·生物信息学·

基于进化足迹法结合位置打分函数的转录因子结合位点预测*

王世缘 王艳秋△

(哈尔滨医科大学生物信息科学与技术学院 黑龙江 哈尔滨 150000)

摘要 目的:预测转录因子结合位点。方法:本文从 ABS 数据库上下载了人类和啮齿类动物的直系同源的启动子序列,首先使用位置打分函数对这些启动子序列进行打分,找到候选的转录因子结合位点,并进一步利用进化足迹法对这些候选的转录因子结合位点进行筛选,只有结合位点同时出现在人类和啮齿类动物的启动子序列才认为是真正的结合位点。结果:在对人类和啮齿类动物进行转录因子结合位点预测时,与单独使用打分函数的结果相比,进化足迹法与打分函数结合的方法有效的提高了预测结果的性能,大幅度提供所得预测结果的特异性。结论:进化足迹法结合位置打分矩阵的方法能较为准确有效的预测转录因子结合位点。

关键词:进化足迹法;位置打分函数;特异性

中图分类号:Q753,Q-33,Q61 **文献标识码**:A **文章编号**:1673-6273(2012)19-3725-03

Prediction of Transcription Factor Binding Sites by Phylogenetic Footprinting Combined with Position Score Function*

WANG Shi-yuan, WANG Yan-qiu△

(College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang,150000, China)

ABSTRACT Objective: Reliable prediction of transcription factor binding sites can help to identify the target genes of transcription factors and infer the relationship between the binding sites. But the performance of some algorithms has been unreliable with respect to poor specificity, more efficient algorithms are required. Methods: In this paper, the human and rodent orthologous promoters sequences are download from the ABS database, position scoring function is used to predict the transcription factor binding sites, phylogenetic footprinting is to construct alignment of the orthologous promoter sequences, and only the binding site appeared in both the promoters of human and rodent are selected. Phylogenetic footprinting are used to reduce the number of predicted binding sites. Results: The predictive results show that compared with position scoring function, the prediction result are improved highly, and the specificity is also improved significantly. Conclusion: The algorithm presented in this study could be provided some help for the prediction of transcription factor binding sites.

Key words: Phylogenetic footprinting; Position scoring function; Specificity

Chinese Library Classification: Q753,Q-33, Q61 Document code: A

Article ID: 1673-6273(2012)19-3725-03

前言

基因的转录被转录因子结合蛋白所调控。转录因子和它们所对应的结合位点之间的相互作用控制着许多重要的生命进程,例如,对环境压力应答的控制[1]。为了全面的理解这些相互作用,我们不仅需要知道一条序列上出现了哪些转录因子结合位点,还需要知道哪些转录因子结合到这些位点。所以,在 DNA 序列上找到结合位点区域是分析基因调控的一个重要的任务[2]。传统上,生物学家仅依靠实验的方法来寻找转录因子结合位点。随着更多的基因序列和表达数据的出现,我们需要应用计算机辅助的方法来识别更多的转录因子结合位点。通常,我们用共有序列来表示已知的转录因子结合位点[3]。但共有序列的概念存在一些缺陷,所以后来人们又提出了用位置权重矩阵的方法来表示转录因子结合位点。TRANSFAC[4]就是基于位置权重矩阵的数据库,该数据库包含一些识别潜在转录因子结合位点的程序,例如 MATCH[5],P-Match[6]和 MatInspector[7]。虽然位置权重矩阵相对于共有序列在预测转录因子结合位点的敏感性上有所提高,但它所预测的转录因子结合位点只有很小一部分是真实的。所以,我们需要更为准确的预测算法。

最近几年,进化足迹法[8,9]成为了预测转录因子结合位点的有效方法。进化足迹法基于的理论是启动子中的功能位点相对于其它区域进化的更加缓慢,因为这些区域具有保守的功能。

* 基金项目:黑龙江省卫生厅项目(2010-213)

作者简介:王世缘(1984),女,硕士,助教,主要研究方向:转录因子结合位点的识别。

E-mail: yuki.net.5@163.com

△通讯作者:王艳秋,电话:0451-86669617,E-mail: wangyq@ems.hrbmu.edu.cn

(收稿日期:2012-03-09 接受日期:2012-04-06)

因此，在进化保守的启动子区域所发现的转录因子结合位点相对于其它区域中发现的更为可靠。虽然进化足迹法在预测转录因子结合位点方面取得了一系列进步，但它仍有很多不足之处，比如，选取何种物种进行比较，利用何种比对算法，选择何种参数鉴别保守区域等等。针对这些问题，在本文中，我们利用进化足迹法来预测转录因子结合位点，并从 ABS 数据库[10]上下载直系同源的启动子区域，利用 CLUSTALW 程序对这些区域进行了比对，之后利用位置打分函数来搜寻实验上已知的结合位点，预测结果显示，我们得到了更好的结果。

1 材料与方法

1.1 材料

我们首先需要选择适当的生物序列去进行序列比对。人类和啮齿类动物大约在 5-7 千万年前分离，它们基因组之间的分离距离既足够小，以致于可以进行直系同源的序列比对，又足够大以致于可以发现功能保守的区域[11]。因此在本文中，我们从 ABS 数据库下载了人类和啮齿类动物的直系同源的启动子区域。

1.2 构建位置打分函数

我们下载 ABS 数据库和 TRANSFAC 数据库中所共有的位置权重矩阵。这些矩阵的编号和对应的转录因子的名称见表 1。

表 1 转录因子名称和矩阵编号

Table 1 The names of transcription factor and the correspondence matrices

Transcription factor	Matrices
AP-1	M00172, M00173, M00174, M00188, M00199, M00517
CREB	M00039, M00113, M00177, M00178
MYOD	M00001, M00184
SRF	M00152, M00186, M00215
TBP	M00471
USF	M00121, M00122, M00187, M00217

根据 Stormo[12]和其他人[13,14]所提出的矩阵理论，位置权重矩阵的估计概率 P(b,j)定义如下：

$$P(b,j)=\frac{f(b,j)+s(b)}{N+\sum s(b)}$$

其中 j 表示矩阵的第 j 列；b 是四个核酸之一；f(b,j) 是在矩阵第 j 列上的 b 的个数；s 为伪计数；N 是全部序列的个数。这里伪计数取 $s=(N)^{1/2}/4$。

接下来计算矩阵中不同列上结合位点序列的保守性。我们结合以前的工作及文献上的资料[15]，对矩阵第 j 列的位点保守性指数 M (j)定义如下：

$$M(j)=\sum_{k_i}\frac{(P(b,j)-1/4)^2}{1/4}$$

其中 p' (b, j) 是第 b 个碱基在矩阵中第 j 个位点出现的频率。上式是 M(j)随位点 j 的变化关系，它表示序列在该位点与完全随机的序列在该位点的偏离程度。可以看出，如果位点保守性越强，则 M (j)的值越大，因此，我们可以用 M(j)表征序列在该位点的保守性。

使用位置打分函数进行预测时，需要对给定的片断进行相似度打分，结合文献上的资料[5]，我们定义了如下的位置打分函数：

$$S=\frac{\sum_{j=1}^{n}(m(j)P(b,j)-M(j)MinP(j))}{\sum_{j=1}^{n}(M(j)MaxP(j)-M(j)MinP(j))}$$

其中 MinP (j), MaxP(j)分别为矩阵第 j 列中的最小概率及最大概率值；n 为矩阵的列数。

1.3 序列比对程序

进化足迹法的标准方法是构建直系同源启动子序列的比对，并以此识别比对中的保守区域[16]。在本文中，我们选用了 CLUSTALW 程序来比对人类和啮齿类的直系同源的启动子序列。

1.4 PI 值的定义

预测出的结合位点的保守性利用比对得到的序列一致性来计算。转录因子结合位点的一致性百分比(PI)值定义为长度为 m 的第 j 列中一致序列的比例，m 为预测的结合位点的长度：

$$PI=\frac{1}{m}\sum_{j=1}^{m}a_j$$

如果 $X_j=Y_j$，即所对应碱基相同，则 $a_j=1$；如果 $X_j\neq Y_j$，则 a=0, 即所对应碱基不同。$X_j, Y_j\in\{A,C,G,T,-\}$为长度为 m 的比对中第 j 个位置的核酸或空位。PI 值越大，表明在同一个转录因子结合位点上，人类和啮齿类动物序列比对所重合的区域越多。通常，PI 值如果大于或等于 0.5，则认为预测的结合位点是保守的。如果在两条或三条同源的启动子区域，一个转录因子所预测得到的结合位点与真实的结合位点重叠至少一半，那么就认为这个预测是正确的。所以，如果 PI 值大于或等于 0.5，此结合位点就认为预测正确。

2 结果与分析

在这里我们分析了位置打分函数在 0.75-0.95 的阈值之下所得到的结果(表 2)。

从上表中可以看出，进化足迹法能够减少预测所得的结合位点的数目。一些预测出来的实验上所证实的结合位点可能被进化足迹法所排除，但我们应用进化足迹法却能大幅度提高所得结合的准确率。通过使用进化足迹法结合位置打分函数，我

表 2 预测的与真实的结合位点数

Table 2 The number of predicted binding sites

Cutoff	Position scoring function		Position scoring function combined with phylogenetic footprinting	
	The number of predicted binding sites	The number of corrected predicted binding sites	The number of predicted binding sites	The number of corrected predicted binding sites
0.75	1582	311	743	284
0.80	946	282	459	251
0.85	594	249	304	218
0.90	363	214	214	177
0.95	140	104	88	79

们所得到的转录因子结合位点涵盖了大部分实验上所证实的结合位点。而预测出的没有被实验证实的结合位点也有可能是真实的结合位点。

本文使用敏感性和特异性去评价预测结果的好坏。敏感性为预测正确的结合位点的数目与实验上已知的结合位点数目之比；特异性为正确预测的结合位点数目与总的预测数目之比。在不同阈值之下所得的敏感性和特异性见表 3。

表 3 不同阈值之下的敏感性和特异性

Table 3 The sensitivity (Sn) and specificity (Sp) under different cutoffs

Cutoff	Position scoring function		Position scoring function combined with phylogenetic footprinting	
	Specificity	Sensitivity	Specificity	Sensitivity
0.75	0.197	0.818	0.382	0.747
0.80	0.298	0.742	0.547	0.661
0.85	0.419	0.655	0.717	0.574
0.90	0.589	0.563	0.827	0.466
0.95	0.743	0.274	0.898	0.232

由上表可知，进化足迹法结合位置打分函数大幅度提高了预测结果的特异性。一般来说，进化足迹法结合位置打分函数与单独使用位置打分函数相比，预测结果的特异性性提高了 22.5%，但敏感性只下降了 7.4%。在 0.75%的阈值之下，进化足迹法结合位置打分函数能够发现 74.7%的实验上已知的结合位点。

3 结论

随着基因组测序工作的完成，有关转录因子结合位点的数据越来越多，针对转录因子结合位点预测的算法也在逐年增多。但目前预测转录因子结合位点的算法所得结果的特异性普遍较低，因此有必要提出一种新的能够准确预测转录因子结合位点的算法。在本研究中，我们利用进化足迹法结合位置打分函数的方法来预测转录因子结合位点。与普通的位置打分函数的方法相比，本研究中的方法由于考虑了多个同源物种的相关启动子序列信息和进化保守性信息，因此所得的预测结果更为准确。所以，此方法可以用于对转录因子结合位点的可靠预测。在将来的研究工作中，我们准备加入更多的有效特征信息以提高预测的准确率，并将我们的方法扩展到其它的直系同源的哺乳动物的转录因子结合位点序列的识别上去。

参 考 文 献(References)

[1] Bulyk ML. Computational prediction of transcription-factor binding site location [J]. Genome Biol.,2003,5:201

[2] Reid JE, Ott S, Wernisch L. Transcriptional programs: modelling higher order structure in transcriptional control [J]. BMC Bioinformatics,2009,10:218

[3] Liu LA, Bader JS. Structure-based ab initio prediction of transcription factor-binding sites [J]. Methods Mol Biol,2009,541:23-41

[4] Wingender E, Chen X, Hehl R, et al. Transfac: an integrated sy- stem for gene expression regulation [J]. Nucleic Acids Res,2000,28:316-319

[5] Kel A E, Gossling E, Reuter I, et al. Matchtm: a tool for searching transcription factor binding sites in DNA sequences [J]. Nucleic Acids Res,2003,31:3576-3579

[6] Chekmenev D S, Haid C, Kel A E. P-Match: transcription factor binding site search by combining patterns and weight matrices [J]. Nucleic Acids Res,2005,33:432-437

[7] Cartharius K, Frech K, Grote K, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites [J]. Bioinformatics,2005,21:2933-2942

[8] Wu J, Sieglaff DH, Gervin J, et al. Discovering regulatory motifs in the Plasmodium genome using comparative genomics [J]. Bioinformatics,2008,24(17):1843-1849

[9] Bergman CM, Quesneville H. Discovering and detecting transposable elements in genome sequences [J]. Brief Bioinform, 2007,8(6):382-392

（下转第 3793 页）

osome position map of C. elegans reveals a lack of universal sequence-dictated positioning[J]. Genome Res,2008,18(7):1051-1063

[4] Thomas R K, Nickerson E, Simons J F, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing[J]. Nat Med,2006,12(7):852-855

[5] Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual[J]. Nature,2008,456(7218):60-65

[6] Shen Y, Sarin S, Liu Y, et al. Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing[J]. PLoS One,2008,3(12):4012

[7] Rothberg J M, Hinz W, Rearick T M, et al. An integrated semiconductor device enabling non-optical genome sequencing [J]. Nature,2011,475(7356):348-352

[8] Howden B P, McEvoy C R, Allen D L, et al. Evolution of multidrug resistance during Staphylococcus aureus infection involves mutation of the essential two component regulator WalKR [J]. PLoS Pathog, 2011,7(11):e1002359

[9] Rohde H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4 [J]. N Engl J Med,2011,365 (8):718-724

[10] Miller W, Hayes V M, Ratan A, et al. Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil)[J]. Proc Natl Acad Sci USA,2011,108(30): 12348-12353

[11] Harris T D, Buzby P R, Babcock H, et al. Single-molecule DNA sequencing of a viral genome[J]. Science, 2008,320(5872):106-109

[12] Bowers J, Mitchell J, Beer E, et al. Virtual terminator nucleotides for next-generation DNA sequencing [J]. Nat Methods,2009,6(8):593-595

[13] Pastor W A, Pape U J, Huang Y, et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells [J]. Nature,2011, 473(7347):394-397

[14] Goren A, Ozsolak F, Shoresh N, et al. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA [J]. Nat Methods,2010,7(1):47-49

[15] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules[J]. Science, 2009, 323(5910): 133-138

[16] Flusberg B A, Webster D R, Lee J H, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing [J]. Nat Methods, 2010, 7(6): 461-465

[17] Uemura S, Aitken C E, Korlach J, et al. Real-time tRNA transit on single translating ribosomes at codon resolution[J]. Nature, 2010, 464 (7291):1012-1017

[18] Grad Y H, Lipsitch M, Feldgarden M, et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011 [J]. Proc Natl Acad Sci U S A,2012,109(8):3065-3070

[19] Clarke J, Wu H C, Jayasinghe L, et al. Continuous base identification for single-molecule nanopore DNA sequencing [J]. Nat Nanotechnol, 2009,4(4):265-270

[20] Stoddart D, Heron A J, Mikhailova E, et al. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore[J]. Proc Natl Acad Sci U S A,2009,106(19):7702-7707

[21] Branton D, Deamer D W, Marziali A, et al. The potential and challenges of nanopore sequencing [J]. Nat Biotechnol,2008,26(10):1146-1153

[22] Sims P A, Greenleaf W J, Duan H, et al. Fluorogenic DNA sequencing in PDMS microreactors[J]. Nat Methods,2011,8(7):575-580

[23] Susan H, James B, Richard W. Methods for real-time single molecule sequence determination[P]. US patent 7329492,2008

[24] Postma H W. Rapid sequencing of individual DNA molecules in graphene nanogaps[J]. Nano Lett,2010,10(2):420-425

[25] Lagerqvist J, Zwolak M, Di Ventra M. Fast DNA sequencing via transverse electronic transport[J]. Nano Lett,2006,6(4):779-782

[26] Tanaka H, Kawai T. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope [J]. Nat Nanotechnol,2009,4 (8):518-522

（上接第 3727 页）

[10] Blanco E, Farre D, Alba M M, et al. ABS: a database of annotated regulatory binding sites from orthologous promoters [J]. Nucleic Acids Res,2006,34:D63-D67

[11] He X, Ling X, Sinha S. Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution [J]. PLoS Comput Biol,2009,5(3):1000299

[12] Stormo G D. DNA binding sites: representation and discovery [J]. Bioinformatics,2000,16:16-23

[13] PairoE, Maynou J, Marco S, et al. A subspace method for the detection of transcription factor binding sites [J]. Bioinformatics, 2012[Epub ahead of print]

[14] Qian J, Ferguson TM, Shinde DN, et al. Sequence dependence of isothermal DNA amplification via EXPAR [J]. Nucleic Acids Res, 2012,40(11):e87

[15] Li Q Z, Lin H. The recognition and prediction of σ 70 promoters in Escherichia coli K-12 [J]. J. Theor. Biol,2006,242:135-141