一种基于片段组装的蛋白质构象空间优化算法

郝小虎[1] 张贵军[1] 周晓根[1] 程正华[1] 张启鹏[2]
(浙江工业大学信息工程学院 杭州 310023)[1] (浙江工业大学经贸管理学院 杭州 310023)[2]

摘 要 针对蛋白质构象空间优化问题,提出一种基于片段组装的构象空间优化算法。算法利用基于Rosetta粗粒度的知识能量模型有效地提高了收敛速度;同时,借助片段组装技术可以有效弥补因能量函数不精确而导致的预测精度不足的缺陷;此外,差分进化算法的引入使得算法具有较好的全局搜索能力。5种测试蛋白的实验结果表明,所提算法具有较好的搜索性能和预测精度。

关键词 蛋白质结构预测,片段组装,差分进化算法,Rosetta粗粒度能量模型
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.049

Protein Conformational Space Optimization Algorithm Based on Fragment-assembly

HAO Xiao-hu[1] ZHANG Gui-jun[1] ZHOU Xiao-gen[1] CHENG Zheng-hua[1] ZHANG Qi-peng[2]
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(College of Economics and Management, Zhejiang University of Technology, Hangzhou 310023, China)[2]

Abstract An optimization algorithm based on fragment-assembly was proposed for the optimization problems of protein conformational space. The algorithm employs Rosetta energy model based on the knowledge and coarse-grained to improve the convergence rate. Simultaneously, fragment-assembly techniques are able to compensate the defect of prediction accuracies caused by the inaccuracy of energy functions. The introduction of differential evolution algorithm successfully improves the global searching capability of the algorithm as well. The experiments on five test proteins verify the superior searching performance and prediction accuracy of the proposed algorithm.

Keywords Protein structure prediction, Fragment-assembly, Differential evolution algorithm, Rosetta knowledge-based coarse-grained energy model

人类基因组序列图的成功绘制意味着人类基因组计划的所有预定目标全部实现,标志着人类基因组计划的胜利完成和后基因组时代的来临[1]。在后基因组时代,研究工作的重心从基因测序转向了基因组功能的识别:根据蛋白质分子的氨基酸序列预测其空间结构。这将使人们更系统地理解生物信息从DNA到具有生物活性蛋白质的遗传信息传递过程,使中心法则得到更为详尽的阐明,进而对生命过程中的各种现象有进一步的深刻认识,最终推动生命科学向前发展。蛋白质结构预测除了其自身的理论意义外,还具有很重要的实际应用意义。人们希望通过研究蛋白质的空间结构来了解其内在机理,这样不仅可以对疾病进行有效的预测和控制,还可以按照人们的设想设计出符合特定需求的非天然蛋白质[2]。

蛋白质结构预测是指从蛋白质的一级结构(即氨基酸序列)建立蛋白质的三维结构模型,并且进一步对蛋白质结构与功能进行研究。当前蛋白质结构预测方法大致分为两类:1)基于模板的蛋白质结构预测[3];2)不基于模板的蛋白质结构预测。其中,不基于模板的从头预测方法[4]应用最为广泛,它建立在热力学理论基础之上,即蛋白质的天然结构对应于自由能最小的结构。从头预测方法不需要氨基酸序列以外的其它更多信息,仅从一条蛋白质的氨基酸序列出发,就可得到蛋白质的空间结构,因而被称为蛋白质结构预测领域的"圣杯"。根据Anfinsen假设[5],直接从氨基酸序列出发,基于分子力场模型,采用全局优化方法,在其势能曲面上搜索蛋白质分子系统的最小能量状态,已经成为生物信息学最重要的研究课题之一。

蛋白质构象优化问题现在面临最大的挑战是对极其复杂的蛋白质能量函数曲面进行搜索。蛋白质能量模型考虑了分子体系成键作用以及范德华力、静电、氢键、疏水等非成键作用,致使其形成的能量曲面极其粗糙,构象对应局部极小解数目随序列长度的增加呈指数增长。而蛋白质构象预测算法能够找到蛋白质稳定结构的机理是,大量蛋白质亚稳定结构构成了低能量区域,所以能否找到蛋白质全局最稳定结构的关键是算法能够找到大量蛋白质亚稳定结构,即增加算法的种群多样性。因此,针对更加精确的蛋白质力场模型,选取有效的构象空间优化算法,使新的蛋白质结构预测算法更具有普遍性和高效性成为生物信息学中蛋白质结构预测的焦点问题。

到稿日期:2014-04-08 返修日期:2014-08-06 本文受国家自然科学基金(61075062,61379020),浙江省自然科学基金(LY13F030008),浙江省科技厅公益项目(2014C33088),浙江省重中之重学科开放基金(20120811),杭州市产学研合作项目(20131631E31)资助。
郝小虎(1990-),男,硕士,主要研究方向为智能优化计算、生物信息学;张贵军(1974-),男,博士,教授,主要研究方向为智能信息处理、全局优化理论及算法设计、生物信息学,E-mail:zgj@zjut.edu.cn(通信作者);周晓根(1987-),男,博士生,主要研究方向为智能优化;程正华(1986-),男,硕士,主要研究方向为现代智能优化算法、生物信息学;张启鹏(1994-),男,主要研究方向为智能优化计算。

片段组装作为一种有效的从头预测方法，在蛋白质结构预测中有着广泛的应用。Lee(2004)采用了一种全新的片段组装技术 PROFESY[5]，针对 CASP5 中的目标蛋白质 betanova 和 1fsd 进行构象空间优化，预测结果分别达到 3.0、4.0Å。Bradley (2005)将片段组装技术和能量极小化过程加入 Rosetta 服务器中[6]，针对 CASP6 中的 3 个目标蛋白，平均预测精度达到 1.5Å，其中目标蛋白 T0281 的预测精度达到 1.6Å。Lee(2011)提出的综合全局优化算法 DFA-CSA 算法[7]，基于 CSA 算法采用片段组装，针对 CASP8 中的目标蛋白进行构象空间优化，实验结果证明了该算法的有效性。Lee (2013)基于片段组装，通过精确的枚举法针对测试蛋白 1m2z 得到很好的结果[8]。

针对蛋白质构象空间优化这个高维优化难题，本文提出了一种基于片段组装的蛋白质构象空间优化算法 FDE(fragment-assembly differential evolution algorithm)。FDE 在片段组装的基础上，采用基于知识的 Rosetta 粗粒度能量模型，融入差分进化算法(DE)，利用 DE 算法较强的全局搜索能力对蛋白质构象空间进行搜索。本文用 5 种蛋白作为测试蛋白，并与 Rosetta 从头预测方法对比，对 FDE 算法性能进行测试验证。

1 Rosetta 力场模型

目前，蛋白质结构预测方法都尽可能多地利用从已知结构得到的基于知识的结构信息。首先，从 PDB 库中挑选目标蛋白具有相关性的蛋白质片段，这些蛋白质片段的使用能够降低蛋白质构象空间的维度，同时还能够降低构象搜索时造成的熵变，进而提高蛋白质结构预测的精度；其次，通过利用这些来源于大量统计所得到的基于知识的势能，能够很好地理解分子中各种能量项之间的复杂相关性怎样精细地保持平衡[9]。蛋白质构象搜索方法加入了这些参数化的基于知识的势能项，算法预测的精度能够得到很大的提高。

Rosetta 平台利用蛋白质 PDB 库中的相关蛋白质片段来模拟蛋白质的全局结构和局部片段的相互作用，从而找到决定蛋白质稳定结构的因素，使用一种基于知识的力场模型计算目标蛋白质的能量值，力场模型在保持氨基酸序列重要结构信息的前提下，只保留了氨基酸中 N、C、C_α 等原子的信息，将每个侧链等效成一个位于质心的伪原子，有效地减小了计算空间的复杂度[10]。

Rosetta 从头预测算法分 4 个过程，每个过程采用不同的能量函数配置。本文采用 Rosetta 从头预测算法的第 4 过程所采用的配置——Score3 配置，此配置考虑了 10 种不同的能量项，相对于其它 3 种配置更加详细。除了确保 Score3 配置具有 Score2 配置相同的函数意义，还必须要求它在某些层次上更为普遍化，以便应对计算结构生物学新的挑战：(1) Score3 配置应该能够代表新的化学基团；(2)Score3 配置必须能够增添新的能量项；(3)Score3 配置应该鼓励新算法的开发与融入[11]。这 3 方面的执行应以松耦合的形式来尽量减少向某一项扩展的工作量，同时添加一个新的能量项，其对应的能量函数不需要更新整体结构的化学表示。

Rosetta 的 Score3 能量模型不同于依赖于原子三维坐标的经验势能函数，是一种粗粒度蛋白质能量模型，其能量函数是 10 种能量项独立加权计算的线性和。本文采用的 Rosetta 力场模型的能量函数表示形式如下：

$$\begin{aligned} E_{protein} = {} & W_{repullsion} E_{repullsion} + W_{attraction} E_{attraction} + W_{solvation} \\ & E_{solvation} + W_{bb\text{-}sc\ hb} E_{bb\text{-}sc\ hb} + W_{bb\text{-}bb\ hb} E_{bb\text{-}bb\ hb} + W_{sc\text{-}sc\ hb} \\ & E_{sc\text{-}sc\ hb} + W_{pair} E_{pair} + W_{dunbrack} E_{dunbrack} + W_{rama} E_{rama} \\ & + W_{reference} E_{reference} \end{aligned}$$

其中各能量项的具体能量函数表达式和具体参数配置请见参考文献[12,13]。

2 片段组装

同源模建在预测蛋白质三维结构时，通过结构比对找到与已知结构同源的蛋白质，再以此同源蛋白质的结构为模板，构建出待测蛋白质的三维结构。片段组装借鉴同源模建的思想，利用已知结构的蛋白质片段预测蛋白质的三维结构。片段组装的核心思想是充分利用蛋白质库中和目标蛋白质相关的一系列蛋白质片段定义一组结构参数，再采用启发式优化算法利用这些片段通过变异等操作组装得到目标蛋白质的结构[14]。

通过片段组装，一方面可以减少算法搜索空间，提高算法收敛速度；另一方面，由于蛋白质的空间结构表现出一定的层次性和规律性，许多序列同源性较低的蛋白质也存在和目标蛋白质具有相关性的结构片段，因此，片组装技术避免了同源模建方法必须使用具有很高同源性蛋白质作为模板的缺陷，利用一切有用的先验知识构建出合理的蛋白质结构模型，可以有效地提高预测精度。

由于蛋白质构象空间的高维特性、能量模型的不精确性，使得片段组装成为从头预测蛋白质结构的重要方法。片段组装过程指将优化待测蛋白质的某个片段与从蛋白质片段库中随机选择的相应位置且具有相关性的片段进行替换，即 3 种二面角 ϕ、φ、ω 的替换。首先在优化目标蛋白质上随机选取一个氨基酸 i，即确定需要替换的片段$[i, i+L-1]$，L 为蛋白质片段长度，然后随机从片段库中选择 $L\times 3$ 个相匹配的氨基酸对应的二面角 ϕ、φ、ω 进行替换[15]。片段组装技术很好地继承了已知蛋白质的稳定片段，能够有效地降低算法的搜索空间，并且一定程度上降低了蛋白质势能函数中局部作用的敏感性，同时结合启发式优化算法的自学习性，可提高整体种群的质量，进而提高算法的整体性能。

3 算法描述

3.1 差分进化算法

1995 年 Price 和 Storn 提出了一种基于群体的启发式全局优化算法，即差分进化算法(DE)，它具有高效、鲁棒的特性，可以求解非线性不可微连续的函数，并成功地应用到了多个科学领域。1996 年，在日本举行的 ICEO 国际会议上，众多实验结果证明 DE 是一种除确定性优化算法外收敛最快的群体进化算法[16]。同时研究结果表明了 DE 类型的算法比 GA、PSO 类型的算法具有更好的全局搜索能力和局部增强能力。

差分进化算法不仅具有较强的全局搜索能力，还具有简单、通用和可并行处理等特点。但是在使用差分进化这种群体优化算法解决多模态优化问题时，由于其贪婪特性较强，算法只能收敛到全局最优解，而丢失了众多局部极值解；其次，问题模型的复杂性也造成这些算法极易陷入某个局部解。单种优化算法总存在一些不可避免的缺点，如果将两种或多种优化算法融合到一起或在一种优化算法中引入其他优化算法

的思想，则可以有效地扬长避短，既能发挥某种优化算法的优点，又能克服其缺点，从而提高算法的各项性能。

3.2 基于片段组装的构象空间优化算法

Rosetta从头预测方法，首先从已知的蛋白质结构的数据库中选择与目标蛋白质相关的蛋白质片段，然后随机组合这些蛋白质片段，形成一个粗粒度的蛋白质构象集，最后通过Monte Carlo模拟退火法将侧链的构象添加到目标蛋白的骨干链上。Rosetta从头预测方法作为一种国际领先的蛋白质结构预测方法，在历届CASP大赛取得了相当不错的成绩，其寻找最低能量形状的过程大致如下：

1. 从没有任何折叠的氨基酸序列开始；

2. 移动序列中的一部分，产生一个新的结构；

3. 通过能量模型计算新结构的能量；

4. 判断能量的变化来决定是否保留这次的移动(否则就淘汰)；

5. 迭代步骤2－4直到链中每一部分都得到足够多次数的移动。

上面的过程称为一条算法轨迹，每条轨迹的最终结果就是预测对应的一个结构。Rosetta会保存每条轨迹中找到的最低能量结构。每条轨迹都是唯一的，因为每次尝试的移动方向都是随机决定的。

本文提出的蛋白质结构预测算法FDE，针对Rosetta粗粒度力场模型，在片段组装的基础上，使用差分进化算法[17]替换Monte Carlo模拟退火方法，利用其较强的全局搜索能力对蛋白质构象空间进行优化。算法FDE的流程描述如下：

FDE算法

1. $t \leftarrow 0$
2. 初始化种群：从蛋白质片段库中随机选取片段产生popSize个种群个体P_{int}，并设置算法参数：种群大小popSize，蛋白质序列长度Length(即优化问题的维数)，算法的迭代次数T，算法的交叉因子CR，蛋白质片段的长度L。
3. 根据评分函数f计算每个种群的函数值大小，并进行排序，其中P_{max}的函数值最优。
4. while not termination condition do
   - 4.1. for 种群P_{int}中每个个体P_i do
     - 4.1.1. 设i=1，其中i∈{1,2,3,…,popSize}。
     - 4.1.2. 其中$P_{origin}=P_i$。
     - 4.1.3. 随机生成正整数rand1，rand2，rand3；其中，rand1，rand2∈{1,2,…,Length}，rand1≠rand2，rand3∈{1,2,3,…,popSize}。
     - 4.1.4. 针对个体P_j做变异操作，其中：j∈{min(rand1，rand2)，…，max(rand1，rand2)}；
       - a)令P_{origin}.phi(j)←P_{rand3}.phi(j)；
       - b)令P_{origin}.psi(j)←P_{rand3}.psi(j)；
       - c)令P_{origin}.omega(j)←P_{rand3}.omega(j)；
     - 4.2. 通过变异操作得到个体S_{new}。
     - 4.3. 根据下式执行算法交叉过程：

$$P_{new}=\begin{cases} S_{new,k} \leftarrow P_{origin,k}, & \text{if rand}(k) \leqslant CR \\ S_{new,k}, & \text{otherwise} \end{cases}$$

     其中k∈{0,0+L,0+2L,…,length}，L为蛋白质片段的长度。
     - 4.4. 对所得到的P_{new}执行选择操作，若$f(P_{new})>f(P_{max})$，则P_{new}替换P_{max}，否则保持种群不变。
   - 4.5. end for
5. 判断是否达到算法的终止条件(算法迭代执行T次)，若未达到，则$t \leftarrow t+1$，转至第一步继续循环执行算法。
6. end while

注：(1)步骤4.1.3中随机数*rand*1、*rand*2、*rand*3的选取，其中*rand*1≠*rand*2，*rand*3≠*i*(步骤4.1中的*i*值)

(2)步骤4.1.4中氨基酸*j*值大小在*rand*1和*rand*2之间。

(3)步骤4.1.4中变异操作将P_{origin}的氨基酸*j*所对应的二面角*phi*、*psi*、*omega*替换为P_{rand3}的相同位置所对应的二面角。

(4)步骤4.3中的交叉操作，若随机数$rand(K) \leqslant CR$，个体S_{new}的片段*K*替换为个体P_{origin}中对应的第*k*个片段，否则直接继承个体S_{new}的第*k*个片段。

4 实验结果分析

本文采用的片段库构建过程如下：首先，通过PISCES服务器[18]以*sequence similarity*≤30%，且*resolution*≤3.0Å，*R-factor*≤0.3为参数对现有的蛋白质数据库(vall. apr24. 2008)进行搜索，选择得到非冗余的蛋白质子集；然后将得到的蛋白质子集中的蛋白质链分解成片段长度为*L*的小片段；最后根据Rosetta片段能量函数[19]从这些小片段中挑选出一部分构成查询序列结构片段库。构建片段库过程使用的工具包括序列对比工具PSI-BLAST，二级结构预测服务器PsiPred、Jufo、SAM，Robetta片段库服务器等。30%的*sequence similarity*可保证拓扑结构的分散性；2002年Kolodny研究团队研究表明：片段长度越短，越容易得到新型的优良结构[15]；片段长度越长，则越需要一个大的蛋白质数据库来保持构象的多样性。综合考虑，本文*L*取3。

采用5种蛋白质对FDE算法的有效性进行测试，并与Rosetta从头预测方法进行比较。测试蛋白信息从蛋白质PDB库(http://www.rcsb.org/pdb/)下载，这些蛋白质的三维结构已由实验室测定方法成功测得，它们被广泛应用于生物信息学的验证实验。根据经验，设定实验参数如下：*popSize*=200，*CR*=0.3，迭代次数*T*为50000次，算法独立运行30次。由于算法中片段组装过程和种群进化过程具有较强的随机性，很有可能因为陷入局部极值解导致得到的实验结果存在较大偏差，本文取实验所得结果中RMSD值最小的目标蛋白作为该蛋白质的实验预测结构，为表明算法的可靠性，文中一并给出30次实验预测结果的平均RMSD值及偏差。如表1所列，表中给出了测试蛋白的具体信息及算法FDE、Rosetta从头预测方法针对5种测试蛋白的实验结果：表中ID为蛋白质的PDB ID号，Length为序列长度，Folding为折叠类型，Rosetta为用Rosetta测试得到的最小RMSD值，FDE^{min}是用FDE算法得到的最小RMSD值，FDE^{avg}是30次运行结果的平均RMSD值及标准差。通过与物理实验方法测得的结构进行比对可知，FDE算法用与Rosetta从头预测方法差不多的时间得到了更好的预测结果：FDE算法所得到的实验结果整体上比Rosetta从头预测方法的精度要高。其中测试蛋白质1GYZ平均使用941s可以得到精度为2.3Å的结构；2JUJ平均使用969s可以得到精度为3.4Å的结构；4ICB平均使用963s可以得到实验精度为2.8Å的结构；2L0G平均使用819s可以得到精度为1.3Å的结构；2LMZ平均使用910s可以得到精度为2.4Å的结构。测试结果表明，差分进化算法能够很好地改善片段组装技术的整体性能。

表 1　测试蛋白质具体信息及其实验结果

ID	Length	Folding	Rosetta	FDE^{min}	FDE^{avg}
1GYZ	60	α	3.0	2.3	3.2±0.5
2JUJ	56	α	4.5	3.4	4.0±0.4
4ICB	76	α	3.2	2.8	3.7±0.5
2L0G	32	α	2.4	1.3	1.5±0.1
2LMZ	42	α	3.0	2.4	3.1±0.3

N=60
RMSD(Rosetta)=3.0

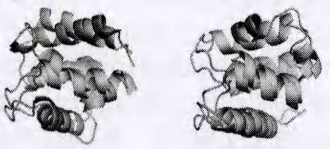
N=60
RMSD(FDE)=2.3

(a)采用 Rosetta 和 FDE 方法所得到的三维结构对比图(1GYZ)

N=56
RMSD(Rosetta)=4.5

N=56
RMSD(FDE)=3.4

(b)采用 Rosetta 和 FDE 方法所得到的三维结构对比图(2JUJ)



N=76
RMSD(Rosetta)=3.2

N=76
RMSD(FDE)=2.8

(c)采用 Rosetta 和 FDE 方法所得到的三维结构对比图(4ICB)

N=32
RMSD(Rosetta)=2.4

N=32
RMSD(FDE)=1.3

(d)采用 Rosetta 和 FDE 方法所得到的三维结构对比图(2L0G)

N=42
RMSD(Rosetta)=3.0

N=42
RMSD(FDE)=2.4

(e)采用 Rosetta 和 FDE 方法所得到的三维结构对比图(2LMZ)

图 1

图 1(a)－(e)所示的是 FDE、Rosetta 方法、物理实验方法所得到的 3 种测试蛋白质的三维结构对比图，从图中可以直观地看出 3 种实验方法所得到的结果之间的误差。

结束语　本文提出了基于片段组装的蛋白质构象空间优化算法 FDE，针对 Rosetta 粗粒度力场模型，在片段组装的基础上，利用差分进化算法较强的全局搜索能力对蛋白质构象空间进行优化。实验结果表明，FDE 算法整体具有较好的性能和预测精度，是一种有效的构象空间优化算法。

参 考 文 献

[1] Collins F, Patrinos A, Jordan E, et al. New goals for the US Human Genome Project[J]. Science, 1998-2003, 282(5389): 682-689

[2] 李娜. 人类基因组计划十年反思[J]. 科技导报, 2010, 28(13): 11

[3] 黄俊峰, 段鹏, 吴文言. 基于模板的蛋白质结构预测[J]. 生物物理学报, 2011, 27(1): 28-37

[4] Werner T, Morris M B, Dastmalchi S, et al. Structural modeling and dynamics of proteins for insights into drug interactions[J]. Advanced Drug Delivery Review, 2012, 64(4): 323-343

[5] Lee J, Joo K, Kim I, et al. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing [J]. Proteins: Structure, Function, and Bioinformatics, 2004, 56(4): 704-714

[6] Bradley P, Misura K M, Baker D. Toward high-resolution de novo structure prediction for small proteins[J]. Science, 2005, 309 (5742): 1868-1871

[7] Lee J, Sasaki T N, Sasai M, et al. De novo protein structure prediction by dynamic fragment assembly and conformational space annealing[J]. Proteins: Structure, Function, and Bioinformatics, 2011, 79(8): 2403-2417

[8] Lee J. Exact Enumeration of Protein Conformations from Fragment Assembly[J]. Journal of Physics: Conference Series, 2013, 410: 1-5

[9] Lee J, Wu S, Zhang Y. Ab initio protein structure prediction[M]// From Protein Structure to Function with Bioinformatics, 2009: 3-25

[10] Saleh S, Olson B, Shehu A. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction[J]. BMC Structural Biology, 2013, 13 (1): 1-28

[11] Rohl C A, Strauss C E, Misura K M, et al. Protein structure prediction using Rosetta[J]. Numerical Computer Methods, 2004, 383: 66-93

[12] Kortemme T, Morozov A V, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes[J]. Journal of molecular biology, 2003, 326(4): 1239-1259

[13] Handl J, Knowles J, Baker D, et al. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2012, 80(2): 490-504

[14] Handl J, Knowles J, Baker D, et al. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2012, 80(2): 490-504

[15] Kolodny R, Koehl P, Guibas L, et al. Small libraries of protein fragments model native protein structures accurately[J]. Journal of Molecular Biology, 2002, 323(2): 297-307

[16] Storn R. Differential evolution design of an IIR-filter[C]// Proceedings of IEEE International Conference on Evolutionary Computation, 1996. Nagoya, 1996, 268-173

[17] 程正华, 张贵军, 邓勇跃, 等. 一种新的蛋白质结构预测多模态优化算法[J]. 计算机科学, 2013, 40(9): 212-215, 229

[18] Wang G, Dunbrack R L. a protein sequence culling server[J]. Bioinformatics, 2003, 19(12): 1589-1591

[19] Gront D, Kulp D W, Vernon R M. Generalized Fragment Picking in Rosetta: Design, Protocols and Applications[J]. PLoS One, 2011, 6(8): e23294