مقایسه مشبکه های مختلف در مساله نشاندن پروتئین‌ها

میثم عقیقی

زیر نظر

دکتر محمد قدسی

تابستان ۱۳۸۹

دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

تهران

مقایسه مشبکه های مختلف در مدل سازی پروتئین ها در مشبکه ها

چکیده

یک جواب موفق برای مساله نشاندن پروتئین‌ها مشخص کننده رشته باقیمانده برای تولید مولکول یک ساختار پروتئین مشخص می‌باشد، از این رو مساله نشاندن پروتئین یک ابزار بالقوه در صنعت تولید دارو می‌باشد. به این صورت که می‌تواند اولین تقریب را برای یک رشته اسید آمینه مقدماتی از یک پروتئین بدست بیاورد. مقاله ما یک تحقیق آماری و محاسباتی روی یک پایگاه داده از پروتئین‌ها است برای مقایسه و در نهایت فهمیدن این که کدام مشبکه‌ها بستر بهتری برای نشاندن پروتئین‌ها بر روی خود هستند و در نشاندن پروتئین‌ها تقریب بهتری را به ما می‌دهند.

در این تحقیق ما ابتدا به مطالعه خواصی از مشبکه‌ها پرداختیم که در نشاندن پروتئین‌ها تأثیرگذار و پراهمیت می‌باشند. سپس چندین مشبکه با این خاصیت‌ها کاندیدا کردیم، و با استفاده از الگوریتم موجود در [۱]، پروتئین‌های نشانده شده در این مشبکه‌ها را با هم مقایسه کردیم. در نهایت به این نتیجه رسیدیم که بهترین مشبکه‌ها برای نشاندن پروتئین مشبکه‌های FCC (Face Centered Cubic) و e-FCC (extended FCC) می‌باشند.

واژه‌های کلیدی: *پروتئین، مشبکه، تاکردن پروتئین، نشاندن پروتئین، بانک داده پروتئین، ساختار پروتئین.*

فهرست مطالب

فهرست شکل ها

فهرست جدول ها و نمودارها

فصل ۱

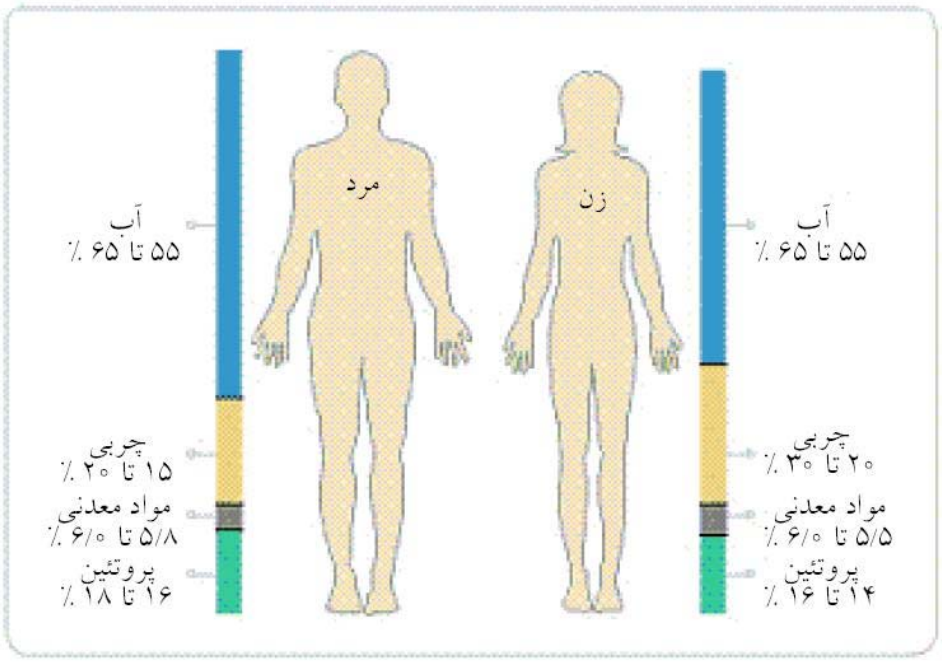
مقدمه

زیست‌شناسی دانش مربوط به مطالعه موجودات زنده است. این دانش به بررسی ویژگی‌ها و رفتار سازواره‌ها، چگونگی پیدایش گونه ها و افراد، و نیز به بررسی برهمکنش جانداران با یکدیگر و محیط پیرامونشان می‌پردازد. شاخه‌ای از زیست‌شناسی به نام زیست‌شناسی مولکولی، مطالعهٔ زیست‌شناسی در سطح مولکولی است. این حوزه دارای وجوه مشترکی بـا زیست‌شناسی، شیمی، و به طور خاص، با علم ژنتیک و بیوشیمی است. بحث عمده در زیست‌شناسی مولکولی استنباط برهم‌کنش بین سیستم‌های درون سلولی، من جمله، برهم‌کنش‌های ,RNA ،DNA، و پروتئین‌سازی است. به‌علاوه، چگونگی تنظیم این برهم‌کنش ها مورد بررسی قرار می‌گیرد.

امروزه استفاده از دانش رایانه در مسائل زیست‌شناسی مولکولی، منجر به تولد شاخه‌ی جدیدی به نام زیست‌انفورماتیک شده است. زیست‌انفورماتیک یا بیوانفورماتیک دانش استفاده از علوم کامپیوتر و آمار و احتمالات در شاخه زیست‌شناسی مولکولی است. در چند دههٔ اخیر ، پیشرفت در زیست‌شناسی مولکولی و تجهیزات مورد نیاز تحقیق در این زمینه باعث افزایش سریع تعیین توالی ژنوم بسیاری از گونه‌های موجودات شد، تا جایی که پروژه‌های تعیین توالی ژنوم‌ها از پروژه‌های بسیار رایج به حسب می‌آیند.امروزه توالی ژنوم بسیاری از موجودات ساده مانند باکتری‌ها تا موجودات بسیار پیشرفته چون یوکاریوت‌های پیچیده شناسایی شده‌است. پروژهٔ شناسایی ژنوم انسان در سال ۱۹۹۰ آغاز شد و در سال ۲۰۰۳ پایان یافت و اکنون اطلاعات کامل مربوط به توالی هر ۲۴ کروموزوم انسان موجود است.

ما در این پایان‌نامه به بررسی و مقایسه مسائلی حول یک موضوع خاص در شاخه زیست‌انفورماتیک می‌پردازیم.

بدن موجودات زنده و به خصوص انسان‌ها از انواع مختلفی از مولکول‌ها تشکیل شده است که هر یک نقش و رفتار خاص خود را دارند. یک رده ویژه از این مولکول‌ها، پروتئین‌ها هستند که بخش مهمی از وظایف زیستی را در بدن موجودات زنده برعهده دارند و همان‌طور که در شکل ۱-۱ هم دیده می‌شود، قسمت قابل توجهی از مولکول‌های سازنده‌ی بدن انسان را شامل می‌شوند.



شکل ۱-۱: درصد حصور انواع ملاول‌ها در بدن انسان در وصعیت طبیعی

شکل از [۲۰]

۱. اهمیت مطالعه و شناخت پروتئین‌ها

در باب اهمیت مطالعه‌ی پروتئین‌ها قسمتی از [۲] را در ادامه می‌آوریم:

پروتئین‌ها یکی از داغ‌ترین مساله‌های بین رشته‌ای هستند که متخصصین زیست شناسی، شیمی، فیزیک، علم کامپیوتر، علم نانو و متخصصین شاخه‌هایی که ذات بین‌رشته‌ای‌شان حتی در نام‌شان آشکار است، مانند زیست‌فیزیک، زیست‌انفورماتیک و زیست‌شیمی که در آن مشغول تحقیق هستند.

سه دلیل عمده برای جذابیت این مساله از دیدگاه فیزیک وجود دارد. اول آن که پروتئین-ها مثال کلاسیکی از سیستم‌های پیچیده هستند. سیستم‌هایی که اجزا و روابط بین اجزای آن‌ها، چندان دشوار و تحلیل‌ناپذیر به نظر نمی‌رسند، ولی جمع شدن تعداد کافی از این اجزا، رفتارهایی

را سبب می‌شود که با رفتار ساده‌ی اجزا بسیار متفاوت است، مانند مغز، لانه‌ی مورچه و ... پروتئین‌ها نیز علی‌رغم کوچکی (چند یا چنده نانومتر) و سادگی (چندهزار یا چندصدهزار اتم)، رفتارهای پیچیده‌ای دارند.

دلیل دوم اهمیت پروتئین‌ها، قرارگرفتن آن‌ها در مرز زنده بودن و نبودن است. از یک سو آن‌قدر کوچک‌اند که با روش‌های فیزیک و شیمی قابل بررسی هستند و از سوی دیگر به اندازه کافی برای نشان دادن بعضی از خواص موجودات زنده بزرگ هستند. پروتئین‌ها کارکرد زیستی دارند و در حقیقت آجرهای سازنده و ماشین‌آلات بدن موجودات زنده هستند.

دلیل دیگر اهمیت پروتئین‌ها، رابطه‌ی آن‌ها با علم نانو و فن‌آوری نانو است. در واقع میلیون‌ها سال پیش از این که تکنولوژی مدرن به صرافت استفاده از امکانات فن‌آوری نانو بیفتد، طبیعت از ماشین‌های نانو استفاده می‌کرده است. مطالعه‌ی پروتئین‌ها به معنی بررسی نحوه‌ی فعالیت، محدودیت‌ها و امکانات نانوابزارهایی است که فعال‌اند و زحمت طراحی و بررسی امکان فعال‌بودن‌شان از دوش انسان برداشته شده است. این مطالعه به خصوص در درک محدودیت‌ها در پیشرفت فن‌آوری نانو بسیار اهمیت دارد. درهم تنیدگی این دو فن‌آوری آن‌قدر زیاد است که هنوز تفاهمی در مورد استفاده از نام نانوزیست‌فن‌آوری یا زیست‌نانوفن‌آوری وجود ندارد.

اهمیت دیگری نیز که می‌توان برای مطالعه‌ی رفتار پروتئین‌ها بیان کرد استفاده آن در روش محاسبات با DNA می‌باشد، که در این روش محاسبه به جای استفاده از فن‌آوری‌های رایانه‌ای مبتنی‌بر مدارهای سنتی سیلیکونی، از DNAها، زیست‌شیمی و زیست‌شناسی مولکولی استفاده می‌شود. [۳] و برای اولین بار لئونارد آدلمن از این روش برای حل مساله فروشنده دوره‌گرد با ۷ رأس استفاده کرد.

از مهم‌ترین خواص محاسبات با DNA، توانایی بالایی آن در پردازش موازی است. تعداد زیادی مولکول DNA به طور همزمان می‌توانند راه‌های مختلفی را برای حل یک مساله بیازمایند. در نتیجه مسائل سختی مانند SAT را که امکان حل‌شان در زمان معقول با رایانه‌های فعلی وجود ندارد، می‌توان به کمک محاسبات با DNA در زمان معقولی حل نمود. خاصیت دیگر محاسبات با DNA، مصرف کم انرژی برای انجام محاسبات شمرده می‌شود.

۲. مدل‌سازی پروتئین‌ها

پروتئین‌ها از اجزای ساده‌ای تشکیل شده‌اند، ولی طریقه کنارهم قرار گرفتن آن‌ها ساختارهای پیچیده‌ای به وجود آورده است، تا جایی که مدل کردن و شبیه‌سازی آن‌ها از توان بهترین رایانه‌های امروزی نیز خارج است. از این رو برای مدل‌سازی پروتئین‌ها، ابتدا از ساده‌سازی آن‌ها استفاده می‌شود.

به طور کلی، در مقالات مربوطه و این مقاله، پروتئین‌ها به شکل زنجیره‌ای از اسیدهای آمینه ساده-سازی شده‌اند که هر اسید آمینه با یک رأس مدل شده است. حتی در این شرایط نیز وضعیت‌های مختلفی که پروتئین می‌تواند به خود بگیرد، فضایی پیوسته و غیرقابل بررسی است، لذا دانشمندان عمل زیست-انفورماتیک برای گسسته کردن این فضا به سراغ مشبکه‌ها می‌روند.

۳.معرفی مسائل و خلاصه کارهای انجام شده

مساله نشاندن پروتئین بر روی یک مشبکه، در حالت کلی بدون نیاز به استفاده از مدل HP بیان می‌شود. این مساله به این صورت بیان می‌شود که یک پروتئین دلخواه از فضای واقعی به دست ما رسیده است، ولی ما نیاز داریم که جایگاه رئوس رشته اسید آمینه‌ای را بر روی یک مشبکه بدانیم تا بتوانیم بر روی این رشته الگوریتم‌هایی اجرا کنیم که در حالت کلی نمی‌توان این کار را انجام داد. از این رو می‌خواهیم زیرمجموعه‌ای از رئوس مشبکه را بیابیم که بیشترین شباهت را به پروتئین اولیه داشته باشد. معیار این شباهت نیز به چند طریق تعریف شده است (c-RMS، d-RMS و a-RMS)، که همه این معیارها از جنس میانگین گیری میان طول‌ها (یا زاویه‌ها)ی پروتئین اصلی و نشانده شده آن روی مشبکه می‌باشد. به یافتن بهترین زیرمجموعه متناظر با پروتئین از رئوس مشبکه، نشاندن پروتئین بر روی مشبکه می‌گویند.

کارهایی که در زمینه نشاندن پروتئین‌ها بر روی مشبکه‌ها انجام شده است را می‌توان به چند طریق نگاه کرد. دسته‌ای از کارهای انجام شده در زمینه بررسی سختی این مساله است، به عنوان مثال [۲۱] می-آید و ثابت می‌کند که نشاندن پروتئین روی مشبکه مکعبی NP-Complete است. مقاله [۲۰] با اثباتی شبیه به همان اثبات می‌آید و ثابت می‌کند که این مساله برای مشبکه مثلثی سه‌بعدی نیز NP-

Complete است. هم‌چنین مقاله‌های بسیاری آمده‌اند و الگوریتم‌هایی تقریبی برای نشاندن پروتئین ارائه کرده‌اند. چند مقاله هم به بررسی مشبکه‌های مختلف و تلاش برای یافتن مشبکه‌ای مناسب‌تر برای نشاندن پروتئین‌ها کرده‌اند، در این راستا مشبکه‌های جدیدی مانند s-FCC و e-FCC به وجود آمده‌اند. و ما در این مقاله این مشبکه‌ها را علاوه بر مشبکه‌های ارائه شده توسط خودمان، مورد بررسی قرار می‌دهیم.

هم چنین در چندین مقاله مانند مقاله [۲۰] خلاصه کارهای انجام شده در این زمینه به طور کامل و جامع بیان شده است.

۴.ساختار مقاله

در این مقاله ما قصد داریم مشبکه‌های مختلف را با یکدیگر مقایسه کنیم و ببینیم که با هر کدام از سه معیار c-RMS، d-RMS و a-RMS پروتئین‌هایی که بر روی این مشبکه‌ها نشانده می‌شوند، در کدام یک از این مشبکه‌ها به پروتئین اصلی نزدیک‌تر خواهد بود.

برای این منظور، ابتدا در فصل ۲ خواننده را با مفاهیم مربوط به این مبحث آشنا کرده، سپس در فصل ۳ که بخش اصلی کار ما است، ابتدا با مطالعه پایگاه‌داده پروتئین‌ها و مرور کارهای قبلی انجام شده در این زمینه، خواصی را برای مشبکه‌ای که بهترین جواب را به ما می‌دهد پیدا می‌کنیم، سپس مشبکه‌هایی با این خواص را بررسی کرده و نتایج آن‌ها را با یکدیگر مقایسه می‌کنیم. سپس در فصل ۴ خصوصیات انجام این مقایسه را توضیح و شرح می‌دهیم. از این جمله مشخصات پایگاه‌داده پروتئین‌ها و الگوریتمی که برای نشاندن پروتئین‌ها بر روی مشبکه‌ها استفاده شده بود، می‌باشند.

فصل ۲

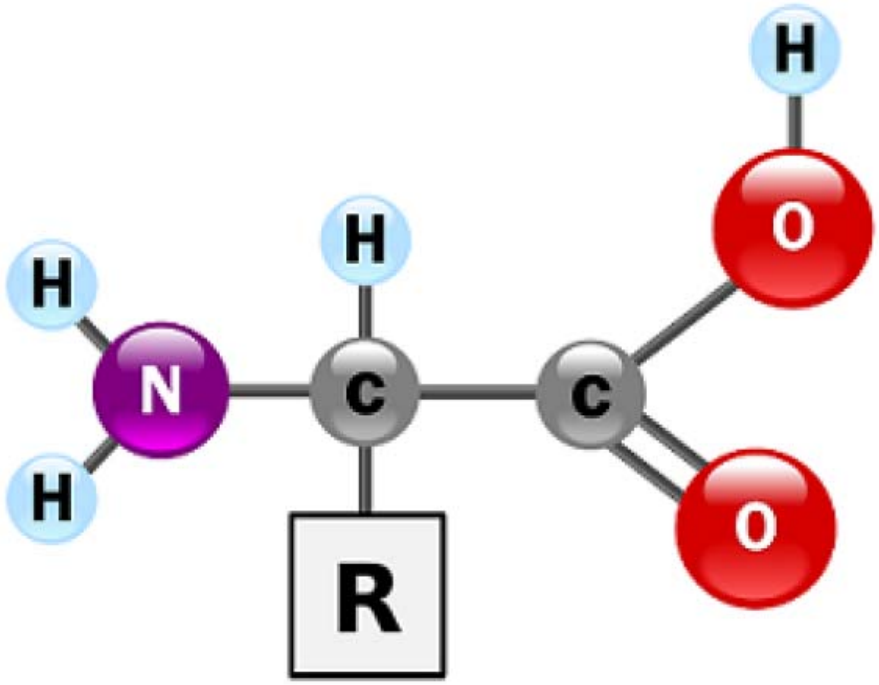
تعاریف و مفاهیم اولیه

در تدوین این بخش به علت مشابهت فراوان مفاهیم اولیه لازم و پیش‌نیازهای علم زیست-انفورماتیک این مقاله به [۲۰]، با اطلاع نویسنده آن مقاله، فصل ۲ از آن عیناً آورده شده است:

۱. پروتئین‌ها و اجزای سازنده

از نظر اندازه، پروتئین‌ها در رده‌ی «درشت مولکول‌ها» قرار دارند. درشت مولکول‌ها، مولکول‌های خیلی بزرگی (از نظر جرم مولکولی، تعداد اتم و ...) هستند که معمولاً از کنار هم قرار گرفتن قطعات کوچک‌تر مولکولی ساخته می‌شوند. قطعات سازنده‌ی این‌گونه درشت‌مولکول‌ها را «مونومر» می‌نامند، و به مولکول‌هایی که از ترکیب چند مونومر درست شده باشند، «پلیمر» می‌گویند.

پروتئین‌ها نیز پلیمر هستند و از قطعات مولکولی کوچک‌تری تشکیل شده‌اند. مونومرهای سازنده پروتئین‌ها «اسیدهای آمینه» هستند. علت نام‌گذاری اسیدهای آمینه به این اسم، ساختار شیمیایی آن-هاست: هم دارای یک بخش اسیدی کربوکسیل (COOH) هستند و هم یک بخش آمینه (NH_2) دارند. در شکل ۲-۱، شمای کلی ساختار یک اسید آمینه در یک مدل دوبعدی نشان داده شده است. بخش اسیدی آن ($COOH$) در سمت راست، و بخش آمینه‌ی آن (NH_2) در سمت چپ تصویر دیده می‌شود. این دو بخش توسط یک اتم کربن مرکزی که به C_α معروف است به هم متصل‌اند. اتم C_α از بالا با یک اتم هیدروژن (H) و از پایین با یک شاخه‌ی جانبی پیوند دارد. این شاخه‌ی جانبی که با R نمایش داده شده، به «زنجیره‌ی جانبی» معروف است.



شکل ۲-۱ شمای کلی یک اسید آمینه

زنجیره‌ی جانبی همه اسیدهای آمینه یکسان نیست. اسیدهای آمینه‌ی مختلف زنجیره‌های جانبی متفاوتی دارند و به طور کلی، انواع مختلف اسیدهای آمینه را از روی زنجیره‌ی جانبی آن‌ها تعیین می‌کنند. در مجموع ۲۲ نوع اسید آمینه با زنجیره‌های جانبی متفاوت شناسایی شده‌اند. در جدول ۲-۱، بعضی مشخصات و خصوصیات این ۲۲ اسید آمینه آورده شده است. یکی از این خصوصیات، «آب گریزی» یک اسید آمینه که در قسمت‌های بعدی در مورد آن توضیح داده خواهد شد.

با ثابت نگه‌داشتن مکان اتصال C_α به بخش اسیدی و بخش آمینه در فضا، به دو روش می‌توان اتم هیدروژن و زنجیره‌ی جانبی را به C_α متصل نمود که تعیین می‌کنند اسید آمینه به اصطلاح چپ‌دست یا راست‌دست باشد. ولی به شکل کاملاً نامتقارن، همه‌ی اسیدهای آمینه‌ای که در طبیعت یافت می‌شوند، چپ‌دست هستند.

برای تشکیل یک پروتئین، مونومرهای آن (اسیدهای آمینه) در کنار یکدیگر قرار می‌گیرند و به شکلی خاص با هم پیوند برقرار می‌کنند تا پلیمر مورد نظر ساخته شود. در هنگام ایجاد پیوند بین دو اسید آمینه (نه لزوماً مشابه)، قسمت OH از بخش اسیدی ($COOH$) یک اسید آمینه، و یک H از بخش آمینه‌ی (NH_2) اسید آمینه‌ی دیگر، جدا می‌شوند و با هم یک مولکول آب تشکیل می‌دهند. باقیمانده‌ی دو اسید آمینه‌ی مذکور، از محل قطعات جداشده، با هم پیوند پپتیدی برقرار می‌کنند. شکل ۲-۲ فرآیند کلی ایجاد یک پیوند پپتیدی بین دو اسید آمینه را نمایش داده است.

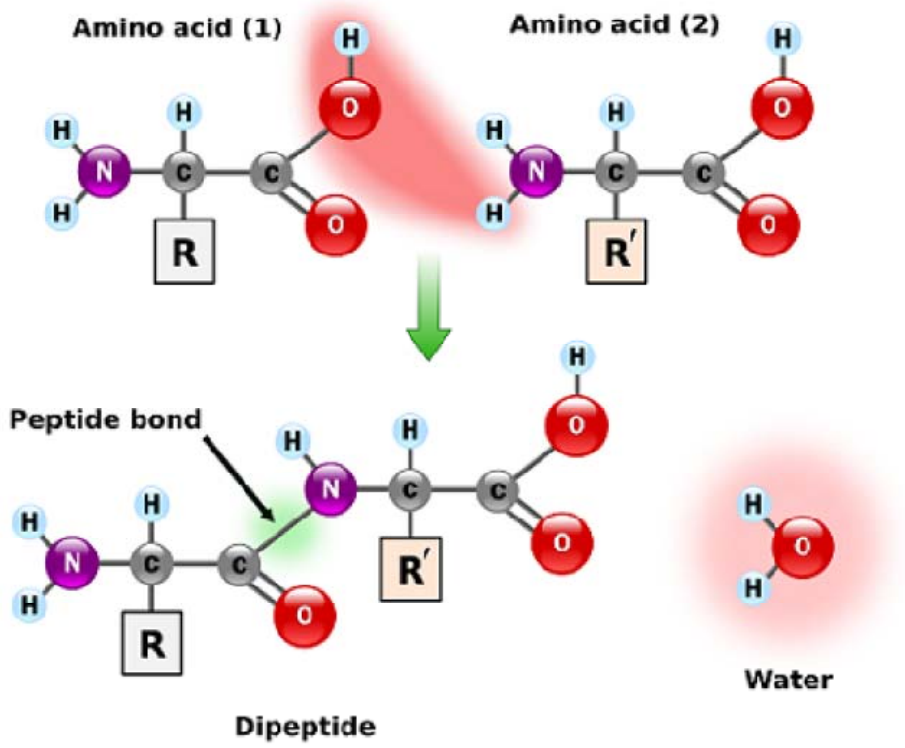
به خاطر ساختار اسیدهای آمینه و چگونگی امکان اتصال آن‌ها به‌هم، اسیدهای آمینه تنها در یک ردیف می‌توانند در کنار هم قرار بگیرند و لذا پروتئین‌ها را نوعی پلیمر خطی می‌دانند. یک سر این پلیمر خطی به عامل اسیدی ($COOH$) و سر دیگر آن به عامل آمینه (NH_2) ختم می‌شود. در نتیجه ترتیب اسیدهای آمینه در پروتئین‌ها ترتیبی جهت‌دار است و اگر این ترتیب را برعکس کنیم پروتئین متفاوتی حاصل می‌شود. گرچه خیلی از پروتئین‌ها خطی هستند، ولی گاهی یک ردیف طولانی از اسیدهای آمینه، پس از چرخیدن در فضا، می‌تواند دو سرش را در کنار هم قرار دهد و با ایجاد یک پیوند پپتیدی دیگر بین دو سرش، این ساختار خطی را به یک ساختار حلقوی تبدیل کند. ولی ما در این‌جا از پروتئین‌های حلقوی عمدتاً صرف‌نظر می‌کنیم.

۲. انرژی آزاد و ساختارهای پروتئین‌ها

عمده‌ی نیروها و پیوندهایی که بین اتم‌ها و مولکول‌ها وجود دارند را می‌توان با نیروهای الکترومغناطیسی توضیح داد. ولی برای توجیح بعضی رفتارها، روش ساده‌تری نیز وجود دارد که گرچه توصیف نیرویی واقعی نیست، تمامی رفتارها را به خوبی توضیح می‌دهد. این توصیف، تمایل سیستم به کاهش انرژی آزاد یا همان انتروپی است و نیروی عامل در این توصیف را نیروی انتروپیک می‌نامند. به طور کلی، یکی سیستم تمایل دارد انرژی آزاد خود را تا جای ممکن کاهش دهد. به کمک این نیرو، آب شدن یخ و پخش شدن جوهر روی کاغذ را نیز می‌توان توجیح کرد. پروتئین‌های شناور در آب نیز سعی می‌کنند سطح انرژی آزاد خود را کاهش دهند و برای این کار در فضا به شکلی به خود می‌پیچند و ساختار خاصی در فضا پیدا می‌کنند. برای آشنایی با جزئیات این ساختار لازم است ابتدا با پیوندهای درون پروتئینی و پیوند پروتئین‌ها با آب آشنا شویم.

نام	نماد	مخفف	آب‌گریزی	درصد حضور در پروتئین‌ها
Alanine	A	Ala	آب‌گریز	۷.۸
Cysteine	C	Cys	آب‌گریز	۱.۹
Aspartic acid	D	Asp	آب‌دوست	۵.۳
Glutamic acid	E	Glu	آب‌دوست	٦.۳
Phenylalanine	F	Phe	آب‌گریز	۳.۹
Glycine	G	Gly	آب‌گریز	۷.۲
Histidine	H	His	آب‌دوست	۲.۳
Isoleucine	I	Ile	آب‌گریز	۵.۳
Lysine	K	Lys	آب‌دوست	۵.۹
Leucine	L	Leu	آب‌گریز	۹.۱
Methionine	M	Met	آب‌گریز	۲.۳
Asparagine	N	Asn	آب‌دوست	۴.۳
Pyrrolysine	O	Pyl	آب‌دوست	
Proline	P	Pro	آب‌گریز	۵.۲
Glutamine	Q	Gln	آب‌دوست	۴.۲
Arginine	R	Arg	آب‌دوست	۵.۱
Serine	S	Ser	آب‌دوست	٦.۸
Threonine	T	Thr	آب‌دوست	۵.۹
Selenocysteine	U	Sec	آب‌گریز	
Valine	V	Val	آب‌گریز	٦.٦
Tryptophan	W	Trp	آب‌گریز	۱.۴
Tyrosine	Y	Tyr	آب‌دوست	۳.۲

جدول ۲-۱: لیست اسیدهای آمینه



شکل ۲-۲: چگونگی ایجاد پیوند پپتیدی بین
دو اسید آمینه

پیوند پپتیدی سخت‌ترین پیوند بین اجزای یک پروتئین است. انرژی این پیوندها که نوعاً کووالانسی هستند، ۵۰ تا ۱۵۰ کیلوکالری بر مول است. پیوندهای پپتیدی یک پروتئین، در دماها و محیط-های گوناگون، تا آستانه‌ی تجزیه‌ی پروتئین برقرار می‌مانند. لذا پیوندهایی پپتیدی را به نام پیوندهای سخت درون پروتئینی رده‌بندی می‌کنند. در مقابل این پیوندها، پیوندهای درون پروتئینی دیگری نیز تعریف می-شوند که به پیوندهای نرم معروف‌اند.

پیوندهای نرم درون پروتئینی پیوندهای ضعیف‌تری هستند که بین دو اسید آمینه‌ی نه لزوماً متوالی در یک پروتئین برقرار می‌شوند. اندازه‌ی انرژی این پیوندها از مرتبه‌ی چند کیلوکالری بر مول است و درنتیجه برخلاف پیوندهای سخت، با تغییر دما و دیگر عوامل محیطی به راحتی دست‌خوش تغییر می‌گردند. مهم‌ترین نمونه‌ی پیوندهای نرم، پیوند هیدروژنی است. پیوند هیدروژنی بین هیدروژن و اتم‌هایی مانند اکسیژن و نیتروژن برقرار می‌شود. در این پیوند، ابر الکترونی اطراف یک اتم هیدروژن، به سمت اتم دیگری که با آن هیدروژن پیوند کووالانسی دارد، کشیده می‌شود و در نتیجه، هسته‌ی اتم هیدروژن که دیگر الکترون ندارد، عریان می‌شود. در این شرایط، بار مثبت هسته‌ی این هیدروژن الکترون‌های اتم‌های دیگری مانند اکسیژن را جذب می‌نماید و با آن اتم‌ها پیوند ضعیفی برقرار می‌کند که پیوند هیدروژنی نامیده شده است.

پیوند هیدروژنی می‌تواند بین دو مولکول متفاوت برقرار شود، مانند پیوند بین هیدروژن یک مولکول آب و اکسیژن یک مولکول دیگر آب. خواص قطبی مولکول‌های آب با این پیوندهای هیدروژنی عجین است. نمونه‌ی دیگر پیوندهای هیدروژنی بین دو مولکول متفاوت، پیوند هیدروژنی بین مولکول‌های آب و اتم‌های یک پروتئین شناور در آب است. این‌گونه پیوندها در تعیین ساختار پروتئین‌ها تأثیر به‌سزایی دارند.

علاوه‌بر پیوند بین دو مولکول متفاوت، پیوند هیدروژنی هم‌چنین می‌تواند بین اتم‌های هیدروژن یک مولکول با اتم‌های دیگر همان مولکول که با کمی چرخش در مجاورت آن هیدروژن قرارگرفته‌اند برقرار شود. نمونه‌ی مهم این‌گونه پیوندهای هیدروژنی در پروتئین‌ها است. اسیدهای آمینه‌ی غیرمجاور در یک پروتئین پس از چرخیدن در فضا در مجاورت هم قرار می‌گیرند و هیدروژن یک اسید آمینه با اتم‌های اسید آمینه‌ی دیگر پیوند هیدروژنی برقرار می‌کند. این‌گونه پیوندهای نرم درون پروتئینی، بعد از پیوندهای سخت درون پروتئینی، دومین منشأ در تعیین ساختار پروتئین‌ها هستند.

پروتئین‌ها در شرایط طبیعی با مولکول‌های آب احاطه شده‌اند. زنجیره‌ی جانبی بعضی اسیدهای آمینه، قطبی نیستند و با آب پیوند هیدروژنی برقرار نمی‌کنند. حضور چنین ساختارهایی در آب، شبکه‌ی پیوندهای هیدروژنی را دچار مشکل می‌کند، چرا که مولکول‌های آب یا باید از بعضی پیوندهای هیدروژنی

صرف نظر کنند یا باید آرایشی خاص به خود بگیرند. در هر دو حالت، انرژی آزاد مولکول‌های آب افزایش می‌یابد.

همانند قطره‌ی روغن روی آب که با جمع شدن، سعی در کاهش سطح تماس خود با آب را دارد، ساختارهای غیرقطبی در یک پروتئین هم سعی می‌کنند با کنار هم قرارگرفتن و پنهان شدن در پشت اسیدهای آمینه‌ی قطبی، تا جای ممکن سطح تماس خود را با مولکول‌های آب کاهش دهند، تا انرژی آزاد سیستم را تا جای ممکن کاهش دهند. این رفتار این دسته از اسیدهای آمینه، این تصور را ایجاد می‌کند که می‌خواهند به نوعی از مولکول‌های آب بگریزند، و به همین دلیل این دسته از اسیدهای آمینه را آب‌گریز یا آب‌ترس می‌نامند.

در مقابل اسیدهای آمینه‌ی آب‌گریز، اسیدهای آمینه‌ی آب‌دوست قرار دارند. زنجیره‌ی جانبی اسیدهای آمینه‌ی آب‌دوست، قطبی است و این اسیدهای آمینه به راحتی با مولکول‌های قطبی آب پیوند برقرار می‌کنند. در جدول ۲-۱، مشخص شده است که کدام اسیدهای آمینه آب‌گریز، و کدام‌ها آب‌دوست هستند.

معمولاً بعد از پیچش و شکل‌گیری پروتئین در فضا (در محیط آب)، اسیدهای آمینه‌ی آب‌گریز بیشتر در بخش‌های مرکزی یا به اصطلاح هسته‌ی پروتئین قرار می‌گیرند، و اسیدهای آمینه‌ی آب‌دوست بیشتر در بخش خارجی پروتئین یا به اصطلاح پوسته‌ی پروتئین ظاهر می‌شوند. بعد از پیوندهای سخت درون پروتئینی، آب‌گریزی و آب‌دوستی اسیدهای آمینه، در کنار پیوندهای نرم درون‌پروتئینی، از مهم‌ترین عوامل شناخته‌شده در تعیین شکل فضایی پروتئین‌ها هستند.

بر اساس اطلاعات موجود، نیازها، مدل‌ها و فضای بحث در مورد پروتئین‌ها، می‌توان ساختار پروتئین‌ها را با جزئیات کم‌تر یا بیش‌تری در نظر گرفت. با توجه به این موضوع برای پروتئین‌ها چهار ساختار در نظر گرفته‌اند که به ساختارهای نخستین، دومین، سومین و چهارمین معروف‌اند. شکل ۲-۳، شمای کلی از این ساختار را نشان می‌دهد. در ادامه، این چهار ساختار را به طور مختصر معرفی می‌کنیم.

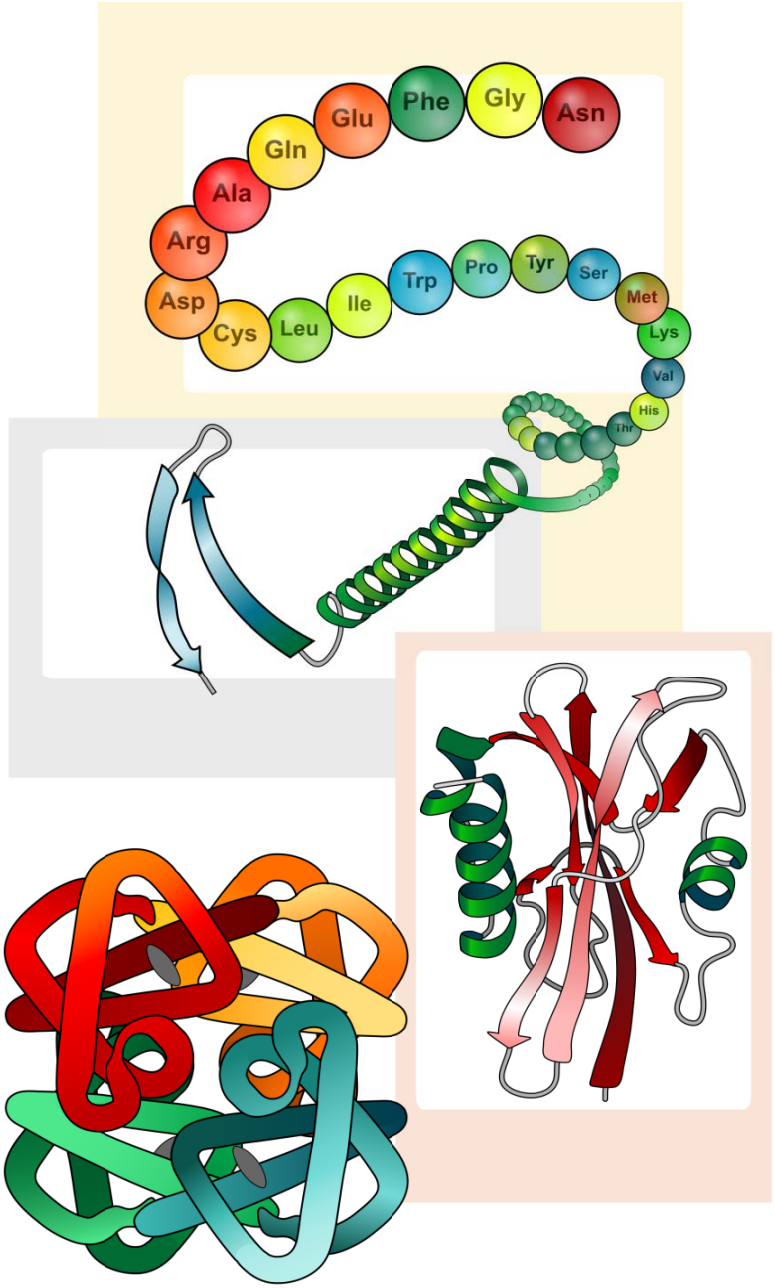
ساختار نخستین یک پروتئین تنها پیوندهای سخت آن را درنظر میگیرد. در نتیجه می‌توان آن را با دنباله-ی اسیدهای آمینه‌ی تشکیل‌دهنده‌ی پروتئین متناظر دانست. پس اگر اسیدهای آمینه‌ی پروتئین را با حروف نمادشان نشان دهیم، و این حروف را به ترتیب حضور اسید آمینه‌ی متناظرشان در پروتئین کنار هم قرار دهیم، رشته‌ی حاصل نیز نمایش‌گر ساختار نخستین پروتئین می‌باشد. در ادامه، منظور از ساختار نخستین پروتئین همین رشته است. در ساختار نخستین یک پروتئین، دو اسید آمینه مجاورند اگر و تنها اگر جایگاه‌هایی متوالی در رشته‌ی مذکور داشته باشند.

ساختار دومین یک پروتئین علاوه بر پیوندهای سخت، بخشی از پیوندهای نرم درون پروتئینی را نیز در نظر می‌گیرد. این پیوندهای نرم مربوط به پیچه‌های آلفا و صفحه‌های بتا است. پیچه‌های آلفا می‌توانند در یک یا چند قسمت از یک پروتئین ظاهر شوند و به این شکل تشکیل می‌گردند که قسمتی از دنباله‌ی اسیدهای آمینه، ساختاری مانند سیم‌پیچ پیدا می‌کند تا همه یا بخشی از اسیدهای آمینه در این پیچه بتوانند با اسیدهای آمینه‌ای که در دور بعدی پیچه در مجاورتشان قرار می‌گیرند، پیوند هیدروژنی برقرار کنند. صفحه‌های بتا نیز می‌توانند در یک یا چند قسمت از یک پروتئین ایجاد شوند و شیوه‌ی ایجادشان به این صورت است که بخشی از دنباله‌ی اسیدهای آمینه، شکلی مارپیچی را (در یک صفحه‌ی فرضی در فضا) به خود می‌گیرد و به این شکل، دو یا چند ردیف از اسیدهای آمینه در کنار هم قرار می‌گیرند. در این وضعیت، اسیدهای آمینه‌ی ردیف‌های کنارهم که در مجاورت یکدیگر قرار گرفته‌اند، با هم پیوند هیدروژنی برقرار می‌کنند.

ساختار سومین یک پروتئین، همان ساختار نهایی آن پروتئین در فضای سه‌بعدی می‌باشد. همه‌ی پیوندهای سخت و نرم درون پروتئینی، به همراه نیروهای مربوط به آب‌گریزی و دیگر عوامل مربوط به کاهش سطح انرژی آزاد پروتئین، در ساختار سوم آن در نظر گرفته می‌شوند. ساختار سوم یک پروتئین همان ساختاری است که پروتئین، در آب با دما و دیگر شرایط معمولی به خود می‌گیرد. چنین ساختاری به این که پروتئین در بدن موجودات زنده است یا نه، ربطی ندارد و در شرایط مشابه بدن نیز همان ساختار را به خود می‌گیرد.

ساختار چهارمین پروتئین‌ها مربوط به چگونگی قرار گرفتن چند پروتئین در کنار هم است که با چسبیدن به یکدیگر، در مجموع یک فعالیت خاص را انجام می‌دهند. نمونه‌ی معروف چنین ساختارهایی، هموگلوبین است.

در این مقاله و مقاله‌های مشابه تأکید بیشتر بر روی ساختارهای نخستین و سومین خواهد بود و از ساختارهای دومین و چهارمین صحبت خاصی نمی‌کنیم. ساده‌ترین دلیل اهمیت ساختار نخستین پروتئین-ها، چگونگی ساخت آن‌ها در بدن موجودات زنده است. توالی اسیدهای آمینه‌ی یک پروتئین، مبتنی بر کدهای سه‌تایی DNAها و RNAها مشخص می‌شود. در ریبوزوم‌های یک سلول که کارخانه‌های تولید پروتئین هستند، از روی رشته‌ی اطلاعات یک RNA، مشخص می‌شود که چه اسیدهای آمینه‌ای و با چه ترتیبی باید در کنار هم قرار بگیرند، و در نتیجه‌ی فعالیت «ترجمه» در ریبوزوم، دنباله‌ی اسیدهای آمینه با ترتیب مذکور تولید می‌شود. این دنباله‌ی اسیدهای آمینه، همان پروتئین تازه متولد شده است که در ابتدا در همان وضعیت ساختار نخستین می‌باشد. بعد از تولد پروتئین، فرآیند شکل‌گیری پروتئین آغاز می‌شود و پروتئین به وضعیت نهای‌اش در فضا می‌رسد که همان ساختار سومین آن است و به آن حالت طبیعی پروتئین هم گفته می‌شود.



شکل ۲-۳: ساختارهای چهارگانه پروتئین

همان‌طور که گفته شد، فرآیند شکل‌گیری پروتئین‌ها به این که در بدن موجودات زنده باشند یا در محیط مشابه آن، ربطی ندارد، و لذا این فرآیند را می‌توان صرفاً یک فرآیند فیزیکی و شیمیایی دانست. پروتئین‌ها تنها پس از رسیدن به شکل نهایی‌شان در فضا می‌توانند به فعالیت‌های خاص خود به خصوص در بدن موجودات زنده بپردازند و قبل از رسیدن به ساختار سومین‌شان نمی‌توانند وظیفه‌ی خود را انجام دهند. به همین علت، ساختار سومین پروتئین‌ها دارای اهمیت زیادی می‌باشد.

۳.مدل‌های تمام‌اتمی و درشت‌دانه

در مطالعه و بررسی پروتئین‌ها می‌توان از مدل‌های مختلفی از پروتئین‌ها استفاده کرد. در گروهی از این مدل‌سازی‌ها که به مدل‌های تمام‌اتمی معروف‌اند، کلیه‌ی جزئیات مربوط به همه‌ی اتم‌های پروتئین در نظر گرفته می‌شود. مشکل بزرگی که مدل‌های تمام‌اتمی به‌خصوص در شبیه‌سازی پروتئین‌ها دارند، این است که در این مدل‌ها، نه‌تنها جزئیات زیاد پروتئین‌ها هزینه‌ی شبیه‌سازی را بالا می‌برد، بلکه زمان وقوع تحولات اتمی (که از مرتبه‌ی 10^{-14} ثانیه است) در مقایسه با مدت زمان شبیه‌سازی (مثلاً شبیه‌سازی شکل‌گیری) پروتئین‌ها (که در حد میکروثانیه و بیشتر است) خیلی کوچک است و در نتیجه، تعداد قدم‌های لازم برای شبیه‌سازی، بسیار زیاد، و فرآیند شبیه‌سازی، بسیار کند می‌شود. مشکل بزرگ دیگری که مدل-های تمام‌اتمی دارند پیچیدگی زیاد آن‌ها است. تحلیل و پیش‌بینی رفتار در مدل‌های پیچیده‌ای مانند مدل-های تمام‌اتمی برای پروتئین‌ها بسیار سخت است.

چاره‌ای که برای حل مشکلات بالا اندیشیده شده، صرف‌نظر از توجه زیاد به جزئیات است. در مدل-های درشت‌دانه همین کار را انجام می‌دهند. هرچه از جزئیات مدل خود بکاهیم، هزینه‌ی مطالعه‌هایی مانند شبیه‌سازی پروتئین‌ها یا تحلیل و پیش‌بینی رفتارشان کاهش می‌یابد و در مقابل، از دقت نتایج نیز کاسته می‌شود. باید از آن بخشی از جزئیات صرف‌نظر کرد که به دقت کار لطمه‌ی کم‌تری بخورد. جزئیاتی برای حذف بهتر هستند که خیلی دست‌خوش تغییر نمی‌شوند و حذف آن‌ها تأثیر خاصی روی بقیه‌ی جاها نمی-گذارد. جزئیات مربوط به داخل اسیدهای آمینه گزینه‌ی مناسبی برای این کار است. اجزا و پیوندهای کوالانسی درون اسیدهای آمینه و زاویه‌هایی که ایجاد می‌کنند، به اندازه‌ی کافی پایدار و سخت هستند که بتوان اسیدهای آمینه را قطعاتی ثابت با خاصیت‌هایی ویژه فرض نمود و با این کار، جاهای دیگر هم دچار

مشکل خاصی نشوند. با همین ایده، در دسته‌ای از مدل‌های درشت‌دانه، هر اسید آمینه را با نماینده‌ای از آن جایگزین می‌کنند. این نماینده می‌تواند یک حجم بیضوی یا حتی تنها یک نقطه باشد که در محل اسید آمینه‌ی متناظرش قرار گرفته است.

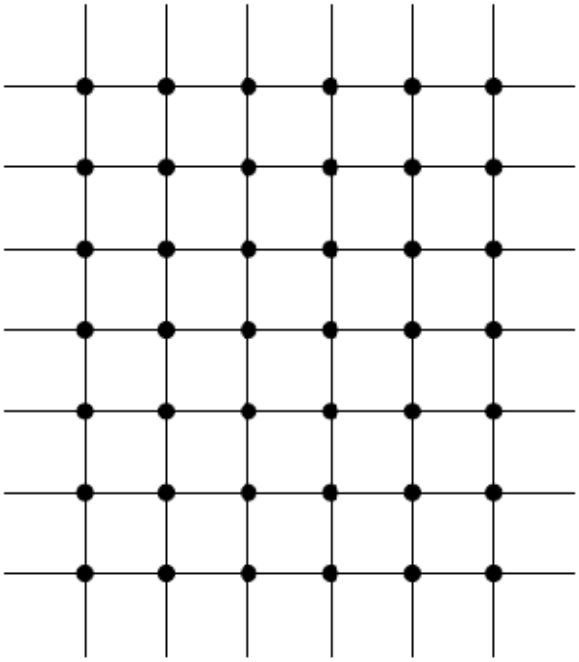
مدلی که در این مقاله و مقاله‌های مشابه برای پروتئین‌ها از آن استفاده می‌شود نیز مدلی درشت-دانه است. به ازای هر اسید آمینه، یک نقطه در فضا در نظر می‌گیریم که دارای خواصی ویژه (مانند قطبی بودن یا آب‌گریزی) است. محل این نقطه در فضا، مکان سابق اتم کربن مرکزی (C_α) اسید آمینه‌ی متناظرش می‌باشد. به ازای هر پیوند سخت (پپتیدی) بین دو اسید آمینه‌ی متوالی در پروتئین، نقطه‌های متناظر آن دو اسید آمینه را در مدل خود به هم وصل می‌کنیم، تا نتیجه، مسیری از نقطه‌ها در فضا باشد. پس در ادامه، منظور از پروتئین، به طور عمده همین مسیر می‌باشد که هم‌چنان می‌توان برای آن ساختارهای نخستین تا چهارمین را تصور نمود. با ساختارهای دومین و چهارمین که کار خاصی نداریم. ساختار نخستین، مانند قبل، رشته‌ی حروف نمایان‌گر اسیدهای آمینه است و ساختار سومین، همان مسیری است که در بالا تعریف کردیم.

برای ساده کردن مدل‌های پروتئینی، علاوه‌بر درشت‌دانه کردن آن‌ها، می‌توان فضایی را که پروتئین در آن، ساختار سومین خود را پیدا می‌کند، به نوعی ساده‌سازی نمود. یکی از روش‌های خوب برای ساده-سازی این فضای پیوسته، گسسته کردن آن است. برای این کار از مشبکه‌ها استفاده می‌شود که در قسمت بعد به معرفی آن‌ها می‌پردازیم.

۴. مشبکه‌ها

مشبکه، گرافی نسبتاً بزرگ (یا نامتناهی) با ساختاری هندسی (در فضای دوبعدی یا سه‌بعدی) است که با الگوهایی تکرارشونده ایجاد شده است. بهترین راه برای آشنایی با مشبکه‌ها، مشاهده‌ی نمونه‌های رایج آن‌ها است. در ادامه نمونه‌هایی از مشبکه‌ها را می‌بینیم.

اولین و رایج‌ترین مشبکه‌ی مورد استفاده، مشبکه‌های مربعی هستند که نمونه‌ی دوبعدی و نمونه‌ی سه‌بعدی آن‌ها را در شکل‌های ۲-۴ و ۲-۵ می‌بینید. به مشبکه‌ی مربعی سه‌بعدی، مشبکه‌ی مکعبی هم می‌گویند.

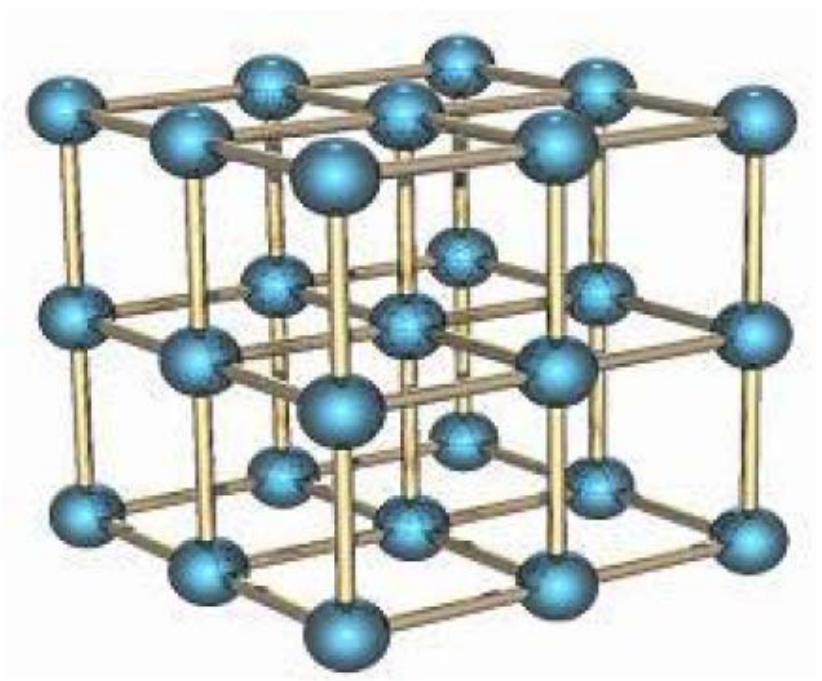


شکل ۲-۴: یک نمونه مشبکه‌ی مربعی دوبعدی

از ویژگی‌های مثبت مشبکه‌های مربعی، سادگی آن‌ها است. به بیان ریاضی، نقاط با مختصات صحیح، رأس‌ها را تشکیل می‌دهند و بین دو نقطه یال گذاشته می‌شود اگر و فقط اگر فاصله‌ی منهتن‌شان یک باشد. هم‌چنین زاویه بین یال‌های مجاور در مشبکه‌های مربعی ۹۰ درجه و ۱۸۰ درجه می‌باشد.

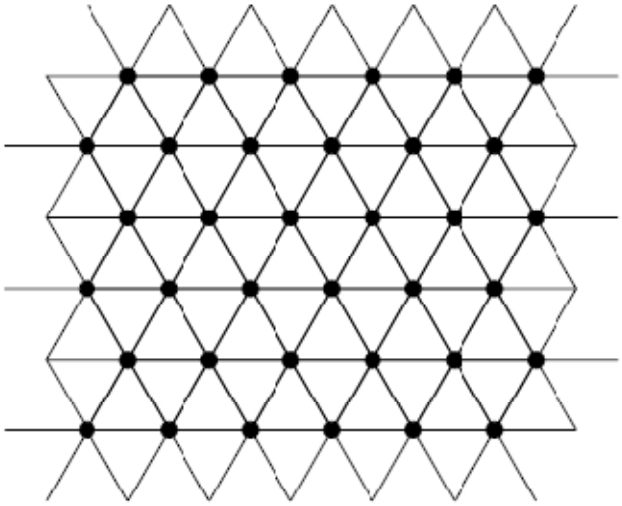
همه‌ی مشبکه‌های مربعی، گراف‌هایی دوبخشی هم هستند که این خصیصه نیز در بعضی جاها مساله‌ی مهمی می‌شود. یک نمونه از تأثیرات دوبخشی بودن مشبکه‌های مربعی را در بخش‌های بعدی مشاهده می‌کنید. مشبکه‌های مربعی را می‌توان به این شکل گسترش داد که در هر مربع (یا وجه)، قطرها را

نیز به مجموعه‌ی یال‌ها اضافه کرد. با این کار، مشبکه دیگر دوبخشی نخواهد بود ولی به علت تقاطع یال‌های آن، مشبکه‌ی مناسبی برای مدل‌سازی پروتئین‌ها نخواهد شد.

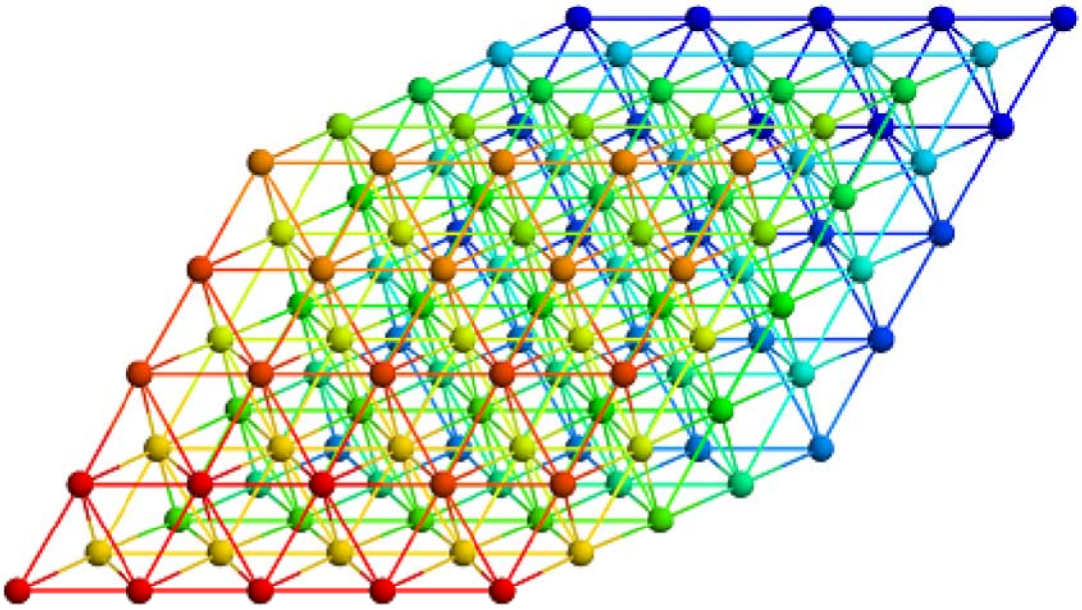


شکل ۲-۵: یک نمونه مشبکه‌ی مربعی سه-
بعدی

از مشبکه‌های دیگری که پس از مشبکه‌های مربعی دارای شهرت زیادی هستند، مشبکه‌های مثلثی هستند. نمونه‌هایی از مشبکه‌های مثلث دوبعدی و سه‌بعدی را به ترتیب در شکل‌های ۲-۶ و ۲-۷ می‌توان دید.

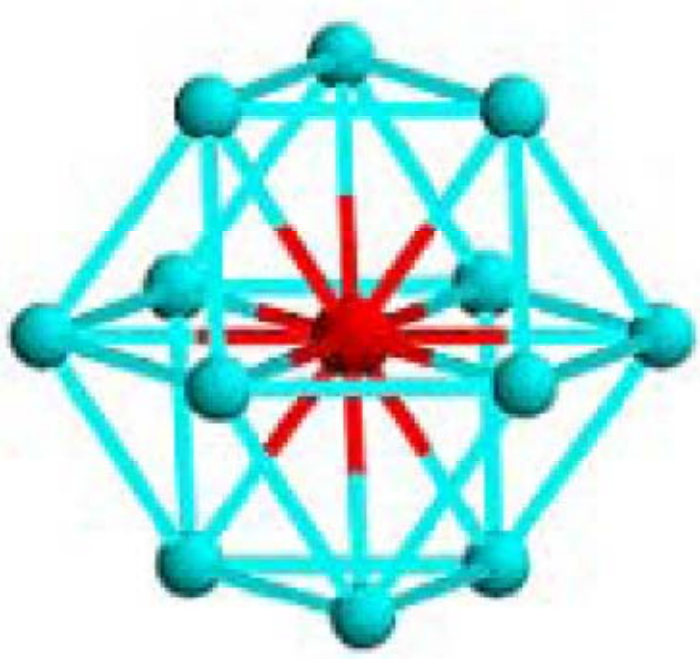


شکل ۲-۶: مشبکه‌ی مثلثی دوبعدی



شکل ۲-۷: مشبکه‌ی مثلث سه‌بعدی

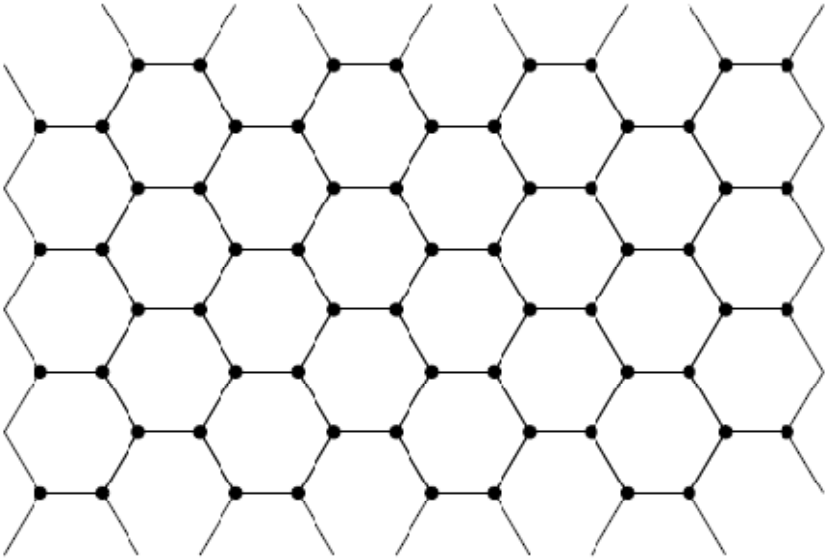
همان‌طور که در شکل ۲-۶ دیده می‌شود، در مشبکه‌های مثلثی دوبعدی، هر رأس، ۶ همسایه دارد. این تعداد، در مشبکه‌های مثلثی سه‌بعدی به ۱۲ همسایه می‌رسد. این ۱۲ همسایه برای یک رأس را در شکل ۲-۸ مشاهده می‌کنید. اگر یک رأس این مشبکه را در یک سطح افقی در نظر بگیریم، ۶ همسایه در همان سطح افقی، ۳ همسایه در سطح افقی بالاتر، و ۳ همسایه در سطح افقی پایین‌تر خواهد داشت.



شکل ۲-۸: همسایه‌های یک رأس در مشبکه-
ی مثلث سه‌بعدی

در مشبکه‌های مثلثی دوبعدی، یال‌های مجاور، با هم زاویه‌های ۶۰، ۱۲۰، یا ۱۸۰ درجه می‌سازند. تعداد انواع زاویه‌های مختلفی که یال‌های مجاور در مشبکه‌های مثلثی سه‌بعدی می‌سازند، خیلی بیشتر می-شود که علاوه بر ۶۰، ۱۲۰، یا ۱۸۰ درجه، زاویه‌ی ۹۰ درجه هم یکی از آن‌هاست. وجود زوایای ۹۰ درجه در مشبکه‌های مثلثی سه‌بعدی دارای اهمیت زیادی است، چون به کمک آن و دقت در ساختار این مشبکه-ها، می‌توان نشان داد که آن‌ها، مشبکه‌ی مربعی دوبعدی را هم شامل می‌شوند و آن را به عنوان زیرمجموعه‌ای از خود دارند.

مشبکه‌های دیگری هم طراحی شده‌اند که البته به اندازه‌ی موارد قبلی پرکاربرد نیستند. یکی از آن‌ها مشبکه‌ی لانه‌زنبوری است که یک مشبکه‌ی دوبعدی است و نمونه‌ای از آن را در شکل ۲-۹ می‌بینید.



شکل ۲-۹: مشبکه‌ی لانه زنبوری

۵. کاربرد مشبکه‌ها در مدل‌سازی پروتئین‌ها

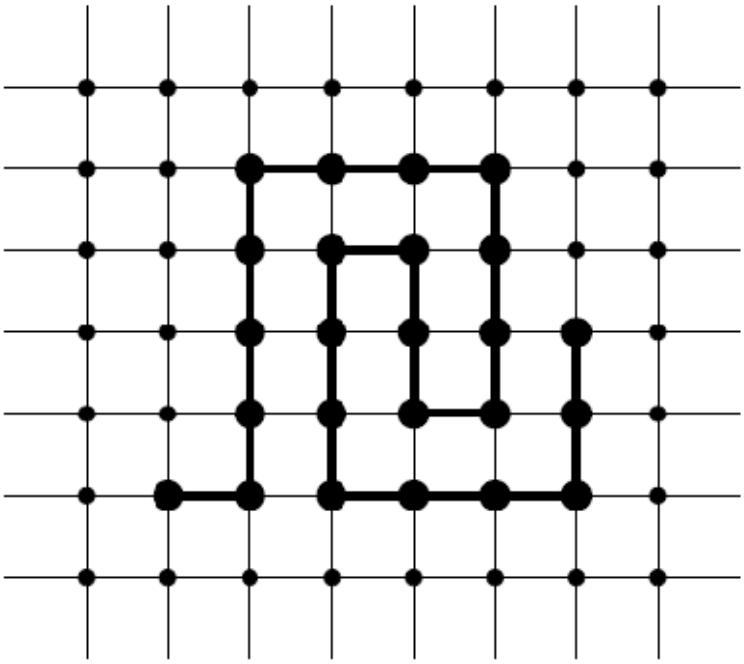
همان‌طور که گفته شد، برای ساده کردن مدل‌های پروتئینی، علاوه بر درشت‌دانه کردن آن‌ها، می‌توان فضایی را که پروتئین در آن، ساختار سومین خود را پیدا می‌کند، به نوعی ساده‌سازی نمود، و یکی از روش‌های خوب برای ساده‌سازی این فضای پیوسته، گسسته کردن آن است. مشبکه‌ها گزینه‌های مناسبی برای گسسته کردن فضایی هستند که پروتئین می‌تواند در آن فضا ساختار طبیعی و تاشده‌ی خود را پیدا کند. دانشمندان در مقالات مختلفی مانند [۱۵]، [۱۶] و [۱۷] در زمینه‌ی استفاده از مشبکه‌ها برای مدل‌سازی فضای ساختار پروتئین صحبت‌های زیادی کرده‌اند.

نشان داده شده که فاصله‌ی دو اسید آمینه‌ی متوالی در یک پروتئین، معمولاً حول مقدار ثابت $3.8\ \dot{A}$ در نوسان است و تغییر کمی دارد. هم‌چنین به طور معمول، زاویه‌ای که یک اسید آمینه با دو اسید آمینه‌ی مجاورش در پروتئین می‌سازد، مقادیری (محدوده‌هایی) تقریباً مشخص دارد. پس شکل تاشده‌ی نهایی و ساختار سوم پروتئین‌ها خیلی هم درجه‌ی آزادی زیادی ندارد. چنین خواصی ایده‌ی استفاده از مشبکه‌ها را برای مدل‌سازی فضایی که ساختار سومین پروتئین در آن تعریف می‌شود، توجیح می‌کند.

حال، ساختار سومین یک پروتئین در یک مشبکه را معرفی می‌کنیم. همان‌طور که گفته شد، مدل ساده‌شده‌ی ما از پروتئین‌ها دنباله‌ای از نقاط در فضا می‌باشد که این نقاط، نماینده‌ی اسیدهای آمینه‌ی پروتئین هستند. در فضای مشبکه، این نقطه‌ها فقط می‌توانند برروی رأس‌های مشبکه قرار بگیرند. در فضای پیوسته‌ی طبیعی، دو اسید آمینه‌ی متمایز نباید برروی یک رأس مشبکه قرار بگیرند، مگر این که عکس آن را صراحتاً گفته باشیم.

همان گونه که گفته شده، در فضای پیوسته‌ی طبیعی، فاصله‌ی اسیدهای آمینه‌ی متوالی، مقداری تقریباً ثابت و مشخص است. بالطبع باید ضابطه‌ی مشابهی را در فضای مشبکه‌ها هم داشته باشیم. ضابطه‌ی مذکور این است که دو اسید آمینه‌ی متوالی در پروتئین باید برروی دو رأس مجاور در مشبکه قرار بگیرند.

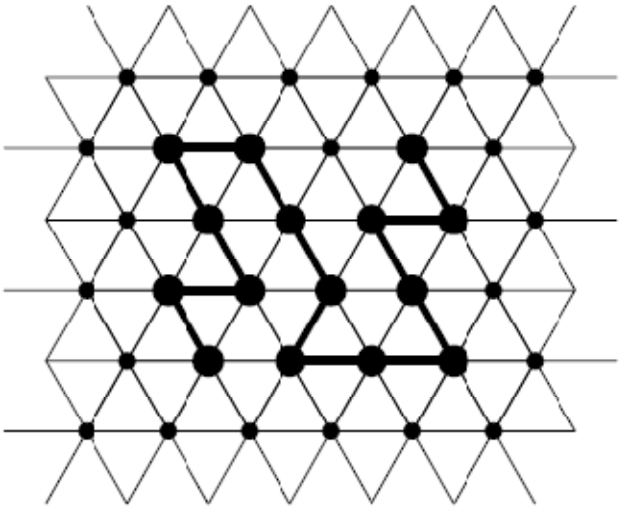
در شکل‌های ۲-۱۰ و ۲-۱۱ نمونه‌هایی را از پروتئین مدل‌شده در مشبکه مشاهده می‌کنید. دایره-های بزرگ‌تر، نمایان‌گر اسیدهای آمینه هستند و خطوط پررنگ‌تر پیوندهای سخت بین اسیدهای آمینه را نشان می‌دهند. همان‌طور که مشاهده می‌کنید، پروتئین مدل‌شده در یک مشبکه، یک مسیر در آن گراف می‌شود. وقتی یک پروتئین را در یک مشبکه مدل می‌کنیم، غیر از یال‌های مربوط به پیوندهای سخت، یال-های دیگری نیز در مشبکه پیدا می‌شوند که دو سرشان اسید آمینه است. مثلاً در شکل ۲-۱۰، ۱۴ یال وجود دارد که دو سرشان اسید آمینه است ولی بین آن دو اسید آمینه پیوند سختی نیست (یعنی در ساختار نخستین پروتئین در مکان‌های متوالی قرار نگرفته‌اند). این یال‌ها مجاورت اسیدهای آمینه را در فضای مشبکه تعریف می‌کنند. می‌گوییم دو اسید آمینه در مشبکه با هم «تماس مکانی» دارند اگر در مشبکه بین آن دو یال وجود داشته باشد و در عین حال پیوند سختی هم با هم نداشته باشند.



شکل ۲-۱۰: یک نمونه پروتئین مدل‌شده در
مشبکه‌ی مربعی دوبعدی

اکنون می‌توانیم مدل HP را تعریف کنیم. مدل HP در عین حال که خیلی ساده به نظر می‌رسد، خیلی از خواص پروتئین‌ها را به خوبی نشان می‌دهد. مقاله‌ی [۱۵] در همین زمینه صحبت کرده است.

مدل HP پروتئین‌ها را مشابه آن‌چه تعریف کردیم یک مسیر برروی مشبکه در نظر می‌گیرد. در عین حال، بحث آب‌گریزی و آب‌دوستی اسیدهای آمینه را نیز مطرح می‌کند. نشان داده شده مهم‌ترین عامل در تعیین ساختار سومین پروتئین‌ها آب‌گریزی و آب‌دوستی اسیدهای آمینه است. پس به جای ۲۲ نوع اسید آمینه، که با ۲۲ حرف نشان داده می‌شوند، تنها از دو حرف H و P استفاده می‌کنیم: از حرف H برای اسیدهای آمینه‌ی آب‌گریز، و از حرف P برای اسیدهای آمینه‌ی آب‌دوست که قطبی هستند. لذا ساختار نخستین پروتئین رشته‌ای از حروف H و P می‌شود و ساختار سومین آن هم مسیری برروی مشبکه می‌شود که رأس‌های آن یکی از دو رنگ H و P را دارند.



شکل ۲-۱۱: یک نمونه پروتئین مدل‌شده در مشبکه مثلثی دوبعدی

در مدل HP، انرژی آزاد پروتئین که سیستم تمایل به کاهش آن را دارد، از روی مکان‌های رأس‌های H و P تعیین می‌شود. فرض بر این است که جایگاه‌هایی از مشبکه که پر نشده‌اند (رأس‌هایی که اسید آمینه‌ای روی آن‌ها نیست)، با مولکول آب پر شده‌اند. مولکول‌های آب نیز قطبی هستند و مانند رأس‌های P می‌باشند. مهم‌ترین عامل در انرژی آزاد سیستم مجاورت رأس‌های H با رأس‌های P و خالی است. هرچه تعداد یال‌هایی از مشبکه که یک سرشان H و سر دیگرشان P یا خالی است، بیشتر باشد، انرژی آزاد سیستم هم بیشتر است. برای راحتی بیشتر، کمیت دیگری را معرفی می‌کنند که ساده‌تر است: تعداد یال‌هایی از مشبکه که هر دو سرشان H است. این کمیت به «اتصالات $H-H$» معروف است. با توجه به این که رشد اتصالات $H-H$، قرینه‌ی رشد انرژی آزاد سیستم است، به جای استفاده از مفهوم انرژی آزاد سیستم که می‌خواهیم کمینه شود، بیشتر از اتصالات $H-H$ که می‌خواهیم بیشینه شود استفاده می‌گردد.

در این مقاله و مقاله‌های مشابه، در مسائلی که با مدل HP سروکار دارند، از اتصالات $H-H$ استفاده می‌کنیم. منتها به جای استفاده از عبارت «اتصالات $H-H$»، جاهایی نیز از لفظ «امتیاز» استفاده

می‌شود. مثلاً گفته می‌شود پروتئین‌ها در مدل HP تمایل دارند ساختاری را به خود بگیرند که امتیازشان را بیشینه کند.

برای یک رشته پروتئینی که ساختار نخستین آن را در اختیار داریم، در مدل HP، تاشدگی‌های (ساختارهای) مختلفی را می‌توان در نظرگرفت. هر یک از این تاشدگی‌ها امتیازی را دارد. هرچه این امتیاز برای یک تاشدگی بیشتر باشد، آن تاشدگی بهتر است. پس «تاشدگی بهینه» برای (ساختار نخستین) یک رشته‌ی پروتئینی در مدل HP را می‌توان تاشدگی‌ای از آن تعریف نمود که بیشترین امتیاز را دارد.

اگر یک تاشدگی برای یک رشته‌ی پروتئینی بهینه باشد، با انتقال، دوران ۹۰ درجه و ۱۸۰ درجه، و یا تقارن آن، تاشدگی‌های دیگری به دست می‌آیند که عموماً با تاشدگی اولیه متفارت‌اند، ولی امتیاز آن‌ها با امتیاز تاشدگی اولیه یکی است و در نتیجه، تاشدگی‌های جدید نیز بهینه هستند. به همین علت، برای جلوگیری از بروز پاره‌ای مشکلات، لازم است تاشدگی‌هایی را که با چنین اعمالی به هم تبدیل می‌شوند، یکی بدانیم.

فصل ۳

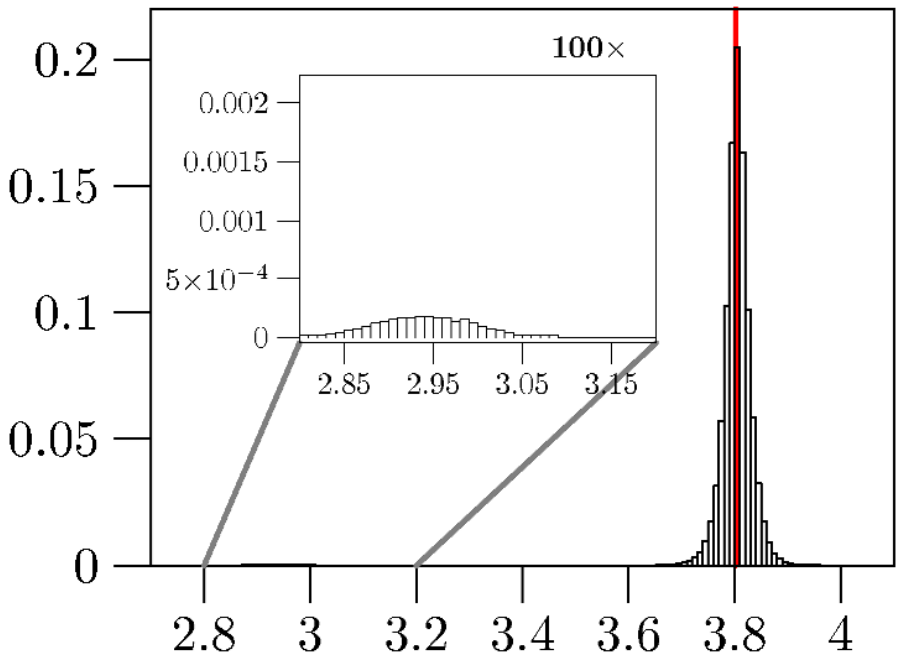
نتایج بدست آمده

هدف ما تعیین کردن مشبکه‌هایی است که بهترین تقریب را در نشاندن بلورهای پروتئین‌های اساسی به ما می‌دهند. ما به این‌ها مشبکه‌های ایده‌آل می‌گوییم. مقاله ما به دو قسمت تقسیم می‌شود، قسمت اول مربوط به یک مطالعه بر روی زیرمجموعه‌ای از پایگاه‌داده پروتئین‌ها است برای این که بتوانیم خواصی برای مشبکه‌های ایده‌آل پیدا کنیم. در قسمت دوم مجموعه‌ای از مشبکه‌های کاندیدا ساخته و تولید می‌شوند، و توانایی آن‌ها در نشاندن پروتئین‌های پایگاه‌داده مقایسه می‌شود. این روش برای مقایسه مشبکه-های در مطالعات قبلی هم آمده است. ([۱]، [۴]، [۵] و [۱۹]) برای نشاندن پروتئین‌ها بر روی مشبکه‌ها از

الگوریتم پارک و لویت [۱] استفاده می‌کنیم. همچنین فرض می‌کنیم پروتئین‌ها در ساختار دومین خود به سر می‌برند و شامل دو مجموعه زاویه سازگار ϕ و ψ هستند. و در نهایت به سراغ مشبکه‌هایی می‌رویم که تقریب خوبی برای پروتئین‌های فرد ساختار دومین می‌دهند، همان طور که برای اکثر پروتئین‌ها جواب خوبی می‌دادند.

۱. مشخص کردن خواص مشبکه‌ها

برای این کار ما به سراغ یک زیرمجموعه ۳۷۰۴ عضوی از پایگاه‌داده ساختار پروتئین‌ها رفتیم. فاصله میان اسیدهای آمینه معمولاً برابر $\dot{A}$ ۳.۸ می‌باشد. ([۶] و [۷]) این مقدار با بررسی فاصله‌های اسیدهای آمینه متوالی در زیرمجموعه ما از پروتئین‌ها تأیید شده است. نمودار پراکندگی همه فاصله‌ها یک قله در مقدار $\dot{A}$ ۳.۸ نشان می‌دهد. این اطلاعات را می‌توانید در نمودار ۳-۱ ملاحظه کنید:



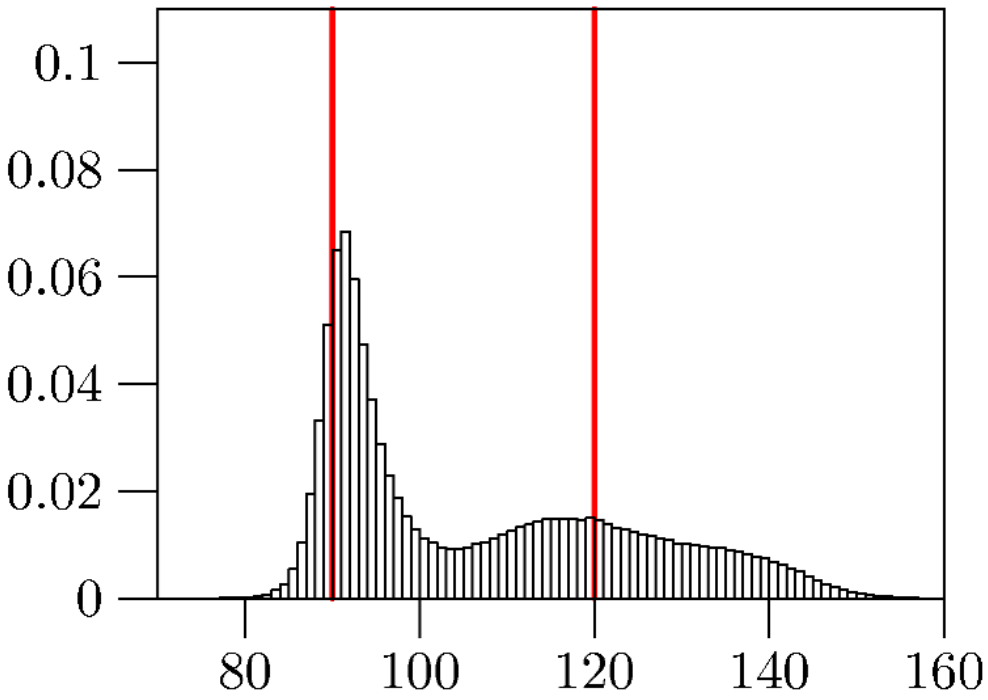
نمودار ۳-۱: فاصله میان اسیدهای آمینه متوالی

تعداد جفت‌ها: ۱۲۸۵۹۴۳ ، میانگین: ۳.۸۰۲۱۷

از این رو، ما فاصله میان دو یال از رشته یک پروتئین را برابر با $\dot{A}$ ۳.۸ فرض می‌کنیم.

بعد از این ما فاصله میان اسیدهای آمینه غیرمتوالی را اندازه‌گیری کردیم تا ببینیم آیا مقادیری کمتر از $\dot{A}$ ۳.۸ پیدا خواهند شد یا خیر. اگر فاصله‌هایی این چنینی پیدا شود یعنی مشبکه ایده‌آل ما نباید لزوماً ساختار منظمی داشته باشد و علاوه‌براین رئوس غیرمجاور در مشبکه می‌توانند نزدیک‌تر از رئوس مجاور باشند. نتایج محاسبه ما نشان داد که از میان ۱۶۵ میلیون جفت اسید آمینه بررسی شده، کمترین فاصله غیرمجاور برابر با $\dot{A}$ ۳.۰۶ می‌باشد و تنها ۱۰ فاصله کمتر از $\dot{A}$ ۳.۵، و تنها ۱۹۹۹ فاصله (برابر با 0.0012%) کمتر از $\dot{A}$ ۳.۸ می‌باشند. این نتایج ما را بر آن داشتند تا فرض کنیم که کمترین فاصله میان دو اسید آمینه متوالی برابر با $\dot{A}$ ۳.۸ است.

برای این که زاویه موجود در رئوس مشبکه ایده‌آل را پیدا کنیم، ما می‌توانیم زوایای میان سه اتم C_αی متوالی را در پایگاه داده پروتئین‌ها بررسی کنیم. این مطالعات بر پایه مطالعات قبلی که بر روی ۱۳ پایگاه داده انجام شده بود، می‌باشد. [۸] نمودار ۳-۲ اطلاعات مربوط به این بخش را نشان می‌دهد:



نمودار ۳-۲: پراکندگی زوایای C_α، خط‌های قرمز نشان‌دهنده ۹۰ و ۱۲۰ هستند.

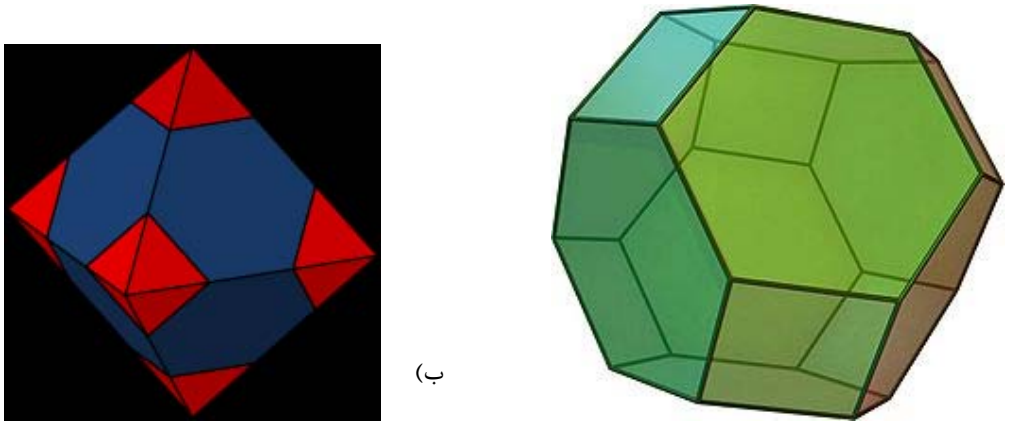
یکی دیگر از کارهایی که در این زمینه انجام شده است [۹]، نشان داده است که از میان مشبکه-های تصادفی و مشبکه‌های دوره‌ای، مشبکه‌های دوره‌ای (تناوبی) به مشبکه‌های ایده‌آل نزدیک‌تر می‌باشند. از این رو ما از این به بعد برای پیدا کردن مشبکه‌های نزدیک به مشبکه ایده‌آل به سراغ مشبکه‌هایی می-رویم که دارای خاصیت‌های زیر باشند:

۱- دارای طول یکنواخت $\dot{A}$ ۳.۸ باشند.
۲- کمترین فاصله میان هر دو رأس آن برابر با $\dot{A}$ ۳.۸ باشد.
۳- دارای زاویه‌های ۹۰ درجه و ۱۲۰ درجه باشد.
۴- ساختار آن تناوبی باشد.

۲. مقایسه مشبکه‌های کاندیدا

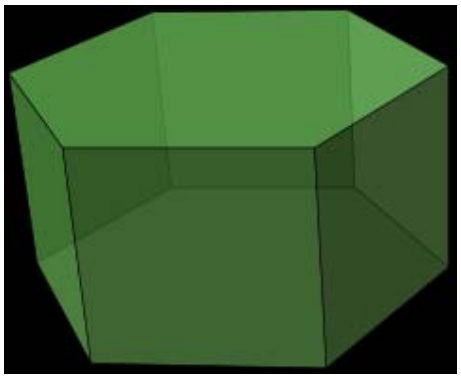
در این بخش ما چندین مشبکه معرفی می‌کنیم که خاصیت‌های گفته شده در بالا را دارا باشند. اولین لیست از مشبکه‌های کاندیدا مشبکه‌هایی هستند که از ترکیب‌های پوشاننده‌ی فضا از چندوجهی‌های منتظم تولید شده‌اند. با دقت به این نکته که هنگامی که می‌خواهیم یک وجه از یک چندوجهی را به یک وجه از چندوجهی دیگر متصل کنیم، این دو وجه باید کاملاً روی یکدیگر قرار بگیرند، و کل فضا باید به طور کامل پوشانده شود. از این رو ما مشبکه‌ها را به دو دسته تقسیم می‌کنیم:

مشبکه‌هایی با پوشش یکتا: هشت‌وجهی ناقص (فضای پوشیده شده با هشت‌وجهی ناقص)، منشور شش-گوش (فضای پوشیده شده با منشور شش‌گوش)، مکعب ساده (مشبکه مکعبی) و هشت‌وجهی مکعبی (فضای پوشیده شده با هشت‌وجهی مکعبی و هشت‌وجهی).

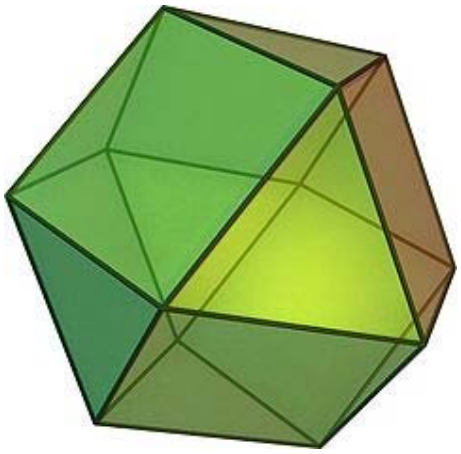


ب)

شکل ۳-۱: الف) هشت وجهی ناقص
ب) طریقه بدست آمدن هشت وجهی ناقص

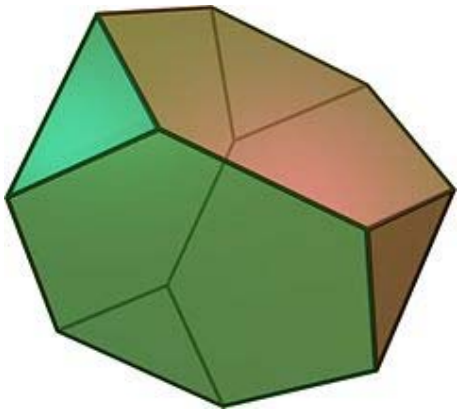


شکل ۳-۲: منشور شش گوش



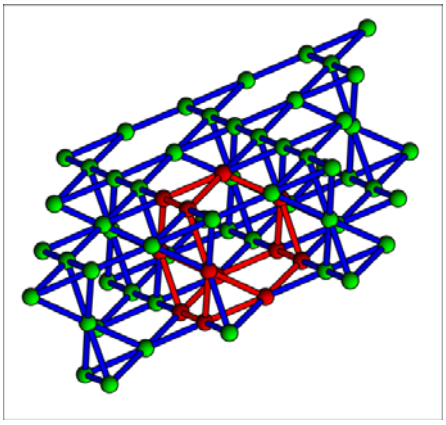
شکل ۳-۳: هشت وجهی مکعبی

مشبکه‌هایی با چند پوشش: چهاروجهی ناقص (فضای پوشیده شده با چهاروجهی ناقص و چهاروجهی)



شکل ۳-۴: چهاروجهی ناقص

روش‌های زیادی برای پوشاندن فضا با چهاروجهی ناقص وجود دارد، ما در این جا با قرار دادن وجوه مانده از چهاروجهی روی هم، هم چنین قرار دادن وجوه مثلثی آن روی هم، مشبکه را تولید کردیم. (شکل ۳-۵) نکته قابل توجه این است که این‌گونه ساختن مشبکه دو شرط تناوبی بودن و منظم بودن مشبکه را تضمین می‌کند.



شکل ۳-۵: طریقه پوشاندن فضا با چهاروجهی ناقص

یک مشبکه دیگر برپایه چهاروجهی ساخته می‌شود که به مشبکه مثلثی سه‌بعدی معروف است. ولی این مشبکه فضاپرکن نیست و به علت سختی محاسبات مربوط به بررسی نشاندن پروتئین‌ها روی این مشبکه، ما این مشبکه را در این مقاله وارد نمی‌کنیم.

در نهایت ما به سراغ سه مشبکه که بر روی مشبکه مکعبی ساخته می‌شوند و اخیراً در این ضمینه درباره آن‌ها صحبت‌هایی شده، می‌رویم. این سه مشبکه عبارت‌اند از: FCC,s-FCC,e-FCC. رئوس یک مشبکه مکعبی را می‌توانیم با مختصات صحیح به صورت (x, y, z) نشان دهیم با این فرض که طوری مقیاس شده‌اند که فاصله میان هر دو رأس از مشبکه برابر با $\dot{A}$ ۳.۸ باشد. این سه مشبکه جدید شامل آن نقاطی از دستگاه مختصات هستند که مجموع مؤلفه‌های آن‌ها عددی زوج باشد، برای مثال سه بردار $v_1 = (\pm1, \pm1, 0), v_2 = (\pm2, 0, 0), v_3 = (\pm2, \pm1, \pm1)$. یال‌های این سه مشبکه نیز به صورت زیر تعریف می‌شوند:

بین دو رأس از صفحه مختصات یال قرار دارد اگر اختلاف بین مؤلفه‌های آن‌ها برابر باشد با:

بردار v_1 یا یکی از جایگشت‌های آن، برای مشبکه Face Centered Cubic (FCC).

بردار v_1 یا v_2، یا یکی از جایگشت‌های آن‌ها، برای مشبکه side-FCC (s-FCC).

بردار v_1، v_2 یا v_3، یا یکی از جایگشت‌های آن‌ها، برای مشبکه extended-FCC (e-FCC).

دو مشبکه s-FCC و e-FCC دارای یال‌هایی با طول‌های متفاوت‌اند و این با قانون اول تولید مشبکه‌ها در تناقض است ولی به علت قوی بودن بقیه فاکتورها و همچنین گرفتن نتیجه‌ای مناسب ما این دو مشبکه را حذف نکردیم. در شبکه s-FCC طول یال‌های بدست آمده از v_1 (مقدار ۶۷٪ یال‌ها) برابر $\dot{A}$ ۳.۸ است. و طول یال‌های بدست آمده از v_2 (مقدار ۳۳٪ یال‌ها) برابر $\dot{A}$ ۵.۴ است. در شبکه e-FCC طول یال‌های بدست آمده از v_2 (مقدار ۱۴٪ یال‌ها) برابر $\dot{A}$ ۳.۸ است. و طول یال‌های بدست آمده از v_1 (مقدار ۲۹٪ یال‌ها) برابر با $\dot{A}$ ۲.۷ و طول یال‌های بدست آمده از v_3 (مقدار ۵۷٪ یال‌ها) برابر $\dot{A}$ ۴.۷ می‌باشد.

برای این که کدام یک از مشبکه‌ها برای نشاندن پروتئین‌ها مناسب‌تر است، ما مجموعه پروتئین‌های پایگاه‌داده را بر روی این مشبکه‌ها می‌نشانیم. اختلاف میان مختصات پروتئین نشانده شده در مشبکه و مختصات واقعی آن به سه طریق محاسبه می‌شود:

c-RMS: coordinate Root Mean Square

d-RMS: relative all-to-all Root Mean Square

a-RMS: angle Root Mean Square

هر کدام از پروتئین‌ها را سه بار بر روی مشبکه‌ها می‌نشانیم، هر بار یکی از اندازه‌های بالا را کمینه می‌کنیم. در نهایت با میانگین‌گیری بر روی همه آن‌ها، نتایج را در جدول زیر مشاهده می‌کنید:

	درجه	c-RMS	d-RMS	a-RMS
هشت‌وجهی ناقص	4	4.4756	3.1293	13.0982
منشور شش‌گوش	5	3.2833	2.2952	10.0313
چهاروجهی ناقص	6	3.1076	2.2381	19.9030
مکعب ساده	6	2.7579	1.9705	21.1005
هشت‌وجهی مکعبی	8	2.1909	1.5961	8.3526
FCC	12	1.6728	1.2431	8.3346
s-FCC	18	1.5311	1.1528	6.2022
e-FCC	42	1.1029	0.8475	2.5700

جدول ۳-۱: میانگین اندازه‌های c-RMS، d-RMS و a-RMS
برای مشبکه‌های مختلف. مرتب شده بر حسب درجه آن‌ها

علاوه بر نتایج مشاهده شده در جدول فوق، مشاهده شد که مقادیر c-RMS و d-RMS با افزایش درجه مشبکه، کاهش می‌یابند. یکی از نتایج قابل بررسی مقایسه نتایج مشبکه مکعبی و مشبکه حاصل از چهاروجهی ناقص می‌باشد (هر دو از درجه ۶). هر دو این مشبکه‌ها تقریباً مقدار c-RMS و d-RMS یکسان دارند وقتی که مسیرهایی که خود را قطع می‌کنند را حذف نکنیم (یعنی بیشتر از یک C_α بتواند در یک رأس قرار گیرد). این یک تأیید بر کارهای قبلی است [۱] که مبتنی بر این مفهوم است که اندازه c-RMS وابستگی زیادی به درجه مشبکه دارد. وقتی که مسیرهایی که خود را قطع می‌کنند را حذف کنیم، مشبکه مکعبی ساده ۱۹٪ بهتر از مشبکه چهاروجهی ناقص است. از آن جایی که مشبکه مکعبی دارای زاویه‌های ۹۰ و ۱۲۰ درجه‌ی بیشتری نسبت به مشبکه حاصل از چهاروجهی ناقص است، حدس ما بر این است که این همان دلیل بهتر بودن نسبی مشبکه مکعبی است برای نشاندن پروتئین.

از این تحلیل ما نتیجه می‌گیریم که از میان مشبکه‌هایی که شرایط چهارگانه گفته شده را ارضاء می‌کنند، مشبکه FCC بهترین نتیجه را می‌دهد. اما اگر شرط اول را حذف کنیم، مشبکه e-FCC بهترین نتیجه را می‌دهد. در آخر تحقیق ما نتیجه می‌دهد که دو مشبکه، مشبکه‌های ایده‌آل خواهند بود: FCC و e-FCC.

۳.بحث و کارهای آینده

با توجه به تحقیقی که ما در این مقاله انجام دادیم، به این نتیجه رسیدیم که مشبکه‌هایی که طول یال‌های آن‌ها یکنواخت و برابر با $\dot{A}$ ۳.۸ باشد، کمترین فاصله میان دو رأس از آن برابر با $\dot{A}$ ۳.۸ باشد، متناوبی باشند، و دارای زاویه‌های ۹۰ و ۱۲۰ درجه باشند، بهترین نتیجه را برای نشاندن پروتئین‌ها می‌دهند. مشبکه FCC بهترین مشبکه از میان مشبکه‌هاست که ضوابط مشبکه ایده‌آل را رعایت می‌کند. ما همچنین مشبکه‌های s-FCC و e-FCC را نیز مورد آزمایش قرار دادیم که این‌ها طول یال‌های یکنواخت نداشتند.

این محاسبات را می‌توان به روش جالبی ادامه داد، به این صورت که مشبکه‌های درجه بالاتری را مورد آزمایش قرار دهیم که شامل خصوصیات مشبکه ایده‌آل ما تا به این جا می‌باشند، تا جایی که

خصوصیت دیگری غیر از درجه نقش مهم‌تری را در نشاندن پروتئین‌ها ایفا کند. هرچند برای تحقیقات آینده ما پیشنهاد می‌کنیم به سراغ یک مشبکه ساده بروند که بتوان یک تقریب سریع از یک پروتئین به ما بدهد و پیچیدگی محاسبات آن کم باشد. یعنی در حالت کلی بهترین مشبکه، مشبکه‌ای است که (همانند مشبکه-های مکعبی و چهاروجهی) هم ساده باشد، هم زاویه‌های به درد بخوری داشته باشد، که علاوه بر مناسب بودن برای نشاندن پروتئین‌ها، بتوان با محاسبات کم و در زمان کوتاه نتایج نشاندن پروتئین را روی آن بررسی کرد.

این نتایج نماینده ساختارهای پروتئینی هستند که امروزه در پایگاه‌داده پروتئین‌ها موجودند، و این اطلاعات از طریق بلورشناسی توسط اشعه ایکس شناسایی شده‌اند. اگرچه این حائز اهمیت است که پروتئین‌ها ساختارهایی سخت و شکننده نیستند. هم‌چنین این اطلاعات که از پایگاه‌داده پروتئین‌ها گرفته شده و مورد استفاده قرار می‌گیرد، متمایل شده به سمت آن دسته پروتئین‌هایی است که ساختار آن‌ها راحت‌تر قابل شناسایی توسط ابزارهای آزمایشگاهی امروزی است.

این کار جالب خواهد بود که پروتئین‌ها را با توجه به خواص آن‌ها دسته‌بندی کنیم و خواص هر دسته را بررسی کنیم و برای هر دسته مشبکه‌های مناسب برای آن دسته را بیابیم. در این مقاله ما پروتئین-هایی را که با آن‌ها کار می‌کردیم به آن دسته پروتئین‌هایی محدود کردیم که در شناسایی آن‌ها از پراش اشعه ایکس استفاده می‌کنند. در ادامه ما کارمان را به پروتئین‌هایی محدود کردیم که به راحتی قابلیت تبلور داشتند، به این علت که این کاملاً منطقی است که مشبکه‌های ایده‌آل ما مشبکه‌هایی ایده‌آل برای نشاندن پروتئین‌هایی باشند که به راحتی قابلیت تبلور دارند: پروتئین‌های کروی، با تعداد زیاد اجزای سخت و محکم، که مشخصات شیمیایی آن‌ها بیشتر از همه با خالص‌سازی و تبلور در تکنیک‌های جدید سازگار است. این ممکن است که انواع مختلف پروتئین‌ها، مشبکه‌های ایده‌آل مختلفی را پذیرا باشند.

فصل ٤

اصول و روش‌های به کار گرفته شده

۱. پایگاه‌داده پروتئین‌ها

یک زیرمجموعه از فایل‌های پایگاه‌داده پروتئین‌ها انتخاب شده است که شبیه به مجموعه‌ای است که قبلاً در [۱۰] کارپلوس استفاده کرده است. فایل‌ها از پایگاه داده پروتئین‌ها استخراج شده‌اند. (نسخه ۱۳ آوریل ۲۰۰۴) همچنین پروتئین‌هایی انتخاب شده‌اند که در خصوصیات آن‌ها پراش اشعه ایکس قید شده

باشد. ما فایل‌هایی که طول زنجیر آن‌ها کمتر مساوی ۲ بود را حذف کردیم، و در نهایت ۳۷۰۴ فایل پایگاه-داده باقی ماندند.

هر فایل پایگاه‌داده در مجموعه ما بعد از تجزیه و استجراج مختصات اتم C_α، نوع اسید آمینه آن و شماره رشته از آن فایل بدست آمده‌اند. همچنین زاویه C_α تعریف می‌شود زاویه‌ای است که میان سه اتم C_αی متوالی تولید می‌شود. با این وجود تعداد زاویه‌های بدست آمده برابر با ۱۰۲۵۲۸۵ شدند.

۲. ارزیابی مشبکه‌ها

همان‌طور که پیش‌تر نیز توضیح دادیم، مشبکه‌ها استفاده شدند که پروتئین‌های پایگاه‌داده را بر روی آن‌ها بنشانیم. پروتئین‌ها به این طریق بر روی مشبکه می‌نشینند که اسیدهای آمینه مجاور بر روی رئوس مجاور از مشبکه قرار گیرند. این روند برای هر جفت پروتئین-مشبکه سه بار انجام گرفت، هر بار برای کمینه کردن یکی از اندازه‌ها میانگین (RMS metrics). به یک نشاندن ۱۰۰٪ موفق می‌گوییم اگر نتیجه آن رسیدن به RMS برابر با ۰ باشد. سپس مشبکه‌ها با توجه به مقدار RMSشان مرتب می‌شوند که این نمایانگر نزدیکی به ایده‌آل بودن برای آن‌هاست. در نهایت بهترین مشبکه، مشبکه‌ای است که همه RMSها برای آن‌ها مینیمم شود. مقادیر RMS که در تحقیق ما استفاده شدند، عبارتند از:

- c-RMS: (coordinate root mean square deviation) بیانگر این است که چه میزان دو شیء بر یکدیگر منطبق هستند.

$$\text{c-RMS}=\sqrt{\frac{\sum_{i=1}^{n}|a_i-b_i|^2}{n}}$$

که در رابطه فوق a_i ها مختصات خوانده شده از پروتئین‌ها توسط اشعه ایکس هستند و b_i ها مختصات نقطه‌ای از شبکه است که a_i در آن نشانده می‌شود.

- d-RMS: (distance root mean square deviation) بیانگر این است که شکل یک شیء چقدر خوب حفظ می‌شود. (دقت کنید که اگر مثلاً یک شیء را به صورت آینه‌ای برگردانیم، در واقع شکل آن تغییری نکرده است، و از این رو d-RMS برای این دو شکل ۰ می‌باشد، ولی c-

RMS برای این دو شکل غیرصفر خواهد بود چون با هیچ دورانی نمی‌توان آن‌ها را بر یکدیگر منطبق کرد.)

$$\text{d-RMS}= \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(|a_i-a_j|-|b_i-b_j|\right)^2}{n(n-1)/2}}$$

که در رابطه فوق a_i و a_j مختصات دو جفت نقطه از مختصات اصلی پروتئین، و b_i و b_j مختصات نقاط متناظر آن‌ها در مشبکه می‌باشد.

- a-RMS: (angle root mean square distance) بیانگر این است که چه میزان زاویه‌های میان اتم‌های مجاور C_α باقی نگه‌داشته شده‌اند. (دقت کنید که اندازه a-RMS نیز در تبدیل آینه‌ای تغییری نمی‌کند.)

a-RMS دقیقاً شبیه به c-RMS تعریف می‌شود، با این تفاوت که در آن a_i ها و b_i ها زاویه‌های میان اتم‌های مجاور C_α می‌باشد.

۳.الگوریتم نشاندن پروتئین بر روی مشبکه

الگوریتم‌های زیادی برای نشاندن پروتئین‌ها بر روی مشبکه‌ها مطرح شده‌اند. [۱]، [۱۹]، [۱۱]، [۱۲] و [۱۳]. در این جا ما از الگوریتم مطرح شده توسط پارک و لویت برای این کار استفاده می‌کنیم. [۱] خلاصه این الگوریتم به صورت زیر است:

مرحله ۱: یک رشته پروتئین با یک سری حرکات دوران و انتقال در فضا به حالتی درمی‌آید که فاصله آن (c-RMS) تا نزدیک‌ترین نقاط مشبکه کمینه شود. نکته‌ای که این جا وجود دارد این است که این فاصله بدست آمده یک کران پایین برای فاصله نهایی (c-RMS) با پروتئین نشانده شده است.

مرحله ۲: نگاشتی که در مرحله قبل تعریف کردیم را در نظر بگیرید، یک زنجیر، یک مسیری تعریف می‌شود از رئوس مشبکه که از یک رأس حداکثر یک بار می‌گذرد. بزرگ‌ترین زنجیر از میان نقاط نگاشت شده از پروتئین به مشبکه انتخاب می‌شود.

مرحله ۳: از بزرگ‌ترین زنجیر شروع می‌کنیم، این زنجیر را از دو طرف آنقدر ادامه می‌دهیم که تمام پروتئین بر روی مشبکه نشانده شده باشند. برای این کار یک روش i-امین بردار استفاده می‌شود، به این صورت که بهترین نشاندن برای (k+1)-امین مونومر را پیدا می‌کنیم، سپس همه مسیرهایی با طول *ahead* را در نظر بگیرید، از k-امین اسید آمینه نشانده شده شروع می‌کنیم. از میان همه مسیرها، آن مسیر را انتخاب می‌کنیم که کمترین اختلاف محلی را با مقدار RMS دارد. اسید آمینه بعدی، اولین اسید آمینه‌ای است که از مسیر بهینه انتخاب می‌شود. هر چه مقدار *ahead* بیشتر باشد، نتیجه دقیق‌تر خواهد بود، منتها زمان اجرای این الگوریتم یک تابع نمایی از مقدار *ahead* می‌باشد. مقدار *ahead* برای c-RMS از ۳ یا ۴، برای d-RMS از ۶، و برای a-RMS از ۲ یا ۳ تجاوز نمی‌کند. ما برای همه آن‌ها از مقدار ۴ استفاده کردیم.

منابع

[1]. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249(2):493-507.

[2]. Nima Hamedani. Geometric properties of real proteins. *PhD Thesis on Physics in Sharif U.T.2006.*

[3]. Wikipedia: http://en.wikipedia.org/wiki/DNA_computing

[4]. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243(4):668-682.

[5]. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci U S A 1992;89(7):2536-2540.

[6]. Creighton TE. Proteins : Structures and Molecular Properties. New York: Freeman; 1993.

[7]. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104(1):59-107.

[8]. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283-438.

[9]. C.Mead, J.Manuch, X.Huang, . Bhattacharyya, L. Stacho, A. Gupta. Investigating lattice structure for Inverse Protein Folding. *FEBS Journal.*

[10]. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 1996;5(7):1406-1420.

[11]. Godzik A, Skolnick J, Kolinski A. Regularities in interaction patterns of globular proteins. Protein Eng 1993;6(8):801-810.

[12]. Rabow AA, Scheraga HA. Improved genetic algorithm for the protein folding problem by use of a Cartesian combination operator. Protein Sci 1996;5(9):1800-1815.

[13]. Koehl P, Delarue M. Building protein lattice models using self-consistent mean field theory. The Journal of Chemical Physics 1998;108(22):9540-9549.
[14]. Wikipedia: http://en.wikipedia.org/wiki/Bioinformatics
[15]. Hue Sun Chan, Ken A.Dill. “Sequence space soup” of proteins and copolymers. In *Journal of Chemican Physics,* 95(5): 3775-3787, 1991.
[16]. D.A.Hinds, M.Levitt. A lattice model for protein structure prediction at low resolution. In *Proceedings of the National Academy of Sciences,* 89(7): 2536-2540, 1992.
[17]. Adam Godzik, Andrzej Kolinski, Jeffrey Skolnick. Lattice representations of globular proteins: how good are they? In *Journal of Computational Chemistry,* 14(10): 1194-1202, 1993.
[18]. Wikipedia: http://en.wikipedia.org/wiki/Biology
[19]. Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. Biochemistry 1990;29(13):3287-3294.
[20]. Kian Mirjalali. Analysis and proposing algorithms for lattice modeling of proteins. *MS Thesis on Computer Engineering in Sharif U.T.2010.*
[21]. Jan Manuch, Daya Ram Gaur. Fitting protein chains to cubic lattice is NP-Complete. In *Journal of bioinformatics and computational biology,* 6(1): 93-106, 2008.
[22]. Wikipedia: http://en.wikipedia.org/wiki/Protein

A Comparison of Different Lattices in the Protein Fitting Problem

Abstract

A successful solution to the Inverse Protein Folding problem (IPF) would determine the necessary residue sequence to produce a stable in-vivo molecule for a specified protein structure. IPF has great potential as a tool in drug design as it could provide a first approximation of a primary amino acid sequence for a target protein. Our investigation incorporates statistical and computational analyses of a subset of the Protein Data Bank (PDB) to define lattice criteria and rank candidate lattices by how close the lattice approximates underlying protein crystal structure.

Keywords: protein, protein folding, protein fitting, protein data bank, protein structure.

A Comparison of Different Lattices in the Protein Fitting Problem

by
Meysam Aghighi

Under supervision of
Prof. Mohammad Ghodsi

Summer 2010
Computer Engineering Department
Sharif University of Technology
Tehran