G-四联体识别工具介绍和对比

# 摘要

DNA的二级结构除了典型的右手双螺旋结构以外，还有其他特殊的二级结构，其中G-四联体结构因其在基因表达调控中的重要作用而受到广泛的关注。G-四联体的检测方法主要分为实验方法检测和计算方法预测，本文重点在于总结现有的常用G-四联体预测工具，分析其工作原理并且对其进行比较。现有的预测工具大多是采用motif的方式进行预测，少量时利用序列的GC含量来进行打分辅助预测，除此之外2Dmeet4G将G-四联体放置于长序列当中，通过对整个序列的二级结构分析来预测G-四联体的形成。随后本文还利用G4RNA数据库，将部分工具在同一批数据上处理比较，发现cGcC计分方法与实验数据结果拟合较好，较优于QGRS打分方法。

目录

[摘要 2](#_Toc491853052)

[一、介绍： 3](#_Toc491853053)

[二、算法介绍： 4](#_Toc491853054)

[1、Quad\_parser: 4](#_Toc491853055)

[2、QGRS mapper： 5](#_Toc491853056)

[3、G4P calculator： 6](#_Toc491853057)

[4、G4-Hunter： 7](#_Toc491853058)

[5、cGcC： 8](#_Toc491853059)

[6、2D meets 4G： 8](#_Toc491853060)

[7、G4RNA： 9](#_Toc491853061)

[三、G4RNA简单分析： 10](#_Toc491853062)

[四、软件对比和总结： 12](#_Toc491853063)

[五、参考文献 14](#_Toc491853064)

# 一、介绍：

1910年，I.Bang发现只有高浓度下的鸟嘌呤才会形成胶体。随后在1962年M. Gellert发现四个鸟嘌呤可以自组装成氢键链接的平面，就是后来所称为的G-四分体，如图1A。G-四分体在阳离子驱动下通过堆叠形成了四链体结构，也就是G-四联体结构。G-四联体具有拓扑结构多样性，根据参与形成G-四联体的DNA链的数目分为单分子、双分子和四分子结构；根据DNA链的方向分为平行、反平行和平行-反平行结构，如图1B。

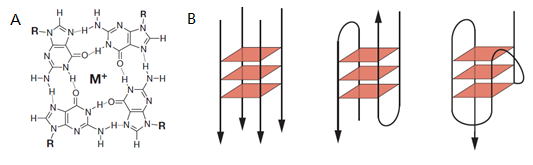


图1：A：G-四分体化学分子式 B：G-四联体三维结构

G-四联体结构具有稳定的热力学结构，而且在基因组中的功能区域如端粒区，启动子区等也发现G-四联体的形成(D. J. Patel, 2007; J. L. Huppert, 2007)。在端粒中，G-四联体的形成会抑制端粒酶的活性，有85%的癌症发生于端粒酶活性上升有关，因此G-四联体也有一定的药学研究价值(L. Oganesian and T. M. Bryan, 2007)。G-四联体在基因启动子区时对基因的转录活性有一定调控作用，如原癌基因c-kit(S. Rankin, 2005)，在体外实验中发现G-四联体在不同情况下调控结果也各不相同。在5‘-UTR区的G-四联体会降低翻译活性，还发现G-四联体会和许多蛋白质特定结合，而且也在许多体内实验中发现了他们的存在。

除了许多实验工作外，还有许多计算方法用来寻找可能形成G-四联体的序列，尽管有一些算法被广泛采纳和使用，但仍有许多计算方法被完善和创新。但是并没有足够的证据能够证明这些算法和程序是完全正确的，现在许多算法也不再预测一段序列是否能形成G-四联体，而是判断这段序列所形成的G-四联体的稳定性。在这里我们将会介绍七种计算工具或算法，并对他们进行了简单的比较。

# 二、算法介绍：

## 1、Quad\_parser:

Quad\_parser[1]算法主要考虑四个方面：核苷酸链的数目，G-四分体堆叠个数，G-四联体中的变异和loop段的长度和位置。算法可以找到能形成G-四联体的序列，但是这些序列也可能形成其他的二级结构。

核苷酸链的数目：虽然参与组成的G-四联体的核苷酸链可以是一条，两条甚至是三条，但是由于绝大多数在生理条件下，DNA链的浓度较低，所以算法不考虑多链组合形成的G-四联体。

G-四分体堆叠个数：理论上G-四联体中可以包含任意个数的G-四分体，一般来说G-四联体的稳定性随G-四分体个数增加而增加只有在浓度高达mM浓度下的鸟嘌呤溶液能够形成单G-四分体的G-四联体，因此生理条件下不能形成。所以一般认为在单层和双层条件下G-四联体稳定性不足，因此在算法中只考虑三层及以上的G-四联体。

G-四联体中的变异：许多研究表明G-四联体不一定需要在G-四分体中全都是鸟嘌呤，但是绝大多数这样的序列都是人为设计出来的，一个包含变异的G-四分体需要由其他正常四分体来稳定，也有研究发现了一些具有滑动结构的G-四联体，即两个G-四联体相对滑动，亦或是一层G-四分体由多段序列的鸟嘌呤组成，但是这些序列也都是人为构造的。G-四联体中如果出现变异，无论是插入删除还是替换都会显著降低G-四联体的稳定性，因此在算法中不考虑G-四联体的变异。

Loop长度和位置：单分子G-四联体必须由loop来连接G-tetrads(连续的鸟嘌呤片段)。这些loop在G-四联体稳定性和G-四联体折叠方式上的影响极大。根据分子建模和生物技术检测，loop长度在1-7的会有助于形成G-四联体，并随着loop增加稳定性会有所下降。理论上G-四联体可以形成长度大于等于8的loop，但是由于稳定性的降低在算法中不会考虑这些特殊情况。

基于以上四个方面认为序列为会在接近生理条件下形成G-四联体。其中N是指包含鸟嘌呤在内的任意的碱基，生理条件是指pH为7.4的100mM的KCl。

## 2、QGRS mapper：

QGRS mapper[2]程序用于寻找可能形成G-四联体的序列。潜在的G-四联体序列的motif被定义为，在式中x是指G-tetrads的鸟嘌呤数目，y1, y2, y3是指三段loop的长度，潜在的G-四联体包含四段相同长度的G-tetrads和三段任意碱基组成的loop。

其中：，即G-tetrads长度大于1。虽然一般认为3个及以上长度的G-tetrads认为是稳定的G-四联体，但是也有许多序列形成了长度为2的G-tetrads。QGRS Mapper是一个灵活的工具，所以也考虑G-tetrads长度为2的情况。

默认情况下QGRS Mapper只寻找最大长度为30bp的序列，但是最大可以设置到45bp。30bp长度下的G-四联体会自动约束G-tetrads的最大长度为6。

在序列总长度的约束下，G-四联体中的loop可以是任意长度。当然可以通过设置使得loop处于一定的长度范围内，例如可以将长度约束在1-7内。也可以指定一段序列，使得找到的G-四联体都包含这样一段序列，但是输入时这段序列需要以正则表达式的形式输入。例如输入T{3，5}时，会寻找包含一段或多段3到5个连续的胸腺嘧啶的G-四联体序列。除此之外，程序允许其中一段loop长度为0。

QGRS mapper还提出了一个打分方法G-scores来描述序列形成稳定G-四联体的可能性。分数越高表示可能行越大，整体的打分原则基于以下三点：1、loop越短分数越高；2、G-四联体倾向于拥有相同长度的loop；3、G-tetrads长度越长G-四联体越稳定。

QGRS mapper会在寻找G-四联体时默认消除重复的G-四联体。由于序列中G-四联体相似行很高，在一段序列中可能形成多种G-四联体，这个时候就认为这些G-四联体是重复的，因为这些G-四联体占据了同一段序列，在实际中，这些G-四联体只会由一段G-四联体形成，因此QGRS mapper默认消除那些G-score较低的G-四联体，只显示在这些序列中G-score最高的那条序列。也可以通过设置显示所有序列。

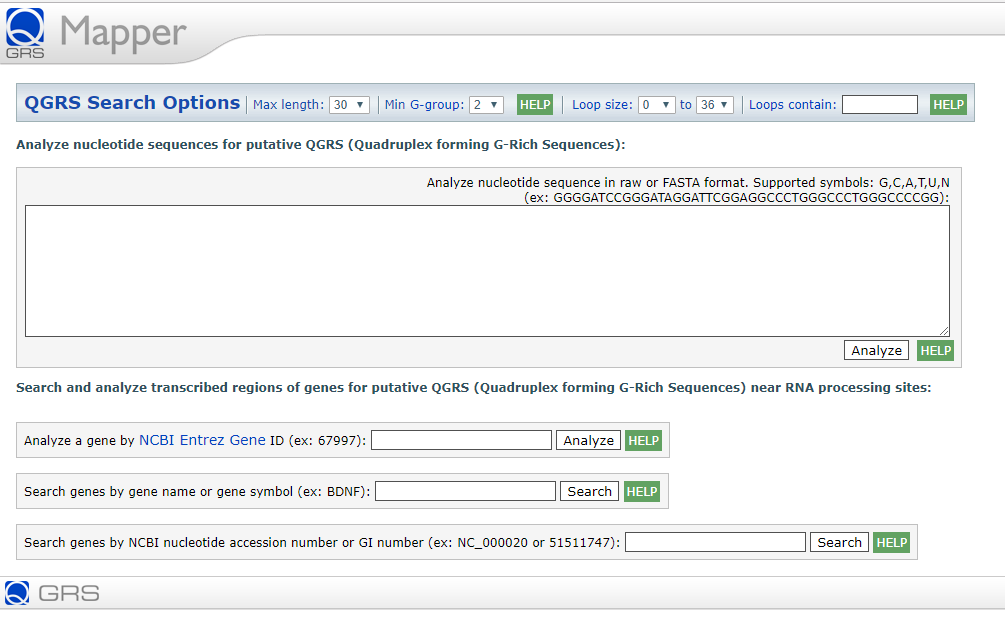


图2：QGRS mapper工具界面：可以自定义序列，也可以直接在基因中进行搜索预测，预测选项设置较多，从G-tetrads长度到loop长度和序列总长度，都可以进行自定义设置。

## 3、G4P calculator：

G4P calculator[3]是根据连续的鸟嘌呤序列(G-run)的密度来推测形成G-四联体的可能性。G4P calculator在一个滑动窗内评估G-run的个数，并且计算符合特定条件的序列占滑动窗的百分比。这里的特定条件是指包含有4个及以上的G-run(G-tetrads)，每个G-run包含3个及以上的鸟嘌呤。滑动窗大小为100bp，每次滑动的距离为20bp。滑动窗为100bp有利于快速分析同时也可以改变从而有利于个性化的分析。20bp的滑动距离是为了让可能形成多条G-四联体的序列在打分时贡献更高。每条序列在最后四次滑动窗时窗的大小会逐渐变小(80bp, 60bp, 40bp, 20bp)，但是由于基因长度远大于窗的大小，所以并不影响最终结果。G4P calculator以百分比的形式对序列进行打分，使得分数独立于其长度，打分标准与quad\_parser算法相类似。这个分数也可以用来在不同序列之间比较，但是比较结果不具有绝对性，即不认为得分高的序列一定比得分低的序列更容易形成G-四联体。

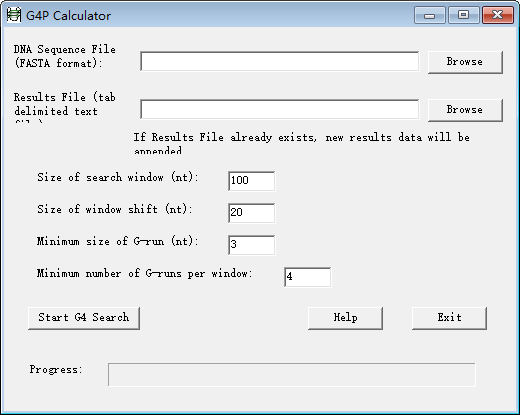


图3：G4P calculator的工具界面：G4P calculator是一个Windows系统的执行文件，输入和输出都是fasta格式的文件，设置选项较少，关于G-四联体序列的自定义设置也较少。

## 4、G4-Hunter：

G4-Hunter[4]考虑到鸟嘌呤的密度和倾斜度(形成二级结构的可能性)，序列中的每一个碱基都被给定一个介于-4到4的分数。对于碱基腺嘌呤(A)和胸腺嘧啶(T)，分数均为0，而对于鸟嘌呤(G)，分数为正，胞嘧啶(C)，分数为负。考虑到G的丰度(或是C的丰度，也就意味着相反链上G的丰度)，一个单个的鸟嘌呤(G)分数为1，两个连续的鸟嘌呤(GG)，每个鸟嘌呤的分数为2，三个时每个为3，四个及以上时，分数均为4。对于胞嘧啶(C)，分数给定方式一致，但是给定的分值为负。在这种打分下，GC富集序列的得分接近为0，考虑到GC富集序列本身更容易形成稳定的双螺旋结构而不是G-四联体结构，这种打分方式相对正确。而且这种打分方式还同时计算了互补链的分数(直接取负即可)。对于一个给定的序列，将每个碱基的得分加和后去算数平均以当作序列本身的分数。通过人为构建序列发现，随机序列的分数与GC含量相关，接近于0。这种方法也证实了基因序列本身并不是随机序列。

对于长序列或基因组序列，首先计算25bp滑动窗内核苷酸序列的分值。随后提取在设定阈值以上的区域，然后对于这些区域去除他们末端的非G(或非C)，如果第一个或最后一个碱基是G(或C)也可以延展这条序列。如果需要的话也可以考虑序列的上一个或下一个碱基从而避免加窗时将连续的G(或C)切割分离。随后将两段具有相同片段的序列进行融合，随后在进行计算得分，但这样共享高分片段会导致序列得分减少甚至低于阈值。在这里以阈值x找到的G-四联体序列称作G4Hx(阈值x理论上可以为0-4的任意值，但一般设置在1-2之间)。

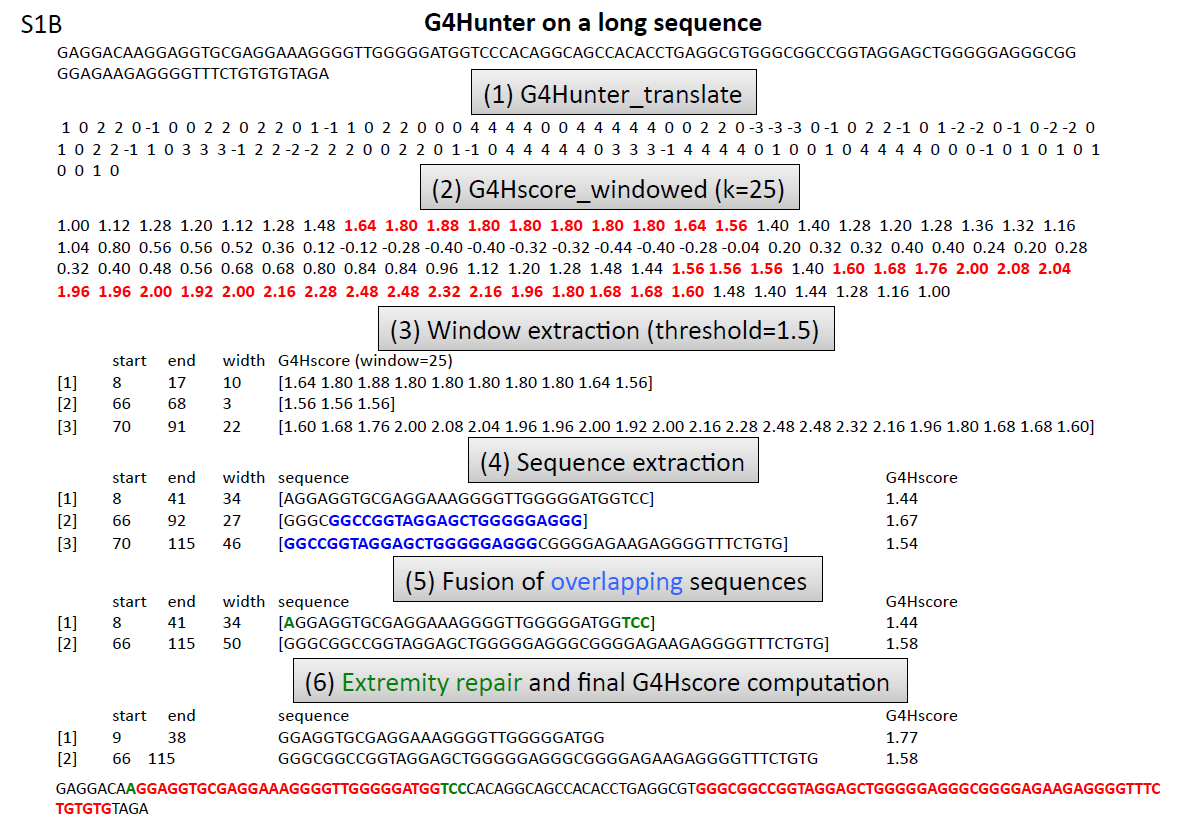


图4：G4-Hunter的长序列计算方法：1,、对每一个碱基进行打分，2、提取在阈值之上的序列片段，3、对提取出的序列进行腐蚀和融合操作。

## 5、cGcC：

cGcC[5]是一种新的打分系统，可以与标准模式结合从而更好地预测RNA中G-四联体的形成。cGcC分数分为两个部分，cG 分数是根据字符串s的长度n决定的： 其中s表示序列字符串，Gs是s所有非空的连续的鸟嘌呤所组成的子串的集合，Gs(i)表示长度为i的子串的集合，|Gs(i)|表示集合Gs(i)的大小。注：Gs中可以包含序列完全相同但位置不同的子串，例如：序列s=‘GGGCGGG’有两个’GGG’子串，所以|Gs(3)|=2。

cG分数就定义为：。

cC分数定义相似为。

也就是说，对于一段PG4序列，每一个G或是C给予分值10，每一段GG或是CC给予分值20，每一段GGG或是CCC给予分值30，以此类推。例如序列s=’GGG’时，cGscore=3(G)\*10+2(GG)\*20+1(GGG)\*30=100。s=’GG’时，cGscore=2(G)\*10+1(GG)\*20=40。最后。

## 6、2D meets 4G：

和其他计算方法不同的是2D meets 4G[6]算法不只是考虑序列的碱基组成，而是将G-四联体序列放置于长序列中，通过预测整个长序列的二级结构，再结合算法中G-四联体的热力学模型从而判断序列中是否具有可能形成G-四联体的序列。所以在短序列中由于其他序列所给予的参数较少，对G-四联体的影响也较少，所以短序列中的G-四联体预测与普通的计算方法相类似。由于算法原理较为复杂，在这里并不详细介绍。



图5：同一段序列采用2D meets 4G算法(RNAfold 2.0g)和不采用2D meets 4G算法(RNAfold 2.0)的二级结构预测结果对比。

## 7、G4RNA：

G4RNA[7]是包含已经经过实验验证了的G4折叠的人类RNA序列的参考数据库，以及相关的实验数据，来源和相关预测措施。 这个综合数据集的创建是将宝贵的研究信息集中到一个关键的一步，为进行系统的比较和序列分析提供可靠和专业的参考工具。

G4RNA数据库是采用MySQL(5.5.40)建立的关系数据库，其核心为核苷酸序列及其主要属性，例如在相关发表文献中的序列识别码，长度和在hg38基因组中的位置及所在的基因。G4RNA完全是有人工对文献进行管理，因此其中只包含有权威出版物和实验验证了的序列。每条序列的实验结果和实验性质都可以通过参考文献进行深入研究。

除此之外，G4RNA还使用G-四联体预测工具提供了这些序列的结果。这些值对于预测G-四联体形成的可能性非常有用，将这些预测结果集中在实验验证的序列上有利于系统的比较和分析，甚至可以促进在RNA生物学领域的发展。

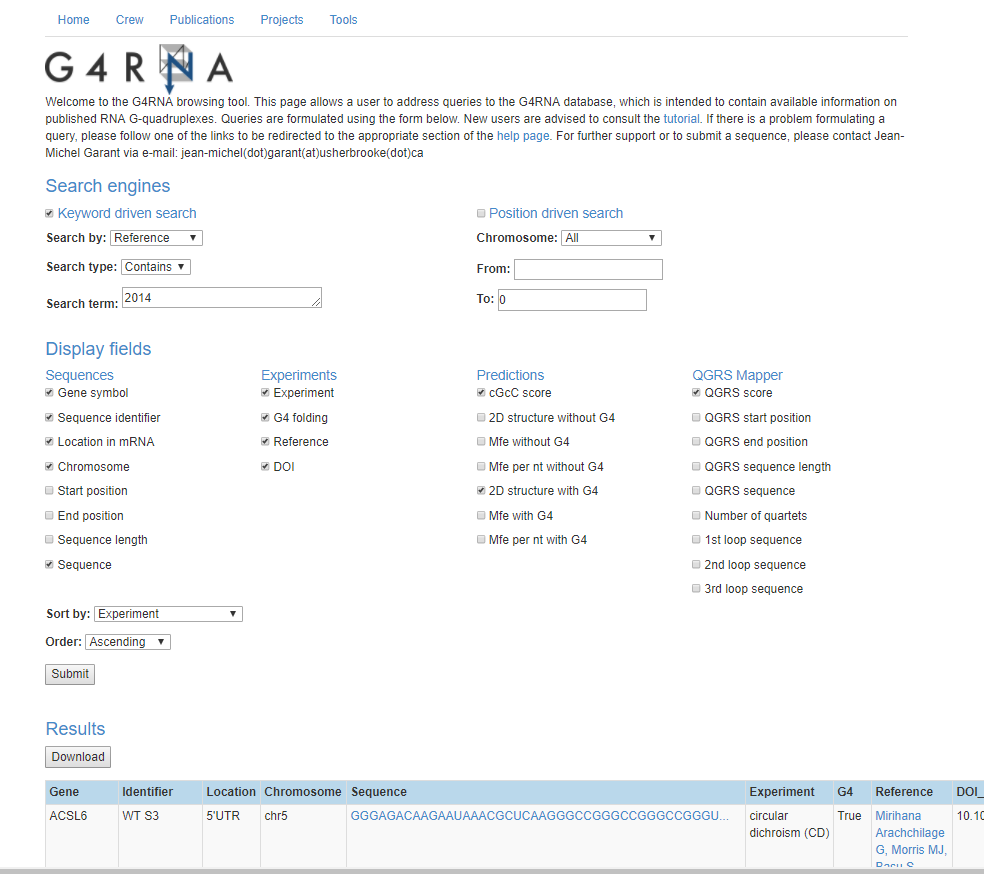


图6：G4RNA数据库网页界面：数据库主要采用参考文献索引或是实验内容索引，也可以通过序列在染色体的位置进行搜索。

# 三、G4RNA简单分析：

作为几个预测工具中唯一一个和实验数据结合的工具，我们在这里将其简单对G4RNA中的数据进行分析，首先根据文献的发表年限，下载从2010年至2016年的所有RNA G-四联体的数据，可以根据下载数据查看不同年限里RNA G-四联体实验相关文献发表数量，如图7所示。G4RNA数据库中收录的序列以正样本居多，而负样本序列大多与G-四联体motif相类似，因为这些序列大多是由于一些特殊变异而无法形成G-四联体，或是被实验者认为能形成但是在实验过程中没有形成的序列。

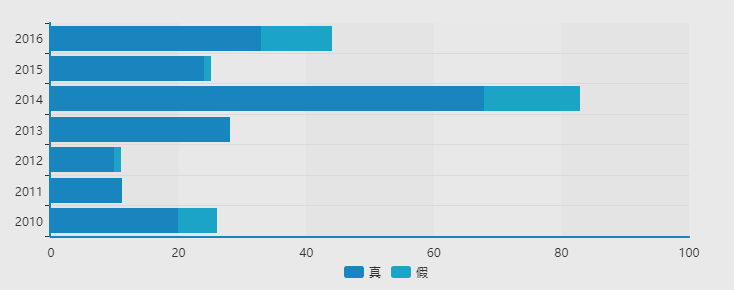


图7：G4RNA数据库中不同年限收录数据，其中真指实验中验证会形成G-四联体的序列，假是指验证了序列不会形成G-四联体。

G4RNA数据库中所收录的实验数据大多是采用环二色光谱(CD)，核磁共振法(NMR)，线性指针(line probing)或是融化温度(melting temperature)，虽然这些实验无法提供准确的G-四联体在体内的形成情况，但是这些实验可以为其他预测工具提供简单的形成情况分析，表1是直接将序列进行RNAfold的G-四联体二级结构分析。

|  |  |  |
| --- | --- | --- |
| 实验结果 预测结果 | 真 | 假 |
| 真 | 158 | 36 |
| 假 | 11 | 23 |

表1：实验结果即实验中是否会形成G-四联体，预测结果是指在二级结构预测中结果是否形成了G-四联体结构。

从表中可以看到真阳性实验的个数远大于其他实验个数，但是这是由于G4RNA数据库中收录的实验结果为真的数目远大于假的数目(194: 34)，在这种样本分布不平衡的情况下讨论精度或是召回率(recall)结果的迷惑性较大，无法带来真实的对预测结果或是实验结果的评价。但是可以通过这样进行分类，从分类结果中查看cGcC和QGRS评分。

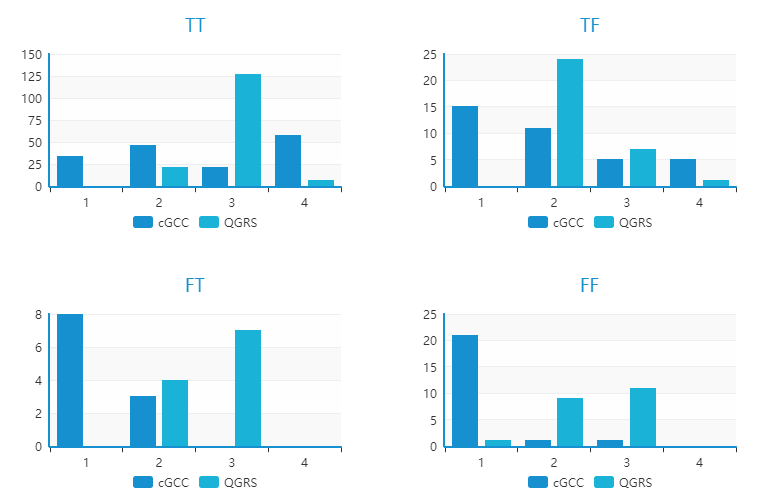
无论是cGcC还是QGRS评分，都是指分值越高序列形成G-四联体的可能性越高，从上图可以看出在四种分类中QGRS评分和RNAfold相性较好，即当序列中存在G-四联体 motif时QGRS分数相对较高，而没有时分数较低，但是与实验的结果匹配相对较差，而cGcC评分本身并不涉及motif，因此评分相对均衡，但是当实验结果为假时，其评分也均较低，在不考虑样本量较小的情况下，这是一个非常理想的结果了，但是对于一些实验结果为真的序列，其评分也存在有较低的情况，因此可以认为cGcC评分在一定水平上通过抑制分数来使预测结果偏向实验假的情况，虽然这样使得假阳性率较低，但是真阳性率也因此降低。

图8：cGcC和QGRS分值与实验结果和RNAfold预测结果的关系。其中将序列分为四类，名称分别为TT，TF，FT，FF，其中第一个T或F为实验验证的结果，第二个表示为RNAfold预测的结果。其中1,2,3,4,分别指分数从低到高四个等级。

从以上结果可以看出，虽然G4RNA收集了许多G-四联体的实验结果，但是受实验研究的限制，所收集的正样本远多于负样本，而且由于发表的序列多为只包含G-四联体片段的短序列，在这种情况下，RNAfold通过二级结构预测G-四联体的方法收到较大限制，难以通过这种方法描述RNAfold预测的准确性。而对于两种评分方法QGRS和cGcC来说，由于QGRS评分建立在G-四联体的motif之上因此预测真阳性样本时准确率较高，但是对于其他三种情况准确率较低；而cGcC评分则是在预测实验结果为假的所有样本中表现良好，而在为真的样本中表现相对较差。

为了更清晰的看出各种预测工具和实验结果之间的关系，我们随后分别利用三者结果采用SVC算法对实验结果构建分类器进行预测，再利用三者同时作为特征输入构建分类器，结果如图9所示。可以看出在三者中cGcC计分方法预测结果最好，其次就是RNAfold方法，但是由于RNAfold本身就是二值化数据，所以在ROC曲线中呈现为只有一个拐点的折线，最后是QGRS算法，可以看出QGRS算法与其他算法差距较大，所以在实验数据中，只查看G-四联体motif的方法适用性较差。将三者结合后发现正确率又有所提升，所以三者结合已经可以得到一个较好的结果。

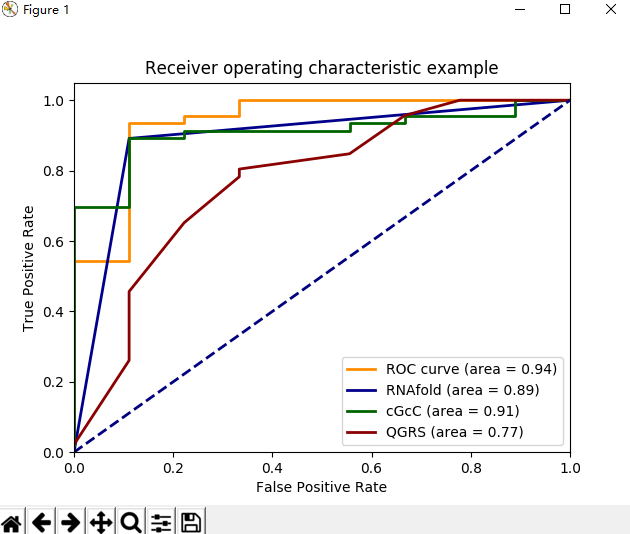


图9：利用RNAfold，cGcC和QGRS评分对实验结果进行分类的ROC曲线。

# 四、软件对比和总结：

G-四联体由于其本身形成原理导致其具有一定的序列相似性，即理论上G-四联体均应符合motif，因此初始设计的G-四联体识别预测工具均是在此基础上预测，再结合少量实验得到的结论，对motif本身进行约束。以Quad\_parser为例，要求G-tetrads大于等于3或是loop长度在1到7之间。但是这种方法找到的序列极多，在实验验证后发现这种方法预测的G-四联体假阳性率较高，并不能直接使用其预测结果，因此许多预测工具开始对序列进行打分如QGRS mapper和G4P calculator，而打分的依据一般都根据GC含量或是loop长度等影响G-四联体稳定性的因素。随着实验技术如环二色谱法(CD)，核磁共振法(NMR)等发展，越来越多的实验发现有部分序列在实验中出现了G-四联体的反应，但是这些序列本身并不符合G-四联体的motif，因此有少量工具如G4-Hunter开始不使用motif，而是利用一定的算法直接对序列进行打分，并生成超出一定分值的序列，一般这样的打分都倾向于生成富G序列，同时这一段序列的C较少。随着RNA G-四联体的研究，越来越多专门预测RNA G-四联体的工具开始出现，在这种趋势下，cGcC打分方法将分数与motif相结合，同时不设阈值，对每条序列进行打分，方便进一步分析，但是并没有一个绝对的参数来判断是否形成了G-四联体结构。而对于RNA G-四联体，也有工具将G-四联体放置在整个RNA二级结构中，以viennaRNA包为例，采用算法2Dmeet4G利用序列最小自由能来预测一段序列是形成沃森-克里克二级结构还是G-四联体二级结构。但是以上这些预测工具多是利用motif和序列本身碱基组成来预测G-四联体，即使是根据最小自由能的方法也是根据序列本身计算得到。为了弥补工具几乎不使用实验数据中的不足，因此出现了G4RNA数据库，这个数据库通过文献挖掘，在权威文献中找到实验支持的G-四联体，不只是实验验证的可以形成的G-四联体，在其中也有无法形成G-四联体的序列。在这些数据的基础上，后续分析可以得到实验的支持，也有利于验证前人的结果。具体对比见表2。

表2：各预测工具对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 工具 | Motif | 其他影响因素 | 程序类型 | G4RNA是否包含 |
| Quad parser |  | None | 现已不可用 | 否 |
| QGRS mapper |  | None | 网页 | 是 |
| G4P calculator |  | None | 程序 | 否 |
| G4-Hunter | None | GC含量 | 脚本 | 否 |
| cGcC | None | GC含量 | 无 | 是 |
| 2Dmeet4G |  | 二级结构 | 程序 | 是 |
| G4RNA | None | 实验结果 | 网页 |  |

续表1：

|  |  |  |
| --- | --- | --- |
| 工具 | 优点 | 缺点 |
| Quad parser | 早期预测工具，采用的motif是其他工具的基础 | 网页接口不可用，预测扩展性不足 |
| QGRS mapper | 将motif和打分相结合，现在也是主要的G-四联体预测工具 | 仅考虑序列本身 |
| G4P calculator | 采用百分比值打分，方便序列比较 | 仅能在长序列中预测G-四联体 |
| G4-Hunter | 采用新的预测方法，是探索G-四联体预测的崭新的尝试 | 只考虑GC连续序列的影响，分值间比较无意义 |
| cGcC | 不考虑motif只是对所有序列进行打分，是一种新的辅助预测方法 | 打分方式较为粗糙，G-四联体的形成和分值关系较小 |
| 2Dmeet4G | 考虑G-四联体在二级结构中的形成，在预测序列二级结构的同时判断G-四联体的形成 | 仅在RNA长序列中有意义，对于短序列程序与其他工具类似 |
| G4RNA | 将实验G-四联体总结，把实验和部分预测工具结合 | 不对序列进行预测，仅展示结果 |

# 五、参考文献

[1]. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Research. 2005;33(9):2908-2916.

[2]. Kikin, O., L. D'Antonio and P.S. Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Research, 2006. 34(Web Server): p. W676-W682.

[3]. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Research. 2006;34(14):3887-3896.

[4]. Bedrat, A., L. Lacroix and J. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Research, 2016. 44(4): p. 1746-1759.

[5]. Beaudoin, J., R. Jodoin and J. Perreault, New scoring system to identify RNA G-quadruplex folding. Vol. 42. 2013.

[6]. Lorenz, R., et al., 2D Meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013. 10(4): p. 832-844.

[7]. Garant, J., et al., G4RNA: an RNA G-quadruplex database. Vol. 2015. 2015..

[8]. I. Bang, “Untersuchungen ¨uber die Guanyls¨are,” Biochemisce Zeitschrift, vol. 26, pp. 293–311, 1910.

[9]. M. Gellert, M. N. Lipsett, and D. R. Davies, “Helix formation by guanylic acid,” Proceedings of the National Academy of Sciences of the United States of America, vol. 48, pp. 2013–2018, 1962.

[10]. D. J. Patel, A. T. Phan, and V. Kuryavyi, “Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics,” Nucleic Acids Research, vol. 35, no. 22, pp. 7429–7455, 2007.

[11]. J. L. Huppert, “Four-stranded DNA: cancer, gene regulation and drug development,” Philosophical transactions. Series A, vol. 365, no. 1861, pp. 2969–2984, 2007

[12]. L. Oganesian and T. M. Bryan, “Physiological relevance of telomeric G-quadruplex formation: a potential drug target,” BioEssays, vol. 29, no. 2, pp. 155–165, 2007.

[13]. S. Rankin, A. P. Reszka, J. Huppert, et al., “Putative DNA quadruplex formation within the human c-kit oncogene,” Journal of the American Chemical Society, vol. 127, no. 30, pp. 10584–10589, 2005.