پنجمين 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

تركيبي از مدل هاي ماركفي پنها ن و شبكه هاي عصبي براي تعيين سا ختا ر دوم پروتئين ها

پزشك حميد [1و2] *** صادقي مهدي [3] اصلاحچي چنگيز [4] ملك پور سيد امير [1و2]
1 قطب بيومتمتيك وبخش آمار دا نشكده رياضي ، آمار و علوم كامپيوتر، پرد يس علوم دانشگاه تهران
2 هسته تحقيقاتي بيو انفور ماتيك ، پژوهشكده كامپيوتر ، پژوهشگاه دا نشهاي بنيادي (IPM)
3 پژوهشگاه ملي مهندسي ژنتيك و زيست فناوري
4 دانشكده علوم رياضي دانشگاه شهيد بهشتي

چكيده :
پرو تئين ها از توالي اسيد هاي آ مينه ساخته شده اند و براي هر يك از اسيد هاي آ مينه يكي از سه ساختار Helix (H) ، Strand (S) يا Coil (C) در نظر گرفته مي شود كه ساختار دوم پروتئين ها را مي سا زد . اما اين ساختار براي بسياري از توالي هاي پروتئيني نا مشخص است . پيش بيني ساختار دوم پروتئين ها يكي از مراحل مهم در تعيين ساختمان سه بعدي و عملكرد پروتئين ها به شمار مي رود. شبكه هاي عصبي معمولا به طور مستقيم براي محاسبه احتمال هاي نگاشت در مدل هاي ماركفي پنهان به كار گرفته مي شوند. در اين مقاله از ساختار تركيبي جديدي ازمدل هاي ماركفي پنهان و شبكه هاي عصبي استفاده شده است به طوري كه ساختار پيش بيني شده توسط شبكه هاي عصبي را به عنوان يك پيش بيني اوليه براي محاسبه احتمال نگاشت اسيد هاي آمينه در مدل ماركفي پنهان به كار گرفته ايم و پيش بيني جديدي بر اسا س مدل هاي ماركفي پنهان انجام داده ايم . با استفاده از اين ساختار تركيبي دقت پيش بيني مدل هاي ماركفي پنهان را در تعيين ساختار دوم پرو تئين ها افزايش داده ايم.

واژه هاي كليدي:
مدل هاي ماركفي پنهان ، شبكه هاي عصبي ، ساختار دوم پرو تئين ، روشها ي بيزي

1- مقدمه :

امروزه روشهاي متعددي براي پيش بيني ساختار دوم پروتئين بر اساس توالي اسيد هاي آمينه پيشنهاد شده است . پيش بيني سا ختا ر دوم پروتئين از منظر مدل هاي مولد در (1) و(2) بررسي شده است كه يك مدل مولد آماري را بر اساس مدل هاي ماركفي پنهان گسترش يا فته [1] توسعه داده است و از يك دوره به طول متغير براي وضعيت هاي آن استفاده مي شود چنين مدلي اجازه مي دهد در هر وضعيت پنهان طول متغيري از مشاهدات نگاشته شود و اين به مدل نيمه ماركف قطاعي[2] معروف شده است يكي از فوايد استفاده از چنين روشهاي احتمالي اين است كه مي توان منابع متفاوتي از اطلاعات موجود در توالي را با استفاده از توزيع احتمال توام ساختار- توالي و بر اساس ساختار هاي قطاعي وارد مدل كرد . در (3) ، (4) با استفاده از اطلا عات تكاملي پيش بيني بهتري براي SSMM بدست آمده است.

در اين مقاله مدل جديدي براي بهبود دقت SSMM ها به كار گرفته شده است كه يك ساختار تركيبي از شبكه هاي عصبي وSSMMها رابدون در نظر گرفتن اطلاعات تكاملي به كار مي گيرد . ابتدا اسيد آمينه ها را به صورت پنجره هايي به شبكه عصبي داده ايم و سپس پيش بيني هاي انجام شده توسط شبكه عصبي را براي تخمين

---

[1] Generalized Hidden Markov Models(GHMM)
[2] Segmental Semi-Markov Models(SSMM)

پارامتر هاي نگاشت در مدل ماركفي پنها ن و بالا بردن دقت پيش بيني بدست آمده در مرحله اول توسط شبكه عصبي ، به كا ر گرفته ايم .

2- شبكه هاي عصبي :

در اين مقاله شبكه هاي عصبي براي انجام يك پيش بيني اوليه مورد استفاده قرار گرفته است به اين صورت كه توالي اسيد هاي آمينه را به صورت پنجره هاي 9 تايي به عنوان ورودي شبكه عصبي در نظر گرفته ايم و يك پيش بيني اوليه را براي اسيد آمينه اي كه در مركز اين پنجره قرار دارد بدست آورده ايم . براي كاهش تعداد نرون هاي ورودي شبكه عصبي از يك بردار 6 تايي براي كد كردن اسيد هاي آمينه استفاده كرده ايم . همچنين خروجي شبكه عصبي شا مل سه نرون است . شبكه عصبي مورد استفاده يك شبكه پيشرو دو لايه است كه خطا هاي بدست آمده در هر مرحله را در شبكه گسترش مي دهد[3] همچنين 20 نرون براي لايه پنهان در نظر گرفته شده است.

3- مدل هاي ماركفي پنهان :

ساختار دوم پروتئين مي تواند توسط يك سه تايي (m ,e,T) تعريف شود كه m نشان دهنده تعداد قطاع[4] هاي توالي است . e نقاط انتهايي قطاع ها و T ساختارقطاع ها را نشان مي دهد. در شكل(1) يك مدل گرافيكي از ساختار دوم پروتئين براي T=(C,S,C,H,...) و e=(3,5,7,11,...) نشان داده شده است.

هرمشاهده O_i شامل اسيد آمينه R_i وساختار پيش بيني شده توسط شبكه عصبي، P_i است همچنين كل مشاهدات را با O نشان مي دهيم.

براي محاسبه توزيع پسين (m ,e,T) به شرط توالي مشاهدات ، محاسبه توزيع پيشين براي متغير هاي (m ,e,T) كه ساختار دوم پروتئين را مشخص مي كند ضروري است و آن را به صورت زير مي نويسيم :

$$P(\mathrm{m}, \mathrm{e}, \mathrm{T}) = P(m)P(e,T \mid m) = P(m)\prod_{i=1}^{m} P(e_i \mid e_{i-1}, T_i)P(T_i \mid T_{i-1})$$

P(m) احتمال پيشين مشاهده m قطاع را در يك توالي نشان مي دهد . $P(e_i| e_{i-1}, T_i)$ به ما اجازه مي دهد كه طول قطاع ها را به الگو در آوريم وآن را به شكل زيربيان مي كنيم :

$$P(e_i| e_{i-1}, T_i) = P(e_i-e_{i-1} \mid T_i)$$

كه احتمال نگاشت يك مشاهده به طول e_i-e_{i-1} رادر وضعيت T_i نشان مي دهد . چون در مرحله استنباط از m استفاده اي نمي كنيم يك توزيع يكنواخت براي آ ن در نظر مي گيريم . براي محاسبه P(m ,e,T|O) بايد مقدار P(O| m ,e,T) را نيز محاسبه كنيم كه به صورت زيرتعريف مي شود :

$$P(O \mid \mathrm{m}, \mathrm{e}, \mathrm{T}) = \prod_{i=1}^{m} P(S_i \mid S_{-i}, T_i)$$

---

3 feed forward back-propagated network

4 segment

S_i مشاهدات قطاع i ام را نشان مي دهد S_{-i} مشاهدات قطاع هاي قبلي را نشان مي دهد. براي $P(Si \mid S\text{-}i, Ti)$ داريم :

$$P(S_i \mid S_{-i}, T_i) = \prod_{k=e_{i-1}+1}^{e_i} P(O_k \mid O_{[1;k-1]}, T_i) = \prod_{k=e_{i-1}+1}^{e_i} P(P_k \mid O_{[1;k-1]}, T_i) P(R_k \mid P_k, O_{[1;k-1]}, T_i)$$

برآورد احتمال هاي شرطي رابطه بالا حتي بوسيله بزرگترين مجموعه از توالي هاي پروتئيني ممكن نيست به همين جهت ما P_k را فقط روي T_i، P_{k-1} و P_{k-2} شرطي مي كنيم . همچنين براي محاسبه $P(R_k \mid P_k, O_{[1;k-1]}, T_i)$ از شرطي كردن روي $O_{[1;k-1]}$ صرف نظر مي كنيم . مدل هاي ويژه اي نيز براي حذف مشاهد ات زايد و شرطي كردن در (5) ارائه شده است كه از آزمون هاي آماري χ^2 براي يافتن همبستگي بين مشاهدات استفاده مي كند.

با محاسبه توزيع پسين $P(m, e, T|O)$ به كمك روشهاي بيزي مي توانيم توزيع پسين[5] براي ساختار هر مشاهده ، $P(T_{Oi} \mid O)$ ، را بدست آوريم كه T_{Oi} نشان دهنده ساختاري است كه براي O_i در نظر گرفته مي شود و وضعيت هاي S، H، C را اخذ مي كند. براي پيش بيني ساختار O_i بايد ماكزيمم سازي زير را انجام دهيم :

$$\arg\max_T P(T_{O_i} \mid O)$$

4- نتايج :

به دليل محدوديت هاي محاسباتي در اجراي شبكه هاي عصبي ، اين روش را روي مجموعه داده هاي RS126 تست كرده ايم . با استفاده از اطلاعات يك توالي تنها و بدون به كارگيري پروفايل انطباق چندگانه و از تقسيم توالي ها به پنج قسمت تخمين پارامتر ها انجام شده و نتايج بدست آمده در جدول1 آمده است :

جدول1: دقت پيش بيني به روش احتمال پسين

Q_H	Q_E	Q_C	Q_3
65.36	38.14	80.51	67.21

Q_C، Q_E و Q_H به ترتيب دقت پيش بيني را براي وضعيت هاي E, C وH نشان مي دهند و Q_3 دقت پيش بيني براي هر سه وضعيت است .
دقت پيش بيني بدست آمده براي C و Q_3 قابل رقابت با بهترين روشهاي موجود است . همچنين پيش بيني مربوط به وضعيت E كه داراي همبستگي با فاصله هاي طولاني هستند در مقايسه با ساير روشها مناسب است .
ثابت شده است كه استفاده از اطلاعات تكاملي براي توالي ها مي تواند به عنوان راهي براي بهبود پيش بيني كمك بسياري كند كه در كار بعدي آن را مورد بررسي قرار مي دهيم. همچنين با بزرگتر شدن مجموعه داده ها شبكه هاي عصبي پيش بيني اوليه بهتري را ارائه مي كنند.

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	…..

T_1 =C e_1=3 T_2 =S e_2=5 T_3 =C e_3=7 T_4 =H e_4=11

شكل (1) : ساختار دوم پروتئين براي T=(C,S,C,H,...) و e=(3,5,7,11,...) نشان داده شده است.

[5] posterior distribution

قدرداني و تشكر : بخشي از هزينه اين كار از محل طرح شماره CS1385-1-02 پژوهشگاه دانشهاي بنيادي (IPM) تامين شده است .

منابع :

[1] Schmidler C.S, Liu J.S, and Brutlag D.L , "Bayesian Segmentation of Protein Secondary Structure," J. Computational Biology, vol. 7,nos. 1/2, pp. 233-248, 2000.
[2] Schmidler C.S, Liu J.S, and Brutlag D.L, "Bayesian Protein Structure Prediction, " Case Studies in Bayesian Statistics, pp. 363-378, Springer, 2002.
[3] Chu W, Ghahramani Z,Podtelezhnikov A, Wild D,''bayesian segmental models with multiple sequence alignment profile for protein secondary structure and contact map prediction ,''Proc.IEEE,Vol.3,pp. 98-113 ,2006.
[4] Won K.J, Hamerlik T ,Bennet A.P,Krogh A.'' evolving hidden markov models for protein secondary structure prediction,'' Proc.IEEE,pp. 33-44,2005.
[5] Aydin Z, Altunbasak Y, and Borodovsky M, "Protein Secondary Structure Prediction with Semi Markov HMMs," Proc. IEEE Int'l Conf. Acoustics Speech, and Signal Processing, 2004.

Hidden Markov Models together with Neural Networks for Protein Secondary Structure Prediction

Hamid Pezeshk[1,2,*] Mehdi Sadeghi [3] Changize Eslahchi [4] Seyed Amir Malekpour[1,2]**

1 School of Mathematics, Statistics and Computer Science, College of Science
and Center of Excellence in Biomathematics, University of Tehran
2 Bioinformatics Research Group, Institute for Studies in Theoretical Physics and Mathematics (IPM)
3 National Institutes of Genetic Engineering and Biotechnology
4 Faculty of Mathematical Science, Shahid Beheshti University

Abstract:

Proteins are built from amino acids sequences. For each amino acid one of the three structures, helix (H), beta strand(S) or coil(C) are considered as protein secondary structures. For many proteins the secondary structure are not known. Protein secondary structure prediction is an important step in three dimensional structure and function determination of proteins. Neural networks are usually implemented directly for estimating emission probabilities in hidden Markov models. In this paper a new combined structure of hidden Markov models and neural networks is applied. We use the predicted structure of neural network as an initial prediction for estimating emission probabilities of amino acids in a hidden Markov model. Using this combined structure we have increased prediction precision of hidden Markov models for protein secondary structure.

Key Words: Hidden Markov Models, Neural Networks, Protein Secondary Structure, Bayesian Methods

WWW.IBP.IR iranian bioinformatics portal