مجله کنترل
ISSN 2008-8345
جلد ۷، شماره ۴، زمستان ۱۳۹۲، صفحه ۶۲-۵۳

پیش بینی تاخوردگی دامنه پروتئین‌ها مبتنی بر روش DBSCAN

محمدرضا محمدیان[۱]، محمدتقی حمیدی بهشتی[۲] ، کاوه کاوسی[۳]

[۱]دانشجوی کارشناسی ارشد، دانشگاه تربیت مدرس، دانشکده مهندسی برق و کامپیوتر، گروه کنترل، m.mohamadian@modares.ac.ir
[۲]دانشیار، دانشگاه تربیت مدرس، دانشکده مهندسی برق و کامپیوتر، گروه کنترل، mbehesht@modares.ac.ir
[۳]استادیار، دانشگاه تهران، دانشکده تحقیقات بیوشیمی-بیوفیزیک، گروه کنترل، kkavousi@alumni.ut.ac.ir

(تاریخ دریافت مقاله ۱۳۹۲/۹/۱۶، تاریخ پذیرش مقاله ۱۳۹۲/۱۲/۵)

چکیده: در این مقاله یک روش طبقه‌بندی تغییر یافته بر مبنای روش‌های طبقه‌بندی بر اساس چگالی برای طبقه‌بندی تاخوردگی پروتئین‌ها ارائه شده است که این روش در برابر وجود نویز مقاوم بوده و از سرعت بالایی برخوردار خواهد بود. طبقه‌بندی پروتئین‌ها بمنظور پیش‌بینی عملکرد آنها و شناسایی خواص پروتئین‌ها یکی از مسائل بزرگ در حوزه طبقه‌بندی است. با توجه به پیشرفت علم و دستگاه‌های توالی یابی پروتئین‌های بسیاری کشف شده‌اند که نیازمند یک سیستم طبقه‌بندی خودکار هستند. این مقاله روش طبقه‌بندی خودکاری را برای طبقه‌بندی پروتئین‌ها براساس تاخوردگی و اطلاعات مستخرج از توالی ارائه خواهد نمود. همچنین در این مقاله از روش ترکیب اطلاعات فازی برای ترکیب اطلاعات طبقه‌بندی کننده‌ها جهت افزایش قطعیت عمل طبقه‌بندی استفاده شده است.

کلمات کلیدی: طبقه‌بندی تاخوردگی پروتئین، طبقه‌بندی براساس حذف نویز، طبقه‌بندی براساس چگالی

Protein Domain Folding Prediction Based on DBSCAN

Mohammad Reza Mohammadian, Mohammad Taghi Hamidi Beheshti, Kaveh Kavousi

Abstract: This paper presents a density-based clustering approach for data classification of protein folding. The method is shown to perform better as compared with the conventional methods with respect to computational speed and robustness against noise. Protein clustering is known to be one of the important challenges of prediction and identification of protein's properties and its performances. Due to recent advances in sequence detection devices, many proteins have been discovered which requires an automatic protein clustering system. An automatic protein clustering method is thus presented here which is based on folding and information extracted from the output features of the sequence. Further, fuzzy data fusion is used to combine the clustered information in order to improve the clustering performance.

Keywords: Density-Based Clustering, Protein-Fold Classification, Robust Clustering

۱- مقدمه

در سیستم‌های بیولوژی طبقه‌بندی ساختاری پروتئین‌ها به قصد پیشگویی عملکرد آنها یکی از چالش‌های بزرگ و مسائل مطرح شده می‌باشد.

در این زمینه طبقه‌بندی تاخوردگی پروتئین‌ها مساله‌ای است که در سالهای اخیر بسیار به آن پرداخته شده است. چرا که در علوم زیست‌شناسی پروتئین‌های بسیاری کشف شده‌اند که این پروتئین‌ها نیازمند یک روش طبقه‌بندی و شناسایی خودکار می‌باشند.

پیش‌بینی تاخوردگی پروتئین‌ها همواره با عدم قطعیت همراه است و این عدم قطعیت‌ها به دلیل نبود تعریف روشن و بدون ابهام از الگوی تاخوردگی پروتئین‌ها است. طبقه‌بندی تاخوردگی پروتئین‌ها حتی با استفاده از ساختمان سوم پروتئین‌ها کاری دشوار است. در حالی که برای طبقه‌بندی و شناسایی پروتئین‌ها تنها از اطلاعات بدست آمده از توالی پروتئین‌ها استفاده می‌شود.

در طبقه‌بندی پروتئین‌ها دو حالت را می‌توان در نظر گرفت، حالتی که برای طبقه‌بندی به ساختار سوم پروتئین نیاز است و حالتی که بر مبنای اطلاعات استخراج شده از توالی پروتئین‌ها کار طبقه‌بندی انجام می‌شود.

روش‌های طبقه‌بندی پروتئین‌ها را می‌توان به سه دسته کاملاً دستی، نیمه خودکار و خودکار دسته‌بندی کرد که در این روش‌ها روش کاملاً دستی بر مبنای نظر افراد خبره و دانشمندان حوزه بیولوژی انجام می‌شود روش طبقه‌بندی SCOP[1] یک روش سلسله مراتبی از این نوع روش‌ها است[۱]. روش نیمه خودکار هم بر مبنای نظر افراد خبره و هم به صورت خودکار طبقه‌بندی را انجام می‌دهد روش طبقه‌بندی CATH[2] یک روش نیمه خودکار است[۲] روش خودکار بدون در نظر گرفتن نظر افراد خبره و به صورت کاملاً خودکار و بدون سرپرست طبقه‌بندی را انجام می‌دهد.

روش‌های طبقه‌بندی پروتئین‌ها هر یک براساس یکی از ویژگی‌های پروتئین‌ها عمل می‌کنند که این ویژگی‌ها می‌تواند ساختارهای چهارگانه پروتئین، عملکرد پروتئین، زنجیره تکاملی و یا معیارهای دیگر باشد.

در [۳, ۴] اولین تحقیقات شناسایی ساختار سوم پروتئین‌ها از روی توالی آنها و با استفاده از اصل کمترین انرژی آزاد انجام گرفت. این روش‌ها اگرچه در پاره‌ای موارد نتایج درخشانی داشته‌اند اما در حالت کلی از معضل کمینه‌ی محلی رنج می‌بردند و به عنوان روشی قابل اعتماد برای تعیین ساختار سوم پروتئین از روی توالی قابل استفاده نیستند.

در [۵] همترازی دوبه دوی توالی‌ها در مورد دامنه‌هایی که رابطه‌ی تکاملی نزدیکی با یکدیگر دارند نتایج مطلوبی گرفته شده است.

یک روش برای طبقه‌بندی پروتئین‌ها نگاشت مسئله به یک مسئله یادگیری ماشینی با سرپرست است. این روش‌ها نیازمند یک استاندارد طلائی هستند تا بتوانند تا حد ممکن خود را با استاندارد مورد استفاده هماهنگ کنند. در این روش‌ها پروتئین ناشناخته براساس استاندارد مورد نظر و آنچه طبقه‌بندی کننده از آن می‌آموزد باید به کلاس الگوی تاخوردگی صحیح و یا کلاس ساختار دوم مناسب منتسب شود. در [۶] لیستی از روش‌های یادگیری ماشینی را می‌توان مشاهده کرد.

تعیین صحیح کلاس ساختار دوم دامنه پروتئین اولین گام در ایجاد سیستم‌های طبقه‌بندی خودکار پروتئین می‌باشد. در [۷] نیشیکاوا[3] و همکاران ارتباط قوی بین ویژگی مستخرج از ترکیب اسیدهای آمینه و کلاس ساختار دوم را نشان دادند.

تعدادی از روش‌ها برای پیش‌بینی ساختار پروتئین‌ها از ایده‌های فازی استفاده می‌کنند. که در آنها به جای تخصیص یک ساختار قطعی به یک مانده اندازه‌ی فازی تعلق مانده‌ی مورد نظر به هر یک از کلاس‌های ساختاری تخمین زده می‌شود. در [۸] سه اندازه‌ی مهم برای ارزیابی کارآیی الگوریتم‌های پیشگوئی ساختار دوم پروتئین معرفی شده‌اند. اندازه‌های سنتی امتیاز Q، اندازه‌ی SOV[4] و ضرایب همبستگی k-حالته[5] (Corr) به ترتیب به سه اندازه‌ی فازی امتیاز F، FOV[6] و ضرایب همبستگی فازی[7] (Forr) تعمیم داده شدند.

استفاده از مدل مخفی مارکوف (HMM[8]) با تعداد حالات کاهش یافته با قابلیت یادگیری برای پیشگویی کلاس ساختار دوم پروتئین و نیز طبقه‌بندی الگوی تاخوردگی پروتئین در [۹] گزارش شده است.

دوبچک[9] و همکارانش یک روش مبتنی بر استفاده از توصیف‌گرهای سراسری زنجیره‌ی پروتئین و یک فرآیند رای‌گیری را برای طبقه‌بندی الگوی تاخوردگی دامنه پروتئین پیشنهاد کردند[۱۰]. آنها از یک شبکه عصبی مصنوعی برای اتخاذ توصیف‌گرها بر پایه‌ی چگونگی ترکیب[10]، گذار[11] و توزیع[12] برای ویژگی‌های مختلف اسیدهای آمینه مانند آبگریزی، ساختار دوم پیشگویی شده و ... را مورد استفاده قرار دادند. همچنین ایده‌ی خود را با استفاده از مفهوم توصیف‌گرهای

[1] Structural Classification Of Proteins
[2] Class- Architecture- Topology- Homologous

[3] Nishikawa
[4] Segment OVerlap measure
[5] k-state Correlation Coeficient
[6] Fuzzy OVerlap measure
[7] Fuzzy Correlation Coeficient
[8] Hidden Markov Models
[9] Dubchak
[10] Composition
[11] Transition
[12] Distribution

سراسری مبتنی بر خواص فیزیکی، شیمیایی و ساختاری اسیدهای آمینه-ی تشکیل دهنده‌ی ساختار اول پروتئین پیگیری کردند[۱۱].

دینگ۱ و دوبچک از روش های منحصر به فرد "یک در مقابل دیگران۲" و "همه در برابر همه۳" به عنوان جدا کننده برای پیش بینی کلاس پروتئین جهت کاهش میزان خطا در مقایسه با روش‌های عادی "یک در مقابل دیگران" استفاده کردند. و از ماشین بردار پشتیبان SVM۴ و شبکه های عصبی به عنوان طبقه‌بندی کننده های پایه استفاده کردند[۱۲].

در [۱۳] از روش طبقه‌بندی کننده پایه‌ای OET-KNN۵ [۱۴, ۱۵] استفاده می‌شود. این روش یک طبقه‌بندی کننده با N طبقه را در نظر می‌گیرد و k نزدیک‌ترین همسایگی را بدست آورده و با استفاده از قاعده ترکیب دمپستر شافر[۱۶-۱۸] آن را طبقه‌بندی می‌کند.

در [۱۹] مفاهیم فولدهای درهم تنیده و ابر فولدها مطرح شده است. که نوعی از طبقه‌بندی کننده را براساس تشکیل ساختارهای اعتقادی مبتنی بر هر یک از ویژگی‌ها، در تصمیم‌گیری هر دامنه‌ی پروتئینی به جای تخصیص به یک طبقه‌ی مشخص، به یک ابر فولد تخصیص می‌دهد و سپس از طریق ترکیب این ساختارهای اعتقادی و پالایش نتایج در نهایت دامنه‌ی ناشناخته، حتی الامکان به یک طبقه‌ی الگوی تاخوردگی تخصیص می‌یابد و در غیر این صورت یک ابر فولد کاندید مقصد نهایی خواهد شد.

۲- داده‌ها و روش مورد استفاده

در این مقاله یک طبقه‌بندی کننده خودکار برای پیش‌بینی تاخوردگی پروتئین‌ها با در دست داشتن اطلاعات مستخرج از توالی پروتئین‌ها بیان می‌شود. طبقه‌بندی بدون استفاده از نظر افراد خبره و بصورت خودکار کار طبقه‌بندی را انجام می‌دهد. که از روش طبقه‌بندی تغییر یافته DBSCAN۶ استفاده شده است. برای طبقه‌بندی پروتئین‌ها دو معیار شباهت و عدم شباهت در نظر گرفته شده که عمل طبقه‌بندی براساس این دو معیار عمل خواهند کرد معیار عدم شباهت می‌تواند در جداسازی انواع تاخوردگی‌هایی که از یکدیگر فاصله دارند مفید واقع شود. چرا که می‌توان دو پروتئینی که به مقدار کافی از یکدیگر فاصله دارند را به طور یقین از هم جدا کرد و در تاخوردگی‌های متفاوتی قرار داد. برای حذف عدم قطعیت‌های موجود از روش ترکیب اطلاعات انتگرال فازی سوگنو برای ترکیب اطلاعات بدست آمده استفاده می-شود.

۲-۱- داده‌های مورد استفاده

مجموعه داده‌های مورد استفاده در این مقاله داده‌های مستخرج از توالی پروتئین‌ها است که از بانک اطلاعاتی مربوط به پژوهش دینگ و دوبچک [20 ,12] تامین شده است.

مجموعه داده‌های اولیه آموزش و تست هر یک به ترتیب شامل ۳۱۱ و ۳۸۳ پروتئین می‌باشند.

طبق گزارش دینگ و دوبچک برای پروتئین‌های با طول زنجیره‌ی ۸۰ و یا بالاتر هیچ پروتئینی در مجموعه آموزش از لحاظ توالی بیش از ۳۵ درصد همانندی با سایر پروتئین‌های مجموعه آموزش ندارد. براساس طبقه‌بندی SCOP مجموعه داده‌های آموزش و تست در همان مرجع به ۲۷ الگوی تاخوردگی پروتئینی مختلف تقسیم شده‌اند.

این طبقات تاخوردگی در ردیف پر جمعیت‌ترین طبقات تاخوردگی SCOP (با حداقل ۷ عضو در هر طبقه‌ی الگوی تاخوردگی) قرار دارند و از میان چهار کلاس ساختاری اصلی انتخاب شده‌اند.

مجموعه داده‌های آزمون شامل دامنه‌هایی از SCOP است که کمتر از ۴۰ درصد مشابهت با یکدیگر دارند. هیچ یک از دامنه‌های مجموعه داده‌های آموزش و آزمون بیش از ۳۵ درصد مشابهت با یکدیگر ندارند و بیش از ۹۰ درصد دامنه‌های مجموعه آزمون کمتر از ۲۵ درصد مشابهت توالی با داده‌های مجموعه آموزش دارند.

جدول ۱ طبقات داده‌های مورد استفاده و نیز کلاس‌های داده به همراه تعداد داده‌های مورد استفاده در قسمت آموزش و تست را نشان می‌دهد[۲۱].

جدول۱: تاخوردگی و کلاس داده‌های آموزش و تست مورد استفاده

Fold no	Fold Name	Class	Train Sample	Test Sample
1	Globin-like	All α	13	**6**
2	Cytochrome c	All α	7	**9**
3	DNAbinding3-helical bundle	All α	12	**20**
4	4-helical up-and-down bundle	All α	7	**8**
5	4-helical cytokines EF-hand	All α	9	**9**
6	EF-hand	All α	6	**9**
7	Immunoglobulin-like	All β	30	**44**
8	cupredoxins	All β	9	**12**
9	Viral coat and capsid protein	All β	16	**13**
10	Cona-like lectin/glucanases	All β	7	**6**
11	SH3-like barrel	All β	8	**8**
12	OB-fold	All β	13	**19**
13	Beta-trefoil	All β	8	**4**
14	Trypsin-like serine proteases	All β	9	**4**
15	lipocalins	All β	9	**7**
16	(TIM)-barrel	α,β	29	**48**
17	FAD(also NAD)-binding motif	α,β	11	**12**
18	Flavodoxin-like	α,β	11	**13**
19	NAD(P)-binding rossmann-fold	α,β	13	**27**
20	P-loop	α,β	10	**12**
21	Thioredoxin-like	α,β	9	**8**
22	ribonuclease H-like motif	α,β	10	**12**
23	Hydrolases	α,β	11	**7**
24	Periplasmic binding protein-like	α,β	11	**4**
25	Grasp β	α,β	7	**8**
26	Ferredoxin-like	α,β	13	**27**
27	Small inhibitors,toxins,lectins	α,β	13	**27**

در جدول ۲، ۱۰ ویژگی متفاوت جهت بررسی طبقه‌بندی تاخوردگی پروتئین‌ها نشان داده شده است. این ویژگی‌ها برای داده‌های تست و آزمون مورد استفاده قرار می‌گیرند که در مقابل هر ویژگی ابعاد

1 Ding
2 one-against-others
3 All against-all
4 Support vector machine
5 Optimized Evidence Theoretic – K-Nearest Neighbor
6 Density-Based Spatial Clustering of Applications with Noise

آن نیز قرار دارد.

جدول ۲: ویژگی‌ها و ابعاد ویژگی‌های مورد استفاده در طبقه‌بندی تغییر یافته[۱۳]

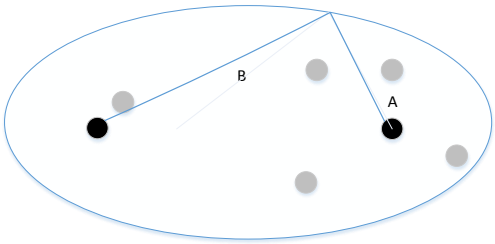
NO.	Feature	Dimension
1	Hydrophobicity	**21**
2	Predicted secondary structure	**21**
3	Normalized van der waals volume	**21**
4	Polarity	**21**
5	Polarizability	**21**
6	Pseudo amino acid composition(λ=1)	**22**
7	Pseudo amino acid composition(λ=4)	**28**
8	Pseudo amino acid composition(λ=14)	**48**
9	Pseudo amino acid composition(λ=30)	**80**
10	PSSM representative	**20**

۲-۲- روش طبقه‌بندی تغییر یافته DBSCAN

فرض می‌شود مجموعه‌ای $\{P_i\}_{i=1}^{N}$ که یک مجموعه‌ای N عضویی از پروتئین‌های ناشناخته است وجود دارد که باید این مجموعه را بدرستی طبقه‌بندی کرد. همچنین در این مجموعه پروتئین یک مجموعه $\{F_j\}_{j=1}^{M}$ وجود دارد که مجموعه تاخوردگی‌های پروتئین-های موجود خواهد بود. که M تعداد انواع تاخوردگی پروتئین‌ها است. روش طبقه‌بندی باید پروتئین‌های مجموعه قبل را به این M نوع تاخوردگی افراز کند.

برای طبقه‌بندی تاخوردگی پروتئین‌ها از طبقه‌بندی‌های مبتنی بر توزیع چگالی استفاده شده است. در این طبقه‌بندی کننده‌ها با استفاده از این خاصیت که توزیع چگالی بردارهای ویژگی هر طبقه در یک ناحیه از فضا متمرکز است و هر طبقه با استفاده از ناحیه‌های با چگالی توزیع کم از دیگر طبقه‌ها جدا می شود، طبقه‌بندی انجام خواهد شد[۲۲]. چگالی را می‌توان به روش‌های متفاوتی تعریف کرد. در این مقاله روش تعریف چگالی براساس بیضی ارائه شده است.

چگالی هر نقطه در فضای دو بعدی برابر است با تعداد نقاط موجود در یک بیضی که مجموع فواصل این نقاط از مراکز بیضی یک مقدار مثبت ε است. نمایشی از روش بهینه پایه‌ی مرکز در شکل ۱ نیز نشان داده شده است.



شکل ۱: روش بیضی برای تعریف چگالی هر نقطه

بنابراین طبقه‌بندی دامنه پروتئین‌ها با این روش تابعی از انتخاب صحیح مقدار ε است. در این طبقه‌بندی دامنه پروتئین‌های موجود را می‌توان به سه دسته تقسیم بندی کرد.

پروتئین‌های مرکزی: پروتئین‌های مرکزی پروتئین‌های هستند که در مرکز بیضی قرار می‌گیرند و در واقع هسته اصلی یک طبقه را تشکیل می‌دهند. چگالی پروتئین‌های مرکزی باید از یک مقدار MinPts بزرگتر باشد. MinPts کمترین تعداد نقاط موجود در حداقل همسایگی یک نقطه است.

پروتئین‌های حاشیه‌ای: پروتئین‌های حاشیه‌ای پروتئین‌های هستند که در مرز طبقه‌ها قرار می‌گیرد و چگالی این پروتئین‌ها کمتر از مقدار MinPts است ولی این نقاط در همسایگی یک پروتئین مرکزی قرار دارند. همسایگی یک پروتئین به تمام پروتئین‌هایی گفته می‌شود که در مجموع فاصله ε از پروتئین‌های مرکزی قرار می‌گیرند. همسایگی پروتئین p را توسط رابطه زیر می‌توان بیان کرد.

(۱)

به طوری که در رابطه قبل D نشان دهنده تمام پروتئین‌های آموزش است.

همچنین در این روش می‌توان پروتئین‌هایی را که نه در نقطه همسایگی قرار دارند و نه جزو نقاط مرکزی هستند را برای طبقه مورد نظر نویز در نظر گرفت.

الگوریتم DBSCAN از الگوریتم‌های مبتنی بر توزیع چگالی است. برای توضیح این الگوریتم ابتدا یک سری مفاهیم و لم بر پایه توزیع چگالی تعریف می‌شود.

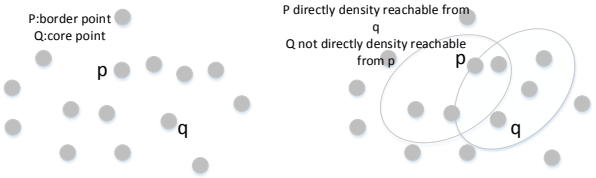
تعریف قابلیت دسترسی مستقیم چگالی: نقطه p را قابل دسترسی مستقیم چگالی[1] از نقطه q می‌گوییم هرگاه دو شرط زیر برقرار باشد.

1- $p \in N_\varepsilon(q)$
2- $|N_\varepsilon(p)| > MinPts$

خاصیت قابلیت دسترسی مستقیم چگالی دارای تقارن نیست. این بدین معنی است که ممکن است نقطه p قابل دسترسی مستقیم از نقطه q باشد ولی نقطه q قابل دسترسی مستقیم از نقطه p نباشد. عدم تقارن این

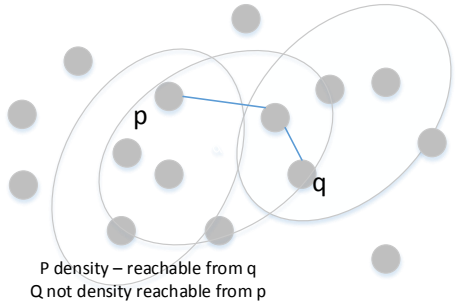
[1] Directly density-reachable

خاصیت در شکل ۲ نیز نشان داده شده است.



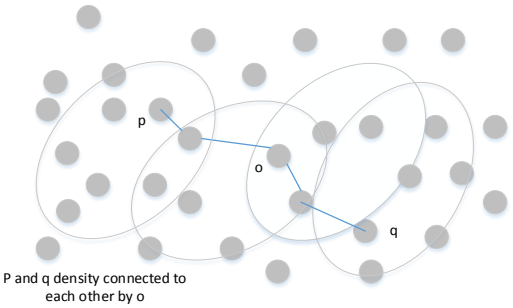
شکل ۲: خاصیت عدم تقارن قابلیت دسترسی مستقیم چگالی

تعریف قابلیت دسترسی چگالی: نقطه p قابل دسترسی چگالی[1] از نقطه q است هرگاه یک زنجیره از نقاط $p_1 = q, p_2, \dots, p_n = p$ وجود داشته باشد به قسمی که p_{i+1} قابل دسترسی مستقیم چگالی از p_i باشد. این خاصیت نیز دارای عدم تقارن است. عدم تقارن این خاصیت در شکل ۳ نشان داده شده است.



شکل ۳: خاصیت عدم تقارن قابلیت دسترسی چگالی

تعریف اتصال چگالی: دو نقطه p و q متصل چگالی[2] هستند هرگاه یک نقطه o وجود داشته باشد به گونه‌ای که p و q هر دو قابل دسترسی چگالی از نقطه o باشند. این خاصیت دارای تقارن است. نمایش این خاصیت در شکل ۴ ارائه شده است.



شکل ۴: نقطات متصل چگالی

تعریف طبقه: طبقه C با پارامترهای ε و MinPts یک زیرمجموعه ناتهی از داده‌ها است که دو خاصیت زیر در آن برقرار باشد.

۱- به ازای هر p،q و r، اگر $p, q \in C$ باشد و r قابل دسترسی چگالی با پارامترهای ε و MinPts از نقطه p یا q باشد آنگاه r نیز عضو طبقه C خواهد بود.

۲- به ازای هر p،q و r عضو C، p،q و r متصل چگالی باشند.

تعریف نویز: با فرض اینکه $C_1, C_2, \dots, C_k$ طبقه های موجود در داده‌های D با پارامترهای ε_i و $MinPts_i$ به ازای $i = 1,2, \dots, k$ باشند آنگاه نویز به نقاطی گفته می‌شود که به هیچکدام از طبقه‌ها متعلق نباشد.

$$noise = [p \in D \mid \forall i: p \notin (C_i]] \quad (۲)$$

در ادامه برای مشخص کردن طبقه‌ها چند لم ارائه شده است.

لم۱: در صورتی که نقطه p در داده‌های D دارای $|N_\varepsilon(p)| > MinPts$ باشد آنگاه تمام نقاط مجموعه‌ای که قابل دسترسی چگالی از نقطه p (با پارامترهای $MinPts$ و ε) هستند یک طبقه را تشکیل می-دهند.

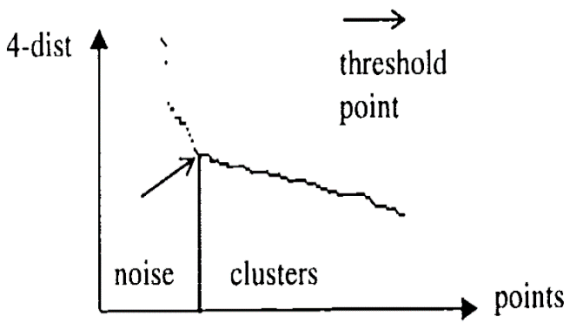
لم۲: در صورتی که C یک طبقه با پارامترهای $MinPts$ و ε باشد و p یک نقطه در آن طبقه با $|N_\varepsilon(p)| > MinPts$ باشد آنگاه طبقه C برابر با مجموعه تمام نقاطی است که قابل دسترسی چگالی از نقطه p هستند.

۲-۲-۱- تعیین ε و MinPts

در روش DBSCAN برای بدست آوردن کارآیی قابل قبول، باید دو پارامتر ε و $MinPts$ به طور مناسب تعیین شوند. یک روش که به طور معمول برای تعیین مناسب دو پارامتر ε و $MinPts$ به کار می‌رود بر اساس فاصله k امین نزدیکترین همسایه نقاط موجود در داده-های آموزش است[۲۳, ۲۴]؛ به فاصله k امین نزدیکترین همسایه هر نقطه $k - dist$ آن نقطه گفته می‌شود. در صورتی که نقاط انتخابی برای محاسبه $k - dist$ متعلق به طبقه‌ها باشند، و مقدار k بزرگتر از سایر طبقه‌ها نباشد و همچنین چگالی طبقه مورد نظر تفاوت زیادی با دیگر طبقه‌ها نداشته باشد آنگاه رنج تغییرات $k - dist$ کم خواهد بود. در حالیکه برای نقاط نویز که در قسمت قبل اشاره شد $k - dist$ مقدار بزرگی است. در صورتی که مقدار $k - dist$ برای تعداد زیادی از نقاط موجود در داده‌ها حساب شود و به طور افزایشی مرتب شود و همچنین k یک مقدار مناسب باشد آنگاه یک تغییر شدید در یک مقدار $k - dist$ مشاهده می‌شود. مقداری از $k - dist$ که تغییر شدید در مقدار آن بوجود می‌آید به عنوان ε شناخته می‌شود و $MinPts$ برابر با k در نظر گرفته می‌شود[۲۳]. مقدار k به طور معمول در روش DBSCAN تصادفی انتخاب می‌شود در پژوهش انجام شده

[1] Density-reachable
[2] Density-connected

مقدار k برای هر طبقه، جداگانه در نظر گرفته شده و به طور معمول برابر با ۴ انتخاب شده است.



شکل ۵: تعیین ε و MinPts

برای طبقه‌بندی دامنه پروتئین‌ها ابتدا ماتریس داده‌های پروتئین تشکیل می‌شود این ماتریس شامل تمامی بردارهای ویژگی استخراج شده از توالی پروتئین‌های در دسترس است. این ماتریس دارای ابعاد $N \times q$ می‌باشد به طوریکه N نشان‌دهنده تعداد بردار ویژگی در دسترس و q برابر با تعداد ویژگی موجود در هر بردار ویژگی است.

$$\boldsymbol{M}=\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1q} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2q} \\ & & \vdots & & \\ x_{N1} & x_{N2} & x_{N3} & \dots & x_{Nq} \end{bmatrix}_{N\times q} \quad (۳)$$

برای طبقه‌بندی از یک ماتریس تشابه و یک ماتریس عدم تشابه استفاده شده است. ماتریس شباهت ماتریسی با ابعاد $N \times N$ است که عضو p_{ij} آن میزان شباهت پروتئین iام به پروتئین jام است (سطر i و j ام ماتریس داده‌ها).

ماتریس عدم شباهت نیز ماتریسی با ابعاد $N \times N$ است که عضو p_{ij} آن میزان عدم شباهت پروتئین iام به پروتئین jام می‌باشد.

۱- برای ماتریس عدم تشابه خواهیم داشت:

$$\begin{cases} p_{ii} = d_0 & for\ all\ \ i \\ d_0 \le p_{ij} & for\ all\ \ i, j \end{cases} \quad (۴)$$

و برای ماتریس تشابه داریم:

$$\begin{cases} p_{ii} = s_0 & for\ all\ \ i \\ p_{ij} \le s_0 & for\ all\ \ i, j \end{cases} \quad (۵)$$

(در حالتی که از ماتریس تشابه استفاده شود، به طور معمول مقادیر ماتریس تشابه بین صفر و یک هستند و p_{ii} ها برابر با یک خواهند بود.)

۲- عناصر ماتریس مجاورت دارای ویژگی تقارن هستند.

$$p_{ij} = p_{ji} \quad (۶)$$

عناصر ماتریس مجاورت در حالت ماتریس تشابه برابر با d_0 و در حالت عدم تشابه برابر با s_0 هستند اگر و فقط اگر i برابر با j باشد.

همچنین در روشهای اندازه‌گیری مجاورت دو خاصیت زیر نیز برقرار هستند.

۱- در حالت ماتریس عدم تشابه

$$p_{ik} \le (p_{ij} + p_{jk}) \quad for\ all\ i.\ j\ and\ k \quad (۷)$$

۲- در حالت ماتریس تشابه

$$k \quad (۸)$$

برای معیار شباهت و عدم شباهت از متریک cosine برای اندازه‌گیری شباهت دو بردار ویژگی استفاده می‌شود که از رابطه (۹) بدست می‌آید[۲۲].

$$\cos\left(\underline{x_i}, \underline{x_j}\right) = \frac{\underline{x_i}.\underline{x_j}}{\left\|\underline{x_i}\right\| \left\|\underline{x_j}\right\|} \quad (۹)$$

در رابطه (۹) $\underline{x_i}.\underline{x_j}$ ضرب داخلی دو بردار $\underline{x_i}$ و $\underline{x_j}$ است و $\left\|\underline{x_j}\right\|$ نشان‌دهنده نرم بردار $\underline{x_j}$ است که مطابق با رابطه‌های (۱۰) محاسبه می‌شود[۲۲].

$$\left\|x_j\right\| = \left(\sum_{k=1}^{q} x_{jk}^2\right)^{\frac{1}{2}} \quad (۱۰)$$

ماتریس داده بوسیله رابطه (۱۱) یکنواخت سازی می‌شود. که در این رابطه μ_j و σ_j به صورت رابطه‌های (۱۲)و (۱۳) بدست می‌آیند.

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (۱۱)$$

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad (۱۲)$$

$$\sigma_j = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(x_{ij} - \mu_j\right)^2} \quad (۱۳)$$

۲-۳- ترکیب اطلاعات بدست آمده از روش طبقه‌بندی

برای بهبود طبقه‌بندی دامنه پروتئین‌ها می‌توان با توجه به ویژگی‌های مطرح شده در جدول۲ طبقه‌بندی را انجام داده و سپس با استفاده از روش‌های ترکیب اطلاعات برای کاهش عدم قطعیت‌های موجود نتایج بدست آمده را با یکدیگر ترکیب کرد[۲۰, ۲۵].

برای ترکیب اطلاعات ابتدا تاخوردگی پروتئین‌ها را براساس ویژگی‌های موجود در جدول۲ طبقه‌بندی کرده و به هر طبقه مقدار C_{ij} به عنوان درجه طبقه‌بندی i در ویژگی j نسبت داده می‌شود. حال باید با

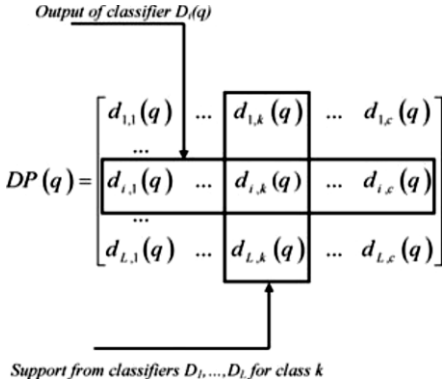
استفاده از روش‌های ترکیب اطلاعات نتایج حاصله را ترکیب کرده و طبقه‌بندی کلی را بر مبنای ترکیب اطلاعات بدست آورد.

ایده روش ترکیب اطلاعات انتگرال فازی بر اساس میزان خبرگی ویژگی‌های جدول۲ در روش استفاده شده برای طبقه‌بندی کننده‌هاست. این میزان خبرگی نه تنها برای هر طبقه‌بندی کننده بلکه برای هر زیر مجموعه از مجموعه‌های طبقه‌بندی می‌باشد. در نتیجه به هر زیر مجموعه از طبقه‌بندی‌ها یک مقدار خبرگی تخصیص داده می‌شود که این میزان خبرگی نشان‌دهنده میزان تصمیم برای کلاس‌های مختلف است.

فرض می‌شود که $\begin{cases} D=\{D_i\} \\ i=1,2,3,\dots,L \end{cases}$ مجموعه‌ی L طبقه‌بندی کننده حاصل از روش تغییر یافته DBSCAN باشد و P(D) مجموعه‌ی توانی D باشد. λ یک اندازه‌ی یکنوای ویژه مانند تابع مجموعه‌ای g است که به وسیله‌ی فرضیات زیر توصیف می‌شود (۱۴):

$$\begin{cases} 1.\ g: P(D) \to [0,1] \\ 2.\ g(0)=0\ , g(D)=1 \\ 3.\ \forall D_i, D_j \in D, D_i \subset D_j \to g(D_i) \le g(D_j) \\ 4.\ \forall D_i, D_j \in D, D_i \cap D_j = \emptyset \to g(D_i \cup D_j) = g(D_i) + g(D_j) + \lambda g(D_i) g(D_j), \lambda \in (-1,\infty) \end{cases}$$

اگر q یک بردار n بعدی از نماینده پروتئین ناشناخته Q باشد.که هر طبقه‌بندی کننده اندازه احتمال را به آن تخصیص می‌دهد. که این اندازه احتمال به تعداد ویژگی‌های مورد استفاده در جدول۲ بستگی دارد. تعداد طبقه‌ها در ویژگی‌های مورد استفاده برابر C و تعداد ویژگی‌ها برابر L در نظر گرفته می‌شود. با توجه به این اندازه احتمالات می‌توان ماتریس پروفایل تصمیم (DP[1]) را تشکیل داد.



(۱۵)

در رابطه (۱۵) هر طبقه‌بندی کننده D_i بردار ویژگی را به عنوان ورودی دریافت می‌کند و C عدد احتمال را به عنوان درجه‌ی پشتیبانی در خروجی تولید می‌نماید.

(۱۶)

$$D_i: R^n \to [0,1]^C$$

هر مقدار $d_{i,k}$ در رابطه‌ی (۱۵) میزان پشتیبانی طبقه‌بندی کننده‌ی D_i از فرضیه‌ی تعلق Q به طبقه‌ی تاخوردگی K می‌باشد. λ اندازه‌ی فازی g را می‌توان از مجموعه‌ای از L مقدار g^{ik} که چگالی فازی نامیده می‌شود از رابطه (۱۷) بدست آورد.

(۱۷)

$$\begin{cases} g^{ik} = N_\varepsilon(i,k) \\ i = 1,2,3,\dots,L \end{cases}$$

با در نظر گرفتن رابطه‌ی (۱۷) g^{ik} میزان خبرگی طبقه‌بندی کننده D_i در طبقه‌بندی تاخوردگی k می‌باشند که از مجموعه داده‌های آموزش به دست می‌آیند.

$$\begin{aligned} N_\varepsilon(i,k) &= \frac{1}{N_{ik}} \sum_t \sum_{j=1}^{C} \left[\left(-\frac{1}{C} \ln \frac{1}{C}\right) - \left(-p_{itj} \ln p_{itj}\right) \right] \\ &= \frac{1}{N_{ik}} \sum_t \sum_{j=1}^{C} \ln \left[C^{\frac{1}{C}} \times (p_{itj})^{p_{itj}} \right] \end{aligned}$$

با به کارگیری چگالی‌های فازی، یک λ اندازه‌ی فازی می‌توان یافت که با این چگالی‌ها سازگار باشد. مقدار λ به عنوان تنها ریشه‌ی حقیقی بزرگتر از ۱- چند جمله‌ای زیر به دست می‌آید:

(۱۸)

$$\begin{cases} \lambda_k + 1 = \prod_{i=1}^{L} \left(1 + \lambda_k g^{ik}\right), \quad \lambda_k \neq 0 \\ i = 1,2,3,\dots L\ ;\ k = 1,2,3,\dots C \end{cases}$$

انتگرال فازی سوگنو شواهد مربوط به یک فرضیه را با انتظار پیشین از اهمیت آن قطعه از شواهد را با هم ترکیب می‌نماید. الگوریتم انتگرال فازی سوگنو به صورت زیر عمل می‌کند.

چگالی‌های فازی $g^{1k}, g^{2k}, g^{3k}, \dots, g^{Lk}$ را با برابر قرار دادن g^{ik} با مهارت محاسبه شده برای طبقه‌بندی کننده D_i در کلاس k تصحیح می‌کند($k = 1,2,3,\dots,C$). سپس مقدار $\lambda_k > -1$ را برای هر کلاس $k = 1,2,3,\dots,C$ از رابطه (۱۹) بدست می‌آورد.

(۱۹)

$$\text{target class} = \arg \operatorname{Max}_{K=1}^{C} \left(\mu_k(q)\right)$$

$\mu_k(q)$ درجه پشتیبانی برای حضور پروتئین q در کلاس k می‌باشد. برای هر پروتئین داده شده q، ستون kام ماتریس DP(q) را جهت بدست آوردن $\left[d_{i_1,k}(q), d_{i_2,k}(q), \dots, d_{i_L,k}(q)\right]^T$ مرتب می‌کند. که در این رابطه $d_{i_1,k}(q)$ بالاترین میزان پشتیبانی و $d_{i_L,k}(q)$ کمترین مقدار پشتیبانی را شامل می‌شود.

انتگرال فازی در مرحله بعد چگالی‌های فازی را با توجه به ارتباط آنها مرتب می‌کند و $g^{1k}, g^{2k}, g^{3k}, \dots, g^{Lk}$ را برابر با $g_k(1) = g^{i_1 k}$ قرار خواهد داد.

سپس برای T=2 تا L رابطه بازگشتی (۲۰) را محاسبه می‌کند.

[1] Decision Profile

(۲۰)

$$g_k(t) = g^{i_t k} + g_k(t-1) + \lambda g^{i_t k} g_k(t-1)$$

در پایان میزان نهایی پشتیبانی برای کلاس k از رابطه (۲۱) بدست می‌آید[۳۲].

(۲۱)

$$\mu_k(q) = \max_{t=1}^{L}\{\min\{d_{i_t,k}(q), g_k(L)\}\}$$

۳- نتیجه‌گیری

در این مقاله روش ارائه شده با توجه به تعریف چگالی جدید می‌تواند قطعیت بیشتری را در زمینه مراکز طبقات از خود نشان دهد. با افزایش مراکز طبقات می‌توان از قطعیت وجود همسایگی‌های بدست آمده نیز اطمینان حاصل کرد. در روش مورد استفاده برای طبقه‌بندی تاخوردگی پروتئین‌ها از ۱۰ ویژگی برای طبقه‌بندی پروتئین‌ها استفاده شده است. استفاده از ۱۰ ویژگی برای طبقه‌بندی پروتئین‌ها می‌تواند خبرگی بیشتری را برای طبقه‌بندی ایجاد کند. در پایان طبقه‌بندی نتایج حاصل از طبقه‌بندی کننده‌ها با استفاده از عملگر ترکیب اطلاعات انتگرال فازی سوگنو با هم ترکیب شده و میزان خبرگی هر ویژگی در این عملگر مد نظر قرار می‌گیرد.

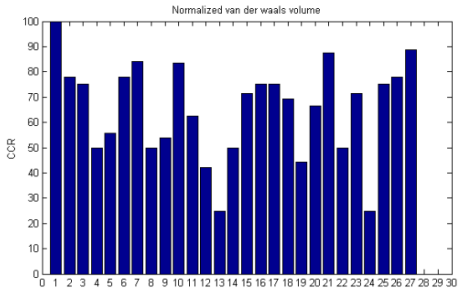
روش طبقه‌بندی ارائه شده یک روش خودکار برای طبقه‌بندی تاخوردگی‌های پروتئین است که می‌تواند بدون ناظر عمل طبقه‌بندی را با دقت بالایی انجام دهد. روش طبقه‌بندی ارائه شده با فرض نامحدود بودن طبقات عمل طبقه‌بندی را انجام می‌دهد. در نتیجه می‌توان از این روش برای پروتئین‌های کشف نشده استفاده کرده و آنها را طبقه‌بندی کرد.

روش طبقه‌بندی ذکر شده بر روی داده‌های قسمت قبل اعمال گردید. شکل۶ نشان دهنده‌ی نتایج طبقه‌بندی هر یک از ویژگی‌های جدول ۲ برای کلاس‌های تاخوردگی مختلف پروتئین‌ها است. همانطور که در شکل‌ها مشخص است نتایج حاصله از هر یک از ویژگی‌ها با ویژگی دیگر متفاوت است. که این امر می‌تواند در تشخیص یک نوع تاخوردگی مفید واقع شود. چرا که ممکن است نتایج حاصل از یک ویژگی در یک طبقه نتیجه ضعیفی داشته باشد ولی ویژگی دیگر در همان طبقه نتیجه قابل قبولی را ارائه دهد. با استفاده از ترکیب اطلاعات می‌توان ضعف طبقه‌بندی حاصل از یک ویژگی را در یک نوع تاخوردگی بوسیله طبقه‌بندی ویژگی‌های دیگر جبران کرد.

نتایج اعمال شده به صورت نرخ طبقه‌بندی صحیح (CCR)[1] بیان شده است. همانطور که از شکل‌ها مشخص است بهترین ویژگی برای طبقه‌بندی پروتئین‌ها طبقه‌بندی براساس ویژگی Predicted secondary structure و PSSM است که می‌توان از آن برای خبرگی بیشتری در ترکیب اطلاعات انتگرال فازی سوگنو استفاده کرد.

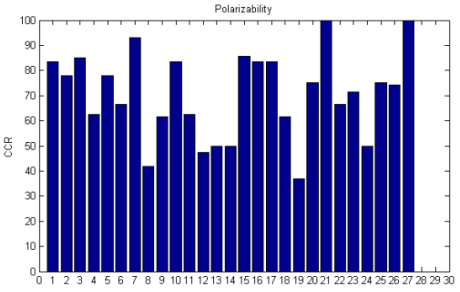
شکل ۷ نشان دهنده مقایسه روش طبقه‌بندی با کارهای پیشین است. در این شکل روش طبقه‌بندی اعمال شده با سه روش PFRES[26]، Information Theoretic Classifier Fusion[13] و Heperfold[19] که بیشترین نرخ صحت طبقه‌بندی را دارا هستند مقایسه شده است، همانطور که از شکل ۷ مشخص است روش اعمالی در شش نوع تاخوردگی globin-like(1)، 4-helical up-1nd-down bundle(4)، conA-like lectin/glucanases(10)، periplasmic binding protein-like(24)، grasp-β(25) و small inhibitors,toxins,lectins(27) بالاترین نرخ صحت طبقه‌بندی را دارا است.

[1] Correct Classification Rate



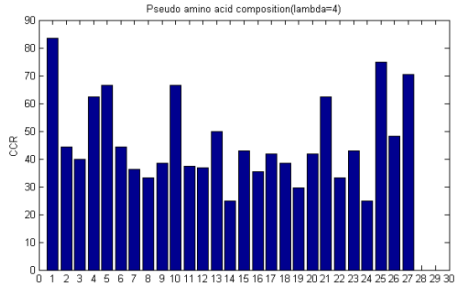
نتیجه طبقه‌بندی Hydrophobicity

نتیجه طبقه‌بندی Normalized van der waals volume



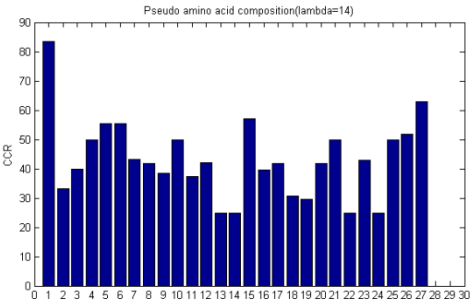
نتیجه طبقه‌بندی Polarity

نتیجه طبقه‌بندی Polarizability



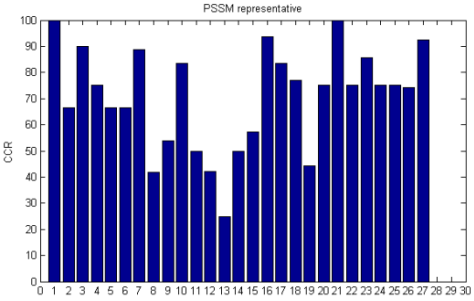
نتیجه طبقه‌بندی Predicted secondary structure

نتیجه طبقه‌بندی Pseude amino acid (λ=4)



نتیجه طبقه‌بندی Pseude amino acid (λ=1)

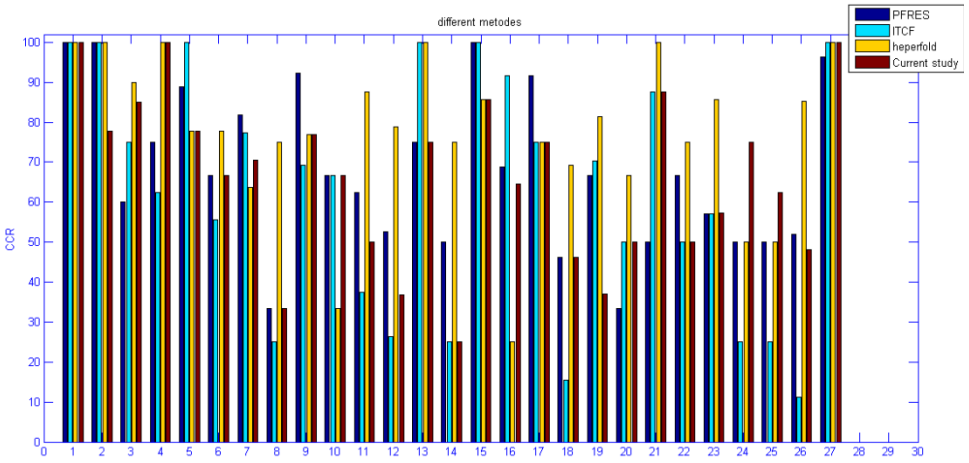
نتیجه طبقه‌بندی Pseude amino acid (λ=14)



نتیجه طبقه‌بندی Pseude amino acid (λ=30)

نتیجه طبقه‌بندیPSSM

شکل ۶:نتایج طبقه‌بندی ویژگی‌ها با استفاده از روش DBSCAN



شکل ۷: مقایسه روش اعمالی با روش‌های پیشین

مراجع

[1] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of molecular biology*, vol. 247, pp. 536-540, 1995.

[2] C. A. Orengo, A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton, "CATH–a hierarchic classification of protein domain structures," *Structure*, vol. 5, pp. 1093-1109, 1997.

[3] C. B. Anfinsen, "Studies on the principles that govern the folding of protein chains," ed, 1972.

[4] C. Anfinsen and H. Scheraga, "Experimental and theoretical aspects of protein folding," *Adv Protein Chem*, vol. 29, pp. 205-300, 1975.

[5] L. Holm and C. Sander, "Protein folds and families: sequence and structure alignments," *Nucleic acids research*, vol. 27, pp. 244-247, 1999.

[6] P. Jain, J. M. Garibaldi, and J. D. Hirst, "Supervised machine learning algorithms for protein structure classification," *Computational Biology and Chemistry*, vol. 33, pp. 216-223, 2009.

[7] K. NISHIKAWA and O. Tatsuo, "Correlation of the amino acid composition of a protein to its structural and biological characters," *Journal of Biochemistry*, vol. 91, pp. 1821-1824, 1982.

[8] J. Lee, "Measures for the assessment of fuzzy predictions of protein secondary structure," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, pp. 453-462, 2006.

[9] C. Lampros, C. Papaloukas, T. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "Sequence-based protein structure prediction using a reduced state-space hidden Markov model," *Computers in Biology and Medicine*, vol. 37, pp. 1211-1224, 2007.

[10] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences*, vol. 92, pp. 8700-8704, 1995.

[11] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, pp. 401-407, 1999.

[12] C. H. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349-358, 2001.

[13] K. Kavousi, B. Moshiri, M. Sadeghi, B. N. Araabi, and A. A. Moosavi-Movahedi, "A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM," *Computational biology and chemistry*, vol. 35, pp. 1-9, 2011.

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, pp. 21-27, 1967.

[15] L. I. Kuncheva, "Combining pattern classifiers: Methods and algorithms (kuncheva, li; 2004)[book review]," *Neural Networks, IEEE Transactions on*, vol. 18, pp. 964-964, 2007.

[16] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The annals of mathematical statistics*, vol. 38, pp. 325-339, 1967.

[17] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 25, pp. 804-813, 1995.

[18] G. Shafer, *A mathematical theory of evidence* vol. 1: Princeton university press Princeton, 1976.

[19] K. Kavousi, M. Sadeghi, B. Moshiri, B. N. Araabi, and A. A. Moosavi-Movahedi, "Evidence Theoretic Protein Fold Classification Based on the Concept of Hyperfold" *Mathematical Biosciences*, 2012.

[20] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical reviews in biochemistry and molecular biology*, vol. 30, pp. 275-349, 1995.

[21] ک.کاوسی, "طبقه بندی خودکار دامنه پروتئین با رویکرد ترکیب اطلاعات ", دانشگاه تهران, ۱۳۹۱.

[22] P.-N. Tan, M. Steinbach, and V. Kumar, "khdaw. com," 2006.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996, pp. 226-231.

[24] H.-P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 672-677.

[25] M. Kazemian, B. Moshiri, H. Nikbakht, and C. Lucas, "A new expertness index for assessment of secondary structure prediction engines," *Computational biology and chemistry*, vol. 31, pp. 44-47, 2007.

[26] K. Chen and L. Kurgan, "PFRES: protein fold classification by using evolutionary information and predicted secondary structure," *Bioinformatics*, vol. 23, pp. 2843-2850, 2007.