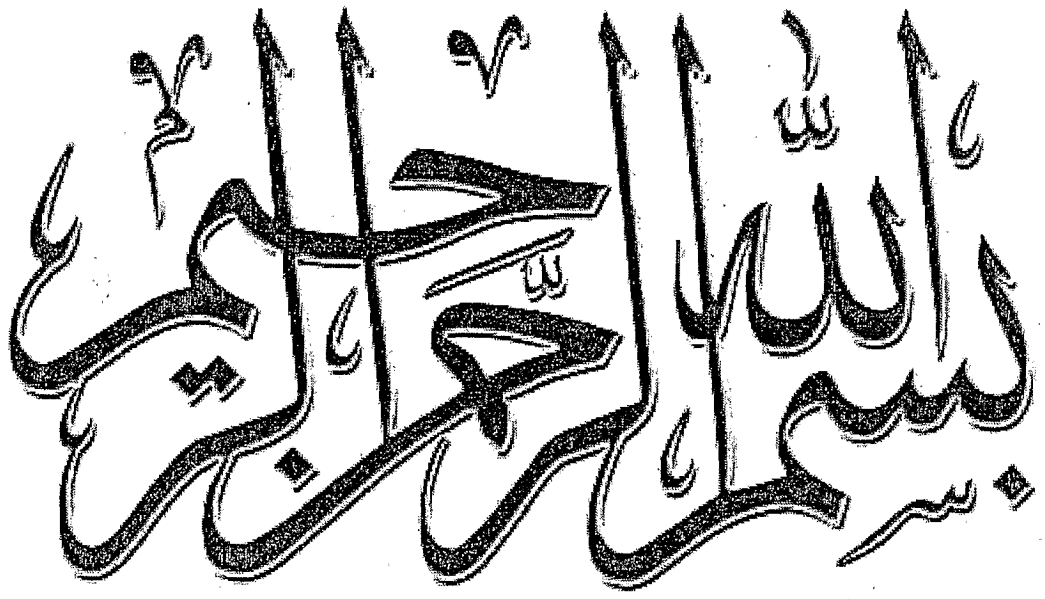
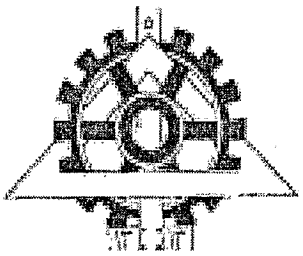
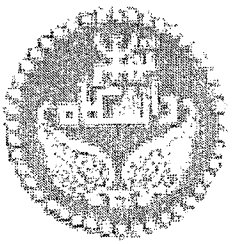

بسم الله الرحمن الرحيم





دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

تجمیع پایگاه‌های داده بیوانفورماتیک با رویکرد ترکیب داده

نگارش:

عادل اردلان

استاد راهنما: دکتر بهزاد مشیری

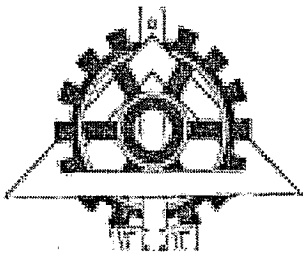
استاد مشاور: دکتر مسعود رهگذر

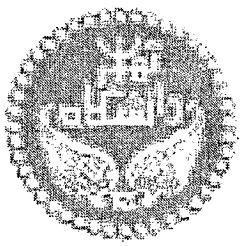
پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد

در

رشتهٔ مهندسی کامپیوتر- گرایش مهندسی فناوری اطلاعات

اسفند ماه ۱۳۸۶





دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد

در رشتهٔ مهندسی کامپیوتر- گرایش مهندسی فناوری اطلاعات

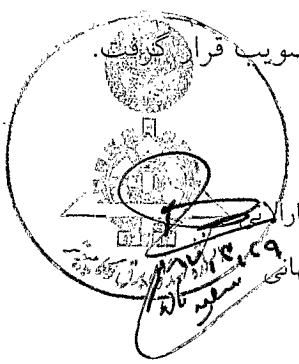
عنوان:

تجمیع پایگاه‌های داده بیوانفورماتیک با رویکرد ترکیب داده

نگارش:

عادل اردلان

این پایان‌نامه در تاریخ ۸۶/۱۱/۱۴ در مقابل هیأت داوران دفاع گردید و مورد تصویب قرار گرفت.



معاونت آموزشی و تحصیلات تکمیلی پردیس دانشکده‌های فنی	دکتر جواد فیض
رئیس دانشکده مهندسی برق و کامپیوتر	دکتر پرویز جبه‌دار مارالانی
معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر	دکتر سعید نادر اصفهانی
استاد راهنما	دکتر بهزاد مشیری
استاد مشاور	دکتر مسعود رهگذر
عضو هیأت داوران	دکتر مهدی صادقی
عضو هیأت داوران	دکتر بابک نجار اعرابی
عضو هیأت داوران	دکتر کارو لوکس

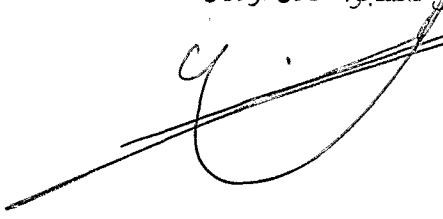
تعهد نامه اصالت اثر

اینجانب **عادل اردلان** تأیید می‌نمایم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان-نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

کلیه حقوق مادی و معنوی این اثر متعلق به پردیس دانشکده‌های فنی دانشگاه تهران می باشد.

نام و نام خانوادگی دانشجو: **عادل اردلان**

امضای دانشجو:



به

کامران، فہیمہ و سحر

به

سہیل و حسام الدین

چکیده

تعیین ساختار مولکول‌های زیستی از مهم‌ترین و پیچیده‌ترین مسائل مطرح شده در حیطه زیست-شناسی مولکولی است. تشخیص این ساختارها به شناخت کارکرد این عوامل در مولفه‌های حیاتی کمک می‌کند. از این رو تلاش فراوانی برای بررسی قابلیت پیش‌بینی این ساختارها با استفاده از روش‌های محاسباتی صورت پذیرفته است. پیچیدگی این مساله ضرورت بررسی هرچه بیشتر را در جهت شناخت دقیق‌تر فرآیندهای حیاتی افزون می‌نماید.

در این پایان‌نامه از بستر تئوری شواهد دمپستر–شافر بعنوان مبنایی برای ارائه‌ی روشی جهت مدلسازی ساختار دوم پروتئین‌ها استفاده شده است. بر اساس این مدل ساختار دوم پروتئین‌ها از طریق دو منبع داده‌ای مهم مورد بررسی و ارزیابی قرار می‌گیرند و از نتایج این بررسی در پیش‌بینی ساختار پروتئین‌هایی که ساختار آنها ناشناخته است، استفاده می‌گردد. منابع اطلاعاتی مورد استفاده با توجه به این مطلب انتخاب شده‌اند که اطلاعات ساختاری قابل توجهی را در ارتباط با آرایش درشت‌مولکول‌های پروتئینی در اختیار قرار می‌دهند. این منابع عبارتند از ساختار اول پروتئین (پایگاه داده‌های پروتئین‌ها[1]) و اطلاعات جابجایی شیمیایی (پایگاه داده‌های تشدید مغناطیسی زیستی[2]).

در مدل ارائه‌شده رشته‌های پروتئینی بر مبنای توالی اسیدهای آمینه بررسی و شواهد موجود استخراج می‌شوند. سپس برای هر اسید آمینه اعداد جابجایی شیمیایی اندازه‌گیری‌شده با دقت ازپیش‌تعیین‌شده، از داده‌های موجود استخراج می‌گردند. در مرحله‌ی بعد به هریک از ساختارهای مورد نظر برای انجام عملیات پیش‌بینی، جرم احتمال تخصیص داده می‌شود. نتایج بدست‌آمده نشان‌دهنده بهبود قابل ملاحظه-ای در صحت پیش‌گویی هریک از ساختارها و نیز صحت پیش‌بینی کلی (Q_3) است. از این جهت این روش می‌تواند برای پیش‌بینی قابل اطمینان ساختار دوم پروتئین‌ها مورد استفاده قرار گیرد.

کلمات کلیدی: تئوری شواهد دمپستر–شافر، بیوانفورماتیک، ساختار دوم پروتئین‌ها، طیف‌سنجی مغناطیسی هسته.

---

[1] Protein Data Bank (PDB) - http://www.rcsb.org/pdb/
[2] Biological Magnetic Resonance Data Bank (BMRB) - http://www.bmrb.wisc.edu/

فهرست مطالب

۱ مقدمه ۱۲

۱-۱ موضوع پایان نامه ۳

۱-۲ محدودیت های این پژوهش ٤

۱-۳ ساختار پایان نامه ٥

۲ پروتئینها وساختار دوم آنها ٦

۲-۱ مقدمه ۷

۲-۲ رده بندی اطلاعات ساختاری مولکولهای زیستی ۱۰

۲-۲-۱ ساختار اول ۱۰

۲-۲-۲ ساختار دوم ۱۱

۲-۲-۳ ساختار سوم ۱۸

۲-۲-۴ ساختار چهارم ۱۹

۲-۳ پیش بینی ساختارهای پروتئین ها ۲۰

۲-۳-۱ روش های مبتنی بر پایداری ترمودینامیکی ۲۱

۲-۳-۲ روش های مبتنی بر شیوه های آماری (مدلسازی تطبیقی) ۲۲

۲-۳-۳ مطالعه چو-فسمن ۲۳

۲-۳-۴ بررسی هم ترازی دنباله ها (الگوریتم اسمیت-واترمن) ۲٤

۲-۳-۵ استفاده از تکنیک های بازشناخت الگو ۲٦

۳ تئوری شواهد دمپستر-شافر و قوانین ترکیب ۲۸

۳-۱ تئوری شواهد دمپستر-شافر ۲۹

۳-۱-۱ چارچوب مشاهدات ۲۹

۳-۱-۲ تخصیص احتمال پایه ۲۹

۳-۱-۳ تابع باور ۳۰

۳-۱-۴ تابع محتمل بودن ۳۱

۳-۲ قواعد ترکیب شواهد ۳۱

۳-۲-۱ قاعده ترکیب شواهد دمپستر ۳۲

۳-۲-۲ قاعده ترکیب یاگر ۳۴

٤ طیف سنجی تشدید مغناطیسی هسته و جابجایی شیمیایی ۳٦

۴-۱ طیف سنجی تشدید مغناطیسی هسته ۳۷

۴-۲ پدیده تشدید مغناطیسی هسته ۳۸

۴-۳ جابجایی شیمیایی ٤۱

۴-۴ ارتباط جابجایی شیمیایی و ساختار دوم پروتئین ٤۲

۴-۵ مدل تحلیل ترکیبی انرژی پروتئین از دادههای جابجایی شیمیایی ٤٤

۵ ارائه روش ملهم از تئوری شواهد دمپستر-شافر برای پیش بینی ساختار دوم پروتئین ها ٤۵

۵-۱ مقدمه ٤٦

۵-۲ اهداف تحقیق ٤۷

۵-۳ الهام از تئوری شواهد دمپستر-شافر در پیش بینی ساختار دوم پروتئین ها ٤۸

۵-۴ آماده سازی داده ها ٤۹

۵-۴-۱ استخراج اطلاعات فایل PDB ٤۹

۵-۴-۲ استخراج اطلاعات فایل BMR ۵۰

۵-۵ آموزش ۵۰

۵-۵-۱ استخراج شواهد ۵۰

۵-۵-۲ تشکیل چارچوب ادراک پایه ۵۲

۵-۶ آزمون ۵۵

۵-۶-۱ آزمون شواهد تک مانده ای ۵۵

۵-۶-۲ آزمون شواهد چند مانده ای ۵۷

۶ پیاده سازی ها و نتایج طبقه بندی ها ۵۹

۶-۱ ساختار نرم افزار پیاده سازی شده ۶۰

۶-۲ بررسی حالت تک مانده ای ۶۱

۶-۳ استفاده از روش ساده سازی شده در امتیازدهی ۶۲

۶-۳-۱ نتایج حاصل از پایگاه PDB ۶۳

۶-۳-۲ نتایج حاصل از پایگاه DSSP ۶۶

۶-۴ بررسی حالت چند مانده ای ۶۸

۶-۴-۱ نتایج حاصل از پایگاه PDB ۶۹

۶-۴-۲ نتایج حاصل از پایگاه DSSP ۷۲

۶-۵ مقایسه و تحلیل ۷۴

۷ نتیجه گیری و پیشنهادها ۷۷

۷-۱ جمع بندی ۷۸

۷-۲ پیشنهاد کارهای آینده ۷۹

پیوست ۱: جابجایی های شیمیایی اندازه گیری شده برای آمینو اسید های مختلف ۸۰

واژه نامه (انگیسی- فارسی) ۸۹

واژه نامه (فارسی- انگیسی) ۹۲

اختصارات ۹۵

فهرست منابع ۹۶

فهرست شکل‌ها

شکل ۲-۱ برخی وظایف مهم پروتئین در موجودات زنده ۸
شکل ۲-۲ساختار سه بعدی پروتئین ۹
شکل ۲-۳ ساختار دوم یک مولکول RNA ۱۲
شکل ۲-۴ ساختار دوم یک مولکول پروتئین ۱۲
شکل ۲-۵ نمای سه بعدی یک مارپیچ آلفا ۱٤
شکل ۲-۶ نمای سه بعدی یک مارپیچ 3_{10} ۱۵
شکل ۲-۷ نمای سه بعدی یک مارپیچ پی ۱۵
شکل ۲-۸ نمایش جاگیری رشته ها در صفحات بتای موازی ۱٦
شکل ۲-۹ نمایش جاگیری رشته ها در صفحات بتای ضدموازی ۱۷
شکل ۲-۱۰ نمای یک خم گاما ۱۷
شکل ۲-۱۱ نمای یک خم بتا ۱۸
شکل ۲-۱۲ ساختار سوم یک مولکول پروتئین ۱۹
شکل ۲-۱۳ مقایسه ساختارهای اول تا چهارم پروتئینها ۲۰
شکل ۲-۱۴ مثالی از هم ترازی دو دنباله ۲۵
شکل ۲-۱۵ نمایی از یک شبکه عصبی طبقه بندی کننده ساختار دوم پروتئین ۲۷
شکل ۳-۱ شمایی از تعاریف تئوری شواهد ۳۱
شکل ۴-۱ قواعد دست راست و چپ از الکترومغناطیس کلاسیک ۳۸
شکل ۴-۲ میله مغناطیسی شبه هسته در میدان مغناطیسی ۳۹
شکل ۴-۳ نمایی از یک طیف سنج NMR ٤۱
شکل ۴-۴ ارتباط جابجایی شیمیایی H^{α} [1] با ساختار دوم ٤۳
شکل ۶-۱ دیاگرام کلاس های نرم افزار پیاده سازی شده ٦۰
شکل ۶-۲ نمونه ای از فایل پیکربندی ٦۱
شکل ۶-۳ مقایسه روش های KNN+MV و PW+OWA ٦٤

شکل ۶-۴ صحت پیش بینی با ضریب باور خم ۱ ..... ۶۵
شکل ۶-۵ صحت پیش بینی با ضریب باور خم ۲ ..... ۶۶
شکل ۶-۶ صحت پیش بینی با ضریب باور خم ۱۰ ..... ۶۶
شکل ۶-۷ تاثیر ضریب باور خم روی صحت نتایج پیش بینی ۶۶
شکل ۶-۸ مقایسه دو روش برای ساختارهای چهارگانه در DSSP برای دقت 10^{0} ۶۷
شکل ۶-۹ مقایسه دو روش برای ساختارهای چهارگانه در DSSP برای دقت 10^{-1} ۶۷
شکل ۶-۱۰ صحت پیش بینی با ضریب باور خم ۱ ..... ۶۸
شکل ۶-۱۱ صحت پیش بینی با ضریب باور خم ۲ ..... ۶۸
شکل ۶-۱۲ دقت پیش بینی با ضریب باور خم ۱۰ ..... ۶۸
شکل ۶-۱۳ صحت پیش بینی مارپیچ آلفا برای مقادیر مختلف S_T برای پایگاه PDB ۶۹
شکل ۶-۱۴ صحت پیش بینی صفحه بتا برای مقادیر مختلف S_T برای پایگاه PDB ۷۰
شکل ۶-۱۵ صحت پیش بینی مارپیچ تصادفی برای مقادیر مختلف S_T برای پایگاه PDB ۷۰
شکل ۶-۱۶ صحت پیش بینی خم برای مقادیر مختلف S_T برای پایگاه PDB ۷۰
شکل ۶-۱۷ صحت پیش بینی کلی برای مقادیر مختلف S_T برای پایگاه PDB ۷۱
شکل ۶-۱۸ صحت پیش بینی مارپیچ آلفا برای مقادیر مختلف S_T برای پایگاه DSSP ۷۲
شکل ۶-۱۹ صحت پیش بینی صفحه بتا برای مقادیر مختلف S_T برای پایگاه DSSP ۷۲
شکل ۶-۲۰ صحت پیش بینی مارپیچ تصادفی برای مقادیر مختلف S_T برای پایگاه DSSP ۷۳
شکل ۶-۲۱ صحت پیش بینی خم برای مقادیر مختلف S_T برای پایگاه DSSP ۷۳
شکل ۶-۲۲ صحت پیش بینی کلی برای مقادیر مختلف S_T برای پایگاه DSSP ۷۳

فهرست جدول‌ها

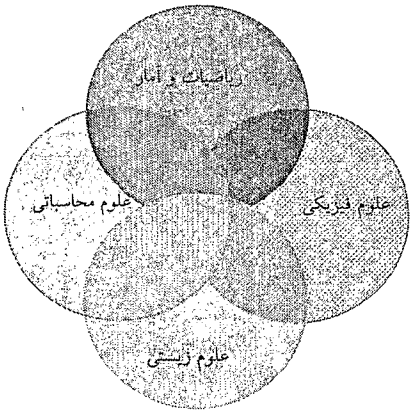
جدول ۲-۱ نام اسیدهای آمینه و مخفف های آنها ۱۰
جدول ۲-۲ مثال هایی از ابزارهای مبتنی بر مدلسازی تطبیقی ۲۳
جدول ۵-۱ ساختارهای هشت گانه پروتئین ها ٤٨
جدول ۵-۲ ماتریس جانشینی BLOSUM62 ..... ٥٦
جدول ۵-۳ شبه برنامه روش پیشنهادی ٥٧
جدول ۶-۱ پارامترهای متغیر آزمون ٦٢
جدول ۶-۲ پارامترهای امتیازدهی در حالت ساده سازی شده ٦٣
جدول ۶-۳ تعداد مانده های موجود در هر ساختار ٦٣
جدول ۶-۴ مقایسه نتایج طبقه بندی ٧٦

۱ مقدمه

بیوانفورماتیک[1] به ایجاد و توسعه الگوریتم‌ها، راهکارهای محاسباتی و آماری، و تئوری برای حل مسائل صوری و عملی ارائه‌شده و یا ملهم از مدیریت و تحلیل داده‌های زیست‌شناختی اطلاق می‌گردد. به عبارت دیگر بیوانفورماتیک علم استخراج دانش از منابع داده‌ای زیست‌شناختی به کمک تحلیل‌های کامپیوتری و حل مسایل زیست‌شناختی با استفاده از روش‌های مجاسباتی و ارائه الگوریتم‌ها و روش‌های عملی و کارآ برای بررسی و نتیجه‌گیری از داده‌های زیستی است.

حجم بالای داده‌های بیوانفورماتیک از یک سو و عدم وجود مدل‌های دقیق که این سیستم‌ها را (با توجه به پیچیدگی بسیار زیاد آنها) مدل نمایند از سوی دیگر، باعث شده است که شناسایی فرآیندها و اندرکنش‌های زیست‌شناختی مولکولی جزو مسایل روز و مورد بررسی وتحقیق فراوان باشند. به عنوان مقیاسی از حجم داده‌ها در این حوزه میتوان به طول ژنوم انسان (۳،۲ میلیون جفت پایه[2])، تعداد ژن‌های موجود (۳۲۰۰۰) و طول هر ژن (۲۷۰۰۰ جفت پایه) اشاره نمود. در این میان استفاده از روش‌های نوین آماری برای استخراج الگوهای رفتاری با استفاده از داده‌های آزمایشگاهی موجود برای شناخت و پیش-بینی فرآیندهای طبیعی با اقبال قابل ملاحظه‌ای مواجه است. همچنین افزایش روبه رشد حجم داده‌های استخراج شده در نتیجه ابداع روش‌های جدید استخراج داده‌های زیست‌شناختی منابع عظیمی از دانش را در اختیار قرار داده‌اند که استخراج این دانش روش‌های کارآمد آماری را طلب می‌کند. هدف از مطالعه بیوانفورماتیک، نه تعیین قوانین ریاضی حاکم بر فیزیک سیستم‌های زیست‌شناختی، بلکه ایجاد ابزار مورد نیاز برای تحلیل داده‌های زیست‌شناختی است. بعنوان مثال یک زیست‌شناس نیازمند ابزاری برای تشخیص ژن‌ها در *DNA* و نیز برای تخمین تفاوت میان بیان ژن‌های گوناگون در بافت‌های مختلف است. تحلیل‌هایی از این دست نیازمند ارائه‌ی مدل‌های آماری و احتمالاتی است که با توجه به طبیعت داده‌ها و نیز با تکیه بر دانش موجود بتوانند زیست‌شناس را در تشخیص یاری دهند.

[1] Bioinformatics
[2] Base Pair



در ایجاد داده‌های مورد استفاده در مطالعات بیوانفورماتیک سازوکارهای اتفاقی زیادی مانند فرآیندهای تصادفی دخیل در تکامل طبیعی و تصادفی‌بودن ذاتی فرآیندهای نمونه‌برداری تاثیرگذار هستند. تئوری فرآیندهای تصادفی[1] سعی در توضیح چگونگی تغییرات فرآیندهای تصادفی در زمان یا مکان دارد. بدین ترتیب استفاده از این تئوری در مورد فرآیند تکامل طبیعی که از پیچیده‌ترین فرآیندهای تصادفی شناخته‌شده است، جایگاه خاصی دارد.

داده‌های بیوانفورماتیک دارای طبیعتی ناهمگون هستند؛ بدان معنا که یک مفهوم واحد در آنها در قالب‌های مختلفی بروز می‌کند. از جمله‌ی این قالب‌ها می‌توان به داده‌های رقمی، داده‌های رشته‌ای و تصاویر (دوبعدی و سه‌بعدی) اشاره نمود. این طبیعت ناهمگون استفاده از انباره‌های داده‌های زیستی را از لحاظ برخورداری از یک دید واحد روی مفاهیم و شناسایی اندرکنش آن‌ها را با مشکل مواجه کرده‌است.

۱-۱ موضوع پایان نامه

در این پایان‌نامه از تئوری شواهد دمپستر-شافر بعنوان یک روش مدلسازی نایقینی بمنظور ارائه راه‌حلی برای مساله‌ی پیش‌بینی ساختار دوم پروتئین‌ها استفاده شده است. از این رو ابتدا مساله‌ی مورد اشاره با استفاده از تئوری شواهد مدلسازی شده و ارکان اصلی مدل که عبارت از قالب شواهد، نحوه

[1] Stochastic Process Theory

تخصیص احتمال پایه و قانون ترکیب شواهد هستند، تعیین گشته‌اند. بر مبنای این مدل شواهد موجود در داده‌های مقصد[1] استخراج (و غنی‌سازی[2]) شده‌اند. بدین اعتبار از آنجا که داده‌ها مورد استفاده از بیش از یک منبع داده‌ای ناهمگن مورد استفاده قرار گرفته‌اند، این روش را می‌توان نوعی ترکیب داده‌ها با استفاده از روشی ملهم تئوری شواهد در نظر گرفت.

پس از تعیین بدنه‌ی شواهد مرجع[3] از روی داده‌های ورودی آموزش، بخش آزمون بر روی شواهد استخراج‌شده از داده‌های آزمون انجام پذیرفته است. برای این منظور ابتدا سنجه‌ای برای ارزیابی میزان شباهت شواهد تعیین شده است که دارای پارامترهای قابل تغییر است. سپس با استفاده از سنجه پیش‌گفته و بر مبنای بدنه شواهد مرجع برای هر شهود آزمون، مجموعه‌ای از شواهد مرجع که بیشترین شباهت را شهود آزمون دارا هستند، تعیین گشته و نتیجه‌گیری از روی این شواهد جهت تصمیم‌سازی درباره شهود آزمون با استفاده از چند روش مرسوم ارزیابی گشته است. عمده‌ی این روش‌ها روش رای‌گیری بیشینه[4] و میانگین‌گیری مرتب وزن‌دار[5] بر مبنای پنجره‌ی پارزن[6] یا نزدیکترین همسایه‌ها[7] بوده است.

۱-۲ محدودیت های این پژوهش

محدودیت‌های موجود در این پژوهش عبارتند از:

- تئوری شواهد دمپستر-شافر بعنوان منبع الهام برای ارائه راه‌حل مورد استفاده قرار گرفته و گسترش‌های مطالعه‌شده این تئوری برای اعمال در این مسالهِ مناسب تشخیص داده نشده‌اند. به عمین جهت این گسترش‌ها مورد واکاوی بیشتر قرار نگرفته‌اند.
- داده‌های مورد استفاده از پایگاه داده‌های تشدید مغناطیسی زیستی استخراج شده‌اند و در

[1] Target Data
[2] Enrich
[3] Reference Body of Evidence
[4] Majority Voting
[5] Ordered Weighted Averaging (OWA)
[6] Parzen Window
[7] (K) Nearest Neighbors

نتیجه تمام داده‌های پروتئین‌های موجود را شامل نمی‌شوند.

۱-۳ ساختار پایان نامه

مقدمات این پایان‌نامه در فصل‌های ابتدایی مورد بررسی اجمالی قرار گرفته‌اند. در فصل دوم ساختار پروتئین‌ها به عنوان دسته مهمی از عناصر زیستی تشکیل‌دهنده ارگانیزم‌های حیاتی تبیین شده است. تئوری شواهد دمپستر-شافر و انواع قواعد ترکیب بعنوان راهبرد مورد نظر برای حل مساله مورد بحث این پایان‌نامه در فصل سوم توضیح داده شده است. طیف‌سنجی مغناطیسی هسته و جابجایی شیمیایی در فصل چهارم به اختصار مورد بررسی واقع شده‌اند. مدل پیشنهادی ملهم از تئوری شواهد در فصل پنجم واکاوی شده است و نتایج حاصل از آن در فصل ششم تشریح شده است. در انتها نتیجه‌های حاصل از این تحقیق و پیشنهادهایی برای کارهای آتی ارائه شده‌اند.

۲ پروتئین‌ها وساختار دوم آن‌ها

۲-۱ مقدمه

در قرن هجدهم آنتوان فورکروی[1] توانست مولکولهایی را در بدن موجودات زنده کشف کند که از لحاظ خواص شیمیایی، قدری متفاوت با دیگر مولکولها بوده و می‌توانستند دسته مجزایی را تشکیل دهند [۱]. این مولکولها برای اولین بار در موادی مانند سفیدهٔ تخم‌مرغ و خون یافت شدند. در سال ۱۸۳۸ به این دسته از مولکول‌ها نام پروتئین دادند و سپس کشف کردند که این مولکولها از واحدهای کوچکتری به نام اسیدهای آمینه تشکیل شده‌اند [۱]. اما تا سال ۱۹۲۶ به نقش مهم این مولکولها در بدن یک موجود زنده پی نبرده بودند. در آن سال نشان داده شد که آنزیم *Urease* یک پروتئین است و از آنجا که وظایف مهم آنزیم‌ها در بدن، در آن موقع شناخته شده بود، پروتئین‌ها در رأس بسیاری از تحقیقات زیست شناسی قرار گرفتند [۱]. دانشمندان به سراغ استخراج کردن پروتئین‌ها از مواد مختلف رفتند و در پی کشف ساختار فیزیکی و واکنش‌های مختلف و وظایف گوناگون این مواد، در بدن بودند. بعدها که به وسیله‌ی اشعه‌ی X در کریستالوگرافی، ساختار برخی پروتئین‌ها کشف شد و ساختار پلیمری و خطی‌بودن این مواد و توالی دقیق واحدهای مونومری آنها (که همان اسیدهای آمینه هستند) مشخص شد، مسائل دیگری از جمله کشف ارتباط بین توالی اسیدآمینه‌ها و ساختار دو بعدی و سه بعدی پروتئین برای زیست شناسان مطرح شدند [۲]، [۳].

امروزه زیست شناسان می‌دانند که مثلاً در بدن انسان ده‌ها هزار نوع پروتئین وجود دارد [۴] که هر کدام وظایف خاصی را بر عهده دارند که برخی از این وظایف، در شکل ۲-۱ نشان داده شده اند. این عملکرد پروتئینها در بدن، با ساختار دو بعدی و سه بعدی و شکل فضایی آنها ارتباط دارد و این ساختار سه بعدی نیز متأثر از توالی اسیدهای آمینه‌ای است که پروتئین را می‌سازند [۵] (در شکل ۲-۲، ساختار

[1] Antoine Fourcroy