蛋白质构象空间局部增强差分进化搜索方法

董 辉 郝小虎 张贵军
(浙江工业大学信息工程学院 杭州 310023)

摘 要 针对蛋白质构象空间搜索问题,提出一种蛋白质构象空间局部增强差分进化搜索方法。在差分进化算法框架下,采用Rosetta Score3粗粒度知识能量模型有效降低构象空间的搜索维数,加快算法收敛速度;引入基于知识的片段组装技术可以有效提高预测精度;利用Monte Carlo算法良好的局部搜索性能对种群做局部增强,以得到更为优良的局部构象;结合差分进化算法较强的全局搜索能力,可以对构象空间进行更为有效的采样。5个测试蛋白实验结果表明,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,差分进化算法,粗粒度能量模型,片段组装,Monte Carlo
中图法分类号 TP301.6 **文献标识码** A

Local Enhancement Differential Evolution Searching Method for Protein Conformational Space

DONG Hui HAO Xiao-hu ZHANG Gui-jun
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract A local enhancement differential evolution searching method for protein conformational space was proposed to address the searching problem of protein conformational space. On the framework of differential evolution algorithm, Rosetta Score3 coarse-grained energy model was employed for decreasing the dimension of searching space and improving the convergence rate of algorithm. The knowledge-based fragment assembly technique was introduced for improving the accuracy of prediction. For getting better local near-native conformation, local enhancement operation was done with taking advantage of the well local search performance of Monte Carlo algorithm. The well global searching capacity of the differential evolution algorithm was combined for sampling the whole conformational space effectively. The experiment results on 5 test proteins verify the superior searching performance and prediction accuracy of the proposed method.

Keywords Protein structure prediction, Differential evolution algorithm, Coarse-grained energy model, Fragment-assembly, Monte Carlo

蛋白质结构预测问题自20世纪50年代以来就一直备受关注,尤其是从头预测构象空间优化方法,是生物信息学和计算生物学领域的热点研究课题[1],因为蛋白质的三维空间结构决定了它所承载的生物功能,想要了解其功能进而对许多由蛋白质结构改变而引起的疾病进行有效的控制和预防,就必须获得其三维结构[2]。从头预测方法直接从蛋白质的氨基酸序列出发,根据Anfinsen原则[3],以计算机为工具,运用适当算法,通过计算得到蛋白质的天然构象,适用于同源性小于25%的大多数蛋白质[4]。而制约从头预测方法预测精度的瓶颈因素主要有两个方面:第一,蛋白质构象空间的高维复杂性以及能量表面的粗糙性,使其成为一个难解的NP-Hard问题;第二,力场模型的不精确性也使得预测结果难以达到较高的精度。设计有效的算法增强对构象空间的采样是解决蛋白质结构从头预测瓶颈问题的有效途径[5]。

在CASP(Critical Assessment of Techniques for Protein Structure Prediction)竞赛的推动下,国内外学者相继提出了很多有效的构象空间采样方法,如遗传算法[6-9]、分子动力学模拟[10-12]、Monte Carlo[13-15]、模拟退火[16-18]、副本交换[19,20]、构象树指导搜索方法[21-25]等。针对蛋白质高维构象空间搜索问题,基于差分进化算法,提出一种蛋白质构象空间局部增强差分进化搜索方法(Local Enhancement Differential Evolution Algorithm, LEDE):在差分进化算法框架下,采用Rosetta Score3粗粒度知识能量模型来有效降低构象空间搜索维数,提高算法的收敛速度;引入基于知识的片段组装技术可以有效提高预测精度;利用Monte Carlo算法良好的局部搜索性能对种群做局部增强,以得到更为优良的局部构象,结合差分进化算法较强的全局搜索能力,可以对构象空间进行更为有效的采样。本文采用5个测试蛋白对所提方法进行验证。

1 理论分析

蛋白质折叠问题本质上是典型的多尺度问题。基于这一特性,Rosetta应用粗粒度蛋白质表达模型构建了基于知识的能量模型,在构象优化过程中,基于粗粒度能量模型能够快速优化得到蛋白"诱饵构象",然后可以进一步基于精度更高的

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。

董 辉(1979-),男,博士,副教授,主要研究方向为智能信息处理、嵌入式系统先进控制技术;张贵军(1974-),男,博士,教授,博士生导师,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)。

全原子能量模型对“诱饵构象”进行修正，蛋白质粗粒度能量模型优化是目前从头预测的一个基本环节。但是，在粗粒度能量模型下进行系统搜索仍然面临着海量的局部极值点和高维复杂性问题，搜索过程极易陷入局部极值解。本文所提蛋白质构象空间局部增强差分进化搜索方法在进化算法的框架下，采用 Rosetta Score3 粗粒度能量模型对构象空间进行降维，有效降低空间搜索维数，提高算法收敛速度；引入片段组装技术来进一步大幅减小构象空间搜索范围，并且能够构建更为合理的蛋白质三维结构模型；利用 Monte Carlo 算法良好的局部搜索性能对种群做局部增强，以得到更为优良的局部构象，与差分进化算法较强的全局搜索能力相结合，可以对构象空间进行更为有效的采样。

1.1 Rosetta Score3 粗粒度能量模型

粗粒度蛋白质表达模型在不丢失氨基酸序列的重要结构信息的前提下，只保留了 N、C、C_α、O 这些主链的骨干原子以及侧链替代原子，用一系列设置为 $-120°$ 到 $120°$ 的二面角 $\varphi(N-C_\alpha)$、$\psi(C_\alpha-C)$ 表示氨基酸链，有效地减小了计算空间的复杂度[26]。Rosetta Score3 应用以上粗粒度蛋白质表达模型构建了下式表示的能量模型：

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

Rosetta Score3 能量模型不同于依赖于原子三维坐标的经验势能函数，它是 10 种能量项独立加权计算的线性和，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[27,28]。这是一种基于知识的能量函数，它利用蛋白质结构数据库(PDB)中已知结构数据作为学习样本，计算得到具有统计性质的区分参数，根据能量按 Boltzmann 分布的原理反推得到[29]。这种能量函数隐式地体现了形成蛋白质天然结构的内在物理化学作用，计算成本较低，而且非常有效[30]。

1.2 片段组装

片段组装通过 PISCES 服务器构建片段库，利用蛋白质库中与目标蛋白相关的蛋白质片段定义一组结构参数。从现有蛋白质 PDB 数据库中搜索得到 14071 条非冗余蛋白质子集，以 L(Rosetta 采用的长度为 3 和 9[31])为单位长度将这些蛋白质分解为短的片段，根据 Rosetta 片段计分函数[32]从这些片段中挑选出得分高的一部分组成具有代表性的位置特异的查询序列结构片段库。组装过程大致有 3 个步骤：(1)选择片段组装插入的位置 i(氨基酸序列中第 i 个残基)；(2)在片段库中通过序列比对寻找局部拟合已知蛋白质结构片段，用于替换位置 i 的残基片段；(3)从选定的片段中选择插入长度(3 或 9 片段长度)。

通过片段组装，一方面可以减小搜索空间，提高算法的收敛速度；另一方面，在一定程度上降低了蛋白质势能函数中局部作用的敏感性，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；其次，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的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[33]。

1.3 DE、Monte Carlo 算法

差分进化算法(DE)[34]是一种基于群体的启发式全局优化算法，可以求解非线性不可微连续的函数，其因高效、鲁棒的特性，成功地应用到了多个科学领域。它是除确定性优化算法外，收敛最快的群体进化算法[35]。它不仅具有较强的全局搜索能力，还具有简单、通用、可并行处理等特点。但是，由于其较强的贪婪特性，在求解多模态函数时往往会使得算法只收敛到全局最优解，从而丢失了众多局部极值解。

Monte Carlo 算法广泛应用于蛋白质构象优化。算法首先随机产生一个构象，对构象进行扰动后生成一个新的构象，通过计算两个构象的能量差值，按照设定温度下的 Boltzmann 概率接收新产生的构象，该过程一直迭代从而得到能量最低的构象。假定算法各态遍历，产生的 Markov 过程将收敛到正则分布[36]。然而，数目众多的构象势垒使得 Monte Carlo 算法往往存在某些不可达状态，极易陷入局部极小值。但是，这也体现了 Monte Carlo 算法良好的局部搜索能力。

在构象空间搜索过程中，不仅要得到能量最低的构象，在算法搜索得到最低能量构象的同时，也希望能够得到一些局部能量最低的构象，这些构象被称为亚稳态构象，它们可以用来在全原子力场模型下进一步“精修”得到全原子级别的天然态构象。Monte Carlo 算法良好的局部搜索性能使得算法能够得到局部能量最低的亚稳态构象，与具有较强全局搜索能力的差分进化算法相结合，便可以达到对构象空间更为有效的搜索采样的目的。

2 算法描述

本文所提蛋白质构象空间局部增强差分进化搜索方法以基本的 DE 算法为框架，利用 Monte Carlo 算法较强的局部搜索能力对种群进行局部增强，结合片段组装技术，对蛋白质构象空间进行有效的搜索。LEDE 算法流程描述如下：

算法 1 LEDE 算法

1. Input ← 序列信息；
2. 参数设置：种群大小 popSize，算法的迭代次数 T，交叉因子 CR，片段的长度 L；
3. 种群初始化：由输入序列产生 popSize 个种群个体 P_{init}；
4. while not end condition do：
   4.1. for 个体 P_i in 种群 P_{int}：
      4.1.1. 设 i=1，其中 i∈{1,2,3,…,popSize}；令 $P_{target}=P_i$
      4.1.2. 随机生成正整数 rand1，rand2，rand3，其中 rand1∈{1,2,3,…, popSize}，rand1 ≠ i；rand2 ≠ rand3 ∈ {1, 2, …, Length}；
      4.1.3. 针对个体 P_j 做变异操作，其中 j = rand1；令 a = min(rand2,rand3)，b=max(rand2,rand3)，k∈[a,b]；
      4.1.4. *for* k *in range*(a,b)：
         a. 令 P_{target}. phi(k) ← P_j. phi(k)；
         b. 令 P_{target}. psi(k) ← P_j. psi(k)；
         c. 令 P_{target}. omega(k) ← P_j. omega(k)；
      4.1.5. end for；
      4.1.6. 通过变异得到测试个体 P_{trail}；
      4.1.7. 生成随机数 rand4，rand5，其中 rand4∈(0,1)，rand5∈(1, Length)；
      4.1.8. 根据下式执行交叉过程：

$$P_{trail}=\begin{cases}P_{trail,rand5}\leftarrow P_{target,rand5}, & \text{if (rand4}\leqslant\text{CR)}\\ P_{trail,rand5}, & \text{otherwise}\end{cases}$$

      4.1.9. 计算 P_{target} 和 P_{trail} 的能量：$E(P_{target})$ 和 $E(P_{trail})$；
      4.1.10. 若 $E(P_{target}) > E(P_{trail})$，则用 P_{trail} 替换 P_{target}，否则保持种群不变；

4.2. end for;

4.3. 得到更新种群 P_{update};

4.4. for 个体 P_i in 种群 P_{update}:

4.4.1. 调用 Monte Carlo 方法对个体做局部增强;

4.4.2. 计算增强过程中产生的构象的能量 E(MC);

4.4.3. 若 $E(P_i) > E(MC)$,则更新种群,否则保持种群不变;

4.5. 得到局部增强后的种群 $P_{enhance}$;

4.6. end for;

5. if end condition ?Y,goto 6,N,return 4;

6. end while.

注:(1)步骤 4.1.4 中变异操作将 P_{target} 的氨基酸 k 所对应的二面角 phi、psi、omega 替换为 P_j 的相同位置所对应的二面角 phi、psi、omega。

(2)步骤 4.1.8 交叉操作,若随机数 $rand4 \leqslant CR$,个体 P_{trail} 的片段 $rand5$ 替换为个体 P_{target} 中对应的片段,否则直接继承个体 P_{trail}。

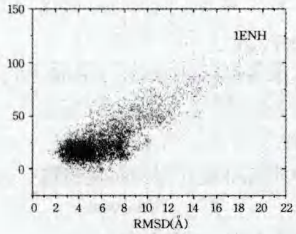
3 实验结果分析

3.1 测试环境及测试蛋白的选取

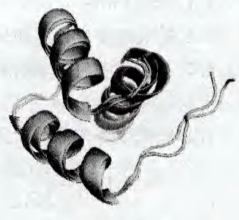
蛋白质构象空间局部增强差分进化搜索方法基于 Rosetta,在原有功能的基础上进行扩展,采用 Python 语言实现算法。运行环境为 64 位 Windows 系统,内存 16GB(实际程序运行占用内存为 2GB 左右)。迭代次数设置为 10000,温度参数设置为 1.0(在这种设置下接近于室温),片段插入长度设置为 3,交叉因子 $CR=0.5$,种群大小设置为 50,算法独立运行 20 次。本文针对折叠类型为 α 的蛋白质进行了测试,分别是 1ENH、1GYZ、2JUJ、2MU2、4ICB,它们从蛋白质 PDB 数据库(http://www.rcsb.org)下载获得,序列长度从 54 到 76。

3.2 测试结果分析

算法在各蛋白质上测试的结果如图 1(a)—图 5(a)所示,图中横坐标表示预测与实验方法测定的结构之间的相似性指标 RMSD,单位为 Å,纵坐标表示算法运行过程中产生的构象的能量得分,图中的散点表示算法搜索过程中产生的所有构象;图 1(b)—图 5(b)所示为蛋白质预测结构与实验结构的三维对比图,图中深色的为实验测定的结构,浅色为预测结构,图中重叠部分越多表明预测结果越好。测试蛋白的信息及测试结果列于表 1 中,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,文中给出了 20 次实验预测结果中去除异常值后的平均 RMSD 值及偏差。表 1 中所列 Length 为序列长度,Folding 为蛋白质折叠类型,$LEDE^{min}$ 表示用 LEDE 算法得到的最小 RMSD 值,$LEDE^{avg}$ 是 20 次运行结果的平均 RMSD 值及标准差。

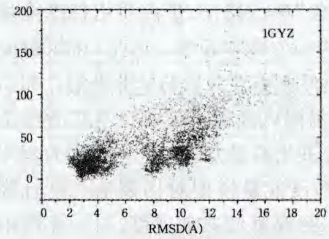


(a)测试结果图

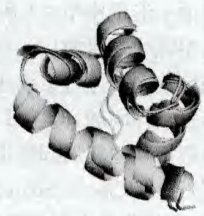


(b)与三维结构的比对

图 1 蛋白质 1ENH

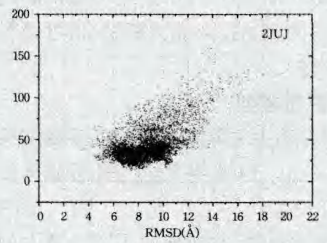


(a)测试结果图

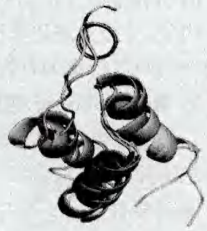


(b)与三维结构的比对

图 2 蛋白质 1GYZ

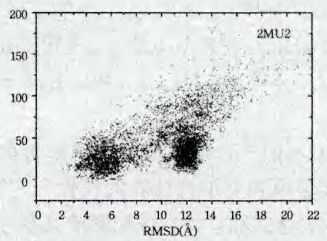


(a)测试结果图

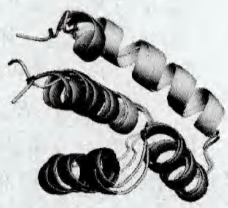


(b)与三维结构的比对

图 3 蛋白质 2JUJ

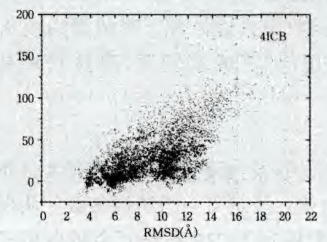


(a)测试结果图

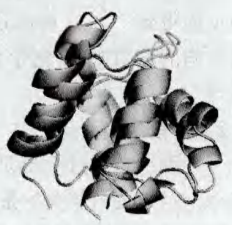


(b)与三维结构的比对

图 4 蛋白质 2MU2



(a)测试结果图



(b)与三维结构的比对

图 5 蛋白质 4ICB

表 1 测试蛋白质的具体信息及其实验结果

ID	Length	Folding	$LEDE^{min}$	$LEDE^{avg}$
1ENH	54	α	1.445	1.965±0.286
1GYZ	60	α	1.734	2.389±0.317
2JUJ	56	α	3.461	4.129±0.399
2MU2	71	α	2.167	2.893±0.373
4ICB	76	α	2.764	3.290±0.343

通过与实验方法测得的结构进行比对可知,LEDE 算法能够较有效地搜索得到蛋白质的近天然态构象。

从蛋白的测试结果图可以看到,对于 5 种测试蛋白,蛋白质构象空间局部增强差分进化搜索方法能够对构象空间进行有效的采样,找到了能量较低的近天然态构象,同时可以明显看出算法搜索到了一些其他的局部极值解,验证了所提算法思路的有效性;三维结构对比图清晰地展示了预测结构与实验测定结构的相似性;在预测精度上,所提算法达到了与目前

主流算法相同甚至更高的预测精度：Baker 研究小组基于 Monte Carlo 的 MMC 算法针对 3 个长度为 49～72 的目标蛋白，平均预测精度为 1.5Å[37]；Zhang 研究小组基于 REMC 算法开发的 Touchstone II 服务器针对 43 个长度为 36～157 的目标蛋白，平均预测精度为 6.7Å[38]；Zhang 研究小组基于 PHSMC 算法开发的 I-TASSER 服务器针对 16 个长度为 49～118 的目标蛋白，平均预测精度为 3.5Å[39]。而本文提出的构象空间搜索方法最小可以达到 1.4Å，平均预测精度达到 2.3Å，测试结果表明了所提算法的有效性。

结束语 本文提出的蛋白质构象空间局部增强差分进化搜索方法在差分进化算法的框架下，采用 Rosetta Score3 粗粒度知识能量模型，有效降低了构象空间搜索维数，提高了算法收敛速度；基于知识的片段组装技术的引入有效提高了预测精度；利用 Monte Carlo 算法良好的局部搜索性能对种群做局部增强，得到了更为优良的局部构象，与差分进化算法较强的全局搜索能力相结合，对构象空间进行了更为有效的采样，得到了粗粒度级别的全局极小点。5 个测试蛋白实验结果表明，LEDE 算法具有较好的搜索性能和预测精度，是一种有效的构象空间搜索算法。

参 考 文 献

[1] 许忠能. 生物信息学[M]. 北京：清华大学出版社，2008

[2] Dill K A, Mac Callum J L. The Protein Folding Problem, 50 Years on [J]. Science 2012,11(338):1042-1046

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

[4] Beliakov G, Lim K F. Challenges of continuous global optimization in molecular structure prediction [J]. European Journal of Operational Research,2007,181(3):1198-1213

[5] Kim D E, Blum B, Bradley P, et al. Sampling Bottlenecks in De novo Protein Structure Prediction [J]. Journal of molecular biology,2009,393(1):249-260

[6] Tantar A A, Melab N, Talbi E G, et al. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid [J]. Future Generation Computer Systems, 2007, 23(3): 398-409

[7] Hoque M T, Chetty M, Lewis A, et al. Twin removal in genetic algorithms for protein structure prediction using low-resolution model [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2011,8(1):234-245

[8] Islam M K, Chetty M. Clustered memetic algorithm with local heuristics for ab initio protein structure prediction [J]. IEEE Transactions on Evolutionary Computation, 2013, 17(4): 558-576

[9] Custódio F L, Barbosa H J C, Dardenne L E. A multiple minima genetic algorithm for protein structure prediction [J]. Applied Soft Computing,2014,15:88-99

[10] Duan Y, Kollman P A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution [J]. Science,1998,282(5389):740-744

[11] Scheraga H A, Khalili M, Liwo A. Protein folding dynamics: overview of molecular simulation techniques [J]. Annu. Rev. Phys. Chem. ,2007,58:57-83

[12] Lindorff-Larsen K, Trbovic N, Maragakis P, et al. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation[J]. Journal of the American Chemical Society, 2012,134(8):3787-3791

[13] Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding [J]. Proteins: Structure, Function, and Bioinformatics, 2002, 48(2):192-201

[14] Shen Y, Picord G, Guyon F, et al. Detecting Protein Candidate Fragments Using a Structural Alphabet Profile Comparison Approach [J]. PloS one,2013,8(11):e80493

[15] Xu D, Zhang Y. Toward optimal fragment generations for ab-initio protein structure assembly [J]. Proteins: Structure, Function, and Bioinformatics,2013,81(2):229-239

[16] Dotu I, Cebrian M, Van Hentenryck P, et al. On lattice protein structure prediction revisited [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(6): 1620-1632

[17] Tyka M D, Jung K, Baker D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers [J]. Journal of computational chemistry,2012,33(31):2483-2491

[18] Joo K, Lee J, Sim S, et al . Protein structure modeling for CASP10 by multiple layers of global optimization [J]. Proteins: Structure, Function, and Bioinformatics,2014,82(S2):188-195

[19] Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding [J]. Chemical Physics Letters, 1999, 314(1):141-151

[20] Sugita Y, Okamoto Y. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape [J]. Chemical Physics Letters,2000,329(3):261-270

[21] Shehu A. An Ab-initio tree-based exploration to enhance sampling of low-energy protein conformations [C]// Robotics: Science and Systems. 2009:241-248

[22] Shehu A, Olson B. Guiding the search for native-like protein conformations with an Ab-inito tree-based exploration [J]. Robotics Research,2010,29(8):1106-1127

[23] Olson B, Molloy K, Shehu A. In search of the protein native state with a probabilistic sampling approach [J]. Journal of Bioinformatics and Computational Biology,2011,9(3):383-398

[24] Olson B, Shehu A. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface [J]. Proteome Sci,2012,10(Suppl 1):S5

[25] Molloy K, Saleh S, Shehu A. Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab initio Protein Structure Prediction[J]. IEEE/ACM Transactions on Computational

Biology and Bioinformatics (TCBB), 2013, 10(5): 1162-1175

[26] 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

[27] Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures [J]. Proceedings of the National Academy of Sciences of the Unitized States of America, 2000, 97(19): 10383-10388

[28] 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

[29] 张贵军,郝小虎,周晓根,等. 动态步长蛋白质构象空间搜索方法[J]. 吉林大学学报(工学报)

[30] Huang E S, Samudrala R, Park B H. Scoring functions for ab initio protein structure prediction [M]// Protein Structure Prediction: Methods and Protocols. 2000

[31] Keaver-Fay A, Tyka M, Lewis S M, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules [J]. Methods in Enzymology, 2011, 487: 545-574

[32] Gront D, Kulp D W, Vernon R M, et al. Generalized fragment picking in Rosetta: design, protocols and applications [J]. PLoS One, 2011, 6(8): e23294

[33] 郝小虎,张贵军,周晓根,等. 一种基于片段组装的蛋白质构象空间优化算法[J]. 计算机科学, 2015, 42(3): 237-240

[34] Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of Global Optimization, 1997, 11(4): 341-359

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

[36] Berg B A, Neuhaus T. Multicanonical ensemble: a new approach to simulate first-order phase transitions [J]. Physical Review Letters, 1992, 68(1): 9-12

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

[38] Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction [J]. Biophysics Journal, 2003, 85(2): 1145-1164

[39] Wu S T, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations [J]. BMC Biology, 2007, 5(1): 1-10

(上接第 9 页)

其中,不同实验组使用不同的数据集,同一组使用的是相同的数据集,即 1、2 使用相同数据集,一、二使用不同数据集。为了更直观地表示结果,将上述实验数据改为柱状图表示,如图 2 所示。

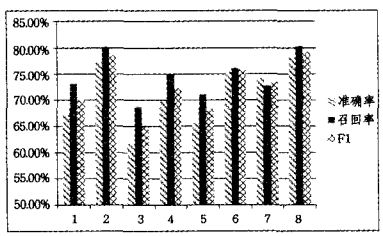


图 2 实验结果柱状图

从图 2 可以看出,在不同的数据集上,本文所设计的算法的歧义消解效果都有较为明显的提升,且效果稳定。具体来讲,基于本文改进的 LN-TF-IDF 算法的歧义消解方法的 F1 比基于传统的 TF-IDF 算法的歧义消解方法的 F1 平均提升了 7.31%。

实验结果说明,本文所设计的算法在解决食品安全领域专有名词的歧义问题中有较好的稳定性和有效性。

结束语 总体来看,本文设计的歧义消解算法的歧义消解的关键步骤在于利用 SVM 分类器消歧,主要工作在于 TF-IDF 算法的优化。该改进的算法增加考虑了特征词的上下文关系及特征词的长度在特征词权重计算中的影响,这样计算得到的特征词权重弥补了传统的 TF-IDF 算法的不足,在一定程度上提高了特征词选择的准确性,进而提高了歧义消解的有效性。这说明特征选择在歧义消解中的重要性。

参 考 文 献

[1] 龚凌晖. 中文命名实体识别与歧义消解研究[D]. 上海:复旦大学, 2011

[2] 何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010(6): 1287-1295

[3] Pedersen T. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense [C]// Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics(NAACL-01). Pittsburgh, PA, 2001

[4] Hoffart J, Yosef M A, Bordino H, et al. Robust Disambiguation of Named Entities in Text[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK, 2011: 782-792

[5] 戴祥鹰. 文本聚类在话题检测与人名消歧中的应用研究[D]. 哈尔滨:哈尔滨工业大学, 2010

[6] 韩伟. 人名消歧研究与实现[D]. 北京:北京大学, 2014

[7] 李永亮,黄曙光,鲍蕾. 一种基于 PageRank 算法和知网的词义消歧方法[J]. 计算机应用与软件, 2011, 28(4): 213-215

[8] 徐钟. 隐含马尔科夫模型在中文实体分类中的应用及研究[D]. 南昌:南昌大学, 2012

[9] Mena B H, van K M. A Hybrid Approach for Robust Multilingual Toponym Extraction and Disambiguation [C]// International Conference on Language Processing and Intelligent Information Systems. Warsaw, Poland, 2013

[10] 廖浩,李志蜀,王秋野,等. 基于词语关联的文本特征词提取方法[J]. 计算机应用, 2007, 27(12): 3009-3012

[11] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京:北京邮电大学, 2012

[12] 范昕炜. 支持向量机算法的研究及其应用[D]. 杭州:浙江大学, 2003