生物信息学应用笔记

Vol. 29 no.;22 2013, 2938-2940页
doi: 10.1093 /生物信息学/ btt494

结构生物信息学

Advance Access出版于2013年8月23日

RNAfbinv:一个交互式Java应用程序，用于基于片段的RNA序列设计

Lina Weinbrand[1]，Assaf Avihoo[2]和DannyBarash[1,*]

[1]内盖夫本古里安大学计算机系，比尔舍瓦84105，以色列和[2]微软以色列研究院，赫兹利亚46733，以色列

副主编:伊沃·Hofacker

摘要。

摘要:在RNA设计问题中，出于生物学原因，假设用户对保留特定的RNA二级结构基序或片段感兴趣是合理的。保存可能是在结构或序列上，或两者兼而有之。因此，考虑片段约束可以使RNA逆向折叠问题受益。我们开发了一个新的交互式Java应用程序，称为基于RNA片段的逆，它允许用户以点括号符号插入RNA二级结构。然后，它执行符合输入二级结构形状、指定的热力学稳定性、指定的突变鲁棒性和形状分解后用户选择的片段的序列设计。在这种基于形状的设计方法中，具有已知生物学功能的特定RNA结构基序被严格执行，而其他结构基序可以在其结构上具有更大的灵活性，从而有利于保留物理属性和附加约束。

可用性:RNAfbinv可以在http://www.cs.bgu.ac.il/*RNAexinv/RNAfbinv网站上免费下载。该网站包含一个帮助文件，其中有关于确切用法的解释。

联系:dbarash@cs.bgu.ac.il

2013年3月11日收稿;2013年6月28日修订;2013年8月19日接受

1 介绍

用于设计折叠成给定RNA二级结构的序列的逆RNA折叠问题于20世纪90年代初在维也纳提出(Hofacker等，1994)。通过随机优化来解决它的方法依赖于使用RNA折叠预测web服务器中可用的软件来解决直接问题，例如RNAfold服务器(Hofacker, 2003)或mfold (Zuker, 2003)，通过使用热力学参数执行能量最小化(Mathews等人，1999)。首先选取种子序列，然后利用局部搜索策略对种子进行突变，并重复应用能量最小化预测RNA折叠的直接问题。然后，在种子序列附近，根据优化问题公式化中的目标函数，找到具有期望折叠性质的设计序列。在INFO-RNA (Busch和 Backofen, 2006)、RNA-SSD (Aguirre-Herna ' ndez 等，2007) 和NUPACK:Design (Zadeh等，2011)中已经研究出了这种方法的性能改进。进化问题的一些一般扩展动机是在Dromi等人(2008)中提出的，并使用不考虑效率的进化计算进行。其他解决相关问题的试验，如设计具有多个目标结构的 RNA，包括 Frnakenstein (lyngsot et al.，2012)。Heitsch et al. (2003)在研究中描述了一些与编码理论有趣的类比。最近，RNAiFOLD中开发了一种约束规划方法(Garcia-Martin et al.，2013)，RNA-ensign中引入了一种全局采样方法(Levin et al.，2012)。在局部搜索策略的框架内，使用RNA二级结构的粗粒树形图表示，我们扩展了RNA逆向折叠问题，以包括热力学稳定性和突变鲁棒性等约束，开发了一个名为RNAexinv的程序(Avihoo等人，2011)。与Dromi等人(2008)的研究相比，该程序提高了效率，但它对用户想要保存的特定RNA二级结构基序没有偏好。

为了激发使用户能够指定从业者想要保存的特定RNA二级结构基序的偏好的需求，可以考虑RNA病毒领域的一个例子。众所周知，特定的丙型肝炎病毒茎环对病毒复制很重要(例如Yi和Lemon, 2003)。在未来，更多这类RNA结构基序在不同类型RNA中的功能重要性将被揭示。用户根据这些给定信息设计序列的能力应该会变得有用。

RNAfbinv(基于RNA片段的逆)是一个交互式Java应用程序，它允许用户通过解决扩展的RNA逆折叠问题，插入RNA二级结构并基于用户选择的片段设计序列。给定RNA二级结构，RNAfbinv程序将该结构分解为RNA二级结构基序的片段。然后，用户可以选择保留某些被认为具有生物学重要性的片段，之后，该程序通过考虑突变鲁棒性和热力学稳定性的附加约束来解决扩展的逆RNA折叠问题(Avihoo等人，2011)。此外，还可以插入序列上的核苷酸约束。基于粗粒树形图的结构基序分解(夏皮罗，1988)允许一种自下而上的方法，在这种方法中，rna设计的序列是由片段、序列约束以及进化和物理属性构建的。在此背景下，Dromi等人(2008)在研究中首次提出了粗粒度树形图的使用

*To whom correspondence should be addressed.

受希格斯(2000)关于RNA二级结构的物理方面的回顾的启发，物理属性被插入作为约束。RNAfbinv是粗粒度思想的自然扩展，它使用粗粒度树形图表示将结构分解为片段。然后，用户可以选择应该保留的结构约束片段，并与用户选择的序列约束片段一起组成构建块，在此基础上设计序列。对于粗粒度，我们选择使用树形图(夏皮罗，1988)，但同样，可以使用抽象形状(Giegerich et al.， 2004)。

该应用程序由三个独立的程序组成:RNAattributes, RNAfbinv和RNAfbinv-flex，支持灵活的长度设计序列。RNAattributes程序是一个辅助工具，它接受RNA序列，并提供其预测的最小自由能二级结构及其最小自由能和突变鲁棒性。RNAfbinv程序是接受RNA结构、热力学稳定性、突变鲁棒性以及基于粗粒树形图表示的用户选择序列和结构片段的主要程序。输出由设计的序列组成，这些序列根据其预测结构与输入结构的碱基对距离进行排序。RNAfbinv-flex程序是RNAfbinv的扩展，它也输出设计的序列，但长度可变。由于这种灵活性，RNAfbinv-flex在当前实现中不接受序列约束，因为当序列大小变化时，序列约束的位置会变得模糊。将来，更复杂的程序可能允许在RNAfbinv-flex中插入序列约束，但这需要特殊的开发。在RNAfbinv-flex的当前实现中，最多删除了2个nt，并且可以通过过滤后处理步骤中获得的结果来考虑序列约束。

2 方法

RNAfbinv使用与RNAexinv类似的方法(Avihoo等人，2011)，除了目标函数最小化的主要差异。一般来说，RNAfbinv的求解方法包括两个阶段:(1)确定一个良好的初始候选者，如果该基序在初始候选者中不存在，则重新选择该基序;(2)使用4-nt前瞻性局部搜索函数模拟退火。下面对这两个阶段进行更详细的阐述。对于第一阶段，根据我们的需要，使用随机起始点的RNAinverse (Hofacker等人，1994年)比像INFO-RNA中使用的确定性第一阶段(Busch和Backofen, 2006年)更可取。随机起始点将为扩展搜索产生不同的起始序列，而不是固定起始点。这样做的目的是为了避免每次重复达到相同的局部最小值。在第二阶段，正如RNAexinv (Avihoo等人，2011)更详细地描述的那样，执行迭代突变以搜索局部最小值，并使用具有4-nt前瞻性局部搜索函数的模拟退火方法对序列附近进行采样(默认前瞻性为4，不穷尽，仅采样)。在目标函数中，在rnafbinv中添加一个相对于RNAexinv输出二进制值的额外项，以说明用户选择的motif存在(或不存在):

f - ðinitial, targetÞ¼
jneutralitytargetneutralityinitialj100 þjdGtargetdGinitialj þ
target motif existsðinitialÞ1000
到树编辑距离夏皮罗表示& target, initialÞ100
þ碱基对距离dotBracket representation & target, initialÞ0:01

目标母题存在的额外项是最重要的约束，应该在没有任何妥协的情况下精确地满足。因此，相对于所有其他项，我们为这一项选择了一个更大的权重1000。其他术语与RNAexinv (Avihoo等人，2011)中相同，描述如下。测量突变鲁棒性的中立性是一个介于0和1之间的数字;因此，赋值权重为100。最小自由能dG是用来测量热力学稳定性的，单位是千卡每摩尔;因此，赋一个单位权值。二级结构之间的距离是使用维也纳RNA包中的RNA距离计算的(支持称为夏皮罗表示的粗粒度树形图和二级结构的点括号表示)。对于夏皮罗表示之间的树编辑距离，选择了一个相对较大的权重100来保持形状，而对于最后一项的碱基对距离，分配了一个非常小的权重0.01。正如Avihoo等人(2011)所解释的那样，最后一个术语是原始RNAinverse中用于保留精确的二级结构的术语，其目的是保护解决方案免受过于基于形状的支配。

3 实现

RNAfbinv目前可以在Linux平台(Ubuntu、OpenSuse和Fedora都被检查过)、Mac和Windows上使用。所有的准备和编译都应该在安装了Java和“GNU CC”编译器的情况下进行。包内容包括维也纳RNA包中的程序。RNAfbinv在GNU许可下是免费的。

有关准备和编译的详细说明可在ReadMe文件中获得，该文件可以从web指针轻松访问。在输入屏幕中，用户以点括号形式插入一个结构，按下“Fragment”键，程序将结构分解为片段。出现图1所示的下一个屏幕，然后让用户通过滚动包含所有结构图案的组合框下拉列表来选择应该保留的结构图案(片段)。用户选择结构基序后，在给出点括号结构的输入屏幕中按下“Process”，将出现一个新窗口，点括号位于顶部，后面是可以轻松填充的物理参数和核苷酸约束信息。综上所述，输入数据和参数列表包含:

(1)点括号表示法的输入结构。

(2)用户选择的结构母题(片段)。

(3)物理参数:以千卡/摩尔为单位的期望自由能和期望中性，这是一个介于0到1之间的数字。

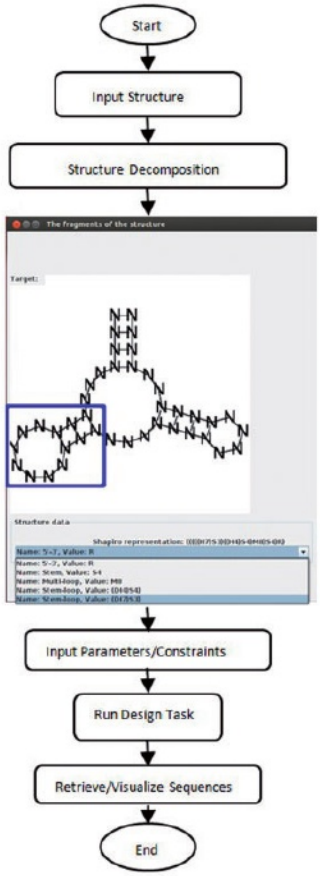

图1所示。结构分解到图案和选择画面

(4)核苷酸约束:需要的核苷酸(s)，起始位置和允许单核苷酸滑移的可能性。

按下“Results”，程序开始执行。输出屏幕上出现了20种不同的可能设计序列。用户可以按下每个序列下面的按钮来获取有用的信息。整个工作流程如图1所示。

计算时间取决于用户选择作为输入的序列长度。对于图1中的示例，在标准PC上运行时间为10分钟。默认的迭代次数设置为100次。平均而言，RNAfbinv比rnaxinv贵10% (Avihoo等人，2011)。在不考虑非结构约束的情况下，有可能将计算时间减少6%。

4 结果

通过在设计的序列上运行RNAfold (Hofacker, 2003)或mfold (Zuker, 2003)，可以很容易地检查RNAattributes、RNAfbinv和RNAfbinv-flex的输出。

为了测试从我们的程序中获得的反问题解决方案是否令人信服地满足所需的约束，我们在人工示例和自然示例上运行它。例如，在取自Krol等人(2004年)、Dromi等人(2008年)和Avihoo等人(2011年)的miRNA前体样本中，我们能够获得满足所有约束条件并完全保存所选片段的设计序列。

5 未来发展

RNAfbinv的基本实现是通用的，可用于解决各种RNA设计问题。它还可以通过选择符合若干约束条件的RNA设计序列来搜索比基于序列的搜索更通用的RNA模式。在未来，将有可能扩展RNAfbinv来处理更多的约束，并通过使用更复杂的采样方法来提高其效率，例如Levin等人(2012)和其他方式。

致谢。

作者感谢Alexander Churkin和Idan Gabdank的技术支持。

资助:本-古里安大学Lynne and William Frankel计算机科学中心。

利益冲突:未声明。

参考文献。

Aguirre-Herna´ndez,R. et al. (2007) Computational RNA secondary structure design: empirical complexity and improved methods. BMC Bioinformatics, 8, 34.

Avihoo,A. et al. (2011) RNAexinv: an extended inverse RNA folding from shape and physical attributes to sequences. BMC Bioinformatics, 12, 319.

Busch,A. and Backofen,R. (2006) INFO-RNA-a fast approach to inverse RNA folding. Bioinformatics, 22, 1823–1831.

Dromi,N. et al. (2008) Reconstruction of natural RNA sequences from RNA shape, thermodynamic stability, mutational robustness, and linguistic complexity by evolutionary computation. J. Biomol. Struct. Dyn., 26, 147–162.

Garcia-Martin,J.A. et al. (2013) RNAiFOLD: a constraint programming algorithm for RNA inverse folding and molecular design. J. Bioinform. Comput. Biol., 11, 1350001.

Giegerich,R. et al. (2004) Abstract shapes of RNA. Nucleic Acids Res., 32, 4843–4851.

Heitsch,C.E. et al. (2003) From RNA secondary structure to coding theory: a combinatorial approach. In: Proceedings of the 8th international meeting on DNA based computers, Vol. 2568, Lecture Notes in Computer Science, Springer-Verlag, pp. 215–228.

Higgs,P.G. (2000) RNA secondary structure: physical and computational aspects. Q. Rev. Biophys., 33, 199–253.

Hofacker,I.L. et al. (1994) Fast folding and comparison of RNA secondary struc-tures. Monatsh. Chem., 125, 167–188.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. Nucleic Acids Res., 31, 3429–3431.

Krol,J. et al. (2004) Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. J. Biol. Chem., 279, 42230–42239.

Levin,A. et al. (2012) A global sampling approach to designing and reengineering RNA secondary structures. Nucleic Acids Res., 40, 10041–10052.

Lyngsø,R.B. et al. (2012) Frnakenstein: multiple target inverse RA folding. BMC Bioinformatics, 13, 260.

Mathews, D. H. et al. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol., 288, 911–940.

Shapiro,B.A. (1988) An algorithm for comparing multiple RNA secondary struc-tures. Comput. Appl. Biosci., 14, 387–393.

Yi,M. and Lemon,S.M. (2003) Structure-function analysis of the 3’ stem-loop of hepa-titis C virus genomic RNA and its role in viral RNA replication. RNA, 9, 331–345.

Zadeh,J.N. et al. (2011) Nucleic acid sequence design via efficient ensemble defect optimization. J. Comput. Chem, 32, 439–452.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31, 3406–3415.