2017 年 全国数学建模竞赛

北航赛区 A 题

miRNA 调控靶基因模型

李子睿 15231030 18810725311 lzrwork@qq.com

梁心雨 15231028 15001091942 872890039@qq.com

陈天羽 14031027 13020036267 610349348@qq.com

miRNA 调控靶基因模型

摘要

RNA 分子对生物活动有重要的调控作用，其中 miRNA 是典型的一类，通过与靶基因相互作用调控生物活动，因而研究 miRNA-靶基因的识别方式成为热点。实验表明其配对方式由许多相关因素的影响，呈现出复杂结构。本文以统计方法分析实验已验证的 miRNA-靶基因组合数据，分析其配对方式是否是碱基配对，再利用动态规划算法计算碱基对匹配进行验证。并基于实验已有结论，根据基本结合规则，对公开的 miRNA-靶基因数据库进一步分析，用 SPSS 建立了多元线性回归模型，实现预测 miRNA 的靶基因。

关键词：miRNA； 靶基因；预测；多元回归；动态规划；

1　问题重述

研究表明，RNA 分子对生物活动有重要的调控作用，其中 miRNA 是有较稳定长度的一类，大约 22 个核苷酸左右，我们希望探究这样的长度的碱基配对是否足以确定对应的靶基因，及二者的识别方式有何规律。

我们将①查阅 miRNA 及其靶基因的相关资料，建立数学模型，分析其配对方式是否是由碱基配对决定。②根据查阅的基因、miRNA 数据库，建立 miRNA 与对应靶基因间的识别模型，并预测模型准确性，做误差分析。

此题的解题关键是，①寻找合适的算法计算碱基对的匹配，②合理的定义参量描述相关因素，通过与实验数据拟合得出较好的数学模型。相关因素有：miRNA-靶基因互补性，结合双链的热稳定性，miRNA 5′ 端与靶基因的结合能力强于 3′ 端，miRNA 与 3′ UTR 作用的自由能。

从第一个靶基因预测软件 miRanda 到现在，相继出现了十余种预测软件，可大致分为第一、二代。第一代预测软件主要从种子互补这一规则出发，结合 miRNA 靶基因跨物种间保守性设计算法；第二代主要以机器学习方法训练参数。（参考文献[1]）

本文将①以统计方法分析实验已验证的 miRNA-靶基因组合数据，分析其配对方式是否是碱基配对，再利用动态规划算法计算碱基对匹配进行验证。②通过对 miRNA-靶基因数据库进一步分析，建立多元线性回归模型。

2　假设与符号说明

2.1 假设

1. 所得实验数据可靠性良好，无异常值。

   由于所用公开数据库为核心数据库，且使用数据选取及其保守，只选取了经过强实验验证的部分，我们认为数据可信度高，不引入误差。

2. 数据严格符合基本结合原则：①miRNA 与靶基因的互补性；②miRNA-mRNA 双链之间的热稳定性；③miRNA 5′ 端于靶基因的结合能力强于 3′ 端；④miRNA 与靶基因结合处不应有复杂二级结构；

   基本结合原则是经典假设（参考文献[2]），也是大量实验总结的结果，具有合理性。

2.2 符号定义

x_1：假设成熟 miRNA 片段共由a_1个氨基酸组成。miRNA 与靶位点有a_2个互补，则记$x_1 = \frac{a_2}{a_1}$。故有：$0 \le x_1 \le 1$。

x_2：若 miRNA 与靶基因的 5′ 端结合，则记$x_2 = 1$；若 miRNA 与靶基因的 3′ 端结合，则记$x_2 = 0$；

x_3：由于因为 GC 之间三个氢键，而 AU 之间只有两个氢键，氢键越多两条链之间的吸引力就越大，双链结构就越不容易被高温破坏。故我们定义 GC 结合的数量为 b，AU 结合的数量为 c，则$x_3 = 3b + 2c$；

x'_3：标准化后的x_3，取值在[0,1]。

x_4：用 Vienna 软件包计算 miRNA 与 3′ UTR 作用的自由能(ΔG)，x_4=ΔG；

x'_4：标准化后的x_4，取值在[0,1]。

y:是对应靶基因的可能性。为x_1, x_2, x'_3, x'_4的线性函数，取值在[0,1]。

3　模型的建立

3.1 初步分析判断

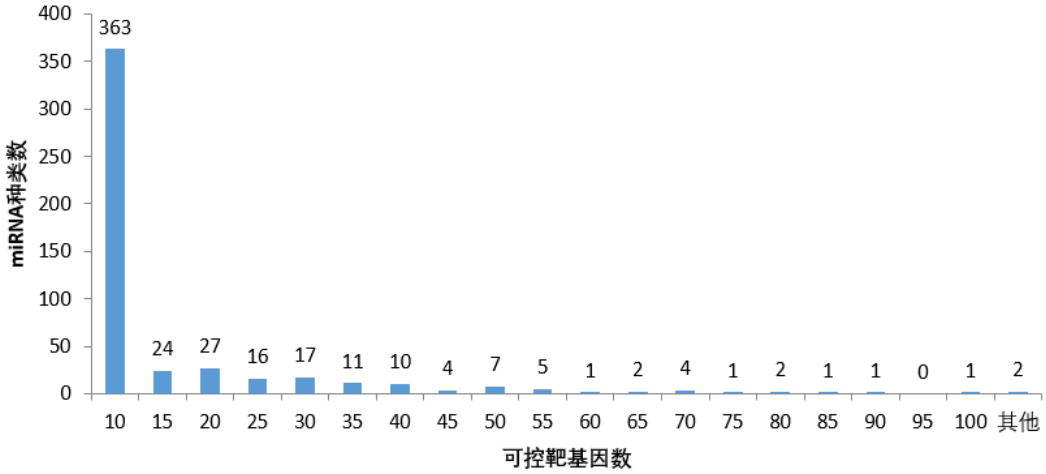
1.　数据描述

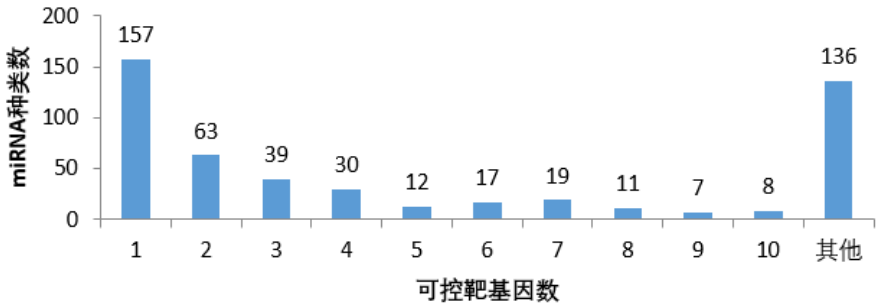
经查阅、筛选资料，我们获得了 5492 个有效可靠的 miRNA-靶基因结合数据，做了如下统计分析。

我们对数据中共 499 种 miRNA 可调控的靶基因数量进行了统计，结果如下：

	平均	中位数	众数	最小	最大
可调控靶基因数	11	3	1	1	205

以可控靶基因数为横坐标，相应 miRNA 种类数为纵坐标，作图如下：



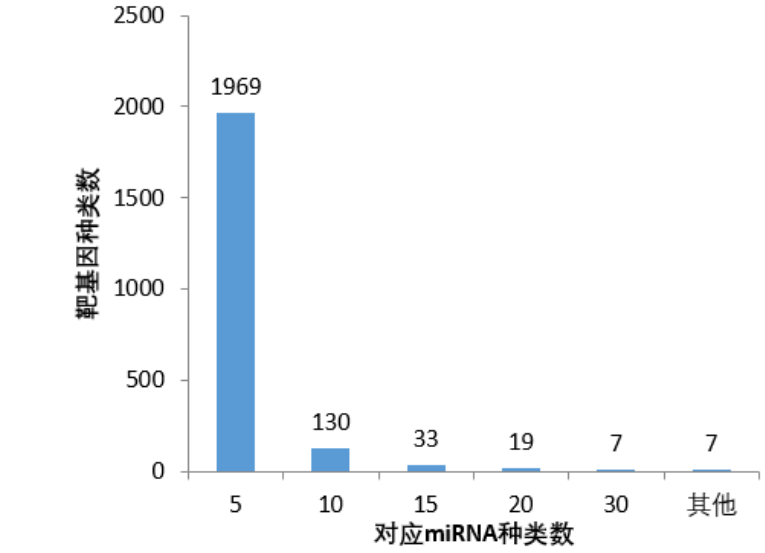


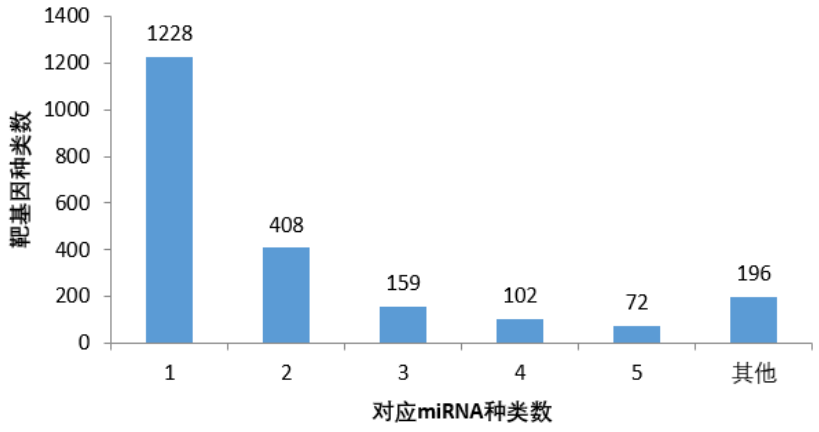
由图表可知，一种 miRNA 可调控多种靶基因，且可达到 200 多种；72.7%集中在 10 种以内，57.9%集中在 4 种以内，其中可调控靶基因为 1 种的最多，占 31.5%。

我们对 2615 种靶基因对应的 miRNA 数同样进行了统计，结果如下：

	平均	中位数	众数	最小	最大
miRNA 种类数	2.5	1	1	1	51

以对应 miRNA 种类数为横坐标，相应靶基因种类数为纵坐标，作图如下：





由图表可知，一种靶基因可由多种 miRNA 调控，可达 51 种；75.3%集中在 5 种以内，62.6%集中在 2 种以内，其中可调控靶基因为 1 种的最多，占 47.0%。

综合看来，miRNA-靶基因的配对很可能不是单纯的碱基配对原则，因此我们以动态规划算法计算了经典的 Watson-Crick 碱基对匹配。

2. 动态规划算法计算

具体算法：

Algorithm 1 Watson-Crick Matching Algorithm

Input: $miRNA$, probable $gene$
Output: saved matched sequence

LCS-LENGTH($miRNA, gene$)

```
1: m ← miRNA.length
2: n ← gene.length
3: let b[1..m, 1..n] and c[0..m, 0..n]be new tables
4: for i ← 1 to m do
5:    c[i, 0] ← 0
6: end for
7: for j ← 0 to n do
8:    c[0, j] ← 0
9: end for
10: for i ← 1 to m do
11:    for j ← 1 to n do
12:       if miRNA[i] == "U" and gene[i] == "A"
          or miRNA[i] == "A" and gene[i] == "T"
          or miRNA[i] == "G" and gene[i] == "C"
          or miRNA[i] == "C" and gene[i] == "G" then
13:          c[i, j] ← c[i − 1, j − 1] + 1
14:          b[i, j] ← " ↖ "
15:       else if c[i − 1, j] ≥ c[i, j − 1] then
16:          c[i, j] ← c[i − 1, j]
17:          b[i, j] ← " ↑ "
18:       else
19:          c[i, j] ← c[i, j − 1]
20:          b[i, j] ← " ⟵ "
21:       end if
22:    end for
23: end for
24: return c and b
```

SAVE-LCS($b, miRNA, i, j$)

```
1: if i == 0 or j == 0 then
2:    return
3: end if
4: if b[i, j] == " ↖ " then
5:    SAVE-LCS(b, miRNA, i − 1, j − 1)
6:    save miRNA[i]
7: else if b[i, j] == " ↑ " then
8:    SAVE-LCS(b, miRNA, i − 1, j)
9: else
10:   SAVE-LCS(b, miRNA, i, j − 1)
11: end if
```

计算结果如下图：

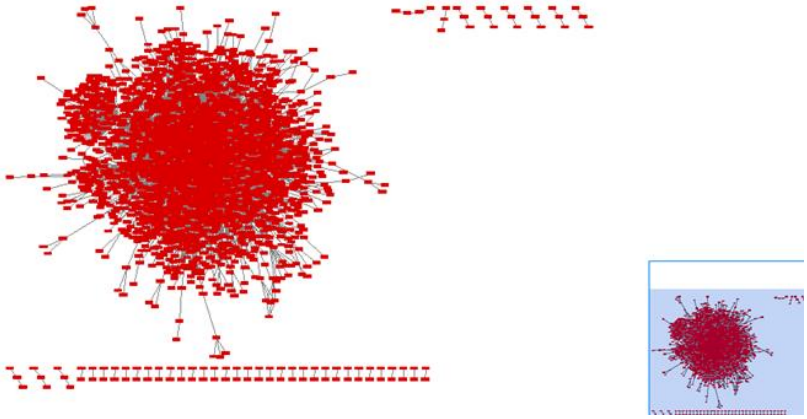
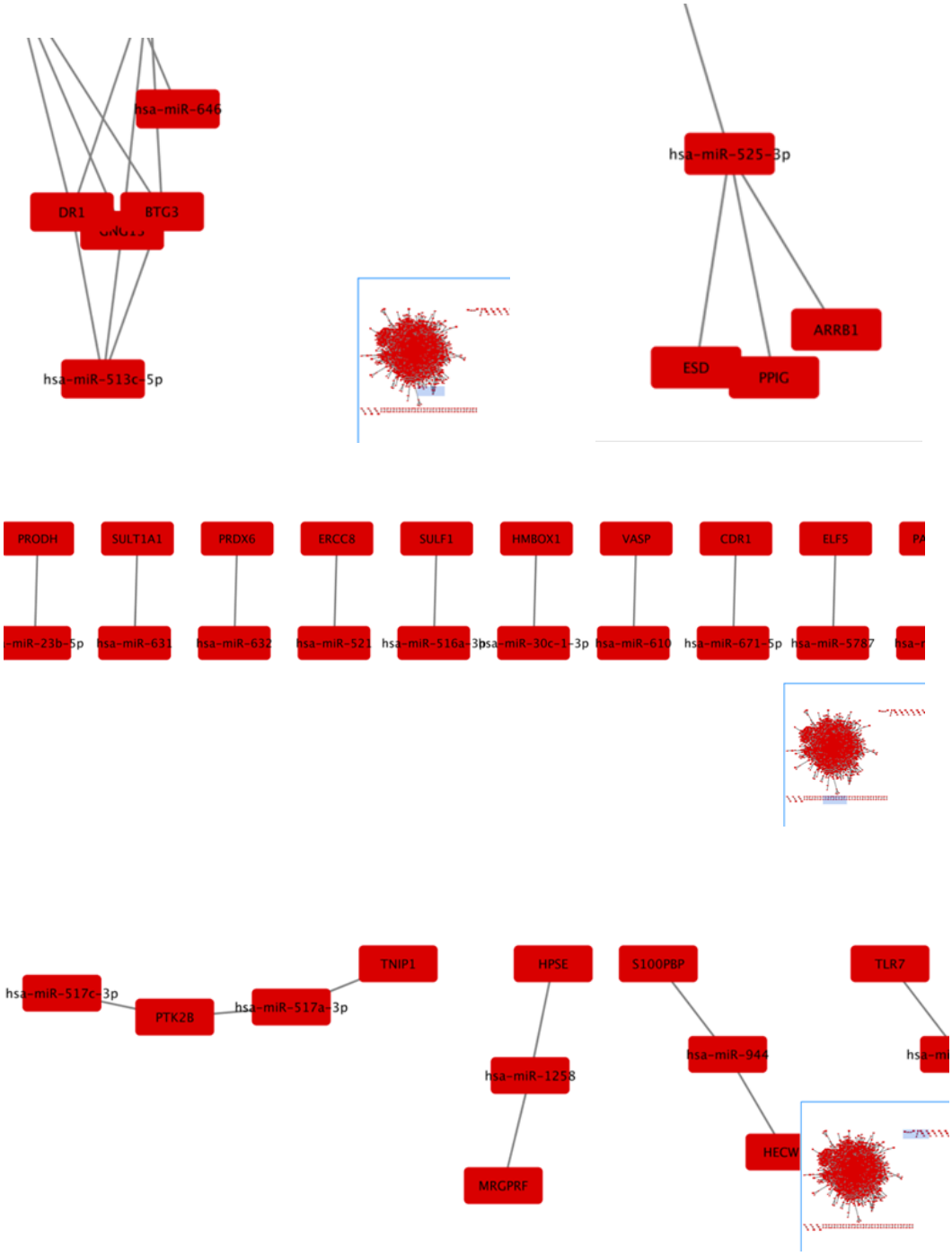


图 1-1

局部放大图如下，其中带“has-”前缀的为 miRNA，其余为靶基因：



从图中可以看出，存在大量多个 miRNA 和多个基因进行配对的情况，除此之外，也确实有仅靠简单的 Watson-Crick 碱基对匹配就成功完成一一对应的节点，以及多个 miRNA 对应一个基因和一个基因对应多个 miRNA 的情况。

然后，通过比较 miRNA 相对的 3’UTR 以及 miRNA 与靶基因结合的自由能大小，写出改进的匹配算法如下。该算法使用贪心算法筛选每条 miRNA 相对的 3’UTR 中得分排名前十位的基因作为 miRNA 的候选靶基因；同时考虑到多个

miRNA 对应于同一靶基因的情况，利用贪心算法筛选出自由能最低且得分最高的一对以完成匹配。

改进匹配算法：

Algorithm 2 Improved Matching Algorithm

Input: k miRNAs used as gene probes whose probable genes are saved in arrays $m_1, m_2, ...m_k$

Output: miRNA target gene array

```
for i ∈ [1, k] do
    sort m_i by its 3'UTR score in descending order
    g_i[1, ..., 10] ← m_i[1, ..., 10]
    g_i.value ← m_i.score
end for
sort g_1, ..., g_k by g_i.value in descending order, i ∈ [1, k]
p ← 1
for i ∈ [1, k − 1] do
    if compare(g_i, g_{i+1}) then
        G[p].add(g_i, g_{i+1})
    else
        p ← p + 1
    end if
end for
i ← 1
while i ≤ p do
    for all g_j such that g_j ∈ G[i] do
        if g_j.value > g_{j+1}.value then
            remove g_{j+1}
            G[i].e ← g_j.freeEnergy
        end if
        j ← j + 1
    end for
    i ← i + 1
end while
sort G by G[i].e in increasing order, i ∈ [1, p]
return G.head
```

结果如下：

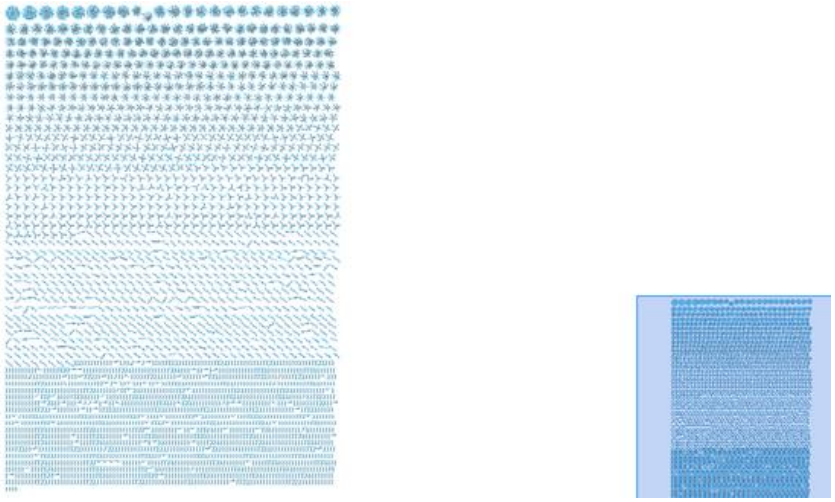
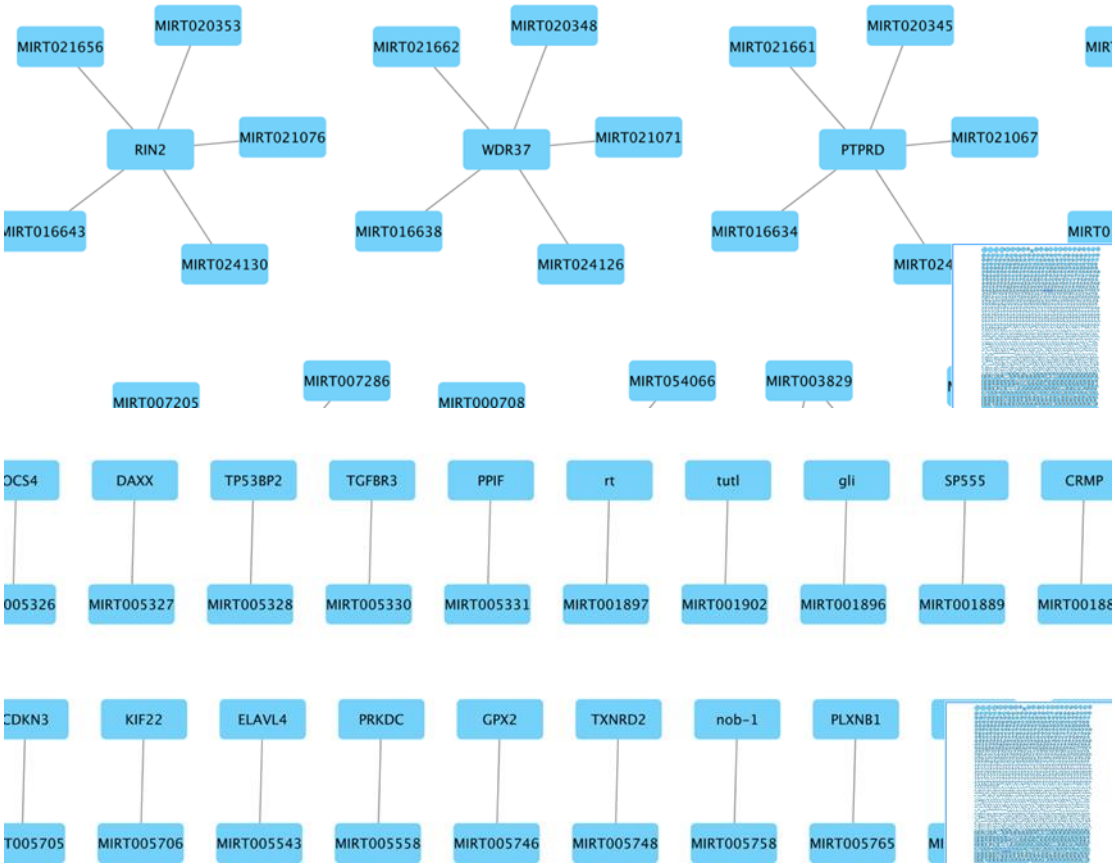


图 2-1

可以看图 2-1 比图 1-1 规整有序多了，少了图 1-1 中一个巨大而杂乱的“互配团”，取而代之的是大量简化后的 miRNA 和基因的匹配对。既有靠碱基互配一一对应的组合，又有多个 miRNA 对应一个基因的情况。

局部放大图，其中有“MIR”前缀的为 miRNA，其余为靶基因：



从图中可以更清晰直观的看到，miRNA-靶基因的配对可以一对一，一对多，多对一，极有可能不依赖碱基互补。

3. 结合靶位的碱基配对情况

由上述分析，不妨取 hsa-miR-93（miRNA）与 PTEN（靶基因）的配位为例，查询其具体配对信息，得到如下结果：

hsa-miR-93/PTEN Alignment

```
    3' gaUGGACGUGCUUGUCGUGAAAc 5' hsa-miR-93
         |::  :|  :|::|||||||
257:5' ggAUUAAUAAAGAUGGCACUUUc 3' PTEN
```

由图，显然其中有非碱基配对部分的存在。

4. 结论

miRNA-靶基因的配对不是单纯的碱基配对原则，还有部分结合等其他配对方式。

3.2 建立配对模型

1. 参量归一化

由于x_1到x_4四个变量的变化范围不一致，故需要将变量进行标准化至[0,1]，需要标准化的变量为x_3和x_4。具体操作如下：

取已知 miRNA-靶基因对中x_3、x_4的最大值与最小值，求出极差Δx_3，Δx_4，则得到：

$$x'_3 = x_3/\Delta x_3;$$

$$x'_4 = x_4/\Delta x_4;$$

最终得到x_1、x_2、x'_3、x'_4四个变量，取值范围均为[0,1]，令这四个变量为自变量，定义一个范围为[0,1]的分值 y 为因变量，来衡量某基因为 miRNA 的靶基因的可能性。

2. 多元回归求解

经过试验验证的人类 miRNA-靶基因对共 5000 个左右，我们选择在 targetscan 上查询该 miRNA-靶基因对的匹配程度（以 score 表示，score 越高，binding 越强），将该匹配程度标准化至[0,1],即得到因变量 y。

从 microRNA.org 可以查询得到x_1、x_2、x_3的值，我们就此得到了 5000 组数据。我们采用多元回归模型来预测 y，有：

$$\hat{y} = \widehat{b_1}x_1 + \widehat{b_2}x_2 + \widehat{b_3}x_3 + \widehat{b_4}x_4$$

用 spss 完成多元回归的过程，得到$\widehat{b_1}$到$\widehat{b_4}$的估计值为：0.83264；0.38754；0.64287；0.87394；

F-value=0.176345 检验通过。证明不能拒绝该假设，即模型符合度较好。由此得到打分模型：

$$\hat{y}=0.83264x_1+0.38754x_2z+0.64287x_3+0.87394x_4$$

3. 模型接受区间

需要确定 y 为何值时认为该基因为 miRNA 的靶基因。计算 5000 组数据 y 的平均值$\bar{y}$=0.738642 和标准差 σ =0.081253，得到置信度为 95%的区间为：

（0.604575,0.872709），由于 y 越接近 1，越可能为靶基因，故只要 y>0.604575，便可以认为该基因为 miRNA 的靶基因。

4　模型分析

本模型的优点是，使用的样本数据是经过试验验证得到的可靠度较高的数据，得到的回归方程可信度较高，但与此同时，由于我们是通过前人的研究得到的可能自变量，因此很有可能考虑的不够全面，即很多影响因素没有列入考虑内，这是缺点之一，此外，当使用该模型进行预测的时候，需要全体人类 mRNA x_1到x_4的变量数据，每一个都进行计算对应的 y 值，再筛选出 y>0.604575 的基因，可以认为得到的结果为预测 miRNA 的靶基因。这种算法所耗时间过长，针对不同的 miRNA 需要计算上万个数据，这是本模型的第二个缺点。

5 参考文献

[1]茹松伟,申卫红,杨鹏程,赵屹,邵启祥. microRNA 靶基因预测算法研究概况及发展趋势[J]. 生命科学,2007,(05):562-567.

[2]夏伟,曹国军,邵宁生. MicroRNA 靶基因的寻找及鉴定方法研究进展[J]. 中国科学(C 辑:生命科学),2009,(01):121-128.

附录

数据来源：

http://mirtarbase.mbc.nctu.edu.tw/php/download.php

http://www.targetscan.org/vert_71/

http://www.microrna.org/microrna/home.do