生物物理学报　第二十二卷　第二期　二〇〇六年四月
ACTA BIOPHYSICA SINICA　Vol.22 No.2　Apr. 2006

基于 *k*-tuple 组合的酵母 ncRNA 与 mRNA 的比较研究

李　华，　应晓敏，　查　磊，　李伍举
（军事医学科学院基础医学研究所，北京 100850）

摘要：ncRNA 和 mRNA 一样，都是重要的功能分子。以 *k*-tuple（*k* 字）含量为特征，对酵母 ncRNA 成熟序列和 mRNA 的编码区、上游序列与下游序列进行了分类与比较研究，结果显示：基于 ncRNA 成熟序列与 mRNA 编码区的 3-tuple 的含量，ncRNA 和 mRNA 的交叉有效性分类精度（leave-one out cross-validation, LOOCV）平均值达到 93.93%；基于上游序列 4-tuple 和 5-tuple 的含量，分类精度分别为 92.49%和 92.76%；基于下游序列 4-tuple 和 5-tuple 的含量，分类精度分别为 91.58%和 90.60%；利用上游序列和下游序列的 4-tuple 与 5-tuple 的含量，其平均分类精度分别为 94.68%和 94.83%；通过 *t* 检验，得到了在 ncRNA 和 mRNA 上、下游序列中具有显著统计学差异的 *k*-tuple。上述结果表明，基于 ncRNA 成熟序列与 mRNA 编码区的 3-tuple 含量和基于 ncRNA 与 mRNA 上、下游序列的 4 或 5-tuple 含量可以有效地区分 ncRNA 与 mRNA。此研究结果不仅有助于准确识别 ncRNA 与 mRNA，还有助于发现 ncRNA 特异的转录因子结合位点。

关键词：酵母；非编码 RNA；*k* 字含量；分类
中图分类号：Q61

0　引　　言

非编码 RNA（ncRNA）是指不直接编码蛋白质的 RNA。作为一种重要的功能分子，非编码 RNA 越来越多地为人们所重视。随着大规模转录组的深入研究，人们发现被转录的基因组比例远远高于蛋白编码基因的比例。最近的研究表明，在小鼠的基因组中，转录的比例达到 62%，而转录产物中约有一半是 ncRNA[1,2]。这说明在生物体中存在大量的 ncRNA，而目前我们发现的 ncRNA 是很少的，远远没有达到如此巨大的数量[3,4]。

ncRNA 的种类很多（如表 1 所示），而且在生命活动中行使着重要的功能，如导向 RNA（gRNA）是指导 mRNA 编辑的小 RNA 分子，多用来指导在 mRNA 转录产物中加入 U 的过程[5]；微 RNA（miRNA）参与了真核基因的转录后调控，在真核生物的生长发育中起着重要的作用[6,7]；小干扰 RNA（siRNA）可以引起转录后基因沉默，它在 RNA 干扰（RNAi）中介导特异的 mRNA 的降解[8,9]；细胞核小分子 RNA（snRNA），是 mRNA 前体剪接体的必要组分；核仁小分子 RNA（snoRNA）参与 rRNA 的加工，并指导 rRNA 上特异位点的甲基化或假尿嘧啶化。近几年的研究表明，越来越多的 ncRNA 在生命活动中发挥着重要的功能[10~13]。这些 ncRNA 和 mRNA 一样，都是经过转录与加工后得到的。那么，它们在转录与加工方面究竟有什么区别呢？ncRNA 是否拥有自己特异的调控元件呢？目前，此方面的研究主要是从碱基组成的角度出发，对 ncRNA 的特征进行分析，并据此对 ncRNA 进行预测。如 Schattner 和 Eddy

Table 1　The classes of the ncRNAs

RNA	For shot
Guide RNA	gRNA
MicroRNA	miRNA
Ribosomal RNA	rRNA
Small interfering RNA	siRNA
Small non-mRNA	snmRNA
Small nuclear RNA	snRNA
Small nucleolar RNA	snoRNA
Small temporal RNA	stRNA
Transfer RNA	tRNA

收稿日期: 2006-01-20
基金项目：国家自然科学基金项目（30470411）和北京市自然科学基金项目（5042021）
通讯作者：李伍举，电话 / 传真：(010)66931324，
E-mail：liwj@nic.bmi.ac.cn

等[14,15]从碱基组成的角度出发，对基因组中的 ncRNA 进行了预测。另外，我们还注意到 Hao 等[16,17]利用 *k*-tuple 探讨了基因组水平的进化树构建问题。但目前为止，还未见到系统地利用 *k*-tuple 组合来考虑 ncRNA 与 mRNA 的差别及相关的 ncRNA 识别的研究报道。

k-tuple 是 DNA 序列中一个长度为 *k* 的核苷酸片段，也称为 *k* 字。本文以 *k*-tuple 含量为基础，以酿酒酵母（*Saccharomyces cerevisiae*）基因组为研究对象，分别对成熟 ncRNA 与 mRNA 的编码区、ncRNA 与 mRNA 的上游、下游序列进行比较研究。

1　数据和方法

1.1　数据

从 NONCODE[18]数据库中得到 189 条酿酒酵母的 ncRNA 序列，这些 ncRNA 中不包括 rRNA 和 tRNA。去除冗余序列后，得到 136 条非冗余 ncRNA 序列。在这 136 条非冗余的 ncRNA 中，有 15 条位于蛋白质编码基因的内含子区。剩余的 121 条 ncRNA 为独立转录的 ncRNA。利用 Blat[19]对这 121 条 ncRNA 进行酵母基因组定位，从而得到 ncRNA 序列上游和下游的 1 kb 的序列。

酿酒酵母的 mRNA 序列来源于 Genbank 数据库，通过去除推定（putative）基因和被剪切的基因，得到 4 058 条 mRNA 序列。利用数据库中的注释信息，得到 mRNA 编码区起始密码子上游 1 kb 和终止密码子下游 1 kb 的序列。

最后，将上述序列分别保存到 BioSun[20]软件的数据库中，便于计算。

1.2　方法

1.2.1　*k*-tuple 含量计算

运用 BioSun[20]软件计算数据库中每条序列的 *k*-tuple 含量，*k*-tuple 含量计算过程如下：对每条序列，首先考虑所有可能的 *k* 个碱基组合，如 *k*=3，则有 64 种组合，然后考虑每种组合在序列中的出现次数，最后计算每种组合出现的相对频数，即为对应的 *k*-tuple 含量。于是每条序列均对应一个向量，其元素个数为所有可能的 *k*-tuple 数目，如 *k*=3，则向量的长度为 4^3=64，基于这些向量构成的矩阵，我们进行了下面的分类与比较研究。

1.2.2　分类研究

运用 Naïve Bayes 分类方法，以交叉有效性分类精度 LOOCV[21]为目标函数，采用逐步优化方法进行变量筛选，找出具有较高分类精度的 *k*-tuple 组合，并以之为基础构建分类函数，整个分类方法由我们自行开发的 Tclass 分类系统[21]完成（http://www.biosun.org.cn/tclass）。

2　结　　果

2.1　编码区的比较结果

在这里我们将成熟的 ncRNA 和成熟 mRNA 的编码区进行比较。由于 ncRNA 和 mRNA 功能的不同，mRNA 主要通过三联体碱基编码氨基酸来行使功能，于是我们考虑通过比较 3-tuple 的含量来对 ncRNA 和 mRNA 进行分类。3-tuple 的含量由 Biosun[20]软件计算。

在这里我们利用 136 条非冗余的 ncRNA 和 4058 条 mRNA 进行比较，由于 ncRNA 序列数目远小于 mRNA 序列数目，为了避免样本数目的不对称引起分类精度的偏差，我们采用不放回抽样方法，将 4 058 条 mRNA 的编码区随机分成 29 组，每组 136 条序列，余下的 114 条 mRNA 的编码区另作一组，一共得到 30 组 mRNA 的编码区序列，分别与 136 条 ncRNA 成熟序列组成训练集，从而获得了 30 个训练集。

对于每个训练集来说，以 LOOCV 为目标函数，采用逐步优化算法进行变量选择，运用 Tclass 程序进行分类[21]，结果表明：开始时，随着变量个数的增加，分类精度逐渐提高，在变量个数增加到一定数目时，分类精度就不再提高，甚至有下降的趋势，我们将最高的分类精度作为此训练集的分类精度。具体结果如表 2 所示。结果显示，平均分类精度达到 93.93%。

2.2　上下游序列的比较结果

不管是 ncRNA 还是 mRNA，上游序列都是重要的调控区，富含各种调控元件。由于一般调控元件的核心区域为 4~5 个碱基，所以利用上游序列的 4-tuple 和 5-tuple 的含量为特征，对 ncRNA 和 mRNA 进行比较。

其次，下游序列对转录终止和调节、翻译终止和调节以及 RNA 加工修饰也有重要的调控作用，因此我们又以下游序列的 4-tuple 和 5-tuple 含量为

Table 2 The classification results

No.	Classification accuracy (%)						
	Coding region	Upstream		Downstream		Upstream and downstream	
	3-tuple	4-tuple	5-tuple	4-tuple	5-tuple	4-tuple	5-tuple
1	91.54	92.56	91.74	90.91	88.84	90.08	96.28
2	93.75	90.50	94.22	87.60	88.84	94.22	96.69
3	94.49	88.02	92.98	92.15	92.56	94.22	94.63
4	94.49	92.15	91.32	90.91	87.60	96.69	89.67
5	94.12	94.22	95.46	92.56	95.04	97.11	94.63
6	94.85	94.63	92.56	90.91	92.15	92.98	95.04
7	93.02	91.74	90.08	95.04	85.12	94.63	90.50
8	93.75	88.84	92.98	92.56	91.32	95.46	95.04
9	96.69	94.63	90.91	94.22	93.39	96.28	92.98
10	94.49	94.22	91.32	90.50	89.26	93.39	91.74
11	93.75	93.39	90.50	91.74	88.84	94.63	93.80
12	95.22	93.80	91.74	91.74	90.91	95.87	90.50
13	93.02	93.39	95.04	92.98	85.54	96.28	94.22
14	93.02	90.08	94.22	88.84	90.08	96.28	91.32
15	94.49	91.74	94.22	93.80	91.74	94.63	96.28
16	93.38	92.15	94.22	88.84	90.08	93.80	94.22
17	93.75	92.15	92.56	89.26	94.63	94.63	96.69
18	93.02	89.26	92.56	92.56	90.50	95.46	94.63
19	95.22	93.39	92.98	91.32	92.56	92.98	98.76
20	94.12	92.15	96.28	94.22	92.98	97.52	93.80
21	95.22	95.87	93.80	95.04	93.80	94.63	94.63
22	93.02	91.74	92.56	88.84	90.91	92.56	96.28
23	92.28	94.63	92.56	92.56	89.67	96.28	96.28
24	93.75	92.15	90.50	92.98	91.32	94.63	95.87
25	95.96	91.74	92.98	91.32	91.74	96.28	95.04
26	93.38	92.56	91.32	90.50	90.91	92.15	97.93
27	94.12	92.98	95.46	92.15	90.50	94.22	96.28
28	93.02	95.46	94.63	93.39	90.50	95.87	97.11
29	92.28	92.98	91.74	91.74	88.84	93.80	95.04
30	94.80	91.32	92.56	88.43	91.32	93.80	97.52
31		92.98	92.15	90.91	91.74	93.39	94.22
32		89.26	88.02	90.91	82.23	92.98	94.22
33		93.39	94.63	90.91	92.98	92.98	94.63
34		94.62	93.01	91.40	91.94	98.39	97.85
Average	93.93	92.49	92.76	91.58	90.60	94.68	94.83

The classification accuracies derived from the coding regions, upstream sequences, downstream sequences, and both upstream and downstream sequences

特征，对 ncRNA 和 mRNA 进行比较。

最后，考虑到上游序列和下游序列在 RNA 转录和加工过程中作为一个整体起作用，所以，我们将 ncRNA 与 mRNA 的上、下游序列的 4-tuple 和 5-tuple 含量组合进行比较，此时，每条序列对应的向量长度加倍，期望可以较好地将 ncRNA 和 mRNA 分开。

在这里我们利用 121 条独立转录的 ncRNA 和 4 058 条 mRNA 的上下游序列进行比较，同样为了避免样本数目的不对称引起分类精度的偏差，我们采用不放回抽样方法，将 4 058 条 mRNA 的上游序列（下游序列）随机分成 33 组，每组 121 条序列，余下的 65 条 mRNA 上游序列（下游序列）另作一组，一共得到 34 组 mRNA 上游序列（下游序列），分别与 121 条 ncRNA 上游序列（下游序列）组成训练集，从而获得了 34 个上游序列、下游序列和上下游序列组合的训练集。

对于每个训练集来说，以 LOOCV 为目标函数，采用逐步优化算法进行变量选择，运用 Tclass 程序进行分类[21]，分类精度的变化趋势与编码区的变化趋势相同。因此我们也将最高的分类精度作为此训练集的分类精度。具体结果如表 2 所示。从表 2 可以看出，上游序列 4-tuple 和 5-tuple 的平均分类精度达到 92.49%和 92.76%，下游序列 4-tuple 和 5-tuple 的平均分类精度达到 91.58%和 90.60%，同时考虑上下游序列的 4-tuple 和 5-tuple 的平均分类精度达到 94.68%和 94.83%。

2.3　*t*- 检验分析

利用 *t*- 检验对上下游序列的 4 和 5-tuple 进行分析，目的是找出 ncRNA 与 mRNA 上下游序列中有显著统计学差异的 4 和 5-tuple。通过 *t*- 检验，我们发现分别有 182 个 4-tuple 和 396 个 5-tuple 在 ncRNA 与 mRNA 上下游序列中有显著的统计学差异，具体分布见表 3。我们将 *P* 值小于 10^{-10} 的 4、5-tuple 列于表 4。

Table 3　The number of the *k*-tuples with statistically significant difference (P<0.0001)

k-tuple	Number	
	Upstream	Downstream
4-tuple	87	95
4-tuple	188	208

Table 4　The *k*-tuples with statistically significant difference

k-tuple	Upstream	Downstream
4-tuple	AAGC、ACAA、ATTG、AGAA、AGAT、AGAG、AGGC、CAAC、CACA、CCAC、CTCG、CTTG、TATC、TAGG、TCTT、TCTG、TTCG、TTTC、TTTT、TTTG、TGAA、TGGC、TGGG、GACA、GATA、GAGA、GAGC、GTTG、GGAA、GGAC、GGAG、GGCA、GGTG、GGGA	AAGC、ACAA、ACAC、AGAA、AGAC、AGGC、AGGT、CAAT、CACA、CAGC、CCAC、CTTG、CTGA、TATC、TAGG、TCAG、TCTG、TTCG、TTTC、TTTT、TTTG、TTGG、TGTC、TGGG、GAAC、GACA、GCAA、GCCC、GGAT、GGTC、GGGA
5-tuple	AAAGC、AAGAT、AAGAG、AAGCA、ACGAG、ATTTG、ATGGT、AGACC、AGATG、CAAGC、CACAA、CCAAC、CCTTG、CCGCG、CTCTG、CTTTT、CTTTG、CGCCC、CGCTC、CGGGA、TATAC、TCACA、TCTCG、TCTTG、TTCTT、TTCTG、TTTAG、TTTCG、TTTTC、TTTTT、TTTTG、TTGAA、TTGGC、TGACA、GAACT、GAAGA、GAAGC、GACGC、GATGC、GCTGA、GTATC、GTTGC、GGAGA、GGAGC、GGGAA	AACAC、AAGAC、AAGGC、ACAAT、ACACA、ATTTG、ATGGG、AGACA、AGCAA、AGCCC、AGGAT、CAAGC、CACAA、CAGAG、CAGCA、CCGAA、CTACA、CTCTG、CTTTG、CTTGG、TAATG、TATTG、TAGTG、TCACA、TCTTG、TCTGA、TTCAG、TTCTG、TTTCG、TTTTT、TTTTG、TTGTC、TGACA、TGTAC、TGGGA、TGGGC、GAAAC、GAAGC、GAGGT、GCCAC、GCTCT、GGCAC

The *k*-tuples with statistically significant difference (P <10^{-10}) in the upstream and downstream sequences of the yeast ncRNAs and the mRNAs

我们通过查找 TRANSFAC 数据库[22]发现，在这些存在显著统计学差异的 4 和 5-tuple 中，有一部分 4 和 5-tuple 为酵母转录因子结合位点的核心区域。如 CAAGC 是酵母转录因子 TEF1 结合位点的核心区域，它在 mRNA 上游序列中的含量要显著高于 ncRNA 上游序列中的含量。在这些 4 和 5-tuple 中，也有在 ncRNA 上游序列中含量要显著高于 mRNA 的，如 CCTTG、CTTTG 等。

3 讨 论

我们对编码区的比较结果显示，3-tuple 的含量可以很好地区分 ncRNA 和 mRNA，其 LOOCV 的平均分类精度可以达到 93.93%。我们也比较过编码区 4-tuple 和 5-tuple 的含量，并不能达到以 3-tuple 含量为特征的分类效果。这就与 mRNA 靠三联体碱基编码氨基酸行使功能相吻合。

在上游序列和下游序列的比较中，我们也利用了 3-tuple 含量进行分类，但是精度约为 70%，远低于 4-tuple 和 5-tuple 的平均分类精度，如基于上游序列的 4-tuple 和 5-tuple 的 LOOCV 平均分类精度达到了 92.49%和 92.76%，这表明仅利用 3-tuple 含量不足以区分 ncRNA 与 mRNA，鉴于一般调控元件的核心区域为 4 到 5 个碱基，此结果表明 ncRNA 与 mRNA 在调控元件的利用方式上存在明显差别。另外，同时考虑上下游序列的 4-tuple 和 5-tuple 的 LOOCV 平均分类精度达到 94.68%和 94.83%，稍高于单独比较上游序列或下游序列的精度，这一点与 RNA 的转录和加工过程可能需要上游序列和下游序列中的调控元件同时作用的观点相一致，表明 ncRNA 和 mRNA 是通过上下游调控区同时作用的结果。

最后，我们在 ncRNA 与 mRNA 上游和下游序列中找出了含量存在显著统计学差异的 4-tuple 和 5-tuple。这些 4 和 5-tuple 可能是转录因子的结合位点，或是一些剪切信号。由于在酵母中存在大量未知的转录因子[23]，这些转录因子可能参与了 ncRNA 的转录过程，我们找出的这些 4 和 5-tuple 对未知转录因子的识别和对这些转录因子功能的研究具有一定的辅助作用。

综上所述，本文通过利用 k-tuple 的含量对酵母的 ncRNA 和 mRNA 进行了比较，结果表明，基于 k-tuple 组合可以有效地区分 ncRNA 与 mRNA，这一结果可以运用于 ncRNA 的预测。我们的工作还有助于发现 ncRNA 特异的调控元件，从而有助于探讨 ncRNA 基因的表达调控规律。

参考文献：

[1] Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engstrom PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui S, Liang Z, Lenhard B, Wahlestedt C. Antisense transcription in the mammalian transcriptome. *Science*, 2005,309(5740):1564~1566

[2] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SPT, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CAM, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo

S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. The transcriptional landscape of the mammalian genome. *Science*, 2005,309(5740):1559~1563

[3] Claverie JM. Fewer genes, more noncoding RNA. *Science*, 2005,309(5740):1529~1530

[4] Kulkarni OC, Vigneshwar R, Jayaraman VK, Kulkarni BD. Identification of coding and non-coding sequences using local Holder exponent formalism. *Bioinformatics*, 2005,21 (20): 3818~3823

[5] Kable ML, Seiwert SD, Heidmann S, Stuart K. RNA editing: a mechanism for gRNA-specified uridylate insertion into precursor mRNA. *Science*, 1996,273(5279):1189~1195

[6] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. *Cell*, 2003, 115(7):787~798

[7] Croce CM, Calin GA. miRNAs, cancer, and stem cell division. *Cell*, 2005,122(1):6~7

[8] Ahlquist P. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, 2002,296(5571):1270~1273

[9] Filipowicz W. RNAi: the nuts and bolts of the RISC machine. *Cell*, 2005,122(1):17~20

[10] Eddy SR. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2001,2(12):919~929

[11] Eddy SR. Noncoding RNA genes. *Curr Opin Genet Dev*, 1999,9(6):695~699

[12] Mattick JS. The functional genomics of noncoding RNA. *Science*, 2005,309(5740):1527~1528

[13] Mattick JS. RNA regulation: a new genetics? *Nat Rev Genet*, 2004,5(4):316~323

[14] Eddy SR. Computational genomics of noncoding RNA genes. *Cell*, 2002,109(2):137~140

[15] Schattner P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res*, 2002,30(9):2076~2082

[16] Liu C, Bai B, Skogerbo G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*, 2005,33 (1):D112~D115

[17] Hao BL, Qi J. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *J Bioinformatics and Computational Biology*, 2004,2(1):1~19

[18] Qi J, Luo H, Hao BL. CVTree: a phylogenetic tree reconstruction tool based on whole gemomes. *Nucleic Acids Res*, 2004,32(Web Server issue):W45-W47

[19] Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*, 2002,12(4):656~664

[20] 李伍举，应晓敏. Biosun：计算机辅助分子生物学实验的软件系统. 军事医学科学院院刊, 2004,28(5):401~405

[21] Li WJ, Xiong MM. Tclass: tumor classification system based on gene expression profile. *Bioinformatics*, 2002,18(2): 325~326

[22] Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 1996,24(1):238~241

[23] Peng WT, Robinson MD, Mnaimneh S, Krogan NJ, Cagney G, Morris Q, Davierwala AP, Grigull J, Yang X, Zhang W, Mitsakakis N, Ryan OW, Datta N, Jojic V, Pal C, Canadien V, Richards D, Beattie B, Wu LF, Altschuler SJ, Roweis S, Frey BJ, Emili A, Greenblatt JF, Hughes TR. A panoramic view of yeast noncoding RNA processing. *Cell*, 2003,113(7): 919~933

THE COMPARISON OF THE YEAST ncRNA AND mRNA BASED ON *k*-tuple COMBINATIONS

LI Hua, YING Xiao-min, ZHA Lei, LI Wu-ju
(*Beijing Institute of Basic Medical Sciences, Beijing 100850, China*)

Abstract: Both ncRNAs and mRNAs are important functional molecules. In this report, the Naïve Baye's classification method has been used to classify ncRNA and mRNA based on their *k*-tuple content, composition of coding regions, upstream sequences and downstream sequences, and the vector concatenation of both upstream and downstream sequences. The results show that the average leave-one-out cross-validation (LOOCV) classification accuracy is 93.93% for 3-tuple content in ncRNA sequences and mRNA coding regions. In upstream 1 kb sequences of the ncRNAs and mRNAs for 4-tuple and 5-tuple content, the classification accuracies are 92.49% and 92.76%, respectively. For the downstream 1 kb sequences of the ncRNAs and mRNAs, the classification accuracy are 91.58% and 90.60% for 4-tuple and 5-tuple content, respectively. For the vector concatenation of upstream and downstream sequences, the average classification accuracies are 94.68% and 94.83% for 4-tuple and 5-tuple content respectively. Finally, the *t*-test has been used to verify if *k*-tuples are statistically different between the upstream and downstream sequences of the ncRNAs and mRNAs. The approach may be used to identify ncRNAs from other genomes and to identify binding motifs within ncRNA sequences.

Key Words: Yeast; ncRNA; *k*-tuple content; Classification

This work was supported by grants from The National Natural Sciences Foundation of China (30470411) and The Natural Sciences Foundation of Beijing (5042021)

Received: Jan 20, 2006

Correspoding author: LI Wu-ju, Tel: +86(10)66931324, E-mail: liwj@nic.bmi.ac.cn