(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109360608 B
(45)授权公告日 2022.05.24

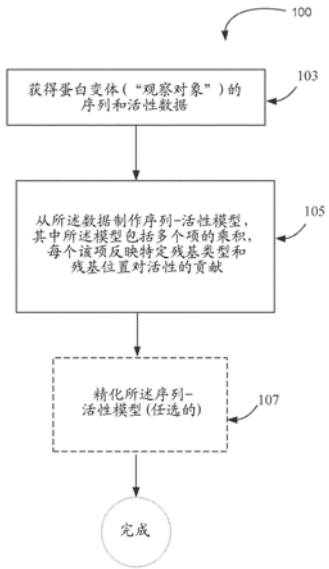
(21)申请号 201811237910.6
(22)申请日 2014.01.29
(65)同一申请的已公布的文献号
申请公布号 CN 109360608 A
(43)申请公布日 2019.02.19
(30)优先权数据
61/759,276 2013.01.31 US
61/799,377 2013.03.15 US
(62)分案原申请数据
201480019389.2 2014.01.29
(73)专利权人 科德克希思公司
地址 美国加利福尼亚州
(72)发明人 格雷戈里·艾伦·科普
尼古拉斯·约翰·阿加德
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 王玮玮 郑霞
(51)Int.Cl.
G16B 30/00(2019.01)
G16B 40/00(2019.01)
G01N 33/50(2006.01)
G01N 33/53(2006.01)
G01N 33/68(2006.01)
(56)对比文件
US 2011257023 A1,2011.10.20
US 2004161796 A1,2004.08.19
CN 101353372 A,2009.01.28
CN 101792934 A,2010.08.04
CN 1672160 A,2005.09.21
审查员 张楠霞
权利要求书3页 说明书66页 附图22页

(54)发明名称

使用相乘形式的模型鉴定生物分子的方法、系统和软件

(57)摘要

本申请提供了用于从复杂的生物分子文库或此类文库的组鉴定具有期望的特性或最适于获得此类特性的生物分子的方法。更特别地,本申请的一些实施方案提供了建立包含相乘项的序列-活性模型并利用这些模型指导定向进化的方法。在一些实施方案中,这些序列-活性模型包括一个或更多个交互作用项,其每个包括表示两个或更多个特定残基对活性的贡献的交互作用系数。在一些实施方案中,这些模型描述了蛋白或核酸序列与蛋白活性之间的关系。在一些实施方案中,本申请还提供了用于制作序列-活性模型的方法,包括但不限于递加或递减技术(stepwise addition or subtraction technique)、贝叶斯回归、集成回归和其他方法。本申请还提供了用于执行本文提供的方法的数字化系统和软件。



CN 109360608 B

1.一种进行定向进化的方法,所述方法包括,

(a)获得多个蛋白变体中每一个的序列和活性数据;

(b)从针对所述多个蛋白变体中的每一个的所述序列和活性数据生成相乘性序列-活性模型,其中所述相乘性序列-活性模型包括:

多个非交互作用相乘项,每一个非交互作用相乘项包括系数和自变量的乘积,所述系数代表蛋白序列或编码所述蛋白序列的核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献,和所述自变量代表蛋白序列或编码所述蛋白序列的核酸序列中特定位置处的特定氨基酸或核苷酸的存在或不存在,所述非交互作用相乘项仅代表特定的氨基酸或核苷酸,并且所述多个非交互作用相乘项通过相乘合并,和

代表所述蛋白变体的活性的因变量;

(c)用所述相乘性序列-活性模型来选择用于一轮定向进化的一个或更多个多肽或多核苷酸。

2.根据权利要求1所述的方法,其中(c)包括:通过评价所述相乘性序列-活性模型的一个或更多个系数鉴定对所述活性有贡献的一个或更多个特定位置处的一个或更多个特定氨基酸或核苷酸,其中所述一个或更多个系数代表所述一个或更多个特定位置的特定氨基酸或核苷酸对活性的贡献。

3.根据权利要求2所述的方法,其中(c)还包括:

选择用于一轮定向进化的一个或更多个突变,其中所述一个或更多个突变处于所述一个或更多个特定位置;以及

选择所述一个或更多个多肽或多核苷酸,其中所述一个或更多个多肽或多核苷酸包含所述一个或更多个突变。

4.根据权利要求2所述的方法,所述方法还包括:对所述一个或更多个多肽或多核苷酸在所述鉴定的一个或更多个位置处进行饱和诱变。

5.根据权利要求2所述的方法,其中所述一个或更多个系数被确定为大于其他系数。

6.根据权利要求1所述的方法,其中(c)包括:

将多个蛋白序列或多个氨基酸序列应用于所述相乘性序列-活性模型并确定所述相乘性序列-活性模型针对所述多个蛋白序列或核酸序列的每一个预测的活性值;以及

通过评价由所述相乘性序列-活性模型针对所述多个蛋白序列或核酸序列预测的活性值,从所述多个蛋白序列或多个氨基酸序列中选择所述一个或更多个多肽或多核苷酸。

7.根据权利要求1-6中任一项所述的方法,所述方法还包括:制备或测定所述一个或更多个多肽或多核苷酸。

8.根据权利要求1-7中任一项所述的方法,所述方法还包括使用所述一个或更多个多肽或多核苷酸进行一轮定向进化。

9.根据权利要求8所述的方法,其中进行一轮定向进化包括片段化并重组被所述模型预测为具有期望的活性水平的所述一个或更多个多肽或多核苷酸。

10.根据权利要求8所述的方法,其中进行一轮定向进化包括对被所述模型预测为具有期望的活性水平的所述一个或更多个多肽或多核苷酸进行饱和诱变。

11.根据权利要求1-10中任一项所述的方法,其中所述非交互作用相乘项以系数×自变量的形式提供。

12.根据权利要求1-10中任一项所述的方法,其中所述非交互作用相乘项以(1+系数×自变量)的形式提供。

13.根据权利要求1-12中任一项所述的方法,其中所述系数被提供于查阅表中。

14.根据权利要求1-13中任一项所述的方法,所述方法还包括对所述一个或更多个多肽或多核苷酸重排。

15.根据权利要求1-14中任一项所述的方法,所述方法还包括形成蛋白变体文库。

16.根据权利要求15所述的方法,所述方法还包括测定活性和对所述蛋白变体文库的成员测序。

17.根据权利要求16所述的方法,所述方法还包括使用所述蛋白变体文库的活性和序列信息以通过对所述蛋白变体文库的活性和序列信息进行操作(b)-(c)来生成并利用新序列-活性模型。

18.根据权利要求1-17中任一项所述的方法,其中所述相乘性序列-活性模型的所述多个相乘项包含至少一个交互作用相乘项,所述交互作用相乘项包括代表(i)所述序列中第一位置处的第一氨基酸或核苷酸和(ii)所述序列中第二位置处的第二氨基酸或核苷酸的特定组合对活性的贡献的交互作用系数,并且

其中所述交互作用系数代表所述特定组合对活性的贡献。

19.根据权利要求18所述的方法,其中生成所述相乘性序列-活性模型包括进行包含交互作用系数的项的递加或递减。

20.根据权利要求18所述的方法,其中生成所述相乘性序列-活性模型包括使用遗传算法选择一个或更多个包含交互作用系数的项。

21.根据前述权利要求1-20中任一项所述的方法,其中生成所述相乘性序列-活性模型包括利用遗传算法精化所述系数的值。

22.根据权利要求21所述的方法,其中生成所述相乘性序列-活性模型包括:

(i)制作一代模型,所述一代模型包含多个模型,其中的每一个模型包含如(b)中所述的多个项的乘积和因变量;

(ii)使用所述一代模型中的每一个模型,预测所述多个蛋白变体中的至少一些的活性;

(iii)基于所述模型准确预测所述多个蛋白变体的活性的能力从所述一代模型中选择一个或更多个模型;

(iv)改进所选择的一个或更多个模型以产生下一代模型;以及

(v)用所述下一代模型多次重复(ii)至(iv)以生成(b)的相乘性序列-活性模型。

23.根据权利要求1-22中任一项所述的方法,其中生成所述相乘性序列-活性模型包括使用先验信息确定所述模型的后验概率分布。

24.根据权利要求1-23中任一项所述的方法,所述方法还包括生成各自具有(b)中形式的相乘性序列-活性模型的集。

25.根据权利要求24所述的方法,所述方法还包括生成集成模型,所述集成模型包括来自两个或更多个序列-活性模型的项,其中所述集成模型的所述项通过所述两个或更多个模型预测活性的能力被加权。

26.根据权利要求25所述的方法,所述方法还包括使用所述集成模型选择用于所述一

轮定向进化的突变。

27.一种进行生物分子的定向进化的方法,所述方法包括:

(a)获得针对多个生物分子的序列和活性数据,每个生物分子包含在多个序列位置的多种类型的亚单位;

(b)从所获得的数据构建相乘性序列-活性模型,其中

所述相乘性序列-活性模型将活性预测为亚单位的类型和序列位置的函数,

所述相乘性序列-活性模型包含多个非交互作用相乘项的乘积,每一个所述非交互作用相乘项包括哑自变量和系数的乘积,其中(1)所述哑自变量代表特定序列位置处的特定类型的特定亚单位的存在/不存在,(2)所述系数代表所述特定亚单位对活性的贡献,和(3)所述非交互作用相乘项仅代表在特定序列位置处的特定类型的特定亚单位;以及

(c)使用所述相乘性序列-活性模型来选择用于一轮定向进化的一个或更多个生物分子。

28.根据权利要求27所述的方法,其中所述序列为完整的基因组、完整的染色体、染色体片段、相互作用基因的基因序列的集合、基因、蛋白或多糖,或其任何组合。

29.根据权利要求27所述的方法,其中所述亚单位为染色体、染色体片段、单体型、基因、密码子、突变、核苷酸、氨基酸、单糖、脂质、或其任何组合。

30.一种计算机系统,所述计算机系统包括:

一个或更多个处理器;

系统存储器;以及

具有计算机可执行指令存储于其上的一个或更多个计算机可读存储介质,当所述计算机可执行指令被所述一个或更多个处理器执行时,导致所述计算机系统执行根据权利要求1-29中任一项所述的方法。

使用相乘形式的模型鉴定生物分子的方法、系统和软件

[0001] 本申请是申请日为2014年1月29日，申请号为201480019389.2，发明名称为“使用相乘形式的模型鉴定生物分子的方法、系统和软件”的申请的分案申请。

[0002] 相关申请的交叉引用

[0003] 该申请基于35U.S.C.§119(e)要求2013年1月31日提交的题为“鉴定具有相互作用的组分的生物分子的方法、系统和软件”的美国临时专利申请号61/759,276和2013年3月15日提交的题为“使用乘法型模型(MODEL OFMULTIPLICATIVE FORM)鉴定生物分子的方法、系统和软件”的美国临时专利申请号61/799,377的权益，其被为了所有目的通过引用全部并入本文。

[0004] 背景

[0005] 本公开内容涉及分子生物学、分子进化、生物信息学和数字化系统的领域。更具体地，本公开内容涉及用于通过计算预测生物分子的活性和/或指导定向进化的方法。还提供了用于执行这些方法的系统，包括数字化系统和系统软件。本公开内容的方法在优化用于工业和治疗用途的蛋白方面具备实用性。

[0006] 蛋白设计长久以来被认为是艰巨的任务，只因为一个原因，构成可搜索的序列空间的可能分子的激增。蛋白的序列空间是巨大的并且利用本领域已知的现有方法不可能探索详尽。由于这种复杂性，很多近似的方法已被用于设计更好的蛋白；其中首当其冲的是定向进化法。如今，蛋白的定向进化被通过往往迭代进行的多种高通量筛选和重组方式来主导。

[0007] 并行地，已提出了多种计算技术用于搜索序列-活性空间。虽然每种计算技术在特定情况下具有优势，但是有效地搜索序列空间以鉴定功能性蛋白的新方法是高度期望的。

[0008] 概述

[0009] 本发明提供了用于从复杂的生物分子文库或此类文库的组鉴定具有期望的特性或最适于获得此类特性的生物分子的方法。更特别地，本发明的一些实施方案提供了建立包括多个相乘项的乘积的序列-活性模型并利用该模型指导定向进化的方法。

[0010] 模型的形式

[0011] 本文公开的一些实施方案提供了用于建立序列-活性模型的方法。每个模型包括多个相乘项的乘积。所述相乘项中的至少一些为非交互作用相乘项，其每个包含代表特定的氨基酸或核苷酸对感兴趣的活性的贡献的系数。所述特定氨基酸或核苷酸在蛋白或和核酸序列中特定位置处具有特定的残基类型。每个非交互作用相乘项还包括自变量(通常只有一个自变量)，其可采取哑变量的形式。自变量代表了在特定的序列位置处特定类型的特定氨基酸或核苷酸的存在或不存在。这些相乘项被称为非交互作用项，因为其每一个代表了蛋白或核酸序列中单个位置处的单个残基的贡献。非交互作用项形式的自变量不代表两个或更多个相互作用残基。另外，一些实施方案中的每个序列-活性模型包括代表蛋白变体的活性的因变量，其中模型描述蛋白变体的活性和多个相乘项的乘积之间的关系。

[0012] 在本发明的一些实施方案中，序列-活性模型还包括多个交互作用项，每个交互作用项包括代表(i)蛋白序列中第一位置处的第一氨基酸或核苷酸和(ii)蛋白序列中第二位

置处的第二氨基酸或核苷酸的特定组合对活性的贡献的交互作用系数。通过交互作用系数代表的贡献是独立的并且不同于第一残基和第二残基中的每个的单独贡献。

[0013] 在一些实施方案中，模型通过相乘组合交互作用项。在一些实施方案中，模型通过相加组合交互作用项。在一些实施方案中，模型通过相乘或相加将交互作用项与非交互作用相乘项组合。在一些实施方案中，模型单纯地相乘，将非交互作用项和交互作用项组合成一个乘积。在其他实施方案中，模型包括通过加法与一个或更多个其他项组合的多个项的至少一个乘积。

[0014] 在本发明的一些实施方案中，序列-活性模型具有相加形式，所述相加形式包括一个或更多个非交互作用项和至少一个交互作用项的和。在与相加模型有关的背景下，非交互作用项也被称为线性项，而交互作用项也被称为非线性项或交叉乘积项。每个非交互作用项代表特定序列位置处特定类型的特定残基在蛋白变体文库的训练组中的存在。所述至少一个交互作用项是交叉乘积项，其包含代表一个相互作用残基的存在的一个变量和代表另一个相互作用残基的存在的另一个变量的乘积。

[0015] 在一些实施方案中，非交互作用相乘项和/或交互作用项具有(系数×自变量)的形式。在其他实施方案中，这些项采取(1+系数×自变量)的形式。当保持模型的相乘性质的情况下，本领域技术人员可采用其它的项表达式。在一些实施方案中，系数被提供于查阅表中。

[0016] 在一些实施方案中，所述方法采用核苷酸序列产生这些模型并预测活性，而不使用氨基酸序列。多组核苷酸例如密码子中的变异影响这些核苷酸序列编码的多肽的活性。在一些实施方案中，所述模型可提供取决于所采用的表达肽的宿主的对优先表达(与编码相同氨基酸的其他密码子相比)的密码子的偏好。

[0017] 在一些实施方案中，提供了用于定向进化的方法。虽然定向进化可被应用于蛋白或编码蛋白的核酸，但是在一些情况中，定向进化被应用于上述生物分子和蛋白之外的生物分子。在此类实施方案中，序列-活性模型可被用来表征多个生物分子的活性和序列之间的关系。例如，序列可以是完整的基因组、完整的染色体、染色体片段、相互作用基因的基因序列的集合、基因、核酸序列、蛋白、多糖等的序列。在一个或更多个实施方案中，序列的亚单位可以是染色体、染色体片段、单体型、基因、核苷酸、密码子、突变、氨基酸、碳水化合物(单体碳水化合物、二体碳水化合物、三体碳水化合物或寡聚碳水化合物)、脂质等。

[0018] 在一些实施方案中，用于生物分子的定向进化的方法可如下表征：

[0019] (a)获得针对多个生物分子的序列和活性数据，每个生物分子包含具有多种类型的亚单位和序列位置的序列；

[0020] (b)从所获得的数据构建序列-活性模型，其中所述序列-活性模型将活性预测为亚单位的类型和序列位置的函数，所述序列-活性模型包含多个非交互作用相乘项的乘积，每个非交互作用相乘项包括(1)代表特定序列位置处特定类型的特定亚单位的存在/不存在的哑变量，和(2)代表所述特定亚单位对活性的贡献的系数；以及

[0021] (c)使用所述序列-活性模型鉴定影响生物分子的期望活性的用于变异的特定位置处的特定类型的一个或更多个亚单位。

[0022] 建立和优化模型

[0023] 在符合以上的实施方案的一个或更多个实施中，用于生成序列活性模型的训练组

源自多个蛋白变体,所述多个蛋白变体可作为蛋白文库被提供。蛋白变体文库可包括来自多种来源的蛋白。在一个实例中,成员包括天然存在的蛋白,诸如由一个基因家族的成员编码的蛋白。在另一个实例中,序列包括通过利用基于重组的多样性生成机制获得的蛋白。例如,可对编码用于该目的的一种或更多种天然存在的亲本蛋白的全长或部分的核酸进行DNA片段化介导的重组、合成的寡核苷酸介导的重组或其组合。又在另一个实例中,这些成员可通过实施鉴定系统地变异的(systematically varied)序列的实验设计(DOE)方案获得。

[0024] 在一些实施方案中,针对多个蛋白变体中的每一个的序列和活性数据通过以下获得:(i)测定所述多个蛋白变体中的每一个以确定其活性,和(ii)对所述多个蛋白变体中的每一个测序以确定其序列。所获得的针对每个序列的数据也被称为观察值。总体地,这些观察值构成训练组。

[0025] 一些实施方案提供了用于为最佳地描述序列的活性的模型选择项和系数的技术。注意,经常存在远多于残基之间存在的实际相互作用的可能的成对的或更高阶的交互作用项。因此,为了避免过度拟合,通常只有有限数目的交互作用项被考虑并且所使用的那些非线性项应反映影响活性的相互作用。

[0026] 一些实施方案提供了通过进行交互作用项和/或非交互作用项的逐步添加、逐步减去或逐步相乘生成序列-活性模型的方法。

[0027] 遗传算法处于可被用来生成具有如以上所述的交互作用项或非交互作用项的乘积形式的模型的技术之中。回归技术和遗传算法处于可被用来生成具有如以上所述的交互作用项或非交互作用项的和的形式的模型的技术之中。

[0028] 本公开内容的一方面提供了制作序列-活性模型的逐步方法,所述序列-活性模型能够帮助鉴定影响期望活性的生物分子,所述方法包括:(a)接收针对多个生物分子的序列和活性数据;(b)由所述序列和活性数据制作基础模型,其中所述基础模型将活性预测为序列的亚单位的存在或不存在的函数;(c)通过对所述基础模型加上/减去或乘以至少一个新交互作用项来制作至少一个新模型,其中所述新交互作用项代表两个或更多个相互作用的亚单位之间的相互作用;(d)确定所述至少一个新模型将活性预测为亚单位的存在或不存在的函数的能力;和(e)基于在(d)中确定的所述至少一个新模型的预测活性的能力以及对添加新交互作用项的偏倚来决定是否对所述基础模型加上/减去所述新交互作用项。然后所得到的模型可被用于多种应用,诸如用于蛋白文库的定向进化以鉴定具有期望的生物活性和特性的蛋白。

[0029] 一些实施方案提供了使用遗传算法选择序列-活性模型的一个或更多个项的方法。其他的实施方案提供了使用遗传算法调整系数的值以使模型符合所获得的数据的方法。

[0030] 在一个或更多个实施方案中,包含交互作用项的模型使用贝叶斯回归技术对观察数据拟合,其中先前的知识被用于确定模型的后验概率分布。

[0031] 在一个或更多个实施方案中,生成了两个或更多个新模型,其各自包括至少一个不同的交互作用项。在此类实施方案中,所述方法还包括制作基于两个或更多个新模型的集成模型。所述集成模型包括来自所述两个或更多个新模型的交互作用项。集成模型根据所述两个或更多个新模型预测感兴趣的活性的能力对交互作用项加权。

[0032] 序列-活性模型可通过很多不同的技术由训练组产生。在某些实施方案中，所述模型为回归模型，诸如偏最小二乘模型、贝叶斯回归模型或主成分回归模型。在另一个实施方案中，所述模型为神经网络。

[0033] 除非明确地指出或固有地不相容，精化本文描述的模型和模型的各种形式的方法彼此相容。其可以多种组合使用以生成期望的序列-活性模型。在本发明的一些实施方案中，所生成的模型可被用来指导定向进化。

[0034] 使用模型指导定向进化

[0035] 在某些实施方案中，具有期望的活性的蛋白变体通过定向进化来鉴定。一些实施方案提供了使用所生成的序列-活性模型指导蛋白变体的定向进化的方法。根据以上描述的方法制作并精化的各种序列-活性模型适用于指导蛋白或生物分子的定向进化。作为该过程的部分，所述方法可鉴定待被用来生成新的蛋白变体文库的序列。此类序列包括对以上鉴定的特定残基的变异或是用来随后引入此类变异的前体。可通过进行诱变或基于重组的多样性生成机制来修饰这些序列，以生成新的蛋白变体文库。任一种或两种生成多样性的形式形成定向进化程序的一部分。新的文库还可被用于开发新的序列-活性模型。分析新的蛋白变体文库以评价对特定活性诸如稳定性、催化活性、治疗活性、对病原体或毒素的耐受、毒性等的影响。

[0036] 在一些实施方案中，寡核苷酸或核酸序列的制备通过使用核酸合成仪合成所述寡核苷酸或核酸序列来完成。本发明的一些实施方案包括使用所制备的寡核苷酸或蛋白序列作为用于定向进化的组成模块(building block)进行一轮定向进化。本发明的多个实施方案将重组和/或诱变应用于这些组成部件以产生多样性。

[0037] 作为一个实例，一些实施方案将重组技术用于寡核苷酸。在这些实施方案中，所述方法包括通过评价序列-活性模型的项的系数来选择用于一轮定向进化的一个或更多个突变。所选择的突变基于如由模型预测的其对蛋白活性的贡献代表在特定位置处特定类型的特定氨基酸或核苷酸。在一些实施方案中，突变的选择包括鉴定一个或更多个系数，所述一个或更多个系数被确定为大于其他系数(或以其他方式指示对活性的强烈影响)的，以及选择由如此鉴定的一个或更多个系数代表的特定位置处的特定氨基酸或核苷酸。在一些实施方案中，在根据序列-活性模型选择突变之后，所述方法包括制备含有或编码所述一个或更多个突变的多个寡核苷酸，并使用如此制备的寡核苷进行一轮定向进化。在一些实施方案中，定向进化技术包括组合和/或重组寡核苷酸。可使用寡核苷酸进行重排反应。

[0038] 其他实施方案应用蛋白序列重组技术。在一些实施方案中，所述方法包括鉴定新的蛋白序列或新的核酸序列，以及制备并测定所述新的蛋白或由所述新的核酸序列编码的蛋白。在一些实施方案中，所述方法还包括使用所述新的蛋白或由所述新的核酸序列编码的蛋白作为用于进一步的定向进化的起始点。在一些实施方案中，定向进化过程包括片段化并重组由模型预测的具有期望的活性水平蛋白序列。

[0039] 在一些实施方案中，所述方法基于被所述模型预测为重要的各个突变来鉴定和/或制备新的蛋白或新的核酸序列。这些方法包括：通过评价序列-活性模型的项的系数选择一个或更多个突变，以鉴定对活性有贡献的特定位置处的一个或更多个特定的氨基酸或核苷酸；鉴定包含以上选择的一个或更多个突变的新蛋白序列或新核酸序列，并制备和测定所述新蛋白或由所述新核酸序列编码的蛋白。

[0040] 在其他的实施方案中，所述方法基于整个序列(而不是各个突变)的预测活性来鉴定和/或制备新蛋白或新核酸序列。在这些实施方案的一些中，所述方法包括将多个蛋白序列或多个氨基酸序列应用于序列-活性模型，以及确定所述序列-活性模型预测的对于所述多个蛋白序列或核酸序列中的每一个的活性值。所述方法还包括通过评价由所述序列-活性模型针对多个序列预测的活性值，从以上应用的多个蛋白序列或多个氨基酸序列中选择新蛋白序列或新核酸序列。所述方法还包括制备并测定具有所述新蛋白序列的蛋白或由所述新核酸序列编码的蛋白。

[0041] 一些实施方案包括选择蛋白序列或核酸序列中的一个或更多个位置，以及在如此鉴定的一个或更多个位置处进行饱和诱变。在一些实施方案中，这些位置通过评价序列-活性模型的项的系数来选择，以鉴定对活性有贡献的特定位置处的一个或更多个特定的氨基酸或核苷酸。相应地，在一些实施方案中，一轮定向进化包括在使用所述序列-活性模型选择的位置处对蛋白序列进行饱和诱变。在一些涉及包括一个或更多个交互作用项的模型的实施方案中，所述方法包括在两个或更多个相互作用残基处同时应用诱变。

[0042] 在一些实施方案中，所述方法涉及选择待产生的新的蛋白变体文库中的一个或更多个成员。然后可在表达系统中合成和/或表达这些成员中的一个或更多个。在特定的实施方案中，所述方法以以下方式继续：(i)提供表达系统，所述新的蛋白变体文库的该所选择的成员可从该表达系统表达；以及(ii)表达所述新的蛋白变体文库的该所选择的成员。

[0043] 一些实施方案使用序列-活性模型鉴定骨架或参考序列，在所述骨架或参考序列中，残基被鉴定用于固定或变异。在一些情况中，参考序列是由模型预测具有最高值(或最高值中的一个)的期望的活性的序列。在另一种情形中，参考序列是初始蛋白变体文库的成员。根据参考序列，所述方法可选择实现变异的子序列。另外地或可选地，序列-活性模型按影响期望的活性的次序排列残基的位置(或处于特定位置的特定残基)。

[0044] 本公开内容的另一方面涉及包括机器可读的介质的设备和计算机程序产品，用于执行以上描述的方法和软件系统的程序指令和/或数据排列被提供于所述机器可读的介质上。经常地，程序指令作为用于执行特定方法的操作的代码提供。如果被用来实现本公开内容的特征，数据则可作为数据结构、数据库表、数据对象或其他合适地设置的特定信息提供。本文描述的任何方法或系统可整体地或部分地呈现为被提供在任何合适的机器可读的介质上的此类程序指令和/或数据。

[0045] 本申请提供了以下内容：

[0046] 项目1.一种进行定向进化的方法，所述方法包括，

[0047] (a)获得针对多个蛋白变体中的每一个的序列和活性数据；

[0048] (b)从针对所述多个蛋白变体中的每一个的所述序列和活性数据生成序列-活性模型，其中所述序列-活性模型包括：

[0049] 多个项的乘积，其中所述项中的至少一些的每一个包括代表蛋白或核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献的系数，和

[0050] 代表所述蛋白变体的活性的因变量；

[0051] (c)通过评价所述序列-活性模型的所述项的系数鉴定对所述活性有贡献的所述特定位置处的一个或更多个所述特定氨基酸或核苷酸来选择用于一轮定向进化的一个或更多个突变；

[0052] (d)制备多个寡核苷酸,所述多个寡核苷酸包含或编码(c)中选择的所述一个或更多个突变;以及

[0053] (e)使用(d)中制备的所述多个寡核苷酸进行一轮定向进化。

[0054] 项目2.根据项目1所述的方法,其中选择用于一轮定向进化的突变包括鉴定一个或更多个系数,所述一个或更多个系数被确定为大于其他系数,以及选择由如此鉴定的所述一个或更多个系数代表的特定位置处的特定氨基酸或核苷酸。

[0055] 项目3.根据前述项目中任一项所述的方法,其中制备包含或编码(c)中选择的突变的多个寡核苷酸包括使用核酸合成仪合成所述寡核苷酸。

[0056] 项目4.根据前述项目中任一项所述的方法,其中进行一轮定向进化包括片段化并重组被所述模型预测为具有期望的活性水平的所述蛋白序列。

[0057] 项目5.根据前述项目中任一项所述的方法,其中进行一轮定向进化包括对被所述模型预测为具有期望的活性水平的蛋白序列进行饱和诱变。

[0058] 项目6.一种进行定向进化的方法,所述方法包括:

[0059] (a)获得针对多个蛋白变体中的每一个的序列和活性数据;

[0060] (b)从针对所述多个蛋白变体中的每一个的所述序列和活性数据生成序列-活性模型,其中所述序列-活性模型包括:

[0061] 多个项的乘积,其中所述项中的至少一些的每一个包括代表蛋白序列或编码所述蛋白序列的核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献的系数;和

[0062] 代表所述蛋白变体的活性的因变量;

[0063] (c)通过评价所述序列-活性模型的所述项的系数鉴定对所述活性有贡献的所述特定位置处的一个或更多个所述特定氨基酸或核苷酸来选择一个或更多个突变;

[0064] (d)鉴定包含(c)中选择的所述一个或更多个突变的新蛋白或新核酸序列;以及

[0065] (e)制备并测定所述新蛋白或由所述新核酸序列编码的蛋白。

[0066] 项目7.根据项目6所述的方法,所述方法还包括将(d)中鉴定的所述新蛋白或新核酸序列用作另外的定向进化的起始点。

[0067] 项目8.根据项目6所述的方法,所述方法还包括在(c)中选择的突变的一个或更多个位置处进行饱和诱变。

[0068] 项目9.一种进行定向进化的方法,所述方法包括:

[0069] (a)获得针对多个蛋白变体中的每一个的序列和活性数据;

[0070] (b)从针对所述多个蛋白变体中的每一个的所述序列和活性数据生成序列-活性模型,其中所述序列-活性模型包括:

[0071] 多个项的乘积,其中所述项中的至少一些的每一个包括代表蛋白序列或编码所述蛋白序列的核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献的系数;和

[0072] 代表所述蛋白变体的活性的因变量;

[0073] (c)通过评价所述序列-活性模型的所述项的系数鉴定对所述活性有贡献的所述特定位置处的一个或更多个所述特定氨基酸或核苷酸来选择所述蛋白序列或核酸序列中的一个或更多个位置;以及

[0074] (d)在(c)中鉴定的所述一个或更多个位置处进行饱和诱变。

[0075] 项目10.一种进行定向进化的方法,所述方法包括:

[0076] (a)获得针对多个蛋白变体中的每一个的序列和活性数据；

[0077] (b)从针对所述多个蛋白变体中的每一个的所述序列和活性数据生成序列-活性模型，其中所述序列-活性模型包括：

[0078] 多个项的乘积，其中所述项的至少一些的每一个包括代表蛋白序列或编码所述蛋白序列的核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献的系数；和

[0079] 代表所述蛋白变体的活性的因变量；

[0080] (c)将多个蛋白序列或多个氨基酸序列应用于所述序列-活性模型并确定所述序列-活性模型针对所述多个蛋白序列或核酸序列的每一个预测的活性值；

[0081] (d)通过评价由所述序列-活性模型针对所述多个序列预测的活性值，从应用于(c)中的所述多个蛋白序列或多个氨基酸序列中选择新蛋白序列或新核酸序列；以及

[0082] (e)制备并测定具有所述新蛋白序列的蛋白或由所述新核酸序列编码的蛋白。

[0083] 项目11.根据项目10所述的方法，其中制备具有所述新蛋白序列的所述蛋白或由所述新核酸序列编码的所述蛋白包括合成所述新蛋白或所述新核酸序列。

[0084] 项目12.根据前述项目中任一项所述的方法，其中包括系数的项的每一个包括系数和自变量的乘积，其中所述系数代表限特定位置处的特定氨基酸或核苷酸对活性的贡献，并且所述自变量代表对应所关联的系数的所述特定位置处的所述特定氨基酸或核苷酸的存在或不存在。

[0085] 项目13.根据项目12所述的方法，其中包括所述系数的项的每一个以(1+系数×自变量)的形式提供。

[0086] 项目14.根据前述项目中任一项所述的方法，其中所述系数被提供于查阅表中。

[0087] 项目15.根据前述项目中任一项所述的方法，其中所述序列-活性模型包括代表所述蛋白变体的活性的所述因变量和所述多个项的乘积之间的关系。

[0088] 项目16.根据前述项目中任一项所述的方法，其中进行一轮定向进化包括对包含或编码(c)中选择的所述突变的多个寡核苷酸重排。

[0089] 项目17.根据前述项目中任一项所述的方法，其中进行一轮定向进化包括形成蛋白变体文库。

[0090] 项目18.根据项目17所述的方法，所述方法还包括测定活性和对所述蛋白变体文库的成员测序。

[0091] 项目19.根据项目18所述的方法，所述方法还包括使用所述蛋白变体文库的活性和序列信息以通过对所述蛋白变体文库的活性和序列信息进行操作(b)-(e)来生成并利用新序列-活性模型。

[0092] 项目20.根据前述项目中任一项所述的方法，其中所述序列-活性模型的所述多个项中的至少一个包括代表(i)所述序列中第一位置处的第一氨基酸或核苷酸和(ii)所述序列中第二位置处的第二氨基酸或核苷酸的特定组合对活性的贡献的交互作用系数，并且

[0093] 其中所述交互作用系数代表所述特定组合对活性的贡献。

[0094] 项目21.根据项目20中任一项所述的方法，其中生成所述序列-活性模型包括进行包含交互作用系数的项的递加或递减。

[0095] 项目22.根据项目20所述的方法，其中生成所述序列-活性模型包括使用遗传算法选择一个或更多个包含交互作用系数的项。

[0096] 项目23.根据前述项目中任一项所述的方法，其中生成所述序列-活性模型包括利使遗传算法精化所述系数的值。

[0097] 项目24.根据项目22或23所述的方法，其中生成序列-活性模型包括：

[0098] (i)制作一代模型，所述一代模型中的每一个包含如(b)中所述的多个项的乘积和因变量；

[0099] (ii)使用所述一代模型中的每一个模型，使用所述蛋白变体的所述序列预测所述多个蛋白变体中的至少一些的活性；

[0100] (iii)基于所述模型准确预测所述多个蛋白变体的活性的能力从所述一代模型中选择一个或更多个模型；

[0101] (iv)改进所选择的一个或更多个模型以产生下一代模型；以及

[0102] (v)用所述下一代模型多次重复(ii)至(iv)以生成(b)的序列-活性模型。

[0103] 项目25.根据前述项目中任一项所述的方法，其中生成所述序列-活性模型包括使用先验信息确定所述模型的后验概率分布。

[0104] 项目26.根据前述项目中任一项所述的方法，所述方法还包括生成各自具有(b)中所述的形式的序列-活性模型的集。

[0105] 项目27.根据项目26所述的方法，所述方法还包括生成集成模型，所述集成模型包括来自两个或更多个序列-活性模型的项，其中所述集成模型的所述项通过所述两个或更多个模型预测活性的能力被加权。

[0106] 项目28.根据项目26所述的方法，所述方法还包括使用所述集成模型选择用于所述一轮定向进化的突变。

[0107] 项目29.一种进行生物分子的定向进化的方法，所述方法包括：

[0108] (a)获得针对多个生物分子的序列和活性数据，每个生物分子包含具有多种类型的亚单位和序列位置的序列；

[0109] (b)从所获得的数据构建序列-活性模型，其中

[0110] 所述序列-活性模型将活性预测为亚单位的类型和序列位置的函数，

[0111] 所述序列-活性模型包含多个基本相乘项的乘积，所述基本相乘项的每一个包括(1)代表特定序列位置处的特定类型的特定亚单位的存在/不存在的哑变量，和(2)代表所述特定亚单位对活性的贡献的系数；以及

[0112] (c)使用所述序列-活性模型鉴定影响生物分子的期望活性的用于变异的特定位置处的特定类型的一个或更多个亚单位。

[0113] 项目30.根据项目29所述的方法，其中所述序列为完整的基因组、完整的染色体、染色体片段、用于相互作用基因的基因序列的集合、基因、蛋白或多糖，或其任何组合。

[0114] 项目31.根据项目29所述的方法，其中所述亚单位为染色体、染色体片段、单体型、基因、密码子、突变、核苷酸、氨基酸、单糖、脂质、或其任何组合。

[0115] 项目32.一种计算机系统，所述计算机系统包括：

[0116] 一个或更多个处理器；

[0117] 系统存储器；以及

[0118] 具有计算机可执行指令存储于其上的一个或更多个计算机可读存储介质，当所述计算机可执行指令被所述一个或更多个处理器执行时，导致所述计算机系统执行用于进行

生物分子的定向进化的方法，所述方法包括：

[0119] (a)获得针对多个生物分子的序列和活性数据，每个生物分子包含具有多种类型的亚单位和序列位置的序列；

[0120] (b)从所获得的数据构建序列-活性模型，其中

[0121] 所述序列-活性模型将活性预测为亚单位的类型和序列位置的函数，

[0122] 所述序列-活性模型包含多个基本相乘项的乘积，所述基本相乘项的每一个包括(1)代表特定序列位置处的特定类型的特定亚单位的存在/不存在的哑变量，和(2)代表所述定特定亚单位对活性的贡献的系数；以及

[0123] (c)使用所述序列-活性模型鉴定影响生物分子的期望的活性的用于变异的特定位置处的特定类型的一个或更多个亚单位。

[0124] 这些和其他特征在下文的详述中并且结合以下附图被更详细地描述。

[0125] 附图简述

[0126] 图1A为描绘生成相乘性序列-活性模型的常规过程的流程图。

[0127] 图1B为描绘用于按照本发明的实施方案使相乘性序列-活性模型符合所观察的数据的遗传算法的流程图。

[0128] 图2是描绘生成一代或更多代蛋白变体文库的定向进化程序的流程图，其中所述操作使用相乘性序列活性模型，诸如在图1中获得的那些相乘性序列活性模型中之一，指导蛋白变体文库的生成。所生成的变体文库可提供序列和活性数据以制作一个或更多个新的序列活性模型。

[0129] 图3A-3H是展示比较线性与非线性相加模型的预测能力的实例的图。

[0130] 图3I-3J是显示比较某些相乘模型和相加模型的预测能力的实例的图。

[0131] 图4A-4B示出了实施递加和递减法来制作序列活性模型的程序的流程图。图4A示出了用于制作模型的递加法的具体实例；且图4B示出了用于制作模型的递减法的具体实例。

[0132] 图5示出了根据一个实施方案在序列变体的定向进化中实施贝叶斯回归的程序的流程图。

[0133] 图6示出了根据一个实施方案在序列变体的定向进化中实施集成回归的程序的流程图。

[0134] 图7是根据一个实施方案描绘生成蛋白变体文库的靴襻p-值方法的流程图。

[0135] 图8是示例性数字化装置的示意图。

[0136] 图9A-9C是代表性地显示线性相加模型、非交互作用相乘模型和交互作用相乘模型的预测力的图。

[0137] 详细描述

[0138] I.定义

[0139] 除非本文另外定义，否则本文使用的所有技术和科学术语具有如本领域普通技术人员通常理解的相同含义。包含本文包括的术语的多本科学词典是本领域技术人员熟知并可获得的。与本文描述的方法和材料相似或等同的任何方法和材料在本文公开的实施方案的实践中具备实用性。

[0140] 通过整体地参考说明书更充分地理解下文紧接着定义的术语。这些定义只是为了

描述特定的实施方案并帮助理解本说明书中描述的复杂概念的目的。这些定义并非意图限制本公开内容的全部范围。特别地，应理解，本公开内容不限于所描述的特定序列、组合物、算法、系统、方法学、操作说明和试剂，因为这些可根据本领域技术人员使用它们的背景而变化。

[0141] 如本说明书和所附的权利要求书中使用的单数形式“一个(a)”、“一个(an)”和“该(the)”包括复数的指代对象，除非该部分内容和上下文明确相反指示。因此，例如，提及“一个装置(a device)”包括两个或更多个此类装置的组合，诸如此类。

[0142] 除非另外指明，“或”的连接意图以其作为布尔逻辑算符的正确含义使用，包括交替地选择特征(A或B，其中相互地，选择A则排除B)和一起选择特征(A或B，其中A和B均被选择)。在该文本中的一些地方，术语“和/或”被用于相同的目的，其将不应被解释为意味着，“或”被用来指相互排除的选择项。

[0143] “生物分子(bio-molecule)”或“生物分子(biological molecule)”指通常在生物有机体中发现的分子。在一些实施方案中，生物分子包含具有多个亚单位的聚合生物大分子(即，“生物聚合物”)。典型的生物分子包括但不限于，与天然存在的聚合物诸如RNA(由核苷酸亚单位形成)、DNA(由核苷酸亚单位形成)和肽或多肽(由氨基酸亚单位形成)共有一些结构特征的分子，包括例如RNA、RNA类似物、DNA、DNA类似物、多肽、多肽类似物、肽核酸(PNA)、RNA和DNA的组合(例如，嵌合体(chimeraplast))等。不期望生物分子局限于任何特定的分子，因为任何合适的生物分子在本发明中具备实用性，包括但不限于，例如，脂质、碳水化合物或由一种或更多种遗传学上可编码的分子(例如，一种或更多种酶或酶通路)组成的其他有机分子等。

[0144] 术语“多核苷酸”和“核酸”指，脱氧核糖核苷酸或核糖核苷酸及其单链或双链形式的聚合物(例如，寡核苷酸、多核苷酸等)。这些术语包括但不限于单链、双链或三链的DNA，基因组DNA，cDNA，RNA，DNA-RNA杂合体，包含嘌呤和嘧啶碱基的聚合物，和/或其他天然的、化学上或生物化学上修饰的碱基、非天然的或衍生的核苷酸碱基。以下为多核苷酸的非限制性实例：基因、基因片段、染色体片段、EST、外显子、内含子、mRNA、tRNA、rRNA、核糖体、cDNA、重组多核苷酸、支化的多核苷酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。在一些实施方案中，多核苷酸包括修饰的核苷酸，诸如甲基化的核苷酸和核苷酸类似物、尿嘧啶(uracyl)、其它糖类和连接基团诸如氟核糖(fluororibose)和硫代酯(thioate)和/或核苷酸分支。在一些可选的实施方案中，核苷酸的序列被非核苷酸组分中断。

[0145] 除非被明确限制，该术语包括含有与参考核酸具有相似的结合特性并且以与天然存在的核苷酸相似的方式被代谢的天然核苷酸的已知类似物的核酸。除非另外指明，特定的核酸序列还隐含地包括其保守性修饰的变体(例如，简并密码子的替换)和互补序列以及明确指示的序列。特别地，简并密码子替换可通过生成其中一个或更多个选择的(或全部的)密码子的第三位被混合碱基和/或脱氧肌苷残基取代的序列来实现(Batzer等人(1991) Nucleic Acid Res.19:5081;Ohtsuka等人(1985)J.Biol.Chem.260:2605-2608;Rossolini等人(1994)Mol.Cell.Probes 8:91-98)。术语核酸与例如寡核苷酸、多核苷酸、cDNA和mRNA可互换地使用。

[0146] 术语“蛋白”、“多肽”和“肽”在可互换地用来表示，不管长度或翻译后修饰(例如，

糖基化、磷酸化、脂质化、豆蔻酰化、泛素化等)如何,通过酰胺键共价连接的至少两个氨基酸的聚合物。在一些情形中,所述聚合物具有至少约30个氨基酸残基,并且通常至少约50个氨基酸残基。更通常地,它们含有至少约100个氨基酸残基。这些术语包括常规被认为是全长蛋白或肽的片段的组合物。D-氨基酸和L-氨基酸以及D-氨基酸和L-氨基酸的混合物被包括在该定义内。本文所述的多肽不局限于遗传学编码的氨基酸。事实上,除了遗传学编码的氨基酸,本文所述的多肽可以全部或部分地由天然存在的和/或合成的非编码的氨基酸组成。在一些实施方案中,多肽为全长的原型(ancestral)多肽或亲本多肽的一部分,与全长亲本多肽的氨基酸序列相比包含氨基酸添加或缺失(例如,空位)或取代,同时仍然保持功能活性(例如,催化活性)。

[0147] 如本文所用,术语"纤维素酶"是指能够将纤维素(β-1,4-葡聚糖或β-D-葡糖苷键)水解成较短的纤维素链、低聚糖、纤维二糖和/或葡萄糖的一类酶。在一些实施方案中,术语"纤维素酶"包括β-葡糖苷酶、内切葡聚糖酶、纤维二糖水解酶、纤维二糖脱氢酶、内切木聚糖酶、β-木糖苷酶、阿拉伯呋喃糖酶、α-葡糖醛酸酶、乙酰基木聚糖酯酶、阿魏酸酯酶、和/或α-葡糖醛酸酯酶(alpha-glucuronyl esterase)。

[0148] 在一些实施方案中,术语"纤维素酶"包括半纤维素水解酶,包括但不限于内切木聚糖酶、β-木糖苷酶、阿拉伯呋喃糖酶、α-葡糖醛酸酶、乙酰基木聚糖酯酶、阿魏酸酯酶和α-葡糖醛酸酯酶。"产生纤维素酶的真菌细胞"为表达并分泌至少一种纤维素水解酶的真菌细胞。在一些实施方案中,产生纤维素酶的真菌细胞表达并分泌纤维素水解酶的混合物。"分解纤维素的"、"纤维素水解的"、"纤维素降解的"和相似的术语指协同地起到将纤维素分解成可溶性二糖或低聚糖诸如纤维二糖(其然后被β-葡糖苷酶进一步水解成葡萄糖)的作用的酶,诸如内切葡聚糖酶或纤维二糖水解酶(后者也称为"外切葡聚糖酶")。在一些实施方案中,纤维素酶为选自β-葡糖苷酶(BGL)、1型纤维二糖水解酶(CBH1)、2型纤维二糖水解酶(CBH2)、糖苷水解酶61s(GH61)和/或内切葡聚糖酶(EG)的重组纤维素酶。在一些实施方案中,纤维素酶为选自β-葡糖苷酶(BGL)、1型纤维二糖水解酶(CBH1)、2型纤维二糖水解酶(CBH2)、糖苷水解酶61s(GH61)和/或内切葡聚糖酶(EG)的重组毁丝霉属(Myceliophthora)纤维素酶。在一些另外的实施方案中,纤维素酶为选自EG1b、EG2、EG3、EG4、EG5、EG6、CBH1a、CBH1b、CBH2a、CBH2b、GH61a和/或BGL的重组纤维素酶。

[0149] 本文使用的术语"序列"指包括但不限于以下的任何生物序列的顺序和身份:完整的基因组、完整的染色体、染色体片段、对于相互作用的基因的基因序列的集合、基因、核酸序列、蛋白、多糖等。在一些背景中,序列指蛋白中的氨基酸残基的顺序和身份(即,蛋白序列或蛋白字符串),或指核酸中核苷酸的顺序和身份(即,核酸序列或核酸字符串)。序列可由字符串表示。"核酸序列"指构成核酸的核苷酸的顺序和身份。"蛋白序列"指构成蛋白或肽的氨基酸的顺序和身份。

[0150] "密码子"指三个连续的核苷酸的特定序列,其为遗传密码的一部分并限定蛋白中的特定氨基酸或者蛋白合成的起始或终止。

[0151] "天然序列"或"野生型序列"指从天然存在的来源分离的多核苷酸或多肽。具有与天然形式相同的序列的天然多肽或多核苷酸的重组形式被包括在"天然序列"内。

[0152] 术语"基因"被广泛地用来指DNA或与生物功能有关的其他核酸的任何片段。因此,基因包括编码序列以及任选地其表达所需的调控序列。基因还任选地包括,例如,形成其他

蛋白的识别序列的未被表达的核酸片段。基因可从多种来源获得，包括从感兴趣的来源克隆或从已知的或预测的序列信息合成，并且可包含被设计为具有期望的参数的序列。

[0153] “基序”指在生物分子内或之间的亚单位形式。例如，术语“基序”可被用来指，非编码的生物分子的亚单位形式或编码呈现的生物分子的亚单位形式。

[0154] 术语“染色体”被用来指细胞中存在的DNA及相关蛋白的有组织的结构，包括含有许多基因、调控元件和其它核苷酸序列的单片缠绕的DNA。该术语还被用来指具有结构的DNA序列。

[0155] 在遗传算法的背景中，术语“染色体”被用作模型群中的单个模型(或一组模型参数)的别称。其被如此使用是因为，来自亲代的模型将其参数(或基因)传递至子代模型，这类似亲本染色体将其基因传递至子代染色体的方式。

[0156] “筛选”指确定一个或更多个生物分子的一个或更多个特性的方法。例如，典型的筛选方法包括确定一个或更多个文库的一个或更多个成员的一个或更多个特性的那些筛选方法。“表达系统”是用于表达由基因或其它核酸编码的蛋白或肽的系统。

[0157] “宿主细胞”或“重组宿主细胞”指含有至少一个重组核酸分子的细胞。因此，例如在一些实施方案中，重组宿主细胞表达在天然形式的(即，非重组的)细胞内未发现的基因。

[0158] “定向进化”、“导向进化”或“人工进化”指通过人工选择、突变、重组或其他操作人工地改变一个或更多个生物分子的序列(或表示该序列的字符串)的体外或体内方法。在一些实施方案中，定向进化发生在繁殖性群体中，其中存在(1)个体差异，有一些差异是(2)可遗传的，其中一些差异(3)在适应性(fitness)方面不同。繁殖成功通过选择预先确定的特性诸如有益特性的结果来确定。繁殖性群体可以是例如，物质群体或计算机系统中的虚拟群体。

[0159] 在某些实施方案中，定向进化方法通过重组编码亲本蛋白变体文库的变体的基因来生成蛋白变体文库。所述方法可使用含有编码亲代变体文库的蛋白的序列或亚序列的寡核苷酸。亲代变体文库的一些寡核苷酸可以是密切相关的，只在选择替代性氨基酸的密码子方面不同，所述替代氨基酸被选择为通过与其他变体重组而被改变。所述方法可被执行一个或多个循环，直到实现期望的结果。如果使用多个循环，则每个循环包括筛选步骤以鉴定待用于随后的重组循环的具有可接受表现的那些变体。在一些实施方案中，定向进化方法通过在由序列-活性模型鉴定的特定位置处的定点定向诱变来产生蛋白变体。一些实施方案采用饱和诱变，其中技术人员尝试在特定位点或基因的狭窄区域产生所有可能的(或尽可能接近所有可能的)突变。

[0160] “重排”和“基因重排”指通过重组亲本多核苷酸的片段的集合通过一系列链延伸循环来引入多样性的定向进化方法。在某些实施方案中，一个或更多个链延伸循环是自引发的；即，除了片段自身之外不添加引物而进行。每个循环包括，通过杂交使单链片段退火，随后通过链延伸延长退火的片段，以及变性。在重排进程中，在有时被称作“模板转换”的过程中，增长的核酸链通常暴露于多个不同的退火配偶体(partner)。如本文使用的，“模板转换”指使来自一个核酸的一个核酸结构域与来自第二核酸的第二结构域转换的能力(即，第一核酸和第二核酸在重排程序中充当模板)。

[0161] 模板转换经常产生嵌合序列，所述嵌合序列因不同起源的片段之间的交叉的引入造成。交叉通过转换的模板在多个循环的退火、延伸和变性期间的重组产生。因此，重排通

常导致变体多核苷酸序列的产生。在一些实施方案中，变体序列构成变体的“文库”。在这些文库的一些实施方案中，变体包含来自两个或更多个亲本多核苷酸的序列片段。

[0162] 当使用两个或更多个亲本多核苷酸时，各个亲本多核苷酸是足够同源的以使得来自不同亲本的片段在重排循环中使用的退火条件下杂交。在一些实施方案中，重排允许具有相对有限的同源性的亲本多核苷酸重组。经常地，各个亲本多核苷酸具有不同的和/或独特的结构域和/或感兴趣的其他序列特征。当使用具有不同的序列特征的亲本多核苷酸时，重排可产生高度多样性的变体多核苷酸。

[0163] 多种重排技术在本领域是已知的。参见，例如，美国专利号6,917,882、7,776,598、8,029,988、7,024,312和7,795,030，其都被通过引用全部并入本文。

[0164] “重叠区延伸基因拼接法”或“基因SOEing”为不依赖限制位点重组DNA序列并且直接体内生成突变的DNA片段的基于PCR的方法。通过修饰被并入引物的5’末端的序列，可使任何成对的聚合物链式反应产物在一个末端共有共同序列。在聚合物链式反应的条件下，该共同序列允许来自两个不同片段的链彼此互补并且彼此杂交，形成具有在每个末端通过共有序列的重叠区连接的两个片段的新序列。该重叠区通过DNA聚合酶的延伸产生重组分子。

[0165] “诱变”是将突变引入标准或参考序列，例如亲本核酸或亲本多肽的程序。

[0166] 定点诱变是对于引入突变有用的技术的一个实例，但是任何合适的方法具备实用性。因此，可选地或另外，突变可通过基因合成、饱和随机诱变、残基的半合成组合文库、定向进化、重现性序列重组(recursive sequence recombination)(“RSR”)(参见例如美国专利申请号2006/0223143，其被通过引用全部并入本文)、基因重排、易错PCR和/或任何其他合适的方法来提供。

[0167] 合适的饱和诱变程序的一个实例被描述于美国公布的专利申请号20100093560，其被通过引用全部并入本文。

[0168] “片段”为核苷酸或氨基酸序列的任何部分。片段可利用本领域已知的任何合适的方法产生，包括但不限于，裂解多肽或多核苷酸序列。在一些实施方案中，片段通过使用裂解多核苷酸的核酸酶产生。在一些另外的实施方案中，片段利用化学技术和/或生物合成技术生成。在一些实施方案中，片段包含至少一个亲本序列的亚序列，所述亚序列利用互补核酸的部分链延长生成。

[0169] “亲本多肽”、“亲本多核苷酸”、“亲本核酸”和“亲本”通常被用来指在多样性生成程序诸如定向进化中用作起点的野生型多肽、野生型多核苷酸或变体。在一些实施方案中，亲本本身经由重排或其他多样性生成程序产生。在一些实施方案中，定向进化中使用的突变体与亲本多肽直接相关。在一些实施方案中，亲本多肽在暴露于极端温度、pH和/或溶剂条件时是稳定的并可充当生成用于重排的变体的基础。在一些实施方案中，亲本多肽对于极端温度、pH和/或溶剂条件是不稳定的，并且亲本多肽被演变以制备稳健的变体。

[0170] “亲本核酸”编码亲本多肽。

[0171] 如本文使用的“突变体”、“变体”和“变体序列”指在一些方面与标准或参考序列不同的生物序列。差异可被称为“突变”。在一些实施方案中，突变体为已通过至少一个取代、插入、交换(cross-over)、缺失和/或其他遗传操作被改变的氨基酸(即，多肽)序列或多核苷酸序列。为了本公开内容的目的，突变体和变体不限于特定的产生所述突变体和变体的

方法。在一些实施方案中，突变体或变体序列与亲本序列相比具有增加的、减少的或基本上相似的活性或特性。在一些实施方案中，变体多肽与野生型多肽(例如亲本多肽)的氨基酸序列相比包含一个或更多个已突变的氨基酸残基。在一些实施方案中，在变体多肽占多数时，与亲本多肽相比，所述多肽的一个或更多个氨基酸残基被保持不变、是不变的、或未被突变。在一些实施方案中，亲本多肽被用作生成具有提高的稳定性、活性或其他特性的变体的基础。

[0172] “文库”或“群”指至少两个不同分子、字符串和/或模型，诸如核酸序列(例如，基因、寡核苷酸等)或来自其的表达产物(例如，酶或其他蛋白)的集合。文库或群通常包括很多不同的分子。例如，文库或群通常包括至少约10个不同的分子。大的文库通常包括至少约100个不同的分子，更通常地至少约1000个不同的分子。对于一些应用，文库包括至少约10000或更多个不同的分子。在某些实施方案中，文库包括通过定向进化程序产生的很多变异或嵌合的核酸或蛋白。

[0173] 当来自两种核酸的每一种的序列被组合在子代核酸中时，所述两种核酸被“重组”。当两种核酸均是重组的底物时，所述两种核酸被“直接”重组。

[0174] “选择”是指，其中一种或更多种生物分子被鉴定为具有一种或更多种感兴趣的特性的程序。因此，例如，技术人员可筛选文库以确定一个或更多个文库成员的一种或更多种特性。如果一个或更多个该文库的成员被鉴定为拥有感兴趣的特性，则其被选择。选择可包括对文库成员的分离，但所述分离不是必需的。另外，选择和筛选可以并且经常是同时的。

[0175] “因变量”表示输出或结果，或者被测试以查看其是否是所述结果。“自变量”表示输入或原因，或被测试以查看其是否是所述原因。因变量可被研究以查看其是否随着自变量变化而变化以及变化多少。

[0176] 在如下的简单的随机线性模型中：

[0177] $y_i = a + bx_i + e_i$，

[0178] 其中，项y_i是因变量的第i个值并且x_i是自变量的第i个值。项e_i被称为“误差”且包含不由自变量解释的因变量的变异性。

[0179] 自变量也被称为“预测变量”、“回归量”、“控制变量”、“操纵变量”、“解释变量”或“输入变量”。

[0180] “相加模型”为描述因变量y作为多个自变量x_i的函数之间的关系的模型，其中该模型通过添加多个项将自变量对因变量的贡献组合，每个项包括自变量的表达式。自变量的表达式反映了自变量对因变量的贡献。

[0181] 在一些实施方案中，数学模型被用来描述一个或更多个自变量(IV)和一个因变量(DV)之间关系。该模型可被写为DV＝IV的代数表达式。“代数表达式”可包括变量、系数、常数和运算符号，诸如加号和减号。$4x^2+3xy+7y+5$为二变量代数表达式。

[0182] 在相加模型中，项为被加号或减号隔开的元素。以上实例采取相加模型的形式。其具有四个项$4x^2$、3xy、7y和5。项可由变量和系数($4x^2$、3xy和7y)或常数(5)组成。在代数表达式中，变量可具有多个值来表示系统的变化情况。例如，其可以是表示行驶中的汽车的速度的连续变量，或表示氨基酸类型的具有多个非连续值的离散变量。变量可以是表示实体的存在或不存在，例如特定位置处特定类型的残基的存在或不存在的比特值变量。在以上代数表达式中，变量为x和y。

[0183] 在一些实施方案中，表达式的“项”可以是表达式的由其他符号诸如由乘号划分开的元素。例如，相乘模型具有由乘法联接的项，如下文进一步说明的。

[0184] “系数”指被因变量或含有因变量的表达式乘的标量值。在以上实例中，“系数”为代数表达式中的项的数字部分。在$4x^2+3xy+7y+5$中，第一项的系数为4。第二项的系数为3，且第三项的系数为7。如果一个项只由变量组成，则其系数为1。

[0185] “常数”为在代数表达式中只含有数字的项。即，其为无变量的项。在表达式$4x^2+3xy+7y+5$中，常数项为“5”。

[0186] “线性项”为具有1的次的项，或升至1的幂的单变量。在以上实例中，项7y为线性项，因为其次为1(y^1或仅y)。与之相比，项$4x^2$是二次项，因为x具有2的次，并且3xy为二元二次项，因为x和y各自具有1的次，乘积导致2的次。

[0187] 相加模型可包括线性项和/或非线性项。“线性项”指包括单个自变量和相关系数的乘积的相加模型的项，其中模型的量随着自变量改变而线性地改变。术语“线性模型”或“线性相加模型”指，其项全部为线性项的任何相加模型。其假定自变量之间没有交互作用(即，每个自变量独立地对蛋白的整体适应性做出贡献)。以上描述的简单随机线性模型为线性相加模型的实例。

[0188] 在相加模型的背景中，除非另外指定，“非线性项”、“交叉乘积项”和“交互作用项”可互换使用，并且它们指包含表达式的模型的项，所述表达式包括两个或更多个自变量的乘积。在一些实施方案中，该表达式可以简单地是常数系数乘以乘积：$c_{i,j}\times x_i x_j$。在相加模型的背景中，术语“非线性模型”或“非线性相加模型”指，包括以上意义上的至少一个非线性项的任何相加模型。但是，在该文本中的一些地方，“非线性项”指更广泛的意义，其中项包含多种形式的单自变量，包括但不限于自变量的幂函数或指数函数。

[0189] “相乘模型”是描述因变量y作为多个自变量x_i的函数之间的关系的模型，其中该模型通过乘以多个项将自变量对因变量的贡献组合，每个项包括自变量的表达式。所述自变量的表达式反映了自变量对因变量的贡献。

[0190] 注意，“线性项”和“线性模型”不适用于相乘模型。所以相乘模型只被描述为交互作用或非交互作用，而不是线性或非线性。相乘模型可包括非交互作用项和/或交互作用项。在相乘模型中，“非交互作用项”指包括单个自变量的表达式的模型的项。

[0191] 与相加模型相似，在相乘模型中，“交互作用项”指包括表达式的模型的项，所述表达式包括两个或更多个自变量的乘积。在一些实施方案中，该表达式可以简单地是常数系数乘以乘积：$c_{i,j}\times x_i x_j$。在其他实施方案中，表达式可以是$(1+c_{i,j}\times x_i x_j)$。

[0192] “相互作用的亚单位”指对所模拟的序列的活性具有协同作用的序列的两个或更多个亚单位，所述协同作用独立且不同于亚单位对所模拟的活性的单独作用。

[0193] “乘积”指两个或更多个变量或项相乘的结果。

[0194] “系数”指因变量或含有因变量的表达式乘的标量值。

[0195] “正交的/正交性”指与模型中的其他自变量或其他关系不相关的自变量。

[0196] 术语“序列-活性模型”指描述一方面的生物分子的活性、特征或特性与另一方面的多种生物序列之间的关系的任何数学模型。

[0197] 术语“编码字符串”指对生物分子的一种表示方式，所述表示方式保存了关于该分子的序列/结构信息。在一些实施方案中，编码字符串包含关于变体文库中的序列突变的信

息。生物分子的编码字符串连同生物分子的活性信息可被用作序列活性模型的训练组。生物分子的非序列特性可被储存或以其他方式与针对生物分子的编码字符串相关。

[0198] “参考序列”为从其产生序列变异的序列。在一些情形中，“参考序列”被用来限定变异。此类序列可以是被模型预测为具有期望的活性的最高值(或最高值中的一个)的序列。在另一种情形中，参考序列可以是原始蛋白变体文库的成员。在特定实施方案中，参考序列为亲本蛋白或亲本核酸的序列。

[0199] “训练组”指一个或更多个模型所符合或基于其建立的一组序列活性数据或观察值。例如，对于蛋白序列-活性模型，训练组包括原始的或改进的蛋白变体文库的残基序列。通常，这些数据包括完整的或部分的残基序列信息，以及文库中每个蛋白的活性值。在一些情形中，在训练组中同时提供有多种类型的活性(例如，速率常数数据和热稳定性数据)。所述活性有时是有益的特性。

[0200] 术语“观察值”是关于蛋白或其他生物实体的的信息，所述信息可被用于训练组来生成诸如序列活性模型的模型。术语“观察值”可指任何测序的且测定的生物分子，包括蛋白变体。在特定实施方案中，每个观察值是对于文库中变体的活性值和相关序列。通常，生成序列-活性模型使用的观察值越多，该序列-活性模型的预测力越好。

[0201] 如本文使用的，术语“有益特性”意图指为蛋白或目的组合物或与蛋白有关的方法提供某种益处的表型或其他可鉴定的特征。有益特性的实例包括当与亲本蛋白比较时，变体蛋白的催化特性、结合特性、暴露于极端温度、pH等时的稳定性、对刺激的敏感性、抑制力等特性的增加或降低。其他有益特性可包括响应特定刺激改变的曲线。下文列出了有益特性的另外的实例。有益特性的值可用作用于序列活性模型的训练组中使用的观察对象的活性值。

[0202] “下一代测序”或“高通量测序”是使测序过程并行化的测序技术，一次产生数千计或数百万计的序列。合适的下一代测序方法的实例包括但不限于，单分子实时测序(例如，Pacific Biosciences,Menlo Park,California)、离子半导体测序(例如，Ion Torrent,South San Francisco,California)、焦磷酸测序(例如，454,Branford,Connecticut)、连接测序(例如，SOLid sequencing of Life Technologies,Carlsbad,California)、合成测序和可逆性末端终止测序(例如，Illumina,San Diego,California)、诸如透射电子显微术的核酸成像技术等。在本公开内容的详述中描述了对示例性技术的进一步说明。

[0203] “预测力”指一种模型在多种条件下正确地预测数据的因变量的值的能力。例如，序列活性模型的预测力指模型由序列信息预测活性的能力。

[0204] “交叉验证”指，用于测试模型预测感兴趣的值(即自变量的值)的能力的普遍性的方法。所述方法利用一组数据制作模型，并利用不同的一组数据测试模型的误差。第一组数据被视为训练组，而第二组数据为验证组。

[0205] “系统方差”指以不同的组合被改变的一个项或一组项的不同描述符。

[0206] “系统地变化的数据”指由以不同的组合被改变的一个项或一组项的不同描述符产生、推导或得到的数据。很多不同的描述符可同时但以不同的组合被改变。例如，从其中氨基酸的组合已被改变的多肽收集的活性数据是系统地变化的数据。

[0207] 如本文使用的，术语“系统地变异的序列”指其中每个残基见于多种背景的一组序列。原则上，系统变异的水平可通过序列彼此正交(即，与平均值相比最大程度地不同)的程

度来定量。

[0208] 术语“切换(toggling)”指将多个氨基酸残基类型引入优化的文库中的蛋白变体序列中的特定位置。

[0209] 术语“回归”和“回归分析”指用来理解自变量中与因变量有关的那些自变量,并被用来探索这些关系的形式的技术。在有限的情形中,回归分析可被用来推导自变量和因变量之间的因果关系。它是用于估计变量之间的关系的统计技术。当焦点是关于因变量和一个或更多个自变量之间的关系时,其包括很多用于对若干个变量建模和分析的技术。更特别地,回归分析帮助技术人员理解当任一个自变量变化而其他自变量保持固定时,因变量的典型值(typical value)如何改变。回归技术可被用于从包括多个观察值的训练组生成序列活性模型,所述多个观察值可包括序列和活性信息。

[0210] 偏最小二乘法或PLS是通过将预测变量(例如,活性)和可观察变量(例如,序列)投射到新的空间而发现线性回归模型的方法族。PLS也被称为潜在结构投射法。X(自变量)和Y(因变量)数据均被投射至新的空间。PLS被用来找出两个矩阵(X和Y)之间的基本关系。潜在变量方法被用来对X和Y空间中的协方差结构建模。PLS模型将尝试找出在X空间中的多维方向,所述多维方向解释Y空间中的最大多维变化的方向。当预测器(predictor)的矩阵具有比观察值更多的变量时并且当在X中存在多重共线性时,PLS回归是特别适合的。

[0211] “描述符”指用来描述或辨识一个项的事物。例如,字符串中的字符可以是该字符串表示的多肽中的氨基酸的描述符。

[0212] 在回归模型中,因变量通过项的和与自变量相关。每个项包括自变量和相关回归系数的乘积。在纯粹的线性回归模型的实例中,回归系数由以下表达形式中的β表示:

[0213] $y_i=\beta_1 x_{i1}+\ldots+\beta_p x_{ip}+\varepsilon_i=x_i^T\beta+\varepsilon_i$

[0214] 其中y_i是因变量,x_i是自变量,ε_i是误差变量,并且T表示转置矩阵(transpose),即向量x_i和β的内积。

[0215] “主成分回归”(PCR)指估计回归系数时使用主成分分析的回归分析。在PCR中,使用自变量的主成分,而不是直接将因变量对自变量回归。PCR在回归中通常只使用主成分的子集。

[0216] “主成分分析”(PCA)指使用正交变换将可能相关的变量的一组观察值转换成称为主成分的线性不相关的变量的一组值的数学程序。主成分的数目小于或等于最初变量的数目。该转换以使得第一主成分具有最大的可能方差(即,在数据中占尽可能多的变化性),并且每个之后的成分在其与之前的成分正交(即,与之不相关)的约束下转而具有最高的可能方差的方式被定义。

[0217] “神经网络”为含有互相连接的处理单元或“神经元”的组的模型,所述处理单元或“神经元”使用联结法(connectionist approach)计算处理信息。神经网络被用来对输入和输出之间的复杂关系建模并被用来找出数据中的模式。大多数神经网络以非线性、分散式、平行的方式处理数据。在大多数情形中,神经网络是在学习阶段期间改变其结构的自适应系统(adaptive system)。处理元件统一且平行地执行多种功能,而不存在对被分配了多个单元的子任务的清晰描述。

[0218] 通常,神经网络包括简单处理单元的网络,所述网络呈现出通过处理元件和元件参数之间的关系确定的复杂性整体行为。神经网络与被设计为改变网络中的联系强度的算

法一起使用以产生期望的信号流。所述强度在训练或学习期间被改变。

[0219] “随机森林”指分类树预测因子的组合，使得每棵树依赖于独立取样的随机向量的值，并且森林中所有的树具有相同的分布。随机森林是由在决策树的每个分叉处具有随机选择特征的许多未修剪的学习决策树组成的学习整体。随机森林生长出大量的分类树，每个分类树对最受欢迎的类别投票。然后随机森林通过从森林中的所有预测树中选取最受欢迎的投票类别来对变量分类。

[0220] 具有不确定的量p的“先验概率分布”或“先验”是在感兴趣的数据(例如，蛋白序列的训练组)被考虑之前表示关于p的不确定性的概率分布。未知的量可以是参数、系数、变量、潜在变量等(例如，多元回归模型中的系数)。

[0221] 不确定的量p的“后验概率分布”或“后验”是在感兴趣的数据被考虑之后表示关于p的不确定性的概率分布。

[0222] 术语“贝叶斯线性回归”指，其中在贝叶斯推理的背景中进行统计分析的线性回归方法。将关于线性回归模型的先验知识(prior belief)，包括模型参数的先验概率分布函数与根据贝叶斯定理的数据的似然函数结合以获得关于参数的后验概率分布。

[0223] “过度拟合”指当统计模型描述随机误差或噪声时，而不是描述潜在关系时发生的情况。过度拟合通常发生当模型极度复杂时，诸如相对于观察对象的数目具有太多的参数。已被过度拟合的模型通常将具有差的预测性能，因为其会夸大数据中的小幅波动。

[0224] 术语“基础模型”被用来指，在改进模型的过程开始时提供的序列-活性模型。

[0225] 术语“更新的模型”被用来指，从基础模型直接或间接地获得的序列-活性模型，其与基础模型和/或其所来源的另一个模型相比具有提高的预测力。

[0226] 模型的“似然函数”或“似然性”为统计模型的参数的函数。提供某些观察结果的一组参数值的似然性等于提供那些参数值的那些观察结果的概率，即$L(\theta|x)=P(x|\theta)$。

[0227] “蒙特卡罗模拟”为依赖大量的随机抽样获得模仿真实现象的数值结果的模拟。例如，从区间(0,1]抽取大量的伪随机均匀变量，并将小于或等于0.50的值指定为正面而将大于0.05的值指定为反面，是对重复掷硬币行为的蒙特卡罗模拟。

[0228] “Metropolis算法”或“Metropolis-Hastings算法”为用于从概率分布获得随机样品的序列的马尔可夫链蒙特卡罗(MCMC)法，对于概率分布，直接抽样是困难的。该抽样序列可被用来估计所述分布(即，生成直方图)，或用来计算积分(例如，期望值)。Metropolis-Hastings或其他MCMC算法通常被用于从多维分布中抽样，特别是当维数高时。Metropolis-Hastings算法的目的是，渐进地生成根据期望分布P(x)的状态x并利用随机过程实现它。该算法的思想是调节该随机过程以使得其渐进地收敛至独特的分布P(x)。

[0229] “马尔可夫链”是具有马尔可夫特性的随机变量$X_1,X_2,X_3...$的序列。换言之，给定当前状态，则未来和过去状态是独立的。形式上是，

[0230] $Pr(X_{n+1}=x|X_1=x_1;X_2=x_2,...,X_n=x_n)=Pr(X_{n+1}=x|X_n=x_n)$.

[0231] X_i的可能值形成称为链的状态空间的可计算组S。“马尔可夫链”系统是在有限的或可计算数目的可能状态之间经历从一个状态转换成另一个状态的数学系统。它是通常表征为无记忆的随机过程：下一个状态只取决于当前的状态而不取决于其之前的事件的序列。

[0232] “赤池信息量准则”(AIC)是对统计模型的相对拟合优度的测量，并且其经常被用

作在有限组的模型中选择模型的标准。AIC建立在信息熵的概念上，事实上当给定的模型被用来描述真实性时提供对信息丢失的相对测量。其可被说成是描述模型结构中偏差和方差之间的权衡，或不严格地讲，模型的准确性和复杂性之间的权衡。AIC可如以下被计算：$A_{IC}=-2\log_e L+2k$，其中L是函数的最大似然性且k是待建立的模型的自由参数的数目。

[0233] “贝叶斯信息准则”为在有限组的模型中选择模型的准则，且与AIC紧密相关。BIC可按照下式计算：$BIC=-2\log_e L+k\log_e(n)$，其中n为数据观察值的数目。由于观察值的数目增加，BIC往往比AIC对更多的自由参数的额外数目罚分。

[0234] “遗传算法”是模仿进化过程的程序。遗传算法(GA)被用于很多领域来解决未被完全表征或太复杂以致不能被完全表征的问题，但是对于所述问题一些分析评价是可获得的。即，GA被用来解决可通过一些对解的相对值(或至少一个可能的解相对于另一个解的相对值)的定量测量来评价的问题。在本公开内容的上下文中，遗传算法为在计算机中选择或操作字符串的过程，通常其中该字符串对应于一个或更多个生物分子(例如，核酸、蛋白等)。

[0235] 术语“遗传操作”(或GO)指生物的和/或计算的遗传操作，其中任何类型的字符串的任何群中(以及由此而来的由此类字符编码的物理对象的任何物理特性中)的所有改变可被描述为随机和/或预先确定地应用有限组的逻辑代数函数的结果。GO的实例包括但不限于扩增、交换、重组、突变、连接、片段化等。

[0236] “集成模型”为其项包括一组模型的所有项的模型，其中集成模型的项的系数基于该组中的各个模型的相应项的加权系数。系数的加权基于各个模型的预测力和/或适应性。

[0237] II.搜寻改进的蛋白变体的程序的概览

[0238] 在探索蛋白序列的引导型进化方法中，序列-活性模型被用来引导蛋白变体的生成。本公开内容的一方面提供了用于制备序列-活性模型的多种方法，所述多种方法基于蛋白文库并可用来搜寻新的且改进了的蛋白文库。该部分首先提供了用于搜寻新的且改进了的蛋白的方法的概览，然后提供了与选择起始文库、建立序列-活性模型和使用该模型引导新蛋白的探索有关的问题的进一步的细节。

[0239] 该公开内容提供了涉及氨基酸残基序列和蛋白活性的示例性实例，但应当理解，本文描述的方法还可对其他生物序列和活性实施。例如，在多个实施方案中，序列可以是完整的基因组、完整的染色体、染色体片段、对于相互作用的基因的基因序列的集合、基因、核酸序列、蛋白、多糖等。在一个或更多个实施方案中，序列的亚单位可以是染色体、染色体片段、单倍型、基因、核苷酸、密码子、突变体、氨基酸、单体碳水化合物、二聚碳水化合物、三聚碳水化合物或寡聚碳水化合物等。

[0240] 通常，在特定的一轮序列定向进化的开始，获得待测序或测定的蛋白变体的训练组。既定的一轮定向进化产生很多变体蛋白，所述变体蛋白与该轮定向进化开始时使用的亲本肽或多个亲本肽相比发生一个或更多个突变而不同。测定该轮定向进化期间产生的变体蛋白的活性。选择具有期望的活性和/或与亲本肽相比具有提高的活性的那些肽被用于至少另外的一轮定向进化。

[0241] 被测序和测定的蛋白变体还可用来制作序列-活性模型。通常，如果事实上对它们进行测序，则它们将被用于序列活性模型。每个被测序和测定的蛋白变体被称为“观察对象”。通常，生成序列-活性模型使用的观察对象越多，该序列-活性模型的预测力越好。

[0242] 在下一代大规模平行测序技术出现之前，经济地对任何一轮定量进化中产生的多于10至30个变体肽测序是困难的。现在随着下一代测序的应用，在一轮定向进化中产生的许多更多的变体蛋白可被测序。结果，可使用大得多的训练组数据的集制作序列活-性模型。现在，序列活性模型可使用这样的训练组产生，其不仅包括一轮中表现最好的肽，而且包括进一步的多轮定向进化不感兴趣但是其序列活性信息可被应用于制作更稳健的序列活性模型的一些肽。

[0243] 在一些实施方案中，通常期望制作具有良好的预测任意序列的活性的能力的序列-活性模型。预测力可通过预测的精确性以及模型精确地预测活性的一致性来表征。另外，可通过其在宽广范围的序列空间精确地预测活性的能力来表征模型。例如，预测力可根据对于肽的给定的测试和/或验证组，计算的和实际的活性之间的残差(residual)来表征。具有较高的普遍性预测力的模型往往遍布不同的验证数据组产生较小且更一致的残差。与测试数据组过度符合的模型往往对于验证数据产生较大且不太一致的残差，如通过以下实例表明的。本公开内容的一方面提供了有效地找出遍布不同的数据组具有高的预测力的模型的方法。

[0244] 如本文描述的序列-活性模型可被用来协助鉴定经历定向进化的起始变体文库中的一个或更多个亲本“基因”。在一轮进化进行之后，鉴定新的变体文库，提供新的一组观察值，然后新的一组观察值可作为数据反馈以制作新的或精化的序列-活性模型。这种在基于新的观察值制作序列-活性模型和基于序列-活性模型进行定向进化之间交替的程序可形成建模-探索的迭代循环，该程序可被重复直到得到期望的蛋白和文库。

[0245] 由于序列-活性模型和变体文库之间的反馈循环，在探索具有改进的活性的蛋白中，更好的模型和更好的变体文库彼此依赖。因此，建模和/或测序领域的瓶颈和改进可影响到这两个领域。在本发明的一些实施方案中，因更好的建模技术得到的建模效率的改进提供更好的模型以引导序列探寻。在一些实施方案中，下一代测序技术被用来改进体外(in vitro)测序速率，以及被用来提供交叉验证数据以经由计算机模拟(in silico)改进计算模型。

[0246] 在本发明的一些实施方案中，有用的序列-活性模型需要稳健的数学建模技术和大量的“观察值”。这些观察值是训练组中提供的用于模型的数据。特别地，每个观察值是文库中变体的活性值和相关序列。在过去，测序是开发大的训练组的限制步骤，并且因此也是开发更加稳健的序列-活性模型的限制步骤。在目前通常使用的方法中，生成了可能具有数以百计的变体的变体文库。但是，事实上这些变体中只有少部分被测序。在一轮典型的定向进化中，事实上只有具有最高活性的约10至30个变体被测序。理想地，文库中更大比例的变体会被测序，包括具有相对低的活性的一些变体。下一代测序工具已经大大改进了测序速率，使得在训练组中包括低活性的变体和高活性的变体成为可能。在一些实施方案中，包括具有一定范围的活性水平的变体导致产生表现更好的模型和/或在对宽广范围的序列和活性空间的预测能力方面更好的模型。

[0247] 本文提及的非交互作用序列-活性模型包括不同的残基作为自变量来预测任何感兴趣的活性。这些非交互作用序列-活性模型不包括解释两个或更多个残基之间的相互作用的项。如果两个残基之间的相互作用对活性具有协同作用，那么非交互作用模型或线性模型可提供与这两个相互作用残基相关的系数的人为地夸大的值。结果，使用模型工作的

人员可能错误地得出结论，通过简单地按照相对高的系数值建议的进行残基取代，得到的肽的活性将高于预期。这是因为，由于使用非交互作用模型或线性模型，研究者不能理解，与残基取代关联的增加的活性主要是该取代与另一取代相互作用的结果。如果研究者理解了这种相互作用的重要性，那么他或她会同时进行这两个取代并得到非交互作用模型所暗示的活性的增加。

[0248] 如果两个残基相互作用以非线性方式抑制活性，那么非交互作用模型将比如果以单纯地彼此分离的方式考虑这两个残基将会是适当的值低的值归因于与这两个残基有关的系数。换言之，对于相互作用的残基进行一种取代而不进行另一个将产生比非交互作用模型或线性模型所暗示的大的活性的结果。

[0249] 由于当残基间相互作用对活性具有强烈影响时，非交互作用模型可能是不恰当的，所以具有解释残基间的相互作用的交互作用项的交互作用模型对于活性的精确预测经常是必需的。然而，利用交互作用项的模型提出了计算方面的挑战和对经验主义的挑战。最值得注意的是，在开发/利用模型方面存在有许许多多可能的交互作用项要考虑，这需要相当大量的计算。更大的限制是，制作具有相当数量的残基-残基交互作用项的模型所需的观察值的潜在数目。另外，给定特定数目的可得的观察值，对于模型创建技术可能存在过度拟合数据的倾向。为了解决该挑战，谨慎地选择并限制序列-活性模型中提供的交互作用项是很多模型的开发中重要考虑的。

[0250] 图1展示了，显示制作序列-活性模型的程序的一种实施方式的流程图。如所描绘的，程序100在提供变体基因或蛋白的序列和活性数据(“观察值”)的模块103处开始。序列数据可例如从包含原始的或改进的蛋白变体文库的残基序列的训练组获得。通常，这些数据包括完整的或部分的残基序列信息，以及文库中每个蛋白的活性值。在一些情形中，多种类型的活性(例如，速率常数数据和热稳定性数据)被共同提供在训练组中。由期望的结果确定的其他的数据来源也可被考虑。一些合适的数据来源包括但不限于，描述与正在构建的序列活性模型有关的特定肽的信息的参考文献。另外的信息来源包括但不限于同一项目中的较早的或不同的多轮定向进化。事实上，期望的是(利用任何合适的方法，包括但不限于本文提供的那些方法)源自先前的多轮定向进化的信息将在后期产生的文库、变体等的开发方面具备实用性。

[0251] 在很多实施方案中，蛋白变体文库的个体成员体现大量的序列和活性。这有利于生成在序列空间的广泛区域可应用的序列-活性模型。用于生成如此多样化的文库的技术包括但不限于蛋白序列的系统变异和如本文描述的定向进化技术。但是，在一些可选的实施方案中，从特定基因家族(例如，在多个物种或有机体中发现的特定激酶)的基因序列产生模型是期望的。由于在该家族的所有成员中很多残基将是相同的，该模型仅描述变异的那些残基。因此，在一些实施方案中，基于此类相对小的训练组(与所有可能变体的组相比)的统计模型在局部上是有效的。即，这些模型只对给定变体的给定观察值有效。在一些实施方案中，目标不是找到全部符合的函数，因为应认识到，在一些模型中，这超出了所考虑的模型系统的能力和/或需要。

[0252] 活性数据可利用本领域已知的任何合适的方法获得，包括但不限于被得当地设计以计算感兴趣的一种活性/多种活性的量级(magnitude)的测定和/或筛选方法。此类技术是熟知的并且对于本发明不是必需的。设计合适的测定或筛选方法的原理是本领域广泛理

解和知晓的。用于获得蛋白序列的技术也是熟知的且对于本发明不是关键的。如所提到的，可使用下一代测序技术。在一些实施方案中，感兴趣活性可以是蛋白的稳定性(例如，热稳定性)。但是，很多重要的实施方案考虑其他的活性诸如催化活性、对病原体和/或毒素的抗性、治疗活性、毒性等。事实上，不期望本发明被限制于任何特定的测定/筛选方法和/或测序方法，因为本领域已知的任何合适的方法在本发明中具备实用性。

[0253] 在已生成或获得训练组数据之后，所述程序利用它生成将活性预测为序列信息的函数的基础序列活性模型。参见模块105。该模型当被提供以特定蛋白的序列信息时，其是预测该蛋白的相对活性的方程式、算法或其他工具。换言之，蛋白序列信息是输入，而活性预测是输出。

[0254] 在一些实施方案中，每个模型包括多个相乘项的乘积，每个相乘项反映特定序列位置处的特定类型的特定残基对活性的贡献。在其他实施方案中，每个模型包括多个项的和，每个相乘项反映特定序列位置处的特定类型的特定残基对活性的贡献。

[0255] 在一些实施方案中，基础模型不包括任何交互作用项。在此类情形中，基础模型可被描述为“线性或非交互作用模型”。在其他实施方案中，基础模型包括所有可获得的交互作用项，在该情形中，基础模型可被描述为非线性模型或交互作用模型。

[0256] 对于很多实施方案，基础模型可对多个残基对活性的贡献排序。下文讨论了全部落在机器学习的范畴之下的产生此类模型的方法(例如，偏最小二乘回归(PLS)、主成分回归(PCR)和多元线性回归(MLR)、贝叶斯线性回归)，以及自变量的形式(序列信息)、因变量的形式(活性)和模型自身的形式(例如，线性一阶表达式、或多个项的乘积、或项的相乘和相加组合的混合)。

[0257] 基础序列活性模型生成之后，该程序通过调整模型的项的系数值以使模型的预测值和观察值之间的剩余误差最小化来精化模型。参见模块107。该类型的调整也被称作模型拟合。可使用本领域已知的多种模型拟合方法。例如，遗传算法可用来调整系数值。对于相加模型，多种回归技术可用来拟合模型。

[0258] 在本发明的一些实施方案中，该程序还通过选择待被包括在模型内或待被从模型排除的合适的项以使得剩余误差最小化和/或改进模型的预测力来精化模型。参见模块107。由于所考虑的模型具有的项都选自相同的项集合，该精化过程也被称为嵌套模型中的模型选择。本发明的一些实施方案使用遗传算法选择合适的项。另外地或可选地，本发明的一些实施方案对基础模型项迭代地加上或减去来自可用的交互作用项的集合的交互作用项，并评价所得到的新模型相比基础模型的改进以产生最终模型。当基础模型包括所有可用的交互作用项时，所述程序以逐步的方式减去这些项。当基础模型不包括交互作用项时，所述程序以逐步的方式加上这些项。调整模型系数值和选择模型的项以改进模型的预测力两者均被称为模型优化技术。

[0259] 在模型选择过程中，本发明的一些实施方案提供了这样的方法：不只考虑了模型对于给定的一组数据考虑的方差，还考虑了该模型预测新数据的能力。在一些实施方案中，该模型的选择途径对与具有较少系数/参数的模型相比具有较多系数/参数的等同模型罚分，以避免将该模型过度符合到给定的数据组。选择方法的实例包括但不限于赤池信息量准则(AIC)和贝叶斯信息准则(BIC)及其变化形式。

[0260] 在一系列嵌套模型中，如在具有渐进地多于基础模型的交互作用项(和相关系数)

的回归模型中，即使额外的系数是假的，较复杂的模型也提供与较简单的模型同样好或比之更好的拟合，因为较复杂的模型享有额外的自由度。本公开内容的某些实施方案采用在拟合优度的增益超过被假参数的消耗所抵消的程度上对较复杂的模型的模型选择方法罚分。

[0261] 下文展示了按照模块105和107中的运算产生序列-活性模型的示例性算法。此类技术包括但不限于遗传算法和对模型中包括另外的交互作用项有偏倚的逐步技术(step-wise techniques)。但是，不期望本公开内容被限制于这些特定的实例。

[0262] 在一方面，本公开内容提供了进行定向进化的方法，所述方法包括：(a)获得针对多个蛋白变体中的每一个的序列和活性数据；(b)从针对所述多个蛋白变体中的每一个的序列和活性数据生成序列-活性模型，其中所述序列-活性模型包括：(1)多个项的乘积，其中所述项的至少一些的每一个包括代表蛋白或核酸序列中特定位置处的特定氨基酸或核苷酸对活性的贡献的系数，和(2)代表蛋白变体的活性的因变量；以及(c)使用该模型指导一轮定向进化。

[0263] 在一些实施方案中，使用该模型指导一轮定向进化包括通过以下操作来选择用于一轮定向进化的一个或更多个突变：评价序列-活性模型的项的系数，以鉴定对活性有贡献的特定位置处的一个或更多个特定氨基酸或核苷酸。所述方法还包括制备含有或编码以上选择的一个或更多个突变的多个寡核苷酸；以及使用以上制备的寡核苷进行一轮定向进化。

[0264] 在其他实施方案中，所述方法还包括鉴定包含以上选择的一个或更多个突变的新蛋白序列或新核酸序列，以及制备并测定所述新蛋白或由所述新核苷酸序列编码的蛋白。在一些实施方案中，所述方法还包括使用所述新蛋白或由所述新核酸序列编码的蛋白作为用于另外的定向进化的起始点。

[0265] 在替代实施方案中，使用该模型指导一轮定向进化包括基于模型对完整序列(而不是对独特的特定亚单位)对该序列的活性的贡献鉴定完整的蛋白序列。所述方法包括将多个蛋白序列或多个氨基酸序列应用于序列-活性模型以及确定序列-活性模型针对所述多个蛋白序列或核酸序列中的每一个预测的活性值。所述方法还包括通过评价序列-活性模型针对多个序列预测的活性值，从以上应用的多个蛋白序列或多个氨基酸序列中选择新蛋白序列或新核酸序列。所述方法还包括制备并测定具有所述新蛋白序列的蛋白或由所述新核苷酸序列编码的蛋白。

[0266] 在其他的实施方案中，所述方法应用饱和诱变技术。所述方法包括通过评价序列-活性模型的项的系数鉴定对活性有贡献的特定序列位置处的特定类型的特定氨基酸或核苷酸中的一个或更多个来选择蛋白序列或核酸序列中的一个或更多个位置特定。然后所述方法在所鉴定的一个或更多个位置处进行饱和诱变。

[0267] 在另一方面，本公开内容提供了用于制作能有助于鉴定影响期望的活性的生物分子的序列活性模型的方法。在一些实施方案中，所述方法包括：(a)接收针对多个生物分子的序列数据和活性数据；(b)根据序列数据和活性数据制作基础模型，其中所述基础模型将活性预测为序列的亚单位的存在或不存在的函数；(c)通过对所述基础模型加上/减去至少一个新的交互作用项来制作至少一个新模型，其中所述新的交互作用项代表两个或更多个相互作用的亚单位之间的相互作用；(d)确定所述至少一个新模型将活性预测为亚单位的

存在或不存在的函数的能力;和(e)基于在(d)中确定的所述至少一个新模型的预测活性的能力以及对添加所述新的交互作用项的偏倚来决定是否对所述基础模型加上/减去所述新交互作用项。然后所得到的模型可被用于多种应用,诸如用于蛋白文库的定向进化以鉴定具有期望的生物活性和特性的蛋白。

[0268] 在一些实施方中,其中所述方法确定了新的交互作用项应被添加至基础模型以产生更新的模型,所述方法还包括:(f)使用更新的模型取代基础模型重复(c),并加上或减去与(c)中加上/减去的交互作用项不同的交互作用项;以及(g)使用更新的模型取代基础模型重复(d)和(e)。在一些实施方案中,所述方法还包括:(h)利用进一步更新的模型重复(f)和(g)。

[0269] 在选择用于训练组的观察对象并选择用于产生序列-活性模型的数学技术之后,基础模型被创建。基础模型的制备通常不考虑其预测能力。如本文所描述的,其按照用于从可用的多个观察对象(即,观察对象组)产生基础模型所定义的程序简单地产生。如以上指出的,序列模型可描述多种序列,虽然在一些实施方案中,该模型描述了蛋白。在后一种情形中,基础模型仅仅是具有针对用于生成训练组的肽的集合中存在的每个突变的一个项的非交互作用/线性模型。在这些实施方案中,基础模型不包括任何代表肽中的残基之间的相互作用的项。在一些实施方案中,基础模型不包括单独的针对观察组中存在的每个和各个突变的项。

[0270] 在替代性方法中,基础模型不仅包括孤立地描述每个突变的项,还另外地包括针对所有潜在的相互作用残基的项。在极端的情形中,所指出的突变之间的每个可想到的相互作用被用于基础模型。这包括针对突变之间的每种和各种成对相互作用的项,以及针对每个和各个可能的三残基相互作用的项,以及针对各个可能的四残基相互作用的项等。一些实施方案只包括成对相互作用,或包括成对相互作用和三向相互作用。三向相互作用为三个不同的亚单位之间的影响活性的相互作用。

[0271] 在本公开内容的使用简单非交互作用模型或线性模型作为基础模型的一个或更多个实施方案中,随后的改进该模型的努力包括添加代表不同的相互作用的新的项。在替代性实施方案中,其中基础模型包括所有的非交互作用项和交互作用项,随后改进该模型的努力包括选择性地去除一些交互作用项。

[0272] 在本发明的一个或更多个实施方案中,改进基础模型的程序在确定所得到的模型是否充分地改进了该模型的质量时包括对基础模型迭代地加或减交互作用项。在每个迭代中,当前模型的预测力被确定并与另一个模型例如基础模型或更新的模型相比较。

[0273] 在预测力的测量已经考虑了模型推广到其他数据组的能力的实施方案中,单独的该测量即可确定候选模型是否应被选择。例如,诸如AIC或BIC的测量考虑模型的似然性(或剩余误差)和参数数目两者。模型的“似然函数”或“似然性”是统计模型的参数的函数。提供一些观察结果的一组参数值的似然性等于提供那些参数值的那些观察结果的概率,即$L(\theta|x)=P(x|\theta)$。模型似然性的示例性计算被描述下文以下的一个部分。如果具有较多参数的模型捕获与具有较少参数的模型相同量的数据差异性,则诸如AIC和BIC的测量偏倚具有较多参数的模型。如果预测力的测量只考虑剩余误差,那么剩余误差改进的量级必须被考虑以确定是否将与当前迭代相关的改变整合到目前最佳的更新的模型中。这可通过将改进的量级与阈值相比来完成。如果量级小于阈值,则在当前迭代中正在考虑的改变不被接受。可

选地，如果改进的量级超过阈值，那么正在考虑的改变被整合到更新的模型中并且所述更新的模型充当新的最佳模型继续用于剩余的迭代。

[0274] 在某些实施方案中，每个迭代考虑对正在考虑的当前最佳模型加或减单一迭代项。在当基础模型只包含非交互作用项时的情形中，所有可用的交互作用项的集可被考虑。这些交互作用项中的每一个被连续地考虑，直到所述程序完成并最终获得最佳模型。

[0275] 在一些情形中，在确定所述程序已经有效地收敛并且不可能进一步改进之后，所述模型生成程序终止于所述集中的所有可用的交互作用项被考虑之前。

[0276] 图2示出了在一个程序(参见，200)中，为了探索蛋白序列和活性空间的目的，模型如何能够被用于反复地引导新的蛋白变体文库的创建。在一个实例中，在最终的模型产生之后，该最终模型被用于选择被预测为影响活性的多个残基位置(例如，位置35)的突变或特定的残基类型(例如，在位置35处的谷氨酸)。在一些实施方案中，所述确定基于项的系数。参见模块207。除了鉴定此类位置之外，该模型可被用来对这些残基位置或残基类型基于其对活性的贡献例如其系数值“排序”。这样的函数是期望的。例如，该模型可预测位置35处的谷氨酸对活性具有最显著的、积极的影响；位置208处的苯丙氨酸对活性具有第二最显著的、积极的影响等。在下文描述的一种特定方法中，PLS或PCR回归系数被采用对特定残基的重要性排序。在另一种特定的方法中，PLS荷载矩阵被用来对特定残基位置的重要性排序。

[0277] 残基使用很多不同的选择方案中的任一种方案来选择，下文描述了这些方案中的一些方案。在一个例证性实例中，被预测为对活性具有最有利的影响的特定残基被保留(即，未变异的)。然而，一定数目的被预测为具有较小影响的其他残基被选择用于变异。在另一个例证性实例中，发现对活性具有最大影响的残基位置被选择用于变异，但是只在其被发现在训练组的高表现成员中不同时。例如，在该方法中，如果所述模型预测残基位置197对活性具有最大的影响，但是所有的或大多数的具有高活性的蛋白在该位置具有亮氨酸，则位置197不被选择用于变异。换言之，下一代文库中所有的或大多数蛋白会在位置197处具有亮氨酸。但是，如果一些“好的”蛋白在该位置处具有缬氨酸但是其他的蛋白具有亮氨酸，那么该程序会选择改变该位置处的氨基酸。在一些情形中，会发现两个或更多个相互作用残基的组合对活性具有最大影响。因此，在一些策略中，这些残基是被共同改变的。共变的实例是对这些残基同时应用饱和诱变。

[0278] 在所述程序已鉴定突变之后，所述程序制备包含或编码所鉴定的突变的寡核苷酸，如在模块209处所示(图2)。多种方法可被用于制备所述寡核苷酸。在一些实施方案中，所述寡核苷酸通过序列合成仪来制备。

[0279] 在鉴定了用于变异的残基后，所述方法接下来产生具有指定的残基变异的新的变体文库。参见模块211(图2)。多种方法学可用于该目的。在一个实例中，进行基于体外或体内重组的多样性生成机制产生新的变体文库。这些程序可使用含有用于编码亲代变体文库的蛋白的序列或亚序列的寡核苷酸。这些寡核苷酸中的一些是密切相关的，只是在选择被选定的用于在209处的改变的替代氨基酸的密码子的方面不同。基于重组的多样性生成机制可进行一个或多个循环。如果使用多个循环，则每个循环都包括筛选步骤以鉴定哪些变体具有待用于随后的重组循环的可接受的表现。这是一种形式的定向进化。但是，不期望本发明被局限于基于重组的多样性生成方法的任何特定方法，因为任何合适的方法/技术在

本发明中具备实用性。在一些实施方案中，例如，本发明在由序列-活性模型的非交互作用项或交互作用项的系数值指示的相互作用的各个序列位置或序列位置的组合处进行饱和诱变如。

[0280] 在另外的例证性实例中，选择“参考”蛋白序列并且在图2的207处被选择的残基被“切换”以鉴定变体文库的各个成员。如此鉴定的新蛋白通过适当的技术合成以产生新的文库。在一个实例中，参考序列可以是训练组的表现最佳的成员或是通过PLS或PCR模型预测的“最佳”序列。

[0281] 在另一个方法中，序列活性模型在用于探索序列空间的遗传算法中被用作“适应性函数”。在一轮或更多轮遗传算法(每轮使用适应性函数选择用于遗传操作的一个或更多个可能的序列)后，如该流程图中描述的鉴定待使用的下一代文库。从非常现实的意义上说，该策略可被视为计算机模拟的定向进化。在理想的情况中，如果技术人员手上具有准确的、精确的整体或局部适应性函数，其可在计算机模拟中进行所有的进化并合成一个最佳变体用于最终的商业或研究应用。虽然在大多数情况中这不可能实现，但是对该程序的该视点为使用机器学习技术用于定向进化的目标和途径提供了清晰性。

[0282] 在另一个例证性实例中，在一个亲本序列中选择一轮定向进化中用于变异的残基。亲本可利用从在先的一轮定向进化得到的模型或通过利用鉴定具有最佳测定表现的文库成员的数据来鉴定。用于下一轮定向进化的寡核苷酸可被限定为包括所选择的亲本的骨架的部分，具有通过用于当前的一轮进化的序列活性模型根据算法预测的一个或更多个突变。这些寡核苷酸可利用任何合适的方法产生，包括但不限于合成方法。

[0283] 新的文库产生之后，如在模块213中所示(图2)，针对活性对其进行筛选。理想地，新的文库提供活性优于之前的文库中观察到的活性的一个或更多个成员。但是，即使没有该优势，新的文库可提供有益的信息。其成员可被用于生成解释新变体的影响的改进的模型，并且从而跨越序列空间的更宽的区域更精确地预测活性。另外，所述文库可代表序列空间从局部最优通向整体最优(例如，在活性方面)的渠道。

[0284] 取决于程序200(图2)的目标，在一些实施方案中，产生一系列新的蛋白变体文库是期望的，每一个蛋白变体文库提供训练组的新成员。然后，更新的训练组被用于产生改进的模型。为了获得该改进的模型，程序200被示为具有如模块215中所示的决策操作，该决策操作决定是否还应产生另一个蛋白变体文库。多种准则可被用于做该决定。决策准则的实例包括但不限于，至今为止产生的蛋白变体文库的数目、来自当前文库的最佳蛋白的活性、期望的活性的量级和在目前的新文库中观察到的提高的度。

[0285] 假定所述序程被用来继续处理新文库，所述程序则返回到操作模块100(图2)，在这里由针对当前的蛋白变体文库而获得的序列和活性数据产生新的序列活性模型。换言之，针对当前的蛋白变体文库的序列和活性数据用作新的模型的训练组的一部分(或其可作为完整训练组)。其后，按照上文描述的执行模块207、209、211、213和215中所示的操作(图2)，但是是对于新的模型。

[0286] 当确定该方法已达到终点时，图2中示出的循环结束且不产生新的文库。在那时，所述程序完全终止，或在一些实施方案中，选择来自一个或更多个文库的一个或更多个序列用于开发和/或生产。参见模块217。

[0287] III.产生观察对象

[0288] 蛋白变体文库是多种蛋白的组,文库中的这些蛋白具有一个或更多个因成员而异的残基。这些文库可利用本文描述的方法和/或本领域已知的任何合适的方式产生。这些文库在提供用来产生根据本发明的多个实施方案的序列活性模型的训练组的数据方面具备实用性。蛋白变体文库中包含的蛋白的数目往往取决于与其生成相关的应用程式(application)和花费。不期望本发明受限于本发明的方法中使用的蛋白文库中的任何特定数目的蛋白。也不期望本发明受限于任何特定的一个或多个蛋白变体文库。

[0289] 在一个实例中,蛋白变体文库由一种或更多种天然存在的蛋白产生,所述蛋白可由一个基因家族编码。可使用其他起点,包括但不限于已知蛋白或新型合成蛋白的重组体。文库可通过多种技术从这些种子(seed)蛋白或起始蛋白产生。在一种情形中,文库通过如Stemmer(1994)Proceedings of the National Academy of Sciences,USA,10747-10751和WO 95/22625(其均被通过引用并入本文)中描述的DNA片段化介导的重组、如Ness等人(2002)Nature Biotechnology 20:1251-1255和WO 00/42561(其均被通过引用并入本文)中描述的合成的寡核苷酸介导的重组、或编码一个或更多个亲本蛋白的一部分或全长的核酸来产生。也可利用这些方法的组合(例如,DNA片段和合成的寡核苷酸的重组)以及例如均被通过引用并入本文的WO97/20078和WO98/27230中描述的其他基于重组的方法。用于产生蛋白变体文库的任何合适的方法在本发明中具备实用性。实际上,不期望本发明受限于用于产生变体文库的任何特定方法。

[0290] 在一些实施方案中,单个的“起始”序列(其可以是“祖”序列)可被采用用于定义建模过程中使用的一组突变体的目的。在一些实施方案中,至少一个起始序列是野生型序列。

[0291] 在某些实施方案中,突变(a)在文献中被鉴定为影响底物特异性、选择性、稳定性或其它有利特性,和/或(b)通过计算被预测为改进蛋白的折叠形式(例如,压紧蛋白的内部残基)、配体结合、亚基相互作用、多个不同的同源体之间的家族重排等。可选地,突变可自然地地引入起始序列并表达待筛选有利特性的产物。定点诱变是对于引入突变有用的技术的一个实例,但是任何合适的方法具备实用性。因此,可选地或另外,突变可通过基因合成、饱和随机诱变、残基的半合成组合文库、定向进化、重现性序列重组(recursive sequence recombination)(“RSR”)(参见例如美国专利申请号2006/0223143,其被通过引用全部并入本文)、基因重排、易错PCR和/或任何其他合适的方法来提供。合适的饱和诱变程序的一个实例被描述于美国公布的专利申请号20100093560,其被通过引用全部并入本文。

[0292] 起始序列不必与野生型蛋白的氨基酸序列相同。但是,在一些实施方案中,起始序列是野生型蛋白的序列。在一些实施方案中,起始序列包含野生型蛋白中不存在的突变。在一些实施方案中,起始序列是源自具有共同特性的一组蛋白,例如一个家族的蛋白的共有序列。

[0293] 可用作亲本序列来源的多个家族或多种类别的酶的非限制性的代表性清单包括但不限于以下:氧化还原酶(E.C.1);转移酶(E.C.2);水解酶(E.C.3)、裂解酶(E.C.4);异构酶(E.C.5)和连接酶(E.C.6)。氧化还原酶的更具体但非限制性的亚组包括,脱氢酶(例如,醇脱氢酶(羰基还原酶)、木酮糖还原酶、醛还原酶、法尼醇脱氢酶、乳酸脱氢酶、阿拉伯糖脱氢酶、葡萄糖脱氢酶、果糖脱氢酶、木糖还原酶和琥珀酸盐/酯脱氢酶)、氧化酶(例如,葡萄糖氧化酶、己糖氧化酶、半乳糖氧化酶和漆酶)、单胺氧化酶、脂肪氧合酶、过氧物酶、醛脱氢酶、还原酶、长链酰基-[酰基-载体-蛋白]还原酶、酰基-CoA脱氢酶、烯-还原酶、合酶(例如,

谷氨酸合酶)、硝酸盐还原酶、单加氧酶和双加氧酶和过氧化氢酶。转移酶的更具体但非限制性的亚组包括甲基转移酶、脒基转移酶和羧基转移酶、转酮醇酶、转醛缩酶、酰基转移酶、糖基转移酶、转氨酶、转谷氨酰胺酶和聚合酶。水解酶的更具体但非限制性的亚组包括酯水解酶、肽酶、糖基化酶、淀粉酶、纤维素酶、半纤维素酶、木聚糖酶、几丁质酶、葡萄糖苷酶、葡聚糖酶、葡糖淀粉酶、酰化酶、半乳糖苷酶、支链淀粉酶、植酸酶、乳糖酶、阿拉伯糖苷酶、核苷酶、腈水解酶、磷酸酶、脂酶、磷脂酶、蛋白酶、ATP酶和脱卤酶。裂解酶的更具体但非限制性的亚组包括脱羧酶、醛缩酶、水合酶、脱水酶(例如,碳酸酐酶)、合酶(例如,异戊二烯、蒎烯和法呢烯合成酶)、果胶酶(例如,果胶裂解酶)和卤代醇脱氢酶。异构酶的更具体但非限制性的亚组包括消旋酶、差向异构酶、异构酶(例如,木糖异构酶、阿拉伯糖异构酶、核糖异构酶、葡萄糖异构酶、半乳糖异构酶和甘露糖异构酶)、互变异构酶和变位酶(例如酰基转移变位酶、磷酸变位酶和氨基变位酶)。连接酶的更具体但非限制性的亚组包括酯合成酶。可被用作亲本序列的来源的其他家族或类别的酶包括,转氨酶、蛋白酶、激酶和合成酶。该清单虽然例示了本公开内容的可能的酶的某些特定方面,但是仍未考虑详尽并且没有描述对本公开内容的限制或限制本公开内容的范围。

[0294] 在一些情形中,在本文描述的在所述方法中有用的候选酶能催化对映选择性反应,诸如,例如对映选择性还原反应。此类酶可被用来制备在例如药物化合物的合成中有用的中间体。

[0295] 在一些实施方案中,候选酶选自内切木聚糖酶(EC 3.2.1.8)、β-木糖苷酶(EC 3.2.1.37)、α-L-阿拉伯呋喃糖苷酶(EC 3.2.1.55)、α-葡糖醛酸酶(EC3.2.1.139)、乙酰基木聚糖酯酶(EC 3.1.1.72)、阿魏酰酯酶(EC 3.1.1.73)、香豆酰酯酶(EC 3.1.1.73)、α-半乳糖苷酶(EC 3.2.1.22)、β-半乳糖苷酶(EC3.2.1.23)、β-甘露聚糖酶(EC 3.2.1.78)、β-甘露糖苷酶(EC 3.2.1.25)、内切-聚半乳糖醛酸酶(EC 3.2.1.15)、果胶甲基酯酶(EC 3.1.1.11)、内切-半乳聚糖酶(EC 3.2.1.89)、果胶乙酰基酯酶(EC 3.1.1.6)、内切-果胶裂解酶(EC4.2.2.10)、果胶酸裂解酶(EC 4.2.2.2)、α鼠李糖苷酶(EC 3.2.1.40)、外切-聚-α-半乳糖醛酸苷酶(EC 3.2.1.82)、1,4-α-半乳糖醛酸苷酶(EC 3.2.1.67)、外切聚半乳糖醛酸裂解酶(EC 4.2.2.9)、鼠李糖半乳糖醛酸聚糖内切裂解酶(EC 4.2.2.B3)、鼠李糖半乳糖醛酸聚糖乙酰酯酶(EC 3.2.1.B11)、鼠李糖半乳糖醛酸聚糖半乳糖醛酸水解酶(rhamnogalacturonan galacturonohydrolase)(EC 3.2.1.B11)、内切-阿拉伯聚糖酶(EC 3.2.1.99)、漆酶(EC 1.10.3.2)、锰依赖型过氧化物酶(EC 1.10.3.2)、淀粉酶(EC 3.2.1.1)、葡萄糖淀粉酶(EC 3.2.1.3)、蛋白酶、脂酶和木质素过氧化物酶(EC1.11.1.14)。1种、2种、3种、4种、5种、或多于5种的酶的任何组合在本发明的组合物中具备实用性。

[0296] 在本发明的一个或更多个实施方案中,一种起始序列以多种方式被修饰以产生文库。在一些实施方案中,文库通过系统地改变起始序列的各个残基来产生。在一个例证性实例中,实验设计(DOE)方法学被用以鉴定被系统地改变的序列。在另一个实例中,“湿实验室”(wet lab)程序诸如寡核苷酸介导的重组被用来引入一定程度的系统变异。不期望本发明被限制于用于产生系统地变异的序列的任何特定方法,因为任何合适的方法具备实用性。

[0297] 如本文使用的,术语“系统地变异的序列”指,其中每个残基见于多种背景的一组序列。原则上,系统变异的水平可通过这些序列彼此正交的程度(即,与平均值相比最大程

度地不同)来定量。在一些实施方案中,该程序不依赖于具有最大程度地正交的序列。但是,模型的质量的改进将与所测试的序列空间的正交性成正比关系。在简单的例证性实例中,通过鉴定两个残基位置,每个残基位置可具有两个不同的氨基酸中的一个,则肽序列被系统地变异。最大程度地不同的文库包括所有四种可能的序列。此类最大程度的系统变异随可变异的位置的数目以指数形式增加;例如当在N个残基位置中的每一个处有2个选择时,最大程度的系统变异以2^N增加。但是,本领域技术人员将容易地认识到,并不需要最大程度的系统变异。系统变异提供了提供良好的序列空间抽样的相对小的待测试序列组的机制。

[0298] 具有系统地变异的序列的蛋白变体可利用本领域普通技术人员所熟知的技术以很多种方式获得。如所指出的,合适的方法包括但不限于基于一个或更多个"亲本"多核苷酸序列生成变体的基于重组的方法。多核苷酸序列可利用多种技术重组,包括例如DNA酶消化待重组的多核苷酸,接下来是核酸的连接和/或PCR重装配。这些方法包括但不限于在例如以下文献中描述的那些方法:Stemmer(1994)Proceedings of the National Academy of Sciences USA,91:10747-10751;美国专利号5,605,793,"Methods for In Vitro Recombination";美国专利号5,811,238,"Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination";美国专利号5,830,721;"DNA Mutagenesis by Random Fragmentation and Reassembly";美国专利号5,834,252,"End Complementary Polymerase Reaction";美国专利号5,837,458,"Methods and Compositions for Cellular and Metabolic Engineering";WO98/42832,"Recombination of Polynucleotide Sequences Using Random or Defined Primers";WO 98/27230,"Methods and Compositions for Polypeptide Engineering";WO 99/29902,"Method for Creating Polynucleotide and Polypeptide Sequences"等,其均被通过引用并入本文。

[0299] 合成重组方法也特别适于产生具有系统变异的蛋白变体文库。在合成重组方法中,共同编码待重组的多个基因的多个寡核苷酸被合成。在一些实施方案中,寡核苷酸共同编码源自同源亲本基因的序列。例如,使用利用序列比对程序,诸如BLAST(参见例如,Atschul,等人,Journal of Molecular Biology,215:403-410(1990)比对感兴趣的同源基因。同源体之间的相应氨基酸变异的核苷酸被指出。这些变异任选地被进一步局限于基于亲本序列的共变异分析、亲本序列的功能信息、亲本序列之间的保守性或非保守性改变的选择或其他合适的准则的全部可能变异的子集。任选地将变异进一步增加至编码通过亲本序列的共变异分析、亲本序列的功能信息、亲本序列之间的保守性或非保守性改变的选择或变异位置的表观公差(apparent tolerance)鉴定的位置处的另外的氨基酸多样性。结果是简并基因序列编码源自亲本基因序列的共有氨基酸序列,简并核苷酸位于编码氨基酸变异的位置。设计这样的寡核苷酸:含有聚集简并基因中存在的多样性所需的核苷酸。关于此类方法的细节可在下列文献中找到:例如,Ness等人(2002),Nature Biotechnology,20:1251-1255;WO 00/42561,"Oligonucleotide Mediated Nucleic Acid Recombination";WO 00/42560,"Methods for Making Character Strings,Polynucleotides and Polypeptides having Desired Characteristics,";WO 01/75767,"In Silico Cross-Over Site Selection"和WO01/64864,"Single-Stranded Nucleic Acid Template-Mediated Recombination and Nucleic Acid Fragment Isolation",其各自被通过引用

并入本文。所鉴定的多核苷酸变异序列可在体外或体内转录和翻译，创建蛋白变异序列的组或文库。

[0300] 系统地变异的序列的组也可利用限定数据组中的序列的实验设计(DOE)方法学通过推理设计。对DOE方法的描述可在Diamond,W.J.(2001)Practical Experiment Designs: for Engineers and Scientists,John Wiley&Sons中和William J Drummond(1981)的“Practical Experimental Design for Engineers and Scientists”Van Nostrand Reinhold Co New York、“Statistics for experimenters”George E.P.Box,William G Hunter和J.Stuart Hunter(1978)、John Wiley和Sons,New York中找到或例如在万维网itl.nist.gov/div898/handbook/上找到。存在可用来进行相关数学运算的若干计算包，包括Statistics Toolbox(MATLAB®)、JMP®、STATISTICA®和STAT-EASE® DESIGN EXPERT®。结果是序列的系统地变异的和正交分布的数据组，所述数据组适于建立本发明的序列-活性模型。基于DOE的数据组也可利用如本领域已知的Plackett-Burman或Fractional Factorial Designs容易地产生。Diamond,W.J.(2001)。

[0301] 在工程化和化学科学中，与完全析因设计相比，分式析因设计被用来定义较少的实验。在这些方法中，因子是在两种或更多种水平之间变化(即，“切换”)。优化技术被用来确保所选择的实验在解释因子空间变化方面是信息量最大的。相同的设计方法(例如，分式析因设计、D-型最适设计)可被应用于蛋白工程以构建较少的序列，在这些序列中给定数目的位置在两个或更多个残基之间切换。在一些实施方案中，该序列组提供了对正在研究的蛋白序列中存在的系统性变异的最佳描述。

[0302] 应用于蛋白工程化的DOE方法的例证性实例包括以下操作：

[0303] 1)基于本文描述的原则鉴定待切换的位置(例如，在亲本序列中存在的位置、保守度等)；

[0304] 2)利用通常可获得的一种统计软件包通过限定因子的数目(即，可变异的位置)、度(level)的数目(即，在每个位置的选择)和运行的实验数目来创建DOE实验以提供输出矩阵。输出矩阵的信息内容(通常由代表每个位置处的残基选择的1和0组成)直接地取决于待运行的实验的数目(通常地，越多越好)。

[0305] 3)使用编码返回至每个位置处的特定残基选择的1和0的输出矩阵创建蛋白比对结果。

[0306] 4)合成编码蛋白比对结果中呈现的蛋白的基因。

[0307] 5)在相关测定中测试所合成的基因编码的蛋白。

[0308] 6)基于所测试的基因/蛋白建立模型。

[0309] 7)按照本文描述的步骤来鉴定重要位置并建立具有提高的适应性的一个或更多个后续文库。

[0310] 在例证性实例中，研究了这样的蛋白，其中20个位置处的功能最佳的氨基酸残基待被确定(例如，其中在每个位置处有2个可能的氨基酸可选)。在该实例中，分辨度IV析因设计将是合适的。分辨度IV设计被定义为，能阐明所有单变量的影响而没有双因素效应与之重叠(with no two-factor effects overlapping them)的设计。然后，所述设计将指定覆盖总多样性为2^{20}(～1百万)个可能的序列中的40个特定氨基酸序列的组。然后利用任何标准基因合成方案制备这些序列并确定这些克隆的功能和适合性。

[0311] 以上方法的替代方法是采用一些或所有可用的序列(例如,GENBANK®数据库和其他公共资源)来提供蛋白变体文库。该方法提供了对感兴趣的序列空间区域的指示。

[0312] IV.测序方法

[0313] 在过去,测序是开发大的训练组和由其得来的越来越稳健的序列活性模型的限制步骤。对变体测序所需的高昂的花费和漫长的时间将观察值的数目限制到几十个变体。下一代测序工具具有大大地降低的花费以及增加的测序速度和容量,使得训练组中包括低活性变体和高活性变体两者成为可能。

[0314] 下一代测序工具可花费不多地在一次运行中对大量的碱基对(例如,至少约1,000,000,000个碱基对)测序。这种能力可用于当在单次运行中对通常长度仅几千碱基对的变异蛋白测序时。下一代测序工具往往被优化以便在单次运行中测序单个的大基因组(例如,人基因组),而不是很多较小的序列。为了实现下一代测序工具平行地对很多观察值测序的潜能,在单次运行中被测序的每个观察值的起始点应被独特地鉴定。在一些实施方案中,条形码标识的序列被用在被供给到下一代测序仪进行单次运行的每个和各个片段上。在一个实例中,条形码独特地鉴定特定平板(例如,96孔板)上的特定孔。在这些实施方案中的一些中,每个板的每个孔含有一个独特的变体。通过对每个变体或更特别地对每个变体的每个片段标以条形码,多个不同变体的基因序列可在单次运行中被测序并被鉴定。在所述过程中,具有相同条形码的所有片段读段通过确定变体的序列长度的算法被一起鉴定并处理。

[0315] 在一些实施方案中,从给定的孔中具有变体的细胞提取DNA,然后将其片段化。然后对这些片段标以条形码以鉴定至少与该变体有关的孔,并且有时候是与该变体有关的孔和板。然后对所得到的片段进行尺寸选择以产生具有适当长度的序列用于下一代测序仪。在一个例证性实例中,读段长度为约200碱基对。在一些实施方案中,直到来自板的多个孔的DNA片段被第一次收集之后才应用板条形码。然后对所收集的DNA标以条形码以辨识所述板。在一些实施方案中,每个片段将具有相同的板条形码,不论所述每个片段源自哪个孔。但是,在一些替代实施方案中,这些片段具有不同的条形码。另外,可应用孔条形码和板条形码以辨识从给定的孔中提取的DNA。

[0316] 在一个或更多个实施方案中,序列数据可利用被认为是第一代测序方法的大量(bulk)测序方法来获得,包括例如桑格测序或Maxam-Gilbert测序。涉及使用带标记的双脱氧链终止剂的桑格测序是本领域熟知的;参见例如,Sanger等人,Proceedings of the National Academy of Sciences of the United States of America 74,5463-5467(1997)。涉及在核酸样品的部分上进行多个部分化学降解反应,接者检测并分析这些片段来推断序列的Maxam-Gilbert测序也是本领域熟知的;参见例如Maxam等人,Proceedings of the National Academy of Sciences of the United States of America 74,560-564(1977)。另一种大量测量方法是通过杂交测序,其中样品的序列基于其在例如微阵列或基因芯片上与多个序列的杂交特性来推导;参见,例如,Drmanac等人,Nature Biotechnology 16,54-58(1998)。

[0317] 在一个或更多个实施方案中,序列数据利用下一代测序方法来获得。下一代测序也被称为“高通量测序”。这些技术使测序过程平行化,一次产生数以千计或数百万计的序列。合适的下一代测序方法的实例包括但不限于,单分子实时测序(例如,Pacific

Biosciences,Menlo Park,California)、离子半导体测序(例如,Ion Torrent,South San Francisco,California)、焦磷酸测序(例如,454,Branford,Connecticut)、连接测序(例如,SOLid sequencing of Life Technologies,Carlsbad,California)、合成测序和可逆性末端终止测序(例如,Illumina,San Diego,California)、核酸成像技术诸如透射电子显微术,等等。

[0318] 一般地,下一代测序方法通常利用体外克隆步骤以扩增各个DNA分子。乳液PCR(emPCR)分离伴随在油相内的水滴中的引物所包被的珠中的各个DNA分子。PCR产生与珠上的引物结合的DNA分子的拷贝,接着被固定用于稍后的测序。emPCR被用于Marguilis等人(由454Life Sciences,Branford,CT商业化)、Shendure和Porreca等人(也被称为“聚合酶克隆测序”)以及SOLiD测序(Applied Biosystems Inc.,Foster City,CA)的方法中。参见M.Margulies等人,例如(2005)“Genome sequencing in microfabricated high-density picolitre reactors”Nature 437:376-380;J.Shendure等人(2005)“Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome”Science 309(5741):1728-1732。体外克隆扩增还可通过“桥式PCR”进行,在所述“桥式PCR”中,片段在引物连接至固体表面后被扩增。Braslavsky等人开发了省去该扩增步骤、直接将DNA分子固定至表面的单分子方法(由Helicos Biosciences Corp.,Cambridge,MA商业化)。I.Braslavsky等人(2003)“Sequence information can be obtained from single DNA molecules”Proceedings of the National Academy of Sciences of the United States of America 100:3960-3964。

[0319] 与表面自然结合的DNA分子可被平行地测序。在“合成测序”中,基于模板链的序列使用DNA聚合酶形成互补链,如染料-终止电泳测序,“可逆”终止剂法(由Illumina,Inc.,San Diego,CA和Helicos Biosciences Corp.,Cambridge,MA商业化)利用可逆形式的染料-终止剂,通过重复移除封端基团来以允许另一个核苷酸的聚合而一次添加一个核苷酸,并实时检测每个位置处的荧光。“焦磷酸测序”也利用DNA聚合,一次添加一个核苷酸并通过由所连接的焦磷酸的释放发射的光来检测和定量被添加到给定位置的核苷酸的数目(由454Life Sciences,Branford,CT商业化)。参见M.Ronaghi,等人(1996).“Real-time DNA sequencing using detection of pyrophosphate release”Analytical Biochemistry 242:84-89。

[0320] 下文更详细地描述了下一代测序方法的具体实例。本发明的一个或更多个实施方案可利用以下测序方法的一种或更多种而不背离本发明的原理。

[0321] 单分子实时测序(也被称为SMRT)是Pacific Biosciences开发的通过并行化单分子DNA合成测序技术。单分子实时测序利用零模式波导(ZMW)。一个DNA聚合酶被附着在ZMW的底部,一分子的DNA作为模板。ZMW是创建足够小以观察被DNA聚合酶并入的DNA的仅一个核苷酸(也被称为碱基)的照射的观察体积的结构。四种DNA碱基的每个被连接至四种不同的荧光染料中的一种。当核苷酸被DNA聚合酶并入时,荧光标签被裂解并从ZMW的观察区域扩散出来,而其荧光不再能够观察得到。检测器检测核苷酸并入的荧光信号,并根据相应的染料荧光进行序时碱基响应(base call)。

[0322] 另一种可应用的单分子测序技术是Helicos True Single Molecule Sequencing(tSMS)技术(例如,如Harris T.D.等人,Science 320:106-109[2008]中描述的)。在tSMS技

术中,DNA样品断裂成约100至200个核苷酸的链,并将多聚A序列添加至每条DNA链的3'末端。通过添加荧光标记的腺苷核苷酸来标记每条链。然后使DNA链与流动池杂交,所述流动池包含数以百万计的被固定至所述流通池表面的寡聚T捕获位点。在某些实施方案中,模板可以是约1亿模板/cm^2的密度。然后将流动池装载至设备例如HeliScope™测序仪,并且激光照亮流通池的表面,揭示每条模板的位置。CCD摄像机可绘制模板在流动池表面上的位置。然后模板荧光标签被裂解掉并被冲走。测序反应通过引入DNA聚合酶和荧光标记的核苷酸开始。寡聚T核酸作为引物。聚合酶以模板指导的方式将带标记的核苷酸整合至引物上。聚合酶和未并入的核苷酸被去除。具有荧光标记的核苷酸的指导性并入的模板通过对流通池表面成像来识别。成像后,裂解步骤去除荧光标签,并且用其他的荧光标记的核苷酸重复该程序直到达到期望的读段长度。随着每个核苷酸的添加步骤收集序列信息。在测序文库的制备中,借助于单分子测序技术的整个基因组测序不包括或通常排除基于PCR的扩增,并且所述方法允许直接测量样品,而不是测量该样品的拷贝。

[0323] 离子半导体测序是基于检测DNA的聚合期间释放的氢离子的DNA测序方法。这是"合成测序"房法,在该方法期间,互补链基于模板链的序列形成。将包含待测序的模板DNA链的微孔用单一种类的脱氧核糖核苷酸三磷酸(dNTP)充满。如果所引入的dNTP与引导性模板核苷酸互补,则其被并入正在增长的互补链。这导致引发ISFET离子传感器的氢离子的释放,所述氢离子的释放指示反应已经发生。如果均聚物的重复片段存在于模板序列中,则多个dNTP分子将在单循环中被并入。这导致相应数目的释放的氢和成比例地更高的电子信号。该技术不同于其他测序技术,因为没有使用被修饰的核苷酸或光学。离子半导体测序也可被称作离子激流测序(ion torrent sequencing)、pH介导的测序、硅测序或半导体测序。

[0324] 在焦磷酸测序中,聚合反应释放的焦磷酸根离子通过ATP硫酸化酶与腺苷5'磷酰硫酸反应以产生ATP;然后ATP驱动荧光素通过荧光素酶转化为氧化荧光素加光。由于荧光是瞬时的,所以在该方法中不需要单独的消除荧光的步骤。一次添加一种类型的脱氧核糖核苷酸三磷酸(dNTP),并根据哪种dNTP在反应位点产生明显信号来识别序列信息。市购可得的Roche GS FLX仪器利用这种方法获得序列。例如在Ronaghi等人,Analytical Biochemistry 242,84-89(1996)和Margulies等人,Nature 437,376-380(2005)(勘误表在Nature 441,120(2006))中详细地讨论了该技术及其应用。市购可得的焦磷酸测序技术为454测序(Roche)(例如,如在Margulies,M.等人Nature 437:376-380[2005]中描述的)。

[0325] 在连接测序中,连接酶被用来将具有突出端的部分双链的寡核苷酸连接至具有突出端的正被测序的核酸;为了使连接发生,这些突出端必须是互补的。部分双链的寡核苷酸的突出端中的碱基可根据偶联至该部分双链的寡核苷酸和/或偶联至与该部分双链的寡核苷酸的另一部分杂交的第二寡核苷酸的荧光团而被鉴定。在获得荧光数据之后,连接的复合体在连接位点上游被裂解,诸如通过在离其识别位点(其被包含于该部分双链的寡核苷酸中)固定距离的位点处切割的II型限制酶例如Bbvl,Bbvl。该裂解反应使刚好在之前的突出端上游的新突出端暴露,并重复所述程序。例如,在Brenner等人,Nature Biotechnology 18,630-634(2000)中详细讨论了该技术及其应用。在一些实施方案中,连接测序通过获得环状核酸分子的滚环扩增产物并将该滚环扩增产物用作连接测序的模板而适合于本发明的方法。

[0326] 连接测序技术的商业上可用的实例是SOLiD™技术(Applied Biosystems)。在

SOLiD™连接测序中，基因组DNA被剪切成片段，并且衔接子被连接至片段的5’和3’末端以产生片段文库。可选地，内部衔接子可通过以下步骤而被引入：将衔接子连接至片段的5’和3’末端、使片段环化、消化环化的片段以产生内部衔接子，并将衔接子连接至所得到的片段的5’和3’末端以产生配对文库。接下来，在包含珠、引物、模板和PCR组分的微反应器中制备克隆珠群。伴随PCR，模板被变性并且珠被富集以分离带有延伸的模板的珠。对所选择的珠上的模板进行3’修饰，所述3’修饰允许与载玻片结合。序列可通过特定的荧光团被鉴定的部分随机的寡核苷酸与中心的已确定的碱基(或碱基对)顺序性杂交和连接来确定。记录颜色之后，连接的寡核苷酸被裂解和去除，然后重复该程序。

[0327] 在可逆终止测序中，荧光染料标记的核苷酸类似物(由于封端基团的存在而成为可逆的链终止剂)被掺入单碱基延伸反应。碱基的身份(identity)根据荧光团来确定；换言之，每个碱基与不同的荧光团配对。获得荧光/序列数据之后，荧光团和封端基团通过化学方法被去除，并重复该循环以获得序列信息的下一个碱基。Illumina GA仪器通过该方法运行。例如，Ruparel等人，Proceedings of the National Academy of Sciences of the United States of America 102,5932-5937(2005)和Harris等人，Science320,106-109(2008)中详细讨论了该技术及其应用。

[0328] 可逆终止剂测序方法的商业上可获得的实例是Illumina的合成测序和基于可逆终止剂的测序(例如，在Bentley等人，Nature 6:53-59[2009]中描述的)。Illumina的测序技术依赖于片段化的基因组DNA与平面，任选地其上结合寡核苷酸锚的透明表面的连接。模板DNA的末端被修复以生成5’磷酸化的平末端，并且Klenow片段的聚合酶活性被用来将单个A碱基添加至平端的磷酸化DNA片段的3’末端。该添加制备了连接至寡核苷酸衔接子的DNA片段，所述寡核苷酸衔接子在其3’末端具有单个T碱基的突出端以增加连接效率。衔接子寡核苷酸与流动池的锚互补。在有限稀释条件下，衔接子修饰的单链模板DNA被添加至流动池并通过与锚的杂交而被固定。连接的DNA片段被延伸并被桥式扩增以生成具有数以亿计的簇的超高密度测序流动池，每个簇包含相同模板的～1,000个拷贝。这些模板利用稳健的四色DNA合成测序技术来测序，所述四色DNA合成测序技术采用具有可移除的荧光染料的可逆终止剂。高密度荧光检测利用激光激发和全内反射光学来完成。将约20-40bp例如36bp的短序列读段与重复片段-掩盖的(repeat-masked)参考基因组比对，并利用专门开发的数据分析流水线软件鉴定短序列读段在参考基因组上的独特位置。也可使用非重复片段-掩盖的参考基因组。不论使用重复片段-掩盖的参考基因组或非重复片段-掩盖的参考基因组，只计算独特地匹配至参考基因组的读段。第一次读取完成以后，可通过计算机模拟(in situ)重新生成模板使得能够进行来自片段的相反末端的第二次读取。因此，可使用DNA片段的单末端或成对末端测序。进行对存在于样品中的DNA片段的部分测序，并对被绘制至已知参考基因组的包含预定长度例如36bp的读段的序列标签计数。

[0329] 在纳米孔测序中，例如利用电泳驱动力使单链核酸分子穿过孔，并通过分析单链核酸分子通过孔时获得的数据来推导序列。数据可以是离子电流数据，其中每个碱基，通过例如部分地阻断通过孔的电流来将电流改变至不同的可辨别的程度，来改变电流。

[0330] 在另一个示例性但非限制性实施方案中，本文描述的方法包括利用透射电子显微术(TEM)获得序列信息。所述方法包括，利用选择性标记有重原子标记物的高分子量(150kb或更大)DNA的单原子分辨率透射电子显微成像，并将这些分子以具有一致的碱基间间距的

超密(链和链之间3nm)平行阵列排列在超薄的膜上。电镜被用来对薄膜上的分子成像,以确定重原子标记物的位置并提取DNA中的碱基序列信息。所述方法被进一步描述于PCT专利公布WO 2009/046445中。

[0331] 在另一个示例性但非限制性实施方案中,本文描述的方法包括,利用第三代测序获得序列信息。在第三代测序中,具有很多小的(～50nm)孔的铝涂层的玻片被用作零模式波导(参见,例如,Levene等人,Science 299,682-686(2003))。铝表面防止DNA聚合酶通过聚磷酸酯化学例如聚乙烯膦酸酯化学(参见例如Korlach等人,Proceedings of the National Academy of Sciences of the United States of America105,1176-1181(2008))附着。这导致DNA聚合酶分子优先附着至铝涂层的孔中暴露的硅。该设置允许待使用的渐失波现象(evanescent wave phenomena)减少荧光背景,允许使用较高浓度的荧光标记的dNTP。荧光团被连接至dNTP的末端磷酸根,以使得荧光在并入dNTP后释放,但荧光团不保持与新并入的核苷酸连接,意味着复合体立即准备好另一轮并入。通过该方法,dNTP至铝涂层的孔中存在的各个引物-模板复合体的并入可被检测到。参见例如Eid等人,Science323,133-138(2009)。

[0332] V.生成序列活性模型

[0333] 如以上所述,用于本文的实施方案的序列-活性模型将蛋白序列信息和蛋白活性联系起来。该模型使用的蛋白序列信息可采取很多种形式。在一些实施方案中,其是蛋白中的氨基酸残基的完整序列。但是,在一些实施方案中,完整的氨基酸序列是不必要的。例如,在一些实施方案中,只提供在特定的研究工作中是改变的那些残基已足够。在一些涉及后续研究阶段的实施方案中,很多残基是固定的,并且只有有限的序列空间的区域仍有待探索。在此类情况的一些中,提供这样的序列活性模型是方便的:作为输入,其只需要鉴定蛋白的需继续探索的区域中的那些残基。在一些另外的实施方案中,这些模型不要求知晓所述残基位置处的残基的准确身份。在一些此类实施方案中,表征特定残基位置处的氨基酸的一个或更多个物理或化学特性被鉴定。在一个示例性实例中,模型需要通过体积、疏水性、酸性等说明残基位置。此外,在一些模型中,采用此类特性的组合。事实上,不期望本发明被限于任何特定方法,因为这些模型对于序列信息、活性信息和/或其他物理特性(例如,疏水性等)的多种设置具备实用性。

[0334] 因此,序列-活性模型的形式可极大地变化,只要其如所期望的基于序列信息提供用于正确地评估蛋白的相对活性的工具。在一些实施方案中,这些模型通常把活性当作因变量,而把序列/残基值当作自变量。模型的数学/逻辑形式的实例包括,相加、相乘数学表达式和多阶的线性和非线性数学表达式、神经网络、分类和回归树/图、聚类方法、递归分区、支持向量机等。这些模型的一些实施方案包括多个相乘项的乘积。相乘项中的至少一些为非交互作用相乘项,其每个包含代表特定氨基酸或核苷酸对感兴趣的活性的贡献的系数。蛋白或和核酸序列中特定位置处的特定氨基酸或核苷酸为特定的残基类型。每个非交互作用相乘项还包括单一的自变量或哑变量,其代表特定氨基酸或核苷酸的存在或不存在。另外,在本发明的一些实施方案中,每个序列-活性模型包含代表蛋白变体的活性的因变量。

[0335] 在本发明的一些实施方案中,序列-活性模型还包括交互作用项,每个交互作用项包含代表(i)蛋白序列中第一位置处的第一氨基酸或核苷酸和(ii)蛋白序列中第二位置处

的第二氨基酸或核苷酸的特定组合对活性的贡献的交互作用系数。

[0336] 在本发明的一些实施方案中，模型通过相乘或相加组合交互作用项。在一些实施方案中，模型通过相乘或相加将交互作用项与非交互作用相乘项组合。在一些实施方案中，模型单纯地相乘，将非交互作用项和交互作用项组合成一个乘积。在其他实施方案中，模型包括通过相加与一个或更多个其他项组合的多个项的至少一个乘积。

[0337] 在本发明的一些替代实施方案中，序列-活性模型具有相加形式，所述相加形式包括一个或更多个非交互作用项和至少一个交互作用项的和。所述至少一个交互作用项是交叉乘积项，其包含代表一个相互作用残基的存在的一个变量和代表另一个相互作用残基的存在的另一个变量的乘积。

[0338] 在一些实施方案中，模型由提供活性和序列之间的数学/逻辑关系的活性对序列信息的训练组开发。该关系通常在用于预测新的序列的活性或残基对感兴趣的活性的影响之前被验证。

[0339] 用于生成模型的多种技术是可获得的并且在本发明中具备实用性。在一些实施方案中，这些技术涉及优化模型或使模型误差最小化。具体实例包括偏最小二乘法、集成回归、随机森林、多种其他回归技术以及神经网络技术、递归分区、支持向量机技术、CART(分类和回归树)等。通常，所述技术应产生能将对活性具有显著影响的残基与那些对活性没有显著影响的残基区分开的模型。在一些实施方案中，这些模型还对各个残基或残基位置基于其对活性的影响排序。不期望本发明局限于用于产生模型的任何特定技术，因为本领域已知的任何合适的方法在本发明中具备实用性。

[0340] 在涉及相加模型的一些实施方案中，这些模型通过回归技术产生，所述回归技术鉴定训练组中自变量和因变量的共同改变。多种回归技术是已知的并被广泛地使用。实例包括多元线性回归(MLR)、主成分回归(PCR)和偏最小二乘回归(PLS)。在一些实施方案中，模型利用涉及多个组分的技术，包括但不限于集成回归和随机森林产生。这些方法和任何其他合适的方法在本发明中具备实用性。不期望本发明局限于任何特定的技术。

[0341] MLR是这些技术中最基础的。其被用于简单地对训练组的成员的一组系数方程求解。每个方程涉及随特定位置处特定残基的存在或不存(即，自变量)的改变训练组成员的活性(即，因变量)。取决于训练组中残基选择的数目，这些方程的数目可以是相当大的。

[0342] 像MLR一样，PLS和PCR从将序列活性和残基值联系起来的方程产生模型。但是，这些技术以不同的方式产生模型。它们先进行坐标转换来减少自变量的数目。然后它们进行对所转换的变量的回归。在MLR中，存在可能非常大量的自变量：在训练组内变异的每个残基位置有两个或三个自变量。假定感兴趣的蛋白和肽往往相当大并且训练组可提供很多不同的序列，那么自变量的数目可能迅速地变得非常大。通过减少变量的数目以集中在数据组中提供最多变异的那些变量上，PLS和PCR通常需要较少的样品并简化了参与产生模型的步骤。

[0343] 实际的回归针对通过原始自变量(即，残基值)的坐标转换获得的相对少数目的潜在变量进行，在这方面PCR与PLS回归相似。PLS和PCR之间的不同在于，PCR中的潜在变量通过最大化自变量(即，残基值)之间的共变来构建。在PLS回归中，潜在变量以最大化自变量和因变量(即，活性值)之间的共变的方式来构建。偏最小二乘回归被描述于Hand,D.J.,等人(2001)Principles of Data Mining(Adaptive Computation and Machine Learning)，

Boston,MA,MIT Press和Geladi等人(1986)“Partial Least-Squares Regression:a Tutorial,”Analytica Chimica Acta,198:1-17中。这些参考均为了所有的目被通过引用并入本文。

[0344] 在PCR和PLS中,回归分析的直接结果是活性是加权的潜在变量的函数的表达式。通过进行将潜在变量转变回初始自变量的坐标转换,该表达式可转换成活性作为初始自变量的函数的表达式。

[0345] 大体上,PCR和PLS两者均先减少训练组中包含的信息的维度,然后对已被转换以产生的新的自变量但保留了初始因变量值的经转换的数据组进行回归分析。转换形式的数据组可产生仅相对少的用于进行回归分析的表达式。在未进行维度减少的方案中,必须考虑可能存在变异的每个单独的残基。这会是相当大的一组系数(例如,对于双向相互作用为2^N个系数,其中N为在训练组中可能变异的残基位置的数目)。在典型的主成分分析中,只采用3、4、5、6个主成分。

[0346] 机器学习技术拟合训练数据的能力往往被称为“模型拟合”,并且在诸如MLR、PCR和PLS的回归技术中,模型拟合通常通过测量值和预测值之间的方差和来测量。对于给定的训练组,最优的模型拟合将利用MLR来完成,而PCR和PLS往往具有较差的模型拟合(较高的测量和预测之间的误差平方和)。但是,利用潜在变量回归技术诸如PCR和PLS的主要优势在于此类模型的预测能力。获得具有很小的误差平方和的模型拟合绝对不能保证模型将能精确地预测训练组中未观察到的新样本-事实上,往往是相反的情况,特别是当存在很多变量并且只有少数观察值(即样本)时。因此潜在变量回归技术(例如,PCR、PLS)虽然往往具有对训练数据较差的模型拟合,但通常更稳健并且能更精确地预测训练组之外的新样本。

[0347] 能被用于产生根据本公开内容的模型的另一种工具类型是支持向量机(SVM)。这些数学工具将已基于活性分类成两个或更多个组的训练组序列作为输入。支持向量机通过根据训练组的不同成员如何接近超平面界面来不同地加权训练组的不同成员,所述超平面界面将训练组的“活性”和“非活性”成员分隔开。该技术需要科学家先决定将哪些训练组成员置于“活性”组以及将哪些训练组成员置于“非活性”组。在一些实施方案中,这通过选择针对活性水平的适当数值来完成,所述数值作为训练组的“活性”和“非活性”成员之间的分界线。支持向量机根据该分类生成向量W,其可为限定训练组中的活性和非活性组的成员的序列的各个自变量提供系数值。这些系数可被用于对如本文其他处描述的各个残基“排序”。该技术被用于鉴定超平面,所述超平面将超平面的相反侧最靠近的训练组成员之间的距离最大化。在另一个实施方案中,进行支持向量回归建模。在该情形中,因变量为连续的活性值的向量。支持向量回归模型产生系数向量W,其可被用来对各个体残基排序。

[0348] 在很多研究中SVM已被用来检查大的数据组,并且已广泛地与DNA微阵列一起应用。其潜在优势包括细微地辨别(通过加权)将样本彼此分开的因素的能力。在SVM能够精确地梳理出哪些残基对功能有贡献的程度上,其可以成为对残基排序特别有用的工具。SVM被描述于S.Gunn(1998)“Support Vector Machines for Classification and Regressions”技术报告,南安普敦大学工程与应用科学学院电子和计算机科学系,其为了所有目的被通过引用并入本文。

[0349] 在本发明的一些实施方案中,能用于产生模型的另一类工具基于分类树的集成利用随机输入来分类和回归,其实例是随机森林。参见Breiman(2001).“Random Forests”

Machine Learning 45(1):5-32。随机森林是树预测器的组合，以使得每个树取决于独立取样的随机向量的值并且森林中所有的树具有相同的分布。随机森林是由在决策树的每个分叉处具有随机选择特征的许多未修剪的决策树学习者组成的学习集成。随着森林中树的数目变大，森林的泛化误差收敛至极限值。

[0350] 随机森林可由下方式构建：

[0351] 1)如果训练组中案例的数目为N，随机地对N个案例抽样-但伴随从原始数据置换。该样本将是用于使树生长的训练组。

[0352] 2)如果有M个输入自变量，指定数值m<<M，以使得在树中的每个节点处，m个变量被随机地从M个变量随机选择出来并且对这些m的最佳分割被用于分割所述节点。m的值在森林生长期间保持恒定。

[0353] 3)在一些实施方案中，每个树生长至可能的最大程度。无修剪。

[0354] 4)然后产生大量的树，k＝1,…,K(通常K>＝100)。

[0355] 5)在大量的树产生之后，其全部对感兴趣的变量的类别投票。例如，其可各自促成活性或特定突变的贡献的最终预测。

[0356] 6)然后随机森林通过从森林中的所有树预测器选取最受欢迎的投票类别来对x(例如，突变或其他自变量的序列)分类。

[0357] 森林误差率取决于森林中任何两个树之间的相关性。增加相关性会增加森林误差率。森林误差率取决于森林中每个单独的树的强度。具有低的误差率的树是强的分类器(classifier)。增加各个树的强度会减小森林误差率。减少m会减少相关性和强度两者。增加m会增加两者。介于两者之间的是m的“最佳”范围-通常是相当宽的。

[0358] 随机森林技术可用于类别变量以及回归模型中的连续变量。在本发明的一些实施方案中，随机森林模型具有与SVM和神经网络模型相当的预测力，但是往往具有较高的计算效率，因为除了其他原因之外，交叉验证被构建到建模程序中且交叉验证的单独程序不是必要的。

[0359] A.相乘模型

[0360] 本发明的一些实施方案提供了构建包括多个相乘项的乘积的序列-活性模型并使用这些模型指导定向进化的方法。在一些实施方案中，序列-活性模型将蛋白变体的活性预测为序列的两个或更多个亚单位的存在或不存在的函数。在一些实施方案中，亚单位为组成蛋白变体的氨基酸。在一些实施方案中，亚单位为编码蛋白变体的核酸或密码子。

[0361] 在一些实施方案中，模型的相乘项中的至少一些为非交互作用相乘项，其每个包括代表特定氨基酸或核苷酸对感兴趣的活性的贡献的系数。蛋白或和核酸序列中特定位置处的特定氨基酸或核苷酸为特定的残基类型。每个非交互作用相乘项还包括单一的自变量或哑变量，其代表特定序列位置处特定类型的特定氨基酸或核苷酸的存在或不存在。在多个实施方案中，哑变量可作为比特值执行，诸如1代表特定残基的存在以及1/系数代表特定残基的不存在。在一些实施方案中，对密码子而非氨基酸或核酸建模。

[0362] 在本发明的一些实施方案中，序列-活性模型还包括交互作用项，每个交互作用项包括代表(i)蛋白序列中第一位置处的第一氨基酸或核苷酸和(ii)蛋白序列中第二位置处的第二氨基酸或核苷酸的特定组合对活性的贡献的交互作用系数。

[0363] 如以上解释的，交互作用项的集通常是大的。包括大量的交互作用项不利地影响

定向进化的效率。另外的项也趋于导致数据的过度拟合，因此降低模型的预测力。相乘模型可能能减少交互作用项的数目并改进模型的预测力，因为该模型的多个非交互作用项形成乘积，允许多个非交互作用项的自变量以相乘的方式对因变量做出贡献。不被本文假设的任何理论束缚，相乘模型可改进模型的预测力和/或定向进化的效率。

[0364] 当两个或更多个密码子编码相同的氨基酸时，存在密码子简并性。在一些实施方案中，密码子简并性可任选地通过使用相同的自变量代表两个或更多个简并的密码子而被剔除。但是，在一些实施方案中密码子简并性也可保留在模型中。例如，一个模型可将具有同义突变的两种不同的核酸与相同的蛋白活性相关联。相乘模型可充分地捕获活性和蛋白序列信息之间的关系的原因可使用Arrhenius方程式将蛋白活性描述为变体蛋白中的多个突变的函数来阐明。原始的Arrhenius方程式：$k = A\,e^{-\frac{\Delta G}{RT}}$ 将化学反应速率常数k描述为活化能ΔG和反应温度T(开氏温度)的函数，其中A为预指数因子(经验常数)，并且R为通用气体常数8.314J/mol.K。Arrhenius方程式可被改写为：$\ln(k) = -\frac{\Delta G}{RT} + \ln(A)$，显示出反应速率常数与温度T正相关而与活化能负相关。换言之，较高的反应温度导致较快的反应，而较高的活化能导致较慢的反应。

[0365] 从概念上讲将化学反应的反应速率与蛋白活性对应并行，并且使活化能与突变对蛋白活性的贡献并行是明智的。利用所述对应，Arrhenius方程式可被用于将蛋白活性k描述为突变的影响ΔG的函数。如果蛋白的骨架残基具有ΔG0的影响，突变1具有ΔG1的影响，且突变2具有ΔG2的影响，则突变1和2的组合影响可表示为：

[0366] $$k = Ae^{-\frac{(\Delta G_0+\Delta G_1+\Delta G_2)}{RT}} = A(e^{-\frac{\Delta G_0}{RT}} \times e^{-\frac{\Delta G_1}{RT}} \times e^{-\frac{\Delta G_2}{RT}})$$。

[0367] 如在该方程式中观察到的，两个突变的影响可组合成相乘形式的乘积。在本发明的一些实施方案中，相乘模型可表示为多个项的乘积。每个项代表序列中的一个位置和该位置处特定残基或核苷酸/密码子的特定组合对活性的贡献。这些项中的每个包括系数和自变量。自变量可采取以下两个值中的一个：当所研究的突变存在时的一个值并且当研究的突变不存在时的一个不同的值。

[0368] 在一些实施方案中，序列活性模型中的每个项可以简单地是其系数和相关的自变量的简单积。在本发明的一些实施方案中，当与该项相关的核苷酸或残基存在时，自变量的值被设为1当自变量不存在时，变量的值为该项的系数值分之一，或$1/C_n$。对于交互作用项，提供对于相互作用对活性的贡献的系数。当所有的相互作用成员(m和n)存在时，相关系数具有1的值，并且当它们不存在时，具有值$1/C_{m,n}$。

[0369] 在另一种形式的模型中，在每个项中当所研究的残基或核苷酸存在于所述位置时自变量被给予1的值，并且当其不存在于所述位置时被给予0。在该形式的模型中，该项被给予1加上系数乘以自变量的值：$(1+C_nx_n)$。与第一种情况相似地处理交互作用项：$(1+C_{m,n}x_mx_n)$。在一个实施方案中，模型可采用这种形式：

[0370] $y = (1+C_1x_1) \times (1+C_2x_2) \times (1+C_3x_3) \times \ldots \times (1+C_nx_n)$。

[0371] 在该模型中，交互作用项可存在或可不存在。就其本质而言，乘积对相互作用残基之间的相互作用做出解释是可能的。在相乘模型不固有地对此类相互作用的所有贡献做出

解释的情况中，交互作用项可作为项被包括在相乘模型中。在该情况中，一个或更多个交互作用项可通过包括代表存在两个或更多个突变的系数和代表两个突变的哑变量而被并入：

[0372] $(1+C_{1,2}x_1x_2)\times(1+C_{2,3}x_2x_3)\times\ldots\times(1+C_{m,n}x_mx_n)$。

[0373] 在一些实施方案中，交互作用项与非交互作用项相乘。在其他实施方案中，交互作用项可通过求和而与非交互作用项组合。在以上实施方案中，交互作用项可通过乘法而在其自身间组合。在替代实施方案中，交互作用项可通过加法而在其自身间组合。

[0374] 表1提供了包括蛋白骨架和具有以下4个感兴趣的位置的变体的实例：10、166、175和340。骨架残基在10处为Ala、在166处为Phe、在175处为Gly且在340处为Phe。对于骨架，蛋白的活性被测量为y，而对于变体，蛋白的活性被测量为y_n。

[0375] 表I.交互作用模型的示例性序列活性数据*

[0376]

	位置				
	10	166	175	340	y(活性)
骨架	Ala	Phe	Gly	Phe	y
变体 1	Ala	Ser	Gly	Phe	y_1
变体 2	Asp	Phe	Val	Ala	y_2
变体 3	Lys	Leu	Gly	Ala	y_3
变体 4	Asp	Ile	Val	Phe	y_4
变体 5	Ala	Ile	Val	Ala	y_5

[0377]

[0378] *在每个变异位置处的所有可能的氨基酸可被考虑。在该实例中，只有变体内的突变(相对于骨架)被考虑。

[0379] 以下形式的相乘模型被用于该实例，

[0380] $y_n=(1+C_{10Asp}X_{10Asp})*(1+C_{10Lys}X_{10Lys})*(1+C_{166Ser}X_{166Ser})$

[0381] $*(1+C_{166Leu}X_{166Leu})*(1+C_{166Ile}X_{166Ile})$

[0382] $*(1+C_{175Val}X_{175Val})*(1+C_{340Ala}X_{340Al}a)$

[0383] 技术人员可通过以下模型获得变体2的预测活性：

[0384] $y_2(1+C_{100Asp}*1)*((1+C_{10Lys}*0)*(1+C_{166Ser}*0)$

[0385] $*(1+C_{166}L_{eu}*0)*((1+C_{166Ile}*0)$

[0386] $*(1+C_{175Val}*1)*(1+C_{340Ala}*1)$

[0387] $=(1+C_{10Asp}*1)*(1+C_{175Val}*1)*(1+C_{340Ala}*1)$ 然而模型方程式应包括变异位置处的

所有氨基酸,以使得:

[0388] $y_n=(1+C_{10Ala}X_{10Ala})*(1+C_{10Asp}X_{10Asp})*(1+C_{10Lys}X10_{Lys})*(1+C_{166Ser}X_{166Ser})$

[0389] $*(1+C_{166phe}X_{166Phe})*(1+C_{166Leu}X_{166Leu})*(1+C_{166Ile}X_{166Ile})$

[0390] $*(1+C_{175Gly}X_{175}Gly)*(1+C_{175Val}X_{175Val})*(1+C_{340phe}X_{340phe})*(1+C_{340Ala}X_{340Ala})$

[0391] 骨架中的项往往被从方程式中剔除。该方法是可接受的,因为往往所述方法使用变体的改进倍数(FIOP)拟合模型,因此骨架氨基酸系数为0(以使得骨架的y=1)。

[0392] 在一些替代实施方案中,另一种方法是使用查阅表,在所述查阅表中对于位置和相关突变的每一个和各个组合提供系数值。为了预测变体序列(突变的集合)的活性,该算法简单地鉴定存在于变体中的那些突变并从查阅表中选取那些突变的相关系数值。然后所选择的系数彼此相乘以产生预测的活性值,所述预测的活性值为因变量。

[0393] 图3I-3J显示了这样的示例,其中相乘模型比相加模型更好地拟合观察值,在某些情况下提供较高的预测力。观察数据包含含有趋于导致对蛋白活性的严重有害影响的突变的蛋白变体。图3I绘制了非交互作用相加模型的预测活性值对蛋白变体的训练组的观察到的活性值的图。明显地,相加模型的预测低估了具有高的活性值的观察对象。相比而言,如图3J中显示的相乘模型的预测活性值更好地匹配观察到的活性值,具有显著地小于相加模型小的误差。

[0394] B.线性相加模型

[0395] 在一些实施方案中,线性模型被用作产生非线性模型的逐步过程中的"基本"模型。通常,活性对序列的线性回归模型具有以下形式:

[0396] $$y=c_0+\sum_{i=1}^{N}\sum_{j=1}^{M}c_{ij}x_{ij} \qquad (1)$$

[0397] 在该线性表达式中,y为预测的响应,而c_{ij}和x_{ij}分别为序列中位置i处的回归系数和用于表示残基选择的比特值或哑变量。在蛋白变体文库的序列中存在N个残基位置,并且这些残基位置中的每个可被一种或更多种残基占据。在任何给定的位置处,可能存在j=1至M个单独的残基类型。该模型假定每个位置处的残基之间的线性(相加)关系。方程1的展开形式如下:

[0398] $y=c_0+c_{11}x_{11}+c_{12}x_{12}+\cdots c_{1M}x_{1M}+c_{21}x_{21}+c_{22}x_{22}+\cdots c_{2M}x_{2M}+\ldots+c_{NM}x_{NM}$

[0399] 如所示的,活性和序列信息的形式的数据源自初始蛋白变体文库并被用于确定模型的回归系数。哑变量首先被从蛋白变异序列的比对中鉴定。氨基酸残基位置被从蛋白变异序列中鉴定,在所述蛋白变异序列中那些位置的氨基酸残基因序列而异。在这些可变残基位置的一些或全部位置的氨基酸残基信息可被整合到序列-活性模型中。

[0400] 表II包含对于10个例证性变体蛋白的可变残基位置和残基类型的形式的序列信息,以及对应于每个变体蛋白的活性值。这些是生成足以解决所有系数的方程式所需的更大的组的代表性成员。因此,例如,对于表II中的例证性蛋白变体序列,位置10、166、175和340为可变残基位置并且所有其他位置,即在该表中未示出的那些位置,包含在变体1-10之间相同的残基。

[0401] 在该实例中,10个变体可包含或可不包含野生型骨架序列。在一些实施方案中,被开发以考虑包含野生型骨架序列的所有变体的数据的模型可能引出完美多重共线性或哑

变量陷阱的问题。该问题可通过多种技术来解决。一些实施方案可在开发该模型时将野生型骨架数据排除在外。一些实施方案可放弃代表野生型骨架的那些系数。一些实施方案可使用诸如PLS回归的技术来解决多重共线性。

[0402] 表II:示例性序列和活性数据

[0403]

变异残基位置	10	166	175	340	y (活性)
变体 1	Ala	Ser	Gly	Phe	y_1
变体 2	Asp	Phe	Val	Ala	y_2
变体 3	Lys	Leu	Gly	Ala	y_3
变体 4	Asp	Ile	Val	Phe	y_4
变体 5	Ala	Ile	Val	Ala	y_5
变体 6	Asp	Ser	Gly	Phe	y_6
变体 7	Lys	Phe	Gly	Phe	y_7
变体 8	Ala	Phe	Val	Ala	y_8
变体 9	Lys	Ser	Gly	Phe	y_9
变体 10	Asp	Leu	Val	Ala	y_{10}

[0404] 因此,基于方程式1,回归模型可从表II中的系统变异的文库推导出。

[0405] $y=c_0+c_{10\ Ala}X_{10Ala}+c_{10Asp}X_{10Asp}+c_{10Lys}X_{10Lys}+c_{166Ser}X_{166ser}+c_{166Phe}X_{166Phe}+c_{166Leu}X_{166Leu}+c_{166Ile}X_{166Ile}+c_{175Gly}X_{175Gly}+c_{175Val}X_{175Val}+c_{340Phe}X_{340phe}+c_{340Ala}X_{340Ala}$ (方程2)

[0406] 比特值(x哑变量)可表示为反映所指定的氨基酸残基的存在或不存在的1或0,或者可选地1或-1,或一些其他替代表示。例如,使用1或0指定方式,x_{10Ala}对于变体1可以为"1"而对于变体2可以为"0"。使用1或-1指示,x_{10Ala}对于变体1可以为"1"而对于变体2可以为"-1"。回归系数可由此基于针对文库中的所有变体的序列活性信息从回归方程推导出。对于变体1-10的此类方程(对于x使用1或0指示)的实例遵循:

[0407] $y_1=c_0+c_{10Ala}(1)+c_{10Asp}(0)+c_{10Lys}(0)+c_{166Ser}(1)+c_{166Phe}(0)+c_{166Leu}(0)+c_{166Ile}(0)+c_{175Gly}(1)+c_{175Val}(0)+c_{340Phe}(1)+c_{340Ala}(0)$

[0408] $y_2=c_0+c_{10Ala}(0)+c_{10Asp}(1)+c_{10Lys}(0)+c_{166Ser}(0)+c_{166Phe}(1)+c_{166Leu}(0)+c_{166Ile}(0)+c_{175Gly}(0)+c_{175Val}(1)+c_{340Phe}(0)+c_{340Ala}(1)$

[0409] $y_3=c_0+c_{10Ala}(0)+c_{10Asp}(0)+c_{10Lys}(1)+c_{166Ser}(0)+c_{166Phe}(0)+c_{166Leu}(1)+c_{166Ile}(0)+c_{175Gly}(1)+c_{175Val}(0)+c_{340Phe}(0)+c_{340Ala}(1)$

[0410] $y_4=c_0+c_{10Ala}(0)+c_{10Asp}(1)+c_{10Lys}(0)+c_{166Ser}(0)+c_{166Phe}(0)+c_{166Leu}(0)+c_{166Ile}(1)+c_{175Gly}(0)+c_{175Val}(1)+c_{340Phe}(1)+c_{340Ala}(0)$

[0411] $y_5=c_0+c_{10Ala}(1)+c_{10Asp}(0)+c_{10Lys}(0)+c_{166Ser}(0)+c_{166Phe}(0)+c_{166Leu}(0)+c_{166Ile}(1)+c_{175Gly}(0)+c_{175Val}(1)+c_{340Phe}(0)+c_{340Ala}(1)$

[0412] $y_6=c_0+c_{10Ala}(0)+c_{10Asp}(1)+c_{10Lys}(0)+c_{166Ser}(1)+c_{166Phe}(0)+c_{166Leu}(0)+c_{166Ile}(0)+c_{175Gly}(1)+c_{175Val}(0)+c_{340Phe}(1)+c_{340Ala}(0)$

[0413] $y_7=c_0+c_{10Ala}(0)+c_{10Asp}(0)+c_{10Lys}(1)+c_{166Ser}(0)+c_{166Phe}(1)+c_{166Leu}(0)+c_{166Ile}(0)+$

$c_{175Gly}(1)+c_{175Val}(0)+c_{340Phe}(1)+c_{340Ala}(0)$

[0414] $y_8=c_0+c_{10Ala}(1)+c_{10Asp}(0)+c_{10Lys}(0)+c_{166Ser}(0)+c_{166Phe}(1)+c_{166Leu}(0)+c_{166Ile}(0)+c_{175Gly}(0)+c_{175Val}(1)+c_{340Phe}(0)+c_{340Al}a(1)$

[0415] $y_9=c_0+c_{10Ala}(0)+c_{10Asp}(0)+c_{10Lys}(1)+c_{166Ser}(1)+c_{166Phe}(0)+c_{166Leu}(0)+c_{166Ile}(0)+c_{175Gly}(1)+c_{175Val}(0)+c_{340Phe}(1)+c_{340Ala}(0)$

[0416] $y_{10}=c_0+c_{10Ala}(0)+c_{10Asp}(1)+c_{10Lys}(0)+c_{166Ser}(0)+c_{166Phe}(0)+c_{166Leu}(1)+c_{16Ile}(0)+c_{175Gly}(0)+c_{175Val}(1)+c_{340Phe}(0)+c_{340Ala}(1)$

[0417] 方程的全集可利用任何合适的回归技术(例如,PCR、PLS或MLR)容易地解决,以确定相应于感兴趣的每个残基和位置的回归系数的值。在该实例中,回归系数的相对量级与特定位置处的特定残基对活性的贡献的相对量级相互关联。然后可对回归系数排序或以其他方式分类以确定哪些残基更可能对期望的活性做出有有利的贡献。表III提供了相应于表II中示例的系统变异文库的例证性回归系数值:

[0418] 表III:回归系数的示例性排列次序

[0419]

回归系数	值
c_{166Ile}	62.15
c_{175Gly}	61.89
c_{10Asp}	60.23
$c_{340\ Ala}$	57.45
$c_{10\ Ala}$	50.12
$c_{166\ Phe}$	49.65
c_{166Leu}	49.42
$c_{340\ Phe}$	47.16
c_{166Ser}	45.34
$c_{175\ Val}$	43.65
$c_{10\ Lys}$	40.15

[0420] 回归系数的排列次序表可被用于构建对于期望的活性被优化的新的蛋白变体文库(即,改进的适应性)。这可以多种方式进行。在一个实施方案中,其通过保留具有最高观察值的系数的氨基酸残基来完成。这些氨基酸残基是被回归模型指示对期望的活性贡献最多的残基。如果采用负描述符来鉴定残基(例如,对于亮氨酸为1以及对于甘氨酸为-1),则基于系数的绝对值排列残基位置变得成为必要。注意在此类情况中,对于每个残基通常只有一个系数。系数量级的绝对值提供对相应的残基位置的排序。然后,考虑各个残基的符号来确定它们中的每个在期望的活性方面是否是有害的或有益的变得成为必要。

[0421] C.非线性相加模型

[0422] 采用非线性建模来解释蛋白中对活性有贡献的残基-残基相互作用。N-K格局

(landscape)描述了这个问题。参数N指在相关多肽序列的集合中可变残基的数目。参数K表示在这些多肽的每个的各个残基之间的相互作用。相互作用通常是在多肽的初级、二级或三级结构中多个残基之间的紧密物理接近度的结果。相互作用可归因于直接相互作用、间接相互作用、物理化学相互作用、由折叠中间体引起的相互作用、平移效应(translational effect)等。参见Kauffman,S.和Levin,S.(1987),"Towards a general theory of adaptive walks on rugged landscapes",Journal of Theoretical Biology 128(1)11-45。

[0423] 参数K被定义为使得对于值K=1,每个可变残基(例如,它们中有20个可变残基)恰好与其序列中的一个其他残基相互作用。在所有的残基与所有其他残基的影响物理地且化学地分隔开的情况中,K的值为0。明显地,取决于多肽的结构,K可具有大量不同的值。对于正在研究的多肽的难以解决的结构,K的值可被评估。但是,往往并非如此。

[0424] 多肽活性的纯线性、相加模型(如以上描述的)可通过包括表示2个或更多个残基之间的特定相互作用的一个或更多个非线性、交互作用项来改进。在以上呈现的模型形式的背景中,这些项被描述为"交叉乘积",所述"交叉乘积"包含表示相互作用而对活性具有显著的有利或不利影响的两个或更多个特定残基(每个与序列中的特定位置有关)的两个或更多个哑变量。例如,交叉乘积项可具有$c_{ab}x_a x_b$的形式,其中x_a为哑变量,表示序列中特定位置处特定残基的存在或不存在,并且变量x_b表示多肽序列中不同位置(其与第一个位置相互作用的位置)处特定残基的存在。下文展示了该模型的详细示例形式。

[0425] 交叉乘积项中表示的所有残基(即,在特别鉴定的位置处两个或更多个特定类型的残基中的每个)的存在影响多肽的整体活性。所述影响可以以很多种方式表现。例如,各个相互作用残基的每一个当单独存在于多肽中时,可能对活性具有不利影响,但是当它们都存在于多肽中时,整体影响是有利的。在其他情形中,相反的情况可能是正确的。另外,可能产生协同效应,其中各个残基的每一个单独地对活性具有相对有限的影响,但当它们都存在时,对活性的影响大于所有单独残基的累积影响。

[0426] 在一些实施方案中,非线性模型包括对于序列中相互作用的可变残基的每种可能的组合的交叉乘积项。但是,这并不代表物理现实(physicalreality),因为事实上只有可变残基的子集彼此相互作用。另外,这会导致"过度拟合"而产生提供虚假结果的模型,所述虚假结果是用来创建模型的特定多肽的表现方式并且不代表多肽内的真实相互作用。代表物理现实并避免过度拟合的模型的交叉乘积项的正确数目通过K值指定。例如,如果K=1,则交叉乘积交互作用项的数目等于N。

[0427] 在构建非线性模型时,在一些实施方案中,鉴定代表对活性具有显著影响的真实结构相互作用的那些交叉乘积交互作用项是重要的。这可通过多种方式完成,包括但不限于正向相加(forward addition),其中候选交叉乘积项被一次一个地添加至仅初始的线性模型直到项的添加不再是统计学上显著的;以及反向相减(reverse substraction),其中所有可能的交叉乘积项被提供在初始模型中并一次去除一个。以下呈现的示例性实例涉及使用逐步加法和减法技术以鉴定有用的非线性项。

[0428] 在一些实施方案中,产生包含此类交互作用项的非线性模型的方法与以上描述的用于产生线性模型的方法相同。换言之,采用训练组将数据"拟合"成模型。但是,一个或更多个非线性项,优选地以上讨论的交叉乘积项被加至模型。进一步地,可采用所得到的非线

性模型，如以上描述的线性模型，来对多个残基对多肽的整体活性的重要性排序。多种技术可用来鉴定由非线性方程预测的可变残基的最佳组合。下文描述了对残基排序的方法。在一些实施方案中，即使当局限于仅由两个残基导致的相互作用时，使用了非常大数量的对于可变残基的可能的交叉乘积项。随着更多的相互作用发生，非线性模型要考虑的潜在的相互作用的数目以指数方式生长。如果该模型包括包含三个或更多个残基的相互作用的可能性，则潜在项的数目生长得甚至更快。

[0429] 在简单的示例性实例中，其中存在20个可变残基并且K＝1(该实例假设每个可变残基与一个其他的可变残基相互作用)，该模型中应有20个交互作用项(交叉乘积)。如果存在任何更少的交互作用项，则该模型将不能充分描述这些相互作用(尽管一些相互作用可能对活性不具有显著影响)。相比而言，如果有任何更多的交互作用项，则该模型可能过度拟合所述数据组。在该实例中，存在N*(N-1)/2或190个可能的相互作用对。找到描述序列中的20种相互作用的20个独特的对的组合是重要的计算问题，因为存在约5.48x 10^{26}种可能的组合。

[0430] 可采用很多技术来鉴定相关的交叉乘积项。取决于问题的容量和可用的计算力，探索所有可能的组合并从而鉴定出最佳符合数据的一个模型是可能的。但是，往往问题在计算方面是高要求的。因此，在一些实施方案中，其利用有效搜索算法或近似法。如本文表明的，一种合适的搜索技术是逐步技术。但是，不期望本发明被限于用于鉴定相关交叉乘积项的任何特定方法。

[0431] 示例性实例在以下表IV中展示，显示了并入到从序列信息预测活性的模型中的非线性交叉乘积项的值。该实例是这样的非线性模型：其中假定在序列中的每个可变位置处只有两个残基选项。在该实例中，通过使用哑变量将蛋白序列映射到编码序列中，所述哑变量分别使用+1和-1对应选择A或选择B。该模型不受任意地选择哪些数值用于指定每个残基选择的影响。表IV第一行中显示的变量位置不指示蛋白序列的实际序列位置。相反，它们是代表蛋白序列中可具有如表IV对于残基选择A和残基选择B在第二和第三行显示的两种选项之一的变化的任意10个假定位置的任意标记。

[0432] 表IV：在每个具有两种选择的位置处的编码残基的实例

[0433]

可变位置标记	1	2	3	4	5	6	7	8	9	10
残基选择A	I	L	L	M	G	W	K	C	S	F
残基选择B	V	A	I	P	H	N	R	T	A	Y
蛋白残基选择	V	A	L	P	G	W	K	T	S	F
模型代码值	-1	-1	1	-1	1	1	1	-1	1	1

[0434] 对于该编码方案，用于将蛋白序列与活性联系起来的线性模型可如下书写：

[0435] $y＝c_1x_1+c_2x_2+c_3x_3...+c_nx_n+...+c_Nx_N+c_0$(方程3)

[0436] 其中y为响应(活性)，c_n是在位置n处残基选择的回归系数，x是在位置n处对于残基选择的哑变量编码(+1/-1)，且c_0是响应的平均值。这种形式的模型假定不存在可变残基之间的相互作用(即，每个残基选择独立地对蛋白的整体适合性做出贡献)。

[0437] 非线性模型包括一定数目的(尚未确定的)交叉乘积项来考虑残基之间的相互作用：

[0438] $y＝c_1x_1+c_2x_2+c_3x_3+\cdots+c_nx_n+c_{1,2}x_1x_2+c_{1,3}x_1x_3+c_{2,3}x_2x_3+...+c_0$(方程4)

[0439] 其中变量与在方程(3)中的那些变量相同，但是现在存在非线性项，例如，$c_{1,2}$为可变位置1和2之间相互作用的回归系数。

[0440] 为了评价线性和非线性模型的性能，使用被称为NK格局的合成数据来源(Kauffman和Levin,1987)。如以上提到的，N为模拟蛋白中可变位置的数目并且K为残基之间的强性偶联(epistatic coupling)。另外，合成数据组通过计算机模拟产生。

[0441] 该数据组被用来产生S＝40个合成样本、N＝20个可变位置并且K＝1(再次重申，对于K＝1，每个可变位置与一个其他可变位置功能性地偶联)的初始训练组。在产生随机蛋白时，每个可变位置具有相等的含有哑变量+1或-1的概率。对于合成训练组的每个成员，残基-残基相互作用(通过交叉乘积表示)和实际活性是已知的。产生另外的V＝100个样本用于验证。再次地，对于验证组的每个成员，残基-残基相互作用和活性是已知的。

[0442] 训练组被用于构建线性和非线性模型。生成选择交叉乘积项的一些非线性模型并生成不选择此类项的另外的非线性模型。图3A-F的模型利用遗传逻辑建模方法生成，而图3G-3H的模型利用逐步建模方法生成。尽管具有线性项和非线性项两者的模型相对于只具有线性项的模型的量性优势在遗传逻辑建模方法和逐步建模方法间不同，但是不论是哪种建模方法，结果表明具有非线性项的模型的概括能力的优势。事实上，不期望本发明被限于任何特定的方法，因为任何合适的建模方法在本发明中具备适应性。

[0443] 对于以上描述的S＝40的训练组容量，线性模型能相当好地关联测量值和预测值，但是当未对在训练组中未见到的数据进行验证时，线性模型显示出较弱的关联性(参见图3A)。如显示的，暗数据点代表40个训练数据点的观察到的活性对线性模型所做的预测。亮数据点代表从40个训练样本构建并用来预测验证样本V的相同模型进行的预测，所述验证样本无一见于初始训练组。验证组提供了对模型的实际预测能力的良好测量，与训练组相反，其可能经受模型过度拟合的问题，特别是对于以下描述的非线性情况。

[0444] 考虑到线性模型被用于模拟非线性适应情形，以上描述的S＝40的训练组的这种结果是值得注意的。在该情形中，线性模型最多能捕获选择给定残基对适应性的平均贡献。假定组合考虑足够数目的平均贡献，线性模型粗略地预测了实际测量到的响应。当训练组大小增加至S＝100时，线性模型的验证结果稍微转好(参见图3B)。相对简单的模型不充分拟合数据的趋势被称为“偏差”。

[0445] 当使用只有S＝40个样本训练非线性模型时，训练组成员的相关性是很好的(参见图3C)。不幸的是，在该示例性实例中，模型在训练组外提供有限的预测力，如通过其在验证组中与测量值的有限相关性所证实的。当存在很多(210种可能)潜在的变量，而有助于鉴定合适的交叉乘积项的训练数据有限时，该非线性模型基本上只能记住其被训练的数据组。这种高复杂性模型过度拟合数据的趋势被称为“方差”。偏差-方差权衡代表了机器学习中的基本问题，并且当处理新的或非典型的机器学习问题时几乎总是需要一些形式的验证来解决它。

[0446] 但是，当利用如在图3D中显示的较大的训练组(S＝100)训练非线性模型时，对于训练预测和更重要的验证预测两者，非线性模型表现得均极为出色。验证预测是足够精确的以致于大部分的数据点被用于标绘训练组的黑圆圈遮盖。

[0447] 为了比较的目的，图3E和3F显示了未仔细挑选交叉乘积项时所制作的非线性模型的性能。不像图3C和3D中的模型，每个可能的交叉乘积项被选择(即，对于N＝20，190个交叉

乘积项)。如在这些图中显示的,预测验证组活性的能力与仔细挑选交叉乘积项时生成的非线性模型的能力相比相对差。这种差的预测验证数据的能力是过度拟合的体现。

[0448] 图3G和3H分别显示了对于计算机模拟的数据,线性模型和逐步、非线性模型的通过剩余误差指示的预测力。逐步非线性模型如以上概括性地描述的以及以下更具体地描述的实施。

[0449] 为了测试这些模型,创建模拟数据。随机数发生器R基于具有平均数MN和标准偏差SD的正态分布创建。然后限定10个突变的组。这些突变的命名为M1、M2…M10(该命名方案是任意的)。该步骤模拟多样性的生成。

[0450] 每个突变代表蛋白序列中给定位置处的氨基酸变化,并且每个位置独立于其他位置。以上每个突变具有基于R(MN=0、SD=0.2)分配的随机活性值A。选择以上的6个突变被并一起配对成三对P。这些对代表突变之间的强性相互作用。

[0451] 基于R(MN=0、SD=0.2)对每对P分配活性值AP。构建50个变体的文库L,其中每个变体包含随机数目的以上限定的突变M-突变M的随机数目由经四舍五入的R(MN=4、S=0.25)的绝对值限定。该步骤模拟文库构建和测序。

[0452] L中每个变体的活性通过以下步骤来计算:首先将来自每个成对突变(如果两个突变均存在)的活性值AP加至1.0(野生型、无突变序列的限定活性),然后加上剩余的单突变的值(A)。测定噪音通过将来自R(MN=0、SD=0.005)的随机值加至针对每个变体的最终值来模拟。该步骤模拟变体的筛选。

[0453] 线性模型LM基于来自最后的步骤的数据构建。该模型含有10个自变量/系数,每个代表M的一个突变。然后线性模型利用普通的最小二乘回归和以上获得的数据拟合。

[0454] 递加法被用来基于以上获得的数据选择模型MM、基础模型为LM、利用AIC作为选择标准,并且选择只含有代表单突变和成对相互作用的系数的模型。对于模型选择方法的详细描述,参见以下的模型选择说明。通过AIC选择的最佳模型利用普通最小二乘回归拟合。

[0455] 为了评价线性模型和非线性模型的预测能力,重复以上描述的程序20次。针对模拟数据测算模型的预测力,其中图3G显示了线性模型并且图3H显示了逐步非线性模型。以上描述了用来预测单突变的值的模型。该预测通过利用预测只含有一个感兴趣的突变并减去1.0(野生型)的变体的模型来进行。如从图3G和3H明显看出的,非线性模型更精确地预测具有较大线性度趋势和较小残差的值。

[0456] D.模型选择

[0457] 基础序列活性模型生成之后,程序通过调整模型的项的系数值以使模型的预测值和观察值之间的剩余误差最小化来精化模型。参见模块107。该类型的调整也被称为模型拟合。可使用本领域已知的多种模型拟合方法。例如,遗传算法可用来调整系数值。对于相加模型,多种回归技术可用来拟合模型。

[0458] 在本发明的一些实施方案中,该程序还通过选择待被包括在模型内或待被从模型排除的合适的项以使得剩余误差最小化和/或改进模型的预测力来精化模型。参见模块107。由于所考虑的模型具有的项都选自相同的项集合,该精化过程也被称为嵌套模型中的模型选择。本发明的一些实施方案使用遗传算法选择合适的项。另外地或可选地,本发明的一些实施方案对基础模型项迭代地加上或减去来自可用的交互作用项的集合的交互作用项,并评价所得到的新模型相比基础模型的改进以产生最终模型。

[0459] 调整模型系数值和选择模型的项以提高模型的预测力两者均被称为模型优化技术。下文展示了用于根据模块105和107中的操作生成序列-活性模型的示例性算法。此类技术包括但不限于偏向于将另外的交互作用项纳入模型中的的遗传算法和逐步技术。但是，不期望本公开内容被限制于这些特定的实例。

[0460] 遗传算法

[0461] 本发明的一些实施方案提供了使用遗传算法选择序列-活性模型的一个或更多个项的方法。其他的实施方案提供了使用遗传算法调整系数的值以使模型符合所获得的数据的方法。

[0462] 在遗传算法中，合适的适应性函数和合适的交配程序(mating procedure)被定义。适应性函数提供了用于确定哪些模型(交叉乘积项的组合)“最适合”或具有最高的预测力(即，可能提供最佳结果)的准则。在一些实施方案中，该算法提供了搜遍参数空间以找到最佳参数值(即序列-活性模型的系数)的机制。在这些实施方案中，群中的个体或染色体中的每一个包括代表被测试的所有系数的基因，和对于这些系数具有在特定范围内选择的值的基因。例如，染色体可具有代表针对位置131处的Gly、具有0.4的值的系数的基因。

[0463] 在一些实施方案中，遗传算法还可被用来为模型选择合适的项。该遗传算法的一个实例可以类似于之前的实例，除非所有的个体/染色体包括代表所有合格参数(系数)的所有基因，并且这些基因的值允许被假定为0的值。如果在运算结束时，最适合的个体中一个项的系数收敛至0，则该项被从模型剔除。相反地，该项被保留。

[0464] 在其他实施方案中，交配程序提供了用于引入来自前一代的成功的“亲本”模型的交互作用项的新组合的机制。

[0465] 遗传算法的一个实例为通过调整系数的值，并任选地选择合适的项以使模型的预测力最大化来使模型符合数据。该实例被参考图1B描述。该算法从包含均具有可变的系数值的多个可能的模型的第一代开始，其中一些模型比其他模型更好地代表物理实体。参见模块131。第一代和每个后续的代被表示为“群”中的很多模型(也被称为个体或染色体)。每个模型/染色体包括代表该代中被测试的所有项的系数值的基因。该遗传算法在通过后续代的模型时继续趋向收敛，每个后续的代通过序列-活性模型的项的不同的一组系数值表征。

[0466] 计算一代中的每个模型对于训练组的多肽(具有已知的序列和相关活性)的适应性。参见图2的模块133、135、141和143。在一些实施方案中，适应性通过均方误差来测量。在其他实施方案中，适应性通过似然性来测量。在另外的实施方案中，适应性通过AIC或BIC来测量。

[0467] 评价特定代中的项的每种组合的预测力(即，决定143被否定回答)之后，检查遗传算法的收敛性或其他标准(诸如，固定的代数)以决定该过程是否应继续用于另外的代。参见模块145。假定该遗传算法尚未满足终止的标准，则对当前代的模型排序。最好地预测活性的那些模型可被保留并用于下一代。参见模块147。例如，可采用10％的精选率(elitism rate)。换言之，前10％的模型(如使用拟合函数确定并通过例如均方误差或AIC测量的)被保留以成为下一代的成员。下一代中剩余的90％的成员通过与前一代“亲本”交配来获得。参见模块149、151和153。

[0468] 如所指出的，“亲本”为选自前一代的模型。参见模块149。通常，虽然选择倾向于前

一代的更合适成员，但是在其选择中存在随机的组分。例如，亲本模型可使用线性加权(例如，比另一个模型表现好1.2倍的模型多20％的可能被选择)或几何加权(即，模型中的预测差异被提高至幂以获得选择的概率)来选择。在一些实施方案中，通过从前一代中的模型排名中简单地选择表现最佳的两个或更多个模型来选择亲本，而不选择其他模型。在这些实施方案中，对选自前一代的所有模型交配。在其他实施方案中，来自前一代的一些模型在不交配的情况下被选择纳入下一代中，并且来自前一代的表现较差的模型被随机地选择作为亲本。这些亲本彼此交配和/或与如此被选择待纳入下一代的表现较好的模型交配。

[0469] 已选择一组亲本模型之后，通过提供来自一个亲本的一些基因(系数值)和来自另一亲本的其他系数值，将成对的此类模型交配(模块151)以产生子代模型。在一个方法中，比对两个亲本的系数并接连地考虑每个值以确定所述子代应采用来自亲本A的项还是来自亲本B的项。在一个实施方案中，交配程序从亲本A开始并且随机地确定“交换(crossover)”事件是否应发生在所遇到的第一个项处。如果是，所述项取自亲本B。如果否，所述项取自亲本A。接连地考虑下一个项的交换，等等。这些项继续来自正在考虑的提供之前的项的亲本，直至交换事件发生。在那个点，下一个项由另一亲本提供，并且所有的后续项均由该亲本提供直至另一个交换事件发生。为了保证在子代模型中两个不同位置处不选择同一项，可采用多种技术，例如，部分交配交换技术。在一些实施方案中，子代染色体可采用基因的值的平均值，而不使用来自亲本之一的基因的系数值。

[0470] 在本发明的一些实施方案中，遗传算法还采用一个或更多个突变机制以生成模型的进一步的多样性(模块152)，这有助于探索未被亲本中的任何现存基因覆盖的参数空间的区域。另一方面，突变机制影响收敛性，以使得突变率越高或突变范围越大，收敛(如果有的话)所用的时间越长。在一些实施方案中，突变通过随机选择染色体/模型和随机选择所述染色体的系数而实施，然后所述系数被随机改变。在一些实施方案中，随机改变的系数值从具有特定范围的随机均匀分布抽取。在其他实施方案中，随机改变的系数值从具有特定范围的随机正态分布抽取。

[0471] 已考虑每个交互作用项之后，子代“模型”被确定为用于下一代。然后，选择另两个亲本以产生另一个子代模型，等等。最后，在以这种方式(模块153)选择了完成的代之后，下一代准备进化，并且然后程序控制返回至模块133，其中下一代的成员按照以上描述被评价。

[0472] 该过程一代又一代地继续，直到满足终止标准(即，决定模块145被肯定回答。在那个点，当前的代中选择至少一个排名靠前的模型作为整体最佳模型。参见模块155。收敛性可通过很多常规技术测试。通常地，其包括确定，来自很多后续的代的最佳模型的表现没有明显地改变。决定标准或收敛性的实例包括但不限于到目前为止生成的代的数目、来自当前文库的最佳蛋白的活性、期望的活性的量级和在最后一代模型中观察到的改进的水平。

[0473] 接下来展示了一个实例以展示遗传算法通过调整模型的系数的值来精化模型的用途。对于每个模型，选择针对每个可能的突变(即，针对模型中的每个项)的系数。种子群中的系数值在某些最大和最小边界内被随机选择。然后每个模型被用于预测来自变体的训练组中的每个变异序列活性。对于给定的模型，将计算的所有活性值与活性的观察值比较并生成均方误差的残差。针对第一代中随机生成的模型的每一个进行该程序。具有最小剩余误差的那些模型被选择为纳入下一代。

[0474] 表V.针对表I的序列的遗传算法的染色体

[0475]

	C_{10Asp}	C_{10Lys}	C_{166Ser}	C_{166Leu}	C_{166Ile}	C_{175Val}	C_{340Ala}
Ch_1	1.15	0.21	-0.1	1.1	2.3	-0.001	1.0
Ch_2	1.2	-0.8	-0.5	0.1	1.9	0.1	1.8
Ch_3	0.1	1.1	0.05	0.65	3	1.2	1.4

[0476]

…							
Ch_{n-1}	0.17	2.01	2	-0.1	0.2	0.76	0.37
Ch_n	3.1	1.6	0.55	1.2	2.6	2.0	-0.3

[0477] 对于该实例,遗传算法使用模型方程式评价每个染色体的适应性以计算单个染色体的预期活性。

[0478] $y_n(1+C_{10Asp}X_{10Asp})*((1+C_{10Lys}X_{10Lys})*(1+C_{166Ser}X_{166Ser})$

[0479] $*(1+C_{166Leu}X_{166Leu}))*(1+C_{166Ile}X_{166Il}e)$

[0480] $*(1+C_{175Val}X_{175Val})*(1+C_{340Ala}X_{340Ala})$

[0481] 对于染色体/模型2,预期活性y＝(1+1.15)*(1-0.001)*(1+1.0)。将该模型的该预期活性与观察到的序列活性比较,并由均方误差(MSE)计算个体的适应性。

[0482] 然后,遗传算法选择具有当前代的最低MSE的前20％的模型。通常,下一代涉及对从前一代选择的模型交配。交配可简单地包括从一个“亲本”模型选择一些系数并从不同的“亲本”选择剩余的系数。可按照“交换”或其他遗传操作进行项的选择。

[0483] 然后下一代模型(通过匹配产生的)各自被用于预测训练组的变体的活性。将所预测的活性与观察到的活性值和所产生剩余误差比较。具有最佳活性的第二代模型被选择用于另外的代的交配和选择。该程序继续进行直至模型表现收敛。例如,遗传算法运行50至100代。在那个点,从当前代中选择至少一个排名靠前的模型作为整体最佳模型。最合适的模型确定序列-活性模型的系数值。

[0484] 逐步选择

[0485] 在一些实施方案中,递加法或递减法被用来制作具有交互作用项的模型。通过实施图1的模块107中显示的操作,具有高预测力、包含交互作用项的最终模型通过从基础模型逐步添加或减去交互作用项来提供。图4A提供了通过将交互作用项加至基础模型并评价新的模型来产生最终的最佳模型来实施图1的模块107的操作的流程图。

[0486] 在该实例中,基础序列模型不包括交互作用项。所述方法首先将当前序列模型和最佳序列模型设定为基础序列模型,模块409。所述方法定义了序列变体的交互作用项的集。这些交互作用项可包括任何数目的两个或更多个氨基酸残基的配对或更高级别的相互作用。参见模块411。尽管模块409被例证为在模块411之前发生,但是两个步骤的次序并不

重要。在一些实施方案中，交互作用项的集包括所有感兴趣的氨基酸残基的析因组合(factorial combination)。在一些另外的实施方案中，至少所有成对的交互作用项被包括。在一些另外的实施方案中，成对的和三向交互作用项被包括。

[0487] 创建基础模型之后，所述方法从所述集中选择尚未经测定的交互作用项。然后所述方法通过将所选择的交互作用项加至当前序列模型来形成新的序列模型。参见模块413。然后所述方法利用对包括另外的交互作用项偏倚的模型选择方法来评价新的序列模型的预测力。参见模块415。所述方法确定新的序列模型的预测力是否比最佳序列模型的预测力大。参见模块417。例如，所述方法可使用采用“似然性”测定(例如，AIC)作为模型选择准则的技术。在此类情况中，只有具有比之前测定的模型小的AIC值的模型被认为具有较高的预测力。

[0488] 在一些实施方案中，所述选择方法倾向于具有更多参数的模型。此类选择方法的实例包括但不限于赤池信息量准则(AIC)和贝叶斯信息准则(BIC)，及其变化形式。例如，AIC可如下计算：

[0489] $AIC = -2\log_e L + 2k$

[0490] 其中L为给出数据组的模型的似然性，并且k为模型中自由参数的数目。

[0491] 在一些实施方案中，给出数据组的模型的似然性可通过多种方法计算，包括但不限于最大似然性法。例如，对于二进制因变量(其中对于一个观察值，活性可以是存在的或不存在的)，模型的似然性可如下计算：

[0492] $$L(model|data) = \prod_{i=1}^{n} \frac{(a_i + b_i)!}{a_i!\, b_i!} p_i^{a_i} (1 - p_i)^{b_i}$$

[0493] 其中n为数据组中数据点的总数；a_i和b_i为包含第i种条件的观察到的试验的数目；p为所述模型预测到正在观察的因变量的概率。

[0494] 在包括一系列嵌套模型的一些实施方案中，如在具有比基础模型渐进地增多的交互作用项(和相关系数)的回归模型中，较复杂的模型提供比较简单的模型同样好或更好的拟合，即使额外的系数是假的，因为较复杂的模型享有额外的自由度。在一些实施方案中，AIC在拟合优度的增益超出被假参数的消耗所抵消的程度上对较复杂的模型罚分。在模型选择中，较小的AIC值指示了较好的模型。

[0495] 在图4A中显示的实例中，如果新的序列模型的预测力大于最佳序列模型的预测力，那么所述方法将新的模型设定为最佳模型。参见模块419。然后，所述方法检查尚未测试的集中是否遗留有任何另外的交互作用项。参见模块421。如果是，程序返回至模块413，从而形成内循环以测试相互作用集中所有可用的交互作用项。通过迭代的内循环，可找到一个最佳交互作用项并将其添加至模型。

[0496] 在所有的交互作用项已被测试并且内循环结束之后，假定存在具有比之前的最佳模型大的预测力的模型，找出具有一个另外的交互作用项的最佳模型。参见模块423。在此类实施方案中，所述方法将当前模型设定为最佳模型，并将该最佳模型的交互作用项从交互作用项的可用集中排除。参见模块425。然后，所述方法回环至模块413。该外循环搜索可改进模型的预测力的下一个最佳交互作用项。如果发现这样的交互作用项，则在外循环中继续下一个最佳交互作用项的搜寻，直到没有具有比之前的最佳序列模型的预测力大的预测力的新的模型被鉴定到。

[0497] 当不在能找到改进所述模型的更多交互作用项时，所述方法将最佳模型设定为最终模型。参见模块427。提供序列和活性数据的最佳模型的搜索结束。然后该模型被用来预测新序列的活性。此类预测可指导用于进一步的变异和测试的序列的选择。

[0498] 在某些实施方案中，交互作用项的集中每个可用的交互作用项被视为对模型的质量或预测力具有可能相等的影响。换言之，在实施时，所述集中的每个可用交互作用项平等地可能被选择为在特定的迭代期间考虑。在一些实施方案中，可用的交互作用项可随机地或以一些任意的次序被选择。在一些其他实施方案中，交互作用项以这样的方式被偏倚或加权：在给定的迭代期间一些项比其他项更可能被选择为被考虑。在某些实施方案中，偏倚或加权可基于关于相互作用的物理或理论信息被应用。例如，可能已知的是在蛋白的两个特定区域中的突变可能自然地彼此靠近并且从而相互作用。与这两个常规区域中的残基有关的交互作用项可能是精化模型的迭代过程期间偏爱的。

[0499] 与图4A的虚拟码相似的例证过程的虚拟码如下：

[0500]
```
SET Coeff = Interaction Terms to Test
Best = Baseline Model
count = 1
WHILE count > 0
        count = 0
        BestFromRound = Best
         BestCoefficient = NULL
        FOR each Interaction Term in Coeff
                TestModel = (best + Interaction Term)[1]
                IF TestModel BETTER THAN BestFromRound  THEN[2]
                        BestFromRound = TestModel
                        Count++
                        BestCoefficient = Interaction Term
                ENDIF
        ENDFOR
        IF count > 0 THEN
                Best = BestFromRound
                Remove BestCoefficient FROM Coeff[3]
        ENDIF
ENDWHILE
```

[0501] 项1将测试交互作用项加至回归模型

[0502] 项2代表赤池信息准则(AIC)、贝叶斯信息准则(BIC)、交叉验证(平均误差)、ANOVA或系数贡献(coefficient contribution)中的一个或更多个的模型比较。

[0503] 项3被提供为避免重复的交互作用项测试。

[0504] 图4B提供了显示图1的模块107中显示的操作的实施方案的流程图。在该程序中，交互作用项被从基础模型中减去以形成最终的最佳模型，所述基础模型包括来自这些项的集中的所有可能的交互作用项。

[0505] 在该实施方案中，基础序列模型包括所限定的集中的所有交互作用项。所述方法首先将当前序列模型和最佳序列模型设定为与程序开始时的基础序列模型相同，模块439。该实施方案与以上描述的最后一个模型相似，因为交互作用项的整个集可包括两个或更多个氨基酸残基的任何数目的配对或更高级别的相互作用。在一些实施方案中，交互作用项的集包括所有感兴趣的氨基酸残基的因子组合。

[0506] 创建基础模型之后，所述方法从已被包括在基础模型中的项的集中选择尚未经测

定的交互作用项。然后所述方法通过将所选择的相互作用从当前序列模型减去来形成新的序列模型。参见模块441。然后所述方法利用对另外的交互作用项偏倚的模型选择方法来评价新的序列模型的预测力。参见模块443。所述方法评价新的序列模型的预测力是否比最佳序列模型的预测力大。参见模块445中显示的决策操作。在一些实施方案中,AIC被用作模型选择准则,以使得具有比之前测定的模型小的AIC值的模型被认为具有较高的预测力。

[0507] 在该示例性实例中,如果新的序列模型的预测力大于最佳序列模型的预测力,那么所述方法将该新的模型设定为最佳模型。参见模块447。接下来,所述方法检查所述集中是否遗留有尚未测试的任何另外的交互作用项(即,从当前序列模型减去)。参见模块449。如果存在任何未测试的项,所述方法返回至模块441,从而形成内循环以测试相互作用集中所有可用的交互作用项。通过迭代的内循环,一个交互作用项被鉴定。将其从模型除去在最大程度上改进了模型(并且如果AIC被用来测量模型的预测力,则通过最大限度地降低AIC)。

[0508] 在所有的交互作用项已被测试并且内循环结束之后,假定存在具有比之前的最佳模型大的预测力的模型,找出具有少一个交互作用项的最佳模型。参见模块451。在该情形中,所述方法将当前模型设定为最佳模型。参见模块453。然后,所述方法回环至模块441。该外循环搜索可以最大限度地提高模型的预测力的下一个交互作用项。如果发现这样的交互作用项,则在外循环中继续搜寻下一个待减去的交互作用项,直到没有具有比之前的最佳序列模型的预测力大的预测力的新模型被鉴定到。

[0509] 当内循环完成并且不能发现待减去的更多交互作用项以改进模型(即,模块451中显示的决策操作被否定回答)时,所述方法将最后的最佳模型设定为最终模型。参见模块455。对提供序列和活性数据的最佳模型的搜索结束。

[0510] E.可选的建模选项

[0511] 关于上述方法的多种另外的变化形式在本公开内容的范围内。事实上,不期望本发明被限制于任何特定的模型,因为任何合适的模型在本发明中具备实用性。作为一个示例性实例,xij变量为氨基酸的物理或化学特性的表现-而不是氨基酸自身(亮氨酸对缬氨酸对脯氨酸...)的确切身份。此类特性的实例包括亲油性、体积和电子特性(例如,形成电荷、与部分电荷有关的范德瓦耳斯表面积等)。为了实施该方法,代表氨基酸残基的xij值可以其特性或由这些特性构建的主成分的形式呈现。不期望本发明被限制于氨基酸、肽和/或多肽的任何特定的特性,因为任何合适的特性在本发明的方法中具备实用性。

[0512] 在一些另外的实施方案中,xij变量代表核苷酸,而不是氨基酸残基。在这些实施方案中,目标是为了鉴定编码用于蛋白变体文库的蛋白的核苷酸序列。通过使用核苷酸而不是氨基酸,可如所期望的优化除了活性之外的参数(例如,比活性)。例如,在特定宿主或载体中的蛋白表达可以是核苷酸序列的函数。两种不同的核苷酸序列可编码具有相同氨基酸序列的蛋白,但是一种核苷酸序列可导致更大量的蛋白的产生和/或所述蛋白活性更大。通过使用核苷酸序列而不是氨基酸序列,本文描述的方法可被用来优化表现出改进的基因表达特性和/或改进的特性(例如,比活性、稳定性等)的微生物菌株。

[0513] 在一些实施方案中,核苷酸序列被表示为密码子的序列。在一些实施方案中,模型将密码子用作核苷酸序列的基本单位,以使得所预测的活性是存在于核苷酸序列中的多种密码子的函数。每个密码子连同其在整个核苷酸序列中的位置充当用于生成序列-活性模

型的自变量。应注意在一些情形中，对于给定的氨基酸的不同密码子在给定的生物体中表达不同。在一些实施方案中，对于给定的氨基酸，每个生物体具有偏好的密码子或密码子的频率分布。通过将密码子用作自变量，该实施方案解释了这些偏好。因此，该实施方案可被用来生成表达变体的文库(例如，其中“活性”包括特定宿主生物体的基因表达水平)。

[0514] 在一些实施方案中，所述方法包括以下操作：(a)接收表征蛋白变体文库的训练组的数据；(b)基于在(a)中获得的数据开发这样的交互作用序列-活性模型：将活性预测为核苷酸类型和核苷酸序列中的相应位置的函数。(c)利用序列-活性模型对核苷酸序列中的位置和/或核苷酸序列中特定位置处的核苷酸类型按照对期望的活性的影响的次序排序；以及(d)利用该排序来鉴定核苷酸序列中待被改变或固定的一个或更多个核苷酸，以提高期望的活性。如所述的，在一些实施方案中，待改变的核苷酸编码特定的氨基酸。

[0515] 在一些其他的实施方案中，所述方法包括利用不同的技术用于排序或以其他方式表征残基与特定特性相关的重要性。如以上针对线性或非交互作用模型描述的，回归系数的量级被用来对残基排序。具有大的量级的系数的残基(例如，166Ile)被视为排序高的残基。该表征被用来决定是否在生成新的、改进的蛋白变体文库时改变特定的残基。如本文描述的，对于交互作用模型，敏感性分析更复杂。

[0516] PLS和其他技术提供除了回归系数量级以外的另外的信息，所述另外的信息可被用来对特定残基或残基位置排序。诸如PLS和主成分分析(PCA)或PCR的技术以主成分或潜在向量的形式提供信息。这些技术通过多维数据组，诸如本文公开的本发明的实施方案采用的蛋白序列-活性空间，表示最大变异的方向或向量。这些潜在向量为多个序列维度，即，构成蛋白序列的各个残基或残基位置的函数，所述蛋白序列构成被用来构建训练组的变体文库。因此潜在向量包括来自训练组中的每个残基位置的贡献的和。一些位置对向量的方向贡献更强。这些贡献通过相对大的“载荷(load)”，即，用来描述向量的系数来表现。作为一个简单的示例性实例，训练组可包括三肽。在该实例中，第一潜在向量包括来自所有三个残基的贡献。

[0517] 向量1＝a1(残基位置1)+a2(残基位置2)+a3(残基位置3)

[0518] 系数a1、a2、和a3为载荷。因为这些载荷反应了相应的残基位置对数据组中的变异的重要性，所以它们可被用来按照如上所述的对各个残基位置对于“切换”决策的目的的重要性排序。载荷像回归系数一样可被用来对每个切换位置处的残基排序。多个参数描述这些载荷的重要性。一些实施方案使用诸如变量重要性投影(VIP)的方法来利用载荷矩阵。该载荷矩阵包括取自训练组的多个潜在变量的载荷。在PLS变量重要性投影(Variable Importance for PLS Projection)方法中，变量(例如，残基位置)的重要性通过计算VIP来计算。对于给定的PLS维度a，$(VIN)_{ak}^2$等于变量的PLS权重的平方$(w_{ak})^2$乘以该PLS维度的y(因变量，例如特定函数)解释的可变性百分比。对所有的PLS维度(成分)的$(VIN)_{ak}^2$求和。然后通过将所述和除以PLS模型解释的y的总变异性百分比，并乘以模型中变量的数目来计算VIP。具有大于1的VIP的变量是对于关联特定函数(y)最相关的，并且因此对于做转换决策的目的的排名最高。

[0519] 在很多实施方案中，本发明利用常规线性回归方法鉴定组合文库中的突变对感兴趣的序列-活性的影响。可选的建模选项和技术例如贝叶斯回归、集成回归、靴襻法(bootstrapping)可被用于联合或取代以上所述的方法。事实上，不期望本发明限制于任何

特定的建模选择和/或技术，因为任何合适的方法在本发明中具备实用性。

[0520] 贝叶斯线性回归

[0521] 在本发明的一些实施方案中，贝叶斯线性回归具备实用性。该方法是一种线性回归方法，其中在贝叶斯推理的背景中进行统计分析。当回归模型具有具备正态分布的误差时，并且如果假定特定形式的先验分布，则模型的参数的后验概率分布可利用贝叶斯推理技术来确定。

[0522] 线性回归模型的普通最小二乘解基于数据的似然性函数利用诸如Moore-Penrose伪逆(Moore-Penrose pseudo inverse)的分析计算方法来评价系数向量和模型误差。这是频概论方法，其假定存在足够的数据观察值来代表对于所有序列的序列-活性关系。但是，样本的实际观察值几乎从来都不足以代表群中的所有成员。当样本(或训练组)容量有限时，这特别成问题。在贝叶斯方法中，对样本数据补充先验概率分布形式的另外的信息。将关于参数的先验知识与根据贝叶斯定理的数据的似然性函数结合来获得关于参数的后验知识。先验知识可取决于领域和可用的先验信息而采用不同的函数形式。

[0523] 例如在一些实施方案中，贝叶斯回归可在模型拟合之前使用先验信息来对系数加权。在一些实施方案中，取自之前的一轮定向进化，例如使用亲本骨架或参考骨架和在之前的多轮中使用的至少一些突变进行的一轮定向进化的序列/活性数据可被用来对线性系数加权。另外，对两个或更多个突变之间的强性关系的预测可被用来对交互作用系数加权。该方法的主要优势之一是纳入先验信息来指导模型预测。

[0524] 先验信息的来源的一个示例性实例是具有针对参考骨架的多个突变中的每一个的独立项和交互作用项的模型。在一些实施方案中，数据获取自每个变体包含一个突变的变体的集合。

[0525] 在本发明中具备实用性的先验信息的另外的实例包括但不限于关于某些突变或突变类型的作用的直观或物理信息。不论来源，先验信息充当对序列和活性之间关系的预先设想的见解。

[0526] 在用于评估模型的参数的一些实施方案中，贝叶斯线性回归利用蒙特卡罗模拟诸如吉布斯抽样或Metropolis算法来基于数据拟合模型。当直接抽样困难时，吉布斯抽样是马尔可夫蒙特卡罗算法，以便获得由特定的多变量概率分布(即，由两个或更多个随机变量的联合概率分布)估算的一系列观察值。

[0527] 图5为例证贝叶斯回归在变体文库的引导进化中的使用的流程图。每轮序列进化包括基于来自之前的一轮的序列的突变，所述之前的一轮序列进化可通过对诸如序列-活性模型的知识引导。例如，如在模块501中，在第n轮进化中，每个变体有一个突变。下一轮或第n+1轮进化为当前的轮，如在模块503中显示的。对于每个变体有至少一个新的突变，总计每变体两个或更多个突变。在该示例性实例中，贝叶斯回归在该轮实施。

[0528] 第n+1轮的序列变体为新模型提供数据训练组。新模型可包括基础模型或完成的模型，所述基础模型只包括对于各个残基的非交互作用项，所述完成的模型包含所有的可能交互作用项/系数，如在模块507中所示。新模型还可包括通过多种技术选择的模型，包括以上解释的逐步添加或减去技术，参见模块505。可选地，可使用以下讨论的遗传算法或靴襻技术被选择模型。这些模型全部基于来自第n+1轮的训练组数据的当前/新数据。贝叶斯推理技术可应用于这些模型，以使得模型基于当前数据的概率函数和先验信息的分布两

者。先验信息可来自之前的一轮，如模块501所示的第n轮的序列变体的数据。信息还可来自之前的任何一轮进化的序列-活性数据或其他先验直观知识，如模块513所示。由模块509所示的贝叶斯回归模型基于由当前数据和先验信息提供的信息预测活性，参见模块511。尽管图5只例证了贝叶斯回归技术对第n+1轮的应用，但是贝叶斯回归技术可应用在多个阶段。也不期望本发明被限制于图5中提供的特定步骤，因为任何合适的方法在本发明中具备实用性。

[0529] 集成回归

[0530] 在一些实施方案中，本发明利用集成回归技术制作序列-活性模型。集成回归模型基于若干回归模型。每个模型的预测值基于特定信息准则(IC)被加权，并且集成预测值是其包含的所有模型的预测值的加权和。在一些实施方案中，模型开发从包含所有交互作用项的基础模型开始。随后的模型通过以一些或所有可能的组合添加交互作用系数来构建。在一些实施方案中，交互作用系数以逐步的程序被添加。每个模型拟合数据并生成IC。对每个模型的加权基于IC，所述IC可以是IC自身或转换形式，例如对数值、否定值(negated value)等。预测可通过以下步骤针对观察值进行：产生所述集中的每个模型的预测值，并通过采用来自每个模型的预测值的加权均值来确定集成预测值。完整的集包含所有可能的模型，但可被修剪以通过对其包含的模型数目或基于IC设定阈值来去除表现差的模型。

[0531] 所述集的成分模型可利用多种技术产生。例如，在一些实施方案中，使用遗传算法来形成成分模型。序列/活性数据被用来产生多个回归模型，每个回归模型具有其自己的系数组。根据适应性准则(例如，AIC或BIC)选择最佳模型。这些模型被“配对”以产生新的混合模型，然后评价所述混合模型的适应性并相应地做出选择。在一些实施方案中，该程序被重复多轮“计算进化”以产生最佳模型的集。可选地，在一些实施方案中，集的组分通过以上描述的逐步回归形成，并选择最佳的n个模型以形成所述集。

[0532] 图6提供了根据本发明的实施方案的在序列变体的定向进化中实施集成回归的过程的流程图。在该实施方案中，集成回归技术可被应用在多轮序列进化轮的任何阶段。例如，在第n轮，模块601中显示的序列变体提供了用于多个模型的训练组数据以形成如模块603所示的模型集。模型集中的模型可以是通过遗传算法和/或逐步选择法生成的模型。在其他实施方案中，模型集包括n折交叉验证模型和/或靴襻模型。在一些实施方案中，只有基于多种模型选择准则诸如AIC或BIC具有较好预测力的模型被选择进入所述集。

[0533] 可选地或另外地，在一些实施方案中，尚未通过模型选择筛选的模型也进入模型集。在一个实施方案中，具有所有非交互作用或交互作用项的所有模型均进入模型集。对于大量的残基以及更大量的残基间的析因相互作用，该实施方案可能是计算十分密集的。在一些可选的实施方案中，只有包含非交互作用项和成对的交互作用项的模型进入所述模型集。不论模型集的纳入方法如何，集成模型包括其成分的所有项。模型集可包含任何数目的模型，包括但不限于贝叶斯模型，就该模型来讲，先验信息可被并入所述集。

[0534] 在一些实施方案中，所述集基于所述集中的每个模型的系数的加权平均值预测序列活性，其中如模块605所示的，权重通过相应的模型的预测力确定。

[0535] 在一些实施方案中，集成回归利用以下工作流程：(1)提供空的集；(2)选择为1或更大的组容量n；(3)将数据点分类为具有容积为为n的组，其中数据点被无置换地分组；和(4)制作预测各个系数和交互作用系数的集成模型。在一些实施方案中，制作集成模型的步

骤(4)还包括:a)移除每个组的数据点,其中剩余的数据形成训练组并且删除的数据形成验证组;b)通过利用逐步回归拟合训练组来制作模型;c)使用验证组测试该模型,这提供了对模型预测能力的指示;d)将该模型添加至用来生成如以上所述的集成模型的模型集。

[0536] 靴襻法

[0537] 用于表征正在考虑的模型在给定的迭代中的预测力的其他技术在本发明中具备实用性。在一些实施方案中,这些技术包括交叉验证技术或靴襻技术。在一些实施方案中,交叉验证采用用来生成模型的一组观察值而删除评价模型的强度的一些观察值。在一些实施方案中,靴襻技术包括利用被置换地测定的一组样本。在一些实施方案中,通过交叉验证或靴襻法生成的模型可被组合到如上描述的集成模型中。

[0538] 在一些另外的实施方案中,这些方法不仅通过残基对活性的预测的贡献的量级来对残基排序,而且还通过那些预测的贡献的置信度对残基排序。在一些情形中,研究者关注模型从一组数据到另一组数据的概括能力。换言之,研究者想知道系数或主成分的值是否是假的。交叉验证和靴襻技术提供了指示模型概括多个数据能力的置信水平的测量。

[0539] 在一些实施方案中,利用统计上更严格的方法,在所述方法中排序基于量级和分布的组合。在这些实施方案的一些中,具有高量级和紧密分布两者的系数提供最高排名。在一些情形中,具有比另一个低的量级的一个系数由于具有较少的变异可被给予较高的排名。因此,一些实施方案基于量级和标准偏差或方差两者对氨基酸残基或核苷酸排序。多种技术可被用来完成排序。事实上,不期望本发明被限制于任何特定的用于排序的技术。利用靴襻p值方法的一个实施方案在以下被描述。

[0540] 采用靴襻法的方法的示例性实例在图7中被描绘。如图7中显示的,方法725开始于模块727,在模块727中提供了初始数据组S。在一些实施方案中,如以上描述的,这是训练组。例如在一些实施方案中,其通过以任何方式(例如,如以上描述的)系统地改变起始序列的各个残基来生成。在由方法725示例的情况中,数据组S具有M个不同的数据点(从氨基酸或核苷酸序列收集的活性和序列信息)用于在分析中使用。

[0541] 由数据组S创建多个靴襻组B。这些靴襻组中的每个通过以置换的方式从组S抽样以形成具有M个成员-全部取自初始组S的新组来获得。参见模块729。“以置换的方式”的条件产生对初始组S的改变。有时候,新的靴襻组B将包含来自S的重复样本。在一些情况中,靴襻组B也不包含原始被包含于S的某些样本。

[0542] 作为示例性实例,提供了具有100条序列的组S。靴襻组B通过从初始组S中的100条序列中随机选择100个成员序列来创建。被用于所述方法的每个靴襻组B包括100条序列。因此,一些序列将被多于一次地被选择并且其他序列将根本不被选择是可能的。所述方法接下来利用从来自组S的100条序列产生的靴襻组B建立模型。参见模块731。如以上描述的,所述模型可利用PLS、PCR、SVM、逐步回归等建立。事实上,期望任何合适的方法将在建立该模型中具备实用性。该模型提供对来自组B的多个样本中出现的残基或核苷酸的排序的系数或其它指标。如在模块733显示的,这些系数或其他指标被记录以便随后使用。

[0543] 接下来,在决策模块735,所述方法确定是否应创建另一个靴襻组。如果是,所述方法返回至模块729,在模块729中如以上描述的创建新的靴襻组B。如果否,所述方法进行至以下讨论的模块737。模块735处的决策转向关于在评价那些值的贡献时要使用多少不同组的系数值。组B的数目应足以产生精确统计。在一些实施方案中,制作并分析100至1000个靴

襻组。这通过约100至1000次通过方法725的模块729、731和733体现。但是，不期望本发明被限制于任何特定数目的靴襻组，因为适于期望的分析的任何数目具备实用性。

[0544] 在制作并分析足够数目的靴襻组B之后，决定735被回答否定。如所示，所述方法然后进行至模块737。此处，对于每个残基或核苷酸(包括密码子)使用系数值(例如，100至1000个值，每个靴襻组一个)计算系数(或由模型产生的其他指标)的平均值或标准偏差。所述方法可从该信息计算t统计值并确定测量值不同于0的置信区间。所述方法从t统计值计算针对置信区间的p值。在该示例性情形中，p值越小，所测量的不同于0的回归系数越可信。

[0545] 要注意的是，p值不过是可对残基重要性的系数或其他指标的统计学变化做出解释的很多不同类型的表征方式中的一种。实例包括但不限于计算回归系数的95％置信区间并排除考虑95％置信区间穿过零线的任何回归系数。基本上，在一些实施方案中，对数据贡献的标准偏差、方差或其他统计上的相关测量做出解释的任何表征方式具备实用性。在一些实施方案中，该表征步骤还对系数的量级做出解释。

[0546] 在一些实施方案中，得到大的标准偏差。该大的标准偏差可归因于多种原因，包括但不限于数据组中差的测量值，和/或初始数据组中的特定残基或核苷酸的有限的代表性。在后一种情况中，一些靴襻组将不包括特定残基或核苷酸的存在。在此类情况中，该残基的系数的值将为0。其他的靴襻组将包括所述残基或核苷酸的至少部分存在并给予非0值的相应系数。但是给予0值的组将导致系数的标准偏差变得相对大。这降低了系数值的置信度并导致较低的排名。但是如果关于所包括的残基或核苷酸存在相对少的数据，这则是意料之中的。

[0547] 接下来，在模块739处，所述方法将回归系数(或其他指标)从最低的(最佳的)p值到最高的(最差的)p值排序。该排名与回归系数自身的绝对值高度相关，因为事实上绝对值越大，越大的标准偏差远离0。因此，对于给定的标准偏差，随着回归系数变大，p值变小。但是，绝对排名将不会总是与p值法和纯粹的量级法两者相同，特别是当在组S中可用来开始的数据点相对少时。

[0548] 最后，如模块741显示的，所述方法基于在模块739的操作中观察到的排名固定和切换某些残基。这本质上是对以上对于其他实施方案描述的排序方法的相同应用。在一个方法中，所述方法固定最佳残基(目前具有最低的p值的那些残基)而切换其他残基(具有最高p值的那些残基)。

[0549] 已显示该方法725在计算机模拟中表现良好。而且，在一些实施方案中，p值排名方法天然地处理单个或极少的偶现残基(instance residue)：p值通常将较高(较差)，因为在靴襻程序中初始数据组中不常出现的那些残基将不太可能被随机选取。即使其系数大，其可变性(以标准偏差衡量)也将是相当高的。在一些实施方案中，这是期望的结果，因为未被很好地代表(即，未以足够的频率被观察到或具有较低的回归系数)的那些残基在下一轮文库设计中可能是用于切换的好的候选物。

[0550] VI.通过修饰模型预测的序列生成优化的蛋白变体文库

[0551] 本发明的目标之一是，通过定向进化产生成优化的蛋白变体文库。本发明的一些实施方案提供了利用所产生的序列-活性模型指导蛋白变体的定向进化的方法。根据以上描述的方法制作并精化的多种序列-活性模型适于指导蛋白或生物分子的定向进化。作为该程序的一部分，所述方法可鉴定待被用于产生新的蛋白变体文库的序列。此类序列包括

对以上鉴定的特定残基的变异，或者是用来随后引入此类变异的前体。可通过进行诱变或基于重组的多样性生成机制来修饰这些序列，以生成新的蛋白变体文库。该新的文库还可被用于开发新的序列活性模型。

[0552] 在一些实施方案中，寡核苷酸或核酸序列的制备通过利用核酸合成仪合成所述寡核苷酸或核酸序列来完成。本发明的一些实施方案包括利用所制备的寡核苷酸或蛋白序列作为用于定向进化的组成模块(building block)进行一轮定向进化。本发明的多个实施方案可将重组和/或诱变应用于这些组成部件以产生多样性。

[0553] 作为一个具体的实例，一些实施方案应用寡核苷酸重组技术。在这些实施方案中，所述方法包括通过评价序列-活性模型的项的系数来选择用于一轮定向进化的一个或更多个突变。突变从特定位置处的特定氨基酸或特定类型的核苷酸的组合中基于通过所述模型预测的它们对蛋白活性的贡献选择。在一些实施方案中，突变的选择包括鉴定被确定为比其他系数大的一个或更多个系数，以及选择由如此鉴定的一个或更多个系数表示的特定位置处的特定氨基酸或核苷酸。在一些实施方案中，在根据序列-活性模型选择突变之后，所述方法包括制备含有或编码所述一个或更多个突变的多个寡核苷酸，并利用所制备的寡核苷进行一轮定向进化。在一些实施方案中，定向进化技术包括组合和/或重组所述寡核苷酸。

[0554] 本发明的其他实施方案应用蛋白序列重组技术。在一些实施方案中，所述方法包括鉴定新的蛋白序列或新的核酸序列，以及制备并测定所述新的蛋白或由所述新的核酸序列编码的蛋白。在一些实施方案中，所述方法还包括使用所述新的蛋白或由所述新的核酸序列编码的蛋白作为用于进一步的定向进化的起始点。在一些实施方案中，定向进化过程包括片段化并重组由模型预测的具有期望的活性水平蛋白序列。

[0555] 在一些实施方案中，所述方法基于被所述模型预测为重要的各个突变来鉴定和/或制备新的蛋白或新的核酸序列。这些方法包括：通过评价序列-活性模型的项的系数选择一个或更多个突变，以鉴定对活性有贡献的特定位置处的一个或更多个特定的氨基酸或核苷酸；鉴定包含以上选择的一个或更多个突变的新蛋白序列或新核酸序列，并制备和测定所述新蛋白或由所述新核酸序列编码的蛋白。

[0556] 在其他的实施方案中，所述方法基于整个序列(而不是各个突变)的预测活性来鉴定和/或制备新蛋白或新核酸序列。在这些实施方案的一些中，所述方法包括将多个蛋白序列或多个氨基酸序列应用于序列-活性模型，以及确定所述序列-活性模型预测的对于所述多个蛋白序列或核酸序列中的每一个的活性值。所述方法还包括通过评价由所述序列-活性模型针对多个序列预测的活性值，从以上应用的多个蛋白序列或多个氨基酸序列中选择新蛋白序列或新核酸序列。所述方法还包括制备并测定具有所述新蛋白序列的蛋白或由所述新核酸序列编码的蛋白。

[0557] 在一些实施方案中，不是简单地合成单个最佳预测蛋白，而是基于对蛋白中每个位置处的残基选择的最佳改变的敏感性分析产生组合的蛋白文库。在该实施方案中，对于所预测的蛋白给定的残基选择越敏感，预测到的适应性改变将更大。在一些实施方案中，这些敏感性从最高至最低并且敏感性得分被用于在随后的轮次中创建组合蛋白文库(即，通过基于敏感性并入那些残基)。在使用非交互作用模型的一些实施方案中，通过简单地考虑与该模型中给定的残基项相关的系数的大小来确定敏感性。但是，对于交互作用模型这是

不可能的。相反，在利用交互作用模型的实施方案中，残基敏感性通过使用模型计算当“最佳”预测的序列中单个残基被改变时的活性改变来确定。

[0558] 本发明的一些实施方案包括选择蛋白序列或核酸序列中的一个或更多个位置，以及在如此鉴定的一个或更多个位置处进行饱和诱变。在一些实施方案中，这些位置通过评价序列-活性模型的项的系数来选择，以鉴定对活性有贡献的特定位置处的一个或更多个特定的氨基酸或核苷酸。相应地，在一些实施方案中，一轮定向进化包括在使用所述序列-活性模型选择的位置处对蛋白序列进行饱和诱变。在一些涉及包括一个或更多个交互作用项的模型的实施方案中，所述方法包括在两个或更多个相互作用残基处同时应用诱变。

[0559] 在一些实施方案中，残基以其排列顺序被考虑。在一些实施方案中，对于所考虑的每个残基，所述程序确定是否“切换”那个残基。术语“切换(toggling)”指，将多种氨基酸残基类型引入优化的文库中的蛋白变体的序列中的特定位置。例如，丝氨酸可出现在一个蛋白变体的位置166，然而苯丙氨酸可出现在相同文库的另一个蛋白变体的位置166。在训练组中的蛋白变体序列之间不发生变化的氨基酸残基在优化的文库中通常保持固定。但是，情况并非总是这样，因为在优化的文库中可能存在变异。

[0560] 在一些实施方案中，优化的蛋白变体文库被设计成使得所有被鉴定的回归系数排序“高”的残基被固定，而剩余的回归系数排序较低的残基被切换。该实施方案的基本原理是，‘最佳’预测的蛋白周围的局部空间应被搜索。应注意，其中切换被引入的起点“骨架”可能是模型预测的最佳蛋白和/或已经被验证为被筛选的文库中的‘最佳’蛋白。事实上，不期望起点骨架被限制于任何特定蛋白。

[0561] 在可选的实施方案中，至少一个或更多个(但并非全部的被鉴定的回归系数排序高的残基)在优选的文库中保持固定，而其他的残基被切换。在一些实施方案中，如果不期望通过一次性并入太多改变而显著地改变其他氨基酸残基的背景，则推荐该方法。再次地，用于切换的起点可以是所述模型预测的最佳的残基组、来自现有文库的最佳验证的蛋白或模拟得很好的“平均”克隆。在后一种情况中，切换被预测为较高重要性的残基可能是期望的，因为在对之前被从抽样中遗漏的活性峰的搜索中应探索更大的空间。该类型的文库通常在早期的多轮文库制备中更为重要，因为其产生了对随后的轮次的更精确的描绘(picture)。不期望起始点骨架被限于任何特定蛋白。

[0562] 以上实施方案的一些备选方案涉及在确定切换哪些残基时使用残基重要性(排名)的不同程序。在一个这样的备选实施方案中，排名较高的残基位置是切换更强烈地偏爱的。该方法中所需的信息包括来自训练组的最佳蛋白的序列、PLS或PCR预测的最佳基序列和来自PLS或PCR模型的残基排名。“最佳”蛋白为数据组中经湿实验室验证的“最佳”克隆(即，具有最高的测量功能但是由于其相对接近交叉验证的预测值仍模拟得很好的克隆)。所述方法将来自该蛋白的每个残基与来自具有期望活性的最高值的“最佳预测”序列的相应残基比较。如果具有最高载荷或回归系数的残基不存在于‘最佳’克隆中，则所述方法将引入该位置作为切换位置用于随后的文库。如果所述残基存在于最佳克隆中，则所述方法不将该位置视作切换位置，并且将按顺序移至下一个位置。对多个残基重复该程序，连续地穿过较低的载荷值，直到生成具有足够容量的文库。

[0563] 在一些实施方案中，待保留的残基的回归系数的数值和待切换的残基的回归系数的数值是变化的。决定待切换的哪些残基以及待保留的哪些残基基于多个因素，包括但不

限于期望的文库容量、回归系数之间差异的量级和认为将存在的非线性度的程度。保留具有小的(中性)系数的残基可能在随后的进化轮次中揭露重要的非线性度。在一些实施方案中,优化的蛋白变体文库包括约2^N个蛋白变体,其中N代表在两个残基之间切换的位置的数目。换言之,通过每个另外的切换增加的多样性使文库的容量翻倍,使得10个切换位置产生～1,000个克隆(1,024),13个位置～10,000个克隆(8,192)而20个位置～1,000,000个克隆(1,048,576)。适当的文库容量取决于诸如筛选成本、格局(landscape)的强度、优选的空间抽样百分比等的因素。在一些情形中,已发现相对大数目的改变的残基产生非常大百分比的无功能的克隆的文库。因此,在一些实施方案中,用于切换的残基的数目从约2至约30变化;即,文库容量在约4和2^{30}～10^9个克隆之间变化。

[0564] 另外,设想到多个随后轮次的文库策略被同时利用,一些策略更有挑战性(固定更多的"有利"残基)而其他策略更为保守(为了更彻底地探索空间的目的,固定较少的"有利"残基)。

[0565] 在一些实施方案中,在大多数天然存在或以其他形式成功的肽中存在的基团或残基或"基序"被鉴定和/或保留,因为它们可能在蛋白的功能性(例如,活性、稳定性等)方面是重要的。例如,可能发现在天然存在的肽中,可变位置3处的Ile总是与可变位置11处的Val成对。因此,在一个实施方案中,在任何切换策略中需要保留此类组。换言之,仅有的被接受的切换是在基础蛋白中保留了特定组合(grouping)的那些切换或者生成不同的也被发现于有活性的蛋白的组合的那些切换。在后一种情况中,切换两个或更多个残基是必要的。

[0566] 在一些另外的实施方案中,湿实验室验证的当前优化的文库中的'最佳'(或几个最佳的之一)蛋白(即,具有测量功能最高的、或几个最高的之一的、仍很好地模拟,即,相对接近交叉验证中的预测值的蛋白)被用作其中多个改变被并入的骨架。在另一个方法中,湿实验室验证的当前最优文库中的'最佳'(或几个最佳的之一)的、可能模拟不好的蛋白被用作其中多个改变被并入的骨架。在一些其他的方法中,序列-活性模型预测为具有最高的期望活性值(或最高值之一)的序列被用作为骨架。在这些方法中,用于"下一代"文库(以及可能地相应的模型)的数据组通过改变一个或几个最佳蛋白中的残基来获得。在一个实施方案中,这些改变构成骨架中的残基的系统变异。在一些情形中,这些改变包括各种诱变、重组和/或亚序列选择技术。这些改变中的每个可体外、体内和/或提供计算机模拟进行。事实上,不期望本发明被限于任何特定的形式,因为任何合适的格式具备实用性。

[0567] 在一些实施方式中,虽然由非交互作用模型预测的最优序列可如以上描述的通过检验来鉴定,但是对于交互作用模型其实不然。在非交互作用项和交互作用项中均出现的某些残基及其在其他残基的很多可能的组合的背景下对活性的整体影响是有疑问的。因此,正如对于交互作用模型的交互作用项的选择,由交互作用模型预测的最优序列可通过用该模型测试所有可能的序列(假定计算资源足够)或通过利用搜索算法诸如逐步算法来鉴定。

[0568] 在一些实施方案中,在计算机中获得的信息-按照如上所述鉴定的进化的蛋白被用来合成新型蛋白并基于物理测定测试它们。精确的对实际的湿实验室确定的适应性功能的计算机模拟表示方式允许研究者减少进化循环的数目和/或实验室中需筛选的变体的数目。在一些实施方案中,优化的蛋白变异文库利用本文描述的重组方法或可选地通过基因

合成方法随后通过体内或体外表达来产生。在一些实施方案中，在优化的蛋白变体文库被筛选期望的活性之后，对其测序。如以上在对图1和2的讨论中指出的，来自优化的蛋白变体文库的活性和序列信息可被采用以生成另一个序列-活性模型，可利用本文描述的方法由所述另一个序列-活性模型设计进一步优化的文库。在一个实施方案中，所有来自该新文库的蛋白被用作数据组的一部分。

[0569] VII.数字化装置和系统

[0570] 明显的是，本文描述的实施方案采用在指令的控制下行动的程序和/或储存在一个或更多个计算机系统中或经其转换的数据。本文公开的实施方案还涉及用于进行这些操作的装置。在一些实施方案中，所述装置针对所需的目的而被专门设计和/或构建，或其可以是通过计算机程序和/或储存于计算机中的数据结构选择性地激活或重新配置的通用型计算机。本发明提供的程序并不固有地与任何特定的计算机或其他特定装置有关。具体地，多种通用型机器在根据本文的教导书写的程序中具备实用性。但是，在一些实施方案中，构建专门的装置以进行所需方法的操作。下文描述了用于各种各样的这些机器的特定结构的一个实施方案。

[0571] 另外，本发明的某些实施方案涉及包括用于进行多种计算机实施的操作的程序指令和/或数据(包括数据结构)的计算机可读介质或计算机程序产品。计算机可读介质的实例包括但不限于磁性介质，诸如硬盘、软盘、磁带；光学介质，诸如CD-ROM设备或全息设备；磁光介质；半导体存储器装置；和被特别地配置以储存并执行程序指令的硬件设备，诸如只读存储器装置(RPM)和随机存取存储器(RAM)、专用集成电路(ASIC)、和可编程的逻辑装置(PLD)。数据和程序指令还可体现在载波或其他传输介质(例如，光学线路、电子线路和/或空中电波)上。事实上，不期望本发明被限制于包含用于执行计算机实施的操作的指令和/或数据的任何特定的计算机可读介质或任何其他计算机程序产品。

[0572] 程序指令的实例包括但不限于诸如由编译器产生的低阶码和可由计算机利用解释器执行的包含较高阶代码的文件。另外，程序指令包括但不限于直接或间接地控制根据本发明的计算机的操作的机器代码、源代码和任何其他代码。代码可规定输入、输出、计算、条件式、分支、迭代循环等。

[0573] 在一个示例性实施方案中，本文公开的表现方法在含有逻辑指令和/或数据的固定的介质或可传输程序部件中体现，所述逻辑指令和/或数据当被加载到合适地配置的计算装置时导致所述装置对一个或更多个字符串进行模拟的遗传操作(GO)。图8显示了示例性数字化装置800，其是一种逻辑装置，能够从介质817、网络端口819、用户输入键盘809、用户输入811或其他输入装置读取指令。之后装置800能利用那些指令在数据空间中指导统计操作，例如构建一个或更多个数据组(例如，确定数据空间的多个代表性成员)。可体现所公开的实施方案的一种类型的逻辑装置是如包括CPU 807、光学用户键盘输入设备809和GUI定点设备811，以及周围部件诸如磁盘驱动器815和监视器805(其显示GO修饰的字符串并提供用户对此类字符串的子集的简化的选择的计算机系统800的计算机系统。固定介质817被任选地用来为整个系统提供程序并可包括例如，盘式光学或磁性介质(disk-type optical or magnetic media)或其他电子存储元件。通信端口819可被用来为系统提供程序并可代表任何类型的通信连接。

[0574] 在一些实施方案中，本公开内容提供了包括一个或更多个处理器、系统存储器和

计算机可执行指令存储于其上的一个或更多个计算机可读存储介质的计算机系统，当所述计算机可执行指令被一个或更多个处理器执行时，导致计算机系统执行用于进行生物分子的定向进化的方法。所述方法包括：(a)获得针对多个生物分子的序列和活性数据，每个生物分子包含具有多种类型的亚单位和序列位置的序列；(2)从所获得的数据建立序列-活性模型；以及(c)使用所述序列-活性模型鉴定影响生物分子的期望活性的用于变异的特定位置处的特定类型的一个或更多个亚单位。在一些实施方案中，序列-活性模型包括多个基本相乘项的乘积，基本相乘项的每一个包括(1)代表特定序列位置处的特定类型的特定亚单位的存在/不存在的哑变量，和(2)代表特定亚单位对活性的贡献的系数。

[0575] 某些实施方案还可体现于专用集成电路(ASIC)或可编程的逻辑设备(PLD)的电路中。在该情况中，所述实施方案以可被用来生成ASIC或PLD的计算机可读描述符实施。本发明的一些实施方案在各种各样的其他数字化装置的电路或逻辑处理器，例如PDA、笔记本电脑系统、显示器、图像编辑设备等内实施。

[0576] 在一些实施方案中，本发明涉及包括计算机可执行指令存储于其上的一个或更多个计算机可读存储介质的计算机程序产品，所述计算机可执行指令当被计算机系统的一个或更多个处理器执行时，导致计算机系统实施用于鉴定影响期望活性的生物分子方法。该方法是本文描述的任何方法，诸如由附图和伪代码涵盖的那些方法。在一些实施方案中，所述方法接收对于多个生物分子的序列和活性数据，并从所述序列和活性数据制作基础模型和改进的模型。在一些实施方案中，所述模型将活性预测为序列的亚单位的存在或不存在的函数。

[0577] 在本发明的一些实施方案中，由计算机程序产品实施的方法通过对基础模型加上/减去至少一个新的交互作用项来制作至少一个新的模型，其中所述新的交互作用项代表两个或更多个相互作用的亚单位之间的相互作用。在一些实施方案中，所述方法确定所述至少一个新的模型将活性预测为亚单位的存在或不存在的函数的能力。该方法还基于如以上确定的所述至少一个新的模型预测活性的能力以及以对包括新的交互作用项的偏倚来决定是否对基础模型加上、减去所述新的交互作用项。

[0578] VIII.实施例

[0579] 以下实施例展示了定向进化的程序，实施了本发明的多个实施方案。所述程序包括创建蛋白变体文库，基于所述文库的变体生成多种形式的序列-活性模型，和进行蛋白的定向进化以实现感兴趣的蛋白活性的期望水平。

[0580] 首先，构建组合的蛋白变体文库，使用重叠延伸自动平行剪接(SOEing)或APS技术，所述蛋白变体具有11个突变。该文库以每个变体约3个突变的速率随机并入突变。蛋白变体的11个突变被示于表VI的前11行。表VI线性相加模型、非交互作用相乘模型和交互作用相乘模型的系数值

[0581]

突变(位置、核苷酸)	相加模型的系数	相乘模型的系数	相乘&交互作用模型的系数

[0582]

1A	0.33	0.48	0.35
1G	0.28	0.49	0.53
2A	-0.63	-0.43	0.1
2G	-0.92	-0.67	-0.35
3A	0.035	0.13	0.037
4A	-0.32	-0.06	-0.17
5A	-1.5	-0.86	-0.84
6A	0.50	1.3	1.4
6G	0.58	1.15	0.95
7A	0.22	0.42	0.84
8A	-1.6	-0.88	-0.83
截距	1.9	-	-
2A*7A	-	-	0.89
2G*7A	-	-	0.80
2A*3A	-	-	0.28
2G*4A	-	-	0.65

[0583] 然后该程序获得针对该文库的序列/活性数据。使用具标条形码能力的下一代测序方法对所构建的变体测序。测序试验包括6x 96孔板，每个板包含6个骨架阳性对照孔和6个阴性对照孔。该试验还测量了蛋白的适应性或期望的活性。适应性被记录为相对于阳性对照的比，作为相对于阳性对照(FIOP)测量的改进倍数。

[0584] 以上获得的序列/活性数据被用来构建代表突变的存在/不存在的数据矩阵。对于总的n行(179)，数据矩阵的每行代表一种变体(序列)。矩阵的第一列包含代表变体活性的FIOP值。所有其他列包含代表突变存在/不存在的值。

[0585] 所获得的数据矩阵使用以下形式的非交互作用相加基础模型来拟合：

[0586] $y=\beta+C_{1A}X_{1A}+C_{1G}X_{1G}+C_{2A}X_{2A}+C_{2G}X_{2G}+C_{3A}X_{3A}$

[0587] $+C_{4A}X_{4A}+C_{5A}X_{5A}+C_{6A}X_{6A}+C_{6G}X_{6G}+C_{7A}X_{7A}+C_{8A}X_{8A}$其中y为由模型预测的活性，并且β为线性方程式的“截距”。

[0588] 使用下列程序，使用伴随置换的随机抽样的靴襻技术被用来使模型符合所获得的数据：

[0589] (a)对数据矩阵的行的数据伴随置换地抽样行数n(179)次。

[0590] (b)使用普通最小二乘回归拟合所抽取的数据。

[0591] (c)重复步骤(a)和(b)500次。

[0592] (d)对来自(c)的每次的系数求平均值以获得最终模型的系数值。

[0593] (e)从(d)获得的模型被用来预测数据矩阵中的序列数据的FIOP值。对于每个观察对象,观察值和预测值被一起绘制在图9A中。

[0594] 该非交互作用相加模型的系数的值被显示于表VI的第二列中。如在图9A中可看到的,非交互作用相加模型高估了对于很多观察对象,特别是具有低的观察到的活性水平的那些观察对象的活性值。

[0595] 第二个实例使用相乘模型来拟合相同的数据。所述模型具有以下形式:

[0596] $y=\beta\times(1-C_{1A}X_{1A})\times(1-C_{1G}X_{1G})\times(1-C_{2A}X_{2A})\times(1-C_{2G}X_{2G})\times(1-C_{3A}X_{3A})\times(1-C_{4A}X_{4A})\times(1-C_{5A}X_{5A})\times(1-C_{6A}X_{6A})\times(1-C_{6G}X_{6G})\times(1-C_{7A}X_{7A})\times(1-C_{8A}X_{8A})$。

[0597] 备选地,如果系数值被设定为有效地将常数1并入y值的计算,则模型可采取以下形式。例如,如果以上模型的系数被设定为横跨-0.9至1.5的范围,那么对于以下备选模型的系数可被设定为横跨0.1至2.5的范围,以完成相似的模型输出:

[0598] $y=\beta\times C_{1A}X_{1A}\times C_{1G}X_{1G}\times C_{2A}X_{2A}\times C_{2G}X_{2G}\times C_{3A}X_{3A}\times C_{4A}X_{4A}\times C_{5A}X_{5A}\times C_{6A}X_{6A}\times C_{6G}X_{6G}\times C_{7A}X_{7A}\times C_{8A}X_{8A}$。

[0599] 该模型通过使用靴襻技术和遗传算法技术调整系数值来精化。按照以下步骤、通过使预测活性值(或适应性)相对于观察值的均方误差最小化来进行模型拟合:

[0600] (a)对行的序列/活性数据伴随置换地抽样行数n次以提供靴带数据(bootstrapping data)。

[0601] (b)使用遗传算法使模型符合所抽取的数据,这将使系数值限制在-0.9和1.5之间。遗传算法具有200的群容量,并且其运行100代。

[0602] (c)重复步骤(a)和(b)500次。

[0603] (d)对来自(c)的每次的前10个染色体/个体求平均值以获得最终模型的系数。

[0604] (e)对于数据矩阵的每行,使用(d)预测FIOP值。对于每行数据(一个观察对象),观察值和预测值被一起绘制在图9B中。该相乘基础模型的系数值被显示于表VI的第三列。

[0605] 如在图9B中看到的,该非交互作用相乘模型具有更好地匹配变体文库的观察到的活性值的预测值,使得其对具有低的观察到的活性水平的蛋白变体没有显著的、一贯的高估。但是,剩余误差仍是相对大的,数据点从图点的对角线散布开来。

[0606] 下一个实例阐明了使用逐步方法鉴定形成交互作用相乘模型的合适的交互作用项精化的相乘基础模型。该程序以下步骤开始精化来自前述实例中概述的相乘基础模型的相乘模型。

[0607] (a)以上获得的基础模型被设定为最佳模型。

[0608] (b)所有可能的成对交互作用系数被放入系数集。(1A*2A,1A*3A,1A*4A,…,6G*8A,7A*8A)

[0609] (c)每个系数被加至最佳模型,所述最佳模型使用具有前述实例中描述的相同参数的遗传算法拟合。

[0610] (d)使用AIC获得来自(c)的每个新模型的适应性。

[0611] (e)来自(d)的具有最低AIC的模型被设定为当前轮的最佳模型。

[0612] (f)如果来自(e)的模型比最佳模型好，则最佳模型被设定为该模型，并且该模型的系数被从系数集中去除，并且运算进入(c)-否则，没有新的模型并且运算结束。

[0613] (g)来自(f)的新模型使用前面的实例中概述的靴襻法/拟合法拟合。

[0614] 该交互作用模型的系数值被显示于表VI的第四列，其中最后4个系数是针对交互作用项的。如在所述表中通过三列系数值可看到的，模型内的系数的相对量级保持相似的模式，表明对于所有的模型优化程序的正确收敛。

[0615] 图9C绘制了如由交互作用相乘模型预测的蛋白变体的活性对蛋白变体的观察到的活性水平的图。此处，在三个模型中，交互作用相乘模型在预测观察对象的活性方面做得最好。遍及活性水平的整个范围，剩余误差一贯地小，数据点分布在对角线附近。

[0616] 下一个实例显示了以上获得的交互作用相乘模型如何被用于指导蛋白变体的定向进化以达到期望的蛋白活性。

[0617] 一个实例使用已经存在于当前文库中的突变。在该实例中，所有貌似合理的突变组合的活性使用以上获得的交互作用相乘模型预测。然后通过计算机模拟构建所有可能的变体的序列，并使用交互作用相乘模型计算其活性(FIOP)。预测的FIOP从最高到最低排序，并列于表VII中。

[0618] 以下显示了前5次预测，以及其预测值和观察值(如果可得的话)。合成并重新测未观察的变体，并与观察的变体一起比较。在重复测试中表现最佳的变体被推进至创建待进化的新骨架。

[0619] 表VII.由交互作用相乘模型预测的最合适的序列

[0620]

预测的	观察的	突变
7	NA	1G,3A,6A,7A
6.7	5.99	1G,6A,7A
6.2	5.8	1A,3A,6A,7A
5.9	NA	1A,6A.7A
5.8	NA	1G,3A,4A,6A,7A

[0621] 在该定向进化实例中，有害突变和突变的组合得以记录，并从在即的几轮进化中排除。而且，被鉴定到但未被包括在下一轮骨架中的有益的多样性以组合的方式与之前鉴定的有益的多样性(如果可得的话)重组在一起。

[0622] 定向进化的另一个实例使用诱变多样性机制生成新的多样性。通过生成多样性来搜索序列空间，所述多样性通过以组合的方式在被交互作用相乘模型鉴定为对感兴趣的活性具有显著的相互作用性影响的位置处进行饱和诱变生成。这些位置为具有高的交互作用系数值的那些位置：在该情况中是位置2、3、4和7。在这些位置同时地进行饱和诱变。筛选所得到的文库针对相对于骨架的活性，这有助于鉴定到表现较好的变体。

[0623] 虽然前面已经为了清楚和理解的目的以一定的细节进行了描述，但对于本领域技术人员通过阅读该公开内容将清楚的是，可进行多种形式和细节上的多种改变而不偏离本公开内容的真实范围。例如，以上描述的所有技术和装置可以多种组合利用。本申请中提到的所有出版物、专利、专利申请或其他文件，在如同每个单独的出版物、专利、专利申请和其他文件被单独地指明为了所有目的被通过引用并入的程度上，被为了所有目的通过引用全部并入。

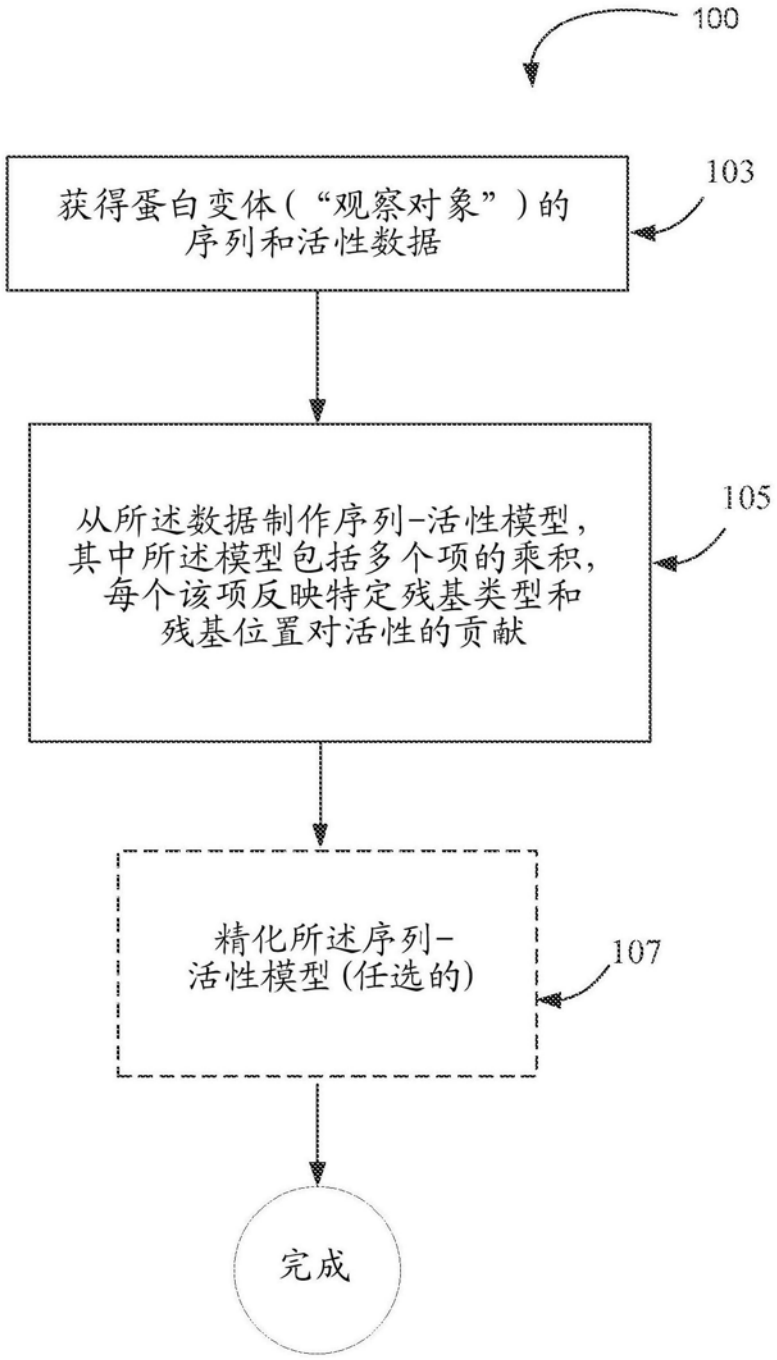

100
获得蛋白变体(“观察对象”)的序列和活性数据
103
从所述数据制作序列-活性模型，其中所述模型包括多个项的乘积，每个该项反映特定残基类型和残基位置对活性的贡献
105
精化所述序列-活性模型(任选的)
107
完成

图1A

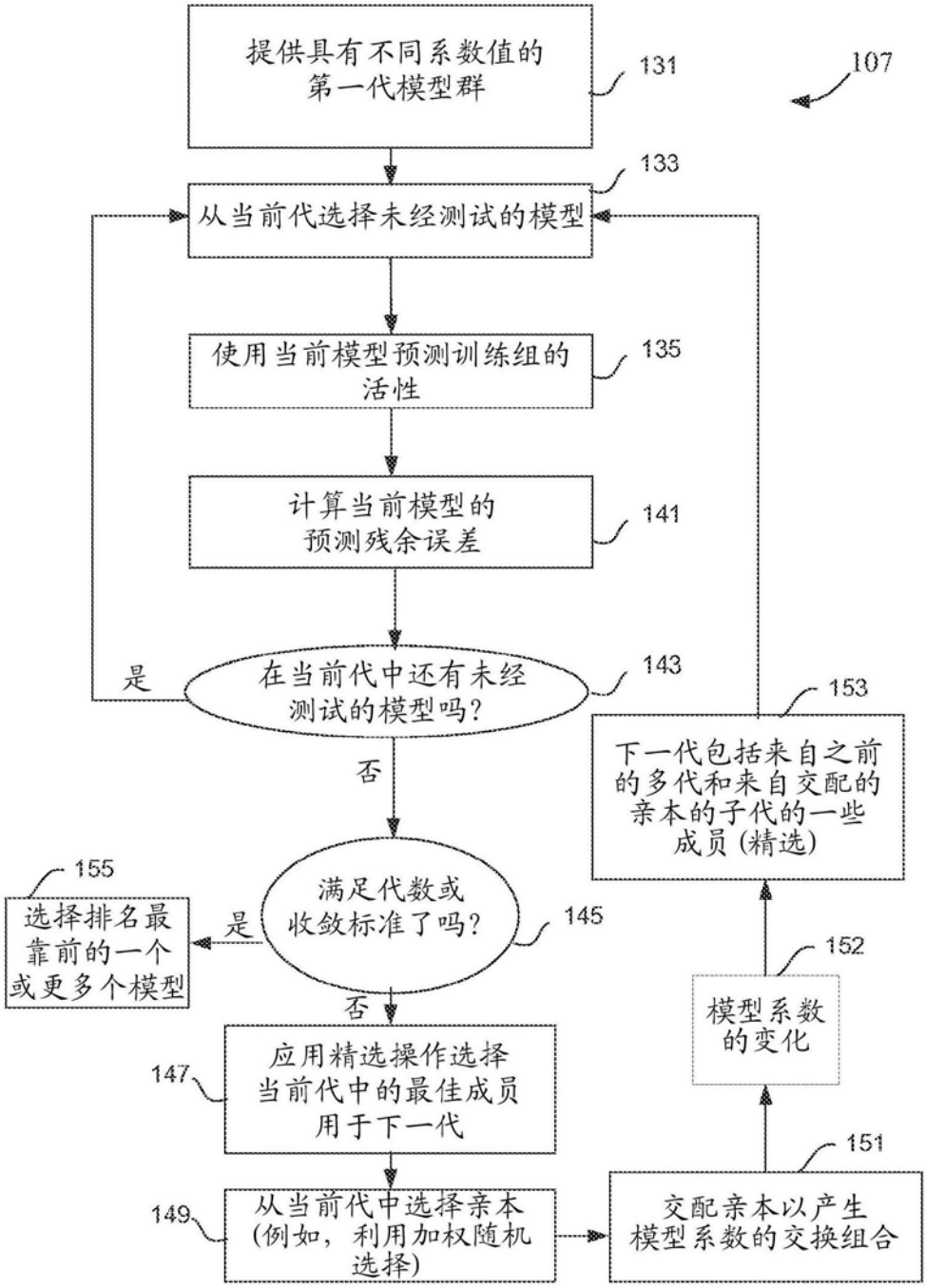

提供具有不同系数值的第一代模型群
131
107
133
从当前代选择未经测试的模型
使用当前模型预测训练组的活性
135
计算当前模型的预测残余误差
141
是
在当前代中还有未经测试的模型吗?
143
否
153
下一代包括来自之前的多代和来自交配的亲本的子代的一些成员(精选)
155
选择排名最靠前的一个或更多个模型
是
满足代数或收敛标准了吗?
145
否
152
模型系数的变化
147
应用精选操作选择当前代中的最佳成员用于下一代
151
149
从当前代中选择亲本(例如,利用加权随机选择)
交配亲本以产生模型系数的交换组合

图1B

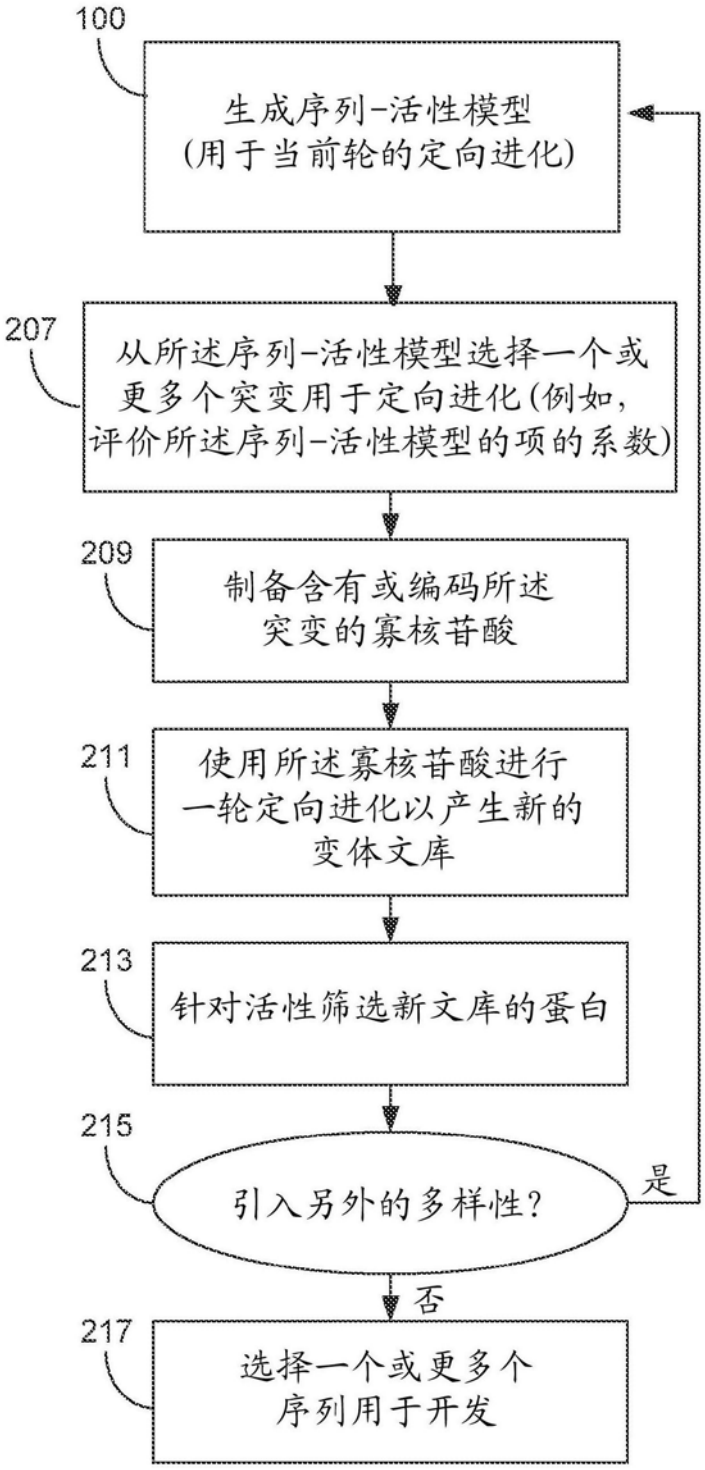

100
生成序列-活性模型
(用于当前轮的定向进化)
207
从所述序列-活性模型选择一个或
更多个突变用于定向进化(例如,
评价所述序列-活性模型的项的系数)
209
制备含有或编码所述
突变的寡核苷酸
211
使用所述寡核苷酸进行
一轮定向进化以产生新的
变体文库
213
针对活性筛选新文库的蛋白
215
引入另外的多样性?
是
否
217
选择一个或更多个
序列用于开发

图2

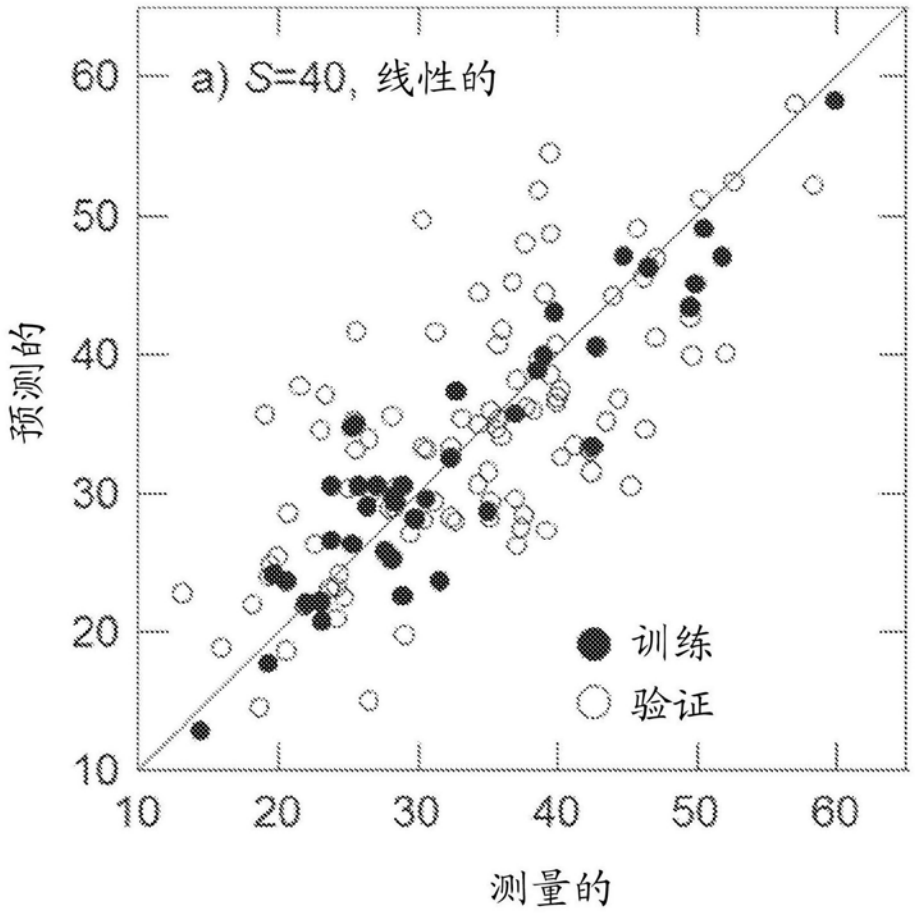


图3A

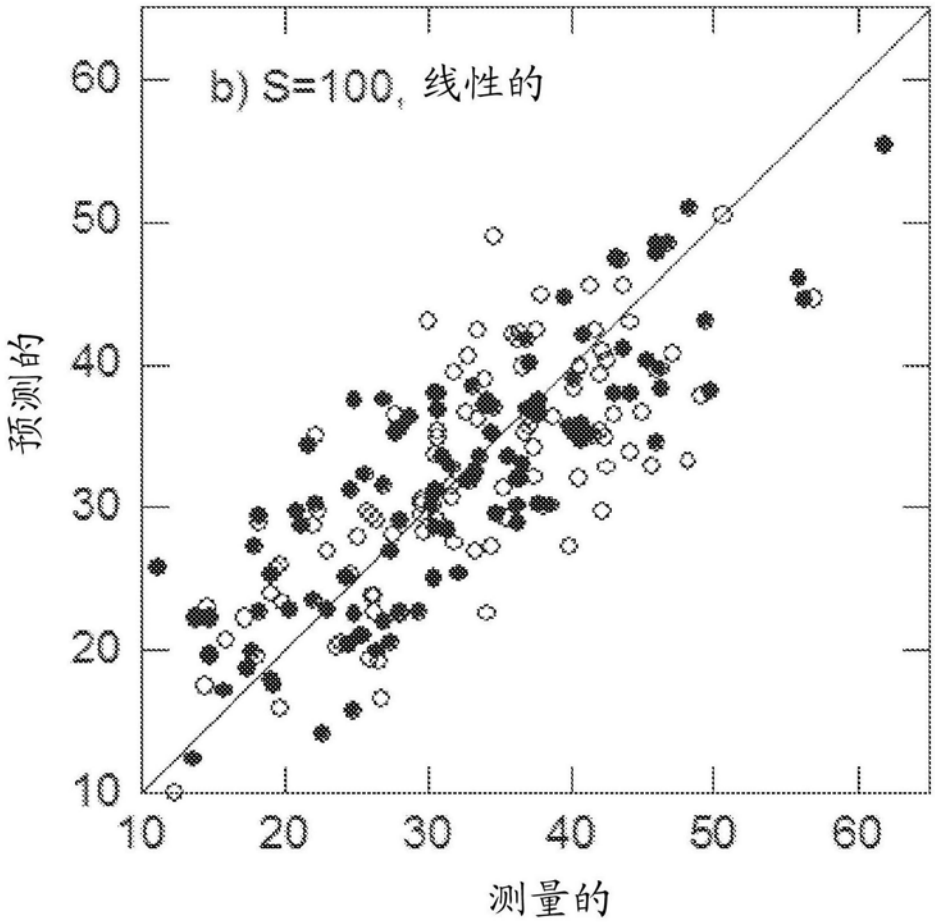


图3B

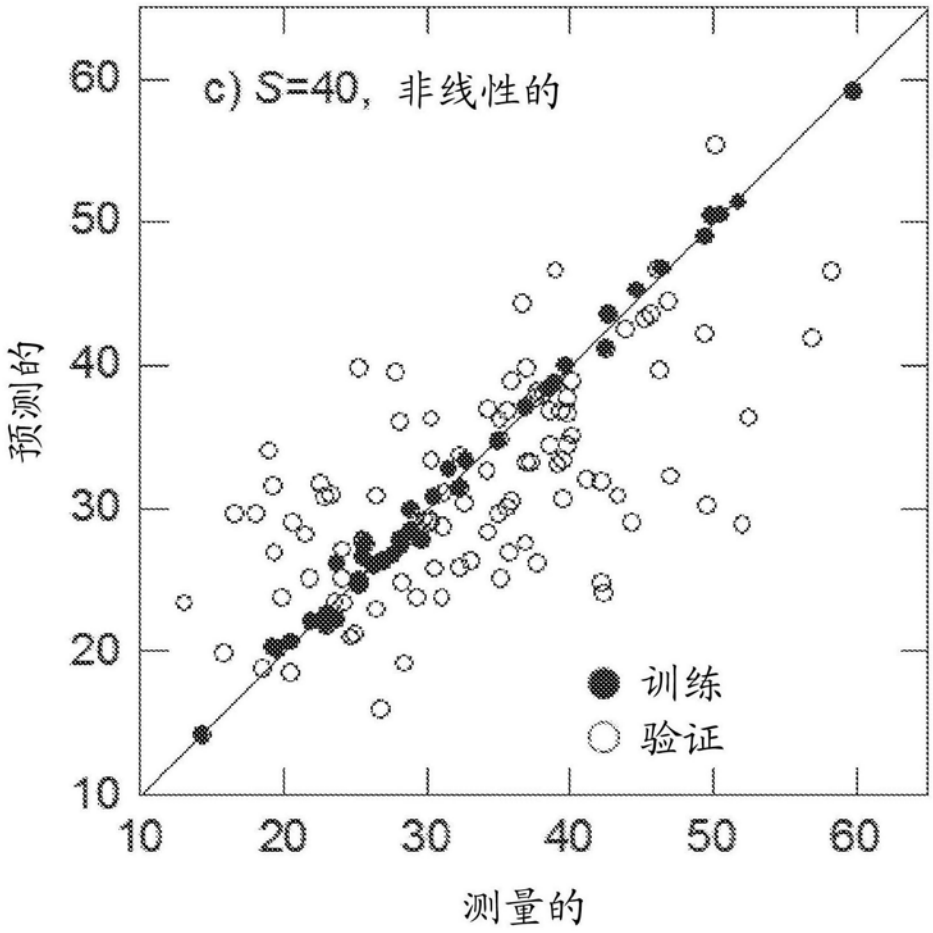


图3C

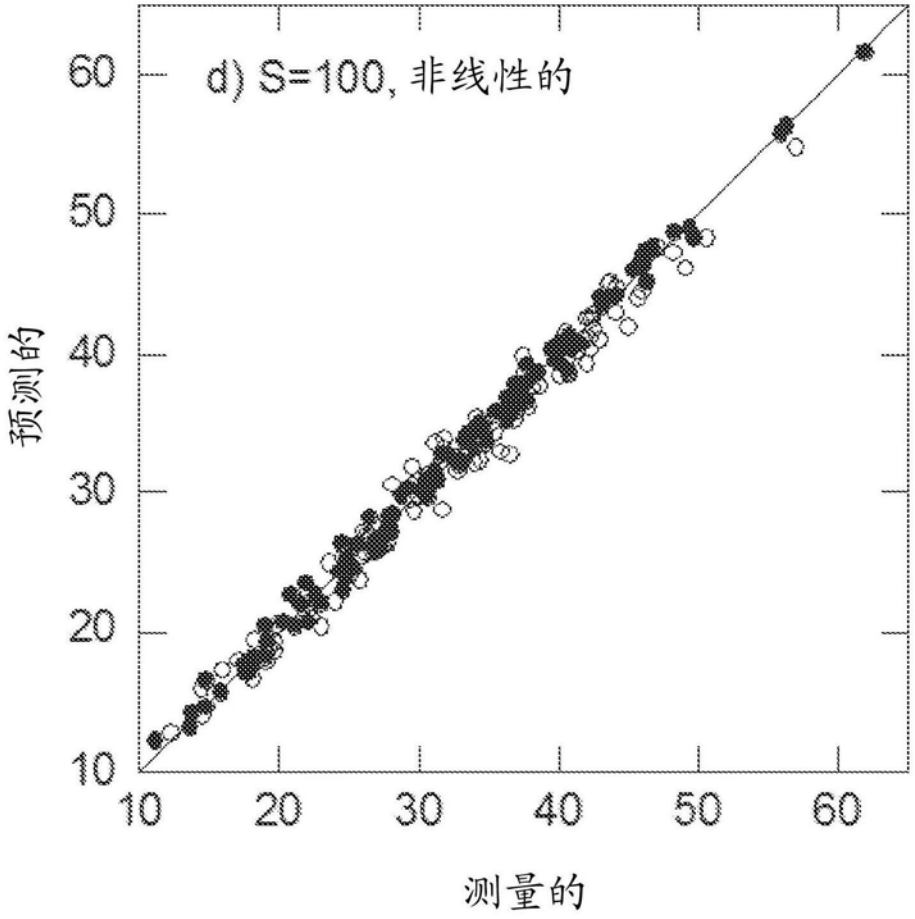


图3D

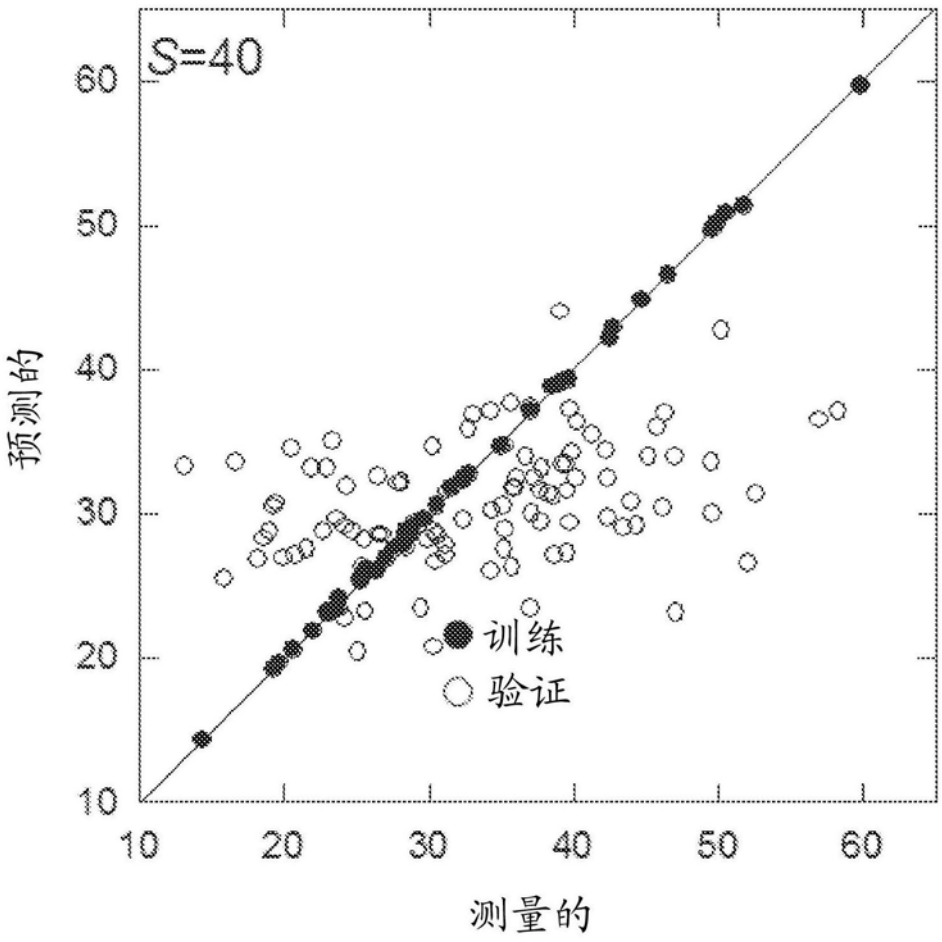


图3E

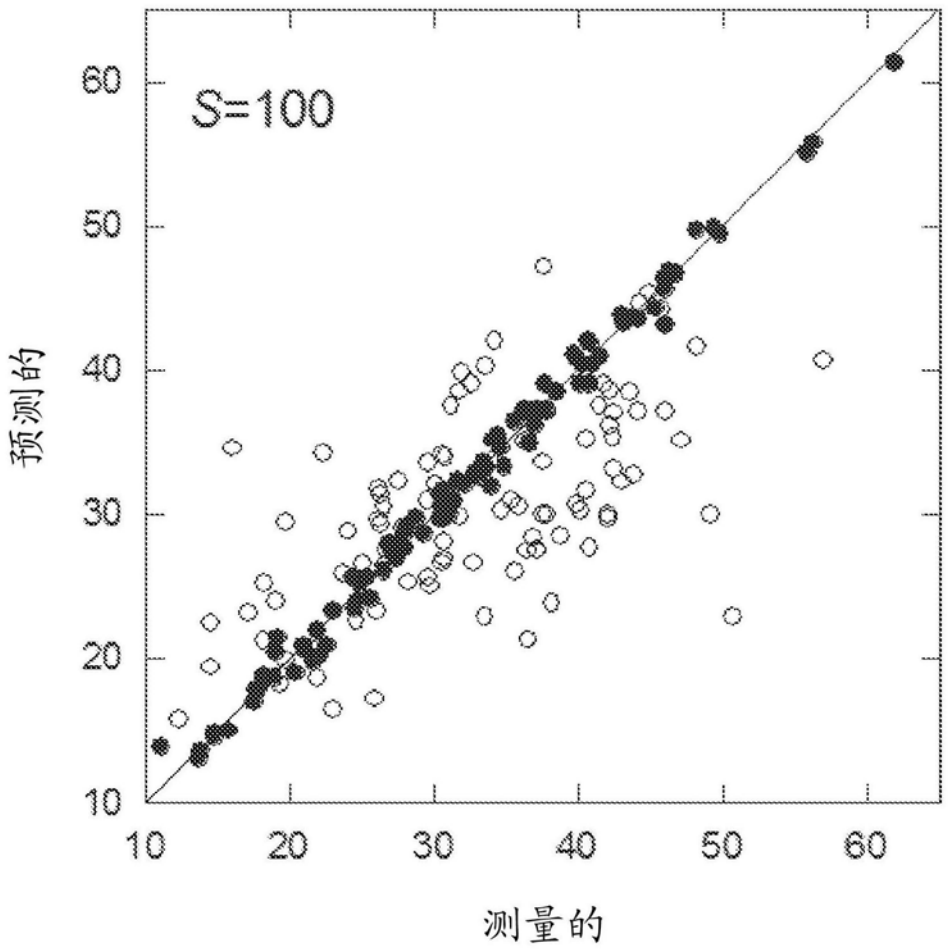


图3F

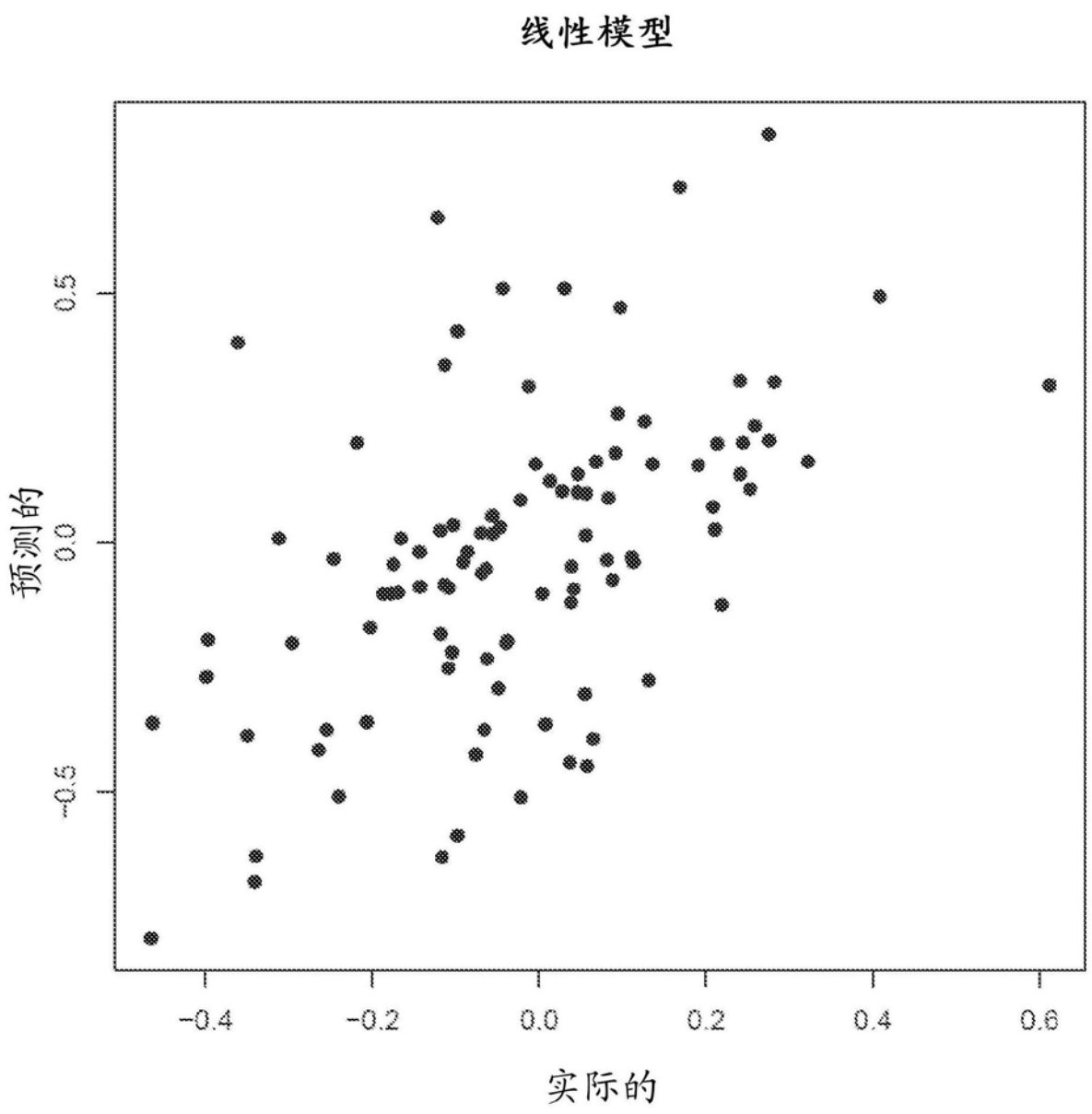


图3G

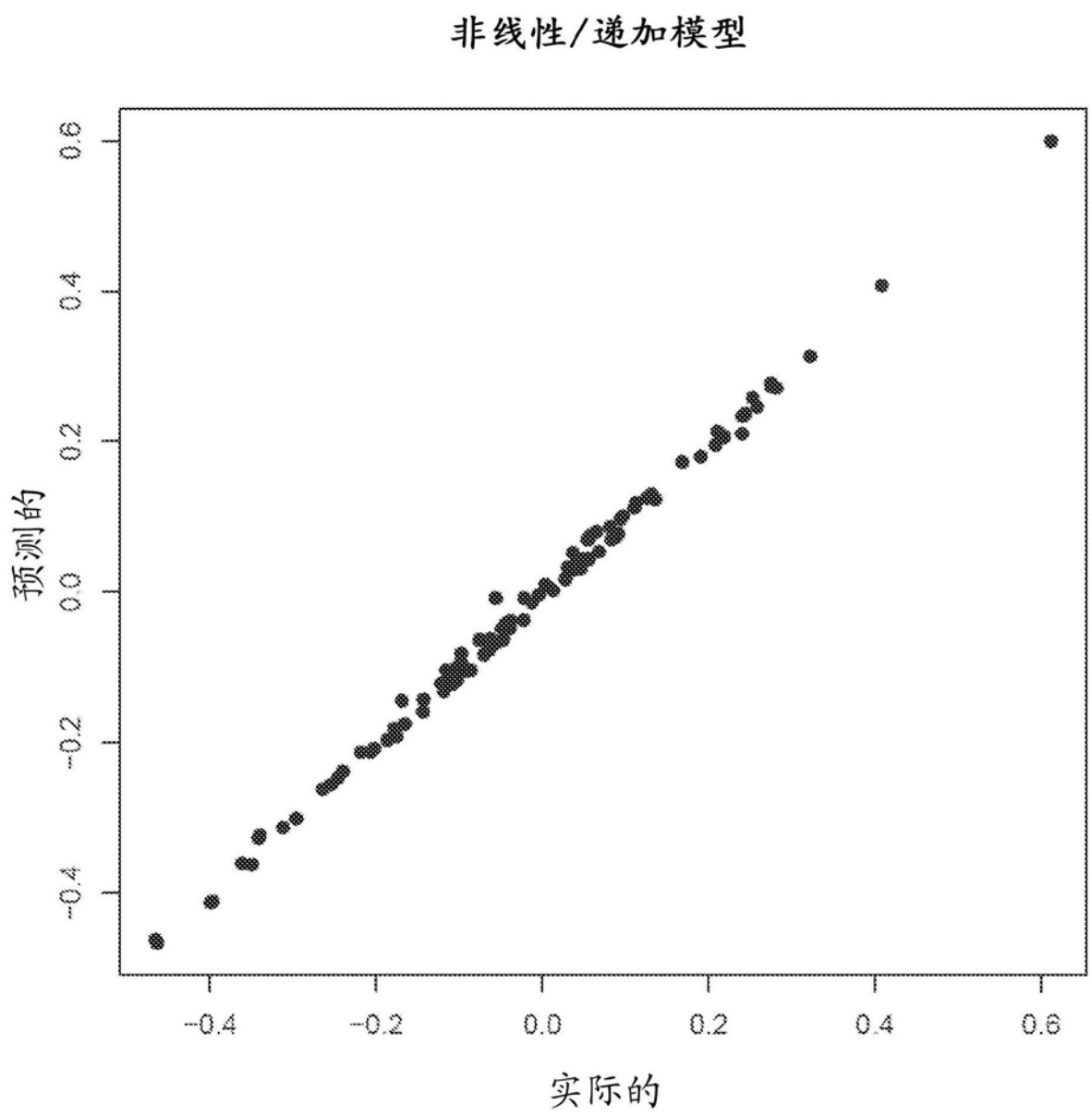


图3H

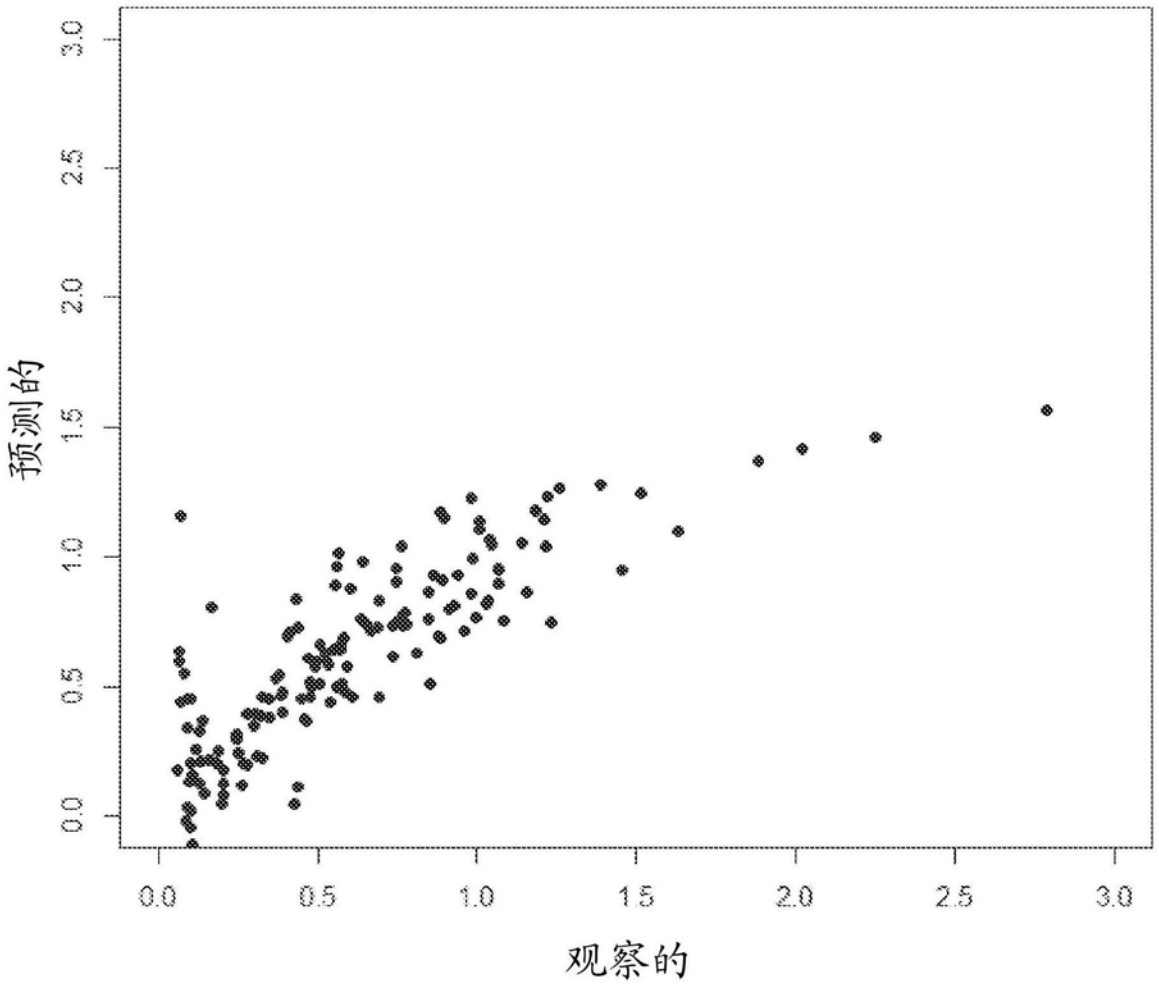


图3I

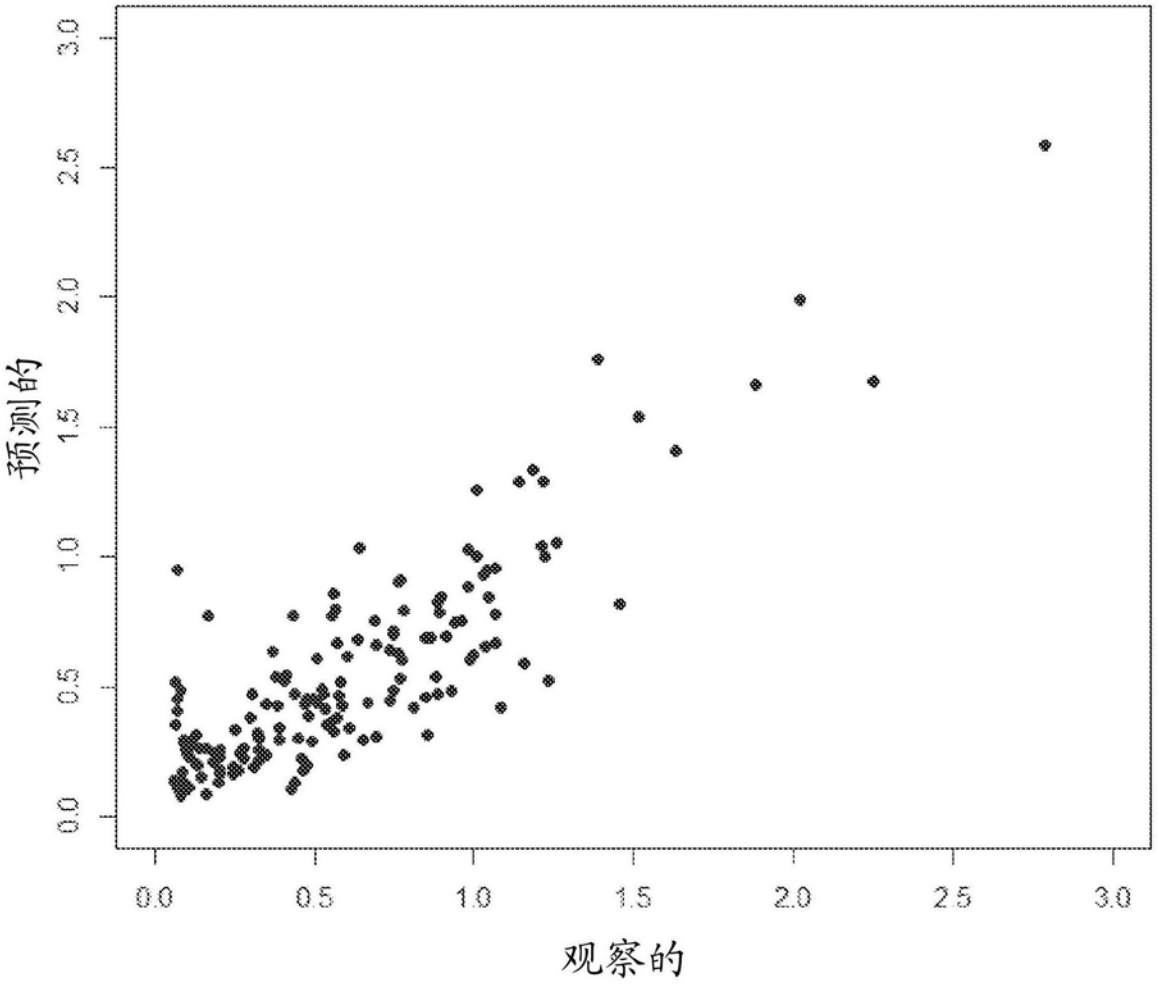


图3J

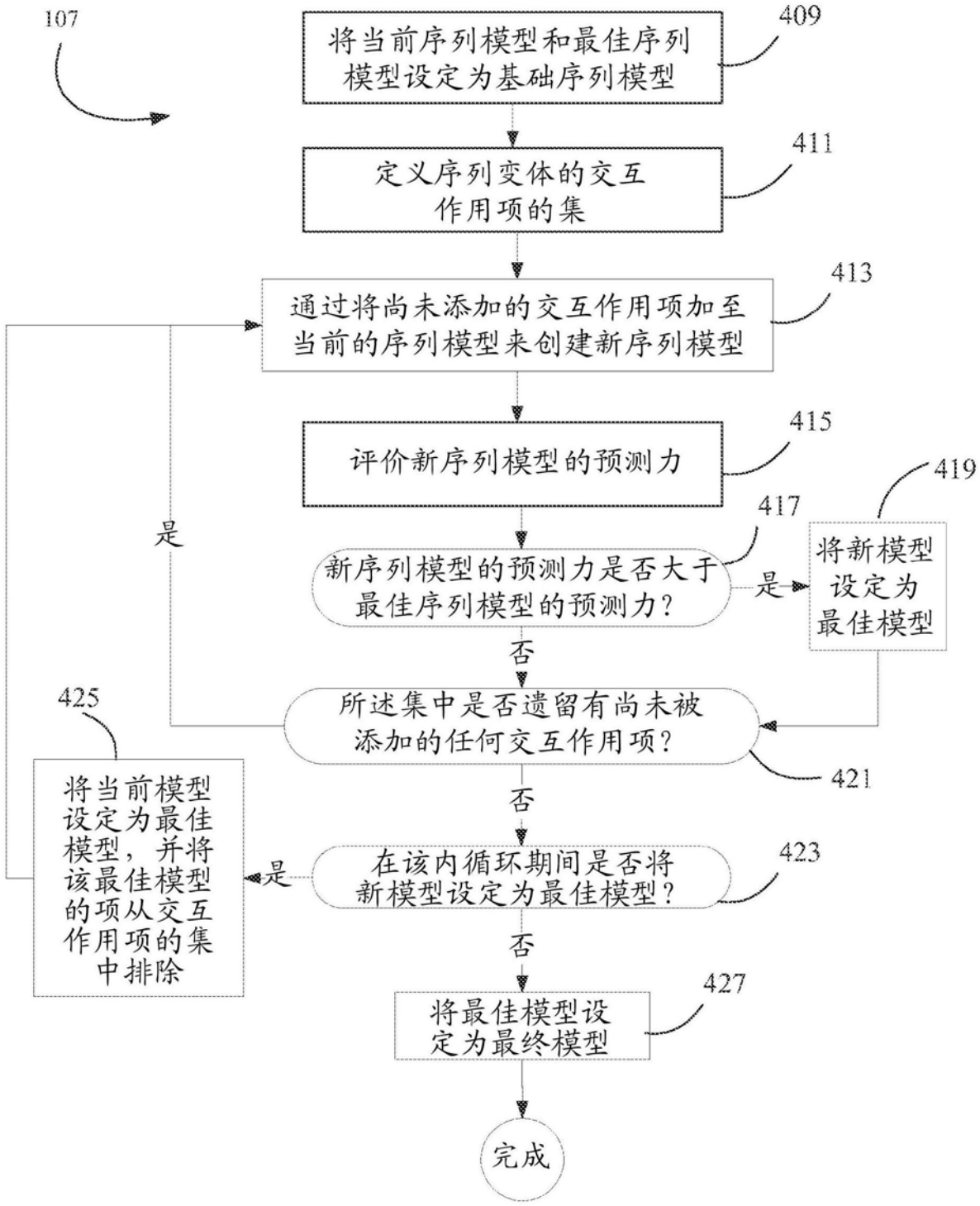

107
将当前序列模型和最佳序列模型设定为基础序列模型
409
定义序列变体的交互作用项的集
411
通过将尚未添加的交互作用项加至当前的序列模型来创建新序列模型
413
评价新序列模型的预测力
415
新序列模型的预测力是否大于最佳序列模型的预测力?
417
是
将新模型设定为最佳模型
419
否
所述集中是否遗留有尚未被添加的任何交互作用项?
421
是
否
在该内循环期间是否将新模型设定为最佳模型?
423
是
将当前模型设定为最佳模型，并将该最佳模型的项从交互作用项的集中排除
425
否
将最佳模型设定为最终模型
427
完成

图4A

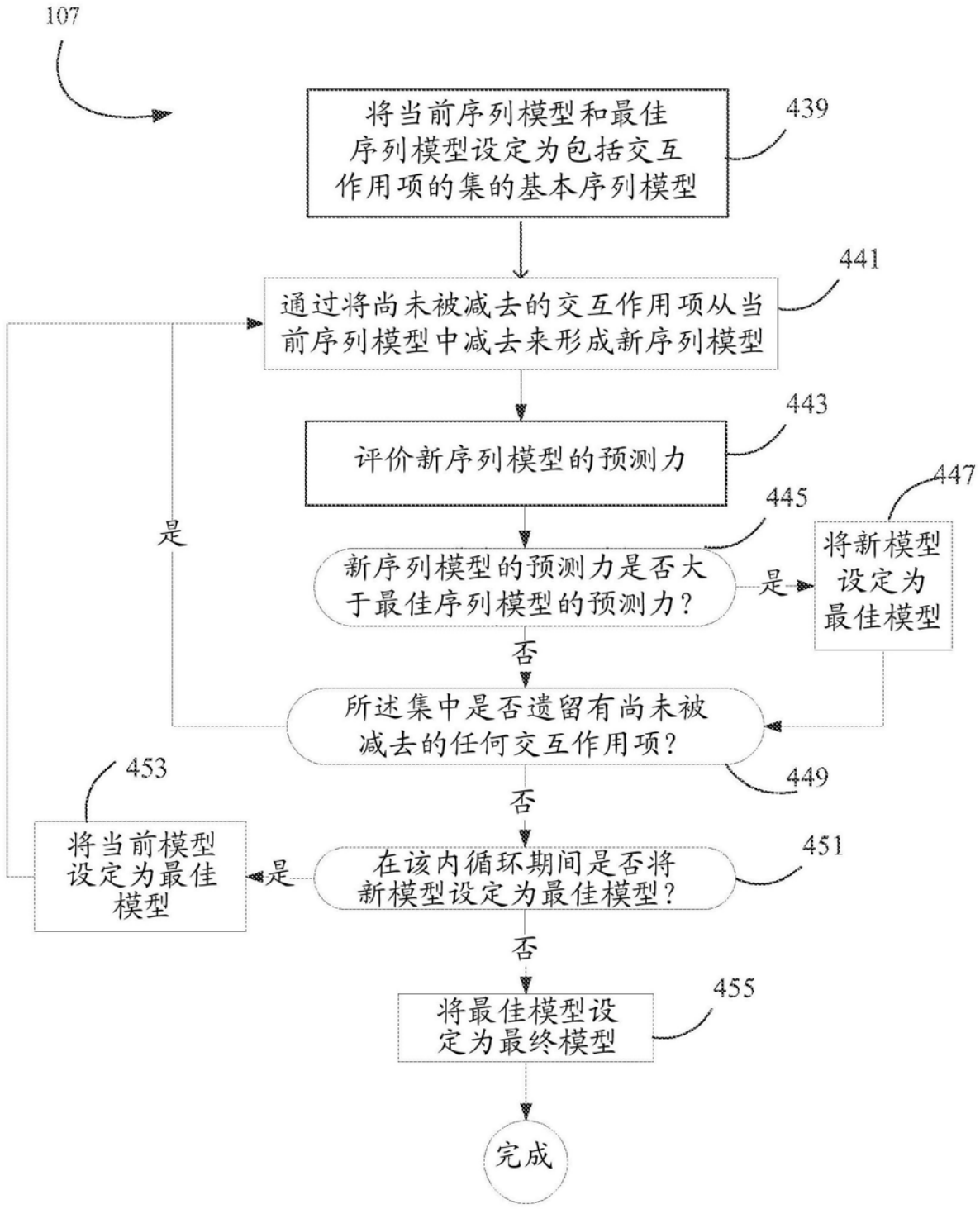

107
将当前序列模型和最佳序列模型设定为包括交互作用项的集的基本序列模型
439
通过将尚未被减去的交互作用项从当前序列模型中减去来形成新序列模型
441
评价新序列模型的预测力
443
新序列模型的预测力是否大于最佳序列模型的预测力？
445
是
将新模型设定为最佳模型
447
否
所述集中是否遗留有尚未被减去的任何交互作用项？
449
是
否
在该内循环期间是否将新模型设定为最佳模型？
451
是
将当前模型设定为最佳模型
453
否
将最佳模型设定为最终模型
455
完成

图4B

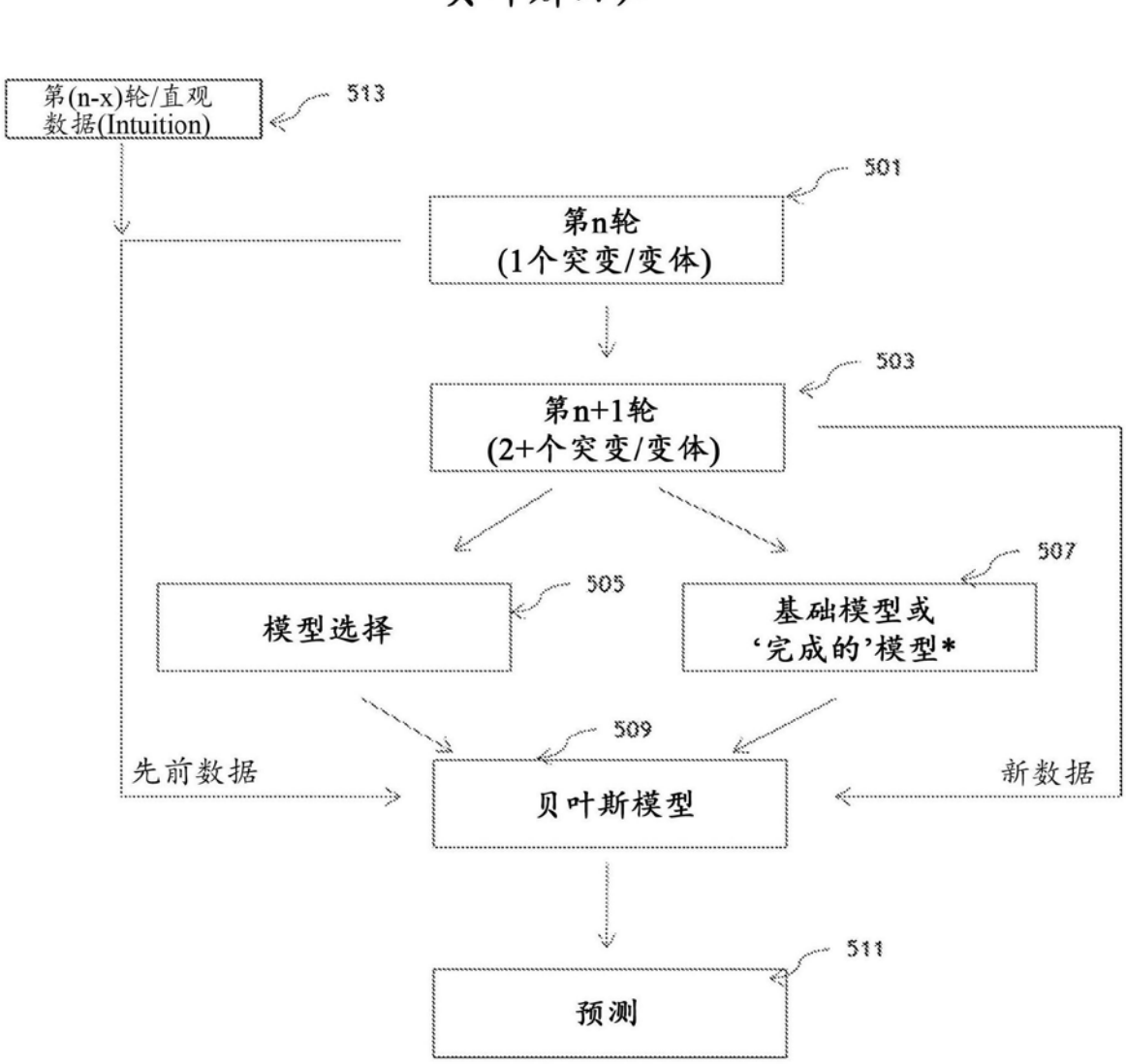

贝叶斯回归
第(n-x)轮/直观数据(Intuition)
513
501
第n轮
(1个突变/变体)
503
第n+1轮
(2+个突变/变体)
505
模型选择
507
基础模型或
'完成的'模型*
509
先前数据
贝叶斯模型
新数据
511
预测

图5

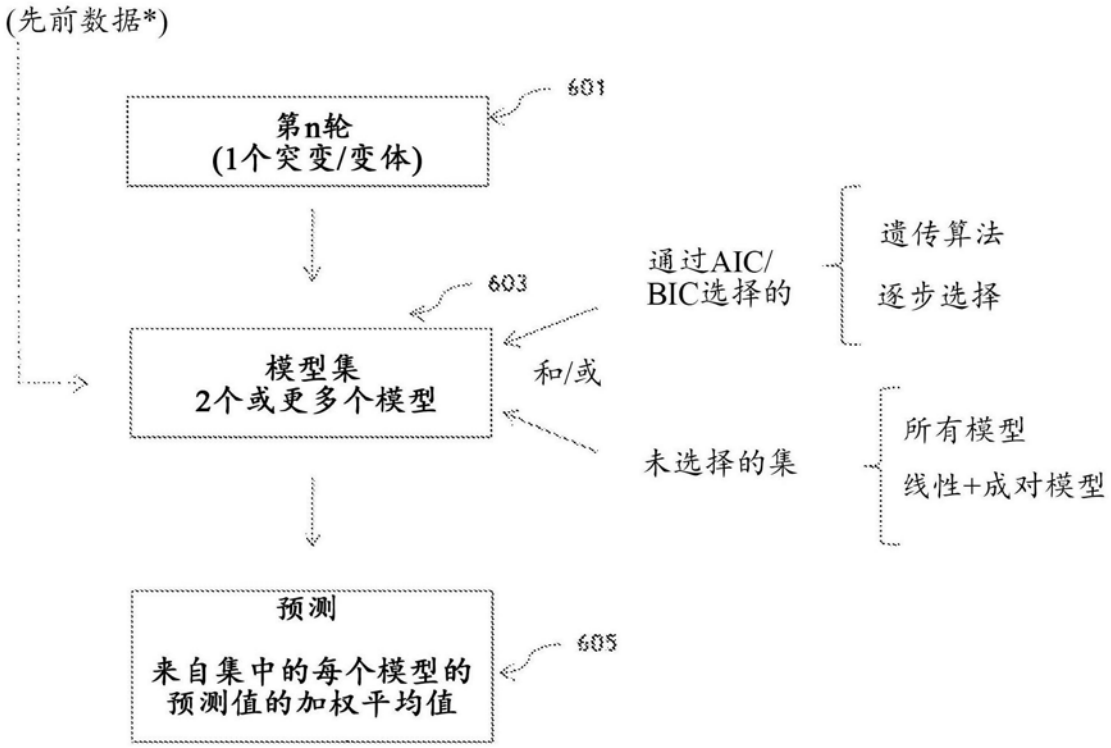

集成回归
(先前数据*)
601
第n轮
(1个突变/变体)
603
模型集
2个或更多个模型
通过AIC/
BIC选择的
遗传算法
逐步选择
和/或
未选择的集
所有模型
线性+成对模型
预测
来自集中的每个模型的
预测值的加权平均值
605

图6

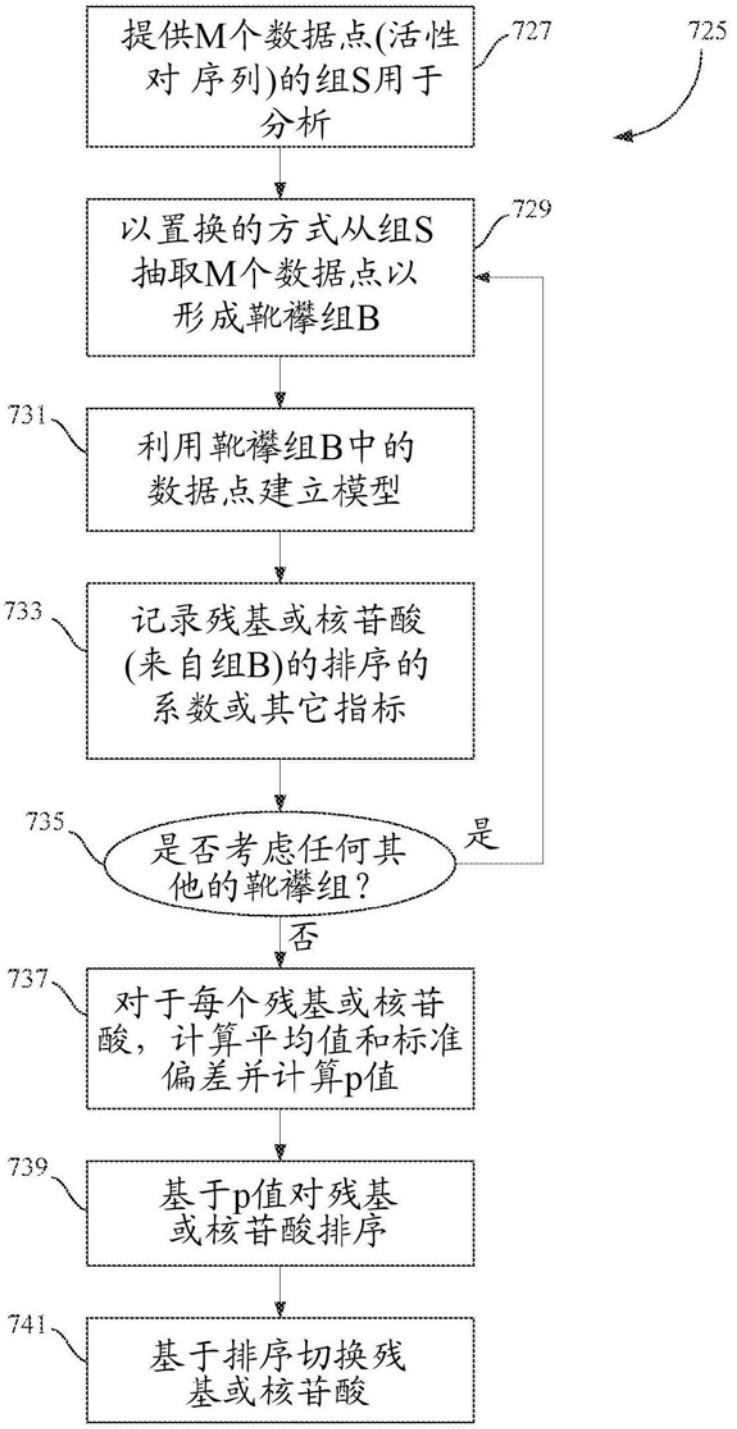

提供M个数据点(活性对 序列)的组S用于分析
727
725
以置换的方式从组S抽取M个数据点以形成靴襻组B
729
731
利用靴襻组B中的数据点建立模型
733
记录残基或核苷酸(来自组B)的排序的系数或其它指标
735
是否考虑任何其他的靴襻组?
是
否
737
对于每个残基或核苷酸，计算平均值和标准偏差并计算p值
739
基于p值对残基或核苷酸排序
741
基于排序切换残基或核苷酸

图7

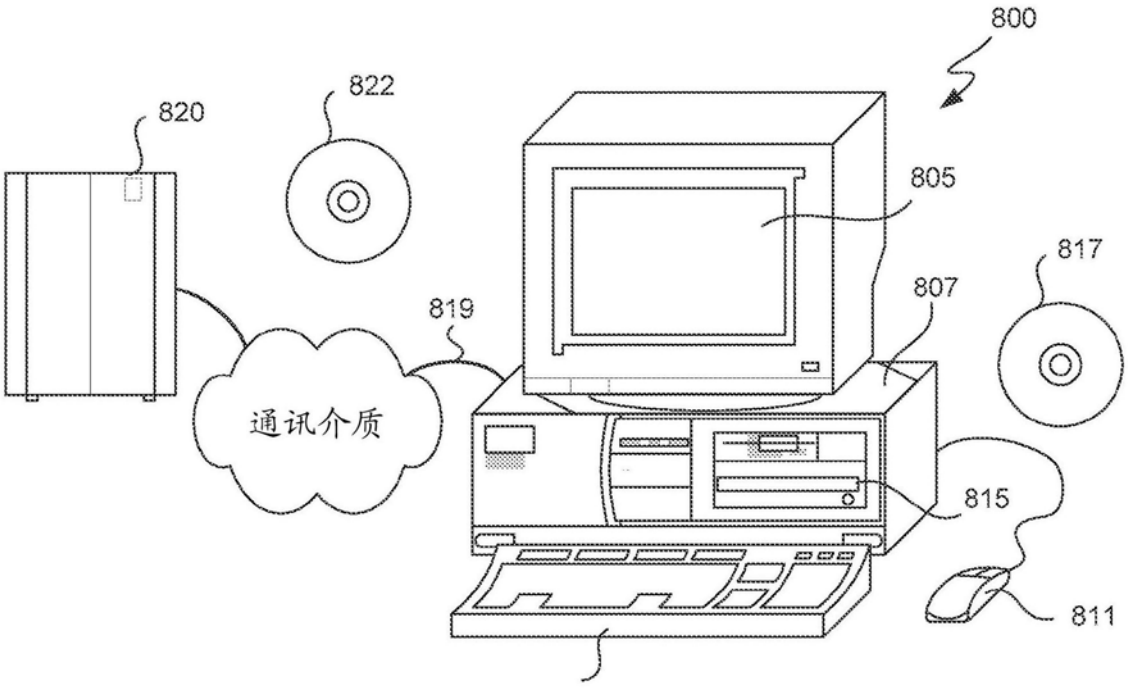


图8

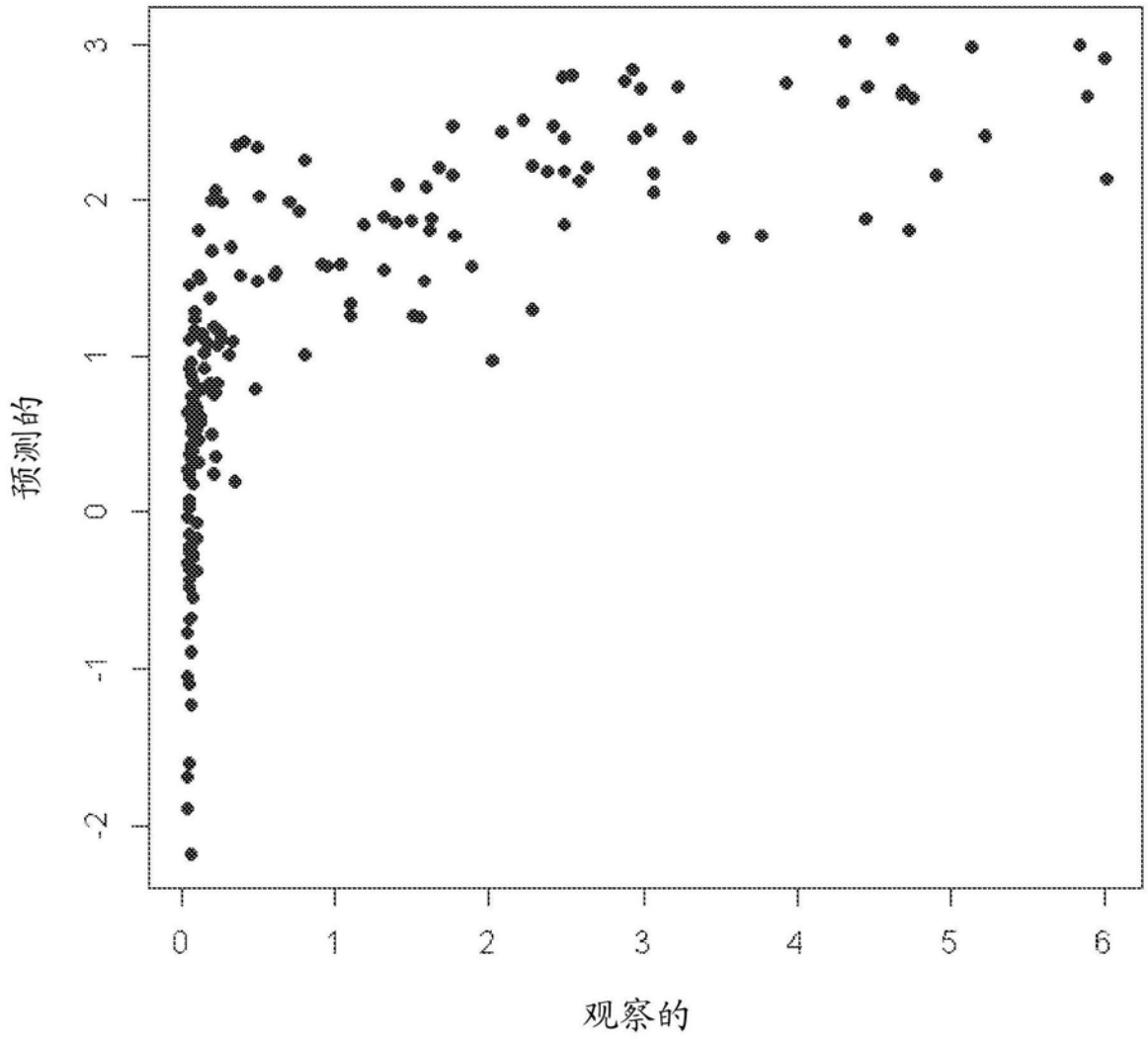


图9A

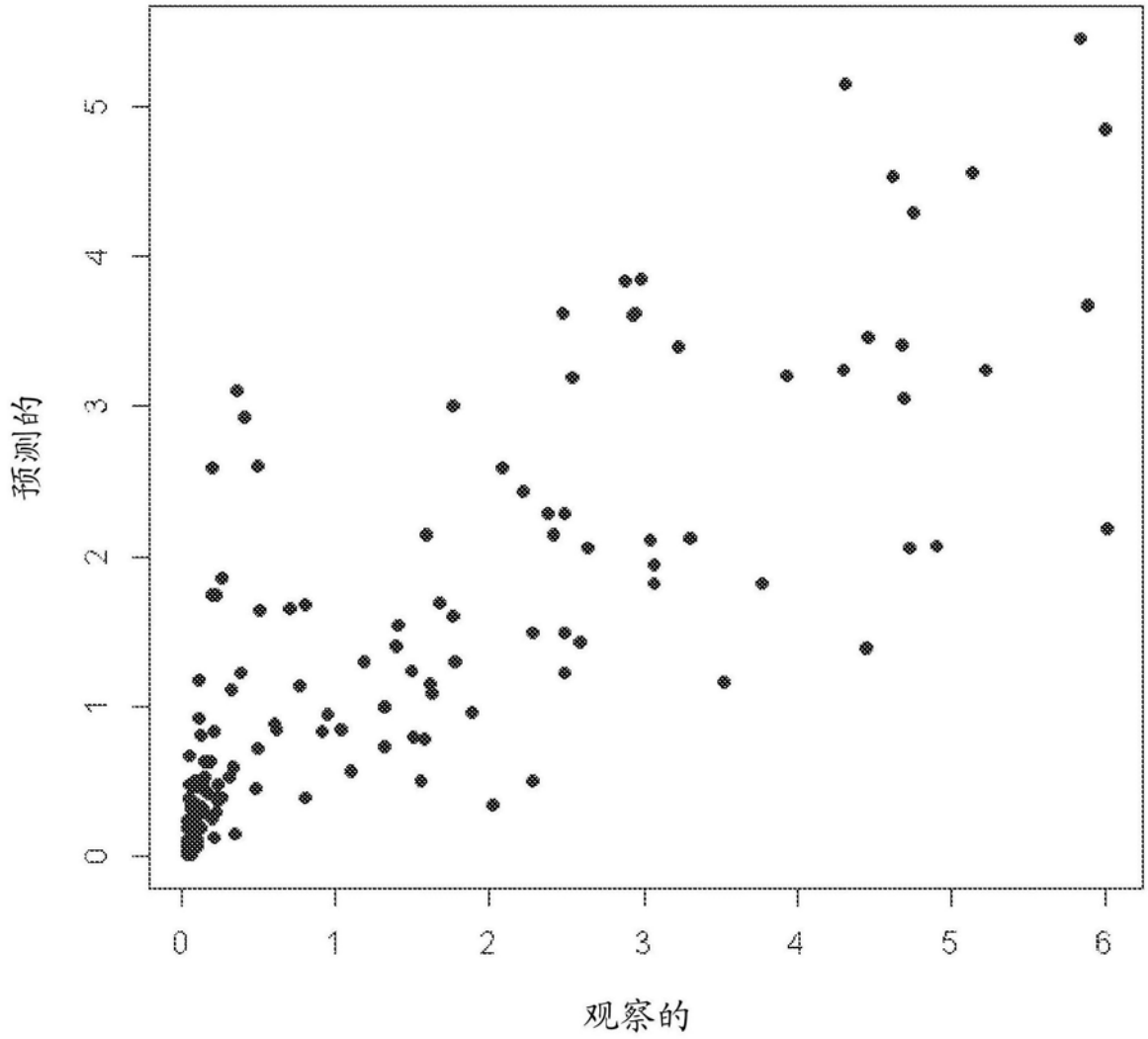


图9B

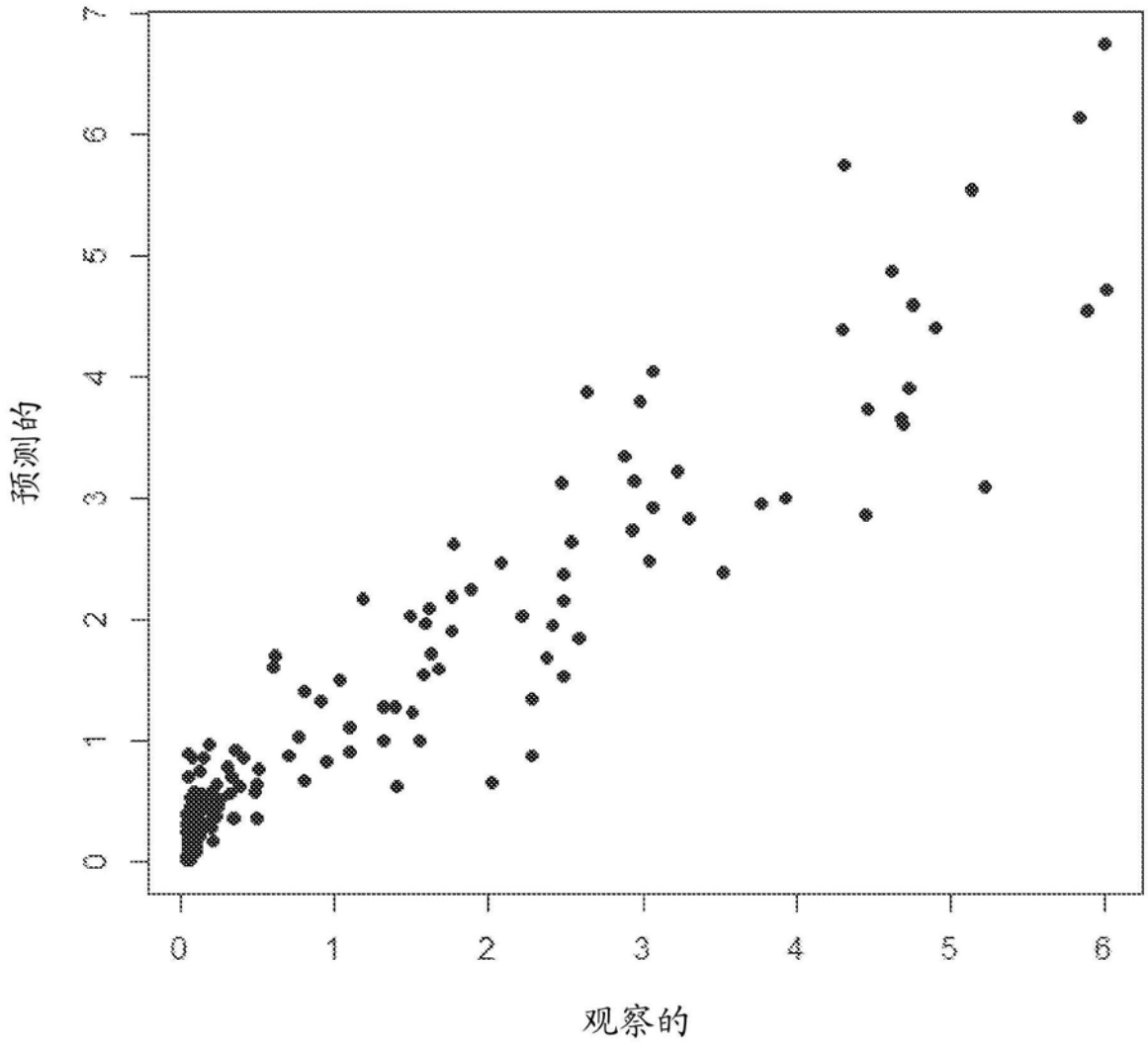


图9C