



(12) 发明专利申请

(10) 申请公布号 CN 116420191 A

(43) 申请公布日 2023. 07. 11

(21) 申请号 202180069819.1

(22) 申请日 2021.11.23

(30) 优先权数据

63/118,919 2020.11.28 US

(85) PCT国际申请进入国家阶段日

2023.04.12

(86) PCT国际申请的申请数据

PCT/EP2021/082707 2021.11.23

(87) PCT国际申请的公布数据

W02022/112260 EN 2022.06.02

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 J·江珀 M·菲古尔诺夫

(74) 专利代理机构 北京市柳沈律师事务所

11105

专利代理师 金玉洁

(51) Int.Cl.

G16B 30/00 (2019.01)

G16B 40/00 (2019.01)

G06N 3/045 (2023.01)

G06N 3/082 (2023.01)

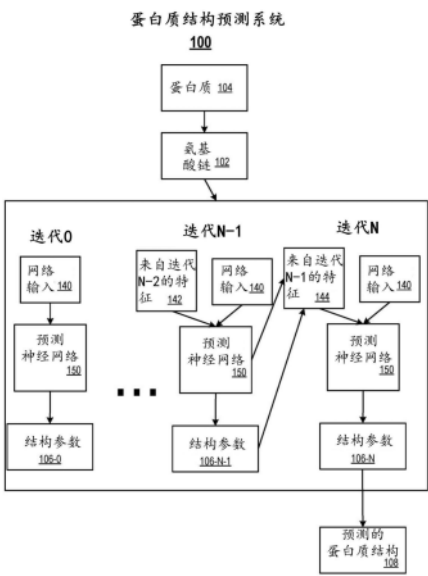
权利要求书4页 说明书30页 附图13页

(54) 发明名称

利用循环通过多次迭代预测蛋白质结构

(57) 摘要

本公开提供了用于预测包括一条或多条链的蛋白质结构的方法、系统和装置,包括编码在计算机存储介质上的计算机程序。在一个方面,一种方法包括,在迭代序列中的第一次迭代之后的每次后续迭代中:获得表征蛋白质的后续迭代的网络输入;从(i)在序列中后续迭代之前的先前迭代中生成的结构参数,(ii)在最后迭代中生成结构参数时由蛋白质结构预测神经网络生成的一个或中间输出,或(iii)两者,生成后续迭代的特征;以及使用蛋白质结构预测神经网络处理用于后续迭代的特征和网络输入,以生成定义蛋白质的另一预测的结构的后续迭代的结构参数。



1. 一种由一个或多个计算机执行的方法,用于预测包括一条或多条链的蛋白质的结构,其中每条链包含氨基酸序列,所述方法包括:

在包括第一次迭代以及随后的一次或多次后续迭代的迭代序列的第一次迭代中:

获得表征蛋白质的第一次迭代的网络输入;

使用蛋白质结构预测神经网络处理第一次迭代的网络输入,以生成定义蛋白质的初始预测的结构的用于第一次迭代的结构参数;

在迭代序列中的每次后续迭代中:

获得表征蛋白质的后续迭代的网络输入;

从(i)在序列中后续迭代之前的先前迭代中生成的结构参数,(ii)在最后迭代中生成结构参数时由蛋白质结构预测神经网络生成的一个或中间输出,或(iii)两者,生成后续迭代的特征;和

使用蛋白质结构预测神经网络处理用于后续迭代的特征和网络输入,以生成定义蛋白质的另一预测的结构的用于后续迭代的结构参数;和

根据序列中最后迭代的结构参数确定蛋白质的最终预测的结构。

2. 根据权利要求1所述的方法,其中,迭代序列中的每次迭代的网络输入是相同的网络输入。

3. 根据任一前述权利要求所述的方法,其中,使用蛋白质结构预测神经网络处理用于后续迭代的特征和网络输入以生成定义蛋白质的预测的结构的用于后续迭代的结构参数包括:

根据用于后续迭代的特征和网络输入来生成组合的输入;和

使用蛋白质结构预测神经网络处理组合的输入,以生成用于后续迭代的结构参数。

4. 根据任一前述权利要求所述的方法,其中:

每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入,

在每次迭代中,结构预测神经网络被配置为在生成迭代的结构参数时重复更新初始对嵌入,

生成特征包括从在先前迭代中生成结构参数时生成的更新的对嵌入,生成对嵌入的变换的集合;和

生成组合的输入包括为后续迭代组合对嵌入的变换的集合和初始对嵌入。

5. 根据任一前述权利要求所述的方法,其中:

每次迭代的相应网络输入包括代表与蛋白质中每条链相对应的相应多序列比对(MSA)的初始MSA表示,

在每次迭代中,结构预测神经网络被配置为生成单个嵌入的一个和多个集合,每个单个嵌入集合包括在生成迭代的结构参数时蛋白质的每个氨基酸的相应单个嵌入,

生成特征包括从在先前迭代中生成结构参数时生成的单个嵌入集合之一,生成单个嵌入的变换的集合;和

生成组合的输入包括为后续迭代组合单个嵌入的变换的集合和初始MSA表示。

6. 根据权利要求5所述的方法,其中,为后续迭代组合单个嵌入的变换的集合和初始MSA表示包括将单个嵌入的变换的集合相加到初始MSA表示的第一行。

7. 根据任一前述权利要求所述的方法,其中:

每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入，
在每次迭代中，结构参数为每个氨基酸指定蛋白质结构中氨基酸中指定的原子的预测的3-D空间位置；

生成特征包括：

根据先前迭代中由结构参数指定的氨基酸的预测的3-D空间位置生成距离图，所述距离图为蛋白质中的每对氨基酸表征蛋白质结构中氨基酸对之间的相应估计的距离；和

根据距离图生成具有与初始对嵌入相同的维度的变换的距离图；和生成组合的输入包括为后续迭代组合变换的距离图和初始对嵌入。

8. 根据权利要求1-6中任一项所述的方法，其中：

每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入，

在每次迭代中，结构参数指定距离图，所述距离图为蛋白质中的每对氨基酸表征蛋白质结构中氨基酸对之间相应估计的距离；

生成特征包括：

根据先前迭代中由结构参数指定的距离图，生成具有与初始对嵌入相同的维度变换的距离图；和

生成组合的输入包括为后续迭代组合变换的距离图和初始对嵌入。

9. 根据任一前述权利要求所述的方法，其中：

每次迭代的相应网络输入包括使用一个或多个模板序列的集合和每个模板序列的对应已知结构生成的蛋白质中每对氨基酸的相应初始对嵌入；和

生成组合的输入包括通过将蛋白质和由先前迭代中的嵌入定义的结构预测相加到一个或多个模板序列的集合和相应的已知结构中，来修改用于后续迭代的初始对嵌入。

10. 一种训练任一前述权利要求的结构预测神经网络的方法，所述方法包括：

选择要执行的迭代次数；

为每次迭代获得相应的训练网络输入，每个训练网络输入表征训练蛋白质；

在包括所选的迭代次数的迭代序列中的第一次迭代中：

使用蛋白质结构预测神经网络处理第一次迭代的网络输入，以生成定义训练蛋白质的初始预测的结构的用于第一次迭代的结构参数；

当所选的迭代次数大于1时，在迭代序列中的每次后续迭代中，直到已经执行了所选的迭代次数：

从(i)在序列中后续迭代之前的先前迭代中生成的结构参数，(ii)在最后迭代中生成结构参数时由蛋白质结构预测神经网络生成的一个或中间输出，或(iii)两者，生成后续迭代的特征；和

使用蛋白质结构预测神经网络处理用于后续迭代的特征和训练网络输入，以生成定义训练蛋白质的另一预测的结构的用于后续迭代的结构参数；和

确定包括结构损失项的目标函数的梯度，所述结构损失项表征(i)由序列中最后迭代生成的结构参数指定的靶蛋白的预测的结构和(ii)训练蛋白质的靶蛋白结构之间的相似性；和

使用梯度更新结构预测神经网络的参数值，从序列中最后迭代的结构参数确定蛋白质的最终预测的结构。

11. 根据权利要求10所述的方法, 其中, 确定梯度包括在由结构预测神经网络在最后迭代中作为输入处理的特征处应用停止梯度。

12. 根据权利要求10或11所述的方法, 其中, 选择迭代次数包括:

从概率分布中采样迭代次数, 所述概率分布为每个迭代次数分配相应概率, 迭代次数在从1到结构预测神经网络被训练后使用的总迭代次数的范围内。

13. 根据权利要求10-12中任一项所述的方法, 还包括: 当所选的迭代次数等于1时:

确定包括结构损失项的目标函数的梯度, 所述结构损失项表征 (i) 由序列中第一次迭代生成的结构参数定义的靶蛋白的初始预测的结构和 (ii) 训练蛋白质的靶蛋白结构之间的相似性; 和

使用梯度来更新结构预测神经网络的参数值。

14. 一种系统, 包括:

一个或多个计算机; 和

一个或多个存储设备, 其通信地耦合到所述一个或多个计算机, 其中所述一个或多个存储设备存储指令, 当指令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权利要求1-13中任一项所述的相应方法的操作。

15. 一个或多个存储指令的非暂时性计算机存储介质, 当指令被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权利要求1-13中任一项所述的相应方法的操作。

16. 一种获得配体的方法, 其中, 配体是药物或工业酶的配体, 所述方法包括:

执行权利要求1-13中任一项所述的方法, 以确定靶蛋白的预测的结构;

评估一个或多个候选配体与靶蛋白的预测的结构的相互作用;

以及根据评估结果来选择一个或多个候选配体作为配体。

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

18. 根据权利要求16或17所述的方法, 其中, 配体是药物, 并且所述方法包括:

执行权利要求1-13中任一项所述的方法, 以确定多个靶蛋白中每个靶蛋白的预测的结构;

评估一个或多个候选配体与每个靶蛋白的预测的结构的相互作用; 和

选择一个或多个候选配体作为配体, 以 i) 获得与每个靶蛋白相互作用的配体, 或 ii) 获得仅与一种靶蛋白相互作用的配体。

19. 一种获得多肽配体的方法, 其中, 配体是药物或工业酶的配体, 所述方法包括:

对于一个或多个候选多肽配体中的每一个, 执行权利要求1-13中任一项所述的方法, 以确定候选多肽配体的预测的结构;

获得靶蛋白的靶蛋白结构;

评估所述一个或多个候选多肽配体的每一个的预测的结构和靶蛋白结构之间的相互作用; 和

根据评估结果选择所述一个或多个候选多肽配体中的一个作为多肽配体。

20. 根据权利要求19所述的方法, 其中, 靶蛋白包括受体或酶, 并且其中, 配体是所述受体或酶的激动剂或拮抗剂; 或者其中多肽配体包含抗体, 靶蛋白包含抗体靶, 特别是病毒或癌细胞蛋白, 并且其中抗体结合抗体靶以提供治疗效果。

21. 一种获得疾病的诊断抗体标记的方法,所述方法包括:

对于一个或多个候选抗体中的每一个,执行权利要求1-13中任一项所述的方法以确定候选抗体的预测的结构;

获得靶蛋白的靶蛋白结构;

评估所述一个或多个候选抗体中的每一个的预测的结构与靶蛋白结构之间的相互作用;和

根据评估结果选择所述一个或多个候选抗体中的一个作为诊断抗体标记。

22. 根据权利要求16-21中任一项所述的方法,其中,评估候选配体之一的相互作用包括确定候选配体的相互作用分数,其中相互作用分数包括候选配体和靶蛋白之间相互作用的量度。

23. 根据权利要求16-22中任一项所述的方法,还包括合成配体。

24. 根据权利要求23所述的方法,还包括在体外和体内测试配体的生物活性。

25. 一种标识蛋白质错误折叠疾病存在的方法,包括:

执行权利要求1-13中任一项所述的方法以确定蛋白质的预测的结构;

获得从人体或动物体获得的蛋白质版本的结构;

将蛋白质的预测的结构与从人体或动物体获得的蛋白质版本的结构进行比较;和

根据比较的结果标识蛋白质错误折叠疾病的存在。

利用循环通过多次迭代预测蛋白质结构

[0001] 相关申请的交叉引用

[0002] 本申请要求2020年11月28日提交的美国临时专利申请序列号63/118,919的申请日的权益,该申请通过引用整体结合于此。

背景技术

[0003] 本说明书涉及预测蛋白质结构。

[0004] 蛋白质由一个或多个氨基酸序列(“链”)来确定。氨基酸是一种有机化合物,它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)。蛋白质折叠是指一个或多个氨基酸序列折叠成三维(3-D)构型的物理过程。蛋白质的结构定义蛋白质进行蛋白质折叠后蛋白质氨基酸序列中原子的三维构型。当在由肽键连接的序列中时,氨基酸可以被称为氨基酸残基。

[0005] 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接收到的输入生成输出,例如预测输出。一些机器学习模型是参数模型,并基于接收的输入和模型的值生成输出。一些机器学习模型是采用多层模型来为接收到的输入生成输出的深度模型。例如,深度神经网络是深度机器学习模型,其包括输出层和一个或多个隐藏层,每个隐藏层对接收的输入应用非线性变换以生成输出。

发明内容

[0006] 本说明书描述了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现的神经网络系统,用于预测蛋白质结构。

[0007] 本说明书中使用的术语“蛋白质(protein)”可以理解为指由一个或多个氨基酸序列(或“链”)指定的任何生物分子。例如,术语“蛋白质”可以指蛋白质域,例如蛋白质的氨基酸链的一部分,其可以几乎独立于蛋白质的其余部分进行蛋白质折叠。作为另一个示例,术语“蛋白质”可以指蛋白质复合物,即包括共同折叠成蛋白质结构的多个氨基酸链。

[0008] 蛋白质中氨基酸链的“多序列比对”(multiple sequence alignment,MSA)指氨基酸链与多个附加氨基酸链(例如来自其它蛋白质,例如同源蛋白质)的序列比对。更具体地,MSA可以定义氨基酸链中的位置和多个附加氨基酸链中的对应位置之间的对应关系。氨基酸链的MSA可以例如通过使用任何合适的计算序列比对技术(例如渐进比对构建)处理氨基酸链的数据库来生成。MSA中的氨基酸链可以理解为具有进化关系,例如,MSA中的每个氨基酸链可以共享共同的祖先。氨基酸链的MSA中氨基酸链中氨基酸之间的相关性可以编码与预测氨基酸链结构相关的信息。

[0009] 实体(例如氨基酸对)的“嵌入(embedding)”可以指将实体表示为数值的有序集,例如数值的向量或矩阵。

[0010] 蛋白质的结构可以由结构参数的集合来定义。定义蛋白质结构的结构参数的集合可以表示为数值的有序集。接下来更详细地描述用于定义蛋白质结构的可能结构参数的几个示例。

[0011] 在一个示例中,定义蛋白质结构的结构参数包括:蛋白质中每个氨基酸的(i)位置参数,和(ii)旋转参数。

[0012] 氨基酸的位置参数可以指定蛋白质结构中氨基酸中特定的原子的预测的3-D空间位置。例如,特定的原子可以是氨基酸中的 α 碳原子,即氨基酸中与氨基官能团、羧基官能团和侧链键合的碳原子。作为另一个示例,特定的原子可以是氨基酸中的 β 碳原子。氨基酸的位置参数可以在任何合适的坐标系中表示,例如3-D[x,y,z]笛卡尔坐标系。

[0013] 氨基酸的旋转参数可以指定氨基酸在蛋白质结构中的预测的“取向(orientation)”。更具体地说,旋转参数可以指定3-D空间旋转操作,如果将其应用于位置参数的坐标系,则导致氨基酸中的三个“主链(main chain)”原子相对于旋转坐标系呈现固定位置。氨基酸中的三个主链原子可以指氨基酸中连接的一系列氮、 α 碳和羰基碳原子。氨基酸的旋转参数可以表示为例如行列式等于1的标准正交 3×3 矩阵。

[0014] 通常,氨基酸的位置和旋转参数定义氨基酸的以自我为中心的参照系。在该参考系中,每个氨基酸的侧链可以从起点开始,并且沿着侧链的第一键(即 α 碳- β 碳键)可以沿着定义的方向。

[0015] 在另一个示例中,定义蛋白质结构的结构参数可以包括“距离图(distance map)”,其表征蛋白质中每对氨基酸之间相应估计的距离(例如,以埃度量)。距离图可以表征氨基酸对之间的估计的距离,例如,通过氨基酸对之间的可能距离的集合的概率分布。

[0016] 在另一个示例中,定义蛋白质结构的结构参数可以定义蛋白质结构中每个氨基酸中每个原子的三维(3D)空间位置。

[0017] 本文描述的蛋白质结构预测系统可用于获得配体,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(例如药物靶标)的氨基酸序列,并使用蛋白质结构预测系统处理基于靶氨基酸序列的输入,以确定靶蛋白的(三级)结构,即预测的蛋白质结构。然后,该方法可以包括评估一个或多个候选配体与靶蛋白结构的相互作用。该方法还可以包括根据相互作用的评估结果选择一个或多个候选配体作为配体。

[0018] 在一些实现中,评估相互作用可以包括评估候选配体与靶蛋白结构的结合。例如,评估相互作用可以包括标识以足够的亲和力结合生物效应的配体。在一些其他实现中,评估相互作用可以包括评估候选配体与对靶蛋白(例如酶)的功能有影响的靶蛋白的结构关联。评估可以包括评估候选配体和靶蛋白结构之间的亲和力,或评估相互作用的选择性。可以根据具有最高亲和力的配体来选择(多个)候选配体。

[0019] (多个)候选配体可以来源于候选配体数据库,和/或可以通过修改候选配体数据库中的配体来获得,例如通过修改候选配体的结构或氨基酸序列,和/或可以通过候选配体的逐步或反复组装/优化来获得。

[0020] 候选配体与靶蛋白结构的相互作用的评估可使用计算机辅助方法执行,其中显示候选配体和靶蛋白结构的图形模型供用户操作,和/或评估可部分或完全自动执行,例如使用标准分子(蛋白质-配体)对接软件。在一些实现中,评估可以包括确定候选配体的相互作用分数,其中相互作用分数包括候选配体和靶蛋白之间相互作用的量度。相互作用分数可以取决于相互作用的强度和/或特异性,例如,分数取决于结合自由能。候选配体可以根据其分数来选择,例如作为具有最高相互作用分数的候选配体。

[0021] 在一些实现中,靶蛋白包括受体或酶,配体是受体或酶的激动剂或拮抗剂。在一些实现中,该方法可用于标识细胞表面标记的结构。然后,这可用于标识配体,例如抗体或标记,诸如荧光标记,其结合细胞表面标记。这可以用于标识和/或治疗癌细胞。

[0022] 在一些实现中,配体是药物,并且确定多个靶蛋白中每一个的预测的结构,并且评估一个或多个候选配体与每个靶蛋白的预测的结构的相互作用。然后可以选择一个或多个候选配体,以获得与每个靶蛋白(功能上)相互作用的配体,或者获得仅与一个靶蛋白(功能上)相互作用的配体。例如,在一些实现中,可能希望获得对多种药物靶标有效的药物。此外或取而代之的是,可能需要针对脱靶效应筛选药物。例如在农业中,确定设计用于一种植物物种的药物不与另一种不同的植物物种和/或动物物种相互作用可能是有用的。

[0023] 在一些实现中,(多个)候选配体可以包括小分子配体,例如分子量<900道尔顿的有机化合物。在一些其他实现中,(多个)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。

[0024] 在一些情况下,蛋白质结构预测系统可用于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;可以使用结构预测神经网络或使用常规的物理研究技术如x射线晶体学和/或磁共振技术和/或低温电子显微镜来确定靶蛋白的结构。

[0025] 另一方面,提供了使用蛋白质结构预测系统获得多肽配体(例如,分子或其序列)的方法。该方法可包括获得一个或多个候选多肽配体的氨基酸序列。该方法可以进一步包括使用蛋白质结构预测系统来确定候选多肽配体的(三级)结构。该方法可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,并评估一个或多个候选多肽配体中每一个的结构与靶蛋白结构之间的相互作用。该方法还可以包括根据评估结果选择一个或多个候选多肽配体作为多肽配体。

[0026] 如前所述,评估相互作用可包括评估候选多肽配体与靶蛋白结构的结合,例如,标识对生物效应具有足够亲和力的配体,和/或评估候选多肽配体与对靶蛋白(例如酶)功能有影响的靶蛋白结构的关联,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。在一些实现中,多肽配体可以是适体。同样,(多个)多肽候选配体可以根据具有最高亲和力的配体进行选择。

[0027] 如前所述,所选的多肽配体可以包含受体或酶,并且配体可以是受体或酶的激动剂或拮抗剂。在一些实现中,多肽配体可包含抗体,靶蛋白可包含抗体靶,例如病毒,特别是病毒外壳蛋白,或癌细胞上表达的蛋白。在这些实现中,抗体与抗体靶结合以提供治疗效果。例如,抗体可以与靶结合,并作为特定受体的激动剂;或者,抗体可以阻止另一个配体与靶的结合,从而阻止相关生物途径的激活。

[0028] 该方法的实施可以进一步包括合成,即制造小分子或多肽配体。配体可以通过任何常规化学技术合成和/或已经可以获得,例如,可以来自化合物库或可以使用组合化学合成。

[0029] 该方法可以进一步包括在体外和/或体内测试配体的生物活性。例如,可以测试配体的ADME(吸收、分布、代谢、排泄)和/或毒理学特性,以筛选出不合适的配体。测试可包括,例如,使候选小分子或多肽配体与靶蛋白接触,并度量蛋白表达或活性的变化。

[0030] 在一些实现中,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

变域抗体、双或多特异性抗体、多价抗体、双可变域抗体、免疫缀合物、纤连蛋白分子、附着蛋白、DARPin、抗体、亲和体、抗卡林、亲和蛋白、蛋白表位模拟物或其组合。候选(多肽)配体可包括具有突变或化学修改的氨基酸Fc区的抗体,例如,当与野生型Fc区相比时,其防止或降低ADCC(antibody-dependent cellular cytotoxicity,抗体依赖性细胞毒性)活性和/或增加半衰期。候选(多肽)配体可以包括具有不同CDR (Complementarity-Determining Region,互补决定区)的抗体。

[0031] 本文描述的蛋白质结构预测系统也可用于获得疾病的诊断抗体标记。还提供了一种方法,对于例如如上所述的一个或多个候选抗体中的每一个,使用蛋白质结构预测系统来确定候选抗体的预测的结构。该方法还可以包括获得靶蛋白的靶蛋白结构,评估一个或多个候选抗体中每一个的预测的结构与靶蛋白结构之间的相互作用,以及根据评估结果选择一个或多个候选抗体中的一种作为诊断抗体标记,例如选择一个或多个对靶蛋白结构具有最高亲和力的候选抗体。该方法可以包括制造诊断抗体标记。诊断抗体标记可用于通过检测其是否与从患者获得的样品(例如体液样品)中的靶蛋白结合来诊断疾病。如上所述,对应的技术可用于获得治疗性抗体(多肽配体)。

[0032] 错误折叠的蛋白质与许多疾病有关。因此,在另一方面,提供了使用蛋白质结构预测系统来标识蛋白质错误折叠疾病的存在的方法。该方法可以包括获得蛋白质的氨基酸序列,并使用蛋白质结构预测系统来确定蛋白质的结构。该方法还可以包括获得从人体或动物体获得的蛋白质的版本的版本的结构,例如通过常规(物理)方法。然后,该方法可以包括将蛋白质的结构与从身体获得的版本的版本的结构进行比较,并根据比较的结果标识蛋白质错误折叠疾病的存在。也就是说,通过与计算机确定的结构进行比较,可以确定来自体内的蛋白质版本的错误折叠。

[0033] 一般来说,标识蛋白质错误折叠疾病的存在可以包括获得蛋白质的氨基酸序列,使用蛋白质的氨基酸序列来确定蛋白质的结构,如本文所述,并且将蛋白质的结构与蛋白质的基线版本的结构进行比较,根据比较结果来标识蛋白质错误折叠疾病的存在。例如,比较的结构可以是突变体和野生型蛋白质的结构。在实现中,野生型蛋白质可以用作基线版本,但是原则上也可以用作基线版本。

[0034] 在一些其他方面,如上所述或本文所述的计算机实施的方法可用于从其氨基酸序列中标识靶蛋白上的活性/结合/封闭位点。

[0035] 根据一个方面,提供了一种由一个或多个计算机执行的方法,用于预测包含一条或多条链的蛋白质的结构,其中每条链包含氨基酸序列,该方法包括:在包括第一次迭代以及随后的一次或多次后续迭代的迭代序列的第一次迭代中:获得表征蛋白质的第一次迭代的网络输入;使用蛋白质结构预测神经网络处理第一次迭代的网络输入,以生成定义蛋白质的初始预测的结构的结构参数;在迭代序列中的每次后续迭代中:获得表征蛋白质的后续迭代的网络输入;从(i)在序列中后续迭代之前的先前迭代中生成的结构参数,(ii)在最后迭代中生成结构参数时由蛋白质结构预测神经网络生成的一个或中间输出,或(iii)两者,生成后续迭代的特征;以及使用蛋白质结构预测神经网络处理用于后续迭代的特征和网络输入,以生成定义蛋白质的另一预测的结构的用于后续迭代的结构参数;以及根据序列中最后迭代的结构参数确定蛋白质的最终预测的结构。

[0036] 在一些实现中,迭代序列中的每次迭代的网络输入是相同的网络输入。

[0037] 在一些实现中,使用蛋白质结构预测神经网络处理用于后续迭代的特征和网络输入,以生成定义蛋白质的预测的结构的用于后续迭代的结构参数,包括:根据用于后续迭代的特征和网络输入来生成组合的输入;以及使用蛋白质结构预测神经网络处理组合的输入,以生成用于后续迭代的结构参数。

[0038] 在一些实现中,每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入(pair embedding);在每次迭代中,结构预测神经网络被配置为在生成迭代的结构参数时重复更新初始对嵌入;生成特征包括从在先前迭代中生成结构参数时生成的更新的对嵌入中生成对嵌入的变换的集合;以及生成组合的输入包括为后续迭代组合对嵌入的变换的集合和初始对嵌入。

[0039] 在一些实现中,每次迭代的相应网络输入包括代表与蛋白质中每条链相对应的相应多序列比对(MSA)的初始MSA表示;结构预测神经网络被配置为生成单个嵌入的一个和多个集合,每个单个嵌入集合包括在生成迭代的结构参数时蛋白质的每个氨基酸的相应单个嵌入;生成特征包括从在先前迭代中生成结构参数时生成的单个嵌入集合之一生成单个嵌入的变换的集合;并且生成组合的输入包括为后续迭代组合单个嵌入的变换的集合和初始MSA表示。

[0040] 在一些实现中,为后续迭代组合单个嵌入的变换的集合和初始MSA表示包括将单个嵌入的变换的集合相加到初始MSA表示的第一行。

[0041] 在一些实现中,每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入;在每次迭代中,结构参数为每个氨基酸指定了蛋白质结构中氨基酸中特定的原子的预测的3-D空间位置;生成特征包括:根据先前迭代中由结构参数指定的氨基酸的预测的3-D空间位置生成距离图,所述距离图为蛋白质中的每对氨基酸表征蛋白质结构中氨基酸对之间的相应估计的距离;以及根据距离图生成具有与初始对嵌入相同的维度的变换的距离图;以及生成组合的输入包括为后续迭代组合变换的距离图和初始对嵌入。

[0042] 在一些实现中,每次迭代的相应网络输入包括蛋白质中每对氨基酸的相应初始对嵌入;在每次迭代中,结构参数指定距离图,该距离图为蛋白质中的每对氨基酸表征蛋白质结构中氨基酸对之间相应估计的距离;生成特征包括:根据在先前迭代中由结构参数指定的距离图,生成具有与初始对嵌入相同的维度变换的距离图;以及生成组合的输入包括为后续迭代组合变换的距离图和初始对嵌入。

[0043] 在一些实现中,每次迭代的相应网络输入包括使用一个或多个模板序列的集合和每个模板序列的对应已知结构生成的蛋白质中每对氨基酸的相应初始对嵌入;以及生成组合的输入包括通过将蛋白质和由先前迭代中的嵌入定义的结构预测相加到一个或多个模板序列的集合和相应的已知结构中,来修改用于后续迭代的初始对嵌入。

[0044] 根据另一方面,提供了一种训练结构预测神经网络的方法,该方法包括:选择要执行的迭代次数;为每次迭代获得相应的训练网络输入,每个训练网络输入表征训练蛋白质;在包括所选的迭代次数的迭代序列中的第一次迭代中:使用蛋白质结构预测神经网络处理第一次迭代的网络输入,以生成定义训练蛋白质的初始预测的结构的用于第一次迭代的结构参数;当所选的迭代次数大于1时,在迭代序列中的每次后续迭代中,直到已经执行了所选的迭代次数:从(i)在序列中后续迭代之前的先前迭代中生成的结构参数,(ii)在最后迭代中生成结构参数时由蛋白质结构预测神经网络生成的一个或中间输出,或(iii)两者,生

成后续迭代的特征;和使用蛋白质结构预测神经网络处理用于后续迭代的特征和训练网络输入,以生成定义训练蛋白质的另一预测的结构用于后续迭代的结构参数;以及确定包括结构损失项的目标函数的梯度,所述结构损失项表征(i)由序列中最后迭代生成的结构参数指定的靶蛋白的预测的结构和(ii)训练蛋白质的靶蛋白结构之间的相似性;以及使用梯度更新结构预测神经网络的参数值,从序列中最后迭代的结构参数得到蛋白质的最终预测的结构。

[0045] 在一些实现中,确定梯度包括在由结构预测神经网络在最后迭代中作为输入处理的特征处应用停止梯度。

[0046] 在一些实现中,选择迭代次数包括:从概率分布中采样迭代次数,所述概率分布为每个迭代次数分配相应概率,迭代次数在从1到结构预测神经网络被训练后使用的总迭代次数的范围内。

[0047] 在一些实现中,该方法还包括:当所选的迭代次数等于1时:确定包括结构损失项的目标函数的梯度,所述结构损失项表征(i)由序列中第一次迭代生成的结构参数定义的靶蛋白的初始预测的结构和(ii)训练蛋白质的靶蛋白结构之间的相似性;以及使用梯度来更新结构预测神经网络的参数值。

[0048] 可以实现本说明书中描述的主题的特定实施例,以便实现一个或多个以下优点。

[0049] 本说明书描述了通过多次迭代预测蛋白质结构的技术。特别地,在每次迭代中,系统使用神经网络进行关于蛋白质结构的预测,并且在第一次迭代后的每次特定迭代中,当使用神经网络生成特定迭代的预测时,系统使用(“循环(recycles)”)来自先前迭代的特征。这样生成预测的方式通常被称为“循环”。这有效地生成了“更深”的神经网络,即,具有更大数量的神经网络层的神经网络,而不需要修改神经网络的结构,并且可以极大地增加由神经网络做出的预测的准确性。此外,本说明书描述了利用循环来训练神经网络的技术,该技术不增加用于训练神经网络的训练过程的存储器消耗(相对于用在单次迭代中生成最终预测的相同架构来训练神经网络),并且最小化附加迭代所必需的训练时间的增加。

[0050] 蛋白质的结构决定了蛋白质的生物功能。因此,确定蛋白质结构可以有助于理解生命过程(例如,包括许多疾病的机制)和设计蛋白质(例如,作为药物,或作为用于工业过程的酶)。例如,哪些分子(例如,药物)将与蛋白质结合(以及结合将在哪里发生)取决于蛋白质的结构。由于药物的有效性会受到它们与蛋白质(例如,在血液中)结合程度的影响,因此确定不同蛋白质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)来确定蛋白质结构可能是耗时且非常昂贵的。因此,本说明书中描述的蛋白质预测系统可以促进涉及蛋白质的生物化学研究和工程领域(例如,药物开发)。

[0051] 本说明书主题的一个或多个实施例的细节在附图和以下描述中阐述。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

[0052] 图1A显示了蛋白质结构预测系统的示例。

[0053] 图1B示出了示例结构预测神经网络。

[0054] 图1C是用于预测蛋白质结构的示例过程的流程图。

[0055] 图2示出了嵌入神经网络的示例架构。

- [0056] 图3示出了嵌入神经网络的更新块的示例架构。
- [0057] 图4示出了MSA更新块的示例架构。
- [0058] 图5示出了更新块的对的示例架构。
- [0059] 图6示出了折叠神经网络的示例架构。
- [0060] 图7说明了氨基酸中键之间的扭转角。
- [0061] 图8是未折叠的蛋白质和折叠的蛋白质的示意图。
- [0062] 图9是用于预测蛋白质结构的另一个示例过程的流程图。
- [0063] 图10示出了用于生成蛋白质中氨基酸链的MSA表示的示例过程。
- [0064] 图11示出了用于为蛋白质中的每对氨基酸生成相应的对嵌入的示例过程。
- [0065] 不同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

[0066] 图1A示出了示例蛋白质结构预测系统100。蛋白质结构预测系统100是在一个或多个位置的一个或多个计算机上实现为计算机程序的系统的示例,其中实现了下述系统、组件和技术。

[0067] 系统100被配置为处理定义蛋白质104的一个或多个氨基酸链102的数据,以生成定义预测的蛋白质结构108的结构参数的集合106,即蛋白质104的结构预测。也就是说,蛋白质104的预测的结构108可以由结构参数的集合106来定义,这些结构参数共同定义了蛋白质进行蛋白质折叠后的预测的3-D结构。

[0068] 定义预测的蛋白质结构108的结构参数可以如前所述。例如,它们可以包括例如蛋白质104中每个氨基酸的位置参数和旋转参数、表征蛋白质中每对氨基酸之间的估计的距离的距离图、蛋白质结构中每个氨基酸中每个原子或主链原子相应的空间位置、或它们的组合,如上所述。

[0069] 为了生成定义预测的蛋白质结构108的结构参数106,系统100至少从定义蛋白质104的氨基酸链102的数据生成表征蛋白质104的网络输入140。

[0070] 作为特定的示例,表征蛋白质104的网络输入140可以包括:(i)蛋白质的多序列比对(MSA)表示,和/或(ii)蛋白质的“对(pair)”嵌入的集合。这些将在下面参照图1B至图11更详细地描述。

[0071] 系统100然后生成结构参数,该结构参数定义多次($N+1$,其中 N 是正整数)迭代(即在图1A的示例中的迭代0到 N)的预测的蛋白质结构108。在多次迭代的每一次,系统100生成结构参数的集合,其定义蛋白质104的预测蛋白质结构。然后,系统使用在最后迭代中生成的结构参数作为定义预测的蛋白质结构108的最终结构参数。

[0072] 在一些实现中,在训练之后,即在对新蛋白质进行预测时的推断时,系统100生成最终预测的迭代次数可以是固定的。也就是说,在训练之后,可以将 N 设置为固定的整数,使得总执行 $N+1$ 次迭代,例如,以平衡计算效率(例如,延迟)和预测质量之间的期望折衷的方式。

[0073] 在一些其他实现中,在训练之后,系统100可以改变执行的迭代次数。例如,系统100可以基于分配给处理的时间预算或资源预算来改变迭代次数,例如,通过执行符合预算的最大迭代次数。作为另一个示例,系统100可以采用置信度估计技术来估计在任何给定迭

代中结构预测是准确的置信度,并且使用置信度估计来确定执行多少次迭代。例如,一旦置信度估计超过阈值或者一旦置信度估计已经停止改善,系统可以停止执行迭代。下面参考图6描述生成置信度估计的示例。

[0074] 在训练期间,系统100可以改变所执行的迭代次数,即,使得N的值对于不同的网络输入而变化。这将在下面参考图1B和图1C更详细地描述。

[0075] 更具体地,在第一次迭代(“迭代0”),系统100使用结构预测神经网络150处理网络输入140,以生成第一次迭代的结构参数的集合106-0。

[0076] 结构预测神经网络150是被配置为处理网络输入140以生成定义蛋白质104的预测的结构的结构参数的集合的神经网络。也就是说,即使处理发生在多次迭代中,在每次迭代中,神经网络150生成定义蛋白质104的完整结构的结构参数的完整的集合。下面参照图1B-图11更详细地描述结构预测神经网络150的具体示例。

[0077] 在每次后续迭代中,即在第一次迭代之后的每次迭代中,系统100从在先前迭代中执行的处理的结果中生成特征。

[0078] 特别地,系统100从(i)由结构预测神经网络150在先前迭代中生成的结构参数,(ii)由结构预测神经网络150在先前迭代中生成结构参数期间生成的中间输出,或(iii)两者生成特征。

[0079] 下面将参考图1B更详细地描述这些特征的生成。

[0080] 然后,系统100使用结构预测神经网络150处理网络输入140和由在先前迭代中执行的处理的结果生成的特征,以生成用于迭代的结构参数的集合。因此,在第一次迭代之后的每次迭代中,系统100生成结构参数的集合,结构参数的集合不仅取决于网络输入140,还取决于来自先前迭代的特征。

[0081] 例如,在迭代N-1,系统100使用结构预测神经网络150处理网络输入140和来自迭代N-2 142的特征,以生成用于下一次迭代的结构参数106-N-1。

[0082] 在最后迭代,即迭代N,系统100使用结构预测神经网络150处理网络输入140和倒数第二次迭代N-1 144的特征,以生成最后迭代的结构参数106-N。

[0083] 系统100然后使用这些结构参数106-N作为定义预测的蛋白质结构108的最终结构参数。因此,即使在每次迭代中生成完整的预测,也只有来自最后迭代的预测被用作定义预测的蛋白质结构108的最终预测。

[0084] 图1B示出了示例结构预测神经网络150。特别地,图1B示出了由神经网络150在多次迭代0至N中的特定一次迭代中执行的处理。

[0085] 如图1B所示,神经网络150处理网络输入140,其包括(i)蛋白质的多序列比对(MSA)表示110,和/或(ii)蛋白质的“对”嵌入的集合110。

[0086] 蛋白质的MSA表示110包括蛋白质中每个氨基酸链的MSA的相应表示。蛋白质中氨基酸链的MSA表示可以表示为 $M \times N$ 嵌入阵列(即,具有M行和N列的2-D嵌入阵列),其中N是氨基酸链中氨基酸的数量。MSA表示的每一行可以与蛋白质中氨基酸链的相应MSA序列相对应。参考图10描述了用于生成蛋白质中氨基酸链的MSA表示的示例过程。

[0087] 系统100从蛋白质中氨基酸链的MSA表示生成蛋白质104的MSA表示110。

[0088] 如果蛋白质仅包括单个氨基酸链,则系统100可以将蛋白质104的MSA表示110标识为蛋白质中单个氨基酸链的MSA表示。

[0089] 如果蛋白质包括多个氨基酸链,则系统100可以通过将蛋白质中氨基酸链的MSA表示组装成嵌入的块对角线2-D阵列来生成蛋白质的MSA表示110,即,其中蛋白质中氨基酸链的MSA表示形成对角线上的块。系统100可以将对角线上的块之外的2-D阵列中的每个位置处的嵌入初始化为默认嵌入,例如零向量。蛋白质中的氨基酸链可以任意排序,蛋白质中氨基酸链的MSA表示可以在块对角矩阵中相应地排序。例如,第一个氨基酸链的MSA表示(即,根据排序)可以是对角线上的第一个块,第二个氨基酸链的MSA表示可以是对角线上的第二个块,等等。

[0090] 通常,蛋白质的MSA表示110可以表示为嵌入的2-D阵列。在整个说明书中,蛋白质的MSA表示的“行”是指定义蛋白质的MSA表示的2-D嵌入阵列的行。类似地,蛋白质的MSA表示的“列”是指定义蛋白质的MSA表示的2-D嵌入阵列的列。

[0091] 对嵌入的集合112包括与蛋白质104中每对氨基酸相对应的相应对嵌入。一般来说,对嵌入表示即编码关于蛋白质104中氨基酸对之间关系的信息。氨基酸对是指蛋白质中包括第一个氨基酸和第二个氨基酸的有序元组,即蛋白质中可能的氨基酸对的集合由下式给出:

$$[0092] \quad \{(A_i, A_j) : 1 \leq i, j \leq N\} \quad (1)$$

[0093] 其中N是蛋白质中的氨基酸数量, $i, j \in \{1, \dots, N\}$ 索引蛋白质中的氨基酸, A_i 是i索引的蛋白质中的氨基酸, A_j 是j索引的蛋白质中的氨基酸。如果蛋白质包含多个氨基酸链,则蛋白质中的氨基酸可以根据蛋白质中氨基酸链的排序顺序索引 $\{1, \dots, N\}$ 。也就是说,首先对第一个氨基酸链的氨基酸进行顺序索引,接着是第二个氨基酸链的氨基酸,接着是第三个氨基酸链的氨基酸,依此类推。对嵌入的集合112可以表示为对嵌入的2-D $N \times N$ 阵列,例如,其中2-D阵列的行由 $i \in \{1, \dots, N\}$ 索引,2-D阵列的列由 $j \in \{1, \dots, N\}$ 索引,并且2-D阵列中的位置 (i, j) 由氨基酸对 (A_i, A_j) 的对嵌入占据。

[0094] 参考图11描述了用于生成(初始化)与蛋白质中每对氨基酸相对应的相应对嵌入的示例过程。

[0095] 系统100可以基于对嵌入来确定蛋白质的预测的结构。然而,在实现中,系统100使用MSA表示110和对嵌入112两者来生成定义预测的蛋白质结构108的结构参数106,因为两者都具有互补的属性。MSA表示110的结构可以显式地依赖于蛋白质中每个氨基酸链相对应的MSA中氨基酸链的数量。因此,MSA表示110可能不适合用于直接预测蛋白质结构,因为蛋白质结构108对MSA中氨基酸链的数量没有显式地依赖性。相比之下,对嵌入112表征蛋白质104中各对氨基酸之间的关系,并且在没有显式参考MSA的情况下表达,因此是用于预测蛋白质结构108的方便有效的数据表示。

[0096] 神经网络150包括嵌入神经网络200和折叠神经网络600。

[0097] 系统100根据嵌入神经网络200的参数值的集合,使用嵌入神经网络200处理MSA表示110和对嵌入112,以更新MSA表示110和对嵌入112。也就是说,嵌入神经网络200处理MSA表示110和对嵌入112,以生成更新的MSA表示114和更新的对嵌入116。

[0098] 嵌入神经网络200通过在MSA表示110和对嵌入112之间共享信息来更新MSA表示110和对嵌入112。更具体地,嵌入神经网络200在基于当前对嵌入112更新当前MSA表示110和基于当前MSA表示110更新当前对嵌入112之间交替。

[0099] 参考图2更详细地描述了嵌入神经网络200的示例架构。

[0100] 更一般地,虽然图2和后续的附图的描述描述了嵌入神经网络200的架构,其中每个块更新对嵌入和MSA表示两者,但是所描述的用于将辅助折叠神经网络并入结构预测系统的技术可以与包括每个更新对嵌入的多个块的嵌入神经网络的任何适当架构一起使用。作为下述体系结构的替代体系结构的特定示例,嵌入神经网络200可以采用这样的体系结构,其中使用MSA表示110来初始化对嵌入112,并且其中每个块仅接收当前的对嵌入,并且通过如下所述应用自注意力(self-attention)来更新那些嵌入,但是不依赖于MSA表示。

[0101] 神经网络150根据更新的对嵌入116、更新的MSA表示114或两者生成折叠神经网络600的输入,并使用折叠神经网络600处理输入以生成定义预测的蛋白质结构的结构参数106。

[0102] 在一些实现中,折叠神经网络600处理更新的对嵌入116以生成距离图,该距离图包括蛋白质中每对氨基酸在蛋白质结构中氨基酸对之间的可能距离的集合上的概率分布。例如,为了在蛋白质结构中氨基酸对之间的可能距离的集合上生成概率分布,折叠神经网络可以将一个或多个完全连接的神经网络层应用于与氨基酸对相对应的更新的对嵌入116。

[0103] 在一些实现中,折叠神经网络600通过使用几何注意力操作处理从更新的MSA表示114和更新的对嵌入116两者得出的输入来生成结构参数106,该几何注意力操作显式地推理蛋白质结构中氨基酸的3-D几何。参考图6描述了实现几何注意力机制的折叠神经网络600的示例架构。

[0104] 如上所述,在N+1次迭代的第一次迭代中,神经网络150仅处理网络输入140。

[0105] 然而,在每次后续迭代中,系统100使用由神经网络150在先前迭代中生成的输出来生成特征113。

[0106] 特别地,系统100使用由神经网络150在先前迭代中生成的结构参数、由神经网络150在先前迭代中生成的中间输出或两者来生成特征113。

[0107] 更具体地,在任何给定的后续迭代中,系统100通过首先根据特征113和网络输入140生成组合的输入,然后使用神经网络150处理组合的输入以生成后续迭代的结构参数106,来使用神经网络150处理特征113和网络输入140两者。

[0108] 如果后续的迭代是最后迭代,则系统100使用在迭代中生成的结构参数106作为定义蛋白质104的预测的结构108的最终结构参数。

[0109] 一般而言,在由神经网络150生成的中间输出用于生成特征113的实现中,系统100可以使用神经网络150的任何层的输出(除了折叠神经网络600的最终层的输出)、神经网络150的任何层的输入(除了折叠神经网络150的输入层)或两者来生成特征。

[0110] 作为特定示例,系统100可以使用由嵌入神经网络200的最后更新块生成的更新的对嵌入116或者由折叠神经网络600的中间层之一生成的更新的对嵌入作为生成特征113的至少一部分,这将在下面更详细地描述。

[0111] 在该示例中,系统100可以至少部分地通过将变换应用于来自先前迭代的更新的对嵌入来生成特征,然后将特征与初始对嵌入112组合,例如相加、平均或级联,作为为后续迭代生成组合的输入的一部分。

[0112] 作为一个示例,系统100可以通过对更新的对嵌入应用归一化运算,例如层归一化运算(例如,如参考吉米巴雷等人的“层归一化”,arXiv:1607.06450所描述的),来变换更新

的对嵌入。

[0113] 作为另一个示例,系统100可以通过使用多层感知器(MLP)神经网络处理更新的对嵌入来变换更新的对嵌入,以生成初始变换的嵌入,然后对初始变换的嵌入应用归一化运算来生成变换的嵌入。

[0114] 作为另一个特定的示例,系统100可以使用作为折叠神经网络600的输入提供的蛋白质中每个氨基酸的“单个”嵌入,或者由折叠神经网络600的中间层之一生成的单个嵌入,作为生成特征113的一部分。下面参考图6详细描述了为蛋白质中的氨基酸生成这些单个嵌入以及使用折叠神经网络600更新单个嵌入。

[0115] 在该示例中,系统100可以变换来自先前迭代的单个嵌入,例如,如上文参考对嵌入所描述的,然后将变换的单个嵌入与初始MSA表示110组合,作为生成后续迭代的输入的一部分。例如,系统100可以将变换的单个嵌入与初始MSA表示110的第一行进行组合,例如相加或平均。

[0116] 在系统100使用在先前迭代中生成的结构参数106作为生成特征113的一部分的一些实现中,系统100可以将来自先前迭代的结构参数106变换成具有与对嵌入相同维度的嵌入,然后将变换的结构参数与初始对嵌入110组合,作为生成组合得网络输入的一部分。

[0117] 在一些实现中,系统100直接组合变换的结构参数和初始对嵌入,例如,系统100相加、平均或级联变换的结构参数和初始对嵌入110。

[0118] 在一些其他实现中,系统100可以通过应用学习的组合,即与嵌入神经网络200和折叠神经网络600的训练一起学习的组合,来组合初始对嵌入110和变换的结构参数。作为特定的示例,系统100可以应用学习的交叉注意力(cross-attention)机制来结合来自变换的结构参数的信息。与下面描述的自注意力机制不同,在交叉注意力机制中,从更新的对嵌入中生成查询,而从变换的结构参数中生成键和值。

[0119] 特别地,如上所述,在一些实现中,结构参数106是距离图,其表征蛋白质中每对氨基酸之间相应估计的距离。在这些情况下,系统100可以将距离图投影到与对嵌入相同的维度,以生成变换的结构参数。

[0120] 例如,系统可以通过将每个距离分配给多个箱中的箱来离散化距离图中的每个距离,每个箱与不同范围的距离值相对应。系统然后可以嵌入每个离散化得距离,例如,通过将离散化距离的独热编码表示乘以嵌入矩阵。独热编码表示将离散化距离表示为向量,该向量在与离散化距离相对应的位置包括“1”,在与所有其他离散化距离相对应的所有其他位置包括“0”。

[0121] 作为另一个示例,系统可以将核函数应用于距离图中的每个距离,以生成距离的表示,然后嵌入这些表示。例如,核函数可以等于函数 $\exp(-\text{distance})$,或者更一般地,与函数 $\exp(-\text{distance})$ 成正比。

[0122] 作为另一个示例,如上所述,在一些其他实现中,结构参数包括为每个氨基酸指定蛋白质结构中氨基酸中特定的原子的预测的3-D空间位置的结构参数。系统100可以基于由结构参数定义的位置,通过计算蛋白质中氨基酸中特定的原子对之间的距离来构建距离图。在这些情况下,系统100可以将距离图变换到与更新的对嵌入相同的维度,以生成如上所述的变换的结构参数。

[0123] 如下文将更详细描述,在一些实现中,系统100至少部分基于一个或多个模板序

列的集合来生成对嵌入112。每个模板序列是蛋白质中氨基酸链的MSA序列，其中模板序列1110的折叠结构是已知的，例如通过物理实验。

[0124] 在系统100使用在先前迭代中生成的结构参数106作为生成特征113的一部分的一些实现中，系统100可以相加蛋白质104和由来自先前迭代的结构参数106定义的蛋白质的折叠的结构，作为用于生成对嵌入112的模板序列中的另一个，对嵌入112是后续迭代的组合的网络输入的一部分。下面参考图11描述使用模板序列来生成对嵌入112。

[0125] 上面的描述一般描述了相同的网络输入140用于所有的多次迭代。然而，在一些情况下，系统对于不同的迭代使用不同的网络输入140。作为一个特定示例，系统可以为每次迭代从较大的MSA序列集合中采样不同的MSA的集合，导致为多次迭代中的每次迭代生成不同的MSA表示110。作为另一个示例，在使用模板序列来更新生成对嵌入112的实现中，对于第一次迭代（“迭代0”），系统可以在不使用任何模板序列的情况下生成对嵌入112。

[0126] 训练引擎可以端到端地训练蛋白质结构预测神经网络150，以优化目标函数，该目标函数包括本文称为结构损失的项。

[0127] 训练引擎可以在包括多个训练示例的训练数据的集合上训练系统100。每个训练示例可以指定：(i) 包括蛋白质的初始MSA表示和初始对嵌入的训练输入，以及(ii) 应该由系统100通过处理训练输入而生成的靶蛋白结构。用于训练系统100的靶蛋白结构可以使用实验技术（例如x射线晶体学或冷冻电子显微镜）来确定。

[0128] 结构损失可以表征以下两者之间的相似性：(i) 由系统100生成的预测蛋白质结构，即，由最后迭代的最终结构参数定义的预测蛋白质结构，和(ii) 应该已经由系统生成的靶蛋白结构。

[0129] 例如，如果预测的结构参数定义蛋白质中每个氨基酸的预测的位置参数和预测的旋转参数，那么结构损失 $L_{\text{structure}}$ 可以由下式给出：

$$[0130] \quad L_{\text{structure}} = \frac{1}{N^2} \sum_{i,j=1}^N \left(1 - \frac{|t_{ij} - \tilde{t}_{ij}|}{A} \right)_+ \quad (2)$$

$$[0131] \quad t_{ij} = R_i^{-1}(t_j - t_i) \quad (3)$$

$$[0132] \quad \tilde{t}_{ij} = \tilde{R}_i^{-1}(\tilde{t}_j - \tilde{t}_i) \quad (4)$$

[0133] 其中，N是蛋白质中氨基酸的数量， t_i 表示氨基酸i的预测的位置参数， R_i 表示由氨基酸i的预测的旋转参数指定的 3×3 旋转矩阵， $\tilde{t}_i$ 是氨基酸i的目标位置参数， $\tilde{R}_i$ 表示由氨基酸i的目标旋转参数指定的 3×3 旋转矩阵，A是常数， R_i^{-1} 表示由预测的旋转参数 R_i 指定的 3×3 旋转矩阵的逆， $\tilde{R}_i^{-1}$ 表示由目标旋转参数 $\tilde{R}_i$ 指定的 3×3 旋转矩阵的逆，并且 $(\cdot)_+$ 表示校正的线性单元(ReLU)操作。

[0134] 参照等式(2) - (4)定义的结构损失可以理解为蛋白质中每对氨基酸的平均损失 $|t_{ij} - \tilde{t}_{ij}|$ 。项 t_{ij} 定义氨基酸j在氨基酸i预测的参照系中的预测的空间位置，并且 $\tilde{t}_{ij}$ 定义了氨基酸j在实际氨基酸i参照系中的实际空间位置。这些项对氨基酸i和j的预测的和实际旋转敏感，因此比只对氨基酸之间的预测和实际距离敏感的损失项携带更丰富的信息。

[0135] 作为另一个示例,如果预测的结构参数定义蛋白质的每个氨基酸中每个原子的预测的空间位置,那么结构损失可以是(i)原子的预测的空间位置和(ii)原子的目标(例如,地面实况)空间位置之间的平均误差(例如,平方误差)。

[0136] 优化结构损失促使系统100生成精确接近真实蛋白质结构的预测蛋白质结构。

[0137] 除了结构损失之外,目标函数可以包括一个或多个辅助损失。例如,基于预测的结构中氨基酸中原子之间的键的键角和/或键长,或者基于预测的结构中不同氨基酸中原子的接近度,辅助损失可能不利于具有在自然界中不可能出现的特征的预测的结构。

[0138] 在一些实现中,辅助损失还包括度量由倒数第二次迭代生成的结构参数定义的结构预测和由最后迭代的结构参数定义的结构预测之间的差的辅助损失,即,惩罚神经网络150在连续迭代生成类似预测的辅助损失。这可以促使神经网络150不仅仅在每次后续迭代中改进初始预测,而是生成可以更准确地表示蛋白质实际结构的新预测。在这些实现中,在训练之后,如果需要多个似是而非的结构预测,系统100可以在多次迭代中输出结构预测作为相同蛋白质的不同结构预测。

[0139] 在训练期间,系统100在最后迭代中用作输入的特征113处应用停止梯度(stopgradient)。停止梯度防止梯度流过特征113回到在任何早期迭代中执行的处理中。

[0140] 更详细地,在最后迭代中生成的预测的结构不仅取决于在最后迭代中执行的处理,而且因为特征113也作为最后迭代中的输入而被提供,所以还取决于在每个较早的迭代中执行的处理。在没有停止梯度的情况下,训练引擎可以通过特征113反向传播梯度,并且对于神经网络150的每个参数,不仅对于最后迭代,而且对于每个先前迭代,计算目标函数相对于该参数的梯度。因为应用了停止梯度,所以系统不通过特征113反向传播任何梯度,并且仅计算关于每个参数的单个梯度,即基于在最后迭代中发生的处理计算的梯度。相对于在迭代0生成的预测总是被用作最终预测的框架,应用停止梯度可以防止再循环框架导致训练时间和计算(例如处理器周期)的显著增加。

[0141] 图1C是用于预测蛋白质结构的示例过程1300的流程图。为了方便起见,过程1300将被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来执行。例如,根据本说明书适当编程的蛋白质结构预测系统,例如图1A的蛋白质结构预测系统100,可以执行过程1300。

[0142] 该系统可以在推断期间,即在训练神经网络之后,执行过程1300,以对新蛋白质进行预测,或者在训练期间,使用具有已知目标结构的蛋白质来确定对结构预测神经网络的参数值的更新。

[0143] 系统选择预测蛋白质结构要执行的迭代次数(步骤1302)。

[0144] 具体地,如上所述,在一些情况下,当在推断期间执行过程1300时,系统为需要新预测的每个新蛋白质选择相同次数的迭代。

[0145] 在这些情况下,当在训练期间执行过程1300时,系统可以为过程1300的不同迭代选择不同的迭代次数。特别地,选择不同的迭代次数允许系统确保接收到在所有迭代中执行的处理的梯度反馈,而不仅仅是推断中最后迭代的迭代,即使如上所述应用了停止梯度。

[0146] 作为特定示例,在训练期间,系统可以从概率分布中对迭代次数进行采样,该概率分布将相应概率分配给从1次总迭代到N+1次(即,在训练之后执行的迭代次数是固定的实现中在推断期间使用的总迭代次数)的每个迭代次数。在训练后执行的迭代次数不固定的

实现中, $N+1$ 可以被设置为等于在推断期间可以执行的最大迭代次数或者代表可能的最大迭代次数的某个其他固定值。

[0147] 在一些实现中, 概率分布是均匀分布, 其将相同的概率分配给从1到 $N+1$ 的迭代次数中的每个迭代次数。

[0148] 在一些其他实现中, 概率分布为一些迭代次数分配比其他迭代次数更高的概率。例如, 概率分布可以将比任何其他可能的迭代次数更高的概率分配给1, 即, 仅执行初始迭代。这可以确保神经网络在初始迭代时学习生成高质量的输出, 然后可以在后续迭代中改进该输出。这也可以节省训练时间, 因为仅执行初始迭代比执行多次迭代耗时更少。

[0149] 作为另一个示例, 随着训练的进行, 即随着神经网络学会以相对较少的迭代次数做出相对准确的预测, 系统可以调整概率分布以越来越倾向于更大的迭代次数。

[0150] 在一些实现中, 系统将神经网络的训练分布在多个计算设备上, 每个计算设备并行执行过程1300。在这些实现中, 系统可以约束选择, 使得对于过程1300的任何给定迭代, 每个计算设备选择相同的迭代次数。选择相同的迭代次数可以防止选择较小迭代次数的设备在选择较大迭代次数的设备完成处理时处于空闲状态。

[0151] 在一些其他情况下, 如上所述, 系统可以在不同情况下选择不同的迭代次数。在这些情况下, 如上所述, 系统还在训练期间选择不同的迭代次数, 除了概率分布为从1次总迭代到固定迭代次数的每个迭代次数分配相应的概率, 该固定迭代次数被确定为代表训练后可以执行的迭代次数的上限。

[0152] 系统为所选迭代次数中的每一次获得相应网络输入(步骤1304)。特别地, 如上所述, 在一些实现中, 网络输入对于所有迭代都是相同的。在其他实现中, 不同迭代之间的网络输入可以不同。

[0153] 在第一次迭代中, 系统使用结构预测神经网络处理第一次迭代的网络输入, 以生成第一次迭代的结构参数(步骤1306)。

[0154] 如果第一次迭代是最后迭代, 即迭代次数被选择为等于1, 则系统使用第一次迭代的结构参数作为定义蛋白质预测的结构最终结构参数(步骤1312)。

[0155] 如果第一次迭代不是最后迭代, 系统执行步骤1308和1310, 直到生成最后迭代的结构参数。

[0156] 特别地, 该系统根据由结构预测神经网络在先前迭代中生成的输出生成特征(步骤1308)。

[0157] 由于上述停止梯度, 系统然后可以丢弃在先前迭代中生成的所有输出, 即, 而不是将输出存储在存储器中, 因为即使在训练期间执行过程1300, 也不需要这些输出来计算梯度。

[0158] 因此, 因为系统不在存储器中存储任何“内部”迭代的输出, 所以当通过多次迭代而不是在单次迭代中生成最终预测时, 训练过程的存储器消耗不会增加。

[0159] 然后, 系统处理迭代的特征和网络输入, 以生成迭代的结构参数(步骤1310)。通常, 如上所述, 系统根据迭代的特征和网络输入生成组合的输入, 然后使用结构预测神经网络处理组合的输入, 以生成迭代的结构参数。

[0160] 如果迭代不是最后迭代, 则过程1300返回到步骤1308。

[0161] 如果迭代是最后迭代, 则系统使用在最后迭代中生成的结构参数作为最终结构参

数(步骤1312)。

[0162] 如果过程1300在训练期间被执行,则系统使用最终结构参数来确定相对于目标函数的结构预测神经网络的参数的梯度(步骤1314),即上面参考图1B描述的目标函数。具体而言,该系统通过反向传播来计算梯度,即,通过神经网络执行反向传递,然后在倒数第二次迭代的特征处应用停止梯度。

[0163] 因此,如上所述,系统不会将梯度反向传播到先前迭代中。

[0164] 反向通过神经网络(计算目标函数相对于神经网络所有参数的梯度)是训练过程中计算开销最大的部分。由于停止梯度,对于过程1300的每次迭代,系统仍然执行相同数量的反向传递,即,仅单次反向传递,如果系统总是在单次迭代中生成预测,则执行该过程1300的每次迭代。因此,该系统可以实现推断期间附加迭代的好处(就预测精度而言),而不会相应降低训练过程的计算效率。

[0165] 一旦满足了阈值条件,例如已经计算了整批训练输入的梯度,则系统可以根据用于训练的优化器的更新规则,例如Adam、rmsProp或SGD,根据梯度计算更新。然后,系统可以将更新应用于参数的当前值,即,从当前值减去更新或将更新相加到当前值。

[0166] 图2示出了嵌入神经网络200的示例架构,其被配置为处理MSA表示110和对嵌入112,以生成更新的MSA表示114和更新的对嵌入116。

[0167] 嵌入神经网络200包括更新块202-A-N的序列。贯穿本说明书,“块”指神经网络的一部分,例如包括一个或多个神经网络层的神经网络的子网。

[0168] 嵌入神经网络中的每个更新块被配置为接收包括MSA表示和对嵌入的块输入,并处理该块输入以生成包括更新的MSA表示和更新的对嵌入的块输出。

[0169] 嵌入神经网络200向第一更新块(即,在更新块序列中)提供被包括在嵌入神经网络200的网络输入中的MSA表示110和对嵌入112。第一更新块处理MSA表示110和对嵌入112,以生成更新的MSA表示和更新的对嵌入。

[0170] 对于第一更新块之后的每个更新块,嵌入神经网络200向更新块提供由先前更新块生成的MSA表示和对嵌入,并且对于除了最后更新块之外的每个更新块,它向下一个更新块提供由更新块生成的更新的MSA表示和更新的对嵌入。

[0171] 嵌入神经网络200通过使用更新块202-A-N的序列重复更新MSA表示110和对嵌入112来逐渐丰富它们的信息内容。

[0172] 嵌入神经网络200可以提供由最终更新块(即,在更新块序列中)生成的更新的MSA表示114和更新的对嵌入116作为网络输出。

[0173] 图3示出了嵌入神经网络200的更新块300的示例架构,即,如参考图2所描述的。

[0174] 更新块300接收包括当前MSA表示302和当前对嵌入304的块输入,并处理该块输入以生成更新的MSA表示306和更新的对嵌入308。

[0175] 更新块300包括MSA更新块400和对更新块500。

[0176] MSA更新块400使用当前对嵌入304更新当前MSA表示302,并且更新块500使用更新的MSA表示306(即,由MSA更新块400生成的)更新当前对嵌入304。

[0177] 通常,MSA表示和对嵌入可以编码互补信息。例如,MSA表示可以编码关于进化相关的氨基酸链的集合中不同位置的氨基酸标识之间的相关性的信息,而对嵌入可以编码关于蛋白质中氨基酸之间的相互关系的信息。MSA更新块400使用在对嵌入中编码的互补信息来

丰富MSA表示的信息内容,并且对更新块500使用在MSA表示中编码的互补信息来丰富对嵌入的信息内容。作为这种丰富的结果,更新的MSA表示和更新的对嵌入编码与预测蛋白质结构更相关的信息。

[0178] 更新块300在本文被描述为首先使用当前对嵌入304更新当前MSA表示302,然后使用更新的MSA表示306更新当前对嵌入304。该描述不应被理解为将更新块限于以该顺序执行操作,例如,更新块可以首先使用当前MSA表示来更新当前对嵌入,然后使用更新的对嵌入来更新当前MSA表示。

[0179] 更新块300在本文被描述为包括MSA更新块400(即,更新当前MSA表示)和对更新块500(即,更新当前对嵌入)。不应将描述理解为将更新块300限制为仅包含一个MSA更新块或仅包含一个对更新块。例如,更新块300可以包括多个MSA更新块,其在MSA表示被提供给对更新块以用于更新当前对嵌入之前多次更新MSA表示。作为另一个示例,更新块300可以包括使用MSA表示多次更新对嵌入的多个对更新块。

[0180] MSA更新块400和对更新块500可以具有使它们能够执行它们所描述的功能的任何适当的架构。

[0181] 在一些实现中,MSA更新块400、对更新块500或两者都包括一个或多个“自注意力”块。如贯穿本文所使用的,自注意力块通常指更新嵌入的集的神经网络块,即,接收嵌入集合并输出更新的嵌入。为了更新给定嵌入,自注意力块可以确定给定嵌入和一个或多个所选嵌入中的每一个之间的相应“注意力权重”,然后使用(i)注意力权重和(ii)所选嵌入来更新给定嵌入。为了方便起见,自注意力块可以说是使用“超过”所选嵌入的注意力来更新给定嵌入。

[0182] 例如,自注意力块可以接收输入嵌入的集 $\{x_i\}_{i=1}^N$,其中N是蛋白质中氨基酸的数量,并且为了更新嵌入 x_i ,自注意力块可以确定注意力权重 $[a_{i,j}]_{j=1}^N$,其中 $a_{i,j}$ 表示 x_i 和 x_j 之间的注意力权重,如:

$$[0183] \quad [a_{i,j}]_{j=1}^N = \text{softmax}\left(\frac{(W_q x_i) K^T}{c}\right) \quad (5)$$

$$[0184] \quad K^T = [W_k x_j]_{j=1}^N \quad (6)$$

[0185] 其中 W_q 和 W_k 是学习的参数矩阵, $\text{softmax}(\cdot)$ 表示软最大归一化运算,并且c是常数。使用注意力权重,自注意力层可以如下更新嵌入 x_i :

$$[0186] \quad x_i \leftarrow \sum_{j=1..N} a_{i,j} \cdot (W_v x_j) \quad (7)$$

[0187] 其中 W_v 是学习的参数矩阵。(对于输入嵌入 x_i , $W_q x_i$ 可以称为“查询嵌入”,对于输入嵌入 x_i , $W_k x_j$ 可以称为“关键字嵌入”,对于输入嵌入 x_i , $W_v x_j$ 可以称为“值嵌入”)。

[0188] 参数矩阵 W_q (“查询嵌入矩阵”)、 W_k (“关键字嵌入矩阵”)和 W_v (“值嵌入矩阵”)是自注意力块的可训练参数。被包括在MSA更新块400和对更新块500中的任何自注意力块的参数可以理解为更新块300的参数,其可以作为参考图1A-图1C描述的蛋白质结构预测系统100的端到端训练的一部分被训练。通常,对于不同的自注意力块,查询、关键字和值嵌入矩

阵的(训练的)参数是不同的,例如,使得被包括在MSA更新块400中的自注意力块可以具有不同的查询、关键字和值嵌入矩阵,这些矩阵具有与被包括在对更新块500中的自注意力块不同的参数。

[0189] 在一些实现中,MSA更新块400、对更新块500或两者都包括以对嵌入为条件的一个或多个自注意力块,即,实现以对嵌入为条件的自注意力操作。为了对对嵌入进行自注意力操作,自注意力块可以处理对嵌入,以生成与每个注意力权重相对应的相应“注意力偏差”。

例如,除了根据等式(5)-(6)确定注意力权重 $[a_{i,j}]_{j=1}^N$ 之外,自注意力块可以生成对应的注意力偏差的集合 $[b_{i,j}]_{j=1}^N$,其中 $b_{i,j}$ 表示 x_i 和 x_j 之间的注意力偏差。自注意力模块可以通过将学习的参数矩阵应用于对嵌入 $h_{i,j}$,即,对于由(i,j)索引的蛋白质中的氨基酸对,来生成注意力偏差 $b_{i,j}$ 。

[0190] 自注意力块可以例如通过对注意力权重和注意力偏差求和(或以其他方式组合)来确定“偏差的注意力权重”的集合 $[c_{i,j}]_{j=1}^N$,其中 $c_{i,j}$ 表示 x_i 和 x_j 之间的偏差的注意力权重。例如,自注意力块可以确定嵌入 x_i 和 x_j 之间的偏差的注意力权重 $c_{i,j}$:

$$[0191] \quad c_{i,j} = a_{i,j} + b_{i,j}$$

[0192] 其中 $a_{i,j}$ 是介于 x_i 和 x_j 之间的注意力权重, $b_{i,j}$ 是介于 x_i 和 x_j 之间的注意力偏差。自注意力块可以使用偏差的注意力权重来更新每个输入嵌入 x_i ,例如:

$$[0193] \quad x_i \leftarrow \sum_{j=1..N} c_{i,j} \cdot (W_v x_j) \quad (8)$$

[0194] 其中 W_v 是学习的参数矩阵。

[0195] 通常,对嵌入编码表征蛋白质结构和蛋白质结构中氨基酸对之间关系的信息。将对嵌入为条件的自注意力操作应用于输入嵌入的集合,允许输入嵌入以被编码在对嵌入中的蛋白质结构信息所通知的方式被更新。嵌入神经网络的更新块可以使用以对嵌入为条件的自注意力块来更新和丰富MSA表示和对嵌入本身。

[0196] 可选地,自注意力块可以具有多个“头”,每个“头”生成与每个输入嵌入相对应的相应更新的嵌入,即,使得每个输入嵌入与多个更新的嵌入相关联。例如,每个头可以根据参考等式(5)-(7)描述的参数矩阵 W_q 、 W_k 和 W_v 的不同值来生成更新的嵌入。具有多个头的自注意力块可以实现“门控”操作,以组合由头为输入嵌入生成的更新的嵌入,即,生成与每个输入嵌入相对应的单个更新的嵌入。例如,自注意力块可以使用一个或多个神经网络层(例如,完全连接的神经网络层)来处理输入嵌入,以生成每个头的相应门控值。自注意力块然后可以根据门控值组合与输入嵌入相对应的更新的嵌入。例如,自注意力块可以为输入嵌入 x_i 生成更新的嵌入,如下所示:

$$[0197] \quad \sum_{k=1}^K \alpha_k \cdot x_i^{next} \quad (9)$$

[0198] 其中,k索引头, α_k 是头k的门控值, x_i^{next} 是头k为输入嵌入 x_i 生成的更新的嵌入。

[0199] 参考图4描述了使用以对嵌入为条件的自注意力块的MSA更新块400的示例架构。参考图4描述的示例MSA更新块通过使用以当前对嵌入为条件的自注意力块处理当前MSA表示的行,基于当前对嵌入更新当前MSA表示。

[0200] 参考图5描述了使用以对嵌入为条件的自注意力块的对更新块500的示例架构。参考图5描述的示例对更新块通过计算更新的MSA表示的外积均值,将外积均值的结果加到当前对嵌入,并使用以当前对嵌入为条件的自注意力块处理当前对嵌入,基于更新的MSA表示更新当前对嵌入。

[0201] 图4示出了MSA更新块400的示例架构。MSA更新块400被配置为接收当前MSA表示302,以(至少部分地)基于当前对嵌入来更新当前MSA表示306。

[0202] 为了更新当前MSA表示302,MSA更新块400使用以当前对嵌入为条件的自注意力操作(即“逐行”自注意力操作)来更新当前MSA表示的每一行中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每一行中的嵌入提供给以当前对嵌入为条件的“逐行”自注意力块402,例如,如参考图3所述,以便为当前MSA表示302的每一行生成更新的嵌入。可选地,MSA更新块可以将逐行自注意力块402的输入相加到逐行自注意力块402的输出。根据当前对嵌入来调节逐行自注意力块402使得MSA更新块400能够使用来自当前对嵌入的信息来丰富当前MSA表示302。

[0203] MSA更新块然后使用不以当前对嵌入为条件的自注意力操作(即“逐列”自注意力操作)来更新当前MSA表示的每一列中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每一列中的嵌入提供给不以当前对嵌入为条件的“逐列”自注意力(self-attention)块404,以生成当前MSA表示302的每一列的更新的嵌入。作为不以当前对嵌入为条件的结果,逐列自注意力块404使用注意力权重(例如,如参考等式(5)-(6)所描述的)而不是偏差的注意力权重(例如,如参考等式(8)所描述的)来为当前MSA表示的每一列生成更新的嵌入。可选地,MSA更新块可以将对逐列自注意力块404的输入相加到逐列自注意力块404的输出。

[0204] MSA更新块然后使用转换块处理当前MSA表示302,例如,该转换块将一个或多个全连接神经网络层应用于当前MSA表示302。可选地,MSA更新块400可以将转换块406的输入相加到转换块406的输出。

[0205] MSA更新块可以输出由逐行自注意力块402、按列自注意力块404和转换块406执行的操作生成的更新的MSA表示306。

[0206] 图5示出了对更新块500的示例架构。对更新块500被配置为接收当前对嵌入304,并且基于(至少部分地)更新的MSA表示306来更新当前对嵌入304。

[0207] 为了更新当前对嵌入304,对更新块500将外积平均值运算502应用于更新的MSA表示306,并将外积平均值运算502的结果相加到当前对嵌入304。

[0208] 外积均值运算定义一系列运算,当应用于表示为 $M \times N$ 嵌入阵列的MSA表示时,生成 $N \times N$ 嵌入阵列,即其中 N 是蛋白质中氨基酸的数量。当前对嵌入304也可以表示为 $N \times N$ 嵌入阵列,并且将外积平均值502的结果加到当前对嵌入304指的是对两个 $N \times N$ 嵌入阵列求和。

[0209] 为了计算外积均值,对更新块生成张量 $A(\cdot)$,例如,由下式给出:

$$[0210] \quad A(\text{res1}, \text{res2}, \text{ch1}, \text{ch2}) = \frac{1}{|\text{rows}|} \sum_{\text{rows}} \text{LeftAct}(\text{row}, \text{res1}, \text{ch1}) \cdot \text{RightAct}(\text{row}, \text{res2}, \text{ch2})$$

[0211] 其中 $\text{res1}, \text{res2} \in \{1, \dots, N\}$, $\text{ch1}, \text{ch2} \in \{1, \dots, C\}$ 、其中 C 是MSA表示的每次嵌入中

的通道数, $|\text{rows}|$ 是MSA表示中的行数, $\text{LeftAct}(\text{row}, \text{res1}, \text{ch1})$ 是应用于位于由“row”索引的行和由“res1”索引的列的MSA表示的嵌入的通道ch1的线性运算(例如, 由矩阵乘法定义), 并且 $\text{RightAct}(\text{row}, \text{res2}, \text{ch2})$ 是应用于位于由“row”索引的行和由“res2”索引的列的MSA表示的嵌入的通道ch2的线性运算(例如, 由矩阵乘法定义)。外积平均值的结果是通过展平和线性投影张量A的 $(\text{ch1}, \text{ch2})$ 维度生成的。可选地, 作为计算外积均值的一部分, 对更新块可以执行一个或多个层归一化运算(例如, 如参考吉米巴雷等人的“层归一化”, arXiv: 1607.06450所描述的)。

[0212] 通常, 更新的MSA表示306编码关于在进化相关的氨基酸链的集合中不同位置的氨基酸标识之间的相关性的信息。在更新的MSA表示306中编码的信息与预测蛋白质的结构相关, 并且通过将在更新的MSA表示中编码的信息结合到当前的对嵌入中(即, 通过外积均值502), 对更新块500可以增强当前的对嵌入的信息内容。

[0213] 在使用更新的MSA表示(即, 通过外积均值502)更新当前对嵌入304之后, 对更新块500使用以当前对嵌入为条件的自注意力操作(即, “逐行”自注意力操作)将当前对嵌入的排列的每一行中的当前对嵌入更新到 $N \times N$ 阵列中。更具体地, 对更新块500将当前对嵌入的每一行提供给“逐行”自注意力块504, 该自注意力块504也以当前对嵌入为条件, 例如, 如参考图3所述, 以便为每一行生成更新的对嵌入。可选地, 对更新块可以将对逐行自注意力块504的输入相加到逐行自注意力块504的输出。

[0214] 然后, 对更新块500使用同样以当前对嵌入为条件的自注意力操作(即“逐列”自注意力操作)来更新当前对嵌入 $N \times N$ 阵列的每一列中的当前对嵌入。更具体地, 对更新块500将当前对嵌入的每一列提供给“逐列”自注意力块506, 该自注意力块506也以当前对嵌入为条件来为每一列生成更新的对嵌入。可选地, 对更新块可以将对逐列自注意力块506的输入相加到逐列自注意力块506的输出。

[0215] 然后, 对更新块500使用转换块处理当前对嵌入, 例如, 将一个或多个全连接神经网络层应用于当前对嵌入。可选地, 对更新块500可以将转换块508的输入相加到转换块508的输出。

[0216] 对更新块可以输出由逐行自注意力块504、按列自注意力块506和转换块508执行的操作生成的更新的对嵌入308。

[0217] 图6示出了折叠神经网络600的示例架构, 其生成了定义预测的蛋白质结构108的结构参数的集合106。例如, 在实现中, 折叠神经网络600通过处理包括蛋白质中每对氨基酸的相应对嵌入312的输入, 基于对嵌入确定蛋白质的预测的结构, 以生成结构参数314的值。折叠神经网络600可以被包括在参照图1A-图1C描述的蛋白质结构预测系统100中。

[0218] 折叠神经网络600为蛋白质中的每个氨基酸生成结构参数, 这些参数可以包括: (i) 位置参数, 和(ii) 旋转参数。如前所述, 氨基酸的位置参数可以指定蛋白质结构中氨基酸中特定的原子的预测的3-D空间位置。氨基酸的旋转参数可以指定氨基酸在蛋白质结构中的预测的“取向”。更具体地, 旋转参数可以指定3-D空间旋转操作, 如果将其应用于位置参数的坐标系, 则导致氨基酸中的三个“主链”原子相对于旋转坐标系呈现固定位置。

[0219] 在实现中, 折叠神经网络600接收来自最终MSA表示、最终对嵌入116或两者得到的输入, 并生成定义蛋白质预测的结构的结构参数106的最终值。例如, 折叠神经网络600可以接收输入, 该输入包括: (i) 蛋白质中每对氨基酸的各自的对嵌入116, (ii) 蛋白质中每个氨

基酸的“单个”嵌入602的初始值,以及(iii)蛋白质中每个氨基酸的结构参数604的初始值。折叠神经网络600处理输入以生成结构参数106的最终值,这些参数共同表征蛋白质的预测的结构108。

[0220] 蛋白质结构预测系统100可以向折叠神经网络600提供作为嵌入神经网络的输出生成的对嵌入,如参考图1A-图1C所述。

[0221] 蛋白质结构预测系统100可以为来自MSA表示114的氨基酸生成初始单个嵌入602,即,如参考图1A-图1C所描述的,生成为嵌入神经网络的输出。例如,如上所述,MSA表示114可以被表示为嵌入的2-D阵列,其具有的列数等于蛋白质中的氨基酸数,其中每列与蛋白质中的相应氨基酸相关联。蛋白质结构预测系统100可以通过对来自与氨基酸相关联的MSA表示114的列的嵌入进行求和(或者组合)来为蛋白质中的每个氨基酸生成初始单个嵌入。作为另一个示例,蛋白质结构预测系统100可以通过从与其结构被估计的蛋白质的氨基酸序列相对应的MSA表示114的行中提取嵌入来生成蛋白质中氨基酸的初始单个嵌入。

[0222] 蛋白质结构预测系统100可以生成具有默认值的初始结构参数604,例如,其中每个氨基酸的位置参数被初始化为原点(例如,在笛卡尔坐标系中 $[0,0,0]$),并且每个氨基酸的旋转参数被初始化为 3×3 单位矩阵。

[0223] 折叠神经网络600可以通过重复更新单个嵌入606和结构参数608的当前值,即从它们的初始值开始,来生成最终结构参数106。更具体地,折叠神经网络600包括更新神经网络块610的序列,其中每个更新块610被配置为更新当前单个嵌入606(即,生成更新的单个嵌入612)和更新当前结构参数608(即,生成更新的结构参数614)。除了更新块之外,折叠神经网络600可包括其它神经网络层或块,例如,其可与更新块交错。

[0224] 每个更新块610可以包括:(i)几何注意力块616,以及(ii)折叠块618,接下来将更详细地描述它们中的每一个。

[0225] 几何注意力块616使用“几何”自注意力操作更新当前的单个嵌入,该操作显式地推理蛋白质结构中氨基酸的3-D几何,即由结构参数定义的。更具体地,为了更新给定的单个嵌入,几何注意力块616确定给定的单个嵌入和一个或多个所选的单个嵌入中的每一个之间的相应注意力权重,其中注意力权重取决于当前的单个嵌入、当前的结构参数和对嵌入。几何注意力块616然后使用(i)注意力权重,(ii)所选的单个嵌入,以及(iii)当前结构参数来更新给定的单个嵌入。

[0226] 为了确定注意力权重,几何注意力块616处理每个当前的单个嵌入,以生成对应的“符号查询”嵌入、“符号关键字”嵌入和“符号值”嵌入。例如,几何注意力块616可以为与第*i*个氨基酸相对应的单个嵌入 h_i 生成符号查询嵌入 q_i 、符号关键字嵌入 k_i 和符号值嵌入 v_i ,如下所示:

$$[0227] \quad q_i = \text{Linear}(h_i) \quad (10)$$

$$[0228] \quad k_i = \text{Linear}(h_i) \quad (11)$$

$$[0229] \quad v_i = \text{Linear}(h_i) \quad (12)$$

[0230] 其中 $\text{Linear}(\cdot)$ 是指具有独立学习的参数值的线性层。

[0231] 几何注意力块616另外处理每个当前的单个嵌入,以生成相应的“几何查询”嵌入、“几何关键字”嵌入和“几何值”嵌入。每个单个嵌入的几何查询、几何关键字和几何值嵌入都是最初在相应氨基酸的局部参考系中生成的3-D点,然后使用氨基酸的结构参数将其旋

转和转换到全局参考系。例如,几何注意力块616可以为与第*i*个氨基酸相对应的单个嵌入 h_i 生成几何查询嵌入 q_i^p 、几何关键字嵌入 k_i^p 和几何值嵌入 v_i^p ,如下所示:

$$[0232] \quad q_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (13)$$

$$[0233] \quad k_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (14)$$

$$[0234] \quad v_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (15)$$

[0235] 其中 $\text{Linear}_p(\cdot)$ 表示具有投影 h_i 到3-D点的独立学习的参数值的线性层(上标表示该量是3-D点), R_i 表示由第*i*个氨基酸的旋转参数指定的旋转矩阵,并且 t_i 表示第*i*个氨基酸的位置参数。

[0236] 为了更新与氨基酸*i*相对应的单个嵌入 h_i ,几何注意力块616可以生成注意力权重 $[a_j]_{j=1}^N$,其中*N*是蛋白质中氨基酸的总数,并且 a_j 是氨基酸*i*和氨基酸*j*之间的注意力权重,如下:

$$[0237] \quad [a_j]_{j=1}^N = \text{softmax} \left(\left[\frac{q_i \cdot k_j}{\sqrt{m}} + \alpha |q_i^p - k_j^p|_2^2 + (b_{i,j} \cdot w) \right]_{j=1}^N \right) \quad (16)$$

[0238] 其中 q_i 表示用于氨基酸*i*的符号查询嵌入, k_j 表示用于氨基酸*j*的符号关键字嵌入, m 表示 q_i 和 k_j 的维度, α 表示学习的参数, q_i^p 表示用于氨基酸*i*的几何查询嵌入, k_j^p 表示用于氨基酸*j*的几何关键字嵌入, $|\cdot|_2$ 是 L_2 范数,并且 $b_{i,j}$ 是与包括氨基酸*i*和氨基酸*j*的氨基酸对相对应的对嵌入116,并且 w 是学习的权重向量(或一些其他学习的投影运算)。

[0239] 通常,氨基酸对的对嵌入隐式地编码了与该对中氨基酸之间的关系相关的信息,例如,该对中氨基酸之间的距离。通过部分基于氨基酸*i*和*j*的对嵌入来确定氨基酸*i*和氨基酸*j*之间的注意力权重,折叠神经网络600用来自对嵌入的信息来丰富注意力权重,从而提高预测的折叠结构的准确性。

[0240] 在一些实现中,几何注意力块616生成多组几何查询嵌入、几何关键字嵌入和几何值嵌入,并使用生成的几何嵌入的集合来确定注意力权重。

[0241] 在生成与氨基酸*i*相对应的单个嵌入 h_i 的注意力权重之后,几何注意力块616使用注意力权重来更新单个嵌入 h_i 。具体地,几何注意力块616使用注意力权重来生成“符号返回”嵌入和“几何返回”嵌入,然后使用符号返回嵌入和几何返回嵌入来更新单个嵌入。几何注意力块124可以为氨基酸*i*生成符号返回嵌入 o_i ,例如:

$$[0242] \quad o_i = \sum_j a_j v_j \quad (17)$$

[0243] 其中 $[a_j]_{j=1}^N$ 表示注意力权重(例如,参照等式(16)定义),并且每个 v_j 表示氨基酸*j*的符号值嵌入。几何注意力块616可以为氨基酸*i*生成几何返回嵌入 o_i^p ,例如,如下:

$$[0244] \quad o_i^p = R_i^{-1} \cdot \left(\sum_j a_j v_j^p - t_i \right) \quad (18)$$

[0245] 在几何返回嵌入 o_i^p 是3-D点的情况下, $[a_j]_{j=1}^N$ 表示注意力权重(例如,参照等式(16)定义), R_i^{-1} 是由氨基酸i的旋转参数指定的旋转矩阵的逆,并且 t_i 是氨基酸i的位置参数。可以理解,几何返回嵌入最初在全局参考系中生成,然后旋转并平移到相应氨基酸的局部参考系。

[0246] 几何注意力块616可以使用对应的符号返回嵌入 o_i (例如,根据等式(17)生成的)和几何返回嵌入 o_i^p (例如,根据等式(18)生成的)来更新氨基酸i的单个嵌入 h_i ,例如:

$$[0247] \quad h_i^{next} = \text{LayerNorm} \left(h_i + \text{Linear}(o_i, o_i^p, |o_i^p|) \right) \quad (19)$$

[0248] 其中 h_i^{next} 是用于氨基酸i的更新的单个嵌入, $|\cdot|$ 是范数,例如 L_2 范数,并且 $\text{LayerNorm}(\cdot)$ 表示层归一化运算,例如,如参考J.L.Ba, J.R.Kiros, G.E.Hinton, “层归一化”, arXiv:1607.06450 (2016)所描述的。

[0249] 使用具体的3-D几何嵌入来更新氨基酸的单个嵌入606,例如,如参考等式(13)-(15)所描述的,使得几何注意力块616能够在更新单个嵌入时推理3-D几何。此外,每个更新块以对整个蛋白质结构的旋转和平移不变的方式更新单个嵌入和结构参数。例如,将相同的全局旋转和平移操作应用于提供给折叠神经网络600的初始结构参数将导致折叠神经网络600生成预测的结构,该预测的结构以相同的方式被全局旋转和平移,但在其他方面是相同的。因此,应用于初始结构参数的全局旋转和平移操作不会影响由折叠神经网络600从初始结构参数开始生成的预测蛋白质结构的准确性。折叠神经网络600生成的表示的旋转和平移不变性有助于训练,例如,因为折叠神经网络600自动学习概括蛋白质结构的所有旋转和平移。

[0250] 在被提供给折叠块618之前,用于氨基酸的更新的单个嵌入可以由几何注意力块616中的一个或多个附加神经网络层,例如线性神经网络层进一步转换。

[0251] 在几何注意力块616更新氨基酸的当前单个嵌入606之后,折叠块618使用更新的单个嵌入612更新当前结构参数608。例如,折叠块618可以将氨基酸i的当前位置参数 t_i 更新为:

$$[0252] \quad t_i^{next} = t_i + \text{Linear}(h_i^{next}) \quad (20)$$

[0253] 其中 t_i^{next} 是更新的位置参数, $\text{Linear}(\cdot)$ 表示线性神经网络层,并且 h_i^{next} 表示氨基酸i的更新的单个嵌入。在另一个示例中,氨基酸i的旋转参数 R_i 可以指定旋转矩阵,并且折叠块618可以将当前旋转参数 R_i 更新为:

$$[0254] \quad w_i = \text{Linear}(h_i^{next}) \quad (21)$$

$$[0255] \quad R_i^{next} = R_i \cdot \text{QuaternionToRotation}(1 + w_i) \quad (22)$$

[0256] 其中 w_i 是3-D向量, $\text{Linear}(\cdot)$ 是线性神经网络层, h_i^{next} 是针对氨基酸i的更新的

单个嵌入, $1+w_i$ 表示具有实部1和虚部 w_i 的四元数, 并且 $\text{QuaternionToRotation}(\cdot)$ 表示将四元数转换成等效 3×3 旋转矩阵的运算。使用等式(21) - (22)更新旋转参数确保了更新的旋转参数定义有效的旋转矩阵, 例如行列式为1的标准正交矩阵。

[0257] 折叠神经网络600可以提供由最终更新模块610生成的更新的结构参数, 作为定义预测的蛋白质结构108的最终结构参数106。折叠神经网络600可包含任何适当数目的更新块, 例如, 5个更新块、25个更新块或125个更新块。可选地, 折叠神经网络的每个更新块可以共享在折叠神经网络的训练期间联合更新的单个参数值的集合。在更新块610之间共享参数值减少了折叠神经网络的可训练参数的数量, 并且因此可以例如通过稳定训练和减少过拟合的可能性来促进折叠神经网络的有效训练。

[0258] 在训练期间, 训练引擎可以基于评估最终结构参数106的准确性的结构损失来训练结构预测系统的参数, 包括折叠神经网络600的参数, 如上所述。在一些实现中, 训练引擎可以进一步评估在最终更新块之前的一个或多个更新块610的辅助结构损失(即, 生成最终结构参数)。更新块的辅助结构损失评估由更新块生成的更新结构参数的准确性。

[0259] 可选地, 在训练期间, 训练引擎可以应用“停止梯度”操作, 以防止梯度通过每个更新块的某些神经网络参数反向传播, 例如, 用于计算更新的旋转参数的神经网络参数(如等式(21) - (22)中所述)。应用这些停止梯度操作可以提高在训练期间计算的梯度的数值稳定性。

[0260] 通常, 由折叠神经网络600生成的预测的蛋白质结构108和对应的基础真实蛋白质结构之间的相似性可以被度量, 例如, 通过相似性度量, 该相似性度量为预测的蛋白质结构中的多个原子中的每一个分配相应的准确度分数。例如, 相似性度量可以为预测的蛋白质结构中的每个碳 α 原子分配各自的准确度分数。预测蛋白质结构中原子的准确度分数可以表征预测蛋白质结构中原子的位置与真实蛋白质结构中原子的实际位置有多接近。可以将预测的蛋白质结构与基础真实蛋白质结构进行比较以生成预测的蛋白质结构中的原子的准确度分数的相似性度量的示例是1DDT相似性度量, 参考: V. Mariani等人, “1DDT: 使用距离差测试比较蛋白质结构和模型的局部无叠加分数”, *Bioinformatics*, 2013年11月1日; 29(21) 2722-2728。

[0261] 折叠神经网络600可以被配置为对预测的蛋白质结构108中的一个或多个原子中的每一个生成相应的置信度估计650。预测的蛋白质结构中的原子的置信度估计650表征了预测的蛋白质结构中的原子的预测的准确度分数(例如, 1DDT准确度分数), 即, 将预测的蛋白质结构与(潜在未知的)基础真实蛋白质结构进行比较的相似性度量所生成的分数。在一个示例中, 预测的蛋白质结构中的原子的置信度估计650可以定义区间集合上的离散概率分布, 该组区间形成该原子的准确度分数的可能值范围的划分。离散概率分布可以将相应的概率与每个区间相关联, 该概率定义了实际准确度分数被包括在该区间中的可能性。例如, 准确度分数的可能值的范围可以是 $[0, 100]$, 并且置信度估计650可以定义区间的集合 $\{[0, 2), [2, 4), \dots, [98, 100]\}$ 上的概率分布。在另一个示例中, 预测的蛋白质结构中的原子的置信度估计650可以是数值, 即, 直接预测该原子的准确度分数。

[0262] 在一些实现中, 折叠神经网络600为蛋白质的每个氨基酸中的特定的原子(例如, α 碳原子)生成相应的置信度估计650。折叠神经网络600可以为蛋白质中的氨基酸中的特定的原子生成置信度估计650, 例如, 通过使用一个或多个神经网络层(例如, 全连接层)处理

由折叠神经网络中的最后更新块生成的氨基酸的更新的单个嵌入。

[0263] 结构预测系统可以基于预测的蛋白质结构中的原子的置信度估计650,生成与蛋白质中每个氨基酸相对应的相应置信度分数。例如,结构预测系统可以生成氨基酸的置信度分数,作为氨基酸中 α 碳原子的准确度分数的可能值的概率分布的期望值。

[0264] 结构预测系统可以生成整个预测的结构置信度分数,例如,作为蛋白质中氨基酸的置信度分数的平均值。

[0265] 在结构预测系统的训练期间,训练引擎可以通过反向传播辅助损失的梯度来调整结构预测系统的参数值,该辅助损失度量(i)由折叠神经网络600生成的置信度估计和(ii)通过比较预测的蛋白质结构和基础真实蛋白质结构生成的准确度分数之间的误差。该误差可以是例如交叉熵误差。

[0266] 由结构预测系统生成的置信度估计可以以多种方式使用。例如,对预测的蛋白质结构中原子的置信度估计可以指示该结构的哪些部分已经被可靠地估计,因此适合于进一步的下流处理或分析。作为另一个示例,每个蛋白质的置信度分数可以用于对蛋白质结构的预测的集合进行排序,例如,由相同的结构预测系统通过处理表征相同蛋白质的不同输入而生成的预测,或者由不同的结构预测系统生成的预测。

[0267] 由结构参数106指定的位置和旋转参数可以定义蛋白质的氨基酸中主链原子的空间位置(例如,在笛卡尔坐标中 $[x, y, z]$)。然而,结构参数106不一定限定蛋白质氨基酸中剩余原子的空间位置,例如氨基酸侧链中的原子。具体而言,氨基酸中剩余原子的空间位置取决于氨基酸中键之间的扭转角的值,例如 ω 角(omega-angle)、 ϕ 角(phi-angle)、 ψ 角(psi-angle)、 χ_1 角(chi1-angle)、 χ_2 角(chi2-angle)、 χ_3 角(chi3-angle)和 χ_4 角(chi4 angle),如参考图7所示。

[0268] 可选地,折叠神经网络600的一个或多个更新块610可以生成定义蛋白质的每个氨基酸中每个原子的相应预测的空间位置的输出。为了生成氨基酸中原子的预测的空间位置,更新块可以使用一个或多个神经网络层来处理氨基酸的更新单个嵌入,以生成氨基酸中原子之间的键的扭转角的预测值。神经网络层可以是例如嵌入有剩余连接的全连接神经网络层。每个扭转角可以表示为例如2-D向量。

[0269] 更新块可以基于:(i)氨基酸的扭转角值,和(ii)氨基酸的更新的结构参数(例如,位置和旋转参数)来确定氨基酸中原子的空间位置。例如,更新块可以根据预定义函数处理扭转角,以在氨基酸的局部参考系中生成氨基酸中原子的空间位置。通过根据氨基酸的更新结构参数旋转和平移原子的空间位置,更新块可以在全局参考系(即,蛋白质中所有氨基酸共有的参考系)中生成氨基酸中原子的空间位置。例如,更新块可以通过将由更新的旋转参数定义的旋转操作应用于原子在局部参考系中的空间位置以生成旋转的空间位置,然后将由更新的位置参数定义的平移操作应用于旋转的空间位置,来确定原子在全局参考系中的空间位置。

[0270] 在一些实现中,替代输出最终结构参数或者与输出最终结构参数相结合,折叠神经网络600输出由最终更新块生成的蛋白质的氨基酸中的原子的预测的空间位置。

[0271] 参考图6描述的折叠神经网络600在本文被表征为接收基于MSA表示114和由嵌入神经网络生成的对嵌入116的输入,例如,如参考图2所描述的。然而,一般来说,折叠神经网络的输入(例如,单个嵌入602和对嵌入116)可以使用任何适当的技术来生成。此外,由折叠

神经网络执行的操作的各个方面(例如,预测蛋白质的每个氨基酸中的原子的空间位置)可以由其他折叠神经网络执行,例如,具有接收不同输入的不同架构。

[0272] 图7说明了氨基酸中键之间的扭转角,例如 ω 角、 ϕ 角、 ψ 角、 χ_1 角、 χ_2 角、 χ_3 角、 χ_4 角和 χ_5 角。

[0273] 图8是未折叠的蛋白质和折叠的蛋白质的示意图。未折叠的蛋白质是氨基酸的随机卷曲。未折叠的蛋白质进行蛋白质折叠并折叠成3D构型。蛋白质结构通常包括稳定的局部折叠模式,如 α 螺旋(如802所示)和 β 折叠。

[0274] 图9是用于预测包括一条或多条链的蛋白质结构的另一个示例过程900的流程图,其中每条链指定一个氨基酸序列。为了方便起见,过程900将被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来执行。例如,根据本说明书适当编程的蛋白质结构预测系统,例如图1A的蛋白质结构预测系统100,可以执行过程900。

[0275] 系统获得代表与蛋白质中每条链相对应的相应多序列比对(MSA)的初始MSA表示(902)。

[0276] 系统获得蛋白质中的每对氨基酸的相应初始对嵌入(904)。

[0277] 该系统使用嵌入神经网络处理包括初始MSA表示和初始对嵌入的输入,以生成包括最终MSA表示和蛋白质中每对氨基酸的相应最终对嵌入的输出。

[0278] 嵌入神经网络包括一系列更新块。每个更新块具有相应更新块参数的集合,并被配置为接收当前MSA表示和蛋白质中每对氨基酸的相应当前对嵌入。每个更新块:(i)基于当前对嵌入更新当前MSA表示,以及(ii)基于更新的MSA表示更新当前对嵌入(906)。

[0279] 系统基于使用最终MSA表示、最终对嵌入或两者来确定蛋白质的预测的结构(908)。

[0280] 图10示出了用于生成蛋白质中氨基酸链的MSA表示1010的示例过程1000。参考图1A描述的蛋白质结构预测系统100可以实现过程1000的操作。

[0281] 为了生成蛋白质中氨基酸链的MSA表示1010,系统100获得蛋白质的MSA 1002,其可以包括例如数千个MSA序列。

[0282] 系统100将MSA序列的集合分成“核心”MSA序列的集合1004和“额外”MSA序列的集合1006。核心MSA序列的集合可以比额外MSA序列的集合1006小(例如,小一个数量级)。系统100可以将MSA序列的集合划分为核心MSA序列1004和额外MSA序列1006,例如,通过随机选择预定数量的MSA序列作为核心MSA序列,并将剩余的MSA序列标识为额外MSA序列1006。

[0283] 对于每个额外MSA序列1006,系统100可以确定额外MSA序列和每个核心MSA序列1004之间的相应相似性度量(例如,基于汉明距离)。系统100然后将每个额外MSA序列1006与额外MSA序列1006最相似的对核心MSA序列1004相关联(即,根据相似性度量)。与核心MSA序列1004相关联的额外MSA序列的集合1006可以被称为“MSA序列簇(MSA sequence cluster)”1008。也就是说,系统100确定与每个核心MSA序列1004相对应的相应MSA序列簇1008,其中与核心MSA序列1004相对应的MSA序列簇1008包括与核心MSA序列1004最相似的额外MSA序列的集合1006。

[0284] 系统100可以基于核心MSA序列和MSA序列簇1008生成蛋白质中氨基酸链的MSA表示。MSA表示1010可以由嵌入 $M \times N$ 阵列来表示,其中 M 是核心MSA序列的数量(即,使得每个核心MSA序列与MSA表示的相应行相关联),并且 N 是氨基酸链中氨基酸的数量。MSA表示中的嵌

入可以由 $(i, j) \in \{(i, j) : i=1, \dots, M, j=1, \dots, N\}$ 索引。

[0285] 为了生成MSA表示1010中位置 (i, j) 处的嵌入, 系统100可以获得定义核心MSA序列 i 中位置 j 处的氨基酸标识的嵌入(例如, 独热嵌入)。系统100还可以基于与核心MSA序列 i 相对应的MSA序列簇1008中的额外MSA序列1006中的位置 j 处的每个可能氨基酸的相对出现频率来确定可能氨基酸集合上的概率分布。系统100然后可以通过组合(例如, 连接)来确定MSA表示中位置 (i, j) 处的嵌入: (i) 定义核心MSA序列 i 中位置 j 处的氨基酸标识的嵌入, 以及 (ii) 与核心MSA序列 i 中位置 j 相对应的可能氨基酸的概率分布。

[0286] 在某些情况下, 一个或多个核心MSA序列的(基本事实)蛋白质结构可能是已知的。具体而言, 对于一个或多个核心MSA序列, 核心MSA序列中氨基酸中的键之间的扭转角的值(例如, ω 角、 ϕ 角、 ψ 角等)可能是已知的。如果已知核心MSA序列 i 中氨基酸的扭转角值, 则系统100可以至少部分基于核心MSA序列 i 中氨基酸 j 的扭转角值, 在MSA表示中的位置 (i, j) 生成嵌入。例如, 系统可以使用一个或多个神经网络层生成扭转角值的嵌入, 并将扭转角值的嵌入连接到MSA表示中的位置 (i, j) 处的嵌入。

[0287] 图11示出了用于为蛋白质中的每对氨基酸生成(初始化)相应的对嵌入112的示例过程1100。参考图1A描述的蛋白质结构预测系统100可以实现过程1100的操作。

[0288] 系统100可以使用蛋白质的MSA表示1102来生成对嵌入112。参考图1A-图1C和图10更详细地描述了生成蛋白质的MSA表示。系统100可以基于蛋白质中每个氨基酸链的相应MSA表示来生成蛋白质的MSA表示1102。为了生成蛋白质中氨基酸链的MSA表示, 系统100可以使用比用于生成参照图1A-图1C描述的MSA表示110更多的MSA序列(例如, 数量级)。因此, 系统100用于生成对嵌入112的MSA表示1102可以具有比参照图1A-图1C描述的MSA表示110更多的行(例如, 数量级)。在一些实现中, 系统100可以使用参考图10描述的额外MSA序列1006来生成MSA表示1102。

[0289] 在生成MSA表示1102之后, 系统100处理MSA表示1102以从MSA表示1102生成对嵌入1104, 例如, 通过对MSA表示1102应用外积平均值运算, 并将对嵌入1104标识为外积平均值运算的结果。

[0290] 系统100使用嵌入神经网络1106处理MSA表示1102和对嵌入1104。嵌入神经网络1106可以通过在MSA表示1102和对嵌入1104之间共享信息来更新MSA表示1102和对嵌入1104。更具体地, 嵌入神经网络1106可以在基于对嵌入1104更新MSA表示1102和基于MSA表示1102更新对嵌入1104之间交替。

[0291] 嵌入神经网络1106可以具有基于参考图2-图5描述的嵌入神经网络架构的架构, 即, 使用逐行和逐列自注意力块更新对嵌入1104和MSA表示1102。

[0292] 在一些实现中, 嵌入神经网络1106可以使用逐列“全局”自注意力操作来更新MSA表示1102的每一列中的嵌入。更具体地, 嵌入神经网络1106可以向逐列全局自注意力块提供MSA表示的每一列中的嵌入, 以生成当前MSA表示的每一列的更新嵌入。为了实现全局的逐列自注意力, 自注意力块可以为列中的每个嵌入生成相应查询嵌入, 然后对查询嵌入进行平均, 以生成该列的单个“全局”查询嵌入。然后, 逐列自注意力块使用单个全局查询嵌入来执行自注意力操作, 这可以将自注意力操作的复杂度从二次(即, 每列嵌入的数量)降低到线性。使用全局自注意力操作可以降低逐列自注意力操作的计算复杂度, 以使得逐列自注意力操作能够在包括大量(例如, 数千个)嵌入的MSA表示1102的列上执行。

[0293] 在使用嵌入神经网络1106更新对嵌入1104和MSA表示1102之后,系统100可以将对嵌入112标识为由嵌入神经网络1106生成的更新的对嵌入。系统100可以丢弃由嵌入神经网络1106生成的更新的MSA表示,或者以任何适当的方式使用它。

[0294] 作为生成对嵌入112的一部分,系统100可以在与蛋白质中每对氨基酸相对应的相对对嵌入中包括相对位置编码信息。该系统可以通过以下方式将被包括在相同氨基酸链中的氨基酸对的相对位置编码信息包括在对应的对嵌入中:计算表示在氨基酸链中分隔氨基酸对的氨基酸数量的有符号差,将结果裁剪到预定的区间,使用独热编码向量表示裁剪的值,对独热编码向量应用线性变换,并将线性变换的结果相加到对应的对嵌入中。该系统可以通过将默认编码向量相加到对应的对嵌入中来将未被包括在相同氨基酸链中的氨基酸对的相对位置编码信息包括在对应的对嵌入中,该默认编码向量指示氨基酸对未被包括在相同氨基酸链中。

[0295] 系统100还可以至少部分地基于一个或多个模板序列的集合1110来生成对嵌入112。每个模板序列1110是蛋白质中氨基酸链的MSA序列,其中模板序列1110的折叠的结构是已知的,例如通过物理实验。

[0296] 系统100可以生成每个模板序列1110的相应模板表示1112。模板序列1110的模板表示1112包括与模板序列中每对氨基酸相对应的相应嵌入,例如,使得长度 n (即,具有 n 氨基酸)的模板序列的模板表示1112可以表示为嵌入的 $n \times n$ 阵列。系统100生成模板序列1110的模板表示1112中位置 (i, j) 处的嵌入,基于例如:(i)代表模板序列中位置 i 和位置 j 处的氨基酸标识的相应嵌入(例如,独热嵌入),(ii)由模板序列中位置 i 和位置 j 处,即模板序列的折叠的结构中的氨基酸中相应碳 α 原子的空间位置差定义的单位向量,其中单位向量是在氨基酸 i 或氨基酸 j 的参照系中计算的,和(iii)模板序列中位置 i 和位置 j 处的氨基酸中相应碳 α 原子的空间位置之间距离的离散化/分格表示。

[0297] 系统100可以使用一个或多个模板更新块1114的序列来处理每个模板表示,以生成与每个模板序列1110相对应的相应更新的模板表示1116。模板更新块可以包括例如逐行自注意力块(例如,更新模板表示的每一行中的嵌入)、逐列自注意力块(例如,更新模板表示的每一列中的嵌入)、以及转换块(例如,将一个或多个神经网络层应用于模板表示中的每个嵌入)。

[0298] 在生成更新的模板表示1116之后,系统100使用更新的模板表示1116来更新对嵌入112。例如,系统100可以使用对更新的模板表示1116的相应 (i, j) 位置处的嵌入的“交叉注意力”来更新每个位置 (i, j) 处的相应的对嵌入112。在更新位置 (i, j) 处的对嵌入112的交叉注意力操作中,从位置 (i, j) 处的对嵌入生成查询嵌入,并且从更新的模板表示1116的相应位置 (i, j) 处的嵌入生成关键字和值嵌入。

[0299] 使用模板序列1110更新对嵌入112使得系统100能够用表征进化相关的模板序列1110的蛋白质结构的信息来丰富对嵌入,从而增强对嵌入112的信息量,并提高使用对嵌入预测的蛋白质结构的准确性。

[0300] 本说明书结合系统和计算机程序组件使用术语“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,意味着该系统已经在其上安装了软件、固件、硬件或它们的组合,这些软件、固件、硬件或它们的组合在操作中使得该系统执行这些操作或动作。一个或多个计算机程序被配置为执行特定的操作或动作意味着一个或多个程序包括指

令,当由数据处理装置执行时,这些指令使得该装置执行这些操作或动作。

[0301] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路、有形体现的计算机软件或固件、计算机硬件中实现,包括本说明书中公开的结构及其结构等同,或者它们中的一个或多个的组合。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序,即编码在有形非易失性存储介质上的计算机程序指令的一个或多个模块,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质可以是机器可读存储设备、机器可读存储基底、随机或串行存取存储设备,或者它们中的一个或多个的组合。可替换地或附加地,程序指令可以被编码在人工生成的传播信号上,例如,机器生成的电、光或电磁信号,其被生成以编码信息,用于传输到合适的接收器设备,以由数据处理装置执行。

[0302] 术语“数据处理装置”指的是数据处理硬件,并且包括用于处理数据的所有种类的装置、设备和机器,例如包括可编程处理器、计算机或多个处理器或计算机。该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,该装置可以可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。

[0303] 计算机程序也可以被称为或描述为程序、软件、软件应用、应用、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,或者声明性或过程性语言;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、组件、子例程或适用于计算环境的其他单元。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中,例如,存储在标记语言文档中的一个或多个脚本,存储在专用于所讨论的程序的单个文件中,或者存储在多个协作文件中,例如,存储一个或多个模块、子程序或代码部分的文件。计算机程序可以被部署为在一个计算机或位于一个地点或分布在多个地点并通过数据通信网络互连的多个计算机上执行。

[0304] 在本说明书中,术语“引擎”被广泛地用来指被编程来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。通常,引擎将被实现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组件。在某些情况下,一个或多个计算机将专用于特定的引擎;在其他情况下,可以在相同一个或多个计算机上安装和运行多个引擎。

[0305] 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来执行,所述一个或多个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。这些过程和逻辑流程也可以由专用逻辑电路(例如FPGA或ASIC)来执行,或者由专用逻辑电路和一个或多个编程计算机的组合来执行。

[0306] 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,或者任何其他种类的中央处理单元。通常,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指令的中央处理器和用于存储指令和数据的一个或多个存储设备。中央处理单元和存储器可以由专用逻辑电路来补充或并入其中。通常,计算机还将包括或可操作地耦合到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,以从该大容量存储设备接收数据或向其传送数据,或两者兼有。然而,计算机不需要有这样的设备。此外,计算机可以嵌入在另一个设备中,例如,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、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,仅举几个示例。

[0307] 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、介质和存储设备,包括例如半导体存储设备,例如EPROM、EEPROM和闪存设备;磁盘,例如内部硬盘或可移动磁盘;磁光盘;以及CDROM和DVD-ROM盘。

[0308] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在计算机上实现,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,以及用户可以用来向计算机提供输入的键盘和定点设备,例如鼠标或轨迹球。也可以使用其他类型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,向用户设备上的网络浏览器发送网页。此外,计算机可以通过向个人设备(例如,运行消息应用程序的智能手机)发送文本消息或其他形式的消息,并接收来自用户的响应消息来与用户交互。

[0309] 用于实现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的通用和计算密集型部分,即推理、工作负载。

[0310] 机器学习模型可以使用机器学习框架来实现和部署,例如TensorFlow框架、微软认知工具包框架、Apache Singa框架或Apache MXNet框架。

[0311] 本说明书中描述的主题的实施例可以在计算系统中实现,该计算系统包括后端组件,例如作为数据服务器,或者包括中间件组件,例如应用服务器,或者包括前端组件,例如具有图形用户界面的客户端计算机、网络浏览器或用户可以通过其与本说明书中描述的主题的实现交互的应用,或者一个或多个这样的后端、中间件或前端组件的任何组合。该系统的组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)来互连。通信网络的示例包括局域网(LAN)和广域网(WAN),例如因特网。

[0312] 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由于在各自的计算机上运行的计算机程序而生成的,并且彼此具有客户端-服务端关系。在一些实施例中,服务端向用户设备发送数据,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。可以在服务端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。

[0313] 虽然本说明书包含许多具体的实施细节,但是这些不应该被解释为对任何发明的范围或可能要求保护的范围的限制,而是对特定发明的特定实施例所特有的特征的描述。本说明书中在独立实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反,在单个实施例的上下文中描述的各种特征也可以在多个实施例中单独或以任何合适的子组合来实现。此外,尽管特征可能上面被描述为在某些组合中起作用,并且甚至最初被如此要求保护,但是来自所要求保护的组合的一个或多个特征在一些情况下可以从该组合中删除,并且所要求保护的组合可以针对子组合或子组合的变体。

[0314] 类似地,尽管在附图中以特定顺序描述了操作,并且在权利要求中以特定顺序叙述了操作,但是这不应该被理解为要求这些操作以所示的特定顺序或顺序来执行,或者要求所有示出的操作都被执行,以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中

都需要这样的分离,并且应该理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或者打包成多个软件产品。

[0315] 已经描述了主题的特定实施例。其他实施例在以下权利要求的范围内。例如,权利要求中所述的动作可以以不同的顺序来执行,并且仍然可以获得期望的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定顺序或连续顺序来实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。

蛋白质结构预测系统

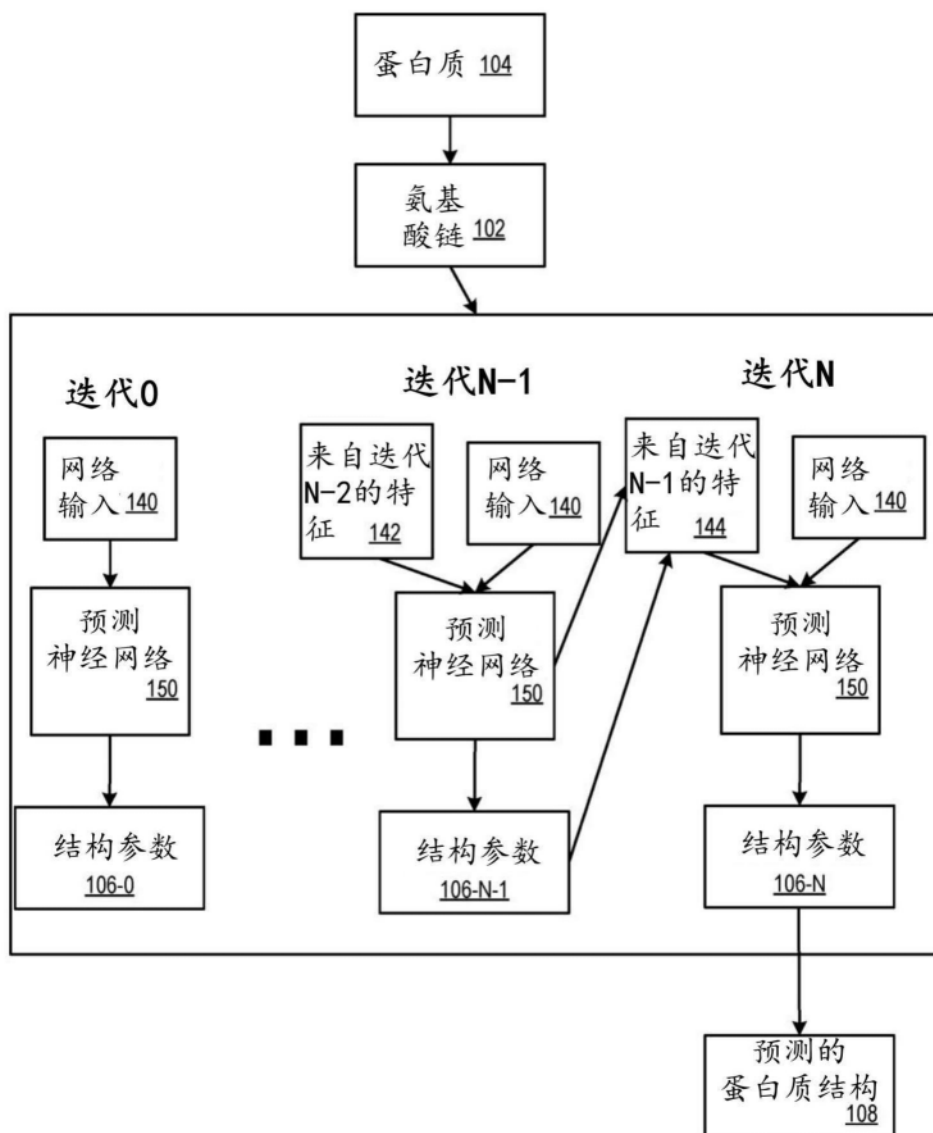
100

图1A

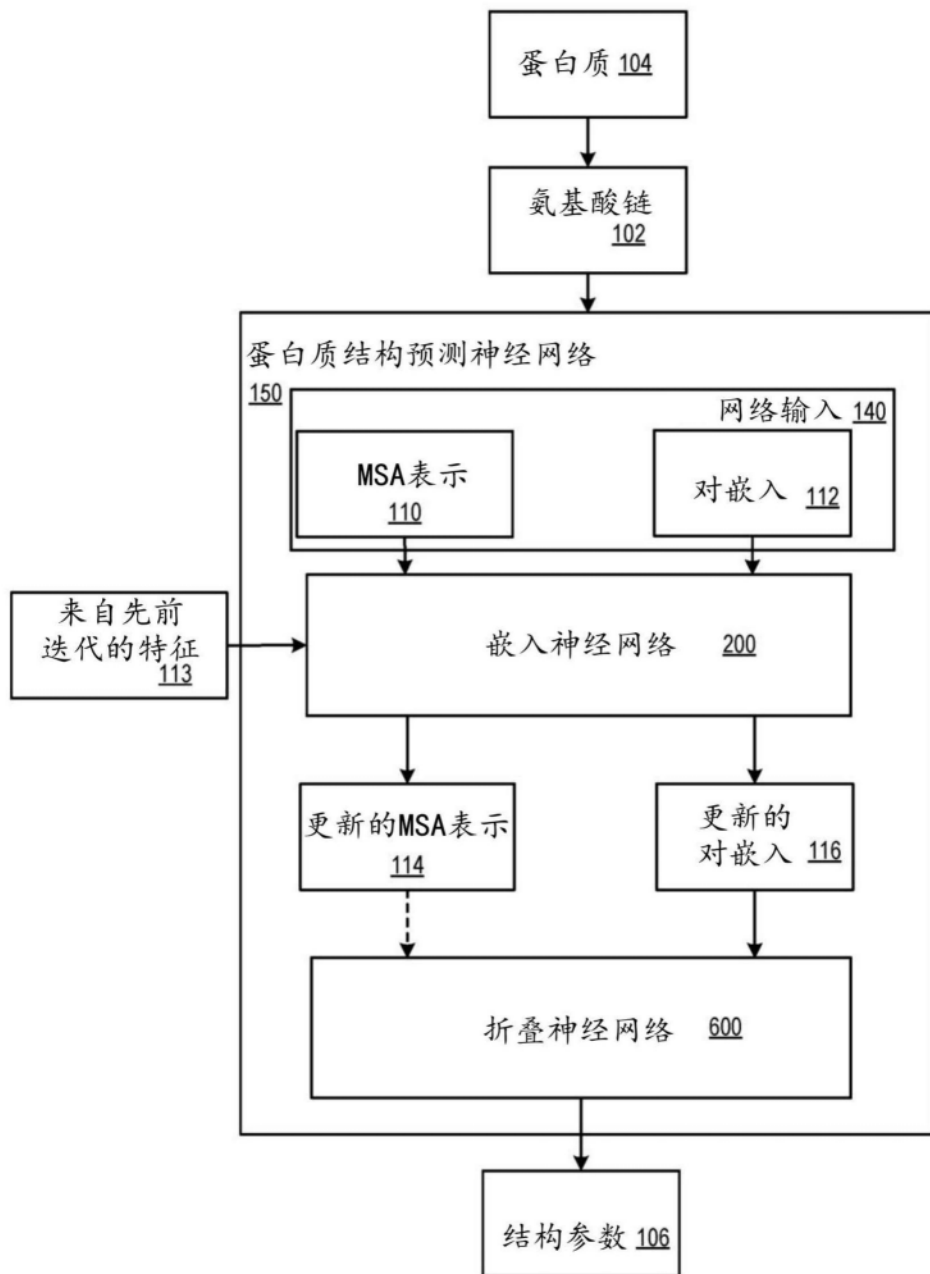


图1B

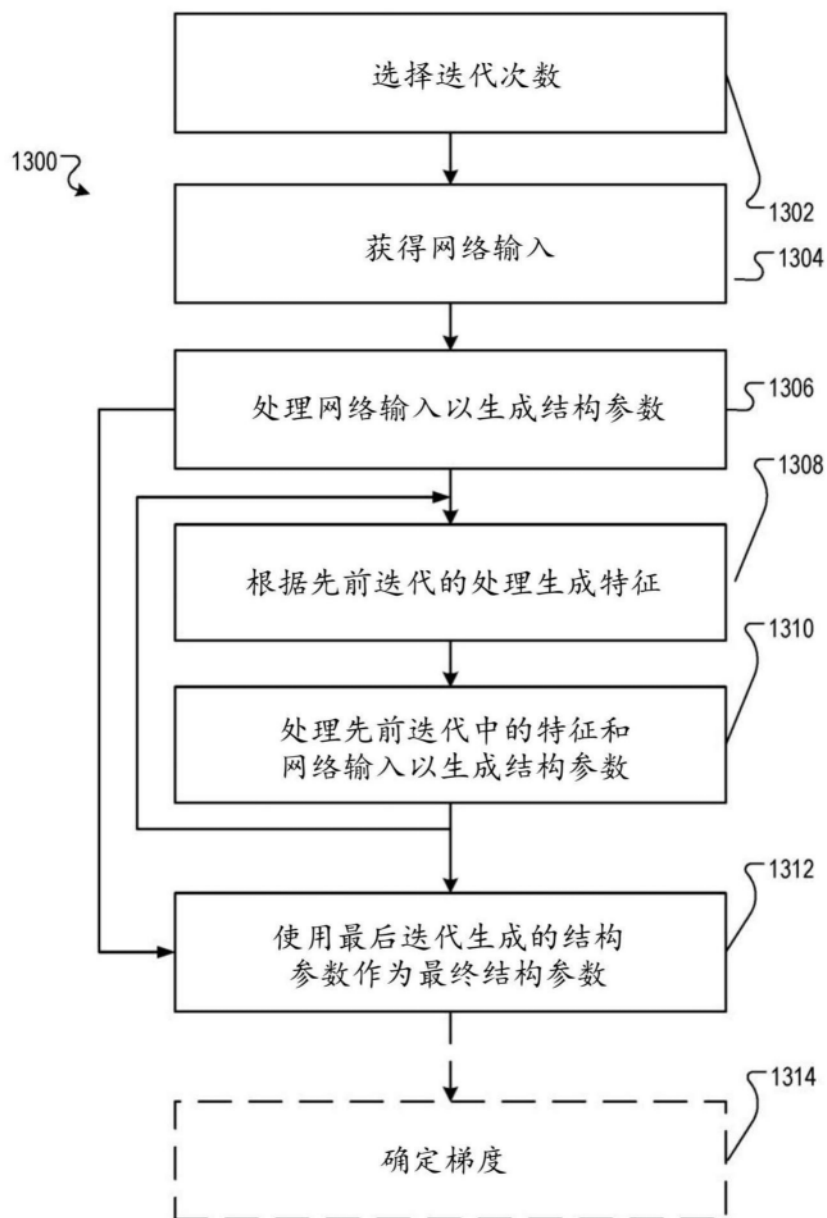


图1C

嵌入神经网络 200

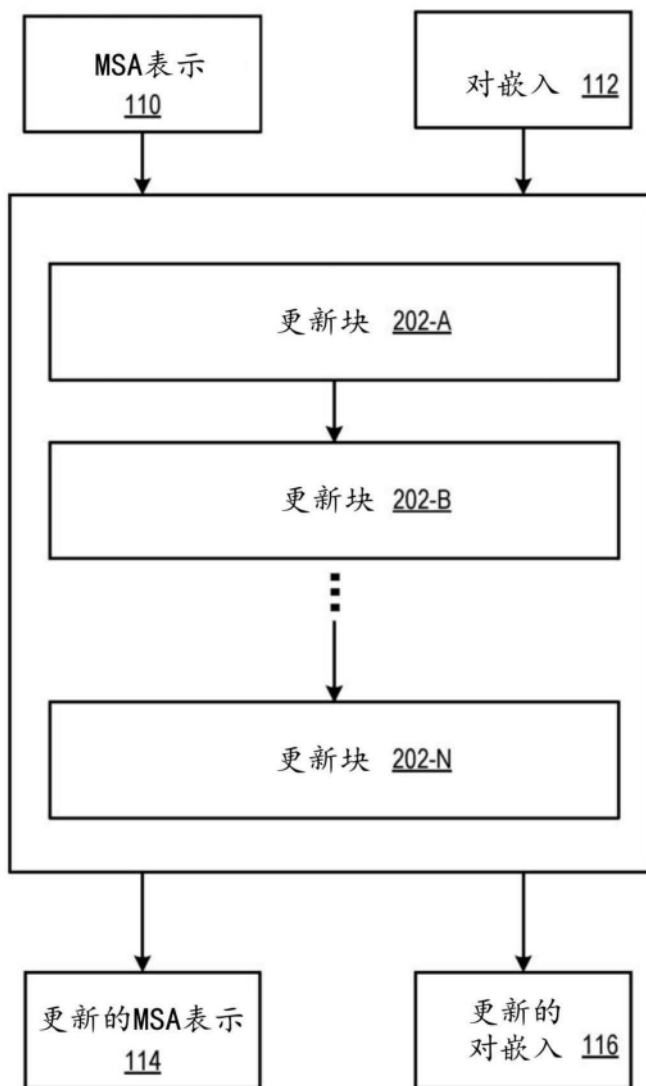


图2

嵌入神经网络更新块

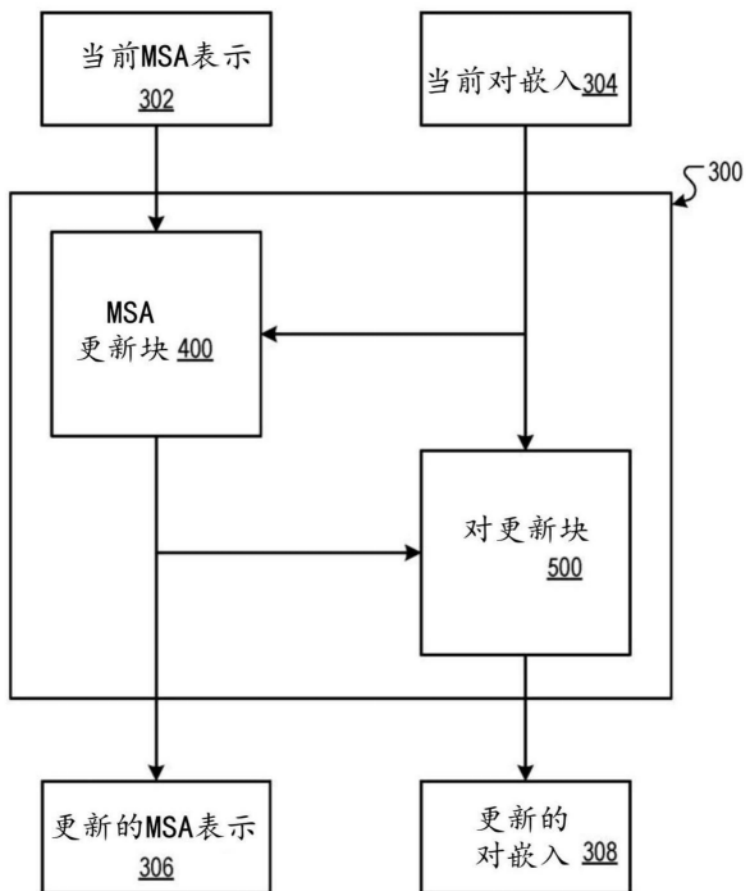


图3

MSA更新块 400

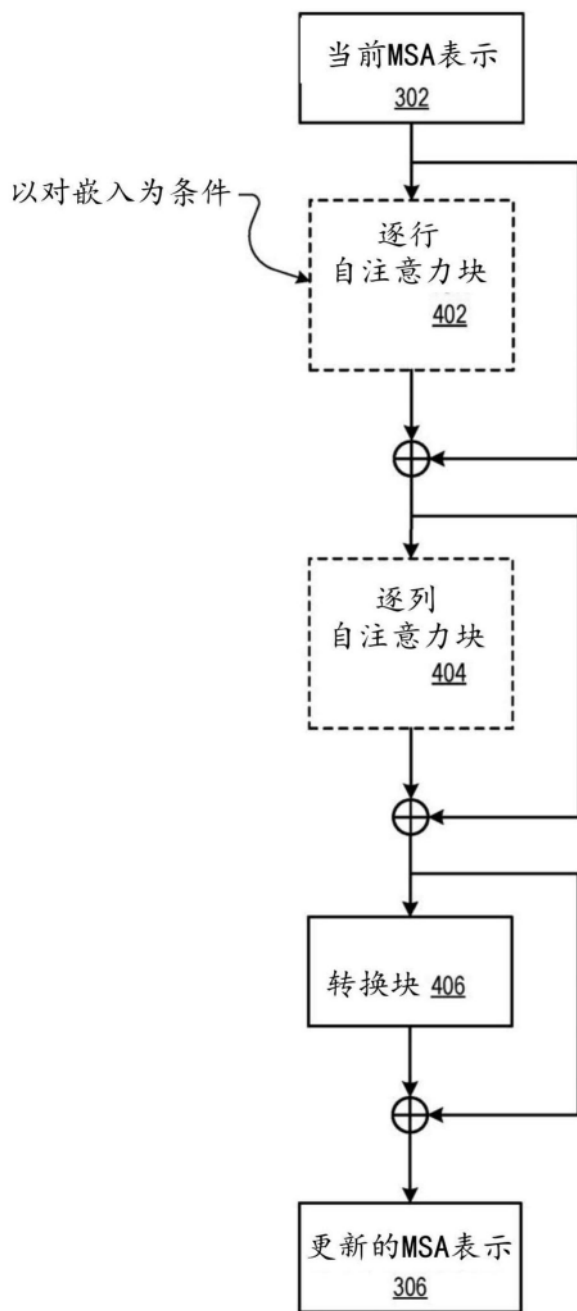


图4

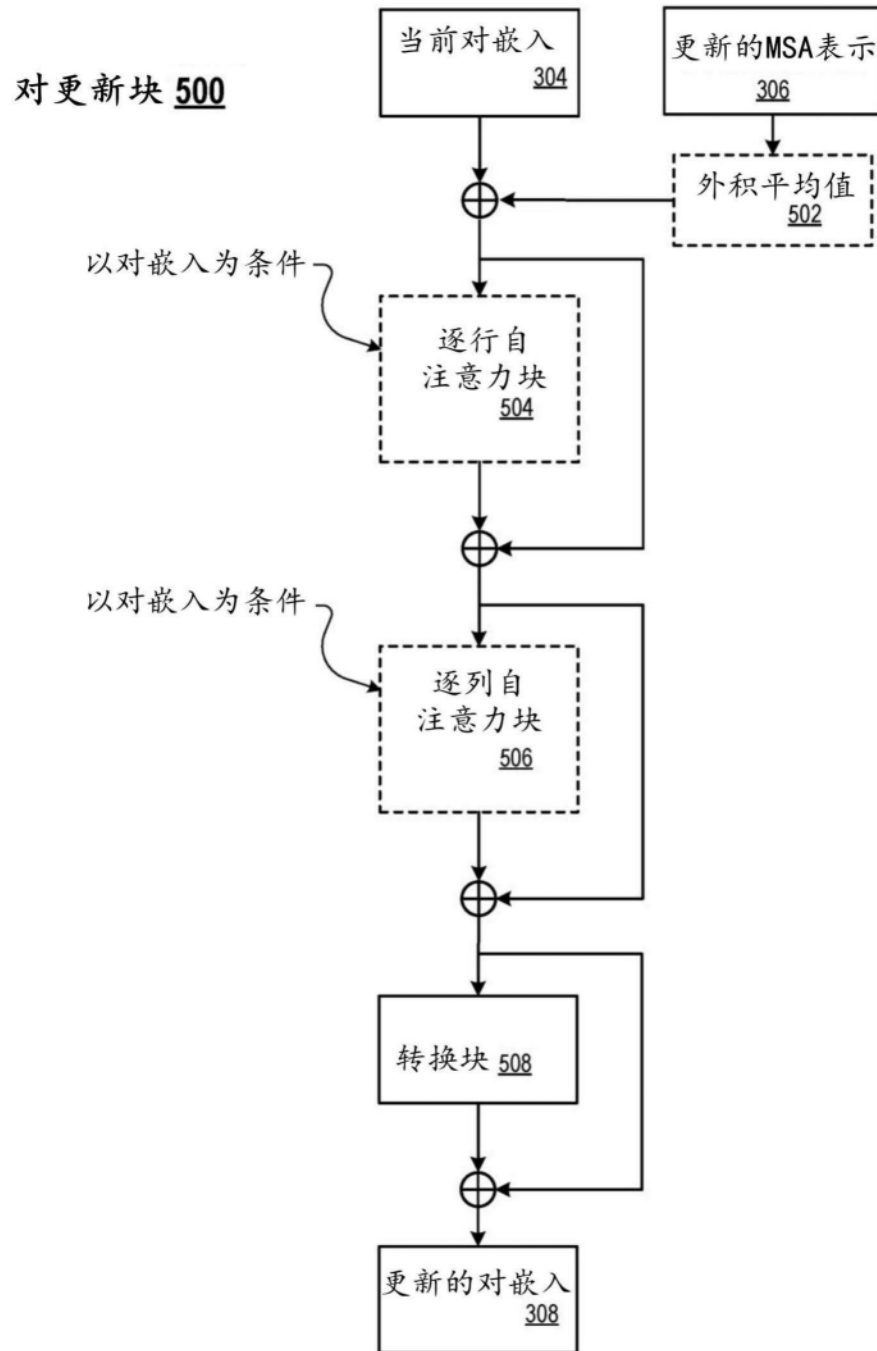


图5

折叠神经网络 600

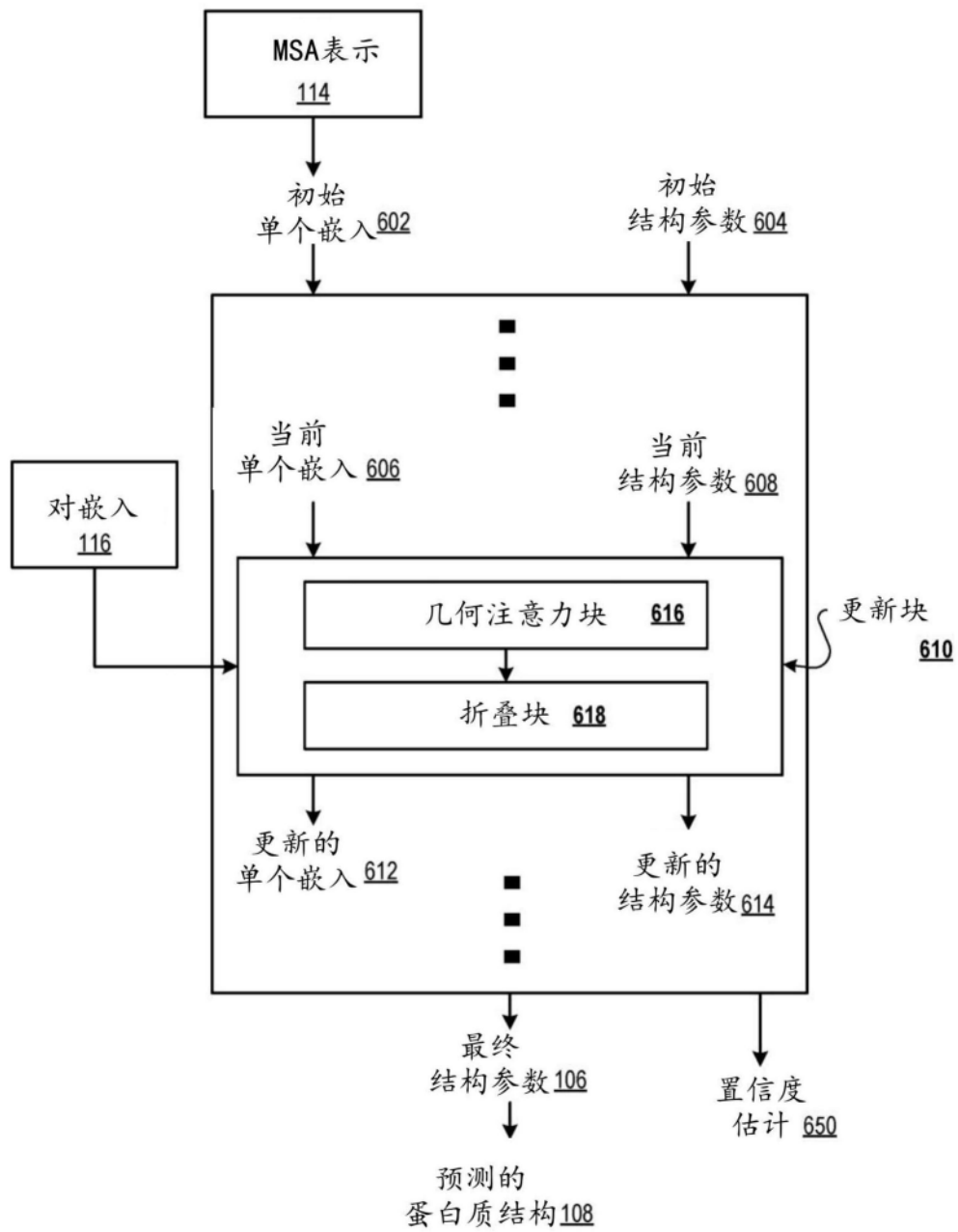


图6

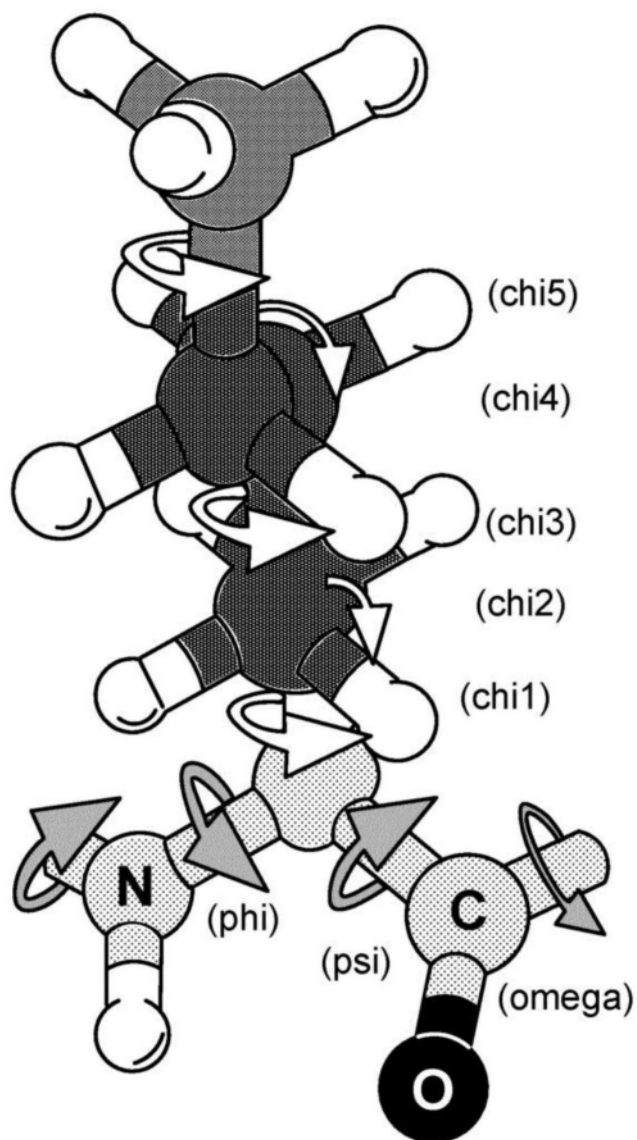


图7

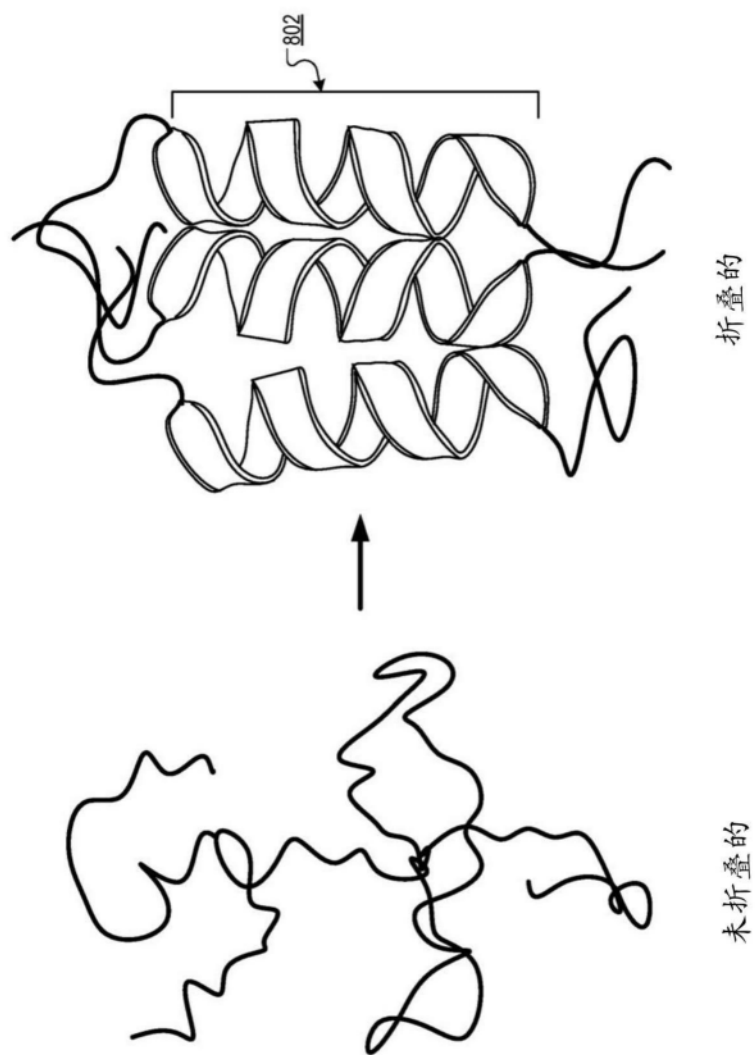


图8

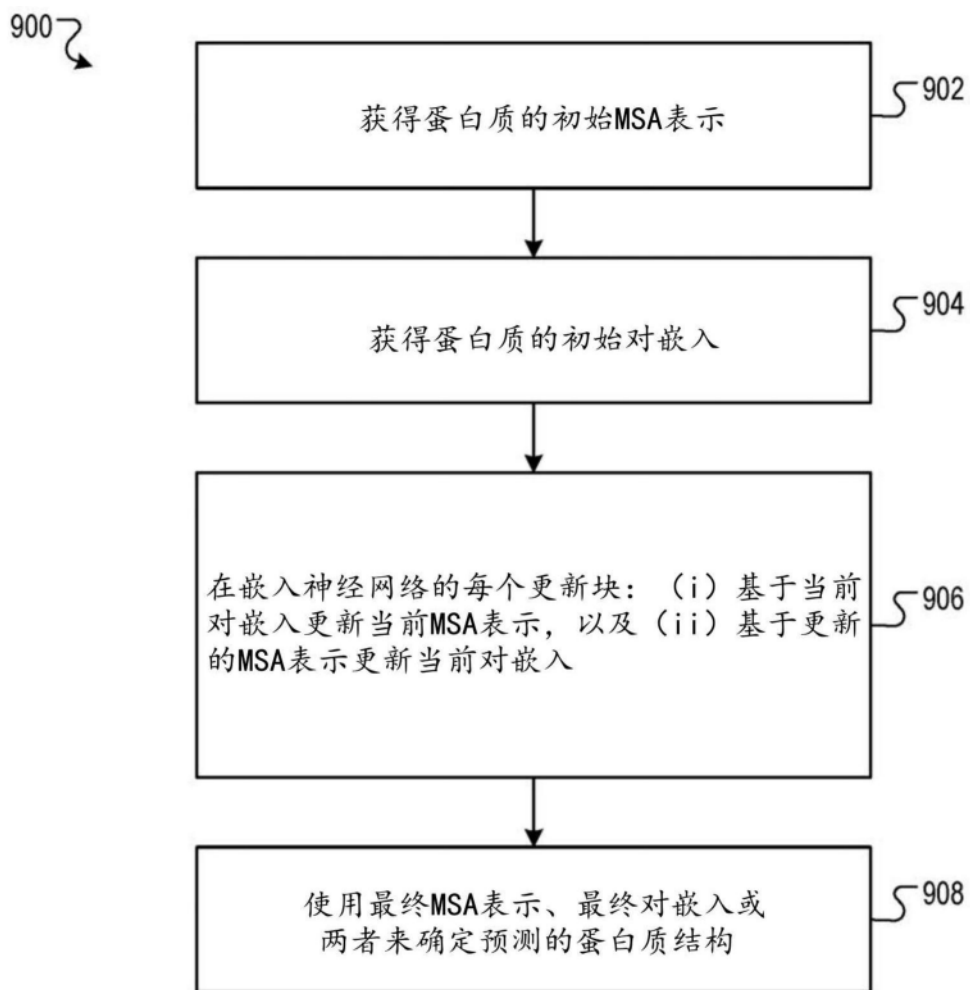


图9

生成蛋白质中氨基酸链的MSA表示

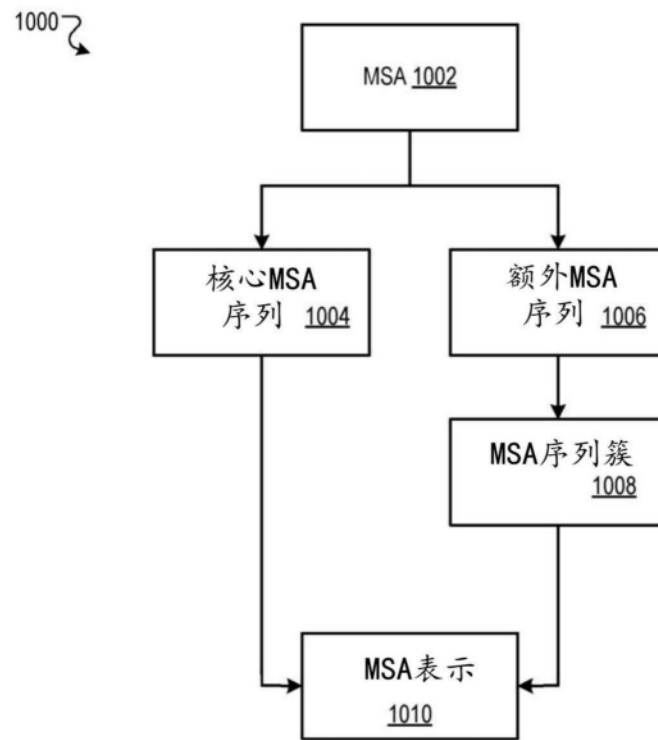


图10

生成初始对嵌入

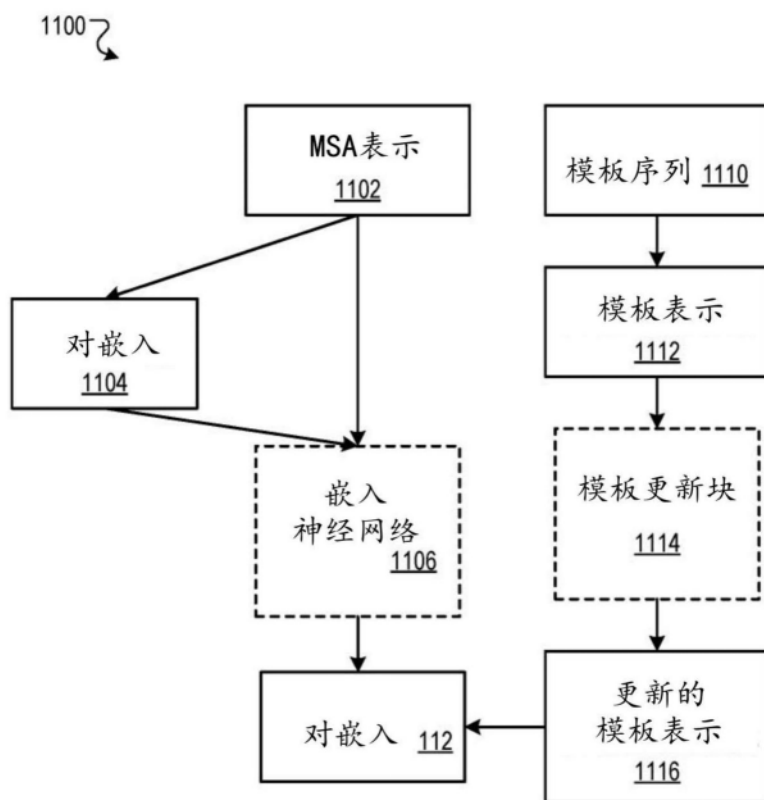


图11