物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于氨基酸位置特异性的蛋白质Loop区结构预测改进方法

袁飞　张传彪　周昕　黎明

An improved algorithm for prediction of protein loop structure based on position specificity of amino acids

Yuan Fei　Zhang Chuan-Biao　Zhou Xin　Li Ming

引用信息 Citation: Acta Physica Sinica, 65, 158701 (2016)　DOI: 10.7498/aps.65.158701
在线阅读 View online: http://dx.doi.org/10.7498/aps.65.158701
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2016/V65/I15

您可能感兴趣的其他文章
Articles you may be interested in

甲型流感病毒DNA序列的长记忆ARFIMA模型
Long-memory ARFIMA model for DNA sequences of influenza A virus
物理学报.2011, 60(4): 048702　http://dx.doi.org/10.7498/aps.60.048702

基于氨基酸位置特异性的蛋白质Loop区结构预测改进方法*

袁飞 张传彪 周昕 黎明†

(中国科学院大学物理科学学院, 北京 100049)

(2016 年 4 月 22 日收到; 2016 年 5 月 13 日收到修改稿)

蛋白质 loop 区的结构预测是理解蛋白质功能的重要一环, 而长 loop 区的结构预测至今还是生物信息学中的难题. 目前已经出现了多种 loop 结构的算法, 其中 LEAP 是预测精度最高的算法之一, 但它在长 loop 区初始主链构象采样上仍有较大的改进余地. 本文中我们将蛋白质二级结构预测算法 SPINE X 与 LEAP 算法结合起来, 构建了新的主链扭转角分布图 (拉氏图), 在主链初始构象采样中引入氨基酸在蛋白序列中的位置特异性信息, 使得初始构象的采样更具针对性. 对取自 CASP10 单链蛋白的 loop 测试集的分析表明, 对长度为 10, 11, 12 个氨基酸的长 loop 区, 改进后算法都比原始 LEAP 算法的预测精度有显著提升. 这种引入氨基酸位置特异性从而提高预测精度的思路有望进一步推广至 loop 结构预测的其他算法.

关键词: loop 区结构预测, 初始主链构象, 氨基酸位置特异性, 拉氏图

PACS: 87.10.Vg, 87.15.bd, 87.15.bg **DOI:** 10.7498/aps.65.158701

1 引 言

蛋白质是人类生命活动主要的物质承担者, 对蛋白质功能的研究一直是生物物理学的研究热点, 而蛋白质的结构对其功能有决定性意义. 蛋白质结构可以划分为一级结构、二级结构、三级结构、四级结构等四个层次. 一级结构是组成蛋白质的氨基酸序列. 二级结构通常指局部的规则有序的结构, 例如 α-helix, β-sheet. 多个二级结构通过 loop 区 (通常含有一个到多个转角) 串联在一起就形成了三级结构. 四级结构则是指多个蛋白质亚基相互作用形成的有特定功能的复合物结构. 目前, 实验上测定蛋白质的方法主要有 X 射线晶体衍射、核磁共振、低温电镜等. 这些方法一般耗时长、耗资高, 且很多蛋白质的结构目前还难以通过实验手段测定. 另外, 目前新蛋白质的发现速度也远超结构测定的速度, 大量蛋白的结构仍然未知. 而 Anfensen 等 [1] 关于 RNASE 折叠的工作以及大量的后续研究显示, 蛋白质的高级结构可能由其氨基酸序列 (一级结构) 完全决定. 因此, 可以从蛋白质的一级结构出发, 通过计算方法预测二级或三级结构. 这一蛋白质结构的研究思路与实验方法形成互补.

至于蛋白质的生物功能, 除规则的二级结构外,loop 区往往也起着非常重要的作用. loop 区通常处于蛋白质表面, 是信号传递、蛋白质-配体识别等过程的重要参与者. 例如, 近年来发现 loop 区参与构成蛋白质的活性位点与结合位点, 调控抗原与免疫球蛋白的结合 [2]、毒素与蛋白质受体的结合 [3]、金属离子与蛋白质的结合 [4] 等. 因此, 对其结构的测定或预测具有重要意义. 然而, 不少 loop 区往往具有较高柔性, 其结构难以实验测定. 因此, 通过计算方法辅助预测 loop 区结构已成为一种重要的手段.

早期预测 loop 区结构的方法是基于 loop 结构数据库, 例如 COMPOSER [5] 以及 Tossato 等 [6],

* 国家自然科学基金 (批准号: 11105218, 11347614) 资助的课题.

† 通信作者. E-mail: liming@ucas.ac.cn

© 2016 中国物理学会 Chinese Physical Society

Lee等[7]提出的方法. 但是由于目前loop区结构数据不完整, 对待测loop区序列进行同源比对时, 难以找到匹配程度较高的序列和结构, 而且随着loop区长度的增加, 匹配的难度更高, 最终的预测效果也更差. 目前, 对于loop区的结构预测一般都采用基于能量采样的方法, 其主要流程如下: 产生大量loop区主链初始构象, 考虑几何约束及能量高低, 进行初次筛选; 利用旋转异构体文库, 对初筛出的主链构象添加侧链, 进行能量优化之后, 利用更精细的能量打分函数进行二次筛选; 对选定的loop构象, 结合相应的力场, 进行全局的能量最小化, 得到最终的构象. 其中所使用的能量函数主要分为基于物理和基于统计的两大类, 这两类中均有预测精度较高的算法. 本文中预测精度均指预测构象与天然构象的主链重原子(O, C, C_α, N)之间的RMSD (root-mean-square deviation). 在基于物理能量函数的算法中, Spassov等发展的LOOPER算法, 针对Fiser数据库[8]中长度为10, 11, 12个氨基酸的loop 集合, 其预测构象RMSD的平均值分别达到2.66, 3.35, 4.08 Å[9]. 文献[10—12]开发的PLOP (protein local optimization program), 在蛋白质晶体结构信息较完整的情况下, 对较长loop区的预测构象RMSD的平均值均达到2 Å以下[12]. 在基于统计势能的算法中, Soto等开发的LoopBuilder[13,14], 采用Rohl等[15]的数据库, 对长度为8, 12个氨基酸的loop, 其预测构象RMSD的平均值和中位值分别达到1.35 Å/0.99 Å, 3.54 Å/3.11 Å. 对于取自CASP (Critical Assessment of Protein Structure Prediction) 10单链蛋白质的长度为10, 11, 12 个氨基酸的loop 测试集, PLOP 预测精度RMSD的中位值分别为3.2, 3.7, 4.2 Å[16], LoopBuilder的RMSD中位值分别为1.9, 2.0, 2.8 Å[16]. 而Liang等[16]开发的基于物理能量函数的LEAP (loop prediction by energy-assisted protocol) 算法, 其预测构象RMSD的中位值分别达到1.0, 1.5, 2.0 Å. 可见LEAP算法是目前预测精度最高的算法之一. 本文的目的就是对该算法进行改进, 进一步提高预测精度.

LEAP算法的主要流程为[16]: 基于loop区主链扭转角(ϕ, ψ)统计分布图(拉氏图), 对氨基酸的扭转角进行采样, 在满足几何约束的条件下获得大量的初始主链构象 (loop长度为10, 11, 12个氨基酸时, 初始构象采样数目达10^6量级), 采用较粗糙的能量函数由低到高对这些构象进行打分排序, 在此基础上结合其他判据筛选出约10^3个初始候选构象; 利用上述能量函数进一步对初始候选构象进行结构弛豫, 然后添加侧链, 利用精细的侧链势能函数对侧链构象进行初步优化; 固定侧链, 利用较精细的能量函数对主链构象进行能量最小化, 并打分筛选出10个最低能量构象; 采用包含更多势能信息的混合力场, 对筛选出的构象(包括主链和侧链)进行全局能量最小化, 并打分排序, 确定能量最小的最终构象. 从上述流程可以看出, 后续步骤的筛选与能量优化都基于初始构象的选取, 而loop的可能构象数随其长度指数增长. 对于长loop, 这会给初始构象的充分采样带来极大的困难. 在LEAP算法中, 每种氨基酸对应的(ϕ, ψ)统计分布图(下文简称为L拉氏图)都是基于大量loop区结构数据统计得到的, 并不针对特定loop. 这会丢失蛋白结构的一些细节信息, 例如, 即使同种氨基酸, 当其处于蛋白序列的不同位置时, 其构象扭转角的统计分布也并不相同. 如果在生成loop主链的初始构象时能计及这种位置依赖性, 就有可能采到更接近天然构象的构象, 从而提高最终构象的预测精度. 虽然目前对于loop区还缺乏这种细致的统计信息, 但对于蛋白质二级结构, 由于其结构数据非常丰富, 这类信息更容易获取和应用. 有些二级结构预测算法能够对每个氨基酸的(ϕ, ψ)给出粗略的预测结果, 如SPINE X (prediction of structural properties of proteins by integrated neural networks 5th edition)[17], 其平均偏差约为35°[17]. 本文尝试将SPINE X提供的位置相关的氨基酸构象信息引入到LEAP算法中, 生成新的扭转角分布图(简称L-S拉氏图), 以此进行主链初始构象的采样. 下面我们先介绍生成L-S拉氏图的具体方法, 举例说明此改进对预测的影响, 然后采用CASP10测试集中的loop数据, 对总体预测效果进行对比分析.

2 研究方法

L拉氏图以10°为一个角度区间, 统计氨基酸的扭转角(ϕ, ψ)在不同区间的出现频次. 不同类别氨基酸对应不同的L拉氏图, 但对处于特定loop区的氨基酸而言, 这种拉氏图不能反映近邻氨基酸对中间氨基酸(ϕ, ψ)取值的影响. 而SPINE X以该氨基酸及近邻氨基酸的位置特异性打分矩阵和7类

物理参数 (立体参数、疏水性、体积、极化率、等电点、helix 倾向性、sheet 倾向性) 作为输入量, 给出该氨基酸扭转角的最可能值. 我们可以在拉氏图中提高 SPINE X 所预测的主链 (ϕ, ψ) 所占比重, 生成新的拉氏图作为主链初始构象的采样依据.

具体做法如下:

1) 对特定氨基酸, 将 SPINE X 预测出的 ϕ, ψ 在拉氏图中定位, 判断出它所处的区间;

2) 将定位区间对应的统计频次直接提升为 L 拉氏图中的最高出现频次, 进一步, 考虑到 SPINE X 预测的误差范围, 我们还将该区域最近邻的 8 个区间对应的统计频次也赋为这个最大值, 以此提高 SPINE X 预测值被采样到的概率;

3) 依据上一步得到的新的 L-S 拉氏图, 采样产生大量主链初始构象, 并按照 L 算法 (即 LEAP 算法) 的后继流程对 loop 区构象进行筛选和优化, 确定最终的预测构象.

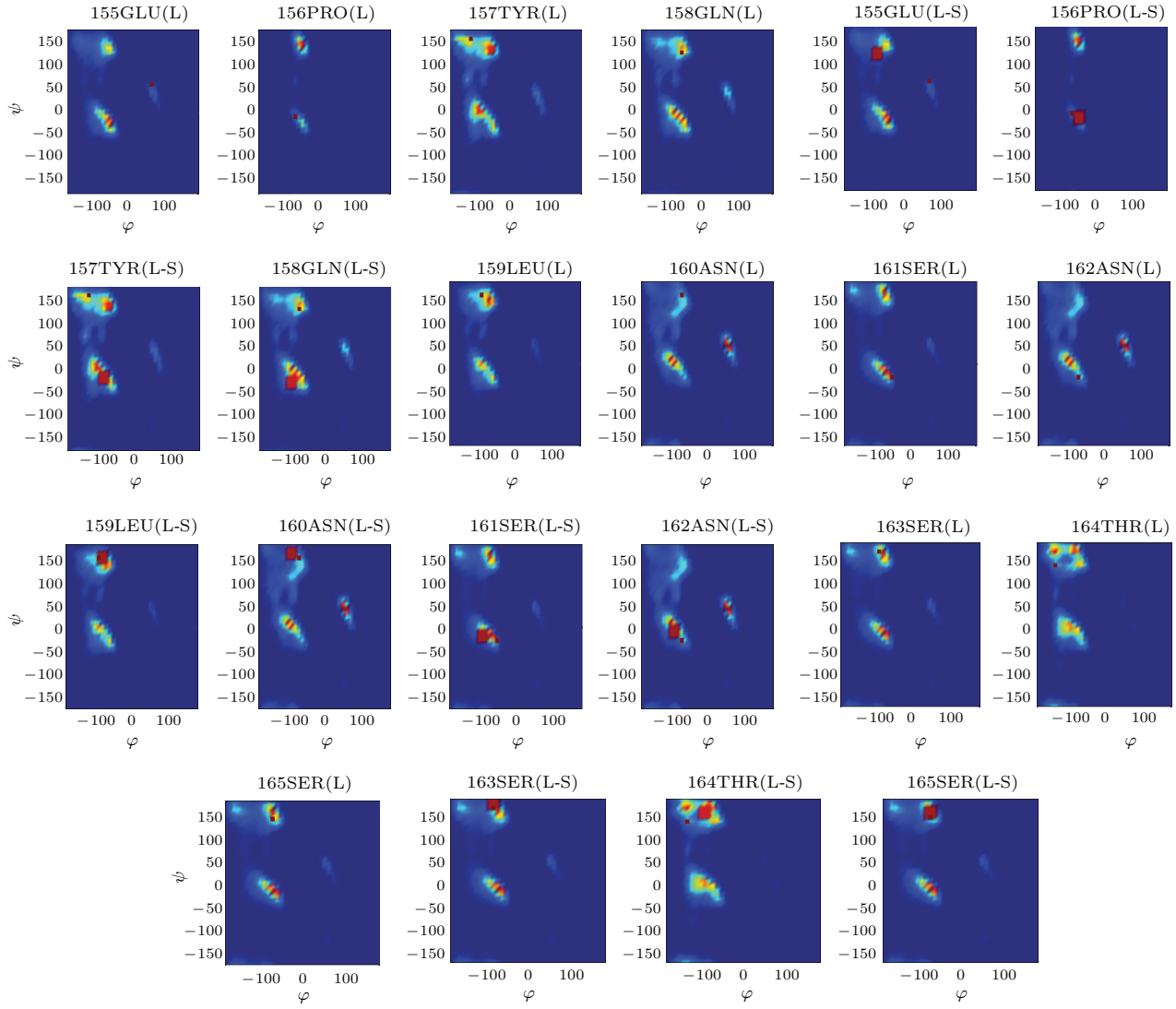

图 1 (网刊彩色) 蛋白质 4h0a 某 loop 区 (氨基酸编号 155—165) 上每个氨基酸扭转角的 L 拉氏图与 L-S 拉氏图. L-S 拉氏图中的方形红色色块表示统计频次提升至最高的区域, 深红色斑点为天然构象扭转角所处区间, 各图中的颜色均表示出现次数的绝对值, 但色度标尺并不一致 (细节从略)

Fig. 1. (color online) L and L-S Ramachandran plots of torsion angle of each amino acid in the loop region (amino acid ID 155-165, of protein 4h0a). The red squares in L-S Ramachandran plots indicate the regions whose occurrence probabilities have been elevated to the maximum, and the deep red spots indicate the torsion angles in the native conformation. The color of each plot represents the absolute occurrence number in each region, and the scale bar differs for each plot (details not give here).

下面我们以蛋白质4h0a的某个loop区(氨基酸编号为155—165)为例, 对上述流程做进一步阐释. 利用SPINE X预测出该loop区所有氨基酸的主链扭转角, 判断出它们在拉氏图所处的区间. 图1显示了对应的L拉氏图与L-S拉氏图, 并在拉氏图中标注了天然构象扭转角所处的位置. 我们在拉氏图中使用不同的颜色区分不同的出现频次, 颜色由蓝到红, 表示频次越来越高.

由图1可知, 氨基酸156PRO, 159LEU, 160ASN, 161SER, 162ASN, 163SER, 164THR, 165SER的SPINE X预测结果均覆盖或者紧邻该氨基酸天然构象的扭转角, 其L-S拉氏图的扭转角分布比相应的L拉氏图更加接近天然构象. 对155GLU, 157TYR, 158GLN, 虽然L拉氏图分布比L-S拉氏图更加接近天然构象, 但两者都远离天然构象. 因此, 统计上看, 利用L-S拉氏图会采到更优(更接近天然构象)的主链初始构象. 图1还显示, 即使对于同一种氨基酸(例如其中的三个SER和两个ASN), 由于在蛋白序列中所处位置不同, SPINE X给出的扭转角都会有所差异, 这正是SPINE X考虑到位置特异性的结果. 值得一提的是, 160ASN的天然构象扭转角在L拉氏图中出现概率很小, 但在L-S拉氏图中的出现概率较大. 从图2中可以看出, 正是从160ASN开始, L算法和L-S算法(即改进后算法)采样所得的初始构象出现了较大偏离. 该图还显示, 无论是L算法还是L-S算法, 最终筛选出的loop 构象与其对应的初始构象相比只有很小偏差. 换句话说, 初始主链构象越接近天然构象, 最终就可能获得更优的预测构象, 从而提高loop结构的预测精度.

从上例可以看出, 当SPINE X对loop区各氨基酸的(ϕ, ψ)给出较好的预测时, 我们通过L-S算法就能采到更优的主链初始构象, 并得到更优的最终构象. 如果SPINE X的预测结果很差, 则可能适得其反. 蛋白质4gpv中氨基酸编号为190—199的loop区就是一个较为极端的例子. 由图3可看出, 对于氨基酸193GLN, 195THR, 196GLY, 197ALA, 199ASP, L拉氏图的扭转角分布比L-S拉氏图更加接近天然构象, 而对氨基酸191PHE, 192ASN, 194TYR则反之. 虽然两组拉氏图中接近天然构象的分布都不多, 但是L-S拉氏图总体上更加偏离天然构象, 因此L-S算法更可能采到远离天然态的主链构象, 这一点明确体现在图4中. 这也反过来说明初始构象采样的重要性. 此外, 与上例类似, 我们也发现主链最终构象与它所对应的初始构象非常接近. 由这两个例子可以看出, 主链的初始构象基本上就决定了最终的预测构象, 其他步骤(添加侧链, 以及各类不同力场的能量优化)都只是在此基础上对结构进行更精细的调整. 图5表明, 最终预测结果更好的算法, 往往也是在初始阶段采到更优候选构象的算法. 如果利用L-S拉氏图能够更大概率地采到这种初始构象, 那就有可能提高loop区的预测精度. 下文我们将对loop区集合进行测试. 为定量刻画主链构象之间的偏差程度, 我们不仅会用到空间构象RMSD这一常用指标, 而且由于本文的讨论是基于主链扭转角分布的, 下面我们还将直接用到主链扭转角的RMSE (root-mean-square error, 由空间构象文件计算所得).

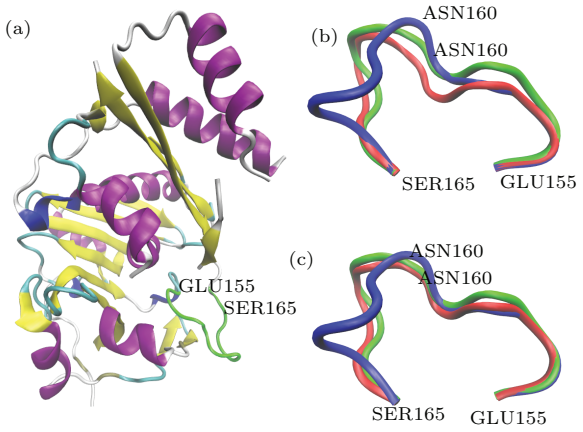

图2 (网刊彩色) 蛋白质4h0a及loop区的构象 (a) 4h0a的蛋白质构象, loop区(氨基酸编号155—165)用绿色标识; (b) 绿色表示该loop的天然构象, 蓝色表示(c)图中L算法最终构象所对应的初始构象, 红色表示(c)图中L-S算法最终构象所对应的初始构象; (c) 绿色表示该loop的天然构象, 蓝色为L算法预测的最终构象, 红色为L-S算法预测的最终构象.

Fig. 2. (color online) The conformations of protein 4h0a and the loop region: (a) The conformation of protein 4h0a, with the loop region (amino acid ID. 155-165) colored in green; (b) the native conformation of the loop is colored in green; the blue one is the initial conformation corresponding to the final conformation predicted by L algorithm, the red one is the initial conformation corresponding to the final conformation predicted by L-S algorithm; (c) the green line denotes the native conformation of the loop, the blue one is the final conformation predicted by L algorithm, and the red one is the final conformation predicted by L-S algorithm.

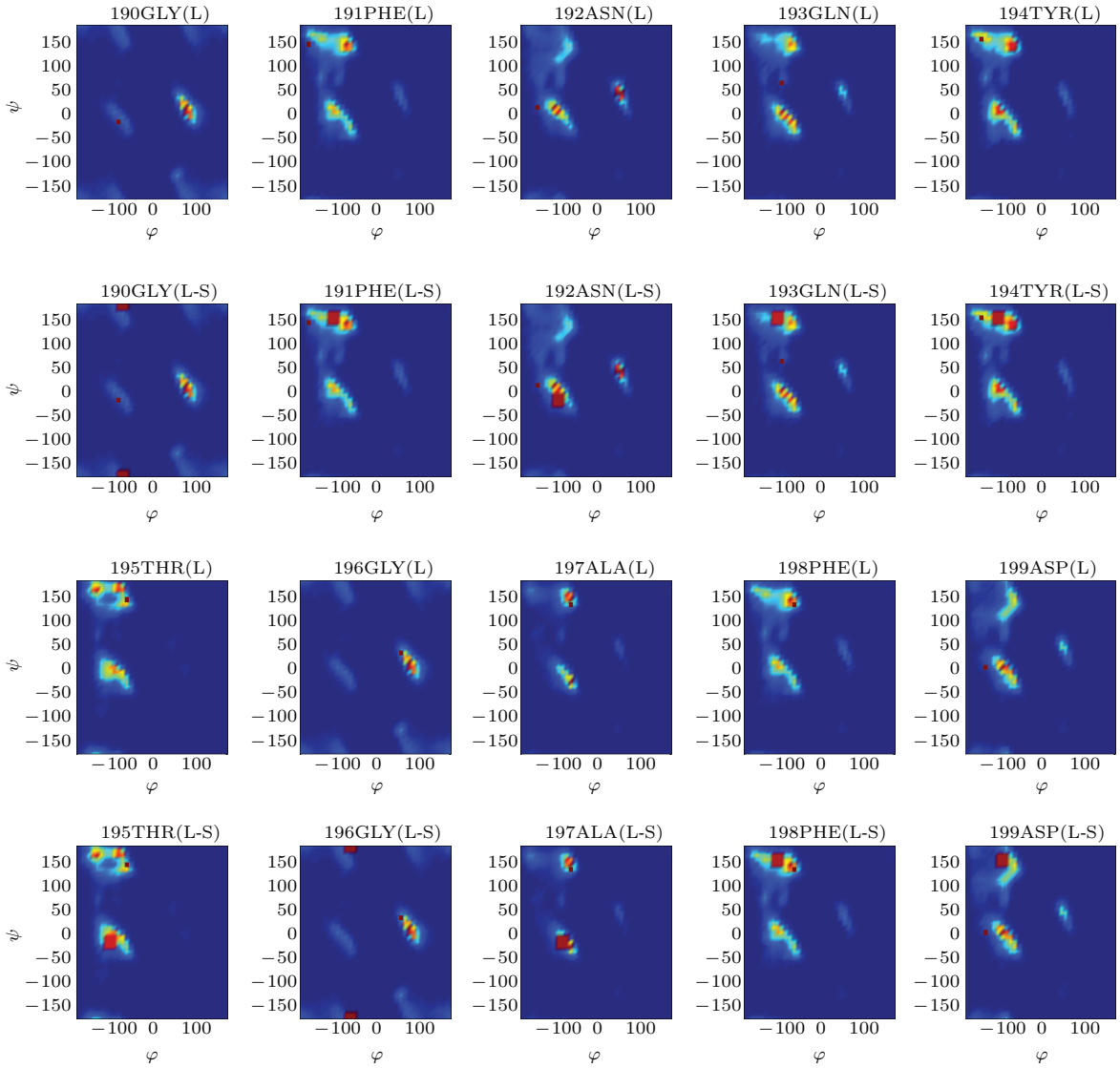

图 3 (网刊彩色) 蛋白质 4gpv 的 loop 区 (氨基酸编号 190—199) 中各氨基酸扭转角的拉氏图 (颜色说明同图 1)

Fig. 3. (color online) Ramachandran plots of each amino acid in the loop region (amino acid ID.190—199) of protein 4gpv. For the explanations of the color, see Fig.1.

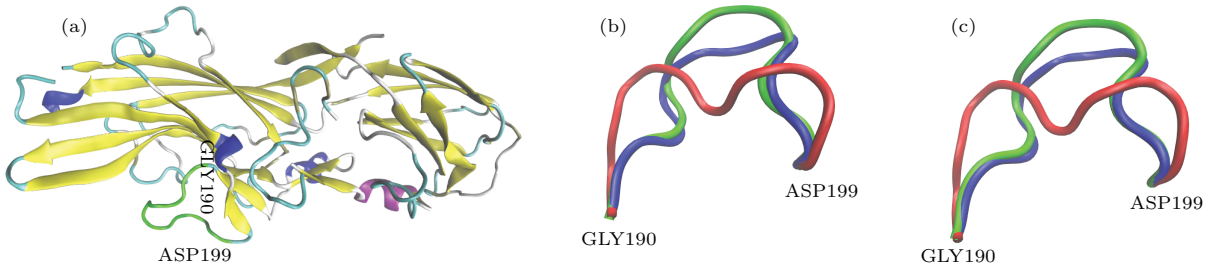

图 4 (网刊彩色) 蛋白质 4gpv 及 loop 区构象 (a) 蛋白 4gpv 的结构, 其 loop 区 (氨基酸编号 190—199) 用绿色标识; (b), (c) 的说明同图 2

Fig. 4. (color online) The conformations of protein 4gpv and the loop region: (a) The conformation of protein 4gpv, with the loop region (amino acid ID. 190–199) colored in green; for the explanations of the colors in (b) and (c), see Fig.2.

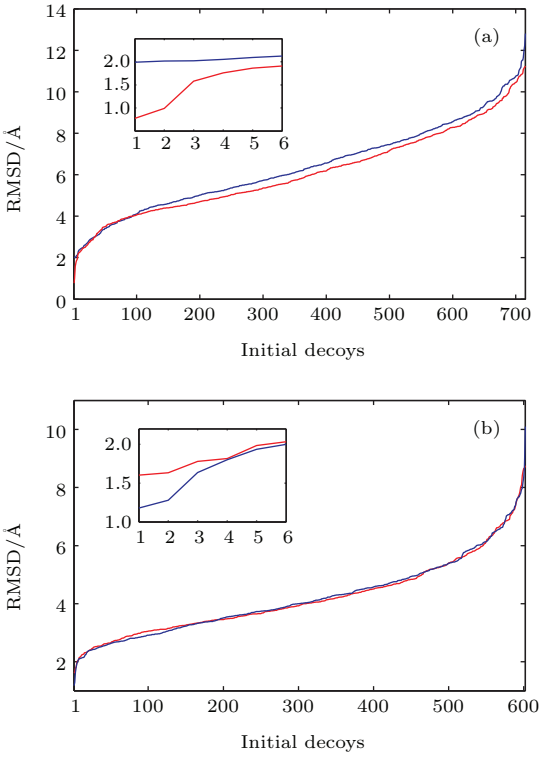

图5 (网刊彩色) 初始候选构象与天然构象之间的RMSD (a) 蛋白4h0a; (b) 蛋白4gpv; 红线表示L-S算法的结果, 蓝线表示L算法的结果, 横轴initial decoy为初始候选构象编号 (按RMSD大小编号); 两插图所示为最优、次优等初始构象相应的RMSD, 可以看出, (a) 中L-S算法采到的最优初始构象明显好于L算法, (b) 中则相反

Fig. 5. (color online) The RMSDs between initial decoys and native conformation, for (a) protein 4h0a and for (b) protein 4gpv. Blue lines denote the results given by L algorithm, red ones denote the results given by L-S algorithm. The horizontal axis indicates the initial decoy numbered by their RMSD. The inset in either figure denotes the RMSD of the optimal initial conformation and several suboptimal ones. It can be seen that L-S algorithm outperforms L algorithm in sampling the better initial decoys for (a), and L algorithm outperforms L-S algorithm for (b).

3 结果与讨论

我们以Liang等[16]构建的取自CASP10单链目标蛋白 (4f67, 4fmw, 4hqf, 2ymv, 2luz, 4ftd, 4gl6, 4hg2, 4gpv, 4epz, 4fgm, 4f54, 4fd0, 4fr9, 4fs7, 4g2a, 4gt6, 4h09, 4e6f, 4fdy, 4h0a) 的loop集合为例, 对L算法和L-S算法进行进一步的统计分析和对比. 表1中列出了不同长度的loop相应的数量.

我们的关注点主要在长链loop, 对长度为10, 11, 12个氨基酸的每个loop, 分别用两种算法独立预测10次. 首先来看最终预测构象与其对应的初始构象之间的依赖关系. 以长度为12个氨基酸的loop集合为例, 我们在图6中给出初始构象、最终构象、天然构象两两之间扭转角的RMSE. 可以看出, 无论哪种算法, 其初始构象与最终构象之间的RMSE都非常小, 两者与天然构象之间的差别都非常接近. 对于长度为10, 11个氨基酸的loop也有类似结论. 表2显示了对于不同长度loop区集合, 两种算法给出的三种构象之间RMSE的平均值. 可以看出, 不论是最终构象还是与之对应的初始构象, L-S算法总体上都比L算法更接近天然构象.

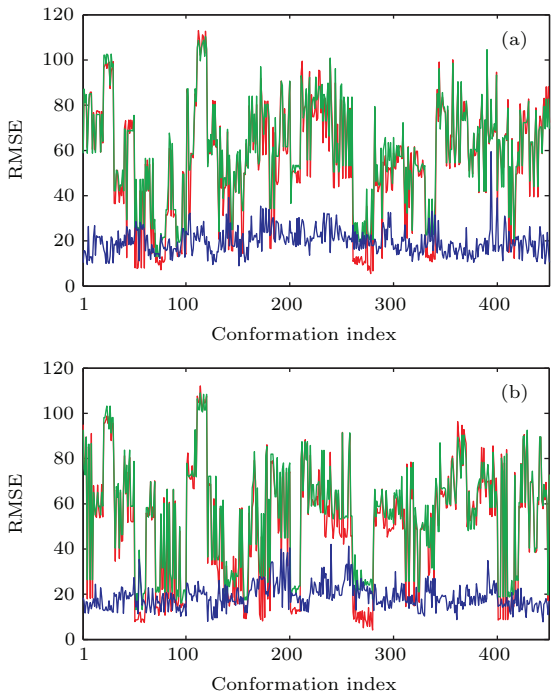

图6 (网刊彩色) 长度为12个氨基酸的loop集合的初始构象、最终预测构象、天然构象之间的RMSE (a) L算法; (b) L-S算法; 其中红线为最终构象与天然构象之间的RMSE, 绿线为初始构象与天然构象之间的RMSE, 蓝线为初始构象与最终构象之间的RMSE; 横轴conformation index为所有loop的预测构象的编号 (每个loop预测10次), 例如第1个loop的构象所占编号为1到10

Fig. 6. (color online) RMSEs between the initial conformations, the final conformations and the native conformations of loops of length 12: (a) The results given by L algorithm; (b) the results given by L-S algorithm. The red lines denote the RMSEs between the final conformation and the native conformation, the green ones denote the RMSEs between the initial conformation and the native conformation, and the blue ones denote the RMSEs between the initial conformation and the final conformation. The horizontal axis is the index of all predicted conformations of all loops (the algorithm is performed 10 times for each loop), e.g. the index 1 to 10 correspond to the 10 predicted conformations of loop 1.

表 1 不同长度的 loop 的数量

Table 1. The numbers of loops of different length.

loop 长度/氨基酸	4	5	6	7	8	9	10	11	12
loop 数量	413	276	225	146	126	83	74	51	45

图 6 和表 2 中的初始构象均为最终构象所对应的初始构象, 而这些初始构象存在于大量初始候选构象之中, 因此要想通过逐级筛选得到更优的最终构象, 就应当尽可能高概率地采集到接近天然构象的初始构象. 我们统计出两种算法的初始候选构象, 计算出这些候选构象与天然构象之间的 RMSE, 并做出相应的统计分布图 (图 7 (a)). 按该图所示, 如果只考察初始候选构象与天然构象的总体接近程度, 则 L-S 算法并不明显优于 L 算法. 但正如图 5 所示, 决定最终构象的往往只是初始采样所得的几个最关键的近天然构象, 而非初始构象总

表 2 初始构象、最终构象和天然构象三者之间 RMSE 的平均值 (loop 长度为 10, 11, 12)

Table 2. The mean RMSEs between the initial conformation, the final conformation and the native conformation (for loops of length 10, 11, 12).

	RMSE 的平均值		
	10	11	12
初始构象 -天然构象 (L/L-S)	48.7/45.8	55.0/50.6	59.6/55.2
最终构象 -天然构象 (L/L-S)	42.4/39.7	50.8/45.6	56.3/51.3
初始构象 -最终构象 (L/L-S)	19.3/18.2	18.7/18.7	19.7/19.0

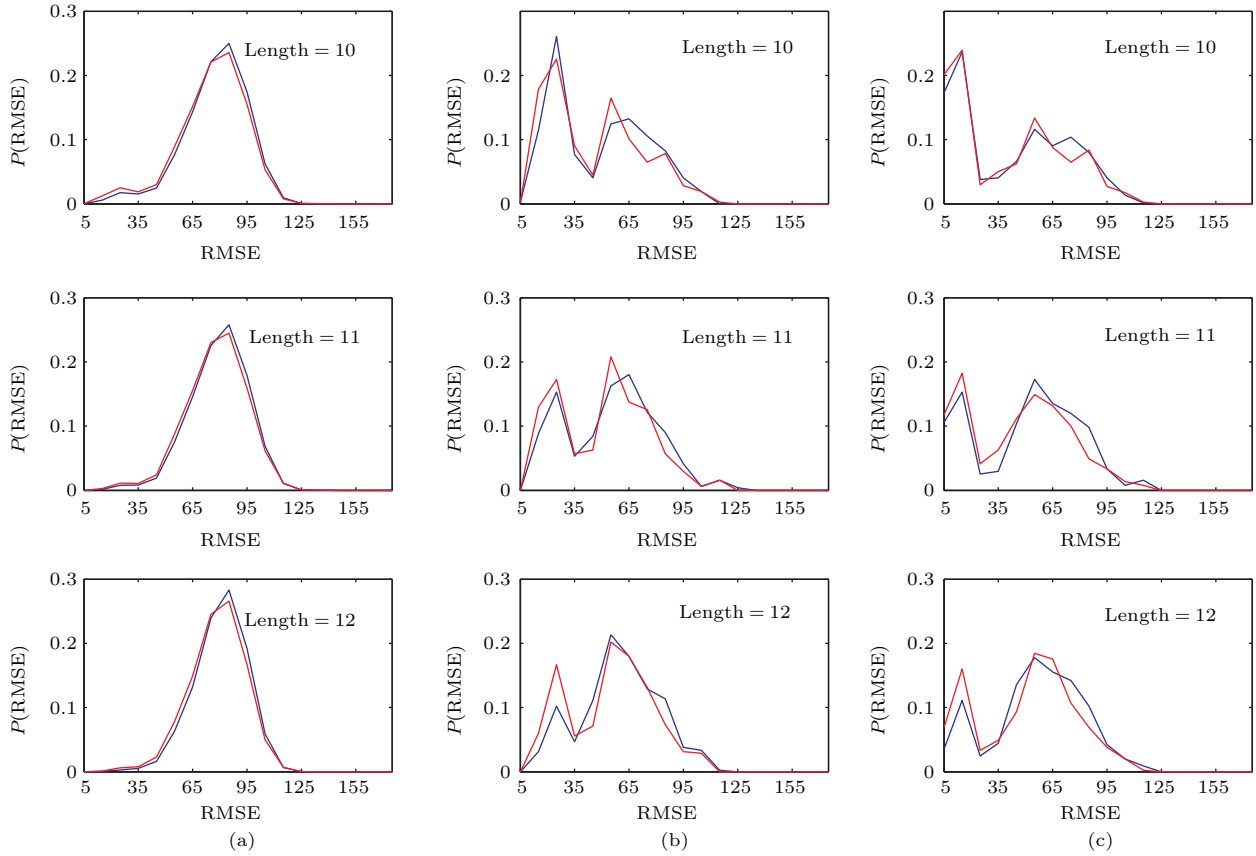

图 7 (网刊彩色) 各类构象 RMSE 的统计分布 (a) 初始候选构象的 RMSE 分布, 每个 loop 均预测 10 次, 每次产生近 10^3 个候选构象, 各子图的统计样本包括该长度下所有 loop 的所有初始候选构象, loop 长度为 10, 11, 12 个氨基酸时, 样本总量分别约为 4.3×10^5, 3.2×10^5, 3.0×10^5 个; (b) 最终构象所对应的初始构象的 RMSE 分布, 统计样本量同 (c); (c) 最终预测构象的 RMSE 分布, 每个 loop 共预测 10 次, 对应 10 个最终预测构象, loop 长度为 10, 11, 12 个氨基酸时, 统计样本量分别为 740, 510, 450 个; 其中蓝线表示 L 算法的结果, 红线表示 L-S 算法的结果

Fig. 7. (color online) The RMSE distribution of conformations: (a) The RMSE distribution of initial decoys, each loop is predicted 10 times and about 10^3 initial decoys are generated each time, the sample of each subgraph contains all the initial decoys of all loops of the same length, the sample size is 4.3×10^5, 3.2×10^5, 3.0×10^5 respectively for loop length 10, 11, 12; (b) the RMSE distribution of initial conformations corresponding to final conformations given in (c), the sample size is the same as (c); (c) the RMSE distribution of the final conformations, ten final conformations are obtained for each loop, the sample size is 740, 510, 450 respectively for loop length 10, 11, 12. Blue lines denote the results given by L algorithm, red lines denote the results given by L-S algorithm.

体. 图7(b)显示了与最终预测构象相应的初始构象接近天然构象的程度, 可以看到, L-S 算法能采到这类最关键初始构象的概率明显高于L算法. 图7(c)给出最终预测构象与天然构象的接近程度, 如图所示, L-S算法最终预测构象也明显优于L算法. 图7(c)与图7(b)的相似性再次表明, 无论L算法还是L-S算法, 能否采到接近天然构象的初始主链构象, 是决定最终预测精度的最关键因素.

为了更直观地说明L-S算法对长loop区预测精度的改善程度, 我们在表3中给出了L算法与L-S算法预测精度RMSD的中位值和平均值. 可以看出, 对于不同长度的loop, 无论是取中位值还是平均值, L-S算法的预测精度均有不同程度的提高.

表3 L算法和L-S算法最终构象RMSD的中位值和平均值

Table 3. The medians/means of RMSDs of the final conformations predicted by L algorithm and L-S algorithm.

	RMSD(Å)(中位值/平均值)		
	10	11	12
L算法	0.97/1.51	1.48/1.95	2.24/2.54
L-S算法	0.85/1.38	1.23/1.68	1.77/2.27

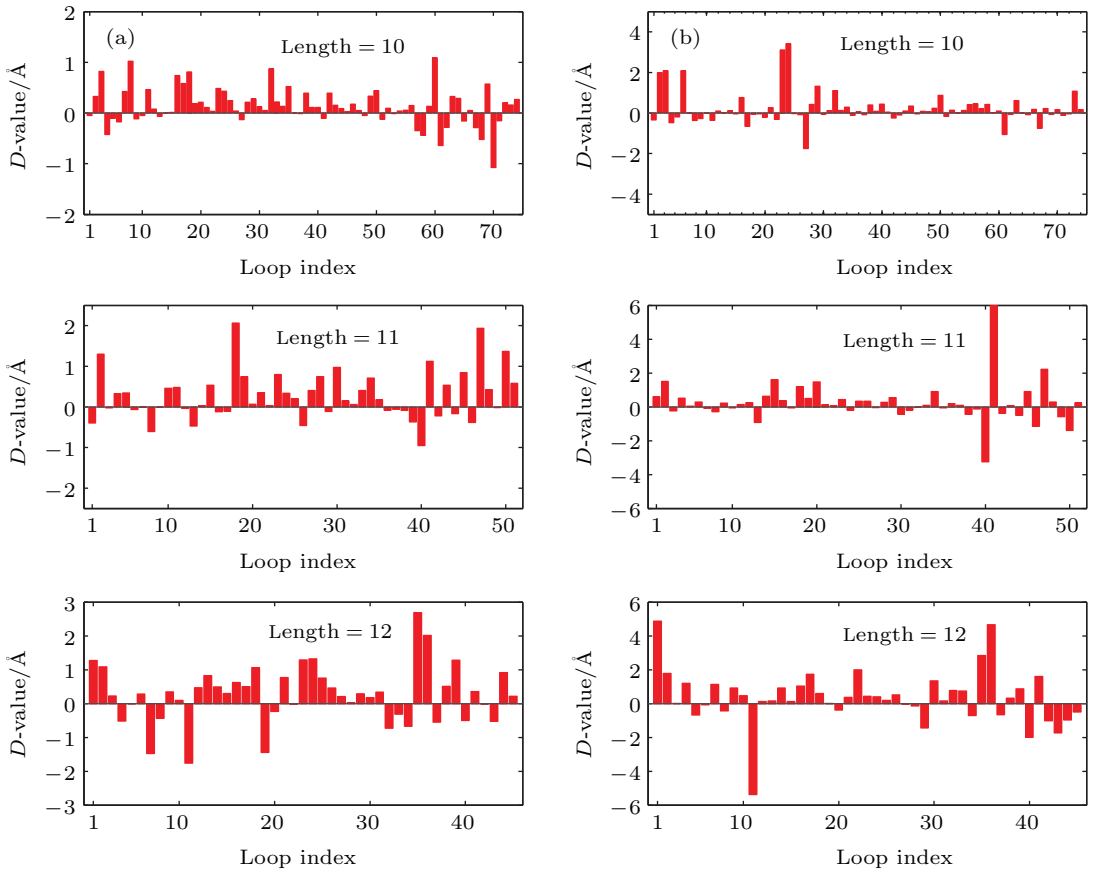

图8 (网刊彩色) L算法与L-S算法预测的loop最终主链构象RMSD之差值 (a)RMSD均值之差; (b)相对最优构象的RMSD之差; loop index为loop的编号

Fig. 8. (color online) *D*-values of backbone RMSDs predicted by L algorithm and L-S algorithm: (a) *D*-value of the mean RMSDs; (b) *D*-value of the RMSDs of the optimal conformations. Loop index shows the numbering of the loops.

为获得更精细的对比信息, 对于各个loop区, 我们计算了L算法与L-S算法所预测的最终构象RMSD之差(即, $\mathrm{RMSD_L}$-$\mathrm{RMSD_{L\text{-}S}}$, 简记为*D*-value), 以此来表征L-S算法相对L算法的改进程度. 由图8(a)可知, 对每个loop做10次独立预测, 对于loop长度为10, 11, 12个氨基酸的情形, 就其RMSD均值而言, L-S算法与L算法各自预测更准的loop个数之比(即,D-value> 0的loop数与D-value< 0的loop数之比)分别为51 : 23, 30 : 21, 30 : 15, 可见L-S算法预测得更准的概率比原始L算法高出约1倍左右. 对每个loop, 10次独立预测结果中还有一个相对最优构象, 考察其对应的RMSD(图8(b))也会发现, 无论是预测精度还是预测更准的概率, L-S算法也都优于L算法.

4 总结与展望

loop区在蛋白质的生物功能中扮演着重要角色, 关于其结构预测已发展出多种算法, 目前所有常见算法中预测精度最高的是LEAP 算法. loop区结构预测本质上是针对特定目标函数的全局优化问题. LEAP 算法是从大量(最高至百万量级)可能初始构象中寻找局部优化构象、并从中选取相对最优构象, 其最终预测结果可能依赖初始构象的选取. 我们针对loop测试集的计算分析证明, LEAP算法的确具有极其显著的初始构象依赖性. 因此, 初始主链构象的采样在整个流程中是至关重要的一环. 然而, 原始LEAP算法在生成初始构象时所采用的是非常粗略的拉氏图, 未能针对性地考虑特定loop 区的序列信息, 在采样时可能会偏离主链天然构象较远, 影响最终的预测精度. 为克服这一困难, 我们尝试在LEAP算法中利用蛋白质二级结构预测的结果, 即引入SPINE X算法将原始的L 拉氏图改造为新的L-S拉氏图, 在此基础上进行采样和后继的各种优化. 我们对较长loop区测试集的计算表明, 与原始L算法相比, L-S 算法的确能采到更加接近天然构象的主链初始构象, 从而实质性地提高最终构象的预测精度, 例如, 对长度为10, 11, 12的loop, 就预测精度RMSD的中位值/平均值而言, L-S算法比L算法分别有0.12 Å/0.13 Å, 0.25 Å/0.27 Å, 0.47 Å/0.27 Å的提高. 这说明在拉氏图中引入氨基酸位置特异性信息对于提升预测精度是有效的. 按照这一思路, 未来可考虑如何进一步细化或优化这种信息, 例如, 最近Heffernan等[18] 开发的的迭代深度学习方法对蛋白质主链扭转角预测精度比SPINE X又有进一步提高, 未来可以考虑用这个算法的结果改造拉氏图, 有可能进一步提高loop 结构的预测精度. 此外, 还可以结合loop区结构数据库信息进一步提高初始构象采样有效性, 例如, 可将待测loop与数据库进行序列同源性比对. 对于具有很高序列相似性的loop 片段, 其构象可直接采用数据库中的同源结构; 对于其他片段的氨基酸, 其构象仍可依据L-S 拉氏图进行采样.

目前所有的全局优化方法都不同程度上依赖初始构象选取, 本文对初始构象采样的改进思路能与现有全局优化方法结合, 例如, 代替LEAP中的局部优化方法, 采用模拟退火方法等, 同时采用本文类似的更有效的初始构象采样方案, 这可能会进一步增加构象优化的效率.

感谢澳大利亚格里菲斯大学的周耀旗教授、日本大阪大学梁世德博士的讨论以及中国科学院大学物理学院宋永顺、杨成、李阳、徐顺的帮助和讨论.

参考文献

[1] Anfinsen C B, Redfield R R, Choate W L, Page J, Carroll W R 1954 *J. Biol. Chem.* **207** 201
[2] Decanniere K, Muyldermans S, Wyns L 2000 *J. Mol. Biol.* **300** 83
[3] Likitvivatanavong S, Aimanova K G, Gill S S 2007 *FEBS Lett.* **583** 2021
[4] Lepsik M, Field M J 2007 *J. Phys. Chem. B* **111** 10012
[5] Sutcliffe M J, Haneef I, Carney D, Blundell T L 1987 *Protein Eng.* **1** 377
[6] Tossato C E, Bindewald E, Hesser J, Maenner R 2002 *Protein Eng.* **15** 279
[7] Lee J, Lee D, Park H, Coutsias E A, Seok C 2010 *Proteins: Struct., Funct., Bioinf.* **78** 3428
[8] Fiser A, Do R K, Sali A 2000 *Protein Sci.* **9** 1753
[9] Spassov V Z, Flook P K, Yan L 2008 *Protein Eng., Des. Sel.* **21** 91
[10] Jacobson M P, Pincus D L, Rapp C S, Day T J F, Honig B, Shaw D W, Friesner R A 2004 *Proteins: Struct., Funct., Bioinf.* **55** 351
[11] Zhu K, Pincus D L, Zhao S W, Friesner R A 2006 *Proteins: Struct., Funct., Bioinf.* **65** 438
[12] Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner R A 2011 *Proteins: Struct., Funct., Bioinf.* **79** 2794
[13] Xiang Z, Soto C S, Honig B 2002 *Proc. Natl. Acad. Sci. U. S. A.* **99** 7432
[14] Soto C S, Fasnacht M, Zhu J, Forrest L, Honig B 2008 *Proteins: Struct., Funct., Bioinf.* **70** 834
[15] Rohl C A, Strauss C E M, Chivian D, Baker D 2004 *Proteins: Struct., Funct., Bioinf.* **55** 656
[16] Liang S, Zhang C, Zhou Y 2014 *J. Comput. Chem.* **35** 335
[17] Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y 2012 *J. Comput. Chem.* **33** 259
[18] Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y 2015 *Sci. Rep.* **5** 11476

An improved algorithm for prediction of protein loop structure based on position specificity of amino acids*

Yuan Fei Zhang Chuan-Biao Zhou Xin Li Ming†

(*School of Physical Science, University of Chinese Academy of Sciences, Beijing 100049, China*)

(Received 22 April 2016; revised manuscript received 13 May 2016)

Abstract

Loop region is necessary structural element of protein molecule, and plays significant roles in protein functioning, e.g., in signaling, ligand recognition. Unlike the well-defined secondary structures (i.e., helix, sheet), however, loop regions vary in structure and some of them are even not able to be measured by ordinary experimental methods. For these reasons, computer-aided prediction of loop structure became a hotspot in bioinformatics and biophysics. Sorts of algorithms have been developed for this purpose. So far, however, the prediction of long loop is still a challenge. Among all the common algorithms, LEAP algorithm achieves the highest precision on long loop prediction. Our investigation on a test data set with LEAP algorithm reveals that the ultimate loop structure predicted by LEAP is almost entirely determined by the initial sampling of the conformation of the loop backbone. If all the backbone conformations in the initial sampling are quite distant from the real (native) conformation, the ultimately predicted structure is also distant from the native conformation, and the prediction accuracy cannot be improved obviously only by increasing the computation time. In the original LEAP, the initial sampling is based on the rough distribution of the backbone torsion angle (Ramachandran plot, R-plot) which doesn't consider the sequence information of the loop region. Many conformations which are far from the native conformation are most likely generated in the sampling. So there raises the open question, is it possible to enhance the initial sampling to be more targeted to the native conformation? In this paper, we suggest an approach to introduce the position-specific amino-acid sequence information into the initial sampling of the backbone conformation, which may generate more targeted initial decoys. An algorithm of protein secondary structure prediction, SPINE X, is used to generate rough but reasonable estimates of torsion angles of each amino acid of the loop backbone in sequence-dependent way. We then combine these values with the original R-plot to reconstruct a new R-plot for each amino acid in the loop, and the initial sampling is performed according to the new R-plot. We applied this new algorithm to a test set of loops (generated from single-chain proteins in CASP 10), and found the medians/means of RMSDs can reduce about 0.12 Å/0.13 Å, 0.25 Å/0.27 Å, 0.47 Å/0.27 Åfor loop sets of length 10, 11, 12, respectively. Comparing to the original LEAP algorithm, the probability of making more accurate predictions is almost doubled when using the refined algorithm. The logic of our approach is not limited to LEAP, and can be extended to other algorithms which are also significantly dependent on initial sampling.

Keywords: loop structure prediction, initial conformation of peptide backbone, position specificity of amino acids, Ramachandran plot

PACS: 87.10.Vg, 87.15.bd, 87.15.bg **DOI:** 10.7498/aps.65.158701

* Project supported by the National Natural Science Foundation of China (Grant Nos.11105218, 11347614).

† Corresponding author. E-mail: liming@ucas.ac.cn