



(12) 发明专利申请

(10) 申请公布号 CN 118318273 A

(43) 申请公布日 2024. 07. 09

(21) 申请号 202280078754.1

(22) 申请日 2022.10.04

(30) 优先权数据

63/252,137 2021.10.04 US

(85) PCT国际申请进入国家阶段日

2024.05.28

(86) PCT国际申请的申请数据

PCT/EP2022/077595 2022.10.04

(87) PCT国际申请的公布数据

W02023/057455 EN 2023.04.13

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 R·A·埃文斯 M·J·奥尼尔

A·普里策尔 N·O·安特洛波瓦

T·F·G·格林 J·江珀

(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 15/20 (2006.01)

G16B 40/20 (2006.01)

G16B 30/10 (2006.01)

G16B 5/20 (2006.01)

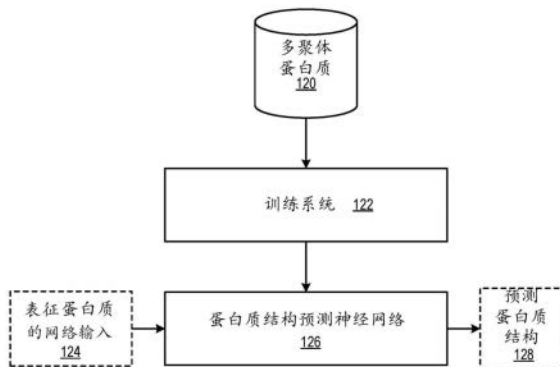
权利要求书7页 说明书38页 附图19页

(54) 发明名称

训练神经网络预测多链蛋白质结构

(57) 摘要

包括编码在计算机存储介质上的计算机程序的方法、系统和装置,用于使用蛋白质结构预测神经网络来预测包括多条氨基酸链的蛋白质的结构,其中,每条链包括相应的氨基酸序列。在一个方面,一种方法包括:接收蛋白质结构预测神经网络的网络输入,其中,网络输入表征蛋白质;使用蛋白质结构预测神经网络来处理表征蛋白质的网络输入,以生成表征蛋白质的预测结构的网络输出;以及基于网络输出来确定蛋白质的预测结构。



1. 一种由一个或多个计算机执行的方法,用于使用蛋白质结构预测神经网络来预测包括多条氨基酸链的蛋白质的结构,其中,每条链包括相应的氨基酸序列,所述方法包括:

接收蛋白质结构预测神经网络的网络输入,其中,所述网络输入表征蛋白质;

使用蛋白质结构预测神经网络来处理表征蛋白质的网络输入,以生成表征蛋白质的预测结构的网络输出;以及

基于网络输出来确定蛋白质的预测结构;

其中,所述蛋白质结构预测神经网络已经通过包括以下各项的操作进行了训练:

生成多个训练示例,其中,每个训练示例与包括多条氨基酸链的相应的训练蛋白质相对应,并且其中,生成每个训练示例包括:

获得标识以下各项的数据:(i) 训练蛋白质的每条氨基酸链的氨基酸序列,以及(ii) 训练蛋白质的靶结构;

选择包括训练蛋白质中包括的氨基酸的真子集的训练蛋白质的裁剪;以及

基于训练蛋白质的裁剪来生成训练示例,包括:

基于包括在训练蛋白质的裁剪中的氨基酸来生成蛋白质结构预测神经网络的网络输入;以及

基于与训练蛋白质的裁剪相对应的训练蛋白质的靶结构的真子集来生成网络输入的靶输出;以及

在所述多个训练示例上训练蛋白质结构预测神经网络。

2. 根据权利要求1所述的方法,其中,对于一个或多个训练示例,选择训练蛋白质的裁剪包括:

对于从训练蛋白质中的多条氨基酸链的排序中的第一条氨基酸链开始的训练蛋白质中的每条氨基酸链:

确定要从氨基酸链中裁剪的氨基酸序列的长度;以及

从氨基酸链中裁剪所确定的长度的氨基酸序列。

3. 根据权利要求2所述的方法,其中,确定要从氨基酸链中裁剪的氨基酸序列的长度包括:

基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及(ii) 当前从训练蛋白质中裁剪的氨基酸的数量。

4. 根据权利要求3所述的方法,其中,基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及(ii) 当前从训练蛋白质中裁剪的氨基酸的数量,包括:

确定要从氨基酸链中裁剪的氨基酸序列的最小长度;

确定要从氨基酸链中裁剪的氨基酸序列的最大长度;以及

根据在最小长度和最大长度之间的长度范围内的概率分布来确定要从氨基酸链中裁剪的氨基酸序列的长度。

5. 根据权利要求4所述的方法,其中,要从氨基酸链中裁剪的氨基酸序列的最小长度被确定为:

$$\min(n_k, \max(0, N_{\text{res}} - (n_{\text{added}} + n_{\text{remaining}})))$$

其中, n_k 是氨基酸链中的氨基酸的数量, N_{res} 是要从训练蛋白质中裁剪的氨基酸的最大

数量, n_{added} 是当前从训练蛋白质中裁剪的氨基酸的数量, 并且 $n_{\text{remaining}}$ 是在氨基酸链的排序中在氨基酸链之后的氨基酸链的组合长度。

6. 根据权利要求4-5中任一项所述的方法, 其中, 要从氨基酸链中裁剪的氨基酸序列的最大长度被确定为:

$$\min(N_{\text{res}} - n_{\text{added}}, n_k)$$

其中, N_{res} 是要从训练蛋白质中裁剪的氨基酸的最大数量, n_{added} 是当前从训练蛋白质中裁剪的氨基酸的数量, 并且 n_k 是氨基酸链中的氨基酸的数量。

7. 根据权利要求4-6中任一项所述的方法, 其中, 根据在最小长度和最大长度之间的长度范围内的概率分布来确定要从氨基酸链中裁剪的氨基酸序列的长度包括:

根据最小长度和最大长度之间的均匀概率分布来随机采样要从氨基酸链中裁剪的氨基酸序列的长度。

8. 根据任一前述权利要求所述的方法, 其中, 对于一个或多个训练示例, 选择训练蛋白质的裁剪包括:

从包括在训练蛋白质中的界面氨基酸集合当中采样界面氨基酸;

对于训练蛋白质中的多个氨基酸, 确定氨基酸与所采样的界面氨基酸之间的相应的空间距离; 以及

从训练蛋白质中的多个氨基酸当中裁剪预定数量的氨基酸, 所述预定数量的氨基酸与所采样的界面氨基酸具有最低空间距离。

9. 根据权利要求1所述的方法, 所述蛋白质结构预测神经网络的训练还包括:

获得多个训练示例, 其中, 每个训练示例与包括多条氨基酸链的相应的训练蛋白质相对应, 包括至少两条具有完全相同的氨基酸序列的氨基酸链, 其中, 每个训练示例包括: (i) 表征训练蛋白质的网络输入, 以及 (ii) 训练蛋白质的靶结构;

在所述多个训练示例中的每个训练示例上训练结构预测神经网络, 包括, 对于每个训练示例:

使用蛋白质结构预测神经网络来处理训练输入, 以生成训练蛋白质的预测结构;

(i) 从训练蛋白质的靶结构中的多条氨基酸链中选择锚链, 以及 (ii) 从训练蛋白质的预测结构中的多条氨基酸链中选择锚链;

对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐;

在对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐之后, 确定预测结构中的每条链到靶结构中的对应的链的一对一分配;

基于预测结构中的链到靶结构中的对应的链的分配, 使用目标函数的梯度来更新结构预测神经网络参数集合的当前值, 所述目标函数度量 (i) 预测结构与 (ii) 靶结构之间的误差。

10. 根据权利要求9所述的方法, 其中, 训练蛋白质的靶结构中的锚链与训练蛋白质的预测结构中的锚链具有相同的氨基酸序列。

11. 根据权利要求9-10中任一项所述的方法, 其中, 确定预测结构中的每条链到靶结构中的对应的链的分配包括:

直到预测结构中的每条链已经被分配给靶结构中的对应的链时, 迭代地执行包括以下各项的操作:

选择未分配的预测结构链；

对于与未分配的预测结构链具有相同的氨基酸序列的每条未分配的靶结构链，确定

(i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差；以及

基于误差将未分配的预测结构链分配给对应的未分配的靶结构链。

12. 根据权利要求11所述的方法，其中，基于误差将未分配的预测结构链分配给对应的未分配的靶结构链包括：

将未分配的预测结构链分配给与最低误差相关联的对应的未分配的靶结构链。

13. 根据权利要求11-12中任一项所述的方法，其中，确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差包括：

确定未分配的预测结构链中的氨基酸坐标的集中趋势的度量；

确定未分配的靶结构链中的氨基酸坐标的集中趋势的度量；以及

基于 (i) 未分配的预测结构链中的氨基酸坐标的集中趋势的度量与 (ii) 未分配的靶结构链中的氨基酸坐标的集中趋势的度量之间的差值的大小，确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差。

14. 根据权利要求13所述的方法，其中，所述集中趋势的度量是均值。

15. 根据权利要求9-14中任一项所述的方法，其中，更新结构预测神经网络参数集合的当前值包括：

对于每条靶结构链，确定靶结构链中的每个氨基酸与被分配给靶结构链的预测结构链中的对应的氨基酸之间的相应的定位误差；以及

基于定位误差来度量 (i) 预测结构与 (ii) 靶结构之间的误差。

16. 根据前述权利要求中任一项所述的方法，其中，表征蛋白质的网络输入包括交叉链遗传数据。

17. 根据任一前述权利要求所述的方法，其中，接收表征蛋白质的网络输入包括：

获得表示与蛋白质中的每条链相对应的相应的多序列对齐 (MSA) 的初始MSA表示；以及

获得蛋白质中的每对氨基酸的相应的初始成对嵌入；

其中，所述网络输入包括初始MSA表示和初始成对嵌入。

18. 根据权利要求1所述的方法，其中，所述蛋白质结构预测神经网络包括嵌入神经网络，并且其中，使用蛋白质结构预测神经网络来处理表征蛋白质的网络输入以生成表征蛋白质的预测结构的网络输出包括：

使用所述嵌入神经网络来处理包括初始MSA表示和初始成对嵌入的输入，以生成包括最终MSA表示和蛋白质中的每对氨基酸的相应的最终成对嵌入的输出，

其中，所述嵌入神经网络包括更新块的序列，

其中，每个更新块具有相应的更新块参数集合，并且执行包括以下各项的操作：

接收当前MSA表示和蛋白质中的每对氨基酸的相应的当前成对嵌入；

基于当前成对嵌入，根据更新块的更新块参数值来更新当前MSA表示；以及

基于更新的MSA表示，根据更新块的更新块参数值来更新当前成对嵌入；以及

使用最终MSA表示、最终成对嵌入或两者来确定蛋白质的预测结构。

19. 一种由一个或多个计算机执行的方法，用于使用蛋白质结构预测神经网络来预测包括多条氨基酸链的蛋白质的结构，其中，每条链包括相应的氨基酸序列，所述方法包括：

接收蛋白质结构预测神经网络的网络输入,其中,所述网络输入表征蛋白质;
使用蛋白质结构预测神经网络来处理表征蛋白质的网络输入,以生成表征蛋白质的预测结构的网络输出;以及

基于网络输出来确定蛋白质的预测结构;

其中,所述蛋白质结构预测神经网络已经通过包括以下各项的操作进行了训练:

获得多个训练示例,其中,每个训练示例与包括多条氨基酸链的相应的训练蛋白质相对应,包括至少两条具有完全相同的氨基酸序列的氨基酸链,其中,每个训练示例包括:(i) 表征训练蛋白质的网络输入,以及(ii) 训练蛋白质的靶结构;

在所述多个训练示例中的每个训练示例上训练结构预测神经网络,包括,对于每个训练示例:

使用蛋白质结构预测神经网络来处理训练输入,以生成训练蛋白质的预测结构;

(i) 从训练蛋白质的靶结构中的多条氨基酸链中选择锚链,以及(ii) 从训练蛋白质的预测结构中的多条氨基酸链中选择锚链;

对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐;

在对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐之后,确定预测结构中的每条链到靶结构中的对应的链的一对一分配;

基于预测结构中的链到靶结构中的对应的链的分配,使用目标函数的梯度来更新结构预测神经网络参数集合的当前值,所述目标函数度量(i) 预测结构与(ii) 靶结构之间的误差。

20. 根据权利要求19所述的方法,其中,训练蛋白质的靶结构中的锚链与训练蛋白质的预测结构中的锚链具有相同的氨基酸序列。

21. 根据权利要求19-20中任一项所述的方法,其中,确定预测结构中的每条链到靶结构中的对应的链的分配包括:

直到预测结构中的每条链已经被分配给靶结构中的对应的链时,迭代地执行包括以下各项的操作:

选择未分配的预测结构链;

对于与未分配的预测结构链具有相同的氨基酸序列的每条未分配的靶结构链,确定(i) 未分配的预测结构链与(ii) 未分配的靶结构链之间的相应的误差;以及

基于误差将未分配的预测结构链分配给对应的未分配的靶结构链。

22. 根据权利要求21所述的方法,其中,基于误差将未分配的预测结构链分配给对应的未分配的靶结构链包括:

将未分配的预测结构链分配给与最低误差相关联的对应的未分配的靶结构链。

23. 根据权利要求21-22中任一项所述的方法,其中,确定(i) 未分配的预测结构链与(ii) 未分配的靶结构链之间的相应的误差包括:

确定未分配的预测结构链中的氨基酸坐标的集中趋势的度量;

确定未分配的靶结构链中的氨基酸坐标的集中趋势的度量;以及

基于(i) 未分配的预测结构链中的氨基酸坐标的集中趋势的度量与(ii) 未分配的靶结构链中的氨基酸坐标的集中趋势的度量之间的差值的大小,确定(i) 未分配的预测结构链与(ii) 未分配的靶结构链之间的相应的误差。

24. 根据权利要求23所述的方法, 其中, 所述集中趋势的度量是均值。

25. 根据权利要求19-24中任一项所述的方法, 其中, 更新结构预测神经网络参数集合的当前值包括:

对于每条靶结构链, 确定靶结构链中的每个氨基酸与被分配给靶结构链的预测结构链中的对应的氨基酸之间的相应的定位误差; 以及

基于定位误差来度量 (i) 预测结构与 (ii) 靶结构之间的误差。

26. 一种系统包括:

一个或多个计算机; 以及

一个或多个存储设备, 与所述一个或多个计算机通信耦合, 其中, 所述一个或多个存储设备存储指令, 当所述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根据权利要求1-25中任一项所述的相应方法的操作。

27. 一个或多个存储指令的非暂时性计算机存储介质, 当所述指令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1-25中任一项所述的相应方法的操作。

28. 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:

执行根据权利要求1-25中任一项所述的方法, 以确定靶蛋白质的预测结构;

评估一个或多个候选配体与靶蛋白质的预测结构的相互作用; 以及

取决于评估的结果来选择候选配体中的一个或多个作为配体。

29. 根据权利要求28所述的方法, 其中, 所述靶蛋白质包括受体或酶, 并且其中, 所述配体是受体或酶的激动剂或拮抗剂; 或者其中, 所述配体包括抗体, 并且靶蛋白质包括抗体靶, 特别是病毒或癌细胞蛋白质, 并且其中, 所述抗体与所述抗体靶键合以提供治疗效果。

30. 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:

对于一个或多个候选多肽配体中的每个候选多肽配体, 执行根据权利要求1-25中任一项所述的方法以确定候选多肽配体的预测结构;

获得靶蛋白质的靶蛋白质结构;

评估所述一个或多个候选多肽配体中的每个候选多肽配体的预测结构与靶蛋白质结构之间的相互作用; 以及

取决于评估的结果来选择候选多肽配体中的一个或多个中的一个作为多肽配体。

31. 一种获得疾病的诊断抗体标记物的方法, 所述方法包括:

对于一个或多个候选抗体中的每个候选抗体, 执行根据权利要求1-25中任一项所述的方法以确定候选抗体的预测结构;

获得靶蛋白质的靶蛋白质结构;

评估所述一个或多个候选抗体中的每个候选抗体的预测结构与靶蛋白质结构之间的相互作用; 以及

取决于评估的结果来选择候选抗体中的一个或多个中的一个作为诊断抗体标记物。

32. 根据权利要求28-31中任一项所述的方法, 其中, 评估候选配体或候选抗体之一的相互作用包括确定候选配体或候选抗体的相互作用分数, 其中, 所述相互作用分数包括候选配体或候选抗体与靶蛋白质之间的相互作用的度量。

33. 根据权利要求28-32中任一项所述的方法, 还包括合成配体或抗体。

34. 根据权利要求33所述的方法,还包括在体外和体内测试配体或抗体的生物活性。

35. 一种识别蛋白质错误折叠疾病的存在的方法,包括:

执行根据权利要求1-25中任一项所述的方法以确定蛋白质的预测结构;

获得从人体或动物体中获得的蛋白质版本的结构;

将蛋白质的预测结构与从人体或动物体中获得的蛋白质版本的结构进行比较;以及
取决于比较的结果来识别蛋白质错误折叠疾病的存在。

36. 一种获得与配体相互作用的蛋白质的方法,其中,所述配体是药物或工业酶的配体,所述方法包括:

执行根据权利要求1-25中任一项所述的方法以确定多个候选蛋白质中的每个候选蛋白质的相应预测结构;

评估配体与候选蛋白质中的每一个的预测结构的相互作用;以及

取决于评估的结果来选择候选蛋白质中的一个或多个。

37. 根据权利要求36所述的方法,其中,所选择的蛋白质包括受体或酶,并且其中,所述配体是受体或酶的激动剂或拮抗剂。

38. 根据权利要求36或37所述的方法,还包括合成所选择的蛋白质。

39. 根据权利要求38所述的方法,还包括在体外和体内测试所选择的蛋白质的生物活性。

40. 根据权利要求36-39中任一项所述的方法,其中,评估配体与候选蛋白质的预测结构的相互作用包括确定候选蛋白质的相互作用分数,其中,所述相互作用分数包括候选蛋白质与配体之间的相互作用的度量。

41. 一种由一个或多个计算机执行的方法,所述方法包括:

获得标识训练蛋白质的每条氨基酸链的氨基酸序列的数据;

选择包括训练蛋白质中包括的氨基酸的真子集的训练蛋白质的裁剪,包括,对于从训练蛋白质中的多条氨基酸链的排序中的第一条氨基酸链开始的训练蛋白质中的每条氨基酸链:

确定要从氨基酸链中裁剪的氨基酸序列的长度;以及

从氨基酸链中裁剪所确定的长度的氨基酸序列。

42. 根据权利要求41所述的方法,其中,确定要从氨基酸链中裁剪的氨基酸序列的长度包括:

基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及(ii) 当前从训练蛋白质中裁剪的氨基酸的数量。

43. 根据权利要求42所述的方法,其中,基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及(ii) 当前从训练蛋白质中裁剪的氨基酸的数量,包括:

确定要从氨基酸链中裁剪的氨基酸序列的最小长度;

确定要从氨基酸链中裁剪的氨基酸序列的最大长度;以及

根据在最小长度和最大长度之间的长度范围内的概率分布来确定要从氨基酸链中裁剪的氨基酸序列的长度。

44. 一种由一个或多个计算机执行的方法,所述方法包括:

获得标识训练蛋白质的每条氨基酸链的氨基酸序列的数据；
选择包括训练蛋白质中包括的氨基酸的真子集的训练蛋白质的裁剪,包括：
从包括在训练蛋白质中的界面氨基酸集合当中采样界面氨基酸；
对于训练蛋白质中的多个氨基酸,确定氨基酸与所采样的界面氨基酸之间的相应的空间距离;以及
从训练蛋白质中的多个氨基酸当中裁剪预定数量的氨基酸,所述预定数量的氨基酸与所采样的界面氨基酸具有最低空间距离。

训练神经网络预测多链蛋白质结构

[0001] 相关申请的交叉引用

[0002] 本申请要求于2021年10月4日提交的美国临时专利申请序列第63/252137号的提交日的权益,其全部内容通过引用结合于此。

背景技术

[0003] 本说明书涉及预测蛋白质结构。

[0004] 蛋白质是由一个或多个氨基酸序列(“链”)确定的。氨基酸是一种包括氨基官能团和羧基官能团以及氨基酸所特有的侧链(即,原子团)的有机化合物。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。蛋白质的结构定义了蛋白质在经历蛋白质折叠之后在蛋白质的氨基酸序列中的原子的3-D构型。当处于由肽键联接的序列中时,氨基酸可以被称为氨基酸残基。

[0005] 预测可以使用机器学习模型来进行。机器学习模型接收输入,并且基于接收到的输入来生成输出,例如预测输出。一些机器学习模型是参数模型,并且基于接收到的输入和模型的值来生成输出。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,深度神经网络是包括一个输出层和一个或多个隐藏层的深度机器学习模型,每个隐藏层对接收到的输入应用非线性变换以生成输出。

发明内容

[0006] 本说明书描述了被实现为一个或多个位置中的一个或多个计算机上的用于预测多链蛋白质结构的计算机程序的神经网络系统。

[0007] 本说明书中使用的术语“蛋白质”可以被理解是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,该部分能够几乎独立于蛋白质的其余部分而经历蛋白质折叠。作为另一示例,术语蛋白质可以指蛋白质复合物(多聚体),即,包括共同折叠成蛋白质结构的多条氨基酸链。

[0008] “多链”蛋白质或“多聚体”蛋白质可以指包括多条氨基酸链的蛋白质。“单链”蛋白质可以指仅包括单个氨基酸链的蛋白质。

[0009] 具有多条氨基酸链的蛋白质中的氨基酸链的“多重性”可以指氨基酸链在蛋白质中重复的次数。多重性为二、三或四等的蛋白质被称为二聚体、三聚体或四聚体等。

[0010] 针对蛋白质中的氨基酸链的“多序列对齐”(multiple sequence alignment,MSA)指定该氨基酸链与例如来自其他蛋白质(例如,同源蛋白质)的多条附加的氨基酸链的序列对齐。更具体地,MSA可以定义该氨基酸链中的定位与多条附加的氨基酸链中的对应的定位之间的对应关系。针对氨基酸链的MSA可以例如通过使用任何适当的计算序列对齐技术(例如,渐进对齐构造)处理氨基酸链数据库而生成。MSA中的氨基酸链可以被理解为具有进化关系,例如,在进化关系中,MSA中的每条氨基酸链可以共享共有的祖先。在针对氨基酸链的MSA中氨基酸链中的氨基酸之间的相关性可以编码与预测氨基酸链的结构相关的信息。

[0011] 实体(例如,一对氨基酸)的“嵌入”可以指将实体表示为数值的有序集合,例如,数

值的向量或矩阵。

[0012] 蛋白质结构可以由结构参数集合来定义。定义蛋白质结构的结构参数集合可以被表示为数值的有序集合。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。

[0013] 在一个示例中,定义蛋白质结构的结构参数包括蛋白质中的每个氨基酸的(i)位置参数和(ii)旋转参数。

[0014] 氨基酸的位置参数可以指定蛋白质结构中的氨基酸中的特定原子的预测3-D空间位置。特定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。氨基酸的位置参数可以以任何适当的坐标系(例如,三维[x,y,z]笛卡尔坐标系)来表示。

[0015] 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“朝向”。更具体地,旋转参数可以指定一种3-D空间旋转操作,如果将这种3-D空间旋转操作应用于位置参数的坐标系,则使得氨基酸中的三个“主链”原子相对于旋转后的坐标系而呈现固定的定位。氨基酸中的三个主链原子可以指氨基酸中联接的一系列氮原子、 α 碳原子和羰基碳原子。氨基酸的旋转参数可以被表示为例如行列式等于1的 3×3 标准正交矩阵。

[0016] 通常,氨基酸的位置参数和旋转参数定义了氨基酸的自我中心参考系。在该参考系中,每个氨基酸的侧链可以从原点开始,并且沿着侧链的第一个键(bond)(即, α 碳- β 碳键)可以是沿着所定义的方向的。

[0017] 在另一示例中,定义蛋白质结构的结构参数可以包括“距离图”,“距离图”表征蛋白质中的每对氨基酸之间的相应估计距离(例如,以埃(angstroms)为单位来度量)。距离图可以例如通过一对氨基酸之间的可能距离的集合内的概率分布来表征该对氨基酸之间的估计距离。

[0018] 在另一示例中,定义蛋白质结构的结构参数可以定义蛋白质结构中的每个氨基酸中的每个原子的三维(3D)空间位置。

[0019] 根据第一方面,提供了一种由一个或多个计算机执行的方法,用于使用蛋白质结构预测神经网络来预测包括多条氨基酸链的蛋白质的结构,其中,每条链包括相应的氨基酸序列,该方法包括:接收蛋白质结构预测神经网络的网络输入,其中,网络输入表征蛋白质;使用蛋白质结构预测神经网络来处理表征蛋白质的网络输入,以生成表征蛋白质的预测结构的网络输出;以及基于网络输出来确定蛋白质的预测结构。

[0020] 在一些实施方式中,蛋白质结构预测神经网络是通过包括以下各项的操作来训练的:获得多个训练示例,其中,每个训练示例与包括多条氨基酸链的相应的训练蛋白质相对应,包括至少两条具有完全相同的氨基酸序列的氨基酸链,其中,每个训练示例包括:(i)表征训练蛋白质的网络输入,以及(ii)训练蛋白质的靶结构;在多个训练示例中的每个训练示例上训练结构预测神经网络,包括,对于每个训练示例:使用蛋白质结构预测神经网络来处理训练输入,以生成训练蛋白质的预测结构;(i)从训练蛋白质的靶结构中的多条氨基酸链中选择锚链,以及(ii)从训练蛋白质的预测结构中的多条氨基酸链中选择锚链;对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐;在对靶结构进行变换以将靶结构中的锚链与预测结构中的锚链对齐之后,确定预测结构中的每条链到靶结构中的对应的链的一对一分配;基于预测结构中的链到靶结构中的对应的链的分配,使用目标函数的梯

度来更新结构预测神经网络参数集合的当前值,该目标函数度量 (i) 预测结构与 (ii) 靶结构之间的误差。

[0021] 在一些实施方式中,训练蛋白质的靶结构中的锚链与训练蛋白质的预测结构中的锚链具有相同的氨基酸序列。

[0022] 在一些实施方式中,确定预测结构中的每条链到靶结构中的对应的链的分配包括:到预测结构中的每条链已经被分配给靶结构中的对应的链时,迭代地执行包括以下各项的操作:选择未分配的预测结构链;对于与未分配的预测结构链具有相同的氨基酸序列的每条未分配的靶结构链,确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差;以及基于误差将未分配的预测结构链分配给对应的未分配的靶结构链。

[0023] 在一些实施方式中,基于误差将未分配的预测结构链分配给对应的未分配的靶结构链包括:将未分配的预测结构链分配给与最低误差相关联的对应的未分配的靶结构链。

[0024] 在一些实施方式中,确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差包括:确定未分配的预测结构链中的氨基酸坐标的集中趋势的度量;确定未分配的靶结构链中的氨基酸坐标的集中趋势的度量;以及基于 (i) 未分配的预测结构链中的氨基酸坐标的集中趋势的度量与 (ii) 未分配的靶结构链中的氨基酸坐标的集中趋势的度量之间的差值的大小,确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差。

[0025] 在一些实施方式中,集中趋势的度量是均值。

[0026] 在一些实施方式中,更新结构预测神经网络参数集合的当前值包括:对于每条靶结构链,确定靶结构链中的每个氨基酸与被分配给靶结构链的预测结构链中的对应的氨基酸之间的相应的定位误差;以及基于定位误差来度量 (i) 预测结构与 (ii) 靶结构之间的误差。

[0027] 在一些实施方式中,蛋白质结构预测神经网络是通过包括以下各项的操作来训练的:生成多个训练示例,其中,每个训练示例与包括多条氨基酸链的相应的训练蛋白质相对应,并且其中,生成每个训练示例包括:获得标识以下各项的数据:(i) 训练蛋白质的每条氨基酸链的氨基酸序列,以及 (ii) 训练蛋白质的靶结构;选择包括训练蛋白质中包括的氨基酸的真子集的训练蛋白质的裁剪;以及基于训练蛋白质的裁剪来生成训练示例,包括:基于包括在训练蛋白质的裁剪中的氨基酸来生成蛋白质结构预测神经网络的网络输入;以及基于与训练蛋白质的裁剪相对应的训练蛋白质的靶结构的真子集来生成网络输入的靶输出;以及在多个训练示例上训练蛋白质结构预测神经网络。

[0028] 在一些实施方式中,对于一个或多个训练示例,选择训练蛋白质的裁剪包括:对于从训练蛋白质中的多条氨基酸链的排序中的第一条氨基酸链开始的训练蛋白质中的每条氨基酸链:确定要从氨基酸链中裁剪的氨基酸序列的长度;以及从氨基酸链中裁剪所确定的长度的氨基酸序列。

[0029] 在一些实施方式中,确定要从氨基酸链中裁剪的氨基酸序列的长度包括:基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及 (ii) 当前从训练蛋白质中裁剪的氨基酸的数量。

[0030] 在一些实施方式中,基于以下各项来确定要从氨基酸链中裁剪的氨基酸序列的长度:(i) 要从训练蛋白质中裁剪的氨基酸的最大数量,以及 (ii) 当前从训练蛋白质中裁剪的

氨基酸的数量,包括:确定要从氨基酸链中裁剪的氨基酸序列的最小长度;确定要从氨基酸链中裁剪的氨基酸序列的最大长度;以及根据在最小长度和最大长度之间的长度范围内的概率分布来确定要从氨基酸链中裁剪的氨基酸序列的长度。

[0031] 在一些实施方式中,要从氨基酸链中裁剪的氨基酸序列的最小长度被确定为:

[0032] $\min(n_k, \max(0, N_{\text{res}} - (n_{\text{added}} + n_{\text{remaining}})))$

[0033] 其中, n_k 是氨基酸链中的氨基酸的数量, N_{res} 是要从训练蛋白质中裁剪的氨基酸的最大数量, n_{added} 是当前从训练蛋白质中裁剪的氨基酸的数量, 并且 $n_{\text{remaining}}$ 是在氨基酸链的排序中在该氨基酸链之后的氨基酸链的组合长度。

[0034] 在一些实施方式中,要从氨基酸链中裁剪的氨基酸序列的最大长度被确定为:

[0035] $\min(N_{\text{res}} - n_{\text{added}}, n_k)$

[0036] 其中, N_{res} 是要从训练蛋白质中裁剪的氨基酸的最大数量, n_{added} 是当前从训练蛋白质中裁剪的氨基酸的数量, 并且 n_k 是氨基酸链中的氨基酸的数量。

[0037] 在一些实施方式中,根据在最小长度和最大长度之间的长度范围内的概率分布来确定要从氨基酸链中裁剪的氨基酸序列的长度:根据最小长度和最大长度之间的均匀概率分布来随机采样要从氨基酸链中裁剪氨基酸序列的长度。

[0038] 在一些实施方式中,对于一个或多个训练示例,选择训练蛋白质的裁剪包括:从包括在训练蛋白质中的界面氨基酸集合当中采样一个界面氨基酸;对于训练蛋白质中的多个氨基酸,确定氨基酸与所采样的界面氨基酸之间的相应的空间距离;以及从训练蛋白质的多个氨基酸中裁剪预定数量的氨基酸,该预定数量的氨基酸与所采样的界面氨基酸具有最低空间距离。

[0039] 在一些实施方式中,表征蛋白质的网络输入包括交叉链遗传数据。

[0040] 在一些实施方式中,接收表征蛋白质的网络输入包括:获得表示与蛋白质中的每条链相对应的相应的多序列对齐(MSA)的初始MSA表示;以及获得蛋白质中的每对氨基酸的相应的初始成对嵌入;其中,网络输入包括初始MSA表示和初始成对嵌入。

[0041] 在一些实施方式中,本文所述的方法还包括:获得表示与蛋白质中的每条链相对应的相应的多序列对齐(MSA)的初始MSA表示;获得蛋白质中的每对氨基酸的相应的初始成对嵌入;使用嵌入神经网络来处理包括初始MSA表示和初始成对嵌入的输入,以生成包括最终MSA表示和蛋白质中的每对氨基酸的相应的最终成对嵌入的输出,其中,嵌入神经网络包括更新块的序列,其中,每个更新块具有相应的更新块参数集合,并且执行包括以下各项的操作:接收当前MSA表示和蛋白质中的每对氨基酸的相应的当前成对嵌入;基于当前成对嵌入,根据更新块的更新块参数值来更新当前MSA表示;以及基于更新的MSA表示,根据更新块的更新块参数值来更新当前成对嵌入;以及使用最终MSA表示、最终成对嵌入或两者来确定蛋白质的预测结构。

[0042] 在一些实施方式中,当前MSA表示包括多个嵌入,并且基于当前成对嵌入来更新当前MSA表示包括:使用对MSA表示中的嵌入的注意力来更新当前MSA表示,其中,注意力以当前成对嵌入为条件。

[0043] 在一些实施方式中,使用对当前MSA表示中的嵌入的注意力来更新当前MSA表示包括:基于当前MSA表示来生成多个注意力权重;基于当前成对嵌入来生成与每个注意力权重相对应的相应的注意力偏差;基于注意力权重和注意力偏差来生成多个有偏差的注意力权

重;以及

[0044] 基于有偏差的注意力权重,使用对当前MSA表示中的嵌入的注意力来更新当前MSA表示中的嵌入。

[0045] 在一些实施方式中,基于有偏差的注意力权重使用注意力来更新当前MSA表示中的嵌入包括,对于当前MSA表示中的每个嵌入:基于有偏差的注意力权重,仅使用对MSA表示中与当前MSA表示中的嵌入的二维阵列布置中的嵌入位于同一行的嵌入的注意力来更新嵌入。

[0046] 在一些实施方式中,基于更新的MSA表示来更新当前成对嵌入包括:对更新的MSA表示应用变换操作;以及通过将变换操作的结果加到当前成对嵌入来更新当前成对嵌入。

[0047] 在一些实施方式中,变换操作包括外积均值运算。

[0048] 在一些实施方式中,基于更新的MSA表示来更新当前成对嵌入还包括,在将变换操作的结果加到当前成对嵌入之后:使用对当前成对嵌入的注意力来更新当前成对嵌入,其中,注意力以当前成对嵌入为条件。

[0049] 在一些实施方式中,使用对当前成对嵌入的注意力来更新当前成对嵌入包括:基于当前成对嵌入来生成多个注意力权重;基于当前成对嵌入来生成与每个注意力权重相对应的相应的注意力偏差;基于注意力权重和注意力偏差来生成多个有偏差的注意力权重;以及基于有偏差的注意力权重,使用对当前成对嵌入的注意力来更新当前成对嵌入。

[0050] 在一些实施方式中,基于有偏差的注意力权重使用对当前成对嵌入的注意力来更新当前成对嵌入包括,对于每个当前成对嵌入:基于有偏差的注意力权重,仅使用对与当前成对嵌入的二维阵列布置中的当前成对嵌入位于同一行的当前成对嵌入的注意力来更新当前成对嵌入。

[0051] 在一些实施方式中,基于有偏差的注意力权重使用对当前成对嵌入的注意力来更新当前成对嵌入包括,对于每个当前成对嵌入:基于有偏差的注意力权重,仅使用对与当前成对嵌入的二维阵列布置中的当前成对嵌入位于同一列的当前成对嵌入的注意力来更新当前成对嵌入。

[0052] 在一些实施方式中,蛋白质包括多条链,并且获得表示与蛋白质中的每条链相对应的相应多MSA的初始MSA表示包括:获得与蛋白质中的每条链相对应的MSA的相应表示作为嵌入的二维阵列;以及将与链相对应的MSA的表示组装成块对角线阵列。

[0053] 在一些实施方式中,确定蛋白质的预测结构包括使用折叠神经网络来处理包括最终成对嵌入的输入以生成定义蛋白质的预测结构的输出,包括,对于蛋白质中多对氨基酸中的每对氨基酸:根据折叠神经网络参数值来处理该对氨基酸的最终成对嵌入,以生成指定蛋白质结构中的该对氨基酸之间的可能距离的集合内的概率分布的输出。

[0054] 在一些实施方式中,确定蛋白质的预测结构包括使用折叠神经网络来处理包括最终成对嵌入的输入以生成定义蛋白质的预测结构的输出,包括:获得蛋白质中的每个氨基酸的初始单个嵌入和初始结构参数值,其中,每个氨基酸的结构参数包括指定蛋白质结构中的氨基酸的预测三维空间位置的位置参数;使用折叠神经网络来处理包括最终成对嵌入、蛋白质中的每个氨基酸的初始嵌入和蛋白质中的每个氨基酸的初始结构参数值的折叠神经网络输入,以生成包括蛋白质中的每个氨基酸的最终结构参数值的网络输出,其中,折叠神经网络包括多个更新块,其中,每个更新块包括多个神经网络层,并且被配置为:接收包括

最终成对嵌入、蛋白质中的每个氨基酸的当前单个嵌入和蛋白质中的每个氨基酸的当前结构参数值的更新块输入;以及处理更新块输入以更新蛋白质中的每个氨基酸的当前单个嵌入和当前结构参数值;其中,氨基酸序列中的每个氨基酸的最终结构参数值共同表征蛋白质的预测结构。

[0055] 在一些实施方式中,获得蛋白质中的每个氨基酸的初始单个嵌入包括:基于最终MSA表示来确定蛋白质中的每个氨基酸的初始单个嵌入。

[0056] 在一些实施方式中,每个氨基酸的结构参数还包括指定蛋白质结构中的氨基酸的预测空间朝向的旋转参数。

[0057] 在一些实施方式中,旋转参数定义了 3×3 旋转矩阵。

[0058] 在一些实施方式中,处理更新块输入以更新蛋白质中的每个氨基酸的当前单个嵌入和当前结构参数值包括:更新蛋白质中的每个氨基酸的当前单个嵌入;以及基于蛋白质中的氨基酸的更新的单个嵌入来更新蛋白质中的每个氨基酸的当前结构参数值。

[0059] 在一些实施方式中,多个更新块的最终更新块还被配置为生成定义蛋白质中的每个氨基酸中的每个原子的预测三维空间位置的输出。

[0060] 在一些实施方式中,生成定义蛋白质中的每个氨基酸中的每个原子的相应三维空间位置的输出包括,对于每个氨基酸:处理氨基酸的更新的单个嵌入以生成氨基酸中的原子之间的键的多个扭转角中的每个扭转角的相应值;以及基于多个扭转角的值来确定氨基酸中的每个原子在氨基酸的局部参考系中的空间位置。

[0061] 在一些实施方式中,该方法还包括,对于每种氨基酸,基于以下各项来确定氨基酸中的每个原子在蛋白质的全局参考系中的空间位置:(i)原子在氨基酸的局部参考系中的空间位置,以及(ii)氨基酸的更新的结构参数值。

[0062] 在一些实施方式中,更新蛋白质中的靶氨基酸的当前单个嵌入包括:确定靶氨基酸与蛋白质中的每个源氨基酸之间的相应注意力权重,包括:基于靶氨基酸的当前单个嵌入来生成靶氨基酸的三维查询嵌入;对于蛋白质中的每个源氨基酸,基于源氨基酸的当前单个嵌入来生成源氨基酸的三维关键字嵌入;以及至少部分基于(i)靶氨基酸的三维查询嵌入与(ii)源氨基酸的三维关键字嵌入之间的差值来确定靶氨基酸与蛋白质中的每个源氨基酸之间的注意力权重;以及使用注意力权重来更新靶氨基酸的当前单个嵌入。

[0063] 在一些实施方式中,基于靶氨基酸的当前单个嵌入来生成靶氨基酸的三维查询嵌入包括:使用生成三维输出的线性神经网络层来处理靶氨基酸的当前单个嵌入;以及对线性神经网络层的三维输出应用旋转操作和平移操作,其中旋转操作由靶氨基酸的当前旋转参数值来指定,并且平移操作由靶氨基酸的当前位置参数值来指定。

[0064] 在一些实施方式中,基于源氨基酸的当前单个嵌入来生成源氨基酸的三维关键字嵌入包括:使用生成三维输出的线性神经网络层来处理源氨基酸的当前单个嵌入;以及对线性神经网络层的三维输出应用旋转操作和平移操作,其中,旋转操作由源氨基酸的当前旋转参数值来指定,并且平移操作由源氨基酸的当前位置参数值来指定。

[0065] 在一些实施方式中,确定靶氨基酸与蛋白质中的每个源氨基酸之间的注意力权重还包括,对于蛋白质中的每个源氨基酸:确定与包括靶氨基酸和源氨基酸的一对氨基酸相对应的最终成对嵌入的投影;以及至少部分基于与包括靶氨基酸和源氨基酸的该对氨基酸相对应的最终成对嵌入的投影来确定靶氨基酸与源氨基酸之间的注意力权重。

[0066] 在一些实施方式中,使用注意力权重来更新靶氨基酸的当前单个嵌入包括:对于蛋白质中的每个氨基酸,基于氨基酸的当前单个嵌入来生成氨基酸的三维值嵌入;使用注意力权重来确定氨基酸的三维值嵌入的加权线性组合;通过对加权线性组合应用旋转操作和平移操作来生成几何回报嵌入,其中,旋转操作反转由靶氨基酸的当前旋转参数值所指定的旋转操作,并且平移操作由靶氨基酸的当前位置参数值的负值来指定;以及基于几何回报嵌入来更新靶氨基酸的当前嵌入。

[0067] 在一些实施方式中,基于蛋白质中的氨基酸的更新的单个嵌入来更新蛋白质中的每个氨基酸的当前结构参数值包括,对于每个氨基酸:将氨基酸的更新的位置参数值确定为以下各项之和:(i)氨基酸的当前位置参数值,以及(ii)氨基酸的更新的单个嵌入的线性投影;将氨基酸的更新的旋转参数值确定为以下各项的组合:(i)由氨基酸的当前旋转参数值所指定的旋转操作,以及(ii)由具有实部和虚部的四元数所指定的旋转操作,其中,实部为1,虚部由氨基酸的更新的单个嵌入的线性投影来指定。

[0068] 在一些实施方式中,该方法还包括,对于蛋白质中的每个氨基酸:处理由最终更新块生成的氨基酸的更新的当前单个嵌入,以生成针对蛋白质的预测结构中的氨基酸中的原子的定位的置信度估计,其中,置信度估计定义了原子的定位的可能准确度范围内的概率分布。

[0069] 本文所述的蛋白质结构预测系统可以用于获得配体,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,特别是靶蛋白质(可以是多聚体蛋白质)的氨基酸序列,例如药物靶,以及使用蛋白质结构预测系统来处理基于靶氨基酸序列的输入,以确定靶蛋白质的(三级和/或四级)结构,即预测蛋白质结构。该方法可以包括评估一个或多个候选配体与靶蛋白质结构的相互作用。该方法还可以包括取决于对相互作用的评估的结果来选择候选配体中的一个或多个作为配体。该方法还可以包括合成所选择的配体。

[0070] 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质结构的键合。例如,评估相互作用可以包括识别利用对生物效应足够的亲和力进行键合的配体。在一些其他实施方式中,评估相互作用可以包括评估候选配体与对靶蛋白质(例如,酶)的功能有效效应的靶蛋白质结构的关联。评估可以包括评估候选配体与靶蛋白质结构之间的亲和力,或者评估相互作用的选择性。候选配体可以根据哪些具有最高亲和力来选择。

[0071] 候选配体可以从候选配体的数据库中导出,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)来导出,和/或可以通过对候选配体的逐步或迭代的组装/优化来导出。

[0072] 对候选配体与靶蛋白质结构的相互作用的评估可以使用计算机辅助方法来执行,其中,显示候选配体和靶蛋白质结构的图形模型以供用户操纵,和/或评估可以(例如,使用标准分子(蛋白质-配体)对接软件)部分或完全自动执行。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分数,其中,相互作用分数包括候选配体与靶蛋白质之间的相互作用的度量。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,取决于键合自由能的分数。候选配体可以取决于其分数来选择,例如,具有最高相互作用分数的候选配体。

[0073] 在一些实施方式中,靶蛋白质包括受体或酶,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,该方法可以用于识别细胞表面标志物的结构。然后,这可以用于识别配体,例如,与细胞表面标记物键合的抗体或标记(诸如荧光标记)。这可以用于识别和/或治疗癌细胞。

[0074] 在一些实施方式中,配体是药物,确定多个靶蛋白质中的每个靶蛋白质的预测结构,并且评估一个或多个候选配体与每个靶蛋白质的预测结构的相互作用。然后,可以选择候选配体中的一个或多个,以或者获得与每个靶蛋白质(在功能上)相互作用的配体,或者获得仅与一个靶蛋白质(在功能上)相互作用的配体。例如,在一些实施方式中,可能希望获得针对多个药物靶有效的药物。此外或相反,可能希望筛选药物的脱靶效应。例如,在农业中,确定为用于一个植物物种而设计的药物不与另一不同的植物物种和/或动物物种相互作用可能是有用的。

[0075] 在一些实施方式中,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。

[0076] 在一些情况下,蛋白质结构预测系统可以用于确定候选多肽配体(在一些情况下可以是多聚体蛋白质)的结构,例如,药物或工业酶的配体的结构。然后,可以评估其与靶蛋白质结构(也可以是多聚体)的相互作用;可能已经使用结构预测神经网络或使用常规物理调研技术(诸如x射线晶体学和/或磁共振技术或低温电子显微镜)来确定了靶蛋白质结构。

[0077] 另一方面,提供了一种使用蛋白质结构预测系统来获得多肽配体(例如,分子或其序列)的方法。该方法可以包括获得一个或多个候选多肽配体的氨基酸序列(候选多肽配体可以是多聚体,在这种情况下,可以获得一个以上的氨基酸序列,例如,多聚体中的每个氨基酸链的氨基酸序列)。该方法还可以包括使用蛋白质结构预测系统来确定候选多肽配体的(三级和/或四级)结构。该方法还可以包括:在计算机上和/或通过物理调研获得靶蛋白质的靶蛋白质结构,以及评估一个或多个候选多肽配体中的每个候选多肽配体的结构与靶蛋白质结构之间的相互作用。该方法还可以包括取决于评估的结果来选择候选多肽配体中的一个或多个作为多肽配体。该方法还可以包括合成所选择的多肽配体。

[0078] 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质结构的键合,例如,识别利用对生物效应足够的亲和力进行键合的配体,和/或评估候选多肽配体与对靶蛋白质(例如,酶)的功能有效应的靶蛋白质结构的关联,和/或评估候选多肽配体与靶蛋白质结构之间的亲和力,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,例如,多聚体适体。同样,多肽候选配体可以根据那些具有最高亲和力来选择。

[0079] 如前所述,所选择的多肽配体可以包括受体或酶,并且配体可以是受体或酶的激动剂或拮抗剂。在一些实施方式中,多肽配体可以包括抗体(可以是多聚体抗体),并且靶蛋白质包括抗体靶(也可以是多聚体),例如,病毒(特别是病毒外壳蛋白质)或在癌细胞上表达的蛋白质。在这些实施方式中,抗体与抗体靶键合以提供治疗效果。例如,抗体可以与靶键合,并且充当特定受体的激动剂;替代地,抗体可以防止另一配体与靶的键合,从而防止相关生物通路的激活。

[0080] 该方法的实施方式还可以包括合成(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来合成和/或已经可以是可用的,例如可以来自化合物库或可以使用组合化学来合成。

[0081] 该方法还可以包括在体外和/或体内测试配体的生物活性。例如,可以测试配体的ADME(absorption distribution metabolism excretion,吸收、分布、代谢、排泄)和/或毒理学属性,以筛选出不合适的配体。测试可以包括,例如,使候选小分子或多肽配体与靶蛋白质接触,并且度量蛋白质的表达或活性的改变。

[0082] 在一些实施方式中,候选(多肽)配体可以包括:隔离抗体、隔离抗体的片段、单可变域抗体、双特异性或多特异性抗体、多价抗体、双可变域抗体、免疫缀合物、纤连蛋白质分子、附着蛋白质(adnectin)、DARPin、亲和多聚体(avimer)、亲和体(affibody)、抗运载蛋白质(anticalin)、亲和蛋白质(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区域的抗体,例如,这种Fc区域与野生型Fc区域相比防止或降低了ADCC(antibody-dependent cellular cytotoxicity,抗体依赖性细胞毒性)活性和/或增加了半衰期。候选(多肽)配体可以包括具有不同CDR(Complementarity-Determining Region,互补确定区域)的抗体。在任何这些实施方式中,候选(多肽)配体可以是例如多聚体。

[0083] 本文所述的蛋白质结构预测系统也可以用于获得疾病的诊断抗体标记物。还提供了一种对于例如如上所述的一个或多个候选抗体中的每个候选抗体使用蛋白质结构预测系统来确定候选抗体的预测结构的方法。该方法还可以涉及:获得靶蛋白质的靶蛋白质结构,评估一个或多个候选抗体中的每个候选抗体的预测结构与靶蛋白质结构之间的相互作用,以及取决于评估的结果来选择候选抗体中的一个或多个中的一个作为诊断抗体标记物,例如,选择对靶蛋白质结构具有最高亲和力的一个或多个候选抗体。该方法可以包括制备诊断抗体标记物。诊断抗体标记物可以用于通过检测其是否与从患者获得的样品(例如,体液的样品)中的靶蛋白质键合来诊断疾病。如上所述,对应的技术可以用于获得治疗抗体(多肽配体)。

[0084] 错误折叠的蛋白质与多种疾病相关联。因此,在另一方面,提供了一种使用蛋白质结构预测系统来识别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(或者在多聚体蛋白质的情况下获得多个氨基酸序列),以及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(例如,通过常规(物理)方法)获得从人体或动物体获得的蛋白质版本的结构。该方法然后可以包括:将该蛋白质的结构与从身体中获得的版本的结构进行比较,以及取决于比较的结果来识别蛋白质错误折叠疾病的存在。也就是说,来自身体的蛋白质版本的错误折叠可以通过与在计算机上确定的结构进行比较来确定。

[0085] 本文所述的蛋白质结构预测系统还可以用于获得与配体(诸如药物或工业酶的配体)相互作用的蛋白质。例如,获得蛋白质的方法可以包括:获得多个靶蛋白质(可以是多聚体蛋白质)的相应的一个或多个氨基酸序列,以及使用蛋白质结构预测系统基于一个或多个氨基酸序列中的每个氨基酸序列来处理相应输入,以确定对应的靶蛋白质的相应结构。所选择的蛋白质可以包括受体或酶,并且其中,配体是受体或酶的激动剂或拮抗剂。该方法然后可以包括评估配体与每个候选蛋白质的预测结构的相互作用。该方法然后可以包括取决于评估的结果来选择候选蛋白质中的一个或多个。该方法还可以包括:合成所选择的蛋白质,以及可选地在体外和体内测试蛋白质的生物活性。

[0086] 在一些实施方式中,评估配体与候选蛋白质的预测结构的相互作用可以包括确定

候选蛋白质的相互作用分数,其中相互作用分数包括候选蛋白质与配体之间的相互作用的度量。

[0087] 一般而言,识别蛋白质错误折叠疾病的存在可以涉及:获得蛋白质的氨基酸序列(或者在多聚体蛋白质的情况下获得多个氨基酸序列),使用蛋白质的氨基酸序列(或多个氨基酸序列)来如本文所述确定蛋白质的结构,以及将蛋白质的结构与蛋白质的基线版本的结构进行比较,取决于比较的结果来识别蛋白质错误折叠疾病的存在。例如,被比较的结构可以是突变体和野生型蛋白质的结构。在实施方式中,野生型蛋白质可以用作基线版本,但是原则上也可以不用作基线版本。

[0088] 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以用于从其氨基酸序列中识别靶蛋白质上的活性/键合/阻断位点。

[0089] 可以实施本说明书中描述的主题的特定实施例,以便实现以下优点中的一个或多个。

[0090] 本说明书描述了一种可以预测蛋白质结构(特别是“多聚体”(即,具有多条氨基酸链的蛋白质)的结构)的系统。该系统使用被称为蛋白质结构预测神经网络的神经网络来预测蛋白质结构。该系统训练蛋白质结构预测神经网络以生成与“靶”(例如,基准真值)蛋白质结构匹配的预测蛋白质结构。特别地,该系统训练蛋白质结构预测神经网络以对度量蛋白质的预测结构(即,由蛋白质结构预测神经网络生成)与靶蛋白质结构之间的误差的目标函数进行优化。一般来说,蛋白质的预测结构包括定义三级结构(即,蛋白质中的每条氨基酸链的3-D形状)和四级结构(即,氨基酸链相对于彼此的布置)的信息。

[0091] 在一些情况下,蛋白质可以包括多条具有完全相同的氨基酸序列的“重复”链,并且预测蛋白质结构与靶蛋白质结构的比较可能是模糊的。特别地,预测蛋白质结构中的每条重复链可以被映射到靶蛋白质结构中共享完全相同的氨基酸序列的多条链上。在这些情况下,为了准确地比较预测蛋白质结构与基准真值蛋白质,该系统高效地确定预测蛋白质结构中的链到靶蛋白质结构中的对应的链的分配,以最小化预测蛋白质结构与靶蛋白质结构之间的误差。因此,该系统提高了预测蛋白质结构与靶蛋白质结构的比较的准确度,这能够实现对于蛋白质结构预测神经网络的更高效训练,从而减少训练期间的计算资源(例如,内存和计算能力)的消耗。

[0092] 处理大型蛋白质(特别是复杂的多聚体)可能是计算密集型的,并且计算资源的消耗随着蛋白质中的氨基酸总数的增加而迅速增加。为了解决这个问题,本说明书中描述的系统可以在多聚体蛋白质的仅包括整个蛋白质的一小部分的“裁剪”上训练蛋白质结构预测神经网络。系统可以通过从蛋白质中的每条链裁剪氨基酸序列来生成用于训练蛋白质结构预测神经网络的蛋白质裁剪,从而实现链覆盖和裁剪多样性。在表现出更大的链覆盖和裁剪多样性的蛋白质裁剪上训练蛋白质结构预测神经网络可以加速对蛋白质结构预测神经网络的训练,从而减少训练期间的计算资源的消耗。如本文所使用的,“裁剪”氨基酸序列是指选择氨基酸序列的一部分以供包括在蛋白质裁剪中。

[0093] 蛋白质的结构确定蛋白质的生物功能。因此,确定蛋白质结构可以促进理解生命过程(例如,包括许多疾病的机制)和设计蛋白质(例如,作为药物或作为工业过程的酶)。例如,哪些分子(例如,药物)将与蛋白质键合(以及键合将在哪里发生)取决于蛋白质的结构。由于药物与蛋白质(例如,血液中的蛋白质)的键合程度能够影响药物的有效性,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)确定蛋白质结构可能既耗时又非常昂贵。因此,本说明书中描述的蛋白质预测系统可以促进生物化学研究和涉及蛋白质的工程领域(例如,药物开发)。

[0094] 在附图和以下描述中阐述了本说明书的主题的一个或多个实施例的细节。根据说明书、附图和权利要求书,本主题的其他特征、方面和优点将变得清晰。

附图说明

[0095] 图1示出了蛋白质结构预测神经网络和训练系统的示例。

[0096] 图2示出了示例训练系统。

[0097] 图3A图示了使用聚焦多样性的裁剪方法来裁剪多聚体蛋白质的示例。

[0098] 图3B图示了使用聚焦界面的裁剪方法来裁剪多聚体蛋白质的示例。

[0099] 图4图示了蛋白质的预测结构与靶结构之间的链分配。

[0100] 图5是通过迭代多聚体蛋白质的氨基酸链并且从每条氨基酸链中选择连续的裁剪直到满足终止标准来生成裁剪出的蛋白质的示例过程的流程图。

[0101] 图6是用于确定具有重复的氨基酸链的多聚体蛋白质的预测蛋白质结构与靶蛋白质结构之间的链分配的示例过程的流程图。

[0102] 图7是用于确定从预测蛋白质结构到靶蛋白质结构的链分配的示例过程的流程图,该链分配将预测蛋白质结构中的预测锚链分配给靶蛋白质结构中的靶锚链。

[0103] 图8示出了证明根据本说明书中描述的技术而训练的蛋白质结构神经网络的优异预测准确度的实验结果。

[0104] 图9示出了示例蛋白质结构预测系统。

[0105] 图10示出了嵌入神经网络的示例架构。

[0106] 图11示出了嵌入神经网络的更新块的示例架构。

[0107] 图12示出了MSA更新块的示例架构。

[0108] 图13示出了一对更新块的示例架构。

[0109] 图14示出了折叠神经网络的示例架构。

[0110] 图15图示了氨基酸中的键之间的扭转角。

[0111] 图16是未折叠蛋白质和折叠蛋白质的图示。

[0112] 图17是用于预测蛋白质结构的示例过程的流程图。

[0113] 图18示出了为蛋白质中的氨基酸链生成MSA表示的示例过程。

[0114] 图19示出了为蛋白质中的每对氨基酸生成相应成对嵌入的示例过程。

[0115] 不同附图中相同的附图标记和名称指示相同的元素。

具体实施方式

[0116] 图1示出了蛋白质结构预测神经网络126和训练系统122的示例。

[0117] 蛋白质结构预测神经网络126被配置为处理表征蛋白质的网络输入124,以生成定义蛋白质的预测结构128的数据。下面参考图9-图19更详细地描述了蛋白质结构预测神经网络126的示例。(特别地,参考图9-图19描述的嵌入神经网络200和折叠神经网络600共同定义了蛋白质结构预测神经网络126)。

[0118] 训练系统122被配置为在训练示例132的集合上训练蛋白质结构预测神经网络126的神经网络参数集合。每个训练示例与蛋白质相对应,并且包括:(i) 表征蛋白质的网络输入,以及(ii) 蛋白质的靶结构。对于每个训练示例,训练系统122训练蛋白质结构预测神经网络126以处理训练示例的网络输入以生成与由训练示例指定的靶结构匹配的预测蛋白质结构。因此,对于每个训练示例,包括在训练示例中的靶结构定义了应当由蛋白质结构预测神经网络通过处理包括在训练示例中的网络输入而生成的蛋白质结构。蛋白质的靶结构可以以各种适当方式中的任何一种(例如,通过物理实验,例如,x射线晶体学实验)来获得。

[0119] 由训练系统122用于训练蛋白质结构预测神经网络126的至少一些训练示例与多聚体蛋白质120相对应,即,包括多条氨基酸链的蛋白质。多聚体蛋白质120可以包括任何适当数量的氨基酸链,例如,3条氨基酸链、10条氨基酸链、100条氨基酸链或1000条氨基酸链。多聚体蛋白质120中的每条氨基酸链可以包括任何适当数量的氨基酸,例如,10个氨基酸、100个氨基酸或1000个氨基酸。

[0120] 在一些实施方式中,训练系统122仅在与多聚体蛋白质相对应的训练示例上训练蛋白质结构预测神经网络126。在其他实施方式中,训练系统122在与单链蛋白质和多聚体蛋白质的混合物相对应的训练示例上训练蛋白质结构预测神经网络。

[0121] 通过在与多聚体蛋白质120相对应的训练示例上训练蛋白质结构预测神经网络126,训练系统122可以调整结构预测神经网络126的神经网络参数集合,以准确稳健地预测复杂的多聚体的结构。特别地,与例如通过仅在与单链蛋白质相对应的训练示例上训练蛋白质结构预测相比,训练系统122可以使蛋白质结构预测神经网络126能够在预测多聚体蛋白质结构时实现更好的性能。训练系统122被设计和实现为克服在多聚体训练蛋白质上训练结构预测神经网络的上下文中出现的多种技术挑战,这将在下文中更详细地描述。

[0122] 下面参考图2-图8更详细地描述了训练系统122的示例。本说明书中描述的训练系统122是普遍适用的,特别是可以用于训练任何适当的蛋白质结构预测神经网络,并且不限于训练具有参考图9-图19描述的架构的蛋白质结构预测神经网络。

[0123] 图2示出了示例训练系统122。训练系统122是被实现为一个或多个位置中的一个或多个计算机上的计算机程序的系统的示例,其中,实现了下面描述的系统、组件和技术。

[0124] 训练系统122被配置为训练蛋白质结构预测神经网络136(例如,参考图9-图19描述的蛋白质结构预测神经网络)。训练系统122包括裁剪引擎129、排列对齐引擎140和训练引擎144,接下来分别更详细地描述它们。在一些实施方式中,可以省略裁剪引擎或排列对齐引擎。

[0125] 裁剪引擎129被配置为处理多聚体蛋白质120的集合以生成裁剪出的蛋白质130的集合,其中,每个裁剪出的蛋白质130是相应的多聚体蛋白质120的裁剪版本。更具体地,每个裁剪出的蛋白质130与相应的多聚体蛋白质120相对应,并且仅包括多聚体蛋白质120中包括的氨基酸的真子集。也就是说,裁剪引擎129通过从包括在对应的多聚体蛋白质120中的一条或多条氨基酸链中裁剪一个或多个氨基酸来生成每个裁剪出的蛋白质130。(通常,裁剪出的蛋白质130本身是多聚体蛋白质,即,每个裁剪出的蛋白质仍包括多条氨基酸链)。

[0126] 裁剪引擎129可以以各种可能方式中的任何一种来处理多聚体蛋白质120以生成对应的裁剪出的蛋白质130。接下来描述用于处理多聚体蛋白质120以生成裁剪出的蛋白质130的几种示例技术。

[0127] 在一些实施方式中,为了从多聚体蛋白质120中生成裁剪出的蛋白质130,训练引擎144迭代多聚体蛋白质120的链,并且选择要从多聚体蛋白质120的每条链中裁剪的连续的氨基酸序列(即,选择以供包括在裁剪出的蛋白质130中)。裁剪引擎129可以随机采样要从多聚体蛋白质120的每条链中裁剪的氨基酸序列的长度。裁剪引擎129可以迭代多聚体蛋白质120的链,直到满足终止标准,例如,直到来自多聚体蛋白质120的预定数量的氨基酸已经被指定以供包括在裁剪出的蛋白质130中。参考图5更详细地描述了通过迭代多聚体蛋白质120的氨基酸链并且从每条链中选择连续的裁剪直到满足终止标准来生成裁剪出的蛋白质130的示例过程。

[0128] 在一些实施方式中,为了从多聚体蛋白质120中生成裁剪出的蛋白质130,训练引擎144可以从多聚体蛋白质120中采样“界面”氨基酸。界面氨基酸可以指多聚体蛋白质120中在来自多聚体蛋白质120的不同链的氨基酸的预定阈值距离内的氨基酸。第一氨基酸与第二氨基酸之间的“距离”可以指第一氨基酸中的指定原子与第二氨基酸中的指定原子之间的三维空间距离。指定原子可以是例如 α 碳原子。阈值距离可以是例如10埃、15埃、20埃或任何其他适当的阈值距离。

[0129] 在采样界面氨基酸之后,裁剪引擎129可以为多聚体蛋白质120中的每个氨基酸确定氨基酸与所采样的界面氨基酸之间的距离。裁剪引擎129然后可以基于多聚体蛋白质120中的氨基酸与所采样的界面氨基酸的距离来指定多聚体蛋白质120中的氨基酸的真子集以供包括在裁剪出的蛋白质130中。例如,裁剪引擎129可以指定多聚体蛋白质120中与所采样的界面氨基酸具有最低距离的预定数量的氨基酸以供包括在裁剪出的蛋白质130中。作为另一示例,裁剪引擎129可以将多聚体蛋白质120中距所采样的界面氨基酸在预定阈值距离内的每个氨基酸指定为包括在裁剪出的蛋白质130中。(在一些实施方式中,裁剪引擎129可以将小的随机值加到计算出的多聚体蛋白质中的氨基酸之间的距离,以降低“平局”(例如,其中多个氨基酸与所采样的界面氨基酸等距)的可能性)。

[0130] 训练系统122生成与每个裁剪出的蛋白质130相对应的相应训练示例132,然后在训练示例132上训练蛋白质结构预测神经网络126,这将在下面更详细地描述。蛋白质结构预测神经网络126预测蛋白质结构所消耗的内存和计算资源的量随着包括在蛋白质中的氨基酸的数量而迅速增加。因此,可以被包括在用于训练结构预测神经网络136的蛋白质中的氨基酸的数量可能受到内存和计算考虑的限制。某些训练蛋白质(特别是复杂的多聚体)可以包括大量氨基酸,因此可能不适合用于训练蛋白质结构预测神经网络126。裁剪引擎129使多聚体蛋白质120能够被裁剪成包括较少氨基酸并且可以轻易用于训练蛋白质结构预测神经网络126的蛋白质。

[0131] 裁剪引擎129可以使用被设计成加速对蛋白质结构预测神经网络126的训练并且提高由蛋白质结构预测神经网络126在训练之后生成的蛋白质结构预测的准确度的技术来裁剪多聚体蛋白质120。例如,裁剪引擎129可以使用“聚焦多样性”的裁剪方法(例如,通过如上所述迭代多聚体蛋白质120的氨基酸链并且从每条氨基酸链中选择具有随机长度的连续的裁剪)来增加链覆盖和裁剪多样性的可能性。作为另一示例,裁剪引擎129可以使用“聚焦界面”的裁剪方法(例如,通过裁剪多聚体蛋白质的以界面氨基酸为中心的区域)来裁剪多聚体蛋白质120的部分,从而促使蛋白质结构预测神经网络126学习管控氨基酸链之间的界面的复杂的化学和物理动力学。作为另一示例,裁剪引擎129可以使用各种裁剪方法(例

如,使用聚焦多样性的裁剪方法和聚焦界面的裁剪方法两者)来生成用于训练蛋白质结构预测神经网络126的裁剪出的蛋白质。

[0132] 在一些实施方式中,裁剪引擎129从单个多聚体蛋白质120中生成多个裁剪出的蛋白质130。例如,裁剪引擎129可以重复地从多聚体蛋白质120中采样界面残基,并且(例如,通过如上所述裁剪多聚体蛋白质的以所采样的界面残基为中心的区域)生成与每个所采样的界面残基相对应的相应的裁剪出的蛋白质130。

[0133] 训练系统122可以生成与每个裁剪出的蛋白质130相对应的相应训练示例132,以及生成与一些或所有多聚体蛋白质120相对应的训练示例132。例如,如果多聚体蛋白质120中的氨基酸的数量超过预定阈值(例如,如上所述基于内存和计算考虑),训练系统122可以可选地抑制生成与多聚体蛋白质120相对应的训练示例132。

[0134] 每个训练示例与“训练”蛋白质(例如,裁剪出的蛋白质130或多聚体蛋白质120)相对应,并且包括:(i)蛋白质结构预测神经网络126的网络输入,其表征训练蛋白质,以及(ii)训练蛋白质的靶结构。

[0135] 蛋白质结构预测神经网络126的网络输入可以具有任何适当的格式。例如,参考图9-图19描述的示例蛋白质结构预测神经网络被配置为接收包括MSA表示和成对嵌入集合的网络输入。下面(例如,参考图9)更详细地描述了用于生成包括MSA表示和成对嵌入集合的网络输入的示例技术。

[0136] 训练蛋白质的靶结构是指应当由蛋白质结构预测神经网络通过处理表征训练蛋白质的网络输入而生成的蛋白质结构。用于训练蛋白质的靶结构可以由结构参数集合来表示。例如,定义训练蛋白质的靶结构的结构参数集合可以包括训练蛋白质中的每个氨基酸的(i)位置参数和(ii)旋转参数,如上所述。作为另一示例,定义训练蛋白质的靶结构的结构参数集合可以包括定义训练蛋白质的氨基酸中的键之间的扭转角的扭转参数集合。作为另一示例,定义训练蛋白质的靶结构的结构参数集合可以包括:(i)训练蛋白质中的氨基酸的位置参数和旋转参数,以及(ii)定义训练蛋白质的氨基酸中的键之间的扭转角的扭转参数。(例如,通过裁剪对应的多聚体蛋白质120而生成的)裁剪出的蛋白质130的靶结构可以由定义多聚体蛋白质120的靶结构的结构参数集合的真子集来定义(集合的真子集是不包括该集合的至少一个成员的子集)。

[0137] 在实施方式中,训练示例132的集合包括至少一些与多聚体蛋白质相对应的训练示例,如上所述。训练系统122可以在训练示例132的集合上训练蛋白质结构预测神经网络126,从而使得蛋白质结构预测神经网络126学习准确地预测多聚体蛋白质的结构。

[0138] 更具体地,训练系统122在训练示例132的集合上训练蛋白质结构预测神经网络126以对包括结构损失和可选的一个或多个辅助损失的目标函数进行优化。对于每个训练示例,结构损失可以表征(i)由蛋白质结构预测神经网络通过处理由训练示例指定的网络输入134而生成的预测蛋白质结构138与(ii)由训练示例指定的靶蛋白质结构142之间的相似性。

[0139] 评估训练示例的目标函数(特别地,例如,结构损失)需要比较预测蛋白质结构138和靶蛋白质结构142。然而,当训练蛋白质包括具有完全相同的氨基酸序列的多条重复链时,那么预测蛋白质结构138与靶蛋白质结构142的比较可能是模糊的。特别地,预测蛋白质结构中的每条重复链可以被映射到靶蛋白质结构中共享完全相同的氨基酸序列的多条链

上。

[0140] 为了解决这个问题,如果训练蛋白质包括重复的氨基酸链(即,具有相同的氨基酸序列的多条链),则排列对齐引擎140确定预测蛋白质结构中的每条氨基酸链到靶蛋白质结构中的对应的氨基酸链的一对一分配。通常,排列对齐引擎140被配置为确定预测靶蛋白质结构与靶蛋白质结构之间的氨基酸链分配,该氨基酸链分配近似地或精确地对预测靶蛋白质结构与靶蛋白质结构之间的相似性进行优化。排列对齐引擎140可以以各种可能方式中的任何一种来确定预测蛋白质结构与靶蛋白质结构之间的氨基酸链分配。接下来描述链对齐技术的几个示例。

[0141] 在一些实施方式中,排列对齐引擎140被配置为评估预测蛋白质结构与靶蛋白质结构之间的每一种可能的氨基酸链分配的结构损失(或可选的整个目标函数)。(如果预测蛋白质结构中的每条氨基酸链被分配给靶蛋白质结构中具有相同的氨基酸序列的唯一链,则对齐被称为“可能的”)。然而,将会理解,评估所有可能的链分配的复杂度随着氨基酸链数量而组合地增长,并且可能需要大量计算资源,特别是对于训练具有许多重复链的蛋白质而言。

[0142] 在一些实施方式中,排列对齐引擎140实现操作以高效地将预测蛋白质结构138中的链分配给靶蛋白质结构142中的对应的链,而无需评估所有可能的链分配。参考图6-图7更详细地描述了用于高效地生成链分配的示例过程。

[0143] 包括在目标函数中的结构损失基于预测结构138中的链到对应的靶结构142的分配来度量:(i) 预测结构138与(ii) 靶结构142之间的误差。结构损失可以以各种可能方式中的任何一种来度量预测结构138与靶结构142之间的误差。接下来描述结构损失的几个示例。

[0144] 在一些实施方式中,为了评估结构损失,训练系统122为靶结构142中的每条链确定靶结构链中的每个氨基酸与被分配给靶结构链的预测结构链中的对应的氨基酸之间的相应的定位误差。(定位误差可以是例如 L_1 误差、 L_2 误差或任何其他适当的误差)。训练系统122可以通过对预测蛋白质结构138和靶蛋白质结构142中的对应的链中的氨基酸之间的定位误差进行组合(例如,求和或求平均)来评估结构损失。

[0145] 在一些实施方式中,预测蛋白质结构和靶蛋白质结构都由包括以下各项的结构参数来定义:(i) 蛋白质中的每个氨基酸的位置参数和旋转参数,以及(ii) 蛋白质中的氨基酸中的键之间的扭转角。在这些实施方式中,结构损失 $\mathcal{L}_{structure}$ 可以由下式给出:

$$[0146] \quad \mathcal{L}_{structure} = \frac{1}{Z} \text{mean}_{ij} \left(\text{minimum}(d_{\text{clamp}}, d_{ij}) \right)$$

$$[0147] \quad d_{ij} = \sqrt{\|x_{ij} - x_{ij}^{\text{true}}\|^2 + \epsilon}$$

$$[0148] \quad x_{ij} = R_i^{-1}(x_j - t_i)$$

$$[0149] \quad x_{ij}^{\text{true}} = R_i^{\text{true}^{-1}}(x_j^{\text{true}} - t_i^{\text{true}})$$

[0150] 其中,i索引蛋白质中的氨基酸,j索引蛋白质中的原子, d_{clamp} 是箝位阈值,Z是归一化常数, ϵ 是小的正常数, $t_i \in \mathbb{R}^3$ 表示氨基酸i的预测位置参数, $R_i \in \mathbb{R}^{3 \times 3}$ 表示氨基酸i

的预测旋转参数, $t_i^{\text{true}} \in \mathbb{R}^3$ 表示氨基酸i的靶旋转参数, $R^{\text{true}} \in \mathbb{R}^{3 \times 3}$ 表示氨基酸i的靶旋转参数, $x_j \in \mathbb{R}^3$ 表示原子j的预测定位, 并且 $x_j^{\text{true}} \in \mathbb{R}^3$ 表示原子j的靶定位。每个氨基酸中的原子(包括侧链原子)的预测定位可以从氨基酸的位置参数、旋转参数和扭转角中导出, 如下面参考图14更详细描述。

[0151] 可选地, 当氨基酸i和原子j被包括在蛋白质中的同一氨基酸链中时, 箝位阈值 d_{clamp} 可以采用 d_{ij} 的第一值, 并且当氨基酸i和原子j被包括在蛋白质中的不同氨基酸链中时, 箝位阈值 d_{clamp} 可以采用 d_{ij} 的第二(不同的)值。第二值可以高于第一值, 这能够导致蛋白质结构预测神经网络因为预测界面不正确而受到比预测链内结构不正确更强烈的惩罚。在一个示例中, 第一箝位值可以是10埃, 并且第二箝位值可以是30埃。

[0152] 在一些实施方式中, 目标函数包括“重叠”辅助损失, “重叠”辅助损失惩罚蛋白质的预测结构中的氨基酸链之间的间距的误差, 例如, 惩罚具有重叠链的预测结构。例如, 重叠辅助损耗可以由下式给出:

$$[0153] \quad \mathcal{L}_{\text{overlap}} = \frac{1}{N(N-1)} \sum_{i,j} \left[\min \left(\frac{1}{20} (\|c_i^{\text{pred}} - c_j^{\text{pred}}\| - \|c_i^{\text{gt}} - c_j^{\text{gt}}\| + 4), 0 \right)^2 \right]$$

[0154] 其中, N是蛋白质中的氨基酸链数量, $i, j \in \{1, \dots, N\}$, c_i^{pred} 是预测链i的 α 碳原子的质心, c_i^{gt} 是基准真值(靶)链i的 α 碳原子的质心。值 $\frac{1}{20}$ 和4是可以通过超参数扫描而选择的超参数。

[0155] 训练引擎144评估每个训练示例的目标函数, 并且(例如, 通过反向传播)确定目标函数相对于蛋白质结构预测神经网络126的神经网络参数集合的梯度。训练引擎144使用梯度(例如, 根据适当的梯度下降优化算法(例如, RMSprop或Adam)的更新规则)来调节神经网络参数集合的值。

[0156] 训练系统122可以在训练示例132的集合上训练蛋白质结构预测神经网络126, 直到满足适当的训练终止标准, 例如, 直到蛋白质结构预测神经网络126达到阈值预测准确度。

[0157] 图3A图示了使用聚焦多样性的裁剪方法来裁剪多聚体蛋白质的示例, 例如, 该裁剪方法迭代多聚体蛋白质的氨基酸链, 并且从每条氨基酸链中选择具有随机长度的连续的裁剪以供包括在裁剪出的蛋白质中。更具体地, 在图3A中, 多聚体蛋白质包括被标记为“链#1”、……、“链#N”的N条氨基酸链。多聚体蛋白质中的每条链的深色阴影部分指示多聚体蛋白质的氨基酸链的被指定以供包括在裁剪出的蛋白质中的连续部分。将会理解, 裁剪出的蛋白质比原始的多聚体蛋白质包括显著更少的氨基酸。参考图5更详细地描述了用于实现聚焦多样性的蛋白质裁剪的示例过程。

[0158] 图3B图示了使用聚焦界面的裁剪方法来裁剪多聚体蛋白质的示例, 例如, 该裁剪方法裁剪多聚体蛋白质的以从多聚体蛋白质中采样的界面氨基酸为中心的区域。例如, 裁剪出的蛋白质可以包括多聚体蛋白质的被包括在由图3B中的圆圈表示的裁剪出的区域中的所有氨基酸。参考图2更详细地描述了用于实现聚焦界面的裁剪方法的示例过程。

[0159] 图4图示了蛋白质的预测结构与靶结构之间的链分配的示例。在该示例中, 蛋白质

包括五条氨基酸链,包括三条具有第一氨基酸序列的氨基酸链和两条具有第二(不同的)氨基酸序列的链。需要链分配(即,来自预测结构的每条链到靶结构中的对应的链的一对一分配)以便评估(例如,比较预测结构和靶结构的)结构损失,作为训练结构预测神经网络的一部分。参考图6-图7更详细地描述了用于确定链分配的示例过程。

[0160] 图5是用于通过迭代多聚体蛋白质的氨基酸链并且从每条氨基酸链中选择连续的裁剪直到满足终止标准来生成裁剪出的蛋白质的示例过程146的流程图。为了方便起见,过程146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,根据本说明书适当地编程的训练系统(例如,图2的训练系统122)可以执行过程146。

[0161] 在实施方式中,系统接收多聚体蛋白质和裁剪预算(148)。裁剪预算定义了要包括在将通过裁剪多聚体蛋白质而生成的裁剪出的蛋白质中的氨基酸的最大数量。多聚体蛋白质中的氨基酸链被理解为与(任意)排序相关联,例如,从1到N的排序,其中,N是多聚体蛋白质中的链数量。

[0162] 系统从多聚体蛋白质的第一条链开始对多聚体蛋白质的每条链执行步骤150-156,直到满足终止标准。为了方便起见,步骤150-156将被描述为对多聚体蛋白质中的“当前”链执行。

[0163] 系统确定要从当前链中裁剪的氨基酸序列的最小长度(150)。系统可以确定要从当前链中裁剪的氨基酸序列的最小长度(即,被选择以供包括在裁剪出的蛋白质中),例如:

[0164] $\min(n_k, \max(0, N_{\text{res}} - (n_{\text{added}} + n_{\text{remaining}})))$

[0165] 其中, n_k 是当前链中的氨基酸的数量, N_{res} 是裁剪预算, n_{added} 是从多聚体蛋白质的之前的链中裁剪的氨基酸的数量,并且 $n_{\text{remaining}}$ 是在多聚体蛋白质中的链的排序中在当前链之后的链的组合长度。

[0166] 系统确定要从当前链中裁剪的氨基酸序列的最大长度(152)。系统可以确定要从当前链中裁剪的氨基酸序列的最大长度,例如:

[0167] $\min(N_{\text{res}} - n_{\text{added}}, n_k)$

[0168] 其中, N_{res} 是裁剪预算, n_{added} 是从多聚体蛋白质的之前的链中裁剪的氨基酸的数量,并且 n_k 是当前链中的氨基酸的数量。

[0169] 系统通过从在最小裁剪长度和最大裁剪长度之间的长度范围内的概率分布中进行采样来确定要从当前链中裁剪的氨基酸序列的长度(154)。在一些实施方式中,系统根据最小裁剪长度和最大裁剪长度之间的均匀概率分布来随机采样要从当前链中裁剪的氨基酸序列的长度。

[0170] 系统从具有所确定的裁剪长度的当前链中裁剪连续的氨基酸序列(156)。系统可以确定要从当前链中裁剪的氨基酸序列的位置,例如,通过从区间 $[0, n_k - \text{CropLength} + 1]$ 内的均匀分布中采样当前链中要包括在裁剪出的序列中的第一个氨基酸的索引来确定,其中, n_k 表示当前链中的氨基酸的数量,并且CropLength表示所确定的裁剪长度。

[0171] 系统确定是否满足终止标准(158)。例如,如果从多聚体蛋白质中裁剪的氨基酸总数等于裁剪预算,则系统可以确定满足终止标准。

[0172] 响应于确定不满足终止标准,系统可以进行到下一个氨基酸序列并且返回到步骤150。

[0173] 响应于确定满足终止标准,系统可以输出裁剪出的蛋白质(160)。根据参考步骤

150和152所描述的最小裁剪长度和最大裁剪长度从多聚体蛋白质的每条链中选择裁剪长度,增加了裁剪程序将使用大部分或所有裁剪预算而不超过裁剪预算的可能性。

[0174] 图6是用于确定具有重复的氨基酸链的多聚体蛋白质的预测蛋白质结构与靶蛋白质结构之间的链分配的示例过程162的流程图。为了方便起见,过程162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,根据本说明书适当地编程的训练系统(例如,图2的训练系统122)可以执行过程162。

[0175] 系统接收蛋白质的预测蛋白质结构和靶蛋白质结构(164)。例如,如参考图2描述的,预测蛋白质结构由蛋白质结构预测神经网络来生成。靶蛋白质结构表示蛋白质的实际结构。

[0176] 系统将靶蛋白质结构中的氨基酸链之一指定为“靶锚链”(166)。系统使用靶锚链作为用于将预测结构与靶结构对齐的参考链,这将在下面更详细地描述。系统可以根据任何适当的标准来选择靶锚链。例如,为了选择靶锚链,系统可以从蛋白质中的链当中识别蛋白质中具有最低多重性的一条或多条氨基酸链的集合。如果每条最低多重性链具有相同的长度,则系统可以随机选择最低多重性链之一作为靶锚链。如果最低多重性链的集合包括不同长度的链,则系统可以从具有最长长度的最低多重性链的真子集当中随机选择靶锚链。

[0177] 对每条有效的预测锚链执行步骤168-170。如果来自预测蛋白质结构的链与靶锚链共享相同的氨基酸序列,则该链被称为有效的预测锚链。

[0178] 系统确定来自预测蛋白质结构的每条链到靶蛋白质结构中的对应的链的一对一分配,包括将预测蛋白质结构中的预测锚链分配给靶蛋白质结构中的靶锚链(168)。更具体地,系统对靶蛋白质结构进行变换以将靶蛋白质结构的靶锚链与预测蛋白质结构的预测锚链对齐。然后,系统贪婪地将预测蛋白质结构中的每条剩余的链分配给靶蛋白质结构中的同一序列的其最近邻链。参考图7更详细地描述了用于在将预测锚链分配给靶锚链之后确定链分配的示例过程。

[0179] 系统确定链分配的对齐误差(170)。链分配的对齐误差是指由根据链分配对预测蛋白质结构和靶蛋白质结构进行比较而生成的预测蛋白质结构与靶蛋白质结构之间的误差。根据链分配对预测蛋白质结构和靶蛋白质结构进行比较是指将预测蛋白质结构中的每条链到靶蛋白质结构中的对应的所分配的链进行比较。

[0180] 系统可以以任何适当的方式确定对齐误差。例如,为了确定对齐误差,系统可以对靶蛋白质结构进行变换(例如,使用旋转和平移操作)以将靶蛋白质结构的靶锚链与预测蛋白质结构的预测锚链对齐。然后,系统可以将对齐误差s确定为:

$$[0181] \quad \bar{x}_k^{\text{pred}} = \text{mean}(x_k^{\text{pred}}), k \in \{1, \dots, N\}$$

$$[0182] \quad \bar{x}_k^{\text{gt}} = \text{mean}(x_k^{\text{gt}}), k \in \{1, \dots, N\}$$

$$[0183] \quad s_k = \text{rmsd}(\bar{x}_k^{\text{pred}}, \bar{x}_k^{\text{gt}}), k \in \{1, \dots, N\}$$

$$[0184] \quad s = \sum_{k=1}^N s_k$$

[0185] 其中, k 索引蛋白质中的氨基酸链, N 是蛋白质中的氨基酸链数量, $\text{mean}(x_k^{\text{pred}})$ 是指预测蛋白质结构的链 k 中的原子的坐标的平均值, $\text{mean}(x_k^{\text{gt}})$ 是指靶蛋白质结构的链 k 中的原子的坐标的平均值, $\text{rmsd}(\cdot)$ 是指均方根偏差度量, 并且预测蛋白质结构中的氨基酸链的索引已经被转置, 以使得预测蛋白质结构中的每条氨基酸链具有与靶蛋白质结构中的对应的所分配的链相同的索引。

[0186] 系统基于对齐误差来选择预测蛋白质结构中的链到靶蛋白质结构中的对应的链的最终链分配 (172)。例如, 系统可以选择最终链分配作为与最低对齐误差相关联的链分配。

[0187] 图7是用于确定从预测蛋白质结构到靶蛋白质结构的链分配的示例过程174的流程图, 示例过程174将预测蛋白质结构中的预测锚链分配给靶蛋白质结构中的靶锚链。为了方便起见, 过程174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, 根据本说明书适当地编程的训练系统 (例如, 图2的训练系统122) 可以执行过程174。

[0188] 系统对靶蛋白质结构进行变换以将靶蛋白质结构的靶锚链与预测蛋白质结构的预测锚链对齐 (176)。特别地, 系统通过对靶蛋白质结构应用旋转和平移操作来对靶蛋白质结构进行变换。选择旋转和平移操作的参数以 (近似地或精确地) 最小化靶锚链和预测锚链中的对应的原子的坐标之间的误差 (例如, 均方误差)。

[0189] 系统选择预测蛋白质结构中的未分配链, 即, 预测蛋白质结构中尚未被分配给靶蛋白质结构中的对应的链的链 (178)。

[0190] 系统为靶蛋白质结构中未分配链与预测蛋白质结构中的未分配链具有相同的氨基酸序列的每条未分配链确定 (i) 未分配的预测结构链与 (ii) 未分配的靶结构链之间的相应的误差 (180)。例如, 为了确定未分配的预测结构链与未分配的靶结构链之间的误差, 系统可以确定未分配的预测结构链中的氨基酸的坐标的集中趋势的度量 (例如, 均值、中值或众数)。系统还可以确定未分配的靶结构链中的氨基酸的坐标的集中趋势的度量。然后, 系统可以基于 (i) 未分配的预测结构链中的氨基酸的坐标的集中趋势的度量与 (ii) 未分配的靶结构链中的氨基酸的坐标的集中趋势的度量之间的差值的大小来确定未分配的预测结构链与未分配的靶结构链之间的误差。

[0191] 系统基于误差将未分配的预测结构链分配给对应的未分配的靶结构链 (182)。例如, 系统可以将未分配的预测结构链分配给相对于未分配的预测结构链具有最低误差的对应的未分配的靶结构链。

[0192] 系统确定在预测结构中是否存在任何链尚未被分配给靶结构中的对应的链 (184)。

[0193] 响应于确定预测结构中仍然有链未被分配, 系统选择该链并且返回到步骤178。

[0194] 响应于确定预测结构中的所有链已被分配, 系统确定链分配完成并且输出链分配 (186)。

[0195] 图8示出了证明根据本说明书中描述的技术 (“AlphaFold-多聚体”) 而训练的蛋白质结构神经网络与其他方法相比的优异预测准确度的实验结果。“DockQ” 分数 (在图8中的 y 轴上引用) 是指表征预测蛋白质结构的准确度的度量。误差条表示均值周围95%的置信度区间。

[0196] 图9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是被实现为一个或多个位置中的一个或多个计算机上的计算机程序的系统的示例,其中,实现了下面描述的系统、组件和技术。

[0197] 系统100被配置为处理定义蛋白质104的一条或多条氨基酸链102的数据,以生成定义预测蛋白质结构108(即,蛋白质104的结构的预测)的结构参数106的集合。也就是说,蛋白质104的预测结构108可以由共同定义在蛋白质经历蛋白质折叠之后的蛋白质的预测三维结构的结构参数106的集合来定义。

[0198] 定义预测蛋白质结构108的结构参数106可以如前所述。例如,它们可以包括例如蛋白质104中的每个氨基酸的位置参数和旋转参数、表征蛋白质中的每对氨基酸之间的估计距离的距离图、蛋白质结构中的每个氨基酸中的每个原子或主干原子的相应空间位置、或其组合,如上所述。

[0199] 为了生成定义预测蛋白质结构108的结构参数106,系统100生成:(i)蛋白质的多序列对齐(MSA)表示110,以及(ii)蛋白质的“成对”嵌入112的集合,这将在下面更详细地描述。

[0200] 蛋白质的MSA表示110包括针对蛋白质中的每条氨基酸链的MSA的相应表示。蛋白质中的氨基酸链的MSA表示可以被表示为 $M \times N$ 嵌入阵列(即,具有M行和N列的2-D嵌入阵列),其中,N是氨基酸链中的氨基酸的数量。MSA表示的每行可以与蛋白质中的氨基酸链的相应MSA序列相对应。参考图18描述了用于生成蛋白质中的氨基酸链的MSA表示的示例过程。

[0201] 系统100从蛋白质中的氨基酸链的MSA表示中生成蛋白质104的MSA表示110。

[0202] 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识别为蛋白质中的单个氨基酸链的MSA表示。

[0203] 如果蛋白质包括多条氨基酸链,则系统100可以通过以各种可能方式中的任何一种组装蛋白质中的氨基酸链的MSA表示来生成蛋白质的MSA表示110。接下来描述用于组装蛋白质中的氨基酸链的MSA表示以生成蛋白质的MSA表示110的方法的几个示例。

[0204] 在一些实施方式中,系统100可以通过将蛋白质中的氨基酸链的MSA表示组装成块对角线2-D嵌入阵列(即,其中蛋白质中的氨基酸链的MSA表示形成对角线上的块)来生成蛋白质的MSA表示110。系统100可以将2-D阵列中对角线上的块之外的每个定位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被分配任意的排序,并且蛋白质中的氨基酸链的MSA表示可以在块对角线矩阵中相应地排序。例如,第一氨基酸链(即,根据排序)的MSA表示可以是对角线上的第一块,第二氨基酸链的MSA表示可以是对角线上的第二个块,依此类推。

[0205] 在一些实施方式中,蛋白质是同聚体,即,包括都具有相同的氨基酸序列的多条氨基酸链。在这些实施方式中,系统100可以通过(例如,从左到右或从上到下)级联蛋白质中的氨基酸链的MSA表示来生成蛋白质的MSA表示110。

[0206] 在一些实施方式中,蛋白质是异聚体,即,包括其中至少一些氨基酸链具有不同的氨基酸序列的多条氨基酸链。在这些实施方式中,为了生成蛋白质的MSA表示,系统可以对齐氨基酸链的MSA表示的行,使得同源的MSA序列位于氨基酸链的MSA表示之间的相同行上。对于部分对齐,例如,在MSA不包括一行或多行的适当同源序列的情况下,缺少同源序列的

行可以由默认嵌入来取代。然后,系统可以(例如,从左到右)级联行对齐的MSA表示。

[0207] 剩余未分配的MSA序列(即,在氨基酸链的MSA表示中尚未分配给同源序列的一行的序列)可以以块对角线的方式堆叠在氨基酸链的级联的行对齐的MSA表示之下。系统100可以将对角线上的块之外的每个定位处的嵌入初始化为默认嵌入,例如,零向量。例如,在Tian-ming Zhou等人的“Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis”(bioRxiv,第240754页,2018)中描述了用于对氨基酸链的MSA表示进行行对齐的示例技术。以这种方式生成蛋白质的MSA表示编码了蛋白质的MSA表示中的交叉链遗传(进化)信息。

[0208] 通常,蛋白质的MSA表示110可以被表示为2-D嵌入阵列。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。

[0209] 成对嵌入112的集合包括与蛋白质104中的每对氨基酸相对应的相应成对嵌入。一般来说,成对嵌入表示(即,编码)关于蛋白质中的一对氨基酸之间的关系的信息。一对氨基酸是指蛋白质中包括第一氨基酸和第二氨基酸的有序元组,即,使得蛋白质中的可能氨基酸对的集合由下式给出:

[0210] $\{(A_i, A_j) : 1 \leq i, j \leq N\}$ (1)

[0211] 其中,N是蛋白质中的氨基酸的数量, $i, j \in \{1, \dots, N\}$ 索引蛋白质中的氨基酸, A_i 是由i索引的蛋白质中的氨基酸,并且 A_j 是由j索引的蛋白质中的氨基酸。如果蛋白质包括多条氨基酸链,则蛋白质中的氨基酸可以根据蛋白质中的氨基酸链的排序而从 $\{1, \dots, N\}$ 中顺序地索引。也就是说,首先来自第一氨基酸链的氨基酸被顺序地索引,接着是来自第二氨基酸链的氨基酸,接着是来自第三氨基酸链的氨基酸,依此类推。成对嵌入112的集合可以被表示为2-D的 $N \times N$ 成对嵌入阵列,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阵列中的定位 (i, j) 由该对氨基酸 (A_i, A_j) 的成对嵌入所占据。

[0212] 在一些实施方式中,系统将每对嵌入与相对定位嵌入进行组合。成对嵌入的相对定位嵌入可以表征例如对应的一对氨基酸是否位于蛋白质中的同一链中,以及对应的一对氨基酸是否位于具有完全相同的氨基酸序列的链中。

[0213] 参考图19描述了用于生成(初始化)与蛋白质中的每对氨基酸相对应的相应成对嵌入的示例过程。

[0214] 系统100使用MSA表示110和成对嵌入112两者来生成定义预测蛋白质结构108的结构参数106,因为两者具有互补的属性。MSA表示110的结构可以显式地取决于MSA中与蛋白质中的每条氨基酸链相对应的氨基酸链数量。因此,MSA表示110可能不适用于直接预测蛋白质结构,因为蛋白质结构108对MSA中的氨基酸链数量没有显式的依赖性。相比之下,成对嵌入112表征蛋白质104中的各对氨基酸之间的关系,并且在没有显式地引用MSA的情况下被表达,因此是用于预测蛋白质结构108的方便有效的数据表示。

[0215] 系统100根据嵌入神经网络200的参数集合的值使用嵌入神经网络200来处理MSA表示110和成对嵌入112,以更新MSA表示110和成对嵌入112。也就是说,嵌入神经网络200处理MSA表示110和成对嵌入112,以生成更新的MSA表示114和更新的成对嵌入116。

[0216] 嵌入神经网络200通过在MSA表示110与成对嵌入112之间共享信息来更新MSA表示

110和成对嵌入112。更具体地,嵌入神经网络200在基于当前成对嵌入112更新当前MSA表示110与基于当前MSA表示110更新当前成对嵌入112之间交替。

[0217] 参考图10更详细地描述了嵌入神经网络200的示例架构。

[0218] 系统100从更新的成对嵌入116、更新的MSA表示114或两者中生成折叠神经网络600的网络输入,并且使用折叠神经网络600处理网络输入以生成定义预测蛋白质结构的结构参数106。

[0219] 在一些实施方式中,折叠神经网络600处理更新的成对嵌入116以生成距离图,对于蛋白质中的每对氨基酸,距离图包括蛋白质结构中的该对氨基酸之间的可能距离的集合内的概率分布。例如,为了生成蛋白质结构中的一对氨基酸之间的可能距离的集合内的概率分布,折叠神经网络可以对与该对氨基酸相对应的更新的成对嵌入116应用一个或多个全连接神经网络层。

[0220] 在一些实施方式中,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和更新的成对嵌入116两者中导出的网络输入来生成结构参数106,几何注意力操作显式地推理蛋白质结构中的氨基酸的3-D几何构造。参考图14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。

[0221] 图10示出了被配置为处理MSA表示110和成对嵌入112以生成更新的MSA表示114和更新的成对嵌入116的嵌入神经网络200的示例架构。

[0222] 嵌入神经网络200包括更新块202-A-N的序列。在整个说明书中,“块”是指神经网络的一部分,例如,神经网络的包括一个或多个神经网络层的子网。

[0223] 嵌入神经网络中的每个更新块被配置为接收包括MSA表示和成对嵌入的块输入,并且处理块输入以生成包括更新的MSA表示和更新的成对嵌入的块输出。

[0224] 嵌入神经网络200向第一更新块(即,在更新块的序列中)提供嵌入神经网络200的包括在网络输入中的MSA表示110和成对嵌入112。第一更新块处理MSA表示110和成对嵌入112以生成更新的MSA表示和更新的成对嵌入。

[0225] 对于第一更新块之后的每个更新块,嵌入神经网络200向更新块提供由前一更新块生成的MSA表示和成对嵌入,并且向下一更新块提供由更新块生成的更新的MSA表示和更新的成对嵌入。

[0226] 嵌入神经网络200通过使用更新块202-A-N的序列重复地更新MSA表示110和成对嵌入112来逐渐丰富它们的信息内容

[0227] 嵌入神经网络200可以提供由最终更新块(即,在更新块的序列中)生成的更新的MSA表示114和更新的成对嵌入116作为网络输出。

[0228] 图11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10描述的。

[0229] 更新块300接收包括当前MSA表示302和当前成对嵌入304的块输入,并且处理块输入以生成更新的MSA表示306和更新的成对嵌入308。

[0230] 更新块300包括MSA更新块400和成对更新块500。

[0231] MSA更新块400使用当前成对嵌入304来更新当前MSA表示302,并且成对更新块500使用更新的MSA表示306(即,其由MSA更新块400生成)来更新当前成对嵌入304。

[0232] 通常,MSA表示和成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氨基酸链集合当中不同定位中的氨基酸身份之间的相关性的信息,并且成对嵌入可以编

码关于蛋白质中的氨基酸之间的相互关系的信息。MSA更新块400使用编码在成对嵌入中的互补信息来丰富MSA表示的信息内容,并且成对更新块500使用编码在MSA表示中的互补信息来丰富成对嵌入的信息内容。作为这种丰富的结果,更新的MSA表示和更新的成对嵌入编码了与预测蛋白质结构更相关的信息。

[0233] 更新块300在本文中被描述为首先使用当前成对嵌入304来更新当前MSA表示302,然后使用更新的MSA表示306来更新当前成对嵌入304。该描述不应被理解为将更新块限制为以该顺序执行操作,例如,更新块可以首先使用当前MSA表示来更新当前的成对嵌入,然后使用更新的成对嵌入来更新当前MSA表示。

[0234] 更新块300在本文中被描述为包括MSA更新块400(即,其更新当前MSA表示)和成对更新块500(即,其更新当前成对嵌入)。该描述不应被理解为将更新块300限制为仅包括一个MSA更新块或仅包括一个成对更新块。例如,更新块300可以包括在MSA表示被提供给成对更新块以用于更新当前成对嵌入之前多次更新MSA表示的多个MSA更新块。作为另一示例,更新块300可以包括使用MSA表示多次更新成对嵌入的多个成对更新块。

[0235] MSA更新块400和成对更新块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。

[0236] 在一些实施方式中,MSA更新块400、成对更新块500或两者包括一个或多个“自注意力”块。如整个文档中所使用的,自注意力块通常是指更新嵌入集合(即,接收嵌入集合并且输出更新嵌入)的神经网络块。为了更新给定嵌入,自注意力块可以确定给定嵌入与一个或多个所选择的嵌入中的每个嵌入之间的相应“注意力权重”,然后使用(i)注意力权重和(ii)所选择的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“对(over)”所选择的嵌入的注意力来更新给定嵌入。

[0237] 例如,自注意力块可以接收输入嵌入集合 $\{x_i\}_{i=1}^N$,其中,N是蛋白质中的氨基酸的数量,并且为了更新嵌入 x_i ,自注意力块可以将注意力权重 $[a_{i,j}]_{j=1}^N$ 确定如下,其中, $a_{i,j}$ 表示 x_i 与 x_j 之间的注意力权重:

$$[0238] \quad [a_{i,j}]_{j=1}^N = \text{softmax}\left(\frac{(W_q x_i) K^T}{c}\right) \quad (5)$$

$$[0239] \quad K^T = [W_k x_j]_{j=1}^N \quad (6)$$

[0240] 其中, W_q 和 W_k 是学习参数矩阵, $\text{softmax}(\cdot)$ 表示soft-max归一化运算,并且c是常数。使用注意力权重,自注意力层可以将嵌入 x_i 更新为:

$$[0241] \quad x_i \leftarrow \sum_{j=1..N} a_{i,j} \cdot (W_v x_j) \quad (7)$$

[0242] 其中, W_v 是学习参数矩阵。 $(W_q x_i)$ 可以被称作输入嵌入 x_i 的“查询嵌入”, $W_k x_j$ 可以被称作输入嵌入 x_j 的“关键字嵌入”,并且 $W_v x_j$ 可以被称作输入嵌入 x_j 的“值嵌入”。

[0243] 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。包括在MSA更新块400和成对更新块500中的任何自注意力块的参

数可以被理解为能够作为参考图9描述的蛋白质结构预测系统100的端到端训练的一部分进行训练的更新块300的参数。通常,查询嵌入矩阵、关键字嵌入矩阵和值嵌入矩阵的(经训练的)参数对于不同的自注意力块是不同的,例如,使得包括在MSA更新块400中的自注意力块可以具有参数与包括在成对更新块500中的自注意力块不同的不同查询嵌入矩阵、关键字嵌入矩阵和值嵌入矩阵。

[0244] 在一些实施方式中,MSA更新块400、成对更新块500或两者包括一个或多个以成对嵌入为条件(即,实现以成对嵌入为条件的自注意力操作)的自注意力块。为了使自注意力操作以成对嵌入为条件,自注意力块可以处理成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 $[b_{i,j}]_{j=1}^N$ 的集合,其中, $b_{i,j}$ 表示 x_i 与 x_j 之间的注意力偏差。自注意力块可以通过将学习参数矩阵应用于成对嵌入 $h_{i,j}$ (即,由(i,j)索引的蛋白质中的氨基酸对)来生成注意力偏差 $b_{i,j}$ 。

[0245] 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(或以其他方式进行组合))确定“有偏差的注意力权重” $[c_{i,j}]_{j=1}^N$ 的集合,其中, $c_{i,j}$ 表示 x_i 与 x_j 之间的有偏差的注意力权重。例如,自注意力块可以将嵌入 x_i 与 x_j 之间的有偏差的注意力权重 $c_{i,j}$ 确定为:

$$[0246] \quad c_{i,j} = a_{i,j} + b_{i,j}$$

[0247] 其中, $a_{i,j}$ 是 x_i 与 x_j 之间的注意力权重, $b_{i,j}$ 是 x_i 与 x_j 之间的注意力偏差。自注意力块可以使用有偏差的注意力权重来更新每个输入嵌入 x_i ,例如:

$$[0248] \quad x_i \leftarrow \sum_{j=1..N} c_{i,j} \cdot (W_v x_j) \quad (8)$$

[0249] 其中, W_v 是学习参数矩阵。

[0250] 通常,成对嵌入编码表征蛋白质结构和蛋白质结构中的氨基酸对之间的关系的信。将以成对嵌入为条件的自注意力操作应用于输入嵌入集合,允许了输入嵌入以通过被编码在成对嵌入中的蛋白质结构信息进行通知的方式来更新。嵌入神经网络的更新块可以使用以成对嵌入为条件的自注意力块来更新和丰富MSA表示和成对嵌入本身。

[0251] 可选地,自注意力块可以具有多个“头部(head)”,每个“头部”生成与每个输入嵌入相对应的相应更新嵌入,即,使得每个输入嵌入与多个更新嵌入相关联。例如,每个头部可以根据参考等式(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嵌入。具有多个头部的自注意力块可以实现“门控”操作以组合由头部为输入嵌入生成的更新嵌入,即,生成与每个输入嵌入相对应的单个更新嵌入。例如,自注意力块可以使用一个或多个神经网络层(例如,全连接神经网络层)处理输入嵌入以生成每个头部的相应门控值。自注意力块然后可以根据门控值来组合与输入嵌入相对应的更新嵌入。例如,自注意力块可以将输入嵌入 x_i 的更新嵌入生成为:

$$[0252] \quad \sum_{k=1}^K \alpha_k \cdot x_i^{next} \quad (9)$$

[0253] 其中, k 索引头部, α_k 是头部 k 的门控值,并且是由头部 k 为输入嵌入 x_i 生成的更新嵌入 x_i^{next} 。

[0254] 参考图12描述了使用以成对嵌入为条件的自注意力块的MSA更新块400的示例架构。通过使用以当前成对嵌入为条件的自注意力块处理当前MSA表示的行,参考图12描述的示例MSA更新块基于当前成对嵌入来更新当前MSA表示。

[0255] 参考图13描述了使用以成对嵌入为条件的自注意力块的成对更新块500的示例架构。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成对嵌入并且使用以当前成对嵌入为条件的自注意力块处理当前成对嵌入,参考图13描述的示例成对更新块基于更新的MSA表示来更新当前成对嵌入。

[0256] 图12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)基于当前成对嵌入来更新当前MSA表示306。

[0257] 为了更新当前MSA表示302,MSA更新块400使用以当前成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表示的每行中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每行中的嵌入提供给以当前成对嵌入为条件的“逐行”自注意力块402,例如,如参考图11描述的,以生成当前MSA表示302的每行的更新嵌入。可选地,MSA更新块可以将逐行自注意力块402的输入加到逐行自注意力块402的输出。使逐行自注意力块402以当前成对嵌入为条件使得MSA更新块400能够使用来自当前成对嵌入的信息来丰富当前MSA表示302。

[0258] 然后,MSA更新块使用不以当前成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表示的每列中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每列中的嵌入提供给不以当前成对嵌入为条件的“逐列”自注意力块404,以生成当前MSA表示302的每列的更新嵌入。作为不以当前成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)描述的)而不是有偏差的注意力权重(例如,如参考等式(8)描述的)来为当前MSA表示的每列生成更新嵌入。可选地,MSA更新块可以将对逐列自注意力块404的输入加到逐列自注意力块404的输出。

[0259] 然后,MSA更新块使用过渡块来处理当前MSA表示302,例如,过渡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可以将过渡块406的输入加到过渡块406的输出。

[0260] MSA更新块可以输出通过由逐行自注意力块402、逐列自注意力块404和过渡块406执行的操作而生成的更新的MSA表示306。

[0261] 图13示出了成对更新块500的示例架构。成对更新块500被配置为接收当前成对嵌入304,并且(至少部分)基于更新的MSA表示306来更新当前成对嵌入304。

[0262] 为了更新当前成对嵌入304,成对更新块500对更新的MSA表示306应用外积均值运算502,并且将外积均值运算502的结果加到当前成对嵌入304。

[0263] 外积均值运算定义了这样一种运算序列,当该运算序列被应用于表示为 $M \times N$ 嵌入阵列的MSA表示时,生成 $N \times N$ 嵌入阵列,即,其中, N 是蛋白质中的氨基酸的数量。当前成对嵌入304也可以被表示为 $N \times N$ 嵌入阵列,并且将外积均值502的结果加到当前成对嵌入304是指对两个嵌入 $N \times N$ 阵列求和。

[0264] 为了计算外积均值,成对更新块生成张量 $A(\cdot)$,例如, $A(\cdot)$ 由下式给出:

$$[0265] \quad A(res1, res2, ch1, ch2) = \frac{1}{|rows|} \sum_{rows} LeftAct(row, res1, ch1) \cdot RightAct(row, res2, ch2)$$

[0266] 其中, $res1, res2 \in \{1, \dots, N\}$, $ch1, ch2 \in \{1, \dots, C\}$, 其中, C 是MSA表示的每个嵌入中的通道数量, $|rows|$ 是MSA表示中的行数, $LeftAct(row, res1, ch1)$ 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 $RightAct(row, res2, ch2)$ 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)。外积均值的结果是通过对张量 A 的 $(ch1, ch2)$ 维度进行展平和线性投影而生成的。可选地, 成对更新块可以执行一个或多个层归一化运算(例如, 如参考Jimmy Lei Ba等人的“Layer Normalization”(arXiv:1607.06450)描述的), 作为计算外积均值的一部分。

[0267] 通常, 更新的MSA表示306编码关于在进化相关氨基酸链集合当中不同位置中的氨基酸身份之间的相关性的信息。编码在更新的MSA表示306中的信息与预测蛋白质的结构相关, 并且通过将编码在更新的MSA表示中的信息结合到当前成对嵌入中(即, 借由外积均值502), 成对更新块500可以增强当前成对嵌入的信息内容。

[0268] 在使用更新的MSA表示(即, 借由外积均值502)来更新当前成对嵌入304之后, 成对更新块500使用以当前成对嵌入为条件的自注意力操作(即, “逐行”自注意力操作)将当前成对嵌入的布置的每行中的当前成对嵌入更新成 $N \times N$ 阵列。更具体地, 成对更新块500将当前成对嵌入的每行提供给也以当前成对嵌入为条件的“逐行”自注意力块504, 例如, 如参考图11描述的, 以生成每行的更新的成对嵌入。可选地, 成对更新块可以将逐行自注意力块504的输入加到逐行自注意力块504的输出。

[0269] 然后, 成对更新块500使用也以当前成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 $N \times N$ 当前成对嵌入阵列的每列中的当前成对嵌入。更具体地, 成对更新块500将当前成对嵌入的每列提供给也以当前成对嵌入为条件的“逐列”自注意力块506, 以生成每列的更新的成对嵌入。可选地, 成对更新块可以将自注意力块506的输入加到逐列自注意力块506的输出。

[0270] 然后, 成对更新块500使用(例如, 对当前成对嵌入应用一个或多个全连接神经网络层的)过渡块来处理当前成对嵌入。可选地, 成对更新块500可以将过渡块508的输入加到过渡块508的输出。

[0271] 成对更新块可以输出通过由逐行自注意力块504、逐列自注意力块506和过渡块508执行的操作而生成的更新的成对嵌入308。

[0272] 图14示出了生成定义预测蛋白质结构108的结构参数106的集合的折叠神经网络600的示例架构。例如, 在实施方式中, 通过处理包括蛋白质中的每对氨基酸的相应成对嵌入116的输入以生成结构参数106的值, 折叠神经网络600基于成对嵌入来确定蛋白质的预测结构。折叠神经网络600可以被包括在参考图9描述的蛋白质结构预测系统100中。

[0273] 在实施方式中, 折叠神经网络600生成结构参数, 结构参数可以包括蛋白质中的每个氨基酸的(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的氨基酸中的特定原子的预测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“朝向”。更具体地, 旋转参数可以指定: 如果3-D空间旋转操作被应用于

位置参数的坐标系,则使得氨基酸中的三个“主链”原子相对于旋转后的坐标系而呈现固定的定位。

[0274] 在实施方式中,折叠神经网络600接收从最终MSA表示、最终成对嵌入或两者中导出的输入,并且生成定义蛋白质的预测结构的结构参数106的最终值。例如,折叠神经网络600可以接收包括以下各项的输入:(i)蛋白质中的每对氨基酸的相应成对嵌入116,(ii)蛋白质中的每个氨基酸的“单个”嵌入602的初始值,以及(iii)蛋白质中的每个氨基酸的结构参数604的初始值。折叠神经网络600处理输入以生成共同表征蛋白质的预测结构108的结构参数106的最终值。

[0275] 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而生成的成对嵌入,如参考图9描述的。

[0276] 蛋白质结构预测系统100可以从MSA表示114(即,作为嵌入神经网络的输出而生成)中生成氨基酸的初始单个嵌入602,如参考图9描述的。例如,如上所述,MSA表示114可以被表示为列数等于蛋白质中的氨基酸的数量的2-D嵌入阵列,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。作为另一示例,蛋白质结构预测系统100可以通过从与估计了结构的蛋白质的氨基酸序列相对应的MSA表示114的行中提取嵌入来为蛋白质中的氨基酸生成初始单个嵌入。

[0277] 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,笛卡尔坐标系中的 $[0,0,0]$),并且每个氨基酸的旋转参数被初始化为 3×3 单位矩阵。

[0278] 折叠神经网络600可以通过重复地更新单个嵌入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)。折叠神经网络600可以包括除了更新块之外的其他神经网络层或块,例如,可以与更新块交错的其他神经网络层或块。

[0279] 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。

[0280] 几何注意力块616使用“几何”自注意力操作来更新当前单个嵌入,“几何”自注意力操作显式地推理蛋白质结构中地氨基酸的3-D几何构造,即,如结构参数定义的。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与一个或多个所选择的单个嵌入中的每个单个嵌入之间的相应注意力权重,其中,注意力权重取决于当前单个嵌入、当前结构参数和成对嵌入。几何注意力块616然后使用(i)注意力权重、(ii)所选择的单个嵌入和(iii)当前结构参数来更新给定的单个嵌入。

[0281] 为了确定注意力权重,几何注意力块616处理每个当前单个嵌入以生成相应的“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*氨基酸相对应的单个嵌入 h_i 将符号查询嵌入 q_i 、符号关键字嵌入 k_i 和符号值嵌入 v_i 生成成为:

$$[0282] \quad q_i = \text{Linear}(h_i) \quad (10)$$

$$[0283] \quad k_i = \text{Linear}(h_i) \quad (11)$$

[0284] $v_i = \text{Linear}(h_i)$ (12)

[0285] 其中, $\text{Linear}(\cdot)$ 是指具有独立的学习参数值的线性层。

[0286] 几何注意力块616另外处理每个当前单个嵌入以生成对应的“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嵌入、几何关键字嵌入和几何值嵌入均是最初在对应的氨基酸的局部参考系中生成、然后使用氨基酸的结构参数而旋转和转换到全局参考系的3-D点。例如,几何注意力块616可以为与第i氨基酸相对应的单个嵌入 h_i 将几何查询嵌入 q_i^p 、几何关键字嵌入 k_i^p 和几何值嵌入 v_i^p 生成:

[0287] $q_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (13)

[0288] $k_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (14)

[0289] $v_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (15)

[0290] 其中, $\text{Linear}_p(\cdot)$ 是指具有将 h_i 投影到3-D点的独立的学习参数值的线性层(上标 p 指示该量是3-D点), R_i 表示由第i氨基酸的旋转参数所指定的旋转矩阵, 并且 t_i 表示第i氨基酸的位置参数。

[0291] 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可以如下生成注意力权重 $[a_j]_{j=1}^N$, 其中, N是蛋白质中的氨基酸总数, 并且 a_j 是氨基酸i与氨基酸j之间的注意力权重:

[0292] $[a_j]_{j=1}^N = \text{softmax} \left(\left[\frac{q_i \cdot k_j}{\sqrt{m}} + \alpha |q_i^p - k_j^p|_2^2 + (b_{i,j} \cdot w) \right]_{j=1}^N \right)$ (16)

[0293] 其中, q_i 表示氨基酸i的符号查询嵌入, k_j 表示氨基酸j的符号关键字嵌入, m表示 q_i 和 k_j 的维数, α 表示学习参数, q_i^p 表示氨基酸i的几何查询嵌入, k_j^p 表示氨基酸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成对嵌入 116, 并且 w 是学习权重向量(或一些其他学习投影运算)。

[0294] 通常, 一对氨基酸的成对嵌入隐式地编码了与该对氨基酸中的氨基酸之间的关系相关的信息, 例如, 该对氨基酸中的氨基酸之间的距离。通过部分基于氨基酸i和j的成对嵌入确定氨基酸i与氨基酸j之间的注意力权重, 折叠神经网络600利用来自成对嵌入的信息来丰富注意力权重, 从而提高预测折叠结构的准确度。

[0295] 在一些实施方式中, 几何注意力块616生成几何查询嵌入、几何关键字嵌入和几何值嵌入的多个集合, 并且使用每个生成的几何嵌入集合来确定注意力权重。

[0296] 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 。特别地, 几何注意力块616使用注意力权重来生成“符号回报”嵌入和“几何回报”嵌入, 然后使用符号回报嵌入和几何回报嵌入来更新单个嵌入。几何注意力块124可以生成氨基酸i的符号回报嵌入 o_i , 例如:

[0297] $o_i = \sum_j a_j v_j$ (17)

[0298] 其中, $[a_j]_{j=1}^N$ 表示注意力权重 (例如, 参考等式 (16) 而定义的), 并且每个 v_j 表示氨基酸 j 的符号嵌入值。几何注意力块616可以将氨基酸 i 的几何回报嵌入 o_i^p 生成为例如:

$$[0299] \quad o_i^p = R_i^{-1} \cdot \left(\sum_j a_j v_j^p - t_i \right) \quad (18)$$

[0300] 其中, 几何回报嵌入 o_i^p 是3-D点, $[a_j]_{j=1}^N$ 表示注意力权重 (例如, 参考等式 (16) 而定义的), R_i^{-1} 是由氨基酸 i 的旋转参数所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, 几何回报嵌入最初是在全局参考系中生成的, 然后被旋转和平移到对应的氨基酸的局部参考系。

[0301] 几何注意力块616可以使用对应的符号回报嵌入 o_i (例如, 根据等式 (17) 而生成的) 和几何回报嵌入 o_i^p (例如, 根据等式 (18) 而生成的) 将氨基酸 i 的单个嵌入 h_i 更新为例如:

$$[0302] \quad h_i^{next} = \text{LayerNorm} \left(h_i + \text{Linear}(o_i, o_i^p, |o_i^p|) \right) \quad (19)$$

[0303] 其中, h_i^{next} 是氨基酸 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 的“Layer Normalization” (arXiv:1607.06450 (2016)) 描述的。

[0304] 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) 描述的, 使几何注意力块616能够在更新单个嵌入时推理3-D几何构造。此外, 每个更新块以对总体蛋白质结构的旋转和平移不变的方式来更新单个嵌入和结构参数。例如, 对提供给折叠神经网络600的初始结构参数应用相同的全局旋转和平移操作将会使得折叠神经网络600生成以相同方式全局地旋转和平移但是其他方面相同的预测结构。因此, 应用于初始结构参数的全局旋转和平移操作将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度。由折叠神经网络600生成的表示的旋转和平移不变性促进了训练, 例如, 因为折叠神经网络600自动地学习对蛋白质结构的所有旋转和平移进行泛化。

[0305] 在被提供给折叠块618之前, 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的神经网络层 (例如, 线性神经网络层) 进一步变换。

[0306] 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来更新当前结构参数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:

$$[0307] \quad t_i^{next} = t_i + \text{Linear}(h_i^{next}) \quad (20)$$

[0308] 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:

$$[0309] \quad w_i = \text{Linear}(h_i^{\text{next}}) \quad (21)$$

$$[0310] \quad R_i^{\text{next}} = R_i \cdot \text{QuaternionToRotation}(1 + w_i) \quad (22)$$

[0311] 其中, w_i 是三维向量, $\text{Linear}(\cdot)$ 是线性神经网络层, h_i^{next} 是氨基酸i的更新的单个嵌入, $1+w_i$ 表示实部为1且虚部为 w_i 的四元数,并且 $\text{QuaternionToRotation}(\cdot)$ 表示将四元数变换成等效的 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。

[0312] 折叠神经网络600可以提供由最终更新块610生成的更新的结构参数,作为定义预测蛋白质结构108的最终结构参数106。折叠神经网络600可以包括任何适当数量的更新块,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共同更新的单个参数值集合。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数量,因此可以(例如,通过稳定化训练和减少过拟合的可能性)促进折叠神经网络的有效训练。

[0313] 在训练期间,训练引擎可以基于如上所述评估最终结构参数106的准确度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。在一些实施方式中,训练引擎还可以评估在最终更新块(即,产生最终结构参数的更新块)之前的更新块610中的一个或多个的辅助结构损失。更新块的辅助结构损失评估由更新块生成的更新的结构参数的准确度。

[0314] 可选地,在训练期间,训练引擎可以应用“停止梯度”操作来防止梯度反向传播通过每个更新块的某些神经网络参数,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)中描述的)。应用这些停止梯度运算可以提高在训练期间计算的梯度的数值稳定性。

[0315] 通常,由折叠神经网络600生成的预测蛋白质结构108与对应的基准真值蛋白质结构之间的相似性可以例如通过为预测蛋白质结构中的多个原子中的每个原子分配相应的准确度分数的相似性度量来度量。例如,相似性度量可以为预测蛋白质结构中的每个碳 α 原子分配相应的准确度分数。预测蛋白质结构中的原子的准确度分数可以表征预测蛋白质结构中的原子的位置与基准真值蛋白质结构中的原子的实际位置有多接近。可以将预测蛋白质结构与基准真值蛋白质结构进行比较以生成预测蛋白质结构中的原子的准确度分数的相似性度量示例是参考V.Mariani等人的“lDDT:a local superposition-free score for comparing protein structures and models using distance difference tests”(Bioinformatics,2013年11月1日;29(21)2722-2728)所描述的lDDT相似性度量。

[0316] 折叠神经网络600可以被配置为:为预测蛋白质结构108中的一个或多个原子中的每个原子生成相应的置信度估计650。预测蛋白质结构中的原子的置信度估计650表征了预测蛋白质结构中的原子的预测准确度分数(例如,lDDT准确度分数),即,将会通过将预测蛋白质结构与(潜在未知的)基准真值蛋白质结构进行比较的相似性度量而生成的预测准确度分数。在一个示例中,预测蛋白质结构中的原子的置信度估计650可以定义在形成该原子的准确度分数的可能值范围的分区的一个或多个区间的集合内的离散概率分布。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范围可以是 $[0, 100]$,并且置信度估计650可以定义在区间集合 $\{[0, 2), [2, 4), \dots, [98, 100]\}$ 内的概率分布。在另一示例中,预测蛋白质结构中的原子的置信度估计650可以是数值,即,直接预测原子的准确度分数的数值。

[0317] 在一些实施方式中,折叠神经网络600为蛋白质的每个氨基酸中的特定原子(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(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一个更新块所生成的氨基酸的更新的单个嵌入)生成蛋白质中的氨基酸的指定原子的置信度估计650。

[0318] 结构预测系统可以基于预测蛋白质结构中的原子的置信度估计650来生成与蛋白质中的每个氨基酸相对应的相应置信度分数。例如,结构预测系统可以将氨基酸的置信度分数生成成为氨基酸中的 α 碳原子的准确度分数的可能值内的概率分布的期望值。

[0319] 结构预测系统可以将整个预测结构的置信度分数生成成为例如蛋白质中的氨基酸的置信度分数的平均值。

[0320] 在对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节结构预测系统的参数值,辅助损失度量(i)由折叠神经网络600生成的置信度估计与(ii)通过将预测蛋白质结构与基准真值蛋白质结构进行比较而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。

[0321] 在一些实施方式中,为了生成预测蛋白质结构的置信度分数,折叠神经网络可以处理每对嵌入 z_{ij} (与蛋白质中的氨基酸i和氨基酸j相对应)以生成误差量 e_{ij} 的估计,误差量 e_{ij} 捕获当使用氨基酸i的旋转参数和位置参数对蛋白质的预测结构和靶结构进行对齐时氨基酸j的 α 碳原子的定位的误差:

$$[0322] \quad e_{ij} = \|T_i^{-1}x_j - T_i^{true^{-1}}x_j^{true}\|$$

$$[0323] \quad T_i^{-1}x_j = R_i^{-1}(x_j - t_i)$$

$$[0324] \quad T_i^{true^{-1}}x_j^{true} = R_i^{true^{-1}}(x_j^{true} - t_i^{true})$$

[0325] 其中, x_j 是原子j的预测定位, x_j^{true} 是原子j的靶定位, R_i 是氨基酸i的预测旋转矩阵, t_i 表示氨基酸i的预测位置参数, R_i^{true} 是氨基酸i的靶旋转矩阵,并且 t_i^{true} 表示氨基酸i的靶位置参数。折叠神经网络可以通过使用一个或多个神经网络层处理对应的成对嵌入 z_{ij} 以生成 e_{ij} 的可能值范围的集合内的分数分布来生成每个 e_{ij} (如上定义)的估计。

[0326] 结构预测系统可以处理由折叠神经网络生成的误差量 $\{e_{ij}\}$ 的估计,以生成预测蛋白质结构的置信度分数。例如,结构预测系统可以将置信度分数pTM生成成为:

$$[0327] \quad pTM = \max_i \frac{1}{N_{res}} \sum_j \mathbb{E}[f(e_{ij})]$$

$$[0328] \quad f(e_{ij}) = \left(1 + \left(\frac{e_{ij}}{c}\right)^2\right)^{-1}$$

[0329] 其中, $i, j \in \{1, \dots, N_{res}\}$,其中, N_{res} 是蛋白质中的氨基酸的数量, $\mathbb{E}[\cdot]$ 表示在由折叠神经网络生成的分数分布内进行的期望运算,并且c是常数。

[0330] 在一些实施方式中,结构预测系统可以处理由折叠神经网络生成的误差量 $\{e_{ij}\}$ 的估计,以生成预测蛋白质结构的置信度分数,置信度分数被设计成对不同链中的氨基酸之间的相互作用进行评分。例如,结构预测系统可以将置信度分数ipTM生成成为:

$$[0331] \quad ipTM = \max_i \frac{1}{|\mathcal{D}_{\text{-chain}(i)}|} \sum_{j \in \mathcal{D}_{\text{-chain}(i)}} \mathbb{E}[f(e_{ij})]$$

$$[0332] \quad f(e_{ij}) = \left(1 + \left(\frac{e_{ij}}{c'}\right)^2\right)^{-1}$$

[0333] 其中, $i \in \{1, \dots, N_{\text{res}}\}$,其中, N_{res} 是蛋白质中的氨基酸的数量, c' 是常数, $\mathbb{E}[\cdot]$ 表示在由折叠神经网络生成的分数分布内进行的期望运算,并且 $\mathcal{D}_{\text{-chain}(i)}$ 是蛋白质中除了氨基酸 i 的链中的氨基酸以外的所有氨基酸的集合。

[0334] 在一些实施方式中,结构预测系统可以生成置信度分数以考虑链内置信度(例如,通过如上定义的pTM)和链间置信度(例如,通过如上定义的ipTM)。例如,结构预测系统可以将置信度分数生成成为:

$$[0335] \quad \alpha_1 \cdot pTM + \alpha_2 \cdot ipTM$$

[0336] 其中,pTM和ipTM如上定义,并且 α_1 和 α_2 是缩放因子,例如,总和为1的正缩放因子。

[0337] 由结构预测系统生成的置信度估计可以以各种方式来使用。例如,预测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而因此适于进一步的下游处理或分析。作为另一示例,每个蛋白质的置信度分数可以用于对蛋白质结构的预测(例如,已经由同一结构预测系统通过处理表征同一蛋白质的不同输入而生成的预测,或者已经由不同的结构预测系统生成的预测)的集合进行排序。

[0338] 由结构参数106指定的位置参数和旋转参数可以定义蛋白质的氨基酸中的主链原子(例如,在 $[x, y, z]$ 笛卡尔坐标中)的空间位置。然而,结构参数106不一定定义蛋白质的氨基酸中的剩余原子(例如,氨基酸的侧链中的原子)的空间位置。特别地,氨基酸中的剩余原子的空间位置取决于氨基酸中的键之间的扭转角(例如, ω 角、 Φ 角、 Ψ 角、 χ_1 角、 χ_2 角、 χ_3 角和 χ_4 角,如参考图15所示)的值。

[0339] 可选地,折叠神经网络600的更新块610中的一个或多个可以生成定义蛋白质的每个氨基酸中的每个原子的相应预测空间位置的输出。为了生成氨基酸中的原子的预测空间位置,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的单个嵌入,以生成氨基酸中的原子之间的键的扭转角的预测值。神经网络层可以是例如被嵌入有残差连接的全连接神经网络层。每个扭转角可以被表示为例如2-D向量。

[0340] 更新块可以基于以下各项来确定氨基酸中的原子的空间位置:(i)氨基酸的扭转角值,以及(ii)氨基酸的更新的结构参数(例如,位置参数和旋转参数)。例如,更新块可以根据预定义函数来处理扭转角,以生成氨基酸中的原子在氨基酸的局部参考系中的空间位置。通过根据氨基酸的更新的结构参数旋转和平移原子的空间位置,更新块可以生成氨基酸中的原子在全局参考系(即,蛋白质中的所有氨基酸共有的参考系)中的空间位置。例如,通过对原子在局部参考系中的空间位置应用由更新的旋转参数定义的旋转操作以生成旋转后的空间位置,更新块可以确定原子在全局参考系中的空间位置,然后将由更新的位置

参数定义的平移操作应用于旋转后的空间位置。

[0341] 在一些实施方式中,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输出蛋白质的氨基酸中由最终更新块生成的原子的预测空间位置。

[0342] 参考图14描述的折叠神经网络600在本文中被表征为接收基于由嵌入神经网络生成的MSA表示114和成对嵌入116的输入,例如,参考图10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和成对嵌入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,为蛋白质的每个氨基酸中的原子预测空间位置)可以由其他折叠神经网络(例如,具有接收不同输入的不同架构的其他折叠神经网络)来执行。

[0343] 图15图示了氨基酸的中键之间的扭转角,例如, ω 角、 Φ 角、 Ψ 角、 χ_1 角、 χ_2 角、 χ_3 角、 χ_4 角和 χ_5 角。

[0344] 图16是未折叠蛋白质和折叠蛋白质的图示。未折叠蛋白质是氨基酸的随机卷曲。未折叠蛋白质经历了蛋白质折叠,并且折叠成3D构型。蛋白质结构往往包括稳定的局部折叠样式,诸如 α 螺旋(例如,如802所示)和 β 折叠。

[0345] 图17是用于预测包括一条或多条链的蛋白质的结构的示例过程900的流程图,其中,每条链指定氨基酸序列。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,根据本说明书适当地编程的蛋白质结构预测系统(例如,图1的蛋白质结构预测系统100)可以执行过程900。

[0346] 系统获得表示与蛋白质中的每条链相对应的相应MSA的初始多序列对齐(MSA)表示(902)。

[0347] 系统为蛋白质中的每对氨基酸获得相应的初始成对嵌入(904)。

[0348] 系统使用嵌入神经网络来处理包括初始MSA表示和初始成对嵌入的输入,以生成包括蛋白质中的每对氨基酸的最终MSA表示和相应的最终成对嵌入的输出。

[0349] 嵌入神经网络包括更新块的序列。每个更新块具有相应的更新块参数集合,并且被配置为接收蛋白质中的每对氨基酸的当前MSA表示和相应的当前成对嵌入。每个更新块:(i) 基于当前成对嵌入来更新当前MSA表示,以及(ii) 基于更新的MSA表示来更新当前成对嵌入(906)。

[0350] 系统基于使用最终MSA表示、最终成对嵌入或两者来确定蛋白质的预测结构(908)。

[0351] 图18示出了用于生成蛋白质中的氨基酸链的MSA表示1010的示例过程1000。参考图1描述的蛋白质结构预测系统100可以实现过程1000的操作。

[0352] 为了生成蛋白质中的氨基酸链的MSA表示100,系统100获得蛋白质的MSA 1002,MSA 1002可以包括例如数千个MSA序列。

[0353] 系统100将MSA序列集合划分成“核心”MSA序列1004的集合和“额外”MSA序列1006的集合。核心MSA序列集合可以比额外MSA序列集合1006更小(例如,小一个数量级)。系统100可以(例如,通过随机选择预定数量的MSA序列作为核心MSA序列,并且将剩余的MSA序列标识为额外MSA序列1006)将MSA序列集合划分成核心MSA序列1004和额外MSA序列1006。

[0354] 对于每个额外MSA序列1006,系统100可以确定额外MSA序列与每个核心MSA序列1004之间的相应相似性度量(例如,基于汉明(Hamming)距离的相似性度量)。系统100然后

可以将每个额外MSA序列1006关联到与额外MSA序列1006最相似(即,根据相似性度量最类似于)的对应的核心MSA序列1004。与核心MSA序列1004相关联的额外MSA序列1006的集合可以被称为“MSA序列聚类”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序列聚类1008,其中,与核心MSA序列1004相对应的MSA序列聚类1008包括与核心MSA序列1004最相似的额外MSA序列1006的集合。

[0355] 系统100可以基于核心MSA序列和MSA序列聚类1008来生成蛋白质中的氨基酸链的MSA表示。MSA表示1010可以由 $M \times N$ 嵌入阵列来表示,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 是氨基酸链中的氨基酸的数量。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。

[0356] 为了生成MSA表示1010中的定位 (i, j) 处的嵌入,系统100可以获得定义核心MSA序列 i 中的定位 j 处的氨基酸身份的嵌入(例如,独热嵌入)。系统100还可以基于MSA序列聚类1008中与核心MSA序列 i 相对应的额外MSA序列1006中的定位 j 处的每个可能氨基酸的相对发生频率来确定可能氨基酸集合内的概率分布。然后,系统100可以通过对以下各项进行组合(例如,级联)来确定MSA表示中的定位 (i, j) 处的嵌入:(i)定义核心MSA序列 i 中的定位 j 处的氨基酸身份的嵌入,以及(ii)与核心MSA序列 i 中的定位 j 相对应的可能氨基酸内的概率分布。

[0357] 在一些情况下,核心MSA序列中的一个或多个的(基准真值)蛋白质结构可能是已知的。特别地,对于核心MSA序列中的一个或多个,核心MSA序列中的氨基酸中的键之间的扭转角(例如, ω 角、 Φ 角、 Ψ 角等)的值可能是已知的。如果核心MSA序列 i 中的氨基酸的扭转角值是已知的,则系统100可以至少部分基于核心MSA序列 i 中的氨基酸 j 的扭转角值生成MSA表示中的定位 (i, j) 处的嵌入。例如,系统可以使用一个或多个神经网络层来生成扭转角值的嵌入,并且将扭转角值的嵌入级联到MSA表示中的定位 (i, j) 处的嵌入。

[0358] 图19示出了用于为蛋白质中的每对氨基酸生成(初始化)相应的成对嵌入112的示例过程1100。参考图1描述的蛋白质结构预测系统100可以实施过程1100的操作。

[0359] 系统100可以使用蛋白质的MSA表示1102来生成成对嵌入112。参考图9和图18更详细地描述了生成蛋白质的MSA表示。系统100可以基于蛋白质中的每条氨基酸链的相应MSA表示来生成蛋白质的MSA表示1102。为了为蛋白质中的氨基酸链生成MSA表示,系统100可以使用比用于生成参考图9描述的MSA表示110更多的MSA序列(例如,多一个数量级)。因此,由系统100用来生成成对嵌入112的MSA表示1102可以比参考图9描述的MSA表示110具有更多的行(例如,多一个数量级)。在一些实施方式中,系统100可以使用参考图18描述的额外MSA序列1006来生成MSA表示1102。

[0360] 在生成MSA表示1102之后,系统100处理MSA表示1102以(例如,通过对MSA表示1102应用外积均值运算,并且将成对嵌入1104标识为外积均值运算的结果)从MSA表示1102生成成对嵌入1104。

[0361] 系统100使用嵌入神经网络1106来处理MSA表示1102和成对嵌入1104。嵌入神经网络1106可以通过在MSA表示1102与成对嵌入1104之间共享信息来更新MSA表示1102和成对嵌入1104。更具体地,嵌入神经网络1106可以在基于成对嵌入1104来更新MSA表示1102与基于MSA表示1102来更新成对嵌入1104之间交替。

[0362] 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,使用逐行和逐列的自注意力块来更新成对嵌入1104和MSA表示1102的架构。

[0363] 在一些实施方式中,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表示的每列中的嵌入,以生成当前MSA表示的每列的更新嵌入。为了实施全局逐列自注意力,自注意力块可以为列中的每个嵌入生成相应的查询嵌入,然后对查询嵌入求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,在每列的嵌入数量方面)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。

[0364] 在使用嵌入神经网络1106来更新成对嵌入1104和MSA表示1102之后,系统100可以将成对嵌入112标识为由嵌入神经网络1106生成的更新的成对嵌入。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来使用它。

[0365] 作为生成成对嵌入112的一部分,系统100可以在与蛋白质中的每对氨基酸相对应的相应成对嵌入中包括相对定位编码信息。系统可以通过以下操作而将包括在同一氨基酸链中的一对氨基酸的相对定位编码信息包括在对应的成对嵌入中:计算表示在氨基酸链中将该对氨基酸分隔开的氨基酸的数量的带符号差值;将结果剪辑到预定义区间;使用独热编码向量来表示剪辑的值;对独热编码向量应用线性变换;以及将线性变换的结果加到对应的成对嵌入中。系统可以通过将默认编码向量加到对应的成对嵌入而将未被包括在同一氨基酸链中的一对氨基酸的相对定位编码信息包括在对应的成对嵌入中,该对应的成对嵌入指示该对氨基酸未被包括在同一氨基酸链中。

[0366] 系统100还可以至少部分基于一个或多个模板序列1110的集合来生成成对嵌入112。每个模板序列1110是蛋白质中的氨基酸链的MSA序列,其中模板序列1110的折叠结构是已知的,例如,通过物理实验而已知。

[0367] 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包括与模板序列中的每对氨基酸相对应的相应嵌入,例如,使得长度为 n (即,具有 n 个氨基酸)的模板序列的模板表示1112可以被表示为 $n \times n$ 嵌入阵列。系统100基于例如以下各项来生成模板序列1110的模板表示1112中的定位 (i, j) 处的嵌入:(i)表示模板序列中的定位 i 和定位 j 处的氨基酸身份的相应嵌入(例如,独热嵌入),(ii)由模板序列(即,模板序列的折叠结构)中的定位 i 和 j 处的氨基酸中的相应碳 α 原子的空间定位差值所定义的单位向量,其中,单位向量是在氨基酸 i 或氨基酸 j 的参考系中计算的,以及(iii)模板序列中的定位 i 和 j 处的氨基酸中的相应碳 α 原子的空间定位之间的距离的离散化/分格化表示。

[0368] 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相对应的相应的更新的模板表示1116。模板更新块可以包括例如逐行自注意力块(例如,其更新模板表示的每行中的嵌入)、逐列自注意力块(例如,其更新模板表示的每列中的嵌入)和过渡块(例如,其对模板表示中的每个嵌入应用一个或多个神经网络层)。

[0369] 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来更新成对嵌入112。例如,系统100可以使用对更新的模板表示1116的对应的定位 (i, j) 处的嵌入的“交叉注意力”来更新每个定位 (i, j) 处的相应成对嵌入112。在更新定位 (i, j) 处的成对嵌

入112的交叉注意力操作中,从定位(i,j)处的成对嵌入中生成查询嵌入,并且从更新的模板表示1116的对应的定位(i,j)处的嵌入中生成关键字和值嵌入。

[0370] 使用模板序列1110来更新成对嵌入112使系统100能够利用表征进化相关模板序列1110的蛋白质结构的信息来丰富成对嵌入,从而增强成对嵌入112的信息内容,并且提高使用成对嵌入预测的蛋白质结构的准确度。

[0371] 本说明书结合系统和计算机程序组件来使用术语“配置”。对于一个或多个计算机的系统而言,被配置为执行特定的操作或动作意味着该系统已经在其上安装了软件、固件、硬件或其组合,这些软件、固件、硬件或其组合在操作中使得该系统执行这些操作或动作。对于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,当由这些指令数据处理装置执行时,使得该装置执行这些操作或动作。

[0372] 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(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例可以被实现为一个或多个计算机程序,即,编码在有形非暂时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器设备、或者它们中的一个或多个的组合。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,机器生成的电信号、光信号或电磁信号)上,生成传播信号是为了对信息进行编码以传输到合适的接收器装置供数据处理装置执行。

[0373] 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,例如包括可编程处理器、计算机、或者多个处理器或计算机。该装置也可以是或者还包括专用逻辑电路,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、或者它们中的一个或多个的组的代码。

[0374] 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,可以以任何形式的编程语言来编写,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来部署,包括作为独立程序,或者作为模块、组件、子例程或适用于计算环境的其他单元。程序可以但不需要与文件系统中的文件相对应。程序可以被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,被存储在专用于所讨论的程序的单个文件,或者被存储在多个协调文件(例如,存储一个或多个模块、子程序或代码部分的文件)中。计算机程序可以被部署为在位于一个地点的一个计算机或多个计算机上执行,或者在分布在多个地点并通过数据通信网络互连的一个计算机或多个计算机上执行。

[0375] 在本说明书中,术语“引擎”被广泛用于指代被编程为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中的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下,一个或多个计算机将专用于特定的引擎;在其他情况下,可以在相同的一个或多个计算机上安装和运行多个引擎。

[0376] 本说明书中描述的过程和逻辑流可以由执行一个或多个计算机程序的一个或多个可编程计算机来执行,以通过对输入数据进行操作并且生成输出来执行功能。这些过程

和逻辑流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。

[0377] 适于执行计算机程序的计算机可以基于通用微处理器或专用微处理器,或者两者,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中接收指令和数据。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一个或多个用于存储指令和数据的存储器设备。中央处理单元和存储器可以由专用逻辑电路补充或结合到专用逻辑电路中。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),以从其接收数据,或者向其发送数据,或两者兼有。然而,计算机不需要具有这种设备。此外,计算机可以被嵌入另一设备中,例如,移动电话、个人数字助理(personal digital assistant,PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(Global Positioning System,GPS)接收器或便携式存储设备(例如,通用串行总线(universal serial bus,USB)闪存驱动器),仅举几例。

[0378] 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,包括例如:半导体存储器设备,例如,EPROM、EEPROM和闪存设备;磁盘,例如,内置硬盘或可移动磁盘;磁光盘;以及DVD-ROM光盘。

[0379] 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)的计算机上实施。也可以使用其他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官反馈,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,包括声学输入、语音输入或触觉输入。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进行交互;例如,通过响应于从网络浏览器接收的请求而向用户设备上的网络浏览器发送网页来与用户进行交互。此外,计算机可以通过向个人设备(例如,运行消息传递应用的智能电话)发送文本消息或其他形式的消息并且从用户接收响应消息来与用户进行交互。

[0380] 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)的共有部分和计算密集型部分。

[0381] 机器学习模型可以使用机器学习框架(例如,TensorFlow框架、“微软认知工具包”框架、Apache Singa框架或Apache MXNet框架)来实施和部署。

[0382] 本说明书中描述的主题的实施例可以在计算系统中实施,计算系统包括后端组件(例如,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、网络浏览器或用户与本说明书中描述的主题的实施方式进行交互所可以通过的app的客户端计算机),或者一个或多个这种后端组件、中间件组件或前端组件的任何组合。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(local area network,LAN)和广域网(wide area network,WAN)(例如,互联网)。

[0383] 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相应的计算机上运行并且彼此具有客户端-服务端关系的计算机程序而生成的。在一些实施例中,服务端将数据(例如,HTML页

面) 发送到用户设备, 例如, 以便向与充当客户端的设备进行交互的用户显示数据并且从该用户接收用户输入。在用户设备处生成的数据 (例如, 用户交互的结果) 可以在服务器处从该设备接收。

[0384] 虽然本说明书包含许多具体的实施细节, 但是这些不应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。本说明书中在分开的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施。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分开实施或以任何合适的子组合实施。此外, 尽管特征可能在上文被描述为在某些组合中起作用, 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指向子组合或子组合的变体。

[0385] 类似地, 尽管操作以特定次序在附图中示出和在权利要求中叙述, 但这不应被理解为要求这些操作以所示的特定次序或顺序次序执行, 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种分离, 并且应当理解, 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打包到多个软件产品中。

[0386] 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, 权利要求中列举的动作可以以不同的次序执行, 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次序或顺序次序来获得期望的结果。在一些情况下, 多任务和并行处理可能是有利的。

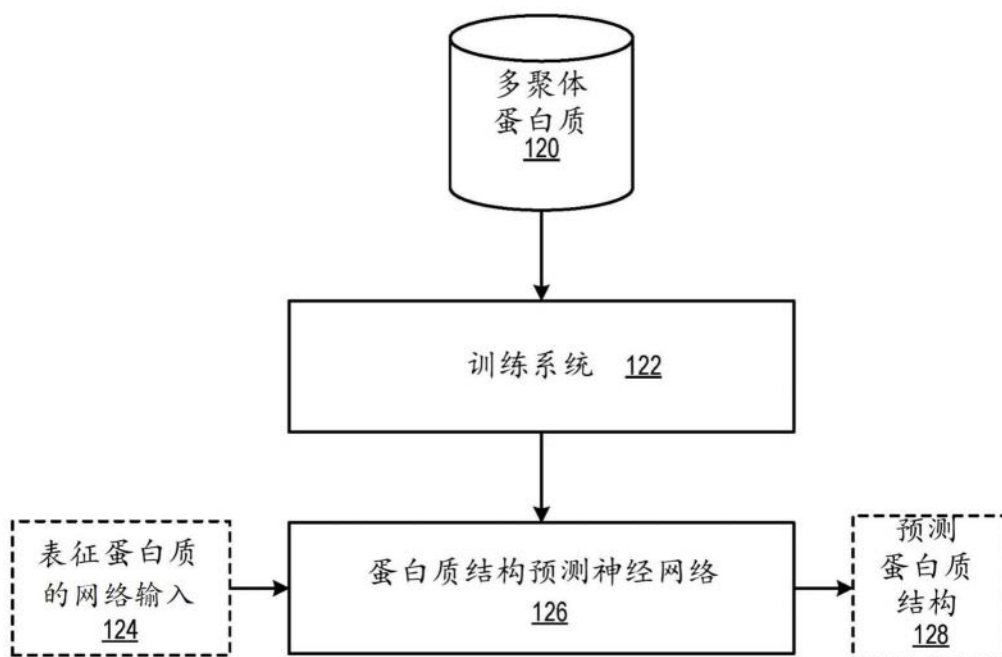


图1

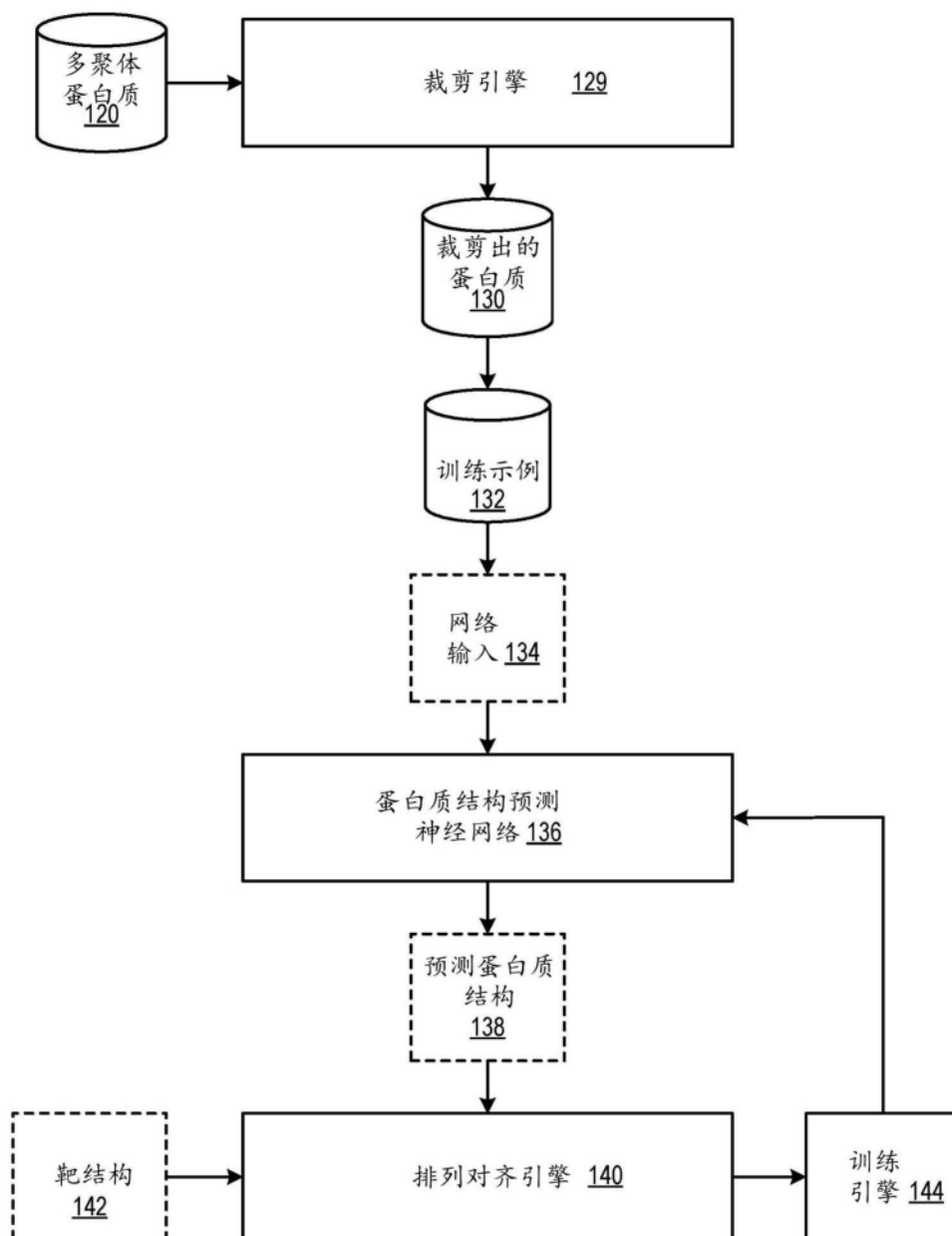
训练系统 122

图2

聚焦多样性的裁剪

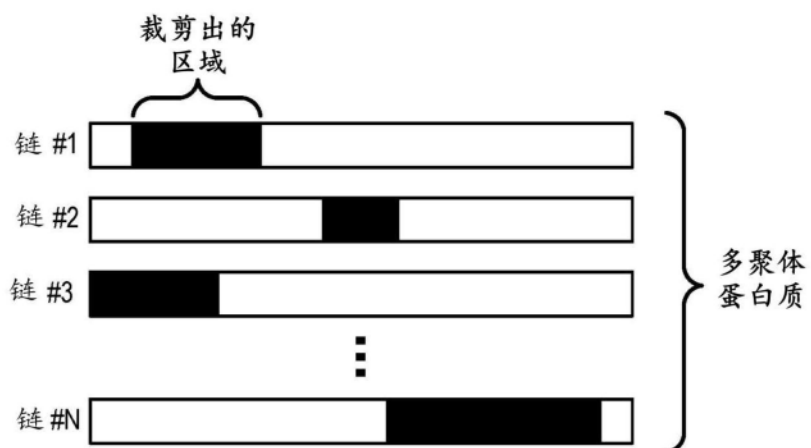


图3A

聚焦界面的裁剪

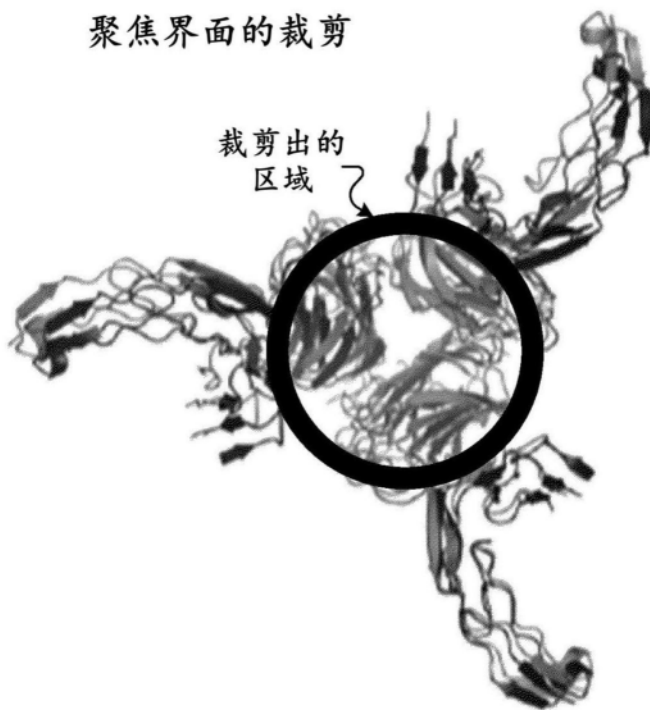


图3B

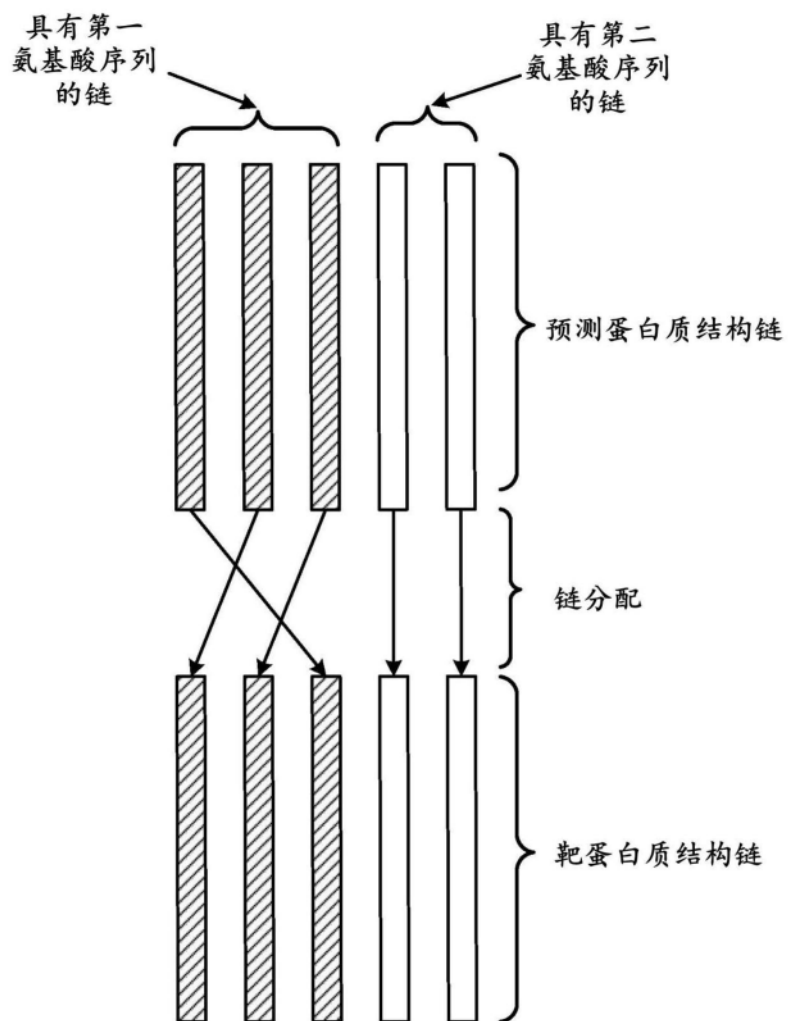


图4

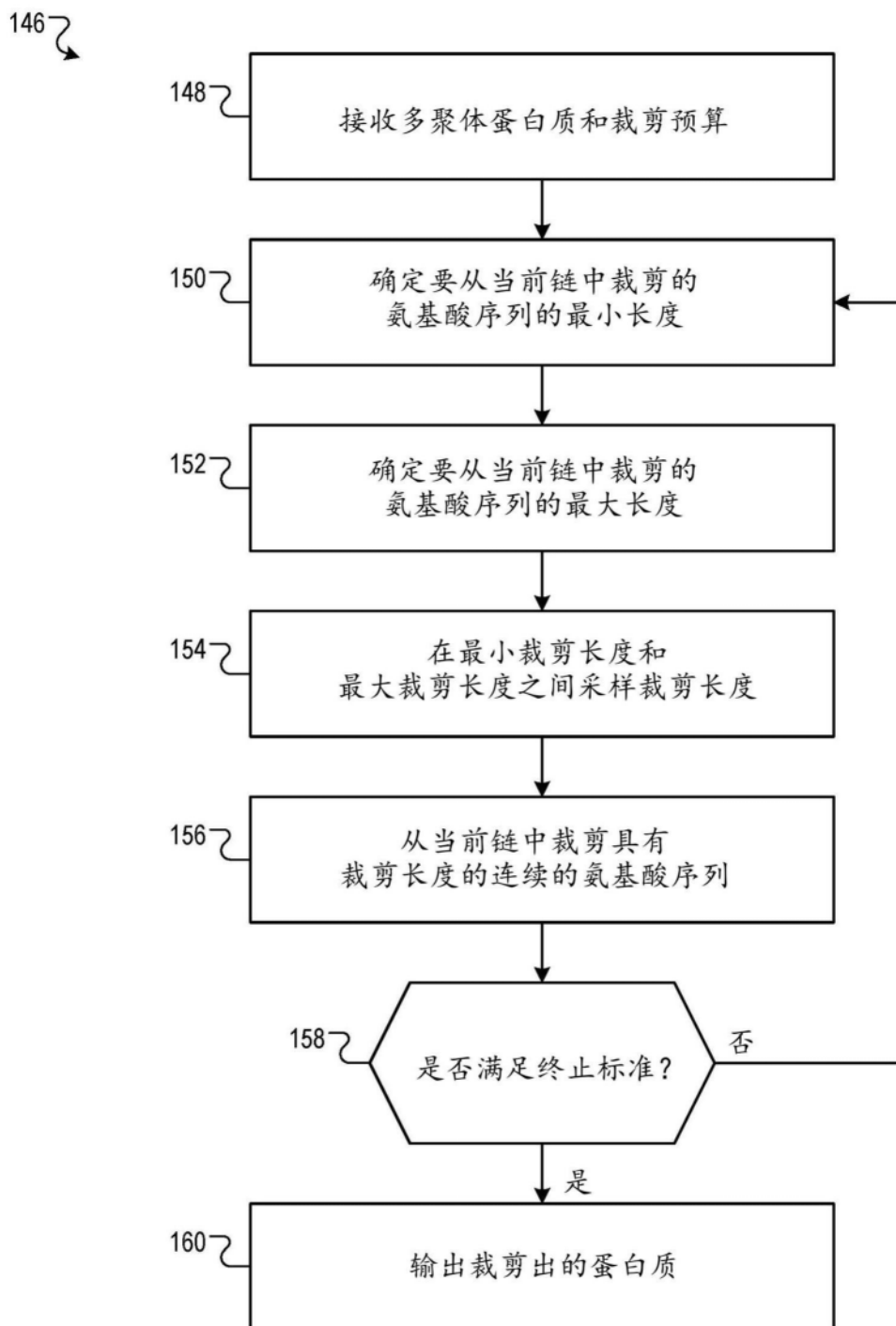


图5

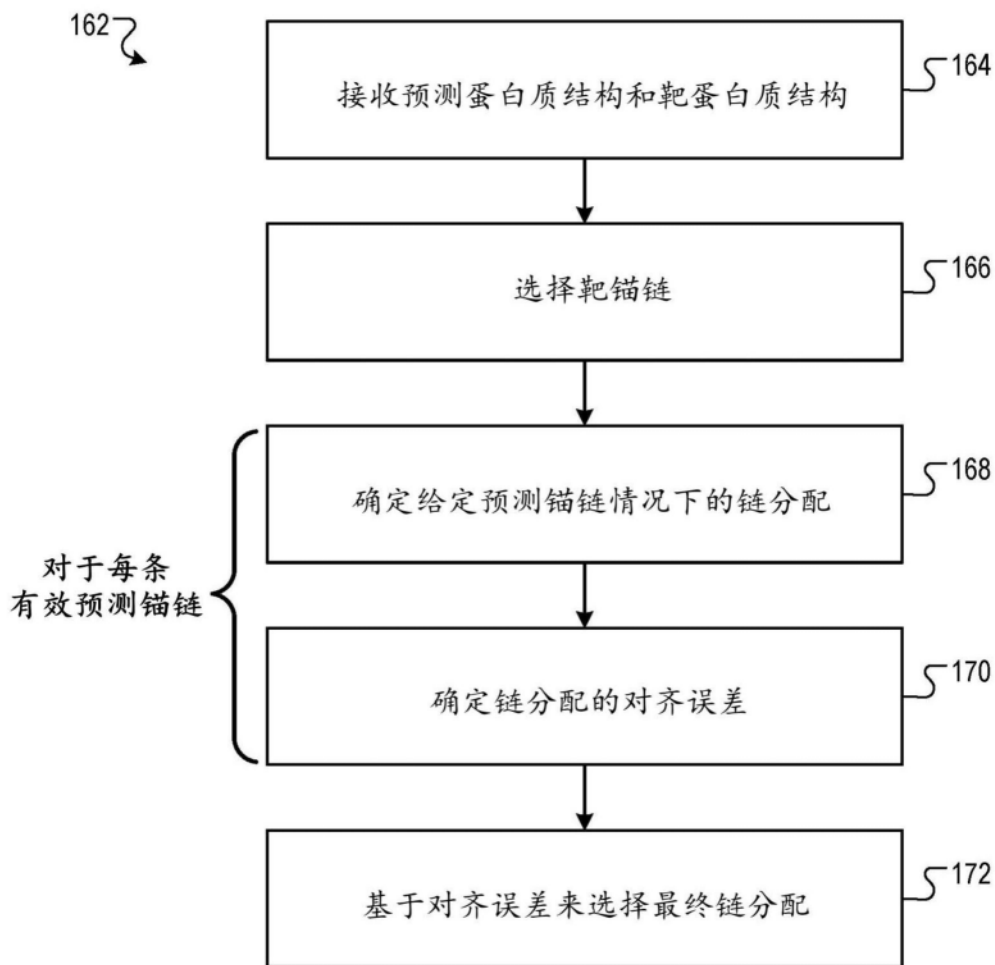


图6

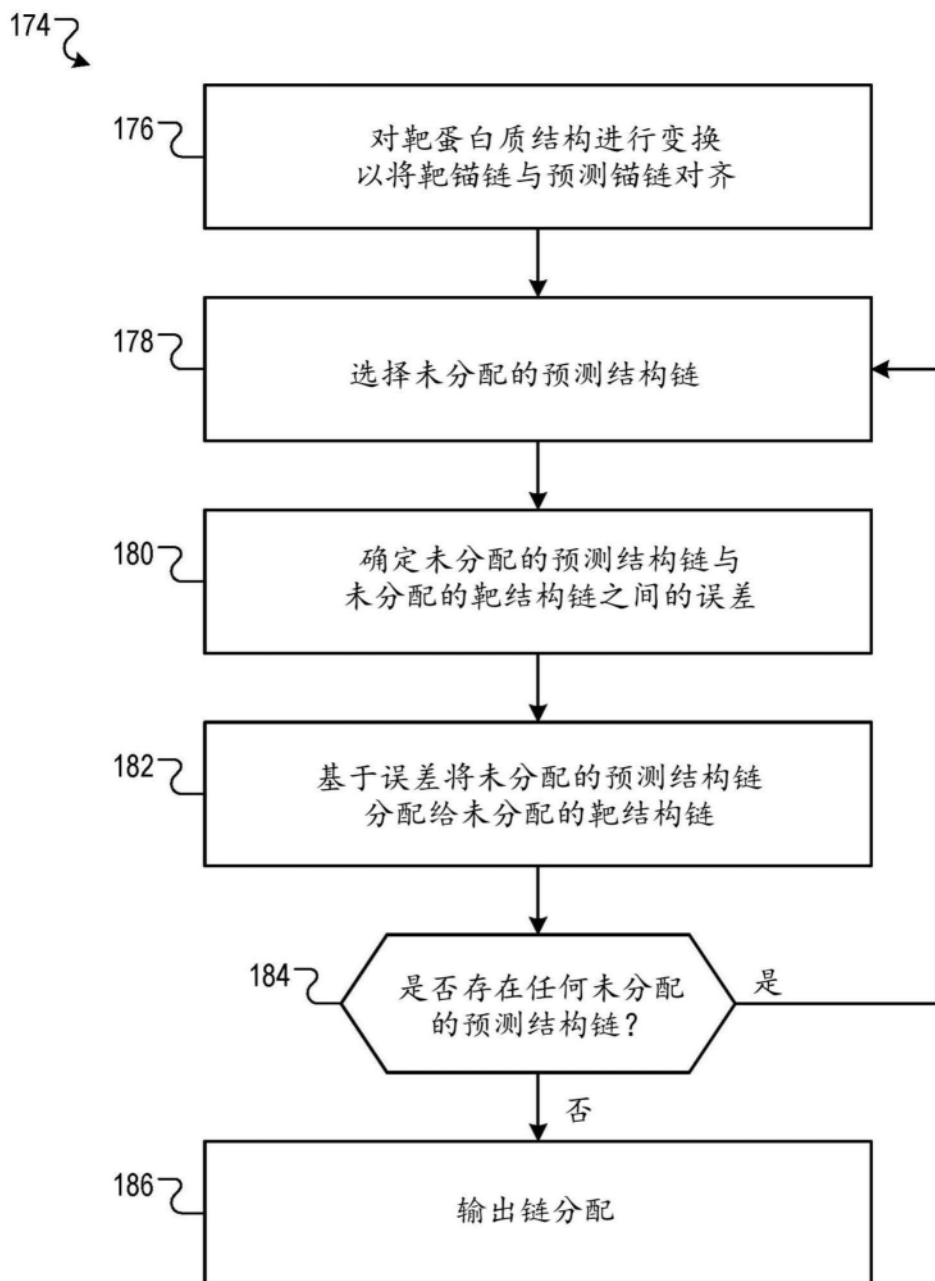


图7

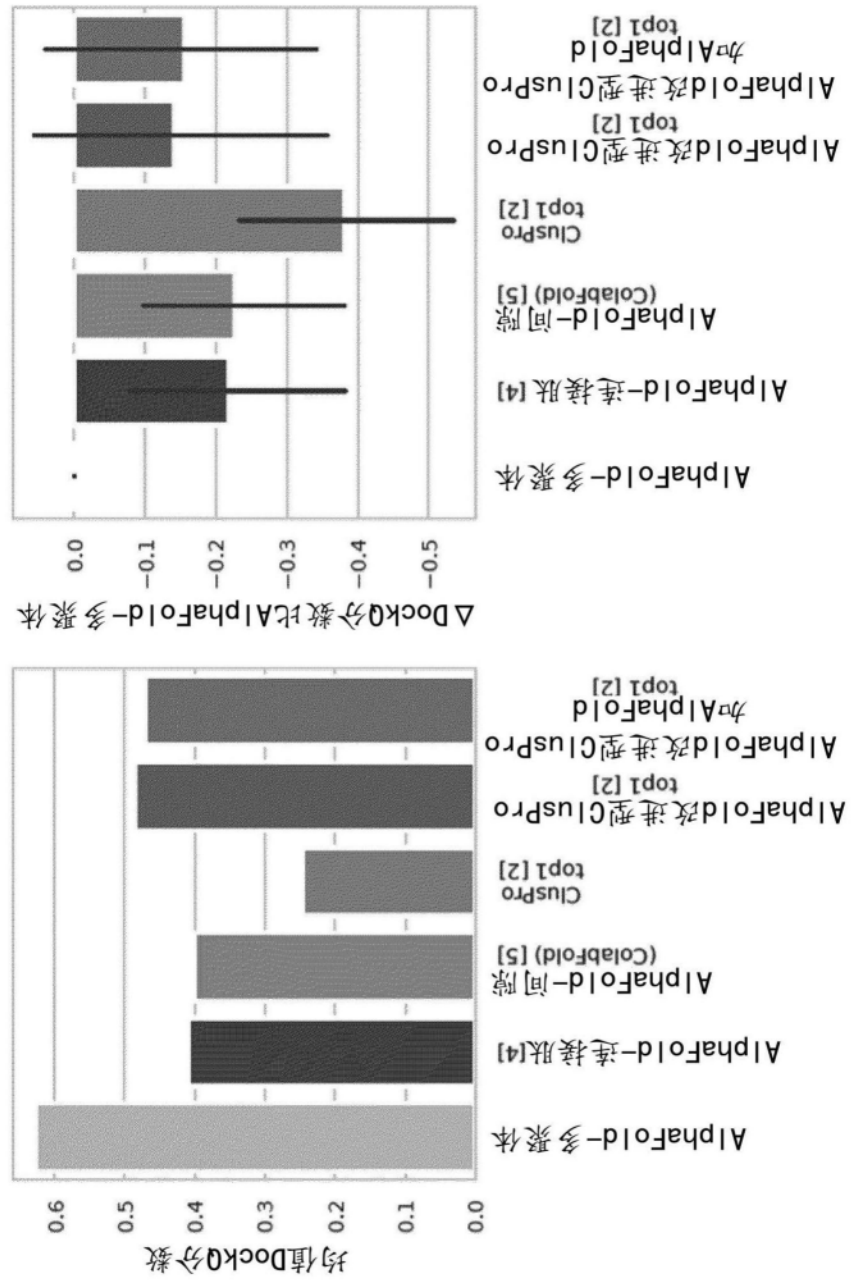


图8

蛋白质结构预测系统

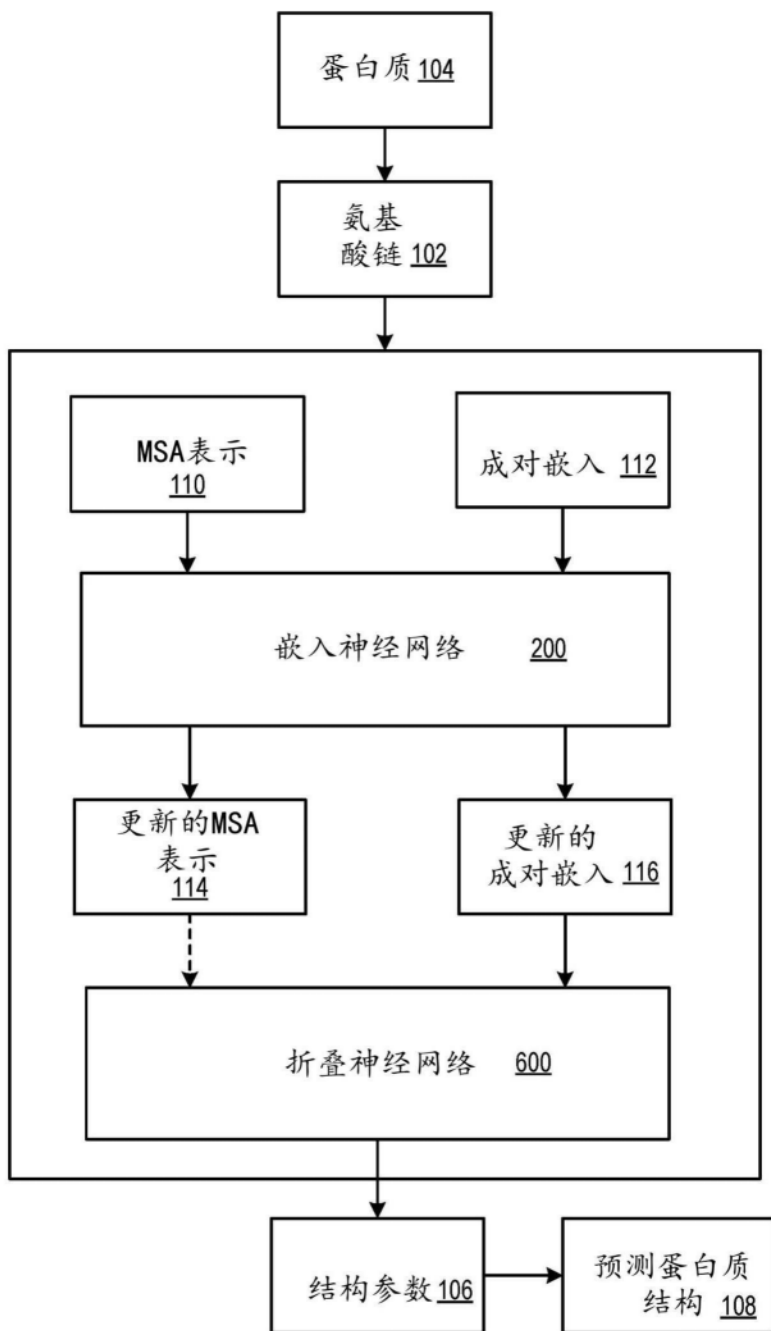
100

图9

嵌入神经网络

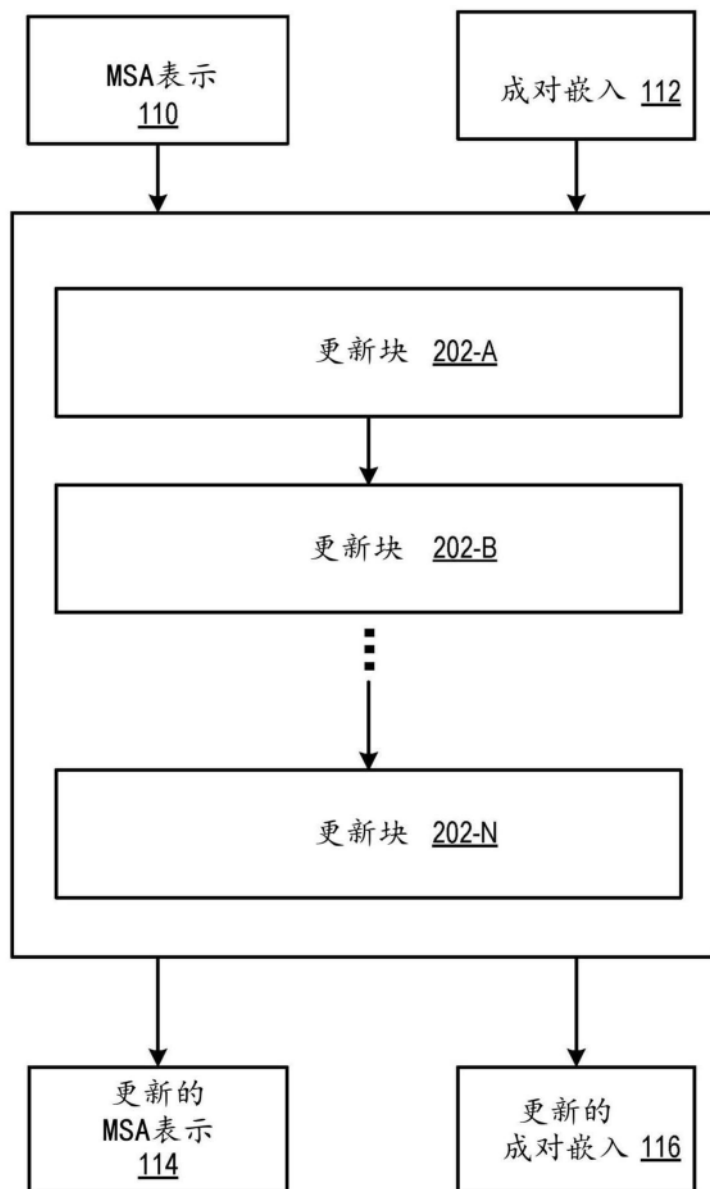
200

图10

嵌入神经网络更新块

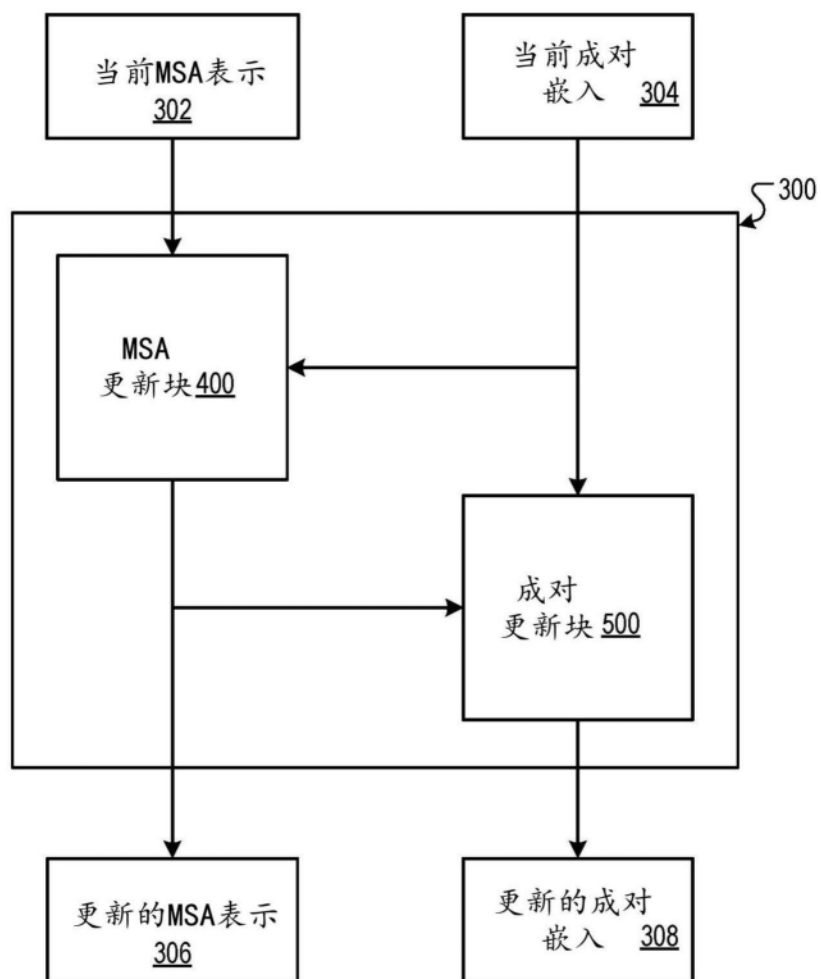


图11

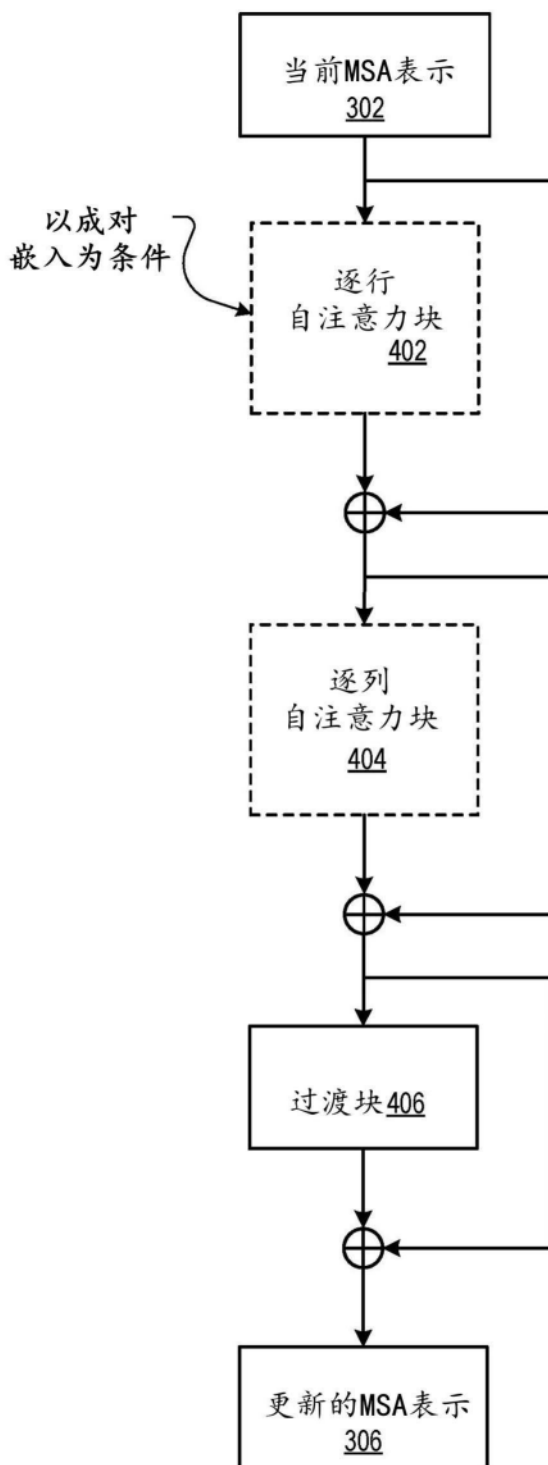
MSA更新块 **400**

图12

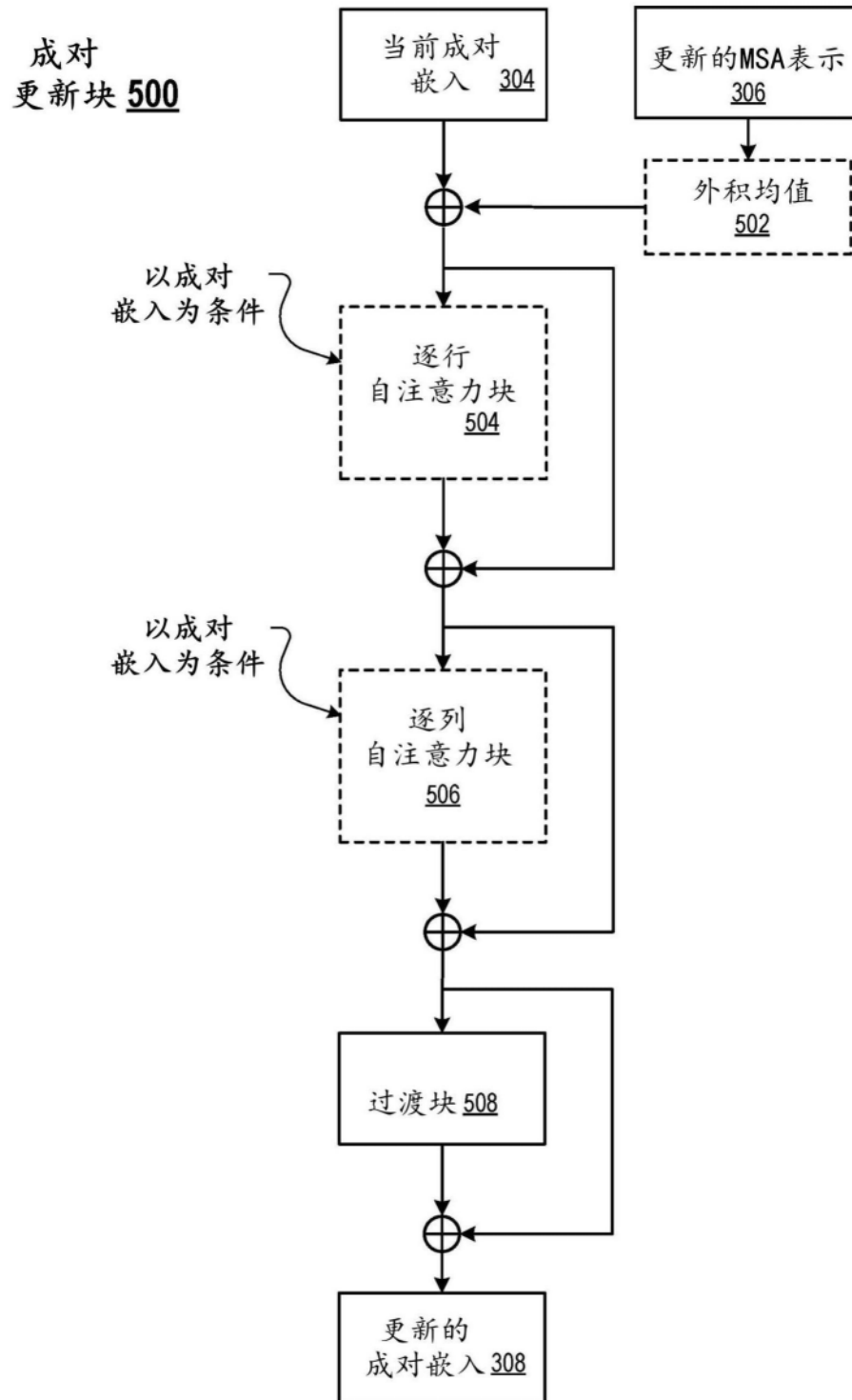


图13

折叠神经网络 600

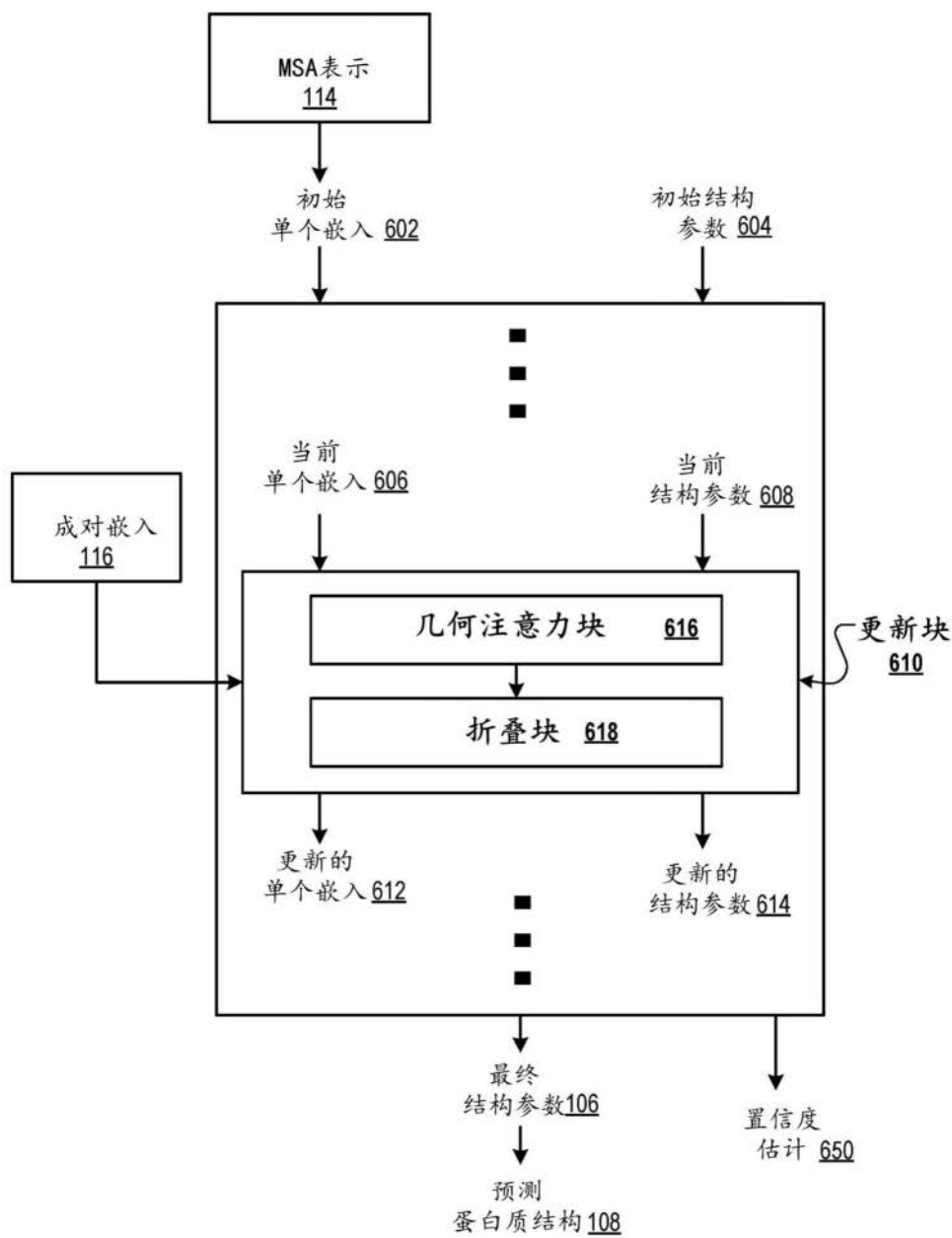


图14

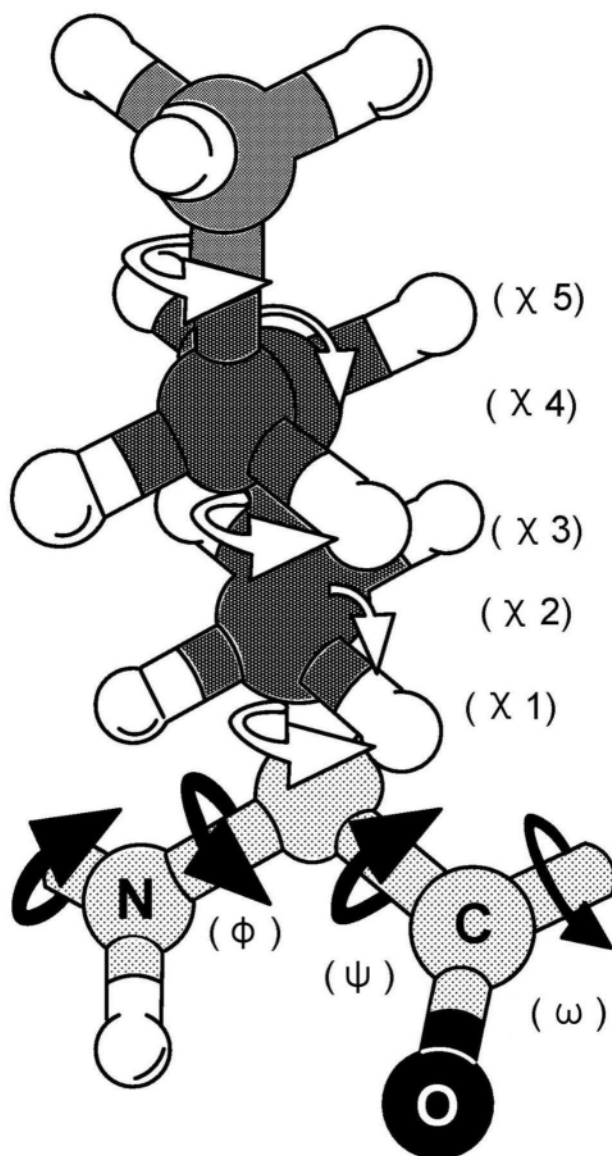


图15

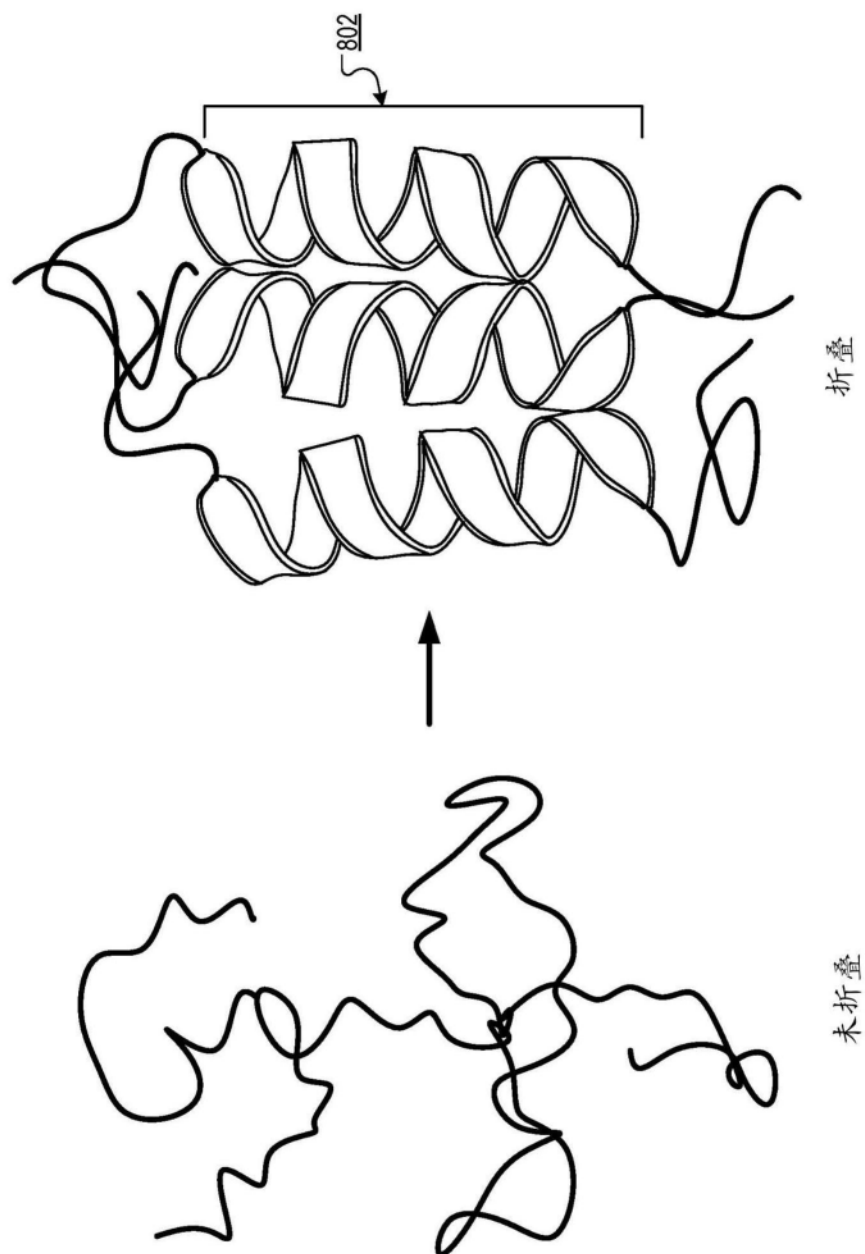


图16

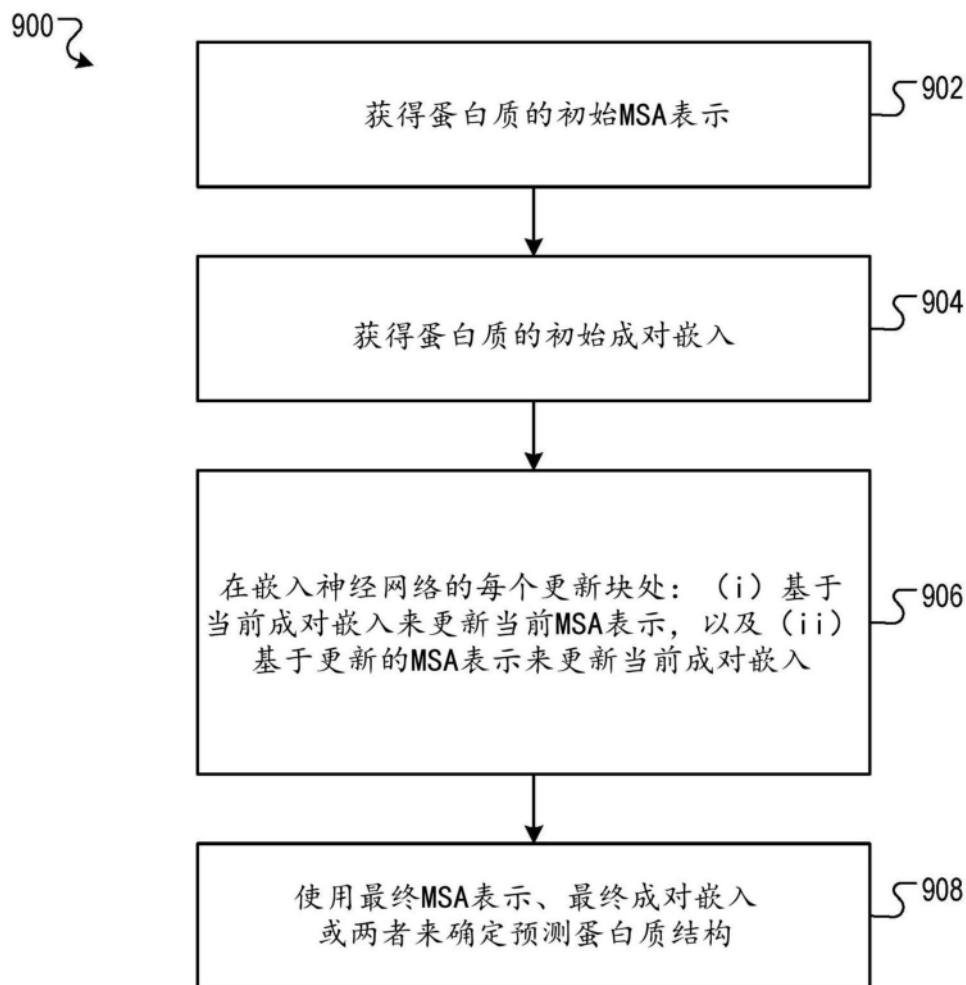


图17

为蛋白质中的氨基酸链生成MSA表示

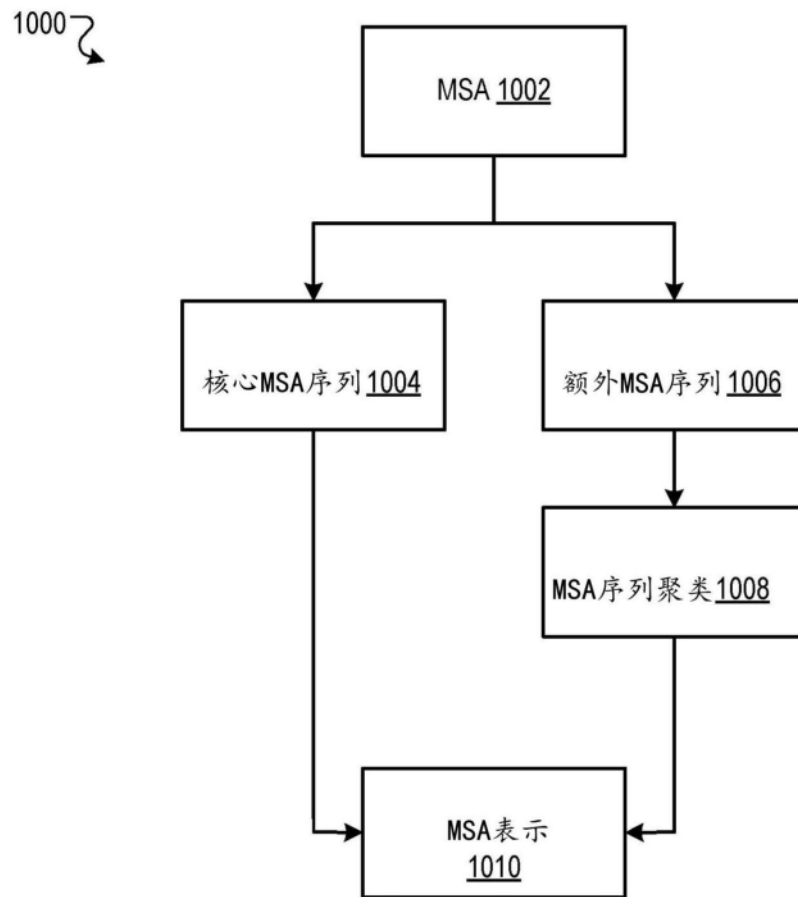


图18

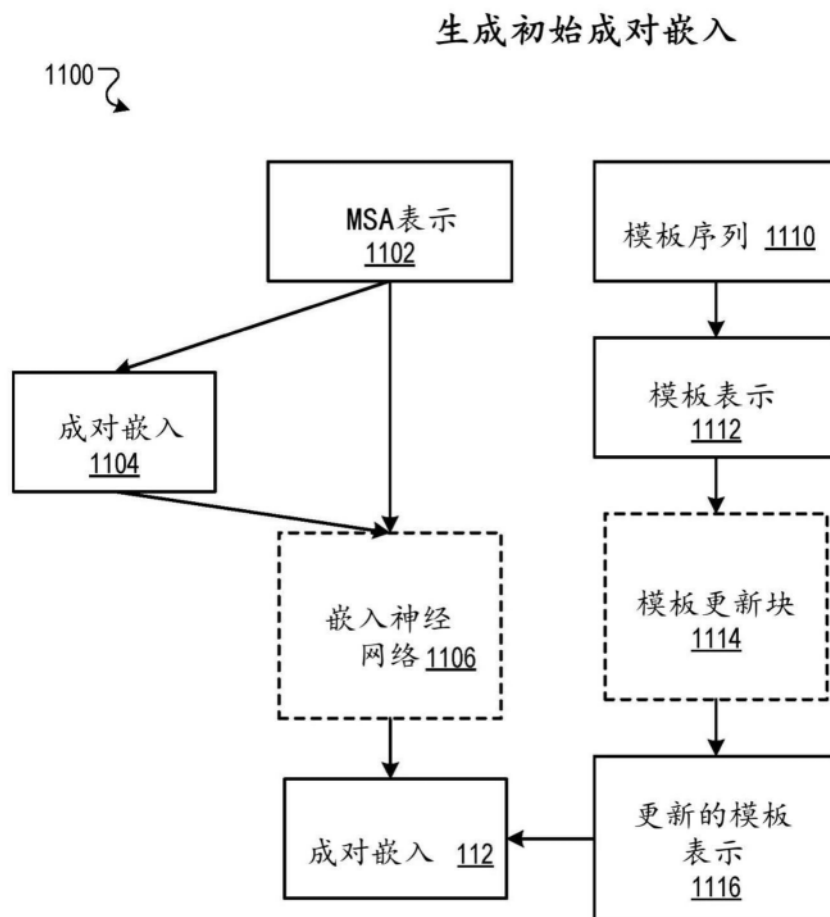


图19