

文章编号 :1004-0374(2007)02-0122-05

ncRNA 研究技术进展

肖章奎, 薛良义*

(宁波大学生命科学与生物工程学院, 宁波 315211)

摘要: ncRNA 通过多种机制调控着基因的表达, 生物信息学、基因组 SELEX 技术及微阵列分析等方法在 ncRNA 的研究中发挥了重要作用, 导致在最近 5 年发现了大量的新 ncRNA, 本文就研究 ncRNA 的各种方法作一简要介绍。

关键词: ncRNA; 生物信息学; 生物芯片; 基因组 SELEX; 质谱法; RNA 组学; RNA 鉴定
中图分类号: R318.04; Q75 **文献标识码:** A

Advances in the research technology of ncRNA

XIAO Zhangkui, XUE Liangyi*

(College of Life Science and Biotechnology, Ningbo University, Ningbo 315211, China)

Abstract: ncRNA plays an important role in regulating gene expression by various mechanisms. Many approaches, such as bioinformatics, genomic SELEX technology, and microarray analysis, were applied to study ncRNA, resulting in the discovery of a large number of new ncRNAs in recent five years. Here we briefly summarize these methods for identifying and validating ncRNA.

Key words: ncRNA; bioinformatics; microarray; genomic SELEX; mass spectrometry; momics; RNA identification

近年来, 在原核生物、真核生物以及古细菌中鉴定了许多 ncRNA(non-coding RNA), 如小分子核仁 RNA(small nucleolar RNA, snoRNA)、小 RNA(microRNA, miRNA)、小干扰 RNA(short interfering RNA, siRNA)等, 并发现它们具有多种功能, 在多个水平上调节着基因的表达, 如对染色体结构的影响, 对 RNA 加工修饰及稳定性的影响, 对转录和翻译的影响, 甚至对蛋白质的稳定性和转运都有影响^[1]。本文对 ncRNA 的研究方法作一简要介绍。

1 生物信息学方法(Bioinformatics)预测

在利用 gene-finding 软件预测基因编码区的同时, 就尝试着用生物信息学方法对 ncRNA 进行鉴定; 但由于 ncRNA 缺少编码蛋白质的基因所具有的典型特征, 如启动子和终止子、开放阅读框、特

异的剪切位点、多聚腺苷酸化位点和 CG 岛等, 且 ncRNA 基因较小, 用于 gene-finding 软件的基序(motif)变动较大等, 因此, 到目前为止, 还没有高效且通用的 ncRNA 基因的预测算法。

现在能成功对 ncRNA 预测的 gene-finding 编程软件一般被设计成只能搜索单一种类的 ncRNA, 如 tRNAScan-SE 搜索 tRNA^[2]、snoScan 搜索带 C/D 盒的 snoRNAs^[3]、SnoGps 搜索带 H/ACA 盒的 snoRNAs^[4]、mirScan 搜索 microRNA^[5]等等。

一些基于基序聚类的软件, 如 RNAmotifs^[6]、Erpin^[7]以及 Patsearch^[8]也用于对 ncRNA 的搜索, 但是这些软件同搜索单一种类的 ncRNA 软件相比, 灵敏度和特异性都较差。实际上, 用实验方法已证实的 ncRNA 很少是用这类软件鉴定出来的。

收稿日期: 2006-12-10; 修回日期: 2007-03-12

基金项目: 浙江省自然科学基金(Y306295); 宁波市自然科学基金(2006A610083)

作者简介: 肖章奎(1974—), 男, 硕士研究生; 薛良义(1962—), 男, 博士, 教授, 博士生导师, * 通讯作者, E-mail: xueliangyi@nbu.edu.cn

随着各种生物物种基因组计划的实施,基因组的序列比较分析可用来检测 ncRNA 和 cis-regulatory RNA 的二级结构^[9-10],如用 QRNA 已检测出在大肠杆菌、酿酒酵母菌和激烈火球菌中的 ncRNA,并在随后的实验中得到了证实。贺华良等^[11]通过比较基因组和分子生物学方法分析了 5 种果蝇全基因组内含子区域的保守序列,获得了 3 个全新的非编码 RNA 基因:1 个典型的带 C/D 盒的 snoRNA 基因和两个 miRNA 基因。

2 对 ncRNA 的测序鉴定

2.1 直接测序 单一的 ncRNA 通过变性凝胶电泳(如含 EB 的聚丙烯酰胺电泳)可从总 RNA 中分离出来。分离出来的 ncRNA 5' 端在多聚核苷酸激酶的作用下加上 r-³²P 标记的 AMP,或在 T4DNA 连接酶的作用下在 3' 端连接上 ³²P 标记的 CMP。5' 端和 3' 端标记的 ncRNA 用酶或化学方法测序。酶法测序是指标记过的 ncRNA 在 50 - 55 °C, 7 mol/L 尿素存在的条件下,由核糖核酸酶,如 RnaseT1、T2、U2、PHYM、CL3、A、M1 等对其进行碱基特异性水解,得到大小不一的 RNA 片段,随后通过变性聚丙烯酰胺凝胶电泳和放射自显影得到其序列。化学方法测序是指将标记过的 RNA 中的每一种碱基分别进行特异性化学修饰后,采用苯胺催化而进行条带剪切,产生大小不一的标记性片段,通过变性聚丙烯酰胺凝胶电泳和放射自显影得到其序列。

最早对 ncRNA 直接测序的研究是在 tRNA 和 rRNA 上进行的^[12-15],如 16S rRNA^[15]。利用直接测序来鉴定新的 ncRNA 种类还远未过时,如最近通过直接测序在真核生物中发现的 snoRNA,可能与 rRNA 的修饰有关^[16]。这种技术也用于对革兰氏阳性细菌中丰富的 RNA 进行分析^[17-18]。

2.2 cDNA 文库的测序 ncRNA 分离方法主要有两种:第一种将某种生物的总 RNA 通过变性凝胶电泳(如变性聚丙烯酰胺凝胶电泳)和胶纯化可以得到所需的 ncRNA (<500nt 的 RNA);第二种则是利用 ncRNA 结合蛋白的抗体进行免疫沉淀反应来分离出 ncRNA,即首先从细胞中纯化出核糖核蛋白颗粒(RNPS),然后与已知的 ncRNA 结合蛋白的特异抗体进行免疫沉淀反应,最后通过酚提取就可得到相关的 ncRNA。

许多从生物体内分离出来的 ncRNA 长度上都比 mRNA 小得多,为 20 - 500 nt,并且也不具有多聚 A 尾巴,不能直接用寡聚 dT 引物来反转录形成

cDNA,因此,要对分离出来的 ncRNA 进行处理。这种处理是指在分离出来的 ncRNA 的 3' 端,用 poly (A)多聚酶加上寡聚 C 或 A 尾巴,或在 T4 RNA 连接酶作用下连接一个寡聚核苷酸连接子,随后 ncRNA 的 5' 端也可通过 T4RNA 连接酶连接一个寡聚核苷酸连接子,这些寡聚核苷酸连接子是已知序列的 RNA 或 DNA。处理过的 ncRNA 通过 RT-PCR 形成 cDNA。

cDNA 与载体连接,构建 cDNA 文库。对文库进行测序并对测序结果进行多方面分析,如通过 BLAST 进行染色体定位、Northern 杂交分析 ncRNA 的表达、原位杂交找出其在细胞和亚细胞的位置以及分析 ncRNA 结合蛋白等。这些分析有助于发现新的 ncRNA 种类并对其进行功能上的鉴定,如通过构建特异性 cDNA 文库,Marker 等^[19]在拟南芥 (*Arabidopsis thaliana*)中鉴定出 140 个 ncRNA,包括 88 个 snoRNAs、2 个 7SL RNA、13 个 U snRNAs 和 1 个类 tRNA-RNA 等,还有 29 个在基因间隔区域,3 个在内含子区域,4 个在 ORFs 的候选 ncRNA 分子。罗俊等^[20]通过构建特异性的 cDNA 文库,发现并鉴定了新的贾第虫 box H/ACAsnoRNA。同样的技术也在黑腹果蝇 (*Drosophila melanogaster*)^[21]、真菌 (*Archaeoglobus fulgidus*)和古细菌 (*Sulfolobus solfataricus*)^[22-23]、大肠杆菌 (*E. coli*)^[24-25]和嗜热菌 (*Aquifex aeolicus*)^[26]中成功运用,均鉴定出了大量的 ncRNA。

这种方法很灵敏,但逆转录酶转录时出错率较高,同时由于 RNA 尤其是 tRNA 和 rRNA 的二级结构和碱基的修饰会造成转录的提前终止,因此,这种方法也有一定的局限性。

3 生物芯片分析

生物芯片技术也可用于 ncRNA 研究,如 Genetix 的一系列微阵列产品,适用于 RNA 干扰 (RNA interference)的研究,能有效筛选 siRNA;又如在细菌中,绝大多数具有一定功能的 ncRNA 被编码在基因间隔区域 (IGRs)。第一个既包括编码区又包括 IGRs 的 DNA 芯片已用于对模式生物大肠杆菌的研究,它不仅可分析出细菌所有的 mRNA、tRNA 和 rRNA,而且还可分析出细菌中大于 40 bp 的 IGRs。因此,使用这种芯片不仅可分析细菌 mRNA 的表达水平,还可特异性地分析来自细菌中 IGRs 的转录产物,如用来分析与大肠杆菌 Hfq 蛋白结合的 ncRNA^[27]。

生物芯片技术也可用于鉴定真核生物中的

ncRNA, 并可研究它们在不同组织中的表达。Inada 和 Guthrie^[28]利用生物芯片技术对酵母中与 La 蛋白(Lhp1)结合的 ncRNA 进行分析, 发现了至少三种新的 H/ACA snoRNAs, 它们存在于基因间隔区域, 且表达程度较高。设计好的生物芯片还可在真核生物中搜索具有一定功能的 ncRNA, 如提取存在于人、小鼠和大鼠的 3 478 个基因内和基因间的 ncRNA 序列^[29], 再设计出含这些序列的生物芯片。利用这些生物芯片与野生型小鼠的 16 种组织中分离出的 RNA 进行 Northern 杂交, 检测出了 55 种新 ncRNA, 进一步证实其中的 8 种 ncRNA 在小鼠所有组织中高表达, 更有趣的是, 这些 ncRNA 只有 5 种在大鼠的组织中表达, 却没有一种在人类组织或培养的细胞中发现, 这 5 种 ncRNA 在大鼠和小鼠中的保守性表达可能暗示它们在这两种生物中具有一定的功能而在人类中缺乏这种功能。目前看来, 利用生物芯片技术来发现 ncRNAs 是一条具有巨大潜力的新途径。Bertone 等^[30]推出的 Tiling 芯片, 可以全面系统地研究基因组转录的所有 RNA 分子, 这一技术已在人^[31-32]和大肠杆菌^[33]等基因组里成功运用。

4 基因组 SELEX 技术(genomic systematic evolution of ligands by exponential enrichment)

许多 ncRNA 在生命过程的不同时期常形成核糖核酸蛋白颗粒(RNPs), 与 ncRNA 结合的蛋白质起着帮助 ncRNA 折叠成它的活性形式, 或在 ncRNA 起作用前防止核酸酶的分解等作用, 还有一些蛋白质与 ncRNA 的相互作用则直接调节着两者的行为。基因组 SELEX 技术是利用分子生物学技术, 构建人工合成的某一个生物体基因组的单链随机 RNA 文库, 其中随机序列长度在 20 - 40 nt。单链随机 RNA 片段易形成发卡、口袋、假节、G-四聚体等二级结构, 能与蛋白质结合, 形成具有很强结合力的复合物。利用这一原理, 将随机 RNA 文库与 ncRNA 结合蛋白相互作用, 洗脱筛选出特异寡核苷酸配基(aptamer), 经 RT-PCR 及体外转录生成新的次一级文库, 再与该靶蛋白结合。反复数个循环, 即可筛选出能与 ncRNA 结合蛋白特异结合的寡核苷酸片段。该片段的序列一旦被确定, 就可获得在基因组中的相应位置, 可检测出有可能进行某种 ncRNA 表达的区域。基因组 SELEX 技术已成功的运用于挑选与特异蛋白结合的 mRNA。近来, Schroeder 实验室已经利用这种方法来鉴定大肠杆菌中的与 Hfq 蛋白结合的 ncRNA, 并初步鉴定出了大量的 ncRNA,

如反义 RNA 和一些存在于基因间区域的候选 ncRNA 分子^[34]。

5 质谱法(mass spectrometry)

质谱法可对复杂的混合物进行快速而灵敏的分析, 并有相当大的自动化操作过程。目前主要有 MALDI 质谱法(matrix-assisted laser desorption ionization mass spectrometry)、ESI 质谱法(electrospray ionization mass spectrometry)和质谱联用法(MS/MS)。在对 RNA 的研究中, 它们可用来对 RNA 进行测序, 如用 ESI、MS/MS 法测序或用酶和化学方法对 RNA 进行消化后产生序列梯度, 然后用 MALDI 质谱法进行测序, 如对嗜水气单胞菌(*Aeromonas hydrophila*) 16 S rRNA 进行的测序^[35]。由于碱基 U 和 C 的相对分子量只有 1 的区别, 使 RNA 测序的最大长度有一定限制, 一般 10 - 20 个 nt 的寡核苷酸片段用这些方法是很精确的, 如对一种长的 RNA 要精确测序, 必须先消化后再多次测序。质谱法还可对转录后 RNA 的修饰进行检测, 它主要是通过通过对修饰后的 RNA 与预期的 RNA 质量进行比较得到一个修饰后的增量, 对增量进行分析可检测其修饰, 同时也可对 RNA 3D 结构和 RNA 与蛋白质的相互作用进行分析, 如 MS 3D 技术已成功用于对 HIV-1 的 RNA 3D 结构和 HIV-1 病毒装配时 RNA 与蛋白质的相互作用分析^[36]。当然这种技术也可用于对 ncRNA 的分析, 如对 tRNA 和 rRNA 结构的分析。

6 功能性的 RNA 组学方法——RNA 鉴定后技术

ncRNA 的鉴定仅仅是作为对它们功能解释的第一步, 要进一步阐明其功能, 可采用以下几种方法:(1)大多数具有一定功能的 ncRNA 是 RNPs 的组成部分, 鉴定后的 ncRNA 可用来捕获细胞提取物中的 ncRNA 结合蛋白, 对蛋白质结构和成分的分析可能揭示出 RNPs 的功能, 因为有些蛋白质可能含有我们已了解的具有催化活性的区域。(2)迄今为止有许多已被鉴定的 ncRNA 能通过反义机制结合特异的靶 RNA, 靶 RNA 包括 mRNA 或别的 ncRNA, 如 rRNA、snRNA、tRNA 等, 对靶 RNA 的分析有助于了解 ncRNA 的功能。(3)对 ncRNA 表达模式的分析也有助于对其功能的了解。ncRNA 或 RNPs 在细胞或亚细胞中的位置, 如存在于核仁、核或细胞质中, 也许能提示这种 ncRNA 或 RNPs 可能涉及在这些细胞间隔区域里所起的作用, 荧光标记技术可对所要研究的 ncRNA 进行定位^[37]; 又如, 对来自不同组织或发育时期的细胞总 RNA 进行 Northern 杂

交,可分析ncRNA组织特异性或不同发育时期的表达,从而间接分析其功能。(4)基因敲除方法可用于验证ncRNA的功能。对于某些模式生物如大肠杆菌,基因敲除常在几天内就可完成^[38-39],但对绝大多数其他生物体说,传统的基因敲除技术耗时较长。近来, RNA 干扰可用于快速的ncRNA敲除^[40-41],但是RNAi靶向ncRNA的作用机制还没有完全弄清楚,近来证实化学修饰的反义miRNA能用于某种miRNA种类的敲除^[42]。

7 展望

以上这些技术手段各有其优缺点,随着各种技术手段的不断改进和更新,对ncRNA基因进行识别及其结构和功能研究,有可能发现新的ncRNA基因及其在基因表达调控中的作用,从而使人们对基因组的结构与功能的理解登上一个新的台阶。

[参 考 文 献]

- [1] 秦云霞,田 娥,刘志昕,等. 非编码RNA及其研究进展. 生物技术通报, 2004, 5: 9-12
- [2] <http://lowelab.ucsc.edu/tRNAscan-SE/>
- [3] <http://lowelab.ucsc.edu/snoscan/>
- [4] <http://lowelab.ucsc.edu/snoGPS/>
- [5] <http://genes.mit.edu/mirscan/>
- [6] <http://www.scripps.edu/mb/case/casegr-sh-3.5.html>
- [7] <http://tagc.univ-mrs.fr/erpin/>
- [8] <http://www.ba.itb.cnr.it/BIG/PatSearch/>
- [9] www.genetics.wustl.edu/eddy/software/
- [10] Rivas E, Klein R J, Jones T A, et al. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol*, 2001, 11: 1369-1373
- [11] 贺华良,周 惠,肖振东,等. 果蝇3个新的小分子非编码RNA的鉴定. 科学通报, 2006, 51(20): 2393-2398
- [12] Donis-Keller H, Maxam A M, Gilbert W. Mapping adenines, guanines, and pyrimidines in RNA. *Nucleic Acids Res*, 1977, 4: 2527-2538
- [13] Yarus M, Barrell B G. The sequence of nucleotides in tRNA Ile from *E. coli*. *Biochem Biophys Res Commun*, 1971, 43: 729-734
- [14] Brownlee G G, Cartwright E, McShane T, et al. The nucleotide sequence of somatic 5S RNA from *Xenopus laevis*. *FEBS Lett*, 1972, 25: 8-12
- [15] Ehresmann C, Stiegler P, Carbon P, et al. Recent progress in the determination of the primary sequence of the 16S RNA of *Escherichia coli*. *FEBS Lett*, 1977, 84: 337-341
- [16] Balakin A G, Smith L, Fournier M J. The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell*, 1996, 86: 823-834
- [17] Pichon C, Felden B. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc Natl Acad Sci USA*, 2005, 102: 14249-14254
- [18] Trotochaud A E, Wassarman K M. A highly conserved 6S RNA structure is required for regulation of transcription. *Nature Struct Mol Biol*, 2005, 12: 313-319
- [19] Marker C, Zemann A, Terhorst T, et al. Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr Biol*, 2002, 12: 2002-2013
- [20] 罗 俊,周 惠,陈重建,等. 贾第虫4种新的box H/ACAsnoRNA的鉴定及其进化意义. 科学通报, 2006, 51(17): 2018-2023
- [21] Yuan G H, Klambt C, Bachelierie J P, et al. RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res*, 2003, 31: 2495-2507
- [22] Tang T H, Bachelierie J P, Rozhdestvensky T, et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci USA*, 2002, 99: 7536-7541
- [23] Tang T H, Polacek N, Zywicki M, et al. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol*, 2005, 55: 469-481
- [24] Vogel J, Bartels V, Tang T H, et al. RNomics in *Escherichia coli* detects new RNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res*, 2003, 31: 6435-6443
- [25] Kawano M, Reynolds A A, Miranda-Rios J, et al. Detection of 50- and 30-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res*, 2005, 33: 1040-1050
- [26] Willkomm D K, Minnerup J, Huttenhofer A, et al. Experimental RNomics in *Aquifex aeolicus*: identification of small non-coding RNAs and the putative 6S RNA homolog. *Nucleic Acids Res*, 2005, 33: 1949-1960
- [27] Zhang A X, Wassarman K M, Rosenow C, et al. Global analysis of small RNA and mRNA targets of Hfq. *Mol Microbiol*, 2003, 50: 1111-1124
- [28] Inada M, Guthrie C. Identification of Lhp1p-associated RNAs by microarray analysis in *Saccharomyces cerevisiae* reveals association with coding and noncoding RNAs. *Proc Natl Acad Sci USA*, 2004, 101: 434-439
- [29] Babak T, Blencowe B J, Hughes T R. A systematic search for new mammalian non-coding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, 2005, 6: 104
- [30] Bertone P, Gerstein M, Snyder M. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chrom Res*, 2005, 13: 259-274
- [31] Kapranov P, Drenkow J, Cheng J, et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res*, 2005, 15: 987-997
- [32] Cheng J L, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 2005, 308: 1149-1154

- [33] Selinger D W, Cheung K J, Mei R, et al. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat Biotechnol*, 2000, 18: 1262-1268
- [34] Huttenhofer A, Vogel J. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*, 2006, 34: 635-646
- [35] Ni J, Pomerantz S C, Rozenski J, et al. Interpretation of oligonucleotide mass spectra for determination of sequence using electrospray ionization and tandem mass spectrometry. *Anal Chem*, 1996, 68: 1989-1999
- [36] Yu E, Fabris D. Direct probing of RNA structures and RNA-protein interactions in the HIV-1 packaging signal by chemical modification and electrospray ionization Fourier transform mass spectrometry. *J Mol Biol*, 2003, 330: 211-223
- [37] Vitali P, Basyuk E, Le Meur E, et al. ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *Cell Biol*, 2005, 169: 745-753
- [38] Datsenko K A, Wanner B L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA*, 2000, 97: 6640-6645
- [39] Yu D G, Ellis H M, Lee E C, et al. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc Natl Acad Sci USA*, 2000, 97: 5978-5983
- [40] Willingham A T, Orth A P, Batalo S, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, 2005, 309: 1570-1573
- [41] Nakamoto M, Jin P, O'Donnell W T, et al. Physiological identification of human transcripts translationally regulated by a specific microRNA. *Hum Mol Genet*, 2005, 14: 3813-3821
- [42] Krutzfeldt J, Rajewsky N, Braich R, et al. Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature*, 2005, 438: 685-689

· 简讯 ·

上海药物所计算生物学研究取得重要进展

上海药物所药物发现与设计中心(DDDC)近年来在计算生物学、计算化学和药物设计研究方面取得了可喜的成绩, 相关研究成果分别在《美国科学院院刊》(PNAS)、《美国化学会志》(JACS)、《核酸研究》(NAR)、《生物化学杂志》(JBC)、《分子生物学杂志》(JMB)、《美国化学会药物化学杂志》(JMC)等国际一流刊物上发表。最近, 中心研究人员在蛋白质-蛋白质相互作用及其网络预测方法学发展方面取得重要进展。蒋华良研究员带领学生张健和沈菊文等经过两年努力, 发展仅根据蛋白质的序列即可预测蛋白质-蛋白质相互作用的新理论预测方法。研究结果于 2007 年 3 月 5 日发表在《美国科学院院刊》(PNAS)在线版上(<http://www.pnas.org/papbyrecent.shtml>)。

蛋白质-蛋白质相互作用(PPI)决定着从转录调节到酶级连反应的几乎所有的生物功能, 这方面的研究具有重要的科学价值和应用前景。然而, 目前的实验方法, 如 GST pull down 和免疫共沉淀方法的通量还不足以满足蛋白质组相互作用网络研究的需要, 酵母双杂交测定 PPI 的速度虽快, 但精度不够。因此, 发展理论方法在基因组水平上预测 PPI 及其相互作用网络, 对功能基因组研究具有十分重要的意义, 也是目前生命科学的前沿领域, 为此 *Nature Biotechnology* 等杂志专门设立了计算生物学(Computational Biology)栏目。目前大多数蛋白质-蛋白质相互作用预测方法需要同源蛋白信息或者蛋白相互作用标识物信息, 这类方法能应用的范围有限, 不能对一般化的蛋白质-蛋白质相互作用给出较好预测结果, 更不能应用于大规模 PPI 网络的预测。

蒋华良研究员等发展的方法是支持向量机算法——一种机器学习算法。他们首先将 20 种氨基酸根据极性和大小分成 7 类, 并用连续的三个氨基酸作为一个单位(三联子)来描述蛋白质序列, 以降低蛋白质相互作用空间的复杂性; 他们还发展了新的内核函数, 该函数考虑了蛋白相互作用的对称性, 因此比现有支持向量算法的内核函数更适合于表征蛋白质-蛋白质的相互作用; 然后他们用超过 16000 对实验测定的蛋白质-蛋白质相互作用结果构造了通用性 PPI 预测模型。他们方法的预测精确性大于 80%, 并能用于不同类型 PPI 网络的预测, 意味着即使只获得蛋白质序列信息, 他们的方法依然能够用于任意新蛋白的功能研究或预测老蛋白质的新功能。蒋华良研究员等发展的蛋白质-蛋白质相互作用方法为蛋白质功能研究提供了较好的理论工具, 是计算生物学研究领域的重要进展。同时他们的方法也可能应用于设计新的药物, 即设计新的化合物或蛋白质调控蛋白质相互作用网络, 而不是抑制或激动单一的靶标蛋白。

摘自 <http://www.sibs.ac.cn>