



(10) 授权公告号 CN 112585684 B

(45) 授权公告日 2024. 07. 19

(21) 申请号 201980054143.1

(22) 申请日 2019.09.16

(65) 同一申请的已公布的文献号

申请公布号 CN 112585684 A

(43) 申请公布日 2021.03.30

(30) 优先权数据

62/734,757 2018.09.21 US

62/734,773 2018.09.21 US

62/770,490 2018.11.21 US

(85) PCT国际申请进入国家阶段日

2021.02.18

(86) PCT国际申请的申请数据

PCT/EP2019/074670 2019.09.16

(87) PCT国际申请的公布数据

W02020/058174 EN 2020.03.26

(73) 专利权人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 安德鲁·W·塞尼尔

詹姆斯·柯克帕特里克

劳伦特·西弗尔

理查德·安德鲁·埃文斯

雨果·佩内多尼斯 秦翀立

孙若溪 凯伦·西蒙尼扬

约翰·将姆佩尔

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

专利代理师 金海霞 刘慧

(51) Int.Cl.

G16B 15/20 (2019.01)

G16B 40/20 (2019.01)

(56) 对比文件

Haiyou Deng 等,Protein structure
prediction,《World Scientific》.2018,第1-16
页.Henrik Fredholm, A Novel Approach to
Prediction of the 3-Dimensional Structure
of Protein Backbones by Neural Networks.
《NIPS》.1990,第524-528页.

审查员 徐倩

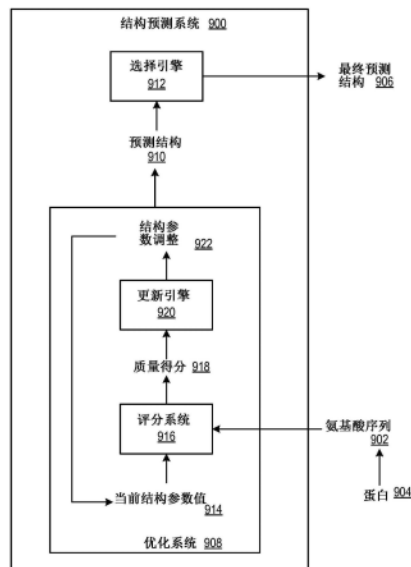
权利要求书5页 说明书55页 附图20页

(54) 发明名称

使用质量得分的梯度的迭代蛋白结构预测

(57) 摘要

用于执行蛋白结构预测和蛋白结构域分割的方法、系统和装置,其包括在计算机存储介质上编码的计算机程序。在一个方面,一种方法包括生成蛋白的多个预测结构,其中生成蛋白的预测结构包括:更新蛋白的多个结构参数的初始值,包括在多个更新迭代中的每一个处:确定结构参数的当前值的质量得分相对于结构参数的当前值的梯度;以及使用所述梯度更新结构参数的当前值。



1. 一种由一个或多个数据处理装置执行的用于确定给定蛋白的最终预测结构的方法, 其中所述给定蛋白包括氨基酸序列, 其中所述给定蛋白的预测结构由多个结构参数的值来定义, 所述方法包括:

生成所述给定蛋白的多个预测结构, 其中生成所述给定蛋白的预测结构包括:

获得定义所述预测结构的多个结构参数的初始值;

更新所述多个结构参数的初始值, 包括, 在多个更新迭代中的每一个处:

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分, 其中所述质量得分基于一个或多个评分神经网络的相应输出, 每个评分神经网络被配置为处理: (i) 所述结构参数的当前值, (ii) 所述给定蛋白的氨基酸序列的表示, 或 (iii) 两者; 和

对于所述多个结构参数中的一个或多个:

确定所述质量得分相对于所述结构参数的当前值的梯度; 以及

使用所述质量得分相对于所述结构参数的当前值的梯度来更新所述结构参数的当前值; 以及

在所述多个更新迭代中的最终更新迭代之后, 将所述给定蛋白的预测结构确定为由所述多个结构参数的当前值定义; 以及

选择所述给定蛋白的所述多个预测结构中的一个预测结构作为所述给定蛋白的最终预测结构, 其中:

所述一个或多个评分神经网络包括距离预测神经网络, 所述距离预测神经网络被配置为处理包括所述给定蛋白的氨基酸序列的表示的输入以生成所述给定蛋白的距离图;

所述距离图对于所述给定蛋白的氨基酸序列中的多对氨基酸中的每一对, 定义在该对氨基酸之间的可能距离范围内的相应概率分布; 和

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括, 对于所述给定蛋白的氨基酸序列中的每一对氨基酸:

使用在由所述距离图定义的该对氨基酸之间的可能距离范围内的相应概率分布, 确定该对氨基酸被由所述结构参数的当前值定义的距离隔开的概率, 以及

基于在所述给定蛋白的氨基酸序列中的每一对氨基酸上, 根据在由所述距离图定义的该对氨基酸残基之间的可能距离范围内的相应概率分布的、该对氨基酸被由所述结构参数的当前值定义的距离隔开的概率的乘积, 确定所述质量得分。

2. 根据权利要求1所述的方法, 其中确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括:

对于所述给定蛋白的氨基酸序列中的每一对氨基酸:

使用在由参考距离图定义的该对氨基酸之间的可能距离范围内的相应概率分布, 确定该对氨基酸被由所述结构参数的当前值定义的距离隔开的概率,

其中所述参考距离图对于所述给定蛋白的氨基酸序列中的每一对氨基酸, 定义在该对氨基酸之间的可能距离范围内的相应概率分布, 所述相应概率分布基于所述氨基酸对中的氨基酸在给定蛋白的氨基酸序列中的位置、所述氨基酸对中的氨基酸的相对偏移或两者来确定; 和

基于在所述给定蛋白的氨基酸序列中的每一对氨基酸上, 根据在由所述参考距离图定义的该对氨基酸残基之间的可能距离范围内的相应概率分布的、该对氨基酸被由所述结构

参数的当前值定义的距离隔开的概率的乘积,确定所述质量得分。

3. 根据权利要求1所述的方法,其中:

所述一个或多个评分神经网络包括结构预测神经网络,所述结构预测神经网络被配置为处理包括所述给定蛋白的氨基酸序列的表示的输入,以生成输出,所述输出对于所述多个结构参数中的每一个,定义在所述结构参数的可能值上的概率分布;并且

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括,对于所述多个结构参数中的每一个:

使用在由所述结构预测神经网络生成的所述结构参数的可能值上的相应概率分布,确定所述结构参数的当前值的概率。

4. 根据权利要求3所述的方法,其中:

对于所述多个结构参数中的每一个,在所述结构参数的可能值上的概率分布是基于所述结构预测神经网络的输出而确定的参数概率分布。

5. 根据权利要求4所述的方法,其中所述参数概率分布是冯·米塞斯概率分布。

6. 根据权利要求1所述的方法,其中:

所述一个或多个评分神经网络包括几何神经网络,所述几何神经网络被配置为处理包括所述给定蛋白的氨基酸序列的表示和所述结构参数的当前值的输入以生成几何得分;

所述几何得分是由所述结构参数的当前值定义的所述给定蛋白的预测结构与所述给定蛋白的实际结构之间的相似性度量的估计;并且

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括:

基于由所述几何神经网络生成的几何得分确定所述质量得分。

7. 根据权利要求1所述的方法,其中确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括:

基于所述结构参数的当前值,确定表征所述结构参数的当前值的似然性的物理约束得分,所述物理约束得分基于所述结构参数的当前值与对所述给定蛋白的结构的物理约束的符合程度。

8. 根据权利要求1所述的方法,其中使用所述质量得分相对于所述结构参数的当前值的梯度来更新所述结构参数的当前值包括:

根据包括动量的梯度下降更新规则,使用所述质量得分相对于所述结构参数的当前值的梯度来更新所述结构参数的当前值。

9. 根据权利要求1所述的方法,其中获得定义所述预测结构的多个结构参数的初始值包括:

使用结构预测神经网络处理包括所述给定蛋白的氨基酸序列的表示的输入,以生成输出,所述输出对于所述多个结构参数中的每一个,定义在所述结构参数的可能值上的概率分布;和

基于在所述结构参数的可能值上的概率分布,对所述多个结构参数中的每个结构参数的初始值进行采样。

10. 根据权利要求1所述的方法,其中获得定义所述预测结构的多个结构参数的初始值包括:

获得对应于所述给定蛋白的先前生成的预测结构的多个结构参数的值;和

通过随机噪声值,对所获得的对应于所述给定蛋白的先前生成的预测结构的多个结构参数的值进行扰动。

11.根据权利要求1所述的方法,其中选择所述给定蛋白的所述多个预测结构中的一个预测结构作为所述给定蛋白的最终预测结构包括:

选择所述给定蛋白的所述多个预测结构中的具有最高质量得分的所述一个预测结构作为所述给定蛋白的最终预测结构。

12.根据权利要求1-11中的任一项所述的方法,其中所述多个结构参数包括多个扭转角。

13.根据权利要求1-11中的任一项所述的方法,其中所述多个结构参数包括多个原子坐标。

14.一种由一个或多个数据处理装置执行的用于确定给定蛋白的预测结构的方法,其中所述给定蛋白包括氨基酸序列,其中所述给定蛋白的预测结构由多个结构参数的值定义,所述方法包括:

获得定义所述预测结构的多个结构参数的初始值;

更新所述多个结构参数的初始值,包括,在多个更新迭代中的每一个处:

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分,其中所述质量得分基于一个或多个评分神经网络的相应输出,每个评分神经网络被配置为处理:(i)所述结构参数的当前值,(ii)所述给定蛋白的氨基酸序列的表示,或(iii)两者;

对于所述多个结构参数中的一个或多个:

确定所述质量得分相对于所述结构参数的当前值的梯度;以及

使用所述质量得分相对于所述结构参数的当前值的梯度来更新所述结构参数的当前值;

在所述多个更新迭代中的最终更新迭代之后,将所述给定蛋白的预测结构确定为由所述多个结构参数的当前值定义,其中:

所述一个或多个评分神经网络包括距离预测神经网络,所述距离预测神经网络被配置为处理包括所述给定蛋白的氨基酸序列的表示的输入以生成所述给定蛋白的距离图;

所述距离图对于所述给定蛋白的氨基酸序列中的多对氨基酸中的每一对,定义在该对氨基酸之间的可能距离范围内的相应概率分布;和

确定表征由所述结构参数的当前值定义的所述预测结构的的质量的质量得分包括,对于所述给定蛋白的氨基酸序列中的每一对氨基酸:

使用在由所述距离图定义的该对氨基酸之间的可能距离范围内的相应概率分布,确定该对氨基酸被由所述结构参数的当前值定义的距离隔开的概率,以及

基于在所述给定蛋白的氨基酸序列中的每一对氨基酸上,根据在由所述距离图定义的该对氨基酸残基之间的可能距离范围内的相应概率分布的、该对氨基酸被由所述结构参数的当前值定义的距离隔开的概率的乘积,确定所述质量得分。

15.一种获得配体的方法,其中所述配体是药物或工业酶的配体,所述方法包括:

获得靶氨基酸序列,其中所述靶氨基酸序列是靶蛋白的氨基酸序列;

将所述靶氨基酸序列用作所述氨基酸序列或氨基酸残基序列来执行根据权利要求1-14中的任一项所述的方法,以确定所述靶蛋白的结构,其中所述靶蛋白的结构是所述最终

预测结构;

评估一种或多种候选配体与所述靶蛋白的结构的作用;以及
取决于所述评估的结果,将一种或多种候选配体选择为所述配体。

16. 根据权利要求15所述的方法,其中所述靶蛋白包括受体或酶,并且其中所述配体是所述受体或酶的激动剂或拮抗剂。

17. 根据权利要求15所述的方法,其中评估所述候选配体中的一种的相互作用包括确定所述候选配体的相互作用得分,其中所述相互作用得分包括所述候选配体与所述靶蛋白之间的相互作用的度量。

18. 根据权利要求15-17中的任一项所述的方法,所述方法进一步包括合成所述配体。

19. 根据权利要求18所述的方法,所述方法进一步包括在体外和体内测试所述配体的生物学活性。

20. 一种获得多肽配体的方法,其中所述配体是药物或工业酶的配体,所述方法包括:
获得一种或多种候选多肽配体的氨基酸序列;

对于所述一种或多种候选多肽配体中的每一种,将所述候选多肽配体的氨基酸序列用作所述氨基酸序列或氨基酸残基序列来执行根据权利要求1-14中的任一项所述的方法,以确定所述候选多肽配体的结构,其中所述候选多肽配体的结构是所述最终预测结构;

获得靶蛋白的靶蛋白结构;

评估所述一种或多种候选多肽配体中的每一种的结构与所述靶蛋白结构之间的相互作用;以及

取决于所述评估的结果,将所述一种或多种候选多肽配体中的一种选择为所述多肽配体。

21. 根据权利要求20所述的方法,其中评估所述候选配体中的一种的相互作用包括确定所述候选配体的相互作用得分,其中所述相互作用得分包括所述候选配体与所述靶蛋白之间的相互作用的度量。

22. 根据权利要求20-21中的任一项所述的方法,所述方法进一步包括合成所述配体。

23. 根据权利要求22所述的方法,所述方法进一步包括在体外和体内测试所述配体的生物学活性。

24. 一种用于鉴定蛋白错折叠疾病的存在计算机系统,所述计算机系统包括:

存储设备;

通信地耦合到所述存储设备的处理器,其中,所述处理器执行存储在所述存储设备中的指令来使得所述计算机系统:

获得蛋白的氨基酸序列;

将所述蛋白的氨基酸序列用作所述氨基酸序列或氨基酸残基序列来执行根据权利要求1-14中的任一项所述的方法,以确定所述蛋白的结构,其中所述蛋白的结构是所述最终预测结构;

获得从人体或动物体获得的蛋白版本的结构;

将所述蛋白的结构与从人体或动物体获得的蛋白版本的结构进行比较;以及

取决于所述比较的结果,鉴定蛋白错折叠疾病的存在。

25. 一种系统,所述系统包括一个或多个计算机以及一个或多个存储设备,所述存储设

备存储指令,所述指令当由所述一个或多个计算机执行时,使得所述一个或多个计算机执行根据权利要求1-17和20-21中的任一项所述的方法。

26.一种或多种计算机可读存储介质,所述计算机存储介质存储指令,所述指令当由一个或多个计算机执行时,使得所述一个或多个计算机执行根据权利要求1-17和20-21中的任一项所述的方法。

使用质量得分的梯度的迭代蛋白结构预测

背景技术

[0001] 本说明书涉及确定蛋白结构。

[0002] 蛋白由氨基酸序列组成。氨基酸是一种有机化合物,包括氨基官能团和羧基官能团,以及对该氨基酸特定的侧链(即原子团)。蛋白折叠是指氨基酸序列折叠成三维构型的物理过程。如本文所使用的,蛋白的结构定义了蛋白经历蛋白折叠后,蛋白的氨基酸序列中的原子的三维构型。当在通过肽键连接的序列中时,氨基酸可以被称为氨基酸残基。

[0003] 可以使用机器学习模型进行预测。机器学习模型接收输入,并基于所接收的输入生成输出,例如预测输出。一些机器学习模型是参数模型,并且基于所接收的输入和模型的参数的值生成输出。可以通过由其氨基酸序列预测结构来确定蛋白的结构。

[0004] 一些机器学习模型是深度模型,其采用模型的多个层来为所接收的输入生成输出。例如,深度神经网络是一种深度机器学习模型,其包括输出层和一个或多个隐藏层,每个隐藏层将非线性变换应用于所接收的输入以生成输出。

发明内容

[0005] 本说明书描述了在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统,该系统执行蛋白三级结构预测和蛋白结构域分割。描述了许多技术。这些技术可以组合或孤立地使用。

[0006] 在第一方面,描述了一种由一个或多个数据处理装置执行的用于确定给定蛋白的最终预测结构的方法。给定蛋白包括氨基酸序列,并且给定蛋白的预测结构由多个结构参数的值定义。生成给定蛋白的预测结构可以包括获得定义预测结构的多个结构参数的初始值,以及更新多个结构参数的初始值。更新可以包括,在多个更新迭代中的每一个处:确定得分,例如,表征由结构参数的当前值定义的预测结构的的质量的质量得分。质量得分可以表示预测的结构正确性程度和/或预测的结构似然性有多大,例如,质量得分可以表征蛋白的预测结构与实际结构之间的估计相似性和/或预测结构的似然性。质量得分可以基于一个或多个评分神经网络的相应输出,每个评分神经网络被配置为处理:(i) 结构参数的当前值,或(ii) 给定蛋白的氨基酸序列的表示,或(iii) 两者。

[0007] 该方法可以进一步包括,对于多个结构参数中的一个或多个:确定质量得分相对于结构参数的当前值的梯度;以及使用质量得分相对于结构参数的当前值的梯度来更新结构参数的当前值。因此,该方法的一些实施方式可以将基于得分的优化系统用于结构预测。

[0008] 该方法可以进一步包括,在多个更新迭代中的最终更新迭代之后,将给定蛋白的预测结构确定为由多个结构参数的当前值定义。

[0009] 该方法可以包括使用上述方法生成给定蛋白的多个预测结构。然后,该方法可以进一步包括将给定蛋白的特定预测结构选择为给定蛋白的最终预测结构。

[0010] 结构参数是定义蛋白结构的参数。它们可以包括主链扭转角(二面角 φ , ψ) 集合和/或可以包括蛋白的一些或全部原子(例如,碳原子,例如 α 或 β 碳原子)的(3D)原子坐标。

[0011] 在实施方式中,这种方法通过优化质量得分,在实施方式中通过梯度下降,有助于

高度准确地预测给定蛋白的结构。质量得分可以被视为将通过梯度下降而被最小化的“势能”。

[0012] 在一些实施方式中,一个或多个评分神经网络包括距离预测神经网络,该距离预测神经网络被配置为处理氨基酸序列的表示以生成给定蛋白的距离图。在实施方式中,距离图对于序列中的多对氨基酸中的每一对,定义了在该对氨基酸之间的可能距离范围内的相应概率分布。例如,可以量化可能距离范围,或者可以由参数化的概率分布来表示在可能距离范围内的概率分布。可以由氨基酸(残基)的特定的相应原子(诸如 α 和/或 β 碳原子)之间的距离来定义该对氨基酸之间的范围。

[0013] 然后,该方法可以进一步包括通过对于每一对氨基酸,使用在由距离图定义的该对氨基酸之间的可能距离范围内的相应概率分布,确定氨基酸被由结构参数的当前值定义的距离隔开的概率来确定质量得分。

[0014] 在实施方式中,预测距离有助于汇聚成准确的预测结构。距离图共同预测了许多距离,并促进了将关于协变、局部结构以及氨基酸残基同一性的距离信息传播到附近残基的方法。更具体地,预测距离概率分布还通过对预测中的不确定性进行建模,进一步促进了这一点。

[0015] 在一些实施方式中,质量得分取决于在序列中的每一对氨基酸上,根据在由距离图定义的可能距离范围内的相应概率分布,氨基酸被由结构参数的当前值定义的距离隔开的概率的乘积(即质量得分可能取决于这些概率的乘积)。

[0016] 确定质量得分可以进一步包括,对于每一对氨基酸,使用在由参考距离图定义的该对氨基酸之间的可能距离范围内的相应概率分布,确定氨基酸被由结构参数的当前值定义的距离隔开的概率。参考距离图可以基于氨基酸对中的氨基酸在给定蛋白的氨基酸序列中的位置、氨基酸对中的氨基酸的相对偏移或两者来定义概率分布;但是在实施方式中,不以氨基酸序列为条件,尽管任选地以序列的长度为条件。该方法可以进一步包括基于在给定蛋白的氨基酸序列中的每一对氨基酸上,根据由参考距离图定义的可能距离范围内的相应概率分布,氨基酸被由结构参数的当前值定义的距离隔开的概率的乘积来确定质量得分。例如,可以使用该乘积,对先前距离分布的过度表示校正质量得分,例如,通过从质量得分的对数中减去该乘积的对数(或等同地,概率的对数的总和)。

[0017] 在实施方式中,评分神经网络可以包括结构预测神经网络,以处理氨基酸序列(的表示)并对于多个结构参数的每一个,生成在该结构参数的可能值上的概率分布。然后,确定质量得分可以包括,对于多个结构参数中的每一个,使用相应概率分布来确定结构参数的当前值的概率。这样的质量得分可以表示结构参数的当前值的似然性;同样地,使用概率分布对此建模可以通过对结构预测的不确定性进行建模来帮助提高准确性。

[0018] 在一些实施方式中,结构参数由离散范围限定,在这种情况下,将在结构参数的可能值上的概率分布表示为参数概率分布以提供平滑、可微分的分布可能是有利的。这有助于确定质量得分相对于结构参数值的梯度。参数概率分布可以是冯·米塞斯(von Mises)(或圆法线)概率分布,这在结构参数可以包括主链扭转角集合的情况下很方便。

[0019] 可以将以此方式确定的质量得分与从距离图得出的质量得分相组合,例如,通过对(负)对数似然性求和,使得质量得分表示可以例如通过梯度下降而被最小化的组合的、可微分的“势能”。结构预测神经网络的输出和距离预测神经网络的输出可以包括公共神经

网络上的单独头。任选地,结构预测神经网络和距离预测神经网络之一或两者的输入可以包括从序列的MSA(多序列比对)中得出的一个或多个特征。

[0020] 在实施方式中,评分神经网络可以包括几何神经网络,以处理氨基酸序列(的表示)并且生成几何得分,该几何得分表示由结构参数的当前值定义的预测结构与给定蛋白的实际结构之间的相似性度量的估计。然后,质量得分可以全部或部分地基于几何得分。

[0021] 确定质量得分可以进一步包括基于结构参数的当前值,确定表征结构参数的当前值的似然性的物理或物理约束得分,该物理或物理约束得分取决于结构参数的当前值与对给定蛋白结构的生化或物理约束的符合程度。例如,可以通过范德华项对结构上的空间约束进行建模。

[0022] 在例如通过梯度下降进行优化之前,可以通过使用结构预测神经网络处理氨基酸序列并从每个结构参数的概率分布进行采样来获得结构参数的初始值。如果先前已经预测了给定蛋白预测结构的结构,则可以通过例如通过随机噪声值扰动这些参数来获得结构参数的初始值。

[0023] 在另一方面,描述了一种用于生成给定蛋白的距离图的计算机实施方法。给定蛋白的(3D)结构由排列在该结构中的氨基酸(更具体地说是氨基酸残基)序列定义,并且距离图表征了结构中氨基酸残基之间的估计距离。

[0024] 该方法可以包括生成多个距离图裁剪(crop),每个距离图裁剪表征蛋白的结构中在(i)序列中的一个或多个相应第一位置的每一个中的氨基酸残基与(ii)序列中的一个或多个相应第二位置的每一个中的氨基酸残基之间的估计距离。生成距离图裁剪可以包括鉴定序列中的一个或多个第一位置和序列中的一个或多个第二位置;第一位置可以是序列的适当子集。生成距离图裁剪可以进一步包括由序列中的第一位置中的氨基酸残基和序列中的第二位置中的氨基酸残基确定网络输入。生成距离图裁剪可以进一步包括将网络输入提供给距离预测神经网络,其被配置为根据距离预测神经网络权重的当前值来处理网络输入以生成包括距离图裁剪的网络输出。然后可以使用所述多个距离图裁剪生成给定蛋白的距离图。

[0025] 在实施方式中,使用裁剪可以大大减少存储器(memory)和处理需求。这也可以促进将更复杂的架构用于距离预测神经网络,这继而允许更精确的表示,以及任选地,允许辅助特征的预测(以及任选地使用这些特征进行训练),如下所述。另外,使用裁剪促进了工人在其中生成距离图裁剪的分布式处理,并且在训练期间,促进了示例的批处理。

[0026] 在一些实施方式中,距离图/裁剪使用二进制值的距离估计(例如,定义接触/非接触)来定义一对氨基酸残基之间的距离;在其他实施方式中,距离图/裁剪定义连续值的距离估计;在其他实施方式中,距离图/裁剪使用距离范围概率分布,即在一对氨基酸之间的可能距离范围内的概率分布,定义该对氨基酸残基之间的距离。在后一种情况下,如前所述,可以对可能距离范围进行量化,或者可以通过参数化的概率分布来表示在可能距离范围内的概率分布。该对氨基酸之间的距离或距离范围可以例如由氨基酸残基的特定相应原子(例如 α 和/或 β 碳原子)之间的距离来定义。

[0027] 在实施方式中,距离图裁剪生成重叠预测。它们可以通过求平均来组合,这可以提高重叠区域的准确性,和/或它们可以利用后续的融合神经网络进行组合。融合神经网络的输出可以具有包括由距离图裁剪覆盖的完整区域的感受野,并且可以被配置为处理具有不

同偏移的输入。

[0028] 在一些实施方式中,鉴定序列中的一个或多个第一位置和序列中的一个或多个第二位置可以包括将第一位置随机采样为第一预定长度的连续位置的第一序列,和/或将第二位置随机采样为第二预定长度的连续位置的第二序列。因此,裁剪可以对应于连续残基的组,从而对结构的(长程)区域之间的距离进行建模。

[0029] 在一些实施方式中,距离预测神经网络包括一个或多个扩张卷积神经网络层、一个或多个残差块(residual block),以及任选地一个或多个关注层。这有利于使用具有大感受野的深层神经网络,从而改善了预测。

[0030] 确定网络输入可以包括提取下述的分量:(i)氨基酸残基序列的表示,和(ii)由包括氨基酸残基序列的多序列比对(MSA)得出的比对特征。比对特征可以包括(MSA中的序列之间的)协变特征,其有助于鉴定接触的残基。

[0031] 距离预测神经网络可以具有辅助输出,该辅助输出表征序列中的第一和第二位置中的氨基酸残基的二级结构,和/或表征残基的扭转角。对这类辅助输出进行训练可以帮助提高距离图裁剪的准确性,并且这些输出本身就很有用。

[0032] 距离图可以被用于确定给定蛋白的预测结构。例如,这可能涉及获得定义蛋白结构的结构参数的初始值,并且基于由距离图定义的结构的质量得分来更新这些初始值。更新可以包括,对于一个或多个或每个结构参数:通过调整结构参数的当前值,例如通过确定质量得分相对于结构参数的当前值的梯度,然后使用质量得分的梯度更新结构参数的当前值,来优化质量得分;或通过其他优化过程。给定蛋白的预测结构可以由最终更新迭代后的结构参数值来定义。如前所述,可以确定质量得分的任选的其他分量。

[0033] 在另一方面,描述了一种方法,该方法包括获得数据,所述数据定义:(i)给定蛋白中的氨基酸序列,和(ii)给定蛋白的预测结构,其中给定蛋白的预测结构由多个结构参数的值定义;由给定蛋白的氨基酸残基序列确定网络输入;根据距离预测神经网络权重的当前值,使用距离预测神经网络处理网络输入以生成给定蛋白的距离图,其中该距离图对于给定蛋白中的氨基酸残基序列中的多对氨基酸残基中的每一对,定义在给定蛋白的结构中的该对氨基酸残基之间的可能距离范围内的相应概率分布;并且使用由距离图定义的概率分布来确定表征给定蛋白的预测结构的质量的得分。

[0034] 如前所述,使用在可能距离范围内的概率分布可以显著提高得分表征预测结构的质量的准确性,因此可以显著提高使用该得分确定的蛋白结构的准确性。

[0035] 该得分可以被用于确定给定蛋白的预测结构。例如,这可能涉及获得定义预测结构的结构参数的初始值,并且基于得分来更新它们。更新可以包括,对于多个结构参数中的一个或多个或每一个:通过调整结构参数的当前值,例如通过确定该得分相对于结构参数的当前值的梯度,并且使用该梯度来更新结构参数的当前值,来优化得分;或通过其他优化过程。给定蛋白的预测结构可以由最终更新迭代之后的多个结构参数的值来定义。如前所述,可以确定得分的任选的其他分量。

[0036] 通常,该方法的其他特征可以如前所述。例如,网络输入可以由下述信息确定:(i)氨基酸残基序列的表示;(ii)由包括氨基酸残基序列的多序列比对得出的比对特征,例如,定义所述多序列比对中不同于给定蛋白的一种或多种蛋白的氨基酸残基序列的数据。比对特征包括多序列比对的二阶统计量,例如残基对之间的相关性或协变。

[0037] 在另一方面,描述了一种计算机实施方法,该方法包括,在一个或多个迭代中的每一个迭代处,维护数据,该数据包括:(i)由多个结构参数的当前值定义的给定蛋白的当前预测结构,以及(ii)基于(即取决于)当前几何得分的表征当前预测结构的的质量的质量得分,所述当前几何得分是给定蛋白的当前预测结构与实际结构之间的相似性度量的估计。该方法可以进一步包括,在一个或多个迭代处,基于当前预测结构确定给定蛋白的替代预测结构,其中该替代预测结构由结构参数的替代值定义。该方法可以进一步包括,在一个或多个迭代处,使用几何神经网络并根据几何神经网络权重的当前值,处理网络输入以生成表征替代几何得分的输出,该网络输入包括:(i)给定蛋白中的氨基酸残基序列的表示,以及(ii)结构参数的替代值,该替代几何得分是给定蛋白的替代预测结构与实际结构之间的相似性度量的估计。该方法可以进一步包括,在一个或多个迭代处,基于替代几何得分确定表征替代预测结构的的质量的质量得分。该方法可以进一步包括,在一个或多个迭代处,使用表征当前预测结构的的质量的质量得分和表征替代预测结构的的质量的质量得分,确定是否将当前预测结构更新为替代预测结构。

[0038] 该方法的一些示例适合于由使用一个或多个搜索计算单元的结构预测系统来实施。例如,可以在多个搜索计算单元的每一个上实施确定替代预测结构、使用几何神经网络、确定质量得分以及确定是否更新的过程。所维护的数据可以是本地的和/或共享的,例如,每个搜索计算单元可以将具有高质量得分的预测折叠结构存储在共享存储器中。因此,搜索计算单元可以基于使用质量得分的模拟退火来实施结构优化系统。

[0039] 在一些实施方式中,该方法获得由结构参数的子集的值定义(对应于结构参数的子集的值)的结构片段,并且使用当前预测结构的一部分和结构片段来生成替代预测结构。可以使用生成神经网络和/或从不同蛋白的实际折叠结构和/或通过来自先前迭代的预测折叠结构进行分段来获得结构片段。使用生成神经网络是有利的,因为它可以生成许多不同的结构片段,这有助于探索搜索空间,并因此可以更快地生成更准确的结构。

[0040] 几何得分估计的相似性度量可以是蛋白结构之间的任何相似性度量例如全局距离测试(GDT)度量(基于 α 碳原子)或均方根偏差(RMSD)度量(结构参数的当前值和与替代值之间的相似性的度量)或一些其他度量。

[0041] 质量得分可以取决于几何得分和估计未来迭代中的预测结构的的质量的值得分的组合,例如加权组合。值得分可以从值神经网络得出,该值神经网络被配置为处理给定蛋白的氨基酸序列的表示和结构参数的当前值。这可以帮助该方法将短期几何得分不足换成长期总体利益。

[0042] 通常,该方法的其他特征以及质量得分的任选的其他分量可以包括先前描述的那些特征。

[0043] 该方法可以被用于确定给定蛋白的预测结构。例如,这可能涉及获得定义预测结构的结构参数的初始值,并且基于质量得分来更新这些值。例如,更新可以包括,在多个更新迭代中的每一个迭代处,响应于确定是否将当前预测结构更新为替代预测结构来更新结构参数的当前值。给定蛋白的预测结构可以由最终更新迭代后的结构参数的值来定义。

[0044] 在另一方面,描述了一种计算机实施方法,该方法包括:接收定义蛋白的氨基酸残基序列和由多个结构参数的值定义的蛋白的预测结构的数据,以及使用几何神经网络并且根据几何神经网络权重的当前值来处理该数据,以生成表征几何得分的输出,其中几何得

分是蛋白的预测结构与蛋白的实际结构之间的相似性度量的估计。该方法的其他特征可以包括先前描述的那些特征。例如,几何神经网络的输入可以包括MSA得出的比对特征。

[0045] 该方法可用于确定包括氨基酸残基序列的给定蛋白的预测结构。例如,这可能涉及获得定义预测结构的结构参数的初始值,并且基于几何得分更新这些值。更新可以包括,对于多个结构参数中的一个或多个或每一个:通过调整结构参数的当前值,例如通过确定几何得分相对于结构参数的当前值的梯度,并且使用该梯度来更新结构参数的当前值,来优化几何得分;或通过其他优化过程。给定蛋白的预测结构可以由最终更新后的多个结构参数的值来定义。可以如先前所述确定得分的任意的其他分量。

[0046] 在另一方面,描述了一种计算机实施方法,所述方法包括:接收数据,所述数据定义:(i)蛋白的氨基酸残基序列、(ii)由多个结构参数的第一值定义的蛋白的第一预测结构,以及(iii)由多个结构参数的第二值定义的蛋白的第二预测结构。该方法可以进一步包括使用几何神经网络并且根据几何神经网络权重的当前值来处理所接收的数据,以生成表征相对几何得分的输出。相对几何得分定义了关于蛋白的第一预测结构与蛋白的实际结构之间的相似性度量是否超过蛋白的第二预测结构与蛋白的实际结构之间的相似性度量的预测。该方法的其他特征可以包括先前描述的那些特征,例如,几何神经网络的输入可以包括MSA得出的比对特征。

[0047] 该方法可以用于确定包括氨基酸残基序列的给定蛋白的预测结构。例如,这可能涉及获得定义预测结构的结构参数的初始值并对其进行更新。更新可以包括,在多个更新迭代中的每一个处:基于当前预测结构,确定由结构参数的替代值定义的给定蛋白的替代预测结构;确定关于结构参数的当前值和替代值的相对几何得分;使用相对几何得分来确定是否将当前预测结构更新为替代预测结构;以及将给定蛋白的预测结构确定为由最终更新迭代后的结构参数的值定义的结构。任选地,相对几何得分可以与其他得分分量组合,如前所述。

[0048] 根据另一方面,提供了一种系统,其包括被配置为存储定义给定蛋白的预测结构集合的数据的中央存储器,其中每个结构由结构参数集合的相应值限定。该系统进一步包括一个或多个搜索计算单元,其中所述一个或多个搜索计算单元中的每一个:(i)维护定义给定蛋白的相应当前预测结构的数据,并且(ii)包括被配置为存储结构片段集合的相应本地存储器。每个结构片段由多个结构参数的相应子集的相应值定义。所述一个或多个搜索计算单元中的每一个被配置为执行操作,包括,在一个或多个搜索迭代中的每一个处:使用存储在搜索计算单元的相应本地存储器中的结构片段来更新由搜索计算单元所维护的数据定义的相应当前预测结构;确定是否满足中央存储器更新条件;如果满足中央存储器更新条件,则将相应当前预测结构存储在中央存储器中;确定是否满足本地存储器更新条件;如果满足本地存储器更新条件,则更新搜索计算单元的相应本地存储器,包括:(i)选择存储在中央存储器中的预测结构,(ii)由所选择的预测结构中确定一个或多个结构片段,以及(iii)将所确定的结构片段存储在搜索计算单元的相应本地存储器中。

[0049] 在一些实施方式中,每个结构片段由定义给定蛋白中的氨基酸残基的连续序列的结构的结构参数集合的相应子集的相应值定义。

[0050] 在一些实施方式中,使用存储在搜索计算单元的相应本地存储器中的结构片段来更新由搜索计算单元所维护的数据定义的相应当前预测结构包括更新相应当前预测结构

以包括：(i) 一部分当前预测结构；以及(ii) 结构片段。

[0051] 在一些实施方式中，确定是否满足中央存储器更新条件包括确定表征当前预测结构的的质量的当前质量得分。如果当前质量得分高于先前由搜索计算对给定蛋白的先前预测结构确定的任何质量得分，则确定满足中央存储器更新条件。

[0052] 在一些实施方式中，确定是否满足本地存储器更新条件包括：确定搜索计算单元是否已经执行了预定数量的搜索迭代；以及如果搜索计算单元已经执行了预定数量的搜索迭代，则确定满足本地存储器更新条件。

[0053] 在一些实施方式中，选择存储在中央存储器中的预测结构包括基于全原子质量得分来选择存储在中央存储器中的预测结构，全原子质量得分基于包括在给定蛋白中的每个原子，表征当前预测结构的质量。每个搜索计算单元被配置为基于主链原子质量得分，更新由搜索计算单元所维护的数据定义的相应当前预测结构，该主链原子质量得分基于包括在给定蛋白中的主链原子，表征当前预测结构的质量。

[0054] 在一些实施方式中，由所选择的预测结构确定一个或多个结构片段包括确定将所选择的预测结构分成结构片段集合，其中每个结构片段定义了给定蛋白中的氨基酸残基的连续序列的结构。

[0055] 在一些实施方式中，该系统进一步通过由生成神经网络生成的网络输出定义的一个或多个结构片段来更新搜索计算单元的相应本地存储器。

[0056] 在一些实施方式中，该系统进一步包括子系统，该子系统被配置为执行操作，该操作包括基于存储在中央存储器中的每个预测结构的全原子质量得分，将存储在中央存储器中的预测结构作选择为最终预测结构，其中预测结构的全原子质量得分是基于包括在给定蛋白中的每个原子。

[0057] 在一些实施方式中，结构参数集合包括主链原子扭转角集合。

[0058] 在一些实施方式中，结构参数集合包括主链原子坐标集合。

[0059] 根据另一方面，提供了一种方法，该方法包括，在一个或多个迭代中的每一个迭代处，维护数据，该数据包括：(i) 由多个结构参数的当前值定义的给定蛋白的当前预测结构，以及(ii) 表征当前预测结构的的质量的质量得分。确定给定蛋白的替代预测结构集合，其中每个替代预测结构由结构参数集合的相应替代值定义，包括使用生成神经网络并且根据生成神经网络权重的当前值来处理包括给定蛋白中的氨基酸残基序列的表示的网络输入，以生成定义结构片段集合的网络输出。每个结构片段由结构参数集合的相应子集的相应值定义。对于每个结构片段，生成相应替代预测结构，包括(i) 当前预测结构的一部分和(ii) 结构片段。确定表征每一个替代预测结构的的质量的相应质量得分。该方法包括基于表征当前预测结构的质量得分和表征替代预测结构的质量得分，确定是否将当前预测结构更新为任一替代预测结构。

[0060] 在一些实施方式中，表征预测结构的的质量的质量得分是基于根据距离图的预测结构的似然性。距离图表征给定蛋白的实际结构中氨基酸残基之间的估计距离。

[0061] 在一些实施方式中，表征预测结构的的质量的质量得分是基于给定蛋白的预测结构与实际结构之间的相似性度量的估计。

[0062] 在一些实施方式中，如果当前预测结构是给定预测结构，则表征给定预测结构的的质量的质量得分是基于表征在未来的迭代中的未来预测结构的的质量的质量得分的估计。

[0063] 在一些实施方式中,网络输入包括由多序列比对得出的比对特征,其中多序列比对包括给定蛋白中的氨基酸残基序列。

[0064] 在一些实施方式中,比对特征被表示为二维数据,并且使用生成神经网络处理网络输入包括:通过一个或多个卷积神经网络层来处理比对特征;以及使用池化层(pooling layer)来处理卷积层的输出。

[0065] 在一些实施方式中,使用生成神经网络来处理包括给定蛋白中的氨基酸残基序列的表示的网络输入以生成定义结构片段集合的网络输出包括:处理网络输入以对于多个结构参数中的多结构参数的每一个,生成定义在结构参数的可能值上的相应概率分布的数据。通过从在结构参数的可能值上的相应概率分布中采样结构参数的相应值来确定多个结构参数中的每个结构参数的相应值。

[0066] 在一些实施方式中,处理网络输入以生成定义在特定结构参数的可能值上的相应概率分布的数据包括处理给定蛋白中的氨基酸残基序列的表示和定义为一个或多个在前结构参数确定的相应值的数据,以生成定义特定结构参数可能值上的相应概率分布的数据,所述一个或多个在前结构参数在结构参数的排序中均在该特定结构参数之前。

[0067] 在一些实施方式中,生成神经网络的架构由WaveNet神经网络的架构得出。

[0068] 在一些实施方式中,处理网络输入以对于多个结构参数的多结构参数中的每一个生成定义在结构参数的可能值上的相应概率分布的数据包括:根据一个或多个潜变量的先验概率分布,从潜变量空间中采样一个或多个潜变量。处理给定蛋白中的氨基酸残基序列的表示以及一个或多个采样的潜变量,以对于多个结构参数的多结构参数中的每一个生成定义结构参数的可能值的相应概率分布的数据。

[0069] 在一些实施方式中,生成神经网络的架构由变分自编码器的架构得出。

[0070] 在一些实施方式中,生成神经网络的架构由DRAW神经网络的架构得出。

[0071] 在一些实施方式中,在包括实际蛋白结构集合的训练数据集合上对生成神经网络进行训练。

[0072] 在一些实施方式中,每个结构片段由定义给定蛋白中的氨基酸残基的连续序列的结构的结构参数集合的相应子集的相应值定义。

[0073] 在一些实施方式中,确定是否将当前预测结构更新为任一替代预测结构包括:如果给定替代预测结构具有最高质量得分,则将当前预测结构更新为给定替代预测结构。

[0074] 在另一方面,提供了一种方法,该方法包括接收定义蛋白的氨基酸残基序列的数据。使用生成神经网络处理包括定义蛋白的氨基酸残基序列的数据的输入,以对于表征蛋白的氨基酸残基序列的结构的结构参数集合中的每一个,生成定义在结构参数的可能值上的相应概率分布的数据。确定蛋白的氨基酸残基序列的预测结构,其中:由多个结构参数的预测值定义预测结构;并且所述确定包括,对于所述多个结构参数中的每一个,根据在结构参数的可能值上的相应概率分布,对所述结构参数的预测值进行采样。

[0075] 在一些实施方式中,对于每个结构参数,定义在结构参数的可能值上的相应概率分布的数据包括冯·米塞斯概率分布的混合的参数的相应值。

[0076] 根据另一方面,提供了一种方法,该方法包括,在一个或多个迭代中的每一个处,维护数据,该数据包括:(i)由多个结构参数的当前值定义的给定蛋白的当前预测结构,以及(ii)基于当前值得分来表征当前预测结构的的质量的质量得分,该当前值得分是在给定当

前预测结构的情况下,表征在未来迭代中的未来预测结构的的质量的质量得分的估计。基于当前预测结构确定给定蛋白的替代预测结构,其中由结构参数的替代值定义替代预测结构。该方法包括使用值神经网络并且根据当前值神经网络权重来处理网络输入以生成表征替代值得分的输出,该网络输入包括:(i) 给定蛋白中的氨基酸残基序列的表示,以及(ii) 结构参数的替代值,该替代值得分是如果在当前时间步,将当前预测结构更新为替代预测结构,表征未来时间步的未来预测结构的的质量的质量得分的估计。基于替代值得分来确定表征替代预测结构的的质量的质量得分。该方法包括使用表征当前预测结构的的质量的质量得分和表征替代预测结构的的质量的质量得分来确定是否将当前预测结构更新为替代预测结构。

[0077] 在一些实施方式中,值神经网络包括一个或多个二维残差卷积块、一个或多个关注层或两者。

[0078] 在一些实施方式中,网络输入包括距离图,该距离图表征给定蛋白的实际结构中的氨基酸对之间的估计距离。

[0079] 在一些实施方式中,值神经网络的输出包括在可能值得分的预定集合上的概率分布。

[0080] 在一些实施方式中,在训练示例集合上使用机器学习训练技术来训练值神经网络。每个训练示例包括:(i) 蛋白的训练预测结构,和(ii) 目标值得分,该目标值得分是通过重复地更新蛋白的训练预测结构而确定的表征蛋白的未来预测结构的的质量的质量得分。

[0081] 在一些实施方式中,重复地更新蛋白的训练预测结构包括:使用基于根据值神经网络权重的当前值由值神经网络生成的值得分的质量得分来重复地更新蛋白的训练预测结构。

[0082] 在一些实施方式中,使用对比发散训练程序来训练值神经网络。

[0083] 根据另一方面,提供了一种方法,该方法包括接收数据,所述数据定义:(i) 蛋白的氨基酸残基序列,以及(ii) 由多个结构参数的值定义的蛋白的特定预测结构。使用值神经网络并根据值神经网络权重的当前值来处理包括定义蛋白的氨基酸残基序列的数据和定义蛋白的特定预测结构的数据的输入,以生成表征值得分的输出。值得分是表征蛋白的未来预测结构的的质量的质量得分的估计,其中通过使用结构修饰程序,在一个或多个时间步上迭代地修饰蛋白的特定预测结构来确定蛋白的未来预测结构。

[0084] 在一些实施方式中,值神经网络包括一个或多个二维残差卷积块、一个或多个关注层或两者。

[0085] 在一些实施方式中,由值神经网络处理的输入包括由多序列比对得出的比对特征,其中多序列比对包括蛋白的氨基酸残基序列。

[0086] 在一些实施方式中,比对特征包括多序列比对的二阶统计量。

[0087] 在一些实施方式中,由值神经网络处理的输入包括距离图,该距离图表征蛋白的实际结构中的氨基酸对之间的估计距离。

[0088] 在一些实施方式中,值神经网络的输出包括定义在可能值得分的预定集合上的概率分布的数据。

[0089] 在一些实施方式中,结构修饰程序包括,在多个迭代中的每一个处,通过用结构片段替代蛋白的当前预测结构的一部分来修饰蛋白的当前预测结构,其中结构片段由多个结

构参数的子集的值定义。

[0090] 根据另一方面,提供了一种方法,该方法包括接收数据,所述数据定义:(i)蛋白的氨基酸残基序列、(ii)由结构参数集合的第一值定义的蛋白的第一预测结构,以及(iii)由该结构参数集合的第二值定义的蛋白的第二预测结构。该方法包括使用值神经网络并根据值神经网络权重的当前值来处理输入以生成表征相对值得分的输出,该输入包括:(i)定义蛋白的氨基酸残基序列的数据、(ii)定义蛋白的第一预测结构的数据,以及(iii)定义蛋白的第二预测结构的数据。相对值得分定义了关于表征蛋白的第一未来预测结构的的质量的质量得分是否超过表征蛋白的第二未来预测结构的的质量的质量得分的预测。通过使用结构修饰程序,在一个或多个迭代中迭代地修饰蛋白的第一预测结构来确定蛋白的第一未来预测结构。通过使用结构修饰程序,在一个或多个迭代中迭代地修饰蛋白的第二预测结构来确定蛋白的第二未来预测结构。

[0091] 根据另一方面,提供了一种用于确定蛋白的结构域分割的方法,其中结构域分割定义了将蛋白的氨基酸序列分成多个结构域。每个结构域定义了蛋白的氨基酸序列的氨基酸子序列。该方法包括获得蛋白的候选结构域分割集合,其中每个候选结构域分割定义相应候选结构域集合。为每个候选结构域分割确定相应结构域分割得分,包括对于每个候选结构域分割,确定用于由候选结构域分割定义的每个候选结构域的相应结构域得分,包括对于每个候选结构域:获得表征蛋白结构中由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据;以及基于表征蛋白结构中由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据,确定候选结构域的结构域得分;从由候选结构域分割定义的每个候选结构域的相应结构域得分中确定候选结构域分割的结构域分割得分。基于候选结构域分割的相应结构域分割得分,将蛋白的结构域分割确定为来自该候选结构域分割集合的候选结构域分割。

[0092] 在一些实施方式中,获得表征蛋白结构中由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据包括:对于由候选结构域定义的氨基酸子序列中的每一对氨基酸,获得表征预测该对氨基酸在蛋白结构中是否被小于阈值的距离隔开的的数据。

[0093] 在一些实施方式中,基于表征蛋白结构中由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据来确定候选结构域的结构域得分包括对于候选结构域中的每个给定氨基酸,确定候选结构域中据预测与给定氨基酸隔开小于阈值的距离的多个其他氨基酸。该方法包括获得定义概率分布的数据,该概率分布对于多个非负整数值中的每一个,定义了与目标结构域相同长度的训练结构域中的给定氨基酸与训练结构域中由整数值定义的多个其他氨基酸隔开小于阈值的距离的相应似然性。结构域的长度是定义该结构域的氨基酸子序列的长度,并且训练蛋白结构域是相应蛋白的实际结构域。基于以下内容确定候选结构域的结构域得分:(i)对于候选结构域中的每个给定氨基酸,候选结构域中据预测与给定氨基酸隔开小于阈值的距离的其他氨基酸的数量,以及(ii)概率分布。

[0094] 在一些实施方式中,获得定义概率分布的数据包括获得定义高斯概率分布的平均值和标准偏差值。平均值是非负整数值集合的平均值,该非负整数值集合包括对于多个训练结构域的每个给定训练结构域中的每个给定氨基酸,给定训练结构域中与给定训练结构域隔开小于阈值的距离的其他氨基酸的数量。标准偏差值是非负整数值集合的标准偏差,该非负整数值集合包括对于多个训练结构域的每个给定训练结构域中的每个给定氨基酸,

给定训练结构域中与给定训练结构域隔开小于阈值的距离的其他氨基酸的数量。

[0095] 在一些实施方式中,基于对于候选结构域中的每个氨基酸,根据概率分布,候选结构域中被预测为与给定氨基酸隔开小于阈值的距离的其他氨基酸的数量的似然性的乘积,确定候选结构域的结构域得分。

[0096] 在一些实施方式中,该方法进一步包括,除了表征蛋白结构中由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据之外,还基于由候选结构域定义的氨基酸子序列的长度的似然性,确定候选结构域的结构域得分。结构域长度是定义结构域的氨基酸子序列的长度。

[0097] 在一些实施方式中,该方法进一步包括:除了由候选结构域分割定义的每个候选结构域的相应结构域得分之外,还基于由候选结构域分割定义的候选结构域的数量们的似然性来确定候选结构域分割的结构域分割得分。

[0098] 在一些实施方式中,从距离图获得表征蛋白结构中由候选结构域指定的氨基酸子序列中的每一对氨基酸之间的估计距离的数据。距离图表征了蛋白结构中蛋白的氨基酸序列中的每一对氨基酸之间的估计距离。距离图是距离图裁剪的加权平均值,其中:每个距离图裁剪表征蛋白的结构中在蛋白的氨基酸序列的相应氨基酸子序列中的每一对氨基酸之间的估计距离;以及通过处理与该距离图裁剪对应的氨基酸子序列的多序列比对(MSA)来生成每个距离图裁剪。

[0099] 在一些实施方式中,基于MSA中包括的氨基酸序列的数量来确定距离图裁剪的加权平均值中的每个距离图裁剪的权重,所述MSA被处理以生成距离图裁剪。

[0100] 在一些实施方式中,将氨基酸序列的结构域分割确定为具有最高结构域分割得分的候选结构域分割。

[0101] 在实施方式中,可以使用一种或多种上述技术的组合来预测蛋白结构。因此,蛋白结构预测系统的一些实施方式可以使用一个或多个神经网络来预测结构中的残基对之间的距离和/或直接估计候选结构的准确性,和/或直接生成蛋白结构。将所有这些技术一起使用并不是必需的。这些方法可以与一种或多种优化技术(诸如模拟退火)相组合,例如使用多个计算单元或梯度下降,例如以优化得分。这样的得分因此可以被认为是将被最小化的势能,诸如距离势或平均力势。

[0102] 本文所述的方法和系统可以被用来获得配体,诸如药物或工业酶的配体。例如,获得配体的方法可以包括获得靶氨基酸序列,特别是靶蛋白的氨基酸序列,并且将靶氨基酸序列用作氨基酸序列来执行如上所述或本文所述的计算机实施方法,以确定靶蛋白的(三级)结构,即,最终预测结构。然后,该方法可以包括评估一种或多种候选配体与靶蛋白的结构相互作用。该方法可以进一步包括取决于相互作用的评估结果,将一种或多种候选配体选择为配体。

[0103] 在一些实施方式中,评估相互作用可以包括评估候选配体与靶蛋白的结构结合。例如,评估相互作用可以包括鉴定以足够亲和力结合以获得生物学效应的配体。在其他实施方式中,评估相互作用可以包括评估候选配体与靶蛋白的结构的相关性,该相关性对靶蛋白例如酶的功能有影响。评估可以包括评估候选配体与靶蛋白的结构之间的亲和力,或评估相互作用的选择性。

[0104] 候选配体可以源自于候选配体的数据库,和/或可以通过修饰候选配体的数据库

中的配体(例如通过修饰候选配体的结构或氨基酸序列)来得到,和/或可以通过逐步或迭代组装/优化候选配体而得到。

[0105] 可以使用计算机辅助方法来评估候选配体与靶蛋白结构的相互作用,其中显示候选配体和靶蛋白结构的图形模型以供用户操纵,和/或可以例如使用标准分子(蛋白-配体)对接软件,部分或完全自动地执行该评估。在一些实施方式中,评估可以包括确定候选配体的相互作用得分,其中相互作用得分包括候选配体与靶蛋白之间的相互作用的度量。相互作用得分可以取决于相互作用的强度和/或特异性,例如取决于结合自由能的得分。可以取决于其得分来选择候选配体。

[0106] 在一些实施方式中,靶蛋白包含受体或酶,并且配体是受体或酶的激动剂或拮抗剂。在一些实施方式中,该方法可以被用来鉴定细胞表面标志物的结构。然后可以将其用于鉴定与细胞表面标志物相结合的配体,例如抗体或标记物(诸如荧光标记物)。这可以用于鉴定和/或治疗癌细胞。

[0107] 在一些实施方式中,候选配体可以包含小分子配体,例如分子量<900道尔顿的有机化合物。在一些其他实施方式中,候选配体可以包括多肽配体,即由氨基酸序列定义的多肽配体。

[0108] 该方法的一些实施方式可以被用来确定候选多肽配体的结构,例如药物或工业酶的配体。然后可以评估其与靶蛋白结构的相互作用。可以使用本文所述的计算机实施方法或使用常规的物理研究技术(诸如X射线晶体学和/或磁共振技术)确定靶蛋白的结构。

[0109] 因此,在另一方面,提供了获得多肽配体(例如分子或其序列)的方法。该方法可以包括获得一种或多种候选多肽配体的氨基酸序列。该方法可以进一步包括将候选多肽配体的氨基酸序列作为氨基酸序列来执行如上所述或本文所述的计算机实施方法,以确定候选多肽配体的(三级)结构。该方法可以进一步包括,在计算机上和/或通过物理研究获得靶蛋白的靶蛋白结构,以及评估所述一种或多种候选多肽配体中的每一种的结构与靶蛋白结构之间的相互作用。该方法可以进一步包括取决于评估结果,将所述一种或多种候选多肽配体中的一种选择为多肽配体。

[0110] 如前所述,评估相互作用可以包括评估候选多肽配体与靶蛋白结构的结合,例如,鉴定以足够亲和力结合以获得生物学效应的配体,和/或评估候选多肽配体与靶蛋白结构的关联性,该关联性对靶蛋白例如酶的功能有影响,和/或评估候选多肽配体与靶蛋白结构之间的亲和力,或评估相互作用的选择性。在一些实施方式中,多肽配体可以是适体。

[0111] 该方法的实施方式可以进一步包括合成,即制备小分子或多肽配体。配体可以通过任何常规化学技术合成和/或可能已经是现有的,例如可以来自化合物库或可以使用组合化学方法合成。

[0112] 该方法可以进一步包括在体外和/或体内测试配体的生物学活性。例如,可以测试该配体的ADME(吸收、分布、代谢、排泄)和/或毒理学性质,以筛选掉不合适的配体。该测试可以包括例如使候选小分子或多肽配体与靶蛋白接触,并测量蛋白的表达或活性的变化。

[0113] 在一些实施方式中,候选(多肽)配体可以包括:分离的抗体、分离的抗体的片段、单可变结构域抗体、双或多特异性抗体、多价抗体、双可变结构域抗体、免疫缀合物、纤连蛋白分子、阿德奈汀(adnectin)、DARPin、阿维默(avimer)、亲和体(affibody)、抗运载蛋白(anticalin)、亲和素(affilin)、蛋白质表位模拟物或其组合。候选(多肽)配体可以包括具

有突变或化学修饰的氨基酸Fc区的抗体,例如与野生型Fc区相比,防止或降低ADCC(抗体依赖性细胞毒性)活性和/或增加半衰期的Fc区。

[0114] 错折叠的蛋白与多种疾病有关。因此,在另一方面,提供了一种鉴定蛋白错折叠疾病的存在的方法。该方法可以包括获得蛋白的氨基酸序列,并使用该蛋白的氨基酸序列来执行如上所述或本文所述的计算机实施方法,以确定该蛋白的结构。该方法可以进一步包括例如通过常规(物理)方法,获得从人体或动物体获得的蛋白版本的结构。然后,该方法可以包括将蛋白的结构与从体内获得版本的结构进行比较,并且取决于比较的结果来鉴定蛋白错折叠疾病的存在。即,可以通过与在计算机上确定的结构进行比较来确定来自体内的蛋白版本的错折叠。

[0115] 在一些其他方面,如上所述或本文所述的计算机实施方法可以被用来从靶蛋白的氨基酸序列中鉴定靶蛋白上的活性/结合/阻断位点。

[0116] 可以实施本说明书中描述的主题的特定实施方式,以实现以下一个或多个其他优点。

[0117] 本说明书描述了一种结构预测系统,该系统可以使用并行运行的多个搜索计算单元来确定蛋白的最终预测结构。每个搜索计算单元被配置为在一系列搜索迭代中确定预测蛋白结构的相应轨迹,并且可以将特别“有希望的”预测结构(即,具有高质量得分的预测结构)保存到中央存储器。为了确定对当前预测结构的更新,搜索计算单元可以使用来自中央存储器中存储的先前预测结构的结构片段来扰动当前预测结构。以这种方式,搜索计算单元可以连续地(即,在每个搜索迭代处)利用先前发现的关于蛋白结构的知识来更新当前预测结构。此外,在一些实施方式中,搜索计算单元可以共享单个中央存储器,在这种情况下,每个搜索计算单元可以利用关于由每一其他搜索计算单元发现的蛋白结构的知识来更新其当前预测结构。相反,在一些常规系统中,使用从(例如,通过实验确定的)已知蛋白结构的数据库中获得的结构片段来仅更新蛋白的预测结构。通过使用更多样的结构片段集合(包括来自在中央存储器中存储的先前预测结构的结构片段)来更新蛋白的预测结构,与一些常规系统相比,本说明书中所述的系统可以在更少的搜索迭代中潜在地生成更准确的预测折叠结构。与一些常规系统相比,通过需要较少的搜索迭代来生成准确的预测折叠结构,本说明书中描述的系统可以由此消耗较少的计算资源(例如,存储器和计算能力)。

[0118] 本说明书描述了一种结构预测系统,该系统可以使用在已知蛋白结构的数据库上训练的生成神经网络来生成结构片段。生成神经网络可以生成大量的不同结构片段,它们中的许多可能不对应于任何已知的(例如,通过实验确定的)蛋白结构。在本说明书中描述的系统中包括的搜索计算单元可以使用由生成神经网络生成的结构片段来更新其相应的当前预测结构,从而潜在地生成比局限于从已知蛋白结构中提取的结构片段的常规系统更准确的预测折叠结构。此外,通过使用由生成神经网络生成的结构片段来更新蛋白的预测结构,与一些常规系统相比,本说明书中描述的系统可能需要较少的搜索迭代来确定蛋白的准确预测结构,从而可能消耗较少的计算资源(例如,存储器和计算能力)。

[0119] 本说明书描述了一种结构预测系统,该系统可以使用几何得分和值得分来确定预测结构的质量得分。预测结构的几何得分是蛋白的预测结构与蛋白的实际结构之间的相似性度量的估计。如果当前预测结构是给定的预测结构,则给定的预测结构的值得分是表征由搜索计算单元在未来的搜索迭代中生成的未来预测折叠结构的的质量的质量得分的估计。

通过使用几何得分和值得分确定预测结构的质量得分,与一些常规系统相比,本说明书中描述的系统可以确定更准确的预测蛋白结构。例如,使用值得分可以使得搜索计算单元能够以可能会在短期内(例如,经过几个搜索迭代)降低当前预测结构的的质量的方式来更新其当前预测结构,但是可能导致更高质量的最终预测结构。此外,与一些常规系统相比,通过在确定预测结构的质量得分中使用几何得分和值得分,本说明书中描述的系统可能需要更少的搜索迭代来确定准确的预测蛋白结构,从而可能消耗更少的计算资源(例如,存储器和计算能力)。

[0120] 本说明书描述了一种结构预测系统,该系统可以通过使用距离预测神经网络生成距离图的多个“裁剪”,并且随后融合裁剪来生成蛋白的距离图(该距离图表征蛋白结构中的氨基酸残基之间的估计距离)。通过生成距离图裁剪而不是整个距离图,距离预测神经网络的架构不受必须建模的最长氨基酸序列的限制。因此,本说明书中描述的距离预测神经网络可以具有比其他方式更复杂的架构(例如,具有更多的神经网络层),从而能够实现更精确的距离图估计。

[0121] 本说明书描述了一种结构预测系统,该系统可以使用连续值的距离估计或距离范围概率分布来生成表征氨基酸对之间的距离的距离图。相反,在一些常规系统中,距离图使用二进制变量(例如,指示氨基酸对之间的距离是否小于预定阈值)来表征氨基酸对之间的距离。通过生成传达关于预测结构对实际蛋白结构的符合程度的更精确信息的距离图,与一些常规系统相比,本说明书中描述的系统可以确定更准确的预测折叠结构。

[0122] 通过生成比由一些常规系统生成的距离图更准确和精确的距离图,与一些常规系统相比,本说明书中描述的系统可以消耗更少的计算资源(例如,存储器和计算能力)。例如,本说明书中描述的系统可以使用所生成的距离图来确定质量得分,其使得当前预测结构比在一些常规系统中更有效地被更新,从而减少了确定精确的预测结构所需的搜索迭代的数量。

[0123] 本说明书描述了一种结构预测系统,该系统可以被用来生成预测的蛋白结构,该结构可以准确地近似于不同蛋白的实际结构。如下文更详细所述,准确地预测蛋白结构可以有助于理解生命过程(例如,包括许多疾病的机理)和设计蛋白(例如,作为药物或作为工业过程的酶)。通过由氨基酸序列预测蛋白结构,本说明书中描述的系统可以促进涉及蛋白的生物化学研究和工程(例如,药物开发)的领域,并且不需要确定蛋白结构的昂贵且费时的物理实验。

[0124] 本说明书描述了一种结构预测系统,该系统可以使用“暖”梯度下降优化程序来优化定义蛋白的预测结构的当前结构参数值。具体地,暖梯度下降优化程序为当前结构参数值赋予了动量,使它们可以在优化期间“滚动”质量得分表面,而不是直接找到质量得分表面的局部最小值。质量得分表面是指通过从相应结构参数值到由相应结构参数值定义的预测蛋白结构的质量得分的映射而定义的高维表面。通过使用暖梯度下降优化程序,本说明书中描述的系统可以有效地探索可能的结构参数值的空间,以确定蛋白的不同预测结构集合,每个预测结构均近似对应于质量得分表面的局部最小值。可以将预测结构集合中的最佳预测结构选择为蛋白的最终预测结构。以这种方式,与如果系统使用产生蛋白的较少不同预测结构集合的不同的梯度下降优化程序相比,本说明书中描述的系统可以确定具有更高质量得分的蛋白的最终预测结构。

[0125] 本说明书描述了一种结构预测系统,该系统可以确定最终预测结构,该结构(近似)优化基于一个或多个评分神经网络的输出的质量得分。评分神经网络可以包括例如结构预测神经网络(其生成在结构参数的可能值上的相应概率分布)、几何神经网络(其生成蛋白的预测结构与实际结构之间的相似性度量的估计)以及距离预测神经网络(其生成表征氨基酸间距离的距离范围概率分布)。可以使用机器学习训练技术,在(例如,使用物理实验确定的)已知蛋白结构的数据库上训练评分神经网络。通过基于使用机器学习技术直接从已知蛋白结构中获悉的质量得分,而不是根据启发式约束手动制作的质量得分来确定最终预测结构,本说明书中描述的系统可以生成更准确的最终预测蛋白结构。

[0126] 本说明书描述了一种结构预测系统,其可以通过生成更准确的预测最终蛋白结构,同时消耗更少的计算资源(例如,存储器、计算能力或两者)来胜过一些其他结构预测系统。例如,一些结构预测系统通过使用随机蛋白结构片段来迭代地调整当前预测结构以生成预测结构。这些结构预测系统可以确定应当将随机蛋白结构片段插入当前预测结构中,如果这样做会增加预测结构的总体质量得分的话。与这些系统相比,本说明书描述了一种结构预测系统,该系统可以使用梯度下降来直接优化关于预测结构的质量得分,并且以这种方式,可以通过较少的迭代来生成更准确的预测结构。

[0127] 本说明书描述了一种结构域分割系统,其可以生成蛋白的结构域分割,该结构域分割可以被用于准确和有效地预测蛋白的结构。例如,结构预测系统可以确定蛋白的每个结构域的预测结构(即,如由结构域分割定义),并且随后通过组合每个结构域的预测结构来确定整个蛋白的预测结构。与一次性确定整个蛋白的预测结构相比,确定蛋白的结构域的预测结构通常是一个“更容易”的问题。具体地,由于氨基酸序列的可能结构的数量随氨基酸序列的长度呈指数增加,所以结构域的可能预测结构的搜索空间通常将成指数地小于整个蛋白的可能预测结构的搜索空间。通过使得结构预测系统能够分开地确定蛋白结构域的预测结构,而不是一次性预测整个蛋白结构,本说明书中描述的系统使得结构预测系统能够生成更准确的预测,同时消耗更少的计算资源(例如,存储器、计算能力,或两者)。

[0128] 本说明书描述了一种系统,该系统可以通过组合大量距离图裁剪来生成蛋白的距离图。该系统通过处理与蛋白的氨基酸序列的相应子序列对应的相应多序列比对(MSA)来生成每个距离图裁剪。通过基于由多个不同的MSA计算得到的距离图裁剪来生成蛋白的距离图,与通过处理与蛋白的整个氨基酸序列对应的单个MSA相比,本说明书中描述的系统可以生成更鲁棒和准确的距离图。

[0129] 可以意识到,本说明书中描述的结构预测系统可以被用来预测蛋白结构域、整个蛋白或蛋白复合物的结构。在蛋白复合物中,一组多种蛋白折叠在一起形成全局结构(例如,各个蛋白可以通过非共价蛋白-蛋白相互作用连接在一起)。

[0130] 在附图和下文的描述中阐述本说明书的主题的一个或多个实施方式的细节。根据说明书、附图和权利要求书,本主题的其他特征、方面和优点将变得显而易见。

附图说明

[0131] 图1是使用一个或多个搜索计算单元的示例性结构预测系统的框图。

[0132] 图2是未折叠蛋白和折叠蛋白的图示。

[0133] 图3是示例性本地存储器更新系统的框图。

- [0134] 图4是示例性片段生成系统的框图。
- [0135] 图5是例如可以由参考图1描述的结构预测系统使用的示例性评分系统的框图。
- [0136] 图6是示例性距离预测系统的框图。
- [0137] 图7是用于通过迭代地更新蛋白的当前预测结构来生成蛋白的预测结构的相应轨迹的示例性过程的流程图。
- [0138] 图8是用于确定是否将蛋白的当前预测结构更新为蛋白的替代预测结构的示例性过程的流程图。
- [0139] 图9是使用优化系统的示例性结构预测系统的框图。
- [0140] 图10是例如可以由参考图9描述的结构预测系统使用的示例性评分系统的框图。
- [0141] 图11是用于确定蛋白的预测结构的示例性过程的流程图。
- [0142] 图12是示例性结构域分割系统的框图。
- [0143] 图13是用于确定蛋白的氨基酸序列的结构域分割的示例性过程的流程图。
- [0144] 图14是用于生成表征蛋白中的每一对氨基酸之间的估计距离的距离图的示例性过程的流程图。
- [0145] 图15是示例性计算系统的框图。
- [0146] 图16示出了使用优化系统的结构预测系统的示例性性能。
- [0147] 图17示出了使用距离预测系统的结构预测系统的示例性性能。
- [0148] 图18示出了用于确定蛋白的预测结构的示例性数据流。
- [0149] 图19示出了示例性距离图的各方面。
- [0150] 图20示出了距离预测神经网络的残差块的示例性架构。
- [0151] 图21示出了被配置为生成蛋白结构片段的DRAW生成神经网络的示例性架构。
- [0152] 在各个附图中,相同的附图标记和名称表示相同的元件。

具体实施方式

[0153] 图1示出了示例性结构预测系统100。结构预测系统100是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下文所述的系统、部件和技术。

[0154] 结构预测系统100被配置为处理定义蛋白104的氨基酸序列102的数据,以生成蛋白104的最终预测结构106。氨基酸序列102中的每个氨基酸是有机化合物,其包括氨基官能团和羧基官能团,以及对该氨基酸特定的侧链(即原子团)。最终预测结构106定义在蛋白104经历蛋白折叠之后,蛋白104的氨基酸序列102中的原子的三维构型的估计。蛋白折叠是指序列例如氨基酸的无规卷曲(例如,由蛋白104的氨基酸序列102定义的)折叠成独特的三维构型(例如,如由最终预测结构106所估计的)的物理过程。尽管构型被描述为独特的,但这并不意味着蛋白在一些情况下不可以不同地折叠。图2提供了未折叠蛋白和折叠蛋白的图示。

[0155] 蛋白的结构决定了蛋白的生物学功能。因此,确定蛋白结构可以有助于理解生命过程(例如,包括许多疾病的机理)和设计蛋白(例如,作为药物或作为工业过程的酶)。例如,哪些分子(例如药物)将与蛋白结合(以及在何处发生结合)取决于蛋白的结构。由于药物的有效性会受到它们与蛋白(例如在血液中)结合的程度的影响,因此确定不同蛋白的结

构可能是药物开发的重要方面。然而,使用物理实验(例如,通过X射线晶体学)确定蛋白结构可能耗时且非常昂贵的。因此,通过由氨基酸序列预测蛋白结构,系统100可以促进涉及蛋白的生化研究和工程领域(例如,药物开发)。

[0156] 氨基酸序列102可以以任何适当的数字格式表示。例如,氨基酸序列102可以被表示为单热向量(one-hot vector)的序列。在该示例中,单热向量序列中的每个单热向量表示氨基酸序列102中的相应氨基酸。单热向量对于每个不同的氨基酸(例如,预定数量的氨基酸中的)具有不同的分量。表示特定氨基酸的单热向量在与该特定氨基酸相对应的分量中具有值1(或一些其他预定值),而在其他分量中具有零(或一些其他预定值)。

[0157] 氨基酸序列102的结构(例如,系统100输出的最终预测结构106)由结构参数集合的值定义。在一些实施方式中,结构参数是三维(3D)数字坐标的序列(例如,表示为3D向量),其中每个坐标表示来自氨基酸序列102的氨基酸中的相应原子的位置(在一些给定参考系中)。例如,结构参数可以是表示该结构中氨基酸中的 α 碳原子的相应位置的3D数字坐标的序列。在本说明书中被称为主链原子的 α 碳原子是指氨基酸中与氨基官能团、羧基官能团和侧链键合的碳原子。在一些实施方式中,结构参数是该结构中,氨基酸中的特定原子之间的扭转角(即,二面角)序列。例如,结构参数可以是该结构中的氨基酸的主链原子之间的 ϕ (ϕ)、 ψ (ψ)以及任选的 ω (ω)二面角的序列,尤其是在 α 碳的任一侧(肽键通常将 ω 约束到接近0或180度)。

[0158] 系统100包括一个或多个搜索计算单元(例如,搜索计算单元108和110)。每个搜索计算单元被配置为维护当前预测结构(例如,当前预测结构112和114),并且在一个或多个搜索迭代中的每一个迭代处,确定是否将当前预测结构更新为替代预测结构(例如,替代预测结构116或118之一)。通常,每个搜索计算单元的任务是确定具有最高可能质量得分(例如,质量得分120或122)的预测结构。如下文更详细所述(例如,参考图7),每个搜索计算单元可以将具有高质量得分的预测折叠结构存储在中央存储器124中。搜索计算单元可以对存储在中央存储器124中的预测结构进行分段,以确定用于更新其相应当前预测结构的结构片段,并且系统可以通过选择存储在中央存储器124中的预测结构来确定蛋白104的最终预测结构106。

[0159] 预测结构的质量得分通常表征预测结构的质量。例如,如将参考图5更详细地所述,预测结构的质量得分可以表征蛋白104的预测结构与实际结构之间的估计相似性、基于预测结构中的氨基酸的主链原子之间的距离的预测结构的似然性,或两者。通常,可以基于氨基酸序列102中的氨基酸的原子的适当子集(例如,仅基于主链原子而不基于任何其他原子)或基于氨基酸序列102中的每个氨基酸中的所有原子来确定预测结构的质量得分。仅基于氨基酸中的主链原子的预测结构的质量得分可以被称为主链原子得分,而参考每个氨基酸中的所有原子确定的质量得分可以被称为全原子得分。

[0160] 每个搜索计算单元可以是例如计算机、具有多个内核的计算机内的内核,或者是能够独立执行操作(例如,搜索迭代)的计算机内的其他硬件或软件,例如专用线程。搜索计算单元可以包括处理器内核、处理器、微处理器、专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路),或任何其他适当的计算单元。在一些示例中,搜索计算单元都是相同类型的计算单元。在其他示例中,搜索计算单元可以是不同类型的计算单元。例如,一个搜索计算单元可以是CPU,而其他搜索计算单元可以是GPU。搜索计算单元可以被配置

为异步地操作。更具体地,每个搜索计算单元可以被配置为与每个其他搜索计算单元无关地执行搜索迭代。

[0161] 每个搜索计算单元通过迭代地更新相应的当前预测结构来生成蛋白104的预测结构的相应轨迹(即,序列)(每个轨迹与不同的质量得分相关联)。在每个搜索迭代处,每个搜索计算单元可以确定是否满足中央存储器更新条件,并且响应于确定满足中央存储器更新条件,可以将其当前预测结构存储在中央存储器124中。将预测结构存储在中央存储器124中是指将定义预测结构的结构参数的值存储在中央存储器124中。例如,如果当前预测结构的质量得分是直到该搜索迭代为止,由搜索计算单元生成的预测结构的轨迹中的最高质量得分,则搜索计算单元可以确定满足中央存储器更新条件。在该示例中,动态地更新中央存储器124以存储由搜索计算单元生成的最“有希望的”预测结构(例如,最有可能准确地近似于氨基酸序列102的实际结构)。中央存储器124是可以作为例如逻辑数据存储区或物理数据存储设备实施的(集成的或分布式的)数据存储。

[0162] 每个搜索计算单元维护被配置为存储结构片段的相应本地存储器(例如,本地存储器126或128)。每个结构片段表征了氨基酸序列102中的氨基酸的子序列(即片段)的预测结构。每个结构片段由下述(i)和(ii)定义:(i)指示氨基酸序列102的相应子序列的数据,和(ii)表征氨基酸序列102的相应子序列中的氨基酸的结构的结构参数集合的值。在特定示例中,如果氨基酸序列102由[A,C,E,F,D,G,G,A]给出(例如,其中A是氨基酸丙氨酸的符号,C是氨基酸半胱氨酸的符号等),则结构片段可以表征氨基酸序列102中的氨基酸[C,E,F]的连续子序列的预测结构。在该示例中,结构片段可以由下述(i)和(ii)定义:(i)指示氨基酸[C,E,F]的子序列的数据,和(ii)包括氨基酸C、E和F中的主链原子之间的扭转角的结构参数的值。如稍后更详细所述,系统100包括本地存储器更新系统130,其被配置为初始化和更新每个搜索计算单元的相应本地存储器。

[0163] 每个搜索计算单元在由搜索计算单元执行的多个搜索迭代中的每一个处,确定是否将其相应当前预测结构更新为替代预测结构(例如,替代预测结构116或118)(如前所述)。特别地,在每个搜索迭代处,每个搜索计算单元从由搜索计算单元维护的本地存储器中获得一个或多个相应结构片段(例如,结构片段132或134),并且由每个获得的结构片段生成相应替代预测结构。搜索计算单元可以通过从其本地存储器中随机采样预定数量的结构片段来从其本地存储器获得结构片段。搜索计算单元可以确定每个替代预测结构以包括:(i)由搜索计算单元维护的当前预测结构的一部分,以及(ii)所获得的结构片段。以这种方式,搜索计算单元使用相应结构片段来生成作为由搜索计算单元维护的当前预测结构的“扰动”的每个替代预测结构。参考图8进一步描述用于确定是否将蛋白的当前预测结构更新为蛋白的替代预测结构的示例性过程。

[0164] 每个结构片段的大小(例如,对应于该结构片段的氨基酸序列102的子序列中的氨基酸的数量)和在每个迭代处由搜索计算单元生成的替代预测结构的数量都是搜索计算单元超参数。通常,这样的超参数在搜索计算单元之间可以不同,并且可以在单个搜索计算单元内的迭代之间改变(例如,根据固定的时间表或如由超参数选择神经网络的输出所指定的)。

[0165] 每个搜索计算单元使用评分系统(例如,评分系统136)来确定每个相应替代预测结构的相应数值质量得分(例如,质量得分120或122)。如前所述,预测结构的质量得分通常

表征该结构的质量。参考图5描述评分系统136的示例。除了替代预测结构的质量得分之外，搜索计算单元还获得由搜索计算单元维护的当前预测结构的质量得分。例如，搜索计算单元可以维护在先前的搜索迭代中使用评分系统136确定的当前预测结构的质量得分。

[0166] 每个搜索计算单元包括更新引擎(例如，更新引擎138)，其使用替代预测结构的质量得分和当前预测结构的质量得分来确定是否将其相应当前预测结构更新为任一替代预测结构。

[0167] 在一些实施方式中，更新引擎可以基于质量得分，使用确定性程序来确定是否将当前预测结构更新为替代预测结构。例如，如果特定替代预测结构具有比当前预测结构和任何其他替代预测结构更高的质量得分，则更新引擎可以确定将当前预测结构更新为特定替代预测结构。在该示例中，如果当前预测结构具有比任何替代预测结构更高的质量得分，则更新引擎可以确定不将当前预测结构更新为任何替代预测结构。

[0168] 在一些实施方式中，更新引擎可以基于质量得分，使用随机程序(即，涉及一些随机性)来确定是否将当前预测结构更新为替代预测结构。例如，更新引擎可以使用质量得分来确定包括当前预测结构和每个替代预测结构的结构集合上的概率分布。在特定示例中，更新引擎可以通过使用soft-max函数，处理当前预测结构和每个替代预测结构的相应质量得分来确定概率分布。更新引擎可以使用概率分布，确定将当前预测结构更新为从包括当前预测结构和每个替代预测结构的结构集合中采样的结构。通过使用随机程序确定是否将当前预测结构更新为替代预测结构，系统100可以“探索”可能的蛋白结构的空間，从而潜在地确定更准确的预测结构(如参考图8进一步所述)。

[0169] 系统100包括本地存储器更新系统130，其被配置为初始化和更新搜索计算单元的本地存储器。初始化搜索计算单元的本地存储器是指在由搜索计算单元执行的第一搜索迭代之前，将多个结构片段存储在本地存储器中。更新搜索计算单元的本地存储器是指在由搜索计算单元执行的搜索迭代之间，将不同的结构片段包括在本地存储器中(并且可能从本地存储器中去除结构片段)。如将参考图3更详细所述，本地存储器更新系统130可以例如使用从由搜索计算单元存储在中央存储器中的预测结构获得的结构片段来更新搜索计算单元的本地存储器。由此，本地存储器更新系统130可以使得每个搜索计算单元能够利用由其他搜索计算单元生成的逐步完善的预测结构来更新其当前预测结构。

[0170] 为了确定蛋白104的最终预测结构106，系统100选择存储在中央存储器124中的预测结构。通常，存储在中央存储器124中的每个预测结构与得分相关联，并且系统100通过基于得分来选择中央存储器124中的预测结构而确定最终预测结构106。例如，系统100可以将最终预测结构106确定为存储在中央存储器124中的具有最高得分的预测结构。与存储在中央存储器124中的预测结构相关联的得分可以是例如，主链原子质量得分或全原子质量得分。

[0171] 在一些情况下，确定全原子质量得分可能比确定主链原子质量得分在计算上更为密集。也就是说，与确定预测结构的主链原子质量得分相比，确定预测结构的全原子质量得分可能需要更多的存储器、执行更多的算术运算或两者。为了减少计算资源消耗，每个搜索计算单元可以被配置为更新由搜索计算单元维护的相应当前预测结构，并且基于主链原子质量得分来更新中央存储器124。然而，为了确定最终预测结构106，系统可以对存储在中央存储器124中的每个预测结构确定全原子质量得分，并且基于这些全原子得分选择最终预

测结构106。

[0172] 通常,每个搜索计算单元的本地存储器和中央存储器124可以分别以集成或分布式格式实施。例如,搜索计算单元的本地存储器可以被实施为物理上远离搜索计算单元的其他部件(例如,搜索计算单元的评分系统和更新系统)的数据存储。作为另一示例,中央存储器124可以被实施为分布式数据存储。在该示例中,每个搜索计算单元可以维护独立的“中央存储器”,并且中央存储器124可以被理解为由每个搜索计算单元维护的相应中央存储器的组合。

[0173] 在一些情况下,存储在中央存储器中的每个预测结构可以与下述(i)和(ii)相关联:(i)生成预测结构的搜索计算单元的超参数集合,以及(ii)预测结构的质量得分(例如,主链原子质量得分或全原子质量得分)。搜索计算单元的超参数可以包括指定例如温度超参数的参数,所述温度超参数表征搜索计算单元更新当前预测结构的容易程度(如稍后将更详细描述)、由搜索计算单元执行的搜索迭代的次数等。每个搜索计算单元可以根据基于相应质量得分的、存储在中央存储器中的在预测结构上的概率分布,通过对与存储在中央存储器中的预测结构相关联的超参数值集合进行采样来重复地更新其超参数值。这可以导致搜索计算单元的超参数的自动优化。

[0174] 图2是未折叠的蛋白和折叠的蛋白的图示。未折叠的蛋白是序列,例如氨基酸的无规卷曲。未折叠的蛋白经历蛋白折叠并折叠成独特的3D构型。蛋白结构通常包括稳定的局部折叠模式,其可以被称为二级结构,例如 α 螺旋(例如,如202所示)和 β 折叠。折叠蛋白的结构可以由结构参数集合的值来定义。例如,如204所描绘,结构参数可以是表示给定参考系中,折叠蛋白的氨基酸中的主链原子的相应位置的3D数字坐标(例如,[x,y,z]坐标)的序列。

[0175] 图3是示例性本地存储器更新系统130的框图。本地存储器更新系统130是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施了下文所述的系统、部件和技术。

[0176] 本地存储器更新系统130被配置为初始化并随后连续更新搜索计算单元的本地存储器(例如,如参考图1所述)。搜索计算单元302应当被理解为包括在结构预测系统100(如参考图1所述的)中的任何搜索计算单元。初始化搜索计算单元302的本地存储器304是指在由搜索计算单元302执行的第一搜索迭代之前,将多个结构片段306存储在本地存储器304中。更新搜索计算单元302的本地存储器304是指在由搜索计算单元302执行的搜索迭代之间,将不同的结构片段306包括在本地存储器304中(并且可能从本地存储器304中去除结构片段)。为了方便起见,以下对本地存储器更新系统130的描述是指更新搜索计算单元302的本地存储器304。然而,用于更新搜索计算单元302的本地存储器304的相同程序(如下所述)可以类似地应用于初始化搜索计算单元的本地存储器304。

[0177] 每当满足搜索计算单元302的本地存储器更新条件时,本地存储器更新系统130就更新搜索计算单元302的本地存储器304。例如,如果自从上次本地存储器更新系统130更新搜索计算单元302的本地存储器304以来,搜索计算单元302已经执行了预定数量的搜索迭代,则可以满足搜索计算单元302的本地存储器更新条件。作为另一示例,如果在预定数量的迭代中,由搜索计算单元302维护的当前预测结构的质量得分的平均增加低于预定阈值,则可以满足搜索计算单元302的本地存储器更新条件。在该示例中,当前预测结构的质量得

分的平均增加低于阈值,可以指示本地存储器304需要用不同的结构片段“刷新”以使得搜索计算单元302能够继续改善其当前预测结构。

[0178] 在一些实施方式中,本地存储器更新系统130通过获得不同的结构片段306并且将当前在本地存储器304中的结构片段替换为该不同的结构片段来更新本地存储器304。在一些实施方式中,本地存储器更新系统130通过获得不同的结构片段306并使用替换策略来部分地替换当前在本地存储器304中的结构片段来更新本地存储器304。例如,替换策略可以是先进先出(FIFO)替换策略,其中最早包括在本地存储器304中的结构片段用该不同的结构片段306替换。

[0179] 本地存储器更新系统130可以以各种不同的方式产生要包括在搜索计算单元302的本地存储器304中的结构片段306。例如,如下文更详细所述,本地存储器更新系统130可以通过以下方式产生结构片段:(i)对存储在中央存储器124中的预测结构进行分段,(ii)对存储在不同蛋白的已知结构的结构数据库308中的实际结构进行分段,(iii)使用包括在片段生成系统310中的生成神经网络生成结构片段,或(iv)其组合。对蛋白的结构进行分段是指从蛋白的结构中提取一个或多个结构片段。

[0180] 在一些实施方式中,本地存储器更新系统130使用由搜索计算单元存储在中央存储器124中的预测结构来产生结构片段306。更具体地,本地存储器更新系统130可以使用采样引擎314,从中央存储器124采样结构312,并且使用分段引擎316处理结构312以确定来自结构312的多个结构片段。

[0181] 采样引擎314可以根据与存储在中央存储器124中的预测结构相关联的得分来从中央存储器124采样预测结构。如前所述,存储在中央存储器124中的每个预测结构可以与质量得分(例如,主链原子质量得分或全原子质量得分)相关联。例如,为了从中央存储器124采样预测结构,采样引擎314可以使用关联得分来确定在中央存储器124中存储的结构上的概率分布(例如,通过使用soft-max函数来处理得分)。然后,采样引擎可以根据所确定的概率分布从中央存储器124采样预测结构。

[0182] 分段引擎316可以处理从中央存储器124采样的预测结构,以生成多个结构片段306。例如,分段引擎316可以通过将预测结构划分成多个结构片段来生成结构片段306,其中每个结构片段定义了氨基酸序列102中的氨基酸残基的连续序列的结构。

[0183] 在一些实施方式中,本地存储器更新系统130使用存储在结构数据库308中的不同蛋白的实际结构来产生结构片段306。可能已经使用物理实验方法例如X射线晶体学确定了存储在结构数据库308中的不同蛋白的实际结构。本地存储器更新系统可以使用采样引擎314从结构数据库308采样(例如,随机地)结构,并且使用分段引擎316处理所采样的结构以确定结构片段306。

[0184] 在一些实施方式中,本地存储器更新系统130使用片段生成系统310来生成结构片段。如参考图4进一步所述,片段生成系统包括生成神经网络,其被训练以生成定义现实结构片段306的网络输出。可以使用机器学习训练技术,在不同蛋白的实际结构的数据库(例如,结构数据库308)上训练片段生成系统310。

[0185] 图4是示例性片段生成系统310的框图。片段生成系统310是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下文所述的系统、部件和技术。

[0186] 片段生成系统310被配置为接收包括氨基酸序列102的输入,并且处理该输入以生成与氨基酸序列102相对应的现实结构片段402。片段生成系统310生成的每个结构片段402由生成神经网络406生成的结构参数404定义,并且表征氨基酸序列102中的氨基酸子序列的现实结构。通常,片段生成系统310可以被用来对氨基酸序列102中的任何氨基酸子序列生成大量不同的现实结构片段402。如果片段大致符合自然界中发现的实际蛋白结构所满足的物理和生化约束,则该结构片段可以被理解为“现实的”。例如,现实结构片段可能不太可能包括会暴露于水的疏水部分,因为这种构型可能不太可能出现在自然界中发现的实际蛋白结构中。

[0187] 为了生成结构片段402,片段生成系统310使用特征生成引擎408来生成用于生成神经网络406的网络输入410。网络输入410可以包括氨基酸序列102的表示、与氨基酸序列102相对应的比对特征412或两者。比对特征412由其他蛋白(例如同源蛋白)的氨基酸序列与相似氨基酸序列102的多序列比对(MSA)中得出。MSA是指来自多个其他蛋白的氨基酸序列中的氨基酸与氨基酸序列102的相似序列(例如,除了未比对的插入间隙和/或残基缺失以外,氨基酸相同)中的氨基酸之间的对应关系。跨MSA序列的两个氨基酸残基位置的相关变化可以指示哪些氨基酸可能接触。特征生成引擎408可以通过使用任何适当的计算序列比对技术(例如,渐进式比对构建)处理其他蛋白的氨基酸序列(例如,存储在数据库中的)来生成MSA。特征生成引擎408可以生成由MSA得出的比对特征412,包括MSA的表示和统计特征(例如,二阶统计特征),诸如参考S.Seemayer,M.Gruber和J.Soding:“CCMpred:fast and precise prediction of protein residue-residue contacts from correlated mutations”,Bioinformatics,2014年所描述的那些。比对特征412可以是1D、2D或具有任何其他适当的维度。后面描述了一些示例性特征。

[0188] 在一些情况下,网络输入410包括表征整个氨基酸序列102的数据(例如,网络输入410包括氨基酸序列102中的每个氨基酸的表示)。在一些其他情况下,网络输入410包括仅表征氨基酸序列102中的氨基酸的适当子集的数据(即,不是氨基酸序列102中的每个氨基酸)。

[0189] 片段生成系统310将网络输入410提供给生成神经网络406,生成神经网络406被配置为根据生成神经网络权重的当前值来处理网络输入410以生成结构参数404。结构参数可以是例如氨基酸序列102的氨基酸集合的主链原子的主链原子扭转角或主链原子坐标。在一些实施方式中,结构参数404定义了整个氨基酸序列102的结构,并且片段生成系统310通过对结构参数404进行分段来生成结构片段。在一些实施方式中,结构参数404定义氨基酸序列102的片段(即,适当的子集)的结构。

[0190] 为了生成结构参数404,生成神经网络被配置为处理网络输入410,以对于每个结构参数404生成定义在结构参数的可能值上的相应概率分布的数据。例如,如果结构参数404是主链原子扭转角,则定义在结构参数的可能值上的概率分布的数据可以是冯·米塞斯概率分布的混合的参数值。作为另一示例,如果结构参数404是主链原子坐标,则定义在结构参数的可能值上的概率分布的数据可以是高斯概率分布的平均值和标准偏差的值。为了生成每个结构参数404,片段生成系统310根据在结构参数的可能值上的概率分布来采样结构参数的值。

[0191] 由于片段生成系统310不确定地生成结构参数404(即,通过从概率分布采样),因

此片段生成系统可以由给定网络输入410产生许多不同的现实结构片段402。

[0192] 通常,可以使用任何适当的神经网络架构来实施生成神经网络406。以下是一些示例。

[0193] 在一些情况下,生成神经网络406可以被配置为顺序地生成与每个结构参数404相对应的概率分布(即,根据结构参数404的排序)。在这些情况下,为了生成在特定结构参数的可能值上的概率分布,可以将生成神经网络406配置为处理网络输入410和定义一个或多个先前结构参数的先前确定值的数据。先前结构参数是指在结构参数的排序中在特定结构参数之前的结构参数。

[0194] 例如,生成神经网络406的架构可以源自于WaveNet生成神经网络的架构,如参考A.Van Den Oord,S.Dieleman,H.Zen,et al.:“A generative model for raw audio”,arXiv:1609.03499v2,2016所描述的。在该示例中,生成神经网络406架构可以包括卷积子网络(包括一个或多个屏蔽卷积层)和输出层。为了生成在给定结构参数的可能值上的概率分布,卷积子网络可以处理包括网络输入410和一个或多个先前结构参数的先前确定的值的卷积子网络输入,以生成卷积子网络输出。输出层可以是soft-max层,其被配置为处理对给定结构参数生成的卷积子网络输出,以生成定义在给定结构参数的可能值上的概率分布的输出。

[0195] 在一些情况下,生成神经网络406可以被配置为通过根据先验概率分布从潜变量空间中采样一个或多个潜变量来生成与每个结构参数404相对应的概率分布。生成神经网络406可以处理采样的潜变量和网络输入410(例如,包括氨基酸序列102的表示)以生成与每个结构参数404相对应的概率分布。在这些情况下,先验概率分布可以是在训练生成神经网络406之前选择的任意概率分布(例如,平均值为0且标准偏差为1的标准高斯分布)。

[0196] 例如,生成神经网络406的架构可以源自于变分自编码器的架构,特别是DRAW生成神经网络的架构,如参考K.Gregor,I.Danihelka,A.Graves,et al.:“DRAW:a recurrent neural network for image generation”,arXiv:1502.04623v2,2015所描述的。在该示例中,生成神经网络架构可以包括循环解码器子网络。在多个内部时间步的每一个处,生成神经网络406可以被配置为根据先验概率分布对潜变量进行采样,并且在内部时间步内,将所采样的潜变量作为输入提供给循环解码器子网络。循环解码器子网络可以被配置为使用采样的潜变量和网络输入410来更新循环解码器子网络的内部状态。生成神经网络406可以被配置为在时间步内,使用循环解码器子网络的更新的内部状态来更新生成神经网络406的“画布(canvas)”内部状态的值。在最后一个内部时间步之后,生成神经网络406的画布内部状态的值定义了与每个结构参数404相对应的概率分布。例如,画布内部状态的值可以包括定义每个结构参数404的相应概率分布的参数的相应值。参考图21,示出了DRAW生成神经网络406的示例性架构。

[0197] 在一些情况下,生成神经网络406可以使用一个或多个卷积层(例如,2D卷积层),然后是池化层来处理网络输入410中包括的比对特征412。

[0198] 可以基于不同蛋白的实际结构(例如,通过实验确定的)的数据库308,由训练引擎414来训练生成神经网络406。更具体地,可以使用机器学习训练技术(例如,随机梯度下降)来重复地更新生成神经网络权重的值,以使生成神经网络406通过处理网络输入410来生成定义现实结构片段402的结构参数404。

[0199] 生成神经网络406,以及更一般地,片段生成系统310,可以被用作结构预测系统100中的本地存储器更新系统310的一部分,如前所述。然而,本说明书中的描述不应当被解释为将生成神经网络406和片段生成系统310限制为在结构预测系统100中的本地存储器更新系统310内使用。

[0200] 图5是示例性评分系统136的框图。评分系统136是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下文所述的系统、部件和技术。

[0201] 结构预测系统100中的搜索计算单元使用评分系统136来确定预测结构的相应质量得分。评分系统136接收输入,该输入包括:(i) 氨基酸序列102,和(ii) 氨基酸序列102的预测结构502(例如,如参考图1所述的替代预测结构)。评分系统136被配置为处理输入以生成预测结构502的质量得分504,该质量得分504通常表征预测结构的质量。例如,系统136可以使用下述中的一个或多个来生成质量得分:(i) 几何得分506、(ii) 值得分508、(iii) 距离似然性得分510、(iv) 其他附加得分512(所有这些都将下文进行详细描述)。

[0202] 如参考图1所述,可以以任何适当的数字格式表示氨基酸序列102和预测结构502。例如,氨基酸序列102可以被表示为单热编码向量的序列,其中每个单热编码向量表示氨基酸序列102的不同氨基酸。预测结构502可以由结构参数(例如,氨基酸序列102中的氨基酸的主链原子之间的扭转角)集合的值表示。

[0203] 任选地,系统136可以包括序列比对引擎514,例如,HHblits (Remmert等,Nature Methods 9,173,2012),其被配置为处理氨基酸序列102以生成比对特征516,比对特征516由来自其他蛋白的氨基酸序列与氨基酸序列102的相似序列的多序列比对(MSA)(如参考图4所述)中得出。

[0204] 任选地,系统136可以包括距离引擎518,其被配置为处理预测结构502以生成结构距离图520。当根据预测结构502折叠氨基酸序列102时,结构距离图520表征了氨基酸序列102中的每一对氨基酸之间的相应距离(例如,以埃为单位测量)。结构中的第一氨基酸与第二氨基酸之间的距离是指当根据该结构折叠氨基酸序列102时,第一氨基酸与第二氨基酸中的特定原子(例如主链原子)之间的距离。在一个示例中,结构距离图可以直接指定氨基酸序列102中的每一对氨基酸之间的距离。在另一示例中,结构距离图可以指定氨基酸序列102中的每一对氨基酸是否接触,即是否被小于预定阈值的距离隔开。结构距离图520可以以任何适当的数字格式表示,例如,表示为矩阵,其中矩阵的(i,j)条目表示氨基酸序列102中的第i个氨基酸与氨基酸序列102中的第j个氨基酸之间的距离。

[0205] 该系统使用下述的一项或多项:(i) 氨基酸序列102的表示,(ii) 定义预测结构502的结构参数,(iii) 比对特征516,以及(iv) 结构距离图520,以生成用于几何神经网络524、值神经网络526和距离预测系统528的相应输入522。在一些情况下,氨基酸序列102的表示和定义预测结构502的结构参数被表示为一维(1D)特征(即,表示为线性序列的特征),而一些比对特征516和结构距离图520被表示为二维特征(2D)(即,被表示为矩阵的特征)。为了生成输入522,系统136可以通过沿着2D特征的矩阵表示的行和列广播和级联1D特征来组合1D和2D特征。

[0206] 系统136根据几何神经网络权重的当前值处理为几何神经网络524生成的相应输入522,以生成指示几何得分506的输出。几何得分506是氨基酸序列102的预测结构502与氨基酸序列102的实际结构之间的相似性度量的估计。例如,相似性度量可以是定义氨基酸序

列102的预测结构502的结构参数与定义氨基酸序列102的实际结构的结构参数之间的均方根偏差(RMSD)。作为另一示例,相似性度量可以是全局距离测试(GDT)相似性度量。GDT相似性可以被定义为氨基酸序列102中这样的氨基酸的主链原子的分数,所述氨基酸在预测结构502中的位置落入其在实际结构中的位置的预定距离内。

[0207] 在一些实施方式中,几何神经网络524被配置为直接输出几何得分506。在一些实施方式中,几何神经网络524被配置为生成输出,该输出定义了了几何得分的预定集合上的概率分布。在这些实施方式中,几何神经网络524的输出包括在几何得分的预定集合中的几何得分的相应概率值。系统136可以将几何得分506确定为在几何得分的预定集合上的概率分布的集中趋势(例如,平均值、中位数或模式)的度量。

[0208] 系统136根据值神经网络权重的当前值处理为值神经网络526生成的相应输入522,以生成指示值得分508的输出。如果在当前搜索迭代中,由搜索计算单元维护的当前预测结构是预测结构502,则值得分508是表征在未来的搜索迭代中由搜索计算单元生成的预测结构的的质量的质量得分的估计。在一些实施方式中,未来搜索迭代是在当前搜索迭代之后的预定数量的迭代。在一些实施方式中,未来搜索迭代是搜索计算单元的最后一个搜索迭代(例如,满足搜索终止标准的搜索迭代)。值得分508使得使用评分系统136的搜索计算单元能够以可以在短期内(例如,经过几个搜索迭代)降低当前预测结构的的质量的方式更新其当前预测结构,但是这可能导致更高质量的最终预测结构。

[0209] 在一些实施方式中,值神经网络526被配置为直接输出值得分508。在一些实施方式中,值神经网络526被配置为生成输出,该输出定义了预定值得分集合上的概率分布。在这些实施方式中,值神经网络526的输出包括预定值得分集合中的每个值得分的相应概率值。系统136可以将值得分508确定为在预定值得分集合上的概率分布的集中趋势(例如,平均值、中位数或模式)的度量。

[0210] 在一些实施方式中,评分系统136可以生成用于几何神经网络524和值神经网络526的相应输入,所述输入包括表征氨基酸序列102的实际结构中的氨基酸残基对之间的估计距离的距离图。可以通过包括在距离预测系统528中的距离预测神经网络来生成距离图,如将参考图6进一步所述。

[0211] 几何神经网络524和值神经网络526可以以任何适当的神经网络配置来实施。例如,几何神经网络524和值神经网络526可以包括多个卷积神经网络层、关注层和残差块(例如,2D残差卷积块)。在一些情况下,卷积层可以包括扩张卷积滤波器以增加其相应感受野的大小。在特定的示例中,几何神经网络524和值神经网络526可以具有包括一系列多个(例如200个)卷积残差块(例如,如参考图20所述),随后是输出具有与氨基酸序列的长度无关的维度的单一向量的平均池化层,接着是具有用于范围[0,100]的100个bin的soft-max层的架构。

[0212] 可以基于包括多个训练示例的训练数据集合,使用标准机器学习训练技术(例如,随机梯度下降)来训练几何神经网络524。每个训练示例可以包括:(i)蛋白的训练预测结构,和(ii)目标几何得分,该目标几何得分是蛋白的训练预测结构与蛋白的实际结构之间的相似性度量。训练预测结构的目标几何得分表示通过处理与所述训练预测结构相对应的输入522,应当由几何神经网络生成的几何得分。例如,可能已经通过物理实验确定了蛋白的实际结构。

[0213] 在一些实施方式中,在静态训练数据集合上训练几何神经网络524,该静态训练数据集合在几何神经网络524的整个训练期间保持固定。在这些实施方式中,可以例如通过随机地扰动蛋白的实际预测结构来确定训练示例中包括的训练预测结构。在一些实施方式中,在几何神经网络524的整个训练中被重复地更新的训练数据集合上训练几何神经网络524。在这些实施方式中,在任何给定的训练迭代处,包括在训练示例中的一些训练预测结构可以是由搜索计算单元根据几何神经网络权重的当前值,使用基于由几何神经网络524生成的几何得分506的质量得分生成的预测结构。通过在整个训练期间重复地更新训练数据集合,可以训练几何神经网络524以识别和校正由搜索计算单元使用根据几何神经网络权重的当前值生成的质量得分生成的不准确的预测结构。

[0214] 可以基于包括多个训练示例的训练数据集合,使用标准机器学习训练技术(例如,随机梯度下降)来训练值神经网络526。每个训练示例可以包括:(i)蛋白的训练预测结构,和(ii)目标值得分,该目标值得分是通过重复地更新蛋白的训练预测结构而确定的表征蛋白的未来预测结构的的质量的质量得分。训练预测结构的目标值得分表示应该由值神经网络526通过处理与训练预测结构相对应的输入522生成的值得分。

[0215] 在一些实施方式中,在静态训练数据集合上训练值神经网络526,该静态训练数据集合在值神经网络526的整个训练期间保持固定。在这些实施方式中,可以通过使用搜索计算单元(例如,如参考图1所述),为不同蛋白计算大量的预测结构轨迹(即,预测结构序列)来确定训练数据集合。为了确定特定的训练示例,可以将来自预测结构轨迹中的特定预测结构选择为特定训练示例的训练预测结构。可以将特定训练示例的目标质量值得分选择为预测结构轨迹中的后续预测结构的质量得分。例如,可以将目标值得分选择为预测结构轨迹中的最后预测结构的质量得分。

[0216] 在一些实施方式中,在训练数据集合上训练值神经网络526,该训练数据集合在值神经网络526的整个训练中被重复地更新。在这些实施方式中,可以通过根据值神经网络权重的当前值,使用基于由值神经网络526生成的值得分508的质量得分,生成用蛋白的大量预测结构轨迹,来更新训练数据集合。可以通过由这些预测结构轨迹生成新的训练示例来更新训练数据集合(例如,使用用于从预测结构轨迹确定训练示例的先前描述的方法)。通过在整个训练过程中重复地更新训练数据集合,可以训练值神经网络526来识别和校正由搜索计算单元使用根据值神经网络权重的当前值生成的质量得分生成的不准确的预测结构。

[0217] 通常,用于训练几何神经网络524的训练数据集合包括训练示例,该训练示例具有对应于蛋白的训练预测结构,对该蛋白,该蛋白的“真相(ground-truth)”结构是已知的(例如通过物理实验)。具体来说,必须知道蛋白的真相结构,才能确定训练示例的目标几何得分。相反,用于训练值神经网络的训练数据集合可以包括训练示例,该训练示例具有与蛋白相对应的训练预测结构,对该蛋白,该蛋白的真相结构可能是未知的。具体而言,不需要知道蛋白的真相结构即可确定训练示例的目标值得分。

[0218] 可以使用对比发散训练程序来训练几何神经网络524和值神经网络526。在这种情况下,由几何神经网络生成的几何得分可能不是蛋白的预测结构502与实际结构之间的特定相似性度量的直接估计。类似地,由值神经网络生成的值得分可能不是未来预测结构的特定质量得分的直接估计。本说明书中对几何神经网络524和值神经网络526的描述应当被

理解为包括使用对比发散训练程序来训练几何神经网络526和值神经网络526的情况。

[0219] 在训练几何神经网络和值神经网络期间,可以使用任何适当的监督损失目标函数,例如,交叉熵损失目标函数或平方误差损失目标函数。在特定示例中,可以使用逻辑损失目标函数来训练被配置为生成相对几何得分的几何神经网络,该逻辑损失目标函数表征由相对重要性得分定义的预测结构的排名是否与预测结构的实际排名一致。

[0220] 系统136使用距离预测系统528来处理为距离预测系统528生成的相应输入522,以生成距离似然性得分510。如将参考图6进一步描述,距离似然性得分510基于下述(i)和(ii)之间的差异来定义预测结构502的似然性:(i)预测结构502中的氨基酸序列102中的氨基酸对之间的距离,以及(ii)氨基酸序列102的实际结构中的氨基酸对之间的估计距离。

[0221] 除了几何得分506、值得分508和距离似然性得分510之外,系统136还可以生成附加得分512。例如,附加得分可以包括基于如果根据预测结构502折叠氨基酸序列102,预测结构502的疏水部分是否会暴露于水的得分。在该示例中,包括会暴露于水的疏水区的预测结构502将导致较低的得分,因为该预测结构502可能与氨基酸序列102的实际结构不同。作为另一示例,附加得分可以包括基于下述(i)和(ii)之间的相似性度量的得分:(i)预测结构502,和(ii)被生成成为单次(one-shot)预测神经网络的输出的蛋白的预测结构。在该示例中,单次预测神经网络可以被配置为处理包括氨基酸序列102的表示和比对特征516的输入以直接生成蛋白的预测结构。

[0222] 系统136可以通过组合以下各项中的一项或多项来确定预测结构502的质量得分504:几何得分506、值得分508、距离似然性得分510和附加得分512。例如,系统136可以将质量得分确定为几何得分506、值得分508、距离似然性得分510和附加得分512的线性组合,其中线性组合的系数是可调节的系统超参数。

[0223] 如前所述,可以将几何神经网络524、值神经网络526和距离预测系统528用作评分系统136的一部分。但是,本说明书中的描述不应当解释为将几何神经网络524、值神经网络526和距离预测系统528限制到在评分系统136内使用。此外,如下文将更详细所述,可以实施几何神经网络524和值神经网络526的不同变形,以在评分系统136之内或之外使用。

[0224] 在一些实施方式中,几何神经网络524可以被配置为处理输入522,该输入包括:(i)定义蛋白的氨基酸序列的数据,(ii)定义蛋白的第一预测结构的数据,和(iii)定义蛋白第二预测结构的数据。通常,蛋白的第一预测结构不同于蛋白的第二预测结构。定义蛋白的第一预测结构的数据和定义蛋白的第二预测结构的数据可以包括结构参数的相应值(例如,蛋白的氨基酸序列中的主链原子的扭转角或坐标)。此外,几何神经网络524的输入522可以包括比对特征516和与蛋白的第一和第二预测结构相对应的相应结构距离图520。几何神经网络524可以被配置为处理输入,以生成表征蛋白的第一预测结构与蛋白的第二预测结构之间的相对几何得分的输出。

[0225] 相对几何得分可以定义关于蛋白的第一预测结构与实际结构之间的相似性度量是否超过蛋白的第二预测结构与实际结构之间的相似性度量的预测。即,相对几何得分可以定义关于第一预测结构还是第二预测结构更准确的预测。在特定的示例中,相对几何得分可以为正(或者可以为负),以表示蛋白的第一预测结构与实际结构之间的相似性度量超过(或者不超过)蛋白的第二预测结构与实际结构之间的相似性度量的预测。

[0226] 当评分系统136正由搜索计算单元使用时,在每个搜索迭代中,几何神经网络524

可以被配置为对每个替代预测结构生成相对几何得分。为了对替代预测结构生成相对几何得分,几何神经网络524可以被配置为共同处理定义在搜索迭代中搜索计算单元的当前预测结构的数据和定义替代预测结构的数据。此后,可以使用相对几何得分(如前所述)来确定替代预测结构的质量得分504。

[0227] 在一些实施方式中,值神经网络526可以被配置为处理输入522,输入522包括:(i)定义蛋白的氨基酸序列的数据,(ii)定义蛋白的第一预测结构的数据,和(iii)定义蛋白的第二预测结构的数据。通常,蛋白的第一预测结构不同于蛋白的第二预测结构。定义蛋白的第一预测结构的数据和定义蛋白的第二预测结构的数据可以包括结构参数的相应值(例如,蛋白的氨基酸序列中的主链原子的扭转角或坐标)。此外,值神经网络526的输入522可以包括比对特征516和与蛋白的第一和第二预测结构相对应的相应结构距离图520。值神经网络526可以被配置为处理输入以生成表征蛋白的第一预测结构与蛋白的第二预测结构之间的相对值得分的输出。

[0228] 相对值得分可以定义关于表征蛋白的第一未来预测结构的的质量的质量得分是否超过表征蛋白的第二未来预测结构的的质量的质量得分的预测。蛋白的第一未来预测结构是指通过在多个搜索迭代中迭代地修饰蛋白的第一预测结构(例如,由搜索计算单元)而生成的预测结构。蛋白的第二未来预测结构是指通过在多个搜索迭代中迭代地修饰蛋白的第二预测结构(例如,由搜索计算单元)而生成的预测结构。也就是说,相对值得分可以定义关于迭代地修饰(例如,由搜索计算单元)第一预测结构还是第二预测结构可以导致“更好”(例如,更准确)的未来预测结构的预测。在特定示例中,相对值得分可以为正(或者为负),以指示蛋白的第一未来预测结构的质量得分超过(或者不超过)蛋白的第二未来预测结构的质量得分的预测。

[0229] 当评分系统136正由搜索计算单元使用时,在每个搜索迭代中,值神经网络526可以被配置为对每个替代预测结构生成相对值得分。为了对替代预测结构生成相对值得分,值神经网络526可以被配置为共同处理定义在搜索迭代中搜索计算单元的当前预测结构的数据和定义替代预测结构的数据。此后,可以使用相对值得分(如前所述)来确定替代预测结构的质量得分504。

[0230] 如前所述,预测结构的主链原子质量得分仅基于氨基酸中的主链原子,而全原子质量得分是参考每个氨基酸中的所有原子确定的。评分系统136可以包括被用来生成全原子质量得分的几何神经网络、值神经网络和距离预测系统的相应实施方式,以及被用来生成主链原子质量得分的几何神经网络、值神经网络以及距离预测系统的相应实施方式。例如,几何神经网络的全原子实施方式可以被配置为处理预测结构502,该结构指定氨基酸序列102中的每个氨基酸中的每个原子的位置,以生成全原子几何得分。全原子几何得分可以基于蛋白的每个氨基酸中的每个原子来定义蛋白的预测结构502与实际结构之间的相似性度量。几何神经网络的主链原子实施方式可以被配置为处理预测结构502,该结构指定氨基酸102中的氨基酸的主链原子的位置,以生成主链原子几何得分。主链原子几何得分可以基于蛋白的每个氨基酸中的主链原子来定义蛋白的预测结构502与实际结构之间的相似性度量。

[0231] 图6是示例性距离预测系统528的框图。距离预测系统528是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下述系统、部件和技术。

[0232] 距离预测系统528被配置为接收输入602(例如,如参考图5所述),输入602包括氨基酸序列102的表示、比对特征604和氨基酸序列102的预测结构606。系统528被配置为处理输入602以生成距离似然性得分510。距离似然性得分510基于下述(i)和(ii)之间的差异来定义预测结构606的似然性:(i)预测结构606中的氨基酸序列102中的氨基酸对之间的距离,和(ii)氨基酸序列102的实际结构中的氨基酸序列102中的氨基酸对之间的估计距离。

[0233] 比对特征是基于氨基酸序列的MSA来确定的,并且可以包括具有每个残基的位置特异性取代概率的表达谱(profile)特征(可能是1D)以及协变特征((例如,在MSA上拟合的经过正规伪似然性训练的Potts模型(例如,与CCMPred类似)的参数)。这些特征可以包括Potts模型的参数的弗罗贝纽斯范数(Frobenius norm)和每个残基对*i, j*的原始参数值。比对特征可以包括明确表示MSA中的间隙和缺失的特征,例如其中第(*i, j*)个元素对在第*i*和第*j*个位置中出现间隙的次数进行计数的间隙矩阵(间隙越大,协变的方差越大),和/或缺失概率,例如,在残基位置右侧发生缺失的概率。

[0234] 可以使用PSI-BLAST来提取其他表达谱特征。MSA中的氨基酸序列的数量可能是其他比对特征。比对特征和氨基酸序列的表示可以被表示为特征的二维阵列,其中每个*i, j*特征是*i*和*j*两个残基的一维特征以及残基对*i, j*的二维特征的级联。在此描述的比对特征可以被用作本说明书中描述的每个评分神经网络例如几何神经网络、值神经网络和结构预测神经网络的输入。

[0235] 因此,在特定示例性实施方式中,距离预测系统528的输入特征可以包括MSA中的氨基酸序列的数量;包括氨基酸序列102的表示(21D)的序列长度特征、缺失概率(1D)、残基索引和表达谱,诸如PSI-BLAST表达谱(21D)、HHblits表达谱(22D)、HHblits偏差、无间隙表达谱(21D)、HMM表达谱(30D)和Potts模型偏差(30D);以及序列长度特征,诸如间隙矩阵(1D)、Frobenius范数(1D)和Potts模型参数(484D)。

[0236] 为了生成距离似然性得分510,系统528生成距离图608,该距离图608表征了氨基酸序列102的实际(例如,通过实验确定的)结构中的氨基酸序列102中的每一对氨基酸之间的估计距离。距离图608可以被表示为矩阵,其中矩阵的(*i, y*)条目包括表征氨基酸序列102中的第*i*个氨基酸与氨基酸序列102的实际结构中的氨基酸序列102中的第*y*个氨基酸之间的估计距离的数据。在一些实施方式中,对于氨基酸序列102中的每一对氨基酸,距离图608将氨基酸对之间的估计距离表征为二进制变量。在这些实施方式中,如果氨基酸对之间的估计距离小于预定阈值(例如,8埃),则二进制变量可以具有值1,否则,二进制值可以具有值0。在一些实施方式中,对于氨基酸序列102中的每一对氨基酸,距离图608将氨基酸对之间的估计距离表征为连续值数量(例如,以埃为单位表示距离)。在一些实施方式中,对于氨基酸序列102中的每一对氨基酸,距离图608通过在预定距离范围集合内的概率分布(例如,概率分布610)来表征该对氨基酸之间的估计距离。在这些实施方式中,概率分布包括预定距离范围集合中的每个距离范围的相应概率值。例如,在一个特定的实施方式中,该对氨基酸之间的距离是氨基酸残基的C_β原子(对于甘氨酸来说为C_α,缺少β碳)之间的距离。在一个特定的实施方式中,将距离范围2-22Å量化为64个相等的bin(如稍后在图19中所示)。

[0237] 为了生成距离图608,系统528可以生成距离图裁剪612的集合。每个距离图裁剪612是全距离图608的适当子集的估计。更具体地,每个距离图裁剪612表征(i)氨基酸序列102的一个或多个第一位置的每一个中的氨基酸残基与(ii)氨基酸序列102的一个或多个

第二位置的每一个中的氨基酸残基之间的估计距离,其中第一位置、第二位置或两者均为氨基酸序列102中的总位置的适当子集。第一位置可以被理解为鉴定距离图矩阵608的相应行,并且第二位置可以被理解为鉴定距离图矩阵608的相应列。如下文将更详细所述,系统528使用融合引擎614来组合距离图裁剪612以生成全距离图608。

[0238] 将每个距离图裁剪生成成为距离预测神经网络616的输出。距离预测神经网络616被配置为通过处理裁剪特征618来生成距离图裁剪,裁剪特征618表征与距离图裁剪相对应的氨基酸序列中的第一位置和第二位置的每一个中的氨基酸。裁剪特征618可以由裁剪特征生成引擎624生成。裁剪特征生成引擎624被配置为提取与对应于距离图裁剪的氨基酸序列102中的第一位置和第二位置中的氨基酸相对应的下述(i)和(ii)的分量:(i)氨基酸序列102的表示和(ii)比对特征604。在一些情况下,裁剪特征生成引擎另外被配置为提取与氨基酸序列中的第一位置中的氨基酸和第二位置中的氨基酸相对应的比对特征的对角线分量。使用接近对角线($i = j$)的特征可以帮助编码局部结构,例如二级结构。

[0239] 如果距离预测神经网络616被配置为通过处理氨基酸序列102的整个表示和比对特征604来直接生成距离图608,则距离预测神经网络616的架构将受到必须建模的最长氨基酸序列的限制。通过处理裁剪特征618以生成距离图裁剪612(被系统528融合以生成全距离图608),距离预测神经网络616可以具有比其他方式更复杂的架构(例如,具有更多的神经网络层),从而可以实现更精确的距离图608估计。

[0240] 距离预测神经网络616可以以任何适当的神经网络配置来实施。例如,距离预测神经网络616可以包括多个卷积神经网络层、关注层和残差块。在一些情况下,卷积层可以包括扩张卷积滤波器以增加其相应感受野的大小。在特定示例中,距离预测神经网络可以是深度二维扩张卷积残差网络。在该示例中,距离预测网络始终是二维的,并且使用带有扩张卷积的220个残差块。参考图20更详细地描述残差块的示例性架构。

[0241] 可以基于包括多个训练示例的训练数据集合来训练距离预测神经网络616。每个训练示例包括训练网络输入和与训练网络输入相对应的目标距离图。训练网络输入源自于具有已知结构的训练蛋白。在一些实施方式中,目标距离图表征训练蛋白的结构中氨基酸残基之间的实际距离。在一些其他实施方式中,目标距离图表征通过处理训练输入,由“教师”神经网络生成的训练蛋白的结构中的氨基酸残基之间的估计距离。在这些实施方式中,教师神经网络可以是比距离预测神经网络616更复杂的神经网络(例如,具有更多层、更多参数或两者兼有的神经网络),其被训练以生成表征蛋白的结构中的氨基酸残基之间的实际距离的距离图。通过训练距离预测神经网络616以生成与教师神经网络的输出匹配的输出生(即,在蒸馏学习框架中),可以比其他方式更有效地训练距离预测神经网络616。用于训练距离预测神经网络616的目标函数可以是交叉熵目标函数。

[0242] 在一些情况下,系统528可以通过生成包括训练距离图的新训练示例来扩充训练数据集合,该训练距离图从最初包括在训练数据集合中的训练距离图随机地扰动。在一些情况下,系统528可以通过生成新的训练示例来扩充训练数据集合,其中使用氨基酸序列的完整MSA的随机子采样来生成训练示例中包括的比对特征。

[0243] 系统528包括裁剪鉴定引擎620,其被配置为选择要由距离预测神经网络616生成的距离图裁剪612。更具体地说,裁剪鉴定引擎620生成包括与每个距离图裁剪(例如,距离图608的距离图裁剪628)相对应的第一位置和第二位置的输出。在一些实施方式中,裁剪鉴

定引擎620被配置为对于每个距离图裁剪生成随机选择的第一位置和第二位置。如果在训练期间应用,则该方法可以充当数据扩充的一种形式。在一些实施方式中,裁剪鉴定引擎620被配置为鉴定预定的距离图裁剪集合,选择该距离图裁剪集合以“覆盖”距离图608(即,使得氨基酸序列102中的每一可能的氨基酸对之间的距离由至少一个距离图裁剪612表征)。任选地,距离预测神经网络的输入可以包括裁剪的相对和/或绝对位置的编码。

[0244] 在一个示例中,裁剪鉴定引擎620选择尺寸为 64×64 的距离图裁剪,即,表征了两组64个连续残基之间的成对距离的距离图裁剪。但是,裁剪不必是正方形的。

[0245] 在生成距离图裁剪612之后,系统528使用融合引擎614“融合”(即,组合)距离图裁剪612以确定全距离图608。在一些实施方式中,系统528将全距离图608确定为距离图裁剪612的平均值。例如,对于氨基酸序列102中的给定氨基酸对,系统528可以将由距离图608表征的给定氨基酸对之间的估计距离的分布确定为由每个距离图裁剪612表征的给定氨基酸对之间的估计距离的分布的平均值。在该示例中,平均值不包括不表征给定氨基酸对之间的距离的距离图裁剪。同样在该示例中,平均值可能更偏重于更以给定氨基酸对为中心的距离图裁剪。这可以帮助减少边缘效应。在一些实施方式中,系统528通过根据融合神经网络参数的当前值,使用融合神经网络处理距离图裁剪612来确定全距离图608,以生成包括全距离图608的输出。

[0246] 在一些情况下,系统528使用多个距离预测神经网络来生成距离图裁剪612的集合,所述多个距离预测神经网络以不同的超参数独立地被训练。使用这样的系综可以进一步提高准确性。

[0247] 该系统包括评估引擎622,该评估引擎被配置为由预测结构606和距离图608确定距离似然性得分626,例如距离似然性得分510。例如,评估引擎622可以基于对于氨基酸序列102中的每一对氨基酸,对根据距离图608的氨基酸对之间的距离的连续值估计与预测结构606中的该对氨基酸之间的距离之间的平方差的和,确定距离似然性得分626。作为另一示例,评估引擎622可以基于对于氨基酸序列102中的每一对氨基酸,根据距离图608该对氨基酸被由预测结构606定义的距离隔开的概率来确定距离似然性得分626。在特定示例中,评估引擎622可以将距离似然性得分 s 确定为:

$$[0248] \quad s = \prod_{(i,j)} p_{i,j}(d_{i,j}) \quad (1)$$

[0249] 其中乘积是在由 (i,j) 索引的氨基酸序列102中的氨基酸对上,并且 $p_{i,j}(d_{i,j})$ 表示根据在由距离图608定义的该对氨基酸残基 (i,j) 之间的可能距离范围内的相应概率分布 $p_{i,j}$,由 (i,j) 索引的氨基酸对被由预测结构606定义的距离 $d_{i,j}$ 隔开的概率。作为另一示例,评估引擎622可以部分地基于“参考”距离图来确定距离似然性得分,如下文将更详细地描述。

[0250] 参考距离图表征氨基酸序列102中的每一对氨基酸之间通常期望的估计距离,但是该估计距离是在不参考氨基酸序列102中的特定氨基酸的同一性的情况下确定的。例如,对于氨基酸序列102中的每一对氨基酸,参考距离图可以基于氨基酸对中的氨基酸的位置和相对偏移来表征该对氨基酸之间的估计距离。给定氨基酸的位置是指氨基酸序列中在给定氨基酸与第一氨基酸之间的其他氨基酸的数量。两个氨基酸之间的相对偏移是指氨基酸序列中的两个氨基酸之间的其他氨基酸的数量。参考距离图可以以与距离图608相同的方

式表征氨基酸对之间的估计距离(例如,作为连续值数量或作为在预定的距离范围集合内的概率分布)。

[0251] 距离预测系统528可以使用不同蛋白的实际结构的蛋白结构数据库来生成参考距离图。在特定示例中,距离预测系统528可以基于包括在蛋白结构数据库的相应蛋白结构中的具有相同位置和相对偏移的每一对氨基酸,确定在一对氨基酸之间的预定距离范围集合内的参考概率分布。在该示例中,参考距离图可以包括相应参考概率分布,其表征氨基酸序列102中的每一对氨基酸之间的估计距离。

[0252] 评估引擎622可以使用参考距离图来确定距离似然性得分626,如下:

$$[0253] \quad s = \log \prod_{(i,j)} p_{i,j}(d_{i,j}) - \log \prod_{(i,j)} p_{i,j}^r(d_{i,j}) \quad (2)$$

[0254] 其中乘积是在由(i,j)索引的氨基酸序列102中的氨基酸对上, $p_{i,j}(d_{i,j})$ 表示根据由距离图608定义的氨基酸残基对(i,j)之间的可能距离范围内的相应概率分布 $p_{i,j}$,由(i,j)索引的氨基酸对被由预测结构606定义的距离 $d_{i,j}$ 隔开的概率, $p_{i,j}^r(d_{i,j})$ 表示根据由参考距离图(如上所述)定义的氨基酸残基对(i,j)之间的可能距离范围内的相应参考概率分布 $p_{i,j}^r$,由(i,j)索引的氨基酸对被由预测结构606定义的距离 $d_{i,j}$ 隔开的概率。通过使用参考距离图确定距离似然性得分626,距离似然性得分626可以表征参考距离图和由距离预测神经网络616生成的距离图之间的偏差(如进一步参考图6所述)。例如,如稍后所述,可以在对数结构域中从使用氨基酸同一性确定的距离似然性得分中减去使用参考距离图确定的距离似然性得分626,以校正先验距离分布的过度表示。

[0255] 通常,距离似然性得分626可以是来自预测结构606和表征预测结构的的质量的距离图608确定的任何数值。

[0256] 通过使用连续值的距离估计值或距离范围概率分布(即,而不是二进制变量)来表征氨基酸对之间的距离,系统528可以生成距离似然性得分626,该距离似然性得分传达关于预测结构606与氨基酸序列102的实际结构的符合程度的更精确的信息。

[0257] 在一些情况下,距离预测神经网络616可以被配置为生成附加的辅助输出(例如,除了距离图裁剪之外)。例如,距离预测神经网络可以被配置为生成表征与距离图裁剪相对应的氨基酸序列中的第一位置和第二位置中的每一个中的氨基酸之间的扭转角的输出。作为另一个示例,距离预测神经网络可以被配置为生成表征与对应于距离图裁剪的氨基酸序列中的第一位置和第二位置中的每一个的氨基酸相对应的估计的二级结构(β 折叠或 α -螺旋二级结构和/或卷曲结构)的输出。作为另一个示例,距离预测神经网络可以被配置为生成表征蛋白的可及表面积(即,溶剂可及的蛋白的表面积)的输出。训练距离预测神经网络以准确地生成附加的辅助输出可能使得距离预测神经网络生成更准确的距离图裁剪。

[0258] 图7是示例性迭代过程的流程图,该示例性迭代过程用于通过迭代地更新蛋白的当前预测结构来生成蛋白的预测结构的轨迹(即序列)。为了方便起见,将过程700描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如,根据本说明书适当编程的搜索计算单元,例如图1的搜索计算单元108和110可以执行过程700。

[0259] 该系统使用被配置为存储结构片段的集合的本地存储器来更新蛋白的当前预测结构(702)。每个结构片段表征蛋白的氨基酸序列中的氨基酸的子序列(即片段)的预测结

构。为了更新蛋白的当前预测结构,系统使用来自本地存储器的结构片段,通过“扰动”蛋白的当前预测结构,生成蛋白的替代预测结构。参考图8,描述了用于确定是否将当前预测结构更新为替代预测结构的示例性过程。

[0260] 系统确定是否满足搜索终止标准(704)。作为示例,当系统已经执行了过程700的步骤的预定数量的迭代时,系统可以确定满足搜索终止标准。作为另一示例,如果在预定数量的迭代(例如,前十个迭代)中,当前预测结构的质量得分的平均增加低于某一阈值,则系统可以确定满足搜索终止标准。预测结构的质量得分通常表征预测结构的质量。例如,预测结构的质量得分可以表征蛋白的预测结构与实际结构之间的估计相似性、基于预测结构的氨基酸中的主链原子之间的距离的预测结构的似然性,或两者。参考806进一步描述了确定预测结构的质量得分。

[0261] 响应于确定满足搜索终止标准,系统确定蛋白的最终预测结构(706)。为了确定蛋白的最终预测结构,系统选择存储在中央存储器中的预测结构。通常,存储在中央存储器中的每个预测结构都与得分相关联,并且系统通过基于得分来选择中央存储器中的预测结构而确定最终预测结构。例如,系统可以将最终预测结构确定为存储在中央存储器中的具有最高得分的预测结构。与存储在中央存储器中的预测结构相关联的得分可以是例如主链原子质量得分或全原子质量得分(如前所述)。

[0262] 响应于确定不满足搜索终止标准,系统通过当前预测结构确定是否满足中央存储器更新条件(708)。如果例如当前预测结构的质量得分是直到当前迭代为止系统生成的预测结构轨迹中的最高质量得分,则系统可以确定满足中央存储器更新条件。在该示例中,系统动态地更新中央存储器以存储最“有希望”的预测结构(例如,最有可能准确地近似于蛋白的实际结构)。

[0263] 响应于确定满足中央存储器更新条件,系统更新中央存储器(710)。中央存储器是被配置为存储预测结构的数据存储(例如,逻辑数据存储区或物理数据存储设备)。系统可以通过将当前预测结构存储在中央存储器中来更新中央存储器。将当前预测结构存储在中央存储器中是指将定义当前预测结构的结构参数的值存储在中央存储器中。

[0264] 系统确定是否满足本地存储器更新条件(712)。例如,如果自从上次系统确定满足本地存储器更新条件以来,系统已经执行了预定数量的迭代,则系统可以确定满足本地存储器更新条件。作为另一示例,如果在预定数量的迭代中,当前预测结构的质量得分的平均增加低于预定阈值,则系统可以确定满足本地存储器更新条件。

[0265] 响应于确定满足本地存储器更新条件,系统更新本地存储器(714)。在一些实施方式中,系统通过获得不同的结构片段并且用不同的结构片段替换当前在本地存储器中的结构片段来更新本地存储器。在一些实施方式中,系统通过获得不同的结构片段并且使用替换策略来部分替换当前在本地存储器中的结构片段来更新本地存储器。例如,替换策略可以是先进先出(FIFO)替换策略,其中最早包括在本地存储器中的结构片段被不同的结构片段替换。

[0266] 系统可以以各种不同的方式获得要包括在本地存储器中的结构片段。例如,如下文将更详细所述,该系统可以通过以下方式生成结构片段:(i)对存储在中央存储器中的预测结构进行分段,(ii)对存储在不同蛋白的已知结构的结构数据库中的实际结构进行分段,(iii)使用包括在片段生成系统中的生成神经网络生成结构片段,或(iv)其组合。

[0267] 在一些实施方式中,系统通过对从中央存储器采样的预测结构进行分段来获得要包括在本地存储器中的结构片段。系统可以根据与存储在中央存储器中的预测结构相关联的得分来从中央存储器采样预测结构。如前所述,存储在中央存储器中的每个预测结构可以与质量得分(例如,主链原子质量得分或全原子质量得分)相关联。作为示例,为了从中央存储器采样预测结构,系统可以使用它们的关联得分来确定存储在中央存储器中的预测结构的概率分布(例如,通过使用soft-max函数来处理得分)。然后,系统可以根据所确定的概率分布,从中央存储器采样预测结构。

[0268] 在一些实施方式中,系统通过对存储在结构数据库中的不同蛋白的实际结构进行分段来获得要包括在本地存储器中的结构片段。可能已经使用物理实验方法例如X射线晶体学确定了存储在结构数据库中的不同蛋白的实际结构。

[0269] 在一些实施方式中,系统使用经训练以生成现实的结构片段的生成神经网络来获得要包括在本地存储器中的结构片段。参考图4进一步描述了使用生成神经网络生成结构片段。

[0270] 在更新本地存储器之后,系统可以返回到702并重复前面的步骤。

[0271] 图8是用于确定是否将蛋白的当前预测结构更新为蛋白的替代预测结构的示例性过程的流程图。为了方便起见,将过程800描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如,根据本说明书适当编程的搜索计算单元(例如,图1的搜索计算单元108和110)可以执行过程800。

[0272] 该系统维护蛋白的当前预测结构和表征当前预测结构的的质量的当前质量得分(802)。当前预测折叠结构由结构参数集合的值定义。例如,结构参数可以是蛋白的氨基酸中的主链原子之间的扭转角序列。当前质量得分是表征当前预测结构的的质量的数值。

[0273] 该系统确定蛋白的替代预测结构(804)。为了确定蛋白的替代预测结构,系统从本地存储器获得结构片段。例如,系统可以从本地存储器中随机采样结构片段。在从本地存储器获得结构片段之后,系统可以通过所获得的结构片段来生成作为对当前预测结构的“扰动”的替代预测结构。更具体地,系统可以确定替代预测结构以包括当前预测结构的一部分和所获得的结构片段。

[0274] 系统为每个替代预测结构确定相应质量得分(806)。例如,系统可以使用以下一项或多项来确定质量得分:(i)几何得分、(ii)值得分、(iii)距离似然性得分、(iv)其他附加得分。例如,系统可以将质量得分确定为几何得分、值得分、似然性得分和一个或多个附加得分的线性组合,其中线性组合的系数是可调节的系统超参数。

[0275] 蛋白的预测结构的几何得分是蛋白的预测结构与蛋白的实际结构之间的相似性度量(例如,RMSD或GDT)的估计。为了确定替代预测折叠结构的几何得分,系统可以根据几何神经网络权重的当前值,使用几何神经网络处理表征替代预测折叠结构的网络输入。网络输入包括蛋白氨基酸序列的表示以及定义替代预测折叠结构的折叠结构参数的值。在一些实施方式中,几何神经网络被配置为直接输出几何得分。在一些实施方式中,几何神经网络被配置为生成输出,该输出定义了预定几何得分集合上的概率分布。在这些实施方式中,几何神经网络的输出包括预定几何得分集合中的几何得分的相应概率值。系统可以将几何得分确定为在预定几何得分集合上的概率分布的集中趋势(例如,平均值、中位数或模式)的度量。

[0276] 蛋白的预测结构的值得分是如果在当前搜索迭代中的当前预测结构(例如,由搜索计算单元维护的)是预测结构,表征在未来的搜索迭代(例如过程700的迭代)中生成(例如,由搜索计算单元生成)的预测结构的的质量的质量得分的估计。为了确定替代预测折叠结构的值得分,系统可以根据值神经网络权重的当前值,使用值神经网络处理表征替代预测折叠结构的网络输入。在一些实施方式中,值神经网络被配置为直接输出值得分。在一些实施方式中,值神经网络被配置为生成输出,该输出定义了预定值得分集合上的概率分布。在这些实施方式中,值网络的输出包括预定值得分集合中的每个值得分的相应概率值。系统可以将值得分确定为在预定值得分集合上的概率分布的集中趋势(例如,平均值、中位数或模式)的度量。

[0277] 蛋白的预测折叠结构的距离似然性得分基于以下(i)与(ii)之间的差异定义了预测结构的似然性:(i)蛋白的预测结构中的氨基酸对之间的距离,以及(ii)蛋白的实际结构中的氨基酸对之间的估计距离。为了确定替代预测折叠结构的距离似然性得分,系统可以使用距离预测系统(如参考图6所述的)处理氨基酸序列的表示,以及(任选地)对应于蛋白的由MSA得出的比对特征。

[0278] 除了几何得分、值得分和距离似然性得分之外,系统还可以为替代预测折叠结构确定附加得分。例如,附加得分可以包括基于如果根据预测结构折叠氨基酸序列,则预测结构的疏水部分是否会暴露于水的得分。

[0279] 基于当前预测结构和替代预测结构的相应质量得分,系统确定是否将当前预测结构更新为替代预测结构(808)。

[0280] 在一些实施方式中,系统可以基于质量得分,使用确定性程序来确定是否将当前预测结构更新为替代预测结构。例如,如果特定替代预测结构具有比当前预测结构和任何其他替代预测结构更高的质量得分,则系统可以确定将当前预测结构更新为特定替代预测结构。在该示例中,如果当前预测结构具有比任何替代预测结构更高的质量得分,则系统可以确定不将当前预测结构更新为任何替代预测结构。

[0281] 在一些实施方式中,系统可以基于质量得分,使用随机程序(即,涉及一些随机性)来确定是否将当前预测结构更新为替代预测结构。例如,系统可以使用质量得分来确定在包括当前预测结构和每个替代预测结构的结构集合上的概率分布。系统可以使用概率分布来确定将当前预测结构更新为从包括当前预测结构和每个替代预测结构的结构集合中采样的结构。

[0282] 在特定示例中,系统可以通过使用soft-max函数,处理当前预测结构和每个替代预测结构的相应质量得分来确定概率分布。在另一特定示例中,系统可以通过根据温度超参数,使用soft-max函数处理当前预测结构和每个替代预测结构的相应质量得分来确定概率分布。可以通过基于以下关系式,根据温度超参数 T ,使用soft-max函数,处理得分集合 $\{z_i\}_{i=1}^n$,确定由离散概率值 $\{q_i\}_{i=1}^n$ 定义的概率分布:

$$[0283] \quad q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_{j=1}^n \exp\left(\frac{z_j}{T}\right)} \quad (3)$$

[0284] 在该示例中,较高的温度参数导致更均匀的得分分布,因此导致系统可以将当前预测结构更新为质量得分低于当前预测结构的替代预测结构的似然性增加。

[0285] 通过根据升高的温度超参数,使用soft-max函数确定是否将当前预测结构更新为替代预测结构,系统可以“探索”可能的蛋白结构的空間。搜索计算单元(例如,如参考图1所述的)可以根据预定时间表来改变用于更新其当前预测折叠结构的温度的值。例如,预定时间表可以将温度值设置为初始高值(导致对可能结构空間的初始探索),该初始高值随着所执行的搜索迭代的数量增加而逐渐减小。

[0286] 图9是示例性结构预测系统900的框图。结构预测系统900是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下文描述的系统、部件和技术。

[0287] 结构预测系统900被配置为处理定义蛋白904的氨基酸序列902的数据,以生成蛋白904的最终预测结构906。氨基酸序列902中的每个氨基酸是有机化合物,其包括氨基官能团和羧基官能团以及对该氨基酸特异的侧链(即原子团)。最终预测结构906定义在蛋白904经历蛋白折叠之后,蛋白904的氨基酸序列902中的原子的三维构型的估计。蛋白折叠是指序列例如氨基酸的无规卷曲(例如,由蛋白904的氨基酸序列902定义的)折叠成独特的三维构型(例如,如由最终预测结构906所估计的)的物理过程。

[0288] 氨基酸序列902可以以任何合适的数字格式表示。例如,氨基酸序列902可以被表示为单热向量的序列。在该示例中,单热向量序列中的每个单热向量表示氨基酸序列902中的相应氨基酸。单热向量对于每种可能的氨基酸(例如,预定数量的可能氨基酸中的)具有不同的分量。表示特定氨基酸的单热向量在与该特定氨基酸相对应的分量中具有值1(或一些其他预定值),而在其他分量中具有值0(或一些其他预定值)。

[0289] 氨基酸序列902的结构(例如,系统900输出的最终预测结构906)由结构参数集合的值定义。在一些实施方式中,结构参数是三维(3D)数字坐标的序列(例如,表示为3D向量),其中每个坐标表示来自氨基酸序列902的氨基酸中的相应原子的位置(在一些给定参考系中)。例如,结构参数可以是表示该结构的氨基酸中的 α 碳原子的相应位置的3D数字坐标的序列。在本说明书中被称为主链原子的 α 碳原子是指氨基酸中与氨基官能团、羧基官能团和侧链键合的碳原子。在一些实施方式中,结构参数是结构中的氨基酸中的特定原子之间的扭转角(即,二面角)序列。例如,结构参数可以是结构中的氨基酸的主链原子之间的 $\phi(\varphi)$ 、 $\psi(\varphi)$ 以及任选地 $\omega(\omega)$ 二面角的序列。

[0290] 为了生成最终预测结构906,系统900使用优化系统908来生成蛋白904的多个预测结构910,每个预测结构910都是蛋白904的最终预测结构906的候选。在生成多个预测结构910之后,系统900使用选择引擎912来将预测结构910之一选择为最终预测结构906(将在下文更详细地描述)。

[0291] 为了生成预测结构910,优化系统908首先获得定义蛋白904的结构的结构参数集合的初始值。通常,优化系统908使用涉及一些随机性的过程来确定结构参数集合的初始值,从而使得优化系统908能够“探索”可能的预测结构的空間。在特定示例中,如果优化系统908先前已经为蛋白生成了一个或多个预测结构910,以确定结构参数的初始值,则优化系统908可以获得为蛋白定义先前生成的预测结构910的结构参数的值。随后,优化系统908可以通过使用随机噪声值,对定义先前生成的预测结构的结构参数的值进行扰动来确定结构参数的初始值(如将参考图11更详细地描述)。

[0292] 在确定了结构参数的初始值之后,优化系统908在多个更新迭代中,迭代地更新

(即,调整)结构参数的值。当优化系统908确定满足终止标准时,优化系统908在最终更新迭代之后输出由结构参数的当前值定义的预测结构910。

[0293] 优化系统908被配置为在多个更新迭代中更新结构参数的值,以生成具有高质量得分的预测结构。如下文将更详细所述,预测结构的质量得分表征了预测结构的质量,例如,预测结构与蛋白904的实际结构的符合程度。为了方便起见,在本说明书中,较高的质量得分将被理解为表征预测结构的较高质量。

[0294] 在每个更新迭代处,优化系统908使用评分系统916来处理当前结构参数值914和氨基酸序列902的表示,以生成表征由当前结构参数值914定义的预测结构的的质量的质量得分918。如将参考图10更详细地所述,评分系统916可以基于以下一项或多项来确定质量得分918: (i) 结构参数似然性得分、(ii) 几何得分、(iii) 距离似然性得分或 (iv) 一个或多个附加得分。

[0295] 评分系统916可以通过确定在每个结构参数的可能值上的相应概率分布,并且根据这些概率分布确定当前结构参数值914的似然性,来确定结构参数似然性得分。评分系统916可以通过生成由当前结构参数值914定义的预测结构与蛋白904的实际结构之间的相似性度量的估计来确定几何得分。评分系统916可以通过确定在氨基酸序列102中的每一对氨基酸之间的可能距离范围上的相应概率分布,并且根据这些概率分布确定由当前结构参数值定义的预测结构的似然性,来确定距离似然性得分。评分系统916可以基于由当前结构参数值914定义的预测结构与现实世界蛋白结构上的生化约束的符合程度,来确定附加得分。如在整个说明书中所使用的,术语“似然性”(例如,如在结构参数似然性得分和距离似然性得分中)可以是指源自于概率分布的任何数值。

[0296] 在确定由当前结构参数值914定义的预测结构的质量得分918之后,优化系统908使用更新引擎920来确定结构参数调整922,优化系统908随后将结构参数调整922用于调整当前结构参数值914。例如,结构参数调整922可以包括与每个结构参数相对应的相应数值。在该示例中,优化系统908可以通过将相应的结构参数调整值922添加到每个当前结构参数值914来调整当前结构参数值914。

[0297] 如将参考图11进一步所述,更新引擎920被配置为通过确定质量得分918相对于每个当前结构参数值914的相应梯度来确定结构参数调整922。更新引擎920使用质量得分918相对于当前结构参数值914的梯度来确定结构参数调整922。质量得分918相对于当前结构参数值914的梯度指示可以调整当前结构参数值914以逐渐提高由所得到的结构参数值定义的预测结构的质量得分918的“方向”。

[0298] 通常,由优化系统908生成的每个预测结构910彼此不同。特别地,定义由优化系统908生成的每个预测结构910的结构参数值是不同的,因为它们源自于不同的(例如,随机确定的)初始结构参数值。选择引擎912可以被配置为将具有最高相应质量得分918的预测结构910选择为最终预测结构906。以这种方式,系统900输出由优化系统908生成的“最佳”预测结构910。

[0299] 通常,优化系统908可以联合优化整个蛋白的预测结构,而不依赖于结构域分割。

[0300] 图10是示例性评分系统916的框图。评分系统916是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下述的系统、部件和技术。评分系统916可以被理解为参考图5描述的评分系统136的替代实施方式。通常,评分系统可以包括

参考图5和图10描述的评分系统的任何部件的任何组合。

[0301] 评分系统916被配置为在优化系统908的每个更新迭代中处理蛋白904的氨基酸序列902的表示和当前结构参数值914,以生成质量评分918。质量得分918是表征由当前结构参数值914定义的蛋白904的预测结构的的质量的数值。评分系统916生成以下一项或多项: (i) 结构参数似然性得分, (ii) 几何得分, (iii) 距离似然性得分, 或 (iv) 一个或多个附加得分, 然后将它们组合 (例如作为加权线性组合) 以生成质量得分918。

[0302] 为了生成结构参数似然性得分1002, 评分系统916可以使用结构预测神经网络1004, 处理包括氨基酸序列902的表示的输入。如前所述, 氨基酸序列902的表示可以是表示氨基酸序列902中的每个氨基酸的单热向量序列。除了氨基酸序列902的表示之外, 结构预测神经网络1004可以被配置为处理附加输入, 该附加输入包括例如由来自其他蛋白的氨基酸序列与氨基酸序列902的多序列比对 (MSA) 得出的数据。MSA是指来自多个其他蛋白的氨基酸序列中的氨基酸与氨基酸序列902中的氨基酸之间的对应关系。可以通过使用任何适当的计算序列比对技术 (例如, 渐进式比对构建) 处理其他蛋白的氨基酸序列 (例如, 存储在数据库中的) 来生成MSA。由MSA得出的结构预测神经网络1004的其他输入可以包括MSA本身的表示、由MSA得出的统计特征 (例如, 二阶统计特征), 诸如参考S. Seemayer, M. Gruber和J. Soding: “CCMpred: fast and precise prediction of protein residue-residue contacts from correlated mutations”, Bioinformatics, 2014所述的那些统计特征, 或两者。

[0303] 结构预测神经网络1004被配置为根据结构预测神经网络权重的值来处理结构预测神经网络输入, 以生成对每个结构参数, 定义在结构参数的可能值上的相应概率分布的输出。在每个结构参数的可能值上的概率分布在本说明书中被称为 (以及在附图中标识为) 结构参数分布1006。例如, 如果结构参数是氨基酸序列902的主链原子之间的扭转角集合, 则对于每个扭转角, 结构预测神经网络1004可以对可能的角度范围集合中的每一个生成相应概率。在特定示例中, 结构预测神经网络1004可以对每个角度范围: $[0, \frac{\pi}{3}), [\frac{\pi}{3}, \frac{2\pi}{3}), [\frac{2\pi}{3}, \pi), [\pi, \frac{4\pi}{3}), [\frac{4\pi}{3}, \frac{5\pi}{3}), [\frac{5\pi}{3}, 2\pi)$ 生成相应概率。在另一特定示例中, 结构预测神经网络1004可以生成定义在氨基酸序列902中的每个氨基酸的结构参数 (例如, 扭转角) 的可能值集合上的相应联合概率分布的数据。在另一特定示例中, 结构预测神经网络1004可以生成在结构参数 (例如, 扭转角) 的可能值上的参数概率分布 (例如, 冯·米塞斯概率分布) 的参数值。

[0304] 为了从结构参数分布1006确定结构参数似然性得分1002, 评估引擎1008使用在由结构预测神经网络1004生成的结构参数的可能值上的相应概率分布, 确定每个当前结构参数值914的概率。此后, 评估引擎1008可以基于每个当前结构参数值的相应概率, 例如基于由式表示的数量, 确定结构参数似然性得分1002:

$$[0305] \quad \prod_i p_i^{sp}(\tau_i) \quad (4)$$

[0306] 其中乘积是在每个结构参数*i*上, 并且 $p_i^{sp}(\tau_i)$ 表示根据在由结构预测神经网络1004生成的结构参数的可能值上的概率分布 $p_i^{sp}(\cdot)$, 当前结构参数值 τ_i 的概率。在特定示例

中,评估引擎1008可以将结构参数似然性得分1002确定为由等式(4)表示的数量,或者为由等式(4)表示的数量的函数(例如,对数)。

[0307] 在一些情况下,例如(如上所述),当结构参数分布1006是在可能结构参数值的范围上的离散概率分布时,参考结构参数分布1006确定的结构参数似然性得分1002可能不是可微分的。为了使结构参数似然性得分1002可微分(即相对于当前结构参数值914),评估引擎1008可以将可微分函数拟合到每个结构参数分布1006。随后,评估引擎1008可以参考拟合到每个结构参数分布1006的可微分函数来确定结构参数似然性得分1002。使可微分函数拟合到给定结构参数分布1006是指确定定义可微分函数的参数的值,该参数的值使得可微分函数尽可能接近地匹配给定结构参数分布1006。评估引擎1008可以使用任何适当的方法来使可微分函数拟合到结构参数分布,例如,矩量法或最大似然法。

[0308] 在一些情况下,评估引擎1008可以将相应参数概率分布(具有可微分概率密度函数)拟合到由结构预测神经网络1004生成的每个结构参数分布1006。例如,参数概率分布可以是具有分布参数 μ 和 κ 的单峰冯·米塞斯概率分布。在一些其他情况下,评估引擎1008可以将相应样条(例如三次样条)拟合到由结构预测神经网络1004生成的每个结构参数分布1006。即使当由结构预测神经网络1004最初生成的结构参数分布是非凸的(例如,多峰)时,由评估引擎1008拟合到结构参数分布的可微分函数也可以是凸函数(例如,单峰冯·米塞斯概率分布)。参考拟合到结构参数分布1006的凸可微分函数来确定结构参数似然性得分1002有助于对质量得分918执行梯度下降,如进一步参考图11所述。

[0309] 为了生成几何得分1010,评分系统916可以使用几何神经网络1012处理包括氨基酸序列902的表示和当前结构参数值914的输入。几何神经网络可以处理附加输入,诸如由来自其他蛋白的氨基酸序列与氨基酸序列902的MSA得出的数据(如前所述)。几何神经网络1012被配置为根据几何神经网络权重的值来处理几何神经网络输入,以生成几何得分1010。几何得分1010是由当前结构参数值914定义的预测结构与蛋白904的实际结构之间的相似性度量的估计。例如,相似性度量可以是当前结构参数值914与定义蛋白904的实际结构的结构参数值之间的均方根偏差(RMSD)。作为另一示例,相似性度量可以是全局距离测试(GDT)相似性度量。GDT相似性可以被定义为氨基酸序列902中这样的氨基酸的主链原子的分数,所述氨基酸在由当前结构参数值914定义的预测结构中的位置落入其在蛋白904的实际结构中的位置的预定距离内。

[0310] 为了生成距离似然性得分1014,评分系统916可以使用距离预测神经网络1016来处理包括氨基酸序列902的表示的输入。距离预测神经网络1016可以处理附加的输入,诸如由来自其他蛋白的氨基酸序列与氨基酸序列902的MSA得出的数据(如前所述)。距离预测神经网络1016被配置为根据距离预测神经网络权重的当前值来处理距离预测神经网络输入,以生成距离图1018。距离图1018对氨基酸序列902中的每一对氨基酸(即,来自氨基酸序列902的两个不同氨基酸的每个集合)定义了在该对氨基酸之间的可能距离范围内的相应概率分布。氨基酸序列902中的第一氨基酸与第二氨基酸之间的距离是指第一氨基酸中的特定原子(例如,主链原子例如 α 碳原子或 β 碳原子)与第二氨基酸中的特定原子之间的物理距离(例如,以埃为单位测量)。在特定示例中,对于氨基酸序列902中的每一对氨基酸,距离图1018可以包括该对氨基酸被距离范围:[0, 2A), [2A, 4A), [4A, 6A), [6A, ∞) 隔开的相应概率,其中A表示埃。在另一个示例中,如前所述,使用64个距离范围。在一些实施方式中,距离

预测神经网络1016可以以被用来生成距离图子集或裁剪,例如如前所述,然后可以将其组合以获得距离图1018。因此,距离预测系统528可以被用在结构预测系统900中以生成距离图1018。

[0311] 评估引擎1008使用距离图1018来确定氨基酸序列902中的每一对氨基酸被由当前结构参数值定义的相应距离 $d_{i,j}$ 隔开的相应概率。当结构参数是数字坐标的序列时,其中每个坐标表示氨基酸序列902中的氨基酸的相应主链原子的位置,则当前结构参数值914直接定义每一对氨基酸之间的距离。例如,一对氨基酸之间的距离可以由表示该对氨基酸中的每个氨基酸中的相应原子例如该对氨基酸的 β 碳原子的位置的相应数字坐标之间的欧几里得距离来定义。当结构参数是氨基酸序列902中的每个氨基酸的相应原子之间的扭转角序列时,则当前结构参数值914间接地定义氨基酸序列902中的每一对氨基酸的位置,因此间接地定义该对氨基酸之间的距离(如前所述)。

[0312] 评估引擎1008使用氨基酸序列902中的每一对氨基酸被由当前结构参数值914定义的相应距离隔开的相应概率来生成距离似然性得分1014。例如,评估引擎1008可以基于由下式表示的数量,生成距离似然性得分1014:

$$[0313] \quad \prod_{(i,j), i \neq j} p_{i,j}^{d_i}(d_{i,j}) \quad (5)$$

[0314] 其中乘积是在由 (i, j) 索引的氨基酸序列902中的氨基酸对上,并且 $p_{i,j}^{d_i}(d_{i,j})$ 表示根据在由距离图1018定义的该对氨基酸残基 (i, j) 之间的可能距离范围内的相应概率分布 $p_{i,j}^{d_i}$,由 (i, j) 索引的氨基酸对被由当前结构参数值914定义的距离 $d_{i,j}$ 隔开的概率。在特定示例中,评估引擎1008可以将距离似然性得分1014确定为由等式(5)表示的数量,或者为由等式(5)表示的数量(例如,负对数概率)的函数(例如,负对数)。

[0315] 在一些情况下,评估引擎1008可以另外使用“参考”距离图来确定距离似然性得分1014。参考距离图表征氨基酸序列902中的每一对氨基酸之间通常期望的估计距离,而不是参考氨基酸序列902中的特定氨基酸的同一性确定的。例如,对于氨基酸序列902中的每一对氨基酸,参考距离图可以基于氨基酸对中的氨基酸的位置和相对偏移来表征氨基酸对之间的估计距离。给定氨基酸的位置是指在给定氨基酸和氨基酸序列中的第一氨基酸之间的其他氨基酸的数量。两个氨基酸之间的相对偏移是指氨基酸序列中的两个氨基酸之间的其他氨基酸的数量。与距离图1018类似,参考距离图通过在可能距离范围集合内的相应概率分布来表征氨基酸序列902中的每一对氨基酸之间的估计距离。

[0316] 评分系统916可以使用不同蛋白的实际结构的蛋白结构数据库来生成参考距离图。在特定示例中,评分系统916可以基于包括在蛋白结构数据库中的相应蛋白结构中的具有相同位置和相对偏移的每一对氨基酸,确定在一对氨基酸之间的可能距离范围集合内的参考概率分布。在另一特定示例中,评分系统916可以基于单独的距离预测神经网络(或距离预测系统528)的输出来确定参考概率分布,该距离预测神经网络被配置为处理表征蛋白的氨基酸序列的长度(但不包括例如比对特征或鉴定氨基酸的特征)的输入。可以使用被用来训练距离预测神经网络616或1016的相同训练数据集(但是考虑到预测其间的距离的原子,具有指示氨基酸是否为缺少 β 碳的甘氨酸的输入特征)来训练这样的单独的距离预测神经网络。

[0317] 评估引擎1008可以根据参考距离图,使用氨基酸序列902中的每一对氨基酸被由当前结构参数值914定义的相应距离隔开的相应概率,以生成距离似然性得分1014。例如,评估引擎1008可以基于由下式表示的数量,生成距离似然性得分1014:

$$[0318] \quad \prod_{(i,j), i \neq j} p_{i,j}^{dl,r}(d_{i,j}) \quad (6)$$

[0319] 其中乘积是在由 (i, j) 索引的氨基酸序列902中的氨基酸对上,并且 $p_{i,j}^{dl,r}(d_{i,j})$ 表示根据在由参考距离图定义的该对氨基酸残基 (i, j) 之间的可能距离范围内的相应概率分布 $p_{i,j}^{dl,r}(\cdot)$, 由 (i, j) 索引的氨基酸对被由当前结构参数值914定义的距离 $d_{i,j}$ 隔开的概率。在特定示例中,评估引擎1008可以将距离似然性得分1014确定为由下式表示的数量:

$$[0320] \quad -\log \prod_{(i,j), i \neq j} p_{i,j}^{dl}(d_{i,j}) + \log \prod_{(i,j), i \neq j} p_{i,j}^{dl,r}(d_{i,j}) \quad (7)$$

[0321] 其中等式 (7) 中的变量具有与等式 (5) 和 (6) 相同的定义。

[0322] 在一些情况下,例如,当距离图1018和参考距离图定义在氨基酸对之间的可能距离范围的范围内的离散概率分布时,参考距离图1018和参考距离图确定的距离似然性得分1014可能不是可微分的。为了使距离似然性得分1014可微分(即,相对于当前结构参数值914),评估引擎1008可以使可微分函数拟合到由距离图1018和参考距离图定义的概率分布。随后,评估引擎1008可以参考拟合到由距离图1018和参考距离图定义的概率分布的可微分函数来确定距离似然性得分1014。如上参考结构参数分布1006所述,评估引擎1008可以将相应参数概率分布(例如,单峰冯·米塞斯概率分布)或样条拟合到由距离图1018和参考距离图定义的概率分布。例如,评估引擎1008可以用样条(诸如三次样条)对离散概率进行插值,例如,负对数概率,这可以称为距离势。在一些实施方式中,距离势可以具有大于阈值距离例如18Å的恒定外推,因为距离越远,越难准确预测。

[0323] 如前所述,评分系统916可以通过生成作为整个距离图1016的适当子集的距离图裁剪集合,然后融合距离图裁剪(例如,通过对重叠的距离图裁剪进行平均)来确定距离图1016,如参考图6所述。

[0324] 评分系统916可以基于一个或多个附加得分,例如物理或物理约束得分,来确定质量得分918。物理得分可以基于由当前结构参数值914定义的预测结构与现实世界蛋白结构上的生化约束的符合程度,来表征当前结构参数值914的似然性。例如,评分系统916可以基于表征与预测结构相关联的原子间势能的范德华势来确定物理得分。这个术语可以帮助抑制空间冲突。在该示例中,评分系统916可以将物理得分确定为:

$$[0325] \quad \sum_i \sum_{j>i} \frac{(r_{ij}^2 - d_{ij}^2)^2}{r_{ij}} \quad (8)$$

[0326] 其中i和j索引氨基酸序列902中的氨基酸, $r_{i,j}$ 表示氨基酸i和氨基酸j的总范德华半径,并且 $d_{i,j}$ 表示氨基酸i和氨基酸j之间的原子间距离。

[0327] 如前所述,例如,如在等式 (7) 中表示的距离似然性得分1014可以被认为是将例如通过梯度下降最小化的基于距离的势能。这可以与基于结构参数似然性得分的势能例如根

据冯·米塞斯概率分布的每个残基的扭转角的负对数似然性的和($-\sum_i \log p(\varphi_i, \psi_i)$), 和/或与基于等式(8)的物理得分的势能相组合,例如求和。因为(7)和(8)是 d_{ij} 而不是 (φ, ψ) 的函数,所以当执行梯度下降以优化 (φ, ψ) 时, d_{ij} 可以通过蛋白几何结构的可微分模型 $x = G(\varphi, \psi)$ 与 (φ, ψ) 相关联,其中 x 表示原子,例如 β 碳坐标,并且 $d_{ij} = ||x_i - x_j||$ 。因此,距离似然性得分1014(“基于距离的势能”)可以被表示为 (φ, ψ) 的函数,以有助于通过梯度下降算法优化这些扭转角。

[0328] 通常,结构预测神经网络1004、几何神经网络1012和距离预测神经网络1016可以以任何适当的神经网络配置来实施。例如,结构预测神经网络1004、几何神经网络1012和距离预测神经网络1016可以包括多个卷积神经网络层、关注层和残差块。在一些情况下,卷积层可以包括扩张卷积滤波器以增加其相应感受野的大小。在一些实施方式中,结构预测神经网络1004可具有自回归架构(例如,源自于WaveNet神经网络架构,同上),该自回归架构顺序地生成以先前结构参数的结构参数值为条件的每个结构参数分布。在一些实施方式中,结构预测神经网络1004可以具有源自于变分自编码器的架构(例如,源自于DRAW神经网络架构,同上),该架构通过在多个内部时间步的每一个处,处理随机采样的潜变量来生成结构参数分布。

[0329] 在一些情况下,一个或多个评分神经网络可能会共享权重值。如果每个神经网络中的权重值都相同,则认为神经网络共享一个权重值,并且在任何一个神经网络中更改权重值(例如,在训练过程中)会导致所有神经网络中的权重值都改变。例如,结构预测神经网络1004和距离预测神经网络1016可以在一个或多个神经网络层中共享相同的权重值。

[0330] 可以使用机器学习训练技术(例如,随机梯度下降)在相应训练数据集合上训练结构预测神经网络1004、几何神经网络1012和距离预测神经网络1016(如下文将更详细描述)。

[0331] 可以基于包括多个训练示例的训练数据集合来训练结构预测神经网络1004。每个训练示例可以包括:(i)源自于具有已知结构的训练蛋白的训练网络输入,以及(ii)定义训练蛋白的已知结构的目标结构参数值。训练网络输入包括训练蛋白的氨基酸序列的表示,以及任选地,由来自其他蛋白的氨基酸序列与氨基酸序列902的MSA得出的数据。目标结构参数值表示应当由结构预测神经网络1004通过处理训练网络输入而生成的结构参数分布1006赋予高概率的参数值。可能已经通过实验方法(例如,X射线晶体学)获得训练蛋白的结构。

[0332] 可以基于包括多个训练示例的训练数据集合来训练几何神经网络1012。每个训练示例可以包括:(i)源自于具有已知结构的训练蛋白的训练网络输入、(ii)蛋白的训练预测结构;以及(ii)作为蛋白的训练预测结构与蛋白的实际结构之间的相似性度量的目标几何得分。训练网络输入包括训练蛋白的氨基酸序列的表示、定义训练蛋白的结构的结构参数值,以及任选地,由来自其他蛋白的氨基酸序列与氨基酸序列902的MSA得出的数据。训练预测结构的目标几何得分表示应当由几何神经网络1012通过处理训练网络输入生成的几何得分。可能已经通过实验方法(例如,X射线晶体学)获得训练蛋白的结构。

[0333] 在一些情况下,可以使用对比发散训练程序来训练几何神经网络1012。在这种情况下,由几何神经网络生成的几何得分可能不是蛋白的预测结构与实际结构之间的特定相似性度量的直接估计。本说明书中几何神经网络1012的描述应当被理解为包括使用对比发

散训练程序来训练几何神经网络的情况。

[0334] 可以基于包括多个训练示例的训练数据集合来训练距离预测神经网络1016。每个训练示例可以包括：(i) 源自于具有已知结构的训练蛋白的训练网络输入，以及(ii) 定义训练蛋白的氨基酸序列中的每一对氨基酸之间的相应距离的目标距离图。训练网络输入包括训练蛋白的氨基酸序列的表示，以及任选地，由来自其他蛋白的氨基酸序列与氨基酸序列902的MSA得出的数据。目标距离图表示氨基酸对之间的距离，该距离应当由距离预测神经网络1016通过处理训练网络输入而生成的距离图1018的概率分布赋予高概率。

[0335] 图11是用于确定蛋白的预测结构的示例性过程1100的流程图。为了方便起见，将过程1100描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如，根据本发明适当编程的优化系统，例如，图9的优化系统908可以执行过程1100。特别地，优化系统908可以多次执行过程1100以确定蛋白的多个预测结构。

[0336] 系统为定义蛋白的结构的每个结构参数确定相应初始值(1102)。通常，系统使用涉及一些随机性的过程来确定结构参数的初始值，如下文更详细所述。

[0337] 在一些实施方式中，为了确定结构参数的初始值，系统获得结构参数值，该结构参数值定义了先前由系统(例如，通过执行过程1100)生成的蛋白的预测结构。然后，系统通过使用随机噪声值，对所获得的定义先前生成的预测结构的结构参数的值进行扰动来确定结构参数的初始值。例如，系统可以通过从预定概率分布(例如，零平均高斯分布)中进行采样来生成随机噪声值，并且可以通过将所生成的随机噪声值添加到定义先前生成的预测结构的结构参数值来确定结构参数的初始值。

[0338] 在一些实施方式中，为了确定结构参数的初始值，系统获得定义系统先前生成的蛋白的多个预测结构的结构参数值。然后，系统通过将定义蛋白的每个先前预测结构的结构参数值组合来确定结构参数的初始值。例如，系统可以通过将定义蛋白的每个先前预测结构的结构参数值进行平均来确定结构参数的初始值。作为另一个示例，系统可以通过提取定义蛋白的每个先前预测结构的结构参数值的不相交部分来确定结构参数的初始值。

[0339] 在一些实施方式中，为了确定结构参数的初始值，系统使用系统的结构预测神经网络来处理输入，该输入包括蛋白的氨基酸序列的表示。如参考图10所述，结构预测神经网络被配置为处理输入以生成对于每个结构参数定义在结构参数的可能值上的相应概率分布的输出。然后，系统通过从在结构参数的可能值上的相应概率分布中采样值来确定每个结构参数的初始值。例如，如果结构参数是氨基酸序列的主链原子之间的扭转角集合，则对于每个扭转角，结构预测神经网络可以生成在可能的角度范围集合内的相应概率。在该示例中，对于每个结构参数，系统可以根据与该结构参数相对应的概率分布对角度范围进行采样，并且将结构参数的初始值确定为从采样的角度范围中随机选择的特定扭转角。

[0340] 系统确定表征由当前结构参数值定义的蛋白的预测结构的的质量的质量得分(1104)。对于过程1100的第一个迭代，当前结构参数值是初始结构参数值(即，如参考1102所述)。如参考图10更详细地所述，系统基于一个或多个评分神经网络的相应输出来确定由当前结构参数值定义的预测结构的的质量得分。每个评分神经网络都被配置为处理：(i) 结构参数的当前值，(ii) 蛋白的氨基酸序列的表示，或(iii) 两者。评分神经网络可以包括以下一项或多项：结构预测神经网络、几何神经网络和距离预测神经网络。

[0341] 通常，一些评分神经网络(例如，结构预测神经网络和距离预测神经网络)可以被

配置为处理不包括当前结构参数值的输入。系统可以一次处理提供给这些评分神经网络的输入(即,在过程1100的第一个迭代之前)。此后,在过程1100的每个迭代中,系统可以使用由这些评分神经网络生成的相应输出来确定质量得分,而无需重新处理其相应输入。对于处理包括当前结构参数值的输入的那些评分神经网络(例如,几何神经网络),系统必须在过程1100的每个迭代中处理它们的相应输入,以便确定该迭代的质量得分。

[0342] 系统确定质量得分相对于每个当前结构参数值的相应梯度(1106)。为了确定质量得分相对于当前结构参数值的梯度,系统可以确定被用来确定质量得分的每个独立得分相对于当前结构参数值的梯度。用于确定质量得分的独立得分可以包括以下一项或多项:结构参数似然性得分、几何得分、距离似然性得分和物理得分。由于质量得分是独立得分的函数(例如,加权线性组合),因此可以从每个独立得分相对于当前结构参数值的梯度确定质量得分相对于当前结构参数值的梯度。例如,如果质量得分由下式给出:

$$[0343] \quad QS = \alpha_1 \cdot SPLS + \alpha_2 \cdot GS + \alpha_3 \cdot DLS \quad (9)$$

[0344] 其中QS是质量得分,SPLS是结构参数似然性得分,GS是几何得分,DLS是距离似然性得分,并且 $\{\alpha_i\}_{i=1}^3$ 是常数,则质量得分相对于当前结构参数值r的梯度由下式给出:

$$[0345] \quad \nabla_r QS = \alpha_1 \cdot \nabla_r SPLS + \alpha_2 \cdot \nabla_r GS + \alpha_3 \cdot \nabla_r DLS \quad (10)$$

[0346] 其中 $\nabla_r QS$ 是质量得分相对于当前结构参数值r的梯度, $\nabla_r SPLS$ 是结构参数似然性得分相对于当前结构参数值r的梯度, $\nabla_r GS$ 是几何得分相对于当前结构参数值r的梯度,并且 $\nabla_r DLS$ 是距离似然性得分相对于当前结构参数值r的梯度。任选地,还可以包括物理得分的术语。任选地,可以省略其他术语,例如 $\nabla_r GS$ 。

[0347] 该系统可以使用任何适当的计算方法来确定结构参数似然性得分、几何得分、距离似然性得分和物理得分相对于当前结构参数值的梯度。(通常,这些得分中的每一个都是结构参数值的可微分函数)。例如,为了确定几何得分相对于当前结构参数值的梯度,系统可以使用反向传播算法的变体,该算法通常用于确定相对于神经网络的当前权重值的梯度。更具体地,系统可以将几何神经网络的权重值视为常量,并且使用反向传播来确定几何得分相对于当前结构参数值的梯度,该当前结构参数值被提供为几何神经网络的输入。作为另一示例,为了确定结构参数似然性得分、距离似然性得分和物理得分的梯度,系统可以使用数值微分方法(有限差分方法)或自动微分方法(例如,如在张量流软件库中实施的)。

[0348] 系统使用质量得分相对于当前结构参数值的梯度来更新当前结构参数值(1108)。通常,系统可以根据来自任何合适的梯度下降优化算法(例如,Adam、RMSprop、Adagrad、Adadelta、AdaMax和L-BFGS等)的更新规则,使用质量得分的梯度,更新当前结构参数值。在一些情况下,系统可以基于包括“动量”的“暖”梯度下降更新规则,使用质量得分的梯度来更新当前结构参数值。当更新规则包括动量时,部分地基于对过程1100的先前迭代中的结构参数值的更新来确定过程1100的当前迭代中的当前结构参数值的更新。参考R.M.Neal, “MCMC using Hamiltonian dynamics”, Ch.5, Handbook of Markov Chain Monte Carlo, Chapman&Hall/CRC Press, 2011, 描述了包括动量的暖梯度下降更新规则的示例。以这种方式,系统可以在优化期间,使得当前结构参数值能够“滚动”质量得分表面,而不是直接找到质量得分表面的局部最小值。质量得分表面是指通过从结构参数值到由结构参数值定义的

预测蛋白结构的质量得分的映射所定义的高维表面。

[0349] 系统确定是否满足终止标准(1110)。例如,如果当前结构参数值已经被更新了至少预定次数(即,已经执行了步骤1104、1106和1108的预定数量的迭代),则系统可以确定满足终止标准。作为另一示例,如果由更新当前结构参数值引起的当前结构参数值的变化(即,如参考1108所述)小于预定阈值,则系统可以确定满足终止标准。

[0350] 响应于确定不满足终止标准,系统可以返回到步骤1104并重复前面的步骤。响应于确定满足终止标准,系统可以在最后一次更新当前参数值之后输出由当前参数值定义的预测结构(1112)。如参考图9所述,系统可以通过重复地执行过程1100以生成蛋白的多个预测结构,并且将该最终预测结构选择为具有最高质量得分的所生成的预测结构,来确定该蛋白的最终预测结构。

[0351] 图12是示例性结构域分割系统1200的框图。结构域分割系统1200是在一个或多个位置的一个或多个计算机上作为计算机程序实施的系统的示例,其中实施下述的系统、部件和技术。

[0352] 结构域分割系统1200被配置为处理定义蛋白1204的氨基酸序列1202的数据以生成蛋白1204的结构域分割1206。氨基酸序列1202中的每个氨基酸是有机化合物,其包括氨基官能团和羧基官能团以及对该氨基酸特异的侧链(即原子团)。通常,蛋白的结构域分割定义了将蛋白的氨基酸序列划分成多个结构域。蛋白的结构域定义了蛋白的氨基酸序列中可以经历蛋白折叠的氨基酸子序列,而与蛋白的氨基酸序列的其余部分(几乎或完全)无关。此外,蛋白结构域可以是独立稳定的,也就是说,可以独立于蛋白的其余部分以稳定形式存在。蛋白折叠是指序列例如氨基酸的无规卷曲(例如,由蛋白1204的氨基酸序列1202定义的)折叠成独特的三维构型的物理过程。

[0353] 在特定的示例中,蛋白的氨基酸序列可以由[A,I,L,M,V,A,A,M,L]给出,其中A表示氨基酸丙氨酸,I表示氨基酸异亮氨酸,L表示氨基酸亮氨酸,M表示氨基酸蛋氨酸,并且V表示氨基酸缬氨酸。该蛋白的示例性结构域分割可以由以下给出:[A,I,L],[M,V,A,A],[M,L]。

[0354] 可以将由系统100为蛋白1204生成的结构域分割1206提供给结构预测系统1208,该结构预测系统1208被配置为生成定义蛋白1204的预测结构1210的输出。结构预测系统1208可以确定由结构域分割1206指定的每个结构域的预测结构,并且随后通过将每个结构域的预测结构组合来确定整个蛋白1204的预测结构。与一次性确定整个蛋白1204的预测结构相比,确定蛋白1204的结构域的预测结构通常是一个“更容易”的问题。具体地,由于氨基酸序列的可能结构的数量随氨基酸序列的长度呈指数增加,所以结构域的可能预测结构的搜索空间通常将成指数地小于整个蛋白1204的可能预测结构的搜索空间。通过分开地确定蛋白1204的每个结构域的预测结构,与直接预测整个蛋白1204的结构相比,结构预测系统1208可以生成更准确的预测,同时消耗更少的计算资源(例如,存储器、计算能力或两者)。

[0355] 参考图1和图9,描述了可以被用来生成蛋白的每个结构域的预测结构的结构预测系统的示例。

[0356] 为了生成结构域分割1206,系统1200生成多个候选结构域分割1230(分别指定蛋白1204的多个候选结构域),并为每个候选结构域分割1230确定相应结构域分割得分1214。如下文更详细地所述,系统1200随后使用结构域分割得分1214来将候选结构域分割1230之

一选择为由系统1200输出的结构域分割1206。

[0357] 系统1200使用结构域分割引擎1216生成候选结构域分割1230。在一些实施方式中,结构域分割引擎1216生成与蛋白1204的每个可能的结构域分割相对应的候选结构域分割1230。在一些其他实施方式中,结构域分割引擎1216生成与蛋白1204的可能结构域分割的适当子集相对应的候选结构域分割1230(例如,通过对蛋白1204的预定数量的可能结构域分割进行随机采样)。

[0358] 系统1200使用评分引擎1218为每个候选结构域分割1230确定相应结构域分割得分1214。为了确定候选结构域分割1230的结构域分割得分1214,评分引擎1218确定由候选结构域分割1230定义的每个候选结构域的相应结构域得分。评分引擎1218随后使用由候选结构域分割1230定义的每个候选结构域的相应结构域得分来确定候选结构域分割1230的结构域分割得分1214。例如,评分引擎1218可以通过将由候选结构域分割1230定义的每个候选结构域的相应结构域得分求和,确定候选结构域分割1230的结构域分割得分1214。

[0359] 在由1220所示的特定示例中,示例性结构域分割1222定义将蛋白1204的氨基酸序列划分为3个结构域:1224-A、1224-B和1224-C。评分引擎1218确定结构域1224-A的结构域得分1226-A、结构域1224-B的结构域得分1226-B和结构域1224-C的结构域得分1226-C。评分引擎1218通过将结构域得分1226-A、1226-B和1226-C求和来确定示例性结构域分割1222的结构域分割得分1228。

[0360] 对于由每个候选结构域分割1230定义的每个候选结构域,评分引擎1218使用以下(i)和(ii)来确定候选结构域的结构域得分:(i)距离图1232,以及(ii)源自于训练结构域1236集合的接触分布数据1234,如下文将更详细所述。

[0361] 距离图1232表征蛋白1204中的每一对氨基酸之间的估计距离。蛋白1204中的第一氨基酸与第二氨基酸之间的距离是指蛋白1204的结构中的第一氨基酸中的特定原子(例如,碳- α 原子或碳- β 原子)与第二氨基酸中的特定例如相应原子之间的物理距离(例如,以埃为单位测量的)。例如,对于蛋白1204中的每一对氨基酸,距离图1232可以包括相应二进制变量,其定义该对氨基酸之间的距离是否被预测为小于预定阈值距离(例如8埃)。系统1200可以通过使用距离图生成引擎1238处理蛋白1204的氨基酸序列1202来生成距离图1232。在下文中,参考图14描述用于生成距离图1232的示例性过程。

[0362] 训练结构域1236定义了相应训练蛋白(其不同于蛋白1204)的实际(即,真相)结构域。可以由人类专家(例如,受过训练以鉴定蛋白结构域的生物学家)手动地确定训练结构域1236。

[0363] 系统1200处理训练结构域1236以生成接触分布数据1234。接触分布数据1234为预定数量的可能长度中的每个给定长度定义了给定长度的训练结构域中的每一氨基酸的接触数上的概率分布。如果隔开两个氨基酸的距离小于预定阈值(例如8埃),则认为蛋白中的两个氨基酸接触。结构域中的给定氨基酸的接触数是指结构域中与给定氨基酸接触的其他氨基酸的数量。结构域的长度是指由结构域定义的氨基酸子序列中的氨基酸的数量。在给定长度的训练结构域中每一氨基酸的接触数上的概率分布定义了对于多个非负整数值中的每一个,给定长度的训练结构域中的给定氨基酸具有由非负整数值定义的训练结构域中的多个接触的相应似然性。

[0364] 例如,接触分布数据1234可以包括对于给定长度的训练结构域,定义在给定长度

的训练结构域中的每一氨基酸的接触数的平均值和标准偏差的数据。平均值和标准偏差定义了给定长度的训练结构域中的每一氨基酸的接触数上的高斯概率分布。在特定示例中,接触分布数据1234可以指定,对于长度为20个氨基酸的训练结构域:(i) 训练结构域中的每一氨基酸的平均接触数为5,(ii) 训练结构域中的每一氨基酸的接触数的标准偏差为1.8。

[0365] 为了确定候选结构域的结构域得分,评分引擎1218(从距离图1232)获得表征由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据。评分引擎1218处理表征由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据,以确定候选结构域中的每一氨基酸的接触数。评分引擎1218从接触分布数据1234获得定义与候选结构域相同长度的训练结构域中的每一氨基酸的接触数上的概率分布的数据。随后,评分引擎1218根据在相同长度的训练结构域中的每一氨基酸的接触数上的概率分布,基于候选结构域中的每一氨基酸的接触数的似然性,确定候选结构域的结构域得分。在下文中,参考图13更详细地描述确定候选结构域的结构域得分。

[0366] 任选地,系统1200可以处理训练结构域1236以生成表征下述(i)和(ii)的附加数据:(i) 训练结构域的长度的分布,以及(ii) 每一训练蛋白中的训练结构域的数量分布。如将参考图13更详细地所述,评分系统1200可以部分地基于该附加数据来确定候选结构域的结构域得分。

[0367] 在对每个候选结构域分割1230生成相应结构域分割得分1214之后,选择引擎1240将候选结构域分割1230之一选择为由系统1200输出的结构域分割1206。例如,选择引擎1240可以将具有最高结构域分割得分1214的候选结构域分割1230选择为由系统1200输出的结构域分割1206。

[0368] 图13是用于确定蛋白的氨基酸序列的结构域分割的示例性过程1300的流程图。为了方便起见,将过程1300描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如,根据本说明书适当编程的结构域分割系统例如图1200的结构域分割系统1200可以执行过程1300。

[0369] 该系统获得蛋白的多个候选结构域分割(1302)。每个候选结构域分割定义了将蛋白的氨基酸序列划分成多个相应候选结构域。在一些实施方式中,系统生成与蛋白的每个可能的结构域分割相对应的候选结构域分割。在一些其他实施方式中,系统生成与蛋白的可能结构域分割的适当子集相对应的候选结构域分割(例如,通过对蛋白的预定数量的可能结构域分割进行随机采样)。

[0370] 该系统获得对于预定数量的可能长度中的每个给定长度,定义在给定长度的训练结构域中的每一氨基酸的接触数上的概率分布的数据(1304)。该系统可以通过处理训练结构域集合来获得数据,其中每个训练结构域定义了训练蛋白的实际(即真相)结构域。在特定的示例中,系统可以确定给定长度的训练结构域中的每一氨基酸的接触数的平均值和标准偏差。更具体地,在该示例中,系统可以确定非负整数值集合的平均值和标准偏差,每个非负整数值定义给定长度的训练结构域中的相应氨基酸的接触数。平均值和标准偏差定义了给定长度的训练结构域中的每一氨基酸的接触数上的高斯概率分布。

[0371] 任选地,系统可以处理训练结构域集合以确定附加数据。例如,系统可以确定表征训练结构域的长度的分布的数据。在该示例中,系统可以确定训练结构域的长度的平均值

和标准偏差。作为另一个示例,系统可以确定表征每一训练蛋白中的训练结构域的数量分布的数据。在该示例中,系统可以确定与每一训练蛋白相对应的训练结构域的数量平均值和标准偏差。

[0372] 该系统获得表征蛋白中的每一对氨基酸之间的估计距离的距离图(1306)。例如,对于蛋白中的每一对氨基酸,距离图可以包括相应二进制变量,其定义该对氨基酸是否被预测为接触。在下文中,将参考图14描述用于生成距离图的示例性过程。

[0373] 系统为每一候选结构域分割的每个候选结构域确定相应结构域得分(1308)。为了确定候选结构域的结构域得分,系统获得(从距离图获得)表征由候选结构域定义的蛋白部分中的每一对氨基酸之间的估计距离的数据。该系统处理表征由候选结构域定义的氨基酸子序列中的每一对氨基酸之间的估计距离的数据,以确定候选结构域中的每一氨基酸的接触数。随后,系统根据在相同长度的训练结构域中的每一氨基酸的接触数上的概率分布,基于候选结构域中的每一氨基酸的接触数的似然性,确定候选结构域的结构域得分。

[0374] 例如,在相同长度的训练结构域中的每一氨基酸的接触数上的概率分布可以由相同长度的训练结构域中的每一氨基酸的接触数的平均值和标准偏差定义的高斯分布。在该示例中,系统可以将候选结构域的结构域得分L确定为:

$$[0375] \quad L = \log \prod_{j=1}^l \mathcal{N}(d_j | \mu_l, \sigma_l) \quad (11)$$

[0376] 其中 d_j 是候选结构域中的第j个氨基酸的接触数,l是候选结构域的长度, μ_l 是长度为l的训练结构域中的每一氨基酸的平均接触数, σ_l 是长度为l的训练结构域中的每一氨基酸的接触数的标准偏差,并且 $\mathcal{N}(d_j | \mu_l, \sigma_l)$ 表示根据由平均参数 μ_l 和标准偏差参数 σ_l 参数化的高斯概率分布的值 d_j 的概率。

[0377] 作为另一示例,系统可以基于候选结构域的长度来另外确定候选结构域的结构域得分。在特定示例中,系统可以将候选结构域的结构域得分L确定为:

$$[0378] \quad L = \log \mathcal{N}(l | \mu^*, \sigma^*) + \log \prod_{j=1}^l \mathcal{N}(d_j | \mu_l, \sigma_l) \quad (12)$$

[0379] 其中l是候选结构域的长度, μ^* 是训练结构域的长度的平均值, σ^* 是训练结构域的长度的标准偏差, $\mathcal{N}(l | \mu^*, \sigma^*)$ 表示根据由平均参数 μ^* 和标准偏差参数 σ^* 参数化的正态概率分布的值l的概率,并且 $\mathcal{N}(d | \mu_l, \sigma_l)$ 参考等式(11)来定义。

[0380] 对于每个候选结构域分割,系统从为由候选结构域分割定义的候选结构域确定的相应结构域得分中确定结构域分割得分(1310)。例如,系统可以将候选结构域分割的结构域分割得分S确定为:

$$[0381] \quad S = \sum_{i=1}^n L_i \quad (13)$$

[0382] 其中i索引由候选结构域分割定义的n个候选结构域,并且 L_i 是为候选结构域i确定的结构域得分。作为另一示例,系统可以基于由候选结构域分割定义的候选结构域的数量来另外确定候选结构域分割的结构域分割得分S。在特定示例中,系统可以将候选结构域

分割的结构域分割得分S确定为:

$$[0383] \quad S = \log \mathcal{N}(n|\hat{\mu}, \hat{\sigma}) + \sum_{i=1}^n L_i \quad (14)$$

[0384] 其中n是由候选结构域分割定义的候选结构域的数量, $\hat{\mu}$ 是具有与蛋白相同长度的每一训练蛋白中的训练结构域的平均数, $\hat{\sigma}$ 是具有与蛋白相同长度的每一训练蛋白中的训练结构域的数量标准偏差, $\mathcal{N}(n|\hat{\mu}, \hat{\sigma})$ 表示根据由平均参数 $\hat{\mu}$ 和标准偏差参数 $\hat{\sigma}$ 参数化的正态概率分布的值n的概率,并且 $\{L_i\}_{i=1}^n$ 参考等式(13)来定义。

[0385] 系统基于为候选结构域分割确定的相应结构域分割得分来确定蛋白的结构域分割(1312)。例如,系统可以将具有最高结构域分割得分的候选结构域分割作选择为蛋白的结构域分割。在确定蛋白的结构域分割之后,系统可以将结构域分割提供给结构预测系统,该结构预测系统被配置为使用结构域分割来生成定义蛋白的预测结构的输出。

[0386] 图14是用于生成距离图的示例性过程1400的流程图,该距离图表征蛋白中的每一对氨基酸之间的估计距离。为了方便起见,将过程1400描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如,根据本说明书适当编程的结构域分割系统例如图12的结构域分割系统1200可以执行过程1400。

[0387] 该系统鉴定蛋白的氨基酸序列的多个氨基酸子序列(1402)。通常,由系统鉴定的氨基酸子序列之一可以是蛋白的完整氨基酸序列。由系统鉴定的多个氨基酸子序列通常“覆盖”蛋白的完整氨基酸序列。也就是说,蛋白中的每个氨基酸通常被包括在由系统鉴定的一个或多个氨基酸子序列中。在一些实施方式中,系统随机地鉴定氨基酸子序列。例如,系统可以通过在蛋白的氨基酸序列中随机地选择氨基酸子序列的起点,并且随机地选择氨基酸子序列的长度来随机地鉴定氨基酸子序列。在这些实施方式中,系统可以继续随机地鉴定氨基酸子序列,直到随机鉴定的氨基酸子序列的集合覆盖蛋白的完整氨基酸序列为止。在一些其他实施方式中,该系统系统地鉴定氨基酸子序列。例如,对于多个不同子序列长度中的每一个(例如64、128和256个氨基酸),系统可以鉴定具有该长度并且被预定偏移量(例如32个氨基酸)隔开的氨基酸子序列。

[0388] 系统对每个鉴定的氨基酸子序列获得相应的多序列比对(MSA)(1404)。蛋白的氨基酸子序列的MSA是指定义氨基酸子序列与来自多个其他蛋白中的每一个的氨基酸(子)序列之间的对应关系的数据。通过使用任何适当的计算序列比对技术(例如,渐进式比对构建)处理来自其他蛋白的氨基酸(子)序列(例如,存储在数据库中的),可以生成用于蛋白的氨基酸子序列的MSA。

[0389] 对于蛋白的每个鉴定的氨基酸子序列,系统处理对该氨基酸子序列获得的MSA,以生成相应距离图裁剪(1406)。与该蛋白的氨基酸子序列相对应的距离图裁剪是指表征蛋白的氨基酸子序列中的每一对氨基酸之间的估计距离的数据。例如,与蛋白的氨基酸子序列相对应的距离图裁剪可以包括用于氨基酸子序列中的每一对氨基酸的相应二进制变量,所述二进制变量定义了该对氨基酸是否被预测为接触。为了生成与氨基酸子序列相对应的距离图裁剪,系统获得了与氨基酸子序列相对应的由MSA得出的特征。由MSA得出的特征可以包括MSA本身的表示、由MSA得出的统计特征(例如,二阶统计特征),诸如参考S.Seemayer,

M.Gruber,和J.Soding:“CCMpred:fast and precise prediction of protein residue-residue contacts from correlated mutations”,Bioinformatics,2014所述,或两者。该系统可以使用神经网络处理由MSA得出的特征,以生成与氨基酸子序列相对应的距离图裁剪。

[0390] 该系统生成距离图,该距离图使用距离图裁剪来表征蛋白中的每一对氨基酸之间的估计距离(1408)。例如,系统可以将距离图生成为距离图裁剪的加权平均值,其中赋予距离图裁剪的权重是基于经过处理以生成距离图裁剪的MSA中的氨基酸(子)序列的数量。在特定示例中,每个距离图裁剪可以包括二进制变量,所述二进制变量指示与该距离图裁剪相对应的氨基酸子序列中的每一对氨基酸是否被预测为接触。对于蛋白中的给定氨基酸对,系统可以确定与每个距离图裁剪中的给定氨基酸对相对应的二进制变量的加权平均值,并将加权平均值四舍五入为0或1以生成二进制变量。从加权平均值中排除了没有表征给定氨基酸对之间的估计距离的距离图裁剪。

[0391] 图15是可以被用于执行上述操作(例如,如参考图1所述的搜索计算单元的操作)的示例性计算机系统1500的框图。系统1500包括处理器1510、存储器1520、存储设备1530和输入/输出设备1540。部件1510、1520、1530和1540中的每一个都可以例如使用系统总线1550互连。处理器1510能够处理用于在系统1500内执行的指令。在一个实施方式中,处理器1510是单线程处理器。在另一实施方式中,处理器1510是多线程处理器。处理器1510能够处理存储在存储器1520中或存储设备1530上的指令。

[0392] 存储器1520将信息存储在系统1500内。在一个实施方式中,存储器1520是计算机可读介质。在一个实施方式中,存储器1520是易失性存储器单元。在另一实施方式中,存储器1520是非易失性存储器单元。

[0393] 存储设备1530能够为系统1500提供大容量存储。在一个实施方式中,存储设备1530是计算机可读介质。在各种不同的实施方式中,存储设备1530可以包括例如硬盘设备、光盘设备、由多个计算设备在网络上共享的存储设备(例如,云存储设备),或一些其他大容量存储设备。

[0394] 输入/输出设备1540为系统1500提供输入/输出操作。在一种实施方式中,输入/输出设备1540可以包括一个或多个网络接口设备,例如以太网卡、串行通信设备(例如,RS-232端口)和/或无线接口设备(例如,802.11卡)。在另一实施方式中,输入/输出设备可以包括驱动器设备,该驱动器设备被配置为接收输入数据并将输出数据发送到其他输入/输出设备,例如键盘、打印机和显示设备1560。但是,也可以使用其他实施方式,诸如移动计算设备、移动通信设备、机顶盒电视客户端设备等。

[0395] 尽管在图15中描述了示例性处理系统,但是,可以以其他类型的数字电子电路或计算机软件、固件或硬件(包括本说明书中公开的结构及其等同结构)或它们中的一种或多种的组合来实施本说明书中所描述的主题和功能操作的实施方式。

[0396] 图16示出了可以通过使用本说明书中所描述的结构预测系统实现的性能增益的示例。特别地,图16示出了与提交给蛋白结构预测关键评估(CASP13)竞赛的其他结构预测系统的性能相比,参考图9描述的结构预测系统的性能的示例。CASP13竞赛是对蛋白结构预测领域的状态进行的盲测评估,以对蛋白结构预测准确性的进展进行基准测试。

[0397] 曲线图1602示出了通过参考图9描述的结构预测系统(由线1604所示)和通过提交

给CASP13的其他结构预测系统(由其余线所示)对给定模板建模(TM)得分阈值预测的自由建模(FM)蛋白结构域的数量。FM蛋白结构域是指先前未确定(例如,通过物理实验)类似蛋白结构域的结构域。TM得分是指介于0和1之间的得分,该得分测量蛋白的拟议结构的主链形状与蛋白的天然(即实际)结构的匹配程度。可以意识到,参考图9描述的结构预测系统几乎在所有TM得分截止值方面都优于其他结构预测系统。

[0398] 图表1606针对六个新确定的蛋白结构(对应于图表1606的水平轴)示出了由参考图9描述的结构预测系统生成的结构预测的TM得分(由黑圈所示,例如1608)和由提交给CASP13的其他结构预测系统生成的结构预测的TM得分(由亮圆圈表示)。可以意识到,参考图9描述的结构预测系统总体上优于其他结构预测系统。

[0399] 图17示出了可以通过使用参考图6描述的距离预测系统实现的性能增益的示例。表1700示出了CASP13中对最可能的L、L/2或L/5个氨基酸残基接触的长程接触预测的精度,其中L是结构域的长度。由参考图6(AF)描述的距离预测系统生成的氨基酸对之间的距离范围内的概率分布作为接触预测的阈值,并且与由CASP13中的排名最高的两种接触预测方法:032(TripletRes)和498(RaptorX-Contact)提交的结果进行比较。表1700示出了对于自由建模(FM)蛋白结构域、基于模板的建模(TBM)蛋白结构域(其中具有相似序列的蛋白结构域具有已知结构)和中间FM/TBM蛋白结构域,参考图6描述的距离预测系统的接触预测准确性。可以意识到,参考图6描述的距离预测系统总体上优于其他接触预测系统。

[0400] 图18是例如使用参考图9描述的结构预测系统来确定蛋白的预测结构的示例性数据流1800的图示。

[0401] 将LxL2D协变特征以及平铺的Lx11D序列以及表达谱特征(其中L是氨基酸序列的长度)级联以生成序列和MSA特征1802。序列和MSA特征可以被表示为数值的3D阵列。

[0402] 使用具有220个残留卷积块的距离预测神经网络1804(例如,如参考图6所述)处理来自特征1802的64x64个裁剪,以生成全距离图1806的裁剪。融合(例如,平均)全距离图的裁剪以生成全距离图1806。全距离图可以指定蛋白中的每一对氨基酸之间的64个可能距离范围内的相应概率分布。距离预测神经网络1804的单独的输出头为蛋白中的每个氨基酸生成结构参数分布(例如,扭转角分布)(即,在该示例中,结构预测网络和距离预测网络共享一些参数值)。

[0403] 在梯度下降的多个迭代中更新定义蛋白的预测结构的结构参数的初始值,以生成蛋白的最终预测结构。在每个迭代中,基于(i)距离图1806、(ii)结构参数分布和(iii)基于表征与预测结构相关联的原子间势能的范德华势能的物理得分来确定由结构参数的当前值定义的预测结构的质量得分。质量得分相对于结构参数的当前值是可微分的,并且相对于结构参数的当前值确定质量得分的梯度。梯度下降优化技术被用于使用质量得分的梯度来调整结构参数的当前值,以确定结构参数的更新值。

[0404] 曲线图1808示出了在每个梯度下降步骤中的蛋白的预测结构与实际结构之间的TM得分1810和RMSD 1812。可以意识到,在梯度下降步骤的顺序中,蛋白的预测结构更准确地近似于蛋白的实际结构。可以从下述(i)和(ii)的覆盖图的3D可视化1814意识到蛋白的最终预测结构准确地近似于蛋白的实际结构:(i)在最后的梯度下降步骤之后的蛋白的预测结构和(ii)蛋白的实际结构。

[0405] 曲线图1816示出了TM得分的改进,其可以通过以不同的初始化多次执行梯度下降

程序以生成不同的预测结构并且将最佳预测结构选择为蛋白的最终预测结构来实现。

[0406] 图19示出了对于蛋白1900,由参考图6描述的距离预测系统生成的距离图的各方面。图示1902是示出蛋白1900的实际(即天然)残基间距离的距离图。图示1904是使用距离预测系统生成并且示出蛋白1900的在残基间距离范围内的概率分布的模式距离图。可以意识到,预测距离图1904是实际距离图1902的准确近似。图示1906显示了在蛋白1900的残基29与所有其他残基之间的可能的距离范围内的预测概率分布。曲线图1908针对距离 $\leq 22\text{\AA}$ 的蛋白1900中的所有残基对,绘制相对于真实距离的预测距离分布的模式,不包括标准偏差 $>3.5\text{\AA}$ 的分布。误差条图示出了对14个bin计算的平均值和标准偏差。曲线图1910绘制了模式距离预测相对于距离分布的标准偏差的误差,不包括原始距离 $>22\text{\AA}$ 的残基对。

[0407] 图20示出了距离预测神经网络(例如,如参考图6所述的)的残差块的示例性架构2000。残差块由一系列神经网络层、交错的三个batchnorm层、两个 1×1 投影层、 3×3 的扩张卷积层和ELU非线性组成。连续层通过1、2、4和8像素的扩张循环,以允许信息快速传播通过序列和MSA特征的裁剪区域。在一个示例性架构中,距离预测神经网络可以包括220个这样的残差块的序列。在最后一个残差块之后,距离预测神经网络可以包括输出层,该输出层具有与由距离预测神经网络生成的距离图裁剪的每个 i, j 分量相对应的相应soft-max函数。

[0408] 图21是被配置为生成蛋白结构片段的DRAW生成神经网络2100的示例性架构的图示,如参考图4所述。生成神经网络2100使用嵌入神经网络2104来处理包括氨基酸序列的序列和MSA特征(例如,蛋白的较长氨基酸序列的子序列)的2-D条件信息2102以生成条件向量2106。嵌入神经网络2104可以包括一个或多个卷积残差块(例如,如参考图20所述),其后是输出条件向量2106的均值池化层。然后将条件向量2106传递到1-D卷积长短期存储器(LSTM)卷积解码器子网络2108中。

[0409] 在128个内部时间步的每一个处,解码器子网络2108根据潜在空间2110上的先验概率分布从潜在空间2110采样潜变量,并且处理该潜变量和条件向量2106以更新解码器子网络的内部状态。先验概率分布可以是例如潜在空间2110上的标准正态分布。在每个内部时间步,生成神经网络2100可以将时间步处的解码器子网络2108的更新的内部状态添加到生成神经网络2100的“画布”内部状态2112。

[0410] 在最终的内部时间步之后,生成神经网络2100的画布内部状态2112的值定义了与蛋白结构片段的每个结构参数相对应的相应概率分布2114(例如,冯·米塞斯分布)。此后,可以根据结构参数值上的概率分布2114来采样任何期望数量的蛋白结构片段的结构参数值。

[0411] 本说明书结合系统和计算机程序部件来使用术语“配置”。对于被配置为执行特定操作或动作的一个或多个计算机的系统,这意味着该系统已经在其上安装了在操作时使得系统执行操作或动作的软件、固件、硬件或它们的组合。对于被配置为执行特定操作或动作的一个或多个计算机程序,这意味着该一个或多个程序包括指令,指令当由数据处理装置执行时,使得该装置执行操作或动作。

[0412] 本说明书中描述的主题和功能操作的实施方式可以在数字电子电路、有形体现的计算机软件或固件、计算机硬件(包括本说明书中公开的结构及其结构等同物)、或者它们中的一个或多个的组合中实施。本说明书中描述的主题的实施方式可以被实施为一个或多

个计算机程序,即,在有形的非瞬态存储介质上编码的计算机程序指令的一个或多个模块,以由数据处理装置执行或控制其操作。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行访问存储设备或它们中的一个或多个的组合。可替代地或附加地,程序指令可以被编码在人工生成的传播信号(例如,机器生成的电、光或电磁信号)上,生成这些信号是为了对信息进行编码以传输到合适的接收器装置以由数据处理装置执行。

[0413] 术语“数据处理装置”是指数据处理硬件,并且涵盖用于处理数据的所有种类的装置、设备和机器,包括例如可编程处理器、计算机或多个处理器或计算机。装置还可以是或进一步包括专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外,装置可以任选地包括为计算机程序创建执行环境的代码,例如构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组的代码。

[0414] 计算机程序(也可以被称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码)可以用任何形式的编程语言编写,包括编译或解释语言、或者声明或过程语言;并且它可以以任何形式部署,包括作为独立程序或作为模块、部件、子例程或适合在计算环境中使用的其他单元。程序可以但不需要对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分(例如标记语言文档中存储的一个或多个脚本)中,存储在专用于所讨论的程序的单个文件中,或者存储在多个协调文件(例如存储一个或多个模块、子程序或部分代码的文件)中。计算机程序可以被部署为在一个计算机上或者在位于一个站点或分布在多个站点并通过数据通信网络互连的多个计算机上执行。

[0415] 在本说明书中,术语“引擎”广泛地用于指代被编程以执行一个或多个特定功能的基于软件的系统、子系统或过程。一般而言,引擎将被实施为安装在一个或多个位置的一个或多个计算机上的一个或多个软件模块或部件。在一些情况下,一个或多个计算机将专用于特定引擎;在其它情况下,可以在同一个或多个计算机上安装并运行多个引擎。

[0416] 本说明书中描述的过程和逻辑流程可以由一个或多个可编程计算机执行,所述可编程计算机执行一个或多个计算机程序,以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程还可以由专用逻辑电路(例如,FPGA或ASIC)执行,或者由专用逻辑电路和一个或多个编程的计算机的组合来执行。

[0417] 适于执行计算机程序的计算机可以基于通用或专用微处理器或两者,或者任何其它种类的中央处理单元。一般而言,中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的基本元件是用于执行或运行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。中央处理单元和存储器可以由专用逻辑电路补充或结合在其中。一般而言,计算机还将包括一个或多个用于存储数据的大容量存储设备(例如,磁盘、磁光盘或光盘),或者被可操作地耦合到一个或多个大容量存储设备以从其接收数据或向其传送数据或两者兼有。但是,计算机不需要具有这样的设备。而且,计算机可以嵌入到另一个设备中,例如,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏机、全球定位系统(GPS)接收器或便携式存储设备(例如通用串行总线(USB)闪存驱动器)等等。

[0418] 适于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备,例如包括:半导体存储器设备,例如,EPROM、EEPROM和闪存设备;磁盘,例如,内部硬盘或可移动盘;磁光盘;以及CD ROM和DVD-ROM盘。

[0419] 为了提供与用户的交互,本说明书中描述的主题的实施方式可以在计算机上实施,该计算机具有用于向用户显示信息的显示设备(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户可以通过其向计算机提供输入的键盘和指点设备(例如,鼠标或轨迹球)。也可以使用其它种类的设备来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感觉反馈,例如,视觉反馈、听觉反馈或触觉反馈;并且可以以任何形式接收来自用户的输入,包括声音、语音或触觉输入。此外,计算机可以通过向用户使用的设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从网络浏览器接收的请求而向用户的设备上的网络浏览器发送网页。而且,计算机可以通过将文本消息或其它形式的消息发送到个人设备(例如,运行消息收发应用的智能电话)并从用户接收响应消息来与用户交互。

[0420] 用于实施机器学习模型的数据处理装置还可以包括例如专用硬件加速器单元,用于处理机器学习训练或生产(即,推理)工作负载的公共部分和计算密集部分。

[0421] 可以使用机器学习框架(例如,TensorFlow框架、微软认知工具包(Microsoft Cognitive Toolkit)框架、Apache Singa框架或Apache MXNet框架)来实施和部署机器学习模型。

[0422] 本说明书中描述的主题的实施方式可以在计算系统中实施,该计算系统包括后端部件(例如,作为数据服务器),或者包括中间件部件(例如,应用服务器),或者包括前端部件(例如,具有图形用户界面或网络浏览器或app的客户端计算机,用户可以通过该图形用户界面或网络浏览器或app与本说明书中描述的主题的实施方式进行交互),或者一个或多个这样的后端、中间件或前端部件的任意组合。系统的部件可以通过任何形式或介质的数字数据通信(例如,通信网络)相互连接。通信网络的示例包括局域网(LAN)和广域网(WAN),例如互联网。

[0423] 计算系统可以包括客户端和服务端。客户端和服务端通常彼此远离,并且通常通过通信网络进行交互。客户端和服务端的关系是通过在相应计算机上运行并且彼此之间具有客户端-服务器关系的计算机程序产生的。在一些实施方式中,服务器向用户设备传输数据,例如,HTML页面,例如为了向与充当客户端的设备交互的用户显示数据并从该用户接收用户输入的目的。可以在服务器处从设备接收在用户设备处生成的数据,例如,用户交互的结果。

[0424] 虽然本说明书包含许多具体的实施细节,但这些不应当被解释为对任何发明的范围或所要求保护的范围的限制,而是对可能特定于特定发明的特定实施方式的特征的描述。本说明书中在单独实施方式的上下文中描述的某些特征也可以在单个实施方式中组合实施。相反,在单个实施方式的上下文中描述的各种特征也可以在多个实施方式中单独地实施或以任何合适的子组合来实施。而且,尽管特征可能在上文描述为以某些组合起作用,甚至最初被这样要求保护,但是在某些情况下,可以从所要求保护的组合中删除该组合中的一个或多个特征,并且所要求保护的组合可以涉及子组合或子组合的变体。

[0425] 类似地,尽管操作以特定次序在附图中描绘并在权利要求中记载,但这不应当被理解为要求以所示的特定次序或以顺序的次序执行这些操作,或者要求执行所有示出的操作以获得期望的结果。在某些情况下,多任务处理和并行处理可能是有利的。而且,上述实施方式中各种系统模块和部件的分离不应当被理解为在所有实施方式中都要求这种分离,

并且应当理解的是,所描述的程序部件和系统一般可以一起集成在单个软件产品中或者封装到多个软件产品中。

[0426] 已经描述了本主题的特定实施方式。其它实施方式在以下权利要求的范围内。例如,权利要求中记载的动作可以以不同的次序执行,并且仍然实现期望的结果。作为一个示例,附图中描绘的过程不一定要要求所示的特定次序或顺序的次序来实现期望的结果。在一些情况下,多任务处理和并行处理可能是有利的。

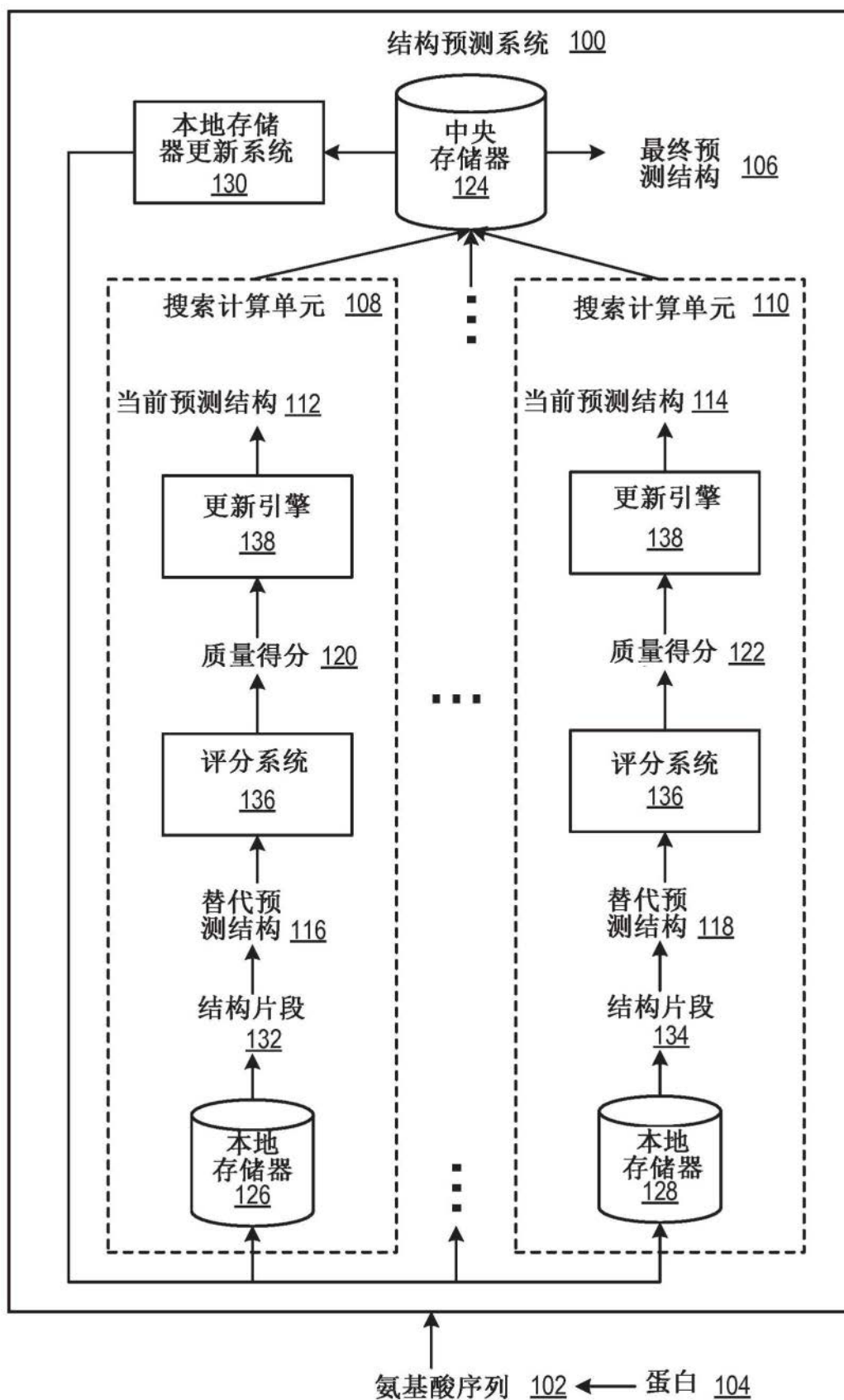


图1

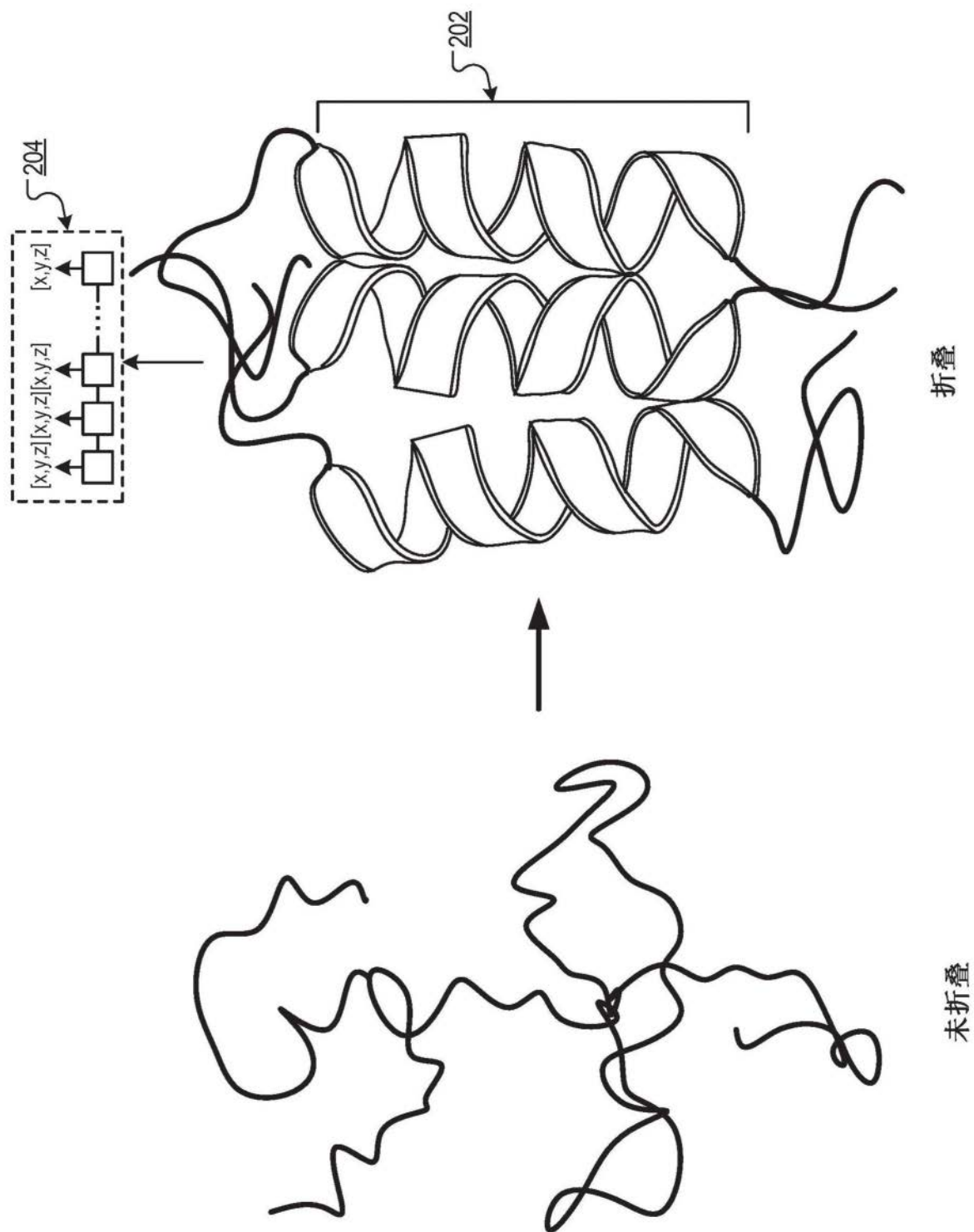


图2

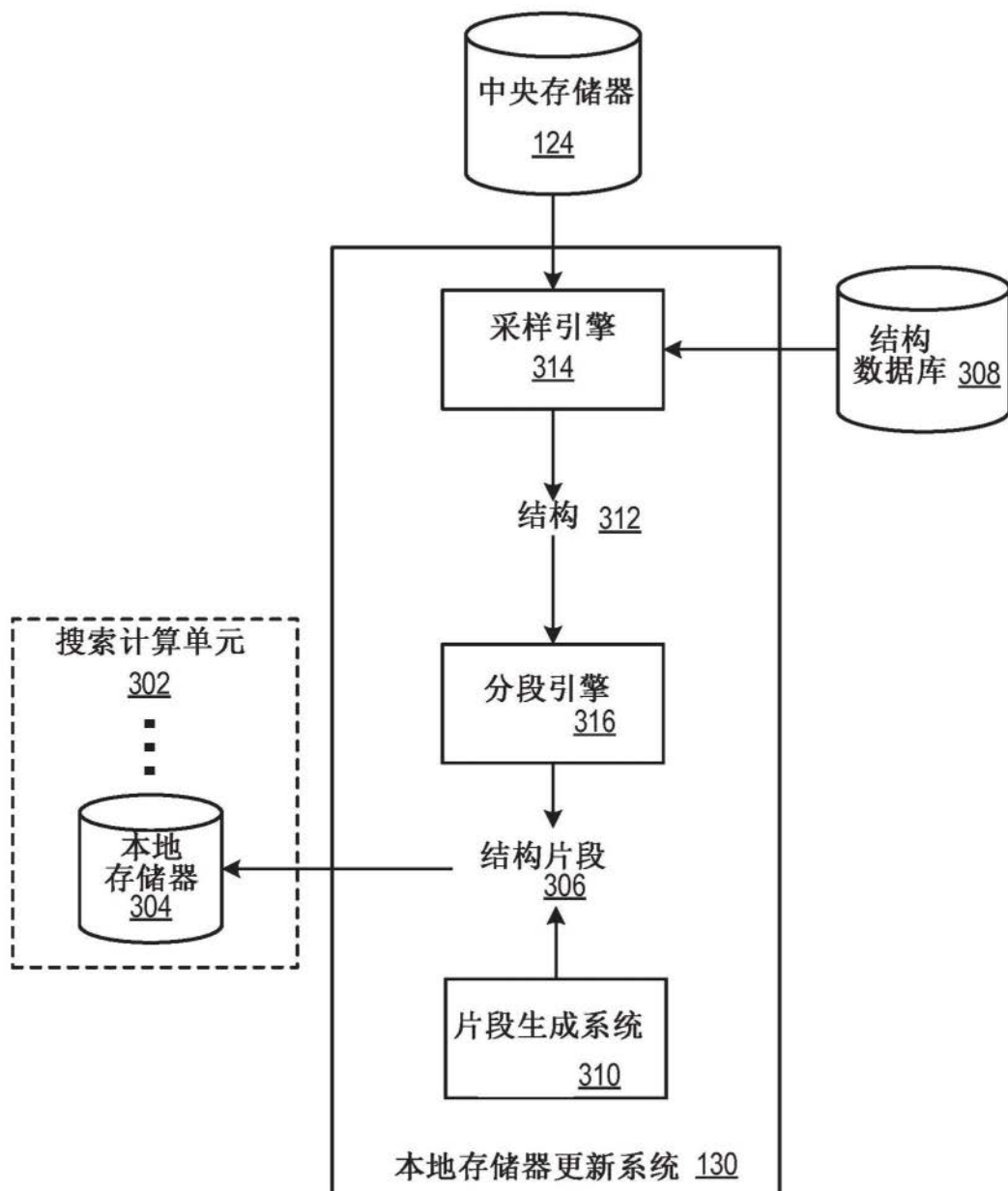


图3

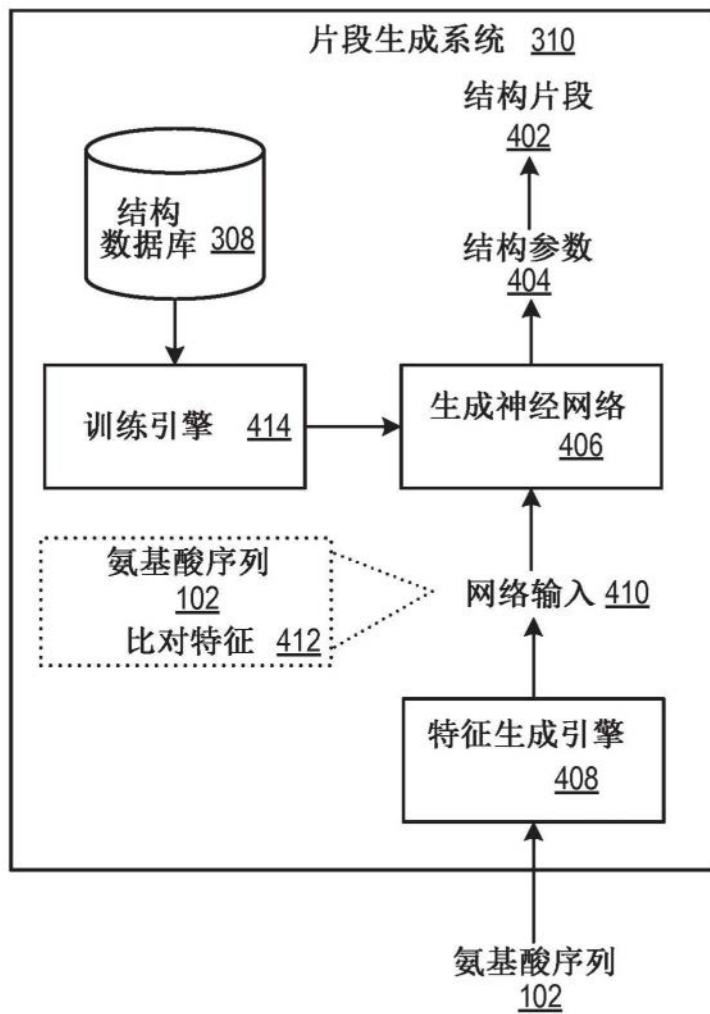


图4

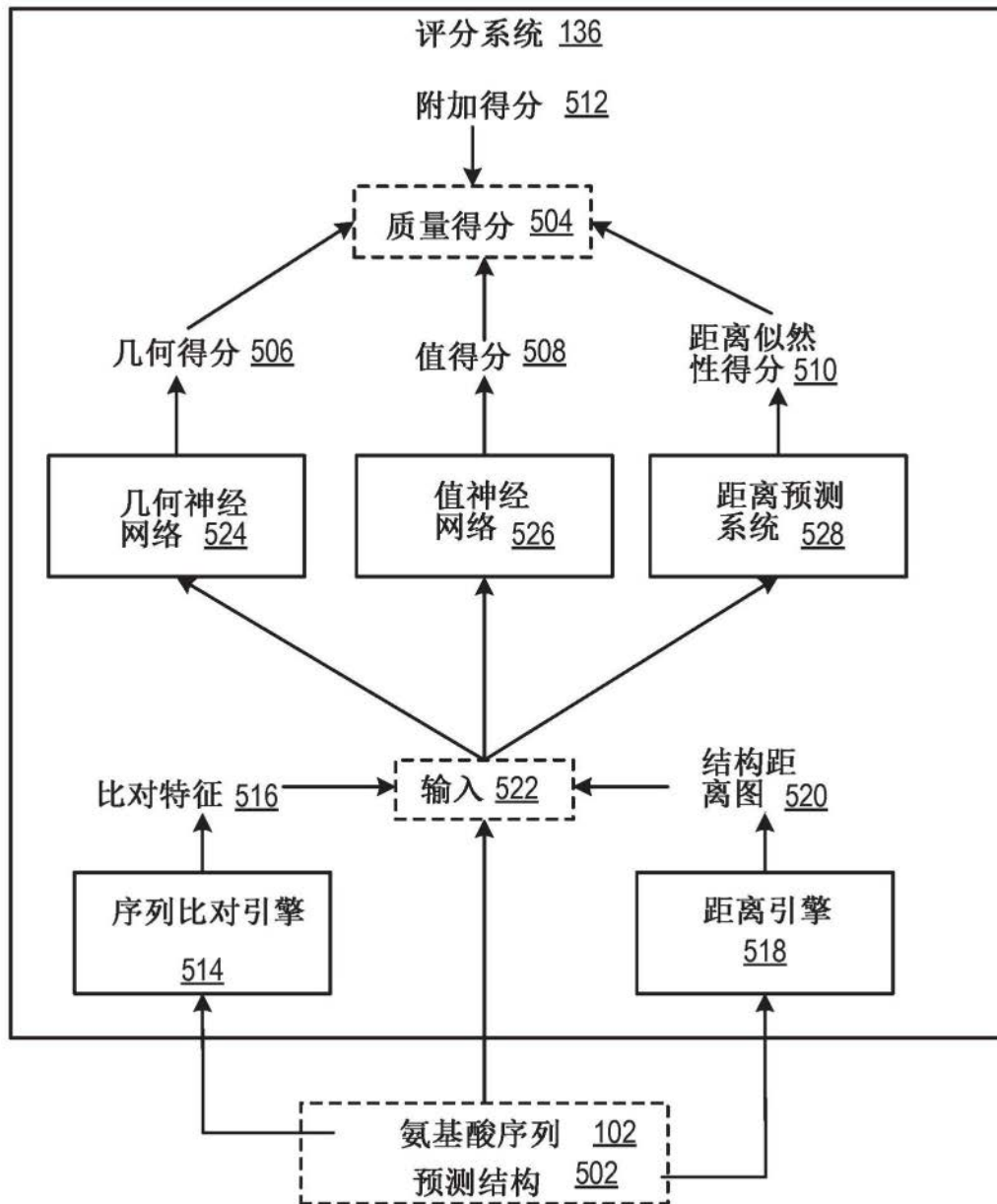


图5

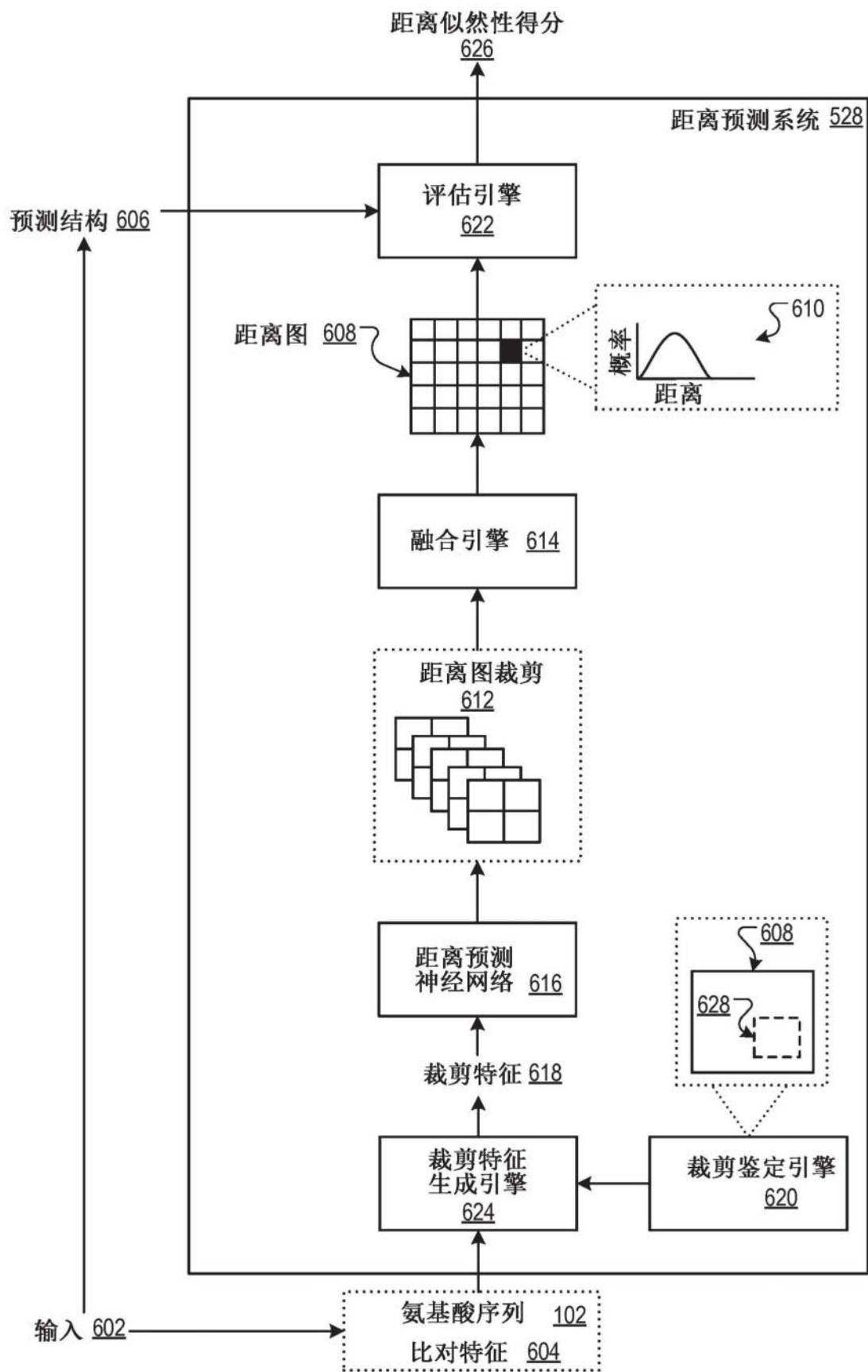


图6

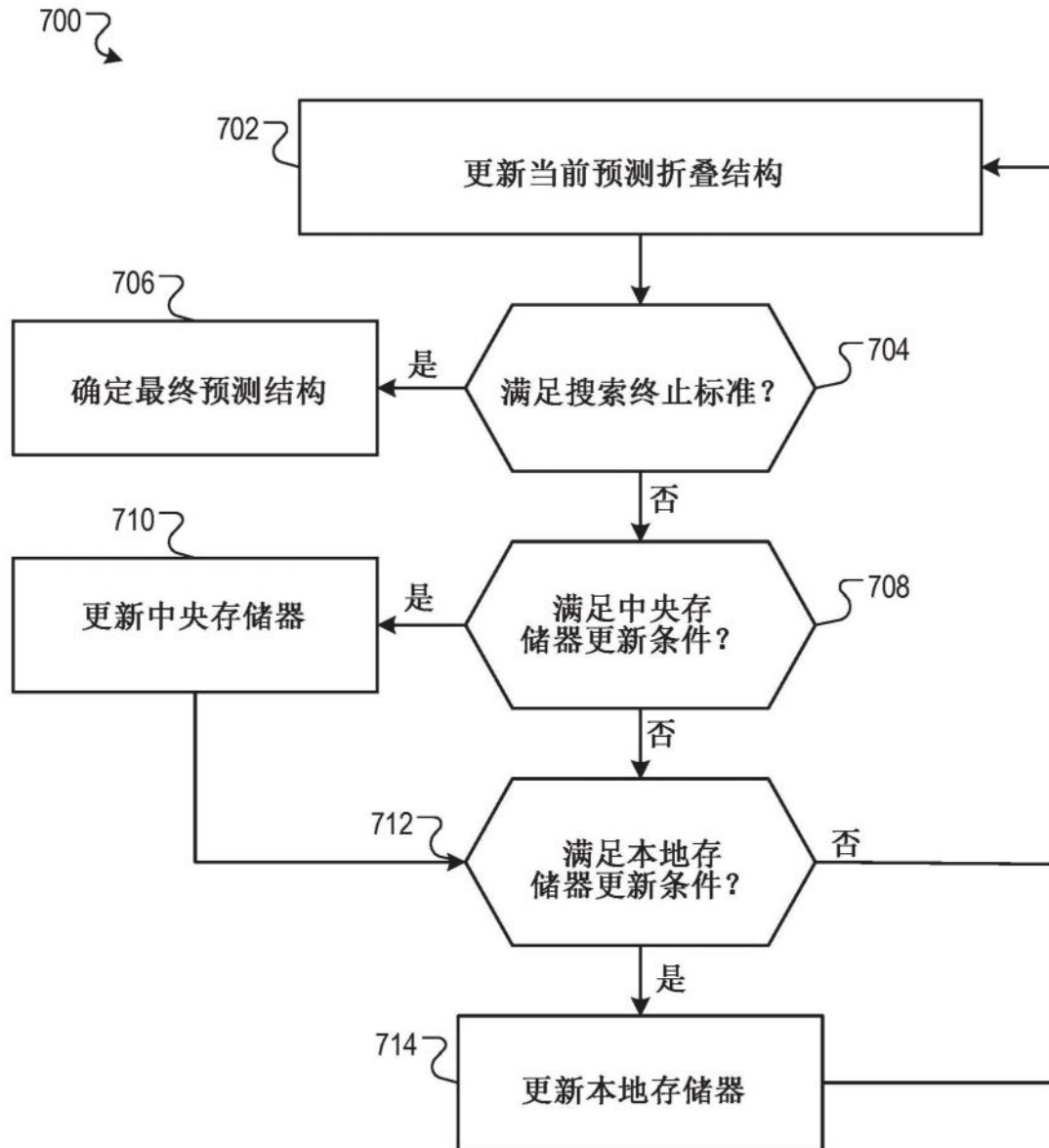


图7

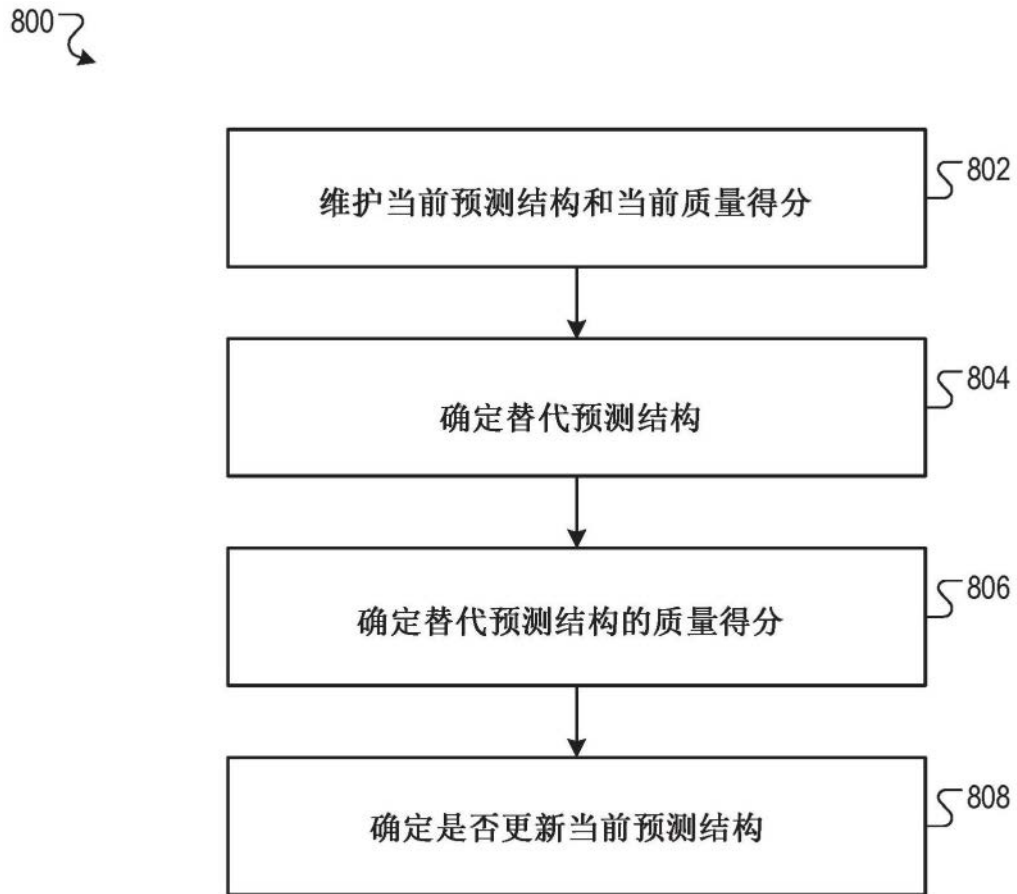


图8

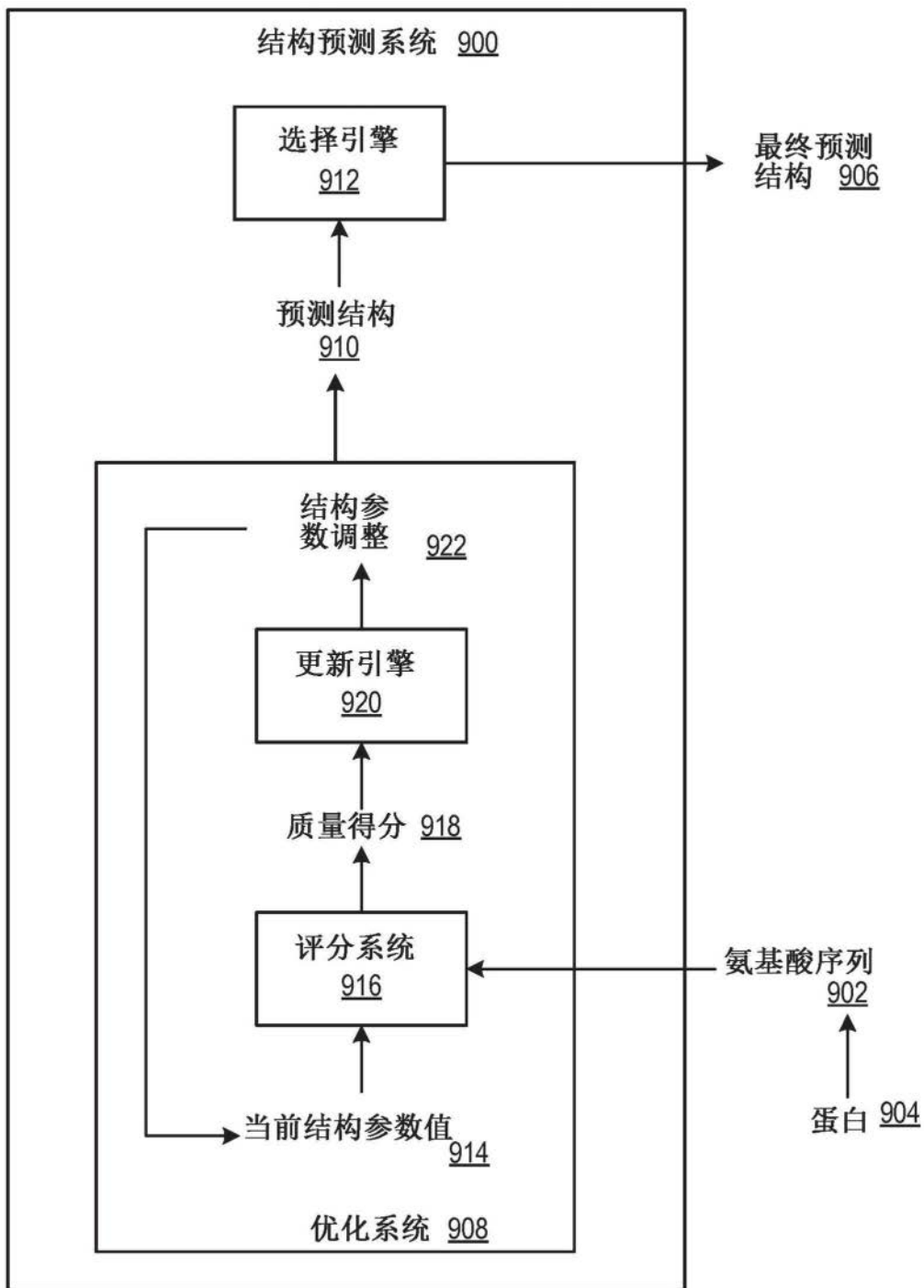


图9

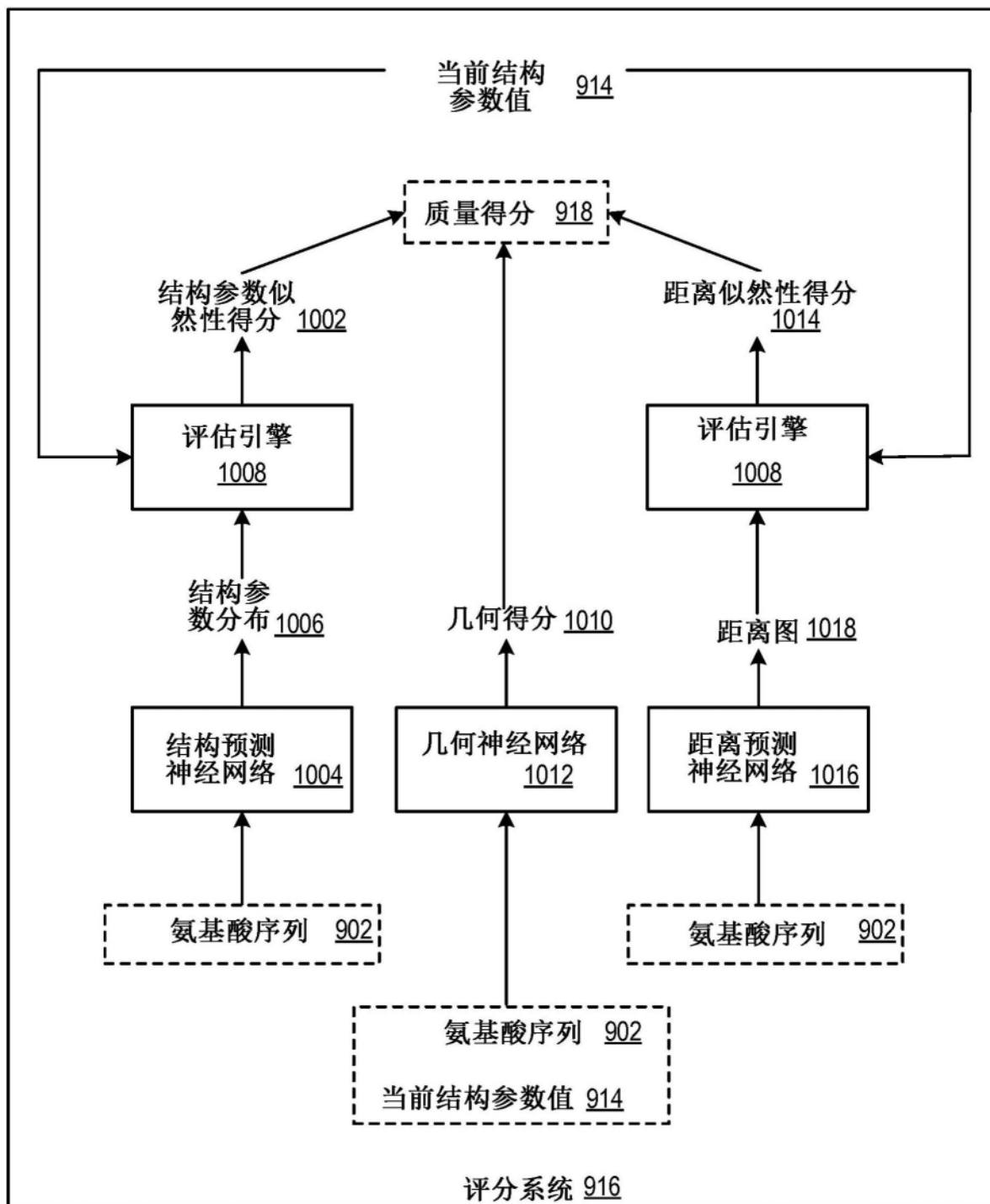


图10

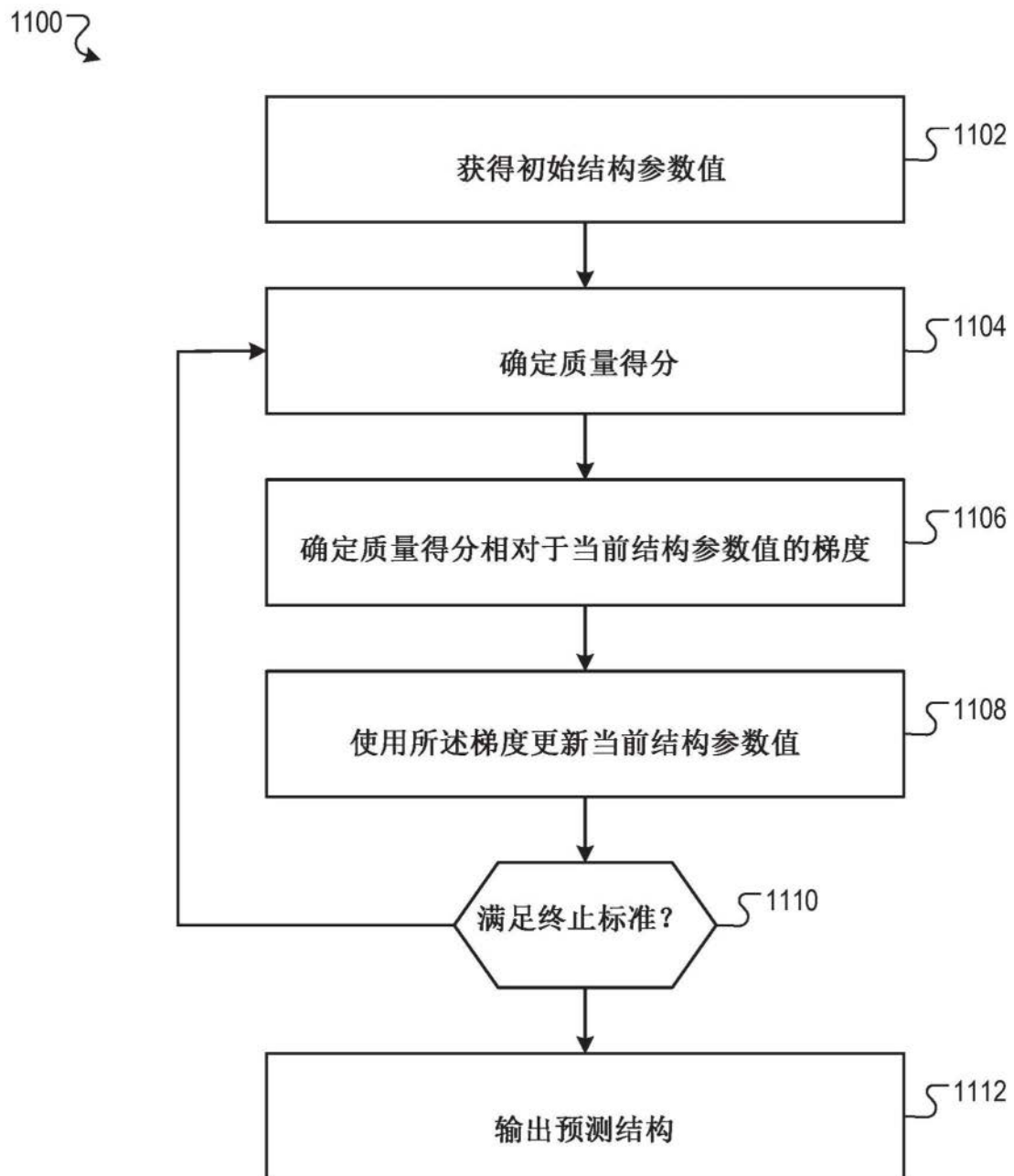


图11

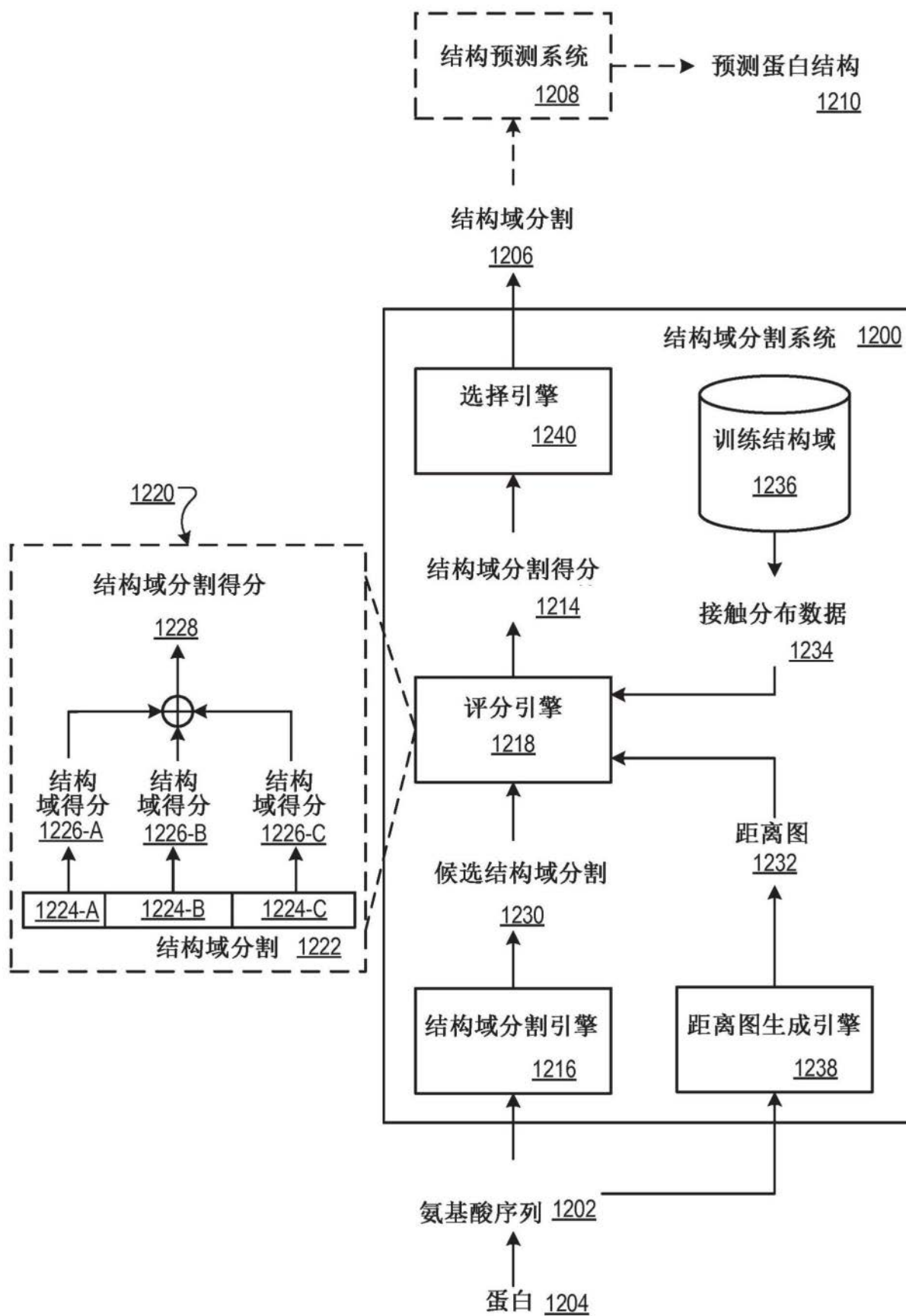


图12

1300

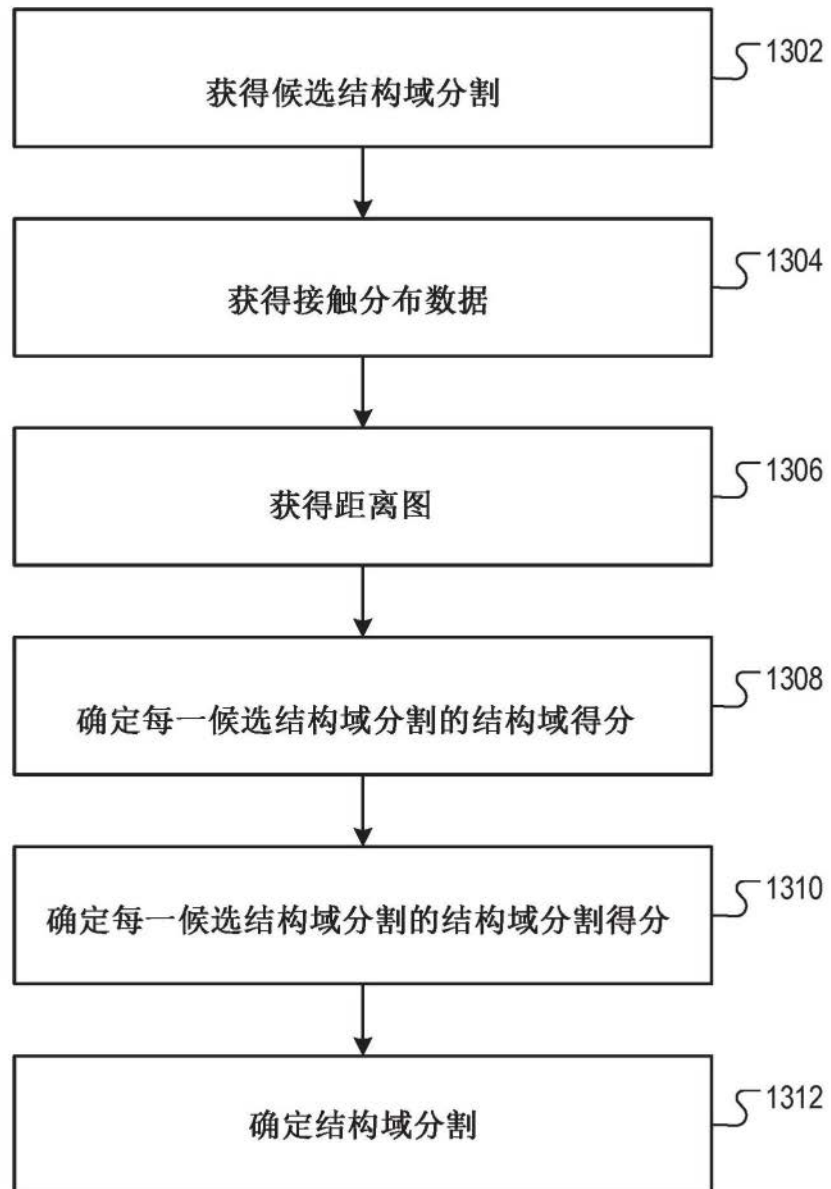


图13

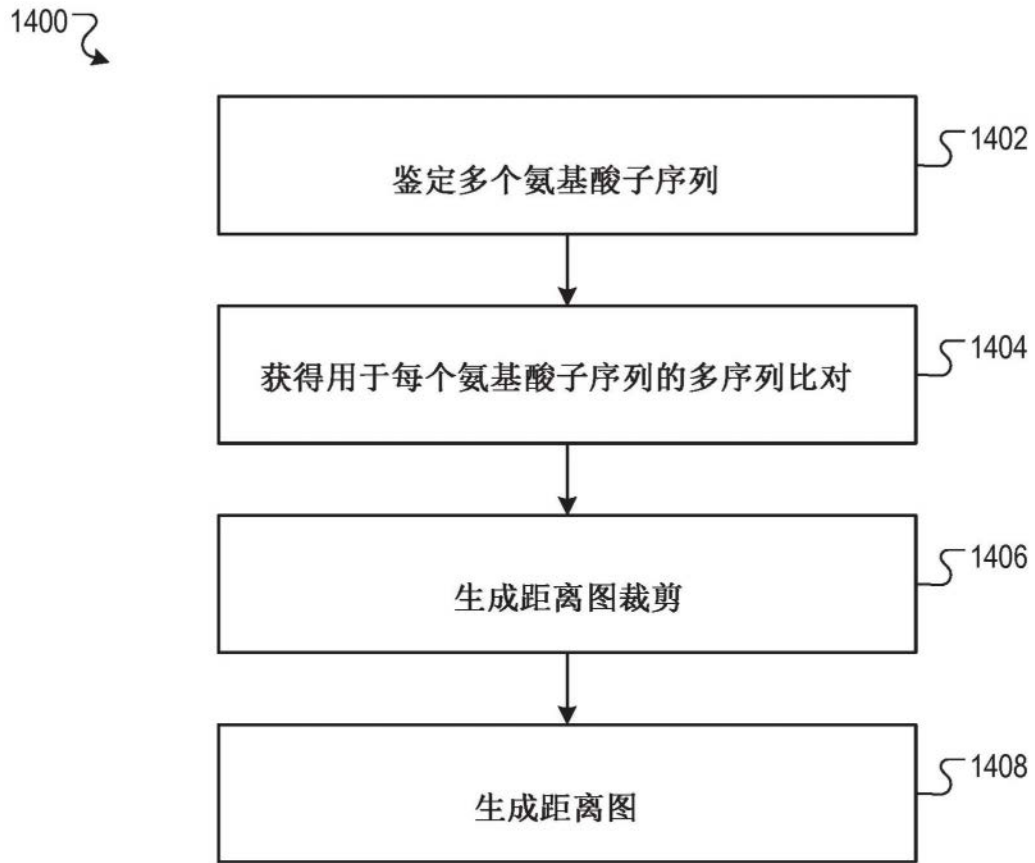


图14

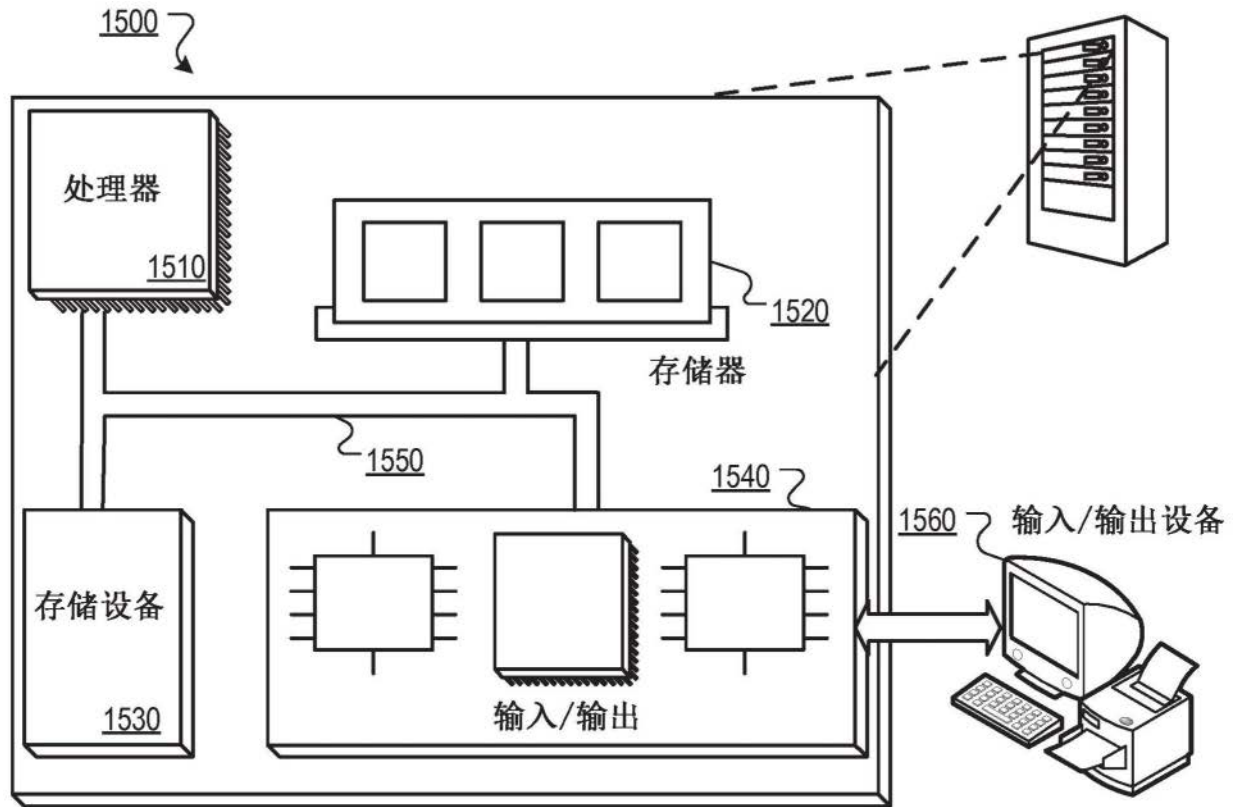


图15

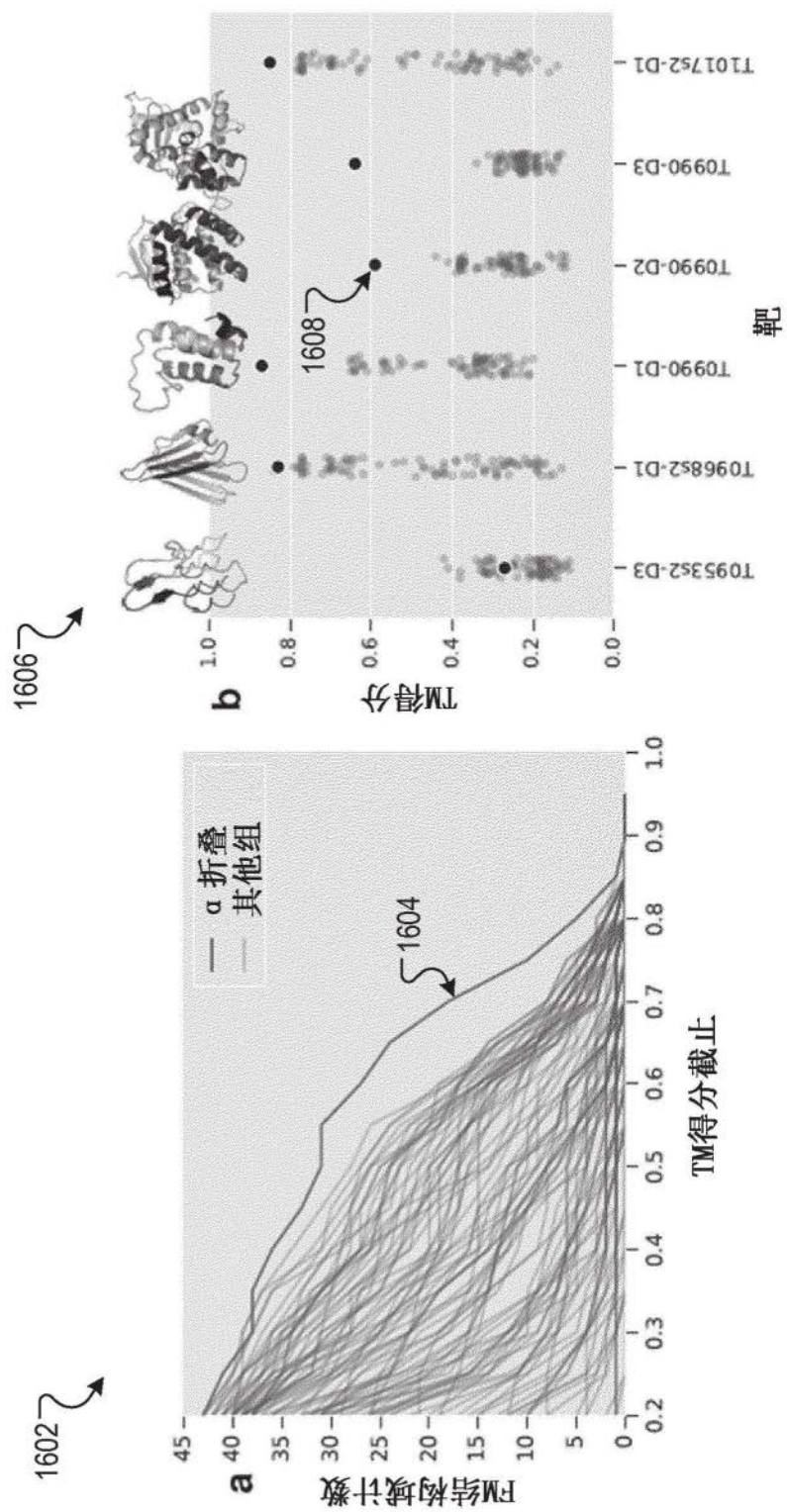


图16

1700 ↷

c	接触精度 集合	N	L长			L/2长			L/5长		
			AF	032	498	AF	032	498	AF	032	498
	FM	31	45.5	42.9	39.8	58.0	55.1	51.7	70.1	67.3	61.6
	FM/TBM	12	59.1	53.0	48.9	74.2	64.5	64.2	85.3	81.0	79.6
	TBM	61	68.3	65.5	61.9	82.4	80.3	76.4	90.6	90.5	87.1

图17

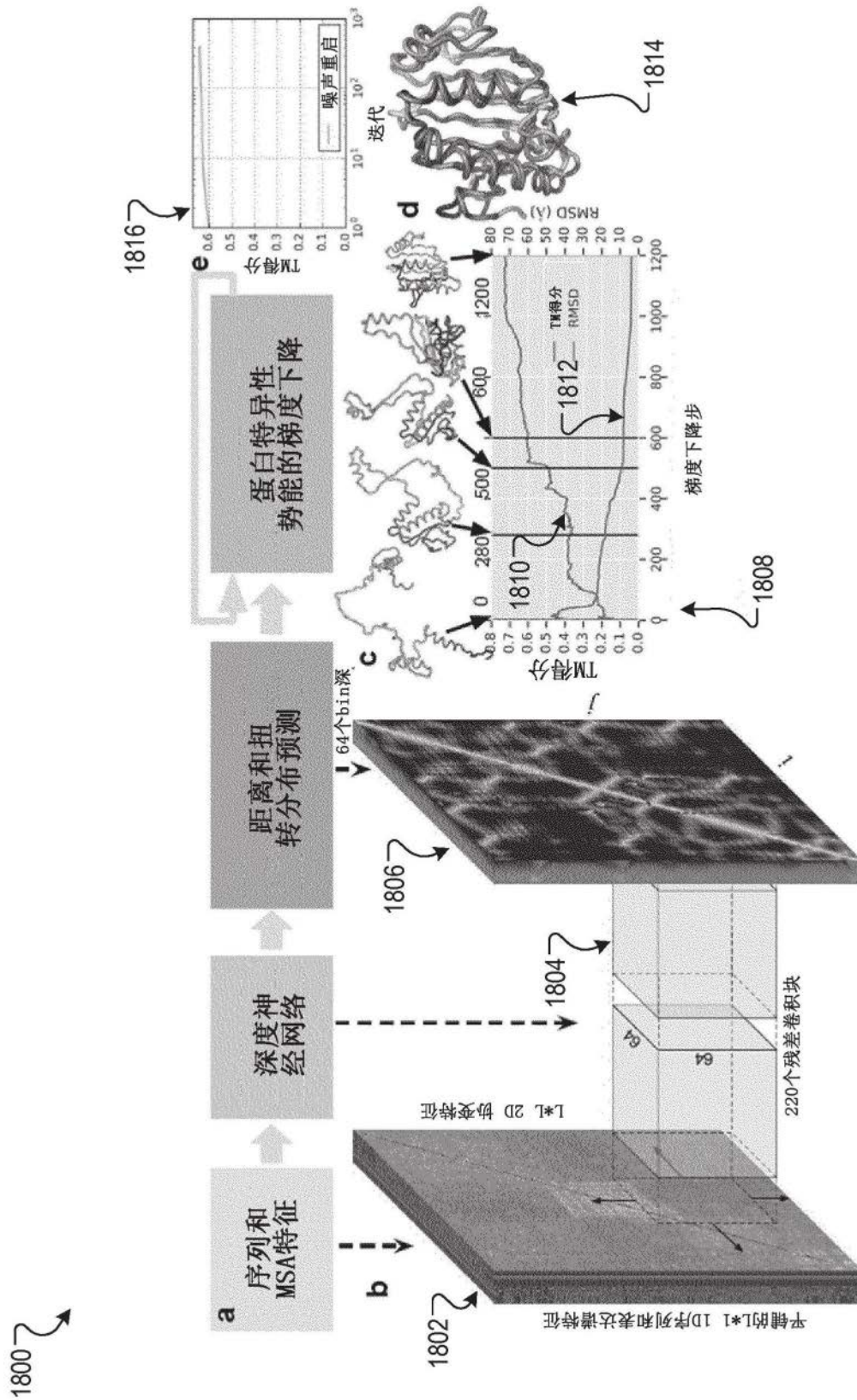


图18

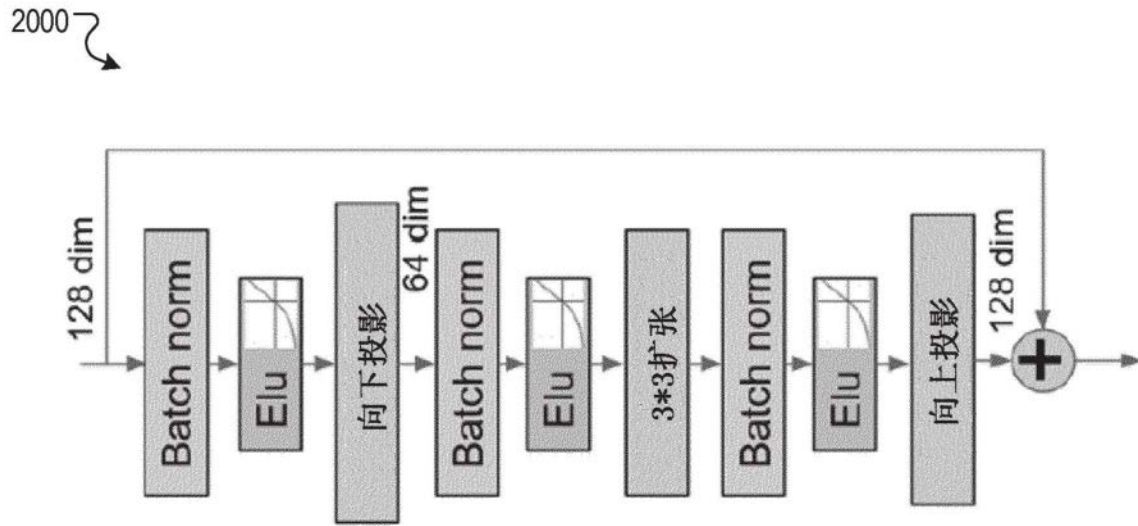


图20

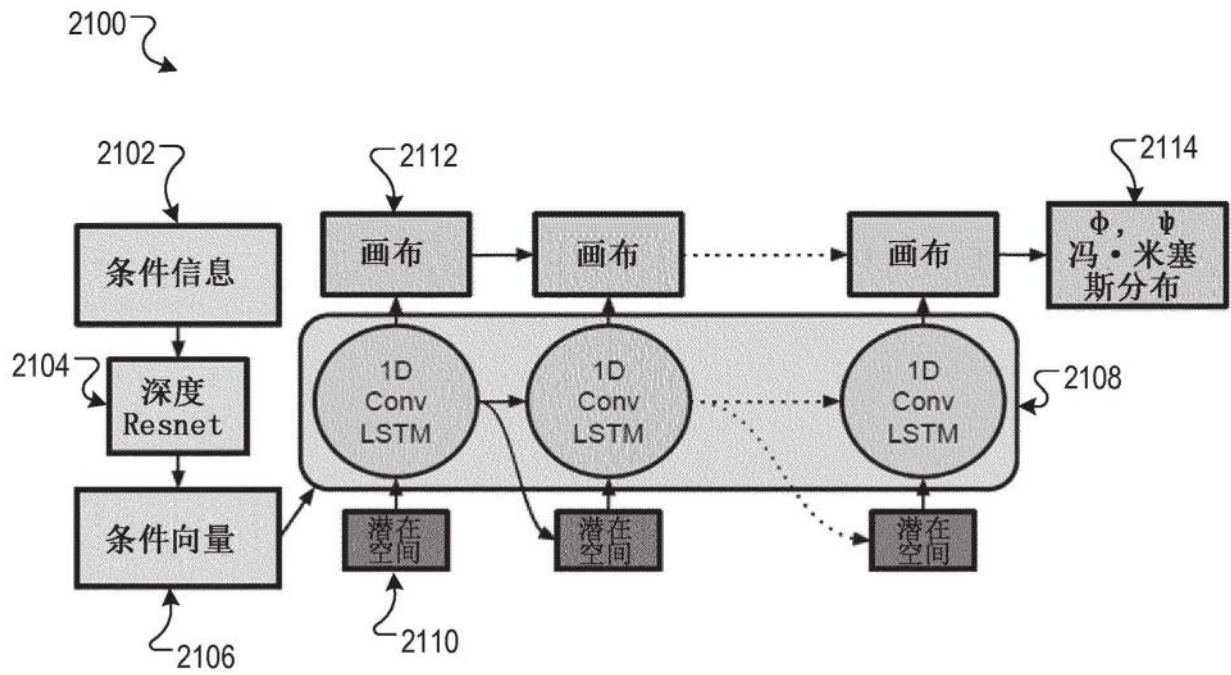


图21