综述与专论

doi: 10 3969/j issn. 1001- 5094. 2009. 11. 001

生物信息学方法在判断 DNA结合蛋白质和预测结合位点中的应用

刘子朋[1], 章宏九[2], 李雅晴[2], 朱云蛟[2], 胡 健[1], 方慧生[1*]

(1. 中国药科大学生命科学与技术学院, 江苏 南京 210009; 2 中国药科大学药学院, 江苏 南京 210009)

[摘要] 综述生物信息学方法在判断 DNA结合蛋白质和预测结合位点中的应用研究进展。蛋白质与 DNA间的相互作用是基因表达调控的分子生物学基础, 因此 DNA结合蛋白的判断以及 DNA与蛋白质间作用位点的预测一直以来都是分子生物学和生物信息学的前沿领域。采用生物信息学方法进行这类判断和预测, 具有省时、省力的特点, 近年来吸引了众多科学家的关注。

[关键词] 生物信息学; 蛋白质–DNA 相互作用; DNA结合蛋白; 结合位点预测
[中图分类号] Q 513 [文献标识码] A [文章编号] 1001- 5094(2009) 11- 0486- 05

The Application of Bioinformatical Method in Distinguishing the DNA-binding Protein and Predicting the Binding Site in Protein-DNA Interface

LIU Zi-peng[1], ZHANG Hong-jiu[2], LI Ya-qing[2], ZHU Yun-jiao[2],
HU Jian[1], FANG Hui-sheng[1]

(1. School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China; 2. School of Pharmacy, China Pharmaceutical University, Nanjing 210009, China)

[Abstract] The research progress in application of bioinformatical method in distinguishing the DNA-binding protein from the other proteins and predicting the binding sites between DNA and protein was reviewed. Genetic regulation is completed by protein-DNA interaction and therefore how to distinguish DNA-binding protein from the other proteins and further predict the binding sites between protein and DNA is always the frontier field in the molecular biology and bioinformatics. In this field, bioinformatical method which has certain features such as time-saving and labour-saving has attracted many scientists.

[Key words] bioinformatics; protein-DNA interaction; DNA-binding protein; prediction of binding site

[接受日期] 2009-06-02

* 通讯作者: 方慧生, 教授;
研究方向: 计算生物学, 生物信息学, 虚拟生命科学;
Tel: 025-83271001; E-mail: hsfang889@ 163. com

近年来,随着生命科学的不断发展,生物技术的长足进步,尤其是人类基因组计划的启动,使得生物学相关信息在量(海量特征)和质(复杂特征)方面都发生了革命性的巨变,人们对海量生物信息进行高效处理的需求也更加迫切。与此同时,计算机技术的发展也使之具备了处理海量生物信息的能力。生物信息学便是在综合计算生物学的研究和生物学信息的计算机处理的基础上成功发展起来的一门新兴学科[1]。它通过对生物学实验数据进行提取、加工、存储、检索和分析,进而获得其中所蕴含的生物学意义。

蛋白质和核酸是生命的物质基础,二者的相互作用广泛存在于生物体内各种生命活动中,如在DNA复制过程中,除链的引发、延伸、终止所涉及的反应都由相应的酶催化外,还需许多具有调节功能的蛋白质对DNA复制进行调节[2]。可见,蛋白质与核酸的相互作用是分子生物学研究的中心问题之一,是许多生命活动的重要组成部分。故近年来,关于蛋白质–核酸相互作用的研究逐渐受到重视,且已涌现出一系列新的研究方法,如核酸适体技术、生物芯片技术、纳米技术及生物信息学方法[3]。

当前,在分子生物学和信息科学快速发展的影响下,生物信息学在生物研究领域中具有指导性作用。与其他方法相比,利用生物信息学方法可大大缩短研究蛋白质–DNA相互作用所需的时间,达到事半功倍的效果。其主要的研究领域有:如何在众多的蛋白质中确定DNA结合蛋白及如何预测蛋白质与DNA相互作用的结合位点。本文就生物信息学方法在上述领域中的最新研究进展作一综述。

1 DNA结合蛋白的判定

与DNA结合的蛋白质在维持基因信息和参与一些生命活动(如DNA的受控转录、复制、修复、折叠和重组等)中有着重要的作用[4]。因此,如何从众多的蛋白质中判定DNA结合蛋白有着非常重要的意义。

为此,人们提出了一系列的预测方法,主要有“基于3D结构的预测方法”和“基于蛋白质电学特性的预测方法”两类。第一类方法是利用已知3D结构的蛋白质推断与其序列相似的未知蛋白质的空间结构。尽管人们对远距离同源蛋白质(即氨基酸序列相似性小于35%)的预测已有初步的解决方案[5],但目前最常用且准确性较高的仍是近距离同源蛋白质(即氨基酸序列相似性大于35%)的预测方法。由于与相应的已知序列的蛋白质相比,已测定三维构象的蛋白质数量较少,故基于3D结构的预测方法应用范围较窄。此处着重介绍基于蛋白质电学特性的预测方法。

该方法是基于“蛋白质残基上所带的电荷与DNA中碱基上的电荷的相互作用可能是蛋白质与DNA相结合的主要作用力之一”这一基本事实而构建的。例如,Stawiski等[6]提出了一种基于人工神经网络的预测方法,其步骤如下:首先从高分辨的蛋白原子结构中提取12个参数作为神经网络的输入层,对作用界面的缝隙处进行分析,并利用可描述处在离子溶液中的分子之间相互静电作用,以模拟不同离子浓度溶液中溶剂对蛋白质结构产生的影响的微分方程——Poisson–Boltzmann方程对作用界面上的正电荷进行修正;再用比对不同蛋白质的氨基酸序列或不同基因的DNA序列时常用的算法——PSI-BLAST搜索完成12个参数中3个参数的序列保守性分析(美国国立生物技术信息中心提供此项服务),在此基础上搭建一个三层人工神经网络(其中一个隐含层有3个单元)。结果表明:该法预测的准确度可达81%,其不足是会忽略很多电学属性[7]。

Ahmad等[8]则设计了一种很简便的预测方法,他们在大量的静电特性基础上构建了一个没有隐含层的双层人工神经网络,应用其预测某个蛋白质是否为DNA结合蛋白。他们测定了62种具有代表性的已知结构的DNA结合蛋白的78种氨基酸序列,计算各种氨基酸序列的静电荷、电偶极矩和四极矩张量,结果发现DNA结合蛋白中三者的值均明显高于非DNA结合蛋白。若只作单变量比较,即分别运用净电荷、净偶极矩和净四极矩作为变量进行预测,准确度分别为82.6%、77.4%和73.7%;若同时运用三者进行预测,则不进行交叉验证的预测准确度最大为85.6%,进行交叉验证的为83.9%,与其他能达到相同准确度的预测方法相比,该预测方法要简单的多。

而有另一种理论认为:蛋白质能否与带负电的DNA结合,取决于蛋白质中所含的带正电的残基簇[9],依据有两个:其一是蛋白质的双链DNA结合位点的物理特性:这种双链DNA结合位点包含正电性原子,可与DNA产生点电荷间、偶极矩、四极矩和

氢键的相互作用[10] (Ohlendorf等, *Adv Biophys*, 1985年)。如果不能与对应的 DNA 中电负性原子结合并产生稳定的静电作用, 这些正电性原子会处于静电不稳定状态[11], 因此, 用一些电负性的氨基酸 (如 Asp/Glu)替换处于静电不稳定状态的极性或碱性氨基酸, 可得到更稳定的结合蛋白[12]; 其二则源于进化理论: 能与 DNA 作用的氨基酸残基及其附近的残基会形成具有一定空间结构的残基簇, 这种残基簇在 DNA 结合蛋白家族里是高度保守的[13-15]。所以该理论的具体实现过程包括两步: 1)输入蛋白结构, 对蛋白质的 Asp/Glu突变体的表面静电稳定性进行分析; 2)找到相应的足够数量的、序列上同源的蛋白结构以确定氨基酸残基的保守性。Chen等[16]运用该原理, 同时考虑 DNA 结合蛋白质的电学性质和残基的保守性两个特性时, 其预测准确度高达 83%, 而分别只考虑其中一个特性时的准确度为 53%和 50%。

2 结合位点的识别

DNA 与蛋白质间的相互作用在基因调节过程中起着重要作用, 所以对其结合位点的准确预测对于理解复杂多样的生物过程具有重要意义。

2.1 转录因子结合位点的预测

转录过程既是 DNA 翻译成蛋白质的关键步骤, 同时也是调控基因表达的关键阶段。转录调控通常是在转录起始步骤实现的。除启动子外, 在几乎所有基因的上游区域中都还存在着激活基因所需的一段特定的 DNA 序列 (转录因子结合位点, TFBS)。这种序列本身并不执行任何功能, 只有当其被调控蛋白 (即转录因子)识别、结合后才能发挥作用。TFBS和转录因子共同控制着基因的转录, 二者的结合具有高度的专一性。

由一组已知的 TFBS 就可设计一个模序的模型, 以预测其他结合相同转录因子的位点。这项工作一般是通过位置特异得分矩阵 (PSSM)完成的。近年来该方法已广泛应用于许多不同功能和种类基因 (如球蛋白基因和细胞周期依赖性基因)的调节区域的分析, 但预测结合位点的灵敏度不高。Naughton等[17]分析了大量真核模序中 *k*-mers[DNA 序列中的简单重复片段, 如 $(A)_n$, $(CA)_n$ 或者 $(CGG)_n$ 等]的分布, 寻找不易被统计学模拟的复杂依赖性, 指出 *k*-mers分布的内在复杂性是由 DNA 功能和进化的双重影响造成的。当决定一个候选的 *k*-mer是不是模序一部分的时候, Naughton等更关注它与模序中其他已知的、具有特异性的 *k*-mer的相似度, 而不是其能否完美地匹配一个模序模型。基于这点, 他们开发出了一个新的、基于图论的名为 MotifScan的算法来测定 TFBS, 并以目前已知的一些真核生物模序为样本, 对该算法的预测准确度进行验证, 同时与基于 PSSM 的主流预测方法进行了比较。结果表明: 该算法在真核模序预测方面较后者有很大改进, 预测准确度至少提高了 5%, 对某些模序的预测准确度甚至提高了 50%以上。

可变阶数马尔科夫模型 (VOM)和可变阶数的贝叶斯树 (VOBAT)在 TFBS识别方面已被证明比传统的位置权重矩阵 (PWM)、马尔科夫模型和贝叶斯树要好[18-19]。基于前两个模型, Grau等[7]提供了一项名为 VOMBAT 的网络服务 (网址为: https://www2.informatik.uni-halle.de:8443/VOMBAT/faces/pages/choose.jsp)。该网站可提供以下功能: 1)给定标明结合位点的数据集和基因组序列, 可进行 VOM 和 VOBAT的训练; 2)给定一组经过训练的模型, 可预测基因组序列中可能存在的 DNA 结合位点; 3)给定标明结合位点的数据集和基因组序列的数据集, 可对不同的模型组合进行交叉验证的试验。

值得一提的是, 目前虽已有对 TFBS较为准确的预测方法, 但基因的调节是以多个转录因子的特异性组合的形式进行的, 而非单个因子单独的作用, 故必须考虑复合模序的预测问题。Waleev等[20]提供了一项名为 CMA的网络服务, 帮助识别启动子-增强子, 其网址为 http://www.gene-regulation.com/pub/programs/cma/CMA.html。该软件基于 TFBS的组成和配对, 利用由 TRANSFAC 数据库收集的 PWM 文库, 可用于确定较长调节区域中复杂的复合模序。

2.2 与 DNA 结合的氨基酸残基的预测

由蛋白质结构分析所得的信息已能用于预测结构已知的蛋白中与 DNA结合的残基, 但目前尚不能直接采用氨基酸序列来预测。Wang等[21]提供了一项名为 BindN 的网络服务 (网址为: http://bioinfo.ggc.org/bindn/), 试图解决这个问题。使用该工具时可先输入氨基酸序列, 然后利用支持向量机

(support vector machines, SVM), 通过侧链 pKa值、疏水指数和氨基酸的分子质量这 3个特征值来预测可能的结合位点的残基。若以 TP、TN、FP、FN 分别代表被正确预测的 DNA 结合残基数、被正确预测的非 DNA 结合残基数、被误测为 DNA 结合残基的非 DNA 结合残基数以及被误测为非 DNA 结合残基的 DNA 结合残基数, 则可用 TP/(TP+ FN)和 TN/(TN + FP) 计算预测的灵敏度和专一性。结果显示, 氨基酸链中 DNA 和 RNA 结合位点的预测灵敏度分别为 69.40% 和 66.28%, 专一性分别为 70.47% 和 69.84%。

锌指蛋白模型的设计对基础学科和应用学科而言是一项很有前景的技术。锌指结构域起到模块单位的作用, 最开始是识别 DNA 3个碱基的位点。C2H2型的锌指由大约 30个氨基酸组成, 形成两条β链和一个 α螺旋, 可与特异的 DNA 结合。锌指具有能与 DNA 结合的结构域, 该结构域又可嵌入到各种影响 DNA 生化特性的因子(如转录激活因子和抑制因子, 核酸酶和整合酶)的结构中。锌指设计的目标是获得具有高度特异性和亲和力的、能识别所有 64种 DNA 三联体的锌指结构域。目前已运用噬菌体库、理性设计和自然发生结构域获得了可特异性识别一些三联体的锌指。Mandell等[22]证明, 把这些锌指结构域嵌入到蛋白质中时, 产生的蛋白可识别 9~18个碱基对的 DNA 序列, 并具有高度特异性和亲和力。他们开发了名为 ZF Tools的工具并公布在互联网上(http://www.zincfingertools.org), 可根据用户提供的 DNA 序列寻找合适的基因调节或核酸酶的作用位点。

2.3 DNA-蛋白质作用的模型设计

DNA-蛋白质作用的模型设计一般涉及作用处的能量和空间构象方面的研究。DNA-蛋白质复合物的结构复杂多样, Luscombe等[23]曾将目前已知的复合物分为 8大类, 共 54个家族。因此, 很难找到具有普遍适用性的模型。

通常预测和设计作用模型的方法是研究出一种氨基酸和碱基间的识别编码。该方法虽简单, 但存在一个明显的不足, 即设计出的模型具有专属性, 只适用于一个蛋白家族。另一种分析 DNA-蛋白质作用的方法是采用简单的大分子动力学模型以及对蛋白质侧链构象进行取样的快速算法, 得到高质量的 DNA-蛋白质作用模型。在此基础上, Havranek等[24]根据已确定的 DNA 和蛋白质主链构象以及基于旋转异构体库的侧链构象信息, 用一个简单的能量方程设计了一个 DNA-蛋白质间作用的模型。该模型可成功预测已知复合物中 DNA 与蛋白质的结合情况。

另外, 一种建立在对 DNA-蛋白质复合物结构进行统计分析的基础之上的方法, 即基于知识的方法已成功应用于计算 DNA-蛋白质相互作用过程中的能量和特异性。Ahmad等[25]就利用该方法做成一项名为 ReadOut的在线网络服务(网址为: http://gibk26.bse.kyutech.ac.jp/jouhou/readout/), 用户只需输入待预测蛋白质的坐标文件(须为 PDB格式, 可在 PDB数据库网站上获得该格式的具体内容)或 DNA-蛋白质复合物在 PDB数据库中的代码(该代码也可在 PDB数据库中查到, 如糖皮质激素受体突变体-DNA 复合物的 PDB代码是 1lat), 就可得到总能量的计分 Z[($Z=(x-\mu)/\sigma$, 其中 x是特定 DNA 序列的能量, μ是随机选取的 DNA 序列能量的平均值, σ是随机选取的序列能量的标准偏差]。该计分可用于评估 DNA-蛋白质作用的特异性程度, 其值越负, 表示 DNA 与蛋白质结合越好。该项网络服务也可用于检查 DNA-蛋白质复合物结构以及设计蛋白质和目的 DNA。

3 结语

全面认识蛋白质和核酸相互作用的规律是生命科学研究的核心工作之一。后基因组时代不断增加的大量相关信息, 使这项工作的进行需要生物学、计算机科学及应用数学等多学科的协作。生物信息学则很好地符合了这个需求, 它不仅可利用现代网络与计算机数据库技术构建精准模型, 同时还能大大提高工作效率, 已在 DNA-蛋白质相互作用的研究中充分体现出其优越性, 但仍存在许多问题, 如分析软件的输出结果存在较大的偏差, 模型不具备普遍性等, 仍需通过生物学试验对理论预测结果进一步验证。另外, 目前研究多集中于转录过程, 较少有对复制和翻译过程中蛋白质与核酸相互作用的研究, 而肿瘤和病毒性疾病的防治恰恰与复制及翻译过程中蛋白质和核酸的相互作用有较大关系。笔者所在课题组根据已有的研究基础, 应用隐马尔科夫链方法

(HMM)预测蛋白质与 DNA相互作用位点,同时准备在分子动力学基础上进行更深入的研究。希望在不久的将来,能比较准确快速地确定蛋白质与 DNA相互作用位点及相应的作用机制。

[参考文献]

[1] 赵国屏. 生物信息学[M]. 北京: 科学出版社, 2001: 2-20.

[2] 王恩多. 蛋白质与核酸的相互作用[J]. 生物科学信息, 1991, 3(5): 1-4.

[3] 王成刚, 莫志宏. 蛋白质/核酸相互作用研究方法进展[J]. 生命科学, 2006, 18(2): 195-198.

[4] Szilgyi A, Skolnick J. Efficient prediction of nucleic acid binding function from low-resolution protein structures[J]. *J Mol Biol*, 2006, 358(3): 922-933.

[5] 方慧生, 陈凯先. 国际性难题暨远距离同源蛋白建模及比对解决的基本策略[C]//中国化学会第 26届学术年会化学信息学与化学计量学分会场论文集. 天津: [出版社不详], 2008.

[6] Stawiski E W, Gregoret L M, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure[J]. *J Mol Biol*, 2003, 326(4): 1065-1079.

[7] Grau J, Ben-Gal I, Posch S, *et al*. VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees[J]. *Nucleic Acids Res*, 2006, 34(Web Server issue): W529-W533.

[8] Ahmad S, Sarai A. Moment-based prediction of DNA-binding proteins[J]. *J Mol Biol*, 2004, 341(1): 65-71.

[9] Wintjens R, Lievin J, Rooman M, *et al*. Contribution of cation-π interactions to the stability of protein-DNA complexes[J]. *J Mol Biol*, 2000, 302(2): 395-410.

[10] Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites[J]. *Nucleic Acids Res*, 1998, 26(10): 2306-2312.

[11] Jones S, van Heyningen P, Berman H M, *et al*. protein-DNA interactions: a structural analysis[J]. *J Mol Biol*, 1999, 287(5): 877-896.

[12] Elcock A H. Prediction of functionally important residues based solely on the computed energetics of protein structure[J]. *J Mol Biol*, 2001, 312(4): 885-896.

[13] Luscombe N M, Thornton J M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity[J]. *J Mol Biol*, 2002, 320(5): 991-1009.

[14] Mimy L A, Gelfand M S. Structural analysis of conserved base pairs in protein-DNA complexes[J]. *Nucleic Acids Res*, 2002, 30(7): 1704-1711.

[15] Sathyapriya R, Vishveshwara S. Interaction of DNA with clusters of amino acids in proteins[J]. *Nucleic Acids Res*, 2004, 32(14): 4109-4118.

[16] Chen Y C, Wu C Y, Lin C. Predicting DNA-binding amino acid residues from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation[J]. *Proteins*, 2007, 67(3): 671-680.

[17] Naughton B T, Fratkin E, Batzoglou S, *et al*. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites[J]. *Nucleic Acids Res*, 2006, 34(20): 5730-5739.

[18] Shmilovici A, Ben-Gal I. Using a VOM model for reconstructing potential coding regions in EST sequences[J]. *Comput Stat*, 2007, 22(1): 49-69.

[19] Ben-Gal I, Hani A, Gohr A, *et al*. Identification of transcription factor binding sites with variable-order bayesian networks[J]. *Bioinformatics*, 2005, 21(11): 2657-2666.

[20] Waleev T, Shtokalo D, Konovalova T, *et al*. Composite module analyst: identification of transcription factor binding site combinations using genetic algorithm[J]. *Nucleic Acids Res*, 2006, 34(Web Server issue): W541-W545.

[21] Wang L J, Brown S J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences[J]. *Nucleic Acids Res*, 2006, 34(Web Server issue): W243-W248.

[22] Mandell J G, Barbas C F. Zinc finger tools: custom DNA-binding domains for transcription factors and nucleases[J]. *Nucleic Acids Res*, 2006, 34(Web Server issue): W516-W523.

[23] Luscombe N M, Austin S E, Berman H M, *et al*. An overview of the structures of protein-DNA complexes[J]. *Genome Biol*, 2000, 1(1): 1-37.

[24] Havranek J J, Duarte C M, Baker D. A simple physical model for the prediction and design of protein-DNA interactions[J]. *J Mol Biol*, 2004, 344(1): 59-70.

[25] Ahmad S, Kono H, Araz-Bravo M J, *et al*. ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition[J]. *Nucleic Acids Res*, 2006, 34(Web Server issue): W124-W127.