهایت سپاس را دارم. پیش از هر چیز جای دارد قدردانی عمیق خود را نسبت به استادم دکتر حمید پزشک برای دانش، بینش، حمایت و راهنمایی های ایشان در طول تحصیلاتم اظهار دارم و به خاطر بینشی که از علم آمار به من داده‌اند از ایشان تشکر نمایم. دکتر پزشک در مراحل مختلف تحقیق صمیمانه همراهی ام کردند و در بهبود کار از هیچ تلاشی دریغ نورزیدند. بدون هدایت ایشان قادر به پیشرفت و انجام این کار نبودم.

از دکترمهدی صادقی و دکتر چنگیز اصلاحچی که از راهنمایی های ارزشمند ایشان استفاده کردم و با نگاه ظریفشان زوایای مختلف تحقیق را بر من گشودند تشکر ویژه ای دارم. همچنین از دکتر عمید رسولیان که داوری پایان نامه‌ام را به عهده گرفتند سپاسگذارم.

در پایان از آقای امیر لکی زاده و خانم سیما نقی زاده کمال تشکر را دارم.

سید امیر ملک پور

شهریور ۱۳۸۷

فهرست مندرجات

فصل ۱

تاریخچه و دورنما

۱.۱ بیوانفورماتیک ساختاری در عصر پس از ژنوم

نقشه ژنوم انسان در سال ۲۰۰۲ کامل شده است، علاوه براین توالی ژنومی بسیاری از موجودات دیگر یا کد گشایی شده است و یا در حال انجام است. در دست داشتن این داده ها خبر از انقلابی در پزشکی و زیست شناسی می دهد و اساس یک بینش به سوی شناخت مکانیسم بیماری ها و گسترش روشها و دارو های جدید برای مهار آنها است و لازمه این کار تلاش بسیاری است که باید انجام شود. زیرا توالی ژنومی موجودی مانندانسان دارای ۳ میلیارد جز است که به طور تخمینی شامل ۳۵۰۰۰ ژن می‌شود و جایگاه عمل ژنها پروتئین ها به حساب می آیند. فرآیند کشف ساختار و عملکرد این پروتئین‌ها و پیشرفت های بعدی در کشف داروهای جدید وابستگی زیادی به شناسایی ساختار سه بعدی و فضایی پروتئین ها دارد. متاسفانه هرچند می توان توالی های پروتئینی را به سادگی با استفاده از توالی ژنومی تعیین کرد ولی کشف ساختار سه بعدی آنها به یکی از سخت ترین مسائلی مبدل شده است که علم تا کنون با آن رو به رو بوده است. با توجه به اینکه روش های آزمایشگاهی تعیین ساختار

بسیار گران و وقت گیر هستند گسترش روش های محاسباتی برای تعیین و پیش بینی این ساختارها بسیار ضروری به نظر می رسد. شاخه ای که به بررسی این مسائل به کمک روش های محاسباتی می پردازد بیو انفورماتیک ساختاری نام دارد. و نه تنها ساختار پروتئین ها بلکه گستره وسیعی از مسائل دیگر را نیز در بر می گیرد. با توجه به پایان یافتن پروژه ژنوم انسان، تلاش های بسیاری برای گسترش و ساخت یک مجموعه بزرگ از ساختار های دومی که با دقت بالا تعیین شده‌اند و انواع متفاوت ساختار های مشاهده شده در طبیعت را شامل می شوند افزایش یافته است. در راستای این هدف کشف روشهای محاسباتی و آماری جدید نیز سهم زیادی در پیشبرد آن خواهد داشت.

۲.۱ تاریخچه پیش بینی ساختار دوم پروتئین ها

پیش بینی ساختار دوم پروتئین ها یکی از مراحل کلیدی در پیش بینی ساختار فضایی و عملکرد پروتئین ها به شمار می رود. در مجموع ۲۰ نوع اسید آمینه وجود دارد و در ساختار دوم پروتئین هابه هر اسید آمینه یکی از سه ساختار $Helix$ ، $\beta - strand$ و $Coil$ نظیر می شود. که این ساختار ها نشان دهنده نظم فضایی محلی اسید های آمینه نسبت به یکدیگر هستند. مسئله پیش بینی ساختار دوم بدون هیچ اطلاعی از ساختار فضایی انجام می شود و تصویر کردن یک فضای ۲۰^L بعدی ، که درآن L طول توالی پروتئینی است ، از اسیدهای آمینه به فضای ۳^L بعدی از ساختار های دوم است. از آنجایی که اطلاع در مورد ساختار دوم می تواند در الگوریتم هایی که به مسئله ساختار فضایی می‌پردازند به کار رود پیش بینی ساختار دوم بسیار مورد علاقه است. این مسئله در طول چند دهه توجه زیادی را به خود جلب کرده است اما هنوز به عنوان مسئله ای مشکل باقی مانده است. ما در اینجا چند نگرش کلاسیک مهم و روشهای مدرن را به طور خلاصه ذکر می کنیم.

۱.۲.۱ روش های کلاسیک

به طور ضروری همه روش های پیش بینی ساختار دوم بر اساس این قاعده هستند که اسید های آمینه مختلف دارای فراوانی های مختلفی در هر یک از ساختار های دوم هستند. این اختلاف فراوانی ها می‌تواند به کمک روشهای آماری مدل شود و به صورت توزیع شرطی اسید های آمینه روی ساختارهای دوم بیان شود. در میان قدیمی ترین روش های انتشار یافته برای استفاده از اختلاف فراوانی ها می توان به روش چو—فسمن [۱] ، [۲],[۱] اشاره کرد. چو و فسمن برای هر اسید آمینه یک پارامتر بر اساس فراوانی مشاهده شده در هر نوع از ساختار های دوم در نظر گرفتند که بر اساس انتقال‌های بین $helix - strand$ است اما به طور ساده به صورت زیر تعریف می شوند:

$$P_s(a) = \frac{n_{a,s}/n_{.,s}}{n_{a,.}/n_{.,.}} \qquad (۱.۲.۱)$$

که در آن a نشان دهنده اسید آمینه است و $s \in \{helix, \beta - strand, coil\}$ نشان دهنده کلاس ساختارهای دوم می باشد. $n_{x,y}$ نشان دهنده تعداد تجربی اسید آمینه x است که در ساختار از نوع y مشاهده می شود و $n_{x,.} = \sum_y n_{x,y}$ است. رابطه بالا به عنوان برآوردی از $P(s|a)$ یا $\frac{P(a|s)}{P(a)}$ با فرض یک توزیع یکنواخت برای s در نظر گرفته می شود. حاصلضرب این پارامتر ها برای یک توالی از اسیدهای آمینه همسایه به عنوان یک معیار ساده برای احتمال زیر توالی با فرض استقلال مشاهدات در نظر گرفته می‌شود. بر اساس این پارامتر ها ۲۰ اسید آمینه به کلاس های مختلفی دسته بندی شدند که در جدول ۱.۲.۱ مشاهده می شود. شایان توجه است که این طبقه بندی با طبقه بندی مدرن که بر اساس ویژگی های فیزیکی اسیدهای آمینه صورت می گیرد کاملا تطابق دارد. بر اساس طبقه بندی بالا و بر اساس تعداد اسید های آمینه هر کلاس در زیر توالی های چهار تایی و پنج تایی نمره هایی به

---

[۱] Chou Fasman

این زیر توالی‌ها در یک پروتئین جدید تخصیص داده شده است. ناحیه های با نمره های بالا به عنوان ناحیه های با پتانسیل $helix$ یا$\beta - strand$ در نظر گرفته شدند. سپس یک سری از قواعد را برای پیش‌بینی ساختار تعریف کردند. این قواعد به عنوان اولین الگوریتم برای پیش بینی ساختار دوم پروتئین ها مطرح هستند.

جدول ۱.۲.۱: تخصیص اسید های آمینه‌به کلاس های ساختار دوم که به وسیله چو و فسمن مورد استفاده قرار گرفته است.

اسید آمینه	کلاس
EAL	تشکیل دهنده $helix$ قوی
$HMQWVF$	تشکیل دهنده $helix$
KI	تشکیل دهنده $helix$ ضعیف
$DTSRC$	بی تفاوت—$helix$
NY	شکن—$helix$
PG	شکن قوی—$helix$
MVI	تشکیل دهنده $\beta - strand$ قوی
$CYFQLTW$	تشکیل دهنده $\beta - strand$
A	تشکیل دهنده $\beta - strand$ ضعیف
RGD	بی تفاوت—$\beta - strand$
$KSHNP$	شکن—$\beta - strand$
E	شکن قوی—$\beta - strand$

دومین روشی که برای پیش بینی ساختار بر اساس فروانی اسید های آمینه پیشنهاد شده است روش $Garnier - Osguthorpe - Robson$ است که دارای نگرش تئوری اطلاعات است. و محتوای اطلاعات نهفته در هراسید آمینه a را در ساختار دوم از نوع s محاسبه می کند. و آن را به صورت $I(s,a) = log\frac{P(a|s)}{P(a)}$ می نویسد. این برابر با لگاریتم پارامتر هایی است که توسط چو و فسمن در نظر

گرفته شدند. سپس تفاوت این اطلاعات $I(\Delta s : a) = I(s,a) - I(-s,a)$ را محاسبه کردند که به سادگی برابر با $log\frac{P(a|s)}{P(a|-s)}$ است . همانند چو—فسمن این نسبت می تواند برای هر اسید آمینه محاسبه شود و ساختار توالی های جدید را پیش بینی کند. گرنیر و دیگران در [۳] این روش را برای پنجره های اطراف هر اسید آمینه به کار گرفتند. و برای هر اسید آمینه بر اساس مشاهدات اطرافش پیش بینی انجام دادند. و بااستفاده از مستقل بودن اسیدهای آمینه در هر پنجره ، تفاوت اطلاعات را در هر پنجره جمع بستند و پیش بینی برای کل توالی پروتئینی با حرکت این پنجره در طول توالی محاسبه می شود. مدل های با وابستگی های مرتبه بالا تر نیز در نسخه های جدید به کار گرفته شدند. دقت دو روش گفته شده در بالا ۵۵٪ گزارش شده اند.

۲.۲.۱ روش های مدرن

شاید یکی از مهمترین تاثیرات روش $Garnier - Osguthorpe - Robson$ معرفی ساختار پنجره ای در پیش بینی باشد که مسئله را به تصویر کردن یک بردار با طول ثابت به یک کلاس از خروجی‌ها کاهش می دهد. گرینر دری به روی کاربرد بسیاری از روشهای طبقه بندی آماری در آمار ، یادگیری ماشینی، و تشخیص الگو گشود. گستره وسیعی از کاربرد این الگو ها در سالهای بعد برای پیش بینی ساختارپروتئین ها به کار گرفته شده است.

بسیاری از موفق ترین این روشها بر اساس مدل های پیش بینی کننده غیر خطی مانند شبکه های عصبی و طبقه بندی بر اساس نزدیکترین همسایگی بوده اند. یکی از منفعت های این روش ها توانایی برای به دام انداختن وابستگی بین مشاهدات در همسایگی یکدیگر است. بسیاری از این روشها مانند روش $Garnier - Osguthorpe - Robson$ از ساختار پنجره ای استفاده کرده اند. یکی از معروفترین روشهای پیش بینی ساختار دوم که مرز ۷۰٪ دقت را شکست، روش PHD است. PHD بر اساس

مدل شبکه های عصبی چند لایه بوده است. که دقت پیش بینی را با استفاده از انطباق چند گانه توالی‌های همریخت [۲] افزایش داده است. با این حال PHD وقتی که روی یک توالی تنها به کار می رود با دقتی در حدود ٦۳٪ ، دقتی بالا تر از سایر روشهای موجود ندارد. برای اطلاع بیشتر در مورد این روش مرجع [۴] را ببینید. به هر حال یکی از مزایایی این روش معرفی توالی های همریخت به عنوان منبعی مناسب از اطلاعات برای پیش بینی ساختار دوم است و به طور تقریبی بهبودی در محدوده ٦ تا۴ درصد رابرای پیش بینی ساختار دوم، نسبت به روشهایی دارد که تنها از یک توالی تنها برای پیش‌بینی استفاده می کنند.

۳.۲.۱ روش های بیزی

استنباط بیزی یک مدل کلی را برای پیش بینی ساختار پروتئین ها با استفاده از توالی پروتئین فراهم می کند. و به صورت سودمندی از طریق گسترش ابزار های زیر عمل می کند:

- یک مدل آماری را بر اساس رابطه های بین ساختار/ توالی فراهم می کند که پارامتر های این مدل از روی ساختار های پروتئینی که از طریق کارهای آزمایشگاهی تعیین شده اند پیش بینی می شوند.
- روشهای محاسباتی را برای استنباط درباره ساختار پروتئین مورد استفاده قرار می دهد.

به کمک دو ابزار بالا ، استنباط بیزی می تواند پیش بینی های دقیقی را برای ساختار پروتئین فراهم کند ، همچنین برآورد های قابل اعتمادی را برای عدم قطعیت پیش بینی به دست می دهد. این در حالی است که برآورد عدم قطعیت پیش بینی به کمک روشهای ممیزی مانند شبکه های عصبی

[۲] $homolog$

ممکن نیست. به کمک مدل های بیزی و استفاده از توزیع توام ساختار– توالی می توان منابع متفاوتی از اطلاعات را در مدل لحاظ کرد. یکی دیگر از مزایای مدل های احتمالی بیزی نسبت به مدل های ممیزی در مدل کردن و پیش بینی اثرات متقابل بین $\beta - strand$ ها است که به الگو در آوردن این اثرات یکی از مشکلات بزرگ در پیش بینی توپولوژی پروتئین ها به شمار می رود و مدل های ممیزی به طور کلی در مدل کردن این اثرات ناتوان یا ضعیف به شمار می روند. مدل های مارکفی پنهان یکی از معروف ترین روشهای بیزی هستند که در زمینه های مختلفی به کار گرفته شده اند. و به عنوان ابزاری برای $sequence\ \ labeling$ به شمار می روند. یکی از اولین روشهایی که برای پیش بینی ساختار دوم به کار رفته است توسط آسای [۳] و دیگران ، [۵]، ارائه شده است که آنها چهار نوع ساختار دوم را در نظر گرفتند و هر یک از این ساختار های دوم به عنوان زیر مدلی مطرح شدندو برای هر زیر–مدل چهار یا پنج وضعیت پنهان در نظر گرفته شد که به طور جداگانه ای با استفاده از توالی‌های پروتئینی از پیش طبقه بندی شده در هر یک از چهار ساختار آموزش می دیدند. سپس این زیر–مدل‌ها که هر یک شامل وضعیت های پنهان می شدند با یکدیگر ادغام شدند. دقت این روش ۵۴٫۷٪ شده است. روش دیگری نیز از این مدل الهام گرفته و در سال ۲۰۰۶ به دقت ۶۸٫۸٪ برای توالی تنها و ۷۵٪ با استفاده از توالی های همریخت رسیده است. مارتین [۴] و دیگران در [۷], [۶] نیز برای هر ساختار از وضعیت های متفاوت و متعددی استفاده کردند و در آن از الگوریتم EM برای یافتن تعداد بهینه وضعیت های پنهان برای هر ساختار استفاده کردند. در دو روش گفته شده در بالا تنها یک مشاهده از هر وضعیت پنهان انتشار می یابد. یکی از دلایل استفاده از وضعیت های مختلف برای هر ساختار مدل کردن توزیع های مختلف طولی قطاع هایی است که در هر یک از ساختار های

---

[۳] $Asai$

[۴] $Martin$

پروتئینی مشاهده می‌شوند. اما اشمیدلر[۵] در [۹], [۸] به جای استفاده از وضعیت های مختلف برای هر ساختار، از مدل‌های مارکفی پنهان گسترش یافته استفاده کردند که به محض ورود به هر وضعیت پنهان یک توالی از مشاهدات از آنها انتشار می یابد. این مدل که به مدل نیمه مارکف قطاعی [۶] معروف شده‌اند تنها از توالی اسید های آمینه برای پیش بینی ساختار پروتئین استفاده می کند دارای دقتی در حدود ۶۷٪ است. مدل های بررسی شده در این پایان نامه نیز مدل های نیمه مارکف قطاعی هستند.

۳.۱ مطالب بررسی شده در این پایان نامه

در فصل دوم مدل های مارکفی پنهان مرور شده و ابزار های استنباط توسط این مدل ها شرح داده شده‌اند. در این فصل مدل های نیمه مارکف پنهان نیز مورد بررسی قرار گرفته اند. این فصل با مقدمه‌ای بر پروتئین ها و نقش آنها در جریان اطلاعات ژنتیک پایان می یابد. در فصل سوم دو مقاله مورد تجزیه و تحلیل قرار گرفته است که هر دوی آنها بر اساس مدل پیشنهادی اشمیدلر هستند. مقاله اول به دنبال بهبود در دقت پیش بینی مدل های $SSMM$ بااستفاده از توالی های همریخت است. این مقاله [۱۰] با عنوان

Chu. W., Ghahramani. Z., Podtelezhnikov. A., David. L. W. (2006). Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 3:98-113.

---

[۵] $Schmidler$

[۶] $Segmental\ \ Semi-Markov\ \ Model(SSMM)$

است و با استفاده از مدل های پارامتری پروفایل انطباق چند گانه پروتئین های همریخت را مدل می‌کند. این روش پارامتری مشکل بزرگ دیگری را نیز حل می کند. در مدل های مارکفی پنهان با افزایش طول پنجره وابستگی تعداد پارامتر های مدل به صورت نمایی افزایش می یابد که با توجه به محدودیت در داده های پروتئینی تعیین ساختار شده برآورد این پارامتر ها عملی نیست. مدل پارامتری در نظر گرفته شده در این مقاله به گونه ای است که با افزایش طول پنجره وابستگی تعداد پارامتر هایش به طور خطی افزایش می یابد.

این مقاله همچنین به بررسی تاثیرات متقابل بین $\beta - strand$ هامی پردازد که یکی از مزایای مدل های احتمالی نسبت به مدل های ممیزی مانند شبکه های عصبی است. پیش بینی این تاثیرات متقابل در بهبود پیش بینی ساختار سه بعدی پروتئین هاکاربرد زیادی دارد و دقت آنها را افزایش می دهد. همچنین مدل کردن این تاثیرات ، به ویژه هنگامی که فقط ازاطلاعات توالی اسید های آمینه استفاده می شود، یکی از ابزارهای ضروری در افزایش دقت پیش بینی پروتئین هابه حساب می آید. اکثر روشهایی که امروزه برای پیش بینی ساختار دوم به کار می روند این اطلاعات را نادیده می گیرند و از اطلاعات محلی استفاده می کنند که یک چهارم اطلاعات مورد نیاز را برای پیش بینی دقیق شامل می‌شود.

مقاله دیگری که مورد بررسی قرار گرفته ساختار های قویتری را برای مدل کردن وابستگی بین مشاهدات در مدل های $SSMM$ پیشنهاد می کند. و در آن سعی شده تا با معرفی این وابستگی‌هادقت پیش بینی ساختار دوم پروتئین ها با استفاده از توالی اسید های آمینه افزایش یابد. دراین مقاله [۱۱] با عنوان

Aydin. Z., Altunbasak. Y., Borodovski. M. (2004). Protein secondary structure prediction with semi markov HMMs. *Proc IEEE Int'l Conf. Acoustics Speech, and Signal*

Processing .

سعی شده تا با استفاده از آزمون های آماری χ^2 همبستگی های معنی دار بین اسید های آمینه جستجو شود و از این همبستگی ها در مدل های $SSMM$ استفاده شود.

در فصل چهارم دو مدل ارائه شده است. در مدل اول سعی شده تا با استفاده از اطلاعات موجود دردو طرف هر مشاهده همبستگی های بیشتری را مدل کنیم. زیرا در مدل های $SSMM$ با توجه به ساختار مارکفی بودن تنها وابستگی های یک طرفه لحاظ می شود. برای استفاده از اطلاعات موجود در دو طرف هر مشاهده مدل های شرطی متفاوتی را در نظر گرفتیم که هر یک از این مدل های شرطی وابستگی بین هر مشاهده با مشاهدات اطرافش را نشان می دهند. به منظور جمع کردن تمامی این مدل‌های شرطی در یک مدل، برای هر یک از این مدل ها با توجه به اهمیت شان در لحاظ کردن وابستگی ها، وزن هایی در نظر گرفته شده‌اند.

مدل دیگری که در فصل چهارم آمده است سعی دراستفاده از اطلاعات موجود در توالی های همریخت در مدل های $SSMM$ دارد. برای انجام این کار از شبکه های عصبی به عنوان ابزار کاهش حجم داده‌ها کمک می گیرد.