**SNF流程使用文档**

Table of Contents

[一、 SNF背景介绍 1](#_Toc7979027)

[二、 SNF算法原理及功能描述 2](#_Toc7979028)

[三、 SNF流程使用方法 3](#_Toc7979029)

[1. 测试示例 3](#_Toc7979030)

[2. 参数说明及注意事项 4](#_Toc7979031)

[四、 结果解读 5](#_Toc7979032)

[1. 所有样本SNF分型结果的热图展示-［K\*\_alpha\*\_T\*\_C\*.pdf］ 5](#_Toc7979033)

[2. 添加所有样本SNF分型结果信息的新的分组文件-[snf.K\*\_alpha\*\_T\*\_C\*.list] 6](#_Toc7979034)

[3.各组学相似度矩阵文件-[\*.matrix.K\*\_alpha\*.xls] 7](#_Toc7979035)

[4. SNF融合矩阵文件-[W.matrix.K\*\_alpha\*\_T\*.xls] 7](#_Toc7979036)

[五、 参考文献 7](#_Toc7979037)

# **SNF**背景介绍

快速更近的技术使得处理海量且多样化的数据以期发现临床和生物学问题成为可能。比如TCGA已经积累了来自上千病人20多个癌种的基因组学、转录组学、表观组学数据。面对如此庞大的数据，必需运用多组学整合方法来发掘生物学过程和表型的异质性，如乳腺癌不同亚型的发现。多组学数据整合方法至少要克服三个来自计算方面的挑战：1.相对于大量的多组学信息，样本数量通常很少；2.不同组学数据数据的数据级差别很大；3.不同组学数据提供的相互补充信息的利用。目前的多组学数据整合方法旨在解决以上挑战。

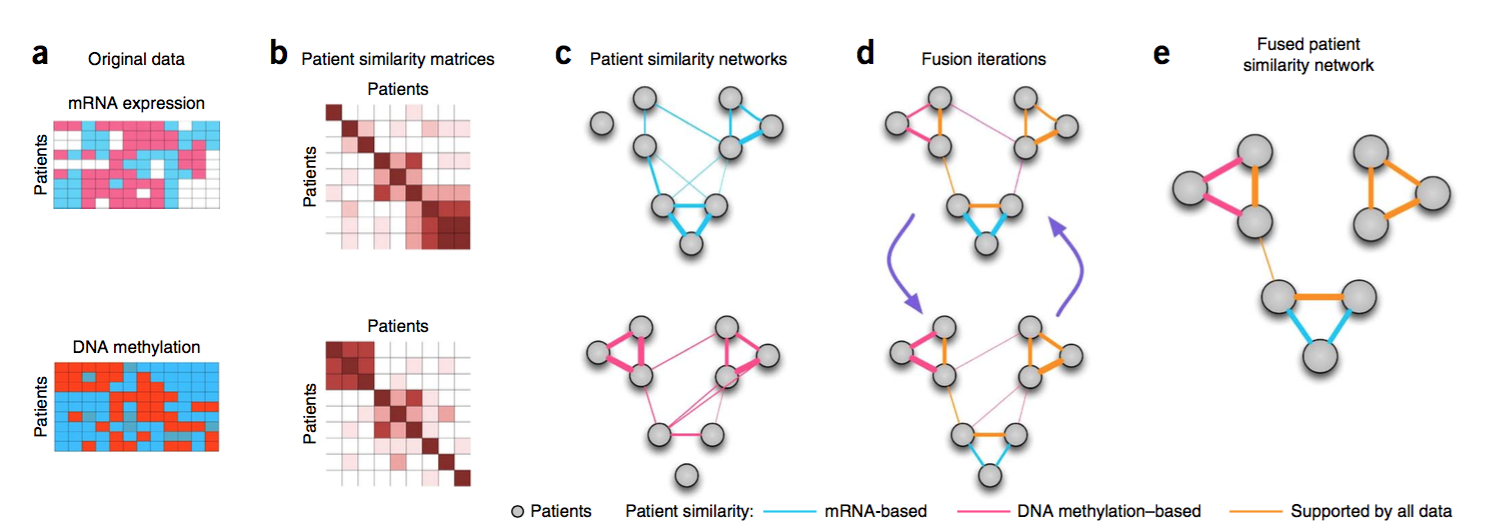
目前的多组学整合方法要么在统一每个样本不同类型数据(如mRNA表达数据、DNA甲基化数据)时会把每种数据类型本来就信噪比很低的信息损失掉，要么先对每种类型的数据单独进行分析，最后却发现得到的结论很不一致难以整合。另外有通过对每种类型的数据预先筛选重要基因以增强信号、再用Consensus Clustering对数据进行整合的方法，然而预先筛选基因会给分析带来偏倚，并且只关注不同数据的共有模式会损失掉不同数据的特有信息。最近的一种机器学习方法iCluster，通过联合水平变量模型进行数据整合分类，该方法和有关的机器学习方法尽管强大，但是不能处理每种数据的所有特征维度，而且对预筛选的基因比较敏感。

SNF(similarity network fusion) 不同于以上方法，它用样本之间的网络关系(相似度)作为整合的基础。比如，整合病人样本数据的时候，SNF会创建一个病人之间的相似度网络。该法第一次提出了通过构建病人相似度网络来整合不同生物学数据：首先构建每种数据的病人相似度网络，再将这些不同的网络用非线性整合方法融合成一个病人相似度网络。

# **SNF算法原理及功能描述**

基于相同的样本集合(如相同病人样本集合)的多组学数据（如各种RNA表达量数据：如果是mRNA来源的数据则输入文件为不同mRNA基因在各病人样本中的表达量矩阵、如果是microRNA来源的数据则输入文件为不同microRNA在各病人样本中的表达量矩阵；又如甲基化数据：不同甲基化位点在各病人中的表达矩阵；…），SNF首先构建每种数据的病人相似度矩阵/网络(计算两两样本相似度距离指标：离散型数据－卡方距离；连续型数据－欧式距离)，再把不同数据来源的相似度矩阵/网络融合为一个相似度矩阵／网络。

具体融合过程如下图所示：通过信号传导理论迭代更新每一个矩阵/网络，使得矩阵/网络之间越来越相似，遍历几轮之后，把所有矩阵/网络融合为一个矩阵/网络。SNF算法的优势在于随着迭代次数的增加，低相似度关联(弱连接关系)逐渐消失，那些在一种或多种矩阵/网络中高相似度的关联(强连接关系)被保留下来，另外，如果低相似度关联(弱连接关系)在多个矩阵／网络中的共同近邻支持度比较高，这样的连接关系也会被保留下来。SNF算法能够充分利用每个矩阵/网络的独有结构，整合所有网络的共有信息以及每个网络的特有信息。



**Figure 1** | Illustrative example of SNF steps.

基于样本之间的相似度矩阵/网络构建，即使对于小的样本集合，SNF也可以从中获取有用信息，并且具有对噪音和数据异质性的鲁棒性，对于海量基因特征矩阵的可延展性。基于SNF整合矩阵，可以有效对样本集合进行亚型划分，也可以对新样本进行亚型预测。基于五个不同癌种的数据测试，SNF得到了一致性高、临床解释度高的病人亚型分类结果，其在亚型划分方面的表现优于当下流行的多组学整合方法以及单组学网络方法。

# **SNF流程使用方法**

## 测试示例

示例一：perl snf.pl -infile inputfile -mf group.mf

示例二：perl snf.pl -K 20:30:10 -Alpha 0.5:0.6:0.1 -T 30:40:10 -C 2:5:1 -infile inputfile -mf group.mf -outdir output --notrun

示例三：perl snf.pl -K 20 -Alpha 0.5 -T 30:40:10 -C 2:5:1 -infile inputfile -mf group.mf -outdir .

## 参数说明及注意事项

［通过perl snf.pl打印help帮助信息］

［标\*参数（--input、--mf）为必需输入文件］

**\*--input** [str] 必需输入文件，不同组学数据的路径及数据类型说明，格式：

第一种组学数据文件的绝对路径 数据类型

第二种组学数据文件的绝对路径 数据类型

第三种组学数据文件的绝对路径 数据类型

注意：1).所有组学数据文件的第一行为样本名，第一列为特征名(如基因/位点名称)，并且所有文件第一行的样本名顺序必须完全一致(有出入的话需要先预处理)。

2).数据类型是指该组学数据是离散型还是连续性，如果是离散型数据，数据类型写“chiDist2”(即通过卡方距离计算样本之间的相似度)，如果是连续型数据，则数据类型写“dist2”（即通过欧式距离计算样本之间的相似度）。

**\*--mf** [str] 必需输入文件，每个样本的分组信息，多种分组可添加多列，如果没有分组信息则只写第一列样本名，注意第一列样本名的顺序必须与—input输入文件的样本名顺序完全一致。格式：

sampleID Group1 Group2 … #表头

sample1 组名 组名 …

sample2 组名 组名 …

**-K** [str] K近邻数目，可以是单独的一个数字或者是一串数字，如果为一串数字，格式为“开始数字：终止数字：步长”，如“10:30:10”表示从数字10开始，以步长10开始遍历到最后一个数字30，即遍历了10、20、30三个数字。该参数设置范围建议：2-N/2，N为样本总数。该参数默认值为20。

**-Alpha** [str] 超参数，用法格式同参数-K。建议设置范围：0.3-0.8，默认值为0.5。

**-T** [str] snf融合迭代的次数，用法格式同参数-K。建议设置范围：10-100，默认值为20。

**-C** [str] cluster划分数目，用法格式同参数-K。建议设置范围：2-8，默认值为0，即由SNF包的自带函数自动计算cluster数目。

**--notrun** 如果配置该参数，则只产生shell不运行。

# **结果解读**

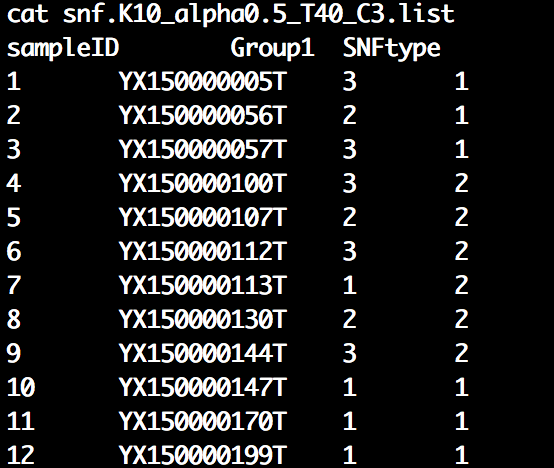
## 1. 所有样本SNF分型结果的热图展示-［K\*\_alpha\*\_T\*\_C\*.pdf］

如下图所示，SNFtype指SNF分型结果的样本标签，Group1代表其他分型结果的样本标签，所有样本按照SNFtype分型结果排序显示。

../../../../../../../Downloads/dNet_temp/doc/K10_alpha0.6_T40

## 2. 添加所有样本SNF分型结果信息的新的分组文件-[snf.K\*\_alpha\*\_T\*\_C\*.list]

该文件向原始分组输入文件(group.mf)的最后一列添加了新的SNFtype分型信息，格式如下：



## 3.各组学相似度矩阵文件-[\*.matrix.K\*\_alpha\*.xls]

如CNV.matrix.K10\_alpha0.5.xls、RNA.matrix.K10\_alpha0.5.xls代表不同类型组学数据在（K:10+alpha:0.5）参数组合下得到的相似度矩阵文件，第一行和第一列均为样本名称，且样本顺序一致。

## 4. SNF融合矩阵文件-[W.matrix.K\*\_alpha\*\_T\*.xls]

如W.matrix.K10\_alpha0.5\_T40.xls代表（K:10+alpha:0.5+T:40）参数组合下得到的融合矩阵文件，第一行和第一列均为样本名称，且样本顺序一致。

# **参考文献**

1.Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature methods, 2014, 11(3): 333-337.