(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106295252 B
(45)授权公告日 2019.05.07

(21)申请号 201610687440.8

(22)申请日 2016.08.18

(65)同一申请的已公布的文献号
申请公布号 CN 106295252 A

(43)申请公布日 2017.01.04

(73)专利权人 杭州布理岚柏科技有限公司
地址 310000 浙江省杭州市拱墅区祥园路38号1幢东部四楼A402

(72)发明人 刘杨

(74)专利代理机构 杭州华鼎知识产权代理事务所(普通合伙) 33217
代理人 项军

(51)Int.Cl.
G16B 50/00(2019.01)
G06F 16/22(2019.01)
G06F 16/953(2019.01)
G06F 16/245(2019.01)
G06F 16/2455(2019.01)

(56)对比文件
CN 101266601 A,2008.09.17,
CN 105630813 A,2016.06.01,
CN 105589936 A,2016.05.18,
CN 1744080 A,2006.03.08,
CN 101539916 A,2009.09.23,
CN 101738196 A,2010.06.16,
CN 105740243 A,2016.07.06,
CN 101201847 A,2008.06.18,
CN 102043812 A,2011.05.04,
CN 104090890 A,2014.10.08,

审查员 何守兵

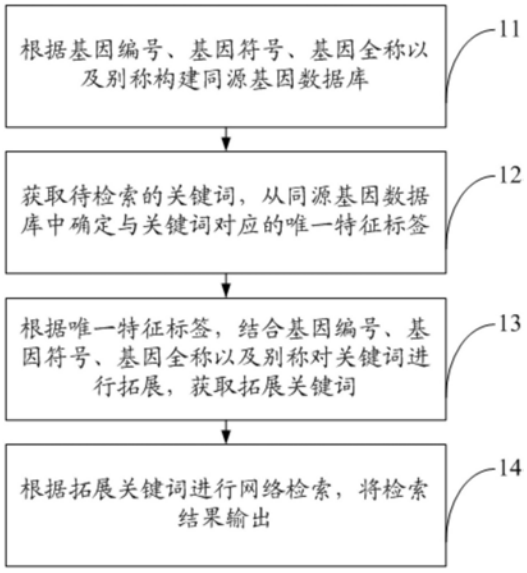
权利要求书2页 说明书4页 附图2页

(54)发明名称

用于基因产品的检索方法

(57)摘要

本发明提供了用于基因产品的检索方法,属于信息检索领域,包括构建同源基因数据库,获取待检索的关键词,确定与关键词对应的唯一特征标签,根据唯一特征标签,对关键词进行拓展,获取拓展关键词,根据拓展关键词进行网络检索。通过根据待检索的关键词获取唯一特征标签,基于唯一特征标签对关键词进行拓展处理,最终根据得到的拓展关键词进行全网检索,由于拓展关键词中包含了与待检索的关键词对应的多重限定,从而保证在互联网上能够搜索到与关键词关联性最强的资源,降低其他无关资源对搜索结果的干扰。



CN 106295252 B

1.用于基因产品的检索方法，其特征在于，所述检索方法包括：

根据基因编号、基因符号、基因全称以及别称构建同源基因数据库；

获取待检索的关键词，从同源基因数据库中确定与关键词对应的唯一特征标签；

根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词；

根据拓展关键词进行网络检索，将检索结果输出；

其中，获取拓展关键词还包括：

从基因子库中获取物种基因数据，结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因；

基于跨物种直接同源基因，以基因全称或基因编号相同为标准在基因子库中进行扩充匹配，得到直接同源基因关键词数据集，根据得到的直接同源基因关键词数据集建立非冗余数据库；

在非冗余数据库中选取与关键词匹配的拓展关键词。

2.根据权利要求1所述的用于基因产品的检索方法，其特征在于，所述检索方法，还包括：

构建包括基因文献、基因产品的检索数据库，在所述检索数据库设有与每个所述基因文献、每个所述基因产品对应的唯一特征标签。

3.根据权利要求2所述的用于基因产品的检索方法，其特征在于，所述检索方法，还包括：

在所述检索数据库中选取与所述唯一特征标签对应的、包括基因文献和/或基因产品的检索结果；

将所述检索结果输出。

4.根据权利要求1所述的用于基因产品的检索方法，其特征在于，所述根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词，包括：

根据唯一特征标签，确定与唯一特征标签对应的目的基因编号、目的基因符号、目的基因全称以及别称；

以关键词为基础，将所述目的基因编号、所述目的基因符号、所述目的基因全称以及别称按或的逻辑结构进行拓展，获取拓展关键词。

5.根据权利要求1所述的用于基因产品的检索方法，其特征在于，还包括：

所述唯一特征标签为字符串，在所述字符串中设有序列字节和验证字节。

6.根据权利要求1所述的用于基因产品的检索方法，其特征在于，在所述同源基因数据库中设有与基因编号、基因符号、基因全称以及别称对应的标签。

7.根据权利要求1或5所述的用于基因产品的检索方法，其特征在于，所述拓展关键词为至少包括基因编号、基因符号、基因全称以及别称在内的字符串。

8.根据权利要求1所述的用于基因产品的检索方法，其特征在于，所述结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因，包括：

从对比数据库中提取与物种基因数据对应的样本基因数据，基于样本基因数据对物种基因数据进行去重筛选，得到筛选后的跨物种直接同源基因。

9.根据权利要求1所述的用于基因产品的检索方法，其特征在于，在所述非冗余数据库中存储的直接同源基因关键词数据具有唯一性。

用于基因产品的检索方法

技术领域

[0001] 本发明属于信息检索领域，特别涉及用于基因产品的检索方法。

背景技术

[0002] 随着测序技术的发展，多物种基因组测序陆续完成，并且由于互联网技术的迅速发展，基于互联网进行基因、以及基因文献、基因产品等相关材料的搜索已经成为业内的趋势。

[0003] 迄今，美国国立卫生院基因数据库(NCBI)里收录基因数目已经超过一千三百万条。但由于命名规则的历史原因和同源基因的存在，每条基因除了具有基因编号(gene ID)之外，还可能有基因全称(gene full name)、基因符号(gene symbol)、别称(aliase，synonym)等业内名称，在收录基因文献、基因产品时不可能按统一的名称进行收录。导致当前基于单一基因名称关键词搜索查询特异性基因相关信息和产品时，查询效率低且查询结果易出现无关数据或遗漏数据等情况。这样给后期的搜索带来了巨大的困难。

发明内容

[0004] 为了解决现有技术中存在的缺点和不足，本发明提供了用于提高检索效率的用于基因产品的检索方法。

[0005] 为了达到上述技术目的，本发明提供了用于基因产品的检索方法，所述检索方法包括：

[0006] 根据基因编号、基因符号、基因全称以及别称构建同源基因数据库；

[0007] 获取待检索的关键词，从同源基因数据库中确定与关键词对应的唯一特征标签；

[0008] 根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词；

[0009] 根据拓展关键词进行网络检索，将检索结果输出。

[0010] 可选的，所述检索方法，还包括：

[0011] 构建包括基因文献、基因产品的检索数据库，在所述检索数据库设有与每个所述基因文献、每个所述基因产品对应的唯一特征标签。

[0012] 可选的，所述检索方法，还包括：

[0013] 在所述检索数据库中选取与所述唯一特征标签对应的、包括基因文献和/或基因产品的检索结果；

[0014] 将所述检索结果输出。

[0015] 可选的，所述根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词，包括：

[0016] 根据唯一特征标签，确定与唯一特征标签对应的目的基因编号、目的基因符号、目的基因全称以及别称；

[0017] 以关键词为基础，将所述目的基因编号、所述目的基因符号、所述目的基因全称以

及别称按或的逻辑结构进行拓展，获取拓展关键词。

[0018] 可选的，还包括：

[0019] 所述唯一特征标签为字符串，在所述字符串中设有序列字节和验证字节。

[0020] 可选的，在所述同源基因数据库中设有与基因编号、基因符号、基因全称以及别称对应的标签。

[0021] 可选的，所述拓展关键词为至少包括基因编号、基因符号、基因全称以及别称在内的字符串。

[0022] 可选的，还包括：

[0023] 从基因子库中获取物种基因数据，结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因；

[0024] 基于跨物种直接同源基因，以基因全称或基因编号相同为标准在基因子库中进行扩充匹配，得到直接同源基因关键词数据集，根据得到的直接同源基因关键词数据集建立非冗余数据库；

[0025] 在非冗余数据库中选取与关键词匹配的拓展关键词。

[0026] 可选的，所述结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因，包括：

[0027] 从对比数据库中提取与物种基因数据对应的样本基因数据，基于样本基因数据对物种基因数据进行去重筛选，得到筛选后的跨物种直接同源基因。

[0028] 可选的，在所述非冗余数据库中存储的直接同源基因关键词数据具有唯一性。

[0029] 本发明提供的技术方案带来的有益效果是：

[0030] 通过根据待检索的关键词获取唯一特征标签，基于唯一特征标签对关键词进行拓展处理，最终根据得到的拓展关键词进行全网检索，由于拓展关键词中包含了与待检索的关键词对应的多重限定，从而保证在互联网上能够搜索到与关键词关联性最强的资源，降低其他无关资源对搜索结果的干扰。

附图说明

[0031] 为了更清楚地说明本发明的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0032] 图1是本发明提供的用于基因产品的检索方法的流程示意图；

[0033] 图2是本发明提供的拓展关键词的获取方式的流程示意图。

具体实施方式

[0034] 为使本发明的结构和优点更加清楚，下面将结合附图对本发明的结构作进一步地描述。

[0035] 实施例一

[0036] 本发明提供了用于基因产品的检索方法，如图1所示，所述检索方法包括：

[0037] 11、根据基因编号、基因符号、基因全称以及别称构建同源基因数据库。

[0038] 12、获取待检索的关键词，从同源基因数据库中确定与关键词对应的唯一特征标

签。

[0039] 13、根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词。

[0040] 14、根据拓展关键词进行网络检索，将检索结果输出。

[0041] 在实施中，为了能够根据关键词获取尽可能丰富、且与基因相关的检索结果，本发明提供了用于基因产品的检索方法，在本检索方法中，首先构建同源基因数据库，在同源基因数据库中包括大量的基因编号，基因符号、基因全称以及别称。以便于在后续步骤中，能够根据关键词的具体内容，在同源基因数据库中确定与关键词关联的基因编号、基因符号、基因全称以及别称。接着根据获取到待检索的关键词，从前一步中构建的同源基因数据库中确定与关键词对应的唯一特征标签。再次根据唯一特征标签对应的基因编号等内容对关键词进行拓展处理，得到处理后的拓展关键词。最终根据拓展关键词进行全网检索，得到检索结果。

[0042] 在上述步骤中，之所以设置获取唯一特征标签的步骤，是为了将包含有基因编号、基因符号、基因全称以及别称的同源基因数据库中的资源对关键词进行拓展，对关键词进行准确的限定，从而保证在互联网上能够搜索到与关键词关联性最强的资源，降低其他无关资源对搜索结果的干扰。

[0043] 值得注意的是，在步骤12中确定唯一特征标签时，在同源基因数据库中存在的关键词组可能会与关键词一一对应，这样，可以对应的关键词组直接确定唯一特征标签；如果在同源基因数据库中，针对待检索的关键词，存在一个以上的关键词组与之对应，这样需要从多个关键词组中选取更为接近的关键词组，进而确定与选出的关键词组对应的唯一特征标签，从而便于根据确定的唯一特征标签完成后续处理步骤。

[0044] 步骤13中获取拓展关键词的步骤具体包括：

[0045] 根据唯一特征标签，确定与唯一特征标签对应的目的基因编号、目的基因符号、目的基因全称以及别称；

[0046] 以关键词为基础，将所述目的基因编号、所述目的基因符号、所述目的基因全称以及别称按或的逻辑结构进行拓展，获取拓展关键词。

[0047] 其中的唯一特征标签为字符串，在所述字符串中设有序列字节和验证字节。以便于在确定唯一特征标签后，通过验证字节对计算出的序列字节进行验证。此外，为了在同源基因数据库中设有与基因编号、基因符号、基因全称以及别称对应的标签。获取到的拓展关键词为至少包括基因编号、基因符号、基因全称以及别称在内的字符串。

[0048] 具体的，所述检索方法，还包括：构建包括基因文献、基因产品的检索数据库，在所述检索数据库设有与每个所述基因文献、每个所述基因产品对应的唯一特征标签。

[0049] 在实施中，除了上述方法中提出的对关键词进行拓展，基于拓展关键词进行全网检索意外，还包括构建检索数据库，进而根据唯一特征标签在检索数据库中进行检索，获取检索后的结果。

[0050] 本步骤中所谓的检索数据库，是包含基因文献、基因产品在内的数据库，事先将可能作为检索结果的基因文献以及基因产品构建数据库，并且为检索数据库中与每个基因对应的内容赋予唯一特征标签。这样在根据关键词确定唯一特征标签后，可以在所述检索数据库中选取与所述唯一特征标签对应的、包括基因文献和/或基因产品的检索结果，进而将

所述检索结果输出，根据唯一特征标签在检索数据库中选出与关键词对应的检索内容，相对于通过互联网进行全网检索，能够实现更为迅速且准确的检索。

[0051] 在第一种检索方式中，提出了根据拓展关键词进行全网检索的方式，下面提出另一种关于拓展关键词的获取方式，具体过程为如图2所示。

[0052] 21、从基因子库中获取物种基因数据，结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因。

[0053] 22、基于跨物种直接同源基因，以基因全称或基因编号相同为标准在基因子库中进行扩充匹配，得到直接同源基因关键词数据集，根据得到的直接同源基因关键词数据集建立非冗余数据库。

[0054] 23、在非冗余数据库中选取与关键词匹配的拓展关键词。

[0055] 在实施中，根据美国国家生物技术信息中心(National Center of Biotechnology Information,NCBI)的基因子库整理多物种的基因数据，结合HomoloGene数据库，筛选跨物种直接同源基因，以基因符号Symbol或全称full name相同为标准在基因子库中匹配扩充直接同源基因数据，最终产生直接同源基因关键词数据集，建立基因符号Symbol名称非冗余数据库，选取与关键词匹配的拓展关键词。

[0056] 步骤21中的结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因的具体方式为：从对比数据库中提取与物种基因数据对应的样本基因数据，基于样本基因数据对物种基因数据进行去重筛选，得到筛选后的跨物种直接同源基因。

[0057] 并且，在非冗余数据库中存储的直接同源基因关键词数据具有唯一性。

[0058] 本发明提供了用于基因产品的检索方法，包括构建同源基因数据库，获取待检索的关键词，确定与关键词对应的唯一特征标签，根据唯一特征标签，对关键词进行拓展，获取拓展关键词，根据拓展关键词进行网络检索。通过根据待检索的关键词获取唯一特征标签，基于唯一特征标签对关键词进行拓展处理，最终根据得到的拓展关键词进行全网检索，由于拓展关键词中包含了与待检索的关键词对应的多重限定，从而保证在互联网上能够搜索到与关键词关联性最强的资源，降低其他无关资源对搜索结果的干扰。

[0059] 上述实施例中的各个序号仅仅为了描述，不代表各部件的组装或使用过程中的先后顺序。

[0060] 以上所述仅为本发明的实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

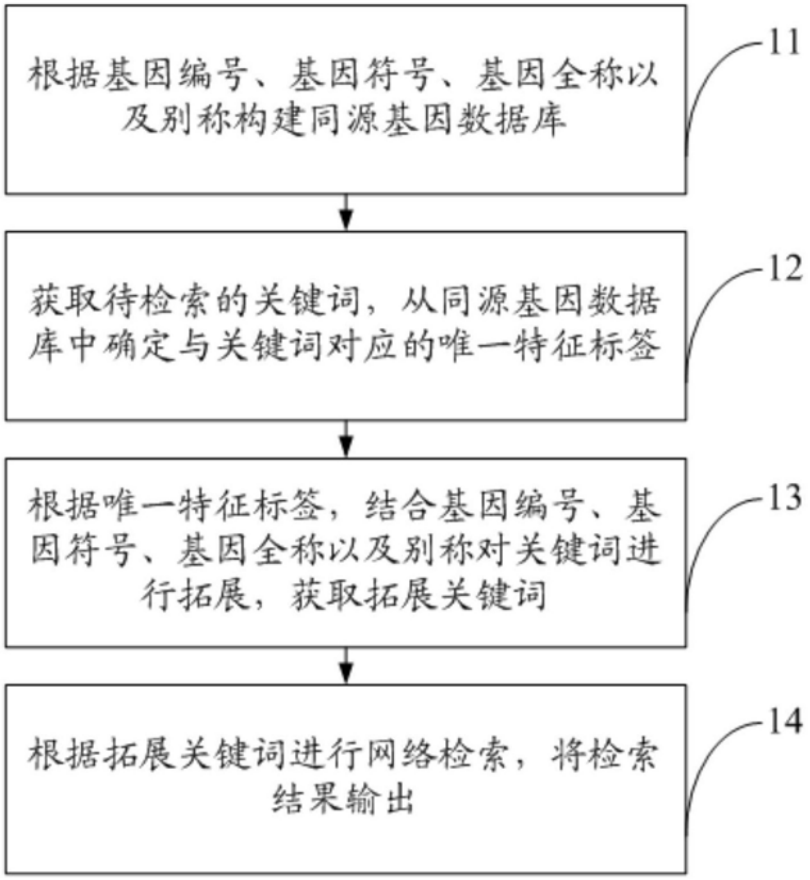

根据基因编号、基因符号、基因全称以及别称构建同源基因数据库
11
获取待检索的关键词，从同源基因数据库中确定与关键词对应的唯一特征标签
12
根据唯一特征标签，结合基因编号、基因符号、基因全称以及别称对关键词进行拓展，获取拓展关键词
13
根据拓展关键词进行网络检索，将检索结果输出
14

图1

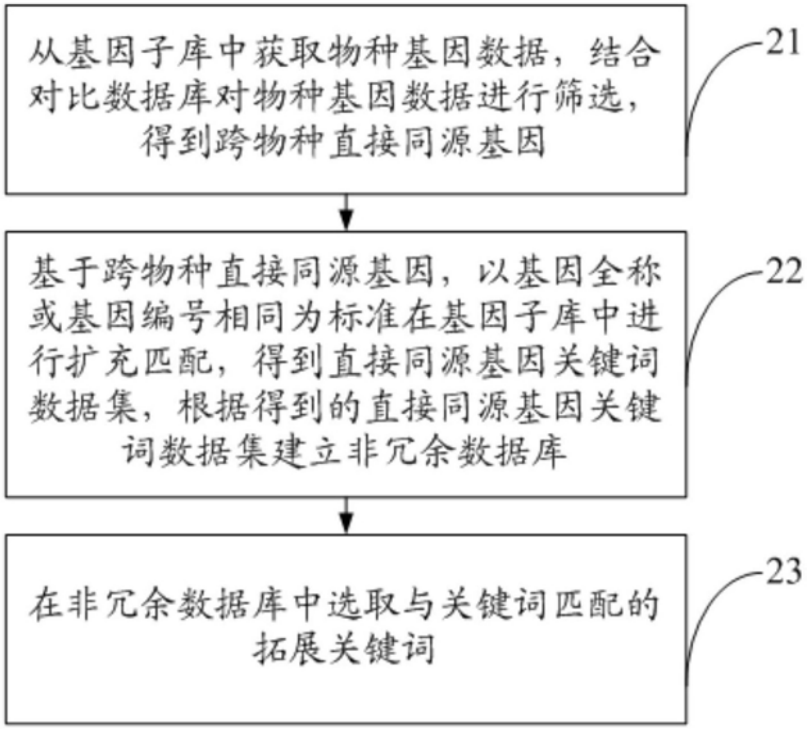

从基因子库中获取物种基因数据，结合对比数据库对物种基因数据进行筛选，得到跨物种直接同源基因
21
基于跨物种直接同源基因，以基因全称或基因编号相同为标准在基因子库中进行扩充匹配，得到直接同源基因关键词数据集，根据得到的直接同源基因关键词数据集建立非冗余数据库
22
在非冗余数据库中选取与关键词匹配的拓展关键词
23

图2