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
(پلی‌تکنیک تهران)

بسمه تعالی

**فرم اطلاعات پایان‌نامه**

**کارشناسی- ارشد و دکترا**

تاریخ:
شماره:

**معاونت پژوهشی**
**فرم پروژه تحصیلات تکمیلی ۷**

**مشخصات دانشجو:**

نام و نام خانوادگی: سپیده بابایی | دانشجوی آزا د ● | بورسیه ○ | معادل ○

شماره دانشجوئی: ۸۵۱۳۳۰۴۳ | دانشکده: مهندسی پزشکی | رشته تحصیلی: مهندسی پزشکی | گروه: بیوالکتریک

**مشخصات استاد راهنما:**

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

نام و نام خانوادگی: -- | درجه و رتبه: --

**مشخصات استاد مشاور: --**

نام و نام خانوادگی: | درجه و رتبه:

نام و نام خانوادگی: | درجه و رتبه:

**عنوان پایان‌نامه به فارسی : پیش بینی ساختار دوم پروتئینها با استفاده از شبکه های عصبی**

**عنوان پایان‌نامه به انگلیسی:**

**Protein secondary structure prediction using neural networks**

نوع پروژه: کارشناسی ○ | ارشد ● | دکتر ا ○ | سال تحصیلی:

کاربردی ● | بنیادی ○ | توسعه‌ای ○ | نظری ○

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

**واژه‌های کلیدی به فارسی:** پیش‌بینی ساختار دوم پروتئین؛ برهمکنش محدوده وسیع؛ شبکه بازگشتی لایه‌ای؛ شبکه بازگشتی دوطرفه؛ هرس

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

| | | | | |
|---|---|---|---|---|
| مشخصات ظاهری | تعداد صفحات ۱۳۰ | تصویر ● جدول ● نمودار ● نقشه ○ واژه‌نامه ○ | تعداد مراجع ۶۶ | تعداد صفحات ضمائم ۱۴ |
| زبان متن | فارسی ● | انگلیسی ○ | چکیده | فارسی ● انگلیسی ○ |
| یادداشت | | | | |

**نظرها و پیشنهادها به منظور بهبود فعالیت‌های پژوهشی دانشگاه**

استاد:

تقدیم به

پدر و مادر مهربانم

# تشکر و قدردانی

با سپاس و ستایش بیکران به درگاه دانای یکتا، بر خود لازم می دانم از استاد راهنمای بزرگوار جناب آقای دکتر سید صالحی که همواره مرا از راهنمایی ها و دانش خویش بهره مند 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 درمورد رشته پروتئین می‌تواند به عنوان سنگ بنای سایر پیش‌بینی‌های پیچیده‌تر ساختار سه‌بعدی به‌کار 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 می‌کنند. 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 درنظر 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 ٪۷۸/۶۶ رسیده است که نسبت به شبکه پایه ٪۶ و نسبت به پیش‌بینی‌کننده‌های معروف ٪۴/۳ 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

# فهرست شکل‌ها

# فهرست جدول‌ها

# 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 ترجمهٔ اطلاعات و انتقال آنها، انتقال فرمانهای مختلف دستگاههای بدن به یکدیگر (هورمونها)، تسریع[۱] فعالیتهای شیمیایی، انتقال و نگهداری مواد، و انجام کارهای مکانیکی (حرکت وتحمل بار) اشاره کرد. دامنهٔ گستردهٔ فعالیتهای بالا، اهمیت زیستی این مولکولها را نشان می‌دهد. هر مولکول پروتئین، از رشته‌ای از آمینواسیدها یا به اصطلاح مانده[۲] 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] سازندهٔ 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 مجموعهٔ چند رشته درهم پیچیده است. ساختاردوم 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[1]

بسیاری از تحقیقات در دهه اخیر برروی 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[2] پرتو X، تشدید مغناطیس هسته چندبعدی[3] و میکروسکوپهای 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 به‌دست نمی‌دهد [۲]. تا امروز تعداد کل ساختارهای تعیین شده به صورت آزمایشگاهی کمتر از ۵۰ هزار ساختار است، درحالی که بیش ازیک 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[4] 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 به‌کار 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 ساختاردوم 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[5]، پارکینسون[6]، آلزایمر[7] و بسیاری ازبیماریهای متابولیسمی و جنینی طراحی شوند [۴].

[1] protein secondary structure prediction
[2] crystallography
[3] multidimentional nuclear magnetic resonance
[4] sequence alignment
[5] sickle-cell anemia
[6] parkinson
[7] Alzeimer

در تمام موارد بالا، درک کارکرد پروتئین بر اساس اطلاعات رشتهٔ سازندهٔ آن اهمیت دارد و مشکل در همین جاست. روش متعارف دستیابی به کار کرد، مبتنی بر دانستن شکل فضایی مولکول پروتئین است، ولی چنانکه ذکر شد، به دست آوردن این شکل بسیار پرهزینه و در مواردی ناممکن است. چارهٔ این مشکل، در استفاده از روشهای میان‌بری است که از اطلاعات رشتهٔ تشکیل‌دهندهٔ پروتئین، ما را مستقیماً به کارکرد آن برسانند. بازشناسی الگو[1]، مهمترین این روشهاست [۱].

## ۱-۳ روشهای موجود و مشکلات آنها

یکی ازمهمترین مسائل در زیست شناسی محاسباتی، پیش‌بینی ساختاردوم پروتئین داده شده از روی توالی آن است و درطی 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[2] [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 باسرپرستی در نظر گرفت [۱]. در نسل دوم اثر همسایگی آمینواسیدها در طول رشته‌های پروتئین، با استفاده از پنجره لغزان[3] روی ورودی در نظر گرفته شده و از روشهای بازشناسی الگو برای پیش‌بینی استفاده می‌شود. اولین روش k همسایه نزدیک[4] 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 آب گریزی آمینواسیدها [۴]، اطلاعات تکاملی صف‌بندی رشته‌های پروتئین [۱۲]، ویژگیهای فیزیکی و شیمیایی آمینواسیدها [۱۳] و گرافهای برهمکنش[5] [۱۴] استفاده کرده‌اند. مدلهای

[1] pattern recognition
[2] Garnier-Osguthorpe-Robson
[3] sliding window
[4] k-nearest neighbour
[5] intraction graph