WWW.IBP.IR iranian bioinformatics portal

جمين همايش ملي بيوتكنولوژي جمهوري اسلامي ايران
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

روشي جديد براي نسبت دادن ساختار دوم پروتئين براساس روابط رياضياتي بين مختصات سه بعدي C_α ها

حسيني سيدرزگار[1و2و*]، صادقي مهدي[3**]، حبيبي مهناز[2و4]، اصلاح چي چنگيز[4]

1-گروه بيوتكنولوژي، پرديس علوم، دانشگاه تهران 2- هسته بيوانفورماتيك، پژوهشكده علوم كامپيوتر، پژوهشگاه دانش هاي بنيادي 3- پژوهشگاه ملي مهندسي ژنتيك و زيست فناوري، تهران 4- دانشكده علوم رياضي، دانشگاه شهيد بهشتي

چكيده فارسي: نسبت دادن ساختار دوم پروتئين، گام اصلي براي آناليز و مدلسازي ساختمان پروتئين است. به همين دليل روشهايي با معيارهاي مختلف به اين منظور بوجود آمده اند. ما روش جديدي را براي نسبت دادن ساختار دوم به پروتئين با پيدا كردن چهار رابطهء رياضياتي بين مختصات سه بعدي c_αهاي اسيد آمينه هاي متوالي ، معرفي كرده ايم كه انطباق با اين روابط را ملاك نسبت دادن ساختار دوم به اسيد آمينه ها قرار مي دهيم.

كلمات كليدي: ساختار دوم پروتئين، مختصات سه بعدي c_α، رابطه هاي رياضياتي

مقدمه : ساختارهای دوم، ساختارهایی در پروتئین هستند که به دلیل برقراری پیوند هیدروژنی بین اسید آمینه ها بر اساس یک الگوی خاص، به وجود می آیند. این ساختارها را در سه دسته کلی طبقه بندی کرده اند: 1.helix ها (که شامل مارپیچ های α، 3_{10} و π می باشد.) 2.β-strandها 3 .coilها که معرف ساختارهای نامنظم موجود در پروتئین است.ساختارهاي دوم به اين دليل كه توصيف ساده و شهودي از ساختمان سه بعدي پروتئين فراهم مي كنند، در زيست شناسي ساختماني كاربردهاي فراواني پيدا كرده اند. از جمله در طبقه بندي پروتئين ها، رديف بندي بهتر توالي ها، رديف بندي ساختمان پروتئينها و مدلسازي ساختمان سه بعدي به كارمي روند.بنابراين وجود روشهايي كه بتواند ساختمان دوم را از مختصات سه بعدي اتمها تعيين كند،ضروري است. اوايل كريستالوگرافها با چشم اين كار را انجام مي دادند اما با افزايش تعداد پروتئينهاي تعيين ساختار شده،لزوم وجود روش هايي كه بصورت اتوماتيك اين كار را انجام دهد در اواسط دههء 70 احساس شد و روشهاي مختلفي براي اين كار بوجود آمدند. از جملهء اين روشها DSSP[1] است كه الگوي تكرار پيوندهاي هيدروژني را با حساب كردن انرژي پيوند زنجيرهء اصلي پيدا مي كند و بر اساس آن ساختار دوم نسبت مي دهد.[2] STRIDE كه مشابه DSSPاست با اين تفاوت كه نحوهء محاسبه پيوند هيدروژني اش متفاوت بوده و معيار زواياي پيوند را نيز به ميان مي آورد. علاوه بر اين دو، روش هاي هندسي فراواني بوجود آمده اند، مثل DEFINE وP-CURVE ، SECSTER,VOTAP.و.... ما روشي را بدين منظور تحت عنوان(MATHREL(MATHmathical RELation based assignment ارائه كرده ايم كه اساس اين روش بر 4 رابطهء رياضياتي كه بين مختصات c_α اسيد آمينه هاي مجاور پيدا كرده ايم، استوار است كه هر كدام از اين رابطه ها به يكي از ساختارهاي منظم (β-strand,π-helix,3_{10}-helix,α-helix) تعلق دارد و در نهايت با استفاده از اين رابطه ها ساختارهاي دوم را به اسيد آمينه هاي موجود در پروتئين نسبت مي دهيم . بمنظور بررسي اعتبار روش خود، آن را با روش هاي رايجSTRIDE,DSSP و فايل هاي[3] PDB مقايسه كرده ايم ونشان داديم كه با آن ها انطباق خوبي دارد.

روش: در اين قسمت مراحل بدست آوردن 4 رابطهء رياضياتي مذكور را معرفي مي كنيم.

بااستفاده از اين خاصيت كه يك مارپيچ در امتداد هر يك از محورهاي مختصات داراي 2 ويژگي نوسان و امتداد خطي است. معادلهء پارامتري مارپيچ را در حالت كلي معرفي كرده ايم:

$$g(t)=f(x(t), y(t), z(t)),$$

$$x(t)=a_x \sin(\omega t+ \theta_0)+ b_x t+ c_x \qquad a_x= R \sin \varphi_x , b_x= R \cos \varphi_x,$$

$y(t)=a_y \sin(\omega t+ \pi/2+ \theta_0)+ b_y t+ c_y$ $\quad a_y= R \sin \varphi_y$, $b_y= R \cos \varphi_y$,
$z(t)=a_z \sin(\omega t+ \pi+ \theta_0)+ b_z t+ c_z$ $\quad a_z= R \sin \varphi_z$, $b_z= R \cos \varphi_z$

كه R شعاع مقطع مارپيچ، φ زاويهء بين محور مارپيچ و محور مختصات متناظر با آن، ω چرخش زاويه اي نوسان، B ارتفاع هر دور مارپيچ، θ_0 زاويهء اوليهء نوسان بوده و c_x ، c_y و c_z ثابت هاي اوليه هستند.

به منظور بدست آوردن روابط كلي كه براي همهء مارپيچ ها صادق باشد، ما پارامترهايي را كه به يک مارپيچ خاص تعلق دارند، مثل a_x، b_x و c_x را طي مراحل زير حذف مي كنيم.

فرض كنيم c(i) و $o \le i \le 4$، پنج نقطهء متوالي در يک مارپيچ باشند آن گاه بر اساس رابطهء (1) داريم:

$$c(i) : x_i= a_x \sin(\omega i+ \theta_0)+ b_x i+ c_x \qquad (2)$$

به منظور حذف b_x و c_x ، A(j) و j=0 را تعريف مي كنيم، به طوري كه

$$A(j)= x_{(j+1)}-2\, x_{(j+1)} + x_{(j)} \qquad (3)$$

حال براي حذف a_x، A(j) را بر A(j+1) تقسيم مي كنيم، بنابراين داريم :

$$\frac{A_x(j)}{A_x(j+1)} = \frac{\lambda(j)\cos\theta_0 + \gamma(j)\sin\theta_0}{\lambda(j+1)\cos\theta_0 + \gamma(j+1)\sin\theta_0} \qquad (4)$$

كه در آن

$$\lambda(j) = \sin((j+2)\omega) - 2\sin((j+1)\omega) + \sin(j\omega)$$
$$\gamma(j) = \cos((j+2)\omega) - 2\cos((j+1)\omega) + \cos(j\omega) \qquad (5)$$

در مرحلهء آخر بمنظور حذف θ_0، رابطهء نهايي را تعريف مي كنيم:

$$\frac{\dfrac{A_x(j)}{A_x(j+1)} - \dfrac{\lambda(j)}{\lambda(j+1)}}{\dfrac{A_x(j+2)}{A_x(j+1)} - \dfrac{\lambda(j+2)}{\lambda(j+1)}} = \frac{\lambda(j+1)\gamma(j) - \lambda(j)\gamma(j+1)}{\lambda(j+1)\gamma(j+2) - \lambda(j+2)\gamma(j+1)} \qquad (6)$$

كه اين رابطه مستقل از پارامترهاي مذكور است و فقط به λ و γ بستگي دارد كه آن ها طبق (5) فقط به ω بستگي دارند و $\omega = \frac{2\pi}{T}$ كه T دورهء تناوب است. اين دورهء تناوب براى مارپيچ هاى موجود در پروتئين بصورت تعداد C_α ها در يک دور تعريف مي شود كه آن را در جدول [1] نمايش داده ايم.كه با استفاده ازω مي توانيم λ و γ را براي هر يک از ساختارها حساب كرده و با استفاده از معادله هاي (3) و (6) معادلهء مخصوص هر يک از اين مارپيچ ها را بدست مي آوريم. براي 5 نقطهء متوالي در يک توالى پروتئينى با N اسيد آمينه $1 \le k \le N-4$، 3 رابطهء زير را براى هر يک از مارپيچ ها داريم. كه براي مارپيچ هاي α, π و 3_{10} m به ترتيب 2.88، 3.51- و 1 وn به ترتيب 2.53- ، 3.23و 0 است.

$$f(x) = \frac{\left(\dfrac{x_{k+2} - 2x_{k+1} + x_k}{x_{k+3} - 2x_{k+2} + x_{k+1}}\right) + m}{\left(\dfrac{x_{k+4} - 2x_{k+3} + x_{k+2}}{x_{k+3} - 2x_{k+2} + x_{k+1}}\right) + n} = -1 \qquad (7)$$

در واقع مي شود β-strand را به عنوان مارپيچي در نظر گرفت كه در هر دور، 2 تا C_α دارد و بنابراين T=2 و ω=π، پس با ساده تر كردن رابطه، براي هر 4 نقطهء متوالي دريك توالي با N اسيد آمينه $1 \le k \le N-3$ داريم:

$$f_\beta(x_k) = \frac{x_{(k+3)} - x_{(k+1)}}{x_{(k+2)} - x_{(k)}} = 1 \,. \qquad (8)$$

الگوريتم: در ابتدا يك پنجره متحرك 5 تايی كه در هر حركت يك آمينو اسيد جلو می رود را در نظر مي گيريم، سپس مقادير (f(x), f(y), f(z)) را براساس معادله (7) برای هر 3 نوع مارپيچ و هر كدام از پنجره ها محاسبه می كنيم.و بعد انحراف هر پنجره را از هر كدام از مارپيچ ها بدست می آوريم . و اين مراحل را برای β-strand با تعريف كردن پنجره متحرك 4 تايی و استفاده از رابطه(8) تكرار می كنيم و در نهايت مقدار انحراف هر اسيد آمينه را از هر كدام از ساختارها با استفاده از انحراف پنجره هايی كه در آن قرار می گيرد حساب می كنيم سپس بر اساس آن،ساختار دوم را به تك تك اسيد آمينه ها نسبت می دهيم.

جدول [1] : خاصيت هاي تناوبي مارپيچ ها

	α -helix	π- helix	3_{10}-helix
T	3.6	4.4	3
ω	2π/3.6	2π/4.4	2π/3

نتايج و بحث: روش ما، MATHRELچون فقط از مختصات c_α ، براي نسبت دادن ساختاردوم استفاده مي كند،اين باعث افزايش سرعت، سادگي استفاده و بازدهي بالاي آن شده است. علاوه بر اينها براي اعتباربخشي روشمان آن را با PDB,STRIDE,DSSP مقايسه كرده ايم . براي اين مقايسه يك مجموعه شامل 1918 پروتئين با شباهت كمتر از 90% و قدرت تفكيك بهتر از2Aرا مورد استفاده قرار داديم و نتايج موجود در جدول [2] بدست آمد.

بر اساس جدول[2]، درصد توافق مارپيچها بين روش ما و هر 3 روش بالاتر از 90% است كه اين بيانگر آن است كه در پيدا كردن مارپيچها،MATHREL با اين روشهاي معتبر،به خوبي موافقت دارد و با استناد به آنها مارپيچ ها را خوب تشخيص مي دهد.از طرف ديگردرصد توافق در پيدا كردن β-strand كمي پايين تر است. دليل اصلي اين مساله اين است كه روشهاي براساس پيوند هيدروژني،فقط قادر به تشخيص β-strandهايي هستند كه در پيوند هيدروژني شركت كرده اند، است و نمي توانند β-bridge ها را پيدا كند.در حاليكه روش ما آن ها را تشخيص مي دهد. ولي در كل با وجود تفاوت زياد در الگوريتم MATHREL با اين روشها، توافق رضايت بخشي با اين روش هاي معتبر بدست آمده است كه اعتبار روش ما را تاييد مي كنند. براي مقايسهء MATHREL با روش هاي DSSP, STRIDE و فايل های موجود در PDB ما از معيارهاي شناخته شده از قبيل درصد توافق (%) Sn (Sensitivity) وSp (Specificity) استفاده كرده ايم، $Sn = \frac{TP}{TP+FN}$ $Sp = \frac{TP}{TP+FP}$.

كه در آن TP(True Positive), TN(True Negative), FN (False Negative) وFP (False Positive)

جدول [2]: نتايج مقايسهء MATHREL با روش های ديگر:

روش هاي مقايسه شده	TP	TN	FP	FN	كل	%	Sn	Sp	ساختار

MATHREL_PDB	165145	298456	11468	39541	514610	90.09	81.20	92.81	Helix
MATHREL-DSSP	152995	316351	23618	21646	514610	91.20	87.61	86.63	Helix
MATHREL- STRIDE	158452	316084	18161	21913	514610	92.21	87.85	87.85	Helix
MATHREL-PDB	63369	350001	46102	55138	514610	80.32	53.47	57.89	β-Strand
MATHREL-DSSP	68334	349306	41137	55833	514610	81.16	55.03	62.42	β-Strand
MATHREL-STRIDE	70997	348882	38474	56257	514610	81.59	55.79	64.85	β-Strand

تشكر و قدرداني: بخشي از هزينه اين كار از محل طرح شماره CS 1385-1-02 پژوهشگاه دانش هاي بنيادي (IPM) تامين 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, proteins, 23(4), 566-579.

3 Sussman JL, Lin DW, Jiang JS, Manning NO, Pirlusky J et al(**1998**). Protein Data Bank (PDB): Database of three-dimentional structural information of biological macromolecules, Acta Crystallographica Section D-Biological Crystallography 54,1078-84

Protein Secondary Structure Assignment Based on Mathematical Relations Between Cα Three Dimensional Coordinates

Sayed Rzgar Hosseini [1,2,*], Mehdi Sadeghi[3], Mahnaz Habibi [2,4], Changiz Eslahchi [4]**

1- Biotechnology group, College of Science, University of Tehran
2- Scool of Computer Science, Institute for Studies in Theoretical Physics and Mathematics
3- National Institute of Genetic Engineering and Biotechnology
4- Faculty of Mathematics, Shahid Beheshti University

Abstract:
The automatic assignment of protein secondary structure from three-dimensional atomic coordinate of proteins is an essential step for the analysis and modeling of protein structures. So different methods regarding different criteria have been designed to perform this task. We introduce a new method for protein secondary structure assignment based solely on cα coordinates. We have found four mathematical relations between three-dimensional coordinates of consecutive residues, each of which applies to one of the four regular secondary structure categories (α-helix, 310-helix, π-helix and ⇓-strand). Our algorithm calculates deviation of the cα coordinate of each residue from each of these relations. Then based on a defined cutoff for deviation from each relation, it assigns secondary structures to all residues of a protein.

Keywords: protein secondary structure assignment, cα three-dimensional coordinate, mathematical relations

WWW.IBP.IR iranian bioinformatics portal