(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113035275 B
(45)授权公告日 2023.08.15

(21)申请号 202110438217.0

(22)申请日 2021.04.22

(65)同一申请的已公布的文献号
申请公布号 CN 113035275 A

(43)申请公布日 2021.06.25

(73)专利权人 广东技术师范大学
地址 510000 广东省广州市天河区石牌中山大道293号

(72)发明人 李振彰 罗文 钟祺楠 翁剑波 黄亮雄 陆海威

(74)专利代理机构 东莞市神州众达专利商标事务所(普通合伙) 44251
专利代理师 周松强

(51)Int.Cl.
G16B 20/50(2019.01)
G06F 16/36(2019.01)
G06F 16/26(2019.01)
G06F 16/28(2019.01)
G06F 17/18(2006.01)

(56)对比文件
CN 105044722 A,2015.11.11
CN 106980763 A,2017.07.25
CN 110379460 A,2019.10.25
US 10052026 B1,2018.08.21
US 2015197785 A1,2015.07.16
US 2016042508 A1,2016.02.11
US 2018060758 A1,2018.03.01
US 2020297323 A1,2020.09.24
罗文 等.基于结合多头注意力机制BiGRU网络的生物医学命名实体识别.计算机应用与软件.2020,151-155.

审查员 杨平勇

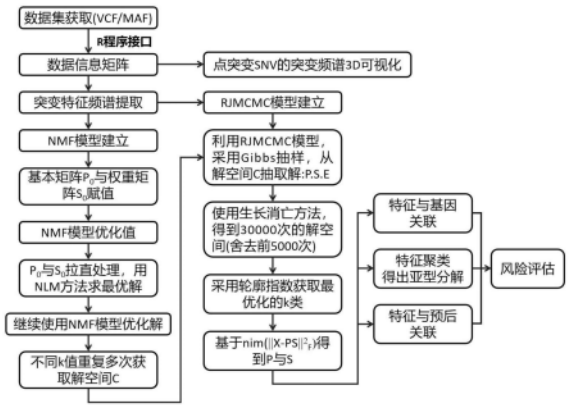
权利要求书5页 说明书11页 附图2页

(54)发明名称

结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法

(57)摘要

本发明提供结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,涉及肿瘤基因特征提取技术领域。该结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,包括以下步骤:S1、数据集获取:突变数据集突变类型包含Somatic SNV和Somatic INDEL,对Somatic SNV和Somatic INDEL进行整体统计使用MuTect软件。该结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,实现注释文件的输入模式,方便使用,节约前期数据处理时间,提高效率,突变频谱3D可视化展示,让研究者可以从空间视觉上直观的看到每个类型的突变情况,增强类型的比较效果展示,创新性结合轮廓系数,进行构建了RJMCMCNMF的模型与算法实现,以及完成代码软件装置设计,实现特征图谱与基因关联获取的软件装置。



CN 113035275 B

1.结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,包括以下步骤:

S1、数据集获取:突变数据集突变类型包含Somatic SNV和Somatic INDEL,对Somatic SNV和Somatic INDEL进行整体统计使用MuTect软件,使用MuTect软件来寻找Somatic SNV和Somatic INDEL位点,对Somatic SNV和Somatic INDEL进行注释使用ANNOVAR或者Oncotator软件,利用ANNOVAR或者Oncotator软件将所检测到SNP以及InDel基因组变异与外部数据库进行注释分析,确定与人类疾病高度相关变异的基因组位置、变异频率、蛋白有害性、基因型杂合性以及所在的功能通路信息;

S2、数据信息矩阵获取:采用具有处理器的计算机,可进行并行运算操作,其中处理器配置成一个R脚本程序接口,基于数据集获取中的文件,选取匹配的参考基因组自动生成信息矩阵;

S3、突变频谱3D可视化展示:采用具有处理器的计算机,可进行并行运算操作,其中处理器配置成一个R脚本程序接口,基于数据信息矩阵获取中获取到的信息矩阵文件,生成该数据集合的突变频谱可视化3D lego图;

S4、突变特征频谱获取:主要包含两个方面,其一是特征提取算法方法,其二是频谱分析软件装置;

S5、特征图谱与基因关联获取:随着特征图谱的分解出来,根据数据集中注释的基因信息实现基因与特征图谱的关联,其实现的途径是确立每个基因非沉默突变对应到某个样本;

S6、特征图谱亚型聚类与预后关联获取:基于系数矩阵信息,获取每个样本对signature的贡献度,基于这些贡献度,可以使用无监督聚类方法对样本进行分类,得到不同的亚型,然后不同亚型与临床信息关联,做预后生存分析,可以找到与预后相关的图谱特征或者与之关联的预后因子。

2.根据权利要求1所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,根据S1中的操作步骤,获取基于参考基因组GRCH37或者GRCH38的注释结果VCF或MAF格式的文件,注释的文件头应该包含至少五列信息:样本名,染色体编号,变异的位点坐标值,参考基因组的碱基和变异后的碱基。

3.根据权利要求1所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,根据S4中的操作步骤,所述特征提取算法方法,包括以下步骤:

S401、确立分析模型:

$X_{m\times n}=P_{m\times k}S_{k\times n}+E_{m\times n}$

约束:$P\geq 0, S\geq 0$

其中$P\in\Re_{+}^{m\times k}$,$S\in\Re_{+}^{k\times n}$,n为样本数目,m为特征类型,$E\in\Re^{m\times n}$;

S402、基于NMF算法的特征解空间的构建:$C_k=\{P,S\}$,表示分类为k的空间集;

S403、可逆跳转蒙特卡罗采样算法模型构建:对于mutational signature分解来讲,其最后得到的类别里边也是96个特征比例图,设最后分解的k个signature就是分层,对于每个signature来讲,其特征是固定的,且每个type对应到signature的概率分配是不一样的,但其分配和为1,对于每个样本来讲,其分配到每个signature的贡献度之和为1,对单个样

本而言，设96个特征为：$y=\{y_1,\ldots,y_{96}\}$

其中y_t为混合数目为k的多元正态混合分布模型$f(y_t)$中抽取的一组随机样本观测值，则包含未知参数Θ的混合模型为：$y_t \sim \sum_{j=1}^{k} w_j f_j(y_t|\Theta)$

由此可得似然函数模型为：

$$p(P,S,E|y,\theta) \propto p(y,P,S,E,\theta) = p(y|P,S,E)p(P|\theta)P(S|\theta)P(E|\theta)$$

$$\propto p(y|P,S,E)\prod_{mk} p(p_{tj})\prod_{kn} p(s_{ji})\prod_{t=1}^{m} p(\varepsilon_t)$$

该模型的先验分布为：

$$s_{ki} \sim Expon(\alpha_{s_{ki}})$$

$$p_{tj} \sim Expon(\alpha_{p_{ij}})$$

$$\varepsilon_t^{-1} \sim G(\gamma,\delta)$$

$i\in[1,n]$，$t\in[1,m]$，$j\in[1,k]$

其中超参数为：$\theta=\{\gamma,\delta,\alpha_{p_{ij}},\alpha_{s_{ji}}\}$；

S404、特征相似性计算方法：$sim(A,B)=\dfrac{\sum_{k=1}^{k} A_k B_K}{\sqrt{\sum_{K=1}^{K}(A_K)^2}\sqrt{\sum_{K=1}^{K}(B_K)^2}}$；

S405、轮廓系数计算：将所有k对应的每个特征作为一个类，通过轮廓系数公式进行这k类数据的评估分析，获取轮廓指数；

S406、运行结果可视化展示方式：将基础矩阵进行归一化后，按照百分比把每个特征属性的柱状图刻画出来，采用不同的颜色进行区分。

4.根据权利要求3所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法，其特征在于，根据S402中的操作步骤，还包括以下步骤：

S4021、随机选取矩阵P0与S0，并且要求P0与S0均是非负定矩阵，归一化处理信息矩阵V0的列，按照V0矩阵的每一分量概率重新生成新的信息矩阵V；

S4022、定义好目标函数模型，模型如下：

$$\arg\min_{P,S} E(\alpha,\beta)=\frac{1}{2}\|V-PS\|_F^2+\alpha\sum_{j\neq i}P_j'P_i+\beta\sum\sum S_{ij}$$；

S4023、获取最优初始解，将矩阵P0与S0进行按列拉直或者按照行拉直，然后按照P0拉直的向量在前，S0拉直的向量在后组成新的向量，该向量作为第二步模型的初始值输入，然后利用R统计软件中的nlm函数进行求最优解；

S4024、处理第三步的最优解，将小于0的分量替换为R统计软件中默认的double型最小的数值，然后根据S4023步骤中的向量拉直规则还原矩阵P与S，得到的P与S作为矩阵分解中的最优初始值；

S4025、获取迭代收敛解，将S4024步骤中得到的P和S，还有S4021步骤中得到的V进行跌

代计算,精度选择为10^-10次方,迭代次数上限约定为100000,计算公式如下:

$$P_{ik} = P_{ik} \cdot \frac{(VS')_{ik}}{(PSS')_{ik}}$$

$$S_{kj} = S_{kj} \cdot \frac{(P'V)_{kj}}{(P'PS)_{kj}}$$;

S4026、选取不同的分解梯度k,重复操作步骤S4021到S4025,针对每个k都重复进行100次试验,记录每次试验的数据结果,结果包括:k,V,P,S,E;

S4027、所有S4026步骤中组成的解空间就是特征提取的解空间。

5.根据权利要求3所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,根据S403中的操作步骤,所述模型的Gibbs抽样约定为如下模型:

I):$p(s_{ji} \mid X, P, S_{-ji}, E) \propto N^{+}(s_{ji} \mid \mu_{s_{ji}}, \sigma^2_{s_{ji}})$

其中$N^{+}(x \mid \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0 \le x < \infty \\ 0, x < 0 \end{cases}$

$X_{ti} = \frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$

$\mu_{s_{ji}}$ *calculate from* $\{S \in C_k\}_{ji}$

$\sigma^2_{s_{ji}} = \frac{1}{N_C} \sum_{c=1}^{N_C} (\frac{\sum_{t=1}^{m} X_{ti} - \frac{1}{n}\sum_{t=1}^{m} \varepsilon_t - \sum_{z=1, z \neq j}^{k} (H^c_{zi} \sum_{t=1}^{m} p_{tz})}{\sum_{t=1}^{m} p_{tj}} - \mu_{s_{ji}})^2$

其中:$H^c_{ji} = s_{ji} \cdot \frac{(P_c^T X)_{ji}}{(P_c^T P S_c)_{ji}}$

$H^c_{ji} = \frac{H^c_{ji}}{\sum_{z=1}^{k} H^c_{zi}}$

II):$p(p_{ij} \mid X, P_{-ij}, S, E) \propto N^{+}(p_{ij} \mid \mu_{p_{ij}}, \sigma^2_{p_{ij}})$

其中$N^{+}(x \mid \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0 \le x < \infty \\ 0, x < 0 \end{cases}$

$X_{ti} = \frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$

$\mu_{p_{ij}}$ *calculate from* $\{P \in C_k\}_{ij}$

$\sigma^2_{p_{ij}} = \frac{1}{N_C} \sum_{c=1}^{N_C} (\frac{\sum_{i=1}^{n} X_{ti} - \varepsilon_t - \sum_{z=1, z \neq j}^{k} (A^c_{iz} \sum_{i=1}^{n} s_{zi})}{\sum_{i=1}^{n} s_{ji}} - \mu_{p_{ij}})^2$

其中:$A^c_{ij} = p_{ij} \cdot \frac{(XS_c^T)_{ij}}{(P_c S_c S_c^T)_{ij}}$

$A^c_{ij} = \frac{A^c_{ij}}{\sum_{z=1}^{m} A^c_{zj}}$

III):$p(\varepsilon_t^{-1} \mid y_t, P_t, S) \propto Gamma(\alpha_t, \beta_t)$

其中:

$$\beta_t = (\delta^{-1} + \frac{1}{2}\sum_{i=1}^{n} (y_{ti} - \sum_{j=1}^{k} p_{tj} s_{ji})^2)^{-1}$$

$\alpha_t = \gamma + \frac{n}{2}$。

6.根据权利要求5所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,根据I)-III),所述抽样实现包括以下步骤:

S4031、使用Gibbs抽样,从$N^+(s_{ji} \mid \mu_{s_{ji}}, \sigma^2_{s_{ji}})$分布中抽取$s_{ji}$;

S4032、使用Gibbs抽样,从$N^+(p_{tj} \mid \mu_{p_{tj}}, \sigma^2_{p_{tj}})$分布中抽取$p_{tj}$;

S4033、使用Gibbs抽样,从Gamma(α_t,β_t)分布中抽取ε_t^{-1};

S4034、更新k,对于k的更新接受规则如下:

设RJMCMCNMF的分解过程,分解维度k的变化看做是状态从C_k跳跃到$C_{k'}$的过程,则设其跳跃的接受概率为:

$$\alpha(k,k') = \min\{1, \frac{p(k',\Theta_{k'} \mid X,\theta) q_k(k,\Theta_k)}{p(k,\Theta_k \mid X,\theta) q_{k'}(k',\Theta_{k'})}\} = \min\{1, \frac{A(k')B(k)}{A(k)B(k')}\}$$

其中

$$A(k) = \ln p(k,\Theta_k \mid X,\theta) \propto \ln p(X \mid k,\theta) + \ln p(P,S,E \mid k,\theta) + \ln p(k)$$

$$\propto -\frac{mn}{2}\ln(2\pi) - \frac{1}{2}n\sum_{t=1}^{m}\ln(\varepsilon_t) - \frac{1}{2}\sum_{i=1}^{n}(X[,i]-PS[,i])^T E^{-1}(X[,i]-PS[,i]) +$$

$$mk\ln(\alpha_p) - \alpha_p\sum_{t=1}^{m}\sum_{j=1}^{k} p_{tj} + nk\ln(\alpha_s) - \alpha_s\sum_{i=1}^{n}\sum_{j=1}^{k} s_{ji} + (\gamma-1)\sum_{t=1}^{m}\ln(\varepsilon_t^{-1}) - \frac{1}{\delta}\sum_{t=1}^{m}\varepsilon_t^{-1} + \ln(\frac{1}{k_{max}-k_{min}})$$

$$B(k) = \ln q_k(k,\Theta_k)$$

$$\propto -\frac{1}{2}\sum_{t=1}^{m}\sum_{j=1}^{k}(\ln(2\pi\sigma^2_{p_{tj}}) + (\frac{p_{tj}-\mu_{p_{tj}}}{\sigma_{p_{tj}}})^2)$$

$$-\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n}(\ln(2\pi\sigma^2_{s_{ji}}) + (\frac{s_{ji}-\mu_{s_{ji}}}{\sigma_{s_{ji}}})^2) +$$

$$\sum_{t=1}^{m}((\alpha_t-1)\ln(\varepsilon_t^{-1}) - (\beta_t\varepsilon_t)^{-1} - \alpha_t\ln(\beta_t) - \ln\Gamma(\alpha_t))$$。

7.根据权利要求6所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,在S4034中的操作步骤中,所述RJMCMCNMF实现包括步骤如下:

1)、设定初始值k0;

2)、计算收敛的初始S0,P0;

3)、通过公式抽样P,S,E;

4)、使用生长消亡方法,u～U(0,1),如果u≤b_k,则进行生长步骤,如果b_k＜u≤b_k+d_k,则进行消亡步骤;

5)、重复以上步骤至设定的迭代步骤(step＝10000,其中前1000次为燃烧期。

8.根据权利要求7所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,在4)中的操作步骤中,所述生长步骤包括以下步骤:

4011)、k＝k0+1;

4012)、执行2),收敛则继续以下步骤;

4013)、从C_k中抽取q_k,即执行3);

4014)、计算α(k0,k);

4015)、计算特征之间的相似性;

4016)、u～U(0,1);

4017)、如果u≤α(k0,k),且两两相似性均小于0.3,则接受k,否则不接受。

9.根据权利要求7所述的结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,其特征在于,在4)中的操作步骤中,所述消亡步骤包括以下步骤:

4021)、k＝k0-1;

4022)、执行2),收敛则继续以下步骤;

4023)、从C_k中抽取q_k,即执行3);

4024)、计算α(k0,k);

4025)、计算特征之间的相似性;

4026)、u～U(0,1);

4027)、如果u≤α(k0,k),且两两相似性均小于0.3,则接受k,否则不接受。

结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法

技术领域

[0001] 本发明涉及肿瘤基因特征提取技术领域,具体为结合轮廓系数和RJMCMC 算法的肿瘤基因点突变的特征提取方法。

背景技术

[0002] 癌症是基因疾病,是由生物体细胞突变引起的。随着基因检测技术例如下一代测序(NGS)的发展,人们发现这些突变是由特定突变特征的组合引起的,这些突变特征通常具有已知的基础过程,它可以更好地提供癌症机制信息,也有助于癌症的预防和治疗。人类基因组是由染色体组成,每个染色体由四种不同的核苷酸组成——A/C/G/T。四个核苷酸实际上形成两对A-T、C-G,当A位于一个链上时,T位于另一个链上,当G位于一个链上时,C必须在同一位置组成。当癌症基因组发生突变时,其中一个核苷酸被另一个核苷酸交换,例如,T被A取代。除了替换(如插入和删除)之外,还有其他突变。突变可能是有缺陷的DNA修复或不同的突变过程的结果,如突变暴露(辐射,吸烟),DNA的酶修饰等。实际上大多数突变都是无害的。按照突变的类型可以分为六大类,分别为C>A(表示有C变异成A),C>G,C>T,T>A, T>C和T>G,按照三碱基核算则可以分为96种不同的突变类型。突变性特征是由不同的突变过程引起的突变类型的某种组合,然后除以该签名引起的突变总数,以便最终考虑每种突变类型的比例贡献。研究表明,某些突变类型在特定癌症中发生更为频繁。例如对肺和皮肤肿瘤中突变的肿瘤基因的分析表明,发现的突变类型与烟草致癌物和紫外光的实验结果相匹配,这主要是已知的外源性致癌物质影响着这些突变类型。值得注意的是,C:G>A:T突变在吸烟相关的肺癌中占主导地位,而C:G>T:A主要发生在dipyrimidines 和CC:GG>TT:AA双核苷酸替代是常见的紫外线光相关皮肤癌的变化特征。因此,从基因组突变数据中寻找这些特征对于发现癌症的基本机制,做好预防和治疗非常重要。

[0003] 目前,NMF即非负矩阵分解法是很多研究者关注的重点。NMF的基本原理是将信号矩阵分解为基本矩阵和相应的系数矩阵,根据代价函数来计算各个信号成分所对应的基本矩阵和系数矩阵,从而实现信号的分离。当下,研究工作者合理地认为在细胞中发生的生化过程通常是独立作用的,因此,可以假设基因组中的突变是细胞中所有突变过程活动的总和,其数据是所有检测样本的不同突变类型的突变计数和,即为观测到的信号矩阵Y。给定模型, Y＝WX,其中W为系数矩阵,也就是不同签名的集合,可以理解为Mutational Signature,X为基本矩阵,也就是决定其活动的强度,代表的是每个样本在每个Mutational Signature的贡献度。

[0004] NMF的优点是稳定性功能,它很好地确定了正确的签名数,由其衍生出一些生物学方法,专门应用于肿瘤特征图谱的提取的,比如NMF、BayeNMF、 SigProfiler以及SignatureAnalyzer等。但在大多数人类癌症类型中,DNA 损伤和修复过程所印迹的突变特征受到非常有限的表征,而且这些方法存在一定的局限性,功能相对单一,且对于一些数据集的分析结果差强人意,尤其是小样本数据或者低深度数据的结果,误差比较大。

发明内容

[0005] 本发明提供的发明目的在于提供结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法，该提供结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法具有更全面的分析内容，适用于大小样本数据集，稳定性高，操作方便，包括从基因突变数据注释结果(MAF或VCF格式，参考基因组可以是GRCH37或GRCH38)生成信号矩阵和突变集合的三维可视化图，基于轮廓系数与RJMCMC方法的图谱特征提取，和基因与突变特征图谱的关联研究功能。

[0006] 为了实现上述的效果，本发明提供如下技术方案：结合轮廓系数和RJMCMC 算法的肿瘤基因点突变的特征提取方法，包括以下步骤：

[0007] S1、数据集获取：突变数据集突变类型包含Somatic SNV和Somatic INDEL，对Somatic SNV和Somatic INDEL进行整体统计使用MuTect软件，使用MuTect 软件来寻找Somatic SNV和Somatic INDEL位点，对Somatic SNV和Somatic INDEL进行注释使用ANNOVAR或者Oncotator软件，利用ANNOVAR或者 Oncotator软件将所检测到SNP以及InDel基因组变异与外部数据库进行注释分析，确定与人类疾病高度相关变异的基因组位置、变异频率、蛋白有害性、基因型杂合性以及所在的功能通路信息；

[0008] S2、数据信息矩阵获取：采用具有处理器的计算机，可进行并行运算操作，其中处理器配置成一个R脚本程序接口，基于数据集获取中的文件，选取匹配的参考基因组自动生成信息矩阵；

[0009] S3、突变频谱3D可视化展示：采用具有处理器的计算机，可进行并行运算操作，其中处理器配置成一个R脚本程序接口，基于数据信息矩阵获取中获取到的信息矩阵文件，生成该数据集合的突变频谱可视化3D lego图；

[0010] S4、突变特征频谱获取：主要包含两个方面，其一是特征提取算法方法，其二是频谱分析软件装置；

[0011] S5、特征图谱与基因关联获取：随着特征图谱的分解出来，根据数据集中注释的基因信息实现基因与特征图谱的关联，其实现的途径是确立每个基因非沉默突变对应到某个样本；

[0012] S6、特征图谱亚型聚类与预后关联获取：基于系数矩阵信息，获取每个样本对signature的贡献度，基于这些贡献度，可以使用无监督聚类方法对样本进行分类，得到不同的亚型，然后不同亚型与临床信息关联，做预后生存分析，可以找到与预后相关的图谱特征或者与之关联的预后因子。

[0013] 进一步的，根据S1中的操作步骤，获取基于参考基因组GRCH37或者GRCH38的注释结果VCF或MAF格式的文件，注释的文件头应该包含至少五列信息：样本名，染色体编号，变异的位点坐标值，参考基因组的碱基和变异后的碱基。

[0014] 进一步的，根据S4中的操作步骤，所述特征提取算法方法，包括以下步骤：

[0015] S401、确立分析模型：

[0016] $X_{m\times n}=P_{m\times k}S_{k\times n}+E_{m\times n}$

[0017] 约束：P≥0，S≥0

[0018] 其中$P\in\Re_{+}^{m\times k}$，$S\in\Re_{+}^{k\times n}$，n为样本数目，m为特征类型，$E\in\Re^{m\times n}$；

[0019] S402、基于NMF算法的特征解空间的构建：$C_k=\{P,S\}$，表示分类为k的空间集；

[0020] S403、可逆跳转蒙特卡罗采样算法模型构建：对于mutational signature 分解来讲，其最后得到的类别里边也是96个特征比例图，设最后分解的k个 signature就是分层，对于每个signature来讲，其特征是固定的，且每个type对应到signature的概率分配是不一样的，但其分配和为1，对于每个样本来讲，其分配到每个signature的贡献度之和为1，对单个样本而言，设 96个特征为：$y=\{y_1,\ldots,y_{96}\}$

[0021] 其中y_t为混合数目为k的多元正态混合分布模型$f(y_t)$中抽取的一组随机样本观测值，则包含未知参数Θ的混合模型为：$y_t \sim \sum_{j=1}^{k} w_j f_j(y_t|\Theta)$

[0022] 由此可得似然函数模型为：

$$p(P,S,E|y,\theta) \propto p(y,P,S,E,\theta) = p(y|P,S,E)p(P|\theta)P(S|\theta)P(E|\theta)$$

[0023]

$$\propto p(y|P,S,E)\prod_{mk} p(p_{tj})\prod_{kn} p(s_{ji})\prod_{t=1}^{m} p(\varepsilon_t)$$

，

[0024] 该模型的先验分布为：

[0025] $$s_{ki} \sim Expon(\alpha_{s_{ki}})$$

[0026] $$p_{tj} \sim Expon(\alpha_{p_{tj}})$$

[0027] $$\varepsilon_t^{-1} \sim G(\gamma,\delta)$$

[0028] $i\in[1,n]$，$t\in[1,m]$，$j\in[1,k]$

[0029] 其中超参数为：$\theta=\{\gamma,\delta,\alpha_{p_{tj}},\alpha_{s_{ji}}\}$；

[0030] S404、特征相似性计算方法：$$sim(A,B)=\frac{\sum_{k=1}^{k} A_k B_K}{\sqrt{\sum_{K=1}^{K}(A_K)^2}\sqrt{\sum_{K=1}^{K}(B_K)^2}}$$；

[0031] S405、轮廓系数计算：将所有k对应的每个特征作为一个类，通过轮廓系数公式进行这k类数据的评估分析，获取轮廓指数；

[0032] S406、运行结果可视化展示方式：将基础矩阵进行归一化后，按照百分比把每个特征属性的柱状图刻画出来，采用不同的颜色进行区分。

[0033] 进一步的，S4021、随机选取矩阵P0与S0，并且要求P0与S0均是非负定矩阵，归一化处理信息矩阵V0的列，按照V0矩阵的每一分量概率重新生成新的信息矩阵V；

[0034] S4022、定义好目标函数模型，模型如下：

[0035] $$\arg\min_{P,S} E(\alpha,\beta)=\frac{1}{2}\|V-PS\|_F^2+\alpha\sum_{j\neq i}P_j'P_i+\beta\sum\sum S_{ij}$$；

[0036] S4023、获取最优初始解，将矩阵P0与S0进行按列拉直或者按照行拉直，然后按照P0拉直的向量在前，S0拉直的向量在后组成新的向量，该向量作为第二步模型的初始值输入，然后利用R统计软件中的nlm函数进行求最优解；

[0037] S4024、处理第三步的最优解，将小于0的分量替换为R统计软件中默认的double型最小的数值，然后根据S4023步骤中的向量拉直规则还原矩阵P 与S，得到的P与S作为矩阵分解中的最优初始值；

[0038] S4025、获取迭代收敛解，将S4024步骤中得到的P和S，还有S4021步骤中得到的V进行跌代计算，精度选择为10^-10次方，迭代次数上限约定为100000，计算公式如下：

[0039] $$P_{ik} = P_{ik} \cdot \frac{(VS')_{ik}}{(PSS')_{ik}}$$

[0040] $$S_{kj} = S_{kj} \cdot \frac{(P'V)_{kj}}{(P'PS)_{kj}}$$ ；

[0041] S4026、选取不同的分解梯度k，重复操作步骤S4021到S4025，针对每个k都重复进行100次试验，记录每次试验的数据结果，结果包括:k,V,P, S,E；

[0042] S4027、所有S4026步骤中组成的解空间就是特征提取的解空间。

[0043] 进一步的，根据S403中的操作步骤，所述模型的Gibbs抽样约定为如下模型：

[0044] Ⅰ)： $$p(s_{ji} \mid X, P, S_{-ji}, E) \propto N^{+}(s_{ji} \mid \mu_{s_{ji}}, \sigma^2_{s_{ji}})$$

[0045] 其中 $$N^{+}(x \mid \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0 \le x < \infty \\ 0, x < 0 \end{cases}$$

[0046] $$X_{ti} = \frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$$

[0047] $$\mu_{s_{ji}} \; calculate \; from \; \{S \in C_k\}_{ji}$$

[0048] $$\sigma^2_{s_{ji}} = \frac{1}{N_C} \sum_{c=1}^{N_C} (\frac{\sum_{t=1}^{m} X_{ti} - \frac{1}{n}\sum_{t=1}^{m} \varepsilon_t - \sum_{z=1, z\neq j}^{k} (H^c_{zi} \sum_{t=1}^{m} p_{tz})}{\sum_{t=1}^{m} p_{tj}} - \mu_{s_{ji}})^2$$

[0049] 其中：$$H^c_{ji} = s_{ji} \cdot \frac{(P_c^T X)_{ji}}{(P_c^T P_c S_c)_{ji}}$$

[0050] $$H^c_{ji} = \frac{H^c_{ji}}{\sum_{z=1}^{k} H^c_{zi}}$$

[0051] Ⅱ)： $$p(p_{tj} \mid X, P_{-tj}, S, E) \propto N^{+}(p_{tj} \mid \mu_{p_{tj}}, \sigma^2_{p_{tj}})$$

[0052] 其中 $$N^{+}(x \mid \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0 \le x < \infty \\ 0, x < 0 \end{cases}$$

[0053] $$X_{ti} = \frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$$

[0054] $$\mu_{p_{tj}} \; calculate \; from \; \{P \in C_k\}_{tj}$$

[0055] $$\sigma^2_{p_{tj}} = \frac{1}{N_C} \sum_{c=1}^{N_C} (\frac{\sum_{i=1}^{n} X_{ti} - \varepsilon_t - \sum_{z=1, z\neq j}^{k} (A^c_{tz} \sum_{i=1}^{n} s_{zi})}{\sum_{i=1}^{n} s_{ji}} - \mu_{p_{tj}})^2$$

[0056] 其中：$$A^c_{tj} = p_{tj} \cdot \frac{(XS_c^T)_{tj}}{(P_c S_c S_c^T)_{tj}}$$

[0057] $$A^c_{tj} = \frac{A^c_{tj}}{\sum_{z=1}^{m} A^c_{zj}}$$

[0058] Ⅲ)： $$p(\varepsilon_t^{-1} \mid y_t, P_t, S) \propto Gamma(\alpha_t, \beta_t)$$

[0059] 其中：

[0060] $$\beta_t = (\delta^{-1} + \frac{1}{2}\sum_{i=1}^{n} (y_{ti} - \sum_{j=1}^{k} p_{tj} s_{ji})^2)^{-1}$$

[0061] $\alpha_t = \gamma + \frac{n}{2}$。

[0062] 进一步的，根据I)-III)，所述抽样实现包括以下步骤：

[0063] S4031、使用Gibbs抽样，从$N^+(s_{ji} \mid \mu_{s_{ji}}, \sigma^2_{s_{ji}})$分布中抽取$s_{ji}$；

[0064] S4032、使用Gibbs抽样，从$N^+(p_{tj} \mid \mu_{p_{tj}}, \sigma^2_{p_{tj}})$分布中抽取$p_{tj}$；

[0065] S4033、使用Gibbs抽样，从Gamma(α_t,β_t)分布中抽取ε_t^{-1}；

[0066] S4034、更新k，对于k的更新接受规则如下：

[0067] 设RJMCMCNMF的分解过程，分解维度k的变化看做是状态从C_k跳跃到$C_{k'}$的过程，则设其跳跃的接受概率为：

[0068] $$\alpha(k,k') = \min\{1, \frac{p(k',\Theta_{k'} \mid X,\theta) q_k(k,\Theta_k)}{p(k,\Theta_k \mid X,\theta) q_{k'}(k',\Theta_{k'})}\} = \min\{1, \frac{A(k')B(k)}{A(k)B(k')}\}$$

[0069] 其中

[0070] $A(k) = \ln p(k,\Theta_k \mid X,\theta) \propto \ln p(X \mid k,\theta) + \ln p(P,S,E \mid k,\theta) + \ln p(k)$

[0071] $$\propto -\frac{mn}{2}\ln(2\pi) - \frac{1}{2}n\sum_{t=1}^{m}\ln(\varepsilon_t) - \frac{1}{2}\sum_{i=1}^{n}(X[,i]-PS[,i])^T E^{-1}(X[,i]-PS[,i]) +$$
$$mk\ln(\alpha_p) - \alpha_p\sum_{t=1}^{m}\sum_{j=1}^{k}p_{tj} + nk\ln(\alpha_s) - \alpha_s\sum_{i=1}^{n}\sum_{j=1}^{k}s_{ji} + (\gamma-1)\sum_{t=1}^{m}\ln(\varepsilon_t^{-1}) - \frac{1}{\delta}\sum_{t=1}^{m}\varepsilon_t^{-1} + \ln(\frac{1}{k_{\max}-k_{\min}})$$

[0072] $$B(k) = \ln q_k(k,\Theta_k)$$
$$\propto -\frac{1}{2}\sum_{t=1}^{m}\sum_{j=1}^{k}(\ln(2\pi\sigma^2_{p_{tj}}) + (\frac{p_{tj}-\mu_{p_{tj}}}{\sigma_{p_{tj}}})^2)$$
$$-\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n}(\ln(2\pi\sigma^2_{s_{ji}}) + (\frac{s_{ji}-\mu_{s_{ji}}}{\sigma_{s_{ji}}})^2) +$$
$$\sum_{t=1}^{m}((\alpha_t-1)\ln(\varepsilon_t^{-1}) - (\beta_t\varepsilon_t)^{-1} - \alpha_t\ln(\beta_t) - \ln\Gamma(\alpha_t))$$。

[0073] 进一步的，在S4034中的操作步骤中，所述RJMCMCNMF实现包括步骤如下：

[0074] 1)、设定初始值k0；

[0075] 2)、计算收敛的初始S0,P0；

[0076] 3)、通过公式抽样P,S,E；

[0077] 4)、使用生长消亡方法，u～U(0,1)，如果u≤b_k，则进行生长步骤，如果 b_k＜u≤b_k+d_k，则进行消亡步骤；

[0078] 5)、重复以上步骤至设定的迭代步骤(step＝10000，其中前1000次为燃烧期。

[0079] 进一步的，在4)中的操作步骤中，所述生长步骤包括以下步骤：

[0080] 4011)、k＝k0+1；

[0081] 4012)、执行2)，收敛则继续以下步骤；

[0082] 4013)、从C_k中抽取q_k，即执行3)；

[0083] 4014)、计算α(k0,k)；

[0084] 4015)、计算特征之间的相似性；

[0085] 4016)、u～U(0,1)；

[0086] 4017)、如果u≤α(k0,k)，且两两相似性均小于0.3，则接受k，否则不接受。

[0087] 进一步的，在4)中的操作步骤中，所述消亡步骤包括以下步骤：

[0088] 4021)、k＝k0-1；

[0089] 4022)、执行2),收敛则继续以下步骤;

[0090] 4023)、从C_k中抽取q_k,即执行3);

[0091] 4024)、计算α(k0,k);

[0092] 4025)、计算特征之间的相似性;

[0093] 4026)、u～U(0,1);

[0094] 4027)、如果u≤α(k0,k),且两两相似性均小于0.3,则接受k,否则不接受。

[0095] 本发明提供了结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,具备以下有益效果:

[0096] 该结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,实现注释文件的输入模式,方便使用,节约前期数据处理时间,提高效率,突变频谱3D可视化展示,让研究者可以从空间视觉上直观的看到每个类型的突变情况,增强类型的比较效果展示,创新性结合轮廓系数,进行构建了RJMCMCNMF 的模型与算法实现,以及完成代码软件装置设计,实现特征图谱与基因关联获取的软件装置,实现特征图谱亚型聚类与预后关联获取的软件装置。

附图说明

[0097] 图1为整体流程图;

[0098] 图2为突变频谱3D可视化展示图;

[0099] 图3为运行结果可视化展示图。

具体实施方式

[0100] 本发明提供一种技术方案:结合轮廓系数和RJMCMC算法的肿瘤基因点突变的特征提取方法,包括以下步骤:

[0101] 步骤一:数据集获取:突变数据集突变类型包含Somatic SNV和Somatic INDEL,对Somatic SNV/InDel进行整体统计使用MuTect软件,使用MuTect 软件来寻找Somatic SNV和InDel位点;对Somatic SNV/InDel进行注释使用ANNOVAR或者Oncotator软件,利用ANNOVAR/Oncotator软件将所检测到 SNP以及InDel等基因组变异与外部数据库进行注释分析,以确定与人类疾病高度相关变异的基因组位置、变异频率、蛋白有害性、基因型杂合性以及所在的功能通路等信息,获取基于参考基因组GRCH37或者GRCH38的注释结果VCF或MAF格式的文件,注释的文件头应该包含至少五列信息:样本名,染色体编号,变异的位点坐标值,参考基因组的碱基,变异后的碱基。

[0102] 步骤二:数据信息矩阵获取:采用具有处理器的计算机,可进行并行运算操作,其中处理器配置成一个R脚本程序接口,步骤一)中的文件,选取匹配的参考基因组就可以自动生成信息矩阵,信息矩阵包含三部分:a)突变信息矩阵,其中行代表属性,比如以6种碱基突变类型为中心,各取5’和3’各一个碱基形成多种组合,该组合有96种类型,以这96种突变类型为基础,确定肿瘤基因组的突变特征信息矩阵,矩阵的列代表每一个样本;b)样本列表文件,与a)中的列一致;c)行属性名称列表,与a)中的行一致。

[0103] 步骤三:突变频谱3D可视化展示:采用具有处理器的计算机,可进行并行运算操作,其中处理器配置成一个R脚本程序接口,步骤二)获取到的信息矩阵文件,生成该数据集

合的突变频谱可视化3D lego图。该部分主要功能是展示该样本数据集中，突变类型在每Mp基因组发生的突变频率，主要计算公式如下：该突变类型的每Mp基因组的突变频率＝该突变类型的突变数据集总数/基因组长度(Mp)；空间转化勾勒函数主要采取勾股定理，根据按比例进行缩放的每Mp基因组发生的突变频率，进行空间描点，从而实现3D的方柱，代表该突变类型的每Mp基因组的突变频率，结果展示图如附图2所示。

[0104] 步骤四：突变特征频谱获取：该部分主要包含两个方面，其一是特征提取算法方法，其二是频谱分析软件装置。

[0105] 关于特征提取算法方法，具体技术方案如下：

[0106] 确立分析模型：

[0107] $X_{m\times n}=P_{m\times k}S_{k\times n}+E_{m\times n}$

[0108] 约束：P≥0，S≥0

[0109] 其中 $P\in\Re_{+}^{m\times k}, S\in\Re_{+}^{k\times n}$，n为样本数目，m为特征类型，$E\in\Re^{m\times n}$。

[0110] 基于NMF算法的特征解空间的构建：

[0111] $C_k=\{P,S\}$，表示分类为k的空间集；

[0112] 其中解空间的定义求解如下：

[0113] 第一步：随机选取矩阵P0与S0，并且要求P0与S0均是非负定矩阵，归一化处理信息矩阵V0的列，按照V0矩阵的每一分量概率重新生成新的信息矩阵V；

[0114] 第二步：定义好目标函数模型，模型如下：

[0115] $$\arg\min_{P,S} E(\alpha,\beta)=\frac{1}{2}\|V-PS\|_F^2+\alpha\sum_{j\neq i}P_j'P_i+\beta\sum\sum S_{ij}$$；

[0116] 第三步：获取最优初始解，将矩阵P0与S0进行按列拉直或者按照行拉直，然后按照P0拉直的向量在前，S0拉直的向量在后组成新的向量，该向量作为第二步模型的初始值输入，然后利用R统计软件中的nlm函数进行求最优解；

[0117] 第四步：适当处理第三步的最优解，将小于0的分量替换为R统计软件中默认的double型最小的数值，然后根据第三步的向量拉直规则还原矩阵P 与S，这步得到的P与S作为矩阵分解中的最优初始值；

[0118] 第五步：获取迭代收敛解，将第四步得到的P，S还有第一步得到的V进行跌代计算，精度选择为10^-10次方，迭代次数上限约定为100000，计算公式如下：

[0119] $$P_{ik}=P_{ik}\cdot\frac{(VS')_{ik}}{(PSS')_{ik}}$$

[0120] $$S_{kj}=S_{kj}\cdot\frac{(P'V)_{kj}}{(P'PS)_{kj}}$$；

[0121] 第六步：选取不同的分解梯度k(范围应该固定在1到30)，重复操作步骤第一到第五步，针对每个k都重复进行100次试验，记录每次试验的数据结果，结果包括：k,V,P,S,E；

[0122] 第七步：所有第六步组成的解空间就是特征提取的解空间。

[0123] 可逆跳转蒙特卡罗采样(RJMCMC)算法模型构建：

[0124] 对于mutational signature分解来讲，其最后得到的类别里边也是96 个特征比

例图，因此这里假设最后分解的k个signature就是分层。理想状态下，对于每个signature来讲，其特征是固定的，且每个type对应到 signature的概率分配是不一样的，但其分配和为1，对于每个样本来讲，其分配到每个signature的贡献度之和为1。对单个样本而言，假设96个特征为：

[0125] $y=\{y_1,\ldots,y_{96}\}$

[0126] 其中y_t为混合数目为k的多元正态混合分布模型$f(y_t)$中抽取的一组随机样本观测值，则包含未知参数Θ的混合模型为：

[0127] $$y_t \sim \sum_{j=1}^{k} w_j f_j(y_t \mid \Theta)$$ ；

[0128] 由此可得似然函数模型为：

$$p(P,S,E \mid y,\theta) \propto p(y,P,S,E,\theta) = p(y \mid P,S,E)p(P \mid \theta)P(S \mid \theta)P(E \mid \theta)$$

[0129] $$\propto p(y \mid P,S,E)\prod_{mk} p(p_{tj})\prod_{kn} p(s_{ji})\prod_{t=1}^{m} p(\varepsilon_t)$$ ；

[0130] 该模型的先验分布为：

[0131] $$s_{ki} \sim Expon(\alpha_{s_{ki}})$$

[0132] $$p_{tj} \sim Expon(\alpha_{p_{tj}})$$

[0133] $$\varepsilon_t^{-1} \sim G(\gamma,\delta)$$

[0134] $i\in[1,n]$，$t\in[1,m]$，$j\in[1,k]$；

[0135] 其中超参数为：$\theta=\{\gamma,\delta,\alpha_{p_{tj}},\alpha_{s_{ji}}\}$。

[0136] 该模型的Gibbs抽样约定为如下模型：

[0137] Ⅰ）：$$p(s_{ji} \mid X,P,S_{-ji},E) \propto N^{+}(s_{ji} \mid \mu_{s_{ji}},\sigma_{s_{ji}}^2)$$

[0138] 其中$$N^{+}(x \mid \mu,\sigma^2)=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0\le x<\infty \\ 0, x<0\end{cases}$$

[0139] $$X_{ti}=\frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$$

[0140] $\mu_{s_{ji}}$ *calculate from* $\{S\in C_k\}_{ji}$

[0141] $$\sigma_{s_{ji}}^2=\frac{1}{N_C}\sum_{c=1}^{N_C}(\frac{\sum_{t=1}^{m}X_{ti}-\frac{1}{n}\sum_{t=1}^{m}\varepsilon_t-\sum_{z=1,z\neq j}^{k}(H_{zi}^{c}\sum_{t=1}^{m}p_{tz})}{\sum_{t=1}^{m}p_{tj}}-\mu_{s_{ji}})^2$$

[0142] 其中：$$H_{ji}^{c}=s_{ji}\cdot\frac{(P_c^T X)_{ji}}{(P_c^T P_c S_c)_{ji}}$$

[0143] $$H_{ji}^{c}=\frac{H_{ji}^{c}}{\sum_{z=1}^{k}H_{zi}^{c}}$$

[0144] ⅠⅠ）：$$p(p_{tj} \mid X,P_{-tj},S,E) \propto N^{+}(p_{tj} \mid \mu_{p_{tj}},\sigma_{p_{tj}}^2)$$

[0145] 其中$$N^{+}(x \mid \mu,\sigma^2)=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2), 0\le x<\infty \\ 0, x<0\end{cases}$$

[0146] $$X_{ti}=\frac{y_{ti}}{\sum_{z=1}^{n} y_{tz}}$$

[0147] $\mu_{p_{ij}}\ calculate\ from\ \{P \in C_k\}_{ij}$

[0148] $$\sigma^2_{p_{ij}} = \frac{1}{N_C}\sum_{c=1}^{N_C}\left(\frac{\sum_{i=1}^{n} X_{ti} - \varepsilon_t - \sum_{z=1,z\neq j}^{k}\left(A_{tz}^{c}\sum_{i=1}^{n} s_{zi}\right)}{\sum_{i=1}^{n} s_{ji}} - \mu_{p_{ij}}\right)^2$$

[0149] 其中：$A_{tj}^{c} = p_{tj}\cdot\frac{(XS_c^T)_{tj}}{(P_cS_cS_c^T)_{tj}}$

[0150] $$A_{tj}^{c} = \frac{A_{tj}^{c}}{\sum_{z=1}^{m} A_{zj}^{c}}$$

[0151] III)： $p(\varepsilon_t^{-1} \mid y_t, P_t, S) \propto Gamma(\alpha_t, \beta_t)$

[0152] 其中：

[0153] $$\beta_t = \left(\delta^{-1} + \frac{1}{2}\sum_{i=1}^{n}\left(y_{ti} - \sum_{j=1}^{k} p_{tj}s_{ji}\right)^2\right)^{-1}$$

[0154] $$\alpha_t = \gamma + \frac{n}{2}$$。

[0155] 以上的I,II,III的具体抽样实现步骤如下：

[0156] 1)、使用Gibbs抽样，从$N^{+}(s_{ji} \mid \mu_{s_{ji}}, \sigma^2_{s_{ji}})$分布中抽取$sj_i$；

[0157] 2)、使用Gibbs抽样，从$N^{+}(p_{tj} \mid \mu_{p_{tj}}, \sigma^2_{p_{tj}})$分布中抽取$p_{tj}$；

[0158] 3)、使用Gibbs抽样，从Gamma(α_t,β_t)分布中抽取ε_t^{-1}；

[0159] 4)、更新k，

[0160] 注意：对每一个k∈[k_{min},k_{max}]，存在一个与之匹配的参数Θ_k={P,S,E}，那么对于同一个k值，则存在此k值的参数集C_k={Θ_k}，那么对于所有的k，则有参数集为$C = \bigcup_{k_{min}}^{k_{max}} C_k$。

[0161] 对于以上的k的更新接受规则如下：

[0162] 假设RJMCMCNMF的分解过程，分解维度k的变化看做是状态从C_k跳跃到 $C_{k'}$的过程，则设其跳跃的接受概率为：

[0163] $$\alpha(k,k') = \min\{1, \frac{p(k',\Theta_{k'} \mid X,\theta)q_k(k,\Theta_k)}{p(k,\Theta_k \mid X,\theta)q_{k'}(k',\Theta_{k'})}\} = \min\{1, \frac{A(k')B(k)}{A(k)B(k')}\}$$

[0164] 其中

[0165] $$A(k) = \ln p(k,\Theta_k \mid X,\theta) \propto \ln p(X \mid k,\theta) + \ln p(P,S,E \mid k,\theta) + \ln p(k)$$
$$\propto -\frac{mn}{2}\ln(2\pi) - \frac{1}{2}n\sum_{t=1}^{m}\ln(\varepsilon_t) - \frac{1}{2}\sum_{t=1}^{n}(X[,i] - PS[,i])^T E^{-1}(X[,i] - PS[,i]) +$$
$$mk\ln(\alpha_p) - \alpha_p\sum_{t=1}^{m}\sum_{j=1}^{k} p_{tj} + nk\ln(\alpha_s) - \alpha_s\sum_{i=1}^{n}\sum_{j=1}^{k} s_{ji} + (\gamma-1)\sum_{t=1}^{m}\ln(\varepsilon_t^{-1}) - \frac{1}{\delta}\sum_{t=1}^{m}\varepsilon_t^{-1} + \ln(\frac{1}{k_{max}-k_{min}})$$

[0166] $$B(k) = \ln q_k(k,\Theta_k)$$
$$\propto -\frac{1}{2}\sum_{t=1}^{m}\sum_{j=1}^{k}(\ln(2\pi\sigma^2_{p_{ij}}) + (\frac{p_{tj}-\mu_{p_{ij}}}{\sigma_{p_{ij}}})^2)$$
$$-\frac{1}{2}\sum_{j=1}^{k}\sum_{t=1}^{n}(\ln(2\pi\sigma^2_{s_{ji}}) + (\frac{s_{ji}-\mu_{s_{ji}}}{\sigma_{s_{ji}}})^2) +$$
$$\sum_{t=1}^{m}((\alpha_t-1)\ln(\varepsilon_t^{-1}) - (\beta_t\varepsilon_t)^{-1} - \alpha_t\ln(\beta_t) - \ln\Gamma(\alpha_t))$$。

[0167] 具体RJMCMCNMF实现步骤如下：

[0168] 1)、设定初始值k0；

[0169] 2)、计算收敛的初始S0,P0;

[0170] 3)、通过公式抽样P,S,E;

[0171] 4)、使用生长消亡方法,u～U(0,1),如果u≤b_k,则进行生长步骤,如果 b_k＜u≤b_k+d_k,则进行消亡步骤;

[0172] 5)、重复以上步骤至设定的迭代步骤(step＝10000,其中前1000次为燃烧期。

[0173] 生长步骤包括以下步骤:

[0174] a)、k＝k0+1;

[0175] b)、执行2),收敛则继续以下步骤;

[0176] c)、从C_k中抽取q_k,即执行3);

[0177] d)、计算α(k0,k);

[0178] e)、计算特征之间的相似性;

[0179] f)、u～U(0,1);

[0180] g)、如果u≤α(k0,k),且两两相似性均小于0.3,则接受k,否则不接受消亡步骤包括以下步骤:

[0181] a)、k＝k0-1;

[0182] b)、执行2),收敛则继续以下步骤;

[0183] c)、从C_k中抽取q_k,即执行3);

[0184] d)、计算α(k0,k);

[0185] e)、计算特征之间的相似性;

[0186] f)、u～U(0,1);

[0187] g)、如果u≤α(k0,k),且两两相似性均小于0.3,则接受k,否则不接受

[0188] 特征相似性计算方法:

[0189] $$sim(A,B)=\frac{\sum_{k=1}^{k}A_kB_K}{\sqrt{\sum_{K=1}^{K}(A_K)^2}\sqrt{\sum_{K=1}^{K}(B_K)^2}}$$。

[0190] 轮廓系数计算:将所有k对应的每个特征作为一个类,通过轮廓系数公式进行这k类数据的评估分析,获取轮廓指数;

[0191] 运行结果可视化展示方式:将基础矩阵进行归一化后,按照百分比把每个特征属性的柱状图刻画出来,采用不同的颜色进行区分,如附图3所示:

[0192] 步骤五:特征图谱与基因关联获取方法:这部分主要随着特征图谱的分解出来,根据数据集中注释的基因信息实现基因与特征图谱的关联,其实现的途径是确立每个基因非沉默突变对应到某个样本,该样本对各个signature 的贡献可算,选定贡献度大于20%作为阈值,定义样本出现signature特征的条件,从而统计检验(Fisher检验)基因与signature出现的可能性。结合基因的功能特征,研究基因在癌症发生发展中的作用,从而间接研究特征图谱在癌症中的作用,甚至可以知道个体用药。比如COSMIC数据库的signature 3这个图谱特征,跟基因BRCA1/2关联,其与铂类化疗敏感相关。基于分解的特征图谱,计算每个非沉默突变对signature的累积贡献概率,从而寻找一些经典的癌基因热点突变与突变特征图谱潜在的因果关系,有助于研究癌症发生发展的机制与变化的过程。同

时可以研究该图谱特征关联密切的热点突变富集在通路(pathway)的情况,有助于寻找潜在的治疗靶点与方法。

[0193] 步骤六:特征图谱亚型聚类与预后关联获取方法:基于系数矩阵信息,获取每个样本对signature的贡献度,基于这些贡献度,可以使用无监督聚类方法对样本进行分类,得到不同的亚型,然后不同亚型与临床信息关联,做预后生存分析,可以找到与预后相关的图谱特征或者与之关联的预后因子 (内因或者外因)。

[0194] 尽管已经示出和描述了本发明的实施例,对于本领域的普通技术人员而言,可以理解在不脱离本发明的原理和精神的情况下可以对这些实施例进行多种变化、修改、替换和变型,本发明的范围由所附权利要求及其等同物限定。

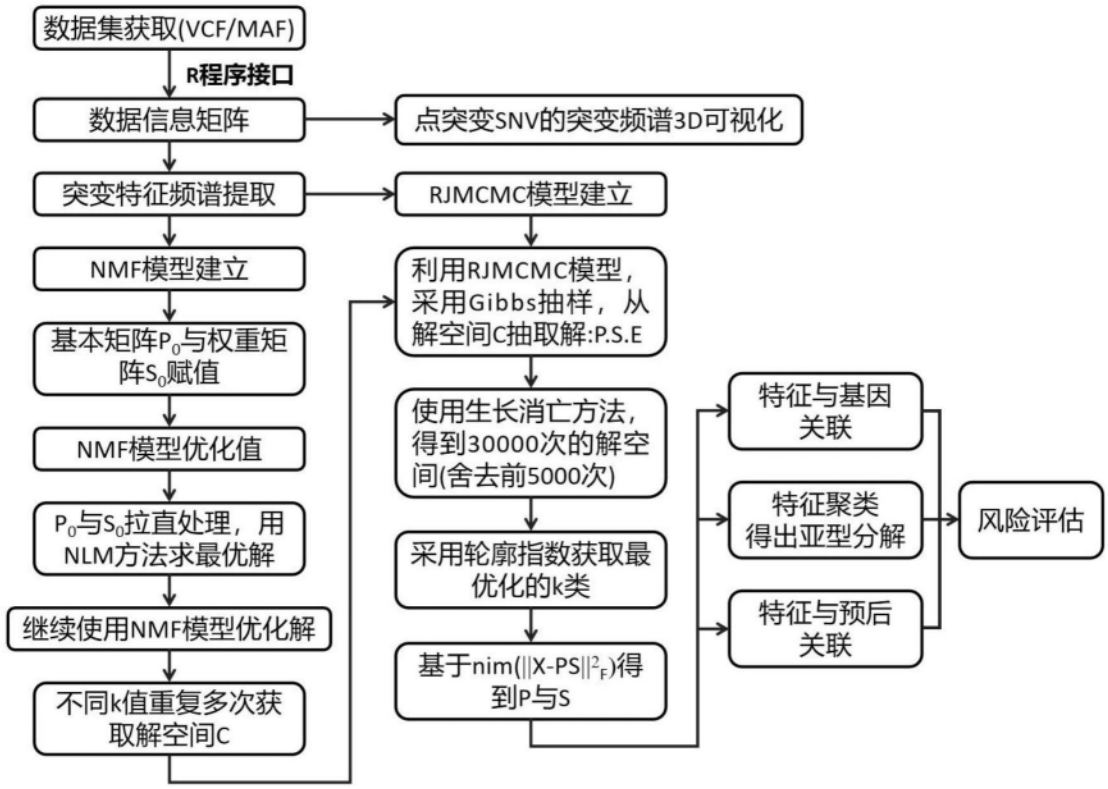

数据集获取(VCF/MAF)
R程序接口
数据信息矩阵
点突变SNV的突变频谱3D可视化
突变特征频谱提取
RJMCMC模型建立
NMF模型建立
基本矩阵P_0与权重矩阵S_0赋值
NMF模型优化值
P_0与S_0拉直处理，用NLM方法求最优解
继续使用NMF模型优化解
不同k值重复多次获取解空间C
利用RJMCMC模型，采用Gibbs抽样，从解空间C抽取解:P.S.E
使用生长消亡方法，得到30000次的解空间(舍去前5000次)
采用轮廓指数获取最优化的k类
基于nim($||X-PS||^2_F$)得到P与S
特征与基因关联
特征聚类得出亚型分解
特征与预后关联
风险评估

图1

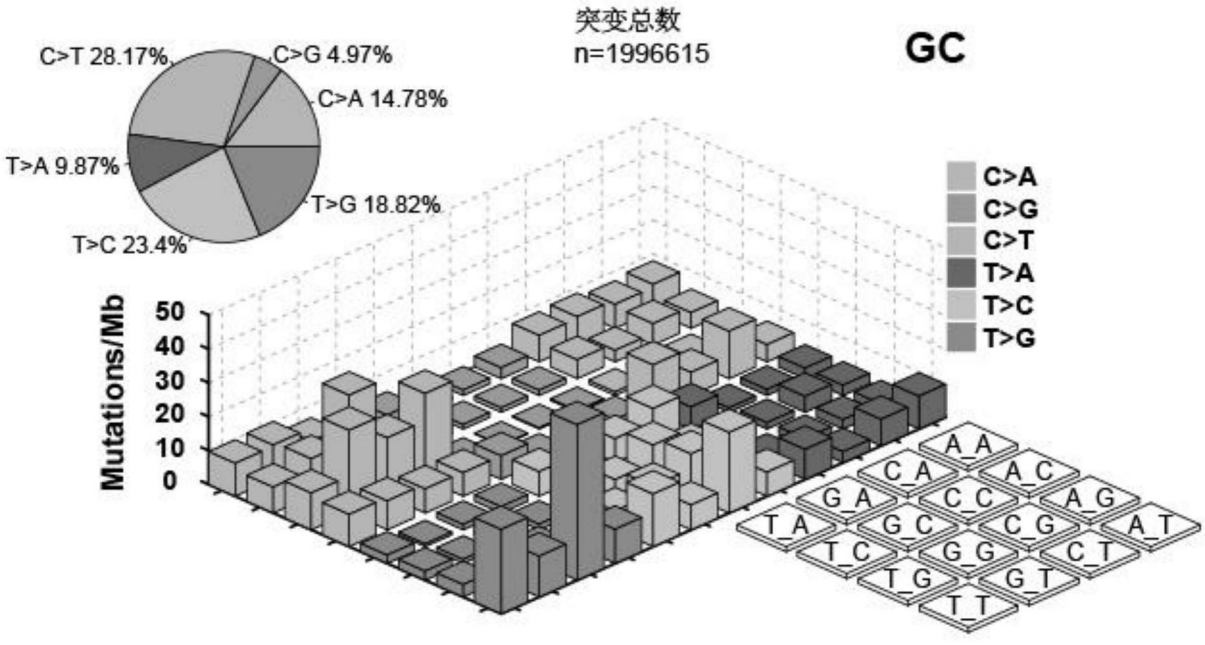

突变总数
n=1996615
GC
C>T 28.17%
C>G 4.97%
C>A 14.78%
T>A 9.87%
T>G 18.82%
T>C 23.4%
Mutations/Mb
50
40
30
20
10
0
C>A
C>G
C>T
T>A
T>C
T>G
A_A
C_A
A_C
G_A
C_C
A_G
T_A
G_C
C_G
A_T
T_C
G_G
C_T
T_G
G_T
T_T

图2

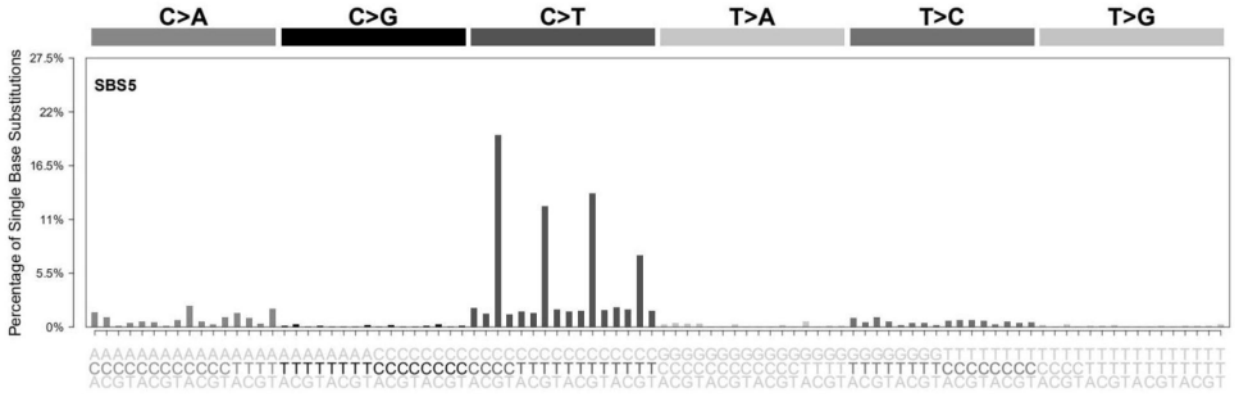

C>A
C>G
C>T
T>A
T>C
T>G
SBS5
Percentage of Single Base Substitutions
27.5%
22%
16.5%
11%
5.5%
0%

图3