大数据层面的 microRNA 功能相似性分析

王莹莹 蔡云鹏

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 随着大数据时代的来临，microRNA 与基因的序列数据不断增加，如何从大量的数据中挖掘有生物学意义的信息成为新的热点问题。研究表明 microRNA 间以协作的方式在疾病中发挥作用，并呈现出网络的结构化趋势。因此，系统分析不同 microRNA 间的相似性将在疾病生物学标记挖掘等研究领域起到关键的桥梁作用。而 microRNA 通过调节其靶基因发挥作用，所以本研究将充分利用现有靶基因数据，从功能角度分析 microRNA 间的相似性。研究选取前期工作所得的靶基因优化列表，利用富集分析将基因集合转化为功能节点集合，并在此基础上利用集合相似性测度计算 microRNA 对在不同层面的功能一致性。结果表明，相同家族的 microRNA 倾向于调控相同或相似的靶基因；类比于非靶基因，microRNA 靶基因倾向于共享较多相似的细胞组分，而在生物学通路及生物学过程中则具有相对较低的相似性。

关键词 microRNA；功能相似性；靶基因
中图分类号 Q 811.4 **文献标志码** A

microRNA Functional Similarity Analysis on Big Data Level

WANG Yingying CAI Yunpeng

(*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract The numbers of microRNA and genes sequences have increased greatly with the advent of big data era. Thus how to explore useful information with biological significances from massive datasets has become a new hot topic. Former researches showed that microRNAs tended to play roles in diseases in a cooperative way and the relationships could be presented in the form of network. As a result, similarity analysis for microRNAs through a system way could play an important role in the field of disease biomarkers discovery. Considering that microRNAs play regulation roles by binding to their target genes, we focused on the available target gene data to analyze the similarity of microRNA pairs on functional levels. The optimization microRNA targets list generated by our former research as input were chosen and the enrichment analysis was used to map gene sets into functional term sets. The similarities between microRNAs were then calculated using similarity metrics on functional levels. Our results show that microRNAs in the same family tend to regulate the same or similar target genes. Compared with non-target genes, microRNA target genes tend to share similar cellular component. However, they show fewer similarities on biological pathway and biological progress levels.

Keywords microRNA; functional similarity; target gene

收稿日期：2014-3-24
作者简介：王莹莹，博士，助理研究员，研究方向为生物信息学；蔡云鹏（通讯作者），博士，硕士研究生导师，研究方向为生物信息学、机器学习和进化计算， E-mail：yp.cai@siat.ac.cn。

1 引　言

随着二代测序等技术的出现与完善，生物数据在基因组学、转录组学、microRNA 组学、蛋白质组学和代谢组学等层面的数据正在以史无前例的速度增长。用于数据的存储、处理和分析所需的人力、物力与财力甚至远远超出了产生数据本身的需求。因此，如何利用生物学数据的特点，系统、合理的分析、挖掘生物数据中蕴含的信息也就成为“大数据时代”生物信息学的新挑战。

microRNAs 是一类长度在 19～24 nt 的非编码 RNA，通过抑制或者降解其数以千计的靶基因在转录后调控层面上行使重要功能[1]。近期研究发现，人体内有超过 1000 个 microRNA，大约可调控人类 1/5 的编码蛋白的基因[2]。microRNA 在多种生物学过程如生长和发育等过程中发挥着重要作用，并且具有家族性及物理位置的聚集性[3-10]。

microRNA 组学(microRNomics)指的是在基因组层面上研究 microRNA 的表达、靶基因及其生物学功能的一门新兴的组学[11]。microRNA 组学的研究内容可分为 microRNA 和靶基因两个互为补充、缺一不可的层面。随着大数据时代的来临，microRNA 与基因的序列数据不断增加，如何从大量的 microRNA—靶基因数据中挖掘有生物学意义的信息也就成为新的热点问题。

由于受到实验技术的限制，无法对多个 microRNA 同步检测靶基因，目前仅有部分人类 microRNA 具有经实验证实的靶基因数据。TarBase[12]和 miRecords[13]数据库中仅存储了人类 125 个 microRNA(约为目前已知人类 microRNA 总数目的 10%)的 836 对实验证实的 microRNA—靶基因信息。而之前的研究表明，每个 microRNA 调控的基因至少有上百个，因此，如何准确预测 microRNA 的靶基因也就成为了microRNA 组学的一个主要挑战[14,15]。目前，已经有若干较为成熟的预测算法如 miRanda[16]、PicTar[17]和 TargetScan[18,19]等。每种预测算法均具有自身的特色与优势。但是，不同的靶基因预测算法对相同的 microRNA 进行预测得到的结果相差很大，这为后续的研究带来了一定的困扰。我们之前的研究发现，尽管不同的预测算法针对同一个 microRNA 预测得到的基因名称不同，但这些基因在功能层面上却具有一致性，即同一个 microRNA 的靶基因倾向于共享某些功能节点，这就为从功能层面重新审视 microRNA 靶基因并对其进行优化奠定了基础[20]。基于此结论，我们选取多个常用的 microRNA 靶基因预测算法，通过功能层面对靶基因进行排序，从生物学角度重新整合和优化，为后续的研究提供了可选资源[21]。

之前的研究表明，不同的 microRNA 间会以协同合作的方式在疾病中发挥作用并呈现出网络的结构化趋势[22]。因此，系统分析不同 microRNA 间的相同之处将在疾病生物学标记挖掘等研究领域起到桥梁作用。而 microRNA 通过调节其靶基因发挥作用，所以充分利用现有靶基因数据，从功能角度分析 microRNA 间的相似性也就成为可能。

基于上述考虑，本研究选取我们前期工作所得到的靶基因优化列表，利用富集分析将基因集合转化为功能节点集合，并在此基础上基于集合相似性测度计算 microRNA 对在不同层面的功能一致性。

2 材料与方法

2.1 microRNA 组学数据

2.1.1 microRNA 家族数据

本文从 miRBase 数据库中下载了 283 个人类 microRNA 的家族信息。

2.1.2 microRNA—靶基因数据

本文研究选取了 1218 个人类成熟的 microRNA，对于每个 microRNA，选取通过多层次功能综合排序后靶基因列表中的前 50% 作为靶基因数据。microRNA 以通用的规则“has（物种缩写）-miR（成熟序列缩写）-数字编号”命名，靶基因的名称则采用 Ensembl Gene ID（以 ENSG 为前缀，后跟数位数字编号）。

2.2 功能层面数据

2.2.1 GO 功能节点数据

GO（Gene Ontology，基因本体论）数据库以有向无环图的方式展示数据。因其包含生物学信息的系统性和广泛性，成为基因功能分析最常用的数据库之一[23]。

在本文中，我们选择了 GO 的三个子库：BP（Biology Process 生物学过程）、MF（Molecular Function 分子功能）以及 CC（Cellular Component 细胞组分）进行分析。其中，BP 包含 5140 个节点、MF 包含 2782 个节点、CC 包含 851 个节点。

2.2.2 生物通路数据（Pathway）

由于 microRNA 通过调节众多靶基因行驶功能，其靶基因也就有可能存在于同一个通路中。研究中，我们选取了 Molecular Signatures Database（v3.0）数据库[24]的 2999 条通路数据信息，其中所含的通路包含来自多个在线数据库以及生物学文献的信息。

2.2.3 转录因子（Transcript Factor，TF）

转录因子和 microRNA 是两个层面的调控子，在不同的转录层面上起到重要的作用。我们从 TransFac 12.1 数据库[25]中获取了 636 个转录因子的 2449 对转录因子—靶基因信息。

2.3 富集分析

对于 GO 的三个子库（BP、CC 和 MF）、转录因子和生物学通路这 5 类功能信息，我们利用 Fisher 精确检验进行富集分析，设定阈值为 0.05，所有 P 值不超过阈值的都被选作成为富集的功能节点，进行下一步的分析。

具体来说，采用单侧 Fisher 精确检验进行富集分析，即利用超几何检验的原理推测多种生物功能数据中包含的 microRNA 靶基因的比例是否与所有输入数据中靶基因所占的比例相同。具体的数值关系可以用表 1 来表示。

Fisher 精确检验的 P 值表示在 k 个microRNA 靶基因中，至少有 x 个被生物功能节点数据 S 注释的概率，具体为：

$$p=1-\sum_{i=0}^{x-1}\frac{\binom{M}{i}\binom{N-M}{k-i}}{\binom{N}{k}}$$

经过该富集分析步骤，保留每个 microRNA 靶基因集合富集的功能节点（P 值不大于 0.05），从而将基因集合转化为功能层面的功能节点集合。

2.4 集合相似性测度

集合相似性是用于衡量两个集合间相似程度的测度，主用是利用两个集合间元素的交集，并

表 1 富集分析原理表

Table 1. The principle table of enrichment analysis

	microRNA 靶基因	非 microRNA 靶基因	总数
属于某生物功能节点数据	a	b	M
不属于某生物学功能节点数据	d	d	N－M
总数	k	N－k	N

集以及集合的大小进行计算的。鉴于不同的集合相似性测度都具有某种程度上的偏向性，本文选取了 7 种经典的测度，取所有测度结果的均值作为最终的计算结果。

具体来说，a 表示两个集合间的交集元素的个数；b 表示第一个集合中包含的而第二个集合中没有的元素集合中的元素个数；c 表示第二个集合中包含的而第一个集合中没有的元素集合中的元素个数。我们采用的相似性测度如下：

(1) Simpson Coefficient：$\dfrac{a}{a+\min(b,c)}$

(2) Second Kulczynski Coefficient：

$$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

(3) Ochiai/Otsuka Coefficient：$\dfrac{a}{\sqrt{(a+b)(a+c)}}$

(4) Dice Coefficient：$\dfrac{a}{a+\dfrac{b+c}{2}}$

(5) Jaccard Coefficient：$\dfrac{a}{a+b+c}$

(6) Sokal and Sneath Coefficient：$\dfrac{a}{a+2b+2c}$

(7) First Kulczynski Coefficient：$\dfrac{a}{b+c}$

2.5　整体框架

本文所进行的分析流程主体如下(见图 1)：

(1) 计算任意两个 microRNA 间在靶基因层

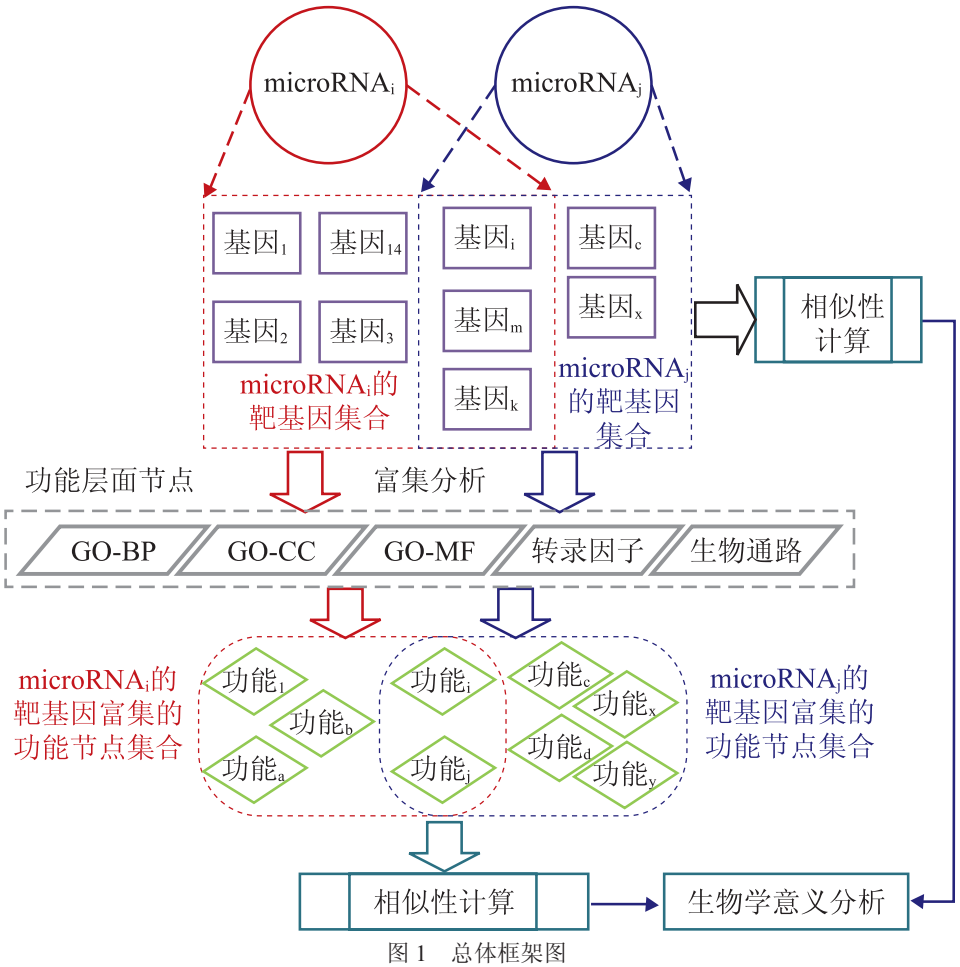

图 1　总体框架图

Fig. 1.　Flowchart of this study

面的相似性。

(2)将任意 microRNA 的靶基因通过富集分析，将其映射到 5 个功能层面(包括 GO 数据库的 3 个子库：BP、CC 和 MF，转录因子层面，生物通路层面)。

(3)计算任意两个 microRNA 的靶基因所富集的功能节点集合间的相似性，并进行生物学意义的分析。

3 结果与讨论

3.1 靶基因相似性分析

对任意两个 microRNA 计算靶基因的相似性的结果表明，microRNA 间共享的靶基因较少。其中，41.12% 的 microRNA-microRNA 对靶基因的相似性在(0,0.1]区间，43.85% 的 microRNA-microRNA 对靶基因的相似性在(0.1,0.2]区间(见图 2 所示)。相似性的平均值为 0.1243。

其中相似性最高的前 10 对 microRNA 组合为：has-miR-519a*与 has-miR-522*、has-miR-519a*与 has-miR-523*、has-miR-519b-5p 与 has-miR-519c-5p、has-miR-522*与 has-miR-523*、has-miR-320a 与 has-miR-320b、has-miR-320c 与 has-miR-320d、has-miR-320a 与 has-miR-320c、has-miR-320b 与 has-miR-320c、has-miR-320b 与 has-miR-320d。其中，前 5 组的 microRNA(has-miR-519a*、has-miR-522*、has-miR-523*、has-miR-519b-5p 和 has-miR-519c-5p)均属于 mir-515 家族；后 5 组的 microRNA(has-miR-320a、has-miR-320b、has-miR-320c、has-miR-320d)均属于 mir-320 家族。说明相同家族的 microRNA 倾向于调控相同的靶基因，且差异碱基的数目越少，越有可能调控相同的靶基因，尤其是在 microRNA“种子区域”几乎无差异碱基的情况下。

3.2 microRNA 靶基因功能相似性分析

通过富集分析，我们将每个 microRNA 的靶基因集合分别富集到 5 个功能层面，包括：GO 数据库的三个子库(BP、CC 和 MF)、生物通路(Pathways)和转录因子(TF)。选取所有 *P* 值不大于 0.05 的功能节点为显著富集的功能节点，将靶基因集合转换为功能节点集合，进而计算功能节点集合之间的相似性。

结果表明，microRNA 在不同的功能层面呈现出不同的相似性趋势(见图 3 所示)。我们将相似性计算结果等分为 10 个区间，并计算每个区间的 microRNA 对数所占总数目的比例。

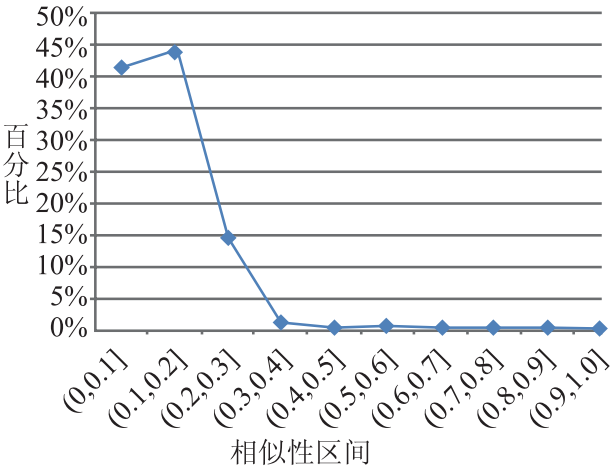

图 2 靶基因相似性结果图

Fig. 2. Results of similarity on targets level

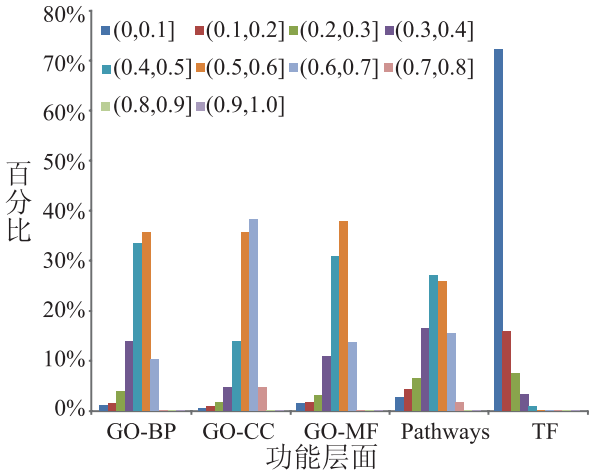

图 3 相似性统计分布图

Fig. 3. Statistic diagram of similarity

在(0,0.1]、(0.1,0.2]、(0.2,0.3]、(0.9,1.0]区间，microRNA 对所占比例的最大值的均来自基于转录因子层面的计算。这一结果说明不同的 microRNA 间倾向于不共享或完全共享靶基因。考虑到转录因子是在转录层面对其靶基因起到调控作用，而 microRNA 是在转录后层面发挥调控作用，因此二者之间的靶基因集合差异较大或完全没有交集则较为合理。有趣的是，尽管只有 0.025% 的 microRNA 对在这一层面呈现出较高的相似性，但对比其他层面仍较高(为 GO-BP 层面的 7.77 倍，GO-CC 层面的 3.97 倍)。因此，我们猜测可能存在转录因子与 microRNA 互为对方调控的情况。然而，由于当前缺乏相应的数据，在将来的工作中，如果具备相应的数据，我们将进一步深入讨论 microRNA 与转录因子间的关系问题。

16.35% 的 microRNA 对在生物学通路的相似性计算结果在(0.3,0.4]区间，但相似性结果处于此区间的 microRNA 对并不都属于相同家族或在临近的物理位置。生物学通路是某个生理过程中多个分子、基因和蛋白等之间复杂调控关系的展现，且通路间多有交互。这也就在一定程度上说明了 microRNA 间在此层面共享一定的靶基因。

与生物学通路相类似，GO-BP 数据库中的功能几点反映的是参与相同生物学过程的基因集合，而每个基因均有可能具有多种功能，故在此层面存在大量非同一家族的 microRNA 共享靶基因，即相似性值在(0.4,0.5]、(0.5,0.6]区间的 microRNA 对的最大值均在 GO-BP 层面。

在 GO-CC 层面，有 42.89% 的 microRNA对的相似性在(0.6,0.9]区间，说明受 microRNA 调控的基因在细胞组分上具有较大程度的相似性(其中，(0.6,0.7]的比例为 38.20%，(0.7,0.8]的比例为 4.66%，(0.8,0.9]的比例为 0.029%，均为与其他功能层面相比在相同区间的最大值)，即受 microRNA 靶基因与非靶基因间在细胞组分层面具有一定的区分度，这为后续的分析提供了一定的理论指导意义。

参 考 文 献

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function [J]. Cell, 2004, 116(2): 281-297.

[2] Perera RJ, Ray A. MicroRNAs in the search for understanding human diseases [J]. BioDrugs, 2007, 21(2): 97-104.

[3] Garofalo M, Quintavalle C, Di Leva G, et al. MicroRNA signatures of TRAIL resistance in human non-small cell lung cancer [J]. Oncogene, 2008, 27(27): 3845-3855.

[4] Johnnidis JB, Harris MH, Wheeler RT, et al. Regulation of progenitor cell proliferation and granulocyte function by microRNA-223 [J]. Nature, 2008, 451(7182): 1125-1129.

[5] Bhattacharyya SN, Habermacher R, Martine U, et al. Relief of microRNA-mediated translational repression in human cells subjected to stress [J]. Cell, 2006, 125(6): 1111-1124.

[6] Lovis P, Gattesco S, Regazzi R. Regulation of the expression of components of the exocytotic machinery of insulin-secreting cells by microRNAs [J]. Biological Chemistry, 2008, 389(3): 305-312.

[7] Makeyev EV, Zhang J, Carrasco MA, et al. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing [J]. Molecular Cell, 2007, 27(3): 435-448.

[8] Calin GA, Dumitru CD, Shimizu M, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia [J]. Proceedings of the National Academy of Sciences, 2002, 99(24): 15524-15529.

[9] Yanaihara N, Caplen N, Bowman E, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis [J]. Cancer Cell, 2006, 9(3): 189-198.

[10] Liao M, Jiang W, Chen X, et al. Systematic analysis of regulation and functions of co-expressed microRNAs in humans [J]. Molecular BioSystems, 2010, 6(10): 1863-1872.

[11] Ghosh Z, Chakrabarti J, Mallick B. miRNomics—The bioinformatics of microRNA genes [J]. Biochemical and Biophysical Research Communications, 2007, 363(1): 6-11.

[12] Papadopoulos GL, Reczko M, Simossis VA, et al. The database of experimentally supported targets: a functional update of TarBase [J]. Nucleic Acids Research, 2009, 37(suppl 1): D155-D158.

[13] Xiao F, Zuo Z, Cai G, et al. miRecords: an integrated resource for microRNA-target interactions [J]. Nucleic acids research, 2009, 37(suppl 1): D105-D110.

[14] Martin G, Schouest K, Kovvuru P, et al. Prediction and validation of microRNA targets in animal genomes [J]. Journal of Biosciences, 2007, 32: 1049-1052.

[15] Bentwich I. Prediction and validation of microRNAs and their targets [J]. FEBS Letters, 2005, 579(26): 5904-5910.

[16] John B, Enright A J, Aravin A, et al. Human microRNA targets [J]. PLoS Biology, 2004, 2(11): e363.

[17] Krek A, Grün D, Poy MN, et al. Combinatorial microRNA target predictions [J]. Nature Genetics, 2005, 37(5): 495-500.

[18] Lewis BP, Shih I, Jones-Rhoades MW, et al. Prediction of mammalian microRNA targets [J]. Cell, 2003, 115(7): 787-798.

[19] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets [J]. Cell, 2005, 120(1): 15-20.

[20] Wang Y, Du L, Li X, et al. Functional homogeneity in microRNA target heterogeneity—a new sight into human microRNomics [J]. Omics: a Journal of Integrative Biology, 2011, 15(1-2): 25-35.

[21] Li J, Zhang Y, Wang Y, et al. Functional combination strategy for prioritization of human miRNA target [J]. Gene, 2014, 533(1): 132-141.

[22] Xu J, Li CX, Li YS, et al. MiRNA–miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features [J]. Nucleic Acids Research, 2011, 39(3): 825-836.

[23] Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology [J]. Nature Genetics, 2000, 25(1): 25-29.

[24] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(43): 15545-15550.

[25] Matys V, Fricke E, Geffers R, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles [J]. Nucleic Acids Research, 2003, 31(1): 374-378.