WWW.IBP.IR iranian bioinformatics portal

پنجمین همایش ملی بیوتکنولوژی جمهوری اسلامی ایران

3-5 آذر ماه 1386، سالن اجلاس سران

The 5th National Biotechnology Congress of Iran

24-26 Nov, 2007, Summit Meeting Conference Hall, Tehran- Iran

تخصیص ساختمان دوم پروتئین با استفاده از انتروپی

حبیبی مهناز [1و2و*]، پزشک حمید [2و3و]، اصلاح چی چنگیز [1]، صادقی مهدی [4]**

1- دانشکده علوم ریاضی، دانشگاه شهید بهشتی 2- هسته بیوانفورماتیک، پژوهشکده علوم کامپیوتر، پژوهشگاه دانش های بنیادی 3- گروه آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران 4- پژوهشگاه ملی مهندسی ژنتیک و زیست فناوری، تهران

چکیده فارسی: تخصیص ساختمان دوم پروتئین از مختصات سه بعدی پروتئین (3D) گام اصلی برای تعین ساختار پروتئین است. هرچند شناسایی عناصر ساختمان دوم از قبیل helices و β-strand به ظاهر آسان است، اما تعاریف متفاوتی از آنها موجود است که متضمن معیارهای متمایز است. ما الگوریتم جدیدی را برای تخصیص ساختمان دوم بر پایهء پارامترهای هندسی و انتروپی ارائه می دهیم. ما نشان می دهیم که قطعاتی با انتروپی پایین تر نظم ساختاری بیشتری دارند.

کلمات کلیدی : آنتروپی، ساختار دوم پروتئین، بردار فاصله

مقدمه: ساختارهای دوم، ساختارهایی در پروتئین هستند که به دلیل پیوند های هیدروژنی بین اسید آمینه ها بر اساس یک الگوی خاصی به وجود می آیند. این ساختارها را در سه دسته کلی طبقه بندی کرده اند. 1- helices (که خود شامل α -helix , 3_{10}- helix و π-helix می باشند.) 2- β-strand (که خود شامل β-sheet و β-bridge است.) 3- ساختارهای تکرار ناپذیر (coil, loop, turn و ساختارهای نامنظم دیگر). تکنیک های تعیین ساختار پروتئین ها از جملهٔ کریستوگرافی و NMR قطع مختصات اتم های یک پروتئین را مشخص می کنند. روش هایی بوجود آمده اند که بتوان از روی مختصات اتم ها، این ساختارها را در پروتئین پیداکنند، که به این روش ها ***تخصیص ساختمان دوم پروتئین*** می گویند. از جمله روش های رایج در این زمینه DSSP[1] است که الگوی تکرار پیوند هیدروژنی را با حساب کردن انرژی پیوند هیدروژنی پیدا می کند و یر اساس آن ساختارهای دوم را نسبت می دهند. روشهای STRIDE[2] ، DEFINE[3] و P-CURVE بر اساس پارامترهای مختلف ساختارهای دوم را تعیین می کنند. روشی که ما برای این منظور ارا ئه کرده ایم (PSE: Protein Segmentation with Entropy) ، یک روش هندسی است و به عبارتی ویژگی های هندسی اسید آمینه های متوالی را بررسی می کنیم و بر اساس انتروپی، میزان انحراف قطعات را در ساختارهای منظم تعیین می کنیم. در این مقاله، PSE با روش های دیگر (DSSP ,STRIDE) و همچنین با ساختار نسبت داده شده در فایل PDB در تعداد زیادی از پروتئین آن را مقایسه کرده ایم و نشان داده ایم که انطباق خوبی دارد.

مواد و روش ها:

پارامترها: فرض کنیم A یک پروتئین دارای s_{i+j+1} اسید آمینه باشد و s_i مختصات C_α اسیدآمینه i- ام باشد. ما به هر اسیدآمینه i، $1 \le i \le n-5$ یک بردار فاصله $\vec{d_i} = (\vec{d_1^i}, \vec{d_2^i}, \vec{d_3^i}, \vec{d_4^i})$ و به هر اسیدآمینه i ، $6 \le i \le n$ یک بردار فاصله

$\overleftarrow{d_i} = (\overleftarrow{d_1^i}, \overleftarrow{d_2^i}, \overleftarrow{d_3^i}, \overleftarrow{d_4^i})$ نسبت می دهیم، که هر $\overrightarrow{d_j^i}$ فاصله بین s_i و s_{i+j+1} و به همین ترتیب $\overleftarrow{d_j^i}$ فاصله بین s_i و s_{i-j-1} است. به همین ترتیب به هر اسیدآمینه i، $1 \le i \le n-3$ سه تایی از زاویه های پیچیدگی نسبت می دهیم، که $\overrightarrow{\phi_1^i}$ زاویه بین $\overrightarrow{\phi_2^i}$ و $\overrightarrow{(i+1)(i+2)}$ و $\overrightarrow{\phi_2^i}$ زاویه بین $\overrightarrow{(i+2)(i+1)}$ و $\overrightarrow{(i+2)(i+3)}$ و $|A_i|\overrightarrow{\phi_3^i}$ زاویهء بین صفحهء گذرا از نقاط s_i و s_{i+1} و صفحهء گذرا از نقاط s_{i+1} و s_{i+2} است. محاسبات آماری نشان می دهد که بردارهای فاصله و زاویه های پیچیدگی اسید آمینه هایی که در ساختارهای منظم β-strand helices شرکت می کنند، در بازه های بدست آمده درجدول [1] صدق می کنند.

الگوریتم PSE: فرض کنید W یک پیغام به طول n و از مجموعه $A=\{A_1, A_2, \ldots, A_N\}$ باشد، به طوری که فراوانی هر A_i در W، $|A_i|$ است، آنگاه انتروپی دنباله W به صورت زیر تعریف می شود:

$$H(W) = \sum_{i=1}^{N} -\frac{|A_i|}{n} \log\left(\frac{|A_i|}{n}\right)$$

حال با توجه جدول [1] به هر اسیدآمینه که در بازهء helices قرار گرفته باشد کد 0 و به هر اسیدآمینه که در بازهء β-strand قرار گرفته باشد کد 1 و در غیر این صورت کد 2 را منسوب می کنیم. بنابراین، دنباله ای از کدهای {0و1,2} خواهیم داشت.

قضیه: فرض کنیم $U=u_1,u_2,\ldots,u_k$ یک دنباله از کدهای {0و2} باشد، به طوری که هرپنجرهء چهارتایی از آن دارای انتروپی کمتر از 0.26 باشد، آنگاه $H(U) \le 0.26$.

در این الگوریتم هدف پیدا کردن قطعاتی با انتروپی کمتر از 0.26 است، که با توجه به قضیه فوق کافی است پنجره های چهارتایی از آن را که دارای انتروپی کمتر از 0.26 باشد، پیدا کنیم. در این صورت دنباله ای از کد 0 به فضای helices و دنباله ای از کد 1 به فضای β-strand منسوب می گردد.

نتایج و بحث: برای مقایسه PSE با روش های دیگر DSSP, STRIDE و PDB-FILE ما از معیارهای شناخته شده از قبیل ضریب همبستگی (CC) و دو پارامتر دیگر Sn (Sensitivity) و Sp (Specificity) استفاده می کنیم،

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TN+FP)(FP+FN)(TP+FP)(TN+FN)}}.$$

$$Sp = \frac{TP}{TP+FP}، \qquad Sn = \frac{TP}{TP+FN}.$$

که در آن TP(True Positive), TN(True Negative), FN (False Negative) و FP (False Positive)

براساس مقایسه 4785 پروتئین موجود در PDB، نتایج بصورت جدول [2] و [3] بدست آمده است.

از آنجایی که روش های مختلف، بر اساس معیارهای متفاوتی این ساختارهای دوم را پیدامی کند، بنابراین نتایج متفاوتی را ارا ئه می دهند. با این حال، این روش ها در پیدا کردن ناحیهء که به یک ساختار خاص تعلق داشته باشد با یکدیگر توافق خوبی دارند و تنها دلیل اختلاف آنها در انتها و ابتدای بازه ها است. از طرفی، چون روش ما هندسی است و پارامترهای $\overleftarrow{d_i}$ و $\overleftarrow{TA_i}$ را برای نسبت دادن ساختارهای دوم بکار می بریم با دقت بیشتری انتها و ابتدای بازه ها را مشخصص می کنیم.

یک ویژگی قوی روش ما، مقدار بالای Sn در مقایسه با PDB، STRIDE و DSSP است که این برتری روش ما را نسبت به روش های دیگر را نشان می دهد. به عبارتی بیشتر از 93٪ از نواحی که این روش ها helix معرفی کرده در روش ما قابل شناسایی است، ولی نواحی در پروتئین وجود دارد که از نظر هندسی شباهت کاملی به helix دارد ولی الگوهایی که، از تکرار پیوند هیدروژنی برای پیدا کردن helix استفاده می کنند آن نواحی را turn معرفی کرده اند. از طرف دیگر توافق کمتر روش ما در پیدا کردن β-strand ها با روش دیگر به این دلیل است که این روش ها، چون بر اساس ساختار پیوند هیدروژنی قرار دارند، قادر به پیدا کردن β-bridge ها که در ساختار پیوند هیدروژنی شرکت نمی کنند نیست و مزیت روش ما این است که این β-bridge را هم تشخیص می دهد. بعلاوه با استفاده از انتروپی می توانیم میزان انحراف را از حالت نرمال به دست آوریم. به عبارتی هراندازه که انتروپی به صفر نزدیک باشد میزان انحراف آن از حالت نرمال کمتر است. (شکل 1 را ملاحظه کنید)

جدول [1]: بازه های فاصله و زاویه های پیچیدگی در helices و β-strand

		α -helix	$\pi-$helix		$\beta-$strand
I^H	$\overrightarrow{d_1}$	[5.1, 5.85]	[5.1, 5.6]	I^S	[6.1, 7.1]
	$\overrightarrow{d_2}$	[4.8,6.3]	[4.7, 6.7]		[9, 10.6]
	$\overrightarrow{d_3}$				
	$\overrightarrow{d_4}$	[5, 9.3]	[4, 8.4]		[9.5, 14]

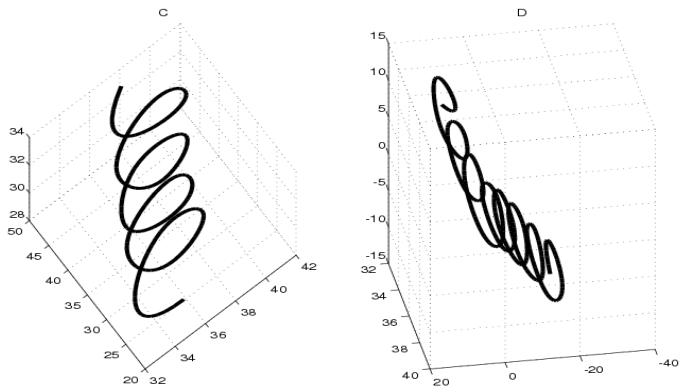
		[5.7, 12]	[5, 13.1]		[12, 17]
J^H	$\vec{\phi}_1$	[84°, 95°]	[80°, 90°]	J^S	[120°, 142°]
	$\vec{\phi}_2$	[84°, 95°]	[80°, 90°]		[120°, 142°]
	$\vec{\phi}_3$	[36°, 63°]	[50°, 78°]		[-180°,-132°]

جدول [2]: مقایسه نتایج تخصیص heices به دست آمده در PSE و روش های دیگر

روش ها	TP	TN	FP	FN	کل	%	حساسیت	تشخیص	CC
PSE_PDB	452685	672894	67040	31798	1224417	91.93	93.4	87.1	0.83
PSE_DSSP	362397	697247	157272	7501	1224417	86.5	97.97	70	0.74
PSE _ STRIDE	374139	696737	145531	8010	1224417	87.5	97.9	72	0.75
PDB_DSSP	405696	731721	71664	15336	1224417	92.8	96.3	84.9	0.85
PDB_STRIDE	414622	726579	62738	20479	1224417	93.2	95.2	85.8	0.85

جدول [3]: مقایسه نتایج تخصیص β -strand به دست آمده در PSE و روش های دیگر

روش ها	TP	TN	FP	FN	کل	%	حساسیت	تشخیص	CC
PSE _PDB	240225	749467	42824	191901	1224417	80.83	84.87	55.6	0.57
PSE _DSSP	239374	753633	185705	45705	1224417	81.1	84	56.3	0.57
PSE_STRIDE	244675	751198	180404	48140	1224417	81.3	83.5	57.5	0.57



شکل .1 شکل فوق نمایش دو helix با انتروپی های متفاوت است. .H(c)=0 و H(d)=0.21 .

تشکر و قدردانی: بخشی از هزینه این کار از محل طرح شماره CS 1385-1-02 پژوهشکده دانش های بنیادی (IPM) تامیین شده است.

منابع:

1 Kabsch,W., Sander,C. **(2001)** Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical Features, Biopolymers, 22(12), 2577-637.

2 Frishman,D., Argos,P. **(1995)** Knowledge-based protein secondary structure assignment, {\it proteins}, 23(4), 566-579.

3 Fodje,MN.,Al-Karadaghi,S.**(2002)** Occurrence, conformational features and amino aci propensities for the pi-helix, proteins Eng, 15(5), 353-358.

ABSTRACT

The automatic assignment of the protein secondary structure from three dimensional coordinates is an essential step in the characterization of protein structure. Although the recognition of secondary structure elements as alpha helices and beta sheets seem straightforward, but there are many different definitions, each regarding different criteria. We introduce a new algorithm for the protein secondary structure assignment based on a number of geometric parameters and by using the entropy, the sequence of protein is

partitioned to segments. Then the secondary structure elements are assigned to each of these segments. It is shown that if the entropy of a segment increases then the regularity in the structure decreases. So it concluded that the concept of entropy could be used as a measure of regularity of the secondary structure.

Keywords*:* secondary structure assignment; entropy; carbon alpha distance; torsion angle

WWW.IBP.IR iranian bioinformatics portal