



(12) 发明专利申请

(10) 申请公布号 CN 102841186 A

(43) 申请公布日 2012. 12. 26

(21) 申请号 201210309385. 0

(22) 申请日 2012. 08. 28

(71) 申请人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路 95 号

(72) 发明人 卢朋 代文 高一波 陈琳 刘西
宋江龙 陈迪 温伟娜

(74) 专利代理机构 中科专利商标代理有限责任
公司 11021

代理人 宋焰琴

(51) Int. Cl.
G01N 33/15(2006. 01)

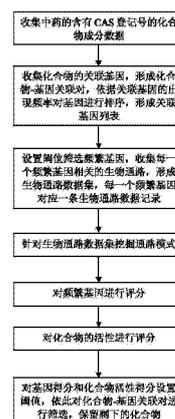
权利要求书 2 页 说明书 7 页 附图 5 页

(54) 发明名称

基于通路模式挖掘的中药活性成分预测方法

(57) 摘要

本发明公开了一种基于通路模式挖掘的中药活性成分预测方法,该方法包括如下步骤:收集中药的含有 CAS 登记号的化合物成分数据;收集化合物的关联基因,形成化合物-基因关联对,依据关联基因的出现频率对基因进行排序,形成关联基因列表;设置阈值筛选频繁基因,形成生物通路数据集,每一个频繁基因对应一条生物通路数据记录;针对生物通路数据集挖掘通路模式;对频繁基因进行评分;对化合物活性进行评分;对基因得分和化合物活性得分设置阈值,依此对化合物-基因关联对进行筛选,保留剩下的化合物,从而预测中药的活性成分。



1. 一种基于通路模式挖掘的中药活性成分预测方法,该方法包括如下步骤:

步骤 1:收集中药的含有登记号的化合物成分数据;

步骤 2:查询第一数据库,收集化合物的关联基因,形成化合物-基因关联对,依据所述关联基因的出现频率对基因进行排序,形成关联基因列表;

步骤 3:设置阈值筛选频繁基因,通过查询第二数据库,收集每一个频繁基因相关的生物通路,形成生物通路数据集;

步骤 4:针对所述生物通路数据集,选取算法,挖掘通路模式;

步骤 5:对频繁基因进行评分;

步骤 6:对化合物活性进行评分;

步骤 7:对所述基因得分和所述化合物活性得分设置阈值,对化合物-基因关联对进行筛选,保留剩下的化合物。

2. 如权利要求 1 所述的方法,其特征在于,所述登记号为 CAS 登记号。

3. 如权利要求 1 所述的方法,其特征在于,所述第一数据库为 CTD 数据库。

4. 如权利要求 1 所述的方法,其特征在于,所述第二数据库为 KEGG 数据库。

5. 如权利要求 1 所述的方法,其特征在于,所述算法为双向关联规则算法。

6. 如权利要求 3 所述的方法,其特征在于,从所述第一数据库收集化合物的关联基因时,去除没有关联基因信息的化合物数据。

7. 如权利要求 4 所述的方法,其特征在于,从所述第二数据库收集基因相关的生物通路时,去除不属于人类的基因和没有生物通路信息的基因。

8. 如权利要求 5 所述的方法,其特征在于,所述步骤 4 包括:遍历所述生物通路数据集的所有数据记录,找到满足支持度要求的 1-项通路频繁集;从 k-项通路频繁集通过连接-剪枝操作生成候选 (k+1)-项通路频繁集;对所述候选 (k+1)-项通路频繁集的检验;对于数据挖掘的结果进行人工筛选,其中 k 为自然数。

9. 如权利要求 1 所述的方法,其特征在于,所述步骤 5 采用如下表达式实现评分:

$$S_g = \frac{1}{k} \sum_{i=1}^k \frac{h_i}{N_i}$$

其中, S_g 表示基因得分, k 表示通路模式中通路频繁集的最高项数, k 为自然数, N_i 表示的是 i-项通路频繁集中双向关联规则的个数, h_i 表示的是基因相关的生物通路中能构成的 i-项双向关联规则的个数, i 为自然数。

10. 如权利要求 1 所述的方法,其特征在于,所述步骤 6 采用如下表达式实现评分:

$$S_c = \frac{1}{2} fun_c + \frac{1}{2} spe_c$$

其中, fun_c 表示化合物关联基因的功能性评价,表达式为 $fun_c = \frac{\sum_{j=1}^m S_{gj}}{S}$;

spe_c 表示化合物关联基因的特异性评价,表达式为 $spe_c = \frac{N_{gt}}{N}$;

其中, m 表示化合物关联基因的个数, S_{gj} 表示化合物第 j 个关联基因的得分, S 表示所有频繁基因的得分之和, j 为自然数;

N 表示化合物关联基因的个数, N_{gt} 表示化合物关联基因中得分大于零的基因个数。

基于通路模式挖掘的中药活性成分预测方法

技术领域

[0001] 本发明涉及计算机领域在中药活性成分研究中的应用领域,特别涉及一种基于通路模式挖掘的中药活性成分预测方法。

背景技术

[0002] 中药发展历史悠久,通过古籍记载和经验传授流传至今。近年来,越来越多的国内外学者开始关注中药的研究。为了研究中药的作用机理,首先需要弄清中药的物质基础。中药成分繁多,成分之间作用机理复杂,往往多种成分协同作用于多个基因靶标。找出中药的活性成分,是中药物质基础研究中十分关键的一步,同时也有助于认识中药复杂的作用机理。在中医药的研究中,将宝贵的中医药经验用这种更加标准更加科学的方式来解释,对于中医药的发扬光大是至关重要的。

[0003] 目前,主要有两类计算模型用于研究药物成分与药物功效的关系,从而找到药物活性成分。第一类计算模型是将化合物的生物活性与化合物的结构和化学特性联系起来,这一类研究建立在“化合物的生物活性依赖其结构和化学特性”的假设之上。但是,获取中药所有成分的结构信息比较困难,这限制了这一模型的有限性。第二类计算模型是将复杂系统的生物活性与系统的化合物构成关联起来,但是,目前仍然缺乏准确的计算模型来量化这种关联关系。

[0004] 针对这种情况,有必要设计更多有效的方法来预测中药的活性成分,从而揭示中药的物质基础,解释中药的作用机理。

发明内容

[0005] (一) 要解决的技术问题

[0006] 本发明所要解决的技术问题在于提供一种基于通路模式挖掘的中药活性成分预测方法能较准确地预测出中药的活性成分。

[0007] (二) 技术方案

[0008] 为解决上述技术问题,本发明提供一种基于通路模式挖掘的中药活性成分预测方法,该方法包括如下步骤:

[0009] 步骤1:收集中药的含有登记号的化合物成分数据;

[0010] 步骤2:查询第一数据库,收集化合物的关联基因,形成化合物-基因关联对,依据所述关联基因的出现频率对基因进行排序,形成关联基因列表;

[0011] 步骤3:设置阈值筛选频繁基因,通过查询第二数据库,收集每一个频繁基因相关的生物通路,形成生物通路数据集;

[0012] 步骤4:针对所述生物通路数据集,选取算法,挖掘通路模式;

[0013] 步骤5:对频繁基因进行评分;

[0014] 步骤6:对化合物活性进行评分;

[0015] 步骤7:对所述基因得分和所述化合物活性得分设置阈值,对化合物-基因关联对

进行筛选,保留剩下的化合物。

[0016] 优选地,所述登记号为 CAS 登记号。

[0017] 优选地,所述第一数据库为 CTD 数据库。

[0018] 优选地,所述第二数据库为 KEGG 数据库。

[0019] 优选地,所述算法为双向关联规则算法。

[0020] 优选地,从所述第一数据库收集化合物的关联基因时,去除没有关联基因信息的化合物数据。

[0021] 优选地,从所述第二数据库收集基因相关的生物通路时,去除不属于人类的基因和没有生物通路信息的基因。

[0022] 优选地,所述步骤 4 还包括:遍历所述生物通路数据集的所有数据记录,找到满足支持度要求的 1-项通路频繁集;从 k-项通路频繁集通过连接-剪枝操作生成候选 (k+1)-项通路频繁集;对所述候选 (k+1)-项通路频繁集的检验;对于数据挖掘的结果进行人工筛选,其中 k 为自然数。

[0023] 优选地,所述步骤 5 采用如下表达式实现评分:

$$[0024] \quad S_g = \frac{1}{k} \sum_{i=1}^k \frac{h_i}{N_i}$$

[0025] 其中, S_g 表示基因得分, k 表示通路模式中通路频繁集的最高项数, k 为自然数, N_i 表示的是 i-项通路频繁集中双向关联规则的个数, h_i 表示的是基因相关的生物通路中能构成的 i-项双向关联规则的个数, i 为自然数。

[0026] 优选地,所述步骤 6 采用如下表达式实现评分:

$$[0027] \quad S_c = \frac{1}{2} fun_c + \frac{1}{2} spe_c$$

[0028] 其中, fun_c 表示化合物关联基因的功能性评价,表达式为 $fun_c = \frac{\sum_{j=1}^m S_{gj}}{S}$;

[0029] spe_c 表示化合物关联基因的特异性评价,表达式为 $spe_c = \frac{N_{gt}}{N}$;

[0030] 其中, m 表示化合物关联基因的个数, S_{gj} 表示化合物第 j 个关联基因的得分, S 表示所有频繁基因的得分之和, j 为自然数;

[0031] N 表示化合物关联基因的个数, N_{gt} 表示化合物关联基因中得分大于零的基因个数。

[0032] (三) 有益效果

[0033] 本发明所提供的方法,着重于对生物信息的数据挖掘,相较于生物辅助软件的操作,更加简便易行;所使用的基因数据和生物通路数据均来源于公开生物数据库,数据可靠、有保障;通过本发明所提供的方法能够比较准确的预测出中药的活性成分,并对成分所作用的基因靶标作出分析。

附图说明

[0034] 图 1 是根据本发明的一种基于通路模式挖掘的中药活性成分预测方法的流程图;

- [0035] 图 2 是根据本发明具体实施例的操作流程示意图；
- [0036] 图 3 是根据本发明具体实施例的基因频率分布示意图；
- [0037] 图 4 是根据本发明具体实施例的化合物 - 基因 - 生物通路映射过程的示意图；
- [0038] 图 5 是根据本发明具体实施例的双向关联规则算法流程的示意图。

具体实施方式

[0039] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明作进一步的详细说明。

[0040] 本发明以化合物 - 基因 - 生物通路的映射顺序，从中药的成分化合物数据关联到生物通路数据，针对生物通路数据进行数据挖掘，提取出通路模式作为中药所治疗疾病病理的描述，以通路模式为标准，分别对基因重要性和化合物活性进行量化衡量，从而筛选出中药的活性成分。

[0041] 图 1 是根据本发明的一种基于通路模式挖掘的中药活性成分预测方法的流程图。如图 1 所示，本发明提供一种基于通路模式挖掘的中药活性成分预测方法，该方法包括如下步骤：

[0042] 步骤 1：收集中药的含有 CAS 登记号 (Chemical Abstracts Service Number) 的化合物成分数据；

[0043] 步骤 2：利用化合物的 CAS 登记号查询 CTD 数据库 (Comparative Toxicogenomics Database)，收集化合物的关联基因，形成化合物 - 基因关联对，同时依据关联基因的出现频率对基因进行排序，形成关联基因列表；

[0044] 步骤 3：设置阈值筛选频繁基因，通过查询 KEGG 数据库 (Kyoto Encyclopedia of Genes and Genomes) 收集每一个频繁基因相关的生物通路，形成生物通路数据集，每一个频繁基因对应一条生物通路数据记录；

[0045] 步骤 4：针对生物通路数据集，利用双向关联规则算法挖掘通路模式。

[0046] 优选地，所述步骤 4 包括：

[0047] 步骤 41：遍历生物通路数据集的所有数据记录，找到满足支持度要求的 1-项通路频繁集；

[0048] 步骤 42：从 k-项通路频繁集通过“连接 - 剪枝”操作生成候选 (k+1)-项通路频繁集，k 为自然数；

[0049] 步骤 43：候选 (k+1)-项通路频繁集的检验；

[0050] 首先进行支持度的检验，去除不满足支持度要求的 (k+1)-项通路；然后进行置信度的检验，依据双向关联规则挖掘的原则，频繁通路共同出现的次数与其中任意一个通路单独出现的次数的比值均要大于或等于置信度。满足以上要求的候选 (k+1)-项通路（亦即一条双向关联规则）构成 (k+1)-项通路频繁集。若 (k+1)-项通路频繁集存在，则转至步骤 42，对新生成的 (k+1)-项通路频繁集进行相同处理，若不存在则停止步骤 42、步骤 43 的循环，将生成的前 k-项通路频繁集代入步骤 44，进行下一步的处理，k 为自然数；

[0051] 步骤 44：对于数据挖掘的结果进行人工筛选；

[0052] 少数生物通路明显与中药所治疗的疾病无关，它们是由数据挖掘引入的杂项，因此通过人工筛选将包含这些通路的双向关联规则去除。最后，筛选后的所有 k-项通路频繁

集构成了通路模式, k 为自然数。

[0053] 步骤 5: 根据频繁基因的相关生物通路与通路模式匹配的情况, 对频繁基因进行评分, 用于衡量基因在疾病病理中的重要程度。基因得分 S_g 定义为表达式 (1):

$$[0054] \quad S_g = \frac{1}{k} \sum_{i=1}^k \frac{h_i}{N_i}$$

[0055] 其中, k 表示通路模式中通路频繁集的最高项数, N_i 表示的是 i -项通路频繁集中双向关联规则的个数, h_i 表示的是基因相关的生物通路中能构成的 i -项双向关联规则的个数。

[0056] 步骤 6: 根据化合物关联基因的得分情况, 对化合物的活性进行评分, 化合物活性得分 S_c 定义为表达式 (4):

$$[0057] \quad S_c = \frac{1}{2} fun_c + \frac{1}{2} spe_c$$

[0058] 其中, fun_c 表示化合物关联基因的功能性评价, 定义为表达式 (2):

$$[0059] \quad fun_c = \frac{\sum_{j=1}^m S_{gj}}{S}$$

[0060] 对于表达式 (2), m 表示化合物关联基因的个数, S_{gj} 表示化合物第 j 个关联基因的得分, S 表示所有频繁基因的得分之和。

[0061] spe_c 表示化合物关联基因的特异性评价, 定义为表达式 (3):

$$[0062] \quad spe_c = \frac{N_{gt}}{N}$$

[0063] 对于表达式 (3), N 表示化合物关联基因的个数, N_{gt} 表示化合物关联基因中得分大于零的基因个数;

[0064] 步骤 7: 对基因得分和化合物活性得分设置阈值, 依此对化合物-基因关联对进行筛选, 保留剩下的化合物, 从而预测了中药的活性成分。

[0065] 下面以预测中药方剂麻杏石甘汤-银翘散的活性成分为例对本发明做进一步阐述。图 2 是根据本发明具体实施例的操作流程示意图, 如图 2 所示, 操作包括如下步骤:

[0066] 步骤 1: 收集成分数据。

[0067] 通过咨询中药专家和查阅文献, 获知麻杏石甘汤-银翘散方剂由 12 味中药组成, 包括: 甘草、麻黄、银花、知母、黄芩、杏仁、连翘、薄荷、浙贝母、牛蒡子、青蒿和石膏。获得这 12 味中药中含有 CAS 登记号的化合物成分共计 541 个。

[0068] 步骤 2: 整理关联基因。

[0069] 通过所述化合物的 CAS 登记号查询 CTD 数据库, 获得所述化合物的关联基因信息。由于所述化合物中有些化合物并不常见, CTD 数据库并未收录或没有“化合物-基因”关联信息, 因此去除这些化合物数据。处理之后, 共获得 153 个化合物、7895 个关联基因和 14603 个化合物-基因关联信息。

[0070] 依据 7895 个基因在不同化合物关联基因中出现的频率 (或次数), 对它们进行排序。基因数目随着基因频率变化, 其分布参见图 3。

[0071] 步骤 3: 准备生物通路数据。

[0072] 设置基因频率阈值为 8,保留频率大于或等于 8 的基因,获得 172 个频繁基因。去除其中没有相关生物通路信息的基因和不属于人类的基因,剩下 152 个频繁基因。

[0073] 查询 KEGG 数据库,准备好 152 个频繁基因的相关生物通路信息。一个频繁基因的相关生物通路数据构成一条数据记录。

[0074] 化合物 - 基因 - 生物通路的映射过程的详细描述参见图 4。

[0075] 步骤 4 :挖掘通路模式。

[0076] 双向关联规则算法的流程参见图 5。将双向关联规则算法中的支持度设置为 0.09,置信度设置为 0.55。挖掘结果包括 38 个 1-项双向关联规则、24 个 2-项双向关联规则和 2 个 3-项双向关联规则。麻杏石甘汤 - 银翘散用于治疗甲流,因此进行人工筛选时只保留感染类疾病、肺部疾病和细胞活动相关的生物通路。这样,筛选后的通路模式由 18 个 1-项双向关联规则、14 个 2-项双向关联规则和 1 个 3-项双向关联规则构成。通路模式的构成情况参见下表。

KEGG通路号	通路描述
<i>1-项通路频繁集</i>	
hsa01100	代谢通路
hsa05152	肺结核
hsa05142	南美锥虫病（美洲锥虫病）
hsa05164	甲型流感病毒
hsa04010	MAPK信号通路
hsa04210	细胞凋亡
hsa05169	EB病毒感染
hsa04060	细胞因子-细胞因子受体交互作用
hsa05168	单纯疱疹病毒感染
hsa04621	NLR信号通路
hsa04510	细胞粘着
hsa05133	百日咳
hsa04660	T细胞受体信号通路
hsa04620	TLR信号通路
hsa05132	沙门氏菌感染
hsa05162	麻疹
hsa04110	细胞周期
hsa05323	风湿性关节炎
<i>2-项通路频繁集</i>	
hsa05142; hsa05164	南美锥虫病（美洲锥虫病）；甲型流感病毒
hsa05142; hsa05152	南美锥虫病（美洲锥虫病）；肺结核
hsa04620; hsa05142	TLR信号通路；南美锥虫病（美洲锥虫病）
hsa05152; hsa05164	肺结核；甲型流感病毒
hsa05142; hsa05168	南美锥虫病（美洲锥虫病）；单纯疱疹病毒感染
hsa05164; hsa05168	甲型流感病毒；单纯疱疹病毒感染
hsa04010; hsa05142	MAPK信号通路；南美锥虫病（美洲锥虫病）
hsa04621; hsa05164	NLR信号通路；甲型流感病毒
hsa04620; hsa05164	TLR信号通路；甲型流感病毒
hsa04210; hsa05152	细胞凋亡；肺结核
hsa04621; hsa05142	NLR信号通路；南美锥虫病（美洲锥虫病）
hsa05132; hsa05133	沙门氏菌感染；百日咳
hsa05133; hsa05152	百日咳；肺结核
hsa05133; hsa05142	百日咳；南美锥虫病（美洲锥虫病）
<i>3-项通路频繁集</i>	
hsa04620; hsa05142; hsa05164	TLR信号通路；南美锥虫病（美洲锥虫病）；甲型流感病毒

[0077]

[0078] 步骤5：计算基因得分。

[0079] 将152个频繁基因的相关通路分别与通路模式进行匹配，依据表达式(1)对152个基因进行评分。

$$[0080] \quad S_g = \frac{1}{k} \sum_{i=1}^k \frac{h_i}{N_i}$$

[0081] 步骤 6 : 计算化合物活性得分。

[0082] 根据以上化合物关联基因的得分情况, 依据表达式 (2) 和 (3) 分别对功能性和特异性进行评分。

$$[0083] \quad fun_c = \frac{\sum_{j=1}^m S_{gj}}{S}$$

$$[0084] \quad spe_c = \frac{N_{gt}}{N}$$

[0085] 然后依据表达式 (4) 对化合物的活性进行评分。

$$[0086] \quad S_c = \frac{1}{2} fun_c + \frac{1}{2} spe_c$$

[0087] 步骤 7 : 筛选活性成分。

[0088] 设置化合物活性得分阈值为 0.5, 基因得分阈值为 0.1。在这两个阈值之上分别有 30 个化合物和 30 个基因。遍历步骤 2 中的 14603 个化合物 - 基因关联对, 筛选出以上 30 个化合物和 30 个基因之间的关联对。如此处理后, 剩下 16 个化合物, 29 个基因和 162 个化合物 - 基因交互对。其中, 16 个化合物就是麻杏石甘汤 - 银翘散的活性成分预测结果, 29 个基因则是活性成分作用的基因靶标, 29 个基因和 162 个化合物 - 基因交互对共同解释了这些活性成分的作用机理。

[0089] 对于预测的麻杏石甘汤 - 银翘散的活性成分, 分子对接实验和医学文献验证它们是比较准确的。其中一些活性成分已经引起了越来越多研究者的兴趣, 并且研究实验证明它们对流感的治疗作用明显。

[0090] 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详细说明, 应理解的是, 以上所述仅为本发明的具体实施例而已, 并不用于限制本发明, 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。

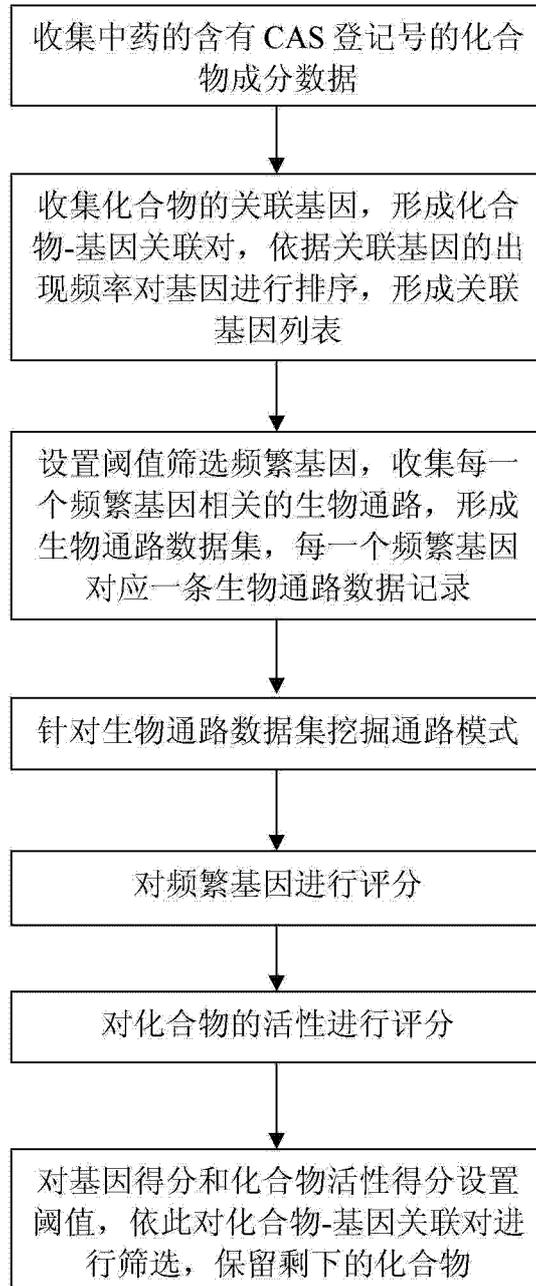


图 1

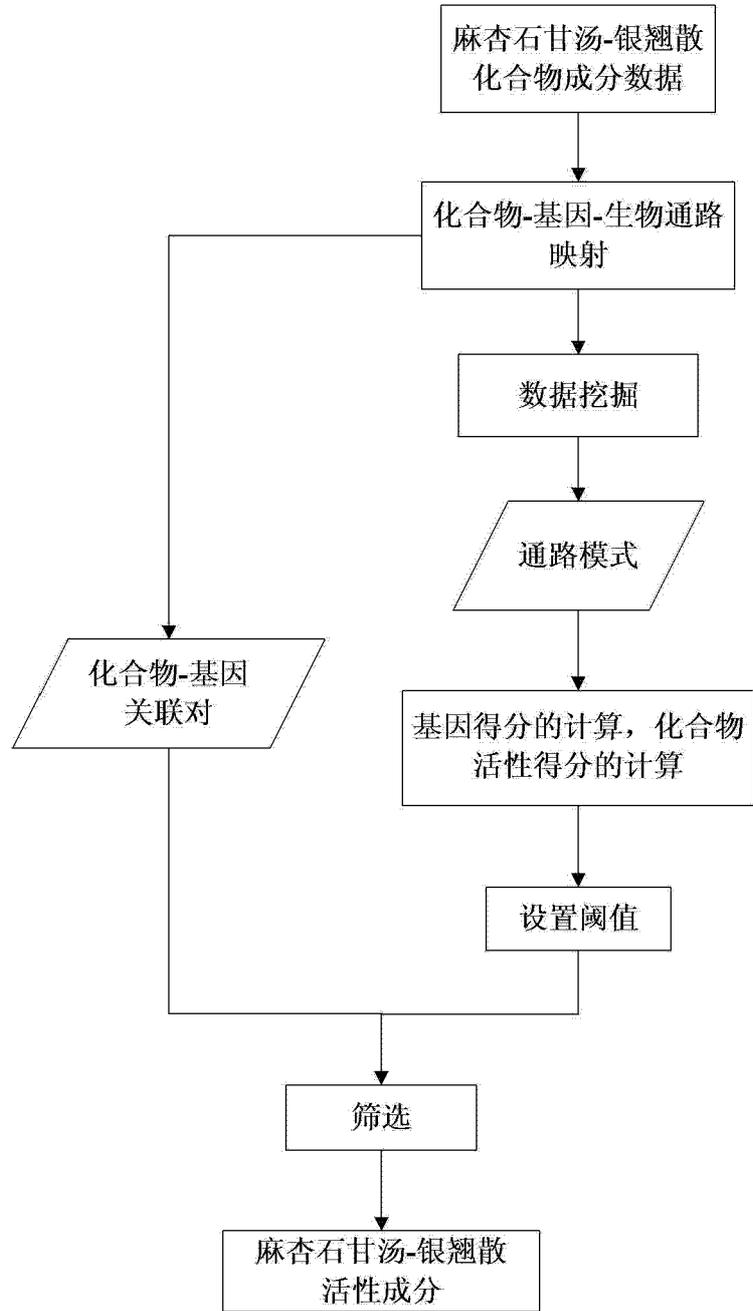


图 2

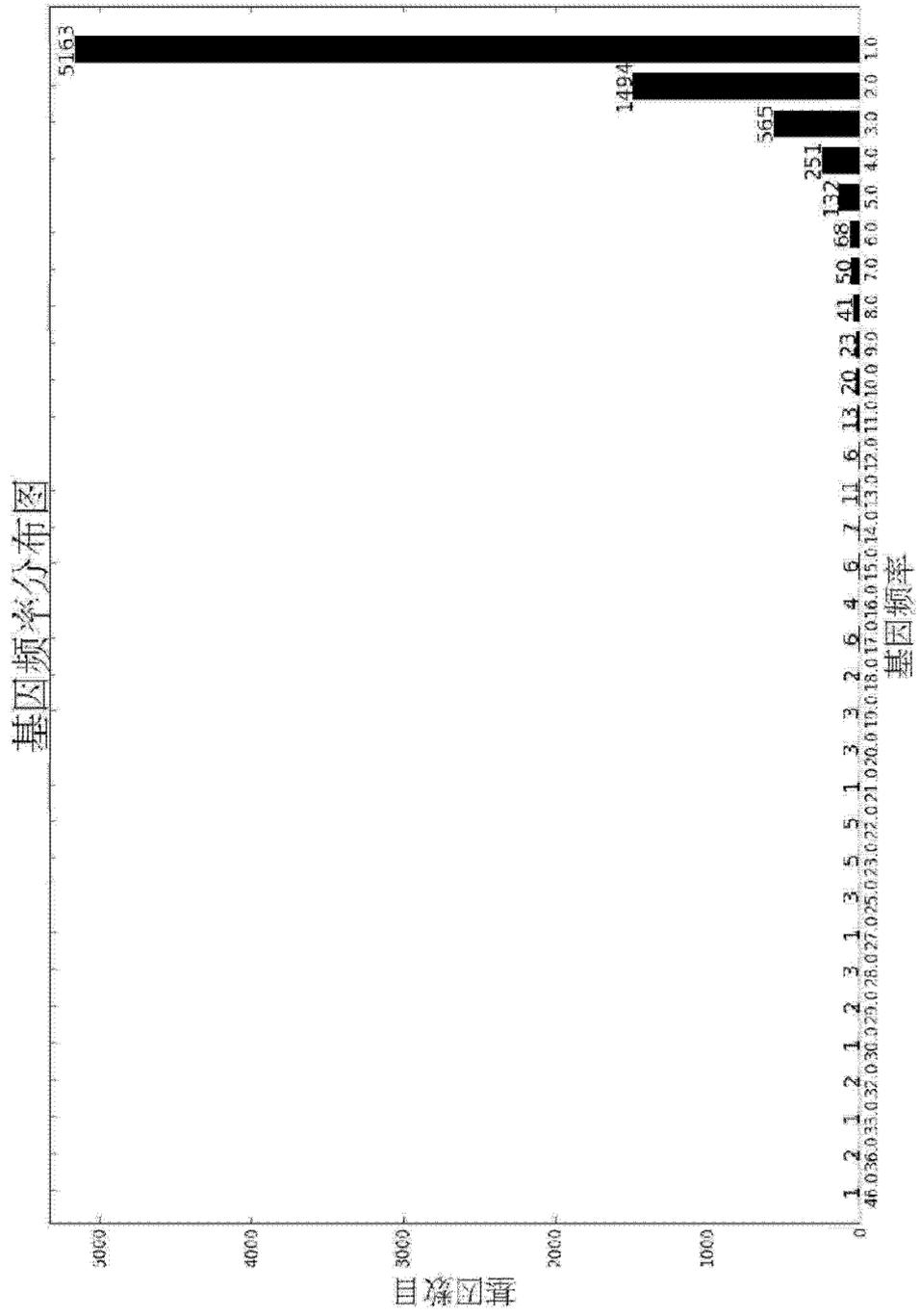


图 3

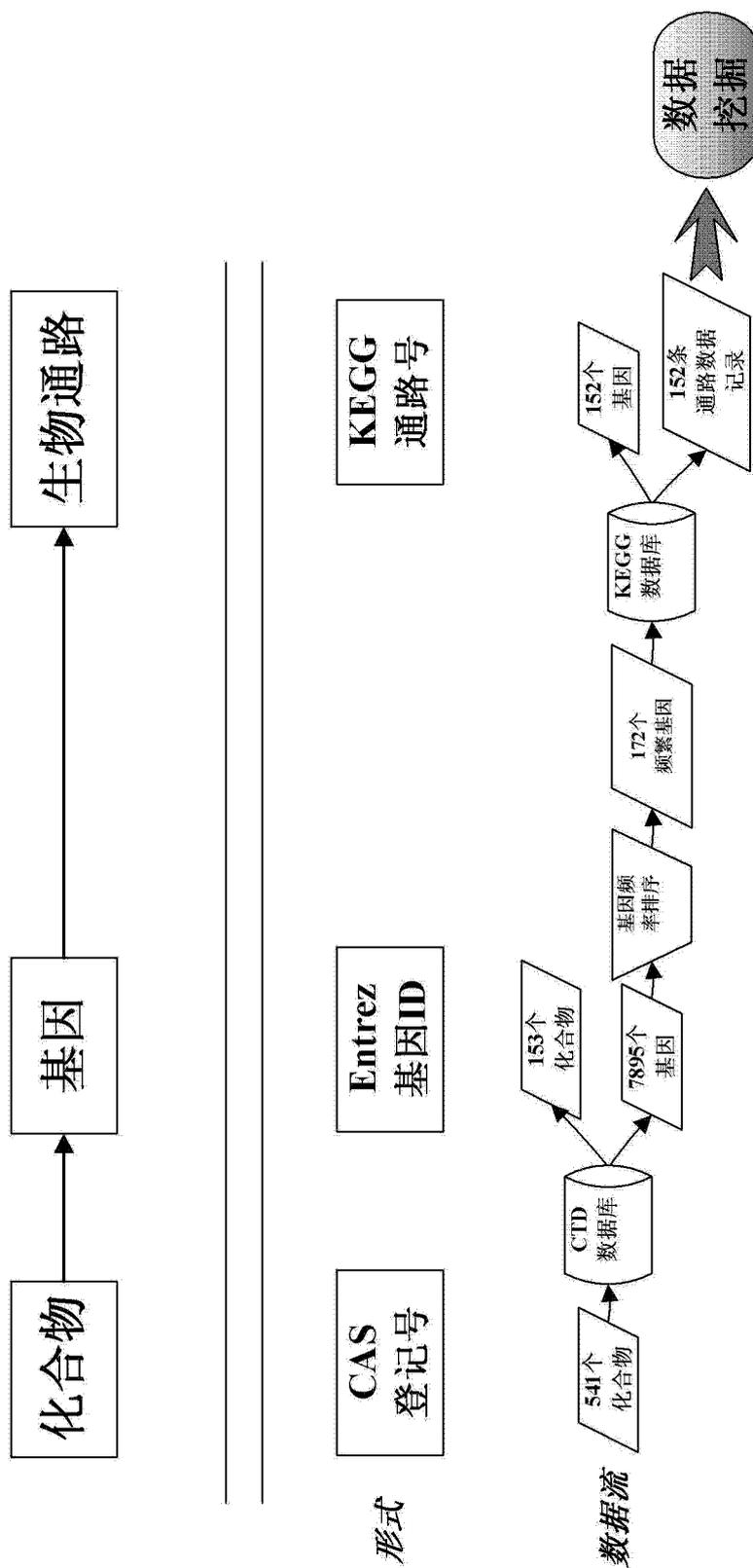


图 4

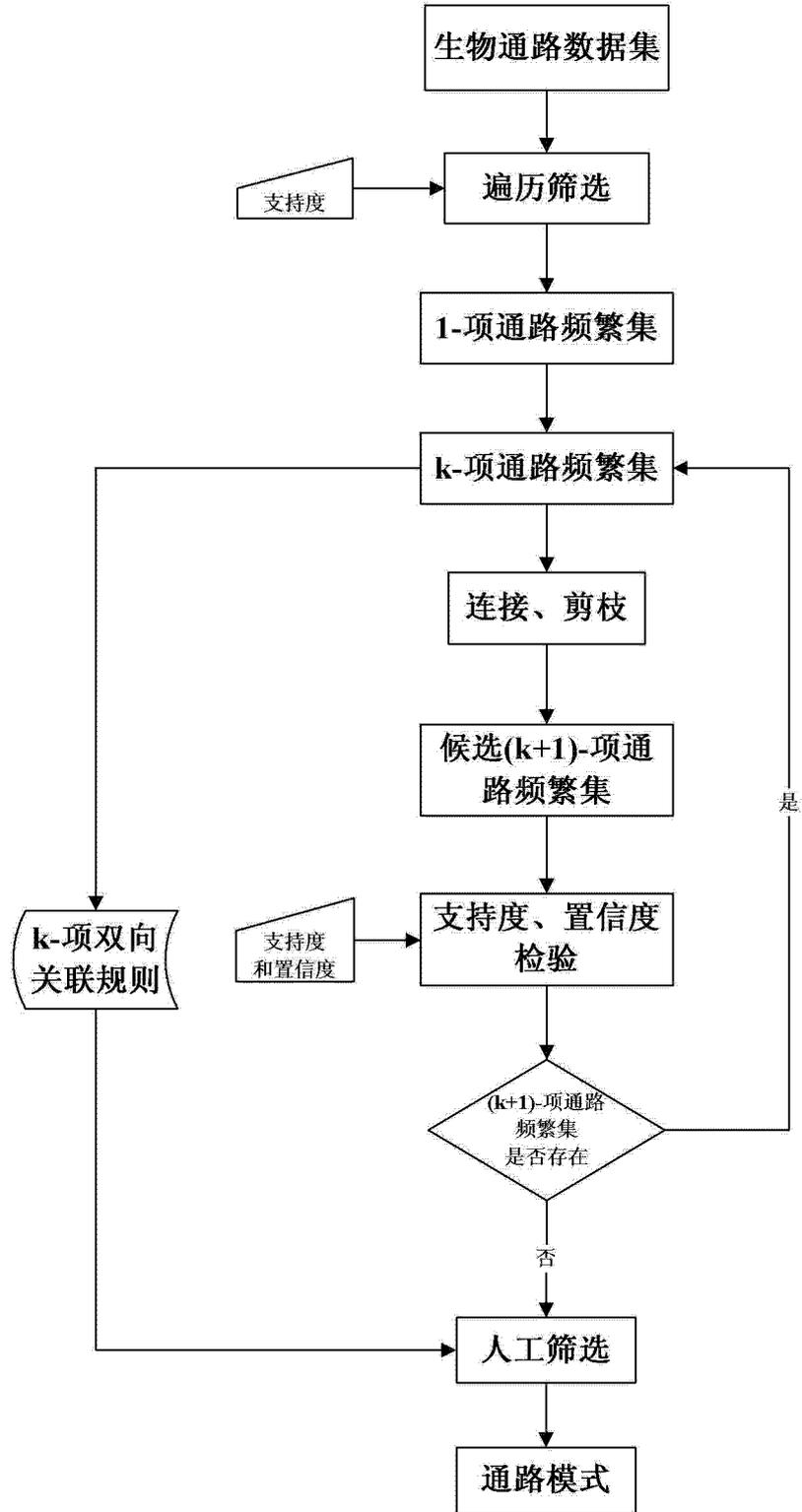


图 5