01Targetscan、microT、miRanda、miRTarget、Pictar和PITA

miRNA靶基因进行优选的方法TarpriGO

这些方法大多基于种子位点的序列匹配、结合位点的自由能以及位点保守性等特征,然而这些特征还不足以对靶基因进行准确识别和预测,所得结果往往具有较高的假阳性。

一种miRNA分子可调控多种基因的表达,同一个基因也可能被多种miRNA调控,这种复杂而精确的调控关系构成了miRNA-靶基因调控网络

本文以与药物成瘾相关的基因及其靶基因为例,构建并分析miRNA-疾病调控网络与miRNA-靶基因调控网络。

02 算法（专利）

<https://www.google.com/patents/CN105808976A?cl=zh>

03 格式

# FASTA序列格式详细说明

Posted on [**2009 年 6 月 23 日**](http://boyun.sh.cn/bio/?p=1192)

fasta序列格式是blast组织数据的基本格式，无论是数据库还是查询序列，大多数情况都使用fasta序列格式，所以首先对fasta格式在做详细说明。

下面是一个来源于NCBI的fasta格式序列：

>gi|187608668|ref|NM\_001043364.2| Bombyx mori moricin (Mor), mRNA

AAACCGCGCAGTTATTTAAAATATGAATATTTTAAAACTTTTCTTTGTTTTTA

TTGTGGCAATGTCTCTGGTGTCATGTAGTACAGCCGCTCCAGCAAAAATACCT

ATCAAGGCCATTAAGACTGTAGGAAAGGCAGTCGGTAAAGGTCTAAGAGCCAT

CAATATCGCCAGTACAGCCAACGATGTTTTCAATTTCTTGAAACCGAAGAAAA

GAAAGCATTAAGAAAAGAAATTGAGTGAATGGTATTAGATATATTACTAAAGG

ATCGATCACAATGATATATAGATAGGTCATAGATGTCAACGTGAATTTATGGA

TTTTTGTTTTCCCCTTTGTAGTACTTACTTATAGTCAGTTCTTAAATTGATTG

CAACGACAACTGTGTACTATTTTTTATATTTGGTTCGAAAAGTTGCATTATTA

ACGATTTTAGAAAATAAAACTACTTTACTTTTACACG

Fasta格式首先以大于号“>”开头，接着是序列的标识符“gi|187608668|ref|NM\_001043364.2|”，然后是序列的描述信息。换行后是序列信息，序列中允许空格，换行，空行，直到下一个大于号，表示该序列的结束。

所有来源于NCBI的序列都有一个gi号“gi|gi\_identifier”，gi号类似与数据库中的流水号，由数字组成，具有绝对唯一性。一条核酸或者蛋白质改变了，将赋予一个新的gi号（这时序列的接收号可能不变）。

gi号后面是序列的标识符，下表是来源于不同数据库的标识符的说明。标识符由序列来源标识、序列标识（如接收号、名称等）等几部分组成，他们之间用“|”隔开，如果某项缺失，可以留空但是“|”不能省略。如上例中标识符为“ref|NM\_001043364.2|”，表示序列来源于NCBI的参考序列库，接收号为“NM\_001043364.2”。

表：序列来源的数据库与对应的标识符

|  |  |
| --- | --- |
| Database Name数据库名称 | Identifier Syntax 标识符 |
| GenBank | gb|accession|locus |
| EMBL Data Library | emb|accession|locus |
| DDBJ, DNA Database of Japan | dbj|accession|locus |
| NBRF PIR | pir||entry |
| Protein Research Foundation | prf||name |
| SWISS-PROT | sp|accession|entry name |
| Brookhaven Protein Data Bank | pdb|entry|chain |
| Patents | pat|country|number |
| GenInfo Backbone Id | bbs|number |
| General database identifier | gnl|database|identifier |
| NCBI Reference Sequence | ref|accession|locus |
| Local Sequence identifier | lcl|identifier |

对于自己构建的序列数据库（序列不是来源与NCBI或其他数据），可以采用“gnl|database|identifier”或者“lcl|identifier”格式，以保证可以使用blast的所有功能。database或者identifier是需要指定的数据库的标识和序列标识，指定的名称可以用大小写字母、数字、下划线“\_”、破折号“-”或者点号“.”。注意名称是区分大小写的，同时不能出现空格，空格表示序列标识符结束。

数据库中的序列标识符必须保证唯一，许多时候格式数据库是formatdb报告错误，就是因为标示符重复，还有一点需要强调的是序列不能为空，否则也会报错。

下面是一个例子，这四个序列的标识符都是唯一。

* gnl|H.sapiens|seq1
* gnl|H.sapiens|seq2
* gnl|M.Mus|seq1
* lcl|seq1

eg1:  
>gi|156623022|dbj|AB006912.1| Bombyx mori T15.180a gene for sex-linked proline rich protein, partial cds

>gi|149243361|pdb|2PQU|G Chain G, Crystal Structure Of Kh1 Domain Of Human Pcbp2 Complexed To Single-Stranded 12-Mer Telomeric Dna