◎工程与应用◎

基于贝叶斯定理的势能函数应用于蛋白质结构预测

史小红[1],肖宏波[1],肖建华[2]
SHI Xiao-hong[1],XIAO Hong-bo[1],XIAO Jian-hua[2]

1.西安工业大学 数理系,西安 710032
2.华中科技大学 控制系,武汉 430074
1.Department of Mathematics and Physics,Xi'an Technological University,Xi'an 710032,China
2.Department of Control Science and Engineering,Huazhong University of Science and Technology,Wuhan 430074,China
E-mail:shixh@mail.hust.edu.cn

SHI Xiao-hong,XIAO Hong-bo,XIAO Jian-hua.Based on Bayesian theorem energy function for protein structure prediction.Computer Engineering and Applications,2008,44(25):196-198.

Abstract: In this article,an S potential energy function based on Bayesian theorem is presented which is used to test the loop decoys.The fine result shows that this S potential energy function is efficient and suited for loop structure prediction and the prediction ability is better than the RAPDF by the discrimination rate of the loop structure.
Key words: Bayesian theorem;S-potential function;loop-structure prediction

摘 要:改进了一种基于贝叶斯定理的S势能函数并应用于对loop结构的测试取得较好结果,证明S势能函数对loop结构预测是可行的和有效的。测定结果说明S势能函数的辨别率高于RAPDF势能函数,对loop结构具有较好的预测能力。
关键词:贝叶斯定理;S势能函数;loop结构预测
DOI:10.3778/j.issn.1002-8331.2008.25.059 **文章编号:**1002-8331(2008)25-0196-03 **文献标识码:**A **中图分类号:**Q71;TP301

1973年Anfinsen提出蛋白质天然构象对应自由能最小三维结构的著名的热力学假说[1]。利用已知的蛋白质结构对序列已知而结构未知的蛋白质进行结构预测成为人们研究的热点,各种预测方法不断涌现[2-3]。用理论方法预测蛋白质结构有两个难点,第一是要有一个合理的势能函数,第二是要有一个有效的寻优方法找到势函数的全局极小点。一个可替代的方法去获得势能函数是基于知识的统计势能,这些函数通过观察由实验方法测定的蛋白质结构的数据库编纂的参量,直接从已知结构的蛋白质中抽取相互作用势能,这种方法简单而且计算量有效。基于知识的辨别函数常常应用于蛋白质折叠的识别问题,也有效地应用于蛋白质结构的从头预测[4]。近年来,基于知识的势能函数也应用在辨别对接引诱物、从晶体相互作用面上区别真实二聚体相互作用面、蛋白质的loop结构预测等方面[5-6]。这些结果说明基于知识的势能函数捕获到伪装在蛋白质表面,核心,和接触面底下的不同氨基酸残基组成成分的普遍物理相互作用的本质[7]。

任何计算方法预测蛋白质结构都要求建立一个势能函数,能够区别正确的结构和不正确的结构。一般情况,基于知识的势能函数使用一个或二个点表示每一个残基。即,空间上的一个点或二个点的位置表示蛋白质序列中的一个残基。势能函数是基于每个残基的倾向性是:埋藏的还是暴露的,或它的倾向性表示一个特殊的二级结构构像,或是接触的距离。最近,Samudrala和Moult(1998年)建立了一个基于残基种类的全原子间概率的势能函数RAPDF应用于蛋白质的同源模建取得成功[8]。

本文在已有的工作基础上,进一步建立了基于贝叶斯原理的S势能函数。应用该函数对loop诱捕数据集进行了测试,发现最小能量对应正确结构,说明S势能函数具有较好的预测能力,用于loop结构(即非规则二级结构[7])预测是可行的和有效的,测试结果表明提出的S势能函数的辨别率接近或高于同类工作。

1 理论与方法

贝叶斯定理在概率论和数理统计中有着多方面的应用,在工程技术、经济分析、投资决策、药物的临床检验诸多方面有极大的实用价值。所谓的贝叶斯公式就是用来计算后验概率的公式,即要在"结果"发生的条件下,推断"原因"发生的可能性大小。把蛋白质结构的可能构象划分为两个子集合,一个是我们认为正确的构象集合$\{C\}$;剩下的是不正确的构像集合$\{I\}$。用集合$\{V_k\}$表示构像的属性。在目前的应用中,这一属性是构像里的原子间距离的简单集合。我们要求的是在这个构像属性发生的情况下,推断这一构像属性是正确构像的可能性大小。假设每一个距离上产生的概率都是独立的。那么:

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.30700162)。
作者简介:史小红,博士,副教授,研究方向蛋白质结构预测、优化算法及其应用。
收稿日期:2008-03-06 **修回日期:**2008-05-16

$$P(\{d_{xy}^{ij}\}|c)=\prod_{ij}P(d_{xy}^{ij}|c) \quad (1)$$

$$P(\{d_{xy}^{ij}\})=\prod_{ij}P(d_{xy}^{ij}) \quad (2)$$

根据贝叶斯定理:

$$P(c|\{d_{xy}^{ij}\})=P(c)\cdot\prod_{ij}\frac{P(d_{xy}^{ij}|c)}{P(d_{xy}^{ij})} \quad (3)$$

d_{xy}^{ij}是氨基酸残基 x 的 i 原子与氨基酸残基 y 的 j 原子之间的距离。$\{d_{xy}^{ij}\}$是一个蛋白质结构中各种 d 的集合。我们希望计算 $P(c|\{d_{xy}^{ij}\})$,这个构像距离集合$\{d_{xy}^{ij}\}$是正确构像集合的概率。为了计算 $P(c|\{d_{xy}^{ij}\})$,把 $P(c|\{d_{xy}^{ij}\})$表达为一个能够从实验结构中产生的概率计算公式。由于对给定的氨基酸序列 $P(c)$是一个常量,可以不再考虑[8]。应用这个原理,对式(3)取对数形式,建立一个势能函数 S,则辨别一个蛋白质结构的势能函数 S 与此结构是正确结构的概率的负对数成正比:

$$S(\{d_{xy}^{ij}\})=-\sum_{ij}\ln\frac{P(d_{xy}^{ij}|c)}{P(d_{xy}^{ij})} \quad (4)$$

这样给定一个氨基酸序列的构像,计算所有原子对之间的距离并求 S 值。公式左边的 S 值等于求每一类原子的每一个距离上的概率分配比率之和。公式(4)中,$P(d_{xy}^{ij}|c)$可以直接从实验结构的统计中获得。使用一个来自 PDB(Protein Data Bank)的三维结构数据。可以计算在天然构像 c 中观测到的氨基酸残基 x 的原子与氨基酸残基 y 的原子在各种距离格段上的概率 $P(d_{xy}|c)$。则:

$$P(d_{xy}|c)=f(d_{xy})=\frac{N(d_{xy})}{d}N(d_{xy}) \quad (5)$$

这里 $N(d_{xy})$是氨基酸残基 x 的原子与氨基酸残基 y 的原子在一个距离格段 d 上出现的个数。分母是 x–y 在所有格段上出现的数目,则:

$$P(d_{xy})=P(d)=f(d)=\frac{\sum_{xy}N(d_{xy})}{d}\sum_{xy}N(d_{xy}) \quad (6)$$

式中 $\sum_{xy}N(d_{xy})$是所有的原子对在距离格段 d 上出现的数目。分母是所有的原子对在所有的距离格段 d 上出现的数目的总和。简单地说,要求的概率通过在一个蛋白质结构数据库中,计算原子类型对的距离的频数而得到。所有的非氢原子都要被考虑,原子的描述是基于残基种类的,导致 167 种原子种类。例如:缬氨酸残基中的 C_a 原子与甘氨酸残基中 C_a 原子是不同的概念。

测定的数据来自于 PDB 数据库 242 个非同源蛋白质晶体结构的实验数据(同源性<25%)(附录 1)。对于有多条链的蛋白质选择 A 条链作为统计数据,对于数据中有多个位置的坐标我们选择 A 坐标值作为统计数据。把距离格段划分为 2.7Å的格子,范围从 0 到 27Å进行了统计分析。为避免出现零,统计初值赋 1。则得到 14 028×10 的一张权值表。

一个主要问题是如何测试这个 S 势能函数的预测能力。理论上有三种策略:第一,测试基于物理原理的函数与来自小的分子系统的蛋白质结构数据进行比较。第二,使用"decoys"数据集合。即设计许多不正确的结构诱捕数据集合,测试一个能量函数能否区别出错误的构像和实验测得的正确构像。第三,使用辨别函数从一些近似的结构中搜索出一个天然的构像。

表 1　测定蛋白质结构数据 242 个非同源蛋白质的数据

项目	最小数	平均数
$N(d_{xy})$	1	1 760
$\sum_{d}N(d_{xy})$	1 365	19 301
$\sum_{xy}N(d_{xy})$	83 851	27 075 384
$\sum_{d}\sum_{xy}N(d_{xy})$	270 753 840	270 753 840

要测试 S 势能函数的预测能力,使用"decoys"数据集合的方法。设计许多不正确的结构集合,测试 S 势能函数能否区别出错误的构像(即有严重的原子空间的碰撞或叠加和键长键角的不合理结构或违背物理化学规律的结构)及测试辨别率 η。N 表示 decoys 内总的结构数目,N_s 表示势能值大于正确结构的 decoys 内结构数目,则:

$$\eta=\frac{N_s}{N}\times100\% \quad (7)$$

2　结果与讨论

使用的 Loop Decoys 数据来自于:Moult(1986 年)和 Fidelis(1994 年)等人设计的系统的 LOOP 的诱捕数据集合[9-10]。下载数据的网址为 http://dd.stanford.edu/。对 1vfa_47–55(残基数为 9)的目标 loop decoys 集合内 175 个诱捕 loop 结构(C_a RMSD 的范围为:0.663Å~5.287Å)进行测试,对应 S 势能函数的范围为–35.191 到–173.388。实验测定的正确结构 1vfa_47–55 的 S 势能函数值对应最小值为–193.733。辨别率达 100%,最小势能差 ΔS=20.35。同时,应用文献[7]DFIRE 势能函数对 1vfa_47–55 的 175 个诱捕 loop 结构也进行对比测试(表 2)。测试结果表明 S 势能函数的辨别率高于 DFIRE 势能函数。

表 2　比较 S 势能函数与 DFIRE 势能函数的辨别率

1vfa_47–55(9)		S 势能函数			DFIRE 势能函数		
C_aRMSD(Å)范围	decoys 数目	平均值	范围	η	平均值	范围	η
0.663~1.995	79	–117.717	–159.198~–35.191	100%	–5.728	–8.901~5.004	93.7%
2.010~2.998	53	–131.120	–167.632~–41.651	100%	–5.748	–8.636~6.189	96.2%
3.098~3.946	28	–137.029	–166.065~–81.818	100%	–2.460	–8.300~25.491	92.9%
4.062~5.287	15	–125.473	–173.388~–72.231	100%	–3.600	–8.402~13.894	80.0%
总计	175	–125.191	–35.191~–173.388	100%	–5.029	–8.901~25.491	93.1%

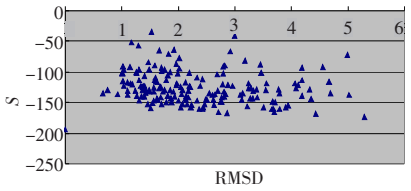

图 1 *S* **势能函数** decoys **测试结果**

图 2 1vfa_47-55 **正确结构(左图)与** decoys **错误结构(右图)**

提出的 S势能函数的预测能力,通过对 loop decoys 的数据进行检测,发现最小能量对应正确结构,说明 S 势能函数具有较好的预测能力,用于 loop 结构预测是可行的和有效的,S 势能函数的预测能力接近或高于同类工作。

3 结论

利用氨基酸序列预测蛋白质结构,需要一个辨别势能函数,能够区别正确结构和不正确结构,建立一个合理的辨别势能函数是关键性的第一步。辨别函数大都基于物理能量函数的概念和统计的原理来建立。论文改进了一类基于知识的统计势能函数,建立了基于贝叶斯定理的 S 势能函数,通过对 242 个正确结构的统计,获得了基于残基种类的 S 势能函数的权值表(14 028×10)。对基于贝叶斯定理的 S 势能函数对1vfa_47-55 的 loop 进行了成功预测,S 最小值对应正确结构。说明提出的 S 势能函数具有较好的预测能力,用于 loop 结构预测是可行的和有效的,测试结果表明提出的 S 势能函数的辨别率接近或高于同类工作。研究结果为进一步研究蛋白质结构预测奠定了理论和实验基础,并扩展了贝叶斯定理的应用范围。

参考文献:

[1] Anfinsen C B.Principles that govern the folding of protein chains[J].Science,1973,181(96):223-230.

[2] 阎隆飞,孙之荣.蛋白质分子结构[M].北京:清华大学出版社,2000:38-212.

[3] 来鲁华.蛋白质的结构预测与分子设计[M].北京:北京大学出版社,1993:2-100.

[4] 赵善荣,唐赟,陈凯先.基于知识的蛋白质结构预测[J].生物化学与生物物理进展,1996,23(5):422-426.

[5] Linhananta A,Zhou H Y,Zhou Y Q.The dual role of a loop with low loop contact distance in folding and domain swapping[J].Protein Science,2002,11(8):1695-1701.

[6] Samudrala R,Moult J.A graph-theoretic algorithm for comparative modeling of protein structure[J].J Mol Biol,1989,279(2):287-302.

[7] Zhang C,Liu S,Zhou Y Q.Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential[J].Protein Science,2004,13(2):391-399.

[8] Samudrala R,Moult J.An all-atom distance-dependent conditional probability discriminatory function for protein structure Prediction[J].J Mol Biol,1998,275(4):985-916.

[9] Moult J,James M N G.An algorithm for determining the conformation of polypeptide segments in proteins by systematic search[J].Proteins,1986,2(1):146-163.

[10] Fidelis K,Stern P S,Bacon D,et al.Critical standard of the prediction of a loop structure[J].Protein Eng,1994,7(7):953-960.

(上接 175 页)

4 结论

为加速规则地形数据场的三维绘制速度,本文建立了双层细节地表模型,并在此基础上给出实时绘制算法。该方法最重要的特点是使用低层次细节层静态表示地形的总体轮廓,用高层次细节层动态对视点所在的局部地形进行详细描述,地表模型表示简单;在实时绘制阶段,通过在图像空间计算高层细节与整个地形投影面积的比例,控制高层次细节覆盖范围,提高绘制效率,并把两层模型的过渡区域推向离视点较远的位置;过渡区域采用了基于图像空间的逐像素纹理混合与偏移方法,避免每一帧直接在图形空间建立几何过渡带计算开销过大的问题,消除了粗、细两层次邻接所产生的视觉裂缝。整个算法在 GPU 进行了设计与实现。实验证明,本文方法在保持逼真视觉效果的同时,能够较大规模地加快三维地形绘制速度。

参考文献:

[1] Duchaineauy M,Wolinsky M.ROAMing terrain:realtime optimally adapting meshes[C]//Proc IEEE Visualization,Phoenix,AZ,USA,1997:81-88.

[2] Nvidia.Vertex Texture Fetch[EB/OL].(2006-03-01).http://developer.nvidia.com/object/using_vertex_textures.html.

[3] Kautz J,Seidel H-P.Hardware accelerated displacement mapping for image based rendering[C]//Graphics Interface.Toronto:Canadian Information Processing Society,2001:61-70.

[4] Engel,Wolfgang F.Direct3D ShaderX vertex and pixel shader tips and tricks[M].[S.l.]:Wordware Publishing,Inc,2002.

[5] Watt A,Policarpo F.3D games animation and advanced real-time rendering[M].London:Pearson Education,2004:145-160.

[6] Fernando R.GPU 精粹-实时图形编程的技术、技巧和技艺[M].北京:人民邮电出版社,2005:4-20.

[7] Gray K.DirectX 9 programmable graphics pipeline[M].Washington:Microsoft Press,2003:67-69.