

به نام خداوند بخشنده مهربان



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی پزشکی

پایان‌نامه‌ی کارشناسی ارشد
(مهندسی پزشکی - بیوالکتریک)

پیش‌بینی ساختار دوم پروتئین‌ها با استفاده از شبکه‌های عصبی

نگارش:

سپیده بابایی

استاد راهنما:

دکتر سید علی سید صالحی

زمستان ۱۳۸۷



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

بسمه تعالی

فرم اطلاعات پایان نامه
کارشناسی - ارشد و دکترا

تاریخ:

شماره:

معاونت پژوهشی
فرم پروژه تحصیلات تکمیلی ۷

مشخصات دانشجو:

نام و نام خانوادگی: سپیده بابایی
شماره دانشجویی: ۸۵۱۳۳۰۴۳
دانشجوی آزاد بورسیه معادل
دانشکده: مهندسی پزشکی رشته تحصیلی: مهندسی پزشکی گروه: بیوالکترونیک

مشخصات استاد راهنما:

نام و نام خانوادگی: دکتر سیدعلی سید صالحی
نام و نام خانوادگی: --
درجه و رتبه: استادیار درجه و رتبه: --

مشخصات استاد مشاور: --

نام و نام خانوادگی:
نام و نام خانوادگی:
درجه و رتبه:
درجه و رتبه:

عنوان پایان نامه به فارسی: پیش بینی ساختار دوم پروتئینها با استفاده از شبکه های عصبی
عنوان پایان نامه به انگلیسی:

Protein secondary structure prediction using neural networks

نوع پروژه: کارشناسی ارشد کاربرد
دکتر توسعه ای بنیادی
سال تحصیلی: نظری

تاریخ شروع: ۱۳۸۶/۹/۲۰ تاریخ خاتمه: ۱۳۸۷/۱۱/۲ تعداد واحد: ۶ سازمان تأمین کننده اعتبار:

واژه های کلیدی به فارسی: پیش بینی ساختار دوم پروتئین؛ برهمکنش محدوده وسیع؛ شبکه بازگشتی لایه ای؛ شبکه بازگشتی دوطرفه؛ هرس

واژه های کلیدی به انگلیسی: protein secondary structure prediction; long-range interaction; layered recurrent network; bidirectional recurrent network; pruning.

تعداد صفحات	تعداد مراجع	تعداد صفحات	مشخصات ظاهری
۱۴ ضمیمه	۶۶	۱۳۰	تصویر <input checked="" type="radio"/> جدول <input checked="" type="radio"/> نمودار <input checked="" type="radio"/> نقشه <input type="radio"/> واژه نامه <input type="radio"/>
زبان متن	فارسی <input checked="" type="radio"/> انگلیسی <input type="radio"/>	چکیده	فارسی <input checked="" type="radio"/> انگلیسی <input type="radio"/>
یادداشت			

نظرها و پیشنهادهای به منظور بهبود فعالیت های پژوهشی دانشگاه
استاد:

تقدیم بہ

پدر و مادر مہربانم

مشکرو قدردانی

با سپاس و ستایش بیکران به درگاه دانای یکتا، بر خود لازم می‌دانم از استاد راهنمای بزرگوار جناب آقای دکتر سید صاحبی که همواره مرا از راهنمایی‌ها و دانش خویش بهره‌مند ساخته اند سپاسگزاری نمایم.

چکیده

شناخت عملکرد پروتئین‌ها با توجه به گستره فعالیت آنها از مسائل مهم زیست‌شناسی محسوب می‌شود. ساختار سه‌بعدی یک پروتئین در نحوه عملکرد آن نقش اساسی دارد. با توجه به رشد روز افزون رشته‌های پروتئین شناخته شده و مشکلات تعیین ساختار آنها به صورت آزمایشگاهی، روشهای محاسباتی برای پیش‌بینی ساختار پروتئین از توالی آمینواسیدهای سازنده آن مورد استفاده قرار می‌گیرند. پیش‌بینی مطمئن ساختار دوم علاوه بر ارائه اطلاعات ارزشمند در مورد رشته پروتئین می‌تواند به عنوان سنگ بنای سایر پیش‌بینی‌های پیچیده‌تر ساختار سه‌بعدی به کار گرفته شود. شبکه‌های عصبی به عنوان ابزاری قدرتمند در طبقه‌بندی الگوهای رشته‌ای در این حوزه کاربرد وسیعی پیدا کرده‌اند. تعیین اندازه و طراحی ساختار مناسب شبکه با توجه به ویژگیهای الگوهای ورودی، نقش مهمی در بهبود کارایی آن ایفا می‌کند. در این کار از روش قاعده‌مند هرس شبکه‌های عصبی برای تعیین اندازه مناسب یک شبکه جلوسو با دو لایه پنهان به عنوان شبکه پایه و حفظ قدرت تعمیم‌دهی آن استفاده شده است. یک شبکه بازگشتی لایه‌ای با دو لایه زمینه که اطلاعات برهمکنش ساختارهای دوم را در طبقه‌بندی وارد می‌کند، برای افزایش صحت پیش‌بینی با الهام از شبکه‌های استفاده شده در بازشناخت گفتار، ارائه شده است. بر این مبنا سه مدل که طول همسایگی مختلفی از خروجی را در شبکه بازگشتی وارد می‌کنند، طراحی شده‌اند که نسبت به شبکه پایه افزایش صحت پیش‌بینی حاصل شده است. سه مدل دیگر بر اساس استفاده از اطلاعات برهمکنش سراسری آمینواسیدها در طول زنجیره پروتئین توسعه یافته‌اند. اتصالات بازگشتی در این مدلها در لایه‌های پنهان حافظه‌ای پویا ایجاد می‌کنند که در دو سوی رشته پروتئین گسترش یافته و اثر همجواری کل زنجیره را برای هر آمینواسید در نظر می‌گیرند. به این ترتیب کارایی شبکه در بازشناسی ساختار دوم رشته ورودی افزایش می‌یابد. در نهایت مدلی ترکیبی از شبکه بازگشتی لایه‌ای و بازگشتی دوطرفه ارائه شده است. این شبکه قادر است اطلاعات همبستگی ساختارهای دوم و همسایگی محلی و سراسری آمینواسیدها را در کل طول رشته پروتئین در نظر گیرد. صحت پیش‌بینی ساختار دوم با استفاده از ساختار جدید ارائه شده افزایش قابل توجهی دارد و به $78/66\%$ رسیده است که نسبت به شبکه پایه 6% و نسبت به پیش‌بینی‌کننده‌های معروف $4/3\%$ بهبود کارایی مشاهده می‌شود.

کلمات کلیدی: پیش‌بینی ساختار دوم پروتئین؛ برهمکنش محدوده وسیع؛ شبکه بازگشتی لایه‌ای؛ شبکه بازگشتی دوطرفه؛ هرس.

فهرست علائم اختصاری

ANN	Artificial Neural Networks
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of Techniques for Protein
CBRNN	Cascade Bidirectional Recurrent Neural Networks
DSSP	Dictionary of Secondary Structure of Protein
MLP	Multi Layer Perceptron
MSA	Multiple Sequence Alignment
PBRNN	Pollastri Bidirectional Recurrent Neural Networks
PDB	Protein Data Bank
PHD	Protein from HeiDelberg
PSSM	Position-Specific Scoring Matrices
RNN	Recurrent Neural Networks
SOV	Segment Overlap

فهرست مطالب

صفحه	عنوان
۱	فصل اول- شرح مساله-----
۲	۱-۱ مقدمه-----
۳	۲-۱ مساله پیش بینی ساختار دوم پروتئین ها-----
۴	۳-۱ روشهای موجود و مشکلات آنها-----
۷	۴-۱ اهداف پروژه-----
۹	۵-۱ ساختار پایان نامه-----
۱۱	فصل دوم- پروتئین ها-----
۱۲	۱-۲ مقدمه-----
۱۲	۲-۲ پروتئین-----
۱۳	۳-۲ ساختار آمینواسیدها-----
۱۵	۲-۳-۱ انواع آمینواسیدها-----
۱۷	۲-۴ نقش پروتئین ها-----
۱۸	۲-۵ پروتئومیکس-----
۱۹	۲-۶ همگنی-----
۱۹	۲-۷ به صف کردن چندگانه رشته های پروتئین-----
۲۰	۲-۸ ساختار پروتئین-----
۲۱	۲-۸-۱ ساختار اول-----
۲۳	۲-۸-۲ ساختار دوم-----
۲۶	۲-۸-۳ ساختار سوم-----
۲۷	۲-۸-۴ ساختار چهارم-----
۲۸	۲-۹ گروه های ساختاری-----
۳۰	۲-۱۰ تعیین ساختار پروتئین-----
۳۱	۲-۱۰-۱ پایگاه داده ساختار پروتئین ها-----
۳۲	۲-۱۱ جمع بندی-----
۳۳	فصل سوم- مروری بر روشهای پیش بینی ساختار دوم-----
۳۴	۳-۱ مقدمه-----

۳۴	-----	۲-۳ پیش‌بینی ساختار دوم
۳۵	-----	۱-۲-۳ روش‌های بهینه‌سازی تابع
۳۶	-----	۲-۲-۳ روش‌های تجربی آماری
۳۶	-----	۳-۲-۳ روش‌های یادگیری ماشینی
۳۶	-----	۱-۳-۲-۳ روش‌های نزدیک‌ترین همسایه
۳۷	-----	۲-۳-۲-۳ مدل‌های مخفی مارکوف
۳۸	-----	۳-۳ شبکه‌های عصبی مصنوعی
۳۸	-----	۱-۳-۳ اساس شبکه‌های عصبی مصنوعی
۳۹	-----	۲-۳-۳ شبکه‌های عصبی جلوسوی چند لایه
۳۹	-----	۳-۳-۳ شبکه‌های عصبی بازگشتی
۳۹	-----	۴-۳ شبکه‌های جلوسو در پیش‌بینی ساختار دوم
۴۲	-----	۱-۴-۳ روش PHD
۴۳	-----	۱-۱-۴-۳ سطح اول: شبکه رشته به ساختار
۴۲	-----	۲-۱-۴-۳ سطح دوم: شبکه ساختار به ساختار
۴۳	-----	۳-۱-۴-۳ سطح سوم: داوری
۴۵	-----	۵-۳ شبکه بازگشتی دوطرفه
۴۸	-----	۶-۳ شبکه بازگشتی دوطرفه متوالی
۵۰	-----	۷-۳ مزایا و معایب روش شبکه عصبی
۵۲	-----	۸-۳ جمع‌بندی
۵۳	-----	فصل چهارم- دادگان و شبکه پایه
۵۴	-----	۱-۴ مقدمه
۵۴	-----	۲-۴ دادگان
۵۶	-----	۳-۴ ماتریس امتیاز ویژه موقعیت
۵۷	-----	۴-۴ تخصیص ساختار دوم
۵۸	-----	۵-۴ ارزیابی
۵۸	-----	۱-۵-۴ درصد صحت
۵۸	-----	۲-۵-۴ قطعه همپوشان
۶۰	-----	۶-۴ تغییر و توسعه ساختاری در شبکه‌های عصبی
۶۱	-----	۷-۴ هرس شبکه‌های عصبی

۶۱	-----۸-۴ شبکه رشته به ساختار-----
۶۲	-----۱-۸-۴ آموزش شبکه-----
۶۴	-----۲-۸-۴ آموزش بیش از حد شبکه-----
۶۵	-----۳-۸-۴ هرس شبکه-----
۶۶	-----۹-۴ شبکه ساختار به ساختار-----
۶۸	-----۱۰-۴ ساختار نهایی-----
۶۹	-----۱۱-۴ شبکه پایه-----
۷۰	-----۱-۱۱-۴ هرس شبکه-----
۷۱	-----۲-۱۱-۴ آموزش شبکه پایه-----
۷۳	-----۱۲-۴ جمع بندی-----
۷۶	-----فصل پنجم- شبکه بازگشتی لایه‌ای و دوطرفه-----
۷۶	-----۱-۵ مقدمه-----
۷۷	-----۲-۵ دوسویه سازی پردازش در شبکه‌های عصبی-----
۷۷	-----۱-۲-۵ شبکه بازگشتی در مدل سازی اثرات هم تولیدی آواها در گفتار-----
۷۷	-----۱-۱-۲-۵ مدل بر مبنای عملکرد نئوکورتکس-----
۷۷	-----۲-۱-۲-۵ مدل بر مبنای عملکرد هیپوکمپوس-----
۷۹	-----۳-۱-۲-۵ مدل ترکیبی-----
۸۰	-----۳-۵ شبکه‌های بازگشتی لایه‌ای-----
۸۱	-----۱-۳-۵ مدل اول، شبکه بازگشتی با دو لایه زمینه-----
۸۳	-----۱-۱-۳-۵ نتایج-----
۸۴	-----۲-۳-۵ مدل دوم، ورودی شبکه بازگشتی خروجی قبلی و بعدی-----
۸۵	-----۱-۲-۳-۵ نتایج-----
۸۶	-----۳-۳-۵ مدل سوم، ورودی شبکه بازگشتی ۷ خروجی قبلی و بعدی-----
۸۷	-----۱-۳-۳-۵ نتایج-----
۸۸	-----۴-۵ شبکه با اتصالات بازگشتی در لایه‌های پنهان-----
۸۹	-----۱-۴-۵ مدل چهارم، شبکه با یک اتصال بازگشتی در لایه پنهان اول-----
۹۰	-----۱-۱-۴-۵ نتایج-----
۹۱	-----۲-۴-۵ مدل پنجم، شبکه با دو اتصال بازگشتی در دو لایه پنهان-----
۹۲	-----۱-۲-۴-۵ نتایج-----

۹۳	----- ۳-۴-۵ مدل ششم، شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان
۹۵	----- ۱-۳-۴-۵ نتایج
۹۶	----- ۵-۵ ترکیب مدلها
۹۸	----- ۱-۵-۵ نتایج
۱۰۰	----- ۶-۵ جمع بندی
۱۰۴	----- فصل ششم - جمع بندی و ارائه پیشنهاد
۱۰۵	----- ۱-۶ جمع بندی
۱۰۹	----- ۲-۶ ارائه پیشنهاد
۱۱۰	----- مراجع
۱۱۵	----- ضمیمه الف
۱۲۱	----- ضمیمه ب

فهرست شکل‌ها

- شکل ۱-۲ دو آمینواسید متصل به هم [۶۳]. ۱۳
- شکل ۲-۲ مثالی ساده برای درک نحوه محاسبه پروفایل صف‌بندی چندگانه رشته پروتئین [۳۹]. ۲۰
- شکل ۳-۲ ساختار اولیه پروتئین G، قسمت B1 با رنگ قرمز مشخص شده است [۶۴]. ۲۲
- شکل ۴-۲ نمایش ماریچ آلفا، الف) نمایش نمادین ب) نمایش گوی و میله [۶۴]. ۲۴
- شکل ۵-۲ نمایش صفحات بتا، الف) نمایش نمادین ب) نمایش گوی و میله [۶۴]. ۲۵
- شکل ۶-۲ ساختار دوم ناحیه B1 از پروتئین G [۶۴]. ۲۵
- شکل ۷-۲ ساختار سوم ناحیه B1 در پروتئین G [۶۴]. ۲۶
- شکل ۸-۲ ساختار چهارم هموگلوبین انسان [۶۴]. ۲۷
- شکل ۹-۲ پروتئین CD8 یک نوع سلول T متعلق به گروه ساختاری بتا [۶۵]. ۲۸
- شکل ۱۰-۲ پروتئین synthase tryptohan، متعلق به گروه ساختاری آلفا بر بتا [۶۵]. ۲۹
- شکل ۱۱-۲ پروتئین 5RNB، متعلق به گروه ساختاری آلفا با بتا [۶۵]. ۲۹
- شکل ۱۲-۲ یک پروتئین غشایی به نام 10PF [۶۵]. ۳۰
- شکل ۱-۳ ساختار شبکه‌ی عصبی جلوسوی چند لایه برای پیش‌بینی ساختار دوم [۲۲]. ۳۹
- شکل ۲-۳ نمای کلی روش PHD. ۴۴
- شکل ۳-۳ شبکه BRNN با دو قسمت پیش سو و پس سو. Ft و Bt دو شبکه بازگشتی هستند [۲۴]. ۴۵
- شکل ۴-۳ شبکه بازگشتی دوطرفه متوالی [۲۷]. ۴۸
- شکل ۵-۳ روند استنتاج الگوریتم CBRNN [۲۷]. ۴۹
- شکل ۱-۴ قسمتی از PSSM یک پروتئین [۶۵]. ۵۷
- شکل ۲-۴ شبکه جلوسو با یک لایه پنهان. ۶۲
- شکل ۳-۴ خطای آموزش و آزمون شبکه اول بر روی دادگان گروه A. به دلیل آموزش بیش از حد شبکه با وجود کاهش خطای آموزش، خطای آزمون پس از مدتی افزایش می‌یابد. ۶۴
- شکل ۴-۴ خطای آموزش شبکه ساختار به ساختار برای دادگان گروه A. ۶۷

- ۷۱ شکل ۴-۵ شبکه پایه، یک شبکه جلو سو با دو لایه پنهان (Net2).
- ۷۲ شکل ۴-۶ خطای آموزش شبکه پایه برای دادگان گروه A.
- ۷۸ شکل ۵-۱ ساختار مدل الهام گرفته شده از نئوکورتکس [۵۲]
- ۷۸ شکل ۵-۲ مدل بازشناس بر مبنای عملکرد هیپوکمپوس [۵۲]
- ۷۹ شکل ۵-۳ ساختار مدل مرکب از مدل‌های الهام گرفته شده از نئوکورتکس و هیپوکمپوس [۵۲]
- ۸۰ شکل ۵-۴ دو نمونه از شبکه‌های بازگشتی (لایه‌ای الف) شبکه جردن و (ب) شبکه المان.
- ۸۲ شکل ۵-۵ مدل اول، یک شبکه بازگشتی با دو لایه زمینه (Net3).
- ۸۴ شکل ۵-۶ مدل دوم، ورودی شبکه بازگشتی خروجی سه حالت فعلی، قبلی و بعدی است (Net4).
- ۸۶ شکل ۵-۷ مدل سوم، ورودی شبکه بازگشتی خروجی ۷ حالت قبلی و بعدی شبکه است (Net5).
- ۸۹ شکل ۵-۸ طرح ساده عملکرد هیپوکمپوس به صورت یک شبکه با یک اتصال بازگشتی.
- ۹۰ شکل ۵-۹ طرح مدل چهارم، شبکه با یک اتصال بازگشتی در لایه پنهان اول (Net6).
- ۹۲ شکل ۵-۱۰ طرح مدل پنجم، شبکه با دو اتصال بازگشتی در دو لایه پنهان (Net7).
- ۹۵ شکل ۵-۱۱ طرح شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان که در زمان باز شده است (Net8). اتصالات پیش سو اطلاعات زمانهای قبل و اتصالات پس سو اطلاعات زمانهای بعد را در لایه‌های پنهان شبکه وارد می‌کنند.
- ۹۷ شکل ۵-۱۲ شبکه ترکیبی (Net9)، ترکیبی از شبکه بازگشتی با دو لایه زمینه (Net4) و شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان (Net8).
- ۱۰۰ شکل ۵-۱۳ صحت پیش‌بینی کل ساختار دوم با مدل‌های ارائه شده و میزان تغییرات آن نسبت به شبکه پایه.
- ۱۰۱ شکل ۵-۱۴ میزان همپوشانی ساختار دوم پیش‌بینی شده با مدل‌های ارائه شده و میزان تغییرات آن نسبت به شبکه پایه.
- ۱۰۲ شکل ۵-۱۵ صحت پیش‌بینی ساختار دوم مارپیچ (H)، صفحات بتا (E) و پیچ (C) با مدل‌های ارائه شده.
- ۱۰۳ شکل ۵-۱۶ میزان تغییرات صحت پیش‌بینی ساختار دوم مارپیچ (H)، صفحات بتا (E) و پیچ (C) با مدل‌های ارائه شده نسبت به شبکه پایه.

شکل ۶-۱ صحت پیش‌بینی و میزان همپوشانی ساختار دوم شبکه‌های بازگشتی مختلف بر ۱۰۸
روی دادگان PSIPRED. سه روش معروف بازگشتی ساده (RNN)، شبکه بازگشتی
Pollastri (PBRNN)، شبکه بازگشتی متوالی (CBRNN) و شبکه‌های ارائه شده بازگشتی
لایه‌ای و دوطرفه.

فهرست جدول‌ها

- ۶ جدول ۱-۱ صحت روشهای مختلف پیش‌بینی ساختار دوم.
- ۱۴ جدول ۱-۲ بیست آمینواسید تشکیل دهنده پروتئین‌ها.
- ۲۳ جدول ۲-۲ انواع ساختارهای دوم و نمادهای آنها.
- ۳۱ جدول ۳-۲ تعداد ساختارهای تعیین شده توسط روشهای مختلف در بانک دادگان پروتئین تا ماه نوامبر سال ۲۰۰۸.
- ۴۷ جدول ۱-۳ تعداد گره‌ها در لایه‌های مختلف شبکه [۲۴].
- ۴۷ جدول ۲-۳ صحت پیش‌بینی برای دادگان RS126 (الف) ۳ نوع (ب) ۸ نوع ساختار دوم.
- ۵۰ جدول ۳-۳ صحت پیش‌بینی شبکه BRNN معمولی، PBRNN و CBRNN.
- ۵۶ جدول ۱-۴ تعداد زنجیره‌های پروتئین و توزیع ساختار دوم.
- ۵۸ جدول ۲-۴ روشهای کاهش ساختار دوم از ۸ به ۳ نوع.
- ۶۳ جدول ۳-۴ صحت پیش‌بینی شبکه با اندازه بیش از حد.
- ۶۵ جدول ۴-۴ صحت پیش‌بینی و تعداد گره‌های بدست آمده بعد از هرس با روش اسکلت‌بندی.
- ۶۶ جدول ۵-۴ صحت پیش‌بینی و تعداد گره‌های بدست آمده بعد از هرس با روش ترکیب گره‌های همبسته.
- ۶۷ جدول ۶-۴ میانگین صحت پیش‌بینی و تعداد گره‌های بدست آمده بعد از هرس با روش اسکلت‌بندی و ترکیب گره‌های همبسته.
- ۶۸ جدول ۷-۴ صحت پیش‌بینی شبکه جلوسو با تعداد گره‌های پنهان به دست آمده پس از هرس با هر دو روش و سپس استفاده از شبکه ساختار به ساختار.
- ۷۰ جدول ۸-۴ درصد صحت پیش‌بینی ساختار دوم (Q3) به تفکیک برای ۳ گروه مجموعه دادگان {A, B, C} و ۳ طول مختلف پنجره ورودی.
- ۷۰ جدول ۹-۴ صحت پیش‌بینی و تعداد گره‌های به دست آمده پس از هرس.
- ۷۳ جدول ۱۰-۴ صحت (Q3) و میزان همپوشانی (SOV) پیش‌بینی شبکه پایه برای سه نوع ساختار دوم در سه گروه دادگان.
- ۸۳ جدول ۱-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان

همپوشانی توسط شبکه بازگشتی با دو لایه زمینه (Net3).

- ۸۵ جدول ۲-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه بازگشتی با دو لایه زمینه، ورودی شبکه بازگشتی خروجی سه حالت فعلی، قبلی و بعدی است (Net4).
- ۸۷ جدول ۳-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی شبکه بازگشتی با دو لایه زمینه، ورودی شبکه بازگشتی خروجی ۷ حالت قبلی و بعدی شبکه است (Net5).
- ۸۸ جدول ۴-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه بازگشتی با دو لایه زمینه، با افزایش گره‌های لایه‌های زمینه، ورودی شبکه بازگشتی خروجی ۷ حالت قبلی و بعدی شبکه است (Net5).
- ۹۱ جدول ۵-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه با یک اتصال بازگشتی (Net6).
- ۹۳ جدول ۶-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه با دو اتصال بازگشتی (Net7).
- ۹۶ جدول ۷-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان (Net8).
- ۹۹ جدول ۸-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه ترکیبی (Net9)، ترکیبی از شبکه بازگشتی با دو لایه زمینه (Net4) و شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان (Net8).
- ۹۹ جدول ۹-۵ صحت کل و صحت پیش‌بینی برای هریک از انواع ساختار دوم و میزان همپوشانی توسط شبکه ترکیبی (Net9)، ترکیبی از شبکه بازگشتی با دو لایه زمینه که ورودی شبکه بازگشتی خروجی ۷ حالت قبلی و بعدی شبکه است (Net5) و شبکه با اتصال بازگشتی دو طرفه در دو لایه پنهان (Net8).

فصل اول

شرح مساله

۱-۱ مقدمه

پروتئین‌ها از مولکولهای مهم زیستی به شمار می‌آیند و اکثر پردازش‌های زیستی را در موجود زنده اداره می‌کنند. از جمله نقشهای زیستی بسیار متنوعی که بر عهده دارند می‌توان به ترجمه^۱ اطلاعات و انتقال آنها، انتقال فرمانهای مختلف دستگاههای بدن به یکدیگر (هورمونها)، تسریع^۲ فعالیتهای شیمیایی، انتقال و نگهداری مواد، و انجام کارهای مکانیکی (حرکت و تحمل بار) اشاره کرد. دامنه^۳ گسترده^۴ فعالیتهای بالا، اهمیت زیستی این مولکولها را نشان می‌دهد. هر مولکول پروتئین، از رشته‌ای از آمینواسیدها یا به اصطلاح مانده^۵ تشکیل شده است و در محیط آبی شکلی سه بعدی به خود می‌گیرد. به این ترتیب ساختار آن در چهار تراز قابل توصیف است. ساختار نوع اول توالی آمینواسیدهای^۳ سازنده پروتئین است. ساختار نوع دوم، ساختار فضایی منظمی است که بر اثر تا خوردگی^۴ توالی آمینواسیدها به صورت موضعی^۵ در ساختار سه بعدی یا ساختار نوع سوم پروتئین پدید می‌آید و ساختار نوع چهارم، ساختار سه بعدی مجموعه^۵ چند رشته درهم پیچیده است. ساختار دوم پروتئین ارتباط نزدیکی با ساختار سوم آن دارد که رفتار، عملکرد و ویژگیهای پروتئین‌ها را تعیین می‌کند [۱]. بنابراین، شناسایی ساختار دوم یک پروتئین سنگ بنای شناخت و توصیف عملکرد آن می‌باشد.

¹ catalysis

² residue

³ amino acids sequence

⁴ folding

⁵ local

۱-۲ مساله پیش‌بینی ساختار دوم پروتئین‌ها^۱

بسیاری از تحقیقات در دهه اخیر بر روی مطالعه و پیش‌بینی ساختار پروتئین بوده است. هدفی مهم در بسیاری از کاربردها، شناخت کارکرد پروتئین است [۱]. تعیین ساختار سه بعدی پروتئین، گامی مهم در آشکار کردن عملکرد آن است. هم اکنون برای تعیین دقیق این ساختار، روشهای آزمایشگاهی بلورنگاری^۲ پرتو X، تشدید مغناطیس هسته چندبعدی^۳ و میکروسکوپیهای الکترونی مورد استفاده قرار می‌گیرند که این روشها، نیازمند صرف هزینه و زمان زیادی هستند. علاوه بر این، در بعضی موارد مثلاً در مورد بعضی از پروتئین‌های غشایی نمی‌توان از بلورنگاری استفاده کرد. در مواردی که پروتئین بسیار بزرگ باشد، تصویر برداری با تشدید مغناطیسی، پاسخ دقیقی به دست نمی‌دهد [۲]. تا امروز تعداد کل ساختارهای تعیین شده به صورت آزمایشگاهی کمتر از ۵۰ هزار ساختار است، درحالی که بیش از یک میلیون توالی پروتئین شناخته شده است [۳].

ساختار دوم یا تاخوردگی زنجیره پروتئین در شکل‌های فضایی مختلف مانند مارپیچ، صفحات بتا و یا پیچ‌های تصادفی است و شناخت آن بنا به این دلایل مهم است (۱) شاخصی از تاخوردگی است. (۲) یک تجسم اولیه از ساختار کل پروتئین را ارائه می‌کند. (۳) صف‌بندی رشته‌های^۴ پروتئینی که ناشی از روند تکامل است را تحت تأثیر قرار می‌دهد. (۴) عملکرد پروتئین به آن وابسته است [۵]. درکنار اطلاعات ارزشمند درمورد رشته پروتئین یک پیش‌بینی مطمئن ساختار دوم پروتئین می‌تواند به عنوان سنگ بنای سایر پیش‌بینی‌های بسیار مشکل‌تر برای پیش‌بینی کامل شکل سه‌بعدی به کار رود. درسالهای اخیر استفاده وسیعی از روشهای محاسباتی برای پیش‌بینی ساختار سه‌بعدی پروتئین‌ها از روی توالی آمینواسیدها صورت گرفته است. تعیین ساختار دوم یک گام اساسی برای این روشهاست [۶]، [۵]. علاوه بر کاربرد پیش‌بینی ساختار سه بعدی، پیش‌بینی ساختار دوم فواید دیگری هم دارد. برای مثال پیش‌بینی می‌تواند برای فهم اطلاعات بیوشیمی و عملکرد زیستی و در تشخیص ناحیه‌ای که پروتئین دستخوش تغییرات شده است استفاده بشود [۷]. از ساختار تعیین شده می‌توان برای مهندسی پروتئین، طراحی دارو و مطالعات ایمنی‌شناسی استفاده شود. با این روش داروها می‌توانند برای درمان بیماریهای خاص مانند کم خونی سلولهای داسی^۵، پارکینسون^۶، آلزایمر^۷ و بسیاری از بیماریهای متابولیسمی و جنینی طراحی شوند [۴].

¹protein secondary structure prediction

²crystallography

³ multidimensional nuclear magnetic resonance

⁴sequence alignment

⁵sickle-cell anemia

⁶parkinson

⁷Alzheimer

در تمام موارد بالا، درک کارکرد پروتئین بر اساس اطلاعات رشته سازنده آن اهمیت دارد و مشکل در همین جاست. روش متعارف دستیابی به کارکرد، مبتنی بر دانستن شکل فضایی مولکول پروتئین است، ولی چنانکه ذکر شد، به دست آوردن این شکل بسیار پرهزینه و در مواردی ناممکن است. چاره این مشکل، در استفاده از روشهای میانبری است که از اطلاعات رشته تشکیل دهنده پروتئین، ما را مستقیماً به کارکرد آن برسانند. بازشناسی الگو^۱، مهمترین این روشهاست [۱].

۱-۳ روشهای موجود و مشکلات آنها

یکی از مهمترین مسائل در زیست شناسی محاسباتی، پیش‌بینی ساختار دوم پروتئین داده شده از روی توالی آن است و در طی دهه گذشته تلاشهای بسیاری برای حل مسأله انجام شده است. سه روش اصلی برای پیش‌بینی ساختار دوم روشهای بهینه سازی تابع، روشهای تجربی آماری و روشهای یادگیری ماشینی هستند. در روشهای بهینه سازی تابع، روش محاسباتی کلی روش *ab-initio* است که در آن هدف کمینه کردن تابع انرژی مولکول است. ولی هم از لحاظ محاسباتی پرهزینه‌اند و هم صورت دقیقی از تابع هدف بهینه‌سازی که همان تابع انرژی مولکول است، وجود ندارد [۸]. روشهای تجربی آماری پیش‌بینی ساختار پروتئین اولین روشهای پیش‌بینی ساختار از روی توالی آمینواسیدها می‌باشد. دو روش پیش‌رو روش *Chou-Fasman* [۹] و *GOR*^۲ [۱۰] است. این راهکارها نسل اول روشهای پیش‌بینی محسوب می‌شوند که فقط از توالی آمینواسیدها برای پیش‌بینی استفاده می‌کنند و صحت پیش‌بینی حاصل کمتر از ۶۰٪ است.

روشهای یادگیری ماشینی پیش‌بینی ساختار دوم پروتئین‌ها از روی ساختارهای شناخته شده را می‌توان یک مساله طبقه‌بندی با سرپرستی در نظر گرفت [۱]. در نسل دوم اثر همسایگی آمینواسیدها در طول رشته‌های پروتئین، با استفاده از پنجره لغزان^۳ روی ورودی در نظر گرفته شده و از روشهای بازشناسی الگو برای پیش‌بینی استفاده می‌شود. اولین روش *k* همسایه نزدیک^۴ است [۱۱]. از آنجا که طبقه‌بندی با این روش بر مبنای شباهت رشته‌هاست، به کارگیری آن وقتی مناسب است که شباهت بین نمونه‌ها و الگو زیاد باشد. صحت پیش‌بینی با مجموعه دادگان که میزان همگنی کمی دارند، پایین است. نسل سوم علاوه بر آمینواسیدهای زنجیره پروتئین از اطلاعات تکمیلی دیگر مانند میزان حلالیت و آب‌گریزی آمینواسیدها [۴]، اطلاعات تکاملی صف‌بندی رشته‌های پروتئین [۱۲]، ویژگیهای فیزیکی و شیمیایی آمینواسیدها [۱۳] و گرافهای برهمکنش^۵ [۱۴] استفاده کرده‌اند. مدل‌های

¹ pattern recognition

² Garnier-Osguthorpe-Robson

³ sliding window

⁴ k-nearest neighbour

⁵ intraction graph