

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2022-509589
(P2022-509589A)

(43)公表日 令和4年1月21日(2022.1.21)

(51)国際特許分類		F I		テーマコード (参考)	
C 1 2 Q	1/6869(2018.01)	C 1 2 Q	1/6869	Z	2 G 0 4 5
C 1 2 M	1/00 (2006.01)	C 1 2 M	1/00	A	2 G 0 6 0
G 0 1 N	33/50 (2006.01)	G 0 1 N	33/50	P	4 B 0 2 9
G 0 1 N	27/00 (2006.01)	G 0 1 N	27/00	Z	4 B 0 6 3
		G 0 1 N	27/00	J	
		審査請求	有	予備審査請求	未請求 (全64頁)
(21)出願番号 特願2021-525061(P2021-525061)		(71)出願人 511252899			
(86)(22)出願日 令和1年11月26日(2019.11.26)		オックスフォード ナノポール テクノロ ジーズ ピーエルシー イギリス国 オックスフォード オーエッ クス4 4ディーキュー オックスフォー ド サイエンс パーク エドモンド ハレ ー ロード ゴズリング ビルディング			
(85)翻訳文提出日 令和3年5月10日(2021.5.10)					
(86)国際出願番号 PCT/GB2019/053334					
(87)国際公開番号 WO2020/109773					
(87)国際公開日 令和2年6月4日(2020.6.4)		(74)代理人 100092783			
(31)優先権主張番号 1819378.9		弁理士 小林 浩			
(32)優先日 平成30年11月28日(2018.11.28)					
(33)優先権主張国・地域又は機関 英国(GB)		(74)代理人 100120134			
(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,A T,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR ,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC, 最終頁に続く		(74)代理人 弁理士 大森 規雄			
		(74)代理人 100187964			
		弁理士 新井 剛			
		(74)代理人 100104282			
		弁理士 鈴木 康仁			
		最終頁に続く			

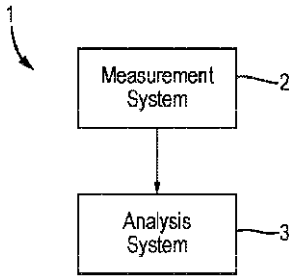
(54)【発明の名称】 機械学習技術を使用するナノ細孔シグナルの分析

(57)【要約】

一連のポリマー単位を含むポリマーによって生じるシグナルが、ナノ細孔に対するポリマーの転位中に機械学習技術を使用して分析される。一連の重み分布が出力され、各々がポリマー内の可能なポリマー単位タイプを表すラベルのセットにわたるラベル間の遷移に関する重みを含む。重み分布からの一連のポリマー単位の推定。遷移に関する重みを使用することは、技術精度を向上させる。

【選択図】 図 1

Fig. 1



【特許請求の範囲】

【請求項 1】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析する方法であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記方法が、

一連の重み分布を出力する機械学習技術を使用して前記シグナルを分析するステップであって、各重み分布が、前記可能なポリマー単位タイプを表すラベルを含む、ラベルのセットにわたるラベル間の遷移に関する重みを含む、ステップと、

前記重み分布から前記一連のポリマー単位の推定値を導出するステップと

10

を含む、前記方法。

【請求項 2】

ラベル間の少なくとも 1 つの遷移が許容されておらず、他の遷移が許容されており、前記重み分布がそれぞれ、許容されている遷移に関する重みを含む、請求項 1 に記載の方法。

【請求項 3】

前記重み分布がそれぞれ、許容されていない遷移に関するヌル重みを含む、請求項 2 に記載の方法。

【請求項 4】

前記一連のポリマー単位の推定値を導出するステップは、ラベル間の遷移が許容されているか許容されていないかを表す遷移行列を考慮に入れる、請求項 2 または 3 に記載の方法。

20

【請求項 5】

前記ラベルのセットが、各ポリマー単位タイプに関する第 1 のラベルおよび第 2 のラベルを含み、前記第 1 のラベルが、前記ポリマー単位タイプのインスタンスの開始を表し、前記第 2 のラベルが、前記ポリマー単位タイプの前記インスタンスにおけるステイを表し、各第 1 のラベルから任意の他のポリマー単位タイプの前記第 1 のラベルへの遷移が許容されており、

各第 1 のラベルから同じポリマー単位タイプの前記第 1 のラベルへの遷移が許容されており、

各第 1 のラベルから同じポリマー単位タイプの前記第 2 のラベルへの遷移が許容されており、

30

各第 1 のラベルから任意の他のポリマー単位タイプの前記第 2 のラベルへの遷移が許容されておらず、

各第 2 のラベルから同じポリマー単位タイプの前記第 1 のラベル、または任意の他のポリマー単位タイプの前記第 1 のラベルへの遷移が許容されており、

各第 2 のラベルから同じポリマー単位タイプの前記第 2 のラベルへの遷移が許容されており、

各第 2 のラベルから任意の他のポリマー単位タイプの前記第 2 のラベルへの遷移が許容されていない、請求項 2 ～ 4 のいずれか一項に記載の方法。

【請求項 6】

前記可能なポリマー単位タイプのセットが、ポリマー単位の既知の配列で常に現れるポリマー単位タイプを含み、前記既知の配列に従った遷移が許容されており、前記既知の配列に反した遷移が許容されていない、請求項 2 ～ 5 のいずれか一項に記載の方法。

40

【請求項 7】

前記一連のポリマー単位における同じタイプのポリマー単位の連続的インスタンスが符号化形式で表される、請求項 1 ～ 6 のいずれか一項に記載の方法。

【請求項 8】

前記ラベルが、各ポリマー単位タイプに関する複数のラベルを含み、各ポリマー単位タイプに関する前記複数のラベルが、前記一連のポリマー単位における前記ポリマー単位タイプの連続的インスタンスを表す、請求項 7 に記載の方法。

50

【請求項 9】

各ポリマー単位タイプの前記複数のラベルが、所定の周期的順序を有し、それによって、ラベル間のいくつかの遷移が、前記所定の周期的順序によって許容されており、間の他の遷移が、前記所定の周期的順序によって許容されておらず、前記重み分布が各々、前記所定の周期的順序によって許容されている遷移に関する重みを含む、請求項 8 に記載の方法。

【請求項 10】

各ポリマー単位タイプの前記複数のラベルが、各ポリマー単位タイプの 2 つのラベルである、請求項 8 または 9 に記載の方法。

【請求項 11】

前記一連のポリマー単位における同じポリマー単位タイプの連続するインスタンスが、ランレングス符号化形式で表される、請求項 7 に記載の方法。

【請求項 12】

前記ラベルが、各ポリマー単位タイプの異なるランレングスに関する複数のラベルを含む、請求項 11 に記載の方法。

【請求項 13】

前記ラベルが、各ポリマー単位タイプに関するラベルを含み、前記重み分布が、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さにわたるさらなる重みを含む、請求項 11 に記載の方法。

【請求項 14】

前記さらなる重みが、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さのセットにわたる重みのカテゴリカル分布を含む、請求項 13 に記載の方法。

【請求項 15】

前記さらなる重みが、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さにわたる、パラメータ化された分布のパラメータを含む、請求項 13 に記載の方法。

【請求項 16】

前記重み分布が、(a) 所与のポリマー単位タイプおよび先行するポリマー単位タイプの可能な対の、(b) 所与のポリマー単位タイプおよび後続のポリマー単位タイプの可能な対の、または (c) 所与のポリマー単位タイプ、先行するポリマー単位タイプ、および後続のポリマー単位タイプの可能なトリプレットの同じポリマー単位タイプの連続的インスタンスの可能な長さにわたるさらなる重みを含む、請求項 13 ～ 15 のいずれか一項に記載の方法。

【請求項 17】

前記可能なポリマー単位タイプが、非修飾形態および修飾形態を有するポリマー単位タイプを含む、請求項 1 ～ 16 のいずれか一項に記載の方法。

【請求項 18】

前記ラベルのセットが、非修飾形態および修飾形態を有する前記ポリマー単位タイプに関するラベルを含む、請求項 17 に記載の方法。

【請求項 19】

各重み分布が、前記非修飾形態および修飾形態を有する各前記ポリマー単位タイプの前記非修飾形態および修飾形態のさらなる重みを含む、請求項 18 に記載の方法。

【請求項 20】

前記ラベルのセットが、各ポリマー単位タイプを表す少なくとも 1 つのラベルを含む、請求項 1 ～ 19 のいずれか一項に記載の方法。

【請求項 21】

前記ラベルのセットが、前記一連のポリマー単位におけるブランクおよび／またはステイを表す少なくとも 1 つのラベルをさらに含む、請求項 1 ～ 20 のいずれか一項に記載の方法。

10

20

30

40

50

【請求項 22】

前記機械学習技術が、少なくとも1つの再帰層を含むニューラルネットワークである、請求項1～21のいずれか一項に記載の方法。

【請求項 23】

前記少なくとも1つの再帰層が、双方向再帰層である、請求項22に記載の方法。

【請求項 24】

前記ニューラルネットワークが、前記一連の重み分布にわたるラベルのすべての経路にわたって前記重み分布のグローバル正規化を適用する、請求項22または23に記載の方法。

【請求項 25】

前記ニューラルネットワークが、前記少なくとも1つの再帰層の前に配置され、前記シグナルのウィンドウ化セクションの畳み込みを実行する、少なくとも1つの畳み込み層を含む、請求項22～24のいずれか一項に記載の方法。

【請求項 26】

前記重みが事後確率を表す、請求項1～25のいずれか一項に記載の方法。

【請求項 27】

前記重み分布から前記一連のポリマー単位の推定値を導出するステップが、接続ノスト時間分類を使用して実行される、請求項1～26のいずれか一項に記載の方法。

【請求項 28】

前記重み分布からポリマー単位の推定値を導出する前記ステップが、それぞれの重み分布に関するラベルを導出することと、前記導出されたラベルをランレングス圧縮することとを含む、請求項1～27のいずれか一項に記載の方法。

【請求項 29】

前記重み分布から前記一連のポリマー単位の推定値を導出する前記ステップが、前記重み分布を基準にして前記一連の重み分布にわたるラベルの最も可能性の高い経路を推定することと、最も可能性が高いと推定されたラベルの前記経路から前記一連のポリマー単位の前記推定値を導出することと、を含む、請求項1～28のいずれか一項に記載の方法。

【請求項 30】

前記重み分布から前記一連のポリマー単位の推定値を導出する前記ステップが、各重み分布に関する最も可能性が高い前記ラベルを推定することと、前記一連の重み分布にわたるラベルのフォワード経路とバックワードの経路を考慮に入れることと、最も可能性が高いと推定された前記ラベルから前記一連のポリマー単位の前記推定値を導出することと、を含む、請求項1～28のいずれか一項に記載の方法。

【請求項 31】

前記ナノ細孔がタンパク質細孔である、請求項1～30のいずれか一項に記載の方法。

【請求項 32】

前記ポリマーがポリヌクレオチドであり、前記ポリマー単位がヌクレオチドである、請求項1～31のいずれか一項に記載の方法。

【請求項 33】

前記シグナルが、イオン電流、インピーダンス、トンネリング特性、電界効果トランジスタ電圧、および光学特性のうちの1つ以上の特性の測定値から導出される、請求項1～32のいずれか一項に記載の方法。

【請求項 34】

前記方法が、コンピュータ装置で実行される、請求項1～33のいずれか一項に記載の方法。

【請求項 35】

ナノ細孔に対する前記ポリマーの転位中に前記ポリマーから前記シグナルを得ることをさらに含む、請求項1～34のいずれか一項に記載の方法。

【請求項 36】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析するた

10

20

30

40

50

めの分析装置であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記分析装置は、

前記シグナルに対して機械技術を実行し、一連の重み分布を出力するように構成された機械学習ユニットであって、各重み分布が、前記可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベル間の遷移に関する重みを含む、前記機械学習ユニットと、

前記重み分布から前記一連のポリマー単位の推定値を導出するように構成された推定ユニットと、を含む、前記分析装置。

【請求項 37】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に導出するように構成された測定デバイスと、

請求項 36 に記載の分析装置と

を含むナノ細孔測定および分析システム。

【請求項 38】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析する方法であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記方法は、

一連の重み分布を出力する機械学習技術を使用して前記シグナルを分析するステップであって、各重み分布が、前記可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含む、ステップと、

前記重み分布から前記一連のポリマー単位の推定値を導出するステップと

を含み、

前記一連のポリマー単位の推定値を導出するステップは、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れ、ラベル間の少なくとも 1 つの遷移が許容されないものとして表され、他の遷移が許容されるものとして表される、前記方法。

【請求項 39】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析するための分析装置であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記分析装置は、

前記シグナルに対して機械技術を実行し、一連の重み分布を出力するように構成された機械学習ユニットであって、各重み分布が、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含む、前記機械学習ユニットと、

前記重み分布から前記一連のポリマー単位の推定値を導出するように構成された推定ユニットと

を含み、

前記推定ユニットは、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れるように構成され、ラベル間の少なくとも 1 つの遷移が許容されないものとして表され、他の遷移が許容されるものとして表される、分析装置。

【請求項 40】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析する方法であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記方法は、

一連の重み分布を出力する機械学習技術を使用してシグナルを分析するステップであって、各重み分布が、前記可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含み、前記一連のポリマー単位における同じポリマー単位タイプの連続的インスタンスが、ランレングス符号化形式で表される、ステップと、

前記重み分布から前記一連のポリマー単位の推定値を導出するステップとを含む、前記方法。

【請求項 41】

ポリマーによって生じるシグナルをナノ細孔に対する前記ポリマーの転位中に分析するための分析装置であって、前記ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、

前記分析装置は、

前記シグナルに対して機械技術を実行し、一連の重み分布を出力するように構成された機械学習ユニットであって、各重み分布が、前記可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含み、前記一連のポリマー単位における同じポリマー単位タイプの連続的インスタンスが、ランレングス符号化形式で表される、前記機械学習ユニットと、

前記重み分布から前記一連のポリマー単位の推定値を導出するように構成された推定ユニットと

を含む前記分析装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、限定されるものではないが、例えばポリヌクレオチドなどのナノ細孔に対するポリマーの転位中のポリマーによって生じるシグナルの分析に関する。

【0002】

ポリマーがナノ細孔に対して転位するナノ細孔を使用して、ポリマー中のポリマー単位の標的配列を推定するための測定システムは、既知である。システムのいくつかの特性は、ナノ細孔のポリマー単位に依存し、その特性の測定値が得られる。この特性は、ナノ細孔に対して転位するポリマー単位の一貫性に依存しており、そのため、経時的なシグナルで、ポリマー単位のシーケンスが推定されることを可能にする。各ポリマー単位は、細孔の寸法と比較して非常に小さいものであり得、複数のポリマー単位が所与の期間でシグナルに影響を及ぼすことを可能にする。ポリマー鎖とナノ細孔との相互作用、巻き取りやスタッキングなどの鎖内の特性、またはポリマー単位とそれらの移動を制御するために使用されるいずれかのシステムとの間の相互作用により、より長距離の影響も存在し得る。シグナルはリードを形成し、これは基礎となるポリマー単位を推定するために復号される必要がある。

【0003】

このようなナノ細孔測定システムは、ヌクレオチドが数百～数十万（および場合によってはそれ超）にも及ぶようなポリヌクレオチドの長い連続リードを表すシグナルを提供することができる。ナノ細孔を使用するこのタイプの測定システムは、特にDNAまたはRNAなどのポリヌクレオチドの配列決定の分野で非常に効果的であることが示されており、最新の開発対象になっている。

【0004】

しかしながら、ポリマー単位の推定の精度は、測定システムの感度が極端に高いために制限される。実際の問題として、高精度の推定は機械学習技術を必要とし、本発明は、推定精度を高めるためにそのような分析を改善することに関する。

【0005】

初期の分析手法は、連続するポリマー単位のグループで構成される可能なk-merを明確にモデル化する隠れマルコフモデル（HMM）が使用された。より最近では、リカレントニューラルネットワーク（RNN）を使用する手法が開発されている。RNNを使用することで、長距離情報を考慮することが可能になり、精度を向上することができる。読み取り速度が向上し、その結果、HMMのような明示的なシグナルモデリングアプローチの基礎となる仮定の有効性が低くなることから、RNNは特に有用である。例として、Teng et al, “Chiron: Translating nanopore raw

10

20

30

40

50

signal directly into nucleotide sequence using deep learning”, Gigascience, 2018年5月1日; 7 (5) [参考文献1] は、生シグナルが入力されるRNNを使用してポリマーヌクレオチドを配列決定する方法を開示する。RNNは、4つの可能性のある塩基タイプとブランクを表すラベルに関する事後確率を含む一連の事後確率分布を出力する。これらは、コネクショニスト時間分類 (connectionist temporal classification) デコーダによって復号され、事後確率から最も可能性の高いポリマー単位を推定することによって、一連のポリマー単位の推定値を導出する。

【0006】

本発明の第1の態様によれば、ポリマーによって生じるシグナルをナノ細孔に対するポリマーの転位中に分析する方法が提供されており、ポリマーは、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含んでおり、本方法は、一連の重み分布を出力する機械学習技術を使用してシグナルを分析することであって、各重み分布が、可能なポリマー単位タイプを表すラベルを含んだラベルのセットにわたるラベル間の遷移に関する重みを含んでいることと、重み分布から一連のポリマー単位の推定値を導出することと、を含む。

10

【0007】

ラベルのセットは、ブランクおよび／またはステイを表すラベルを含むことができる。言い換えると、このセットは、可能なポリマー単位タイプを表すと言うことができる。

【0008】

20

この遷移は、あるラベルと別のラベルとの間のものであり得る。この遷移は、連続するラベル間のものであり得る。

【0009】

したがって、本方法は、k個のポリマー単位を含むk-merを表すのではなく、可能なポリマー単位タイプを表すラベルを示す重みを提供する。しかしながら、本方法は、ラベル自体に関する重みではなく、ラベル間の遷移に関する重みを導出する。そのような方法は、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する一連の重みを導出する比較法よりも利点をもたらす。ラベルのセットにわたるラベル間の遷移に関する重みを提供することにより、より正確な様式で一連のポリマー単位の推定を可能にする追加情報が提供される。これは、重みがラベルの可能な経路に関する情報を提供するのに対し、ラベルに関する重みは提供しないためである。

30

【0010】

例えば、ラベルに関する重みによって予測される特定の位置について、ラベルが正しくない状況があるが、一方で、その位置を通るラベルの経路を考慮することで、正しい別のラベルが予測され得る。このようにして、追加情報が推定に供給され、それによって精度を向上させる。

【0011】

例として、この技術は、例えば、ホモポリマーなど、1つ以上のポリマー単位の短い配列が繰り返される領域を含む反復配列の領域の、より良好な推定を可能にする。

【0012】

40

好ましくは、ラベル間の少なくとも1つの遷移が許容されておらず、他の遷移が許容されており、重み分布は各々、許容されている遷移に関する重みを含んでいる。その場合、重み分布は各々、許容されていない遷移に関するヌル重みを含むことができ、または一連のポリマー単位の推定値を導出するステップは、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れることができる。

【0013】

表現の1つのタイプでは、ラベルのセットは、各ポリマー単位タイプに関する第1および第2のラベルを含むことができ、第1のラベルはポリマー単位タイプのインスタンスの開始を表し、第2のラベルはポリマー単位タイプのインスタンスにおけるステイを表しており、各第1のラベルから任意の他のポリマー単位タイプの第1のラベルへの遷移が許容さ

50

れており、各第1のラベルから同じポリマー単位タイプの第1のラベルへの遷移が許容されており、各第1のラベルから同じポリマー単位タイプの第2のラベルへの遷移が許容されており、各第1のラベルから任意の他のポリマー単位タイプの第2のラベルへの遷移が許容されておらず、各第2のラベルから同じポリマー単位タイプの第1のラベル、または任意の他のポリマー単位タイプの第1のラベルへの遷移が許容されており、各第2のラベルから同じポリマー単位タイプの第2のラベルへの遷移が許容されており、各第2のラベルから任意の他のポリマー単位タイプの第2のラベルへの遷移が許容されていない。

【0014】

「ステイ」は、本方法が、ラベルが変化しないと判断する状況を表し、このことは、ポリマー単位の同じインスタンスに対応する2つの重み分布と見なされ得る。

10

【0015】

可能なポリマー単位タイプのセットは、ポリマー単位の既知の配列で常に現れるポリマー単位タイプを含むことができ、既知の配列に従った遷移が許容されており、既知の配列に反した遷移が許容されていない。

【0016】

一連のポリマー単位の中で同じポリマー単位タイプの連続的インスタンスが、符号化形式で表されることができる。

【0017】

ラベルが、各ポリマー単位タイプに関する、例えば2つのラベルなどの複数のラベルを含むことができ、各ポリマー単位タイプに関する複数のラベルが、一連のポリマー単位におけるポリマー単位タイプの連続的インスタンスを表す。

20

【0018】

各ポリマー単位タイプに関する複数のラベルは、所定の周期的順序を有することができ、それにより、ラベル間のいくつかの遷移が、所定の周期的順序によって許容されており、ラベル間の他の遷移が、所定の周期的順序によって許容されておらず、重み分布が、所定の周期的順序によって許容されている遷移に関する重みを含む。

【0019】

一連のポリマー単位内の同じポリマー単位タイプの連続するインスタンスは、ランレングス符号化形式で表される。

【0020】

これらのラベルは、各ポリマー単位タイプの異なるランレングスに関する複数のラベルを含み得る。

30

【0021】

これらのラベルは、各ポリマー単位タイプに関する1つのラベルを含むことができ、重み分布が、各ポリマー単位タイプについて、同じポリマー単位タイプの連続的インスタンスの可能な長さにわたるさらなる重みを含むことができる。

【0022】

さらなる重みは、各ポリマー単位タイプについて、同じポリマー単位タイプの連続的インスタンスの可能な長さのセットにわたる重みのカテゴリカル分布を含むことができる。

【0023】

さらなる重みは、各ポリマー単位タイプについて、同じタイプのポリマー単位の連続するインスタンスの可能な長さにわたるパラメータ化された分布のパラメータを含むことができる。

40

【0024】

可能なポリマー単位タイプは、非修飾および修飾の形態を有するポリマー単位タイプを含む場合、ラベルのセットが、非修飾形態および修飾形態を有するポリマー単位タイプを表すラベルを含み得、各重み分布は、非修飾形態および修飾形態を有する少なくとも1つのポリマー単位タイプの各々の非修飾形態および修飾形態について、さらなる重みを含み得る。ポリマー単位の非修飾形態は、カノニカルポリマー単位として説明され得、ポリマー単位の修飾形態は、非カノニカルポリマー単位として説明され得る。修飾（または非カノ

50

ニカル) ポリマー単位は、通常、対応する非修飾 (カノニカル) ポリマー単位とは異なる形でシグナルに影響を及ぼす。

【0025】

いくつかの実施形態において、1つ以上の非カノニカルポリマー単位を含むポリマーは、参照が行われ、参照により本明細書に組み込まれている2019年9月4日に出願された国際特許出願第PCT/GB2019/052456号に詳細に記載されるように調製され、続いて分析され得る。一例では、ある割合のカノニカルポリマー単位 (例えば、アミノ酸) は、非決定論的な方法で、例えば、化学変換または酵素変換によって、対応する非カノニカルポリマー単位 (例えば、アミノ酸) に変換され得る。その場合、一連のポリマー単位の推定値を導出 (「呼び出し」) するとき、非カノニカル塩基は、対応するカノニカル塩基であると推定 (「呼び出される」) されることができる。このように、非カノニカルポリマー単位の分析においてカノニカルポリマー単位として認識することにより、初期変換は、例えば、シグナルの分析に存在するいずれかの誤差が非体系的になる結果になるなど、シグナルにより多くの情報を提供する方法を提供することができ、それによって、推定精度の向上をもたらす。

10

【0026】

ラベルのセットは、各ポリマー単位タイプに関する少なくとも1つのラベル、および一連のポリマー単位内のブランクに関する少なくとも1つのラベルを含み得る。

【0027】

機械学習技術は、双方向回帰層であり得る少なくとも1つの回帰層を含むニューラルネットワークであり得る。

20

【0028】

ニューラルネットワークは、一連の重み分布を通るすべての経路にわたって重み分布のグローバル正規化を適用することができる。

【0029】

ニューラルネットワークは、回帰層の前に配置され、シグナルのウィンドウ化されたセクションの畳み込みを実行する複数の畳み込み層を含み得る。

【0030】

これらの重みは事後確率を表し得る。

【0031】

重み分布から一連のポリマー単位の推定値を導出するステップは、コネクションリスト時間分類を使用して実行することができる。

30

【0032】

重み分布から一連のポリマー単位の推定値を導出するステップは、各重み分布に関するラベルを導出すること、および導出されたラベルをランゲルス圧縮することを含み得る。

【0033】

重み分布から一連のポリマー単位の推定値を導出するステップが、重み分布を基準にして一連の重み分布を通るラベルの最も可能性の高い経路を推定することを含み得、一連のポリマー単位の推定値は、最も可能性が高いと推定されたラベルの経路から導出される。

【0034】

代替的に、重み分布から一連のポリマー単位の推定値を導出するステップが、各重み分布に関する最も可能性が高いラベルを推定すること、および一連の重み分布を通るラベルのフォワードとバックワードの経路を考慮に入れることを含み得、一連のポリマー単位の推定値が、最も可能性が高いと推定されたラベルから導出される。

40

【0035】

本発明の第2の態様によれば、ポリマーによって生じるシグナルをナノ細孔に対するポリマーの転位中に分析する方法が提供されており、本方法は、一連の重み分布を出力する機械学習技術を使用してシグナルを分析することであって、各重み分布が、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含んでいることと、重み分布から一連のポリマー単位の推定値を導出することと、を含み、ここで、

50

一連のポリマー単位の推定値を導出するステップは、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れ、ラベル間の少なくとも一方の遷移は、許容されないものとして表されており、他方の遷移は、許容されたものとして表される。

【0036】

本発明の第3の態様によれば、ポリマーによって生じるシグナルをナノ細孔に対するポリマーの転位中に分析する方法が提供されており、ポリマーは、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、本方法は、一連の重み分布を出力する機械学習技術を使用してシグナルを分析することであって、各重み分布が、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含み、一連のポリマー単位内の同じポリマー単位タイプの連続的インスタンスがランレングス符号化形式で表されていることと、重み分布から一連のポリマー単位の推定値を導出することと、を含む。

10

【0037】

第1の態様のいずれかの特徴は、本発明の第2および第3の態様に任意の組み合わせで適用することができる。

【0038】

さらに、本発明によれば、本方法は、コンピュータ装置で実行されるコンピュータプログラムによって実施されてもよく、または同様の方法を、本発明のいずれかの態様に実施するように構成された分析装置が提供されてもよい。

【0039】

さらに本発明によれば、ナノ細孔測定および分析システムが提供され得、ナノ細孔に対するポリマーの転位中にポリマーからシグナルを導出するように構成された測定システムと組み合わせられたような分析装置を含む。

20

【0040】

より良好な理解を可能にするために、本発明の実施形態は、添付の図面を参照して非限定的な例としてここで説明される。

【図面の簡単な説明】

【0041】

【図1】 ナノ細孔測定および分析システムの概略図である。

【図2】 時間の経過に伴う典型的なシグナルのプロットである。

30

【図3】 分析システムにおけるニューラルネットワークの図である。

【図4】 ニューラルネットワークのウィンドウングセクションの動作を示すシグナルの一部のプロットである。

【図5】 RNNの再帰層の図である。

【図6】 非再帰層の図である。

【図7】 一方向層の図である。

【図8】 「フォワード」と「バックワード」の再帰層を組み合わせた双方向再帰層の図である。

【図9】 「フォワード」と「バックワード」の再帰層を交互に組み合わせた代替双方向再帰層の図である。

40

【図10】 4つのタイプのポリヌクレオチドを表すラベル間の遷移に関する重み分布の表である。

【図11】 4つのタイプのポリヌクレオチドとブランクを表すラベル間の遷移に関する重み分布の表である。

【図12】 5つのタイプのポリヌクレオチド（そのうちの1つはメチル化C）とブランクを表すラベル間の遷移に関する重み分布の表である。

【図13】 重みが4つのタイプのポリヌクレオチドの各々について2つのラベルを含むラベル間の遷移に関する、重み分布の表である。

【図14】 重みがフリップフロップ表現を使用してホモポリマーを表している、重み分布の表である。

50

【図 1 5】 6-mer のシグナルモデルと、読み取りヘッドおよびシステムの他のコンポーネントに対するおおよその位置を使用する、4つの塩基の残留電流のプロットである。

【図 1 6】 重みがランレングス符号化表現を使用してホモポリマーを表す重み分布の表である。

【図 1 7】 重みが、ランレングス符号化表現の異なる定式化を使用してホモポリマーを表す重み分布の表である。

【図 1 8】 重み分布のさらなる重みの表であり、各可能なホモポリマータイプについて、可能な長さのセットにわたるカテゴリ分布を表す。

【図 1 9】 重み分布のさらなる重みの表であり、各可能なホモポリマータイプについて、可能な長さにわたるパラメータ化された分布を表す。

【図 2 0】 平均および分散パラメータの異なる値によって表される2つの分布のプロットである。

【図 2 1】 ホモポリマーを表すために使用することができる可能な分布の表である。

【図 2 2】 重み分布のさらなる重みの表であり、各可能なポリマー単位の対について、可能な長さのセットにわたるカテゴリ分布を表す。

【図 2 3】 重み分布のさらなる重みの表であり、各ポリマー単位の可能なトリプレットについて、可能な長さのセットにわたるカテゴリ分布を表す。

【図 2 4】 ラベルのセットが、修飾ポリマー単位に関するラベルを含むように拡張された重み分布の表である。

【図 2 5】 修飾の因数分解された表現における、ポリマー単位タイプの非修飾形態および修飾形態のさらなる重みの表である。

【図 2 6】 5-塩基の表現について、シグナルのプロットおよびそこから推定されたポリマー単位である。

【図 2 7】 ニューラルネットワークのデコーダによって実行される方法のフロー図である。

【図 2 8】 様々な復号アルゴリズムの定義である。

【図 2 9】 様々な復号アルゴリズムの定義である。

【図 3 0】 様々な復号アルゴリズムの定義である。

【図 3 1】 さらなる復号アルゴリズムの定義である。

【図 3 2】 フリップフロップ表現のための目的遷移行列を構築するためのアルゴリズムの定義である。

【図 3 3】 すべての経路にわたるトレーニングのための目的関数の定義である。

【図 3 4】 マルチステイ表現のための目的遷移行列を構築するためのアルゴリズムの定義である。

【図 3 5】 ランレングス符号化表現のための目的遷移行列を構築するためのアルゴリズムの定義である。

【図 3 6】 シグナルのプロットおよびそこから推定されるポリマー単位であり、長いホモポリマーの例を示している。

【図 3 7】 最良経路をトレーニングするための目的関数の定義である。

【図 3 8】 ファンクターの表である。

【図 3 9】 シグナルのプロットおよびそこから推定されるポリマー単位であり、フリップフロップ表現がシャープニングを使用してトレーニングされる一例を示している。

【図 4 0】 シャープニングを用いずにトレーニングされた表現、およびシャープニングを用いてトレーニングされた表現について、推定された一連のポリマー単位の基準に対する整列を示す表である。

【0 0 4 2】

図 1 は、測定システム 2 および分析システム 3 を含むナノ細孔測定および分析システム 1 を示す。測定システム 2 は、ナノ細孔に対するポリマーの転位中に、一連のポリマー単位を含むポリマーからのシグナルを導出する。分析システム 3 は、一連のポリマー単位の推定値を導出するためのシグナル分析の方法を実行する。

10

20

30

40

50

【0043】

一般に、ポリマーは、任意のタイプ、例えば、ポリヌクレオチド（または核酸）、タンパク質などのポリペプチド、または多糖であり得る。ポリマーは、天然または合成であり得る。ポリヌクレオチドは、ホモポリマー領域を含み得る。ホモポリマー領域は、5～15ヌクレオチドを含み得る。

【0044】

ポリヌクレオチドまたは核酸の場合、ポリマー単位はヌクレオチドであり得る。ポリヌクレオチドは、典型的には、デオキシリボ核酸（DNA）、リボ核酸（RNA）、cDNA、または、当該技術分野で既知の任意の合成核酸、例えば、ペプチド核酸（PNA）、グリセロール核酸（GNA）、トレオース核酸（TNA）、ロックド核酸（LNA）、もしくはヌクレオチド側鎖を有する他の合成ポリマーであり得る。PNA骨格は、ペプチド結合によって連結した繰り返しN-（2-アミノエチル）-グリシン単位で構成される。GNA骨格は、ホスホジエステル結合によって連結した繰り返しグリコール単位で構成される。TNA骨格は、ホスホジエステル結合によって共に連結した繰り返しトレオース糖で構成される。LNAは、上で考察されたように、リボース部分における2'酸素および4'炭素を接続する余分な架橋を有するリボヌクレオチドから形成される。核酸は、一本鎖、二本鎖、または一本鎖領域および二本鎖領域の両方を含み得る。ポリヌクレオチドは、DNAの1本の鎖にハイブリダイズされたRNAの一本鎖を含み得る。典型的には、cDNA、RNA、GNA、TNA、またはLNAは一本鎖である。

【0045】

ポリマー単位は、任意のタイプのヌクレオチドであり得る。ヌクレオチドは、天然に存在しても人工であってもよい。例えば、本方法を使用して、製造されたオリゴヌクレオチドの配列を確かめることができる。ヌクレオチドは、典型的には、核酸塩基、糖、および少なくとも1つのリン酸基を含む。核酸塩基および糖は、ヌクレオチドを形成する。核酸塩基は、典型的には、複素環である。好適な核酸塩基には、プリンおよびピリミジン、より具体的にはアデニン、グアニン、チミン、ウラシル、およびシトシンが含まれる。糖は、典型的には、五炭糖である。好適な糖には、リボース、およびデオキシリボースが含まれるが、これらに限定されない。ヌクレオチドは、典型的には、リボヌクレオチドまたはデオキシリボヌクレオチドである。ヌクレオチドは、典型的には、一リン酸、二リン酸、または三リン酸を含む。

【0046】

ヌクレオチドは、損傷した塩基または後成的塩基などの修飾された塩基であり得る。例えば、ポリヌクレオチドは、ピリミジンダイマーを含み得る。かかるダイマーは、典型的には、紫外線による損傷に関連付けられ、皮膚メラノーマの主因である。ヌクレオチドは、明確なシグナルを有するマーカーとして機能するようにラベリングまたは修飾されることができる。この技術は、例えば、ポリヌクレオチド中の無塩基単位またはスパーサーである塩基の不在を同定するために使用されることができる。本方法はまた、あらゆるタイプのポリマーにも適用することができる。

【0047】

ポリペプチドの場合、ポリマー単位は、天然に存在するかまたは合成されるアミノ酸であり得る。

【0048】

多糖の場合、ポリマー単位は単糖であり得る。

【0049】

特に、測定システム2がナノ細孔を含み、ポリマーがポリヌクレオチドを含む場合、ポリヌクレオチドは長く、例えば、少なくとも5kB（キロ塩基）、すなわち少なくとも5000ヌクレオチド、または少なくとも30kB（キロ塩基）、すなわち少なくとも30,000ヌクレオチド、または少なくとも100kB（キロ塩基）、すなわち少なくとも100,000ヌクレオチドであり得る。

【0050】

測定システム 2 の特性と得られたシグナルは以下の通りである。

【0051】

測定システム 2 は、1 つ以上のナノ細孔を含むナノ細孔システムである。単純なタイプでは、測定システム 2 は単一のナノ細孔しか有さないが、より実用的な測定システム 2 は、情報の並列収集を提供するために、典型的にはアレイ内の多くのナノ細孔を使用する。

【0052】

シグナルは、通常はナノ細孔を通る、ポリマーのナノ細孔に対する転位中に記録され得る。

【0053】

ナノ細孔は、典型的にはナノメートルのオーダーのサイズを有する細孔であり、そこを通るポリマーの通過を可能にし得る。

【0054】

ナノ細孔は、タンパク質細孔または固体細孔であり得る。細孔の寸法は、一度に 1 つのポリマーのみが細孔を転位することができるようなものであり得る。

【0055】

ナノ細孔がタンパク質細孔である場合には以下の特性を有し得る。

【0056】

生物学的細孔は、膜貫通タンパク質細孔であり得る。本発明に従って使用するための膜貫通タンパク質細孔は、 β バレル細孔または α ヘリックスバンドル細孔から誘導され得る。 β -バレル細孔は、 β 鎖から形成されるバレルまたはチャンネルを含む。好適な β -バレル細孔としては、 α -溶血毒、炭疽毒素、およびロイコシジンなどの β -毒素、ならびに *Mycobacterium smegmatis* ポリン (Msp)、例えば MspA、MspB、MspC、または MspD、リセニン、外膜ポリン F (OmpF)、外膜ポリン G (OmpG)、外膜ホスホリパーゼ A および *Neisseria* オートトランスポーターリポタンパク質 (NalP) などの細菌の外膜タンパク質／ポリンなどの他の細孔が挙げられるが、これらに限定されない。 α -ヘリックスバンドル細孔は、 α -ヘリックスから形成されるバレルまたはチャンネルを含む。好適な α -ヘリックスバンドル細孔は、内膜タンパク質および α 外膜タンパク質、例えば WZA および ClyA 毒素を含むが、これらに限定されない。膜貫通孔は、Msp または α -溶血素 (α -HL) に由来し得る。膜貫通孔はリセニンに由来してもよい。リセニン由来の好適な細孔は、WO 2013 / 153359 に開示されている。MspA 由来の好適な細孔は、WO 2012 / 107778 に開示されている。細孔は、WO - 2016 / 034591 に開示されているように、CsgG に由来し得る。この細孔は、DNA 折り紙細孔であり得る。

【0057】

タンパク質細孔は、天然に存在する細孔であり得、または変異体細孔であり得る。典型的な細孔は、WO - 2010 / 109197、Stoddart D et al., Proc Natl Acad Sci, 12; 106 (19): 7702-7、Stoddart D et al., Angew Chem Int Ed Engl. 2010; 49 (3): 556-9、Stoddart D et al., Nano Lett. 2010 Sep 8; 10 (9): 3633-7、Butler TZ et al., Proc Natl Acad Sci 2008; 105 (52): 20647-52、および WO - 2012 / 107778 に記載される。

【0058】

タンパク質細孔は、WO - 2015 / 140535 に記載されているタイプのタンパク質細孔の 1 つであり得、そこに開示されている配列を有し得る。

【0059】

タンパク質細孔は、生体膜などの両親媒性層、例えば脂質二重層に挿入することができる。両親媒性層は、親水および親油性の両方を有する、リン脂質などの両親媒性分子から形成された層である。両親媒性層は、単分子層または二重層であり得る。両親媒性層は、Gonzalez - Perez et al., Langmuir, 2009, 25, 1

10

20

30

40

50

0447-10450またはWO2014/064444に開示されているような共ブロックポリマーであってもよい。代替的に、タンパク質細孔は、例えば、WO2012/005857に開示されているように、固体層に設けられた開口に挿入され得る。

【0060】

ナノ細孔のアレイを提供するための好適な装置は、WO-2014/064443に開示されている。ナノ細孔は、各ウェルを横切って提供され得、電極は、各ナノ細孔を通る電流の流れを測定するためのASICと電氣的に接続された各ウェルに提供される。好適な電流測定装置は、WO-2016/181118に開示されるような電流感知回路を含み得る。

【0061】

ナノ細孔は、固体層に形成された開口を含み得、これは、固体細孔と呼ばれ得る。開口は、分析物が通過することがある、またはそこに入ることができる固体層に提供されるウェル、ギャップ、チャネル、トレンチ、またはスリットであり得る。このような固体層は、生物学的起源のものではない。換言すれば、固体層は、有機体または細胞等の生物学的環境、もしくは生物学的に利用可能な構造の合成的に製造されたバージョンに由来しないか、またはそれらから単離されない。固体層は、マイクロ電子材料、Si₃N₄、Al₂O₃およびSiO₂などの絶縁材料、ポリアミドなどの有機および無機ポリマー、テフロン（登録商標）などのプラスチックまたは二成分付加硬化型シリコンゴムなどのエラストマー、ならびにガラスを含むがこれらに限定されない有機材料および無機材料の両方から形成することができる。固体層はグラフェンから形成されてもよい。好適なグラフェン層は、WO-2009/035647、WO-2011/046706、またはWO-2012/138357に開示されている。固体細孔のアレイを準備するための好適な方法は、WO-2016/187519に開示されている。

【0062】

そのような固体細孔は、典型的には、固体層の開口である。開口は、ナノ細孔としての特性を強化するために、化学的またはその他の方法で変更することができる。固体細孔は、トンネル電極（Ivanov AP et al., Nano Lett., 2011 Jan 12; 11(1): 279-85）、または電界効果トランジスタ（FET）デバイス（例えば、WO-2005/124888に開示されている）などのポリマーの代替または追加の測定を提供する追加のコンポーネントと組み合わせて使用することができる。固体細孔は、例えば、WO-00/79257に記載されているものを含む既知のプロセスによって形成され得る。

【0063】

ナノ細孔は、固体細孔とタンパク質細孔のハイブリッドであり得る。

【0064】

測定システム2は、細孔に対して転位するポリマー単位に依存する特性の一連の測定を行う。一連の測定はシグナルから測定され得る。

【0065】

測定された特性は、ポリマーと細孔の間の相互作用に関連付けられ得る。このような相互作用は、細孔の狭窄領域で発生する可能性がある。

【0066】

測定システム2の1つのタイプでは、測定される特性は、ナノ細孔を通して流れるイオン電流であり得る。これらの、および他の電氣的測定は、Stoddart Det al., Proc Natl Acad Sci, 12; 106(19): 7702-7, Lieberman KR et al, J Am Chem Soc. 2010; 132(50): 17961-72、およびWO-2000/28312に記載されているような標準の単一チャネル記録装置を使用して行うことができる。代替的に電氣的特性の測定は、例えば、WO-2009/077734、WO-2011/067559、またはWO-2014/064443に記載されているようなマルチチャネルシステムを使用して実行されてもよい。

10

20

30

40

50

【0067】

イオン性溶液は、膜または固体層のいずれかの側に提供され得、これらのイオン性溶液は、それぞれの区画に存在し得る。目的のポリマー分析物を含むサンプルを膜の片側に追加し、例えば電位差または化学勾配の下で、ナノ細孔に対して移動することを可能にする。シグナルは、ポリマーの細孔に対する移動中に導出されることができ、例えば、ナノ細孔を通るポリマーの転位中に得られ得る。ポリマーは、ナノ細孔を部分的に転位し得る。

【0068】

ポリマーがナノ細孔を通して転位するときに測定値を得ることを可能にするために、転位の速度は、ポリマー結合部分によって制御することができる。典型的には、部分は、適用された電場と共に、またはそれに対して、ナノ細孔を通してポリマーを移動することができる。いくつかの実施形態では、部分は、例えば、部分が酵素、酵素活性である場合、または分子ブレーキとして使用する分子モーターであり得る。ポリマーがポリヌクレオチドであるいくつかの実施形態では、ポリヌクレオチド結合酵素の使用を含む、転位の速度を制御するために提案されたいくつかの方法がある。ポリヌクレオチドの転位の速度を制御するための好適な酵素には、ポリメラーゼ、ヘリカーゼ、エキソヌクレアーゼ、一本鎖および二本鎖結合タンパク質、ならびにジャイレースなどのトポイソメラーゼが含まれるが、これらに限定されない。他のポリマータイプの場合、そのポリマータイプと相互作用する部分をいくつかの実施形態で 사용할ことができる。いくつかの実施形態では、ポリマー相互作用部分は、WO-2010/086603、WO-2012/107778、および Lieberman KR et al, J Am Chem Soc. 2010; 132(50):17961-72)、ならびに電圧ゲート方式(Luan B et al., Phys Rev Lett. 2010; 104(23):238103)に開示されるいずれかであり得る。

【0069】

いくつかの実施形態では、ポリマー結合部分は、ポリマーの動きを制御するためにいくつかの方法で 사용할ことができる。部分は、適用された電場と共に、またはそれに対して、ナノ細孔を通してポリマーを移動することができる。いくつかの実施形態では、部分は、例えば、部分が酵素、酵素活性である場合、または分子ブレーキとして使用する分子モーターであり得る。いくつかの実施形態では、ポリマーの転位は、細孔を通るポリマーの移動を制御する分子ラチェットによって制御され得る。いくつかの実施形態では、分子ラチェットは、ポリマー結合タンパク質であり得る。ポリヌクレオチドに関して、ポリヌクレオチド結合タンパク質は、好ましくは、ポリヌクレオチドハンドリング酵素である。ポリヌクレオチドハンドリング酵素は、ポリヌクレオチドの少なくとも1つの特性と相互作用すること、かつそれを修飾することができるポリペプチドである。酵素は、個々のヌクレオチドまたはヌクレオチドのより短い鎖、例えばジ-またはトリヌクレオチドを形成するために、ポリヌクレオチドを切断することによってポリヌクレオチドを修飾してもよい。酵素は、それを配向するか、またはそれを特定の位置に移動させることによって、ポリヌクレオチドを修飾してもよい。ポリヌクレオチドハンドリング酵素は、それが、標的ポリヌクレオチドに結合し、かつ細孔を通るその移動を制御することができる限り、酵素活性を提示する必要がない。例えば、酵素を、その酵素活性を除去するために修飾してもよく、または酵素として作用することを防ぐ条件下で使用してもよい。そのような条件は以下でより詳しく考察される。

【0070】

いくつかの実施形態では、好ましいポリヌクレオチドハンドリング酵素は、ポリメラーゼ、エキソヌクレアーゼ、ヘリカーゼ、およびトポイソメラーゼ、例えば、ジャイレースである。いくつかの実施形態では、ポリヌクレオチドハンドリング酵素は、例えば、WO-2015/140535またはWO-2010/086603に記載されているタイプのポリヌクレオチドハンドリング酵素のうちの1つであり得る。

【0071】

ナノ細孔を介したポリマーの転位は、印加された電位の有無にかかわらず、シスからトラ

10

20

30

40

50

ンスまたはトランスからシスのいずれかで発生してもよい。転位は、転位を制御する可能性のある印加電位下で発生する可能性がある。

【0072】

二本鎖DNA上で進行的または前進的に作用するエキソヌクレアーゼを細孔のシス側に使用して、印加された電位下で、または逆電位下のトランス側で、残りの一本鎖を貫通接続することができる。同様に、二本鎖DNAを巻き戻すヘリカーゼも類似の様式で 사용할 ことができる。印加電位に対する鎖転位を必要とする配列決定用途の可能性もあるが、DNAは最初に逆電位または非電位下で酵素によって「捕捉され」なければならない。その後、結合に続いて電位が戻されると、鎖は、細孔をシスからトランスへと通り、電流によって拡張された立体構造で保持されることになる。一本鎖DNAエキソヌクレアーゼまたは一本鎖DNA依存性ポリメラーゼは分子モーターとして作用して、印加された電位に対して、新たに転位した一本鎖を、制御された様式で細孔を通してトランスからシスへと引き戻すことができる。代替的に、一本鎖DNA依存性ポリメラーゼは、細孔を通るポリヌクレオチドの動きを遅らせる分子ブレーキとして作用することができる。WO-2012/107778またはWO-2012/033524に記載されている任意の部分、技術または酵素を使用して、ポリマーの動きを制御することができる。

10

【0073】

しかしながら、測定システム2は、1つ以上のナノ細孔を含む代替のタイプのものであり得る。

【0074】

同様に、測定される特性は、イオン電流以外のタイプのものであることがある。代替タイプの特性のいくつかの例には、電気的特性および光学特性が含まれるが、これらに限定されない。蛍光の測定を含む好適な光学的方法は、J. Am. Chem. Soc. 2009, 131, 1652-1653によって開示されている。可能な電気的特性には、イオン電流、インピーダンス、トンネリング特性、例えば、トンネリング電流（例えば、Ivanov AP et al., Nano Lett. 2011 Jan 12; 11(1): 279-85に開示されている）、およびFET（電界効果トランジスタ）電圧（例えば、WO2005/124888に開示されている）が含まれる。1つ以上の光学特性が使用されることができ、任意選択で電気特性と組み合わせられる（Sonigvet al., Rev Sci Instrum. 2010 Jan; 81(1): 014301）。この特性は、ナノ細孔を流れるイオン電流などの膜貫通電流であり得る。イオン電流は典型的には、DCイオン電流であってもよいが、原則として、代替案は、AC電流フロー（すなわち、AC電圧の印加下で流れるAC電流の大きさ）を使用することもできる。

20

30

【0075】

いくつかのタイプの測定システム2では、このシグナルは、一連のイベントからの測定値を含むものとして特徴付けられることができ、各イベントは測定値の群を提供する。図2は、電流を測定する場合のそのようなシグナル10の典型的な例を示す。各イベントの測定値の群のレベルは類似しているが、多少の差異はある。これは、各ステップがイベントに対応するノイズの多いステップ波と考えることができる。イベントは、例えば、測定システム2の所与の状態または相互作用から生じる生化学的重要性を有し得る。これは、場合によっては、ラチェット様式で発生するナノ細孔を介したポリマーの転位から生じる可能性がある。しかしながら、このタイプのシグナルは、すべてのタイプの測定システムによって生成されるわけではなく、本明細書で説明される方法は、シグナルのタイプには依存しない。例えば、転位速度が測定サンプリング速度に近づく、例えば、ポリマー単位の転位速度の1倍、2倍、5倍、または10倍で測定が行われる場合、イベントは、より遅い配列決定速度、またはより速いサンプリングレートと比較して、より顕在化しないか、または存在しないことがある。

40

【0076】

さらに、イベントが存在する場合、通常、群内の測定数に関する先験的な知識はなく、こ

50

れは予測不能に変動する。これら変動および測定値の数の知識不足が要因で、一部の群を区別することが困難になる場合があり、これらは、例えば、群が短い場合、および／または2つの連続する群の測定値のレベルが互いに近接している場合などである。

【0077】

各イベントに対応する測定値の群は、典型的には、イベントの時間スケールにわたって一貫したレベルを持っているが、ほとんどのタイプの測定システム2では、短い時間スケールで変動し得る。このような変動は、例えば電気回路およびシグナル処理から生じる測定ノイズ、特に電気生理学の特定の場合の増幅器から生じ得る。測定される特性の程度が小さいため、このような測定ノイズは避けられない。このような変動は、測定システム2の基礎となる物理的または生物学的システムの固有の変動または広がり、例えば、ポリマー

10

【0078】

ほとんどのタイプの測定システム2は、多かれ少なかれ、そのような固有の変動を経験するであろう。所与のタイプの測定システム2について、両方の変動源が寄与するか、またはこれらのノイズ源の1つが支配的であり得る。

【0079】

ポリマー単位がナノ細孔に対して転位する速度である配列決定速度の増加に伴い、イベントはあまり目立たなくなり、したがって同定が困難になるか、または消える可能性がある。したがって、そのようなイベント検出に依存する分析方法は、配列決定速度が増すにつれて効率が低下し得る。

20

【0080】

しかしながら、本明細書に開示される方法は、そのようなイベントの検出に依存しない。以下に説明する方法は、比較的速い配列決定速度でも有効であり、この配列決定速度には、ポリマー転位が少なくとも毎秒10ポリマー単位、好ましくは毎秒100ポリマー単位、より好ましくは毎秒500ポリマー単位、またはより好ましくは毎秒1000ポリマー単位であるものが含まれる。

【0081】

サンプルレートとは、シグナルにおける測定値の速度である。通常、サンプルレートは配列決定速度よりも速くなる。例えば、サンプルレートは、100Hz～30kHzの範囲であり得るが、これは限定的ではない。実際には、サンプルレートは測定システム2の性質に依存し得る。

30

【0082】

場合によっては、本方法は、関連する一連のポリマー単位の測定値である複数の一連の測定値を使用することができる。例えば、複数の一連の測定値は、関連する配列を有する別個のポリマーの一連の測定値であり得、または関連する配列を有する同一ポリマーの異なる領域の一連の測定値であり得る。

【0083】

ポリヌクレオチドの場合、複数の一連のポリマー単位は相補的であることによって関係し得るので、一連のポリマー単位の1つはテンプレートと呼ばれ、それに対して相補的である他の一連のポリマー単位は相補体と呼ばれる。この場合、テンプレートおよび相補体の測定値は、任意の適切な技術を使用して取得することができ、例えば、ポリヌクレオチド結合タンパク質を使用して、またはポリヌクレオチドサンプル調製を介して順次取得される。好適な方法には、WO-2010/086622またはWO-2013/014451に記載されている方法が含まれる。

40

【0084】

単一の一連のポリマー単位に関連する本明細書で開示された方法のいずれも、例えば、WO-2010/086622またはWO-2013/014451に記載される方法を使用することによって、テンプレートおよび相補体などの複数の一連の測定値に適用され得る。

50

【0085】

一連の測定値は、分析システム3によって分析される生シグナルを形成する。生シグナルは、分析システム3に供給される前に、または分析システム3での初期段階として、例えばノイズを低減するためにフィルタリングされる前に、測定システム2で前処理され得る。このような場合には、以下の分析は、前処理されたシグナルに対して実行される。

【0086】

分析システム3は、測定システム2に物理的に関連付けられてもよく、測定システム2に制御シグナルを提供することもできる。その場合、測定システム2および分析システム3を含むナノ細孔測定および分析システム1は、WO-2008/102210、WO-2009/07734、WO-2010/122293、WO-2011/067559、またはWO2014/04443のいずれかに開示されるように構成され得る。

10

【0087】

代替的に、分析システム3は、別個の装置に実装され得、その場合、測定値の列は、任意の好適な手段、典型的にはデータネットワークによって、測定システム2から分析システム3に転送される。例えば、1つの便利なクラウドベースの実装は、分析システム3が、インターネットを介して入力シグナル11が供給されるサーバーになることである。

【0088】

分析システム3は、コンピュータプログラムを実行するコンピュータ装置によって実装されてもよく、専用のハードウェアデバイス、またはそれらの任意の組み合わせによって実装されてもよい。いずれの場合も、この方法で使用されるデータは、分析システム3のメモリに記憶される。

20

【0089】

コンピュータプログラムを実行するコンピュータ装置の場合、コンピュータ装置は、任意のタイプのコンピュータシステムであり得るが、典型的には、従来の構造である。コンピュータプログラムは、任意の好適なプログラミング言語で書くことができる。コンピュータプログラムは、任意のタイプであり得るコンピュータ可読記憶媒体に記憶され得る。例えば、コンピューティングシステムのドライブに挿入可能であり、磁氣的、光学的または光磁氣的に情報を記憶し得る記録媒体、ハードドライブなどのコンピュータシステムの固定記録媒体、またはコンピュータのメモリであり得る。

【0090】

コンピュータ装置が専用のハードウェアデバイスによって実装されている場合、任意の好適なタイプのデバイス、例えば、FPGA（フィールドプログラマブルゲートアレイ）またはASIC（特定用途向け集積回路）を使用することができる。好ましい実施形態では、コンピュータプログラムの一部は、グラフィックスプロセッシングユニット（GPU）などの並列計算に適したハードウェアを使用して実装され得る。

30

【0091】

ナノ細孔測定および分析システム1を使用する方法は、以下のように実行される。

【0092】

最初に、シグナル10は、測定システム2を使用して導出される。例えば、ポリマーは、例えば細孔を通して、細孔に対して転位させられ、ポリマーが転位する間にシグナルが導出される。ポリマーの転位を可能にする条件を提供することにより、ポリマーを細孔に対して転位させることができ、その結果、転位が自発的に起こり得る。転位する間にシグナル10が導出される。

40

【0093】

次に、分析システム3は、次に説明するように、シグナル10の分析方法を実行する。

【0094】

分析システム3は、ニューラルネットワーク20を使用してシグナルを分析する。再帰型ニューラルネットワーク20のパラメータは、以下でさらに説明するトレーニング中に値を取得するが、このように再帰型ニューラルネットワークは、いずれかの特定形式の測定値、または独自の特性を有する測定システム2に依存しない。例えば、再帰型ニューラル

50

ネットワークは、 $k - m e r$ に依存する測定値に依存しない。

【0095】

好適なニューラルネットワーク20の一例が図3に示されており、ウィンドウイングユニット30、CNN（畳み込みニューラルネットワーク）40、RNN（リカレントニューラルネットワーク）50、およびデコーダ80を含み、これらは以下のようにシグナルを連続的に処理する。

【0096】

ウィンドウイングユニット30は、例えば図4に示されるように、シグナル10の連続的なウィンドウ化セクション31を導出するためにシグナル10をウィンドウ化する。ウィンドウ化セクション11は、CNN40に供給される。

【0097】

ウィンドウ化セクション31は、長さ32、および連続的なウィンドウ化セクション31間のストライド33を有し、これらは両方とも、時間またはシグナル10のサンプル数でカウントされることができる。ストライド33は、単一のサンプルか、または複数のサンプルであってもよい。ストライド33が単一のサンプルである場合、エッジ効果を無視すると、シグナル10内のサンプルと同数のウィンドウ化セクション31が存在する。ストライド33が単一のサンプルよりも大きい場合、ウィンドウイングユニット30はダウンサンプリングを実行し、シグナル10内のサンプルよりもウィンドウ化セクション31が少なくなる。典型的には、ストライド33は長さ32よりも短く、ウィンドウ化セクション10はシグナル10内でオーバーラップするようになっている。

【0098】

例として、長さ32は4.75ミリ秒であり得、ストライドは0.5ミリ秒であり得る。

【0099】

さらなる例として、長さは19個のサンプルであり得、ストライドは2個のサンプルであり得る。

【0100】

CNN40は、少なくとも1つの畳み込み層を含む。少なくとも1つの畳み込み層は、各ウィンドウ化セクション11の畳み込みを実行して、各ウィンドウ化セクション31に関する特徴ベクトル41を導出する。これは、シグナル内で明らかになり得るイベントに関係なく行われるため、そのようなイベントが明らかであるか、または明らかでないシグナルにも、もしくは前処理中にイベントが提供されているシグナルにも同様に適用可能になっている。特徴ベクトル41は、RNN50に供給される。

【0101】

CNN40は、以下で考察されるようにRNN50と一緒にトレーニングされる。

【0102】

CNN40は、任意の形態をとることができる。

【0103】

一例では、CNN40は、重み W およびバイアス b 、ならびにアクティベーション関数 g を有するアフィン変換によって画定される単一の畳み込み層であり得る。ここで $I_{t-j:t+k}$ が、両端を含む、 $t-j \sim t+k$ の測定値を含む生シグナル20の測定ウィンドウを表し、 O_t は、出力特徴ベクトルである。

$Y_t = A I_{t-j:t+k} + b$ アフィン変換

$O_t = g(Y_t)$ アクティベーション

【0104】

双曲線正接は好適なアクティベーション関数であるが、正規化線形単位（ReLU）、指数線形単位（ELU）、ソフトプラス単位、およびシグモイド単位を含むがこれらに限定されない、より多くの代替案が当該技術分野で知られている。複数の畳み込み層もまた、使用されることができる。

【0105】

別の例では、CNN40は、参考文献1のCNNと同じ形式をとることができる。

10

20

30

40

50

【0106】

説明したように、真っ直ぐな畳み込みネットワークには、生のシグナルで検出された特徴の正確な位置に依存するという欠点があり、これは特徴間の間隔にも依存することを意味する。依存性は、最初の畳み込みによって生成された特徴ベクトルの出力配列を、その入力の順序統計に作用する第2の「プーリング」ネットワークへの入力として使用することで軽減できる。

【0107】

例として、プーリングネットワークが単層ニューラルネットワークである場合、次の方程式は、出力が入力ベクトルにどのように関連するかを示す。 f を入力特徴のインデックスとすると、 A_f は特徴 f の重み行列であり、 S をその入力の順序統計の一部またはすべてを返すファンクターとする。

10

【数1】

$$y_t = b + \sum_f A_f S(I_{f,t-j,t+k}) \quad \text{アフィン変換}$$

$$O_t = g(y_t) \quad \text{アクティベーション}$$

【0108】

このような層の有用でありながら計算効率の高い例の1つは、入力特徴の数と同じサイズの特徴ベクトルを返すものである。その要素は、それぞれの特徴ごとに取得される最大値である。ファンクター S_M に最後の順序統計のみを返させると、その入力で取得された最大値であり、そして U_f を、その $(f, 1)$ 要素の単位値以外ゼロで完全に構成され（単一の列）行列とする。

20

【数2】

$$y_t = \sum_f U_f S_M(I_{f,t-j,t+k}) \quad \text{アフィン変換}$$

$$O_t = y_t \quad \text{アクティベーションは適用されない}$$

【0109】

行列 U_f が非常にスパースであるため、計算効率の理由のために、行列乗算は、暗黙的に実行されることができ：ここで $\sum_f U_f x_f$ の効果は、出力特徴ベクトル x_f の要素 f を設定することである。

30

【0110】

畳み込みおよび／またはプーリングは、 n 番目の位置（ n のストライド）ごとにそれらの出力を計算するだけで実行できるため、出力をダウンサンプリングする。ダウンサンプリングは、同様の精度を達成するためにネットワークの残りの部分がより少ないブロック（より高速な計算）を処理する必要があるため、計算の観点から有利であり得る。

【0111】

畳み込み層のスタックを追加すると、上記の問題の多くが解決される。畳み込みによって学習された特徴検出は、システムについて追加の仮定を作製することなく、ナノ細孔固有の特徴検出器および要約統計量の両方として機能することができる。機能の不確実性は、様々な機能の相対的な重みによってネットワークの残りの部分に渡されるため、さらに処理することでこの情報を考慮に入れて、不確実性のより正確な予測および定量化を行うことができる。

40

【0112】

RNN50は、一連の重み分布を出力する。RNN50は、少なくとも1つの回帰層52を含み、この回帰層または各回帰層の後にフィードフォワード層53が続いている。図5は、単一の回帰層52の場合のRNNを示しているが、一般に、任意の複数の回帰層52およびその後のフィードフォワード層53が存在し得る。このことは単位アーキテクチャ

50

の柔軟な選択をもたらす。これらの層は、異なるパラメータを有し得、異なるサイズであり得、または様々な単位タイプでも構成され得る。

【0113】

この回帰層または各回帰層52は、双方向であることが好ましく、各入力特徴ベクトルの影響がRNNを介して両方向に伝播することを可能にする。代替の好ましい実施形態は、交互の方向に配置された複数の一方向回帰層、例えば、リバース、フォワード、リバース、フォワード、リバースの連続的方向に配置された層を含む。これらの双方向アーキテクチャにより、RNN50が、HMMでは利用できない様式で情報を蓄積および伝播することを可能にする。回帰層の追加的利点は、モデル化するために、例えば、反復手順を介して、シグナルからモデルに（またはその逆に）正確なスケールリングを必要としないことである。

10

【0114】

フィードフォワード層53でのサブサンプリングについて、個別のアフィン変換が各列のフォワード層とバックワード層の出力ベクトルに適用され、その後に積算が行われ、このことは、入力と出力の連結によって形成されたベクトルにアフィン変換を適用することと等しくなっている。次いで、アクティベーション関数が要素ごとに結果の行列に適用される。

【0115】

再帰層52は、ここで説明されるように、いくつかのタイプのニューラルネットワーク単位を使用することができる。単位のタイプは、「再帰」であるかどうかに応じて、2つの一般的なカテゴリに分類される。非再帰単位は配列の各ステップを独立して処理するが、再帰単位は配列で使用され、あるステップから次のステップに状態ベクトルを渡すように設計される。

20

【0116】

非再帰単位と再帰単位の違いを図式的に示すために、図6は非再帰単位61の非再帰層60を示し、図7～9はそれぞれの非再帰単位64～66の3つの異なる層62～64を示す。図6～図9の各々で、矢印はベクトルが渡される接続を示し、分割された矢印は複製されたベクトルであり、結合された矢印は連結されたベクトルである。

【0117】

図6の非再帰層60では、非再帰単位61は、分割または連結されない別個の入力および出力を有する。

30

【0118】

図7の再帰層62は、再帰単位65の出力ベクトルが分割され、再帰層内の次の再帰単位65に一方向に渡される一方向の再帰層である。

【0119】

それ自体は個別の単位ではないが、図8および9の双方向再帰層63および64は各々、より単純な再帰単位66および67からそれぞれ作られた繰り返し単位のような構造を有する。

【0120】

図8の双方向再帰層において、双方向再帰層63は、再帰単位66の2つのサブ層68および69からなり、図7の一方向再帰層62と同じ構造を有するフォワードサブ層68と、図7の一方向再帰層62から時間が逆転したかのように逆転して状態ベクトルを1つの単位66から前の単位66に渡す構造を有するバックワードサブ層69になっている。フォワードおよびバックワードのサブ層68および69の両方が同じ入力を受け取り、対応する単位66からのそれらの出力が一緒に連結されて、双方向再帰層63の出力を形成する。フォワードサブ層68内の単位66とバックワードサブ層69内の単位との間に接続がないことに留意されたい。

40

【0121】

図9の代替双方向再帰層64は、同様に、再帰単位67の2つのサブ層70および71からなり、図7の一方向再帰層62と同じ構造を有するフォワードサブ層68、ならびに、

50

図 7 の単方向再帰層 6 2 から時間が逆転したかのように逆転するバックワードサブ層 6 9 である。同様に、フォワードサブ層 6 8 およびバックワードサブ層 6 9 は、同じ入力を受け取るが、図 8 の双方向再帰層 6 3 とは対照的に、フォワードサブ層 6 8 の出力は、バックワードサブ層 6 9 の入力であり、バックワードサブ層 6 9 の出力は、双方向再帰層 6 4 の出力を形成する（フォワードサブ層 6 8 およびバックワードサブ層 6 9 は逆転されることができる）。

【0122】

図 9 に示す双方向再帰層 6 4 の一般化は、複数の「フォワード」および「バックワード」再帰サブ層で構成される再帰層のスタックとなり、ここでは各層の出力は次の層の入力になっている。

【0123】

RNN 5 0 の双方向再帰層 5 2 は、図 8 の双方向再帰層 6 3 および図 9 の双方向再帰層 6 4 のいずれかの形態をとることができる。一般に、図 3 の双方向再帰層 3 4 は、非再帰層、例えば、図 6 の非再帰層 6 0、または一方向再帰層、例えば、図 7 の再帰層 6 2 によって置き換えることができる。しかしながら、双方向の再帰層 3 4 を使用することにより、パフォーマンスが向上する。

【0124】

次に、フィードフォワード層 5 3 について説明する。

【0125】

フィードフォワード層 5 3 は、それぞれのベクトルを処理するフィードフォワード単位 5 4 を備える。フィードフォワード単位 5 4 は、古典的なニューラルネットワークの標準的な単位であり、すなわち、アフィン変換が入力ベクトルに適用され、次に、非線形関数が要素ごとに適用される。フィードフォワード層 5 3 はすべて、非線形関数に双曲線正接を使用するが、ネットワークの全体的な精度をほとんど変化させずに他の多くの層を使用することができる。

【0126】

ステップ t における入力ベクトルが I_t である場合、アフィン変換のための重み行列およびバイアスはそれぞれ A および b であり、出力ベクトル O_t は以下である。

$$y_t = A I_t + b \quad \text{アフィン変換}$$

$$O_t = \tanh(y_t) \quad \text{非線形性}$$

【0127】

RNN 5 0 の重み分布は、グローバルに正規化される。これは、以下でより詳しく考察される。

【0128】

非再帰単位 6 2 および再帰単位 6 5 ~ 6 7 は、各イベントを独立して扱うが、次に説明するような形態を有する長短期記憶単位に置き換えることができる。

【0129】

長短期記憶 (LSTM) 単位は、Hochreiter and Schmidhuber, Long short-term memory, Neural Computation, 9 (8): 1735-1780, 1997 で紹介された。LSTM 単位は再帰単位であるため、配列の 1 つのステップから次のステップに状態ベクトルを渡す。LSTM は、単位がメモリセルであるという表記に基づく。メモリの内容を含む隠れ状態は、あるステップから次のステップに渡され、メモリの更新方法を制御する一連のゲートを介して操作される。1 つのゲートは、メモリの各要素をワイプ (忘れる) するかどうかを制御し、別のゲートは、新しい値に置き換えるかどうかを制御し、最後のゲートは、メモリのリードと出力を決定する。メモリセルを微分可能にするのは、概念的なコンピュータメモリセルのバイナリオン/オフ論理ゲートが、シグモイド関数によって生成される概念的な確率に置き換えられ、メモリセルの内容が期待値を表すことである。

【0130】

第 1 に LSTM の標準実装について説明し、次いで基本的な方法で実際に使用される「の

10

20

30

40

50

ぞき穴」の変更について説明する。

【0131】

標準的なLSTMは次の通りである。

【0132】

LSTM単位の様々な操作に関連する確率は、次の一連の方程式によって定義される。 I_t を、ステップtの入力ベクトルであるとし、 O_t を出力ベクトルとし、バイアス b_x と、それぞれ入力と前の出力に対する重み行列 W_{xI} および W_{xO} を有する、xでインデックス付けされたアフィン変換を行うと、 σ は非線形シグモイド変換である。

$$f_t = \sigma(W_{fI}I_t + W_{fO}O_{t-1} + b_f) \quad \text{忘却確率}$$

$$u_t = \sigma(W_{uI}I_t + W_{uO}O_{t-1} + b_u) \quad \text{更新確率}$$

$$o_t = \sigma(W_{oI}I_t + W_{oO}O_{t-1} + b_o) \quad \text{出力確率}$$

【0133】

上で定義された更新ベクトルがあり、 $\circ$ 演算子が要素ごとの（アダマール）乗算を表すとすると、内部状態 S_t を更新し、新しい出力を決定する数式は次のようになる。

【数3】

$$\begin{aligned} v_t &= \tanh(W_{vI}I_t + W_{vO}O_{t-1} + b_v) && \text{更新する値} \\ S_t &= S_{t-1} \circ f_t + v_t \circ u_t && \text{メモリセルを更新する} \\ O_t &= \tanh(S_t) \circ o_t && \text{メモリセルからのリード} \end{aligned}$$

【0134】

のぞき穴の変更は次の通りである。

【0135】

「のぞき穴」の変更（Gers and Schmidhuber, 2000）は、LSTMアーキテクチャにいくつかの追加接続を追加し、メモリセルの隠れ状態を「のぞき見」（通知）するための確率の忘却、更新、および出力を可能にする。ネットワークの更新式は上記の通りであるが、 P_x を隠れ状態に等しい長さの「のぞき見」ベクトルとすると、確率ベクトルの3つの式は次のようになる。

【数4】

$$\begin{aligned} f_t &= \sigma(W_{fI}I_t + W_{fO}O_{t-1} + b_f + P_{f-} \circ S_{t-1}) && \text{忘却確率} \\ u_t &= \sigma(W_{uI}I_t + W_{uO}O_{t-1} + b_u + P_{u-} \circ S_{t-1}) && \text{更新確率} \\ o_t &= \sigma(W_{oI}I_t + W_{oO}O_{t-1} + b_o + P_{o-} \circ S_t) && \text{出力確率} \end{aligned}$$

【0136】

代替的に、非再帰単位62および再帰単位65～67は、以下のような形態を有するゲート付き再帰単位に置き換えることができる。

【0137】

ゲート付き再帰単位（GRU）は実行が速いことがわかっているが、最初は精度が低くなっている。GRUのアーキテクチャは、LSTMほど直感的ではなく、隠れ状態と出力の分離が不要で、「忘却」ゲートと「入力ゲート」も組み合わせられている。

【数5】

$$\begin{aligned} o_t &= \sigma(W_{oI}I_t + W_{oS}S_{t-1} + b_o) && \text{出力確率} \\ u_t &= S_{t-1} \circ \sigma(W_{uI}I_t + W_{uS}S_{t-1} + b_u) && \text{状態から更新} \\ v_t &= \tanh(W_{vI}I_t + W_{vR}u_t + b_v) && \text{更新する値} \\ S_t &= (1 - o_t) \circ S_{t-1} + o_t \circ v_t && \text{更新状態} \end{aligned}$$

【0138】

イベントと同じ数の列が出力されるが、双方向レイヤーの存在のためにその内容がイベントの入力セット全体によって通知される可能性があるため、ネットワークへの入力で各列が単一のイベントで同定されると想定するのは正しくない。入力イベントと出力列の間の

対応は、トレーニングセット内のシンボルでラベル付けされる方法によって行われる。

【0139】

別の例では、RNN50は、参考文献1のRNNと同じ形式をとることができる。

【0140】

ここで、RNN50によって出力される一連の重み分布51について考察する。

【0141】

重み分布51は、一連の重み分布を形成するために、連続する時間ステップに関して出力されている。時間ステップは、原則として、シグナル10のサンプル周期と同じ長さであり得るが、ニューラルネットワーク20におけるオーバーサンプリングのために、通常、シグナル10のサンプル周期よりも長い。しかしながら、規則的長さの時間ステップは、例えば、ウィンドウイングユニット30のストライド13に対応しており、これは、イベント呼び出しが実行されるシステムとは対照的であり、したがって、連続する重み分布51間の時間ステップは、可変である連続して検出されたイベントの長さに対応する。

【0142】

一般に、重み分布51は、連続的なポリマー単位がナノ細孔に対して転位する速度よりも速い速度で出力されており、すなわち、ポリマー単位よりも多くの重み分布51が存在する。各一連のポリマー単位に対応する複数の重み分布51は、先験的には未知である。

【0143】

各重み分布51は、複数の重みを含む。重みは事後確率を表す。重みは実際の事後確率であり得、または、より一般的には、実際の確率ではないが事後確率を表す重みであり得る。一般に、重みが実際の確率ではない場合、重みの正規化を考慮して、原則として事後確率が、そこから決定され得る。

【0144】

RNN50は、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたってラベル間の遷移に関する重みを出力する。したがって、遷移に関する重みは、それらの遷移の事後確率を表す。ポリマー単位よりも多くの重み分布51があることから、いくつかの表現では、ラベルから同じラベルへの遷移が可能であり、したがって、重み分布51は、そのような遷移に関する重みを含むこと、すなわち、「遷移」という単語が、ラベルを変更する必要があることでも、追加のポリマー単位を放出する必要があることを意味するものでもないことが理解される。

【0145】

RNN50によって出力される重み分布51の様々な例が、以下に示される。これらの例の各々は、ポリマー単位がポリヌクレオチドであり、ポリマー単位タイプが4つの塩基A、C、GおよびTである場合を指している。上で考察されたように、本発明の方法は、より多くのタイプのポリヌクレオチドおよび／またはヌクレオチドではないポリマー単位に等しく適用可能であり、そのためこれらの例はそれに応じて汎用化されることができる。例の各々において、重み分布51は、ラベル間の遷移を表す重みを含む。したがって、重みは W_{ij} と表記されており、 i は、そこから遷移が発生するラベルのためのインデックスであり、 j は、そこへの遷移が発生するラベルのためのインデックスである。したがって、重み W_{ij} は、ラベル i からラベル j への遷移の重みである。図面のそれぞれにおいて、行はそこから遷移が発生するラベル i に対応しており、列はそこへの遷移が発生するラベル j に対応する。

【0146】

このような構成の一例は、生成される重み分布の重みの数に等しい、出力（最終）層内のフィードフォワード要素の数で構成されたRNNである。図7～9はまた、RNNのユニットからの出力の数を示す例を提供しており、これらの構成のいずれか1つ以上がRNN内に存在し得、出力の数が生成される重み分布の重みの数に等しくなるようになることが理解される。

【0147】

図10および11は、RNN50によって出力され得る重み分布51の2つの例を示す。

【0148】

図10の例では、A、C、G、Tとして示されている4つの塩基の各々に関する単一のラベルがある。すべての遷移が許容され、それにより、ラベルの各々からラベルの各々への16の遷移に関する合計16の重み w_{ij} がある。

【0149】

図10の例は、ホモポリマーの適切な表現を提供しておらず、これは一連のポリマー単位内の同じタイプの複数のポリマー単位の連続である。これは、あるラベルから同じラベルへの移行が、特定のポリマー単位タイプの同じインスタンスと特定のポリマー単位タイプのさらなるインスタンスとを区別しないことからである。結果として、あるラベルから同じラベルへの一連の遷移は、一連のポリマー単位（すなわち、単一のポリマー単位といずれかの長さの同じポリマー単位タイプのホモポリマーとの両方）の任意数（1つ以上）のインスタンスを表す。

10

【0150】

しかしながら、図11は、図10の表現を拡張することでホモポリマーの表現を改善する一例であり、ラベルのセットが、(i) 4つの塩基の異なる1つを各々表す単一ラベル、および(ii) 一連のポリマー単位内のブランクを表すラベルを含むようになっている。許容されたすべての遷移が図11の例に示されており、それにより、各ラベルから各ラベルへの25の遷移に関する合計25の重み w_{ij} がある。この表現では、ブランクラベルは、それらが同じタイプであっても、一連の塩基（ポリマー単位）の2つのインスタンス間の分離を表す。

20

【0151】

言い換えると、ポリマー単位の配列では、データ測定値のウィンドウが分析されることがあり得るが、そのデータウィンドウではポリマー単位間の遷移は存在しない。この場合、遷移は、前のラベルから「ブランク」ラベルへの遷移として表され得、これは、ポリマー単位の新規インスタンスがウィンドウ内で遷移されなかったことを表す。

【0152】

いくつかの実施形態では、ポリマー単位が分離しているようにブランクのいずれかの側のポリマー単位を取り扱うために、ポリマー単位の決定された配列内でブランクが存在しなければならないという点で、ブランクは強制的なものとして扱われ得る。例えば、次のように生成されたラベルの配列が、ブランクが「-」で表される：A A A - - - Aの場合、これはポリマー単位の実際のシーケンスは、= A Aとして決定される。最初の3つのインスタンスの各々の「A」ラベルは、同じ実際のポリマー単位「A」のインスタンスとして扱われるが、最後の「A」ラベルは、最初の3つの「A」から2つのブランクラベルで区切られるため、別個のものとして扱われる。

30

【0153】

いくつかの実施形態では、ブランクが、ポリマー単位の間スペースおよびラベルの繰り返しを表すという点で、ブランクは任意選択のものとして扱われ得る。例えば、次のようなブランクが「-」で表される生成されたラベルの配列：A A A - - - Aの場合、これはポリマー単位の実際のシーケンス=A A A Aに決定される。「A」ラベルの最初の3つのインスタンスのそれぞれが、別個のポリマー単位として扱われ、ブランクのラベルは、これらのユニットと最後の「A」ラベルとの間のスペースとして機能する。

40

【0154】

ラベル間の遷移に関する重みを使用したRNN50の出力におけるこの表現は、参考文献1とは対照的であり、RNNが、4つのタイプのポリヌクレオチド（つまり、塩基C、G、A、およびT）の各々を表す4つのラベルで構成されるラベルのセットにわたるラベル、およびブランクを表すラベルに関する事後確率（重みの特定の例）を出力する。ラベル間の遷移に関する重みを使用する表現は、一連のポリマー単位の推定の精度を向上させる追加情報が提供されるため、ラベルに関する重みを使用する表現よりも利点をもたらす。これは、重みが一連のポリマー単位を通る可能な経路に関する情報を提供するのに対し、ラベル自体に関する重みは、さらなる分析を目的とした他のラベルとの関係に関する情

50

報を失うためである。したがって、追加情報がポリマー単位を推定するステップに提供され、このことが復号の精度を向上させている。

【0155】

さらに、この表現は、許容された遷移と許容されていない遷移を表現できる。つまり、ラベルは、ラベル間の1つ以上の遷移が許容されず、他の遷移が許容される様式で、可能なポリマー単位タイプを表すことができる。その場合、重み分布51は、許容された遷移に関する重みを含む。重み分布51は、許容されない遷移に関するヌル重みを含み得る。

【0156】

ヌル重みは、RNN50によって出力された重み分布51において、重みがないことになり得る。図面に示されている例では、ヌル重みは重みがないことで示されているが、代わりに以下の代替方法が適用され得る。

【0157】

代替的に、ヌル重みは、RNN50の実装を容易にするために、RNN50によって出力される重み分布51内に存在する重みであり得るが、公称値を有している。そのような公称値は、ゼロ値または意味のない大きさを有する値であり得、以下に説明するように、デコーダ80によって実行される推定に影響を与えるようになっている。代替的に、そのような公称値は、RNNによって出力された重み分布51内に存在するが、例えば、以下に説明するような遷移行列を使用することによってデコーダ80によって無視される値であり得る。

【0158】

このいくつかの例は以下の通りである。

【0159】

許容遷移と非許容遷移が発生する第1の例は、可能なポリマー単位タイプのセットに、ポリマー単位の既知のシーケンスに常に現れるポリマー単位タイプが含まれる場合である。この場合、既知のシーケンスに従った遷移は許容され、既知のシーケンスに反する遷移は許容されていない。ポリヌクレオチドについて、この一例は、脊椎動物の5-メチルシトシンがグアニンに先行するシトシン（「CpG」）でのみ発生し、このことが、可能な遷移をさらに制限するために使用されることができ、それにより、より少ないRNN50からの重みが必要とされる。つまり、CpGメチル化は、メチル化C（ C^M として本明細書で表される）に常にGが続くことになり、そのため C^M が既知の配列 $C^M G$ で常に生じる。図12は、これを表すために使用される重み分布51の例である。重み分布51は、図11の重み分布から適合されており、メチル化Cを表すラベルを、4つのタイプのポリヌクレオチド（すなわち、塩基C、G、AおよびT）を表す4つのラベルとブランクを表すラベルに追加する。この場合には、 C^M からA、CまたはTへの遷移が許容されず、それにより、これらの遷移にはヌル重みがあり、つまり、重み W_{61} 、 W_{62} 、および W_{64} は、重み分布でヌルになっている。これにより、RNNはメチル化C塩基に関するより良い情報をもたらすことができ、このことがメチル化C塩基の推定の精度を向上させる。

【0160】

任意選択で、 C^M から C^M への遷移の重みがヌルにされることができ、このことはステイの場合にあり得る。これは、測定中に同定され得るが、この遷移はCpGメチル化がメチル化CにGが常につき、つまり $C^M G$ 配列になることから、配列の一部を形成しない。このことのさらなる例は、以下で説明するフリップフロップ表現であり、ここで、改変されたフリップまたは改変されたフロップからグアニンまたは改変されたフロップのラベルへの遷移は、RNN50から必要な重みの数を60から52に減らすことができる（すべての可能性のある遷移に必要な100の重みを参照のこと）。必要とされるネットワーク出力の量の低減は別として、可能であるこれらに遷移を制限することは、本方法が、推定における誤差と誤検出の改変呼び出しの両方である不可能なコンテキストでの改変を伴ったポリマー単位タイプの推定を生成することを防止する。

【0161】

第2の例は、ラベルのセットが改変され、各ポリマータイプが、単一ラベルの代わりに複

10

20

30

40

50

数のラベルで表される表現である。例えば、ラベルのセットは、各ポリマー単位タイプに関する、第1および第2のラベルを含むことができ、ここで、第1のラベルは、ポリマー単位タイプのインスタンスの開始を表し、第2のラベルは、ポリマー単位タイプのインスタンス内のステイを表している。上述のように、「ステイ」は、本方法が、連続する重み分布に関連するラベルが変化しないと判断する状況を表し、このことは、ポリマー単位の同じインスタンスに対応する2つの重み分布と見なされ得る。本明細書では、この例を「マルチステイ」と称することにする。ステイは異なるラベルで表されるため、このことは表現を改善する。このことはポリマー単位の推定精度を改善する。

【0162】

このマルチステイ表現は、一部の遷移が許容され、一部が許容されないという結果をもたらす。例えば、第1のラベル、例えば「A」は、同じポリマー単位タイプに関して、第2のラベル、例えばAS内に、または異なるポリマー単位タイプの第1のラベル内にのみ遷移することが許容される。具体的には、以下の遷移が許容され、また許容されておらず、それらは以下の通りである。

- a) 各第1のラベルから任意の他のポリマー単位タイプの第1のラベルへの遷移が許容され、各第1のラベルから同じポリマー単位タイプの第1のラベルへの遷移が許容されていない。
- b) 各第1のラベルから同じポリマー単位タイプの第2のラベルへの遷移が許容されている。
- c) 各第1のラベルから任意の他のポリマー単位タイプの第2のラベルへの遷移が許容されていない。
- d) 各第2のラベルから同じポリマー単位タイプの第1のラベルへ、または任意の他のポリマー単位タイプの第1ラベルへの遷移が許容されている。
- e) 各第2のラベルから同じポリマー単位タイプの第2のラベルへの遷移が許容されている。
- f) 各第2のラベルから任意の他のポリマー単位タイプの第2のラベルへの遷移が許容されていない。

【0163】

上述の例示的なマルチステイ表現スキームは、ブランクが上述のように強制的または任意選択であると見なされ得る図11のスキームと同様の様式で「強制的」スキームであると見なされ得る。そのために、上記と同様のスキームが、第1のラベルが同じ第1のラベルに遷移することが許容されるマルチステイ表現について想定され得ることが理解されるであろう。かかるスキームは、「任意選択の」マルチステイ表現スキームと見なすことができる。

【0164】

図13は、このタイプの表現を実装するために図10のものから適合された重み分布51の一例を示している。したがって、図13では、ラベルのセットは、A、C、G、およびTとして示される4つのタイプの塩基に関する4つの第1のラベル、ならびにAS、CS、GSおよびTSとして示される4つのタイプの塩基に関する4つの第2のラベルを含む。本明細書では、上付き文字S（「ステイ」を意味する）は、塩基の同じタイプに関して第2のラベルを第1のラベルから区別するために使用され、ステイを表している。図13に示すように、許容される遷移と許容されない遷移を考慮すると、以下の重みが存在するか、またはヌルになっており、それらは以下の通りである。

- a) 各第1のラベル（例えば、A）から任意の他のポリマー単位タイプ（例えば、C、G、およびT）の第1のラベルへの遷移が許容され、各第1のラベル（例えば、A）から同じポリマー単位タイプの第1のラベルへの遷移が許容されており、そのため、左上の象限内にすべての重みが存在している。
- b) 各第1のラベル（例えば、A）から同じポリマー単位タイプ（例えば、AS）第2のラベルへの遷移が許容されており、そのため、右上の象限の重みにW15、W26、W37、およびW48が存在している。

c) 各第 1 のラベル (例えば、A) から任意の他のポリマー単位タイプ (例えば、C S、G S、T S) の第 2 のラベルへの遷移が許容されておらず、そのため、 W_{15} 、 W_{26} 、 W_{37} 、および W_{48} を除いた右上の象限の重みがヌルになっている。

d) 各第 2 のラベル（例えば、A S）から同じポリマー単位タイプ（例えば、A）第 1 のラベルへの遷移、または任意の他のポリマー単位タイプ（例えば、C、G、T）の第 1 のラベルへの遷移が許容されており、そのため、左下の象限内にすべての重みが存在している。

e) 各第 2 のラベル (例えば、A S) から同じポリマー単位タイプ (例えば、A S) の第 2 のラベルへの遷移が許容されており、右下象限の重みに W_{55} 、 W_{66} 、 W_{77} 、および W_{88} が存在している。

f) 各第2のラベル（例えば、A S）から任意の他のポリマー単位タイプ（例えば、C S、G SおよびT S）の第2のラベルへの遷移が許容されておらず、そのため、W₅₅、W₆₆、W₇₇、およびW₈₈を除いた右下の象限内の重みがヌルになっている。

【 0 1 6 5 】

マルチステイ表現は、上記のメチル化Cの表現、または実際には、既知のポリマー単位の配列で常に現れるポリマー単位タイプの同様の表現と組み合わせることができる。

【 0 1 6 6 】

ここでホモポリマーの表現を検討する。ホモポリマーは、一連のポリマー単位内の同じタイプのポリマー単位の連続したインスタンスの配列である。

【 0 1 6 7 】

第2のラベル（例えば、A S）からポリマー単位（例えば、A）の同じタイプの第1のラベルへの遷移は、ポリマー単位の同じタイプの第2のインスタンスを表現することから、ホモポリマーは上で考察されたマルチステイ表現によって適切に表現される。例えば、一連のラベルA A S A S A A S A A S A S A A S A S A S A Sは、長さが4ポリマー単位のホモポリマーを表現し、連続するラベルAまたはA Sの数は任意であり、実際には変化している。しかしながら、推定の精度は、表現を適合させることによって改善されることができ、例えば以下のように、ラベルが符号化形式でホモポリマーを表すようになっている。

【 0 1 6 8 】

符号化形式のホモポリマーの第 1 の表現は、「フリップフロップ」表現と呼ばれ、以下のようになっている。

【 0 1 6 9 】

分析方法の出力が固定長フラグメントをオーバーラップすることの利点の1つは、オーバーラップの量を使用して、ポリマー単位の遷移が発生したかどうか、そしてその場合に発生した回数を判断できることである。オーバーラップに依存する分析方法は、ホモポリマーのようなポリマーの複雑度の低い領域では失敗し、この領域では、オーバーラップが曖昧になり得（例えば、AAA→AAAは、0、1、2、またはそれ以上のAホモポリマーの遷移となり得る）、異なる表現が望ましい。フリップフロップ表現では、ラベルは、各ポリマー単位タイプについての複数のラベルを含むことによってホモポリマーを表し、各ポリマー単位タイプについての複数のラベルは、一連のポリマー単位内のポリマー単位タイプの連続するインスタンスを表す。通常、ポリマー単位タイプごとに2つのラベルがあり、参照しやすいように「フリップ」および「フロップ」と称され得る。

【 0 1 7 0 】

したがって、固定長フラグメントに復号するのではなく、復号のフリップフロップ方式は、ポリマー単位の配列を「フリップ」と「フロップ」のラベルの配列として表すが、次の制限があり、ホモポリマーは「フリップ」ラベルで開始する必要がある、次いで、「フリップ」と「フロップ」のラベル間で終了するまで交互に切り替わる。フリップフロップ表現は、隣接するラベルと同じラベルがないことを確実にし、ホモポリマーを使用した1つのユニットの転位（フリップからフロップへの変更、またはその逆）は、転位なし（フリップからフリップへ、またはフロップからフロップへ）と常に区別可能になっている。例

として、一連のポリマー単位 C A A T A C C T T T A A A A A A A A G A A A C T T T T A G C T C は、C A A F T A C C F T T F T A A F A A F A A F A A F G A A F A C T T F T T F として表され、ここで、ポリマー単位 X のフリップラベルは X で表されており、対応するフロップラベルは X F で表される。

【0171】

フリップフロップ符号化の下では、1つの転位は常に転位がないものと区別可能であり、より多くのポリマー単位の転位は、依然として曖昧であり得る。したがって、連続する重み分布 5 1 によって表されるラベルに関して、塩基 A の 2 つのラベルが A (フリップである) および A F (フロップである) である場合、一連のラベル A A A A A A A F A F A F A F A A A は、長さが 3 ポリマー単位のホモポリマーを表し、連続するラベル A または A F の数は任意であり、実際には変動している。原則として、ポリマー単位タイプごとに 3 つ以上のラベルが存在する場合があるが、2 つのラベルで十分である。

10

【0172】

ポリマー単位の各タイプの複数のラベルは、所定の周期的順序を有し得る。ポリマー単位の各タイプ、フリップとフロップの 2 つのラベルの一例では、所定の周期的順序は、第 1 のポリマー単位が常にフリップとなり、その後、フリップとフロップが交互になるものであり得る。したがって、ラベル間のいくつかの遷移は、所定の周期的順序によって許容されており、間の他の遷移は、所定の周期的順序によって許容されていない。重み分布において所定の周期的順序で許容されていない遷移についてヌル重みがあるが、一方で当然のこととして、所定の周期的順序で許容されている遷移についての重みがある。

20

【0173】

所定の周期的順序が、最初のポリマー単位が常にフリップされ、その後フリップとフロップが交互に行われるような上記の例では、ポリマー単位のいずれか所与のタイプのフリップからポリマーの他のタイプのフロップへの遷移が許容されておらず、同様に、ポリマー単位のいずれか所与のタイプのフロップからポリマーのいずれかの他のタイプのフロップへの遷移は許容されていない。

【0174】

図 1 4 は、このタイプのフリップフロップ表現の重み分布 5 1 の一例を示す。したがって、図 1 4 では、ラベルのセットは、A、C、G、および T として示されている 4 つのタイプの塩基に関する 4 つの第 1 のラベル (フリップ)、ならびに A F、C F、G F および T F として示されている 4 つのタイプの塩基に関する 4 つの第 2 のラベル (フロップ) を含む。図 1 4 に示すように、許容される遷移と許容されない遷移を考慮すると、以下の重みが存在するか、またはヌルになっており、それらは以下の通りである。

30

a) 各第 1 のラベル (フリップ、例えば、A) からポリマー単位 (例えば、A、C、G、および T) のすべてのタイプの第 1 のラベル (フリップ) への遷移が許容されており、そのため、左上の象限内にはすべての重みが存在している。

b) 各第 1 のラベル (フリップ、例えば、A) からポリマー単位の同じタイプ (フリップ、例えば、A F) の第 2 のラベルへの遷移が許容されており、そのため、右上の象限には W 1 5、W 2 6、W 3 7、および W 4 8 が存在している。

c) 各第 1 のラベル (フリップ、例えば、A) からポリマー単位の任意の他のタイプ (例えば、C F、G F、T F) の第 2 のラベルへの遷移は許容されておらず、そのため、右上の象限では W 1 5、W 2 6、W 3 7、および W 4 8 を除いた重みはヌルになっている。

40

d) 各第 2 のラベル (フロップ、例えば A F) からポリマー単位のすべてのタイプ (フリップ、例えば A、C、G、および T) の第 1 のラベルへの遷移は許容されており、左下の象限内にはすべての重みが存在している。

e) 各第 2 のラベル (フロップ、例えば、A F) からポリマー単位の同じタイプ (フロップ、例えば、A F) の第 2 のラベルへの遷移は許容されており、そのため、右下の象限内には W 5 5、W 6 6、W 7 7、および W 8 8 が存在している。

f) 各第 2 のラベル (フロップ、例えば、A F) からポリマー単位の任意の他のタイプ (フロップ、例えば、C F、G F および T F) の第 2 のラベルへの遷移は許容されておらず

50

、そのため、右下の象限ではW 5 5、W 6 6、W 7 7、およびW 8 8を除いた重みはヌルになっている。

【0 1 7 5】

ポリマー単位の転位速度に対する測定が行われる速度に応じて、ポリマーが測定と測定の間に複数の回数で転位する場合、複数の単位の明白な転位が観察される場合がある。これが発生する可能性が高い場合は、各ポリマー単位の追加の冗長ラベル（「フラップ」、「フリップ」、「フレップ」など）が加えられることができ、追加的単位の存在を表わされることができ、例えば、フリップからフラップへ移る配列は、中間フロップラベルの存在を意味している。

【0 1 7 6】

符号化形式のホモポリマーの第2の表現は、ランレングス符号化表現と呼ばれており、以下の通りである。

【0 1 7 7】

フリップフロップ表現は、長いホモポリマーを介して呼び出すことができるが、交互のラベルのパスとして実行し、複数の接続された呼び出しを行う必要がある。より長いホモポリマーの場合、観察されたシグナルの平坦化は、ポリマーがナノ細孔に対して転位して、ラベル内の各変化の位置がより恣意的になることから、シグナルが変化するとき明確な時間がもはやないことを意味し得る。図15は、T-フリップとT-フロップの両方の累積的な証拠が高いにもかかわらず、重みがT-フリップまたはT-フロップの間で分割されている一例示的領域に対するこの特異性の喪失の例を示す。

【0 1 7 8】

このように、ホモポリマーを交互のラベルの列として表すのではなく、代わりに、ホモポリマー全体が、ポリマー単位タイプに関してラベルで表されることができる。したがって、カノニカル配列またはそのフリップフロップ符号化を呼び出すようにRNN50をトレーニングするのではなく、RNN50は、一連のポリマー単位のランレングス符号化を呼び出すようにトレーニングされている。例えば、カノニカル配列T A A T T C A A A C T T T T T T C T G A T A A G C T G G Tのランレングス符号化は、T A 2 T 2 C A 3 C T 7 C T G A T A 2 G C T G 2 Tであり、ランレングスはこの塩基に従い、1つの長さは暗黙的である。可能な限り最長での実行が常に行われるため、同じ塩基での実行で隣接する実行はない。

【0 1 7 9】

ランレングス符号化表現の第1の定式化では、ラベルは、各ポリマー単位タイプの異なるランレングスのラベルを含む。図16は、このような重み分布の一例を示している。この例では、A、C、G、Tとして示されている4つの塩基に関して、およびA 2、A 3などとして示されている各塩基のホモポリマーに関して単一のラベルがある。ホモポリマーのすべての可能な長さを収容するためには多数のラベルがあり、1つのタイプの塩基のホモポリマーに関するラベルから同じタイプの塩基のホモポリマーであるが異なる長さのものへの遷移を除いて、すべての遷移が許容されており、そのため、ラベルの数の二乗にほぼ等しい数のラベル間のほとんどの遷移に関する多数の多数の重み w_{ij} があり（他の可能な遷移スキームも代替的に実装され得る）、扱いが困難になっている。

【0 1 8 0】

大きなゲノムの長いホモポリマーは、偶然に予想されるよりも頻繁に発生するため、定型化した配列の間に遭遇する可能性のあるすべてのホモポリマーの長さを表すために必要なラベルの数は非常に多くなっている。ネットワークによって出力される重みは、ホモポリマーラベル間の遷移を明示的にパラメータ化するため、トレーニングする必要のあるパラメータの数が多いことと、それらが弱く結合されていることの両方の理由からデータのトレーニングが問題になる。ラベルの中でラベルをシャッフルすると（例えば、A 6 → A 3、T 2 → T 7、G 8 → G 1）、同じ性能にトレーニングされる同等のモデルが得られるため、長さ4および6のホモポリマーのトレーニング例は、長さ5のホモポリマーについてモデルに通知しない。

【0181】

ランレングス符号化の代替的な好ましい定式化は、重み分布 51 をいくつかの従属分布に因数分解することである。したがって、ラベルは、各ポリマー単位タイプに関するラベルを含み、重み分布 51 は、遷移に関する重みに加えて、各ポリマー単位タイプのランレングス圧縮ホモポリマーの可能な長さにはわたるさらなる重みを含む。遷移重みは、RNN 50 によって出力され、ランレングス圧縮シーケンス全体の分布、つまり、すべての長さが削除されたランレングス符号化配列、およびポリマー単位が指定された実行（ラン）の長さ（レングス）の条件付き分布の個別のセットを記述する。

【0182】

ランレングス符号化のこの好ましい定式化では、RNN によって出力される重み分布 51 は、異なるポリマー単位タイプ間の遷移を表すために、図 10 に示される形式の重みを含み得る。上で考察されたように、この場合、ラベルから同じラベルへの一連の遷移は、ポリマー単位（すなわち、任意の長さの同じポリマー単位タイプの単一のポリマー単位またはホモポリマー）の一連の任意の数のインスタンスを表す。

【0183】

RNN によって出力される重み分布 51 の代替として、ランレングス符号化のこの好ましい定式化は、各ポリマータイプが、単一のラベルではなく第 1 および第 2 のラベルによって表されるラベルのセット、例えば、第一のポリマー単位タイプに関するラベル A とラベル A_H に対して定義され得る。したがって、上付き文字 H は、同じポリマー単位タイプに関して第 1 のラベルから第 2 のラベルを区別するために使用され、「ホールド」を効果的に表している。

【0184】

これは、以下の点を除いて、図 13 に示すマルチステイ表現に類似している。上述のように、マルチステイ表現では、第 2 のタイプのラベルから第 1 のタイプのラベルへの遷移（例えば、A_S から A_H）が許容され、同じポリマー単位タイプのさらなるインスタンスの発生を表している。結果として、ホモポリマーは、一連のラベルによって表されており、そこでは、A A S A S A A S A A S A S A A S A S A S A S が、長さ 3 のポリマー単位のホモポリマーを表す上の例にあるように、第 1 のタイプのラベルが繰り返されている。対照的に、現在の表現では、許容された遷移が異なるため、第 2 のタイプのラベルから第 1 のタイプのラベルへの遷移（例えば、A_H から A_H）は許容されていない。結果として、1 つのポリマー単位タイプの単一のインスタンスおよびポリマー単位の同タイプの任意の長さのホモポリマーはすべて、第 1 のタイプのラベルおよび任意数の第 2 のタイプのラベルを含む一連のラベルによって表されている。例えば、A A H A H A H A H A H A H A H A H A H A H は、単一の塩基 A または塩基 A のホモポリマーを表し得る。具体的には、このことは以下のように達成されている。

- a) 各第 1 のラベルから任意の他のポリマー単位タイプの第 1 のラベルへの遷移は許容されているが、各第 1 のラベルから同じポリマー単位タイプの第 1 のラベルへの遷移は許容されていない。
- b) 各第 1 のラベルから同じポリマー単位タイプの第 2 のラベルへの遷移が許容されている。
- c) 各第 1 のラベルから任意の他のポリマー単位タイプの第 2 のラベルへの遷移が許容されていない。
- d) 各第 2 のラベルから同じポリマー単位タイプの第 1 のラベルへの遷移は許容されていない。
- e) 各第 2 のラベルから他のポリマー単位タイプの第 1 のラベルへの遷移は許容されている。
- f) 各第 2 のラベルから同じポリマー単位タイプの第 2 のラベルへの遷移は許容されている。
- g) 各第 2 のラベルから他のポリマー単位タイプの第 2 のラベルへの遷移は許容されていない。

【0185】

図17は、図10の重み分布からこのタイプの表現を実施するために適合された、かかる重み分布51の一例を示している。したがって、図17では、ラベルのセットは、A、C、G、およびTとして示される4つのタイプの塩基に関する4つの第1のラベル、ならびにAH、CH、GHおよびTHとして示される4つのタイプの塩基に関する4つの第2のラベルを含む。図17に示すように、許容される遷移と許可されない遷移を考慮して、以下の重みが存在するか、ヌルになっている。

a) 各第1のラベル（例えば、A）から任意の他のポリマー単位タイプ（例えば、C、G、およびT）の第1のラベルへの遷移は許容されるが、各第1のラベル（例えば、A）から同じポリマー単位タイプ（例えば、A）の第1のラベルへの遷移は許容されておらず、そのため、W₁₁、W₂₂、W₃₃、およびW₄₄を除いて2つの左上象限内にある重みはヌルになっている。

10

b) 各第1のラベル（例えば、A）から同じポリマー単位タイプの第2のラベル（例えば、AH）への遷移は許容されており、そのため、右上象限内の重みにはW₁₅、W₂₆、W₃₇、およびW₄₈が存在している。

c) 各第1のラベル（例えば、A）から任意の他のポリマー単位タイプ（例えば、CH、GH、TH）の第2のラベルへの遷移は許容されておらず、そのため、W₁₅、W₂₆、W₃₇、およびW₄₈を除いて右上象限にある重みはヌルになっている。

d) 各第2のラベル（例えば、AH）から同じポリマー単位タイプの第1のラベル（例えば、A）への遷移は許容されておらず、そのため、右上象限にある重みW₅₁、W₆₂、W₇₃、およびW₈₄はヌルになっている。

20

e) 各第2のラベル（例えば、AH）から任意の他のポリマー単位タイプ（例えば、C、G、およびT）の第1のラベルへの遷移は許容されており、そのため左下象限にある重みは、W₅₁、W₆₂、W₇₃、およびW₈₄を除いたものが存在している。

f) 各第2のラベル（例えば、AH）から同じポリマー単位タイプ（例えば、AH）の第2のラベルへの遷移は許容されており、そのため、右下象限内の重みにはW₅₅、W₆₆、W₇₇およびW₈₈が存在している。

g) 各第2のラベル（例えば、AH）から任意の他のポリマー単位タイプ（例えば、CH、GH、TH）の第2のラベルへの遷移は許容されておらず、そのため、W₅₅、W₆₆、W₇₇、およびW₈₈を除いて右下象限にある重みはヌルになっている。

30

【0186】

したがって、所与のポリマー単位タイプの一連のラベルは、常に第1のラベルの単一のインスタンスで始まり、次いで第2のラベルの1つ以上のインスタンスで始まる。例えば、一連のラベルA、AAH、AAHAAHなどのいずれか（任意の数のラベルAHを伴う）は、任意数のポリマー単位のインスタンスの列（つまり、単一のポリマー単位または任意長の同じポリマー単位タイプのホモポリマー）を表す。

【0187】

上述のように、図10の例はホモポリマーの適切な表現を提供せず、同じことが図17の例にも当てはまる。しかしながら、ホモポリマーは、ランレングス圧縮ホモポリマーの可能な長さにわたるさらなる重みによって表されている。ここでは、そのようなさらなる重みのいくつかの可能性が説明され、それらのそれぞれは、図10の形式または図17の形式の重みと組み合わせて適用され得る。

40

【0188】

さらなる重みについての第1の可能性は、それらが、可能なポリマー単位タイプのごとに、ホモポリマーの可能な長さのセットにわたる重みのカテゴリカル分布を含むことである。可能な長さはカテゴリであり、RNN50出力は各カテゴリに重みを割り当てる。一般に、各カテゴリは単一のホモポリマーの長さを表すことができ、またはカテゴリの一部またはすべてがホモポリマーの長さの範囲を表すことができる。カテゴリは、所定長さを超えるすべてのホモポリマーを表すカテゴリを含むことができる。カテゴリは等間隔である必要はない。

50

【0189】

図18は、この第1の可能性に従ったそのようなさらなる重みの一例を示している。この例では、4つの塩基A、C、G、Tのそれぞれの可能な長さごとに重み l_{ij} があり、これら塩基はインデックス*i*によってインデックス付され、長さはインデックス*j*によってインデックス付けされている。この例では、各カテゴリは単一の長さに対応しているが、代わりに、各カテゴリを長さの範囲に対応させて、カテゴリの数を減らすこともできる。図18に示されるさらなる重みは、ラベル間の遷移のための重みと共に重み分布51の一部を形成し、これは、例えば、図10～13のいずれかに示されるような、上述のような形式をとることができる。

【0190】

カテゴリ分布は、すべてのホモポリマーラベル間の遷移を完全に指定するよりも少ないパラメータを必要とし、基礎となるランゲルス圧縮ゲノムが推定され得るようにするが、依然としてトレーニングデータを十分に活用しない弱い結合の問題があり、長いホモポリマーのトレーニングを困難にしている。

【0191】

さらなる重みの第2の可能性は、それらが可能なポリマー単位タイプごとに、ホモポリマーの可能な長さに対応するパラメータ化された分布のパラメータを含むことである。所与のポリマー単位のホモポリマーが所与の長さのいずれかとなる確率を計算するために、このようなパラメータが使用され得る。

【0192】

図19は、この第2の可能性に従ったそのようなさらなる重みの一例を示している。この例では、A、C、G、Tとして示された塩基の4つのタイプそれぞれにインデックス*i*でインデックス付けされた重み p_{ij} がある。これら重みは、分布の*j*個のパラメータである P_1 、 P_2 、…、 P_j を示しており、パラメータはインデックス*j*によってインデックス付けされている。パラメータは、分布を表すいずれかのパラメータであり得る。一般に、分布に応じて、*j*はいずれかの複数の値を有することができる。図19に示されるさらなる重みは、ラベル間の遷移のための重みと共に、重み分布51の一部を形成し、これは上述のような、例えば、図10～13のいずれかに示されるような形をとることができる。

【0193】

例として、図20は、平均と分散の2つのパラメータの異なる値によってそれぞれ表されるホモポリマーの長さの2つの異なる分布の例を示している。

【0194】

ホモポリマーの長さにわたってパラメータ化された分布を使用する利点は、分布がホモポリマーの長さの事後分布として解釈され得ることであり、推定された長さに信頼を置くことができる。例えば、図20では、両方の分布がホモポリマーの長さの同じ事後平均推定値を示すが、信頼度は異なり、分散が大きい分布（左）は、分散が小さい分布（右）よりも信頼度が低くなっている。

【0195】

異なるホモポリマーの長さの予測は、すべて同じネットワーク出力のセットを介して行われるため、この予測は以前よりもはるかに緊密に結合され、ネットワークが1つのホモポリマーの例から同様の長さの例に一般化することを可能にしている。

【0196】

ネットワークの出力と組み合わせて、多くの異なる確率分布が使用され得る。発生する可能性のあるホモポリマーの長さを表すことができる分布を選択することは有利であり、そのため、この分布は、潜在的な長さの大きなセット、または半無限のセットにわたってサポートする必要がある。所与のホモポリマーの長さにおいて高い信頼度（低分散）と低い信頼度（高分散）の両方表すパラメータの値が存在することも望ましい。負の二項分布または幾何分布が使用されることがあり、信頼性の高いケースと低いケースを区別することができない。

10

20

30

40

50

【0197】

幾何分布の分散は平均の関数であり、負の二項には追加の自由度があり、その分散は常に平均より大きくなければならない。これらの両方の基準を満たす分布は、 $[0, \infty]$ にわたってサポートする連続分布を離散化することで見つけることができる。離散化の1つの方法は、ホモポリマーが長さ L である確率を、 $L \sim L+1$ までの密度関数の積分に設定するか、代替的に $L=0$ を適切に処理して $L-0.5 \sim L+0.5$ に設定する。

【0198】

好ましくは、離散化された分布は、明示的な累積密度関数を有する。このような密度の例は、ワイブル分布、対数ロジスティック分布、対数正規分布、ガンマ分布であるが、これらに限定されない。使用されるパラメトリック分布またはその離散化された対応物の平均、最頻値、分散の明示的な表現がある場合は、有利であるが、必須ではない。

10

【0199】

図21は、すべてが

【数6】

N

上でサポートを有している、ホモポリマーの長さを表すためのいくつかの適切な離散分布を示している。 $\Gamma(\alpha)$ はガンマ関数、 $\gamma(\alpha, \beta)$ は不完全ガンマ関数、 $\Phi(x)$ は標準正規分布の累積分布である。

【0200】

20

第1および第2の可能性のそれぞれにおいて、さらなる重みは、可能なポリマー単位タイプごと、すなわち、ホモポリマーの可能なポリマー単位タイプごとに定義される。これは効果的であるが、さらなる改善が変更によってもたらされることができ、この変更には、(a) 所与のポリマー単位タイプおよび先行するポリマー単位タイプの可能な対についてのさらなる重み、(b) 所与のポリマー単位タイプおよび後続のポリマー単位タイプの可能な対についてのさらなる重み、または(c) 所与のポリマー単位タイプ、先行するポリマー単位タイプ、および後続のポリマー単位タイプの可能なトリプレットについてのさらなる重みがある。

【0201】

この変更により、重みは、例えば、第1の可能性に応じたホモポリマーの可能な長さのセットにわたる重みのカテゴリ分布、または第2の可能性に応じたホモポリマーの可能な長さにわたるパラメータ化された分布のパラメータ、など同じ形式を取るが、重みの数は増加する。ケース(a)および(b)について、重みの数が3倍に増やされ、それにより、各可能なポリマー単位タイプの代わりにそれぞれの可能な対のための分布を設定し、例えば、4つのタイプの塩基 $\{A, C, G, T\}$ の代わりに、12の対の塩基 $\{(A, C), (A, G), (A, T), (C, A), (C, T), (C, G), (G, A), (G, C), (G, T), (T, A), (T, C), (T, G)\}$ の分布を設定する。例として、図22は、そのようなさらなる重みの一例を示しており、これは、各ポリマー単位タイプの対に対して設定されたホモポリマーの可能な長さにわたるパラメータ化された分布のパラメータを含んでいる。これは、ケース(a)および(b)に対応し、この対は、ケース(a)では、所与のポリマー単位および先行するポリマー単位タイプであり、ケース(b)では、所与のポリマー単位タイプおよび後続のポリマー単位タイプである。パラメータの形式自体は、図19の場合と同じであり、特定のポリマー単位のホモポリマーが任意の長さであるという確率を計算するために同様に使用されることができる。

30

40

【0202】

同様に、ケース(c)の場合、重みの数を9倍に増やして、各可能なトリプレットについての分布、例えば、4つのタイプの塩基ではなく36個の塩基のトリプレットを設定する。例として、図23は、そのようなさらなる重みの例を示しており、これはポリマー単位タイプの各トリプレットに対して設定されたホモポリマーの可能な長さにわたるパラメータ化された分布のパラメータを含む。これはケース(c)に対応しており、トリプレット

50

は、特定のポリマー単位、先行するポリマー単位タイプ、および後続のポリマー単位タイプになっている。パラメータの形式自体は、図 19 の場合と同じであり、特定のポリマー単位のホモポリマーが任意の長さであるという確率を計算するために同様に使用されることができる。

【0203】

この変更は、長いホモポリマーのエッジを識別する能力が、先行および／または後続のポリマー単位に依存して変化する可能性があるという認識に基づいて、精度を向上させる。例えば、塩基 T から塩基 A のホモポリマーへの遷移は、塩基 C から塩基 A のホモポリマーへの遷移よりもはるかに識別が容易である。したがって、様々な対またはトリプレットの分布を表す異なるさらなる重みを提供することは、ポリマー単位をより正確に推定することができる表現をもたらす。

10

【0204】

重み分布 51 のいくつかの従属分布への同様の因数分解を使用して、ポリマーの他の特性を表すことができる。一例は、非修飾形態および修飾形態を有するポリマー単位タイプ、例えば、塩基の 1 つのタイプおよび同じ塩基の修飾されたタイプを含み得るポリヌクレオチドの表現である。

【0205】

DNA の天然鎖には、例えば、5-メチルシトシンや 6-メチルアデニンなどの修飾塩基が含まれており、それらの存在と位置は、一連のナノ細孔測定を使用して検出できる。フリップフロップおよび他の表現は、容易に一般化して、ラベルのセットを塩基 A、C、G および T から拡張することによって修飾を喚起することができるようになり、修飾塩基を表すための追加のラベル（例えば、C^M は修飾された C を表す）を含む。

20

【0206】

図 24 は、重み分布の一例を示しており、ラベルのセットが、修飾塩基に関してラベル C^M を追加的に含むように拡張されている。同様に、追加のラベル C^M は、図 10、12 ~ 14、または 16 に示される重み分布 51 のいずれかでラベルのセットに追加され得る。

【0207】

ラベルのアルファベットのこの拡張は、当技術分野で説明された以前の方法でも使用でき、これは特定時間のシグナルが塩基の固定長フラグメントによって表されることができると仮定するが、これらは、考慮される修飾の数が増えるにつれて、ネットワークには、固定長の各塩基間の可能な遷移について出力が必要であることから尺度が正しくない。例えば、追加の修飾塩基が許容されると、4 つのカノニカル塩基、3125 (= 5⁵) で構成される長さ 5 の断片については、1024 の可能な組み合わせ (4⁵) があり、2 つの修飾が許容されると、7776 (= 6⁵) の可能な組み合わせがある。RNA には 100 を超える既知の修飾があり、そのため、フラグメントベースモデルは急速に増加する処理量を要する。

30

【0208】

ポリマー単位の非修飾形態は、カノニカルポリマー単位として説明され得、ポリマー単位の修飾形態は、非カノニカルポリマー単位として説明され得る。修飾（または非カノニカル）ポリマー単位は、通常、対応する非修飾（カノニカル）ポリマー単位とは異なる形でシグナルに影響を及ぼす。

40

【0209】

参照が行われ、参照により本明細書に組み込まれる、2019 年 9 月 4 日に出願された国際特許出願第 PCT/GB2019/052456 号には、カノニカルおよび非カノニカル塩基に関する教示が含まれており、これは本明細書に開示される本発明の方法のいずれかに適用され得る。

【0210】

国際特許出願第 PCT/GB2019/052456 号は、本発明の方法のいずれかに適用され得る非カノニカル塩基の例を開示する。

【0211】

50

国際特許出願第 P C T / G B 2 0 1 9 / 0 5 2 4 5 6 号はまた、本発明の方法のいずれかと組み合わせて使用され得る 1 つ以上の非カノニカルポリマー単位を含むポリマーを調製および分析する方法を開示する。

【 0 2 1 2 】

非限定的な例として、本発明の方法のいずれかと組み合わせて使用され得る国際特許出願第 P C T / G B 2 0 1 9 / 0 5 2 4 5 6 号に開示される方法は、カノニカルポリマー単位（例えば、アミノ酸）の比率を対応する非カノニカルポリマー単位（例えば、アミノ酸）に、例えば、化学変換または酵素変換によって、非決定論的な方法で変換する。その場合、一連のポリマー単位の推定値を導出（「呼び出し」）するとき、非カノニカル塩基は、対応するカノニカル塩基であると推定（「呼び出される」）されることができる。これには、国際特許出願第 P C T / G B 2 0 1 9 / 0 5 2 4 5 6 号の図 1 8 b ~ 1 8 k を参照して説明された方法が含まれる。

10

【 0 2 1 3 】

カノニカルおよび非カノニカルのポリマー単位がターゲットポリマーに非決定論的に組み込まれることから、ポリマー単位の基本的な配列は不明であり、鎖ごとに異なる。各鎖が代替的塩基を含む場合でも、依然として関連するカノニカル配列があり、任意の代替物のタイプと場所を推測するのではなく、これを直接呼び出す利点がある。言い換えれば、ターゲットポリマーに追加のポリマー単位があるにもかかわらず、分析はカノニカル値のみをシグナルに帰属させ決定された配列が A、C、G、および T の群からの塩基で構成されるようになっている。このように、非カノニカルポリマー単位を分析においてカノニカルポリマー単位として認識することにより、初期変換は、例えば、シグナルの分析に存在するいずれかの誤差が非体系的になるなど、シグナルにより多くの情報を提供する方法を提供することができ、それによって、推定精度の向上をもたらす。

20

【 0 2 1 4 】

フリップフロップおよび同様の表現ははるかに扱いやすく、それは、遷移重みをパラメータ化するために、各時点で R N N 5 0 から出力される必要な重みの数は、フラグメントの長さに等しい累乗としてではなく、修飾の数で 2 次関数的に定められるためである（4 0 は 4 つのカノニカル塩基を出力し、6 0 は 1 つの追加の修飾された塩基を出力し、8 4 は 2 つの追加の修飾された塩基を出力する、など）。

【 0 2 1 5 】

ニューラルネットワーク 1 0 がフリップフロップ表現を使用する場合、正しい配列の確率を最大化するためにトレーニングが実行され、各リードに対して、推定配列を生成するためにさらに復号されなければならない条件付き確率場を生成する。使用される復号の方法は、リードと呼ばれる塩基の総数やその構成の要約統計量など、バルクメトリックで明らかになる不要なバイアスを最終呼び出しに導入することができる。同じ配列を有するか、または共通のサブ配列を含む鎖のリードから推定された配列が集合的に考慮される場合、さらなるバイアスが明らかになることがある。

30

【 0 2 1 6 】

この問題を低減するために、ペナルティ項がトレーニング済みのニューラルネットワーク 1 0 に組み込まれ、その出力を調整して対象のメトリックのパフォーマンスを向上させることができ、例えば、新しいポリマー単位を放出しないことに対応するすべての重み（同じ塩基内のフリップフリップ、またはフロップフロップ遷移）から定数を減算することは、呼び出されるポリマー単位の数を増やす一方で、特定のポリマー単位の比率は、そのアイデンティティの新しいポリマー単位を放出することになるすべての遷移に定数を追加することによって増やされることができる。

40

【 0 2 1 7 】

使用されるペナルティ項の値は、値のグリッド上の代表的なリードのセットの対象となるメトリックを計算することによって調整されることができ、代替的に、シンプレックス法などのより正式な最適化方法、または当技術分野で既知である他の多くの方法が使用され得る。ペナルティ項は、固定定数よりむしろ、リードに関する事前情報の関数であること

50

がある。

【0218】

ペナルティ項は、任意の層でニューラルネットワーク10に組み込まれることができるが、可能であれば最終層に組み込むことが好ましく、放出される遷移重みに直接影響し、これが最終的な推定シーケンスへの影響を直感的に理解できるという利点があり、そのため、フォームにペナルティを導く。

【0219】

ニューラルネットワーク10の出力の解釈を確率モデルとして保持するために、必須ではないが、「グローバル正規化」が実行される前にペナルティが組み込まれることが望ましい。

【0220】

多くの場合、カノニカル塩基の配列といずれかの修飾の存在の両方を正確に決定することは重要であり、修飾を推定しようとして、基になるカノニカル配列の推定に悪影響を与えることは望ましくない。これがどのように生じ得るかの一例は、カノニカルなシトシンと5-メチルシトシンの間の重みの分割であり、そのため、別の塩基が最も可能性の高い推定値になる。

【0221】

重み分割動作を防ぐために、RNNによって出力された重み分布51は、2つの従属分布に含められることができる。この場合、第1の分布は、上述の形態のいずれかをとり重み分布51であり、単一のラベルが非修飾および修飾形態を有するポリマー単位タイプを表し、第2の分布は、条件付き分布であり、非修飾および修飾形態のさらなる重みを含む。この表現は、任意の数の修飾形態について、かつ、いずれかの可能なポリマー単位タイプの修飾形態について拡張され得る。

【0222】

図25は、塩基Cの非修飾形態および同じ塩基C^Mの修飾形態を表すためのさらなる重みの一例を示す。この場合、さらなる重みは、塩基Cの非修飾形態に関する重みm₁と、塩基C^Mの修飾形態に関する重みm₂である。これは、図24に示されたタイプの重み分布51の代わりに適用され得る。さらなる重みは、ラベル間の遷移の重みと共に、重み分布51の一部を形成し、これは例えば、図10~14、16のいずれかに示されるように、上述のような形をとることができる。

【0223】

この因数分解された表現は、修飾が存在しないかのようにカノニカル配列が決定され得、その後修飾の場所が決定され得ることを意味する。修飾の条件付き分布自体は、おそらく以前の生物学的期待を反映して因数分解され得る。例えば、一方の分布は、シトシンが修飾されているかどうかを表す可能性があり、他方の分布は、修飾が存在する場合、その修飾が5-メチルシトシン(5mC)であるか5-ヒドロキシメチルシトシンであるかを表す可能性がある。

【0224】

例として、図26は、この様式で拡張された4つの塩基のフリップフロップ表現を使用して修飾された塩基5mCを検出した場合に、RNN50の出力によって予測された塩基を示す。この例では、修飾された塩基5mCは、外部予測と一致する場所の3つの位置で推定される。

【0225】

上述したように、RNN50の重み分布は、グローバルに正規化される。このようなグローバル正規化は、一連の重み分布を通るラベルのすべての経路で行われる可能性があるため、すべての可能性のある経路の合計は1つである。グローバル正規化は、重みが事後確率と見なすことができるように、出力空間にわたることができる。

【0226】

グローバル正規化は、ローカル正規化よりも厳密に表現力があり、「ラベルバイアス問題」として当該技術分野で知られている問題を回避する。

10

20

30

40

50

【0227】

ローカル正規化よりもグローバル正規化を使用することの利点は、条件付き確率場 (Lafferty et al., Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data, Proceedings of the International Conference on Machine Learning, June 2001) が、最大エントロピーマルコフモデル (McCallum et al., Maximum Entropy Markov Models for Information Extraction and Segmentation, Proceedings of ICML 2000, 591-598. Stanford, California, 2000) より優位にあることに類似する。ラベルバイアス問題は、ポリマー配列の拡張のような、ラベル間の許容された遷移の行列がスパースであるモデルに影響する。

10

【0228】

グローバル正規化は、配列全体を正規化することでこの問題を軽減し、異なる時間の遷移を相互に交換することを可能にする。グローバル正規化は、ホモポリマーおよびその他の複雑度の低い配列の偏った推定を回避するのに特に有利である。これらの配列は、他の配列と比較して許容される遷移の数が異なり得るためである (モデルによっては多くまたは少なくあり得る)。

20

【0229】

デコーダ80をここで考察する。

【0230】

デコーダ80は、重み分布51から一連のポリマー単位の推定値を導出する。これは、例えば、Graves et al., "Connectionist temporal classification labelling unsegmented sequence data with recurrent neural networks", In Proceedings of the 23rd international conference on Machine Learning, 369-376 (ACM, 2006) において開示されているように、コネクショニスト時間分類を使用して行うことができる。

30

【0231】

デコーダ80は、図27に示すように、以下の3つのステップを実行する。

【0232】

ステップS1において、ラベルの推定値は、それぞれの重み分布51に関して導出される。この推定値については、以下でさらに考察される。

【0233】

ステップS2では、ステップS1で導出されたラベルは、一連のポリマー単位の推定値を導出するためにランレングス圧縮 (これは、復号化とも呼ばれる) される。ポリマー単位よりも多くの重み分布51があるために、これが必要となる。上述のように、同じラベルの連続的配列が、RNN50に固有のポリマーの表現において同じポリマー単位を表すことから、ランレングス圧縮はポリマー単位の推定値を生成する。

40

【0234】

ステップS2はまた、所与のポリマー単位タイプを表すために複数のラベルが使用される表現を考慮に入れる。例えば、上述のマルチステイ表現では、第2のラベルは、同じポリマー単位タイプに関して第1のラベルに圧縮される。同様に、上述のフリップフロップ表現では、第1のラベル (フリップ) の連続的インスタンスは単一のポリマー単位に圧縮され、第1のラベル (フロップ) の連続的インスタンスは別の単一のポリマー単位に圧縮され、その後も同様の圧縮が行われ、それにより、ホモポリマーの推定値を提供する。

【0235】

例えば、図11のスキームでは、同じポリマー単位のインスタンスを区別するためにブラ

50

ンクを復号することは、ステップ S 2 で実行され得る。上で考察されたように、「任意選択的」および「強制的」スキームは、ブランクに対して考慮されることができ、ステップ S 2 は、2 つのスキームのどちらが後に続くかに応じて、ラベルの配列 A A A - - A を A A A または A A のいずれかに復号することができる。

【0236】

フリップフロップ方式の場合、ステップ S 2 は、同じラベルの複数の実行を単一の対応するポリマー単位に折り畳むことを含み得る。例えば、ラベルの配列 C A A F T A C C F T T F は、ステップ S 2 で一連のポリマー単位 C A A T A C C T T に復号され得る。

【0237】

マルチステイスキームに関して、ステップ S 2 は、同じタイプの異なるポリマー単位として同じラベルの連続した配列を識別することによる復号を含み得る。例えば、ラベルの配列 A A S A S T T S C A A S A S は、ステップ S 2 で一連のポリマー単位 A T C A に復号され得る。

【0238】

ランレングス符号化スキームに関して、ステップ S 2 は、同じラベルの実行を折り畳むこと（および、スキームにおいて必要に応じてブランクをドロップすること）による復号化を含み得る。例えば、ラベルの配列 T A 2 T 2 C A 3 は、ステップ S 2 で、一連のポリマー単位 T A A T T C A A A を表すことができる。

【0239】

ステップ S 3 は、重み分布 5 1 が従属分布に因数分解される場合に実行されるが、それ以外の場合は省略される。この場合、ステップ S 1 および S 2 は、遷移に関する重みを使用して実行され、ステップ S 3 では、それによって表されるポリマー単位の品質を推定するために、さらなる重みが使用される。例えば、上述のランレングス符号化表現では、ホモポリマーの長さを推定するためにさらなる重みが使用される。同様に、上述の修飾された形態の因数分解された表現では、ポリマー単位が非修飾形態であるか修飾形態であるかを推定するために、さらなる重みが使用される。

【0240】

ここで、ステップ S 1 におけるラベルの推定について考察する。重みはそれぞれの遷移の事後確率を表すため、重み分布 5 1 を介してラベルのいずれか任意経路の事後確率を導出するために重みを使用されることができ、これは問題の経路に対応する一連の遷移の重みで表される事後確率を組み合わせることによる。このことは、重みが、異なる経路の可能性が考慮されることを可能にすることを意味し、これが推定精度を向上させる。したがって、ステップ S 1 では、重み分布 5 1 を通るラベルの経路に関する遷移のための、組み合わせられた重みの考慮に基づいた技術を利用する。

【0241】

（上で考察されたように）1 つ以上の遷移が許容されない場合、ステップ S 1 でデコーダによって実行される推定は、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れることができる。

【0242】

「最良経路」と「最良ラベル」と呼ばれる 2 つの異なる手法が可能である。

【0243】

最良経路手法では、ラベルの最も可能性の高い経路が、重み分布 5 1 を基準にして一連の重み分布 5 1 を通っている。この場合、それぞれの重み分布 5 1 に関してステップ S 1 で導出されたラベルは、その最も可能性の高い経路のラベルである。

【0244】

重み分布 5 1 は遷移にわたる重みであるため、配列を推定するための復号化の 1 つの方法は、重みの最大合計を有する経路を見つけることである。このような経路は、例えば動的プログラムアルゴリズムを使用して、遷移重みから効率的な様式で見つけられ得る。ピタビアルゴリズムが使用され得る。

【0245】

10

20

30

40

50

例えば、図 28 は、最良経路アルゴリズムを示し、ここで $RNN50$ は、 w_{ijk} の重みをブロック i におけるラベル j からラベル k への遷移に出力する。ベクトル t_i は、トレースバック情報を格納し、これは、特定の現在のラベルから得られる最良のラベルであり、スコア S と最良経路 P を決定するために使用されている。

【0246】

ランレングス符号化の場合、検出される最適経路はランレングス圧縮配列用であり、各実行の長さは、 $RNN50$ によって出力される適切な条件付き分布から決定される必要がある。最良経路が新しいポリマー単位が発生したことを示す場合、実行の長さは、そのポリマー単位に対応する条件付き分布から推定される。この推定を行う適切な方式には、条件付き分布の平均値（四捨五入あり）、最頻値、または中央値を見つけることが含まれ、好適な事前確率が与えられれば、最大ベイズ因子を有する長さも使用されることもある。ネットワークが、存在している可能性のある塩基修飾を表す条件付き分布を出力する場合、修飾は順序ではなくカテゴリであるため、事後平均と中央値は好適な推定量ではないが、それらの存在で最良経路をマークアップするプロセスは同様に進行する。

10

【0247】

ランレングス符号化の場合、ランレングスバイアス補正が適用され得る。このモデルは実際のリードからトレーニングされるため、学習されてモデルの重みに組み込まれたランレングスの事前分布がいくつかある。ランダム鎖または実際の（例えば、ゲノム）鎖から派生したリードの場合、トレーニングデータが含むことになる様々な長さ（レングス）の実行（ラン）の比率には顕著な偏りがあり、例えば、長い実行は非常に稀である。このことは、方法が長い実行を呼び出す機能に影響を及ぼす。実行の長さには曖昧さがあり、短い呼び出しは、長い呼び出しよりも正しいことが多く、そのため、単一リードの精度を最大化すると、実行の呼び出しが短くなる傾向がある。そのため、比較的短いランレングスにバイアス補正を適用することには利点がある。

20

【0248】

最も可能性の高い経路を見つけた場合、カノニカル塩基の配列を導出する必要がある。フリップフロップ表現の場合、ラベルの隣接する繰り返しは、他の CTC のようなモデルのブランクラベルと同じ間隔の役割を果たし、次いで各ラベルのフリップフロップアイデンティティがスクラップされて、カノニカル塩基が残るためにマージされる。ランレングス符号化の場合、ブランクラベルが削除され、各実行が適切な数の塩基に拡張される。

30

【0249】

最良ラベル手法についてここで考察するが、最良経路手法は、正しいラベルが最も可能性の高い経路上にない場合に、いくつかの特定のラベルを不正確に誤って推定し得ることに注意されたい。 $RNN50$ からの重み分布 51 は、ラベルのすべての可能な経路にわたる確率分布、ラベルを位置に割り当てる一貫した方法、および一連のラベルに対応し、それにより、ポリマー単位に対応する（同じ配列をもたらす多くの経路があるので）この対応は一意ではないが）各経路を有効に規定する。最良ラベル手法は、最も可能性の高い一連のラベル（したがってポリマー単位）を推定することにより、最良経路手法よりも改善されている。つまり、最良経路を見つけるのではなく、タイムステップ i の後に経路がラベル j にあったという事後確率は、この条件を満たすすべての経路を合計することで見つけることができる。これは、一連の重み分布 51 を通るラベルのフォワードおよびバックワード経路を考慮に入れることができる。この場合、それぞれの重み分布 51 に関してステップ $S1$ で導出されたラベルは、このように最も可能性が高いラベルとして導出される。

40

【0250】

動的計画法を時間のフォワードとバックワードの両方で使用することで、この計算が、最良経路の場合と同様の再帰を使用して効率的に実行されることを可能にしており、最良経路アルゴリズムがビタビ復号法の形式と見なされ得る場合、事後確率の計算は、フォワードおよびバックワードアルゴリズムの形式と見なされ得る。同様に、タイムステップ i でラベルの変更があった事後確率は、この遷移を満たすすべての経路を合計することで計算

50

されることができ、この計算はまた、効率的な方法で実行されることができる。

【0251】

事後確率は、各位置での可能性のあるラベルについての情報を提供するが、ほとんど同じようなラベルを選択して復号すると一貫性がない経路になり、配列も一貫性がなくなる可能性がある。遷移行列 T を1つのラベルから、遷移が許容されているかどうかに応じてエントリが1または0のいずれかになる別のラベルへと定義することにより、最良経路復号アルゴリズムがこれらの事後確率に適用され、すべての一貫性のある経路の中から、そのラベルの事後確率の合計を最大化する経路を見つけることができる。

【0252】

この一例として、図29は、そのようなアルゴリズムが位置 i にあるラベル k であることの事後確率 P_{ik} に適用されることを示す。ベクトル t_i は、トレースバック情報を格納し、これは、特定の現在のラベルから得られる最良ラベルであり、スコア S と最良経路 P を決定するために使用されている。

10

【0253】

代替的に、すべての一貫した経路にわたってラベルの事後確率の対数の合計を最大化する経路を見つけるために、最良経路アルゴリズムが事後確率の対数に適用されることができる。これは、すべての一貫した経路にわたってラベルの事後確率の積を最大化する経路を見つけることと同じである。

【0254】

この一例として、図30は、位置 i にあるラベル k である対数事後確率 p_{ik} にそのような最良経路アルゴリズムが適用されることを示す。ベクトル t_i は、トレースバック情報を格納し、これは、特定の現在のラベルから得られる最良ラベルであり、スコア S と最良経路 P を決定するために使用されている。

20

【0255】

代替的に、重み分布 51 は遷移にわたって定義されることから、各位置にあるラベルではなく、位置間で取られる遷移の事後確率を計算するために、フォワードアルゴリズムおよびバックワードアルゴリズムが使用されることができる。

【0256】

この一例として、図31は、すべての経路を合計した事後確率の計算を示す。これらの重みは遷移にわたっていることから、遷移行列と同じ形状を有し、その対数が、一貫した経路を見つけるために、遷移重みの代わりに図28で定義された数式に入力されることができる。

30

【0257】

ゲノムの同じ領域をカバーする多数のシグナルからコンセンサス配列を生成するためのより成功した手法の1つは、「ポリシング」と呼ばれており、いくつかの出版物に記載されている。コンセンサス配列のポリシングは反復プロセスであり、ドラフトコンセンサス配列への候補の変更は、すべてのリードがそれらにどの程度一致しているかによってスコアリングされ、高いスコアの変更が保持され、1つのリードによって引き起こされた間違いが他のリードによって修正されることを可能としており、この手順がスコアの高い変更が見つからなくなるまで繰り返される。

40

【0258】

ポリシングがシングルリードへの適用にも有益であり得るかは、明らかではない。前の項で説明したポリマー単位を推定するためのすべての手法は、ネットワーク出力を介して塩基配列が抽出され得る適切な経路を見つけることを目的としているが、単一経路を適切であると識別するのではなく、レジストレーションのないトレーニング目標が特定の配列のすべての経路にわたって合計される。トレーニング基準と一致させるために、RNN50からの出力が、最も可能性の高い配列を見つけて、最も可能性の高い経路ではなく、同じ配列になるすべての経路にわたり合計して復号されるのが理想的である。特定の配列のすべての経路を合計することは、候補の変更が適切かどうかを評価するためにポリシングが使用する基準であり、そのため、ポリシングは、最も可能性の高い配列を見つけるための

50

、反復ヒューリスティック、貪欲山登り法の変形として考えることができる。

【0259】

関連する一連のポリマー単位の測定値である複数の一連の測定値を分析する場合、この方法は基本的に同じであるが、複数の一連の測定値からの測定値は、複数のそれぞれの次元で配置されるものとして扱われている。このことは次元を増加させるが、ニューラルネットワーク10の形態は、それ以外は上で説明したものと同じである。この場合に適用できるいくつかのさらなる考慮事項は以下の通りである。

【0260】

ペナルティ項を使用する場合、すべての遷移に対して一定になっている放出しないペナルティの代替として、ペナルティは、遷移に応じて異なる値を取るか、完全に存在しない可能性がある。例えば、一部の遷移は状態に変化をもたらさず、他のリードでの失われた状態を意味しないことから、自由であり得るか、わずかなペナルティを有し得る。

10

【0261】

使用される1つまたは複数のペナルティは、各リードに対して同じである必要はなく、なぜ2つのリードが異なる特性を持ち得るかという、適切な生物物理学的な理由があることがある。例えば、一方のリードはモーターの上で二本鎖であった分子からのものであるのに対し、他方は一本鎖であった分子からのものであり得、あるいは、2つのリードは異なるモーターを備えた鎖であり得、一方のリードがDNAであるのに対し、他方のリードがRNAである可能性がある、あるいは、2つのリードは、同じフォワードリバース相補鎖の第1と第2の部分であり得、配列決定中のこれら2つの間のハイブリダイゼーションが反応速度を変化させる。

20

【0262】

使用される1つまたは複数のペナルティは、時間に依存する可能性がある。使用される1つまたは複数のペナルティは、リードのローカル統計に依存する可能性がある。これら例には、速度、失速の存在、またはノイズが含まれる。使用される1つまたは複数のペナルティは、他のモデルまたは手法を使用したリード分析の出力、例えば、スリップ（塩基の欠落）の可能性の予測など、に依存する可能性がある。

【0263】

フリップフロップ表現とランレングス符号化（RLE）表現の両方の状態遷移モデルには時間順序があり、状態順序を逆にするには、有効な状態の配列ではないことがある。つまり、RLE表現では、塩基は留まる前に放出される必要がある、フリップフロップ表現では、繰り返しの最初の塩基が「フリップ」である必要がある。この結果、一方のリードが逆補数または逆である他方リードの鎖（または鎖の一部）からのものである場合、分析前に一方のリードを逆にして同じ手順を2つのフォワードリードのものとして適用するだけでは不十分である。

30

【0264】

2つの異なる方向のリードを組み合わせ、両方のリードの状態をペアとして監視するために、より複雑な手順を使用することもできるが、一方のリードと他方のリードで標準モデルを使用することが好都合であり、ここで他方のリードは、「バックワード」にトレーニングされたものとなっている（トレーニング中にはリードからのシグナルとターゲット配列が逆になっている（場合によっては補完されている））。このようなモデルの対を使用すると、フォワードリードとリバースリードの両方がモデルの状態を同じ順序で通過するため、これら両方がフォワードリードであるかのように組み合わせることができる。

40

【0265】

ニューラルネットワーク10は、例えば以下のように、従来技術を使用してトレーニングされることができる。

【0266】

ニューラルネットワーク10は、ラベルの経路上の確率を表す重みを表す分布（ラベルによる測定の一貫したラベリング）を出力し、これは次に、ポリマー単位の配列の推定値に復号される。ニューラルネットワーク10は、この推定値の誤差比率が低いことを確実に

50

することを目的とした基準でトレーニングされている。

【0267】

遷移重みを使用して経路全体の確率分布を規定する重要な側面は、すべての経路の合計が1になるように重みを正規化することである。遷移重みのセットが与えられると、正規化係数は、上で考察された事後確率の計算で使用されるようにフォワードアルゴリズム（またはバックワードアルゴリズム）を適用することにより、動的計画法を使用して計算され得る。1に正規化されるのは、各時点でのネットワークの出力ではなく、すべての可能な経路の合計であるため、この手法はグローバル正規化と呼ばれ、各経路のスコアが確率（の対数）としての判読を有することを確実にする。一貫したラベルを有するすべての経路は、確率に対応しており、これらの確率はすべての経路に分布を形成する。

10

【0268】

グローバル正規化とは対照的に、ニューラルネットワーク10を正規化して、すべての時点での出力の合計が1になるようにすることは、ローカル正規化と呼ばれる。各経路のスコアは計算でき、確率の形式を有するが、確率質量の合計が1未満であるため、分布を形成しない。ローカル正規化は、ラベルが一貫した経路を形成するかどうかに関係なく、ラベルのすべての配列に確率を割り当てる。

【0269】

配列ラベリングのトレーニングは、入力シグナルの対であるトレーニング例、それに対応するラベルの配列、およびトレーニング例を最適化するための目的関数を必要とする。ナノ細孔測定値とポリマー単位の配列との間の真のレジストレーションは不明であるため、Graves et al (2006)に記載されているようなレジストレーションのないトレーニング方法が好まれている。レジストレーションのあるトレーニング方法では、測定配列の各要素にラベルを付ける必要があるが、レジストレーションのない方法では、ポリマー単位の真の配列のみを知ることが必要とされている。リードについてのポリマー単位の真の配列は、ナノ細孔デバイスで既知の配列のポリマーを測定するか、もしくは、リードを基準配列または既知の配列を有する測定値のセットと比較することで決定される。

20

【0270】

既知の配列の測定値の例は、小さなゲノムが含むことができ、ここではラムダファージ（50キロ塩基）のような単一リードにおいて完全なゲノムを配列決定することが可能になっている。制限消化もまた使用されることができ、フラグメントはその長さで識別される。別の例では、既知のフラグメントを実行に順次追加する必要があるが、これはしたがって、フラグメントがデータに表示される時間によって識別できる。シグナルリードに配列を割り当てることができる任意の方法が使用され得ることは明らかになる。

30

【0271】

ニューラルネットワーク10をトレーニングする場合、様々な状況で様々な実験にわたって各ポリマー単位にまたがる測定を行うことが有益であり、そのため、ネットワークは通常の実行状態で遭遇するあらゆる変動の多くにさらされている。理想的には、ナノ細孔によって読み取られるとニューラルネットワーク10は、完全なリード、すなわち、全長ポリマーを網羅するシグナルと配列の対を使用してトレーニングされる。しかしながら、実

40

【0272】

反復型、畳み込み型、注意型のニューラルネットワークユニットには時間順序の概念があり、トレーニングで提示される測定ウィンドウのサイズによって、学習できるコンテキストが制限されている。各ポリマー単位が有することができる影響の範囲が広いため、トレーニングするための測定の大きなウィンドウをニューラルネットワーク10に提示することは有益である。使用されるウィンドウのサイズは、ニューラルネットワーク10が、細孔、ポリマー鎖、および他のシステム構成要素間の相互作用の適切な内部表現を作成できるようにする十分に大きな一連の測定値を提示することと、利用可能な計算処理量との間

50

のバランスである。理想的には、各リード全体が使用されるが、実際には、固定サイズの測定チャンクが適切な妥協点を提示する。適切なチャンクのサイズは、ナノ細孔と鎖の移動速度に依存するが、約200～約300塩基に対応するチャンクサイズが適切であることが証明されている。例えば、このことは、Cs g Gナノ細孔に適していることを証明している。

【0273】

一例示的なトレーニングセットのサイズは、シグナルおよび配列の約300塩基チャンクの約100万セットを含み得る。数千チャンクのより小さなトレーニングセットで十分になることができ、100万チャンクを超えるより大きなトレーニングセットは、トレーニングの多様性をもたらすことができる。

10

【0274】

ニューラルネットワークをトレーニングする多くの技術、または他の機械学習方法が、当技術分野で知られており、ここで適用されることができる。様々な実験的実行に一般化するための方法による能力により、ポリマー配列がトレーニングデータの大規模なセットから恩恵を受け、グラフィックプロセッシングユニット（GPU）またはその他の特殊なハードウェアで計算を実行することが好ましいため、目的関数の方向を最大化しようとすることは、メモリに制限があることから、多くの場合、非現実的である。データの完全なセット全体にわたり目的関数を直接最大化するのではなく、確率的勾配降下法（SGD）または関連する手法を使用して、完全なトレーニングセットのサブセット（「ミニバッチ」）を使用して反復的に目的関数を概ね最大化することが好ましい。推奨されるミニバッチサイズは、使用する計算デバイスで利用可能なメモリと、ミニバッチの各要素の測定数に依存する。

20

【0275】

確率的勾配降下法（SGD）の多くの変形が当技術分野で知られており、それらは例えば、SGD、モメンタムSGD、ネステロフモメンタムSGD、RMSprop、Adamax、Adamなどである。Adamの変形である、「Adamski」では、反復Nのモメンタムがモメンタムランピング係数 r によって0からその中の最大値 μ に増加しており、 $\mu_N = \mu(1 - e^{-rN})$ が好ましい。Adamskiには、学習率、2つの平滑化パラメータ（当技術分野ではdecay1およびdecay2と呼ばれることが多い）、およびモメンタム傾斜率がある。これらのパラメータの多くの選択が有益になっている。好ましいパラメータ化は、 10^{-3} の初期学習率、0.9および0.999の平滑化パラメータ、ならびに0.005のモメンタム傾斜係数を有している。0.95および0.99の平滑化パラメータはまた、初期学習率を 10^{-4} に下げると同様に、すでにトレーニングされたモデルを改良するためにも効果的であることが証明されている。

30

【0276】

SGDおよび関連する手法は反復的に進行し、各反復は次のステップで構成される。

1. 完全なトレーニングデータのサブセットを選択する。
2. このサブセットの目的関数を計算する。
3. 逆伝搬法を使用して、すべてのネットワークパラメータの勾配を計算する。
4. SGDまたはその変形を使用してネットワークパラメータを更新する。
5. 1に進む（次の反復を開始する）。

40

【0277】

ステップ4における更新のサイズは、学習率と呼ばれる係数でスケールされる。学習率が高いということは、パラメータが急速に変化できることを意味しており、そのため最大化がより迅速に進むことができ、各ミニバッチの効果が大きくなることができ、モデルが収束に近いときの更新は、ミニバッチ間の変動によって影響を受けることを意味する。反復ごとに学習率をゆっくりと低減することが好ましく、この低減は動的であることができ、バッチごとの目的関数の変化と変動性に従って、または所定のスケジュールに従って学習率に調整される。選択のために、ある初期学習率 R とミニバッチ数 K に対して、 N thミニバッチの学習率が $R / (1 + (N / K))$ である双曲線減衰が使用される。

50

【0278】

合計を使用して、ミニバッチの各構成要素のスコアを組み合わせるミニバッチのスコアとしたが、他の組み合わせ方法も可能である。合計は、その構成要素のスコアの平均に比例するミニバッチスコアとなり、中心的な傾向の他の測定値に対応する組み合わせもまた、好ましい特性を有している。中央値、トリミングされた平均値、または加重平均などの組み合わせ、もしくはM推定量のフィッティングが、目標の感度を孤立値を有するミニバッチ要素に変更するために使用され得る。

【0279】

ミニバッチの各要素から合計スコアへの寄与は、すべての一貫した経路にわたって合計された真の配列の事後確率（の対数）である。フリップフロップ表現の場合、フリップからフリップまたはフロップからフロップの遷移は、配列の同じ位置に留まっていることを表すが、一方で、他のすべての遷移には位置の移動が含まれる。各時点でRNN50から出力されたラベル間の遷移を表す重みが与えられると、これらは既知の配列の位置間の遷移重みに変換されることができる。

10

【0280】

図32は、ラベル S_1 、 S_2 、...、 S_N のフリップフロップの符号化配列の各時点 i に対する目的遷移行列 m_i の要素をどのように構築するかを示す。図33で説明されている目的関数は、この目的遷移行列を使用して、ミニバッチの各要素のスコアを計算する。

【0281】

目的関数の遷移行列は非常にスパースであり、対角（ステイ）と超対角（位置に移動）に非ゼロ要素のみを有しており、この計算の好ましい実施形態では、ゼロ要素のみを無視し、象限から線形までの真の配列の長さの観点から、各ステップの見かけ上の複雑性を低減する。

20

【0282】

マルチステイ表現の目的関数は、構造的にはフリップフロップの目的に類似しているが、同じ位置にとどまることを表す状態が異なる。ステイ状態または非ステイ状態からいずれかの非ステイ状態への遷移は、位置の変更を意味しており、ステイ状態へのいずれかの遷移は意味しない。マルチステイ表現の場合、新しい位置にとどまる（塩基からステイへの遷移）と古い位置にとどまる（ステイからステイへの遷移）を表す遷移は区別されており、目的関数の効率的な計算には、「残りの」位置のための真の配列の複製： S_1 、 R_1 、 S_2 、 R_2 ...、 S_N 、 R_N の使用が必要とされる。

30

【0283】

図34は、この例について目的遷移行列の要素をどのように構築するかを示す。目的遷移行列を形成することを目的として、元の位置は $1 \dots N$ と列挙されるが、一方に対応する複製位置は $N+1 \dots 2N$ と列挙される。

【0284】

図33で説明される目的関数は、この目的遷移行列を使用して、ミニバッチの各要素のスコアを計算する。目的遷移行列はスパースであり、目的計算の好ましい実施形態は、このスパース性を利用している。

【0285】

各スコアは、目的関数で使用前に重みで乗算でき、この重みは、トレーニングプロセスに対応するミニバッチの要素の値を表すことができる。重みは、例えば、異常な配列構成を有する要素、または以前にトレーニングされたネットワークのテスト中に検出されることがある塩基呼び出しエラーに関与することが知られている要素の場合に大きくなり得る。ミニバッチの要素の重みを決定する1つの方法は、それを最も希少なホモポリマーの頻度の逆数に等しく設定することであり、この頻度は、トレーニングデータのセット全体から、または他の外部基準から決定される。

40

【0286】

ランレングス符号化の目的は、マルチステイモデルの目的と同様に定義されているが、新しい配列位置が遷移するたびに追加の要素が組み込まれ、ネットワークによって出力され

50

る対応する条件付き分布によって、実行（ラン）の長さ（レングス）がどれだけ適切に予測されるかを表している。ランレングス圧縮配列にわたる目的遷移行列の形式は、多くのステイ目的の場合と同じ形式を有し、いずれの塩基も位置間で許容された遷移に暗黙的になっている同じ塩基に続くことができないが、ネットワークがその組成を与えられた各位置でホモポリマーの長さに割り当てる対数確率からの追加的構成要素を備えている。

【0287】

トレーニングデータのホモポリマー含有量が偏っていることが明らかである場合、他のデータセットを表していないことがあるため、ネットワークがこの偏りを学習することは多くの用途で望ましくないことがある。ネットワークがトレーニング目標で直接その組成を与えられた各位置でホモポリマーの長さに割り当てる対数確率を使用する代わりに、最初に別の分布と組み合わせられることができ、この他の分布は、トレーニングデータ（「分布前のトレーニング」）からホモポリマーの頻度を集計することによって取得され得る。この方法でトレーニングすることにより、ネットワークは、トレーニングの事前分布の期待を克服する対数確率を割り当てることを学習する必要がある。

10

【0288】

ベースコールの目的で、トレーニングデータからの事前分布、またはホモポリマーの長さの他の期待値が、「ベイズの定理」などの標準的な方法を使用してネットワークによって割り当てられた対数確率と組み合わせられ、ホモポリマーの長さに関する外部情報によって通知される新しい対数確率を生成し、代替的にネットワークからの対数確率が、バイアスのない呼び出しに直接使用されることもできる。

20

【0289】

図35は、この例について目的遷移行列の要素をどのように構築するかを示している。測定的时间ステップ i における配列の位置 j について、ネットワークが長さ L_j の実行に組成 S_j を割り当てる確率の対数が、 $r^i S_j : L_j$ となるようにする。図33で説明される目的関数は、この目的遷移行列を使用して、ミニバッチの各要素のスコアを計算する。

【0290】

レジストレーションを用いない様式でベースコールモデルをトレーニングすることの利点は数多くあるが、提示されるほとんどの復号アルゴリズムと使用されるトレーニング目標の間には乖離がある。モデルトレーニングの目的関数は、塩基の真の配列の確率を最大化し、それを表す可能性のあるすべての個々の経路の確率を合計することであるが、一方で上述のポリシング以外のすべての復号ルーチンは、高いスコアの経路を見つけに行くことである。図36は、この乖離によって引き起こされる問題の1つを示す。特に、図36は、2410～2600の回数間に長いホモポリマー領域がある場合のフリップフロップ表現の例として時間経過と共に特定のラベルに存在するシグナル（上）と事後確率（下）を示しており、ここではモデルが、Tフリップ状態（赤い実線）と交互になるのではなく、Tフロップ状態（赤い破線）にとどまっている。長いホモポリマーに入った後、領域の開始と終了の前後で推定が行われるが、フリップとフロップの状態はすぐに不明瞭になり、事後確率は均一になる。領域を通る複数の経路があり、この領域は、フリップ塩基とフロップ塩基のレジストレーションがわずかに異なり、事後確率がこのアンサンブルの平均を反映する。

30

40

【0291】

考えられる代替案の1つは、すべての経路を合計するのではなく、トレーニングの目的として最良経路のスコアを使用することであり、これは、レジストレーションが明示的に定義されておらず、ラベル付けとは異なって、最良のレジストレーションがモデルが行うように変化し得ることから、依然としてレジストレーションを用いない方法となる。最良経路へのトレーニングは直感的であるように見えるが、最初の貧弱なモデルには悪い最良経路があり、トレーニングプロセスがそれを強化することにより、モデルを何も無い状態からトレーニングする場合、この手法は大きく失敗することになる。

【0292】

シャープニングは、他のすべての可能性を考慮しながら、レジストレーションを事前に指

50

定することなく、トレーニングを単一の経路に集中させる方法である。初めに、すべての経路の合計のスコア（図33）と最良経路のスコア（図37）を計算するためのアルゴリズムを検討する。これらの両方がそれぞれファンクター $\log \sum_j \exp$ と $\max_j$ を適用し、遷移重みと以前の前方ベクトルを共に組み合わせる。シャープニングの目標は、このファンクターを、すべての可能な経路を合計するが、スコアの高い経路を高く重み付けするファンクターに置き換えることである。

【0293】

図38は、フォワードベクトルと遷移またはマッピング重みを組み合わせるために使用される関数であり得るいくつかのファンクターを示す。図38で「すべての経路をシャープニング」と呼ばれるファンクターが好ましいが、他の多くのファンクターが使用されることができ、実際には、組み合わせられて新規のファンクターを作成する。

10

【0294】

最初からシャープニングを有効にしてトレーニングするよりも、全経路目的関数を使用してトレーニングを開始し、次に、良好なモデルが見つかり、シャープニング係数（ a ）を1からより高い値に増やし、場合により、さらに高いシャープニングの値で繰り返す方が有利であることが分かった。この多段階プロセスはまた、モデルが最良経路目標を使用してトレーニングすることを可能にする。最初にすべての経路の目的に対してトレーニングを行うと、良好なモデルが見つかり、最良経路が良好であり、次に、さらにトレーニングを行うことで、この経路が強化されるようになる。

【0295】

20

図39は、図36に示した同じ領域の例で、フリップフロップ表現をシャープニングする効果を示す。図39は、この例では経時的に特定の状態にあるが、シャープニングを使用したトレーニングを行った場合のシグナル（上）と事後確率（下）を示す。概して、2400～2620回の間には、長い領域があり、ホモポリマーが発生し、モデルは、塩基の配列を呼び出すために、T-フロップ状態とT-フリップ状態を交互に繰り返している。個々の呼び出しはより明確であり、ホモポリマー領域全体でT-フリップとT-フロップの間で交互になっていることが分かる。

【0296】

このモデルを復号すると、シャープニングされていないモデルよりもポリマー単位の推定が優れたものになる。これは、図40に示す例に示されており、ここでは、シャープニングされていないモデルとシャープニングされたモデルからのポリマー単位の推定（ベースコール）が基準配列と比較される。シャープニングされていない呼び出しは、8個のT塩基のみを呼び出すが、シャープニングされた呼び出しは、基準にある27個のT塩基と一致している。

30

【0297】

最良経路またはシャープニングを使用すると、トレーニング対象が置き換えられるが、それらを使用して、それを補強し、テスト中に見つかった望ましくない動作からネットワークをトレーニングすることもできる。そのような望ましくない動作の1つは、ホモポリマーの長さを過小評価する傾向になる可能性があり、このことは、トレーニングデータが短いホモポリマーに大きく偏っている場合に発生することがあり、トレーニング対象にペナルティを追加することで修正され得る。そのようなペナルティの1つは、ホモポリマーが呼び出される位置を見つけるために最良経路を使用し、そして、その真の長さを、その位置でネットワークによって割り当てられた対数確率に基づいた推定値と比較することによって見つけることができ、この比較は、絶対差の合計を使用して実行することができ、この比較は、差の二乗和を使用して実行することができ、他の多くの比較方法が当技術分野で知られている。ペナルティはトレーニング目標に追加される可能性があり、ペナルティは、トレーニング目標に対する重要性を変更するために、所定の係数によって追加的に重み付けされる可能性がある。

40

【0298】

ペナルティ項を重み付けする係数は、所定のものではなくラグランジュ乗数として扱われ

50

ることもある。トレーニングは、ラグランジュ乗数の停留点を見つけながら、トレーニング目標を最適化することによって進行する。これらのポイントにおいて、またはその近くでは、ペナルティは概ねゼロであり、ネットワークは、ペナルティ条件の保持に従ってトレーニングされており、ペナルティが真の長さで推定された長さの絶対差の合計である例の場合、ネットワーク呼び出しは平均した補正長さになる。

【0299】

例えば、ホモポリマーの長さごとに1つのトレーニング目標を強化するために複数のペナルティ項が使用されることができ、各ペナルティは、所定の係数で重み付けされるか、ラグランジュ乗数として扱われることができる。

【0300】

上述の説明は、重み分布51がラベルのセット間の遷移を表す場合を考慮している。代替策として、本明細書に記載の方法は、重み分布51がラベルのセット内のラベルを表す場合に適合され得る。

【0301】

この場合、重み分布51がラベルのセット内のラベルを表しており、デコーダ80は、遷移行列を使用して、ラベル間の遷移が許容されるか許容されないかを表すことができる。遷移行列は、重み分布51の重みの行列と同様の形式を有することができるが、許容されるかまたは許可されない遷移を示すバイナリ要素を有する。遷移行列は、少なくとも1つの遷移を許容されていないものとして表し、他の遷移を許容されているものとして表すことができる。デコーダ80は、この遷移行列を使用して、遷移行列に従って許容されるラベルを通る異なる経路の可能性を考慮に入れて、ラベルを表す重み分布51から一連のポリマー単位の推定値を導出することができる。

【0302】

同様に、この場合、重み分布51がラベルのセット内のラベルを表しており、一連のポリマー単位内の同じポリマー単位タイプの連続するインスタンスは、例えば、フリップフロップ表現、またはランレングス符号化表現を使用して、上述のように符号化形式で表すことができる。

【0303】

上述の説明は、RNN50を含むニューラルネットワーク10に関連しているが、上述の形式および復号化を有する重み分布は、他の形式の機械学習技術、例えば、HMMに等しく適用されることができる。

【0304】

本発明の第2の態様によれば、以下の条項で規定される方法が提供される。

【0305】

条項1. ポリマーによって生じるシグナルをナノ細孔に対するポリマーの転位中に分析する方法であって、ポリマーが、可能なポリマー単位タイプのセットに属する一連のポリマー単位を含み、方法が、一連の重み分布を出力する機械学習技術を使用してシグナルを分析することであって、各重み分布が、可能なポリマー単位タイプを表すラベルを含むラベルのセットにわたるラベルに関する重みを含んでいることと、重み分布から一連のポリマー単位の推定値を導出することと、を含み、一連のポリマー単位の推定値を導出するステップは、ラベル間の遷移が許容されるか許容されないかを表す遷移行列を考慮に入れ、ラベル間の少なくとも一方の遷移は、許容されないものとして表されており、他方の遷移は、許容されたものとして表されている、方法。

【0306】

条項2. ラベル間の少なくとも1つの遷移が許容されておらず、他の遷移が許容されており、重み分布がそれぞれ、許容されているラベルに関する重みを含む、条項1に記載の方法。

【0307】

条項3. 重み分布が各々、許容されていないラベルに関するヌル重みを含む、条項2に記載の方法。

【0308】

条項 4. 一連のポリマー単位の推定値を導出する前記ステップが、ラベル間の遷移が許容されているか許容されていないかを表す遷移行列を考慮に入れる、条項 2 または 3 に記載の方法。

【0309】

条項 5 ラベルのセットが、各ポリマー単位タイプに関する第 1 のラベルおよび第 2 のラベルを含み、第 1 のラベルが、ポリマー単位タイプのインスタンスの開始を表し、第 2 のラベルが、ポリマー単位タイプのインスタンスにおけるステイを表し、各第 1 のラベルから任意の他のポリマー単位タイプの第 1 のラベルへの遷移が許容されており、各第 1 のラベルから同じポリマー単位タイプの第 1 のラベルへの遷移が許容されており、各第 1 のラベルから同じポリマー単位タイプの第 2 のラベルへの遷移が許容されており、各第 1 のラベルから任意の他のポリマー単位タイプの第 2 のラベルへの遷移が許容されておらず、各第 2 のラベルから同じポリマー単位タイプの第 1 のラベル、または任意の他のポリマー単位タイプの第 1 のラベルへの遷移が許容されており、各第 2 のラベルから同じポリマー単位タイプの第 2 のラベルへの遷移が許容されており、各第 2 のラベルから任意の他のポリマー単位タイプの第 2 のラベルへの遷移が許容されていない、条項 2 ～ 4 のいずれか一項に記載の方法。

10

【0310】

条項 6. 可能なポリマー単位タイプのセットが、ポリマー単位の既知の配列で常に現れるポリマー単位タイプを含み、前記既知の配列に従った遷移が許容されており、前記既知の配列に反した遷移が許容されていない、条項 2 ～ 5 のいずれか一項に記載の方法。

20

【0311】

条項 7. 一連のポリマー単位における同じポリマー単位タイプの連続的インスタンスが符号化形式で表される、条項 2 ～ 6 のいずれか一項に記載の方法。

【0312】

条項 8. ラベルが、各ポリマー単位タイプに関する複数のラベルを含み、各ポリマー単位タイプに関する複数のラベルが、一連のポリマー単位におけるポリマー単位タイプの連続的インスタンスを表す、条項 7 に記載の方法。

【0313】

条項 9. ポリマー単位の各タイプのための複数のラベルが、所定の周期的順序を有し、それにより、ラベル間のいくつかの遷移が、所定の周期的順序によって許容されており、間の他の遷移が、所定の周期的順序によって許容されておらず、重み分布がそれぞれ、所定の周期的順序によって許容されている遷移に関する重みを含む、条項 8 に記載の方法。

30

【0314】

条項 10. 各ポリマー単位タイプの複数のラベルが、各ポリマー単位タイプの 2 つのラベルである、条項 8 または 9 に記載の方法。

【0315】

条項 11. 一連のポリマー単位内の同じポリマー単位タイプの連続するインスタンスが、ランレングス符号化された形式で表されている、条項 7 に記載の方法。

【0316】

条項 12. ラベルが、各ポリマー単位タイプの異なるランレングスに関する複数のラベルを含む、条項 11 に記載の方法。

40

【0317】

条項 13. ラベルが、各ポリマー単位タイプに関するラベルを含み、重み分布が、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さにわたるさらなる重みを含む、条項 11 に記載の方法。

【0318】

条項 14. さらなる重みが、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さのセットにわたる重みのカテゴリカル分布を含む、条項 3 に記載の方法。

50

【0319】

条項15．さらなる重みが、各ポリマー単位タイプの同じポリマー単位タイプの連続的インスタンスの可能な長さにわたるパラメータ化された分布のパラメータを含む、条項13に記載の方法。

【0320】

条項16．可能なポリマー単位タイプが、非修飾形態および修飾形態を有するポリマー単位タイプを含む、条項2～15のいずれか一項に記載の方法。

【0321】

条項17．ラベルのセットが、非修飾形態および修飾形態を有する前記ポリマー単位タイプに関するラベルを含む、条項16に記載の方法。

【0322】

条項18．各重み分布が、非修飾形態および修飾形態を有する各ポリマー単位タイプの非修飾形態および修飾形態のさらなる重みを含む、条項17に記載の方法。

【0323】

条項19．ラベルのセットが、各ポリマー単位タイプを表す少なくとも1つのラベルを含む、条項2～18のいずれか一項に記載の方法。

【0324】

条項20．ラベルのセットが、一連のポリマー単位におけるブランクおよび／またはステイを表す少なくとも1つのラベルをさらに含む、条項1～19のいずれか一項に記載の方法。

【0325】

条項21．機械学習技術が、少なくとも1つの再帰層を含むニューラルネットワークである、条項1～20のいずれか一項に記載の方法。

【0326】

条項22．少なくとも1つの再帰層が、双方向再帰層である、条項21に記載の方法。

【0327】

条項23．ニューラルネットワークが、一連の重み分布を介してラベルのすべての経路にわたって重み分布のグローバル正規化を適用する、条項21または22に記載の方法。

【0328】

条項24．ニューラルネットワークが、少なくとも1つの再帰層の前に配置され、シグナルのウィンドウ化セクションの畳み込みを実行する、少なくとも1つの畳み込み層を含む、条項21～23のいずれか一項に記載の方法。

【0329】

条項25．重みが事後確率を表す、条項1～24のいずれか一項に記載の方法。

【0330】

条項26．重み分布から一連のポリマー単位の推定値を導出するステップが、コネクショニスト時間分類を使用して実行される、条項1～25のいずれか一項に記載の方法。

【0331】

条項27．重み分布からポリマー単位の推定値を導出するステップが、それぞれの重み分布に関するラベルを導出することと、導出されたラベルをランレングス圧縮することとを含む、条項1～26のいずれか一項に記載の方法。

【0332】

条項28．重み分布から一連のポリマー単位の推定値を導出するステップが、重み分布を基準にして一連の重み分布を通るラベルの最も可能性の高い経路を推定することと、最も可能性が高いと推定されたラベルの経路から一連のポリマー単位の推定値を導出することと、を含む、条項1～27のいずれか一項に記載の方法。

【0333】

条項29．重み分布から一連のポリマー単位の推定値を導出するステップが、各重み分布に関して最も可能性が高いラベルを推定することと、一連の重み分布を通るラベルのフォワードとバックワードの経路を考慮に入れることと、最も可能性が高いと推定されたラベ

10

20

30

40

50

ルから一連のポリマー単位の推定値を導出することと、を含む、条項 1 ～ 2 8 のいずれか一項に記載の方法。

【 0 3 3 4 】

条項 3 0 . ナノ細孔がタンパク質細孔である、条項 1 ～ 2 9 のいずれか一項に記載の方法。

【 0 3 3 5 】

条項 3 1 . ポリマーがポリヌクレオチドであり、ポリマー単位がヌクレオチドである、条項 1 ～ 3 0 のいずれか一項に記載の方法。

【 0 3 3 6 】

条項 3 2 . シグナルが、イオン電流、インピーダンス、トンネリング特性、電界効果トランジスタ電圧、および光学特性のうちの 1 つ以上の特性の測定値から導出されている、条項 1 ～ 3 1 のいずれか一項に記載の方法。

【 0 3 3 7 】

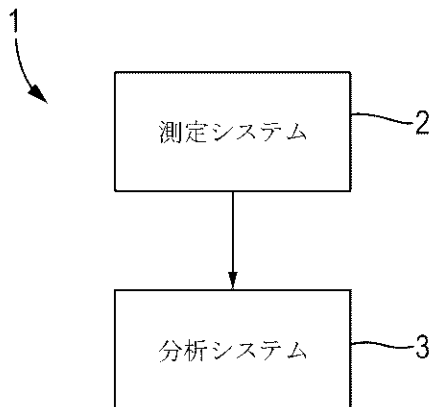
条項 3 3 . 方法が、コンピュータ装置で実行されている、条項 1 ～ 3 2 のいずれか一項に記載の方法。

【 0 3 3 8 】

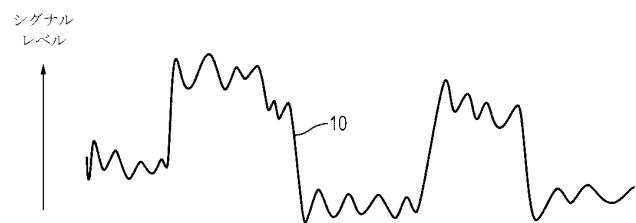
条項 3 4 . ナノ細孔に対するポリマーの転位中にポリマーからシグナルを導出することをさらに含む、条項 1 ～ 3 3 のいずれか一項に記載の方法。

【 図 面 】

【 図 1 】



【 図 2 】



10

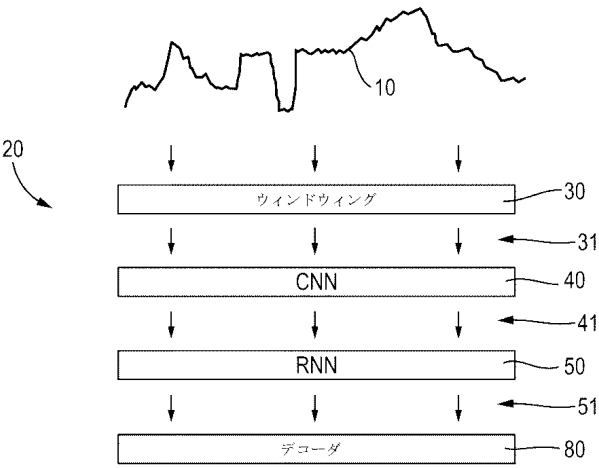
20

30

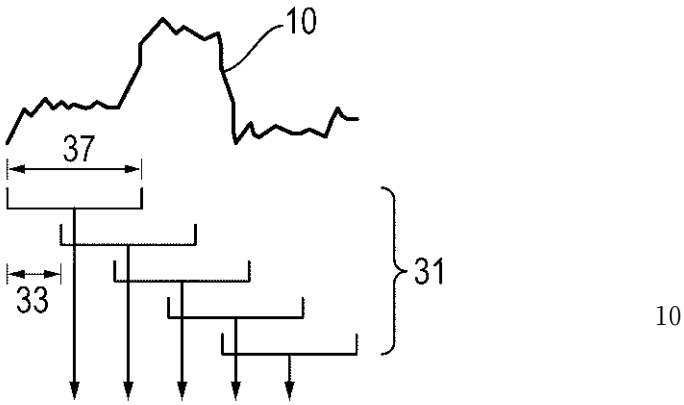
40

50

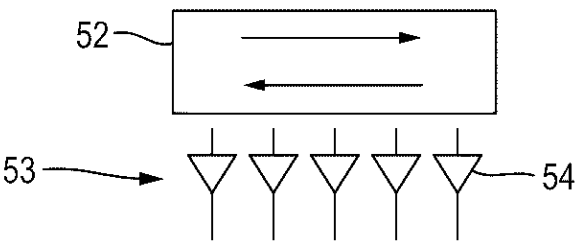
【図 3】



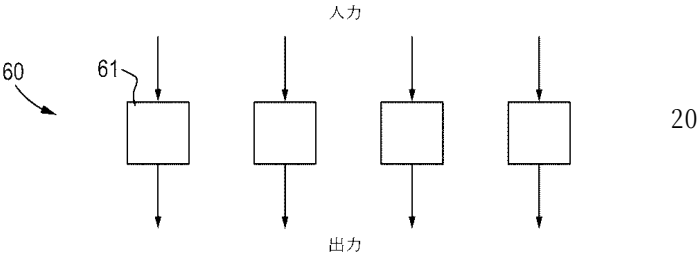
【図 4】



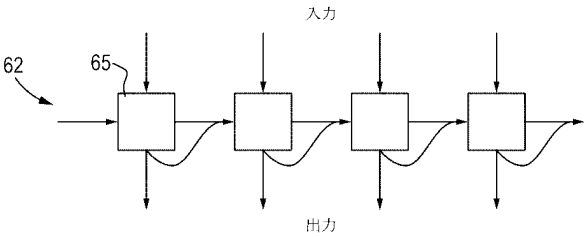
【図 5】



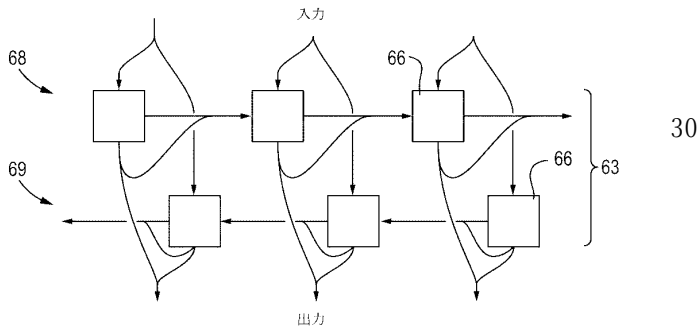
【図 6】



【図 7】



【図 8】



10

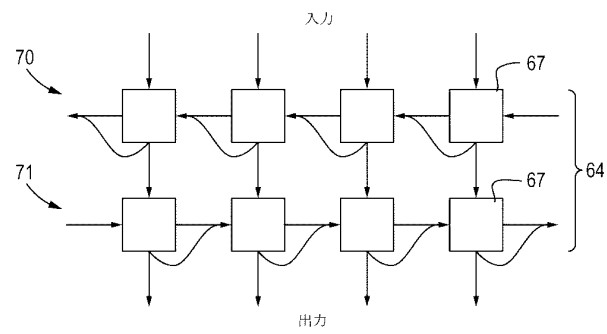
20

30

40

50

【図 9】



【図 1 0】

基本形

	A	C	G	T
A	W_{11}	W_{12}	W_{13}	W_{14}
C	W_{21}	W_{22}	W_{23}	W_{24}
G	W_{31}	W_{32}	W_{33}	W_{34}
T	W_{41}	W_{42}	W_{43}	W_{44}

10

【図 1 1】

ブランク

	A	C	G	T	B
A	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}
C	W_{21}	W_{22}	W_{23}	W_{24}	W_{25}
G	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}
T	W_{41}	W_{42}	W_{43}	W_{44}	W_{45}
B	W_{51}	W_{52}	W_{53}	W_{54}	W_{55}

【図 1 2】

CpG メチル化

	A	C	G	T	B	C^m
A	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}
C	W_{21}	W_{22}	W_{23}	W_{24}	W_{25}	W_{26}
G	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}	W_{36}
T	W_{41}	W_{42}	W_{43}	W_{44}	W_{45}	W_{46}
B	W_{51}	W_{52}	W_{53}	W_{54}	W_{55}	W_{56}
C^m	—	—	W_{63}	—	W_{65}	W_{66}

20

30

40

50

【図 1 3】

マルチステイ

	A	C	G	T	A ^s	C ^s	G ^s	T ^s
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	—	—	—
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	—	W ₂₆	—	—
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	—	—	W ₃₇	—
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	—	—	—	W ₄₈
A ^s	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	—	—	—
C ^s	W ₆₁	W ₆₂	W ₆₃	W ₆₄	—	W ₆₆	—	—
G ^s	W ₇₁	W ₇₂	W ₇₃	W ₇₄	—	—	W ₇₇	—
T ^s	W ₈₁	W ₈₂	W ₈₃	W ₈₄	—	—	—	W ₈₈

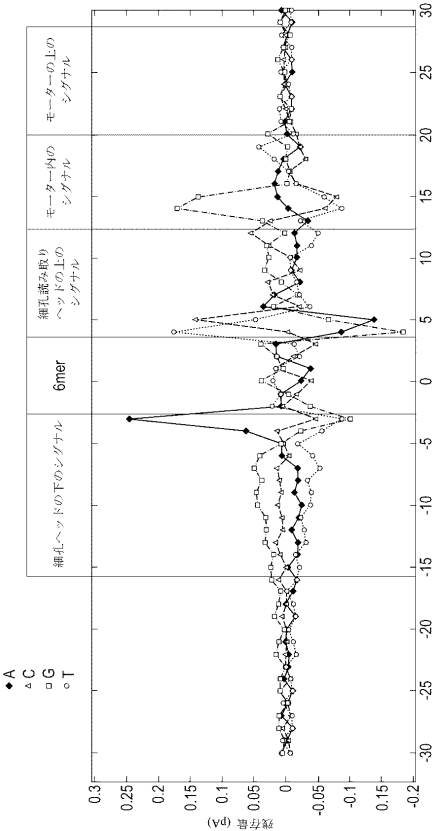
【図 1 4】

フリップフロップ

	A	C	G	T	A ^f	C ^f	G ^f	T ^f
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	—	—	—
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	—	W ₂₆	—	—
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	—	—	W ₃₇	—
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	—	—	—	W ₄₈
A ^f	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	—	—	—
C ^f	W ₆₁	W ₆₂	W ₆₃	W ₆₄	—	W ₆₆	—	—
G ^f	W ₇₁	W ₇₂	W ₇₃	W ₇₄	—	—	W ₇₇	—
T ^f	W ₈₁	W ₈₂	W ₈₃	W ₈₄	—	—	—	W ₈₈

10

【図 1 5】



【図 1 6】

ランレングス符号化-離散

	A ¹	C ¹	G ¹	T ¹	A ²	C ²	G ²	T ²	A ³	...
A ¹	W ₁₁	W ₁₂	W ₁₃	W ₁₄	—	W ₁₆	W ₁₇	W ₁₈	—	...
C ¹	W ₂₁	W ₂₂	W ₂₃	W ₂₄	W ₂₅	—	W ₂₇	W ₂₈	W ₂₉	...
G ¹	W ₃₁	W ₃₂	W ₃₃	W ₃₄	W ₃₅	W ₃₆	—	W ₃₈	W ₃₉	...
T ¹	W ₄₁	W ₄₂	W ₄₃	W ₄₄	W ₄₅	W ₄₆	W ₄₇	—	W ₄₉	...
A ²	—	W ₅₂	W ₅₃	W ₅₄	W ₅₅	W ₅₆	W ₅₇	W ₅₈	W ₅₉	...
C ²	W ₆₁	—	W ₆₃	W ₆₄	W ₆₅	W ₆₆	W ₆₇	W ₆₈	W ₆₉	...
G ²	W ₇₁	W ₇₂	—	W ₇₄	W ₇₅	W ₇₆	W ₇₇	W ₇₈	W ₇₉	...
T ²	W ₈₁	W ₈₂	W ₈₃	—	W ₈₅	W ₈₆	W ₈₇	W ₈₈	W ₈₉	...
A ³	—	W ₉₂	W ₉₃	W ₉₄	—	W ₉₆	W ₉₇	W ₉₈	W ₉₉	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

20

30

40

50

【図 17】

ランレングス符号化ーホールド

	A	C	G	T	A ^H	C ^H	G ^H	T ^H
A	－	W ₁₂	W ₁₃	W ₁₄	W ₁₅	－	－	－
C	W ₂₁	－	W ₂₃	W ₂₄	－	W ₂₆	－	－
G	W ₃₁	W ₃₂	－	W ₃₄	－	－	W ₃₇	－
T	W ₄₁	W ₄₂	W ₄₃	－	－	－	－	W ₄₈
A ^H	－	W ₅₂	W ₅₃	W ₅₄	W ₅₅	－	－	－
C ^H	W ₆₁	－	W ₆₃	W ₆₄	－	W ₆₆	－	－
G ^H	W ₇₁	W ₇₂	－	W ₇₄	－	－	W ₇₇	－
T ^H	W ₈₁	W ₈₂	W ₈₃	－	－	－	－	W ₈₈

【図 18】

ランレングス符号化ーカテゴリカル

	1	2	3	4	
A	l ₁₁	l ₁₂	l ₁₃	l ₁₄	
C	l ₂₁	l ₂₂	l ₂₃	l ₂₄	
G	l ₃₁	l ₃₂	l ₃₃	l ₃₄	
T	l ₄₁	l ₄₂	l ₄₃	l ₄₄	

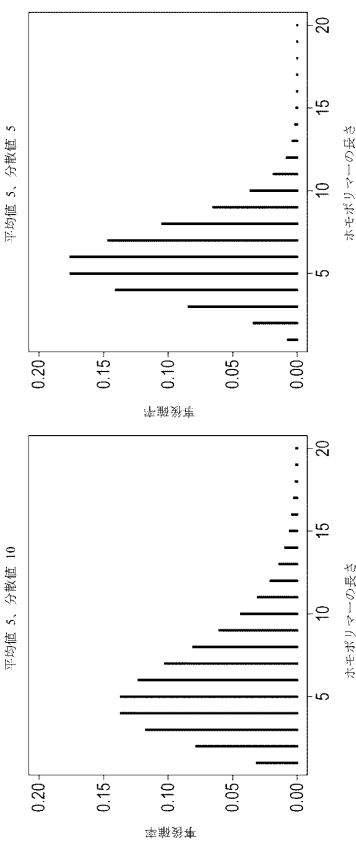
10

【図 19】

ランレングス符号化ーパラメータ化

	P ₁	P ₂	
A	P ₁₁	P ₁₂	
C	P ₂₁	P ₂₂	
G	P ₃₁	P ₃₂	
T	P ₄₁	P ₄₂	

【図 20】



20

30

40

50

【図 2 1】

分布パラメータ ガンマ	$\mathbb{P}(X=x), \text{スケール}$ $\frac{1}{\Gamma(\alpha)} \left[\gamma\left(\alpha, \frac{x+1}{\beta}\right) - \gamma\left(\alpha, \frac{x}{\beta}\right) \right]$ 形状 $\alpha > 0$; 尺度 $\beta > 0$
対数ロジスティック	$\frac{1}{1 + \left(\frac{x+1}{\alpha}\right)^{-\beta}} - \frac{1}{1 + \left(\frac{x}{\alpha}\right)^{-\beta}}$ 形状 $\alpha > 0$; 尺度 $\beta > 0$
対数正規	$\frac{1}{2} \phi\left(\frac{\log(x+1) - \mu}{\sigma\sqrt{2}}\right) - \frac{1}{2} \phi\left(\frac{\log x - \mu}{\sigma\sqrt{2}}\right)$ 対数平均 $\mu \in \mathbb{R}$; 対数標準偏差 $\sigma > 0$
ワイブル	$e^{-\left(\frac{x}{\beta}\right)^\alpha} - e^{-\left(\frac{x+1}{\beta}\right)^\alpha}$ 形状 $\alpha > 0$; 尺度 $\beta > 0$

【図 2 3】

	P ₁	P ₂	...
C→A→C	P ₁₁	P ₁₂	...
G→A→G	P ₂₁	P ₂₂	...
T→A→T	P ₃₁	P ₃₂	...
G→A→C	P ₄₁	P ₄₂	...
T→A→G	P ₅₁	P ₅₂	...
T→A→T	P ₆₁	P ₆₂	...
T→A→C	P ₇₁	P ₇₂	...
C→A→G	P ₈₁	P ₈₂	...
G→A→T	P ₉₁	P ₉₂	...
A→C→A	P ₁₀₋₁	P ₁₀₋₂	...
G→C→G	P ₁₁₋₁	P ₁₁₋₂	...
T→C→T	P ₁₂₋₁	P ₁₂₋₂	...
G→C→A	P ₁₃₋₁	P ₁₃₋₂	...
T→C→G	P ₁₄₋₁	P ₁₄₋₂	...
A→C→T	P ₁₅₋₁	P ₁₅₋₂	...
T→C→A	P ₁₆₋₁	P ₁₆₋₂	...
A→C→G	P ₁₇₋₁	P ₁₇₋₂	...
G→C→T	P ₁₈₋₁	P ₁₈₋₂	...
T→G→T	P ₁₉₋₁	P ₁₉₋₂	...
A→G→A	P ₂₀₋₁	P ₂₀₋₂	...
C→G→C	P ₂₁₋₁	P ₂₁₋₂	...
A→G→T	P ₂₂₋₁	P ₂₂₋₂	...
C→G→A	P ₂₃₋₁	P ₂₃₋₂	...
T→G→C	P ₂₄₋₁	P ₂₄₋₂	...
C→G→T	P ₂₅₋₁	P ₂₅₋₂	...
T→G→A	P ₂₆₋₁	P ₂₆₋₂	...
A→G→C	P ₂₇₋₁	P ₂₇₋₂	...
A→T→A	P ₂₈₋₁	P ₂₈₋₂	...
C→T→C	P ₂₉₋₁	P ₂₉₋₂	...
G→T→G	P ₃₀₋₁	P ₃₀₋₂	...
C→T→A	P ₃₁₋₁	P ₃₁₋₂	...
G→T→C	P ₃₂₋₁	P ₃₂₋₂	...
A→T→G	P ₃₃₋₁	P ₃₃₋₂	...
C→T→A	P ₃₄₋₁	P ₃₄₋₂	...
G→T→C	P ₃₅₋₁	P ₃₅₋₂	...
A→T→G	P ₃₆₋₁	P ₃₆₋₂	...

【図 2 2】

	P ₁	P ₂	...
A→C	P ₁₁	P ₁₂	...
A→G	P ₂₁	P ₂₂	...
A→T	P ₃₁	P ₃₂	...
C→A	P ₄₁	P ₄₂	...
C→G	P ₅₁	P ₅₂	...
C→T	P ₆₁	P ₆₂	...
G→A	P ₇₁	P ₇₂	...
G→C	P ₈₁	P ₈₂	...
G→T	P ₉₁	P ₉₂	...
T→A	P ₁₀₋₁	P ₁₀₋₂	...
T→C	P ₁₁₋₁	P ₁₁₋₂	...
T→G	P ₁₂₋₁	P ₁₂₋₂	...

【図 2 4】

	A	C	G	T	B	C ^M
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	W ₁₆
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	W ₂₅	W ₂₆
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	W ₃₅	W ₃₆
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	W ₄₅	W ₄₆
B	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	W ₅₆
C ^M	W ₆₁	W ₆₂	W ₆₃	W ₆₄	W ₆₅	W ₆₆

10

20

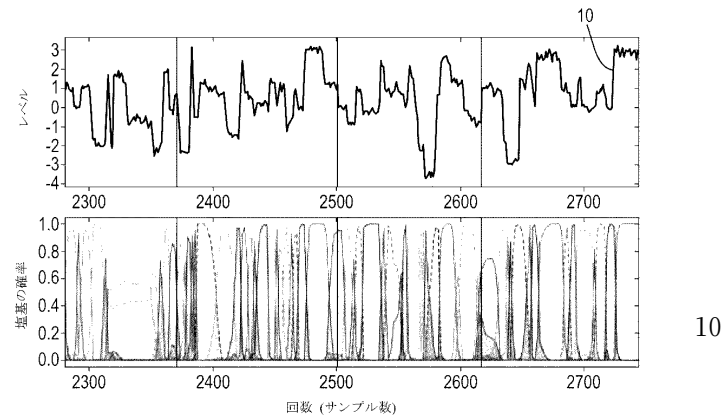
30

40

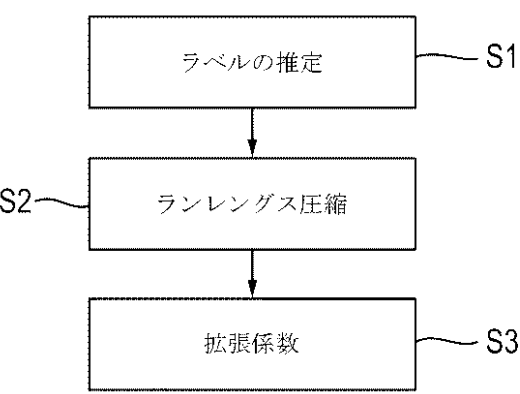
【図 2 5】

	C	C ^M
A	M ₁	M ₂

【図 2 6】



【図 2 7】



【図 2 8】

$f_k^0 = 0 \quad \forall k$	フォワードベクトルの初期化
$f_k^{i+1} = \max_j f_j^i + w_{jk}^i$	最良経路の更新
$t_k^{i+1} = \operatorname{argmax}_j f_j^i + w_{jk}^i$	更新のトレースバック
$S = \max_j f_j^n$	スコアの計算
$P_n = \operatorname{argmax}_j f_j^n$	最終位置における最高スコア
$P_{i-1} = t_{P_i}^i$	最良経路の発見にトレースバック

【図 2 9】

$f_k^0 = p_k^0$	フォワードベクトルの初期化
$f_k^{i+1} = \max_j f_j^i + T_{jk} p_k^{i+1}$	最良経路の更新
$t_k^{i+1} = \operatorname{argmax}_j f_j^i + T_{jk} p_k^{i+1}$	更新のトレースバック
$S = \max_j f_j^n$	スコアの計算
$P_n = \operatorname{argmax}_j f_j^n$	最終位置における最高スコア
$P_{i-1} = t_{P_i}^i$	最良経路の発見にトレースバック

【図 3 0】

$f_k^0 = \log p_k^0$	フォワードベクトルの初期化
$f_k^{i+1} = \max_j f_j^i + T_{jk} \log p_k^{i+1}$	最良経路の更新
$t_k^{i+1} = \operatorname{argmax}_j f_j^i + T_{jk} \log p_k^{i+1}$	更新のトレースバック
$S = \max_j f_j^n$	スコアの計算
$P_n = \operatorname{argmax}_j f_j^n$	最終位置における最高スコア
$P_{i-1} = t_{P_i}^i$	最良経路の発見にトレースバック

10

20

30

40

50

【図 3 1】

f_k^0	$= 0 \quad \forall k$	フォワードベクトルの初期化
f_k^{i+1}	$= \log \sum_j e^{f_j^i + w_{jk}^i}$	フォワードの更新
b_k^N	$= 0 \quad \forall k$	バックワードベクトルの初期化
b_k^{i-1}	$= \log \sum_j e^{b_j^i + w_{kj}^i}$	バックワードの更新
S	$=$	スコア (正規化係数)
p_j^i	$= \frac{1}{S} e^{f_j^i + b_j^i}$	事後確率 (ラベル j)
p_{jk}^i	$= \frac{1}{S} e^{f_j^i + w_{jk}^i + b_k^{i+1}}$	事後確率 (転位 j→k)

【図 3 2】

$m_{j\ j+1}^i$	$= w_{S_j\ S_{j+1}}^i$	状態 S_j から状態 S_{j+1} への移動
$m_{j\ j}^i$	$= w_{S_j\ S_j}^i$	状態 S_j に留まる
$m_{j\ k}^i$	$= 0$	その他

10

【図 3 3】

f_k^0	$= 0 \quad k = 1;$ それ以外なら $-\infty$	フォワードベクトルの初期化
f_k^{i+1}	$= \log \sum_j e^{f_j^i + m_{jk}^i}$	全経路の更新
S	$= f_N^n$	スコアの計算

【図 3 4】

位置 j から j+1 への移動
位置 j から j+1 への移動が維持されている
位置 j に 残留 (新たに到着)
位置 j に 残留 (以前に到着)
その他

20

$$\begin{aligned} &= w_{S_j\ S_{j+1}}^i \\ &= w_{R_j\ S_{j+1}}^i \\ &= w_{S_j\ R_j}^i \\ &= w_{R_j\ R_j}^i \\ &= 0 \end{aligned}$$

$$\begin{aligned} &m_{j\ j+1}^i \\ &m_{j+N+1\ j+1}^i \\ &m_{j\ j+N+1}^i \\ &m_{j+N+1\ j+N+1}^i \\ &m_{j\ k}^i \end{aligned}$$

30

40

50

【図 3 5】

$$\begin{aligned} m_{j+1}^i &= w_{S_j S_{j+1}}^i + r_{S_{j+1} L_{j+1}}^i \\ m_{j+N+1}^i &= w_{R_j S_{j+1}}^i + r_{S_{j+1} L_{j+1}}^i \\ m_{j+N+1}^i &= w_{S_j R_j}^i \\ m_{j+N+1}^i &= w_{R_j R_j}^i \\ m_{j+k}^i &= 0 \end{aligned}$$

位置 j から j+1 への移動

位置 j から j+1 への移動が維持されている

位置 j に 残留 (新たに到着)

位置 j に 残留 (以前に到着)

その他

【図 3 7】

フォワードベクトルの初期化

最良経路の更新

スコアの計算

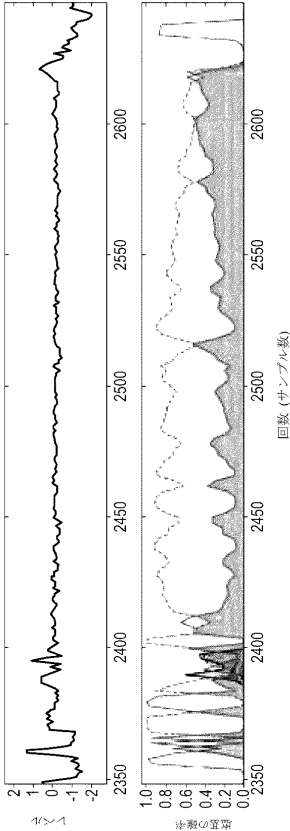
$$\begin{aligned} &= 0 \\ k &= 1; \text{ それ以外なら } -\infty \\ &= \max_j f_j^i + m_{jk}^i \\ S &= f_N^n \end{aligned}$$

f_k^0

f_k^{i+1}

S

【図 3 6】



【図 3 8】

ファンクター

$$\log \sum_j \exp$$

$$\max_j$$

$$\frac{1}{a} \log \sum_j \exp a$$

$$\log \sum_j \exp - \max_j$$

$$p \max_j + (1 - p) \log \sum_j \exp$$

全経路

最良経路

全経路をシャープニング

最良経路を除いたすべて

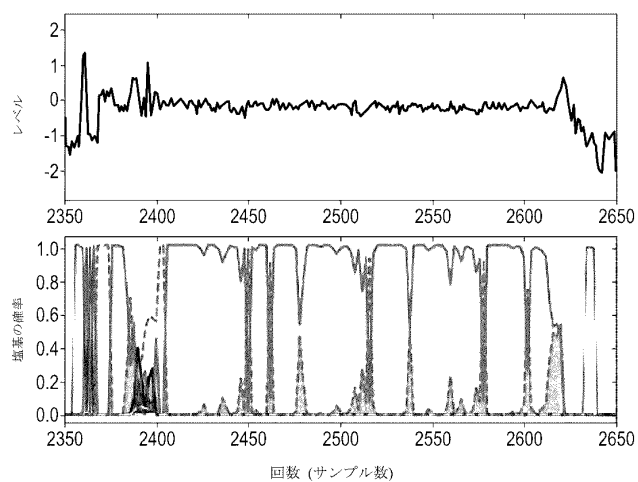
重みを増した最良経路

$L_1 \text{norm}$

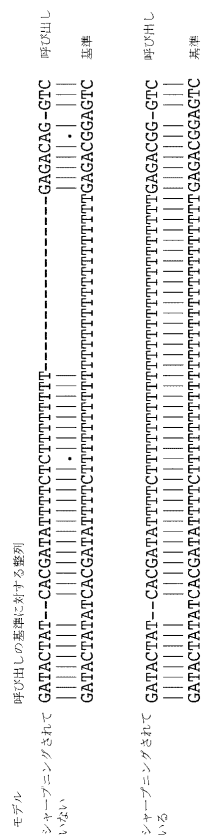
$L_\infty \text{norm}$

$L_a \text{norm}$

【圖 39】



【図 40】



10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2019/053334

A. CLASSIFICATION OF SUBJECT MATTER INV. C12Q1/6869 G01N33/487 ADD.		
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) C12Q G01N		
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, WPI Data, BIOSIS, Sequence Search, EMBASE		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2013/041878 A1 (OXFORD NANOPORE TECH LTD [GB]) 28 March 2013 (2013-03-28) page 37 - page 38; claim 1	1-41
X	LANDRY MATTHEW ET AL: "Analysis of nanopore detector measurements using Machine-Learning methods, with application to single-molecule kinetic analysis", BMC BIOINFORMATICS, BIOMED CENTRAL, LONDON, GB, vol. 8, no. Suppl 7, 1 November 2007 (2007-11-01), page S12, XP021027669, ISSN: 1471-2105, DOI: 10.1186/1471-2105-8-S7-S12 the whole document ----- -/--	1-41
☒ 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 18 February 2020		Date of mailing of the international search report 27/02/2020
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Gabriels, Jan

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2019/053334

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2015/173587 A1 (OXFORD NANOPORE TECH LTD [GB]) 19 November 2015 (2015-11-19) claim 1	1-41
A	----- VLADIMÍR BOZA ET AL: "DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads", PLOS ONE, vol. 12, no. 6, 5 June 2017 (2017-06-05), page e0178751, XP055669713, DOI: 10.1371/journal.pone.0178751 the whole document	1
A	----- Marcus Stoiber ET AL: "BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal", bioRxiv, 1 May 2017 (2017-05-01), XP055472754, DOI: 10.1101/133058 Retrieved from the Internet: URL:https://www.biorxiv.org/content/biorxiv/early/2017/05/01/133058.full.pdf the whole document	1
A	----- YU LI ET AL: "DeepSimulator: a deep simulator for Nanopore sequencing", BIOINFORMATICS., vol. 34, no. 17, 6 April 2018 (2018-04-06), pages 2899-2908, XP055563771, GB ISSN: 1367-4803, DOI: 10.1093/bioinformatics/bty223 -----	1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/GB2019/053334

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
W0 2013041878 A1	28-03-2013	CA 2849624 A1	28-03-2013
		CN 104066850 A	24-09-2014
		EP 2758545 A1	30-07-2014
		EP 3269825 A1	17-01-2018
		JP 6457811 B2	23-01-2019
		JP 2014531901 A	04-12-2014
		JP 2019066484 A	25-04-2019
		US 2016162634 A1	09-06-2016
		US 2017219557 A1	03-08-2017
		US 2019310242 A1	10-10-2019
		W0 2013041878 A1	28-03-2013
W0 2015173587 A1	19-11-2015	EP 3143538 A1	22-03-2017
		US 2017091427 A1	30-03-2017
		W0 2015173587 A1	19-11-2015

10

20

30

40

50

フロントページの続き

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,N
E,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,K
G,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,N
I,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,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

(72)発明者 マッシングム，ティモシー，リー

イギリス国 オックスフォード オーエックス4 4ディーキュー，オックスフォード サイエンス
パーク，エドモンド ハレー ロード，ゴズリング ビルディング

F ターム（参考） 2G045 AA40 DA12
2G060 AA05 AA15 AA19 AD06 AE20 AF06 AF20 HC13 KA09
4B029 AA07 BB20 GB09
4B063 QA08 QA18 QA20 QQ42 QQ52 QS40 QX10