(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106295243 A
(43)申请公布日 2017.01.04

(21)申请号 201610649359.0

(22)申请日 2016.08.10

(71)申请人 华中科技大学
地址 430074 湖北省武汉市洪山区珞喻路1037号

(72)发明人 刘士勇 郑进芳

(74)专利代理机构 华中科技大学专利中心 42201
代理人 胡星驰

(51)Int.Cl.
G06F 19/16(2011.01)

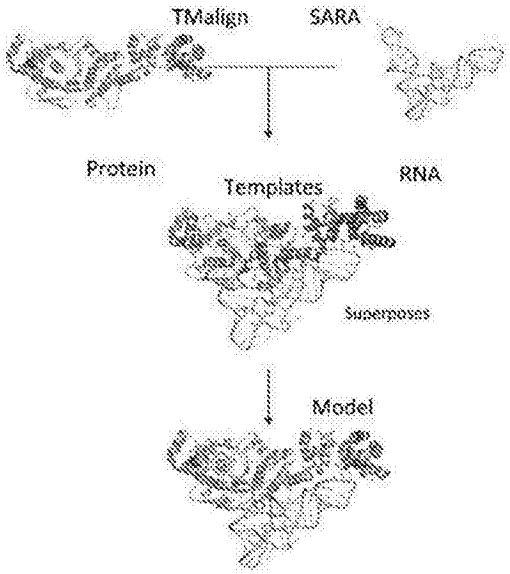
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种蛋白质-RNA复合物结构预测方法

(57)摘要

本发明公开了一种蛋白质-RNA复合物结构预测方法，具体涉及一种基于模板构建蛋白质-RNA相互作用模型的方法，首先通过从PDB数据库中挑选出439个的蛋白质-RNA的模板库，然后使用蛋白质(RNA)的结构比对所有的模板复合物得出相似分数；然后再根据蛋白质(RNA)中的相似分数小的那个值对模型进行排序，最后与给定的阈值进行计较来判断给定的蛋白质-RNA是否能够结合并给出蛋白质-RNA的3D结构。本发明开创性地提出了在基于模板构建蛋白质-RNA的相互作用模型的计算方法，填补了目前的空白，本发明的计算方法比对接的方法成功率增加了40%左右，大大的促进了蛋白质-RNA三维结构领域的发展。



CN 106295243 A

1.一种蛋白质-RNA复合物结构预测方法,其特征在于,包括如下步骤:

(1)计算模板复合物结构分数:将给定的蛋白质和RNA的单体结构分别与模板库中的蛋白质-RNA相互作用模型模板进行比对,分别得到给定的蛋白质与模板蛋白质的相似分数A,以及给定的RNA与模板RNA相似分数B;对所述相似分数A和相似分数B进行比较,取所述相似分数A和所述相似分数B中较小的相似分数作为利用该模板得到的蛋白质-RNA相互作用模型的复合物结构分数,每一个模板得到一个蛋白质-RNA相互作用模型的复合物结构分数;

(2)模型排序:将步骤(1)获得的蛋白质-RNA相互作用模型的复合物结构分数按照降序排列;

(3)模型判断:预先给定一个阈值,将步骤(2)按照降序排列获得的第一个复合物结构分数,即蛋白质-RNA相互作用模型的复合物结构分数的最大值与所述阈值进行比较,当所述复合物结构分数的最大值小于该阈值,则判断该模型结构不正确,所述给定蛋白质和RNA不能结合;当所述复合物结构分数的最大值大于所述阈值,则判断该蛋白质-RNA相互作用模型结构正确,该给定蛋白质和RNA可以结合。

2.如权利要求1所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述模板库的获得方法为:从PDB数据库中下载到所有的蛋白质-RNA复合物结构,然后从中根据晶体结构分辨率和蛋白质残基以及RNA碱基个数挑选确定模板库。

3.如权利要求2所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述模板库中的蛋白质-RNA相互作用模型晶体结构分辨率比3.0好,所述蛋白质残基个数大于30,所述RNA的碱基个数大于20。

4.如权利要求1所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述模板库中一共有439个蛋白质-RNA相互作用模型模板。

5.如权利要求1所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述给定的蛋白质和模板蛋白质的比对方法为使用TMalign程序来比对。

6.如权利要求1所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述给定的RNA与模板RNA的比对方法为使用SARA程序来比对。

7.如权利要求6所述的蛋白质-RNA复合物结构预测方法,其特征在于,所述SARA程序使用一个归一化的向量来代表RNA的结构,结合RNA的二级结构特征,来比对RNA的二级结构。

一种蛋白质-RNA复合物结构预测方法

技术领域

[0001] 本发明属于分子构建模型领域，具体地，涉及一种蛋白质-RNA复合物结构预测方法，更具体地，涉及一种基于模板构建蛋白质-RNA相互作用模型的方法。

背景技术

[0002] 为了揭示蛋白质-RNA的相互作用的机理，有两种方法来获取蛋白质-RNA的三维结构：第一种是实验上的方法，比如用的是结晶蛋白质-RNA的晶体，然后是用X射线的来解析其三维结构；第二种用的是计算机模拟的技术。又可以分成对接的方法和基于模板的方法，目前已经有对接的方法了如3dRPC，然而基于模板的方法在蛋白质-RNA还没有被实现。

[0003] 基于对接的方法是根据几何互补原理，在生物学上中锁钥模型，就是当两种分子之间进行识别时是根据这两种分子形状上的互补，根据几何上的互补得到一个评价分数，并且基于分数的高低判断分子之间取向的合理性。由于计算机能够取样很多很多的构象，因此根据分数来对这么多的构象进行排序，然而仅仅根据几何互补性分数，其前10的构象中至少有一个构象是正确的概率比较低。

[0004] 在蛋白质-蛋白质模型构建之中，对接的方法和基于模板的方法各有其优点，基于模板的在排名前几名的成功率要比对接的方法要高，而且基于模板的方法能够在构象变化比较大的情况下获取较高的成功率。

[0005] 然而由于缺乏RNA的三维结构，因此基于模板的方法来构建蛋白质-RNA的相互作用的能力极其的有限，随着越来越多的RNA的三维结构被解析出来，增加了基于模板的方法来构建蛋白质-RNA的能力。而且随着RNA测序技术的发展，发现了很多的RNA，然而大量的RNA其功能还不清楚。另外蛋白质-RNA的三维结构比非结构能够提供更加详细的蛋白质-RNA相互作用机理，然而目前在蛋白质-RNA基于模板建模的领域的计算方法还为空白，这使得开发基于模板来构建蛋白质-RNA的相互作用模型的方法尤为迫切。

发明内容

[0006] 针对现有技术的以上缺陷或改进需求，本发明提供了一种蛋白质-RNA复合物结构预测方法，其目的在于通过构建基于模板的蛋白质-RNA的相互作用模型，由此解决现有技术蛋白质-RNA复合物结构预测方法准确率低、基于模板的蛋白质-RNA相互作用模型计算方法缺乏的技术问题。

[0007] 为实现上述目的，按照本发明的一个方面，提供了一种蛋白质-RNA复合物结构预测方法，包括如下步骤：

[0008] (1)计算模板复合物结构分数：将给定的蛋白质和RNA的单体结构分别与模板库中的蛋白质-RNA相互作用模型模板进行比对，分别得到给定的蛋白质与模板蛋白质的相似分数A，以及给定的RNA与模板RNA相似分数B；对所述相似分数A和相似分数B进行比较，取所述相似分数A和所述相似分数B中较小的相似分数作为利用该模板得到的蛋白质-RNA相互作用模型的复合物结构分数，每一个模板得到一个蛋白质-RNA相互作用模型的复合物结构分

数；

[0009] (2)模型排序：将步骤(1)获得的蛋白质-RNA相互作用模型的复合物结构分数按照降序排列；

[0010] (3)模型判断：预先给定一个阈值，将步骤(2)按照降序排列获得的第一个复合物结构分数，即蛋白质-RNA相互作用模型的复合物结构分数的最大值与所述阈值进行比较，当所述复合物结构分数的最大值小于该阈值，则判断该模型结构不正确，所述给定蛋白质和RNA不能结合；当所述复合物结构分数的最大值大于所述阈值，则判断该蛋白质-RNA相互作用模型结构正确，该给定蛋白质和RNA可以结合。

[0011] 优选地，所述模板库的获得方法为：从PDB数据库中下载到所有的蛋白质-RNA复合物结构，然后从中根据晶体结构分辨率和蛋白质残基以及RNA碱基个数挑选确定模板库。

[0012] 优选地，所述模板库中的蛋白质-RNA相互作用模型晶体结构分辨率比3.0好，所述蛋白质残基个数大于30，所述RNA的碱基个数大于20。

[0013] 优选地，所述模板库中一共有439个蛋白质-RNA相互作用模型模板。

[0014] 优选地，所述给定的蛋白质和模板蛋白质的比对方法为使用TMalign程序来比对。

[0015] 优选地，所述给定的RNA与模板RNA的比对方法为使用SARA程序来比对。

[0016] 优选地，所述SARA程序使用一个归一化的向量来代表RNA的结构，结合RNA的二级结构特征，来比对RNA的二级结构。

[0017] 总体而言，通过本发明所构思的以上技术方案与现有技术相比，能够取得下列有益效果。

[0018] (1)本发明开创性地提出了在基于模板构建蛋白质-RNA的相互作用模型的计算方法和程序，填补了目前的空白。

[0019] (2)测试了本发明基于模板构建蛋白质-RNA相互作用模型的方法PRIME的性能，PRIME的成功率的比对接的方法RPDOCK增加了40％左右，这大大的促进了蛋白质-RNA三维结构领域的发展。

[0020] (3)由于蛋白质-RNA的相互作用跟许多的疾病相关，比如癌症，所以PRIME有可能揭示由于蛋白质-RNA的相互作用而引起的疾病的分子机制。

附图说明

[0021] 图1是本发明蛋白质-RNA相互作用模型的模板库构建的流程图；

[0022] 图2是本发明实施例1的技术方案流程图。

具体实施方式

[0023] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

[0024] 首先阐述一下本发明的原理：首先通过从PDB数据库中挑选出439个的蛋白质-RNA的模板库，将给定的蛋白质和RNA的单体结构分别与模板库中的蛋白质-RNA相互作用模型模板进行比对，分别得到给定的蛋白质与模板蛋白质的相似分数A，以及给定的RNA与模板

RNA相似分数B;对所述相似分数A和相似分数B进行比较,选择A和B中较小的相似分数作为利用该模板得到的蛋白质-RNA相互作用模型的复合物结构分数,每一个模板得到一个蛋白质-RNA相互作用模型的复合物结构分数;将蛋白质-RNA相互作用模型的复合物结构分数按照降序排列;预先给定一个阈值,将按照降序排列获得的第一个复合物结构分数,即蛋白质-RNA相互作用模型的复合物结构分数的最大值与所述阈值进行比较,当所述复合物结构分数的最大值小于该阈值,则判断该模型结构不正确,所述给定蛋白质和RNA不能结合;当所述复合物结构分数的最大值大于所述阈值,则判断该蛋白质-RNA相互作用模型结构正确,该给定蛋白质和RNA可以结合。

[0025] 一种蛋白质-RNA复合物结构预测方法,具体的,一种基于模板构建蛋白质-RNA相互作用模型的方法PRIME,包括如下步骤:

[0026] (1)从PDB数据库挑选确定模板库

[0027] 从PDB数据库下载到所有的蛋白质-RNA复合物结构总共1574个,之后选择出晶体结构的分辨率比3.0好且蛋白质残基和RNA的碱基个数分别大于30和20的结构,并且计算其相互作用的界面残基个数大于5,保留其结构,这里我们得到了344个复合结构,总共2954个蛋白质-RNA的相互作用模板,之后再去掉那些RNA很相似的RNA序列且留下晶体分辨率的最好的模板结构,最终得到439个相互作用模型,作为模板库。

[0028] 如图1所示从PDB数据库下载到所有的蛋白质-RNA复合物结构总共1574个,之后选择出晶体结构的分辨率比3.0好且蛋白质残基和RNA的碱基个数分别大于30和20的结构,并且计算其相互作用的界面残基个数大于5,保留其结构,这里我们得到了344个复合结构,总共2954个蛋白质-RNA的相互作用模板,之后在去掉那些RNA很相似的RNA序列且留下晶体分辨率的最好的模板结构,最终我们得到了439个相互作用模型,并且作为我们的模板库。

[0029] (2)使用蛋白质(RNA)的结构比对所有的模板复合物

[0030] 将蛋白质和RNA单体结构作为程序的输入,本发明使用TMalign程序来比对蛋白质结构,TMalign是一种比对蛋白质结构的方法,得到给定蛋白质和模板的蛋白质的相似分数A;使用SARA程序来比对RNA结构,SARA使用一个归一化的向量来代表RNA的结构,结合RNA的二级结构特征,来比对RNA的二级结构,得到给定RNA和模板的RNA的相似分数B,根据各自的比对,将蛋白质和RNA结构叠加到一个蛋白质-RNA相互作用的模板结构之上,这样就得到了一个蛋白质-RNA相互作用的模型。有一个模板,就有一个模型,因此一共可以得到439个模型。

[0031] (3)模型排序

[0032] 选择上述A和B中较小的相似分数作为利用该模板得到的蛋白质-RNA相互作用模型的复合物结构分数,每一个模板得到一个蛋白质-RNA相互作用模型的复合物结构分数,一共有439个模型,所以对于给定的蛋白质和RNA单体,利用模板一共得到439个模型的复合物结构分数,按照复合物结构分数对439个蛋白质-RNA相互作用模型进行排序,按照降序排列。

[0033] 相似分数代表着这两个结构之间的相似度,分数越高,那么这两个结构就越相似,选择A和B中较小的分数是为了保证这些相互作用是一致的,从而才能由这个模板构建出来的模型是正确的。

[0034] (4)根据阈值和排名来选择模型

[0035] 在对模型进行排序了之后，预先给定一个阈值0.45来判定这个模型的正确性。这个阈值是由PRIME在模板库上测试给出的。判别模型正确与否的标准是：将按照降序排列后的复合物结构分数的最大值与该阈值进行比较，复合物结构分数最大值比阈值小，这样构建出来的模型是不正确的，如果大于这个阈值，那么这个模型是正确的，我们判断这个蛋白质-RNA可以结合，并给出蛋白质-RNA的3D结构。

[0036] 以下为实施例：

[0037] 实施例1

[0038] 图2显示了构建蛋白质-RNA的基于模板的方法来构建相互作用的模型示意图。图2中最上面的蛋白质和RNA单体结构作为程序的输入，然后使用TMalign这个程序蛋白质比对的程序比对蛋白质结构，TMalign是一种比对蛋白质结构的方法；使用SARA这个程序比对RNA结构，而SARA是一种使用一个归一化的向量来代表RNA的结构，结合RNA的二级结构特征，来比对RNA的二级结构。中间的是一个蛋白质-RNA相互作用的模板结构，分别根据各自的比对，将蛋白质和RNA结构叠加到模板结构之上，最终就得到了图1最下面部分的蛋白质-RNA相互作用的模型。实际上输入一个蛋白质(RNA)的结构我们不仅仅得到一个相互作用模型，有一个模板，则有一个模型，因此我们得到439个模型，根据和模板的相似性，我们对模型进行了一个排序。图2中的一个例子就是1A9N_B和1A9N_C使用本发明根据1N78_AC蛋白质-RNA复合物构建出来模型，而且这个排名是第一且相似分数大于阈值0.45，因此判断这个模型是正确的。与由晶体结构给出的模型相比，本发明给出的模型的配体的rmsd是3.0，这就验证了本发明的基于模板的蛋白质-RNA复合物结构预测方法的准确性和实用性。

[0039] 本领域的技术人员容易理解，以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

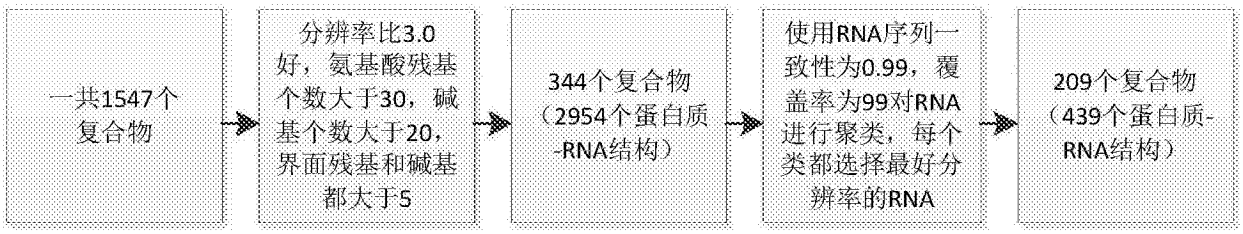

一共1547个复合物
分辨率比3.0好，氨基酸残基个数大于30，碱基个数大于20，界面残基和碱基都大于5
344个复合物（2954个蛋白质-RNA结构）
使用RNA序列一致性为0.99，覆盖率为99对RNA进行聚类，每个类都选择最好分辨率的RNA
209个复合物（439个蛋白质-RNA结构）

图1

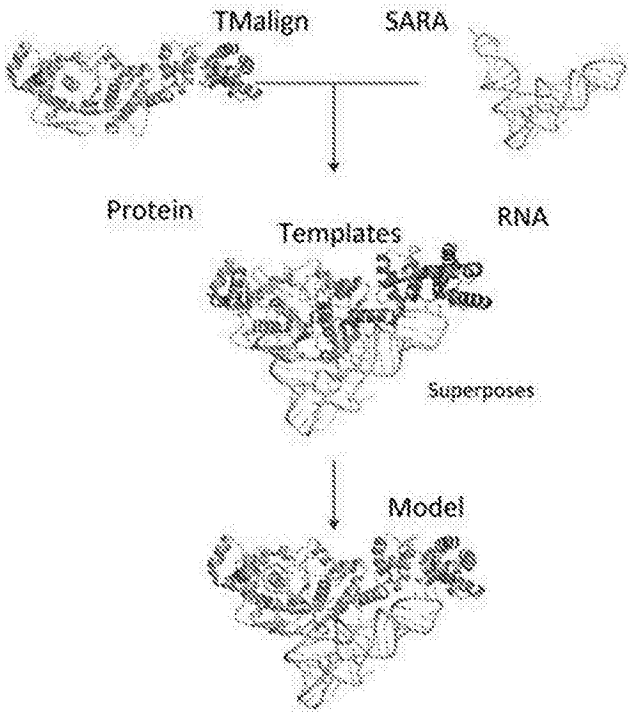

TMalign
SARA
Protein
Templates
RNA
Superposes
Model

图2