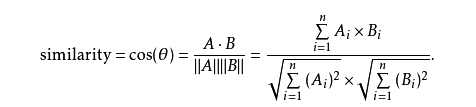
网站建立

该网站（http://www.rnanut.net/lncompare/）使用“Apache，MySQL和PHP框架，搭载在一台Linux服务器上；网站中所有的数据图形可视化使用了开源的G2(https://antv.alipay.com/)完成，数据表格显示及下载用一个名为vis的JavaScript插件（http://visjs.org/）实现。

相似性分析：

为了度量任意两个已知的LncRNA之间的相似性，经过对多种相似度计算结果的对比，最终采用了效果较好的余弦相似度计算方法。



公式中，A,B分别是两个包含了相同特征的值的两个LncRNA的特征向量，即删除了两个向量中有空值的特征后的特征向量。余弦相似度的值就是A,B两个向量的点乘与两个向量二范式乘积的商。

为了进一步刻画两个lncrna间的相似度，在余弦相似度的基础上，又进行了一次互秩相似度计算。

Rank=（AB\*BA）1/2;

其中，AB表示B在与A的余弦相似性高低的降序排序中的序号，BA同理。

富集分析：

Website construction  
This Website（http://www.rnanut.net/lncompare/） is built on a Linux server using the Apache+MySQL+PHP framework; all data graphics visualizations on the site are done using the open source G2 package(https://antv.alipay.com/), and the display and download of the table data is done using a JavaScript plugin called vis（http://visjs.org/）.

Similarity analysis  
In order to measure the similarity between two LncRNAs, after comparing the results of multiple similarity calculation methods, the optimal cosine similarity calculation method is adopted.

In the formula, A and B are two LncRNA feature vectors,which has deleted the NaN values. The value of the cosine similarity is the quotient of the point multiplication of the two vectors of A and B and the product of the Second paradigm of two vectors.

In order to further describe the similarity between two lncrnas, based on the cosine similarity, a mutual rank similarity calculation is performed.

In the formula, denotes the serial number of B in the descending order of Cosine similarity between all lncrnas and A, and is the same.

使用超几何分布算法刻画输入的lncrna在某个特征（数量，分类）上是否有明显富集（p<0.005）.以分类(是否与疾病相关)举例：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Input | Background | Total |
| Related | a | b | a + b |
| Irrelevant | c | d | c + d |
| Total | a + c | b + d | (a+b+c+d)=n |

输入中与疾病相关的lncrnas的数量是a，和疾病无关的lncrnas的数量是c；背景中与疾病相关的lncrnas的个数是b，与疾病无关的lncrnas的个数是d；lncrna的总数量是n。通过该公式计算的p值，断定输入的lncrnas是否在某个特征上有明显富集（p<0.005）。

Use the hypergeometric distribution algorithm to characterize whether the lncrnas inputed have obvious enrichment on a certain feature (quantity , classification). For example: Whether the lncRNAs inputed are related to disease.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Input | Background | Total |
| Related | a | b | a + b |
| Irrelevant | c | d | c + d |
| Total | a + c | b + d | (a+b+c+d)=n |

The number of the inputed lncRNAs that associated with the diseases is a, and that unrelated to the diseases is c; The number of the lncRNAs that included in background and associated with the diseases is b, and that included in background and unrelated to the diseases is d; The total number of lncRNAs is n. The p value calculated by the formula determines whether the input lncrnas are significantly enriched in a certain feature (p < 0.005).