doi:10.3969/j.issn.1672-5565.2013.01.04

# Dehouck-Gilis-Rooman 统计势能函数的简化

陆文伟[1,2],黄日波[2,3],杜丽琴[3],韦宇拓[3]
(1.广西大学数学与信息科学学院,广西 南宁 530004;2. 广西科学院,广西 南宁 530007;
3. 广西大学生命科学与技术学院,广西 南宁 530004)

**摘 要**:统计有效能函数(Statistical effective energy function (SEEF))又称统计势能函数,来源于对已知蛋白质结构数据库进行的统计分析。近年来 Dehouck-Gilis-Rooman (DGR) 小组通过将蛋白质结构的总体统计势能分解成低阶的势能项之和,提出了一种新型的统计势能函数,这种势能函数由有限个可加和的项组成。我们试图对这种势能函数进行简化,即通过诱导蛋白质数据库选择尽可能少的势能项,而且这些势能项的加和可满足一定的性能要求。结果我们得出了六组势能项组合,其性能可和 DGR 小组得出的势能项组合相媲美,但其势能项数目却大大减少,完全可用于蛋白质结构预测。

**关键词**:诱导蛋白数据集,能量函数性能,蛋白质结构预测,统计有效能函数,富集
**中图分类号**:Q518.2 **文献标识码**:A **文章编号**:1672-5565(2013)-01-022-07

# Reduced representation of Dehouck-Gilis-Rooman potential function by selecting against decoy datasets

LU Wen-wei[1,2], HUANG Ri-bo[2,3], DU Li-qin[3], WEI Yu-tuo[3]
(1. *Science and Technology College of Mathematics and Imformation Sciences, Guangxi University, Nanning Guangxi* 530004 *China*;
2. *Guangxi Academy of Sciences, Nanning Guangxi* 530007, *China*; 3. *College of Life Science and Biotechnology, Guangxi University, Nanning Guangxi* 530004, *China*)

**Abstract**: Statistical effective energy function (SEEF) is derived from the statistical analysis of the database of known protein structures. Dehouck-Gilis-Rooman (DGR) group has recently created a new generation of SEEF in which the additivity of the energy terms was manifested by decomposing the total folding free energy into a sum of lower order terms. We tried to optimize the potential function based on their work. By using decoy datasets as screening filter, six new combinations of the energy terms were found to be comparable to DGR potential in performance test.

**Key words**: Decoy Dataset, Performance of Energy Function, Protein Structure Prediction, Statistical Effective Energy Function(SEEF), Enrichment

蛋白质结构决定功能[1],但时至今日,人们仍没办法直接从氨基酸序列预测其三维结构。多年来,人们提出了各种各样的蛋白质结构预测方法,其中就包含统计势能函数[2]。2006年,Dehouck - Gilis - Rooman 小组提出了一种新的统计势能函数[3],这种函数克服了前期统计势能函数的一些缺陷[3-5],即无法使用越来越大的蛋白质结构数据库和仅考虑单一的蛋白质结构特征如残基三联子(Residue triplet)[6]、溶剂可及性(Solvent accessibility)[7]、二面角(Dihedral angle)[8]、以及氢键(Hydrogen bonding)[9]等。这个新的统计势能函数组合了几个蛋白质序列和结构因子:氨基酸类型(Residue type,s)、主链二面角(Backbone torsion angle,t)、溶剂可及性(Solvent accessibility,a)、序列相对位置(Relative positions along the sequence)和残基间距离(Interresidue distance,$d$)。同时将蛋白质折叠总自由能分解成子势能项的总和,每一项子势能由不同的结构因子组合推导得出。这样一来,既可以单独

**收稿日期**:2012-08-25;**修回日期**:2012-09-24.
**资助项目**:国家973项目2009CB724703,国家自然科学基金31060125资助。
**作者简介**:陆文伟,广西大学数学与信息科学学院,博士。Tel: +86-0771-3270027; E-mail: lu_wenwei@163.com.

分析某个子势能,又可避免过高评价其对总能的影响。更为重要的是,可以有效地处理大小有限的数据库。Dehouck – Gilis – Rooman(DGR)统计势能函数一共包含 42 个子势能项,分为局部和距离势能两大类,可以以一定的数量组合加和成总能。

我们的目的是在这 42 个子势能项的基础上筛选某些项进行组合,在同等性能条件下争取更小的组合数目以节省计算资源。前文[10]曾经提出了一个略有缺陷的势能项选优方法,且只考察了 DGR 小组提出的 42 个势能项中的 17 项,所得的几个组合性能也不是很理想。本文是对前文的补充和完善,改进了前文提出的性能优化方法,同时提出另一种筛选方法——组合数学方法,对 DGR 小组提出的所有 42 个势能项均进行了考察,得出了 6 个性能较好的势能项组合,并对选择的组合进行了富集计算。

# 1　方法

## 1.1　序列与结构描述子

序列和结构描述子基于 DGR 小组之前的研究工作[3]:

氨基酸类型(Amino acid type,$s$),$s_i$,指序列上某个位置 i 的氨基酸类型,即 20 种天然 α 氨基酸之一。

主链构象(Backbone conformation,$t$),残基 $i$ 的主链构象,$t_i$,由其二面角描述。每一个二面角都可以按照以下七个 Ramachandran 图[11]上的区域进行归类[12]:A, C, B, P, G, E 和 O。

溶剂可及性(Solvent accessibility,$a$),残基 $i$ 的溶剂可及性,$a_i$,定义为目标蛋白质结构中残基溶剂可及面积[13]与扩展的三肽(Gly – X – Gly)溶剂可及面积[14]之比。

残基间距离(Interresidue distance,$d$),$d_{ij}$,定义为残基重心($C_\mu$)间的距离。本文忠实于 DGR 小组采用的方法,即某个蛋白质结构数据库中所有 20 种残基其重心的统计平均值[15]。

四个描述子按一定规律组合形成耦合,每一个耦合对应一个子势能项,而这些项又可以通过加和形成新的统计势能。DGR 小组所得的 42 个子势能项,可分为两大类:距离势能(共 14 个子能项)和局部势能(共 28 个子能项)[3]。

距离势能项为:Dn2_ad, Dn2_sd, Dn2_td, Dn3_ada, Dn3_asd, Dn3_atd, Dn3_sds, Dn3_tdt, Dn3_tsd, Dn4_atsc, Dn5_asdas, Dn5_atdat, Dn5_tsdts, Dn7_atscats.

局部势能项为:Ln2_aa, Ln2_as, Ln2_at, Ln2_ts, Ln2_tt, Ln3_aaa, Ln3_aas, Ln3_aat, Ln3_ass, Ln3_ats, Ln3_att, Ln3_tss, Ln3_tts, Ln3_ttt, Ln4_aaaa, Ln4_aaas, Ln4_aaat, Ln4_aass, Ln4_aats, Ln4_aatt, Ln4_asss, Ln4_atss, Ln4_atts, Ln4_attt, Ln4_tsss, Ln4_ttss, Ln4_ttts, Ln4_tttt.

为了方便,本文对势能项采用了和前文[10]不同的标记方式,以 $D$ 开头的表示距离势能,以 $L$ 开头的表示局部势能,$n$ 和紧跟的数字表示描述子数目,下划线后的小写字母表示描述子组合。

## 1.2　性能指标

共有三个性能指标,和我们的前文一致[10]:

成功率($S_i$):自由能比相应诱导蛋白都低的天然蛋白数目和所有天然蛋白数目的百分比。

Z – score 平均值($<Z>$):定义为蛋白质天然结构的能量和诱导蛋白集合平均能量之间的差与该诱导蛋白集合能量标准差的比值,用于检测统计势能鉴别天然蛋白质结构的能力。$<Z>$ 是诱导蛋白质集合所有天然蛋白质 Z – score 的平均值。

$S_{-1}$:Z – score 低于 –1 的天然蛋白的百分比,此指标用于评估势能函数的性能,尤其是当天然蛋白结构与诱导蛋白的结构很接近时。

## 1.3　富集的计算

在结构上接近天然蛋白质的诱导蛋白,在能量上也会接近天然蛋白。一个好的势能函数,应该能反映这个特点。富集(enrichment)就是衡量势能函数反映这个特点的好坏程度[16-17]。

针对每个天然蛋白质及其诱导蛋白集合,设其所有诱导蛋白的数目为 $l$,可得到两个子集:一个是在结构比对中具备低 rmsd 的诱导蛋白集合,设其诱导蛋白的数目为 $r$;另一个是在能量计算中具备低能量的诱导蛋白集合,设其诱导蛋白的数目为 $m$。两个子集在整个诱导蛋白集合中所占比例 $a$ 均取 15%,当富集的值大于 1 时,说明组合对低 rmsd 的诱导蛋白富集[16]。设在低能量诱导蛋白中具有 $n$ 个低 rmsd 的诱导蛋白,则富集的计算公式为:

$$\left(\frac{n}{m}\right)\left(\frac{m}{l}\right)^{-1}=\frac{n*l}{m^2}=\frac{n}{a^2*1}$$

式中 $r, m, l$ 分别是两个子集和总集合的诱导蛋白数目。

得到测试集合中每个天然蛋白的富集后,再对所有天然蛋白的富集进行平均,就得到整个测试集合的富集。

## 1.4　诱导蛋白质集合

用于筛选的诱导蛋白质集合有两个:

第一个,Decoys 'R' Us dataset,简记为 Ru,包含 25 个天然蛋白质[18],每一个蛋白质均有数百个诱

导蛋白与之相应,由不同的计算方法产生:1ctf, 1r69, 1sn3, 2cro, 4pti and 4rxn,1fc2 - c, 1hdd - c, 2cro,1bg8 - a, 1bl0, 1jwe,1ctf, 1dkt - a, 1fca, 1nlk, 1pgb, 1trl - a,1ctf, 1dtk, 1fc2 - c, 1igd, 1shf - a, 2cro, 2ovo。

第二个,Baker's models 2007,简记为 Baker,由 Baker's group 在 2007 年为 CASP7 开发,包括 59 个天然蛋白[19], 每个对应 100 个诱导蛋白,这些诱导蛋白通过 the Rosetta denovo structure prediction algorithm 产生,并对所有的原子进行了微调,和天然蛋白结构非常逼近:1a19, 1a32, 1a68 , 1acf, 1ail, 1aiu, 1b3a, 1bjf, 1bk2, 1bkr, 1bm8, 1bq9, 1c8c, 1c9o, 1cc8, 1cei, 1cg5, 1ctf, 1dhm, 1e6i, 1elw, 1enh, 1ew4, 1eyv, 1fkb, 1fna, 1gvp,1hz6, 1ig5, 1iib, 1kpe, 1lis, 1lou, 1nps, 1opd, 1pgx, 1ptq, 1r69, 1rnb, 1scj, 1shf, 1ten, 1tig,1tul, 1ubi, 1ugh, 1urn, 1utg, 1vcc, 1vie, 1vls, 1who, 2acy, 2chf, 2ci2, 2tif, 4ubp, 5cro, 256b。

用于测试筛选组合性能和计算富集的诱导蛋白质集合一个,简记为 Standard,来源于 Dimitri Gilis 提出的标准诱导蛋白质数据库[17],该数据库共有 45 个天然蛋白质,本文挑选了其中 10 个,其诱导蛋白均由 Rosetta2[16] 方法产生:1a32,1ail,1cei,1csp, 1dol,1hyp,1pgx,1r69,1tuc,1vcc。

### 1.5 势能项的筛选方法

#### 1.5.1 性能优化方法

初始组合的挑选以某一个或数个性能指标为标准,在这里我们选取指标 <Z>,原则是在 42 个势能项中选取使值 <Z> 更小的项。

令 $C=\{p_1,p_2,\cdots,p_{42}\}$ 表示所有的势能项集合,$p_i$ 表示势能项,其中 $i\in[1,42]$; $M=\{C_1^*,C_2^*,\cdots,C_k^*\}$ 表示初始组合集合,$C_k^*=\{p_1^*,p_2^*,\cdots,p_i^*\}$ 表示初始组合,$p_i^*$ 表示选取的势能项,其中 $k,i\in[1,42]$;$D=\{D_1,D_2,\cdots,D_n\}$ 表示用于测试的假蛋白集合,其中 $D_n$ 表示某个天然蛋白质及其对应的假蛋白结构集合,即 $D_n=\{d_i^*,d_1,d_2\cdots,d_j\}$,$d_i^*$ 表示天然蛋白结构,$d_j$ 表示假蛋白结构,$i,j,n\in N$ ;又令 z 表示计算所得的 <Z>,$z^*$ 表示当前最小的 <Z>。

从前文[10]知道,当值小于 0 时 z 才有意义,而且 $z^*$ 初始值的选择对选择结果有一定的影响。到目前为止,有关文献[3, 10]得到的组合其值很少超过 -5.0,因此我们特意为 $z^*$ 的初始值定义一组数值:[ -0.2, -0.4, -0.6, -0.8, -1.0, -1.2, -1.4, -1.6, -1.8, -2.0, -2.2, -2.4, -2.6, -2.8, -3.0, -3.2, -3.4, -3.6, -3.8, -4.0, -4.2, -4.4, -4.6, -4.8, -5.0],由此得到选择算法如下:

```
C = {p1,p2,…,p42}
Ck* = {}
M = {}
z* = input (initial_value)
While C ! = {}
    p = random(C)
    Ck* = Ck* + p
    C = C - p
    Ctemp1 = C
    Foreach pi in Ctemp1
      Ck* = Ck* + pi
      z = performance(Ck*)
      If z < z*
        C = C - pi
        z* = z
        Ctemp1 = Ctemp1 - pi
      Else
        Ck* = Ck* - pi
      End if
    End foreach
    If length(Ck*) < = 1
      Ck* = {}
    Else
      M = M + Ck*
      Ck* = {}
      Output z*
    End if
End while
```

伪代码中的函数 input, random, performance 以及 length 分别表示输入数值,随机获取势能项,计算组合性能以及求取集合元素的个数,集合可以通过 + 和 - 操作来添加和删除势能项,通过 {} 操作来清空自身或赋空值。

#### 1.5.2 组合数学方法

给定一个非空集合 $s$,共有 $r$ 个元素,从集合中选取 $n$ 个子元素形成一个无序排列的结果可用组合数学的组合算法[20]来取得。从 DGR 小组所得的 42 个势能项中筛选势能项可变成计算这些势能项的各种组合。通过组合算法得到的势能项组合,同样在诱导蛋白集合 Ru 和 Baker 上计算性能,并从中挑选表现好的组合。

令 $s=\{p_1,p_2,\cdots,p_n\}$ $n\in[1,42]$,式中是势能项。理论上,$n$ 可以取 0 到 42 之间的值,从而得到 42 种组合。当 $n>5$ 和 $n<38$ 时,组合数会非常大,难以计算,所以本文仅考察 $n<5$ 和 $n>38$ 时的组

合,$n<5$ 的组合作为候选组合,$n>38$ 时的组合不符合本文简化的目的,仅作参考和比对之用。

## 2　结果

### 2.1　性能优化结果

#### 2.1.1　基于 Ru 导出的组合

当 $z^*$ 给定的初始值大于 -2.0 时,得到的组合是一样的:Dn2_ad, Dn2_sd, Dn3_sds, Dn4_atsc, Dn5_asdas, Ln2_as, Ln2_at, Ln2_ts, Ln3_aaa, Ln3_aas, Ln3_aat, Ln3_ass, Ln3_ats, Ln3_att, Ln3_tss, Ln3_tts, Ln4_aaaa, Ln4_aaas, Ln4_aaat, Ln4_aass, Ln4_aats, Ln4_asss, Ln4_atss, Ln4_atts, Ln4_tsss, Ln4_ttss. 这样一个组合子势能项的数目太大,其性能指标值分别为:$<Z>$ = -4.5041, $S_1$ = 80, $S_{-1}$ = 100。

当 $z^*$ 给定的初始值在[-3.6, -2.0]之间时,出现不同的组合,其中最好的几个是:

Ln2_at, Ln3_aas, Ln3_aat, Ln3_ass, Ln3_ats, Ln3_tss, Ln3_tts, Ln4_aaas, Ln4_aass, Ln4_aats, Ln4_asss, Ln4_atss, Ln4_atts, Ln4_tsss, Ln4_ttss. 组合子势能项数目偏大,性能指标值为:$<Z>$ = -4.934 8, $S_1$ = 68, $S_{-1}$ = 100。

Ln2_at, Ln3_ats, Ln3_tss, Ln3_tts, Ln4_asss, Ln4_atss, Ln4_tsss, Ln4_ttss. 组合子势能项数目适当减少了,性能指标值为:$<Z>$ = -4.6067, $S_1$ = 60, $S_{-1}$ = 96。

Ln3_aas, Ln3_tss, Ln3_tts, Ln4_aass, Ln4_asss, Ln4_atss, Ln4_tsss, Ln4_ttts. 性能指标值为:$<Z>$ = -4.4578, $S_1$ = 64, $S_{-1}$ = 96。结果和第二组相当。

当 $z^*$ 的值 $z^* < -3.6$ 时,没有可用的组合。

#### 2.1.2　基于 Baker 导出的组合

当 $z^*$ 给定的初始值大于 -2.0 时,得到的组合同样是一样的:Dn2_ad, Dn2_sd, Dn2_td, Dn3_sds, Dn3_tdt, Dn3_tsd, Dn5_atdat, Dn5_tsdts, Ln2_at, Ln3_aaa, Ln3_aas, Ln3_ass, Ln3_ats, Ln3_tss, Ln3_tts, Ln4_aaaa, Ln4_aaas, Ln4_aaat, Ln4_aass, Ln4_aats, Ln4_aatt, Ln4_asss, Ln4_atss, Ln4_atts, Ln4_tsss, Ln4_ttss, Ln4_ttts. 和基于 Ru 导出的组合相似,子势能项数目巨大,性能一般:$<Z>$ = -4.112 0, $S_{-1}$ = 76, $S_{-1}$ = 96。

当 $z^*$ 给定的初始值在[-4.4, -2.0]之间时,出现不同的组合,其中最好的几个是:

*Ln3_ass*, *Ln4_asss*, *Ln4_atss*, *Ln4_tsss*, *Ln4_ttss*. 组合大小数目非常理想,性能也很好:$<Z>$ = -5.4821, $S_1$ = 71, $S_{-1}$ = 93。

Ln3_ass, Ln4_atss, Ln4_tsss, Ln4_ttss. 组合大小仅为 4,性能值为:$<Z>$ = -5.1406, $S_1$ = 72, $S_{-1}$ = 94。

Ln4_asss, Ln4_tsss, Ln4_ttss. 组合大小仅为 3,性能值为:$<Z>$ = -4.951 4, $S_1$ = 69, $S_{-1}$ = 93。

当 $z^*$ 的初始值 $z^* < -4.4$ 时,没有可用的组合。

相比较而言,从 Baker 推导出的组合性能较高,且组合包含的子势能项数目更少。

### 2.2　组合数学结果

先看 $n<6$ 时的情况,表 1 是 $n$ = 2, 3, 4, 5 时在诱导蛋白集合 Ru 和 Baker 上 $<Z>$ 指标最好的组合。在 *Ru* 上,所有组合的 $<Z>$ 值最好的是当 $n$ = 5 时的 -4.766 9,稍弱于性能优化方法得出的最好组合的 $<Z>$ 值 -4.934 83,但性能优化方法得出的最优组合子势能项数目却大得多。至于 Baker, $n$ = 3, 4, 5 时最好的组合和性能优化方法所得的三个最好的组合一致:

Ln4_asss, Ln4_tsss, Ln4_ttss;

Ln3_ass, Ln4_atss, Ln4_tsss, Ln4_ttss;

Ln3_ass, Ln4_asss, Ln4_atss, Ln4_tsss, Ln4_ttss.

这说明我们的性能选优方法是有效的。

**表 1　组合数学算法得出的结果**

Table 1　Results of combinatorial algorithm

| Combinations | Baker | | | Ru | | |
|---|---|---|---|---|---|---|
| | $<Z>$ | $S_1$ | $S_{-1}$ | $<Z>$ | $S_1$ | $S_{-1}$ |
| Ln4_asss_Ln4_tsss | -4.415 6 | 66 | 80 | -1.622 2 | 32 | 56 |
| Ln3_aas_Ln3_tss | -2.596 2 | 54 | 92 | -3.747 1 | 60 | 100 |
| Ln4_asss_Ln4_tsss_Ln4_ttss | -4.951 4 | 69 | 93 | -1.411 1 | 20 | 64 |
| Ln3_aas_Ln3_tss_Ln3_tts | -2.105 7 | 42 | 88 | -4.105 9 | 56 | 100 |
| Ln3_ass_Ln4_asss_Ln4_tsss_Ln4_ttss | -5.201 3 | 73 | 95 | -2.651 2 | 40 | 88 |
| Ln2_ts_Ln3_aas_Ln3_tss_Ln3_tts | -2.165 5 | 44 | 90 | -4.549 7 | 60 | 100 |
| Ln3_ass_Ln4_asss_Ln4_atss_Ln4_tsss_Ln4_ttss | -5.482 1 | 71 | 93 | -2.404 1 | 40 | 80 |
| Ln2_ts_Ln3_aas_Ln3_tss_Ln3_tts_Ln4_asss | -2.725 6 | 56 | 92 | -4.696 0 | 64 | 100 |

当考虑 $n>38$ 时,值没有小于 -4.0 的,但是组合间的性能差异很小,从标准差上就可以看出来:$n=38,39,40,41$ 时的标准差分别为 0.20,0.17,0.14,0.10。$n$ 越大,标准差越小,组合间性能越平均。这说明,组合的子势能项数目达到一定数目后,其性能表现比较平衡,不会因组合的差异和诱导蛋白集合的不同而出现很大的变化,同时性能比较一般,不会出现性能很好和很差的组合。

另外,当 $n=5$ 时,组合数超过了 80 万。在这 80 多万个组合中,在诱导蛋白集合 Ru 和 Baker 上值同时小于 -4.0 的共有 18 个,表 2 列出了其中头 9 个。这 18 个组合几乎全来自局部势能项,且以 Ln3 和 Ln4 占绝大多数,它们同时在两个筛选集合上的性能表现都不错。如果需要通用性较强的组合,这 18 个组合都是很好的选择。

**表 2 n=5 时,对两个诱导蛋白集合的 <Z> 均小于 -4.0 的组合**

Table 1 The combinations whose <Z> are all less than -4.0 on RU and Baker decoy sets, when n=5

| Combinations | Baker | | | Ru | | |
|---|---|---|---|---|---|---|
| | <Z> | S1 | $S_{-1}$ | <Z> | S1 | $S_{-1}$ |
| Ln3_aas_Ln3_tss_Ln4_asss_Ln4_atss_Ln4_ttss | -4.674 7 | 68 | 95 | -4.051 6 | 60 | 100 |
| Ln3_aas_Ln3_tss_Ln4_atss_Ln4_tsss_Ln4_ttss | -4.553 9 | 69 | 95 | -4.017 2 | 60 | 100 |
| Ln3_aas_Ln3_tss_Ln4_asss_Ln4_atss_Ln4_tsss | -4.487 8 | 68 | 95 | -4.123 3 | 56 | 100 |
| Ln2_at_Ln3_ass_Ln3_tss_Ln4_asss_Ln4_ttss | -4.478 3 | 66 | 95 | -4.031 1 | 52 | 100 |
| Ln3_aas_Ln3_tss_Ln4_asss_Ln4_tsss_Ln4_ttss | -4.464 7 | 71 | 95 | -4.190 0 | 52 | 100 |
| Ln2_at_Ln3_ass_Ln3_tss_Ln4_tsss_Ln4_ttss | -4.342 8 | 69 | 95 | -4.029 6 | 64 | 100 |
| Ln3_aas_Ln3_ass_Ln3_tss_Ln4_asss_Ln4_atss | -4.322 3 | 73 | 95 | -4.043 1 | 60 | 100 |
| Ln3_aas_Ln3_tss_Ln4_aass_Ln4_asss_Ln4_ttss | -4.289 9 | 66 | 95 | -4.027 8 | 52 | 96 |
| Ln3_aas_Ln3_ass_Ln3_tss_Ln4_atss_Ln4_tsss | -4.280 3 | 71 | 97 | -4.048 1 | 56 | 100 |

## 2.3 新组合

我们分别在两个诱导蛋白集合 Ru 和 Baker 导出的组合中,挑选性能靠前(new_c1 ~ new_c4),并且在两个集合中性能都不太差的组合(new_c5 和 new_c6),共六个,分别来自组合算法中 n=4,5 类型的组合:

Ln3_ass,Ln4_asss,Ln4_tsss,Ln4_ttss,简记为 new_c1;Ln2_ts,Ln3_aas,Ln3_tss,Ln3_tts,简记为 new_c2;Ln3_ass,Ln4_asss,Ln4_atss,Ln4_tsss,Ln4_ttss,简记为 new_c3;Ln2_ts,Ln3_aas,Ln3_tss,Ln3_tts,Ln4_asss,简记为 new_c4;Ln3_aas,Ln3_tss,Ln4_asss,Ln4_atss,Ln4_ttss,简记为 new_c5;Ln3_aas,Ln3_tss,Ln4_atss,Ln4_tsss,Ln4_ttss,简记为 new_c6。

这些新组合和 DGR 小组提出的性能优异的 local、dist 组合[3]的性能比较见表 3。可以看出,在 Ru 集合上,组合 new_c2、new_c4 的性能比其他组合要好,和 DGR 小组提出的组合相当。

新得到的六个组合在 standard 集合上的性能表现见表 4。

**表 3 新组合的性能比较**

Table 3 The performance of the new combinations

| Decoy set | Combinations | <Z> | S1 | S-1 |
|---|---|---|---|---|
| Ru | Local | -4.16 | 76 | 92 |
| | Dist | -4.65 | 80 | 88 |
| | new_c1 | -2.651 2 | 40 | 88 |
| | new_c2 | -4.549 7 | 60 | 100 |
| | new_c3 | -2.404 1 | 40 | 80 |
| | new_c4 | -4.696 0 | 64 | 100 |
| | new_c5 | -4.051 6 | 60 | 100 |
| | new_c6 | -4.017 2 | 60 | 100 |

由表 4 可以看出,new_c2 和 new_c4 性能较差,这两个组合都包含有势能项 Ln2_ts;new_c5 和 new_c6 性能相近,和在集合 Baker 和 Ru 上的性能差别不大,体现了它们在不同性能测试集合上的均衡性;new_c3 仍保留其在集合 Baker 上的良好性能,但在集合 Ru 上的性能却一般。

至于富集,所有六个组合的富集值均大于 1,说明这些组合均对低 rmsd 的诱导蛋白富集。

表 4　新组合在 standard 集合上的性能和富集

Table 4　The performance and enrichment of the new combination on the standard decoyset

| New combinations | <Z> | S1 | $S_{-1}$ | Enrichment |
|---|---|---|---|---|
| new_c1 | -4.708 8 | 60 | 90 | 1.480 2 |
| new_c2 | -1.537 4 | 0 | 80 | 1.572 4 |
| new_c3 | -5.198 2 | 60 | 90 | 1.420 6 |
| new_c4 | -1.897 8 | 20 | 80 | 1.598 9 |
| new_c5 | -4.448 8 | 50 | 100 | 1.611 6 |
| new_c6 | -4.275 8 | 60 | 100 | 1.592 4 |

## 3　结论

本文采用了两种方法对势能项进行筛选:性能优化方法和组合数学方法,目的在于通过筛选集合选择势能项数目少,但又具备一定性能的 DGR 统计势能函数函数。通过性能优化方法,在集合 ru 上得到的组合,$<Z>$ 最小的是 -4.934 8,但是相应组合的势能项数目却达到 15 个。在组合数学方法中,$<Z>$ 最小的是 -4.766 9,略低于性能优化方法得出的最佳 $<Z>$,但组合的势能项数目仅为 5。在集合 baker 上,性能优化方法得出的最好组合,和组合数学得出的基本一致。这说明我们的性能优化方法可以保证得到性能最优的组合,但不一定保证得到的组合的势能项数目也是合适的,也就是说,相对而言,性能优化方法可以得到最优性能,组合数学方法却可以得到适当的组合势能项数目。

对于富集,六个组合的值都在 1 以上,都对低 rmsd 的诱导蛋白富集。尽管它们在不同的集合上性能不一,但富集值却相差不大。有意思的是,组合 new_c2 和 new_c4 在 standard 集合上的性能并不好,但其富集值却比组合 new_c1 和 new_c3 稍大。而性能在不同集合上表现稳定的组合 new_c5 和 new_c6,其富集值位居前列。

与其他统计势能函数相比较,DGR 统计势能函数除了性能优异外[21],还具备可加性。这使得科学家们可以针对某种特定集合挑选适合的势能项,以组成性能较好,大小适当的组合,new_c1 ~ new_c4 就属于此类。它们在筛选集合上具有很好的性能,但却不保证在其它类型的集合上也表现良好。反之,科学家们也可以使用普适性较强的组合,比如新组合 new_c5 和 new_c6。这些组合对特定的集合不一定具备很好的性能,但却具有一定的通用性,即在不同的测试集合间,其性能之间的差别并不显著。

**参考文献**(References)


[1] Anfinsen, C. B., Principles that govern the folding of protein chains[J]. Science, 1973,181(96):223 -30.

[2] Lazaridis, T. and M. Karplus, Effective energy functions for protein structure prediction[J]. Current Opinion in Structural Biology, 2000,10(2):139 -145.

[3] Dehouck, Y., D. Gilis, and M. Rooman, A new generation of statistical potentials for proteins[J]. Biophys J, 2006,90(11):4010 -7.

[4] Boas, F. E. and P. B. Harbury, Design of Protein - Ligand Binding Based on the Molecular - Mechanics Energy Model[J]. Journal of Molecular Biology, 2008,380(2):415 -424.

[5] Emiko Furuichi, P. K., Influence of protein structure databases on the predictive power of statistical pair potentials[J]. Proteins: Structure, Function, and Genetics, 1998,31(2):139 -149.

[6] Goldstein, R. A., Z. A. Luthey - Schulten, and P. G. Wolynes, Protein tertiary structure recognition using optimized Hamiltonians with local interactions[J]. Proceedings of the National Academy of Sciences of the United States of America, 1992,89(19):9029 -9033.

[7] Jones, D. T., W. R. Taylort, and J. M. Thornton, A new approach to protein fold recognition[J]. Nature, 1992,358(6381):86 -89.

[8] Dewitte, R. S. and E. I. Shakhnovich, Pseudodihedrals: Simplified protein backbone representation with knowledge - based energy[J]. Protein Science, 1994,3(9):1570 -1581.

[9] Sippl, M. J., Helmholtz free energy of peptide hydrogen bonds in proteins[J]. J Mol Biol, 1996,260(5):644 -8.

[10] Lu Wen - Wei, Huang Ri - Bo, Wei Yu - Tuo, Meng Jian - Zong, Du Li - Qin, Du Qi - Shi, , Statistical energy potential: reduced representation of Dehouck - Gilis - Rooman function by selecting against decoy datasets[J]. Amino Acids, 2012,42(6):2353 -2361.

[11] Ramachandran, G. N. and V. Sasisekharan, Conformation of Polypeptides and Proteins, in Advances in Protein Chemistry[M]. J. M. L. A. J. T. E. C. B. Anfinsen and M. R. Frederic, Editors. Academic Press, 1968,283 -437.

[12] Rooman, M. J., J. -P. A. Kocher, and S. J. Wodak, Prediction of protein backbone conformation based on seven structure assignments: Influence of local interactions[J]. Journal of Molecular Biology, 1991,221(3):961 -979.

[13] Lee, B. and F. M. Richards, The interpretation of protein structures: estimation of static accessibility[J]. J Mol Biol, 1971, 55(3):379 -400.

[14] Rose, GD, Geselowitz, AR, Lesser, GJ, Lee, RH, Zehfus, MH, Hydrophobicity of amino acid residues in globular proteins[J]. Science, 1985,229(4716):834 -838.

[15] Kocher, J. -P. A., M. J. Rooman, and S. J. Wodak, Factors Influencing the Ability of Knowledge - based Potentials to Identify Native Sequence - Structure Matches[J]. Journal of Molecular Biology, 1994,235(5):1598 -1613.

[16] Jerry Tsai, R. B., Alexandre V. Morozov, Brian Kuhlman, Carol A. Rohl, David Baker, An improved protein decoy set for testing energy functions for protein structure prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2003,53(1):76 -87.

[17] Gilis, D., Protein decoy sets for evaluating energy functions[J].

J Biomol Struct Dyn, 2004,21(6):725 - 36.

[18] Samudrala, R. and M. Levitt, Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction [In Process Citation][J]. Protein Sci, 2000,9(7):1399 - 1401.

[19] Das, R. Qian, B. Raman, S. Vernon, R. Thompson, J. Bradley, P. Khare, S. Tyka, M. D. Bhat, D. Chivian, D. Kim, D. E. Sheffler, W. H. Malmstrom, L. Wollacott, A. M. Wang, C. Andre, I. Baker, D., Structure prediction for CASP7 targets using extensive all - atom refinement with Rosetta@ home[J]. Proteins, 2007,69 Suppl 8:118 - 28.

[20] Ruskey, F. and A. Williams, The coolest way to generate combinations[J]. Discrete Mathematics, 2009,309(17):5305 - 5320.

[21] Dehouck, Yves, Grosfils, Aline, Folch, Benjamin, Gilis, Dimitri, Bogaerts, Philippe, Rooman, Marianne, Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC - 2.0[J]. Bioinformatics, 2009,25(19):2537 - 2543.