

الله



دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پیش‌بینی نقشه تماس پروتئین توسط روش ماشین گروهی

پایان‌نامه کارشناسی ارشد هوش مصنوعی و رباتیک

نرگس حبیبی

استاد راهنما

دکتر محمد حسین سرائی



دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پایان نامه کارشناسی ارشد رشته هوش مصنوعی و رباتیک خانم نرجس خاتون

حبیبی

تحت عنوان

پیش‌بینی نقشه تماس پروتئین توسط روش ماشین گروهی

در تاریخ 9/1388 توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

دکتر محمد حسین سرائی

1- استاد راهنمای پایان نامه

دکتر حسن کربکندی

2- استاد مشاور پایان نامه

دکتر فرید شیخ‌الاسلام

3- استاد داور پایان نامه

دکتر سید رسول موسوی

4- استاد داور پایان نامه

دکتر علی محمد دوست حسینی

سرپرست تحصیلات تکمیلی دانشکده

((هوالعليم))

پروردگارا، تو را سپاس می‌گویم که به من فرصت آموختن بخشدی. لطف و رحمت توست که در همه لحظه‌های حیاتم جاریست.

دفتر زندگی‌ام ورق خورده و نظاره‌گر آغاز فصلی دیگر از سرنوشتمن هستم. هماره انسان‌های بزرگواری را می‌بینم که وجود هر یک نعمتی است بزرگ از سوی خداوند و اکنون، اینجا، فرصتی است، هرچند کوتاه، برای قدرانی و سپاس.

در ابتدا از خانواده عزیزم بینهایت سپاسگزارم. بعد از پروردگار، زندگی خود را مدیون زحمات، از خودگذشتگی‌ها و صبر آن‌ها می‌دانم.

از استادان ارجمند، جناب آقای دکتر سرائی و جناب آقای دکتر کربکندي کمال تشکر را دارم که زحمت راهنمائی و مشاوره این پایان‌نامه را تقبل نموده و با رهنمودهای ارزشمند و موثر و سخنان دلگرمکننده خود من را در طی انجام کار همراهی فرمودند، نهایت قدردانی را دارم. همچنین از آقای دکتر شیخ‌الاسلام و آقای دکتر موسوی که زحمت داوری این پایان‌نامه را به‌عهده داشتند، سپاسگزارم.

از آقای مهندس مهدویانی که در طول انجام پایان‌نامه همواره سخاوتمندانه وقت خود را در اختیار من قرار دادند و از راهنمائی‌های ارزشمند ایشان بھویژه در زمینه مباحث محاسباتی بسیار بهره‌مند شدم، کمال تشکر را دارم.

از پروفسور Kevin Karplus، استاد دانشگاه Santa Cruz کالیفرنیا که از طریق ایمیل، صبورانه به سوالات بسیاری از جانب من پاسخ داده و راهنمائی‌های موثری نمودند، متشکرم.

از استاد بزرگوار دانشکده برق و کامپیوتر که در طی دوران تحصیل افتخار شاگردی آن‌ها را داشتم، بسیار سپاسگزارم.

از تمامی دوستان عزیزم تشکر می‌کنم. روزهای بسیار خوب و فراموش‌نشدنی را در کنارشان تجربه کردم. در پایان از محضر استادی، عذر تقصیر می‌خواهم به خاطر کوتاهی‌های خود و از این که نتوانستم آن‌طور که شایسته است، کاری در خور جامعه علمی ارائه دهم.

نرگس حبیبی

کلیه حقوق مادی مترتب بر نتایج مطالعات،
ابتكارات و نوآوریهای ناشی از تحقیق موضوع
این پایاننامه متعلق به دانشگاه صنعتی اصفهان
است.

تقدیم به:

عزیزترین‌ها یم،

پدر و مادرم که در تمامی لحظه‌ها دلسوزانه

در کنارم قرار دارند،

و برادرانم که شادی‌بخش و امیدبخش زندگی‌ام

هستند.

فهرست مطالب

14	tRNA 2-6-2
14	کد ژنتیکی 2-7
15	پروتئین 2-8
 اسیدهای آمینه 2-8-1
17	ویژگی‌های اسیدهای آمینه 2-8-2
18	پیوندهای پپتیدی 2-8-3
19	2-8-4 سطوح مختلف ساختار پروتئین هشت
24	2-8-5 اهمیت توالی اسیدهای آمینه در ساختار و واکنش‌های پروتئین
24	2-8-6 روش‌های مطالعه ساختمان پروتئین
25	2-9 قضیه مرکزی زیست مولکولی

فصل سوم: مفاهیم محاسباتی پیش‌زمینه

26	3-1 مقدمه
26	3-2 مفاهیم بیوانفورماتیکی
26	PDB 3-2-1
29	3-2-2 pdbselect
30	3-2-3 فرمت FASTA
30	3-2-4 تراز توالی
32	3-2-5 جهش وابسته
34	3-2-6 پیش‌بینی ساختار دوم
35	3-2-7 پایگاه داده AAindex
36	3-3 مفاهیم یادگیری ماشین، تئوری اطلاعات و آمار و احتمال
36	3-3-1 مروری بر روش‌های ردیابی
37	3-3-2 شبکه‌های عصبی
44	3-3-3 بیش‌پوشش
44	3-3-4 اثر فراموشی خطرناک
45	3-3-5 یادگیری گروهی

.....	3-3-6 مفاهیم تئوری اطلاعات و آمار و احتمال
52	فصل چهارم: نقشه تماس پروتین
.....	4-1 مقدمه
.....	4-2 معرفی نقشه تماس.....
60	4-2-1 تماس‌های محلی
60	4-2-2 خطای نقشه تماس.....
61	4-2-3 ویژگی‌های نقشه تماس
63	4-3 پیش‌بینی نقشه تماس.....
63	4-3-1 نحوه ارزیابی پیش‌بینی نقشه تماس
.....	فصل پنجم: مروری بر کارهای پیشین
65	5-1 مقدمه
65	5-2 پیش‌بینی توسط شبکه‌های عصبی
68	5-3 پیش‌بینی توسط الگوریتم ژنتیک.....
69	5-4 پیش‌بینی توسط ماشین بردار پشتیبان
70	5-5 پیش‌بینی توسط قوانین و استگی
.....	فصل ششم: روش پیشنهادی
71	6-1 مقدمه
71	6-2 فایل‌های داده.....
72	6-2-1 پردازش فایل‌ها
72	6-3 استخراج ویژگی‌ها
73	6-3-1 به دست آوردن MSA
74	6-3-2 لاغر کردن MSA
74	6-3-3 استخراج ویژگی‌های توالی
74	6-3-4 استخراج ویژگی‌های تک ستونی
75	6-3-5 استخراج ویژگی‌های جفت ستونی

78	6-4 آمادهسازی دادههای آموزش و تست
78	6-4-1 تولید دادههای آموزش و تست
.....	6-4-2 استخراج نمونههای دارای اطلاعات
80	6-4-3 متوازن کردن دادههای آموزشی
81	6-5 نحوه ارزیابی کارائی سیستم
82	6-6 آموزش و تست مدل
82	6-6-1 آموزش مدل
85	6-6-2 تست مدل
85	6-7 تحلیل نتایج
86	6-7-1 نتایج روش پایه
86	6-7-2 نتایج روش مبتنی بر میانگین‌گیری گروه
87	6-7-3 نتایج روش پیشنهادی
	فصل هفتم: نتیجه‌گیری و پیشنهادها
90	1-7 نتیجه‌گیری
92	7-2 پیشنهادها
94	پیوست الف
96	مراجع

چکیده

بیوانفورماتیک علمی است بینرشته‌ای که قواعد ریاضی، فیزیک، شیمی و علوم کامپیوتر را به داده‌های وسیع، متنوع و پیچیده زیست‌شناسی، اعمال می‌کند. هدف بیوانفورماتیک، حل مسائل زیست‌شناسی در سطح مولکولی است. پروتئین‌ها از اجزایی اصلی سلول‌های موجودات زنده هستند. هر مولکول پروتئین، از زنجیره‌ای از اسیدهای آمینه تشکیل می‌شود. برای پروتئین چهار ساختار (اول، دوم، سوم، چهارم) تعریف شده است. ساختار اول، همان زنجیره اسیدهای آمینه آن است. ساختارهای دوم، ساختارهای محلی هستند که توسط برقراری پیوندهای نیتروژنی به وجود می‌آیند. رایج‌ترین آن‌ها، مارپیچ‌های آلفا و صفحه‌های بتا هستند. ساختار سوم، شکل کلی یک مولکول پروتئین و در واقع، موقعیت فضائی ساختارهای دوم نسبت به یکدیگر است که در اثر تاشدن زنجیره اسید آمینه شکل می‌گیرد. ساختار چهارم از تجمع چندین پروتئین ایجاد می‌گردد. محققان، پیوسته پروتئین‌های جدیدی کشف و توالی اسیدهای آمینه آن‌ها را تعیین می‌کنند. عمل پروتئین، وابسته به شکل ساختار سوم آن است. مولکول‌هایی که یک پروتئین می‌تواند به آن‌ها متصل شود، بستگی به شکل سبعدی پروتئین دارند. ساختار سوم خود، وابسته به توالی اسید آمینه است.

متاسفانه، تعیین ساختار سوم، به سادگی تعیین ساختار اول پروتئین نیست. روش‌های فعلی تعیین ساختار سوم، بسیار پر هزینه و زمان‌بر هستند. در نتیجه محققان بر روی روش‌هایی کار می‌کنند که بتوانند ساختار سوم پروتئین را صرفاً بر اساس توالی اسید آمینه آن پیش‌بینی نمایند. پیش‌بینی نقشه تماس، یکی از این روش‌های است. با داشتن نقشه تماس، می‌توان ساختار سوم را پیش‌بینی نمود. نقشه تماس پروتئین، یک نمایش ساده شده و دوبعدی از ساختار فضائی پروتئین است. هدف در مساله پیش‌بینی نقشه تماس، محاسبه تقریبی نقشه تماس یک پروتئین با استفاده از توالی اسید آمینه آن و ویژگی‌هایی است که صرفاً از روی توالی قابل محاسبه و یا پیش‌بینی هستند. رویکردهای آماری و یادگیری ماشین متعددی برای پیش‌بینی نقشه تماس ارائه شده است.

ماشین گروهی، یک روش یادگیری ماشین است که در آن وظیفه یادگیری میان چند یادگیر و فضای ورودی به چند زیرفضا تقسیم می‌شود. پاسخ یادگیرها به یک ورودی، به نحوی با یکدیگر ترکیب شده و پاسخ نهایی سیستم را تشکیل می‌دهند. این پاسخ، دقیق‌تر از پاسخ هر یک از یادگیرهاست. هدف این تحقیق، ارائه یک روش نوین پیش‌بینی نقشه تماس بر اساس ایده ماشین گروهی است. گروه یادگیر در روش پیشنهادی، مجموعه‌ای از شبکه‌های عصبی است. ویژگی‌هایی متعددی برای آموزش سیستم استخراج می‌شوند. سپس یک گروه از شبکه‌های عصبی به عنوان مدل پیش‌بینی کننده ایجاد می‌گردد. معیار مهم در ارزیابی پیش‌بینی نقشه تماس، نسبت تماس‌های درست پیش‌بینی شده به تعداد کل تماس‌های پیش‌بینی شده است.

برای تحلیل نتایج مدل پیشنهادی، دو روش دیگر نیز پیاده‌سازی و نتایج آن‌ها مقایسه شده است. نتایج، نشان‌دهنده کارائی روش ماشین‌گروهی در مسئله پیش‌بینی نقشه تماس است.

واژه‌های کلیدی: 1- بیوانفورماتیک 2- یادگیری ماشین 3- ماشین‌گروهی 4- شبکه عصبی 5- نقشه تماس پرونین 6- پیش‌بینی نقشه تماس

فصل اول

مقدمه

1-1 بیوانفورماتیک¹

بیوانفورماتیک علمی است بینرشته‌ای که قواعد ریاضی، فیزیک، شیمی و علوم کامپیوتر را به داده‌های وسیع، متنوع و پیچیده زیستشناسی، اعمال می‌کند. هدف بیوانفورماتیک، حل مسائل زیستشناسی در سطح مولکولی است. دو اصطلاح بیوانفورماتیک و زیستشناسی محاسباتی²، معمولاً به جای یکدیگر به کار می‌روند. هر چند این دو، فعالیت‌های مشترک بسیاری دارند، دو زمینه مجزا از یکدیگر در نظر گرفته می‌شوند. زیستشناسی محاسباتی، مسائل زیستی را توسط فرضیه‌ها و با استفاده از داده‌های تجربی و یا شبیه‌سازی شده بررسی می‌نماید. در واقع، بیوانفورماتیک بر اطلاعات مرکز از داده‌های زیستشناسی محاسباتی بر فرضیه‌ها. تعاریف زیر را برای این دو رشته می‌توان در نظر گرفت:

- بیوانفورماتیک: توسعه یا کاربرد ابزارها و روش‌های محاسباتی به منظور بهره‌برداری بیشتر از داده‌های زیستی و پزشکی است، شامل به دست آوردن³، ذخیره، سازماندهی، تحلیل و نمایش⁴ دادن این داده‌ها.

¹ Bioinformatics

² Computational biology

³ Acquire

⁴ Visualize

• زیست‌شناسی محاسباتی: توسعه و کاربرد روش‌های تئوری و تحلیل داده، مدل‌سازی ریاضی و تکنیک‌های شبیه‌سازی محاسباتی، برای مطالعه سیستم‌های زیستی و اجتماعی¹ است. دو زمینه اصلی که در حوزه بیوانفورماتیک به آن پرداخته می‌شود، ژنومیک و پروتئومیک هستند. ژنومیک شامل تجزیه و تحلیل داده‌های ژنوم موجودات است. ژنوم، توالی کل DNA موجود در سلول‌های یک جاندار است که به عنوان ماده ژنتیکی عمل نموده و سبب بروز صفات وراثتی می‌شود. به طور خلاصه می‌توان گفت ژنومیک شامل توالی‌بایی و تحلیل ژن‌ها در یک موجود زنده است. پروتئومیک به بررسی پروتئین‌های یک موجود زنده گفته می‌شود. علاوه بر ژنومیک و پروتئومیک، شاخه‌های دیگری از علوم زیستی وجود دارند که در بیوانفورماتیک به آن‌ها پرداخته می‌شود، مانند ترانسکریپتومیک² و متابولومیک³. در هر یک از این بخش‌ها سعی می‌شود تا به بخشی از سوالات و پیچیدگی‌های علم زیست‌شناسی پاسخ داده شود [2].

1-1-1 اهداف بیوانفورماتیک

به طور کلی بیوانفورماتیک سه هدف عمده دارد. اولاً بیوانفورماتیک داده‌ها را به گونه‌ای سازمان‌دهی می‌کند که محققان بتوانند به آسانی به اطلاعات موجود دسترسی داشته باشند و نیز نتایج جدید را ارسال نمایند، مانند بانک داده پروتئین⁴ (PDB). هدف دوم توسعه ابزارهایی برای کمک به تحلیل این داده‌هاست. مثالی از تحلیل داده‌ها، مقایسه توالی یک پروتئین با دیگر پروتئین‌های است. توسعه چنین ابزارهایی نیازمند تخصص هم در زمینه محاسبات و هم در زمینه زیست‌شناسی است. هدف سوم استفاده از این ابزارها برای تحلیل داده‌ها، کشف الگوهای جدید و تفسیر نتایج به صورتی است که از نظر زیست‌شناسی بامعا باشد. به عنوان مثال می‌توان پیش‌بینی ساختار پروتئین را نام برد [2].

1-1-2 کاربرد روش‌های محاسباتی در بیوانفورماتیک

امروزه، تقریباً تمامی شاخه‌های علم از روش‌های ریاضی، به عنوان بخشی از ابزارهای تحقیق خود استفاده می‌کنند. می‌توان گفت اکثر روش‌های ریاضی کاربردی در بیوانفورماتیک به‌کار گرفته می‌شوند. در میان آن‌ها، تئوری آمار و احتمال و الگوریتم‌های علوم کامپیوتر دارای اهمیت ویژه هستند. آمار و احتمال نه تنها ابزارهای تحقیقات بیوانفورماتیک به شمار می‌روند، بلکه زبانی برای فرمول‌بندی کردن

¹ Social

² Transcriptomic

³ Metabolomic

⁴ Protein data bank

نتایج نیز هستند. الگوریتم‌های علوم کامپیوتر بستر تکنیکی بیوانفورماتیک را فراهم می‌نمایند. دیگر ابزارهای محاسباتی، مانند روش‌های بهینه‌سازی و روش‌های تحلیل الگو، در مرتبسازی و مدل‌سازی مکانیزم‌های زیستی نقش مهمی ایفا می‌کنند [3].

1-1-3 کاربردهای بیوانفورماتیک

اطلاعات به دست آمده از تحلیل داده‌های زیستی توسط علم بیوانفورماتیک، در موارد بسیاری به زیست‌شناسی کمک می‌کنند، از جمله یافتن ژن‌ها در میان توالی‌های ژنومیک، پیشگویی ساختار و عملکرد محصولات ژن‌ها، توضیح تعامل ژن با محصولاتش و یافتن ارتباط میان ژن‌ها و توالی‌های پروتئینی. پکی از مهمترین کاربردهای بیوانفورماتیک در علوم پزشکی است که در این زمینه می‌توان به طراحی داروها، شناسایی علل ژنتیکی بیماری‌ها و تشخیص بیماری‌ها بر اساس اطلاعات ژنتیکی اشاره نمود. ضرورت بهکارگیری بیوانفورماتیک در مطالعات پزشکی تا حدی است که دانشمندان معتقدند بدون استفاده از بیوانفورماتیک تحقیقات دارویی و زیست‌شناسی مدرن، متوقف خواهد شد [2].

1-2 پیش‌بینی نقشه تماس پروتئین¹

پروتئین‌ها از اجزایی اصلی سلول‌های موجودات زنده هستند. هر مولکول پروتئین از تعدادی اسید آمینه تشکیل می‌شود. برای پروتئین چهار ساختار (اول، دوم، سوم، چهارم) تعریف شده است. ساختار اول²، همان زنجیره اسیدهای آمینه آن است. ساختارهای دوم³، ساختارهای محلی⁴ هستند که توسط برقراری پیوندهای نئدروژنی به وجود می‌آیند. رایج‌ترین آن‌ها، مارپیچ‌های آلفا و صفحه‌های بتا هستند. ساختار سوم⁵، شکل کلی یک مولکول پروتئین و در واقع، موقعیت فضائی ساختارهای دوم نسبت به یکدیگر است که در اثر تاشدن⁶ زنجیره اسیدهای آمینه شکل می‌گیرد. ساختار چهارم⁷ از تجمع چندین پروتئین ایجاد می‌گردد. محققان، پیوسته پروتئین‌های جدیدی کشف و توالی اسیدهای آمینه آن‌ها را تعیین می‌کنند. عمل پروتئین، وابسته به شکل ساختار سوم آن است. مولکول‌هایی که یک پروتئین می‌تواند به آن‌ها متصل شود، بستگی به شکل سبعدی پروتئین دارند. ساختار سوم خود، وابسته به توالی اسید آمینه است.

¹ Protein contact map Prediction

² Primary structure

³ Secondary structures

⁴ Local

⁵ Tertiary structure

⁶ Folding

⁷ Quaternary structure

متاسفانه، تعیین ساختار سوم، به سادگی تعیین توالی پروتئین نیست. روش‌های فعلی تعیین ساختار سوم، بسیار پر هزینه و زمانبر هستند. در نتیجه محققان بر روی روش‌هایی کار می‌کنند که بتوانند ساختار سوم پروتئین را صرفا بر اساس توالی اسید آمینه آن پیش‌بینی نمایند. پیش‌بینی نقشه تماس، یکی از این روش‌های است. با داشتن نقشه تماس، می‌توان ساختار سوم را با دقت مناسب پیش‌بینی نمود. نقشه تماس پروتئین، یک نمایش ساده شده و دو بعدی از ساختار فضایی پروتئین است. هدف در مساله پیش‌بینی نقشه تماس، محاسبه تقریبی نقشه تماس یک پروتئین با استفاده از توالی اسید آمینه آن (ساختار اول) و ویژگی‌هایی¹ است که صرفا از روی توالی قابل محاسبه و یا پیش‌بینی هستند. رویکردهای آماری و یادگیری ماشین² متعددی برای پیش‌بینی نقشه تماس ارائه شده، از جمله مدل پنهان مارکوف³ (HMM)، شبکه‌های عصبی مصنوعی⁴ (ANN)، الگوریتم ژنتیک (GA)، ماشین‌های بردار پشتیبان⁵ (SVM) و قوانین وابستگی⁶ (AR).

روش ماشین گروهی⁷، یک روش یادگیری ماشین است که در آن وظیفه یادگیری میان چند یادگیر و فضای ورودی به چند زیرفضا تقسیم می‌شود. پاسخ یادگیرها به یک ورودی، به نحوی با یکدیگر ترکیب شده و پاسخ نهائی سیستم را تشکیل می‌دهند. کارائی یک ماشین گروهی، بهتر از کارائی هر یک از اعضایش، بعنهایی، است. در این تحقیق از یک روش ماشین گروهی برای پیش‌بینی نقشه تماس استفاده شده است. گروه یادگیر، مجموعه‌ای از شبکه‌های عصبی است. برای آموزش شبکه‌های عصبی، ویژگی‌هایی متعددی از روی توالی پروتئین‌ها استخراج می‌شوند. سپس یک گروه از شبکه‌ها، به عنوان مدل پیش‌بینی‌کننده ایجاد می‌گردد. ایجاد گروه به صورت سریال صورت می‌گیرد، بدین معنی که ابتدا یک شبکه عصبی آموزش دیده و به عنوان اولین عضو گروه قرار می‌گیرد. سپس شبکه‌های دیگر در چند تکرار به گروه اضافه می‌شوند. در هر تکرار یک شبکه جدید ایجاد شده و آموزش داده می‌شود. سپس کارائی گروه با در نظر گرفتن و بدون در نظر گرفتن شبکه جدید، محاسبه می‌گردد. این شبکه به شرطی به گروه اضافه می‌شود که بتواند میانگین کارائی گروه را افزایش دهد. این پروسه ادامه می‌یابد تا جایی که برای یک تعداد تکرار از پیش‌تعیین شده، شبکه‌های عصبی جدید آموزش دیده نتوانند کارائی گروه را افزایش داده و به گروه اضافه شوند. آموزش هر شبکه جدید با یک زیرمجموعه تصادفی از داده‌های

¹ Features

² Machine learning

³ Hidden Markov model

⁴ Artificial Neural networks

⁵ Support vector machines

⁶ Association rules

⁷ Committee machine

آموزشی صورت می‌گیرد. تست مدل بر روی مجموعه داده‌های تست که 20% کل داده‌ها را تشکیل می‌دهند، انجام می‌شود. معیار مهم در ارزیابی پیش‌بینی نقشه تماس، نسبت تماس‌های پیش‌بینی شده صحیح به تعداد کل تماس‌های پیش‌بینی شده است. بر این اساس، دقت پیش‌بینی برای هر پروتئین موجود در مجموعه تست محاسبه و میانگین دقت‌ها به عنوان میزان کارائی سیستم گزارش می‌شود. برای تحلیل نتایج مدل پیشنهادی، دو مدل دیگر نیز پیاده‌سازی و نتایج آن‌ها مقایسه شده است. بررسی نتایج به دست آمده از این سه مدل، کارائی روش ماشین گروهی در مسئله پیش‌بینی نقشه تماس را نشان می‌دهد.

1-3 ساختار گزارش

ساختار گزارش تحقیق به صورت زیر است. در فصل دوم مفاهیم ژنتیکی پیش‌زمینه لازم شامل DNA، ژن و ژنوم، کروموزوم، RNA، کد ژنتیکی، پروتئین و قضیه مرکزی زیست مولکولی شرح داده می‌شوند. فصل سوم شامل مفاهیم محاسباتی پیش‌زمینه لازم است. مباحث این فصل، به دو بخش دسته مفاهیم بیوانفورماتیکی و مفاهیم یادگیری ماشین، تئوری اطلاعات و آمار و احتمال تقسیم می‌گردند. در بخش اول پایگاه داده PDB، فرمت فایل PDB، لیست pdbselect، فرمت فایل FASTA، تراز توالي¹ (جفتی و چندگانه)، جهش وابسته²، پیش‌بینی ساختار دوم و پایگاه داده AAindex و در بخش دوم روش‌های ردبهندی³، شبکه‌های عصبی، تخمين دقت، اثر فراموشی خطرناک⁴، یادگیری گروهی و چند مفهوم تئوری اطلاعات و آمار و احتمال معرفی می‌شوند. در فصل چهارم نقشه تماس پروتئین و ویژگی‌های آن معرفی شده و سپس مساله پیش‌بینی نقشه تماس، اهمیت آن، نحوه ارزیابی پیش‌بینی بررسی می‌گردد. سپس در فصل پنجم، مروری بر کارهای انجام شده در زمینه پیش‌بینی نقشه تماس صورت می‌گیرد. فصل ششم به شرح روش پیشنهادی برای پیش‌بینی نقشه تماس می‌پردازد. این فصل شامل نحوه آماده‌سازی فایل‌های داده، انواع ویژگی‌های مورد استفاده در شبکه‌های عصبی و نحوه استخراج آن‌ها، آماده‌سازی داده‌های آموزشی و تست شبکه‌های عصبی، نحوه ارزیابی کارائی مدل، نحوه آموزش و تست مدل و تحلیل نتایج و مقایسه آن‌ها با نتایج دو روش پیاده‌سازی شده دیگر است. در پایان، فصل هفتم، به جمع‌بندی و نتیجه‌گیری پرداخته و پیشنهاداتی برای ادامه کار مطرح می‌نماید.

¹ Sequence alignment

² Correlated mutation

³ Classification

⁴ Catastrophic forgetting effect

4-1 دستاوردهای پژوهشی تحقیق

- Narjes Khatoon Habibi, Kaveh Mahdaviani, Mohammad Hossein Saraee, “Mining Protein Primary Structure Data Using Committee Machines Approach to Predict Protein Contact Map”, The 4th IEEE International Conference on Emerging Technologies (ICET-08), 18-19 October 2008, Rawalpindi, Pakistan, published.
- Narjes Khatoon Habibi, Mohammad Hossein Saraee, “Protein Contact Map Prediction Based on an Ensemble Learning Method”, IEEE International Conference on Computer Engineering and Technology (ICCET 2009), 22-24 January 2009, Singapore, published.

فصل دوم

مفاهیم زیستی پیش‌زمینه

2-1 مقدمه

در ابتدا لازم است در مورد مفاهیم، اصطلاحات و پروسه‌های اصلی ژنتیک مولکولی توضیح داده شود. موجودات زنده از سلول‌ها ایجاد شده‌اند. اطلاعات مورد نیاز برای ساختن هر موجود زنده، ژنوم¹ آن موجود نامیده می‌شود. تمام فعالیت‌های یک سلول توسط پروتئین‌ها انجام می‌شود. پروتئین‌های هر سلول بر حسب اطلاعات ژنتیکی موجود در ژنوم ساخته می‌شوند. در این بخش ابتدا ژنوم و نحوه ساخت پروتئین‌ها، سپس اجزای تشکیل‌دهنده و ساختار پروتئین‌ها، و در پایان قضیه مرکزی زیست مولکولی² شرح داده شده‌اند.

2-2 سلول

هر موجود زنده از ساختار‌های کوچکی به نام سلول ایجاد شده است. هر سلول خود سیستم پیچیده‌ایست که درون یک غشا³ قرار دارد. بعضی موجودات زنده مانند باکتری‌ها و مخمر نان⁴، تنها از یک سلول و برخی دیگر از چندین نوع سلول تشکیل شده‌اند. به عنوان مثال بدن انسان به طور تقریبی از 60

¹ Genome

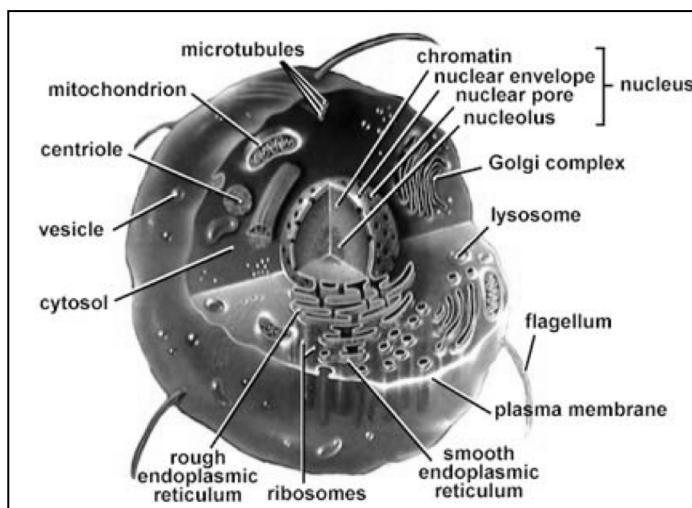
² Central dogma of molecular biology

³ Membrane

⁴ Yeast

تریلیون سلول ساخته شده است. این سلول‌ها مشتمل بر 320 نوع مختلف هستند که هر یک ساختار و عملکرد خاصی دارند. موجودات را از نظر تکامل پیچیدگی به دو دسته پروکاریوت^۱ (که ابتدائی‌تر هستند) و یوکاریوت^۲ (که پیشرفته‌تر هستند) تقسیم‌بندی می‌نمایند.

هر سلول یوکاریوت یک هسته دارد که توسط غشا از بقیه سلول جدا می‌شود. درون هسته کروموزوم‌ها قرار دارند که تمام اطلاعات ژنتیکی موجود زنده در آن‌ها ذخیره شده است. از دیگر اجزای سلول‌های یوکاریوتی می‌توان سانتریول^۳، لیزوژوم^۴، میتوکندری^۵ و ریبوژوم^۶ را نام برد. شکل [4] ساختار یک سلول یوکاریوتی را نشان می‌دهد.



شکل 2-1) یک سلول یوکاریوتی [4]

DNA 2-3

DNA یا اسید دزونوکلئیک⁷، اساس بلوک‌های رمزکننده حیات است. یک مولکول DNA تکرشته‌ای، زنجیرهای از مولکول‌های کوچک به نام نوکلئوتید و در واقع یک پلی‌نوکلئوتید است. هر نوکلئوتید (در DNA)، از یک قند پنج کربنی داکسی‌ریبوز، یک گروه فسفات و یک باز آلی تشکیل شده است. چهار نوع باز آلی با نام‌های آدنین⁸ (A)، سیتوزین⁹ (C)، گوانین¹ (G) و تیمین² (T) وجود دارند.

¹ Prokaryote

² Eukaryote

³ Centriole

⁴ Lysosome

⁵ Mitochondria

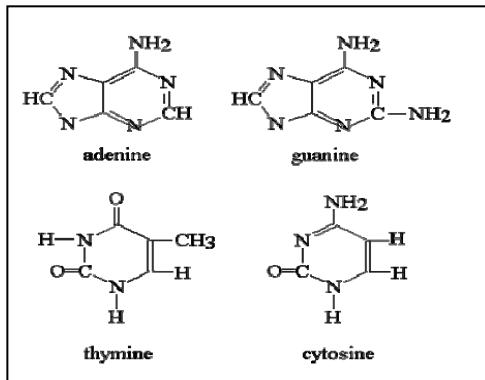
⁶ Ribosome

⁷ Deoxyribonucleic

⁸ Adenosine

⁹ Cytosine

شکل (2-2) ساختار شیمیائی بازهای آلي را نشان میدهد (ساختار نوکلئوتیدهای RNA با ساختار نوکلئوتیدهای DNA متفاوت است که در بخش (6-2) توضیح داده شده است).



شکل 2-2) چهار باز آلي سازنده DNA [5]

اطلاعات لازم برای ایجاد یک موجود زنده، ساده یا پیچیده، یک رشته مشکل از الفبای چهار حرفي بازهای آلي است. از کنار هم قرار گرفتن نوکلئوتیدهای مختلف، یک پلی‌نوکلئوتید به وجود می‌آید. انتهایی پلی‌نوکلئوتیدها از نظر شیمیائی متفاوت و به عبارتی پلی‌نوکلئوتیدها جهتدار هستند. انتهاي توالي به وسیله علامت¹ 3' یا 5' مشخص می‌شود. قرارداد برچسبگذاري توالي از 5' به 3' (از چپ به راست) است. برای نمونه یک پلی‌نوکلئوتید در زیر مشاهده می‌شود:

5' G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3'

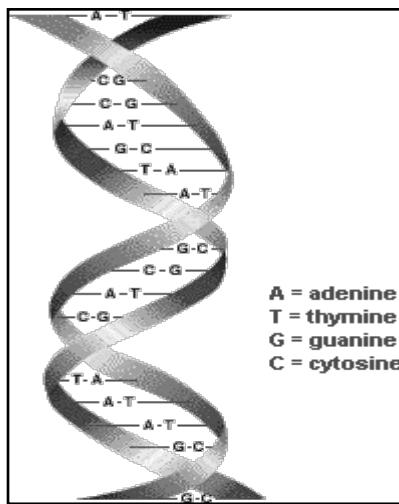
مي تواند تک رشته‌اي يا دو رشته‌اي باشد. در يك DNA دو رشته‌اي، به رشته دوم ((رشته مکمل معکوس)) گفته می‌شود. علت اين است که جهت رشته دوم در جهت عکس رشته اول و بازهای آن، مکمل بازهای رشته اول‌اند. بازهای مکمل به بازهایی گفته می‌شود که می‌توانند با يكديگر پيوند تشکيل دهند. در DNA، A با T و C با G پيوند ایجاد می‌کنند. برای پلی‌نوکلئوتید مثال بالا، پلی‌نوکلئوتید دو رشته‌اي به صورت زير است:

5' G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3'
3' C←A←T←T←C←A←G←G←G←C←A←A←T←C←G 5'

¹ Guanine

² Thymine

دو رشته پلی‌نوکلئوتید مکمل، یک ساختار پایدار به نام مارپیچ دوگانه¹ ایجاد می‌کنند (شکل Rosalind DNA در سال 1953 توسط دکتر Jamse Watson، پروفسور Francis Crick، دکتر Maurice Wilkins و پروفسور Franklin کشف شد. در هر دور مارپیچ، 10 جفت باز وجود دارد.



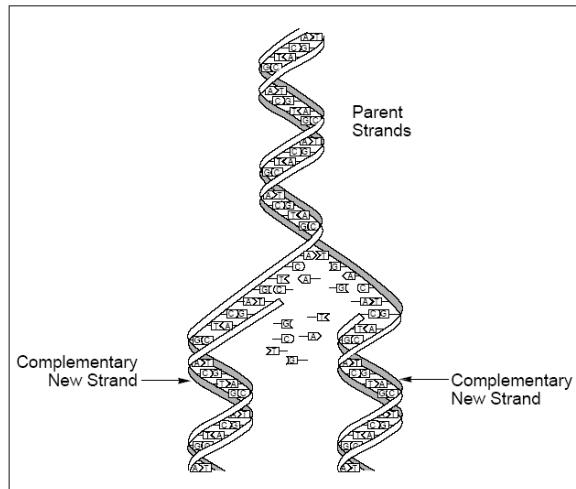
شکل 2-3) مارپیچ دوگانه [5] DNA

((همانندسازی DNA))² فرآیندی است که در آن دو رشته DNA از یکدیگر باز شده و یک رشته مکمل جدید، برای هر رشته مولکول DNA اصلی ایجاد می‌شود. بعد از اتمام پروسه، دو مولکول DNA مشابه مولکول اولیه، به وجود می‌آیند. همانندسازی DNA در طی تقسیم سلولی³ (تقسیم یک سلول به دو سلول دیگر) اتفاق می‌افتد و هر یک از دو مولکول DNA، به یکی از دو سلول حاصل انتقال می‌یابد [5]. شکل (4-2) فرآیند همانندسازی را نشان می‌دهد. همان‌طور که در شکل مشاهده می‌شود، پس از باز شدن دو رشته مارپیچ، با قرار گرفتن یک رشته مکمل معکوس در کنار هر یک، دو مارپیچ جدید ایجاد می‌شوند.

¹ Double helix DNA

² DNA replication

³ Cell division



شکل 2-4) فرآیند همانندسازی DNA [5]

2-4 ژن و ژنوم

هر مولکول DNA شامل ژنهای بسیاری است که واحدهای فیزیکی ارثبری به شمار می‌آیند. یک ژن، توالی خاص و مشخصی از نوکلئوتیدها بوده که در برگیرنده اطلاعات لازم برای ساختن پروتئین‌هاست. پروتئین‌ها اجزای سازنده سلول‌ها، بافت‌ها و آنزیمهای هستند و برای واکنش‌های بیولوژیکی، حیاتی‌اند. ژنوم یک موجود زنده، به تمامی ژن‌ها، به علاوه بخش‌های غیرمرمزگذار¹ مولکول‌های DNA آن موجود گفته می‌شود. تنها حدود ده درصد از ژنوم، حاوی توالی‌های کدگذار برای پروتئین‌ها هستند. این قسمت‌ها را اصطلاحاً اگزون² می‌نامند. در مقابل، اینtron‌ها³ که در بسیاری از ژن‌ها به صورت پراکنده وجود دارند، قابلیت کدگذاری پروتئینی را ندارند. اندازه ژنوم با تعداد بازهای آن مشخص می‌گردد. ژنوم انسان شامل بیش از 3.2 بیلیون جفت باز و بیش از 30000 ژن است. انسان می‌تواند در حدود 100000 نوع پروتئین تولید کند [4].

2-5 کروموزوم

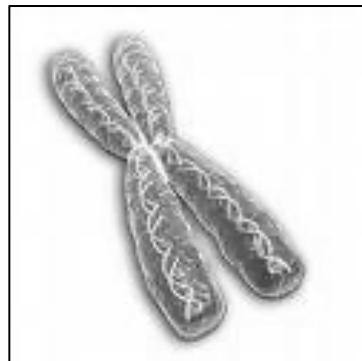
درون هسته، یک یا چند مولکول دو رشته‌ای DNA وجود دارد که در انسان در 23 واحد سازماندهی شده‌اند. این واحدهای میکروسکوپی که از نظر فیزیکی مجزا هستند، کروموزوم نام دارند. یک کروموزوم در شکل (5-2) نمایش داده شده است. تمامی ژن‌ها بر روی کروموزم‌ها قرار دارند. هسته سلول در انسان‌ها دارای دو مجموعه کروموزوم بوده که هر مجموعه از یکی از والدین گرفته می‌شود.

¹ Non-coding

² Exon

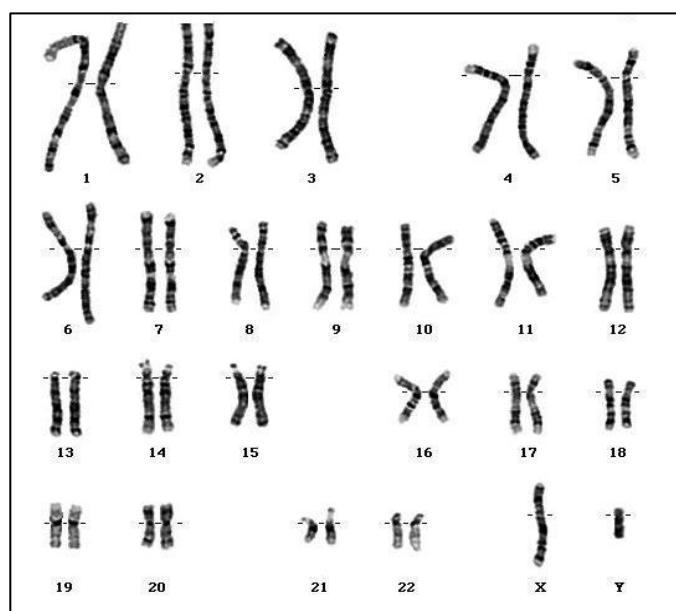
³ Intron

هر مجموعه دارای 23 جفت کروموزوم است که 22 جفت از آن‌ها غیرجنسی¹ و 1 جفت کروموزوم جنسیت X و یا Y می‌باشد (یک زن دارای یک جفت کروموزم X و یک مرد دارای یک کروموزم Y و یک کروموزم X است).



شکل 2-5) کروموزوم [5]

کروموزوم‌ها را می‌توان در زیر یک میکروسکوپ معمولی مشاهده کرد. وقتی کروموزوم‌ها به رنگ خاصی آغشته شوند، الگوهای تیره و روشنی مشاهده خواهد شد که بیانگر تغییرات در میزان پیوندهای A-T نسبت به C-G می‌باشند. توع در اندازه و الگوهای پیوندی موجود در کروموزوم‌ها، آن‌ها را از یکدیگر تمیز می‌دهد. به این تحلیل، کاریوتیپ² گفته می‌شود. شکل 2-6) کاریوتیپ یک انسان سالم را نشان می‌دهد.



¹ Autosome

² Karyotype

شکل 2-6) کاریوتیپ یک مرد سالم [5]

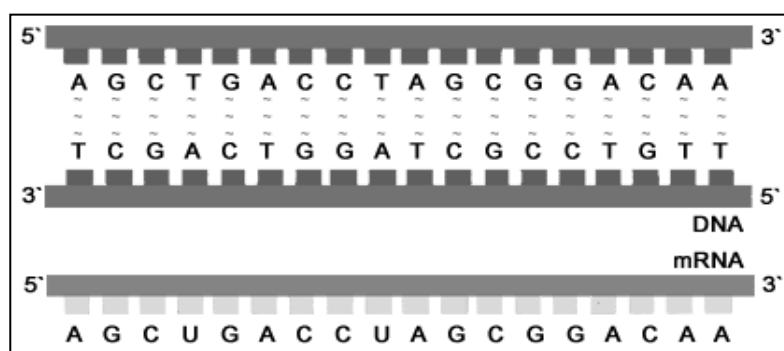
موجوداتی مانند انسان، دارای یک جفت از هر نوع کروموزوم هستند. در این دسته از موجودات که اصطلاحاً به آنها دیپلوايد¹ گفته می‌شود، هر یک از جفت‌های مربوط به یک نوع کروموزوم، از یکی از والدین به ارث برده می‌شود. ژن‌های متناظر در یک جفت کروموزوم، ممکن است دقیقاً همانند و یا با یکدیگر متفاوت باشند [5].

RNA 2-6

RNA یا اسید ریبونوکلئیک² مولکولی است که مشابه DNA از نوکلئوتیدها تشکیل شده است، با این تفاوت که به جای قند داکسی‌ریبوز، قند ریبوز و به جای باز تیمین (T)، بازی به نام یوراسیل³ (U) در ساختار آن وجود دارد. RNA به صورت تکرشته‌ای، دورشته‌ای و به عنوان بخشی از یک مارپیچ ترکیبی DNA-RNA در سلول وجود دارد. دو نوع اصلی مولکول RNA که در ساخت پروتئین دخالت دارند، tRNA و mRNA هستند [5].

mRNA 2-6-1

mRNA، کپی قسمتی از اطلاعات ژنتیکی مولکول‌های DNA است. ((نسخه‌برداری))⁴ فرآیندی است که در آن بخشی از DNA، به صورت یک مولکول mRNA نسخه‌برداری می‌شود. mRNA حاصل، مکمل ایست که نسخه‌برداری از روی آن انجام می‌شود. شکل (7-2) مولکول mRNA ساخته شده از روی یک DNA فرضی را نشان می‌دهد.



¹ Diploid

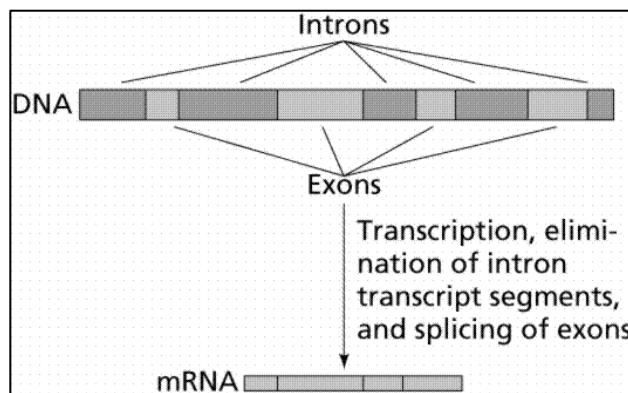
² Ribonucleic

³ Uracil

⁴ Transcription

شکل 7-2) مولکول mRNA ساخته شده از روی یک DNA فرضی [5]

در سلول های یوکاریوتی، پیش از ترجمه mRNA به یک پروتئین، طی فرآیند ((پردازش))¹، نواحی اینtron حذف و اگزون ها به یکدیگر متصل می شوند (شکل 8-2). mRNA تغییر یافته، از هسته خارج و به یک توالي پروتئین ترجمه² می گردد [5].



شکل 8-2) فرآیند پردازش mRNA [5]

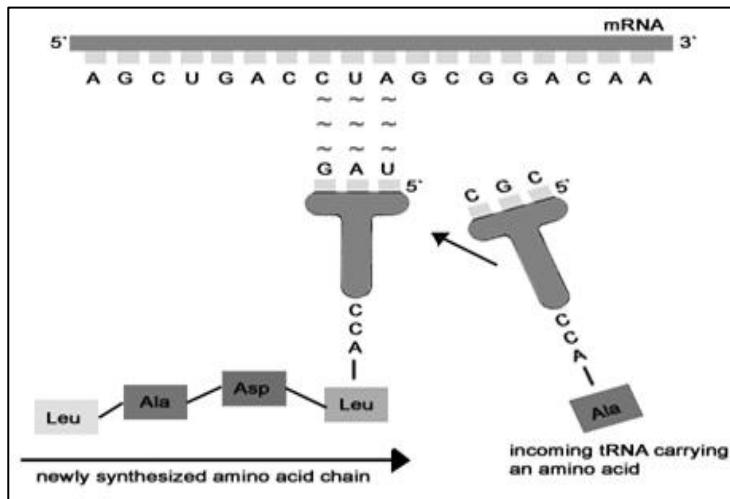
tRNA 2-6-2

tRNA مولکولی سمبودی است که برای ساخت پروتئین ضروری می باشد. به هر مولکول tRNA، یک اسید آمینه متصل می شود. اسیدهای آمینه در بخش (1-8-2) توضیح داده شده اند. نوع اسید آمینه، توسط یک توالي باز سهتائی به نام توالي ((آنتی کدون))³ یا کد ژنتیکی، تعیین می شود. توالي آنتی کدون، مکمل توالي mRNA متناظر است. ترجمه، پروسه ایست که در آن توالي نوکلئوتیدی mRNA، برای تعیین نوع و ترتیب اسید آمینه هایی که باید به یکدیگر متصل شوند، به وسیله tRNA پردازش شده و با کمک ریبوزوم ها، پروتئین ساخته می شود. شکل (2-9) پروسه ترجمه یک مولکول mRNA فرضی را به طور ساده نمایش می دهد [5].

¹ Splicing

² Translate

³ Anti-codon



شکل 2-9) پروسه ترجمه یک mRNA به پروتئین متناظر [5]

2-7 کد ژنتیکی

در بخش (2-6-2)، گفته شد که آنتیکدون یک توالی سهتایی از نوکلئوتیدهای است. به دلیل وجود 4 نوکلئوتید، 64 توالی ($4^4 = 64$) آنتیکدون وجود دارد. آنتیکدون AUG نشانگر آغاز ترجمه و آنتیکدون‌های UAA، UAG و UGA نشانگر پایان ترجمه‌اند. AUG می‌تواند یک اسید آمینه را نیز رمزگذاری کند. در نتیجه 61 ($64 - 3 = 61$) آنتیکدون برای رمزگذاری پروتئین‌ها وجود دارد. اما تنها 20 اسید آمینه در موجودات زنده یافت می‌شوند. پس کد ژنتیکی تکراری¹ (افزونه) است، یعنی یک اسید آمینه می‌تواند توسط چند آنتیکدون مختلف رمز شود. جدول (2-1)، فهرست آنتیکدون‌ها و پروتئین‌هایی که هر یک از آن‌ها رمزگذاری می‌کنند را نشان می‌دهد [5].

جدول 2-1) فهرست آنتیکدون‌ها و پروتئین‌هایی که هر یک رمز می‌کنند [5].

¹ Redundant

		Second base			
		U	C	A	G
First base	U	UUU (Phe/F)Phenylalanine UUC (Phe/F)Phenylalanine UUA (Leu/L)Leucine UUG (Leu/L)Leucine, Start	UCU (Ser/S)Serine UCC (Ser/S)Serine UCA (Ser/S)Serine UCG (Ser/S)Serine	UAU (Tyr/Y)Tyrosine UAC (Tyr/Y)Tyrosine UAA Stop UAG Stop	UGU (Cys/C)Cysteine UGC (Cys/C)Cysteine UGA Stop UGG (Trp/W)Tryptophan
	C	CUU (Leu/L)Leucine CUC (Leu/L)Leucine CUA (Leu/L)Leucine CUG (Leu/L)Leucine, Start	CCU (Pro/P)Proline CCC (Pro/P)Proline CCA (Pro/P)Proline CCG (Pro/P)Proline	CAU (His/H)Histidine CAC (His/H)Histidine CAA (Gln/Q)Glutamine CAG (Gln/Q)Glutamine	CGU (Arg/R)Arginine CGC (Arg/R)Arginine CGA (Arg/R)Arginine CGG (Arg/R)Arginine
	A	AUU (Ile/I)Isoleucine, AUC (Ile/I)Isoleucine AUA (Ile/I)Isoleucine AUG (Met/M)Methionine, Start	ACU (Thr/T)Threonine ACC (Thr/T)Threonine ACA (Thr/T)Threonine ACG (Thr/T)Threonine	AAU (Asn/N)Asparagine AAC (Asn/N)Asparagine AAA (Lys/K)Lysine AAG (Lys/K)Lysine	AGU (Ser/S)Serine AGC (Ser/S)Serine AGA (Arg/R)Arginine AGG (Arg/R)Arginine
	G	GUU (Val/V)Valine GUC (Val/V)Valine GUA (Val/V)Valine GUG (Val/V)Valine,	GCU (Ala/A)Alanine GCC (Ala/A)Alanine GCA (Ala/A)Alanine GCG (Ala/A)Alanine	GAU (Asp/D)Aspartic acid GAC (Asp/D)Aspartic acid GAA (Glu/E)Glutamic acid GAG (Glu/E)Glutamic acid	GGU (Gly/G)Glycine GGC (Gly/G)Glycine GGA (Gly/G)Glycine GGG (Gly/G)Glycine

2-پروتئین

پروتئین، زنجیره‌ای از اسیدهای آمینه است. طول این زنجیره معمولاً کمتر از 2000 واحد و بسیار کوتاه‌تر از مولکول‌های اسید نوکلئیک موجود در طبیعت است. تمام اسیدهای آمینه دارای چند قسمت مشابه هستند. این قسمت‌های مشابه، به یکدیگر متصل شده و ستون فقرات¹ پروتئین را می‌سازند. همان‌طور که گفته شد، ساخت یک مولکول پروتئین از روی ژن متناظر آن، شامل سه مرحله است: نسخه‌برداری، پردازش و ترجمه.

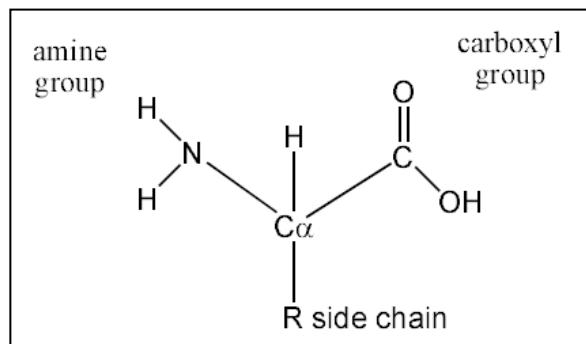
2-اسیدهای آمینه

تعداد 20 نوع اسید آمینه مختلف در مولکول‌های پروتئین یافت می‌شوند. شکل‌های (2-10) و (2-11) ساختار شیمیائی اسیدهای آمینه را نشان می‌دهند. هر اسید شامل یک اتم کربن مرکزی (C_{α}) است که به آن چهار گروه متصل شده‌اند [6]:

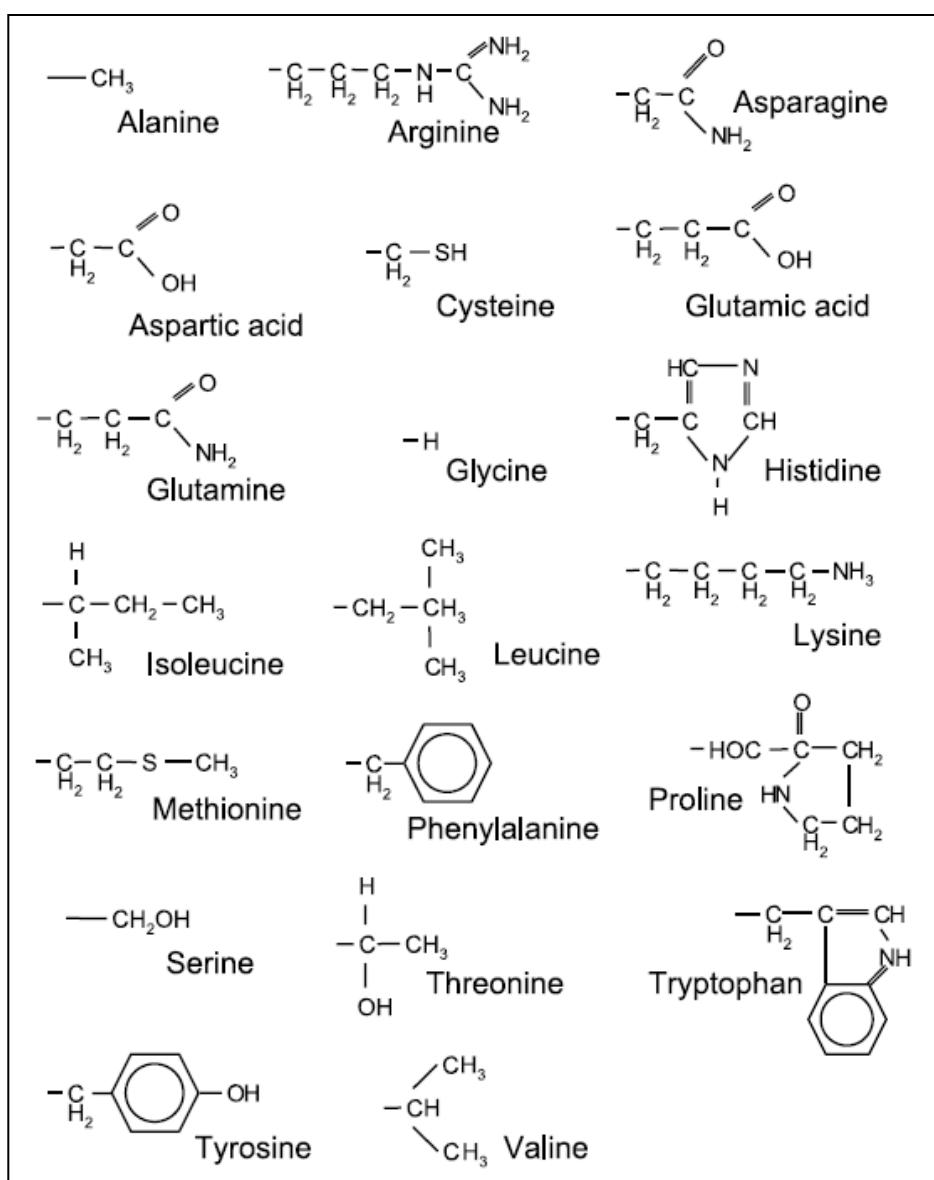
- یک اتم نیتروژن
- یک گروه کربوکسیل (-COO⁻)
- یک گروه آمینو (-NH⁺)₃
- گروه R که در هر اسید آمینه متفاوت است و به آن زنجیره جانبی² نیز گفته می‌شود.

¹ Backbone

² Side chain



شکل 2-10) ساختار عمومی یک اسید آمینه [3]



شکل-2(11) ساختار شیمیائی 20 اسید آمینه سازنده پروتئین‌ها [3]

هر یک از اسیدهای آمینه علاوه بر نام، دارای یک نام اختصاری سه حرفی و یک کد یک حرفی است.

این مشخصات در جدول (2-2) نشان داده شده‌اند.

Amino acid	Abbreviation	Symbol
Alanine	ALA	A
Arginine	ARG	R
Asparagine	ASN	N
Aspartic acid	ASP	D
Cysteine	CYS	C
Glutamine	GLN	Q
Glutamic acid	GLU	E
Glycine	GLY	G
Histidine	HIS	H
Isoleucine	ILE	I
Leucine	LEU	L
Lysine	LYS	K
Methionine	MET	M
Phenylalanine	PHE	F
Proline	PRO	P
Serine	SER	S
Threonine	THR	T
Tryptophan	TRP	W
Tyrosine	TYR	Y
Valine	VAL	V

جدول (2-2) کد یک حرفی، کد سه حرفی و نام کامل اسیدهای آمینه [3]

2-8-2 ویژگی‌های اسیدهای آمینه

ویژگی‌های اسیدهای آمینه توسط خصوصیات گروه‌های R آن‌ها مشخص می‌شود. گروه‌های R از نظر اندازه، ساختار، خصوصیات الکتریکی و واکنش‌پذیری شیمیائی، تفاوت دارند. خصوصیات مهم گروه‌های R، که نحوه واکنش آن‌ها با دیگر ترکیبات شیمیائی و مولکول‌های آب مجاور را تعیین می‌کنند، بار الکتریکی و قطبیت (آب‌دوست یا چرب‌دوست بودن زنجیره جانبی) هستند. قطبیت، در اثر خاصیت الکترونگاتیوی اتم‌ها ایجاد می‌شود. مثلاً یک مولکول آب، از نظر بار الکتریکی خنثی است، اما به دلیل این‌که اکسیژن الکترونگاتیوئر از نیدروژن است، مولکول حالت قطبی پیدا می‌کند. اسیدهای آمینه، بر اساس قطبیت (در pH 6 تا 7) و بار الکتریکی به چند دسته تقسیم می‌شوند [3]:

- اسیدهای آمینه غیرقطبی: غیرقطبی بودن، از توزیع متوازن بارهای الکتریکی + و - بر روی گروه R ناشی می‌شود.

اسیدهای آمینه غیرقطبی با مولکول‌های آب پیوند برقرار

نمی‌کند و به آن‌ها آبگریز¹ یا داخلی² گفته می‌شود. اسیدهای آمینه آبگریز، یک هسته آبگریز در فضای داخلی مولکول پروتئین تشکیل می‌دهند و با مولکول‌های آب مجاور تماس ندارند. اسیدهای آمینه غیرقطبی بیشترین اجزای تشکیل دهنده پروتئین‌های موجودات زنده‌اند. tryptophan، phenylalanine، proline، valine، isoleucine، leucine، alanine و methionine در این دسته قرار دارند.

- اسیدهای آمینه بدون بار قطبی: این اسیدهای آمینه به دلیل قطبی بودن تمايل دارند در سطح

خارجی پروتئین قرار بگیرند و با مولکول‌های آب پیوند برقرار کنند. از این رو به آن‌ها آبدوست گفته می‌شود. این دسته شامل اسیدهای آمینه serine، threonine، cysteine، و tyrosine است.

- اسیدهای آمینه باردار قطبی: این اسیدها دارای گروه‌های قطبی و باردار و بسیار آبدوست

هستند. اسیدهای باردار قطبی در بخش‌های خارجی پروتئین واقع می‌شوند و معمولاً در مکان‌های فعال شیمیائی³ قرار می‌گیرند. ۵ اسید آمینه در این دسته قرار دارند: lysine، glutamine، asparagine، proline و tyrosine.

- اسیدهای آمینه آروماتیک⁴: گروه‌های R اسیدهای این دسته، دارای حلقه‌های کربنی

آروماتیک هستند. اسیدهای آروماتیک، یعنی phenylalanine، tyrosine و tryptophan، از بزرگترین و سنگین‌ترین اسیدهای آمینه هستند. Phenylalanine و tryptophan غیرقطبی و آبگریزند. Tyrosine اسید آمینه‌ای است که قابلیت یونی شدن دارد، اما میزان قطبیت آن ضعیف است. اسیدهای با این میزان قطبیت، بی‌تفاوت⁵ نامیده می‌شوند. اسیدهای آمینه آروماتیک تمايل به قرار گرفتن در داخل مولکول پروتئین دارند. وجود این اسیدها باعث تثیت ساختار یک پروتئین می‌شود.

2-8-3 پیوند‌های پپتیدی

یک پروتئین از به هم پیوستن تعدادی اسید آمینه به کمک پیوند‌های پپتیدی ایجاد می‌شود. پیوند پپتیدی از طریق ایجاد پیوند بین گروه کربوکسیل یک اسید آمینه و گروه آمینوی اسید آمینه دیگر به وجود می‌آید.

¹ Hydrophobe

² Internal

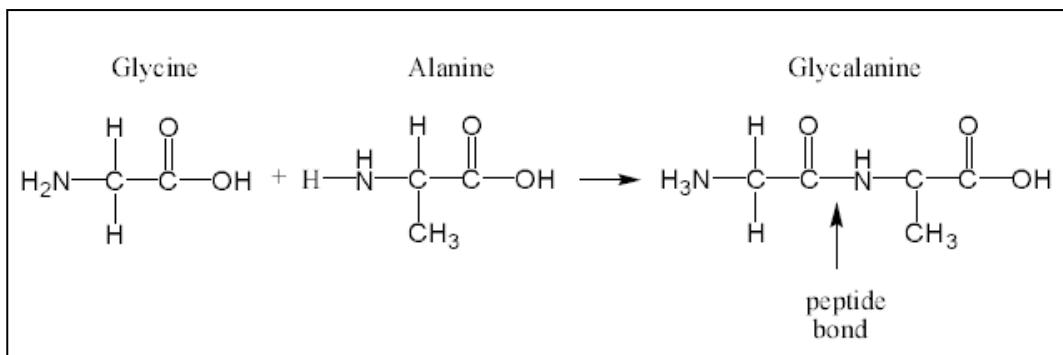
³ Chemically active sites

⁴ Aromatic

⁵ Indifferent

در واقع پروتئین، یک پلیپپتید است. شکل (2-12)، نحوه ایجاد پیوند پپتیدی میان دو اسید آمینه glycine و alanine را نشان می‌دهد.

لازم به ذکر است که مانند یک پلی‌نوكلئوتید، دو انتهای پلی‌پپتید از لحاظ شیمیائی متفاوت و قابل تشخیص‌اند: یک انتها، یک گروه آمینو آزاد به نام پایانه آمینو (NH_2^-) یا پایانه N و انتهای دیگر، یک گروه کربوکسیل آزاد به نام پایانه کربوکسیل (-COOH) یا پایانه C دارد.



شکل 2-12) نحوه ایجاد یک پیوند پپتیدی بین دو اسید آمینه glycine و alanine و تشکیل [3] glycalanine

2-8-4 سطوح مختلف ساختار پروتئین

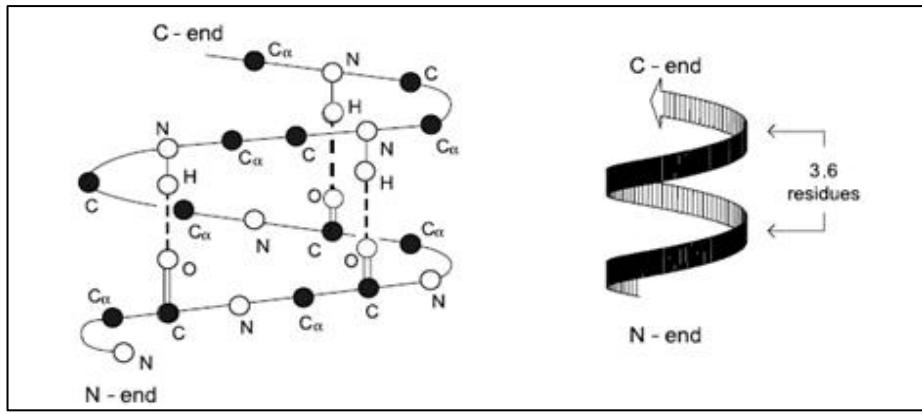
در مولکول‌های پروتئینی چهار سطح از لحاظ ساختاری شناخته شده‌اند:[3]

- ساختار اول: که به توالی اسیدهای آمینه اطلاق می‌شود.
- ساختار دوم: ساختارهای دوم، ساختارهایی هستند که در بسیاری پروتئین‌ها شکل می‌گیرند. علت شکل‌گیری این ساختارها، به وجود آمدن پیوند بین اتم‌های نیدروژن و اکسیژن است. معمولاً نواحی مختلف یک پلی‌پپتید، ساختارهای دوم مختلفی به خود می‌گیرند. ساختارهای دوم رایج عبارتند از:

- مارپیچ‌های آلفا¹: یک ساختار متناوب است که در آن ستون فقرات پروتئین مانند یک پیچ، شکل مارپیچی به خود گرفته و گروه‌های R خارج مارپیچ قرار می‌گیرند. تعداد اسیدهای آمینه در هر دور مارپیچ، تقریباً 3.6 است و هر اسید آمینه متناظر با یک چرخش 100 درجه‌ایست. پایداری فضایی مارپیچ توسط برقراری پیوندهای نیدروژنی بین اتم اکسیژن گروه کربوکسیل اسید آمینه شماره n و اتم نیدروژن گروه آمینوی اسید آمینه $n+4$ حفظ می‌شود. اسید آمینه $n+4$ در بالای اسید آمینه n، در راستای محور

¹ α -helix

مارپیچ، قرار می‌گیرد. این نوع مارپیچ، مارپیچ 4-آلfa نامیده می‌شود. ساختار یک مارپیچ 4-آلfa در شکل (13-2) نمایش داده شده است. ساختارهای مارپیچ دیگری، مشابه 4-آلfa، به نامهای β_{10} ³ و π وجود دارند که تعداد اسیدهای آمینه در هر دور مارپیچ به ترتیب 3 و 5 است. مارپیچ‌ها معمولاً دست-راست¹ هستند و با α_R -helix نشان داده می‌شوند. شکل (13-2) یک مارپیچ دست-راست را نشان می‌دهد. مارپیچ‌های چپ-دست² (α_L -helix) بسیار نادرند.



شکل 2-13) چپ: ساختار یک مارپیچ دست-راست. راست: نمایش نمادین مارپیچ دست-راست

[3]

- صفحه بتا³: یک رشته بتا، زنجیرهای از 5 تا 10 اسید آمینه است که زاویه در ستون فقرات آن‌ها تقریباً 120 درجه است. برخلاف $N-C_\alpha-C-N$ مارپیچ‌های آلفا که توسط یک زنجیره اسید آمینه شکل می‌گیرند، صفحه‌های بتا به وسیلهٔ وجود آمدن پیوند بین جفت‌های بیش از یک رشته بتا ایجاد می‌شوند. این ساختار توسط برقراری پیوندهای نئدروژنی میان اتم اکسیژن گروه کربوکسیل یک اسید آمینه در زنجیره اول و اتم نئدروژن اسید آمینه‌ای دیگر در زنجیره دوم، شکل می‌گیرد. با توجه به جهت قرار گرفتن زنجیره‌ها، صفحه‌های بتا می‌توانند موازی⁴ یا

¹ Right-handed

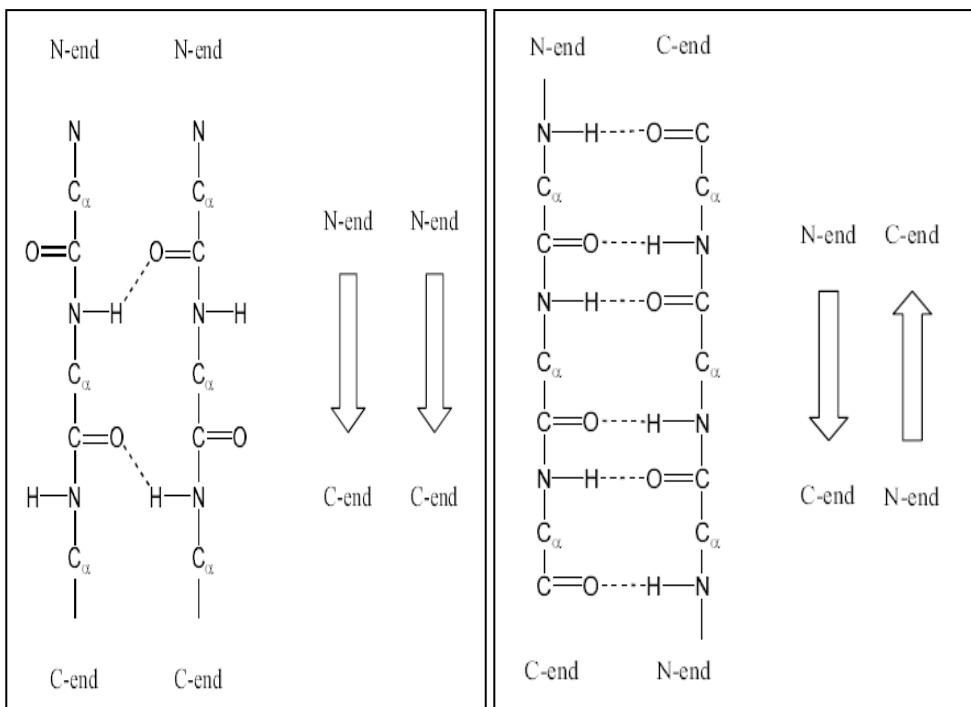
² Left-handed

³ β -helix

⁴ Parallel

ناموازی باشند. شکل (14-2) یک صفحه بتای موازی و یک صفحه بتای ناموازی را نشان می‌دهد. صفحه‌های بتا قابلیت گسترش از عرض، به وسیله پیوستن زنجیرهای دیگر، را دارند.

- ساختارهای دیگر: ساختارهای دوم دیگری وجود دارند، مانند دور¹، حلقه‌های سنجاق‌سری²، پلهای دی‌سولفید³ و زبانه‌های روی⁴. این ساختارها بر نحوه شکل‌گیری و تثبیت شکل پروتئین، بسیار اثرگذارند. به طور کلی به این ساختارها کویل⁵ گفته می‌شود.



شکل 14-2) راست: ساختار یک صفحه بتای موازی. چپ: ساختار یک صفحه بتای ناموازی [3]

- ساختار سوم: ساختار سوم، شکل کلی یک مولکول پروتئین و در واقع، موقعیت فضائی ساختارهای دوم نسبت به یکدیگر است که در اثر تاشدن زنجیره اسیدهای آمینه شکل می‌گیرد. این ساختار با مختصات فضائی مراکز تمام اتم‌های پروتئین مشخص می‌شود. شکل ساختار سوم وابسته به شکل ستون فقرات پروتئین است که خود بستگی به هندسه

¹ Turn

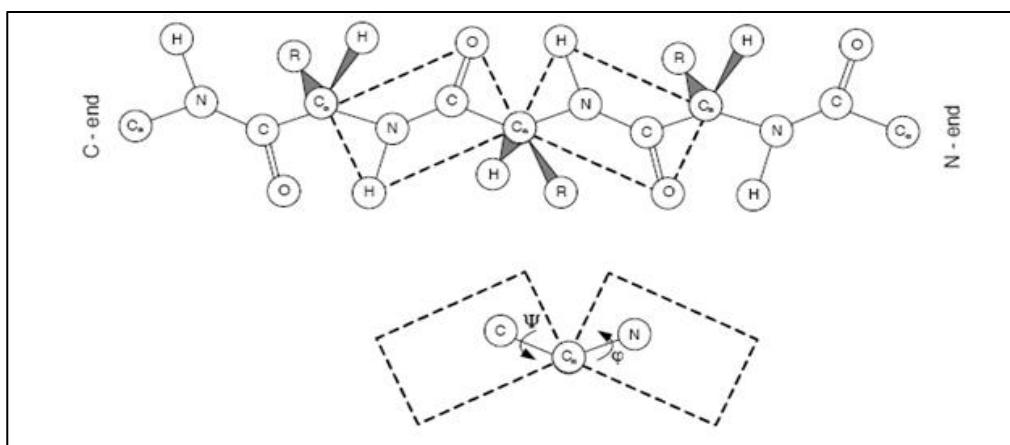
² Hair-pin loop

³ Disulfide bridge

⁴ Zinc finger

⁵ Coil

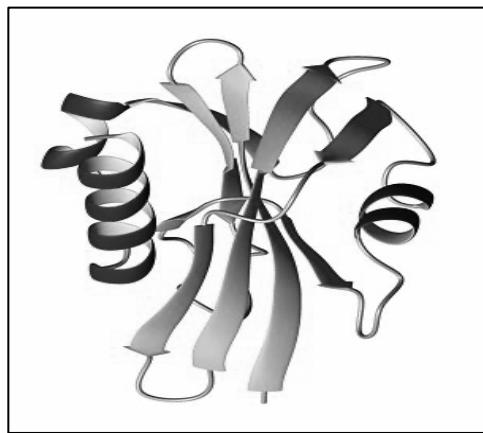
پیوندهای پپتیدی دارد. شکل (2-15) تعریف و نمایش قراردادی هندسه پیوندهای پپتیدی را نشان می‌دهد. بخش بالای تصویر، نمایش سه بعدی بخشی از زنجیره پلی‌پپتیدی است. مشخصه این تصویر این است که اتم‌های شکل دهنده پیوندهای پپتیدی، یعنی C و N و چهار اتم متصل به آن‌ها، O، H، C_α و $C_{\alpha\alpha}$ ، تقریباً در یک صفحه قرار دارند. بخش پائینی تصویر، بازه تاشدن پلی‌پپتید را نشان می‌دهد. برای هر اسید آمینه، پیکربندی¹ زوایای ثابت است و تا شدن، توسط چرخش دو صفحه حول محورهای $C - C_\alpha$ و $C_\alpha - N$ ، صورت می‌گیرد. زوایای چرخش ϕ و ψ که در بازه 180° - درجه تا +180° درجه تغییر می‌کنند، زوایای دوسطحی² نامیده می‌شوند. تاشدن پروتئین از عواملی اثر می‌گیرد. از جمله این عوامل، نیروهای بین اسیدهای آمینه پروتئین است. هر دو اسید آمینه با توجه به خواص فیزیکی-شیمیایی، تراکنشهایی با یکدیگر دارند که در شکل تاشدن پروتئین اثرگذار است. با توجه به خواص اتم‌های اسید آمینه، از آنجایی که پروتئین در محیط طبیعی، در آب قرار دارد، یکی دیگر از عوامل موثر در تاشگی، آبدوستی و یا آبگریزی اسیدهای آمینه است. در نهایت پروتئین به شکلی تا خواهد شد که اسیدهای آمینه آبدوست در سطح خارجی پروتئین قرار گیرند و اسیدهای آمینه آبگریز به دور هم جمع شده و هسته آبگریز را تشکیل دهند. شکل (2-16) تصویری از ساختار سوم یک پروتئین را نشان می‌دهد.



شکل 2-15) بالا: بخشی از زنجیره پلی‌پپتیدی. دو اتم C و N که تشکیل پیوند پپتیدی می‌دهند و چهار اتم مجاور آن‌ها در یک صفحه قرار دارند. پائین: بازه تاشدن پلی‌پپتید [3]

¹ Configuration

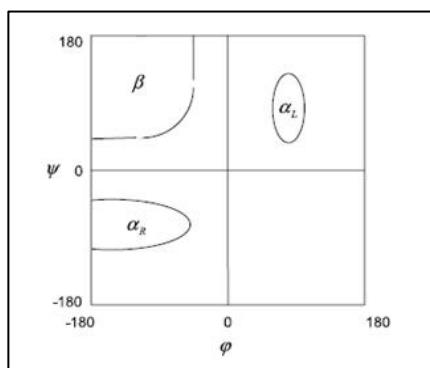
² Dihedral



شکل 2-16) ساختار سوم یک پروتئین [5]

نمودار راماچاندران¹

مجموعه زاوایای دوسری $\{\phi_1, \varphi_1, \dots, \phi_N, \varphi_N\}$ که N طول توالی اسید آمینه است، پیکربندی یک زنجیره پلی پپتیدی را، بدون نمایش مختصات، مشخص می‌کند. به نمودار نقطه‌ای² که مقادیر زوایای دوسری را برای یک زنجیره پلی پپتید، در صفحه $\psi - \varphi$ نمایش می‌دهد، نمودار راماچاندران گفته می‌شود (شکل 2-17). همان‌طور که در شکل دیده می‌شود، مقادیر زوایای ψ و φ به طور یکنواخت توزیع نشده‌اند. برخی پیکربندی‌های ψ و φ ، به علت محدودیت‌های فضائی R، اتفاق نمی‌افتد (پیکربندی‌های ممنوع). بعضی از نواحی نمودار، متناظر با ساختارهای دوم مارپیچ آلفا و صفحه بنا هستند.



شکل 2-17) نمودار راماچاندران برای یک زنجیره پلی پپتیدی [3]

مکان‌های فضائی R ها

¹ Ramachandran plot

² Dot plot

مورد دیگری که ساختار سوم یک زنجیره پلی پپتیدی را شرح می‌دهد، مکان گروهای R نسبت به ستون فقرات پروتئین است. در شکل (15-2) دیده می‌شود که R ها می‌توانند به دو صورت نسبت به ستون فقرات بچرخند، به این ترتیب که R اول به سمت بیرون شکل و R بعدی به سمت داخل شکل قرار گرفته و زنجیره به همین ترتیب ادامه یابد. به این پیکربندی trans گفته می‌شود. پیکربندی دیگری به نام cis وجود دارد که در آن تمام R ها به یک سمت اشاره می‌کنند. trans و cis دارای خصوصیات متفاوت هستند.

- ساختار چهارم: ساختار چهارم از تجمع دو یا چند پلی‌پپتید به وجود می‌آید. ساختار چهارم می‌تواند شامل چند مولکول از یک نوع پلی‌پپتید یا از پلی‌پپتیدهای مختلف باشد، مانند RNA-پلیمراز. این ساختار، توسط شکل زنجیره‌های تشکیل‌دهنده و واکنش‌های شیمیائی آن‌ها مشخص می‌گردد. برخی از ساختارهای چهارم از تعداد بسیار زیادی پلی‌پپتید تشکیل می‌شوند.

5-8-2 اهمیت توالی اسیدهای آمینه در ساختار و واکنش‌های پروتئین

ساختارهای دوم، سوم و چهارم پروتئین به وسیله ساختار اول یعنی توالی اسیدهای آمینه تعیین می‌شوند. این موضوع در ساختار دوم بهتر قابل درک است، زیرا مشخص شده که بعضی اسیدهای آمینه به خاطر خصوصیات فیزیکی و شیمیائی گروههای R خود، باعث تشکیل مارپیچ آلفا، بعضی دیگر موجب تشکیل صفحه بتا و برخی موجب تشکیل ساختارهای دوم دیگر می‌شوند. این عوامل امروزه به خوبی شناخته شده‌اند، به طوری که بر اساس آن‌ها روش‌هایی برای پیش‌بینی ساختار دوم حاصل از توالی اسیدهای آمینه به وجود آمده‌اند. ساختارهای سوم و چهارم یک پروتئین نیز به توالی اسیدهای آمینه آن بستگی دارند. اثر متقابل بین اسیدهای آمینه و این ساختارها به قدری پیچیده است که قوانین پیش‌بینی کننده موجود هنوز قابل اعتماد نیست. با این وجود معلوم شده است چنانچه پروتئینی مثلا در اثر حرارت ساختمان خود را از دست بدده و شکل غیر منظمی به خود بگیرد، همچنان قادر است (مثلا در اثر سرد شدن) دوباره ساختار صحیح سوم خود را بازیابد. پس از تشکیل ساختار سوم، تجمع پلی‌پپتیدها برای تشکیل ساختار چهارم خود به خود انجام می‌شود. این مساله نشان می‌دهد که اطلاعات لازم برای تشکیل ساختارهای سوم و چهارم باید در توالی اسیدهای آمینه نهفته باشد.

از بررسی‌های قبل می‌توان نتیجه‌گیری کرد که واکنش‌های یک پروتئین نیز به وسیله اسیدهای آمینه آن تعیین می‌شوند. برای روشن شدن موضوع، پروتئین‌هایی را در نظر می‌گیریم که باید خود را

به یک مولکول DNA متصل کنند تا بتوانند فعالیت خود را در سلول انجام دهند. این پروتئین‌های متصل شونده به DNA، دسته بزرگ و گوناگونی از قبیل RNA-پلیمراز و تعدادی از پروتئین‌های تنظیم‌کننده هستند که رونویسی ژن‌ها را تنظیم و گاهی از آن جلوگیری می‌کنند. مثالی از این پروتئین‌ها، پروتئین تنظیمی کرو¹ است. این پروتئین شامل دو پلیپپتید یکسان است که هر یک شامل سه مارپیچ آلفا و یک صفحه بتا می‌باشد. در حالت فعال پروتئین، دو مارپیچ آلفا دقیقاً 34 Å (34 آنگstrom) از هم فاصله دارند و لذا برای استقرار در داخل دو بخش فرورفته اصلی یا بزرگ مولکول DNA مناسب هستند. اگر مارپیچ‌های آلفا وجود نداشتند یا جهت دیگری داشتند، قابلیت اتصال به DNA در این پروتئین از بین می‌رفت. بنابراین خاصیت اتصال به پروتئین بستگی به ساختارهای دوم، سوم و چهارم پروتئین دارد که همه این‌ها به نوبه خود به توالی اسیدآمینه‌ای وابسته است. همان طور که گفته شد، توالی اسید آمینه‌ای از توالی نوکلوتیدی در mRNA، یعنی نسخه‌ای از ژن، تبعیت می‌کند [6].

2-8-2 روش‌های مطالعه ساختمان پروتئین

آزمایش‌های گوناگونی برای مطالعه ساختمان پروتئین‌ها مورد استفاده قرار می‌گیرند، از جمله [6]:

- 1- نوع اسیدهای آمینه، از طریق شکستن پروتئین‌ها به اجزای اسیدآمینه آن‌ها با قرار دادن نمونه در اسید غلیظ و دمای بالا به مدت چند ساعت، به دست می‌آید. سپس اسیدهای آمینه به کمک کروماتوگرافی جدا و مورد ارزیابی قرار می‌گیرند.
- 2- توالی اسیدهای آمینه را می‌توان به چند روش تعیین کرد. این روش‌ها مشتمل بر فرآیند تجزیه گام به گام و جدا کردن یک به یک اسیدهای آمینه از انتهای پلیپپتید است. هر اسید آمینه به وسیله خواص کروماتوگرافی خود تشخیص داده می‌شود و چرخه تا آخرین اسیدآمینه تکرار می‌شود. توالی‌یاب‌های پلیپپتیدی خودکار در سال 1967 به وجود آمدند.
- 3- ساختار سوم به کمک روش‌های مانند تجزیه پراش اشعه ایکس² و طیف‌شناصی NMR³ تعیین می‌شود.

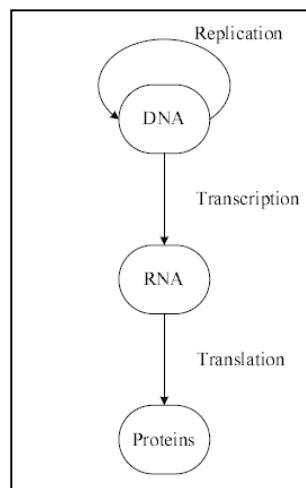
¹ Cro

² X-ray crystallography

³ NMR spectroscopy

2- قضیه مرکزی زیست مولکولی

با کنار هم قرار دادن مطالب گفته شده، جریان اطلاعات ژنتیکی مشخص می‌شود. به این صورت که ساخت RNA و RNA به نوبه خود ساخت پروتئین را هدایت می‌کند. جریان اطلاعات ژنتیکی از اسید نوکلئیک به پروتئین، قضیه مرکزی زیست مولکولی نامیده می‌شود (شکل (18-2)). [5]



شکل 2-18) قضیه مرکزی زیست مولکولی [3]

فصل سوم

مفاهیم محاسباتی پیش‌زمینه

3-1 مقدمه

در این فصل مفاهیم محاسباتی پیش‌زمینه لازم، شرح داده شده‌اند. مباحثت، به دو بخش مفاهیم بیوانفورماتیکی و مفاهیم یادگیری ماشین، تئوری اطلاعات و آمار و احتمال تقسیم می‌شوند. در بخش اول به توضیح پایگاه داده PDB، فرمت فایل PDB، لیست pdbselect، فرمت فایل FASTA، تراز توالی (دوگانه و چندگانه)، جهش وابسته، پیش‌بینی ساختار دوم و پایگاه داده AAindex پرداخته شده است. در بخش دوم روش‌های رده‌بندی، شبکه‌های عصبی، تخمین دقت، اثر فراموشی خطرناک، یادگیری گروهی و چند مفهوم تئوری اطلاعات و آمار و احتمال معرفی شده‌اند. لازم به ذکر است که تمامی آدرس‌های وب ذکر شده، در تاریخ "15 April 2009" معتبر هستند.

3-2 مفاهیم بیوانفورماتیکی

در این بخش مفاهیم بیوانفورماتیکی لازم که در این تحقیق از آن‌ها استفاده شده است، شرح داده شده‌اند.

PDB 3-2-1

بانک داده پروتئین (PDB)، یک پایگاه داده برای داده‌های ساختاری سه‌بعدی مولکول‌های بیولوژیکی بزرگ مانند پروتئین‌ها و اسیدهای نوکلئیک است. داده‌ها معمولاً توسط روش‌های تجزیه پراش اشعه ایکس

و طیفبینی NMR به دست آمده و توسط زیستشناسان و بیوشیمیست‌ها از سراسر جهان به این پایگاه داده ارسال می‌شوند. PDB، فایل‌های ارسالی را برای خطاهای ناسازگاری‌های احتمالی بررسی می‌نماید و سپس آن‌ها را با یک فرمت متغیر استاندارد به طور رایگان بر روی وب قرار می‌دهد. PDB توسط سازمانی به نام بانک داده پروتئین جهانی (wwPDB) مدیریت می‌شود.

PDB یک منبع اصلی در زمینه زیست‌شناسی ساختاری مانند ژنومیک ساختاری است. اکثر ژورنال‌های علمی مهم از دانشمندان می‌خواهند ساختارهای خود را به PDB ارسال نمایند. اگر محتویات PDB به عنوان داده اولیه تصور شود، هزاران پایگاه داده مشتق شده از آن وجود دارند که داده‌ها را به روش‌های مختلف طبقه‌بندی می‌کنند، مانند SCOP و CATH که ساختارها را بر اساس نوع و رابطه تکاملی آن‌ها طبقه‌بندی نموده‌اند.

PDB به طور هفتگی به‌亨گام می‌شود. در حال حاضر (10 March 2009)، 56366 ساختار در PDB وجود دارد که 52079 ساختار، بر وتنین هستند. پایگاه داده PDB در آدرس زیر قابل دسترسی است:

PDB: <http://www.pdb.org>

فایل های PDB

فایل‌های ساختاری PDB، به فرمت PDB هستند. در شکل(3-1) بخشی از یک فایل PDB مشاهده می‌شود.

HEADER EXTRACELLULAR MATRIX 22-JAN-98 1A3I
 TITLE X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
 TITLE 2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
 ...
 EXPDTA X-RAY DIFFRACTION
 AUTHOR R.Z.KRAMER, L.VITAGLIANO, J.BELLA, R.BERISIO, L.MAZZARELLA,
 AUTHOR 2 B.BRODSKY, A.ZAGARI, H.M.BERMAN
 ...
 REMARK 350 BIOMOLECULE: 1
 REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
 REMARK 350 BIOMT1 1 1.000000 0.000000 0.000000 0.000000
 REMARK 350 BIOMT2 1 0.000000 1.000000 0.000000 0.000000
 ...
 SEQRES 1 A 9 PRO PRO GLY PRO PRO GLY PRO PRO GLY
 SEQRES 1 B 6 PRO PRO GLY PRO PRO GLY
 SEQRES 1 C 6 PRO PRO GLY PRO PRO GLY
 ...
 ATOM 1 N PRO A 1 8.316 21.206 21.530 1.00 17.44
 N
 ATOM 2 CA PRO A 1 7.608 20.729 20.336 1.00 17.44
 C
 ATOM 3 C PRO A 1 8.487 20.707 19.092 1.00 17.44
 C
 ATOM 4 O PRO A 1 9.466 21.457 19.005 1.00 17.44
 O
 ATOM 5 CB PRO A 1 6.460 21.723 20.211 1.00 22.26
 C
 ...
 HETATM 130 C ACY 401 3.682 22.541 11.236 1.00 21.19
 C
 HETATM 131 O ACY 401 2.807 23.097 10.553 1.00 21.19
 O
 HETATM 132 OXT ACY 401 4.306 23.101 12.291 1.00 21.19
 O
 ...

شکل 3-1) بخشی از یک فایل PDB [7]

علاوه بر لیست مختصات، یک فایل PDB شامل یک هدر یا بخش سرآغاز، شامل اطلاعاتی در مورد مقالات منتشر شده در مورد ساختار، جزئیات کار تجربی و اطلاعات مفید دیگر است. گفته شد که یک پروتئین ممکن است شامل چند زنجیره پلیپپتید باشد (ساختار چهارم)، در نتیجه در یک فایل ممکن است بیش از یک زنجیره وجود داشته باشد. مختصات داده شده در فایل‌های PDB، در سیستم مختصات کاتزین سهبعدی، در واحد آنگستروم هستند. برای مبدا مختصات استانداردی تنظیم نشده است. بعضی مرکز را بر روی میانگین هندسی اتم‌ها، بعضی بر روی میانگین وزن‌دار و بعضی نقاط دیگری را انتخاب می‌کنند. سیستم مختصات به راحتی قابل انتقال به هر مکان جدید است. با وجود بررسی‌های صورت گرفته بر روی فایل‌های PDB، اشتباهاتی در آن‌ها وجود دارند. در ادامه توضیح مختصری در رابطه با محتوای فایل‌های PDB داده می‌شود. نوع خطها، با حروف بزرگ در سمت چپ هر خط فایل نوشته می‌شود. بعضی از انواع خط، به تمام مدل‌ها اعمال نمی‌شوند و بنابراین در تمام فایل‌ها وجود ندارند. محتویات یک فایل PDB به ترتیب عبارت است (تمامی بخش‌ها ذکر نشده‌اند) [7]:

- TITLE و HEADER: اسم فایل، تاریخ و یک عنوان کوتاه.
- COMPND: اسم پروتئین.
- SOURCE: موجود زنده‌ای که پروتئین از آن به دست آمده است.
- KEYWDS: کلمات کلیدی که به جستجو در فایل کمک می‌کنند.
- EXPDTA: روش تجربی به دست آوردن ساختار.
- AUTHOR: لیست افرادی که این داده‌ها را در PDB قرار داده‌اند.
- REVDAT: لیست تمام تاریخ‌های بازنگری داده‌های این پروتئین.
- JRNL: رفرنس ژورنال به مقاله اصلی در مورد این مدل.
- REMARK: رفرنس به مقالات ژورنال در مورد ساختار این پروتئین و اطلاعات کلی دیگر در مورد محتویات این فایل.
- SEQRES: توالی اسید آمینه پروتئین که اسیدهای آمینه با کد سه‌حرفی نشان داده شده‌اند.
- FORMUL و HET: لیست مواد غیرپروتئینی موجود در ساختار (heteromer).
- TURN، SHEET، HELIX: لیست عناصر ساختار دوم پروتئین.

• MODEL: شماره سریال مدل (ساختارهای به دست آمده توسط NMR ، معمولاً چند مدل دارند).

• ATOM: مختصات اتمی تمام اتمهای پروتئین. ترتیب اتمها به این صورت است که ابتدا اتمهای ستون فقرات لیست می‌شوند، شامل نیتروژن (N)، کربن آلفا (CA)، کربن کربوکسیل (C) و اکسیژن کربوکسیل (O). سپس اتمهای زنجیره جانبی در ادامه قرار می‌گیرند، شامل کربن بتا (CB)، کربن گاما (CG) و به همین ترتیب. در زنجیره‌های جانبی شاخه‌ای یا حلقه‌ها، اتمها در دو شاخه با اعداد 1 و 2 شماره‌گذاری می‌شوند. برای مثال اتمهای اسید aspartic به ترتیب N، CA، C، O، CG، CB، OE1 و OE2 هستند. در فایل علامتی وجود ندارد که شروع هر اسید را مشخص کند و هر N نشان‌گر آغاز اسید آمینه بعدی است. هر سطر ATOM شامل فیلد‌های (ATOM، شماره سریال اتم، نام اتم، مشخصه مکان فرعی اتم، نام اسید آمینه، مشخصه زنجیره، شماره توالی اسید آمینه، کد برای درج اسیدهای آمینه¹، مختصه X، مختصه Y، مختصه Z،occupancy، فاکتور دما، مشخصه بخش²، نماد عنصر، بار روی اتم) است. وجود برخی فیلد‌ها اجباری و برخی اختیاری است.

• ANISOU: فاکتورهای دمای ناهمسان‌گرد.³

• TER: مشخص‌کننده پایان زنجیره‌های مجزا در مدل.

• HETATM: شامل اطلاعات مشابه خطوط ATOM برای هر مولکول غیرپروتئین موجود در پروتئین که در خطوط HET و FORMULA ذکر شده‌اند.

• CONECT: لیست پیوندهای میان اتمهای غیرپروتئینی موجود در فایل.

• END MASTER و مشخص‌کننده انتهاي فایل.

pdbselect 3-2-2

لیست‌های pdbselect، هر یک شامل زیرمجموعه‌ای از زنجیره‌های پروتئین موجود در پایگاه داده PDB هستند. این لیست‌ها با ارائه یک زیرمجموعه بیانگر⁴ از زنجیره‌های پروتئین، در زمان بسیاری از تحقیقات پروتئینی صرفه‌جوئی می‌نمایند. در حال حاضر pdbselect تقریباً 15 برابر از PDB کوچکتر است. هر لیست با نگه داشتن تنها یک زنجیره از میان زنجیره‌های مشابه، با یک آستانه شباهت خاص، به

¹ Code for insertion of residues

² Segment identifier

³ Anisotropic temperature factors

⁴ Representative

وجود می‌آید. رایج‌ترین لیست، لیست 25% است که در آن هیچ دو زنجیره پروتئین بیش از 25% با یکدیگر شbahت ندارند. فرمت لیست، نحوه محاسبه شbahت میان توالی‌ها و الگوریتم انتخاب توالی‌ها، در مستندات مربوطه بیان شده است. لیست pdbselect در آدرس وب زیر قابل دستیابی است:

pdbselect: <http://bioinfo.tg.fh-giessen.de/pdbselect>

3-2-3 فرمت FASTA

در بیوانفورماتیک فرمت FASTA یک فرمت متنی برای بیان توالی‌های نوکلئوتیدی یا پپتیدی است که در آن جفت‌های باز یا اسید‌های آمینه توسط کدهای یک حرفی نشان داده می‌شوند. یک توالی در FASTA با یک خط توضیح آغاز می‌شود و خطوط داده‌های توالی در ادامه می‌آیند. خط توضیح با یک نماد ">" در اولین ستون، از خطوط داده متمایز می‌شود. کلمه بعد از ">", مشخصه¹ یکتای توالی است که تا آخرین نماد "]" ادامه دارد و بقیه خط، توضیحات است. مشخصه و توضیحات اختیاری هستند. میان ">" و اولین حرف مشخصه، نباید فاصله وجود داشته باشد. توصیه می‌شود که همه خطوط طول کمتر از 80 کارکتر داشته باشند. در یک فایل بیش از یک توالی می‌تواند وجود داشته باشد که آغاز هر یک با نماد ">" مشخص می‌گردد. سادگی فرمت FASTA، پردازش آن را توسط ابزارهای پردازش متن به سادگی میسر می‌سازد. نمونه‌ای از یک فایل FASTA برای یک توالی اسید آمینه در شکل (3-2) مشاهده می‌شود.

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVHCTNLMNTTTGLLNGSYSNRT
QIWQKHRTSDNSLILLNKHYNLTVTCKRPGNKTVLPTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNMLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

شکل 3-2) نمونه‌ای از یک فایل [8] FASTA

3-2-4 تراز توالی

در بیوانفورماتیک تراز توالی روши برای منظم کردن توالی‌های DNA، RNA یا پروتئین به منظور تشخیص نواحی مشابه آن‌ها است. این نواحی مشابه می‌توانند نتیجه ارتباط عملیاتی، ساختاری یا تکاملی

¹ Identifier

آن‌ها باشند. توالی‌های تراز شده معمولاً به صورت سطرهای یک ماتریس نشان داده می‌شوند. بین عناصر توالی‌ها در مکان‌های لازم فاصله¹ درج می‌شود تا عناصر یکسان یا شبیه در ستون‌های بعدی تراز شوند. اگر دو توالی دارای جد مشترک باشند، می‌توان عدم تطابق‌ها در تراز را به عنوان نقاط جهش و فاصله‌ها را به عنوان ایندل‌ها² (جهش‌های درج یا حذف) که در یک یا هر دو توالی اتفاق افتاده، تفسیر نمود. در توالی‌های پروتئین، درجه شباهت میان اسیدهای آمینه یک مکان توالی می‌تواند به عنوان یک معیار غیردقیق از میزان حفاظت³ یک ناحیه مورد استفاده قرار گیرد. میزان ((حفظت)) یک ناحیه، به مفهوم درجه تغییر عناصر آن ناحیه در طول زمان است. ناحیه‌های حفاظت شده⁴، معمولاً اهمیت عملیاتی یا ساختاری دارند.

تراز کردن توالی‌های طولانی نیازمند روش‌های محاسباتی است. این روش‌ها به دو نوع تراز‌های سراسری و محلی تقسیم می‌شوند. ترازهای سراسری کل طول توالی‌ها را تراز می‌کنند. در مقابل ترازهای محلی، ابتدا نواحی مشابه را درون توالی‌های طولانی تعیین و سپس آن نواحی را تراز می‌نمایند. ترازهای محلی معمولاً به ترازهای سراسری ترجیح داده می‌شوند، اما به دلیل تعیین نواحی مشابه، محاسبه آن‌ها مشکل‌تر است. ترازهای سراسری هنگامی مفید هستند که توالی‌های مورد سوال، شبیه و طول آن‌ها تقریباً یکسان باشند. یک الگوریتم تراز سراسری Needleman-Wunsch است که بر مبنای برنامه‌نویسی پویا عمل می‌نماید. ترازهای محلی برای تراز کردن توالی‌های غیرمشابه که ممکن است نواحی مشابه داشته باشند، مفید هستند. روش Smith-Waterman یک الگوریتم تراز محلی بر اساس برنامه‌نویسی پویاست [3].

تراز جفتی⁵

روش‌های تراز جفتی برای یافتن بهترین تطابق سراسری یا محلی بین دو توالی بهکار می‌روند. از جمله روش‌های اصلی تراز جفتی، روش ماتریس نقطه⁶ و برنامه‌نویسی پویا هستند.

تراز توالی چندگانه⁷

تراز توالی چندگانه (MSA)، توسعه تراز جفتی است. روش‌های MSA چندین توالی را با یکدیگر را تراز می‌کنند. MSA از نظر محاسباتی دشوارتر از تراز جفتی است و معمولاً منجر به مسائل بهینه‌سازی -NP-

¹ Gap

² Indel

³ Conservation

⁴ Conserved

⁵ Pairwise alignment

⁶ Dot-matrix

⁷ Multiple sequence alignment

Complete می‌شود. روش‌های رایج به کار رفته در MSA عبارتند از برنامه‌نویسی پویا، روش‌های جلورونده¹، روش‌های تکراری²، الگوریتم ژنتیک و مدل پنهان مارکوف.

HSSP یک پایگاه داده است که برای هر پروتئین موجود در PDB، دارای یک فایل حاوی MSA برای آن پروتئین، میزان حفاظت آن و اطلاعاتی دیگر است. این پایگاه داده همواره به‌هنگام می‌شود. از نرم‌افزارهای رایج MSA، ClustalW، Blast و SAM را می‌توان نام برد. موارد ذکر شده در آدرس‌های زیر قابل دستیابی هستند:

HSSP: <ftp://ftp.embl-heidelberg.de/pub/databases/hssp/>

ClustalW: <http://www.clustal.org/>

Blast: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

SAM: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

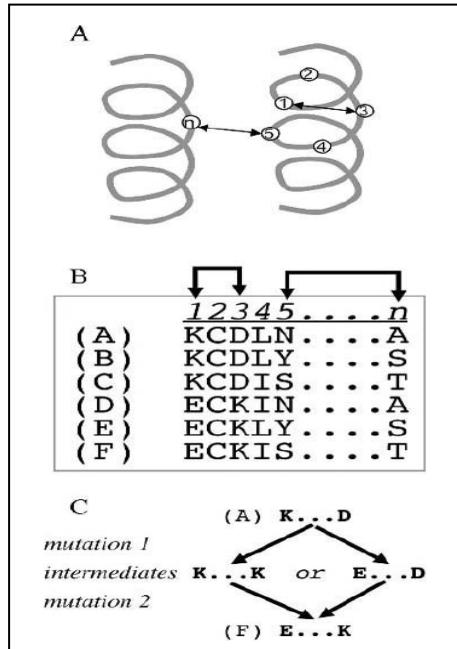
3-2-5 جهش وابسته

هنگامی که یک اسید آمینه در یک پروتئین جهش پیدا می‌کند، این امکان وجود دارد که در مکانی دیگر یک جهش جبران‌کننده³ اتفاق بیافتد. به چنین جهش‌هایی، جهش وابسته گفته می‌شود. دو مکان جهش کننده ممکن است از نظر عملیات، انرژی یا نزدیکی فیزیکی با یکدیگر مرتبط باشند. شکل (3-3) نمونه‌ای از یک جهش وابسته را برای یک پروتئین فرضی نمایش می‌دهد [9].

¹ Progressive

² Iterative

³ Compensating mutation



شکل 3-3 (A) چند اسید آمینه در ساختار یک پروتئین (B) یک MSA برای پروتئین. مکان‌های 1 و 3 و نیز 5 و n با یکدیگر جهش وابسته دارند. (C) تئوری این است که وقتی یک اسید آمینه که از نظر ساختاری با اهمیت است، جهش می‌کند، ساختار پروتئین ناپایدار می‌شود. در نتیجه جهش دیگری برای بازگرداندن پایداری رخ می‌دهد. حالات میانی به دلیل ناپایدار بودن حذف می‌شوند [9].

الگوریتم‌های متعددی برای اندازه‌گیری میزان وابستگی میان مکان‌های یک پروتئین وجود دارند. یکی از الگوریتم‌های تحلیل جهش وابسته در جدول (2-3) ارائه شده است.

جدول 2-3) الگوریتم محاسبه جهش وابسته با استفاده از ماتریس McLachlan [10]

- فرض شود پروتئین هدف طول lenP و MSA آن N توالی دارد.
- (1) برای هر مکان، یک ماتریس $N \times N$ را محاسبه کن. این ماتریس، شباهت جفت‌های اسید آمینه را برای آن مکان برای همه توالی‌ها نشان می‌دهد. شباهت با استفاده از ماتریس McLachlan محاسبه می‌شود.
- (2) میانگین ماتریس هر مکان را محاسبه کن و در آرایه avgArr به طول lenP ذخیره کن.
- (3) برای هر جفت مکان، ضریب همبستگی (کوواریانس) را محاسبه کن، به این صورت که کوواریانس در ایهای منتظر ماتریس‌های دو مکان را محاسبه و سپس میانگین تمام کوواریانس‌ها را به عنوان ضریب همبستگی دو مکان در نظر بگیر. کوواریانس دو متغیر X و Y به صورت زیر تعریف می‌شود:

$$\text{Covariance}(x,y) = E[(x - \text{mean}X)(y - \text{mean}Y)] / (\sigma_X \cdot \sigma_Y)$$

این الگوریتم از ماتریسی به نام McLachlan استفاده می‌کند. ماتریس McLachlan، از جمله ماتریس‌های جایگزینی¹ (ماتریس‌های شباخت یا ماتریس‌های جهش) است که نرخ تبدیل یک اسید آمینه به اسید آمینه دیگری را در طول زمان نشان می‌دهند. یک ماتریس جایگزینی، شامل 210 مقدار عددی است (20 مقدار قطری و $20^*19/2$ مقدار غیرقطري). از این ماتریس‌ها برای تراز توالي‌ها و جستجوی توالي‌های مشابه استفاده می‌شود [10].

3-2-3 پيش‌بياني ساختار دوم

ساختار‌های دوم در بخش (4-8-2) شرح داده شدند. یک عنصر ساختار دوم متناظر با یک ناحیه محلي در ساختار سوم بوده که ويزگي‌های هندسي متمايزی نشان مي‌دهد. دو نوع ساختار دوم پايه‌اي، ماريپيج آلفا و رشته بتا هستند که منظم بوده و به راحتی در ساختار سوم قابل تشخيص هستند. دیگر انواع ساختار دوم عموماً به سختي تشخيص داده مي‌شوند. به همين دليل، بيشر روش‌های پيش‌بياني ساختار دوم، از یک الفاي سه‌حرفي برای پيش‌بياني استفاده مي‌نمایند: H برای ماريپيج آلفا ، E برای رشته بتا و C برای كوييل

¹ Substitution matrix

که منظور، دیگر ساختارهای دوم است. الفهای دیگری با تعداد حروف بیشتر نیز وجود دارند، مانند DSSP و STR4، STR2 هستند و ۵۰٪ اسیدهای آمینه در یک پروتئین جز مارپیچهای آلفا و رشته‌های بتا ۵۰٪ بقیه نامنظمند. تحلیل ساختارهای دوم و ویژگی‌های مرتبط پروتئین‌ها، منجر به ایجاد یک مجموعه از قوانین در مورد این ساختارها شده که برای پیش‌بینی ساختارهای دوم حائز اهمیت هستند. بیشتر روش‌های پیش‌بینی، از این قوانین به طور مستقیم یا غیرمستقیم استفاده می‌کنند.

سه گام استانداردی که امروزه تقریباً توسط تمامی روش‌های پیش‌بینی ساختار دوم است دنبال می‌شوند

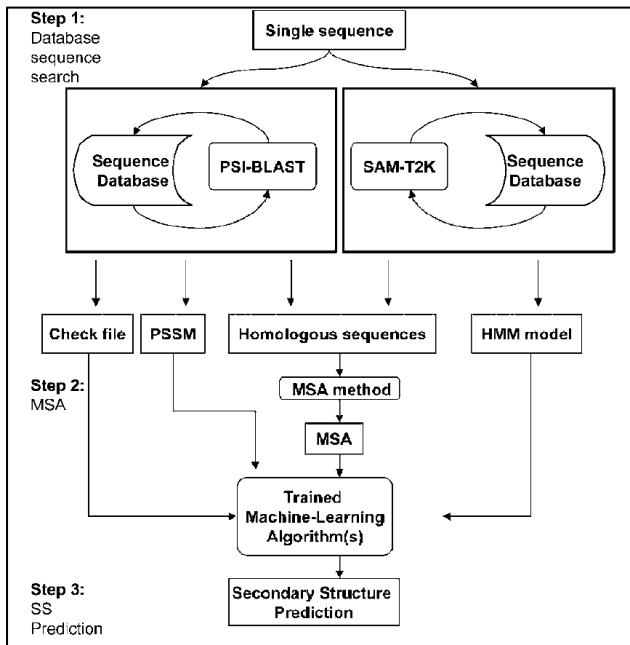
عبارتند از [9]:

- استخراج توالی‌های شبیه به توالی مورد نظر از یک پایگاه داده.
- محاسبه MSA برای توالی‌ها
- پیش‌بینی ساختار دوم توسط یکی از روش‌های یادگیری ماشین با کمک اطلاعات MSA به دست آمده.

شکل (4-3) این مراحل را نشان می‌دهد.

در پیش‌بینی ساختار دوم از روش‌های یادگیری ماشین مختلفی استفاده می‌شود. رایج‌ترین آن‌ها، شبکه‌های عصبی، مدل پنهان مارکوف و ماشین بردار پشتیبان هستند. نرمافزارهای مختلفی بر همین اساس به وجود آمده‌اند، مانند PHD، MNOSTER، PROBCONS، MUSCLE، YASPIN، PSIPRED و SAMT02. به عنوان مثال PSIPRED اطلاعات MSA به دست آمده توسط Blast را به یک شبکه عصبی دو لایه می‌دهد. دقت PSIPRED در حدود ۷۶.۵٪ است و در زمره بهترین روش‌های پیش‌بینی ساختار دوم به شمار می‌رود. PSIPRED در آدرس زیر قابل دستیابی است:

PSIPRED: <http://bioinf.cs.ucl.ac.uk/psipred>



شکل 3-4) مراحل پیش‌بینی ساختار دوم [9]

3-2-7 AAindex پایگاه داده

AAindex یک پایگاه داده از خصوصیات مختلف اسیدهای آمینه است. این پایگاه داده در حال حاضر شامل سه بخش است: AAindex1، AAindex2 و AAindex3. تمام داده‌ها، از مقالات منتشر شده استخراج شده‌اند.

یک اندیس¹ اسید آمینه، مجموعه‌ای از 20 مقدار عددی است که یکی از ویژگی‌های بیوشیمیائی و شیمیائی-فیزیکی اسیدهای آمینه را بیان می‌کند. بخش AAindex1 شامل تعدادی از اندیس‌های منتشر شده به همراه نتیجه تحلیل کلاستر اندیس‌ها، با استفاده از ضریب همبستگی به عنوان فاصله دو اندیس، است. این بخش در حال حاضر 544 اندیس دارد. بخش AAindex2 شامل مجموعه‌ای از ماتریس‌های شباهت منتشر شده به همراه نتیجه تحلیل کلاستر آن‌ها است. این بخش هم‌اکنون 94 ماتریس دارد. ماتریس پتانسیل AAindex²، ماتریسی است که احتمال تماس داشتن هر جفت اسید آمینه را نشان می‌دهد. بخش AAindex3 شامل مجموعه‌ای از این ماتریس‌هاست. این بخش در حال حاضر دارای 47 ماتریس است. پایگاه داده AAindex در آدرس زیر قابل دستیابی است:

AAindex: <http://www.genome.jp/aaindex>

¹ Index

² Contact potential

3-3 مفاهیم یادگیری ماشین، تئوری اطلاعات و آمار و احتمال

در این بخش مفاهیم یادگیری ماشین، تئوری اطلاعات و آمار و احتمال به کار رفته در تحقیق، معرفی شده‌اند.

3-3-1 مروری بر روش‌های رده‌بندی

رده‌بندی، یک جز اساسی سیستم‌های شناسایی الگو است که در آن از ویژگی‌های داده برای تعیین دسته مناسب آن استفاده می‌شود. یک مثال ساده از مسئله رده‌بندی این است که ما می‌خواهیم با استفاده از ویژگی‌های هوا مانند وضعیت، دما، رطوبت و شرایط باد، تصمیم بگیریم که بیرون برویم یا خیر. در این مثال دو رده وجود دارد: بیرون رفتن یا نرفتن. یک رده‌بند هوا به عنوان ورودی ویژگی‌ها هوا را می‌گیرد و خروجی ((بیرون رفتن)) یا ((بیرون نرفتن)) را تولید می‌کند. به طور کلی رده‌بند یاد خواهد گرفت که چگونه روزهای بد و خوب را برای بیرون رفتن از روی ویژگی‌های هوا تشخیص دهد.

یک رده‌بند یکتابع است که یک نمونه بدون برچسب را به یک برچسب نگاشت می‌کند. یک استنتاج کننده، یا یک الگوریتم استنتاج، یک رده‌بند را از روی یک مجموعه داده، ایجاد می‌نماید. به عنوان مثال C4.5 استنتاج کننده درخت تصمیم‌گیری هستند که رده‌بندهای درخت تصمیم‌گیری را ایجاد می‌کنند. فرض کنید V فضای نمونه‌های بدون برچسب باشد و Y مجموعه برچسب‌های ممکن باشد. فرض کنید که $X \in V \times Y$ فضای نمونه‌های برچسبدار و $D = \{x_1, x_2, \dots, x_n\}$ یک پایگاه داده شامل n نمونه برچسب خورده باشد، که $x = \langle v_i \in V, y_i \in Y \rangle$. یک رده‌بند C یک نمونه بدون برچسب $v \in V$ را به برچسب $y \in Y$ نگاشت و یک استنتاج کننده I پایگاه داده D را به رده‌بند C نگاشت می‌کند. عبارت $(v, I(D, v))$ برچسب نسبت داده شده به یک نمونه بدون برچسب را با رده‌بند ساخته شده با استنتاج کننده I روی پایگاه داده D نشان می‌دهد. فرض می‌شود که یک توزیع روی مجموعه نمونه‌های بدون برچسب وجود دارد و همچنین پایگاه داده، شامل نمونه‌های مستقل و یکسان توزیع شده است [1].

روش‌های متعددی برای رده‌بندی وجود دارند. از جمله این روش‌ها رده‌بند بیز، درخت‌های تصمیم‌گیری، شبکه‌های عصبی و ماشین بردار پشتیبان را می‌توان نام برد. رده‌بند تصادفی¹ به رده‌بندی گفته می‌شود که بدون یادگیری، به طور تصادفی برای ورودی‌های خود، خروجی تولید می‌نماید. دقت یک رده‌بند تصادفی، $1/2$ است و به منظور ارزیابی دقت دیگر رده‌بندها استفاده می‌گردد. یک رده‌بند در پاسخ

¹ Random predictor

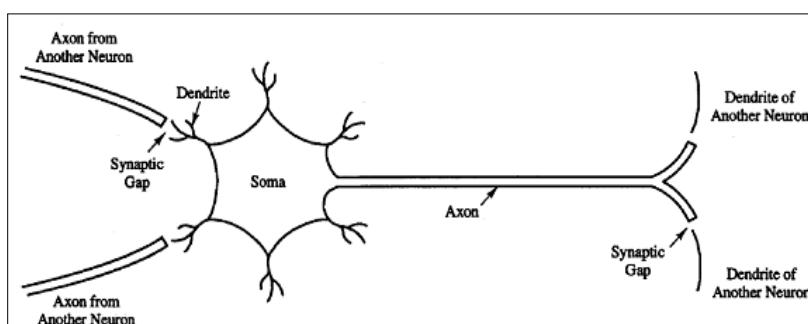
به الگوهایی که بر روی آنها آموزش ندیده است، مانند ردهبند تصادفی عمل میکند. در بخش بعد، شبکههای عصبی شرح داده شده‌اند.

3-3-2 شبکههای عصبی

شبکه عصبی مصنوعی ایده‌ای است برای پردازش اطلاعات که از سیستم عصبی زیستی الهام گرفته شده و مانند مغز به پردازش اطلاعات می‌پردازد. عنصر کلیدی این ایده، ساختار جدید سیستم پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی به هم پیوسته تشکیل شده (نورون‌ها¹) که برای حل یک مسئله با هم هماهنگ عمل می‌کنند. شبکههای عصبی نظیر انسان‌ها، با مثل یاد می‌گیرند. یک شبکه عصبی برای انجام وظیفه‌ای مشخص، مانند شناسایی الگوها و دسته‌بندی اطلاعات، در طول یک پروسه یادگیری، تنظیم می‌شود. در سیستم‌های زیستی، یادگیری با تنظیماتی در اتصالات سیناپسی که بین اعصاب قرار دارد همراه است. شبکههای عصبی نیز از همین روش استفاده می‌کنند [12].

ساختار نورون طبیعی

گفته شد که سیستم عصبی بیولوژیکی و شبکههای عصبی مصنوعی از نورون‌ها تشکیل شده‌اند. ساختار یک نورون طبیعی در شکل(3-5) نشان داده شده است.



شکل 3-5) ساختار یک نورون طبیعی [12]

هر نورون طبیعی از سه قسمت اصلی تشکیل شده است: بدنه سلول²، دندریت³ و اکسون⁴. دندریت‌ها به عنوان مناطق دریافت سیگنال‌های الکتریکی، آن‌ها را به هسته سلول منتقل می‌کنند. بدنه سلول انرژی لازم را برای فعالیت نورون فراهم کرده و بر روی سیگنال‌های دریافتی عمل می‌کند که با یک عمل ساده

¹ Neuron

² Soma

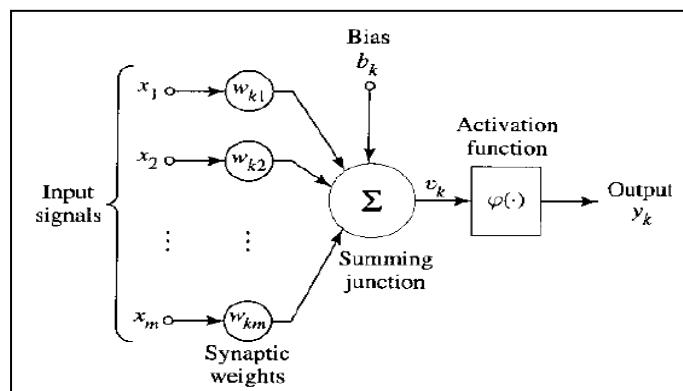
³ Dendrite

⁴ Axon

جمع و مقایسه با یک سطح آستانه مدل می‌گردد. اکسون سیگنال‌های الکتروشیمیایی دریافتی از هسته سلول را به نورون‌های دیگر منتقل می‌نماید. محل تلاقی یک اکسون از یک سلول به دندانیت‌های سلول‌های دیگر را سیناپس می‌گویند. توسط سیناپس‌ها ارتباطات بین نورون‌ها برقرار می‌شود. زمانی که سیگنال عصبی از اکسون به نورون‌ها میرسد، باعث تحریک آن‌ها می‌شود. نورون‌ها از هریک از اتصالات ورودی خود ولتاژی دریافت می‌کنند (توسط سیگنال عصبی ورودی) و آن‌ها را با یکدیگر جمع می‌زنند. اگر این حاصل جمع به یک مقدار آستانه رسید، اصطلاحاً نورون آتش می‌کند¹ و روی اکسون خود یک ولتاژ خروجی ارسال می‌نماید که این ولتاژ به دندانیت‌های متصل به این اکسون رسیده و باعث یکسری فعل و انفعال‌های شیمیایی در اتصالات سیناپسی می‌شود و می‌تواند باعث آتش کردن نورون‌های دیگر گردد. تمامی فعالیت‌های مغزی انسان توسط همین آتش‌کردن‌ها انجام می‌شوند [12].

ساختار نورون مصنوعی

نورون مصنوعی با الهام از نورون طبیعی به صورت شکل (6-3) مدل می‌شود.



شکل (6-3) ساختار یک نورون مصنوعی [12]

سه جزء مدل نورونی عبارتند از [12]:

- مجموعه‌ای از لینک‌های ارتباطی که هر یک با وزن خود مشخص می‌شوند. یک سیگنال ورودی x_j در ورودی سیناپس j متصل شده به نورون k در وزن w_{kj} ضرب می‌شود. مقادیر وزن می‌توانند مثبت یا منفی باشند.
- جمع‌کننده خطی برای جمع کردن سیگنال‌های ورودی وزن دار.
- تابع فعالیت برای محدود کردن دامنه خروجی نورون که با $\varphi(\cdot)$ نمایش داده می‌شود.

¹ Excite

مدل ذکر شده، دارای یک بایاس اعمال شده خارجی نیز است (b_k). اثر بایاس، بسته به مثبت یا منفی بودن، افزایش یا کاهش ورودی تابع فعالیت است. به طور ریاضی نوروون k را با معادلات زیر می‌توان توصیف نمود:

(3-1)

$$u_k = \sum_{i=1}^m w_{ki} x_i \quad (3-2)$$

$$v_k = u_k + b_k \quad (3-3)$$

$$y_k = \varphi(v_k)$$

خروجی نوروون و „خروجی ترکیب‌کننده خطی است. توابع فعالیت به سه نوع کلی تقسیم می‌شوند: تابع آستانه¹، تابع تکه‌ای-خطی² و تابع سیگموید³. فرمول‌های (4-3) و (5-3) دو تابع آستانه و تکه‌ای-خطی را توصیف می‌کنند. تابع سیگموید رایج‌ترین تابع فعالیت استفاده شده در شبکه‌های عصبی است. مثالی از سیگموید، تابع logistic است که معادله آن در فرمول (6-3) ارائه شده است.

(3-4)

$$\varphi(v) = \begin{cases} 1 & v \geq 0 \\ 0 & v < 0 \end{cases}$$

(3-5)

$$\phi(v) = \begin{cases} 1 & v \geq +1/2 \\ v & -1/2 > v > -1/2 \\ 0 & v \leq -1/2 \end{cases} \quad (3-6)$$

$$\varphi(v) = \frac{1}{1 + e^{-v}}$$

¹ Threshold

² Piecewise-linear

³ Sigmoid

معماری‌های مختلف شبکه‌های عصبی را می‌توان به صورت زیر طبقه‌بندی نمود [12]:

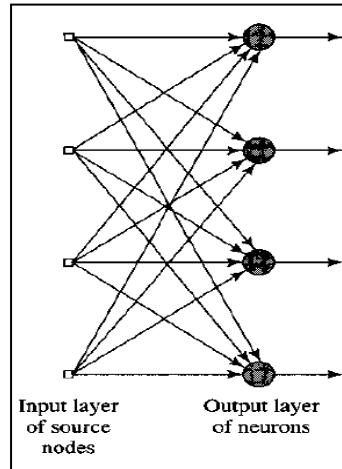
- شبکه‌های feed-forward: در یک شبکه عصبی لایه‌ای، نورون‌ها به صورت لایه‌ها سازماندهی می‌شوند. در ساده‌ترین حالت، یک شبکه لایه‌ای، شامل یک لایه ورودی (بردار ورودی شبکه عصبی) است که به یک لایه خروجی از گره‌ها (نورون‌های محاسباتی) متصل می‌گردد. جریان حرکت سیگنال‌ها یک‌طرفه از ورودی به خروجی است، به عبارت دیگر شبکه بدون سیکل یا feed-forward است. به این شبکه، شبکه feed-forward تک‌لایه گفته می‌شود. مثالی از یک شبکه feed-forward تک لایه در شکل (7-3) نمایش داده شده است. تک‌لایه، در واقع لایه خروجی است و لایه ورودی شمارش نمی‌گردد، زیرا محاسبه‌ای در آن صورت نمی‌گیرد. نوع دیگر شبکه‌های feed-forward، شبکه‌های چند‌لایه هستند. این شبکه‌ها دارای یک یا چند لایه پنهان¹ هستند که نورون‌های محاسباتی آن‌ها نورون‌های پنهان نامیده می‌شوند. عمل نورون‌های پنهان، مداخله بین گره‌های ورودی و نورون‌های خروجی شبکه به نحوی مفید است. اضافه کردن لایه‌های پنهان قدرت شبکه را افزایش می‌دهد. مثالی از یک شبکه چند لایه با یک لایه پنهان در شکل (8-3) مشاهده می‌شود. این شبکه کاملاً متصل² است، یعنی هر گره در هر لایه شبکه به هر گره در لایه بعدی متصل است. اگر برخی از لینک‌ها وجود نداشته باشند، شبکه متصل جزئی³ نامیده می‌شود. نمونه‌های مختلفی از شبکه‌های feed-forward طراحی شده‌اند، مانند RBFNN⁴.

¹ Hidden

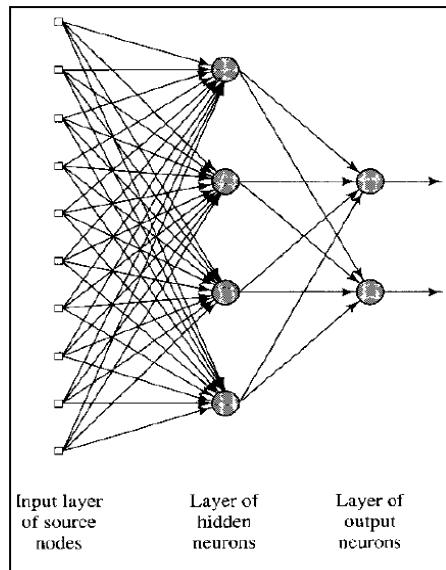
² Fully connected

³ Partially connected

⁴ Radial Basis Function Neural Network

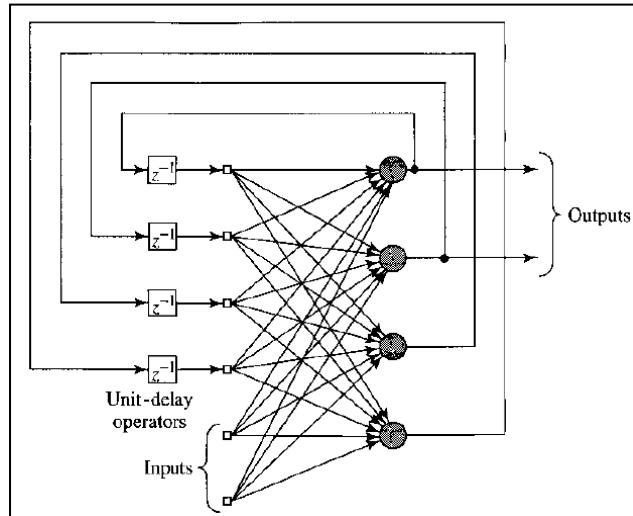


شکل ۳-۷) مثالی از یک شبکه تک لایه [12] feed-forward



شکل ۳-۸) مثالی از یک شبکه چند لایه [12] feed-forward

- شبکه‌های recurrent: این شبکه‌ها دارای حداقل یک حلقه فیدبک هستند. وجود حلقه، اثر عمیقی بر قابلیت یادگیری و کارائی شبکه دارد. حلقه‌های فیدبک، شامل واحدهای تاخیرند.
- نمونه‌ای از یک شبکه recurrent در شکل (9-3) مشاهده می‌شود.



شکل 9-3) نمونه‌ای از یک شبکه recurrent [12]

- معماری‌هایی مانند شبکه Kohonen که قدمت کمتری نسبت به دو نوع معماری قبلی داشته و برخی از آن‌ها هنوز در مرحله معرفی قرار دارند.

پرسپترون و پرسپترون چند لایه

شبکه پرسپترون، ساده‌ترین شبکه feed-forward تک‌لایه است و یک ردیبد خطي به شمار می‌رود.تابع فعالیت این شبکه به صورت زیر تعریف می‌شود:

(3-7)

$$f(x) = \begin{cases} 1 & \text{if } wx + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

شبکه پرسپترون چند لایه (MLP)، توسعه پرسپترون استاندارد است. MLP دارای یک یا چند لایه پنهان است. توابع فعالیت نورون‌ها در این شبکه غیرخطی‌اند. قابلیت MLP بیش از پرسپترون است، زیرا می‌تواند داده‌هایی که جداینیر خطی نیستند را جدا نماید [12].

یادگیری در شبکه‌های عصبی

هر شبکه عصبی شامل دانشی است که به صورت وزن‌ها ذخیره شده است. یادگیری به معنای تعیین این وزن‌هاست. از الگوریتم‌های یادگیری به منظور تغییر وزن‌ها در طول آموزش استفاده می‌شود. روش‌های یادگیری عبارتند از [12]:

- یادگیری با معلم¹: در این روش به هر نورون خروجی گفته می‌شود که پاسخ دلخواه به هر ورودی چیست. در این روش هدف تعیین مجموعه‌ای از وزن‌های خطا برین مقادیر خروجی دلخواه و محاسبه شده را حداقل نمایند. برای آموزش شبکه، یک نمونه آموزشی به آن داده می‌شود و بر اساس پاسخ شبکه به نمونه، وزن‌ها اصلاح می‌گردند. آموزش بر روی نمونه‌های آموزشی تکرار می‌شود تا جائی که شبکه به یک وضعیت پایدار برسد و تغییر محسوسی در وزن‌ها اتفاق نیافتد. تابع خطا که هدف، یافتن مینیمم آن است، تابع هزینه² یا تابع کارائی³ نامیده می‌شود. از جمله توابع هزینه، جمع مربع خطاهای⁴ (SSE) و میانگین مربع خطاهای⁵ (MSE) را می‌توان نام برد. یکی از معروف‌ترین الگوریتم‌های یادگیری با معلم، الگوریتم ((انتشار به عقب))⁶ است. این الگوریتم برای آموزش شبکه‌های پرسپترون چند لایه، به کار می‌رود. گونه‌های تغییر یافته مختلفی از الگوریتم انتشار به عقب توسعه یافته‌اند، مانند الگوریتم ((انتشار به عقب انعطاف‌پذیر)).⁷
- یادگیری بدون معلم: این روش تنها بر پایه اطلاعات محلی عمل می‌کند و به خودسامانده⁸ نیز معروف است. به عنوان نمونه‌هایی از روش یادگیری بدون معلم می‌توان Hebbian و competitive را نام برد.
- یادگیری semi-supervised: روش‌های جدید یادگیری که زمان زیادی از معرفی آن‌ها نمی‌گذرد.
- هر گذر⁹ بر روی مجموعه داده آموزشی یک epoch نام دارد و آموزش شامل تعدادی epoch است. وزن‌ها و بایاس‌های شبکه می‌توانند بعد از اعمال هر نمونه اصلاح شوند (یادگیری درون خطی¹⁰) و یا بعد از هر epoch (یادگیری دسته‌ای¹¹). همچنین اگر فاز یادگیری و فاز عمل یک شبکه عصبی مجزا

¹ Supervised learning

² Cost function

³ Performance function

⁴ Sum squared error

⁵ Mean squared error

⁶ Back-propagation

⁷ Resilient back-propagation

⁸ Self-organizing

⁹ Pass

¹⁰ Online

¹¹ Batch

باشد، گفته می‌شود که شبکه به صورت برون خطی¹ یاد می‌گیرد و اگر یادگیری و عمل شبکه در یک زمان باشد، یادگیری درون خطی نامیده می‌شود.

مزایا و معایب شبکه‌های عصبی

برخی از مزیت‌های شبکه‌های عصبی عبارتند از [12]:

- یادگیری انطباق‌پذیر: شبکه‌های عصبی قابلیت یادگیری نحوه انجام وظایف بر پایه اطلاعات داده شده برای آموزش را دارند.

• خودسازماندهی: یک شبکه عصبی می‌تواند سازماندهی‌اش را با توجه به اطلاعاتی که در طول دوره یادگیری دریافت می‌کند، خود ایجاد نماید و نیاز به ایجاد یک الگوریتم برای انجام وظیفه‌ای خاص نیست.

• عملکرد بلادرنگ²: محاسبات شبکه عصبی می‌تواند به صورت موازی انجام شود و سخت‌افزارهای مخصوصی طراحی و ساخته شده‌اند که می‌توانند از این قابلیت استفاده کنند و زمان محاسبات را به شدت کاهش دهند.

• تحمل اشتباه: خرابی جزئی یک شبکه منجر به تنزل کارایی متناظر با آن می‌شود ولی قابلیت های شبکه می‌توانند تا حد زیادی حتی با خرابی‌های بزرگ حفظ شوند.

شبکه‌های عصبی معایبی هم دارند، از جمله فقدان قابلیت تفسیر نحوه عملکرد و امکان بیش‌پوشش³ شدن. بیش‌پوشش در بخش (3-3-3) توضیح داده شده است.

خصوصیات تئوری شبکه‌های عصبی

از جمله خصوصیات تئوری مهم شبکه‌های عصبی که هنگام استفاده از آن‌ها بایستی مد نظر قرار گیرند، به موارد زیر می‌توان اشاره نمود [12]:

- قدرت محاسباتی: پرسپترون چند لایه یک تخمین‌زننده جهانی است.
- ظرفیت: ظرفیت به توانایی شبکه عصبی در مدل کردن هر تابع داده شده گفته می‌شود و در واقع میزان اطلاعاتی است که در شبکه می‌تواند ذخیره شود.

• همگرائی: به طور کلی چیزی در مورد همگرائی یک شبکه عصبی نمی‌توان گفت، زیرا همگرائی وابسته به چندین عامل زیر است:

¹ Offline

² Real-time

³ Overfitting

- ممکن است بر روی سطح خط، چند مینیمم محلی وجود داشته باشد که این مساله به تابع هزینه و مدل بستگی دارد.

- روش بهینه‌سازی مورد استفاده، هنگامی که از یک مینیمم محلی خیلی دور باشد، ممکن است به مینیمم همگرا نشود.

- برای مقادیر زیاد داده و پارامتر بعضی روش‌ها غیر عملی‌اند.

3-3-3 بیشپوشش

بیشپوشش شدن از مفاهیم مهم در یادگیری ماشین است. معمولاً یک الگوریتم یادگیری با استفاده از یک مجموعه از نمونه‌های آموزشی، آموزش داده می‌شود. فرض می‌شود یادگیر به مرحله‌ای می‌رسد که قادر خواهد بود خروجی صحیح را برای نمونه‌های دیده نشده در طول آموزش نیز پیش‌بینی نماید و در واقع بتواند به شرایطی که در طول آموزش ندیده، تعمیم پیدا کند. با این وجود، بهویژه در شرایطی که آموزش بسیار طولانی می‌شود و یا وقتی تعداد نمونه‌های آموزشی بسیار محدودند، یادگیر به ویژگی‌های خاصی از داده‌های آموزشی منطبق می‌شود که چندان به تابع هدف مرتبط نیستند. در پروسه بیشپوشش شدن، کارائی مدل روی نمونه‌های آموزشی همچنان افزایش می‌یابد، اما روی داده‌های دیده نشده کاهش پیدا می‌کند. به پروسه بیشپوشش شدن شبکه‌های عصبی در طول آموزش، بیشآموزش¹ هم گفته می‌شود. بهمنظور اجتناب از بیشپوشش شدن، نیاز به استفاده از روش‌های دیگر مانند ارزیابی متقطع²، توقف زودهنگام³ و تنظیم⁴ است. این روش‌ها می‌توانند الگوریتم را وقتی که آموزش بیشتر منجر به تعمیم بهتر نمی‌شود، متوقف کنند. به عنوان مثال در روش توقف زودهنگام، مجموعه داده آموزشی به دو زیرمجموعه آموزش و تست تقسیم‌بندی می‌شود. آموزش، با استفاده از زیرمجموعه آموزش صورت می‌گیرد. از زیرمجموعه تست برای ارزیابی دقت مدل در حین آموزش استفاده می‌شود. روند آموزش تا زمانی ادامه پیدا می‌کند که کارائی بر روی زیرمجموعه آموزش، کمتر از کارائی بر روی زیرمجموعه تست باشد.

3-3-4 اثر فراموشی خطرناک

اثر فراموشی خطرناک از مشکلات رایج الگوریتم‌های یادگیری شبکه‌های عصبی است. هنگامی که اطلاعات جدیدی توسط شبکه عصبی یاد گرفته می‌شود، معمولاً اطلاعات یاد گرفته شده قبلی خراب و یا

¹ Overtraining

² Cross-validation

³ Early stopping

⁴ Regularization

به طور کامل از بین می‌روند. این مساله، تحت عنوان مشکل ((اثر فراموشی خطرناک¹)) ((تداخل¹ خطرناک)) و یا ((یادگیری سریال)) شناخته شده است. علت این امر را می‌توان به صورت زیر توضیح داد. تشخیص یک مجموعه اولیه از الگوها توسط شبکه عصبی، بدین معنی است که شبکه یک نقطه در فضای وزن‌ها یافته است ($W_{initial}$) که در آن نقطه، وزن‌های شبکه به گونه‌ای تعریف شده‌اند که شبکه را قادر به تشخیص این الگوها می‌سازند. حال اگر یادگیری با مجموعه جدیدی از الگوها ادامه یابد، شبکه به یک نقطه جدید در فضای وزن‌ها (W_{new}) حرکت می‌کند که متاظر با مجموعه‌ای از وزن‌هاست که امکان تشخیص الگوهای جدید را به شبکه می‌دهد. فراموشی خطرناک وقتی اتفاق می‌افتد که W_{new} ، نقطه مناسبی برای تشخیص الگوهای یادگرفته شده اولیه نباشد. یک راه حل عملی موثر برای رفع این مشکل، تکرار² است، بدین مفهوم که اطلاعات قبلی یاد گرفته شده، مجدداً همراه با اطلاعات جدید یاد گرفته شوند [13].

3-3-3 یادگیری گروهی

در برخی از کاربردها به ویژه آن‌ها که با مقادیر زیادی از داده برخورد دارند، مدل‌های یادگیری ماشین معمولاً دچار ضعف هستند. علت اصلی این است که با افزایش سایز داده‌ها، هزینه محاسباتی به صورت غیرخطی³ رشد می‌کند. برای مثال در شبکه‌های عصبی، در پروسه یادگیری به صورت دسته‌ای، تمام داده‌ها باقیستی به مدل داده شوند و خروجی متاظر هر یک محاسبه شود تا شبکه بتواند خطای کلی و جهت مناسب برای تغییر مکان مدل در فضای وزن‌اش را بیاید. این پروسه وقتی که مثلاً شامل ضرب ماتریس‌ها باشد، می‌تواند با مشکل مواجه شود، به ویژه برای شبکه‌هایی که دارای تعداد زیادی نورون هستند.

قابلیت تعمیم⁴ مدل، مشکل دیگر برخورد با داده‌های زیاد و پیچیده است. قابلیت اطلاعاتی یک مدل محدود است. برای مثال یک شبکه عصبی با یک ساختار خاص و تعداد نورون مشخص در هر لایه، می‌تواند تعداد محدودی الگو را ردپنده کند [14]. بر اساس این واقعیت، با افزایش تنوع الگوها، نیاز به افزایش تعداد پارامترهای مدل است (مانند تعداد نورون‌های هر لایه). افزایش تعداد پارامترها هم بر هزینه محاسباتی و هر بر قابلیت تعمیم مدل اثر نامطلوبی می‌گذارد. از یک طرف، یک مدل کوچک با تعداد کمی پارامتر قادر به یادگیری تمام الگوها نیست و از طرف دیگر، یک مدل با تعداد زیادی پارامتر با مشکل بیشپوشش شدن مواجه می‌شود [15].

¹ Interference

² Rehearsal

³ Non-linear

⁴ Generalization

برای برطرف کردن مشکلات ذکر شده، ایده ماشین گروهی مطرح شد. علاقمندی جامعه یادگیری ماشین در زمینه ماشین‌های گروهی در اواسط دهه 1990 آغاز شد و همچنان این زمینه تحقیقاتی بسیار فعال است. یک ماشین گروهی، شامل گروهی از تخمین‌زننده‌های خروجی ماشین (پیش‌بینی) از ترکیب خروجی‌های اعضاش تولید می‌شود. ایده ماشین گروهی بر مبنای قاعده ((تقسیم کن و پیروز شو))¹ است.

ماشین‌های گروهی از چند نظر می‌توانند مفید واقع شوند. اولاً کارائی کلی گروه، توسط هر یک از اعضاش به تنهایی دست نیافتنی است. علت آن است که با ترکیب پیش‌بینی‌های اعضا گروه، خطای پیش‌بینی هر یک از آن‌ها تا اندازه‌ای برطرف می‌شود. در واقع ماشین‌های گروهی، تخمین‌زننده‌های جهانی² هستند. دوماً به جای آموزش یک تخمین‌زننده بهوسیله تمام داده‌ها، از نظر محاسباتی کارترست که مجموعه داده به چند زیرمجموعه تقسیم شده، هر تخمین‌زننده بر روی یک زیرمجموعه آموزش داده شود و سپس پاسخ‌های تخمین‌زننده‌ها ترکیب گردند. به عنوان نمونه‌هایی از تخمین‌زننده‌هایی که این پروسه برای آن‌ها سودمند است، رگرسیون پرسه گوسی³، شبکه‌های عصبی، اسپلاین‌های هموار کننده⁴ و ماشین بردار پشتیبان را می‌توان نام برد، زیرا در این سیستم‌ها زمان آموزش به شدت با افزایش سایز داده‌های آموزشی افزایش می‌یابد. با استفاده از روش ماشین گروهی، پیچیدگی محاسباتی تنها به صورت خطی با سایز داده‌ها افزوده می‌شود [16]. همچنین نشان داده شده که ایده ماشین گروهی تاثیر خوبی بر قابلیت تعمیم مدل دارد [15].

ماشین‌های گروهی دارای معایبی نیز هستند، از جمله این‌که برای ذخیره چند یادگیر نیاز به فضای بیشتری هست و تحلیل مدل نیز دشوارتر است.

ماشین‌های گروهی را می‌توان به دو نوع ماشین‌های با ساختار‌های ایستا⁵ و ماشین‌های با ساختار پویا⁶ تقسیم کرد [12]. در ادامه این ساختار‌ها شرح داده شده‌اند.

ساختار‌های ایستا

¹ Divide and conquer

² Universal approximator

³ Gaussian process regression

⁴ Smoothing splines

⁵ Static

⁶ Dynamic

در این نوع ماشین گروهی، پاسخ‌های چند یادگیر توسط مکانیزمی که شامل سیگال ورودی نیست، ترکیب می‌شوند. ماشین‌های گروهی ایستا که ساختار ساده‌تری نسبت به نوع پویا دارند، شامل روش‌های میانگین‌گیری گروه¹ و بوستینگ² هستند.

میانگین‌گیری گروه

در این روش خروجی‌های یادگیرها به طور خطی ترکیب می‌شوند تا پاسخ کلی را تولید کنند. شکل (3-10) تعدادی از شبکه‌های عصبی آموزش دیده به طور مختلف (متخصصین³) را نشان می‌دهد. ورودی شبکه‌ها مشترک است و خروجی‌های آن‌ها به نحوی ترکیب می‌شوند تا خروجی نهائی یا تولید شود. به دو علت از چنین روشی استفاده می‌گردد:

- اگر ترکیب متخصصین با یک شبکه عصبی واحد جایگزین شود، تعداد زیادی پارامتر قابل تنظیم وجود خواهد داشت. زمان آموزش برای چنین شبکه بزرگی احتمالاً طولانی‌تر از زمان آموزش موازی مجموعه‌ای از متخصصین است.
- وقتی تعداد پارامترها نسبت به اندازه مجموعه داده‌های آموزشی زیاد باشد، خطر بیشپوشش شدن وجود دارد.

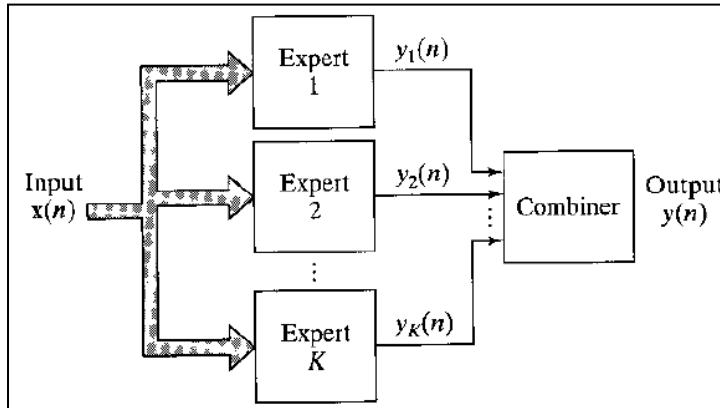
در روش میانگین‌گیری گروه انتظار می‌رود که شبکه‌های مختلف بر روی سطح خط⁴، به مینیمم‌های محلی مختلف همگرا شوند و با ترکیب خروجی‌ها، کارائی کلی بهبود یابد. روش‌های مختلفی برای آموزش هر یک از شبکه‌ها و ترکیب خروجی‌ها وجود دارد. یک روش آن است که شبکه‌ها ساختار و تنظیمات یکسان داشته باشند، اما آموزش آن‌ها از شرایط اولیه مختلف آغاز شود (وزن‌ها و بایاس‌های اولیه هر یک متفاوت باشند). خروجی‌ها را نیز می‌توان با یک میانگین‌گیری ساده ترکیب و پاسخ نهائی مدل را محاسبه نمود [12].

¹ Ensemble averaging

² Boosting

³ Experts

⁴ Error surface



شکل 10-3) روش میانگین‌گیری گروه [12]

بوستینگ

روش بوستینگ کاملاً با روش میانگین‌گیری گروه متفاوت است. در روش میانگین‌گیری گروه، تمام اعضا بر روی یک مجموعه داده آموزشی یکسان، آموزش داده می‌شوند و فقط شرایط اولیه آموزش آن‌ها با یکدیگر متفاوت است. در حالی که در روش بوستینگ، اعضا بر روی مجموعه‌هایی با توزیع کاملاً متفاوت آموزش داده می‌شوند. در این روش یک مجموعه یادگیر ضعیف، تبدیل به یک مجموعه یادگیر قوی با دقت زیاد به میزان دلخواه می‌شود. بوستینگ یک روش عمومی است که برای بهبود کارائی هر الگوریتم یادگیری می‌تواند استفاده شود و به سه صورت مختلف پیاده‌سازی می‌شود: بوستینگ با فیلتر کردن¹، بوستینگ با زیرنمونه‌برداری² و بوستینگ با وزن‌دهی مجدد³. در این بخش، دو نوع روش اول بوستینگ، شرح داده شده‌اند.

در روش بوستینگ با فیلتر کردن، ماشین گروهی شامل سه یادگیر است که به صورت زیر آموزش داده می‌شوند:

- یادگیر اول بر روی یک مجموعه با N_1 نمونه آموزش داده می‌شود. N_1 نمونه به طور تصادفی از مجموعه داده‌های آموزشی انتخاب می‌شوند.
- از یادگیر آموزش دیده اول، برای ساختن مجموعه آموزشی یادگیر دوم به نحوی که در زیر توضیح داده شده است، استفاده می‌شود:

- یک سکه را پرتاب می‌کنیم. اگر سکه رو⁴ آمد، نمونه‌های جدید یک به یک به یادگیر اول داده می‌شوند، سپس از نمونه‌های درست ردیابی شده صرف‌نظر می‌کنیم تا به یک

¹ Boosting by filtering² Boosting by sub-sampling³ Boosting by reweighting⁴ Head

نمونه غلط ردهبندی شده برسیم. نمونه مذکور به مجموعه داده آموزشی یادگیر دوم اضافه می‌شود.

- اگر سکه پشت¹ آمد، عکس عمل بالا انجام می‌شود، یعنی نمونه‌های درست ردهبندی شده در مجموعه داده آموزشی یادگیر دوم قرار می‌گیرند.

- این پروسه ادامه پیدا می‌کند تا N_1 نمونه توسط یادگیر اول فیلتر شوند. این نمونه‌ها مجموعه داده آموزشی یادگیر دوم را ایجاد می‌کنند.

با پرتاب سکه، این اطمینان حاصل می‌شود که اگر یادگیر اول بر روی مجموعه داده یادگیر دوم نست شود، خطای آن $1/2$ ، یعنی معادل خطای ردهبند تصادفی، خواهد بود. این مطلب بدین معنی است که خطای یادگیر اول بر روی مجموعه آموزشی یادگیر دوم، از خطای ردهبند تصادفی فاصله نگرفته و بنابراین الگوئی در مجموعه دوم را فرانگرفته است. به عبارت دیگر مجموعه داده دوم، یک توزیع کاملاً متفاوت از توزیع اول دارد و به این ترتیب یادگیر دوم ناچار می‌شود توزیعی متمایز از توزیع اول را یاد بگیرد.

• هنگامی که یادگیر دوم آموزش داده شد، یک مجموعه آموزشی برای یادگیر سوم به صورت زیر ایجاد و در انتها با این مجموعه داده، آموزش داده می‌شود:

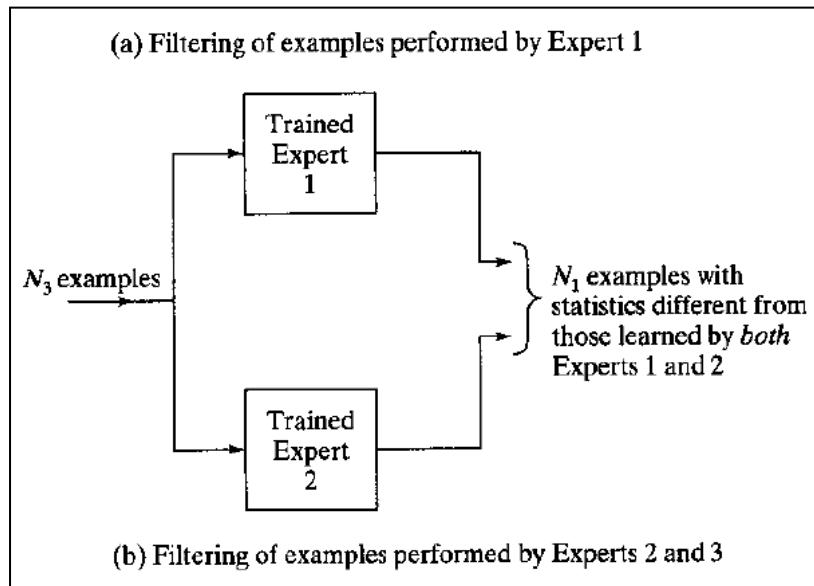
- یک نمونه جدید به دو یادگیر اول و دوم داده می‌شود. اگر هر دو آن نمونه را یکسان ردهبندی کردند، از نمونه صرف نظر می‌شود، در غیر این صورت نمونه به مجموعه داده آموزشی یادگیر سوم اضافه می‌گردد.

- این پروسه ادامه پیدا می‌کند تا N_1 نمونه به طور مشترک توسط یادگیر اول و دوم فیلتر شوند. این نمونه‌ها مجموعه داده آموزشی یادگیر سوم را تشکیل می‌دهند.

دقت شود که در هر مرحله، مجموعه داده آموزشی تشکیل شده برای هر یادگیر، از مجموعه داده‌های آموزشی کل حذف می‌گردد. این الگوریتم به نوعی هوشمند است، زیرا ماشین گروهی نیاز به یک مجموعه داده آموزشی بزرگ برای عملیات خود دارد، اما تنها زیرمجموعه‌ای از این داده‌ها را برای آموزش استفاده می‌کند (N_1). به علاوه، دو مرحله فیلترسازی بر روی بخش‌های ((مشکل برای یادگیری))² تمرکز دارند. برای ترکیب پاسخ سه یادگیر می‌توان از روش رایگیری ساده استفاده نمود (تعداد یادگیرها فرد در نظر گرفته می‌شود که در نهایت بتوان تصمیمگیری کرد).

¹ Tail

² Hard-to-learn



شکل 11-3) ساختار روش بوستینگ با فیلتر کردن [12]

یک الگوریتم معروف برای روش بوستینگ با زیرنمونه‌برداری، آدابوست¹ است. آدابوست، یک مجموعه از داده‌های آموزشی را دریافت و یک الگوریتم یادگیری پایه یا ضعیف داده شده را T دور فراخوانی می‌کند. ایده اصلی، نگهداری یک توزیع یا مجموعه‌ای از وزن‌ها بر روی مجموعه آموزشی است. در ابتدا تمام وزن‌ها مساوی‌اند، اما در هر دور وزن نمونه‌های اشتباه ردبندی شده افزوده می‌شود تا یادگیر ضعیف بیشتر مرکز بر مثال‌های سخت باشد و به این ترتیب، به تدریج قادر به یادگیری الگوهای مختلف گردد. در عمل، یادگیر ضعیف ممکن است الگوریتمی باشد که بتواند از وزن‌های نمونه‌های آموزشی استفاده کند. اگر این کار ممکن نباشد، یک زیرمجموعه از نمونه‌های آموزشی می‌تواند به طور تصادفی با توجه به وزن‌ها انتخاب شود و این نمونه‌های انتخاب شده، بدون وزن برای آموزش یادگیر ضعیف استفاده شوند.

بوستینگ به نحوی متفاوت از روش میانگین‌گیری گروه، خط را کاهش می‌دهد. در این روش، هر عضو گروه لازم است تنها کمی بهتر از حدس تصادفی عمل کند. قابلیت یادگیری ضعیف هر از اعضاء قابلیت یادگیری قوی تبدیل می‌شود، در حالی که خطای ماشین گروهی به اندازه دلخواه کوچک می‌گردد. این تبدیل قابل توجه، با فیلتر کردن توزیع داده‌های ورودی به نحوی که باعث شود یادگیرهای ضعیف نهایتاً کل توزیع را یاد بگیرند، یا با نمونه‌برداری مجدد از نمونه‌های آموزشی بر اساس یک توزیع احتمال

¹ Adaboost

در آدابوست، صورت می‌گیرد. مزیت آدابوست نسبت به بوستینگ با فیلتر کردن این است که آدابوست از مجموعه داده‌های آموزشی با اندازه ثابت استفاده می‌کند [12].

ساختارهای پویا

در نوع دوم از ماشین‌های گروهی، سیگنال ورودی مستقیما در مکانیزم مجتمع کردن خروجی‌های اعضا نقش دارد. در اینجا دو نوع ساختار پویا معرفی می‌شود: ترکیب یادگیرها¹ و ترکیب سلسله مراتبی یادگیرها². در مدل اول قاعده ((تقسیم کن و پیروز شو)) تنها یکبار و در مدل سلسله مراتبی چندین بار اعمال می‌شود که هر بار متناظر با یکی از سطوح سلسله مراتب است.

ترکیب یادگیرها

فرض شود که نمونه‌های مجموعه داده‌ها از چند توزیع مختلف بیرون کشیده شده‌اند. برای رده‌بندی چنین داده‌هایی می‌توان از مدل ترکیب یادگیرها (ME) استفاده نمود. شکل (12-3) ساختار این مدل را نشان می‌دهد. مدل ME شامل K یادگیر و یک واحد مجتمع‌سازی به نام شبکه دروازه³ است که به عنوان یک واسطه میان یادگیرها عمل می‌کند. فرض می‌شود که هر یادگیر در یک ناحیه از فضای ورودی، بهترین کارائی را دارد. در اینجا، یادگیرها شبکه‌های عصبی در نظر گرفته شده‌اند.

شکل (13-3) ساختار شبکه دروازه و یک نورون آن را نمایش می‌دهد. شبکه دروازه شامل یک لایه با K نورون است که هر نورون به یک یادگیر اختصاص داده شده است. برخلاف یادگیرها، نورون‌های شبکه دروازه غیرخطی‌اند و تابع فعالیت آن‌ها به صورت زیر تعریف می‌شود:

(3-8)

exp(u_i)

„ ضرب داخلی بردار ورودی x و بردار وزن u_i است: „

(3-9)

$$u_k = a_k^T x \quad k = 1, 2, \dots, K$$

¹ Mixture of experts

² Hierarchical mixture of experts

³ Gating network

دقیق شود که وابستگی خطی u_k به بردار ورودی x ، خروجی شبکه دروازه را یک تابع غیرخطی از x می‌سازد. خروجی‌های شبکه دروازه، دارای خصوصیات زیر هستند:

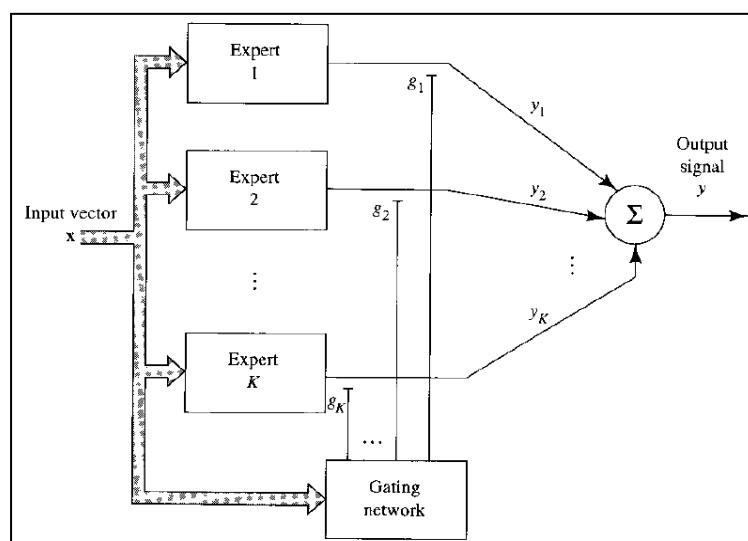
(10-3)

$$\begin{cases} 0 \leq g_k \leq 1 & \text{for all } k \\ \sum_{k=1}^K g_k = 1 \end{cases}$$

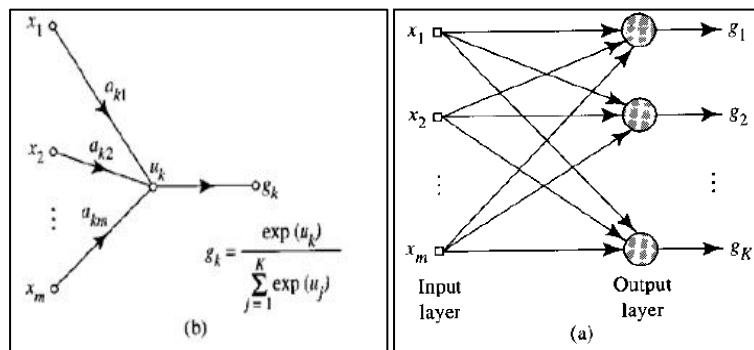
اگر خروجی یادگیر k ام در پاسخ به ورودی x باشد، خروجی کلی مدل ME به صورت زیر به دست می‌آید:

(3-11)

$$y = \sum_{k=1}^K g_k y_k$$



شکل 3-12) ساختار روش ترکیب یادگیرها [12]



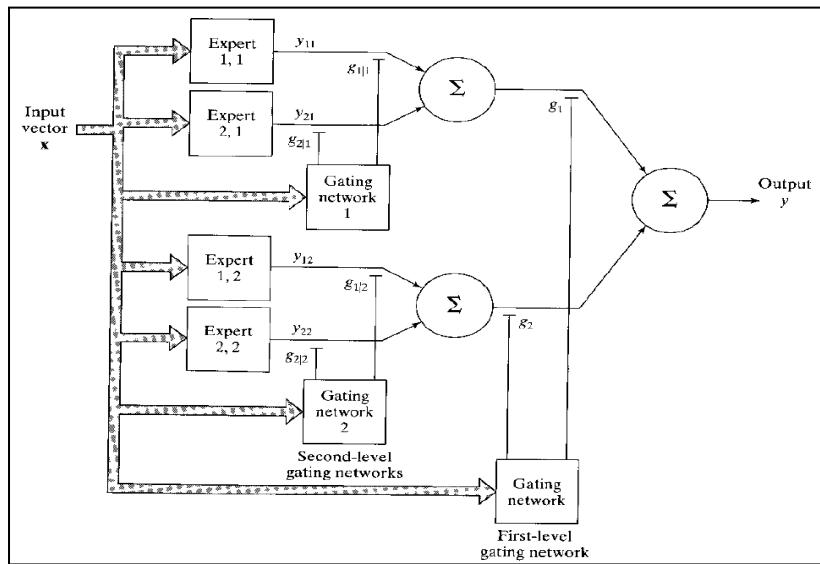
شکل 13-3 (a) ساختار شبکه دروازه (b) ساختار هر نورون آن [12]

ترکیب سلسله‌مراتبی یادگیرها

مدل ME فضای ورودی را به چند زیرفضا تقسیم و از یک شبکه دروازه برای توزیع اطلاعات (جمع‌آوری شده از داده‌های آموزشی) به یادگیرهای مختلف استفاده می‌کند. مدل ترکیب سلسله مراتبی یادگیرها (HME) توسعه ME است. معماری HME به صورت یک درخت است که شبکه‌های دروازه در گره‌های غیربرگ و یادگیرها در برگ‌ها هستند. HME فضای ورودی را به یک مجموعه تودرتو¹ از زیرفضاهای تقسیم می‌کند. اطلاعات میان یادگیرهای مختلف تحت نظرات چندین شبکه دروازه که به صورت سلسله‌مراتبی سازمان یافته‌اند، ترکیب و دوباره توزیع می‌شوند. شکل (14-3) یک HME با چهار عضو و دو سطح را نشان می‌دهد. همان‌طور که در شکل مشخص است، از سه شبکه دروازه در آن استفاده شده است.

مدل HME محصول استراتژی ((تقسیم کن و پیروز شو)) است. اگر معتقد باشیم که تقسیم فضای ورودی به چند ناحیه استراتژی خوبی است، پس به طور معادل تقسیم هر یک از آن نواحی به چند زیرناحیه نیز استراتژی خوبی بهشمار می‌رود. تقسیم نواحی را به طور بازگشتی می‌توان ادامه داد تا جایی که پیچیدگی سطوح تخمین‌زننده، یک برازش خوب به پیچیدگی محلي داده‌های آموزشی باشد [12].

¹ Nested



شکل 3-14) ساختار مدل ترکیب سلسله‌مراتبی یادگیرها

3-3-6 مفاهیم تئوری اطلاعات و آمار و احتمال

در این بخش چند مفهوم مرتبط به تئوری اطلاعات و آمار و احتمال که از آن‌ها در این تحقیق استفاده شده، معرفی شده‌اند.

آنتروپی اطلاعات

فرض کنید یک مجموعه گستره از نمادها $\{v_1, v_2, \dots, v_n\}$ وجود دارد. احتمال رخداد P_i با نمایش داده می‌شود. آنتروپی این توزیع، میزانی از تصادفی بودن یا غیرقابل پیش‌بینی بودن توالي نمادهایی است که از آن توزیع بیرون کشیده می‌شوند. آنتروپی برای توزیع‌های گستره به صورت زیر تعریف می‌گردد:

(3-12)

$$H = -\sum_{i=1}^n P_i \log P_i$$

برای هر نماد v_i که احتمال رخداد آن صفر باشد، بر اساس $\lim_{P_i \rightarrow 0} P_i \log P_i = 0$ ، تعریف می‌شود در فرمول بالا، اگر آنتروپی به بیت اندازه‌گیری شود، پایه لگاریتم 2 است. اگر پایه e باشد (لگاریتم طبیعی)، گفته می‌شود که آنتروپی به nats اندازه‌گیری می‌گردد. دقت شود که آنتروپی به خود نمادها وابسته نیست، بلکه به احتمال رخداد آن‌ها بستگی دارد. اگر توزیع یکنواخت باشد (احتمال رخداد همه یکسان باشد)، ((توزیع آنتروپی حداقل)) است، بدین معنا که حداقل عدم قطعیت در مورد هویت هر نماد بیرون کشیده شده از توزیع وجود دارد. اگر تمام P_i ‌ها به غیر از یکی از آن‌ها صفر باشند، ((توزیع

آنتروپی حاصل)) است، یعنی این اطمینان وجود دارد که همواره نماد بیرون کشیده شده از توزیع، نماد مذکور است. برای توزیع های پیوسته، آنتروپی به صورت زیر تعریف می شود [17]:

(3-13)

$$H = - \int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

اطلاعات دو جانبی^۱

اگر X و Y دو متغیر تصادفی باشند، به محتوای اطلاعاتی نسبی Y که در متغیر X وجود دارد، اطلاعات دو جانبی X و Y گفته می شود و به صورت زیر تعریف می گردد:

(3-14)

$$MI(X, Y) = \sum_{j=1}^n \sum_{k=1}^m p_{jk} \log \frac{p_{jk}}{p_j p_k}$$

در فرمول (14-3)، متغیرهای X و Y می توانند به ترتیب n و m مقدار متفاوت بگیرند. احتمال رخداد p_j مقدار j برای متغیر X ، احتمال رخداد مقدار k برای متغیر Y و احتمال رخداد همزمان مقدار j برای X و مقدار k برای Y است. با توجه به فرمول، $MI(X, Y) = MI(Y, X)$. اگر X و Y مستقل باشند، $MI(X, Y) = 0$ و این متناظر با این حقیقت است که با به دست آوردن اطلاعاتی در مورد X ، هیچ اطلاعاتی در مورد Y نمی توان به دست آورد [17].

تخمین پارامترهای توزیع احتمال

پارامترهای یک توزیع احتمال را می توان به وسیله نمونه هایی از آن توزیع تقریب زد و سپس از مقادیر تقریبی به جای مقادیر واقعی استفاده نمود. روش های متعددی برای تقریب زدن وجود دارد، مانند احتمال حداقل² (ML)، تخمین بیز و تطبیق ممان. در این بخش دو روش تطبیق ممان و ML معرفی شده اند .[17]

تطبیق ممان یک روش تخمین پارامترهای جمعیت است. ممان n ام توزیع احتمال $f(X)$ (حول صفر)، به صورت $E(X^n)$ ، یعنی امید " X " تعریف می شود. ممان k ام جمعیت را می توان توسط ممان k ام نمونه تخمین زد:

¹ Mutual information (MI)

² Maximum likelihood

(3-15)

$$E(X^n) \approx \frac{1}{n} \sum_{i=1}^n X_i^n$$

نمونه‌هایی هستند که از توزیع بیرون کشیده شده‌اند. در روش تطبیق ممان، ممان‌های نمونه $\{x_1, x_2, \dots, x_n\}$ با ممان‌های مشاهده نشده جمعیت معادل قرار داده شده و با حل معادلات، پارامترهای مورد نظر به دست می‌آیند. به عنوان مثال روند تخمین پارامترهای توزیع گاما به روش تطبیق ممان در ادامه توضیح داده شده است. فرمول (3-16) توزیع گاما را نشان می‌دهد.

(3-16)

$$\frac{x^{\alpha-1} e^{-x/\beta}}{\gamma(\alpha)}$$

ممان‌های اول و دوم توزیع گاما به صورت فرمول‌های (3-17) و (3-18) هستند:

(3-17)

$$E(X_1) = \alpha\beta$$

(3-18)

$$E(X_1^2) = \beta^2\alpha(\alpha+1)$$

فرمول‌های (3-17) و (3-18)، ((ممان‌های جمعیت)) هستند. ممان‌های اول و دوم ((نمونه)) به صورت فرمول‌های (3-19) و (3-20) تعریف می‌شوند:

(3-19)

$$m_1 = \frac{X_1 + \dots + X_n}{n}$$

(3-20)

$$m_2 = \frac{X_1^2 + \dots + X_n^2}{n}$$

با مساوی قرار دادن ممان‌های جمعیت و نمونه، تخمین دو پارامتر توزیع گاما به صورت زیر به دست می‌آید:

(3-21)

$$m^2 - m^2$$

تخمین ML، یک روش آماری برای برآورد یک مدل ریاضی به داده‌های جهان واقعی با استفاده از تخمین ML، پارامترهای آزاد مدل به گونه‌ای که تنظیم می‌شوند که یک برآورد مناسب به داده‌ها فراهم شود. برای یک مجموعه ثابت از داده‌ها و یک مدل، ML پارامترهای مدل را به نحوی انتخاب می‌کند که محتمل‌ترین حالت برای رخداد داده‌های داده شده، باشد. مجموعه D_θ از توزیع‌های احتمال را که با یک پارامتر ناشناخته θ (می‌تواند بردار باشد) با یک توزیع احتمال گسته یا پیوسته مرتبط است، در نظر می‌گیریم. این توزیع با f_θ نمایش داده می‌شود. n نمونه $\{x_1, x_2, \dots, x_n\}$ را از این توزیع بیرون می‌کشیم. سپس چگالی احتمال (چند متغیره) مرتبط با داده‌های مشاهده شده، با استفاده از محاسبه می‌گردد ($\ell(\theta)$). تابع likelihood به عنوان تابعی از θ با ثابت، $\{x_1, x_2, \dots, x_n\}$ تابع باشد، را حداکثر می‌کند، تخمین می‌زنند:

(3-22)

$$\ell(\theta) = f_\theta(x_1, x_2, \dots, x_n)$$

روش ML، پارامتر θ را با یافتن $\hat{\theta}$ ای که $\ell(\theta)$ را حداکثر می‌کند، تخمین می‌زنند:

(3-23)

$$\hat{\theta} = \arg \max \ell(\theta)$$

معمولًا فرض می‌شود که داده‌های بیرون کشیده شده، توزیع شده به طور یکسان و مستقل از یکدیگر¹ هستند. به این ترتیب مساله بسیار ساده شده، زیرا $\ell(\theta)$ را می‌توان به صورت ضرب n چگالی احتمال تک متغیره نوشت:

(3-24)

$$\ell(\theta) = \prod f_\theta(x_i) \Rightarrow \ell^*(\theta) = \sum \log f_\theta(x_i)$$

¹ Independent and identically distributed

ماکزیم عبارت بالا به طور عددی توسط روش‌های بهینه‌سازی مختلف می‌تواند محاسبه شود. باید توجه داشت که تخمین ML ممکن است یکتا نباشد یا اصلاً وجود نداشته باشد. به عنوان نمونه، فرض شود که یک سکه را n بار پرتاب می‌کنیم. در t پرتاب شیر می‌آید. احتمال شیر آمدن را p مینامیم (توزیع برنولی). با استفاده از ML می‌توان p را محاسبه نمود. بدین منظور،تابع likelihood باید روی تمام مقادیر

$$\text{ماکزیم شود: } n < t < n$$

$$(3-25) \quad (\theta) =$$

$$f_D(H = t|p) = \binom{n}{t} p^t (1-p)^{n-t}$$

یک راه ماکزیم کردن تابع بالا، مشتق گرفتن نسبت به p و مساوی قرار دادن با صفر است:

$$(3-26)$$

$$0 = \frac{\partial}{\partial p} \ell(\theta) = p^{t-1} (1-p)^{n-t-1} (t-np)$$

مقداری که $\ell(\theta)$ را حداقل می‌کند، $p=t/n$ می‌باشد. بنابراین می‌توان گفت که تخمین ML برای p ، مساوی مقدار t/n است.

P-value و فرضیه تهی¹

((فرض آماری)), ادعایی در مورد یک جمعیت مورد بررسی است که ممکن است درست یا نادرست باشد. به عبارت دیگر فرض آماری یک ادعا یا گزاره در مورد توزیع یک جمعیت یا پارامترهای توزیع یک متغیر تصادفی است. در استدلال‌های استقرائی و از جمله مسائل آماری، معمولاً نتایج حاصل از مشاهدات با در نظر گرفتن احتمالات تعمیم داده می‌شوند و هیچگاه هدف محقق اثبات مطلبی نیست، بلکه در مقام رد یا عدم رد فرضیه‌هاست (یعنی از برهان خلف استفاده می‌کند). به این‌گونه فرضیه‌ها که محقق در صدد رد یا تبرئه آن‌هاست، ((فرضیه صفر)) یا ((فرضیه تهی)) گفته می‌شود و آن را با H_0 نشان می‌دهند. در مقابل فرضیه تهی، ((فرضیه مخالف)) یا ((جایگزین²)) است (H_1) که در صورت رد H_0 مورد قبول واقع می‌شود. مثلاً چنانچه فردی در معرض اتهامی باشد، وی را به دادگاه برد و در دادگاه با فرض اینکه شخص بی‌گناه است، محاکمه وی آغاز می‌شود. بنابراین بی‌گناهی متهم فرضیه تهی است و اگر دادگاه اتهام او را رد نکند، او را گناهکار می‌شناسد (فرضیه جایگزین) [18].

¹ Null hypothesis

² Alternative

((آزمون آماری)), یک روش آماری است که محقق را به درک درستی یا نادرستی یک پارامتر از طریق جم‌آوری داده‌ها را هنمایی می‌کند (مثلاً در دادگاه از طریق سوال کردن از شاهدان عینی موضوع بررسی می‌شود). اگر نتیجه آزمایش، با فرض صحیح بودن فرضیه تهی، تفاوت معنی‌داری با آنچه که انتظار داریم داشته باشد، فرضیه تهی را رد می‌کنیم و در غیر این صورت آن را می‌پذیریم. چون در آزمون، نمونه‌های انتخاب شده و پارامتر محاسبه شده، حالتی از حالات ممکن است، می‌توان برای سهولت در قضاؤت، قبل از شروع آزمایش حدی را تعیین نمود که در خارج از آن فرض تهی رد گردد. بدین ترتیب ((قانون تصمیم‌گیری)) وضع می‌گردد. مثلاً اگر فرضیه تهی این باشد که پارامتر جمعیت 1000 است، می‌توان ((حد تصمیم‌گیری)) را 980 در نظر گرفت. یعنی هر میانگین بین 980-1000 به عنوان 1000 پذیرفته است، اما کمتر از آن پذیرفته نیست و گفته می‌شود ((اختلاف معنادار)) یا ((اختلاف آماری)) با حد 980 دارد. به عبارت دیگر، اگر برآورد پارامتر بیش از 980 باشد، بر اساس قانون تصمیم‌گیری، تفاوت آن با 1000 جزئیست و بیشتر در اثر اغتشاش و تصادف است و فرضیه تهی رد نمی‌گردد.

انتخاب حد تصمیم‌گیری بستگی به سلیقه محقق دارد. به منظور تعیین اصولی برای استاندارد نمودن نحوه وضع قوانین تصمیم‌گیری، می‌توان از مفهوم اشتباه نوع اول استفاده نمود. در بررسی یک فرضیه، دو نوع اشتباه نوع اول و دوم تعریف شده است. ((اشتباه نوع اول)) وقتی اتفاق می‌افتد که فرض تهی درست است ولی محقق به هر دلیل براساس آزمون تجربی آن را رد می‌کند. ((اشتباه نوع دوم)), خطای پذیرفتن یک فرض تهی غلط بر اساس آزمون آماریست. در یک تحقیق علمی می‌توان حداقل احتمال ارتکاب اشتباه نوع اول را تعیین و به عنوان حد تصمیم‌گیری استفاده نمود. این احتمال، ((سطح معنی‌دار بودن آزمون^۱)) نامیده می‌شود. در عمل از سطوح معنی‌دار 5% و 1% برای سهولت محاسبه استفاده می‌شود. به طور کلی هر عددی که منجر به رد فرضیه تهی می‌گردد، با پارامتر مطرح شده در فرضیه تهی اختلاف معنی‌دار دارد [13].

P-value، احتمال رخداد یک مشاهده با فرض درست بودن فرضیه تهی است. اگر P-value یا مساوی سطح معنی‌دار باشد، نتیجه‌گیری می‌شود که آزمایش با فرضیه تهی سازگار نیست و فرضیه مخالف قبول می‌شود و بالعکس (18). به عنوان مثال فرض می‌کنیم سکه‌ای داریم که می‌خواهیم اریب بودن یا نبودن آن را بررسی کنیم:

¹ Significant level

- فرضیه تهی: سکه نااریب است.
- فرضیه مخالف: سکه اریب است.
- سطح اهمیت: 5%
- آزمایش: سکه را 20 مرتبه پرتاب می‌کنیم، 14 مرتبه رو می‌آید.
- احتمال مشاهده 14 رو یا بیشتر در 20 پرتاب، یا 14 پشت یا بیشتر در 20 پرتاب، با فرض درست بودن فرضیه تهی (P-value): 0.1154
- نتیجه آزمون: مقدار محاسبه شده بیش از 0.05 است، پس نتیجه آزمایش با فرضیه تهی سازگار است و مشاهده 14 شیر در 20 پرتاب مربوط به شанс می‌شود.
- نتیجه‌گیری: در این مثال، ما نتوانستیم فرضیه تهی را در سطح اهمیت 5% رد کنیم.

فصل چهارم

نقشه تماس پروتئین

4-1 مقدمه

پروتئین‌ها از اجزایی اصلی سلول‌های موجودات زنده هستند. عمل پروتئین، وابسته به شکل ساختار سوم آن است. مولکول‌هایی که یک پروتئین می‌تواند به آن‌ها متصل شود، بستگی به شکل سه‌بعدی پروتئین دارد. پروتئین‌ها نقش اساسی در بیماری از بیماری‌هایی که امروزه با آن‌ها مواجه هستیم، دارند. بیماری آزاریمر، جنون گاوی و تعداد زیادی از سرطان‌ها، توسط پروتئین‌هایی به وجود می‌آیند که به درستی تا نمی‌شوند. مشخص کردن علت تاشدن‌های ناهنجار، می‌تواند به محققان در مبارزه با این بیماری‌ها کمک کند. اما تعیین ساختار سوم، به سادگی تعیین توالی پروتئین نیست. روش‌های فعلی تعیین ساختار سوم، پرهزینه و زمان‌بر هستند و محققان بر روی روش‌هایی کار می‌کنند که بتوانند ساختار سوم پروتئین را پیش‌بینی نمایند.

همان‌طور که در بخش (4-8) عنوان شد، ساختار سوم پروتئین با مختصات تمام اتم‌های آن در فضای سه‌بعدی مشخص می‌شود. ساختار سوم را می‌توان به نحوی ساده‌تر نمایش داد. این نمایش ساده‌تر که نقشه تماس نام دارد، مشاهده ویژگی‌های ساختار و طراحی الگوریتم‌های پیش‌بینی ساختار را تسهیل می‌کند. نقشه تماس یک ماتریس دو بعدی بولی¹ است که اندازه هر بعد آن به تعداد اسیدهای آmine زنجیره

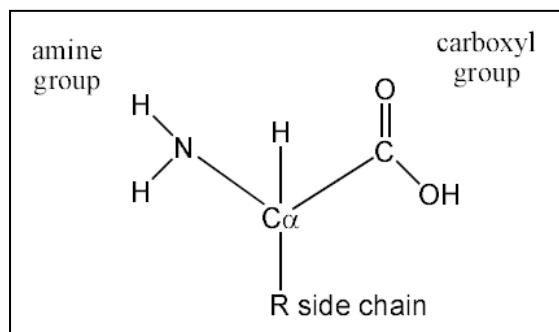
¹ Boolean

پروتئین است. ماتریس نشان می‌دهد که آیا اسیدهای آمینه، دو به دو، در فضای مجاورت یکدیگر قرار دارند یا خیر.

پیش‌بینی نقشه تماس در سال‌های اخیر در حوزه بیوانفورماتیک بسیار مورد توجه بوده است و روش‌های بسیاری برای پیش‌بینی نقشه تماس یک پروتئین از روی توالی آن پیشنهاد شده، زیرا نقشه تماس یک ساختمان داده مناسب برای پیش‌بینی ساختار پروتئین، توسط روش‌های آماری، یادگیری ماشین، و داده‌کاوی^۱ است. پیشرفت‌هایی در ساخت مدل‌های پیش‌بینی ساختار و بینشی جدید در مورد نحوه تاشدن پروتئین‌ها با استفاده از این ساختمان داده حاصل شده است. در واقع نقشه تماس، پلی‌بین ساختار اول و پیش‌بینی ساختار سوم به شمار می‌رود [9].

4-2 معرفی نقشه تماس

ساختار یک اسید آمینه در شکل (1-4) مجدداً نمایش داده شده است. به کربن مرکزی که چهار گروه به آن متصل شده‌اند^۲ و به کربن گروه R، گفته می‌شود.



شکل 4-1) ساختار اسید آمینه [3]

به طور معمول، تماس بین اسیدهای آمینه در فضای سه‌بعدی، بر حسب فاصله بین آن‌ها بیان می‌شود.

فاصله بین دو اسید آمینه معمولاً به یکی از صورت‌های زیر در نظر گرفته می‌شود [9]:

- مینیمم فاصله بین تمام اتم‌ها
- مینیمم فاصله بین تمام اتم‌های ستون فقرات
- مینیمم فاصله بین تمام اتم‌های زنجیره جانبی
- فاصله بین اتم‌های^۳

¹ Data mining

- فاصله بین اتم‌های C برای اسید آمینه (glycine)

- فاصله بین اتم‌های کربن گروه‌های کربوکسیل

- مراکز هندسی¹ دو اسید آمینه

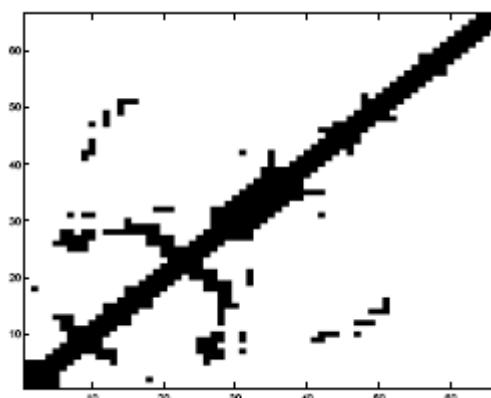
اگر فاصله تعریف شده کمتر از یک مقدار آستانه (thr) باشد، گفته می‌شود که بین دو اسید آمینه تماس وجود دارد.

برای یک پروتئین با N اسید آمینه، نقشه تماس، یک ماتریس $C = N \times N$ است که هر عضو آن به صورت زیر تعریف می‌شود [9]:

(4-1)

$$C_{ij} = \begin{cases} 1 & \text{if } D_{ij} < thr \\ 0 & \text{otherwise} \end{cases} \quad i, j = 1, \dots, N$$

که فاصله بین اسیدهای آمینه i و j است. به ماتریسی که فاصله بین هر جفت اسید آمینه یک پروتئین را نشان می‌دهد، ماتریس فاصله² گفته می‌شود. بنابراین می‌توان نقشه تماس را به عنوان یک ماتریس فاصله آستانه‌گذاری شده در نظر گرفت. شکل (4-2) نقشه تماس یک پروتئین را نمایش می‌دهد.



شکل (4-2) نقشه تماس پروتئینی به طول 70 اسید آمینه [4]

دقت شود که برای هر جفت اسید آمینه دو نوع فاصله وجود دارد، یکی ((فاصله توالی)) و دیگری ((فاصله فضائی)). فاصله توالی دو اسید a_j و a_i که با $|i-j|$ نمایش داده می‌شود، تعداد اسیدهای بین این دو در زنجیره پلی‌پپتید است.

¹ Geometrical center

² Distance matrix

4-2-1 تماس‌های محلی

تماس‌های یک پروتئین به دو دسته محلی و غیرمحلی تقسیم‌بندی می‌شوند. تماس‌های محلی به تماس‌های بین اسیدهای آمینه‌ای گفته می‌شود که فاصله توالی آن‌ها کمتر از یک مقدار آستانه است.

4-2-2 خطای نقشه تماس

به میانگین اختلاف بین ماتریس‌های فاصله دو پروتئین، خطای ماتریس فاصله^۱ (DME) گفته می‌شود. دو ماتریس a و b به صورت فرمول (4-2) تعریف می‌شود. از آستانه loc برای حذف فاصله‌های محلی استفاده می‌گردد. DME می‌تواند به صورت میانگین اختلاف‌ها (فرمول (4-2)) و یا به صورت جذر میانگین مربعات اختلاف‌ها² در نظر گرفته شود. به طور مشابه، خطای نقشه تماس³ (CME) به صورت فرمول (3-4) تعریف می‌شود. CME تقریبی از DME است [9].

$$(4-2) \quad DME(a,b) = \frac{\sum_{i=i}^{N-loc} \sum_{j=i+loc}^N |D_{ij}^a - D_{ij}^b|}{0.5(N-loc-1)(N-loc)}$$

$$(4-3) \quad CME(a,b) = \frac{\sum_{i=i}^{N-loc} \sum_{j=i+loc}^N |C_{ij}^a - C_{ij}^b|}{0.5(N-loc-1)(N-loc)}$$

4-2-3 ویژگی‌های نقشه تماس

نقشه تماس از مختصات سه‌بعدی اتم‌ها مستقل است. استقلال از مختصات به علاوه خاصیت بولی، امکان استخراج الگوهای موجود در نقشه تماس را توسط الگوریتم‌های داده‌کاوی و یادگیری ماشین، تسهیل می‌نماید.

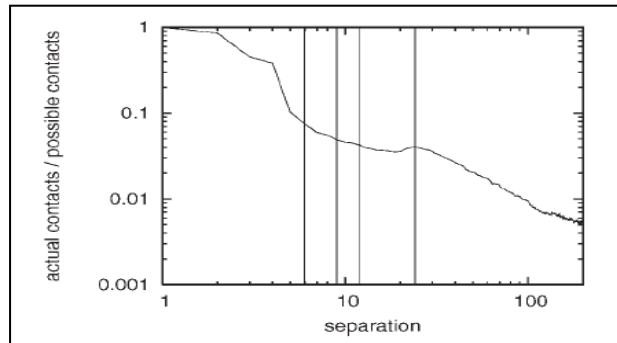
تعداد جفت‌هایی که با یکدیگر تماس ندارند، به صورت توان دوم طول زنجیره پلی‌پپتید افزایش می‌یابند. تعداد تماس‌های، به صورت خطی با طول زنجیره افزوده می‌شوند. شب خط و استگی، فقط به چگونگی تعریف تماس استگی دارد. با استفاده از فاصله‌های آستانه فاصله 8 آنگستروم و صرف

¹ Distance matrix error

² Root-mean-square distance difference

³ Contact map error

نظر از تماس‌های محلی $3|j-i|$ ، تعداد تماس‌ها در یک پروتئین کروی¹ تقریباً سه برابر طول پروتئین با انحراف معیار استاندارد 0.4 ± 0.4 است. از آنجا که هر تماس شامل دو اسید آمینه است، این عدد نشان می‌دهد که به طور تقریبی به ازای هر اسید آمینه، (0.8 ± 0.8) تماس وجود دارد. شکل (3-4) چگونگی کاهش تعداد تماس‌ها با افزایش فاصله توالی را نشان می‌دهد [9].



شکل 4(3) چگونگی کاهش احتمال تماس با افزایش فاصله توالی [9]

بیشتر تماس‌ها در پروتئین‌ها محلی هستند. در نظر گرفتن این موضوع در ارزیابی دقت پیش‌بینی نقشه تماس اهمیت دارد، زیرا پیش‌بینی تماس‌های محلی ساده‌تر از تماس‌های غیر محلی است. مرتبه تماس² یک پروتئین به میانگین فاصله توالی اسیدهای آمینه‌ای که با یکدیگر تماس دارند، گفته می‌شود. در برخی مطالعات، از مرتبه تماس به عنوان معیاری از پیچیدگی فضائی تای پروتئین استفاده می‌شود [9].

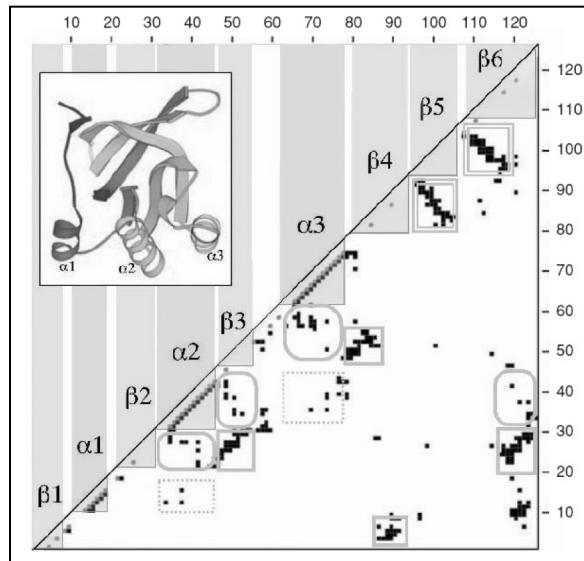
ساخترهای دوم یک پروتئین در نقشه تماس آن به آسانی قابل تشخیص هستند. مارپیچ‌های آلفا به صورت خطوطی پیوسته و موازی با قطر اصلی، بین اسیدهای $i+k$ و $i+j$ نمایان می‌شوند که k معمولاً 4 است. رشته‌های بتای موازی و ناموازی خطوطی پیوسته در ناحیه غیرقطري³ ایجاد می‌کنند. رشته‌های ناموازی عمود بر قطر اصلی قرار گرفته و تماس‌های بین اسیدهای $i-k$ و $i+j$ ، که k از صفر تا طول رشته تغییر می‌کند را نشان می‌دهند. رشته‌های موازی در امتداد قطر اصلی ظاهر می‌شوند و تماس‌های بین اسیدهای $i+k$ و $i+j$ را نشان می‌دهند. صفحات بتای ناموازی و موازی به صورت مجموعه‌ای از خطوط عمود بر قطر اصلی و موازی با آن دیده می‌شوند. تماس‌های بین مارپیچ‌های آلفا و دیگر

¹ Globular

² Contact order

³ Off-diagonal

ساختارها، به صورت خطوط شکسته در نقشه تماس ظاهر می‌گردند [9]. شکل (4-4) نقشه تماس آنزیم گلوتاتئون دیاستاز¹ را نشان می‌دهد. ساختارهای دوم در این شکل مشخص شده‌اند.



شکل 4-4) نقشه تماس یک آنزیم. نقاط خاکستری تماس‌های محلی $|i-j| < 3$ ، مستطیل‌های خاکستری رشته‌های بتای موازی، مستطیل‌های دوگانه رشته‌های بتای ناموازی، مستطیل‌های نقطه چین تماس‌های بین دو مارپیچ و مستطیل‌های گرد تماس‌های بین مارپیچ و رشته‌های بتا را نشان می‌دهند [9].

4-3 پیش‌بینی نقشه تماس

تعیین ساختار سوم پروتئین، به سادگی تعیین توالی پروتئین نیست و روش‌های آزمایشگاهی موجود، بسیار پرهزینه و زمان‌بر هستند و در نتیجه محققان بر روی روش‌هایی کار می‌کنند که بتوانند ساختار سوم پروتئین را صرفا بر اساس توالی اسیدهای آمینه آن پیش‌بینی نمایند. پیش‌بینی نقشه تماس، یکی از این روش‌های است. با داشتن نقشه تماس، می‌توان ساختار سوم را با دقت مناسب پیش‌بینی نمود [9]. هدف در مساله پیش‌بینی نقشه تماس، محاسبه تقریبی نقشه تماس یک پروتئین با استفاده از توالی اسید آمینه آن (ساختار اول) و ویژگی‌هایی است که صرفا از روی توالی قابل محاسبه و یا پیش‌بینی هستند.

4-3-1 نحوه ارزیابی پیش‌بینی نقشه تماس

معیارهایی که به طور معمول برای ارزیابی پیش‌بینی نقشه تماس به کار می‌روند، عبارتند از [9]:

¹ Glutathione reductase

- دقت^۱: دقت نشان می‌دهد چند درصد از تماس‌های پیش‌بینی شده صحیح هستند. در واقع دقت، نشان‌دهنده پیش‌بینی‌های مثبت درست^۲ (TP) است و به صورت زیر محاسبه می‌شود:

$$\text{دقت} = \frac{\text{تعداد تماس‌های پیش‌بینی شده صحیح}}{\text{تعداد کل تماس‌های پیش‌بینی شده}} \quad (4-4)$$

پیش‌بینی شده

- پوشش^۳: پوشش نشان می‌دهد چه درصدی از کل تماس‌های واقعی، پیش‌بینی شده‌اند و به صورت زیر تعریف می‌شود:

$$\text{پوشش} = \frac{\text{تعداد تماس‌های پیش‌بینی شده صحیح}}{\text{تعداد کل}} \quad (4-5)$$

تماس‌های واقعی

- بهبود نسبت به تصادف^۴: این معیار نشان می‌دهد که مدل پیش‌بینی کننده چه مقدار نسبت به پیش‌بینی تصادفی، بهتر عمل کرده است و به صورت زیر به دست می‌آید:

$$\text{بهبود نسبت به} \quad (4-6)$$

تصادف = دقت/دقت تصادفی

- که دقت تصادفی، احتمال پیدا کردن یک تماس را نشان می‌دهد و به روش زیر محاسبه می‌شود:

$$\text{دقت تصادفی} = \frac{\text{تعداد تماس‌های واقعی}}{\text{تعداد}} \quad (4-7)$$

جفت‌های اسید آمینه

- تعداد بلوک‌های پیش‌بینی شده صحیح: یک نقشه تماس پیش‌بینی شده خوب، باید بتواند به طور یکتا به یک ساختار سه‌بعدی صحیح تصویر شود. نقشه تماس را می‌توان به بلوک‌هایی که هر یک نشان‌دهنده تماس بین اجزای ساختارهای دوم هستند، تقسیم نمود (شکل (4-4)). اگر یک مدل پیش‌بینی اغلب بلوک‌ها را تشخیص دهد ولی دقت کلی آن کم باشد، می‌تواند پیش‌بینی کننده خوبی بهشمار رود و نقشه تماس پیش‌بینی شده، قابلیت تصویر به ساختار سه‌بعدی مناسب را دارد. از طرف دیگر، اگر دقت کلی مدل زیاد باشد اما تماس‌های پیش‌بینی شده در نواحی محلی یا در بعضی از بلوک‌ها مرکز باشند، ساخت ساختار سه‌بعدی مناسب ممکن نیست. پس می‌توان تعداد بلوک‌های پیش‌بینی شده صحیح را به عنوان معیار دیگری در نظر

¹ Accuracy

² True Positive

³ Coverage

⁴ Improvement over random

گرفت. از آنجا که پیش‌بینی تماس‌های غیر محلی مشکل‌تر از پیش‌بینی تماس‌های محلی است، می‌توان به تماس‌های غیر محلی وزن بیشتری اختصاص داد.

- نسبت تماس‌های پیش‌بینی شده صحیح به تعداد کل پیش‌بینی‌ها، در مقابل تعداد کل پیش‌بینی‌ها: این معیار در بخش (5-6) توضیح داده شده است.

در فصل بعد، مروری بر کارهای انجام شده در زمینه پیش‌بینی نقشه تماس صورت گرفته است.

فصل پنجم

مروری بر کارهای پیشین

5-1 مقدمه

رویکردهای آماری و یادگیری ماشین متعددی برای پیش‌بینی نقشه تماس ارائه شده‌اند. از رویکردهای آماری می‌توان به جهش وابسته [20]، پتانسیل‌های مبتنی بر دانش¹ [21] و مدل پنهان مارکوف [22] اشاره کرد. روش‌های یادگیری ماشین ارائه شده، مبتنی بر مدل‌های شبکه عصبی [23-26] ، الگوریتم ژنتیک [27-28] ، ماشین‌های بردار پشتیبان [29] و قوانین وابستگی [30] هستند. از اطلاعات معمول مورد استفاده در روش‌ها جهش وابسته، حافظت توالی² و ساختارهای دوم را می‌توان نام برد. در ادامه چند روش ارائه شده برای پیش‌بینی نقشه تماس بررسی شده‌اند.

5-2 پیش‌بینی توسط شبکه‌های عصبی

شبکه‌های عصبی برای پیش‌بینی تماس‌های غیر محلی کارایی نسبتاً مناسبی دارند. به علاوه عملکرد شبکه‌های عصبی بر روی پروتئین‌های با طول زیاد، نسبت به روش‌های دیگر، بهتر است. Fariselli و Cascadio [23] یک روش پیش‌بینی مبتنی بر شبکه عصبی ارائه نموده‌اند. شبکه عصبی به عنوان مدل پیش‌بینی، برای هر جفت اسید آمینه یک پروتئین، ویژگی‌های استخراج شده برای آن دو را گرفته و جفت

¹ Knowledge-based potentials

² Sequence conservation

اسید آمینه را در رده ((تماسدار¹)) یا رده ((بدون تماس²)) ردهبندی می‌کند. ساختار شبکه‌های عصبی مورد استفاده، feed-forward و الگوریتم آموزش، بازگشت به عقب استاندارد است. برای محاسبه فاصله بین دو اسید آمینه، کمینه فاصله میان اتم‌های سنگین آن‌ها (تمام اتم‌ها، بهجز ظیروژن) با آستانه 4.5 آنگستروم در نظر گرفته شده و بازه $3|j-i|$ به عنوان تماس‌های محلی تعریف شده است. مجموعه داده‌ها از 1996 pdbselect-oct-1997 و pdbselect-oct-1997 انتخاب شده است. برای از بین بردن عدم تعادل بین تعداد جفت‌های تماسدار (نمونه‌های مثبت) و جفت‌های بدون تماس (نمونه‌های منفی)، از فاکتور احتمالی متوازن کننده [31] استفاده شده است. این فاکتور تعداد سیکل‌های الگوریتم بازگشت به عقب را برای نمونه‌های منفی، کاهش می‌دهد. 5 شبکه عصبی با ویژگی‌های متفاوت آموزش داده شده‌اند. هر شبکه یک لایه میانی با دو نورون دارد. لایه خروجی دارای یک نورون است که احتمال تماس یک جفت اسید آمینه را نشان می‌دهد. تعداد نورون‌های لایه ورودی بسته به ویژگی‌های مورد استفاده، متفاوت است. تمام شبکه‌ها یک ورودی برای طول توالی و یک ورودی برای فاصله توالی دو اسید آمینه دارند. طول و فاصله توالی هر دو به صورت نرمال شده به شبکه داده می‌شوند. نرمال کردن، با تقسیم این دو ویژگی بر حداقل طول پروتئین‌های مجموعه داده‌ها که 1000 در نظر گرفته شده، صورت می‌گیرد. با استفاده از این دو ویژگی، خصوصیات ویژه هر توالی در نظر گرفته می‌شود. شبکه 1، 212 نورون ورودی دارد. 210 نورون اول، تمام ترکیبات دوتائی ممکن اسیدهای آمینه را نشان می‌دهند ($210 = \frac{20+1}{2} \times 20^*$). به 209 نورون ورودی‌ای که نشان‌دهنده جفت اسید آمینه واقع در مکان‌های i و j است، مقدار یک و به 2 نورون دیگر، مقدار صفر داده می‌شود. 2 نورون دیگر، همان‌طور که گفته شد، طول توالی و فاصله توالی دو اسید i و j را نشان می‌دهند. شبکه 2، 214 نورون ورودی دارد. 212 نورون اول مشابه شبکه 1 هستند و 2 ورودی جدید، میزان آبگریزی همسایگی دو اسید آمینه را نشان می‌دهند. همسایگی هر اسید، پنجره‌ای به طول 7 در نظر گرفته شده است. شبکه 3، 216 ورودی دارد. 2 ورودی جدید، میزان حافظت توالی دو اسید آمینه را نشان می‌دهند. شبکه 4، پنجره‌ای از جفت‌های موازی و ناموازی همسایگی مکان‌های i و j را در نظر می‌گیرد. این پنجره شامل جفت‌های $\{i-1, j\}$, $\{i, j\}$, $\{i+1, j\}$, $\{i-1, j+1\}$ و $\{i+1, j+1\}$ است. این شبکه، 1050 ورودی برای پنجره ذکر شده (210^*5), 2 ورودی برای میزان آبگریزی همسایگی و 2 ورودی برای طول توالی و فاصله توالی، یعنی مجموعاً 1054 ورودی دارد. شبکه 5، دارای 1056 ورودی است. 1054 ورودی مشابه شبکه 4 هستند و 2 ورودی دیگر میزان

¹ Contact² Non-contact

حافظت توالی دو اسید آمینه را نشان می‌دهند. معیارهای ارزیابی، دقت و بهبود نسبت به تصادف در نظر گرفته شده‌اند. نتایج برای 5 شبکه به صورت زیر است:

- شبکه 1: دقت=0.144 و بهبود نسبت به تصادف=5.4.
- شبکه 2: دقت=0.147 و بهبود نسبت به تصادف=5.5.
- شبکه 3: دقت=0.150 و بهبود نسبت به تصادف=5.6.
- شبکه 4: دقت=0.151 و بهبود نسبت به تصادف=5.7.
- شبکه 5: دقت=0.160 و بهبود نسبت به تصادف=6.

نتایج نشان می‌دهد که هر چه میزان اطلاعات ورودی بیشتر باشد، پیش‌بینی دقیق‌تری حاصل می‌گردد. در ادامه، Fariselli و همکاران [24] به بهبود روش قبلی خود [23] پرداخته و اثر افزودن ویژگی‌های بیشتر بر دقت پیش‌بینی نقشه تماس را بررسی نموده‌اند. 5 شبکه عصبی با ویژگی‌های مختلف آموزش داده شده‌اند که پیچیدگی ویژگی‌ها به تدریج افزایش می‌یابد. معیار تماس، فاصله بین اتم‌های C_β با آستانه 8 آنگستروم در نظر گرفته شده است. محدوده $6 \leq |j-i|$ به عنوان تماس‌های محلی در نظر گرفته شده است. ساختار شبکه‌های عصبی مشابه روش قبل [23] است، با این تفاوت که لایه پنهان 8 نورون دارد. نرخ یادگیری شبکه عصبی و مومنتم (دو پارامتر الگوریتم بازگشت به عقب) به ترتیب به 0.1 و 0.9 مقداردهی شده‌اند. MSA و حافظت توالی از پایگاه داده HSSP برای هر پروتئین استخراج و جهش وابسته، بر اساس روش Gobel و همکاران [10] محاسبه شده است. شبکه 1 مشابه روش قبل [23، 210] ورودی برای همه جفت‌های ممکن اسید آمینه دارد، با این تفاوت که به جای مقداردهی نورون متناظر با جفت i وز به یک و مقداردهی بقیه نورون‌ها به صفر، هر نورون با فرکانس رخداد جفت متناظر در MSA مقداردهی می‌شود. به این صورت که دفعات رخداد i و j در توالی‌های MSA شمارش و بر تعداد توالی‌ها تقسیم می‌گردد. شبکه 2، علاوه بر ورودی‌های شبکه 1، 1 ورودی برای جهش وابسته اسیدهای i و j و 2 ورودی برای میزان حفاظت این دو مکان دارد. شبکه 3، مشابه شبکه 2 است، با این تفاوت که جهش وابسته و حفاظت توالی را برای همسایگی‌های i و j در نظر می‌گیرد. این همسایگی شامل اسیدهای $i-1$ ، $i+1$ ، $i-j$ و $i+j+1$ است. در نتیجه 6 ورودی برای حفاظت توالی و 9 ورودی برای جهش وابسته دارد. شبکه 4، علاوه بر ورودی‌های شبکه 2، دارای 1 ورودی برای فاصله توالی نرمال شده است. نرمال‌سازی توسط تقسیم بر طول توالی صورت می‌گیرد. شبکه 5، مشابه شبکه 2 است، اما 18 ورودی نیز برای ساختارهای دوم پیش‌بینی شده همسایگی‌های i و j دارد. برای هر مکان، سه حالت مارپیچ آلفا،

صفه بتا و کویل در نظر گرفته شده است. پیش‌بینی ساختار دوم توسط روش Jacoboni و همکاران [32] صورت گرفته است. برای تعیین تعداد تماس‌های پیش‌بینی شده، خروجی‌های شبکه عصبی به صورت نزولی مرتب شده، سپس به تعداد نصف طول توالی از بیشترین مقادیر، تماس در نظر گرفته شده‌اند. این مقادیر معمولاً بیش از 0.75-0.8 هستند. نتایج پیش‌بینی بر اساس روش Olmea و Valencia [20] فیلتر شده‌اند. نتایج پیش‌بینی بر اساس دقت، بهبود نسبت به تصادف و توزیع فاصله تماس‌های پیش‌بینی شده ارزیابی شده‌اند. بهترین نتیجه را شبکه 5 با دقت 0.25 و بهبود نسبت به تصادف 8.05 دارد.

روشی برای پیش‌بینی نقشه تماس با استفاده از شبکه عصبی RBFNN ارائه Han و Zhang [25] نموده‌اند. مجموعه داده مورد استفاده، پروتئین‌های ویروس هپاتیت C هستند. در این مطالعه، فاصله بین مراکز هندسی اسیدهای آمینه در نظر گرفته شده است. پارامترهای شبکه عصبی با استفاده از الگوریتم ژنتیک تعیین شده‌اند. ویژگی‌های مورد استفاده شامل جفت اسید آمینه، ساختار دوم، بازه طول توالی، بازه فاصله توالی و رده جفت اسید آمینه است. طول توالی پروتئین به 4 بازه و فاصله توالی به 3 بازه تقسیم‌بندی شده است. بر حسب چند خصوصیت بیوشیمیائی، 10 دسته برای یک جفت اسید آمینه تعریف شده است. برای کاهش پیچیدگی محاسباتی، به جای پیش‌بینی ساختارهای دوم، ساختارهای واقعی موجود در فایل‌های PDB مورد استفاده قرار گرفته‌اند. این ساختارهای دوم، 8 حالت دارند که طبق روش Gue و همکاران [33] به 3 حالت کاهش داده شده‌اند. برای هر ویژگی به تعداد مناسب بیت در نظر گرفته شده و از یک ورودی دودوئی برای شبکه عصبی استفاده شده است. مدل با چند آستانه فاصله مختلف آزمایش و نقشه‌های تماس پیش‌بینی شده برای دو پروتئین آزمون، ترسیم شده‌اند.

در روش ارائه شده توسط Shackelford و Karplus [26]، از یک شبکه عصبی feed-forward با یک لایه پنهان به عنوان مدل پیش‌بینی‌کننده، استفاده شده است. الگوریتم آموزش، بازگشت به عقب انعطاف‌پذیر است. ویژگی‌های استخراج شده در این روش، عمدها از MSA هر پروتئین استخراج شده‌اند. اندازه بردار ورودی شبکه عصبی 449 می‌باشد. ویژگی اصلی این روش، تعریف معیاری کاراتر برای محاسبه جهش‌های وابسته است. این روش، بهترین نتیجه را در مسابقات CASP 2007¹ در زمینه پیش‌بینی تماس کسب نموده است (CASP مسابقه‌ای است که به منظور ارزیابی روش‌های پیش‌بینی ساختار پروتئین، هر دو سال یکبار از سال 1994 برگزار می‌شود). این مسابقه دارای چند گروه مانند پیش‌بینی ساختار دوم،

¹ Critical Assessment of Techniques for Protein Strucutre Prediction

پیش‌بینی ساختار سوم و پیش‌بینی تماس است). مدل پیشنهاد شده در این تحقیق، براساس همین روش است که جزئیات آن در فصل ششم شرح داده شده است.

5-3 پیش‌بینی توسط الگوریتم ژنتیک

Gupta و همکاران [27]، یک روش پیش‌بینی نقشه تماس با استفاده از الگوریتم ژنتیک ارائه داده‌اند. تماس‌های با فاصله توالی $3 < |z_i|$ به عنوان تماس‌های محلی در نظر گرفته شده‌اند. در این روش کروموزوم‌ها، نقشه‌های تماس مختلف برای یک توالی اسید آمینه هستند. در ابتدا ساختار دوم برای هر اسید آمینه پیش‌بینی می‌شود. مراحل روش به صورت زیر است:

- تولید جمعیت اولیه: یک جمعیت اولیه تصادفی تولید می‌شود. در نقشه تماس‌های تولید شده، بیت‌های مربوط به ساختارهای دوم پیش‌بینی شده یکسان و بقیه بیت‌ها تصادفی هستند.
 - تولید فرزند:
 - دو والد با برآزنده‌گی¹ بیشتر انتخاب می‌شوند. روش تعیین برآزنده‌گی در ادامه توضیح داده شده است.
 - عمل cross-over بر روی دو والد انتخاب شده انجام می‌شود.
 - عمل جهش بر روی هر بیتی که منتظر با ساختارهای دوم نباشد، اعمال می‌گردد.
- گفته شد که در مرحله انتخاب والد، میزان برآزنده‌گی نقشه تماس والد ارزیابی می‌شود. برآزنده‌گی، میزان محتمل² بودن یک نقشه تماس را نشان می‌دهد و توسط یک شبکه عصبی و بر اساس چند پارامتر تعیین می‌گردد. این پارامترها عبارتند از:

- مشابه یا مخالف بودن بار الکتریکی جفت اسید آمینه‌ای که با یکدیگر در تماسند.
- میزان آبگریزی همسایگی‌های جفت اسید آمینه
- میانگین فاصله بین اسیدهای آمینه‌ای که با یکدیگر تماس دارند.
- درجه رئوس اسیدهای آمینه (برای هر اسید آمینه، به صورت تعداد اسیدهای آمینه‌ای که با آن اسید آمینه در تماسند، محاسبه می‌شود).

از نقشه تماس پیش‌بینی شده برای پیش‌بینی ساختار سوم استفاده شده است.

¹ Fitness

² Feasible

5-4 پیش‌بینی توسط ماشین بردار پشتیبان

ماشین بردار پشتیبان (SVM) روشی برای رده‌بندی نمونه‌ها در دو رده و در فضای ویژگی‌های دلخواه است و از این‌رو روشی مناسب برای مساله پیش‌بینی نقشه تماس به‌شمار می‌رود. در روش Zhao و Karypis [29]، جفت اسیدهای آمینه‌ای که با یکدیگر تماس دارند، به عنوان نمونه‌های مثبت و جفت‌هایی که تماس ندارند، به عنوان نمونه‌های منفی در نظر گرفته شده‌اند. SVM یک ابرصفحه چندبعدی بهینه در فضای ویژگی‌ها، برای جدا کردن نمونه‌های مثبت و منفی از یکدیگر تعریف می‌کند. معیار تماس، فاصله بین اتم‌های C_α با آستانه ۸ آنگستروم است و محدوده $|r_i - r_j| \leq 6$ به عنوان تماس‌های محلی در نظر گرفته شده است. ویژگی‌های به کار رفته، فاصله توالی دو اسید، ساختارهای دوم پیش‌بینی شده، جهش وابسته و محافظت توالی هستند. پیش‌بینی ساختار دوم با استفاده از نرم افزار PSIPRED و تراز توالی چندگانه با نرم‌افزار ClustalW انجام شده است. جهش وابسته و حافظت توالی یکبار به صورت استاندارد و یکبار با استفاده از اندیس‌های AAindex محاسبه شده‌اند. مقایسه نتایج نشان می‌دهد که استفاده از تعریف استاندارد برای محاسبه جهش وابسته و اندیس‌های AAindex حافظت توالی نتایج بهتری در بردارند. این روش، ۱۵ مدل SVM را با استفاده از ترکیب‌های مختلف ویژگی‌های ذکر شده آموزش داده و نتایج را مقایسه نموده است. هر SVM توسط SVM_{light} [19]، با یک کرنل خطی و مقدار C پیش‌فرض آموزش داده شده است. برای ساخت مجموعه داده‌ها از پروتئین‌هایی استفاده شده که شواهد آن‌ها با یکدیگر کمتر از ۲۵% باشد و برای هر یک حداقل ۱۵ پروتئین شبیه یافت شود (توسط Blast). به دلیل متوازن نبودن تعداد نمونه‌های مثبت و منفی، از نمونه‌های منفی نمونه برداری^۱ شده تا تعداد نمونه‌های مثبت و منفی مساوی گردد. مدل‌ها بر روی کلاس‌های پروتئین پایگاه داده CATH آموزش داده شده‌اند. نتایج، با دو معیار دقت و بهبود نسبت به تصادف مورد بررسی قرار گرفته‌اند. نتایج بیان‌گر آن است که ساختار دوم در پیش‌بینی نقشه تماس پروتئین‌هایی که صفحات بتا دارند، مهمترین ویژگی است. از طرف دیگر، جهش وابسته و حافظت توالی برای پیش‌بینی نقشه تماس پروتئین‌هایی که دارای مارپیچ‌های آلفا هستند، حائز اهمیت هستند. به علاوه، نتایج نشان‌دهنده آن هستند که ساختارهای دوم پیش‌بینی شده و جهش وابسته هر یک مجموعه‌های متفاوتی از تماس‌ها را تشخیص می‌دهند. به علاوه آموزش مدل‌های متفاوت به ازای هر کلاس پروتئین CATH، می‌تواند منجر به بهبود نتایج گردد.

^۱ Sampling

5-5 پیش‌بینی توسط قوانین وابستگی

Zaki و همکاران [30] روشهایی برای پیش‌بینی نقشه تماس با استفاده از قوانین وابستگی ارائه نموده‌اند. فاصله، بین اتم‌های C_α با آستانه ۷ آنگستروم و بازه $|j-i| < 4$ به عنوان تماس‌های محلی در نظر گرفته شده است. در این روش، پایگاه داده‌ای از داده‌های آموزشی ایجاد شده که هر رکورد آن شامل جفت اسید آمینه، خروجی مدل پنهان مارکوف HMMSTR [34] و رده رکورد (تماس یا عدم تماس) است. قوانین وابستگی از این پایگاه داده استخراج شده و برای پیش‌بینی رده جفت‌های آزمون مورد استفاده قرار می‌گیرند. نتایج به دست آمده به صورت زیر هستند (N طول توالی است):

$$\bullet \quad N < 100 : \text{دقت}=0.26, \text{پوشش}=0.63 \text{ و بهبود نسبت به تصادف}=4$$

$$\bullet \quad 100 \leq N < 170 : \text{دقت}=21.5, \text{پوشش}=0.1 \text{ و بهبود نسبت به تصادف}=6$$

$$\bullet \quad 170 \leq N < 300 : \text{دقت}=0.13, \text{پوشش}=0.075 \text{ و بهبود نسبت به تصادف}=6.5$$

$$\bullet \quad N > 300 : \text{دقت}=0.097, \text{پوشش}=0.075 \text{ و بهبود نسبت به تصادف}=7.8$$

افزودن اطلاعات بیشتر به رکوردها، می‌تواند نتایج را بهبود بخشد. با وجود پوشش کم، مجموعه تماس‌های پیش‌بینی شده، شامل تماس‌هایی است که از نظر فیزیکی غیرممکنند و برای بهبود نتایج، لازم است حذف شوند.

6-1 مقدمه

فصل ششم روش پیشنهادی

روش پیشنهادی پیش‌بینی نقشه تماس، بر اساس ایده ماشین گروه، شبکه‌های عصبی هستند. استخراج ویژگی‌ها و طراحی شبکه‌ها بر مبنای روش Shackelford و Karplus [26] صورت گرفته است. برای پیش‌بینی نقشه تماس یک پروتئین، مجموعه‌ای از ویژگی‌ها برای هر جفت اسید آمینه پروتئین محاسبه و به عنوان ورودی به شبکه‌های عصبی داده می‌شوند. شبکه‌های عصبی، هر جفت اسید آمینه را در یکی از دو رده جفت‌های تماسدار (نمونه‌های مثبت) یا جفت‌های بدون تماس (نمونه‌های منفی) ردپندي مي‌نمایند. پاسخ نهائی گروه به عنوان رده هر جفت، در نظر گرفته می‌شود. پیاده‌سازی روش، با استفاده از زبان برنامه‌نویسی Java، محیط توسعه¹ Eclipse و نرم‌افزار MATLAB انجام شده است. بخش پیش‌پردازش داده‌ها و استخراج ویژگی‌ها عمدتاً توسط زبان Java و بقیه مراحل توسط نرم‌افزار MATLAB صورت گرفته است. در ادامه جزئیات مراحل روش پیشنهادی بیان شده‌اند.

¹ Integrated development environment (IDE)

6- فایل‌های داده

فایل‌های لازم برای آموزش و تست سیستم، از پایگاه داده PDB [35] بازیابی شده‌اند. مشابه برخی از روش‌های ارائه شده ([23-24] و [36])، انتخاب فایل‌های PDB مورد نظر بر اساس لیست-pdbselect- [37] صورت گرفته است. در نتیجه میزان شباهت بین فایل‌ها کمتر از 25% است که این 25%-May2008 امر باعث می‌شود تعداد الگوهای موجود در داده‌های آموزش و تست افزایش یابد. 80% از فایل‌ها برای آموزش سیستم و 20% باقی‌مانده برای تست سیستم مورد استفاده قرار گرفته‌اند. انتخاب مجموعه داده‌های آموزش و تست به صورت تصادفی صورت گرفته است.

6-2-1 پیش‌پردازش فایل‌ها

در ایجاد مجموعه داده‌ها، از فایل‌های PDB با مشخصات زیر صرف نظر شده است:

- فایل‌های چند مدلی¹
- فایل‌های که در ابتدا یا در میان رکوردهای ATOM آن‌ها، یک یا چند رکورد HETATM قرار دارد.
- فایل‌های که در ابتدا یا در میان رکوردهای ATOM آن‌ها، یک یا چند رکورد ANISOU قرار دارد.
- فایل‌هایی که شامل یک یا چند رکورد ATOM با کمتر از 12 فیلد هستند.
- فایل‌هایی که یک یا چند اسید آمینه مفقود² دارند.
- فایل‌هایی که در آن‌ها شماره یک یا چند اسید آمینه، یک عدد به اضافه یک حرف است، مانند .1A
- فایل‌هایی که در آن‌ها یک یا چند اسید آمینه، اتم ندارند (به غیر از اسید آمینه Glycine)
- فایل‌هایی که در آن‌ها یک یا چند اسید آمینه، جز 20 اسید آمینه استاندارد نیستند، مانند BMET و AMET

همان طور که گفته شد، فایل‌های PDB ممکن است شامل بیش از یک زنجیره باشند. در این تحقیق تنها اولین زنجیره هر فایل PDB که معمولاً زنجیره A است، در نظر گرفته شده است.

¹ Multi-model

² Missing

استخراج داده‌ها از هر فایل PDB که دارای مشخصات ذکر شده در بالا نباشد، به این ترتیب انجام شده که از هر اسید آمینه، فقط مختصات سه بعدی اتم C_β آن در نظر گرفته شده است. برای اسید آمینه Glycine که اتم C_α ندارد، اتم C_β آن در نظر گرفته شده است. به هر یک از اسیدهای آمینه به ترتیب کدی از 1 تا 20 نسبت داده شده است. فایل‌های نهائی، مجموعه‌ای از رکوردها با فیلدهای زیر هستند:

(مختصه_x, مختصه_y, مختصه_z, کد اسید آمینه)

6-3 استخراج ویژگی‌ها

ویژگی‌های مورد استفاده در شبکه‌های عصبی، عمدتاً از MSA پروتئین ورودی محاسبه شده‌اند. این ویژگی‌ها به سه دسته تقسیم می‌شوند، ویژگی‌های توالی، ویژگی‌های تک ستونی و ویژگی‌های جفت ستونی.

6-3-1 MSA به دست آوردن

پیش‌بینی نقشه تماس مشابه روش Shackelford و Karplus [26] بر اساس MSA های توالی‌های پروتئین صورت گرفته است. در این تحقیق، MSA های پروتئین‌های مجموعه داده، از سرویس دهنده وب SAM (SAM T-08) [26] و [38-42] به دست آمده‌اند. این سرویس توسط مدل پنهان مارکوف در میان توالی‌های NR (پایگاه داده پروتئین غیرتکراری¹ NCBI) [43]، توالی‌های شبیه به توالی هدف را یافته و آن‌ها را تراز می‌نماید. سیستم مدل‌سازی و تراز توالی² SAM، مجموعه‌ای از ابزارهای انعطاف‌پذیر برای ایجاد و استفاده از مدل‌های پنهان مارکوف به منظور تحلیل توالی‌های بیولوژیکی است. مدل‌ها می‌توانند بر روی یک خانواده از پروتئین‌ها یا اسیدهای نوکلئیک آموزش داده شده و سپس هم به منظور تولید MSA و هم به منظور جستجوی پایگاه داده برای اعضاً جدید خانواده به کار روند.

سرور SAM، یک توالی پروتئین به فرمت FASTA را به عنوان ورودی گرفته و اطلاعات بسیاری، از جمله MSA و ساختار دوم پیش‌بینی شده، در مورد آن تولید می‌کند. فرمت فایل‌های MSA سرور SAM a2m است. یک فایل a2m شامل هر تعداد توالی می‌تواند باشد. هر توالی با یک خط شناسائی³ شروع می‌شود. این خط حاوی اطلاعاتی در مورد توالی است و در سطر بعد از آن، خود توالی قرار دارد. برای پروتئین‌ها، الفبای به کار رفته به شرح زیر است:

ACDEFGHIKLMNPQRSTVWY • برای اسیدهای آمینه استاندارد

¹ NCBI's non-redundant protein database

² Sequence alignment and modeling system

³ Identifying

X برای هر اسید آمینه •

D برای N یا B •

E برای Q یا Z •

ترازبندی توسط کارکترهای کوچک، بزرگ و کارکتر ویژه فاصله "-" بیان شده است. کارکترهای بزرگ و کارکتر "-", نشان‌دهنده ستون‌های تراز هستند. تعداد ستون‌های تراز در تمام توالی‌ها باید یکسان باشد. کارکترهای کوچک نشان‌دهنده نقاط درج در بین ستون‌های تراز یا در دو انتهای توالی هستند. در پیاده‌سازی روش، بر روی فایل‌های a2m دریافت شده از سرور SAM پیش‌پردازش انجام شده است، به این ترتیب که ابتدا کلیه خطوط شناسائی از فایل حذف شده‌اند. سپس کارکترهای کوچک از توالی‌ها حذف شده‌اند. کارکترهای B به طور تصادفی با N یا D، کارکترهای Z به طور تصادفی با Q یا E و کارکترهای X به طور تصادفی با یکی از اسیدهای آمینه استاندارد جایگزین گردیده‌اند.

6-3-2 لاغر کردن¹ MSA

اگر توالی‌های یک MSA شباهت زیادی به یکدیگر داشته باشند، سیگنال‌هایی که به دنبال آن‌ها هستیم، به آسانی توسط سیگنال‌های دیگری که از تکرار توالی‌هایی از زیرخانواده‌های خاص ایجاد می‌شوند، مشوش می‌گردند. برای کاهش این اثر تصنیعی، تعداد توالی‌های هر MSA، توسط یک پروسه حریصانه² کاهش یافته‌اند [26]. کاهش تعداد به این صورت انجام می‌شود که اولین توالی به لیست توالی‌های نگه داشته شده اضافه می‌گردد. سپس هر توالی با تمام توالی‌های لیست مقایسه می‌شود. اگر میزان شباهت آن با هر یک از توالی‌های نگه داشته شده زیاد باشد، توالی حذف و در غیر این صورت به لیست اضافه می‌گردد. آستانه شباهت در این تحقیق، براساس روش Shackelford و Karplus [26]، 50% در نظر گرفته شده است.

6-3-3 استخراج ویژگی‌های توالی

ویژگی‌های توالی شامل طول توالی پروتئین و فاصله توالی دو اسید آمینه از یکدیگر است که به سادگی قابل محاسبه هستند.

¹ Thinning

² Greedy

6-3-4 استخراج ویژگی‌های تک ستونی

برای هر ستون MSA، چند ویژگی شامل توزیع اسیدهای آمینه، آنتروپی و ساختار دوم پیش‌بینی شده، استخراج شده است. در ستون‌های MSA، از توالی‌هایی که کارکتر فاصله دارند در محاسبه تمامی ویژگی‌ها، صرف نظر گردیده است.

توزیع اسیدهای آمینه

در روش Shackelford و Karplus [26]، با استفاده از یک تنظیم کننده ترکیبی دریکله¹ [44]، توزیع اسیدهای آمینه برای هر ستون MSA محاسبه شده است. در این تحقیق از یک روش تقریبی ولی ساده‌تر استفاده شده، به این صورت که برای هر ستون، تعداد رخداد هر یک از 20 اسید آمینه، شمارش و با تقسیم بر تعداد کل اسیدهای آمینه نرمال می‌شود. به این ترتیب احتمال تقریبی رخداد هر اسید آمینه در هر مکان پروتئین به دست می‌آید.

آنتروپی

برای هر ستون MSA، بر اساس فرمول (12-3)، آنتروپی توزیع اسیدهای آمینه محاسبه شده است. هر چه نوع اسیدهای آمینه یک ستون بیشتر باشد و در واقع آن مکان کمتر حفاظت شده باشد، مقدار آنتروپی متناظرش بیشتر خواهد بود.

ساختار دوم پیش‌بینی شده

برای هر پروتئین، توسط سرور SAM ساختارهای دوم پیش‌بینی شده‌اند. SAM چندین نوع (چندین الفای) ساختار دوم بر اساس تعاریف مختلف پیش‌بینی می‌کند. در این تحقیق از الفای STR4 [45] استفاده شده است. STR4 ترکیبی از چند الفای مختلف و شامل کارکترهای روبرو است:

ABCDEFGHIJKLMQQRSTUWV

6-3-5 استخراج ویژگی‌های جفت ستونی

گفته شد که برای پیش‌بینی اینکه آیا یک جفت اسید آمینه در تماس هستند یا نه، ویژگی‌های متعددی از ستون‌های متناظر MSA استخراج شده است. این ویژگی‌ها ممکن است خاص یک ستون یا جفتی از ستون‌ها باشند. اما بیشتر ویژگی‌های دارای اطلاعات²، آن‌هایی هستند که از جفت ستون‌ها استخراج می‌شوند [26].

¹ Dirichlet mixture regularizer

² Informative

تمام ویژگی‌های جفتی برای یک جفت ستون را می‌توان از جدول وابستگی¹ آن‌ها استخراج کرد. در آمار از جدول وابستگی برای ثبت و تحلیل رابطه بین دو یا چند متغیر تصادفی که معمولاً گستته هستند، استفاده می‌شود. جدول وابستگی برای دو ستون MSA، یک جدول 20 در 20 است که نشان می‌دهد هر جفت از اسیدهای آمینه x و y، چند بار در آن دو ستون همزمان اتفاق افتاده‌اند. هر چند می‌توان کارکتر فاصله را نیز در نظر گرفته و جدول‌های 21 در 21 ایجاد نمود، در این تحقیق از توالی‌هائی که در هر یک از دو ستون، کارکتر فاصله دارند، در ساختن جدول‌ها و محاسبه ویژگی‌ها صرف نظر شده است. ویژگی‌های جفتی استخراج شده بر دو نوع هستند: آن‌هائی که به دنبال نوع‌هائی از اسیدهای آمینه هستند که احتمال برقراری تماس بین آن‌ها بیشتر است (تمایل) و آن‌هائی که به دنبال جهش وابسته در دو ستون می‌گردد (MI E-Value و آنتروپی مشترک).

MI E-value

در روش Shackelford و Karplus [26] از یک ویژگی به نام MI E-value برای تعیین میزان وابستگی بین دو ستون MSA استفاده شده است. MI E-value بر اساس ویژگی اطلاعات دو جانبی (MI) تعریف شده است. در توالی‌های بیولوژیکی، MI میزان همبستگی² یا ارتباط³ میان اسیدهای آمینه مکان‌های X و Y پروتئین، که ممکن است از محدودیت‌های⁴ ساختاری، عملیاتی⁵ یا تکاملی⁶ به وجود بباید را بیان می‌کند (هر مکان در توالی یک پروتئین را می‌توان یک متغیر تصادفی در نظر گرفت که 20 مقدار می‌تواند بگیرد) [46].

برای محاسبه MI بین دو ستون MSA، در فرمول (14-3) اگر فرض شود که در ستون اول n نوع اسید آمینه مختلف و در ستون دوم m نوع اسید آمینه مختلف وجود دارد، احتمال وجود اسید آمینه نوع j در ستون اول، احتمال وجود اسید آمینه نوع k در ستون دوم و احتمال رخداد همزمان اسید آمینه j در ستون اول و اسید آمینه k در ستون دوم است. اگر هر یک از دو ستون کاملاً حفاظت شده باشند، و $MI = 0$. در محاسبه p_{jk} ، q_k و p_j تنها توالی‌هائی در نظر گرفته می‌شوند که کارکتر فاصله در آن‌ها وجود نداشته باشد [47].

¹ Contingency table

² Correlation

³ Association

⁴ Constraint

⁵ Functional

⁶ Evolutionary

برخلاف جذابیت تئوری MI، نتایج نشان می‌دهند که MI ویژگی ضعیفی برای پیش‌بینی تماس است. علت می‌تواند خلوتی¹ جدول‌های وابستگی باشد که باعث می‌شود MI های مشاهده شده، بیشتر تابع اثرات کوچکی نمونه باشند تا سیگنال‌های جهش وابسته (در آمار، هنگامی که نمونه آماری کافی نیست، به نتایج حاصل نمی‌توان اعتماد کرد. مثلاً ممکن است معیار مورد استفاده در محاسبه MSA، کاملاً دقیق و مناسب نبوده است). در نتیجه باید از روشی برای تعیین میزان معنی‌دار بودن اطلاعات MI مشاهده شده بین دو ستون (با توجه به محدود بودن نوع و تعداد اسیدهای آمینه موجود در دو ستون که باعث کوچک شدن بازه تغییرات MI می‌شود) استفاده کرد [26].

برای رفع مشکل ذکر شده، در روش Shackelford و Karplus [26] به این ترتیب عمل می‌شود که یک توزیع برای مقادیر MI، بر اساس جداول وابستگی تصادفی تخمین می‌زنند، سپس از این توزیع، برای تخمین احتمال مقدار MI مشاهده شده به طور اتفاقی (P-value) استفاده می‌کنند. نتیجه، با عنوان E-value، به صورت حاصل‌ضرب P-value در تعداد جفت‌های دو ستون مورد تست، گزارش می‌شود.

با آزمایش معلوم شده که توزیع‌های MI، از نوع توزیع گاما هستند. برای تخمین توزیع هر جفت ستون‌های MSA، به علت وجود تعداد زیاد جفت ستون‌هایی که باید ارزیابی شوند، تنها 50 جدول وابستگی تصادفی با جایگردانی² تصادفی ستون دوم به دست می‌آیند. سپس برای هر جدول مقدار MI آن محاسبه شده و مقادیر محاسبه شده به یک توزیع گاما برازش می‌شوند. در نهایت برای $4L$ (طول توالی پروتئین است) بالاترین جفت‌هایی که کمترین مقدار P-value تخمین زده شده را دارند، P-value توسط 500 جدول مجدداً محاسبه می‌شود.

در بخش (6-3-3) فرضیه تهي و P-value توضیح داده شده است. در مورد MI برای هر جفت ستون، فرضی تهي این است که مقدار MI محاسبه شده، معنی‌دار نیست. با این فرض، احتمال مشاهده MI اصلی، طبق توزیع گامای محاسبه شده، به دست می‌آید. هر چه این احتمال کوچکتر باشد، مدرک قوی‌تری علیه فرضیه تهي وجود دارد، یعنی MI مشاهده شده بامعناتر است.

در پیاده‌سازی روش پیشنهادی، محاسبه MI برای هر دو ستون، با استفاده از جدول وابستگی آن‌ها انجام شده است. تخمین پارامترهای توزیع گاما در روش Shackelford و Karplus [26] با استفاده از روش تطبیق ممان انجام شده، اما در این تحقیق، تخمین توسط احتمال حداقل (ML) صورت گرفته است.

¹ Sparseness

² Permute

آنتروپی مشترک¹

آنتروپی مشترک میزان تغییرات در دو ستون MSA را اندازه می‌گیرد و به صورت زیر تعریف می‌شود:

(6-1)

$$JE = -\sum_{(x,y)} \rho(x,y) \log \rho(x,y)$$

نسبت تعداد رخداد همزمان اسید آمینه x در ستون اول و اسید آمینه y در ستون دوم به تعداد کل

جفت‌ها در دو ستون است. محاسبه آنتروپی مشترک از روی جدول وابستگی صورت گرفته است.

احتمال در تماس بودن جفت‌های اسید آمینه در مکان‌های حافظت شده که آنتروپی مشترک پائینی دارند، بیش از جفت‌هایی است که آنteroپی مشترک آن‌ها زیاد است. همچنین، طبق قضیه‌ای در تئوری اطلاعات، آنتروپی مشترک، یک حد بالا برای MI است، زیرا همواره $\text{Antrropy مشترک} \leq MI$.

تمایل²

تمایل دو نوع اسید آمینه a و b برای برقراری تماس به صورت لگاریتم نسبت احتمال تماس بین دو اسید آمینه به احتمال رخداد مستقل آن‌ها تعریف می‌شود:

(6-2)

$$\log \frac{P(a,b)}{P(a)P(b)}$$

در روش Karplus و Shackelford [26] از تعریفی کمی متفاوت برای تمایل استفاده شده است که به تماس‌های با فاصله توالی زیاد، وزن بیشتری نسبت به تماس‌های با فاصله توالی کم نسبت می‌دهد. برای

محاسبه مقادیر تمایل، بر روی یک مجموعه داده بزرگ (2191 توالی و در حدود دو میلیون جفت)، برای

هر جفت از انواع اسید آمینه، یک شمارش وزن‌دار از تماس‌ها انجام می‌شود:

(6-3)

$$W_{a,b} = \sum_{r_i, r_j} \frac{\delta(r_i = a, r_j = b, i \text{ contacts } j)}{N}$$

و a و b انواع اسیدهای آمینه و r_i و r_j اسیدهای آمینه‌های مکان‌های i و j هستند. تماس‌های محلی 8 کمتر از $|i-j|$ در نظر گرفته نمی‌شوند (توضیح فرمول (6-3): تمام جفت مکان‌های i و j بررسی می‌شوند).

¹ Joint entropy

² Propensity

اگر در مکان a و در مکان b اسید i بوده و با هم در تماس باشند، یک مقدار به صورت وزن دار به تعداد تماس ها اضافه می شود. وزن، عکس تعداد تماس ها با فاصله توالی $|j-i|$ است). سپس تعداد تماس های ممکن شمارش می گردد:

(6-4)

$$N_{a,b} = \sum \delta(r_i = a, r_j = b)$$

تمایل، لگاریتم نسبت این دو مقدار است:

(6-5)

$$\text{Pr } op_{a,b} = \log(W_{a,b}/N_{a,b})$$

ویژگی تمایل برای دو ستون، با میانگین گیری از میزان تمایل تمام جفت های دو ستون به دست می آید. در پیاده سازی روش پیشنهادی، از مقادیر تمایل محاسبه شده توسط Karplus و Shackelford استفاده شده است. جدول مقادیر، در پیوست (الف) ارائه شده است.

تعداد جفت ها

این ویژگی برای یک جفت اسید آمینه، تعداد جفت های دو ستون MSA را نشان می دهد. هر چه تعداد جفت ها بیشتر باشد، اطلاعات جهش وابسته قابل اعتمادتر خواهد بود.

6-4 آماده سازی داده های آموزش و تست

پس از استخراج ویژگی های مورد نظر، این ویژگی ها به طور مناسب کنار یکدیگر قرار گرفته تا بردار ورودی شبکه عصبی ایجاد شود. سپس نمونه هایی که دارای اطلاعات نیستند حذف شده و در نهایت نسبت نمونه های مثبت و منفی برای داده های آموزشی متوازن شده است. در ادامه جزئیات این مراحل شرح داده شده اند.

6-4-1 تولید داده های آموزش و تست

در این تحقیق تماس بین دو اسید آمینه، بر اساس فاصله بین اتم های C_β آن دو تعریف شده است. اگر این فاصله کوچکتر یا مساوی 8 آنگستروم باشد، جفت اسید آمینه، در تماس با یکدیگر در نظر گرفته شده اند. مقدار آستانه برای تعریف تماس های محلی، 8 در نظر گرفته شده است. به عبارت دیگر، در آموزش و

تست تنها جفت‌هایی از هر پروتئین در نظر گرفته شده‌اند که حداقل به اندازه 8 اسید آمینه در توالی از یکدیگر فاصله داشته باشند [26].

مشابه روش Karplus و Shackelford [26]، در تولید بردار ورودی، برای اسیدهای آمینه مکان‌های i و j ، ویژگی‌های جفت ستونی، تنها خود جفت (i,j) در نظر گرفته شده‌اند، در حالی که برای ویژگی‌های تک ستونی i و j (به غیر از آنتروپی)، از پنجره‌ای به عرض 5 استفاده شده است. یعنی، مقادیر ویژگی‌های تک ستونی مکان‌های $i \pm 2$ و $j \pm 2$ در بردار ورودی قرار گرفته‌اند. بدین ترتیب تاثیر همسایگی‌های جفت اسید آمینه در نظر گرفته می‌شوند.

در مورد ویژگی‌های جفت ستونی، به جای قرار دادن مقادیر خام این ویژگی‌ها در بردار ورودی، رتبه مقدار در لیست مقادیر تمام جفت ستون‌ها قرار داده شده است [26] (با توجه به مقادیر ویژگی‌های جفت ستونی، تتنوع رتبه‌ها بیش از خود مقادیر است و چون حساسیت تابع فعالیت سیگموید زیاد نیست و هدف، نشان دادن اختلاف به شبکه عصبی است، بهتر است از رتبه‌ها استفاده شود. البته این روش در Hallati مناسب است که همه مقادیر به یکدیگر نزدیک باشند. اگر بعضی مقادیر فاصله زیادی با بقیه داشته باشند، با در نظر گرفتن رتبه‌ها، آن‌ها نیز به بقیه نزدیک شده و اختلاف فاصله، از بین می‌رود).

بدین ترتیب، بردار ورودی شبکه عصبی برای دو اسید آمینه در مکان‌های i و j از مقادیر زیر

تشکیل می‌شود:

- لگاریتم طول توالی
- لگاریتم فاصله توالی
- توزیع اسید آمینه $i \pm 2$ (هر مکان 20 مقدار)
- توزیع اسید آمینه $j \pm 2$ (هر مکان 20 مقدار)
- ساختار دوم پیش‌بینی شده $i \pm 2$ (هر مکان 21 مقدار)
- ساختار دوم پیش‌بینی شده $j \pm 2$ (هر مکان 21 مقدار)
- آنتروپی i
- آنتروپی j
- تعداد جفت‌های (i,j)
- مقدار (i,j) MI E-value
- لگاریتم رتبه (i,j) MI E-value

- لگاریتم رتبه تمایل (j,i)
 - لگاریتم رتبه آنتروپی مشترک (i,j)
 - کلاس (j,i) (خروجی مطلوب)
- اندازه بردار ورودی 419 است.

6-4-2 استخراج نمونه‌های دارای اطلاعات

برای کاهش هزینه محاسباتی و حافظه در شبکه‌های عصبی در روش Shackelford و Karplus [26]، ویژگی‌ها برای تمامی جفت‌های یک پروتئین استخراج می‌شوند. سپس نمونه‌ها به طور جداگانه بر اساس هر یک از ویژگی‌های جفت ستونی مرتب‌سازی شده و نمونه‌هایی که جز بالاترین $4L$ (طول پروتئین است) جفت حداقل یکی از ویژگی‌ها هستند، برای آموزش و تست سیستم مورد استفاده قرار می‌گیرند. مجموعه حاصل همواره بزرگتر از $4L$ است و نمونه مناسبی برای آموزش و تست فراهم می‌نماید. استفاده از این مجموعه غنی شده، مشکل داده‌های آموزشی نامتوازن را کاهش‌می‌دهد، مشروط بر این‌که از پروسه غنی‌سازی یکسانی برای آموزش و تست استفاده شود.

پروسه مذکور را به این صورت می‌توان توجیه نمود که در داده‌ها، تعداد زیادی از نمونه‌های رده منفی اثربخشی روی مرز جداگانه دو رده ندارند، زیرا نسبت به مرز جداگانه بسیار دور واقع شده‌اند. در واقع این نمونه‌ها دارای اطلاعاتی نیستند. در حالت ایده‌آل، حذف این نمونه‌ها از داده‌های آموزشی نه تنها مشکلی برای آموزش شبکه‌های عصبی ایجاد نمی‌کند، بلکه باعث کاهش هزینه محاسباتی می‌گردد (البته در عمل، امکان حذف مقداری از نمونه‌های لب مرز هم وجود دارد). با مرتب کردن داده‌ها بر اساس چند ویژگی و انتخاب $4L$ بالاترین، این نمونه‌های بی‌اثر حذف می‌شوند.

ویژگی‌هایی که جفت‌ها بر اساس آن‌ها مرتب می‌شوند، بایستی جز ویژگی‌های جفت ستونی باشند. نحوه مرتب‌سازی (صعودی یا نزولی) بستگی به نوع ویژگی دارد. در این تحقیق ویژگی‌های زیر برای مرتب‌سازی انتخاب و 4 مجموعه ایجاد شده‌اند:

- تعداد جفت‌ها: مرتب‌سازی به صورت نزولی
 - لگاریتم رتبه MI E-value: مرتب‌سازی به صورت صعودی
 - لگاریتم رتبه تمایل: مرتب‌سازی به صورت نزولی
 - لگاریتم رتبه آنتروپی مشترک: مرتب‌سازی به صورت صعودی
- سپس $4L$ بالاترین جفت‌ها از هر مجموعه انتخاب و اجتماع آن‌ها محاسبه گردیده است.

6-4-3 متوزن کردن داده‌های آموزشی

در یک پروتئین، تعداد جفت‌های بدون تماس (رده منفی) بسیار بیشتر از جفت‌های تماس‌دار (رده مثبت) است. نسبت اندازه این دو رده به طور تقریبی بیش از 20 به 1 است [36]. بنابراین، آموزش شبکه عصبی با تمام جفت‌ها، شبکه را به سمت رده منفی بایاس می‌کند و باعث می‌شود شبکه تعداد بسیار کمی از رده مثبت را پیش‌بینی نماید. در واقع می‌توان گفت اختلاف زیاد مقادیر نمونه‌های دو رده، مشکل اثر فراموشی را برای شبکه در پی دارد. برای رفع این مشکل در این تحقیق، بعد از استخراج جفت‌های دارای اطلاعات، با انتخاب تصادفی زیرمجموعه‌ای از نمونه‌ها، نسبت نمونه‌های بدون تماس به نمونه‌های تماس‌دار، 4 به 1 نگه داشته می‌شود [36].

مشکل متوزن کردن داده‌های آموزشی این است که شبکه در طول آموزش مقادیر زیادی از نمونه‌های منفی را نمی‌بیند و هنگام تست سیستم، نمی‌تواند پاسخ صحیح به این دسته از الگوهای مشاهده نشده بدهد. برای برطرف کردن این مشکل، در روش Shackelford و Karplus [26] در هر epoch آموزش شبکه، 5% از نمونه‌های منفی به طور تصادفی تعویض می‌شوند. به این ترتیب به احتمال زیاد شبکه در طول آموزش اکثر نمونه‌های منفی را می‌بیند و در عین حال توازن کافی برای ممانعت از بایاس شدن شبکه به سمت رده منفی فراهم می‌گردد. تعویض تصادفی نمونه‌های منفی در طول آموزش، در این تحقیق انجام نشده است. توجه شود که متوزن کردن نسبت دو رده در مورد داده‌های تست نمی‌تواند انجام شود، زیرا در این صورت آمار خطای سیستم، واقعی نخواهد بود.

6-5 نحوه ارزیابی کارائی سیستم

معیار مهم در ارزیابی پیش‌بینی نقشه تماس، نسبت تماس‌های پیش‌بینی شده صحیح به تعداد کل پیش‌بینی‌ها، یعنی $\frac{TP}{TP+FP}$ (محور Y)، در مقابل تعداد کل پیش‌بینی‌ها، یعنی $\frac{TP}{TP+FP}$ (محور X) است (محور X نمایش‌دهنده تعداد تماس‌های پیش‌بینی شده صحیح و FP نشان‌گر تعداد تماس‌های پیش‌بینی شده غیرصحیح است). برای هر پروتئین تست، نمودار تغییرات کمیت اول نسبت به کمیت دوم ایجاد می‌گردد. محور X به صورت (طول پروتئین/ $\log(\frac{TP}{TP+FP})$) در نظر گرفته می‌شود [26] (لگاریتم، باعث کوچک شدن مقادیر X و در نتیجه نزدیک شدن مقادیر دو محور X و Y می‌گردد).

به دلیل این‌که خروجی هر شبکه و در نتیجه پاسخ گروه (میانگین پاسخ‌ها) پیوسته است، برای تعیین رده هر نمونه مورد آزمون، دو رویکرد قابل استفاده است. یک روش، آستانه‌گذاری خروجی است که چندان صحیح به نظر نمی‌رسد. روشی که در این تحقیق استفاده شده است، معیار کارائی را بر حسب

آستانه‌های مختلف ارزیابی می‌کند. به این صورت که خروجی مدل به طور نزولی مرتب می‌شود، هر بار آستانه در نظر گرفتن خروجی مدل به عنوان یک تماس، مقدار بعدی خروجی در نظر گرفته شده است. حال با داشتن آستانه، (طول پروتئین/آستانه) Log مقدار بعدی بر روی محور X است و مقدار متناظر بر روی محور Y، با شمارش تعداد تماس‌های پیش‌بینی شده صحیح و تقسیم بر آستانه به دست می‌آید.

میانگین نمودار‌های کارائی برای تمامی پروتئین‌های تست، به عنوان نمودار کارائی سیستم گزارش می‌شود. در غیر این صورت، این امکان وجود دارد که نتایج خوب یا بد غیرواقعی به دست آیند، زیرا ممکن است برای برخی از پروتئین‌های تست، به طور استثنای نتایج بسیار خوب و یا بسیار بد حاصل شوند و نتوان بهدرستی، میزان کارائی مدل را ارزیابی نمود.

کاربردهای مختلف، نیازمند تعداد پیش‌بینی‌های صحیح متفاوت بوده و تحمل خطای متفاوتی دارند. برای مثال، در کاربردی ممکن است 1 یا 2 پیش‌بینی صحیح، بتواند دو مدل را از یکدیگر متمایز نماید. بنابراین، کل نمودار برای تفسیر نتایج لازم است.

6-6 آموزش و تست مدل

در این بخش ابتدا مراحل ساخت مدل پیش‌بینی کننده و سپس نحوه تست آن شرح داده شده است.

5-6-1 آموزش مدل

روش پیشنهادی، بر اساس تکنیک ماشین گروهی است. ساختار ماشین پیشنهادی، از نوع ساختارهای ایستا است (بخش (5-3-3)) و در آن از گروهی از شبکه‌های عصبی برای رده‌بندی جفت‌های اسید آمینه به عنوان تماس‌دار یا بدون تماس استفاده می‌شود. مدل در دو مرحله ایجاد می‌گردد. در هر مرحله یک گروه ساخته می‌شود. در مرحله اول گروه 1 ایجاد می‌شود. ایجاد گروه به صورت سریال صورت می‌گیرد، بدین معنی که ابتدا یک شبکه عصبی، به نحوی که در ادامه شرح داده شده است، آموزش دیده و به عنوان اولین عضو گروه قرار می‌گیرد. سپس شبکه‌های دیگر در چند تکرار به گروه اضافه می‌شوند. در هر تکرار یک شبکه جدید ایجاد شده و آموزش داده می‌شود. این شبکه به شرطی به گروه اضافه می‌شود که بتواند میانگین کارائی گروه را افزایش دهد. مفهوم افزایش کارائی در ادامه توضیح داده شده است. این پروسه ادامه می‌یابد تا جایی که برای یک تعداد تکرار از پیش تعیین شده، شبکه‌های عصبی جدید آموزش دیده، نتوانند کارائی گروه را افزایش داده و به گروه اضافه شوند. حداقل تعداد تکرار برای مرحله اول 5 و برای مرحله دوم 20 در نظر گرفته شده است.

همان‌طور که در بالا شرح داده شد، آخرین شبکه‌ای که به گروه₁ اضافه می‌شود، بایستی کارائی مدل را بهبود داده و بنابراین باید بهتر از میانگین گروه₁ باشد (کارائی آخرین شبکه را نمی‌توان با یکایک شبکه‌های قبلی مقایسه کرد، زیرا ممکن است عضوهایی با کارائی شبکه آخر و یا حتی بهتر از آن قبلا در گروه وجود داشته‌اند، اما رای‌های صحیح تا قبل از اضافه شدن شبکه آخر کافی نبوده است). در مرحله دوم، با در نظر گرفتن آخرین عضو گروه₁ به عنوان اولین عضو گروه اصلی که گروه₂ نامیده می‌شود، از اضافه شدن شبکه‌های نامناسب جلوگیری می‌شود. در واقع ایجاد گروه₂ با یک عضو خوب شروع می‌شود که تا حدودی تضمین می‌کند که اعضای بعدی گروه نیز که در آینده اضافه می‌شوند، کارائی خوبی داشته باشند. اگر اولین عضو گروه خوب نباشد، شبکه‌های بعدی که کمی بهتر هستند، به گروه اضافه می‌شوند و این شبکه‌های نه‌چندان خوب، روی کارائی کل گروه اثر منفی می‌گذارند. شرط خاتمه پروسه ایجاد گروه₂، مشابه گروه₁ است.

ساختار شبکه‌های عصبی مورد استفاده بر اساس روش Skachelford و Karplus [26] طراحی شده است. شبکه‌ها از نوع feed-forward هستند. هر شبکه یک لایه پنهان دارد. تعداد نورون‌های لایه ورودی 419، لایه پنهان 45 و لایه خروجی 1 است. الگوریتم آموزش، بازگشت به عقب انعطاف‌پذیر، حداقل تعداد epoch های آموزش، 500 و تابع کارائی شبکه‌ها، MSE است.

برای آموزش یک شبکه جدید، K پروتئین به طور تصادفی از مجموعه داده‌های آموزشی انتخاب می‌شوند (تقسیم فضای ورودی). برای انتخاب K باید توجه داشت که هنگام آموزش هر شبکه، هر چه میزان داده‌ها بیشتر باشد، احتمال وجود الگوهای بیشتری در داده‌ها و احتمال دیده شدن هر الگو توسط تعداد بیشتری از اعضا وجود دارد. هنگام تست مدل، برای این‌که گروه پاسخ صحیح به یک ورودی بدهد، حداقل نیمی از اعضا باید در طی آموزش الگوی مورد سوال را دیده باشند و بتوانند پاسخ صحیح بدene. پس از یک طرف آموزش هر شبکه با تعداد داده زیاد مناسب‌تر است، اما از طرف دیگر، چون ظرفیت یادگیری هر شبکه محدود است، با زیاد بودن میزان داده‌ها، ممکن است شبکه نتواند همه الگوهای موجود در آن‌ها را فرابگیرد. در روش پیشنهادی، چون شبکه‌ها تا حد دلخواه می‌توانند به گروه اضافه شوند، لزومی به آموزش هر شبکه با میزان زیاد داده، نیست و K به 10 مقداردهی شده است.

برای اجتناب از مشکل اثر فراموشی، آموزش شبکه‌ها با K پروتئین به طور سریال انجام نمی‌شود. هر شبکه K بار آموزش داده می‌شود. در هر بار، پروتئین i ام ($K \leq i \leq 1$) به طور تصادفی انتخاب می‌گردد. سپس احتمال انتخاب داده i ام برای دفعات بعدی کاهش و احتمال انتخاب دیگر داده‌ها افزایش

می‌یابد. این کار به وسیله استفاده از یک تابع توزیع احتمال متغیر روی K داده انجام می‌شود. بنابراین، هر داده آموزشی می‌تواند در هر بار آموزش، با احتمال متغیر انتخاب شود و به این ترتیب از اثر فراموشی تا حد زیادی اجتناب می‌گردد.

در آموزش هر شبکه با یک پروتئین، از 90% آن برای آموزش (زیرمجموعه اول) و از 10% بقیه برای جلوگیری از بیشپوشش، توسط روش توقف زودهنگام (زیرمجموعه دوم)، استفاده می‌شود. 25% از زیرمجموعه دوم، از ابتدای پروتئین و 75% آن از انتهای پروتئین استخراج می‌شود. به این ترتیب در زیرمجموعه دوم، اکثریت جفت‌ها، فاصله توالی زیاد دارند.

پس از آموزش هر شبکه جدید، دو نمودار کارائی محاسبه می‌شوند. در بخش (5-6) گفته شد که نمودار کارائی برای یک پروتئین تست، از میانگین‌گیری پاسخ اعضاً یک گروه به آن پروتئین به دست می‌آید. نمودار کل، خود میانگین نمودارهای تمام پروتئین‌های تست است. گروهی که برای محاسبه نمودار اول استفاده می‌شود، گروه جاری، یعنی تمام شبکه‌های اضافه شده تا به این لحظه است. نمودار دوم، از گروه جاری به اضافه شبکه جدید که هنوز به گروه نپیوسته است، به دست می‌آید.

سپس دو نمودار کارائی مقایسه می‌شوند تا مشخص شود که آیا شبکه جدید کارائی گروه را بهبود می‌دهد یا خیر. برای مقایسه دو نمودار، نمودار تفاضل محاسبه می‌شود. نمودار تفاضل، اختلاف دو نمودار کارائیست. از مجموع وزن‌دار¹ مقادیر نمودار تفاضل به عنوان معیار بهبود کارائی استفاده می‌گردد. اگر مجموع وزن‌دار، مثبت باشد، شبکه جدید به گروه اضافه می‌شود، در غیر این صورت از آن صرف نظر شده و شبکه بعدی ایجاد و بررسی می‌گردد. تابع وزن پیشنهادی به صورت زیر است:

$$Weight(d) = \frac{\exp\left\{ \frac{-(d - C)^2}{2\sigma^2} \right\}}{6-6}$$

که

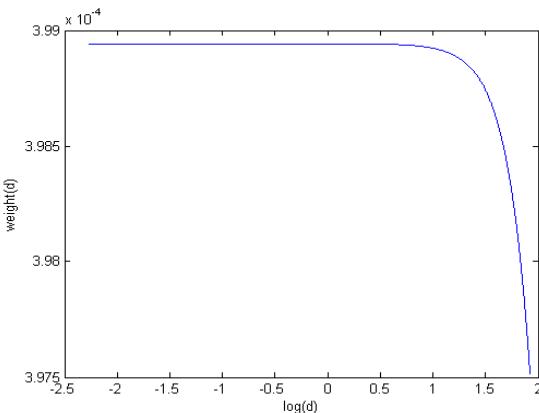
(7-6)

$$d = \frac{TP + FP}{TP + FN}$$

در فرمول (6-6)، $C = L_p/2$ و $\sigma = 1000$ در نظر گرفته می‌شود که L_p میانگین طول پروتئین‌هاست.

¹ Weighted sum

تابع وزن پیشنهادی در شکل (1-6) نمایش داده شده است. شکل نشان می‌دهد که تمرکز بر آستانه نزدیک به نقطه C که در کارهای قبلی به آن توجه شده است ([24] و [26]), می‌باشد. برای آستانه‌های کوچکتر که امکان داشتن تماس‌های پیش‌بینی شده بیشتری را می‌دهد، از وزن‌های کوچکتر استفاده می‌شود. همان‌طور که در شکل مشخص است، محور افقی لگاریتمی است.



شکل 1-6) تابع وزن پیشنهادی

توجه به این نکته لازم است که در مدل پیشنهادی، مجموعه داده‌های آموزشی خود به دو زیرمجموعه آموزش و تست تقسیم می‌شود. از زیرمجموعه آموزش جهت آموزش هر شبکه عصبی جدید استفاده می‌گردد. از زیرمجموعه تست برای بررسی این‌که آیا شبکه جدید به گروه اضافه شود یا خیر، استفاده می‌شود. در واقع کارائی گروه، بدون در نظر گرفتن و با در نظر گرفتن شبکه جدید، روی تمام داده‌های زیرمجموعه تست، ارزیابی می‌شود. بنابراین منظور از ((داده‌های تست)) در این بخش، زیرمجموعه تست مجموعه داده‌های آموزشی است. به دلیل این‌که این زیرمجموعه در جریان آموزش مدل نقش دارند، از آن‌ها در تست نهایی مدل استفاده نمی‌شود. تست نهایی مدل (بخش 6-2) بر روی مجموعه داده‌های تست که 20% کل مجموعه داده‌ها را تشکیل می‌دهند، انجام می‌گردد.

عمل استخراج نمونه‌های دارای اطلاعات (بخش 4-2) بر روی مجموعه داده‌های آموزشی و تست صورت می‌گیرد. به علاوه، بر اساس آنچه در بخش 4-3) شرح داده شد، تنها زیرمجموعه آموزش مجموعه داده‌های آموزشی، متوازن می‌گردد.

دلیل استفاده از ماشین گروهی در روش پیشنهادی، وجود حجم بالای مقادیر داده و تعداد ویژگی‌های زیاد است. با استفاده از یک شبکه، هر چند پیچیده، احتمال یادگیری تمامی الگوهای ورودی بسیار کم است. به علاوه، حتی اگر یک شبکه بتواند تمام الگوها را یاد بگیرد، محاسبات آن هزینه بسیار بالائی

خواهد داشت و حتی می‌تواند غیرممکن باشد. با استفاده از چند شبکه که هر یک روی بخشی از الگوها آموزش می‌بینند و احتمالاً با یکدیگر همپوشانی¹ دارند، پوشش بهتری بر روی فضای ورودی صورت می‌گیرد و به علاوه با میانگین‌گیری از پاسخ‌ها، خطا کاهش پیدا می‌کند.

6-6-2 تست مدل

برای تست نهائی مدل، کارائی گروه اصلی ساخته شده در مرحله آموزش (گروه 2)، بر روی هر پروتئین مجموعه تست محاسبه و میانگین کارائی‌های تمام پروتئین‌های تست، به عنوان کارائی مدل گزارش می‌شود. نحوه محاسبه کارائی در بخش (5-6) توضیح داده شد. همان‌طور که گفته شد، پروسه استخراج جفت‌های دارای اطلاعات بر روی مجموعه داده‌های تست اعمال می‌شود، اما این داده‌ها متوازن نمی‌گردند تا دقیق به دست آمده صحیح باشد.

6-7 تحلیل نتایج

به منظور ارزیابی مدل پیشنهادی، دو مدل دیگر پیش‌بینی نقشه تماس پیاده سازی شده‌اند. در این بخش نتایج سه روش زیر ارائه و با یکدیگر مقایسه شده‌اند:

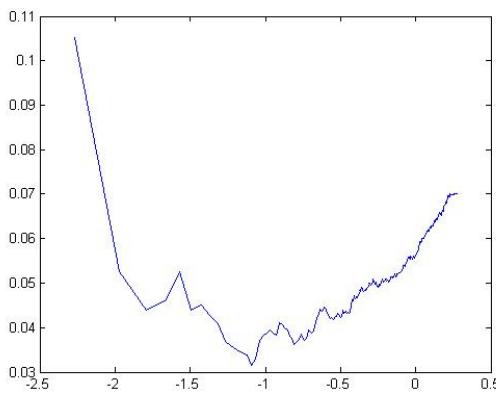
- روش پایه
- روش مبتنی بر میانگین‌گیری گروه
- روش پیشنهادی

توجه به این نکته لازم است که چون وزن‌دهی اولیه هر شبکه به طور تصادفی انجام می‌شود، در هر سه روش، در هر بار اجرا، نتیجه تا حدی متفاوت خواهد بود. می‌توان میانگین چندین اجرا را به عنوان نتیجه هر روش در نظر گرفت.

6-7-1 نتایج روش پایه

در این مدل، بر اساس روش Karplus و Shackelford [26]، یک شبکه عصبی به وسیله تمام داده‌های آموزشی به طور سریال، آموزش داده شده است. ساختار، ورودی‌ها و الگوریتم آموزش شبکه، مانند شبکه‌های روش پیشنهادی هستند. نمودار کارائی مدل که در شکل (2-6) مشاهده می‌شود، بر اساس تعریف بخش (5-6) محاسبه شده است.

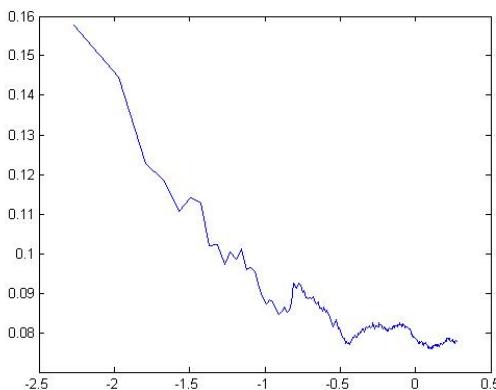
¹ Overlap



شکل 6-2) نمودار کارائی روش پایه

6-7-2 نتایج روش مبتنی بر میانگین‌گیری گروه

تکنیک میانگین‌گیری گروه در بخش (5-3-3) توضیح داده شده است. برای پیاده‌سازی روش، 10 شبکه ایجاد و هر یک با تمامی داده‌های آموزشی به طور سریال، آموزش داده شده‌اند. ساختار شبکه‌ها، الگوریتم آموزش و ورودی‌های شبکه، مانند شبکه‌های روش پیشنهادی هستند. پاسخ مدل به یک پروتئین ورودی، میانگین پاسخ شبکه‌ها در نظر گرفته شده است. چون وزن‌دهی اولیه شبکه‌ها به طور تصادفی صورت می‌گیرد، امکان همگرا شدن شبکه‌ها به نقاط بهینه محلی متفاوت وجود دارد و انتظار می‌رود پاسخ گروه، بهتر از پاسخ هر از شبکه‌ها باشد. نمودار کارائی مدل در شکل (6-3)، بهبود پاسخ را نسبت به روش پایه نشان میدهد.

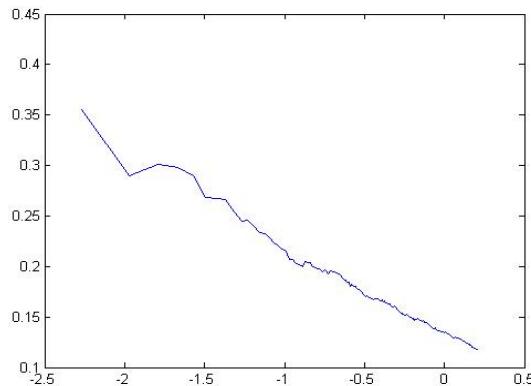


شکل 6-3) نمودار کارائی روش میانگین‌گیری گروه

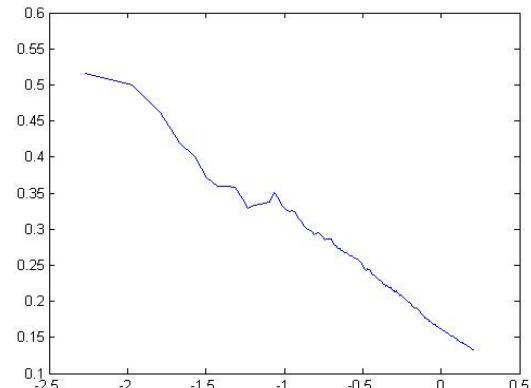
6-7-3 نتایج روش پیشنهادی

در این بخش، نتایج روش پیشنهادی برای مدلی با 105 عضو بررسی شده‌اند. نمودارهای (6-4) و (6-5) کارائی آخرین اعضایی گروه‌های 1 و 2 را بر روی زیرمجموعه تست داده‌های آموزشی نشان میدهند.

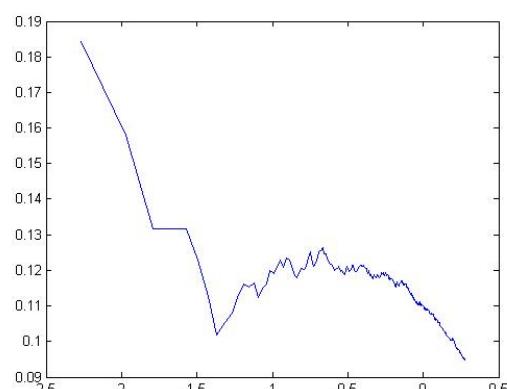
تغییر در کارائی مدل با افزایش اعضای گروه در نمودارهای (6-6) تا (6-11) نمایش داده شده است. با مقایسه کارائی کل گروه (نمودار (11-6)) با کارائی دو روش ذکر شده قبلی (نمودارهای (2-6) و (6-6)، می‌توان به کارائی تکنیک یادگیری گروهی و روش پیشنهادی در زمینه پیش‌بینی نقشه تماس پی برد.



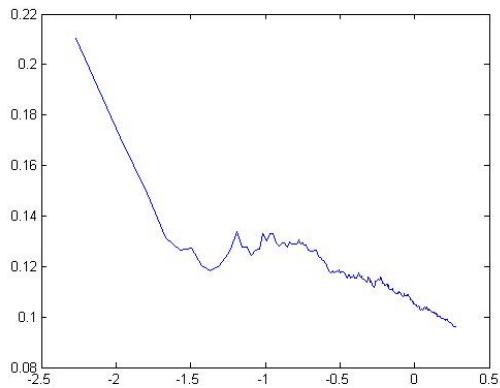
شکل 6-4) نمودار کارائی گروه 1 بر روی زیرمجموعه تست داده‌های آموزشی



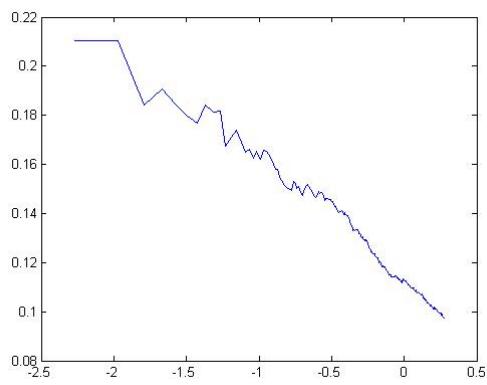
شکل 6-5) نمودار کارائی گروه 2 بر روی زیرمجموعه تست داده‌های آموزشی



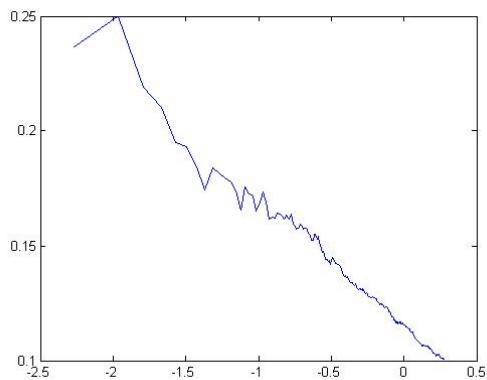
شکل 6-6) نمودار کارائی گروه 2 با 5 عضو



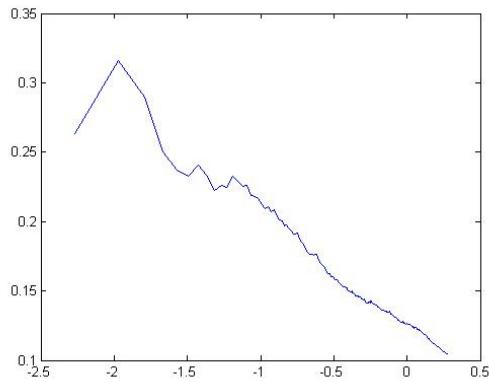
شکل 6-7) نمودار کارائی گروه 2 با 20 عضو



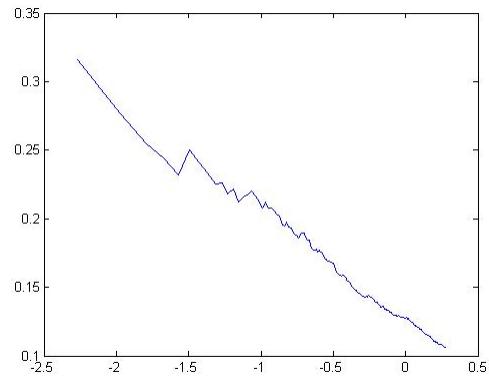
شکل 6-8) نمودار کارائی گروه 2 با 30 عضو



شکل 6-9) نمودار کارائی گروه 2 با 50 عضو



شکل 6-10) نمودار کارائی گروه 2 با 90 عضو



شکل 6-11) نمودار کارائی مدل

فصل هفتم

نتیجه‌گیری و پیشنهادها

1-7 نتیجه‌گیری

نقشه تماس پروتئین، یک نمایش ساده از ساختار سه بعدی پروتئین است و نشان می‌دهد که هر جفت اسید آمینه در فضای یک مقدار آستانه تعریف شده، به یکدیگر نزدیکتر هستند یا خیر. پیش‌بینی نقشه تماس پروتئین از روی توالی آن یکی از مباحث مهم در بیوانفورماتیک است، زیرا با داشتن نقشه تماس، پیش‌بینی ساختار سوم پروتئین امکان‌پذیر می‌گردد که از اهمیت بسیار برخوردار است. پیش‌بینی ساختار سوم، مساله پیش‌بینی نقشه تماس را می‌توان به صورت یک مساله رده‌بندی در نظر گرفت که هر جفت اسید آمینه، در یکی از دو رده ((تماس‌دار)) و ((بدون تماس)) قرار می‌گیرند.

در این تحقیق، نقشه تماس پروتئین بر اساس تکنیک ماشین گروهی پیش‌بینی شده است. یادگیری گروهی یک روش یادگیری ماشین است که در آن وظیفه یادگیری، میان چند یادگیر و فضای ورودی به چند زیرفضا تقسیم می‌شود. پاسخ یادگیرها به یک ورودی، با یکدیگر ترکیب شده و پاسخ نهائی سیستم را تشکیل می‌دهند. این پاسخ، دقیق‌تر از پاسخ هر یک از یادگیرهاست. کارائی روش ماشین گروهی، در

بسیاری از زمینه‌ها و کاربردها به اثبات رسیده است. هدف این تحقیق، بررسی کارائی آن در مساله پیش‌بینی نقشه تماس است.

گروه یادگیر در این تحقیق، مجموعه‌ای از شبکه‌های عصبی با ساختار یکسان می‌باشد. شبکه‌ها از نوع feed-forward هستند. هر شبکه یک لایه پنهان دارد. تعداد نورون‌های لایه ورودی 419، لایه پنهان 45 و لایه خروجی 1 است. الگوریتم آموزش بازگشت به عقب انعطاف‌پذیر است. حداقل تعداد epoch های آموزش، 500 در نظر گرفته شده است.تابع کارائی شبکه‌ها، MSE است.

مجموعه‌ای از ویژگی‌ها برای هر جفت اسید آمینه پروتئین محاسبه و به عنوان ورودی به شبکه‌های عصبی داده می‌شوند. این ویژگی‌ها عمدتاً بر اساس MSA هر پروتئین استخراج می‌گردند. MSA پروتئین‌ها از وب سرویس SAM به دست می‌آیند. ویژگی‌ها بر سه نوع هستند. نوع اول، ویژگی‌های توالی است که شامل طول توالی پروتئین و فاصله توالی جفت اسید آمینه می‌باشد. نوع دوم، ویژگی‌های تک ستونی است که شامل ساختار دوم پیش‌بینی شده هر مکان، توزیع اسیدهای آمینه (احتمال رخداد هر اسید آمینه برای هر مکان) و آنتروپی (میزان حفاظت هر مکان) است. نوع سوم، ویژگی‌های جفت ستونی هستند که بر اساس جفت مکان‌های پروتئین استخراج می‌شوند. یکی از ویژگی‌های این گروه، تمایل است. تمایل دو اسید آمینه، بیان‌گر احتمال تماس داشتن آن دو است. ویژگی تمایل برای دو ستون، با میانگین‌گیری از میزان تمایل تمام جفت‌های دو ستون به دست می‌آید. ویژگی دیگر آنتروپی مشترک دو مکان است که میزان تغییرات دو ستون را نشان می‌دهد. احتمال در تماس بودن جفت‌های اسید آمینه در مکان‌های حفاظت شده که آنتروپی مشترک پائینی دارند، بیش از جفت‌هایی است که آنتروپی مشترک آن‌ها زیاد است. مهمترین ویژگی این گروه MI E-value است که از آن برای تعیین میزان وابستگی دو ستون MSA استفاده شده است. هر چه میزان وابستگی بیشتر باشد، احتمال در تماس بودن نیز بیشتر است. معیار مهم برای ارزیابی پیش‌بینی نقشه تماس، نسبت تعداد تماس‌های پیش‌بینی‌های صحیح به کل پیش‌بینی‌هاست. به دلیل این‌که خروجی هر شبکه و در نتیجه پاسخ گروه پیوسته است، برای تعیین کلاس هر جفت مورد آزمون، دو رویکرد قابل استفاده است. یک روش، آستانه‌گذاری خروجی است که چنان صلح نیست. روشنی که در این تحقیق استفاده شده است، معیار کارائی را بر حسب آستانه‌های مختلف ارزیابی می‌کند. به این صورت که خروجی مدل به طور نزولی مرتب می‌شود. هر بار آستانه، مقدار بعدی در نظر گرفته شده و برای آن آستانه نسبت تعداد تماس‌های پیش‌بینی شده درست به تعداد کل تماس‌ها محاسبه می‌گردد.

ایجاد گروه یادگیر، به صورت سریال صورت می‌گیرد. بدین معنی که ابتدا یک شبکه عصبی آموزش دیده و به عنوان اولین عضو گروه قرار می‌گیرد. سپس شبکه‌های دیگر در چند تکرار به گروه اضافه می‌شوند. در هر تکرار یک شبکه جدید ایجاد شده و آموزش داده می‌شود. این شبکه به شرطی به گروه اضافه می‌شود که بتواند میانگین کارائی گروه را افزایش دهد. این پروسه ادامه می‌یابد تا جایی که برای یک تعداد تکرار از پیش تعیین شده، شبکه‌های عصبی جدید آموزش دیده، نتوانند کارائی گروه را افزایش داده و به گروه اضافه شوند. برای آموزش یک شبکه جدید، K پروتئین به طور تصادفی از مجموعه داده‌های آموزشی انتخاب می‌شوند. برای اجتناب از مشکل اثر فراموشی، هر شبکه K بار به طور تصادفی، با یکی از K داده آموزش داده می‌شود. K به 10 مقداردهی شده است.

برای ارزیابی روش پیشنهادی، دقت سه روش پایه، میانگین‌گیری گروه و مدل پیشنهادی بر روی یک مجموعه تست یکسان محاسبه شده است. روش پایه تنها از یک شبکه عصبی که با تمام داده‌های آموزشی، آموزش داده شده، برای ردبندی جفت‌های اسید آمینه استفاده می‌کند. دقت این روش به طور میانگین در اجرای مختلف در حدود 0.05 است. در روش میانگین‌گیری گروه، 10 شبکه هر یک با تمام داده‌های آموزشی، آموزش داده شده و میانگین پاسخ گروه به هر داده تست، به عنوان پاسخ نهائی در نظر گرفته شده است. چون وزن‌دهی اولیه شبکه‌ها به طور تصادفی صورت می‌گیرد، امکان همگرا شدن شبکه‌ها به نقاط بینه محلی متفاوت وجود دارد و انتظار می‌رود پاسخ گروه، بهتر از پاسخ هر از شبکه‌ها باشد. دقت این روش بهتر از روش پایه و در حدود 0.1 است. دقت روش پیشنهادی بهتر از دو روش قبلی و به طور میانگین در اجرای مختلف 0.2 است. در طول آموزش، دقت مدل به تدریج با افزایش تعداد اعضای گروه بهبود می‌یابد. نتایج، نشان‌دهنده کارائی روش یادگیری گروهی برای پیش‌بینی نقشه تماس است.

7-2 پیشنهادها

مدل پیشنهادی می‌تواند با استفاده از ترکیب متفاوتی از ویژگی‌های ورودی برای شبکه‌های عصبی و با آموزش بر روی یک مجموعه داده بزرگتر بهبود یابد. در زیر پیشنهادهایی برای ادامه کار ارائه شده است:

- پیش‌پردازش متفاوت فایل‌های PDB با مشخصات ذکر شده در بخش (6-1) به جای حذف

آن‌ها از مجموعه داده‌ها

- لاغر کردن فایل‌های MSA با دو آستانه شباht متفاوت برای استخراج ویژگی‌های تک ستونی و جفت ستونی (بخش (2-3-6))
- محاسبه ویژگی توزیع اسیدهای آمینه به روش تنظیم کننده ترکیبی دریکله [44]، به جای محاسبه تقریبی احتمال رخداد آن‌ها (بخش (4-3-6)-توزیع اسیدهای آمینه)
- استفاده از الفباهای روش‌های دیگر پیش‌بینی ساختار دوم (بخش (4-3-6)-ساختار دوم پیش‌بینی شده)
- استفاده از آندیس‌های AAindex3 که در بخش (7-2-3) توضیح داده شده‌اند، به جای ویژگی تمایل (بخش (6-3-6)-تمایل)
- استفاده از انواع دیگر ماشین گروهی، مانند روش‌های پویا، به جای مدل پیشنهادی (بخش (6-1))
- استفاده از مدل‌های یادگیری دیگر مانند SVM و مدل پنهان مارکوف به جای شبکه عصبی (بخش (1-6-6))
- استفاده همزمان از چند نوع مدل یادگیری در گروه (بخش (1-6-6))
- تعریف معیاری کاراتر برای اضافه کردن شبکه جدید به گروه (بخش (1-6-6))
- حذف تماس‌های غیرمحتمل¹ از مجموعه تماس‌های پیش‌بینی شده بر اساس قوانین بیوشیمی. بخشی از این قواعد در جدول (1-7) آورده شده است.

جدول (1-7) قوانین بیوشیمی برای حذف تماس‌های پیش‌بینی شده غیرمحتمل [27]

<p>قانون حداکثر همسایه (maximum neighbor rule) : یک اسید آمینه می‌تواند حداکثر با 12 اسید آمینه دیگر در تماس باشد.</p>
<p>قانون حداکثر تماس دو جانبه (maximum mutual contact rule): اگر اسید آمینه α و β با یکدیگر در تماس باشند، حداکثر 6 اسید آمینه در تماس با هر دوی آن‌ها وجود دارد.</p>
<p>قانون تماس دو جانبه مارپیچ (helix mutual contact rule): یک اسید آمینه نمی‌تواند در یک زمان با اسید آمینه‌های واقع بر دو طرف مخالف یک مارپیچ در تماس باشد.</p>

¹ Improbable

قانون مارپیچ (helix rule): درون یک مارپیچ، تنها تماس بین اسیدهای i و $i+4$ مجاز است.

قانون بتا (β -rule): هیچ تماси درون یک رشته بتا مجاز نیست.

قانون دفن رشته (strand burial rule): اگر یک رشته بتواند با یکی از دو رشته دیگر جفت شود، رشته‌ای را انتخاب می‌کند که غیرقطبی¹‌تر است.

قاعده جفت شدن بتا (β -pairing rule): یک رشته بتا می‌تواند با حداقل دو رشته بتای دیگر در تماس باشد.

¹ Non-polar

پیوست الف

جدول مقادیر تمایل

AA	A	R	N	D	C	E	Q	G
A	0.0049	0.00262	0.00275	0.00254	0.00418	0.00217	0.00243	0.00402
R	0.00262	0.00187	0.00227	0.00308	0.00289	0.00271	0.00188	0.00288
N	0.00275	0.00227	0.00341	0.00273	0.00302	0.00203	0.00223	0.00348
D	0.00254	0.00308	0.00273	0.00196	0.0022	0.00133	0.00169	0.00293
C	0.00418	0.00289	0.00302	0.00219	0.0113	0.0018	0.00229	0.00399
E	0.00217	0.00271	0.00203	0.00133	0.0018	0.00111	0.00147	0.00213
Q	0.00243	0.00188	0.00223	0.00169	0.00229	0.00147	0.00197	0.00268
G	0.00402	0.00288	0.00348	0.00293	0.00399	0.00213	0.00268	0.00455
H	0.00361	0.00253	0.00306	0.00356	0.00407	0.00246	0.00209	0.00376
I	0.00554	0.00272	0.0028	0.00234	0.00453	0.00237	0.00298	0.0036
L	0.00555	0.00301	0.00271	0.00212	0.00467	0.0022	0.00262	0.00355
K	0.00224	0.00132	0.00217	0.00337	0.00186	0.00263	0.00183	0.00245
M	0.0043	0.00264	0.00279	0.00231	0.00428	0.00188	0.00248	0.00373
F	0.00579	0.00335	0.00335	0.0027	0.00541	0.00231	0.00315	0.00431
P	0.00327	0.00259	0.00297	0.00259	0.00284	0.00235	0.00263	0.00353
S	0.00364	0.00266	0.00324	0.00316	0.0038	0.0024	0.00258	0.00414
T	0.00364	0.00257	0.0032	0.0028	0.00346	0.00228	0.00265	0.00378
W	0.00383	0.00303	0.00292	0.00234	0.00514	0.00226	0.00276	0.00347
Y	0.00507	0.0035	0.00342	0.00299	0.00468	0.00269	0.003	0.00433
V	0.00543	0.00295	0.00286	0.00255	0.00556	0.00228	0.00278	0.00389

AA	H	I	L	K	M	F	P	S
A	0.00361	0.00554	0.00555	0.00224	0.0043	0.00579	0.00327	0.00364
R	0.00253	0.00272	0.00301	0.00132	0.00264	0.00335	0.00259	0.00266
N	0.00306	0.0028	0.00271	0.00217	0.00279	0.00335	0.00297	0.00324
D	0.00356	0.00234	0.00212	0.00337	0.00231	0.0027	0.00259	0.00316
C	0.00407	0.00453	0.00467	0.00186	0.00428	0.00541	0.00284	0.0038
E	0.00246	0.00237	0.0022	0.00263	0.00188	0.00231	0.00235	0.0024
Q	0.00209	0.00298	0.00262	0.00183	0.00248	0.00315	0.00263	0.00258
G	0.00376	0.0036	0.00355	0.00245	0.00373	0.00431	0.00353	0.00414
H	0.00444	0.00366	0.00366	0.00199	0.00332	0.00447	0.00307	0.00387
I	0.00366	0.00882	0.00851	0.00262	0.00626	0.00774	0.00341	0.0036
L	0.00366	0.00851	0.00786	0.00244	0.0057	0.00695	0.00358	0.00337
K	0.00199	0.00262	0.00244	0.00135	0.00213	0.00291	0.00212	0.00235
M	0.00332	0.00626	0.0057	0.00213	0.00509	0.00666	0.00315	0.00331
F	0.00447	0.00774	0.00695	0.00291	0.00666	0.0088	0.00434	0.00424
P	0.00307	0.00341	0.00358	0.00212	0.00315	0.00434	0.00342	0.00312
S	0.00387	0.0036	0.00337	0.00235	0.00331	0.00424	0.00312	0.00387
T	0.00371	0.00433	0.00401	0.00218	0.00352	0.00425	0.0032	0.00362
W	0.00365	0.00584	0.00445	0.0027	0.00501	0.00532	0.00405	0.0034
Y	0.0043	0.00646	0.00584	0.00326	0.00552	0.00656	0.00464	0.00366
V	0.00361	0.00827	0.00813	0.00254	0.00583	0.007	0.0035	0.00381

AA	T	W	Y	V
A	0.00364	0.00383	0.00507	0.00543
R	0.00257	0.00303	0.0035	0.00295
N	0.0032	0.00292	0.00342	0.00286
D	0.0028	0.00234	0.00299	0.00255
C	0.00346	0.00514	0.00468	0.00556
E	0.00228	0.00226	0.00269	0.00228
Q	0.00265	0.00276	0.003	0.00278
G	0.00378	0.00347	0.00433	0.00389
H	0.00371	0.00365	0.0043	0.00361
I	0.00433	0.00584	0.00646	0.00827
L	0.00401	0.00445	0.00584	0.00813
K	0.00218	0.0027	0.00326	0.00254
M	0.00352	0.00501	0.00552	0.00583
F	0.00425	0.00532	0.00656	0.007
P	0.0032	0.00405	0.00464	0.0035
S	0.00362	0.0034	0.00366	0.00381
T	0.00354	0.00324	0.00384	0.00434
W	0.00324	0.00334	0.00523	0.00452
Y	0.00384	0.00523	0.00611	0.00609
V	0.00434	0.00452	0.00609	0.00834

مراجع

- [1] NIH Working Definition of Bioinformatics and Computational Biology, <http://www.bisti.nih.gov/docs/CompuBioDef.pdf>, April 2009.
- [2] Luscombe, N. M., Greenbaum, D. and Gerstein, M. "What is Bioinformatics? A Proposed Definition and Overview of the Field", *Methods of Information in Medicine*, Vol. 40(4), pp. 346-358, 2001.
- [3] Polanski, A. and Kimmel, M. *Bioinformatics*, Springer, New York, 2007.
- [4] A Quick Intro to Elements of Biology: http://www.ebi.ac.uk/microarray/biology_intro.html, April 2009.
- [5] Alberts B., Johnson A., Lewis J., Raff M., Roberts K. and Walter P., *Molecular Biology of the Cell*. 4th edition, Garland Science, New York, 2002.
- [6] یزدی صمدی، ب و ولیزاده، م، ژنتیک از دیگاه مولکولی، انتشارات دانشگاه تهران، تهران، 1383.
- [7] Rhodes, G., *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, 2nd edition, Academic Press, USA, 2000.
- [8] Fasta Format Description: <http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>
- [9] Xu Y., Xu D. and Liang J., *Computational Methods for Protein Structure Prediction and Modeling, Volume 1: Basic Characterzation*, Springer, USA, 2007.
- [10] Gobel, U., Sander, C., Schneider, R. and Valencia, R., "Correlated Mutations and Residue Contacts in Proteins", *Proteins*, Vol. 18(4), pp. 309-317. 1994.
- [11] Mitra, S. and Acharya, T., *Data Mining Multimedia, Soft Computing and Bioinformatics*, John Wiley & Sons Inc, New Jersey, 2003.
- [12] Haykin, S. *Neural Networks, a Comprehensive Foundation*, 2nd edition, Prentice Hal, USA, 1999.
- [13] Robins, A., "Sequential Learning in Neural Networks: A Review and a Discussion of Pseudorehearsal Based Methods", *Intelligent Data Analysis*, Vol. 8(I3), pp. 301-322, 2004.

- [14] Ji, C. and Psaltis, D., "Capacity of Two-layer Feedforward Neural Networks with Binary Weights", *IEEE Transaction on Information Theory*, Vol. 44(I1), pp. 256-268., 1998.
- [15] Mahdaviani, K., Mazyar, H., Majidi, S. and Saraee, M. H., "A Method to Resolve the Overfitting Problem in Recurrent Neural Networks for Prediction of Complex Systems' Behavior ", *IEEE International Joint Conference on Neural Networks (IJCNN 2008)*, pp. 3723-3728, Hong Kong, China, 2008.
- [16] Tresp, V., *Committee machines*, in Hu, Y., H. and Hwang, J., N. (Eds.): *Handbook for Neural Network Signal Processing*, CRC Press, USA, 2001.
- [17] Duda, R. O., Hart, P. E. and Stork, D. G., *Pattern Classification*, 2nd edition, John Wiley & Sons, New York, 2001.
- [18] رضانی، ع و میرمحمدی میدی، س ع م، آمار و احتمالات (کاربرد در کشاورزی)، انتشارات جهاد دانشگاهی واحد صنعتی اصفهان، اصفهان، 1384.
- [19] Joachims, T., *Making Large-scale SVM Learning Practical*, in Burges, C., Smola, A. and Schlkopf, B. [Eds.], *Advances in Kernel Methods-Support Vector Learning*, MIT Press, 1999.
- [20] Olmea, O. and Valencia, A., "Improving Contact Predictions by Combination of Correlated Mutations and Other Sources of Sequence Information", *Folding and Design*, Vol. 2(3), pp. 25-32, 1997.
- [21] Park, K., Vendruscolo, M. and Domany, E., "Toward an Energy Function for the Contact Map Representation of Proteins", *Proteins: Structure, Function, and Genetics*, Vol. 40(2), pp. 237-248, 2000.
- [22] Shao, Y. and Bystroff, C., "Predicting Interresidue Contacts Using Templates and Pathways", *Proteins: Structure, Function and Genetics*, Vol. 53(S6), pp. 497-502, 2003.
- [23] Fariselli, P. and Cascadio, R., "A Neural Network Based Prediction of Residue Contacts in Protein", *Protein Engineering*, Vol. 12(1), pp. 15-21, 1999.

- [24] Fariselli, P., Olmea, O., Valencia, A. and Casadio, R., "Prediction of Contact Maps with Neural Networks and Correlated Mutations", *Protein Engineering*, Vol. 14(11), pp. 835-843, 2001.
- [25] Zhang, G. Z. and Han, k., "Hepatitis C Virus Contact Map Prediction Based on an Binary Encoding Strategy", *Computational Biology and Chemistry*, Vol. 31(I3), pp. 233-238, 2007.
- [26] Shackelford, G. and Karplus, K., "Contact Prediction Using Mutual Information and Neural Nets", *Proteins: Structure, Function and Bioinformatics*, Vol. 69(S8), pp. 159-164, 2007.
- [27] Gupta, N., Mangal, N. and Biswas, S., "Evaluation and Similarity Evaluation of Protein Structures in Contact Map Space", *Proteins: Structure, Function and Bioinformatics*, Vol. 59(2), pp. 196-204, 2005.
- [28] MacCallum, R., M., "Striped Sheets and Protein Contact Prediction", *Bioinformatics*, Vol. 20(S1), pp. 224-231, 2004.
- [29] Zhao, Y., Karypis, G., "Prediction of Contact Maps Using Support Vector Machines", *Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*, (BIBE'03), pp. 26-36, Bethesda, MD, USA, 2003.
- [30] Zaki, M. J., Jin, S. and Bystroff, C., "Mining Residue Contacts in Proteins Using Local Structure Predictions", *IEEE Transaction on System, Man and Cybernetics:Part B*, Vol. 33(5), pp. 789-801, 2003.
- [31] Fariselli, P., Compiani, M. and Casadio, R., "Predicting Secondary Structures of Membrane Proteins with Neural Networks", *European Biophysics Journal*, Vol. 22(1), pp. 41-51, 1993.
- [32] Jacoboni, I., Martelli, PL., Fariselli, P., Compiani, M. and Casadio, R., "Predictions of Protein Segments with the Same Amino Acid Sequence and Different Secondary Structure: a Benchmark for Predictive Methods", *Proteins*, 41(4): pp. 535-544, 2000.
- [33] Guo, J., Chen, H., Sun, Z. and Lin, Y., "A Novel Method for Protein Secondary Structure Prediction Using Dual-layer SVM and Profiles", *Protein: Structure, Function, Bioinformatics*, Vol. 54(4), pp. 738-743, 2004.

- [34] Bystroff, C. and Baker, D., “HMMSTR: A Hidden Markov Model for Local Sequence-structure Correlations in Proteins”, *Molecular Biology*, Vol. 301(1), pp. 173-190, 2000.
- [35] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M., “The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures”, *European Journal of Biochemistry*, Vol. 112, pp. 535-542, 1977.
- [36] Wu, S. and Zhang, Y., “A Comprehensive Assessment of Sequence-based and Template-based Methods for Protein Contact Prediction”, *Bioinformatics*, Vol. 24(7), pp. 924-931, 2008.
- [37] Hobohm, U., Scharf, M., Schneider, R. and Sander, C., “Selection of a Representative Set of Structures from the Brookhaven Protein Data Bank”, *Protein Science*, Vol. 1, pp. 409-417, 1992.
- [38] Karplus, K., Barrett, C. and Hughey, R., “Hidden Markov Models for detecting Remote Protein Homologies”, *Bioinformatics*, Vol. 14(10), pp. 846-856, 1998.
- [39] Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, M., Diekhans, M., Grate, L., Casper, J. and Hughey, R., “What Is the Value Added by Human Intervention in Protein Structure Prediction?”, *Proteins: Structure, Function, and Genetics*, Vol. 45(S5), pp. 86-91, 2001.
- [40] Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. and Hughey, R., “Combining Local-structure, Fold-recognition, and New-fold Methods for Protein Structure Prediction”, *Proteins: Structure, Function, and Genetics*, Vol. 53(S6), pp. 491-496, 2003.
- [41] Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M. and Hughey, R., “SAM-T04: What's New in Protein-structure Prediction for CASP6”, *Proteins: Structure, Function, and Bioinformatics*, Vol. 61(S7), pp. 135-142, 2005.

- [42] Katzman, S., Barrett, C., Thiltgen, G., Karchin, R. and Karplus, K., “Predict-2nd: a Tool for Generalized Protein Local Structure Prediction”, *Bioinformatics*, Vol. 24(21), pp. 2453-2459, 2008.
- [43] NR (All non-redundant GenBank CDS translation+PDB+SwissProt+PIR+PRF Database), Distributed via anonymous FTP from <ftp://ftp.ncbi.nlm.nih.gov/blast/db>.
- [44] Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, S. And Haussler, D., “Dirichlet Mixtures: a Method for Improving Detection of Weak But Significant Protein Sequence Homology”, *Computer Applications in the Biosciences*, Vol. 12(4), pp. 327–345, 1996.
- [45] SAM T08 Frequently Asked Questions: http://compbio.soe.ucsc.edu/SAM_T08/faq.html, April 2009.
- [46] Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W and Dress, A. W., “Correlations Among Amino Acid Sites in bHLH Protein Domains: an Information Theoretic Analysis”, *Molecular Biology and Evolution*, Vol. 17(1), pp. 164-178, 2000.
- [47] Fodor, A. and Aldrich, R., “Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments”, *Proteins: Structure, Function and Bioinformatics*, Vol. 56(2), pp. 211-221, 2004.

Protein Contact Map Prediction Using Committee Machine Approach

Narjes Khatoon Habibi

nhabibi@ec.iut.ac.ir

Date of Submission April 2009

Department of Electrical and Computer Engineering
Isfahan University of Technology, Isfahan 84156-83111, Iran

Degree: M.Sc

Language: Farsi

Mohammad H. Saraee, saraee@cc.iut.ac.ir

Abstract

Bioinformatics is a multi-disciplinary field that applies principles from Mathematics, Physics, Chemistry and Computer Science to widespread, numerous and complex biological data. The aim of bioinformatics is to solve biological problems in molecular level. Proteins are the basic functional units of living organism's cells which carry out almost all the activities of life. Each protein molecule comprises of an amino acid chain. Four structures are defined for protein, namely primary, secondary, tertiary (3-D) and quaternary. Primary structure is the amino acid chain. Secondary structures are local structures which are generated by hydrogen bonds. The most common types of the secondary structures are alpha helices and beta sheets. Tertiary structure is the final shape of a protein molecule which is formed by folding amino acid chain. In fact, it is the spatial status of the secondary structures in relation to each other. The quaternary structure is formed by collection of several proteins. Protein function is dependent on its tertiary structure. The determination of the 3-D structure of proteins, is an important step toward understanding the behavior of them. Tertiary structure itself is dependent on amino acid chain.

The determination of tertiary structure is not as straightforward process as the primary one. Existing experimental processes to determine tertiary structure are costly and time-consuming which encourage researchers to find methods to predict protein tertiary structure only based on its amino acid chain. Contact map prediction is one of such method. Protein contact map is a simplified, 2-D representation of protein spatial structure. The purpose of contact map prediction problem is to compute an estimate of contact map of a protein based on its primary structure and features that are computable or predictable from primary structure. Over the years, a variety of statistical and machine learning methods have been developed to predict contact map.

Committee machine is a machine learning method which divides the learning task among a number of learners and input spaces into some sub-spaces. Learner's responses to an input, are combined to produce the system's final response that is more accurate than of every individual's response. The aim of this research is to propose a novel method for contact map prediction based on committee machine. In the proposed method, learner group is a set of neural networks. Different features are extracted and then in two phases, the learner group is generated as predictive model. The important principle in evaluating contact map prediction is the ratio of correct predicted contacts to all predicted ones. To analyze the results of the proposed model, two other methods are implemented and their results are compared and presented. The results show considerable gain which is achievable by committee machine approach in contact map prediction problem.

Keywords: Bioinformatics, Machine Learning, Committee Machine, Neural Network, Protein Contact Map, Contact Map Prediction



Isfahan University of Technology

Department of Electrical and Computer Engineering

Protein Contact Map Prediction Using Committee Machine Approach

A Thesis

Submitted in partial fulfillments of the requirements
for the degree of Master of Science

By

Narjes Khatoon Habibi

Evaluated and Approved by the Thesis Committee, on April 29, 2009

1- M.H. Saraee, Assist. Prof. (Supervisor)

2- H. Korbekandi, Assist. Prof. (Advisor)

3- F. Sheiokhleslam, Assoc. Prof. (Examiner)

4- S. R. Mousavi, Assist. Prof. (Examiner)

A. M. Doost-Hoseini, Assoc. Prof. (Department Graduate Coordinator)

