(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2021-523479
(P2021-523479A)
(43)公表日　令和3年9月2日(2021.9.2)

(51)Int.Cl.		F I			テーマコード（参考）
G16B 30/20	(2019.01)	G16B	30/20		4B029
C12Q 1/68	(2018.01)	C12Q	1/68		4B063
C12M 1/34	(2006.01)	C12M	1/34	Z	

審査請求 未請求　予備審査請求 未請求　（全 59 頁）

(21)出願番号　特願2020-564123 (P2020-564123)
(86)(22)出願日　令和1年5月13日 (2019.5.13)
(85)翻訳文提出日　令和3年1月5日 (2021.1.5)
(86)国際出願番号　PCT/US2019/032065
(87)国際公開番号　WO2019/222120
(87)国際公開日　令和1年11月21日 (2019.11.21)
(31)優先権主張番号　62/671,884
(32)優先日　平成30年5月15日 (2018.5.15)
(33)優先権主張国・地域又は機関
米国 (US)
(31)優先権主張番号　62/671,260
(32)優先日　平成30年5月14日 (2018.5.14)
(33)優先権主張国・地域又は機関
米国 (US)

(71)出願人　516144164
クアンタム－エスアイ　インコーポレイテッド
ＱＵＡＮＴＵＭ－ＳＩ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国　０６４３７　コネチカット州　ギルフォード　オールド　ウィットフィールド　ストリート　５３０
(74)代理人　100105957
弁理士　恩田　誠
(74)代理人　100068755
弁理士　恩田　博宣
(74)代理人　100142907
弁理士　本田　淳

最終頁に続く

(54)【発明の名称】機械学習可能な生物学的ポリマーアセンブリ

(57)【要約】
本明細書には、高分子の生物学的ポリマーアセンブリを生成するための機械学習技術が記載されている。例えば、システムは、機械学習技術を使用して、生物のＤＮＡのゲノムアセンブリ、生物のＤＮＡの一部の遺伝子配列、またはタンパク質のアミノ酸配列を生成し得る。システムは、シークエンシングデバイスによって生成された生物学的ポリマー配列および配列から生成されたアセンブリにアクセスし得る。システムは、配列およびアセンブリを使用して機械学習モデルへの入力を生成し得る。システムは、入力を機械学習モデルに提供して、対応する出力を取得し得る。システムは、対応する出力を使用して、アセンブリ内の位置において生物学的ポリマーを同定し、次にアセンブリ内の位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得し得る。

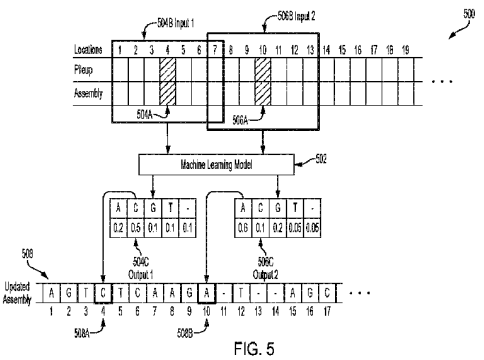

FIG. 5

【特許請求の範囲】
【請求項１】
高分子の生物学的ポリマーアセンブリを生成する方法であって、
少なくとも１つのコンピュータハードウェアプロセッサを使用して、
複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、
前記複数の生物学的ポリマー配列および前記アセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、
前記第１の入力を前記トレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、
前記トレーニングされた深層学習モデルの前記第１の出力を使用して、前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、
前記第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すように前記アセンブリを更新して、更新されたアセンブリを取得するステップとを実行するステップを含む方法。
【請求項２】
前記高分子がタンパク質を含み、前記複数の生物学的ポリマー配列が複数のアミノ酸配列を含み、前記アセンブリが個々のアセンブリ位置におけるアミノ酸を示す、請求項１に記載の方法。
【請求項３】
前記高分子が核酸を含み、前記複数の生物学的ポリマー配列が複数のヌクレオチド配列を含み、前記アセンブリが個々のアセンブリ位置におけるヌクレオチドを示す、請求項１または任意の他の先行する請求項に記載の方法。
【請求項４】
前記アセンブリは、前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、
前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、前記第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、
前記アセンブリの更新することは、前記第１のアセンブリ位置において前記第２のヌクレオチドを示すように前記アセンブリを更新することを含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項５】
前記アセンブリを更新して、前記更新されたアセンブリを取得した後、
前記複数のヌクレオチド配列を前記更新されたアセンブリに整列させるステップと、
前記複数のヌクレオチド配列および前記更新されたアセンブリを使用して、前記トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、
前記第２の入力を前記トレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、
前記トレーニングされた深層学習モデルの前記第２の出力に基づいて、前記第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、
前記第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように前記更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとをさらに含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項６】
前記複数のヌクレオチド配列を前記アセンブリに整列させることをさらに含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項７】
前記複数のヌクレオチド配列が少なくとも９個のヌクレオチド配列を含む、請求項６ま

たは任意の他の先行する請求項に記載の方法。
【請求項８】
前記トレーニングされた深層学習モデルへの前記第１の入力を生成するステップは、
前記第１の複数のアセンブリ位置を選択すること、
選択された第１の複数のアセンブリ位置に基づいて前記第１の入力を生成することを含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項９】
前記アセンブリ内の前記第１の複数の位置を選択することは、
前記アセンブリが前記第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、
決定された尤度を使用して、前記第１の複数のアセンブリ位置を選択することを含む、請求項８または任意の他の先行する請求項に記載の方法。
【請求項１０】
前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、前記複数のヌクレオチド配列の個々の１つを前記アセンブリと比較することを含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項１１】
前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定するために前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置の各々における複数のヌクレオチドの各々に関して、
ヌクレオチドがその位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定すること、
前記アセンブリがその位置においてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、
前記カウントと前記参照値との差異を示すエラー値を決定すること、
前記参照値および前記エラー値を前記第１の入力に含ませることを含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項１２】
前記アセンブリがその位置においてヌクレオチドを示しているすかどうかに基づいて前記参照値を決定することは、
前記アセンブリがその位置においてヌクレオチドを示している場合、前記参照値が第１の値であると決定すること、
前記アセンブリがその位置においてヌクレオチドを示していない場合、前記参照値が第２の値であると決定することを含む、請求項１１または任意の他の先行する請求項に記載の方法。
【請求項１３】
前記第１の値は、前記複数のヌクレオチド配列の数であり、
前記第２の値は０である、請求項１２または任意の他の先行する請求項に記載の方法。
【請求項１４】
前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、
第１の列は、第１のアセンブリ位置において複数のヌクレオチドに対して決定された参照値およびエラー値を保持し、
第２の列は、前記第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する、請求項１１または任意の他の先行する請求項に記載の方法。
【請求項１５】
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置が、前記第１のア

センブリ位置とは別個の少なくとも２つのアセンブリ位置を含む、請求項１１または任意の他の先行する請求項に記載の方法。
【請求項１６】
１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度が、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置において存在する尤度を含み、
前記第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを決定することを含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項１７】
前記複数のヌクレオチド配列から前記アセンブリを生成することをさらに含む、請求項３または任意の他の先行する請求項に記載の方法。
【請求項１８】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、前記アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む、請求項１７または任意の他の先行する請求項に記載の方法。
【請求項１９】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを前記複数のヌクレオチド配列に適用することを含む、請求項１７または任意の他の先行する請求項に記載の方法。
【請求項２０】
参照高分子のシークエンシングから取得された生物学的ポリマー配列と、前記参照高分子の所定のアセンブリとを含むトレーニングデータにアクセスするステップと、
前記トレーニングデータを使用して深層学習モデルをトレーニングして、トレーニングされた深層学習モデルを取得するステップとをさらに含む、請求項１または任意の他の先行する請求項に記載の方法。
【請求項２１】
前記参照高分子は前記高分子とは異なる、請求項２０または任意の他の先行する請求項に記載の方法。
【請求項２２】
深層学習モデルが畳み込みニューラルネットワーク（ＣＮＮ）を含む、請求項１または任意の他の先行する請求項に記載の方法。
【請求項２３】
高分子の生物学的ポリマーアセンブリを生成するためのシステムであって、
少なくとも１つのコンピュータハードウェアプロセッサと、
命令を格納する少なくとも１つの非一時的なコンピュータ可読記憶媒体とを備え、前記命令は、前記少なくとも１つのコンピュータハードウェアプロセッサによる実行時に、前記少なくとも１つのコンピュータハードウェアプロセッサに、
複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、
前記複数の生物学的ポリマー配列および前記アセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、
前記第１の入力を前記トレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、
前記トレーニングされた深層学習モデルの前記第１の出力を使用して、前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、

前記第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すように前記アセンブリを更新して、更新されたアセンブリを取得するステップとを実行させる、システム。

【請求項２４】

前記高分子がタンパク質を含み、前記複数の生物学的ポリマー配列が複数のアミノ酸配列を含み、前記アセンブリが個々のアセンブリ位置におけるアミノ酸を示す、請求項２３に記載のシステム。

【請求項２５】

前記高分子が核酸を含み、前記複数の生物学的ポリマー配列が複数のヌクレオチド配列を含み、前記アセンブリが個々のアセンブリ位置におけるヌクレオチドを示す、請求項２３または任意の他の先行する請求項に記載のシステム。

【請求項２６】

前記アセンブリは、前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、

前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、前記第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、

前記アセンブリを更新することは、前記第１のアセンブリ位置において前記第２のヌクレオチドを示すように前記アセンブリを更新することを含む、請求項２５または任意の他の先行する請求項に記載のシステム。

【請求項２７】

前記命令は、前記アセンブリを更新して、前記更新されたアセンブリを取得した後、前記少なくとも１つのコンピュータハードウェアプロセッサに、

前記複数のヌクレオチド配列を前記更新されたアセンブリに整列させるステップと、

前記複数のヌクレオチド配列および前記更新されたアセンブリを使用して、前記トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、

前記第２の入力を前記トレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、

前記トレーニングされた深層学習モデルの前記第２の出力に基づいて、前記第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、

前記第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように前記更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとをさらに実行させる、請求項２５または任意の他の先行する請求項に記載のシステム。

【請求項２８】

前記命令は、前記少なくとも１つのコンピュータハードウェアプロセッサに、前記複数のヌクレオチド配列を前記アセンブリに整列させるステップを実行させる請求項２５または任意の他の先行する請求項に記載のシステム。

【請求項２９】

前記複数のヌクレオチド配列が少なくとも９個のヌクレオチド配列を含む、請求項２８または任意の他の先行する請求項に記載のシステム。

【請求項３０】

前記トレーニングされた深層学習モデルへの前記第１の入力を生成するステップは、

前記第１の複数のアセンブリ位置を選択すること、

選択された第１の複数のアセンブリ位置に基づいて前記第１の入力を生成することを含む、請求項２５または任意の他の先行する請求項に記載のシステム。

【請求項３１】

前記アセンブリ内の前記第１の複数の位置を選択することは、

前記アセンブリが前記第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、

決定された尤度を使用して、前記第１の複数のアセンブリ位置を選択することを含む、

請求項３０または任意の他の先行する請求項に記載のシステム。
【請求項３２】
前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、前記複数のヌクレオチド配列の個々の１つを前記アセンブリと比較することを含む、請求項２５または任意の他の先行する請求項に記載のシステム。
【請求項３３】
前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定するためにトレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置の各々における複数のヌクレオチドの各々に関して、
ヌクレオチドがその位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定すること、
前記アセンブリがその位置においてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、
前記カウントと前記参照値との差異を示すエラー値を決定すること、
前記参照値および前記エラー値を前記第１の入力に含ませることを含む、請求項２５または任意の他の先行する請求項に記載のシステム。
【請求項３４】
前記アセンブリがその位置においてヌクレオチドを示すかどうかに基づいて前記参照値を決定することは、
前記アセンブリがその位置においてヌクレオチドを示している場合、前記参照値が第１の値であると決定すること、
前記アセンブリがその位置においてヌクレオチドを示していない場合、前記参照値が第２の値であると決定することを含む、請求項３３または任意の他の先行する請求項に記載のシステム。
【請求項３５】
前記第１の値は、前記複数のヌクレオチド配列の数であり、
前記第２の値は０である、請求項３４または他の先行する請求項に記載のシステム。
【請求項３６】
前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、
第１の列は、第１のアセンブリ位置において複数のヌクレオチドに対して決定された参照値およびエラー値を保持し、
第２の列は、前記第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する、請求項３３または任意の他の先行する請求項に記載のシステム。
【請求項３７】
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置が、前記第１のアセンブリ位置とは別の少なくとも２つのアセンブリ位置を含む、請求項３３または任意の他の先行する請求項に記載のシステム。
【請求項３８】
１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度が、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置において存在する尤度を含み、
前記第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複

数のヌクレオチドのうちの第１のヌクレオチドであることを決定することを含む、請求項２５または任意の他の先行する請求項に記載のシステム。
【請求項３９】
前記命令は、前記少なくとも１つのコンピュータハードウェアプロセッサに、前記複数のヌクレオチド配列から前記アセンブリを生成することを実行させる、請求項２５または任意の他の先行する請求項に記載のシステム。
【請求項４０】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、前記アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む、請求項３９または任意の他の先行する請求項に記載のシステム。
【請求項４１】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを前記複数のヌクレオチド配列に適用することを含む、請求項３９または任意の他の先行する請求項に記載のシステム。
【請求項４２】
前記命令は、前記少なくとも１つのコンピュータハードウェアプロセッサに
参照高分子のシークエンシングから取得された生物学的ポリマー配列と、前記参照高分子の所定のアセンブリとを含むトレーニングデータにアクセスするステップと、
前記トレーニングデータを使用して深層学習モデルをトレーニングして、トレーニングされた深層学習モデルを取得するステップとをさらに実行させる、請求項２３または任意の他の先行する請求項に記載のシステム。
【請求項４３】
前記参照高分子は前記高分子とは異なる、請求項４２または任意の他の先行する請求項に記載の方法。
【請求項４４】
深層学習モデルが畳み込みニューラルネットワーク（ＣＮＮ）を含む、請求項２３または任意の他の先行する請求項に記載のシステム。
【請求項４５】
命令を格納する少なくとも１つの非一時的なコンピュータ可読記憶媒体であって、前記命令は、少なくとも１つのコンピュータハードウェアプロセッサによる実行時に、前記少なくとも１つのコンピュータハードウェアプロセッサに高分子の生物学的ポリマーアセンブリを生成する方法を実行させ、前記方法は、
複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、
前記複数の生物学的ポリマー配列および前記アセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、
前記第１の入力を前記トレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、
前記トレーニングされた深層学習モデルの前記第１の出力を使用して、前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、
前記第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すように前記アセンブリを更新して、更新されたアセンブリを取得するステップとを含む、少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項４６】
前記高分子がタンパク質を含み、前記複数の生物学的ポリマー配列が複数のアミノ酸配列を含み、前記アセンブリが個々のアセンブリ位置におけるアミノ酸を示す、請求項４５に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項４７】
前記高分子が核酸を含み、前記複数の生物学的ポリマー配列が複数のヌクレオチド配列

を含み、前記アセンブリが個々のアセンブリ位置におけるヌクレオチドを示す、請求項４５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項４８】

前記アセンブリは、前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、

前記第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、前記第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、

前記アセンブリを更新することは、前記第１のアセンブリ位置において前記第２のヌクレオチドを示すように前記アセンブリを更新することを含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項４９】

前記方法は、前記アセンブリを更新して、前記更新されたアセンブリを取得した後、

前記複数のヌクレオチド配列を前記更新されたアセンブリに整列させるステップと、

前記複数のヌクレオチド配列および前記更新されたアセンブリを使用して、前記トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、

前記第２の入力を前記トレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、

前記トレーニングされた深層学習モデルの前記第２の出力に基づいて、前記第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、

前記第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように前記更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとをさらに含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項５０】

前記方法が、前記複数のヌクレオチド配列を前記アセンブリに整列させるステップをさらに含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項５１】

前記複数のヌクレオチド配列が少なくとも９個のヌクレオチド配列を含む、請求項５０または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項５２】

前記トレーニングされた深層学習モデルへの前記第１の入力を生成することは、

前記第１の複数のアセンブリ位置を選択すること、

選択された第１の複数のアセンブリ位置に基づいて前記第１の入力を生成することを含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項５３】

前記アセンブリ内の前記第１の複数の位置を選択することは、

前記アセンブリが前記第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、

決定された尤度を使用して、前記第１の複数のアセンブリ位置を選択することを含む、請求項５２または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。

【請求項５４】

前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、前記複数のヌクレオチド配列の個々の１つを前記アセンブリと比較することを含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコン

ピュータ可読記憶媒体。
【請求項５５】
前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定するためにトレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置の各々における複数のヌクレオチドの各々に関して、
ヌクレオチドがその位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定すること、
前記アセンブリがその位置においてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、
前記カウントと前記参照値との差異を示すエラー値を決定すること、
前記参照値および前記エラー値を前記第１の入力に含ませることを含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項５６】
前記アセンブリがその位置においてヌクレオチドを示すかどうかに基づいて参照値を決定することは、
前記アセンブリがその位置においてヌクレオチドを示している場合、前記参照値が第１の値であると決定すること、
前記アセンブリがその位置においてヌクレオチドを示していない場合、前記参照値が第２の値であると決定することを含む、請求項５５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項５７】
前記第１の値は、前記複数のヌクレオチド配列の数であり、
前記第２の値は０である、請求項５６または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項５８】
前記トレーニングされた深層学習モデルに提供される前記第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、
第１の列は、第１のアセンブリ位置において複数のヌクレオチドに対して決定された参照値およびエラー値を保持し、
第２の列は、前記第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する、請求項５５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項５９】
前記第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置が、前記第１のアセンブリ位置とは別の少なくとも２つのアセンブリ位置を含む、請求項５５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６０】
１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度が、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置において存在する尤度を含み、
前記第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって前記第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを決定することを含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ

可読記憶媒体。
【請求項６１】
前記方法が、前記複数のヌクレオチド配列から前記アセンブリを生成するステップをさらに含む、請求項４７または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６２】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、前記アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む、請求項６１または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６３】
前記複数のヌクレオチド配列から前記アセンブリを生成することは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを前記複数のヌクレオチド配列に適用することを含む、請求項６１または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６４】
前記方法が
参照高分子のシークエンシングから取得された生物学的ポリマー配列と、前記参照高分子の所定のアセンブリとを含むトレーニングデータにアクセスするステップと、
前記トレーニングデータを使用して深層学習モデルをトレーニングして、トレーニングされた深層学習モデルを取得するステップとをさらに含む、請求項４５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６５】
前記参照高分子は前記高分子とは異なる、請求項６４または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【請求項６６】
深層学習モデルが畳み込みニューラルネットワーク（ＣＮＮ）を含む、請求項４５または任意の他の先行する請求項に記載の少なくとも１つの非一時的なコンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、高分子（例えば、核酸またはタンパク質）の生物学的ポリマー（例えば、ゲノムアセンブリ、ヌクレオチド配列、またはタンパク質配列）のアセンブリを生成することに関する。
【背景技術】
【０００２】
シークエンシングデバイスは、アセンブリを生成するために使用することができるシークエンシングデータを生成し得る。一例として、シークエンシングデータは、ゲノムを（全体的または部分的に）組み立てるために使用することができる生物学的サンプルからのＤＮＡのヌクレオチド配列を含み得る。別の例として、シークエンシングデータは、タンパク質配列を（全体的または部分的に）組み立てるために使用することができるアミノ酸配列を含み得る。
【発明の概要】
【０００３】
一態様によれば、高分子の生物学的ポリマーアセンブリを生成する方法が提供される。方法は、少なくとも１つのコンピュータハードウェアプロセッサを使用して、複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、複数の生物学的ポリマー配列およびアセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、

第１の入力をトレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、トレーニングされた深層学習モデルの第１の出力を使用して、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得するステップとを含む。

【０００４】

一実施形態によれば、高分子はタンパク質を含み、複数の生物学的ポリマー配列は複数のアミノ酸配列を含み、アセンブリは個々のアセンブリ位置におけるアミノ酸を示す。

一実施形態によれば、高分子は核酸を含み、複数の生物学的ポリマー配列は複数のヌクレオチド配列を含み、アセンブリは個々のアセンブリ位置におけるヌクレオチドを示す。

【０００５】

一実施形態によれば、アセンブリは、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、アセンブリを更新するステップは、第１のアセンブリ位置における第２のヌクレオチドを示すようにアセンブリを更新することを含む。

【０００６】

一実施形態によれば、方法は、アセンブリを更新して、更新されたアセンブリを取得した後、複数のヌクレオチド配列を更新されたアセンブリに整列させるステップと、複数のヌクレオチド配列および更新されたアセンブリを使用して、トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、第２の入力をトレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、トレーニングされた深層学習モデルの第２の出力に基づいて、第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとを含む。

【０００７】

一実施形態によれば、方法は、複数のヌクレオチド配列をアセンブリに整列させるステップをさらに含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも５個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも９個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも１０個のヌクレオチド配列を含む。

【０００８】

一実施形態によれば、トレーニングされた深層学習モデルへの第１の入力を生成するステップは、第１の複数のアセンブリ位置を選択すること、選択された第１の複数のアセンブリ位置に基づいて第１の入力を生成することを含む。一実施形態によれば、アセンブリ内の第１の複数のアセンブリ位置を選択することは、アセンブリが第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、および決定された尤度を使用して、第１の複数のアセンブリ位置を選択することを含む。

【０００９】

一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数のヌクレオチド配列の個々の１つをアセンブリと比較することを含む。一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成して、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定することは、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置における複数のヌクレオチドの各々に関して、ヌクレオチドがその位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定すること、アセンブリがその位置に

おいてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、カウントと参照値との間の差異を示すエラー値を決定すること、第１の入力に参照値およびエラー値を含ませることを含む。

【００１０】

一実施形態によれば、アセンブリがその位置においてヌクレオチドを示すかどうかに基づいて参照値を決定することは、アセンブリがその位置においてヌクレオチドを示している場合、参照値が第１の値であると決定すること、アセンブリがその位置においてヌクレオチドを示していない場合、参照値が第２の値であると決定することを含む。一実施形態によれば、第１の値は、複数のヌクレオチド配列の数であり、第２の値は０である。

【００１１】

一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、第１の列は、第１のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持し、第２の列は、第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する。一実施形態によれば、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置は、第１のアセンブリ位置とは別の少なくとも２つのアセンブリ位置を含む。

【００１２】

一実施形態によれば、１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度は、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置に存在する尤度を含み、第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを同定することを含む。

【００１３】

一実施形態によれば、方法は、複数のヌクレオチド配列からアセンブリを生成するステップをさらに含む。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを複数のヌクレオチド配列に適用することを含む。

【００１４】

一実施形態によれば、方法は、参照高分子のシークエンシングから取得された生物学的ポリマー配列と、参照高分子の所定のアセンブリとを含むトレーニングデータにアクセスするステップと、トレーニングデータを使用して深層学習モデルをトレーニングして、トレーニングされた深層学習モデルを取得するステップとをさらに含む。一実施形態によれば、参照高分子は、高分子とは異なる。一実施形態によれば、深層学習モデルは、畳み込みニューラルネットワーク（ＣＮＮ）を含む。

【００１５】

別の態様によれば、高分子の生物学的ポリマーアセンブリを生成するためのシステムが提供される。システムは、少なくとも１つのコンピュータハードウェアプロセッサと、命令を格納する少なくとも１つの非一時的なコンピュータ可読記憶媒体とを備え、命令は、少なくとも１つのコンピュータハードウェアプロセッサによる実行時に、少なくとも１つのコンピュータハードウェアプロセッサに、複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、複数の生物学的ポリマー配列およびアセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、第１の入力をトレーニングされた

深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、トレーニングされた深層学習モデルの第１の出力を使用して、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得するステップとを実行させる。

【００１６】

一実施形態によれば、高分子はタンパク質を含み、複数の生物学的ポリマー配列は複数のアミノ酸配列を含み、アセンブリは個々のアセンブリ位置におけるアミノ酸を示す。

一実施形態によれば、高分子は核酸を含み、複数の生物学的ポリマー配列は複数のヌクレオチド配列を含み、アセンブリは個々のアセンブリ位置におけるヌクレオチドを示す。

【００１７】

一実施形態によれば、アセンブリは、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、アセンブリを更新するステップは、第１のアセンブリ位置における第２のヌクレオチドを示すようにアセンブリを更新することを含む。

【００１８】

一実施形態によれば、命令はさらに、少なくとも１つのコンピュータハードウェアプロセッサに、アセンブリを更新して更新されたアセンブリを取得した後、複数のヌクレオチド配列を更新されたアセンブリに整列させるステップと、複数のヌクレオチド配列および更新されたアセンブリを使用して、トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、第２の入力をトレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、トレーニングされた深層学習モデルの第２の出力に基づいて、第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとを実行させる。

【００１９】

一実施形態によれば、命令はさらに、少なくとも１つのコンピュータハードウェアプロセッサに、複数のヌクレオチド配列をアセンブリに整列させるステップを実行させる。一実施形態によれば、複数のヌクレオチド配列は、少なくとも５個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも９個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも１０個のヌクレオチド配列を含む。

【００２０】

一実施形態によれば、トレーニングされた深層学習モデルへの第１の入力を生成するステップは、第１の複数のアセンブリ位置を選択すること、選択された第１の複数のアセンブリ位置に基づいて第１の入力を生成することを含む。一実施形態によれば、アセンブリ内の第１の複数の位置を選択することは、アセンブリが第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、および決定された尤度を使用して、第１の複数のアセンブリ位置を選択することを含む。

【００２１】

一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数のヌクレオチド配列の個々の１つをアセンブリと比較することを含む。一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成して、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定することは、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置における複数のヌクレオチドの各々に関して、ヌクレオチドがその位置にあることを示

す複数のヌクレオチド配列の数を示すカウントを決定すること、アセンブリがその位置においてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、カウントと参照値との間の差異を示すエラー値を決定すること、第１の入力に参照値およびエラー値を含ませることを含む。一実施形態によれば、アセンブリがその位置においてヌクレオチドを示すかどうかに基づいて参照値を決定することは、アセンブリがその位置においてヌクレオチドを示している場合、参照値が第１の値であると決定すること、アセンブリがその位置においてヌクレオチドを示していない場合、参照値が第２の値であると決定することを含む。一実施形態によれば、第１の値は、複数のヌクレオチド配列の数であり、第２の値は、０である。一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、第１の列は、第１のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持し、第２の列は、第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する。一実施形態によれば、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置は、第１のアセンブリ位置とは別の少なくとも２つのアセンブリ位置を含む。

【００２２】

一実施形態によれば、１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度は、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置に存在する尤度を含み、第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを同定することを含む。

【００２３】

一実施形態によれば、命令はさらに、少なくとも１つのコンピュータハードウェアプロセッサに、複数のヌクレオチド配列からアセンブリを生成するステップを実行させる。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを複数のヌクレオチド配列に適用することを含む。

【００２４】

一実施形態によれば、命令はさらに、少なくとも１つのコンピュータハードウェアプロセッサに、参照高分子および参照高分子の所定のアセンブリのシークエンシングから取得された生物学的ポリマー配列を含むトレーニングデータにアクセスするステップと、トレーニングデータを使用して深層学習モデルをトレーニングし、トレーニングされた深層学習モデルを取得するステップとを実行させる。一実施形態によれば、参照高分子は高分子とは異なる。一実施形態によれば、深層学習モデルは、畳み込みニューラルネットワーク（ＣＮＮ）を含む。

【００２５】

別の態様によれば、非一時的なコンピュータ可読記憶媒体が提供される。非一時的なコンピュータ可読記憶媒体は、少なくとも１つのコンピュータハードウェアプロセッサによる実行時に、少なくとも１つのコンピュータハードウェアプロセッサに高分子の生物学的ポリマーアセンブリを生成する方法を実行させる命令を格納する。方法は、複数の生物学的ポリマー配列と、個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリとにアクセスするステップと、複数の生物学的ポリマー配列およびアセンブリを使用して、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップと、第１の入力をトレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置

の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がその位置に存在する１つまたは複数の尤度を示す対応する第１の出力を取得するステップと、トレーニングされた深層学習モデルの第１の出力を使用して、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップと、第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得するステップとを含む。

【００２６】

一実施形態によれば、高分子はタンパク質を含み、複数の生物学的ポリマー配列は複数のアミノ酸配列を含み、アセンブリは個々のアセンブリ位置におけるアミノ酸を示す。

一実施形態によれば、高分子は核酸を含み、複数の生物学的ポリマー配列は複数のヌクレオチド配列を含み、アセンブリは個々のアセンブリ位置におけるヌクレオチドを示す。

【００２７】

一実施形態によれば、アセンブリは、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチドを示し、第１の複数のアセンブリ位置における生物学的ポリマーを同定するステップは、第１のアセンブリ位置において第２のヌクレオチドを同定することを含み、アセンブリを更新するステップは、第１のアセンブリ位置における第２のヌクレオチドを示すようにアセンブリを更新することを含む。

【００２８】

一実施形態によれば、方法は、アセンブリを更新して、更新されたアセンブリを取得した後、複数のヌクレオチド配列を更新されたアセンブリに整列させるステップと、複数のヌクレオチド配列および更新されたアセンブリを使用して、トレーニングされた深層学習モデルに提供される第２の入力を生成するステップと、第２の入力をトレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がその位置に存在する１つまたは複数の尤度を示す対応する第２の出力を取得するステップと、トレーニングされた深層学習モデルの第２の出力に基づいて、第２の複数のアセンブリ位置におけるヌクレオチドを同定するステップと、第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するステップとを含む。

【００２９】

一実施形態によれば、方法は、複数のヌクレオチド配列をアセンブリに整列させるステップをさらに含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも５個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも９個のヌクレオチド配列を含む。一実施形態によれば、複数のヌクレオチド配列は、少なくとも１０個のヌクレオチド配列を含む。

【００３０】

一実施形態によれば、トレーニングされた深層学習モデルへの第１の入力を生成するステップは、第１の複数のアセンブリ位置を選択すること、選択された第１の複数のアセンブリ位置に基づいて第１の入力を生成することを含む。一実施形態によれば、アセンブリ内の第１の複数の位置を選択することは、アセンブリが第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定すること、および決定された尤度を使用して、第１の複数のアセンブリ位置を選択することを含む。

【００３１】

一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数のヌクレオチド配列の個々の１つをアセンブリと比較することを含む。一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成して、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドを同定することは、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置における複数のヌクレオチドの各々に関して、ヌクレオチドがその位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定すること、アセンブリがその位置においてヌクレオチドを示しているかどうかに基づいて参照値を決定すること、カウントと

参照値との間の差異を示すエラー値を決定すること、第１の入力に参照値およびエラー値を含ませることを含む。一実施形態によれば、アセンブリがその位置においてヌクレオチドを示すかどうかに基づいて参照値を決定することは、アセンブリがその位置においてヌクレオチドを示している場合、参照値が第１の値であると決定すること、アセンブリがその位置においてヌクレオチドを示していない場合、参照値が第２の値であると決定することを含む。一実施形態によれば、第１の値は、複数のヌクレオチド配列の数であり、第２の値は、０である。一実施形態によれば、トレーニングされた深層学習モデルに提供される第１の入力を生成するステップは、複数の列を有するデータ構造に値を配置することを含み、第１の列は、第１のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持し、第２の列は、第１のアセンブリ位置の近傍にある１つまたは複数のアセンブリ位置のうちの第２のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する。一実施形態によれば、第１のアセンブリ位置の近傍の１つまたは複数のアセンブリ位置は、第１のアセンブリ位置とは別の少なくとも２つのアセンブリ位置を含む。

【００３２】

一実施形態によれば、１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度は、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置に存在する尤度を含み、第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のヌクレオチドが第１の位置に存在する尤度が複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを同定することを含む。

【００３３】

一実施形態によれば、方法は、複数のヌクレオチド配列からアセンブリを生成するステップをさらに含む。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することを含む。一実施形態によれば、複数のヌクレオチド配列からアセンブリを生成するステップは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを複数のヌクレオチド配列に適用することを含む。

【００３４】

一実施形態によれば、方法は、参照高分子のシークエンシングから取得された生物学的ポリマー配列と、参照高分子の所定のアセンブリとを含むトレーニングデータにアクセスするステップと、トレーニングデータを使用して深層学習モデルをトレーニングして、トレーニングされた深層学習モデルを取得するステップとをさらに含む。一実施形態によれば、参照高分子は、高分子とは異なる。一実施形態によれば、深層学習モデルは、畳み込みニューラルネットワーク（ＣＮＮ）を含む。

【図面の簡単な説明】

【００３５】

以下の図面を参照して、本出願の様々な態様および実施形態に関して説明する。図面は必ずしも一定の縮尺で描かれているわけではないことを理解されたい。複数の図面に表示されている構成要素は、表示されている全ての図面で同じ参照番号で示されている。

【図１Ａ】本明細書に記載の技術のいくつかの実施形態による、本明細書に記載の技術の態様を実施し得るシステムを示す図である。

【図１Ｂ】本明細書に記載の技術のいくつかの実施形態による、本明細書に記載の技術の態様を実施し得るシステムを示す図である。

【図１Ｃ】本明細書に記載の技術のいくつかの実施形態による、本明細書に記載の技術の態様を実施し得るシステムを示す図である。

【図２Ａ】本明細書に記載の技術のいくつかの実施形態による、アセンブリシステムの実施形態を示す図である。

【図２Ｂ】本明細書に記載の技術のいくつかの実施形態による、アセンブリシステムの実施形態を示す図である。
【図２Ｃ】本明細書に記載の技術のいくつかの実施形態による、アセンブリシステムの実施形態を示す図である。
【図２Ｄ】本明細書に記載の技術のいくつかの実施形態による、アセンブリシステムの実施形態を示す図である。
【図３Ａ】本明細書に記載の技術のいくつかの実施形態による、生物学的ポリマーアセンブリを生成するための機械学習モデルをトレーニングするための例示的なプロセス３００を示す図である。
【図３Ｂ】本明細書に記載の技術のいくつかの実施形態による、図３Ａのプロセスによって取得された機械学習モデルを使用して生物学的ポリマーアセンブリを生成するための例示的なプロセス３１０を示す図である。
【図４Ａ】本明細書に記載の技術のいくつかの実施形態による、機械学習モデルへの入力を生成する例を示す図である。
【図４Ｂ】本明細書に記載の技術のいくつかの実施形態による、機械学習モデルへの入力を生成する例を示す図である。
【図４Ｃ】本明細書に記載の技術のいくつかの実施形態による、機械学習モデルへの入力を生成する例を示す図である。
【図５】本明細書に記載の技術のいくつかの実施形態による、生物学的ポリマーアセンブリを更新する例を示す図である。
【図６】本明細書に記載の技術のいくつかの実施形態による、生物学的ポリマーアセンブリを生成するために使用される例示的な畳み込みニューラルネットワーク（ＣＮＮ）モデルの構造を示す図である。
【図７】従来の技術と比較した、本明細書に記載の技術のいくつかの実施形態により具体化されたアセンブリ技術の性能を示す図である。
【図８】本明細書に記載の技術のいくつかの実施形態を実施する際に使用し得る例示的なコンピューティングデバイス８００のブロック図である。
【発明を実施するための形態】
【００３６】
高分子は、タンパク質またはタンパク質フラグメント、（任意のタイプのＤＮＡの）ＤＮＡ分子またはフラグメント、または（任意のタイプのＲＮＡの）ＲＮＡ分子またはフラグメントであり得る。生物学的ポリマーは、アミノ酸（例えば、高分子がタンパク質またはそのフラグメントである場合）、またはヌクレオチド（例えば、高分子がＤＮＡ、ＲＮＡ、またはそのフラグメントである場合）であり得る。
【００３７】
本発明者らは、機械学習技術を使用して高分子の生物学的ポリマーアセンブリを生成するシステムを開発した。例えば、本発明者らによって開発されたシステムは、機械学習技術を使用して、生物のＤＮＡのゲノムアセンブリを生成するように構成され得る。別の例として、本発明者らによって開発されたシステムは、機械学習技術を使用してタンパク質のアミノ酸配列を生成するように構成され得る。
【００３８】
いくつかの実施形態では、システムは、１つまたは複数の生物学的ポリマー配列（例えば、シークエンシングデバイスによって生成される）および配列から生成された初期アセンブリにアクセスし得る。アセンブリは、個々のアセンブリの位置において生物学的ポリマー（例えば、ヌクレオチド、アミノ酸）が存在することを示し得る。システムは、（１）配列と初期アセンブリとを使用して、機械学習モデルに提供される入力を生成し、（２）入力をトレーニング済みの機械学習モデルに提供して、対応する出力を取得し、（３）機械学習モデルから取得した出力を使用して初期アセンブリを更新し、更新されたアセンブリを取得することによって、初期アセンブリの生物学的ポリマーの表示のエラーを修正し得る。更新されたアセンブリは、初期アセンブリよりも生物学的ポリマーの表示におけ

るエラーが少なくなり得る。

【００３９】

いくつかの実施形態では、アセンブリは、複数の位置と、個々の位置における生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）の表示とを含み得る。例として、アセンブリは、生物のゲノム内の位置におけるヌクレオチドを示すゲノムアセンブリであり得る。別の例として、アセンブリは、生物のＤＮＡの一部のヌクレオチドの配列を示す遺伝子配列であり得る。別の例として、アセンブリは、タンパク質のアミノ酸配列（「タンパク質配列」とも呼ばれる）であり得る。生物学的ポリマーは、ヌクレオチド、アミノ酸、または他の任意のタイプの生物学的ポリマーであり得る。生物学的ポリマー配列は、本明細書では「配列」または「リード（ｒｅａｄ）」と呼ばれ得る。

【００４０】

いくつかの従来の生物学的ポリマーアセンブリ技術は、シークエンシング技術を利用して高分子（例えば、ＤＮＡ、ＲＮＡ、またはタンパク質）の生物学的ポリマー配列を生成し、生成された配列を使用して高分子のアセンブリを生成し得る。例えば、シークエンシングデバイスは、生物のＤＮＡサンプルからヌクレオチド配列を生成し得、その配列を使用して、生物のＤＮＡのゲノムアセンブリを生成し得る。別の例として、シークエンシングデバイスは、タンパク質サンプルのアミノ酸配列を生成し得、その配列を使用して、タンパク質のより長いアミノ酸配列を組み立て得る。コンピューティングデバイスは、シークエンシングデバイスによって生成された配列にアセンブリアルゴリズムを適用してアセンブリを生成し得る。例えば、コンピューティングデバイスは、ＤＮＡサンプルのヌクレオチド配列にオーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アセンブリアルゴリズムを適用して、生物のゲノムアセンブリまたはその一部を生成し得る。

【００４１】

核酸サンプルからヌクレオチド配列を生成するために使用されるシークエンシング技術の１つのタイプは、１０００個未満のヌクレオチドのヌクレオチド配列（即ち、「ショートリード」）を生成する第２世代シークエンシング（「ショートリードシークエンシング」としても知られる）である。シークエンシング技術は、１０００個以上のヌクレオチドのヌクレオチド配列（即ち、「ロングリード」）を生成し、かつ第二世代シークエンシングよりもアセンブリの大きな部分を提供する第三世代シークエンシング（「ロングリードシークエンシング」とも呼ばれる）に進化した。しかしながら、本発明者らは、第三世代シークエンシングは第二世代シークエンシングよりも精度が低く、その結果、ロングリードから生成されたアセンブリはショートリードから生成されたアセンブリよりも精度が低いことを認識した。本発明者らはまた、アセンブリの精度を向上するための従来のエラー訂正技術は、計算コストおよび時間がかかることを認識した。従って、本発明者らは、（１）第三世代シークエンシングから生成されたアセンブリの精度を向上させ、（２）従来のエラー訂正技術よりも効率的であるアセンブリのエラーを修正するための機械学習技術を開発した。

【００４２】

本明細書に記載のいくつかの実施形態は、発明者がアセンブリの生成に関して認識した上記の問題の全てに対処する。しかしながら、本明細書に記載される全ての実施形態がこれらの問題の全てに対処するわけではないことを理解されたい。本明細書に記載の技術の実施形態は、生物学的ポリマーアセンブリの上記の問題に対処する以外の目的に使用し得ることも理解されたい。一例として、本明細書に記載の技術の実施形態を使用して、アミノ酸配列から生成されたタンパク質配列の精度を向上し得る。別の例として、本明細書に記載の技術の実施形態を使用して、ショートリードから生成されたアセンブリの精度を向上し得る。

【００４３】

いくつかの実施形態では、システムは、（１）個々のアセンブリ位置に存在する生物学的ポリマーを示すアセンブリ（例えば、複数の生物学的ポリマー配列から生成される）にアクセスし、（２）複数の生物学的ポリマー配列およびアセンブリを使用して、トレーニ

ングされた深層学習モデルに提供される第１の入力を生成し、（３）第１の入力をトレーニングされた深層学習モデルに提供して、第１の複数のアセンブリ位置の各々に関して、１つまたは複数の個々の生物学的ポリマーの各々がそのアセンブリ位置に存在する１つまたは複数の尤度（例えば、確率）を示す対応する第１の出力を取得し、（４）トレーニングされた深層学習モデルの第１の出力を使用して、第１の複数のアセンブリ位置における生物学的ポリマーを同定し、（５）第１の複数のアセンブリ位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得するように構成されている。いくつかの実施形態では、システムは、複数の生物学的ポリマー配列をアセンブリに整列させるように構成され得る。

【００４４】

いくつかの実施形態では、高分子はタンパク質であり得、複数の生物学的ポリマー配列は複数のアミノ酸配列であり得、アセンブリは個々のアセンブリ位置におけるアミノ酸を示す。いくつかの実施形態において、高分子は、核酸（例えば、ＤＮＡ、ＲＮＡ）であり得、複数の生物学的配列は、複数のヌクレオチド配列であり得、アセンブリは、個々のアセンブリ位置におけるヌクレオチドを示す。

【００４５】

いくつかの実施形態では、アセンブリは、複数のアセンブリ位置のうちの第１のアセンブリ位置における第１のヌクレオチド（例えば、アデニン）を示す。第１の複数のアセンブリ位置における生物学的ポリマーを同定することは、第１のアセンブリ位置において第１のヌクレオチドとは異なる第２のヌクレオチド（例えば、チミン）を同定することを含み、アセンブリを更新することは、第１のアセンブリ位置における第２のヌクレオチド（例えば、チミン）を示すようにアセンブリを更新することを含む。

【００４６】

いくつかの実施形態では、システムは、複数の更新の反復を実行するように構成され得る。システムは、アセンブリを更新して、更新されたアセンブリを取得した後、（１）複数のヌクレオチド配列を更新されたアセンブリに整列させ、（２）複数のヌクレオチド配列および更新されたアセンブリを使用して、トレーニングされた深層学習モデルに提供される第２の入力を生成し、（３）第２の入力をトレーニングされた深層学習モデルに提供して、第２の複数のアセンブリ位置の各々に関して、１つまたは複数の個々のヌクレオチドの各々がそのアセンブリ位置に存在する１つまたは複数の尤度（例えば、確率）を示す対応する第２の出力を取得し、（４）トレーニングされた深層学習モデルの第２の出力に基づいて、第２の複数のアセンブリ位置におけるヌクレオチドを同定し、（５）第２の複数のアセンブリ位置において同定されたヌクレオチドを示すように更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するように構成され得る。

【００４７】

いくつかの実施形態では、システムは、（１）第１の複数のアセンブリ位置を選択し、（２）選択された第１の複数のアセンブリ位置に基づいて第１の入力を生成することによって、トレーニングされた深層学習モデルへの第１の入力を生成するように構成され得る。いくつかの実施形態では、システムは、（１）アセンブリが第１の複数のアセンブリ位置においてヌクレオチドを不正確に示す尤度を決定し、（２）決定された尤度を使用して、第１の複数のアセンブリ位置を選択することによって、第１の複数のアセンブリ位置を選択するように構成され得る。

【００４８】

いくつかの実施形態では、システムは、（例えば、１つまたは複数の特徴の値を決定するために）複数のヌクレオチド配列の個々の１つをアセンブリと比較することによって、トレーニングされた深層学習モデルに提供される第１の入力を生成するように構成され得る。いくつかの実施形態では、システムは、第１の入力の近傍にある１つまたは複数のアセンブリ位置の各々における複数のヌクレオチドの各々に関して、（１）ヌクレオチドがそのアセンブリ位置にあることを示す複数のヌクレオチド配列の数を示すカウントを決定し、（２）アセンブリがそのアセンブリ位置においてヌクレオチドを示しているかどうか

に基づいて参照値を決定し、（３）カウントと基準値との間の差異を示すエラー値を決定し、（４）第１の入力に基準値およびエラー値を含ませることによって、第１の複数のアセンブリ位置の第１のアセンブリ位置におけるヌクレオチドを同定するための第１の入力を生成するように構成され得る。いくつかの実施形態では、システムは、アセンブリがそのアセンブリ位置においてヌクレオチドを示すかどうかに基づいて、（１）アセンブリがそのアセンブリ位置においてヌクレオチドを示している場合、参照値が第１の値（例えば、複数のヌクレオチド配列の数）であると決定し、（２）アセンブリがそのアセンブリ位置においてヌクレオチドを示していない場合、参照値が第２の値（例えば、０）であると決定することにより、参照値を決定するように構成され得る。いくつかの実施形態では、システムは、３個、４個、５個、６個、７個、８個、９個、１０個、１５個、２０個、２５個、３０個、３５個、４０個、４５個、または５０個の位置の近傍を使用するように構成され得る。

【００４９】

いくつかの実施形態では、システムは、行／列を有するデータ構造に値を配置することによって、第１のアセンブリ位置におけるヌクレオチドを同定するための第１の入力を生成するように構成され得、（１）第１の行／列は、第１のアセンブリ位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持し、（２）第２の行／列は、第１のアセンブリ位置の近傍にある第２の位置において複数のヌクレオチドに関して決定された参照値およびエラー値を保持する。

【００５０】

いくつかの実施形態では、１つまたは複数の個々の生物学的ポリマーの各々がアセンブリ位置に存在する１つまたは複数の尤度は、複数のヌクレオチドの各々に関して、ヌクレオチドがアセンブリ位置において存在する尤度（例えば、確率）を含む。システムは、第１の複数のアセンブリ位置のうちの第１のアセンブリ位置におけるヌクレオチドが複数のヌクレオチドのうちの第１のヌクレオチドであることを同定することによって、アセンブリ内の第１の複数のアセンブリ位置における生物学的ポリマーを同定するように構成され得る。システムは、第１のヌクレオチドが第１のアセンブリ位置に存在する尤度が、複数のヌクレオチドのうちの第２のヌクレオチドが第１のアセンブリ位置に存在する尤度よりも大きいことを決定することによって、第１のアセンブリ位置におけるヌクレオチドが第１のヌクレオチドであることを同定し得る。

【００５１】

いくつかの実施形態では、システムは、複数のヌクレオチド配列からアセンブリ（例えば、初期アセンブリ）を生成するように構成され得る。いくつかの実施形態では、システムは、アセンブリとなる複数のヌクレオチド配列からコンセンサス配列を決定することによって（例えば、多数決を取ることによって）アセンブリを生成するように構成され得る。いくつかの実施形態では、システムは、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ）アルゴリズムを複数のヌクレオチド配列に適用することによって、複数のヌクレオチド配列からアセンブリを生成するように構成され得る。いくつかの実施形態では、システムは、（１）参照高分子のシークエンシングから取得された生物学的ポリマー配列と、参照高分子の所定の生物学的ポリマーアセンブリとを含むトレーニングデータにアクセスし、（２）トレーニングデータを使用して深層学習モデル（畳み込みニューラルネットワークまたは再帰型ニューラルネットワークなど）をトレーニングして、トレーニングされた深層学習モデルを取得するように構成されている。いくつかの実施形態では、深層学習モデルをトレーニングするために使用される参照高分子は、アセンブリが生成されている高分子とは異なり得る。

【００５２】

上記で導入され、以下でより詳細に説明される技術は、技術が特定の実施形態に限定されないことから、多数の方法のいずれかで実施され得ることを理解されたい。実施形態の詳細の例は、説明のみを目的として本明細書に提供されている。さらに、本明細書に記載の技術の態様は、特定の技術または技術の組み合わせの使用に限定されないことから、本

明細書に開示される技術は、個別にまたは任意の適切な組み合わせで使用され得る。

【００５３】

図１Ａは、本明細書に記載の技術の態様を具体化し得るシステム１００を示す。システム１００は、１つまたは複数のシークエンシングデバイス１０２、アセンブリシステム１０４、モデルトレーニングシステム１０６、およびデータストア１０８Ａを含み、これらの各々は、ネットワーク１１１に接続されている。

【００５４】

いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、高分子の１つまたは複数のサンプル標本１１０のシークエンシングによってシークエンシングデータを生成するように構成され得る。例えば、サンプル標本１１０は、核酸（例えば、ＤＮＡおよび／またはＲＮＡ）、またはタンパク質（例えば、ペプチド）を含む生物学的サンプルであり得る。シークエンシングデータは、サンプル標本（単数または複数）１１０の生物学的ポリマー配列を含み得る。生物学的ポリマー配列は、高分子サンプル中に存在する生物学的ポリマーの順序および位置を示す英数字記号の配列として表され得る。いくつかの実施形態では、生物学的ポリマー配列は、生物学的サンプルのシークエンシングから生成されたヌクレオチド配列であり得る。例として、ヌクレオチド配列は、（１）アデニンを表す「Ａ」、（２）シトシンを表す「Ｃ」、（３）グアニンを表す「Ｇ」、（４）チミンを表す「Ｔ」、（５）ウラシルを表す「Ｕ」、（６）配列内の位置にヌクレオチドが存在しないことを表す「－」を使用し得る。いくつかの実施形態では、生物学的ポリマー配列は、タンパク質サンプル（例えば、ペプチド）のシークエンシングから生成されたアミノ酸配列であり得る。一例として、アミノ酸配列は、タンパク質に存在し得る個々の異なるアミノ酸を表すために異なる英数字を使用する英数字配列であり得る。

【００５５】

いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、核酸サンプル（例えば、ＤＮＡサンプル）のシークエンシングからヌクレオチド配列を生成するように構成され得る。いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、合成によって核酸サンプルをシークエンシングするように構成され得る。シークエンシングデバイス（単数または複数）１０２は、ヌクレオチドが、シークエンシングされている核酸に相補的である核酸の新たに合成された鎖に取り込まれるときに、ヌクレオチドを同定するように構成され得る。シークエンシング中に、重合酵素（例えば、ＤＮＡポリメラーゼ）は、ターゲット核酸分子のプライミング位置（「プライマー」と呼ばれる）に結合（例えば、付着）して、重合酵素の作用を介してヌクレオチドをプライマーに取り込み得る。シークエンシングデバイス（単数または複数）１０２は、取り込まれている各ヌクレオチドを検出するように構成され得る。いくつかの実施形態において、ヌクレオチドは、励起に応答して発光する個々の発光分子（例えば、フルオロフォア）と結合され得る。発光分子は、個々のヌクレオチドと結合している発光分子が取り込まれているときに励起され得る。シークエンシングデバイス（単数または複数）１０２は、発光を検出するための１つまたは複数のセンサを含み得る。各タイプのヌクレオチドは、個々のタイプの発光分子と結合され得る。シークエンシングデバイス（単数または複数）１０２は、検出された発光に基づいて発光分子のタイプを同定することによって、取り込まれているヌクレオチドを同定し得る。例えば、シークエンシングデバイス（単数または複数）１０２は、発光強度、寿命、波長、または他の特性を使用して、異なる発光分子を区別し得る。いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、ヌクレオチドの取り込み中に生成された電気信号を検出して、取り込まれているヌクレオチドを同定するように構成され得る。シークエンシングデバイス（単数または複数）１０２は、電気信号を検出し、それらを使用して取り込まれているヌクレオチドを同定するためのセンサ（単数または複数）を含み得る。

【００５６】

いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、本明細書に記載されているものとは異なる技術を使用して核酸をシークエンシングするよう

に構成され得る。いくつかの実施形態は、本明細書に記載の核酸シークエンシングの特定の技術に限定されない。

【００５７】

いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、タンパク質サンプル（例えば、ペプチド）のシークエンシングからアミノ酸配列を生成するように構成され得る。いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、個々のアミノ酸に選択的に結合する試薬を使用してタンパク質サンプルをシークエンシングするように構成され得る。試薬は、他のタイプのアミノ酸よりも１つまたは複数のタイプのアミノ酸に選択的に結合し得る。いくつかの実施形態において、試薬は、個々の発光分子と結合され得る。発光分子は、発光分子と結合されている試薬とアミノ酸との間の相互作用に応答して励起され得る。いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、発光分子の発光を検出することによってアミノ酸を同定するように構成され得る。シークエンシングデバイス１０２は、発光を検出するための１つまたは複数のセンサを含み得る。いくつかの実施形態において、各タイプのアミノ酸は、個々のタイプの発光分子と結合され得る。シークエンシングデバイス（単数または複数）１０２は、検出された発光に基づいて発光分子のタイプを同定することによってアミノ酸を同定し得る。一例として、シークエンシングデバイス（単数または複数）１０２は、発光強度、寿命、波長、または他の特性を使用して、異なる発光分子を区別し得る。いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、試薬とアミノ酸との間の結合相互作用の間に生成される電気信号を検出するように構成され得る。シークエンシングデバイス（単数または複数）１０２は、電気信号を検出するためのセンサ（単数または複数）を含み得、かつその信号を使用して、個々の結合相互作用に関与するアミノ酸を同定し得る。

【００５８】

いくつかの実施形態では、シークエンシングデバイス（単数または複数）１０２は、本明細書に記載されているものとは異なる技術を使用してタンパク質をシークエンシングするように構成され得る。いくつかの実施形態は、本明細書に記載のタンパク質のシークエンシングの特定の技術に限定されない。

【００５９】

図１Ａの実施形態に示されるように、シークエンシングデバイス（単数または複数）１０２は、デバイス（単数または複数）１０２によって生成されたシークエンシングデータを、格納のためにデータストア１０８Ａに送信するように構成され得る。シークエンシングデータは、高分子サンプルのシークエンシングから生成された配列を含み得る。シークエンシングデータは、１つまたは複数の他のシステムによって使用され得る。一例として、シークエンシングデータは、高分子のアセンブリを生成するためにアセンブリシステム１０４によって使用され得る。別の例として、シークエンシングデータは、アセンブリシステム１０４によって使用されるための機械学習モデルをトレーニングするためのトレーニングデータとして、モデルトレーニングシステム１０６によって使用され得る。シークエンシングデータの使用例が本明細書に記載される。

【００６０】

いくつかの実施形態では、アセンブリシステム１０４は、シークエンシングデバイス（単数または複数）１０２によって生成されたシークエンシングデータを使用してアセンブリ１１２を生成するように構成されたコンピューティングデバイスであり得る。アセンブリシステム１０４は、アセンブリシステム１０４がアセンブリを生成するために使用する機械学習モデル１０４Ａを含む。いくつかの実施形態では、機械学習モデル１０４Ａは、モデルトレーニングシステム１０６から得られるトレーニングされた機械学習モデルであり得る。アセンブリシステム１０４によって使用され得る機械学習モデルの例は、本明細書に記載されている。

【００６１】

いくつかの実施形態では、アセンブリシステム１０４は、初期アセンブリを更新するこ

とによってアセンブリ１１２を生成するように構成され得る。初期アセンブリは、従来のアセンブリアルゴリズムをシークエンシングデータに適用することで取得され得る。いくつかの実施形態では、アセンブリシステム１０４は、初期アセンブリを生成するように構成され得る。アセンブリシステム１０４は、シークエンシングデバイス（単数または複数）１０２から取得されたシークエンシングデータにアセンブリアルゴリズムを適用することによって初期アセンブリを生成するように構成され得る。一例として、アセンブリシステム１０４は、オーバーラップ・レイアウト・コンセンサス（ＯＬＣ：Ｏｖｅｒｌａｐ Ｌａｙｏｕｔ Ｃｏｎｓｅｎｓｕｓ）アセンブリまたはド・ブラウン・グラフ（ＤＢＧ：Ｄｅ Ｂｒｕｉｊｎ Ｇｒａｐｈ）アセンブリを、データストア１０８Ａからのシークエンシングデータ（例えば、ヌクレオチド配列）に適用して、初期アセンブリを生成し得る。いくつかの実施形態では、アセンブリシステム１０４は、アセンブリシステム１０４とは別のシステムによって生成された初期アセンブリを取得するように構成され得る。一例として、アセンブリシステム１０４は、シークエンシングデバイス（単数または複数）１０２によって生成されたシークエンシングデータにアセンブリアルゴリズムを適用したアセンブリシステム１０４とは別のコンピューティングデバイスによって生成された初期アセンブリを受信し得る。

【００６２】

いくつかの実施形態では、アセンブリシステム１０４は、トレーニングされた機械学習モデル１０４Ａを使用して、アセンブリ（例えば、アセンブリアルゴリズムの適用から取得された初期アセンブリ）を更新または改良するように構成され得る。アセンブリシステム１０４は、アセンブリ内の１つまたは複数のエラーを修正することによって、かつ／またはアセンブリ内の生物学的ポリマーの表示を確認することによって、アセンブリを更新するように構成され得る。いくつかの実施形態では、アセンブリシステム１０４は、（１）シークエンシングデータおよびアセンブリを使用して機械学習モデル１０４Ａへの入力を生成すること、（２）生成された入力を機械学習モデル１０４Ａに提供して、対応する出力を取得すること、（３）機械学習モデル１０４Ａから取得された出力を使用してアセンブリを更新することによってアセンブリを更新するように構成され得る。いくつかの実施形態では、機械学習モデル１０４Ａの出力は、アセンブリ内の複数の位置の各々に関して、１つまたは複数の個々の生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）の各々がアセンブリ内のその位置に存在する１つまたは複数の尤度を示し得る。一例として、出力は、位置の各々に関して、個々のヌクレオチドがその位置に存在する確率を示し得る。いくつかの実施形態では、アセンブリシステム１０４は、（１）機械学習モデル１０４Ａから取得された出力を使用して、アセンブリの位置における生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）を同定し、（２）位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、更新されたアセンブリを取得するように構成され得る。本明細書では、機械学習モデルを使用してアセンブリを更新するための例示的な技術に関して説明している。

【００６３】

いくつかの実施形態では、アセンブリシステム１０４は、更新される（例えば、修正または確認される）べきアセンブリ内の位置を識別するように構成され得る。アセンブリシステム１０４は、選択された位置を使用して機械学習モデル１０４Ａへの入力を生成するように構成され得る。いくつかの実施形態では、アセンブリシステム１０４は、（１）個々のアセンブリの位置における生物学的ポリマーの表示が不正確である尤度を決定すること、および（２）決定された尤度に基づいて修正されるべき位置を選択することによって更新されるべき位置を識別するように構成され得る。いくつかの実施形態では、アセンブリシステム１０４は、個々の位置に示される生物学的ポリマーが不正確である尤度を示す数値を決定し、尤度値に基づいて更新されるべき位置を選択するように構成され得る。一例として、アセンブリシステム１０４は、閾値よりも大きな不正確である尤度を有する位置を選択し得る。

【００６４】

いくつかの実施形態では、アセンブリシステム１０４は、アセンブリ内の位置に関する特徴値を決定することによって、機械学習モデル１０４Ａへの入力を生成するように構成され得る。アセンブリシステム１０４は、アセンブリおよびアセンブリが生成された配列を使用して特徴値を決定するように構成され得る。例示的な特徴を本明細書において記載する。いくつかの実施形態では、アセンブリシステム１０４は、複数の位置の各々に関して機械学習モデル１０４Ａへの入力を生成するように構成され得る。各位置に関して、アセンブリシステム１０４は、特徴値を決定し、機械学習モデル１０４Ａへの入力として特徴値を提供して、対応する出力を取得するように構成され得る。アセンブリシステム１０４は、位置に関して提供された入力に対応する出力を使用して、その位置に示された生物学的ポリマーを修正するか、またはその位置において示された生物学的ポリマーが正確であることを確認するように構成され得る。いくつかの実施形態では、複数の位置は、アセンブリ内の全ての位置であり得る。いくつかの実施形態では、複数の位置は、アセンブリ内の一部の位置であり得る。

【００６５】

一部の位置が更新される実施形態では、アセンブリシステム１０４は、一部の位置を選択するように構成され得る。アセンブリシステム１０４は、（１）アセンブリが複数の位置において生物学的ポリマーを不正確に示す尤度を決定すること、（２）尤度を使用して、複数の位置から一部の位置を選択することを含むいくつかの方法で一部の位置を選択するように構成され得る。例えば、アセンブリシステム１０４は、（１）閾値の尤度を超える尤度を有する位置を特定し、（２）特定された位置を一部の位置として選択し得る。

【００６６】

いくつかの実施形態では、アセンブリシステム１０４は、位置の近傍の１つまたは複数の位置において決定された特徴値を使用して修正されるべき位置に関する入力を生成するように構成され得る。選択された位置に関して、機械学習モデル１０４Ａは、アセンブリ内の周囲の位置からのコンテキスト情報を利用して、選択された位置に関する出力を生成し得る。いくつかの実施形態では、近傍の位置は、（１）選択された位置、および（２）選択された位置の周囲の一組の位置を含み得る。一例として、近傍は、機械学習モデル１０４Ａが出力を生成することになる選択された位置を中心とする複数の位置のウィンドウであり得る。アセンブリシステム１０４は、５個の位置、１０個の位置、１５個の位置、２０個の位置、２５個の位置、３０個の位置、３５個の位置、４０個の位置、４５個の位置、および／または５０個の位置のウィンドウを使用し得る。

【００６７】

いくつかの実施形態では、アセンブリシステム１０４は、最終的なアセンブリ１１２を生成するために複数の更新の反復を実行するように構成され得る。一例として、アセンブリシステム１０４は、（１）初期アセンブリで１回目の反復を実行して、第１の更新されたアセンブリを取得し、（２）第１の更新されたアセンブリに対して２回目の反復を実行して、第２の更新されたアセンブリを取得し得る。いくつかの実施形態では、アセンブリシステム１０４は、更新を反復して実行するように構成され得る。アセンブリシステム１０４は、条件が満たされるまで更新の反復を実行するように構成され得る。例示的な条件が本明細書において記載されている。

【００６８】

いくつかの実施形態では、モデルトレーニングシステム１０６は、データストア１０８Ａに格納されたデータにアクセスし、アクセスされたデータを使用して、アセンブリを生成する際に使用するための機械学習モデルをトレーニングするように構成されたコンピューティングデバイスであり得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、異なるアセンブリシステムに対する別個の機械学習モデルをトレーニングするように構成され得る。個々のアセンブリシステム用にトレーニングされた機械学習モデルは、アセンブリシステムの固有の特性に合わせて調整され得る。一例として、モデルトレーニングシステム１０６は、（１）第１のアセンブリシステム用の第１の機械学習モデルをトレーニングし、（２）第２のアセンブリシステム用の第２の機械学習モデルをトレー

ニングするように構成され得る。アセンブリシステムの各々に対する個別の機械学習モデルは、個々のアセンブリシステムの固有のエラープロファイルに合わせて調整され得る。例えば、異なるアセンブリシステムは、初期アセンブリを生成するために異なるアセンブリアルゴリズムを採用し得、各アセンブリシステム用にトレーニングされた機械学習モデルは、アセンブリアルゴリズムのエラープロファイルに合わせて調整され得る。

【００６９】

いくつかの実施形態では、モデルトレーニングシステム１０６は、単一のトレーニングされた機械学習モデルを複数のアセンブリシステムに提供するように構成され得る。一例として、モデルトレーニングシステム１０６は、複数のアセンブリシステムからのアセンブリを集約して、単一の機械学習モデルをトレーニングし得る。複数のアセンブリシステムで採用されているアセンブリ技術における変動に起因するモデルの変動を軽減するために、単一の機械学習モデルが複数のアセンブリシステムに対して正規化され得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、複数のシークエンシングデバイスに対して単一のトレーニングされた機械学習モデルを提供するように構成され得る。一例として、モデルトレーニングシステム１０６は、複数のシークエンシングデバイスからのシークエンシングデータを集約し、単一の機械学習モデルをトレーニングし得る。単一の機械学習モデルは、デバイスの変動に起因するモデルの変動を軽減するために、複数のシークエンシングデバイスに対して正規化され得る。

【００７０】

いくつかの実施形態では、モデルトレーニングシステム１０６は、（１）１つまたは複数の参照高分子（例えば、ＤＮＡ、ＲＮＡ、タンパク質）のシークエンシングから取得された生物学的ポリマー配列と、（２）参照高分子（単数または複数）の１つまたは複数の所定のアセンブリとを含むトレーニングデータを使用することによって機械学習モデルをトレーニングするように構成され得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、所定のアセンブリ内の生物学的ポリマーの表示を、機械学習モデルをトレーニングするためのラベルとして使用するように構成され得る。ラベルは、アセンブリの位置における正確な表示または所望の表示を表し得る。一例として、トレーニングデータは、生物のＤＮＡサンプルのシークエンシングから所得されるヌクレオチド配列、および生物の所定のゲノムアセンブリを含み得る。この例では、モデルトレーニングシステム１０６は、所定のゲノムアセンブリ内のヌクレオチドの表示を、トレーニングデータに教師あり学習アルゴリズムを適用するためのラベルとして使用し得る。

【００７１】

いくつかの実施形態では、モデルトレーニングシステム１０６は、外部データベースのトレーニングデータにアクセスするように構成され得る。一例として、モデルトレーニングシステム１０６は、（１）パシフィック・バイオサイエンシズ社（Ｐａｃｉｆｉｃ Ｂｉｏｓｃｉｅｎｃｅｓ）のＲＳ ＩＩ（パックバイオ（Ｐａｃｂｉｏ（登録商標））データベースおよび／またはオックスフォード・ナノポア社（Ｏｘｆｏｒｄ Ｎａｎｏｐｏｒｅ）のＭｉｎｉＩＯＮ（ＯＮＴ）データベースのシークエンシングデータ、（２）米国国立バイオ技術情報センター（ＮＣＢＩ）の参照ゲノムデータベースの所定のゲノムアセンブリにアクセスし得る。別の例として、モデルトレーニングシステム１０６は、ユニットプロット（ＵｎｉｔＰｒｏｔ）データベースおよび／またはヒト・プロテオーム・プロジェクト（ＨＰＰ：Ｈｕｍａｎ Ｐｒｏｔｅｏｍｅ Ｐｒｏｊｅｃｔ ）データベースからタンパク質シークエンシングデータおよび関連するプロテオームアセンブリにアクセスし得る。

【００７２】

いくつかの実施形態では、モデルトレーニングシステム１０６は、ラベル付けされたトレーニングデータを使用して教師あり学習トレーニングアルゴリズムを適用することによって機械学習モデルをトレーニングするように構成され得る。一例として、モデルトレーニングシステム５０４は、確率的勾配降下法を使用することによって、深層学習モデル（例えば、ニューラルネットワーク）をトレーニングし得る。別の例として、モデルトレー

ニングシステム１０６は、コスト関数を最適化することによってサポートベクターマシン（ＳＶＭ）の決定境界を特定するためにＳＶＭをトレーニングし得る。一例として、モデルトレーニングシステム１０６は、（１）シークエンシングデータと、シークエンシングデータへのアセンブリアルゴリズムの適用により生成されたアセンブリとを使用して、機械学習モデルへの入力を生成し、（２）高分子の所定のアセンブリ（例えば、公開データベースからの）を使用して入力にラベルを付け、（３）生成された入力および対応するラベルに教師ありトレーニングアルゴリズムを適用し得る。

【００７３】

いくつかの実施形態では、モデルトレーニングシステム１０６は、教師なし学習アルゴリズムをトレーニングデータに適用することによって機械学習モデルをトレーニングするように構成され得る。一例として、モデルトレーニングシステム１０６は、ｋ平均クラスタリングを実行することによって、クラスタリングモデルのクラスタを特定し得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、（１）シークエンシングデータと、シークエンシングデータへのアセンブリアルゴリズムの適用により生成されたアセンブリとを使用して、機械学習モデルへの入力を生成し、（２）生成された入力に教師なし学習アルゴリズムを適用し得る。一例として、モデルトレーニングシステム１０６は、モデルの各クラスタが個々のヌクレオチドを表すクラスタリングモデルをトレーニングし得、クラスタ分類は、ゲノムアセンブリまたは遺伝子配列内のある位置におけるヌクレオチドを示し得る。別の例として、モデルトレーニングシステム１０６は、モデルの各クラスタが個々のアミノ酸を表すクラスタリングモデルをトレーニングし得、クラスタ分類は、タンパク質配列内のある位置におけるアミノ酸を示し得る。

【００７４】

いくつかの実施形態では、モデルトレーニングシステム１０６は、半教師あり学習アルゴリズムをトレーニングデータに適用することによって機械学習モデルをトレーニングするように構成され得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、（１）教師なし学習アルゴリズム（例えば、クラスタリング）をトレーニングデータに適用することによって一組のラベル付けされていないトレーニングデータにラベルを付けること、および（２）ラベル付けされたトレーニングデータに教師あり学習アルゴリズムを適用することによって、半教師あり学習アルゴリズムをトレーニングデータに適用するように構成され得る。一例として、モデルトレーニングシステム１０６は、（１）シークエンシングデータと、シークエンシングデータへのアセンブリアルゴリズムの適用により生成されたアセンブリとを使用して、機械学習モデルへの入力を生成し、（２）生成された入力に教師なし学習アルゴリズムを適用して入力にラベルを付け、（３）ラベル付けされたトレーニングデータに教師あり学習アルゴリズムを適用し得る。

【００７５】

いくつかの実施形態では、機械学習モデルは、深層学習モデル（例えば、ニューラルネットワーク）を含み得る。いくつかの実施形態では、深層学習モデルは、畳み込みニューラルネットワーク（ＣＮＮ：ｃｏｎｖｏｌｕｔｉｏｎａｌ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ）を含み得る。いくつかの実施形態では、深層学習モデルは、再帰型ニューラルネットワーク（ＲＮＮ：ｒｅｃｕｒｒｅｎｔ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ）、多層パーセプトロン、オートエンコーダ、および／またはＣＴＣ適合ニューラルネットワークモデルを含み得る。いくつかの実施形態では、機械学習モデルは、クラスタリングモデルを含み得る。一例として、クラスタリングモデルは、複数のクラスタを含み得、クラスタの各々は、生物学的ポリマー（例えば、ヌクレオチド、またはアミノ酸）に関連付けられている。

【００７６】

いくつかの実施形態では、モデルトレーニングシステム１０６は、複数のシークエンシングデバイスの各々に対する別個の機械学習モデルをトレーニングするように構成され得る。個々のシークエンシングデバイス用にトレーニングされた機械学習モデルは、シークエンシングデバイスの固有の特性に合わせて調整され得る。一例として、モデルトレーニングシステム１０６は、（１）第１のシークエンシングデバイス用の第１の機械学習モデ

ルをトレーニングし、（２）第２のシークエンシングデバイス用の第２の機械学習モデルをトレーニングし得る。個々のシークエンシングデバイス用にトレーニングされた機械学習モデルは、シークエンシングデバイスによって生成されたシークエンシングデータとともに使用するために最適化され得る。例えば、機械学習モデルは、シークエンシングデバイスによって使用される特定のシークエンシング技術（例えば、第三世代シークエンシング）のために最適化され得る。

【００７７】

いくつかの実施形態では、モデルトレーニングシステム１０６は、以前にトレーニングされた機械学習モデルを定期的に更新するように構成され得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、新たなトレーニングデータを使用して機械学習モデルの１つまたは複数のパラメータの値を更新することによって、以前にトレーニングされたモデルを更新するように構成され得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、以前に取得されたトレーニングデータと新たなトレーニングデータとの組み合わせを使用して新たな機械学習モデルをトレーニングすることによって、機械学習モデルを更新するように構成され得る。

【００７８】

いくつかの実施形態では、モデルトレーニングシステム１０６は、異なるタイプのイベントのいずれか１つに応答して機械学習モデルを更新するように構成され得る。例えば、いくつかの実施形態では、モデルトレーニングシステム１０６は、ユーザコマンドに応答して機械学習モデルを更新するように構成され得る。一例として、モデルトレーニングシステム１０６は、ユーザがトレーニングプロセスの実行を命令し得るユーザインターフェースを提供し得る。いくつかの実施形態では、モデルトレーニングシステム１０６は、例えば、ソフトウェアコマンドに応答して、機械学習モデルを自動的に（即ち、ユーザコマンドに応答することなく）更新するように構成され得る。別の例として、いくつかの実施形態では、モデルトレーニングシステム１０６は、１つまたは複数の条件の検出に応答して機械学習モデルを更新するように構成され得る。例えば、モデルトレーニングシステム１０６は、期間の満了を検出することに応答して、機械学習モデルを更新し得る。別の例として、モデルトレーニングシステム１０６は、閾値量（例えば、配列の数および／またはアセンブリの数）の新たなトレーニングデータを受信することに応答して、機械学習モデルを更新し得る。

【００７９】

図１Ａに示される例示的な実施形態では、モデルトレーニングシステム１０６は、アセンブリシステム１０４から分離されているが、いくつかの実施形態では、モデルトレーニングシステム１０６は、アセンブリシステム１０４の一部であり得る。図１Ａに示される例示的な実施形態では、アセンブリシステム１０４は、シークエンシングデバイス（単数または複数）１０２から分離されているが、いくつかの実施形態では、アセンブリシステム１０４は、シークエンシングデバイスの構成要素であり得る。いくつかの実施形態では、シークエンシングデバイス１０２、モデルトレーニングシステム１０６、およびアセンブリシステム１０４は、各々、単一のシステムの構成要素であり得る。

【００８０】

いくつかの実施形態では、データストア１０８Ａは、データを格納するためのシステムであり得る。いくつかの実施形態では、データストア１０８Ａは、１つまたは複数のコンピューティングデバイス（例えば、サーバ）によってホストされる１つまたは複数のデータベースを含み得る。いくつかの実施形態では、データストア１０８Ａは、１つまたは複数の物理ストレージデバイスを含み得る。一例として、物理ストレージデバイス（単数または複数）は、１つまたは複数のソリッドステートドライブ、ハードディスクドライブ、フラッシュドライブ、および／または光学ドライブを含み得る。いくつかの実施形態では、データストア１０８Ａは、データを格納する１つまたは複数のファイルを含み得る。一例として、データストア１０８Ａは、データを格納する１つまたは複数のテキストファイルを含み得る。別の例として、データストア１０８Ａは、１つまたは複数のＸＭＬファイ

ルを含み得る。いくつかの実施形態では、データストア１０８Ａは、コンピューティングデバイスのストレージ（例えば、ハードドライブ）であり得る。いくつかの実施形態では、データストア１０８Ａは、クラウドストレージシステムであり得る。

【００８１】

いくつかの実施形態では、ネットワーク１１１は、無線ネットワーク、有線ネットワーク、またはそれらの任意の適切な組み合わせであり得る。一例として、ネットワーク１１１は、インターネットなどのワイドエリアネットワーク（ＷＡＮ）であり得る。いくつかの実施形態では、ネットワーク１１１は、ローカルエリアネットワーク（ＬＡＮ）であり得る。ローカルエリアネットワークは、シークエンシングデバイス（単数または複数）１０２、アセンブリシステム１０４、モデルトレーニングシステム１０６、およびデータストア１０８Ａの間の有線接続および／または無線接続によって形成され得る。いくつかの実施形態は、本明細書に記載の特定のタイプのネットワークに限定されない。

【００８２】

図１Ｂは、遺伝子アセンブリを生成するように構成された場合の例示的なシステム１００を示す。遺伝子アセンブリは、ゲノムアセンブリまたは遺伝子配列であり得る。例えば、出力されるアセンブリ１１２は、遺伝子アセンブリであり得る。シークエンシングデバイス（単数または複数）１０２は、核酸サンプル１１０をシークエンシングしてヌクレオチド配列を生成するように構成され得る。一例として、シークエンシングデバイス（単数または複数）１０２は、生物からのＤＮＡサンプルをシークエンシングして、ヌクレオチド配列を生成し得る。シークエンシングデバイス（単数または複数）１０２によって生成されたヌクレオチド配列は、データストア１０８Ｂに格納され得る。アセンブリシステム１０４は、機械学習モデル１０４Ａを使用して遺伝子アセンブリを生成するように構成され得る。一例として、アセンブリシステム１０４は、（１）シークエンシングデバイス（単数または複数）１０２によって生成されたヌクレオチド配列にアセンブリ技術（例えば、ＯＬＣ）を適用することによって初期遺伝子アセンブリを取得し、（２）機械学習モデル１０４Ａを使用して初期遺伝子アセンブリを更新して、遺伝子アセンブリ１１２を取得し得る。

【００８３】

図１Ｃは、タンパク質配列を生成するように構成された場合の例示的なシステム１００を示す。例えば、出力されるアセンブリ１１２は、タンパク質配列であり得る。シークエンシングデバイス（単数または複数）１０２は、タンパク質サンプル１１０をシークエンシングしてアミノ酸配列を生成するように構成され得る。一例として、シークエンシングデバイス（単数または複数）１０２は、タンパク質からペプチドをシークエンシングして、アミノ酸配列を生成し得る。シークエンシングデバイス（単数または複数）１０２によって生成されたアミノ酸配列は、データストア１０８Ｃに格納され得る。アセンブリシステム１０４は、機械学習モデル１０４Ａを使用してタンパク質配列を生成するように構成され得る。一例として、タンパク質シークエンシングシステム１０４は、（１）シークエンシングデバイス（単数または複数）１０２によって生成されたアミノ酸配列にアセンブリアルゴリズムを適用することによってタンパク質配列を取得し、（２）機械学習モデル１０４Ａを使用してタンパク質配列を更新して、タンパク質配列を取得し得る。

【００８４】

図２Ａは、本明細書に記載の技術のいくつかの実施形態による、アセンブリを生成するためのアセンブリシステム２００を示す。アセンブリシステム２００は、図１Ａ～図１Ｃを参照して上記で説明したアセンブリシステム１０４であり得る。アセンブリシステム２００は、シークエンシングデータ２０２を使用してアセンブリ２０４を生成するように構成されたコンピューティングデバイスであり得る。アセンブリシステム２００は、特徴生成器２００Ａおよび機械学習モデル２００Ｂを含む複数の構成要素を含む。アセンブリシステム２００Ｃは、任意選択的に、アセンブラ２００Ｃを含み得る。

【００８５】

いくつかの実施形態では、特徴生成器２００Ａは、機械学習モデルへの入力として提供

され得る１つまたは複数の特徴の値を決定するように構成され得る。特徴生成器２００Ａは、（１）配列データ２０２、および（２）アセンブリ（例えば、配列データ２０２へのアセンブリアルゴリズムの適用により得られる）から特徴（単数または複数）の値を決定するように構成され得る。配列データ２０２は、アセンブリを生成するためにアセンブリアルゴリズムによって使用される複数の配列を含み得る。いくつかの実施形態では、特徴生成器２００Ａは、配列の各々をアセンブリと比較することによって特徴（単数または複数）の値を決定するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、配列をアセンブリの一部と整列させるように構成され得る。例えば、特徴生成器２００Ａは、配列をアセンブリ内の一組の位置に整列させ得、アセンブリ内の一組の位置における生物学的ポリマーの表示は、整列された配列から決定されたものである。特徴生成器２００Ａは、整列された配列を、アセンブリ内の一組の位置において示される生物学的ポリマー（例えば、ヌクレオチド、アミノ酸）と比較することによって、特徴（単数または複数）の値を決定するように構成され得る。特徴（単数または複数）の値を決定するための例示的な技術は、図４Ａ～図４Ｃを参照して以下に説明される。

【００８６】

図２Ａの実施形態に示されるように、特徴生成器２００Ａは、機械学習モデル２００Ｂに提供される入力を生成するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、アセンブリ内の複数の位置の各々に対して入力を生成するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、複数の位置を選択し、選択された複数の位置を使用して入力を生成するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、アセンブリが複数の位置において生物学的ポリマーを不正確に示す複数の尤度を決定し、決定された複数の尤度を使用して複数の位置を選択することによって複数の位置を選択するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、アセンブリ内に示された生物学的ポリマーとは異なる生物学的ポリマーを特定する位置に整列された配列の数に基づいて、アセンブリが、ある位置において生物学的ポリマーを不正確に示す尤度を決定するように構成され得る。特徴生成器２００Ａは、尤度が閾値尤度を超えると決定されたときに、その位置に対する入力を生成するように構成され得る。

【００８７】

いくつかの実施形態では、特徴生成器２００Ａは、（１）ターゲット位置において同定される生物学的ポリマー、（２）ターゲット位置の近傍の１つまたは複数の他の位置において同定される生物学的ポリマーを使用して、アセンブリ内のターゲット位置に関して機械学習モデル２００Ｂに提供される入力を生成するように構成され得る。いくつかの実施形態では、特徴生成器２００Ａは、ターゲット位置およびターゲット位置の近傍にある他の位置（単数または複数）における特徴値を決定するように構成され得る。近傍の他の位置（単数または複数）における特徴値は、ターゲット位置に関する出力を生成するために機械学習モデル２００Ａにコンテキスト情報を提供し得る。いくつかの実施形態では、近傍のサイズは、設定可能なパラメータであり得る。例えば、近傍のサイズは、ソフトウェアアプリケーションにおけるユーザ入力によって指定され得る。

【００８８】

いくつかの実施形態では、特徴生成器２００Ａは、ターゲット位置の近傍の位置において決定された特徴値を含むウィンドウとして入力を生成するように構成され得る。ターゲット位置の近傍は、ターゲット位置と、ターゲット位置のウィンドウ内の１つまたは複数の他の位置とを含み得る。いくつかの実施形態では、ウィンドウのサイズは、２個の位置、３個の位置、５個の位置、１０個の位置、１５個の位置、２０個の位置、２５個の位置、３０個の位置、３５個の位置、４０個の位置、４５個の位置、または５０個の位置であり得る。いくつかの実施形態では、特徴生成器２００Ａは、６０個の位置、７０個の位置、８０個の位置、９０個の位置、または１００個の位置の近傍のサイズを使用するように構成され得る。いくつかの実施形態では、ウィンドウは、ターゲット位置を中心にして配置され得る。

【００８９】
いくつかの実施形態では、機械学習モデル２００Ｂは、図１Ａ～図１Ｃを参照して上記で説明した機械学習モデル１０４Ａであり得る。図１Ａの実施形態に示されるように、機械学習モデル２００Ｂは、特徴生成器２００Ａからの入力を受信するように構成され得る。機械学習モデル２００Ｂは、特徴生成器２００Ａによって提供される個々の入力に対応する出力を生成するように構成され得る。機械学習モデル２００Ｂは、アセンブリ内の複数の位置における生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）を同定するためにアセンブリシステム２００によって使用される出力を生成するように構成され得る。いくつかの実施形態では、機械学習モデル２００Ｂは、位置に関して、複数の生物学的ポリマーの各々がその位置に存在する尤度を出力するように構成され得る。一例として、機械学習モデル２００Ｂは、複数のヌクレオチドの各々に関して、ヌクレオチドがその位置に存在する確率を出力し得る。別の例として、機械学習モデル２００Ｂは、複数のアミノ酸の各々に関して、アミノ酸がその位置に存在する確率を出力し得る。いくつかの実施形態では、アセンブリシステム２００は、アセンブリ内のある位置における生物学的ポリマーを、機械学習モデル２００Ｂの出力によって示されるような、生物学的ポリマーのその位置において存在する尤度が最も高い生物学的ポリマーであると同定するように構成され得る。一例として、アセンブリシステム２００は、複数のヌクレオチドの中から、その位置に存在する可能性が最も高いヌクレオチドを選択し得る。別の例として、アセンブリシステム２００は、複数のアミノ酸の中から、その位置に存在する可能性が最も高いアミノ酸を選択し得る。

【００９０】
いくつかの実施形態では、アセンブリシステム２００は、機械学習モデル２００Ｂから取得した出力を使用して、出力アセンブリ２０４を生成するように構成され得る。アセンブリシステム２００は、機械学習モデル２００Ｂから取得された出力からアセンブリ内の位置において同定された生物学的ポリマーを使用してアセンブリを更新するように構成され得る。アセンブリシステム２００は、アセンブリ内の位置において同定された生物学的ポリマーを示すようにアセンブリを更新して、出力アセンブリ２０４を取得するように構成され得る。一例として、アセンブリは、アセンブリ内の第１の位置においてアデニンを示し、アセンブリ内の第２の位置においてグアニンを示し得る。この例では、アセンブリシステム２００は、（１）機械学習モデル２００Ｂから取得された出力を使用して、第１の位置におけるヌクレオチドがチミンであり、第２の位置におけるヌクレオチドがグアニンであることを同定し、（２）アセンブリ内の第１の位置をチミンを示すように更新し、第２の位置において示されたヌクレオチドを変更せずに維持して、出力アセンブリ２０４を生成し得る。上記の例によって示されるように、アセンブリシステム２００は、他の位置（単数または複数）における生物学的ポリマーの表示を変更せずに、機械学習モデル２００Ｂから取得された出力を使用して、アセンブリ内の位置（単数または複数）における生物学的ポリマーの表示を変更し得る。例えば、アセンブリシステム２００は、アセンブリ内のある位置において同定された生物学的ポリマーが、アセンブリで示された生物学的ポリマーと一致することを決定して、更新されたアセンブリ内でその位置における表示を変更せずに維持し得る。

【００９１】
図１Ａの実施形態に示されるように、アセンブラ２００Ｃは、アセンブリを特徴生成器２００Ａに提供するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、アセンブリアルゴリズムを（例えば、高分子サンプルのシークエンシングから受信される）配列データ２０２に適用することによって、特徴生成器２００Ａに提供されるアセンブリを生成するように構成され得る。一例として、アセンブラ２００Ｃは、アセンブリアルゴリズムを、配列データ２０２に含まれるヌクレオチド配列に適用して、アセンブリを生成するように構成され得る。次に、アセンブリ内の位置における生物学的ポリマーを同定するための出力を取得するために機械学習モデル２００Ｂに提供される入力を生成するために、アセンブリが特徴生成器２００Ａに提供され得る。アセンブラ２００Ｃによっ

て生成されたアセンブリは、出力アセンブリ２０４を生成するために、機械学習モデル２００Ｂから取得された出力を使用してアセンブリシステム２００によって更新され得る。

【００９２】

いくつかの実施形態では、アセンブラ２００Ｃは、オーバーレイ・レイアウト・コンセンサス（ＯＬＣ：ｏｖｅｒｌａｙ　ｌａｙｏｕｔ　ｃｏｎｓｅｎｓｕｓ）アルゴリズムを、配列データ２０２に含まれるヌクレオチド配列に適用して、アセンブリを生成するように構成され得る。シークエンシングデバイスは、核酸（単数または複数）を含む生物学的サンプルの複数のコピーをシークエンシングし得る。結果として、配列データ２０２は、アセンブリの各部分（例えば、一組の位置）に関して、アセンブリの一部に整列する複数の配列を含み得る。アセンブリ内の位置をカバーする配列の平均数は、配列の「カバレッジ」と呼ばれ得る。アセンブラ２００Ｃは、（１）配列の重複領域に基づいて重複グラフを生成し、（２）重複グラフを使用して、アセンブリの個々の一部に整列する配列（「コンティグ（ｃｏｎｔｉｇｓ）」とも呼ばれる）のレイアウトを生成し、（３）アセンブリの一部に整列する各組の配列に関して、アセンブリの一部を生成するために組内の配列のコンセンサスを取ることによって、ＯＬＣアルゴリズムを配列に適用するように構成され得る。

【００９３】

いくつかの実施形態では、アセンブラ２００Ｃは、配列のペアを比較して、それらが生物学的ポリマー（例えば、ヌクレオチド）の１つまたは複数の同一の部分配列を含むかどうかを決定することによって、重複領域を有する配列を同定するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、（１）少なくともヌクレオチドの閾値数（例えば、３、４、５、６、８、１０、２０、３０、４０、５０、６０、７０、８０、９０、１００、２００、３００、４００、５００）の同一の部分配列（単数または複数）を共有する配列のペアを重複配列として同定し、（２）各重複領域の長さ（即ち、ヌクレオチドの数）を決定し、（３）同定された重複配列および重複領域の長さに基づいて重複グラフを生成するように構成され得る。重複グラフは、重複する配列の個々のペアを接続する頂点およびエッジとしての配列を含み得る。決定された長さは、重複グラフにおけるエッジのラベルとして使用され得る。

【００９４】

いくつかの実施形態では、アセンブラ２００Ｃは、重複グラフを使用して配列を連結することによって、アセンブリの個々の一部に整列された複数組の配列のレイアウトを生成するように構成され得る。アセンブラ２００Ｃは、配列を連結するために重複グラフを通るパスを発見するように構成され得る。一例として、アセンブラ２００Ｃは、連結された配列を取得するためにヌクレオチドを表す一組の英数字を連結し得る。いくつかの実施形態では、アセンブラ２００Ｃは、グリーディアルゴリズム（ｇｒｅｅｄｙ　ａｌｇｏｒｉｔｈｍ）を重複グラフに適用して、連結された配列を同定し得る。一例として、アセンブラ２００Ｃは、グリーディアルゴリズムを適用して、最短共通超文字列（ｓｈｏｒｔｅｓｔ　ｃｏｍｍｏｎ　ｓｕｐｅｒｓｔｒｉｎｇ）を連結された配列として同定し得る。

【００９５】

いくつかの実施形態では、アセンブラ２００Ｃは、レイアウト配列を使用してアセンブリを生成するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、各組がアセンブリの一部と整列する、複数の組のレイアウト配列を同定し得る。アセンブラ２００Ｃは、アセンブリの一部と整列するレイアウト配列のコンセンサスを取ることによって、アセンブリの一部を生成するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、アセンブリの一部内のある位置における生物学的ポリマー（例えば、ヌクレオチド）が、アセンブリの一部に整列した配列の大多数がその位置にあることを示す生物学的ポリマーであると決定することによって、コンセンサスを取るように構成され得る。一例として、アセンブラ２００Ｃは、ヌクレオチド配列の重複グラフを生成し、アセンブリ内の一組の４個の位置に対応する４個のヌクレオチド配列「ＴＡＧＡ」、「ＴＡＧＡ」、「ＴＡＧＴ」、「ＴＡＧＡ」、および「ＴＡＧＣ」を同定し得る。この例では、

アセンブラ２００Ｃは、４個のヌクレオチド配列の全てが最初の３個の位置が「ＴＡＧ」であることを示し、ヌクレオチド配列の大多数が４番目の位置が「Ａ」であることを示すので、４個のヌクレオチド配列間のコンセンサスを「ＴＡＧＡ」と決定し得る。

【００９６】

いくつかの実施形態では、アセンブリシステム２００は、機械学習技術を使用してＯＬＣアルゴリズムのコンセンサスステップを実行するように構成され得る。アセンブラ２００Ｃがアセンブリを生成するために使用されるレイアウトを生成すると、システムは、レイアウトおよびレイアウトから取得されたコンセンサスアセンブリを使用して機械学習モデルへの入力を生成するように構成され得る。いくつかの実施形態では、アセンブリシステム２００は、出力アセンブリ２０４を得るために、本明細書に記載の技術を使用してコンセンサスアセンブリを更新するように構成され得る。

【００９７】

いくつかの実施形態では、アセンブラ２００Ｃは、参照により本明細書に組み込まれる、ゲノミクス（Ｇｅｎｏｍｉｃｓ）、第９５巻、第６号、２０１０年６月に公開された「次世代シークエンシングデータのためのアセンブリアルゴリズム（Ａｓｓｅｍｂｌｙ Ａｌｇｏｒｉｔｈｍｓ ｆｏｒ Ｎｅｘｔ－Ｇｅｎｅｒａｔｉｏｎ Ｓｅｑｕｅｎｃｉｎｇ Ｄａｔａ）」に記載されたシークエンシングデータ２０２にアルゴリズムを適用するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、ＯＬＣアルゴリズム以外のアセンブリアルゴリズムを配列データ２０２に適用してアセンブリを生成するように構成され得る。いくつかの実施形態では、アセンブラ２００Ｃは、ド・ブラウン・グラフ（ＢＢＧ）アセンブリを配列データ２０２に適用するように構成され得る。いくつかの実施形態は、特定のタイプのアセンブリアルゴリズムに限定されない。いくつかの実施形態では、アセンブラ２００Ｃは、配列データ２０２を使用してアセンブリを生成するように構成されたソフトウェアアプリケーションを含み得る。一例として、システムは、ＨＧＡＰアセンブラ、ファルコン（Ｆａｌｃｏｎ）アセンブラ、カヌ（Ｃａｎｕ）アセンブラ、ヒンジ（Ｈｉｎｇｅ）アセンブラ、ミニアスム（Ｍｉｎｉａｓｍ）アセンブラ、またはフライ（Ｆｌｙｅ）アセンブラを含み得る。別の例として、システムは、ＳＰＡｄｅｓアセンブリアプリケーション、レイ（Ｒａｙ）アセンブリアプリケーション、ＡＢｙＳＳアセンブリアプリケーション、ＡＬＬＰＡＨＳＴＳ－ＬＧアセンブリアプリケーション、またはトリニティ（Ｔｒｉｎｉｔｙ）アセンブリアプリケーションを含み得る。いくつかの実施形態は、特定のアセンブラに限定されない。

【００９８】

図２Ａの破線によって示されるように、いくつかの実施形態では、アセンブラ２００Ｃは、アセンブリシステムに含まれなくてもよい。アセンブリシステム２００は、別個のシステムからアセンブリを受信し、受信したアセンブリを更新して出力アセンブリ２０４を生成するように構成され得る。一例として、別個のコンピューティングデバイスは、アセンブリアルゴリズム（例えば、ＯＬＣ）を配列データ２０２に適用して、アセンブリを生成し、生成されたアセンブリをアセンブリシステム２００に送信し得る。

【００９９】

図２Ｂは、図２Ａを参照して上記のアセンブリシステム２００の実施形態を示し、アセンブリシステム２００は、機械学習モデル２００Ｂから特徴生成器２００Ａへのフィードバック矢印によって示されるように、アセンブリに対する更新の複数の反復を実行するように構成される。いくつかの実施形態では、アセンブリシステム２００は、第１の更新されたアセンブリを取得した後、機械学習モデル２００Ｂへの入力として提供され得る１つまたは複数の特徴の値を決定するように構成され得る。特徴生成器２００Ａは、（１）配列データ２０２と、（２）アセンブリアルゴリズムの配列データ２０２への適用から取得された初期アセンブリを更新することから取得された第１の更新されたアセンブリとから特徴（単数または複数）の値を決定するように構成され得る。特徴生成器２００Ａは、出力を得るために決定された特徴（単数または複数）の値を機械学習モデル２００Ｂへの入力として提供するように構成され得る。アセンブリシステム２００は、機械学習モデル２

００Ｂからの出力を使用して、（１）第１の更新されたアセンブリ内の個々の位置における生物学的ポリマーを同定し、（２）個々の位置において同定された生物学的ポリマーを示すように第１の更新されたアセンブリを更新して、第２の更新されたアセンブリを取得するように構成され得る。第２の更新されたアセンブリは、アセンブリシステム２００によって出力されたアセンブリ２０４であり得る。

【０１００】

いくつかの実施形態では、アセンブリシステム２００は、条件が満たされるまで更新の反復を実行するように構成され得る。いくつかの実施形態では、アセンブリシステム１０４は、閾値回数の反復が実行されたとシステムが判定するまで、更新の反復を実行するように構成され得る。いくつかの実施形態では、反復の閾値回数は、ユーザ入力（例えば、ソフトウェアコマンド、またはハードコードされた値）によって設定され得る。いくつかの実施形態では、アセンブリシステム１０４は、反復の閾値回数を決定するように構成され得る。一例として、アセンブリシステム２００は、初期アセンブリを取得するために使用されたアセンブリ技術のタイプに基づいて、更新の反復の閾値回数を決定し得る。いくつかの実施形態では、アセンブリシステム２００は、指定された停止基準が満たされるまで、アセンブリを反復して更新するように構成され得る。一例として、アセンブリシステム２００は、（１）最新の更新の反復から取得された現在のアセンブリと前のアセンブリとの間の差異の数を決定し、（２）差異の数が差異の閾値数より少ない場合、および／または差異のパーセンテージが閾値パーセンテージより少ない場合、アセンブリの反復した更新を停止するように決定し得る。

【０１０１】

図２Ｃは、図２Ａを参照して上記のアセンブリシステム２００の実施形態を示し、アセンブリシステム２００は、特徴生成器２００Ａから機械学習モデル２００Ｂへの複数の矢印によって示されるように、アセンブリの複数の位置を並列に修正するように構成される。図２Ａを参照して説明したように、いくつかの実施形態では、特徴生成器２００Ａは、複数の位置の各々に関して、機械学習モデル２００Ｂに提供される入力を生成するように構成され得る。図２Ｃの実施形態では、アセンブリシステム２００は、アセンブリの複数の位置を並列に更新するように構成され得る。アセンブリシステム２００は、（１）アセンブリ内の第１の位置を更新し、（２）アセンブリ内の第１の位置の更新を完了する前に、アセンブリ内の第２の位置の更新を開始するように構成され得る。いくつかの実施形態では、アセンブリシステム２００は、複数の入力を並列に生成すること、かつ／または複数の個々の位置に対して生成された複数の入力を機械学習モデル２００Ｂに並列に提供することによって、複数の位置を並列に更新するように構成され得る。一例として、特徴生成器２００Ａは、（１）機械学習モデル２００Ｂへの第１の位置に関する第１の入力を生成および／または提供し、（２）機械学習モデル２００Ｂから第１の入力に対応する出力を取得する前に、機械学習モデル２００Ｂへの第２の位置に関する第２の入力を生成および／または提供し得る。

【０１０２】

いくつかの実施形態では、図２Ｃのアセンブリシステム２００は、アセンブリの複数の位置を並列に更新するように構成された複数のプロセッサを含むコンピューティングデバイスであり得る。いくつかの実施形態では、アセンブリシステム２００は、マルチスレッドアプリケーションを使用するように構成され得、アプリケーションの各スレッドは、アセンブリ内の個々の位置を１つまたは複数の他のスレッドと並列に更新するように構成される。

【０１０３】

図２Ｄは、図２Ａを参照して上記のアセンブリシステム２００の実施形態を示し、アセンブリシステム２００は、（１）機械学習モデル２００Ｂから特徴生成器２００Ａへの矢印によって示されるように、更新の複数の反復を実行し、（２）特徴生成器２００Ａから機械学習モデル２００Ｂへの複数の矢印によって示されるように、アセンブリの複数の位置を並列に修正するように構成されている。いくつかの実施形態では、アセンブリシステ

ム２００は、図２Ｂを参照して上記のように複数の更新の反復を実行し、各更新サイクル中に、図２Ｃを参照して上記のようにアセンブリ内の複数の位置を並列に更新するように構成され得る。

【０１０４】

図３Ａは、本明細書に記載の技術のいくつかの実施形態による、生物学的ポリマーアセンブリを生成するために機械学習モデルをトレーニングするための例示的なプロセス３００を示す。プロセス３００は、任意の適切なコンピューティングデバイス（単数または複数）によって実行され得る。一例として、プロセス３００は、図１Ａ～図１Ｃを参照して説明されたモデルトレーニングシステム１０６によって実行され得る。プロセス３００は、本明細書で説明される機械学習モデルをトレーニングするために実行され得る。一例として、プロセス３００が、図６を参照して説明した畳み込みニューラルネットワーク（ＣＮＮ）６００などの深層学習モデルをトレーニングするために実行され得る。

【０１０５】

いくつかの実施形態では、機械学習モデルは、深層学習モデルであり得る。いくつかの実施形態では、深層学習モデルはニューラルネットワークであり得る。例として、機械学習モデルは、アセンブリ内の複数の位置における生物学的ポリマー（例えば、ヌクレオチド、アミノ酸）を同定する際に使用される出力を生成する畳み込みニューラルネットワーク（ＣＮＮ）であり得る。別の例として、機械学習モデルは、ＣＴＣ適合ニューラルネットワークであり得る。いくつかの実施形態では、深層学習モデルの一部は、個別にトレーニングされ得る。一例として、深層学習モデルは、入力データを１つまたは複数の特徴（単数または複数）の値にエンコードする第１の部分と、特徴（単数または複数）の値を入力として受信して、１つまたは複数の生物学的ポリマーを同定する出力を生成する第２の部分とを有し得る。

【０１０６】

いくつかの実施形態では、機械学習モデルは、クラスタリングモデルであり得る。いくつかの実施形態では、モデルの各クラスタは、生物学的ポリマーに関連付けられ得る。例示的な例として、クラスタリングモデルは５つのクラスタを含み得、各クラスタは個々のヌクレオチドに関連付けられている。例えば、第１のクラスタはアデニンに関連付けられ得、第２のクラスタはシトシンに関連付けられ得、第３のクラスタはグアニンに関連付けられ得、第４のクラスタはチミンに関連付けられ得、第５のクラスタは、（例えば、アセンブリ内のある位置において）ヌクレオチドが存在しないことを示し得る。クラスタおよび関連する生物学的ポリマーの例示的な数は、例示の目的で本明細書に記載されている。

【０１０７】

プロセス３００は、ブロック３０２で開始し、プロセス３００を実行するシステムは、１つまたは複数の参照高分子（例えば、ＤＮＡ、ＲＮＡ、またはタンパク質）のシークエンシングによるシークエンシングデータにアクセスする。いくつかの実施形態では、システムは、参照高分子のシークエンシングによるシークエンシングデータにデータベースからアクセスするように構成され得る。一例として、システムは、細菌のシークエンシングにより取得されたシークエンシングデータにＯＮＧデータベースからアクセスし得る。シークエンシングデータは、高分子の１つまたは複数のサンプルをシークエンシングすることにより取得され得る。一例として、シークエンシングデータは、酵母の一種であるサッカロミセス・セレビシエ（Ｓａｃｃｈａｒｏｍｙｃｅｓ　ｃｅｒｅｖｉｓｉａｅ）の生物学的サンプルから取得され得る。別の例として、シークエンシングデータは、タンパク質のペプチドサンプルをシークエンシングすることから取得され得る。いくつかの実施形態では、シークエンシングデータは、核酸（例えば、ＤＮＡ、ＲＮＡ）を含む生物学的サンプルをシークエンシングすることから取得されたヌクレオチド配列を含み得る。いくつかの実施形態では、シークエンシングデータは、タンパク質サンプル（例えば、タンパク質からのペプチド）をシークエンシングすることから取得されたアミノ酸配列を含み得る。

【０１０８】

いくつかの実施形態では、システムは、機械学習モデルが、ターゲットシークエンシン

グ技術によって生成されたシークエンシングデータから生成されたアセンブリの精度を向上させるようにトレーニングされ得るように、ターゲットシークエンシング技術によるシークエンシングデータにアクセスするように構成され得る。機械学習モデルは、機械学習モデルがターゲットシークエンシング技術の特徴的なエラーを修正するために最適化され得るように、ターゲットシークエンシング技術のエラープロファイルに関してトレーニングされ得る。いくつかの実施形態では、システムは、第三世代シークエンシングにより取得されたデータにアクセスするように構成され得る。いくつかの実施形態では、第三世代シークエンシングは、１分子リアルタイムシークエンシングであり得る。一例として、システムは、ヌクレオチドに結合された発光分子による発光を検出することによって核酸サンプルをシークエンシングするシステムから取得されたデータにアクセスし得る。別の例として、システムは、アミノ酸と選択的に相互作用する試薬に結合された発光分子による発光を検出することによってペプチドをシークエンシングするシステムから取得されたデータにアクセスし得る。いくつかの実施形態では、システムは、第２世代シークエンシングから取得されたデータにアクセスするように構成され得る。一例として、システムは、サンガー・シークエンシング（Ｓａｎｇｅｒ　ｓｅｑｕｅｎｃｉｎｇ）、マキサムギルバート・シークエンシング（Ｍａｘａｍ－Ｇｉｌｂｅｒｔ　ｓｅｑｕｅｎｃｉｎｇ）、ショットガン・シークエンシング（ｓｈｏｔｇｕｎ　ｓｅｑｕｅｎｃｉｎｇ）、パイロ・シークエンシング（ｐｙｒｏｓｅｑｕｅｎｃｉｎｇ）、コンビナトリアル・プローブ・アンカー合成（ｃｏｍｂｉｎａｔｏｒｉａｌ　ｐｒｏｂｅ　ａｎｃｈｏｒ　ｓｙｎｔｈｅｓｉｓ）、またはライゲーション（ｌｉｇａｔｉｏｎ）によるシークエンシングから取得されたシークエンシングデータにアクセスし得る。いくつかの実施形態では、システムは、デノボ・ペプチド・シークエンシング（ｄｅ　ｎｏｖｏ　ｐｅｐｔｉｄｅ　ｓｅｑｕｅｎｃｉｎｇ）から取得されたデータにアクセスするように構成され得る。一例として、システムは、タンデム質量分析（ｔａｎｄｅｍ　ｍａｓｓ　ｓｐｅｃｔｒｏｍｅｔｒｙ）から取得されたアミノ酸配列にアクセスし得る。いくつかの実施形態は、特定のターゲットシークエンシング技術に限定されない。

【０１０９】

次に、プロセス３００はブロック３０４に移行し、システムは、ブロック３０２で取得されたシークエンシングデータの少なくとも一部から生成されたアセンブリにアクセスする。いくつかの実施形態では、システムは、アセンブリアルゴリズム（例えば、ＯＬＣアセンブリ、ＤＢＧアセンブリ）のシークエンシングデータへの適用により取得されたアセンブリにアクセスするように構成され得る。いくつかの実施形態では、システムは、アセンブリアルゴリズムをシークエンシングデータに適用することによってアセンブリにアクセスするように構成され得る。いくつかの実施形態では、システムは、１つまたは複数のアセンブリアルゴリズムのシークエンシングデータへの適用により生成された所定のアセンブリにアクセスするように構成され得る。一例として、アセンブリは、以前に別のコンピューティングデバイスによって実行され、データベースに格納されてもよい。例えば、シークエンシングデータが取得されたデータベースは、１つまたは複数のアセンブリアルゴリズムのシークエンシングデータへの適用により生成されたアセンブリをも格納し得る。

【０１１０】

いくつかの実施形態では、システムは、ターゲットアセンブリ技術により生成されたアセンブリにアクセスするように構成され得、機械学習モデルは、ターゲットアセンブリ技術の特徴的なエラーを修正するようにトレーニングされ得る。機械学習モデルは、機械学習モデルがターゲットアセンブリ技術の特徴的なエラーを修正するために最適化され得るように、ターゲットアセンブリ技術のエラープロファイルに関してトレーニングされ得る。いくつかの実施形態では、システムは、特定のアセンブリアルゴリズムおよび／またはソフトウェアアプリケーションによって生成されたアセンブリにアクセスするように構成され得る。例として、システムは、カヌ（Ｃａｎｕ）アセンブラ、ミニアスム（Ｍｉｎｉａｓｍ）アセンブラ、またはフライ（Ｆｌｙｅ）アセンブラによって生成されたアセンブ

リにアクセスし得る。いくつかの実施形態では、システムは、アセンブラのクラスから生成されたアセンブリにアクセスするように構成され得る。一例として、システムは、グリーディ・アルゴリズムアセンブラまたはグラフメソッド・アセンブラから生成されたアセンブリにアクセスし得る。いくつかの実施形態は、特定のアセンブリ技術に限定されない。

【０１１１】

次に、プロセス３００は、ブロック３０６に移行し、システムは、参照高分子（単数または複数）の１つまたは複数の所定のアセンブリにアクセスする。いくつかの実施形態では、参照高分子（単数または複数）の所定のアセンブリは、個々の高分子（単数または複数）に関する真のまたは正確なアセンブリを表し得る。従って、システムは、参照高分子（単数または複数）の所定のアセンブリを使用してトレーニングデータにラベルを付けるように構成され得る。一例として、システムは、ＮＣＢＩデータベースから生物のＤＮＡの参照ゲノムにアクセスし得る。この例では、システムは参照ゲノムを使用して、ゲノムアセンブリ内のヌクレオチドを同定するための機械学習モデルをトレーニングするための教師あり学習の実行の際に使用するラベルを決定し得る。別の例として、システムは、ユニットプロット（ＵｎｉｔＰｒｏｔ）データベースからタンパク質の参照タンパク質配列にアクセスし、参照タンパク質配列を使用して、タンパク質配列内のアミノ酸を同定するための機械学習モデルをトレーニングするための教師あり学習の実行の際に使用するラベルを決定し得る。

【０１１２】

次に、プロセス３００はブロック３０８に移行し、システムは、ブロック３０２～３０８でアクセスされるデータを使用して機械学習モデルをトレーニングする。いくつかの実施形態では、システムは、（１）ブロック３０２においてアクセスされたシークエンシングデータおよびブロック３０４においてアクセスされたアセンブリを使用して、機械学習モデルへの入力を生成し、（２）ブロック３０６においてアクセスされた所定のアセンブリを使用して、生成された入力にラベルを付け、（３）ラベル付けされたトレーニングデータに教師あり学習アルゴリズムを適用するように構成され得る。いくつかの実施形態では、システムは、シークエンシングデータを使用して１つまたは複数の特徴の値を生成することによって、機械学習モデルへの入力を生成するように構成され得る。いくつかの実施形態では、システムは、アセンブリ内の各位置に対する特徴（単数または複数）の値を決定するように構成され得る。一例として、システムは、（１）個々のヌクレオチドに対するカウントを決定し、各カウントは、ヌクレオチドがその位置に存在することを示すヌクレオチド配列の数を示し、（２）カウントを使用して特徴（単数または複数）の値を決定することによって、位置に関する特徴の値を決定し得る。入力を生成して、入力にラベルを付けるための例示的な技術は、図４Ａ～図４Ｃを参照して本明細書に記載されている。

【０１１３】

いくつかの実施形態では、システムは、ラベル付けされたトレーニングデータを使用して深層学習モデルをトレーニングするように構成され得る。いくつかの実施形態では、システムは、ラベル付けされたトレーニングデータを使用して決定木モデルをトレーニングするように構成され得る。いくつかの実施形態では、システムは、ラベル付けされたトレーニングデータを使用してサポートベクターマシン（ＳＶＭ：ｓｕｐｐｏｒｔ　ｖｅｃｔｏｒ　ｍａｃｈｉｎｅ）をトレーニングするように構成され得る。いくつかの実施形態では、システムは、ラベル付けされたトレーニングデータを使用してナイーブベイズ分類器（ＮＢＣ：Ｎａｉｖｅ　Ｂａｙｅｓ　ｃｌａｓｓｉｆｉｅｒ　）をトレーニングするように構成され得る。

【０１１４】

いくつかの実施形態では、システムは、確率的勾配降下法を使用することによって機械学習モデルをトレーニングするように構成され得る。システムは、目的関数を最適化するために機械学習モデルのパラメータを反復的に変更して、トレーニングされた機械学習モ

デルを取得し得る。例えば、システムは確率的勾配降下法を使用して、畳み込みネットワークのフィルタおよび／またはニューラルネットワークの重みをトレーニングし得る。

【０１１５】

いくつかの実施形態では、システムは、ラベル付けされたトレーニングデータを使用して教師ありトレーニングを実行するように構成され得る。いくつかの実施形態では、システムは、（１）機械学習モデルに生成された入力を提供して、対応する出力を取得し、（２）出力を使用してアセンブリ内の複数の位置に存在する生物学的ポリマーを同定し、（２）同定された生物学的ポリマーと参照アセンブリの複数の位置において示されている生物学的ポリマーとの間の差異に基づいて機械学習モデルをトレーニングすることによって機械学習モデルをトレーニングするように構成され得る。参照アセンブリ内のある位置において示される生物学的ポリマーは、個々の入力に関するラベルであり得る。差異は、機械学習モデルが、現在の組のパラメータで構成された場合に、ラベルを再現する際にどの程度良好に動作するかの尺度を提供し得る。例として、機械学習モデルのパラメータは、確率的勾配降下法および／またはモデルのトレーニングに適した他の反復最適化手法を使用して更新され得る。一例として、システムは、決定された差異に基づいてモデルの１つまたは複数のパラメータを更新するように構成され得る。

【０１１６】

いくつかの実施形態では、システムは、教師なしトレーニングアルゴリズムを一組のラベル付けされていないトレーニングデータに適用し得る。図３Ａの実施形態は、ブロック３０６において参照高分子の所定のアセンブリにアクセスすることを含むが、いくつかの実施形態では、システムは、所定のアセンブリにアクセスすることなくトレーニングを実行するように構成され得る。これらの実施形態では、システムは、教師なしトレーニングアルゴリズムをトレーニングデータに適用して、機械学習モデルをトレーニングするように構成され得る。システムは、（１）シークエンシングデータと、シークエンシングデータから生成されたアセンブリとを使用してモデルへの入力を生成し、（２）生成された入力に教師なしトレーニングアルゴリズムを適用することによってモデルをトレーニングするように構成され得る。いくつかの実施形態では、機械学習モデルはクラスタリングモデルであり得、システムは、教師なし学習アルゴリズムをトレーニングデータに適用することによって、クラスタリングモデルのクラスタを識別するように構成され得る。各クラスタは、生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）と関連付けられ得る。一例として、システムは、トレーニングデータを使用してｋ平均クラスタリングを実行して、クラスタ（例えば、クラスタ重心）を識別し得る。

【０１１７】

いくつかの実施形態では、システムは、半教師あり学習アルゴリズムをトレーニングデータに適用するように構成され得る。システムは、（１）教師なし学習アルゴリズム（例えば、クラスタリング）をトレーニングデータに適用することによって一組のラベル付けされていないトレーニングデータにラベルを付け、（２）ラベル付けされたトレーニングデータに教師あり学習アルゴリズムを適用し得る。一例として、システムは、シークエンシングデータから生成された入力およびシークエンシングデータから取得されたアセンブリにｋ平均クラスタリングを適用して、入力をクラスタリングし得る。次に、システムは、クラスタメンバーシップに基づく分類によって各入力にラベルを付け得る。次に、システムは、確率的勾配降下アルゴリズムおよび／または他の反復最適化手法をラベル付けされたデータに適用することによって、機械学習モデルをトレーニングし得る。

【０１１８】

ブロック３０８において機械学習モデルをトレーニングした後、プロセス３００は終了する。いくつかの実施形態では、システムは、トレーニングされた機械学習モデルを格納するように構成され得る。システムは、機械学習モデルの１つまたは複数のトレーニングされたパラメータの値（単数または複数）を保存し得る。一例として、機械学習モデルは、１つまたは複数のニューラルネットワークを含み得、システムは、ニューラルネットワーク（単数または複数）のトレーニングされた重みの値を格納し得る。別の例として、機

械学習モデルは畳み込みニューラルネットワークを含み、システムは畳み込みニューラルネットワークの１つまたは複数のトレーニングされたフィルタを格納し得る。いくつかの実施形態では、システムは、アセンブリ（例えば、ゲノムアセンブリ、タンパク質配列、またはそれらの一部）を生成する際に使用するためのトレーニングされた機械学習モデルを（例えば、アセンブリシステム１０４内に）格納するように構成され得る。

【０１１９】

いくつかの実施形態では、システムは、新たなデータを取得し、新たなトレーニングデータを使用して機械学習モデルを更新するように構成され得る。いくつかの実施形態では、システムは、新たなトレーニングデータを使用して新たな機械学習モデルをトレーニングすることによって機械学習モデルを更新するように構成され得る。一例として、システムは、新たなトレーニングデータを使用して新たな機械学習モデルをトレーニングし得る。いくつかの実施形態では、システムは、新たなトレーニングデータを使用して機械学習モデルを再トレーニングして、機械学習モデルの１つまたは複数のパラメータを更新することによって機械学習モデルを更新するように構成され得る。一例として、モデルによって生成された出力（単数または複数）および対応する入力データは、以前に取得されたトレーニングデータとともにトレーニングデータとして使用され得る。いくつかの実施形態では、システムは、（例えば、図３Ｂを参照して以下に説明するプロセス３１０を実行することから得られる）アミノ酸を同定するデータおよび出力を使用して、トレーニングされた機械学習モデルを反復して更新するように構成され得る。一例として、システムは、第１のトレーニングされた機械学習モデル（例えば、教師モデル）に入力データを提供して、１つまたは複数のアミノ酸を同定する出力を取得するように構成され得る。次に、システムは、入力データおよび対応する出力を使用して機械学習モデルを再トレーニングして、第２のトレーニングされた機械学習モデル（例えば、学生モデル）を取得し得る。

【０１２０】

いくつかの実施形態では、システムは、複数のシークエンシング技術の各々に関して別個の機械学習モデルをトレーニングするように構成され得る。機械学習モデルは、シークエンシング技術から取得したデータを使用して、個々のシークエンシング技術に関してトレーニングされ得る。機械学習モデルは、シークエンシング技術のエラープロファイルに関して調整され得る。いくつかの実施形態では、システムは、複数のアセンブリ技術の各々に関して別個の機械学習モデルをトレーニングするように構成され得る。機械学習モデルは、アセンブリ技術から取得したアセンブリを使用して、個々のアセンブリ技術に関してトレーニングされ得る。機械学習モデルは、アセンブリ技術のエラープロファイルに関して調整され得る。

【０１２１】

いくつかの実施形態では、システムは、複数のシークエンシング技術に関して使用される一般化された機械学習モデルをトレーニングするように構成され得る。一般化された機械学習モデルは、複数のシークエンシング技術から集約されたデータを使用してトレーニングされ得る。いくつかの実施形態では、システムは、複数のアセンブリ技術に関して使用される一般化された機械学習モデルをトレーニングするように構成され得る。一般化された機械学習モデルは、複数のアセンブリ技術を使用して生成されたアセンブリを使用してトレーニングされ得る。

【０１２２】

図３Ｂは、本明細書に記載の技術のいくつかの実施形態による、アセンブリ（例えば、ゲノムアセンブリ、遺伝子配列、タンパク質配列、またはそれらの一部）を生成するためのプロセス３００から取得されたトレーニングされた機械学習モデルを使用するための例示的なプロセス３１０を示す。プロセス３１０は、任意の適切なコンピューティングデバイスによって実行され得る。一例として、プロセス３１０は、図１Ａ～図１Ｃを参照して上記のアセンブリシステム１０４によって実行され得る。

【０１２３】

プロセス３１０は、ブロック３１２で開始し、システムは、アセンブリを生成するため

に、シークエンシングデータに対するアセンブリアルゴリズム（例えば、ＯＬＣアセンブリまたはＤＢＧアセンブリ）を実行する。一例として、システムは、ＤＮＡサンプルのシークエンシングから生成されたヌクレオチド配列に対してアセンブリアルゴリズムを適用し得る。別の例として、システムは、タンパク質からのペプチドサンプルのシークエンシングから生成されたアミノ酸配列にアセンブリアルゴリズムを適用し得る。システムは、図２Ａ～図２Ｄのアセンブラ２００Ｃを参照して、上記のようなアセンブリアルゴリズムを適用し得る。いくつかの実施形態では、システムは、アセンブリアプリケーションを含み得る。システムは、アセンブリアプリケーションを実行することによってアセンブリを生成するように構成され得る。アセンブリアプリケーションの例は、本明細書に記載されている。

【０１２４】

ブロック３１２の周囲の破線によって示されるように、いくつかの実施形態では、システムは、アセンブリアルゴリズムを実行しなくてもよい。システムは、別個のシステム（例えば、別個のコンピューティングデバイス）によって生成されたアセンブリを取得し、ブロック３１４～３２２のステップを実行して、取得されたアセンブリを更新し得る。

【０１２５】

次に、プロセス３１０は、ブロック３１２に移行し、システムがシークエンシングデータおよびアセンブリにアクセスする。いくつかの実施形態では、システムは、（例えば、ブロック３１２において）システムによって生成されたアセンブリにアクセスするように構成され得る。いくつかの実施形態では、システムは、別個のシステムによって生成されたアセンブリにアクセスするように構成され得る。一例として、システムは、システムとは別のコンピューティングデバイス上で実行されるソフトウェアアプリケーションによって生成されたアセンブリを受信し得る。いくつかの実施形態では、システムは、プロセス３００でトレーニングされた機械学習モデルが更新するのに（例えば、エラーを修正するのに）最適化されたターゲットアセンブリ技術（例えば、アルゴリズムおよび／またはソフトウェアアプリケーション）から生成されたシークエンシングデータにアクセスするように構成され得る。例として、機械学習モデルは、カヌ（Ｃａｎｕ）アセンブリアプリケーションから生成されたアセンブリでトレーニングされ、システムは、カヌアセンブリアプリケーションによって生成されたアセンブリにアクセスし得る。

【０１２６】

いくつかの実施形態では、システムは、アクセスされたアセンブリを生成するために使用された生物学的ポリマー配列を含むシークエンシングデータにアクセスするように構成され得る。一例として、アクセスされるシークエンシングデータは、ゲノムアセンブリまたは遺伝子配列を生成するためにアセンブリアルゴリズムが適用されたヌクレオチド配列を含み得る。別の例として、アクセスされるシークエンシングデータは、タンパク質配列を生成するためにアセンブリアルゴリズムが適用されたアミノ酸配列を含み得る。いくつかの実施形態では、システムは、プロセス３００でトレーニングされた機械学習モデルが更新するのに最適化されたターゲットシークエンシング技術から生成されたシークエンシングデータにアクセスするように構成され得る。例として、機械学習モデルは、第三世代シークエンシングから生成されたシークエンシングデータでトレーニングされ得、システムは、第三世代シークエンシングから生成されたシークエンシングデータにアクセスし得る。

【０１２７】

次に、プロセス３１０は、ブロック３１６に移行し、システムは、シークエンシングデータおよびアセンブリを使用して、機械学習モデルに提供される入力を生成する。いくつかの実施形態では、システムは、アセンブリ内の個々の位置に関する入力を生成するように構成され得る。システムは、（１）シークエンシングデータからの配列をアセンブリ内の一組の位置に整列させ、（２）整列された配列の生物学的ポリマーを、アセンブリ内の位置に示される生物学的ポリマーと比較して、１つまたは複数の特徴の値を決定することによって、アセンブリ内の一組の位置に関する入力を生成するように構成し得る。いくつ

かの実施形態では、システムは、アセンブリ内の一組の位置における生物学的ポリマーを示すシークエンシングデータからの配列を同定することによって、アセンブリ内の一組の位置に配列を整列させるように構成され得る。一例として、アセンブリは、１から１０，０００のインデックスが付けられた位置を含み得、システムは、ヌクレオチド配列「ＴＡＧＧＴＣ」、「ＴＡＧＴＴＣ」、「ＴＡＧＧＣＣ」、「ＴＡＧＧＴＣ」が各々、アセンブリの５～１０にインデックスが付けられた位置に整列することを決定し得る。この例では、システムは、各ヌクレオチド配列を、アセンブリ内の５～１０にインデックスが付けられた位置において示された生物学的ポリマーと比較して、特徴（単数または複数）の値を決定し得る。特徴の例、および特徴の値の生成は、図４Ａ～図４Ｃを参照して説明されている。

【０１２８】

いくつかの実施形態では、システムは、アセンブリ内の個々の位置に関する入力を生成するように構成され得る。システムは、機械学習モデルへの入力として提供する位置に関する入力を生成して、アセンブリ内の位置に存在する生物学的ポリマー（例えば、ヌクレオチド、アミノ酸）を同定するために使用され得る出力を取得するように構成され得る。いくつかの実施形態では、システムは、その位置における生物学的ポリマーの表示、およびその位置の近傍にある１つまたは複数の他の位置における生物学的ポリマーの表示に基づいて、アセンブリ内のある位置に関する入力を生成するように構成され得る。入力は、モデルが対応する出力を生成するために使用するアセンブリ内の位置の周囲のコンテキスト情報を機械学習モデルに提供し得る。システムは、その位置およびその位置の近傍の他の位置（単数または複数）における特徴（単数または複数）の値を決定することによって、その位置の近傍の位置における生物学的ポリマーの表示に基づいて、ある位置に関する入力を生成するように構成され得る。一例として、システムは、（１）位置を選択し、（２）選択された位置を中心とする近傍の位置を特定し、（３）選択された位置および近傍の位置の各々における特徴（単数または複数）の値である入力を生成し得る。

【０１２９】

いくつかの実施形態では、システムは、設定された近傍のサイズを使用するように構成され得る。本明細書において近傍のサイズの例が説明される。いくつかの実施形態では、システムによって使用される近傍の位置の数は、設定可能なパラメータであり得る。例えば、システムは、使用する近傍のサイズを指定するユーザ入力（例えば、ソフトウェアアプリケーションにおける）を受信し得る。いくつかの実施形態では、システムは、近傍のサイズを決定するように構成され得る。一例として、システムは、シークエンシングデータが生成されたシークエンシング技術および／またはアセンブリが生成されたアセンブリ技術に基づいて近傍のサイズを決定し得る。

【０１３０】

いくつかの実施形態では、システムは、（１）アセンブリ内の位置を選択し、（２）選択された位置に関する個々の入力を生成することによって機械学習モデルに提供される入力を生成するように構成され得る。いくつかの実施形態では、システムは、アセンブリがアセンブリ内の位置において生物学的ポリマーを不正確に示す尤度を決定し、決定された尤度を使用して入力を生成する位置を選択することによって、アセンブリ内の位置を選択するように構成され得る。一例として、システムは、アセンブリが位置において生物学的ポリマーを不正確に示す尤度が閾値尤度を超えるかどうかを決定し、尤度が閾値尤度を超える場合、その位置に関する入力を生成し得る。いくつかの実施形態では、システムは、生物学的ポリマーがその位置に存在することを示す整列された配列の数に基づいて、位置が生物学的ポリマーを不正確に示す尤度を決定するように構成され得る。システムは、生物学的ポリマーがその位置にあることを示す配列の数と配列の総数との間の差異である尤度を決定し得る。一例として、アセンブリは、一組の９個のヌクレオチド配列のからのコンセンサスに基づいて、アセンブリ内のある位置においてチミンを示し得、このとき、４個のヌクレオチド配列は、チミンがその位置に存在することを示し、２個のヌクレオチド配列は、グアニンがその位置に存在することを示し、３個のヌクレオチド配列は、アデニ

ンがその位置に存在することを示す。この例では、システムは、アセンブリが、アセンブリ内の位置にある生物学的ポリマーを、チミンを示すヌクレオチド配列の数（４）とヌクレオチド配列の総数（９）との間に差異があると不正確に示す尤度を決定して、５の値を取得し得る。システムは、５が閾値の差異（例えば、１、２、３、４）より大きいと判定し、その結果、位置に関する入力を生成し得る。

【０１３１】

いくつかの実施形態では、システムは、１、２、３、４、５、６、７、８、９、または１０の閾値の差異を使用するように構成され得る。いくつかの実施形態は、特定の閾値の差異に限定されない。いくつかの実施形態では、閾値の差異は、設定可能なパラメータであり得る。システムによって使用される閾値尤度は、システムがモデルに提供される入力を生成する位置の数に影響を与え得る。一例として、システムは、ソフトウェアアプリケーションへのユーザ入力として閾値の値を受信し得る。いくつかの実施形態では、システムは、設定された閾値尤度を使用し得る。一例として、閾値尤度の値がエンコードされ得る。いくつかの実施形態では、システムは、閾値尤度を自動的に決定するように構成され得る。一例として、システムは、アセンブリが生成されたアセンブリ技術および／またはシークエンシングデータが生成されたシークエンシング技術に基づいて閾値尤度を決定し得る。

【０１３２】

いくつかの実施形態では、システムは、位置に関する入力を２次元行列として生成するように構成され得る。いくつかの実施形態では、マトリクスの各行／列は、アセンブリ内の個々の位置において決定された特徴（単数または複数）の値を指定し得る。いくつかの実施形態では、システムは、画像として入力を生成するように構成され得、画像のピクセルは、特徴（単数または複数）の値を保持する。一例として、画像の各行／列は、アセンブリ内の個々の位置において決定された特徴（単数または複数）の値を指定し得る。

【０１３３】

次に、プロセス３１０は、ブロック３１８に移行し、システムは、対応する出力を取得するためにブロック３１６で生成された入力を機械学習モデルに提供する。いくつかの実施形態では、システムは、機械学習モデルへの別個の入力として、アセンブリ内の個々の位置に対して生成された入力を提供するように構成され得る。一例として、システムは、ターゲット位置に対応する出力を取得するために、機械学習モデルへの入力として、ターゲット位置およびその位置の近傍の位置において決定された一組の特徴値を提供し得る。いくつかの実施形態では、システムは、（例えば、図２Ｃ～図２Ｄを参照して上で説明したように）複数の位置に対して並列に生成された入力を提供するように構成され得る。一例として、システムは、（１）第１の位置に対して生成された第１の入力をモデルに提供し、（２）第１の入力に対応する第１の出力を取得する前に、第２の位置に対して生成された第２の入力をモデルに提供し得る。いくつかの実施形態では、システムは、複数の位置に対して生成された入力を順次提供するように構成され得る。例えば、システムは、（１）対応する第１の出力を取得するために、第１の位置に対して生成された第１の入力をモデルに提供し、（２）第１の出力を取得した後、対応する第２の出力を取得するために、第２の位置に対する第２の入力を提供し得る。

【０１３４】

いくつかの実施形態では、機械学習モデルに提供される入力に対応する出力は、アセンブリ内の複数の位置の各々に関して、１つまたは複数の生物学的ポリマーの各々がその位置に存在する尤度を示し得る。一例として、出力は、ゲノムアセンブリ内の複数の位置の各々に関して、１つまたは複数のヌクレオチド（例えば、アデニン、グアニン、チミン、シトシン）の各々がその位置に存在する尤度（例えば、確率）を示し得る。別の例として、出力は、タンパク質配列内の複数の位置の各々に関して、１つまたは複数のアミノ酸の各々がその位置に存在する尤度を示し得る。いくつかの実施形態では、出力は、アセンブリ内のある位置に生物学的ポリマーが存在しない尤度を示し得る。一例として、システムは、「－」文字がアセンブリ内の位置における尤度を示し得る。

【０１３５】
いくつかの実施形態では、モデルは、アセンブリ内の個々の位置に対応する出力を提供し得る。システムは、アセンブリ内のターゲット位置に対して生成された入力を提供し、ターゲット位置に存在する１つまたは複数の生物学的ポリマーの各々の尤度を示す対応する出力を取得し得る。一例として、システムは、ゲノムアセンブリ内の位置に対して生成された入力を提供し、一組の４つの可能性のあるヌクレオチド（例えば、アデニン、グアニン、チミン、シトシン）の各々がその位置に存在する尤度を示す対応する出力を取得し得る。例えば、尤度は、その位置に存在する各ヌクレオチドの確率値であり得る。

【０１３６】
次に、プロセス３１０は、ブロック３２０に移行し、システムは、モデルから取得された出力を使用して、アセンブリ内の位置における生物学的ポリマーを同定する。いくつかの実施形態では、システムは、モデルに提供された対応する入力に応答してその位置に対して取得された出力を使用して、位置の各々に関して、その位置に存在する生物学的ポリマーを特定することによって、アセンブリ内の位置における生物学的ポリマーを特定するように構成され得る。モデルからの出力は、個々の位置に対応する複数組の出力値を含み得る。各組の出力値は、１つまたは複数の生物学的ポリマーの各々がアセンブリ内の個々の位置に存在する尤度を指定し得る。システムは、個々の位置においてその位置に存在する尤度が最も高い生物学的ポリマーである生物学的ポリマーを同定し得る。例として、アセンブリ内の第１の位置に関する一組の出力値は、アデニン（Ａ）０．１、シトシン（Ｃ）０．６、グアニン（Ｇ）０．１、チミン（Ｔ）０．１５、およびブランク（－）０．０５の組のその位置に関する尤度を示し得る。この例では、システムは、アセンブリ内の位置にあるシトシン（Ｃ）を同定し得る。いくつかの実施形態では、位置に関して生成された入力に対応するモデルからの出力は、その位置において生物学的ポリマーを指定する分類であり得る。一例として、モデルからの出力は、アデニン（Ａ）、シトシン（Ｃ）、グアニン（Ｇ）、チミン（Ｔ）、またはブランク（－）の分類であり得る。

【０１３７】
次に、プロセス３１０はブロック３２２に移行し、システムは、アセンブリを更新して、更新されたアセンブリを取得する。システムは、ブロック３２０において同定された生物学的ポリマーに基づいてアセンブリを更新するように構成され得る。いくつかの実施形態では、システムは、アセンブリ内の位置における生物学的ポリマーの表示を更新することによってアセンブリを更新するように構成され得る。いくつかの例では、ブロック３２０において位置に存在すると同定された生物学的ポリマーは、アセンブリ内の生物学的ポリマーの表示とは異なり得る。これらの例では、システムは、アセンブリ内の位置における生物学的ポリマーの表示を変更し得る。一例として、システムは、（１）モデルの出力を使用して、アデニン「Ａ」の表示を有するアセンブリ内の第１の位置にチミン「Ｔ」が存在することを同定し、（２）アデニン「Ａ」の以前の表示からチミン「Ｔ」を表示するようにアセンブリ内の第１の位置を変更し得る。いくつかの例では、ある位置に存在すると同定された生物学的ポリマーは、アセンブリ内のその位置における生物学的ポリマーの表示と同じであり得る。これらの例では、システムは、アセンブリ内のその位置における生物学的ポリマーの表示を変更しない。一例として、システムは、（１）モデルの出力を使用して、チミン「Ｔ」の表示を有するアセンブリ内の第１の位置においてチミン「Ｔ」が存在していることを同定し、（２）第１の位置の表示を変更せずに維持し得る。

【０１３８】
いくつかの実施形態では、システムは、アセンブリ内の複数の位置を並列に更新するように構成され得る。一例として、システムは、（１）アセンブリ内の第１の位置の更新を開始し、（２）第１の位置における更新を完了する前に、アセンブリの第２の位置の更新を開始し得る。いくつかの実施形態では、システムは、アセンブリ内の位置を順次更新するように構成され得る。一例として、システムは、（１）アセンブリの第１の位置を更新し、（２）アセンブリの第１の位置における更新を完了した後、アセンブリの第２の位置を更新する。

【０１３９】
いくつかの実施形態では、ブロック３２２においてアセンブリを更新して第１の更新されたアセンブリを取得した後、プロセス３１０は、ブロック３２２からブロック３１６への破線によって示されるように、ブロック３１６に戻ってもよい。いくつかの実施形態では、システムは、第１の更新されたアセンブリおよびシークエンシングデータを使用して機械学習モデルへの入力を生成するように構成され得る。一例として、システムは、シークエンシングデータの一組のヌクレオチド配列および第１の更新されたアセンブリを使用して、モデルへの入力を生成し得る。システムは、ヌクレオチド配列を第１の更新されたアセンブリの個々の位置に整列させて、上記のように機械学習モデルへの入力を生成し得る。次に、システムは、ブロック３１６から３２２における動作を実行して、第２の更新されたアセンブリを取得し得る。いくつかの実施形態では、アセンブリシステムは、条件が満たされるまで反復を実行するように構成され得る。

【０１４０】
いくつかの実施形態では、システムは、閾値の反復回数が実行されたとシステムが判定するまで、更新の反復を実行するように構成され得る。いくつかの実施形態では、反復の閾値回数は、ユーザ入力（例えば、ソフトウェアコマンド、またはハードコードされた値）によって設定され得る。いくつかの実施形態では、システムは、反復の閾値回数を決定するように構成され得る。一例として、システムは、初期アセンブリを取得するために使用されたアセンブリ技術のタイプに基づいて、更新の反復の閾値回数を決定し得る。いくつかの実施形態では、システムは、アセンブリが収束したことをシステムが検出するまで更新の反復を実行するように構成され得る。一例として、アセンブリシステムは、（１）最新の反復から取得された現在のアセンブリと前のアセンブリとの間の差異の数を決定し、（２）差異の数が差異の閾値数または差異のパーセンテージよりも少ない場合、更新の反復の実行を停止するように決定し得る。

【０１４１】
いくつかの実施形態では、システムは、アセンブリへの単一の更新を実行するように構成され得、プロセス３１０は、アセンブリへの単一の更新を実行した後、ブロック３２２において終了し得る。更新されたアセンブリは、システムによって出力アセンブリとして出力され得る。一例として、システムは、出力アセンブリがブロック３１４においてアクセスされる初期アセンブリよりも正確であるように、アセンブリ内のエラーが修正されたゲノムアセンブリを出力し得る。別の例として、システムは、出力タンパク質配列がブロック３１４においてアクセスされる初期タンパク質配列よりも正確であるように、エラーが修正されたタンパク質配列を出力し得る。

【０１４２】
いくつかの実施形態では、システムは、アセンブリの第１の部分に対して第１の数の更新の反復を実行し、アセンブリの第２の部分に対して第２の数の更新の反復を実行するように構成され得る。例として、システムは、（例えば、ブロック３１６～３２２で動作の複数の反復を実行することによって）ゲノムアセンブリの１～１００のインデックスが付けられた位置を複数回更新し、（例えば、ブロック３１６～３２２で動作を１回実行することによって）ゲノムアセンブリの１０１～２００のインデックスが付けられた位置を１回更新する。システムは、生物学的ポリマーを不正確に示し得る一部内の位置の数に基づいて、複数回更新するためのアセンブリの一部を決定するように構成され得る。一例として、システムは、（１）ウィンドウ位置（例えば、２５個、５０個、７５個、１００個、または１０００個の位置）内で閾値の尤度を超える不正確な生物学的ポリマーの表示の尤度を有する位置の数を決定し、（２）数が位置の閾値数を超えたときに、ウィンドウ位置に対して更新サイクルを実行することを決定し得る。

【０１４３】
図４Ａ～図４Ｃは、本明細書に記載の技術のいくつかの実施形態による、機械学習モデルに提供される入力を生成する例を示す。
図４Ａは、ヌクレオチド配列４０１（図４Ａにおいて「パイルアップ」とラベル付けさ

れている）、ヌクレオチド配列４０１から生成された生物学的ポリマーのアセンブリ４０２、およびアセンブリ内の個々の位置に関する生物学的ポリマーのラベル４０４を含むアレイ４００を示す。一例として、図４Ａに示されるデータは、機械学習モデルをトレーニングするためのプロセス３００を実行することから取得されたトレーニングデータであり得、（１）シークエンシングデータ４０１およびアセンブリ４０２は、ブロック３０２および３０４において取得され、（２）ラベル４０４は、ブロック３０６において取得される。別の例として、シークエンシングデータ４０１およびアセンブリ４０２は、トレーニングされた機械学習モデルを使用してアセンブリを生成するために、プロセス３１０のブロック３１２および／または３１４において取得され得る。

【０１４４】

図４Ａの実施形態に示されるように、シークエンシングデータ４０１は、ＤＮＡをシークエンシングすることから生成されたヌクレオチド配列を含む。シークエンシングデータ４０１の各行はヌクレオチド配列である。図４Ａの例に示すように、ヌクレオチド配列は英数字の配列として表され、「Ａ」はアデニンを表し、「Ｃ」はシトシンを表し、「Ｇ」はグアニンを表し、「Ｔ」はチミンを表し、「－」はその位置にヌクレオチドは存在しないことを表す。いくつかの実施形態は、個々のヌクレオチドまたはその欠如を表すための特定の組の英数字に限定されないことから、本明細書に記載の例示的な英数字は、例示の目的のためである。

【０１４５】

図４Ａの実施形態では、アセンブリ４０２は、ヌクレオチド配列４０１から生成される。いくつかの実施形態では、アセンブリ４０２は、シークエンシングデータ４０１にアセンブリアルゴリズム（例えば、ＯＬＣアセンブリ）を適用することにより取得され得る。図４Ａの実施形態では、アセンブリ４０２は、ヌクレオチド配列のコンセンサスを取ることにより取得される。コンセンサスは、アセンブリ４０２内の各位置に関するヌクレオチド配列の多数決によって決定され、システムは、その位置に最大数のヌクレオチド配列によって示される生物学的ポリマーを同定する。システムは、複数のヌクレオチドの各々に関して、（１）（例えば、ヌクレオチドがその位置に存在することを示すことによって）ヌクレオチドを選出するヌクレオチド配列の数を決定し、（２）その位置において示される選出数が最も多いヌクレオチドを同定するように構成され得る。例として、強調表示された列４０６の位置に関して、（１）４個の配列はアデニンを示し、３個の配列はシトシンを示し、２個の配列はグアニンを示し、（２）アセンブリ４０２内の位置はアデニンを示す。別の例として、アセンブリ４０２の第１の位置に関して、全てのヌクレオチド配列はシトシンを示し、従って、アセンブリ４０２は、第１の位置においてシトシンを示す。

【０１４６】

図４Ａの実施形態では、ラベル４０４は、アセンブリ４０２内の位置に対する所望の生物学的ポリマーを示し得る。いくつかの実施形態において、システムは、参照ゲノムからラベルを決定するように構成され得る。例えば、システムは、生物からのＤＮＡサンプルをシークエンシングすることによりヌクレオチド配列を取得し、ヌクレオチド配列へのアセンブリアルゴリズムの適用によりアセンブリ４０２を取得し、生物の既知の参照ゲノムから（例えば、ＮＣＢＩデータベースから）ラベル４０４を取得し得る。ラベル４０４は、教師ありトレーニングのために使用され、かつ／または生成されたアセンブリの精度を決定するために使用される各位置に関する真のまたは正確な生物学的ポリマーの表示を表し得る。

【０１４７】

図４Ｂは、図４Ａに示されるデータ４００から決定された値のアレイ４１０を示す。アレイ４１０は、アセンブリ４０２内の列４０６の位置に関する機械学習モデルへの入力の生成の際の中間ステップを示す。アレイ４１０は、図４Ａのヌクレオチド配列を表す「パイルアップ」とラベル付けされた一組の行を含む。アセンブリ内の各位置に関して、システムは、複数のヌクレオチドの各々のカウントを決定する。カウントは、ヌクレオチドがアセンブリ内の位置にあることを示すヌクレオチド配列の数を示す。アレイ４１０の「パ

イルアップ」セクションの各エントリは、ヌクレオチドに関するカウントを保持する。例として、図４Ｂにおける列４１２のカウントは、アデニンが４、シトシンが３、グアニンが２、チミンが０、ヌクレオチド無しが０である。別の例として、アレイ４１０の第１の列のカウントは、アデニンが０、シトシンが９、グアニンが０、チミンが０、ヌクレオチド無しが０である。

【０１４８】

アレイ４１０はさらに、図４Ｂのアセンブリ４０２を表す、図４Ｂにおいて「アセンブリ」とラベル付けされた一組の行を含む。アセンブリ４０２内の各位置に関して、アレイ４１０は、その位置に示されたヌクレオチドから決定された列の値を含む。各位置に関して、システムは、複数のヌクレオチドの各々に参照値を割り当て得、参照値は、ヌクレオチドがアセンブリ内の位置において示されているかどうかを示す。一例として、図４Ｂの４１２とラベル付けされた列において、アセンブリセクションは、（１）アデニンはアセンブリ４０２内の対応する位置に示されているヌクレオチドであるため、アデニンに対する９の値を有し、（２）他のヌクレオチドの各々はアセンブリ４０２内の対応する位置に示されていないため、他のヌクレオチドの各々に対する０の値を有する。別の例として、アレイ４１０の第１の列において、アセンブリセクションは、（１）シトシンはアセンブリ４０２内の対応する位置に示されているヌクレオチドであるため、シトシンに対する９の値を有し、（２）他のヌクレオチドの各々はアセンブリ４０２内の対応する位置に示されていないため、他のヌクレオチドの各々に対する０の値を有する。図４Ｂの例に示されるように、いくつかの実施形態では、ヌクレオチドがアセンブリ位置に示されるときにアセンブリ位置においてヌクレオチドに割り当てられる参照値は、整列されたヌクレオチド配列の数に等しい（例えば、図４Ａの例では９）。

【０１４９】

図４Ｃは、図４Ｂのアレイ４１０の値を使用して生成された特徴値のアレイ４２０を示す。いくつかの実施形態では、アレイ４２０は、対応する出力を得るために機械学習モデルへの入力として提供され得る。図４Ｃの例では、アレイ４２０は、列４２２に対応するアセンブリ内の位置に関してモデルに提供される入力である。アレイ４２０は、列４２２に対応するターゲット位置において決定された特徴の値、およびターゲット位置の近傍における２４個の位置に関して決定された特徴の値を含む。アレイ４２０は、ターゲット位置の左側にある１２個の位置、およびターゲット位置の右側にある１２個の位置に関する特徴の値を含む。

【０１５０】

アレイ４２０のパイルアップセクションにおいて、各列は、複数のヌクレオチドの各々に関するエラー値を指定する。列におけるヌクレオチドに関するエラー値は、（１）ヌクレオチドが列に対応するアセンブリ４０２内の位置にあることを示すヌクレオチド配列の数と、（２）アレイ４２０のアセンブリセクション内のヌクレオチドに割り当てられた参照値との間の差異を示す。例として、図４Ｃの列４２２に関して、値は、（１）アデニンが４－９＝－５であり、（２）シトシンが３－０＝３であり、（３）グアニンが２－０＝２であり、（４）チミンが０－０＝０であり、（５）ブランクが０－０＝０であるとして決定される。アレイ４２０のアセンブリセクションは、図４Ｂのアレイ４１０のアセンブリセクションと同じであり得る。

【０１５１】

いくつかの実施形態では、アレイ４２０内のパイルアップの値は、アセンブリ４０２がある位置においてヌクレオチドを不正確に同定する尤度を示し得る。システムは、値を使用して機械学習モデルへの入力を生成する位置を選択し得る。図４Ｃに示すように、パイルアップの非ゼロの値が強調表示されている。いくつかの実施形態では、システムは、ある位置におけるパイルアップ値が閾値を超えたときに、その位置に関して機械学習モデルに提供される入力を生成することを決定するように構成され得る。例えば、システムは、アデニンに関して決定された５の差異が４の閾値の差異を超えると決定することによって、列４２２に対応するアセンブリ４０２内の位置に関する入力を生成することを決定し得

る。閾値の差異の例は本明細書において説明されている。

【０１５２】

いくつかの実施形態では、アレイ４２０は、アセンブリ内の位置（例えば、列４２２に対応する位置）を更新するための機械学習モデルへの入力として提供され得る。システムは、機械学習モデルから取得した対応する出力を使用して、アセンブリ内の位置に存在するヌクレオチドを同定し、それに応じてアセンブリを更新し得る。いくつかの実施形態では、アレイ４２０は、モデルのトレーニングの一部として機械学習モデルに提供される複数の入力のうちの１つであり得る。システムは、機械学習モデルおよびラベル４０４から取得された対応する出力を使用して、機械学習モデルの１つまたは複数のパラメータへの調整を決定し得る。一例として、機械学習モデルはニューラルネットワークであり得、システムは、機械学習モデルの出力から同定されたヌクレオチドとラベルとの間の差異を使用して、ニューラルネットワークの重みに対する１つまたは複数の調整を決定し得る。

【０１５３】

図４Ａの例示的な実施形態は、核酸に関連するデータを示しているが、いくつかの実施形態では、データは、タンパク質に関連し得る。例えば、配列４０１はアミノ酸配列であり得、アセンブリ４０２はタンパク質配列であり得、ラベル４０４はタンパク質配列中の各位置に関する参照アミノ酸であり得る。システムは、アミノ酸配列、タンパク質配列、および／またはラベルに基づいて、図４Ｂ～図４Ｃに示される値を決定し得る。

【０１５４】

図５は、本明細書に記載の技術のいくつかの実施形態による、アセンブリを更新するプロセスを示す。図５は、更新されたアセンブリ５０８を生成するために機械学習モデル５０２に提供されるアセンブリデータ５００からの入力の生成を示す。アセンブリデータ５００は、例えば、図４Ｃを参照して上記で説明したデータの形式であり得る。図示された更新のプロセスは、図１Ａ～図１Ｃを参照して上記で説明されたアセンブリシステム１０４によって実行され得る。

【０１５５】

図５の実施形態に示されるように、システムは、更新されるべきアセンブリ内の位置５０４Ａおよび５０６Ａを選択する。一例として、システムは、（１）アセンブリがアセンブリ内の位置において生物学的ポリマー（例えば、ヌクレオチド、アミノ酸）を不正確に示す尤度を決定し、（２）位置５０４Ａ、５０６Ａにおける尤度が各々位置５０４Ａ、５０６Ａを選択するための閾値尤度を超えると決定することによって位置５０４Ａ、５０６Ａを選択し得る。システムが位置５０４Ａ、５０６Ａを選択すると、システムは、機械学習モデル５０２に提供される対応する入力を生成することを決定し得る。

【０１５６】

図５の実施形態に示されるように、システムは、位置５０４Ａに対応する第１の入力５０４Ｂと、位置５０６Ａに対応する第２の入力５０６Ｂとを生成する。システムは、図４Ａ～図４Ｃを参照して上記のように入力５０４Ｂ、５０６Ｂの各々を生成し得る。例えば、システムは、（１）その位置を中心とする位置の近傍を選択し、（２）近傍の位置の各々において１つまたは複数の特徴の値を決定し、（３）特徴（単数または複数）の値を位置に関する入力として使用することによって、入力５０４Ｂ、５０６Ｂの各々を生成し得る。いくつかの実施形態では、システムは、特徴（単数または複数）の値をデータ構造に格納するように構成され得る。一例として、システムは、図４Ｃに示されるように、値を２次元アレイまたは画像内に格納し得る。

【０１５７】

図５の実施形態に示されるように、システムは、対応する出力を得るために、生成された入力５０４Ｂ、５０６Ｂの各々を機械学習モデル５０２への入力として提供する。出力５０４Ｃは、位置５０４Ａに対して生成された入力５０４Ｂに対応し、出力５０６Ｃは、位置５０６Ａから生成された入力５０６Ｂに対応する。いくつかの実施形態では、システムは、入力５０４Ｂ、５０６Ｂを機械学習モデル５０２に順次提供するように構成され得る。一例として、システムは、（１）入力５０４Ｂを機械学習モデル５０２に提供して、

対応する出力５０４Ｃを取得し、（２）出力５０４Ｃを取得した後、入力５０６Ｂを機械学習モデル５０２に提供して、対応する出力５０６Ｃを取得する。いくつかの実施形態では、システムは、入力５０４Ｂ、５０６Ｂを機械学習モデル５０２に並列に提供するように構成され得る。一例として、システムは、（１）入力５０４Ｂを機械学習モデル５０２に提供し、（２）入力５０４Ｂに対応する出力５０４Ｃを取得する前に、入力５０６Ｂを機械学習モデル５０２に提供する。

【０１５８】

図５の実施形態に示されるように、出力５０４Ｃ、５０６Ｃの各々は、１つまたは複数のヌクレオチドの各々がアセンブリ内の位置に存在する尤度を示す。図５の実施形態では、尤度は確率である。例として、出力５０４Ｃは、（１）４個の異なるヌクレオチドの各々に関して、ヌクレオチドが位置５０４Ａに存在する確率と、（２）位置５０４Ａにおいてヌクレオチドが存在しない確率（「－」文字によって表される）とを指定する。出力５０４Ｃにおいて、アデニンは０．２の確率を有し、シトシンは０．５の確率を有し、グアニンは０．１の確率を有し、チミンは０．１の確率を有し、ヌクレオチドが位置５０４Ａにおいて存在しない確率は０．１である。別の例として、出力５０６Ｃは、（１）４個の異なるヌクレオチドの各々に関して、ヌクレオチドが位置５０６Ａに存在する確率と、（２）位置５０６Ａにおいてヌクレオチドが存在しない確率（「－」文字によって表される）とを指定する。この例では、アデニンは０．６の確率を有し、シトシンは０．１の確率を有し、グアニンは０．２の確率を有し、チミンは０．０５の確率を有し、ヌクレオチドが位置５０４Ａにおいて存在しない確率は０．０５である。

【０１５９】

図５の実施形態に示されるように、システムは、機械学習モデル５０２から取得された出力を使用して、アセンブリ内の位置を更新して、更新されたアセンブリ５０８を取得する。いくつかの実施形態では、システムは、（１）機械学習モデルから取得した出力を使用して、位置において存在するヌクレオチドを同定し、（２）同定されたヌクレオチドを示すようにアセンブリ内の位置を更新して、更新されたアセンブリ５０８を取得することによってアセンブリを更新するように構成され得る。図５の例に示すように、システムは、（１）出力５０４Ｃを使用して、シトシンがその位置に存在する尤度が最も高いと判定し、（２）その位置においてシトシン「Ｃ」を示すように、更新されたアセンブリ５０８内の対応する位置５０８Ａを設定することによって、初期アセンブリの位置５０４Ａを更新する。別の例として、システムは、（１）出力５０６Ｃを使用して、アデニンがその位置に存在する尤度が最も高いと判定し、（２）アデニン「Ａ」を示すように、更新されたアセンブリ５０８内の対応する位置５０８Ｂを設定することによって、初期アセンブリの位置５０６Ａを更新する。いくつかの例では、システムは、（１）機械学習モデル５０２から取得した出力を使用して、ある位置において同定されたヌクレオチドが、その位置において既に示されていることを決定し、（２）更新されたアセンブリ５０８において位置における表示を変更せずに維持し得る。

【０１６０】

更新されたアセンブリ５０８は、初期アセンブリとは別に示されているが、いくつかの実施形態では、更新されたアセンブリ５０８は、初期アセンブリの更新されたバージョンであり得る。例えば、システムは、初期アセンブリをメモリに格納し、メモリ内の初期アセンブリの値を更新して、更新されたアセンブリ５０８を取得し得る。いくつかの実施形態では、システムは、更新されたアセンブリ５０８を、初期アセンブリとは別個のアセンブリとして生成し得る。例えば、システムは、初期アセンブリを第１のメモリ位置に格納し、更新されたアセンブリ５０８を別個のアセンブリとして第２のメモリ位置に格納し得る。

【０１６１】

いくつかの実施形態では、システムは、初期アセンブリ内の複数の位置において更新を順次実行するように構成され得る。一例として、システムは、（１）出力５０４Ｃを使用して、更新されたアセンブリ５０８内の位置５０８Ａを更新し、（２）位置５０８Ａにお

ける更新を完了した後、出力５０６Ｃを使用して、更新されたアセンブリ５０８内の位置５０８Ｂを更新する。いくつかの実施形態では、システムは、初期アセンブリ内の複数の位置において並列に更新を実行するように構成され得る。一例として、システムは、（１）出力５０４Ｃを使用して位置５０８Ａの更新を開始し、（２）位置５０８Ａにおける更新を完了する前に、出力５０６Ｃを使用して位置５０８Ｂの更新を開始する。

【０１６２】

いくつかの実施形態では、システムは、アセンブリ内の個々の位置に関する入力を生成し、機械学習モデル５０２に入力を提供し、機械学習モデルからの出力を使用してアセンブリ内の複数の位置を並列に更新するプロセスを実行するように構成され得る。一例として、システムは、（１）初期アセンブリの位置５０４Ａに関する入力の生成を開始し、（２）位置５０４Ａにおける位置に対する更新を完了する前に、初期アセンブリの位置５０６Ａに関する入力の生成を開始し得る。アセンブリの更新を並列化することにより、システムは、（例えば、必要な時間が短縮されることによって）アセンブリを生成するプロセスをより効率的にする。システムは、複数のプロセッサを使用し、かつ／または複数のアプリケーションスレッドを使用することにより、プロセスを並列化し得る。

【０１６３】

図５の実施形態は、ゲノムアセンブリの一部を更新することを示しているが、いくつかの実施形態は、タンパク質配列またはその一部を更新するために、図示されたプロセスを実施し得る。例えば、初期アセンブリはタンパク質配列であり得る。次に、システムは、タンパク質配列内の位置に関する入力を生成して、機械学習モデル５０２に提供し得る。システムは、複数のアミノ酸の各々が位置において存在する尤度（例えば、確率）を示す出力を取得し得る。次に、システムは、初期タンパク質配列を更新して、更新されたタンパク質配列を取得し得る。

【０１６４】

図６は、本明細書に記載の技術のいくつかの実施形態による、アセンブリを生成するための例示的な畳み込みニューラルネットワークモデル６００を示す。いくつかの実施形態では、畳み込みニューラルネットワークモデル６００は、図３Ａを参照して上記のプロセス３００を実行することによってトレーニングされ得る。いくつかの実施形態では、プロセス３００から取得されたトレーニングされた畳み込みニューラルネットワークモデル６００を使用して、図３Ｂを参照して上記のようにアセンブリを生成するためにプロセス３１０を実行し得る。

【０１６５】

いくつかの実施形態では、モデル６００は、シークエンシングデータから生成された入力、およびシークエンシングデータから生成されたアセンブリを受信するように構成される。一例として、モデル６００は、図１Ａ～図１Ｃを参照して上記のアセンブリシステム１０４によって使用される機械学習モデルであり得る。シークエンシングデータは、生物学的ポリマー配列（例えば、ヌクレオチド配列またはアミノ酸配列）を含み得る。いくつかの実施形態では、システムは、１つまたは複数の特徴の値を決定し、決定された値をモデル６００への入力として提供するように構成され得る。一例として、システムは、アセンブリ内の位置の近傍における特徴の値を決定し、位置の近傍において決定された値をモデル６００への入力として提供し得る。入力の例および入力を生成するための技術が本明細書で説明されている。

【０１６６】

図６の例示的な実施形態では、モデル６００は、モデル６００に提供された入力を受信する第１の畳み込み層６０２を含む。第１の層６０２において、システムは、モデル６００に提供された入力を、３ｘ５ｘ６４の行列として表される６４個の３ｘ５フィルタにより畳み込む。例えば、システムは、出力を得るために、３ｘ５ｘ６４の行列の各チャネルにより（例えば、図４Ｃに示されるような）１０ｘ２５の入力マトリクスを畳み込み得る。第１の層６０２は、システムが畳み込みからの出力に適用する活性化関数としてＲｅＬｕ関数を含む。いくつかの実施形態では、第１の層６０２はまた、畳み込みの出力のサイ

ズを縮小するためのプーリング層を含み得る。

【０１６７】

図６の例示的な実施形態では、モデルは、第１の層６０２の出力を受信する第２の畳み込み層６０４を含む。第２の層６０４において、システムは、３ｘ５ｘ１２８の行列として表される一組の１２８個の３ｘ５フィルタにより入力を畳み込む。システムは、第１の畳み込み層６０２からの出力を３ｘ５ｘ１２８のフィルタセットにより畳み込み得る。第２の畳み込み層６０４は、システムが畳み込みからの出力に適用する活性化関数としてＲｅＬＵ関数を含む。いくつかの実施形態では、第２の層６０４はまた、畳み込みの出力のサイズを縮小するためのプーリング層を含み得る。次に、第２の畳み込み層６０４の出力は、第３の畳み込み層６０６に渡される。第３の層６０６において、システムは、３ｘ５ｘ２５６の行列として表される一組の２５６個の３ｘ５フィルタにより入力を畳み込む。次に、システムは畳み込みからの出力にＲｅＬｕ活性化関数を適用する。いくつかの実施形態では、第３の層６０６はまた、畳み込みの出力のサイズを縮小するためのプーリング層を含み得る。

【０１６８】

図６の例示的な実施形態では、モデル６００は、５つの完全に接続された層を有する高密度層６０８を含み、各々が２５６の入力値を受信する。システムは、第３の畳み込み層６０６から取得された出力を凝縮して（ｃｏｎｄｅｎｓｅ）、高密度層６０８への入力として提供し得る。高密度層６０８は、複数の値を出力することができ、各値は、入力がモデル６００に提供された位置において個々の生物学的ポリマー（例えば、ヌクレオチドまたはアミノ酸）が存在する尤度を示す。一例として、高密度層は５個の値を出力し得、各値は、ヌクレオチド（例えば、アデニン、シトシン、グアニン、チミン、および／またはヌクレオチド無し）がその位置に存在する尤度を示す。システムは、ソフトマックス（ｓｏｆｔｍａｘ）関数を高密度層６０８の出力に適用して、合計が１になる一組の確率値を取得し得る。図６の例示的な実施形態に示されるように、システムは、ソフトマックス関数を高密度層６０８の出力に適用して、個々のヌクレオチドがアセンブリ内のある位置に存在する確率を示す５個の確率の出力６１０を取得する。出力６１０は、（例えば、図５を参照して上で説明したように）アセンブリを更新するために使用し得る。

【０１６９】

図７は、本明細書に記載の技術のいくつかの実施形態による技術の性能結果を示している。各プロットは、従来の手法と比較して、技術によって提供される精度の向上を示す。図７では、カヌ（Ｃａｎｕ）およびミニアスム（Ｍｉｎｉａｓｍ）は２つの従来のアセンブリ技術である。ミニアスム（Ｍｉｎｉａｓｍ）＋レコン（Ｒａｃｏｎ）は、レコン・エラー訂正を適用したミニアスムを表す。カヌ（Ｃａｎｕ）＋クォーラムＱｕｏｒｕｍ）は、カヌから生成されたアセンブリを修正するために本明細書で説明する技術の実施である。ミニアスム＋クォーラムは、ミニアスムから生成されたアセンブリを修正するために本明細書で説明する技術の実施である。

【０１７０】

図７に示すように、ミニアスム＋クォーラムは、データの各サンプルに関して、ミニアスム＋レコンよりもエラー率が大幅に低くなっている。例として、３０Ｘ　パックバイオ（Ｐａｃｂｉｏ）データからの大腸菌の場合、ミニアスム＋クォーラム（連結点で表される）の各反復のエラー率は、１００エラー／１００キロベース（ｋｉｌｏ－ｂａｓｅｓ）満であるが、ミニアスム＋レコンの最小エラー率は約２００エラー／１００キロベースである。別の例として、３０Ｘ　ＯＮＴデータからの大腸菌の場合、ミニアスム＋クォーラムの各反復のエラー率は約４００エラー／１００キロベースであるが、ミニアスム＋レコンのエラー率は約５００エラー／１００キロベースである。

【０１７１】

図７に示すように、カヌ＋クォーラムは、カヌのみの結果よりも精度が向上している。カヌには従来のエラー訂正技術が組み込まれているが、本明細書で説明する技術により、アセンブリ生成の精度が向上する。例として、３０Ｘ　ＯＮＴデータからの大腸菌の場合

、カヌのエラー率は５００エラー／１００キロベースを超えるが、カヌ＋クォーラムの各反復のエラー率は３５０エラー／１００キロベース未満である。

【０１７２】

図７に示されるように、本明細書に記載される技術は、エラー訂正を実行するために実質的に大量の計算時間を追加することなく、アセンブリの向上された精度を提供し得る。例として、ミニアスム＋クォーラムは、実質的に同じＣＰＵ時間数で、ミニアスム＋レコンよりも優れた精度を実現する。別の例として、カヌ＋クォーラムは、アセンブリを修正するためのＣＰＵ時間数を大幅に増加させることなく、カヌ単独よりも高い精度を実現する。

【０１７３】

いくつかの実施形態では、本明細書で説明されるシステムおよび技術は、１つまたは複数のコンピューティングデバイスを使用して実施され得る。しかしながら、実施形態は、特定のタイプのコンピューティングデバイスによる動作に限定されない。さらなる例として、図８は、例示的なコンピューティングデバイス８００のブロック図である。コンピューティングデバイス８００は、１つまたは複数のプロセッサ８０２および１つまたは複数の有形の非一時的なコンピュータ可読記憶媒体（例えば、メモリ８０４）を含み得る。メモリ８０４は、有形の非一時的なコンピュータ記録可能媒体に、実行時に上記の機能のいずれかを実施するコンピュータプログラム命令を格納し得る。プロセッサ８０２は、メモリ８０４に接続され、そのようなコンピュータプログラム命令を実行して、機能を実現および実行させる。

【０１７４】

コンピューティングデバイス８００はまた、コンピューティングデバイスが他のコンピューティングデバイスと（例えば、ネットワークを介して）通信することができるネットワーク入力／出力（Ｉ／Ｏ）インタフェース８０６を含み、かつ、１つまたは複数のユーザＩ／Ｏインタフェース８０８も含み、コンピューティングデバイスは、１つまたは複数のユーザＩ／Ｏインタフェース８０８を介してユーザに出力を提供し、かつユーザから入力を受信する。ユーザＩ／Ｏインタフェースは、キーボード、マウス、マイクロフォン、ディスプレイデバイス（例えば、モニタまたはタッチスクリーン）、スピーカ、カメラ、および／または他の様々なタイプのＩ／Ｏデバイスなどのデバイスを含み得る。

【０１７５】

上述した実施形態は、多くの方法で実施することができる。例として、実施形態は、ハードウェア、ソフトウェア、又はそれらの組み合わせを用いて実施し得る。ソフトウェアで実施する場合、ソフトウェアコードは、単一のコンピューティングデバイスで提供されるか、複数のコンピューティングデバイスに分散されるかに関係なく、任意の適切なプロセッサ（例えば、マイクロプロセッサ）またはプロセッサの集合上で実行することができる。上述した機能を実行する任意の構成要素又は構成要素の集合は、上述の機能を制御する１つまたは複数のコントローラとして一般的に考えられることを理解されたい。１つまたは複数のコントローラは、専用ハードウェア、またはマイクロコードまたはソフトウェアを使用して上記の機能を実行するようにプログラムされた汎用ハードウェア（例えば、１つまたは複数のプロセッサ）など、様々な方法で実施することができる。

【０１７６】

この点に関して、本明細書で説明される実施形態の１つの実施は、１つまたは複数のプロセッサ上での実行時に、１つまたは複数の実施形態の上記の機能を実行するコンピュータプログラム（即ち、複数の実行可能な命令）がエンコードされた少なくとも１つのコンピュータ可読記憶媒体（例えば、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリまたは他のメモリ技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）、または他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージまたは他の磁気ストレージデバイス、または他の有形の非一時的なコンピュータ可読記憶媒体）を含むことを理解されたい。コンピュータ可読媒体は、本明細書で説明される技術の態様を実施するために、記憶されているプログラムが任意のコンピューティングデバイスにロードでき

るように移送可能である。加えて、実行時に、上述した機能の任意の１つを実行するコンピュータプログラムの参照は、ホストコンピュータ上で動作するアプリケーションプログラムに限定されないことを理解されたい。むしろ、コンピュータプログラムおよびソフトウェアという用語は、本明細書では一般的な意味で使用され、１つまたは複数のプロセッサをプログラムして本明細書で説明する技術の態様を実施するために使用することができる任意のタイプのコンピュータコード（例えば、アプリケーションソフトウェア、ファームウェア、マイクロコード、または他の形式のコンピュータ命令）を指す。

【０１７７】

本開示の様々な特徴および態様は、単独で、２以上の任意の組み合わせにおいて、または前述の実施形態において具体的に開示されていない様々な構成で使用することができ、従って、その用途において、上述の説明または図面に示されている構成要素の詳細および構成に限定されない。例として、一実施形態で説明された態様は、別の実施形態で説明された態様と任意の方法で組み合わせることができる。

【０１７８】

「ほぼ」、「実質的に」および「約」という用語は、いくつかの実施形態では目標値の±２０％以内、いくつかの実施形態では目標値の±１０％以内、いくつかの実施形態では目標値の±５％以内、およびいくつかの実施形態では目標値の±２％以内を意味するために使用され得る。「ほぼ」および「約」という用語は、目標値を含むことができる。

【０１７９】

また、本明細書で開示されるコンセプトは、方法として具現化されてもよく、その一例が提供されている。方法の一部として実行される処理は、任意の適切な方法で順序が付けられてもよい。従って、実施形態は、例示的な実施形態では逐次的な工程として示されているが、図示されている順序とは異なる順序で工程を実施すること、及びいくつかの工程を同時に実施することも可能である。

【０１８０】

請求項の要素を修飾するために、請求項に「第１」、「第２」、「第３」等の順序を示す用語が使用されているが、これは、請求項のある１つの要素の優先度や、先行性や、順序を示すか、又はある方法を実施する時間的な順序を示すものではなく、単なる標識として同じ名称を有する（但し、通常の用語を使用する）他の要素からある名前を有する別の請求項の要素を区別するために使用されている。

【０１８１】

また、本明細書で使用されている言い回しや用語は、説明を目的としたものであり、限定的なものと見なすべきではない。本明細書における「含む」、「備える」、「有する」、「含有する」、「含む」、およびそれらの変形の使用は、その後に列挙される項目およびその均等物ならびに追加の項目を包含することを意味する。

【図１Ａ】

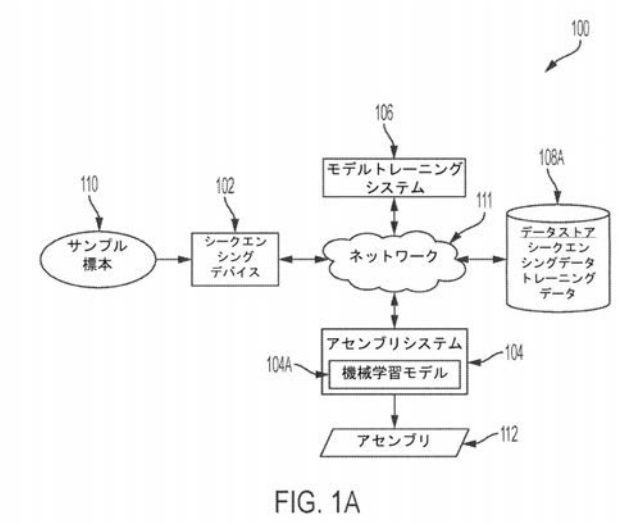

FIG. 1A

【図１Ｂ】

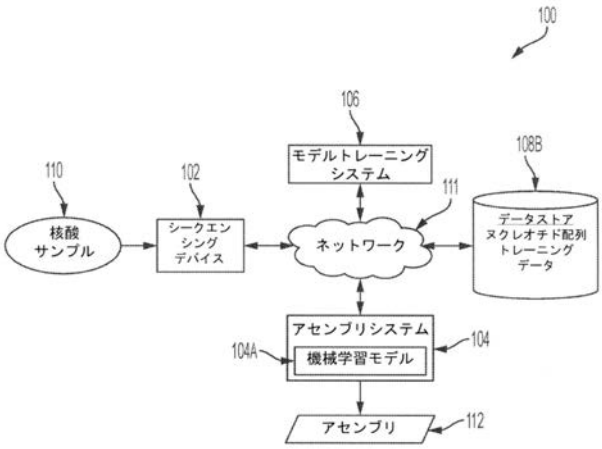

FIG. 1B

【図１Ｃ】

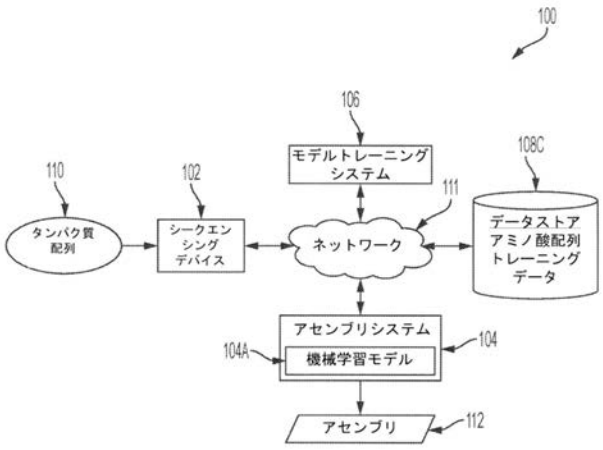

FIG. 1C

【図２Ａ】

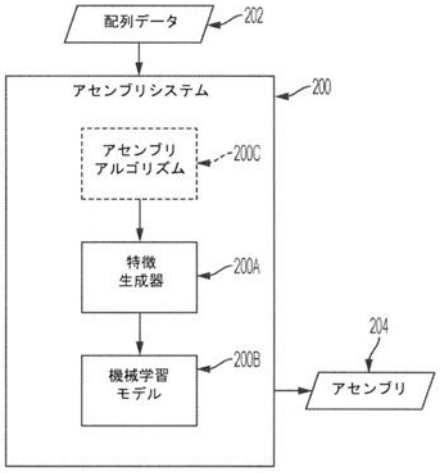

FIG. 2A

【図２Ｂ】

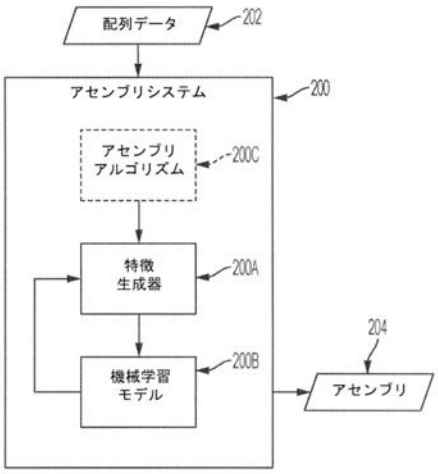

FIG. 2B

【図２Ｃ】

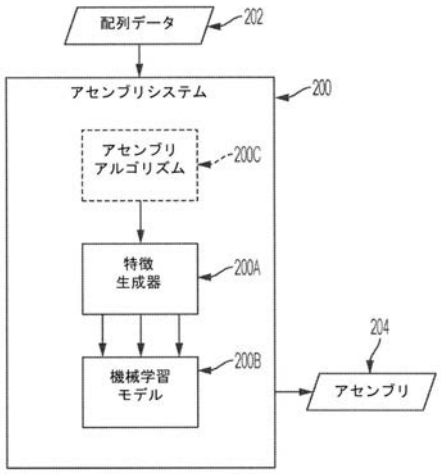

FIG. 2C

【図２Ｄ】

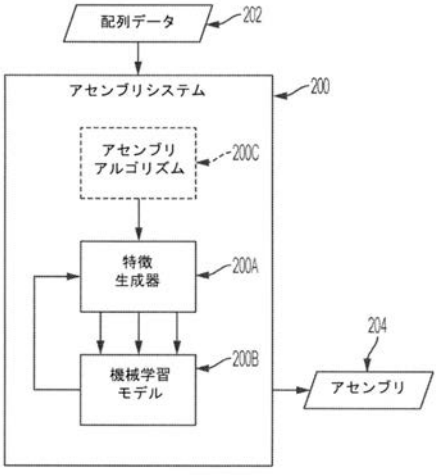

FIG. 2D

【図３Ａ】

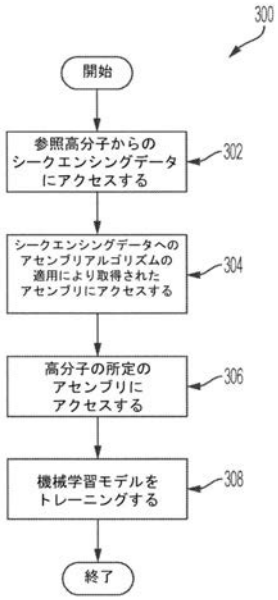

FIG. 3A

【図３Ｂ】

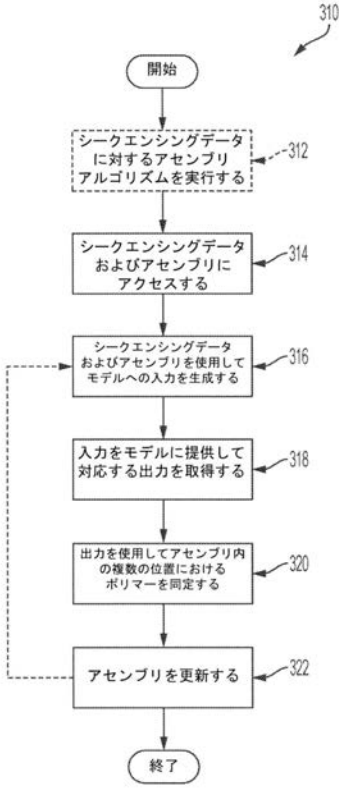

FIG. 3B

【図４Ａ】

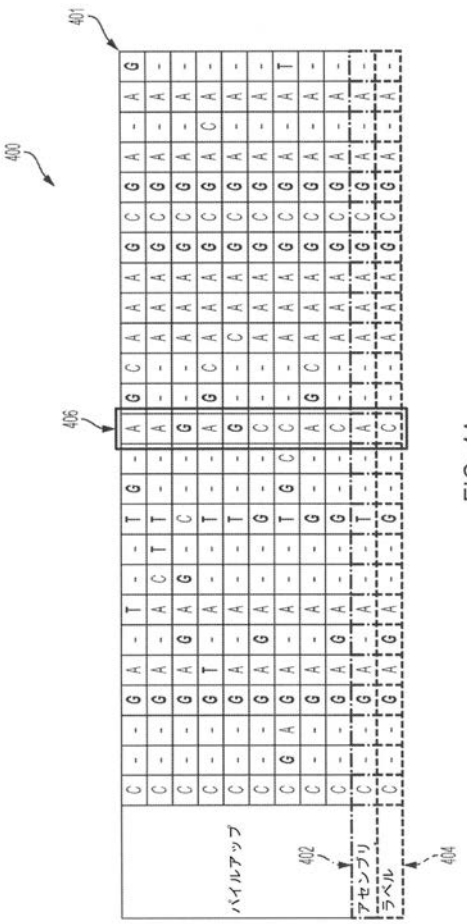

FIG. 4A

【図４Ｂ】

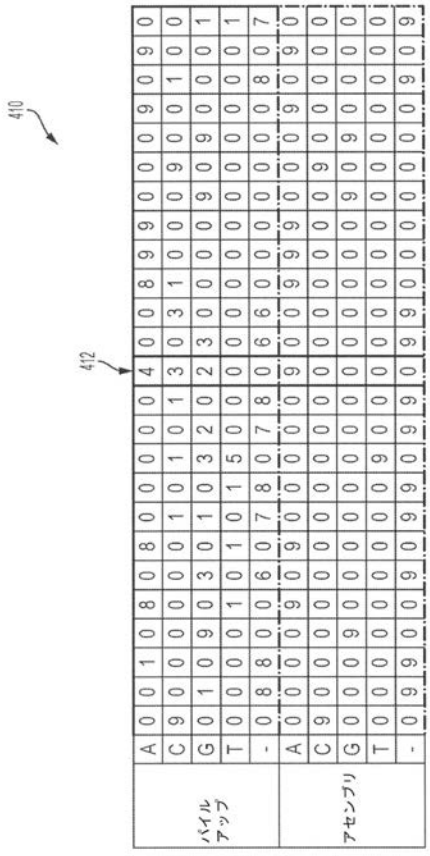

FIG. 4B

【図４Ｃ】

FIG. 4C

【図５】

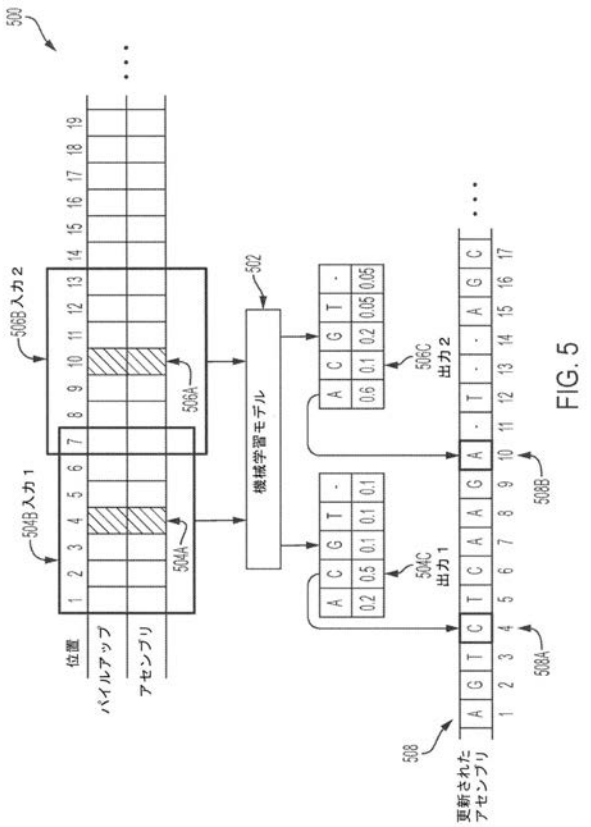

FIG. 5

【図６】

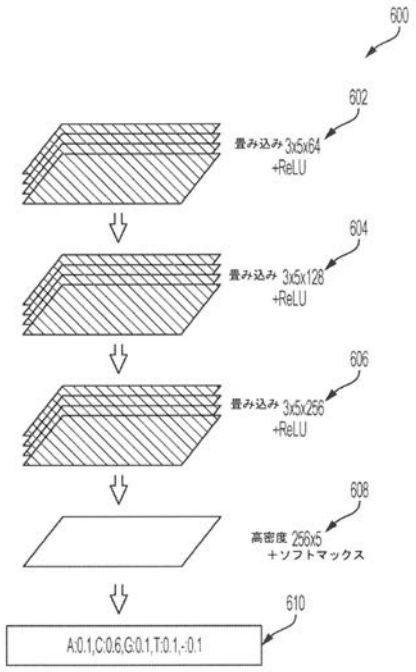

FIG. 6

【図７】

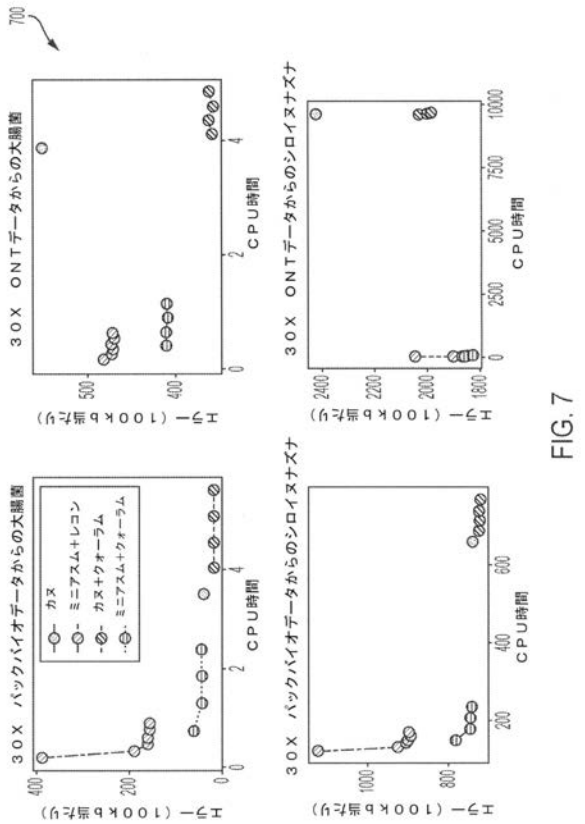

FIG. 7

【図８】

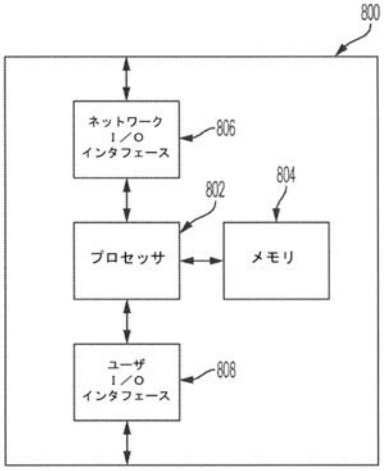

FIG. 8

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/032065

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B30/20
ADD. G16B40/20 G16B40/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NICHOLAS J LOMAN ET AL: "A complete bacterial genome assembled de novo using only nanopore sequencing data", NATURE METHODS, vol. 12, no. 8, 15 June 2015 (2015-06-15), pages 733-735, XP055609123, New York ISSN: 1548-7091, DOI: 10.1038/nmeth.3444 abstract p. 736, section "Online Methods", sub-section "Assembly pipeline and software" p. 736, section "Online Methods", sub-section "Computing the consensus sequence using signal data" -/--	1-66

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 July 2019	09/08/2019

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Heidrich, Alexander

2

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/032065

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	-& NICHOLAS J LOMAN ET AL: "A complete bacterial genome assembled de novo using only nanopore sequencing data - Supplementary Material", NATURE METHODS, vol. 12, no. 8, 15 June 2015 (2015-06-15), pages 733-735, XP055609143, New York ISSN: 1548-7091, DOI: 10.1038/nmeth.3444 p. 7, section "1.2 Computing the Consensus Sequence" p. 8, "Consensus Algorithm": p. 9 -----	1-66
X	ROBERT VASER ET AL: "Fast and accurate de novo genome assembly from long uncorrected reads", GENOME RESEARCH, vol. 27, no. 5, 18 January 2017 (2017-01-18), pages 737-746, XP055608901, US ISSN: 1088-9051, DOI: 10.1101/gr.214270.116 p. 737, section "Results" pp. 741 to 745, section "Methods" -----	1-66
X	Ryan Poplin ET AL: "Creating a universal SNP and small indel variant caller with deep neural networks", bioRxiv, 20 March 2018 (2018-03-20), XP055585250, DOI: 10.1101/092890 Retrieved from the Internet: URL:https://www.biorxiv.org/content/biorxiv/early/2018/03/20/092890.full.pdf [retrieved on 2019-05-03] abstract pages 4,6 -----	1-66
X	Ruibang Luo ET AL: "Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing", bioRxiv, 26 April 2018 (2018-04-26), XP055608907, DOI: 10.1101/310458 Retrieved from the Internet: URL:https://www.biorxiv.org/content/biorxiv/early/2018/04/28/310458.full-text.pdf [retrieved on 2019-07-25] abstract page 4 - page 8 ----- -/--	1-66

2

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/032065

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ALLEX C F ET AL: "Neural network input representations that produce accurate consensus sequences from DNA fragment assemblies", BIOINFORMATICS, OXFORD UNIVERSITY PRESS, SURREY, GB, vol. 15, no. 9, 1 September 1999 (1999-09-01), pages 723-728, XP002267211, ISSN: 1367-4803, DOI: 10.1093/BIOINFORMATICS/15.9.723 figures 1-8 -----	1-66

2

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

フロントページの続き

(81)指定国・地域 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT

(74)代理人 100152489
弁理士 中村 美樹
(72)発明者 ツァオ、ミン ドゥック
アメリカ合衆国 ０２４７８ マサチューセッツ州 ベルモント フレデリック ストリート ８

Ｆターム(参考) 4B029 AA07 BB15 FA15
4B063 QA11 QQ42 QQ52 QQ79 QS39