PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(7): 573~579
www.pibb.ac.cn

C2H2 型锌指蛋白结合的 DNA 序列预测方法的研究进展 *

沈 磐 杨 冬** 贺福初
(军事医学科学院放射与辐射医学研究所，国家蛋白质科学中心(北京)，
北京蛋白质组研究中心，蛋白质组学国家重点实验室，北京 102206)

摘要 C2H2 型锌指蛋白是哺乳动物中数量最多的一类转录调控因子．C2H2 型锌指蛋白中含有的 C2H2 型锌指基序多是不相同的，表明它们很可能结合不同的 DNA 序列，从而调控不同的基因，行使多样化的调控功能．然而，目前大多数 C2H2 型锌指蛋白结合的 DNA 序列仍不明确，这阻碍了 C2H2 型锌指蛋白的功能研究．目前，针对 C2H2 型锌指蛋白的靶序列预测已有一些初步的研究．本文介绍了 C2H2 型锌指基序与 DNA 结合的经典模式，并对 C2H2 型锌指蛋白靶序列预测方法中所用到的算法、训练集、金标准数据集及相应工具进行了全面系统的总结归纳，旨在丰富对 C2H2 型锌指蛋白靶序列预测原理和工具的认识，为 C2H2 型锌指蛋白靶序列的精确预测和更深入的功能研究打下基础．

关键词 转录因子，C2H2 型锌指蛋白，靶序列预测
学科分类号 Q51
DOI: 10.16476/j.pibb.2017.0047

基因表达和调控的研究对于揭示生命奥秘、了解生命本质至关重要．自从 1982 年 Bogenhagen 等[1]在爪蟾卵母细胞中首次发现了锌指蛋白 TFⅢA 参与 5S RNA 基因的调控，现已在各类生物(如酵母、植物、哺乳动物等)中发现了锌指蛋白参与基因调控的过程[2-5]．锌指蛋白中含有数量不等的锌指基序，每一个锌指基序都要通过结合 Zn^{2+} 形成稳定的类似“手指”结构，从而根据自身特性选择性结合特异的下游靶标．锌指蛋白能够通过与生物大分子(如 DNA、RNA)结合来实现基因表达的调控．根据锌指基序序列的特点(如半胱氨酸 (Cys) 和组氨酸 (His)的数量和位置)可以将锌指蛋白分为不同的类别，如 Jeremy Berg 等[6]划分的 9 类锌指蛋白(表 1)．

目前已知的人类 C2H2 型锌指蛋白基因约占全部锌指蛋白基因的 45%(751/1652)，是数量最多的一类转录因子．许多 C2H2 型锌指蛋白还含有其他的结构域，比如 KRAB、SCAN、BTB 以及 SET 结构域[7-8]，暗示着它们的功能是多样化的[9-11]．

目前大多数 C2H2 型锌指蛋白的功能是不明确的[12]．C2H2 型锌指蛋白作为转录因子，研究其调控的靶序列即可对其功能给出重要的线索和信息．染色质免疫共沉淀与高通量测序相结合的技术 (ChIP-seq)[13]对于转录因子与其结合的靶序列的研究十分重要[14]．但由于 C2H2 型锌指蛋白数量众多，其抗体的制备仍是一大阻碍．即使抗体齐全，在不同的生理条件下，同一个 C2H2 型锌指蛋白调控的靶序列很可能不同，很难检测所有生理状态下 C2H2 型锌指蛋白的靶序列．针对这一问题，将 ChIP-seq 实验与 C2H2 型锌指蛋白靶序列预测方法相结合，来获得 C2H2 型锌指蛋白的靶序列信息是一个非常好的解决方案．C2H2 型锌指蛋白靶序列预测结果对于 ChIP-seq 实验结果的筛选具有重要参考价值，能够极大地提高 C2H2 型锌指蛋白靶序

* 国家自然科学基金(31671376)，北京市科技新星计划(Z161100004916148)，国家国际科技合作专项(2014DFB30020，2014DFB30010)，国家重大科学研究计划(2015CB910700，2014CBA02001)和蛋白质组学国家重点实验室开放课题(SKLP-O201404，SKLP-O201507)资助项目．
** 通讯联系人．
Tel: 010-61777052, E-mail: yangdongbprc@163.com
收稿日期：2017-02-13，接受日期：2017-06-20

Table 1　The motif sequences and representative proteins of the nine types of zinc finger proteins[6]

表 1　9 类锌指蛋白的序列及其代表蛋白质[6]

锌指基序类型	序列	代表蛋白质
C2H2	C-X_{2-4}-C-X_{12}-H-X_{3-5}-H	TFⅢA
C8	C-X_2-C-X_{13}-C-X_2-C-X_{15}-C-X_5-C-X_{12}-C-X_4-C	Steroid-thyroid receptor
C6	C-X_2-C-X_6-C-X_6-C-X_2-C-X_6-C	GAL4
C3HC4	C-X_2-C-X_{9-27}-C-X_{1-3}-H-X_{2-3}-C-X_2-C-X_{4-48}-C-X_2-C	RING finger
C2HC	C-X_2-C-X_4-H-X_4-C	Retroviral nucleocapsid
C2HC5	C-X_2-C-X_{17-19}-H-X_2-C-X_2-C-X_2-C-X_{16-20}-C-X_{2-3}-C	LIM domain
C4	C-X_2-C-X_{17}-C-X_2-C	GATA-1
C3H	C-X_{6-14}-C-X_{4-5}-C-X_3-H	Nup475
C4HC3	C-X_2-C-X_{11-21}-C-X_2-C-X_4-H-X_2-C-X_{14-17}-C-X_2-C	Requium

列的准确性，为后续的功能和机制研究提供重要参考信息并起到辅助推断作用．本文将对 C2H2 型锌指蛋白靶序列预测方法进行介绍，并对各方法中的算法、训练集、金标准数据集等进行全面系统的总结归纳．

1　C2H2 型锌指蛋白简介及其与 DNA 结合的一般规律

一般来说，C2H2 型锌指基序与 DNA 的结合遵循以下规律[15−17]：以 C2H2 型锌指基序中 α 螺旋起始位氨基酸为 1 号氨基酸，那么−1、2、3、6 号这 4 个氨基酸就是结合 DNA 的关键氨基酸；−1、3、6 号 3 个关键氨基酸分别结合正义链 DNA 中 N3、N2、N1 3 个位置，而 2 号则结合反义链中的碱基，即与正义链 N4 互补配对的碱基(图 1)．

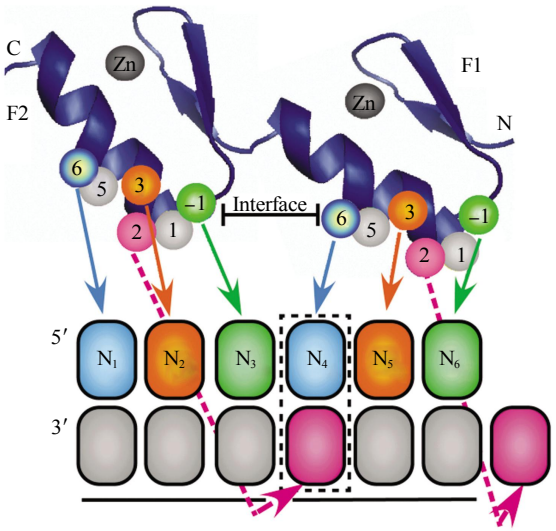

Fig. 1　The schematic diagram of the canonical recognition pattern of C2H2 zinc fingers binding DNA[15]

图 1　C2H2 型锌指基序与 DNA 序列结合原理图[15]

C2H2 型锌指基序与 DNA 结合的体外实验数据给 C2H2 型锌指蛋白靶序列预测方法提供了宝贵的数据来源．根据 C2H2 型锌指基序与 DNA 进行结合的原理，可以设计体外实验(如细菌单杂交技术和蛋白质结合芯片技术)，得到 C2H2 型锌指基序与 DNA 结合的实验证据．下面将根据预测方法得到的结果对目前已有的预测方法进行介绍．

2　通过 C2H2 型锌指序列预测结合的 DNA 序列

2.1　随机森林模型 (random forest model)

随机森林是一种基于分类树(classification tree)的算法．2001 年 Breiman 等把分类树组合成随机森林，即随机化调用变量(列)和数据(行)，生成一系列分类树，再汇总分类树的结果[18−19]．那么，随机森林模型是如何预测 C2H2 型锌指基序结合的 DNA 序列呢？

如果共有 n 个 C2H2 锌指基序(即观测数据)，有 k 类与之相结合的 DNA 序列，那么在构建每个分类树的时候，随机森林会随机地有放回地在训练集中重新选择 m 个($m < n$)观测值，即 Bootstrap 重新抽样．为了更大地增加随机性，随机森林每次会随机选取部分 C2H2 型锌指基序的特征变量进行分类树的构建．这样，每次构建的分类树的多样性会大大增加．在重复生成若干个分类树后，将这些分类树得到的每一类 C2H2 锌指基序结合的 DNA 序列平均结果作为最终结果．当再输入一个新的 C2H2 型锌指基序时，随机森林会根据已构建好的分类树对该 C2H2 型锌指基序进行投票分类，从而将该类别得到的最终结果作为预测的 DNA 序列(图 2)．

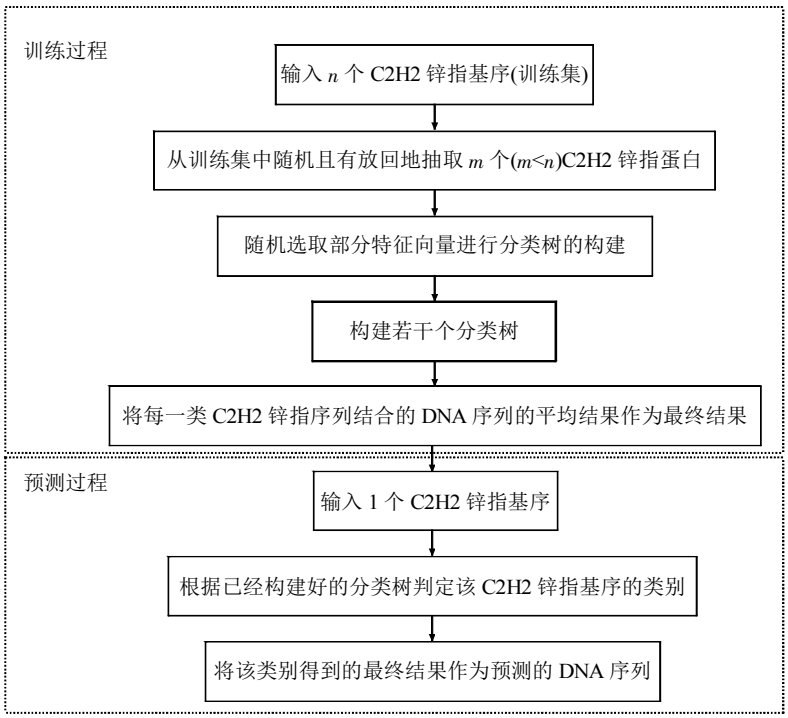

Fig. 2　The schematic diagram of random forest model of DNA-binding preferences of C2H2 zinc fingers

图 2　基于随机森林模型的 C2H2 型锌指基序靶序列预测方法基本过程

目前，已有 ZF Models (Zinc Finger Specificity prediction is based upon random forest model) [15] 和 ZifRC (Zinc Finger Recognition Code)[20]这两种方法通过随机森林模型构建了基于 C2H2 型锌指基序的靶序列预测方法.

ZF Models 基于 B1H(bacterial one-hybrid system)得到的 1 209 个单个 C2H2 型锌指基序结合的 DNA 序列以及 678 种 2 个串联 C2H2 型锌指基序结合的 DNA 序列的体外实验数据构建而成. ZF Models 可以通过识别 C2H2 型锌指蛋白中的 C2H2 型锌指基序，并对其中的关键氨基酸进行提取后，预测该 C2H2 型锌指蛋白结合的 DNA. 此外，ZF Models 还能够通过输入 4 位 C2H2 型锌指关键氨基酸获得该 C2H2 型锌指结合的 DNA. ZF Models收集了许多通过 SELEX (systematic evolution of ligands by exponential enrichment)技术得到的锌指基序与 DNA 结合的体外实验数据[21−24]作为金标准数据集，并通过计算预测 DNA 序列 PFM (position frequency matrix)与实验 DNA 序列 PFM 间的均方差(mean square error，MSE)来评估预测的准确性，对于某一个位点，MSE 可以从 0～1，对于随机碱基的位点(每一个碱基可能的概率为 0.25)与一个确定碱基的位点(1 种碱基概率为 1，其余 3 种为 0)间，MSE 为 0.1875.

ZifRC 所用的训练集为基于 B1H 技术所得到的 47 072 个 C2H2 型锌指基序与长度为 4 的随机碱基 DNA 序列结合的体外实验数据. 连接 2 个 C2H2 型锌指基序的标准长度为 4～6 个氨基酸. 当输入 C2H2 型锌指蛋白时，ZifRC 能够识别标准长度的锌指连接区，从而对 C2H2 型锌指蛋白中所有的 C2H2 型锌指基序结合的 DNA 序列进行预测，最终得到 C2H2 型锌指蛋白结合的 DNA，而对于非标准长度的连接区，则需要将其连接的 2 个锌指基序拆开后分别预测，再手动整合得到 C2H2 型锌指蛋白结合的 DNA. ZifRC 按照 C2H2 型锌指蛋白中锌指连接区均为标准长度的筛选原则，从 CIS-BP 数据库[25]中筛选得到了 64 个 C2H2 型锌指蛋白，并去除了每一个 C2H2 型锌指蛋白中重复的 motif，最终得到金标准数据集. 通过计算预测 DNA 序列 PWM (position weight matrix)与相应金标准中 DNA 序列 PWM 间的皮尔逊相关系数和 *P* 值，从而评估预测方法的准确性. 按照 ZifRC 的金标准数据库和评估方法，ZifRC 的准确性明显优于 ZF Models[20].

2.2　最近邻法(nearest neighbor approach)

最近邻法是数据挖掘分类技术中重要的方法之一，其核心思想为：如果一个样本在特征空间中的 k 个最相邻样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性[26]. 当我们将 C2H2 型锌指基序与其结合的 DNA 序列信息(训练集)输入到最近邻法时，最近邻法会计算训练集中每个 C2H2 型锌指基序间的距离(例如欧氏距离)从而将它们进行分类，将每一类 C2H2 锌指基序结合的 DNA 序列的平均结果作为最终结果. 当再输入一个新的 C2H2 型锌指基序时，最近邻法首先会计算 C2H2 型锌指基序与每一个 C2H2 型锌指基序间的距离，若 k 个最相邻的 C2H2 型锌指基序中的大多数属于某一类 C2H2 型锌指基序，则将那一类 C2H2 型锌指基序结合的 DNA 序列作为该 C2H2 型锌指基序结合的 DNA 序列，从而完成预测的过程(图 3).

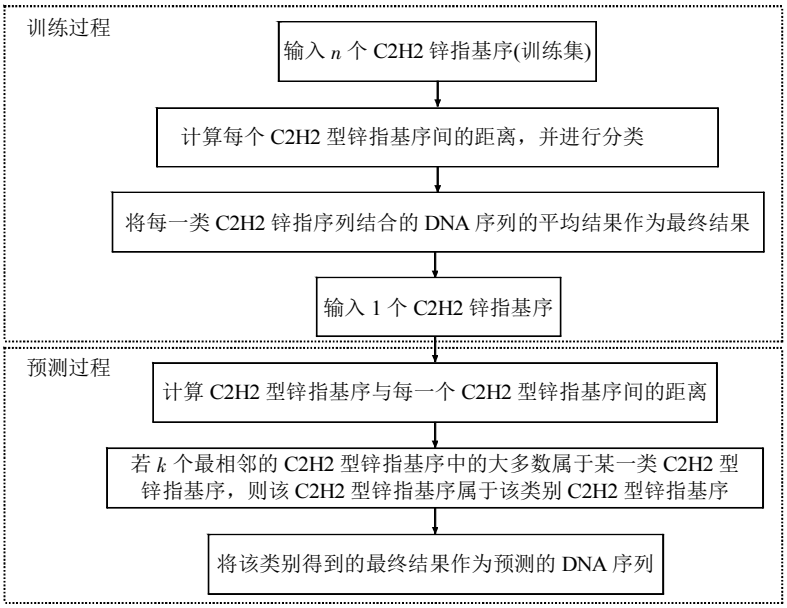

Fig. 3　The schematic diagram of nearest neighbor approach of DNA-binding preferences of C2H2 zinc fingers

图 3　基于最近邻法的 C2H2 型锌指基序靶序列预测方法基本过程

B1H screen of C2H2-ZF domain[27]就是基于最近邻法构建的. 同样的，B1H screen 所用的训练集也是基于 B1H 得到的 C2H2 型锌指基序与 DNA 结合的体外实验数据，只是在设计上有一些改变. 本方法设计了两种实验方案：a. F2 union. 设计 3 个串联的 C2H2 型锌指，仅改变处于中间的 C2H2 型锌指的氨基酸序列(F2)，得到 F2 与长度为 3 的随机碱基 DNA 序列结合的实验数据. b. F3 union. 仍然是先设计 3 个串联的 C2H2 型锌指，仅改变处于最末尾 C2H2 型锌指的氨基酸序列(F3)，得到 F3 与长度为 3 的随机碱基 DNA 序列结合的实验数据. 在得到了实验数据后，通过最近邻法构建预测模型. B1H screen of C2H2-ZF domain 能够识别 C2H2 型锌指蛋白中的 C2H2 型锌指基序，并预测 C2H2 型锌指基序结合的 DNA，但是它不能直接获得 C2H2 型锌指蛋白结合的 DNA，需要用户自己将预测得到的 C2H2 型锌指基序结合的 DNA 进行拼接获得. B1H screen 收集了 JASPAR [28]、UniProbe [29]、Jolma database of human transcription factors[24]、FlyFactorSurvey[30]数据库中 C2H2 型锌指蛋白及其结合 DNA 序列的信息，去重复后共包含 143 个 C2H2 型锌指蛋白，以此作为金标准数据库. B1H screen 所用的评估方法与 ZifRC 相似，均是通过计算预测 DNA 序列 PWM (position weight matrix)与相应金标准中 DNA 序列 PWM 间的皮尔逊相关系数和 P 值来评估预测的准确性，不同之处为 B1H screen 将皮尔逊相关系数大于 0.5 的预测结果作为可信的结果.

3 通过 DNA 序列预测结合的 C2H2 型锌指序列

3.1 人工神经网络(artificial neural network, ANN)

人工神经网络是通过模仿生物的神经网络结构构建的数学模型[31]. 其基本过程为：将原始数据输入人工神经网络后，会进行多个隐含层(hidden layers)的非线性处理，最终得到目标输出. 输入 C2H2 型锌指基序与其结合的 DNA 序列信息(训练集)至人工神经网络后，人工神经网络会挖掘其中深层次的信息，并得到隐含层，然后通过隐含层的信息处理传递挖掘数据的深层次结构，最终输出结果(即 DNA 序列结合的 C2H2 锌指基序). 将输出结果与真实结果的误差作为目标函数，当该目标函数达到最小值时即得到该人工神经网络的最佳参数，从而完成训练过程. 当输入一个新的 DNA 序列时，网络会输出该 DNA 序列对应的 C2H2 型锌指基序，完成预测.

ZiF-Predict 通过人工神经网络预测 DNA序列中可能存在的 C2H2 型结合位点. ZiF-Predict 整合了 ZifBASE 等[32-34]数据库中 C2H2 型锌指基序与 DNA 结合的体外实验数据，并以此为训练集和金标准数据库. 当输入一段 DNA 序列时，ZiF-Predict 会识别 DNA 序列中可能会被 C2H2 型锌指基序结合的 3 个连续碱基，并给出相应的 C2H2 型锌指基序的序列信息.

3.2 模块组装方法 (modular assembly approach)

体外实验能够获得单个 C2H2 型锌指蛋白与 DNA 结合的数据，那么若将所有可能的 3 个连续碱基与全部 C2H2 型锌指基序进行实验验证，就可以得到全部 C2H2 型锌指基序结合的 DNA 序列信息. 如果想知道一段 DNA 序列中有没有 C2H2 型锌指基序的结合位点，就可以检索这段 DNA 序列中有没有能够与 C2H2 型锌指基序相结合的 3 个连续碱基.

但实际结合过程中总往往并不是只有一个 C2H2 型锌指基序参与了 C2H2 型锌指蛋白与 DNA 结合的过程，因此 ZiFiT[35]采用的是 ZifDB 数据库[36]中 3 个连续的 C2H2 型锌指基序与 DNA 结合的体外实验数据. 输入了 DNA 序列后，ZiFiT 能够检索其中是否存在能够与 C2H2 型锌指基序相结合的 9 个连续碱基，并给出结合该 9 个连续碱基的 C2H2 型锌指基序序列信息. 此外，ZiFiT 还支持 BLAST 在线比对工具，方便寻找潜在的靶序列.

4 总结与展望

C2H2 型锌指基序与 DNA 结合的实验数据已经较为庞大，相应的预测方法也能够提供一些 C2H2 型锌指蛋白结合 DNA 的信息，但总体来说，本领域还处于一个初级的阶段. 通过体外实验数据得到的结果很难反应体内的真实情况，从而基于体外实验数据构建的预测方法存在一定的缺陷，而通过 ChIP-seq 等体内实验手段又难以判断具体是 C2H2 型锌指基因编码的哪一个 C2H2 型锌指蛋白与 DNA 发生了结合，更难以判断其中发生结合的 C2H2 型锌指基序. 相信对以上问题的继续探索会激起研究人员对 C2H2 型锌指蛋白预测的兴趣，促进实验手段的不断进步和预测算法的持续完善，使得预测准确性大大提高，最终揭开 C2H2 型锌指蛋白的神秘面纱.

参 考 文 献

[1] Bogenhagen D F, Wormington W M, Brown D D. Stable transcription complexes of *Xenopus* 5S RNA genes: a means to maintain the differentiated state. Cell, 1982, **28**(2): 413-421

[2] Corkins M E, May M, Ehrensberger K M, *et al*. Zinc finger protein Loz1 is required for zinc-responsive regulation of gene expression in fission yeast. Proc Natl Acad Sci USA, 2013, **110** (38): 15371-15376

[3] Gupta S K, Rai A K, Kanwar S S, *et al*. Comparative analysis of zinc finger proteins involved in plant disease resistance. PloS One, 2012, **7**(8): e42578

[4] Zhao X Q, Bai F W. Zinc and yeast stress tolerance: micronutrient plays a big role. Journal of Biotechnology, 2012, **158**(4): 176-183

[5] Tian C, Xing G, Xie P, *et al*. KRAB-type zinc-finger protein Apak specifically regulates p53-dependent apoptosis. Nature Cell Biology, 2009, **11**(5): 580-591

[6] Berg J M, Shi Y. The galvanization of biology: a growing appreciation for the roles of zinc. Science, 1996, **271** (5252): 1081-1085

[7] 马占福, 杨 冬, 贺福初, 等. KRAB型锌指蛋白在高等脊椎动物胚胎发育和肿瘤发生、发展中的调控功能. 遗传, 2010, **32**(5): 431-436
Ma Z F, Yang D, He F C, *et al*. Hereditas (Beijing), 2010, **32**(5): 431-436

[8] Collins T, Stone J R, Williams A J. All in the family: the BTB/POZ, KRAB, and SCAN domains. Molecular and Cellular Biology, 2001, **21**(11): 3609-3615

[9] Yang D, Ma Z, Lin W, *et al*. Identification of KAP-1-associated complexes negatively regulating the Ey and beta-major globin genes in the beta-globin locus. Journal of Proteomics, 2013, **80**: 132-144

[10] 杨 冬, 姜 颖, 贺福初. KAP-1：转录调控中的一个桥梁分子.

遗传, 2007, **29**(2): 131-136
Yang D, Jiang Y, He F C. Hereditas (Beijing), 2007, **29**(2): 131-136

[11] 田春艳, 张令强, 贺福初. KRAB型锌指蛋白(KZNF)的研究进展. 遗传, 2006, **11**(5): 1451-1456
Tian C Y, Zhang L Q, He F C. Hereditas (Beijing), 2006, **11**(5): 1451-1456

[12] Schmitges F W, Radovani E, Najafabadi H S, *et al*. Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Research, 2016, **26**(12): 1742-1752

[13] Furey T S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet, 2012, **13**(12): 840-852

[14] Jaini S, Lyubetskaya A, Gomes A, *et al*. Transcription factor binding site mapping using ChIP-Seq. Microbiology Spectrum, 2014, **2**(2): 1-21

[15] Gupta A, Christensen R G, Bell H A, *et al*. An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. Nucleic Acids Research, 2014, **42**(8): 4800-4812

[16] Emerson R O, Thomas J H. Adaptive evolution in zinc finger transcription factors. PLoS Genet, 2009, **5**(1): e1000325

[17] Liu H, Chang L H, Sun Y, *et al*. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome Biol Evol, 2014, **6**(3): 510-525

[18] Breiman L. Random forest. Machine Learning, 2001, **45**(1): 5-13

[19] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics, 2012, **99**(6): 323-329

[20] Najafabadi H S, Mnaimneh S, Schmitges F W, *et al*. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nature Biotechnology, 2015, **33**(5): 555-562.

[21] Hockemeyer D, Soldner F, Beard C, *et al*. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. Nature Biotechnology, 2009, **27**(9): 851-857

[22] Wood A J, Lo T W, Zeitler B, *et al*. Targeted genome editing across species using ZFNs and TALENs. Science, 2011, **333**(6040): 307

[23] Soldner F, Laganiere J, Cheng A W, *et al*. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. Cell, 2011, **146**(2): 318-331

[24] Jolma A, Yan J, Whitington T, *et al*. DNA-binding specificities of human transcription factors. Cell, 2013, **152**(1-2): 327-339

[25] Weirauch M T, Yang A, Albu M, *et al*. Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 2014, **158**(6): 1431-1443

[26] Hu L Y, Huang M W, Ke S W, *et al*. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus, 2016, **5**(1): 1304-1312

[27] Persikov A V, Wetzel J L, Rowland E F, *et al*. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. Nucleic Acids Research, 2015, **43**(3): 1965-1984

[28] Sandelin A, Alkema W, Engstrom P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research, 2004, **32** (Database issue): D91-94

[29] Newburger D E, Bulyk M L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Research, 2009, **37**(Database issue): D77-82

[30] Zhu L J, Christensen R G, Kazemian M, *et al*. FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Research, 2011, **39**(Database issue): D111-117

[31] Scotti L, Ishiki H, Mendonca Junior F J, *et al*. Artificial neural network methods applied to drug discovery for neglected diseases. Combinatorial Chemistry & High Throughput Screening, 2015, **18**(8): 819-829

[32] Jayakanthan M, Muthukumaran J, Chandrasekar S, *et al*. ZifBASE: a database of zinc finger proteins and associated resources. BMC Genomics, 2009, **10**: 421-427

[33] Dreier B, Beerli R R, Segal D J, *et al*. Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. The Journal of Biological Chemistry, 2001, **276**(31): 29466-29478.

[34] Maeder M L, Thibodeau-Beganny S, Osiak A, *et al*. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Molecular Cell, 2008, **31** (2): 294-301

[35] Sander J D, Zaback P, Joung J K, *et al*. Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. Nucleic Acids Research, 2007, **35**(Web Server issue): W599-605

[36] Fu F, Sander J D, Maeder M, *et al*. Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. Nucleic Acids Research, 2009, **37** (Database issue): D279-283

The Advancement of The Prediction Methods for DNA-binding Preferences of C2H2 Zinc Finger Proteins*

SHEN Pan, YANG Dong**, HE Fu-Chu

(*State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Radiation Medicine, Beijing* 102206, China)

Abstract C2H2 zinc finger proteins represent the largest family of transcription factors in mammalian. Their C2H2 zinc finger arrays are highly variable, indicating that most of them have unique DNA binding motifs, regulating different genes and playing diversified roles. However, the detailed regulatory functions of many C2H2 zinc finger proteins are unknown because of the unclear target sequences. The prediction of DNA-binding preferences of C2H2 zinc finger proteins is a commendable approach to figure it out. In this review, the canonical recognition pattern of C2H2 zinc fingers binding DNA was described. The prediction models of DNA-binding preferences of C2H2 zinc finger proteins according to their methods, training datasets, and golden standard datasets were summarized. This review is of great benefit to the comprehensive understanding of the prediction models of DNA-binding preferences of C2H2 zinc finger proteins. All of these information will facilitate the further theoretical and applied studies of C2H2 zinc finger proteins.

Key words transcription factor, C2H2 zinc finger proteins, prediction of DNA-binding preferences

DOI: 10.16476/j.pibb.2017.0047

* This work was supported by grants from The National Natural Science Foundation of China (31671376), the Beijing Nova Program (Z161100004916148), the International Science & Technology Cooperation Program of China (2014DFB30020, 2014DFB30010), the National Major Scientific Research Program (2015CB910700, 2014CBA02001) and the Foundation of State Key Lab of Proteomics (SKLP-O201404, SKLP-O201507).

**Corresponding author.

Tel: 86-10-61777052, E-mail: yangdongbprc@163.com

Received: February 13, 2017 Accepted: June 20, 2017