



Impact of Hydrophobic Amino Acids in Protein Fold Recognition

ARTICLE INFO

Article Type

Original Research

Authors

Mirzaie M.* PhD

How to cite this article

Mirzaie M. Impact of Hydrophobic Amino Acids in Protein Fold Recognition. Modares Journal of Biotechnology. 2019;10(3):433-440.

ABSTRACT

Aims Prediction of three-dimensional structure from a sequence of amino acids is one of the important problems in structural bioinformatics. Proteins select a special structure among many possible conformations in order of seconds. Levinthal paradox expresses that random searches could not be an effective way to form a native structure and a principal mechanism should be available. Reduced alphabet fewer than 20 have been interested in protein structure because it could sufficiently simplify the protein folding problem. It is generally assumed that the native structure form in the lowest free energy among all conformational states. Therefore, it is needed to design a trustworthy potential function that could discriminate protein fold from incorrect ones.

Materials and Methods Knowledge-based potential functions are one type of energy functions derived from a database of known protein structures. In this study, we introduce a knowledge-based potential and assess the power of five amino acids ALA, LEU, ILE, VAL, and PHE in discrimination of native structure using the reduced model. In the reduced model only the energy between the aforementioned amino acids are calculated.

Finding The reduced model was evaluated using four criteria. The results indicate that there is no significant difference between the 20- amino acid model and the reduced model.

Conclusion The presented model indicates that the power of discrimination of native structure is originally from the interaction between the aforementioned amino acids. Therefore, it needed a new strategy to capture the remaining interactions to improve the power of knowledge-based potential function.

Keywords Protein Structure, Tertiary; Hydrophobic and Hydrophilic Interactions; Protein Folding; Atomic Energy

CITATION LINKS

[1] Principles that govern the folding of protein ... [2] Levinthal's ... [3] The protein folding problem: when will it be ... [4] Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief ... [5] Medium and long-range interaction parameters between amino ... [6] Estimation of effective interresidue contact energies from protein crystal ... [7] An empirical energy potential with a reference state for protein fold and ... [8] Nature of driving force for protein folding: a result from analyzing ... [9] Knowledge-based potentials in protein fold ... [10] Potential energy functions for protein ... [11] Energy-based de novo protein folding by ... [12] The I-TASSER suite: protein structure and function ... [13] Predicting protein-protein interactions between human ... [14] A distance-dependent atomic knowledge-based potential and force for ... [15] Distance-dependent atomic knowledge-based force in ... [16] Delaunay-based nonlocal interactions are ... [17] A computational approach to simplifying the protein folding ... [18] Phase determination in the X-ray diffraction patterns ... [19] Random packings and the structure of simple ... [20] The quickhull algorithm for convex ... [21] Structure validation by Calpha geometry: phi,psi and Cbeta ... [22] Calculation of conformational ensembles potentials of mean ... [23] Energy functions that discriminate X-ray and near-native ... [24] Assembly of protein tertiary structures from fragments with similar ... [25] Improved recognition of native-like protein structures ... [26] Ab initio construction of protein tertiary structures using a ... [27] A novel approach to decoy set generation: designing a physical energy function ... [28] Accurate modeling of protein conformation by automatic segment ... [29] Comparative protein structure modeling by iterative alignment, model ... [30] Structure prediction for CASP7 targets using extensive all-atom refinement ... [31] 3DRobot: Automated Generation of Diverse and Well-packed Protein ... [32] Knowledge-based entropies improve the identification ... [33] New knowledge-based scoring function with inclusion of backbone ...

*Applied Mathematics Department, Mathematical Sciences Faculty, Tarbiat Modares University, Tehran, Iran

*Correspondence

Address: Tarbiat Modares University, Nasr Bridge, Jalal-Al-Ahmad Highway, Tehran, Iran
Phone: +98 (21) 28884267
Fax: +98 (21) 28884267
mirzaie@modares.ac.ir

Article History

Received: July 31, 2018

Accepted: September 25, 2018

ePublished: September 21, 2019

نقش اسیدآمینه‌های هیدروفوب در تشخیص ساختار طبیعی پروتئین

مهدی میرزائی* MD

گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه تربیت مدرس، تهران، ایران

چکیده

اهداف: پیش‌بینی ساختار سوم پروتئین‌ها از توالی اسیدآمینه‌ها یکی از مسایل مهم در بیوانفورماتیک ساختاری است. پروتئین‌ها ساختار طبیعی خود را از میان آرایش‌های متعدد فضایی در کسری از ثانیه انتخاب می‌کنند. پارادوکس *لوینتال* بیان می‌کند جست‌وجوی تصادفی در میان فضای تمام آرایش‌های فضایی ممکن امکان‌پذیر نیست و باید مکانیزمی برای آن وجود داشته باشد. الفبای کمتر از ۲۰ اسیدآمینه در ساختار پروتئین می‌تواند پیچیدگی مساله تاخوردگی پروتئین را تا حدودی کاهش دهد. عموماً فرض می‌شود که ساختار طبیعی در مینیمم انرژی خود شکل می‌گیرد. بنابراین نیاز به یک تابع انرژی مناسب برای ارزیابی ساختار ضروری است.

مواد و روش‌ها: توابع پتانسیل دانش-پایه یکی از انواع توابع انرژی است که از داده‌های ساختار پروتئین‌های شناخته‌شده به دست می‌آید. در این مطالعه یک تابع انرژی دانش-پایه ارایه می‌کنیم و تاثیر اسیدآمینه‌های آلانین، لوسین، ایزولوسین، والین و فنیل‌آلانین در تشخیص ساختار طبیعی را بررسی می‌کنیم. در مدل کاهش‌یافته تنها انرژی بین اسیدآمینه‌های مذکور را در نظر می‌گیریم. **یافته‌ها:** مدل کاهش‌یافته را با چهار معیار ارزیابی کردیم. نتایج نشان می‌دهد که تفاوت معنی‌داری بین نتایج مدل ۲۰- اسیدآمینه و مدل کاهش‌یافته وجود ندارد. **نتیجه‌گیری:** این مدل نشان می‌دهد که توان تابع انرژی به‌دست‌آمده حاصل قدرت انرژی میان این پنج اسیدآمینه است و بنابراین برای افزایش قدرت توابع پتانسیل نیازمند نگرش جدیدی هستیم که بتوانیم به نحو موثرتری از اندرکنش تمام اسیدآمینه‌ها در ساختار بهره ببریم.

کلیدواژه‌ها: ساختار سوم پروتئین، اندرکنش هیدروفوبیک و هیدروفیلیک، تاخوردگی پروتئین، انرژی اتمی

تاریخ دریافت: ۱۳۹۷/۵/۹

تاریخ پذیرش: ۱۳۹۷/۷/۳

*نویسنده مسئول: mirzaie@modares.ac.ir

مقدمه

توالی پروتئین‌ها از ترکیبات مختلفی از ۲۰ نوع اسیدآمینه تشکیل می‌شود. این درشت‌مولکول‌ها، عملکردهای مختلفی از کاتالیزور در واکنش‌های بیوشیمیایی تا موتور مولکولی را بر عهده دارند که ناشی از شکل سه‌بعدی ساختاری آنها است. بنابراین دانستن عملکرد یک پروتئین نیازمند اطلاعات ساختاری آن است. تاخوردگی پروتئین، فرآیندی فیزیکی است که در آن زنجیره پلی‌پپتیدی یا همان توالی اسیدآمینه‌ها به ساختار سه‌بعدی مشخصی تبدیل می‌شود. *کریستین /فینسن* نشان داد که تحت شرایط مشخصی، پروتئینی که از حالت تاخوردگی خارج شده می‌تواند دوباره به حالت تاخوردگی باز گردد و کارایی خود را از سر بگیرد^[1]. این آزمایش، این مطلب مهم را روشن ساخت که توالی اسیدهای آمینه تمامی اطلاعات مورد نیاز برای تعیین ساختار سه‌بعدی پروتئین را در خود دارند. *فینسن* به خاطر این کار ارزشمند، جایزه نوبل شیمی سال ۱۹۷۲ را اخذ کرد.

اما این که چگونه و تحت چه فرآیندی توالی اسیدهای آمینه ساختار پروتئین را مشخص می‌کند، یکی از مهم‌ترین سؤالات حل‌نشده در علوم زیستی است. پروتئین‌ها از ۴۰ تا ۳۵۰۰۰ اسیدآمینه و در نتیجه تعداد زیادی اتم، تشکیل شده‌اند و این انبوه زیاد از اتم‌ها در مدت‌زمان بسیار کمی (گاهی اوقات کسری از ثانیه) در محل مناسب خود قرار می‌گیرند و پروتئین ساختار طبیعی خود را می‌گیرد. سؤال مهم این است که چگونه در مدت‌زمان بسیار کم، پروتئین شکل طبیعی خود را می‌گیرد؟ در سال ۱۹۶۸ *لوینتال*^[2] نشان داد که اگر هر اسیدآمینه، سه درجه آزادی داشته باشد، 3^{30} آرایش فضایی مختلف برای یک پروتئین با ۱۰۱ اسیدآمینه به دست می‌آید و اگر هر تغییر کانفورماسیون، 10^3 بر ثانیه زمان ببرد، آنگاه تقریباً 10^{30} سال (که بیش از عمر هستی است) طول می‌کشد تا تمام حالت‌های ممکن برای آرایش نهایی جست‌وجو شود. او نتیجه گرفت که به ناچار باید مسیری برای فولشدن پروتئین وجود داشته باشد. متأسفانه تاکنون مسیری برای این مساله ارایه نشده است. درواقع مساله فولشدن پروتئین خود سه مساله مختلف است^[3]:

(۱) چگونه ساختار طبیعی پروتئین از طریق نیروهای بین‌اتمی که بر هر اتم در اسیدآمینه‌های مختلف وارد می‌شود، شکل می‌گیرد؟
(۲) مساله پیشگویی ساختار پروتئین و این که چگونه می‌توان ساختار سه‌بعدی پروتئین را از توالی اسیدهای آمینه پیشگویی کرد؟
(۳) پارادوکس *لوینتال* و این که چطور پروتئین در مدت‌زمان بسیار کم، شکل طبیعی و نهایی خود را می‌گیرد؟
تعداد توالی‌های پروتئینی شناخته‌شده به مراتب بیشتر از پروتئین‌هایی است که ساختار سوم آنها به‌صورت تجربی تعیین شده است. تعداد ساختارهای PDB تعیین‌شده و گزارش‌شده در سایت RCSB تا سال ۲۰۱۸ تقریباً ۱۳۱۰۰۰ ساختار است در حالی که بیش از ۱۱۰۰۰۰۰۰۰ توالی در UniProt تاکنون گزارش شده است. تعیین ساختار پروتئین‌ها به‌صورت تجربی به‌وسیله اشعه X یا روش NMR بسیار زمان‌بر و با هزینه بالا است. تقریباً هزینه تعیین تجربی یک ساختار پروتئین ۱۰۰۰۰۰ دلار برآورد می‌شود، در حالی که هزینه تعیین توالی یک پروتئین تقریباً ۱۰۰۰ دلار است^[4]. هزینه بالای زمانی و مالی برای تعیین تجربی ساختار که باعث فاصله زیاد میان تعداد توالی‌های شناسایی‌شده با تعداد ساختارهای تعیین شده است، اهمیت ابزارهای کارآمد محاسباتی برای تعیین ساختار سوم پروتئین‌ها را روشن می‌سازد. راهکارهای موجود برای پیشگویی ساختار پروتئین، عموماً بر مبنای این فرضیه ترمودینامیکی است که ساختار طبیعی سه‌بعدی پروتئین در وضعیتی شکل می‌گیرد که انرژی آن در مینیمم گلوبال خود قرار داشته باشد. بنابراین با توجه به این نظریه، تمام مطالعات روی ساختار پروتئین‌ها، چه پیشگویی ساختار و چه شبیه‌سازی آن کاملاً وابسته به کارگیری تابع انرژی مناسب برای پروتئین است. در حال حاضر دو نوع متفاوت از توابع انرژی موجود است که نوع اول توابع انرژی هستند که در آن انرژی بین دو اتم براساس قوانین و روابط فیزیکی محاسبه می‌شود و نوع دوم توابع انرژی هستند که در آن انرژی بین دو اتم براساس تابع

انرژی کل ساختار لحاظ می‌کنند. در سال ۲۰۰۹ تابع انرژی با چهار اتم تایپ کربن، اکسیژن، نیتروژن و گوگرد که در آن تنها اندرکنش‌های مستقیم که با روش مفروش‌سازی دلونی ساختار به دست می‌آید، معرفی کردیم و در سال ۲۰۱۲ مدل را تعمیم دادیم^[14]. در سال ۲۰۱۴ این مدل با ۱۶۷ اتم تایپ و در نظر گرفتن تنها اندرکنش‌های با فاصله توالی بیشتر از ۵ در تشخیص ساختار طبیعی به کار گرفته شد^[16]. در این مقاله نشان داده شد که حذف اندرکنش‌های نزدیک در توالی، یعنی اندرکنش‌هایی با فاصله توالی کمتر از ۵ منجر به نتایج بهتری در تشخیص فولد صحیح می‌شود. دلایل متعددی وجود دارد که محققین را علاقه‌مند به مدلی با الفبای کاهش‌یافته از اسیدآمین‌ها می‌کند. از جمله مهم‌ترین دلایل، کاهش پیچیدگی در ساختار پروتئین است که می‌تواند سرخ‌هایی را برای مساله تاخوردگی پروتئین و توجیهی برای پارادوکس لوینتال ارائه دهد. نتایج تجربی متعددی از انتخاب سه، چهار یا بیشتر از اسیدآمین‌ها برای معرفی مدل کاهش‌یافته از ساختار پروتئین‌ها وجود دارد که در آنها معیارهای مختلفی برای ساخت مدل کاهش‌یافته وجود داشته است. مثلاً وانگ و همکاران مدل کاهش‌یافته‌ای مبتنی بر استراتژی کمترین عدم تطابق ارائه دادند^[17]. در این استراتژی اسیدآمین‌ها با ویژگی‌های متفاوت در یک گروه قرار می‌گیرند. روش آنها مجموعه کاهش‌یافته را IAGEK گزارش می‌کرد. گرچه روش‌هایی برای معرفی الفبای کاهش‌یافته از اسیدآمین‌ها ارائه شده، اما تاکنون روشی برای ساخت مدلی کاهش‌یافته در تشخیص فولد صحیح معرفی نشده است. در این مقاله نشان می‌دهیم نتایج مدل کاهش‌یافته به‌دست‌آمده از ۵ اسیدآمین VALIF در تشخیص ساختار طبیعی پروتئین بسیار نزدیک به مدل ۲۰- اسیدآمین‌ای است به‌گونه‌ای که تفاوت معنی‌داری میان نتایج مدل کاهش‌یافته با مدل ۲۰- اسیدآمین‌ای وجود ندارد. برای ارزیابی مدل از چهار معیار تعداد ساختارهای طبیعی تشخیص داده شده، Z-Score انرژی ساختارهای موجود در یک مجموعه داده، RMSD ساختار با کمترین انرژی و ضریب همبستگی میان انرژی ساختارها و RMSD ساختارها استفاده کردیم.

مواد و روش‌ها

عموماً فرض می‌شود که ساختار سه‌بعدی پروتئین در مینیمم انرژی قرار دارد. برای محک‌زدن تابع انرژی معرفی‌شده، تابع انرژی باید قدرت تشخیص ساختار طبیعی را از میان ساختارهای غلط داشته باشد تا بتوان از آن در پیشگویی ساختار نیز استفاده کرد. برای بررسی این مطلب، انرژی ساختار طبیعی و ساختارهای غلط را محاسبه می‌کنند و انتظار دارند که انرژی ساختار طبیعی از ساختارهای غلط کمتر باشد. همان‌طور که می‌دانیم اتم‌های تشکیل‌دهنده پروتئین‌ها شامل کربن، اکسیژن، نیتروژن، گوگرد و هیدروژن هستند. اما انتخاب اتم تیپ‌ها بر مبنای اتصال و ماهیت شیمیایی و محل حضور آنها (در زنجیره جانبی یا بدنه) است. زیرا

بولتزمن و از روی ساختار پروتئین‌های شناخته‌شده به دست می‌آید. به توابع انرژی نوع اول، توابع انرژی "فیزیکی" و به توابع انرژی نوع دوم توابع انرژی "آماري" یا "دانش- پایه" گفته می‌شود. در توابع انرژی فیزیکی، از یک میدان نیروی مکانیک مولکولی استفاده می‌شود که در آن پارامترهای میدان نیرو از اطلاعات مولکول‌های کوچک به دست می‌آید. این نوع توابع انرژی، مجموعی از انرژی‌های الکترواستاتیکی و واندروالسی، پیوندی و انرژی حاصل از تغییر زاویه پیوندی و چرخش زاویه دایه‌درال هستند. اما توابع انرژی "آماري" یا "دانش- پایه" همه میان‌کنش‌های فیزیکی را یک‌جا و در یک جمله جمع می‌کنند. از آنجایی که این نوع توابع انرژی از طریق ساختار سوم‌های پروتئینی شناخته‌شده به دست می‌آیند، با غنی‌شدن پایگاه داده‌های ساختاری، بیشتر مورد توجه قرار گرفته‌اند. به‌کارگیری تابع انرژی دانش- پایه به کار ارزشمند *Tanaka* و *Shiraga* در سال ۱۹۷۶ باز می‌گردد که برای اولین بار موفق شدند قدرت اندرکنش میان دو اسیدآمین را با فراوانی تعداد تماس میان آنها در ساختار پروتئین‌های طبیعی مشخص‌شده با اشعه X اندازه‌گیری کنند^[5]. بعداً *میاوا* و *جرینگان* اندرکنش میان دو اسیدآمین را به کمک تقریب شبه‌شیمیایی فرموله کردند^[6,7]. در این مدل، هر اسیدآمین با مرکز زنجیر جانبی نمایش داده می‌شود. یک جفت اسیدآمین در تماس در نظر گرفته می‌شود، اگر فاصله فضایی آن کمتر از ۶ آنگستروم باشد، قدرت اندرکنش میان اسیدآمین‌های در تماس در این مدل با ماتریسی ۲۰ در ۲۰ نمایش داده می‌شود که شامل میانگین انرژی بین جفت اسیدآمین‌های مختلف است. لی و همکاران این ماتریس را با استفاده از تجزیه مقدار ویژه تحلیل کردند^[8]. نتایج آنها نشان داد که علی‌رغم ماهیت پیچیده اندرکنش‌ها در ساختار سوم پروتئین‌ها، نیروی اصلی در پروتئین‌ها حاصل اندرکنش‌های هیدروفوبیک در ساختار است. در نوع دیگری از توابع انرژی دانش- پایه، انرژی وابسته به متغیر فاصله است که مقادیر انرژی به کمک عکس رابطه بولتزمن از توزیع احتمال فاصله جفت اتم‌ها در فضای مرجع و پروتئین‌های موجود در پایگاه داده به دست می‌آید^[9]. با غنی‌تر شدن پایگاه داده‌های ساختار پروتئینی، توابع انرژی دانش- پایه از دقت بیشتری برخوردار شده‌اند و بنابراین بیشتر نیز مورد توجه قرار گرفته‌اند. امروزه با توجه به دقت بالا و سرعت محاسبات مناسب از این نوع توابع در طراحی پروتئین^[10]، شبیه‌سازی تاخوردگی پروتئین^[11] و پیش‌بینی ساختار سوم^[12] استفاده می‌شود. باید توجه داشت که گرچه ساختار سوم پروتئین، اطلاعات مهمی در زمینه عملکرد به ما می‌دهد اما دانستن اندرکنش پروتئین- پروتئین نیز کمک شایانی به درک بیماری‌ها می‌کند. ابزارهای محاسباتی مختلفی با هدف پیش‌بینی اندرکنش پروتئینی میان میزبان و ویروس ارائه شده است^[13].

تاکنون توابع متعددی برحسب وضعیت مرجع و انتخاب تعداد و نوع اتم تایپ معرفی شده‌اند^[14, 15]. معمولاً در این نوع توابع دو اتم تایپ که فاصله فضایی آنها از مقدار آستانه‌ای (معمولاً ۷ آنگستروم) کمتر باشد را در تماس در نظر گرفته و انرژی آن را در

به دست آوردن تابع انرژی آماری Top500H است. این پایگاه داده شامل ۵۰۰ پروتئین با درجه تفکیک حداقل ۱/۸ آنگستروم است. به علاوه هر دو پروتئین در این پایگاه داده حداقل ۶۰٪ توالی یکسان دارند [21]. انرژی بین جفت اتم i و j که در فاصله فضایی d از یکدیگر قرار گرفته اند، با توجه به رابطه ارایه شده توسط سیل به صورت زیر است [22]:

$$\Delta E^{ij}(d) = RT \ln([1 + M_{ij}\sigma] - \ln[1 + M_{ij}\sigma \left(\frac{f_{ij}(d)}{f_{xx}(d)}\right)])$$

که در آن T برابر ۲۹۳ کلوین و KT برابر ۵۸۲٪ است. M_{ij} تعداد جفت اتم های i و j است. یعنی:

$$M_{ij} = \sum_{d=1}^{29} f(i, j, d)$$

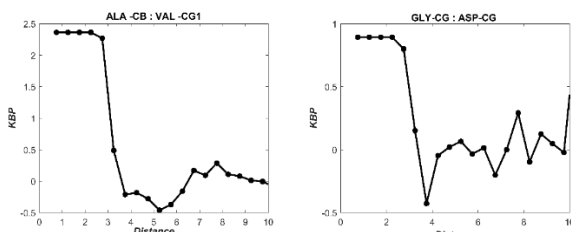
که در آن $f(i, j, d)$ تعداد جفت اتم های i و j که با هم در تماس هستند و در فاصله فضایی d از هم قرار دارند. σ را همان گونه که در مطالعه سیل [22] بیان شده است برابر با $\frac{1}{50}$ در نظر می گیریم. $f_{ij}(d)$ برابر است با فراوانی نسبی رخداد جفت اتم های i و j که با هم در تماس هستند و در فاصله فضایی d از هم قرار دارند؛ یعنی:

$$f_{ij}(d) = \frac{f(i, j, d)}{M_{ij}}$$

و $f_{xx}(d)$ برابر است با فراوانی نسبی رخداد تمام جفت اتم ها که با هم در تماس هستند و در فاصله فضایی d از هم قرار دارند؛ یعنی:

$$f_{xx}(d) = \frac{\sum_{i=1}^{167} \sum_{j=1}^{167} f(i, j, d)}{\sum_{i=1}^{167} \sum_{j=1}^{167} \sum_{d=1}^{29} f(i, j, d)}$$

برای محاسبه انرژی در ابتدا باید توزیع تجربی فاصله میان جفت اتم ها را با دسته بندی فاصله به دست آورد. برای این منظور فاصله بین دو اتم را از ۷۰٪ تا ۱۵۰ آنگستروم در دسته های ۵٪ آنگستروم تقسیم بندی می کنیم و رخداد هر اتم در فاصله کمتر از ۷۰٪ و بیشتر از ۱۵۰ آنگستروم را نادیده می گیریم. همچنین به منظور جلوگیری از وقوع فراوانی صفر در برخی بین ها، به فراوانی ها عدد یک را اضافه می کنیم. اندرکنش میان اتم های موجود در یک اسید آمینه را نیز نادیده می گیریم. دو نمونه از توابع انرژی به دست آمده در نمودار ۱ بیان شده است.



نمودار (۱) تابع انرژی دانش- پایه به دست آمده بین دو جفت اتم تایپ مشخص که در عنوان شکل قید شده است. محور افقی فاصله و محور عمودی انرژی است.

انرژی کل یک ساختار که با G نمایش می دهیم به صورت زیر محاسبه می شود:

$$G = \frac{1}{2} \sum_{i,j} \Delta E_k^{ij}(l)$$

اتم ها از نظر پیرامون، نحوه اتصال و ماهیت شیمیایی با یکدیگر متفاوتند. با در نظر گرفتن چنین تفاوت هایی تعداد کل اتم تیپ ها در ۲۰ اسید آمینه برابر با ۱۶۷ است. در این مطالعه ۱۶۷ نوع مختلف از اتم را برای به دست آوردن تابع انرژی در نظر می گیریم. برای به دست آوردن انرژی نیازمند تعریف مناسبی از تماس میان اتم ها هستیم.

تماس بین اتم ها

یک نگرش رایج آن است که دو اتم را در تماس در نظر می گیرند، هر گاه فاصله فضایی آنها از یک عدد ثابت کمتر باشد، بنابراین هر دو اتمی که فاصله آنها از یک عدد ثابت کمتر باشد، دارای میان کنش هستند. این مطلب همیشه درست نیست زیرا ممکن است حائلی میان این دو اتم که بسیار به هم نزدیک هستند از طریق سایر اتم ها ایجاد شده باشد و بنابراین با اشغال فضای میانی بین این دو اتم توسط اتم های دیگر، میان کنش مستقیمی بین دو اتم مزبور وجود نداشته باشد. دیگرام های ورونی و در نتیجه مفروض سازی دلونی این معضل را حل می کند. درواقع مطالعه ساختارهای نامنظم با روش آماری و هندسی اولین بار توسط برنال بررسی شد [18]. برنال پیشنهاد بررسی ساختار را با تحلیل های آماری روی چهاروجهی های نامنظم که حاصل مفروض سازی فضای سه بعدی است، ارایه کرد. بعداً شخصی به نام فینی الگوریتم را در حالت ورونی پیاده سازی و عملی ساخت [19]. درواقع ناحیه ورونی p ، یک چندوجهی محدب است به طوری که تمام نقاط در این چندوجهی نزدیکترین نقاط به p هستند. مفروض سازی دلونی S که با $DT(S)$ نشان می دهیم، دوگان گراف نمودار ورونی $VD(S)$ است. یعنی هر ناحیه ورونی متناظر با راسی است و دو راس با هم مجاور هستند، اگر نواحی ورونی نظیرشان در یالی مشترک باشند. روش های مختلفی برای به دست آوردن مفروض سازی دلونی وجود دارد. یکی از روش های شناخته شده روش $Qhull$ است [20]. در محاسبات مطالعه حاضر از نرم افزار متلب استفاده شد که این نرم افزار از روش $Qhull$ استفاده می کند. با نشان دادن موقعیت هر اتم در فضا به عنوان یک نقطه، ساختار پروتئین را به عنوان مجموعه ای از نقاط در فضای سه بعدی در نظر می گیریم. حال هر چندوجهی ورونی، ناحیه ای در فضا حول اتم A است به طوری که تمام نقاط در این ناحیه نزدیکترین نقاط به A هستند. از طرفی با توجه به روش ساختن مفروض سازی دلونی، هر ۴ اتمی که چندوجهی ورونی آنها یک وجه مشترک داشته باشند، تشکیل یک چهاروجهی دلونی می دهند. بنابراین مفروض سازی دلونی چهاروجهی هایی را به ما می دهد که در آن راس های هر چهاروجهی نزدیکترین همسایه به همدیگر هستند. بنابراین گوئیم دو اتم با هم در تماسند، اگر دو راس یک چهاروجهی دلونی در ساختار پروتئین باشند.

انرژی بین اتمی

بعد از تعیین در تماس بودن دو اتم i و j با استفاده از الگوریتم دلونی و تعیین فاصله فضایی و فاصله توالی آنها نوبت به محاسبه انرژی آماری بین جفت اتم i و j می رسد. پایگاه داده مورد استفاده برای

صورتی که قادر به تشخیص ساختار طبیعی نبودیم، انتظار می‌رود که ساختار انتخاب شده توسط سیستم نمره‌دهی ما RMSD کمی با ساختار طبیعی داشته باشد. در حالت ایده‌آل RMSD ساختار با مینیمم انرژی صفر است، یعنی آن ساختار، ساختار طبیعی است. ساختارهایی با RMSD کمتر از ۲ به عنوان ساختارهای شبه طبیعی شناخته می‌شوند. بنابراین اگر مدل در تشخیص ساختار طبیعی موفق نباشد اما ساختار شبه طبیعی را به عنوان ساختار با مینیمم انرژی گزارش کند، نشان می‌دهد که مدل موفق بوده است.

CC: ضریب همبستگی میان نمره و RMSD: تشخیص ساختار طبیعی در میان ساختارهای غلط یکی از معیارهای لازم برای بررسی کارایی روش ارایه شده است، اما کافی نیست، زیرا همبستگی و رابطه میان انرژی یک ساختار غلط و میزان شباهت ساختاری آن را با ساختار طبیعی نشان نمی‌دهد. درواقع نظریه قیفی شکل بودن تابع انرژی در فولدینگ پروتئین چنین همبستگی را ایجاد می‌کند. انتظار داریم هر چه RMSD یک ساختار از ساختار طبیعی بیشتر باشد، انرژی آن نیز بیشتر باشد. بنابراین باید ضریب همبستگی میان انرژی و RMSD نزدیک به یک باشد.

یافته‌ها

در ابتدا به بررسی نتایج مدل ۲۰- اسیدآمینه‌ای می‌پردازیم. در این مدل تمام اسیدآمینه‌ها و اندرکنش میان اتم‌های آنها که از روش دلونی به دست آمده، در نظر گرفته می‌شود. همان‌طور که در جدول ۲ دیده می‌شود معیارهای ارزیابی مدل به خوبی تابع انرژی را تایید می‌کنند. ۸۴٪ ساختارهای طبیعی در این مدل تشخیص داده شده‌اند و میانگین Z-Score نیز بزرگ‌تر از ۳ با انحراف معیار ۰/۷۹ است. به علاوه میانگین RMSD ساختار با مینیمم انرژی کمتر از یک است که عدد قابل توجهی است. برای مجموعه دیکوی‌های 4-moulder, hg-structal, ig-structal-hire, state-reduced و 3DROBOT که اکثر ساختارها در محدوده یک تا ۵ آنگستروم هستند ضریب همبستگی پیرسون بین انرژی و RMSD عدد قابل قبولی است. اما باید توجه داشت که انحراف معیار برای PC برابر ۰/۲۸ و برای RMSD برابر ۰/۹۹ است و این به دلیل پراکندگی در محدوده RMSD مجموعه دیکوی‌ها است. بنابراین معیارهای ارزیابی تابع انرژی مدل ۲۰- اسیدآمینه را تایید می‌کند.

و/نگ و همکاران الفبای کاهش یافته‌ای با استراتژی عدم تطابق کمینه ارایه دادند^[17]. در این استراتژی اسیدآمینه‌ها با ویژگی‌های متفاوت در یک گروه قرار می‌گیرند. این گروه مجموعه شامل پنج اسیدآمینه گلوتامیک اسید، گلیسین، ایزولوسین، آلانین و لایزین را ارایه دادند. نتایج مدل حاصل از کاهش به این پنج اسیدآمینه در جدول ۳ بیان شده است. همان‌طور که در جدول مشاهده می‌شود قدرت تشخیص این مدل برای مجموعه دی‌کوی‌های lmds و fisa برابر صفر است. در بقیه دیکوی‌ها نیز نهایتاً قدرت تشخیص ۵۰٪ دیده می‌شود. میانگین Z-score برای تمام دیکوی‌ها از ۳ کمتر

که در آن مجموع بالا روی تمام جفت اتم‌هایی است که با روش دلونی با هم در تماس بوده‌اند. به علاوه جفت اتم با فاصله فضای بیش از ۶ آنگستروم و فاصله توالی بیش از ۵ را نادیده گرفته‌ایم. در مدل ۲۰- اسیدآمینه‌ای تمام اسیدآمینه‌ها را برای محاسبه انرژی کل در نظر می‌گیریم، در حالی که در مدل کاهش یافته تنها اسیدآمینه‌های مشخصی را در نظر می‌گیریم. مثلاً در کاهش به پنج اسیدآمینه F, I, L, A و V تنها این ۵ اسیدآمینه و انرژی بین اتم‌هایی از این پنج اسیدآمینه که با هم اندرکنش دارند، را در نظر می‌گیریم. برای ارزیابی مدل، از مجموعه دیکوی‌های جدول ۱ استفاده می‌کنیم که با روش‌های مختلفی به دست آمده‌اند.

جدول ۱) مجموعه داده دیکوی: ستون اول نام مجموعه داده، ستون دوم تعداد پروتئین‌ها، ستون سوم محدوده RMSD و ستون چهارم محدوده تعداد ساختارها را بیان می‌کند.

#Structures	RMSD Range	#Protein	Decoy Set
۶۸۵-۶۳۰	۰/۸-۱/۵	۷	4state-reduced ^[23]
۲۴۱۲-۹۷۱	۳/۵-۸/۵	۶	fisa-casp3 ^[24]
۵۰۰	۲/۷۵-۴/۷۵	۴	fisa ^[25]
۲۰۰۰	۴/۷۵-۷	۸	lattice-ssfit ^[26]
۵۰۰-۲۱۵	۲/۵-۴/۷۵	۱۰	lmds ^[27]
۲۹	۰/۵-۱۲	۲۹	hg-structal ^[28]
۱۹	۰/۷-۲	۲۰	ig-structal-hires ^[28]
۵۹	۰/۷-۲	۶۱	ig-structal ^[28]
۴۰۰-۳۰۰	۱/۸۵-۶/۲۵	۲۰	moulder ^[29]
۱۲۰	۰/۵-۱۱	۵۹	ROSETTA ^[30]
۳۰۰	۰/۲-۱/۴	۲۰۰	3DRobot ^[31]

معیارهای ارزیابی مدل

چهار معیار برای ارزیابی مدل مطالعه حاضر در نظر گرفتیم که به شرح زیر است:

Top1: تعداد ساختارهای طبیعی پروتئین که انرژی آنها از انرژی سایر ساختارها کمتر به دست آمده است.

Z-score: ساختار طبیعی برای یک پروتئین همراه با ساختارهای غلط آن که برابر است با:

$$Z - score = \frac{(score_{decoys}) - score_{native}}{\sigma_{decoys}}$$

که در آن $score_{native}$ انرژی به دست آمده برای ساختار طبیعی است و $\langle score_{decoys} \rangle$ و σ_{decoys} به ترتیب میانگین و انحراف معیار انرژی ساختارهای غلط است. درواقع، Z-score تعداد انحراف معیار میان انرژی ساختار طبیعی و میانگین انرژی ساختارهای غلط است. در توزیع نرمال، ۹۹٪ داده‌ها با فاصله ۳ انحراف معیار از میانگین قرار دارند. با این فرض که توزیع انرژی ساختارها نرمال است، Z-score برابر یا بیشتر از ۳ به این معنی است که ساختار طبیعی در میان ۱٪ بهترین ساختارهای غلط قرار گرفته و این نشان می‌دهد که تابع انرژی معرفی شده در تشخیص ساختار طبیعی موفق بوده است.

RMSD: از آنجایی که ساختارهای غلط ساخته شده به گونه‌ای هستند که در میان آنها ساختار شبه طبیعی وجود دارد، بنابراین در

است. میانگین RMSD نیز عدد بزرگی است. ضریب همبستگی پیرسون نیز مدل را تایید نمی‌کند.

جدول ۲) نتایج ارزیابی مدل ۲۰- اسیدآمینة روی مجموعه دیکوی‌ها؛ ستون اول نشان‌دهنده مجموعه دیکوی و ستون‌های دوم تا چهارم نتایج معیارهای ارزیابی مدل را نشان می‌دهند. ستون آخر تعداد پروتئین در هر مجموعه داده است.

#Target	PC	RMSD	Z-score	Top1	Decoy Set
۷	۰/۵۹	۰/۰۰	۳/۸۶	۷	4state-reduced
۸	۰/۰۴	۰/۰۰	۴/۴۱	۸	lattice-ssfit
۱۰	۰/۰۷	۱/۷۸	۴/۰۵	۷	Lmds
۴	۰/۲۳	۰/۸۸	۲/۲۶	۳	Fisa
۶	۰/۰۲	۱/۵۳	۴/۱۷	۵	fisa-Casp3
۲۰	۰/۶۵	۰/۰۰	۴/۰۵	۲۰	ig-structural-hire
۲۹	۰/۷۰	۰/۳۲	۲/۸۳	۲۲	hg-structural
۶۱	۰/۴۷	۰/۳۰	۳/۲۴	۵۱	ig-structural
۲۰	۰/۵۸	۰/۱۶	۴/۶۵	۱۹	moulder
۵۹	۰/۲۹	۳/۱۱	۲/۵۹	۳۵	ROSETTA
۲۰۰	۰/۶۲	۰/۲۰	۴/۲۹	۱۷۹	3DROBOT
۴۲۴	۰/۳۷±۰/۲۸	۰/۷۵±۰/۹۹	۳/۶۶±۰/۷۹	۳۵۶	#Total

جدول ۳) نتایج ارزیابی مدل ۵- اسیدآمینة حاصل از اسیدآمینة‌های گلايسين، ايزولوسين، گلوتاميك‌اسيد، لايزين و آلانين روی مجموعه دیکوی‌ها

#Target	PC	RMSD	Z-score	Top1	Decoy Set
۷	۰/۲۲	۴/۵۱	۱/۵۳	۲	4state-reduced
۸	۰/۰۲	۷/۶۳	۱/۸۵	۲	lattice-ssfit
۱۰	۰/۰۷	۱۰/۵۰	۱/۴۰	۳	Lmds
۴	۰/۰۱	۱/۱۸	۱/۱۹	۰	Fisa
۶	-۰/۰۱	۲/۵۱	۱/۵۵	۰	fisa-Casp3
۲۰	۰/۳۹	۱/۷۴	۱/۵۸	۹	ig-structural-hire
۲۹	۰/۳۰	۲/۰۴	۱/۵۸	۱۰	hg-structural
۶۱	۰/۲۰	۷/۲۵	۱/۰۴	۱۰	ig-structural
۲۰	۰/۳۰	۲/۹۶	۲/۵۷	۱۳	moulder
۵۹	۰/۰۹	۴/۵۱	۱/۱۱	۹	ROSETTA
۲۰۰	۰/۳۶	۷/۶۳	۲/۰۱	۷۳	3DROBOT
۴۲۴	۰/۱۷±۰/۱۵	۴/۶۹±۳/۰۷	۱/۵۸±۰/۴۲	۱۳۱	#Total

طبیعی در این مدل تشخیص داده شده‌اند و میانگین Z-Score نیز بزرگ‌تر از ۳ است. بنابراین معیارهای ارزیابی تابع انرژی این مدل کاهش‌یافته ۵- اسیدآمینة‌ای را تایید می‌کند و نتایج نشان می‌دهد که معیارها بسیار نزدیک به مدل ۲۰- اسیدآمینة است. همچنین آزمون ویلکاکسون روی مقادیر PC، RMSD، Z-score برای دو گروه یعنی برای مدل ۵- اسیدآمینة بالا و مدل ۲۰- اسیدآمینة، P-value‌هایی با مقادیر ۰/۲۶، ۰/۱۱ و ۰/۵۱ تاکید می‌کند که تفاوت معنی‌داری بین نتایج این دو مدل وجود ندارد.

به‌منظور بررسی بیشتر، مدل کاهش‌یافته ۵- اسیدآمینة‌ای با اسیدآمینة‌های آسپاراژین، گلوتامیک‌اسید، آسپارتیک‌اسید، لایزین، آرژنین، گلوتامین را ارزیابی کردیم. نتایج این مدل در جدول ۵ بیان شده است. این اسیدآمینة‌ها، قطبی هستند. همان‌طور که در جدول مشاهده می‌شود قدرت تشخیص این مدل کاهش‌یافته نزدیک صفر است و هیچ‌کدام از معیارهای ارزیابی این مدل را تایید نمی‌کند، در حالی که نتایج به‌دست‌آمده با مدل کاهش‌یافته با پنج اسیدآمینة هیدروفوب لوسین، والین، ایزولوسین، فنیل‌آلانین و آلانین بسیار نزدیک به مدل ۲۰- اسیدآمینة بود.

جدول ۴) نتایج ارزیابی مدل ۵- اسیدآمینة حاصل از اسیدآمینة‌های فنیل‌آلانین، ايزولوسين، والين، لوسين و آلانين روی مجموعه دیکوی‌ها

#Target	PC	RMSD	Z-score	Top1	Decoy Set
۷	۰/۴۳	۲/۶۱	۳/۳۹	۴	4state-reduced
۸	۰/۰۰	۰/۰۰	۳/۵۴	۸	lattice-ssfit
۱۰	۰/۰۵	۳/۱۳	۳/۶۱	۵	Lmds
۴	۰/۲۰	۱/۳۴	۳/۹۷	۳	Fisa
۶	۰/۰۲	۱/۵۶	۲/۶۱	۵	fisa-Casp3
۲۰	۰/۶۰	۰/۱۳	۳/۱۱	۱۸	ig-structural-hire
۲۹	۰/۶۵	۰/۵۲	۲/۶۵	۲۲	hg-structural
۶۱	۰/۳۶	۰/۹۸	۲/۴۲	۳۴	ig-structural
۲۰	۰/۳۷	۰/۱۸	۵/۰۷	۱۹	moulder
۵۹	۰/۲۷	۳/۴۲	۲/۵۷	۳۱	ROSETTA
۲۰۰	۰/۴۷	۰/۳۵	۳/۶۷	۱۷۳	3DROBOT
۴۲۴	۰/۳۰±۰/۲۳	۱/۲۹±۱/۲	۳/۲۲±۰/۷۸	۳۲۲	#Total

جدول ۵) نتایج ارزیابی مدل ۵- اسیدآمینة حاصل از اسیدآمینة‌های آرژنین، آسپاراژین، گلوتامیک‌اسید، آسپارتیک‌اسید، لایزین و گلوتامین روی مجموعه دیکوی‌ها

#Target	PC	RMSD	Z-score	Top1	Decoy Set
۷	۰/۱۷	۵/۲۶	۰/۷۲	۰	4state-reduced
۸	۰/۰۱	۹/۷۴	۱/۱۱	۰	lattice-ssfit
۱۰	۰/۰۳	۷/۲۱	۰/۳۹	۰	Lmds
۴	۰/۰۳	۵/۰۴	۰/۱۷	۰	Fisa
۶	۰/۰۰	۱۱/۱۷	۰/۰۲	۰	fisa-Casp3
۲۰	۰/۲۱	۱/۹۰	۰/۹۱	۳	ig-structural-hire
۲۹	۰/۱۲	۴/۴۵	۰/۱۲	۱	hg-structural
۶۱	۰/۰۹	۲/۱۸	۰/۰۸	۵	ig-structural
۲۰	۰/۳۵	۸/۳۹	۱/۱۶	۱	moulder
۵۹	۰/۰۱	۸/۱۷	۰/۰۶	۴	ROSETTA
۲۰۰	۰/۰۹	۶/۷۸	۰/۸۹	۴	3DROBOT
۴۲۴	۰/۱۰±۰/۱۰	۶/۴۰±۲/۹۵	۰/۵۰±۰/۴۷	۱۸	#Total

حال به بررسی نتایج حاصل از مدل کاهش‌یافته به پنج اسیدآمینة هیدروفوب آلانین، فنیل‌آلانین، ايزولوسين، والين و لوسين می‌پردازیم. همان‌طور که در جدول ۴ مشاهده می‌شود، نتایج این مدل بسیار به مدل ۲۰- اسیدآمینة نزدیک است. نتایج این مدل در تشخیص ساختار طبیعی روی دیکوی‌های Fisa، Lattice-ssfit، fisa-casp3، hg-structural و moulder با مدل ۲۰ اسیدآمینة برابری می‌کند. به‌علاوه نتایج روی مجموعه دیکوی‌های ig-structural-hire، ROSETTA و 3DROBOT بسیار شبیه به نتایج مدل ۲۰- اسیدآمینة است. میانگین RMSD نیز کوچک‌تر از ۱/۳ است که از ۲ کوچک‌تر است. ضریب همبستگی بین انرژی و RMSD نیز برای مجموعه دیکوی‌های 4-state-reduced، ig-structural-hire که شامل ساختارهایی در محدوده یک تا ۵ آنگستروم هستند، عدد قابل قبولی است. ۷۵٪ ساختارهای

- 2- Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. *Proc Natl Acad Sci USA*. 1992;89(1):20-2.
- 3- Dill KA, Ozkan SB, Weiik TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved?. *Curr Opin Struct Biol*. 2007;17(3):342-6.
- 4- Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K4, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief Bioinform*. 2018;19(3):482-94.
- 5- Tanaka S, Scheraga HA. Medium and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*. 1976;9(6):945-50.
- 6- Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*. 1985;18(3):534-52.
- 7- Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins*. 1999;36(3):357-69.
- 8- Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys Rev Lett*. 1997;79:765-8.
- 9- Mirzaie M, Sadeghi M. Knowledge-based potentials in protein fold recognition. *J Paramed Sci*. 2010;1(4):65-75.
- 10- Boas FE, Harbury PB. Potential energy functions for protein design. *Curr Opin Struct Biol*. 2007;17(2):199-204.
- 11- Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci USA*. 1999;96(5):2025-30.
- 12- Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: protein structure and function prediction. *Nat Methods*. 2015;12(1):7-8.
- 13- Emamjomeh A, Goliaei B, Zahiri J, Ebrahimpour R. Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol Biosyst*. 2014;10(12):3147-54.
- 14- Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M. A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys. *Proteins*. 2009;77(2):454-63.
- 15- Mirzaie M, Sadeghi M. Distance-dependent atomic knowledge-based force in protein fold recognition. *Proteins*. 2012;80(3):683-90.
- 16- Mirzaie M, Sadeghi M. Delaunay-based nonlocal interactions are sufficient and accurate in protein fold recognition. *Proteins*. 2014;82(3):415-23.
- 17- Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. *Nat Struct Biol*. 1999;6(11):1033-8.
- 18- Bernal JD. Phase determination in the X-ray diffraction patterns of complex crystals and its application to protein structure. *Nature*. 1952;169(4311):1007-8.
- 19- Finney JL. Random packings and the structure of simple liquids II. The molecular geometry of simple liquids. *Proc Royal Soc A*. 1970;319(1539):495.
- 20- Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. *ACM Trans Math Softw*. 1996;22(4):469-83.
- 21- Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha

از محدودیت‌های مطالعه حاضر می‌توان به وابستگی تابع پتانسیل به مجموعه داده آموزش اشاره کرد. با افزایش ساختار سوم‌های شناسایی‌شده از طریق روش‌های تجربی این محدودیت به مرور مرتفع می‌شود. پیشنهاد می‌شود برای مطالعات آتی، این روش به‌صورت محاسباتی در روش‌های رایج در پیش‌بینی ساختار سوم ارزیابی شود.

نتیجه‌گیری

نتایج مدل کاهش‌یافته نشان می‌دهد که علی‌رغم استفاده از تنها ۵ نوع اسیدآمینه ذکرشده که به‌طور متوسط ۲ اندرکنش‌های ساختار را شامل می‌شود، می‌توان به نتایجی بسیار نزدیک به نتایج مدل ۲۰- اسیدآمینه در تشخیص ساختار طبیعی پروتئین رسید. این مطلب بیان می‌کند که انرژی بین سایر اسیدآمینه‌ها نقش موثری در تشخیص ساختار طبیعی ندارد. بنابراین برای آن که بتوانیم از قدرت اندرکنش میان بقیه اسیدآمینه‌ها نیز به نحو موثرتری استفاده کنیم، باید استراتژی مناسب‌تری در تعریف انرژی میان اسیدآمینه برگزینیم. در این زمینه کارهای جدیدی در حال پیگیری است. اخیراً *سانکار* و همکاران^[32] در مطالعه‌ای نشان دادند که اندرکنش‌های میان اسیدآمینه‌های قطبی در آنتروپی نقش دارند. همان‌طور که می‌دانیم انرژی آزاد یک پروتئین مجموعی از انرژی و آنتروپی است. *سانکار* و همکاران انرژی آزاد یک ساختار را با انرژی دانش- پایه و آنتروپی دانش- پایه تقریب زدند. آنتروپی دانش- پایه براساس فراوانی تغییرات تماس میان اسیدآمینه‌ها در یک مجموعه پروتئینی به دست آمد. به‌علاوه نتایج آنها تایید کرد که پایداری پروتئین حاصل توازن میان انرژی حاصل از اندرکنش‌های هیدروفوبیک و آنتروپی حاصل از اندرکنش‌های قطبی و باردار است. بنابراین یکی از استراتژی‌های ممکن برای اصلاح توابع دانش- پایه می‌تواند اضافه کردن جمله آنتروپی باشد. *وانگ* و همکاران^[33] اخیراً تابع پتانسیل جدیدی ارائه دادند که حاصل تلفیق آنتروپی و پتانسیل دانش- پایه است. در این مطالعه آنتروپی از داده‌های ساختار پروتئین‌های شناخته‌شده به دست می‌آید. این توابع پتانسیل در پیش‌بینی ساختار سوم و تعیین تاثیر جهش در پایداری ساختار پروتئین کاربرد دارد.

تشکر و قدردانی: از دانشگاه تربیت‌مدرس به‌خاطر در اختیار قراردادن امکانات برای انجام این تحقیق سپاسگزارم.

تأییدیه اخلاقی: موردی از سوی نویسنده گزارش نشده است.

تعارض منافع: نویسنده مقاله اعلام می‌دارد هیچ‌گونه تعارض منافی وجود ندارد.

منابع مالی: مطالعه حاضر تحت حمایت مالی دانشگاه تربیت‌مدرس بوده است.

منابع

- 1- Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223-30.

local minima with native structure characteristics. *J Mol Biol.* 2003;329(1):159-74.

28- Levitt M. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol.* 1992;226(2):507-33.

29- John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* 2003;31(14):3982-92.

30- Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins.* 2007;69 Suppl 8:118-28.

31- Deng H, Jia Y, Zhang Y. 3DRobot: Automated Generation of Diverse and Well-packed Protein Structure Decoys. *Bioinformatics.* 2016;32(3):378-87.

32- Sankar K, Jia K, Jernigan RL. Knowledge-based entropies improve the identification of native protein structures. *Proc Natl Acad Sci USA.* 2017;114(11):2928-33.

33- Wang X, Zhang D, Huang SY. New knowledge-based scoring function with inclusion of backbone conformational entropies from protein structures. *J Chem Inf Model.* 2018;58(3):724-32.

geometry: phi,psi and Cbeta deviation. *Proteins.* 2003;50(3):437-50.

22- Sippl MJ. Calculation of conformational ensembles potentials of mean force - an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol.* 1990;213(4):859-83.

23- Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol.* 1996;258(2):367-92.

24- Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol.* 1997;268(1):209-25.

25- Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Protein.* 1999;34(1):82-95.

26- Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol.* 2000;300(1):171-85.

27- Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having