

مقایسه کارایی یک تابع انرژی دانش‌پایه و یک تابع نیروی دانش‌پایه در نمره‌دهی

کمپلکس‌های پروتئین - پروتئین

رحیم جعفری^۱، مهدی میرزایی^۲، مجید عرفانی مقدم^۳، مهدی صادقی^۴

^۱ تهران، دانشگاه تربیت مدرس، دانشکده علوم زیستی، گروه نانوبیوتکنولوژی

^۲ تهران، دانشگاه علوم پزشکی شهید بهشتی، دانشکده پیراپزشکی

^۳ تهران، دانشگاه تربیت مدرس، دانشکده علوم زیستی، گروه بیوفیزیک

^۴ تهران، پژوهشگاه دانش‌های بنیادی، پژوهشکده علوم کامپیوتر، گروه بیوانفورماتیک

^۵ تهران، پژوهشگاه ملی مهندسی ژنتیک و زیست فناوری

تاریخ دریافت: ۹۱/۶/۱۹ تاریخ پذیرش: ۹۱/۱۲/۲۱

چکیده

تابع نیروی دانش‌پایه نوعی از توابع نمره‌دهی می‌باشد که از آن در زمینه تشخیص فولد پروتئینها با موفقیت قابل توجهی استفاده شده است. در این مطالعه کارایی نوعی تابع انرژی دانش‌پایه و تابع نیروی هم ارز آن در تشخیص کمپلکس‌های درست پروتئین-پروتئین از کمپلکس‌های نادرست، با یکدیگر مقایسه شده است. مقدار نیروی کل که از یک جزء کمپلکس (گیرنده/لیگاند) بر جزء دیگر وارد می‌شود به عنوان معیار پایداری کمپلکس، مورد استفاده قرار گرفت. چنین انتظار می‌رود که این نیرو در ساختار طبیعی کمترین مقدار را داشته باشد. جهت ارزیابی کارایی هر روش، دو مجموعه مورد استفاده قرار گرفت که یکی از آنها با الگوریتم داکینگ جسم نرم و دیگری با الگوریتم داکینگ جسم سخت ایجاد شده بودند. نتایج حاصل از این مقایسه نشان می‌دهد نرخ موفقیت مدل انرژی در انتخاب ساختارهای طبیعی و نزدیک به طبیعی بالاتر از روش نیرو می‌باشد. ظاهراً وابستگی مقدار نیرو به شکل ناحیه اتصال کمپلکسها باعث ایجاد خطاهایی در مدل نیرو شده که سبب نامناسب گردیدن آن در نمره‌دهی کمپلکس‌های داک شده می‌شود.

واژه های کلیدی: انرژی دانش‌پایه، نیرو، تابع نمره‌دهی، داکینگ پروتئین-پروتئین

* نویسنده مسئول، تلفن: ۴۴۷۸۷۳۷۳، پست الکترونیکی: sadeghi@nigeb.ac.ir

مقدمه

تعیین نگرديده، زیرا طبیعت ناپایدار و گذرای برهمکنش پروتئینها سبب ایجاد اختلال در فرآیند ساخت کریستال از آنها می‌شود. از این رو با توجه به اهمیت موضوع، نیاز به ابزارهای کمکی جهت مطالعه برهمکنش پروتئینها بدون استفاده از ساختارهایی که با روشهای تجربی به دست آمده‌اند، احساس می‌شود.

پروتئینها تقریباً در تمام فرآیندهای زیستی نقش محوری دارند و در بسیاری از موارد، تنها در صورتی قادر به انجام عمل اختصاصی خود هستند که با برخی پروتئینهای دیگر برهمکنش داشته باشند. جهت بررسی دقیق برهمکنش پروتئینها در مقیاس اتمی، لازم است از ساختارهایی که با روشهای تجربی به دست آمده‌اند استفاده شود ولی متأسفانه ساختار کمپلکس بسیاری از پروتئینها تا کنون

در این رابطه $u_{ij}(r)$ انرژی بین اتمهای از نوع i و j با فاصله r از دو مولکول مختلف، k_B ثابت بولتزمن، T دمای محیط بر حسب کلونین، $\rho_{ij}(r)$ و $\rho_{ij}^*(r)$ احتمال جفت اتمهای i و j در فاصله r که به ترتیب در ساختارهای واقعی و "وضعیت مرجع" محاسبه شده‌اند. وضعیت مرجع یک وضعیت فرضی است که در آن برهمکنش اتمها بر حسب تصادف و نه بر حسب تمایل واقعی آنها به یکدیگر انجام می‌گیرد.

توابع نمره‌دهی آماری در ابتدا به منظور تشخیص فولد طبیعی در پروتئینهای تک زیرواحدی ابداع شدند و در طول سالها تلاشهای زیادی جهت بهبود آنها صورت گرفته است (۱۹). یکی از تغییرات بنیادی در چنین توابعی، تبدیل آنها از فرم انرژی به فرم نیرو بوده که توسط میرزایی و همکاران صورت گرفته است (۱۷). آنها نشان دادند، با ایجاد این تغییر و همچنین به کارگیری یک الگوریتم مناسب در مقایسه نیروهای وارد بر هر اتم در ساختارهای مختلف، می‌توان قدرت تشخیص ساختار طبیعی را از ساختارهای غیرطبیعی افزایش داد.

در بررسی حاضر سعی شده توانایی مدل انرژی و مدل نیرو در تشخیص کمپلکسهای طبیعی و نزدیک به طبیعی از کمپلکسهای غیرطبیعی با یکدیگر مقایسه شود. با فرض اینکه ساختار طبیعی در وضعیت تعادل قرار دارد، انتظار می‌رفت نیروهایی که دو مولکول بر یکدیگر وارد می‌کنند در کمترین حد ممکن یعنی نزدیک به صفر بوده و ساختارهای نادرست متناسب با میزان انحراف آنها از ساختار طبیعی نیروی بیشتری بر یکدیگر وارد کنند. بنابراین روش نمره‌دهی در این بررسی، محاسبه برآیند تمام نیروهای وارد بر اتمهای گیرنده/لیگاند در ناحیه اتصال و مقایسه آنها با یکدیگر است.

مواد و روشها

داکینگ، یک روش محاسباتی می‌باشد که در آن سعی می‌شود با استفاده از ساختارهای اجزای سازنده (گیرنده و لیگاند)، شکل نهایی کمپلکس مورد نظر پیشگویی شود. الگوریتمهای داکینگ از دو بخش اصلی تشکیل شده‌اند که عبارتند از الگوریتمهای جستجو و توابع نمره‌دهی. الگوریتم جستجو باید بتواند حالت‌های مختلف برهمکنش دو مولکول را در یک بازه زمانی قابل قبول به وجود آورد و از توابع نمره‌دهی جهت ارزیابی کیفیت ساختارهای ایجاد شده به منظور تشخیص ساختار درست (طبیعی) یا "نزدیک به درست" از ساختارهای غلط استفاده می‌شود. لازم به ذکر است، توابع نمره‌دهی هم در زمان جستجو و هم در زمان پس از آن مورد استفاده قرار می‌گیرند (۱۰).

توابع نمره‌دهی را بر اساس ماهیت و نحوه ساخت، معمولاً در سه گروه مختلف طبقه‌بندی می‌کنند. در توابع نمره‌دهی فیزیکی، از روابط مربوط به مکانیک مولکولی جهت محاسبه انرژی ساختارها استفاده می‌شود (۴ و ۱۳). توابع نمره‌دهی تجربی، ترکیبی خطی از انواع ویژگیهای مرتبط به انرژی مثل پیوند هیدروژنی، الکترواستاتیک، وان‌دروالس و غیره هستند که وزن یا ضرایب آنها متناسب با نوع کاربرد تنظیم می‌شود (۵، ۸، ۱۱ و ۱۸). دسته سوم، توابع دانش‌پایه هستند که در ساخت آنها از خصوصیات ساختاری و فیزیکوشیمیایی تعداد نسبتاً زیادی پروتئین با ساختار مشخص، استفاده می‌شود.

توابع پتانسیل آماری نوعی از توابع دانش‌پایه هستند که در آنها فراوانی خصوصیت‌های ذکر شده با استفاده از عکس رابطه بولتزمن به یک سری تابع انرژی موثر تبدیل می‌شوند (۲، ۳، ۶، ۱۲ و ۲۱). در بسیاری از موارد، از خصوصیات مثل برهمکنش بین اتمها یا زنجیره‌های جانبی اسیدهای آمینه در ساخت این نوع توابع استفاده می‌شود و شکل کلی آن به صورت زیر است:

$$u_{ij}(r) = -k_B T \ln [\rho_{ij}(r) / \rho_{ij}^*(r)]$$

مجموعه هر دو به صورت جداگانه تعیین ساختار شده‌اند. این کمپلکسها بر اساس مشکل بودن داکینگ به سه گروه تقسیم می‌شوند: آسان، متوسط و سخت. اجرای داکینگ جسم سخت به وسیله *zdock* بر روی کمپلکسهایی که جزء گروه سخت هستند نمی‌تواند ساختار نزدیک به طبیعی تولید کند. از آنجایی که تعداد دکویهای ۱۲۴ کمپلکس نسبتاً زیاد است، برای افزایش سرعت محاسبه انرژی و نیرو، یک مجموعه کوچکتر ساخته شد که اعضای آن به صورت کاملاً تصادفی انتخاب شده بودند. از این زیر مجموعه، کمپلکسهایی که جزء گروه سخت بودند و همچنین آنهایی که توالی‌شان بیشتر از ۶۵ درصد با توالی پروتئینهای مجموعه آموزش یکسان بودند، حذف گردیدند. در نتیجه با اعمال فیلترهای ذکر شده در نهایت ۴۶ کمپلکس باقی ماند.

در مرحله بعد با استفاده از نرم افزار *zdock-3.0.1* (۱۵)، عمل داکینگ جسم سخت بر روی هر یک از ۴۶ جفت انجام گرفته و به ازای هر یک از آنها تعداد ۳۶۰۰ کمپلکس تولید شد. در بین ساختارهای مربوط به ۴۲ کمپلکس تست، حداقل یک ساختار نزدیک به طبیعی وجود داشت. ۴ مورد دیگر که فاقد ساختار نزدیک به طبیعی بودند از مجموعه حذف شدند. معیار به کار رفته در تعریف ساختارهای نزدیک به طبیعی در بخش نتایج توضیح داده شده است.

تعریف انواع اتم (Atom types): با وجود آنکه نوع اتمهای سازنده پروتئینها مشخص و محدود می‌باشند، به دلیل نوع اتصالات و محیط پیرامونی اتمها، تنوع آنها به لحاظ خصوصیات فیزیکوشیمیایی زیاد است. بر این اساس تعداد "انواع اتمی" در ۲۰ اسید آمینه طبیعی ۱۶۷ نوع می‌باشد. مسلماً در ساخت توابع، هر اندازه اتمهای مورد استفاده متنوع‌تر باشند، تابع نهایی بهتر خواهد بود. اما تفکیک اتمها به انواع مختلف سبب می‌شود فراوانی آنها در ساختارهای مجموعه آموزش بسیار کمتر از حدی بشود که

مجموعه آموزش: جهت آموزش از مجموعه‌ای که توسط Huang و Zou ساخته شده بود استفاده گردید (۶). این مجموعه دارای ۸۵۰ کمپلکس پروتئین-پروتئین می‌باشد که تفکیک‌پذیری ساختاری آنها حداکثر ۲/۵ آنگستروم است. زیرواحدهای هر کمپلکس حداقل ۱۰ اسید آمینه داشته و حداقل ۳۰ برهمکنش اسید آمینه-اسید آمینه در ناحیه برهمکنش آنها وجود دارد. توالی هر جفت مولکول گیرنده-گیرنده یا لیگاند-لیگاند از دو کمپلکس مختلف، حداکثر ۷۰ درصد یکسان هستند.

مجموعه‌های تست: معمولاً به منظور ارزیابی دقت توابع نمره‌دهی از مجموعه‌هایی به نام "مجموعه دکوی" (Decoy set) استفاده می‌شود. این مجموعه‌ها شامل چندین پروتئین تست هستند که به ازای هر یک از آنها تعداد زیادی ساختار غلط یا دکوی و تعداد بسیار کمی ساختار "نزدیک به درست" وجود دارد و توابع نمره‌دهی باید بتوانند این ساختارها را از ساختارهای غلط تشخیص دهند. در این مطالعه از دو مجموعه دکوی استفاده گردید که با روشهای داکینگ جسم سخت و داکینگ جسم نرم تولید شده‌اند. در داکینگ جسم سخت، پیکربندی اجزای سازنده کمپلکس در طول فرآیند داکینگ ثابت باقی می‌ماند ولی در داکینگ جسم نرم سعی می‌شود علاوه بر تغییر موقعیت دو مولکول نسبت به یکدیگر، تغییرات پیکربندی ناشی از برهمکنش آنها نیز شبیه‌سازی شود.

مجموعه داکینگ جسم نرم (soft body docking): مجموعه دکوی مورد استفاده با استفاده از نرم افزار RosettaDock تولید شده است. این مجموعه شامل ۵۴ کمپلکس می‌باشد که به ازای هر یک از آنها حدود ۱۰۰۰ ساختار وجود دارد (۵)

مجموعه داکینگ جسم سخت (rigid body docking): مجموعه‌های تست (Benchmark) شماره ۲ و ۳ ساخته شده توسط Weng و همکاران (۷ و ۱۶) شامل ۱۲۴ کمپلکس می‌باشد. مولکول‌های گیرنده و لیگاند در این

بروز چنین مشکلی شود. در این مطالعه ۱۶۷ نوع اتم در ۱۶ گروه قرار داده شد (جدول ۱) که دو به دو مجموعاً ۱۳۶ "جفت تیپ اتمی" را تشکیل می‌دهند.

بتوان از آنها در ساخت تابع پتانسیل استفاده کرد زیرا اساس این نوع توابع، مشاهده آماری می‌باشد. در چنین شرایطی لازم است با دسته‌بندی اتمهای دارای خصوصیات فیزیکوشیمیایی مشابه در انواع اتمی کمتر، تا اندازه‌ای از

جدول ۱- لیست ۱۶ تیپ اتمی متعلق به ۲۰ اسید آمینه استاندارد که در ساخت توابع انرژی استفاده شده‌اند.

شماره نوع اتم	توضیح	مثال
۱	اکسیژن‌های گروه کربونیل در زنجیره اصلی	GLY-O, PHE-O, ...
۲	اکسیژن‌های گروه کربونیل در زنجیره جانبی	GLN-OE1, ASN-OD1
۳	اکسیژن‌های گروه هیدروکسیل	SER-OG, TER-OG1, TYR-OH
۴	اکسیژن‌های گروه کربوکسیل	ASP-OD1, ASP-OD2, GLU-OE1, GLU-OE2
۵	نیترژن‌های زنجیره اصلی	GLY-N, PHE-N, ...
۶	نیترژن‌های باردار زنجیره جانبی	ARG-NH1, ARG-NH2, LYS-NZ, HIS-ND1, HIS-NE2
۷	نیترژن‌های خنثی زنجیره جانبی (با یک هیدروژن آزاد)	TRP-NE1, ARG-NE
۸	نیترژن‌های خنثی زنجیره جانبی (با دو هیدروژن آزاد)	ASN-ND2, GLN-NE2
۹	گوگردها	MET-SD, CYS-SG
۱۰	کربن‌های گروه کربونیل در زنجیره اصلی	GLY-C, PHE-C, ...
۱۱	کربن‌های گروه کربونیل در زنجیره جانبی	ASN-CG, GLN-CD
۱۲	کربن‌های گروه کربوکسیل در زنجیره جانبی	ASP-CG, GLU-CD
۱۳	کربن‌های آروماتیک	PHE-CE1, TYR-CG, TRP-CZ2, HIS-CE1, ...
۱۴	کربن‌های آلفا (زنجیره اصلی)	GLY-CA, PHE-CA, ...
۱۵	کربنی که به نیترژن با بار مثبت متصل شده	ARG-CZ
۱۶	بقیه کربن‌های زنجیره جانبی (کربن‌های آلیفاتیک)	VAL-CG1, PRO-CD, ILE-CD2, ARG-CB, HIS-CB, ...

انرژی بین جفت اتمهایی از نوع i و j در فاصله r ، که یکی از آنها متعلق به گیرنده و دیگری متعلق به لیگاند باشد، از رابطه زیر محاسبه گردید که بر گرفته از کار Sippl می‌باشد (۲۰).

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

σ وزنی است که به هر مشاهده داده می‌شود و مقدار آن ۰/۰۲ انتخاب گردید.

در این رابطه، M_{ij} تعداد کل جفت اتمهای ij است.

$$M_{ij} = \sum_{r=1}^{34} n(i, j, r)$$

محاسبه توابع انرژی: در ساخت توابع انرژی با کاربرد داکینگ، معمولاً از اطلاعات موجود در ناحیه اتصال کمپلکسها استفاده می‌شود. دو اسید آمینه از دو زیرواحد مختلف، متعلق به ناحیه اتصال خواهند بود اگر فاصله حداقل یک جفت از اتمهای سنگین آنها (اتمهایی به جز هیدروژن) ۴/۵ آنگستروم و یا کمتر باشد.

در ابتدا فراوانی جفت اتمهای ناحیه اتصال در محدوده‌های مشخصی از فاصله‌ها (مثلاً ۱/۲۵-۱ آنگستروم، ۱/۲۵-۱/۵ آنگستروم و ...) حساب می‌شود. در مطالعه حاضر از ۳۴ "محدوده فاصله"، هر یک با پهنای ۰/۲۵ آنگستروم استفاده گردید و از جفت اتمهایی که فاصله آنها بیش از ۸/۵ آنگستروم بود صرف نظر شد.

بردار نیرو بین اتمهای i و j و در فاصله d برابر خواهد بود با:

$$\vec{F} = \sum_{ij} \vec{f}_{ij}$$

$$\vec{f}_{i,j,d} = f_{i,j,d} \cdot \vec{u}$$

که در آن $f_{i,j,d}$ مقدار عددی نیرو در فاصله d می‌باشد که قبلاً از طریق مشتق عددی محاسبه شده بود. نیروی کل (\vec{F}) که از سمت اتمهای گیرنده بر اتمهای لیگاند وارد می‌شود جمع برداری تمام نیروهای حاصل از جفت اتمهای گیرنده-لیگاند است. بدیهی است، نیروی کل گیرنده بر لیگاند دقیقاً برابر ولی برخلاف جهت نیرویی است که از لیگاند بر گیرنده وارد می‌شود. بنابراین محاسبه یکی از آنها کافی است.

در نهایت برای مقایسه کمپلکسها با یکدیگر لازم است اندازه بردار نیروی کل که همان نمره هر ساختار می‌باشد حساب گردد:

$$\text{Score} = \|\vec{F}\| = \sqrt{F_i^2 + F_j^2 + F_k^2}$$

F_i ، F_j و F_k مؤلفه‌های بردار F هستند.

نتایج

پس از نمره‌دهی با روش مربوطه و سپس مرتب کردن ساختارها بر اساس نمره‌شان، انتظار می‌رود ساختارهای با کیفیت بهتر در رتبه‌های بالاتر قرار گیرند. اگر در مورد یک پروتئین تست، حداقل یک ساختار با کیفیت مورد نظر در چندتای اول (مثلاً در ۱۰ تای اول) مشاهده شود در این صورت روش مورد نظر در مورد آن موفق بوده است. درصد مواردی که در آنها موفق به یافتن حداقل یک ساختار مطلوب که حداکثر دارای رتبه مورد نظر باشد، نشان دهنده میزان توانایی روش مورد مطالعه به ازای آن رتبه خاص ($top\ n$) می‌باشد. به عنوان مثال اگر ۵۰ عدد پروتئین تست وجود داشته باشد که هر یک دارای چندین

$n(i, j, r)$ تعداد جفت اتمهای ij در فاصله r می‌باشد (فراوانی مطلق)

$f_{ij}(r)$ فراوانی نسبی جفت اتمهای ij در فاصله r است که برابر است با:

$$f_{ij}(r) = \frac{n(i, j, r)}{M_{ij}}$$

$f_{xx}(r)$ فراوانی نسبی همه جفت اتمها در فاصله r است که برابر می‌باشد با:

$$f_{xx}(r) = \frac{\sum_{i=1}^{16} \sum_{j=1}^{16} n(i, j, r)}{\sum_{i=1}^{16} \sum_{j=1}^{16} M_{ij}}$$

$T = 293\text{ }^\circ\text{K}$ و در نتیجه مقدار RT برابر است با 0.582 kcal/mol

انرژی کل یک کمپلکس، مجموع انرژی جفت اتمهای گیرنده-لیگاند در ناحیه اتصال است.

محاسبه توابع نیرو: نیرو در حقیقت تغییرات انرژی به ازای تغییرات فاصله است ($F = \frac{\Delta u}{\Delta r}$). برای محاسبه مقدار نیرو در نقطه مورد نظر از روش مشتق عددی استفاده شد.

$$F_i = \frac{p_{i+1} - p_{i-1}}{r_{i+1} - r_{i-1}}$$

در این رابطه F_i مقدار نیرو برای "محدوده فاصله" شماره i ، p_{i+1} و p_{i-1} به ترتیب مقدار انرژی در محدوده‌های بعدی و قبلی و r_{i+1} و r_{i-1} به ترتیب فاصله تا مرکز محدوده‌های بعدی و قبلی هستند.

محاسبه نیروی وارد بر اتمها: برای محاسبه نیرویی که از اتم گیرنده A_i با مختصات (x_i, y_i, z_i) به اتم لیگاند A_j با مختصات (x_j, y_j, z_j) وارد می‌شود، در ابتدا لازم است بردار یگه A_i به A_j حساب شود.

$$u = \frac{(x_j - x_i, y_j - y_i, z_j - z_i)}{d}$$

$$d = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

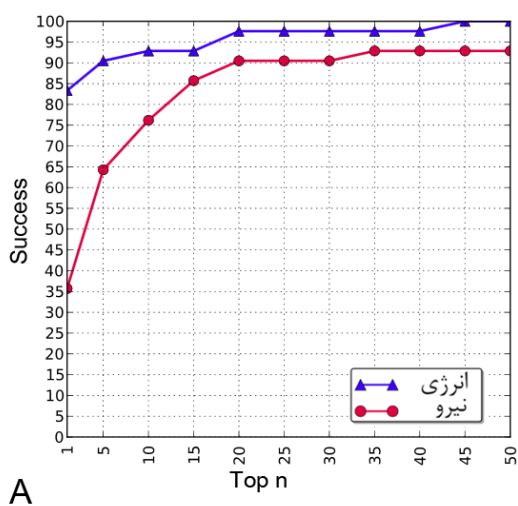
همین دلیل فایل pdb آنها به صورت دستی اضافه شد. بعد از نمره‌دهی با دو روش انرژی و نیرو، "نرخ موفقیت" برای هر کدام از مجموعه‌ها محاسبه گردید (شکل ۱). در مجموعه‌ای که با استفاده از داکینگ جسم سخت ساخته شده، هر دو روش توانسته‌اند ساختارهای طبیعی را تا حد زیادی از ساختارهای غیرطبیعی جدا کنند، به طوری که درصد موفقیت در *top 10* برای هر دو بیش از ۷۵ درصد است. اما در مجموع، دقت روش انرژی بهتر از روش نیرو است زیرا درصد موفقیت انرژی برای رتبه اول (*top 1*) بالاتر بوده و همچنین زودتر به درصد موفقیت ۱۰۰ می‌رسد.

در مجموعه دوم که با روش داکینگ جسم نرم ساخته شده، نرخ موفقیت هر دو روش در مقایسه با مجموعه قبل کاهش زیادی می‌یابد. در این مجموعه نیز روش انرژی بهتر از نیرو عمل می‌کند. همانطور که در شکل ملاحظه می‌شود روش نیرو در هیچ یک از *top n* ها قادر به یافتن ساختار طبیعی نیست.

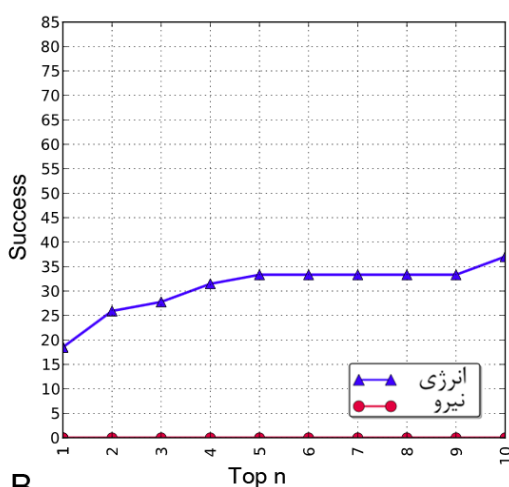
کمپلکس باشند و مثلاً اگر در ۵ مورد از آنها بتوان حداقل یک کمپلکس با ساختار مناسب در بین مثلاً ۳ ساختاری که بیشترین نمره را دارند یافت، در این صورت درصد موفقیت به ازای *top 3* معادل ۱۰ درصد خواهد بود. به این ترتیب با تغییر *top n* و محاسبه مجدد درصد موفقیت، می‌توان منحنی "نرخ موفقیت" را به دست آورد.

یافتن ساختار طبیعی: از آنجا که محاسبه ساختاری، دقیقاً شبیه به آنچه که با روشهای تجربی تعیین می‌شود، عملاً از طریق داکینگ غیر ممکن است، معمولاً هدف مورد جستجو نه ساختار طبیعی بلکه ساختارهای نزدیک به طبیعی هستند. اگر ساختار طبیعی در بین ساختارهای دیگر موجود باشد، در این صورت یافتن آن بسته به نوع مجموعه، در بسیاری از موارد نسبتاً راحت است. با این وجود از آنجایی که هدف از انجام این پژوهش مقایسه روشها می‌باشد، نتیجه چنین آزمونی جهت قضاوت نهایی مفید خواهد بود.

مجموعه‌های مورد استفاده در ابتدا فاقد ساختارهای کمپلکس طبیعی (تعیین شده با روش تجربی) بودند به



A



B

شکل ۱) نرخ موفقیت در یافتن ساختار طبیعی. (A) مجموعه تولید شده با روش داکینگ جسم سخت. (B) مجموعه تولید شده با روش داکینگ جسم نرم (منحنی نیرو با محور افقی نمودار مماس شده و به سختی دیده می‌شود).

گیرنده‌های آنها با استفاده از اتمهای زنجیره اصلی (ستون فقرات پروتئین) به طور بهینه بر روی یکدیگر قرار گیرند. برای محاسبه I_{rms} ، ابتدا اسیدهای آمینه موجود در محل اتصال کمپلکس طبیعی را مشخص کرده و سپس $rmsd$ بین این اسیدهای آمینه و معادل آنها در کمپلکس پیشگویی شده را بعد از انطباق آنها با استفاده از اتمهای زنجیره اصلی حساب می‌کنند. لازم به ذکر است که در این مورد برای تعیین ناحیه اتصال در ساختار طبیعی از cut off معادل ۱۰ آنگستروم استفاده می‌شود. در نهایت کیفیت کمپلکسها با توجه به مقادیر این سه ویژگی بیان می‌گردد (جدول ۲).

یافتن ساختار نزدیک به طبیعی: تعریف ساختار نزدیک به طبیعی بر اساس شاخصهای مختلفی انجام می‌گیرد که یکی از آنها اصطلاحاً "کیفیت ساختار" است. بر اساس معیارهای CAPRI، کیفیت ساختار تلفیقی از سه پارامتر f_{nat} ، I_{rms} و L_{rms} می‌باشد (۹ و ۱۴). f_{nat} عبارت است از تعداد تماسهای اسیدآمینه-اسیدآمینه طبیعی در ساختار پیشگویی شده، تقسیم بر تعداد تماسهای اسیدآمینه-اسیدآمینه در ساختار طبیعی (ساختار کریستاله). اسید آمینه‌های گیرنده و لیگاندی که در فاصله کمتر از ۵ آنگستروم باشند، در تماس با هم فرض می‌شوند. L_{rms} عبارت است از $rmsd$ بین لیگاند در ساختار پیشگویی شده و لیگاند در حالت طبیعی، بعد از آنکه

جدول ۲- کیفیت ساختارهای پیش بینی شده بر اساس ترکیب سه پارامتر f_{nat} ، I_{rms} و L_{rms} (۹ و ۱۴)

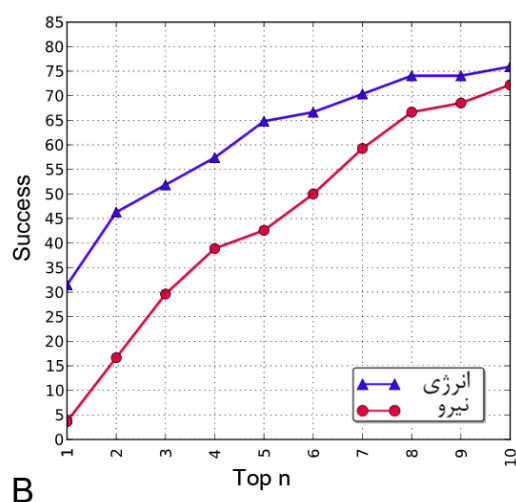
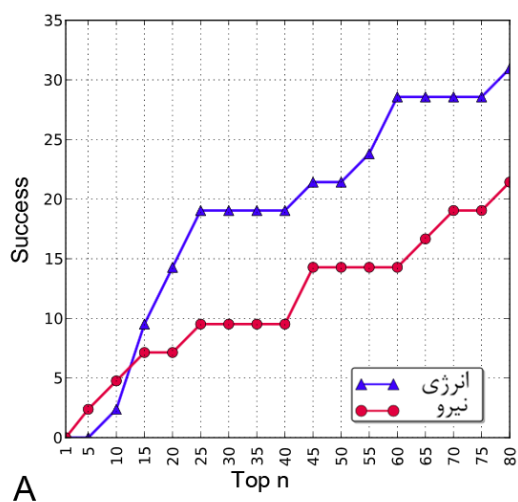
کیفیت	ترکیب سه ویژگی
بالا	$(f_{nat} > 0.5) \text{ and } (L_{rms} < 1.0 \text{ or } I_{rms} < 1.0)$
متوسط	$(0.3 < f_{nat} < 0.5) \text{ and } (L_{rms} < 5.0 \text{ or } I_{rms} < 0.2) \text{ or } (f_{nat} > 0.5 \text{ and } L_{rms} > 1.0 \text{ and } I_{rms} > 1.0)$
قابل قبول	$(0.1 < f_{nat} < 0.3) \text{ and } (L_{rms} < 10.0 \text{ or } I_{rms} < 4.0) \text{ or } (f_{nat} > 0.3 \text{ and } L_{rms} > 5.0 \text{ and } I_{rms} > 2.0)$
غلط	$(f_{nat} > 0.1) \text{ and } (L_{rms} < 10.0 \text{ or } I_{rms} < 4.0)$

می‌باشد، اما این مقدار به قدری ناچیز است که بر اساس آن نمی‌توان قضاوت قابل اطمینانی داشت. در مجموع، قدرت تمایز روش انرژی در این مجموعه نیز بهتر از روش نیرو است.

در مجموعه دوم که با روش داکینگ جسم نرم ساخته شده، توانایی روش نیرو در یافتن ساختارهای نزدیک به طبیعی بهتر از ساختار طبیعی می‌باشد که دلیل آن بیشتر بودن تعداد آنها است. با این حال در این مجموعه نیز مجدداً روش انرژی عملکرد بهتری نسبت به روش نیرو دارد.

نتایج ارائه شده در این قسمت بر مبنای توانایی روشها در یافتن ساختارهایی است که بر اساس جدول فوق، کیفیت آنها در حد "قابل قبول" و یا بهتر باشد. در این آزمون ساختارهای طبیعی از مجموعه‌های تست خارج شدند زیرا بر اساس معیار ذکر شده کیفیت آنها "بالا" است که نباید با ساختارهای نزدیک به طبیعی که چنین کیفیتی دارند مخلوط شوند.

همانطور که در شکل ۲ مشاهده می‌شود، در مجموعه ساخته شده با روش داکینگ جسم سخت، به ازای $top n$ های ابتدایی، توانایی روش نیرو کمی بیشتر از روش انرژی



شکل ۲- نرخ موفقیت در یافتن ساختار نزدیک به طبیعی. (A) مجموعه تولید شده با روش داکینگ جسم سخت. (B) مجموعه تولید شده با روش داکینگ جسم نرم

بر خلاف روش معمول که نمره یک ساختار بر اساس مجموع انرژی برهمکنش جفت اتمهای گیرنده-لیگاند محاسبه می‌شود، روش به کار رفته در این مطالعه، محاسبه نیروی کلی است که از سمت یک مولکول بر دیگری وارد می‌شود. با این فرض که کمپلکس طبیعی در یک وضعیت تعادل قرار دارد انتظار می‌رود نیرویی که از سمت گیرنده بر لیگاند و برعکس وارد می‌شود در کمترین حد ممکن باشد. جهت مقایسه این دو روش، از دو آزمون مختلف یعنی یافتن ساختار طبیعی و یافتن ساختارهای نزدیک به طبیعی در بین مجموعه‌ای از ساختارهای غلط استفاده شد. مجموعه‌های مورد استفاده با دو روش داکینگ جسم سخت و نرم ساخته شده بودند.

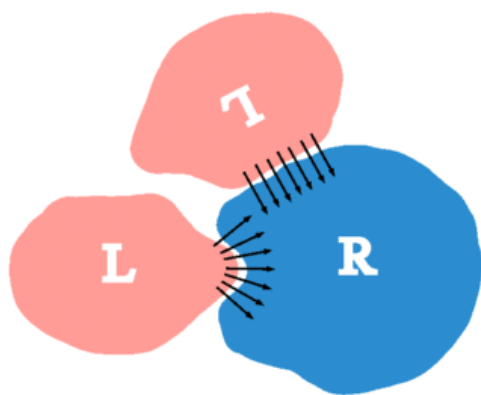
نتیجه آزمون یافتن ساختار طبیعی در مجموعه داکینگ جسم سخت نشان می‌دهد، روش نیرو توانسته ساختارهای طبیعی را برای حدود سه چهارم از کمپلکسهای تست در ۱۰ تای اول رتبه‌بندی قرار دهد. چنین نتیجه‌ای، بیانگر آن است که فرضیه به کار رفته در روش نیرو تا حدودی درست بوده است. با این وجود عملکرد روش انرژی در مقایسه با روش نیرو، در یافتن ساختار طبیعی در این

بحث

در بسیاری از کاربردها، انرژی یک شاخص پذیرفته شده در اندازه‌گیری و مقایسه میزان پایداری سیستمها می‌باشد. با این فرض که شکل طبیعی مولکولهای زیستی اغلب در حداقل انرژی یا نزدیک به آن قرار دارد، در کاربردهای بیوانفورماتیکی نیز جهت مقایسه میزان طبیعی بودن ساختارها، از انرژی چه به صورت مستقیم (از طریق توابع فیزیکی) و چه به صورت ضمنی (در غالب توابع انرژی دانش پایه)، به طور وسیعی استفاده می‌شود. اما به کارگیری نیرو به جای انرژی با چنین هدفی سابقه و عمومیت چندانی نداشته و تنها موردی که از آن به طور موفقیت‌آمیز توسط میرزایی و همکاران استفاده شده، تمایز ساختارهای غیرطبیعی از ساختار طبیعی در پروتئینها بوده است (۱۷). با در نظر گرفتن موفقیت نسبتاً خوب روش نیرو در کاربرد ذکر شده و همچنین نبود هیچ گزارشی در مورد استفاده از آن در مسئله داکینگ، در این مطالعه سعی شد کارایی روش انرژی و نیرو در مسئله نمره‌دهی کمپلکسهای پروتئین-پروتئین با یکدیگر مقایسه شود.

از این رو به جای محاسبه نمره کمپلکسها به روش مقایسه‌ای، در مطالعه حاضر نمره هر ساختار به صورت مستقل و بدون در نظر گرفتن وضعیت نیروها در سایر ساختارها محاسبه شد. همانگونه که گفته شد این نمره، برآیند همه نیروهایی است که از سمت گیرنده بر لیگاند و یا برعکس وارد می‌شود.

به دلیل آنکه نیرو یک ویژگی برداری است، علاوه بر نوع و فاصله اتمها، نحوه پراکندگی آنها نسبت به هم نیز در مقدار نیروی برآیند تأثیر می‌گذارد. نحوه پراکندگی و موقعیت فضایی اتمها نسبت به یکدیگر، متأثر از شکل ناحیه اتصال دو پروتئین است که در کمپلکسهای مختلف متفاوت می‌باشد. در یک مجموعه دکوی، شکل ناحیه اتصال در برخی از ساختارها ممکن است مسطح و در برخی دیگر غیر مسطح (مقعر/محدب یا تاب‌دار) باشد (شکل ۳). نواحی اتصال که شکل غیر مسطح دارند بردارهای نیرو در آنها پراکنده تر از نواحی اتصال با شکل مسطح می‌باشد. واضح است جمع بردارهای پراکنده منجر به ایجاد برداری با طول کمتر نسبت به جمع بردارهای موازی خواهد شد.



شکل ۳- وابستگی پراکندگی بردارهای نیرو به شکل ناحیه اتصال

این وابستگی نیروی برآیند به شکل ساختار موجب می‌شود، در صورت شبیه بودن سایر عوامل، کمپلکسهایی که ناحیه اتصال آنها اعوجاج و انحنای بیشتری دارد، در مقایسه با آنهایی که شکل ناحیه اتصالشان مسطح‌تر است،

مجموعه و نیز مجموعه دوم که با روش داکینگ جسم نرم ایجاد شده بود، بهتر است.

بر اساس این نتیجه و همچنین نتایج به دست آمده در آزمون یافتن ساختارهای "نزدیک به طبیعی" در هر دو مجموعه، می‌توان چنین جمع بندی نمود که روش انرژی در نمره‌دهی و جداسازی کمپلکسهای پروتئین-پروتئین با ساختار درست از کمپلکسهای غیرطبیعی، توانایی بالاتری در مقایسه با روش نیرو دارد.

این نتیجه‌گیری بر خلاف یافته قبلی می‌باشد که در آن روش نیرو به شکل موفقیت‌آمیزی در زمینه تشخیص فولد استفاده شده است. در فولد شدن پروتئینها و همچنین اتصال آنها به یکدیگر، برهمکنش‌های فیزیکی مشابهی بین اسیدهای آمینه و مولکولهای آب وجود دارد. با توجه به این شباهت بعید به نظر می‌رسد که دلیل این اختلاف در یافته‌ها، ناشی از وجود یک تفاوت بنیادین در ماهیت فیزیکی این دو پدیده زیستی باشد.

ظاهراً دلیل این تفاوت، نحوه به کارگیری نیرو در این دو مسئله بوده است. در مسئله تشخیص فولد، نمره هر ساختار نه به صورت مستقل بلکه از طریق مقایسه با سایر ساختارها تعیین می‌شود. به این معنی که مقدار نیروی وارد بر هر اتم در تمام ساختارها با یکدیگر مقایسه شده و ساختاری که تعداد اتمهای دارای حداقل نیرو در آن بیش از بقیه باشد امتیاز بیشتری کسب کرده و به عنوان ساختار طبیعی انتخاب می‌گردد (جهت مطالعه جزئیات روش، به تحقیق حاضر مراجعه شود). متأسفانه استفاده از چنین روشی در مسئله داکینگ، عملاً غیر ممکن است زیرا کمپلکسهای یک مجموعه از نواحی متفاوت به هم متصل بوده و به همین دلیل، یک اتم در یک ناحیه اتصال مشخص، ممکن است در ناحیه اتصال بسیاری از کمپلکسهای دیگر وجود نداشته و در نتیجه عمل مقایسه نیروی این اتم در تمام ساختارها امکان‌پذیر نباشد.

به دلیل آنکه برآیند نیروها مستقل از شکل ساختار نیست، استفاده از روش نیرو به صورت ذکر شده با پیچیدگی بسیار زیادی روبرو می‌باشد. اما در روش انرژی چنین پیچیدگی وجود نداشته و صرفاً نوع جفت اتمها و فاصله آنها از یکدیگر در مقدار انرژی کل مؤثر می‌باشد.

رتبه بهتری را در نمره‌دهی با روش نیرو کسب کنند. این می‌تواند یک عامل ایجاد خطا باشد زیرا بررسیهای قبلی نشان داده، به جز برخی استثناها، ناحیه اتصال در بیشتر کمپلکسهای پروتئین-پروتئین کم و بیش مسطح است (۱).

منابع

1. Bahadur RP, Zacharias M. (2008), The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci*, 65:1059-1072
2. Chuang GY, Kozakov D, Brenke R, Comeau SR, Vajda S. (2008), DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J*, 95:4217-4227
3. Feliu E, Aloy P, Oliva B. (2011), On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Sci*, 20:529-541
4. Fiorucci S, Zacharias M. (2010), Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins*, 78: 3131-3139
5. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. (2003), Protein-Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J Mol Biol*, 331:281-299
6. Huang SY, Zou X. (2008), An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, 72:557-579
7. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. (2008), Protein-protein docking benchmark version 3.0. *Proteins*, 73:705-709
8. Jackson RM, Gabb HA, Sternberg MJ. (1998), Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem. *J Mol Biol*, 276:265-85
9. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. (2003), CAPRI: a critical assessment of predicted interactions. *Proteins*, 52:2-9
10. Lee K, Lee JW. (2008), Computational Approaches to Protein-Protein Docking. *Current Proteomics*, 5:10-19
11. Liang S, Liu S, Zhang C, Zhou Y. (2007), A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. *Proteins*, 69:244-53
12. Liu S, Vakser IA. (2011), DECK Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics*, 12:280
13. Mandell JG, Roberts VA, Pique ME, Kotlovyy V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. (2001), Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14:105-113
14. Méndez R, Leplae R, Lensink MF, Wodak SJ. (2005), Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60:150-169
15. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. (2005), Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, 60:214-216
16. Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. and Weng, Z. (2007), Integrating statistical pair potentials into protein complex prediction. *Proteins*, 69: 511-520
17. Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M. (2009), A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys. *Proteins*, 77:454-463
18. Pierce B, Weng Z. (2007), ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins*, 67: 1078-1086
19. Shen MY, Sali A. (2006), Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15:2507-2524.
20. Sippl MJ. (1990), Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213:859-883

21. Zhang C, Liu S, Zhou H, Zhou Y. (2004), An accurate residue-level pair potential of mean force for folding and binding based on the distance- scaled ideal-gas reference state. *Protein Sci*, 13:400-411

Comparing the performance of a knowledge-based potential and a knowledge-based force functions in the scoring of protein-protein complexes

Jafari R.¹, Mirzaie M.^{2,4}, Erfani Moghaddam M.³ and Sadeghi M.⁵

¹ Nanobiotechnology Dept., Faculty of Biological Sciences, Tarbiat Modares University, Tehran, I.R. of Iran

² Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, I.R. of Iran

³ Biophysics Dept., Faculty of Biological Sciences, Tarbiat Modares University, Tehran, I.R. of Iran

⁴ Bioinformatics Dept., School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, I.R. of Iran

⁵ Bioinformatics Dept., National Institute of Genetic Engineering and Biotechnology, Tehran, I.R. of Iran

Abstract

The knowledge-based force function is a new type of the scoring functions that has been used in the field of protein fold recognition with a noticeable success. In this study we compared the performance of a knowledge-based potential function and its corresponding force function in discrimination of the correct protein-protein complexes from the incorrect ones. The total force imposed by one component (receptor/ligand) upon another was used as a measure of complex stability. This force is expected to be the lowest in the native structure. To test the performance of each method, two decoy sets were used; one generated by soft body docking and the other by rigid body docking algorithms. The results of this comparison show that, for both decoy sets, the success rates in native and near native selections of the energy model are higher than that of the force model. It seems, the dependence of amount of the force on shape of the interface region introduces errors in the later model and therefore makes it inappropriate for the scoring of docked complexes.

Key words: knowledge-based potential, force, scoring function, protein-protein docking