(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7583153号
(P7583153)

(45)発行日　令和6年11月13日(2024.11.13)

(24)登録日　令和6年11月5日(2024.11.5)

(51)国際特許分類			FI		
G16B	*40/20*	*(2019.01)*	G16B	40/20	
C12N	*15/63*	*(2006.01)*	C12N	15/63	Z ZNA
C12N	*15/12*	*(2006.01)*	C12N	15/12	
G16B	*30/00*	*(2019.01)*	G16B	30/00	

請求項の数　15　（全57頁）

(21)出願番号　特願2023-512747(P2023-512747)
(86)(22)出願日　令和3年8月20日(2021.8.20)
(65)公表番号　特表2023-538139(P2023-538139A)
(43)公表日　令和5年9月6日(2023.9.6)
(86)国際出願番号　PCT/US2021/046975
(87)国際公開番号　WO2022/040573
(87)国際公開日　令和4年2月24日(2022.2.24)
審査請求日　令和5年6月9日(2023.6.9)
(31)優先権主張番号　63/068,654
(32)優先日　令和2年8月21日(2020.8.21)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　597160510
リジェネロン・ファーマシューティカルズ・インコーポレイテッド
ＲＥＧＥＮＥＲＯＮ　ＰＨＡＲＭＡＣＥＵＴＩＣＡＬＳ，　ＩＮＣ．
アメリカ合衆国１０５９１－６７０７ニューヨーク州タリータウン、オールド・ソー・ミル・リバー・ロード７７７番
(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
(74)代理人　100152489

最終頁に続く

(54)【発明の名称】　配列生成および予測のための方法およびシステム

(57)【特許請求の範囲】

【請求項１】

遺伝的データをコンピュータにより受信することであって、前記遺伝的データは、第一の複数のヌクレオチド配列を含み、前記複数のヌクレオチド配列の各ヌクレオチド配列は、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、受信することと、

前記第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列をコンピュータにより決定することと、

閾値を満たす前記関連発現スコアに基づいて、前記第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列をコンピュータにより決定することと、

前記第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列をコンピュータにより生成することと、

コアプロモーターとして標識された前記第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された前記第四の複数のヌクレオチド配列に基づいて、訓練データセットをコンピュータにより生成することと、

前記訓練データセットに基づいて、生成モデルをコンピュータにより訓練することと、

前記生成モデルをコンピュータにより出力することと、を含む、コンピュータ実装方法。

【請求項２】

前記第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された前記第二の複数のヌクレオチド配列をコンピュータにより決定することが、

閾値を満たす前記関連発現スコアに基づいて、前記第一の複数のヌクレオチド配列から複数のＴＳＳをコンピュータにより決定することと、
前記複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基をコンピュータにより決定することと、
前記複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基をコンピュータにより決定することと、
各サミットヌクレオチド塩基および前記関連する複数の周辺塩基を、コアプロモーターとして標識された前記第二の複数のヌクレオチド配列としてコンピュータにより保存することと、を含む、請求項１に記載の方法。
【請求項３】
前記複数のＴＳＳに基づいて、前記複数のサミットヌクレオチド塩基をコンピュータにより決定することは、前記複数のＴＳＳのそれぞれについて、最強の遺伝子発現のキャップ解析（ＣＡＧＥ）シグナルを有するヌクレオチド塩基をコンピュータにより決定することを含む、請求項２に記載の方法。
【請求項４】
前記複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、前記関連する複数の周辺塩基をコンピュータにより決定することが、前記複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基をコンピュータにより決定することを含む、請求項２に記載の方法。
【請求項５】
５’方向の前記第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み、３’方向の前記第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む、請求項４に記載の方法。
【請求項６】
前記第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された前記第四の複数のヌクレオチド配列をコンピュータにより生成することが、
前記第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基をコンピュータにより決定することと、
コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として、各関連する複数のシフト塩基をコンピュータにより保存することと、を含む、請求項１～５のいずれか一項に記載の方法。
【請求項７】
前記第三の複数のヌクレオチド配列の各ヌクレオチド配列について、前記関連する複数のシフト塩基をコンピュータにより決定することが、前記第三の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をコンピュータによりシフトさせることを含み、前記ヌクレオチド塩基の前記量は、前記第三の複数のヌクレオチド配列と類似のクロマチンランドスケープを持たせるのに十分近い距離を保つことと、その一方で、近隣にある調節要素を拾わない程度に十分に離れていることとのバランスを表す、請求項６に記載の方法。
【請求項８】
前記訓練データセットに基づいて、前記生成モデルをコンピュータにより訓練することが、
前記訓練データセットの各ヌクレオチド配列について、複数のシード配列および標的ヌクレオチド対をコンピュータにより生成することであって、各シード配列および標的ヌクレオチド対は、定義された長さを有するシード配列、および所定のヌクレオチド配列上の前記シード配列の直後に標的ヌクレオチドを含む、生成することと、
前記複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をコンピュータによりベクター化することと、
前記ベクター化されたシード配列および標的ヌクレオチド対に基づいて、前記生成モデ

ルをコンピュータにより訓練することと、を含む、請求項１～７のいずれか一項に記載の方法。
【請求項９】
前記訓練データセットの各ヌクレオチド配列について、前記複数のシード配列および標的ヌクレオチド対をコンピュータにより生成することが、
前記関連発現スコアに基づいて、前記複数のＴＳＳをコンピュータによりクラスタリングすることと、
ＴＳＳの各クラスターについて、四分位幅をコンピュータにより決定することと、
前記四分位幅に基づいて、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳとしてコンピュータにより標識することと、
シャープＴＳＳまたはブロードＴＳＳ標識に基づいて、前記訓練データセット中の前記ヌクレオチド配列をシャープＴＳＳ群またはブロードＴＳＳ群にコンピュータにより分割することと、
前記定義された長さで、定義された工程サイズを有するスライディングウィンドウを、各ヌクレオチド配列にコンピュータにより適用することと、
前記スライディングウィンドウの各工程で、シード配列および標的ヌクレオチド対をコンピュータにより保存することと、を含む、請求項８に記載の方法。
【請求項１０】
前記複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をコンピュータによりベクター化することは、各ヌクレオチドをそれぞれの番号としてコンピュータによりコード化することを含む、請求項８に記載の方法。
【請求項１１】
前記生成モデルが、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を含む、請求項１～１０のいずれか一項に記載の方法。
【請求項１２】
前記生成モデルに基づいて、ヌクレオチド配列をコンピュータにより生成することをさらに含み、前記生成モデルに基づいて、前記ヌクレオチド配列をコンピュータにより生成することが、
ａ）シード配列をコンピュータにより受信することと、
ｂ）前記シード配列に基づいて、次のヌクレオチドをコンピュータにより予測することと、
ｃ）前記シード配列に前記次のヌクレオチドをコンピュータにより付加することと、
ｄ）前記ヌクレオチド配列の所望の長さに到達するまでｂ～ｃをコンピュータにより繰り返すことであって、前記ヌクレオチド配列がコアプロモーター配列である、繰り返すことと、を含む、請求項１～１１のいずれか一項に記載の方法。
【請求項１３】
前記所望の長さは、５０ヌクレオチド～１００ヌクレオチドである、請求項１２に記載の方法。
【請求項１４】
前記コアプロモーター配列に基づいてプロモーターを操作すること、および前記プロモーターを核酸構築物に挿入することをさらに含む、請求項１２に記載の方法。
【請求項１５】
予測モデルに、前記ヌクレオチド配列を提供することと、
前記予測モデルに基づいて、前記ヌクレオチド配列がコアプロモーターであることをコンピュータにより決定することと、を含む、請求項１２に記載の方法。
【発明の詳細な説明】
【背景技術】
【０００１】
関連出願の相互参照
本出願は、２０２０年８月２１日に出願された米国仮特許出願第６３／０６８，６５４

号の出願日における優先権の利益を主張するものである。これ以前の出願の内容は、参照によりその全体が本明細書に組み込まれる。

【０００２】

配列表への参照

２０２１年８月２０日に提出された配列表は、２０２１年８月２０日に「３７５９５＿００３３Ｐ１＿Ｓｅｑｕｅｎｃｅ＿Ｌｉｓｔｉｎｇ．ｔｘｔ」というファイル名のテキストファイル（サイズ２，８０５バイト）として作成されたものであり、連邦規則法典第３７巻特許法第１．５２条（ｅ）（５）に従い、本明細書において参照により援用されている。

【０００３】

アデノ随伴ウイルス（ＡＡＶ）は、遺伝子治療において導入遺伝子の送達のためのゴールドスタンダードである。免疫原性が低く感染性が強いなど、多くの利点を提供する一方で、一つの制限は、その厳密なＤＮＡパッケージング能力である。多くの治療法がすでにこの限界に近づいている。組換えＡＡＶベクターにコードされた他の特徴とともに、これは制御配列のためのスペースをほとんど残さない。一般的に使用されるウイルス性および内因性哺乳類プロモーターは両方ともこれらの制限を超えており、ＡＡＶを介した大きな導入遺伝子の送達には使用できない。したがって、短く効率的な制御配列に対する強いニーズがある。

【発明の概要】

【０００４】

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、第二の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練することと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、を含む方法が開示される。

【０００５】

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、第二の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練することと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、を含む方法も開示される。

【０００６】
遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、遺伝的データを正規化することと、関連発現スコアに基づいて、ＴＳＳをクラスター化することと、ＴＳＳの各クラスターについて、四分位幅を決定することと、四分位幅に基づいて、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳに標識することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットにおける各ヌクレオチド配列に対して、複数のシード配列および標的ヌクレオチド対を生成することと、複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクトル化することと、ベクトル化されたシード配列および標的ヌクレオチド対に基づいて、生成モデルを訓練することと、および生成モデルを出力することと、を含む方法も開示される。

【０００７】
遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、生成モデルを訓練することと、および生成モデルを出力することと、を含む方法も開示される。

【０００８】
また、ヌクレオチド配列を受信することと、訓練された予測モデルにヌクレオチド配列を提供することと、および予測モデルに基づいて、ヌクレオチド配列がコアプロモーターであることを決定することと、を含む方法が開示される。

【０００９】
また、（ａ）ヌクレオチド配列と配列長を受信すること、（ｂ）訓練された生成モデルに、ヌクレオチド配列を提供すること、（ｃ）生成モデルに基づいて、ヌクレオチド配列に関連付けられた次のヌクレオチドを決定すること、（ｄ）ヌクレオチド配列に次のヌクレオチドを付与すること、（ｅ）ヌクレオチド配列の長さが配列長に等しくなるまでｂ～ｄを繰り返すこと、および（ｆ）ヌクレオチド配列をコアプロモーター配列として出力すること、を含む方法も開示される。

【００１０】
遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、閾値を

満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、生成モデルを訓練することと、を含む方法も開示される。

【００１１】

開示される方法のいずれかを実施するよう構成された装置が開示される。

装置が開示される方法のいずれかを実施するよう構成された、プロセッサが実行可能な指示実施形態を有する、コンピュータ可読媒体が開示される。

【００１２】

開示される方法および組成物のさらなる利点は、一部が、以下の記載において記載されるか、一部が、記載から理解されるか、または開示される方法および組成物の実施によって学んでもよい。開示される方法および組成物の利点は、添付の特許請求の範囲において特に指摘されている要素および組み合わせによって実現され、達成されるであろう。前述の一般的な説明および以下の詳細な説明は両方とも、請求される本発明の、あくまで例示的かつ説明的なものであって、限定的なものではないことを理解されたい。

【図面の簡単な説明】

【００１３】

本明細書において援用され、かつ本明細書の一部を成す添付の図面は、開示される方法および組成物の一部の実施形態を例証し、説明と共に、開示される方法および組成物の原理を説明する役割を果たすものである。

【図１】図１は、例示的な操作環境を示す。

【図２】図２は、例示的な方法を示す。

【図３】図３は、例示的な方法を示す。

【図４Ａ】図４Ａは、例示的なＲＮＮブロックのコンパクトな表記を示す。

【図４Ｂ】図４Ｂは、ＲＮＮブロックの拡張表記を示す。

【図５】図５は、例示的なＬＳＴＭ－ＲＮＮブロックを示す。

【図６】図６は、例示的な方法を示す。

【図７】図７は、例示的な方法を示す。

【図８】図８は、方法の一例を示す。８０２は、配列番号１を示し、８０６は、配列番号２を示し、８１０は、配列番号３を示し、８１４は、配列番号４を示し、８１８は、配列番号５を示す。

【図９】図９は、例示的な方法を示す。

【図１０】図１０は、予測モデルの例示的な特徴を示す。

【図１１】図１１は、例示的な方法を示す。

【図１２】図１２は、例示的な方法を示す。

【図１３】図１３は、プロモーターアッセイの一例を示す。

【図１４】図１４は、コアプロモーターを制御する生成コアプロモーターの性能の比較を示す。

【図１５】図１５は、例示的な操作環境を示す。

【図１６－１】図１６は、例示的な方法を示す。

【図１６－２】同上。

【図１７】図１７は、例示的な方法を示す。

【図１８－１】図１８は、例示的な方法を示す。

【図１８－２】同上。

【図１９】図１９は、例示的な方法を示す。

【図２０】図２０は、例示的な方法を示す。

【図２１】図２１は、例示的な方法を示す。

【発明を実施するための形態】
【００１４】
下記の特定の実施形態およびそれに含まれる実施例についての発明を実施するための形態、ならびに図面およびその前後の説明を参照することによって、開示される方法および組成物についての理解を容易にすることができる。
【００１５】
Ａ．用語の定義
当然のことながら、本開示の方法および組成物は、記載されている特定の方法論、プロトコルおよび試薬に限定されるものではない。理由はこれらが、変更される可能性があるからである。本明細書中に使用されている用語は、あくまで特定の実施形態を説明することを目的としたものであって、もっぱら添付の特許請求の範囲により限定される本発明の範囲を限定するものではないことも、理解すべきである。
【００１６】
本明細書および添付の特許請求の範囲で使用される場合、単数形「ａ」、「ａｎ」、および「ｔｈｅ」は、文脈が明白に指示しない限り、複数の参照を含むことに留意されたい。したがって、例えば、「ある配列」への言及は、複数の配列を含み、「その配列」への言及は、一つまたは複数の配列および当業者に公知のその均等物などへの言及である。
【００１７】
本明細書で使用される場合、用語「配列決定」または「シーケンサー」は、生体分子、例えば、ＤＮＡまたはＲＮＡなどの核酸の配列を決定するために使用される多数の技術のいずれかを指す。例示的な配列決定方法としては、標的配列決定、単一分子のリアルタイム配列決定、エクソン配列決定、電子顕微鏡ベースの配列決定、パネル配列決定、トランジスタ介在性配列決定、直接配列決定、ランダムショットガン配列決定、サンガージデオキシ末端配列決定、全ゲノム配列決定、ハイブリダイゼーションによる配列決定、パイロシークエンシング、二本鎖配列決定、サイクルシーケンシング、単一塩基伸長配列決定、固相配列決定、ハイスループット配列決定、超平行シグネチャシーケンシング、エマルションＰＣＲ、より低い変性温度ＰＣＲ（ＣＯＬＤ－ＰＣＲ）での共増幅、マルチプレックスＰＣＲ、可逆的染料ターミネーターによる配列決定、対末端配列決定、短期配列決定、エキソヌクレアーゼ配列決定、ライゲーションによる配列決定、ショートリードシーケンシング、一分子配列決定、合成による配列決定、リアルタイムシーケンシング、逆ターミネーター配列決定、ナノポア配列決定、４５４配列決定、Ｓｏｌｅｘａ　Ｇｅｎｏｍｅ　Ａｎａｌｙｚｅｒ配列決定、ＳＯＬｉＤ（商標）配列決定、ＭＳ－ＰＥＴ配列決定、およびその組み合わせが挙げられるが、これらに限定されない。一部の実施形態では、配列決定は、例えば、ＩｌｌｕｍｉｎａまたはＡｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓから市販されている遺伝子アナライザーなどの遺伝子アナライザーによって行うことができる。
【００１８】
「ポリヌクレオチド」、「核酸」、「核酸分子」、または「オリゴヌクレオチド」は、ヌクレオシド間結合によって結合されたヌクレオシド（デオキシリボヌクレオシド、リボヌクレオシド、もしくはそのアナログを含む）の直鎖ポリマーを指す。典型的には、ポリヌクレオチドは、少なくとも三つのヌクレオシドを含む。オリゴヌクレオチドは、通常、数個の単量体単位、例えば、３～４個から数百個の単量体単位までのサイズ範囲に及ぶ。ポリヌクレオチドが、「ＡＴＧＣＣＴＧ」などの文字の配列で表される場合、ヌクレオチドは、左から右に５’→３’の順であり、別段示されない限り、「Ａ」は、アデノシンを示し、「Ｃ」は、シトシンを示し、「Ｇ」は、グアノシンを示し、「Ｔ」は、チミジンを示すことは、理解されるだろう。文字Ａ、Ｃ、Ｇ、およびＴは、当該技術分野で標準的なように、塩基自体、ヌクレオシド、または塩基を含むヌクレオチドを指すように使用され得る。
【００１９】
用語「ＤＮＡ（デオキシリボ核酸）」は、それぞれが、四つの核酸塩基、すなわち、アデニン（Ａ）、チミン（Ｔ）、シトシン（Ｃ）、およびグアニン（Ｇ）のうちの一つを含

む、デオキシリボヌクレオシドを含むヌクレオチドの鎖を指す。用語「ＲＮＡ（リボ核酸）」は、それぞれが、四つの核酸塩基、すなわち、Ａ、ウラシル（Ｕ）、Ｇ、およびＣのうちの一つを含む、四つのタイプのリボヌクレオシドを含むヌクレオチドの鎖を指す。ヌクレオチドの特定の対は、相補的な様式で互いに特異的に結合する（相補的塩基対と呼ばれる）。ＤＮＡでは、アデニン（Ａ）は、チミン（Ｔ）と対形成し、シトシン（Ｃ）は、グアニン（Ｇ）と対形成する。ＲＮＡでは、アデニン（Ａ）は、ウラシル（Ｕ）と対形成し、シトシン（Ｃ）は、グアニン（Ｇ）と対形成する。第一の核酸鎖が、第一の鎖のヌクレオチドに相補的であるヌクレオチドからなる第二の核酸鎖に結合するとき、この二つの鎖は、結合して、二本鎖を形成する。本明細書で使用される場合、「核酸配列決定データ」、「核酸配列決定情報」、「核酸配列」、「ヌクレオチド配列」、「ゲノム配列」、「遺伝子配列」または「フラグメント配列」もしくは「核酸配列決定読み取り」は、ＤＮＡまたはＲＮＡなどの核酸の分子（例えば、全ゲノム、全トランスクリプトーム、エキソーム、オリゴヌクレオチド、ポリヌクレオチド、またはフラグメント）におけるヌクレオチド塩基の順序（例えば、アデニン、グアニン、シトシン、およびチミンまたはウラシル）示す任意の情報またはデータを示す。本教示は、キャピラリー電気泳動、マイクロアレイ、ライゲーションベースのシステム、ポリメラーゼベースのシステム、ハイブリダイゼーションベースのシステム、直接的または間接的ヌクレオチド識別システム、パイロシーケンシング、イオンベースもしくはｐＨベースの検出システム、および電子署名ベースのシステムを含むが、これらに限定されない、すべての利用可能な様々な技術、プラットフォームまたはテクノロジーを使用して得られる配列情報を企図すると、理解されるべきである。

【００２０】

「ベクター」は、プラスミド、ファージ、ウイルス構築物、またはコスミドなどのレプリコンであり、これに別のＤＮＡセグメントが付着する場合がある。ベクターを使用して、ＤＮＡセグメントを細胞内に形質導入し、発現させる。

【００２１】

「プロモーター」または「プロモーター配列」は、細胞内でＲＮＡポリメラーゼを結合し、メッセンジャーＲＮＡ、リボソームＲＮＡ、核小体ＲＮＡの核内低分子、または任意のＲＮＡポリメラーゼＩ、ＩＩ、またはＩＩＩの任意のクラスによって転写された任意の種類のＲＮＡなどのポリヌクレオチドまたはポリペプチドコード配列の転写を開始することができるＤＮＡ制御領域である。

【００２２】

「任意選択的な」または「任意選択的に」は、後述されている事象、状況または材料が起こる場合もあれば起こらない場合もあるか、存在する場合もあれば存在しない場合もあることを意味すると共に、この記載には、前述の事象、状況または材料が起こる場合の例および起こらない場合の例、または存在する場合の例および存在しない場合が包含されることを意味する。

【００２３】

この明細書の記載および特許請求の範囲を通じて、語「含む（ｃｏｍｐｒｉｓｅ）」およびこの語の変形、例えば「含む（ｃｏｍｐｒｉｓｉｎｇ）」および「含む（ｃｏｍｐｒｉｓｅｓ）」などは、「～を含むがこれに限定されない」を意味し、例えば、他の追加のもの、コンポーネント、整数、または工程を除外することを意図するものではない。特に、一つまたは複数の工程または動作を含むものとして記載される方法では、それぞれの工程が、列挙されているものを含むこと（その工程が、「からなる」などの限定する用語を含まない限り）が具体的に企図されており、それは、それぞれの工程が、例えば、工程に挙げられていない他の追加のもの、コンポーネントまたは工程を排除することが意図されていないことを意味している。

【００２４】

「例示的な」は、「の一例」を意味し、好ましい構成または理想的な構成の表示を伝達することを意図するものではない。「など」は、限定的な意味で使用されるものではなく

、説明を目的に使用される。

【００２５】

本明細書では、範囲は、「約」一つの特定の値から、かつ／または「約」別の特定の値までとして表現される場合がある。こうした範囲が表されるとき、具体的に企図され、開示されることが考慮される範囲は、文脈が別途具体的に示さない限り、一つの特定の値からおよび／または他の特定の値の範囲である。同様に、値が近似値として表現されている場合には、先行する「約」を使用することにより、特定の値が別の実施形態を形成することが理解されるであろうし、具体的には、文脈が別途具体的に示さない限り、開示されることが考慮されるべき実施形態が企図される。これらの範囲の各々の終点は、文脈が別途具体的に示さない限り、他の終点と関連して、かつ他の終点とは独立して有意であることがさらに理解されるであろう。最後に、明示的に開示された範囲内に含まれる個々の値および値のサブレンジの全ても、具体的に企図されており、文脈が別段示さない限り、開示されているとみなされるべきであることが理解されるべきである。前述は、特定の事例において、これらの実施形態の一部またはすべてが明示的に開示されているか否かにかかわらず、適用される。

【００２６】

Ｂ．制御配列を設計するためのアプローチ

図１は、ＡＡＶベクター内にパッケージングされたＡＡＶベクターおよびＤＮＡの概略図を示す。ＡＡＶ遺伝子治療ベクターのパッケージ化されたＤＮＡには、ＡＡＶゲノムの３’末端に一つおよび５’末端に一つの、二つの逆位末端反復（ＩＴＲ）配列が必要であることが判明した。ＡＡＶの二つのＩＴＲは合計で約０．２～０．３ｋｂであるため、これら二つのＩＴＲの間に導入できる外来性ＤＮＡ（対象遺伝子を含む）は４．４ｋｂよりも小さくなければならない。一例として、ＩＴＲは、２ｘ１４５ｂｐの長さであってもよい（左ＩＴＲと右ＩＴＲは同一である）。２つのＩＴＲ間の外来性ＤＮＡの長さが許容される最大値（４～４．４ｋｂ）に近い場合、パッケージング効率は大幅に低下する。外来性ＤＮＡは、図１に示すように、プロモーター（エンハンサー要素を含むまたは含まない）、目的の導入遺伝子／遺伝子、ポリＡを含むことができるが、これらに限定されない。含まれる外来性ＤＮＡ要素が多いほど、目的の遺伝子は小さくなり得る。したがって、目的の遺伝子のサイズを大きくすることができる一つの選択肢は、プロモーターなどの制御配列を設計するための機械学習アプローチにおいて本明細書に記載されるプロモーターサイズを小さくすることである。

【００２７】

本明細書では、制御配列を設計するための機械学習アプローチを記載する。記載された方法は、前処理のためのデータ方法（訓練データの生成）、および予測モデルおよび生成モデルを生成するための方法に分離され得る。したがって、図２に示すように、２１０でプロモーター配列データセットを決定することを含む、方法２００が本明細書に記載される。プロモーター配列データセットの一部、全部、またはバリアントを使用して、２２０で生成モデルのための訓練データセットを生成してもよい。生成モデルは、プロモーター配列データセットに従って訓練されることに基づいて、プロモーター配列を生成するように構成されてもよい。プロモーター配列データセットの一部、全部、またはバリアントを使用して、２２０で予測モデルのための訓練データセットを生成してもよい。生成モデルのための訓練データセットを使用して、２４０で生成モデルを訓練してもよい。予測モデルのための訓練データセットを使用して、２４０で予測モデルを訓練してもよい。予測モデルは、生成モデルの品質管理機構としての役目を果たし得る。生成モデルを使用して、２６０でコアプロモーター配列を生成してもよい。予測モデルを使用して、２７０でコアプロモーター配列を、コアプロモーターとして、またはコアプロモーターではないとして分類してもよい。したがって、生成されたコアプロモーター配列を実験設定で試験する前に、予測モデルを使用して、生成モデルの設定をベンチマークし、生成された配列が、内因性配列で訓練されたモデルに基づいて、コアプロモーター活性に対して陽性であると予測されるかどうかを試験してもよい。予測モデルと生成モデルとの間のデータ漏洩を避け

るために、すべての配列は、いかなる重複も含まないように相互参照されてもよい。
【００２８】
Ｃ．訓練データの生成方法
機械学習では、訓練データセットは、さらなる適用および利用のためのベースラインとしての役割を果たす、初期データセットであってもよい。一部の実施形態では、訓練データセットは、標識される。一部の実施形態では、訓練データセットは、標識されない。一つまたは複数の機械学習技術を使用して、訓練データセットを分析し、特徴（説明変数または独立変数と呼んでもよい）と結果（必要に応じて、目的変数または従属変数と呼んでもよい）との間の関係を一般化するモデルを作成してもよい。
【００２９】
したがって、プロモーター配列データから訓練データセットを作成するための方法が説明される。訓練データセットは、訓練されるモデル（例えば、生成モデル対予測モデル）に基づいて、異なる方法に従って作成されてもよい。ヒトゲノムの候補コアプロモーターのリストを含む候補コアプロモーターデータ（プロモーター配列データ）を決定してもよい。一実施形態では、候補コアプロモーターデータは、公開されているソースからダウンロードしてもよい。候補コアプロモーターデータは、転写開始点（ＴＳＳ）プロファイリングデータとして、例えば、ＦＡＮＴＯＭ５からダウンロードしてもよい。一実施形態では、プロモーター配列データは、遺伝子発現のキャップ解析（ＣＡＧＥ）データを含んでもよい。ＣＡＧＥは、転写開始点およびそのプロモーターをマッピングするための技術である。ＣＡＧＥは、ゲノム全体の転写開始検出を可能にし、プロモーター使用解析を含む、組織／細胞／条件特異的転写開始点（ＴＳＳ）の同時同定によるハイスループット遺伝子発現プロファイリングをもたらす。ＣＡＧＥは、ｍＲＮＡの５’末端から最初の２０ヌクレオチドに由来するＤＮＡタグのコンカテマーの調製および配列決定に基づいており、これは分析されたサンプル中のｍＲＮＡの元の濃度（ＲＮＡ頻度）を反映する。ＣＡＧＥデータでは、生物学的状態（サンプル）のパネル全体のＴＳＳピークは、ＴＳＳピークの各々が、隣接し、関連するＴＳＳを含む、ＤＰＩ（分解ベースのピーク識別）によって識別されてもよい。ＴＳＳピークは、プロモーターを定義するためのアンカーとして、およびプロモーターレベルの発現分析の単位として使用され得る。ＴＳＳサミットは、所与のコアプロモーター内で最も強いシグナルを有するヌクレオチド塩基の座標を指し得る。
【００３０】
プロモーター配列データは、生成モデル用の訓練データセットおよび予測モデル用の訓練データセットを生成するために使用され得る。
１．生成モデルのための訓練データの生成
一実施形態では、プロモーター配列データは、必要に応じてフィルタリングされてもよい。例えば、プロモーター配列データは、種、成人／小児の状態、器官／組織の関連性などによって配列のみを保持するようにフィルタリングされてもよい。例えば、プロモーター配列データは、ヒト、成人、および／または肝臓データに関連付けられた配列のみを保持するようにフィルタリングされてもよい。プロモーター配列データは、正規化されてもよい。例えば、べき乗分布は正規化に使用され得る。一実施形態では、複数のライブラリ（例えば、成人肝臓から一つと成人腎臓から一つ）を使用する場合、正規化が用いられてもよい。ライブラリ内で、配列リードまたはタグの数が、ＴＳＳの相対的な強度を示している場合がある。しかしながら、シークエンスされたリードの総数は各実験で異なる可能性があるため、リードまたはタグの絶対数は、異なるシークエンシング実験間で比較できない。それらを比較できるようにするために、シークエンスのリードまたはタグのカウントが正規化される。この正規化は、例えば、リード数を倍率で割る（例えば、リードの総数を１００万で割る）ことによって行われてもよい。別の実施例では、正規化は、べき乗分布に基づいてもよい。ＣＡＧＥタグの数以下の数（＜＝）によってサポートされるＴＳＳの数は、べき乗分布によって近似され得る逆累積分布に従う。例えば、典型的なＣＡＧＥライブラリでは、１０個のＣＡＧＥタグでサポートされる１，０００個のＴＳＳと、１，０００個のＣＡＧＥタグでサポートされる１０個のＴＳＳがある。各ライブラリは、そ

の実験的に決定された分布を、分析中のライブラリとほぼ同様の仮想参照分布に当てはめることによって正規化することができる。例えば、肝臓と腎臓の両方のサンプルの上位１，０００個のＴＳＳ（ＣＡＧＥタグ数による）のみが分析される場合、正規化が有用であり得る。肝臓サンプルがより深く配列決定されている（総リード数が多い）場合、ランキングは腎臓で活性なＴＳＳによって支配されるか、またはその逆であり得る。適切に正規化した後、腎臓と肝臓の両方で上位約５００のピークが、組み合わせたサンプルの上位１，０００のＴＳＳを構成するはずである。

【００３１】

次いで、ＴＳＳをコアプロモーターと区別してもよい。コアプロモーターは、近くのＴＳＳのクラスターとみなすことができる。これらのＴＳＳは、わずかに異なる５’末端を有する機能的に等価なｍＲＮＡを生じさせる。個別のコアプロモーターは、例えば１００ｂｐのウィンドウ内に複数の小さなＴＳＳの「ピーク」があるようなブロード転写開始パターン、または例えば、１０ｂｐのウィンドウ内に単一の高いＴＳＳピークといくつかの非常に小さなＴＳＳピークがあるようなシャープ転写開始パターンを有することができる（これは一例である）。コアプロモーター、その後クラスを決定するために、どのＴＳＳが共通クラスターに属するかを決定してもよい。したがって、プロモーター配列データ中のＴＳＳは、クラスター化されてもよく、結果として生じるコアプロモーター（またはＴＳＳクラスター）の四分位幅が決定されてもよい（例えば、下分位点＝０．１、上分位点＝０．９）。一実施形態では、二つの独立したＴＳＳが互いに２０ｂｐ以下（＜＝）離れている場合、二つの独立したＴＳＳが同一のクラスター（したがってコアプロモーター）に属する、距離ベースのクラスタリングを使用してもよい。さらに、ＴＳＳは、クラスタリングに含まれる最小限のリード数によってサポートされるように要求されてもよい。

【００３２】

クラスタリング後、ＴＳＳクラスター（＝コアプロモーター）幅を決定することができる。そのためには、クラスター全体に沿って移動し、ＴＳＳタグの総数を累積和として数えることができる。その和が下分位点（例えば、合計の０．１または１０％）に当たる位置は、コアプロモーター開始として定義されてもよく、その和が、上分位点（例えば、合計の０．９または９０％）に当たる位置は、コアプロモーター終了として定義されてもよい。これらの二つの位置の間の幅は、四分位幅と呼んでもよい。

【００３３】

ＴＳＳは、四分位幅に基づいてプロモータークラスにビン化してもよい。プロモータークラスは、例えば、狭いゲノム領域内に転写が起こるシャープ型プロモーター、およびより大きなゲノム領域にＴＳＳが分散されるブロード型プロモーターがあり得る。コアプロモーターは、四分位幅によってランク付けされてもよく、下半分はシャープ（幅が小さい）、上半分はブロード（幅が広い）として標識されてもよい。シャープ型プロモーターおよびブロード型プロモーターは、それぞれＴＡＴＡボックスおよびＣｐＧアイランドと関連している可能性が高い。次いで、候補コアプロモーター配列は、各ＴＳＳについて、ＴＳＳサミットを５’方向に数塩基、および３’方向に数塩基拡張することによって決定され得る。例えば、長さ１００ｂｐの候補コアプロモーター配列を作成するために、５’方向に４９ｂｐ、および３’方向に５０ｂｐのヌクレオチドを決定することができる。候補コアプロモーター配列は、ＣＡＧＥシグナルに従ってフィルタリングされてもよい。閾値未満のＣＡＧＥシグナルを有する候補コアプロモーター配列は除外されてもよく、得られたコアプロモーター配列はコアプロモーターとして標識されてもよい。閾値は、例えば、１０を超える正規化されたカウントであってもよい。別の実施例では、閾値は、約５～約１５であり得る。カウント分布は、非常に長い「尾」を持つ分布であってもよく、ほとんどのコアプロモーターは、少数のＣＡＧＥタグのみによってサポートされているのに対し、多くのタグによってサポートされるコアプロモーターはごくわずかであることを意味する。１０を超えるカットオフを選択すると、最強のコアプロモーターのみが検討されるようになる。あるいは、例えば、約１，０００ピーク～約３，０００ピークまでの最高数のピークを使用することができる。閾値として選択されたピークの数は、総数と強度との間のトレ

ードオフを表す。コアプロモーターを多く選択するほど、弱いコアプロモーターを含むことによってシグナルが「希釈」される。このカットオフは、これら二つの相反する力の間のバランスである。他の閾値の例としては、限定されるものではないが、５（約上位５０００）、１０（約上位３０００）、２５（約上位１５００）、５０（約上位１０００）などの絶対数が挙げられる。

【００３４】

一実施形態では、制御配列のセットは、コアプロモーター配列のセットを、５’方向または３’方向に塩基数だけシフトさせることによって生成されてもよい。例えば、候補コアプロモーター配列を５’方向に５０，０００ｂｐシフトさせることによって。シフトした塩基数は、類似のクロマチンランドスケープを有するのに十分近い距離を保ちつつ、隣接する調節要素を選ばないほど十分に離れていることのバランスを表す。５’方向へのシフトは、遺伝子本体（３’方向に延び、哺乳類ゲノムでは５０ｋｂ超の長さであることが多い）へのシフトを防止する。別の実施形態では、制御配列のセットは、ゲノム全体からランダム配列を選択することによって生成されてもよい。制御配列は、任意のＣＡＧＥピークと重複する任意の制御配列を除去するためにフィルタリングされてもよく、制御配列は、コアプロモーターではないとして標識されてもよい。

【００３５】

一実施形態では、遺伝的データを受信することを含む生成モデル用の訓練データセットを生成するための方法が記述されている。遺伝的データは、第一の複数のヌクレオチド配列を含むことができる。第一の複数のヌクレオチド配列は、プロモーター配列を含み得る。複数のヌクレオチド配列の各ヌクレオチド配列は、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含み得る。関連発現スコアは、ＣＡＧＥピークを含み得る。遺伝的データは、プロモーター配列データを含んでもよい。遺伝的データは正規化されてもよい。べき乗法の適用を含む、当技術分野で公知の任意の正規化技術を使用してもよい。ＴＳＳは、関連発現スコアに基づいてクラスター化されてもよく、ＴＳＳの各クラスターについて、四分位幅が決定されてもよい。四分位幅は、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳとして標識するために使用され得る。複数のサミットヌクレオチド塩基が、ＴＳＳで決定されてもよい。サミットヌクレオチド塩基を決定することは、最も強いＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含み得る。各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定してもよい。関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定し、それによって候補コアプロモーター配列を形成することを含み得る。５’方向の第一の複数のヌクレオチド塩基は、４９個のヌクレオチド塩基を含むことができ、３’方向の第二の複数のヌクレオチド塩基は、５０個のヌクレオチド塩基を含むことができる。各サミットヌクレオチド塩基およびその関連する複数の周辺塩基は、コアプロモーターとして標識された第二の複数のヌクレオチド配列（候補コアプロモーター配列）として保存されてもよい。第三の複数のヌクレオチド配列は、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から決定され得る。閾値は、例えば、１０を超える正規化されたカウントであってもよい。別の実施例では、閾値は、約５～約１５であり得る。カウント分布は、非常に長い「尾」を持つ分布であってもよく、ほとんどのコアプロモーターは、少数のＣＡＧＥタグのみによってサポートされているのに対し、多くのタグによってサポートされるコアプロモーターはごくわずかであることを意味する。１０を超えるカットオフを選択すると、最強のコアプロモーターのみが検討されるようになる。あるいは、例えば、約１，０００ピーク～約３，０００ピークまでの最高数のピークを使用することができる。閾値として選択されたピークの数は、総数と強度との間のトレードオフを表す。コアプロモーターを多く選択するほど、弱いコアプロモーターを含むことによってシグナルが「希釈」される。このカットオフは、これら二つの相反する力の間のバランスである。他の閾値の例としては、限定されるものではないが、５（約上位５０００）、１０（約上位３０００）、２５（約上位１５００）、５０（約上位１０００）

などの絶対数が挙げられる。第三の複数のヌクレオチド配列は、コアプロモーター配列のセットを含んでもよい。コアプロモーター配列のセットは、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する任意の配列に対してさらにフィルタリングされてもよい。

【００３６】

制御配列のセットは、第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定し、関連する複数のシフト塩基のそれぞれを、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として保存することによって生成され得る。

【００３７】

コアプロモーター配列のセットおよび制御配列のセットは、生成モデルのための訓練データセットとして保存されてもよい。

２．予測モデルのための訓練データの生成

プロモーター配列データは、ＣＡＧＥピーク閾値を適用することによってフィルタリングされてもよい。例えば、プロモーター配列データの上位ＣＡＧＥピークのみを使用してもよい。一実施形態では、ＣＡＧＥピーク閾値は、予測モデル用に選択されたコアプロモーターの強度が、生成モデル用に選択されたコアプロモーターの強度と合致するように設定されてもよい。コアプロモーターの強度が合致する場合、新規に生成されたコアプロモーターの分類は、より信頼性が高い場合がある。しかしながら、予測モデルに採用される機械学習モデル（例えば、ロジスティック回帰モデル）が、生成モデルに採用される機械学習モデル（例えば、ニューラルネットワーク）よりもバイアスが大きい可能性があるため、コアプロモーターの数は、予測モデルの方が少なくてもよい。そのため、予測モデルは過剰適合する可能性が低く、より少ない例で訓練することができる。生成モデルにおける過剰適合を避けるために、予測モデルよりも多くの実施例を訓練に使用してもよい。予測モデルと生成モデルの両方に対して同様の強度のコアプロモーターを確保するために、閾値はそれに応じて選択されるべきであるが、予測モデルに対してそれほど重要ではない可能性がある生成モデルに対して十分な数のコアプロモーターを確保するために、入力データを選択すべきである。

【００３８】

フィルタリングされたプロモーター配列データをさらにフィルタリングして、生成モデルのために生成された訓練データセットのピークのいずれかと重複する任意の配列データを除去してもよい。次いで、コアプロモーター配列のセットは、各ＴＳＳについて、ＴＳＳサミットを５’方向に数塩基、および３’方向に数塩基拡張することによって決定され得る。例えば、長さ１００ｂｐのコアプロモーター配列を作成するために、５’方向に４９ｂｐ、および３’方向に５０ｂｐのヌクレオチドを決定することができる。コアプロモーター配列のセットは、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する任意の配列に対してさらにフィルタリングされてもよい。

【００３９】

制御配列のセットは、コアプロモーター配列のセットを、５’方向または３’方向に塩基数だけシフトさせることによって生成されてもよい。例えば、候補コアプロモーター配列を５’方向に５０，０００ｂｐシフトさせることによって。制御配列をフィルタリングして、任意のＣＡＧＥピークと重複する任意の制御配列、および生成モデルのための制御配列のセットと重複する任意の制御配列を除去することができ、制御配列は、コアプロモーターではないとして標識されてもよい。

【００４０】

一実施形態では、遺伝的データを受信することを含む予測モデル用の訓練データセットを生成するための方法が説明されている。遺伝的データは、第一の複数のヌクレオチド配列を含むことができる。第一の複数のヌクレオチド配列は、プロモーター配列を含み得る。複数のヌクレオチド配列の各ヌクレオチド配列は、関連発現スコアを有する少なくとも一つのＴＳＳを含み得る。関連発現スコアは、ＣＡＧＥピークを含み得る。遺伝的データは、プロモーター配列データを含んでもよい。第一の複数のヌクレオチド配列をフィルタ

リングして、生成モデルのために生成された訓練データセットのピークのいずれかと重複する任意の配列データを除去することができる。複数のサミットヌクレオチド塩基が、ＴＳＳで決定されてもよい。サミットヌクレオチド塩基を決定することは、最も強いＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含み得る。各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定してもよい。関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定し、それによって候補コアプロモーター配列を形成することを含み得る。５’方向の第一の複数のヌクレオチド塩基は、４９個のヌクレオチド塩基を含むことができ、３’方向の第二の複数のヌクレオチド塩基は、５０個のヌクレオチド塩基を含むことができる。各サミットヌクレオチド塩基およびその関連する複数の周辺塩基は、コアプロモーターとして標識された第二の複数のヌクレオチド配列（コアプロモーター配列のセット）として保存されてもよい。コアプロモーター配列のセットは、例えば、ヒトゲノムアセンブリ（ｈｇ１９、ｈｇ３８など）中にＮｓを含有する任意の配列に対してさらにフィルタリングされてもよい。

【００４１】

制御配列のセットは、第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定し、関連する複数のシフト塩基のそれぞれを、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列（制御配列のセット）として保存することによって生成され得る。制御配列のセットをフィルタリングして、任意のＣＡＧＥピークと重複する任意の制御配列、および生成モデルのための制御配列のセットと重複する任意の制御配列を除去することができる。

【００４２】

コアプロモーター配列のセットおよび制御配列のセットは、予測モデルの訓練データセットとして保存されてもよい。

Ｄ．生成モデル

１．生成モデルを生成する方法

本明細書において、一つまたは複数のリカレントニューラルネットワーク（ＲＮＮ）、例えば、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＬＳＴＭ－ＲＮＮ）などを使用して、長さの異なる新規のコアプロモーター配列を生成する技術が開示される。ＬＳＴＭ－ＲＮＮモデルをシード配列に適用して、シード配列から可能性の高い次のヌクレオチドを予測することができる。可能性の高い次のヌクレオチドを、シード配列に連結し、得られた配列を、ＬＳＴＭ－ＲＮＮモデルに戻して、シード配列と以前に決定した可能性の高い次のヌクレオチドから別の可能性の高い次のヌクレオチドを予測することができる。ＬＳＴＭ－ＲＮＮモデルを使用して、任意の長さのコアプロモーター配列を生成してもよい。

【００４３】

ＲＮＮは、ユニット間の接続が有向サイクルを形成する人工ニューラルネットワークの一種である。ＲＮＮは、ネットワークが動的な時間的挙動を示すことができる内部状態を有する。例えば、フィードフォワードニューラルネットワークとは異なり、ＲＮＮは、その内部メモリを使用して、任意の入力シークエンスを処理できる。ＬＳＴＭ－ＲＮＮは、標準的なニューラルネットワークユニットの代わりに、またはそれに加えて、ＬＳＴＭユニットをさらに含む。ＬＳＴＭユニット、またはブロックは、任意の長さの時間にわたって値を記憶または保存できる、「スマート」ユニットである。ＬＳＴＭブロックには、その入力が記憶するのに十分重要である時、値を記憶し続けるべきまたは忘れるべき時、および値を出力するべき時を決定するゲートを含む。

【００４４】

議論の明確化のために、本開示全体を通して議論される実施形態は、ＬＳＴＭ－ＲＮＮに関して議論される。しかしながら、様々なタイプのＲＮＮが、例えば、メモリ細胞シークエンシング操作に関するバリアント（例えば、双方向、単方向、後方視単方向、または

後方視ウィンドウを有する前方視単方向）またはメモリ細胞タイプに関するバリアント（例えば、ＬＳＴＭバリアントまたはゲーティングされた反復単位（ＧＲＵ））を含む、記載した実施形態で使用され得る。ＢＬＳＴＭ－ＲＮＮでは、出力は過去と未来両方の状態情報に依存する。さらに、乗算ユニットのゲートにより、メモリ細胞は、過去と未来両方のイベントの長いシークエンスにわたって情報を保存し、アクセスすることができる。さらに、他のタイプの位置認識ニューラルネットワーク、または連続予測モデルを、ＬＳＴＭ－ＲＮＮまたはＢＬＳＴＭ－ＲＮＮの代わりに使用することができる。本開示全体を通して示されるように、ＬＳＴＭ－ＲＮＮは、入力層、出力層、および一つまたは複数の非表示層を含む。

【００４５】

図３、図４Ａ、図４Ｂ、および図５はそれぞれ、ニューラルネットワーク３００、ＲＮＮブロック４００、およびＬＳＴＭ　ＲＮＮブロック５００の概要を提供するために示される。図３は、ニューラルネットワーク３００の一例を示す。ニューラルネットワーク３００は、入力ノード、ブロック、またはユニット３０２、出力ノード、ブロック、またはユニット３０４、および非表示ノード、ブロック、またはユニット３０４を含む。入力ノード３０２は、接続３０８を介して非表示ノード３０６に接続され、非表示ノード３０６は、接続３１０を介して出力ノード３０４に接続される。

【００４６】

入力ノード３０２は、入力データに対応する一方、出力ノード３０４は、入力データの関数として出力データに対応する。例えば、入力ノード３０２は、入力配列に対応してもよく、出力ノード１０４は、出力配列またはヌクレオチドに対応してもよい。ノード３０６は、ニューラルネットワークモデル自体がノードを生成する非表示ノードである。ノード３０６の一つの層のみが図示されるが、実際には、ノード３０６の二つ以上の層が通常存在する。

【００４７】

したがって、ニューラルネットワーク３００を構築するために、手動でまたは別の方法で既に出力データにマッピングされた入力データ形式で訓練データが、ネットワーク３００を生成するニューラルネットワークモデルに提供される。したがって、モデルは、非表示ノード３０６、入力ノード３０２と非表示ノード３０６との間の接続３１０の重み、非表示ノード３０６と出力ノードとの間の接続３１０の重み、および非表示ノード３０６自体の層間の接続の重みを生成する。その後、ニューラルネットワーク３００は、出力データが未知である入力データに対して使用して、所望の出力データを生成することができる。

【００４８】

ＲＮＮは、ニューラルネットワークの一種である。一般的なニューラルネットワークは、入力データを処理して出力データを生成する間、いかなる中間データも保存しない。比較すると、ＲＮＮはデータを持続し、そうでない一般的なニューラルネットワークよりもその分類能力を向上させることができる。

【００４９】

図４Ａは、ＲＮＮであるニューラルネットワーク３００の非表示ノード３０６を代表する、ＲＮＮブロック４００の一例をコンパクトに表記したものを示す。ＲＮＮブロック４００は、入力ノード３０２のうちの一つからつながる図３の接続３０８であり得る、または別の非表示ノード３０６からつながる接続であり得る、入力接続４０２を有する。ＲＮＮブロック４００は同様に、出力ノード３０４のうちの一つにつながる図３の接続３１０であり得る、または別の非表示ノード３０６につながる接続であり得る、出力接続４０４を有する。

【００５０】

ＲＮＮブロック４００は概して、出力接続４０４上に提供される情報を生成するために、入力接続４０２上に提供される情報に対して（少なくとも）実行される処理４０６を含むと言われる。処理４０６は、典型的には関数の形態である。例えば、関数は、出力接続４０４を入力接続４０２にマッピングする、同一性活性化関数であってもよい。関数は、

入力接続４０２に基づいて、範囲（０、１）内の値を出力できる、ロジスティックシグモイド関数などのシグモイド活性化関数であってもよい。関数は、入力接続４０２に基づいて、範囲（－１、１）内の値を出力できる、ハイパボリックロジスティックタンジェント関数などのハイパボリックタンジェント関数であってもよい。

【００５１】

ＲＮＮブロック４００はまた、それ自体の時間的後継回路に戻る時間的ループ接続４０８を有する。接続４０８は、ＲＮＮブロック４００を再帰的にするものであり、複数のノード内のかかるループの存在は、ニューラルネットワーク３００を再帰的にするものである。したがって、ＲＮＮブロック４００が接続４０４上に出力する情報（または他の情報）は、接続４０８上に持続することができ、これに基づいて、接続４０２上で受信された新しい情報を処理することができる。すなわち、ＲＮＮブロック４００が接続４０４上に出力する情報は、ＲＮＮブロック４００が次に入力接続４０２上で受信する情報とマージ、または連結され、処理４０６を介して処理される。

【００５２】

図４Ｂは、ＲＮＮブロック４００の拡張表記を示す。ＲＮＮブロック４００’および接続４０２’、４０４’、４０６’、４０８’は、同じＲＮＮブロック４００および接続４０２、４０４、４０６、４０８であるが、時間的には遅い時間のものである。したがって、図４Ｂは、ＲＮＮブロック４００’が、早い時間で（同じ）ＲＮＮブロック４００によって提供される接続４０６に提供される情報を、遅い時間で受信することを示す。遅い時間でのＲＮＮブロック４００’は、それ自体が接続４０６’上でさらに遅い時間にそれ自体に情報を提供することができる。

【００５３】

ＬＳＴＭ－ＲＮＮは、ＲＮＮの一種である。理論上の一般的なＲＮＮは、短期的および長期的に情報を維持することができる。しかしながら、実際には、こうしたＲＮＮは、長期的に情報を維持する能力が証明されていない。より厳密には、一般的なＲＮＮは、実質的に長期的な依存関係を学習することができない。つまり、ＲＮＮは、以前に比較的長い期間処理した情報に基づいて情報を処理することができない。比較すると、ＬＳＴＭ－ＲＮＮは、長期的な依存関係を学習できる特別なタイプのＲＮＮであり、それゆえ、長期的に情報を維持することのできるタイプのＲＮＮである。

【００５４】

図５は、例示的なＬＳＴＭ－ＲＮＮブロック５００’を示す。ＬＳＴＭ－ＲＮＮブロック５００’は、図４Ａおよび図４ＢのＲＮＮブロック４００／４００’の接続４０２／４０２’および４０４／４０４’ならびに処理４０６／４０６’と同等の入力接続５０２’、出力接続５０４’および処理５０６’を有する。しかしながら、ＲＮＮブロック４００／４００’の時間的インスタンスを接続する単一の時間的ループ接続４０８／４０８’を有するのではなく、ＬＳＴＭ－ＲＮＮブロック５００’は、ＬＳＴＭ－ＲＮＮブロック５００の時間的インスタンス間で情報が持続する、二つの時間的ループ接続５０８’および５１０’を有する。

【００５５】

入力接続５０２’上の情報は、ＬＳＴＭ－ＲＮＮブロックの以前の時間的インスタンスから接続５０８上に提供された持続的情報にマージされ、処理５０６’を受ける。処理５０６’の結果は、もしあるとしても、ＬＳＴＭ－ＲＮＮブロックの以前の時間的インスタンスから接続５１０上に提供された持続的情報とどのように組み合わされるかは、ゲート５１２’および５１４’を介して制御される。接続５０２’および５０８’のマージ情報に基づいて動作するゲート５１２’は、接続５１０上の持続情報の通過（または非通過）を許可する、要素別積演算子５１６’を制御する。同じ基準で動作するゲート５１４’は、処理５０６’の出力の通過（または非通過）を許可する要素別演算子５１８’を制御する。

【００５６】

演算子５１６’および５１８’の出力は、追加演算子５２０’を介して合計され、ＬＳＴＭ－ＲＮＮブロック５００’の現在のインスタンスの接続５１０’上の持続的情報として渡される。したがって、接続５１０’の持続情報が接続５１０の持続情報を反映する程度、およ

び接続５１０’のこの情報が処理５０６’の出力を反映する程度は、ゲート５１２’および５１４’によって制御される。そのため、情報は、必要に応じて、ＬＳＴＭ－ＲＮＮブロックの複数の時間的インスタンス全体、またはそれらにわたって持続し得る。

【００５７】

ＬＳＴＭ－ＲＮＮブロック５００’の現在のインスタンスの出力は、それ自体が、ＲＮＮの次の層への接続５０４’上で提供され、また接続５０８’上でＬＳＴＭ－ＲＮＮブロックの次の時間的インスタンスにも持続する。この出力は、別の要素別積演算子５２２’によって提供され、ゲート５２４’および５２６’によってそれぞれ制御されるように、接続５１０’上にも提供される情報と、接続５０２’および５０８上のマージされた情報との組み合わせを通過させる。このようにして、次に、図５のＬＳＴＭ－ＲＮＮブロック５００’は、長期情報および短期情報の両方を持続させることができる。

【００５８】

以下でさらに詳細に説明するように、例えばＬＳＴＭ－ＲＮＮなどの一つまたは複数の（ＲＮＮ）を使用して、プロモーター配列を決定してもよい。ＬＳＴＭ－ＲＮＮは、訓練プロセスを介して、一連の入力特徴に基づいて一連のパラメータを予測するモデルを提供する。図６は、ＬＳＴＭ－ＲＮＮを訓練するための方法６００の一例を示す。一部の変形では、訓練プロセスは、専用ソフトウェア（例えば、ｋｅｒａｓ、ＴｅｎｓｏｒＦｌｏｗスクリプト、ＣＵＲＲＥＮＮＴツールキット、または市販のツールキットの修正版など）および／または専用ハードウェアを使用して実施され得る。

【００５９】

工程６１０で、コンピューティングデバイスは、本明細書に記述されたように、訓練データセットを決定し得る。訓練データセットは、コアプロモーターとして標識されたコアプロモーター配列のセットと、コアプロモーターではないとして標識された制御配列のセットとを含んでもよい。訓練データセットは、ＬＳＴＭ－ＲＮＮの訓練および検証に使用され得る。一部の変形では、訓練データの第一の部分を訓練に使用してもよく、訓練データの第二の部分を試験／検証に使用してもよい。コアプロモーター配列のセットは、異なるクラス（シャープ／ブロード）の配列を、シード配列と予測標的の対に分割してもよい。シード配列は、例えば、１０個のヌクレオチドなど、任意の長さのヌクレオチド配列を含んでもよい。予測標的は、シード配列のすぐ後にあるヌクレオチドを含んでもよい。シード配列／予測標的対は、工程サイズ１のスライディングウィンドウアプローチを使用してコアプロモーター配列を分割することによって生成され得る。次いで、配列対は、数値符号化（例えば、「Ａ」：０、「Ｃ」：１、「Ｇ」：２、「Ｔ」：３）を使用してベクター化されてもよい。コアプロモーターの異なるクラスは、その生物学において根本的に異なる。典型的には、プラスミドベースまたはウイルスベクターベースの導入遺伝子では、「シャープ」プロモータークラスのみが使用される。二つのクラスは、その配列内容が異なるため、二つのクラスは、訓練または配列生成のために混合することができない。言い換えれば、両方のクラスのコアプロモーターが訓練に使用された場合、「シャープ」コアプロモーターに通常見られるコアプロモーターモチーフからのシグナルは希釈され、非機能性配列の生成につながる可能性がある。代わりに、二つのタイプが分離され、その後、二つの別々に訓練されたモデルに基づいて新しい配列が生成され、それによって、「シャープ」または「ブロード」コアプロモーターの新規のインスタンスを生成することができる。どのタイプのコアプロモーターが生成されるかは、用途により異なるが、最も一般的には、「シャープ」タイプである。結論として、分離は用途にはあまり関係しないが（通常、シャープコアプロモーターのみが使用されるため）、シャープとブロードの混合が正しいシグナルを希釈することを考えると、適切な訓練にはより関係がある。しかしながら、別個のクラスを使用して、「ブロード」タイプのコアプロモーターを生成することもできる。

【００６０】

工程６２０で、コンピューティングデバイスは、ＬＳＴＭ－ＲＮＮのモデルセットアップを実行し得る。モデルセットアップには、ＬＳＴＭ－ＲＮＮの接続の重みを初期化する

ことが含まれ得る。一部の配置では、事前訓練は、重みを初期化するために実施され得る。その他の場合、重みは、分布スキームに従って初期化されてもよい。一つの分布スキームは、平均が０、標準偏差が０．１の正規分布に従って、重みの値を無作為化することを含む。一実施形態では、各ヌクレオチドは重みと関連付けられてもよい。例えば、重みは、Ａ：０．２、Ｃ：０．０５、Ｇ：０．６、Ｔ：０．１５のように割り当てることができる。

【００６１】

工程６３０で、コンピューティングデバイスは、ＬＳＴＭ－ＲＮＮのモデル訓練を実行し得る。一部の変形では、モデル訓練は、例えば、最急降下、確率的勾配降下、または弾性逆伝播（ＲＰＲＯＰ）を含む、一つまたは複数の訓練技術を実施することを含み得る。訓練技術は、訓練セットをモデルに適用し、モデルを調整する。モデル訓練をスピードアップするために、並列訓練を実施することができる。例えば、訓練セットは、他のバッチと並列に処理される１００個の配列のバッチに分割されてもよい。

【００６２】

工程６４０で、コンピューティングデバイスは、ＬＳＴＭ－ＲＮＮのモデル検証を実行し得る。一部の変形では、検証セットを、訓練されたモデルに適用し、ヒューリスティックを追跡してもよい。ヒューリスティックは、一つまたは複数の停止条件と比較されてもよい。停止条件が満たされる場合、訓練プロセスは終了し得る。一つまたは複数の停止条件のいずれも満たされない場合、工程６３０のモデル訓練および工程６４０の検証が繰り返される場合がある。工程６３０および６４０の各反復は、訓練エポックと呼ばれることがある。追跡され得るヒューリスティックの一部には、二乗誤差の合計（ＳＳＥ）、加重二乗誤差の合計（ＷＳＳＥ）、回帰ヒューリスティック、または訓練エポックの数が含まれる。

【００６３】

２．生成モデルを使用する方法

図７は、プロモーター配列を決定するための訓練されたＬＳＴＭ－ＲＮＮ７１０を使用する例示的な流れを示す。ＬＳＴＭ－ＲＮＮ７１０は、入力配列７２０（例えば、「シード」）を受けるように構成されてもよい。入力配列７２０は、ヌクレオチド配列を含んでもよい。ヌクレオチド配列は、プロモーター配列（例えば、コアプロモーター配列）を含んでもよい。入力配列７２０は、長さを有してもよい。長さは、例えば、約５ヌクレオチド～約１００ヌクレオチドであってもよい。入力配列７２０は、１０ヌクレオチド長であってもよい。他の入力配列長、例えば、１、２、３、４、５、６、７、８、９、１１、１２、１３、１４、１５、１６、１７、１８、１９、または２０ヌクレオチド長が企図される。一実施形態では、入力配列７２０のソースは、訓練データセットからのランダムに選択されたコアプロモーターのランダム配列または第一の１０ｎｔ（または任意の他の入力配列長）であってもよい。一実施形態では、実世界の例からの特定のコアプロモーターモチーフにつながる１０ｎｔ（または任意の他の入力配列長）は、そのモチーフを有するコアプロモーターの生成を強制するために、入力配列７２０として選択されてもよい。別の実施形態では、ランダム配列を入力配列として使用してもよく、出力配列を特定のモチーフの存在についてスクリーニングしてもよい。

【００６４】

入力配列７２０は、入力層、一つまたは複数の非表示層、および出力層を介して、入力配列７２０を処理するＬＳＴＭ－ＲＮＮ７１０に入力される。ＬＳＴＭ－ＲＮＮ７１０は、出力層を介して、可能性の高い次のヌクレオチドを出力する。したがって、ＬＳＴＭ－ＲＮＮ７１０は、出力配列７３０を生成するために入力配列７２０に付加され得る、可能性の高い次のヌクレオチドを予測するように構成されてもよい。ＬＳＴＭ－ＲＮＮ７１０は、ヌクレオチド確率に基づいて、可能性の高い次のヌクレオチドを予測するように構成されてもよい。ヌクレオチド確率は、例えば、Ａ：０．２、Ｃ：０．０５、Ｇ：０．６、Ｔ：０．１５であり得る。ＬＳＴＭ－ＲＮＮ７１０は、生成された出力配列７３０を入力として取り、出力配列７３０を新しい入力配列７２０として効果的に処理するように構成

されてもよい。ＬＳＴＭ－ＲＮＮ７１０は、出力配列７３０に対して所望の長さが達成されるまで、可能性の高い次のヌクレオチドを繰り返し予測するように構成されてもよい。所望の出力長は、例えば、約２０ヌクレオチド～約１００ヌクレオチドであってもよい。所望の出力長は、５０ヌクレオチドであってもよい。ＬＳＴＭ－ＲＮＮ７１０は、前のヌクレオチドのプロモーター配列が与えられた新しいプロモーター配列において、次のヌクレオチドの確率分布を生成する。これにより、ＬＳＴＭ－ＲＮＮ７１０は、一度に一つの新しいプロモーター配列を生成することができる。一実施形態では、任意の数のコアプロモーターを生成することができ、その後、ＧＣ含量またはコアプロモーターモチーフ含量（予測モデルで使用されるものと類似した特徴）などの特定の態様についてスクリーニングすることができる。

【００６５】

図８は、コアプロモーター配列の生成の視覚的描写である。シード配列８０２は、ＬＳＴＭ－ＲＮＮ７１０に入力されてもよく、所望の最終コアプロモーター配列長が指定されてもよい。図８の実施例では、シード配列８０２は、１０ヌクレオチド長であり、所望の最終コアプロモーター配列長は、１４ヌクレオチドである。ＬＳＴＭ－ＲＮＮ７１０は、シード配列８０２が与えられると、次に可能性の高いヌクレオチド８０４を予測し得る。次に可能性の高いヌクレオチド８０４をシード配列８０２に添加（連結）して、配列８０６を作成してもよい。ＬＳＴＭ－ＲＮＮ７１０は、配列８０６の少なくとも一部が与えられると、次に可能性の高いヌクレオチド８０８を予測し得る。例えば、ｎ個のヌクレオチドのスライディングウィンドウを使用して、次に可能性の高いヌクレオチド８０８を予測してもよい。スライディングウィンドウは、例えば、５、６、７、８、９、１０、１１、１２、１３、１４、または１５ヌクレオチド長であってもよい。次に可能性の高いヌクレオチド８０８を配列８０６に添加（連結）して、配列８１０を作成してもよい。ＬＳＴＭ－ＲＮＮ７１０は、配列８１０の少なくとも一部が与えられると、次に可能性の高いヌクレオチド８１２を予測し得る。例えば、ｎ個のヌクレオチドのスライディングウィンドウを使用して、次に可能性の高いヌクレオチド８１２を予測してもよい。次に可能性の高いヌクレオチド８１２を配列８１０に添加（連結）して、配列８１４を作成してもよい。ＬＳＴＭ－ＲＮＮ７１０は、配列８１４の少なくとも一部が与えられると、次に可能性の高いヌクレオチド８１６を予測し得る。例えば、ｎ個のヌクレオチドのスライディングウィンドウを使用して、次に可能性の高いヌクレオチド８１６を予測してもよい。次に可能性の高いヌクレオチド８１６を配列８１６に添加（連結）して、最終コアプロモーター配列８１８を作成してもよい。最終コアプロモーター配列８１８の長さは、所望の最終コアプロモーター配列長と等しく、最終コアプロモーター配列８１８は、コアプロモーターとして出力されてもよい。

【００６６】

一実施形態では、ａ）ヌクレオチド配列と配列長を受信すること、ｂ）訓練された生成モデルに、ヌクレオチド配列を提供すること、ｃ）生成モデルに基づいて、ヌクレオチド配列に関連付けられた次のヌクレオチドを決定すること、ｄ）ヌクレオチド配列に次のヌクレオチドを付与すること、ｅ）ヌクレオチド配列の長さが配列長に等しくなるまでｂ～ｄを繰り返すこと、およびｆ）ヌクレオチド配列をコアプロモーター配列として出力すること、を含む方法が記述される。配列長は、約５０ヌクレオチド～約１００ヌクレオチドとすることができる。

【００６７】

本明細書に記載されるように、定義された長さおよび／または定義された配列含量を有するコアプロモーター配列を生成するための方法およびシステムが提供される。生成されたコアプロモーター配列は、例えば、前臨床研究および遺伝子治療の両方におけるＡＡＶ、アデノ、またはレンチウイルスなどのウイルスベクター；Ｃａｓ９発現、Ｃｒｅ－ｌｏｘ、ＴＡＬＥＮまたはジンクフィンガーヌクレアーゼを駆動するものなどのゲノム編集ベクター；発光または蛍光レポータープラスミド；高収率の抗体発現プラスミド；クローニングおよびエンジニアリングプラスミド；化学遺伝学（例えばＤＲＥＡＤＤ）、および光

遺伝学などの任意の種類の導入遺伝子に使用することができる。一部の態様では、生成されたコアプロモーターのうちの一つまたは複数を含む構築物またはベクターは、核酸構築物と呼んでもよい。したがって、開示された核酸構築物は、生成されたコアプロモーターおよび一つまたは複数の導入遺伝子を含み得る。一部の態様では、開示される核酸構築物は、任意の発現ベクター、ウイルスまたは非ウイルスであってもよい。一部の態様では、核酸構築物は、直鎖状または環状であってもよい。環状核酸構築物は、プラスミドまたはベクターと呼んでもよい。

【００６８】

したがって、生成されたコアプロモーター配列は、遺伝子治療のためのウイルスベクター上の導入遺伝子に使用され得る。現在、ＡＡＶは遺伝子治療のゴールドスタンダードだが、ゲノムサイズが非常に限られているため、ＡＡＶベクターで使用されるいかなる要素も、サイズを最適化する必要がある。例えばＣａｓ９などの一部の導入遺伝子のコード配列は、サイズを最適化することができないが、他の調節要素（コアプロモーターなど）は可能である。したがって、１００ｎｔ長であり得る内因性コアプロモーターを使用する代わりに、開示された方法およびシステムは、５０ｎｔ長であり、定義されたモチーフ含量を有し、したがって、任意の内因性コアプロモーターよりも遺伝子発現を駆動するのにさらに効率的なコアプロモーターを生成することができる。

【００６９】

一実施形態では、生成されたコアプロモーター配列の配列含量は、遺伝子治療設定における自然免疫反応を回避するために定義されてもよい。コアプロモーターは、多くの場合、高い含量のＣｐＧジヌクレオチドを有し、これはＴＬＲベースの自然免疫反応をトリガすることができる。説明された生成モデルを使用して、ＣｐＧジヌクレオチドを欠くコアプロモーター配列を生成することができる。

【００７０】

方法は、コアプロモーター配列に基づいてプロモーターを操作することをさらに含んでもよい。方法は、プロモーターを核酸構築物に挿入することをさらに含んでもよい。プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含む。方法は、核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することをさらに含んでもよい。一部の態様では、本開示の核酸構築物を含む、任意の公知のウイルスベクターを作製することができる。一部の態様では、方法は、生成されたコアプロモーターを含む任意の公知の非ウイルスベクター（例えば、ＤＮＡベースのベクター）を作製することを含んでもよい。

【００７１】

用語「発現ベクター」は、（例えば、転写制御要素に連結された）細胞による発現に適した形態の導入遺伝子を含有する任意のベクター（例えば、プラスミド、コスミドまたはファージ染色体）を含む。一部の態様では、プラスミドは一般的に使用されるＤＮＡベクターの形態であるため、「プラスミド」と「ベクター」は互換的に使用される。さらに、本発明は、同等の機能を果たす他のベクターを含むことを意図している。

【００７２】

Ｅ．予測モデル

１．予測モデルを生成する方法

ここで図９を参照すると、予測モデルを生成するための方法が説明されている。説明された方法は、所定の配列に対するプロモーター状態（例えば、プロモーター／非プロモーター）を予測するよう構成されている少なくとも一つのＭＬモジュール９３０である、訓練モジュール９２０による、一つまたは複数の訓練データセット９１０の分析に基づき、訓練するための機械学習（「ＭＬ」）技術を使用してもよい。

【００７３】

訓練データセット９１０は、コアプロモーター配列（ＹＥＳ）として標識されたコアプロモーター配列のセット、およびコアプロモーター配列（ＮＯ）ではないとして標識された制御配列のセットを含んでもよい。このようなデータは、全体的または部分的に、本明

細書に記載のプロモーター配列データから導出されてもよい。

【００７４】

コアプロモーター配列のセットおよび制御配列のセットのサブセットは、訓練データセット９１０または試験データセットにランダムに割り当てられてもよい。一部の実施では、訓練データセットまたは試験データセットへのデータの割り当ては完全に無作為ではない場合がある。この場合、一つまたは複数の基準が、割り当て中に使用され得る。一般に、任意の好適な方法を使用して、データを訓練データセットまたは試験データセットに割り当ててもよい一方で、はいおよびいいえの標識分布が、訓練データセットおよび試験データセットにおいていくらか類似していることを保証し得る。

【００７５】

訓練モジュール９２０は、一つまたは複数の特徴選択技術により、訓練データセット９１０における複数のコアプロモーター配列（例えば、はいとして標識された）および／または複数の制御配列（例えば、いいえとして標識された）から特徴セットを抽出することによって、ＭＬモジュール９３０を訓練してもよい。訓練モジュール９２０は、正の例（例えば、はいであると標識された）の統計上有意な特徴および負の例（例えば、いいえであると標識された）の統計上有意な特徴を含む訓練データセット９１０から、特徴セットを抽出することによって、ＭＬモジュール９３０を訓練してもよい。

【００７６】

訓練モジュール９２０は、様々な方法で、訓練データセット９１０から特徴セットを抽出してもよい。訓練モジュール９２０は、異なる特徴抽出技術を使用して、各回に特徴抽出を複数回実施し得る。一例では、異なる技術を使用して生成される特徴セットは各々が、異なる機械学習ベースの分類モデル９４０を生成するために使用され得る。例えば、最も高い品質の測定基準を伴う特徴セットが、訓練における使用のために選択され得る。訓練モジュール９２０は、新規の配列（例えば、未知のプロモーター状態を有する）が、プロモーターである可能性が高いか、または低いかを示すよう構成されている、一つまたは複数の機械学習ベースの分類モデル９４０Ａ～９４０Ｎを構築するための特徴セットを使用してもよい。

【００７７】

訓練データセット９１０を分析して、訓練データセット９１０における特徴とはい／いいえの標識の間の任意の依存性、関連性、および／または相関を決定してもよい。識別された相関は、異なるはい／いいえの標識と関連する特徴のリストの形態を有してもよい。本明細書で使用される場合、用語「特徴」は、データのある項目が、一つまたは複数の特定のカテゴリー内にあるか否かを決定するために使用され得るデータの項目の任意の特徴を指し得る。例として、本明細書に記載される特徴は、一つまたは複数の配列パターン、ＧＣ含量、ＣｐＧ含量、公知のコアプロモーター配列モチーフ、ＡＴＧ頻度、および／または相対エントロピーを含んでもよい。公知のコアプロモーター配列モチーフの発生については、ＴＳＳに対する相対位置決め方式も考慮され得る。相対エントロピーは、固定ヌクレオチド分布に基づき、ある配列がランダム配列とどれほど類似しているかの尺度である（コアプロモーター配列はランダムではない）。ランダムＤＮＡの確率の例は、ランダムＤＮＡの確率：Ａ：０．３、Ｃ：０．２、Ｔ：０．２、Ｇ：０．３である。図１０は、プロモーター状態を予測する様々な特徴の相対的有意性を示す。例えば、ＣｐＧジヌクレオチド含量は、コアプロモーターの同一性と最も強い正の相関を有する。ランダムＤＮＡは、比較的少数のＣｐＧを有する一方で、コアプロモーターは、多数を有することができる（コアプロモーターのサブセットは、ＣｐＧアイランドと呼ばれる）。ＴＡＴＡボックスの有無も重要である。

【００７８】

図９に戻ると、特徴選択技術は、一つまたは複数の特徴選択ルールを含み得る。一つまたは複数の特徴選択ルールは、特徴発生ルールを含み得る。特徴発生ルールは、訓練データセット９１０においていずれの特徴が閾値の回数にわたって生じるかを決定すること、および閾値を満たすそれらの特徴を、特徴として特定することを含み得る。

【００７９】
単一の特徴選択ルールを、特徴を選択するために適用してもよく、または複数の特徴選択ルールを、特徴を選択するために適用してもよい。特徴選択ルールは、カスケード方式で適用されてもよく、特徴選択ルールは、特定の順序で適用され、以前のルールの結果に適用される。例えば、特徴発生ルールは、訓練データセット９１０に適用されて、特徴の第一のリストを生成し得る。特徴の最終リストは、一つまたは複数の特徴群（例えば、プロモーター状態を予測するために使用され得る特徴の群）を決定するためのさらなる特徴選択技術により分析されてもよい。任意の好適な計算技術を使用して、フィルター方法、ラッパー方法、および／または埋め込み方法などの任意の特徴選択技術を使用して、特徴群を特定し得る。一つまたは複数の特徴群は、フィルター方法に従い選択されてもよい。フィルター方法には、例えば、ピアソンの相関、線形判別分析、分散分析（ＡＮＯＶＡ）、カイ二乗、それらの組み合わせなどが含まれる。フィルター方法に従った特徴の選択は、任意の機械学習アルゴリズムから独立している。代わりに、特徴は、転帰変数（例えば、はい／いいえ）との相関について、様々な統計検定におけるスコアに基づいて選択され得る。

【００８０】
別の例として、一つまたは複数の特徴群は、ラッパー方法により選択されてもよい。ラッパー方法は、特徴のサブセットを使用し、特徴のサブセットを使用して機械学習モデルを訓練するように構成され得る。以前のモデルから引き出された推論に基づいて、特徴は、サブセットから追加および／または削除され得る。ラッパー方法は、例えば、前方特徴量選択、後方特徴量削減、再帰的特徴量削減、それらの組み合わせなどを含む。一例として、前方特徴選択を使用して、一つまたは複数の特徴群を識別してもよい。前方特徴量選択は、機械学習モデルにおける特徴なしに始まる反復方法である。各反復において、モデルを最良に改善する特徴が、新たな変数の追加によって機械学習モデルの性能が改善されなくなるまで加えられる。一例として、後方排除を使用して、一つまたは複数の特徴群を識別してもよい。後方削減は、機械学習モデルにおける全ての特徴で始まる反復方法である。各反復では、最下位の特徴が、特徴の除去時に改善が観察されなくなるまで除去される。再帰的特徴除去を使用して、一つまたは複数の特徴群を識別してもよい。再帰的特徴量削減は、性能が最良である特徴サブセットを見出すことを目指す貪欲最適化アルゴリズムである。再帰的特徴量削減によって、モデルが反復的に作成され、各反復で最良または最悪の性能の特徴を別にしておく。再帰的特徴量削減によって、全ての特徴が消耗するまで、特徴が残っている次のモデルが構築される。再帰的特徴量削減によって、次に、それらの削減の順序に基づいて特徴がランク付けされる。

【００８１】
さらなる例として、一つまたは複数の特徴群は、埋め込み方法により選択されてもよい。埋め込み方法によって、フィルター方法とラッパー方法の質が組み合わされる。埋め込み方法には、例えば、過学習を低下させるためのペナルティ機能を実施する、最小絶対収縮および選択演算子（ＬＡＳＳＯ）およびリッジ回帰が含まれる。例えば、ＬＡＳＳＯ回帰によって、係数の大きさの絶対値に相当するペナルティを加えるＬ１正則化が実施され、リッジ回帰によって、係数の大きさの二乗に相当するペナルティを加えるＬ２正則化が実施される。

【００８２】
訓練モジュール９２０によって特徴セットが生成された後、訓練モジュール９２０によって、特徴セットに基づいて、機械学習ベースの分類モデル９４０が生成され得る。機械学習ベースの分類モデルは、機械学習技術を使用して生成される、データ分類のための複雑な数学的モデルを指し得る。一例では、機械学習ベースの分類モデル９４０は、境界特徴を表すサポートベクトルのマップを含み得る。この例では、境界特徴は、ある特徴セット内の最高ランクの特徴から選択されても、かつ／またはそれらを表してもよい。

【００８３】
訓練モジュール９２０は、それぞれの分類カテゴリー（例えば、はい、いいえ）につい

ての機械学習ベースの分類モデル９４０Ａ～９４０Ｎを構築するための訓練データセット９１０から決定または抽出された特徴セットを使用してもよい。いくつかの例では、機械学習ベースの分類モデル９４０Ａ～９４０Ｎを、単一の機械学習ベースの分類モデル９４０に組み合わせてもよい。同様に、ＭＬモジュール９３０は、単一もしくは複数の機械学習ベースの分類モデル９４０を含有する単一の分類指標、および／または単一もしくは複数の機械学習ベースの分類モデル９４０を含有する複数の分類指標を表し得る。

【００８４】

特徴を、機械学習アプローチ、例えば判別分析；決定木；最近傍（ＮＮ）アルゴリズム（例えば、ｋ－ＮＮモデル、レプリケーターＮＮモデルなど）；統計アルゴリズム（例えば、ベイジアンネットワークなど）；クラスタリングアルゴリズム（例えば、ｋ平均値、平均値シフトなど）；ニューラルネットワーク（例えば、リザーバネットワーク、人工ニューラルネットワークなど）；サポートベクター機械（ＳＶＭ）；ロジスティック回帰アルゴリズム；線形回帰アルゴリズム；マルコフモデルまたはチェーン；主成分分析（ＰＣＡ）（例えば、線形モデルについて）；多層パーセプトロン（ＭＬＰ）ＡＮＮ（例えば、非線形モデルについて）；リザーバネットワークの複製（例えば、非線形モデルについて、通常は時系列について）；ランダムフォレスト分類；それらの組み合わせおよび／または同様のものを使用して訓練された分類モデルにおいて組み合わせてもよい。得られたＭＬモジュール９３０は、プロモーター状態を新規の配列に割り当てるための、それぞれの特徴についての決定ルールまたはマッピングを含んでもよい。

【００８５】

一実施形態では、訓練モジュール９２０は、畳み込みニューラルネットワーク（ＣＮＮ）として機械学習ベースの分類モデル９４０を訓練してもよい。ＣＮＮは、少なくとも一つの畳み込み特徴層および最終の分類層（ｓｏｆｔｍａｘ）につながる三つの完全に連結した層を含む。最終の分類層を最終的に適用して、当該技術分野で公知のｓｏｆｔｍａｘ関数を使用して、完全に結び付けられた層の出力を組み合わせてもよい。

【００８６】

特徴およびＭＬモジュール９３０は、試験データセット内の配列のプロモーター状態を予測するために使用され得る。一例では、それぞれの配列の予測結果は、配列がプロモーターである可能性または確率に対応する信頼レベルを含む。信頼レベルは、ゼロから一の間の値であってもよく、それは、配列が、はい／いいえのプロモーター状態に属する可能性を表してもよい。一例では、二つの状態（例えば、はいおよびいいえ）があるとき、信頼レベルは、値ｐに対応してもよく、それは、特定の配列が、第一の状態（例えば、はい）に属する可能性を指す。この場合では、値１－ｐは、特定の配列が、第二の状態（例えば、いいえ）に属する可能性を指し得る。一般に、複数の信頼レベルは、試験データセットの各配列について、および三つ以上の状態がある場合、各特徴について提供され得る。最も高性能の特徴は、各試験配列について得られた結果を、各試験配列についての公知のはい／いいえのプロモーター状態と比較することによって決定されてもよい。一般に、最も高性能の特徴は、公知のはい／いいえプロモーター状態と密接に一致する結果を有するであろう。最も高性能の特徴を使用して、配列のはい／いいえプロモーター状態を予測してもよい。例えば、新規の配列が、決定／受信されてもよい。新規の配列は、最も高性能の特徴に基づき、新規の配列を、プロモーター（はい）またはプロモーターではない（いいえ）のいずれかとして分類し得るＭＬモジュール９３０に適用されてもよい。

【００８７】

図１１は、訓練モジュール９２０を使用して、ＭＬモジュール９３０を生成するための例となる訓練方法１１００を説明するフローチャートである。訓練モジュール９２０によって、教師あり、教師なし、および／または半教師あり（例えば、補強ベース）の機械学習ベースの分類モデル９４０を実施することができる。図１１に例証する方法１１００は、教師あり学習方法の例であり；訓練方法のこの例の変形を以下で考察するが、しかし、他の訓練方法は、教師なしおよび／または半教師ありの機械学習モデルを訓練するために類似的に実施することができる。

【００８８】
訓練方法１１００は、工程１１１０において第一の配列データを決定（例えば、アクセス、受信、検索など）してもよい。配列データは、コアプロモーター配列の標識されたセットおよび制御配列の標識されたセットを含んでもよい。標識は、プロモーター状態（例えば、はいまたはいいえ）に対応してもよい。
【００８９】
訓練方法１１００は、工程１１２０において、訓練データセットおよび試験データセットを生成してもよい。訓練データセットおよび試験データセットは、標識された配列を訓練データセットまたは試験データセットのいずれかに無作為に割り当てることによって、生成されてもよい。一部の実施では、訓練または試験データとしての標識された配列の割り当ては、完全に無作為でなくてもよい。一例として、標識された配列の大部分を使用して、訓練データセットを生成してもよい。例えば、標識された配列の７５％を使用して、訓練データセットを生成してもよく、２５％を使用して、試験データセットを生成してもよい。別の例では、標識された受容体配列の８０％を使用して、訓練データセットを生成してもよく、２０％を使用して、試験データセットを生成してもよい。
【００９０】
訓練方法１１００は、工程１１３０において、例えば、プロモーター状態（例えば、はい対いいえ）の異なる分類の中で区別するための分類指標によって使用することができる一つまたは複数の特徴を決定（例えば、抽出、選択など）してもよい。一例として、訓練方法１１００は、標識された配列からセットの特徴を決定してもよい。さらなる例では、特徴のセットは、訓練データセットまたは試験データセットのいずれかにおいて標識された配列以外の標識された配列から決定されてもよい。言い換えると、標識された配列は、機械学習モデルの訓練のためよりむしろ、特徴の決定のため使用され得る。このような標識された配列を使用して、特徴の初期のセットを決定してもよく、それは、訓練データセットを使用してさらに低減されてもよい。例として、本明細書に記載される特徴は、一つまたは複数の配列パターン、ＧＣ含量、ＣｐＧ含量、公知のコアプロモーター配列モチーフ、ＡＴＧ頻度、および／または相対エントロピーを含んでもよい。公知のコアプロモーター配列モチーフの発生については、ＴＳＳに対する相対位置決め方式も考慮され得る。相対エントロピーは、固定ヌクレオチド分布に基づき、ある配列がランダム配列とどれほど類似しているかの尺度である（コアプロモーター配列はランダムではない）。ランダムＤＮＡの確率の例は、ランダムＤＮＡの確率：Ａ：０．３、Ｃ：０．２、Ｔ：０．２、Ｇ：０．３である。図１０は、プロモーター状態を予測する様々な特徴の相対的有意性を示す。例えば、ＣｐＧジヌクレオチド含量は、コアプロモーターの同一性と最も強い正の相関を有する。ランダムＤＮＡは、比較的少数のＣｐＧを有する一方で、コアプロモーターは、多数を有することができる（コアプロモーターのサブセットは、ＣｐＧアイランドと呼ばれる）。ＴＡＴＡボックスの有無も重要である。
【００９１】
図１１に戻ると、訓練方法１１００は、工程１１４０で、一つまたは複数の特徴を使用して、一つまたは複数の機械学習モデルを訓練し得る。一例では、機械学習モデルは、教師あり学習を使用して訓練され得る。別の例では、教師なし学習および半教師ありを含む、他の機械学習技術が用いられてもよい。１１４０で訓練された機械学習モデルは、解決される問題および／または訓練データセットで利用可能なデータに応じて、異なる基準に基づいて選択され得る。例えば、機械学習分類器は、異なる程度のバイアスを受け得る。したがって、二つ以上の機械学習モデルを、１１４０で訓練し、工程１１５０で最適化し、改善し、相互検証することができる。
【００９２】
訓練方法１１００は、１１６０で予測モデルを構築するために、一つまたは複数の機械学習モデルを選択し得る。予測モデルは、試験データセットを使用して評価してもよい。予測モデルは、試験データセットを分析し、工程１１７０において予測されるプロモーター状態を生成してもよい。予測されるプロモーター状態を、工程１１８０において評価し

て、こうした値が、所望の精度レベルを達成したかどうかを決定することができる。予測モデルの性能は、予測モデルによって示される複数のデータ点の多数の真の陽性、偽陽性、真の陰性、および／または偽陰性の分類に基づいて、多数の方法で評価され得る。

【００９３】

例えば、予測モデルの偽陽性は、予測モデルが、実際にはプロモーターではない配列を、誤ってプロモーターとして分類した回数を指し得る。逆に、予測モデルの偽陰性は、機械学習モデルが、実際には配列がプロモーターであるのに、プロモーター配列をプロモーターではないと分類した回数を指し得る。真陰性および真陽性は、予測モデルによって一つまたは複数の配列が、プロモーター、またはプロモーターではないとして正しく分類された回数を指し得る。これらの測定に関連するのは、想起および精度の概念である。一般に、想起とは、真陽性および偽陰性の合計に対する真陽性の比率を指し、それによって予測モデルの感度が定量化される。同様に、精度は、真の陽性と偽陽性との合計の正陽性の比を指す。このような所望の精度レベルに達すると、訓練期が終了し、予測モデル（例えば、ＭＬモジュール９３０）が、工程１１９０において出力されてもよく、しかしながら、所望の精度レベルに達していないとき、訓練方法１１００のその後の反復は、例えば、配列データのより大きな収集を考慮するなどの変動を伴って、工程１１１０において開始して行われてもよい。

【００９４】

２．予測モデルを使用する方法

図１２は、機械学習ベースの分類器を使用して、ヌクレオチド配列がプロモーターであるかどうかを決定するための例示的なプロセスフローの図である。図１２に図示するように、非分類配列１２１０は、ＭＬモジュール９３０への入力として提供されてもよい。ＭＬモジュール９３０は、分類結果１２２０に到達するために、機械学習ベースの分類器を使用して、未分類の配列１２１０を処理してもよい。

【００９５】

分類結果１２２０は、非分類配列１２１０の一つまたは複数の特徴を識別し得る。例えば、分類結果１２２０は、非分類配列１２１０のプロモーター状態を識別してもよい（例えば、非分類配列１２１０がプロモーター機能を実行する可能性が高いかどうか）。

【００９６】

ＭＬモジュール９３０を使用して、生成モデル（例えば、ＬＳＴＭ－ＲＮＮ７１０）によって生成された配列を分類してもよい。予測モデル（例えば、ＭＬモジュール９３０）は、生成モデル（例えば、ＬＳＴＭ－ＲＮＮ７１０）の品質管理機構として機能し得る。生成モデルによって生成された配列を実験設定で試験する前に、予測モデルを使用して、生成された配列がコアプロモーター活性に対して陽性であると予測されるかどうかを試験してもよい。

【００９７】

Ｆ．使用方法

一部の態様では、特定のプロモーター配列は、本明細書に記載される方法のうちの一つまたは複数によって生成および／または識別される。プロモーター配列が識別されると、プロモーターを産生または操作することができる。一部の態様では、産生された、操作された、および合成されたという用語は、互換的に使用され得る。一部の態様では、プロモーターは、共通使用の技術に従って化学的に合成することができる。例えば、Ｂｅａｕｃａｇｅ ｅｔ ａｌ．（１９８１）Ｔｅｔ．Ｌｅｔｔ．２２：１８５９、および米国特許第４，６６８，７７７号を参照のこと。その全体が参照により本明細書に取り込まれる。かかる化学オリゴヌクレオチド合成は、Ｐｅｒｋｉｎ－Ｅｌｍｅｒ Ｃｏｒｐ．，Ｆｏｓｔｅｒ Ｃｉｔｙ，Ｃａｌｉｆ．，ＵＳＡの一部門であるＡｐｐｌｉｅｄ ＢｉｏｓｙｓｔｅｍｓによるＢｉｏｓｅａｒｃｈ ４６００または８６００ ＤＮＡ合成装置、およびＰｅｒｃｅｐｔｉｖｅ Ｂｉｏｓｙｓｔｅｍｓ，Ｆｒａｍｉｎｇｈａｍ，Ｍａｓｓ．，ＵＳＡによるＥｘｐｅｄｉｔｅなどの市販の装置を使用して実施することができる。プロモーターはまた、Ｙｕ ｅｔ ａｌ．Ｒｅｃｅｎｔ Ｐａｔ ＤＮＡ Ｇｅｎｅ Ｓｅｑ，２０１２，

Ａｐｒ；６（１）：１０－２１に開示された特許に記載された技術のいずれかを使用して合成することができる。各文献は参照によりその全体が組み込まれる。

【００９８】

プロモーターが産生されると、プロモーターは核酸構築物に挿入され得る。一部の態様では、核酸構築物は、ウイルスベクターを産生するために使用されるプラスミドを含むがこれに限定されないプラスミドであってもよい。一部の態様では、プロモーターは、ウイルスの作製に使用されるプラスミドに既に存在する導入遺伝子（すなわち、対象遺伝子）の上流に挿入され得る。一部の態様では、プロモーター配列は、核酸配列を形成する導入遺伝子の上流に挿入されてもよく、次いで、核酸配列は、ウイルスの作製に使用されるプラスミドに挿入されてもよい。一部の態様では、プロモーターは、導入遺伝子を挿入する前にプラスミド内に挿入することができる。任意の公知のクローニング方法を使用して、プロモーターを含むプラスミドまたは核酸配列を作製することができる。例えば、一部の態様では、ＡＡＶなどのウイルスベクターの作製に使用されるプラスミドを、一つまたは複数の制限エンドヌクレアーゼで切断することができる。５’末端および３’末端に制限エンドヌクレアーゼ部位を含み、間に特定のプロモーターを有する核酸配列を、制限エンドヌクレアーゼ部位に特異的な制限エンドヌクレアーゼを用いて切断することができる。一部の態様では、５’末端および３’末端に制限エンドヌクレアーゼ部位を含み、間に特定のプロモーターを有する核酸配列は、プロモーターの下流の導入遺伝子をさらに含み、制限エンドヌクレアーゼ部位の間にもある。一部の態様では、プラスミドの切断に使用される制限エンドヌクレアーゼは、５’末端および３’末端に制限エンドヌクレアーゼ部位を含み、間に特定のプロモーターを有する核酸配列の切断に使用される制限エンドヌクレアーゼと同一である。同じ制限エンドヌクレアーゼで切断することで、プラスミドに付着末端、および５’末端および３’末端に制限エンドヌクレアーゼ部位を含み、間に特定のプロモーターを有する核酸配列が産生される。一部の態様では、特定のプロモーターを含む核酸配列は、制限エンドヌクレアーゼで切断されたプラスミドに既に付着している特定の制限エンドヌクレアーゼ部位を末端に既に含有するように化学的に合成することができるため、制限エンドヌクレアーゼで核酸配列を切断する工程を省くことができる。最終的に、核酸配列および類似の付着末端を有するプラスミドを互いに接触させ、導入遺伝子の上流に特定のプロモーターを含む環状プラスミドの形成を可能にする。次いで、環状プラスミドを使用して、公知のウイルス産生方法を使用して、ＡＡＶなどのウイルスを産生することができる。

【００９９】

Ｇ．実施例

以下の実施例は、本方法およびシステムを例証する。以下の実施例は、その限定を意図するものではない。

【０１００】

本明細書に開示される方法を使用してコアプロモーター配列を識別した後、コアプロモーター活性を、生物学的システムを使用して試験することができる。図１３は、プロモーターアッセイがどのように設計され得るかの一例の概要を示す概略図である。個々のコアプロモーター（ＣＰ）候補（対照または生成されたコアプロモーター）を試験するために、二つのレポーター構築物を設計した。レポーター構築物は、ＮａｎｏＬｕｃルシフェラーゼ（Ｎｌｕｃ、Ｐｒｏｍｅｇａ）のコード配列、ＳＶ４０後期ポリアデニル化シグナル、および肝特異的エンハンサー（Ｋｈｅｒａｄｐｏｕｒ　ｅｔ　ａｌ．，ｄｏｉ：ｈｔｔｐ：／／ｄｘ．ｄｏｉ．ｏｒｇ／１０．１１０１／ｇｒ．１４４８９９．１１２から）、またはエンハンサー省略のいずれかを含有し、ユニバーサルリバースプライマー結合部位が続く。さらに、ＰＣＲ中にコアプロモーター候補を追加するために、ＮａｎｏＬｕｃコード配列の上流にプライマー結合部位を導入した。活性化エンハンサーをレポーター遺伝子の下流に配置して、エンハンサー内で転写が開始されないようにすることができる。

【０１０１】

レポーター構築物を、二本鎖ＤＮＡサンプル（ｇＢｌｏｃｋ、ＩＤＴ）として注文し、

水中で１０ｎｇ／μＩに希釈した。ルシフェラーゼ活性をアッセイするために細胞内に導入できるレポーター構築物を生成するために、これら二つのテンプレートをＰＣＲのユニバーサルテンプレートとして使用し、その間、それぞれのコアプロモーター候補を５’オリゴヌクレオチドを介して導入した。そのために、コアプロモーター配列、続いてレポーター構築物の５’プライマー結合部位に対応する配列を含有するオリゴヌクレオチドを注文した。得られたｄｓＤＮＡ　ＰＣＲ産物は、コアプロモーター候補、Ｎｌｕｃ　ＣＤＳ、ＳＶ４０後期ポリＡ、およびエンハンサー（図１３下）、またはバックグラウンド対照としてのエンハンサーなし（図１３上）のいずれかからなる。この直鎖状ｄｓＤＮＡ産物は、ルシフェラーゼ活性を評価するために、選択した細胞株へのトランスフェクションに直接使用することができる。ＰＣＲについては、２５μＬの２ｘ　Ｑ５　ｈｏｔｓｔａｒｔ　ｍａｓｔｅｒｍｉｘ（ＮＥＢ）、２．５μＬのフォワードオリゴ、２．５μＬのリバースオリゴ、１ｎｇのｇＦＭ１５テンプレート２０μＬのH_2Oを使用した。この反応混合物を、以下のＰＣＲプログラムで増幅した：９８℃で３０秒、９８℃で１０秒、６８℃で３０秒、７２℃で１５秒、工程２へさらに２４回、７２℃で２分間、４℃で保持。最後に、ＰＣＲ反応を、Ａｍｐｕｒｅ　ＸＰビーズ（Ｂｅｃｋｍａｎ　Ｃｏｕｌｔｅｒ）を使用して、０．５５ビーズ対ＰＣＲ反応比で精製した。コアプロモーター候補をプラスミド上ではなく、ＰＣＲ産物のコンテキスト内で試験することにより、他の交絡配列が存在しないことを保証する。

【０１０２】

図１４は、生成されたコアプロモーター（表１に示す配列）を対照コアプロモーターと比較したインビトロプロモーターアッセイを示す。図１４に示すように、「ｎｏ－ＣＰ」は対照、「ＳｅｒｐｉｎＡ１」は内因性コアプロモーター、「ＳＣＰ１」は合成コアプロモーター、「ＧＣＰ７」、「ＧＣＰ１０」、「ＧＣＰ＿ＭＴＥ」、および「ＧＣＰ＿ＭＴＥ＿Ｖ２」は、開示された方法に従って生成されるシャープコアプロモーター、「ＧＣＰ１８」は、開示された方法に従って生成されるランダムコアプロモーターである。生成されたコアプロモーター（ＣＰ）および対照の倍率変化値（平均±９５％信頼区間、ＣＩ）を示す。すべてのデータポイントは、エンハンサーなしのノーコアプロモーターコントロールの平均値に対して正規化される。コアプロモーターはタイプ別にグループ化される：対照はコアプロモーターを含まない陰性対照である；内因性は肝臓発現ＳｅｒｐｉｎＡ１遺伝子のコアプロモーターである；合成はＳＣＰ１（Ｋａｄｏｎａｇａ　ｌａｂ）などの操作されたコアプロモーターである；ｇｅｎｅｒａｔｅｄ＿ｓｈａｒｐは、内因性シャープコアプロモーターで訓練されたモデルに基づいて生成されたコアプロモーターである；ｇｅｎｅｒａｔｅｄ＿ｒａｎｄｏｍは、内因性ランダム配列で訓練されたモデルに基づくコアプロモーターである（追加の陰性対照として機能する）。左の二つのパネルは、エンハンサーを含まないレポーター構築物に由来する倍率変化（ベースライン）を示し、右の二つのパネルは、エンハンサー（Ｈｕｈ－７の内因性肝エンハンサー、ＨＥＫ２９３－ＨＺの最小ＳＦＦＶエンハンサー）を含むレポーター構築物に由来する倍率変化を示す。上の二つのパネルは、Ｈｕｈ－７細胞から得られたデータであり、下の二つのパネルは、ＨＥＫ２９３細胞で行われた実験から得られたデータである。

【０１０３】

【表１】

表１：生成されたコアプロモーターの配列。

ＩＤ	配列	配列番号
ＳＥＲＰＩＮＡ１	ＣＧＴＴＧＣＣＣＣＴＣＴＧＧＡＴＣＣＡＣＴＧＣＴＴＡＡＡＴＡＣＧＧＡＣＧＡＧＧＡＣＡＧＧＧＣＣＣＴＧＴＣＴＣＣＴＣＡＧＣＴＴＣＡＧＧＣＡＣＣＡＣＣＡＣＴＧＡＣＣＴＧＧＧＡＣＡＧＴＧＡＡ	６
ＳＣＰ１	ＣＣＣＴＡＧＧＧＴＡＣＴＴＡＴＡＴＡＡＧＧＧＧＧＴＧＧＧＧＧＣＧＣＧＴＴＣＧＴＣＣＴＣＡＧＴＣＧＣＧＡＴＣＧＡＡＣＡＣＴＣＧＡＧＣＣＧＡＧＣＡＧＡＣＧＴＧＣＣＴＡＣＧＧＡＣＣＧＧ	７
ＧＣＰ７	ＡＧＣＴＧＡＡＡＣＣＡＡＣＴＣＴＴＧＡＧＣＡＡＴＡＴＡＡＡＡＧＣＴＧＣＴＧＣＣＣＧＧＧＡＣＣＣＡＧＣＧＣＡＧＡＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＧＣＧＣＡＧＣＧＣＴ	８
ＧＣＰ１０	ＣＧＣＣＧＴＧＧＣＴＡＴＡＡＡＡＧＣＡＣＴＧＣＡＣＡＣＣＣＣＧＣＣＡＡＣＣＣＡＡＡＣＣＣＣＧＧＣＡＡ	９
ＧＣＰ＿ＭＴＥ	ＧＧＧＡＡＣＴＧＧＴＡＴＡＡＡＡＧＧＧＣＣＧＧＣＧＣＴＧＧＴＴＡＣＣＣＡＧＴＣＣＴＴＧＧＣＧＣＣＣＣＣＴＣＧＡＧＣＣＧＡＧＣＡＧＡＣＧＴＧＴＣＴＡＧＴＡＧＡＴＣＴＣＡＣ	１０
ＧＣＰ＿ＭＴＥ＿Ｖ２	ＧＧＧＡＡＣＴＧＧＴＡＴＡＡＡＡＧＧＧＣＣＧＧＣＧＣＴＣＧＴＧＧＣＧＴＴＡＣＣＣＡＧＴＣＣＴＴＧＧＣＧＣＣＣＣＣＴＣＧＡＧＣＣＧＡＧＣＡＧＡＣＧＴＧＴＣＴＡＧＴＡＧＡＴＣＴＣＡＣ	１１
ＧＣＰ１８	ＧＴＣＴＣＴＣＣＡＧＴＴＧＧＡＴＣＡＧＧＴＡＧＡＴＡＡＣＴＴＴＴＴＴＧＡＡＡＣＡＴＴＴＴＣＴＴＡＴＴＧＧＧＡＡＧＡＴＣＴＧＧＧＴＴＣＣＡＴＴＣＴＧＣＴＣＴＣＴＧＧＧＡＴＴＧＣＡＧＧＴＧＴＧＡＧＣＣＡＣＡ	１２

【０１０４】

図１４に示すアッセイは、Ｈｕｈ－７細胞およびＨＥＫ２９３－ＨＺ細胞のトランスフェクション、続いてルシフェラーゼアッセイを伴う。１ウェル当たり1×10^4個のＨｕｈ－７細胞またはＨＥＫ２９３－ＨＺ細胞を、ＤＭＥＭ＋１０％ＦＢＳ中、９６ウェルプレートに播種し、２４時間後に、Ｍｉｒｕｓ　ＴｒａｎｓＩＴ－ＬＴ１トランスフェクション試薬（Ｍｉｒｕｓ　Ｂｉｏ、＃ＭＩＲ　２３０４）を使用して、０．１μｇのレポーター構築物でトランスフェクトした。トランスフェクション対照として、ホタルルシフェラーゼプラスミドを１：９の比率で共トランスフェクトした（ホタルプラスミド：ＮａｎｏＬｕｃ　ＰＣＲ産物）。ルシフェラーゼ活性をアッセイするために、細胞を溶解し、Ｎａｎｏ－Ｇｌｏデュアルルシフェラーゼアッセイシステム（Ｐｒｏｍｅｇａ社、＃Ｎ１６１０）を用いてトランスフェクションの２４時間後に処理した。ＳｐｅｃｔｒａＭａｘ　ｉ３プレートリーダー（Ｍｏｌｅｃｕｌａｒ　Ｄｅｖｉｃｅｓ）を使用して、ルシフェラーゼ活性を測定した。

【０１０５】

計算方法

訓練データ

ヒトゲノムの候補コアプロモーターのリストは、ＲパッケージＣＡＧＥｒ（ｄｏｉ：１０．１８１２９／Ｂ９．ＢＩＯＣ．ＣＡＧＥＲ）を使用してＦＡＮＴＯＭ５（ｄｏｉ：１０．１０３８／ｓｄａｔａ．２０１７．１１２）から転写開始点（ＴＳＳ）プロファイリングデータとしてダウンロードするか、またはＦＡＮＴＯＭ$_5$から直接ダウンロードした。このデータから、予測モデルおよび生成モデルに対して、コアプロモーターの二つの別個のリストを作成した（以下のモデルの説明を参照）。生成モデルについて、ＦＡＮＴＯ

Ｍ５データセットをフィルタリングして、ヒトの成人肝データ（サンプル：ｌｉｖｅｒ＿＿ａｄｕｌｔ＿＿ｐｏｏｌ１）のみを保持した。データを正規化し（方法＝「ｐｏｗｅｒＬａｗ」、ｆｉｔＩｎＲａｎｇｅ＝ｃ（５、１０００）、アルファ＝１．０５、Ｔ＝１＊１０＾６）、ＴＳＳをクラスタリングし、四分位幅を計算した（ｑＬｏｗ＝０．１、ｑＵｐ＝０．９）。その後、その四分位幅に基づいて、シャープＴＳＳおよびブロードＴＳＳにビン化された。次に、コアプロモーター候補を、ＴＳＳサミットを５’方向に４９ｂｐ、および３’方向に５０ｂｐ延長することによって作成した。ＣＡＧＥシグナルに基づいて、２，９５０の最強のコアプロモーターのみが維持された。対照セットとして、これらのコアプロモーターを５’方向に５０，０００ｂｐシフトさせ、シフト後に任意のＣＡＧＥピークと重複するものをフィルタリングして除去した（ｎ＿ｃｏｎｔｒｏｌ＝２９１５）。

【０１０６】

予測モデルでは、ＦＡＮＴＯＭ５データセット全体の上位ＣＡＧＥピーク（ｈｇ１９．ｃａｇｅ＿ｐｅａｋ＿ｐｈａｓｅ１ａｎｄ２ｃｏｍｂｉｎｅｄ＿ｃｏｏｒｄ．ｂｅｄの列５でカットオフ＞５０，０００）を取り、生成モデルの任意のピークと重複するものをフィルタリングして除外した。生成モデルのピークと同様に、ＴＳＳサミットを、５’方向に４９ｂｐ、３’方向に５０ｂｐ延長して、コアプロモーターの最終リストを作成した。これらのコアプロモーターを５’方向に５０，０００ｂｐシフトさせて、陰性対照領域のリストを作成し、これを生成モデルについて任意のＣＡＧＥピークまたは陰性対照領域に対してさらにフィルタリングした。上記のコアプロモーター配列の全てを、ヒトゲノムアセンブリ（ｈｇ１９）中のＮｓを含有する任意の配列に対してフィルタリングした。予測モデルでは、コアプロモーターの各カテゴリーを、標識に関連付けられた配列として保存した（１＝コアプロモーター、０＝陰性対照）。

【０１０７】

予測モデル

予測モデルを訓練するために、標識された配列を取り（「訓練データ」のセクションを参照）、コアプロモーター生物学に関連する特徴を抽出した。ＧＣ含量、ＡＴおよびＣＧジヌクレオチド頻度、ＡＴＧ頻度、コアプロモーターモチーフの発生、および相対エントロピーが計算された（それぞれ、Ａ、Ｃ、Ｇ、Ｔの頻度が、０．３、０．２、０．２、０．３のランダム配列と比較して）。モチーフの発生については、ＴＳＳに対する相対位置決め方式も考慮された。次に、訓練データセットを、ｓｃｉｋｉｔ－ｌｅａｒｎのｔｒａｉｎ＿ｔｅｓｔ＿ｓｐｌｉｔを使用して、訓練セットおよび検証セットに分割した（８０％訓練、２０％検証、ラベル階層化あり、Ｓｃｉｋｉｔ－ｌｅａｒｎ：Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ ｉｎ Ｐｙｔｈｏｎ、Ｐｅｄｒｅｇｏｓａ ｅｔ ａｌ．，ＪＭＬＲ １２，ｐｐ．２８２５－２８３０，２０１１，バージョン０．２１．２）。上記は、１９の特徴を有する１１，３３９例の訓練セットおよび２，８３５の検証例をもたらした。イエローブリック（ｄｏｉ：１０．５２８１／ｚｅｎｏｄｏ．３６８７３３０、バージョン０．９．１）に実装されたｆ１加重スコアを使用して、Ｌ１正則化を伴うロジスティック回帰モデルのハイパーパラメータを選択した：ｐｅｎａｌｔｙ＝「ｌ１」、ｓｏｌｖｅｒ＝「ｌｉｂｌｉｎｅａｒ」、ｍｕｌｔｉ＿ｃｌａｓｓ＝「ａｕｔｏ」、Ｃ＝０．５。ＲＯＣや特徴の重要性などのすべての指標は、イエローブリックを使用して可視化された。

【０１０８】

生成モデル

生成モデルを訓練するために、異なるクラスのコアプロモーター配列を、予測標的として以下のヌクレオチドを有する１０ｎｔシード配列の対に分割した。シード／標的対を生成するために、コアプロモーター配列を、工程サイズ１のスライディングウィンドウアプローチを使用して分割した。これにより、シャープＣＰについては９４，８６９対、ブロードＣＰについては１７０，６４０対、ランダムな対照については２６１，７２０対となった。これらの配列対を、単純な数値符号化（「Ａ」：０、「Ｃ」：１、「Ｇ」：２、「Ｔ」：３）を使用してベクター化した。

【０１０９】

ＴｅｎｓｏｒＦｌｏｗバックエンド（１．１３．０－ｄｅｖ２０１９０１２６）を備えたｋｅｒａｓ（２．２．４）を使用して、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を実装した。ＬＳＴＭは１２８ユニットで、その後、ソフトマックスアクティベーション関数を有する単一の高密度出力層が続いた。ＲＭＳｐｒｏｐ（ｌｒ＝０．００１）オプティマイザおよびカテゴリカルな交差エントロピーを、損失関数として使用した。このモデルは、それぞれの対（上記参照）を入力として使用し、バッチサイズ１２８で２５エポック訓練させたが、早期停止を採用した（損失監視、忍耐＝１、分＿デルタ＝０．００１）。新しい配列を生成するために、学習した確率分布からのサンプリングを使用して、「シード」配列の新しいヌクレオチドを予測することができる。次いで、新たに生成された配列を、配列生成の反復プロセスにおける配列生成の別のサイクルの別のシードとして使用することができる。このプロセスに確率を加えるために、最初に学習した確率を、ソフトマックス温度によって再計量してから、この新しく導出された分布からサンプリングした（温度＝０．８）。最後に、このアプローチを使用して、シャープでブロードかつランダムなコアプロモーター配列からの入力を使用して訓練されたモデルに基づいて、新規のコアプロモーター配列を生成した。これらの配列の長さは、５０～１００ｎｔの範囲であった。

【０１１０】

図１５は、ネットワーク１５０４を通じて接続された計算デバイス１５０１およびサーバ１５０２の非限定的な例を含む環境１５００を描写するブロック図である。一態様では、いずれの記載の方法のいくつかまたは全ての工程も、本明細書に記載の計算デバイスで実行することができる。コンピューティングデバイス１５０１は、配列データ１５２０（例えば、ＣＡＧＥデータなどのプロモーター配列データ）、訓練データ１５２２（例えば、標識された配列データ：コアプロモーター配列および制御配列）、生成モジュール１５２４（例えば、任意の補助的な訓練モジュールを含む、ＬＳＴＭ－ＲＮＮ７１０）、予測モジュール１５２６（例えば、任意の補助的訓練モジュールを含む、ＭＬモジュール９３０）などの一つまたは複数を保存するように構成された一つまたは複数のコンピュータを備え得る。サーバ１５０２は、配列データ１５２０を保存するように構成した一つまたは複数のコンピュータを含むことができる。複数のサーバ１５０２は、ネットワーク１５０４を通じて計算デバイス１５０１と通信することができる。一実施形態では、サーバ１５０２は、ＣＡＧＥ実験によって生成されたデータのためのリポジトリを備えてもよい。

【０１１１】

計算デバイス１５０１およびサーバ１５０２は、ハードウェアアーキテクチャに関して、一般にプロセッサ１５０８、メモリシステム１５１０、入力／出力（Ｉ／Ｏ）インターフェース１５１２、およびネットワークインターフェース１５１４を含む、デジタルコンピュータであってもよい。これらの構成要素（１５０８、１５１０、１５１２、および１５１４）は、ローカルインターフェース１５１６を介して通信的に連結される。ローカルインターフェース１５１６は、例えば、当該技術分野で公知の一つまたは複数のバスまたは他の有線もしくは無線接続であってもよいが、これに限定されない。ローカルインターフェース１５１６は、コントローラ、バッファ（キャッシュ）、ドライバ、リピータ、およびレシーバなどの、通信を可能にするための追加の要素（簡略化のために省略される）を有してもよい。さらに、ローカルインターフェースは、前述の構成要素間の適切な通信を可能にするためのアドレス、制御、および／またはデータ接続を含んでもよい。

【０１１２】

プロセッサ１５０８は、特にメモリシステム１５１０に保存される、ソフトウェアを実行するためのハードウェアデバイスであってもよい。プロセッサ１５０８は、任意のカスタム作製または市販のプロセッサ、中央処理ユニット（ＣＰＵ）、計算デバイス１５０１およびサーバ１５０２に関連付けられたいくつかのプロセッサの中の補助プロセッサ、半導体ベースのマイクロプロセッサ（マイクロチップもしくはチップセットの形態）、またはソフトウェア命令を実行するための一般に任意のデバイスとすることができる。計算デバイス１５０１および／またはサーバ１５０２が動作中である時、プロセッサ１５０８は

、メモリシステム１５１０内に保存されているソフトウェアを実行して、メモリシステム１５１０へのおよびそこからのデータを通信し、ソフトウェアに従って、計算デバイス１５０１およびサーバ１５０２の動作を一般に制御するように構成されてもよい。

【０１１３】

Ｉ／Ｏインターフェース１５１２を使用して、一つまたは複数のデバイスまたは構成要素からユーザ入力を受信する、かつ／またはそれらへとシステム出力を提供することができる。ユーザ入力は、例えば、キーボードおよび／またはマウスを介して提供されてもよい。システム出力は、表示デバイスおよびプリンタ（図示せず）を介して提供されてもよい。Ｉ／Ｏインターフェース１５１２は、例えば、シリアルポート、パラレルポート、小型コンピュータシステムインターフェース（ＳＣＳＩ）、赤外（ＩＲ）インターフェース、無線周波数（ＲＦ）インターフェース、および／またはユニバーサルシリアルバス（ＵＳＢ）インターフェースを含んでもよい。

【０１１４】

ネットワークインターフェース１５１４は、計算デバイス１５０１および／またはネットワーク１５０４上のサーバ１５０２から送信および受信するために使用することができる。ネットワークインターフェース１５１４は、例えば、１０ＢａｓｅＴ Ｅｔｈｅｒｎｅｔアダプタ、１００ＢａｓｅＴ Ｅｔｈｅｒｎｅｔアダプタ、ＬＡＮ ＰＨＹ Ｅｔｈｅｒｎｅｔアダプタ、Ｔｏｋｅｎ Ｒｉｎｇアダプタ、ワイヤレスネットワークアダプタ（例えば、ＷｉＦｉ、セルラー、サテライト）、または任意の他の好適なネットワークインターフェースデバイスを含んでもよい。ネットワークインターフェース１５１４は、ネットワーク１５０４上での適切な通信を可能にするためのアドレス、制御、および／またはデータ接続を含んでもよい。

【０１１５】

メモリシステム１５１０は、揮発性メモリ素子（例えば、ランダムアクセスメモリ（ＤＲＡＭ、ＳＲＡＭ、ＳＤＲＡＭなどのＲＡＭ））および不揮発性メモリ素子（例えば、ＲＯＭ、ハードドライブ、テープ、ＣＤＲＯＭ、ＤＶＤＲＯＭなど）のいずれか一つまたはその組み合わせを含んでもよい。さらに、メモリシステム１５１０は、電子、磁気、光学、および／または他の型の保存媒体を組み込んでもよい。メモリシステム１５１０は、様々な構成要素が互いに離れて位置するが、プロセッサ１５０８によってアクセスすることができる、分散型アーキテクチャを有し得ることに留意されたい。

【０１１６】

メモリシステム１５１０内のソフトウェアは、一つまたは複数のソフトウェアプログラムを含んでもよく、これらの各々は、論理機能を実施するための実行可能な命令の順序付けされたリストを含む。図１５の例では、コンピューティングデバイス１５０１のメモリシステム１５１０におけるソフトウェアは、配列データ１５２０、訓練データ１５２２、生成モジュール１５２４、予測モジュール１５２６、および適当な操作システム（Ｏ／Ｓ）１５１８を含むことができる。図１５の例では、サーバ１５０２のメモリシステム１５１０内のソフトウェアは、配列データ１５２０、および好適なオペレーティングシステム（Ｏ／Ｓ）１５１８を含むことができる。オペレーティングシステム１５１８は、他のコンピュータプログラムの実行を本質的に制御し、スケジューリング、入力－出力制御、ファイルおよびデータ管理、メモリ管理、および通信制御、ならびに関連するサービスを提供する。

【０１１７】

例証の目的で、アプリケーションプログラムおよびオペレーティングシステム１５１８などの他の実行可能なプログラム構成要素は、本明細書では別々のブロックとして例証されているが、そのようなプログラムおよび構成要素は、計算デバイス１５０１および／またはサーバ１５０２の異なる保存構成要素内で、様々な時間に存在し得ることが認識される。生成モジュール１５２４および／または予測モジュール１５２６の実施は、何らかの形態のコンピュータ可読媒体に保存されても、それを通過して伝達されてもよい。本開示の方法のいずれも、コンピュータ可読媒体上に具現化されたコンピュータ可読命令によっ

て実行することができる。コンピュータ可読媒体は、コンピュータによってアクセス可能な任意の利用可能媒体とすることができる。例として、かつ限定を意図するものではないが、コンピュータ可読媒体は、「コンピュータストレージ媒体」および「通信媒体」を含み得る。「コンピュータ記憶媒体」は、コンピュータ可読命令、データ構造、プログラムモジュール、または他のデータなどの、情報を記憶するための任意の方法または技術で実施される、揮発性および不揮発性の取り外し可能な媒体および取り外し不能な媒体を含み得る。例示的なコンピュータ記憶媒体は、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリもしくは他の記憶技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）もしくは他の光学記憶装置、磁気カセット、磁気テープ、磁気ディスク記憶デバイスもしくは他の磁気記憶デバイス、または所望の情報の記憶に使用することができ、かつコンピュータによってアクセスすることができる任意の他の媒体を含み得る。

【０１１８】

一実施形態では、予測モジュール１５２６は、図１６に示す方法１６００を実行するように構成され得る。方法１６００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法１６００は、遺伝的データを受信することを含み、遺伝的データは、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列は、１６０１で関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む。関連発現スコアは、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含み得る。

【０１１９】

方法１６００は、閾値を満たす関連発現スコアに基づいて、１６０２で第一の複数のヌクレオチド配列から複数のＴＳＳを決定することを含み得る。

方法１６００は、複数のＴＳＳに基づいて、１６０３で複数のサミットヌクレオチド塩基を決定することを含み得る。複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することは、複数のＴＳＳの各々について、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含み得る。

【０１２０】

方法１６００は、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、１６０４で関連する複数の周辺塩基を決定することを含み得る。複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含み得る。５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み得、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む。

【０１２１】

方法１６００は、１６０５でコアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することを含み得る。

【０１２２】

方法１６００は、第二の複数のヌクレオチド配列の各ヌクレオチド配列について、１６０６で関連する複数のシフト塩基を決定することを含み得る。第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することは、第二の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含み得る。

【０１２３】

方法１６００は、１６０７でコアプロモーターではないとして標識された第三の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することを含み得る。

方法１６００は、コアプロモーターとして標識された第二の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づい

て、１６０８で訓練データセットを生成することを含み得る。
【０１２４】
方法１６００は、訓練データセットに基づいて、１６０９で予測モデルについての複数の特徴を決定することを含み得る。予測モデルについての複数の特徴は、ＧＣ含量、ＡＴおよびＣＧジヌクレオチド頻度、ＡＴＧ頻度、コアプロモーターモチーフの発生、相対エントロピー、および関連するＴＳＳに対する相対的位置決め方式のうちの一つまたは複数を含んでもよい。
【０１２５】
１６１０では、方法１６００は、訓練データセットの第一の部分に基づいて、１６１０で複数の特徴に従って予測モデルを訓練することを含んでもよい。
方法１６００は、訓練データセットの第二の部分に基づいて、１６１１で予測モデルを試験することを含み得る。
【０１２６】
方法１６００は、試験に基づいて、１６１２で予測モデルを出力することを含み得る。
一実施形態では、方法１６００は、生成モデルで使用されるＴＳＳの発現スコアと重複する発現スコアを有する第一の複数のヌクレオチド配列から、複数のＴＳＳの任意のＴＳＳをフィルタリングすることをさらに含むことができる。一実施形態では、方法１６００は、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含むことができる。
【０１２７】
一実施形態では、予測モジュール１５２６は、図１７に示す方法１７００を実行するように構成され得る。方法１７００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法１７００は、遺伝的データを受信することを含み、遺伝的データは、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列は、１７０１で関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む。関連発現スコアは、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含んでもよい。
【０１２８】
方法１７００は、第一の複数のヌクレオチド配列に基づいて、１７０２でコアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することを含み得る。第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することは、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することと、各サミットヌクレオチド塩基および関連する複数の周辺塩基を、コアプロモーターとして標識された第二の複数のヌクレオチド配列として保存することと、を含み得る。
【０１２９】
複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することは、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む場合がある。複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含み得る。５’方向の第一の複数のヌクレオチド塩基は、４９個のヌクレオチド塩基を含み得、３’方向の第二の複数のヌクレオチド塩基は、５０個のヌクレオチド塩基を含む。
【０１３０】
方法１７００は、第二の複数のヌクレオチド配列に基づいて、１７０３でコアプロモーターではないとして標識された第三の複数のヌクレオチド配列を決定することを含み得る。第二の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識され

た第三の複数のヌクレオチド配列を決定することは、第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することと、各関連する複数のシフト塩基を、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として保存することと、を含み得る。

【０１３１】

第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することは、第二の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含み得る。

【０１３２】

方法１７００は、コアプロモーターとして標識された第二の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、１７０４で訓練データセットを生成することを含み得る。

【０１３３】

方法１７００は、訓練データセットに基づいて、１７０５で予測モデルについての複数の特徴を決定することを含み得る。予測モデルについての複数の特徴は、ＧＣ含量、ＡＴおよびＣＧジヌクレオチド頻度、ＡＴＧ頻度、コアプロモーターモチーフの発生、相対エントロピー、および関連するＴＳＳに対する相対的位置決め方式のうちの一つまたは複数を含んでもよい。

【０１３４】

方法１７００は、訓練データセットの第一の部分に基づいて、１７０６で複数の特徴に従って予測モデルを訓練することを含み得る。

方法１７００は、訓練データセットの第二の部分に基づいて、１７０７で予測モデルを試験することを含み得る。

【０１３５】

方法１７００は、試験に基づいて、１７０８で予測モデルを出力することを含み得る。

一実施形態では、クラム１７００の方法はまた、生成モデルで使用されるＴＳＳの発現スコアと重複する発現スコアを有する第一の複数のヌクレオチド配列から、複数のＴＳＳの任意のＴＳＳをフィルタリングすることを含んでもよい。一実施形態では、方法１７００は、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含むことができる。

【０１３６】

一実施形態では、生成モジュール１５２４は、図１８に示す方法１８００を実施するように構成され得る。方法１８００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法１８００は、遺伝的データを受信することを含み、遺伝的データは、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列は、１８０１で関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む。関連発現スコアは、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含んでもよい。

【０１３７】

方法１８００は、１８０２で遺伝的データを正規化することを含み得る。

方法１８００は、関連発現スコアに基づいて、１８０３でＴＳＳをクラスタリングすることを含み得る。

【０１３８】

方法１８００は、ＴＳＳの各クラスターについて、１８０４で四分位幅を決定することを含み得る。

方法１８００は、四分位幅に基づいて、１８０５で各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳとして標識することを含み得る。

【０１３９】

方法１８００は、複数のＴＳＳに基づいて、１８０６で複数のサミットヌクレオチド塩基を決定することを含み得る。複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩

基を決定することは、複数のＴＳＳの各々について、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含み得る。

【０１４０】

方法１８００は、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、１８０７で関連する複数の周辺塩基を決定することを含み得る。複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含み得る。５’方向の第一の複数のヌクレオチド塩基は、４９個のヌクレオチド塩基を含み、３’方向の第二の複数のヌクレオチド塩基は、５０個のヌクレオチド塩基を含む。

【０１４１】

方法１８００は、１８０８でコアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することを含み得る。

【０１４２】

方法１８００は、閾値を満たす関連発現スコアに基づいて、１８０９で第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することを含み得る。

方法１８００は、第三の複数のヌクレオチド配列の各ヌクレオチド配列について、１８１０で関連する複数のシフト塩基を決定することを含み得る。第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することは、第三の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含み得る。

【０１４３】

方法１８００は、１８１１でコアプロモーターではないとして標識された第四の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することを含み得る。

方法１８００は、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、１８１２で訓練データセットを生成することを含み得る。

【０１４４】

方法１８００は、訓練データセットの各ヌクレオチド配列について、１８１３で複数のシード配列および標的ヌクレオチド対を生成することを含み得る。訓練データセット中の各ヌクレオチド配列について、複数のシード配列および標的ヌクレオチド対を生成することは、シャープＴＳＳまたはブロードＴＳＳ標識に基づいて、訓練データセット中のヌクレオチド配列をシャープＴＳＳ群またはブロードＴＳＳ群に分割することと、定義された長さのスライディングウィンドウを適用し、定義された工程サイズを各ヌクレオチド配列に有することと、スライディングウィンドウの各工程でシード配列および標的ヌクレオチド対を保存することと、を含み得る。

【０１４５】

方法１８００は、１８１４で、複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクター化することを含み得る。各シード配列および標的ヌクレオチド対は、定義された長さを有するシード配列、および所定のヌクレオチド配列上のシード配列の直後に標的ヌクレオチドを含み得る。定義された長さは、例えば、１０塩基であってもよい。複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクター化することは、各ヌクレオチドをそれぞれの番号としてコード化することを含み得る。

【０１４６】

方法１８００は、ベクター化されたシード配列および標的ヌクレオチド対に基づいて、１８１５で生成モデルを訓練することを含み得る。生成モデルは、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を含んでもよい。

【０１４７】
方法１８００は、１８１６で生成モデルを出力することを含んでもよい。
一実施形態では、方法１８００は、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることを含み得る。
【０１４８】
一実施形態では、方法１８００は、生成モデルに基づいて、ヌクレオチド配列を生成することを含み得る。ヌクレオチド配列は、例えば、コアプロモーター配列であってもよい。方法１８００は、コアプロモーター配列に基づいてプロモーターを操作することを含み得る。
【０１４９】
生成モデルに基づいて、ヌクレオチド配列を生成することは、（ａ）シード配列を受信することと、（ｂ）シード配列に基づいて、次のヌクレオチドを予測することと、（ｃ）シード配列に次のヌクレオチドを付加することと、（ｄ）ヌクレオチド配列の所望の長さに達するまでｂ～ｃを繰り返すことと、を含み得る。所望の長さは、例えば、約５０ヌクレオチド～約１００ヌクレオチドであってもよい。
【０１５０】
一実施形態では、方法１８００は、プロモーターを核酸構築物に挿入することを含み得る。プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含み得る。
【０１５１】
一実施形態では、方法１８００は、核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することを含んでもよい。
一実施形態では、生成モジュール１５２４は、図１９に示す方法１９００を実施するように構成され得る。方法１９００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法１９００は、遺伝的データを受信することを含み、遺伝的データは、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列は、１９０１で関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む。関連発現スコアは、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含んでもよい。
【０１５２】
方法１９００は、第一の複数のヌクレオチド配列に基づいて、１９０２でコアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することを含み得る。第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することは、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することと、各サミットヌクレオチド塩基および関連する複数の周辺塩基を、コアプロモーターとして標識された第二の複数のヌクレオチド配列として保存することと、を含み得る。
【０１５３】
複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することは、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む。複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することは、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含み得る。５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み得、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む。
【０１５４】
方法１９００は、閾値を満たす関連発現スコアに基づいて、１９０３で第二の複数のヌ

クレオチド配列から第三の複数のヌクレオチド配列を決定することを含み得る。
方法１９００は、第三の複数のヌクレオチド配列に基づいて、１９０４でコアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することを含み得る。第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することは、第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することと、各関連する複数のシフト塩基を、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として保存することと、を含み得る。
【０１５５】
第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することは、第三の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含み得る。
【０１５６】
方法１９００は、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、１９０５で訓練データセットを生成することを含み得る。
【０１５７】
方法１９００は、訓練データセットに基づいて、１９０６で生成モデルを訓練することを含んでもよい。生成モデルは、例えば、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を含んでもよい。訓練データセットに基づいて、生成モデルを訓練することは、訓練データセット中の各ヌクレオチド配列について、複数のシード配列と標的ヌクレオチド対を生成することと、複数のシード配列と標的ヌクレオチド対の各シード配列と標的ヌクレオチド対をベクター化することと、ベクター化されたシード配列と標的ヌクレオチド対に基づいて、生成モデルを訓練することと、を含み得る。
【０１５８】
各シード配列および標的ヌクレオチド対は、定義された長さを有するシード配列、および所定のヌクレオチド配列上のシード配列の直後に標的ヌクレオチドを含み得る。定義された長さは、例えば、１０塩基であってもよい。複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクター化することは、各ヌクレオチドをそれぞれの番号としてコード化することを含み得る。
【０１５９】
方法１９００は、１９０７で生成モデルを出力することを含んでもよい。
一実施形態では、方法１９００は、遺伝的データを正規化することを含み得る。一実施形態では、方法１９００は、関連発現スコアに基づいて、ＴＳＳをクラスタリングすることと、ＴＳＳの各クラスターについて、四分位幅を決定することと、四分位幅に基づいて、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳとして標識することと、を含み得る。
【０１６０】
一実施形態では、方法１９００は、生成モデルに基づいて、ヌクレオチド配列を生成することを含み得る。ヌクレオチド配列は、例えば、コアプロモーター配列であってもよい。
【０１６１】
一実施形態では、方法１９００は、コアプロモーター配列に基づいてプロモーターを操作することを含み得る。方法１９００は、プロモーターを核酸構築物に挿入することを含み得る。プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含み得る。
【０１６２】
一実施形態では、方法１９００は、核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することを含んでもよい。
一実施形態では、方法１９００は、ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることを含み得る。

【０１６３】
一実施形態では、方法１９００は、生成モデルに基づいて、ヌクレオチド配列を生成することを含み得る。ヌクレオチド配列は、例えば、コアプロモーター配列であってもよい。方法１９００は、コアプロモーター配列に基づいてプロモーターを操作することを含み得る。
【０１６４】
生成モデルに基づいて、ヌクレオチド配列を生成することは、（ａ）シード配列を受信することと、（ｂ）シード配列に基づいて、次のヌクレオチドを予測することと、（ｃ）シード配列に次のヌクレオチドを付加することと、（ｄ）ヌクレオチド配列の所望の長さに達するまでｂ～ｃを繰り返すことと、を含み得る。所望の長さは、例えば、約５０ヌクレオチド～約１００ヌクレオチドであってもよい。
【０１６５】
一実施形態では、方法１９００は、プロモーターを核酸構築物に挿入することを含み得る。プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含み得る。
【０１６６】
一実施形態では、方法１９００は、核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することを含んでもよい。
一実施形態では、予測モジュール１５２６は、図２０に示す方法２０００を実行するように構成され得る。方法２０００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法２０００は、２０１０でヌクレオチド配列を受信することを含んでもよい。ヌクレオチド配列を受信することは、複数のヌクレオチド配列を受信することを含み得、複数のヌクレオチド配列は、生成モデルによって生成された。
【０１６７】
方法２０００は、訓練された予測モデルに、２０２０でヌクレオチド配列を提供することを含み得る。
方法２０００は、予測モデルに基づいて、２０３０でヌクレオチド配列がコアプロモーターであることを決定することを含み得る。
【０１６８】
一実施形態では、方法２０００は、ヌクレオチド配列がコアプロモーターであるという決定に基づいて、一つまたは複数の基準に従ってヌクレオチド配列をフィルタリングすることを含み得る。一つまたは複数の基準は、例えば、ＧＣ含量またはモチーフのうちの一つまたは複数を含んでもよい。
【０１６９】
一実施形態では、方法２０００は、遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、第二の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練す

ることと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、を含み得る。

【０１７０】

一実施形態では、生成モジュール１５２４は、図２１に示す方法２１００を実施するように構成され得る。方法２１００は、単一の計算デバイス、複数の電子デバイス、および同様のものによって、全体的または部分的に実施されてもよい。方法２１００は、２１１０でヌクレオチド配列および配列長を受信することを含んでもよい。

【０１７１】

方法２１００は、訓練された生成モデルに、２１２０でヌクレオチド配列を提供することを含み得る。

方法２１００は、生成モデルに基づいて、２１３０でヌクレオチド配列と関連付けられた次のヌクレオチドを決定することを含み得る。

【０１７２】

方法２１００は、２１４０で、次のヌクレオチドをヌクレオチド配列に付加することを含んでもよい。

方法２１００は、２１５０でヌクレオチド配列の長さが配列長と等しくなるまで、工程２１２０～２１４０を繰り返すことを含み得る。配列長は、例えば、約５０ヌクレオチド～約１００ヌクレオチドであってもよい。

【０１７３】

方法２１００は、２１６０でコアプロモーター配列としてヌクレオチド配列を出力することを含み得る。

一実施形態では、方法２１００は、コアプロモーター配列に基づいてプロモーターを操作することを含み得る。

【０１７４】

一実施形態では、方法２１００は、プロモーターを核酸構築物に挿入することを含み得る。プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含み得る。

【０１７５】

一実施形態では、方法２１００は、核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することを含んでもよい。

一実施形態では、方法２１００は、遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、および訓練データセットに基づいて、生成モデルを訓練することと、を含み得る。

【０１７６】

一実施形態では、方法２１００は、一つまたは複数の基準に従ってヌクレオチド配列をフィルタリングすることを含み得る。一つまたは複数の基準は、ＧＣ含量またはモチーフのうちの一つまたは複数を含んでもよい。

【０１７７】

記載された方法、システム、および装置、ならびにそれらの変形に照らして、本明細書では、以下に本発明のより具体的に記述された特定の実施形態を説明する。しかし、これらの特に列挙された実施形態は、本明細書に記載される異なるまたはより一般的な教示を含む任意の異なる特許請求の範囲に対して何らかの限定効果を有すると解釈されるべきで

はなく、または「特定の」実施形態が、その中に文字通り使用される言語の固有の意味以外の何らかの方法で、何らかの形で限定されると解釈されるべきでもない。

【０１７８】

実施形態１：

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、第二の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練することと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、を含む方法。

【０１７９】

実施形態２：

関連発現スコアが、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８０】

実施形態３：

複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することが、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８１】

実施形態４：

複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することが、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８２】

実施形態５：

５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む、実施形態４に記載の実施形態。

【０１８３】

実施形態６：

第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することが、第二の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８４】

実施形態７：

予測モデルについての複数の特徴が、ＧＣ含量、ＡＴおよびＣＧジヌクレオチド頻度、ＡＴＧ頻度、コアプロモーターモチーフの発生、相対エントロピー、および関連するＴＳＳに対する相対的位置決め方式のうちの一つまたは複数を含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８５】

実施形態８：

生成モデルで使用されるＴＳＳの発現スコアと重複する発現スコアを有する第一の複数のヌクレオチド配列から、複数のＴＳＳの任意のＴＳＳをフィルタリングすることをさらに含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８６】

実施形態９：

ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含む、先行する実施形態のいずれか一つに記載の実施形態。

【０１８７】

実施形態１０：

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、第二の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練することと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、を含む方法。

【０１８８】

実施形態１１：

第二の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列を決定することが、第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することと、各関連する複数のシフト塩基を、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として保存することと、を含む、実施形態１０に記載の実施形態。

【０１８９】

実施形態１２：

第二の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することが、第二の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含む、実施形態１１に記載の実施形態。

【０１９０】

実施形態１３：

関連発現スコアが、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含む、実施形態１０～１２のいずれかに記載の実施形態。

【０１９１】

実施形態１４：

第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することが、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩

基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することと、各サミットヌクレオチド塩基および関連する複数の周辺塩基を、コアプロモーターとして標識された第二の複数のヌクレオチド配列として保存することと、を含む、実施形態１０～１３のいずれかに記載の実施形態。

【０１９２】

実施形態１５：

複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することが、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む、実施形態１４に記載の実施形態。

【０１９３】

実施形態１６：

複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することが、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含む、実施形態１４に記載の実施形態。

【０１９４】

実施形態１７：

５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む、実施形態１６に記載の実施形態。

【０１９５】

実施形態１８：

予測モデルについての複数の特徴が、ＧＣ含量、ＡＴおよびＣＧジヌクレオチド頻度、ＡＴＧ頻度、コアプロモーターモチーフの発生、相対エントロピー、および関連するＴＳＳに対する相対的位置決め方式のうちの一つまたは複数を含む、実施形態１０～１７のいずれかに記載の実施形態。

【０１９６】

実施形態１９：

生成モデルで使用されるＴＳＳの発現スコアと重複する発現スコアを有する第一の複数のヌクレオチド配列から、複数のＴＳＳの任意のＴＳＳをフィルタリングすることをさらに含む、実施形態１４～１８のいずれかに記載の実施形態。

【０１９７】

実施形態２０：

ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含む、実施形態１０～１９のいずれかに記載の実施形態。

【０１９８】

実施形態２１：

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、遺伝的データを正規化することと、関連発現スコアに基づいて、ＴＳＳをクラスター化することと、ＴＳＳの各クラスターについて、四分位幅を決定することと、四分位幅に基づいて、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳに標識することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、

コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットにおける各ヌクレオチド配列に対して、複数のシード配列および標的ヌクレオチド対を生成することと、複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクトル化することと、ベクトル化されたシード配列および標的ヌクレオチド対に基づいて、生成モデルを訓練することと、および生成モデルを出力することと、を含む方法。

【０１９９】

実施形態２２：

関連発現スコアが、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含む、実施形態２１に記載の実施形態。

【０２００】

実施形態２３：

複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することが、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む、実施形態２１～２２のいずれかに記載の実施形態。

【０２０１】

実施形態２４：

複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することが、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含む、実施形態２１～２３のいずれかに記載の実施形態。

【０２０２】

実施形態２５：

５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む、実施形態２４に記載の実施形態。

【０２０３】

実施形態２６：

第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することが、第三の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含む、実施形態２１～２５のいずれかに記載の実施形態。

【０２０４】

実施形態２７：

ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含む、実施形態２１～２６のいずれかに記載の実施形態。

【０２０５】

実施形態２８：

各シード配列および標的ヌクレオチド対は、定義された長さを有するシード配列、および所定のヌクレオチド配列上のシード配列の直後に標的ヌクレオチドを含む、実施形態２１～２７のいずれかに記載の実施形態。

【０２０６】

実施形態２９：

定義された長さが１０塩基である、実施形態２８に記載の実施形態。

実施形態３０：

訓練データセット中の各ヌクレオチド配列について、複数のシード配列および標的ヌクレオチド対を生成することは、シャープＴＳＳまたはブロードＴＳＳ標識に基づいて、訓練データセット中のヌクレオチド配列をシャープＴＳＳ群またはブロードＴＳＳ群に分割することと、定義された長さのスライディングウィンドウを適用し、定義された工程サイズを各ヌクレオチド配列に有することと、スライディングウィンドウの各工程でシード配列および標的ヌクレオチド対を保存することと、を含む、実施形態２１～２９のいずれかに記載の実施形態。

【０２０７】

実施形態３１：

複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクター化することは、各ヌクレオチドをそれぞれの番号としてコード化することを含む、実施形態２１～３０のいずれかに記載の実施形態。

【０２０８】

実施形態３２：

生成モデルが、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を含む、実施形態２１～３１のいずれかに記載の実施形態。

【０２０９】

実施形態３３：

生成モデルに基づいて、ヌクレオチド配列を生成することをさらに含む、実施形態２１～３２のいずれかに記載の実施形態。

【０２１０】

実施形態３４：

生成モデルに基づいて、ヌクレオチド配列を生成することは、（ａ）シード配列を受信することと、（ｂ）シード配列に基づいて、次のヌクレオチドを予測することと、（ｃ）シード配列に次のヌクレオチドを付加することと、（ｄ）ヌクレオチド配列の所望の長さに達するまでｂ～ｃを繰り返すことと、を含む、実施形態３３に記載の実施形態。

【０２１１】

実施形態３５：

所望の長さは、約５０ヌクレオチド～約１００ヌクレオチドである、実施形態３４に記載の実施形態。

【０２１２】

実施形態３６：

ヌクレオチド配列がコアプロモーター配列である、実施形態３３～３５のいずれかに記載の実施形態。

【０２１３】

実施形態３７：

コアプロモーター配列に基づいてプロモーターを操作することをさらに含む、実施形態３６に記載の実施形態。

【０２１４】

実施形態３８：

プロモーターを核酸構築物に挿入することをさらに含む、実施形態３７に記載の実施形態。

【０２１５】

実施形態３９：

プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含む、実施形態３８に記載の実施形態。

【０２１６】

実施形態４０：

核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することをさらに

含む、実施形態３８～３９のいずれかに記載の実施形態。
【０２１７】
実施形態４１：
遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、生成モデルを訓練することと、および生成モデルを出力することと、を含む、方法。
【０２１８】
実施形態４２：
遺伝的データを正規化することをさらに含む、実施形態４１に記載の実施形態。
実施形態４３：
関連発現スコアに基づいて、ＴＳＳをクラスタリングすることと、ＴＳＳの各クラスターについて、四分位幅を決定することと、四分位幅に基づいて、各ＴＳＳをシャープＴＳＳまたはブロードＴＳＳとして標識することと、をさらに含む、実施形態４１～４２のいずれかに記載の実施形態。
【０２１９】
実施形態４４：
第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することが、第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することと、各関連する複数のシフト塩基を、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列として保存することと、を含む、実施形態４１～４３のいずれかに記載の実施形態。
【０２２０】
実施形態４５：
訓練データセットに基づいて、生成モデルを訓練することは、訓練データセット中の各ヌクレオチド配列について、複数のシード配列と標的ヌクレオチド対を生成することと、複数のシード配列と標的ヌクレオチド対の各シード配列と標的ヌクレオチド対をベクター化することと、ベクター化されたシード配列と標的ヌクレオチド対に基づいて、生成モデルを訓練することと、を含む、実施形態４１～４４のいずれかに記載の実施形態。
【０２２１】
実施形態４６：
関連発現スコアが、遺伝子発現のキャップ解析（ＣＡＧＥ）ピークを含む、実施形態４１～４５のいずれかに記載の実施形態。
【０２２２】
実施形態４７：
第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することは、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することと、各サミットヌクレオチド塩基および関連する複数の周辺塩基を、コアプロモーターとして標識された第二の複数のヌクレオチド配列として保存することと、を含む、実施形態４１～４６のいずれかに記載の実施形態。
【０２２３】

実施形態４８：
複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することが、複数のＴＳＳのそれぞれについて、最強のＣＡＧＥシグナルを有するヌクレオチド塩基を決定することを含む、実施形態４７に記載の実施形態。
【０２２４】
実施形態４９：
複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、関連する複数の周辺塩基を決定することが、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基について、５’方向の第一の複数のヌクレオチド塩基および３’方向の第二の複数のヌクレオチド塩基を決定することを含む、実施形態４７～４８のいずれかに記載の実施形態。
【０２２５】
実施形態５０：
５’方向の第一の複数のヌクレオチド塩基が４９個のヌクレオチド塩基を含み、３’方向の第二の複数のヌクレオチド塩基が５０個のヌクレオチド塩基を含む、実施形態４９に記載の実施形態。
【０２２６】
実施形態５１：
第三の複数のヌクレオチド配列の各ヌクレオチド配列について、関連する複数のシフト塩基を決定することが、第三の複数のヌクレオチド配列の各ヌクレオチド配列から、ある量のヌクレオチド塩基をシフトさせることを含む、実施形態４１～５０のいずれかに記載の実施形態。
【０２２７】
実施形態５２：
ヒトゲノムアセンブリ（ｈｇ１９）中にＮｓを含有する第二の複数のヌクレオチド配列の任意のヌクレオチド配列をフィルタリングすることをさらに含む、実施形態４１～５１のいずれかに記載の実施形態。
【０２２８】
実施形態５３：
各シード配列および標的ヌクレオチド対は、定義された長さを有するシード配列、および所定のヌクレオチド配列上のシード配列の直後に標的ヌクレオチドを含む、実施形態４５～５２のいずれかに記載の実施形態。
【０２２９】
実施形態５４：
定義された長さが１０塩基である、実施形態５３に記載の実施形態。
実施形態５５：
訓練データセット中の各ヌクレオチド配列について、複数のシード配列および標的ヌクレオチド対を生成することは、シャープＴＳＳまたはブロードＴＳＳ標識に基づいて、訓練データセット中のヌクレオチド配列をシャープＴＳＳ群またはブロードＴＳＳ群に分割することと、定義された長さのスライディングウィンドウを適用し、定義された工程サイズを各ヌクレオチド配列に有することと、スライディングウィンドウの各工程でシード配列および標的ヌクレオチド対を保存することと、を含む、実施形態４５～５４のいずれかに記載の実施形態。
【０２３０】
実施形態５６：
複数のシード配列および標的ヌクレオチド対の各シード配列および標的ヌクレオチド対をベクター化することは、各ヌクレオチドをそれぞれの番号としてコード化することを含む、実施形態４５～５５のいずれかに記載の実施形態。
【０２３１】
実施形態５７：

生成モデルが、長短期メモリ（ＬＳＴＭ）リカレントニューラルネットワーク（ＲＮＮ）を含む、実施形態４１～５６のいずれかに記載の実施形態。

【０２３２】

実施形態５８：

生成モデルに基づいて、ヌクレオチド配列を生成することをさらに含む、実施形態４１～５７のいずれかに記載の実施形態。

【０２３３】

実施形態５９：

生成モデルに基づいて、ヌクレオチド配列を生成することは、（ａ）シード配列を受信することと、（ｂ）シード配列に基づいて、次のヌクレオチドを予測することと、（ｃ）シード配列に次のヌクレオチドを付加することと、ｄ）ヌクレオチド配列の所望の長さに達するまでｂ～ｃを繰り返すことと、を含む、実施形態５８に記載の実施形態。

【０２３４】

実施形態６０：

所望の長さは、約５０ヌクレオチド～約１００ヌクレオチドである、実施形態５９に記載の実施形態。

【０２３５】

実施形態６１：

ヌクレオチド配列がコアプロモーター配列である、実施形態４１～６１のいずれかに記載の実施形態。

【０２３６】

実施形態６２：

コアプロモーター配列に基づいてプロモーターを操作することをさらに含む、実施形態６１に記載の実施形態。

【０２３７】

実施形態６３：

プロモーターを核酸構築物に挿入することをさらに含む、実施形態６２に記載の実施形態。

【０２３８】

実施形態６４：

プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含む、実施形態６３に記載の実施形態。

【０２３９】

実施形態６５：

核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することをさらに含む、実施形態６４に記載の実施形態。

【０２４０】

実施形態６６：

ヌクレオチド配列を受信することと、訓練された予測モデルにヌクレオチド配列を提供することと、および予測モデルに基づいて、ヌクレオチド配列がコアプロモーターであることを決定することと、を含む方法。

【０２４１】

実施形態６７：

ヌクレオチド配列を受信することは、複数のヌクレオチド配列を受信することを含み、複数のヌクレオチド配列は、生成モデルによって生成された、実施形態６６に記載の実施形態。

【０２４２】

実施形態６８：

ヌクレオチド配列がコアプロモーターであるという決定に基づいて、一つまたは複数の

基準に従ってヌクレオチド配列をフィルタリングすることをさらに含む、実施形態６６～６７のいずれかに記載の実施形態。

【０２４３】

実施形態６９：

一つまたは複数の基準が、ＧＣ含量またはモチーフのうちの一つまたは複数を含む、実施形態６８に記載の実施形態。

【０２４４】

実施形態７０：

遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、閾値を満たす関連発現スコアに基づいて、第一の複数のヌクレオチド配列から複数のＴＳＳを決定することと、複数のＴＳＳに基づいて、複数のサミットヌクレオチド塩基を決定することと、複数のサミットヌクレオチド塩基の各サミットヌクレオチド塩基に対して、関連する複数の周辺塩基を決定することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列として、各サミットヌクレオチド塩基および関連する複数の周辺塩基を保存することと、第二の複数のヌクレオチド配列の各ヌクレオチド配列に対して、関連する複数のシフト塩基を決定することと、コアプロモーターではないとして標識された第三の複数のヌクレオチド配列として、各関連する複数のシフト塩基を保存することと、コアプロモーターとして標識された第二の複数のヌクレオチド配列、およびコアプロモーターではないとして標識された第三の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、訓練データセットに基づいて、予測モデルに対する複数の特徴を決定することと、訓練データセットの第一の部分に基づいて、複数の特徴に従って予測モデルを訓練することと、訓練データセットの第二の部分に基づいて、予測モデルを試験することと、および試験に基づいて、予測モデルを出力することと、をさらに含む、実施形態６６～６９のいずれかに記載の実施形態。

【０２４５】

実施形態７１：

（ａ）ヌクレオチド配列と配列長を受信すること、（ｂ）訓練された生成モデルに、ヌクレオチド配列を提供すること、（ｃ）生成モデルに基づいて、ヌクレオチド配列に関連付けられた次のヌクレオチドを決定すること、（ｄ）ヌクレオチド配列に次のヌクレオチドを付与すること、（ｅ）ヌクレオチド配列の長さが配列長に等しくなるまでｂ～ｄを繰り返すこと、および（ｆ）ヌクレオチド配列をコアプロモーター配列として出力すること、を含む方法。

【０２４６】

実施形態７２：

コアプロモーター配列に基づいてプロモーターを操作することをさらに含む、実施形態７１に記載の実施形態。

【０２４７】

実施形態７３：

プロモーターを核酸構築物に挿入することをさらに含む、実施形態７１～７２のいずれかに記載の実施形態。

【０２４８】

実施形態７４：

プロモーターを核酸構築物に挿入することは、導入遺伝子の上流の核酸構築物にプロモーターを挿入して、導入遺伝子の発現を駆動することを含む、実施形態７３に記載の実施形態。

【０２４９】

実施形態７５：

核酸構築物を含む、アデノ随伴ウイルスまたはレンチウイルスを作製することをさらに含む、実施形態７３～７４のいずれかに記載の実施形態。

【０２５０】
実施形態７６：
配列長は、約５０ヌクレオチド～約１００ヌクレオチドである、実施形態７１～７５のいずれかに記載の実施形態。
【０２５１】
実施形態７７：
遺伝的データが、第一の複数のヌクレオチド配列を含み、複数のヌクレオチド配列の各ヌクレオチド配列が、関連発現スコアを有する少なくとも一つの転写開始点（ＴＳＳ）を含む、遺伝的データを受信することと、第一の複数のヌクレオチド配列に基づいて、コアプロモーターとして標識された第二の複数のヌクレオチド配列を決定することと、閾値を満たす関連発現スコアに基づいて、第二の複数のヌクレオチド配列から第三の複数のヌクレオチド配列を決定することと、第三の複数のヌクレオチド配列に基づいて、コアプロモーターではないとして標識された第四の複数のヌクレオチド配列を決定することと、コアプロモーターとして標識された第三の複数のヌクレオチド配列およびコアプロモーターではないとして標識された第四の複数のヌクレオチド配列に基づいて、訓練データセットを生成することと、および訓練データセットに基づいて、生成モデルを訓練することと、をさらに含む、実施形態７１～７６のいずれかに記載の実施形態。
【０２５２】
実施形態７８：
一つまたは複数の基準に従ってヌクレオチド配列をフィルタリングすることをさらに含む、実施形態７１～７７のいずれかに記載の実施形態。
【０２５３】
実施形態７９：
一つまたは複数の基準が、ＧＣ含量またはモチーフのうちの一つまたは複数を含む、実施形態７８に記載の実施形態。
【０２５４】
実施形態８０：
実施形態１～７９に記載のいずれかを行うため構成された、装置。
実施形態８１：
装置が実施形態１～７９に記載のいずれかを行うよう構成された、プロセッサが実行可能な指示実施形態を有する、コンピュータ可読媒体。
【０２５５】
当業者は、通常の実験だけを用いることで、本明細書に記載の方法および組成物の特定の実施形態の多数の同等物を認識し、または確認できる。かかる同等物は、以下の特許請求の範囲に包含されることが意図される。

【図面】

【図１】

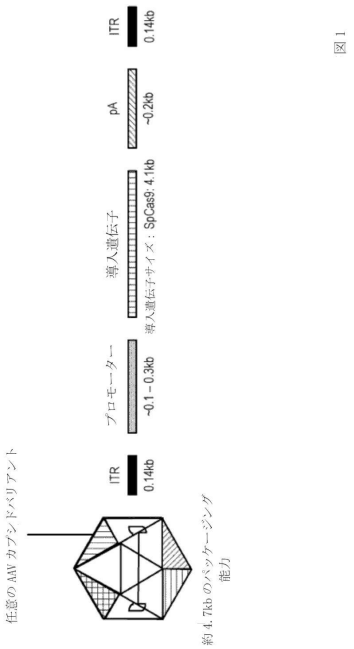
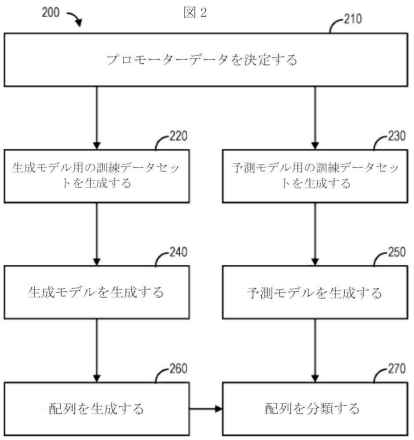
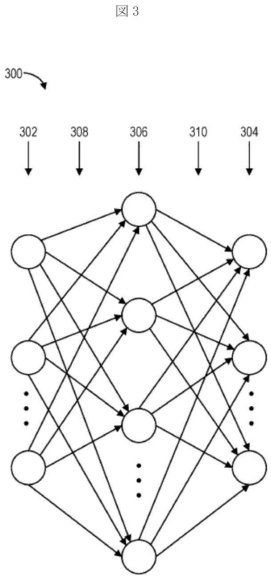


図１

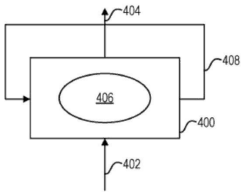
【図２】



【図３】

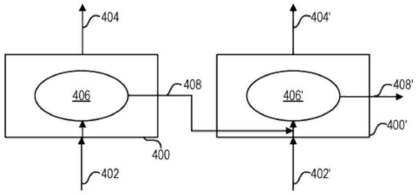


【図４Ａ】



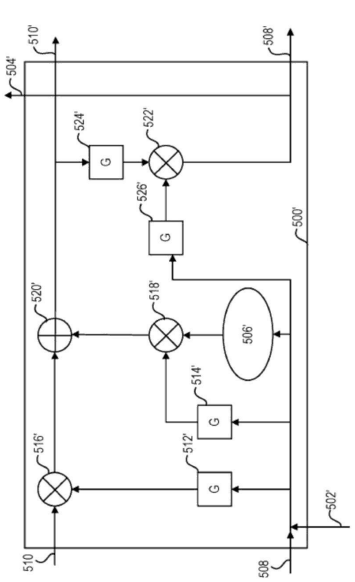
【図４Ｂ】

図 4B



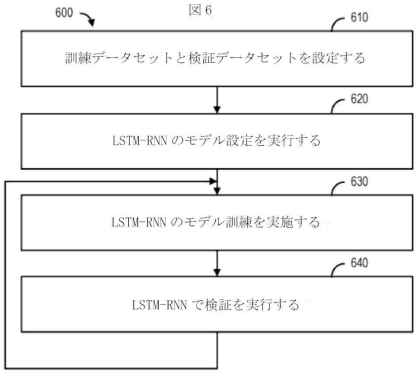
【図５】

図 5



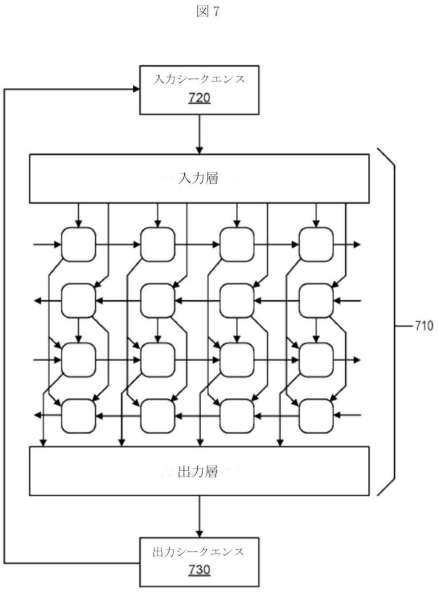
【図６】

図 6



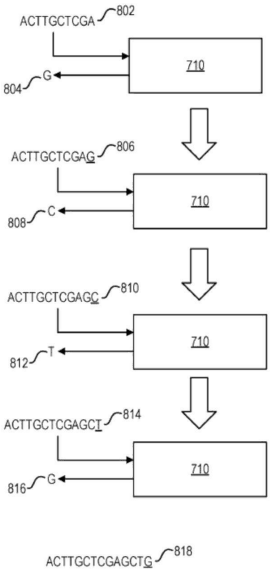
【図７】

図 7



【図８】

図８



【図９】

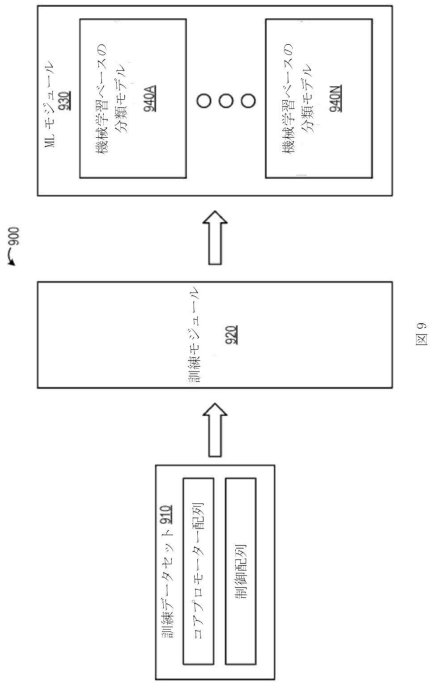
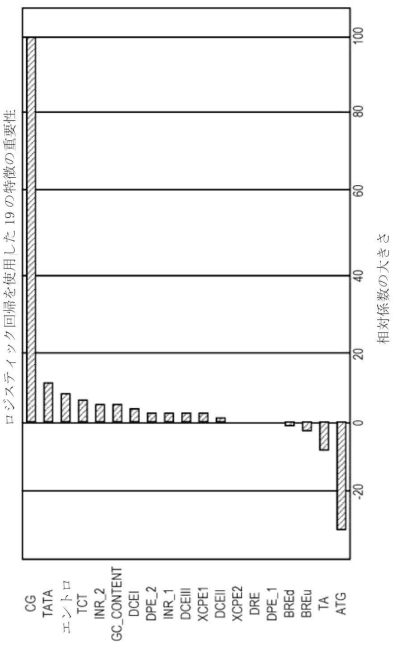


図９

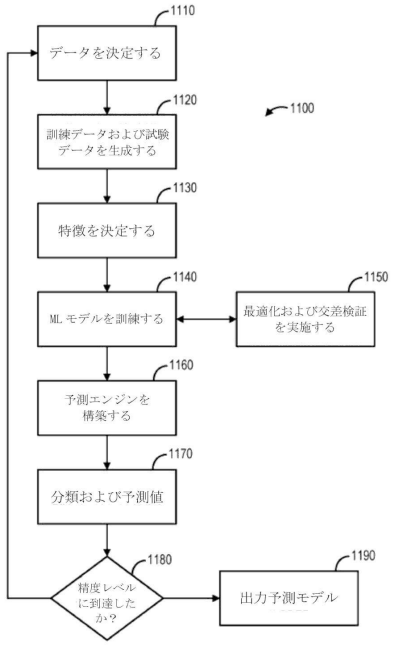
【図１０】

図 10



【図１１】

図 11



【図１２】

図 12

【図１３】

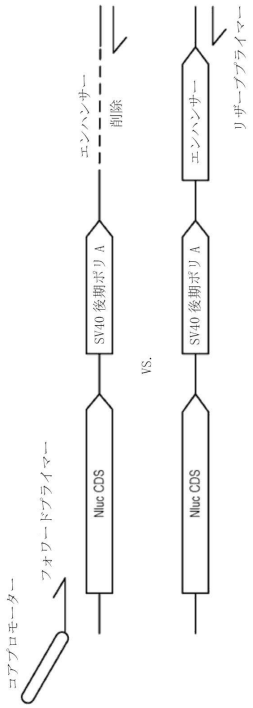


図 13

【図１４】

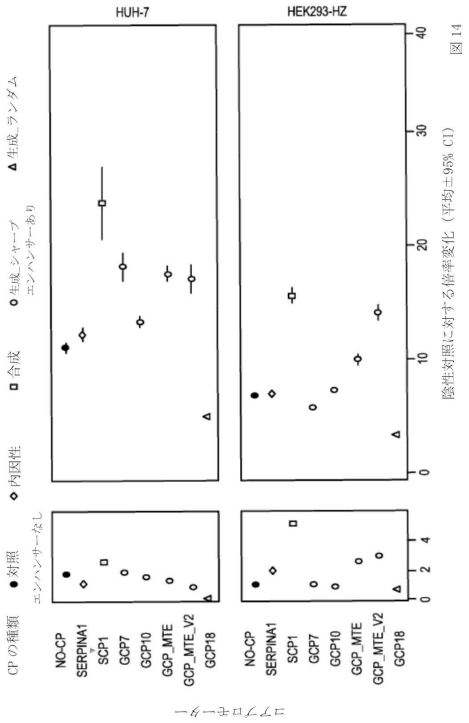


図 14

【図１５】

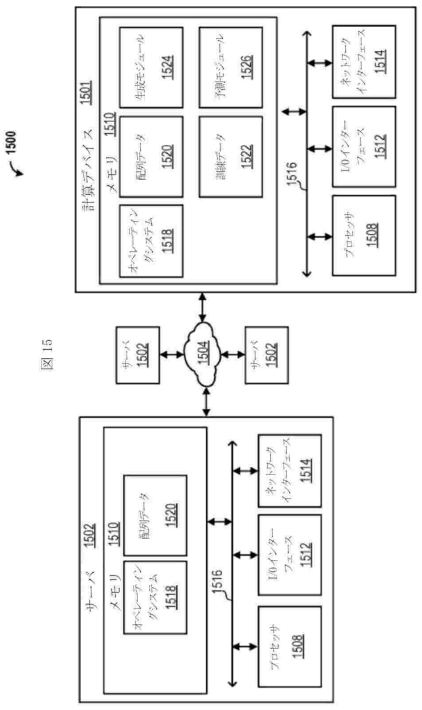
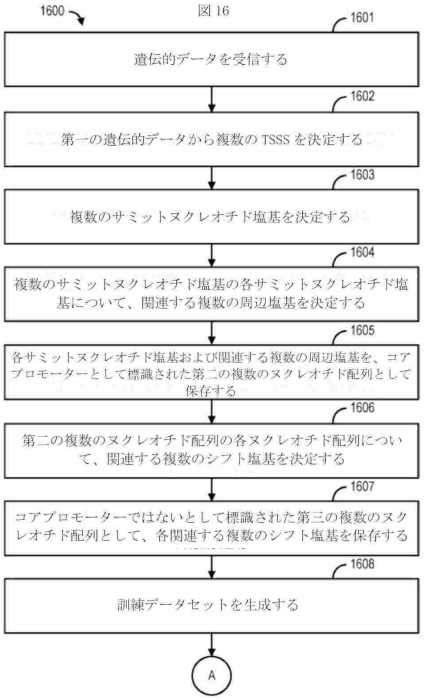
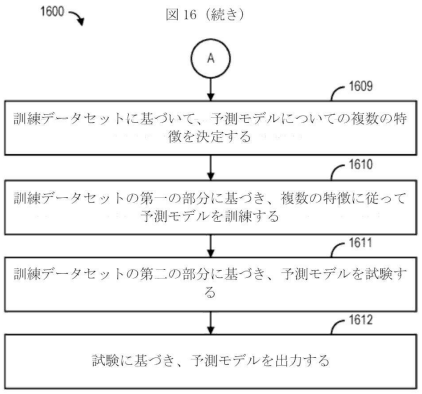


図 15

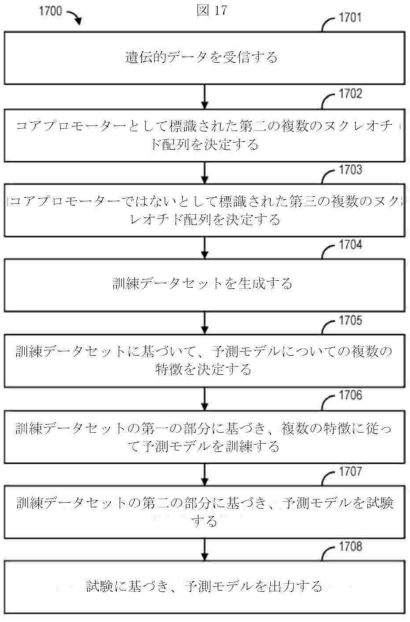
【図１６－１】



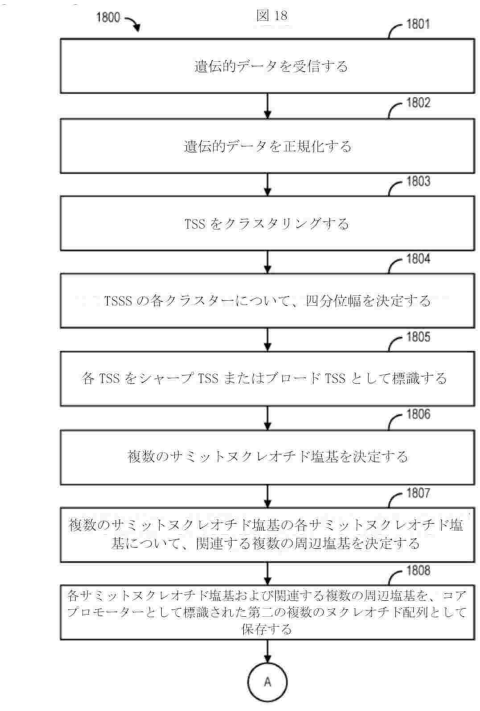
【図１６－２】



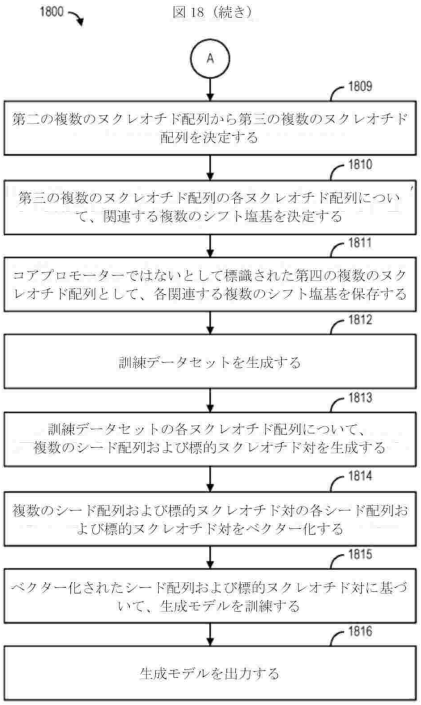
【図１７】



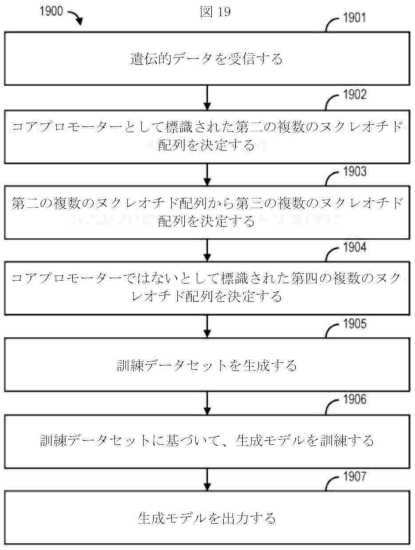
【図１８－１】



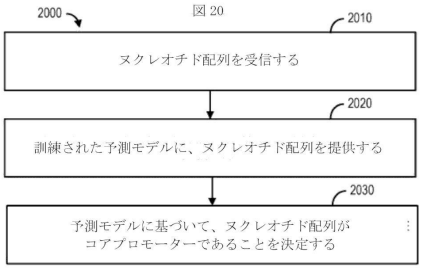
【図１８－２】



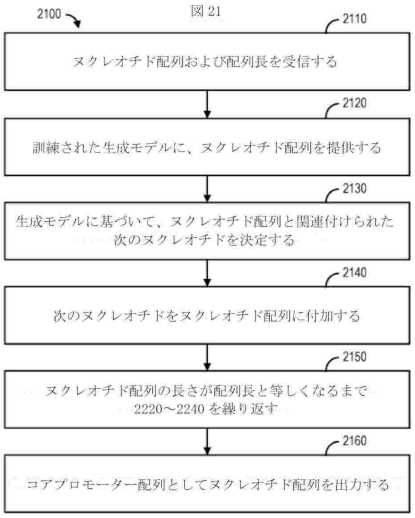
【図１９】



【図２０】



【図２１】



【配列表】
0007583153000001.app

フロントページの続き

弁理士　中村 美樹
(72)発明者　ミュルター、フェリクス
アメリカ合衆国 １０５９１ ニューヨーク州 タリータウン オールド ソー ミル リバー ロード ７７７
(72)発明者　シェーンヘル、クリストファー
アメリカ合衆国 １０５９１ ニューヨーク州 タリータウン オールド ソー ミル リバー ロード ７７７

審査官　山田 倍司

(56)参考文献
米国特許出願公開第２０１９／００７３４４３（ＵＳ，Ａ１）
国際公開第１９９９／０６６３０２（WO，Ａ２）
Ye Wang et al.，Synthetic Promoter Design in Escherichia coli based on Generative Adversarial Network，bioRxiv，2019年04月25日，https://doi.org/10.1101/563775
Georgios K. Georgakilas et al.，Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data，Scientific Reports，2020年01月21日，Vol.10，No.877，https://doi.org/10.1038/s41598-020-57811-3
Mhaned Oubounyt et al.，DeePromoter: Robust Promoter Predictor Using Deep Learning，Frontiers in Genetics，2019年04月05日，Vol.10，No.286，https://doi.org/10.3389/fgene.2019.00286
David R. Kelley，Cross-species regulatory sequence activity prediction，PLoS Computational Biology，2020年07月20日，Vol.16，No.7，https://doi.org/10.1371/journal.pcbi.1008050

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　５／００－９９／００
Ｃ１２Ｎ　１５／００－１５／９０
Ｃ１２Ｑ　１／００－ ３／００