بسم الله الرحمن الرحيم

دانشکده‌ی ریاضی، آمار و علوم کامپیوتر

دانشگاه تهران

تهران

پیشگویی ساختار سوم پروتئین‌ها بر اساس مدل آب‌گریز-آب‌دوست با استفاده از شبکه‌های عصبی

نگارش

صغری میکائیل‌نژاد

استاد راهنما

دکتر عباس نوذری دالینی

استاد مشاور

دکتر هایده اهرابیان

پایان‌نامه جهت دریافت درجه‌ی کارشناسی ارشد

در علوم کامپیوتر

بهمن ۸۷

University of Tehran
College of Science
School of Mathematics, Statistics and Computer Science

Prediction of Tertiary Structure of Proteins in the HP Model Using Neural Networks

by

Soghra Mikaeyl Nejad

Under supervision of

Dr. Abbas Nowzari Dalini

Co-supervisor

Dr. Hayedeh Ahrabian

A thesis submitted to the Graduate Studies Office
In partical fulfillment of the requirements for
The degree of Master of Science
Computer Science

January 2009

تقدیر و تشکر

وَسَخَّرَ لَكُم مَّا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ جَمِيعًا مِّنْهُ إِنَّ فِي ذَٰلِكَ لَآيَاتٍ لِّقَوْمٍ يَتَفَكَّرُونَ

حمد و سپاس فراوان خدای را که توفیق کسب علم و دانش نصیب بنده خویش نمود تا وجودش را به زینت علم بیاراید. اکنون که با لطف و استعانت پروردگار گامی دیگر را در زندگی خود پشت‌سر نهاده‌ام، بر خود لازم می‌دانم که مراتب سپاس و قدردانی خویش را تقدیم به کسانی کنم که در طول مدت تحصیل مرا یاری نموده‌اند.

با تشکر از اساتید محترم، آقای دکتر نوذری دالینی و خانم دکتر هایده اهرابیان به خاطر راهنمایی‌هایشان در این پایان‌نامه که مطمئنا بدون هدایت ایشان این کار به انجام نمی رسید.

از پدر و مادر عزیزم که در تمام مراحل زندگی یاور و پشتیبان من بوده و همواره مرا مرهون لطف و محبت‌های خود نموده‌اند سپاس‌گزاری می‌کنم.

هم‌چنین از همکاری و کمک‌های دوستان عزیزم در گردآوری و ویرایش پایان‌نامه صمیمانه تشکر می‌کنم.

همتم بدرقه راه کن ای طایر قدس که دراز است ره مقصد و من نوسفرم

تقدیم به

دو ستاره همیشه درخشان زندگیم، پدر و مادر عزیز و مهربانم

به پاس تمامی فداکاری‌ها، محبت‌ها، مهربانی‌ها، حمایت‌ها، تشویق‌ها و امیدهایی که همواره قدمهایم را در پیمودن مراحل مختلف زندگی استوارتر می‌کرد.

و تقدیم به

برادر و خواهرانم

که آسمان نگاه پرمهرشان شوق پیمودن این راه را در من زنده می‌کرد.

چکیده

با توجه به نقش و کارکرد حیاتی پروتئین‌ها در بدن انسان و وابستگی عمل‌کرد یک پروتئین به ساختار سه‌بعدی آن پیش‌بینی ساختار توالی پروتئین اهمیت فوق‌العاده‌ای می‌یابد. تاکنون روش‌های مختلفی برای این مسئله ارائه شده است. در برخی از این روش‌ها از شبکه‌های عصبی برای پیش‌گوئی ساختار پروتئین‌ها استفاده شده است که در عمل به نتایج قابل قبولی منجرشده‌اند. یکی از این روش‌ها استفاده از یک شبکه عصبی به نام شبکه کشسان است. همچنین در برخی از این الگوریتم‌ها از خاصیت آب‌دوستی و آب‌گریزی اسیدهای‌آمینه پروتئین‌ها برای حل مسئله پیش‌گوئی ساختار پروتئین‌ها استفاده می‌کنند.

در این پایان‌نامه پس از معرفی مقدمات زیست‌شناسی مولکولی و شبکه‌های عصبی، الگوریتم‌هایی که برای تاکردن پروتئین‌ها با استفاده از شبکه کشسان وجود دارند بررسی می‌شوند. در این الگوریتم‌ها یک توالی پروتئین به عنوان ورودی به شبکه کشسان داده می‌شود. خروجی شبکه بهترین تاخوردگی برای توالی می‌باشد. در نهایت یک شبکه کشسان با معیار جدیدی برای انرژی شبکه با استفاده از سطح دسترسی اسیدهای‌آمینه توالی ارائه خواهد شد.

پیش گفتار

یکی از اجزای مهم و اساسی در سلول‌های موجودات زنده پروتئین‌ها هستند که انواع مختلفی دارند. هریک از این انواع پروتئین‌ها فعالیت مخصوص به خود دارند و فعالیت‌های ضروری موجودات زنده را کنترل می‌کنند. پروتئین‌ها از مولکول‌های پلیمر ناهمگن که از زیر واحدهایی که در یک رشته خطی به هم وصل شده‌اند، تشکیل یافته‌اند. زیرواحدهای تشکیل دهنده پروتئین‌ها اسیدهای‌آمینه می باشند و در کل ۲۰ نوع اسید آمینه در طبیعت وجود دارد. چگونگی تاخوردن رشته پروتئین و نحوه قرار گرفتن این زیرواحدها، به روش‌های مختلفی انجام می‌گیرد که فقط یکی از این تاخوردگی‌ها باعث می شود پروتئین فعالیت اصلی خود را به درستی انجام دهد و در پایین‌ترین سطح انرژی قرار بگیرد. نحوه تاخوردگی یک پروتئین به پیش‌گوئی تاخوردگی پروتئین یا پیش‌گوئی ساختار پروتئین معروف است. بیماری‌هایی مانند آلزایمر، جنون گاوی و حتی برخی از سرطان‌ها از تا شدن نادرست برخی از پروتئین‌ها ناشی می‌شوند. اگر پروتئینی به طرز ناصحیحی تا بخورد برای سلول‌های مجاور نیز ایجاد مسمومیت می‌کند. اگر یک پروتئین تاخورده باز شود، دوباره به همان حالت اصلی خود تا می‌خورد. تشخیص نحوه تاخوردن یک پروتئین به پیش‌گوئی تاخوردن پروتئین یا پیش‌گوئی ساختار پروتئین معروف است. تاکنون رویکردهای مختلفی برای پیش‌گوئی تاخوردن پروتئین‌ها ارائه شده است. در این روش‌ها به طور کلی دنباله‌ای از اسیدآمینه‌ها به عنوان ورودی دریافت می‌شود و الگوریتم سعی می‌کند بهترین روش تاخوردن آن را که به ساختار اصلی این پروتئین نزدیک‌تر است به عنوان خروجی تولید کند. یکی از عواملی که در چگونگی تاخوردن پروتئین‌ها اثر می‌گذارد، ویژگی‌های اسیدآمینه‌های موجود در توالی است. اسیدآمینه‌های موجود در رشته پروتئین دارای خاصیت آب‌گریزی یا آب‌دوستی هستند. اسیدآمینه‌های آب‌دوست برخلاف اسیدآمینه‌های آب‌گریز تمایل زیادی به تماس با مولکول‌های آب

دارند، از این‌رو به هنگام تاخوردن پروتئین اسیدآمینه‌های آب‌دوست تمایل دارند که به سمت بیرونی ساختار رفته و نزدیک به مولکول‌های آب قرار بگیرند. اسیدآمینه‌های آب‌گریز نیز در مرکز ساختار یک هسته آب‌گریز تشکیل می‌دهند.

ویژگی دیگری که اسیدآمینه‌ها دارند سطح دسترسی آن‌هاست. منظور از سطح‌دسترسی یک اسیدآمینه، نسبت مساحتی از آن که در محیط آبی با آب در تماس است به کل مساحت سطح آن اسیدآمینه می‌باشد. سطح‌دسترسی اسیدآمینه‌ها نیز نقش مهمی در پیش‌گوئی ساختار پروتئین‌ها دارد.

در سال‌های اخیر از تکنیک‌های ماشین یادگیری از قبیل شبکه‌های عصبی به طور گسترده‌ای در روش‌های پیش‌گوئی ساختار پروتئین‌ها استفاده می‌شود. در این روش‌ها توالی پروتئین به عنوان ورودی به شبکه وارد می‌شود و شبکه بهترین تاخوردگی آن را با کم‌ترین سطح انرژی به عنوان خروجی تولید می‌کند. یکی از شبکه‌های عصبی که در این زمینه به کار گرفته شده است، شبکه کشسان نام دارد. این شبکه که بیشتر جنبه هندسی دارد اغلب برای مسائل بهینه‌سازی استفاده می‌شود. شبکه‌های کشسانی که تاکنون برای حل مسئله تاکردن پروتئین‌ها استفاده شده‌اند یک رشته پروتئین را با استفاده از خاصیت آب‌گریز-آب‌دوستی اسیدآمینه‌های آن بر روی یک لاتیس تا می‌کنند. انرژی ساختار پروتئین در این روش‌ها بر حسب این‌که هر جفت اسیدآمینه آب‌گریزی که بر روی لاتیس مجاور باشند و در توالی هم‌جوار نباشند محاسبه می‌شود.

در این پایان‌نامه راه‌حل‌های بهینه برای مسئله تا کردن پروتئین‌ها با استفاده از شبکه عصبی کشسان بررسی شده‌اند به‌طوری که یک رشته پروتئین به عنوان ورودی به شبکه وارد می‌شود و این شبکه بهترین تاخوردگی آن را با کمترین سطح انرژی به عنوان خروجی تولید می‌کند. در روش‌هائی که بحث شده‌اند از مدل آب‌گریز-آبدوست استفاده شده است. در انتها روشی نوین با استفاده از سطح دسترسی هر اسیدآمینه در یک رشته پروتئین پیشنهاد شده است.

در فصل اول به تعریف مفاهیم اولیه زیستی و معرفی شبکه‌های عصبی پرداخته می‌شود. سپس در فصل دوم شبکه کشسان معرفی شده و نحوه عمل‌کرد آن توضیح داده می‌شود. فصل سوم شامل الگوریتم‌های شبکه کشسان دوبعدی و سه‌بعدی موجود برای تاکردن پروتئین‌هاست و در فصل آخر مشکلات روش‌های قبلی بررسی شده و راه‌حل‌هایی برای رفع این مشکلات ارائه شده است. همچنین

یک الگوریتم شبکه کشسان با معیاری جدید به‌نام سطح دسترسی برای تعیین سطح انرژی ساختار ارائه خواهد شد.

فهرست مطالب

۱ مفاهیم اولیه ۱

۱-۱ مفاهیم زیستی ۱

۱-۱-۱ اسیدهای نوکلئیک ۱

۱-۱-۲ پروتئین‌ها ۵

۱-۱-۳ پروتئومیکس ۱۴

۱-۱-۴ تاکردن پروتئین‌ها ۱۶

۱-۱-۵ مدل آب‌گریز-آب‌دوست ۱۸

۱-۲ شبکه‌های عصبی ۲۰

۱-۲-۱ مدل محاسباتی شبکه‌های عصبی ۲۰

۱-۲-۲ شبکه عصبی هاپفیلد ۲۴

۱-۳ روش شبیه‌سازی تبرید ۲۷

۲ شبکه کشسان ۳۲

۲-۱ معرفی شبکه ۳۲

۲-۲ پیکربندی شبکه ۳۳

۲-۳ بررسی انرژی شبکه کشسان ۳۶

۲-۴ آرایش فضای اولیه شبکه کشسان و بسط آن ۳۸

۲-۵ الگوریتم شبکه کشسان برای مسئله فروشنده دوره‌گرد ۳۹

۳ تاکردن پروتئین‌ها با روش شبکه کشسان ۴۳

۳-۱ معرفی روش ۴۳

۳-۲ ساختار کلی شبکه ۴۵

۳-۳ تا کردن پروتئین‌ها در حالت دوبعدی ۴۹

۳-۴ جستجوی محلی ۵۰

۳-۵ تا کردن پروتئین‌ها در حالت سه بعدی ۵۳

۳-۶ جستجوی محلی **LS1** ۵۵

۳-۷ جستجوی محلی **LS2** ۵۹

۴ پیش‌گوئی‌تاکردن پروتئین‌هابا سطح دسترسی ۶۳

۴-۱ مشکلات روش‌های قبلی ۶۳

۴-۲ تابع انرژی با استفاده از سطح دسترسی ۶۶

۴-۳ الگوریتم ارائه شده برای تاکردن پروتئین‌ها در حالت سه‌بعدی ۶۷

۴-۴ پیاده‌سازی و نتایج ۶۹

۴-۵ پیشنهادات ۷۲

مراجع ۷۳

واژه‌نامه‌ی فارسی ۷۸

واژه‌نامه‌ی انگلیسی ۸۳

فصل ۱

مفاهیم اولیه

در این فصل مفاهیم مقدماتی که برای پایان نامه لازم است بیان می‌شود. در این قسمت ابتدا به بررسی پروتئین‌ها پرداخته می‌شود[۷, ۳۸] و در ادامه شبکه‌های عصبی مورد مطالعه قرار می‌گیرند[۹].

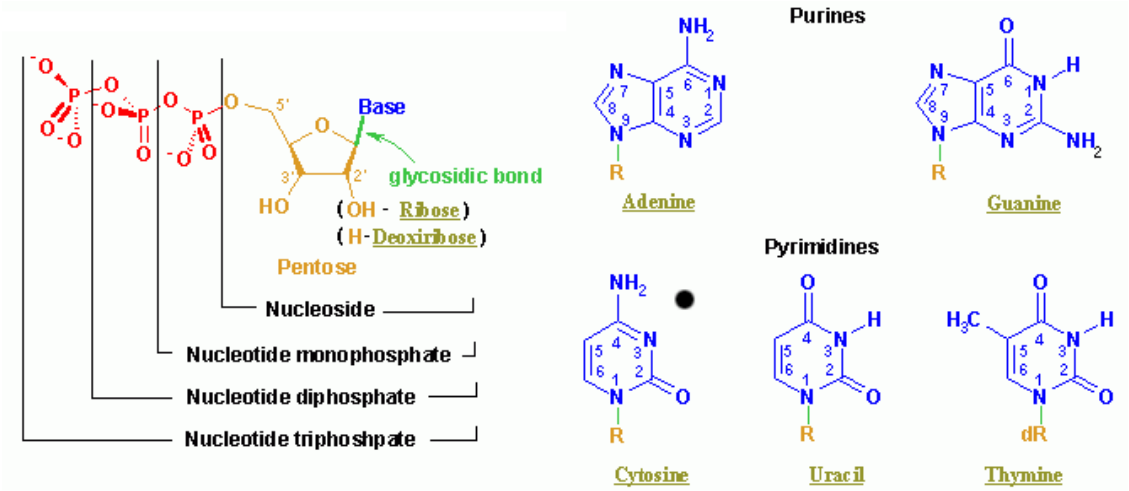
۱-۱ مفاهیم زیستی

زیست‌شناسی یا به عبارت دیگر علوم زیستی همواره در پی شناخت بیشتر و بهتر ساختار موجودات زنده و ارتباط آن‌ها با یکدیگر می‌باشد. ساختار موجودات زنده از اجزا متنوعی تشکیل شده است که در این بخش مواردی از آن‌ها که در این پایان‌نامه مورد نیاز است تعریف می‌گردد.

۱-۱-۱ اسیدهای نوکلئیک

در هسته هر سلول مولکول‌هایی قرار دارند که اساسی‌ترین اطلاعات حیات را در خود ذخیره کرده‌اند. این مولکول‌ها، داکسی ریبونوکلئیک‌اسید نام دارند. داکسی ریبونوکلئیک‌اسیدها از چهار نوع مولکول ساخته شده‌اند که به آن‌ها نوکلئوتید گفته می‌شود. این چهار نوکلئوتید آدنین، گوانین، سیتوزین و تیمین نام دارند. در شکل ۱-۱ این نوکلئوتیدها نشان داده شده‌اند. این مولکول‌ها خود به دو زیر گروه تقسیم می‌شوند: آدنین و گوانین در گروه پیورین قرار می‌گیرند و سیتوزین و تیمین در گروه پیریمیدین جای داده می‌شوند.

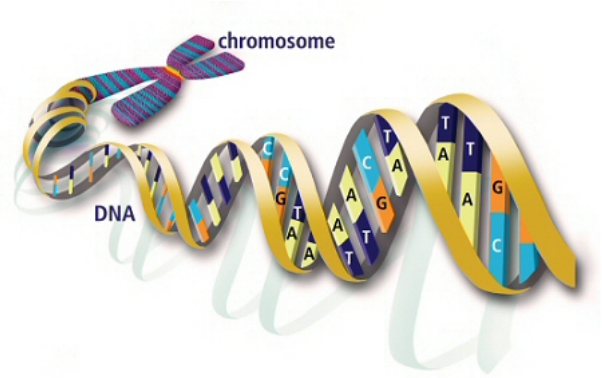
هنگام سنتز مولکول *DNA*، نوکلئوتیدها به اسیدهای نوکلئیک تبدیل می‌شوند که با پیوند فسفودی‌استر آن‌هابه‌یکدیگر، رشته‌های*DNA* ساخته می‌شوند. *DNA* یک مارپیچ دوتایی است، این مارپیچ در شکل



شکل ۱-۱ : ساختار نوکلئوتیدهای رشته DNA.

۱-۲ نشان داده شده است. نوکلئوتیدها حلقه‌های مسطحی دارند که اندازه آن‌ها بین ۳ تا ۴ آنگستروم می‌باشد. وقتی مارپیچ دوتایی تشکیل می‌شود، مولکول‌های A با مولکول‌های T رشته مقابل و مولکول‌های G با مولکول‌های C رشته روبه رو پیوند هیدروژنی برقرار می‌کنند و جفت‌های قلیایی ایجاد می‌شود. این جفت‌های قلیایی باعث می‌شوند که مارپیچ پایدار باقی بماند. تصاویری که با استفاده از اشعه X از مولکول DNA گرفته شده است، نشان می‌دهد که در هر دور از مارپیچ ۱۰ جفت قلیایی وجود دارد. ساختمان سه‌بعدی DNA در سال ۱۹۵۳، تقریبا ۲۰ سال پس از کشف ماهیت شیمیایی DNA به‌وسیله جمیز واتسون[۱] و فرانسیس کریک[۲] کشف شد. مدل مارپیچی می‌تواند نحوه رونویسی از مولکول DNA در هنگام تقسیم سلول را نیز توجیه کند. در هنگام تقسیم سلول، باید نسخه مشابهی از DNA تهیه شود تا در سلول‌های جدید قرار گیرد. این فرآیند اصطلاحاً رونویسی نامیده می‌شود. برای انجام این کار، اتصالات هیدروژنی بین نوکلئوتیدها باز شده و دو رشته مارپیچ از هم جدا می‌گردند. سپس هر یک از این رشته‌ها به عنوان پایه‌ای برای ساخت رشته مقابل استفاده می‌شوند. به این ترتیب دو مارپیچ کاملا یکسان DNA ساخته می‌شوند و هر کدام از آن‌ها در یکی از دو سلول نوزاد قرار می‌گیرد. چون در هر بار رونویسی، نیمی از مولکول DNA قبلی حفظ می‌شود، گفته می‌شود که رونویسی DNA ، نیمه

۱ James Watson
۲ Francis Crick



شکل ۱-۲ : مارپیچ رشته *DNA*

پایستار است. اگرچه *DNA* اطلاعات ژنتیکی جاندار زنده را در خود دارد، اما برای عمل‌کرد موفق به وجود ریبونوکلئیک اسید یا *RNA* نیاز دارد. *RNA* هم مانند *DNA* از رشته‌های اسید نوکلئیکی تشکیل شده که با پیوندهای مشابهی به هم متصل شده اند، اما دو تفاوت عمده با *DNA* دارد. اول اینکه در ساختار آن به جای تیمین، از اوراسیل[۳] استفاده شده است و دوم اینکه مارپیچی نیست و فقط از یک رشته تنها ساخته شده است. برای انجام بعضی کارها *RNA* از روی *DNA* ساخته می‌شود. ساخته شدن *RNA* از روی *DNA* را رونویسی می‌گویند. رونویسی به وسیله آنزیم *RNA*پلی‌مراز صورت می‌گیرد. رونویسی براساس وجود رابطه مکملی بین بازها انجام می‌شود. در شکل ۱-۳ نمونه‌ای از رونویسی *RNA* از روی *DNA* دیده می‌شود. اساس رونویسی در پروکاریوت‌ها و یوکاریوت‌ها یکسان است. در این‌جا ابتدا رونویسی در پروکاریوت‌ها بررسی می‌شود. رونویسی در سه مرحله انجام می‌شود.

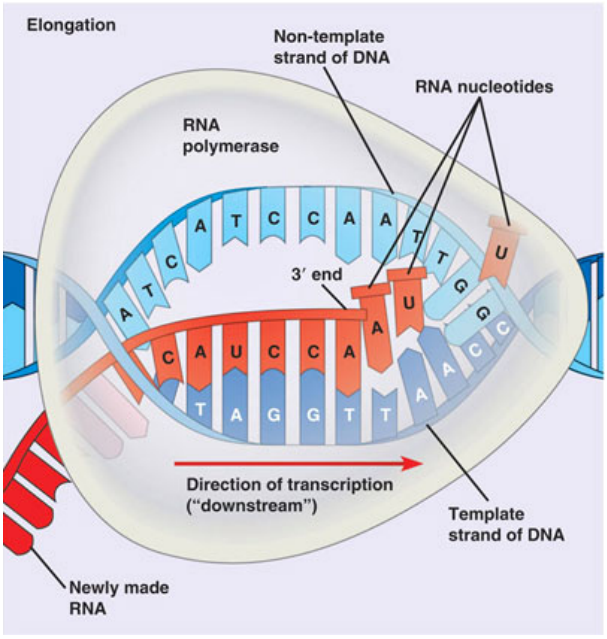
مرحله اول: آغاز

مرحله دوم: ادامه

مرحله سوم: پایان

مرحله اول رونویسی با اتصال آنزیم *RNA*پلی‌مراز به قسمتی از *DNA* به نام راه‌انداز آغاز می‌شود. راه‌انداز قسمتی از *DNA* است که به آنزیم امکان می‌دهد رونویسی را از محل درست آغاز کند و این کار را مثلا از وسط ژن آغاز نکند. اگر *RNA*پلی‌مراز روی قسمتی از *DNA* بنشیند که آن‌جا راه‌انداز نباشد، آن‌قدر روی *DNA* حرکت می‌کند تا به راه‌انداز برسد.

[۳] Uracil



شکل ۱-۳ : رونویسی *RNA* از روی رشته *DNA*.

در مرحله دوم *RNA*پلی‌مراز دو رشته‌ای *DNA* را از یکدیگر باز می‌کند و یکی از دو رشته *DNA* را رونویسی می‌کند. به این ترتیب که در مقابل هریک از رئوکسی ریبونوکلئوتیدهای آن، ریبونوکلئوتید مکمل را قرار می‌دهد و هر ریبونوکلئوتید را به ریبونوکلئوتید قبلی وصل می‌کند. اولین نوکلئوتیدی از *DNA* را که رونویسی می‌شود جایگاه رونویسی می‌نامند.

در انتهای ژن، قسمت مخصوصی وجود دارد که به آن جایگاه پایان رونویسی می‌گویند، با رونویسی شدن این جایگاه *DNA*، *mRNA* و آنزیم *RNA*پلی‌مراز از یکدیگر جدا می‌شوند و رونویسی پایان می‌پذیرد. نحوه عمل آنزیم *RNA*پلی‌مراز در کلیه مراحل رونویسی بدون درنگ به‌دنبال هم اجرا می‌شوند. *RNA*هایی که در حال ساخته شدن هستند ساختارهای پر مانندی را به نمایش می‌گذارند. علت ایجاد چنین ساختاری این است که به محض ترک راه اندازتوسط *RNA*پلی‌مراز، آنزیم دیگری می‌تواند روی آن بنشیند و رونویسی را آغاز کند. بنابراین در حالی که آنزیم اول قدری از رونویسی را به پیش برده است آنزیم دوم تازه رونویسی را آغاز کرده است. در نتیجه طول *RNA*ای که آنزیم دوم ساخته است از طول *RNA*ای که آنزیم اول ساخته است کوتاه‌تر است. این روند هم‌چنان ادامه پیدا می‌کند. به این معنی که وقتی آنزیم دوم راه انداز را ترک کرد آنزیم سوم روی راه‌انداز می‌نشیند و *RNA*ای که می‌سازد از *RNA*ای

که آنزیم دوم در حال ساختن آن است کوتاه‌تر است.

به‌طور کلی دستگاه رونویسی یوکاریوتی پیچیده‌تر از دستگاه پروکاریوتی است سلول‌های یوکاریوتی برخلاف سلول‌های پروکاریوتی که تنها یک نوع آنزیم *RNA* پلی‌مراز دارند دارای سه نوع آنزیم‌اند. تفاوت دیگری که رونویسی در یوکاریوت‌ها با رونویسی در پروکاریوت‌ها دارد این است که در یوکاریوت‌ها آنزیم *RNA* پلی مراز قادر نیست به تنهایی راه‌انداز را شناسایی کند. اتصال آنزیم به راه‌انداز به کمک پروتئین‌های ویژه‌ای به نام عوامل رونویسی صورت می‌گیرد. همچنین در یوکاریوت‌ها *RNA* قبل از خروج از هسته دست‌خوش تغییر می‌شود و به *RNA* بالغ تبدیل می‌گردد.

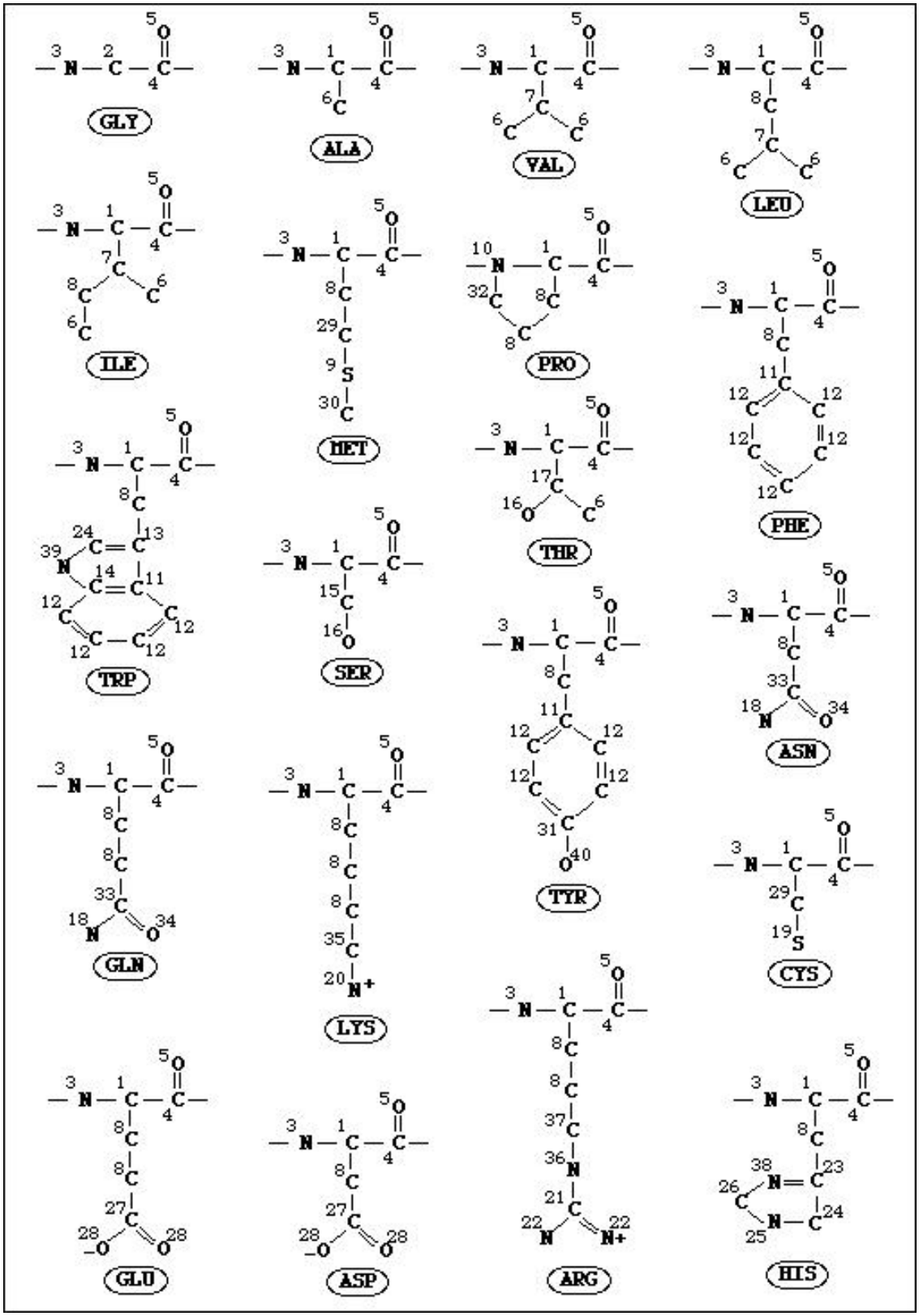
۱-۱-۲ پروتئین‌ها

تمام فعالیت‌هایی که دریک سلول انجام می‌شود بر عهده پروتئین‌ها است. تقریباً همه پروتئین‌ها اثر خود را با همکاری پروتئین‌های دیگر در سلول اعمال می‌کنند و هیچ پروتئینی نیست که در سلول به‌تنهایی عمل کند. اصلی ترین مواد مغذی هر سلول زنده پروتئین‌ها هستند. در ساختمان آن‌ها نه تنها کربن، هیدروژن و اکسیژن وجود دارد، بلکه ازت و گاهی گوگرد نیز در بعضی ترکیبات آن‌ها دیده می‌شود. پروتئین‌ها مسئول انجام اعمال گوناگونی هستند. آن‌ها در تشکیل ماده انقباضی عضلات نقش مهمی ایفاء می‌کنند و همچنین در ساختن بعضی ازهورمون‌ها، آنزیم‌ها و آنتی‌کورها و تبدیل انرژی شیمیایی به کار به طور گسترده استفاده می‌شوند. یکی از مسائل مهمی که موجب ایجاد انگیزه برای شناخت پروتئین‌ها در دانشمندان می‌شود، تولید دارو است. در صورتی که به درستی ساختار عوامل حمله کننده شناسایی شوند و چنان‌چه ساختاری متناسب با عامل مورد نظر پیدا کرده و تولید شود، با بسیاری از بیماری‌ها به سادگی می‌توان مبارزه کرد. این امر بسیاری از شرکت‌های داروسازی را بر آن داشته است که به طراحی پروتئین‌ها به عنوان دارو بپردازند. پروتئین‌ها از زیرواحدهایی به‌نام اسیدآمینه ساخته شده‌اند و همان‌طور که قبلا اشاره شد در فرآیند رونویسی و ترجمه از روی مولکول‌های *DNA* موجود در هسته به‌وجود می‌آیند. از آن‌جا که اسیدهای‌آمینه با ترتیب‌های مختلف و با طول نامحدود در یک توالی می‌توانند قرار بگیرند، از این‌رو انواع بی‌شماری از پروتئین‌ها می‌توانند وجود داشته باشند. هر موجود زنده توانایی ساخت تعداد زیادی پروتئین را دارد. به عنوان مثال، هر انسان بین ۳۰۰۰۰ تا ۵۰۰۰ پروتئین

مختلف می‌سازد.

به صورت کلی پروتئین از تعدادی اسید آمینه تشکیل شده است که در یک رشته پشت‌سر یکدیگر قرار گرفته‌اند. ترتیب و ماهیت اسیدهای آمینه، ویژگی‌های هر پروتئین را تعیین می‌کند. یک پروتئین با طول متوسط حدود ۱۰۰ اسید آمینه دارد. تعداد کل اسید آمینه‌های موجود در طبیعت ۲۰ عدد می‌باشد که هر کدام از ترکیب‌های متفاوتی از اتم‌های پیوند خورده به وجود آمده‌اند و با توجه به ترکیب و ترتیب اتم‌ها، هر اسید آمینه خواص خاص خود را داراست. به رشته اسید آمینه که پروتئین از آن تشکیل شده ساختار اول یا توالی پروتئین می‌گویند. تمامی اسیدهای آمینه دارای یک قسمت شبیه به هم هستند که این قسمت‌های شبیه، به یکدیگر متصل شده و ستون فقرات پروتئین را می‌سازند. تفاوت اسیدهای آمینه در باقی قسمت‌های آن‌هاست که به آن زنجیر جانبی می‌گویند. هر اسید آمینه، از یک کربن نامتقارن به نام کربن آلفا تشکیل شده که با چهار گروه مختلف کربوکسیل، اتم هیدروژن، گروه آمینه بازی و یک زنجیره غیر جانبی پیوند برقرار می‌کند. شکل ۱-۴ ساختمان مولکولی برخی از اسیدهای آمینه را به تصویر می‌کشد. زنجیره غیرجانبی یا گروه R ممکن است یک زنجیره کربنی و یا یک حلقه کربنی باشد. عوامل دیگری مانند الکل، آمین، کربوکسیل و نیز گوگرد می‌توانند در ساختمان ریشه R شرکت کنند. این زنجیره جانبی خود چندین اتم کربن دارد و آن‌ها را به ترتیبی که از کربن آلفا فاصله می‌گیرند، با حروف بتا β، گاما γ و دلتا δ نشان می‌دهند. اگر در حالی که عامل کربوکسیل روی کربن آلفا قرار دارد، گروه آمینه روی کربن‌های غیرآلفا قرار گیرد، نوع اسید آمینه به γ، β یا δ تغییر خواهد کرد. اسیدهای آمینه آزاد، به مقدار بسیار ناچیز در سلول‌ها وجود دارند. بیشتر اسیدهای آمینه آلفا در سنتز پروتئین می‌کنند، در صورتی که اسیدهای‌آمینه بتا، گاما و دلتا واسطه‌های شیمیایی هستند.

هر یک از اسیدهای‌آمینه علاوه بر نام، دارای یک کد سه حرفی، یک کد یک حرفی و یک کد عددی هستند. نام این اسیدها و کدهای آن‌ها در شکل ۱-۵ نشان داده شده است. کدهای سه حرفی در حقیقت توالی سه نوکلئوتیدی است که این اسیدآمینه را کد می‌کند و در فرآیند ساخت اسیدآمینه، این توالی نوکلئوتیدی باعث تولید اسیدآمینه متناظر می‌شود. اسیدهای آمینه در آغاز تشکیل زمین، به همراه سایر مواد آلی پیدا شدند. اسیدهای آمینه‌ای که در حضور پرتوهای فرابنفش بوجود آمدند،



شکل ۱-۴ : ساختمان مولکولی اسیدهای‌آمینه.

Name	Integer	Code	Abberviation	codons
Alanine	1	A	Ala	GCU GCC GCA GCG
Arginine	2	R	Arg	CGU CGC CGA CGG AGA AGG
Asparagine	3	N	Asn	AAU AAC
Aspartic acid (Aspartate)	4	D	Asp	GAU GAC
Cysteine	5	C	Cys	UGU UGC
Glutamine	6	Q	Gln	CAA CAG
Glutamic acid (Glutamate)	7	E	Glu	GAA GAG
Glycine	8	G	Gly	GGU GGC GGA GGG
Histidine	9	H	His	CAU CAC
Isoleucine	10	I	Ile	AUU AUC AUA
Leucine	11	L	Leu	UUA UUG CUU CUC CUA CUG
Lysine	12	K	Lys	AAA AAG
Methiomine	13	M	Met	AUG
Phenylalamine	14	F	Phe	UUU UUC
Proline	15	P	Pro	CCU CCC CCA CCG
Serine	16	S	Ser	UCU UCC UCA UCG AGU AGC
Threonine	17	T	Thr	ACU ACC ACA ACG
Tryptophan	18	W	Trp	UGG
Tyrosine	19	Y	Tyr	UAU UAC
Valine	20	V	Val	GUU GUC GUA GUG

شکل ۱-۵ : لیست کدهای داده شده اسیدهای آمینه.

گوناگونی بسیار داشته‌اند، اما به دلایلی ناشناخته، تنها بیست اسید آمینه در سلول زنده کاربرد پیدا کرده‌اند.تمام پروتئین‌ها از همین بیست اسیدآمینه تشکیل شده‌اند. در شکل ۵-۱ ساختمان مولکولی این اسیدآمینه‌ها نشان داده شده است. در ادامه چگونگی سنتز پروتئین‌ها بررسی می‌شود.

پروتئین‌ها از اتصال اسیدهای آمینه به یکدیگر از طریق پیوند پپتیدی بدست می‌آیند. تشکیل پیوند پپتیدی و قرار گرفتن ترتیب اسیدهای‌آمینه که برای هر پروتئین اختصاصی است، به سادگی امکان پذیر نیست. به همین دلیل می‌بایست در سلول مکانیسم ویژه‌ای وجود داشته باشد که بتواند ویژگی پروتئین‌ها را حفظ کند. بیوسنتز پروتئین‌ها در واقع ترجمه ترتیب نوکلئوتیدهای DNA به مولکول پروتئین بوسیله RNA است. انتقال اطلاعات از DNAها به وسیله $mRNA$ امکان‌پذیر است. بدین ترتیب برای هر پروتئین، $mRNA$ اختصاصی آن پروتئین وجوددارد. به عبارت دیگر هر پروتئین در روی DNA ، ژن اختصاصی دارد که اطلاعات آن ژن در $mRNA$ رونویسی و در مولکول ترجمه می‌شود. در بیوسنتز پروتئین‌هایی که در ساختارشان بیش از چند اسید آمینه دارند، وجود یک مکانیسم سنتزی که در آن ترکیبات و عوامل بسیاری دخالت می‌کنند، الزامی است. این مکانیسم به یک سیستم رمزیاب نیاز دارد که بطور خودکار واحد اسید آمینه معینی را در موقعیت ویژه‌ای از زنجیره پروتئینی قرار می‌دهد. در ادامه ترکیباتی که در سنتز پروتئین شرکت می‌کنند شرح داده می‌شود.

mRNA: این RNA اطلاعات مربوط به پروتئین ویژه‌ای را از مولکول DNA می‌گیرد و به ماشین سنتزکننده پروتئین یا ریبوزوم انتقال می‌دهد. در ترتیب نوکلئوتیدهای $mRNA$ هر سه نوکلئوتید مجاور بیان‌گر رمز یا کدون یک اسیدآمینه مشخص هستند و به همین جهت ترتیب نوکلئوتیدها در $mRNA$ بیان‌کننده ترتیب اسیدهای آمینه در پروتئین است. هر اسید آمینه رمز مشخصی دارد. فقط متیونین و تریپتوفان یک رمز دارند، در حالی‌که سایر اسیدهای آمینه واجد دو یا تعداد بیشتری رمز هستند. رمز میتونین همواره AUG است که آغاز سنتز را در همه پروتئین‌ها به عهده دارد. در سلول‌های پروکاریوت و یوکاریوت، پروتئین سازی با میتونین آغاز می‌شود. سه رمز UGA و UAA و UAG برای پروتئین رمزخوانی نمی‌کنند، بلکه رمزهایی هستند که پایان سنتز زنجیره پروتئین را بیان می‌کنند.

tRNA: $tRNA$ ساختار سه‌بعدی L شکل دارد که در آن دو ناحیه پذیرنده و آنتی‌کدون آزادند و بقیه مولکول تاب خورده است. آنتی‌کدون یا پادرمز شامل سه نوکلئوتید است که مکمل رمز ویژه‌ای از