



(21) 申请号 202410924304.0

G16B 40/20 (2019.01)

(22) 申请日 2024.07.11

G06F 18/2135 (2023.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 118471326 A

G06F 18/23 (2023.01)

(43) 申请公布日 2024.08.09

(73) 专利权人 浙江大学

地址 310058 浙江省杭州市西湖区余杭塘路866号

(56) 对比文件

CN 107629124 A, 2018.01.26

CN 109872772 A, 2019.06.11

(72) 发明人 瞿海斌 张胜 陈杭 万宇翔

审查员 靳超

(74) 专利代理机构 广州三环专利商标代理有限公司

44202

专利代理师 许羽冬

(51) Int. Cl.

G16B 20/00 (2019.01)

G16B 15/30 (2019.01)

权利要求书3页 说明书14页 附图6页

(54) 发明名称

一种细胞培养过程数据分析方法、系统、设备及介质

(57) 摘要

本发明涉及细胞培养技术领域,公开了一种细胞培养过程数据分析方法、系统、设备及介质,包括:获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,所述多个批次划分为第一批次集合和第二批次集合;任一批次的细胞培养过程划分为若干阶段;根据所述第一批次集合中第一批次和所述第二批次集合中第二批次各自的各阶段的在线过程数据和离线检验数据,分别确定第一批次特征数据和第二批次特征数据;根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,在所得的目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。本发明提高了商业生产规模的细胞培养过程的分析效果,降低了所需成本。

获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,所述多个批次划分为第一批次集合和第二批次集合;所述第一批次集合的实际蛋白表达量均值高于所述第二批次集合的实际蛋白表达量均值;任一批次的细胞培养过程划分为若干阶段

S100

根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,根据所述第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据

S200

根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,得到目标主成分方向,在所述目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量

S300

1. 一种细胞培养过程数据分析方法,其特征在于,包括:

获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,所述多个批次划分为第一批次集合和第二批次集合;所述第一批次集合的实际蛋白表达量均值高于所述第二批次集合的实际蛋白表达量均值;任一批次的细胞培养过程划分为若干阶段;所述在线过程数据为细胞培养过程中的培养环境数据,且是通过内设于培养罐的检测设备测量得到的实时相关数据;所述离线检验数据为细胞培养过程中的细胞状态检验数据,且是通过外部的检验设备对从培养罐中取的药液进行检验得到的非实时相关数据;

根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,根据所述第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据;

根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,得到目标主成分方向,在所述目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量;

其中,所述根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,得到目标主成分方向,包括:

获取预设的多个候选对比度参数;

确定所述第一批次特征数据和所述第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度,以及所述第一批次特征数据在任一候选对比度参数对应的候选主成分方向上的特征保留度;

基于所述分离度以及所述特征保留度在所述多个候选对比度参数中确定目标对比度参数;

将所述目标对比度参数对应的候选主成分方向作为所述目标主成分方向。

2. 根据权利要求1所述的方法,其特征在于,所述确定所述第一批次特征数据和所述第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度,包括:

将所述第一批次特征数据在所述候选主成分方向上投影,得到第一投影特征数据;

将所述第二批次特征数据在所述候选主成分方向上投影,得到第二投影特征数据;

基于所述第一投影特征数据与所述第二投影特征数据间的直方图交叉核的倒数确定所述分离度。

3. 根据权利要求1所述的方法,其特征在于,通过以下方式确定所述特征保留度:

将所述第一批次特征数据在所述候选主成分方向上投影,得到第一投影特征数据;

计算所述第一投影特征数据的方差,作为所述特征保留度。

4. 根据权利要求3所述的方法,其特征在于,所述基于所述分离度以及所述特征保留度在所述多个候选对比度参数中确定目标对比度参数,包括:

根据所述多个候选对比度参数各自对应的特征保留度与所述第一批次特征数据的方差的比较结果,对所述多个候选对比度参数进行筛选,得到筛选后对比度参数;

在所述筛选后对比度参数对应的分离度中确定最大分离度;

将所述最大分离度对应的筛选后对比度参数作为所述目标对比度参数。

5. 根据权利要求1所述的方法,其特征在于,通过以下方式确定所述任一候选对比度参数对应的候选主成分方向,包括:

$$\mathbf{v}^* = \arg \max_{\mathbf{v}} [\lambda_E(\mathbf{v}) - \alpha \lambda_R(\mathbf{v})]$$

其中, \mathbf{v}^* 为候选主成分方向; α 为候选对比度参数; $\lambda_E(\mathbf{v})$ 的形式为 $\lambda_E(\mathbf{v}) = \mathbf{v}^T C_E \mathbf{v}$, $\lambda_R(\mathbf{v})$ 的形式为 $\lambda_R(\mathbf{v}) = \mathbf{v}^T C_R \mathbf{v}$, 其中, C_E 和 C_R 分别为融合特征数据的第一协方差矩阵和第二批次特征数据的第二协方差矩阵, \mathbf{v} 为矩阵 $C_E - \alpha C_R$ 的第一个特征向量;

所述融合特征数据是根据所述第一批次特征数据和所述第二批次特征数据进行融合得到的。

6. 根据权利要求1至4任一项所述的方法, 其特征在于, 通过以下方式对所述任一批次的细胞培养过程进行划分:

以所述任一批次的每个样本为中心样本分别构建移动窗, 计算所述移动窗中其他样本与所述中心样本之间的距离;

筛选距离最近的若干个样本作为近邻样本, 计算所述近邻样本与所述中心样本间的距离均值, 作为近邻距离;

当所述近邻距离超过分段阈值时, 将所述中心样本对应的时间点作为阶段划分点, 以对所述任一批次的细胞培养过程进行划分。

7. 根据权利要求1至4任一项所述的方法, 其特征在于, 通过以下方式划分所述第一批次集合和所述第二批次集合:

根据所述多个批次各自的生产方法对所述多个批次进行聚类, 得到多个批次簇;

确定每个批次簇对应的实际蛋白表达量均值;

将实际蛋白表达量均值超过预设蛋白表达量阈值的批次簇确定为所述第一批次集合;

将实际蛋白表达量均值未超过预设蛋白表达量阈值的批次簇确定为所述第二批次集合。

8. 根据权利要求1至4任一项所述的方法, 其特征在于, 所述第一批次特征数据的确定方式与所述第二批次特征数据的确定方式相同; 所述根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定所述第一批次特征数据, 包括:

基于每个第一批次中每个阶段的在线过程数据进行特征提取, 得到每个阶段的在线数据特征;

基于每个第一批次中每个阶段的离线检验数据确定每个阶段的离线数据特征;

对所述在线数据特征以及所述离线数据特征进行融合, 得到每个第一批次的融合特征矩阵;

对每个第一批次的融合特征矩阵进行合并, 得到所述第一批次特征数据。

9. 一种细胞培养过程数据分析系统, 其特征在于, 包括:

数据采集模块, 用于获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据; 其中, 所述多个批次划分为第一批次集合和第二批次集合; 所述第一批次集合的实际蛋白表达量均值高于所述第二批次集合的实际蛋白表达量均值; 任一批次的细胞培养过程划分为若干阶段; 所述在线过程数据为细胞培养过程中的培养环境数据, 且是通过内设于培养罐的检测设备测量得到的实时相关数据; 所述离线检验数据为细胞培养过程中的细胞状态检验数据, 且是通过外部的检验设备对从培养罐中取的药液进行检验得到的非实

时相关数据;

特征数据提取模块,用于根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,根据所述第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据;

变量分析模块,用于根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,得到目标主成分方向,在所述目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量;

所述变量分析模块还用于:获取预设的多个候选对比度参数;确定所述第一批次特征数据和所述第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度,以及所述第一批次特征数据在任一候选对比度参数对应的候选主成分方向上的特征保留度;基于所述分离度以及所述特征保留度在所述多个候选对比度参数中确定目标对比度参数;将所述目标对比度参数对应的候选主成分方向作为所述目标主成分方向。

10. 一种计算机设备,其特征在于,包括:存储器和处理器,所述存储器和所述处理器之间互相通信连接,所述存储器中存储有计算机指令,所述处理器通过执行所述计算机指令,从而执行权利要求1至8中任一项所述的细胞培养过程数据分析方法。

一种细胞培养过程数据分析方法、系统、设备及介质

技术领域

[0001] 本发明涉及细胞培养技术领域,尤其涉及一种细胞培养过程数据分析方法、系统、设备及介质。

背景技术

[0002] 生物药是当前医药市场中最具有发展前景的方向之一,约占据1/3的市场份额。细胞培养是生物药生产过程中至关重要的一环,常见的抗体、细胞因子或其他蛋白类的生物药一般都需要通过细胞培养来生产。细胞培养过程终点的蛋白表达量是一个关键性能指标,若蛋白表达量下降,则会直接导致药品产量降低,从而使得企业利润受损。因此如何调控细胞培养过程使蛋白表达量维持在高水平,是整个生物制药行业都面临的问题。

[0003] 然而细胞培养过程极其复杂,许多因素都会影响蛋白表达量,如培养温度、pH、溶氧、CO₂分压、培养基配方和补料方式等。目前一般采用实验设计等方法在小试或者中试规模通过实验的方式来考察各个因素对细胞培养过程的影响,而将同样方法用于商业生产规模的细胞培养过程时,由于培养规模巨大,导致实验过程非常耗费时间、人力和物料,所耗费的实验成本是企业难以承受的。因此,对于商业生产规模的细胞培养过程,如何以较低成本辨析蛋白表达量下降的原因并有的放矢地优化细胞培养过程,是困扰所有生物药企业的关键问题。

发明内容

[0004] 本发明提供了一种细胞培养过程数据分析方法、系统、设备及介质,以提高对商业生产规模的细胞培养过程的分析效果,降低所需成本。

[0005] 为了实现上述技术效果,本发明实施例提供了一种细胞培养过程数据分析方法,该方法包括:

[0006] 获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,所述多个批次划分为第一批次集合和第二批次集合;所述第一批次集合的实际蛋白表达量均值高于所述第二批次集合的实际蛋白表达量均值;任一批次的细胞培养过程划分为若干阶段;

[0007] 根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,根据所述第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据;

[0008] 根据所述第一批次特征数据和所述第二批次特征数据进行对比主成分分析,得到目标主成分方向,在所述目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。

[0009] 本发明实施例还提供了一种细胞培养过程数据分析系统,该系统包括:

[0010] 数据采集模块,用于获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,所述多个批次划分为第一批次集合和第二批次集合;所述第一批次集

合的实际蛋白表达量均值高于所述第二批集合的实际蛋白表达量均值;任一批次的细胞培养过程划分为若干阶段;

[0011] 特征数据提取模块,用于根据所述第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批特征数据,根据所述第二批集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批特征数据;

[0012] 变量分析模块,用于根据所述第一批次特征数据和所述第二批特征数据进行对比主成分分析,得到目标主成分方向,在所述目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。

[0013] 本发明实施例还提供了一种终端设备,包括存储器和处理器,所述存储器和所述处理器之间互相通信连接,所述存储器存储有计算机指令,所述处理器通过执行所述计算机指令,从而执行如上述的一种细胞培养过程数据分析方法。

[0014] 本发明实施例还提供了一种存储介质,所述存储介质上存储有计算机程序,所述计算机程序被计算机调用并执行,实现如上述的一种细胞培养过程数据分析方法。

[0015] 本发明通过采用对比主成分分析对不同批次集合间的差异进行合理调整,并通过确定目标主成分方向从而确定引起蛋白表达量下降的主要影响变量,提升了分析结果的准确性以及分析过程的效率。本发明还通过获取细胞培养过程中的在线过程数据和离线检验数据,并对在线过程数据和离线检验数据均进行分析,扩大了引起蛋白表达量下降的主要影响变量的分析范围,与相关技术相比能够搜索到更多的结果,从而提高了分析结果的准确性。

附图说明

[0016] 图1为本发明实施例提供的一种细胞培养过程数据分析方法的步骤示意图;

[0017] 图2为本发明实施例中目标主成分方向的确定方法的步骤示意图;

[0018] 图3为本发明实施例中目标对比参数的确定方法的步骤示意图;

[0019] 图4为本发明应用场景示例中任一批次的近邻距离和所有批次的阶段长度的分布示意图;

[0020] 图5为本发明应用场景示例中所有批次的蛋白表达量的示意图;

[0021] 图6为本发明应用场景示例中目标主成分方向中变量的载荷值示意图;

[0022] 图7为本发明应用场景示例中所有批次的v9和v11变化趋势示意图;

[0023] 图8为本发明应用场景示例中所有批次第五阶段的v11变化趋势示意图;

[0024] 图9为本发明实施例提供的一种细胞培养过程数据分析系统的模块示意图;

[0025] 其中,说明书附图的附图标记如下:

[0026] 1、数据采集模块,2、特征数据提取模块,3、变量分析模块。

具体实施方式

[0027] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0028] 生物药是当前医药市场中最具有发展前景的方向之一,约占据1/3的市场份额。细胞培养是生物药生产过程中至关重要的一环,常见的抗体、细胞因子或其他蛋白类的生物药一般都需要通过细胞培养来生产。细胞培养过程终点的蛋白表达量是一个关键性能指标,若蛋白表达量下降,则会直接导致药品产量降低,从而使得企业利润受损。因此如何调控细胞培养过程使蛋白表达量维持在高水平,是整个生物制药行业都面临的问题。

[0029] 然而细胞培养过程极其复杂,许多因素都会影响蛋白表达量,如培养温度、pH、溶氧、CO₂分压、培养基配方和补料方式等。目前一般采用实验设计等方法在小试或者中试规模通过实验的方式来考察各个因素对细胞培养过程的影响,非常耗费时间、人力和物料,且在小试/中试规模得到的实验结论可能并不适用于商业生产规模,导致了研发与生产之间的脱节。然而对于商业生产规模的细胞培养过程而言,实施多次实验来进行过程优化非常困难,由于其规模巨大,所耗费的实验成本是企业难以承受的。因此,如何以较低成本辨析蛋白表达量下降的原因并有的放矢地优化细胞培养过程,是困扰所有生物药企业的关键问题。

[0030] 目前许多生物药企业部署了生产信息系统,如制造执行系统(Manufacturing Execution System,MES)、数据采集与监视控制系统(Supervisory Control And Data Acquisition,SCADA)系统、实验室信息管理系统(Laboratory Information Management System,LIMS)等。这些信息系统可以采集海量的细胞培养过程数据,而过程数据反映了生产状态,通过充分的数据分析挖掘可以识别过程模式,找出过程问题原因。

[0031] 细胞培养过程数据可以分为离线检验数据(如细胞密度、细胞活率和葡萄糖浓度等)和在线过程数据(如温度、pH、溶氧和通气量等),目前相关的数据分析方法主要针对离线检验数据展开,缺乏对在线过程参数,因此难以确认蛋白表达量下降的原因。此外,细胞培养过程是一个多阶段的批次生产过程,每个批次的长度(时间)并不相同,存在着多阶段和批次长度不齐的问题,对后续数据分析造成一定的困难。

[0032] 因此,需要致力于开发一种商业生产规模的细胞培养过程数据分析方法,用于辨析蛋白表达量下降的原因。

[0033] 基于上述问题,本发明提供了一种细胞培养过程数据分析方法,首先对细胞培养过程中产生的在线过程数据和离线检验数据进行采集,并对不同批次分别进行阶段划分。对于划分阶段后的各批次,根据生产方法对批次进行聚类,并根据批次簇各自的实际蛋白表达量均值确定第一批次集合和第二批次集合;其中,第一批次集合的实际蛋白表达量均值大于第二批次集合(可以理解的是,第一批次集合可以对应高产批次集合,相应地,第二批次集合可以对应低产批次集合)。在此基础上,通过函数型数据分析方法对任一阶段的在线特征数据进行提取,并根据离线检验数据确定任一阶段的离线特征数据,将在线特征数据和离线检验数据融合得到该阶段的融合特征矩阵;其中,在线特征数据分别与在线过程数据中的变量对应,离线特征数据分别与离线检验数据中的变量对应。根据第一批次集合和第二批次集合的融合特征矩阵,并通过改进的对比主成分分析方法对引起蛋白表达量下降的主要影响变量进行分析,提高了对商业生产规模的细胞培养过程的分析效果,降低了所需成本。

[0034] 在改进的对比主成分分析方法中,首选获取预设多个候选对比度参数,并根据第一批次集合和第二批次集合各自对应的在线过程数据和离线检验数据确定在任一候选对

比度参数对应的主成分方向上的分离度;然后根据第一批次集合对应的在线过程数据和离线检验数据确定在任一候选对比度参数对应的主成分方向上的投影特征数据并基于分离度以及投影特征数据确定目标主成分方向,以在目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。

[0035] 请参照图1所示,图1为本发明实施例提供的一种细胞培养过程数据分析方法的步骤示意图,该方法可以包括以下步骤:

[0036] S100、获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据;其中,多个批次划分为第一批次集合和第二批次集合;第一批次集合的实际蛋白表达量均值高于第二批次集合的实际蛋白表达量均值;任一批次的细胞培养过程划分为若干阶段。

[0037] S200、根据第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,根据第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据。

[0038] S300、根据第一批次特征数据和第二批次特征数据进行对比主成分分析,得到目标主成分方向,在目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。

[0039] 其中,在线过程数据为在细胞培养过程中以一定时间间隔获取的实时相关数据,离线检验数据为在细胞培养过程中在一定时间间隔后通过检验获取的检验相关数据。其中,在线过程数据的时间间隔小于离线检验数据的时间间隔。

[0040] 本实施例中,在线过程数据可以为细胞培养过程中的培养环境数据,包括pH值、溶氧量、温度、底层空气流量、表层空气流量、底层氧气流量和培养液体积等中的至少一个,在线过程数据可由用于培养细胞的培养罐通过内设的检测设备测量得到。离线检验数据可以为细胞培养过程中的细胞状态检验数据,包括活细胞密度、细胞活率、细胞平均直径、细胞聚集率、pH值、葡萄糖浓度和乳酸浓度等中的至少一个,离线检验数据可通过一定时间间隔后通过从培养罐中取一定量的药液,并通过外部的检验设备通过检验得到,可理解为离线检验数据为无法通过培养罐内设的检测设备获得的数据。在其他不同实施例中,在线过程数据和离线检验数据的类型可根据培养的细胞类型和采用的生产方法等进行确定。

[0041] 本实施例中,第一批次集合也可称为高产批次集合,第二批次集合也可称为低产批次集合,其中,第一批次集合的实际蛋白表达量均值高于第二批次集合的实际蛋白表达量均值,可认为第一批次集合中包含的第一批次的产量均高于第二批次集合中包含的第二批次的产量。

[0042] 具体地,对任一批次所划分的阶段数量均相同。其中,任一批次的细胞培养过程的时间长度可以相同也可以不同,任一批次的不同阶段的时间长度可以相同也可以不同。示例性地,阶段划分的标准可根据人为设定或细胞培养过程中的生产操作进行设定,上述生产操作包括但不限于:向包含任一批次的培养罐中增加培养基和调节包含任一批次的培养罐中的pH值等。

[0043] 具体地,在线特征数据中各特征分别与在线过程数据中的各变量对应,离线特征数据中各特征分别与离线检验数据中的各变量对应,可理解为在线特征数据表征了在线过程数据中的各变量在任一阶段的变化特征,离线特征数据表征了离线检验数据中的各变量

在任一阶段的变化特征。

[0044] 具体地,目标主成分中包括在线过程数据中的变量和离线检验数据中的变量,即目标主成分为在线过程数据中的变量和离线检验数据中的变量的线性组合,其中,线性组合的系数即目标主成分方向,表示了任一变量对目标主成分方向的影响程度。

[0045] 具体地,通过计算在目标主成分方向上任一变量对应的载荷值,确定主要影响变量。其中,载荷值表示了任一变量在目标主成分方向上的贡献程度,本实施例中即导致蛋白表达量下降的贡献程度,将载荷值最高的变量作为引起蛋白表达量下降的主要影响变量。在一些实施例中,确定引起蛋白表达量下降的主要影响变量后,可通过对主要影响变量在所有批次细胞培养过程中的变化曲线进行分析,从而判断主要影响变量造成第一批次集合和第二批次集合之间差异的原因。

[0046] 上述细胞培养过程数据分析方法中,通过获取细胞培养过程中的在线过程数据和离线检验数据,并对在线过程数据和离线检验数据均进行分析,扩大了引起蛋白表达量下降的主要影响变量的分析范围,与相关技术相比能够搜索到更多的结果,从而提高了分析结果的准确性。此外,还通过对细胞培养过程进行阶段划分,从而将所有批次的的数据维度调整至一致,在后续分析过程中可根据不同阶段中在线过程数据和离线检验数据中包含的任一变量的具体变化进行进一步分析。

[0047] 在一些实施例中,请参照图2所示,图2为本发明实施例中目标主成分方向的确定方法的步骤示意图,如图所示,根据第一批次特征数据和第二批次特征数据进行对比主成分分析,得到目标主成分方向,包括以下步骤:

[0048] S320、获取预设的多个候选对比度参数。

[0049] S340、确定第一批次特征数据和第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度,以及第一批次特征数据在任一候选对比度参数对应的候选主成分方向上的特征保留度。

[0050] S360、基于分离度以及特征保留度在多个候选对比度参数中确定目标对比度参数。

[0051] S380、将目标对比度参数对应的候选主成分方向作为目标主成分方向。

[0052] 具体地,候选对比度参数为非负数,其范围上限可根据任一批次集合间的差异进行选择,例如:当第一批次特征数据和第二批次特征数据之间的差异较小时,可通过适当增加候选对比度参数的大小从而充分放大第一批次特征数据和第二批次特征数据之间的差异,这时候候选对比度参数的范围上限可达到 10^4 。在一些实施例中,候选对比度参数的数量可根据实际情况进行选择,例如:当用于执行对比主成分分析的设备的运算能力较低时,此时若获取的候选对比度参数过多,会造成运算量增加,运算时间延长,导致分析效率下降;而若获取的候选对比度参数过少,会造成遗漏所需的目标对比度参数,导致最终分析结果的准确性下降。

[0053] 具体地,分离度用于表征第一批次特征数据和第二批次特征数据之间的分离程度,其中,分离度越大,表示第一批次特征数据和第二批次特征数据之间在当前候选对比度参数对应的候选主成分方向上的差异越大,该候选主成分方向能够充分表示第一批次集合和第二批次集合之间产生差异的原因。

[0054] 具体地,特征保留度用于表征第一批次特征数据分离后的特征信息保留程度,其

中,特征保留度越大,表示第一批次特征数据在当前候选对比度参数对应的候选主成分方向上保留的特征信息越多,该候选主成分方向能够充分表示第一批次集合中的特征数据。

[0055] 在一些实施例中,可分别对分离度和特征保留度设置不同的数值条件,例如:选取一定范围内最大值和设定一定阈值并选取超过或不超阈值的值。分离度和特征保留度的条件可根据第一批次集合和第二批次集合之间的差异大小决定:当第一批次集合和第二批次集合之间的差异较大时,此时不需要通过候选对比度参数将第一批次集合和第二批次集合进行分离,因此可选择较小的候选对比度参数,从而尽可能保留第一批次集合的特征数据;当第一批次集合和第二批次集合之间的差异较小时,此时需要通过候选对比度参数将第一批次集合和第二批次集合进行分离,因此可选择较大的候选对比度参数,从而尽可能地放大第一批次集合的特征数据和第二批次集合之间的差异,为后续分析提供数据基础。

[0056] 上述细胞培养过程数据分析方法中,通过设置分离度和特征保留度作为对比度参数选择的条件,其中,通过计算分离度保证不同批次特征数据间能够充分分离,并通过特征保留度减少了第一批次特征数据的损失,使得利用最终选出的对比度参数确定的目标主成分方向能够合理且充分地反映不同批次集合间的差异,从而提升了最终分析结果的准确性。

[0057] 在一些实施例中,确定第一批次特征数据和第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度,包括以下步骤:

[0058] S343、将第一批次特征数据在候选主成分方向上投影,得到第一投影特征数据。

[0059] S346、将第二批次特征数据在候选主成分方向上投影,得到第二投影特征数据。

[0060] S349、基于第一投影特征数据与第二投影特征数据间的直方图交叉核的倒数确定分离度。

[0061] 具体地,可知候选主成分为在线过程数据和离线检验数据中各变量的线性组合,该线性组合的系数即候选主成分方向,将第一批次特征数据投影在候选主成分方向上得到的第一投影特征数据表征了第一批次集合在候选主成分方向上的特征,将第二批次特征数据投影在候选主成分方向上得到的第二投影特征数据表征了第二批次集合在候选主成分方向上的特征。通过第一投影特征数据和第二投影特征数据可得到经过候选对比度参数处理后的第一批次特征数据和第二批次特征数据之间的关系。

[0062] 在一些实施例中,第一投影特征数据的形式如下式:

$$[0063] \quad K'_i = \mathbf{v}^* \cdot K$$

[0064] 其中, K'_i 为第一投影特征数据的矩阵形式; \mathbf{v}^* 为候选主成分方向; K 为第一批次特征数据的矩阵形式。相应地,第二投影特征数据的形式如下式:

$$[0065] \quad R'_i = \mathbf{v}^* \cdot R$$

[0066] 其中, R'_i 为第二投影特征数据的矩阵形式; \mathbf{v}^* 为候选主成分方向; R 为第二批次特征数据的矩阵形式。

[0067] 具体地,第一批次特征数据和第二批次特征数据之间的关系包括分离度,分离度可通过第一投影特征数据和第二投影特征数据间的直方图交叉核计算得到。

[0068] 其中,第一投影特征数据和第二投影特征数据的行表示任一批次,列表示经过候

选对比度参数处理后的在线过程数据和离线检验数据中各变量的特征。将第一投影特征数据转换为多个对应于任一批次的投影特征直方图,其中横轴表示各变量,纵轴表示各变量对应的特征值大小,将所有投影特征直方图进行堆叠,得到对应于第一投影特征数据的第一投影特征直方图;通过相同方法可得到对应于第二投影特征数据的第二投影特征直方图。对第一投影特征直方图和第二投影特征直方图计算直方图交叉核,直方图交叉核表征了第一投影特征直方图和第二投影特征直方图的相似程度,取直方图交叉核的倒数作为分离度,此时分离度表征了第一投影特征直方图和第二投影特征直方图的差差异程度。

[0069] 上述细胞培养过程数据分析方法中,通过将第一批次特征数据和第二批次特征数据投影在候选主成分方向上,得到了在分析过程中对应于不同候选主成分方向的任一批次特征数据,并在此基础上通过所得不同批次特征数据对分离度进行计算,使得能够确定不同批次特征数据间是否足够分离,从而通过选取合适的候选对比度参数适当放大不同批次集合间的差异,提升引起蛋白表达量下降的主要影响变量的分析效果。

[0070] 在一些实施例中,通过以下方式确定特征保留度,包括以下步骤:

[0071] S352、将第一批次特征数据在候选主成分方向上投影,得到第一投影特征数据。

[0072] S355、计算第一投影特征数据的方差,作为特征保留度。

[0073] 其中,方差表征了一组数据的离散程度,通过计算第一投影特征数据的方差可对第一批次集合在候选主成分方向上的特征之间的离散程度进行表示。当第一投影特征数据的方差较大时,表示第一批次集合在候选主成分方向上的特征之间的离散程度较大,此时任一特征之间仍存在一定差异,对第一投影特征数据进行分析时可得到第一批次集合中任一批次的特征,即保留了第一批次集合在经过候选对比度参数处理前的大部分特征。当第一投影特征数据的方差较小时,表示第一批次集合在候选主成分方向上的特征之间的离散程度较小,此时任一特征之间差异已不再明显,对第一投影特征数据进行分析时无法得到第一批次集合中任一批次的特征,即损失了第一批次集合在经过候选对比度参数处理前的大部分特征。

[0074] 上述细胞培养过程数据分析方法中,通过计算第一投影特征数据的方差,确定第一投影特征数据之间的离散程度,使得能够确定第一批次特征数据在该候选主成分方向上的特征保留度,从而通过选取合适的候选对比度参数适当保留第一批次集合的特征,提升引起蛋白表达量下降的主要影响变量的分析效果。

[0075] 在一些实施例中,请参照图3所示,图3为本发明实施例中目标对比参数的确定方法的步骤示意图,如图所示,基于分离度以及特征保留度在多个候选对比度参数中确定目标对比度参数,包括以下步骤:

[0076] S363、根据多个候选对比度参数各自对应的特征保留度与第一批次特征数据的方差的比较结果,对多个候选对比度参数进行筛选,得到筛选后对比度参数。

[0077] S366、在筛选后对比度参数对应的分离度中确定最大分离度。

[0078] S369、将最大分离度对应的筛选后对比度参数作为目标对比度参数。

[0079] 具体地,根据第一批次特征数据的方差设定特征保留度的比较基准,其中,比较基准可为第一批次特征数据的方差的任意倍数,包括但不限于:四分之三、二分之一和三分之二等。在一些实施例中,当分析的重点更偏向于第一批次集合在经过候选对比度参数处理前的特征时,特征保留度的比较基准应当选择更高的倍数;当分析的重点更偏向于第一批

次特征数据和第二批次特征数据之间的差异时,特征保留度的比较基准可适当降低。

[0080] 其中,目标对比度参数的形式如下式:

$$[0081] \quad \alpha = \arg \max D(\alpha_i), s.t. V(\alpha_i) > \frac{V(\alpha_0)}{2}$$

[0082] 其中, α 为目标对比度参数; α_i 为第 i 个候选对比度参数; $D(\alpha_i)$ 为分离度; $V(\alpha_i)$ 为经过第 i 个候选对比度参数处理后的第一批次特征数据的方差; $V(\alpha_0)$ 为经过第 i 个候选对比度参数处理前的第一批次特征数据的方差。

[0083] 上述细胞培养过程数据分析方法中,通过对比度参数将不同批次集合之间充分分开,从而放大不同批次集合间的差异,提升了引起蛋白表达量下降的主要影响变量的分析效果。在确定目标对比度参数时,需要选择符合特征保留度条件,且能使不同批次集合之间分离度最大的候选对比度参数。

[0084] 在一些实施例中,根据多个候选对比度参数各自对应的特征保留度与第一批次特征数据的方差的比较结果,对多个候选对比度参数进行筛选,得到筛选后对比度参数,包括以下步骤:

[0085] S364、根据第一批次特征数据的方差的一半确定比较基准。

[0086] S365、将大于比较基准的特征保留度所对应的候选对比度参数,作为筛选后对比度参数。

[0087] 在一些实施例中,通过以下方式确定任一候选对比度参数对应的候选主成分方向,包括:

$$[0088] \quad \mathbf{v}^* = \arg \max_{\mathbf{v}} [\lambda_E(\mathbf{v}) - \alpha \lambda_R(\mathbf{v})]$$

[0089] 其中, \mathbf{v}^* 为候选主成分方向; α 为候选对比度参数; $\lambda_E(\mathbf{v})$ 的形式为 $\lambda_E(\mathbf{v}) = \mathbf{v}^T C_E \mathbf{v}$, $\lambda_R(\mathbf{v})$ 的形式为 $\lambda_R(\mathbf{v}) = \mathbf{v}^T C_R \mathbf{v}$, 其中, C_E 和 C_R 分别为融合特征数据的第一协方差矩阵和第二批次特征数据的第二协方差矩阵, \mathbf{v} 为矩阵 $C_E - \alpha C_R$ 的特征向量。其中,融合特征数据是根据第一批次特征数据和第二批次特征数据进行融合得到的。

[0090] 具体地,第一协方差矩阵表征了融合特征数据中任一批次的任一变量之间的关联性,第二协方差矩阵表征了第二批次特征数据中任一批次的任一变量之间的关联性。可知 λ_E 为对角矩阵,其对角线上的元素为第一协方差矩阵的特征值; λ_R 也为对角矩阵,其对角线上的元素为第二协方差矩阵的特征值。

[0091] 具体地, \mathbf{v}^* 为矩阵 $C_E - \alpha C_R$ 的第一个特征向量,其表征了融合特征数据和第二批次特征数据之间差异最大时各变量之间的关系,由此可得到候选主成分方向中表示各变量影响的系数。

[0092] 在一些实施例中,根据目标主成分方向、目标对比度参数、融合特征数据和第二批次特征数据,能够对目标主成分方向上任一变量对应的载荷值进行计算,其形式如下式:

$$[0093] \quad w^* = \left| \sqrt{\lambda^*} v^* \right|$$

[0094] 其中, w^* 为载荷值; λ^* 为矩阵 $C_E - \alpha C_R$ 的第一个特征值; v^* 为目标主成分方向。

[0095] 上述细胞培养过程数据分析方法中,通过候选对比度参数将第二批次特征数据与融合特征数据进行分离,放大了实际蛋白表达量均值较低的第二批次集合与整体批次的差异,从而得到候选主成分方向,为后续引起蛋白表达量下降的主要影响变量提供依据。

[0096] 在一些实施例中,通过以下方式对任一批次的细胞培养过程进行划分,包括以下步骤:

[0097] S110、以任一批次的每个样本为中心样本分别构建移动窗,计算移动窗中其他样本与中心样本之间的距离。

[0098] S120、筛选距离最近的若干个样本作为近邻样本,计算近邻样本与中心样本间的距离均值,作为近邻距离。

[0099] S130、当近邻距离超过分段阈值时,将中心样本对应的时间点作为阶段划分点,以对任一批次的细胞培养过程进行划分。

[0100] 具体地,移动窗的长度为时间长度,即移动窗内包括中心样本所对应时间点之前和之后一定时间内的所有样本。其他样本为移动窗内所有样本中除中心样本以外的样本,其他样本与中心样本之间的距离为时间距离。

[0101] 其中,近邻距离的形式如下式:

$$[0102] \quad d(x_t) = \sqrt{\frac{\sum_{i=1}^k d(x_t, x_{t,i})^2}{k}}$$

[0103] 其中, $d(x_t)$ 为中心样本 x_t 的近邻距离; $d(x_t, x_{t,i})$ 为中心样本 x_t 与第 i 个其他样本 $x_{t,i}$ 之间的距离; k 为近邻样本的总数。

[0104] 上述细胞培养过程数据分析方法中,当近邻距离超过分段阈值时,表明移动窗内其他样本与中心样本的距离均较远,该距离表示任一样本之间的时间距离,即不同样本的采样时间之间存在间隔,而采样时间的间隔即表示不同样本位于不同的细胞培养阶段。在不同阶段中不同批次的在线过程数据和离线检验数据会发生不同的变化,通过划分细胞培养过程的不同阶段将不同批次的在线过程数据和离线检验数据变化统一为相同维度,以便后续分析。

[0105] 在一些实施例中,通过以下方式确定第一批次集合和第二批次集合,包括以下步骤:

[0106] S140、根据多个批次各自的生产方法对多个批次进行聚类,得到多个批次簇。

[0107] S150、确定每个批次簇对应的实际蛋白表达量均值。

[0108] S160、将实际蛋白表达量均值超过预设蛋白表达量阈值的批次簇确定为第一批次集合。

[0109] S170、将实际蛋白表达量均值未超过预设蛋白表达量阈值的批次簇确定为第二批次集合。

[0110] 具体地,在细胞培养过程中,生产方法的不同会造成任一批次产生与其他批次不

同的培养过程,即各变量出现不同的变化,导致最终的蛋白表达量不同。任一批次簇对应的实际蛋白表达量均值即该批次簇对应的生产方法的实际蛋白表达量均值,通过将实际蛋白表达量均值与预设蛋白表达量阈值比较将批次簇分为第一批次集合和第二批次集合,其中,第一批次集合即产量较高的高产批次,第二批次集合即产量较低的低产批次。

[0111] 在一些实施例中,第一批次特征数据的确定方式与第二批次特征数据的确定方式相同;根据第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据,包括以下步骤:

[0112] S210、基于每个第一批次中每个阶段的在线过程数据进行特征提取,得到每个阶段的在线数据特征。

[0113] S220、基于每个第一批次中每个阶段的离线检验数据确定每个阶段的离线数据特征。

[0114] S230、对在线数据特征以及离线数据特征进行融合,得到每个第一批次的融合特征矩阵。

[0115] S240、对每个第一批次的融合特征矩阵进行合并,得到第一批次特征数据。

[0116] 具体地,基于函数型数据分析方法对在线过程数据进行特征提取,即通过基函数对任一批次的任一阶段中任一变量的变化曲线进行拟合,得到包含对应基函数的表达式,提取该表达式中的对应基函数的系数作为该变量的在线数据特征。

[0117] 在一些实施例中,基函数的类型和数量根据任一阶段任一变量对应的参考批次中该变量的变化曲线性质决定。其中,对于任一阶段的任一变量,获取所有批次中该变量的变化曲线,并分别计算任一变化曲线的复杂度,选取其中复杂度最大的变化曲线对应的批次即该阶段该变量的参考批次。

[0118] 具体地,复杂度与该阶段的时间长度和该变量的变化曲线的粗糙度有关;当任一阶段的时间长度越长,任一变量的变化曲线的粗糙度越大,表示该阶段可能包含更多关于变量的特征数据。其中,复杂度的形式如下式:

$$[0119] \quad H_p = \left(\frac{L_p}{L_{\max}} + \sum_{m=1}^M \frac{\overline{D_{p,m}^2}}{\overline{D_{m,\max}^2}} \right) / (M + 1)$$

[0120] 其中, H_p 为复杂度; L_p 为该批次中第 p 个阶段的长度; L_{\max} 为该批次中最大的阶段长度; M 为变量数; $\overline{D_{p,m}^2}$ 为该批次第 p 个阶段中第 m 个变量的二阶差分均值;

$\overline{D_{m,\max}^2}$ 为该批次中 $\overline{D_{p,m}^2}$ 的最大值; $\overline{D_{p,m}^2}$ 满足下式:

$$[0121] \quad \overline{D_{p,m}^2} = \frac{\sum_{t=1}^{L_p-2} [(x_{p,m,t+2} - x_{p,m,t+1}) - (x_{p,m,t+1} - x_{p,m,t})]}{L_p - 2}$$

[0122] 其中, $x_{p,m,t}$ 为该批次第 p 个阶段中第 m 个变量在 t 时刻的值。

[0123] 具体地,根据参考批次中该阶段该变量的变化曲线性质确定基函数。在一些实施例中,对于变化曲线为光滑呈曲线性质的变量,选用B-样条基函数进行拟合;对于变化曲线为粗糙无曲线性质的变量,选用多项式基函数进行拟合。其中,以多项式基函数表示的形式

如下式：

$$[0124] \quad x = \sum c_i t^{i-1}$$

[0125] 其中， t^{i-1} 为第*i*个多项式基函数， c_i 为对应的系数。

[0126] 上述细胞培养过程数据分析方法中，对所有批次中任一阶段的任一变量均采用相同类型和数量的基函数，由此可解决由于任一批次的时间长度不同，使得阶段划分后任一阶段的时间长度不同，导致在分析过程中需要加入引入将各相同阶段处理为相同时间长度的步骤，增加了分析方法的复杂度的问题。通过采用相同数量的基函数，使不同批次的任一阶段中任一变量的特征均为相同维度，降低了分析方法的繁琐程度，提高了分析效率。

[0127] 基于上述实施例提供的一种细胞培养过程数据分析方法，以下描述该方法的一种实际应用场景。在本实际应用场景中，所采集的数据为某企业商业规模细胞培养过程，其培养液体积为2000L，基于所采集的数据对蛋白表达量下降的原因进行分析。

[0128] 在本实际应用场景中，从细胞培养产线采集在线过程数据和离线检验数据，其中在线过程数据包括9个变量，离线检验数据包括7个变量，如表1所示。

[0129] 表1在线数据与离线数据变量

No.	在线数据变量	单位	No.	离线数据变量	单位
1	pH值1	—	10	活细胞密度	10 ⁶ /mL
2	pH值2	—	11	细胞活率	%
3	溶氧量1	%	12	细胞平均直径	μm
4	溶氧量2	%	13	细胞聚集率	%
[0130] 5	温度	°C	14	pH	—
6	底层空气流量	L/min	15	葡萄糖浓度	g/L
7	表层空气流量	L/min	16	乳酸浓度	g/L
8	底层氧气流量	L/min			
9	培养液体积	L			

[0131] 该细胞培养过程共持续12天，在线过程数据的采集间隔为10s，离线检验数据采集间隔约为24h，因此任一批次的在线过程数据约有933120个采样点，而任一批次的离线检验数据有84个采样点。本实施例中共采集了32批数据。

[0132] 在阶段划分过程中，针对任一样本均构建长度L为2400的移动窗，并设定近邻样本的选择数量k为300，其中，长度L表示样本长度的个数，单个样本长度为10s，因此移动窗的长度L为400分钟。在移动窗的范围 $[x_{t-1200}, x_{t+1200}]$ 内分别计算其他样本与中心样本的距离，并从中取出与中心样本距离最小的300个样本作为近邻样本计算近邻距离。

[0133] 请参照图4所示，图4中A表示任一批次的近邻距离分布，可以发现近邻距离中出现4个明显的“尖峰”，在“尖峰”内的样本与其他样本的距离相对较大，因此这些样本与其他样本存在明显差异。所以，这些样本可以认为是各个阶段变换时的过渡样本，而该批次则可以

被这4个“尖峰”分为5个阶段。图4中B表示了32个批次中各个阶段的长度分布,其中任一批次的第五阶段相对其他阶段均较长,达到5500min左右,而其他阶段均在2800min左右。

[0134] 对于划分阶段后的所有批次,根据任一阶段中任一变量的变化曲线的复杂度确定对应的基函数类型和数量,如表2所示,其中B表示B-样条基函数,P表示多项式基函数。

[0135] 表2不同阶段不同变量所采用的基函数类型与数目

变量 编号	阶段1		阶段2		阶段3		阶段4		阶段5	
	类型	数目								
1	B	24	B	9	B	40	B	5	B	8
2	B	24	B	11	B	40	B	5	B	8
3	P	1	P	1	P	1	B	28	B	28
[0136] 4	P	1	P	1	P	1	B	28	B	28
5	P	1	P	1	P	1	P	1	P	1
6	P	1	P	1	P	1	P	1	P	1
7	P	1	P	1	P	1	P	1	P	1
8	P	3	P	3	P	3	B	22	B	20
9	P	2	P	2	P	2	P	2	P	2

[0137] 根据表2所示,可以将在线过程数据由368个基函数进行表达。提取这些基函数的系数,即得到在线过程数据对应的在线数据特征,形成 32×368 的在线数据特征矩阵。对于离线检验数据,根据离线检验数据对应的测量值得到 32×84 的离线数据特征矩阵。将两个矩阵融合,得到 32×452 的扩展特征矩阵,作为融合特征矩阵。此时,细胞培养过程中的离线数据和在线数据可以用融合特征矩阵来表示。

[0138] 请参照图5所示,图5为本发明实际应用场景中所有批次的蛋白表达量的示意图,如图所示,其中簇1、簇2和簇3表示通过不同生产方法进行细胞培养的批次簇,红色虚线表示高产批次与低产批次的划分标准。因此,根据蛋白表达量可以将簇1、簇2和簇3进行分类,其中,簇1和簇3为高产的批次簇,所包含批次作为高产批次,簇2为低产的批次簇,所包含批次作为低产批次。

[0139] 首先考察簇1与簇2之间的差异,将簇1与簇2的特征并集作为 E ,簇2的特征作为 R ,并分别计算对应的协方差矩阵 C_E 和 C_R ,此时可确定包含目标对比度参数 α 的目标主成分方向。为确定目标对比度参数 α ,首先从 $(0,10^4]$ 范围内选取40个待选 α 值,然后针对每一个 α 值,计算对比度参数处理后的 K 和 R ,记为 K'_i 和 R'_i ,计算 K'_i 和 R'_i 的直方图交叉核的值,并以其倒数作为分离度 $D(\alpha_i)$ 。在最大化分离度的同时,尽可能保留 K'_i 的方差

$V(\alpha_i)$, 即目标对比度函数需要满足 $v^* = \underset{v}{\operatorname{argmax}}[\lambda_E(v) - \alpha\lambda_R(v)]$, 此时最少可以保留原

始方差的一半。经过以上步骤, 本实施例中, 簇1与簇2的目标对比度参数为26.3665。

[0140] 确定目标对比度参数后, 即可确定目标主成分方向, 并计算在目标主成分方向上任一变量对应的载荷值, 从而确定各变量对簇1和簇2之间差异的贡献, 如图6所示, 可以看出, v9 (培养液体积) 和v11 (细胞活率) 对差异具有非常高的贡献, 因此这两个变量可能是导致蛋白表达量下降的原因。

[0141] v9和v11的趋势如图7所示, 图7中A为所有批次第一阶段的v9变化趋势, B为所有批次第四阶段的v9变化趋势, C为所有批次第五阶段的v9变化趋势, D为所有批次的v11变化趋势。可以看出高产批次与低产批次的v9的趋势基本是重合的, 而从第10天开始, 高产批次的v11明显下降。因此, v11过高(即细胞活率过高)可能是导致蛋白表达量下降的主要影响因素。

[0142] 进一步地, 请参照图8所示, 图8为本发明实际应用场景中所有批次第五阶段的v11变化趋势示意图, 如图所示, 可以发现在第10天, 簇3中的v9上升(即细胞培养液体积上升)是导致蛋白表达量增加的原因。与所采用的生产方法对照, 可以发现细胞培养液体积上升的原因为, 在簇3包含的批次中, 均在第10天向培养罐里额外补加了部分培养基。

[0143] 综合以上结果, 可以认为, 造成簇2蛋白表达量下降的原因是由于细胞活率过高, 导致营养物质被过多地用于细胞生长, 而不是用于蛋白表达。因此向簇3中包含的批次额外加入培养基, 可以为蛋白表达提供足够的营养, 使得蛋白表达量进一步提高。因此, 今后的生产中应当继续在第10天额外补加适量的培养基, 从而保证蛋白表达量保持在一个高水平。

[0144] 本发明实施例还提供了一种细胞培养过程数据分析系统, 请参照图9所示, 该系统包括数据采集模块1、特征数据提取模块2和变量分析模块3。

[0145] 在一些实施例中, 数据采集模块1用于获取在多个批次的细胞培养过程中采集到的在线过程数据和离线检验数据; 其中, 多个批次划分为第一批次集合和第二批次集合; 第一批次集合的实际蛋白表达量均值高于第二批次集合的实际蛋白表达量均值; 任一批次的细胞培养过程划分为若干阶段。

[0146] 在一些实施例中, 特征数据提取模块2用于根据第一批次集合中第一批次的各阶段的在线过程数据和离线检验数据确定第一批次特征数据, 根据第二批次集合中第二批次的各阶段的在线过程数据和离线检验数据确定第二批次特征数据。

[0147] 在一些实施例中, 变量分析模块3用于根据第一批次特征数据和第二批次特征数据进行对比主成分分析, 得到目标主成分方向, 在目标主成分方向的多个相关变量中确定引起蛋白表达量下降的主要影响变量。

[0148] 在另一些实施例中, 变量分析模块3包括候选对比度参数获取单元、条件确定单元、目标对比度参数确定单元和目标主成分方向确定单元。

[0149] 在一些实施例中, 候选对比度参数获取单元用于获取预设的多个候选对比度参数。

[0150] 在一些实施例中, 条件确定单元用于确定第一批次特征数据和第二批次特征数据在任一候选对比度参数对应的候选主成分方向上的分离度, 以及第一批次特征数据在任一

候选对比度参数对应的候选主成分方向上的特征保留度。

[0151] 在一些实施例中,目标对比度参数确定单元用于基于分离度以及特征保留度在多个候选对比度参数中确定目标对比度参数。目标对比度参数确定单元还包括:根据多个候选对比度参数各自对应的特征保留度与第一批次特征数据的方差的比较结果,对多个候选对比度参数进行筛选,得到筛选后对比度参数;在筛选后对比度参数对应的分离度中确定最大分离度;将最大分离度对应的筛选后对比度参数作为目标对比度参数。

[0152] 在一些实施例中,目标主成分方向确定单元用于将目标对比度参数对应的候选主成分方向作为目标主成分方向。

[0153] 本发明实施例还提供了一种终端设备,包括存储器和处理器,存储器和处理器之间互相通信连接,存储器存储有计算机指令,处理器通过执行计算机指令,从而执行如上述任一实施例提供的一种细胞培养过程数据分析方法。

[0154] 本发明实施例还提供了一种存储介质,存储介质上存储有计算机程序,计算机程序被计算机调用并执行,实现如上述任一实施例提供的一种细胞培养过程数据分析方法。

[0155] 以上所述的具体实施例,对本发明的目的、技术方案和有益效果进行了进一步的详细说明,应当理解,以上所述仅为本发明的具体实施例而已,并不用于限定本发明的保护范围。特别指出,对于本领域技术人员来说,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

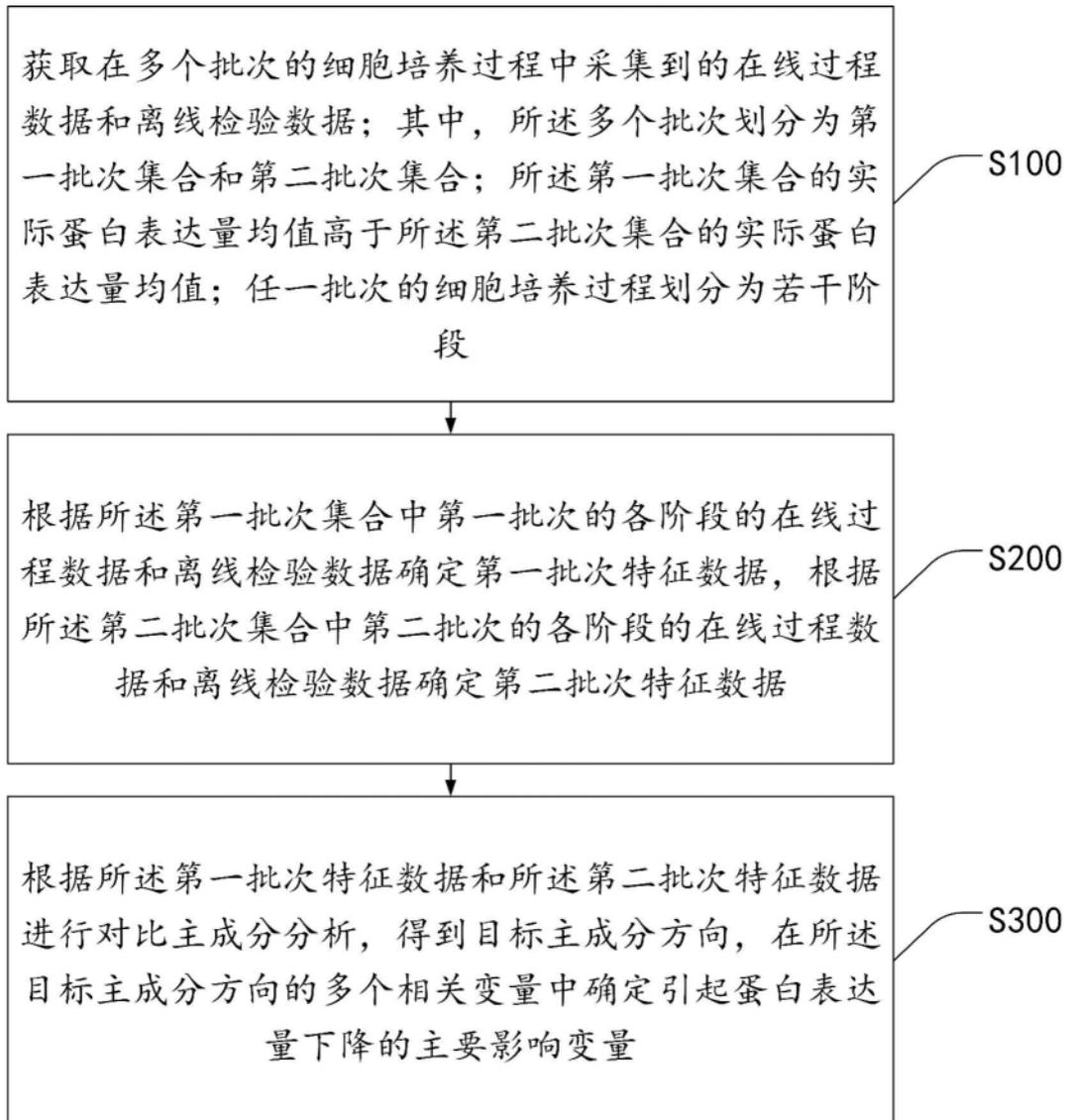


图1

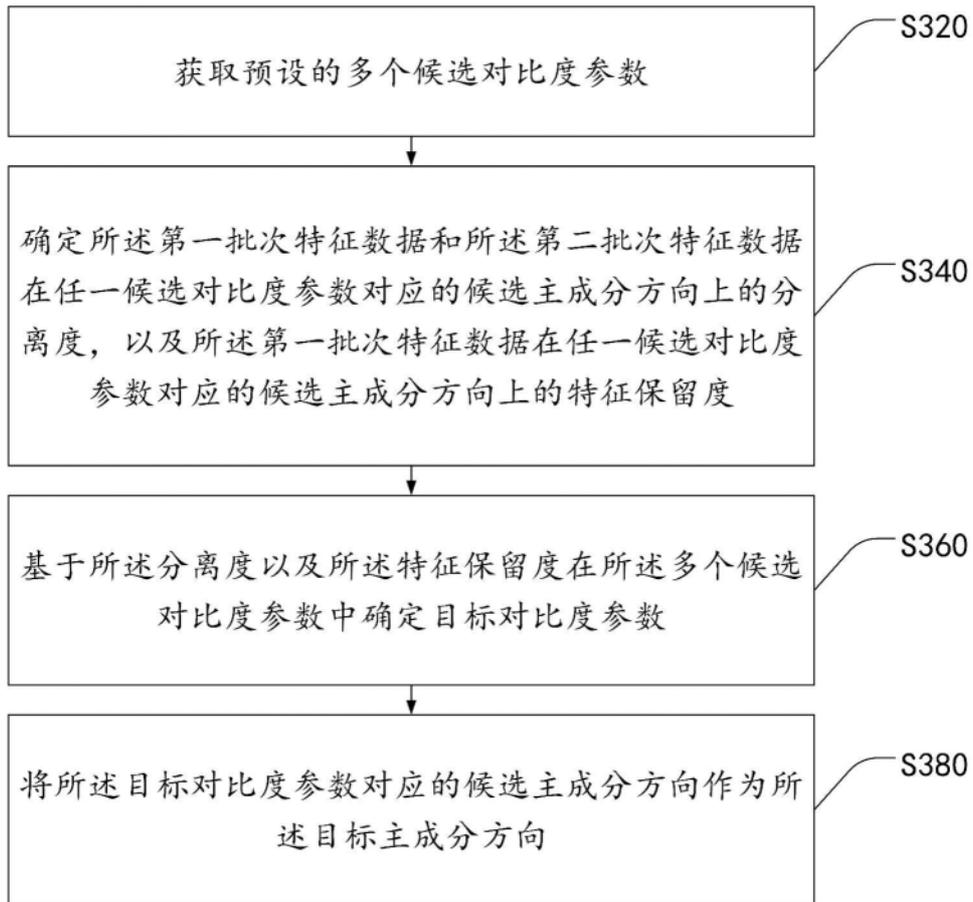


图2

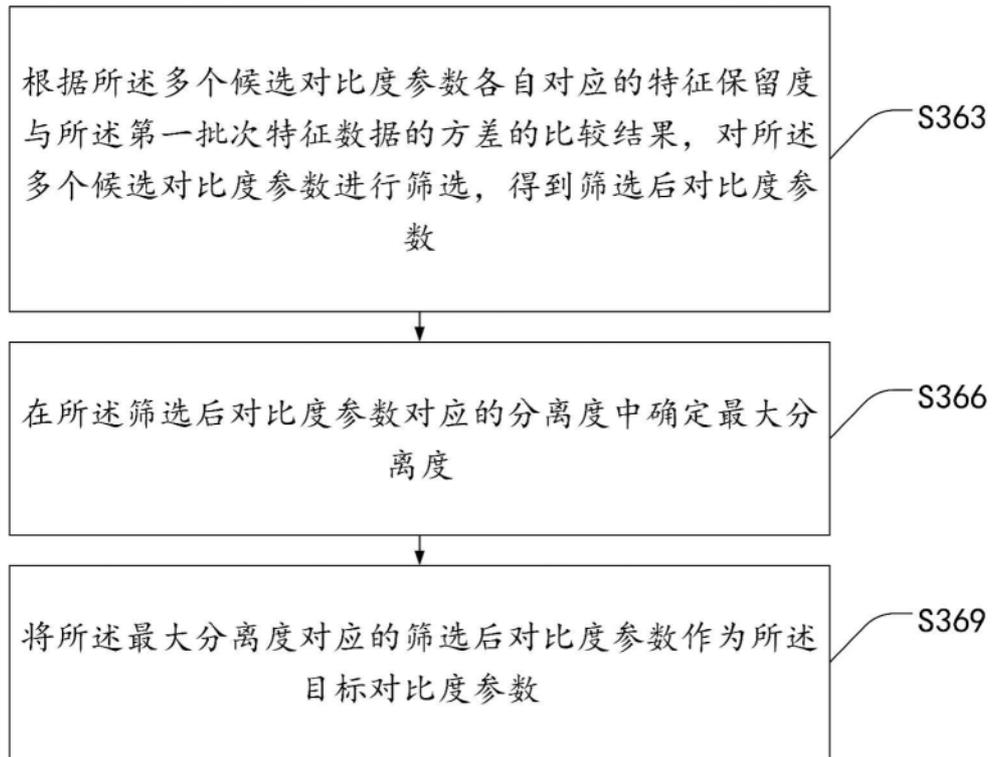


图3

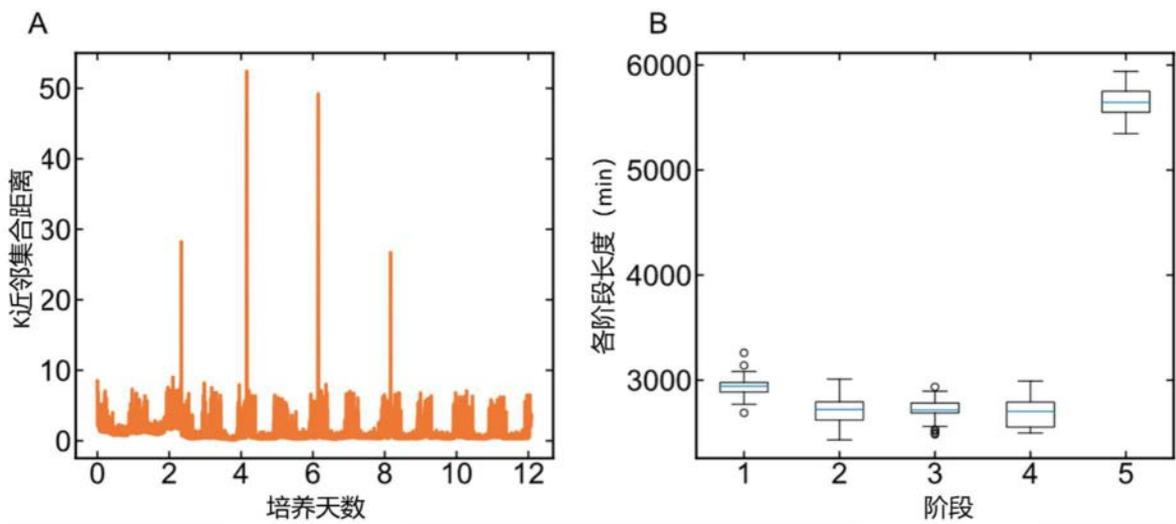


图4

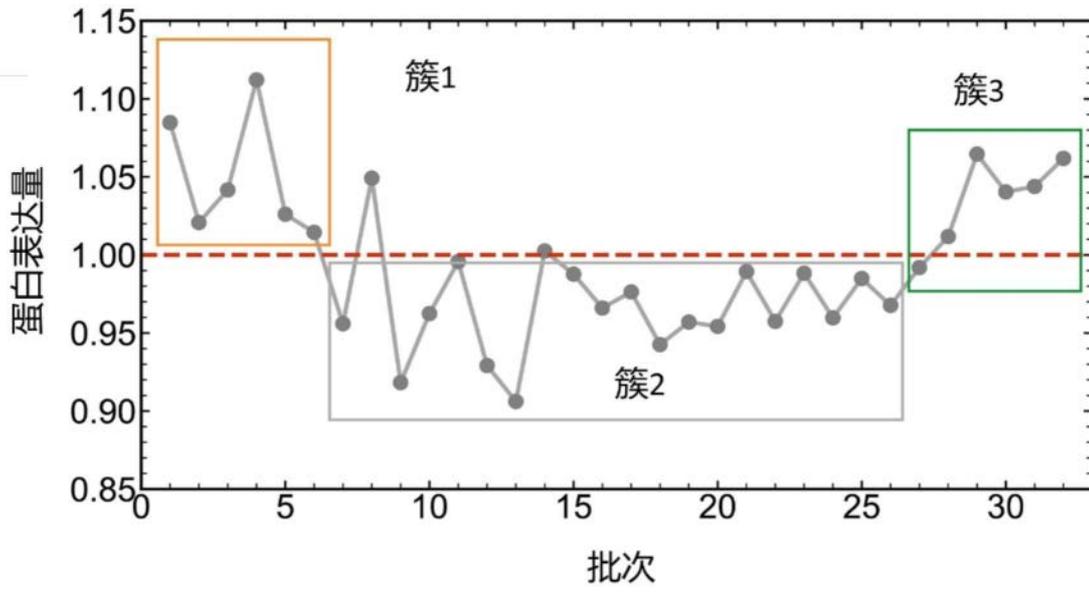


图5

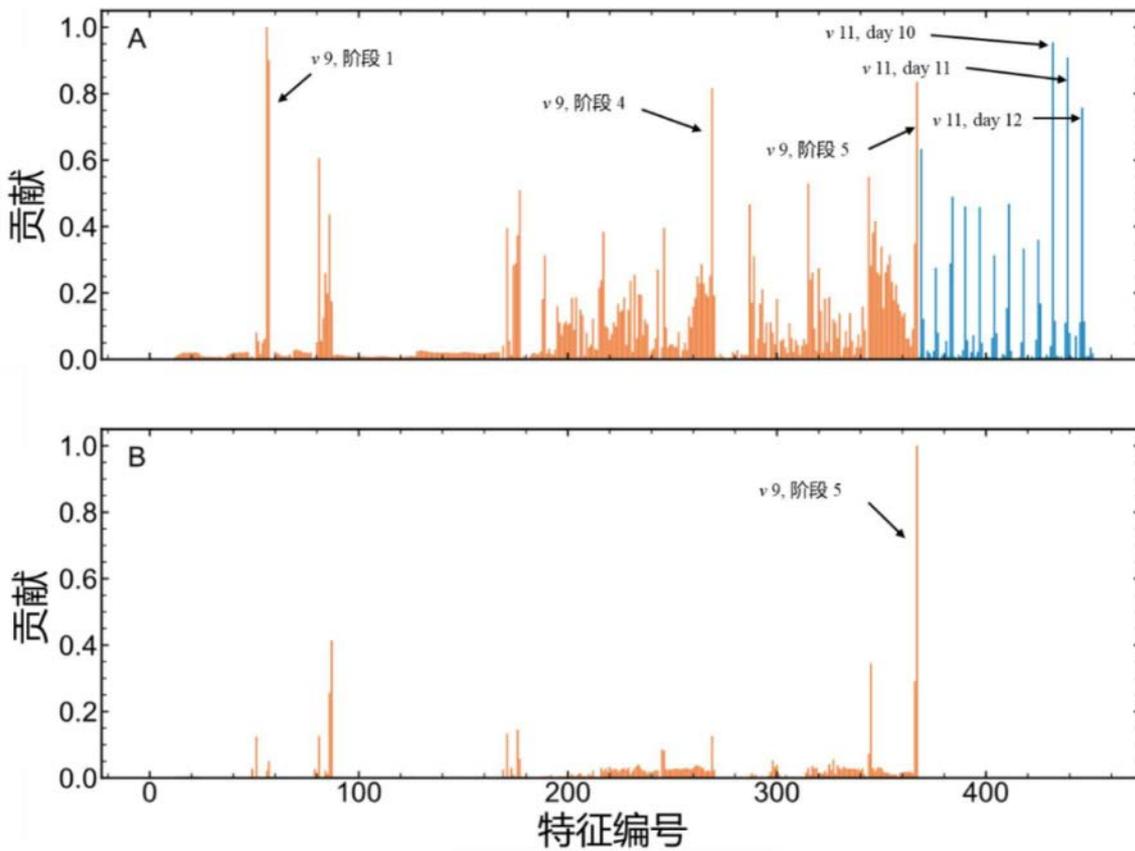


图6

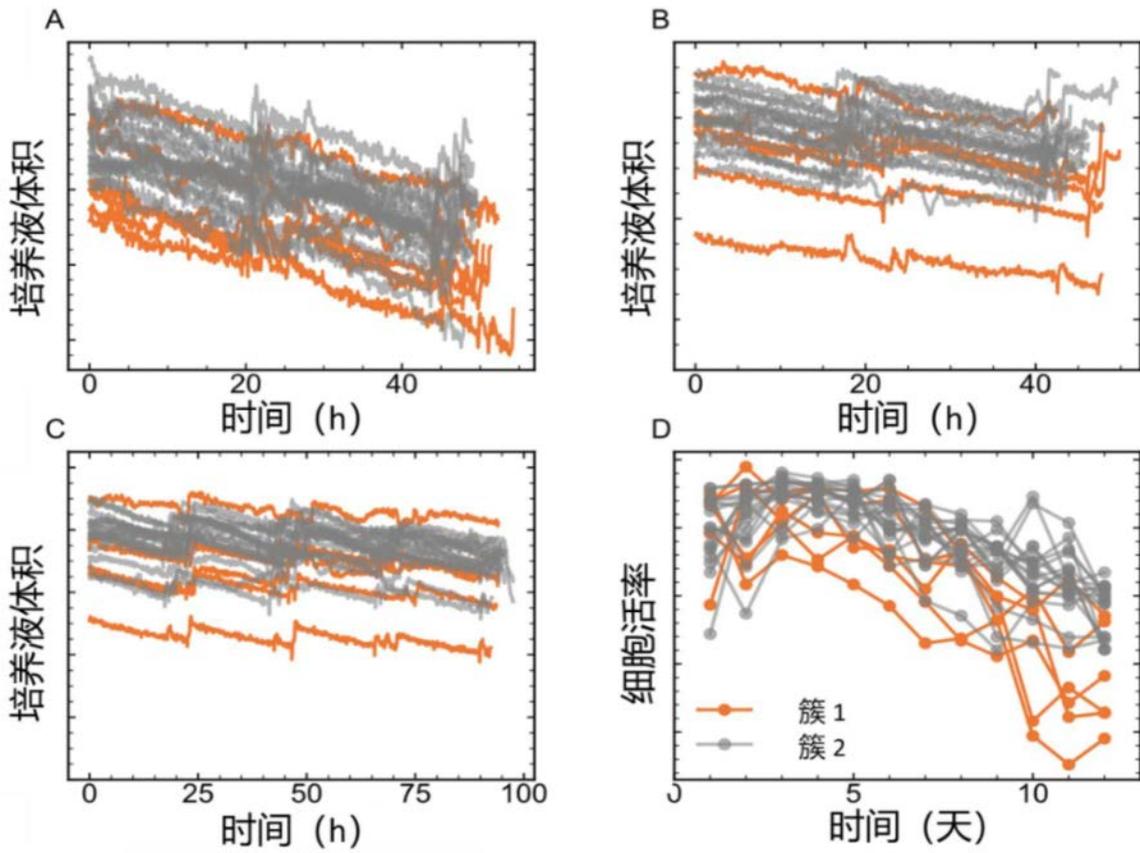


图7

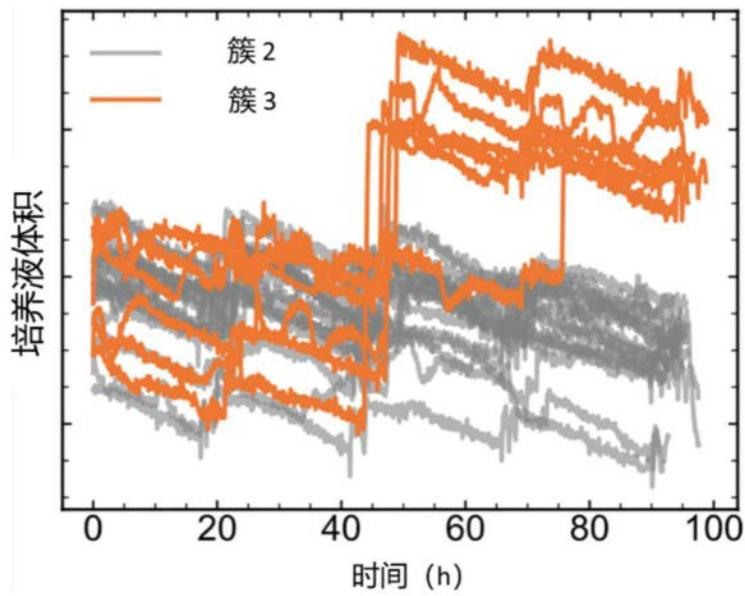


图8

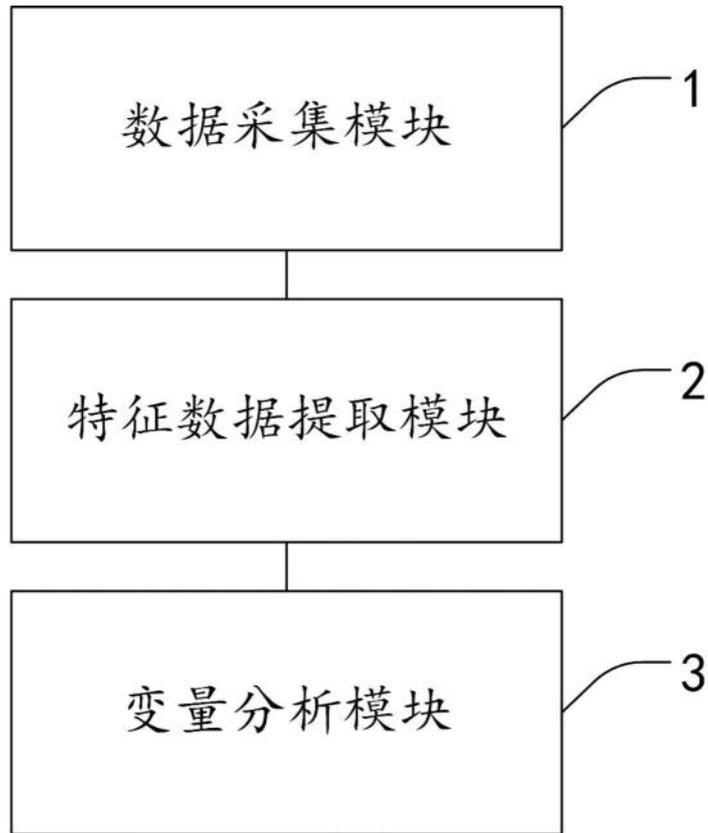


图9