



Flexibility Prediction of Protein Structures Using Support Vector Machine

ARTICLE INFO

Article Type

Original Research

Authors

Mozafari Lagha M.¹ MSc,
Arab S.Sh.^{*2} PhD,
Zahiri J.³ PhD

How to cite this article

Mozafari Lagha M, Arab S Sh, Zahiri J. Flexibility Prediction of Protein Structures Using Support Vector Machine. Modares Journal of Biotechnology. 2018;9(4):549-555.

¹Biological Sciences Faculty, Tarbiat Modares University, Tehran, Iran

²Biophysics Department, Biological Sciences Faculty, Tarbiat Modares University, Tehran, Iran

³Bioinformatics Department, University of Tehran, Tehran, Iran

*Correspondence

Address: Tarbiat Modares University, Nasr Bridge, Jalal-Al-Ahmad Highway, Tehran, Iran

Phone: +98 (21) 82883494

Fax: -

sh.arab@modares.ac.ir

Article History

Received: May 10, 2016

Accepted: June 09, 2017

ePublished: December 21, 2018

ABSTRACT

Aims : Information of the protein structure is essential to understand the protein functions. Flexibility is one of the most important characteristics related to protein functions. Knowledge about flexibility of the protein structures can be helpful to improve protein structure prediction and comprehend their function. This study was conducted with the aim of investigating the flexibility prediction of protein structures, using support vector machine.

Materials & Methods In this study, a balanced dataset containing 95 proteins was used. The features used in the present study for modeling amino acids formed a 33-dimensional vector. Some of them were obtained by crawling a window with the length of 17 focusing on the target amino acid on the protein chain, and some were only related to the target amino acid. To define the flexibility factor, the characteristics based on the information derived from the two-dimensional angular variations was used. The information was calculated for each amino acid by considering the position of each amino acid alone and for the adjacent amino acid pairs in a seventeenth window, and the support vector machine method was used for prediction.

Findings The accuracy was 73.1%, F-measure was 71%, precision was 73%, and sensitivity was 73.2%. Acceptable superiority of the proposed method was confirmed in comparison with the current methods. The angular representation of each protein was able to accurately demonstrate the 3D characteristics and properties of the protein structure.

Conclusion The accuracy is 73.1%, F-measure is 71%, precision is 73%, and sensitivity is 73.2% and angular aspect is the best descriptor for flexibility prediction. Angular representation of each protein can accurately reflect the 3D characteristics and properties of the protein structure.

Keywords Flexibility Prediction; Two-dimensional Angular Variations; Information Theory; Support Vector Machine

CITATION LINKS

[1] Prediction of protein motions from amino acid sequence and its application to protein-protein ... [2] Prediction of flexible/rigid regions from protein ... [3] Prediction of chain flexibility in ... [4] Accuracy of protein flexibility ... [5] Improved amino acid flexibility ... [6] Protein flexibility and rigidity predicted from ... [7] PROFbval: Predict flexible and rigid residues ... [8] Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature ... [9] PredyFlexy: Flexibility and local structure prediction from ... [10] Predicting protein flexibility through the ... [11] Protein conformational flexibility prediction ... [12] Fluctuations of backbone torsion angles obtained from NMR-determined structures and their ... [13] Prediction of flexibility in proteins ... [14] Protein flexibility modeling using kernel ... [15] Prediction of protein flexibility ... [16] FlexPred: A web-server for predicting ... [17] Structural flexibility in proteins: Impact ... [18] CABS-flex predictions of protein flexibility ... [19] Coarse-grained molecular ... [20] Consistent view of protein fluctuations from ... [21] CABS-flex: Server for fast simulation of protein structure ... [22] What is a support vector ... [23] Support vector machines, the interface to libsvm in package ... [24] Protein data ... [25] CD-HIT Suite: A web server for clustering and comparing biological ... [26] Handling imbalanced datasets ... [27] PPIevo: Protein-protein interaction prediction from PSSM based evolutionary ... [28] Prediction of coordination number and relative solvent accessibility in ... [29] The GOR method for predicting secondary structures in proteins. In: Fasman GD, editor. Prediction of protein structure and the principles of protein ... [30] Gapped BLAST and PSI-BLAST: A new generation of protein database search ... [31] Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural ... [32] SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural ...

پیشگویی انعطاف‌پذیری ساختارهای پروتئینی با استفاده ماشین بردار پشتیبان

مژگان مظفری لقا MSc

دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران، ایران

سیدشهریار عرب PhD*

گروه بیوفیزیک، دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران، ایران

جواد ظهیری PhD

گروه بیوفیزیک، دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران، ایران

چکیده

اهداف: اطلاعات موجود در ساختمان پروتئین‌ها برای درک چگونگی فعالیت آنها بسیار مفید است. انعطاف‌پذیری یکی از مهم‌ترین فاکتورهای ساختمانی مرتبط با عملکرد پروتئین‌ها است. دانش درباره انعطاف‌پذیری ساختارهای پروتئینی، کمک بزرگی به کیفیت پیشگویی ساختمان پروتئین‌ها و درک عملکرد پروتئین‌ها می‌کند. مطالعه حاضر با هدف بررسی پیش‌گویی انعطاف‌پذیری ساختارهای پروتئینی با استفاده از ماشین بردار پشتیبان انجام شد.

مواد و روش‌ها: در مطالعه حاضر از یک مجموعه داده متعادل ۹۵ پروتئینی استفاده شد. ویژگی‌های استفاده‌شده در مطالعه حاضر برای مدل‌کردن اسیدآمینه‌ها، یک بردار ۳۳ بعدی را تشکیل داد. برخی از آنها از لغزاندن پنجره‌ای به طول ۱۷ با مرکزیت اسیدآمینه هدف روی زنجیره پروتئین به دست آمده‌اند و برخی تنها مربوط به اسیدآمینه هدف بودند. برای تعریف فاکتور انعطاف‌پذیری، ویژگی‌های مبتنی بر اطلاعات حاصل از تغییرات زوایای دووجهی، استفاده شد. این اطلاعات برای هر اسیدآمینه با در نظر گرفتن موقعیت هر اسیدآمینه به‌تنهایی و برای جفت اسیدآمینه‌های مجاور در یک پنجره هفده‌تایی محاسبه و برای پیشگویی از روش ماشین بردار پشتیبان استفاده شد.

یافته‌ها: میزان صحت ۷۳/۱٪، معیار F ۷۱٪ دقت ۷۳٪ و حساسیت ۷۳/۲٪ به دست آمد. برتری قابل قبول روش پیشنهادی در مقایسه با روش‌های موجود تایید شد. نمایش زاویه‌ای هر پروتئین توانست به‌خوبی خصوصیات و ویژگی‌های ساختار سه‌بعدی پروتئین را نشان دهد.

نتیجه‌گیری: میزان صحت ۷۳/۱٪، معیار F ۷۱٪ دقت ۷۳٪ و حساسیت ۷۳/۲٪ است و بهترین نگاه به بحث انعطاف‌پذیری، نگاه زاویه‌ای است. نمایش زاویه‌ای هر پروتئین می‌تواند به‌خوبی خصوصیات و ویژگی‌های ساختار سه‌بعدی پروتئین را نشان دهد.

کلیدواژه‌ها: پیشگویی انعطاف‌پذیری، تغییرات زوایای دووجهی، تئوری اطلاعات، ماشین بردار پشتیبان

تاریخ دریافت: ۱۳۹۵/۰۲/۲۱

تاریخ پذیرش: ۱۳۹۶/۰۳/۱۹

*نویسنده مسئول: sh.arab@modares.ac.ir

مقدمه

انعطاف‌پذیری ساختارهای پروتئینی یکی از مهم‌ترین فاکتورهای ساختاری مرتبط با عملکرد پروتئین‌ها است. عملکرد پروتئین‌ها غالباً به برهمکنش‌های برگشت‌پذیر بین پروتئین‌ها و محیط اطراف‌شان مانند آب، نمک‌ها، یون‌ها، سایر پروتئین‌ها، اسیدنوکلئیک‌ها و مولکول‌های کوچک و بزرگ بی‌شمار موجود در سیستم زنده بستگی دارد. طبیعت این برهمکنش از اهمیت و جایگاه بالایی برخوردار است و این پدیده به‌ویژه به موجود زنده این امکان را می‌دهد تا سریع و به‌طور برگشت‌پذیر به تغییرات محیطی و شرایط متابولیکی پاسخ دهد. ضمن این تغییرات و اتصالات اغلب ساختمان فضایی پروتئین تغییر می‌کند، این تغییرات ممکن است جزئی باشد و ارتعاشات مولکولی و جنبش‌های کوچک باقی‌مانده‌های اسیدآمینه‌ای را منعکس کند، گاهی تغییرات شکل فضایی، ممکن است بسیار وسیع و ژرف باشد[1].

حرکات و تغییرات شکل فضایی برای عملکرد پروتئین، به

اختلافات موجود در پایداری ساختاری بخش‌های مختلف یک پروتئین بستگی دارد. برای مثال جایگاه‌های اتصال یک پروتئین آلوستریک، شامل قسمت‌های پایدار در مجاورت بخش‌های ناپایدار است که قطعات ناپایدار، مسئول تغییرات شکل فضایی مکرر یا حرکات نامنظم و سازمان‌دهی شده است. این موضوع منجر به تقسیم‌بندی قطعات پروتئین به دو دسته مناطق سخت و انعطاف‌پذیر می‌شود[2]. مطالعات انجام‌شده در مورد انعطاف‌پذیری به‌طور کلی پیرامون دو موضوع شامل تعیین نواحی انعطاف‌پذیر و سخت در پروتئین‌ها و پیشگویی این مناطق فقط با داشتن توالی پروتئین‌ها، تمرکز می‌کنند.

برای پیشگویی انعطاف‌پذیری ساختارهای پروتئینی در دهه‌های اخیر الگوریتم‌های مختلفی ارایه شده است. این الگوریتم‌ها از پارامترهای مختلفی برای بیان انعطاف‌پذیری استفاده می‌کنند، از جمله آنها می‌توان به فاکتور حرارتی حاصل از کریستالوگرافی اشعه ایکس[3-9]، RMSF حاصل از نوسانات اتمی حاصل از شبیه‌سازی مولکولی[10]، پارامترهای نظم به‌دست‌آمده از طیف‌سنجی رزونانس مغناطیسی هسته[11]، پراکندگی زوایای دی‌هیدرال[12] و پروفایل انرژی[13] اشاره کرد. یکی از پرکاربردترین فاکتور مورد نظر، فاکتور حرارتی است که منعکس‌کننده نوسانات اتم‌ها در موقعیت‌های میانگین‌شان می‌باشد. اکثر روش‌های موجود از فاکتورهای حرارتی حاصل از کریستالوگرافی اشعه ایکس به‌عنوان فاکتور انعطاف‌پذیری استفاده می‌کنند[3-10، 14-16]، ولی به‌دلیل اینکه در این روش از فرم کریستاله استفاده می‌شود و شرایط تهیه کریستال و سایر محدودیت‌های ساختاری موجود در این روش، برای استفاده از این پارامتر باید با احتیاط رفتار کرد[17، 18]. پارامترهای نظم حاصل از رزونانس مغناطیسی هسته، میانگین جابه‌جایی اتمی حاصل از روش‌های شبیه‌سازی مولکولی روش‌های جایگزینی هستند که برای پیشگویی استفاده شده‌اند[19-21]. در این مطالعه، برای پیشگویی انعطاف‌پذیری، یک روش جدید مبتنی بر الگوریتم ماشین بردار پشتیبان ارایه شد. همچنین در طی این پژوهش خصیصه جدید، ویژگی مبتنی بر اطلاعات حاصل از تغییرات زوایای دووجهی برای انعطاف‌پذیری ساختارهای پروتئینی معرفی و استفاده شد.

مطالعه حاضر با هدف بررسی پیش‌گویی انعطاف‌پذیری ساختارهای پروتئینی با استفاده از ماشین بردار پشتیبان انجام شد.

اِزار و روش‌ها

ماشین بردار پشتیبان: ماشین بردار پشتیبان، یکی از روش‌های یادگیری با نظارت است که از تئوری یادگیری آماری سرچشمه می‌گیرد و از آن برای طبقه‌بندی و رگرسیون استفاده می‌کند[22]. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی برای طبقه‌بندی نسبت به روش‌های قدیمی‌تر از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری طبقه‌بندی ماشین بردار پشتیبان، طبقه‌بندی خطی داده‌ها است. در تقسیم‌بندی خطی داده‌ها، سعی می‌شود خطی انتخاب شود که حاشیه اطمینان بیشتری داشته باشد. یکی از خصوصیات مهم بردار پشتیبان این است که به‌طور همزمان خطای تجربی طبقه‌بندی را کمینه نموده و حاشیه‌های هندسی را بیشینه می‌کند، بنابراین طبقه‌بندی بیشینه‌کننده حاشیه نیز نامیده می‌شود.

برای یک مسئله دوبعدی با دو دسته پاسخ، خطوط بی‌شماری

$$\Delta(\tau_k^i, \tau_k^j) = \begin{cases} \frac{|\tau_k^i - \tau_k^j|}{180} \text{ if } |\tau_k^i - \tau_k^j| < 180 \\ 2 - \frac{|\tau_k^i - \tau_k^j|}{180} \text{ Otherwise} \end{cases}$$

برای اینکه تغییرات زوایای دوجویی به‌دست‌آمده برای هر اسیدآمین به پروتئین‌های مختلف با یکدیگر قابلیت مقایسه داشته باشند، از یک فاکتور نرمال‌سازی به نام C_m استفاده شد که براساس رابطه زیر به‌دست می‌آید:

$$C_m = \begin{cases} \frac{2(m-1)}{m} & \text{if } m \text{ is even} \\ \frac{2m}{(m+1)} & \text{if } m \text{ is odd} \end{cases}$$

بعد از محاسبه تغییرات زوایای دوجویی، میزان تمایل یک اسیدآمین به قرارگیری در یک منطقه انعطاف‌پذیر و یا سخت به‌وسیله ویژگی‌های مبتنی بر تئوری اطلاعات به‌دست آمد [29]. این مقدار برای هر اسیدآمین به‌تنهایی و برای جفت اسیدآمین‌ها در موقعیت‌های مجاور هم محاسبه شد. در این مطالعه، دو رخداد یکی مربوط به اسید آمینه خاص (R) در توالی پروتئینی و دیگری انعطاف‌پذیری یا سختی اسیدآمین خاص (S) وجود داشت. براساس محاسبات تئوری اطلاعات رابطه زیر به‌دست می‌آید:

$$I(S; R) = \log \left(\frac{P(S|R)}{P(S)} \right)$$

که $P(S|R)$ احتمال شرطی رخداد انعطاف‌پذیری به سختی برای اسیدآمین خاص است و براساس رابطه Information difference، رابطه زیر را داریم:

$$I(\Delta S; R) = I(S; R) - I(n - S; R) = \log[f_{S,R}/f_{n-S,R}] + \log[f_{n-S,S}/f_{S,S}]$$

در رابطه فوق، f_S تعداد کل اسیدآمین‌های در حالت انعطاف‌پذیر، f_{n-S} برابر با تعداد کل اسیدآمین‌های در حالت سخت، $f_{S,R}$ فراوانی اسیدآمین‌های خاص در حالت انعطاف‌پذیربودن و $f_{n-S,R}$ فراوانی اسیدآمین‌های خاص در حالت سخت‌بودن هستند. از این رابطه برای محاسبه اطلاعات هر باقی‌مانده به‌تنهایی استفاده می‌شود.

$$I(\Delta S_j; R_{j-m}, \dots, R_j, \dots, R_{j+m}) \approx I(\Delta S_j; R_j) + \sum_{m,m \neq 0} I(\Delta S_j; R_{j+m})$$

می‌توان رابطه فوق را بسط داد و از آن برای محاسبه اطلاعات جفت اسیدآمین‌ها استفاده کرد. در این مطالعه از یک پنجره هفده‌تایی برای محاسبه حال‌های جفتی استفاده شد:

$$I(\Delta S_j; R_{j-m}, \dots, R_j, \dots, R_{j+m}) \approx I(\Delta S_j; R_j) + \sum_{m,m \neq 0} I(\Delta S_j; R_{j+m} | R_j)$$

پیشگویی انعطاف‌پذیری: برای پیشگویی انعطاف‌پذیری، از روش ماشین بردار پشتیبان استفاده شد. همان گونه که در بخش قبل توضیح داده شد در این روش انتخاب توابع کرنل و همچنین

ممکن است وجود داشته باشد که بتوان توسط آنها طبقه‌بندی را انجام داد ولی فقط یکی از این خطوط، بیشینه تفکیک و جداسازی را دارد. مفهوم آموزشی که نمونه‌ها بتوانند به‌عنوان نقاط در یک فضای با ابعاد بالا طبقه‌بندی شوند و پیداکردن خطی که آنها را جدا کند منحصر به فرد نیست. آنچه که ماشین بردار پشتیبان را از سایر جداکننده‌ها متمایز می‌کند، چگونگی انتخاب ابرصفحه است. در ماشین بردار پشتیبان، بیشینه‌کردن حاشیه بین دو کلاس مدنظر است، بنابراین ابرصفحه‌ای را انتخاب می‌کند که فاصله آن از نزدیک‌ترین داده‌ها در هر دو طرف جداکننده خطی بیشینه باشد. چنین ابرصفحه‌ای به‌عنوان ابر صفحه بیشینه حاشیه شناخته می‌شود. برای اینکه بتوان مسئله ابعاد خیلی بالا را به‌وسیله این روش حل کرد، از توابعی به‌نام هسته استفاده می‌شود. توابع مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگموئید را می‌توان استفاده کرد، که هر کدام از این توابع خود شامل پارامترهای مختلفی هستند که باید توسط کاربر تعیین شوند [23].

مجموعه داده پروتئینی: مجموعه داده‌ای ۴۳۷۸ پروتئین از سایت بانک اطلاعات پروتئینی RCSB انتخاب شد [24]. در این مجموعه پروتئین‌هایی با طول بیشتر از ۴۰ اسیدآمین که به‌وسیله رزونانس مغناطیسی هسته تعیین ساختار شده بودند و بیش از پنج مدل ساختمانی داشتند، انتخاب شدند. برای جلوگیری از حضور پروتئین‌های مشابه در بانک، از نرم‌افزار تحت وب CD-HIT استفاده شد [25]. پس از اعمال حذف در مرحله پیشین ۲۸۳۱ پروتئین باقی ماند. در نهایت، به‌دلیل ماهیت طبیعی ساختارهای پروتئینی و عدم توازن موجود در میزان انعطاف‌پذیری نسبت به سختی که منجر به ایجاد داده‌های مثبت کاذب و منفی کاذب در مجموعه داده مورد استفاده در پیشگویی می‌شود، از روش بازنمونه‌گیری (Under resampling) برای متعادل‌سازی مجموعه داده استفاده شد. در این تکنیک مجموعه داده با کاهش تعداد نمونه‌های افزونه متعادل می‌شود [26, 27]. در نهایت با روش حذف تصادفی نمونه‌های کلاس حداکثر، یک مجموعه داده متعادل شامل ۹۵ پروتئین باقی ماند که ادامه روند پیشگویی روی این مجموعه داده صورت گرفت.

در نظرگرفتن زوایای موجود در ساختارهای پروتئینی: شکل فضایی یک پروتئین به‌وسیله سه زاویه دوجویی تعیین می‌شود که امگا، فی و پسی نامیده شده و منعکس‌کننده چرخش اطراف هر یک از سه پیوند تکراری در اسکلت اصلی پروتئین است. این زاویه‌ها به‌عنوان یکی از مهم‌ترین توصیف‌گرهای انعطاف‌پذیری شناخته می‌شوند. در این مطالعه، برای پیشگویی انعطاف‌پذیری از تغییرات زوایای فی و پسی استفاده شد.

استخراج زوایای دوجویی از فایل‌های پروتئینی: همانگونه که در انتخاب مجموعه داده پروتئینی مطرح شد، در این مطالعه، از پروتئین‌هایی که توسط رزونانس مغناطیسی هسته تعیین ساختار شده بودند، استفاده شد. از برنامه DSSP به‌منظور استخراج زوایای دوجویی هر اسیدآمین در مدل‌های مختلف ساختمانی استفاده شد [28].

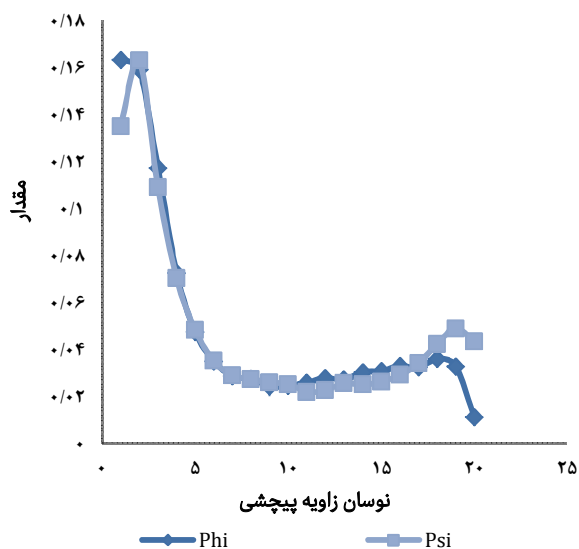
محاسبه تغییرات زوایای دوجویی: برای به‌دست‌آوردن تغییرات زوایا از رابطه $\Delta\tau_k = \frac{C_m}{m(m-1)} \sum_{i < j} \Delta(\tau_k^i, \tau_k^j)$ استفاده شد [10]. که $\Delta\tau_k$ تغییرات زوایای دوجویی برای k ام اسیدآمین می‌باشد، m برابر با تعداد کل مدل‌هایی که در پروتئین وجود دارد، $\Delta(\tau_k^i, \tau_k^j)$ براساس رابطه زیر به‌دست می‌آید و قدر مطلق فاصله نرمال‌شده بین زاویه τ_k^i با زاویه τ_k^j اسیدآمین k ام است:

عدم آزادی چرخش میزان اطلاعات انعطاف‌پذیری اش بسیار پایین بود.

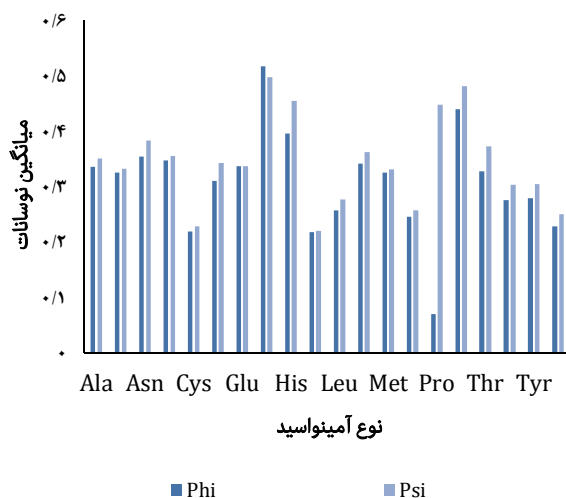
انعطاف‌پذیری اسیدآمینه‌ها فقط به ویژگی‌های فیزیکوشیمیایی آن اسیدآمینه بستگی نداشت، بلکه جایگاه قرارگیری و اسیدآمینه‌های اطراف نیز این میزان را به شدت تحت تاثیر قرار داد (نمودارهای ۷-۳).

اسیدآمینه‌هایی که تمایل بسیار پائینی به قرارگیری در مناطق انعطاف‌پذیر داشتند، زمانی که در ابتدا و انتهای زنجیره‌های پروتئینی بودند، الگوی کاملاً انعطاف‌پذیری از خود نشان دادند. همچنین الگوی انعطاف‌پذیری اسیدآمینه همسایه نقش بسیار تعیین‌کننده در الگوی انعطاف‌پذیری هر اسیدآمینه داشت، به طور مثال اسیدآمینه‌ای مثل سیستئین که الگوی سختی دارد زمانی که در کنار اسیدآمینه انعطاف‌پذیری مثل گلاستین قرار گرفت، انعطاف‌پذیری بیشتری از خود نشان داد (نمودار ۷).

بهترین نگاه به بحث انعطاف‌پذیری، نگاه زاویه‌ای بود، به طوری که میزان صحت ۷۳/۱٪، معیار F ۷۱٪، حساسیت ۷۳/۲٪ و دقت ۷۳٪ به دست آمد.



نمودار (۱) توزیع تغییرات زوایای دووجهی برای کل مجموعه داده مورد مطالعه



نمودار (۲) میانگین نوسانات تغییرات زوایای دووجهی برای هر اسیدآمینه

مقداردهی مناسب پارامترهای کرنل می‌تواند تاثیر بسزایی در دقت نتیجه نهایی داشته باشد. کرنل‌های روش ماشین بردار پشتیبان دو پارامتر C و ϵ دارند. به دست آوردن مقدار بهینه برای این دو پارامتر هنوز یکی از مسایل باز در این حیطه است. تجربه نشان داده هنگامی که مقادیر این پارامترها توانی از ۲ باشد، معمولاً دقت روش ماشین بردار پشتیبان بالا و قابل قبول است. بنابراین مقادیر این دو پارامتر، توانی از ۲ و بین 2^{-10} و 2^{-15} در نظر گرفته شده است. در این پژوهش از توابع چندجمله‌ای استفاده شد.

خصیصه‌هایی که در این پژوهش برای مدل‌کردن اسیدآمینه‌ها استفاده شده‌اند، یک بردار ۴۹ بعدی را تشکیل می‌دهند. برخی از این خصیصه‌ها از لغزاندن پنجره‌ای به طول ۱۷ با مرکزیت اسیدآمینه هدف روی زنجیره پروتئین به دست آمده‌اند و برخی فقط مربوط به اسیدآمینه هدف هستند. در این مطالعه از خصیصه‌های توالی پروتئین مانند ماتریس موقعیت‌دهی جایگاه ویژه^[30] که یک ویژگی ۲۰ بعدی است، ویژگی‌های فیزیکوشیمیایی توالی پروتئینی^[31] در ۵ بعد، ویژگی‌های ساختاری پیشگویی‌شده مثل ساختار دوم و سطح در دسترس^[32] هر کدام در ۲ بعد و توصیف‌گر انعطاف‌پذیری معرفی‌شده در ۱۷ بعد، استفاده شد.

ارزیابی الگوریتم پیشنهادی: برای ارزیابی روش پیشنهادی از روش تقاطع‌سنجی k-گانه استفاده شد. در این روش داده‌ها به k زیرمجموعه افزایش می‌شوند و هر بار یک زیرمجموعه برای آزمون و k-1 زیرمجموعه باقی‌مانده برای آموزش به کار می‌روند. در این مطالعه از روش 10-fold استفاده شد. به طوری که کل داده‌های مجموعه متعادل در ابتدا به ۱۰ زیرمجموعه تقسیم شد که هر زیرمجموعه شامل ۱۰ پروتئین بود. در مرحله بعد ۱۰ پروتئین از مجموعه داده کنار گذاشته شد و آموزش روی ۸۵ پروتئین باقی‌مانده صورت گرفت و در نهایت بررسی روی ۱۰ پروتئین کنار گذاشته‌شده به عنوان مجموعه تست صورت گرفت. این روند با ۱۰ بار تکرار انجام شد.

کارایی روش پیشنهادی با توجه به کیفیت پیش‌گوینده ارزیابی می‌شود. معیارهای زیادی برای ارزیابی وجود دارد. در مطالعه حاضر برای ارزیابی کارایی الگوریتم پیشنهادی از معیارهای تکمیلی صحت، دقت، حساسیت و معیار F استفاده شد.

یافته‌ها

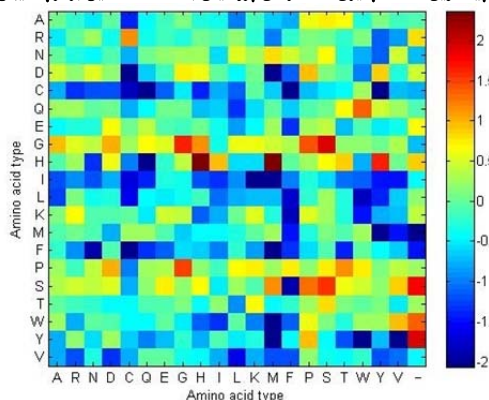
توزیع براساس هر دو زاویه دووجهی، غیریکنواخت مشاهده شد که علت این امر ماهیت طبیعی ساختارهای پروتئینی و عدم توازن میزان انعطاف‌پذیری و سختی در حالت پایدار پروتئین‌ها بود (نمودار ۱).

اسیدآمینه گلاستین بیشترین نوسانات زاویه را داشت و به عنوان انعطاف‌پذیرین اسیدآمینه شناخته شده است.

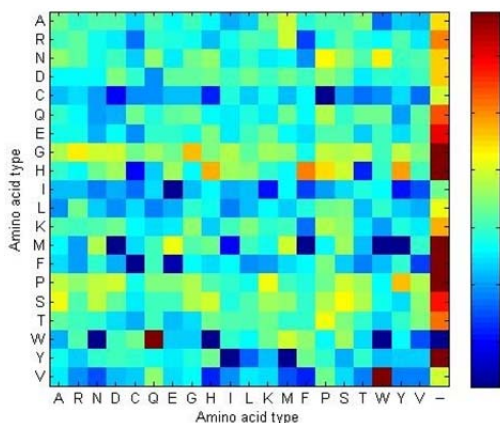
نوسانات زاویه‌ای برای هر دو زاویه دووجهی با یکدیگر مطابقت داشتند به جز برای اسیدآمینه پرولین، به این علت که زنجیره جانبی پرولین با اسکلت اصلی اسیدآمینه متصل شده بود و همین امر منجر به محدودیت و آزادی چرخش آن شد (نمودار ۲).

اسیدآمینه‌هایی مانند گلاستین، سرین، هیستیدین بیشترین اطلاعات انعطاف‌پذیری را داشتند یعنی تمایل قرارگیری این دسته از اسیدآمینه‌ها در مناطق انعطاف‌پذیر بسیار بالا بود (نمودار ۳).

اطلاعات محاسبه‌شده برای هر دو زاویه با یکدیگر تطابق داشت، ولی تنها مورد استثنا اسیدآمینه پرولین بود که در زاویه فی به دلیل



نمودار ۶) نمودار دوبعدی به‌دست‌آمده بر مبنای تئوری اطلاعات با در نظر گرفتن اسیدآمینه‌های مجاور بعد از هر اسیدآمینه براساس تغییرات زوایای پسی (محور Y بیست نوع اسیدآمینه و محور X بیست نوع اسیدآمینه مجاور بعد از هر اسیدآمینه را نشان می‌دهد و علامت منفی نشان‌دهنده این مطلب است که اسیدآمینه مد نظر در انتهای توالی پروتئینی قرار داشته، نوار ابزار رنگی در کنار نمودار میزان انعطاف‌پذیری را نشان می‌دهد و هر چه به رنگ قرمز نزدیک‌تر باشد یعنی انعطاف‌پذیری بیشتر و هر چه به سمت رنگ آبی بزرگ‌تر شود یعنی سختی بیشتر بوده است)

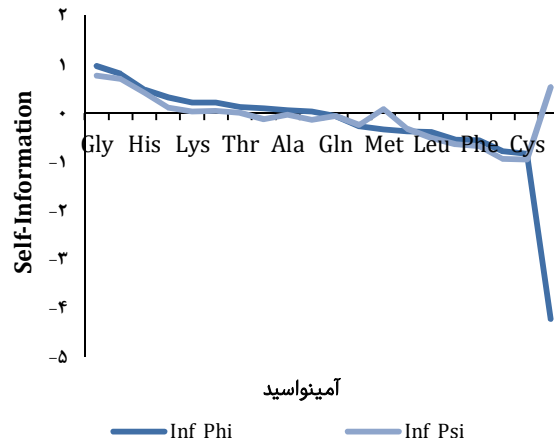


نمودار ۷) نمودار دوبعدی به‌دست‌آمده بر مبنای تئوری اطلاعات با در نظر گرفتن اسیدآمینه‌های مجاور قبل از هر اسیدآمینه براساس تغییرات زوایای پسی (محور Y بیست نوع اسیدآمینه و محور X بیست نوع اسیدآمینه مجاور قبل از هر اسیدآمینه را نشان می‌دهد، علامت منفی نشان‌دهنده این مطلب است که اسیدآمینه مربوط در ابتدای توالی پروتئینی قرار داشته، نوار ابزار رنگی در کنار نمودار نشان‌دهنده میزان انعطاف‌پذیری است و هر چه به رنگ قرمز نزدیک‌تر باشد یعنی انعطاف‌پذیری بیشتر و هر چه به سمت رنگ آبی بزرگ‌تر شود یعنی سختی بیشتر بوده است)

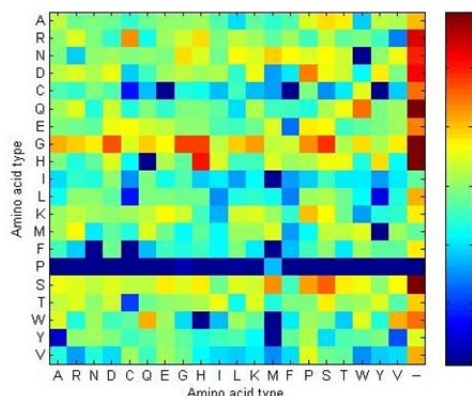
بحث

مطالعه حاضر با هدف بررسی پیشگویی انعطاف‌پذیری ساختارهای پروتئینی با استفاده از ماشین بردار پشتیبان انجام شد.

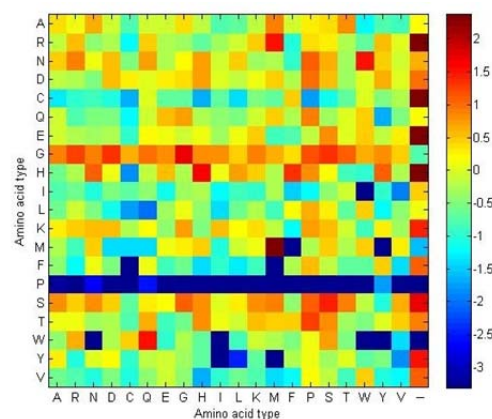
اطلاعات موجود در ساختمان پروتئین‌ها برای درک چگونگی فعالیت آنها ضروری است زیرا ویژگی‌های جایگاه فعال و جایگاه‌های اتصال پروتئین‌ها و همچنین بررسی حرکات داخلی و تحرکات در مقیاس بزرگ‌تر، مثل حرکات لولایی، تغییرات آلوستریک به وضعیت سه‌بعدی دقیق پروتئین وابسته است. انعطاف‌پذیری یکی از مهم‌ترین فاکتورهای ساختمانی مرتبط با عملکرد پروتئین‌ها می‌باشد. به همین علت نیاز به مطالعات ساختاری و ارایه یک تعریف دقیق از انعطاف‌پذیری بیش از پیش احساس می‌شود. یکی از مهم‌ترین و پرکاربردترین توصیف‌کننده‌های انعطاف‌پذیری، فاکتور حرارتی است که



نمودار ۳) اطلاعات مربوط به هر اسیدآمینه براساس تغییرات زوایای دووجهی



نمودار ۴) نمودار دوبعدی به‌دست‌آمده بر مبنای تئوری اطلاعات با در نظر گرفتن اسیدآمینه‌های مجاور بعد از هر اسیدآمینه براساس تغییرات زوایای فی (محور Y بیست نوع اسیدآمینه و محور X بیست نوع اسیدآمینه مجاور بعد از هر اسیدآمینه را نشان می‌دهد. علامت منفی نشان‌دهنده این مطلب است که اسیدآمینه مد نظر در انتهای توالی پروتئینی قرار داشت. نوار ابزار رنگی در کنار نمودار نشان‌دهنده میزان انعطاف‌پذیری بود، هر چه به رنگ قرمز نزدیک‌تر یعنی انعطاف‌پذیری بیشتر و هر چه به سمت رنگ آبی بزرگ‌تر شود یعنی سختی بیشتر بوده است)



نمودار ۵) نمودار دوبعدی به‌دست‌آمده بر مبنای تئوری اطلاعات با در نظر گرفتن اسیدآمینه‌های مجاور قبل از هر اسیدآمینه براساس تغییرات زوایای فی (محور Y بیست نوع اسیدآمینه و محور X بیست نوع اسیدآمینه مجاور قبل از هر اسیدآمینه را نشان می‌دهد. علامت منفی نشان‌دهنده این مطلب است که اسیدآمینه مد نظر در ابتدای توالی پروتئینی قرار داشت. نوار ابزار رنگی در کنار نمودار نشان‌دهنده میزان انعطاف‌پذیری بود و هر چه به رنگ قرمز نزدیک‌تر باشد یعنی انعطاف‌پذیری بیشتر و هر چه به سمت رنگ آبی بزرگ‌تر شود یعنی سختی بیشتر بوده است)

روش‌های موجود در این زمینه مقایسه صورت گرفت. روش پیشنهادی مطالعه حاضر به مراتب نسبت به کارهای انجام شده در این زمینه بهبود پیدا کرده است.

براساس نتایج به دست آمده بهترین نگاه به بحث انعطاف‌پذیری، نگاه زاویه‌ای بود، طبق جدول ۱ روش پیشنهادی پژوهش حاضر نسبت به تحقیقات دیگر در این زمینه بهبود یافته بود [9, 12].

جدول ۱) نتایج پیشگویی انعطاف‌پذیری با روش پیشنهادی و مقایسه با دیگران

معیار F	دقت	حساسیت	صحت	منبع
۵۰/۳	۴۸/۰	۵۰/۵	۴۷/۰	[9]
۵۴/۱	۸۰/۷	۴۰/۰	۴۶/۳	[12]
۷۱/۰	۷۳/۰	۷۳/۲	۷۳/۱	PFP-SVM

یکی از اصلی‌ترین مزایای نمایش زاویه‌ای ساختار پروتئین، نمایش اطلاعات ساختاری پروتئین‌ها در بعد کمتر نسبت به نمایش مختصاتی آن است. به عبارت دیگر با این روش به جای مدل‌سازی مختصات اتم‌ها، دو متغیر زاویه‌ای آنها که ممکن است خاصیت تناوبی داشته باشد، بررسی می‌شود. در نگاه اول به نظر می‌رسد که با کاهش بعد اطلاعات زیادی را از دست خواهیم داد، اما در عمل با در نظر گرفتن ساختارهای اولیه، ساختار دوم و محدودیت‌هایی که در ساختار سه‌بعدی همه پروتئین‌ها مشترک است، می‌توان ساختار پروتئین را از روی هر یک از نمایش‌های زاویه‌ای بازسازی کرد.

نتیجه‌گیری

میزان صحت ۷۳/۱٪، معیار F ۷۱٪ دقت ۷۳٪ و حساسیت ۷۳/۲٪ می‌باشد و بهترین نگاه به بحث انعطاف‌پذیری، نگاه زاویه‌ای است. نمایش زاویه‌ای هر پروتئین می‌تواند به خوبی خصوصیات و ویژگی‌های ساختار سه‌بعدی پروتئین را نشان دهد.

تشکر و قدردانی: موردی از سوی نویسندگان یافت نشد.

تأییدیه اخلاقی: موردی از سوی نویسندگان یافت نشد.

تعارض منافع: موردی از سوی نویسندگان یافت نشد.

سهم نویسندگان: مژگان مظفری لقا (نویسنده اول)، نگارنده مقدمه/پژوهشگر اصلی (۴۰٪)؛ سیدشهریار عرب (نویسنده دوم)، روش‌شناس/پژوهشگر اصلی/نگارنده بحث (۳۰٪)؛ جواد ظهیری (نویسنده سوم)، پژوهشگر کمکی (۳۰٪)

منابع مالی: موردی از سوی نویسندگان یافت نشد.

منابع

- 1- Hirose S, Yokota K, Kuroda Y, Wako H, Endo S, Kanai S, et al. Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. BMC Struct Biol. 2010;10:20.
- 2- Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol. 2007;7:25.
- 3- Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften. 1985;72(4):212-3.
- 4- Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins. 1994;19(2):141-9.
- 5- Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci. 2003;12(5):1060-72.
- 6- Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins. 2005;61(1):115-26.
- 7- Schlessinger A, Yachdav G, Rost B. PROFbval: Predict flexible and rigid residues in proteins. Bioinformatics. 2006;22(7):891-3.

نشان‌دهنده میانگین جابه‌جایی‌های اتمی است. با این وجود به دلیل اینکه فاکتور حرارتی از روش‌های کریستالوگرافی به دست می‌آید، روش‌ها و شرایط آزمایش تهیه کریستال و سایر محدودیت‌های ساختار موجود، منجر به عدم در نظر گرفتن حرکت به خوبی در این روش‌ها می‌شود. در نتیجه احتیاط بیشتر در استفاده از این پارامتر در مطالعات لازم است. در این مطالعه، روش جدیدی بر مبنای ویژگی اطلاعات حاصل از تغییرات زوایای دووجهی ارائه شد. یکی از اصلی‌ترین مزایای این نمایش زاویه‌ای، نمایش اطلاعات ساختاری در بعد کمتر نسبت به نمایش مختصاتی آنها است. این موضوع بیانگر این نکته اساسی است که نمایش زاویه‌ای هر پروتئین به خوبی نشان‌دهنده ویژگی‌های ساختار سه‌بعدی پروتئین‌ها می‌باشد. همان گونه که نتایج نشان دادند، این ویژگی به خوبی توصیف‌کننده انعطاف‌پذیری بوده و نتایج نشان‌دهنده برتری الگوریتم پیشنهادی نسبت به کارهای صورت گرفته در این زمینه بود.

در مطالعه حاضر براساس هر دو زاویه، دووجهی غیریکنواخت بود که علت این امر ماهیت طبیعی ساختارهای پروتئینی و عدم توازن میزان انعطاف‌پذیری و سختی در حالت پایدار پروتئین‌ها است.

براساس نتایج مشاهده شده در مطالعه حاضر، اسیدآمینه گلیسین بیشترین نوسانات زاویه‌ای را داشت و به عنوان انعطاف‌پذیرین اسیدآمینه شناخته شد که با نتایج به دست آمده با تحقیقات انجام شده در این زمینه مطابقت دارد. همچنین اسیدآمینه‌های باردار مثل سرین نیز جز اسیدآمینه‌های انعطاف‌پذیر دسته‌بندی می‌شوند در حالی که اسیدآمینه‌های هیدروفوب مثل ایزولوسین، والین و تریپتوفان کمترین نوسانات زاویه‌ای را دارند و سخت در نظر گرفته می‌شوند.

نوسانات زاویه‌ای برای هر دو زاویه دووجهی با یکدیگر مطابقت دارند به جز برای اسیدآمینه پرولین، به این علت که زنجیره جانبی پرولین به اسکلت اصلی اسیدآمینه متصل شده است و همین امر منجر به محدودیت و آزادی چرخش آن می‌شود.

اسیدآمینه‌هایی مانند گلاسین، سرین، هیستیدین بیشترین اطلاعات انعطاف‌پذیری را دارند یعنی تمایل قرارگیری این دسته از اسیدآمینه‌ها در مناطق انعطاف‌پذیر بسیار بالا است.

اطلاعات محاسبه شده برای هر دو زاویه با یکدیگر تطابق داشت ولی تنها مورد استثنا اسیدآمینه پرولین بود که در زاویه فی به دلیل عدم آزادی چرخش میزان اطلاعات انعطاف‌پذیری آن بسیار پایین به دست آمد.

انعطاف‌پذیری اسیدآمینه‌ها تنها به ویژگی‌های فیزیکوشیمیایی آن اسیدآمینه بستگی ندارد، بلکه جایگاه قرارگیری و اسیدآمینه‌های اطراف نیز این میزان را به شدت تحت تاثیر قرار می‌دهد.

طبق نتایج مطالعه حاضر اسیدهای آمینه‌ای که تمایل بسیار پائینی در قرارگیری در مناطق انعطاف‌پذیر داشتند، زمانی که در ابتدا و انتهای زنجیره‌های پروتئینی قرار می‌گرفتند، الگوی کاملاً انعطاف‌پذیری نشان دادند. همچنین الگوی انعطاف‌پذیری اسیدآمینه همسایه نقش بسیار تعیین‌کننده در الگوی انعطاف‌پذیری هر اسیدآمینه داشت، مثلاً اسیدآمینه‌ای مانند سیستئین که الگوی سختی دارد زمانی که در کنار اسیدآمینه انعطاف‌پذیری مانند گلاسین قرار می‌گرفت، انعطاف‌پذیری بیشتری نشان می‌داد.

در مطالعه حاضر برای پیشگویی انعطاف‌پذیری از ویژگی‌های مبتنی بر اطلاعات حاصل از تغییرات زوایای دووجهی استفاده شد. به منظور سنجش کارایی روش پیشنهادی، بین روش پیشنهادی و

molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J Chem Theory Comput.* 2013;9(1):119-25.

21- Jamroz M, Kolinski A, Kmiecik S. CABS-flex: Server for fast simulation of protein structure fluctuations. *Nucleic Acids Res.* 2013;41(Web Server issue):W427-31.

22- Noble WS. What is a support vector machine?. *Nat Biotechnol.* 2006;24(12):1565-7.

23- Meyer D. Support vector machines, the interface to libsvm in package e1071 [Internet]. Wien: FH Technikum Wien; 2014 [cited 2015 May]. Available from: <https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf>

24- Bank PD. Protein data bank. *Nat New Biol.* 1971;233:223.

25- Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics.* 2010;26(5):680-2.

26- Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. *GESTS Int Trans Comput Sci Eng.* 2006;30.

27- Zahiri J, Yaghoubi O, Mohammad Noori M, Ebrahimpour R, Masoudi Nejad A. PPlevo: Protein-protein interaction prediction from PSSM based evolutionary information. *Genomics.* 2013;102(4):237-42.

28- Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins.* 2002;47(2):142-53.

29- Garnier J, Robson B. The GOR method for predicting secondary structures in proteins. In: Fasman GD, editor. *Prediction of protein structure and the principles of protein conformation.* Berlin: Springer Science & Business Media; 1989. pp. 417-65.

30- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-402.

31- Meiler J, Müller M, Zeidler A, Schmäsche F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol Model Annu.* 2001;7(9):360-9.

32- Magnan CN, Baldi P. SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics.* 2014;30(18):2592-7.

8- Pan XY, Shen HB. Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept Lett.* 2009;16(12):1447-54.

9- De Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly JC. PredyFlexy: Flexibility and local structure prediction from sequence. *Nucleic Acids Res.* 2012;40(Web Server issue):W317-22.

10- Bornot A, Etchebest C, De Brevern AG. Predicting protein flexibility through the prediction of local structures. *Proteins.* 2011;79(3):839-52.

11- Trott O, Siggers K, Rost B, Palmer AG. Protein conformational flexibility prediction using machine learning. *J Magn Reson.* 2008;192(1):37-47.

12- Zhang T, Faraggi E, Zhou Y. Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins.* 2010;78(16):3353-62.

13- Arab SS, Sadeghi M, Eslahchi C, Pezeshk H. Prediction of flexibility in proteins based on sequence of amino acids. *Biotechnol Tarbiat Modares Univ.* 2010;1(1):1-10. [Persian]

14- Chen XW, Chen J. Protein flexibility modeling using kernel based methods. *International Joint Conference on Neural Networks.* Montreal: IEEE; 2005.

15- Yuan Z, Bailey TL, Teasdale RD. Prediction of protein flexibility profile [Internet]. St Lucia: The University of Queensland; 2014 [cited 2015 May]. Available from: https://www.researchgate.net/publication/228899091_Prediction_of_Protein_Flexibility_Profile

16- Kuznetsov IB, McDuffie M. FlexPred: A web-server for predicting residue positions involved in conformational switches in proteins. *Bioinformation.* 2008.3(3):134-6.

17- Hinsen K. Structural flexibility in proteins: Impact of the crystal environment. *Bioinformatics.* 2008;24(4):521-8.

18- Jamroz M, Kolinski A, Kmiecik S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics.* 2014;30(15):2150-4.

19- Markvoort AJ. Coarse-grained molecular dynamics. In: Van Santen RA, Sautet P, editors. *Computational methods in catalysis and materials science: An introduction for scientists and engineers.* New Jersey: John Wiley & Sons; 2009. pp. 151-66.

20- Jamroz M, Orozco M, Kolinski A, Kmiecik S. Consistent view of protein fluctuations from all-atom