



(12)发明专利

(10)授权公告号 CN 106934253 B

(45)授权公告日 2020.07.17

(21)申请号 201710175969.6

G16H 50/50(2018.01)

(22)申请日 2012.06.11

G06N 7/06(2006.01)

(65)同一申请的已公布的文献号

G16B 5/00(2019.01)

申请公布号 CN 106934253 A

G06N 5/02(2006.01)

(43)申请公布日 2017.07.07

(30)优先权数据

11195417.8 2011.12.22 EP

61/495,824 2011.06.10 US

61/525,700 2011.08.19 US

(62)分案原申请数据

201280028435.6 2012.06.11

(73)专利权人 菲利普莫里斯生产公司

地址 瑞士纳沙泰尔

(72)发明人 J·霍恩格 F·马丁 M·派奇

A·塞沃尔

(74)专利代理机构 中国国际贸易促进委员会专

利商标事务所 11038

代理人 宋岩

(51)Int.Cl.

G16B 40/00(2019.01)

G06N 7/00(2006.01)

(56)对比文件

Selventa.com.Anonymous: "Reverse Causal Reasoning Method - White Paper".
《http://www.selventa.com/attachments/white_papers/reverse-causal-reasoning.pdf》.2011,

Selventa.com.Anonymous: "Reverse Causal Reasoning Method - White Paper".
《http://www.selventa.com/attachments/white_papers/reverse-causal-reasoning.pdf》.2011,

Congreso Monteiro.Patricia Gimenez: "Local Influence Analysis Based on the Perturbation Manifold in functional Measurement Error Models".《http://www.matematica.uns.edu.ar/XCongresomonteiro/actas.HTM》.2009,

审查员 涂丹辉

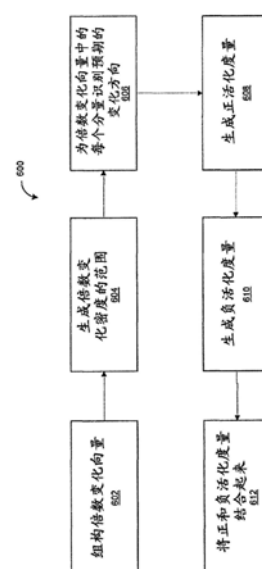
权利要求书4页 说明书32页 附图16页

(54)发明名称

用于基于网络的生物活动评价的系统和方法

(57)摘要

公开了用于基于网络的生物活动评价的系统和方法。本发明提供了用于量化响应于物剂的生物系统的扰动的系统和方法。处理和控制数据在处理器处接收,并且被用来计算在生物系统中的生物实体的计算因果网络模型的节点的活动测量结果。为模型生成扰动得分,从而表征处理物剂对生物系统的影响。



1. 一种用于量化响应于物剂的生物系统的扰动的计算机化的方法,该方法包括:
在至少一个处理器处接收对应于生物系统对于物剂的响应的一组处理数据,其中,生物系统包括多个生物实体,每个生物实体与至少一个其他的生物实体相互作用;
在至少一个处理器处接收对应于没有暴露给物剂的生物系统的一组对照数据;
在至少一个处理器处提供网络模型,该网络模型表示生物系统并且包括:
表示生物实体的节点,
表示生物实体之间的关系的边,和
所述节点的表示对照数据和处理数据之间的预期变化方向的方向值;
其中,所述预期变化方向基于所接收的对照数据和处理数据而被调节;
对于所述节点,基于对照数据和处理数据之间的差异来获得表示倍数变化值的活动测量结果;及
至少部分地基于方向值和活动测量结果,为所述网络模型生成表示响应于物剂的生物系统的扰动的得分。
2. 根据权利要求1所述的方法,其中,每个相应节点的相应倍数变化值包括由相应节点表示的生物实体的对照数据和处理数据之间的差异的对数。
3. 根据权利要求1所述的方法,还包括:至少部分地基于倍数变化表示违反零倍数变化的基础原假设的概率来确定权重值,其中至少部分地基于权重值来生成得分。
4. 根据权利要求1所述的方法,其中,生成得分包括使用活动测量结果、权重值和方向值的线性或非线性组合,以获得一组加权活动测量结果。
5. 根据权利要求4所述的方法,其中,生成得分还包括:
对加权活动测量结果求和以获得总和;及
通过基于在网络模型中表示的生物实体的数量的因子将总和标准化。
6. 根据权利要求5所述的方法,其中,正的得分指示由机制假设描述的过程的上调,为零的得分指示该过程沿着机制假设的方向是不变的,而负的得分指示该过程被下调。
7. 根据权利要求1所述的方法,其中,生成得分包括:
获得表示活动测量结果与方向值一致的程度的正活化度量,及
获得表示活动测量结果与方向值不一致的程度的负活化度量。
8. 根据权利要求7所述的方法,其中,得分是由网络模型表示的一个或多个生物机制在给定活动测量结果的情况下被激活的概率的量化。
9. 根据权利要求7所述的方法,其中:
方向值包括正方向值,
将网络模型中的与正方向值相关联的生物实体识别为生物实体的第一子集,及
将网络模型中的与负方向值相关联的生物实体识别为生物实体的第二子集。
10. 根据权利要求9所述的方法,其中,生成得分包括:
使用权重值对生物实体的第一子集的活动测量结果进行缩放,并且将缩放后的活动测量结果与方向值结合,以获得第一组加权活动测量结果;及
使用权重值对生物实体的第二子集的活动测量结果进行缩放,并且将缩放后的活动测量结果与方向值结合,以获得第二组加权活动测量结果。
11. 根据权利要求10所述的方法,其中:

至少部分地基于第一组加权活动测量结果的总和来生成正活化度量,及
至少部分地基于第二组加权活动测量结果的总和来生成负活化度量。

12. 根据权利要求1所述的方法,其中,方向值包括正方向值和负方向值,并且至少部分地基于具有正方向值的节点的第一总和及具有负方向值的节点的第二总和来生成得分。

13. 根据权利要求12所述的方法,其中,具有正方向值的节点对应于与响应于第二处理相比响应于第一处理活动被预期为增大的生物实体,具有负方向值的节点对应于与响应于第二处理相比响应于第一处理活动被预期为减小的生物实体。

14. 根据权利要求13所述的方法,还包括:对于具有正方向值的节点,

以递增的顺序对活动测量结果进行排序以获得第一组有序活动测量结果;

计算第一组有序活动测量结果中的相邻活动测量结果之间的差异,其中至少部分地基于这些差异以及正方向值来生成第一总和,以及

对于具有负方向值的节点,

以递增的顺序对活动测量结果进行排序以获得第二组有序活动测量结果;

计算第二组有序活动测量结果中的相邻活动测量结果之间的差异,其中至少部分地基于这些差异以及负方向值来生成第二总和,

其中:

第一总和对与第一组有序活动测量结果中的相应活动测量结果乘以如下差异对应的值进行求和,该差异是相应活动测量结果与相邻于相应活动测量结果的另一活动测量结果之间的差异,及

第二总和对与第二组有序活动测量结果中的相应活动测量结果乘以如下差异对应的值进行求和,该差异是相应活动测量结果与相邻于相应活动测量结果的另一活动测量结果之间的差异。

15. 根据权利要求14所述的方法,其中,第一组有序活动测量结果中的第一活动测量结果低于第一组有序活动测量结果中的第二活动测量结果,并且得分被生成以使得与第一活动测量结果相比第二活动测量结果对于得分影响更大。

16. 根据权利要求1所述的方法,还包括:生成得分的置信区间。

17. 根据权利要求16所述的方法,其中,生成置信区间包括确定应用到活动测量结果的一组系数是否是活动测量结果的函数。

18. 根据权利要求17所述的方法,其中,生成置信区间还包括基于活动测量结果的方差来评价得分的方差。

19. 一种用于量化响应于物剂的生物系统的扰动的计算机化的系统,该计算机化的系统包括被配置为执行以下操作的处理器:

接收对应于生物系统对于物剂的响应的一组处理数据,其中,生物系统包括多个生物实体,每个生物实体与至少一个其他的生物实体相互作用;

接收对应于没有暴露给物剂的生物系统的一组对照数据;

提供网络模型,该网络模型表示生物系统并且包括:

表示生物实体的节点,

表示生物实体之间的关系的边,和

所述节点的表示对照数据和处理数据之间的预期变化方向的方向值;

其中,所述预期变化方向基于所接收的对照数据和处理数据而被调节;

对于所述节点,基于对照数据和处理数据之间的差异来获得表示倍数变化值的活动测量结果;及

至少部分地基于方向值和活动测量结果,为所述网络模型生成表示响应于物剂的生物系统的扰动的得分。

20. 根据权利要求19所述的系统,其中,每个相应节点的相应倍数变化值包括由相应节点表示的生物实体的对照数据和处理数据之间的差异的对数。

21. 根据权利要求19所述的系统,其中,处理器还被配置为至少部分地基于倍数变化表示违反零倍数变化的基础原假设的概率来确定权重值,其中至少部分地基于权重值来生成得分。

22. 根据权利要求19所述的系统,其中,处理器通过以下操作生成得分:使用活动测量结果、权重值和方向值的线性或非线性组合,以获得一组加权活动测量结果。

23. 根据权利要求22所述的系统,其中,处理器还通过以下操作生成得分:

对加权活动测量结果求和以获得总和;及

通过基于在网络模型中表示的生物实体的数量的因子将总和标准化。

24. 根据权利要求23所述的系统,其中,正的得分指示由机制假设描述的过程的上调,为零的得分指示该过程沿着机制假设的方向是不变的,而负的得分指示该过程被下调。

25. 根据权利要求19所述的系统,其中,处理器通过以下操作生成得分:

获得表示活动测量结果与方向值一致的程度的正活化度量,及

获得表示活动测量结果与方向值不一致的程度的负活化度量。

26. 根据权利要求25所述的系统,其中,得分是由网络模型表示的一个或多个生物机制在给定活动测量结果的情况下被激活的概率的量化。

27. 根据权利要求25所述的系统,其中:

方向值包括正方向值,

将网络模型中的与正方向值相关联的生物实体识别为生物实体的第一子集,及

将网络模型中的与负方向值相关联的生物实体识别为生物实体的第二子集。

28. 根据权利要求27所述的系统,其中,处理器通过以下操作生成得分:

使用权重值对生物实体的第一子集的活动测量结果进行缩放,并且将缩放后的活动测量结果与方向值结合,以获得第一组加权活动测量结果;及

使用权重值对生物实体的第二子集的活动测量结果进行缩放,并且将缩放后的活动测量结果与方向值结合,以获得第二组加权活动测量结果。

29. 根据权利要求28所述的系统,其中:

至少部分地基于第一组加权活动测量结果的总和来生成正活化度量,及

至少部分地基于第二组加权活动测量结果的总和来生成负活化度量。

30. 根据权利要求19所述的系统,其中,方向值包括正方向值和负方向值,并且至少部分地基于具有正方向值的节点的第一总和及具有负方向值的节点的第二总和来生成得分。

31. 根据权利要求30所述的系统,其中,具有正方向值的节点对应于与响应于第二处理相比响应于第一处理活动被预期为增大的生物实体,具有负方向值的节点对应于与响应于第二处理相比响应于第一处理活动被预期为减小的生物实体。

32. 根据权利要求31所述的系统, 其中, 处理器还被配置为执行以下操作: 对于具有正方向值的节点,

以递增的顺序对活动测量结果进行排序以获得第一组有序活动测量结果;

计算第一组有序活动测量结果中的相邻活动测量结果之间的差异, 其中至少部分地基于这些差异以及正方向值来生成第一总和, 以及

对于具有负方向值的节点,

以递增的顺序对活动测量结果进行排序以获得第二组有序活动测量结果;

计算第二组有序活动测量结果中的相邻活动测量结果之间的差异, 其中至少部分地基于这些差异以及负方向值来生成第二总和,

其中:

第一总和对与第一组有序活动测量结果中的相应活动测量结果乘以如下差异对应的值进行求和, 该差异是相应活动测量结果与相邻于相应活动测量结果的另一活动测量结果之间的差异, 及

第二总和对与第二组有序活动测量结果中的相应活动测量结果乘以如下差异对应的值进行求和, 该差异是相应活动测量结果与相邻于相应活动测量结果的另一活动测量结果之间的差异。

33. 根据权利要求32所述的系统, 其中, 第一组有序活动测量结果中的第一活动测量结果低于第一组有序活动测量结果中的第二活动测量结果, 并且得分被生成以使得与第一活动测量结果相比第二活动测量结果对于得分影响更大。

34. 根据权利要求19所述的系统, 其中, 处理器还被配置为生成得分的置信区间。

35. 根据权利要求34所述的系统, 其中, 处理器通过确定应用到活动测量结果的一组系数是否是活动测量结果的函数来生成置信区间。

36. 根据权利要求35所述的系统, 其中, 处理器还通过基于活动测量结果的方差来评价得分的方差来生成置信区间。

用于基于网络的生物活动评价的系统和方法

[0001] 本申请是申请号为201280028435.6,申请日为2012年6月11日,题为“用于基于网络的生物活动评价的系统和方法”的中国发明专利申请的分案申请。

背景技术

[0002] 人体经常受到暴露于在长期来看能够会造成严重的健康风险的潜在有害物剂(agent)而扰动。暴露于这些物剂下能够危害人体内部的生物机制的正常机能。为了理解和量化这些扰动对人体的影响,研究人员研究生物系统对暴露于物剂作出响应的机制。有些团组已经在活体动物试验方法中广泛地使用。但是,动物试验方法并非总是足够的,因为关于它们的可靠性和关联性存在着疑问。在不同动物的生理方面存在着众多差异。因此,不同的物种对于暴露于试剂下可能会作出不同的响应。因此,关于从动物试验中获得的响应是否可以外推到人类生物学中是存在疑问的。其他方法包括通过对人类志愿者的临床研究来评估风险。但是这些风险评估是以后验方式执行的,并且因为疾病可能要花费数十年才表现出来,所以这些评估可能不足以阐明将有害物质与疾病联系起来的机制。另外一些方法包括体外实验。尽管,基于体外细胞和组织的方法作为对它们的基于动物的相应方法的完全或部分替代方法已经获得普遍接受,但是这些方法价值有限。因为体外方法集中于细胞和组织机制的具体方面;它们并不总是会考虑在整个生物系统中发生的复杂的相互作用。

[0003] 在过去十年中,与传统的剂量相关的疗效和毒性测定结合的核酸、蛋白质及代谢物水平的高通量测量已经作为用于阐明许多生物过程的作用机制的方法而出现。研究人员已经尝试将来自这些全异测量的信息与来自文献的关于生物路径的知识结合以构建(assembly)有意义的生物模型。为了这个目的,研究人员已经开始使用能够挖掘大量数据的数学及计算技术(例如,聚类和统计方法)来识别可能的生物作用机制。

[0004] 之前的工作同样探讨了揭示由对生物过程的一种或多种扰动引起的基因表达变化的特性特征的重要性,以及关于该特征存在于作为对该过程的具体活动幅度的测量结果的附加数据集内的后续评分。在这方面的大部分工作已经涉及对与疾病表型关联的特征的识别和评分。这些表型衍生的特征提供显著的分类能力,但是在单个具体扰动与特征之间缺少机理或因果关系。因此,这些特征可以表示通过通常未知的机制导致相同的疾病表型的或由其引起的多种不同的未知扰动。

[0005] 一个挑战在于理解在生物系统中的各种个体生物实体的活动如何允许激活或抑制不同的生物机制。因为个体实体(例如,基因)会涉及多种生物过程(例如,发炎和细胞增殖),对基因活动的测量并不足以识别触发该活动的基础生物过程。当前技术中没有一种被应用于识别作为微尺度下的生物实体的活动的原因的基础机制,也不响应于潜在有害物剂和实验条件而提供对这些实体于其内起作用的不同生物机制的激活的定量评估。因此,需要用于基于生物机制来分析系统级的生物数据并且量化生物系统随着系统响应于物剂或环境变化的变化的改进的系统和方法。

发明内容

[0006] 在一个方面,本文所描述的系统和方法涉及用于量化生物系统响应于物剂的扰动的计算机化的方法及一个或多个计算机处理器。

[0007] 在一个方面,计算机化的方法包括:在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中该生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用;在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据集;在第三处理器处提供表示生物系统的计算因果网络模型,并且该计算因果网络模型包括或包含:表示生物实体的节点、表示生物实体之间的关系的边以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值;以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果;以第五处理器来计算节点的权重值,其中至少一个权重值不同于至少另一个权重值;并且以第六处理器来为计算模型生成表示生物系统对于物剂的扰动的得分,其中该得分基于方向值、权重值和活动测量结果。该生物系统可以由至少一个机制假设表示。该生物系统可以由多个计算因果网络模型或者包括多个机制假设的至少一个计算因果网络模型表示。方法还可以包括基于在各自的计算模型中的可测量节点的数量来标准化得分。

[0008] 权重值可以表示在处理数据集和控制数据集中的至少一个中的置信度。权重值可以包括或包含局部错误非发现率。方法还可以包括以第七处理器来计算节点的活动测量结果在模型上的近似分布或者在模型中的机制假设;以第八处理器来针对该近似分布计算活动测量结果的期望值;并且以第九处理器为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于期望值。近似分布可以基于活动测量结果。在某些实现方式中,计算期望值可以包括执行矩形近似。方法还可以包括以第十处理器来基于活动测量结果计算正活化度量(positive activation metric)和负活化度量,正和负活化度量分别表示关于模型的在活动测量结果与方向值之间的一致性和不一致性;并且以第十一处理器来为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于正和负活化得分。正活化度量、负活化度量或两者都可以基于局部错误非发现率。活动测量结果可以是倍数变化值,并且每个节点的倍数变化值包括或包含由各自节点表示的生物实体的在处理数据与控制数据之间的差异的对数。生物系统的子集可以包括或包含细胞增殖机制、细胞应激机制、细胞发炎机制和DNA修复机制中的至少一个。物剂可以包括或包含通过加热烟草生成的烟雾剂、通过燃烧烟草生成的烟雾剂、烟草烟雾或香烟烟雾中的至少一个。物剂可以包括或包含非均质物质,包括不存在于生物系统内的或者不是从其中得出的分子或实体。物剂可以包括或包含毒素、治疗化合物、刺激物、弛缓剂、天然产品、制造产品和食物材料。处理数据集可以包括或包含多个处理数据集,使得每个可测量节点包括或包含由第一概率分布定义的多个倍数变化值以及由第二概率分布定义的多个权重值。处理数据集可以包括或包含多个处理数据集,使得每个可测量节点包括或包含多个倍数变化值和相应的权重值。生成得分的步骤可以包括活动测量结果、权重值和方向值的线性或非线性组合;以及通过比例因子对该组合进行的标准化。该组合可以是算术组合,并且比例因子是从其接收测量数据的生物实体的数量的平方根。得分可以通过几何扰动指数评分技术(index scoring technique)、概率扰动指数评分技术或预期扰动指数评分技术来生成。方法还可以包括基于参数或非参数计算引导技术(bootstrapping technique)来确定得分的置信区

间。

[0009] 在另一个方面,还描述了用于量化响应于物剂的生物系统的扰动的计算机系统。该系统包括至少一个处理器,该至少一个处理器被配置用于或适合于:接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用;接收与未暴露于物剂的生物系统对应的控制数据集;提供表示生物系统的计算因果网络模型,并且该计算因果网络模型包括或包含:表示生物实体的节点,表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值;计算节点的表示在处理数据与控制数据之间的差异的活动测量结果;计算节点的权重值,其中至少一个权重值不同于至少另一个权重值;并且为计算模型生成表示生物系统对于物剂的扰动的得分,其中该得分基于方向值、权重值和活动测量结果。生物系统可以由至少一个机制假设表示。生物系统可以由多个计算因果网络模型或者包括多个机制假设的至少一个计算因果网络模型表示。计算机系统还可以包括基于在各自的计算模型中的可评分节点的数量来标准化得分。权重值可以表示在处理数据集和控制数据集中的至少一项中的置信度。权重值可以包括或包含局部错误非发现率。在某些实现方式中,计算机系统还包括计算节点的活动测量结果在模型上的近似分布或者在模型中的机制假设;以第八处理器来针对该近似分布计算活动测量结果的期望值;并且为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于期望值。近似分布可以基于活动测量结果。在计算机系统的某些实现方式中,它还可以包括计算期望值,包括执行矩形近似。系统还可以包括基于活动测量结果来计算正活化度量和负活化度量,正和负活化度量分别表示关于模型在活动测量结果与方向值之间的一致性和不一致性;并且为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于正和负活化得分。正活化度量、负活化度量或两者都可以基于局部错误非发现率。活动测量结果可以是倍数变化值,并且每个节点的倍数变化值可以包括或包含由各自节点表示的生物实体的在处理数据与控制数据之间的差异的对数。生物系统的子集可以包括或包含细胞增殖机制、细胞应激机制、细胞发炎机制和DNA修复机制中的至少一个。物剂可以包括或包含通过加热烟草生成的烟雾剂、通过燃烧烟草生成的烟雾剂、烟草烟雾或香烟烟雾中的至少一个。物剂可以包括或包含非均质物质,包括不存在于生物系统内的或者不是从其中得出的分子或实体。物剂可以包括或包含毒素、治疗化合物、刺激物、弛缓剂、天然产品、制造产品和食物材料。处理数据集可以包括或包含多个处理数据集,使得每个可测量节点包括或包含由第一概率分布定义的多个倍数变化值以及由第二概率分布定义的多个权重值。处理数据集可以包括或包含多个处理数据集,使得每个可测量节点包括或包含多个倍数变化值和相应的权重值。生成得分的步骤可以包括活动测量结果、权重值和方向值的线性或非线性组合;以及通过比例因子对该组合进行的标准化。该组合可以是算术组合,并且比例因子是从其接收测量数据的生物实体的数量的平方根。得分可以通过几何扰动指数评分技术、概率扰动指数评分技术或预期扰动指数评分技术来生成。系统还可以包括基于参数或非参数计算引导技术来确定得分的置信区间。

[0010] 在某些方面,计算机化的方法可以包括在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括多个生物实体,每个生物实体与至少另一个生物实体相互作用,并且在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据

集。计算机化的方法可以包括在第三处理器处提供表示生物系统的计算因果网络模型。计算模型可以包括或包含表示生物实体的节点,表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值。计算机化的方法还可以包括以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果,以及以第五处理器来计算节点的权重值,其中至少一个权重值不同于至少另一个权重值。计算机化的方法还可以包括以第六处理器来为计算模型生成表示生物系统对于物剂的扰动的得分,其中该得分基于方向值、权重值和活动测量结果。在某些实现方式中,计算机化的方法还包括基于在各自的计算模型中的节点数量来标准化得分。在某些实现方式中,第一至第六处理器中的每一个都包括于或包含于单个处理器或单个计算装置之内。在其他实现方式中,第一至第六处理器中的一个或多个跨多个处理器或计算装置分布。

[0011] 在某些实现方式中,计算因果网络模型包括或包含存在于表示可能的原因的节点与表示测得量的节点之间的因果关系集。在这样的实现方式中,活动测量结果可以包括倍数变化。倍数变化可以是用于描述在控制数据与处理数据之间的从初始值变为最终值的节点测量变化的数字。倍数变化数可以表示在控制条件与处理条件之间的生物实体的活动的倍数变化的对数。每个节点的活动测量结果都可以包括或包含由各自节点表示的生物实体的在处理数据与控制数据之间的差异的对数。在这样的实现方式中,权重值可以表示要赋予节点的倍数变化值的权重。权重值可以表示测量节点关于所感兴趣的特征或结果的已知生物学重要性(例如,在癌症研究中的已知致癌物质)。权重值可以表示在扰动数据集和控制数据集中的至少一项内的置信度。更特别地,权重值可以包括或包含局部错误非发现率。在这样的实现方式中,计算机化的方法可以通过将活动测量结果乘以权重值和方向值并对节点进行求和而生成计算模型的得分。在某些实现方式中,计算机化的方法包括或包含以处理器生成每个所生成的得分的置信区间。置信区间可以包括对所生成的得分的分布进行近似。

[0012] 在另一个方面,本文所描述的系统和方法涉及用于量化响应于物剂的生物系统的扰动的计算机化的方法。计算机化的方法可以包括在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用,并且在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据集。计算机化的方法可以包括在第三处理器处提供表示生物系统的计算因果网络模型。计算模型可以包括或包含表示生物实体的节点、表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值。计算机化的方法还可以包括以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果,以及以第五处理器来计算活动测量结果在节点上的近似分布。计算机化的方法还可以包括或包含以第六处理器来计算近似分布的期望值。计算机化的方法还可以包括以第七处理器来为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于期望值。在某些实现方式中,第一至第七处理器每个都包括于或包含于单个处理器或单个计算装置之内。在其他实现方式中,第一至第七处理器中的一个或多个跨多个处理器或计算装置分布。

[0013] 在某些实现方式中,计算因果网络模型包括或包含存在于表示可能的原因的节点与表示测得量的节点之间的因果关系集。在这样的实现方式中,活动测量结果可以包括或

包含倍数变化。倍数变化可以是用于描述在控制数据与处理数据之间的从初始值变为最终值的节点测量变化的数字。倍数变化数可以表示在控制条件与处理条件之间的生物实体的活动的倍数变化的对数。计算机化的方法可以包括或包含以处理器生成倍数变化密度的范围,该范围可以表示倍数变化值在处理条件下能够于生物系统中采用的值的集合的近似。处理器可以生成近似的倍数变化密度,该倍数变化密度可以包括或包含倍数变化值的近似概率分布。在这样的实现方式中,计算机化的方法还包括或包含计算近似的倍数变化密度的近似期望值。计算机化的方法可以基于所算得的期望值来生成计算模型的得分。

[0014] 在某些实现方式中,近似分布可以大体上基于活动测量结果。除此以外且可任选地,期望值可以包括矩形近似。在某些实现方式中,计算机化的方法包括或包含以处理器生成每个所生成的得分的置信区间。生成置信区间可以包括执行参数引导技术。

[0015] 在又一方面,本文所描述的系统和方法涉及用于量化响应于物剂的生物系统的扰动的计算机化的方法。计算机化的方法可以包括在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用,并且在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据集。计算机化的方法可以包括在第三处理器处提供表示生物系统的计算因果网络模型。计算模型可以包括或包含表示生物实体的节点、表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值。计算机化的方法还可以包括以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果,并且以第五处理器来基于活动测量结果计算正活化得分和负活化得分,正和负活化得分分别表示在活动测量结果与方向值之间的一致性和不一致性。计算机化的方法还可以包括以第六处理器来为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于正和负活化得分。在某些实现方式中,第一至第六处理器中的每一个都包括于或包含于单个处理器或单个计算装置之内。在其他实现方式中,第一至第六处理器中的一个或多个跨多个处理器或计算装置分布。

[0016] 在某些实现方式中,计算因果网络模型包括或包含存在于表示可能的原因的节点与表示测得量的节点之间的因果关系集。在这样的实现方式中,活动测量结果可以包括或包含倍数变化。倍数变化可以是用于描述在控制数据与处理数据之间的从初始值变为最终值的节点测量变化的数。倍数变化数可以表示在控制条件与处理条件之间的生物实体的活动的倍数变化的对数。计算机化的方法可以包括或包含以处理器生成倍数变化密度的范围,该范围可以表示在处理条件下倍数变化值能够于生物系统中采用的值的集合的近似。计算机化的方法可以包括以处理器来基于倍数变化值和方向值计算正活化得分。正和负活化得分可以指示所观察到的生物实体的激活/抑制与预期的变化方向是一致还是不一致的。在一个实例中,正活化得分是方向值与活动测量结果一致的概率。负活化得分可以是方向值与活动测量结果不一致的概率。计算机化的方法还可以包括或包含通过结合正和负活化得分来生成计算模型的得分。在某些实现方式中,该得分基于局部错误非发现率。

[0017] 在某些实现方式中,生物系统的子集包括或包含细胞增殖机制、细胞应激机制、细胞发炎机制和DNA修复机制中的至少一个。物剂可以包括或包含通过加热烟草生成的烟雾剂、通过燃烧烟草生成的烟雾剂、烟草烟雾或香烟烟雾中的至少一个。物剂可以包括镉,汞,铬,烟碱,烟草特有的亚硝胺及它们的代谢物(4-甲基亚硝胺基-1-(3-吡啶基)-1-丁酮4

(NNK)、N'-亚硝基去甲基尼古丁(NNN)、N-亚硝基新烟草碱(NAT)、N-亚硝基假木贼碱(NAB)和4-(甲基亚硝氨基)-1-(3-吡啶基)-1-丁醇(NNAL))。在某些实现方式中,物剂包括或包含用于烟碱替代疗法(replacement therapy)的产物。物剂可以包括或包含非均质物质,包括不存在于生物系统内的或者不是从其中得出的分子或实体。物剂还可以包括或包含毒素、治疗化合物、刺激物、弛缓剂、天然产品、制造产品和食物材料。在某些实现方式中,处理数据集包括或包含与生物网络模型的某些节点对应的多个处理数据集,其中每个这样的节点对应于由第一概率分布定义的多个倍数变化值以及由第二概率分布定义的多个权重值。

[0018] 在又一方面,本文所描述的系统和方法涉及用于量化响应于物剂的生物系统的扰动的计算机化的方法以及一个或多个计算机处理器。所述计算机化的方法可以包括在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用,并且在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据集。计算机化的方法可以包括在第三处理器处提供表示生物系统的计算因果网络模型。计算模型可以包括或包含表示生物实体的节点、表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的变化的预期方向的方向值。计算机化的方法还可以包括以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果。计算机化的方法还可以包括以第五处理器来为计算模型生成表示生物系统对于物剂的扰动的得分,其中该得分基于方向值和活动测量结果。在某些实现方式中,计算机化的方法还包括基于在各自的计算模型中的节点数量来标准化得分。计算机化的方法还可以包括以第六处理器来生成每个所生成的得分的置信区间。置信区间可以包括对所生成的得分的分布进行近似,并且t统计量可以从所生成的得分的近似分布的方差中得出。在某些实现方式中,第一至第六处理器中的每一个都包括于或包含于单个处理器或单个计算装置之内。在其他实现方式中,第一至第六处理器中的一个或多个跨多个处理器或计算装置分布。

[0019] 本文所描述的计算机化的方法可以在具有一个或多个计算装置的计算机化系统中实现,每个计算装置包括一个或多个处理器。一般地,本文所描述的计算机化系统可以包括一个或多个引擎,该一个或多个引擎包括或包含被配置为具有硬件、固件和软件以执行本文所描述的一种或多种计算机化的方法的一个或多个处理装置,例如,计算机、微处理器、逻辑装置或者其他装置或处理器。在某些实现方式中,计算机化系统包括或包含系统响应曲线引擎(system response profile engine)、网络建模引擎和网络评分引擎。引擎可以时不时地互连,并且还时不时地与一个或多个数据库连接,包括扰动数据库、可测量数据库、实验数据数据库和文献数据库。本文所描述的计算机化系统可以包括或包含具有通过网络接口通信的一个或多个处理器和引擎的分布式计算机化系统。这样的实现方式可能适合于经由多种通信系统进行的分布式计算。在另一方面,还描述了包含适合于执行本文所描述的方法的程序代码的计算机程序产品。在另一方面,还描述了包括计算机程序产品的计算机或者计算机可记录介质或装置。

附图说明

[0020] 本公开内容的更多特征及其性质和各种优点在考虑了下面结合附图进行的详细描述的情况下将变得显而易见,在附图中同样的参考符号在所有附图中指代相同的部分,

并且在附图中：

[0021] 图1是用于量化生物网络对于扰动的响应的示例性计算机化系统的框图。

[0022] 图2是用于通过计算网络扰动幅度 (NPA) 得分来量化生物网络对于扰动的响应的示例性过程的流程图。

[0023] 图3是作为系统响应曲线的基础的数据的图形表示，其中包括两个物剂、两个参数、N个生物实体的数据。

[0024] 图4是具有若干生物实体以及它们的关系的生物网络的计算模型的图示。

[0025] 图5是用于生成几何扰动指数 (GPI) 得分的示例性过程的流程图。

[0026] 图6是用于生成概率扰动指数 (PPI) 得分的示例性过程的流程图。

[0027] 图7是用于生成预期扰动指数 (EPI) 得分的示例性过程的流程图。

[0028] 图8是用于生成几何扰动指数 (GPI) 得分的置信区间的示例性过程的流程图。

[0029] 图9示出了以本文所公开的系统和方法来分析的生物网络模型。

[0030] 图10-14示出了对基于网络的生物机制的网络扰动幅度 (PA) 评分结果。

[0031] 图15是用于量化生物扰动的影响的示例性分布式计算机化系统的框图；以及

[0032] 图16是可以用来实现本文所描述的任意计算机化系统中的任意构件的示例性计算装置的框图。

具体实施方式

[0033] 词“包括”或“包含”并不排除其他元件或步骤，并且不定冠词“一 (a)”或“一个 (an)”并不排除复数。本文所描述的是用于定量地评估在生物系统受到物剂扰动时于生物系统内的变化大小的计算系统和方法。某些实现方式包括或包含用于计算表示在生物系统的一部分内的变化大小的数字值的方法。该计算将从生物系统受到物剂扰动的一组可控实验中获得的数据集作为输入来使用。数据然后被应用于生物系统的特征的网络模型。网络模型被用作仿真和分析的基础，并且表示在生物系统中启用所感兴趣的特征的生物机制和路径。该特征或者其机制和路径中的某些可能有助于疾病的病理以及对生物系统的不利的健康影响。在数据库中关于生物系统的现有知识被用来构造网络模型，该网络模型由关于众多生物实体在各种条件下 (包括在正常条件下以及在物剂的扰动下) 的状态的数据填充。所使用的网络模型是动态的，因为它表示各种生物实体响应于扰动的状态变化，并且能够得出关于物剂对生物系统的影响的定量及客观的评价。同样提供了用于运行这些计算方法的计算机系统。

[0034] 除了别的以外，由本发明的计算机化的方法生成的数字值能够用来确定由制造产品 (用于安全评价或比较)、包括营养补充在内的治疗化合物 (用于确定疗效或健康效益) 以及环境活动物质 (用于预测长期接触的风险以及与副作用及疾病的发作的关系) 造成的所希望的或不利的生物影响的大小。

[0035] 在一个方面，本文所描述的系统和方法基于扰动的生物机制的网络模型来提供表示扰动的生物系统的变化大小的计算的数字值。在此称为网络扰动幅度 (NPA) 得分的数字值能够用来概要地表示在所定义的生物机制中的各种实体的状态变化。针对不同物剂或不同类型的扰动所获得的数字值能够用来相对地比较不同物剂或扰动对生物机制的影响，该生物机制使它自己能够作为或显现为生物系统的特征。因而，NPA得分可以用来测量生物机

制对于不同扰动的响应。术语“得分”在此一般地用来指代用于提供对生物系统的变化大小的定量度量的值或一组值。这样的得分通过使用本技术领域已知的及根据本文所公开的方法的各种数学及计算算法中的任一种,采用从样品或主体中获得的一个或多个数据集来计算出。

[0036] NPA得分可以帮助研究人员和临床医生改进诊断、实验设计、治疗决策和风险评估。例如,NPA得分可以用来筛选毒理学分析方面的一组候选的生物机制,以识别很可能在暴露于潜在有害的物剂下而受到影响的那些生物机制。通过提供对于扰动的网络响应的测量,这些NPA得分可以允许分子事件(由实验数据测量的)与出现于细胞、组织、器官或有机体层面的表型或生物学结果关联。临床医生可以使用NPA值来将受物剂影响的生物机制与患者的生理条件进行比较,以确定患者在暴露于该物剂时很可能会经历到什么健康风险或效益(例如,免疫功能受损的患者可能尤其容易受到导致强烈的免疫抑制响应的物剂的影响)。

[0037] 图1是用于量化网络模型对于扰动的响应的计算机化系统100的框图。特别地,系统100包括或包含系统响应曲线引擎110、网络建模引擎112和网络评分引擎114。引擎110、112和114时不时地互连,并且还时不时地与一个或多个数据库连接,所述数据库包括扰动数据库102、可测量数据库104、实验数据数据库106和文献数据库108。如同本文所使用的,引擎包括或包含一个或多个处理装置,例如,计算机、微处理器、逻辑装置或者参照图14所描述的一个或多个其他装置,该一个或多个处理装置以硬件、固件和软件来配置以执行一个或多个计算操作。

[0038] 图2是根据一种实现方式的用于通过计算网络扰动幅度(NPA)得分来量化生物网络对于扰动的响应的过程200的流程图。过程200的步骤将被描述为通过图1的系统100的各种构件来执行,但是这些步骤中的任一个都可以通过任意适合的硬件或软件构件(本地或远程)来执行,并且可以按照任何适当的顺序布置或者并行地执行。在步骤210,系统响应曲线(SRP)引擎110接收来自各种不同来源的生物数据,并且数据本身可以是各种不同类型的。数据包括来自其中生物系统受到扰动的实验数据,以及控制数据。在步骤212,SRP引擎110生成代表在生物系统中的一个或多个实体响应于物剂被引入生物系统而改变的程度的系统响应曲线(SRP)。在步骤214,网络建模引擎112提供含有多个网络模型的一个或多个数据库,所述数据库之一被选择为与感兴趣的物剂或特征相关。选择能够在作为系统的生物功能的基础的机制的现有知识的基础上作出。在某些实现方式中,网络建模引擎112可以使用系统响应曲线、在数据库中的网络以及之前在文献中描述的网络来提取在系统内的实体之间的因果关系,由此生成、精炼或扩充网络模型。在步骤216,网络评分引擎114使用由网络建模引擎112在步骤214识别的网络以及由SRP引擎110在步骤212生成的SRP生成每个扰动的NPA得分。NPA得分在生物实体之间的基础关系(由网络表示)的背景下量化对于扰动或处理的生物响应(由SRP表示)。下面的描述为了公开内容的清晰性起见而划分成子部分,而不是为了限制。

[0039] A. 生物系统

[0040] 生物系统在本发明的背景下是有机体或者有机体的一部分,包括功能部分,有机体在此称为主体。主体一般为哺乳动物,包括人类。主体能够是人类群体中的个体人。本文所使用的术语“哺乳动物”包括或包含(但不限于)人类、非人灵长类、小鼠、老鼠、狗、猫、牛、

羊、马和猪。除人之外的哺乳动物能够有利地被用作能够用来提供人类疾病的模型的主体。非人主体能够是未改性的、转基因动物、基因改造的动物或者携带有一个或多个基因突变或沉默基因的动物。主体能够是雄性或雌性。取决于操作的目的,主体能够是已经暴露于所感兴趣的物剂的主体。主体能够是已经在很长一段时间(可选地包括本研究之前的时间)内暴露于物剂的主体。主体能够是已经在一段时间内暴露于物剂的、但是不再与物剂接触的个体。主体能够是已经被诊断或被识别为具有疾病的主体。主体能够是已经经历过或者正在经历疾病或不利的健康状况的处理的主体。主体还可以是展示出针对具体的健康状况或疾病的一个或多个症状或风险因素的主体。主体能够是预感染的但是没有疾病症状的主体。在某些实现方式中,所讨论的疾病或健康状况与在很长的一段时间内暴露于物剂或使用物剂关联。根据某些实现方式,系统100(图1)含有或生成与所感兴趣的扰动类型或结果相关的一个或多个生物系统的计算机化模型及其功能机制(共同地,“生物网络”或“网络模型”)。

[0041] 取决于操作的环境,生物系统能够被定义于不同的水平,因为它涉及总体中的个体有机体的功能,有机体一般为器官、组织、细胞类型、细胞器、细胞组分或具体个体的细胞。每个生物系统包括一个或多个生物机制或路径,该一个或多个生物机制或路径的操作显现为系统的功能特征。用于再现所定义的人类健康状况的特征且适用于暴露于所感兴趣的物剂的动物系统是优选的生物系统。用于反映在疾病病因或病理中所涉及的细胞类型和组织的细胞和器官型系统同样是优选的生物系统。能够对尽可能地概括体内的人体生物学的原始细胞或器官培养赋予优先级。同样重要的是将体外人类细胞培养与从动物模型中得出的最相当的体内培养匹配。这允许使用匹配的体外系统作为参照系统来产生从体内的动物模型到人类生物学的转化连续。因此,为了用于本文所描述的系统和方法而构想出的生物系统能够通过功能特征(生物功能、生理功能或细胞功能)、细胞器、细胞类型、组织类型、器官、发展阶段或者前述项的组合(不限定)来定义。生物系统的实例包括或包含,但不限于,肺、外皮、骨骼、肌肉、神经(中枢和末梢)、内分泌、心血管、免疫、循环、呼吸、泌尿、肾脏、胃肠、结直肠、肝脏和生殖系统。其他的实例生物系统包括或包含,但不限于,在上皮细胞、神经细胞、血液细胞、连接组织细胞、平滑肌细胞、骨骼肌细胞、脂肪细胞、卵子细胞、精子细胞、干细胞、肺细胞、脑细胞、心肌细胞、喉部细胞、咽部细胞、食管细胞、胃部细胞、肾脏细胞、肝细胞、乳腺细胞、前列腺细胞、胰腺细胞、胰岛细胞、精巢细胞、膀胱细胞、宫颈细胞、子宫细胞、结肠细胞和直肠细胞中的各种细胞功能。某些细胞可以是细胞系的细胞,在适当的培养条件下体外培养或体外无限保持。细胞功能的实例包括或包含,但不限于,细胞增殖(例如,细胞分裂)、变性、再生、衰老、由核子对细胞活动的控制、细胞至细胞信令、细胞分化、细胞反分化、分泌、迁移、吞噬、修复、凋亡和发展规划。能够将其视为生物系统的细胞组分的实例包括或包含,但不限于,细胞质、细胞骨架、隔膜、核糖体、线粒体、核子、内质网(ER)、高尔基体、溶酶体、DNA、RNA、蛋白质、肽聚糖和抗体。

[0042] B. 扰动

[0043] 在生物系统中的扰动能够通过暴露或者与生物系统的一个或多个部分接触而由一个或多个物剂在一段时间内导致。物剂能够是单一物质或物质的混合物,包括其中并非全部成分都被识别或被特征化的混合物。物剂或其成分的化学和物理性质可以是没有完全特征化的。物剂能够由其结构、其成分或者在某些条件下会产生物剂的来源来定义。物剂的

实例是非均质物质,即,不存在于生物系统内的或者不是从其中得出的分子或实体,以及在与生物系统接触之后由其产生的任何中间物或代谢物。物剂能够是碳水化合物、蛋白质、脂质、核酸、生物碱、维生素、金属、重金属、矿物质、氧、离子、酶、激素、神经递质、无机化合物、有机化合物、环境物剂、微生物、粒子、环境条件、环境力或体力。物剂的非限制性实例包括或包含(但不限于)营养物、代谢废物、毒物、毒品、毒素、治疗化合物、刺激物、弛缓剂、天然产品、制造产品、食物材料、病原菌(朊蛋白、病毒、细菌、真菌、原生动物)、其尺寸在微米范围内或以下的粒子或实体、上述项目的副产品以及上述项的混合物。物理物剂的非限制性实例包括或包含放射物、电磁波(包括阳光)、温度的升高或降低、剪切力、流体压力、放电或其后果或外伤。

[0044] 有些物剂不会扰动生物系统,除非它达到阈值浓度或者它与生物系统接触达一段时间,或者两者的结合。导致扰动的物剂暴露或接触可以按照剂量来量化。因而,扰动能够由长期暴露于物剂引起。暴露的时长能够由时间的单位、由暴露的频率或者由在主体的实际或估计的寿命内的时间百分比来表示。扰动还能够由从生物系统的一个或多个部分中抑制物剂(如同以上所描述的)或者限制对生物系统的一个或多个部分的物剂供给导致。例如,扰动能够由营养物、水、碳水化合物、蛋白质、脂质、生物碱、维生素、矿物质、氧、离子、酶、激素、神经递质、抗体、细胞因子、光线的减少供应或缺乏,或者通过限制有机体的某些部分的移动,或者通过抑制或要求锻炼导致。

[0045] 物剂可以导致不同的扰动,这取决于生物系统的哪个(哪些)部分被暴露以及暴露条件。物剂的非限制性实例可以包括或包含通过加热烟草生成的烟雾剂、通过燃烧烟草生成的烟雾剂、烟草烟雾或香烟烟雾,以及它们的气态成分或微粒成分中的任一种。物剂的更多非限制性实例包括或包含镉、汞、铬、烟碱、烟草专有的亚硝胺及它们的代谢物(4-甲基亚硝胺基-1-(3-吡啶基)-1-丁酮4 (NNK)、N'-亚硝基去甲尼古丁 (NNN)、N-亚硝基新烟草碱 (NAT)、N-亚硝基假木贼碱 (NAB) 和4-(甲基亚硝氨基)-1-(3-吡啶基)-1-丁醇 (NNAL)), 以及用于烟碱替代疗法的任何产品。用于物剂的暴露方案或复合刺激应当反映在日常设定中暴露的范围和环境。标准的暴露方案的设置能够被设计为用于系统地应用于同样良好定义的实验系统。每个化验能够被设计为用于收集时间和剂量相关的数据,以采集早期的及后期的事件并且确保覆盖典型的剂量范围。但是,本领域技术人员应当理解,本文所描述的系统和方法可以被调整和修改,使其适用于正在处理的应用,并且本文所设计的系统和方法可以在其他合适的应用中使用,并且其它此类添加和修改不应脱离本发明的范围。

[0046] 在各种实现方式中,对基因表达、蛋白质表达或周转、微核糖核酸表达或周转、翻译后修饰(post-translational modifications)、蛋白质修饰、易位、抗体产生代谢物图谱或者上述项的两项或更多项的组合的高输出的系统级测量结果在各种条件(包括各自的控制)下生成。功能性结果测量在本文所描述的方法中是所希望的,因为它们一般地能够用作用于评估的锚定并且表示在疾病病因中的清晰步骤。

[0047] 如同本文所使用的,“样品”指的是独立于主体或实验系统的任何生物样品(例如,细胞、组织、器官或整个动物)。样品能够包括或包含(不作限定)单个细胞或多个细胞、细胞组分、组织活检、切除的组织、组织提出物、组织、组织培养提出物、组织培养介质、呼出气体、全血、血小板、血清、血浆、红细胞、白细胞、淋巴球、中性粒细胞、巨噬细胞、B细胞或其子集、T细胞或其子集、造血细胞子集、内皮细胞、滑体、淋巴液、腹水液、间质液、骨髓、脑脊液、

胸腔积液、肿瘤渗入物、唾液、黏液、痰、精液、汗液、尿液,或者任何别的体液。样品能够通过以下方法从主体中获得,包括但不限于:静脉穿刺、排泄、活检、针刺、灌洗、刮切、手术切除,或者本技术领域已知的其他方法。

[0048] 在操作期间,对于给定的生物机制、结果、扰动或者前述项的组合,系统100能够生成网络幅度(PA)值,该网络幅度值是对在网络中的生物实体响应于处理条件而产生的状态变化的量化度量。

[0049] 系统100(图1)包括与感兴趣的健康状况、疾病或生物结果相关的一个或多个计算机化的网络模型。这些网络模型中的一个或多个基于现有的生物知识,并且能够由外部来源上传并在系统100内管理。还能够基于测量在系统100内重新生成模型。可测量元件因此通过使用现有知识被集成到生物网络模型之内。以下所描述的是表示能够用来生成或精炼网络模型的感兴趣的生物系统的变化的或者表示对于扰动的响应的数据类型。

[0050] 参照图2,在步骤210,系统响应曲线(SRP)引擎110接收生物数据。SRP引擎110可以接收从各种不同来源接收该数据,并且数据本身可以是各种不同类型的。由SRP引擎110使用的生物数据可以从文献数据库(包括来自医药产品或医疗装置的临床前的、临床的及临床后的试验的数据)、基因组数据库(基因组序列及表达数据,例如,国家生物技术信息中心的基因表达库(Gene Expression Omnibus)或者欧洲生物信息研究所的ArrayExpress (Parkinson等,2010,Nucl.Acids Res.,doi:10.1093/nar/gkql040.Pubmed ID 21071405))、可购买的数据库(例如,美国马里兰州盖瑟斯堡的Gene Logic)或实验工作取得。数据可以包括或包含来自一个或多个不同来源的原始数据,例如,使用为了研究特定的处理条件或暴露于特定物剂的影响而特别设计的一个或多个物种的体外实验、间接体内或直接体内实验。体外实验系统可以包括或包含表示人类疾病的关键方面的组织培养或器官型培养(三维培养)。在这样的实现方式中,用于这些实验的物剂剂量和暴露方案可以基本上反映可以针对人类在正常使用或活动条件期间或者在特殊使用或活动条件期间预料的暴露范围和环境。实验参数和试验条件可以根据期望来选择以反映物剂的性质和暴露条件、所讨论的生物系统的分子和路径、所涉及的细胞类型和组织、感兴趣的结果以及疾病病因的各方面。从特定的动物模型中得到的分子、细胞或组织可以与特定的人类分子、细胞或组织培养匹配,以提高基于动物的发现的可译性。

[0051] 除了别的以外,由SRP引擎110接收到的许多都由高通量的实验技术生成的数据包括或包含(但不限于)与核酸(例如,具体DNA或RNA种类的绝对量或相对量、DNA序列的变化、RNA序列、三级结构的变化,或者由序列化、在微阵列上的核酸特定的杂交、定量的聚合酶链式反应或者本技术领域已知的其他技术所确定的甲基化模式)、蛋白质/肽(例如,由本技术领域已知的方法确定的蛋白质的绝对量或相对量、蛋白质的具体片段、肽聚糖、二级或三级结构的变化或者翻译后修饰)以及在某些条件下的机能活动(例如,酶的活动、蛋白质水解活动、转译调整活动、运输活动、与某些结合配对物的结合亲和力)相关的数据。包括蛋白质或肽的翻译后修饰在内的修饰能够包括或包含(但不限于)甲基化、乙酰化、法呢酰化、生物素酰化、硬脂酰化、甲酰化、豆蔻酰化、蛋白质棕榈酰化、香叶酰香叶酰化、聚乙二醇化、磷酸化、硫酸化、糖基化、糖化改变(sugar modification)、脂质化、脂质改变、泛素化、sumolation、二氧化硫键合、cysteinylation、氧化、谷胱甘肽、羧化、醛糖酸化反应和脱酰胺化。另外,蛋白质还能够通过一系列的反应来进行翻译后修饰,例如,产生糖化蛋白质产

物的Amadori反应、席夫碱反应和美拉德反应。

[0052] 数据还可以包括或包含测得的机能结果,例如,但不限于,处于细胞层面的机能结果,包括细胞增殖、发展命运和细胞死亡,处于生理层面的机能结果,包括肺活量、血压、锻炼熟练度。数据还可以包括或包含疾病活动或严重性的测量值,例如,但不限于,肿瘤转移、肿瘤好转、功能丧失以及在疾病的某一阶段的预期寿命。疾病活动能够通过其结果为能够在所定义的条件下来根据来自一个或多个主体的样品(或者样品的总体)的评价来获得的值或一组值的临床评估来测量。临床评估还能够基于由主体提供的对访问或调查问卷的回答。

[0053] 这种数据可以为了明确地在确定系统响应曲线中使用而生成,或者可以在之前的实验中产生或者已发表于文献中。一般地,数据包括或包含与分子、生物结构、生理条件、基因特征或表型相关的信息。在某些实现方式中,数据包括或包含关于条件、位置、数量、活动,或者分子的子结构、生物结构、生理条件、基因特征或表型的描述。如后面将描述的,在临床设定中,数据可以包括或包含从对从人类主体中获得的样品执行的化验或对暴露于物剂的人类主体的观察中获得的原始数据或已处理的数据。

[0054] 在步骤212,系统响应曲线(SRP)引擎110基于在步骤212接收到的生物数据来生成系统响应曲线(SRP)。该步骤可以包括或包含背景校正、标准化、倍数变化计算、显著性确定以及差别响应(例如,表达不同的基因)的识别中的一项或多项。SRP是用于表示在生物系统内的一个或多个测量的实体(例如,分子、核酸、肽、蛋白质、细胞等)响应于施加于生物系统的扰动(例如,暴露于物剂)而单独改变的程度的表示。在一个实例中,为了生成SRP,SRP引擎110收集用于施加于给定的实验系统(“系统-处理”对)的一组给定参数(例如,处理或扰动条件)的一组测量值。图3示出了两个SRP:包括或包含经受到具有变化的参数(例如,暴露于第一处理物剂的剂量和时间)的第一处理306的N个不同的生物实体的生物活动数据的SRP 302,以及包括或包含经受到第二处理308的N个不同的生物实体的生物活动数据的类似的SRP 304。包括或包含于SRP内的数据可以是原始实验数据、已处理的实验数据(例如,被过滤以去除异常值,以置信估计来标记,针对多次试验求均值)、由计算生物模型生成的数据,或者从科学文献中取得的数据。SRP可以以众多方式来表示数据,例如,绝对值、绝对变化、倍数变化、对数变化、函数和表格。SRP引擎110将SRP传递给网络建模引擎112。

[0055] 虽然在前一步骤中得出的SRP表示将用于确定网络扰动的大小的实验数据,但尤其是作为用于计算和分析的基础的生物网络模型。该分析要求开发与生物系统的特征相关的机制和路径的详细网络模型。这样的架构提供除已经在更典型的基因表达分析中使用的基因列表的检查之外的机理层理解。生物系统的网络模型是表示动态生物系统的且通过组装有关生物系统的各种基本性质的定量信息来建立的数学构造。

[0056] 这样的网络架构是迭代过程。关于网络边界的描绘由与感兴趣的过程(例如,在肺中的细胞增殖)相关的机制和路径的文献调研指导。用于描述这些路径的因果关系从现有的知识中取得以使网络集结。基于文献的网络能够使用含有相关的表型端点的高通量的数据集来验证。SRP引擎110能够用来分析数据集,该分析结果能够用来证实,精炼或生成网络模型。

[0057] C.网络

[0058] 参照图2,在步骤214,网络建模引擎112使用来自具有基于以感兴趣的生物系统的

特征为基础的机制或路径的网络模型的SRP引擎110的系统响应曲线。在某些方面,网络建模引擎112被用来识别已经基于SRP而生成的网络。网络建模引擎112可以包括或包含用于接收模型的更新和改变的构件。网络建模引擎112还可以重复网络生成的过程,并加入新的数据并生成附加的或精炼的网络模型。网络建模引擎112还可以促进一个或多个数据集的合并或者一个或多个网络的合并。从数据库中取得的网络集可以手动补充另加的节点、边或者全新的网络(例如,通过挖掘用于描述由特定的生物实体直接调节的附加基因的文献的文字)。这些网络含有可以允许进行过程评分的特征。网络拓扑结构被保持;因果关系的网络能够从网络中的任意点追踪到可测量的实体。此外,模型是动态的,并且用来构建它们的假定能够被修改或被重申并允许适应于不同的组织环境和种类。这允许在有新的知识可用时进行反复的测试和改进。网络建模引擎112可以去除具有低置信度或者作为与科学文献中的实验结果冲突的主体的节点或边。网络建模引擎112还可以包括或包含可以使用有监管或无监管的学习方法(例如,度量学习、矩阵填充、模式识别)推断出的附加的节点或边。

[0059] 在某些方面,生物系统被建模为由顶点(或节点)和用于连接节点的边构成的数学图。例如,图4示出了具有9个节点(包括节点402和404)和边(406和408)的简单网络400。节点能够表示在生物系统中的生物实体或过程,例如,但不限于,化合物、DNA、RNA、基因、蛋白质、肽聚糖、抗体、细胞、组织、器官和细胞或分子过程。生物实体并不一定限定于其处理或控制数据被接收到或可使用的那些生物实体。因而,表示生物实体的节点能够包括或包含该多个生物实体,并且可以包括或包含一个或多个另外的生物实体。至少一些节点是可评分的,并且得分可以表示节点的活动水平。许多节点表示其活动水平能够被测量的生物实体。但是,在某些实现方式中,计算机化的方法并不一定要接收所有此类可测量节点的数据。因而,节点是可评分的和/或可测量的。在某些实现方式中,大多数节点都是可测量的。可测量节点可以含有或包括所测得的数据。边能够表示在节点之间的关系。图中的边能够表示节点之间的各种关系。例如,边可以表示“绑定至”的关系、“用以表达”的关系、“基于表达图谱来共同调控”的关系、“抑制”的关系、“共同发生于手稿中”的关系或者“共用结构元件”的关系。一般地,这些类型的关系描述一对节点间的关系。图中的节点还能够表示节点间的关系。因而,可以表示关系之间的关系或者在关系与表示于图中的另一类型的生物实体之间的关系。例如,在表示化学品的两个节点之间的关系可以表示反应。该反应可以是处于反应与用于抑制该反应的化学品之间的关系中的节点。

[0060] 图可以是无方向性的,意指在与每个边关联的两个顶点之间没有方向。作为选择,图形的边可以从一个顶点指向另一个顶点。例如,在生物的环境中,转录调控网络和代谢网络可以被建模为有向图。在转录调控网络的图形模型中,节点将表示基因,边表示节点之间的转录关系。作为另一个实例,蛋白质-蛋白质相互作用网络描述在有机体的蛋白质组中的蛋白质之间的直接的物理相互作用,并且在此类网络中通常没有与相互作用关联的方向。因而,这些网络可以被建模为无向图。某些网络可以具有有向边和无向边。构成图形的实体和关系(即,节点和边)可以作为相关节点的网络存储于系统100中的数据库内。

[0061] 在数据库中表示的知识可以是各种不同类型的,是从各种不同来源中取得的。例如,某些数据可以表示基因组数据库,包括关于基因的信息以及它们之间的关系。在这样的实例中,节点可以表示致癌基因,而与致癌基因节点连接的另一个节点可以表示用于抑制致癌基因的基因。数据可以表示蛋白质及它们之间的关系、疾病及它们的相互关系,以及各

种疾病状态。存在着能够结合于图形表示中的许多不同类型的数据。计算模型可以表示在表示在例如DNA数据集、RNA数据集、蛋白质数据集、抗体数据集、细胞数据集、组织数据集、器官数据集、医疗数据集、流行病学数据、化学数据集、毒理学数据集、患者数据集和人口数据集集中的知识的节点之间的关系的网络。如同本文所使用的,数据集是由在所定义的条件 下对样品(或一组样品)的评价得出的数字值的合集。数据集能够通过例如在实验上测量样品的可量化实体来获得;或者作为选择,从服务提供者(例如,实验室、临床研究组织)处获得,或者从公共或专有数据库中获得。数据集可以含有数据以及由节点表示的生物实体,并且在每个数据集集中的节点可以与同一数据集集中的其他节点或者其他数据集集中的节点相关。而且,网络建模引擎112可以生成用于将在例如DNA、RNA、蛋白质或抗体的数据集集中的遗传信息表示为在医疗数据集集中的医疗信息,表示为在患者数据集集中的关于个体患者的信息,以及表示为在流行病学数据集集中的关于整个人群的信息的计算模型。除了以上所描述的各种数据集之外,还可以存在许多别的数据集,或者可以在生成计算模型时包括的或包含的生物信息的类型。例如,数据库能够还包括或包含医疗记录数据、结构/活动关系数据、关于传染病理的信息、关于临床试验的信息、暴露模式数据、与产品的使用历史相关的数据,以及任何其他类型的生命科学相关的信息。

[0062] 网络建模引擎112可以生成表示例如在基因之间的调节的相互作用、在蛋白质之间的相互作用或者在细胞或组织内的复杂的生化相互作用的一个或多个网络模型。由网络建模引擎112生成的网络可以包括或包含静态模型和动态模型。网络建模引擎112可以采用任何可应用的数据方案来表示系统,例如,超图和加权二分图,在所述数学方案中使用两种类型的节点来表示反应和化合物。网络建模引擎112还可以使用其他推理技术来生成网络模型,例如,基于在表达不同的基因中的功能相关的基因的过表达的分析、贝叶斯网络分析、图形化高斯模型技术或者基因关联网络技术,以基于一组实验数据(例如,基因表达、代谢物浓度、细胞应答等)来识别相关的生物网络。生物系统可以由多个网络模型表示,包括计算因果网络模型。

[0063] 如同以上所描述的,网络模型基于作为生物系统的功能特征的基础的机制和路径。网络建模引擎112可以生成或含有表示有关与对物剂的长期健康风险或健康效益的研究相关的生物系统的特征的结果的模型。因此,网络建模引擎112可以生成或含有用于细胞功能的各种机制的网络模型,尤其是与在生物系统中的感兴趣的特征有关的或对其有贡献的那些机制,包括(但不限于)细胞增殖、细胞应激、细胞再生、凋亡、DNA破坏/修复或炎症反应。在其他实施例中,网络建模引擎112可以含有或生成与急性全身毒性、致癌性、经皮渗透、心血管疾病、肺部疾病、生态毒性、眼部灌洗/腐蚀、遗传毒性、免疫毒性、神经毒性、药代动力学、药物代谢、器官毒性、生殖和发育毒性、皮肤刺激/腐蚀或皮肤致敏相关的计算模型。一般地,网络建模引擎112可以含有或生成用于核酸(DNA、RNA、SNP、siRNA、miRNA、RNAi)、蛋白质、肽聚糖、抗体、细胞、组织、器官和任何其他生物实体的状态以及它们各自的相互作用的计算模型。在一个实例中,计算网络模型能够用来表示免疫系统的状态以及各种类型的白细胞在免疫响应或炎症反应期间的机能。在其他实例中,计算网络模型能够用来表示心血管系统的性能以及内皮细胞的机能和代谢。

[0064] 在本发明的某些实现方式中,网络是从因果生物知识的数据库得出的。该数据库可以通过对不同的生物机制执行实验研究以提取机制之间的关系(例如,激活或抑制关

系) 而生成, 这些关系中的一些可以是因果关系, 并且可以与市场上可购得的数据库结合, 例如, 由美国马萨诸塞州剑桥市的Selventa公司管理的Genstruct技术平台 (Genstruct Technology Platform) 或Selventa知识库 (Selventa Knowledgebase)。使用因果生物知识的数据库, 网络建模引擎112可以识别用于链接扰动102与可测物104的网络。在某些实现方式中, 网络建模引擎112使用来自SRP引擎110的系统响应曲线以及之前在文献中生成的网络来提取生物实体之间的因果关系。除了其他处理步骤之外, 数据库还可以被进一步处理以通过在不同的生物实体集之间应用同源推理来去除逻辑不一致性并生成新的生物知识。

[0065] 在某些实现方式中, 从数据库中提取出的网络模型基于反向因果推理 (RCR)、用于处理因果关系的网络以制定机制假设并然后针对差别测量的数据集来评价那些机制假设的自动推理技术。每个机制假设将生物实体链接至它能够影响的可测量的量。至少一个机制假设可以被制定, 例如, 多个机制假设。例如, 除了别的以外, 可测量的量还能够包括或包含浓度的增大或减小、生物实体的数量或相对多度、生物实体的激活或抑制或者生物实体的结构、功能或逻辑的改变。RCR使用作为计算的基础的在生物实体之间的在实验上观察到的因果相互作用的有向网络。有向网络可以用Biological Expression Language™ (BEL™) 语言 (用于记录生物实体之间的相互关系的语法) 来表达。RCR计算为网络模型生成指定某些约束, 例如, 但不限于, 路径长度 (连接上游节点与下游节点的边的最大数量) 以及用于将上游节点连接至下游节点的可能的因果路径。RCR的输出是根据用于评价关联性和精确性的统计数据来排名的一组机制假设, 这组假设表示实验测量值的差异的上游控制者。机制假设输出能够被组合到因果链和较大的网络中, 以在较高的互连机制和路径层面解释数据集。

[0066] 一种类型的机制假设包括存在于表示可能原因的节点 (上游节点或控制者) 与表示测得量的节点 (下游节点) 之间的因果关系集。机制假设能够用来进行预测, 例如, 如果由上游节点表示的实体的多度增大, 则由因果增大关系链接的下游节点将被推断为是增大的, 而由因果减小关系链接的下游节点将被推断为减小的。

[0067] 机制假设表示在一组测得的数据 (例如, 基因表达数据) 与作为那些基因的已知控制者的生物实体之间的关系。另外, 这些关系包括或包含在上游实体与下游基因的差别表达之间的影响的符号 (正号或负号)。假设的下游基因从管理文献的因果生物知识的数据库中取出。形式为可计算的因果网络模型的用于将上游实体链接至下游基因的机制假设的因果关系是通过NPA评分方法来计算网络变化的基础。生物系统可以由至少一个机制假设 (例如, 多个机制假设) 表示。该至少一个计算因果网络模型可以包括多个机制假设。

[0068] 生物实体的可评分的复杂因果网络模型能够通过收集表示模型中的实体的个体机制假设并将表示整个复杂因果网络模型的所有下游基因与单个上游过程的连接重新分组而转换成单个因果网络模型; 这实际上是基础图结构的扁平化。以此方式, 通过网络模型来描述的生物实体的活动变化能够经其个体机制假设的结合来评估, 使得基础的基因表达测量在整体上对网络有贡献。

[0069] 为了生成在本发明的方法中使用的可评分的网络, 参考节点首先选自起始的, 典型为复杂的因果网络模型。参考节点能够是在网络中的其水平或活动在整体上与网络的活动是正相关的任何实体 (这与例如其活动可以是与网络活动负相关的抑制剂相反)。然后, 确定在模型中的每个节点与参考节点之间的因果关系。这能够通过首先要求模型为“因果

关系一致”来完成。用于模型中的每个节点的下游可测量实体(在本例中为基因表达)的调节的符号基于在该模型节点与参考节点之间的关系来调整。例如,用于与参考节点具有正向因果关系的模型节点(即,该节点预计将在参考节点增大时被正向调节)的下游基因表达的符号被保持。另一方面,用于与参考节点具有负向因果关系的模型节点(即,该节点预计将在参考节点增大时被负向调节)的下游基因表达的符号是反向的。所有下游基因表达及它们的符号然后被组合到单个机制假设中,并且具有相反符号的下游基因表达(来自多种模型节点)被从机制假设中删除。

[0070] 对于将为因果关系一致的网络模型,对于模型中的任何节点的增加,应当有可能的是通过遵循用于连接节点的因果关系来明确地将“正向调节”或“负向调节”的符号映射于模型中的每个其他节点上。生物解释能够用来解决不明确性,以通过考虑什么过程正通过机制假设来评分以及每个节点按照什么符号与参考节点有效相关来构造因果关系一致的模型。例如,负反馈于其处连接回到模型的节点与正被评分的过程具有特定的关系,并且尽管负反馈可以调节该节点,它也不应改变该关系。因而,在负反馈回路与该节点之间的连接能够被从模型中去除,以按照与已知的事实相符的方式来获得因果一致性。关于以上所描述的方法的变型在美国专利申请公开No.2007/0225956和2009/0099784中进行了讨论,这两个申请通过引用的方式全文并入本文。示例性的因果网络模型在Westra JW、Schlage WK、Frushour BP、Gebel S、Catlett NL、Han W、Eddy SF、Hengstermann A、Matthews AL、Mathis C等人的Construction of a Computable Cell Proliferation Network Focused on Non-diseased Lung Cells,BMC Syst Biol 2011,5:105中进行了描述,该文章通过引用的方式全文并入本文。

[0071] 在某些实现方式中,系统100可以在细胞已经暴露于香烟烟雾时含有或生成用于细胞增殖机制的计算机化模型。在这样的实例中,系统100还可以含有或生成表示与香烟烟雾暴露相关的各种健康状况的一个或多个网络模型,包括(但不限于),癌症、肺部疾病和心血管疾病。在某些方面,这些网络模型基于所施加的扰动(例如,暴露于物剂)、在各种条件下的响应、感兴趣的可测量的量、正在研究的结果(例如,细胞增殖、细胞应激、发炎、DNA修复)、实验数据、临床数据、流行病学数据和文献中的至少一项。

[0072] 作为说明性的实例,网络建模引擎112可以被配置用于生成细胞应激的网络模型。网络建模引擎112可以接收用于描述从文献数据库中获知的应激响应所涉及的相关机制的网络。网络建模引擎112可以基于已知的生物机制来选择一个或多个网络,以响应于在肺和心血管的环境下的应激而操作。在某些实现方式中,网络建模引擎112识别出在生物系统中的一个或多个功能单元,并且通过基于它们的功能来结合较小的网络而构建较大的网络模型。特别地,对于细胞应激模型,网络建模引擎112可以考虑与对氧化应激、基因毒性应激、低氧应激、渗透、外源性应激和剪切应激的响应相关的功能单元。因此,用于细胞应激模型的网络构件可以包括或包含外源性代谢响应、基因毒性应激、内皮剪切应激、低氧响应、渗透应激和氧化应激。网络建模引擎112还可以接收来自对来自在特定的细胞分组中执行的应激关联实验的公开可用的转录数据的计算分析的内容。

[0073] 当生成生物机制的网络模型时,网络建模引擎112可以包括或包含一个或多个规则。这样的规则可以包括或包含用于选择网络内容、节点类型等的规则。网络建模引擎112可以从实验数据数据库106中选出一个或多个数据集,包括体外和体内实验结果的结合。网

络建模引擎112可以使用实验数据来验证在文献中所识别的节点和边。在对细胞应激进行建模的实例中,网络建模引擎112可以基于实验表示在无疾病的肺或心血管组织内的生理上相关的应激的良好程度来选择实验的数据集。数据集的选择可以基于例如表型应激端点数据的可用性、基因表达图谱实验的统计严密性,以及实验环境与正常肺或心血管无疾病的生物的关联性。

[0074] 在识别了相关网络的合集之后,网络建模引擎112还可以处理并精炼那些网络。例如,在某些实现方式中,多种生物实体及它们的连接可以被分组并且由一个或多个新节点来表示(例如,使用聚类或其他技术)。

[0075] 网络建模引擎112还可以包括或包含与所识别的网络中的节点和边有关的描述性信息。节点可以通过其关联的生物实体、关于关联的生物实体是否是可测量的量的指示,或者生物实体的任何其他描述符来描述。某些节点是可评分的,并且得分可以表示节点的活动水平。许多节点表示其活动水平能够被测量的生物实体。但是,在某些实现方式中,计算机化的方法并不一定要接收所有此类可测量的节点的数据。因而,节点是可评分的和/或可测量的。在某些实现方式中,大多数节点是可测量的。可测量的节点可以含有或包括所测得的数据。边可以通过例如它表示的关系的类型(例如,因果关系(例如,上调或下调)、相关性、条件相关或独立)、该关系的强度或者在该关系中的统计置信度来描述。在某些实现方式中,对于每个处理,表示可测量的实体的每个节点与响应于该处理的活动变化的预期方向(即,增大或减小)关联。例如,当支气管上皮细胞暴露于诸如肿瘤坏死因子(TNF)之类的物剂时,特定基因的活动可以增加。这种增加可以由于可从文献中获知的直接调节关系(并且以由网络建模引擎112识别出的一个网络来表示)而出现,或者通过经由网络建模引擎112所识别的一个或多个网络的边来跟踪众多的调节关系(例如,自分泌信令)而出现。在某些情形中,网络建模引擎112可以识别每个可测量实体响应于特定的扰动而改变的预期方向。当在网络中的不同路径指示特定实体的改变的相反的预期方向时,这两个路径可以被更详细地检查,以确定净改变方向,或者该特定实体的测量值可以被丢弃。在某些实施例中,节点的方向值可以表示在控制数据与处理数据之间的变化的预期方向。在某些实施例中,节点的方向值可以表示在控制数据与处理数据之间的预期值变化。在某些实施例中,节点的方向值可以表示控制数据和处理数据的值的预期增加或减小。适当地,该变化表示处理之后的变化。

[0076] D. 网络扰动幅度

[0077] 本文所提供的计算方法和系统将SRP转译成NPA得分。被识别为在网络模型之内的扰动的下游影响的实验测量结果被聚合到专用于网络的响应得分中。因此,在步骤216,网络评分引擎114使用由网络建模引擎112在步骤214识别的网络以及由SRP引擎110在步骤212生成的SRP来为每个扰动生成NPA得分。NPA评分将所定义的一个或多个算法应用于由一系列处理-控制比较构成的实验数据集,在该实验数据集中,实验数据在所定义的生物网络模型的背景下被过滤以表示特定范围的生物(例如,一组特定的基因表达关系)。NPA得分在生物实体之间的基础关系(由所识别的网络表示)的背景下对处理的生物响应(由SRP表示)进行量化。网络评分引擎114包括或包含用于生成包含于网络建模引擎112内的或者由其识别的每个网络的NPA得分的硬件和软件构件。

[0078] 网络评分引擎114可以被配置用于实现众多评分技术中的任一种。这样的技术包

括用于生成标量值得分的技术。这样的技术还包括用于生成向量值得分的技术。向量值得分指示网络对于扰动的响应的大小和拓扑分布。

[0079] 一种所描述的评分技术是强度评分技术。强度得分是作为活动的均值的标量值得分。强度得分是对于在SRP中表示的不同实体的活动观察值的均值。网络响应的强度按照下式来计算：

$$[0080] \quad strength = \frac{\sum_i d_i \beta_i}{N} \quad (1)$$

[0081] 其中 d_i 表示与节点 i 关联的实体的活动改变的预期方向， β_i 表示在处理与控制条件之间的活动的倍数变化的对数（即，用于描述从初始值变为最终值的数量变化程度的数字），并且 N 是具有关联的已测量的生物实体的节点的数量。正强度得分指示SRP与从所识别的网络中得出的预期活动变化匹配，而负强度得分指示SRP与预期活动变化不匹配。

[0082] 得分可以通过几何扰动指数评分技术、概率扰动指数评分技术或预期扰动指数评分技术来生成。一种评分技术是几何扰动指数 (GPI) 评分技术。图5是可以由网络评分引擎114实现的GPI评分技术的流程图500。在步骤502，网络评分引擎组构 (assemble) 倍数变化向量 β 。倍数变化是用于描述在不同的条件下（例如，在扰动与控制条件之间）可测量变化从初始值变为最终值的程度的数字。该倍数变化向量具有 N 个分量，对应于在具有关联的测量的生物实体的网络中的节点的数量。在某些实现方式中，倍数变化向量的第 i 分量 β_i 表示第 i 测量生物实体的活动在扰动与控制条件之间的倍数变化的对数（例如，底为2）（即，实体的活动在两个条件之间变化的因子的对数）。结果， β_i 的值为0指示在扰动与控制条件之间没有观察到活动变化。对数运算无需包含于此，或者可以由任何别的线性或非线性函数来代替。例如，在某些实现方式中， β_i 表示在没有对数运算的情况下于扰动条件之间的活动的倍数变化；在这样的实现方式中， β_i 的值为1指示在扰动与控制条件之间没有观察到活动变化。应当理解，倍数变化仅仅是用于本文所描述的网络评分技术的量化活动的一种可能方法，并且用于表达可测物的变化的其他惯用手段也可以使用。在某些实施例中，生成得分的步骤可以包括活动测量结果、权重值和方向值的线性或非线性组合；以及通过比例因子对该组合进行的标准化。该组合可以是算术组合，并且比例因子可以是接收到其测得数据的生物实体的数量的平方根。在某些实施例中，得分不是标量值得分。

[0083] 在步骤504，网络评分引擎114生成加权向量 r 。加权向量 r 同样具有 N 个分量，针对倍数变化向量 β 的每个分量一个。加权向量 r 的每个分量 r_i 表示待赋予观察到的第 i 倍数变化 β_i 的权重。在某些实现方式中，权重表示第 i 测量实体关于感兴趣的特征或结果（例如，在癌症研究中的已知致癌物质）的已知生物学意义。在某些实现方式中，权重表示与节点关联的生物实体的活动测量值的置信度。通过按置信度估计来给对数倍数变化加权，其置信度低的倍数变化 β_i 对GPI得分贡献较小。改进的实验室条件、增加的生物复制数、较好的再现性、较小的方差和较强的信号可以全都有助于特定 β_i 的较高置信度。

[0084] 可以有利地用来加权的一个值是本地错误非发现率 $fndr_i$ （即，倍数变化值 β_i 表示违反关于零倍数变化的基础原假设的概率，在某些情形中，在所观察到的 p 值的条件下），如同由Strimmer等人在“A general modular framework for gene set enrichment analysis” (BMC Bioinformatics 10:47, 2009) 中以及由Strimmer在“A unified approach

to false discovery rate estimation” (BMC Bioinformatics9:303,2008)中所描述的,这两篇文章均通过引用的方式全文并入本文。在某些实现方式中,fndr_i按照下式来计算:

$$[0085] \quad fndr_i(\beta_1, \dots, \beta_N) = 1 - 2v_i(\beta_1, \dots, \beta_N) \int_{|\beta_i/S_i(\beta_1, \dots, \beta_N)|}^{\infty} t_{df}(x) dx \quad (2)$$

[0086] 其中fdr_i是本地错误发现率(即,倍数变化值 β_i 没有表示违反关于零倍数变化的基础原假设的概率), v_i 是由Benjamini等在“Controlling the false discovery rate: a practical and powerful approach to multiple testing”(Journal of the Royal Statistical Society, Series B 57:289,1995)中描述的Benjamini-Hochberg调整因子,该文章通过引用的方式全文并入本文,p是获得至少如实际观察到的倍数变化 β_i 那样极端的倍数变化的概率(假定零倍数变化的原假设为真),而 t_{df} 是df个自由度的t分布。注意,p是关于 β_i 和标准方差 S_i 的函数,该标准方差 S_i 进而基于全部 β_i 。在可替换的实现方式中,没有对多种测试进行调整;因此, $v_i(\beta_1, \dots, \beta_N)$ 等于1,并且加权向量 $r_i = 1 - p(\beta_i, S_i(\beta_1, \dots, \beta_N))$ 。

[0087] 在步骤506,网络评分引擎114使用加权向量r来缩放倍数变化向量 β 。结果是其中每个分量 β_i 乘以其关联的权重分量 r_i 的缩放的倍数变化向量。实现此类在计算上的缩放的一种方式是在对角线上具有权重分量 r_i 的NxN对角矩阵,并且将该矩阵乘以Nx1向量 β ,如等式3所示:

$$[0088] \quad \begin{bmatrix} r_1 \beta_1 \\ r_2 \beta_2 \\ \vdots \\ r_N \beta_N \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & r_N \end{bmatrix}}_{diag(r)} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}}_{\beta} \quad (3)$$

[0089] 在步骤508,网络评分引擎114识别在倍数变化向量 β 中的每个分量的变化的预期方向。网络评分引擎114可以通过询问网络建模引擎112以检索因果生物网络模型的变化预期方向来这样做。网络评分引擎114然后能够将这些预期的变化方向结合到N个分量的向量d中,其中向量d的第i分量 d_i 表示第i测量生物实体的预期的变化方向(例如,+1用于增大的活动,而-1用于减小的活动)。

[0090] 在步骤510,网络评分引擎114将缩放的倍数变化向量(在步骤506生成的)的分量与每个分量的预期变化方向(在步骤508识别出的)结合。在某些实现方式中,该组合是算术组合,其中所缩放的倍数变化 $r_i \beta_i$ 每个都乘以其相应的预期变化方向 d_i ,并且结果对于所有N个生物实体进行叠加。在算术上,步骤510的这种实现方式能够由下式表示:

$$[0091] \quad \sum_i d_i r_i \beta_i \quad (4)$$

[0092] 在其他实现方式中,向量d、r和 β 可以按照任何线性的或非线性的方式结合。

[0093] 在步骤512,网络评分引擎114将步骤510的组合标准化。在某些实现方式中,标准化包括乘以预定的比例因子。一个这样的比例因子是生物实体数N的平方根。在该实现方式中,GPI得分能够由下式表示:

$$[0094] \quad GPI = \frac{\sum_i d_i r_i \beta_i}{\sqrt{N}} \quad (5)$$

[0095] 可以是预定的也可以不是预定的其他比例因子同样可以使用。在某些实施例中，因果网络模型（例如，机制假设）能够被看作是在N维下游可测量空间中的单位负向量 $s = (1, 1, -1, 1, \dots, -1) / \sqrt{N}$ （其中每个维度表示因果网络模型的下游可测量值，在此为基因表达）。所观察到的扰动对下游基因表达的影响同样是该空间中的向量。因而在几何形状上，在因果网络模型中的扰动的幅度能够通过将差分 \log_2 表达向量投射到假设单元向量上来量化。但是，因果网络模型的下游测量来自一般模型。为了明确地处理支持NPA得分的数据的特异性，每个下游被指派激活的置信，该置信被设定为本地错误非发现率（ $fndr_i = (1 - fdr_i)$ ）。它等同于根据每个差分表达的置信来给下游基因表达空间的维度加权，并且因此考虑以加权标量积来定义基因表达空间的几何形状： $\langle s | \beta \rangle_w = S^T \cdot \text{diag}(fndr) \cdot \beta$ 。因此， $GPI = (\sum s_i - fndr_i \beta_i) / \sqrt{N}$ 。通过用错误非发现率给差分 \log_2 表达加权，置信度很小的个体差别表达值被移至较靠近0（无变化），而置信度较强的值被最低程度地减小。正的GPI得分指示由机制假设描述的过程的上调，为零的GPI得分指示该过程沿着机制假设的方向是不变的，而负的GPI得分指示过程被下调。

[0096] 图6是可以由网络评分引擎114实现的概率扰动指数（PPI）评分技术的流程图600。如同以上关于SRP引擎110（图1）及过程200的步骤212（图2）的描述那样，每个SRP都表示测量生物实体在处理条件下的活动（或活动变化）。然后，每个SRP与所测得的活动数量关联，每个测量的生物实体一个。PPI是由感兴趣的网络表示的生物机制在给定所观察到的SRP的情况下被激活的概率的量化。

[0097] 在步骤602，网络评分引擎114组构倍数变化向量 β 。表示所观察到的N个测量的生物实体的活动的倍数变化的这种倍数变化向量可以如同以上所描述的那样参照图5所示的几何扰动指数（GPI）评分技术的步骤502来组构。在步骤604，网络评分引擎114生成倍数变化密度的范围。倍数变化密度的范围表示在处理条件下倍数变化值能够于生物系统中采用的值的集合的近似，并且可以由范围 $[-W, W]$ 来近似，其中W是 \log_2 倍数变化的理论的预期最大绝对值。通过这样来选择W，所有观察到的倍数变化都将落入范围 $[-W, W]$ 之内。例如，基因芯片的最大预期信号（例如，为 \log_2 尺度的16）可以被用作值W。

[0098] 在步骤606，网络评分引擎114识别倍数变化向量 β 中的每个分量的预期变化方向。该步骤可以如同以上所描述的参照图5所示的GPI评分技术的步骤508来执行，产生一组与所观察到的倍数变化 β_i 相应的预期变化方向 d_i 。

[0099] 在步骤608，网络评分引擎114生成正活化度量。在某些实现方式中，正活化度量表示SRP指示所观察到的生物实体的激活/抑制与由 d_i 表示的预期变化方向一致的程度。一致的行为在本文中称为“正活化”。可以使用的一个正活化度量是一个或多个网络被正向激活的概率。这样的概率（称为PPI+）可以按照以下表达式来计算：

$$[0100] \quad PPI^+ = \Pr(\text{Positively Activated}) = \frac{1}{W} \int_0^W \Pr(\text{Positively Activated} | \phi) d\phi \quad (6)$$

[0101] 其中

$$[0102] \quad \Pr(\text{Positively Activated} | \varphi) = \frac{1}{N} \sum_{0 < d_i \beta_i < \varphi} fndr_i \quad (7)$$

[0103] 其中 $fndr_i$ 是以上参照公式1所讨论的错误非发现率。在某些实现方式中,网络评分引擎114被配置为使用表示0至 W 之间的 φ 值的区段集在数字上对公式6的表达式求积分。可以使用的一个区段集是区段 $[d_{(i-1)}\beta_{(i-1)}, d_{(i)}\beta_{(i)}]$,其中下标 (\cdot) 表示按照从最小倍数变化到最大倍数变化的顺序取得的值,并且具有约定 $d_{(0)}\beta_{(0)} = 0$ 。在这样的实现方式中,网络评分引擎114按照下式来计算正活化度量 PPI^+ 的近似值:

$$[0104] \quad PPI^+ \approx \frac{1}{WN} \sum_{0 < d_i \beta_i} fndr_i d_i \beta_i \quad (8)$$

[0105] 在步骤610,网络评分引擎114生成负活化度量。在某些实现方式中,负活化度量表示SRP指示所观察到的生物实体的激活/抑制与由 d_i 表示的预期变化方向一致的程度。不一致的行为在本文中称为“负活化”。可以使用的一个负活化度量是一个或多个网络被负活化的概率。这样的概率(称为 PPI^-)可以按照以下表达式来计算:

$$[0106] \quad PPI^- = \Pr(\text{Negatively Activated}) = \frac{1}{W} \int_{-W}^0 \Pr(\text{Negatively Activated} | \varphi) d\varphi \quad (9)$$

[0107] 其中

$$[0108] \quad \Pr(\text{Negatively Activated} | \varphi) = \frac{1}{N} \sum_{\varphi < d_i \beta_i < 0} fndr_i \quad (10)$$

[0109] 其中 $fndr_i$ 是以上参照公式1和7所讨论的错误非发现率。如同以上参照正活化度量所讨论的,在某些实现方式中,网络评分引擎114被配置为使用表示 $-W$ 至0之间的 φ 值的区段集在数字上对公式9的表达式求积分。可以使用的一个区段集是区段 $[d_{(i-1)}\beta_{(i-1)}, d_{(i)}\beta_{(i)}]$,其中下标 (\cdot) 表示按照从最小倍数变化到最大倍数变化的顺序取得的值,并且具有约定 $d_{(0)}\beta_{(0)} = 0$ 。在这样的实现方式中,网络评分引擎114按照下式来计算负活化度量 PPI^- 的近似值:

$$[0110] \quad PPI^- \approx \frac{1}{WN} \sum_{d_i \beta_i < 0} fndr_i d_i \beta_i \quad (11)$$

[0111] 在步骤612,网络评分引擎将正活化度量(在步骤608生成)与负活化度量(在步骤610生成)组合以生成称为概率扰动指数或PPI的复合度量。步骤612的组合能够是任何线性的或非线性的组合。在某些实现方式中,PPI是正活化度量与负活化度量的加权线性组合。例如,网络评分引擎114可以被配置用于根据下式来生成PPI:

$$[0112] \quad PPI = \frac{1}{2} (PPI^+ + PPI^-) \quad (12)$$

[0113] 其中 PPI^+ 和 PPI^- 是以上所描述的正活化度量和负活化度量。根据公式12所产生的PPI按以下方式与按照公式5计算出的GPI相关:

$$[0114] \quad GPI = \frac{W}{\sqrt{N}} (PPI^+ - PPI^-) \quad (13)$$

[0115] 另外,网络评分引擎114可以被配置用于通过计算其第i分量按照下式定义的向量的L1范数来计算公式12的PPI:

$$[0116] \quad \left[\frac{1}{2WN} fndr_i d_i \beta_i \right] \quad (14)$$

[0117] 图7是可以由网络评分引擎114实现的预期扰动指数 (EPI) 评分技术的流程图700。如同以上关于SRP引擎110 (图1) 和过程200的步骤212 (图2) 所讨论的,每个SRP表示测量生物实体在处理条件下的活动 (或活动变化)。然后,每个SRP与所测得的活动数量关联,每个测量的生物实体一个。EPI是由SRP表示的在所有生物实体上的平均活动变化的量化。一般地,以SRP表示的测得活动可以从测得活动的分布中随机取得,EPI表示该分布的期望值。如果每个倍数变化 β_i 都是从分布 $p(\cdot)$ 中取得的,则该分布的期望值是

$$[0118] \quad EPI = \int \varphi \cdot p(\varphi) \cdot d\varphi \quad (15)$$

[0119] 由于真实的理论分布 $p(\cdot)$ 不容易知道,因而网络评分引擎114可以被配置用于执行以下所描述的步骤,以基于所观察到的活动以及从系统100中取得的其他信息对EPI值进行近似。

[0120] 在步骤702,网络评分引擎114组构倍数变化向量 β 。这个表示所观察到的N个测量生物实体的活动的倍数变化的倍数变化向量可以如同以上所描述的那样参照图5所示的几何扰动指数 (GPI) 评分技术的步骤502或者图6所示的概率扰动指数 (PPI) 评分技术的步骤602来组构。在步骤704,网络评分引擎114生成倍数变化密度的范围。网络评分引擎114可以如同以上所描述的那样参照图6所示的PPI评分技术的步骤604来生成倍数变化密度的范围。

[0121] 在步骤706,网络评分引擎114识别出倍数变化向量 β 中的每个分量的预期变化方向。该步骤可以如同以上所描述的那样参照图5所示的GPI评分技术的步骤508来执行,从而产生与所观察到的倍数变化 β_i 对应的一组预期变化方向 d_i 。

[0122] 在步骤708,网络评分引擎114生成近似的倍数变化密度。如果从分布 $p(\cdot)$ 中取得的每个倍数变化 β_i ,则分布 $p(\cdot)$ 能够是由下式近似表示:

$$[0123] \quad \hat{p}(\varphi) \propto \begin{cases} \frac{1}{N} \sum_{i|d_i \beta_i > \varphi} \frac{|\beta_i|}{W}, & \varphi > \varepsilon \\ \frac{1}{N} \sum_{i|d_i \beta_i < \varphi} \frac{|\beta_i|}{W}, & \varphi < \varepsilon \end{cases} \quad (16)$$

[0124] 在步骤710,网络评分引擎114生成近似的倍数变化密度的近似期望值,从而产生EPI得分。在某些实现方式中,网络评分引擎114应用计算内插技术 (例如,线性或非线性的内插技术) 来根据公式16的分布生成近似的连续分布,然后使用公式15的公式来计算该分布的期望值。在其他实现方式中,网络评分引擎114被配置用于将公式16的离散分布用作对

连续分布的矩形近似,并且按照下式来计算EPI:

$$EPI \approx \frac{1}{WN} \left[\sum_{d_i \beta_i > 0} (d\beta)_{(i)} \left(\sum_{j=1}^{n^+} (d\beta)_{(j)} \right) ((d\beta)_{(i)} - (d\beta)_{(i-1)}) - \right. \\ \left. \sum_{d_i \beta_i < 0} -(d\beta)_{(i)} \left(\sum_{j=1}^{n^-} -(d\beta)_{(j)} \right) (-(d\beta)_{(i)} - (-(d\beta)_{(i-1)})) \right] \quad (17)$$

[0126] 在公式17中,下标(•)表示按照从最小倍数变化到最大倍数变化的顺序取得的值), n^+ 是预期其活动将响应于处理($d_i \beta_i > 0$)而增大的实体的数量(按步骤706),而 n^- 是预期其活动将响应于处理($d_i \beta_i < 0$)而减小的实体的数量(按步骤706)。在EPI得分中,值较高的倍数变化会比值较低的倍数变化受到更多的考虑,从而提供具有高特异性的活动的测量结果。

[0127] 网络评分引擎114还可以被配置用于确定在网络得分附近的置信区间。这些置信区间可以由临床医生和研究人员用来评价在网络得分中反映出的实验结果,并且可以在更多的数据处理步骤中由系统100的其他构件(例如,由聚合引擎110)使用。用于确定置信区间的一种有用的方法是对于给定的第一类(假阳性(false positive))错误风险 α (例如, $\alpha = 0.05$)来评价关于网络得分为0(或者表示在处理与控制条件之间没有活动差异的其他适当的空值)的原假设。在某些实现方式中,网络评分引擎114使用计算引导技术(例如,参数或非参数引导技术)来对所计算的度量的分布进行近似。许多这样的引导技术都是本技术领域已知的。当能够作出关于基础分布的少数假定时,则可以有利地采用非参数技术。当基础分布被假定时,则可以有利地采用参数技术。在下文所讨论的实例中,假定 β_i 在原假设下由正常的分布引起,具有零均值以及基于 t_{df} 自由度的样品方差 S_i^2 。网络评分引擎可以通过使用统计估计和检验程序来生成这些量,以及表示 β_i 的t统计量和调整的t统计量,例如,由“limma”R包的线性模型方法生成的t统计量和调整的t统计量,通常用于差分基因表达的分析并且由Smyth在“Linear models and empirical Bayes methods for assessing differential expression in microarray experiments”(Statistical Applications in Genetics and Molecular Biology,3:3,2004)进行了描述,该文章通过引用的方式全文并入本文。例如,为了确定EPI得分的置信区间(如同以上参照图7所讨论的),网络评分引擎114可以被配置用于实现参数引导技术以对 β_i 的分布进行近似,假定 β_i 从基础的正常分布中产生。在其中看起来违反了对于百分位自举的应用的假定的实现方式中(该应用可以包括或包含EPI),网络评分引擎114可以另外应用由Efron在“The jackknife, the bootstrap, and other resampling plans”(SIAM,1982)中以及由Diciccio等人在“A review of bootstrap confidence intervals”(Journal of the Royal Statistical Society,50:338,1988)中进行描述的纠偏百分位方法,这两篇文章每篇都以引用的方式全文并入本文。

[0128] 在某些实现方式中,网络评分引擎114可以采用分析方法来确定置信区间,代替引导技术或者与其结合。由网络评分引擎114用来在分析上确定置信区间的特定技术将取决于所使用的特定的网络评分技术以及对于 β_i 的基础统计分布的假定。

[0129] 例如,当网络评分引擎114被配置用于计算强度得分(根据公式1)时,网络评分引

擎114将强度得分视为由独立的近似正态的随机变量的加权和构成的随机变量。结果,强度得分的分布是近似正态的随机变量,具有零均值以及按照下式来计算出的方差

$$[0130] \quad S_{strength}^2 = \frac{1}{N^2} \sum_i S_i^2 \quad (18)$$

[0131] 网络评分引擎114能够使用方差 $S_{strength}^2$ 来按照下式求出t统计量

$$[0132] \quad t = \frac{strength}{S_{strength}} \quad (19)$$

[0133] 其自由度df以由Satterthwaite在“An approximate distribution of estimates of variance components” (Biometrics, 2:110, 1946) 中以及由Welch在“The generalization of student’s problems when several different population variances are involved” (Biometrika, 34:28, 1947) 中描述的Welch-Satterthwaite公式来估计,这两篇文章中的每一篇都以引用的方式全文并入本文。使用这些量,网络评分引擎114可以按照下式生成强度得分的 $(1-\alpha)$ 的置信区间

$$[0134] \quad strength \pm t_{df}^{\alpha/2} S_{strength} \quad (20)$$

[0135] 作为另一个实例,当网络评分引擎114被配置用于计算GPI得分(如同以上参照图5所讨论的)时,网络评分引擎114还可以被配置用于根据图8的流程图800的步骤来计算GPI得分的置信区间。在步骤802,网络评分引擎114按照下式来执行由公式5表示的GPI得分的一阶泰勒展开,作为 β_i 的函数:

$$[0136] \quad GPI(\beta_1, \dots, \beta_N) = GPI(\hat{\beta}_1, \dots, \hat{\beta}_N) + \sum_i \left. \frac{\partial GPI}{\partial \beta_i} \right|_{\hat{\beta}_i} (\beta_i - \hat{\beta}_i) + O(N^2) \quad (21)$$

[0137] 其中 β_i 是所测得的倍数变化值。GPI得分的一阶泰勒近似保留了前两项而丢弃 $O(N^2)$ 项。

[0138] 在步骤804,网络评分引擎114评估在GPI计算中的 β_i 项的系数是否为 β_i 的函数。这些系数包括或包含预期方向项 d_i 和权重 r_i 。当这些系数不取决于 β_i 的值时,在公式21中的一阶项成为关于 β_i 的常数值,并且网络评分引擎114进行到步骤808。但是,当系数确实取决于 β_i 的值时,网络评分引擎114进行到步骤806以对公式21中的一阶项进行近似。特别地,当加权向量 r 是 β_i 的函数并且所预期的方向项 d_i 不是 β_i 的函数时,一阶项可以表示为:

$$[0139] \quad \frac{\partial GPI}{\partial \beta_i} = \frac{1}{\sqrt{N}} \left(d_i r_i + d_i \beta_i \frac{\partial r_i}{\partial \beta_i} \right) \quad (22)$$

[0140] 特别地,当加权向量 r 为错误非发现率值的向量 $fndr_i$ 时,如同以上参照公式2和图5的步骤504所讨论的,网络评分引擎114可以将下列表达式用于公式22的导数项:

$$\begin{aligned}
 \frac{\partial}{\partial \beta_i} fndr_i(\beta_1, \dots, \beta_N) &= -2 \underbrace{\frac{\partial v_i(\beta_1, \dots, \beta_N)}{\partial \beta_i}}_{\text{term1}} \underbrace{\int_{\frac{|\beta_i|}{S_i}}^{\infty} t_{df}(x) dx}_{\text{term2}} \\
 [0141] \quad &- 2v_i(\beta_1, \dots, \beta_N) \frac{\partial}{\partial \beta_i} \left(\int_{\frac{|\beta_i|}{S_i}}^{\infty} t_{df}(x) dx \right)
 \end{aligned} \quad (23)$$

[0142] 在公式23中的标为“term1 (项1)”的导数表示Benjamini-Hochberg调整因子的导数,而标为“term2 (项2)”的积分表示第i生物实体的倍数变化的p值。因为Benjamini-Hochberg项在p值低时是最相关的,所以网络评分引擎114可以被配置用于在步骤806使项1和项2的乘积近似为0。结果,网络评分引擎114可以应用微积分的基本定理并且使用公式23的导数项的下列近似:

$$[0143] \quad \frac{\partial}{\partial \beta_i} fndr_i(\beta_1, \dots, \beta_N) \approx 2 \operatorname{sgn}(\beta_i) \frac{v_i(\beta_1, \dots, \beta_N)}{S_i} t_{df} \left(\left| \frac{\beta_i}{S_i} \right| \right) \quad (24)$$

[0144] 将公式24的近似代入公式21的表达式中得到GPI得分的下列近似:

$$\begin{aligned}
 GPI(\beta_1, \dots, \beta_N) &\approx GPI(\hat{\beta}_1, \dots, \hat{\beta}_N) + \\
 [0145] \quad &(\beta_i - \hat{\beta}_i) \cdot \sum_i \left(d_i fndr_i + d_i \left| \hat{\beta}_i \right| \left[2 \frac{v_i(\hat{\beta}_1, \dots, \hat{\beta}_N)}{\hat{S}_i} t_{df} \left(\left| \frac{\hat{\beta}_i}{\hat{S}_i} \right| \right) \right] \right) \frac{1}{\sqrt{N}}
 \end{aligned} \quad (25)$$

[0146] 在步骤808,网络评分引擎114使用在前述步骤中生成的GPI得分的近似来确定GPI得分的近似方差。如果GPI得分已经被近似为随机变量 β_i 的仿射函数(如在公式21中),则近似的方差将是由下式给出的 β_i 的方差的加权和:

$$[0147] \quad S_{GPI}^2 = \sum_i \left(\frac{\partial GPI}{\partial \beta_i} \right)^2 S_i^2 \quad (26)$$

[0148] 其中 S_i^2 是第i倍数变化 β_i 的方差。因而,公式25的近似的方差可以被写为:

$$[0149] \quad S_{GPI}^2 \approx \sum_i \left(fndr_i + \left| \beta_i \right| \left[2 \frac{v_i(\beta_1, \dots, \beta_N)}{\hat{S}_i} t_{df} \left(\left| \frac{\beta_i}{\hat{S}_i} \right| \right) \right] \right)^2 S_i^2 \frac{1}{N} \quad (27)$$

[0150] 其中当 $d_i = +/ -1$ 时丢弃 d_i 项,因为 $d_i^2 = 1$ 。

[0151] 在步骤810,网络评分引擎114在所观察到的倍数变化值下评价GPI得分的方差(例如,如公式27所表示的)。在步骤812,网络评分引擎114按照下式来生成GPI得分的置信区间:

$$[0152] \quad GPI \pm t_{df}^{\alpha/2} S_{GPI} \quad (28)$$

[0153] 其中 S_{GPI} 如同以上所描述的那样参照公式26和27来计算。公式28可以按照需要来调整以在所观察到的倍数变化值下确定PPI得分的方差。

[0154] 除了以上所描述的标量值得分之外或者作为其的代替,网络评分引擎114还可以生成向量值得分。一个向量值得分是每个测量节点的活动的倍数变化或绝对变化的向量。

[0155] 在某些实现方式中,对于每个扰动(例如,暴露于已知的或未知的物剂),网络评分引擎114都可以生成多个NPA得分。例如,网络评分引擎114可以生成特定网络的、特定的物剂剂量以及特定的暴露时间的NPA得分。

[0156] E. 实验结果

[0157] 用于通过计算网络扰动幅度(NPA)得分来量化生物网络对于扰动的响应的过程200已经被用来使用若干种因果网络模型来分析肿瘤坏死因子(TNF)处理的正常人的支气管上皮(NHBE)细胞。应激响应和免疫响应的转录因子NF- κ B(激活的B细胞的核因子 κ 轻链(kappa-light-chain)增强子)的激活已经被良好地定义为在多种系统中的肿瘤坏死因子 α (TNF α)诱导的信令的主要介质。正常人的支气管上皮(NHBE)细胞以四种不同剂量的TNF α (0.1、1、10和100ng/mL)来处理,并且总的RNA在处理之后的四个不同时间(30分钟、2小时、4小时和24小时)被收集用于微阵列测量。所有处理都与时间匹配的模仿处理对照进行比较,以获得16个对照(4种剂量 \times 4个时间点)。正常人的支气管上皮细胞(Lonza WalkersviUe公司)在标准培养介质(Lonza WalkersviUe公司的Clonetics介质)中进行培养。细胞以TNF α (Sigma)或溶剂对照(HBSS)来处理,并然后在期望的扰动时长之后进行收获。细胞被立即放到冰上并且被划分成使用RNeasy Microkit(Qiagen)从其中提取出总的RNA的三个技术同样的实验。已处理的RNA样品然后被杂混于Affymetrix U133Plus 2.0的微阵列。细胞活力和细胞计数在24小时之后针对所有条件用CellTiter-Glo[®]化验(Promega)来控制。NF- κ B核转位使用Cellomics NF- κ B Activation HCS Reagent Kit(Thermo Scientific)来测量。数据处理和NPA方法在R统计环境中实现。原始RNA表达数据使用可在R统计环境中获得的微阵列分析工具的Bioconductor套件的affy和limma包来分析(Gentleman R: Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer Science+Business Media, 2005; Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J等: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5: R80)。稳健的微阵列分析(RMA)背景校正和分位数标准化被用来生成探测集表达值(Irizarry等, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003, 4: 249-264)。总体线性模型符合于所有复制分组的数据,并且感兴趣的特定对照(“处理的”与“对照”条件的比较)被评价以在表达阵列上生成每个探测集的原始p值。原始p值随后针对多种试验效果使用Benjamini-Hochberg错误发现率(FDR)来校正。

[0158] 探测集使用HG-U133_Plus_2.na30探测集映射及下列准则来与在Selventa知识库中的RNA多度节点匹配。首先,只考虑“at”或“s_at”探测集。其次,丢弃映射到多个基因的探测集。第三,当多个探测集映射到同一基因时,优先选择“at”探测集,而不是“s_at”探测集。最后,当仍然保留有映射到同一基因的多个探测集时,选择在所有感兴趣的对照中具有最

低几何均值FDR校正的p值的探测集。然后针对所有复制分组重新使线性模型符合于只是映射到知识库中的RNA多度节点的那些探测集,并且FDR校正的p值被重新计算。Selventa知识库是含有超过150万个节点(生物学概念和实体)以及超过750万个边(关于节点之间的因果和非因果关系的声明)的库。在Selventa知识库中的声明从同行评审的科学文献以及其他公共和专有数据库中得出。特别地,每个声明都描述来自在人、小鼠和老鼠物种的背景于体外或体内执行的实验的个体实验观察。声明同样采集有关参考来源的信息(例如,列示于MEDLINE中的杂志文章的PubMed ID (PMID)),以及包括物种(人、小鼠或老鼠)以及实验观察由其中得出的组织或细胞系在内的关键背景信息。一个实例因果声明是NF κ B(活化的B细胞的核因子 κ 轻链增强子)增强的转录活动导致CXCL1(趋化因子(C-X-C基元)配体1)的mRNA表达增加[HeLa cell line;Human;PMID 16414985]。知识库含有从健康组织和疾病区(例如,发炎、代谢疾病、心血管损伤、肝损伤和癌症)得出的因果关系。

[0159] GPI、EPI和PPI评分方法首先使用创建为对NF- κ B激活的特定度量的因果网络模型、NF- κ B指向模型来调查。该模型包括已知将由NF- κ B直接调节的155个基因(取自247个不同的参考物,某些基因由多个参照物支持)(其表达以NF- κ B依赖的方式来控制且其启动子序列由NF- κ B直接绑定的基因)。两种评分方法示出了响应于TNF α 的相同模式,已经证明剂量依赖的响应是一直有的,而时间依赖的响应一般在后面的时间变饱和(参见图10a)。EPI方法在性质上与GPI不同,因为EPI得分从2小时到4小时到24小时持续增加,而GPI得分从4小时到24小时达到稳定状态。此外,EPI方法对于0.1ng/mL的TNF α 产生接近零的得分。一般地,EPI得分看起来减少至0(或接近于0)的得分,所述得分通过其他方法会倾向于相对更低。对于EPI方法,已发现除了2小时的时间点外其他所有时间点的最低剂量不是NF- κ B指向网络模型所专有的。

[0160] 然后,将NF- κ B指向模型得分与NF- κ B核转位进行比较。在激活时,NF- κ B被传输到核子之内,在这些核子中它起着调节许多基因的表达的作用。然后,一系列反馈回路导致NF- κ B的后续转位回到细胞质,并且该振荡循环继续进行多次。因为NF- κ B振荡在总体的不同细胞中在稍微不同的时段内发生,所以第一振荡可以是对NF- κ B激活的最可靠的总量测量。尽管第一振荡的时间取决于剂量,在TNF α 处理之后的30分钟可以用于测量所使用的剂量的NF- κ B核转位的现实时间。所有三种评分方法都产生在得分与核转位之间单调的,并且在某些情形中几乎线性的关系,GPI和EPI评分方法的Pearson相关系数为0.85-0.98(图11)。图11示出了在30分钟处的针对30分钟处的NF- κ B核转位绘制的NF- κ B指向NPA得分。在NF- κ B核转位中的误差条表示平均核转位针对相同细胞总体的三个不同视域的标准方差。有趣的是,该剂量依赖的关系在TNF α 处理之后在不同的时间都得到了保持(图13)。这些发现证明,基于因果网络模型的NPA得分能够对NF- κ B转录活动进行量化。

[0161] 因果网络模型的范围和组成对本发明的NPA评分方法的影响同样被调研。首先,评估手动选择已知将由NF- κ B按照TNF α 依赖的方式特别调制的一组测量的效果。子模型根据之前经由逆转录酶的聚合酶链式反应(RT-PCR)来测量的20个基因的集合来构造,以评估响应于在3T3小鼠的成纤维细胞中的TNF α 处理的NF- κ B活动(省略没有直接人类同源性的2个基因)。测量在投以TNF α (范围为从100ng/mL到0.005ng/mL的10种不同浓度)时由3T3细胞中的TNF α 在12小时的时间进程内扰动的这些基因。该子模型产生了与NF- κ B指向模型很相似的激活模式(图14),暗示着其TNF α 依赖的表达尚未被直接验证的基因的纳入对得分的质量

不具有不利影响。图14示出了针对(a) NF- κ B指向模型和(b) 包括20个NF- κ B 调节的基因的子模型使用GPI和EPI来评分的来自TNF α 处理的NHBE细胞的转录数据的结果,所述20个NF- κ B 调节的基因据报告在小鼠3T3的成纤维细胞中(NFKBIA、CASP4、CCL5、TNF A IP3、CCL2、ZFP36、RIPK2、TNFSF10、NFKBIE、IL6、CCL20、ICAM1、TNFRSF1A、TNFRSF1B、SQSTM1、NRG1、SOD1、IL1RL1、HIF1A、ERBB2) 是TNF α 应答的(Tay等, Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. Nature 2010, 466:267-271)。

[0162] 然后,调研使用从与测量不太接近的上游生物实体中得出的因果网络模型的效果。为了这样做,构造了两个附加模型:IKK/NF- κ B 信令模型,该模型包括已知将通过在从IKK B 激酶(IKK) 蛋白质到NF- κ B激活的信令的因果网络模型中的蛋白质的扰动来调制的992个基因(取自414个不同参考物)(图9);以及TNF模型,该模型包括已知将通过以TNF α 来处理细胞来调制的1741个基因(取自589个不同参考物)。虽然NF- κ B指向模型完全包括其表达由单个转录因子(NF- κ B) 直接控制的基因,但是这两个模型每个都含有其直接转录控制器不一定为已知的基因。这些基因的表达可以通过在构造模型时不涉及的转录因子来控制。例如,虽然已知在IKK/NF- κ B 信令模型中的基因将通过在IKK/NF- κ B 信令因果网络模型中的蛋白质的扰动来调制,但是这些基因中的一些能够被调节为由NF- κ B 所直接调制的较小的基因子集的改动表达导致的次级影响。同样地,TNF α 是配体,并且因此不直接转交(mediate)任何基因的转录。以TNF α 处理细胞导致大量转录因子被激活,这些转录因子中任何一个都可以直接地或间接地(例如,通过自分泌信令)改变每个基因在TNF模型中的表达。

[0163] 图9示出了完整因果网络模型(上部),连同基础模型体系架构的示意图(中部)。CHUK、IKBKB和IKBKG充当NFKBIA、NFKBIB和NFKBIE的抑制剂,该NFKBIA、NFKBIB和NFKBIE进而为NFKB1、NFKB2和RELA的抑制剂。在模型中使用的节点被列示于每个部分下方。粗体的节点表示在知识库中具有可测量的下游基因表达的节点,并且可测量的下游基因表达的数量在方括号中给出(因为相同的下游可以在多个节点下找到,所以这1227个下游的可测对象对应于992个唯一的可测量对象)。所使用的符号如下:“CHUK P@S”表示在丝氨酸下磷酸化的CHUK(其中残留物若为已知的则被给出),“CHUK P@ST”表示在丝氨酸或苏氨酸下磷酸化的CHUK(准确的残留物未知),“kaof(CHUK)”表示CHUK的激酶活动,“CHUK:IKBKB”表示CHUK和IKBKB蛋白质的配合物,“IkappaB kinase complex Hs”表示各种IkB 激酶(CHUK、IKBKB和IKBKG)在智人(Hs)中的聚合,“degradationof(NFKBIA)”表示NFKBIA退化的过程,以及“taof(NFKB1)”表示NFKB1的转录活动。

[0164] IKK/NF- κ B信令模型和TNF模型深入洞察了机制假设在与测量间的不同接近水平下的行为。IKK/NF- κ B 信令模型主要包括由NF- κ B(直接地或间接地) 调节的基因(图9),并且它产生很类似于NF- κ B指向模型(图10(b))的响应模式。这种类似的响应模式暗示着在已知将由转录因子直接调节的基因的总层面的行为与其中直接调节的知识为未知的基因的行为之间不存在大的差异。在NF- κ B指向模型中看到的时间及剂量依赖的响应看起来比TNF模型中的稍微较不稳健(图10(c)),例如,在30分钟的时间点处,但是该方法同样产生了很相似的响应。因而,尽管响应的一般模式在模型当中得到良好的保留,但是在与被测量的实体较不接近的模型中也能够观察到较小的但可觉察的响应差异。

[0165] 为了评估因果网络模型特别地响应于相关的TNF α 信令扰动的能力,针对关键细胞

周期分量、转录因子E2F1构建了另一个模型,假定E2F1与NF- κ B相比是TNF α 信令的较不直接的效应剂。E2F1指向模型包括已知将由E2F1直接调节的80个基因(取自54个不同参考物)(表达受E2F1控制,而启动子序列受E2F1绑定)。为了提供不与NF- κ B信令直接相关的生物学的NPA结果的比较,响应于经由CDK抑制剂对细胞周期进程的抑制来评估以上所引入的四个模型(NF- κ B指向、IKK/NF- κ B信令、TNF和E2F1指向)的NPA响应。特别地,公开可用的微阵列数据集被用来以三种不同浓度的CDK抑制剂R547对HCT116结肠癌细胞进行处理(GSE15395)(Berkofsky-Fessler等,Preclinical biomarkers for a cyclin-dependent kinase inhibitor translate to candidate pharmacodynamic biomarkers in phase I patients.Mol Cancer Ther 2009,8:2517-2525)(图12)。所有三种NPA方法都证明剂量及时间依赖性会降低在4小时、6小时和24小时的时间点处的E2F1指向模型的得分。TNF模型示出了与E2F1指向模型相似的响应模式。相比之下,NF- κ B指向模型和IKK/NF- κ B信令模型的得分没有显示出这种相同的剂量和时间依赖的模式,表明这些关注的模型潜在地含有很少的细胞周期调节的基因。

[0166] F. 硬件

[0167] 图15是用于量化生物扰动的影响分布式计算机化系统1500的框图。系统1500的构件与图1的系统100中的那些构件是相同的,但是系统100的布局是这样的:使得每个构件通过网络接口1510来通信。这样的实现方式可能适合于经由包括无线通信系统的多种通信系统进行的分布式计算,该无线通信系统可以共用对共同网络资源的访问,例如,“云计算”范式。

[0168] 图16是计算装置的框图,例如,图1的系统100或者用于执行参照图1-10所描述的过程的图13的系统1300的任意构件。包括SRP引擎110、网络建模引擎112、网络评分引擎114、聚合引擎116以及一个或多个数据库(包括结果数据库、扰动数据库和文献数据库)的系统100的每个构件都可以在一个或多个计算装置1600上实现。在某些方面,多个上述构件和数据库可以包括或包含于一个计算装置1600内。在某些实现方式中,构件和数据库可以跨若干个计算装置1600而实现。

[0169] 计算装置1600包括至少一个通信接口单元、输入/输出控制器1610、系统存储器和一个或多个数据存储装置。系统存储器包括或包含至少一个随机存取存储器(RAM 1602)和至少一个只读存储器(ROM1604)。所有这些元件都与中央处理单元(CPU 1606)通信以促进计算装置1600的操作。计算装置1600可以按照许多不同的方式来配置。例如,计算装置1600可以是常规的独立式计算机,或者作为选择,计算装置1600的功能可以跨多个计算机系统和体系架构分布。计算装置1600可以被配置用于执行建模、评分和聚合操作中的一些或全部。在图10中,计算装置1600经由网络或本地网络链接至其他服务器或系统。

[0170] 计算装置1600可以按照分布式体系架构来配置,其中数据库和处理器被安放于分离的单元或位置。某些此类单元执行初级处理功能并且最低程度地含有通用控制器或处理器和系统存储器。在这方面,这些单元每个都经由通信接口单元1608连结至用作与其他服务器、客户端或用户计算机及其他相关装置间的初级通信链路的通信集线器或端口(未示出)。通信集线器或端口自身可以具有最小限度的处理能力,主要用作通信路由器。多种通信协议可以作为系统的一部分,包括,但不限于:以太网(Ethernet)、SAP、SASTM、ATP、BLUETOOTHTM、GSM和TCP/IP。

[0171] CPU 1606包括处理器,例如,一个或多个常规的微处理器和一个或多个辅助协处理器,例如,用于转移CPU 1606的工作负载的数学协处理器。CPU 1606与通信接口单元1608和输入/输出控制器1610通信,由此CPU 1606与诸如其他服务器、用户终端或装置之类的其他装置通信。通信接口单元1608和输入/输出控制器1610可以包括或包含用于与例如其他处理器、服务器或客户端终端同步通信的多种通信通道。与每个其他通信的装置不需要持续地相互发送信号。相反地,这样的装置只需要在必要时彼此发送信号,可以实际上大部分时间实际都避免交换数据,并且可以需要执行几个步骤来建立装置之间的通信链路。

[0172] CPU 1606同样与数据存储装置通信。数据存储装置可以包括磁存储器、光存储器或半导体存储器的适当组合,并且可以包括或包含例如RAM 1602、ROM 1604、闪存驱动器、光盘(例如,压缩盘)或者硬盘或硬盘驱动器。例如,CPU 1606和数据存储装置每个都可以完全位于单个计算机或其他计算装置之内;或者通过通信介质(例如,USB端口、串口线、同轴线、以太网型网线、电话线、射频收发器或者其他类似的无线或有线介质,或者上述限的组合)彼此连接。例如,CPU1606可以经由通信接口单元1608与数据存储装置连接。CPU 1606可以被配置用于执行一个或多个特定的处理功能。

[0173] 数据存储装置可以存储例如(i)用于计算装置1600的操作系统1612;(ii)适用于根据本文所描述的系统和方法并且尤其是根据针对CPU 1606详细描述的过程来引导CPU 1606的一个或多个应用1614(例如,计算机程序代码或计算机程序产品);或者(iii)适用于存储可以用来存储程序所需的信息的信息的数据库1616。在某些方面,数据库包括或包含用于存储实验数据的以及发表的文献模型的数据库。

[0174] 操作系统1612和应用1614可以按照例如压缩的、不压缩的和加密的格式来存储,并且可以包括或包含计算机程序代码。程序的指令可以从数据存储装置以外的计算机可读介质(例如,ROM 1604或RAM1602)读入处理器的主存储器内。虽然在程序中的指令序列的执行促使CPU 1606执行本文所描述的过程步骤,但是也可以使用硬连线电路来代替用于实现本发明的过程的软件指令或者与其结合。因而,所描述的系统和方法并不限于硬件和软件的任何具体结合。

[0175] 合适的计算机程序代码可以被提供用于执行与本文所描述的建模、评分和聚合相关的一个或多个功能。程序同样可以包括或包含程序单元,例如,操作系统1612、数据库管理系统以及允许处理器经由输入/输出控制器1610与计算机外围装置(例如,视频显示器、键盘、计算机鼠标等)接口连接的“装置驱动器”。

[0176] 本文所使用的术语“计算机可读介质”指的是用于给计算装置1600的处理器(或者本文所描述的装置的任何其他处理器)提供或参与提供用于执行的指令的任何非临时性介质。这样的介质可能要采取许多形式,包括(但不限于)非易失性介质和易失性介质。非易失性介质包括或包含例如光盘、磁盘或光磁盘,或者集成电路的存储器,例如,闪存存储器。易失性介质包括或包含通常构成主存储器的动态随机存取存储器(DRAM)。计算机可读介质的共同形式包括或包含,例如,软盘、柔性盘、硬盘、磁带、任何别的磁介质、CD-ROM、DVD、任何别的光介质、打孔卡、纸带、具有孔图形的任何别的物理介质、RAM、PROM、EPROM或EEPROM(电可擦可编程只读存储器)、FLASH-EEPROM、任何别的存储器芯片或盒,或者计算机能够读取的任何别的非临时性介质。

[0177] 各种形式的计算机可读介质可以涉及将一个或多个指令的一个或多个序列传递

给CPU 1606(或者本文所描述的装置的任何其他处理器),以用于执行。例如,指令最初可以位于远程计算机(未示出)的磁盘上。远程计算机能够将指令装载到其动态存储器内,以及经由以太网连接、电线或甚至是使用调制解调器的电话线来发送指令。位于计算装置1600(例如,服务器)本地的通信装置能够接收在各自的通信线路上的数据,并且将数据放置于处理器的系统总线上。系统总线将数据输送到主存储器,处理器从该主存储器中检索出指令并执行。由主存储器接收到的指令可以在由处理器执行之前或之后任选地存储于存储器内。另外,指令可以经由通信端口作为电信号、电磁信号或光信号来接收,这些信号是用于载送各种类型的信息的无线通信或数据流的示例性形式。更多的方面和实施例在下列段落中进行阐述:

[0178] 1.一种用于量化响应于物剂的生物系统的扰动的计算机化的方法,包括在第一处理器处接收与生物系统对物剂的响应对应的处理数据集,其中生物系统包括或包含多个生物实体,每个生物实体与至少另一个生物实体相互作用;在第二处理器处接收与未暴露于物剂的生物系统对应的控制数据集;在第三处理器处提供用于表示生物系统的计算因果网络模型,并且该计算因果网络模型包括或包含:表示生物实体的节点、表示生物实体之间的关系的边,以及节点的表示在控制数据与处理数据之间的预期变化方向的方向值;以第四处理器来计算节点的表示在处理数据与控制数据之间的差异的活动测量结果;以第五处理器来计算节点的权重值,其中至少一个权重值不同于至少另一个权重值;并且以第六处理器来为计算模型生成表示生物系统对于物剂的扰动的得分,其中该得分基于方向值、权重值和活动测量结果。

[0179] 2.根据段落1所述的计算机化的方法,还包括基于在相应的计算模型中的节点数量来标准化得分。

[0180] 3.根据上述段落中的任一段落所述的计算机化的方法,其中权重值表示在处理数据集和控制数据集中的至少一项中的置信度。

[0181] 4.根据上述段落中的任一段落所述的计算机化的方法,其中权重值包括局部错误非发现率。

[0182] 5.根据段落1所述的计算机化的方法,还包括以第七处理器来计算活动测量结果在节点上的近似分布;以第八处理器来计算近似分布的期望值;以及以第九处理器为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于期望值。

[0183] 6.根据段落5所述的计算机化的方法,其中近似分布基于活动测量结果。

[0184] 7.根据段落5-6中的任一段落所述的计算机化的方法,其中计算期望值包括执行矩形近似。

[0185] 8.根据段落1所述的计算机化的方法,还包括以第十处理器来基于活动测量结果计算正活化得分和负活化得分,正和负活化得分分别表示在活动测量结果与方向值之间的一致性和不一致性;以及以第十一处理器来为每个计算模型生成表示生物系统子集对于物剂的扰动的得分,其中该得分基于正和负活化得分。

[0186] 9.根据段落8所述的计算机化的方法,其中该得分基于本地错误非发现率。

[0187] 10.根据段落8-9中的任一段落所述的计算机化的方法,其中活动测量结果是倍数变化值,并且每个节点的倍数变化值包括由相应节点表示的生物实体的在处理数据与控制数据之间的差异的对数。

[0188] 11. 根据上述段落中的任一段落所述的计算机化的方法, 其中生物系统的子集包括细胞增殖机制、细胞应激机制、细胞发炎机制和DNA修复机制中的至少一个。

[0189] 12. 根据上述段落中的任一段落所述的计算机化的方法, 其中物剂包括通过加热烟草生成的烟雾剂、通过燃烧烟草生成的烟雾剂、烟草烟雾或香烟烟雾中的至少一种。

[0190] 13. 根据上述段落中的任一段落所述的计算机化的方法, 其中物剂包括非均质物质, 包括不存在于生物系统内的或者不是从其中得出的分子或实体。

[0191] 14. 根据上述段落中的任一段落所述的计算机化的方法, 其中物剂包括毒素、治疗化合物、刺激物、弛缓剂、天然产品、制造产品和食物材料。

[0192] 15. 根据上述段落中的任一段落所述的计算机化的方法, 其中处理数据集包括多个处理数据集, 使得每个节点包括由第一概率分布定义的多个倍数变化值以及由第二概率分布定义的多个权重值。

[0193] 虽然本发明的实现方式已经参照具体的实例进行了具体地展示和描述, 但是本领域技术人员应当理解, 在不脱离由所附权利要求所定义的本发明的精神和范围的情况下能够于其中进行形式和细节的各种改变。本发明的范围因而由所附的权利要求所指示, 并且因此希望包含落入权利要求的等价物的意思及范围之内内的所有变化。

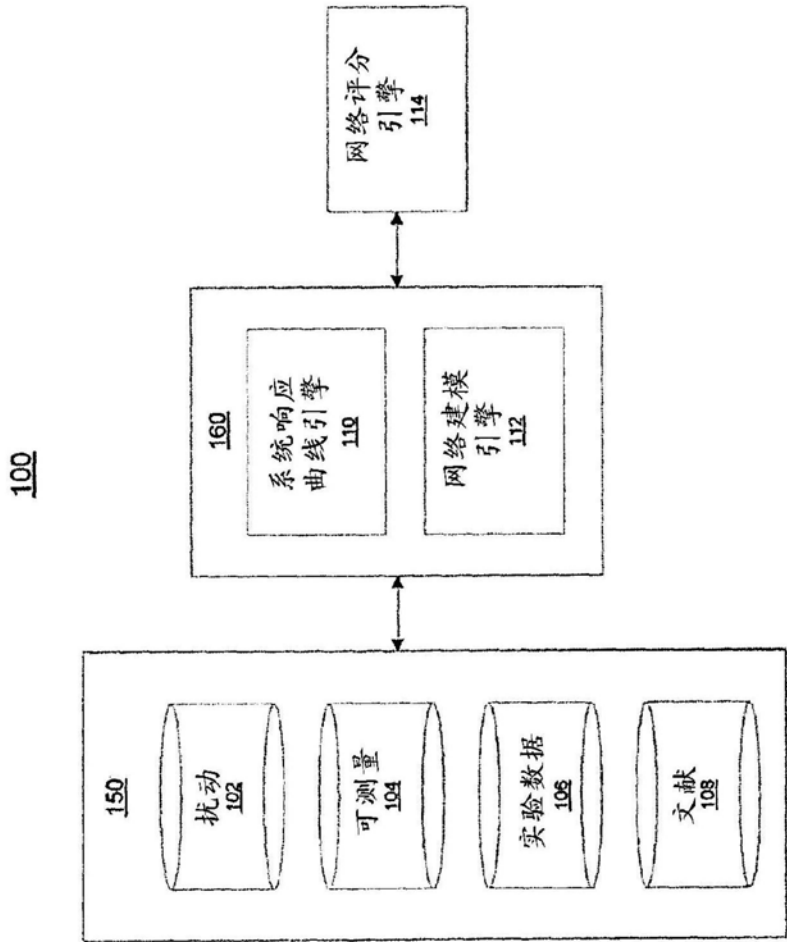


图1

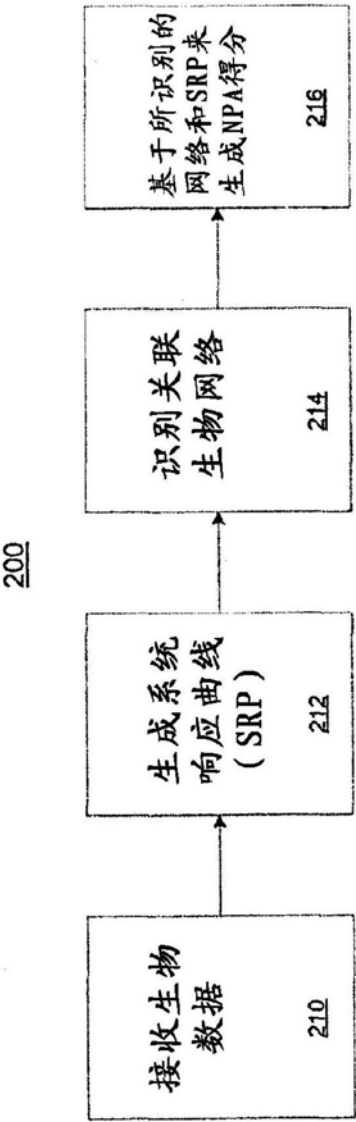


图2

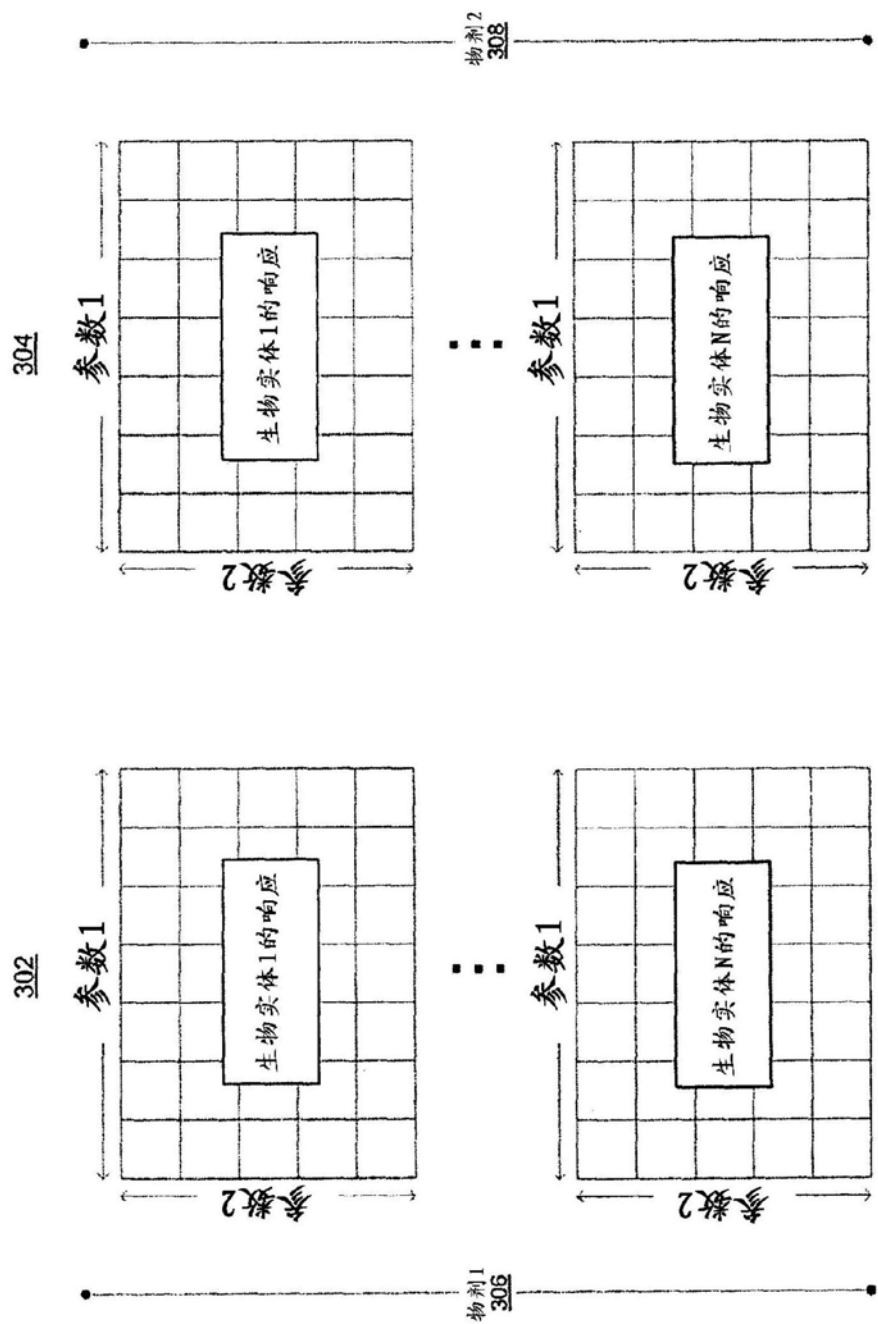


图3

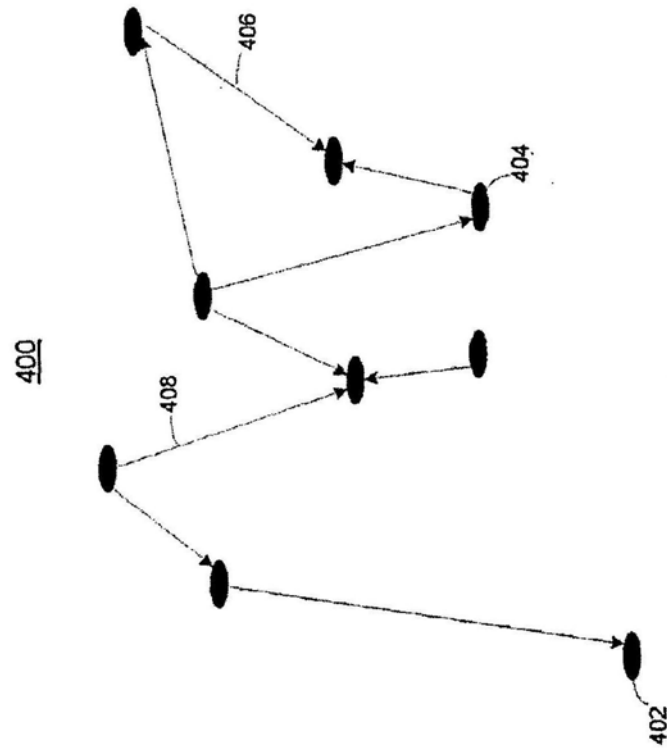


图4

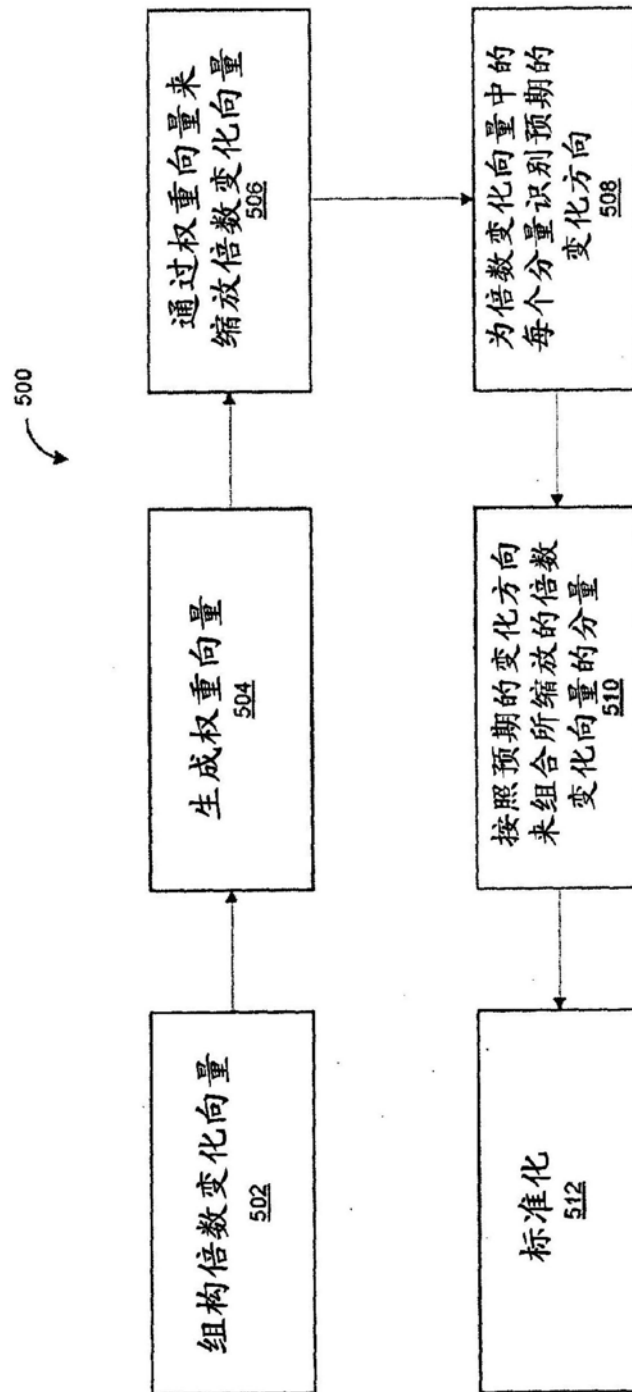


图5

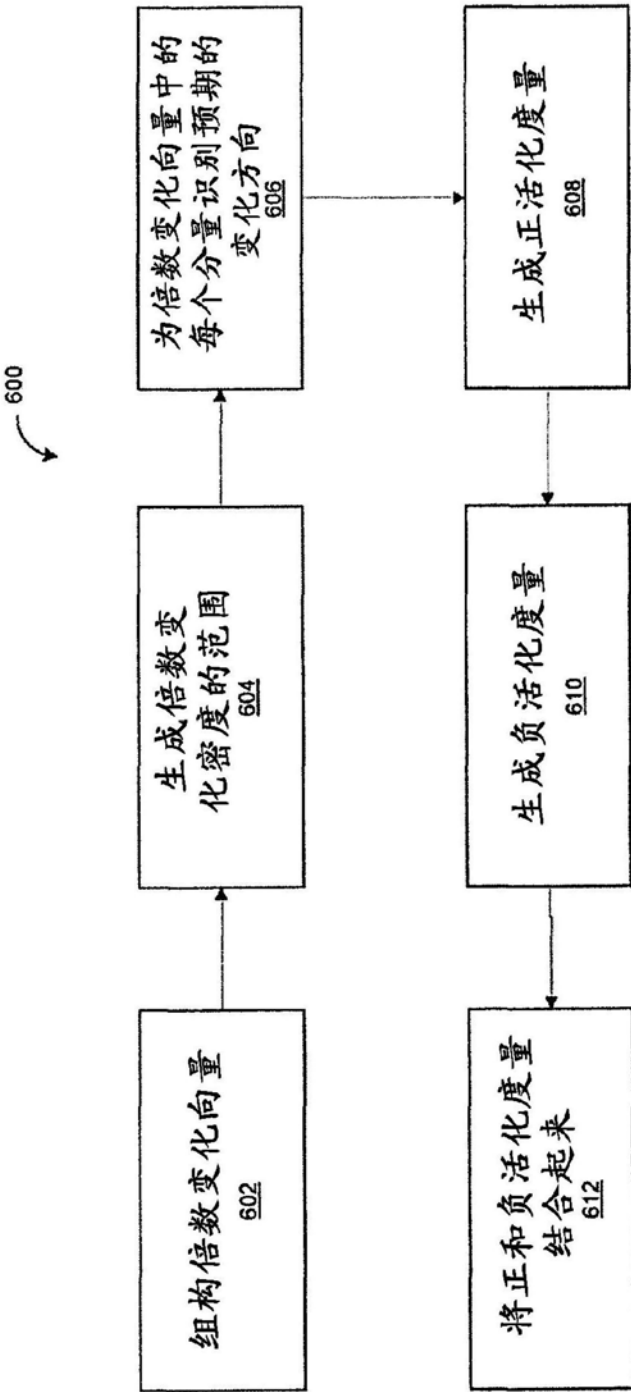


图6

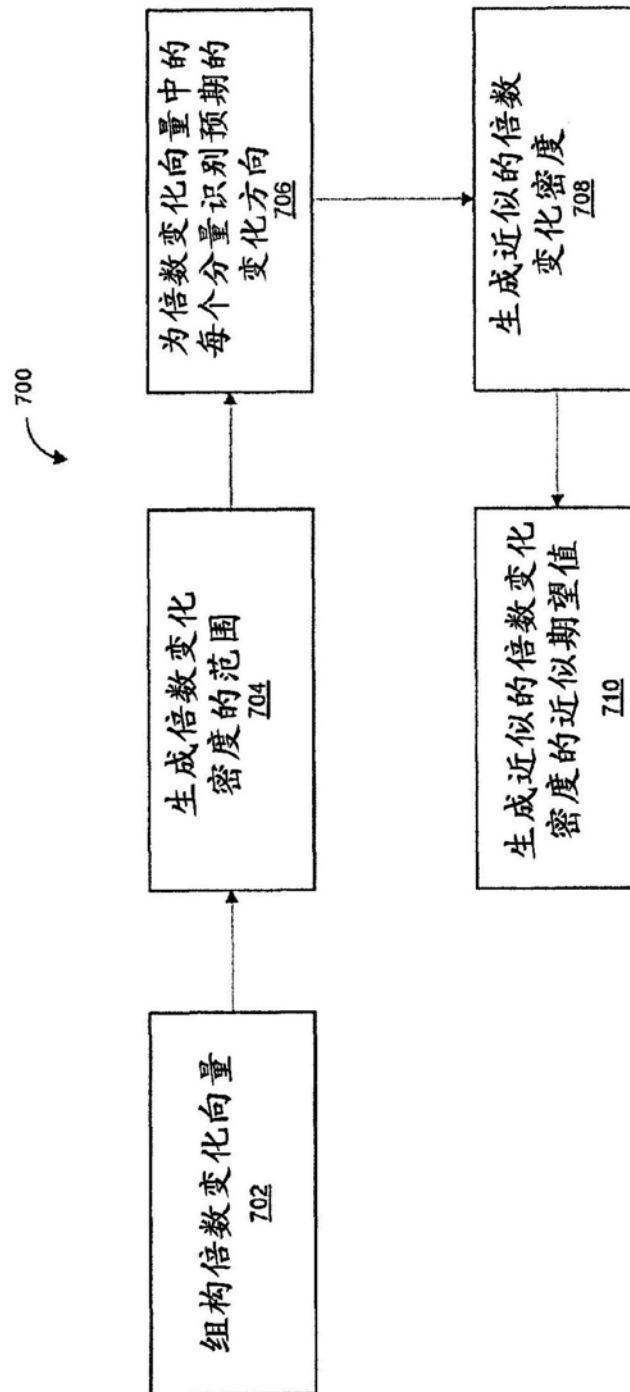


图7

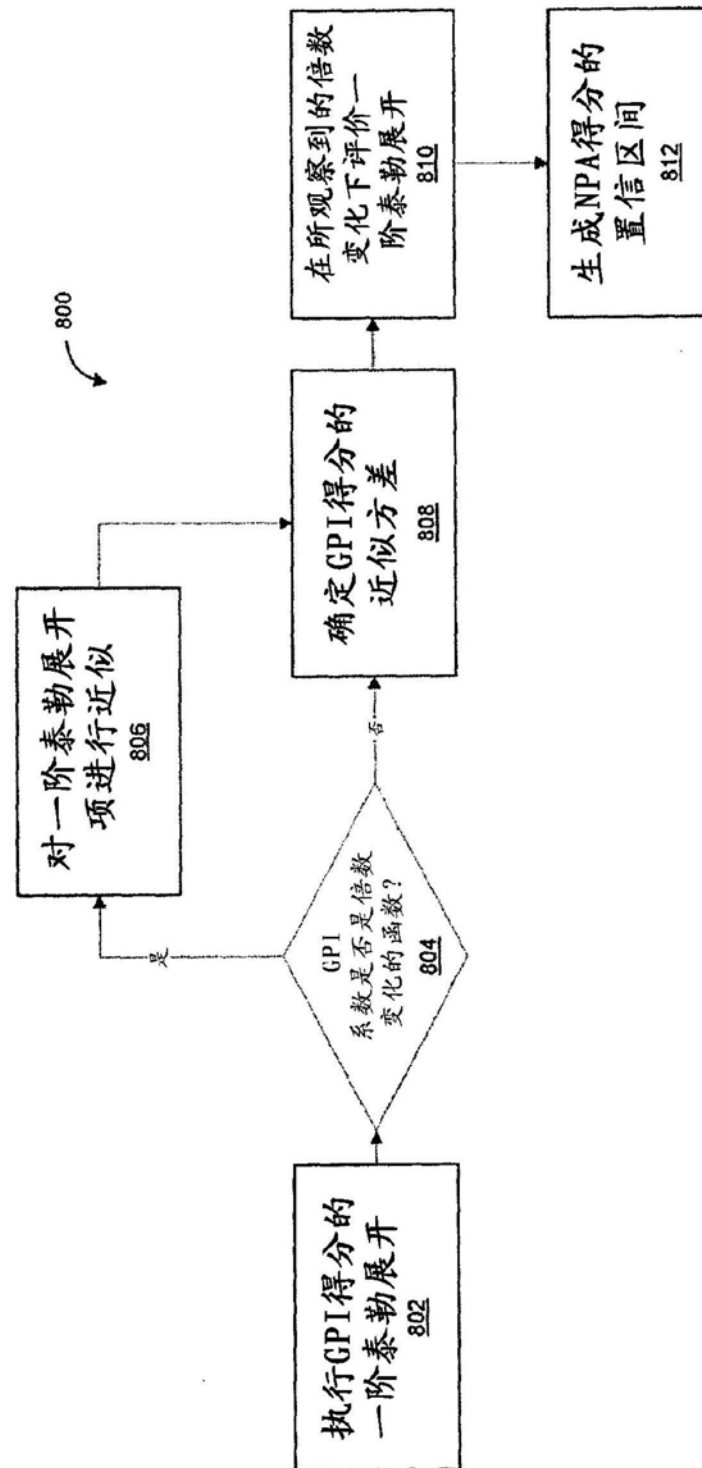


图8

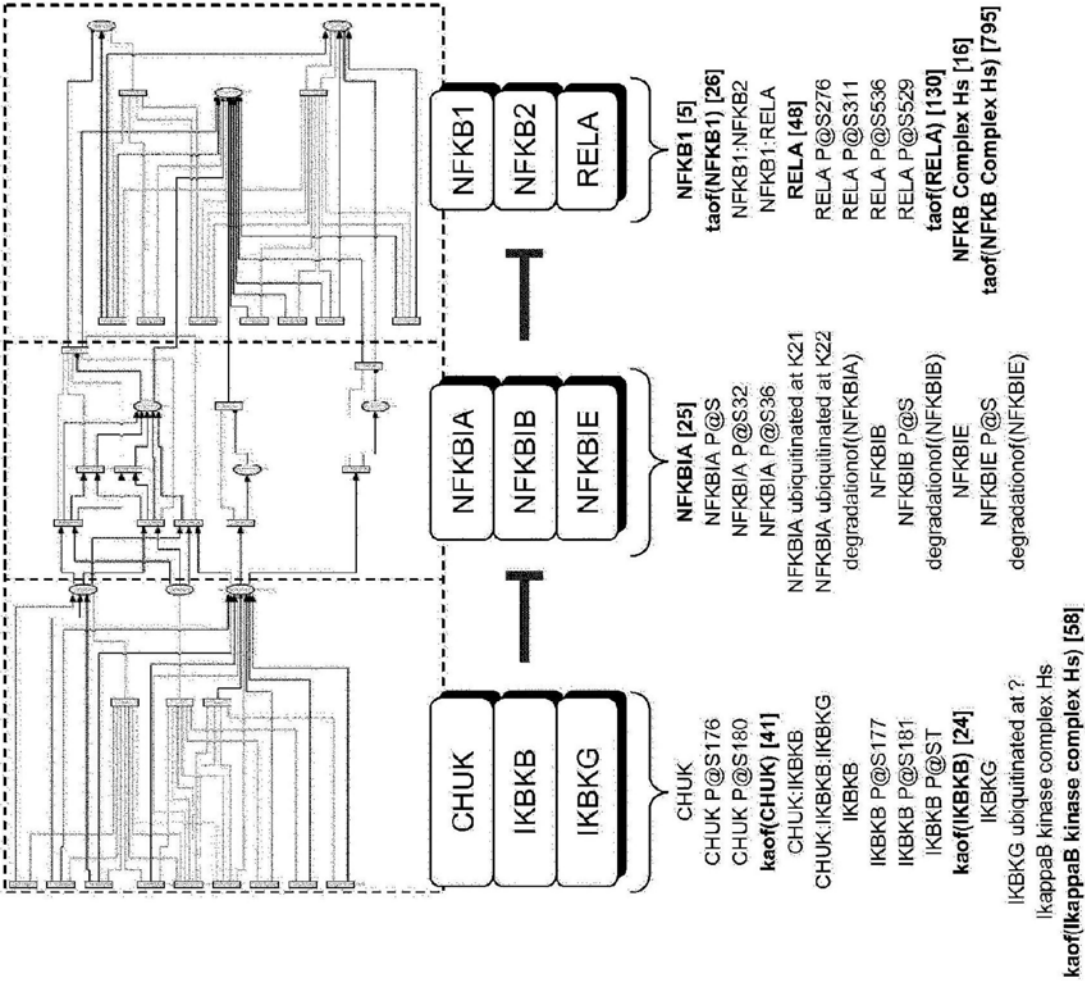


图9

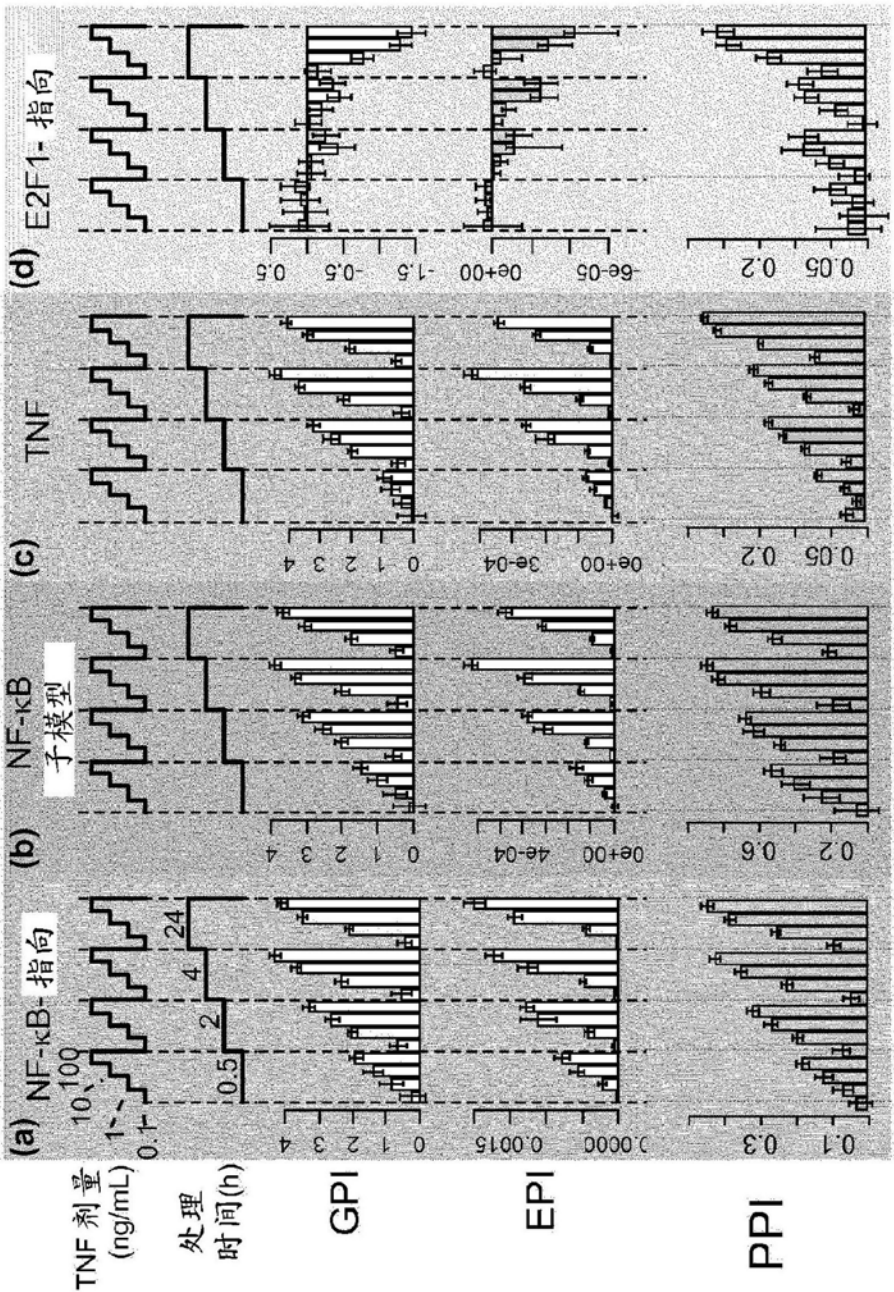


图10

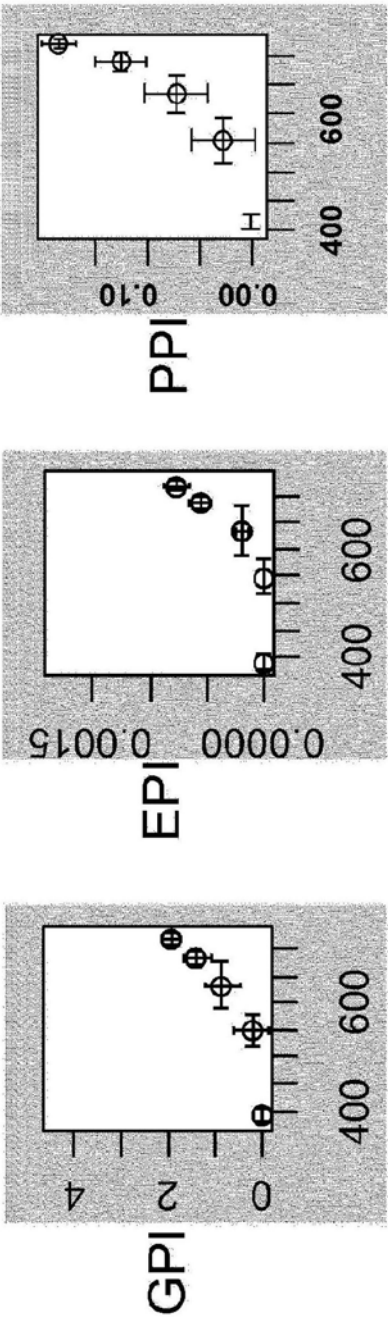


图11

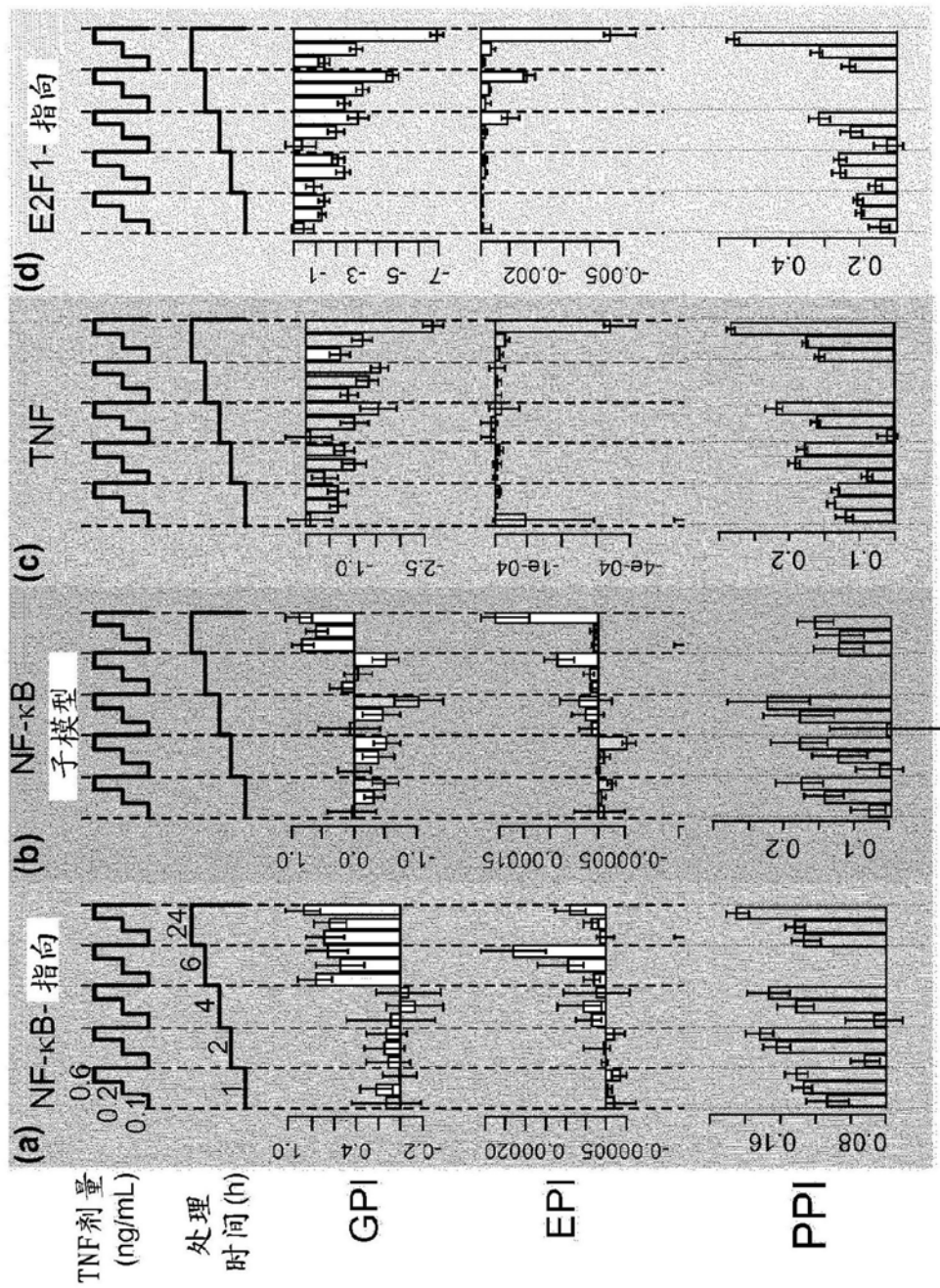


图12

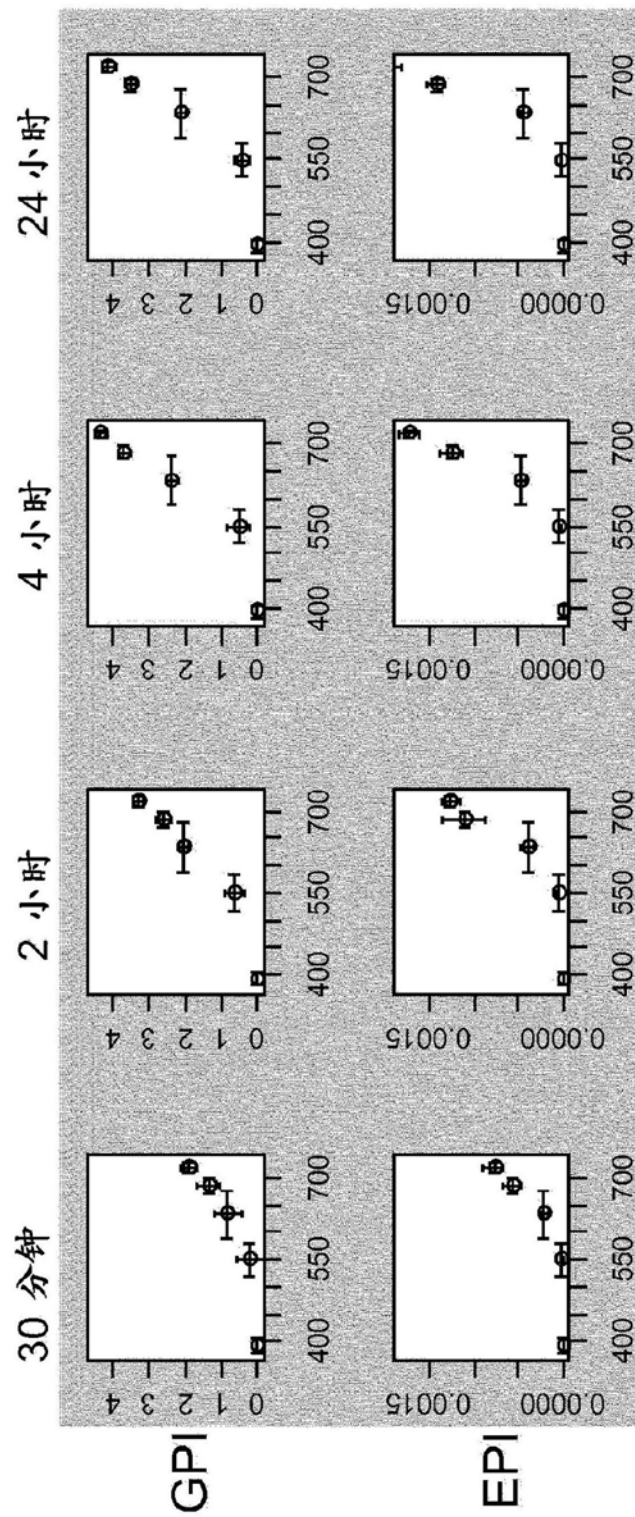


图13

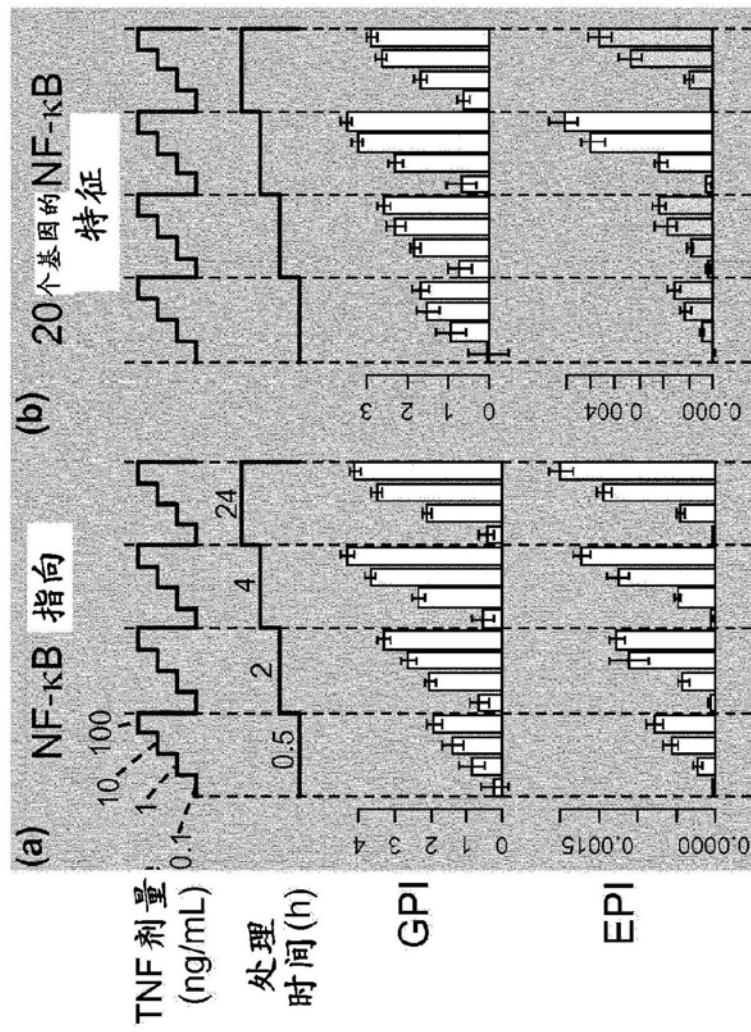


图14

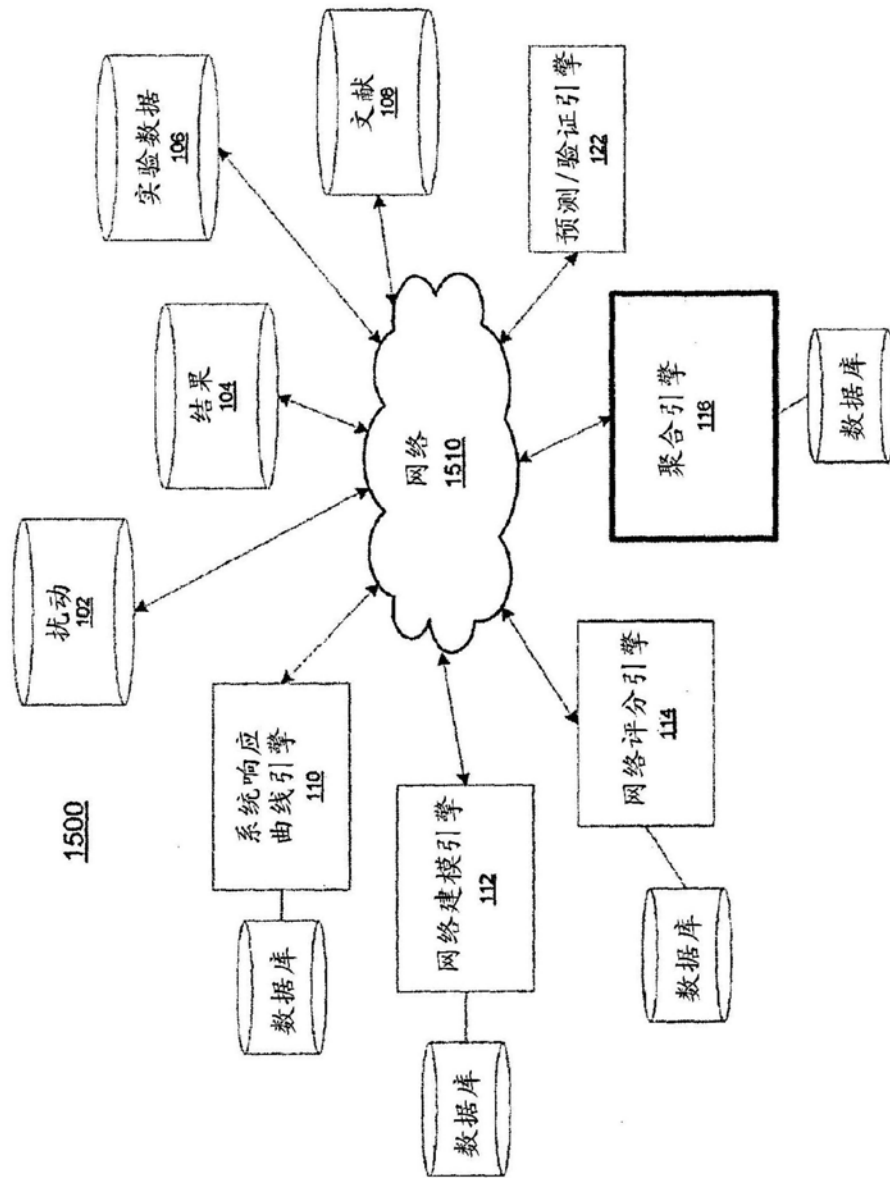


图15

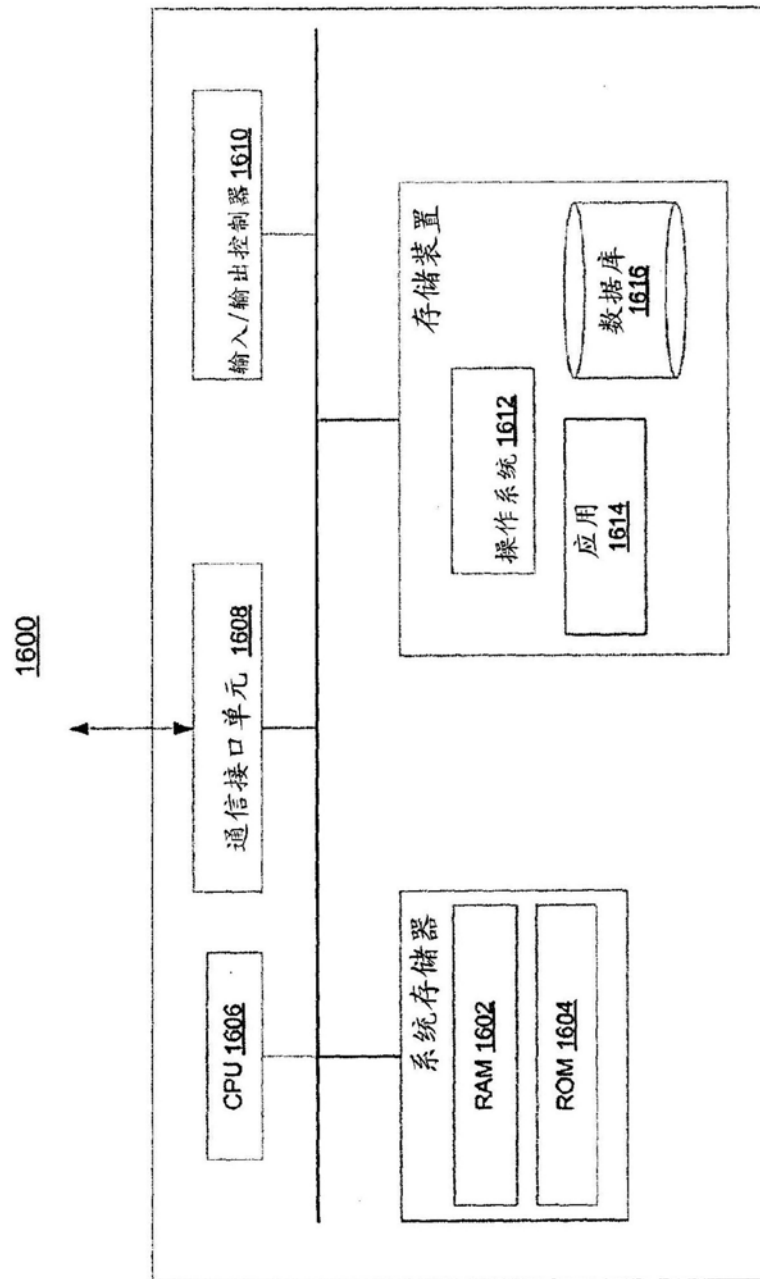


图16