生物物理学报　第二十一卷　第四期　二〇〇五年八月
ACTA BIOPHYSICA SINICA　Vol.21 No.4　Aug. 2005

酵母基因上游区与内含子可能的短程和长程转录协同增效作用

张昆林[2,1]，　张　静[1]，　罗静初[2]
（1. 云南大学统计系，云南大学应用统计中心，昆明 650091；2. 北京大学生物信息中心，北京大学生命科学学院，北京大学蛋白质工程和植物基因工程国家重点实验室，北京 100871）

摘要：根据实验观察到的 DNA 成环和弯折机制，以 140 bp 为分界点，探讨高频转录基因上游区与内含子之间可能存在的短程和长程转录协同增效作用（synergy）。用与随机序列做对比的方法，抽提出最近距离在 140 bp 以下的寡核苷酸对，以及最近距离在 140 bp 以上的寡核苷酸对。仔细分析两种距离下的可能的协同寡核苷酸对的位置特征和碱基组分，发现短程协同作用的寡核苷酸对的平均最近距离都在 110 bp 以下，位于上游区的 CCAA 是一个很明显的特征；而长程协同作用的寡核苷酸对的平均最近距离集中在 250~400 bp，并且在多数寡核苷酸对中，位于上游区的寡核苷酸是 GC 丰富的正调控元件。

关键词：转录调控；协同增效作用；寡核苷酸对；内含子；酵母基因
学科分类号：Q61

1 引　　言

在真核基因的表达调控中，转录因子之间或调控位点之间往往会有协同作用（cooperativity or synergy），并且协同作用可以增强或抑制转录效率[1]。对 DNA 而言，两位点产生协同作用的机制比较典型的是 DNA 成环（looping）或 DNA 弯折（bending）。较早观察到的 DNA 成环机制中两个调控位点的距离比较靠近，一般来说大概不超过 8 个 DNA 整数螺距的范围（84 bp，一个 DNA 螺距大约是 10.5 bp）。后来也有报导说可以达到 20 个 DNA 螺距（210 bp）；此时位点之间的距离可以是整数螺距，也可以不是[2,3]。位点距离在 140 bp 范围内的协同作用还有以 DNA 弯折的机制进行的，但由于 DNA 刚性的限制，在此长度范围内的 DNA 弯折需要借助某些蛋白质的作用才能形成[4]。当 DNA 长度超过 140 bp 时，DNA 的弯折也许比较容易实现。此外，在较远距离上，还有一些别的协同作用机制，如 DNA 追踪（tracking），DNA 伸展 / 成环（spreading/looping）以及 DNA 超卷曲（supercoiling）等[5]。

目前对转录位点或转录因子间组合调控（combinatorial control）的研究较多集中在基因上游[6]。我们前期的工作发现酵母高频转录基因的内含子中存在一些可能的转录正调控位点，并且内含子的位置比较偏向基因的 5' 端，而低频转录的基因内含子却不具有这种明显的特征[7,8]，由此推测高频转录基因的内含子可能与基因上游区的调控有协同增效作用[8]。通过与随机序列作比较，我们曾抽提出一些在 84 bp 范围近程匹配基因数显著高于随机匹配数的寡核苷酸对（一个寡核苷酸位于上游区，另一个位于内含子中），探测到几个非随机出现的寡核苷酸对，并对这些寡核苷酸对的相互作用模式作了一些分析[10]。实际上，这个简单的分析是基于 DNA 成环的机制。为了更全面地了解酵母高频转录基因上游区和内含子之间的转录协同增效作用，还需要对不同距离上调控位点的协同作用做比较全面的探索。本文中我们以 140 bp 作为分界，考虑转录因子作用位点的最近距离分别在≤140 bp（短程）和 >140 bp（长程）范围内有协同增效作用的寡核苷酸对。结果表明以 140 bp 作为长、短程作用的分界点是有意义的。在 140 bp 的范围内，所抽提出的四核苷酸对包含 84 bp 范围内抽提出的大部分四核苷酸对，上游寡核苷酸含 A、T 较多，且 CCAA 仍是上游的一个典型信号。而在

收稿日期：2005-04-04
基金项目：国家自然科学基金项目（30360027），973 项目（2003CB715900），863 项目
通讯作者：张静，电话：(0871)6541419，
E-mail：zhangjing@ynu.edu.cn

>140 bp 的范围内，未抽提出 CCAA 这个片段，上游序列富含 G、C。 由此看来，以 140 bp 为分界的短程和长程范围上的协同作用机制可能有所不同，这是值得注意的现象。

2 材料和方法

2.1 材料

酵母基因组是目前实验研究较多的对象，含有内含子的基因不多，大约只有 234 个。对这些含内含子的基因，实验分析（microroarry analysis）提供了充分的基因转录表达数据。与前期关于协同作用的研究[10]一致，仍用其中的 67 个高频转录基因（转录频率 >30 mRNA/hr）和 71 个低频转录基因（转录频率 ≤10 mRNA/hr）（YIDB, http://www.imb-jena.de/RNA.html）作为分析样本。本文的研究主要是探索内含子与上游调控区的协同增效作用，所以低频转录基因只是作为抽提转录正调控元件的对照样本。起初我们仅是凭直觉选取 30 和 10 作为高低转录频率的域值，后来发现这些高频转录的基因几乎都是核糖体蛋白编码基因，而低频转录的基因没有为核糖体蛋白编码的。这说明我们这样选取的样本有很好的生物学意义，所得结果反映的是核糖体蛋白基因的共性。所选的两类基因序列均包含上游区。前期研究还得到了高频转录基因上游区和内含子中的潜在转录正调控元件，本文将直接利用这些结果。此外，考虑到上游区的一般转录调控元件也可能和内含子中的潜在转录正调控元件有协同作用，这部分调控元件通过对比高频转录基因的上游区和酵母基因组的非编码区得到。所有转录调控元件都与 TRANSFAC 中的“检验位点”[10]做过对比，已经知道它们是否可能与 3 种转录因子 RAP 1、ABF 1 和 TAF（分别简记为 R、A 和 T）结合。

2.2 方法

2.2.1 近程匹配和远程匹配

本文主要研究高频转录基因上游区和内含子的协同增效作用，我们的分析主要基于这样的思想：协同作用的位点对是非随机共出现的。对于一对寡核苷酸或转录位点（$M1$, $M2$），其中 $M1$ 位于上游区，$M2$ 位于内含子，我们沿用最近距离（即距离外显子最近的 $M1$ 和距离外显子最近的 $M2$ 之间的碱基数）的概念来定义它们的距离[10,11]。为了描述寡核苷酸对的相关范围，这里引入“近程匹配”和“远程匹配”的概念：如果一条基因上游区有 $M1$ 且内含子中有 $M2$，且当（$M1$,$M2$）的最近距离小于等于 140 bp 时，称（$M1$,$M2$）近程匹配（short-range match）；当（$M1$,$M2$）的最近距离大于 140 bp，但是小于等于 1 kb 时，称（$M1$,$M2$）远程匹配（long-range match）。

在参考文献[9]中，我们打乱（shuffling）67 条高转录基因中每一条基因上游区的碱基顺序，同时也打乱每一条基因内含子的碱基顺序，得到一个随机序列组，这个随机序列组的序列数仍然是 67；重复这样的操作 N 次，就得到 N 个随机序列组，且每个组包含 67 条序列[9]。现在，我们用同样的方法构造随机序列组，并且在这些随机序列组的序列上以相同的方式定义近程匹配和远程匹配。进一步，对于原高频转录基因序列组 S 和一个随机序列组 S^*，设（$M1$,$M2$）在 S 中近程匹配的序列数为 u，远程匹配的序列数为 v（$u+v \leqslant 67$）；设（$M1$, $M2$）在 S^* 中近程匹配的序列数为 u^*，远程匹配的序列数为 v^*($u^*+v^* \leqslant 67$)。如果 $u>u^*$ 且 $v<v^*$，则称 S 近程匹配一致优于 S^*；如果 $u<u^*$ 且 $v>v^*$，则称 S 远程匹配一致优于 S^*。S 近程匹配一致优于 S^*，表示相对于 S^*，（$M1$,$M2$）在 S 上的最近距离偏向 140 bp 以下；S 远程匹配一致优于 S^*，则表示相对于 S^*，（$M1$,$M2$）在 S 上的最近距离偏向 140 bp 以上。近程匹配一致优于和远程匹配一致优于对固定的 S^* 来说是互不相容的。

2.2.2 提取两类最近距离有显著偏向性的寡核苷酸对

具有转录协同作用的寡核苷酸对应具有明显非随机性特征，即这些寡核苷对近程 / 远程匹配的高频转录基因数应该明显地不同于随机匹配数。这里同样用与随机序列作比较的方法来提取对应于转录协同作用的寡核苷酸对。记寡核苷酸对为（$M1$, $M2$），高频转录基因组为 S，N 个随机序列组为 S_i (i=1,2,…,N)。设 IM=#({S_i|S 近程匹配一致优于 S_i, i=1,2,…,N})，OM=#({S_i|S 远程匹配一致优于 S_i, i=1,2,…,N})。其中“#”表示对集合求元素的个数，IM 和 OM 分别表示 S 近程匹配一致优于的随机序列组的个数和 S 远程匹配一致优于的随机序列组的个数。对于短程协同作用，要检验（$M1$, $M2$）近程匹配的高频转录基因数是否显著一致高于近程匹配的随机序列数，设置检验值 P=1–IM/N；而对应于长程协同作用，要检验（$M1$,$M2$）远程匹配的高频转录基因数是否显著一致高于远程匹配的随机

序列数，检验值 $P=1-OM/N$。对于近程匹配，我们取 $P<0.05$，即对（$M1$,$M2$）来说，要求 S 近程匹配一致优于 95% 以上的随机序列组，才认为（$M1$, $M2$）对应于短程协同作用；不过对于远程匹配的情况，鉴于碱基数目较大，核苷酸在打乱了的随机序列中重现的可能性还是比较大，所以对于远程匹配的情形，我们将 P 值取得更小一些（<0.02 或 0.01），以保证结果的可信度。为了减少随机涨落，我们同样取 N=30、100 和 1 000，要求在这三种情况下，P 值都要小于所确定的阈值，并取 N=1 000 时的 P 值作为检验的 P 值。

抽提出有显著距离偏向性的两类寡核苷酸对之后，我们给出了这些寡核苷酸对在序列上的位置分布信息，如最近距离的平均值和标准差等，以此来研究短程协同作用和长程协同作用的 DNA 序列结构特征。我们对四、五核苷酸对都做了上述分析。

3 结　　果

3.1 短程协同作用的寡核苷酸对

对于短程协同作用，$P<0.05$ 时可能的四核苷酸对见表 1，五核苷酸对见表 2。表中每一行中都列出了潜在的协同作用（短程协同作用）寡核苷酸对、统计检验的 P 值、近程匹配的高频转录基因数、位于上游区的潜在调控位点到翻译起始点的距离平均值、寡核苷酸对的最近距离的平均值、最近距离的标准差、以及可能与潜在调控位点结合的转录因子。所有这些信息揭示了协同作用的主要模式。与参考文献[10]的结果相比，我们新增加了最近距离的标准差，这一指标反映最近距离分布的离散程度。

Table 1 Potential synergistic tetranucleotide pairs within shortrange and related information (nearest distance ≤140 bp, $P<0.05$)

*M*1	*M*2	*P*	*s_match*	*up_mean*	*dis_mean*	*std*	F1	F2
CCAA	TAAA	0.006	34	31	71	32	R/T	R/A
	AATT	0.006	27	31	87	34		A
	TTCA	0.001	31	28	85	32		R/A/T
	TATT	0	36	34	89	28		R/A/T
	TGGT	0.029	17	25	96	31		R/A/T
	GAAT	0.003	28	29	81	30		R/A/T
	CACG	0.019	14	22	82	23		R/A/T
GAAA	TAAA	0.022	42	34	78	34	R/A/T	R/A
	AAAT	0.003	45	30	83	31		R/A
	TATT	0.002	44	34	85	30		R/A/T
	TTAT	0.001	44	33	83	27		R/A
	ATAT	0.026	32	33	89	34		R/A/T
	GAAT	0.016	34	33	83	28		R/A/T
	TTCA	0.002	38	28	81	30		R/A/T
TGAA	TTCA	0.005	36	41	91	26	R/A/T	R/A/T
	TTAT	0.013	42	43	92	30		R/A
TGTA	TTCA	0.012	34	40	94	33	R/A/T	R/A/T

The bold denotes positive regulatory motif. "*M*1" and "*M*2" denote the motifs in upstream regions and introns respectively. "*s_match*" denotes the number of matched genes within 140 bp range. "*up_mean*" represents the average of the distances between *M*1 and translation start site. "*dis_mean*" represents the average of the nearest distances between *M*1 and *M*2 and "*std*" is the standard deviation. "F1" and "F2" represent the transcription factors which may bind to motif *M*1 and *M*2, respectively. "R", "A" and "T" represent RAP1, ABF1 and TAF, respectively. Similar notes in hereinafter

Table 2 Potential synergistic pentanucleotide pairs with shortrange and related information (nearest distance ≤140 bp, P<0.05)

*M*1	*M*2	*P*	*s_match*	*up_mean*	*dis_mean*	*std*	F1	F2
ACCAA	TTAAA	0.012	11	20	57	28	R/T	-
	TATTT	0.016	14	20	81	21		R/A
	TTTAA	0.012	11	20	57	28		-
	AAATA	0.016	14	20	81	21		R/A
	ATAGT	0.019	8	28	80	32		A
ATACC	AATAT	0.037	10	47	97	23	R	R/A/T
	ATATT	0.037	10	47	97	23		R/A/T
AAATT	ATAGT	0.023	12	46	104	27	A	A

虽然我们将近程匹配的最近距离限制在140 bp 以下，但从表 1 和表 2 可以看出，最近距离的平均值都在 110 bp 以下。最近距离的标准差都很小，均在 35 bp 以下，表明最近距离的分布比较集中。就碱基组分来说，表 1 的结果包含了参考文献[10]中大部分的四核苷酸对。上游区的四核苷酸 CCAA 值得注意，它在 84 bp 范围的分析中也得到过。在整个高转录基因序列组上游区中，CCAA 共出现 313 次，分布于 62 条高频转录基因的上游区，平均每条基因上游区有 5 个。在五核苷酸的分析中，也探测到了调控元件 ACCAA（表 2）。显然，CCAA 是具短程协同作用的寡核苷酸对在上游区的一个显著特征。它可能就是 CCAAT box 的构件。事实上，CCAA 的调控功能已经被实验所证实[12,13]。此外，从 *up_mean* 值看，上游的几个寡核苷酸在序列中的位置呈现出一定的分布特征，CCAA 距离翻译起始位点最近；TGAA 和 TGTA 稍远一些，这两个寡核苷酸很相似，有共有序列 TGWA。这样的位置特征可能与转录调控机制有关。从表 1 和表 2 还可以看出，五核苷酸的结果多为四核苷酸的延伸。

3.2 长程协同作用的寡核苷酸对

表 3 列出了对应于长程协同作用的四核苷酸对，这里我们取 P<0.01，即要求 S 远程匹配一致优于 99% 以上的随机序列组。从表 3 可以看出，这些长程协同作用的寡核苷酸对的最近距离的平均值集中在 250~400 bp 范围内；并且最近距离的标准差较大，表明调控位点分布较为分散。此外还有一个值得注意的现象是在长程协同作用的寡核苷酸对中，上游的元件主要是 GC 丰富的转录正调控元件，他们距离翻译起始位点的距离多在 200 bp 以上。这似乎也提示，内含子与基因上游之间的长程协同作用在协同调控中占有较大的优势，而且这些长程协同的作用是以增强基因转录效率为目的的。

在 P<0.01 的阈值范围内，没有得到五核苷酸对的结果，这主要是因为五核苷酸在打乱了的随机序列中重组的机会比起四核苷酸来要小得多。不过我们得到了 3 个 P 值稍大一些（<0.05）的结果，见表 4。

3.3 两种协同作用的比较与联系

对应于两种协同作用的寡核苷酸对的上游元件的组分有较明显的差异。短程协同作用的上游元件富含 AT，而长程协同作用的则富含 GC。表 3 只列出了 P<0.01 的结果，事实上，在 P 值稍大一些（P<0.015）的五核苷酸对中，长程协同作用的上游元件也是富含 GC 的，例如 GCGC、GGCC、GAGG 等。从内含子来看，其中的转录正调控元件多数可参与两种协同作用，典型的有 TAAA、GAAT、ATAT 等，而且一些元件在序列中通过连接或重叠形成较长的序列片断，如 TAAA 和 AAAT 与 AATT 形成 TAAAATT，等等。内含子中的这种调控模块（module）与上游区有可能同时发生着两种协同作用。

Table 3 Potential synergistic tetranucleotide pairs within longrange and related information (nearest distance >140 bp, P<0.01)

*M*1	*M*2	*P*	*l_match*	*up_mean*	*dis_mean*	*std*	F1	F2
CGGC	TAAA	0	52	201	266	95	A	R/A
	AAAT	0	52	201	264	88		R/A
	AATT	0	53	199	302	127		A
	TTAT	0	51	202	276	103		R/A
	TATT	0.004	50	203	269	95		R/A/T
	ATTA	0.001	50	203	291	119		R/A/T
	TTCA	0.006	51	202	280	88		R/A/T
	TATC	0.009	51	202	329	137		A
	GAAT	0	53	199	294	105		R/A/T
ACCC	TAAA	0.004	50	233	300	103	R/A	R/A
	AAAT	0.003	50	235	305	99		R/A
	AATT	0.006	53	225	338	119		A
	TATT	0.007	50	235	309	97		R/A/T
	TTAT	0.009	50	235	315	104		R/A
	ATTA	0.006	50	235	332	127		R/A/T
	ATAT	0.006	54	217	331	113		R/A/T
	TATC	0	55	219	355	122		A
TCCG	TATC	0.003	55	207	335	142	R/A/T	A
TGGA	AAAT	0.009	46	176	251	86	R/A/T	R/A
CGGG	TAAA	0.002	47	226	293	91	R/A	R/A
	AAAT	0.001	47	226	294	93		R/A
	TTAT	0.005	47	225	303	95		R/A
	ATTA	0.003	47	225	322	113		R/A/T
	TTCA	0.002	49	220	296	95		R/A/T
	GAAT	0.008	49	219	321	115		R/A/T
CGCT	TATC	0.008	56	189	314	142	R/A	A

"*l_match*" denotes the number of genes matched by the motif pair within long range. Similarly in **Table 4**.

Table 4 Potential synergistic pentanucleotide pairs within longrange and related information (nearest distance >140 bp, P<0.05)

*M*1	*M*2	*P*	*l_match*	*up_mean*	*dis_mean*	*std*	F1	F2
GCCTA	TAAAA	0.044	38	281	429	199	-	A
	ATTAT	0.048	38	281	413	169	-	A
	ATAAT	0.048	38	281	413	169	-	A

为了直观地了解两类寡核苷酸对在高频转录基因中的分布情况，我们将抽提出的寡核苷酸对在基因序列上标注出来。图 1 显示了 YBL087C 的结果。这条基因上两类寡核苷酸对都有。我们用大写字母标注可能的调控元件，用带下划线的黑体大写字母标注了符合距离条件的两个四核苷酸对（为了简明起见，我们只标注了两个四核苷酸对）。这两个四核苷酸对中的 CCAA 和 TAAA 对应于短程协同作用，最近距离为 69 bp，而 CGGG 和 TAAA 对应于长程协同作用，最近距离为 222 bp。两个四核苷酸对在内含子中的转录正调控元件都是 TAAA。

…(−178)**CGGG**^{L}…(−25)**CCAA**^{S}cACACC…[42]…(+48)T**TAAA**$^{L\&S}$ATTTT…

Fig.1 The illustrations of motifs and potential synergistic motif pairs in YBL087C. Square bracket and the bold number in it denote the 1st exon and its length. Motifs are denoted by upper cases. The number in a round bracket indicates the position of the base on the right of the bracket. The position of the first base of the 1st exon (i.e. the translation start site) is "+1", and its adjacent position upstream is "−1". The underlined bold bases are two synergistic motif pairs. The symbols (L or S) on the shoulders of motifs indicate the types of synergy (longrange or shortrange).

虽然短程和长程协同作用对的上游序列有些差异，但总的说来，协同作用对的上游寡核苷酸 GC 含量较高，而内含子寡核苷酸则主要由 AT 构成，而且这个结果多是非随机的。因为从碱基组成上看，尽管高转录基因上游序列与内含子的 GC 含量有些差异，分别为 38%和 33%，而相应的 AT 含量分别为 62%和 67%。如果这些基因上游和内含子的序列是完全随机的，仅仅 5%的碱基含量差异不会造成如此大的寡核苷酸组成差异。这也说明上游和内含子在序列的组织上是非随机的；协同作用对的寡核苷酸上下游碱基的这种互补性也许与参与协同作用的转录因子的作用机制有关。

4 讨 论

真核基因的内含子中存在转录调控位点的现象已被越来越多的实验所证实[14]。一些实验还特别表明，内含子与基因上游调控有协同调控的作用[15]，但是对这种协同作用的机制还缺乏认识。所以系统研究内含子在基因转录调控中的作用以及如何起作用，对于理解真核基因的调控机理是一项很重要的工作。我们前期的研究亦表明酵母内含子很可能参与了基因转录调控[7-9]，至少调控转录的效率。本文及参考文献[10]对高频转录基因调控位点序列特征的分析进一步提示，内含子对转录的调控与基因上游的调控有一定程度关联性。而且得到了一些可能的协同调控元件组织信息。

此外从得到的结果来看，对应于短程协同作用的寡核苷酸对的平均最近距离都不到 110 bp（最长是 104 bp）；而对应于长程协同作用的寡核苷酸对的平均最近距离都大于 250 bp（最短是 251 bp）。因此，在 110~250 bp 范围内短程和长程协同作用都不明显，这表明两种协同作用在距离上是有显著区别的。同时，对应于两种不同的协同作用的两类寡核苷酸对，它们位于上游区的调控元件明显不同，并且具有各自的特征。这也说明我们以 140 bp 做为短程协同作用和长程协同作用的分界点是有意义的。

真核基因上游调控元件中一般都有一个 TATA-box，然而酵母核糖体基因是个例外。酵母核糖体基因上游调控元件中缺乏 TATA-box，其相应的功能主要由 Rap1 因子的激活子来完成[16]。在我们对酵母高频转录基因（即核糖体基因）的分析中（无论是上游还是内含子），都未抽提出具有统计显著性的 TATA 元件，而是探测到较多的 Rap1 等因子结合位点，这也说明我们的方法是有效的。

基因转录调控是一个复杂的问题，基因序列中的任何区域（包括外显子）都可能含有转录调控位点，事实上我们在这些基因的外显子中也探测到调控位点，协同作用在基因的任何部位的调控位点间都可能发生。本文主要考虑内含子与基因上游的协同增效（即提高转录效率）作用关系，目的是要认识内含子在转录调控方面的功能性特征。对于这些基因全面的复杂调控网络关系的研究正在进行中。

参考文献：

[1] 薛 文，王 进，黄启来，郑伟娟，华子春. 真核基因转录激活的多位点协同调控. 生物化学与生物物理进展，2002,29(4): 510~513

[2] Griffith J, Hochschild A, Ptashne M. DNA loops induced by cooperative binding of λ repressor. *Nature*, 1986,322(6081):

750~752

[3] Hochschild A, Ptashne M. Cooperative binding of λ repressors to sites separated by integral turns of the DNA helix. *Cell*, 1986,44(5):681~687

[4] Carey M, Smale ST. Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001. 38~41, 466~468

[5] Bondarenko VA, Liu YV, Jiang YI, Studitsky VM. Communication over a large distance: enhancers and insulators. *Biochem Cell Biol*, 2003,81(3):241~251

[6] Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 2001,29:153~159

[7] 张 静，石秀凡. 酵母基因中转录正调控内含子序列特征的统计分析. 生物化学与生物物理进展，2003,30(2):231~238

[8] Zhang J, Hu J, Shi XF, Cao H, Liu WB. Detection of potential positive regulatory motifs of transcription in yeast introns by comparative analysis of oligonucleotide frequencies. *Comput Biol Chem*, 2003,27(4~5):497~506

[9] 张 静，石秀凡，杨恒芬. 酵母内含子在基因序列中的分布对基因转录效率的影响. 生物化学与生物物理进展，2003,30(6):945~949

[10] 张昆林，张 静，罗静初. 酵母基因上游与内含子可能存在的转录协同作用. 生物化学与生物物理进展，2005,32(1):46~52

[11] Konopka AK. Biomolecular sequence analysis: pattern acquisition and frequency counts. In: Cooper DN, ed. Nature encyclopedia of the human genome. Vol. 1. London: Nature Publishing Group Reference, 2003. 311~322

[12] Ren Z, Jin H, Whitby PW, Morton DJ, Stull TL. Role of CCAA nucleotide repeats in regulation of hemoglobin and hemoglobin~haptoglobin binding protein genes of Haemophilus influenzae. *J Bacteriol*, 1999,18(18): 5865~5870

[13] Mencía M, Moqtaderi Z, Geisberg JV, Kuras L, Struhl1 K. Activator-specific recruitment of TFIID and regulation of ribosomal protein genes in yeast. *Mol Cell*, 2002,9:823~833

[14] Bhattacharyya N, Banerjee D. Transcriptional regulatory sequences within the first intron of the chicken apolipoprotein AI (apo AI) gene. *Gene*, 1999,234(2):371~380

[15] Schjerven H, Brandtzaeg P, Johansen FE. A novel NF-kappa B/Rel site in intron 1 cooperates with proximal promoter elements to mediate TNF-alpha-induced transcription of the human polymeric Ig receptor. *J Immunol*, 2001,167 (11): 6412~6420

[16] Hartatik T, Okada S, Okabe S, Arima M, Hatano M, Tokuhisa T. Binding of BAZF and Bc16 to STAT6~binding DNA sequences. *Biochem Biophys Res Commun*, 2001,284(1): 26~32

POSSIBLE SHORT- AND LONG-RANGE TRANSCRIPTIONAL SYNERGISTIC REGULATION BETWEEN UPSTREAM REGIONS AND INTRONS IN YEAST GENES

ZHANG Kun-lin[2,1], ZHANG Jing[1], LUO Jing-chu[2]

(*1. Department of Statistics, The Center of Applied Statistics, Yunnan University, Kunming 650091, China;*
2. Center of Bioinformatics, College of Life Science and National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, China)

Abstract: Based on the mechanism of DNA looping and bending obtained in experiment and with the cutoff of 140 bp, the short and long range potential transcriptional synergy in highly-transcribed yeast genes was studied. By comparing with random sequences, tetra- and penta-nucleotide pairs whose nearest distance were within the range of 140 bp and over 140 bp were extracted. The position distributions and base constitutions of these extracted oligonucleotide pairs were analysed, the result showed that the average nearest distances were less than 110 bp for the short-range synergistic oligonucleotide pairs and CCAA was a noticeable upstream instance. For the long-range synergistic oligonucleotide pairs, the average nearest distances were within the range from 250 bp to 400 bp and majority of the oligonucleotides located upstream regions were GC-rich positive regulatory elements.

Key Words: Transcriptional regulation; Synergy; Oligonucleotide pairs; Intron; Yeast