跨膜蛋白结构预测新方法

刘宏德① 王 睿① 卢小泉①* 陈 晶① 刘秀辉① 丁 兰②

(① 西北师范大学化学化工学院, 兰州 730070; ② 西北师范大学生命科学学院, 兰州 730070. * 联系人, E-mail: luxq@nwnu.edu.cn)

摘要 基因组数据中大约有 20%～30%的基因产物被预测为膜蛋白, 膜蛋白是一类具有重要生物功能的蛋白质. 预测膜蛋白跨膜区的数量和位置是生物信息学中重要的研究课题. 提出了一种预测膜蛋白跨膜区的新方法——连续小波变换极大值谱(MSCWT). 该法对 8 种 SARS-CoV 膜蛋白的预测准确度与常用膜蛋白预测软件 TMpred 相当, 对 MPtopo 数据库中 131 种新的已知结构的螺旋束蛋白(共包含 548 个跨膜区)的预测显示, 其跨膜螺旋区预测准确率为 91.6%, 膜蛋白序列的预测准确率为 89.3%.

关键词 连续小波变换极大值谱(MSCWT) 膜蛋白 跨膜区

膜蛋白是一类嵌在生物膜中的蛋白质, 在细胞中具有重要的生物功能, 它们构成了各种神经信号分子、激素和受体, 是各种离子跨膜的通道 [1], 也是许多药物分子的靶点 [2]. 然而, 膜蛋白与生物膜的稳定构象非常不利于用X光晶体衍射方法和核磁共振技术测定其三维结构, 目前仅有少数膜蛋白的结构已知. 因此, 设计准确、高效的预测膜蛋白结构的方法成为生物信息学中重要的研究课题.

Kyte等人 [3]利用氨基酸疏水自由能, 通过一个滑动窗口把蛋白质序列换化为疏水图谱, 再设定合适的阈值来预测跨膜区. 随后, von Heijne[4]提出了著名的正电荷居内规则, 为膜蛋白的结构预测提供了进一步的指导. Rost等人 [5]将膜蛋白多序列比对的信息整合到一个神经网络, 结合正电荷居内规则来预测跨膜区. Sonnhammer等人 [6]提出了基于隐马尔可夫模型的方法, 该方法采用了 7 个状态的模型, 这 7 个状态分别对应于跨膜蛋白的不同区域, 即跨膜核心、跨膜区两边的跨膜末端、膜内的环、膜外的短环和长环、远离膜的区域. Jones等人 [7]通过统计分析, 得到各种氨基酸在膜内、膜外、跨膜核心区、跨膜末端区出现的频率和它们在整个跨膜蛋白序列中出现的频率, 根据这两个频率的比值得到氨基酸出现的偏好性, 最后根据所得到的偏好性通过动态规划算法进行结构预测. 与文献[7]相比, 文献[8,9]的预测也是基于残基出现频率的, 不同的是, 他们实现了对跨膜方向的预测.

小波分析被誉为“数学放大镜”, 它能通过伸缩和平移等运算功能对信号进行多尺度细化分析. 因此使用小波分析, 能够同时得到信号在时、频两域的局部特征. 对于研究非平稳信号、突变信号, 小波分析拥有天然的优越性 [10,11]. 在生物信息学领域, Hirakawa等人 [12]首次利用小波变换的多分辨分析来预测蛋白质的疏水核; Mandell等人 [13]通过对疏水值序列使用小波变换来区分蛋白质的结构; Murray等人 [14]用连续小波变换来描述和检测蛋白质的活性区域; 邱建丁等人 [15]利用单尺度上的连续小波变换预测膜蛋白的跨膜区.

本文利用作者提出的连续小波变换极大值谱(MSCWT)来预测膜蛋白跨膜区. 此前, 该方法已经成功用于色谱、毛细管电泳、差示脉冲伏安信号和紫外光谱信号等重叠信号的解析 [16]. MSCWT不同于单尺度上的极大值检测, 它是基于多尺度下的连续小波变换, 可以同时检测信号在各尺度下小波系数的极大值, 在分析多频率组分信号的变化规律中非常有用.

1 数据和方法

(ⅰ) 数据. 测试数据来源于 MPtopo 数据库(http://blanco.biomol.uci.edu/mptopo/), 库中的所有蛋白质的结构都已经过结晶学、基因融合和其他方法的验证. 8 个 SARS-CoV 膜蛋白(orf 3, orf 4, orf 7, orf 8, orf 9, orf 14, m, s)序列则来自网站: http://athena.bioc.uvic.ca/sars/map/diagram.html.

(ⅱ) 连续小波变换极大值谱(MSCWT). 信号$f(t)$的连续小波变换(CWT)定义为式(1), 其中a, b分别为尺度参数和平移参数; $\Psi a,b(t)$为小波函数; 变换结果$Wf(a, b)$是一个三维向量, 是a与b的函数. 尺度参数

2007-07-31 收稿, 2007-10-24 接受

国家自然科学基金(批准号: 20775060, 20335030)、第 4 届教育部高校青年教师奖和甘肃省高分子重点实验室资助项目

a 与频率成反比.

$$Wf(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}f(t)\psi_{a,b}^{*}(t)\mathrm{d}t. \quad (1)$$

MSCWT是基于CWT衍生出来的一种波谱 [16], 主要用于分析多频率组分信号的变化规律. MSCWT的获取方法如下: 对一个时域信号, 首先进行CWT, 把信号从时域投影到小波域, 这样, 原信号中不同的频率的组分便分布在小波域的不同尺度上. 然后在不同尺度上检测信号的极大值, 并记录极大值所对应的位置(平移参数b). 最后以检测到的极大值和其对应的位置作图, 就得到了MSCWT. 需要指出的是由于分析多频率组分的信号特征, 所以本文使用的CWT是在一个合适的尺度范围内进行的.

MSCWT 伪代码

```
**********************************************
For b=1 to length of signal
  For a = a1 to a2
    Detect the maximum of Wf(a, b)
  Next a
    Plot the maximum
Next b
```

注释: $a1$ 和 $a2$ 是尺度范围的上下限; $Wf(a, b)$是信号的连续小波变换.

```
**********************************************
```

(iii) 参数选择. 在实施CWT时, 需要选择的参数有: (1) 小波函数; (2) 尺度范围. 关于参数①的选择在文献中已有详述 [16], 这里使用文献的结果, 即以Morlet小波作为分析小波. 关于参数②, 一律使用区间[1~64]为尺度范围, 计算步长为 0.5.

(iv) 实验步骤. 用 MSCWT 预测膜蛋白跨膜区的过程为:

(1) 将膜蛋白序列的残基用其疏水自由能表示, 构成新的数字序列(疏水序列);

(2) 对疏水序列进行连续小波变换, 进而提取 MSCWT 谱;

(3) 以 1.5 为阈值对跨膜区域进行预测, 其规则为: 在 MSCWT 谱上, 大于 1.5 的数值所对应的位置被认为是候选区域; 候选区域的长度应大于 20, 小于 20 的片段一般认为不可能具有跨膜功能; 如果两个或 3 个候选区域(不超过 3 个)离得比较近(序列长度<3), 则将它们合并为一个区域.

(v) 预测准确性评价指标. 为了检验预测方法的可靠性, 本文从氨基酸、跨膜区以及蛋白质序列 3 个层次来对测试结果进行评价 [17]:

(1) 氨基酸残基的正确性. 残基准确预测率 $F_{\mathrm{AAcor}}=(N_{\mathrm{AAcor}}/N_{\mathrm{AAall}})\times 100\%$, 其中 N_{AAcor} 是正确预测的残基总数; N_{AAall} 是所有残基数目.

(2) 跨膜区的正确性. 主要采用以下指标来检测所预测跨膜区的正确性:

① 假阳性数: FP;

② 假阴性数: FN;

③ 跨膜区预测准确率: $Q_{\mathrm{P}}=\sqrt{M\times C}\times 100\%$, 其中 $M=N_{\mathrm{cor}}/N_{\mathrm{obs}}$, 可被认为是灵敏度指标; $C=N_{\mathrm{cor}}/N_{\mathrm{prd}}$ 是一种特征性指标; N_{cor} 是正确预测的跨膜区总数; N_{obs} 是实际跨膜区总数; N_{prd} 是预测的跨膜区总数.

(3) 膜蛋白序列预测的正确性. 如果一条膜蛋白序列中所有的跨膜区都被正确预测, 那么就认为这条膜蛋白序列是被正确预测的膜蛋白.

膜蛋白序列预测的准确率: $Q_{\mathrm{t}}=(N_{\mathrm{TT}}/N_{\mathrm{TOR}})\times 100\%$, 其中 N_{TT} 是被正确预测的膜蛋白序列数; N_{TOR} 是所有的膜蛋白序列数.

此外, 本文也将 MSCWT 和其他两种广泛应用的膜蛋白预测软件进行了比较, 一种是 TMpred (http://www.ch.embnet.org/software/TMPRED_form.html); 另外一种是 DAS (http://www.sbc.su.se/~miklos/DAS/).

2 结果与讨论

MSCWT可用来预测任何单一肽链上跨膜区段存在与否、跨膜区段的起始和终止位点等信息. 此外, MSCWT还可用于进一步分析单一肽链上其他更细微的结构.

图 1 是 MSCWT 方法与 TMpred 膜蛋白预测软件分别对 SARS-CoV 的 Orf 9, Orf 14 蛋白跨膜区域的预测图谱比较. TMpred 图中实线和虚线分别表示对同一种蛋白序列 TMpred 给出的两种建议模型(strongly preferred model 和 alternative model)的信号. 而 $i<-o$ 和 $o<-i$ 分别表示建议模型相对膜的取向, 前者为由内到外, 后者为由外到内. 从图中可以发现两点: (1) MSCWT 放大了原疏水序列中的疏水和亲水残基排布的特性, 这有利于观察蛋白疏水性随序列的变化. 所要研究的主要峰显现出来, 这些峰对应着疏水性的改变, 这是跨膜蛋白片段的特征. 事实上, 以 1.5 为阈值, 曲线中大于 1.5 的部分所对应的区域就是该蛋白的跨膜区域. 由 MSCWT 按本文所述方法预测的

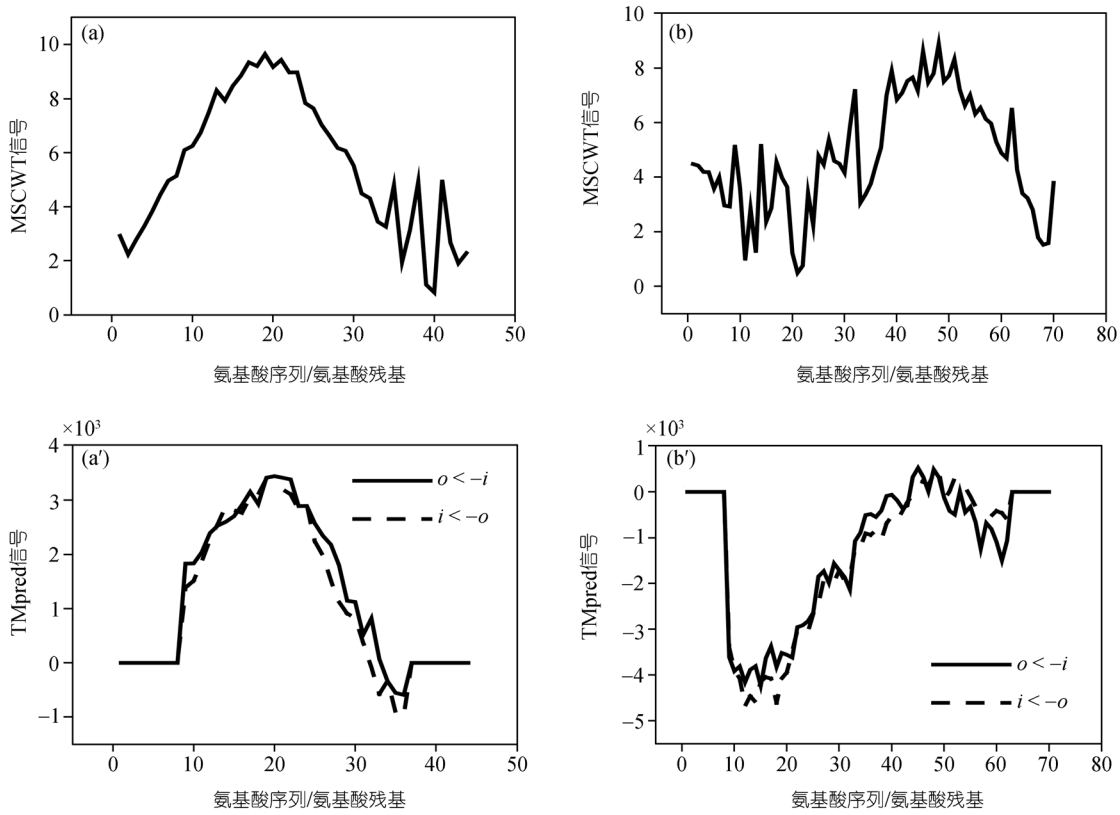

图 1 MSCWT 和 TMpred 对 SARS 的 Orf 9 和 Orf 14 蛋白跨膜区域的预测比较

(a) Orf 9 的 MSCWT(跨膜区: 8~30); (b) Orf 14 的 MSCWT(跨膜区: 36~58); (a)′ Orf 9 的 TMpred 图谱(跨膜区: 9~30); (b)′ Orf 14 的 TMpred 图谱(跨膜区: 36~58). TMpred 图中, 实线和虚线分别表示对同一种蛋白序列 TMpred 给出的两种建议模型的信号, 而 $i<-o$ 和 $o<-i$ 分别表示建议模型相对膜的取向, 前者为由内到外, 后者为由外到内

8 种 SARS-CoV 跨膜片段的结果显示在表 1 中. MSCWT 与 TMpred 预测的结果非常接近. 在图 1 中可以看到 MSCWT 与 TMpred 预测的跨膜信号峰所对应的位置基本一致, 说明 MSCWT 的预测达到了 TMpred 的预测精度, 能够较准确地预测出膜蛋白跨膜区的个数和位置. (2) 由于 CWT 是在一定频率范围内进行的, 所以, 较大的尺度对应信号低频率部分, 较小的尺度对应信号的高频率. 因此, MSCWT 不仅能够揭示信号的整体特点, 而且可以反映局部的变化. 在疏水序列的 MSCWT 中, 大跨度的峰一般不是特别平滑, 这说明在跨膜片段内, 蛋白的疏水性随残基位置的不同而不同, 进而可以推测这些蛋白是以更为复杂的折叠方式存在于跨膜片段内的, 而 MSCWT 可为其折叠结构的预测提供有价值的参考. 虽然图 1 的 MSCWT 图谱中有很多“毛刺”, 有时妨碍了对跨膜区域的发现, 但是这些“毛刺”恰好提供了关于跨膜片段内部的疏水特性, 所以本文并没有将高频率的“毛刺”滤除. 如果只关心序列的整体特征, 只需在进行 CWT 时将起始的尺度设置为较大的值. 我们在实验中发现如果将起始的尺度设为 8, 则 MSCWT 曲线变得比较光滑, 但是这时无法观察局部的任何微小的变化.

为了更进一步检验预测方法的可靠性, 我们将最新的 Mptopo 数据库(2007-07-26)中所有三维结构已知的蛋白序列作为测试集, 共包含 131 种膜蛋白序列, 548 个跨膜区. 这个测试集的记录中含有由实验方法确定的跨膜区段的数目和各区段的起始和终止位置, 因此可以将这些作为可靠的已知样本. MSCWT 的预测结果显示: 548 个跨膜区中有 530 被正确预测出, 假阳性有 81 个, 还有 18 个跨膜区未测出来, 漏测率为 3.3%. 此外, 131 种膜蛋白中有 117 条膜蛋白序列中所有的跨膜区都被正确预测, 我们的预测方法得到的跨膜螺旋区预测准确率为 91.6%, 膜蛋白序列的预测准确率为 89.3%.

表 1 MSCWT 和 TMpred 对 SARS-CoV 膜蛋白跨膜区的预测比较(MSCWT 的阈值为 1.5)

蛋白序列	方法	跨膜区 1	跨膜区 2	跨膜区 3	跨膜区 4
Orf 7	MSCWT	7~29	32~41		
	TMpred	7~29	21~41		
Orf 8	MSCWT	1~14	47~68		99~117
	TMpred	1~18			99~117
Orf 9	MSCWT	8~30			
	TMpred	9~30			
Orf 4	MSCWT	4~20			52~75
	TMpred	1~20			52~75
Orf 14	MSCWT	36~58			
	TMpred	36~58			
Orf 3	MSCWT	37~54		92~121	
	TMpred	41~58	76~99	86~121	
s	MSCWT	1~15	55~85	233~240: 575~594	309~325:530~540
	TMpred	1~17	51~73	233~257:570~588	345~364:524~542
m	MSCWT	21~32	47~63	72~95	125~146
	TMpred	21~38	51~69	74~96	

另外, 我们按顺序从这个测试集中抽出前面 65 个膜蛋白作为检验集(包含 277 个跨膜区), 用 TMpred 和 DAS 对测试集进行预测并将预测结果进行了比较(表 2). 3 种方法的预测结果略有差异. 相比之下, MSCWT 的漏测区段最少, 对预测的 277 个跨膜区, MSCWT 只漏测了 13 个, 而 TMpred 和 DAS 漏测区段分别为 33 和 30 个.

表 2 MSCWT, TMpred 和 DAS 的预测性能

方法	N_{obs}	N_{prd}	N_{cor}	FP	FN	Q_p	N_{TOR}	N_{TT}	Q_t
MSCWT	277	313	264	49	13	89.7%	65	55	84.6%
TMpred	277	248	244	4	33	93.1%	65	49	75.4%
DAS	277	278	247	31	30	89.0%	65	52	80.0 %

对跨膜螺旋区预测准确率而言, TMpred 最高, 达 93.1%, MSCWT 的此值居于 TMpred 和 DAS 中间, 为 89.7%, DAS 最低, 为 89.0%. MSCWT 的膜蛋白序列的预测准确率最高, 为 84.6%, TMpred 和 DAS 在该数据集上的膜蛋白序列的预测准确率只是 75.4%和 80.0%. 结果表明, 本文提出的方法达到了现有预测算法中较高的水平, 能够较为准确地预测出膜蛋白跨膜区.

3 结论

本文提出的研究膜蛋白跨膜区的新方法——连续小波变换极大值谱(MSCWT)能够快速准确地预测跨膜区的位置和数量, 有望成为研究膜蛋白的一种有力的新工具. MSCWT 的缺陷在于无法提供跨膜片段的方向信息.

参 考 文 献

1 Tanford C. How protein chemists learned about the hydrophobic factor. Protein Sci, 1997, 6(6): 1358—1366

2 Adams P D, Arkin I T, Engelman D M, et al. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. Nat Struct Biol, 1995, 2(2): 154—162[DOI]

3 Kyte J, Doolittle R F. A simple method for displaying the hydropathic character of a protein. J Mol Biol, 1982, 157(1): 105—132[DOI]

4 von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive rule. J Mol Biol, 1992, 225(2): 487—494[DOI]

5 Rost B, Casadio R, Fariselli P. Transmembrane helices predicted at 95% accuracy. Pro Sci, 1995, 4(3): 521—533

6 Sonnhammer E L, von Heijne G, Krogh A. A Hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol, 1998, 6: 175—182

7 Jones D T, Taylor W R, Thornton J M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry, 1994, 33(10): 3038—3049[DOI]

8 Persson B, Argos P. Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. J Mol Biol, 1994, 273(2): 182—192[DOI]

9 Persson B, Argos P. Topology prediction of membrane proteins. Pro Sci, 1996, 5(2): 363—371

10 卢小泉, 莫金垣. 分析化学计量学中的新方法——小波分析. 分析化学, 1996, 24(9): 1100—1106

11 卢小泉, 刘宏德. 分析化学中的小波分析技术. 北京: 化学工业出版社, 2006

12 Hirakawa H, Muta S, Kuhara S. The hydrophobic cores of proteins predicted by wavelet analysis. Bioinformatics, 1999, 15(2): 141—148[DOI]

13 Mandell A J, Selz K A, Shlesinger M F. Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families. Physica A, 1997, 244: 254—262[DOI]

14 Murray K, Gorse D, Thornton J. Wavelet transforms for the characterization and detection of repeating motifs. Mol Biol, 2002, 316(2): 341—363[DOI]

15 Qiu J D, Liang R P, Mo J Y, et al. Prediction of transmembrane proteins based on the continuous wavelet transform. Chem Inf Comput Sci, 2004, 44(2): 741—747[DOI]

16 Lu X Q, Liu H D, Xue Z H, et al. Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal. J Chem Inf Comput Sci, 2004, 44(4): 1228—1237[DOI]

17 陈钟强, 刘琪, 朱贻盛, 等. 膜蛋白跨膜区预测方法的评价. 生物化学与生物物理学报, 2002, 34(3): 285—290