一种GPCR跨膜螺旋形变的建模方法

陈 荣[1] 吕 强[1,2] 吴宏杰[1,3] 陈沙沙[1]

(苏州大学计算机科学与技术学院 苏州 215006)[1]

(江苏省计算机信息处理技术重点实验室 苏州 215006)[2]

(苏州科技学院电子与信息工程学院 苏州 215011)[3]

摘 要 跨膜螺旋是GPCR的最主要特征,单个螺旋的预测精度直接影响GPCR整体三维结构的预测。GPCR螺旋形变预测是一个挑战性的难题。该形变用发生形变的残基位置和该位置前后两端螺旋的夹角表示。基于目前已知的所有GPCR的跨膜螺旋结构,根据螺旋序列相似度进行聚类,然后在每类中对形变角度用连续型von Mises概率分布来建模。对建模后GPCR跨膜螺旋的形变角度进行了回归和预测测试。基于本文方法的模型,只需进行15次采样,就会有一次的采样结果近似符合天然螺旋的形变角度,这在很大程度上能够帮助跨膜螺旋空间结构的预测。

关键词 GPCR跨膜螺旋,序列相似性,形变建模,连续概率分布

中图法分类号 TP391.4 **文献标识码** A

Method for Modeling the Distortions of Transmembrane Helix in G-protein Couple Receptor

CHEN Rong[1] LV Qiang[1,2] WU Hong-jie[1,3] CHEN Sha-sha[1]

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)[1]

(Jiangsu Provincial Key Lab for Information Processing Technologies, Suzhou 215006, China)[2]

(College of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215011, China)[3]

Abstract Transmembrane helix is main feature of the GPCR, and the accuracy of single helix prediction directly affects the prediction is of entire GPCR structure. It is a challenge problem to predict the distortion of the GPCR helix. The distortion is represented by using the kink residue place and the angle of two fragment helix around the distortion position. Based on the all known GPCR's helixes structure currently, the helixes are clustered according to the helix sequence similarity, and the bend angles in each cluster are modeled by using continuous von mises probability distribution. The modeled GPCR TM kink angles are tested by using the regression and forecast testing method. Based on this article's model, only fifteen times sampling would have a sample result close to the native TM distortion angle, which will help to improve the TM-helix structure prediction.

Keywords GPCR transmembrane helix, Sequence similarity, Distortion modeling, Continuous probability distribution

1 引言

由于采用生化手段测定蛋白质结构代价高、耗时长,因此采用计算的手段测定蛋白质三维结构成为计算生物学中的重要课题之一[1]。GPCR是一种非常常见且具有重要功能的蛋白质,故对GPCR建模研究对蛋白质的空间结构的解析也是有重大意义的[2]。跨膜螺旋结构是GPCR的一个非常显著的特征,故对跨膜螺旋的空间结构的预测的准确度会直接影响GPCR整体结构的预测,所以预测和微调跨膜螺旋空间结构对了解GPCR的功能也是至关重要的。而由于GPCR跨膜螺旋结构可能发生微小形变[3],因此要很好地解决螺旋结构的预测面临巨大挑战[4]。

现有的形变角度建模过程中,主要依据形变的几何特性来建模[5,6],如文献[7]中先使用形变位置前后的残基的Cα之间的距离来过滤候选螺旋,然后使用形变角度的均值和方差来控制角度范围[8]。但是,由于蛋白质的序列信息决定了它的三维结构,跨膜螺旋的形变角度从根本上依赖于序列信息,几何特性是序列信息导致的一种具体表现,因此在建模过程中除了考虑几何特性外,还应充分考虑跨膜蛋白质的序列特征[9]。

目前跨膜螺旋结构的形变的角度大部分都是采用Monte Carlo方式[10]进行采样,再对产生的螺旋候选集按照能量的高低进行筛选[11],这种基于离散模型的采样方法可以部分地模拟形变角度的变化情况,但对螺旋形变角度建模的精度还不够。实际上,跨膜螺旋的形变角度是在(0,180°)范围内连续变化,所以连续模型更符合跨膜螺旋的实际变化情况。此外,离散模型对能量函数的依赖程度很高,导致采样空间较大,预测代价也就越大。

到稿日期:2011-12-01 返修日期:2012-03-16 本文受国家自然科学基金(60970055,61170125)资助。

陈 荣(1989—),男,硕士生,主要研究方向为生物信息计算,E-mail:20114227017@suda.edu.cn;吕 强 男,教授,博士生导师,主要研究方向为生物信息计算、元启发搜索、并行计算。

另外，文献[12]根据 Pro 残基及其变异残基来确定的形变准确率能够达到 60%左右。本文采用人工专家标定的形变位置，将专家知识融入训练集中，为模型训练提供了高质量的数据基础。

本文提出了一种根据序列相似性进行聚类的混合一元 von Mises 模型[13]，其对跨膜螺旋的形变角度进行建模与采样。von Mises 概率分布模型是一种连续型弧度模型，本文中将多个不同模型参数的一元 von Mises 模型混合使用。在训练集方面，本文螺旋形变位置不同于计算预测[14]，而是采用更具准确性的专家人工标注方法。实验结果显示，本方法预测的形变角度精度更高。

2 形变建模的设计

本文的模型训练集来自 2011 年以前的所有已知天然 GPCR 的跨膜螺旋，从 uniprot 官方网站[15]获取，共 101 个 PDB，含有 707 根天然跨膜螺旋，见表 1。

表 1 训练集所有 pdb 的 ID 号，其中以 a 结尾的表示 a 链，b 结尾的表示 b 链

PDB CODE										
1boja	1boka	1gzma	1gzmb	1hzxa	1hzxb	1jfpa	1kada	1kp1a	1kpna	1kpwa	1kpxa
1i9ha	1i9hb	1in6a	1n3ma	1n3mb	1n3mc	1n3md	1n3me	1n3mf	1nd8a	1ne0a	1oeaa
1opna	1opwa	1ov0a	1ov1a	1oz5a	1ozca	1pb2a	1t78a	1ul9a	1ul9b	1upea	1vz1a
1y5da	1y9ca	1y36a	1z8ea	1ztia	1zv0b	2a0da	2ac6a	2ah3a	2amka	2auia	2b0xa
2b6qa	2b6ra	2b6sa	2b6ua	2b6va	2cdwa	2f75a	2ff9a	2g1xa	2g2aa	2g87a	2g87b
2gfza	2he6a	2hpya	2hpyb	2i35a	2i36a	2i36b	2i36c	2i37a	2i37b	2i37c	2iila
2ik3a	2ik5a	2iqka	2iqma	2iqna	2iqpa	2iqra	2iqsa	2iqua	2iqva	2iqwa	2peda
2pedb	2r4ra	2r4sa	2rh1a	2vt4a	2vt4b	2vt4c	2vt4d	2z73a	2z73b	2ziya	3c9la
3c9ma	3capa	3capb	3d4sa	3dqba							

2.1 形变拓扑表示与计算

每根螺旋的拓扑结构可以表示为两根理想螺旋片段在形变位置前后的拼接，如图 1 左图所示。

首先，对这些单螺旋结合专家的先验的领域知识，人工标注出每个螺旋的形变位置、形变程度等信息。然后，对专家统计出的每根螺旋形变位置数据计算单螺旋的形变角度信息，计算方法可以近似为一个横投影面上的两条直线的夹角，即形变前后两螺旋片段的方向向量的夹角，计算过程如图 1 右图所示。

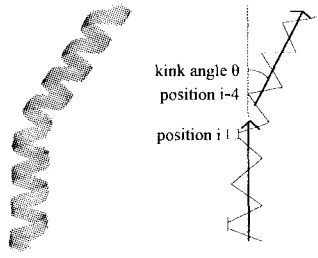


图 1 跨膜螺旋的空间形变角度计算

形变角度计算步骤：1. 取形变位置的前后各两个残基，总共 4 个残基的 Cα 原子空间坐标，计算均值坐标，并将其作为形变位置的空间坐标。2. 取形变位置到 N-terminal 的所有残基的 Cα 原子空间坐标，取其均值坐标作为前片段螺旋的质心空间坐标。3. 取形变位置到 C-terminal 的所有残基的 Cα 原子空间坐标，取其均值坐标作为前片段螺旋的质心空间坐标。4. 将以上取得的 3 点的坐标依次连成两条直线，这两条直线的方向向量的夹角即为形变的角度。

2.2 依据序列相似性聚类

首先，获取每根螺旋的序列信息，根据所有螺旋的序列信息，使用 Blast 来计算螺旋序列两两之间的相似性，相似度采用 0 到 100 之间的数值表示，这样可得到相似性矩阵。

然后，根据相似性矩阵，使用 Repeated Bisection 聚类算法来进行聚类[16]，聚类的时间复杂度为 O($n\log n$)。聚类方法的相似性函数是 cos，标准函数是 I_2，迭代次数为 10，每次迭代选取数据较多的类进行分割。Repeated Bisection 聚类方法的特点是速度快，且聚类的效果均匀，不会出现有些类中的数据很少的情况，这对于本文数据集是均匀分布的情况是非常适合的。同时我们也尝试了 Spectral clustering 方法[18]与 K-means 方法，对于本文的数据集，它们的聚类结果均略差于 Repeated Bisection 聚类方法。

最后，在开放性测试中使用了 35 根已知天然结构的 GPCR 跨膜螺旋序列，见表 2。根据测试螺旋序列与所有类的距离确定每根螺旋所属的类别，这不同于文献[17]中的聚类中心的选择。测试螺旋与某个类之间的距离的计算方法为 $d_i=\sum_{i=1}^{n}S_i/n$，其中 S_i 表示测试的螺旋与该类中的第 i 个螺旋的相似性，n 是这个类的螺旋数目，取与该螺旋相似性最大的类别为该螺旋所属的类别。

表 2 开放测试集的 PDB 的相关信息，列 CODE：PDB 编号，TM_N：PDB 中第几螺旋，K_R：形变残基，K_P：形变位置

CODE	TM_N	K_R	K_P	CODE	TM_N	K_R	K_P	CODE	TM_N	K_R	K_P	CODE	TM_N	K_R	K_P	CODE	TM_N	K_R	K_P
2RH1	1	I	47	3PBL	1	G	46	3ODU	1	L	50	3EML	1	N	24	3RZE	1	V	42
2RH1	2	P	88	3PBL	2	P	84	3ODU	2	P	92	3EML	2	P	61	3RZE	2	I	75
2RH1	3	A	119	3PBL	3	I	118	3ODU	3	Y	121	3EML	3	Q	89	3RZE	3	T	112
2RH1	4	P	168	3PBL	4	P	167	3ODU	4	P	170	3EML	4	P	139	3RZE	4	S	155
2RH1	5	P	211	3PBL	5	P	200	3ODU	5	P	211	3EML	5	P	189	3RZE	5	P	202
2RH1	6	P	288	3PBL	6	P	344	3ODU	6	P	254	3EML	6	P	248	3RZE	6	P	430
2RH1	7	P	323	3PBL	7	P	380	3ODU	7	P	299	3EML	7	P	285	3RZE	7	P	465

2.3 形变弧度的连续 von Mises 建模

本文采用一元 von Mises 分布的 cosine 变形[19]（记为 M1 分布）的连续型概率分布对形变角度进行建模。类似于高斯分布对线距离数据进行建模，von Mises 分布对于角弧度数据进行建模具有相当重要的科学意义和适用性[20]。标准一元 von Mises 分布的概率密度函数如下[21]：

$$g(\varphi)\sim M1(k,\mu)=(2\pi I_0(k))^{-1}\exp(k\cos(\varphi-\mu)) \quad (1)$$

式中，k 代表整个数据集的聚集度，这里可以用方差的倒数来近似表示[21]，μ 代表整个数据集的平均水平，$I_0(k)$ 是 0 阶第一类修正 Bessel 函数。本文估计参数 k 和 μ 时比较了两种不同方法。方法 1 采用正态分布的参数估计，即用均值和方差来代替参数，如文献[7]中使用的方法。方法 2 采用矩估计方式。两种方法的估计过程如下。

方法 1　正态分布估计方法。参数 μ 的估计，$\mu=\sum_{i=1}^{n}\varphi_i/n$，参数 k 的估计 $var=\sum_{i=1}^{n}(\sin\varphi_i)^2/n$，故 $k=1/var$，其中 φ_i 表示统计得出的所有螺旋结构的形变的弧度值，范围为 $(0,\pi/2)$。

方法 2　矩估计方法。参数 μ 的估计，其中减 π 是因为要让参数落在 $(-\pi,\pi)$ 之间。

$$\mu=\arctan(\sum_{i=1}^{n}\sin\varphi_i/\sum_{i=1}^{n}\cos\varphi_i)-\pi \quad (2)$$

参数 k 的估计，这里的 k 表示聚集度，类似正态分布，$1/k$ 接近于 σ^2。

$$R_e^2=\frac{I_1(k)^2}{I_0(k)^2}=\frac{N}{N-1}\{(\frac{\sum_{i=1}^{n}\cos\varphi_i}{n})+(\frac{\sum_{i=1}^{n}\sin\varphi_i}{n})-\frac{1}{n}\} \quad (3)$$

$$\text{因为方差：}var(x)=1-E[\cos(x-\mu)]=1-\frac{I_1(k)}{I_0(k)} \quad (4)$$

根据式(3)可求得 R_e(即 $I_1(k)/I_0(k)$)，这样将 R_e 代入式(4)，求得方差 $var(x)$，那么聚集度 $k=1/var(x)$。

2.4　模型采样

本文使用离散均值概率采样和连续型模型采样对统计出的螺旋形变角度值进行对比。其中离散采样方法流程，首先将所有的角度数据存入数组，其次随机生成一个整数作为数组下标，取出数组中对应的角度值，这样即为采样一次，如此往复直到采样结束。

本文的连续模型的采样算法是现有的一元 von Mises 的采样算法[13]的改进，参见算法 1。

算法 1　对一元 von Mises 分布 $g(\varphi)-M1(k,\mu)$ 采样

$a=1+\sqrt{1+4k^2}$，$b=(a-\sqrt{2a})/(2k)$，$r=(1+b^2)/(2b)$

While 没有找到满足条件的 f do

　随机产生 μ (0,1)

　$z=\cos(\mu*\pi)$，$f=(1+r*z)/(r+z)$，$c=k*(r-f)$

　随机产生 μ(0,1)

　if $c*(2-c)-\mu>0$ OR $\log(c/\mu)+1-c>=0$ then

　f 满足条件，退出 while 循环

　end if

　end while

以概率 0.5 随机选择 $(+/-\arccos(f)+\mu)\bmod(2\pi)$ 两个值之一作为 φ 返回

这里需要注意的是采样算法的输入 μ 和 k 必须在弧度值下。因统计的角度值是在 $(0,\pi/2)$ 之间，需要将统计出的角度值 φ_i 转换到 $(-\pi,\pi)$ 之间，故做变换 $\varphi_i'=4\varphi_i-\pi$。采样算法的输出 φ' 是在 $(-\pi,\pi)$ 之间的，需要转换到 $(0,\pi/2)$ 之间，故做变换 $\varphi=(\varphi'+\pi)/4$。

3　模型验证与讨论

模型验证采用回归性测试和开放性测试来验证，开放性测试采用了 2011 年 9 月后新增中的 7 个 PDB 共 35 根天然 GPCR 跨膜螺旋，见表 2。

3.1　验证实验设计

本文总共设计了 5 组验证实验，其中前 4 组是回归性测试，第 5 组为开放性测试，具体验证实验设计见表 3。

表 3　5 组验证实验的设计

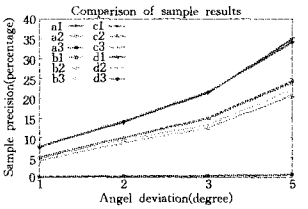
模型类型	是否聚类	参数估计方式	回归测试	开放测试
离散模型	非聚类	无参数估计	实验 1	
连续模型	非聚类	参数估计 1	实验 2	
连续模型	非聚类	参数估计 2	实验 3	
连续模型	聚 类	参数估计 2	实验 4	实验 5

第 1 组实验是离散模型的采样。第 2 组实验是螺旋未聚类前，采用正态分布估计 von Mises 模型的参数。第 3 组实验是采用矩估计方式来估计 von Mises 模型的参数。第 4 组实验是螺旋聚类后，采用正态分布估计 von Mises 模型的参数。第 5 组实验是螺旋聚类后，采用矩估计参数估计的开放性测试。最后与文献[7]的方法进行间接比较。

3.2　验证实验结果

本文一个比较直接的结果是对训练集的 707 根和开放测试集的 35 根天然的 GPCR 跨膜螺旋的形变信息的统计，包括螺旋的 ID、形变位置、形变程度、形变角度等信息。这些重要的数据集可以为其他一些模型提供很好的准确的数据源。

模型的验证测试结果(假设采样 N 次，$N=100,200,500$)，统计预测形变角度与天然形变角度的误差在指定范围($\pm1°,\pm2°,\pm3°,\pm5°$)内出现的比率。前 4 组回归测试实验结果如图 2 所示，第 5 组开放测试实验结果如图 4 所示。



横坐标为角度误差(°)，纵坐标为相应角度误差出现的比率(%)

图 2　封闭式测试的 4 组测试结果的比较

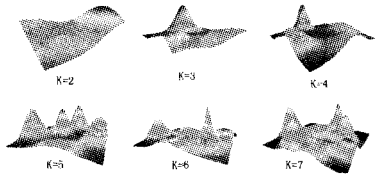
第 1 组实验，离散采样实验结果如图 2 中的接近横坐标的折线 a1($N=100$)，a2($N=200$)，a3($N=500$)所示。实验结果几乎接近零比率，说明了螺旋的形变角度不能简单地使用离散模型进行建模采样。

第 2 组实验，连续采样的实验结果如图 2 中处在中间的折线 b1($N=100$)，b2($N=200$)，b3($N=500$)所示。对比第 1、第 2 两组实验的结果可以看到，第 2 组实验结果有了很大的提高，预测形变角度误差在 1°左右的结果，从第 1 组实验几乎接近 0 的比率提高到 5%以上，说明对于本文的螺旋形变角度的建模，连续模型要明显优于离散模型。

第 3 组实验，采样的实验结果如图 2 中处在中间的折线 c1($N=100$)，c2($N=200$)，c3($N=500$)所示。第 2 和第 3 两组实验结果对比表明，在本文的形变弧度数据集下，连续模型下的方法 2 矩估计方式要稍微优于方法 1 的正态分布估计方式。

第 4 组实验，本文使用 Repeated Bisection 聚类算法的实现[22]，比较了类个数 K 为 2,3,4,5,6,7 时的聚类结果，如图 3 所示。通过可视化分析发现，分成 5 类时聚类结果最佳。

此时，从山坡的高度可知各个类内相似性最高，从山谷的颜色可知类间的差异性最显著，从各山坡的体积可知各类元素分布最为均匀。

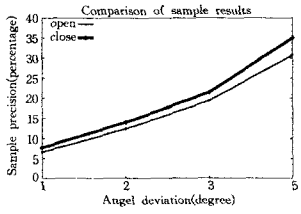


(a)波峰的个数表示类的个数。(b)波峰的高度表示该类中元素相似性的程度,波峰越高就越相似。(c)在同一山坡中元素的相似性高于其它山坡中元素的相似性。(d)山坡的形状是高斯曲线,说明了类内数据的分布。(e)山坡的体积与类中的元素数量成正比。(f)颜色代表了类中内部数据的差异度,红色和黄色代表低差异、蓝色代表高差异性。

图 3　不同聚类数($K=2,3,4,5,6,7$)的三维图结果

聚类后进行回归测试,即使用矩估计方法重新参数估计和采样,采样的结果表明大部分的螺旋的采样精度提高,如图 2 中处在上方的折线 d1($N=100$),d2($N=200$),d3($N=500$)所示。从本文的实验结果中可知,有超过 80%的螺旋,其天然形变角度与采样的形变误差在 1°内比率可达到 10%左右,即这些螺旋只需 10 次采样就能有一次采样结果大致符合天然的形变角度,而且采样精度低的螺旋数不到总螺旋总数的 10%。

第 5 组实验,使用那些不在训练集中的天然螺旋进行测试,这也是本文模型非常重要的结果,其测试的结果与回归测试的结果的比较如图 4 所示。图中说明了开放性测试的结果与回归性测试的结果比较接近,表明了本文模型对于开放性测试也很适用。



横坐标为角度误差(°),纵坐标为相应角度误差出现的比率(%)

图 4　封闭式测试与开放性测试实验结果的比较

同时,将本文方法与文献[7]方法进行比较,本文使用与序列相关的连续的 von Mises5 模型,文献[7]使用与残基 Pro 相关的离散 Monte Carlo 采样。比较了这两种方法分别对高精度模型与低精度模型采样的成功率,从图 4 中可以看出,对于高精度模型,即误差在 1°以内,本文的模型对于任意给定未知结构的 GPCR 跨膜螺旋的氨基酸序列,只需采样 15 次,就有一个预测值与天然的形变角度值的误差在 1°以内。而文献[7]中的方法,需采样 40 次才能获得一个高精度模型;对于低精度模型,即误差为 5°以内,本文采样的比率是 34%,而文献[7]则为 11%。可以看出,针对本实际问题,连续 von Mises5 模型的性能确实优于离散模型的性能。

3.3　讨论

本文的第 1 组与第 2、第 3、第 4 组实验的结果对比表明,连续模型比离散模型更适合本文的数据集,故可以得到更加精确的跨膜螺旋的形变角度值。不同于离散模型,连续模型不需要对螺旋结构的形变角度进行识别,几乎完全依赖于能量函数。使用连续模型的采样可以很好地表征整个天然 GPCR 跨膜螺旋的形变角度的概率分布。本文的连续 von Mises 模型的参数有两种估计方式,即方法 1 的正态分布估计和方法 2 的矩估计。方法 2 的估计方式虽然在结果上稍微优于方法 1,但没有明显的差异,可能的原因是两种估计方式对连续型数据分布的拟合,都已经做得很好了,故采用不同的估计方法对实验的结果影响不大。

本文的第 2 组、3 组与第 4 组的实验结果对比表明,在连续模型中融入序列信息,更可以帮助提高螺旋形变角度的预测精度。其主要原因是加入了螺旋的序列的全部信息,而不仅仅限于考虑形变前后的一些特殊的残基,如文献[7]中只对螺旋形变角度进行两类处理,其中一类是形变在 Pro 残基,另一类是形变在 Pro 残基并且在 $i-1$ 位置是 Ser 残基、$i-3$ 位置是 Thr 残基[23]。两类不同的螺旋采用不同的均值和方差来控制,这种模型只能与个别残基相关,不能与序列相关。这样对于一个未知结构的螺旋序列信息,可以根据序列相似性来定位类别,并使用不同的参数来采样,对螺旋之间进行了真正的有差别对待。

结束语　本文针对已知结构的天然 GPCR 的跨膜螺旋形变建模。使用了 Repeated Bisection 聚类算法对收集的螺旋根据序列相似性进行聚类,然后根据不同的类别进行 von Mises 概率模型的参数估计,最后对模型进行采样。根据本文的概率模型,可以对已知序列未知结构的跨膜螺旋提供一组形变角度采样值,帮助提高螺旋的形变角度预测。本文不仅从蛋白质的三维结构出发,还加入了螺旋序列信息指导采样,使采样更加具有针对性。本文的结果可以直接用于 GPCR 的七螺旋建模,以及 GPCR 整体三维结构的重建。

参 考 文 献

[1] 吴宏杰,吕强,吴进珍,等. 从头预测蛋白质骨架的一种并行蚁群方法及其在 CASP8/9 中应用[J]. 中国科学:信息科学,2011

[2] Zhang Yang, Zhang Jian. GPCRRD: G protein-coupled receptor spatial restraint database for 3-D structure modeling and function annotation[J]. Bioinformatics,2010,26(2):3004-3005

[3] Langelaan D N, Christian B M W, Rainey J K. Improved Helix and Kink Characterization in Membrane Proteins Allows Evaluation of Kink Sequence Predictors [J]. Journal of Chemical Information Modelling,2010,50(12):2213-2220

[4] Wang R Y-R, Han Y, Krassovsky K, et al. Modeling Disordered Regions in Proteins Using Rosetta[J]. PLoS ONE,2011,6(7):e22060

[5] Meher J K, Nisignal M, Pranab K M, et al. Signal processing approach for prediction of kink in transmembrane a-helices [J]. CCIS,2011,147(1):170-177

[6] Zaki N, Bouktif S, Lazarova-Molnar S. A Combination of Compositional Index and Genetic Algorithm for Predicting Transmembrane Helical Segments[J]. PLoS ONE,2011,6(7):e21821

[7] Barth P, Schonbru J, Baker D. Toward high-resolution prediction and design of transmembrane helical protein structures[J]. Proceedings of National Academy of Science of USA,2007,28(6):15682-15687

[8] Cordes F S, Bright J N, Sansom M S P. Proline-induced distorsions of transmembrane helices[J]. Journal Molecular Biology, 2002,323:951-960

[9] Shen Hong-bin, Chou J J. MemBrain: Improving the Accuracy of Predicting Transmembrane Helices[J]. PLoS ONE,2008,3(6):e2399

[10] Robert C P, Casella G. Monte Carlo Statistical Methods(2nd edition)[M]. New York: Springer,2004

[11] Baker D, Jeffrey J, Moughon S. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations [J]. Journal of Molecular Biology, J-567

Health Science, Box 357350, Seattle WA 98195, USA, 2003: 281-299

[12] Sarah Y, Salem F, Duan Yang, et al. The evolution of transmembrane helix kinks and the structural diversity of g proteincoupled receptors[Z]. California Institute of Technology, Pasadena, CA, 2003: 959-963

[13] Best D J, Fisher N I. Efficient simulation of the von mises distribution[J]. Journal ofthe Royal Statistical Society. Series C (Applied Statistics), 1979, 28(2): 152-157

[14] Meruelo A D, James I S, Bowie U. TMKink: A method to predict transmembrane helix kinks[J]. Protein Science, 2011, 20: 1256-1264

[15] Apweiler R, Bairoch A, Wu C. Uniprot: the universal protein knowledgebase[J]. Nucleic acids Research, 2004, 32: 115-119

[16] Zhao Ying, Karypis G. Evaluation of hierarchical clustering algorithms for document datasets[C]// Proc. of Int'l. Conf on Information and Knowledge Management. 2002: 515-524

[17] 黄旭，吕强，钱培德. 一种用于蛋白质结构聚类的聚类中心选择算法[J]. 自动化学报，2011，37(6)：682-692

[18] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416

[19] Mardia K V, Taylor C C, Subramaniam G K. Protein bioinformatics and mixtures of bivariate von mises distributions for angular data[J]. Journal of the Royal Statistical Society, Series C (Applied Statistics), 2007, 63(2): 505-512

[20] Fisher N I. Statistical Analysis of Circular Data[M]. Cambridge University Press, 1996

[21] wikipedia [EB/OL]. http://en.wikipedia.org/wiki/Talk:Von Mises distribution. URL, Oct. 2010

[22] Rasmussen M, Karypis G. gcluto: An interactive clustering, visualization, and analysis system[R]. UMN-CS TR-04-021. 2004

[23] Deupi X, Olivelia M, Govaerts C, et al. Ser and thr resiudes modulate the conformation of pro-kinked transmembrane α-helices [J]. Biophysical Journal, 2004, 86: 105-115

(上接第 176 页)

务描述方式对应的召回率和准确度的函数关系。在评价一个匹配算法的效率时，片面地考虑召回率或准确率都不够全面。为了全面地考察实验结果，在准确率为 Y 轴、召回率为 X 轴的图上画出了不同真实度下的准确率与召回率的函数关系图形。实验统计结果见图 5，两条曲线分别对应了两种匹配算法的召回率-准确率曲线。不难看出，在大多数情况下星点实线所标示的性能曲线性能高于虚线所标示的性能曲线。不难发现一个匹配算法的性能越好，其对应的召回率-准确率曲线越向上凸，具体来讲就是曲线与 X 轴、Y 轴所围成的图形的面积越大，匹配算法的性能越好。最理想的情况是，面积为 1。在不同真实度条件下，召回率和准确率统计实验结果显示了基于 Petri 网的 OWL-S 语义匹配相对于传统 UDDI 匹配机制具有明显优势。

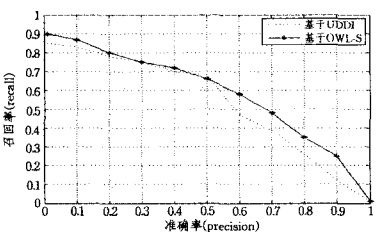


图 5 实验结果

结束语 本文提出了一种基于 Petri 网的 OWL-S 语义匹配机制，即借助 Multi-Agent 服务发现框架，使用 PNSDL 描述语言来发布和请求服务。提出了一个基于 OWL-S 语义的服务匹配机制，给出了相应的匹配算法。该匹配机制以用可能性和必然性程度代表服务能胜任需求的置信度，实现服务的匹配以提高服务发现的召回率和准确度的效率。该机制具有强大的逻辑推理能力，并能对多服务建模，提高匹配的真实度、准确度和召回率。

下一步将在 Multi-Agent 服务发现框架及大型的分布式环境中对服务组合问题进行研究。

参 考 文 献

[1] 岳昆，王晓玲，周傲英. Web 服务核心支撑技术：研究综述[J]. 软件学报，2004，15(3)：428-442

[2] Clement L, Hately A, von Riegen, et al. UDDI Version 3. 0. 2 [EB/OL]. http://uddi.org/pubs/uddi2v3.0.2-20041019.htm, 2004

[3] Martin D, Burstein M, Hobbs J, et al. OWL2S: Semantic Markup for Web Services[EB/OL]. http://www.w3.org/ Submission/ 2004/ SUBM2OWL2S220041122/, 2004

[4] Huhns M N, Singh M P. Ontologies for agents [J]. IEEE Internet Computing, 1997, 1(6): 81-83

[5] Trastour D, Bartolini C, Castillo J G. A service Web approach to service description for matchmaking of services [R]. HP Labs, Tech Rep: HPL2001-183. 2001

[6] Verma K, Sivashanmugam K, Sheth A, et al. METEOR-S WSDI: A scalable P2P infrastructure of registries for semantic publication and discovery of Web services[J]. Journal of Information Technology and Management, 2005, 6(1): 17-39

[7] Chen C W, Gan P S, Yang C H. A service discovery mechanism with load balance issue in decentralized peer-to-peer network[C]// Barolli L, ed. Proc. of the 11th Int'l Conf. on Parallel and Distributed Systems (ICPADS 2005). Washington: IEEE Computer Society Press, 2005: 592-598

[8] 马千里，廖明宏，高振国，等. 普适计算环境下的服务发现协议[J]. 计算机工程，2009，35(17)：247-248

[9] Petri C A. Kommunikation mit Automaten[D]. Bonn, Germany: Institute for Instrumentelle Mathematik, Schriften des IIM, 1962

[10] Vedal J C, Lama M, Bugarin A. A High-level Petri Net Ontology Compatible with PNML[C]// ACSD, Petri Net Markup Language Forum 2006. Turku, 2006

[11] Weber J, Kindler E. The Petri Net Markup Language[J]. Lecture Notes in Computer Science, 2003, 2472: 124-144

[12] 张广胜，蒋昌俊，丁志军. 基于模糊 Petri 网的服务发现框架研究[J]. 计算机研究与发展，2006，43(11)：1886-1894

[13] Wooldrige M. Intelligent Agent: Theory and Practice [J]. Knowledge Engineering Review, 1995, 10(2): 115-152

[14] 郑啸，罗军舟，宋爱波. 基于 Agent 和蚁群算法的分布式服务发现[J]. 软件学报，2010，21(8)：1795-1809

[15] 翟正利，杨扬. 基于模糊 Petri 网和本体的网格服务发现[J]. 北京科技大学学报，2006，28(12)：1196-1200

[16] 邱田，胡晓惠，李鹏飞，等. 基于 OWL-S 的服务发现语义匹配机制[J]. 电子学报，2010，38(1)：42-47