

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7581190号
(P7581190)

(45)発行日 令和6年11月12日(2024.11.12)

(24)登録日 令和6年11月1日(2024.11.1)

(51)国際特許分類

F I

G 1 6 B 40/20 (2019.01)

G 1 6 B 40/20

C 1 2 Q 1/6869(2018.01)

C 1 2 Q 1/6869

Z

請求項の数 15 (全169頁)

(21)出願番号	特願2021-517978(P2021-517978)	(73)特許権者	500358711
(86)(22)出願日	令和2年3月21日(2020.3.21)		イルミナ インコーポレイテッド
(65)公表番号	特表2022-524562(P2022-524562 A)		アメリカ合衆国 カリフォルニア州 9 2
(43)公表日	令和4年5月9日(2022.5.9)		1 2 2 サンディエゴ イルミナ ウェイ
(86)国際出願番号	PCT/US2020/024091	(74)代理人	5 2 0 0
(87)国際公開番号	WO2020/191390		100108453
(87)国際公開日	令和2年9月24日(2020.9.24)		弁理士 村山 靖彦
審査請求日	令和5年3月22日(2023.3.22)	(74)代理人	100110364
(31)優先権主張番号	62/821,602		弁理士 実広 信哉
(32)優先日	平成31年3月21日(2019.3.21)	(74)代理人	100133400
(33)優先権主張国・地域又は機関	米国(US)		弁理士 阿部 達彦
(31)優先権主張番号	62/821,618	(72)発明者	キショール・ジャガナタン
(32)優先日	平成31年3月21日(2019.3.21)		アメリカ合衆国・カリフォルニア・9 2
最終頁に続く			1 2 2・サン・ディエゴ・イルミナ・ウ
			エイ・5 2 0 0
			最終頁に続く

(54)【発明の名称】 人工知能ベースの品質スコアリング

(57)【特許請求の範囲】

【請求項1】

ベースコールの品質スコアを判定するコンピュータ実装方法であって、
クラスターのための入力データをニューラルネットワークベースのベースコーラーによって処理し、前記入力データの特徴マップを生成することであって、前記入力データは、配列決定器によって生成された配列決定画像から導出された画像データである、生成することと、
前記特徴マップを出力層によって処理して分類スコア出力を生成することであって、前記分類スコア出力が、前記クラスターのうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別する、ことと、
前記分類スコア出力に基づいて、前記クラスターに対して塩基をコールすることと、
前記分類スコア出力との相関に基づいて、前記コールされた塩基のビニングされた品質スコアを判定することと、を含み、
前記相関が、
ベースコール誤り率の範囲に対応する、選択されたビニングされた品質スコアのセット、及び
前記選択されたビニングされた品質スコアのセットと、それぞれの分類スコアと、前記それぞれの分類スコアに関するコールされたベースのベースコール誤り率との間の適合を利用し、
前記相関が、前記選択されたビニングされた品質スコアのセットと前記それぞれの分類

スコアとの間の前記適合を使用することを含む、コンピュータ実装方法。

【請求項 2】

前記選択されたビニングされた品質スコアのセットが、訓練中に訓練データを処理することに応じて前記ニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの前記それぞれの分類スコアのサブセットにわたり、

前記それぞれの分類スコアが実数である、請求項 1 に記載のコンピュータ実装方法。

【請求項 3】

前記それぞれの分類スコアは、合計すると 1 になり、かつ前記ニューラルネットワークベースのベースコーラーのソフトマックス出力層によって生成される、指数関数的に正規化されたソフトマックススコアを含む、請求項 1 又は 2 に記載のコンピュータ実装方法。

10

【請求項 4】

推論中に、品質スコアが現在のベースコールサイクルのための設定閾値を下回っているクラスターをベースコールすることを中止すること

を更に含む、請求項 1 ～ 3 のいずれか一項に記載のコンピュータ実装方法。

【請求項 5】

推論中に、継続的なベースコールサイクル後の平均のビニングされた品質スコアが設定閾値を下回っている特定のクラスターをベースコールすることを中止すること

を更に含む、請求項 1 ～ 3 のいずれか一項に記載のコンピュータ実装方法。

【請求項 6】

対応するグラウンドトゥールスペースコールは、複数の配列決定器具、配列決定ケミストリー、及び配列決定プロトコル上で配列決定された十分に特徴付けられたヒト及び非ヒトサンプルから導出される、請求項 1 ～ 5 のいずれか一項に記載のコンピュータ実装方法。

20

【請求項 7】

並列に動作し、かつメモリに連結されている多数のプロセッサと、

前記多数のプロセッサ上で実行され、配列決定器によって生成された配列決定画像のセットから導出された画像データのセットを含む訓練実施例上で訓練され、ニューラルネットワークのベースコール品質予測を、既知の正しいベースコールを識別するベースコール品質グラウンドトゥールスと漸進的に一致させる逆伝播ベースの勾配更新技術を使用して前記ベースコール品質グラウンドトゥールスでラベル付けされる、前記ニューラルネットワークと、

30

前記多数のプロセッサのうちの少なくとも 1 つの上で実行され、1 つ以上の配列決定サイクルにおいて捕捉された配列決定画像から導出されたデータを、1 つ以上のクラスターのためにコールされた 1 つ以上の塩基の品質を判定するための前記ニューラルネットワークに供給する、前記ニューラルネットワークの入力モジュールと、

前記多数のプロセッサのうちの少なくとも 1 つの上で実行され、前記ニューラルネットワークによる分析を前記 1 つ以上のクラスターのためにコールされた前記 1 つ以上の塩基の前記品質を識別する出力に変換する、前記ニューラルネットワークの出力モジュールと、を含む、ニューラルネットワークベースの品質スコアラー。

【請求項 8】

前記ニューラルネットワークが畳み込みニューラルネットワークである、請求項 7 に記載のニューラルネットワークベースの品質スコアラー。

40

【請求項 9】

前記出力モジュールが、高品質、中品質、及び低品質であるベースコール品質の尤度を生成するソフトマックス分類層を更に含み、

前記尤度に基づいて、前記品質を高品質、中品質、又は低品質として分類することを更に含む、請求項 7 に記載のニューラルネットワークベースの品質スコアラー。

【請求項 10】

前記ソフトマックス分類層が、複数の品質スコアを割り当てられている前記品質について尤度を生成し、

前記尤度に基づいて、前記品質に、前記複数の品質スコアのうちの 1 つから品質スコア

50

を割り当てること

を更に含む、請求項 9 に記載のニューラルネットワークベースの品質スコアラー。

【請求項 11】

前記複数の品質スコアが、ベースコール誤り確率に対数的に基づいており、

前記複数の品質スコアが、Q6、Q10、Q15、Q20、Q22、Q27、Q30、Q33、Q37、Q40、及びQ50を含む、請求項 10 に記載のニューラルネットワークベースの品質スコアラー。

【請求項 12】

前記出力モジュールが、前記品質を識別する連続的な値を生成する回帰層を更に含む、請求項 7 ～ 11 のいずれか一項に記載のニューラルネットワークベースの品質スコアラー。

【請求項 13】

補足的な入力モジュールであって、

前記配列決定画像からの前記データを、前記コールされた 1 つ以上 の塩基のための品質予測子値で補足し、

前記品質予測子値を前記配列決定画像からの前記データと共に畳み込みニューラルネットワークに供給する、補足的な入力モジュール

を更に含む、請求項 8 に記載のニューラルネットワークベースの品質スコアラー。

【請求項 14】

前記品質予測子値が、オンライン重複、純度、フェージング、開始 5、6 量体スコア、モチーフ蓄積、エンジネス、近似的なホモポリマー、強度減衰、最後から 2 番目のチャステティ、背景との信号重複 (SOWB)、及び／又はシフトされた純度 G 調整を含む、請求項 13 に記載のニューラルネットワークベースの品質スコアラー。

【請求項 15】

前記品質予測子値が、ピーク高さ、ピーク幅、ピーク位置、相対ピーク位置、ピーク高さ比、ピーク間隔比、及び／又はピーク対応を含む、請求項 13 に記載のニューラルネットワークベースの品質スコアラー。

【発明の詳細な説明】

【技術分野】

【0001】

(優先権出願)

【0002】

本出願は、以下の出願の優先権又は利益を主張する。

【0003】

2019年3月21日に提出された「Training Data Generation for Artificial Intelligence-Based Sequencing」と題する米国仮特許出願第62/821,602号(代理人整理番号ILLM1008-1/IP-1693-PRV)、

【0004】

2019年3月21日に提出された「Artificial Intelligence-Based Generation of Sequencing Metadata」と題する米国仮特許出願第62/821,618号(代理人整理番号ILLM1008-3/IP-1741-PRV)、

【0005】

2019年3月21日に提出された「Artificial Intelligence-Based Base Calling」と題する米国仮特許出願第62/821,681号(代理人整理番号ILLM1008-4/IP-1744-PRV)、

【0006】

2019年3月21日に提出された「Artificial Intelligence-Based Quality Scoring」と題する米国仮特許出願第62/821,724号(代理人整理番号ILLM1008-7/IP-1747-PRV)、

【0007】

2019年3月21日に出願された「Artificial Intelligence-Based Sequencing」と題する米国仮特許出願第62/821,766号（代理人整理番号ILLM1008-9/IP-1752-PRV）、

【0008】

2019年6月14日に出願された「Training Data Generation for Artificial Intelligence-Based Sequencing」と題する蘭国特許出願第2023310号（代理人整理番号ILLM1008-11/IP-1693-NL）、

【0009】

2019年6月14日に出願された「Artificial Intelligence-Based Generation of Sequencing Metadata」と題する蘭国特許出願第2023311号（代理人整理番号ILLM1008-12/IP-1741-NL）、

【0010】

2019年6月14日に出願された「Artificial Intelligence-Based Base Calling」と題する蘭国特許出願第2023312号（代理人整理番号ILLM1008-13/IP-1744-NL）、

【0011】

2019年6月14日に出願された「Artificial Intelligence-Based Quality Scoring」と題する蘭国特許出願第2023314号（代理人整理番号ILLM1008-14/IP-1747-NL）、及び

【0012】

2019年6月14日に出願された「Artificial Intelligence-Based Sequencing」と題する蘭国特許出願第2023316号（代理人整理番号ILLM1008-15/IP-1752-NL）。

【0013】

2020年3月20日に出願された「Training Data Generation for Artificial Intelligence-Based Sequencing」と題する米国特許出願第16/825,987号（代理人整理番号ILLM1008-16/IP-1693-US）、

【0014】

2020年3月20日に出願された「Training Data Generation for Artificial Intelligence-Based Sequencing」と題する米国特許出願第16/825,991号（代理人整理番号ILLM1008-17/IP-1741-US）、

【0015】

2020年3月20日に出願された「Artificial Intelligence-Based Base Calling」と題する米国特許出願第16/826,126号（代理人整理番号ILLM1008-18/IP-1744-US）、

【0016】

2020年3月20日に出願された「Artificial Intelligence-Based Quality Scoring」と題する米国特許出願第16/826,134号（代理人整理番号ILLM1008-19/IP-1747-US）、

【0017】

2020年3月21日に出願された「Artificial Intelligence-Based Sequencing」と題する米国特許出願第16/826,168号（代理人整理番号ILLM1008-20/IP-1752-PRV）、

【0018】

本願と同時に提出され、その後PCT国際公開第WO _____号とし

10

20

30

40

50

て公開されている「Training Data Generation for Artificial Intelligence-Based Sequencing」と題するPCT特許出願第PCT _____号（代理人整理番号ILLM1008-21/IP-1693-PCT）、

【0019】

本願と同時に出願され、その後、PCT国際公開第WO _____号として公開されている「Artificial Intelligence Based Generation of Sequencing Metadata」と題するPCT特許出願第PCT _____号（代理人整理番号第ILLM1008-22/IP-1741-PCT）、

【0020】

本願と同時に出願され、その後PCT国際公開第WO _____号として公開されている「Artificial Intelligence-Based Base Calling」と題するPCT特許出願第PCT _____号（代理人整理番号第ILLM1008-23/IP-1744-PCT）、及び

【0021】

本願と同時に出願され、その後PCT国際公開第WO _____号として公開されている「Artificial Intelligence-Based Sequencing」と題するPCT特許出願第PCT _____号（代理人整理番号ILLM1008-25/IP-1752-PCT）。

【0022】

優先権出願は、本明細書に完全に記載されているかのように、全ての目的のために参照により本明細書に組み込まれる。

【0023】

（組み込み）

【0024】

以下は、本明細書に完全に記載されているかのように、全ての目的のために参照により組み込まれる。

【0025】

2019年5月16日に出願された「Systems and Devices for Characterization and Performance Analysis of Pixel-Based Sequencing」と題する米国仮特許出願第62/849,091号（代理人整理番号ILLM1011-1/IP-1750-PRV）、

【0026】

2019年5月16日に出願された「Base Calling Using Convolutionals」と題する米国特許仮出願第62/849,132号（代理人整理番号ILLM1011-2/IP-1750-PR2）、

【0027】

2019年5月16日に出願された「Base Calling Using Compact Convolutionals」と題する米国仮特許仮出願第62/849,133号（代理人整理番号ILLM1011-3/IP-1750-PR3）、

【0028】

2020年2月20日に出願された「Artificial Intelligence-Based Base Calling of Index Sequences」と題する米国仮特許出願第62/979,384号（代理人整理番号ILLM1015-1/IP-1857-PRV）、

【0029】

2020年2月20日に出願された「Artificial Intelligence-Based Many-To-Many Base Calling」と題する米国仮特許出願第62/979,414号（代理人整理番号ILLM1016-1/IP-18

10

20

30

40

50

58-PRV)、

【0030】

2020年2月20日に出願された「Knowledge Distillation-Based Compression of Artificial Intelligence-Based Base Caller」と題する米国仮特許出願第62/979,385号(代理人整理番号ILLM1017-1/IP-1859-PRV)、

【0031】

2020年2月20日に出願された「Multi-Cycle Cluster Based Real Time Analysis System」と題する米国特許仮出願第62/979,412号(代理人整理番号ILLM1020-1/IP-1866-PRV)、

10

【0032】

2020年2月20日に出願された「Data Compression for Artificial Intelligence-Based Base Calling」と題する米国仮特許出願第62/979,411号(代理人整理番号ILLM1029-1/IP-1964-PRV)、

【0033】

2020年2月20日に出願された「Squeezing Layer for Artificial Intelligence-Based Base Calling」と題する米国仮特許出願第62/979,399号(代理人整理番号ILLM1030-1/IP-1982-PRV)、

20

【0034】

Liu P, Hemani A, Paul K, Weis C, Jung M, Wehn N. 3D-Stacked Many-Core Architecture for Biological Sequence Analysis Problems. Int J Parallel Prog. 2017;45(6):1420-60、

【0035】

Z. Wu, K. Hammad, R. Mittmann, S. Magierowski, E. Ghafar-Zadeh, and X. Zhong, 「FPGA-Based DNA Basecalling Hardware Acceleration」, in Proc. IEEE 61st Int. Midwest Symp. Circuits Syst., Aug. 2018, pp. 1098-1101、

30

【0036】

Z. Wu, K. Hammad, E. Ghafar-Zadeh, and S. Magierowski, 「FPGA-Accelerated 3rd Generation DNA Sequencing」, in IEEE Transactions on Biomedical Circuits and Systems, Volume 14, Issue 1, Feb. 2020, pp. 65-74、

【0037】

Prabhakar et al., 「Plasticine: A Reconfigurable Architecture for Parallel Patterns」, ISCA'17, June 24-28, 2017, Toronto, ON, Canada、

40

【0038】

M. Lin, Q. Chen, and S. Yan, 「Network in Network」, in Proc. of ICLR, 2014、

【0039】

L. Sifre, 「Rigid-motion Scattering for Image Classification, Ph. D. thesis, 2014、

【0040】

50

L. Sifre and S. Mallat, 「Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination」, in Proc. of CVPR, 2013、

[0041]

F. Chollet, 「Xception: Deep Learning with Depthwise Separable Convolutions」, in Proc. of CVPR, 2017、

[0042]

X. Zhang, X. Zhou, M. Lin, and J. Sun, 「Shuffle Net: An Extremely Efficient Convolutional Neural Network for Mobile Devices」, in arXiv:1707.01083, 2017、

10

[0043]

K. He, X. Zhang, S. Ren, and J. Sun, 「Deep Residual Learning for Image Recognition」, in Proc. of CVPR, 2016、

[0044]

S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, 「Aggregated Residual Transformations for Deep Neural Networks」, in Proc. of CVPR, 2017、

20

[0045]

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, 「Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications」, in arXiv:1704.04861, 2017、

[0046]

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, 「MobileNetV2: Inverted Residuals and Linear Bottlenecks」, in arXiv:1801.04381v3, 2018、

30

[0047]

Z. Qin, Z. Zhang, X. Chen, and Y. Peng, 「FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy」, in arXiv:1802.03750, 2018、

[0048]

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587, 2017、

40

[0049]

J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. arXiv preprint arXiv:1611.10012, 2016、

[0050]

S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A

50

. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, 「WAVENET: A GENERATIVE MODEL FOR RAW AUDIO」, arXiv:1609.03499, 2016、

[0051]

S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta and M. Shoeybi, 「DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH」, arXiv:1702.07825, 2017、

[0052]

F. Yu and V. Koltun, 「MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS」, arXiv:1511.07122, 2016、

[0053]

K. He, X. Zhang, S. Ren, and J. Sun, 「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」, arXiv:1512.03385, 2015、

[0054]

R. K. Srivastava, K. Greff, and J. Schmidhuber, 「HIGHWAY NETWORKS」, arXiv:1505.00387, 2015、

[0055]

G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, 「DENSELY CONNECTED CONVOLUTIONAL NETWORKS」, arXiv:1608.06993, 2017、

[0056]

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, 「GOING DEEPER WITH CONVOLUTIONS」, arXiv:1409.4842, 2014、

[0057]

S. Ioffe and C. Szegedy, 「BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT」, arXiv:1502.03167, 2015、

[0058]

J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, 「DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE」, arXiv:1704.03669, 2017、

[0059]

L. C. Piqueras, 「AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION」, Tampere University of Technology, 2016、

[0060]

J. Wu, 「Introduction to Convolutional Neural Networks」, Nanjing University, 2017、

[0061]

10

20

30

40

50

「Illumina CMOS Chip and One-Channel SBS Chemistry」, Illumina, Inc. 2018, 2 pages、

【0062】

「skikit-image/peak.py at master」, GitHub, 5 pages, [2018-11-16に検索]。インターネット<URL:<https://github.com/scikit-image/scikit-image/blob/master/skimage/feature/peak.py#L25>>から検索、

【0063】

「3.3.9.11. Watershed and random walker for segmentation」, Scipy lecture notes, 2 pages, [2018-11-13に検索]。インターネット<URL:http://scipy-lectures.org/packages/scikit-image/automatic_examples/plot_segmentations.html>から検索、

10

【0064】

Mordvintsev, Alexander and Revision, Abid K., 「Image Segmentation with Watershed Algorithm」, Revision 43532856, 2013, 6 pages [2018-11-13に検索]。インターネット<URL:https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_watershed/py_watershed.html>から検索、

20

【0065】

Mzur, 「Watershed.py」, 25 October 2017, 3 pages, [2018-11-13に検索]。インターネット<URL:<https://github.com/mzur/watershed/blob/master/Watershed.py>>から検索、

【0066】

Thakur, Pratibha, et. al. 「A Survey of Image Segmentation Techniques」, International Journal of Research in Computer Applications and Robotics, Vol. 2, Issue. 4, April 2014, Pg. :158-165、

30

【0067】

Long, Jonathan, et. al., 「Fully Convolutional Networks for Semantic Segmentation」, :IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 39, Issue 4, 1 April 2017, 10 pages、

40

【0068】

Ronneberger, Olaf, et. al., 「U-net: Convolutional networks for biomedical image segmentation.」 In International Conference on Medical image computing and computer-assisted intervention, 18 May 2015, 8 pages、

【0069】

Xie, W., et. al., 「Microscopy cell counting and detection with fully convolutional regression networks」, Computer methods in bi

50

omechanics and biomedical engineering: Imaging & Visualization, 6 (3), pp. 283–292, 2018、
【0070】

Xie, Yuanpu, et al., 「Beyond classification: structured regression for robust cell detection using convolutional neural network」, International Conference on Medical Image Computing and Computer-Assisted Intervention. October 2015, 12 pages、
【0071】

10

Snuverink, I. A. F., 「Deep Learning for Pixelwise Classification of Hyperspectral Images」, Master of Science Thesis, Delft University of Technology, 23 November 2017、19 pages、
【0072】

Shevchenko, A., 「Keras weighted categorical_crossentropy」, 1 page, [2019-01-15に検索]。インターネット<URL: <https://gist.github.com/skeeeet/cad06d584548fb45eece1d4e28cfa98b>>から検索、
【0073】

20

van den Assem, D. C. F., 「Predicting periodic And chaotic signals using Wavenets」, Master of Science Thesis, Delft University Of Technology, 18 August 2017, Pages 3–38、
【0074】

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, 「CONVOLUTIONAL NETWORKS」, Deep Learning, MIT Press, 2016、及び
【0075】

30

J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, 「RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS」, arXiv:1512.07108, 2017。
【0076】

(発明の分野)

【0077】

本開示の技術は、人工知能コンピュータ及びデジタルデータ処理システム、並びに、知能をエミュレーションするための対応するデータ処理方法及び製品（すなわち、知識ベースのシステム、推論システム及び知識取得システム）に関すると共に、不確実性（例えば、ファジー論理システム）、適応システム、機械学習システム、及び人工ニューラルネットワークを用いて推論するためのシステムを含む。具体的には、開示される技術は、データを分析するための深層畳み込みニューラルネットワークなどの深層ニューラルネットワークを使用することに関する。

40

【背景技術】

【0078】

このセクションで説明される主題は、単にこのセクションにおける言及の結果として、先行技術であると想定されるべきではない。同様に、このセクションで言及した問題、又は背景として提供された主題に関連する問題は、先行技術において以前に認識されていると想定されるべきではない。このセクションの主題は、単に、異なるアプローチを表し、

50

それ自体はまた、特許請求される技術の実施に対応し得る。

【0079】

深層ニューラルネットワークは、高レベル機能を連続的にモデル化するために、複数の非線形及び複雑な変換層を使用する、人工ニューラルネットワークの類である。深層ニューラルネットワークは、観測された出力と予測出力との間の差を伝達してパラメータを調整する逆伝搬を介してフィードバックを提供する。深層ニューラルネットワークは、大きな訓練データセットの利用可能性、並列分散コンピューティングの能力、及び高度な訓練アルゴリズムと共に進化している。深層ニューラルネットワークは、コンピュータビジョン、音声認識、及び自然言語処理などの多数のドメインにおける主な進歩を促進している。

【0080】

畳み込みニューラルネットワーク (CNN) 及び反復ニューラルネットワーク (RNN) は、深層ニューラルネットワークの構成要素である。畳み込みニューラルネットワークは、特に、畳み込み層、非線形層、及びプーリング層を含む構造で画像認識に成功してきた。反復ニューラルネットワークは、パーセプトロン、長い短期メモリユニット、及びゲートされた反復単位のような構成単位間の周期的接続を有する入力データの連続的な情報を利用するように設計される。加えて、多くの他の出現深層ニューラルネットワークが、深層時空間ニューラルネットワーク、多次元反復ニューラルネットワーク、及び畳み込み自動エンコーダなどの限定された状況に関して提案されてきた。

【0081】

深層ニューラルネットワークの訓練の目標は、各層における重みパラメータの最適化であり、このことは、より好適な階層表現がデータから学習され得るように、より単純な特徴を複雑な特徴に徐々に組み合わせる。最適化プロセスの単一サイクルは、以下のように構成される。まず、訓練データセットと仮定すると、前方へのパスは、各層内の出力を順次計算し、ネットワークを通して機能信号を順方向に伝搬する。最終出力層において、客観的な損失関数は、推論された出力と所与のラベルとの間の誤差を測定する。訓練誤差を最小化するために、後方へのパスは、連鎖ルールを使用して、誤差信号を逆伝搬し、ニューラルネットワーク全体の全ての重みに対する勾配を計算する。最後に、確率勾配降下に基づいて最適化アルゴリズムを使用して確率パラメータが更新される。バッチ勾配降下が完全データセットごとにパラメータ更新するのに対し、確率的勾配降下は、データ例の各々の小さいセットについて更新を実行することによって確率的近似値を提供する。いくつかの最適化アルゴリズムは確率的勾配降下に由来する。例えば、Adagrad and Adam訓練アルゴリズムは、それぞれ、各パラメータの更新頻度及び勾配のモーメントに基づいて学習率を適応的に修正しながら、確率的勾配降下を実行する。

【0082】

深層ニューラルネットワークの訓練における別のコア要素は規則化であり、規則化は、過剰適合を回避し、したがって良好な一般化性能を達成することを意図する戦略を指す。例えば、重み減衰は、重みパラメータがより小さい絶対値に収束するように、客観的な損失関数にペナルティ項を追加する。ドロップアウトは、訓練中にニューラルネットワークから隠れたユニットをランダムに除去し、可能なサブネットワークの集合体と見なすことができる。ドロップアウトの能力を向上させるために、新たな起動関数、maxout、及びrnnDropと呼ばれる反復性ニューラルネットワークに対するドロップアウトのバリエーションが提案されている。更に、バッチ正規化は、ミニバッチ内の各アクティブ化に関するスカラー特徴の正規化を介した新たな規則化方法を提供し、各々の平均及び分散をパラメータとして学習する。

【0083】

配列データが多次元及び高次元であると仮定すると、深層ニューラルネットワークは、それらの広範な適用性及び強化された予測能力により、生物情報科学研究のためにかなり有望である。畳み込みニューラルネットワークは、モチーフ発見、病原性変異体の同定、及び遺伝子発現推論などのゲノミクスにおける配列に基づく問題を解決するために採用されている。畳み込みニューラルネットワークは、DNAを研究するのに特に有用な重み共

10

20

30

40

50

有戦略を使用するが、これは、短い配列モチーフを捕捉することができ、この配列モチーフは、有意な生物学的機能を有すると推定されるDNA中の局所的パターンを再現する。畳み込みニューラルネットワークの顕著な特徴は、畳み込みフィルタの使用である。

【0084】

精巧に設計され、手動で巧妙に作り上げられた特徴に基づく従来の分類アプローチとは異なり、畳み込みフィルタは、知識の情報表現に生入力データをマッピングするプロセスに類似した特徴の適応学習を実行する。この意味では、畳み込みフィルタは、そのようなフィルタのセットが入力内の関連するパターンを認識し、訓練手順中にそれ自体を更新することができるため、一連のモチーフスキャナーとして機能する。反復ニューラルネットワークは、タンパク質又はDNA配列などの様々な長さの連続的データにおける長距離依存性を捉えることができる。

10

【0085】

したがって、テンプレートの生成及びベースコールのための、理にかなった深層学習ベースの枠組みを使用する機会が生じる。

【0086】

ハイスループット技術の時代では、努力ごとに最も低いコストで解釈可能なデータの最高収率を蓄積することは、重大な課題を残している。クラスター形成のためにブリッジ増幅を利用するものなどの核酸配列決定のクラスターベースの方法は、核酸配列決定のスループットを増加させる目的に有益な貢献をしている。これらのクラスターベースの方法は、固体支持体上に固定化された核酸の密集した集団を配列決定することに依存し、典型的には、固体支持体上の別個の場所に位置する複数のクラスターを同時に配列決定する過程で生成される光信号を抑制するための画像解析ソフトウェアの使用を伴う。

20

【0087】

しかしながら、このような固相核酸クラスターベースの配列決定技術は、達成することができるスループットの量を制限する相当な障害に直面している。例えば、クラスターベースの配列決定方法では、空間的に分解されるには互いに物理的に近接し過ぎる、又は実際には、固体支持体上で物理的に重なり合う2つ又はそれ以上のクラスターの核酸配列を決定することは障害物をもたらす可能性がある。例えば、現在の画像解析ソフトウェアは、2つの重複クラスターのうちのどれから光信号が発せられたかを判定するための貴重な時間及び計算リソースを必要とする場合がある。結果として、得られ得る核酸配列情報の量及び／又は質に関して、様々な検出プラットフォームにとって妥協が不可避である。

30

【0088】

高密度核酸凝集体ベースのゲノミクス法は、ゲノム分析の他の領域にも同様に延在する。例えば、核酸クラスターベースのゲノミクスは、配列決定用途、診断及びスクリーニング、遺伝子発現分析、エピジェネティクス分析、多型の遺伝分析などに使用することができる。これらの核酸クラスターベースのゲノミクス技術のそれぞれは、厳密に近接して又は空間的に重複する核酸クラスターから生成されたデータを分解する能力がない場合に制限される。

【0089】

明らかに、ゲノミクス（例えば、任意の及び全ての動物、植物、微生物又は他の生物学的種又は集団のゲノム特性評価のための）、薬理ゲノミクス、トランスクリプトミクス、診断、予後、生物医学的リスク評価、臨床及び研究遺伝学、個人向け医療、薬物有効性及び薬物相互作用評価、獣医医学、農業、進化、及び生物学的研究、水性培養、林業、海洋調査、生態学的管理、及び環境管理、並びに他の目的を含む、様々な用途で迅速かつ費用効果の高い方法で取得できる核酸配列データの質と量を向上させる必要がある。

40

【発明の概要】

【課題を解決するための手段】

【0090】

開示される技術は、ハイスループット核酸配列決定技術におけるスループットのレベルを増加させることを含めて、これら及び同様のニーズに対処するニューラルネットワーク

50

ベースの方法及びシステムを提供すると共に、他の関連する利点を提供する。

【0091】

特許又は出願ファイルは、カラーで創作された少なくとも1つの図面を含む。カラー図面（単数又は複数）を有するこの特許又は特許出願公開のコピーは、必要な料金の要求及び支払いの際に、序によって提供される。カラー図面はまた、補助コンテンツタブを介してPAIR（patent application information retrieval：特許出願情報検索）で利用可能であってもよい。

【0092】

図面では、同様の参照文字は、概して、異なる図全体を通して同様の部分を指す。また、図面は必ずしも縮尺通りではなく、その代わりに、開示された技術の原理を例示することを強調している。以下の説明では、開示される技術の様々な実施態様が、以下の図面を参照して説明される。

【図面の簡単な説明】

【0093】

【図1】一実施態様による、ベースコールするためにRTAベースコーラーによって使用される処理段階を示す。

【図2】開示されたニューラルネットワークベースのベースコーラーを使用してベースコールする一実施態様を示す。

【図3】ニューラルネットワークベースのテンプレート生成器の出力から識別されたクラスター中心の場所／位置情報をサブピクセルドメインからピクセルドメインに変換する一実施態様である。

【図4】サイクル固有及び画像チャネル固有の変換を使用して参照クラスター中心からいわゆる「変換されたクラスター中心」を導き出す一実施態様である。

【図5】ニューラルネットワークベースのベースコーラーに供給される入力データの一部である画像パッチを示す。

【図6】単一のターゲットクラスターがニューラルネットワークベースのベースコーラーによってベースコールされているときに距離チャンネルのための距離値を決定する一実施態様を示す。

【図7】ピクセルとターゲットクラスターとの間で計算される距離値をピクセル単位で符号化する一実施態様を示す。

【図8a】複数のターゲットクラスターがニューラルネットワークベースのベースコーラーによって同時にベースコールされているときに距離チャンネルのための距離値を決定する一実施態様を示す。

【図8b】ターゲットクラスターのそれぞれについて、ピクセル中心－最も近いクラスター中心間の距離に基づいて決定されたいくつかの最も近いピクセルを示す。

【図9】ピクセルとクラスターのうちの最も近いクラスターとの間で計算される最小距離値をピクセル単位で符号化する一実施態様を示す。

【図10】本明細書において「クラスター形状データ」と称される、ピクセル－クラスター間の分類／帰属／カテゴリー化を使用する一実施態様を示す。

【図11】クラスター形状データを使用して距離値を計算する一実施態様を示す。

【図12】ピクセルと割り当てられたクラスターとの間で計算される距離値をピクセル単位で符号化する一実施態様を示す。

【図13】異なる配列決定サイクルのためのデータの処理を分離するために使用されるニューラルネットワークベースのベースコーラーの特殊化したアーキテクチャの一実施態様を示す。

【図14】分離された畳み込みの一実施態様を示す。

【図15a】組み合わせ畳み込みの一実施態様を示す。

【図15b】組み合わせ畳み込みの別の一実施態様を示す。

【図16】各畳み込み層が畳み込みフィルタのバンクを有する、ニューラルネットワークベースのベースコーラーの畳み込み層の一実施態様を示す。

10

20

30

40

50

【図 1 7】画像チャネルを補足するスケーリングチャネルの 2 つの構成を示す。

【図 1 8 a】赤色画像及び緑色画像を生成する単一の配列決定サイクルのための入力データの一実施態様を示す。

【図 1 8 b】画像チャネルから生成された特徴マップに組み込まれる加法的バイアスを供給する距離チャネルの一実施態様を示す。

【図 1 9 a】単一のターゲットクラスターをベースコールする一実施態様を示す。

【図 1 9 b】単一のターゲットクラスターをベースコールする一実施態様を示す。

【図 1 9 c】単一のターゲットクラスターをベースコールする一実施態様を示す。

【図 2 0】複数のターゲットクラスターに同時にベースコールする一実施態様を示す。

【図 2 1】複数の継続的な配列決定サイクルにおける複数のターゲットクラスターを同時にベースコールし、それによって複数のターゲットクラスターのそれぞれのためのベースコール配列を同時に生成する一実施態様を示す。

10

【図 2 2】単一のクラスターのベースコールの実施態様のための次元数ダイアグラムを示す。

【図 2 3】複数のクラスター、単一の配列決定サイクルのベースコールの実施態様のための次元数ダイアグラムを示す。

【図 2 4】複数のクラスター、複数の配列決定サイクルのベースコールの実施態様のための次元数ダイアグラムを示す。

【図 2 5 a】多重サイクル入力データの例示的なアレイ化された入力構成を示す。

【図 2 5 b】多重サイクル入力データの例示的なスタック化された入力構成を示す。

20

【図 2 6 a】ベースコールされているターゲットクラスターの中心を中心ピクセルの中心に配置するように画像パッチのピクセルをリフ্রেームする一実施態様を示す。

【図 2 6 b】(i) 中心ピクセルの中心がターゲットクラスターの中心と一致し、(i i) 非中心ピクセルがターゲットクラスターの中心から等距離である、別の例示的なリフ্রেームされた／シフトされた画像パッチを示す。

【図 2 7】標準的な畳み込みニューラルネットワーク及びリフ্রেームされた入力を使用して、現在の配列決定サイクルにおける単一のターゲットクラスターをベースコールする一実施態様を示す。

【図 2 8】標準的な畳み込みニューラルネットワーク及び整列された入力を使用して、現在の配列決定サイクルにおける複数のターゲットクラスターをベースコールする一実施態様を示す。

30

【図 2 9】標準的な畳み込みニューラルネットワーク及び整列された入力を使用して、複数の配列決定サイクルにおける複数のターゲットクラスターをベースコールする一実施態様を示す。

【図 3 0】ニューラルネットワークベースのベースコーラーを訓練する一実施態様を示す。

【図 3 1 a】ニューラルネットワークベースのベースコーラーとして使用されるハイブリッドニューラルネットワークの一実施態様を示す。

【図 3 1 b】現在の隠れ状態表現を生成するためにハイブリッドニューラルネットワークの反復モジュールによって使用される 3 D 畳み込みの一実施態様を示す。

【図 3 2】ベースコールされる一連の t 個の配列決定サイクルのうちの単一の配列決定サイクルのためのサイクルごとの入力データを畳み込みモジュールの畳み込み層のカスケードによって処理する一実施態様を示す。

40

【図 3 3】単一の配列決定サイクルのサイクルごとの入力データを、畳み込みモジュールの畳み込み層のカスケードによって生成された対応する畳み込み表現と混合する一実施態様を示す。

【図 3 4】継続的な配列決定サイクルの平坦化された混合表現をスタックとして配列する一実施態様を示す。

【図 3 5 a】図 3 4 のスタックに順方向及び逆方向での 3 D 畳み込みの反復適用を受けさせ、一連内の t 個の配列決定サイクルのそれぞれにおけるクラスターのそれぞれのためのベースコールを生成する一実施態様を示す。

50

【図35b】3D量み込みを適用する長・短期記憶（LSTM）ネットワークの入力ゲート、活性化ゲート、忘却ゲート、及び出力ゲートによって3D入力体積 $x(t)$ （平坦化された混合表現のグループを含む）を処理する一実施態様を示す。LSTMネットワークは、ハイブリッドニューラルネットワークの反復モジュールの一部である。

【図36】ニューラルネットワークベースのベースコーラーを訓練するために使用される訓練データ内のトリヌクレオチド（3量体）を平衡化する一実施態様を示す。

【図37】RTAベースコーラーのベースコール精度をニューラルネットワークベースのベースコーラーと比較する。

【図38】RTAベースコーラーのタイル間一般化を同じタイル上でのニューラルネットワークベースのベースコーラーのものと比較する。

【図39】RTAベースコーラーのタイル間一般化を、同じタイル上での、及び異なるタイル上でのニューラルネットワークベースのベースコーラーのものと比較する。

【図40】同様に、RTAベースコーラーのタイル間一般化を異なるタイル上でのニューラルネットワークベースのベースコーラーのものと比較する。

【図41】ニューラルネットワークベースのベースコーラーへの入力として供給される異なるサイズの画像パッチが、ベースコール精度にどのような変化をもたらすかを示す。

【図42】A. baumannii及びE. coliからの訓練データ上でのニューラルネットワークベースのベースコーラーのレーン間一般化を示す。

【図43】A. baumannii及びE. coliからの訓練データ上でのニューラルネットワークベースのベースコーラーのレーン間一般化を示す。

【図44】A. baumannii及びE. coliからの訓練データ上でのニューラルネットワークベースのベースコーラーのレーン間一般化を示す。

【図45】A. baumannii及びE. coliからの訓練データ上でのニューラルネットワークベースのベースコーラーのレーン間一般化を示す。

【図46】図42、図43、図44、及び図45に関して上述したレーン間一般化のための誤りプロファイルを示す。

【図47】図46の誤りプロファイルによって検出された誤りの発生源が緑色チャネルにおける低いクラスター強度に帰属することを示す。

【図48】2回の配列決定実行（リード1及びリード2）についてのRTAベースコーラー及びニューラルネットワークベースのベースコーラーの誤りプロファイルを比較する。

【図49a】4つの異なる器具上でのニューラルネットワークベースのベースコーラーの実行間一般化を示す。

【図49b】同じ器具上で実施された4つの異なる実行上でのニューラルネットワークベースのベースコーラーの実行間一般化を示す。

【図50】ニューラルネットワークベースのベースコーラーを訓練するために使用される訓練データのゲノム統計を示す。

【図51】ニューラルネットワークベースのベースコーラーを訓練するために使用される訓練データのゲノムコンテキストを示す。

【図52】ロングリード（例えば、 2×250 ）のベースコールにおけるニューラルネットワークベースのベースコーラーのベースコール精度を示す。

【図53】ニューラルネットワークベースのベースコーラーが、画像パッチ間にわたって中央クラスターピクセル及びその隣接ピクセルにどのように着目するかの一実施態様を示す。

【図54】一実施態様による、ニューラルネットワークベースのベースコーラーを訓練及び実行するために使用される様々なハードウェア構成要素及び構成を示す。他の実施態様では、異なるハードウェア構成要素及び構成が使用される。

【図55】ニューラルネットワークベースのベースコーラーを使用して実施することができる様々な配列決定タスクを示す。

【図56】t分布型確率的近傍埋め込み法（t-SNE）によって可視化された散布図であり、ニューラルネットワークベースのベースコーラーのベースコール結果を表現する。

10

20

30

40

50

【図57】品質スコアリングのためにニューラルネットワークベースのベースコーラーによって作成されたベースコール信頼確率を選択する一実施態様を示す。

【図58】ニューラルネットワークベースの品質スコアリングの一実施態様を示す。

【図59a】品質スコアとニューラルネットワークベースのベースコーラーによって行われたベースコール信頼性予測との間の対応の一実施態様を示す。

【図59b】品質スコアとニューラルネットワークベースのベースコーラーによって行われたベースコール信頼性予測との間の対応の一実施態様を示す。

【図60】推論中にニューラルネットワークベースのベースコーラーによって行われたベースコール信頼性予測から品質スコアを推論する一実施態様を示す。

【図61】配列決定画像から導出された入力データを処理し、品質指標を直接生成するように、ニューラルネットワークベースの品質スコアラーを訓練する一実施態様を示す。

【図62】推論中にニューラルネットワークベースの品質スコアラーの出力として品質指標を直接生成する一実施態様を示す。

【図63A】配列決定システムの一実施態様を示す。配列決定システムは、構成可能なプロセッサを含む。

【図63B】配列決定システムの一実施態様を示す。配列決定システムは、構成可能なプロセッサを含む。

【図63C】ベースコールセンサー出力など、配列決定システムからのセンサーデータの分析のためのシステムの簡略ブロック図である。

【図64A】ホストプロセッサによって実行される実行時プログラムの機能を含む、ベースコール動作の態様を示す簡略図である。

【図64B】図63Cに示されるものなどの構成可能プロセッサの構成の簡略図である。

【図65】本明細書に開示される技術を実施するために、図63Aの配列決定システムによって使用され得るコンピュータシステムである。

【図66】データ正規化及びデータ増強を含むことができる、データ前処理の異なる実施態様を示す。

【図67】ニューラルネットワークベースのベースコーラーが、細菌データ上で訓練され、ヒトデータ上で試験されるとき（細菌データ及びヒトデータは同じアッセイを共有する（例えば、両方ともイントロンデータを含む））、図66のデータ正規化技術（DeepRTA（norm））及びデータ増強技術（DeepRTA（augment））がベースコール誤り率を減少させることを示す。

【図68】ニューラルネットワークベースのベースコーラーが、非エクソンデータ（例えば、イントロンデータ）上で訓練され、エクソンデータ上で試験されるとき、図66のデータ正規化技術（DeepRTA（norm））及びデータ増強技術（DeepRTA（augment））がベースコール誤り率を減少させることを示す。

【発明を実施するための形態】

【0094】

以下の説明は、開示された技術を当業者が作製及び使用することを可能にするために提示され、特定の用途及びその要件に関連して提供される。開示される実施態様に対する様々な修正は、当業者には容易に明らかとなり、本明細書で定義される一般原理は、開示される技術の趣旨及び範囲から逸脱することなく、他の実施態様及び用途に適用され得る。したがって、開示される技術は、示される実施態様に限定されることを意図するものではなく、本明細書に開示される原理及び特徴と一致する最も広い範囲を与えられるものである。

【0095】

導入

【0096】

塩基がデジタル画像のシーケンスで分類されるとき、ニューラルネットワークは、過去及び将来のサイクルの画像チャネルと共に現在のサイクル内の複数の画像チャネルを処理する。クラスター内において、ストランドのいくつかは、合成の主過程より先に又は遅れ

10

20

30

40

50

て進み得、この位相の外れた標識付けは、プレフェージング（プレフェージング）又はフェージング（phasing）として知られている。低率のプレフェージング及びポストフェージング（post-phasing）が経験的に観察された場合、プレフェージング及びポストフェージングからもたらされる信号内のノイズのほとんど全ては、現在、過去及び将来のサイクルで、すなわち、3つのサイクルだけでデジタル画像を処理するニューラルネットワークによって処理することができる。

【0097】

現在のサイクル内のデジタル画像チャネルの間では、画像をサイクル内で整列する慎重なレジストレーションが、正確な塩基分類に強く寄与する。誤りの発生源の中でも、波長及び不一致照明源の組み合わせは、測定されたクラスター中心位置に小さい補正可能な差を生じさせる。画像タイル全体にわたってクラスター中心を正確に位置合わせするために、並進、回転、及びスケーリングを伴う、一般的なアフィン変換を使用することができる。アフィン変換は、画像データをリフレームするために及びクラスター中心についてオフセットを解消するために使用することができる。

【0098】

画像データをリフレームすることは、典型的には、アフィン変換を適用することによって、画像データを補間することを意味する。リフレームすることにより、対象のクラスター中心をピクセルパッチの中心ピクセルの中央に置くことができる。又は、画像収集時のジッタ及び他の不一致を克服するために、画像をテンプレートと位置合わせすることができる。リフレームすることは、ピクセルパッチ内の全てのピクセルの強度値を調整することを含む。バイリニア及びバイキュービック補間並びに重み付け面積調整は、代替的な方策である。

【0099】

いくつかの実施態様では、クラスター中心座標を、追加の画像チャネルとしてニューラルネットワークに供給することができる。

【0100】

距離信号もまた、塩基分類に寄与することができる。いくつかのタイプの距離信号は、クラスター中心からの領域の分離を反映する。最も強い光信号は、クラスター中心と一致すると見なされる。クラスター周辺部に沿った光信号は、近くのクラスターからの漂遊信号を含むことがある。分類は、信号成分の寄与がクラスター中心からのその分離に従って減衰されると、より正確であることが観察されている。機能する距離信号としては、単一クラスター距離チャネル、マルチクラスター距離チャネル、及びマルチクラスター形状ベース距離チャネルが挙げられる。単一クラスター距離チャネルは、中心ピクセル内にクラスター中心を有するパッチに適用される。そのとき、パッチ内の全ての領域の距離は、中心ピクセル内のクラスター中心からの距離である。中心ピクセルと同じクラスターに属していないピクセルには、計算した距離を与える代わりに、背景としてフラグ付けすることができる。マルチクラスター距離チャネルは、各領域から最も近いクラスター中心までの距離を事前計算する。これは、領域を誤ったクラスター中心に接続する可能性を有するが、その可能性は低いものである。マルチクラスター形状ベース距離チャネルは、領域（サブピクセル又はピクセル）を、同じ塩基分類を生成するピクセル中心に隣接領域を介して関連付ける。これにより、いくらかの計算コストで、誤ったピクセルまでの距離を測定する可能性が回避される。距離信号に対するマルチクラスター及びマルチクラスター形状ベースのアプローチは、事前計算、及び画像内の複数のクラスターとの併用の対象となるという利点を有する。

【0101】

形状情報は、ノイズから信号を分離し、信号対ノイズ比を改善するために、ニューラルネットワークによって使用することができる。上記の考察では、領域分類への及び距離チャネル情報を供給することへのいくつかのアプローチが識別された。いずれのアプローチにおいても、領域は、クラスターエッジを画定するために、背景として（クラスターの一部でないものとして）マーキングすることができる。ニューラルネットワークは、結果と

10

20

30

40

50

して得られる不規則なクラスター形状に関する情報を利用するように訓練することができる。距離情報及び背景分類は、組み合わせても、別個に使用してもよい。当接するクラスターからの信号を分離することは、クラスター密度が増加すると、ますます重要になる。

【0102】

並列処理の規模を増大させるための1つの方向性は、撮像媒体上のクラスター密度を増加させることである。密度を増加させることは、隣接するクラスターを有するクラスターを読み取るときに背景ノイズを増加させるという欠点を有する。任意のパッチ（例えば、 3×3 ピクセル）の代わりに、形状データを使用すると、例えば、クラスター密度が増加するのに応じて信号分離を維持するのに役立つ。

【0103】

開示される技術の一態様を適用すると、塩基分類スコアもまた、品質を予測するために活用することができる。開示される技術は、分類スコアを従来の S a n g e r 又は P h r e d 品質 Q スコアと直接又は予測モデルを介して関連させることを含む。Q 2 0、Q 3 0 又は Q 4 0 などのスコアは、 $Q = -10 \log_{10} P$ によって、塩基分類誤り確率に対数的に関係付けられる。Q スコアとのクラススコアの相関は、多出力ニューラルネットワーク又は多変量回帰分析を使用して実施することができる。品質スコアのリアルタイム計算の利点は、塩基分類中に、欠陥のある配列決定実行を早期に終了させることができることである。出願人は、分析シーケンスが全体の8分の1〜4分の1まで進んだときに、実行を終了させる時折の（まれな）決定が行われ得ることを見出した。終了させる決定は、50 サイクル後又は25〜75 サイクル後に行うことができる。別の方法で300〜1000 サイクルを実行する連続プロセスでは、早期の終了は、実質的なリソースの節約をもたらす。

【0104】

特殊化した畳み込みニューラルネットワーク（CNN）アーキテクチャを使用して、複数のサイクルにわたって塩基を分類することができる。1つの特殊化は、初期層の処理中でのデジタル画像チャンネル間の分離を含む。畳み込みフィルタスタックは、処理をサイクル間で分離するように構造化することができ、異なるサイクルからのデジタル画像セットの間のクロストークを防止する。処理をサイクル間で分離する動機は、異なるサイクルにおいて撮影された画像が残差レジストレーション誤差を有し、したがって位置がずれていて、互いに対してランダムな並進オフセットを有することである。これは、センサーの運動段階の動きの有限の精度に起因し、また、異なる周波数チャンネルで撮影された画像は、異なる光路及び波長を有するためでもある。

【0105】

継続的なサイクルからの画像セットを使用する動機は、特定のサイクルにおける信号へのプレフェージング及びポストフェージングの寄与が、2次の寄与であることである。したがって、畳み込みニューラルネットワークが下位層のデジタル画像セットの畳み込みを画像収集サイクル間で構造的に分離することは有用であり得る。

【0106】

畳み込みニューラルネットワーク構造はまた、クラスター化に関する情報の取り扱いにおいて特殊化することもできる。クラスター中心及び／又は形状のためのテンプレートは、畳み込みニューラルネットワークがデジタル画像データと組み合わせる、追加情報を提供する。クラスター中心の分類及び距離データは、サイクル間にわたって繰り返し適用することができる。

【0107】

畳み込みニューラルネットワークは、画像フィールド内の複数のクラスターを分類するように構造化することができる。複数のクラスターが分類される時、ピクセル又はサブピクセルのための距離チャンネルは、ピクセル又はサブピクセルが属する、最も近いクラスター中心又は隣接するクラスター中心のいずれかに対して相対的な距離情報をよりコンパクトに含むことができる。あるいは、各ピクセル若しくはサブピクセルのために、又は少なくとも、クラスター中心を含む各ピクセルのために、大きな距離ベクトルを供給するこ

10

20

30

40

50

とができ、これにより、所与のピクセルのためのコンテキストである、クラスター中心から全ての他のピクセルまでの完全な距離情報を与える。

【0108】

ベースコールとのテンプレート生成のいくつかの組み合わせは、面積重み付けの変動を使用して距離チャンネルの代わりとなることができる。ここからは、距離チャンネルの代わりに、テンプレート生成器の出力が、どのように直接使用され得るかについて説明する。

【0109】

ピクセル値修正へのテンプレート画像の直接適用に影響を及ぼす3つの検討事項、すなわち、画像セットがピクセル又はサブピクセルドメイン内で処理されるかどうか、いずれのドメインにおいても、面積重みがどのように計算されるか、及びサブピクセルドメインにおいて、補間された強度値を修正するためにテンプレート画像をマスクとして適用すること、について説明する。

【0110】

ピクセルドメイン内で塩基分類を実行することは、アップサンプリングから生じる計算量の増加（16倍など）を必要としない利点を有する。ピクセルドメインでは、畳み込みの最上位層であっても、不要な計算を解消するための論理を追加する代わりに、収集されない計算の実行を正当化するのに十分なクラスター密度を有し得る。まず、ピクセルドメインにおいて距離チャンネルなしでテンプレート画像データを直接使用する実施例から始める。

【0111】

いくつかの実施態様では、分類は、特定のクラスターに焦点を合わせる。これらの例では、クラスターの周辺部上のピクセルは、どの隣接するクラスターが分類の焦点であるかに応じて、異なる修正された強度値を有し得る。サブピクセルドメイン内のテンプレート画像は、重複ピクセルが2つの異なるクラスターに強度値を寄与することを示すことができる。本明細書において、2つ以上の隣接する又は当接するクラスターが両方とも同じ光学ピクセルに重なっているとき、この光学ピクセルを「重複ピクセル」と称し、両方のクラスターが、この光学ピクセルからの強度読み取り値に寄与する。流域解析（雨流を稜線において異なる流域に分けることに因んで名付けられた）は、当接するクラスターでさえも分離するために適用することができる。データがクラスターごとの分類のために受信されるとき、テンプレート画像は、クラスターの周辺部に沿った重複ピクセルのための強度データを修正するために使用することができる。重複ピクセルは、どのクラスターが分類の焦点であるかに応じて、異なる修正された強度を有することができる。

【0112】

ピクセルの修正された強度は、ホームクラスター（すなわち、そのピクセルが属するクラスター又はそのピクセルによって強度発光が主に表現されるクラスター）への重複ピクセル内のサブピクセル寄与に基づいて、アウェイクラスター（すなわち、そのピクセルによって強度発光が表現される非ホームクラスター）とは対照的に、低減させることができる。5つのサブピクセルがホームクラスターの一部であり、2つのサブピクセルがアウェイクラスターの一部であると仮定する。そのとき、7つのサブピクセルは、ホーム又はアウェイクラスターの強度に寄与する。ホームクラスターに焦点がある間、一実施態様では、重複ピクセルは、16個のサブピクセルのうちの7つがホーム又はアウェイクラスターの強度に寄与するため、 $7/16$ で強度を低減させられる。別の実施態様では、強度は、サブピクセルの総数によって除したホームクラスターに寄与するサブピクセルの面積に基づいて、 $5/16$ で低減させられる。第3の実施態様では、強度は、寄与するサブピクセルの総面積によって除したホームクラスターに寄与するサブピクセルの面積に基づいて、 $5/7$ で低減させられる。後者の2つの計算は、焦点がアウェイクラスターに移ると変化し、分子に「2」を有する分数をもたらす。

【0113】

当然ながら、距離チャンネルがクラスター形状のサブピクセルマップと共に考慮されている場合、強度の更なる低減を適用することができる。

【0114】

分類の焦点であるクラスターのためのピクセル強度がテンプレート画像を使用して修正されると、修正されたピクセル値は、ニューラルネットワークベースの分類器の層によって畳み込みされて、修正された画像を生成する。修正された画像は、継続的な配列決定サイクルにおいて塩基を分類するために使用される。

【0115】

あるいは、ピクセルドメイン内の分類は、画像のチャンク内の全てのピクセル又は全てのクラスターについて並列に進行することができる。このシナリオでは、中間計算の再利用性を保証するために、ピクセル値の1つの修正のみを適用することができる。上記の分数のいずれも、強度のより小さい又はより大きい減衰が望まれるかどうかに応じて、ピクセル強度を変更するために使用することができる。

10

【0116】

画像チャンクのためのピクセル強度がテンプレート画像を使用して修正されると、ピクセル及び周囲コンテキストは、ニューラルネットワークベースの分類器の層によって畳み込みされて、修正された画像を生成することができる。画像チャンク上で畳み込みを実行することにより、共有されたコンテキストを有するピクセル間での中間計算の再利用が可能になる。修正された画像は、継続的な配列決定サイクルにおいて塩基を分類するために使用される。

【0117】

この説明は、サブピクセルドメイン内での面積重みの適用のために並列化することができる。この並列とは、個々のサブピクセルについて重みを計算することができるということである。重みは、光学ピクセルの異なるサブピクセル部分について同じになり得るが、必ずしもそうである必要はない。重複ピクセルのサブピクセルがそれぞれ5つ及び2つである、ホーム及びアウェイクラスターの上記のシナリオを繰り返すと、ホームクラスターに属するサブピクセルへの強度の割り当ては、ピクセル強度の $7/16$ 、 $5/16$ 又は $5/7$ とすることができる。この場合もやはり、距離チャンネルがクラスター形状のサブピクセルマップと共に考慮されている場合、強度の更なる低減を適用することができる。

20

【0118】

画像チャンクのためのピクセル強度がテンプレート画像を使用して修正されると、サブピクセル及び周囲コンテキストは、ニューラルネットワークベースの分類器の層によって畳み込みされて、修正された画像を生成することができる。画像チャンク上で畳み込みを実行することにより、共有されたコンテキストを有するサブピクセル間での中間計算の再利用が可能になる。修正された画像は、継続的な配列決定サイクルにおいて塩基を分類するために使用される。

30

【0119】

別の代替例は、サブピクセルドメイン内に補間された画像データにテンプレート画像をサブピクセルドメイン内で2値マスクとして適用することである。テンプレート画像は、クラスター間の背景ピクセルを必要とするか、又は異なるクラスターからのサブピクセルが当接することを可能にするかのいずれかになるように配置することができる。テンプレート画像は、マスクとして適用することができる。このマスクは、補間されたピクセルがテンプレート画像内で背景として分類されると、その補間されたピクセルが、補間によって割り当てられた値を保持するか、又は背景値（例えば、ゼロ）を受け入れるかどうかを決定する。

40

【0120】

この場合もやはり、画像チャンクのためのピクセル強度がテンプレート画像を使用してマスクされると、サブピクセル及び周囲コンテキストは、ニューラルネットワークベースの分類器の層によって畳み込みされて、修正された画像を生成することができる。画像チャンク上で畳み込みを実行することにより、共有されたコンテキストを有するサブピクセル間での中間計算の再利用が可能になる。修正された画像は、継続的な配列決定サイクルにおいて塩基を分類するために使用される。

50

【0121】

開示される技術の特徴は、共有されたコンテキスト内で任意の数のクラスターを分類し、中間計算を再利用するために、組み合わせることができる。光学 픽セル解像度において、一実施態様では、約10パーセントの 픽セルが、分類されるクラスター中心を保持する。レガシーシステムでは、不規則な形状のクラスターが観察されると、 3×3 個の光学 픽セルが、クラスター中心のための潜在的な信号寄与因子として分析のためにグループ化された。最上位の畳み込み層から離れた1つの 3×3 フィルタであっても、クラスター密度は、実質的に半分よりも多くの光学 픽セルからの光信号をクラスター中心の 픽セル内にロールアップする傾向にある。スーパーサンプリングされた解像度においてのみ、最上位の畳み込み層のためのクラスター中心密度は、1パーセント未満に低下する。

10

【0122】

共有されたコンテキストは、いくつかの実施態様ではかなり大きい。例えば、 15×15 光学 픽セルコンテキストが、正確な塩基分類に寄与し得る。同等の4倍アップサンプリングされたコンテキストは、 60×60 個のサブ 픽セルになるであろう。この広大なコンテキストは、ニューラルネットワークが撮像中に不均一な照明及び背景の影響を認識するのに役立つ。

【0123】

開示される技術は、テンプレート入力内のクラスター境界をデジタル画像入力内で検出された境界と組み合わせるために、下位の畳み込み層において小さなフィルタを使用する。クラスター境界は、ニューラルネットワークが背景条件から信号を分離し、画像処理を背景に対して正規化するのを助ける。

20

【0124】

開示された技術は、中間計算を実質的に再利用する。 $20 \sim 25$ 個のクラスター中心が 15×15 個の光学 픽セルのコンテキスト領域内に現れると仮定する。そのとき、第1の層の畳み込みが、ブロック単位の畳み込みロールアップにおいて $20 \sim 25$ 回再利用される。再利用率は、最後から2番目の層に至るまで層ごとに低減されていき、最後から2番目の層において、光学解像度における再利用率が初めて1倍未満に低下する。

【0125】

複数の畳み込み層からのブロック単位のロールアップ訓練及び推論は、픽セル又はサブ 픽セルのブロックに継続的なロールアップを適用する。ブロック周辺部の周囲には、第1のデータブロックのロールアップ中に使用されるデータがロールアップの第2のブロックと重なっており、ロールアップの第2のブロックのために再利用することができる、重複ゾーンが存在する。ブロック内には、重複ゾーンによって囲まれた中心領域において、ロールアップすることができ、かつ再利用することができる、픽セル値及び中間計算が存在する。重複ゾーンを用いると、コンテキストフィールドのサイズを（例えば 3×3 フィルタの適用によって 15×15 から 13×13 に）漸進的に低減させる畳み込み結果は、畳み込みされた値を保持する同じメモリブロック内に書き込むことができ、ブロック内での基礎計算の再利用を損なうことなくメモリを節約する。より大きなブロックでは、重複ゾーン内の中間計算を共有することは、より少ないリソースを必要とする。より小さいブロックでは、複数のブロックを並列に計算して、重複ゾーン内の中間計算を共有することが可能であり得る。

30

40

【0126】

より大きなフィルタ及び拡張は、下位の畳み込み層がテンプレート及び／又はデジタル画像データ内のクラスター境界に反応した後、畳み込み層の数を低減し、これは、分類を損なうことなく速度計算であり得る。

【0127】

テンプレートデータのための入力チャネルは、デジタル画像フィールド内の複数のクラスター中心を分類することと一貫性のあるテンプレート構造体を作製するように選択することができる。上記の2つの代替例は、この一貫性基準、すなわち、コンテキスト全体にわたってリフレイムし、距離マッピングすることを満たさない。リフレイムすることは、

50

1つのクラスターのみを中心を光学ピクセルの中心に配置する。複数のクラスターを分類するためにより良好であるのは、クラスター中心を保持していると分類されるピクセルのために中心オフセットを供給することである。

【0128】

距離マッピングは、提供される場合、全てのピクセルがそれぞれ独自の距離マップをコンテキスト全体にわたって有しない限り、コンテキスト領域全体にわたって実行することは困難である。より単純な距離マップは、デジタル画像入力ブロックから複数のクラスターを分類するための有用な一貫性を提供する。

【0129】

ニューラルネットワークは、クラスターの境界におけるピクセル又はサブピクセルのテンプレート内の分類から学習することができ、そのため、距離チャンネルは、クラスター中心オフセットチャンネルを伴った、2値又は3値分類を供給するテンプレートに置き換えることができる。距離マップは、使用される場合、ピクセル（又はサブピクセル）が属するクラスター中心からのピクセルの距離を与えることができる。又は距離マップは、最も近いクラスター中心までの距離を与えることができる。距離マップは、背景ピクセルに割り当てられたフラグ値で2値分類を符号化することができ、又はピクセル分類からの別個のチャンネルとすることができ。クラスター中心オフセットと組み合わせられると、距離マップは、3値分類を符号化することができ。いくつかの実施態様、特に1又は2ビットでピクセル分類を符号化する実施態様では、少なくとも開発中に、ピクセル分類のために及び距離のために別個のチャンネルを使用することが望ましい場合がある。

【0130】

開示される技術は、上位層内でいくらかの計算リソースを節約するために計算の低減を含むことができる。クラスター中心オフセットチャンネル又は3値分類マップは、ピクセル中心の最終分類に寄与しないピクセル畳み込みの中心を識別するために使用することができる。多くのハードウェア／ソフトウェア実装では、推論中にルックアップを実行し、畳み込みロールアップをスキップすることは、更に9つの乗算及び8つの加算を実行して3×3フィルタを適用するよりも、上位層において効率を向上させることができる。並列実行のために計算をパイプライン処理するカスタムハードウェアでは、全てのピクセルをパイプライン内でそれぞれ分類することができる。そのとき、クラスター中心マップは、クラスター中心と一致するピクセルのみについて結果を収集するために、それらのピクセルに対して最終分類のみが所望されることから、最終の畳み込み後に使用することができる。この場合もやはり、光学ピクセルドメイン内では、現在観察されているクラスター密度において、約10パーセントのピクセルのためのロールアップされた計算が収集される。4倍アップサンプリングされたドメイン内では、最上位層内でのサブピクセル分類のうちの1パーセント未満が収集されるため、より多くの層が、一部のハードウェア上において、スキップされた畳み込みから利益を得ることができる。

【0131】

ニューラルネットワークベースのベースコール

【0132】

図1は、一実施態様による、ベースコールするためにRTAベースコーラーによって使用される処理段階を示す。図1はまた、2つの実施態様による、ベースコールするために、開示されるニューラルネットワークベースのベースコーラーによって使用される処理段階も示す。図1に示すように、ニューラルネットワークベースのベースコーラー218は、RTAベースコーラーによって使用される処理段階の多くを不要にすることによって、ベースコールプロセスを合理化することができる。合理化することにより、ベースコール精度及びスケールが改善する。ニューラルネットワークベースのベースコーラー218の第1の実施態様では、ベースコーラーは、ニューラルネットワークベースのテンプレート生成器1512の出力から識別されたクラスター中心の場所／位置情報を使用してベースコールを実行する。第2の実施態様では、ニューラルネットワークベースのベースコーラー218は、ベースコールするためにクラスター中心の場所／位置情報を使用しない。第

2の実施態様は、パターン化されたフローセル設計がクラスター生成に使用されるときに使用される。パターン化されたフローセルは、既知の基準位置に対して正確に位置付けられ、かつパターン化されたフローセル上に事前配置されたクラスター分布を提供する、ナノウェルを含む。他の実施態様では、ニューラルネットワークベースのベースコーラー218は、ランダムなフローセル上に生成されたクラスターをベースコールする。

【0133】

ここからは、ニューラルネットワークが配列決定画像をマッピングしてベースコールするように訓練されているニューラルネットワークベースのベースコールについて説明する。議論は、以下のように整理される。最初に、ニューラルネットワークへの入力について説明する。次に、ニューラルネットワークの構造及び形態について説明する。最後に、ニューラルネットワークの出力について説明する。

10

【0134】

入力

【0135】

図2は、ニューラルネットワーク206を使用してベースコールする一実施態様を示す。

【0136】

主入力：画像チャンネル

【0137】

ニューラルネットワーク206への主入力は、画像データ202である。画像データ202は、配列決定実行中に配列決定器222によって生成された配列決定画像108から導出される。一実施態様では、画像データ202は、配列決定画像222から抽出された $n \times n$ 個の画像パッチを含み、 n は1~10,000の範囲の任意の数である。配列決定実行は、対応する m 個の画像チャンネルのために配列決定サイクルごとに m 個の画像を生成し、画像パッチは、特定の配列決定サイクルのための画像データを調製するために m 個の画像のそれぞれから抽出される。4、2、及び1チャンネルケミストリーなどの異なる実施態様では、 m は4又は2である。他の実施態様では、 m は、1、3、又は4超である。画像データ202は、いくつかの実施態様では光学的なピクセルドメイン内にあり、他の実施態様ではアップサンプリングされたサブピクセルドメイン内にある。

20

【0138】

画像データ202は、複数の配列決定サイクル（例えば、現在の配列決定サイクル、1つ又はそれ以上の先行する配列決定サイクル、及び1つ又はそれ以上の継続的な配列決定サイクル）のためのデータを含む。一実施態様では、画像データ202は、ベースコールされる現在（時間 t ）の配列決定サイクルのためのデータが、（i）左側フランキング／コンテキスト／前の／先行する／先の（時間 $t-1$ ）の配列決定サイクルのためのデータ及び（ii）右側フランキング／コンテキスト／次の／継続的な／後続の（時間 $t+1$ ）の配列決定サイクルのためのデータを伴うように、3個の配列決定サイクルのためのデータを含む。他の実施態様では、画像データ202は、単一の配列決定サイクルのためのデータを含む。

30

【0139】

画像データ202は、1つ又はそれ以上のクラスター及びそれらの周囲背景の強度発光を表現する。一実施態様では、単一のターゲットクラスターがベースコールされる場合、画像パッチは、本明細書において「ターゲットクラスター中心化パッチ抽出」と称される概念である、各画像パッチがその中心ピクセル内にターゲットクラスターの中心を含むような方法で、配列決定画像108から抽出される。

40

【0140】

画像データ202は、入力データ204内で強度チャンネル（画像チャンネルとも呼ばれる）を使用して符号化される。特定の配列決定サイクルのために配列決定器222から取得された m 個の画像のそれぞれについて、その強度データを符号化するために別個の画像チャンネルが使用される。例えば、配列決定実行が、各配列決定サイクルにおいて赤色画像及び緑色画像を生成する2チャンネルケミストリーを使用し、そのとき、入力データ204は

50

、（i）赤色画像内に捕捉された1つ又はそれ以上のクラスター及びそれらの周囲背景の強度発光を表現する $n \times n$ 個のピクセルを有する第1の赤色画像チャンネルと、（ii）緑色画像内に捕捉された1つ又はそれ以上のクラスター及びそれらの周囲背景の強度発光を表現する $n \times n$ 個のピクセルを有する第2の緑色画像チャンネルと、を含むと、考える。

【0141】

一実施態様では、バイオセンサーは、光センサーのアレイを備える。光センサーは、バイオセンサーの検出表面上の対応するピクセル領域（例えば、反応部位／ウェル／ナノセル）からの情報を感知するように構成されている。ピクセル領域内に配設された分析物は、ピクセル領域、すなわち、関連する分析物と関連付けられると言われる。配列決定サイクルでは、ピクセル領域に対応する光センサーは、関連する検体からの発光／光子を検出／捕捉／検知するように構成され、それに応じて、画像化されたチャンネルごとにピクセル信号を生成するように構成される。一実施態様では、各撮像チャンネルは、複数のフィルタ波長帯域のうちの1つに対応する。別の実施態様では、各撮像チャンネルは、配列決定サイクルで複数の撮像イベントのうちの1つに対応する。更に別の実施態様では、各撮像チャンネルは、特定のレーザーを用いた照明と特定の光学フィルタを通した撮像との組み合わせに対応する。

【0142】

光センサーからのピクセル信号は、（例えば、通信ポートを介して）バイオセンサーに結合された信号プロセッサに伝達される。各配列決定サイクル及び各画像化チャンネルについて、信号プロセッサは、ピクセルが対応する光センサーから得られるピクセル信号をそれぞれ描写／含有／示す／表す／特徴付ける画像を生成する。このようにして、画像内のピクセルは、（i）ピクセルによって表されるピクセル信号を生成したバイオセンサーの光センサーと、（ii）対応する光センサーによって放射が検出され、ピクセル信号に変換された関連分析物と、（iii）関連分析物を保持するバイオセンサーの検出表面上のピクセル領域と、に対応する。

【0143】

例えば、配列決定動作が2つの異なる画像化チャンネル：赤色チャンネル及び緑色チャンネルを使用すると考える。次いで、各配列決定サイクルにおいて、信号プロセッサは、赤色画像及び緑色画像を生成する。このようにして、配列決定実行の一連の k 配列決定サイクルについて、赤色画像及び緑色画像の k 対を有するシーケンスが出力として生成される。

【0144】

赤色画像及び緑色画像（すなわち、異なる画像化チャンネル）内のピクセルは、配列決定サイクル内で1対1の対応を有する。これは、一対の赤色画像及び緑色画像内の対応するピクセルが、異なる画像化チャンネル内にある同じ関連する検体の強度データを示すことを意味する。同様に、赤色画像及び緑色画像の対にわたるピクセルは、配列決定サイクル間に1対1の対応を有する。これは、赤色画像及び緑色画像の異なるペア内の対応するピクセルが、配列決定実行の異なる獲得イベント／タイムステップ（配列決定サイクル）に関して、同じ関連する分析物の強度データを示すことを意味する。

【0145】

赤色画像及び緑色画像（すなわち、異なる画像化チャンネル）内の対応するピクセルは、第1の赤色チャンネル及び第2の緑チャンネル内の強度データを表す、「サイクルごとの画像」のピクセルと見なすことができる。ピクセルがピクセルエリアのサブセットのピクセル信号、すなわち、バイオセンサーの検出面の領域（タイル）を描写するサイクルごとの画像は、「サイクルごとのタイル画像」と呼ばれる。サイクルごとのタイル画像から抽出されたパッチは、「サイクルごとの画像パッチ」と呼ばれる。一実施態様では、パッチ抽出は、入力準備者によって実行される。

【0146】

画像データは、配列決定実行の一連の k シーケンスサイクルのために生成された一連のサイクルごとの画像パッチを含む。サイクルごとの画像パッチ内のピクセルは、関連する検体のための強度データを含み、強度データは、関連付けられた検体からの排出を検出す

るように構成された対応する光センサーによって、1つ又はそれ以上の画像化チャンネル（例えば、赤色チャンネル及び緑色チャンネル）のために取得される。一実施態様では、単一のターゲットクラスターをベースとする場合、サイクルごとの画像パッチは、標的関連検体及び非中心ピクセルに関する強度データを含む中心ピクセルで中心に置かれ、サイクルごとの画像パッチ内の非中心ピクセルは、標的関連検体に隣接する関連する検体の強度データを含む。一実施態様では、画像データは、入力準備者によって調製される。

【0147】

非画像データ

【0148】

別の実施態様では、ニューラルネットワークベースのベースコーラー218及びニューラルネットワークベースの品質スコアラー6102への入力データは、分子延長中の水素イオンの放出によって誘発されるpH変化に基づく。pH変化は検出され、組み込まれた塩基の数に比例する電圧変化に変換される（例えば、Ion Torrentの場合）。

【0149】

更に別の実施態様では、ニューラルネットワークベースのベースコーラー218及びニューラルネットワークベースの品質スコアラー6102への入力データは、塩基を同定しながら分析物がナノ細孔の中又はその開口部付近を通過する際の電流の中断を測定するためにバイオセンサーを使用するナノ細孔検知から構築される。例えば、Oxford Nanopore Technologies (ONT) 配列決定は、以下の概念に基づく：ナノ細孔を介して膜を介してDNA（又はRNA）の単一鎖を通過させ、膜にわたって電位差を印加する。細孔内に存在するヌクレオチドは、細孔の電気抵抗に影響を及ぼし、そのため、経時的な電流測定は、細孔を通過するDNA塩基の配列を示すことができる。この電流信号（プロットされたときにその外観に起因する「押しつぶし」）は、ONTシーケンサによって収集された生データである。これらの測定値は、4kHz周波数（例えば）で取られた16ビットの整数データ取得（DAC）値として記憶される。1秒当たり～450塩基対のDNA鎖速度を用いて、これは、平均して、塩基当たり約9つの生観察を与える。次いで、この信号を処理して、個々の読み取りに対応する開孔信号の破断を特定する。これらの生信号の伸長は、ベースと呼ばれ、DAC値をDNA塩基の配列に変換するプロセスである。いくつかの実施態様では、入力データは、正規化又はスケーリングされたDAC値を含む。

【0150】

補足的な入力：距離チャンネル

【0151】

画像データ202は、補足的な距離データ（距離チャンネルとも呼ばれる）を伴う。距離チャンネルは、画像チャンネルから生成された特徴マップに組み込まれる加法的バイアスを供給する。この加法的バイアスは、ピクセル中心－クラスター中心間の距離に基づいていることからベースコール精度に寄与し、ピクセル中心－クラスター中心間の距離は、距離チャンネル内にピクセル単位で符号化されている。

【0152】

「単一のターゲットクラスター」ベースコール実施態様では、入力データ204内の各画像チャンネル（画像パッチ）について、補足的な距離チャンネルは、その中心ピクセルを含み、かつベースコールされるターゲットクラスターの中心からのそのピクセルの中心の距離を識別する。それにより、距離チャンネルは、画像パッチの中心ピクセルからの画像パッチのピクセルのそれぞれの距離を示す。

【0153】

「マルチクラスター」ベースコール実施態様では、入力データ204内の各画像チャンネル（画像パッチ）について、補足的な距離チャンネルは、ピクセルとクラスターのそれぞれとの間の中心間距離に基づいて選択されるクラスターのうちの最も近いクラスターからの各ピクセルの中心間距離を識別する。

【0154】

「マルチクラスター形状ベース」ベースコール実施態様では、入力データ204内の各画像チャンネル（画像パッチ）について、補足的な距離チャンネルは、1つのクラスターのみに対して各クラスターピクセルを分類することに基づいて選択される割り当てられたクラスターからの各クラスターピクセルの中心間距離を識別する。

【0155】

補足的な入力：スケーリングチャンネル

【0156】

画像データ202は、異なるクラスターサイズ及び不均一な照明条件を考慮する補足的なスケーリングデータ（スケーリングチャンネルとも呼ばれる）を伴う。スケーリングチャンネルもまた、画像チャンネルから生成された特徴マップに組み込まれる加法的バイアスを供給する。この加法的バイアスは、中央クラスターピクセルの平均強度に基づくため、ベースコール精度に寄与し、この平均強度は、スケーリングチャンネル内にピクセル単位で符号化されている。

【0157】

補足的な入力：クラスター中心座標

【0158】

いくつかの実施態様では、ニューラルネットワークベースのテンプレート生成器1512の出力から識別されたクラスター中心の場所／位置情報216（例えば、 $x-y$ 座標）は、ニューラルネットワーク206に補足的な入力として供給される。

補足的な入力：クラスター帰属情報

【0159】

いくつかの実施態様では、ニューラルネットワーク206は、どのピクセル又はサブピクセルが、同じクラスターを表現する／に寄与する／に属する背景ピクセル又はサブピクセル、クラスター中心ピクセル又はサブピクセル、及びクラスター／クラスター内部ピクセル又はサブピクセルであるかを分類するクラスター帰属情報を補足的な入力として受信する。他の実施態様では、減衰マップ、2値マップ、及び／若しくは3値マップ、又はそれらの変形が、ニューラルネットワーク206への補足的な入力として供給される。

【0160】

前処理：強度修正

【0161】

いくつかの実施態様では、入力データ204は距離チャンネルを含んでいないが、代わりにニューラルネットワーク206が、ニューラルネットワークベースのテンプレート生成器1512 1512の出力、すなわち、減衰マップ、2値マップ、及び／又は3値マップに基づいて修正される修正された画像データを入力として受信する。このような実施態様では、画像データ202の強度は、距離チャンネルの不在を考慮するように修正される。

【0162】

他の実施態様では、画像データ202は、1つ又はそれ以上のロスレス変換動作（例えば、畳み込み、逆畳み込み、フーリエ変換）に供され、結果として得られる修正された画像データは、ニューラルネットワーク206への入力として供給される。

【0163】

ネットワーク構造及び形態

【0164】

ニューラルネットワーク206は、本明細書において「ニューラルネットワークベースのベースコーラー」218とも称される。一実施態様では、ニューラルネットワークベースのベースコーラー218は、多層パーセプトロン（MLP）である。別の実施態様では、ニューラルネットワークベースのベースコーラー218は、フィードフォワードニューラルネットワークである。更に別の実施態様では、ニューラルネットワークベースのベースコーラー218は、全結合型ニューラルネットワークである。更なる実施態様では、ニューラルネットワークベースのベースコーラー218は、完全畳み込みニューラルネットワークである。また更なる実施態様では、ニューラルネットワークベースのベースコーラ

10

20

30

40

50

ー 2 1 8 は、セマンティックセグメンテーションニューラルネットワークである。

【 0 1 6 5 】

一実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、複数の畳み込み層を有する畳み込みニューラルネットワーク (CNN) である。別の実施態様では、長・短期記憶ネットワーク (LSTM)、双方向 LSTM (Bi-LSTM)、又はゲート付き回帰型ユニット (GRU) などの反復ニューラルネットワーク (RNN) である。更に別の実施態様では、CNN 及び RNN の両方を含む。

【 0 1 6 6 】

更に他の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、1 D 重畳、2 D 重畳、3 D 重畳、4 D 重畳、5 D 重畳、拡張又は無性重畳、転置重畳、奥行分離可能な重畳、点単位の重畳、 1×1 重畳、グループ重畳、扁平重畳、空間及びクロスチャネルの重畳、シャッフルされたグループ化重畳、空間的な分離可能な重畳、並びに逆重畳を使用することができる。それは、ロジスティック回帰／対数損失、多クラスクロスエントロピー／ソフトマックス損失、二値クロスエントロピー損失、平均二乗誤差損失、L1 損失、L2 損失、平滑 L1 損失、及び Huber 損失などの 1 つ又はそれ以上の損失機能を使用することができる。それは、TFRecord、圧縮符号化 (例えば、PNG)、シャープ化、マップ変換に対する平行コール、バッチング、プリフェッチ、モデル並列、データ並列、及び同期／非同期 SGD のような、任意の並列、効率、及び圧縮方式を使用することができる。これは、アップサンプリング層、ダウンサンプリング層、反復接続、ゲート及びゲートされたメモリユニット (LSTM 又は GRU など)、残留ブロック、残留接続、ハイウェイ接続、スキップ接続、ペエホル接続、アクティブ化機能 (例えば、非線形変換関数は、整流線形ユニット (ReLU)、漏れやすい ReLU、指数関数的ライナーユニット (ELU)、シグモイド及び双曲線正接 (tanh))、バッチ正規化層、規則化層、ドロップアウト、プール層 (例えば、最大又は平均プール)、グローバル平均プール層、及び注意機構のような非線形変換機能を含む。

【 0 1 6 7 】

ニューラルネットワークベースのベースコーラー 2 1 8 は、入力データ 2 0 4 を処理し、入力データ 2 0 4 の代替表現 2 0 8 を生成する。代替表現 2 0 8 は、いくつかの実施態様では畳み込み表現であり、他の実施態様では隠れ表現である。次いで、代替表現 2 0 8 は、出力層 2 1 0 によって処理され、出力 2 1 2 を生成する。出力 2 1 2 は、以下に説明するように、ベースコールを生成するために使用される。

【 0 1 6 8 】

出力

【 0 1 6 9 】

一実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、特定の配列決定サイクルのために、単一のターゲットクラスターのためのベースコールを出力する。別の実施態様では、特定の配列決定サイクルのために、複数のターゲットクラスター内の各ターゲットクラスターのためのベースコールを出力する。更に別の実施態様では、複数の配列決定サイクル内の各配列決定サイクルのために、複数のターゲットクラスター内の各ターゲットクラスターのためのベースコールを出力し、それによって、各ターゲットクラスターのためのベースコール配列を生成する。

【 0 1 7 0 】

距離チャンネル計算

【 0 1 7 1 】

ここからは、距離チャンネルの距離値を計算する際に使用するクラスター中心の適切な場所／位置情報 (例えば、 $x-y$ 座標) がどのように取得されるかについて説明する。

【 0 1 7 2 】

座標のダウンスケーリング

【 0 1 7 3 】

図 3 は、ニューラルネットワークベースのテンプレート生成器 1 5 1 2 の出力から識別

10

20

30

40

50

されたクラスター中心の場所／位置情報をサブピクセルドメインからピクセルドメインに変換する一実施態様である。

【0174】

クラスター中心の場所／位置情報は、ニューラルネットワークベースのベースコールのために、少なくとも、(i) ベースコールされるターゲットクラスターの中心を中心ピクセル内に含む配列決定画像108からの画像パッチを抽出することによって入力データを構築するために、(ii) 画像パッチの中心ピクセルに含まれるターゲットクラスターの中心からの画像パッチのピクセルの中心の距離を識別する距離チャンネルを構築するために、及び／又は(iii) ニューラルネットワークベースのベースコーラー218への補足的な入力216として、使用される。

10

【0175】

いくつかの実施態様では、クラスター中心の場所／位置情報は、アップサンプリングされたサブピクセル解像度でニューラルネットワークベースのテンプレート生成器1512の出力から識別される。しかしながら、いくつかの実施態様では、ニューラルネットワークベースのベースコーラー218は、光学ピクセル解像度の画像データ上で動作する。したがって、一実施態様では、クラスター中心の場所／位置情報は、ニューラルネットワークベースのテンプレート生成器1512への入力として供給される画像データをアップサンプリングするために使用されるのと同じアップサンプリング係数でクラスター中心の座標をダウンスケーリングすることによってピクセルドメインに変換される。

【0176】

20

例えば、ニューラルネットワークベースのテンプレート生成器1512への入力として供給される画像パッチデータが、いくつかの初期配列決定サイクルからの配列決定画像108をアップサンプリング係数 f でアップサンプリングすることによって導出されと考える。そのとき、一実施態様では、ポストプロセッサ1814によるニューラルネットワークベースのテンプレート生成器1512によって生成され、テンプレート／テンプレート画像304に記憶された、クラスター中心302の座標は、 f (除数) で除される。これらのダウンスケーリングされたクラスター中心座標は、本明細書において「参照クラスター中心」308と称され、テンプレート／テンプレート画像304内に記憶される。一実施態様では、ダウンスケーリングは、ダウンスケーラー306によって実施される。

【0177】

30

座標の変換

【0178】

図4は、サイクル固有及び画像チャンネル固有の変換を使用して参照クラスター中心308からいわゆる「変換されたクラスター中心」404を導き出す一実施態様である。このようにする動機について、最初に議論する。

【0179】

異なる配列決定サイクルで撮影された配列決定画像は、位置がずれていて、互いに対してランダムな並進オフセットを有する。これは、センサーの運動段階の動きの有限の精度に起因し、また、異なる画像／周波数チャンネルで撮影された画像は、異なる光路及び波長を有するためでもある。その結果、参照クラスター中心と、配列決定画像内のクラスター中心の場所／位置との間にオフセットが存在する。このオフセットは、異なる配列決定サイクルにおいて捕捉された画像間、及び異なる画像チャンネル内で同じ配列決定サイクルにおいて捕捉された画像内で変化する。

40

【0180】

このオフセットを考慮するために、サイクル固有及び画像チャンネル固有の変換が参照クラスター中心に適用されて、各配列決定サイクルの画像パッチのためのそれぞれの変換されたクラスター中心を生成する。サイクル固有及び画像チャンネル固有の変換は、画像相関を使用して、完全な6パラメータのアフィン変換 (例えば、並進、回転、スケーリング、剪断、右反射、左反射) 又はプロクラステス変換 (例えば、並進、回転、スケーリング、任意選択的にアスペクト比に拡張される) を決定する画像レジストレーションプロセスに

50

よって導出され、その更なる詳細は、付録 1、2、3、及び 4に見出すことができる。

【0181】

例えば、4つのクラスター中心のための参照クラスター中心が (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; (x_4, y_4) であり、配列決定実行が、各配列決定サイクルにおいて赤色画像及び緑色画像を生成する2チャンネルケミストリーを使用する、と考える。そのとき、例示的な配列決定サイクル3では、サイクル固有及び画像チャンネル固有の変換は、赤色画像に対して

【数1】

$$\{\alpha_r^3, \beta_r^3, \chi_r^3, \delta_r^3, \varepsilon_r^3, \phi_r^3\}$$

10

であり、緑色画像に対して

【数2】

$$\{\alpha_g^3, \beta_g^3, \chi_g^3, \delta_g^3, \varepsilon_g^3, \phi_g^3\}$$

20

である。

【0182】

同様に、例示的な配列決定サイクル9では、サイクル固有及び画像チャンネル固有の変換は、赤色画像に対して

【数3】

$$\{\alpha_r^9, \beta_r^9, \chi_r^9, \delta_r^9, \varepsilon_r^9, \phi_r^9\}$$

30

であり、緑色画像に対して

【数4】

$$\{\alpha_g^9, \beta_g^9, \chi_g^9, \delta_g^9, \varepsilon_g^9, \phi_g^9\}$$

である。

【0183】

そして、配列決定サイクル3の赤色画像のための変換されたクラスター中心

40

【数5】

$$(\hat{x}_1, \hat{y}_1); (\hat{x}_2, \hat{y}_2); (\hat{x}_3, \hat{y}_3); (\hat{x}_4, \hat{y}_4)$$

は、参照クラスター中心 (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; (x_4, y_4) に変換

【数6】

50

$$\{\alpha_r^3, \beta_r^3, \chi_r^3, \delta_r^3, \varepsilon_r^3, \phi_r^3\}$$

を適用することによって導き出され、配列決定サイクル 3 の緑色画像のための変換されたクラスター中心

【数 7】

$$(\tilde{x}_1, \tilde{y}_1); (\tilde{x}_2, \tilde{y}_2); (\tilde{x}_3, \tilde{y}_3); (\tilde{x}_4, \tilde{y}_4)$$

10

は、参照クラスター中心 (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; (x_4, y_4)) に変換

【数 8】

$$\{\alpha_g^3, \beta_g^3, \chi_g^3, \delta_g^3, \varepsilon_g^3, \phi_g^3\}$$

20

を適用することによって導き出される。

【0184】

同様に、配列決定サイクル 9 の赤色画像のための変換されたクラスター中心

【数 9】

$$(\vec{x}_1, \vec{y}_1); (\vec{x}_2, \vec{y}_2); (\vec{x}_3, \vec{y}_3); (\vec{x}_4, \vec{y}_4)$$

30

は、参照クラスター中心 (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; (x_4, y_4)) に変換

【数 10】

$$\{\alpha_r^9, \beta_r^9, \chi_r^9, \delta_r^9, \varepsilon_r^9, \phi_r^9\}$$

を適用することによって導き出され、配列決定サイクル 9 の緑色画像のための変換されたクラスター中心

【数 11】

$$(\tilde{x}_1, \tilde{y}_1); (\tilde{x}_2, \tilde{y}_2); (\tilde{x}_3, \tilde{y}_3); (\tilde{x}_4, \tilde{y}_4)$$

40

は、参照クラスター中心 (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; (x_4, y_4)) に変換

【数 12】

50

$$\{\alpha_g^g, \beta_g^g, \chi_g^g, \delta_g^g, \varepsilon_g^g, \phi_g^g\}$$

を適用することによって導き出される。

【0185】

一実施態様では、変換は、変換器402によって実施される。

【0186】

変換されたクラスター中心404は、テンプレート／テンプレート画像304内に記憶され、(i)対応する配列決定画像108からのパッチ抽出を（例えば、パッチ抽出器406によって）行うために、(ii)距離式

【数13】

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

で対応する画像パッチのための距離チャンネルを計算するために、及び(iii)ベースコールされている対応する配列決定サイクルのためのニューラルネットワークベースのベースコーラー218への補足的な入力として、それぞれ使用される。他の実施態様では、距離の2乗、 $e^{-\text{距離}}$ 、及び $e^{\text{距離}}$ （距離の2乗）などの異なる距離式を使用することができる。

【0187】

画像パッチ

【0188】

図5は、ニューラルネットワークベースのベースコーラー218に供給される入力データの一部である画像パッチ502を示す。入力データは、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。

【0189】

例えば、配列決定実行が、各配列決定サイクルにおいて赤色画像及び緑色画像を生成する2チャンネルケミストリーを使用し、入力データが、配列決定実行の一連の3つの配列決定サイクル、すなわち、ベースコールされる現在（時間 t ）の配列決定サイクル、前（時間 $t-1$ ）の配列決定サイクル、及び次（時間 $t+1$ ）の配列決定サイクルにわたるデータを含む、と考える。

【0190】

そのとき、入力データは、以下のサイクルごとの画像パッチセットのシーケンスを含む：現在の赤色画像パッチ及び現在の緑色画像パッチが現在の配列決定サイクルにおいて捕捉された赤色及び緑色の配列決定画像からそれぞれ抽出されている現在のサイクルの画像パッチセット、前の赤色画像パッチ及び前の緑色画像パッチが前の配列決定サイクルにおいて捕捉された赤色及び緑色の配列決定画像からそれぞれ抽出されている前のサイクルの画像パッチセット、並びに次の赤色画像パッチ及び次の緑色画像パッチが次の配列決定サイクルにおいて捕捉された赤色及び緑色の配列決定画像からそれぞれ抽出されている次のサイクルの画像パッチセット。

【0191】

各画像パッチのサイズは $n \times n$ とすることができ、 n は1～10,000の範囲の任意の数とすることができ。各画像パッチは、光学ピクセルドメイン内、又はアップサンプリングされたサブピクセルドメイン内にあることができる。図5に示す実施態様では、抽

出された画像ページ 502 は、複数のクラスター 1～m 及びそれらの周囲背景を覆う／表現するピクセルについてのピクセル強度データを有する。また、図示された実施態様では、画像パッチ 502 は、その中心ピクセル内に、ベースコールされているターゲットクラスターの中心を含むような方法で、抽出される。

【0192】

図 5 では、ピクセル中心は黒い矩形によって表現され、整数の場所／位置座標を有しており、クラスター中心は紫色の円によって表現され、浮動小数点数の場所／位置座標を有する。

【0193】

単一のターゲットクラスターのための距離計算

10

【0194】

図 6 は、単一のターゲットクラスターがニューラルネットワークベースのベースコーラー 218 によってベースコールされているときに距離チャンネルのための距離値 602 を決定する一実施態様を示す。ターゲットクラスターの中心は、ニューラルネットワークベースのベースコーラー 218 への入力として供給される画像パッチの中心ピクセル内に含まれている。距離値は、各ピクセルについて、その中心とターゲットクラスターの中心との間の距離が決定されるように、ピクセルごとに計算される。したがって、距離値は、入力データの一部である画像パッチのそれぞれ内の各ピクセルについて計算される。

【0195】

図 6 は、特定の画像パッチのための 3 つの距離値 d_1 、 d_c 、及び d_n を示す。一実施態様では、距離値 602 は、距離式

20

【数 14】

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

を使用して計算され、これは、変換されたクラスター中心 404 上で動作する。他の実施態様では、距離の 2 乗、 $e^{\wedge}$ 距離、及び $e^{\wedge}$ （距離の 2 乗）などの異なる距離式を使用することができる。

30

【0196】

他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、距離値 602 は、サブピクセルドメイン内で計算される。

【0197】

したがって、単一のターゲットクラスターベースコール実施態様では、距離チャンネルは、ベースコールされているターゲットクラスターに対してのみ計算される。

【0198】

図 7 は、ピクセルとターゲットクラスターとの間で計算される距離値 602 をピクセル単位で符号化すること 702 の一実施態様を示す。一実施態様では、入力データ内において、距離値 602 は距離チャンネルの一部として、対応する各画像チャンネル（画像パッチ）を「ピクセル距離データ」として補足する。赤色画像及び緑色画像が配列決定サイクルごとに生成される実施例に戻ると、入力データは、赤色画像チャンネル及び緑色画像チャンネルをピクセル距離データとしてそれぞれ補足する赤色距離チャンネル及び緑色距離チャンネルを含む。

40

【0199】

他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、距離チャンネルは、サブピクセルごとに符号化される。

【0200】

複数のターゲットクラスターのための距離計算

【0201】

50

図 8 a は、複数のターゲットクラスター 1 ~ m がニューラルネットワークベースのベースコーラー 2 1 8 によって同時にベースコールされているときに距離チャンネルのための距離値 8 0 2 を決定する一実施態様を示す。距離値は、各ピクセルについて、その中心と複数のクラスター 1 ~ m のそれぞれの各中心との間の距離が決定され、最小距離値（赤色）がピクセルに割り当てられるように、ピクセルごとに計算される。

【 0 2 0 2 】

したがって、距離チャンネルは、ピクセルとクラスターのそれぞれとの間の中心間距離に基づいて選択されるクラスターのうちの最も近いクラスターからの各ピクセルの中心間距離を識別する。図示された実施態様では、図 8 a は、2 つのピクセル及び 4 つのクラスター中心に対するピクセル中心 - クラスター中心間の距離を示す。ピクセル 1 はクラスター 1 に最も近く、ピクセル n はクラスター 3 に最も近い。

10

【 0 2 0 3 】

一実施態様では、距離値 8 0 2 は、距離式

【 数 1 5 】

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

を使用して計算され、これは、変換されたクラスター中心 4 0 4 上で動作する。他の実施態様では、距離の 2 乗、 $e^{\text{距離}}$ 、及び $e^{-\text{距離}}$ （距離の 2 乗）などの異なる距離式を使用することができる。

20

【 0 2 0 4 】

他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、距離値 8 0 2 は、サブピクセルドメイン内で計算される。

【 0 2 0 5 】

したがって、マルチクラスターベースコール実施態様では、距離チャンネルは、複数のクラスターの中から最も近いクラスターに対して計算される。

【 0 2 0 6 】

図 8 b は、ターゲットクラスター 1 ~ m のそれぞれについて、ピクセル中心 - 最も近いクラスター中心間の距離 8 0 4（ d_1 、 d_2 、 d_{23} 、 d_{29} 、 d_{24} 、 d_{32} 、 d_n 、 d_{13} 、 d_{14} など）に基づいて決定されたいくつかの最も近いピクセルを示す。

30

【 0 2 0 7 】

図 9 は、ピクセルとクラスターのうちの最も近いクラスターとの間で計算される最小距離値をピクセル単位で符号化すること 9 0 2 の一実施態様を示す。他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、距離チャンネルは、サブピクセルごとに符号化される。

【 0 2 0 8 】

クラスター形状に基づく複数のターゲットクラスターのための距離計算

【 0 2 0 9 】

図 1 0 は、複数のターゲットクラスター 1 ~ m がニューラルネットワークベースのベースコーラー 2 1 8 によって同時にベースコールされているときに、本明細書において「クラスター形状データ」又は「クラスター形状情報」と称される、ピクセル - クラスター間の分類／帰属／カテゴリー化 1 0 0 2 を使用して、距離チャンネルのためのクラスター距離値 1 1 0 2 を決定する一実施態様を示す。最初に、以下において、クラスター形状データがどのように生成されるかについて簡単に検討する。

40

【 0 2 1 0 】

上述したように、ニューラルネットワークベースのテンプレート生成器 1 5 1 2 の出力は、同じクラスターを表現する／に寄与する／に属する背景ピクセル、中心ピクセル、及びクラスター／クラスター内部ピクセルとしてピクセルを分類するために使用される。こ

50

のピクセルークラスター間の分類情報は、ピクセル中心とクラスター中心との間の距離に関係なく、各ピクセルを1つのクラスターのみに帰属させるために使用され、クラスター形状データとして記憶される。

【0211】

図10に示す実施態様では、背景ピクセルは灰色に着色され、クラスター1に属するピクセルは黄色に着色され（クラスター1ピクセル）、クラスター2に属するピクセルは緑色に着色され（クラスター2ピクセル）、クラスター3に属するピクセルは赤色に着色され（クラスター3ピクセル）、クラスターmに属するピクセルは青色に着色されている（クラスターmピクセル）。

【0212】

図11は、クラスター形状データを使用して距離値1102を計算する一実施態様を示す。最初に、クラスター形状を考慮しないで計算された距離情報が誤りを起こしやすい理由について説明する。その後、クラスター形状データがこの制限をどのように克服するかについて説明する。

【0213】

クラスター形状データを使用しない「マルチクラスター」ベースコール実施態様では（図8a～図8b及び図9）、ピクセルのための中心間距離値は、複数のクラスターの中から最も近いクラスターに対して計算される。ここで、クラスターAに属するピクセルが、クラスターAの中心から遠く離れていて、クラスターBの中心にはより近くにある場合のシナリオを考える。そのような場合、クラスター形状データなしでは、ピクセルは、（ピクセルが実際に属している）クラスターAに対する距離値が割り当てられる代わりに、（ピクセルが属していない）クラスターBに対して計算される距離値を割り当てられる。

【0214】

「マルチクラスター形状ベース」ベースコール実施態様は、生画像データにおいて定義され、ニューラルネットワークベースのテンプレート生成器1512によって生成されるような、真のピクセルークラスター間のマッピングを使用することによってこれを回避する。

【0215】

2つの実施態様間の対比は、ピクセル34及び35に関して見ることができる。図8bでは、ピクセル34及び35の距離値は、クラスター形状データを考慮しないで、クラスター3の最も近い中心に対して計算される。しかしながら、図11では、クラスター形状データに基づいて、ピクセル34及び35の距離値1102は、（それらが実際に属する）クラスター2に対して計算される。

【0216】

図11において、クラスターピクセルはクラスター強度を表現し、背景ピクセルは背景強度を表現する。クラスター距離値は、クラスターのうちの1つのみに対して各クラスターピクセルを分類することに基づいて選択されるクラスターのうちの割り当てられたクラスターからの各クラスターピクセルの中心間距離を識別する。いくつかの実施態様では、背景ピクセルは、0若しくは0.1、又は何らかの他の最小値など、所定の背景距離値を割り当てられる。

【0217】

一実施態様では、上述したように、クラスター距離値1102は、距離式

【数16】

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

を使用して計算され、これは、変換されたクラスター中心404上で動作する。他の実施態様では、距離の2乗、 $e^{\wedge}$ 距離、及び $e^{\wedge}$ （距離の2乗）などの異なる距離式を使用

することができる。

【0218】

他の実施態様では、画像パッチがアップサンプリングされたサブピクセル解像度であるとき、クラスター距離値1102はサブピクセルドメイン内で計算され、クラスター及び背景帰属1002はサブピクセルごとに生じる。

【0219】

したがって、マルチクラスター形状ベーススコール実施態様では、距離チャンネルは、複数のクラスターの中から割り当てられたクラスターに対して計算される。割り当てられたクラスターは、生画像データにおいて定義された真のピクセルークラスター間のマッピングに従ってクラスターのうちの1つのみに対して各クラスターピクセルを分類することに基づいて選択される。

10

【0220】

図12は、ピクセルと割り当てられたクラスターとの間で計算される距離値1002をピクセル単位で符号化する一実施態様を示す。他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、距離チャンネルは、サブピクセルごとに符号化される。

【0221】

深層学習は、多層ニューラルネットワークを使用する強力な機械学習技術である。コンピュータビジョン及び画像処理ドメインにおける1つの特に成功したネットワーク構造は、畳み込みニューラルネットワーク(CNN)であり、各層が、入力テンソル(画像様の多次元高密度アレイ)から異なる形状の出力テンソルへのフィードフォワード畳み込み変換を実施する。CNNは、画像の空間コヒーレンス、及び最大3又は4Dのアレイ上で訓練を高速にする汎用のグラフィック処理ユニット(GPU)の出現により、画像様の入力に特に適している。これらの画像様の特性を活用することは、サポートベクターマシン(SVM)又は多層パーセプトロン(MLP)などの他の学習方法と比較して、優れた経験的性能をもたらす。

20

【0222】

画像データと補足的な距離及びスケールデータとの両方を処理するために、標準的なCNNを増強する特殊化したアーキテクチャを導入する。以下の詳細がある。

【0223】

特殊化したアーキテクチャ

【0224】

図13は、異なる配列決定サイクルのためのデータの処理を分離するために使用されるニューラルネットワークベースのベースコーラー218の特殊化したアーキテクチャの一実施態様を示す。特殊化したアーキテクチャを使用する動機について、最初に説明する。

【0225】

上述したように、ニューラルネットワークベースのベースコーラー218は、現在の配列決定サイクル、1つ又はそれ以上の先行する配列決定サイクル、及び1つ又はそれ以上の継続的な配列決定サイクルのためのデータを処理する。追加の配列決定サイクルのためのデータは、配列特異的なコンテキストを提供する。ニューラルネットワークベースのベースコーラー218は、訓練時に配列特異的なコンテキストを学習し、それらをベースコールする。更に、前及び後の配列決定サイクルのためのデータは、現在の配列決定サイクルにプレフェージング信号及びフェージング信号の2次の寄与を提供する。

40

【0226】

空間的な畳み込み層

【0227】

しかしながら、上述したように、異なる配列決定サイクルにおいて及び異なる画像チャンネル内で捕捉された画像は、位置がずれていて、互いに対して残差レジストレーション誤差を有する。この位置ずれを考慮するために、特殊化したアーキテクチャは、配列決定サイクル間で情報を混合せず、配列決定サイクル内でのみ情報を混合する、空間的な畳み込

50

み層を含む。

【0228】

空間的な畳み込み層は、畳み込みの「専用の共有されていない」シーケンスを介して複数の配列決定サイクルのそれぞれのためのデータを独立して処理することによって分離を操作可能にする、いわゆる「分離された畳み込み」を使用する。分離された畳み込みは、所与の配列決定サイクルのみ、すなわち、イントラサイクルのデータ及び結果として得られる特徴マップにわたって畳み込みを行い、任意の他の配列決定サイクルのデータ及び結果として得られる特徴マップにわたって畳み込みを行うことはない。

【0229】

例えば、入力データが、(i) ベースコールされる現在(時間 t)の配列決定サイクルのための現在のデータ、(ii) 前(時間 $t-1$)の配列決定サイクルのための前のデータ、及び(iii) 次(時間 $t+1$)の配列決定サイクルのための次のデータを含む、と考える。そのとき、特殊化したアーキテクチャは、3つの別個のデータ処理パイプライン(又は畳み込みパイプライン)、すなわち、現在のデータ処理パイプライン、前のデータ処理パイプライン、及び次のデータ処理パイプラインを開始させる。現在のデータ処理パイプラインは、現在(時間 t)の配列決定サイクルのための現在のデータを入力として受信し、それを複数の空間的な畳み込み層によって独立して処理して、いわゆる「現在の空間畳み込み表現」を最終の空間的な畳み込み層の出力として生成する。前のデータ処理パイプラインは、前(時間 $t-1$)の配列決定サイクルのための前のデータを入力として受信し、それを複数の空間的な畳み込み層によって独立して処理して、いわゆる「前の空間畳み込み表現」を最終の空間的な畳み込み層の出力として生成する。次のデータ処理パイプラインは、次(時間 $t+1$)の配列決定サイクルのための次のデータを入力として受信し、それを複数の空間的な畳み込み層によって独立して処理して、いわゆる「次の空間畳み込み表現」を最終の空間的な畳み込み層の出力として生成する。

【0230】

いくつかの実施態様では、現在、前、及び次の処理パイプラインは並列に実行される。

【0231】

いくつかの実施態様では、空間的な畳み込み層は、特殊化したアーキテクチャ内の空間的な畳み込みネットワーク(又はサブネットワーク)の一部である。

【0232】

時間的な畳み込み層

【0233】

ニューラルネットワークベースのベースコーラー218は、配列決定サイクル間、すなわち、インターサイクルの情報を混合する時間的な畳み込み層を更に含む。時間的な畳み込み層は、空間的な畳み込みネットワークからそれらの入力を受信し、それぞれのデータ処理パイプラインのための最終の空間的な畳み込み層によって生成された空間畳み込み表現上で動作する。

【0234】

時間的な畳み込み層のインターサイクル動作の自由は、位置ずれ特性(空間的な畳み込みネットワークへの入力として供給される画像データ内に存在する)が、空間的な畳み込み層のシーケンスによって実行される分離された畳み込みのカスケードによって空間畳み込み表現から一掃されるという事実から生じる。

【0235】

時間的な畳み込み層は、スライディングウィンドウに基づいた継続的な入力内の入力チャネルにわたってグループ単位の畳み込みを行う、いわゆる「組み合わせ畳み込み」を使用する。一実施態様では、継続的な入力は、前の空間的な畳み込み層又は前の時間的な畳み込み層によって生成される継続的な出力である。

【0236】

いくつかの実施態様では、時間的な畳み込み層は、特殊化したアーキテクチャ内の時間的な畳み込みネットワーク(又はサブネットワーク)の一部である。時間的な畳み込みネ

10

20

30

40

50

ットワークは、その入力を空間的な畳み込みネットワークから受信する。一実施態様では、時間的な畳み込みネットワークの最初の時間的な畳み込み層が、配列決定サイクル間の空間畳み込み表現をグループ単位で組み合わせる。別の実施態様では、時間的な畳み込みネットワークの後続の時間的な畳み込み層が、前の時間的な畳み込み層の継続的な出力を組み合わせる。

【0237】

最終の時間的な畳み込み層の出力は、出力を生成する出力層に供給される。出力は、1つ又はそれ以上の配列決定サイクルにおける1つ又はそれ以上のクラスターをベースコールするために使用される。

【0238】

以下では、分離された畳み込み及び組み合わせ畳み込みについてより詳細に説明する。

【0239】

分離された畳み込み

【0240】

前方伝搬の間、特殊化したアーキテクチャは、複数の入力からの情報を2つの段階で処理する。第1の段階では、入力間の情報の混合を防止するために、分離畳み込みが使用される。第2の段階では、入力間で情報を混合するために、組み合わせ畳み込みが使用される。第2の段階からの結果は、複数の入力に対する単一の推論を行うために使用される。

【0241】

これは、畳み込み層がバッチ内の複数の入力を同時に処理し、バッチ内の各入力に対して対応する推論を行う、バッチモード技術とは異なる。対照的に、特殊化したアーキテクチャは、複数の入力を単一の推論にマッピングする。単一の推論は、4つの塩基（A、C、T、及びG）のそれぞれについての分類スコアなど、2つ以上の予測を含むことができる。

【0242】

一実施態様では、入力、各入力異なる時間ステップにおいて生成され、複数の入力チャンネルを有するように、時間的順序付けを有する。例えば、複数の入力は、以下の3つの入力を含むことができる：時間ステップ（ t ）における現在の配列決定サイクルによって生成された現在の入力、時間ステップ（ $t-1$ ）における前の配列決定サイクルによって生成された前の入力、及び時間ステップ（ $t+1$ ）における次の配列決定サイクルによって生成された次の入力。別の実施態様では、各入力は、それぞれ、1つ又はそれ以上の前の畳み込み層によって現在の、前の、及び次の入力から導出され、 k 個の特徴マップを含む。

【0243】

一実施態様では、各入力は、以下の5つの入力チャンネルを含むことができる：赤色画像チャンネル（赤色）、赤色距離チャンネル（黄色）、緑色画像チャンネル（緑色）、緑色距離チャンネル（紫色）、及びスケールチャンネル（青色）。別の実施態様では、各入力は、前の畳み込み層によって生成された k 個の特徴マップを含むことができ、各特徴マップは入力チャンネルとして処理される。

【0244】

図14は、分離された畳み込みの一実施態様を示す。分離された畳み込みは、各入力に畳み込みフィルタを並列に適用することによって、複数の入力を一度に処理する。分離された畳み込みでは、畳み込みフィルタは、同じ入力内の入力チャンネルを組み合わせ、異なる入力内の入力チャンネルは組み合わせない。一実施態様では、各入力に同じ畳み込みフィルタが並列に適用される。別の実施態様では、各入力に異なる畳み込みフィルタが並列に適用される。いくつかの実施態様では、各空間的な畳み込み層は、 k 個の畳み込みフィルタのバンクを含み、これらのそれぞれが各入力に並列に適用される。

【0245】

組み合わせ畳み込み

【0246】

10

20

30

40

50

組み合わせ畳み込みは、異なる入力の対応する入力チャネルをグループ化し、各グループに畳み込みフィルタを適用することによって、異なる入力の間で情報を混合する。対応する入力チャネルのグループ化及び畳み込みフィルタの適用は、スライディングウィンドウに基づいて行われる。このコンテキストにおいて、ウィンドウは、例えば、2つの継続的な配列決定サイクルのための出力を表す、2つ以上の継続的な入力チャネルにわたる。このウィンドウはスライディングウィンドウなので、ほとんどの入力チャネルが2つ以上のウィンドウ内で使用される。

【0247】

いくつかの実施態様では、異なる入力は、先行する空間的又は時間的な畳み込み層によって生成される出力シーケンスに由来する。出力シーケンスにおいて、異なる入力は、継続的な出力として配列されており、したがって、次の時間的な畳み込み層によって継続的な入力であると見なされる。次いで、次の時間的な畳み込み層において、組み合わせ畳み込みは、継続的な入力内の対応する入力チャネルのグループに畳み込みフィルタを適用する。

【0248】

一実施態様では、継続的な入力は、現在の入力時間が時間ステップ (t) における現在の配列決定サイクルによって生成され、前の入力が、時間ステップ ($t-1$) における前の配列決定サイクルによって生成され、次の入力が、時間ステップ ($t+1$) における次の配列決定サイクルによって生成されるような、時間的順序付けを有する。別の実施態様では、各継続的な入力は、それぞれ、1つ又はそれ以上の前の畳み込み層によって現在の、前の、及び次の入力から導出され、 k 個の特徴マップを含む。

【0249】

一実施態様では、各入力は、以下の5つの入力チャネルを含むことができる：赤色画像チャネル（赤色）、赤色距離チャネル（黄色）、緑色画像チャネル（緑色）、緑色距離チャネル（紫色）、及びスケールチャネル（青色）。別の実施態様では、各入力は、前の畳み込み層によって生成された k 個の特徴マップを含むことができ、各特徴マップは入力チャネルとして処理される。

【0250】

畳み込みフィルタの深さ B は、スライディングウィンドウに基づいて対応する入力チャネルが畳み込みフィルタによってグループ単位で畳み込みされる、継続的な入力の数に依存する。換言すれば、深さ B は、各スライディングウィンドウ内の継続的な入力の数及びグループサイズに等しい。

【0251】

図15aでは、各スライディングウィンドウ内で2つの継続的な入力からの対応する入力チャネルが組み合わせられており、したがって、 $B=2$ である。図15bでは、各スライディングウィンドウ内で3つの継続的な入力からの対応する入力チャネルが組み合わせられており、したがって、 $B=3$ である。

【0252】

一実施態様では、スライディングウィンドウは、同じ畳み込みフィルタを共有する。別の実施態様では、各スライディングウィンドウに異なる畳み込みフィルタが使用される。いくつかの実施態様では、時間的な畳み込み層は、 k 個の畳み込みフィルタのバンクを含み、これらのそれぞれがスライディングウィンドウに基づいて継続的な入力に適用される。

【0253】

フィルタバンク

【0254】

図16は、各畳み込み層が畳み込みフィルタのバンクを有する、ニューラルネットワークベースのベースコーラー218の畳み込み層の一実施態様を示す。図16では、5つの畳み込み層が示されており、これらのそれぞれは、64個の畳み込みフィルタのバンクを有する。いくつかの実施態様では、各空間的な畳み込み層は、 k 個の畳み込みフィルタのバンクを有し、 k は、1、2、8、64、128、256などの任意の数とすることがで

10

20

30

40

50

きる。いくつかの実施態様では、各時間的な量み込み層は、 k 個の量み込みフィルタのバンクを有し、 k は、1、2、8、64、128、256などの任意の数とすることができる。

【0255】

ここからは、補足的なスケーリングチャンネル、及びそれがどのように計算されるかについて説明する。

【0256】

スケーリングチャンネル

【0257】

図17は、画像チャンネルを補足するスケーリングチャンネルの2つの構成を示す。スケーリングチャンネルは、ニューラルネットワークベースのベースコーラー218に供給される入力データ内にピクセル単位で符号化されている。異なるクラスターサイズ及び不均一な照明条件は、広範囲のクラスター強度が抽出されるという結果をもたらす。スケーリングチャンネルによって供給される加法的バイアスは、クラスター間にわたってクラスター強度を同等にする。他の実施態様では、画像パッチが、アップサンプリングされたサブピクセル解像度であるとき、スケーリングチャンネルは、サブピクセルごとに符号化される。

【0258】

単一のターゲットクラスターがベースコールされているとき、スケーリングチャンネルは、全てのピクセルに同じスケーリング値を割り当てる。複数のターゲットクラスターが同時にベースコールされているとき、スケーリングチャンネルは、クラスター形状データに基づいてピクセルのグループに異なるスケーリング値を割り当てる。

【0259】

スケーリングチャンネル1710は、全てのピクセルに対して同じスケーリング値 (s_1) を有する。スケーリング値 (s_1) は、ターゲットクラスターの中心を含む中心ピクセルの平均強度に基づく。一実施態様では、平均強度は、ターゲットクラスターに対するA及びTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって計算される。

【0260】

スケーリングチャンネル1708は、クラスター形状データに基づいた対応するクラスターに帰属するそれぞれのピクセルグループに対して異なるスケーリング値 (s_1 、 s_2 、 s_3 、 s_m) を有する。各ピクセルグループは、対応するクラスターの中心を含む中央クラスターピクセルを含む。特定のピクセルグループのためのスケーリング値は、その中央クラスターピクセルの平均強度に基づく。一実施態様では、平均強度は、対応するクラスターに対するA及びTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察する中央クラスターピクセルの強度値を平均化することによって計算される。

【0261】

いくつかの実施態様では、背景ピクセルは、0若しくは0.1、又は何らかの他の最小値であり得る、背景スケーリング値 (s_b) を割り当てられる。

【0262】

一実施態様では、スケーリングチャンネル1706及びそれらのスケーリング値は、強度スケーラー1704によって決定される。強度スケーラー1704は、先行する配列決定サイクルからのクラスター強度データ1702を使用して、平均強度を計算する。

【0263】

他の実施態様では、補足的なスケーリングチャンネルは、ニューラルネットワークベースのベースコーラー218の最後の層より前に又はその層に、ニューラルネットワークベースのベースコーラー218の1つ又はそれ以上の中間層より前に又はそれらの層に、及び画像サイズに一致するようにピクセル単位で符号化する代わりに単一の値としてなど、異なる方法で入力として提供することができる。

【0264】

ここからは、ニューラルネットワークベースのベースコーラー218に供給される入力

10

20

30

40

50

データについて説明する。

【0265】

入力データ：画像チャンネル、距離チャンネル、及びスケーリングチャンネル

【0266】

図18aは、赤色画像及び緑色画像を生成する単一の配列決定サイクルのための入力データ1800の一実施態様を示す。入力データ1800は、以下を含む。

- ・赤色画像から抽出された画像パッチ内のピクセルに関する赤色強度データ1802（赤色）。赤色強度データ1802は、赤色画像チャンネル内に符号化される。

- ・ピクセル単位で赤色強度データ1802を補足する赤色距離データ1804（黄色）。赤色距離データ1804は、赤色距離チャンネル内に符号化される。

- ・緑色画像から抽出された画像パッチ内のピクセルに関する緑色強度データ1806（緑色）。緑色強度データ1806は、緑色画像チャンネル内に符号化される。

- ・ピクセル単位で緑色強度データ1806を補足する緑色距離データ1808（紫色）。緑色距離データ1808は、緑色距離チャンネル内に符号化される。

- ・ピクセル単位で赤色強度データ1802及び緑色強度データ1806を補足するスケーリングデータ1810（青色）。スケーリングデータ1810は、スケーリングチャンネル内に符号化される。

【0267】

他の実施態様では、入力データは、より少ない又はより多い数の画像チャンネル及び補足的な距離チャンネルを含むことができる。一実施例では、4チャンネルケミストリーを使用する配列決定実行のために、入力データは、各配列決定サイクルのための4つの画像チャンネル、及び4つの補足的な距離チャンネルを含む。

【0268】

ここからは、距離チャンネル及びスケーリングチャンネルがベースコール精度にどのように寄与するかについて説明する。

【0269】

加法的バイアス

【0270】

図18bは、画像チャンネルから生成された特徴マップに組み込まれる加法的バイアスを供給する距離チャンネルの一実施態様を示す。この加法的バイアスは、ピクセル中心-クラスター中心間の距離に基づいていることからベースコール精度に寄与し、ピクセル中心-クラスター中心間の距離は、距離チャンネル内にピクセル単位で符号化されている。

【0271】

平均して、約 3×3 のピクセルが、1つのクラスターを含む。クラスターの中心における密度は、クラスターが実質的に中央の場所から外向きに成長することから、外辺部におけるよりも高いと予想される。周辺クラスターピクセルは、近くのクラスターからの競合する信号を含むことができる。したがって、中央クラスターピクセルは、最大強度領域と見なされ、クラスターを確実に識別するピーコンとして機能する。

【0272】

画像パッチのピクセルは、複数のクラスター（例えば、10～200個のクラスター）及びそれらの周囲背景の強度発光を表現する。追加のクラスターは、より広い半径からの情報を組み込み、画像パッチ内に強度発光が表現されている、基礎をなす塩基を見分けることによってベースコール予測に寄与する。換言すれば、クラスターのグループからの強度発光は、個別の塩基（A、C、T、又はG）に割り当てることができる強度パターンを累積的に作り出す。

【0273】

補足的な距離チャンネルでクラスター中心からの各ピクセルの距離を畳み込みフィルタに明示的に伝えることは、より高いベースコール精度をもたらすことが観察されている。距離チャンネルは、どのピクセルがクラスター中心を含んでいるか、及びどのピクセルがクラスター中心からより遠く離れているかを畳み込みフィルタに伝える。畳み込みフィルタは

10

20

30

40

50

、（b）周辺クラスターピクセル、背景ピクセル、及びそれらから導出される特徴マップよりも、（a）中央クラスターピクセル、それらの隣接ピクセル、及びそれらから導出される特徴マップに着目することによって配列決定信号をその適切なソースクラスターに割り当てるために、この情報を使用する。着目することの一実施例では、距離チャンネルは、（a）から生じた特徴マップに組み込まれている正の加法的バイアスを供給し、（b）から生じた特徴マップに組み込まれている負の加法的バイアスを供給する。

【0274】

距離チャンネルは、画像チャンネルと同じ次元数を有する。これにより、畳み込みフィルタは、局所受容フィールド内の画像チャンネル及び距離チャンネルを別個に評価し、評価をコヒーレントに組み合わせることができる。

【0275】

単一のターゲットクラスターがベースコールされているとき、距離チャンネルは、画像パッチの中心にある1つの中央クラスターピクセルのみを識別する。複数のターゲットクラスターが同時にベースコールされているとき、距離チャンネルは、画像パッチ間にわたって分布した複数の中央クラスターピクセルを識別する。

【0276】

「単一クラスター」距離チャンネルは、ベースコールされる単一のターゲットクラスターの中心を中心ピクセル内に含む画像パッチに適用される。単一クラスター距離チャンネルは、単一のターゲットクラスターまでの画像パッチ内の各ピクセルの中心間距離を含む。この実施態様では、画像パッチはまた、単一のターゲットクラスターに隣接する追加のクラスターも含むが、追加のクラスターはベースコールされない。

【0277】

「マルチクラスター」距離チャンネルは、ベースコールされる複数のターゲットクラスターの中心をそれぞれの中央クラスターピクセル内に含む画像パッチに適用される。マルチクラスター距離チャンネルは、複数のターゲットクラスターの中から最も近いクラスターまでの画像パッチ内の各ピクセルの中心間距離を含む。これは、誤ったクラスターまでの中心間距離を測定する可能性を有するが、その可能性は低いものである。

【0278】

「マルチクラスター形状ベース」距離チャンネルは、ベースコールされる複数のターゲットクラスターの中心をそれぞれの中央クラスターピクセル内に含み、かつピクセルークラスター間の帰属情報が既知である、画像パッチに適用される。マルチクラスター距離チャンネルは、画像パッチ内の各クラスターピクセルの中心から、複数のターゲットクラスターのうちでそのクラスターピクセルに属する又はそれに帰属するクラスターの中心までの距離を含む。背景ピクセルには、計算された距離を与える代わりに、背景としてフラグ付けすることができる。

【0279】

図18bはまた、画像チャンネルから生成された特徴マップに組み込まれる加法的バイアスを供給するスケーリングチャンネルの一実施態様も示す。この加法的バイアスは、中央クラスターピクセルの平均強度に基づくため、ベースコール精度に寄与し、この平均強度は、スケーリングチャンネル内にピクセル単位で符号化されている。距離チャンネルのコンテキストにおける加法的バイアスについての議論は、スケーリングチャンネルに類似的に適用される。

【0280】

加法的バイアスの実施例

【0281】

図18bは、加法的バイアスがどのようにして距離及びスケーリングチャンネルから導出され、画像チャンネルから生成された特徴マップに組み込まれるかの一実施例を更に示す。

【0282】

図18bでは、畳み込みフィルタ1814は、2つの画像チャンネル1802及び1806、2つの距離チャンネル1804及び1808、並びにスケーリングチャンネル1810

10

20

30

40

50

にわたって局所受容フィールド1812（マゼンタ）を評価する。距離及びスケーリングチャンネルは別個に符号化されるため、加法的バイアスは、チャンネル固有の畳み込みカーネル（又は特徴検出器）1816a～eのそれぞれの中間出力1816a～e（及びバイアス1816f）が局所受容フィールド1812のための最終出力／特徴マップ要素1820としてチャンネル単位で蓄積される（1818）ときに行われる。この実施例では、2つの距離チャンネル1804及び1808によって供給される加法的バイアスは、それぞれ中間出力1816b及び1816dである。スケーリングチャンネル1810によって供給される加法的バイアスは、中間出力1816eである。

【0283】

加法的バイアスは、ベースコール、すなわち、中央クラスターピクセル及びそれらの隣接ピクセルのピクセル強度についてより重要で信頼性が高いと考えられる画像チャンネル内のそれらの特徴をより強調することによって、特徴マップコンパイレーションプロセスを進める。訓練中、グラウンドトゥルースペースコールとの比較から計算された勾配の逆伝播は、畳み込みカーネルの重みを更新して、中央クラスターピクセル及びそれらの隣接ピクセルに対してより強い活性化をもたらす。

【0284】

例えば、局所受容フィールド1812によって覆われた隣接ピクセルのグループ内のピクセルがクラスター中心を含み、そのとき、距離チャンネル1804及び1808は、クラスター中心へのピクセルの近接度を反映する、と考える。結果として、強度中間出力1816a及び1816cがチャンネル単位の蓄積1818において距離チャンネル加法的バイアス1816b及び1816dとマージされると、ピクセルの正にバイアスされた畳み込み表現1820がもたらされる。

【0285】

対照的に、局所受容フィールド1812によって覆われたピクセルがクラスター中心の近くにないと、距離チャンネル1804及び1808は、クラスター中心からのそれらの分離を反映する。結果として、強度中間出力1816a及び1816cがチャンネル単位の蓄積1818において距離チャンネル加法的バイアス1816b及び1816dとマージされると、ピクセルの負にバイアスされた畳み込み表現1820がもたらされる。

【0286】

同様に、スケーリングチャンネル1810から導出されるスケーリングチャンネル加法的バイアス1816eは、ピクセルの畳み込み表現1820を正に又は負にバイアスすることができる。

【0287】

明確にするために、図18bは、単一の配列決定サイクルのための入力データ1800上での単一の畳み込みフィルタ1814の適用を示す。当業者であれば、この説明が、複数の畳み込みフィルタ（例えば、k個のフィルタのフィルタバンクであり、kは8、16、32、64、128、256などとすることができる）に、複数の畳み込み層（例えば、複数の空間的及び時間的な畳み込み層）、及び複数の配列決定サイクル（例えば、t、t+1、t-1）に拡張され得ることは理解するであろう。

【0288】

他の実施態様では、距離及びスケーリングチャンネルは、距離及びスケーリングチャンネル並びに画像チャンネルが同じ次元数を有するので、別々に符号化される代わりに、画像チャンネルに直接適用されて、変調されたピクセル増幅）を生成する。更なる実施態様では、畳み込みカーネルの重みは、要素単位の増幅中に画像チャンネル内の最も重要な特徴を検出するように、距離及び画像チャンネルに基づいて決定される。更に他の実施態様では、距離及びスケーリングチャンネルは、第1の層に供給される代わりに、下流の層及び／又はネットワーク（例えば、全結合ネットワーク又は分類層）に補助入力として提供される。また更なる実施態様では、距離及びスケーリングチャンネルは、第1の層に供給され、下流の層及び／又はネットワークに（例えば、残差接続を介して）再供給される。

【0289】

上記の説明は、 k 個の入力チャネルを有する 2 D 入力データについてのものである。3 D 入力への拡張は、当業者によって理解されるであろう。簡潔に述べると、体積入力は、次元 $k \times l \times w \times h$ を有する 4 D テンソルであり、 l は追加の次元、すなわち、長さである。各個々のカーネルは、4 D テンソル内で掃引された 4 D テンソルであり、3 D テンソル（チャネル次元は、横断方向に掃引されないため、潰れている）。

【0290】

他の実施態様では、入力データ 1800 がアップサンプリングされたサブピクセル解像度にあるとき、距離及びスケールチャネルはサブピクセルごとに別個に符号化され、加法的バイアスはサブピクセルレベルで行われる。

【0291】

特殊化したアーキテクチャ及び入力データを使用したベースコール

【0292】

ここからは、特殊化したアーキテクチャ及び入力データが、ニューラルネットワークベースのベースコールのためにどのように使用されるかについて説明する。

【0293】

単一のクラスターのベースコール

【0294】

図 19 a、図 19 b、及び図 19 c は、単一のターゲットクラスターをベースコールする一実施態様を示す。特殊化したアーキテクチャは、3 つの配列決定サイクル、すなわち、ベースコールされる現在（時間 t ）の配列決定サイクル、前（時間 $t-1$ ）の配列決定サイクル、及び次（時間 $t+1$ ）の配列決定サイクルを処理し、現在（時間 t ）の配列決定サイクルにおける単一のターゲットクラスターのためのベースコールを生成する。

【0295】

図 19 a 及び図 19 b は、空間的な畳み込み層を示す。図 19 c は、時間的な畳み込み層を、いくつかの他の非畳み込み層と共に示す。図 19 a 及び図 19 b では、垂直な点線が、空間的な畳み込み層を特徴マップから分離し、水平な破線が、3 つの配列決定サイクルに対応する 3 つの畳み込みパイプラインを分離している。

【0296】

各配列決定サイクルのために、入力データは、次元数 $n \times n \times m$ （例えば、図 18 a における入力テンソル 1800）のテンソルを含み（ n は正方形テンソルの幅及び高さを表し、 m は入力チャネルの数を表す）、3 つのサイクルのための次元数 $n \times n \times m \times t$ の入力データを作成する。

【0297】

ここで、各サイクルごとのテンソルは、その画像チャネルの中心ピクセル内に、単一のターゲットクラスターの中心を含む。各サイクルごとのテンソルはまた、特定の配列決定サイクルにおいて画像チャネルのそれぞれに捕捉された単一のターゲットクラスターの、いくつかの隣接するクラスターの、及びそれらの周囲背景の強度発光を表現する。図 19 a では、2 つの例示的な画像チャネル、すなわち、赤色画像チャネル及び緑色画像チャネルが示されている。

【0298】

各サイクルごとのテンソルはまた、対応する画像チャネルを補足する距離チャネル（例えば、赤色距離チャネル及び緑色距離チャネル）も含む。距離チャネルは、単一のターゲットクラスターまでの対応する画像チャネル内の各ピクセルの中心間距離を識別する。各サイクルごとのテンソルは、画像チャネルのそれぞれにおいて強度値をピクセル単位でスケールするスケールチャネルを更に含む。

【0299】

特殊化したアーキテクチャは、5 つの空間的な畳み込み層及び 2 つの時間的な畳み込み層を有する。各空間的な畳み込み層は、次元数 $j \times j \times \alpha$ の k 個の畳み込みフィルタのバンク（ j は正方形フィルタの幅及び高さを表し、 α はその深さを表す）を使用して、分離された畳み込みを適用する。各時間的な畳み込み層は、次元数 $j \times j \times \alpha$ の k 個の畳み込

10

20

30

40

50

みフィルタのバンク（ j は正方形フィルタの幅及び高さを表し、 α はその深さを表す）を使用して組み合わせ畳み込みを適用する。

【0300】

特殊化したアーキテクチャは、分類前の層（例えば、平坦化層及び高密度層）及び出力層（例えば、ソフトマックス分類層）を有する。分類前の層は、出力層のための入力を調製する。出力層は、現在（時間 t ）の配列決定サイクルにおける単一のターゲットクラスターのためのベースコールを生成する。

【0301】

一貫して減少する空間次元数

【0302】

図19a、図19b、及び図19cはまた、畳み込みフィルタによって生成される、結果として得られる特徴マップ（畳み込み表現又は中間畳み込み表現又は畳み込みされた特徴又は活性化マップ）も示す。サイクルごとのテンソルから開始して、結果として得られる特徴マップの空間次元数は、1つの畳み込み層から次の畳み込み層までの間に一定ステップサイズで減少する（本明細書において「一貫して減少する空間次元数」と称される概念）。図19a、図19b、及び図19cでは、一貫して減少する空間次元数のために、2の例示的な一定ステップサイズが使用される。

【0303】

一貫して減少する空間次元数は、以下の式によって表される：「現在の特徴マップ空間次元数＝前の特徴マップ空間次元数－畳み込みフィルタ空間次元数＋1」。一貫して減少する空間次元数は、畳み込みフィルタに、中央クラスターピクセル及びそれらの隣接ピクセル上でアテンションの焦点を漸進的に狭めさせ、中央クラスターピクセル及びそれらの隣接ピクセルの間で局所的な依存性を捕捉する特徴を有する特徴マップを生成させる。これは結果として、中央クラスターピクセル内に中心が含まれているクラスターを正確にベースコールするのを助ける。

【0304】

5つの空間的な畳み込み層の分離された畳み込みは、3つの配列決定サイクル間での情報の混合を防止し、3つの別個の畳み込みパイプラインを維持する。

【0305】

2つの時間的な畳み込み層の組み合わせ畳み込みは、3つの配列決定サイクルの間で情報を混合する。第1の時間的な畳み込み層は、最終の空間的な畳み込み層によって次及び現在の配列決定サイクルのためにそれぞれ生成された次及び現在の空間畳み込み表現にわたって畳み込みを行う。これにより、第1の時間的出力が得られる。第1の時間的な畳み込み層はまた、最終の空間的な畳み込み層によって現在及び前の配列決定サイクルのためにそれぞれ生成された現在及び前の空間畳み込み表現にわたって畳み込みを行う。これにより、第2の時間的出力が得られる。第2の時間的な畳み込み層は、第1及び第2の時間的出力にわたって畳み込みを行い、最終の時間的出力を生成する。

【0306】

いくつかの実施態様では、最終の時間的出力は、平坦化層に供給されて、平坦化された出力を生成する。次いで、平坦化された出力は、高密度層に供給されて、高密度出力を生成する。高密度出力は、出力層によって処理されて、現在（時間 t ）の配列決定サイクルにおける単一のターゲットクラスターのためのベースコールを生成する。

【0307】

いくつかの実施態様では、出力層は、現在の配列決定サイクルにおける単一のターゲットクラスターに組み込まれた塩基がA、C、T、及びGであることの尤度（分類スコア）を生成し、これらの尤度に基づいて塩基をA、C、T、又はGとして分類する（例えば、図19aの塩基Aなど、最大尤度を有する塩基が選択される）。このような実施態様では、尤度は、ソフトマックス分類層によって生成される指数関数的に正規化されたスコアであり、合計すると1になる。

【0308】

10

20

30

40

50

いくつかの実施態様では、出力層は、単一のターゲットクラスターのための出力ペアを導出する。出力ペアは、現在の配列決定サイクルにおける単一のターゲットクラスターに組み込まれた塩基がA、C、T、又はGであることのクラスラベルを識別し、クラスラベルに基づいて単一のターゲットクラスターをベースコールする。一実施態様では、1, 0のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、1, 1のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。別の実施態様では、1, 1のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、0.5, 0.5のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。更に別の実施態様では、1, 0のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、0.5, 0.5のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。また更なる実施態様では、1, 2のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、1, 1のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。

【0309】

いくつかの実施態様では、出力層は、現在の配列決定サイクルにおける単一のターゲットクラスターに組み込まれた塩基がA、C、T、又はGであることを識別する単一のターゲットクラスターのためのクラスラベルを導出し、クラスラベルに基づいて単一のターゲットクラスターをベースコールする。一実施態様では、0.33のクラスラベルがA塩基を識別し、0.66のクラスラベルがC塩基を識別し、1のクラスラベルがT塩基を識別し、0のクラスラベルがG塩基を識別する。別の実施態様では、0.50のクラスラベルがA塩基を識別し、0.75のクラスラベルがC塩基を識別し、1のクラスラベルがT塩基を識別し、0.25のクラスラベルがG塩基を識別する。

【0310】

いくつかの実施態様では、出力層は、単一の出力値を導出し、この単一の出力値を塩基A、C、T、及びGに対応するクラス値範囲と比較し、この比較に基づいて、単一の出力値を特定のクラス値範囲に割り当て、この割り当てに基づいて単一のターゲットクラスターをベースコールする。一実施態様では、単一の出力値はシグモイド関数を使用して導出され、単一の出力値は0～1の範囲である。別の実施態様では、0～0.25のクラス値範囲がA塩基を表し、0.25～0.50のクラス値範囲がC塩基を表し、0.50～0.75のクラス値範囲がT塩基を表し、0.75～1のクラス値範囲がG塩基を表す。

【0311】

当業者であれば、他の実施態様において、特殊化したアーキテクチャが、より少ない又はより多い数の配列決定サイクルのための入力データを処理することができ、より少ない又はより多い数の空間的及び時間的な畳み込み層を含むことができることは理解するであろう。また、入力データの次元数、入力データ内のサイクルごとのテンソル、畳み込みフィルタ、結果として生じる特徴マップ、及び出力が異なってもよい。また、畳み込み層内の畳み込みフィルタの数が異なってもよい。それは、異なるパディング及びストライディング構成を使用することができる。それは、異なる分類関数（例えば、シグモイド又は回帰）を使用することができ、全結合層を含んでも含まなくてもよい。それは、1D重畳、2D重畳、3D重畳、4D重畳、5D重畳、拡張又は無性重畳、転置重畳、奥行分離可能な重畳、1×1重畳、グループ重畳、扁平重畳、空間及びクロスチャネルの重畳、シャッフルされたグループ化重畳、空間的な分離可能な重畳、及び逆重畳を使用することができる。それは、ロジスティック回帰／対数損失、多クラスクロスエントロピー／ソフトマックス損失、二値クロスエントロピー損失、平均二乗誤差損失、L1損失、L2損失、平滑L1損失、及びHuber損失などの1つ又はそれ以上の損失機能を使用することができる。それは、TFRecord、圧縮符号化（例えば、PNG）、シャープ化、マップ変換に対する平行コール、バッチング、プリフェッチ、モデル並列、データ並列、及び同期／非同期SGDのような、任意の並列、効率、及び圧縮方式を使用することができる。これは、アップサンプリング層、ダウンサンプリング層、反復接続、ゲート及びゲートされたメモリユニット（LSTM又はGRUなど）、残留ブロック、残留接続、ハイ

10

20

30

40

50

ウェイ接続、スキップ接続、ペエホル接続、アクティブ化機能（例えば、非線形変換関数は、整流線形ユニット（ReLU）、漏れやすいReLU、指数関数的ライナーユニット（ELU）、シグモイド及び双曲線正接（tanh））、バッチ正規化層、規則化層、ドロップアウト、プール層（例えば、最大又は平均プール）、グローバル平均プール層、及び注意機構のような非線形変換機能を含む。

【0312】

単一のクラスターのベースコールについて説明してきたが、ここからは、複数のクラスターのベースコールについて説明する。

【0313】

複数のクラスターのベースコール

10

【0314】

入力データのサイズ及びフローセル上のクラスター密度に応じて、1入力当たり10個～30万個のクラスターが、ニューラルネットワークベースのベースコーラー218によって同時にベースコールされる。これを並列プロセッサ上に実装されたデータ並列戦略及び／又はモデル並列戦略に拡張し、サイズ10のバッチ又はミニバッチを使用すると、1バッチ当たり又は1ミニバッチ当たり100個～300万個のクラスターが同時にベースコールされるという結果になる。

【0315】

配列決定構成（例えば、クラスター密度、フローセル上のタイルの数）に応じて、タイルは、2万個～30万個のクラスターを含む。別の実施態様では、IlluminaのNovaSeq配列決定器が、タイル当たり最大4百万個のクラスターを有する。したがって、タイルの配列決定画像（タイル画像）は、2万個～30万個のクラスター及びそれらの周囲背景からの強度発光を表現することができる。そのため、一実施態様では、タイル画像全体を含む入力データを使用すると、1入力当たり30万個のクラスターが同時にベースコールされるという結果になる。別の実施態様では、入力データ内でサイズ15×15ピクセルの画像パッチを使用すると、1入力当たり100個未満のクラスターが同時にベースコールされるという結果になる。当業者であれば、これらの数が、配列決定構成、並列戦略、アーキテクチャの詳細（例えば、最適なアーキテクチャハイパーパラメータに基づく）、及び利用可能な計算に応じて変化し得ることは理解するであろう。

20

【0316】

図20は、複数のターゲットクラスターに同時にベースコールする一実施態様を示す。入力データは、上述の3つの配列決定サイクルのための3つのテンソルを有する。各サイクルごとのテンソル（例えば、図18aの入力テンソル1800）は、特定の配列決定サイクルにおいて画像チャネルのそれぞれに捕捉されたベースコールされる複数のターゲットクラスター及びそれらの周囲背景の強度発光を表現する。他の実施態様では、ベースコールされない、いくつかの追加の隣接するクラスターもまた、コンテキストのために含まれる。

30

【0317】

マルチクラスターベースコール実施態様では、各サイクルごとのテンソルは、対応する画像チャネルを補足する距離チャネル（例えば、赤色距離チャネル及び緑色距離チャネル）を含む。距離チャネルは、複数のターゲットクラスターの中から最も近いクラスターまでの対応する画像チャネル内の各ピクセルの中心間距離を識別する。

40

【0318】

マルチクラスター形状ベースコール実施態様では、各サイクルごとのテンソルは、対応する画像チャネルを補足する距離チャネル（例えば、赤色距離チャネル及び緑色距離チャネル）を含む。距離チャネルは、対応する画像チャネル内の各クラスターピクセルの中心から、複数のターゲットクラスターのうちでそのクラスターピクセルに属する又はそれに帰属するクラスターの中心までの距離を識別する。

【0319】

各サイクルごとのテンソルは、画像チャネルのそれぞれにおいて強度値をピクセル単位

50

でスケールリングするスケールリングチャンネルを更に含む。

【0320】

図20では、各サイクルごとのテンソルの空間次元数は、図19aに示されるものよりも大きい。すなわち、図19aの単一のターゲットクラスターベースコール実施態様では、各サイクルごとのテンソルの空間次元数が 15×15 であるのに対し、図20の複数のクラスターベースコール実施態様では、各サイクルごとのテンソルの空間次元数は 114×114 である。追加のクラスターの強度発光を表現するより大きな量の画素化データを有することは、いくつかの実施態様によれば、複数のクラスターに対して同時に予測されるベースコールの精度を向上させる。

【0321】

冗長な畳み込みの回避

【0322】

更に、各サイクルごとのテンソル内の画像チャンネルは、配列決定画像から抽出された画像パッチから得られる。いくつかの実施態様では、空間的に連続している（例えば、左部、右部、上部、及び底部で連続している）抽出された画像パッチの間に重複ピクセルが存在する。したがって、一実施態様では、重複ピクセルは冗長な畳み込みを受けず、重複ピクセルが後続の入力の一部であるときは、前の畳み込みからの結果が後のインスタンスで再利用される。

【0323】

例えば、サイズ $n \times n$ ピクセルの第1の画像パッチが配列決定画像から抽出され、サイズ $m \times m$ ピクセルの第2の画像パッチも同じ配列決定画像から抽出され、その結果、第1及び第2の画像パッチが空間的に連続し、 $o \times o$ ピクセルの重複領域を共有する、と考える。更に、 $o \times o$ ピクセルが第1の画像パッチの一部として畳み込みされて、メモリに記憶される第1の畳み込み表現を生成すると考える。そのとき、第2の画像パッチが畳み込みされると、 $o \times o$ ピクセルが再び畳み込みされることはなく、代わりに、第1の畳み込み表現がメモリから取り出され、再利用される。いくつかの実施態様では、 $n = m$ である。他の実施態様では、それらは等しくない。

【0324】

次いで、入力データは、特殊化したアーキテクチャの空間的及び時間的な畳み込み層によって処理されて、次元数 $w \times w \times k$ の最終の時間的出力を生成する。ここでも、一貫して減少する空間次元数の事象下で、空間次元数は、各畳み込み層において2の一定ステップサイズで減少する。すなわち、 $n \times n$ の空間次元数の入力データで開始すると、 $w \times w$ の空間次元数の最終の時間的出力が導出される。

【0325】

次いで、空間次元数 $w \times w$ の最終の時間的出力に基づいて、出力層が、 $w \times w$ 組のユニット内の各ユニットのためのベースコールを生成する。一実施態様では、出力層は、ユニットごとに4つの塩基（A、C、T、及びG）のための4とおりの分類スコアを生成するソフトマックス層である。すなわち、 $w \times w$ 組のユニット内の各ユニットは、図20に示されるように、対応するソフトマックスクアドルプルにおける最大分類スコアに基づいてベースコールが割り当てられる。いくつかの実施態様では、 $w \times w$ 組のユニットは、最終の時間的出力を平坦化層及び高密度層によって処理して平坦化された出力及び高密度出力をそれぞれ生成することの結果として導出される。このような実施態様では、平坦化された出力は、 $w \times w \times k$ 個の要素を有し、高密度出力は、 $w \times w$ 組のユニットを形成する $w \times w$ 個の要素を有する。

【0326】

複数のターゲットクラスターのためのベースコールは、 $w \times w$ 組のユニット内のベースコールされたユニットのうちのどれが、中央クラスターピクセル、すなわち、複数のターゲットクラスターのそれぞれの中心を含む入力データ内のピクセルと一致するか、又はそれに対応するかを識別することによって取得される。所与のターゲットクラスターには、所与のターゲットクラスターの中心を含むピクセルと一致する又はそれに対応するユニッ

10

20

30

40

50

トのベースコールが割り当てられる。換言すれば、中央クラスターピクセルと一致しない又はそれに対応しないユニットのベースコールはフィルタで除外される。この機能は、ベースコールフィルタリング層によって操作可能になっており、ベースコールフィルタリング層は、いくつかの実施態様では特殊化したアーキテクチャの一部であり、又は他の実施態様では後処理モジュールとして実装されている。

【0327】

他の実施態様では、複数のターゲットクラスターのためのベースコールは、 $w \times w$ 組のユニット内のベースコールされたユニットのどのグループが同じクラスターを覆うかを識別することによって、すなわち、同じクラスターを表現する入力データ内のピクセルグループを識別することによって取得される。次いで、各クラスター及びその対応するピクセルグループについて、それぞれの4つの塩基クラス（A、C、T、及びG）の分類スコア（ソフトマックス確率）の平均が、ピクセルグループ内のピクセルにわたって計算され、最も高い平均分類スコアを有する塩基クラスが、クラスターをベースコールするために選択される。

【0328】

訓練中、いくつかの実施態様では、グラウンドトゥルス比較及び誤差計算は、中央クラスターピクセルと一致する又はそれらに対応するユニットに対してのみ行われ、その結果、それらの予測されたベースコールは、グラウンドトゥルスラベルとして識別される正しいベースコールに対して評価される。

【0329】

複数のクラスターのベースコールについて説明してきたが、ここからは、複数のクラスター及び複数のサイクルのベースコールについて説明する。

【0330】

複数のクラスター及び複数のサイクルのベースコール

【0331】

図21は、複数の継続的な配列決定サイクルにおける複数のターゲットクラスターを同時にベースコールし、それによって複数のターゲットクラスターのそれぞれのためのベースコール配列を同時に生成する一実施態様を示す。

【0332】

上述の単一及び複数のベースコール実施態様では、1つの配列決定サイクル（現在（時間 t ）の配列決定サイクル）におけるベースコールは、3つの配列決定サイクル（現在（時間 t ）、前／左側フランキング（時間 $t-1$ ）、及び次／右側フランキング（時間 $t+1$ ）の配列決定サイクル）のためのデータを使用して予測され、右側及び左側フランキングの配列決定サイクルは、塩基トリプレットモチーフのための配列特異的なコンテキスト、並びにプレフェージング信号及びフェージング信号の2次の寄与を提供する。この関係は、以下の式によって表される：「入力データにデータが含まれている配列決定サイクルの数（ t ）＝ベースコールされている配列決定サイクルの数（ y ）＋右側及び左側フランキングの配列決定サイクルの数（ x ）」。

【0333】

図21では、入力データは、 t 個の配列決定サイクルのための t 個のサイクルごとのテンソルを含んでおり、それにより、入力データの次元数は $n \times n \times m \times t$ となり、ここで、 $n=114$ 、 $m=5$ 、及び $t=15$ である。他の実施態様では、これらの次元数は異なる。 t 個の配列決定サイクルのうち、第 t の配列決定サイクル及び第1の配列決定サイクルは、右側及び左側フランキングコンテキスト x として機能し、それらの間の y 個の配列決定サイクルがベースコールされる。したがって、 $y=13$ 、 $x=2$ 、及び $t=y+x$ である。各サイクルごとのテンソルは、図18aの入力テンソル1800など、画像チャンネル、対応する距離チャンネル、及びスケールチャンネルを含む。

【0334】

次いで、 t 個のサイクルごとのテンソルを有する入力データは、特殊化したアーキテクチャの空間的及び時間的な畳み込み層によって処理されて、 y 個の最終の時間的出力を生

10

20

30

40

50

成し、これらのそれぞれは、ベースコールされている y 個の配列決定サイクルのそれぞれ 1 つに対応する。 y 個の最終の時間的出力のそれぞれは、 $w \times w \times k$ の次元数を有する。ここでも、一貫して減少する空間次元数の事象下で、空間次元数は、各畳み込み層において 2 の一定ステップサイズで減少する。すなわち、 $n \times n$ の空間次元数の入力データで開始すると、 $w \times w$ の空間次元数の y 個の最終の時間的出力のそれぞれが導出される。

【0335】

次に、 y 個の最終の時間的出力のそれぞれが、出力層によって並列に処理される。 y 個の最終の時間的出力のそれぞれについて、出力層は、 $w \times w$ 組のユニット内の各ユニットのためのベースコールを生成する。一実施態様では、出力層は、ユニットごとに 4 つの塩基（A、C、T、及び G）のための 4 とおりの分類スコアを生成するソフトマックス層である。すなわち、 $w \times w$ 組のユニット内の各ユニットは、図 20 に示されるように、対応するソフトマックスアダプトルにおける最大分類スコアに基づいてベースコールが割り当てられる。いくつかの実施態様では、 $w \times w$ 組のユニットは、 y 個の最終の時間的出力のそれぞれについて、後のものを平坦化層及び高密度層によってそれぞれに処理して対応する平坦化された出力及び高密度出力を生成することの結果として導出される。このような実施態様では、各平坦化された出力は、 $w \times w \times k$ 個の要素を有し、各高密度出力は、 $w \times w$ 組のユニットを形成する $w \times w$ 個の要素を有する。

【0336】

y 個の配列決定サイクルのそれぞれについて、複数のターゲットクラスターのためのベースコールは、対応する $w \times w$ 組のユニット内のベースコールされたユニットのうちのどれが、中央クラスターピクセル、すなわち、複数のターゲットクラスターのそれぞれの中心を含む入力データ内のピクセルと一致するか、又はそれに対応するかを識別することによって取得される。所与のターゲットクラスターには、所与のターゲットクラスターの中心を含むピクセルと一致する又はそれに対応するユニットのベースコールが割り当てられる。換言すれば、中央クラスターピクセルと一致しない又はそれに対応しないユニットのベースコールはフィルタで除外される。この機能は、ベースコールフィルタリング層によって操作可能になっており、ベースコールフィルタリング層は、いくつかの実施態様では特殊化したアーキテクチャの一部であり、又は他の実施態様では後処理モジュールとして実装されている。

【0337】

訓練中、いくつかの実施態様では、グラウンドトゥールズ比較及び誤差計算は、中央クラスターピクセルと一致する又はそれらに対応するユニットに対してのみ行われ、その結果、それらの予測されたベースコールは、グラウンドトゥールズラベルとして識別される正しいベースコールに対して評価される。

【0338】

入力ごとに、 y 個の配列決定サイクルのそれぞれにおける複数のターゲットクラスターのそれぞれのためのベースコール、すなわち、複数のターゲットクラスターのそれぞれのための長さ y のベースコール配列がもたらされる。他の実施態様では、 y は 20、30、50、150、300 などである。当業者であれば、これらの数が、配列決定構成、並列戦略、アーキテクチャの詳細（例えば、最適なアーキテクチャハイパーパラメータに基づく）、及び利用可能な計算に応じて変化し得ることは理解するであろう。

【0339】

エンドツーエンドの次元数ダイアグラム

【0340】

以下の説明は、画像データからベースコールを生成するのに必要な基礎となるデータ次元数変化の異なる実施態様を、上記データ次元数変化を実現するデータオペレータの次元数と共に例示するために、次元数ダイアグラムを使用する。

【0341】

図 22、図 23、及び図 24 では、矩形は、空間的及び時間的な畳み込み層並びにソフトマックス分類層のようなデータオペレータを表し、角に丸みを帯びた矩形は、データオ

10

20

30

40

50

ペレータによって生成されたデータ（例えば、特徴マップ）を表す。

【0342】

図22は、単一のクラスターのベースコールの実施態様のための次元数ダイアグラム2200を示す。入力「サイクル次元」は3であり、第1の時間的な畳み込み層に至るまでは、結果として得られる特徴マップに対して3であり続けることに留意されたい。3のサイクル次元は、3つの配列決定サイクルを提示しており、その連続性は、3つの配列決定サイクルのための特徴マップが別個に生成及び畳み込みされ、かつ3つの配列決定サイクルの間で特徴が混合されないことを表す。分離された畳み込みパイプラインは、空間的な畳み込み層の深さ単位の分離された畳み込みフィルタによって実現される。空間的な畳み込み層の深さ単位の分離された畳み込みフィルタの「深さ次元数」は1であることに留意されたい。これは、深さ単位の分離された畳み込みフィルタが、所与の配列決定サイクルのみ、すなわち、イントラサイクルのデータ及び結果として得られる特徴マップにわたって畳み込みを行うことを可能にし、かつ深さ単位の分離された畳み込みフィルタが、任意の他の配列決定サイクルのデータ及び結果として得られる特徴マップにわたって畳み込みを行うことを防止するものである。

10

【0343】

対照的に、時間的な畳み込み層の深さ単位の組み合わせ畳み込みフィルタの深さ次元数は2であることに留意されたい。これは、深さ単位の組み合わせ畳み込みフィルタが、複数の配列決定サイクルからの結果として得られた特徴マップにわたってグループ単位の畳み込みを行い、配列決定サイクル間で特徴を混合することを可能にするものである。

20

【0344】

また、2の一定ステップサイズによる「空間次元数」の一貫した減少にも留意されたい。

【0345】

更に、4つの要素を有するベクトルが、ソフトマックス層によって指数関数的に正規化されて、4つの塩基（A、C、T、及びG）のための分類スコア（すなわち、信頼性スコア、確率、尤度、ソフトマックススコア）を生成する。最も高い（最大）ソフトマックススコアを有する塩基は、現在の配列決定サイクルにおいてベースコールされている単一のターゲットクラスターに割り当てられる。

【0346】

当業者であれば、他の実施態様では、例示された次元数が、配列決定構成、並列戦略、アーキテクチャの詳細（例えば、最適なアーキテクチャハイパーパラメータに基づく）、及び利用可能な計算に応じて変化し得ることは理解するであろう。

30

【0347】

図23は、複数のクラスター、単一の配列決定サイクルのベースコールの実施態様のための次元数ダイアグラム2300を示す。単一のクラスターのベースコールに対するサイクル、深さ、及び空間次元数に関する上記の説明は、この実施態様に適用される。

【0348】

ここで、ソフトマックス層は、10,000個のユニットのそれぞれに対して独立して動作し、10,000個のユニットのそれぞれについてソフトマックススコアのそれぞれのクアドルプルを生成する。クアドルプルは、4つの塩基（A、C、T、及びG）に対応する。いくつかの実施態様では、10,000個のユニットは、64,000個の平坦化されたユニットの10,000個の高密度ユニットへの変換から導出される。

40

【0349】

次いで、10,000個のユニットのそれぞれのソフトマックススコアクアドルプルから、各クアドルプルにおいて最も高いソフトマックススコアを有する塩基が、10,000個のユニットのそれぞれ1つに割り当てられる。

【0350】

次いで、10,000個のユニットのうち、現在の配列決定サイクルにおいて同時にベースコールされている2,500個のターゲットクラスターのそれぞれの中心を含んでいる2,500個の中央クラスターピクセルに対応する、それらの2500個のユニットが

50

選択される。選択された2,500個のユニットに割り当てられた塩基は、結果として、2,500個のターゲットクラスターのうちの対応するものに割り当てられる。

【0351】

当業者であれば、他の実施態様では、例示された次元数が、配列決定構成、並列戦略、アーキテクチャの詳細（例えば、最適なアーキテクチャハイパーパラメータに基づく）、及び利用可能な計算に応じて変化し得ることは理解するであろう。

【0352】

図24は、複数のクラスター、複数の配列決定サイクルのベースコールの実施態様のための次元数ダイアグラム2400を示す。単一のクラスターのベースコールに対するサイクル、深さ、及び空間次元数に関する上記の説明は、この実施態様に適用される。

10

【0353】

更に、複数のクラスターのベースコールに対するソフトマックススペースのベースコール分類についての上記の説明は、ここでも適用される。しかしながら、ここでは、2,500個のターゲットクラスターのソフトマックススペースのベースコール分類は、ベースコールされた13個の配列決定サイクルのそれぞれに対して並列に行われ、それによって、2,500個のターゲットクラスターのそれぞれのための13個のベースコールを同時に生成する。

【0354】

当業者であれば、他の実施態様では、例示された次元数が、配列決定構成、並列戦略、アーキテクチャの詳細（例えば、最適なアーキテクチャハイパーパラメータに基づく）、及び利用可能な計算に応じて変化し得ることは理解するであろう。

20

【0355】

アレイ化された入力対スタック化された入力

【0356】

ここからは、ニューラルネットワークベースのコラーへの多重サイクル入力データが配置され得る2つの構成について説明する。第1の構成は「アレイ化された入力」と呼ばれ、第2の構成は「スタック化された入力」と呼ばれる。アレイ化された入力は、図25aに示されており、図19a～図24に関して上に説明されている。サイクルごとの入力内の画像パッチは、残差レジストレーション誤差に起因して互いに対して位置がずれていることから、アレイ化された入力は、各配列決定サイクルの入力を別個の列／ブロック内に符号化する。特殊化したアーキテクチャは、別個の列／ブロックのそれぞれの処理を分離するために、アレイ化された入力と共に使用される。また、距離チャンネルは、サイクル内での画像パッチ間の位置ずれ、及びサイクル間での画像パッチ間の位置ずれを考慮するために、変換されたクラスター中心を使用して計算される。

30

【0357】

対照的に、図25bに示される、スタック化された入力は、異なる配列決定サイクルからの入力を単一の列／ブロック内に符号化する。一実施態様では、これにより、スタック化された入力内の画像パッチが、インターサイクル及びイントラサイクルの残差レジストレーション誤差を排除する、アフィン変換及び強度補間によって互いに位置合わせされているため、特殊化したアーキテクチャを使用する必要性がなくなる。いくつかの実施態様では、スタック化された入力は、全ての入力に対して共通のスケーリングチャンネルを有する。

40

【0358】

別の実施態様では、各画像パッチの中心ピクセルの中心が、ベースコールされている単一のターゲットクラスターの中心と一致するように、画像パッチをリフ্রেーム又はシフトするために、強度補間が使用される。これにより、全ての非中心ピクセルが単一のターゲットクラスターの中心から等距離であるため、補足的な距離チャンネルを使用する必要性がなくなる。距離チャンネルなしのスタック化された入力は、本明細書において「リフ্রেームされた入力」と称され、図27に例示されている。

【0359】

50

しかしながら、複数のクラスターを伴うベースコール実施態様では、画像パッチに、ベースコールされる複数の中央クラスターピクセルが含まれるため、リフレームすることは実行不可能であり得る。距離チャンネルなし、かつリフレームなしのスタック化された入力、本明細書において「整列された入力」と称され、図28及び図29に例示されている。整列された入力は、距離チャンネルの計算が所望されず（例えば、計算制限に起因する）、リフレームすることが不可能であるときに使用され得る。

【0360】

以下のセクションは、特殊化したアーキテクチャ及び補足的な距離チャンネルを使用せず、代わりに標準的な畳み込み層及びフィルタを使用する、様々なベースコール実施態様について説明する。

【0361】

リフレームされた入力：距離チャンネルなしの整列された画像パッチ

【0362】

図26aは、ベースコールされているターゲットクラスターの中心を中心ピクセルの中心に配置するように画像パッチ2602のピクセルをリフレームする（2600a）一実施態様を示す。ターゲットクラスターの中心（紫色）は、画像パッチ2602の中心ピクセル内に収まっているが、図2600aに示すように、中心ピクセルの中心からあるオフセット（赤色）の位置にある。

【0363】

オフセットを排除するために、リフレーマー2604が、ピクセルの強度を補間してリフレームを補正することによって画像パッチ2602をシフトさせ、リフレームされた／シフトされた画像パッチ2606を生成する。シフトされた画像パッチ2606では、中心ピクセルの中心は、ターゲットクラスターの中心と一致する。また、非中心ピクセルは、ターゲットクラスターの中心から等距離である。補間は、最近傍強度抽出、ガウシアンベースの強度抽出、 2×2 サブピクセル領域の平均に基づく強度抽出、 2×2 サブピクセル領域の最大輝度に基づく強度抽出、 3×3 サブピクセル領域の平均に基づく強度抽出、バイリニア強度抽出、バイキュービック強度抽出、及び／又は重み付け面積カバー率に基づく強度抽出によって実行することができる。これらの技術は、「強度抽出方法」と題された付録に詳細に記載されている。

【0364】

図26bは、(i) 中心ピクセルの中心がターゲットクラスターの中心と一致し、(ii) 非中心ピクセルがターゲットクラスターの中心から等距離である、別の例示的なリフレームされた／シフトされた画像パッチ2600bを示す。これらの2つの要因により、全ての非中心ピクセルがターゲットクラスターの中心に対して同じ程度の近接度を有するため、補足的な距離チャンネルを提供する必要性がなくなる。

【0365】

図27は、標準的な畳み込みニューラルネットワーク及びリフレームされた入力を使用して、現在の配列決定サイクルにおける単一のターゲットクラスターをベースコールする一実施態様を示す。図示された実施態様では、リフレームされた入力は、ベースコールされている現在(t)の配列決定サイクルのための現在の画像パッチセット、前(t-1)の配列決定サイクルのための前の画像パッチセット、及び次(t+1)の配列決定サイクルのための次の画像パッチセットを含む。各画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。図27は、2つの画像チャンネル、すなわち、赤色チャンネル及び緑色チャンネルを示す。各画像パッチは、ベースコールされているターゲットクラスター、いくつかの隣接するクラスター、及びそれらの周囲背景を覆うピクセルのためのピクセル強度データを有する。リフレームされた入力はまた、共通のスケールチャンネルも含む。

【0366】

図26a～図26bに関して上述したように、リフレームされた入力は、画像パッチが、中心にターゲットクラスターの中心を配置するようにリフレーム又はシフトされるため

10

20

30

40

50

、いかなる距離チャンネルも含んでいない。また、画像パッチは、インターサイクル及びイントラサイクルの残差レジストレーション誤差を除去するために互いに位置合わせされる。一実施態様では、これは、アフィン変換及び強度補間を使用して行われ、その更なる詳細は、付録 1、2、3、及び 4 に見出すことができる。これらの要因により、特殊化したアーキテクチャを使用する必要性がなくなり、代わりに、標準的な畳み込みニューラルネットワークが、リフレームされた入力と共に使用される。

【0367】

図示された実施態様では、標準的な畳み込みニューラルネットワーク 2700 は、標準的な畳み込みフィルタを使用する 7 つの標準的な畳み込み層を含む。これは、配列決定サイクル間でのデータの混合を防止するための分離された畳み込みパイプラインが存在しないことを意味する（データが整列されており、混合することができるので）。いくつかの実施態様では、中央クラスター中心及びその隣接ピクセルに対して他のピクセルよりも着目するように標準的な畳み込みフィルタを教育するために、一貫して減少する空間次元数の事象が使用される。

【0368】

次いで、リフレームされた入力は、標準的な畳み込み層によって処理されて、最終畳み込み表現を生成する。最終畳み込み表現に基づいて、現在の配列決定サイクルにおけるターゲットクラスターのためのベースコールは、図 19c に関して上述したように、平坦化層、高密度層、及び分類層を使用して同様の様式で取得される。

【0369】

いくつかの実施態様では、プロセスは、複数の配列決定サイクルにわたって反復されて、ターゲットクラスターのためのベースコールの配列を生成する。

【0370】

他の実施態様では、プロセスは、複数のターゲットクラスターのための複数の配列決定サイクルにわたって反復されて、複数のターゲットクラスター内の各ターゲットクラスターのためのベースコールの配列を生成する。

【0371】

整列された入力：距離チャンネル及びリフレームなしの整列された画像パッチ

【0372】

図 28 は、標準的な畳み込みニューラルネットワーク及び整列された入力を使用して、現在の配列決定サイクルにおける複数のターゲットクラスターをベースコールする一実施態様を示す。ここで、リフレームすることは、画像パッチが、ベースコールされている複数の中央クラスターピクセルを含んでいるため、実行不可能である。その結果、整列された入力内の画像パッチは、リフレームされない。更に、一実施態様によれば、補足的な距離チャンネルは、計算考慮事項に起因して含まれていない。

【0373】

次いで、整列された入力は、標準的な畳み込み層によって処理されて、最終畳み込み表現を生成する。最終畳み込み表現に基づいて、ターゲットクラスターのそれぞれのためのベースコールは、図 20 に関して上述したように平坦化層（任意選択）、高密度層（任意選択）、分類層、及びベースコールフィルタリング層を使用して、同様の様式で現在の配列決定サイクルにおいて取得される。

【0374】

図 29 は、標準的な畳み込みニューラルネットワーク及び整列された入力を使用して、複数の配列決定サイクルにおける複数のターゲットクラスターをベースコールする一実施態様を示す。整列された入力は、標準的な畳み込み層によって処理されて、ベースコールされている y 個の配列決定サイクルのそれぞれのための最終畳み込み表現を生成する。y 個の最終畳み込み表現に基づいて、ターゲットクラスターのそれぞれのためのベースコールは、図 21 に関して上述したように平坦化層（任意選択）、高密度層（任意選択）、分類層、及びベースコールフィルタリング層を使用して、ベースコールされている y 個の配列決定サイクルのそれぞれのために同様の様式で取得される。

【0375】

当業者であれば、他の実施態様において、標準的な畳み込みニューラルネットワークが、より少ない又はより多い数の配列決定サイクルのためのリフレームされた入力进行处理することができ、より少ない又はより多い数の標準的な畳み込み層を含むことができることは理解するであろう。また、リフレームされた入力の次元数、リフレームされた入力内のサイクルごとのテンソル、畳み込みフィルタ、結果として生じる特徴マップ、及び出力が異なってもよい。また、畳み込み層内の畳み込みフィルタの数が異なってもよい。それは、1D重畳、2D重畳、3D重畳、4D重畳、5D重畳、拡張又は無性重畳、転置重畳、奥行分離可能な重畳、 1×1 重畳、グループ重畳、扁平重畳、空間及びクロスチャネルの重畳、シャッフルされたグループ化重畳、空間的な分離可能な重畳、及び逆重畳を使用することができる。それは、ロジスティック回帰／対数損失、多クラスクロスエントロピー／ソフトマックス損失、二値クロスエントロピー損失、平均二乗誤差損失、L1損失、L2損失、平滑L1損失、及びHuber損失などの1つ又はそれ以上の損失機能を使用することができる。それは、TFRecord、圧縮符号化（例えば、PNG）、シャープ化、マップ変換に対する平行コール、バッチング、プリフェッチ、モデル並列、データ並列、及び同期／非同期SGDのような、任意の並列、効率、及び圧縮方式を使用することができる。これは、アップサンプリング層、ダウンサンプリング層、反復接続、ゲート及びゲートされたメモリユニット（LSTM又はGRUなど）、残留ブロック、残留接続、ハイウェイ接続、スキップ接続、ペエホル接続、アクティブ化機能（例えば、非線形変換関数は、整流線形ユニット（ReLU）、漏れやすいReLU、指数関数的ライナーユニット（ELU）、シグモイド及び双曲線正接（tanh））、バッチ正規化層、規則化層、ドロップアウト、プール層（例えば、最大又は平均プール）、グローバル平均プール層、及び注意機構のような非線形変換機能を含む。

10

20

【0376】

訓練

【0377】

図30は、ニューラルネットワークベースのベースコーラー218を訓練する（3000）一実施態様を示す。特殊化したアーキテクチャ及び標準的なアーキテクチャの両方で、ニューラルネットワークベースのベースコーラー218は、予測されたベースコール3004を正しいベースコール3008と比較し、比較に基づいて誤差3006を計算する、逆伝播ベースの勾配更新技術を使用して訓練される。そのとき、誤差3006は勾配を計算するために使用され、勾配は、後方伝播3010中にニューラルネットワークベースのベースコーラー218の重み及びパラメータに適用される。訓練3000は、アダムなどの確率的勾配更新アルゴリズムを使用して、訓練器1510によって操作可能である。

30

【0378】

訓練器1510は、訓練データ3002（配列決定画像108に由来する）を使用して、ニューラルネットワークベースのベースコーラー218を、予測されたベースコール3004を生成する前方伝播3012及び誤差3006に基づいて重み及びパラメータを更新する後方伝播3010の数千～数百万回の反復にわたって訓練する。訓練3000に関する更なる詳細は、「深層学習ツール」と題された付録に見出すことができる。

40

【0379】

CNN-RNNベースのベースコーラー

【0380】

ハイブリッドニューラルネットワーク

【0381】

図31aは、ニューラルネットワークベースのベースコーラー218として使用されるハイブリッドニューラルネットワーク3100aの一実施態様を示す。ハイブリッドニューラルネットワーク3100aは、少なくとも1つの畳み込みモジュール3104（又は畳み込みニューラルネットワーク（CNN））及び少なくとも1つの反復モジュール3108（又は反復ニューラルネットワーク（RNN））を含む。反復モジュール3108は

50

、畳み込みモジュール 3104 からの入力を使用及び／又は受信する。

【0382】

畳み込みモジュール 3104 は、1 つ又はそれ以上の畳み込み層によって入力データ 3102 を処理し、畳み込み出力 3106 を生成する。一実施態様では、入力データ 3102 は、「入力」と題されたセクションに上述したように、画像チャネル又は画像データのみを主入力として含む。ハイブリッドニューラルネットワーク 3100 a に供給される画像データは、上述した画像データ 202 と同じであってもよい。

【0383】

別の実施態様では、入力データ 3102 はまた、画像チャネル又は画像データに加えて、「入力」と題されたセクションに上述したように、距離チャネル、スケーリングチャネル、クラスター中心座標、及び／又はクラスター帰属情報などの補足的なチャネルも含む。

【0384】

画像データ（すなわち、入力データ 3102）は、1 つ又はそれ以上のクラスター及びそれらの周囲背景の強度発光を表現する。畳み込みモジュール 3104 は、配列決定実行の一連の配列決定サイクルのための画像データを畳み込み層によって処理し、画像データの 1 つ又はそれ以上の畳み込み表現（すなわち、畳み込み出力 3106）を生成する。

【0385】

一連の配列決定サイクルは、ベースコールされる t 個の配列決定サイクルのための画像データを含むことができ、 t は 1 ～ 1000 の任意の数である。 t が 15 ～ 21 であるときに、正確なベースコール結果が観察されている。

【0386】

反復モジュール 3110 は、畳み込み出力 3106 を畳み込み、反復出力 3110 を生成する。特に、反復モジュール 3110 は、畳み込み表現及び前の隠れ状態表現を畳み込むことに基づいて現在の隠れ状態表現（すなわち、反復出力 3110）を生成する。

【0387】

一実施態様では、反復モジュール 3110 は、畳み込み表現及び前の隠れ状態表現に三次元（3D）畳み込みを適用し、以下のように数式化される、現在の隠れ状態表現を生成する。

【0388】

$$H_t = W_{13D CONV} V_t + W_{23D CONV} h_{t-1}, \text{ 式中}$$

h_t は、現在の時間ステップ t において生成された現在の隠れ状態表現を表し、

V_t は、現在の時間ステップ t における現在のスライディングウィンドウにおいて入力体積を形成する、畳み込み表現のセット又はグループを表し、

$W_{13D CONV}$ は、 V_t に適用された第 1 の 3D 畳み込みフィルタの重みを表し、

h_{t-1} は、前の時間ステップ $t-1$ において生成された前の隠れ状態表現を表し、

$W_{23D CONV}$ は、 h_{t-1} に適用された第 2 の 3D 畳み込みフィルタの重みを表す。

【0389】

いくつかの実施態様では、 $W_{13D CONV}$ 及び $W_{23D CONV}$ は、重みが共有されるため同じである。

【0390】

次いで、出力モジュール 3112 が、反復出力 3110 に基づいてベースコール 3114 を生成する。いくつかの実施態様では、出力モジュール 3112 は、1 つ又はそれ以上の全結合層と分類層（例えば、ソフトマックス）とを含む。このような実施態様では、現在の隠れ状態表現は全結合層によって処理され、全結合層の出力は、分類層によって処理されて、ベースコール 3114 を生成する。

【0391】

ベースコール 3114 は、クラスターのうちの少なくとも 1 つのための、かつ配列決定サイクルのうちの少なくとも 1 つのためのベースコールを含む。いくつかの実施態様では、ベースコール 3114 は、クラスターのそれぞれのための、かつ配列決定サイクルのそれぞれのためのベースコールを含む。したがって、例えば、入力データ 3102 が 25 個

10

20

30

40

50

のクラスターのための、かつ 15 個の配列決定サイクルのための画像データを含むとき、ベースコール 3 1 0 2 は、25 個のクラスターのそれぞれに対する 15 個のベースコールからなるベースコール配列を含む。

【0 3 9 2】

3 D 畳み込み

【0 3 9 3】

図 3 1 b は、現在の隠れ状態表現を生成するためにハイブリッドニューラルネットワーク 3 1 0 0 b の反復モジュール 3 1 1 0 によって使用される 3 D 畳み込み 3 1 0 0 b の一実施態様を示す。

【0 3 9 4】

3 D 畳み込みは、入力体積内に存在する各ボクセルが、畳み込みカーネルの等価位置におけるボクセルによって乗算される数学的操作である。最後には、結果の合計が出力体積に加算される。図 3 1 b では、3 D 畳み込み演算の表現を観察することができ、入力 3 1 1 6 内の強調されたボクセル 3 1 1 6 a が、カーネル 3 1 1 8 内のそれらの対応するボクセルで乗算される。これらの計算後、それらの合計 3 1 2 0 a が、出力 3 1 2 0 に加算される。

【0 3 9 5】

入力体積の座標は (x, y, z) によって与えられ、畳み込みカーネルはサイズ (P, Q, R) を有するので、3 D 畳み込み演算は、以下のように数学的に定義することができる。

【0 3 9 6】

【数 1 7】

$$O_{xyz} = \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} K_{pqr} I_{(x+p)(y+q)(z+r)}$$

【0 3 9 7】

式中

O は、畳み込みの結果であり、

I は、入力体積であり、

K は、畳み込みカーネルであり、

(p, q, r) は、K の座標である。

【0 3 9 8】

バイアス項は、明確さを改善するために上記の式から省略されている。

【0 3 9 9】

3 D 畳み込みは、2 D 畳み込みのようにマトリックスから空間的な情報を抽出することに加えて、連続するマトリックスの間に存在する情報も抽出する。これにより、3 D 畳み込みは、3 D オブジェクトの空間的な情報及び連続画像のセットの時間的な情報の両方をマッピングすることが可能になる。

【0 4 0 0】

畳み込みモジュール

【0 4 0 1】

図 3 2 は、ベースコールされる一連の 7 個の配列決定サイクルのうちの単一の配列決定サイクルのためのサイクルごとの入力データ 3 2 0 2 を畳み込みモジュール 3 1 0 4 の畳み込み層 3 2 0 0 のカスケードによって処理する一実施態様を示す。

【0 4 0 2】

畳み込みモジュール 3 1 0 4 は、サイクルごとの入力データのシーケンス内の各サイクルごとの入力データを畳み込み層のカスケード 3 2 0 0 によって別個に処理する。サイク

10

20

30

40

50

ルごとの入力データのシーケンスは、ベースコールされる配列決定実行の一連の t 個の配列決定サイクルのために生成され、 t は $1 \sim 1000$ の任意の数である。したがって、例えば、この一連に 15 個の配列決定サイクルが含まれるとき、サイクルごとの入力データのシーケンスは、 15 個の異なるサイクルごとの入力データを含む。

【0403】

一実施態様では、各サイクルごとの入力データは、画像チャンネル（例えば、赤色チャンネル及び緑色チャンネル）又は画像データ（例えば、上記の画像データ 202 ）のみを含む。画像チャンネル又は画像データは、一連内のそれぞれの配列決定サイクルにおいて捕捉された 1 つ又はそれ以上のクラスター及びそれらの周囲背景の強度発光を表現する。別の実施態様では、各サイクルごとの入力データは、画像チャンネル又は画像データに加えて、距離チャンネル及びスケーリングチャンネルなどの補足的なチャンネルも含む（例えば、上述の入力データ 1800 ）。

【0404】

図示された実施態様では、サイクルごとの入力データ 3202 は、ベースコールされる一連の t 個の配列決定サイクルのうちの単一の配列決定サイクルのための 2 つの画像チャンネル、すなわち、赤色チャンネル及び緑色チャンネルを含む。各画像チャンネルは、サイズ 15×15 の画像パッチ内に符号化される。畳み込みモジュール 3104 は、 5 つの畳み込み層を含む。各畳み込み層は、サイズ 3×3 の 25 個の畳み込みフィルタのバンクを有する。更に、畳み込みフィルタは、入力画像又はテンソルの高さ及び幅を維持する、いわゆる $SAME$ パディングを使用する。 $SAME$ パディングを使用すると、出力特徴マップが入力特徴と同じサイズを有するように、入力特徴にパディングが追加される。対照的に、いわゆる $VALID$ パディングは、パディングなしを意味する。

【0405】

第 1 の畳み込み層 3204 は、サイクルごとの入力データ 3202 を処理し、サイズ $15 \times 15 \times 25$ の第 1 の畳み込み表現 3206 を生成する。第 2 の畳み込み層 3208 は、第 1 の畳み込み表現 3206 を処理し、サイズ $15 \times 15 \times 25$ の第 2 の畳み込み表現 3210 を生成する。第 3 の畳み込み層 3212 は、第 2 の畳み込み表現 3210 を処理し、サイズ $15 \times 15 \times 25$ の第 3 の畳み込み表現 3214 を生成する。第 4 の畳み込み層 3216 は、第 3 の畳み込み表現 3214 を処理し、サイズ $15 \times 15 \times 25$ の第 4 の畳み込み表現 3218 を生成する。第 5 の畳み込み層 3220 は、第 4 の畳み込み表現 3218 を処理し、サイズ $15 \times 15 \times 25$ の第 5 の畳み込み表現 3222 を生成する。 $SAME$ パディングは、結果として生じる畳み込み表現の空間次元（例えば、 15×15 ）を維持することに留意されたい。いくつかの実施態様では、畳み込み層内の畳み込みフィルタの数は、 2 、 4 、 16 、 32 、 64 、 128 、 256 、 512 、及び 1024 など、 2 のべき乗である。

【0406】

畳み込みがより深くなると、情報が失われる可能性がある。これを考慮するために、いくつかの実施態様では、スキップコネクションを使用して、（ 1 ）元のサイクルごとの入力データを再導入すること、及び（ 2 ）それより前の畳み込み層によって抽出された低レベルの空間的特徴をそれより後の畳み込み層によって抽出された高レベルの空間的特徴と組み合わせることを行う。これにより、ベースコール精度の改善が観察される。

【0407】

図 33 は、単一の配列決定サイクルのサイクルごとの入力データ 3202 を、畳み込みモジュール 3104 の畳み込み層のカスケード 3200 によって生成されたその対応する畳み込み表現 3206 、 3210 、 3214 、 3218 、及び 3222 と混合する（ 3300 ）一実施態様を示す。畳み込み表現 3206 、 3210 、 3214 、 3218 、及び 3222 は連結されて畳み込み表現のシーケンス 3304 を形成し、次に、サイクルごとの入力データ 3202 と連結されて、混合表現 3306 を生成する。他の実施態様では、連結の代わりに総和が使用される。また、混合 3300 は、混合器 3302 によって操作可能になっている。

【0408】

次いで、平坦化器3308が、混合表現3306を平坦化し、サイクルごとの平坦化された混合表現3310を生成する。いくつかの実施態様では、平坦化された混合表現3310は、サイクルごとの入力データ3202並びに畳み込み表現3206、3210、3214、3218、及び3222と少なくとも1つの次元サイズを共有する高次元ベクトル又は二次元（2D）アレイである（例えば、 15×1905 、すなわち、同じ行方向次元）。これは、下流の3D畳み込みにおける特徴抽出を容易にするデータの対称性を生じさせる。

【0409】

図32及び図33は、ベースコールされる一連の t 個の配列決定サイクルのうちの単一の配列決定サイクルのためのサイクルごとの画像データ3202の処理を例示する。畳み込みモジュール3104は、 t 個の配列決定サイクルのそれぞれのためのそれぞれのサイクルごとの画像データを別個に処理し、 t 個の配列決定サイクルのそれぞれのためのそれぞれのサイクルごとの平坦化された混合表示を生成する。

【0410】

スタック化

【0411】

図34は、継続的な配列決定サイクルの平坦化された混合表現をスタック3400として配列する一実施態様を示す。図示された実施態様では、15個の配列決定サイクルのための15個の平坦化された混合表現3204a~3204oがスタック3400内にスタック化される。スタック3400は、利用可能な特徴を、3D畳み込みフィルタの同じ受容フィールド内の空間次元及び時間次元（すなわち、複数の配列決定サイクル）の両方から作成する3D入力体積である。スタック化は、スタック化器3402によって操作可能になっている。他の実施態様では、スタック3400は、任意の次元数（例えば、1D、2D、4D、5Dなど）のテンソルとすることができる。

【0412】

反復モジュール

【0413】

配列決定データにおける長期依存性を捕捉するために、具体的には、クロスサイクルの配列決定画像におけるプレフェージング及びフェージングからの2次の寄与を考慮するために、反復処理を使用する。反復処理は、時間ステップの使用により、連続データの分析に使用される。現在の時間ステップにおける現在の隠れ状態表現は、（i）前の時間ステップからの前の隠れ状態表現、及び（ii）現在の時間ステップにおける現在の入力の関数である。

【0414】

反復モジュール3108は、スタック3400に順方向及び逆方向での3D畳み込みの反復適用（すなわち、反復処理3500）を受けさせ、一連内の t 個の配列決定サイクルのそれぞれにおけるクラスターのそれぞれのためのベースコールを生成する。3D畳み込みは、スライディングウィンドウに基づいてスタック3400内の平坦化された混合表現のサブセットから時空間的な特徴を抽出するために使用される。各スライディングウィンドウ（ w ）は、それぞれの配列決定サイクルに対応しており、図35a中に橙色で強調されている。いくつかの実施態様では、 w は、同時にベースコールされている配列決定サイクルの総数に応じて、1、2、3、5、7、9、15、21などとなるようにパラメータ化される。一実施態様では、 w は、同時にベースコールされている配列決定サイクルの総数の分数である。

【0415】

したがって、例えば、各スライディングウィンドウが、15個の平坦化された混合表現3204a~3204oを含むスタック3400からの3つの継続的な平坦化された混合表現を含むと考える。そのとき、第1のスライディングウィンドウ内の第1の3つの平坦化された混合表現3204a~3204cは、第1の配列決定サイクルに対応し、第2の

10

20

30

40

50

スライディングウィンドウ内の次の3つの平坦化された混合表現 3 2 0 4 b ~ 3 2 0 4 d は、第2の配列決定サイクルに対応し、以下同様である。いくつかの実施態様では、最終の平坦化された混合表現 3 2 0 4 o から開始して、最終の配列決定サイクルに対応する最終のスライディングウィンドウ内に十分な数の平坦化された混合表現を符号化するために、パディングが使用される。

【0416】

各時間ステップにおいて、反復モジュール 3 1 0 8 は、(1) 現在の入力 $x(t)$ 及び (2) 前の隠れ状態表現 $h(t-1)$ を受け取り、現在の隠れ状態表現 $h(t)$ を計算する。現在の入力 $x(t)$ は、現在のスライディングウィンドウ (w)、橙色) 内に入るスタック 3 4 0 0 からの平坦化された混合表現のサブセットのみを含む。したがって、各時間ステップにおける、各現在の入力 $x(t)$ は、複数の平坦化された混合表現 (例えば、 w に応じて、1、2、3、5、7、9、15、又は 21 個の平坦化された混合表現) の 3 D 体積である。例えば、(i) 単一の平坦化された混合表現が次元 15×1905 を有する二次元 (2 D) であり、かつ (i i) w が 7 であるとき、各時間ステップにおける、各現在の入力 $x(t)$ は、次元 $15 \times 1905 \times 7$ の 3 D 体積である。

【0417】

反復モジュール 3 1 0 8 は、第1の 3 D 畳み込み ($W1_{3D CONV}$) を現在の入力 $x(t)$ に適用し、第2の 3 D 畳み込み ($W2_{3D CONV}$) を前の隠れ状態表現 $h(t-1)$ に適用して、現在の隠れ状態表現 $h(t)$ を生成する。いくつかの実施態様では、 $W1_{3D CONV}$ 及び $W2_{3D CONV}$ は、重みが共有されるため同じである。

【0418】

ゲート付き処理

【0419】

一実施態様では、反復モジュール 3 1 0 8 は、長・短期記憶 (LSTM) ネットワーク又はゲート付き回帰型ユニット (GRU) ネットワークなどのゲート付きネットワークによって現在の入力 $x(t)$ 及び前の隠れ状態表現 $h(t-1)$ を処理する。例えば、LSTM の実施態様では、現在の入力 $x(t)$ は、前の隠れ状態表現 $h(t-1)$ と共に、LSTM ユニットの4つのゲート、すなわち、入力ゲート、活性化ゲート、忘却ゲート、及び出力ゲートのそれぞれによって処理される。これは、図 3 5 b に例示されており、この図は、3 D 畳み込みを現在の入力 $x(t)$ 及び前の隠れ状態表現 $h(t-1)$ に適用し、出力として現在の隠れ状態表現 $h(t)$ を生成する LSTM ユニットのによって、現在の入力 $x(t)$ 及び前の隠れ状態表現 $h(t-1)$ を処理する (3 5 0 0 b) 一実施態様を示す。そのような実施態様では、入力、活性化、忘却、及び出力ゲートの重みは、3 D 畳み込みを適用する。

【0420】

いくつかの実施態様では、ゲート付きユニット (LSTM 又は GRU) は、双曲線正接及びシグモイドのような非線形 / スカッシング関数を使用しない。

【0421】

一実施態様では、現在の入力 $x(t)$ 、前の隠れ状態表現 $h(t-1)$ 、及び現在の隠れ状態表現 $h(t)$ は全て、同じ次元数を有する 3 D 体積であり、3 D 体積として入力、活性化、忘却、及び出力ゲートによって処理されるか、又は生成される。

【0422】

一実施態様では、反復モジュール 3 1 0 8 の 3 D 畳み込みは、SAME パディングと共に、サイズ 3×3 の 25 個の畳み込みフィルタのバンクを使用する。いくつかの実施態様では、畳み込みフィルタのサイズは 5×5 である。いくつかの実施態様では、反復モジュール 3 1 0 8 によって使用される畳み込みフィルタの数は、2、4、16、32、64、128、256、512、及び 1024 など、2 のべき乗によって因数分解される。

【0423】

双方向処理

【0424】

反復モジュール 3 1 0 8 は、最初に、スライディングウィンドウに基づいて始まりから終わりまで（トップダウンで）スタック 3 4 0 0 を処理し、順方向トラバーサルのための現在の隠れ状態表現のシーケンス（ベクトル）

【数 1 8】

$$\vec{h}_t = 3DCONV(x_t + \vec{h}_{t-1})$$

を生成する。

【0 4 2 5】

次いで、反復モジュール 3 1 0 8 は、スライディングウィンドウに基づいて終わりから始まりまで（ボトムアップで）スタック 3 4 0 0 を処理し、逆方向／リバーストラバーサルのための現在の隠れ状態表現のシーケンス（ベクトル）

【数 1 9】

$$\overleftarrow{h}_t = 3DCONV(x_t + \overleftarrow{h}_{t-1})$$

を生成する。

【0 4 2 6】

いくつかの実施態様では、両方の方向について、各時間ステップで、処理は、L S T M 又は G R U のゲートを使用する。例えば、各時間ステップにおいて、順方向の現在の入力 $x(t)$ は、L S T M ユニットの入力、活性化、忘却、及び出力ゲートによって処理されて、順方向の現在の隠れ状態表現

【数 2 0】

$$\vec{h}_t$$

を生成し、逆方向の現在の入力 $x(t)$ は、別の L S T M ユニットの入力、活性化、忘却、及び出力ゲートによって処理されて、逆方向の現在の隠れ状態表現

【数 2 1】

$$\overleftarrow{h}_t$$

を生成する。

【0 4 2 7】

次いで、各時間ステップ／スライディングウィンドウ／配列決定サイクルについて、反復モジュール 3 1 0 8 は、対応する順方向及び逆方向の現在の隠れ状態表現を組み合わせ（連結又は総和又は平均化して）、組み合わせられた隠れ状態表現

【数 2 2】

10

20

30

40

50

を生成する。

次いで、組み合わせられた隠れ表現

10

は、1つ又はそれ以上の全結合ネットワークによって処理されて、高密度表現を生成する。次いで、高密度表現は、ソフトマックス層によって処理されて、所与の配列決定サイクルにおけるクラスターのそれぞれに組み込まれた塩基がA、C、T、及びGであることの尤度を生成する。塩基は、尤度に基づいてA、C、T、又はGとして分類される。これは、一連内のt個の配列決定サイクルのそれぞれ（又は各時間ステップ／スライディングウィンドウ）について、並列に又は逐次的にのいずれかで行われる。

20

当業者であれば、他の実施態様において、ハイブリッドアーキテクチャが、より少ない又はより多い数の配列決定サイクルのための入力データを処理することができ、より少ない又はより多い数の畳み込み層及び反復層を含むことができることは理解するであろう。また、入力データの次元数、現在及び前の隠れ表現、畳み込みフィルタ、結果として生じる特徴マップ、並びに出力が異なってもよい。また、畳み込み層内の畳み込みフィルタの数が異なってもよい。それは、異なるパディング及びストライディング構成を使用することができる。それは、異なる分類関数（例えば、シグモイド又は回帰）を使用することができ、全結合層を含んでも含まなくてもよい。それは、1D重畳、2D重畳、3D重畳、4D重畳、5D重畳、拡張又は無性重畳、転置重畳、奥行分離可能な重畳、 1×1 重畳、グループ重畳、扁平重畳、空間及びクロスチャネルの重畳、シャッフルされたグループ化重畳、空間的な分離可能な重畳、及び逆重畳を使用することができる。それは、ロジスティック回帰／対数損失、多クラスクロスエントロピー／ソフトマックス損失、二値クロスエントロピー損失、平均二乗誤差損失、L1損失、L2損失、平滑L1損失、及びHuber損失などの1つ又はそれ以上の損失機能を使用することができる。それは、TFRecord、圧縮符号化（例えば、PNG）、シャープ化、マップ変換に対する平行コール、バッチング、プリフェッチ、モデル並列、データ並列、及び同期／非同期SGDのような、任意の並列、効率、及び圧縮方式を使用することができる。これは、アップサンプリング層、ダウンサンプリング層、反復接続、ゲート及びゲートされたメモリユニット（LSTM又はGRUなど）、残留ブロック、残留接続、ハイウェイ接続、スキップ接続、ペエホル接続、アクティブ化機能（例えば、非線形変換関数は、整流線形ユニット（ReLU）、漏れやすいReLU、指数関数的ライナーユニット（ELU）、シグモイド及び双曲線正接（tanh））、バッチ正規化層、規則化層、ドロップアウト、プール層（例えば、最大又は平均プール）、グローバル平均プール層、及び注意機構のような非線形変換機能を含む。

30

40

図 3 6 は、ニューラルネットワークベースのベースコーラー 2 1 8 を訓練するために使

50

用される訓練データ内のトリヌクレオチド（3量体）を平衡化する一実施態様を示す。平衡化は、訓練データでのゲノムに関する統計の学習を非常に少なくし、ひいては一般化を改善する。ヒートマップ3602は、「A. baumannii」と呼ばれる第1の有機体のための訓練データ内の平衡化された3量体を示す。ヒープマップ（Heap map）3604は、「E. coli」と呼ばれる第2の有機体のための訓練データ内の平衡化された3量体を示す。

【0432】

図37は、RTAベースコーラーのベースコール精度をニューラルネットワークベースのベースコーラー218と比較する。図37に示されるように、RTAベースコーラーは、2回の配列決定実行（リード：1及びリード：2）において、より高い誤り率を有する。すなわち、ニューラルネットワークベースのベースコーラー218は、両方の配列決定実行において、RTAベースコーラーよりも性能が優れている。

10

【0433】

図38は、RTAベースコーラーのタイル間一般化を同じタイル上でのニューラルネットワークベースのベースコーラー218のものと比較する。すなわち、ニューラルネットワークベースのベースコーラー218では、推論（試験）は、訓練に使用されるのと同じタイルのためのデータ上で実行される。

【0434】

図39は、RTAベースコーラーのタイル間一般化を、同じタイル上での、及び異なるタイル上でのニューラルネットワークベースのベースコーラー218のものと比較する。すなわち、ニューラルネットワークベースのベースコーラー218は、第1のタイル上のクラスターのためのデータ上で訓練されるが、推論を第2のタイル上のクラスターからのデータ上で実行する。同じタイルの実施態様では、ニューラルネットワークベースのベースコーラー218は、タイル5上のクラスターからのデータ上で訓練され、タイル5上のクラスターからのデータ上で試験される。異なるタイルの実施態様では、ニューラルネットワークベースのベースコーラー218は、タイル10上のクラスターからのデータ上で訓練され、タイル5上のクラスターからのデータ上で試験される。

20

【0435】

図40もまた、RTAベースコーラーのタイル間一般化を異なるタイル上でのニューラルネットワークベースのベースコーラー218のものと比較する。異なるタイルの実施態様では、ニューラルネットワークベースのベースコーラー218は、タイル10上のクラスターからのデータ上で訓練され、タイル5上のクラスターからのデータ上で試験された後、タイル20上のクラスターからのデータ上で訓練され、タイル5上のクラスターからのデータ上で試験される。

30

【0436】

図41は、ニューラルネットワークベースのベースコーラー218への入力として供給される異なるサイズの画像パッチが、ベースコール精度にどのような変化をもたらすかを示す。両方の配列決定実行（リード：1及びリード：2）において、誤り率は、パッチサイズが3×3から11×11へと増加するにつれて減少している。すなわち、ニューラルネットワークベースのベースコーラー218は、画像パッチが大きいほど、より正確なベースコールを生成する。いくつかの実施態様では、ベースコール精度は、100×100ピクセル以下の画像パッチを使用することによって、計算効率とのバランスが取られる。他の実施態様では、3000×3000ピクセルの大きさの（及びそれより大きい）画像パッチが使用される。

40

【0437】

図42、図43、図44、及び図45は、A. baumannii及びE. coliからの訓練データ上でのニューラルネットワークベースのベースコーラー218のレーン間一般化を示す。

【0438】

図43を参照すると、一実施態様では、ニューラルネットワークベースのベースコーラ

50

ー 2 1 8 は、フローセルの第 1 のレーン上のクラスターからの E . c o l i データ上で訓練され、フローセルの第 1 及び第 2 の両方のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 及び第 2 の両方のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの E . c o l i データ上で訓練され、第 1 及び第 2 の両方のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。また更なる実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 及び第 2 の両方のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。

10

【 0 4 3 9 】

一実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、フローセルの第 1 のレーン上のクラスターからの E . c o l i データ上で訓練され、フローセルの第 1 及び第 2 の両方のレーン上のクラスターからの E . c o l i データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 及び第 2 の両方のレーン上のクラスターからの E . c o l i データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの E . c o l i データ上で訓練され、第 1 のレーン上のクラスターからの E . c o l i データ上で試験される。また更なる実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 及び第 2 の両方のレーン上のクラスターからの E . c o l i データ上で試験される。

20

【 0 4 4 0 】

図 4 3 では、ベースコール精度（誤り率によって測定される）が、これらの実施態様のそれぞれについて、2 回の配列決定実行（例えば、リード：1 及びリード：2）に対して示されている。

【 0 4 4 1 】

図 4 4 を参照すると、一実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、フローセルの第 1 のレーン上のクラスターからの E . c o l i データ上で訓練され、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの E . c o l i データ上で訓練され、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。また更なる実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で試験される。

30

40

【 0 4 4 2 】

一実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、フローセルの第 1 のレーン上のクラスターからの E . c o l i データ上で訓練され、第 1 のレーン上のクラスターからの E . c o l i データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 1 のレーン上のクラスターからの A . b a u m a n n i データ上で訓練され、第 1 のレーン上のクラスターからの E . c o l i データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー 2 1 8 は、第 2 のレーン上のクラスターからの E . c o l i データ上で訓練さ

50

れ、第1のレーン上のクラスターからのE. coli データ上で試験される。また更なる実施態様では、ニューラルネットワークベースのベースコーラー218は、第2のレーン上のクラスターからのA. baumannii データ上で訓練され、第1のレーン上のクラスターからのE. coli データ上で試験される。

【0443】

図44では、ベースコール精度（誤り率によって測定される）が、これらの実施態様のそれぞれについて、2回の配列決定実行（例えば、リード：1及びリード：2）に対して示されている。図43を図44と比較すると、後者によってカバーされる実施態様が50～80パーセントの誤り低減をもたらすことができる。

【0444】

図45を参照すると、一実施態様では、ニューラルネットワークベースのベースコーラー218は、フローセルの第1のレーン上のクラスターからのE. coli データ上で訓練され、第2のレーン上のクラスターからのA. baumannii データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー218は、第1のレーン上のクラスターからのA. baumannii データ上で訓練され、第2のレーン上のクラスターからのA. baumannii データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー218は、第2のレーン上のクラスターからのE. coli データ上で訓練され、第1のレーン上のクラスターからのA. baumannii データ上で試験される。第2の第1のレーン内。また更なる実施態様では、ニューラルネットワークベースのベースコーラー218は、第2のレーン上のクラスターからのA. baumannii データ上で訓練され、第2のレーン上のクラスターからのA. baumannii データ上で試験される。

【0445】

一実施態様では、ニューラルネットワークベースのベースコーラー218は、フローセルの第1のレーン上のクラスターからのE. coli データ上で訓練され、第2のレーン上のクラスターからのE. coli データ上で試験される。別の実施態様では、ニューラルネットワークベースのベースコーラー218は、第1のレーン上のクラスターからのA. baumannii データ上で訓練され、第2のレーン上のクラスターからのE. coli データ上で試験される。更に別の実施態様では、ニューラルネットワークベースのベースコーラー218は、第2のレーン上のクラスターからのE. coli データ上で訓練され、第2のレーン上のクラスターからのE. coli データ上で試験される。また更なる実施態様では、ニューラルネットワークベースのベースコーラー218は、第2のレーン上のクラスターからのA. baumannii データ上で訓練され、第2のレーン上のクラスターからのE. coli データ上で試験される。

【0446】

図45では、ベースコール精度（誤り率によって測定される）が、これらの実施態様のそれぞれについて、2回の配列決定実行（例えば、リード：1及びリード：2）に対して示されている。図43を図45と比較すると、後者によってカバーされる実施態様が50～80パーセントの誤り低減をもたらすことができる。

【0447】

図46は、図42、図43、図44、及び図45に関して上述したレーン間一般化のための誤りプロファイルを示す。一実施態様では、誤りプロファイルは、緑色チャンネル内のA及びT塩基のベースコールにおける誤りを検出する。

【0448】

図47は、図46の誤りプロファイルによって検出された誤りの発生源が緑色チャンネルにおける低いクラスター強度に帰属することを示す。

【0449】

図48は、2回の配列決定実行（リード1及びリード2）についてのRTAベースコーラー及びニューラルネットワークベースのベースコーラー218の誤りプロファイルを比較する。この比較により、ニューラルネットワークベースのベースコーラー218の優れ

10

20

30

40

50

たベースコール精度が確認される。

【0450】

図49aは、4つの異なる器具上でのニューラルネットワークベースのベースコーラー218の実行間一般化を示す。

【0451】

図49bは、同じ器具上で実施された4つの異なる実行上でのニューラルネットワークベースのベースコーラー218の実行間一般化を示す。

【0452】

図50は、ニューラルネットワークベースのベースコーラー218を訓練するために使用される訓練データのゲノム統計を示す。

【0453】

図51は、ニューラルネットワークベースのベースコーラー218を訓練するために使用される訓練データのゲノムコンテキストを示す。

【0454】

図52は、ロングリード（例えば、 2×250 ）のベースコールにおけるニューラルネットワークベースのベースコーラー218のベースコール精度を示す。

【0455】

図53は、ニューラルネットワークベースのベースコーラー218が、画像パッチ間にわたって中央クラスターピクセル及びその隣接ピクセルにどのように着目するかの一実施態様を示す。

【0456】

図54は、一実施態様による、ニューラルネットワークベースのベースコーラー218を訓練及び実行するために使用される様々なハードウェア構成要素及び構成を示す。他の実施態様では、異なるハードウェア構成要素及び構成が使用される。

【0457】

図55は、ニューラルネットワークベースのベースコーラー218を使用して実施することができる様々な配列決定タスクを示す。いくつかの実施例としては、品質スコアリング（Q S c o r i n g）及びバリエーション分類が挙げられる。図55はまた、ニューラルネットワークベースのベースコーラー218がベースコールを実行するいくつかの例示的な配列決定器具も示す。

【0458】

図56は、t分布型確率的近傍埋め込み法（t-SNE）によって可視化された散布図5600であり、ニューラルネットワークベースのベースコーラー218のベースコール結果を表現する。散布図5600は、ベースコール結果が64（ 4^3 ）個のグループにクラスター化され、各グループが主に特定の入力3量体（トリヌクレオチド反復パターン）に対応していることを示す。これは、ニューラルネットワークベースのベースコーラー218が、少なくとも3つの配列決定サイクルのための入力データを処理し、配列特異的なモチーフを学習して、現在のベースコールを前の及び継続的なベースコールに基づいて生成するからである。

【0459】

品質スコアリング

【0460】

品質スコアリングは、各ベースコールに品質スコアを割り当てるプロセスを指す。品質スコアは、配列決定トレースの予測特徴の値を品質テーブルに基づいて確率に変換する、Phredフレームワークに従って定義される。品質テーブルは、校正データセット上で訓練することによって取得され、配列決定プラットフォームの特性が変化するときを更新される。品質スコアの確率的な解釈は、バリエーションコール及び配列アセンブリなどの下流の分析において異なる配列決定リードの適正な統合を可能にする。したがって、品質スコアを定義するための有効なモデルが、いずれのベースコーラーにも不可欠である。

【0461】

10

20

30

40

50

最初に、品質スコアがどのようなものであるかについて説明する。品質スコアは、ベースコールにおける配列決定誤りの確率の尺度である。高い品質スコアは、ベースコールがより信頼でき、不正確である可能性がより低いことを意味する。例えば、ある塩基の品質スコアがQ30である場合、この塩基が不正確にコールされている確率は0.001である。これはまた、ベースコール精度が99.9%であることも示す。

【0462】

以下の表は、ベースコール品質スコアと、それらに対応する誤り確率、ベースコール精度率、及びベースコール誤り率との関係を示す。

【0463】

【表1】

品質スコア	誤り確率	ベースコール誤り率	ベースコール精度率
Q10	0.1 (10に1つ)	10%	90%
Q20	0.01 (100に1つ)	1%	99%
Q30	0.001 (1,000に1つ)	0.1%	99.9%
Q40	0.0001 (10,000に1つ)	0.01%	99.99%
Q50	0.00001 (100,000に1つ)	0.001%	99.999%
Q60	0.000001 (1,000,000に1つ)	0.0001%	99.9999%

【0464】

次に、品質スコアがどのように生成されるかについて説明する。配列決定実行中、品質スコアは、全ての配列決定サイクルのそれぞれについて、全てのタイルのそれぞれの上で、全てのクラスターのそれぞれのための各ベースコールに割り当てられる。Illumina品質スコアが、各ベースコールに対して2ステップのプロセスで計算される。各ベースコールに対して、品質予測子値の数が計算される。品質予測子値は、ベースコールが抽出されるクラスターの観察可能な特性である。これらは、強度プロファイル及び信号対ノイズ比などの特性を含み、ベースコール信頼性の様々な態様を測定する。それらは、ベースコールの品質と相関するように経験的に決定されている。

【0465】

品質モデルは、品質テーブル又はQテーブルとしても知られ、品質予測子値の組み合わせを列挙し、それらに対応する品質スコアに関連付ける。この関係は、経験的データを使用する較正プロセスによって決定される。新たな品質スコアを推定するために、品質予測子値は、新しいベースコールに対して計算され、事前較正された品質テーブル内の値と比較される。

【0466】

次に、品質テーブルがどのように較正されるかについて説明する。較正は、複数の器具上で配列決定された様々な十分に特徴付けられたヒト及び非ヒトサンプルを含む経験的データから統計的な品質テーブルが導出されるプロセスである。Phredアルゴリズムの修正版を使用すると、品質テーブルは、生信号の特性、及びリードを適切な参照に位置合わせすることによって決定された誤り率を使用して開発及び改良される。

【0467】

次に、品質テーブルが時々変化する理由について説明する。品質テーブルは、特定の器具構成及びバージョンのケミストリーによって生成された実行のための品質スコアを提供する。新しいハードウェア、ソフトウェア、又はケミストリーバージョンなど、配列決定プラットフォームの重要な特性が変化するとき、品質モデルは再較正を必要とする。例えば、配列決定ケミストリーの改善は、新たなデータを正確にスコアリングするために品質

テーブルの再校正を必要とし、これは、相当量の処理時間及び計算リソースを消費する。

【0468】

ニューラルネットワークベースの品質スコアリング

【0469】

品質予測子値又は品質テーブルを使用せず、代わりに、十分に較正されたニューラルネットワークの予測に対する信頼性から品質スコアを推論する、品質スコアリングのためのニューラルネットワークベースの技術を開示する。ニューラルネットワークの文脈において、「較正」とは、主観的な予測と経験的な長期にわたる頻度との間の一貫性又は相関を指す。これは、頻度論者の確実性の観念であり、ニューラルネットワークが、時間の90%で、特定のラベルが正しいラベルであると主張する場合、評価中、正しいものとなる確率90%が与えられた全てのラベルの90%は、正しいラベルである。較正は、精度に対して直交性の関係であることに留意されたい。すなわち、ニューラルネットワークの予測は、正確であっても誤較正され得る。

【0470】

開示されるニューラルネットワークは、現実世界の配列決定実行のベースコールドメインを適切にモデル化する多様な配列決定特性を有する大規模な訓練セット上で訓練されるため、十分に較正される。特に、様々な配列決定プラットフォーム、配列決定器具、配列決定プロトコル、配列決定ケミストリー、配列決定試薬、クラスター密度、及びフローセルから得られた配列決定画像が、ニューラルネットワークを訓練するための訓練実施例として使用される。他の実施態様では、異なるベースコール及び品質スコアリングモデルは、異なる配列決定プラットフォーム、配列決定器具、配列決定プロトコル、配列決定ケミストリー、配列決定試薬、クラスター密度、及び／又はフローセルに対してそれぞれ使用される。

【0471】

4つのベースコールクラス（A、C、T、及びG）のそれぞれについて、多数の配列決定画像が、広範な配列決定条件の下で対応のベースコールクラスを表す強度パターンを識別する訓練実施例として使用される。これにより、結果として、ニューラルネットワークの分類能力を訓練中に存在しない新しいクラスにまで拡張する必要性がなくなる。更に、各訓練実施例は、リードを適切な参照に位置合わせすることに基づいて、対応するグラウンドトゥールズで正確にラベル付けされる。結果としてもたらされるのは、予測に対する信頼性を、以下に数学的に表現される、品質スコアリングのための確実性尺度として解釈することができる、十分に較正されたニューラルネットワークである。

【0472】

$Y = \{A, C, T, G\}$ はベースコールクラスA、C、T、及びGのためのクラスラベルのセットを示し、 X は入力空間を示すとする。 $N_{\theta}(y|x)$ は、開示されたニューラルネットワークのうちの1つが入力 $x \in X$ 上で予測する確率分布を示し、 θ はニューラルネットワークのパラメータを示すとする。正しいラベル y_i を有する訓練実施例 x_i に対して、ニューラルネットワークはラベル

【数24】

$$\hat{y}_i = \operatorname{argmax}_{y \in Y} N_{\theta}(y|x_i)$$

を予測する。予測は、

【数25】

$$\hat{y}_i = y_i$$

であれば $c_i = 1$ 、そうでなければ 0 の正確性スコアと、信頼性スコア
【数 2 6】

$$r_i = N_o(\hat{y}_i | x_i)$$

とを取得する。

【0 4 7 3】

ニューラルネットワーク $N_\theta(y | x)$ は、全ての $(x_i, y_i) \in D$ 及び $r_i = \alpha$ にわたって、 $c_i = 1$ である確率が α であるため、データ分布 D に対して十分に較正されている。例えば、 D からのサンプルから、100 個の予測が得られ、それぞれが信頼性 0.8 を有する場合、80 個は、ニューラルネットワーク $N_\theta(y | x)$ によって正確に分類される。より形式的には、 $P_{\theta, D}(r, c)$ は、 D 上でのニューラルネットワーク $N_\theta(y | x)$ の予測の r 及び c 値に対する分布を示し、

【数 2 7】

$$P_{\theta, D}(c=1 | r=I_\alpha) = \alpha \quad \forall \alpha \in [0, 1]$$

として表現され、式中、 I_α は、 α の周囲の小さい非ゼロの間隔を示す。

【0 4 7 4】

十分に較正されたニューラルネットワークは、品質予測子値又は品質テーブルとは異なり、多様な訓練セット上で訓練されるため、それらは、器具構成及びケミストリーバージョンに固有のものではない。これは、2つの利点を有する。第1に、異なる種類の配列決定器具に対して、十分に較正されたニューラルネットワークは、異なる品質テーブルを別個の較正プロセスから導出する必要性をなくす。第2に、同じ配列決定器具に対して、十分に較正されたニューラルネットワークは、配列決定器具の特性が変化するときの再較正の必要性をなくす。以下の詳細がある。

【0 4 7 5】

ソフトマックス信頼確率からの品質スコアの推論

【0 4 7 6】

第1の十分に較正されたニューラルネットワークは、配列決定画像 108 から導出された入力データを処理し、塩基が A、C、T、及び G であることのベースコール信頼確率を生成する、ニューラルネットワークベースのベースコーラー 218 である。ベースコール信頼確率はまた、尤度又は分類スコアと見なすこともできる。一実施態様では、ニューラルネットワークベースのベースコーラー 218 は、ソフトマックス関数を使用して、ベースコール信頼確率をソフトマックススコアとして生成する。

【0 4 7 7】

ソフトマックススコアは、較正され（すなわち、それらはグラウンドトゥルス正確性尤度を表す）、したがって、当然のことながら、品質スコアに対応するため、品質スコアは、ニューラルネットワークベースのベースコーラー 218 のソフトマックス関数によって生成されるベースコール信頼確率から推論される。

【0 4 7 8】

訓練中にニューラルネットワークベースのベースコーラー 218 によって生成されたベースコール信頼確率のセットを選択し、それらのベースコール誤り率（又はベースコール精度率）を決定することによって、ベースコール信頼確率と品質スコアとの間の対応を具体的に説明する。

【0 4 7 9】

それで、例えば、ニューラルネットワークベースのベースコーラー 218 によって生成されたベースコール信頼確率「0.90」を選択する。ニューラルネットワークベースのベースコーラー 218 が 0.90 のソフトマックススコアでベースコール予測を行ったときに多数（例えば、10000～1000000 の範囲）のインスタンスを取得する。多数のインスタンスは、妥当性検査セット又は試験セットのいずれかから取得することができる。次いで、多数のインスタンスのうちの対応のインスタンスに関連付けられた対応するグラウンドトゥールスペースコールとの比較に基づいて、多数のインスタンスのうちのどれだけでベースコール予測が正しかったかを判定する。

【0480】

ベースコールが多数のインスタンスの 90 パーセントで正確に予測され、誤コールは 10 パーセントだったことを観察する。これは、0.90 のソフトマックススコアに対して、ベースコール誤り率が 10 % であり、ベースコール精度率が 90 % であることを意味し、これは、結果として、品質スコア Q10 に対応する（上の表を参照）。同様に、0.99、0.999、0.9999、0.99999、及び 0.999999 のような他のソフトマックススコアに対して、それぞれ、品質スコア Q20、Q30、Q40、Q50、及び Q60 との対応を観察する。これは、図 59a に示されている。他の実施態様では、ソフトマックススコアと、Q9、Q11、Q12、Q23、Q25、Q29、Q37、及び Q39 などの品質スコアとの間の対応を観察する。

【0481】

また、ビニングされた品質スコアとの対応も観察する。例えば、0.80 のソフトマックススコアはビニングされた品質スコア Q06 に対応し、0.95 のソフトマックススコアはビニングされた品質スコア Q15 に対応し、0.993 のソフトマックススコアはビニングされた品質スコア Q22 に対応し、0.997 のソフトマックススコアはビニングされた品質スコア Q27 に対応し、0.9991 のソフトマックススコアはビニングされた品質スコア Q33 に対応し、0.9995 のソフトマックススコアはビニングされた品質スコア Q37 に対応し、0.9999 のソフトマックススコアはビニングされた品質スコア Q40 に対応する。これは、図 59b に示されている。

【0482】

本明細書において使用されるサンプルサイズは、小さなサンプルの問題を回避するために大きくなっており、例えば、10000～1000000 の範囲とすることができる。いくつかの実施態様では、ベースコール誤り率（又はベースコール精度率）を決定するために使用されるインスタンスのサンプルサイズは、評価されているソフトマックススコアに基づいて選択される。例えば、0.99 のソフトマックススコアに対して、サンプルは、100 個のインスタンスを含み、0.999 のソフトマックススコアに対して、サンプルは、1000 個のインスタンスを含み、0.9999 のソフトマックススコアに対して、サンプルは、10000 個のインスタンスを含み、0.99999 のソフトマックススコアに対して、サンプルは、100000 個のインスタンスを含み、0.999999 のソフトマックススコアに対して、サンプルは、1000000 個のインスタンスを含む。

【0483】

ソフトマックスに関しては、ソフトマックスは、多クラス分類のための出力活性化関数である。形式的には、いわゆるソフトマックス分類器を訓練することは、クラス確率への回帰であり、クラスを返すのではなく、各クラスの尤度の信頼性予測を返すので、真の分類器ではない。ソフトマックス関数は、値のクラスを取り、それらを、合計すると 1 になる確率値に変換する。ソフトマックス関数は、任意の実数値の k 次元ベクトルを 0～1 の範囲内の実数値の k 次元ベクトルにスカッシュする。したがって、ソフトマックス関数を使用することは、出力が、有効な指数関数的に正規化された確率質量関数である（非負かつ合計して 1 になる）ことを確実にする。

【0484】

【数 28】

10

20

30

40

50

$$\tilde{y}_i$$

がベクトル

【数 2 9】

$$\tilde{y}=[\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n]$$

10

の i 番目の要素であるとする。

【0 4 8 5】

【数 3 0】

$$\tilde{y}_i = (\text{softmax}(\tilde{z}))_i = \frac{\exp(\tilde{z}_i)}{\sum_{j=1}^N \exp(\tilde{z}_j)}$$

20

【0 4 8 6】

式中

【数 3 1】

$$\tilde{y}$$

30

は長さ n のベクトルであり、 n は分類内のクラスの数である。これらの要素は、 $0 \sim 1$ の値を有し、それらが有効な確率分布を表すように、合計すると 1 になる。

【0 4 8 7】

例示的なソフトマックス活性化関数 5 7 0 6 を図 5 7 に示す。ソフトマックス 5 7 0 6 は、

【数 3 2】

$$z \mapsto \text{softmax}([z, \frac{z}{10}, -2z])$$

40

として 3 つのクラスに適用される。3 つの出力は常に、合計すると 1 になることに留意されたい。したがって、それらは、離散確率質量関数を定義する。

【0 4 8 8】

分類に使用される場合、

【数 3 3】

$$\tilde{y}_i$$

50

は、クラス i 内にある確率を与える。

【数 3 4】

$$P(Y=i|\tilde{z})=(\text{softmax}(\tilde{z}))_i=\tilde{y}_i$$

【0 4 8 9】

「ソフトマックス」という名前は、多少の混乱を与え得る。この関数は、 $\max$ 関数よりも、 $\arg\max$ 関数により密接に関連する。用語「ソフト」は、ソフトマックス関数が連続的かつ微分可能であるという事実に由来する。結果がワンホットベクトルとして表現される、 $\arg\max$ 関数は、連続的でも微分可能でもない。したがって、ソフトマックス関数は、 $\arg\max$ の「ソフト化された」バージョンを提供する。恐らく、ソフトマックス関数を「 softargmax 」と呼ぶことがより適切であるであろうが、現在の名前は、確立した慣習となっている。

【0 4 9 0】

図 5 7 は、品質スコアリングのためにニューラルネットワークベースのベースコーラー 2 1 8 のベースコール信頼確率 3 0 0 4 を選択する (5 7 0 0) 一実施態様を示す。ニューラルネットワークベースのベースコーラー 2 1 8 のベースコール信頼確率 3 0 0 4 は、分類スコア (例えば、ソフトマックススコア又はシグモイドスコア) 又は回帰スコアとすることができる。一実施態様では、ベースコール信頼確率 3 0 0 4 は、訓練 3 0 0 0 中に生成される。

【0 4 9 1】

いくつかの実施態様では、選択 5 7 0 0 は、量子化に基づいて行われ、この量子化は、ベースコール信頼確率 3 0 0 4 にアクセスし、量子化された分類スコア 5 7 0 4 を生成する量子化器 5 7 0 2 によって実行される。量子化された分類スコア 5 7 0 4 は、任意の実数とすることができる。一実施態様では、量子化された分類スコア 5 7 0 4 は、

【数 3 5】

$$0.9 \sum_{i=1}^n 0.1^{(i-1)}$$

として定義される選択式に基づいて選択される。別の実施態様では、量子化された分類スコア 5 7 0 4 は、

【数 3 6】

$$\forall_{i=1}^{n=10} 0.1i$$

として定義される選択式に基づいて選択される。

【0 4 9 2】

図 5 8 は、ニューラルネットワークベースの品質スコアリング 5 8 0 0 の一実施態様を示す。量子化された分類スコア 5 7 0 4 のそれぞれについて、ベースコール誤り率 5 8 0 8 及び／又はベースコール精度率 5 8 1 0 は、そのベースコール予測 3 0 0 4 を対応するグラウンドトゥールスペースコール 3 0 0 8 と比較することによって決定される (例えば、様々なサンプルサイズを有するバッチにわたって)。比較は、比較器 5 8 0 2 によって

実行され、比較器は、結果として、ベースコール誤り率判定器 5804 及びベースコール精度率判定部 5806 を含む。

【0493】

次いで、量子化された分類スコア 5704 と品質スコアとの間の対応を確立するために、適合が、量子化された分類スコア 5704 とそれらのベースコール誤り率 5808（及び／又はそれらのベースコール精度率 5810）との間で適合判定器 5812 によって決定される。一実施態様では、適合判定器 5812 は回帰モデルである。

【0494】

適合に基づいて、品質スコアは、相関器 5814 によって、量子化された分類スコア 5704 と相関させられる。

【0495】

図 59a～図 59b は、品質スコアとニューラルネットワークベースのベースコーラー 218 によって行われたベースコール信頼性予測との間の対応 5900 の一実施態様を示す。ニューラルネットワークベースのベースコーラー 218 のベースコール信頼確率は、分類スコア（例えば、ソフトマックススコア又はシグモイドスコア）又は回帰スコアとすることができる。図 59a は、品質スコアのための品質スコア対応スキーム 5900a である。図 59b は、ビンニングされた品質スコアのための品質スコア対応スキーム 5900a である。

【0496】

推論

【0497】

図 60 は、推論 6000 中にニューラルネットワークベースのベースコーラー 218 によって行われたベースコール信頼性予測から品質スコアを推論する一実施態様を示す。ニューラルネットワークベースのベースコーラー 218 のベースコール信頼確率は、分類スコア（例えば、ソフトマックススコア又はシグモイドスコア）又は回帰スコアとすることができる。

【0498】

推論 6000 の間、予測されたベースコール 6006 には、そのベースコール信頼確率（すなわち、最も高いソフトマックススコア（赤色））が最も対応する品質スコア 6008 が割り当てられる。いくつかの実施態様では、品質スコア対応 5900 は、品質スコア対応スキーム 5900a～5900b をルックアップすることによって作成され、品質スコア推論器 6012 によって操作可能になっている。

【0499】

いくつかの実施態様では、チャステティ（chastity）フィルタ 6010 が、所与のクラスターのベースコールを、そのコールベースに割り当てられた品質スコア 6008、又は継続的なベースコールサイクルにわたる平均品質スコアが、予め設定された閾値を下回るときに終了させる。

【0500】

推論 6000 は、バッティングなどの平行四辺形技術を含む、数百、数千、及び／又は数百万回の前方伝搬 6014 を含む。推論 6000 は、入力データを（配列決定画像 108 から導出された画像チャンネル及び／又は補足的なチャンネル（例えば、距離チャンネル、スクエーリングチャンネル）と共に）含む推論データ 6002 上で実行される。推測 6000 は、テスト 6004 によって動作可能である。

【0501】

ベースコール品質の直接予測

【0502】

第 2 の十分に校正されたニューラルネットワークは、配列決定画像 108 から導出された入力データを処理し、品質指標を直接生成する、ニューラルネットワークベースの品質スコアラー 6102 である。

【0503】

10

20

30

40

50

一実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、多層パーセプトロン（MLP）である。別の実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、フィードフォワードニューラルネットワークである。更に別の実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、全結合型ニューラルネットワークである。更なる実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、完全畳み込みニューラルネットワークである。また更なる実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、セマンティックセグメンテーションニューラルネットワークである。

【0504】

一実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、複数の畳み込み層を有する畳み込みニューラルネットワーク（CNN）である。別の実施態様では、長・短期記憶ネットワーク（LSTM）、双方向LSTM（Bi-LSTM）、又はゲート付き回帰型ユニット（GRU）などの反復ニューラルネットワーク（RNN）である。更に別の実施態様では、CNN及びRNNの両方を含む。

【0505】

更に他の実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、1D重畳、2D重畳、3D重畳、4D重畳、5D重畳、拡張又は無性重畳、転置重畳、奥行分離可能な重畳、点単位の重畳、 1×1 重畳、グループ重畳、扁平重畳、空間及びクロスチャネルの重畳、シャッフルされたグループ化重畳、空間的な分離可能な重畳、並びに逆重畳を使用することができる。それは、ロジスティック回帰／対数損失、多クラスクロスエントロピー／ソフトマックス損失、二値クロスエントロピー損失、平均二乗誤差損失、L1損失、L2損失、平滑L1損失、及びHuber損失などの1つ又はそれ以上の損失機能を使用することができる。それは、TFRecord、圧縮符号化（例えば、PNG）、シャープ化、マップ変換に対する平行コール、バッチング、プリフェッチ、モデル並列、データ並列、及び同期／非同期SGDのような、任意の並列、効率、及び圧縮方式を使用することができる。これは、アップサンプリング層、ダウンサンプリング層、反復接続、ゲート及びゲートされたメモリユニット（LSTM又はGRUなど）、残留ブロック、残留接続、ハイウェイ接続、スキップ接続、ペエホル接続、アクティブ化機能（例えば、非線形変換関数は、整流線形ユニット（ReLU）、漏れやすいReLU、指数関数的ライナーユニット（ELU）、シグモイド及び双曲線正接（tanh））、バッチ正規化層、規則化層、ドロップアウト、プール層（例えば、最大又は平均プール）、グローバル平均プール層、及び注意機構のような非線形変換機能を含む。

【0506】

いくつかの実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、ニューラルネットワークベースのベースコーラー-218と同じアーキテクチャを有する。

【0507】

入力データは、配列決定画像108から導出された画像チャネル及び／又は補足的なチャネル（例えば、距離チャネル、スケーリングチャネル）を含むことができる。ニューラルネットワークベースの品質スコアラ-6102は、入力データを処理し、入力データの代替表現を生成する。代替表現は、いくつかの実施態様では畳み込み表現であり、他の実施態様では隠れ表現である。次いで、代替表現は、出力層によって処理されて、出力を生成する。出力は、品質指標を生成するために使用される。

【0508】

一実施態様では、同じ入力データがニューラルネットワークベースのベースコーラー-218及びニューラルネットワークベースの品質スコアラ-6102に供給されて、（i）ニューラルネットワークベースのベースコーラー-218からのベースコール、及び（ii）ニューラルネットワークベースの品質スコアラ-6102からの対応する品質指標を生成する。いくつかの実施態様では、ニューラルネットワークベースのベースコーラー-218及びニューラルネットワークベースの品質スコアラ-6102は、エンドツーエンドの逆伝播で一緒に訓練される。

【0509】

一実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、特定の配列決定サイクルのために、単一のターゲットクラスターのための品質指標を出力する。別の実施態様では、特定の配列決定サイクルのために、複数のターゲットクラスター内の各ターゲットクラスターのための品質指標を出力する。更に別の実施態様では、複数の配列決定サイクル内の各配列決定サイクルのために、複数のターゲットクラスター内の各ターゲットクラスターのための品質指標を出力し、それによって、各ターゲットクラスターのための品質指標配列を生成する。

【0510】

一実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、配列決定画像108からのデータを含み、ベースコール品質グラウンドトゥールズでラベル付けされた訓練実施例上で訓練された畳み込みニューラルネットワークである。ニューラルネットワークベースの品質スコアラ-6102は、畳み込みニューラルネットワーク6102のベースコール品質予測6104をベースコール品質グラウンドトゥールズ6108と漸進的に一致させる逆伝播ベースの勾配更新技術を使用して訓練される。いくつかの実施態様では、塩基を、それが誤ったベースコールである場合は0として、そうでない場合は1としてラベル付けする。その結果、出力は、誤りの確率に対応する。一実施態様では、これにより、入力特徴として配列コンテキストを使用する必要性がなくなる。

【0511】

畳み込みニューラルネットワーク6102の入力モジュールは、1つ又はそれ以上の配列決定サイクルにおいて捕捉された配列決定画像108からのデータを、1つ又はそれ以上のクラスターのためにコールされた1つ又はそれ以上の塩基の品質を判定するための畳み込みニューラルネットワーク6102に供給する。

【0512】

畳み込みニューラルネットワーク6102の出力モジュールは、畳み込みニューラルネットワーク6102による分析を、1つ又はそれ以上のクラスターのためにコールされた1つ又はそれ以上の塩基の品質を識別する出力6202に変換する。

【0513】

一実施態様では、出力モジュールは、高品質、中品質（任意選択であり、点線で示されている）、及び低品質である品質状態の尤度を生成するソフトマックス分類層を更に含む。別の実施態様では、出力モジュールは、高品質及び低品質である品質状態の尤度を生成するソフトマックス分類層を更に含む。当業者であれば、品質スコアを異なるようにかつ識別可能にバケット化する他のクラスが使用され得ることを理解するであろう。ソフトマックス分類層は、複数の品質スコアが割り当てられている品質について尤度を生成する。これらの尤度に基づいて、品質は、複数の品質スコアのうちの1つから品質スコアが割り当てられる。複数の品質スコアは、ベースコール誤り確率に対数的に基づいている。複数の品質スコアは、Q6、Q10、Q15、Q20、Q22、Q27、Q30、Q33、Q37、Q40、及びQ50を含む。別の実施態様では、出力モジュールは、品質を識別する連続的な値を生成する回帰層を更に含む。

【0514】

いくつかの実施態様では、ニューラルネットワークベースの品質スコアラ-6102は、配列決定画像108からのデータをコールされた塩基のための品質予測子値で補足し、品質予測子値を配列決定画像からのデータと共に畳み込みニューラルネットワーク6102に供給する、補足的な入力モジュールを更に含む。

【0515】

いくつかの実施態様では、品質予測子値としては、オンライン重複、純度、フェージング、開始5（start5）、6量体スコア、モチーフ蓄積、エンジネス（endiness）、近似的なホモポリマー、強度減衰、最後から2番目のチャステティ、背景との信号重複（SOWB）、及び／又はシフトされた純度G調整が挙げられる。他の実施態様では、品質予測子値としては、ピーク高さ、ピーク幅、ピーク位置、相対ピーク位置、ピーク高さ比、ピ

10

20

30

40

50

ーク間隔比、及び／又はピーク対応が挙げられる。品質予測子値に関する更なる詳細は、本明細書に完全に記載されているかのように参照により組み込まれる、米国特許出願公開第2018/0274023号及び同第2012/0020537号に見出すことができる。

【0516】

訓練

【0517】

図61は、配列決定画像108から導出された入力データを処理し、品質指標を直接生成するように、ニューラルネットワークベースの品質スコアラ-6102を訓練する(6100)一実施態様を示す。ニューラルネットワークベースの品質スコアラ-6102は、予測された品質指標6104を正しい品質指標6108と比較し、比較に基づいて誤差6106を計算する、逆伝播ベースの勾配更新技術を使用して訓練される。そのとき、誤差6106は勾配を計算するために使用され、勾配は、後方伝播6110中にニューラルネットワークベースの品質スコアラ-6102の重み及びパラメータに適用される。訓練6100は、アダムなどの確率的勾配更新アルゴリズムを使用して、訓練器1510によって操作可能である。

【0518】

訓練器1510は、訓練データ6112(配列決定画像108に由来する)を使用して、ニューラルネットワークベースの品質スコアラ-6102を、予測された品質指標を生成する前方伝播6116及び誤差6106に基づいて重み及びパラメータを更新する後方伝播6110の数千〜数百万回の反復にわたって訓練する。いくつかの実施態様では、訓練データ6112は、品質予測子値6114で補足される。訓練6100に関する更なる詳細は、「深層学習ツール」と題された付録に見出すことができる。

【0519】

推論

【0520】

図62は、推論6200中にニューラルネットワークベースの品質スコアラ-6102の出力として品質指標を直接生成する一実施態様を示す。推論6200は、バッチングなどの平行四辺形技術を含む、数百、数千、及び／又は数百万回の前方伝播6208を含む。推論6200は、入力データを(配列決定画像108から導出された画像チャネル及び／又は補足的なチャネル(例えば、距離チャネル、スケーリングチャネル)と共に)含む推論データ6204上で実行される。いくつかの実施態様では、推論データ6204は、品質予測子値6206で補足される。推論6200は、テスト6210によって動作可能である。

【0521】

データ前処理

【0522】

いくつかの実施態様では、開示される技術は、画像データ202内のピクセルに適用され、前処理された画像データ202pを生成する前処理技術を使用する。このような実施態様では、画像データ202の代わりに、前処理された画像データ202pが、ニューラルネットワークベースのベースコーラ-218への入力として提供される。データ前処理は、データプリプロセッサ6602によって操作可能になっており、データプリプロセッサは、結果として、データ正規化器6632及びデータ増強器6634を含むことができる。

【0523】

図66は、データ正規化及びデータ増強を含むことができる、データ前処理の異なる実施態様を示す。

【0524】

データ正規化

【0525】

10

20

30

40

50

一実施態様では、データ正規化が、画像データ202内のピクセルに画像パッチごとに適用される。これは、結果として得られる正規化された画像パッチのピクセル強度ヒストグラムが、0の5パーセンタイル及び1の95パーセンタイルを有するように、画像パッチ内のピクセルの強度値を正規化することを含む。すなわち、正規化された画像パッチでは、(i) 5%のピクセルは、0未満の強度値を有し、(ii) 別の5%のピクセルは、1を超える強度値を有する。画像データ202のそれぞれの画像パッチは、別個に正規化することができ、又は画像データ202は、一度に全て正規化することができる。結果としてもたらされるのは、正規化された画像パッチ6616であり、これは、前処理された画像データ202pの一例である。データ正規化は、データ正規化器6632によって操作可能になっている。

10

【0526】

データ増強

【0527】

一実施態様では、データ増強は、画像データ202内のピクセルの強度値に適用される。これは、(i) 画像データ202内の全てのピクセルの強度値に同じスケール係数を乗算することと、(ii) 画像データ202内の全てのピクセルのスケールされた強度値に同じオフセット値を加算することと、を含む。単一のピクセルの場合、これは、以下の式によって表すことができる。

【0528】

$$\text{増強されたピクセル強度 (API)} = aX + b$$

20

式中、 a はスケール係数、 X は元のピクセル強度、 b はオフセット値、 aX はスケールされたピクセル強度である。

【0529】

結果としてもたらされるのは、増強された画像パッチ6626であり、これもまた、前処理された画像データ202pの一例である。データ増強は、データ増強器6634によって操作可能になっている。

【0530】

図67は、ニューラルネットワークベースのベースコーラー218が、細菌データ上で訓練され、ヒトデータ上で試験されるとき（細菌データ及びヒトデータは同じアッセイを共有する（例えば、両方ともイントロンデータを含む））、図66のデータ正規化技術（DeepRTA (norm)）及びデータ増強技術（DeepRTA (augment)）がベースコール誤り率を減少させることを示す。

30

【0531】

図68は、ニューラルネットワークベースのベースコーラー218が、非エクソンデータ（例えば、イントロンデータ）上で訓練され、エクソンデータ上で試験されるとき、図66のデータ正規化技術（DeepRTA (norm)）及びデータ増強技術（DeepRTA (augment)）がベースコール誤り率を減少させることを示す。

【0532】

換言すれば、図66のデータ正規化及びデータ増強技術により、ニューラルネットワークベースのベースコーラー218は、訓練では見られなかったデータ上でより良好に一般化し、ひいては過剰適合を低減することが可能になる。

40

【0533】

一実施態様では、データ増強は、訓練中及び推論中の両方に適用される。別の実施態様では、データ増強は、訓練中にのみ適用される。更に別の実施態様では、データ増強は、推論中にのみ適用される。

【0534】

配列決定システム

【0535】

図63A及び63Bは、配列決定システム6300Aの一実施態様を示す。配列決定システム6300Aは、構成可能プロセッサ6346を含む。構成可能プロセッサ6346

50

は、本明細書に開示されるベースコール技術を実施態様する。配列決定システムは、「シーケンサ」とも称される。

【0536】

配列決定システム6300Aは、生物学的物質又は化学物質のうちの少なくとも1つに関連する任意の情報又はデータを得ることができる。いくつかの実施態様では、配列決定システム6300Aは、ベンチトップデバイス又はデスクトップコンピュータと同様であり得るワークステーションである。例えば、所望の反応を実施するためのシステム及び構成要素の大部分（又は全て）は、共通のハウジング6302内にあってもよい。

【0537】

特定の実施態様では、配列決定システム6300Aは、*de novo sequencing*、全ゲノム又は標的ゲノム領域の再配列、及びメタゲノミクスを含むがこれらに限定されない、様々な用途のために構成された核酸配列決定システムである。シーケンサはまた、DNA又はRNA分析に使用されてもよい。いくつかの実施態様では、配列決定システム6300Aはまた、バイオセンサー内に反応部位を生成するように構成されてもよい。例えば、配列決定システム6300Aは、サンプルを受容し、サンプル由来のクロノウイルス増幅核酸の表面結合クラスターを生成するように構成され得る。各クラスターは、バイオセンサー内の反応部位を構成するか、又はその一部であってもよい。

【0538】

例示的な配列決定システム6300Aは、バイオセンサー6312と相互作用して、バイオセンサー6312内で所望の反応を行うように構成されたシステム容器又はインターフェース6310を含んでもよい。図63Aに関して以下の説明では、バイオセンサー6312はシステム受け部6310内に装填される。しかしながら、バイオセンサー6312を含むカートリッジは、システム受け部6310に挿入されてもよく、一部の状態では、カートリッジは一時的又は永久的に除去され得ることが理解される。上述のように、カートリッジは、とりわけ、流体制御及び流体貯蔵構成要素を含んでもよい。

【0539】

特定の実施態様では、配列決定システム6300Aは、バイオセンサー6312内で多数の平行反応を行うように構成されている。バイオセンサー6312は、所望の反応が生じ得る1つ又はそれ以上の反応部位を含む。反応部位は、例えば、バイオセンサーの固体表面に固定化されてもよく、又はバイオセンサーの対応する反応チャンバ内に位置するビーズ（又は他の可動基材）に固定化されてもよい。反応部位は、例えば、クロノウイルス増幅核酸のクラスターを含むことができる。バイオセンサー6312は、固体撮像装置（例えば、CCD又はCMOSイメージャ）及びそれに取り付けられたフローセルを含んでもよい。フローセルは、配列決定システム6300Aから溶液を受容し、溶液を反応部位に向けて方向付ける1つ又はそれ以上の流路を含んでもよい。任意選択的に、バイオセンサー6312は、熱エネルギーを流路の内外に伝達するための熱要素と係合するように構成することができる。

【0540】

配列決定システム6300Aは、相互に相互作用して、生物学的又は化学的分析のための所定の方法又はアッセイプロトコルを実行する、様々な構成要素、アセンブリ、及びシステム（又はサブシステム）を含んでもよい。例えば、配列決定システム6300Aは、配列決定システム6300Aの様々な構成要素、アセンブリ、及びサブシステムと通信してもよく、またバイオセンサー6312も含むシステムコントローラ6306を含む。例えば、システム容器6310に加えて、配列決定システム6300Aはまた、配列決定システム6300Aの流体ネットワーク及びバイオセンサー6312の流体の流れを制御する流体制御システム6308と、バイオアッセイシステムによって使用され得る全ての流体（例えば、気体又は液体）を保持する流体貯蔵システム6314と、流体ネットワーク、流体貯蔵システム6314及び／又はバイオセンサー6312内の流体の温度を調節し得る温度制御システム6304と、バイオセンサー6312を照明するように構成された照明システム6316と、を備えていてもよい。上述のように、バイオセンサー6312

を有するカートリッジがシステム容器 6310 内に装填される場合、カートリッジはまた、流体制御及び流体貯蔵構成要素を含んでもよい。

【0541】

また、配列決定システム 6300A は、ユーザーと対話するユーザーインターフェース 6318 を含んでもよい。例えば、ユーザーインターフェース 6318 は、ユーザーから情報を表示又は要求するディスプレイ 6320 と、ユーザー入力を受け取るためのユーザー入力デバイス 6322 とを含むことができる。いくつかの実施態様では、ディスプレイ 6320 及びユーザー入力デバイス 6322 は、同じデバイスである。例えば、ユーザーインターフェース 6318 は、個々のタッチの存在を検出し、またディスプレイ上のタッチの場所を識別するように構成されたタッチ感知ディスプレイを含んでもよい。しかしながら、マウス、タッチパッド、キーボード、キーパッド、ハンドヘルドスキャナー、音声認識システム、動き認識システムなどの他のユーザー入力デバイス 6322 が使用されてもよい。以下でより詳細に説明するように、配列決定システム 6300A は、所望の反応を実施するために、バイオセンサー 6312（例えば、カートリッジの形態）を含む様々な構成要素と通信してもよい。配列決定システム 6300A はまた、バイオセンサーから得られたデータを分析して、ユーザーに所望の情報を提供するように構成されてもよい。

【0542】

システムコントローラ 6306 は、マイクロコントローラ、低減命令セットコンピュータ（RISC）、特定用途向け集積回路（ASIC）、フィールドプログラマブルゲートアレイ（FPGA）、粗粒化再構成可能構造（CGRAs）、論理回路、及び本明細書に記載される機能を実行することができる任意の他の回路又はプロセッサと、を備える。上記の実施例は、例示的なものに過ぎず、したがって、システムコントローラという用語の定義及び／又は意味を制限することを意図するものではない。例示の実施態様では、システムコントローラ 6306 は、検出データを取得し分析する少なくとも 1 つのために、1 つ又はそれ以上の記憶要素、メモリ、又はモジュール内に記憶された命令のセットを実行する。検出データは、ピクセル信号の複数のシーケンスを含むことができ、それにより、数百万個のセンサー（又はピクセル）のそれぞれからのピクセル信号のシーケンスを、多くのベースコールサイクルにわたって検出することができる。記憶要素は、配列決定システム 6300A 内の情報源又は物理メモリ要素の形態であってもよい。

【0543】

命令セットは、本明細書に記載される様々な実施態様の方法及びプロセスなどの特定の動作を実行するように配列決定システム 6300A 又はバイオセンサー 6312 に指示する様々なコマンドを含んでもよい。命令のセットは、有形の非一時的コンピュータ可読媒体又は媒体の一部を形成し得るソフトウェアプログラムの形態であってもよい。本明細書で使用するとき、用語「ソフトウェア」及び「ファームウェア」は互換可能であり、RAM メモリ、ROM メモリ、EPROM メモリ、EEPROM メモリ、及び不揮発性 RAM（NVRAM）メモリを含むコンピュータによって実行されるメモリに記憶された任意のコンピュータプログラムを含む。上記メモリタイプは、例示的なものに過ぎず、したがって、コンピュータプログラムの記憶に使用可能なメモリの種類に限定されない。

【0544】

ソフトウェアは、システムソフトウェア又はアプリケーションソフトウェアなどの様々な形態であってもよい。更に、ソフトウェアは、別個のプログラムの集合、又はより大きいプログラム内のプログラムモジュール若しくはプログラムモジュールの一部の形態であってもよい。ソフトウェアはまた、オブジェクト指向プログラミングの形態のモジュール式プログラミングを含んでもよい。検出データを取得した後、検出データは、ユーザー入力に応じて処理された配列決定システム 6300A によって自動的に処理されてもよく、又は別の処理マシン（例えば、通信リンクを介したりリモート要求）によって行われる要求に応じて処理されてもよい。図示の別の実施態様では、システムコントローラ 6306 は分析モジュール 6344 を含む。他の別の実施態様では、システムコントローラ 6306 は分析モジュール 6344 を含まず、代わりに分析モジュール 6344 へのアクセスを有

する（例えば、分析モジュール 6 3 4 4 は、クラウド上で別個にホスティングされ得る）。

【0 5 4 5】

システムコントローラ 6 3 0 6 は、通信リンクを介して、バイオセンサー 6 3 1 2 及び配列決定システム 6 3 0 0 A の他の構成要素に接続されてもよい。システムコントローラ 6 3 0 6 はまた、オフサイトシステム又はサーバに通信可能に接続されてもよい。通信リンクは、配線、コード、又は無線であってもよい。システムコントローラ 6 3 0 6 は、ユーザーインターフェース 6 3 1 8 及びユーザー入力デバイス 6 3 2 2 からユーザー入力又はコマンドを受信してもよい。

【0 5 4 6】

流体制御システム 6 3 0 8 は、流体ネットワークを含み、流体ネットワークを通る 1 つ又はそれ以上の流体の流れを方向付けるように構成されている。流体ネットワークは、バイオセンサー 6 3 1 2 及び流体貯蔵システム 6 3 1 4 と流体連通していてもよい。例えば、流体貯蔵システム 6 3 1 4 から流体を選択し、制御された方法でバイオセンサー 6 3 1 2 に向けてもよく、又は流体は、バイオセンサー 6 3 1 2 から引き出され、例えば、流体貯蔵システム 6 3 1 4 内の廃棄物リザーバに向けられてもよい。図示されていないが、流体制御システム 6 3 0 8 は、流体ネットワーク内の流体の流量又は圧力を検出する流量センサーを含んでもよい。センサーは、システムコントローラ 6 3 0 6 と通信してもよい。

【0 5 4 7】

温度制御システム 6 3 0 4 は、流体ネットワーク、流体貯蔵システム 6 3 1 4 及び／又はバイオセンサー 6 3 1 2 の異なる領域における流体の温度を調節するように構成されている。例えば、温度制御システム 6 3 0 4 は、バイオセンサー 6 3 1 2 と相互作用し、バイオセンサー 6 3 1 2 内の反応部位に沿って流れる流体の温度を制御する熱循環器を含んでもよい。温度制御システム 6 3 0 4 はまた、配列決定システム 6 3 0 0 A 又はバイオセンサー 6 3 1 2 の中実要素又は構成要素の温度を調節してもよい。図示されていないが、温度制御システム 6 3 0 4 は、流体又は他の構成要素の温度を検出するためのセンサーを含んでもよい。センサーは、システムコントローラ 6 3 0 6 と通信してもよい。

【0 5 4 8】

流体貯蔵システム 6 3 1 4 は、バイオセンサー 6 3 1 2 と流体連通しており、所望の反応を行うために使用される様々な反応成分又は反応物質を貯蔵してもよい。流体貯蔵システム 6 3 1 4 はまた、流体ネットワーク及びバイオセンサー 6 3 1 2 を洗浄又は洗淨し、反応物質を希釈するための流体を貯蔵してもよい。例えば、流体貯蔵システム 6 3 1 4 は、試料、試薬、酵素、他の生体分子、緩衝液、水性、及び非極性溶液などを保存するための様々なリザーバを含んでもよい。更に、流体貯蔵システム 6 3 1 4 はまた、バイオセンサー 6 3 1 2 から廃棄物を受容するための廃棄物リザーバを含んでもよい。カートリッジを含む実施態様形態では、カートリッジは、流体貯蔵システム、流体制御システム、又は温度制御システムのうちの 1 つ又はそれ以上を含み得る。したがって、これらのシステムに関する本明細書に記載される構成要素のうちの 1 つ又はそれ以上は、カートリッジハウジング内に収容され得る。例えば、カートリッジは、サンプル、試薬、酵素、他の生体分子、緩衝液、水性、及び非極性溶液、廃棄物などを保存するための様々なリザーバを有し得る。したがって、流体貯蔵システム、流体制御システム、又は温度制御システムのうちの 1 つ又はそれ以上は、カートリッジ又は他のバイオセンサーを介してバイオアッセイシステムと取り外し可能に係合され得る。

【0 5 4 9】

照明システム 6 3 1 6 は、バイオセンサーを照明するための光源（例えば、1 つ又はそれ以上の LED）及び複数の光学構成要素を含んでもよい。光源の例としては、レーザー、アークランプ、LED、又はレーザーダイオードが挙げられる。光学部品は、例えば、反射器、偏光板、ビームスプリッタ、コリマ、レンズ、フィルタ、ウェッジ、プリズム、鏡、検出器などであってもよい。照明システムを使用する実施態様では、照明システム 6 3 1 6 は、励起光を反応部位に向けてるように構成されてもよい。一例として、蛍光団は、緑色の光の波長によって励起されてもよく、そのため、励起光の波長は約 5 3 2 nm であ

10

20

30

40

50

り得る。一実施態様では、照明システム 6316 は、バイオセンサー 6312 の表面の表面法線に平行な照明を生成するように構成されている。別の実施態様では、照明システム 6316 は、バイオセンサー 6312 の表面の表面法線に対してオフアングルである照明を生成するように構成されている。更に別の実施態様では、照明システム 6316 は、いくつかの平行照明及びある程度のオフアングル照明を含む複数の角度を有する照明を生成するように構成されている。

【0550】

システム容器又はインターフェース 6310 は、機械的、電氣的、及び流体的な方法のうちの少なくとも 1 つにおいてバイオセンサー 6312 と係合するように構成される。システム受け部 6310 は、バイオセンサー 6312 を所望の配向に保持して、バイオセンサー 6312 を通る流体の流れを容易にすることができる。システム受け部 6310 はまた、バイオセンサー 6312 と係合するように構成された電気接点を含んでもよく、それにより、配列決定システム 6300A は、バイオセンサー 6312 と通信してもよく、及び／又はバイオセンサー 6312 に電力を供給することができる。更に、システム容器 6310 は、バイオセンサー 6312 と係合するように構成された流体ポート（例えば、ノズル）を含んでもよい。いくつかの実施態様では、バイオセンサー 6312 は、電氣的に、また流体方式で、システム受け部 6310 に取り外し可能に連結される。

【0551】

加えて、配列決定システム 6300A は、他のシステム若しくはネットワークと遠隔で、又は他のバイオアッセイシステム 6300A と通信してもよい。バイオアッセイシステム 6300A によって得られた検出データは、リモートデータベースに記憶されてもよい。

【0552】

図 63B は、図 63A のシステムで使用することができるシステムコントローラ 6306 のブロック図である。一実施態様では、システムコントローラ 6306 は、互いに通信することができる 1 つ又はそれ以上のプロセッサ又はモジュールを含む。プロセッサ又はモジュールのそれぞれは、特定のプロセスを実行するためのアルゴリズム（例えば、有形及び／又は非一時的コンピュータ可読記憶媒体上に記憶された命令）又はサブアルゴリズムを含んでもよい。システムコントローラ 6306 は、モジュールの集合として概念的に例示されるが、専用ハードウェアボード、DSP、プロセッサなどの任意の組み合わせを利用して実施態様されてもよい。あるいは、システムコントローラ 6306 は、単一のプロセッサ又は複数のプロセッサを備えた既製の PC を使用して実施態様されてもよく、機能動作はプロセッサ間に分散される。更なる選択肢として、以下に記載されるモジュールは、特定のモジュール式機能が専用ハードウェアを利用して実施されるハイブリッド構成を利用して実施態様されてもよく、残りのモジュール式機能は、既製の PC などを利用して実施される。モジュールはまた、処理ユニット内のソフトウェアモジュールとして実施態様されてもよい。

【0553】

動作中、通信ポート 6350 は、バイオセンサー 6312（図 63A）及び／又はサブシステム 6308、6314、6304（図 63A）から情報（例えば、データ）に情報（例えば、コマンド）を送信してもよい。実施態様形態では、通信ポート 6350 は、ピクセル信号の複数のシーケンスを出力することができる。通信リンク 6334 は、ユーザーインターフェース 6318 からユーザー入力を受信し（図 63A）、ユーザーインターフェース 6318 にデータ又は情報を送信してもよい。バイオセンサー 6312 又はサブシステム 6308、6314、6304 からのデータは、バイオアッセイセッション中に、システムコントローラ 6306 によってリアルタイムで処理されてもよい。追加的に又は代替的に、データは、バイオアッセイセッション中にシステムメモリ内に一時的に記憶され、リアルタイム又はオフライン操作よりも遅く処理されてもよい。

【0554】

図 63B に示すように、システムコントローラ 6306 は、中心処理装置（CPU）6352 と共に主制御モジュール 6324 と通信する複数のモジュール 6326～6348

を含んでもよい。主制御モジュール6324は、ユーザーインターフェース6318と通信してもよい（図63A）。モジュール6326～6348は、主制御モジュール6324と直接通信するものとして示されているが、モジュール6326～6348はまた、互いに、ユーザーインターフェース6318と、及びバイオセンサー6312と直接通信してもよい。また、モジュール6326～6348は、他のモジュールを介して主制御モジュール6324と通信してもよい。

【0555】

複数のモジュール6326～6348は、サブシステム6308、6314、6304及び6316とそれぞれ通信するシステムモジュール6328～6332、6326を含む。流体制御モジュール6328は、流体ネットワークを通る1つ又はそれ以上の流体の流れを制御するために、流体制御システム6308と通信して、流体ネットワークの弁及び流量センサーを制御してもよい。流体貯蔵モジュール6330は、流体が低い場合、又は廃棄物リザーバが容量又はその近くにあるときにユーザーに通知することができる。流体貯蔵モジュール6330はまた、流体が所望の温度で貯蔵され得るように、温度制御モジュール6332と通信してもよい。照明モジュール6326は、所望の反応（例えば、結合事象）が生じた後など、プロトコル中に指定された時間で反応部位を照明するために、照明システム6316と通信してもよい。いくつかの実施態様では、照明モジュール6326は、照明システム6316と通信して、指定された角度で反応部位を照明することができる。

【0556】

複数のモジュール6326～6348はまた、バイオセンサー6312と通信する装置モジュール6336と、バイオセンサー6312に関連する識別情報を判定する識別モジュール6338とを含んでもよい。装置モジュール6336は、例えば、システム容器6310と通信して、バイオセンサーが配列決定システム6300Aとの電氣的及び流体的接続を確立したことを確認することができる。識別モジュール6338は、バイオセンサー6312を識別する信号を受信してもよい。識別モジュール6338は、バイオセンサー6312の識別情報を使用して、他の情報をユーザーに提供してもよい。例えば、識別モジュール6338は、ロット番号、製造日、又はバイオセンサー6312で実行されることが推奨されるプロトコルを決定し、その後表示してもよい。

【0557】

複数のモジュール6326～6348はまた、バイオセンサー6312から信号データ（例えば、画像データ）を受信及び分析する分析モジュール6344（信号処理モジュール又は信号プロセッサとも呼ばれる）も含む。分析モジュール6344は、検出／画像データを記憶するためのメモリ（例えば、RAM又はフラッシュ）を含む。検出データは、ピクセル信号の複数のシーケンスを含むことができ、それにより、数百万個のセンサー（又はピクセル）のそれぞれからのピクセル信号のシーケンスを、多くのベースコールサイクルにわたって検出することができる。信号データは、その後の分析のために記憶されてもよく、又はユーザーインターフェース6318に送信されて、所望の情報をユーザーに表示することができる。いくつかの実施態様では、信号データは、分析モジュール6344が信号データを受信する前に、固体撮像素子（例えば、CMOS画像センサー）によって処理され得る。

【0558】

分析モジュール6344は、複数の配列決定サイクルのそれぞれにおいて、光検出器から画像データを取得するように構成される。画像データは、光検出器によって検出された発光信号から導出され、ニューラルネットワークベースの品質スコアラー6102及び／又はニューラルネットワークベースのベースコーラー218を介して、複数の配列決定サイクルのそれぞれについて画像データを処理し、複数の配列決定サイクルのそれぞれにおいて分析物のうちの少なくとも一部のためのベースコールを生成する。光検出器は、1つ又はそれ以上のオーバーヘッドカメラ（例えば、バイオセンサー6312上のクラスターの画像を上から撮影するIlluminaのGAIIxのCCDカメラ）の一部であって

もよく、又はバイオセンサー6312自体の一部（例えば、バイオセンサー6312上のクラスターの下にあり、底部からのクラスターの画像を取るIlluminaのiSeqのCMOS画像センサー）であってもよい。

【0559】

光検出器の出力は、それぞれクラスターの強度放出及びそれらの周辺背景を示すシーケンス画像である。シーケンス画像は、配列決定中に配列にヌクレオチドを組み込む結果として生成される強度放出を示す。強度放出は、関連する検体及びそれらの周囲の背景からのものである。シーケンス画像は、メモリ6348に記憶される。

【0560】

プロトコルモジュール6340及び6342は、メイン制御モジュール6324と通信して、所定のアッセイプロトコルを実施する際にサブシステム6308、6314及び6304の動作を制御する。プロトコルモジュール6340及び6342は、所定のプロトコルに従って特定の動作を実行するように配列決定システム6300Aに指示するための命令セットを含み得る。図示のように、プロトコルモジュールは、シーケンスごとの合成プロセスを実行するための様々なコマンドを発行するように構成された、配列合成（SBS）モジュール6340であってもよい。SBSにおいて、核酸テンプレートに沿った核酸プライマーの伸長を監視して、テンプレート中のヌクレオチド配列を決定する。下にある化学プロセスは、（例えば、ポリメラーゼ酵素により触媒される）又はライゲーション（例えば、リガーゼ酵素により触媒される）であり得る。特定のポリマー系SBSの実施態様では、プライマーに付加されるヌクレオチドの順序及び種類の検出を使用してテンプレートの配列を決定することができるように、蛍光標識ヌクレオチドをテンプレート依存様式でプライマー（それによってプライマーを伸長させる）に添加する。例えば、第1のSBSサイクルを開始するために、1つ又はそれ以上の標識されたヌクレオチド、DNAポリメラーゼなどを、核酸テンプレートのアレイを収容するフローセル内に／それを介して送達することができる。核酸テンプレートは、対応する反応部位に位置してもよい。プライマー伸長が、組み込まれる標識ヌクレオチドを、撮像事象を通して検出することができる、これらの反応部位が検出され得る。撮像イベントの間、照明システム6316は、反応部位に励起光を提供することができる。任意に、ヌクレオチドは、ヌクレオチドがプライマーに添加されると、更なるプライマー伸長を終結する可逆終端特性を更に含むことができる。例えば、可逆的ターミネーター部分を有するヌクレオチド類似体をプライマーに添加して、デブロッキング剤が部分を除去するためにデブロッキング剤が送達されるまで続く伸長が生じ得ない。したがって、可逆終端を使用する別の実施態様では、フローセル（検出前又は検出後）にデブロッキング試薬を送達するために、コマンドを与えることができる。1つ又はそれ以上のコマンドは、様々な送達工程間の洗浄（複数可）をもたらすために与えられ得る。次いで、サイクルをn回繰り返してプライマーをn個のヌクレオチドで伸長させることができ、それによって長さnの配列を検出する。例示的な配列決定技術は、例えば、Bentley et al., Nature 456: 53-59 (20063)、国際公開第04/0163497号、米国特許第7,057,026号、国際公開第91/066763号、同第07/123744号、米国特許第7,329,492号、同第7,211,414号、同第7,315,019号、米国特許第7,405,2631号、及び米国特許出願第20063/01470630632号（それぞれ参照により本明細書に組み込まれる）に記載されている。

【0561】

SBSサイクルのヌクレオチド送達工程では、単一の種類のヌクレオチドのいずれかを一度に送達することができる、又は複数の異なるヌクレオチドタイプ（例えば、A、C、T、及びG）を送達することができる。一度に単一の種類のヌクレオチドのみが存在するヌクレオチド送達構成では、異なるヌクレオチドは、個別化された送達に固有の時間的分離に基づいて区別することができるため、異なるヌクレオチドは別個の標識を有する必要はない。したがって、配列決定方法又は装置は、単一の色検出を使用することができる。例えば、励起源は、単一の波長又は単一の波長範囲の励起のみを提供する必要がある。ある

10

20

30

40

50

時点で、送達がフローセル内に存在する複数の異なるヌクレオチドをもたらすヌクレオチド送達構成では、異なるヌクレオチドタイプを組み込む部位は、混合物中のそれぞれのヌクレオチドタイプに結合された異なる蛍光標識に基づいて区別することができる。例えば、4つの異なる蛍光団のうちの1つをそれぞれ有する4つの異なるヌクレオチドを使用することができる。一実施態様では、4つの異なるフルオロフォアは、スペクトルの4つの異なる領域における励起を使用して区別することができる。例えば、4つの異なる励起放射線源を使用することができる。あるいは、4つ未満の異なる励起源を使用することができるが、単一源からの励起放射線の光学的濾過を使用して、フローセルにおいて異なる励起放射線の範囲を生成することができる。

【0562】

いくつかの実施態様では、4つ未満の異なる色を、4つの異なるヌクレオチドを有する混合物中で検出することができる。例えば、ヌクレオチドの対は、同じ波長で検出することができるが、対のうちの1つのメンバーに対する強度の差に基づいて、又は、対の他の部材について検出された信号と比較して明らかなシグナルを出現又は消失させる、対の1つのメンバーへの変化（例えば、化学修飾、光化学修飾、又は物理的改質を行うことを介して）に基づいて区別され得る。4つ未満の色の検出を使用して4つの異なるヌクレオチドを区別するための例示的な装置及び方法が、例えば、米国特許出願第61/5363, 294号及び同第61/619, 63763号に記載されており、これらはその全体が参照により本明細書に組み込まれる。2012年9月21日に出願された米国特許出願第13/624, 200号は、その全体が参照により組み込まれる。

【0563】

複数のプロトコルモジュールはまた、バイオセンサー6312内の製品を増幅するための流体制御システム6308及び温度制御システム6304にコマンドを発行するように構成された試料調製（又は生成）モジュール6342を含んでもよい。例えば、バイオセンサー6312は、配列決定システム6300Aに係合されてもよい。増幅モジュール6342は、バイオセンサー6312内の反応チャンバに必要な増幅成分を送達するために、流体制御システム6308に命令を発行することができる。他の実施態様では、反応部位は、テンプレートDNA及び／又はプライマーなどの増幅のためのいくつかの成分を既に含有していてもよい。増幅成分を反応チャンバに送達した後、増幅モジュール6342は、既知の増幅プロトコルに従って異なる温度段階を通して温度制御システム6304にサイクルするように指示し得る。いくつかの実施態様では、増幅及び／又はヌクレオチドの取り込みは、等温的に実施される。

【0564】

SBSモジュール6340は、クローン性アンプリコンのクラスターがフローセルのチャンネル内の局所領域上に形成されるブリッジPCRを実行するコマンドを発行することができる。ブリッジPCRを介してアンプリコンを生成した後、アンプリコンを「線形化」して、一本鎖テンプレートDNAを作製してもよく、ssDNA及び配列決定プライマーは、関心領域に隣接する普遍配列にハイブリダイズされてもよい。例えば、合成方法による可逆的ターミネーター系配列決定を、上記のように又は以下のように使用することができる。

【0565】

各塩基性コーリング又は配列決定サイクルは、例えば、修飾DNAポリメラーゼ及び4種類のヌクレオチドの混合物を使用することによって達成することができる単一の塩基によってssDNAを延長することができる。異なる種類のヌクレオチドは、固有の蛍光標識を有することができ、各ヌクレオチドは、各サイクルにおいて単一塩基の組み込みのみが生じることを可能にする可逆的ターミネーターを更に有し得る。ssDNA、励起光に単一の塩基を添加した後、反応部位に入射し、蛍光発光を検出することができる。検出後、蛍光標識及びターミネーターは、ssDNAから化学的に切断され得る。別の同様の基本コーリング又は配列決定サイクルは、以下の通りであってもよい。そのような配列決定プロトコルでは、SBSモジュール6340は、バイオセンサー6312を通る試

10

20

30

40

50

薬及び酵素溶液の流れを方向付けるように流体制御システム 6308 に指示することができる。本明細書に記載される装置及び方法と共に利用することができる例示的な可逆性ターミネーターベースの SBS 方法は、米国特許出願公開第 2007/0166705 (A1) 号、米国特許出願公開第 2006/0163639 01 (A1) 号、米国特許第 7,057,026 号、米国特許出願公開第 2006/0240439 (A1) 号、米国特許出願公開第 2006/0263147 14709 (A1) 号、国際公開第 05/065631 4 号、米国特許出願公開第 2005/0147009 00 (A1) 号、国際公開第 06/063B1 99 号及び国際公開第 07/0147025 1 号（それぞれ参照によりその全体が本明細書に組み込まれる）に記載されている。可逆性ターミネーター系 SBS の例示的な試薬は、米国特許第 7,541,444 号、米国特許第 7,057,026 号、同第 7,414,147 16 号、同第 7,427,673 号、同第 7,566,537 号、同第 7,592,435 号、及び国際公開第 07/146335 3663 号に記載されており、これらはそれぞれ参照によりその全体が本明細書に組み込まれる。

10

【0566】

いくつかの実施態様では、増幅及び SBS モジュールは、単一のアッセイプロトコルで動作してもよく、例えば、テンプレート核酸は増幅され、続いて同じカートリッジ内で配列される。

【0567】

配列決定システム 6300A はまた、ユーザーがアッセイプロトコルを再構成することを可能にし得る。例えば、決定システム 6300A は、決定されたプロトコルを修正するために、ユーザーインターフェース 6318 を通じてユーザーにオプションを提供することができる。例えば、バイオセンサー 6312 が増幅のために使用されると判定された場合、配列決定システム 6300A は、アニーリングサイクルの温度を要求し得る。更に、配列決定システム 6300A は、選択されたアッセイプロトコルに対して一般的に許容されないユーザー入力をユーザーが提供した場合に、ユーザーに警告を発行し得る。

20

【0568】

実施態様形態では、バイオセンサー 6312 は、センサー（又はピクセル）のミリオンを含み、それらのそれぞれは、連続するベースコールサイクルにわたって複数のピクセル信号のシーケンスを生成する。分析モジュール 6344 は、センサーのアレイ上のセンサーの行方向及び／又は列方向の位置に従って、ピクセル信号の複数のシーケンスを検出し、それらに対応するセンサー（又はピクセル）に属させる。

30

【0569】

図 63C は、ベースコールセンサー出力などの配列決定システム 6300A からのセンサーデータの分析のためのシステムの簡略ブロック図である。図 63C の例では、システムは構成可能プロセッサ 6346 を含む。構成可能プロセッサ 6346 は、中心処理ユニット (CPU) 6352（すなわち、ホストプロセッサ）によって実行される実行時プログラムと協調して、ベースコーラー（例えば、ニューラルネットワークベースの品質スコアラー 6102 及び／又はニューラルネットワークベースのベースコーラー 218）を実行することができる。配列決定システム 6300A は、バイオセンサー 6312 及びフローセルを含む。フローセルは、遺伝物質のクラスターが、遺伝物質中の塩基を同定するためにクラスター内の反応を引き起こすために使用される一連の分析物フローに曝露される 1 つ又はそれ以上のタイルを含み得る。センサーは、タイルデータを提供するために、フローセルの各タイルにおけるシーケンスの各サイクルの反応を検知する。遺伝的配列決定はデータ集約的操作であり、このデータ集約的操作は、ベースコールセンサーデータを、ベースコール動作中に感知された各遺伝物質群のベースコールのシーケンスに変換する。

40

【0570】

本実施例のシステムは、ベースコール動作を調整するための実行時プログラムを実行する CPU 6352 と、タイルデータのアレイのシーケンスを記憶するメモリ 6348 B と、ベースコール動作によって生成されたベースコールリードと、ベースコール動作で使用する他の情報とを記憶する。また、この図では、システムは、構成ファイル（又はファ

50

イル)、例えば、FPGAビットファイル、並びに構成可能プロセッサ6346を構成及び再構成するために使用されるニューラルネットワークのモデルパラメータなどの構成ファイル(又はファイル)を記憶するメモリ6348Aを含む。配列決定システム6300Aは、構成可能プロセッサを構成するためのプログラムを含むことができ、いくつかの実施形態では、ニューラルネットワークを実行する再構成可能なプロセッサを含み得る。

【0571】

配列決定システム6300Aは、バス6389によって構成可能プロセッサ6346に結合される。バス6389は、PCI-SIG規格(PCI Special Interest Group)によって現在維持及び開発されているPCIe規格(Peripheral Component Interconnect Express)と互換性のあるバス技術などの高スループット技術を使用して実施態様することができる。また、この例では、メモリ6348Aは、バス6393によって構成可能プロセッサ6346に結合される。メモリ6348Aは、構成可能プロセッサ6346を有する回路基板上に配置されたオンボードメモリであってもよい。メモリ6348Aは、ベースコール動作で使用される作業データの構成可能プロセッサ6346による高速アクセスに使用される。バス6393はまた、PCIe規格と互換性のあるバス技術などの高スループット技術を使用して実施態様することもできる。

【0572】

フィールドプログラマブルゲートアレイFPGA、粗いグレー構成可能な再構成可能アレイCGRAs、並びに他の構成可能かつ再構成可能なデバイスを含む構成可能なプロセッサは、コンピュータプログラムを実行する汎用プロセッサを使用して達成され得るよりも、より効率的に又はより高速に様々な機能を実施態様するように構成することができる。構成可能なプロセッサの構成は、時にはビットストリーム又はビットファイルと称される構成ファイルを生成するために機能的な説明を編集することと、構成ファイルをプロセッサ上の構成可能要素に配布することと、を含む。構成ファイルは、データフローパターンを設定するように回路を構成することにより、分散メモリ及び他のオンチップメモリリソースの使用、ルックアップテーブルコンテンツ、構成可能な論理ブロックの動作、及び構成可能な論理ブロックの動作、及び構成可能なアレイの構成可能な相互接続及び他の要素のような構成可能な実行ユニットとを含む。構成ファイルがフィールド内で変更され得る場合、ロードされた構成ファイルを変更することによって構成ファイルを変更することができる場合に再構成可能である。例えば、構成ファイルは、揮発性SRAM要素内に、不揮発性読み書きメモリ素子内に記憶されてもよく、構成可能又は再構成可能なプロセッサ上の構成可能要素のアレイ間に分散されたものであってもよい。様々な市販の構成可能なプロセッサは、本明細書に記載されるようなベースコール動作において使用するのに好適である。例としては、Googleのテンソル処理ユニット(TPU)(商標)、GX4 Rackmount Series(商標)、GX9 Rackmount Series(商標)、NVIDIA DGX-1(商標)、Microsoft' Stratix V FPGA(商標)、GraphcoreのIntelligent Processor Unit(IPU)(商標)、QualcommのZeroth Platform(商標)(Snapdragon processors(商標)、NVIDIA Volta(商標)、NVIDIAのドライブPX(商標)、NVIDIAのJETSON TX1/TX2 MODULE(商標)、Intel's Nirvana™、Movidius VPU(商標)、Fujitsu DPI(商標)、アームDynamicIQ(商標)、IBM TrueNorth(商標)、Lambda GPU Server with Testa V100s(商標)、Xilinx Alveo(商標) U200、Xilinx Alveo(商標) U250、Xilinx Alveo(商標) U280、Intel/Altera Stratix(商標) GX2800、Intel/Altera Stratix(商標) GX2800、及びIntel Stratix(商標) GX10M、が含まれる。いくつかの実施例では、ホストCPUは、構成可能プロセッサと同じ集積回路上に実施態様することができる。

10

20

30

40

50

【0573】

本明細書に記載される実施形態は、構成可能プロセッサ6346を使用して、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218を実装する。構成可能プロセッサ6346の構成ファイルは、高レベルの記述言語HDL又はレジスタ転送レベルRTL言語仕様を使用して実行されるロジック機能を指定することによって実施態様することができる。本明細書は、選択された構成可能プロセッサが構成ファイルを生成するように設計されたリソースを使用してコンパイルすることができる。構成可能なプロセッサではない場合がある特定用途向け集積回路の設計を生成する目的で、同じ又は類似の仕様をコンパイルすることができる。

【0574】

したがって、本明細書に記載される全ての実施形態における構成可能プロセッサ構成可能プロセッサ6346の代替例は、特定用途向けASIC又は専用集積回路又は集積回路のセットを含む構成されたプロセッサを含み、又はシステムオンチップSOCデバイス、又は本明細書に記載されるようなニューラルネットワークベースのベースコール動作を実行するように構成された、システムオンチップSOCデバイス、又はグラフィック処理ユニット(GPU)プロセッサ若しくは粗粒化再構成可能構造(CGRA)プロセッサである。

【0575】

一般に、ニューラルネットワークの実行を実行するように構成された、本明細書に記載の構成可能なプロセッサ及び構成されたプロセッサは、本明細書ではニューラルネットワークプロセッサと称される。

【0576】

構成可能プロセッサ6346は、この例では、CPU6352によって実行されるプログラムを使用して、又は構成可能要素6391のアレイを構成する他のソースによってロードされた構成ファイルによって構成される(例えば、構成論理ブロック(CLB)、例えばルックアップテーブル(LUT)、フリップフロップ、演算処理ユニット(PMU)、及び計算メモリユニット(CMU)、構成可能なI/Oブロック、プログラマブル相互接続)を計算して、ベースコール機能を実行する。この例では、構成は、バス6389及び6393に結合され、ベースコール動作で使用される要素間でデータ及び制御パラメータを分散する機能を実行するデータフローロジック6397を含む。

【0577】

また、構成可能プロセッサ6346は、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218を実行するために、ベースコール実行論理6397を用いて構成される。論理6397は、マルチサイクル実行クラスター(例えば、6379)を含み、この実施例では、実行クラスターXを介した実行クラスター1を含む。多重サイクル実行クラスターの数は、動作の所望のスループットを伴うトレードオフ、及び構成可能プロセッサ6346上の利用可能なリソースに従って選択することができる。

【0578】

多重サイクル実行クラスターは、構成可能なプロセッサ6346上の構成可能な相互接続及びメモリリソースを使用して実施態様されるデータ流路6399によってデータフローロジック6397に結合される。また、多重サイクル実行クラスターは、例えば構成可能プロセッサ6346上の構成可能な相互接続及びメモリリソースを使用して実装される制御経路6395によってもデータフローロジック6397に結合されており、これらの制御経路は、利用可能な実行クラスター、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行を実施するための入力ユニットを利用可能な実行クラスターに提供する準備ができていて、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218のための訓練されたパラメータを提供する準備ができていて、ベースコール分類データの出力パッチを提供する準備ができていて、

10

20

30

40

50

並びにニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行に使用される他の制御データを示す制御信号を提供する。

【0579】

構成可能プロセッサ6346は、訓練されたパラメータを使用してニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行を実行して、ベースコール動作の検知サイクルに関する分類データを生成するように構成される。ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行を実行して、ベースコール動作の被験者検知サイクルの分類データを生成する。ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行は、N個の感知サイクルのそれぞれの検知サイクルからのタイルデータのアレイの数Nを含むシーケンスで動作し、N個の検知サイクルは、本明細書に記載される実施例での時間シーケンスにおける動作ごとの1つの基本位置に対する異なる基本呼び出し動作のためのセンサーデータを提供する。任意選択的に、N個の感知サイクルのうちのいくつかは、実行される特定のニューラルネットワークモデルに従って必要に応じて、シーケンスから出ることができる。数Nは、1を超える任意の数であり得る。本明細書に記載されるいくつかの実施例では、N個の検知サイクルの検知サイクルは、被験者の検知サイクルに先行する少なくとも1つの検知サイクル、及び被験者サイクルの後の少なくとも1回の検知サイクルについての検知サイクルのセットを表す。本明細書では、数Nが5以上の整数である、実施例が記載される。

【0580】

データフローロジック6397は、N個のアレイの空間的に整合されたパッチのタイルデータを含む所与の実行のための入力ユニットを使用して、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行のために、メモリ6348Aから構成可能プロセッサ6346に、タイルデータ及びモデルパラメータの少なくともいくつかの訓練されたパラメータを移動させるように構成される。入力ユニットは、1回のDMA動作におけるダイレクトメモリアクセス動作によって、又は、配備されたニューラルネットワークの実行と協調して、利用可能なタイムスロットの間に移動するより小さいユニット内で移動させることができる。

【0581】

本明細書に記載される感知サイクルのタイルデータは、1つ又はそれ以上の特徴を有するセンサーデータのアレイを含むことができる。例えば、センサーデータは、DNA、RNA、又は他の遺伝物質の遺伝的配列における塩基位置で4塩基のうちの1つを同定するために分析される2つの画像を含むことができる。タイルデータはまた、画像及びセンサーに関するメタデータを含むことができる。例えば、ベースコール動作の実施形態では、タイルデータは、タイル上の遺伝物質群の中心からのセンサーデータのアレイ内の各ピクセルの距離を示す中心情報からの距離などの、クラスターとの画像の位置合わせに関する情報を含むことができる。

【0582】

ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行中に、タイルデータはまた、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行中に生成されるデータも含み得る。ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行中に再計算されるのではなく再計算することができる中間データと称される。例えば、ニューラルネットワークベースの品質スコアラ-6102及び／又はニューラルネットワークベースのベースコーラ-218の実行中に、データフローロジック6397は、タイルデータのアレイの所与のパッチのセンサーデータの代わりに、中間データをメモリ6348Aに書き込むことができる。このような実施形態は、以下により詳細に

記載される。

【0583】

図示されているように、ベースコール動作の検知サイクルからタイルのセンサーデータを含むタイルデータを記憶する実行時プログラムによってアクセス可能なメモリ（例えば、6348A）を含む、ベースコールセンサー出力の分析のためのシステムが説明される。また、システムは、メモリへのアクセスを有する構成可能プロセッサ6346などのニューラルネットワークプロセッサを含む。ニューラルネットワークプロセッサは、訓練されたパラメータを使用してニューラルネットワークの実行を実行して、検知サイクルのための分類データを生成するように構成される。本明細書に記載されるように、ニューラルネットワークの実行は、被験者サイクルを含むN個の感知サイクルのそれぞれの感知サイクルからタイルデータのN個のアレイのシーケンスで動作して、被験者サイクルの分類データを生成する。データフローロジック908は、N個の感知サイクルのそれぞれの感知サイクルからのN個のアレイの空間的に整合されたパッチのデータを含む入力ユニットを使用して、ニューラルネットワークの実行のために、メモリからニューラルネットワークプロセッサにタイルデータ及び訓練されたパラメータを移動させるために提供される。

【0584】

また、ニューラルネットワークプロセッサがメモリへのアクセスを有し、複数の実行クラスターを含み、ニューラルネットワークを実行するように構成された複数の実行クラスター内の実行クラスターを含むシステムも説明される。データフローロジック6397は、メモリへのアクセス、及び複数の実行クラスター内のクラスターを実行して、複数の実行クラスター内の利用可能な実行クラスターにタイルデータの入力ユニットを提供し、入力ユニットは、それぞれの感知サイクルからタイルデータのアレイの空間的に整列されたパッチの数Nを含む、入力ユニットと、被験者検知サイクルを含み、N個の空間的に整合されたパッチをニューラルネットワークに適用して、被験者検知サイクルの空間的に整合されたパッチの分類データの出力パッチを生成させるように、実行クラスターに、ニューラルネットワークにN個の空間的に位置合わせされたパッチを適用させることと、を含み、Nは1より大きい。

【0585】

図64Aは、ホストプロセッサによって実行される実行時プログラムの機能を含む、ベースコール動作の態様を示す簡略図である。この図では、フローセルからの画像センサーの出力は、ライン6400上で画像処理スレッド6401に提供され、画像処理スレッド8001は、個々のタイルのセンサーデータのアレイ内の位置合わせ及び配置、及び画像の再サンプリングなどの画像上のプロセスを実行することができ、フローセル内の各タイルのタイルクラスターマスクを計算するプロセスによって使用することができ、フローセルの対応するタイル上の遺伝子材料のクラスターに対応するセンサーデータのアレイ内のピクセルを識別するプロセスによって使用することができる。画像処理スレッド6401の出力は、CPU内のディスパッチロジック6410に、ライン6402上に提供され、これは、高速バス6403上又は高速バス6405上のデータキャッシュ6404（例えば、SSD記憶装置）に、ベースコール動作の状態に従って、図63Cの構成可能プロセッサ6346などのニューラルネットワークプロセッサハードウェア6420に転送される。処理され、変換された画像は、以前に使用されたサイクルを検知するために、データキャッシュ6404上に記憶され得る。ハードウェア6420は、ニューラルネットワークによって出力された分類データをディスパッチロジック6464に返し、ディスパッチロジック8080は、情報をデータキャッシュ6404に、又はライン6411上でスレッド6402に渡し、分類データを使用してベースコール及び品質スコア計算を実行し、ベースコール読み取りのための標準フォーマットでデータを配置することができる。ベースコール及び品質スコア計算を実行するスレッド6402の出力は、ベースコールリードを集約するスレッド6403に、ライン6412上に提供され、データ圧縮などの他の動作を実行し、結果として得られるベースコール出力を顧客による利用のために指定された宛先に書き込む。

10

20

30

40

50

【0586】

いくつかの実施形態では、ホストは、ニューラルネットワークを支持するハードウェア6420の出力の最終処理を実行する、スレッド（図示せず）を含むことができる。例えば、ハードウェア6420は、マルチクラスターニューラルネットワークの最終層から分類データの出力を提供することができる。ホストプロセッサは、ベースコール及び品質スコアスレッド6402によって使用されるデータを設定するために、分類データを超えて、ソフトマックス関数などの出力起動機能を実行することができる。また、ホストプロセッサは、ハードウェア6420への入力前のタイルデータのバッチ正規化などの入力動作（図示せず）を実行することができる。

【0587】

図64Bは、図63Cの構成などの構成可能プロセッサ6346の構成の簡略図である。図64Bでは、構成可能プロセッサ6346は、複数の高速PCIeインターフェースを有するFPGAを含む。FPGAは、図63Cを参照して説明されるデータフローロジック6397を含むラッパー6490を用いて構成される。ラッパー6490は、CPU通信リンク6477を介してCPU内の実行時プログラムとのインターフェース及び調整を管理し、DRAM通信リンク6497を介してオンボードDRAM6499（例えば、メモリ6348A）との通信を管理する。ラッパー6490内のデータフローロジック6397は、数Nのサイクルのために、オンボードDRAM6499上のタイルデータのアレイをクラスター6485まで横断することによって取得されたパッチデータを提供し、クラスター6485からプロセスデータ6487を取得して、オンボードDRAM6499に配信する。ラッパー6490はまた、タイルデータの入力アレイ、及び分類データの出力パッチの両方について、オンボードDRAM6499とホストメモリとの間のデータの転送を管理する。ラッパーは、ライン6483上のパッチデータを割り当てられたクラスター6485に転送する。ラッパーは、オンボードDRAM6499から取得されたクラスター6485にライン6481の重みやバイアスなどの訓練されたパラメータを提供する。ラッパーは、CPU通信リンク6477を介してホスト上のランタイムプログラムから提供されるか、又はそれに応答して生成されるクラスター6485に、ライン6479上の構成及び制御データを提供する。クラスターはまた、ホストからの制御信号と協働して使用されて、空間的に整列したパッチデータを提供し、クラスター6485のリソースを使用して、パッチデータを介して多重サイクルニューラルネットワークをパッチデータの上で実行するために、ホストからの制御信号と協働して使用されるラッパー6490に、ライン6489上の状態信号を提供することができる。

【0588】

上述のように、タイルデータの複数のパッチのうちの対応するパッチ上で実行するように構成されたラッパー6490によって管理される単一の構成可能なプロセッサ上に複数のクラスターが存在し得る。各クラスターは、本明細書に記載される複数の感知サイクルのタイルデータを使用して、被験者検知サイクルにおけるベースコールの分類データを提供するように構成することができる。

【0589】

システムの例では、フィルタ重み及びバイアスのようなカーネルデータを含むモデルデータをホストCPUから構成可能プロセッサに送信することができ、その結果、モデルは、サイクル数の関数として更新され得る。ベースコール動作は、代表的な例では、数百の感知サイクルの順序で含むことができる。ベースコール動作は、いくつかの実施形態では、ペアリングされた端部読み取りを含むことができる。例えば、モデル訓練されたパラメータは、20サイクルごと（又は他の数のサイクル）ごとに、又は特定のシステム及びニューラルネットワークモデルに実施態様される更新パターンに従って更新されてもよい。いくつかの実施形態では、タイル上の遺伝的クラスター内の所与のストリングのためのシーケンスが、ストリングの第1の端部から（又は上方に）延在する第1の部分と、ストリングの第2の端部から上方（又は下方）に延在する第2の部分とを含む、ペアリングされた端部リードを含むいくつかの実施形態では、訓練されたパラメータは、第1の部分から

10

20

30

40

50

第2の部分への遷移で更新され得る。

【0590】

いくつかの実施例では、タイルのための感知データの複数サイクルの画像データは、CPUから包装材6490に送信され得る。ラッパー6490は、任意選択的に、感知データの一部の前処理及び変換を行い、その情報をオンボードDRAM6499に書き込むことができる。各感知サイクルの入力タイルデータは、タイル当たり4000×3000ピクセル／タイル以上を含むセンサーデータのアレイを含むことができ、2つの特徴はタイルの2つの画像の色を表し、1ピクセル当たり1つ又は2つのバイトを含むセンサーデータのアレイを含むことができる。数Nが、多重サイクルニューラルネットワークの各実施において使用される3回の検知サイクルである実施形態では、多重サイクルニューラルネットワークの各実施のためのタイルデータのアレイは、数当たり数百メガバイトの数で消費することができる。システムのいくつかの実施形態では、タイルデータはまた、タイルごとに1回記憶されたDFCデータのアレイ、又はセンサーデータ及びタイルに関する他のタイプのメタデータも含む。

10

【0591】

動作中、多重サイクルクラスターが利用可能である場合、ラッパーは、パッチをクラスターに割り当てる。ラッパーはタイルの横断面にタイルデータの次のパッチをフェッチし、適切な制御及び構成情報と共に割り当てられたクラスターに送信する。クラスターは、構成可能プロセッサ上の十分なメモリを用いて構成されて、パッチを含むデータのパッチを、定位置に処理されているいくつかのシステム内で複数サイクルから保持するのに十分なメモリを有するように構成することができ、様々な実施形態では、ピンポンバッファ技術又はラスタ走査技術を使用して処理される。

20

【0592】

割り当てられたクラスターが、現在のパッチのニューラルネットワークのその動作を完了し、出力パッチを生成すると、それはラッパーに信号を送る。ラッパーは、割り当てられたクラスターから出力パッチを読み出すか、あるいは割り当てられたクラスターは、データをラッパーにプッシュする。次いで、ラッパーは、DRAM6499内の処理されたタイルのための出力パッチを組み立てる。タイル全体の処理が完了し、データの出力パッチがDRAMに転送されると、ラッパーは、処理された出力アレイを、特定のフォーマットでホスト／CPUに返送する。いくつかの実施形態では、オンボードDRAM6499は、ラッパー6490内のメモリ管理論理によって管理される。ランタイムプログラムは、リアルタイム分析を提供するために連続フローで実行される全てのサイクルについての全てのタイルデータのアレイの分析を完了するために、配列決定動作を制御することができる。

30

【0593】

技術的な改善及び用語

【0594】

ベースコールは、蛍光標識されたタグを分析物と共に組み込む又は取り付けることを含む。検体は、ヌクレオチド又はオリゴヌクレオチドであってよく、タグは、特定のヌクレオチド型（A、C、T、又はG）であってもよい。励起光は、タグを有する検体に向けられ、タグは検出可能な蛍光シグナル又は強度発光を発する。強度発光は、検体に化学的に結合された励起タグによって放出される光子を示す。

40

【0595】

特許請求の範囲を含む本出願全体を通して、「画像、画像データ、又は分析物及びそれらの周囲背景の強度放射を示す画像領域が使用されるとき、それらは、検体に取り付けられたタグの強度放射を指す。当業者であれば、取り付けられたタグの強度放出は、タグが取り付けられている検体の強度放射を表すか、又はそれに相当し、したがって互換的に使用されることを理解するであろう。同様に、検体の特性は、検体に取り付けられたタグ、又は取り付けられたタグからの強度放出の特性を指す。例えば、検体の中心とは、検体に取り付けられたタグによって放出される強度放出の中心を指す。別の実施例では、検体の

50

周囲の背景とは、検体に取り付けられたタグによって放出される強度放射の周囲の背景を指す。

【0596】

特許、特許出願、論文、書籍、木、及びウェブページが挙げられるがこれらに限定されない、本出願において引用された文献及び類似材料は、その全体が参照により明示的に組み込まれる。組み込まれた文献及び類似の材料のうちの1つ又はそれ以上が、定義された用語、用語使用、記載された技術などを含むがこれらに限定されない、本出願とは異なる、又は矛盾する場合には、この出願は制御する。

【0597】

開示される技術は、核酸テンプレート又はその相補体、例えば、DNA若しくはRNAポリヌクレオチド又は他の核酸サンプルなどの核酸サンプルから得ることができる核酸配列情報の品質及び量を改善するためにニューラルネットワークを使用する。したがって、開示される技術の特定の実施は、以前に利用可能な方法と比較して、より高いスループットのポリヌクレオチド配列決定、例えば、より高いDNA又はRNA配列データの収集速度、配列データ収集におけるより高い効率、及び／又はそのような配列データを得る低コストを提供する。

【0598】

開示される技術は、ニューラルネットワークを使用して、固相核酸クラスターの中心を同定し、そのようなクラスターの配列決定中に生成される光信号を解析して、隣接する、隣接する、又は重複するクラスター間で曖昧さなく区別して、単一の離散したソースクラスターに配列決定シグナルを割り当てる。したがって、これら及び関連する実施態様は、高密度クラスターアレイの領域からの、配列データなどの有意義な情報の回収を可能にし、有用な情報は、重複する又は非常に近接して隣間配置された隣接クラスターの影響を混乱させることに起因して、そのような領域から以前に得られなかった場合がある。重複するシグナルの効果（例えば、核酸配列決定において使用されるような）の効果を含む。

【0599】

以下により詳細に記載されるように、特定の実施形態では、本明細書で提供されるように、1つ又は複数の核酸クラスターに固定化された固体支持体を含む組成物が提供される。各クラスターは、同じ配列の複数の固定化された核酸を含み、本明細書で提供されるような検出可能な中心標識を有する識別可能な中心を有し、識別可能な中心は、クラスター内の周囲領域において固定化された核酸と区別可能である。また、識別可能な中心を有するこのようなクラスターを作製及び使用するための方法も本明細書に記載される。

【0600】

本開示の実施態様は、多数の状況での使用が見出され、その利点は、クラスター内の実質的に中心の位置の位置を識別、決定、アノテーション、記録、ないしは別の方法で割り当てる能力から得られる、多くの状況において使用が見出されるであろう。ハイスループット核酸配列決定、光学的又は他のシグナルを個別のソースクラスターに割り当てるための画像解析アルゴリズムの開発、及び固定化された核酸クラスターの中心の認識が望ましい及び有益である他の用途が望ましい。

【0601】

特定の実施態様では、本発明は、核酸配列決定（例えば、「配列決定」）などのハイスループット核酸分析に関連する方法を企図する。例示的なハイスループット核酸解析としては、非限定的に、デノボ配列決定、再配列決定、全ゲノム配列決定、遺伝子発現解析、遺伝子発現モニタリング、エピジェネティクス分析、ゲノムメチル化分析、対立遺伝子特異的プライマー伸長（APSE）、遺伝的多様性プロファイリング、全ゲノム多型発見及び解析、単一ヌクレオチド多型解析、ハイブリダイゼーション系配列決定法などが挙げられる。当業者は、様々な異なる核酸が、本発明の方法及び組成物を使用して分析され得ることを理解するであろう。

【0602】

本発明の実施は核酸配列決定に関連して記載されているが、それらは、異なる時点で取

10

20

30

40

50

得された画像データ、空間的位置、又は他の時間的若しくは物理的観点で取得された画像データが分析される任意の分野において適用可能である。例えば、本明細書に記載される方法及びシステムは、マイクロアレイ、生物学的検体、細胞、生物などからの画像データが取得され、異なる時点又は視点で取得され、分析される、分子生物学及び細胞生物学の分野において有用である。画像は、蛍光顕微鏡法、光学顕微鏡法、共焦点顕微鏡法、光学画像化法、磁気共鳴画像化法、トモグラフィー走査などが挙げられるが、これらに限定されない、技術分野において既知の任意の数の技術を使用して得ることができる。別の例として、本明細書に記載される方法及びシステムは、監視、空中、又は衛星撮像技術などによって取得された画像データが、異なる時点又は視点で取得され、分析される場合に適用することができる。この方法及びシステムは、視野内で取得された画像を分析するのに特に有用であり、この視野内で、観察される検体は、視野内の互いに対して同じ場所に留まる。しかしながら、検体は、別個の画像で異なる特性を有してもよく、例えば、検体は、視野の別々の画像において異なるように見える場合がある。例えば、検体は、異なる画像で検出された所与の検体の色とは異なるように見える場合があり、異なる画像内の所与の分析物のために検出された信号の強度の変化、又は更には、1つの画像中の所与の分析物の信号の外観、及び別の画像内の検体の信号の消失を示し得る。

【0603】

本明細書に記載される例は、学術分析又は商業的分析のための様々な生物学的又は化学的プロセス及びシステムにおいて使用されてもよい。より具体的には、本明細書に記載される例は、指定された反応を示すイベント、特性、品質、又は特性を検出することが望ましい様々なプロセス及びシステムにおいて使用されてもよい。例えば、本明細書に記載される例としては、光検出デバイス、バイオセンサー、及びそれらの構成要素、並びにバイオセンサーと共に動作するバイオアッセイシステムが挙げられる。いくつかの実施例では、装置、バイオセンサー、及びシステムは、フローセルと、実質的に一体型構造と一緒に（取り外し可能に又は固定的に）結合された1つ又はそれ以上の光センサーと、を含み得る。

【0604】

装置、バイオセンサー、及びバイオアッセイシステムは、個別に又は集合的に検出され得る複数の指定された反応を実施するように構成されてもよい。装置、バイオセンサー、及びバイオアッセイシステムは、複数の指定された反応が並行して生じる多数のサイクルを実行するように構成されてもよい。例えば、装置、バイオセンサー、及びバイオアッセイシステムを使用して、酵素操作及び光又は画像検出／捕捉の反復サイクルを通して、DNA特徴の高密度配列を配列することができる。したがって、デバイス、バイオセンサー、及びバイオアッセイシステム（例えば、1つ又はそれ以上のカートリッジを介した）は、試薬又は他の反応成分を反応溶液中に送達する1つ又はそれ以上のマイクロ流体チャネル、バイオセンサー、及びバイオアッセイシステムを含んでもよい。いくつかの実施例では、反応溶液は、約5以下、又は約4以下、又は約3以下のpHを含むなど、実質的に酸性であってもよい。いくつかの他の実施例では、反応溶液は、約8以上、又は約9以上、又は約10以上のpHを含むなど、実質的にアルカリ性／塩基性であってもよい。本明細書で使用するとき、用語「酸性」及びその文法的変異体は、約7未満のpH値を指し、用語「塩基性」、「アルカリ性」及びその文法的変異型は、約7を超えるpH値を指す。

【0605】

いくつかの実施例では、反応部位は、均一又は反復パターンなどの所定の方法で提供又は離間される。いくつかの他の実施例では、反応部位はランダムに分布している。反応部位のそれぞれは、関連する反応部位からの光を検出する1つ又はそれ以上の光ガイド及び1つ又はそれ以上の光センサーと関連付けることができる。いくつかの実施例では、反応部位は、指定された反応を少なくとも部分的に区画化し得る反応凹部又はチャンバ内に位置する。

【0606】

本明細書で使用するとき、「指定された反応」は、対象となる検体などの対象となる化

学物質又は生物学的物質の化学的、電気的、物理的、又は光学的特性（又は品質）のうちの少なくとも1つの変化を含む。特定の実施例では、指定された反応は、例えば、蛍光標識生体分子を対象とする蛍光標識生体分子の組み込みなどの正の結合事象である。より一般的には、指定された反応は、化学変換、化学変化、又は化学的相互作用であってもよい。指定された反応はまた、電気特性の変化であってもよい。特定の実施例では、指定された反応は、検体と蛍光標識された分子を組み込むことを含む。検体はオリゴヌクレオチドであってもよく、蛍光標識分子はヌクレオチドであってもよい。指定された反応は、励起光が標識ヌクレオチドを有するオリゴヌクレオチドに向けられ、蛍光団が検出可能な蛍光シグナルを発するときに、指定された反応が検出され得る。代替例では、検出された蛍光は、化学発光又は生物発光の結果である。指定された反応はまた、例えば、ドナーフルオロフォアをアクセプタ蛍光団に近接させることによって蛍光（又はF o r s t e r）共鳴エネルギー移動（F R E T）を増加させることができ、ドナーとアクセプタ蛍光団とを分離することによってF R E Tを減少させ、消光剤をフルオロフォアから分離することによって蛍光を増加させるか、又は消光剤及び蛍光団を共局在させることによって蛍光を減少させることができる。

【0607】

本明細書で使用するとき、「反応溶液」、「反応成分」又は「反応物質」は、少なくとも1つの指定された反応を得るために使用され得る任意の物質を含む。例えば、潜在的な反応成分としては、例えば、試薬、酵素、サンプル、他の生体分子、及び緩衝液が挙げられる。反応成分は、溶液中の反応部位に送達されてもよく、及び／又は反応部位で固定されてもよい。反応成分は、反応部位で固定化された対象検体などの別の物質と直接又は間接的に相互作用し得る。上記のように、反応溶液は、実質的に酸性であってもよい（すなわち、比較的高い酸性度を含む）（例えば、約5以下のpH、約4以下のpHを含む）、又は約3以下のpH、又は実質的にアルカリ性／塩基性（すなわち、比較的高いアルカリ性／塩基性を含む）（例えば、約8以上のpH、約9以上のpH、又は約10以上のpHを含む）。

【0608】

本明細書で使用するとき、用語「反応部位」は、少なくとも1つの指定された反応が生じ得る局所的領域である。反応部位は、物質がその上に固定され得る反応構造又は基材の支持表面を含んでもよい。例えば、反応部位は、その上に反応成分、例えば、その上に核酸のコロニーを有する反応構造（フローセルのチャンネル内に配置され得る）の表面を含んでもよい。いくつかのこのような実施例では、コロニー中の核酸は同じ配列を有し、例えば、一本鎖又は二本鎖テンプレートのクローンコピーである。しかしながら、いくつかの実施例では、反応部位は、例えば、一本鎖又は二本鎖形態で、単一の核酸分子のみを含有してもよい。

【0609】

複数の反応部位は、反応構造に沿ってランダムに分布してもよく、又は所定の様式で配置されてもよい（例えば、マイクロアレイなどのマトリックス内の並列）。反応部位はまた、指定された反応を区画化するように構成された空間領域又は容積を少なくとも部分的に画定する反応室又は凹部を含むことができる。本明細書で使用するとき、用語「反応チャンバ」又は「反応凹部」は、支持構造体の画定された空間領域（多くの場合、流路と流体連通している）を含む。反応凹部は、周囲環境又は空間領域から少なくとも部分的に分離されてもよい。例えば、複数の反応凹部は、検出表面などの共有された壁によって互いに分離されてもよい。より具体的な例として、反応凹部は、検出表面の内面によって画定された窪み、ウェル、溝、空洞、又は窪みを含むナノセルであってもよく、ナノセルが流路と流体連通することができるように、開口部又は開口部（すなわち、開側面である）を有することができる。

【0610】

いくつかの実施例では、反応構造の反応凹部は、固体がその中に完全に又は部分的に挿入され得るように、固体（半固体を含む）に対してサイズ及び形状を定められる。例えば

、反応凹部は、捕捉ビーズを収容するような大きさ及び形状であってもよい。捕捉ビーズは、クロノウイルス増幅DNA又はその上の他の物質を有してもよい。あるいは、反応凹部は、およその数のビーズ又は固体基材を受容するような大きさ及び形状であってもよい。別の例として、反応凹部は、拡散又はフィルタ流体又は反応凹部に流入し得る溶液を制御するように構成された多孔質ゲル又は物質で充填されてもよい。

【0611】

いくつかの実施例では、光センサー（例えば、フォトダイオード）は、対応する反応部位と関連付けられる。反応部位に関連する光センサーは、関連する反応部位において指定された反応が生じたときに、少なくとも1つの光ガイドを介して、関連する反応部位からの光放射を検出するように構成されている。いくつかの場合では、複数の光センサー（例えば、光検出又はカメラデバイスのいくつかのピクセル）は、単一の反応部位に関連付けられてもよい。他の場合では、単一の光センサー（例えば、単一のピクセル）は、単一の反応部位に、又は反応部位の群と関連付けられてもよい。バイオセンサーの光センサー、反応部位、及び他の特徴は、光の少なくとも一部が反射されることなく光センサーによって直接検出されるように構成されてもよい。

【0612】

本明細書で使用するときに、「生物学的又は化学物質」は、生体分子、対象試料、対象検体、及び他の化学化合物を含む。生物学的物質又は化学物質を使用して、他の化学化合物を検出、同定、若しくは分析するか、又は他の化学化合物を研究又は分析するための仲介として機能してもよい。特定の実施例では、生物学的物質又は化学物質は、生体分子を含む。本明細書で使用するときに、「生体分子」は、バイオポリマー、ヌクレオチド、核酸、ポリヌクレオチド、オリゴヌクレオチド、タンパク質、酵素、ポリペプチド、抗体、抗原、リガンド、受容体、多糖類、炭水化物、ポリリン酸、細胞、組織、生物、若しくはそれらの断片、又は前述の種の類似体若しくは模倣体などの任意の他の生物学的に活性な化学化合物のうちの少なくとも1つを含む。更なる例では、生物学的若しくは化学物質又は生体分子は、酵素又は試薬などの別の反応の生成物、例えば、ピロ配列決定反応においてピロリン酸を検出するために使用される酵素又は試薬などの酵素又は試薬の生成物を検出する。ピロホスフェート検出に有用な酵素及び試薬は、例えば、参照によりその全体が組み込まれる米国特許公開第2005/0244870号に記載されている。

【0613】

生体分子、試料、及び生物学的物質又は化学物質は、天然に存在しても合成であってもよく、反応凹部又は領域内の溶液又は混合物中に懸濁されてもよい。生体分子、試料、及び生物学的物質又は化学物質もまた、固体相又はゲル材料に結合されてもよい。生体分子、試料、及び生物学的物質又は化学物質はまた、医薬組成物を含んでもよい。場合によっては、対象とする生体分子、試料、及び生物学的物質又は化学物質は、標的、プローブ、又は検体と呼ばれる場合がある。

【0614】

本明細書で使用するときに、「バイオセンサー」は、反応部位又は反応部位に近接して生じる指定された反応を検出するように構成された複数の反応部位を有する反応構造を含む装置を含む。バイオセンサーは、固体光検出装置又は「撮像」デバイス（例えば、CCD又はCMOS光検出デバイス）、及び任意選択的に、それに取り付けられたフローセルを含んでもよい。フローセルは、反応部位と流体連通する少なくとも1つの流路を含み得る。1つの特定の例として、バイオセンサーは、生物学的アッセイシステムに流体的かつ電気的に結合するように構成される。バイオアッセイシステムは、所定のプロトコル（例えば、配列番号合成）に従って反応部位に反応溶液を送達し、複数の撮像事象を実施してもよい。例えば、バイオアッセイシステムは、反応溶液を反応部位に沿って流すことができる。反応溶液のうちの少なくとも1つは、同じ又は異なる蛍光標識を有する4種類のヌクレオチドを含んでもよい。ヌクレオチドは、反応部位の対応するオリゴヌクレオチドなどに結合してもよい。次いで、バイオアッセイシステムは、励起光源（例えば、発光ダイオード（LED）などの固体光源）を使用して反応部位を照明することができる。励起光は

10

20

30

40

50

、波長の範囲を含む所定の波長又は波長を有してもよい。入射励起光によって励起された蛍光標識は、光センサーによって検出され得る発光信号（例えば、励起光とは異なる波長又は波長の光、及び潜在的に互いに異なる）を提供することができる。

【0615】

本明細書で使用する時、用語「固定化された」は、生体分子又は生物学的物質又は化学物質に関して使用されるとき、生体分子又は生物学的物質又は化学物質を、光検出デバイス又は反応構造の検出表面などの表面に実質的に付着させることを含む。例えば、生体分子又は生物学的物質又は化学物質は、非共有結合（例えば、静電力、ファンデルワールス、及び疎水性界面の脱水）を含む吸着技術、並びに官能基又はリンカーが生体分子の表面への結合を促進する共有結合技術を含む吸着技術を用いて、反応構造の表面に固定化されてもよい。生体分子又は生物学的物質若しくは化学物質を表面に固定化することは、表面の特性、生体分子又は生物学的物質若しくは化学物質を担持する液体媒体、並びに生体分子又は生物学的物質若しくは化学物質自体の特性に基づいてもよい。いくつかの場合において、表面は、生体分子（又は生物学的物質又は化学物質）を表面に固定化することを容易にするために、表面を官能化（例えば、化学的又は物理的に修飾）してもよい。

10

【0616】

いくつかの実施例では、核酸は、その反応凹部の表面などの反応構造に固定化することができる。特定の実施例では、本明細書に記載される装置、バイオセンサー、バイオアッセイシステム及び方法は、天然ヌクレオチド、及び天然ヌクレオチドと相互作用するように構成された酵素の使用を含んでもよい。天然ヌクレオチドとしては、例えば、リボヌクレオチド又はデオキシリボヌクレオチドが挙げられる。天然ヌクレオチドは、一リン酸、二リン酸、又は三リン酸形態であってよく、アデニン（A）、チミン（T）、ウラシル（U）、グアニン（G）、又はシトシン（C）から選択される塩基を有することができる。しかしながら、上記ヌクレオチドの非天然ヌクレオチド、修飾ヌクレオチド、又は類似体を使用することができることが理解されるであろう。

20

【0617】

上記のように、生体分子又は生物学的物質又は化学物質は、反応構造の反応凹部内の反応部位に固定されてもよい。このような生体分子又は生物学的物質は、干渉嵌め、接着、共有結合、又は捕捉によって、反応凹部内に物理的に保持又は固定化されてもよい。反応凹部内に配置され得る物品又は固体の例としては、ポリマービーズ、ペレット、アガロースゲル、粉末、量子ドット、又は反応チャンバ内で圧縮及び／又は保持され得る他の固体が挙げられる。特定の実施態様では、反応凹部は、DNAオリゴヌクレオチドに共有結合することができるヒドロゲル層でコーティング又は充填されてもよい。特定の実施例では、DNAボールなどの核酸超構造は、例えば、反応凹部の内面に取り付けることによって、又は反応凹部内に液体中に滞留することによって、反応凹部内又は反応凹部に配置することができる。DNAボール又は他の核酸超構造を実施することができ、次いで、反応凹部内又は反応凹部に配置することができる。あるいは、DNAボールは、反応凹部においてその場で合成することができる。反応凹部内に固定された物質は、固体、液体、又は気体状態であり得る。

30

【0618】

本明細書で使用する時、用語「検体」は、相対位置に従って他の点又は領域と区別することができるパターンの点又は領域を意味することを意図する。個々の分析物は、特定の種類の1つ又はそれ以上の分子を含むことができる。例えば、検体は、特定の配列を有する単一の標的核酸分子を含むことができ、又は検体は、同じ配列（及び／又はその相補的配列）を有するいくつかの核酸分子を含むことができる。パターンの異なる検体である異なる分子は、パターン内の検体の場所に従って互いに分化させることができる。例示的な検体としては、基材中のウェル、基材中又は基材上のビーズ（又は他の粒子）、基材からの突出部、基材上の隆起部、基材上のゲル材料のパッド、又は基材内のチャンネルが挙げられる。

40

【0619】

50

検出、特徴付け、又は識別される様々な標的検体のいずれも、本明細書に記載される装置、システム、又は方法で 사용할 ことができる。例示的な検体としては、限定するものではないが、核酸（例えば、DNA、RNA 又はそれらの類似体）、タンパク質、多糖類、細胞、抗体、エピトープ、受容体、リガンド、酵素（例えば、キナーゼ、ホスファターゼ又はポリメラーゼ）、小分子薬物候補、細胞、ウイルス、生物などが挙げられるが、これらに限定されない。

【0620】

用語「検体」、「核酸」、「核酸分子」、及び「ポリヌクレオチド」という用語は、本明細書において互換的に使用される。様々な実施態様では、核酸は、特定の種類の核酸分析のために、本明細書で提供されるようなテンプレート（例えば、核酸テンプレート、又は核酸テンプレートに相補的な核酸相補体）として使用されてもよく、核酸増幅、核酸発現解析、及び／又は核酸配列決定、又はこれらの好適な組み合わせが挙げられるが、これらに限定されない。特定の実施における核酸としては、例えば、3' - 5' ホスホジエステル中のデオキシリボヌクレオチドの直鎖ポリマー、又はデオキシリボ核酸（DNA）、例えば、一本鎖及び二本鎖 DNA、ゲノム DNA、コピー DNA 若しくは相補的 DNA（cDNA）、組み換え DNA、又は任意の形態の合成 DNA 若しくは修飾 DNA が挙げられる。他の実施態様では、核酸としては、例えば、3' - 5' ホスホジエステル中のリボヌクレオチドの直鎖ポリマー、又はリボ核酸（RNA）などの他の結合、例えば、一本鎖及び二本鎖 RNA、メッセンジャー（mRNA）、コピー RNA 又は相補的 RNA（cRNA）、あるいはスプライシングされた mRNA、リボソーム RNA、小核 RNA（snRNA）、microRNA（miRNA）、低干渉 RNA（siRNA）、ピウイ RNA（piRNA）、又は任意の形態の合成若しくは修飾 RNA。本発明の組成物及び方法において使用される核酸は、長さが変化してもよく、無傷又は完全長の分子若しくは断片、又はより大きい核酸分子のより小さい部分であってもよい。特定の実施態様では、核酸は、本明細書の他の箇所に記載されるように、1つ又はそれ以上の検出可能な標識を有してもよい。

【0621】

用語「検体」、「クラスター」、「核酸クラスター」、「核酸コロニー」、及び「DNA クラスター」は互換的に使用され、固体支持体に結合された核酸テンプレート及び／又はその相補体の複数のコピーを指す。典型的には、特定の好ましい実施態様では、核酸クラスターは、それらの 5' 末端を介して固体支持体に結合されたテンプレート核酸及び／又はその相補体の複数のコピーを含む。核酸クラスターを構成する核酸鎖のコピーは、一本鎖又は二本鎖形態であってよい。クラスター内に存在する核酸テンプレートのコピーは、例えば、標識部分の存在に起因して、互いに異なる対応する位置にヌクレオチドを有することができる。対応する位置はまた、異なる化学構造を有するが、ウラシル及びチミンの場合など、類似の Watson-Crick 塩基対形成特性を有するアナログ構造を含むことができる。

【0622】

核酸のコロニーはまた、「核酸クラスター」と呼ばれることもある。核酸コロニーは、本明細書の他の箇所でも更に詳細に記載されるように、クラスター増幅又はブリッジ増幅技術によって任意に作成することができる。標的配列の複数の反復は、ローリングサークル増幅手順を使用して作成された混乱剤などの単一の核酸分子中に存在し得る。

【0623】

本発明の核酸クラスターは、使用される条件に応じて、異なる形状、サイズ、及び密度を有することができる。例えば、クラスターは、実質的に円形、多面、ドーナツ形、又はリング形状の形状を有することができる。核酸クラスターの直径は、約 0.2 μm ~ 約 6 μm 、約 0.3 μm ~ 約 4 μm 、約 0.4 μm ~ 約 3 μm 、約 0.5 μm ~ 約 2 μm 、約 0.75 μm ~ 約 1.5 μm 、又は任意の介在直径であるように設計することができる。特定の実施態様において、核酸クラスターの直径は、約 0.5 μm 、約 1 μm 、約 1.5 μm 、約 2 μm 、約 2.5 μm 、約 3 μm 、約 4 μm 、約 5 μm 、又は約 6 μm である。

核酸クラスターの直径は、クラスターの産生において実施される増幅サイクルの数、核酸テンプレートの長さ、又はクラスターが形成される表面に付着したプライマーの密度を含むが、これらに限定されない多数のパラメータによって影響され得る。核酸クラスターの密度は、典型的には、 $0.1/\text{mm}^2$ 、 $1/\text{mm}^2$ 、 $10/\text{mm}^2$ 、 $100/\text{mm}^2$ 、 $1,000/\text{mm}^2$ 、 $10,000/\text{mm}^2 \sim 100,000/\text{mm}^2$ の範囲であるように設計することができる。本発明は、一部では、より高密度の核酸クラスター、例えば、 $100,000/\text{mm}^2 \sim 1,000,000/\text{mm}^2$ 、及び $1,000,000/\text{mm}^2 \sim 10,000,000/\text{mm}^2$ を更に企図する。

【0624】

本明細書で使用する時、「検体」は、検体又は視野内の対象領域である。マイクロアレイデバイス又は他の分子分析デバイスに関連して使用される場合、分析物は、類似又は同一の分子によって占有される領域を指す。例えば、検体は、増幅オリゴヌクレオチド、又は同じ又は類似の配列を有するポリヌクレオチド又はポリペプチドの任意の他の群であり得る。他の実施態様では、検体は、試料上の物理的領域を占有する任意の要素又は要素群であり得る。例えば、分析物は、ランドのバセル、水の本体などであってもよい。分析物が撮像されると、各検体は、一部の領域を有する。したがって、多くの実施態様では、分析物は、単に1つのピクセルではない。

【0625】

検体間の距離は、任意の数の方法で説明することができる。いくつかの実施態様では、検体間の距離は、1つの分析物の中心から別の分析物の中心まで説明することができる。他の実施態様では、距離は、1つの分析物の縁部から別の分析物の縁部まで、又は各分析物の最も外側の識別可能な点間に記載することができる。分析物の縁部は、チップ上の理論的若しくは実際の物理的境界、又は分析物の境界内のいくつかの点として説明することができる。他の実施態様では、距離は、試料上の固定点、又は試料の画像に関して説明することができる。

【0626】

一般に、分析方法に関して、いくつかの実施態様が本明細書に記載される。自動又は半自動化方法で方法を実行するためのシステムも提供されることが理解されるであろう。したがって、本開示は、ニューラルネットワークベースのテンプレート生成及びベースコールシステムを提供し、システムは、プロセッサと、記憶装置と、画像解析用のプログラムと、を含むことができ、プログラムは、本明細書に記載される方法のうちの1つ又はそれ以上を実行するための命令を含む。したがって、本明細書に記載される方法は、例えば、本明細書に記載されるか又は技術分野において既知の構成要素を有するコンピュータ上で実行することができる。

【0627】

本明細書に記載される方法及びシステムは、様々なオブジェクトのうちのいずれかを分析するのに有用である。特に有用な物体は、固体担体又は付着した検体を有する固相表面である。本明細書に記載される方法及びシステムは、 $x-y$ 平面における分析物の繰り返しパターンを有する物体と共に使用される場合、利点を提供する。一例は、細胞、ウイルス、核酸、タンパク質、抗体、炭水化物、小分子（薬物候補など）、生物学的活性分子、又は他の対象検体の集合を有するマイクロアレイである。

【0628】

核酸及びポリペプチドなどの生物学的分子を有する検体を有するアレイの用途の数が増えてきた。このようなマイクロアレイは、典型的には、デオキシリボ核酸（DNA）又はリボ核酸（RNA）プローブが挙げられる。これらは、ヒト及び他の生物に存在するヌクレオチド配列に特異的である。特定の用途では、例えば、個々のDNA又はRNAプローブをアレイの個々の検体に取り付けることができる。既知のヒト又は生物からのものなどの試験サンプルは、標的核酸（例えば、遺伝子断片、mRNA、又はアンプリコン）が配列中のそれぞれの検体で相補的プローブにハイブリダイズするように、アレイに曝露することができる。プローブは、標的的特異的プロセス（例えば、標的核酸上に存在する標識に

起因して、又は検体においてハイブリダイズした形態で存在するプローブ又は標的の酵素標識に起因して) 標識することができる。次いで、分析物の上の特定の光の周波数を走査して、どの標的核酸が試料中に存在するかを特定することによって検査することができる。

【0629】

生物学的マイクロアレイは、遺伝子配列決定及び類似の用途に使用され得る。一般に、遺伝子配列決定は、DNA又はRNAの断片などの標的核酸の長さのヌクレオチドの順序を決定することを含む。比較的短い配列は、典型的には、各分析物において配列決定され、得られた配列情報は、様々な生物情報科学法において使用されて、フラグメントが由来する多くの広範な長さの遺伝物質の配列を確実に決定するために、様々な生物情報科学法において使用されてもよい。特徴的断片の自動化されたコンピュータベースのアルゴリズムが開発されており、ゲノムマッピング、遺伝子の同定、及びそれらの機能などにおいて、より最近使用されてきた。マイクロアレイは、多数の変異体が存在するため、ゲノム含有量を特徴付けるのに特に有用であり、このことは、個々のプローブ及び標的に対して多くの実験を実施する代替物であるため、ゲノム含有量を特徴付けるのに特に有用である。マイクロアレイは、実用的な方法でこのような調査を実施するための理想的なフォーマットである。

【0630】

技術分野において既知の様々な分析物アレイ(「マイクロアレイ」とも呼ばれる)のいずれも、本明細書に記載される方法又はシステムにおいて使用することができる。典型的なアレイは、それぞれが個々のプローブ又はプローブの集団を有する検体を含む。後者の場合、各検体におけるプローブの集団は、典型的には、単一種のプローブを有する均質である。例えば、核酸配列の場合、各検体は、それぞれ共通の配列を有する複数の核酸分子を有することができる。しかしながら、いくつかの実施態様では、アレイの各分析物における集団は、不均質であり得る。同様に、タンパク質配列は、単一のタンパク質又はタンパク質の集団を有する検体を有することができ、典型的には、同じアミノ酸配列を有するが、必ずしもそうではない。プローブは、例えば、プローブを表面に共有結合することによって、又はプローブと表面との非共有相互作用(複数可)を介して、アレイの表面に取り付けることができる。いくつかの実施態様では、核酸分子などのプローブは、例えば、米国特許出願第13/784,368号及び米国特許出願公開第2011/0059865(A1)号に記載されるように(これらはそれぞれ参照により本明細書に組み込まれる)、ゲル層を介して表面に付着させることができる。

【0631】

例示的なアレイとしては、限定するものではないが、Illumina, Incから入手可能なBeadChipアレイ(San Diego, Calif.)又は、例えば、プローブが、表面上に存在するビーズ(例えば、表面上のウェル内のビーズ)に取り付けられ、それぞれ参照により本明細書に組み込まれる、米国特許第6,266,459号、同第6,355,431号、同第6,770,441号、同第6,859,570号、又は同第7,622,294号、又は国際公開第00/63437号に記載されたものなどの他のものが挙げられる。使用することができる市販のマイクロアレイの更なる例としては、例えば、VLSIPS(商標)(Very Large Scale Immobilized Polymer Synthesis)技術と呼ばれることがある技術に従って合成されたAffymetrix(登録商標)GeneChip(登録商標)マイクロアレイ又は他のマイクロアレイが挙げられる。スポットされたマイクロアレイはまた、本開示のいくつかの実施態様による方法又はシステムにおいて使用することができる。例示的なスポッティングされたマイクロアレイは、Amersham Biosciencesから入手可能なCodeLink(商標)Arrayである。有用な別のマイクロアレイは、Agilent Technologiesから入手可能なSurePrint™ Technologyなどのインクジェット印刷法を使用して製造されるものである。

【0632】

他の有用な配列としては、核酸配列決定用途で使用されるものが挙げられる。例えば、

ゲノム断片（多くの場合、クラスターと呼ばれる）のアンプリコンを有する配列は、それぞれ参照により本明細書に組み込まれる、Bentley et al., Nature 456: 53-59 (2008)、国際公開第04/018497号、国際公開第91/06678号、国際公開第07/123744号、米国特許第7,329,492号、同第7,211,414号、同第7,315,019号、同第7,405,281号、若しくは同第7,057,026号、又は米国特許出願公開第2008/0108082 (A1) 号に記載されているものなどが特に有用である。核酸配列決定に有用な別の種類の配列は、エマルジョンPCR技術から生成される粒子の配列である。実施例は、Dressman et al., Proc. Natl. Acad. Sci. USA 100: 8817-8822 (2003)、国際公開第05/010145号、米国特許出願公開第2005/0130173号又は同第2005/0064460号に記載されており、これらはそれぞれその全体が参照により本明細書に組み込まれる。

【0633】

核酸配列に使用される配列は、多くの場合、核酸分析物のランダムな空間パターンを有する。例えば、Illumina Incから入手可能なHiSeq又はMiSeq配列決定プラットフォーム (San Diego, Calif.) は、核酸配列がランダム播種、続いてブリッジ増幅によって形成されるフローセルを利用する。しかしながら、パターン化された配列は、核酸配列又は他の分析用途にも使用することができる。パターン化されたアレイの例、それらの使用方法及びその使用法は、米国特許出願第13/787,396号、同第13/783,043号、同第13/784,368号、米国特許出願公開第2013/0116153 (A1) 号、及び同第2012/0316086 (A1) 号に記載されており、これらはそれぞれ参照により本明細書に組み込まれる。このようなパターン化された配列の分析物を使用して、単一の核酸テンプレート分子を捕捉して、例えば、ブリッジ増幅を介して、均質なコロニーの後続の形成を行うことができる。このようなパターン化された配列は、核酸配列決定用途に特に有用である。

【0634】

アレイ（又は本明細書の方法又はシステムで使用される他の物体）上の検体のサイズは、特定の用途に適するように選択することができる。例えば、いくつかの実施態様では、アレイの分析物は、単一の核酸分子のみを収容するサイズを有することができる。このサイズ範囲の複数の検体を有する表面は、単一分子分解能で検出するための分子の配列を構築するのに有用である。このサイズ範囲の検体もまた、核酸分子のコロニーをそれぞれ含む検体を有するアレイでの使用にも有用である。したがって、アレイの検体はそれぞれ、約 1mm^2 以下、約 $500\text{ }\mu\text{m}^2$ 以下、約 $100\text{ }\mu\text{m}^2$ 以下、約 $10\text{ }\mu\text{m}^2$ 以下、約 $1\text{ }\mu\text{m}^2$ 以下、約 500 nm^2 以下、又は約 100 nm^2 以下、約 10 nm^2 以下、約 5 nm^2 以下、又は約 1 nm^2 以下の面積を有することができる。代替的に又は追加的に、アレイの検体は、約 1mm^2 以上、約 $500\text{ }\mu\text{m}^2$ 以上、約 $100\text{ }\mu\text{m}^2$ 以上、約 $10\text{ }\mu\text{m}^2$ 以上、約 $1\text{ }\mu\text{m}^2$ 以上、約 500 nm^2 以上、約 100 nm^2 以上、約 10 nm^2 以上、約 5 nm^2 以上、又は約 1 nm^2 以上である。実際に、検体は、上記に例示したものから選択される上限と下限との間の範囲内の大きさを有することができる。表面の検体のいくつかのサイズ範囲が核酸及び核酸のスケールに関して例示されてきたが、これらのサイズ範囲の検体は、核酸を含まない用途に使用できることが理解されるであろう。検体のサイズは、核酸用途に使用されるスケールに必ずしも限定される必要はないことが更に理解されるであろう。

【0635】

分析物のアレイなどの複数の検体を有する物体を含む実施例では、検体は、互いの間の空間で分離されている、別個のものとするることができる。本発明において有用なアレイは、最大で $100\text{ }\mu\text{m}$ 、 $50\text{ }\mu\text{m}$ 、 $10\text{ }\mu\text{m}$ 、 $5\text{ }\mu\text{m}$ 、 $1\text{ }\mu\text{m}$ 、 $0.5\text{ }\mu\text{m}$ 以下の縁部から縁部までの距離によって分離される分析物を有することができる。代替的に又は追加的に、アレイは、少なくとも $0.5\text{ }\mu\text{m}$ 、 $1\text{ }\mu\text{m}$ 、 $5\text{ }\mu\text{m}$ 、 $10\text{ }\mu\text{m}$ 、 $50\text{ }\mu\text{m}$ 、 $100\text{ }\mu\text{m}$ 、又はそれ以上の縁部から縁部までの距離によって分離される分析物を有することができ

る。これらの範囲は、分析物の平均縁部間隔及び縁部間隔、並びに最小又は最大間隔に適用することができる。

【0636】

いくつかの実施態様では、アレイの分析物は、別個である必要はなく、代わりに、隣接する検体は互いに当接することができる。検体が別個であるか否かに関わらず、検体及び／又は検体のピッチの大きさは、アレイが所望の密度を有することができるように変化し得る。例えば、規則的なパターンにおける平均分析物ピッチは、最大で100 μm 、50 μm 、10 μm 、5 μm 、1 μm 、0.5 μm 以下であり得る。代替的に又は追加的に、規則的なパターンにおける平均分析物ピッチは、少なくとも0.5 μm 、1 μm 、5 μm 、10 μm 、50 μm 、100 μm 、又はそれ以上であり得る。これらの範囲は、規則的なパターンの最大ピッチ又は最小ピッチにも適用することができる。例えば、規則的なパターンの最大分析物ピッチは、100 μm 以下、50 μm 以下、10 μm 以下、5 μm 以下、1 μm 以下、0.5 μm 以下とすることができ、かつ／又は規則的なパターンにおける最小分析物ピッチは、少なくとも0.5 μm 、1 μm 、5 μm 、10 μm 、50 μm 、100 μm 、又はそれ以上であり得る。

10

【0637】

アレイ内の検体の密度は、単位面積あたりに存在する検体の数に関しても理解され得る。例えば、アレイに関する検体の平均密度は、少なくとも約 1×10^3 検体/ mm^2 、 1×10^4 検体/ mm^2 、 1×10^5 検体/ mm^2 、 1×10^6 検体/ mm^2 、 1×10^6 検体/ mm^2 、 1×10^7 検体/ mm^2 、 1×10^8 検体/ mm^2 、又は 1×10^9 検体/ mm^2 以上であり得る。あるいは、又はそれに加えて、アレイに関する検体の平均密度は、最大で約 1×10^9 検体/ mm^2 、 1×10^8 検体/ mm^2 、 1×10^7 検体/ mm^2 、 1×10^6 検体/ mm^2 、 1×10^5 検体/ mm^2 、 1×10^4 検体/ mm^2 、又は 1×10^3 検体/ mm^2 以下であり得る。

20

【0638】

上記の範囲は、例えば、分析物のアレイの全て又は一部を含む規則的なパターンの全て又は一部に適用することができる。

【0639】

パターン内の検体は、様々な形状のいずれかを有することができる。例えば、アレイの表面上などの2次元平面で観察される場合、検体は、丸みを帯びた、円形、楕円形、矩形、正方形、対称、非対称、三角形、多角形などに見える場合がある。検体は、例えば六角形又は直線パターンを含む規則的な繰り返しパターンで配置することができる。パターンは、所望のレベルのパッキングを達成するように選択され得る。例えば、円形分析物は、六角形の配置で最適に充填される。当然のことながら、他の包装構成もまた、円形分析物のために使用することができ、逆もまた同様である。

30

【0640】

パターンは、パターンの最小幾何学的単位を形成するサブセット内に存在する検体の数に関して特徴付けることができる。サブセットは、例えば、少なくとも約2、3、4、5、6、10以上の検体を含み得る。分析物のサイズ及び密度に応じて、幾何学的単位は、 1mm^2 、500 μm^2 、100 μm^2 、50 μm^2 、10 μm^2 、1 μm^2 、500 nm^2 、100 nm^2 、50 nm^2 、10 nm^2 以下の面積を占めることができる。代替的に又は追加的に、幾何学的単位は、10 nm^2 、50 nm^2 、100 nm^2 、500 nm^2 、1 μm^2 、10 μm^2 、50 μm^2 、100 μm^2 、500 μm^2 、 1mm^2 以上の面積を占めることができる。形状、サイズ、ピッチなどの幾何学的単位における検体の特性は、アレイ又はパターンの検体に関して、より一般的に本明細書に記載されるものから選択することができる。

40

【0641】

分析物の規則的なパターンを有するアレイは、検体の相対的な場所に対して順序付けられるが、各検体の1つ又はそれ以上の他の特性に対してランダムであってもよい。例えば、核酸配列の場合、核酸検体は、それらの相対的な位置に関して規則的であるが、任意の

50

特定の分析物に存在する核酸種に関する配列の知識に関してランダムであってもよい。より具体的な例として、テンプレート核酸を有する検体の反復パターンを播種し、各検体でテンプレートを増幅して、分析物においてテンプレートのコピーを形成することによって形成される核酸配列（例えば、クラスター増幅又はブリッジ増幅を介して、核酸検体の規則的なパターンを有するが、配列にわたる核酸の配列の分布に関してランダムであろう。したがって、アレイ上の核酸材料の存在の検出は、分析物の繰り返しパターンをもたらすことができるのに対し、配列特異的検出は、アレイにわたる信号の非反復分布をもたらすことができる。

【0642】

本明細書におけるパターン、順序、ランダム性などの説明は、アレイ上の検体などの物体上の検体にも関するだけでなく、画像中の検体にも関連することが理解されるであろう。したがって、パターン、順序、ランダム性などは、限定するものではないが、グラフィカルユーザーインターフェース又は他の出力デバイスなどのコンピュータ可読媒体又はコンピュータ構成要素を含むがこれらに限定されない、画像データを記憶、操作、又は通信するために使用される様々なフォーマットのうちのいずれかに存在することができる。

【0643】

本明細書で使用する時、用語「画像」は、オブジェクトの全て又は一部の表現を意味することを意図する。表現は、光学的に検出された再現であり得る。例えば、蛍光、発光、散乱、又は吸収信号から画像を得ることができる。画像内に存在するオブジェクトの部分は、物体の表面又は他のx y面であり得る。典型的には、画像は2次元表現であるが、場合によっては、画像内の情報は、3つ又はそれ以上の次元から導出することができる。画像は、光学的に検出された信号を含む必要はない。非光信号を代わりに存在させることができる。画像は、本明細書の他の箇所に記載されるもののうちの1つ又はそれ以上などの、コンピュータ可読フォーマット又は媒体に提供することができる。

【0644】

本明細書で使用する時、「画像」は、試料又は他の物体の少なくとも一部分の再現又は表現を指す。いくつかの実施態様では、再現は、例えばカメラ又は他の光学検出器によって生成される光再現である。再現は、非光学的再現、例えば、ナノ細孔分析物のアレイから得られる電気信号の表現、又はイオン感応性CMOS検出器から得られた電気信号の表現であり得る。特定の実施態様では、非光学的再現性は、本明細書に記載される方法又は装置から除外され得る。画像は、例えば、100 µm、50 µm、10 µm、5 µm、1 µm、又は0.5 µm未満離れたものを含む、様々な間隔のいずれかで存在する検体の検体を区別することができる解像度を有することができる。

【0645】

本明細書で使用する時、「取得」、「取得」、及び同様の用語は、画像ファイルを取得するプロセスの任意の部分の指す。いくつかの実施態様では、データ取得は、標本の画像を生成することと、標本内の信号を探すことと、信号の画像を探すか又は生成するように検出デバイスに指示することと、画像ファイルの更なる分析又は変換のための命令、及び画像ファイルの任意の数の変換又は操作のための命令を与えることと、を含むことができる。

【0646】

本明細書で使用する時、用語「テンプレート」は、信号又は検体間の場所又は関係の表現を指す。したがって、いくつかの実施態様では、テンプレートは、検体中の検体に対応する信号の表現を有する物理的グリッドである。いくつかの実施態様では、テンプレートは、チャート、テーブル、テキストファイル、又は分析物に対応する場所を示す他のコンピュータファイルであり得る。本明細書に提示される実施態様では、異なる基準点で捕捉された試料の画像のセットにわたって検体の場所を追跡するためにテンプレートが生成される。例えば、テンプレートは、別の分析物に対する1つの分析物の方向及び／又は距離を記述するx、y座標、又は一連の値であり得る。

【0647】

本明細書で使用する時、用語「標本」は、画像が取り込まれる物体又は物体の領域を指すことができる。例えば、画像が土壌の表面から撮影される実施例では、ランドのパセルは、標本であり得る。生体分子の分析がフローセル内で行われる他の実施態様では、フローセルは、任意の数のサブディビジョンに分割されてもよく、これらのそれぞれは検体であってもよい。例えば、フローセルは、様々な流路又はレーンに分割されてもよく、各レーンは、画像化される2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、110、120、140、160、180、200、400、600、800、1000個以上の別個の領域に更に分割され得る。フローセルの一例は8つのレーンを有し、各レーンは120個の標本又はタイルに分割されている。別の実施態様では、試料は、複数のタイル、又は更にはフローセル全体で作製されてもよい。したがって、各検体の画像は、撮像されるより大きい表面の領域を表すことができる。

10

【0648】

本明細書に記載される範囲及び連続数リストへの言及は、列挙された数だけではなく、列挙された数字間の全ての実数を含むことが理解されるであろう。

【0649】

本明細書で使用する時、「参照点」は、画像間の任意の時間的又は物理的区別を指す。好ましい別の実施態様では、基準点は時間点である。より好ましい実施態様では、参照点は、配列決定反応中の時点又はサイクルである。しかしながら、用語「基準点」は、画像を区別又は分離することができる、角度、回転、時間、又は他の態様などの画像を区別又は分離する他の態様を含むことができる。

20

【0650】

本明細書で使用する時、「画像のサブセット」は、セット内の画像のグループを指す。例えば、サブセットは、画像のセットから選択される1、2、3、4、6、8、10、12、14、16、18、20、30、40、50、60又は任意の数の画像を含んでもよい。特定の別の実施態様では、サブセットは、1、2、3、4、6、8、10、12、14、16、18、20、30、40、50、60以下、又は画像のセットから選択される任意の数の画像を含んでもよい。好ましい別の実施態様では、画像は、各サイクルに相関する4つの画像を有する1つ又はそれ以上の配列決定サイクルから得られる。したがって、例えば、サブセットは、4サイクルにわたって取得された16画像のグループであり得る。

30

【0651】

塩基は、ヌクレオチド塩基又はヌクレオチド、A（アデニン）、C（シトシン）、T（チミン）、又はG（グアニン）を指す。本出願は、「塩基（複数可）」及び「ヌクレオチド（複数可）」を互換的に使用する。

【0652】

用語「染色体」は、DNA及びタンパク質成分（特にヒストン）を含むクロマチンストランドに由来する、生きている細胞の本発明の有効性を有する遺伝子キャリアを指す。本明細書では、従来の国際的に認識されている個々のヒトゲノム染色体番号付けシステムが本明細書で使用される。

40

【0653】

「部位」という用語は、参照ゲノム上の固有の位置（例えば、染色体ID、染色体位置及び配向）を指す。いくつかの実施態様では、部位は、残基、配列タグ、又は配列上のセグメントの位置であってもよい。用語「遺伝子座」は、参照染色体上の核酸配列又は多型の特定の位置を指すために使用されてもよい。

【0654】

本明細書における用語「試料」は、典型的には、配列決定及び／又はフェーズドされる核酸を含有する生物液、細胞、組織、器官、又は生物に由来するサンプル、又は配列決定及び／又はフェーズドされる核酸配列を少なくとも1つ含有する核酸の混合物に由来するサンプルを指す。このような試料としては、痰／口腔流体、羊水、血液、血液画分、細針

50

生検試料（例えば、外科生検、針生検など）、尿、腹膜流体、胸膜流体、組織外植片、臓器培養物、及びこれらの任意の他の組織若しくは細胞調製物、又はこれらの画分若しくは誘導体が挙げられるが、これらに限定されない。サンプルは、多くの場合、ヒト被験者（例えば、患者）から採取されるが、試料は、イヌ、ネコ、ウマ、ヤギ、ヒツジ、ウシ、ブタなどを含むがこれらに限定されない、染色体を有する任意の生物から採取することができる。試料は、生物学的源から得られるように、又は試料の特性を修正する前処理後に、直接使用することができる。例えば、このような前処理は、血漿を血液から調製すること、粘性流体を希釈することなどを含んでもよい。前処理の方法には、濾過、沈殿、希釈、蒸留、混合、遠心分離、凍結、凍結乾燥、濃縮、増幅、核酸断片化、干渉成分の不活性化、試薬の添加、溶解などを含んでもよいが、これらに限定されない。

10

【0655】

用語「配列」は、互いに結合されたヌクレオチドの鎖を含むか、又はそれを表す。ヌクレオチドは、DNA又はRNAに基づくことができる。1つの配列は、複数のサブシーケンスを含んでもよいことを理解されたい。例えば、単一の配列（例えば、PCRアンプリコン）は、350ヌクレオチドを有してもよい。サンプル読み取りは、これらの350ヌクレオチド内に複数のサブシーケンスを含んでもよい。例えば、サンプル読み取りは、例えば、20～50ヌクレオチドを有する第1及び第2のフランキングサブシーケンスを含んでもよい。第1及び第2の隣接するサブシーケンスは、対応するサブシーケンス（例えば、40～100ヌクレオチド）を有する反復セグメントの両側に位置してもよい。隣接するサブ配列のそれぞれは、プライマーサブ配列（例えば、10～30ヌクレオチド）を含んでもよい（又はその一部を含み得る）。読書を容易にするために、用語「サブ配列」は「配列」と称されるが、2つの配列は、共通のストランド上で互いに別個である必要はないことが理解される。本明細書に記載の様々な配列を区別するために、配列は、異なる標識（例えば、標的配列、プライマー配列、隣接配列、参照配列など）を与えられてもよい。「対立遺伝子」などの他の用語は、同様の物体を区別するために異なるラベルを与えられてもよい。アプリケーションは、「読み取り（単数又は複数）」及び「シーケンス読み取り（単数又は複数）」を互換的に使用する。

20

【0656】

用語「paired end sequencing」は、標的フラグメントの両端を配列する配列決定方法を指す。ペレッドエンド配列決定は、ゲノム再構成及び反復セグメントの検出、並びに遺伝子融合及び新規転写物の検出を容易にし得る。パイレッドエンド配列決定の方法は、国際公開第07010252号、国際出願PCT/GB2007/003798号、及び米国特許出願公開第2009/0088327号に記載されており、これらはそれぞれ参照により本明細書に組み込まれる。一実施例では、一連の操作は、以下のように実施されてもよく、（a）核酸のクラスターを生成する、（b）核酸を直鎖化すること、（c）第1の配列決定プライマーをハイブリダイズし、伸長の繰り返しサイクル、走査及び脱ブロッキングを行う。（d）相補的なコピーを合成することによって、フロー細胞表面上の標的核酸を「反転」し、（e）再合成された鎖を直鎖化し、（f）第2配列決定プライマーをハイブリダイズし、伸長の繰り返しサイクル、走査及び脱ブロッキングを行う。反転操作は、ブリッジ増幅の単一サイクルについて上述した試薬を送達することができる。

30

40

【0657】

用語「参照ゲノム」又は「参照配列」は、対象からの同定された配列を参照するために使用され得る任意の生物の部分的又は完全ないずれかの特定の既知のゲノム配列を指す。例えば、ヒト被験者に使用される参照ゲノム、並びに多くの他の生物が、ncbi.nlm.nih.govのNational Center for Biotechnology Informationで見出される。「ゲノム」は、核酸配列で発現される、生物又はウイルスの完全な遺伝子情報を指す。ゲノムは、DNAの遺伝子及び非コード配列の両方を含む。参照配列は、それに位置合わせされたリードよりも大きくてもよい。例えば、それは、少なくとも約100倍大きい、又は少なくとも約1000倍大きい、又は少

50

なくとも約10,000倍大きい、又は少なくとも約105倍大きい、又は少なくとも約106倍大きい、又は少なくとも約107倍大きい場合がある。一実施例では、参照ゲノム配列は、完全長ヒトゲノムのものである。別の例では、参照ゲノム配列は、13番染色体などの特定のヒト染色体に限定される。いくつかの実施態様では、参照染色体は、ヒトゲノムバージョンhg19からの染色体配列である。このような配列は、染色体参照配列と呼ばれる場合があるが、用語参照ゲノムは、そのような配列を網羅することを意図している。参照配列の他の例としては、他の種のゲノム、並びに任意の種の染色体、サブ染色体領域（ストランドなど）などが挙げられる。様々な実施態様では、参照ゲノムは、複数の個体に由来するコンセンサス配列又は他の組み合わせである。しかしながら、特定の用途では、参照配列は、特定の個体から採取されてもよい。他の実施態様では、「ゲノム」はまた、ゲノム配列の特定の記憶形式及び表現を使用するいわゆる「グラフゲノム」も被覆する。一実施態様では、グラフゲノムは、線形ファイルにデータを記憶する。別の実施態様では、グラフゲノムは、代替的な配列（例えば、小さな差異を有する染色体の異なるコピー）がグラフ内の異なる経路として記憶されている表現を指す。グラフゲノムの実施に関する更なる情報は、<https://www.biorxiv.org/content/biorxiv/early/2018/03/20/194530.full.pdf>において見出すことができ、その内容は参照によりその全体が本明細書に組み込まれる。

【0658】

用語「読み取られる」は、ヌクレオチドサンプル又は参照のフラグメントを記述する配列データの集合を指す。用語「読み取られる」は、サンプル読み取り及び／又は参照読み取りを指し得る。典型的には、必ずしもそうではないが、読み取りは、サンプル又は参照における連続的な塩基対の短いシーケンスを表す。読み取りは、サンプル又は参照フラグメントのベース対配列（ATCG）によって記号的に表されてもよい。読み取りが基準シーケンスと一致するか、又は他の基準を満たすかを判定するために、メモリデバイスに記憶され、適切に処理されてもよい。読み取りは、配列決定機器から直接、又はサンプルに関する記憶された配列情報から間接的に得られてもよい。場合によっては、例えば、染色体又はゲノム領域又は遺伝子に位置合わせされ、特異的に割り当てられ得る、より大きな配列又は領域を同定するために使用することができる十分な長さ（例えば、少なくとも約25bp）のDNA配列である。

【0659】

次世代配列決定法としては、例えば、合成技術（Illumina）、ピロ配列決定（454）、イオン半導体技術（Ion Torrent sequencing）、一分子リアルタイム配列決定（Pacific Biosciences）及びライゲーションによる配列決定（SOLiD配列決定）が挙げられる。配列決定法に応じて、各読み取りの長さは、約30bp～10,000bpを超えて変化し得る。例えば、SOLiDシーケンサを用いたDNA配列決定法は、約50bpの核酸リードを生成する。別の例では、Ion Torrent Sequencingは、最大400bpの核酸リードを生成し、454のピロ配列は、約700bpの核酸リードを生成する。更に別の例では、単分子リアルタイム配列決定法は、10,000bp～15,000bpのリードを生成し得る。したがって、特定の実施態様では、核酸配列のリードは、30～100bp、50～200bp、又は50～400bpの長さを有する。

【0660】

用語「サンプル読み取り」、「サンプル配列」又は「サンプル断片」は、サンプルから対象とするゲノム配列に関する配列データを指す。例えば、サンプル読み取りは、順方向及び逆方向プライマー配列を有するPCRアンプリコンからの配列データを含む。配列データは、任意の選択配列手順から得ることができる。サンプルの読み取りは、例えば、配列単位合成（SBS）反応、配列決定・ライゲーション反応、又は反復要素の長さ及び／若しくは同一性を決定することが望ましい任意の他の好適な配列決定方法であり得る。サンプル読み取りは、複数のサンプル読み取りに由来するコンセンサス（例えば、平均又

は加重) 配列であり得る。特定の実施態様では、参照配列を提供する工程は、PCR アンプリコンのプライマー配列に基づいて目的の遺伝子座を同定することを含む。

【0661】

用語「生フラグメント」は、サンプル読み取り又はサンプル断片内の対象とする指定位置又は二次位置に少なくとも部分的に重なり合う、対象とするゲノム配列の一部の配列データを指す。生産物断片の非限定的な例としては、二重ステッチされた断片、単純なステッチされたフラグメント、及び単純な非ステッチの断片が挙げられる。用語「生」は、生のフラグメントがサンプル読み取りにおける配列データとのいくつかの関係を有する配列データを含むことを示すために使用され、生のフラグメントが、サンプル読み取りにおける潜在的要因体に対応し、かつそれを認証又は確認する支持変異体を示すかどうかに関わらず、使用される。用語「生フラグメント」は、フラグメントが、サンプル読み取りにおける変異型コールを検証する支持変異体を必ずしも含むことを示すものではない。例えば、サンプル読み取りが、第1の変異体を呈するために、変異型呼び出しアプリケーションによって判定されるとき、この変異型呼び出しアプリケーションは、1つ又はそれ以上の生のフラグメントが、サンプル読み取りにおける変異体を考慮して、そうでなければ発生することが予想され得る、対応する種類の「支持」変異体を欠くと判定することができる。

【0662】

用語「マッピング」、「整列された」、「整列している」、又は「整列する」という用語は、読み取り又はタグを参照シーケンスと比較し、それによって、参照配列が読み取りシーケンスを含むかどうかを判定するプロセスを指す。参照配列が読み取られた場合、読み取りは参照シーケンスにマップされてもよく、又は特定の別の実施態様では、参照シーケンス内の特定の位置にマッピングされてもよい。いくつかの場合において、整列は、読み取りが特定の参照配列のメンバーであるか否か(すなわち、読み取りが参照配列中に存在するか又は存在していないか)かを単に伝える。例えば、ヒト染色体13についての参照配列に対する読み取りの位置合わせは、13番染色体の参照配列中に読み取りが存在するかどうかを伝える。この情報を提供するツールは、設定メンバシップ試験機と呼ばれることがある。場合によっては、位置合わせは、読み取り又はタグマップがある参照シーケンス内の位置を更に示す。例えば、参照配列がヒトゲノム配列全体である場合、アライメントは、染色体13上にリードが存在することを示してもよく、更に、読み取られたものが染色体13の特定の鎖及び/又は部位にあることを更に示してもよい。

【0663】

用語「インデル」は、生物のDNA中の塩基の挿入及び/又は欠失を指す。マイクロインデルは、1~50ヌクレオチドの正味変化をもたらすインデルを表す。インデルの長さが3の倍数でない限り、ゲノムの領域をコードする際に、フレームシフト変異が生じる。インデルは、点突然変異と対比することができる。インデル挿入物は、配列からヌクレオチドを欠失させるが、点変異は、DNA中の全体的な数を変えなくヌクレオチドのうちの1つを置き換える置換の形態である。インデルはまた、隣接するヌクレオチドにおける置換として定義され得るTandem Base変異(TBM)と対比することもできる(主に2つの隣接するヌクレオチドで置換されるが、隣接する3つのヌクレオチドでの置換が観察された)。

【0664】

用語「変異体」は、核酸参照とは異なる核酸配列を指す。典型的な核酸配列変異体としては、限定するものではないが、単一のヌクレオチド多型(SNP)、短い欠失及び挿入多型(Indel)、コピー数変動(CNV)、マイクロ衛星マーカー、又は短いタンDEM反復及び構造変異が挙げられる。体細胞変異体の呼び出しは、DNAサンプル中に低頻度で存在する変異体を同定するための努力である。体細胞変異体のコーリングは、癌治療の文脈において対象となる。癌は、DNAにおける変異の蓄積によって引き起こされる。腫瘍由来のDNAサンプルは、一般的に不均質であり、いくつかの正常細胞、癌進行の早期段階(より少ない突然変異を伴う)、及び一部の後期細胞(より多くの変異を有する)を含む。この不均一性のため、腫瘍を配列決定するとき(例えば、FFPEサンプルから

）、体細胞変異は、多くの場合、低頻度で現れる。例えば、S N V は、所与の塩基を被覆するリードの10%のみに見られ得る。変異体分類子によって体細胞又は生殖細胞系として分類される変異体は、本明細書では「試験中の変異体」とも称される。

【0665】

用語「ノイズ」は、配列決定プロセス及び／又は変異型呼び出しアプリケーションにおける1つ又はそれ以上の誤差から生じる、誤りのある変異型コールを指す。

【0666】

用語「変異体頻度」は、集団内の特定の遺伝子座における対立遺伝子（遺伝子の変異体）の相対頻度を表し、分画又は割合として表される。例えば、分画又は割合は、その対立遺伝子を保有する集団中の全ての染色体の割合であってもよい。一例として、サンプル変異体頻度は、個体から対象となるゲノム配列について得られたリード及び／又はサンプルの数に対応する「集団」にわたって、対象とするゲノム配列に沿った特定の遺伝子座／位置における対立遺伝子／変異体の相対頻度を表す。別の例として、ベースライン変異体頻度は、1つ又はそれ以上のベースラインゲノム配列に沿った特定の遺伝子座／位置における対立遺伝子／変異体の相対頻度を表し、ここで、1つ又はそれ以上のベースラインゲノム配列について得られた、1つ又はそれ以上のベースラインゲノム配列に沿った特定の遺伝子座／位置における対立遺伝子／変異体の相対頻度を表す。

10

【0667】

用語「変異型対立遺伝子頻度（V A F）」は、変異体を標的位置での全体的な被覆率で割った、配列決定されたリードの割合を指す。V A F は、変異体を保有する配列決定されたリードの割合の尺度である。

20

【0668】

用語「位置」、「指定位置」、及び「遺伝子座」は、ヌクレオチド配列内の1つ又はそれ以上のヌクレオチドの位置又は座標を指す。用語「位置」、「指定位置」、及び「遺伝子座」はまた、ヌクレオチドの配列における1つ又はそれ以上の塩基対の位置又は座標を指す。

【0669】

用語「ハプロタイプ」は、互いに遺伝する染色体上の隣接する部位における対立遺伝子の組み合わせを指す。ハプロタイプは、存在する場合、所与の座の組の間に生じた組み換え事象の数に応じて、1つの遺伝子座、いくつかの座、又は染色体全体であってもよい。

30

【0670】

本明細書における「閾値」という用語は、サンプル、核酸、又はその一部（例えば、読み取り）を特徴付けるためのカットオフとして使用される数値又は数値を指す。閾値は、経験的分析に基づいて変化してもよい。閾値は、そのような値を生じさせる源が特定の方法で分類されるべきかどうかを判定するために、測定値又は計算値と比較することができる。閾値は、経験的又は分析的に識別することができる。閾値の選択は、ユーザーが分類を行う必要があることを望む信頼度に依存する。閾値は、特定の目的（例えば、感度及び選択性のバランスのために）選択されてもよい。本明細書で使用する時、用語「閾値」は、分析の過程が変化し得る点、及び／又はアクションがトリガされ得る点を示す。閾値は、所定の数である必要はない。その代わりに、閾値は、例えば、複数の因子に基づく関数であってもよい。閾値は、状況に適応し得る。更に、閾値は、上限、下限、又は限界間の範囲を示し得る。

40

【0671】

いくつかの実施態様では、配列決定データに基づく指標又はスコアは、閾値と比較され得る。本明細書で使用する時、用語「メトリック」又は「スコア」は、配列決定データから決定された値又は結果を含んでもよく、又は配列決定データから決定された値又は結果に基づく関数を含んでもよい。閾値と同様に、指標又はスコアは、状況に適応し得る。例えば、指標又はスコアは、正規化された値であってもよい。スコア又はメトリックの例として、1つ又はそれ以上の実施態様は、データを分析する際にカウントスコアを使用することができる。計数スコアは、サンプル読み取りの数に基づいてもよい。サンプル読み

50

取りは、サンプル読み取りが少なくとも1つの共通の特性又は品質を有するように、1つ又はそれ以上のフィルタリング段階を経ていてもよい。例えば、計数スコアを決定するために使用されるサンプル読み取りのそれぞれは、参照配列と整列されていてもよく、又は潜在的な対立遺伝子として割り当てられてもよい。共通の特性を有するサンプル読み取りの数をカウントして、読み取りカウントを決定することができる。カウントスコアは、読み取りカウントに基づいてもよい。いくつかの実施態様では、計数スコアは、読み取りカウントと等しい値であってもよい。他の実施例では、計数スコアは、読み取りカウント及び他の情報に基づいてもよい。例えば、計数スコアは、遺伝子座の特定の対立遺伝子の読み取りカウント、及び遺伝子座の合計リード数に基づいてもよい。いくつかの実施態様では、計数スコアは、遺伝子座の読み出しカウント及び以前に得られたデータに基づいてもよい。いくつかの実施態様では、計数スコアは、所定の値間の正規化スコアであってもよい。計数スコアはまた、サンプルの他の遺伝子座からのリードカウントの関数、又は対象となるサンプルと同時に実行された他のサンプルからのリードカウントの関数であってもよい。例えば、計数スコアは、特定の対立遺伝子の読み取りカウント及びサンプル中の他の遺伝子座の読み取りカウント、及び／又は他のサンプルからのリードカウントの関数であってもよい。一例として、他の遺伝子座からのリードカウント及び／又は他のサンプルからのリードカウントを使用して、特定の対立遺伝子についての計数スコアを正規化してもよい。

10

【0672】

用語「被覆率」又は「断片被覆率」は、配列の同じフラグメントに対する多数のサンプルリードの計数又は他の尺度を指す。読み取りカウントは、対応するフラグメントをカバーするリードの数のカウントを表し得る。あるいは、被覆率は、履歴知識、サンプルの知識、遺伝子座の知識などに基づく指定された因子を、読み取りカウントに掛けることによって決定されてもよい。

20

【0673】

用語「読み取り深さ」（従来、「×」が続く数）は、標的位置における重複アラインメントを伴う配列決定されたリードの数を指す。これは、多くの場合、1組の間隔（エキソン、遺伝子、又はパネルなど）のカットオフを超える平均又は百分率として表現される。例えば、臨床報告は、パネル平均被覆率が、標的化されたベースカバー $>100\times$ の98%を有する1,105 $\times$ であると言うことができる。

30

【0674】

用語「ベースコール品質スコア」又は「Qスコア」は、単一の配列決定された塩基が正しい確率に反比例する0~50からの範囲のPHREDスケールされた確率を指す。例えば、20のQを有するTベースコールは、99.99%の確率で正しいと考えられる。Q<20での任意のベースコールは、低品質であると見なされるべきであり、変異体を支持する配列決定されたリードのかなりの割合が低い場合に同定される任意の変異体は、潜在的に偽陽性であると見なされるべきである。

【0675】

用語「変異体リード」又は「変異体リード番号」は、変異体の存在を支持する配列決定されたリードの数を指す。

40

【0676】

「ストリンデディティー」（又はDNAストランド）に関して、DNA中の遺伝的メッセージは、文字A、G、C、及びTの文字、例えば、5'-AGGACA-3'として表すことができる。多くの場合、配列は、本明細書に示される方向、すなわち、5'端を左に、3'端を右に書き込む。DNAは、（特定のウイルスのように）一本鎖分子として生じる場合があるが、通常、二本鎖単位としてDNAを見つける。これは、2つの抗平行ストランドを有する二重螺旋構造を有する。この場合、「逆平行」という語は、2つのストランドが平行に走るが、反対の極性を有することを意味する。二本鎖DNAは、塩基とペアリングによって一緒に保持され、ペアリングは、アデニン（A）対がチミン（T）及びシトシン（C）対とグアニン（G）との対となるように、常に保持される。このペアリングは相

50

補性と呼ばれ、1本のDNA鎖は、他方の相補体であると言われる。したがって、二本鎖DNAは、5'-AGGACA-3'及び3'-TCCGTG-5'と同様に、2つのストリングとして表され得る。2つのストランドは、反対の極性を有することに留意されたい。したがって、2つのDNA鎖のストランド性は、基準ストランド及びその補体、順方向及び逆方向ストランド、トップ及びボトムストランド、センス及びアンチセンスストランド、又はWatson及びCrickストランドと呼ぶことができる。

【0677】

リードアライメント（リードマッピングとも呼ばれる）は、ゲノム中の配列が由来する場合に、参照するプロセスである。整列が行われると、所与の読み取りの「マッピング品質」又は「マッピング品質スコア（MAPQ）」は、ゲノム上のその位置が正しい確率を定量化する。マッピング品質は、位相スケールで符号化され、Pはアライメントが正しくない確率である。確率は、以下のように計算される。式 $P = 10^{(-MAPQ/10)}$ 中、MAPQはマッピング品質である。例えば、-4の電力に対する $40 = 10$ のマッピング品質は、読み取りが不正確に位置合わせされた0.01%の機会が存在することを意味する。したがって、マッピング品質は、読み取りの基本品質、参照ゲノムの複雑性、及びパレドエンド情報などのいくつかの位置合わせ因子と関連付けられる。最初に、読み取りの基本品質が低い場合、観察された配列が誤っている可能性があり、したがってそのアライメントが誤っていることを意味する。第2に、マッピング能力はゲノムの複雑さを指す。反復領域は、これらの領域に含まれるマップ及びリードをマッピングすることがより困難であり、通常、マッピング品質が低くなる。この文脈では、MAPQは、リードが一意的に整列されておらず、それらの実際の原点を決定することができないという事実を反映する。第3に、パレドエンド配列決定データの場合、コンコダント対は、より良好に整列される可能性が高い。マッピング品質が高いほど、アライメントがより良好である。良好なマッピング品質と整合された読み取りは、通常、読み出しシーケンスが良好であり、高いマッピング可能領域内ではわずかな不一致と位置合わせされたことを意味する。MAPQ値は、アライメント結果の品質管理として使用することができる。20よりも高いMAPQと位置合わせされたリードの割合は、通常、下流分析のためである。

【0678】

本明細書で使用する時、「信号」は、例えば画像内の発光、好ましくは発光などの検出可能な事象を指す。したがって、好ましい別の実施態様では、信号は、画像内に捕捉された任意の検出可能な発光（すなわち、「スポット」）を表すことができる。したがって、本明細書で使用する時、「信号」は、検体の分析物からの実際の放出の両方を指すことができ、実際の分析物と相関しない擬似発光を指すことができる。したがって、信号はノイズから生じ得、試験片の実際の分析物を代表しないように後に廃棄することができる。

【0679】

本明細書で使用する時、用語「塊」は、一群の信号を指す。特定の実施態様では、信号は、異なる検体に由来する。好ましい別の実施態様では、信号塊は、一緒にクラスター化する信号群である。より好ましい実施態様では、シグナル凝集は、1つの増幅オリゴヌクレオチドによって覆われた物理的領域を表す。各信号塊は、理想的には、いくつかの信号（テンプレートサイクル当たり1つ、恐らくはクロストークによってより多く）として観察されるべきである。したがって、2つ（又はそれ以上）の信号が同じ信号の塊からテンプレートに含まれる、重複する信号が検出される。

【0680】

本明細書で使用する時、「最小」、「最大」、「最小化」、「最大化」、及びその文法的変異形などの用語は、絶対最大値又は最小値ではない値を含むことができる。いくつかの実施態様では、値は、最大値及び最小値付近を含む。他の実施例では、値は、局所的な最大値及び／又は局所最小値を含むことができる。いくつかの実施態様では、値は、絶対最大値又は最小値のみを含む。

【0681】

本明細書で使用する時、「クロストーク」は、別個の画像においても検出される1つ

の画像内の信号の検出を指す。好ましい別の実施態様では、クロストークは、放出された信号が2つの別個の検出チャンネルで検出されるときに発生し得る。例えば、放射された信号が1つの色で発生する場合、その信号の放射スペクトルは、別の色で別の放射された信号と重なってもよい。好ましい実施態様では、ヌクレオチド塩基A、C、G、及びTの存在を示すために使用される蛍光分子は、別個のチャンネルで検出される。しかし、A及びCの発光スペクトルは重複するため、色チャンネルを使用した検出中に、C色信号の一部が検出され得る。したがって、A信号とC信号との間のクロストークにより、1つのカラー画像からの信号が他のカラー画像に現れることを可能にする。いくつかの実施態様では、G及びTクロストークがある。いくつかの実施態様では、チャンネル間のクロストークの量は非対称である。チャンネル間のクロストークの量は、とりわけ、適切な放射スペクトルを有する信号分子の選択、並びに検出チャンネルのサイズ及び波長範囲の選択によって制御され得ることが理解されるであろう。

10

【0682】

本明細書で使用する時、「レジスタ」、「登録」、「登録」、及び同様の用語は、画像又はデータセット内の信号を、別の時点又は視点からの画像又はデータセット内の信号と関連させるための任意のプロセスを指す。例えば、位置合わせを使用して、画像のセットからの信号を整列させてテンプレートを形成することができる。別の例では、位置合わせを使用して、他の画像からの信号をテンプレートに位置合わせすることができる。1つの信号は、別の信号に直接又は間接的に登録されてもよい。例えば、画像「S」からの信号は、画像「G」に直接登録されてもよい。別の例として、画像「N」からの信号は、画像「G」に直接登録されてもよく、あるいは、画像「N」からの信号は、以前に画像「G」に登録された画像「S」に登録されてもよい。したがって、画像「N」からの信号は、画像「G」に間接的に登録される。

20

【0683】

本明細書で使用する時、用語「基準」は、物体内部又は物体上の区別可能な基準点を意味することを意図する。基準点は、例えば、マーク、第2の物体、形状、縁部、領域、不規則性、チャンネル、ピット、ポストなどであり得る。基準点は、オブジェクトの画像内に、又はオブジェクトを検出することに由来する別のデータセット内に存在することができる。基準点は、物体の平面内のx及びy又はz座標によって指定することができる。代替的に又は追加的に、基準点は、例えば、物体と検出器との相対位置によって定義される、x y平面に直交するz座標によって指定することができる。基準点に対する1つ又はそれ以上の座標は、オブジェクトの1つ又はそれ以上の他の分析物、又はオブジェクトに由来する画像又は他のデータセットに対して指定することができる。

30

【0684】

本明細書で使用する時、用語「光信号」は、例えば、蛍光、発光、散乱、又は吸収信号を含むことを意図する。光信号は、紫外線(UV)範囲(約200~390nm)、可視(VIS)範囲(約391~770nm)、赤外線(IR)範囲(約0.771~25マイクロメートル)、又は電磁スペクトルの他の範囲で検出することができる。これらの範囲のうちの1つ又はそれ以上の全て又は一部を除外する方法で、光信号を検出することができる。

40

【0685】

本明細書で使用する時、用語「信号レベル」は、所望又は所定の特性を有する検出されたエネルギー又は符号化された情報の量又は量を意味することを意図する。例えば、光信号は、強度、波長、エネルギー、周波数、電力、輝度などのうちの1つ又はそれ以上によって定量化することができる。他の信号は、電圧、電流、電界強度、磁場強度、周波数、電力、温度などの特性に従って定量化することができる。信号の不在は、ゼロの信号レベル、又はノイズとは有意に区別されない信号レベルであると理解される。

【0686】

本明細書で使用する時、用語「シミュレートする」は、物理又は行動の特性を予測する物理的又は行動の表現又はモデルを作成することを意味することを意図する。表現又は

50

モデルは、多くの場合、そのもの又は行動と区別可能であり得る。例えば、表現又はモデルは、色、被加工、サイズ、又は形状の全て又は一部から検出される信号の強度などの1つ又はそれ以上の特性に対するものと区別することができる。特定の実施態様では、表現又はモデルは、何か又は行為と比較して、理想化、誇張、ミュート、又は不完全であり得る。したがって、いくつかの実施態様では、モデルの表現は、例えば、上記の特性のうちの少なくとも1つに関して表すものであることを表すものであることができる。表現又はモデルは、本明細書の他の箇所に記載されるもののうちの1つ又はそれ以上などのコンピュータ可読フォーマット又は媒体に提供され得る。

【0687】

本明細書で使用する時、用語「特定の信号」は、背景エネルギー又は情報などの他のエネルギー又は情報にわたって選択的に観察される、検出されたエネルギー又は符号化情報を意味することを意図する。例えば、特定の信号は、特定の強度、波長、若しくは色で検出される光信号、特定の周波数、電力若しくは場強度で検出される電気信号、又は分光及び分析検出に関する技術分野において既知の他の信号であり得る。

【0688】

本明細書で使用する時、用語「スイング」は、物体の矩形部分を意味することを意図する。スイングは、ストリップの最長寸法に平行な方向に、物体と検出器との間の相対移動によって走査される細長いストリップであり得る。一般に、矩形部分又はストリップの幅は、その全長に沿って一定である。物体の複数のスエージは、互いに平行であってもよい。物体の複数のスエージは、互いに重なり合い、互いに隣接するか、又は間質領域によって互いに分離され得る。

【0689】

本明細書で使用する時、用語「分散」は、予想される差、及び観察される差、又は2つ又はそれ以上の観測結果間の差を意味することを意図する。例えば、分散は、期待値と測定値との間の不一致であり得る。標準偏差、標準偏差の二乗、変動係数などの統計関数を使用して、分散を表すことができる。

【0690】

本明細書で使用する時、用語「x y 座標」は、x y 平面内の位置、サイズ、形状、及び／又は向きを指定する情報を意味することを意図する。情報は、例えば、デカルトシステムにおける数値座標であり得る。座標は、x 軸及び y 軸の一方又は両方に対して提供することができ、又は x y 平面内の別の場所に対して提供することができる。例えば、物体の検体の座標は、対象物の基準又は他の分析物の位置に対する検体の場所を指定することができる。

【0691】

本明細書で使用する時、用語「x y 平面」は、直線軸 x 及び y によって画定される2次元領域を意味することを意図する。検出器及び検出器によって観察される物体を参照して使用される場合、検出器と検出されている物体との間の観測方向に直交するように更に指定することができる。

【0692】

本明細書で使用する時、用語「z 座標」は、x y 平面に直交する軸に沿った点、線、又は領域の位置を指定する情報を意味することを意図する。特定の別の実施態様では、z 軸は、検出器によって観察される物体の領域に直交する。例えば、光学系の焦点の方向は、z 軸に沿って指定されてもよい。

【0693】

いくつかの実施態様では、獲得された信号データは、アフィン変換を用いて変換される。いくつかのそのような実施態様では、テンプレートの生成は、色チャンネル間のアフィン変換が走る間に一貫しているという事実を使用する。この一貫性のため、標本中の検体の座標を決定する際に、デフォルトオフセットのセットを使用することができる。例えば、デフォルトオフセットファイルは、A チャンネルなどの1つのチャンネルに対する異なるチャンネルに対する相対変換（シフト、スケール、スキュー）を含むことができる。しかしなが

10

20

30

40

50

ら、他の実施態様では、ラン中及び／又は走る間の色チャンネルドリフト間のオフセットは、オフセット駆動型テンプレート生成を困難にする。このような実施例では、本明細書で提供される方法及びシステムは、オフセットしたテンプレート生成を利用することができ、これについては以下で更に説明する。

【0694】

上記の実施態様のいくつかの態様では、システムはフローセルを含み得る。一部の態様では、フローセルは、レーン、又は他の構成のタイルを含み、タイルの少なくとも一部は、1つ又はそれ以上の検体群を含む。一部の態様では、検体は、核酸などの複数の分子を含む。特定の態様では、フローセルは、標識されたヌクレオチド塩基を核酸の配列に送達し、それによって、核酸を含む検体に対応するシグナルを生成するように、検体内の核酸にハイブリダイズするプライマーを伸長させるように構成される。好ましい実施態様では、検体内の核酸は、互いに同一又は実質的に同一である。

【0695】

本明細書に記載される画像解析システムのいくつかにおいて、画像のセット内の各画像は、色信号を含み、異なる色は、異なるヌクレオチドベースに対応する。一部の態様では、画像のセットの各画像は、少なくとも4つの異なる色から選択される単一の色を有する信号を含む。一部の態様では、画像のセット内の各画像は、4つの異なる色から選択される単一の色を有する信号を含む。本明細書に記載されるシステムのいくつかにおいて、核酸は、4つの異なる画像を生成するように、4つの異なる標識ヌクレオチド塩基を分子の配列に提供することにより、核酸を配列決定することができ、各画像は単一の色を有するシグナルを含み、信号色が、4つの異なる画像のそれぞれに対して異なることにより、核酸内の特定の位置に存在する4つの可能なヌクレオチドに対応する4つのカラー画像のサイクルを生成する。方法。特定の態様では、システムは、追加の標識ヌクレオチド塩基を分子の配列に送達するように構成されたフローセルを含み、それによって複数のカラー画像のサイクルを生成する。

【0696】

好ましい実施態様形態では、本明細書で提供される方法は、プロセッサがアクティブにデータを取得しているか、又はプロセッサが低活動状態にあるかどうかを判定することを含み得る。多数の高品質画像を取得及び記憶することは、典型的には、大量の記憶容量を必要とする。更に、取得され記憶されると、画像データの分析はリソース集約的になり得、追加の画像データの取得及び記憶などの他の機能の処理能力を妨げる可能性がある。したがって、本明細書で使用する時、用語「低活動状態」は、所与の時間におけるプロセッサの処理能力を指す。いくつかの実施態様では、低活動状態は、プロセッサがデータを取得及び／又は記憶していないときに生じる。いくつかの実施態様では、一部のデータ取得及び／又はストレージが行われる場合には、低いアクティビティ状態が生じるが、他の機能に干渉することなく画像解析が同時に生じ得るように、追加の処理能力が残る。

【0697】

本明細書で使用する時、「競合を特定する」とは、複数のプロセスがリソースに対して競合する状況を特定することを指す。いくつかのそのような実施態様では、1つのプロセスは、別のプロセスに対して優先度を与えられる。いくつかの実施態様では、時間、処理能力、記憶能力、又は優先度が与えられる任意の他のリソースの割り当てに対する優先度を与える必要性に関連し得る。したがって、いくつかの実施態様では、処理時間又は容量が、データセットを分析し、データセットを取得及び／又は記憶するかのいずれかなどの2つのプロセス間に分散される場合、2つのプロセス間の不一致が存在し、プロセスのうちの1つに優先度を与えることによって解決することができる。

【0698】

本明細書では、画像解析を実行するためのシステムも提供される。システムは、プロセッサと、記憶容量と、画像解析用のプログラムと、を含むことができ、プログラムは、記憶のための第1のデータセット及び分析のための第2のデータセットを処理するための命令を含み、処理は、記憶装置上の第1のデータセットを取得及び／又は記憶することと、

プロセッサが第1のデータセットを取得していないときに第2のデータセットを解析することと、を含む。特定の態様では、プログラムは、第1のデータセットを取得及び／又は記憶することと、第2のデータセットを解析することとの間の競合の少なくとも1つのインスタンスを識別するための命令を含み、第1のデータセットを取得及び／又は記憶することが優先度与えられるように、画像データを取得及び／又は記憶することが優先される。特定の態様では、第1のデータセットは、光学撮像装置から取得された画像ファイルを含む。特定の態様では、システムは、光学撮像装置を更に備える。一部の態様では、光学撮像装置は、光源と検出デバイスとを備える。

【0699】

本明細書で使用する時、用語「プログラム」は、タスク又はプロセスを実行するための命令又はコマンドを指す。用語「プログラム」は、用語「モジュール」と互換的に使用され得る。特定の実施態様では、プログラムは、同じコマンドセットの下で実行される様々な命令のコンパイルであり得る。他の実施態様では、プログラムは、別個のバッチ又はファイルを参照することができる。

【0700】

以下に記載されるのは、本明細書に記載される画像解析を実行するための方法及びシステムを利用する驚くべき効果の一部である。いくつかの配列決定の実現例では、配列決定システムの有用性の重要な尺度は、その全体的な効率である。例えば、1日当たりに生成されるマッピング可能なデータの量、並びに器具の設置及び実行の総コストは、経済的な配列決定ソリューションの重要な態様である。マッピング可能なデータを生成し、システムの効率を高めるための時間を短縮するために、リアルタイムのベースコールを機器コンピュータ上で有効にすることができ、配列決定ケミストリー及び画像化と並行して実行することができる。これにより、配列決定化学仕上げの前に、データ処理及び分析が完了することを可能にする。更に、中間データに必要な記憶を低減し、ネットワークを横切って移動する必要があるデータの量を制限することができる。

【0701】

シーケンス出力が増加している間、本明細書で提供されるシステムからネットワークに転送された実行ごとのデータ、及び二次分析処理ハードウェアは、実質的に減少している。機器コンピュータ（取得コンピュータ）上でデータを変換することにより、ネットワークロードが劇的に低減される。これらのオン機器、オフネットワークデータ低減技術を伴わずに、DNA配列決定機器のフレッツの画像出力は、ほとんどのネットワークをクリップするであろう。

【0702】

ハイスループットDNA配列決定機器の広範な採用は、使用の容易さ、用途の範囲に対する支持、及び実質的に任意の実験室環境に対する適合性によって、部分的に駆動されてきた。本明細書に提示される高度に効率的なアルゴリズムは、配列決定インスツルメントを制御することができる単純なワークステーションに、有意な分析機能を加えることを可能にする。計算ハードウェアの必要条件のこの低減は、配列決定出力レベルが増加し続けるにつれて、更に重要となる、いくつかの実用的な利点を有する。例えば、単純なタワー、熱生成、実験室設置面積、及び電力消費を最小限に抑えるために、画像解析及びベースコールを行うことによって、最小に保たれる。対照的に、他の商業的な配列決定技術は、1次分析のために、最大5回の処理電力で、そのコンピューティングインフラストラクチャを最近ランプアップして、熱出力及び電力消費の増加を開始する。したがって、いくつかの実施態様では、本明細書で提供される方法及びシステムの計算効率は、サーバハードウェアを最小限に抑えながら、それらの配列決定スループットを増加させることを可能にする。

【0703】

したがって、いくつかの実施態様では、本明細書に提示される方法及び／又はシステムは、状態マシンとして機能し、各試料の個々の状態の追跡を保ち、試料が次の状態に進む準備ができていることを検出すると、適切な処理を行い、試料をその状態に前進させる。

状態マシンがファイルシステムを監視して、好ましい実施例に従って試料が次の状態に進む準備ができているかを判定する方法のより詳細な例が、以下の実施例 1 に記載されている。

【0704】

好ましい実施態様では、本明細書で提供される方法及びシステムは、マルチスレッドであり、構成可能な数のスレッドと協働することができる。したがって、例えば、核酸配列決定の文脈において、本明細書で提供される方法及びシステムは、リアルタイム分析のためのライブ配列決定実行中に背景において作用することができ、又はオフライン分析のために既存の画像データセットを使用して実行することができる。特定の好ましい実施形態では、方法及びシステムは、それぞれのスレッドを、それが関与する検体のそれ自体のサブセットを与えることによって、マルチスレッドを取り扱う。これにより、スレッド保持の可能性が最小限に抑えられる。

【0705】

本開示の方法は、検出装置を使用して物体の標的画像を取得する工程を含むことができ、この画像は、オブジェクト上の検体の繰り返しパターンを含む。表面の高解像度撮像が可能な検出装置が特に有用である。特定の実施態様では、検出装置は、本明細書に記載される密度、ピッチ、及び／又は検体サイズにおける検体を区別するのに十分な分解能を有するであろう。表面から画像又は画像データを得ることができる検出装置が特に有用である。例示的な検出器は、物体と検出器とを静的関係に維持しつつ、面積画像を取得するように構成されたものである。走査装置も使用することができる。例えば、連続領域画像を取得する装置（例えば、「ステップ及びショット」検出器と呼ばれる）を使用することができる。また、物体の表面上の点又は線を連続的に走査して、表面の画像を構築するためにデータを蓄積するデバイスも有用である。点走査検出器は、表面の $x-y$ 平面内のラスタ運動を介してオブジェクトの表面上の点（すなわち、小さい検出領域）を走査するように構成することができる。線走査検出器は、物体の表面の y 次元に沿った線を走査するように構成することができ、この線の最長寸法は、 x 次元に沿って生じる。検出デバイス、物体、又はその両方を移動させて、走査検出を達成できることが理解されるであろう。例えば核酸配列決定用途において特に有用な検出装置は、米国特許出願公開第 2012/0270305 (A1) 号、同第 2013/0023422 (A1) 号、及び同第 2013/0260372 (A1) 号、並びに米国特許第 5,528,050 号、同第 5,719,391 号、同第 8,158,926 号及び同第 8,241,573 号に記載されており、これらはそれぞれ、参照により本明細書に組み込まれる。

【0706】

本明細書に開示される実施態様は、ソフトウェア、ファームウェア、ハードウェア、又はそれらの任意の組み合わせを生成するためのプログラミング技術又は工学技術を使用して、製造方法、装置、システム、又は物品として実施態様されてもよい。本明細書で使用するとき、用語「製造物品」は、光学記憶デバイスなどのハードウェア又はコンピュータ可読媒体、並びに揮発性又は不揮発性メモリデバイス内に実施態様されるコード又は論理を指す。そのようなハードウェアとしては、フィールドプログラマブルゲートアレイ (FPGA)、粗粒度再構成可能構造 (CGRA)、特定用途向け集積回路 (ASIC)、複合プログラマブル論理デバイス (CPLD)、プログラマブルロジックアレイ (PLA)、マイクロプロセッサ、又は他の同様の処理装置が挙げられるが、これらに限定されない。特定の実施態様では、本明細書に記載される情報又はアルゴリズムは、非一過性記憶媒体中に存在する。

【0707】

特定の実施態様形態では、本明細書に記載されるコンピュータ実装の方法は、物体の複数の画像が取得されている間、リアルタイムで発生することができる。このようなリアルタイム分析は、核酸配列が流体及び検出工程の繰り返しサイクルに供される核酸配列決定用途に特に有用である。配列決定データの分析は、多くの場合、本明細書に記載される方法をリアルタイム又は背景で実行するのに有益であり得る一方で、他のデータ獲得又は分

析アルゴリズムがプロセス中である間に、本明細書に記載される方法を実行することが有益であり得る。本方法で使用することができるリアルタイム分析法の例は、Illumina, Inc (San Diego, Calif.) から市販されており、及び／又は、参照により本明細書に組み込まれる米国特許出願公開第2012/0020537 (A1) 号に記載されているMiSeq及びHiSeq配列決定機器に使用されるものである。

【0708】

1つ又はそれ以上のプログラムされたコンピュータによって形成され、本明細書に記載される方法の1つ又はそれ以上のステップを実行するために実行されるコードを有するプログラミングが、1つ又はそれ以上の機械可読媒体上に記憶されている、例示的データ分析システム。一実施態様では、例えば、システムは、標的オブジェクトからデータを取得するように構成された1つ又はそれ以上の検出システム（例えば、光学撮像システム）へのシステムのネットワーキングを可能にするように設計されたインターフェースを含む。インターフェースは、適切な場合には、データを受信及び条件することができる。特定の実施態様では、検出システムは、例えば、アレイ又は他の物体の画像と一緒に形成する個々の画像要素又はピクセルを表す画像データを出力する。プロセッサは、処理コードによって定義された1つ又はそれ以上のルーチンに従って、受信した検出データを処理する。処理コードは、様々な種類のメモリ回路に記憶されてもよい。

【0709】

現時点で企図される実施態様によれば、検出データ上で実行される処理コードは、検出データを分析して、データ内で可視又は符号化された個々の検体の場所、及び分析物が検出されない場所（すなわち、分析物が存在しないか、又は、既存の分析物から有意な信号が検出されない場所）及びメタデータを判定するように設計されたデータ分析ルーチンを含む。特定の実施態様では、アレイ内の検体位置は、典型的には、撮像された検体に付着した蛍光染料の存在に起因して、非検体位置よりも明るく見える。検体は、例えば、検体におけるプローブの標的が検出されているアレイ内に存在しない場合、分析物は、それらの周囲領域よりも明るく見える必要はないことが理解されるであろう。個々の検体が現れる色は、使用される染料、並びに撮像目的のために撮像システムによって使用される光の波長の関数であり得る。標的が結合されていない、又は特定のラベルを有さない検体は、マイクロアレイ内の予想される場所などの他の特性に従って特定することができる。

【0710】

データ分析ルーチンがデータ中に個々の分析物を配置すると、値割り当てが実行され得る。一般に、値割り当ては、対応する場所における検出器構成要素（例えば、ピクセル）によって表されるデータの特性に基づいて、各分析物にデジタル値を割り当てる。すなわち、例えば、画像化データが処理されるとき、値割り当てルーチンは、特定の場所で特定の色又は波長の光が特定の場所で検出されたことを認識するように設計されてもよい。典型的なDNA画像化用途では、例えば、4つの共通ヌクレオチドは、4つの別個の区別可能な色によって表される。次いで、各色は、そのヌクレオチドに対応する値を割り当てられてもよい。

【0711】

本明細書で使用するとき、用語「モジュール」、「システム」、又は「システムコントローラ」は、1つ又はそれ以上の機能を実行するように動作するハードウェア及び／又はソフトウェアシステム及び回路を含み得る。例えば、モジュール、システム、又はシステムコントローラは、コンピュータメモリなどの有形及び非一時的コンピュータ可読記憶媒体上に記憶された命令に基づいて動作を実行する、コンピュータプロセッサ、コントローラ、又は他のログベースのデバイスを含んでもよい。あるいは、モジュール、システム、又はシステムコントローラは、有線論理及び回路に基づいて動作を実行する、有線デバイスを含んでもよい。添付の図面に示されるモジュール、システム、又はシステムコントローラは、ソフトウェア又は結線命令に基づいて動作するハードウェア及び回路、ハードウェアを動作させるように命令するソフトウェア、又はそれらの組み合わせを表し得る。モジュール、システム、又はシステムコントローラは、1つ又はコンピュータマイクロプロ

セッサなどの1つ又はそれ以上のプロセッサを含む、及び／又はそれと接続されるハードウェア回路又は回路を含むか、又は表すことができる。

【0712】

本明細書で使用する時、用語「ソフトウェア」及び「ファームウェア」は互換可能であり、RAMメモリ、ROMメモリ、EPROMメモリ、EEPROMメモリ、及び不揮発性RAM（NVRAM）メモリを含むコンピュータによって実行されるメモリに記憶された任意のコンピュータプログラムを含む。上記メモリタイプは単なる例であり、コンピュータプログラムの記憶に使用可能なメモリの種類に限定されるものではない。

【0713】

分子生物学分野では、使用中の核酸配列決定のためのプロセスの1つは、配列番号合成である。この技術は、非常に平行な配列決定プロジェクトに適用することができる。例えば、自動プラットフォームを使用することにより、数百万の配列決定反応を同時に行うことが可能である。したがって、本発明の実施態様の1つは、核酸配列決定中に生成された画像データを取得、記憶、及び分析するための器具及び方法に関する。

【0714】

取得及び記憶することができるデータ量の莫大なゲインは、合理化された画像解析方法を更により有益にする。例えば、本明細書に記載される画像解析方法は、設計者及びエンドユーザーの両方が、既存のコンピュータハードウェアの効率的な使用を行うことを可能にする。したがって、本明細書では、迅速に増加するデータ出力の面における処理データの計算量を低減する方法及びシステムが本明細書に提示される。例えば、DNA配列決定の分野では、収率は最近の過程で15倍に拡大され、DNA配列決定デバイスの単一の実行において何百ギガーゼに達する可能性がある。計算インフラストラクチャの要件が比例的に増加した場合、大規模なゲノム規模の実験は、ほとんどの研究者に到達していない。したがって、より多くの生シーケンスデータの生成は、二次分析及びデータ記憶の必要性を増加させ、データ輸送及び記憶の最適化を非常に有益にする。本明細書に提示される方法及びシステムのいくつかの実施態様は、使用可能なシーケンスデータを生成するために必要な時間、ハードウェア、ネットワーキング、及び実験室インフラストラクチャ要件を低減することができる。

【0715】

本開示は、方法を実行するための様々な方法及びシステムを説明する。方法のいくつかの例は、一連の工程として記載される。しかしながら、実施態様は、本明細書に記載される特定の工程及び／又は工程の順序に限定されないことを理解されたい。工程は省略されてもよく、工程は修正されてもよく、及び／又は他の工程が追加されてもよい。更に、本明細書に記載される工程を組み合わせることができ、工程は同時に実行されてもよく、工程は同時に実行されてもよく、工程は複数のサブステップに分割されてもよく、工程は、異なる順序で実行されてもよく、又は工程（又は一連の工程）は、反復的に再実行されてもよい。加えて、本明細書には異なる方法が記載されているが、他の実施態様では、異なる方法（又は異なる方法の工程）を組み合わせてもよいことを理解されたい。

【0716】

いくつかの実施態様では、タスク又は動作を実行するように「構成された」処理ユニット、プロセッサ、モジュール、又はコンピューティングシステムは、タスク又は動作を実行するように特に構造化されていると理解され得る（例えば、タスク又は動作を実行するように調整又は意図された、及び／又はタスク若しくは動作を実行するように調整若しくは意図された1つ又はそれ以上のプログラム又は命令を有すること、及び／又はタスク若しくは動作を実行するように調整又は意図された処理回路の配置を有する）。明確さ及び疑義の回避のために、汎用コンピュータ（適切にプログラムされた場合にタスク又は動作を実行するように構成された）となり得る）汎用コンピュータは、タスク又は動作を実行するために具体的にプログラム又は構造的に変更されない限り、タスク又は動作を実行するように「構成されている」ように構成されていない）。

【0717】

更に、本明細書に記載される方法の操作は、操作が、商業的に妥当な時間期間内に、当業者には、平均的なヒト又は当業者によって実施されることができないように、十分に複雑であり得る。例えば、この方法は、そのような人が商業的に妥当な時間内で方法を完了できないように、比較的複雑な計算に依存し得る。

【0718】

本出願全体を通して、様々な刊行物、特許、又は特許出願が参照されている。これらの出版物の全体の開示は、本発明が属する技術分野の状態をより完全に説明するために、本出願において参照により本明細書に組み込まれる。

【0719】

用語「含む (comprising)」は、本明細書では、列挙された要素のみならず、任意の追加の要素を更に包含する、オープンエンドであることが意図される。

【0720】

本明細書で使用する時、用語「それぞれ」は、項目の集合を参照して使用されるとき、集合内の個々の項目を特定することを意図しているが、必ずしも集合内の全ての項目を指すものではない。明示的な開示又は文脈がそうでないことを明確に指示する場合、例外が生じ得る。

【0721】

上記の実施例を参照して本発明を説明したが、本発明から逸脱することなく様々な修正を行うことができることを理解されたい。

【0722】

本出願のモジュールは、ハードウェア又はソフトウェアで実施態様することができ、図に示されるように、正確に同じブロックで分割される必要はない。いくつかは、異なるプロセッサ若しくはコンピュータ上に実施態様されてもよく、又は多数の異なるプロセッサ若しくはコンピュータの中で広がることもできる。加えて、モジュールの一部は、達成される機能に影響を及ぼすことなく、図に示されるものとは並行して、又は異なる順序で操作され得ることが理解されるであろう。また、本明細書で使用する時、用語「モジュール」は、モジュールを構成するために本明細書で考慮することができる、「サブモジュール」を含むことができる。モジュールとして指定された図のブロックはまた、方法におけるフローチャート工程と考えることができる。

【0723】

本明細書で使用する時、情報項目の「識別」は、その情報の項目の直接仕様を必ずしも必要としない。情報は、単に、一方向の1つ又はそれ以上の層を通じた実際の情報を単に参照することによって、又は情報の実際の項目を決定するのに十分である異なる情報の1つ又はそれ以上のアイテムを識別することによって、フィールド内で「識別され得る」ことができる。加えて、用語「指定する」は、本明細書では、「識別する」と同じであることを意味する。

【0724】

本明細書で使用する時、所与の信号、イベント又は値は、「前デセサー信号、イベント又は前デセサー信号の値、所与の信号、イベント、又は値によって影響されるイベント又は値に依存する。介在処理要素、工程又は期間が存在する場合、所与の信号、イベント、又は値は、「前デセサー信号、イベント又は値」に依存して「存在する」ことができる。介在処理要素又はステップが2つ又はそれ以上の信号、イベント、又は値を組み合わせる場合、処理要素又はステップの信号出力は、「信号、イベント、又は値入力のそれぞれ」に依存していると見なされる。所与の信号、イベント又は値が前デセサー信号、イベント又は値と同じである場合、これは単に、所与の信号、イベント、又は値が「前デセサー信号、イベント又は値」に依存して「依存して」又は「依存して」又は「ベースデセサー信号、イベント又は値」に基づいて、「依存して」又は「依存する」と見なされる。別の信号、イベント又は値に対する所与の信号、イベント、又は値の「応答性」は、同様に定義される。

【0725】

本明細書で使用する、「並行して」又は「並行して」は、正確な同時性を必要としない。個人の1人の評価が、個人の別の評価が完了する前に開始する場合に十分である。

コンピュータシステム

【0726】

図65は、本明細書に開示される技術を実施態様するために配列決定システム800Aによって使用され得るコンピュータシステム6500である。コンピュータシステム6500は、バスサブシステム6555を介して多数の周辺デバイスと通信する、少なくとも1つの中心処理装置(CPU)6572を含む。これらの周辺デバイスは、例えば、メモリデバイス及びファイルストレージサブシステム6536、ユーザーインターフェース入力デバイス6538、ユーザーインターフェース出力デバイス6576及びネットワークインターフェースサブシステム6574を含む記憶サブシステム6510を含むことができる。入力及び出力デバイスは、コンピュータシステム6500とのユーザー対話を可能にする。ネットワークインターフェースサブシステム6574は、他のコンピュータシステム内の対応するインターフェースデバイスへのインターフェースを含む外部ネットワークへのインターフェースを提供する。

10

【0727】

一実施態様では、システムコントローラ7806は、記憶サブシステム6510及びユーザーインターフェース入力デバイス6538に通信可能にリンクされている。

【0728】

ユーザーインターフェース入力デバイス6538は、キーボードと、マウス、トラックボール、タッチパッド、又はグラフィックスタブレットなどのポインティングデバイスと、スキャナーと、ディスプレイに組み込まれたタッチスクリーンと、音声認識システム及びマイクロフォンなどのオーディオ入力デバイスと、他の種類の入力デバイスと、を含むことができる。一般に、用語「入力デバイス」の使用は、コンピュータシステム6500に情報を入力するための全ての可能な種類のデバイス及び方法を含むことを意図する。

20

【0729】

ユーザーインターフェース出力デバイス6576は、ディスプレイサブシステム、プリンタ、ファックス装置、又はオーディオ出力デバイスなどの非視覚ディスプレイを含むことができる。ディスプレイサブシステムは、LEDディスプレイ、陰極線管(CRT)、液晶ディスプレイ(LCD)などのフラットパネルデバイス、投影デバイス、又は可視画像を作成するための何らかの他の機構を含むことができる。ディスプレイサブシステムはまた、音声出力デバイスなどの非視覚ディスプレイを提供することができる。一般に、用語「出力装置」の使用は、コンピュータシステム6500からユーザー又は別のマシン若しくはコンピュータシステムに情報を出力するための、全ての可能な種類のデバイス及び方法を含むことを意図する。

30

【0730】

記憶サブシステム6510は、本明細書に記載されるモジュール及び方法のうちのいくつか又は全ての機能を提供するプログラミング及びデータ構築物を記憶する。これらのソフトウェアモジュールは、概して、深層学習プロセッサ6578によって実行される。

【0731】

深層学習プロセッサ6578は、グラフィック処理ユニット(GPU)、フィールドプログラマブルゲートアレイ(FPGA)、特定用途向け集積回路(ASIC)、及び／又は粗粒化再構成可能構造(CGRA)であり得る。深層学習プロセッサ6578は、Google Cloud Platform(商標)、Xilinx(商標)及びCirrascala(商標)などの深層学習クラウドプラットフォームによってホスティングすることができる。深層学習プロセッサ6578の例としては、Googleのテンソル処理ユニット(TPU)(商標)、GX4 Rackmount Series(商標)、GX65 Rackmount Series(商標)、NVIDIA DGX-1(商標)、Microsoft's Stratix V FPGA(商標)、GraphcoreのIntelligent Processor Unit(IPU)(商標)、Snapdr

40

50

agon processors (商標)、NVIDIAのVolta (商標)、NVIDIAのドライブPX (商標)、NVIDIAのJETSON TX1/TX2 MODULE (商標)、IntelのNirvana (商標)、Movidius VPU (商標)、Fujitsu DPI (商標)、アームDynamicIQ (商標)、IBM TrueNorth (商標)、Lambda GPU Server with Testa V100s (商標)、及び他のものが挙げられる。

【0732】

記憶サブシステム6510で使用されるメモリサブシステム6522は、プログラム実行中に命令及びデータを記憶するためのメインランダムアクセスメモリ(RAM)6532と、固定命令が記憶された読み取り専用メモリ(ROM)6534とを含む多数のメモリを含むことができる。ファイル記憶サブシステム6536は、プログラム及びデータファイルのための永続的な記憶装置を提供することができ、ハードディスクドライブ、関連する取り外し可能な媒体、ドライブ、光学ドライブ、又は取り外し可能な媒体カートリッジを含むことができる。特定の実施態様の機能を実施態様するモジュールは、ストレージサブシステム6510内のファイル記憶サブシステム6536によって、又はプロセッサによってアクセス可能な他のマシン内に記憶され得る。

10

【0733】

バスサブシステム6555は、コンピュータシステム6500の様々な構成要素及びサブシステムを、意図されるように互いに通信するための機構を提供する。バスサブシステム6555は、単一のバスとして概略的に示されているが、バスサブシステムの代替実施態様は、複数のバスを使用することができる。

20

【0734】

コンピュータシステム6500自体は、パーソナルコンピュータ、ポータブルコンピュータ、ワークステーション、コンピュータ端末、ネットワークコンピュータ、テレビ、メインフレーム、サーバファーム、緩く分散した一組の緩くネットワーク化されたコンピュータ、又は任意の他のデータ処理システム若しくはユーザーデバイスを含む様々なタイプのものであり得る。コンピュータ及びネットワークの変化の性質により、図65に示されるコンピュータシステム6500の説明は、本発明の好ましい実施態様を例示する目的のための特定の例としてのみ意図される。コンピュータシステム6500の多くの他の構成は、図65に示されるコンピュータシステムよりも多く又は少ない構成要素を有することができる。

30

【0735】

特定の改善

【0736】

我々は、ニューラルネットワークベースのテンプレート生成及びニューラルネットワークベースのベースコールの様々な実施態様を説明している。実施態様の1つ又はそれ以上の特徴を、ベース実施態様と組み合わせることができる。相互に排他的でない実施態様は、組み合わせ可能であると教示されている。実施態様の1つ又はそれ以上の特徴を他の実施態様と組み合わせることができる。本開示は、これらのオプションのユーザーを定期的に通知する。これらの選択肢を繰り返す列挙のいくつかの実施態様からの省略は、前述のセクションで教示されている組み合わせを制限するものとして解釈されるべきではない。これらの記載は、以下の実施のそれぞれに参照することにより本明細書に組み込まれる。

40

【0737】

ベースコールー単一分析物距離チャンネル

【0738】

配列決定実行中にフローセルのタイル上で合成された分析物をベースコールするニューラルネットワーク実装方法を開示し、配列決定実行は、複数の配列決定サイクルを有し、複数の配列決定サイクルのそれぞれは、1つ又はそれ以上の画像を有する画像セットを生成し、画像のそれぞれは、1つ又はそれ以上の画像チャンネルのうちのそれぞれ1つにおいて分析物及びそれらの周囲背景の強度発光を表現する。方法は、複数の配列決定サイクル

50

のうちの初期のサイクルにおいてそれぞれ生成された初期画像セットをテンプレート生成器によって処理して、テンプレート画像内の分析物の参照中心を識別することを含む。方法は、複数の配列決定サイクルのうちの現在のサイクルにおいて生成された現在の画像セットの、複数の配列決定サイクルのうちの現在のサイクルに先行する複数の配列決定サイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の先行する画像セットの、及び複数の配列決定サイクルのうちの現在のサイクルの後に続く複数の配列決定サイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の後続の画像セットのそれぞれにおける1つ又はそれ以上の画像にアクセスすることを含む。方法は、現在の、先行する、及び後続の画像セット内の画像のそれぞれをテンプレート画像と位置合わせして、サイクル固有及び画像チャンネル固有の変換を決定することを含む。方法は、これらの変換を分析物の参照中心に適用して、画像のそれぞれにおける分析物の変換された中心を識別することを含む。方法は、ベースコールされている分析物のうちの特定の1つについて、各画像パッチが、画像のうちの対応の1つにおいて識別された分析物のうちの特定の1つの変換された中心をその中心ピクセル内に含み、画像チャンネルのうちの対応する1つにおいて分析物のうちの特定の1つの、分析物のうちのいくつかの隣接する分析物の、及びそれらの周囲背景の強度発光を表現するように、現在の、先行する、後続の画像セット内の画像のそれぞれから画像パッチを抽出することを含む。方法は、各画像パッチについて、その中心ピクセルに含まれた分析物のうちの特定の1つの変換された中心からのそのピクセルの中心の距離を識別する距離情報を生成することを含む。方法は、距離情報を各画像パッチ内にピクセル単位で符号化することによって入力データを構築することを含む。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、複数の配列決定サイクルのうちの現在のサイクルにおける分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を生成することを含む。方法は、これらの尤度に基づいて塩基をA、C、T、又はGとして分類することを含む。

【0739】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0740】

一実施態様では、方法は、各画像パッチについて、そのピクセルのうちのどれが分析物のうちの特定の1つを覆い、そのピクセルのうちのどれがこれを覆わないかを識別する分析物帰属情報を生成することと、分析物帰属情報を各画像パッチ内にピクセル単位で符号化することによって入力データを構築することと、を含む。一実施態様では、分析物のうちの特定の1つを覆うピクセルは、分析物帰属情報において非ゼロの値が割り当てられる。一実施態様では、分析物のうちの特定の1つを覆わないピクセルは、分析物帰属情報においてゼロの値が割り当てられる。一実施態様では、方法は、畳み込みニューラルネットワークへの入力として、分析物の変換された中心の位置座標を提供することを含む。そのような一実施態様では、入力は、畳み込みニューラルネットワークの第1の層に供給される。別のそのような実施態様では、入力は、畳み込みニューラルネットワークの1つ又はそれ以上の中間層に供給される。更に別のそのような実施態様では、入力は、畳み込みニューラルネットワークの最終層に供給される。一実施態様では、方法は、畳み込みニューラルネットワークへの入力として、画像パッチのピクセルに対応するスケーリング値を有する強度スケーリングチャンネルを提供することを含む。そのような実施態様では、スケーリング値は、分析物のうちの特定の1つの中心を含む画像パッチの中心ピクセルの平均強度に基づく。一実施態様では、強度スケーリングチャンネルは、画像パッチの全てのピクセルについて同じスケーリング値をピクセル単位で含む。一実施態様では、中心ピクセルの

10

20

30

40

50

平均強度は、画像チャネルのうちの対応する1つのそれぞれについて決定される。

【0741】

一実施態様では、中心ピクセルの平均強度は、第1の画像チャネルについて、分析物のうちの特定の1つに対するA及びTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第2の画像チャネルについて、分析物のうちの特定の1つに対するA及びCベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第1の画像チャネルについて、分析物のうちの特定の1つに対するAベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第2の画像チャネルについて、分析物のうちの特定の1つに対するGベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第3の画像チャネルについて、分析物のうちの特定の1つに対するTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第3の画像チャネルについて、分析物のうちの特定の1つに対するCベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。

【0742】

一実施態様では、配列決定実行は、第1のリードプライマー及び第2のリードプライマーを使用して分析物中の断片の両端部を順方向及び逆方向に配列決定するペアードエンド配列決定を実施し、それによって各断片についてリードペアを生成するものであり、リードペアは、順方向リード及び逆方向リードを有する。一実施態様では、断片の両端部は、連続的に配列決定されて、順方向リード及び逆方向リードを交互に生成する。一実施態様では、断片の両端部は、同時に配列決定されて、順方向リード及び逆方向リードを同時に生成する。一実施態様では、順方向リード及び逆方向リードはそれぞれ、断片のうちの1つ又はそれ以上を含む。一実施態様では、断片のうちの1つ又はそれ以上は、連続的に配列決定される。一実施態様では、断片のうちの1つ又はそれ以上は、同時に配列決定される。一実施態様では、配列決定実行は、単一のリードプライマーを使用して1つの方向で断片を配列決定するシングルリード配列決定を実施する。一実施態様では、配列決定実行は、ループ内で断片の二本鎖コピーを配列決定するサーキュラー配列決定(circular sequencing)を実施し、このループは、所与の断片の二本鎖コピーにわたって複数回反復する。一実施態様では、配列決定実行は、断片の積み重ねられたコピーを配列決定するスタックド配列決定(stacked sequencing)を実施し、所与の断片の積み重ねられたコピーは、垂直又は水平に積み重ねられる。一実施態様では、画像パッチのサイズは、 3×3 ピクセルから 10000×10000 ピクセルの範囲である。

【0743】

一実施態様では、変換された中心は、浮動小数点数座標値である。そのような実施態様では、方法は、丸め操作を使用して浮動小数点数座標値を丸めて、変換された中心の整数座標値を生成することと、中心ピクセルを、その整数座標と変換された中心に対して生成された整数座標値との間の重なりに基づいて識別することと、を含む。一実施態様では、丸め操作は、床関数、天井関数、及び／又は丸め関数のうちの少なくとも1つである。一実施態様では、丸め操作は、整数関数及び／又は正整数関数のうちの少なくとも1つである。一実施態様では、テンプレート生成器は、ニューラルネットワークベースのテンプレート生成器である。一実施態様では、出力層はソフトマックス層であり、尤度は、複数の配列決定サイクルのうちの現在のサイクルにおける分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの指数関数的に正規化されたスコア分布である。

【0744】

一実施態様では、画像チャネルのそれぞれ1つは、複数のフィルタ波長帯域のうちの1つに対応する。別の実施態様では、画像チャネルのそれぞれ1つは、複数の画像イベントのうちの1つに対応する。一実施態様では、フローセルは、分析物を占有するウェルのアレイを有する、少なくとも1つのパターン化された表面を有する。別の実施態様では、フローセルは、少なくとも1つのパターン化されていない表面を有し、分析物は、非パターン化表面上で不均一に散乱される。一実施態様では、画像セットは、4つの画像を有する。別の実施態様では、画像セットは2つの画像を有する。更に別の実施態様では、画像セットは1つの画像を有する。一実施態様では、配列決定動作は、4チャンネル化学を利用する。別の実施態様では、配列決定動作は、2チャンネル化学を利用する。更に別の実施態様では、配列決定実行は、1チャンネル化学を利用する。

10

【0745】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0746】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。入力データは、配列決定実行の現在の配列決定サイクルにおいて生成された現在の画像セットの、現在の配列決定サイクルに先行する配列決定実行の1つ又はそれ以上の配列決定サイクルにおいてそれぞれ生成された1つ又はそれ以上の先行する画像セットの、及び現在の配列決定サイクルの後に続く配列決定実行の1つ又はそれ以上の配列決定サイクルにおいてそれぞれ生成された1つ又はそれ以上の後続の画像セットのそれぞれで1つ又はそれ以上の画像から抽出された画像パッチを含む。画像パッチのそれぞれは、ベースコールされているターゲット分析物の、いくつかの隣接する分析物の、及びそれらの周囲背景の強度発光を対応する画像チャネル内で表現する。入力データは、画像パッチの中心ピクセル内に位置するターゲット分析物の中心からの画像パッチのピクセルの中心の距離を識別するために画像パッチのそれぞれにピクセル単位で符号化されている距離情報を更に含む。方法は、畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおけるターゲット分析物を出力に基づいてベースコールすることを含む。

20

30

【0747】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0748】

一実施態様では、方法は、畳み込み表現を出力層によって処理して、現在の配列決定サイクルにおけるターゲット分析物に組み込まれた塩基がA、C、T、及びGであることの尤度を生成することと、塩基をこれらの尤度に基づいてA、C、T、又はGとして分類することと、を含む。一実施態様では、尤度は、ソフトマックス層によって生成される指数関数的に正規化されたスコアである。

40

【0749】

一実施態様では、方法は、出力から、現在の配列決定サイクルにおけるターゲット分析物に組み込まれた塩基がA、C、T、又はGであることのクラスラベルを識別するターゲット分析物の出力ペアを導出することと、クラスラベルに基づいてターゲット分析物をベースコールすることと、を含む。一実施態様では、1, 0のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、1, 1のクラスラベルがT塩基を識別し、

50

0, 0のクラスラベルがG塩基を識別する。別の実施態様では、1, 1のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、0.5, 0.5のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。更に別の実施態様では、1, 0のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、0.5, 0.5のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。また更なる実施態様では、1, 2のクラスラベルがA塩基を識別し、0, 1のクラスラベルがC塩基を識別し、1, 1のクラスラベルがT塩基を識別し、0, 0のクラスラベルがG塩基を識別する。一実施態様では、方法は、出力から、現在の配列決定サイクルにおけるターゲット分析物に組み込まれた塩基がA、C、T、又はGであることを識別するターゲット分析物のためのクラスラベルを導出することと、クラスラベルに基づいてターゲット分析物をベースコールすることと、を含む。一実施態様では、0.33のクラスラベルがA塩基を識別し、0.66のクラスラベルがC塩基を識別し、1のクラスラベルがT塩基を識別し、0のクラスラベルがG塩基を識別する。別の実施態様では、0.50のクラスラベルがA塩基を識別し、0.75のクラスラベルがC塩基を識別し、1のクラスラベルがT塩基を識別し、0.25のクラスラベルがG塩基を識別する。一実施態様では、方法は、出力から、単一の出力値を導出することと、単一の出力値を塩基A、C、T、及びGに対応するクラス値範囲と比較することと、比較に基づいて、単一の出力値を特定のクラス値範囲に割り当てることと、割り当てに基づいてターゲット分析物をベースコールすることと、を含む。一実施態様では、単一の出力値はシグモイド関数を使用して導出され、単一の出力値は0~1の範囲である。別の実施態様では、0~0.25のクラス値範囲がA塩基を表し、0.25~0.50のクラス値範囲がC塩基を表し、0.50~0.75のクラス値範囲がT塩基を表し、0.75~1のクラス値範囲がG塩基を表す。

【0750】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0751】

配列決定実行中にフローセルのタイル上で合成された分析物をベースコールするニューラルネットワーク実装方法を開示し、配列決定実行は、複数の配列決定サイクルを有し、複数の配列決定サイクルのそれぞれは、1つ又はそれ以上の画像を有する画像セットを生成し、画像のそれぞれは、1つ又はそれ以上の画像チャネルのうちのそれぞれ1つにおいて分析物及びそれらの周囲背景の強度発光を表現する。方法は、複数の配列決定サイクルのうちの初期のサイクルにおいてそれぞれ生成された初期画像セットをテンプレート生成器によって処理して、テンプレート画像内の分析物の参照中心を識別することを含む。方法は、複数の配列決定サイクルのうちの現在のサイクルにおいて生成された現在の画像セットの、複数の配列決定サイクルのうちの現在のサイクルに先行する複数の配列決定サイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の先行する画像セットの、及び複数の配列決定サイクルのうちの現在のサイクルの後に続く複数の配列決定サイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の後続の画像セットのそれぞれにおける1つ又はそれ以上の画像にアクセスすることを含む。方法は、現在の、先行する、及び後続の画像セット内の画像のそれぞれをテンプレート画像と位置合わせして、サイクル固有及び画像チャネル固有の変換を決定することを含む。方法は、これらの変換を分析物の参照中心に適用して、画像のそれぞれにおける分析物の変換された中心を識別することを含む。方法は、ベースコールされている分析物のうちの特定の1つについて、各画像パッチが、画像のうちの対応の1つにおいて識別された分析物のうちの特定の1つの変換された中心をその中心ピクセル内に含み、画像チャネルのうちの対応する1つにおいて分析物のうちの特定の1つの、分析物のうちのいくつかの隣接する分析物の、及びそれらの周囲背景の強度発光を表現するように、現在の、先行す

る、後続の画像セット内の画像のそれぞれから画像パッチを抽出することを含む。方法は、各画像パッチについて、その中心ピクセルに含まれた分析物のうちの特定の1つの変換された中心からのそのピクセルの中心の距離を識別する距離情報を生成することを含む。方法は、距離情報を各画像パッチ内にピクセル単位で符号化することによって入力データを構築することを含む。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、複数の配列決定サイクルのうち現在のサイクルにおける分析物のうちの特定の1つを出力に基づいてベースコールすることを含む。

【0752】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0753】

一実施態様では、方法は、畳み込み表現を出力層によって処理して、複数の配列決定サイクルのうち現在のサイクルにおける分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を生成することと、塩基をこれらの尤度に基づいてA、C、T、又はGとして分類することと、を含む。

【0754】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0755】

一実施態様では、コンピュータ実装方法は、入力データをニューラルネットワークによって処理することと、入力データの代替表現を生成することと、を含む。入力データは、配列決定実行の1つ又はそれ以上の配列決定サイクルのそれぞれのためのサイクルごとの画像データを含む。サイクルごとの画像データは、それぞれの配列決定サイクルにおいて捕捉された1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する。方法は、代替表現を出力層によって処理することと、出力を生成することと、を含む。方法は、配列決定サイクルのうち1つ又はそれ以上における分析物のうち1つ又はそれ以上を出力に基づいてベースコールすることを含む。

【0756】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0757】

一実施態様では、方法は、サイクルごとの画像データのピクセルと分析物のうち1つ又はそれ以上との間の距離を識別する補足的な距離情報をサイクルごとの画像データに添付することを含む。そのような実施態様では、距離は、分析物のうち1つ又はそれ以上の中心、形状、及び／又は境界に関するコンテキストをニューラルネットワーク及び出力層による処理に組み込む。一実施態様では、方法は、サイクルごとの画像データのピクセルにスケーリング値を割り当てる補足的なスケーリング情報をサイクルごとの画像データに添付することを含む。そのような実施態様では、スケーリング値は、分析物のうち1つ又はそれ以上の強度の変動を考慮する。

10

20

30

40

50

【0758】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

ベースコールー多分析物距離チャンネル

【0759】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスを含む入力データにアクセスすることを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。各画像パッチは、複数の分析物及びそれらの周囲背景を覆うピクセルのためのピクセル強度データと、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される分析物のうちの最も近い分析物からの各ピクセルの中心間距離を識別するピクセル距離データと、を有する。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、現在の配列決定サイクルにおける分析物のうちのそれぞれ1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別する、分析物のそれぞれのためのスコア分布を生成することを含む。方法は、分析物のそれぞれをこれらの尤度に基づいてベースコールすることを含む。

【0760】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0761】

一実施態様では、ピクセル距離データは、各画像パッチ内にピクセル単位で符号化される。一実施態様では、中心間距離は、分析物の変換された中心の位置座標及びピクセル中心の位置座標を使用する距離式から導出される。一実施態様では、方法は、畳み込みニューラルネットワークへの入力として、各画像パッチのピクセルに対応するスケーリング値を有する強度スケーリングチャンネルを提供することを含み、スケーリング値は、分析物の変換された中心を含む各画像パッチ内の中心ピクセルの平均強度の組み合わせに基づく。一実施態様では、強度スケーリングチャンネルは、画像パッチの全てのピクセルのピクセル強度データに同じスケーリング値をピクセル単位で適用する。一実施態様では、強度スケーリングチャンネルは、第1の中心ピクセルの平均強度から導出された第1のスケーリング値が、第1の中心ピクセルに連続的に隣接している隣接ピクセルの第1のピクセル近傍に適用され、別の中心ピクセルの平均強度から導出される別のスケーリング値が、別の中心ピクセルに連続的に隣接している隣接ピクセルの別のピクセル近傍に適用されるように、画像パッチのピクセルのピクセル強度データにピクセル近傍に基づいて異なるスケーリング値をピクセル単位で適用する。一実施態様では、ピクセル近傍は、中心ピクセルを中心とする $m \times n$ のピクセルパッチであり、ピクセルパッチは 3×3 ピクセルである。一実施態様では、ピクセル近傍は、中心ピクセルを中心とする n 接続されたピクセル近傍である。一実施態様では、中心ピクセルの平均強度は、画像チャンネルのうちの対応する1つのそれぞれについて決定される。一実施態様では、中心ピクセルの平均強度は、第1の画像チャンネルについて、分析物のうちの対応の分析物に対するA及びTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第2の画像チャンネルについて、分析物のうちの対応の分析物に対するA及びCベースコールを生成した2

つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第1の画像チャンネルについて、分析物のうちの対応の分析物に対するAベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第2の画像チャンネルについて、分析物のうちの対応の分析物に対するGベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第3の画像チャンネルについて、分析物のうちの対応の分析物に対するTベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、中心ピクセルの平均強度は、第3の画像チャンネルについて、分析物のうちの対応の分析物に対するCベースコールを生成した2つ以上の先行する配列決定サイクル中に観察された中心ピクセルの強度値を平均化することによって決定される。一実施態様では、方法は、各画像パッチについて、そのピクセルのうちのどれが分析物を覆い、そのピクセルのうちのどれがこれらを覆わないかを識別する分析物帰属情報を生成することと、分析物帰属情報を各画像パッチ内にピクセル単位で符号化することによって入力データを構築することと、を含む。一実施態様では、分析物を覆うピクセルは、分析物帰属情報において非ゼロの値が割り当てられる。一実施態様では、分析物を覆わないピクセルは、分析物帰属情報においてゼロの値が割り当てられる。一実施態様では、各画像パッチのサイズは、 3×3 ピクセルから 10000×10000 ピクセルの範囲である。一実施態様では、出力層はソフトマックス層であり、スコア分布は指数関数的に正規化されたスコア分布である。

【0762】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0763】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスを含む入力データにアクセスすることを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。各画像パッチは、複数の分析物及びそれらの周囲背景を覆うピクセルのためのピクセル強度データと、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される分析物のうちの最も近い分析物からの各ピクセルの中心間距離を識別するピクセル距離データと、を有する。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベース

【0764】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0765】

一実施態様では、方法は、出力から、現在の配列決定サイクルにおける分析物のそれぞれ1つに組み込まれた塩基がA、C、T、及びGであることを識別する分析物のそれぞれ

のためのスコア分布を導出することと、分析物のそれぞれを尤度に基づいてベースコールすることと、を含む。一実施態様では、出力層はソフトマックス層であり、スコア分布は指数関数的に正規化されたスコア分布である。一実施態様では、方法は、出力から、現在の配列決定サイクルにおける分析物のそれぞれ1つに組み込まれた塩基がA、C、T、及びGであることを識別する分析物のそれぞれのための出力ペアを導出することと、分析物のそれぞれをクラスラベルに基づいてベースコールすることと、を含む。一実施態様では、方法は、出力から、単一の出力値を導出することと、単一の出力値を塩基A、C、T、及びGに対応するクラス値範囲と比較することと、比較に基づいて、単一の出力値を特定のクラス値範囲に割り当てることと、分析物のそれぞれを割り当てに基づいてベースコールすることと、を含む。一実施態様では、単一の出力値はシグモイド関数を使用して導出され、単一の出力値は0～1の範囲である。

10

【0766】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

ベースコールー多分析物形状ベース距離チャンネル

【0767】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスを含む入力データにアクセスすることを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。各画像パッチは、分析物強度を表現する分析物ピクセル及び背景強度を表現する背景ピクセルを使用して複数の分析物及びそれらの周囲背景の強度発光を表現する。各画像パッチは、分析物のうちの1つのみに対して各分析物ピクセルを分類することに基づいて選択される分析物のうちの割り当てられた分析物からの各分析物ピクセルの中心間距離を識別する分析物距離データと共に符号化される。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、現在の配列決定サイクルにおける分析物のうちのそれぞれ1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別する、分析物のそれぞれのためのスコア分布を生成することを含む。方法は、分析物のそれぞれをこれらの尤度に基づいてベースコールすることを含む。

20

30

【0768】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

40

【0769】

一実施態様では、分析物は、複数の分析物ピクセルにわたる不規則な形状を有し、ピクセルー分析物間の分類は、不規則な形状に基づく。一実施態様では、全ての背景ピクセルは、分析物距離データ内で同じ最小中心間距離が割り当てられる。一実施態様では、全ての背景ピクセルは、同じ最小強度が割り当てられる。一実施態様では、各分析物ピクセルは、ニューラルネットワークベースのテンプレート生成器によって生成された減衰マップに基づいて、分析物のうちの1つのみに分類される。そのような実施態様では、減衰マップは、分析物を隣接ピクセルからなる不連続領域として、分析物の中心を不連続領域のうちの対応する領域の質量中心にある中心ピクセルとして、及びそれらの周囲背景を不連続領域のいずれにも属していない背景ピクセルとして表現する。一実施態様では、不連続領

50

域のうちの対応する領域内の隣接ピクセルは、隣接ピクセルが属する不連続領域内の中心ピクセルからの隣接ピクセルの距離に従って重み付けされた強度値を有する。一実施態様では、不連続領域のうちの対応する領域内の隣接ピクセルは、同じ分析物に属し、かつそれを共同で表現する分析物内部ピクセルとしてカテゴリー化され、分析物ごとにメモリに記憶される。一実施態様では、中心ピクセルは、不連続領域のうちの対応する領域内で最も高い強度値を有する。一実施態様では、背景ピクセルは全て、減衰マップ内で同じ最低強度値を有する。一実施態様では、分析物距離データは、各画像パッチ内へのピクセル単位の符号化である。一実施態様では、中心間距離は、分析物の変換された中心の位置座標及びピクセル中心の位置座標を使用する距離式から導出される。一実施態様では、分析物の変換された中心は、減衰マップによって識別された分析物の中心にサイクル固有の変換及び画像チャンネル固有の変換を適用することによって導出される。

10

【0770】

一実施態様では、方法は、畳み込みニューラルネットワークへの入力として、各画像パッチのピクセルに対応するスケーリング値を有する強度スケーリングチャンネルを提供することを含む。そのような実施態様では、スケーリング値は、分析物の変換された中心を含む各画像パッチ内の中心ピクセルの平均強度の組み合わせに基づく。一実施態様では、強度スケーリングチャンネルは、第1の分析物の中心を含む第1の中心ピクセルの平均強度から導出された第1のスケーリング値が、第1の分析物に属し、かつそれを共同で表現する隣接ピクセルの第1のピクセルグループに適用され、別の分析物の中心を含む別の中心ピクセルの平均強度から導出された別のスケーリング値が、別の分析物に属し、かつそれを共同で表現する隣接ピクセルの別のピクセルグループに適用されるように、画像パッチのピクセルのピクセル強度データにピクセルグループに基づいて異なるスケーリング値をピクセル単位で適用する。一実施態様では、中心ピクセルの平均強度は、画像チャンネルのうちの対応する1つのそれぞれについて決定される。一実施態様では、方法は、各画像パッチについて、そのピクセルのうちのどれが分析物を覆い、そのピクセルのうちのどれがこれらを覆わないかを識別する分析物帰属情報を生成することと、分析物帰属情報を各画像パッチ内にピクセル単位で符号化することによって入力データを構築することと、を含む。一実施態様では、分析物を覆うピクセルは、分析物帰属情報において非ゼロの値が割り当てられる。別の実施態様では、分析物を覆わないピクセルは、分析物帰属情報においてゼロの値が割り当てられる。

20

30

【0771】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0772】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスを含む入力データにアクセスすることを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャンネルのそれぞれ1つのための画像パッチを有する。各画像パッチは、分析物強度を表現する分析物ピクセル及び背景強度を表現する背景ピクセルを使用して複数の分析物及びそれらの周囲背景の強度発光を表現する。各画像パッチは、分析物のうちの1つのみに対して各分析物ピクセルを分類することに基づいて選択される分析物のうちの割り当てられた分析物からの各分析物ピクセルの中心間距離を識別する分析物距離データと共に符号化される。方法は、入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。

40

50

【0773】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

10

【0774】

特殊化したアーキテクチャ

【0775】

互いに対してレジストレーション誤差を有する配列決定画像を使用して分析物をベースコールするネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスにアクセスすることを含む。シーケンスは、サイクルごとの画像パッチセット間にわたって、及びサイクルごとの画像パッチセット内で画像パッチ間のレジストレーション誤差を有する。シーケンス内の各画像パッチは、一連内の対応する配列決定サイクルにおける対応する画像チャンネル内において、ベースコールされているターゲット分析物の、いくつかの隣接する分析物の、及びそれらの周囲背景の強度情報を表現する。シーケンス内の各画像パッチは、その中心ピクセル内に位置するターゲット分析物の中心からのそのピクセルの中心の距離を識別する距離情報と共に符号化される。方法は、第1の畳み込みサブネットワークによって各サイクルごとの画像パッチセットを別個に処理して、各配列決定サイクルのための中間畳み込み表現を生成することを含み、これは、強度及び距離情報を組み合わせ、かつ結果として得られる畳み込み表現を配列決定サイクル内でのみ組み合わせ、かつ配列決定サイクル間では組み合わせない畳み込みを適用することを含む。方法は、第2の畳み込みサブネットワークによって一連内の継続的な配列決定サイクルのための中間畳み込み表現をグループ単位で処理して、一連のための最終畳み込み表現を生成することを含み、これは、中間畳み込み表現を組み合わせ、かつ結果として得られる畳み込み表現を配列決定サイクルの間で組み合わせる畳み込みを適用することを含む。方法は、最終畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおけるターゲット分析物を出力に基づいてベースコールすることを含む。

20

30

【0776】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

40

【0777】

一実施態様では、シーケンス内の各画像パッチは、複数の分析物及びそれらの周囲背景を覆うピクセルのためのピクセル強度データと、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される分析物のうちの最も近い分析物からの各ピクセルの中心間距離を識別するピクセル距離データと、を有する。そのような実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、シーケンス内の各画像パッチは、分析物強度を表現する分析物ピクセル及び背景強度を表現する背景ピクセルを使用して複数の分析物及びそれらの周囲背景の強度発光を表現し、各分析物ピクセルを分析物のうちの1つのみに分類することに基づいて選択される分析物のうちの割り当てられた分析物からの各分析物ピクセルの中心

50

間距離を識別する分析物距離データと共に符号化される。そのような実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、方法は、第1の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、第2の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、出力層への入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。

【0778】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0779】

レジストレーション誤差を有する画像データを使用して分析物をベースコールするネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのための入力データにアクセスすることを含む。入力データは、各配列決定サイクルのための画像テンソルを有する。各画像テンソルは、1つ又はそれ以上の画像チャンネルのためのデータを有し、これには、画像チャンネルごとに、ベースコールされているターゲット分析物、いくつかの隣接する分析物、及び周囲背景を覆うピクセルのためのピクセル強度データ、並びにターゲット分析物の中心からピクセルの中心までの距離のためのピクセル距離データが含まれる。入力データは、画像テンソル間にわたってのピクセル間のクロスサイクルのレジストレーション誤差と、画像テンソル内でのピクセル間のクロス画像チャンネルのレジストレーション誤差と、を有する。方法は、各入力テンソルを、空間的な畳み込み層のシーケンスを有する空間的な畳み込みネットワークによって別個に処理して、各配列決定サイクルのための空間畳み込み表現を生成することを含み、これは、ピクセル強度及び距離を配列決定サイクル内でのみ組み合わせ、配列決定サイクル間では組み合わせない、第1の空間的な畳み込み層から始めることと、先行する空間的な畳み込み層の出力を一連の配列決定サイクル内の各配列決定サイクル内でのみ組み合わせ、配列決定サイクル間では組み合わせない、継続的な空間的な畳み込み層に進むことと、を含む。方法は、継続的な配列決定サイクルのための空間畳み込み表現を、時間的な畳み込み層のシーケンスを有する時間的な畳み込みネットワークによってグループ単位で処理して、これらの一連のための時間的な畳み込み表現を生成することを含み、これは、空間畳み込み表現を一連の配列決定サイクル内の配列決定サイクルの間で組み合わせる第1の時間的な畳み込み層から始めることと、先行する時間的な畳み込み層の継続的な出力を組み合わせる継続的な時間的な畳み込み層に進むことと、を含む。方法は、時間的な畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおけるターゲット分析物を出力に基づいてベースコールすることを含む。

【0780】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0781】

一実施態様では、グループ単位で処理することは、重複するスライディングウィンドウ内の継続的な中間畳み込み表現にわたって畳み込みを行うことを更に含む。一実施態様では、継続的な時間的な畳み込み層は、重複するスライディングウィンドウ内の継続的な出力を組み合わせる。一実施態様では、ピクセル距離データは、各画像テンソル内へのピク

セル単位の符号化である。一実施態様では、シーケンス内の各画像テンソルは、複数の分析物及びそれらの周囲背景を覆うピクセルのためのピクセル強度データと、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される分析物のうちの最も近い分析物からの各ピクセルの中心間距離を識別するピクセル距離データと、を有する。一実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、シーケンス内の各画像テンソルは、分析物強度を表現する分析物ピクセル及び背景強度を表現する背景ピクセルを使用して複数の分析物及びそれらの周囲背景の強度発光を表現し、各分析物ピクセルを分析物のうちの1つのみに分類することに基づいて選択される分析物のうちの割り当てられた分析物からの各分析物ピクセルの中心間距離を識別する分析物距離データと共に符号化される。一実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、方法は、第1の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、第2の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、出力層への入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。

【0782】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0783】

リフレーム

【0784】

配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、配列決定実行の一連の配列決定サイクルのために生成されたサイクルごとの画像パッチセットのシーケンスにアクセスすることを含む。シーケンス内の各サイクルごとの画像パッチセットは、1つ又はそれ以上の画像チャネルのそれぞれ1つのための画像パッチを有する。各画像パッチは、ベースコールされているターゲット分析物、いくつかの隣接する分析物、及び周囲背景を覆うピクセルのためのピクセル強度データを有する。方法は、各画像パッチのピクセルをリフレームして、ターゲット分析物の中心を中心ピクセルの中心に配置することを含む。方法は、リフレームされた画像パッチを畳み込みニューラルネットワークによって畳み込んで、リフレームされた画像パッチの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、出力を生成することを含む。方法は、現在の配列決定サイクルにおけるターゲット分析物を出力に基づいてベースコールすることを含む。

【0785】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0786】

一実施態様では、リフレームすることは、リフレームすることを補正するための各画像パッチのピクセルの強度補間を更に含む。一実施態様では、強度補間は、最近傍強度抽出、ガウシアンベースの強度抽出、 2×2 サブピクセル領域の平均に基づく強度抽出、 2×2 サブピクセル領域の最大輝度に基づく強度抽出、 3×3 サブピクセル領域の平均に基づく強度抽出、バイリニア強度抽出、バイキュービック強度抽出、及び／又は重み付け面積

カバー率に基づく強度抽出のうちの少なくとも1つを更に含む。一実施態様では、リフレーションする前に、ターゲット分析物の中心は、各画像パッチの中心ピクセル内で中心ピクセルの中心から外れた位置にある。一実施態様では、リフレーションすることは、各画像パッチの非中心ピクセルが、ターゲット分析物のそれぞれの中心から等距離であることを必要とすることを更に含む。一実施態様では、シーケンス内の各画像パッチは、複数の分析物及びそれらの周囲背景を表現するピクセルのためのピクセル強度データと、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される分析物のうちの最も近い分析物からの各ピクセルの中心間距離を識別するピクセル距離データと、を有する。一実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、シーケンス内の各画像パッチは、分析物強度を表現する分析物ピクセル及び背景強度を表現する背景ピクセルを使用して複数の分析物及びそれらの周囲背景の強度発光を表現し、各分析物ピクセルを分析物のうちの1つのみに分類することに基づいて選択される分析物のうちの割り当てられた分析物からの各分析物ピクセルの中心間距離を識別する分析物距離データと共に符号化される。一実施態様では、方法は、現在の配列決定サイクルにおける分析物のそれぞれを出力に基づいてベースコールすることを含む。一実施態様では、方法は、第1の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、第2の畳み込みサブネットワークへの入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。一実施態様では、方法は、出力層への入力として、ターゲット分析物及び／又は隣接する分析物の位置座標を提供することを含む。

10

20

【0787】

フローセル上の分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、フローセル上の分析物を合成する配列決定実行の複数の配列決定サイクルにわたって生成された画像セットのシーケンスにアクセスすることを含む。画像セットのシーケンス内の各画像は、フローセルの非重複領域を覆い、複数の配列決定サイクルのうちのそれぞれ1つにおける対応する画像チャンネル内に捕捉された非重複領域上の分析物のサブセット及びそれらの周囲背景の強度発光を表現する。方法は、複数の配列決定サイクルのうちの特定の1つにおける分析物のサブセットのうちの特定の1つに組み込まれたヌクレオチド塩基（A、C、T、又はG）を、画像セットのシーケンスから、複数の配列決定サイクルのうちの特定の1つにおいて生成された現在の画像セット、複数の配列決定サイクルのうちの特定の1つに先行する複数のシーケンスサイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の先行する画像セット、及び複数の配列決定サイクルのうちの特定の1つの後に続く複数の配列決定サイクルのうちの1つ又はそれ以上においてそれぞれ生成された1つ又はそれ以上の後続の画像セットを選択することによって決定することを含む。方法は、選択された画像セットのそれぞれの中の画像から画像パッチを抽出することを含む。画像パッチは、分析物のサブセットのうちの特定の1つを中心とし、分析物のサブセットからの追加の隣接する分析物を含む。方法は、画像パッチを畳み込みニューラルネットワークの1つ又はそれ以上の層によって畳み込んで、画像パッチの畳み込み表現を生成することを含む。方法は、畳み込み表現を出力層によって処理して、ヌクレオチド塩基がA、C、T、及びGであることの尤度を生成することを含む。方法は、これらの尤度に基づいてヌクレオチド塩基をA、C、T、又はGとして分類することを含む。

30

40

【0788】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0789】

50

一実施態様では、方法は、複数の配列決定サイクルのそれぞれについて選択すること、抽出すること、畳み込むこと、処理すること、及び分類することを反復することによって複数の配列決定サイクルにわたって分析物のサブセットのうちの特定の1つのためのベースコールの配列を生成することを含む。一実施態様では、方法は、サブセット内の複数の分析物のそれぞれのために複数の配列決定サイクルのそれぞれについて選択すること、抽出すること、畳み込むこと、処理すること、及び分類することを反復することによって複数の配列決定サイクルにわたってサブセット内の複数の分析物のためのベースコールの配列を生成することを含む。一実施態様では、フローセルの非重複領域はタイルである。一実施態様では、対応する画像チャネルは、複数のフィルタ波長帯域のうちの1つである。一実施態様では、対応する画像チャネルは、複数の画像イベントのうちの1つである。

10

【0790】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

複数のサイクルにおける複数のクラスターの同時ベースコール

【0791】

フローセル上の分析物をベースコールするニューラルネットワーク実装方法を開示する。方法は、画像セットのシーケンスから入力画像データを取得することを含む。画像セットのシーケンスは、フローセル上の分析物を合成する配列決定実行の複数の配列決定サイクルにわたって生成される。画像セットのシーケンス内の各画像は、フローセルの非重複領域を覆い、複数の配列決定サイクルのうちのそれぞれ1つにおける対応する画像チャネル内に捕捉された非重複領域上の分析物のサブセット及びそれらの周囲背景の強度発光を表現する。方法は、入力画像データをニューラルネットワークの1つ又はそれ以上の層によって処理して、入力画像データの代替表現を生成することを含む。方法は、代替表現を出力層によって処理して、複数の配列決定サイクルのそれぞれのそれぞれにおけるサブセット内の分析物のうちの少なくともいくつかに組み込まれたヌクレオチド塩基（A、C、T、又はG）を識別する出力を生成し、それによって、複数の配列決定サイクルにわたってサブセット内の分析物のうちの少なくともいくつかのためのベースコールの配列を生成することを含む。

20

30

【0792】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0793】

一実施態様では、出力層はソフトマックス層であり、出力は、複数の配列決定サイクルのそれぞれにおいてサブセット内の分析物のうちの少なくともいくつかのそれぞれに組み込まれたヌクレオチド塩基がA、C、T、及びGであることの指数関数的に正規化されたスコア分布である。一実施態様では、入力画像データは、画像セットのシーケンス内の画像を含む。一実施態様では、入力画像データは、画像セットのシーケンス内の画像のそれぞれからの少なくとも1つの画像パッチを含む。一実施態様では、ニューラルネットワークは、畳み込みニューラルネットワークである。別の実施態様では、ニューラルネットワークは、残差ニューラルネットワークである。更に別の実施態様では、ニューラルネットワークは、反復ニューラルネットワークである。

40

【0794】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体

50

を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0795】

反復畳み込みベースのベースコール

【0796】

ベースコールするためのニューラルネットワークベースのシステムを開示する。システムは、反復モジュール及び畳み込みモジュールを有するハイブリッドニューラルネットワークを含む。反復モジュールは、畳み込みモジュールからの入力を使用する。畳み込みモジュールは、配列決定実行の一連の配列決定サイクルのための画像データを1つ又はそれ以上の畳み込み層によって処理し、画像データの1つ又はそれ以上の畳み込み表現を生成する。画像データは、1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する。反復モジュールは、畳み込み表現及び前の隠れ状態表現を畳み込むことに基づいて現在の隠れ状態表現を生成する。出力モジュールは、現在の隠れ状態表現に基づいて、分析物のうちの少なくとも1つのための、かつ配列決定サイクルのうちの少なくとも1つのためのベースコールを生成する。

【0797】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0798】

ベースコールするニューラルネットワーク実装方法を開示する。方法は、サイクルごとの入力データのシーケンス内の各サイクルごとの入力データを畳み込みニューラルネットワークの畳み込み層のカスケードによって別個に処理することを含む。サイクルごとの入力データのシーケンスは、配列決定実行の一連の配列決定サイクルのために生成され、各サイクルごとの入力データは、それぞれの配列決定サイクルにおいて捕捉された1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する画像チャネルを含む。方法は、各配列決定サイクルについて、別個に処理することに基づいて、畳み込み層のそれぞれにおいて畳み込み表現を生成し、それによって畳み込み表現のシーケンスを生成することと、そのサイクルごとの入力データをその対応する畳み込み表現のシーケンスと混合し、混合表現を生成することと、その混合表現を平坦化し、平坦化された混合表現を生成することと、を含む。方法は、継続的な配列決定サイクルの平坦化された混合表現をスタックとして配列することを含む。方法は、スタック内の平坦化された混合表現のサブセットにわたってスライディングウィンドウに基づいて畳み込みを行い（各スライディングウィンドウはそれぞれの配列決定サイクルに対応する）、（i）スタック上の現在のスライディングウィンドウ内の平坦化された混合表現のサブセット及び（ii）前の隠れ状態表現に基づいて、各配列決定サイクルのための各時間ステップにおける現在の隠れ状態表現を継続的に生成する反復ニューラルネットワークによって、スタックを順方向及び逆方向に処理することを含む。方法は、スタックを順方向及び逆方向に処理することの結果に基づいて、配列決定サイクルのそれぞれにおける分析物のそれぞれをベースコールすることを含む。反復ニューラルネットワークは、LSTM及びGRUなど、ゲート付き反復ニューラルネットワークとすることができる。

【0799】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

10

20

30

40

50

【0800】

方法は、所与の配列決定サイクルの順方向及び逆方向の現在の隠れ状態表現を時間ステップに基づいて組み合わせ、組み合わせられた隠れ状態表現を生成すること、組み合わせられた隠れ状態表現を1つ又はそれ以上の全結合ネットワークによって処理し、高密度表現を生成すること、高密度表現をソフトマックス層によって処理して、所与の配列決定サイクルにおける分析物のそれぞれに組み込まれた塩基がA、C、T、及びGであることの尤度を生成すること、並びにこれらの尤度に基づいて塩基をA、C、T、又はGとして分類することによって、所与の配列決定サイクルにおける分析物のそれぞれをベースコールすることを含む。一実施態様では、組み合わせることは連結を含む。別の実施態様では、組み合わせることは総和を含む。更に別の実施態様では、組み合わせることは平均化を含む。

10

【0801】

一実施態様では、各サイクルごとの入力データは、画像チャンネルを補足し、対応する画像チャンネル内のピクセルと1つ又はそれ以上の分析物中心との間の中心間距離を含む、距離チャンネルを含む。一実施態様では、各サイクルごとの入力データは、画像チャンネルを補足し、画像チャンネル内の1つ又はそれ以上のピクセルの平均強度に基づいたスケーリング値を含む、スケーリングチャンネルを含む。一実施態様では、混合することは、畳み込み表現及びサイクルごとの入力データを連結することを更に含む。一実施態様では、混合することは、畳み込み表現及びサイクルごとの入力データの総和をとることを更に含む。一実施態様では、平坦化された混合表現は二次元アレイである。一実施態様では、平坦化された混合表現のサブセットは、三次元体積である。一実施態様では、反復ニューラルネットワークは、三次元畳み込みを三次元体積に適用する。一実施態様では、三次元畳み込みは、SAMEパディングを使用する。一実施態様では、畳み込み層は、SAMEパディングを使用する。一実施態様では、反復ニューラルネットワークは、入力ゲート、活性化ゲート、忘却ゲート、及び出力ゲートを含む長・短期記憶(LSTM)ネットワークである。そのような実施態様では、方法は、(i)スタック上の現在のスライディングウィンドウ内の平坦化された混合表現のサブセット及び(ii)前の隠れ状態表現を、入力ゲート、活性化ゲート、忘却ゲート、及び出力ゲートによって処理することと、各配列決定サイクルのための各時間ステップにおける現在の隠れ状態表現を生成することと、を含む。入力ゲート、活性化ゲート、忘却ゲート、及び出力ゲートは、(i)スタック上の現在のスライディングウィンドウ内の平坦化された混合表現のサブセット及び(ii)前の隠れ状態表現に畳み込みを適用する。

20

30

【0802】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0803】

一実施態様では、ベースコールするニューラルネットワーク実装方法は、配列決定実行の一連の配列決定サイクルのための画像データを畳み込みモジュールの1つ又はそれ以上の畳み込み層によって畳み込むことと、画像データの1つ又はそれ以上の畳み込み表現を生成することとを含む。画像データは、1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する。方法は、畳み込み表現及び前の隠れ状態表現を反復モジュールによって畳み込むことと、現在の隠れ状態表現を生成することとを含む。方法は、現在の隠れ状態表現を出力モジュールによって処理することと、分析物のうちの少なくとも1つのための、かつ配列決定サイクルのうちの少なくとも1つのためのベースコールを生成することとを含む。

40

【0804】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この

50

実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0805】

品質スコアの推論

【0806】

ニューラルネットワークベースのベースコーラーによってコールされた塩基に品質スコアを割り当てるコンピュータ実装方法を開示する。方法は、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの分類スコアを量子化することを含む。方法は、量子化された分類スコアのセットを選択することを含む。方法は、セット内の各量子化された分類スコアについて、その予測されたベースコールを対応するグラウンドトゥルスベースコールと比較することによって、ベースコール誤り率を決定することを含む。方法は、量子化された分類スコアとそれらのベースコール誤り率との間の適合を決定することを含む。すなわち、各量子化された分類スコアについて、その量子化された分類スコアが割り当てられている訓練データ内の訓練実施例のセットが決定される。決定された訓練実施例のセット内の各訓練実施例について、訓練実施例のための予測されたベースコールが、訓練実施例のためのグラウンドトゥルスベースコールと比較され、誤り率が、決定された訓練実施例のセットにわたる比較から決定されて、特定の量子化された分類スコアのための誤り率を提供する。方法は、適合に基づいて、品質スコアを量子化された分類スコアに相関させることを含む。

【0807】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0808】

一実施態様では、量子化された分類スコアのセットは、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの分類スコアのサブセットを含み、分類スコアは実数である。一実施態様では、量子化された分類スコアのセットは、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの全ての分類スコアを含み、分類スコアは実数である。一実施態様では、分類スコアは、合計すると1になり、かつニューラルネットワークベースのベースコーラーのソフトマックス出力層によって生成される、指数関数的に正規化されたソフトマックススコアである。一実施態様では、量子化された分類スコアのセットは、

【数37】

$$0.9 \sum_{i=1}^n 0.1^{(i-1)}$$

10

20

30

40

50

として定義され、ソフトマックススコアに適用される選択式に基づいて選択される。一実施態様では、量子化された分類スコアのセットは、

【数 3 8】

$$\bigvee_{i=1}^{n=10} 0.1i$$

として定義され、ソフトマックススコアに適用される選択式に基づいて選択される。一実施態様では、方法は、相関に基づいて、推論中にニューラルネットワークベースのベースコーラーによってコールされた塩基に品質スコアを割り当てることを含む。一実施態様では、方法は、推論中にニューラルネットワークベースのベースコーラーによってコールされた塩基に品質スコア対応スキームを適用することに基づいて品質スコアを割り当てることを含む。そのような実施態様では、スキームは、推論中に、推論データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成された、分類スコアの範囲を、セット内の対応する量子化された分類スコアにマッピングする。一実施態様では、方法は、推論中に、品質スコアが現在のベースコールサイクルのための設定閾値を下回っている分析物をベースコールすることを中止することを含む。一実施態様では、方法は、推論中に、平均品質スコアが継続的なベースコールサイクルの後に設定閾値を下回っている分析物をベースコールすることを中止することを含む。一実施態様では、予測されたベースコールを対応するグラウンドトゥールスペースコールと比較するために使用されるサンプルサイズは、各量子化された分類スコアに固有である。一実施態様では、予測されたベースコールを対応するグラウンドトゥールスペースコールと比較するために使用されるサンプルサイズは、各量子化された分類スコアに固有である。一実施態様では、適合は、回帰モデルを使用して決定される。一実施態様では、方法は、各量子化された分類スコアについて、その予測されたベースコールを対応するグラウンドトゥールスペースコールと比較することによってベースコール精度率を決定することと、量子化された分類スコアとそれらのベースコール精度率との間の適合を決定することとを含む。一実施態様では、対応するグラウンドトゥールスペースコールは、複数の配列決定器具、配列決定ケミストリー、及び配列決定プロトコル上で配列決定された十分に特徴付けられたヒト及び非ヒトサンプルから導出される。

【0 8 0 9】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0 8 1 0】

品質スコアの予測

【0 8 1 1】

並列に動作し、メモリに連結されている多数のプロセッサ上で実行されるニューラルネットワークベースの品質スコアラーを開示する。システムは、多数のプロセッサ上で実行される畳み込みニューラルネットワークを含む。畳み込みニューラルネットワークは、配列決定画像からのデータを含む訓練実施例上で訓練され、畳み込みニューラルネットワークのベースコール品質予測をベースコール品質グラウンドトゥールスペースと漸進的に一致させる逆伝播ベースの勾配更新技術を使用してベースコール品質グラウンドトゥールスペースでラベル付けされる。システムは、多数のプロセッサのうちの少なくとも1つの上で実行され、1つ又はそれ以上の配列決定サイクルにおいて捕捉された配列決定画像からのデータを、1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を判定するための畳み込みニューラルネットワークに供給する、畳み込みニューラルネットワ

ークの入力モジュールを含む。システムは、多数のプロセッサのうちの少なくとも1つの上で実行され、畳み込みニューラルネットワークによる分析を1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を識別する出力に変換する、畳み込みニューラルネットワークの出力モジュールを含む。

【0812】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0813】

一実施態様では、出力モジュールは、高品質、中品質、及び低品質である品質状態の尤度を生成するソフトマックス分類層を更に含む。そのような実施態様では、これらの尤度に基づいて、品質状態は、高品質、中質、又は低品質として分類される。一実施態様では、ソフトマックス分類層は、複数の品質スコアが割り当てられている品質状態について尤度を生成する。そのような実施態様では、これらの尤度に基づいて、品質状態は、複数の品質スコアのうちの1つから品質スコアが割り当てられる。一実施態様では、品質スコアは、ベースコール誤り確率に対数的に基づいており、複数の品質スコアは、Q6、Q10、Q43、Q20、Q22、Q27、Q30、Q33、Q37、Q40、及びQ50を含む。一実施態様では、出力モジュールは、品質状態を識別する連続的な値を生成する回帰層を更に含む。一実施態様では、システムは、配列決定画像からのデータをコールされた塩基のための品質予測子値で補足し、品質予測子値を配列決定画像からのデータと共に畳み込みニューラルネットワークに供給する、補足的な入力モジュールを含む。一実施態様では、品質予測子値としては、オンライン重複、純度、フェージング、開始5、6量体スコア、モチーフ蓄積、エンジンズ、近似的なホモポリマー、強度減衰、最後から2番目のチャステティ、背景との信号重複(SOWB)、及び／又はシフトされた純度G調整が挙げられる。一実施態様では、品質予測子値としては、ピーク高さ、ピーク幅、ピーク位置、相対ピーク位置、ピーク高さ比、ピーク間隔比、及び／又はピーク対応が挙げられる。

【0814】

このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0815】

品質スコアリングするニューラルネットワーク実装方法も開示する。方法は、1つ又はそれ以上の配列決定サイクルにおいて捕捉された配列決定画像からのデータを、1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を判定するための畳み込みニューラルネットワークに供給することを含む。畳み込みニューラルネットワークは、配列決定画像からのデータを含む訓練実施例上で訓練され、ベースコール品質グラウンドトゥールズでラベル付けされる。訓練することは、畳み込みニューラルネットワークのベースコール品質予測をベースコール品質グラウンドトゥールズと漸進的に一致させる逆伝播ベースの勾配更新技術を使用することを含む。方法は、畳み込みニューラルネットワークによる分析を、1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を識別する出力に変換することを含む。

【0816】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるか

10

20

30

40

50

を理解するであろう。

【0817】

一実施態様では、コンピュータ実装方法は、1つ又はそれ以上の分析物のための入力データをニューラルネットワークによって処理し、入力データの代替表現を生成することと、代替表現を出力層によって処理して出力を生成することであって、出力は、分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別することと、出力に基づいて、分析物のうちの1つ又はそれ以上に対して塩基をコールすることと、出力によって識別された尤度に基づいて、コールされた塩基の品質を判定することと、を含む。

【0818】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。このセクションで説明される方法の他の実施態様は、上述の方法のいずれかを実行するためにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含むことができる。このセクションで説明される方法の更に別の実施態様は、メモリと、メモリ内に記憶された命令を実行するように動作可能な1つ又はそれ以上のプロセッサとを含むシステムを含むことができ、上記の方法のいずれかを実行することができる。

【0819】

ニューラルネットワークベースの品質スコアラーを開示し、このニューラルネットワークベースの品質スコアラーは、並列に動作し、かつメモリに連結されている、多数のプロセッサ上で実行される。システムは、多数のプロセッサ上で実行され、配列決定画像からのデータを含む訓練実施例上で訓練され、ニューラルネットワークのベースコール品質予測をベースコール品質グラウンドトゥールズと漸進的に一致させる逆伝播ベースの勾配更新技術を使用してベースコール品質グラウンドトゥールズでラベル付けされる、ニューラルネットワークを含む。システムは、多数のプロセッサのうちの少なくとも1つの上で実行され、1つ又はそれ以上の配列決定サイクルにおいて捕捉された配列決定画像からのデータを、1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を判定するためのニューラルネットワークに供給する、ニューラルネットワークの入力モジュールを含む。システムは、多数のプロセッサのうちの少なくとも1つの上で実行され、ニューラルネットワークによる分析を1つ又はそれ以上の分析物のためにコールされた1つ又はそれ以上の塩基の品質状態を識別する出力に変換する、ニューラルネットワークの出力モジュールを含む。

【0820】

他の実施態様のための特定の実施態様セクションで説明される特徴のそれぞれは、この実施態様に等しく適用される。上記のように、全ての他の特徴はここでは繰り返されず、参照により繰り返されるべきである。読者は、これらの実施態様において特定された特徴が、他の実施態様で特定されたベース特徴のセットと容易に組み合わせることができるかを理解するであろう。

【0821】

項目

【0822】

本開示はまた、以下の項目も含む。

【0823】

項目セット1

【0824】

1. コンピュータ実装方法であって、

入力データをニューラルネットワークによって処理し、入力データの代替表現を生成す

10

20

30

40

50

ることであって、入力データは、配列決定実行の1つ又はそれ以上の配列決定サイクルのそれぞれのためのサイクルごとのデータを含み、サイクルごとのデータは、それぞれの配列決定サイクルにおける1つ又はそれ以上の分析物を示している、ことと、

代替表現を出力層によって処理し、出力を生成することと、

出力に基づいて、配列決定サイクルのうちの1つ又はそれ以上における分析物のうちの1つ又はそれ以上をベースコールすることと、を含む、コンピュータ実装方法。

【0825】

2. サイクルごとのデータが、それぞれの配列決定サイクルにおける周囲背景を示している、項目1に記載のニューラルネットワーク実装方法。

【0826】

3. 入力データが画像データであり、サイクルごとのデータが、それぞれの配列決定サイクルにおいて捕捉された1つ又はそれ以上の分析物及び周囲背景を示している強度発光を含む、項目1～2のいずれか一項に記載のニューラルネットワーク実装方法。

【0827】

4. サイクルごとのデータのピクセルと、分析物のうちの1つ又はそれ以上を示している強度発光を表現するそれらのピクセルとの間の距離を識別する補足的な距離情報をサイクルごとのデータに添付することを更に含む、項目3に記載のコンピュータ実装方法。

【0828】

5. サイクルごとのデータのピクセルにスケーリング値を割り当てる補足的なスケーリング情報をサイクルごとのデータに添付することを更に含む、項目3に記載のコンピュータ実装方法。

【0829】

6. サイクルごとのデータが、それぞれの配列決定サイクルにおいて検出された電圧変化を示している、項目1に記載のニューラルネットワーク実装方法。

【0830】

7. サイクルごとのデータが、それぞれの配列決定サイクルにおいて測定された電流信号を示している、項目1に記載のニューラルネットワーク実装方法。

【0831】

8. 複数の配列決定サイクルを含む配列決定実行中に合成された分析物をベースコールするニューラルネットワーク実装方法であって、この方法は、

入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することであって、

入力データが、配列決定実行の現在の配列決定サイクルにおいて生成された現在の画像セットの、現在の配列決定サイクルに先行する配列決定実行の1つ又はそれ以上の配列決定サイクルにおいてそれぞれ生成された1つ又はそれ以上の先行する画像セットの、及び現在の配列決定サイクルの後に続く配列決定実行の1つ又はそれ以上の配列決定サイクルにおいてそれぞれ生成された1つ又はそれ以上の後続の画像セットのそれぞれで1つ又はそれ以上の画像から抽出された画像パッチを含み、

画像パッチのそれぞれが、ベースコールされているターゲット分析物の強度発光を表現し、

入力データが、画像パッチの中心ピクセルからの画像パッチのピクセルのそれぞれの距離を示す距離情報を更に含む、ことと、

畳み込み表現を出力層によって処理して、出力を生成することと、

現在の配列決定サイクルにおけるターゲット分析物を出力に基づいてベースコールすることと、を含む、ニューラルネットワーク実装方法。

【0832】

9.

畳み込みニューラルネットワークへの入力としてそれぞれの分析物を表す画像領域の中心の位置座標を提供することを更に含む、

入力が、畳み込みニューラルネットワークの第1の層に提供され、

10

20

30

40

50

入力が、畳み込みニューラルネットワークの1つ又はそれ以上の中間層に提供され、
入力が、畳み込みニューラルネットワークの最終層に提供される、項目8に記載のニューラルネットワーク実装方法。

【0833】

10.

畳み込みニューラルネットワークへの入力として、画像パッチのピクセルに対応するスケーリング値を有する強度スケーリングチャンネルを提供することを更に含み、

スケーリング値が、特定のターゲット分析物をそれぞれ含む画像パッチの中心ピクセルの平均強度に基づいている、項目8～9のいずれか一項に記載のニューラルネットワーク実装方法。

【0834】

11. 強度スケーリングチャンネルが、画像パッチの全てのピクセルについてピクセル単位で同じスケーリング値を含む、項目8～10のいずれか一項に記載のニューラルネットワーク実装方法。

【0835】

12. 各画像パッチが、それぞれのピクセルと複数の分析物のうちの最も近い分析物との間の距離を示すピクセル距離データを更に含み、複数の分析物のうちの最も近い分析物は、ピクセルと分析物のそれぞれとの間の中心間距離に基づいて選択される、項目8に記載のニューラルネットワーク実装方法。

【0836】

13. 各画像パッチが、分析物のうちの1つのみに対して各分析物ピクセルを分類することに基づいて選択された複数の分析物のうちの割り当てられた分析物からの各分析物ピクセルの距離を識別する分析物距離データを更に含み、項目8に記載のニューラルネットワーク実装方法。

【0837】

14. 入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することが、

畳み込みニューラルネットワークの第1の畳み込みサブネットワークによって各サイクルごとの画像パッチセットを別個に処理して、各配列決定サイクルのための中間畳み込み表現を生成することであって、強度及び距離情報を組み合わせ、かつ結果として得られる畳み込み表現を配列決定サイクル内でのみ組み合わせ、かつ配列決定サイクル間では組み合わせない畳み込みを適用することを含む、ことと、

畳み込みニューラルネットワークの第2の畳み込みサブネットワークによって一連内の継続的な配列決定サイクルのための中間畳み込み表現をグループ単位で処理して、一連のための最終畳み込み表現を生成することであって、中間畳み込み表現を組み合わせ、かつ結果として得られる畳み込み表現を配列決定サイクルの間で組み合わせる畳み込みを適用することを含む、ことと、を含み、

出力層によって畳み込み表現を処理して出力を生成することが、出力層によって最終畳み込み表現を処理することを含む、項目8～13のいずれか一項に記載のニューラルネットワーク実装方法。

【0838】

15. ターゲット分析物の中心を中心ピクセルの中心に配置するように各画像パッチのピクセルをリフ্রেームして、リフ্রেームされた画像パッチを生成することを更に含み、

入力データを畳み込みニューラルネットワークによって畳み込んで、入力データの畳み込み表現を生成することが、リフ্রেームされた画像パッチを畳み込みニューラルネットワークによって畳み込んで、畳み込み表現を生成することを含む、項目8～14のいずれか一項に記載のニューラルネットワーク実装方法。

【0839】

16. リフ্রেームすることが、リフ্রেームすることを補正するための各画像パッチのピクセルの強度補間を更に含み、項目15に記載のニューラルネットワーク実装方法。

10

20

30

40

50

【0840】

17. ベースコールするニューラルネットワーク実装方法であって、この方法は、
サイクルごとの入力データのシーケンス内の各サイクルごとの入力データを畳み込みニューラルネットワークの畳み込み層のカスケードによって別個に処理することであって、
サイクルごとの入力データのシーケンスが、配列決定実行の一連の配列決定サイクルのために生成され、

各サイクルごとの入力データが、それぞれの配列決定サイクルにおいて捕捉された1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する画像チャンネルを含む、ことと、

各配列決定サイクルに対して、

別個に処理することに基づいて、畳み込み層のそれぞれにおいて畳み込み表現を生成し、それによって畳み込み表現のシーケンスを生成することと、

そのサイクルごとの入力データをその対応する畳み込み表現のシーケンスと混合し、混合表現を生成することと、

その混合表現を平坦化し、平坦化された混合表現を生成することと、を行うことと、

継続的な配列決定サイクルの平坦化された混合表現をスタックとして配列することと、
スタックを反復ニューラルネットワークによって順方向及び逆方向に処理することであって、この反復ニューラルネットワークが、

スライディングウィンドウに基づいたスタック内の平坦化された混合表現のサブセットにわたって畳み込みを行うことであって、各スライディングウィンドウがそれぞれの配列決定サイクルに対応する、ことと、

(i) スタック上の現在のスライディングウィンドウ内の平坦化された混合表現のサブセット及び(ii) 前の隠れ状態表現に基づいて、各配列決定サイクルのための各時間ステップにおける現在の隠れ状態表現を継続的に生成することと、を行う、ことと、

スタックを順方向及び逆方向に処理することの結果に基づいて、配列決定サイクルのそれぞれにおける分析物のそれぞれをベースコールすることと、を含む、ニューラルネットワーク実装方法。

【0841】

18.

所与の配列決定サイクルにおける分析物のそれぞれを、

所与の配列決定サイクルの順方向及び逆方向の現在の隠れ状態表現を時間ステップに基づいて組み合わせ、組み合わせられた隠れ状態表現を生成することであって、組み合わせることが、連結又は総和又は平均化を含む、ことと、

組み合わせられた隠れ状態表現を1つ又はそれ以上の全結合ネットワークによって処理し、高密度表現を生成することと、

高密度表現をソフトマックス層によって処理して、所与の配列決定サイクルにおける分析物のそれぞれに組み込まれた塩基がA、C、T、及びGであることの尤度を生成することと、

尤度に基づいて塩基をA、C、T、又はGとして分類することと、によってベースコールすることを更に含む、項目17に記載のニューラルネットワーク実装方法。

【0842】

19. ベースコールするためのニューラルネットワークベースのシステムであって、システムは、

反復モジュール及び畳み込みモジュールを有するハイブリッドニューラルネットワークであって、反復モジュールが、畳み込みモジュールからの入力を使用する、ハイブリッドニューラルネットワークと、

配列決定実行の一連の配列決定サイクルのための画像データを1つ又はそれ以上の畳み込み層によって処理し、画像データの1つ又はそれ以上の畳み込み表現を生成する畳み込みモジュールであって、画像データが、1つ又はそれ以上の分析物及びそれらの周囲背景の強度発光を表現する、畳み込みモジュールと、

10

20

30

40

50

畳み込み表現及び前の隠れ状態表現を畳み込むことに基づいて現在の隠れ状態表現を生成する反復モジュールと、

現在の隠れ状態表現に基づいて、分析物のうちの少なくとも1つのための、かつ配列決定サイクルのうちの少なくとも1つのためのベースコールを生成する出力モジュールと、を備える、ニューラルネットワークベースのシステム。

【0843】

20. クラスターをベースコールするコンピュータ実装方法であって、

入力データをニューラルネットワークによって処理し、入力データの代替表現を生成することであって、

入力データが、(i) 配列決定実行の1つ又はそれ以上の配列決定サイクルのそれぞれのためのサイクルごとのデータと、(ii) 補足的な距離情報と、を含み、

サイクルごとのデータが、配列決定サイクルのうちのそれぞれ1つにおいて捕捉された1つ又はそれ以上のクラスター及びその周囲背景を示している強度発光を表現するピクセルを含み、

サイクルごとのデータが、サイクルごとのデータのピクセル間の距離を識別する補足的な距離情報を伴い、

ニューラルネットワークによるサイクルごとのデータのピクセルの処理中に、補足的な距離情報が、サイクルごとのデータのピクセルのうちのどれがクラスターの中心を含んでいるか、及びサイクルごとのデータのピクセルのうちのどれがクラスターの中心からより遠く離れているかをニューラルネットワークに伝える加法的バイアスを供給する、ことと、

代替表現を出力層によって処理し、出力を生成することと、

出力に基づいて、配列決定サイクルのうちの1つ又はそれ以上におけるクラスターのうちの1つ又はそれ以上をベースコールすることと、を含む、コンピュータ実装方法。

【0844】

21. 加法的バイアスが、ベースコールすることの精度を改善する、項目20に記載のコンピュータ実装方法。

【0845】

22. ニューラルネットワークが、周辺クラスターピクセル、背景ピクセル、及びそれらから導出される代替表現よりも、中央クラスターピクセル、それらの隣接ピクセル、及びそれらから導出される代替表現に着目することによって配列決定信号をその適切なソースクラスターに割り当てるために、補足的な距離情報を使用する、項目21に記載のコンピュータ実装方法。

【0846】

項目セット2

【0847】

1. コンピュータ実装方法であって、

1つ又はそれ以上の分析物のための入力データをニューラルネットワークベースのベースコーラーによって処理し、入力データの代替表現を生成することと、

代替表現を出力層によって処理して出力を生成することであって、出力は、分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別する、ことと、

出力に基づいて、分析物のうちの1つ又はそれ以上に対して塩基をコールすることと、出力によって識別された尤度に基づいて、コールされた塩基の品質スコアを判定することと、を含む、コンピュータ実装方法。

【0848】

2. 尤度に基づいて、コールされた塩基の品質スコアを判定することが、

訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成されるベースコールの分類スコアを量子化することと、

量子化された分類スコアのセットを選択することと、

セット内の各量子化された分類スコアについて、その予測されたベースコールを対応す

10

20

30

40

50

るグラウンドトゥールスペースコールと比較することによって、ベースコール誤り率を決定することと、

量子化された分類スコアとそれらのベースコール誤り率との間の適合を決定することと、適合に基づいて、品質スコアを量子化された分類スコアに相関させることと、を含む、項目 1 に記載のコンピュータ実装方法。

【0849】

3. 量子化された分類スコアのセットが、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの分類スコアのサブセットを含み、

分類スコアが実数である、項目 1 ～ 2 のいずれか一項に記載のコンピュータ実装方法。

10

【0850】

4. 量子化された分類スコアのセットが、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成される予測されたベースコールの全ての分類スコアを含み、

分類スコアが実数である、項目 1 ～ 3 のいずれか一項に記載のコンピュータ実装方法。

【0851】

5. 分類スコアが、合計すると 1 になり、かつニューラルネットワークベースのベースコーラーのソフトマックス出力層によって生成される、指数関数的に正規化されたソフトマックススコアである、項目 1 ～ 4 のいずれか一項に記載のコンピュータ実装方法。

【0852】

20

6. 量子化された分類スコアのセットが、

【数 39】

$$0.9 \sum_{i=1}^n 0.1^{(i-1)}$$

として定義され、ソフトマックススコアに適用される選択式に基づいて選択される、項目 1 ～ 5 のいずれか一項に記載のコンピュータ実装方法。

【0853】

30

7. 量子化された分類スコアのセットが、

【数 40】

$$\bigvee_{i=1}^{n=10} 0.1^i$$

として定義され、ソフトマックススコアに適用される選択式に基づいて選択される、項目 1 ～ 6 のいずれか一項に記載のコンピュータ実装方法。

【0854】

40

8.

相関に基づいて、推論中にニューラルネットワークベースのベースコーラーによってコールされた塩基に品質スコアを割り当てることを更に含む、項目 1 ～ 7 のいずれか一項に記載のコンピュータ実装方法。

【0855】

9.

推論中にニューラルネットワークベースのベースコーラーによってコールされた塩基に品質スコア対応スキームを適用することに基づいて品質スコアを割り当てることを更に含む、

スキームが、推論中に推論データを処理することに応じてニューラルネットワークベー

50

スのベースコーラーによって生成された、分類スコアの範囲を、セット内の対応する量子化された分類スコアにマッピングする、項目 8 に記載のコンピュータ実装方法。

【0856】

10.

推論中に、品質スコアが現在のベースコールサイクルのための設定閾値を下回っている分析物をベースコールすることを中止することを更に含む、項目 8 ～ 9 のいずれか一項に記載のコンピュータ実装方法。

【0857】

11.

推論中に、平均品質スコアが継続的なベースコールサイクルの後に設定閾値を下回っている分析物をベースコールすることを中止することを更に含む、項目 8 ～ 10 のいずれか一項に記載のコンピュータ実装方法。

10

【0858】

12. 予測されたベースコールを対応するグラウンドトゥルスベースコールと比較するために使用されるサンプルサイズが、各量子化された分類スコアに固有である、項目 8 ～ 11 のいずれか一項に記載のコンピュータ実装方法。

【0859】

13. 適合が、回帰モデルを使用して決定される、項目 8 ～ 12 のいずれか一項に記載のコンピュータ実装方法。

【0860】

14.

各量子化された分類スコアについて、その予測されたベースコールを対応するグラウンドトゥルスベースコールと比較することによって、ベースコール精度率を決定することと、

20

量子化された分類スコアとそれらのベースコール精度率との間の適合を決定することと、を更に含む、項目 8 ～ 13 のいずれか一項に記載のコンピュータ実装方法。

【0861】

15. 対応するグラウンドトゥルスベースコールが、複数の配列決定器具、配列決定ケミストリー、及び配列決定プロトコル上で配列決定された十分に特徴付けられたヒト及び非ヒトサンプルから導出される、項目 8 ～ 14 のいずれか一項に記載のコンピュータ実装方法。

30

【0862】

16. ニューラルネットワークベースの品質スコアラーであって、

並列に動作し、かつメモリに連結されている多数のプロセッサと、

多数のプロセッサ上で実行され、配列決定画像からのデータを含む訓練実施例上で訓練され、ニューラルネットワークのベースコール品質予測を、既知の正しいベースコールを識別するベースコール品質グラウンドトゥルスと漸進的に一致させる逆伝播ベースの勾配更新技術を使用してベースコール品質グラウンドトゥルスでラベル付けされる、ニューラルネットワークと、

多数のプロセッサのうちの少なくとも 1 つの上で実行され、1 つ又はそれ以上の配列決定サイクルにおいて捕捉された配列決定画像からのデータを、1 つ又はそれ以上の分析物のためにコールされた 1 つ又はそれ以上の塩基の品質を判定するためのニューラルネットワークに供給する、ニューラルネットワークの入力モジュールと、

40

多数のプロセッサのうちの少なくとも 1 つの上で実行され、ニューラルネットワークによる分析を 1 つ又はそれ以上の分析物のためにコールされた 1 つ又はそれ以上の塩基の品質を識別する出力に変換する、ニューラルネットワークの出力モジュールと、を含む、ニューラルネットワークベースの品質スコアラー。

【0863】

17. ニューラルネットワークが畳み込みニューラルネットワークである、項目 16 に記載のニューラルネットワークベースの品質スコアラー。

50

【0864】

18. 出力モジュールが、高品質、中品質、及び低品質である品質の尤度を生成するソフトマックス分類層を更に含み、

尤度に基づいて、品質を高品質、中品質、又は低品質として分類することを更に含む、項目16に記載のニューラルネットワークベースの品質スコアラー。

【0865】

19. ソフトマックス分類層が、複数の品質スコアを割り当てられている品質について尤度を生成し、

尤度に基づいて、品質に、複数の品質スコアのうちの1つから品質スコアを割り当てることを更に含む、項目16に記載のニューラルネットワークベースの品質スコアラー。

【0866】

20. 品質スコアが、ベースコール誤り確率に対数的に基づいており、
複数の品質スコアが、Q6、Q10、Q15、Q20、Q22、Q27、Q30、Q33、Q37、Q40、及びQ50を含む、項目16～19のいずれか一項に記載のニューラルネットワークベースの品質スコアラー。

【0867】

21. 出力モジュールが、品質を識別する連続的な値を生成する回帰層を更に含む、項目16～20のいずれか一項に記載のニューラルネットワークベースの品質スコアラー。

【0868】

22.
補足的な入力モジュールであって、
配列決定画像からのデータを、コールされた塩基のための品質予測子値で補足し、
品質予測子値を配列決定画像からのデータと共に畳み込みニューラルネットワークに供給する、補足的な入力モジュールを更に含む、項目16～21のいずれか一項に記載のニューラルネットワークベースの品質スコアラー。

【0869】

23. 品質予測子値が、オンライン重複、純度、フェージング、開始5、6量体スコア、モチーフ蓄積、エンジネス、近似的なホモポリマー、強度減衰、最後から2番目のチャステティ、背景との信号重複(SOWB)、及び／又はシフトされた純度G調整を含む、項目22に記載のニューラルネットワークベースの品質スコアラー。

【0870】

24. 品質予測子値が、ピーク高さ、ピーク幅、ピーク位置、相対ピーク位置、ピーク高さ比、ピーク間隔比、及び／又はピーク対応を含む、項目22に記載のニューラルネットワークベースの品質スコアラー。

【0871】

25. ベースコールの品質スコアを判定するコンピュータ実装方法であって、
1つ以上の分析物のための入力データをニューラルネットワークベースのベースコーラーによって処理し、入力データの代替表現を生成することと、

代替表現を出力層によって処理して出力を生成することであって、出力が、分析物のうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別することと、

出力に基づいて、分析物のうちの1つ以上に対して塩基をコールすることと、
出力によって識別された尤度に基づいて、コールされた塩基の品質スコアを判定することであって、

訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成されるベースコールの分類スコアを量子化すること、

量子化された分類スコアのセットを選択すること、
セット内の各量子化された分類スコアについて、その予測されたベースコールを対応するグラウンドトゥルースベースコールと比較することによって、ベースコール誤り率を決定すること、

10

20

30

40

50

量子化された分類スコアとそれらのベースコール誤り率との間の適合を決定すること、及び

適合に基づいて、品質スコアを量子化された分類スコアに相関させること、によって行われる、判定することと、を含む、コンピュータ実装方法。

【0872】

26. 適合は量子化された分類スコアと品質スコアとの間の対応を明示している、項目1に記載のコンピュータ実装方法。

【0873】

27. ベースコールの品質スコアを判定するコンピュータ実装方法であって、

1つ以上のクラスターのための入力データをニューラルネットワークベースのベースコーラーによって処理し、入力データの代替表現を生成することと、

代替表現を出力層によって処理して出力を生成することであって、出力が、クラスターのうちの特定の1つに組み込まれた塩基がA、C、T、及びGであることの尤度を識別する、ことと、

出力に基づいて、クラスターのうちの1つ以上に対して塩基をコールすることと、

ニューラルネットワークベースのベースコーラーの訓練に対して較正された量子化スキームに基づいて、出力によって識別された尤度に基づいてコールされた塩基の品質スコアを判定することであって、量子化スキームが、訓練中に訓練データを処理することに応じてニューラルネットワークベースのベースコーラーによって生成されるコールされた塩基の分類スコアを量子化することを含む、ことと、

量子化された分類スコアのセットを選択することと、

セット内の各量子化された分類スコアについて、その予測されたベースコールを対応するグラウンドトゥールスペースコールと比較することによって、ベースコール誤り率を決定することと、

量子化された分類スコアとそれらのベースコール誤り率との間の適合を決定することと、

適合に基づいて、品質スコアを量子化された分類スコアに相関させることと、を含む、コンピュータ実装方法。

【符号の説明】

【0874】

218 ニューラルネットワークベースのベースコーラー

6102 ニューラルネットワークベースの品質スコアラー

6304 温度制御システム

6306 システムコントローラ

6308 流体制御システム

6312 バイオセンサー

6314 流体貯蔵システム

6316 照明システム

6318 ユーザーインターフェース

6324 主制御モジュール

6326 照明モジュール

6328 流体制御モジュール

6330 流体貯蔵モジュール

6332 温度制御モジュール

6336 装置モジュール

6338 識別モジュール

6340 SBSモジュール

6342 増幅モジュール

6344 分析モジュール

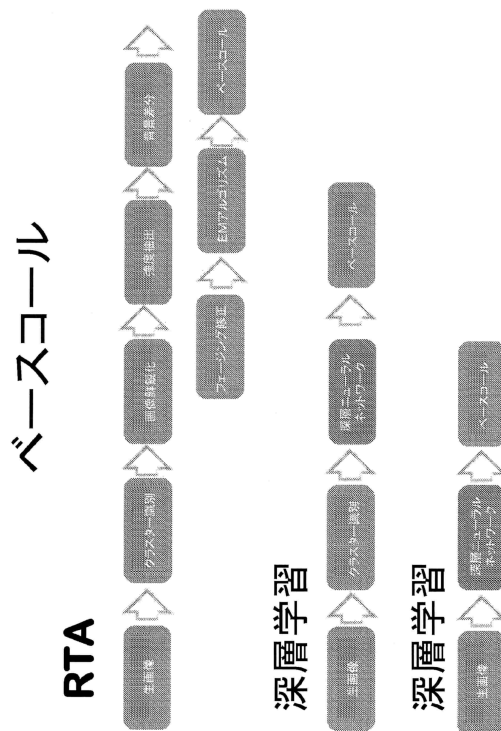
6346 構成可能プロセッサ

6348 メモリ

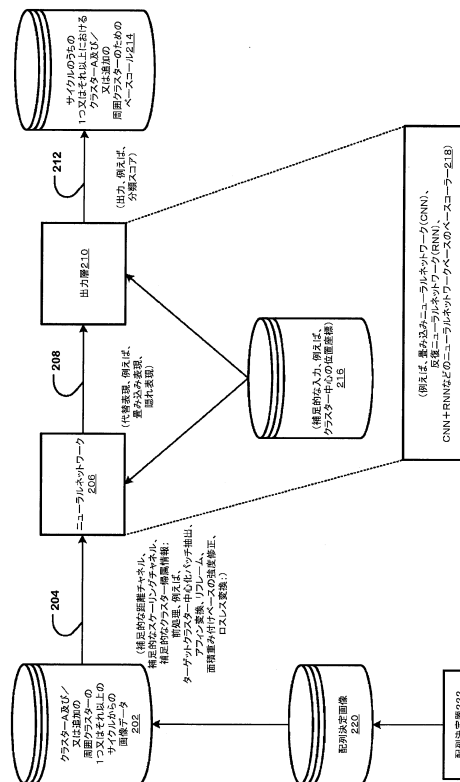
- 6348A メモリ（ビットファイル、ベースコールパラメータ）
 6348B メモリ（画像データ、ベースコールリード）
 6348C メモリ（タイルデータ、モデルパラメータ）
 6352 CPU（実行時）
 6379 多重サイクル実行クラスター1
 6397 データフローロジック
 6500 コンピュータシステム
 6510 ストレージサブシステム
 6522 メモリサブシステム
 6536 ファイル記憶サブシステム
 6538 ユーザーインターフェース入力装置
 6555 バスサブシステム
 6574 ネットワークインターフェースサブシステム
 6576 ユーザーインターフェース出力装置
 6578 深層学習プロセッサ（GPU、FPGA、CGRA）

【図面】

【図1】



【図2】



10

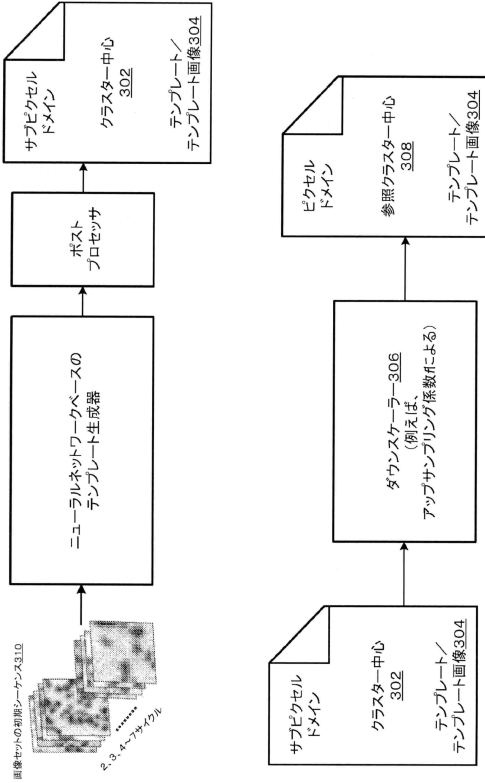
20

30

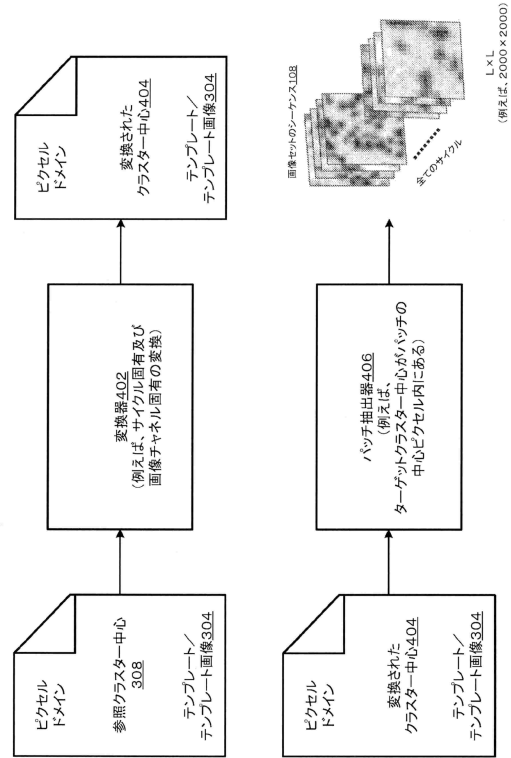
40

50

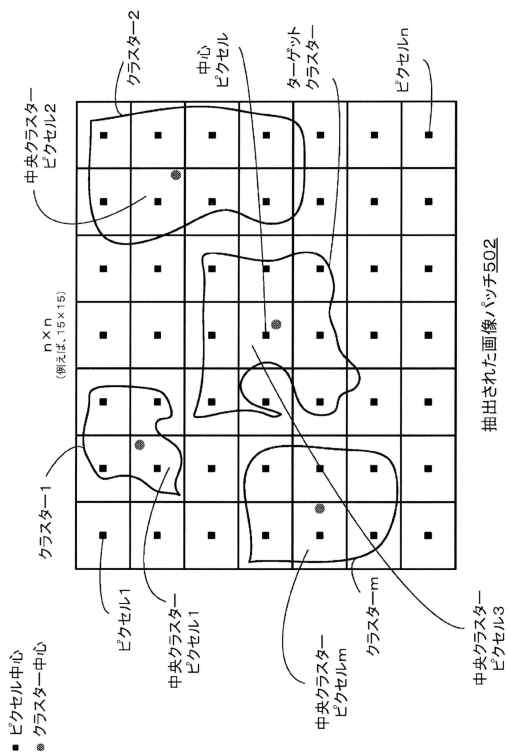
【図 3】



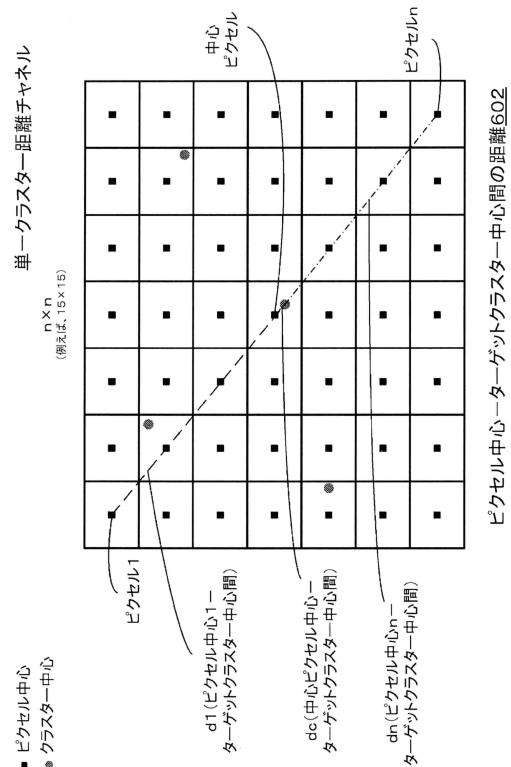
【図 4】



【図 5】



【図 6】



10

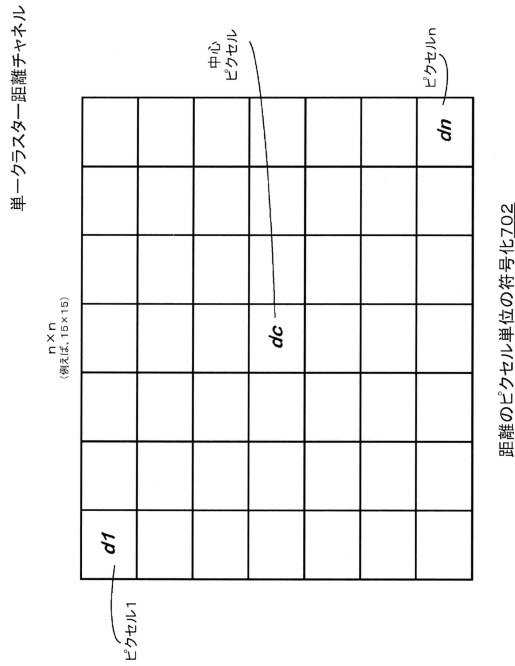
20

30

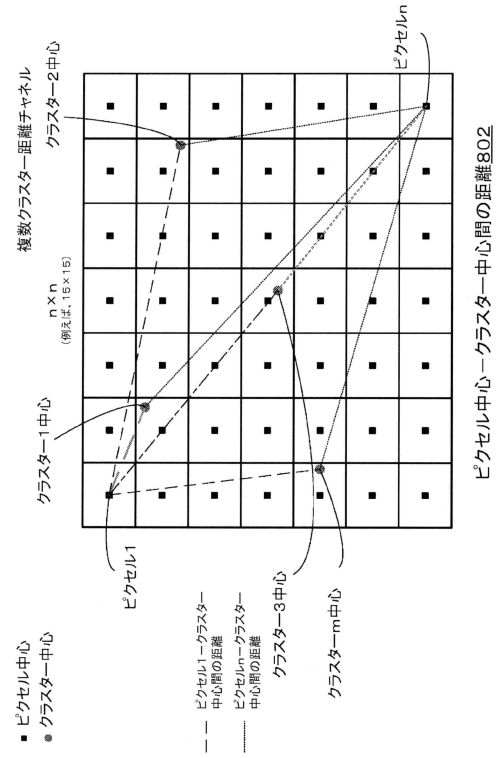
40

50

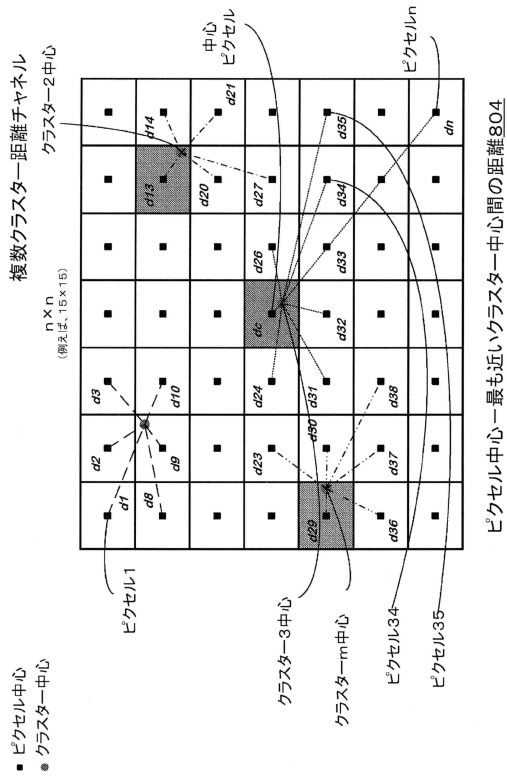
【図 7】



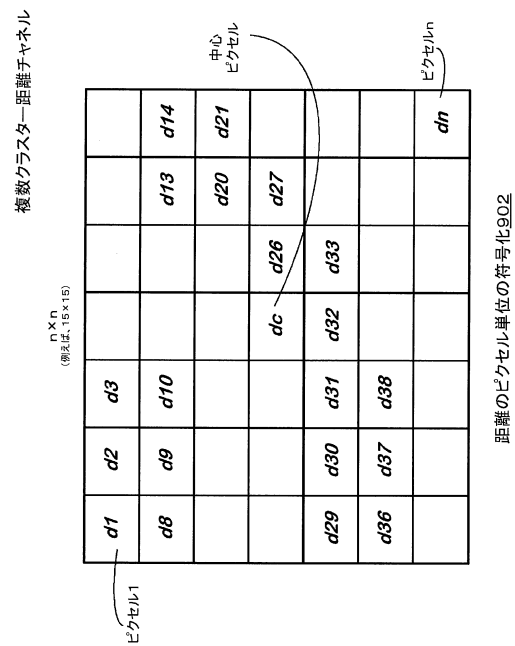
【図 8 a】



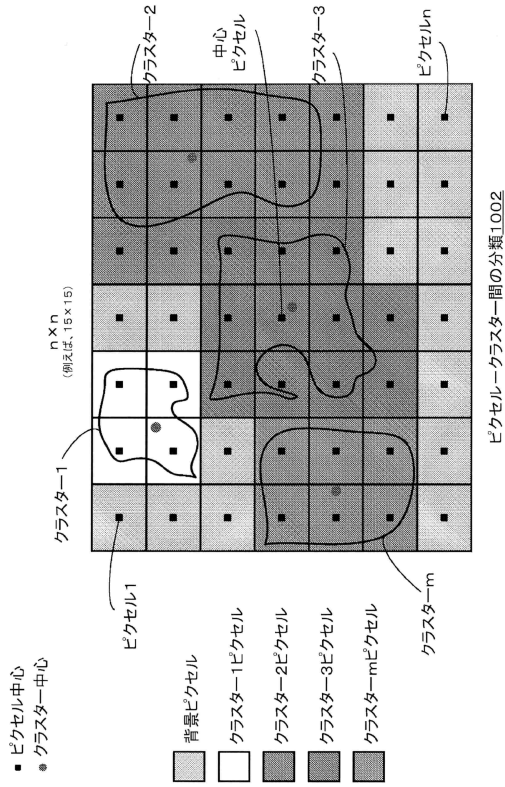
【図 8 b】



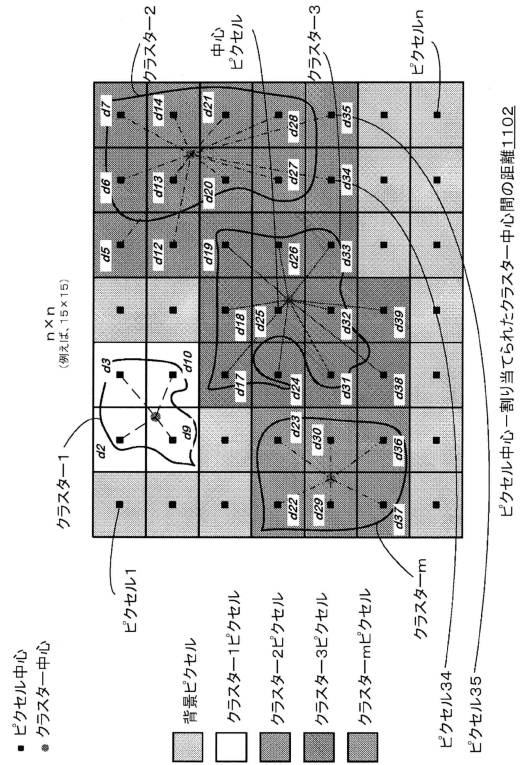
【図 9】



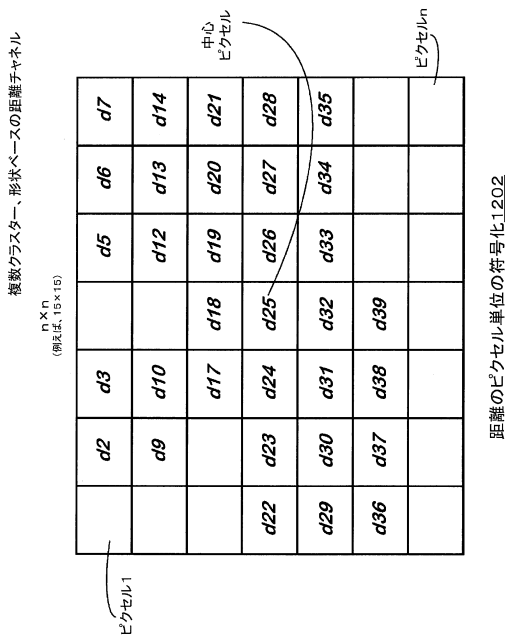
【図 10】



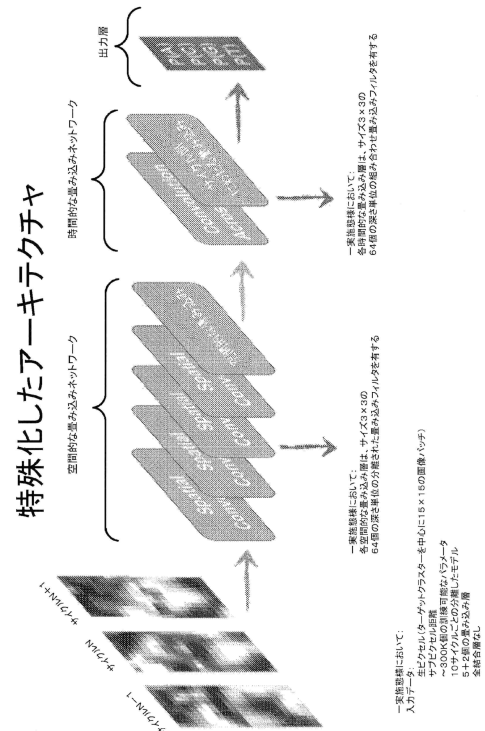
【図 1 1】



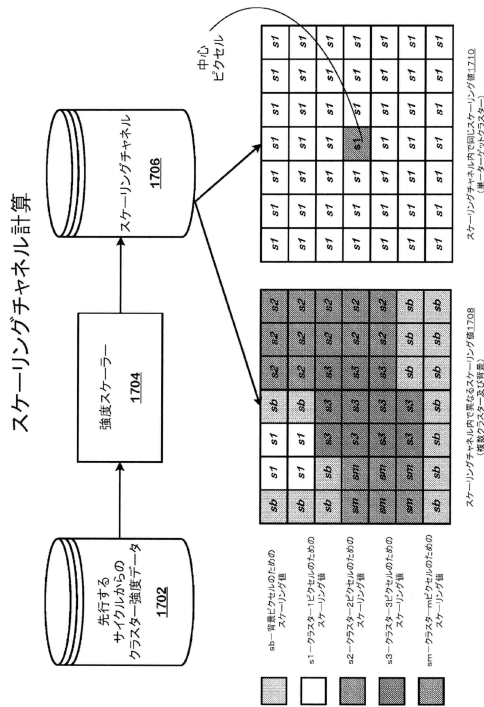
【図 12】



【図 13】

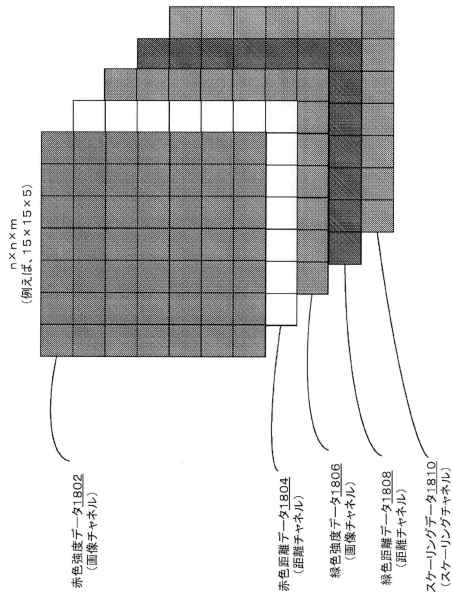


【図 1 7】

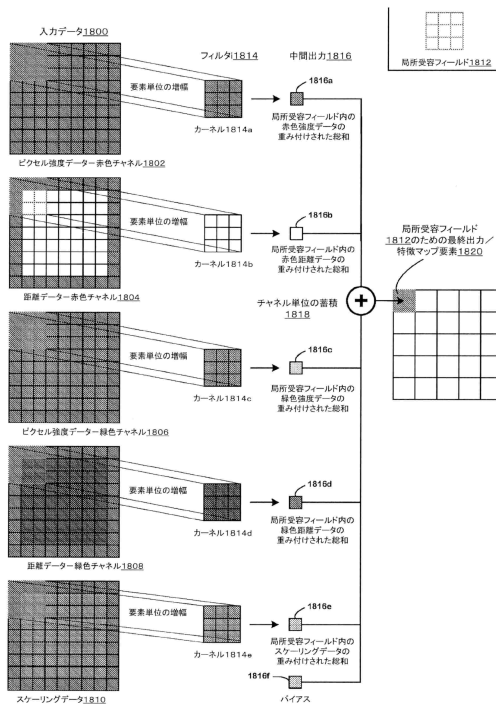


【図 1 8 a】

赤色画像及び緑色画像を生成する単一の配列決定サイクルのための入力データ1800

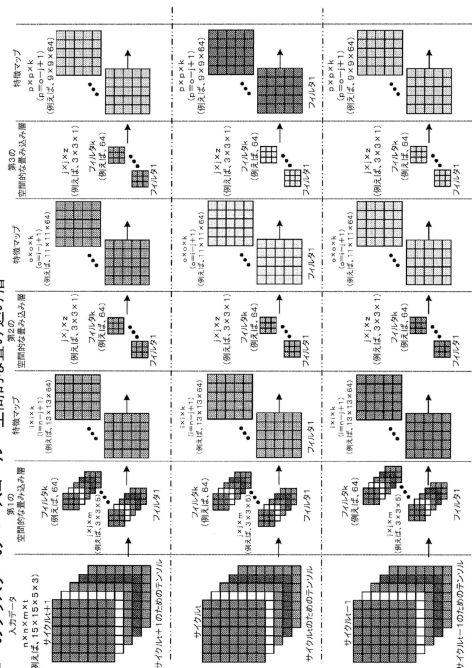


【図 1 8 b】



【図 1 9 a】

単一のクラスターのベースコーネル空間的な量込み層



10

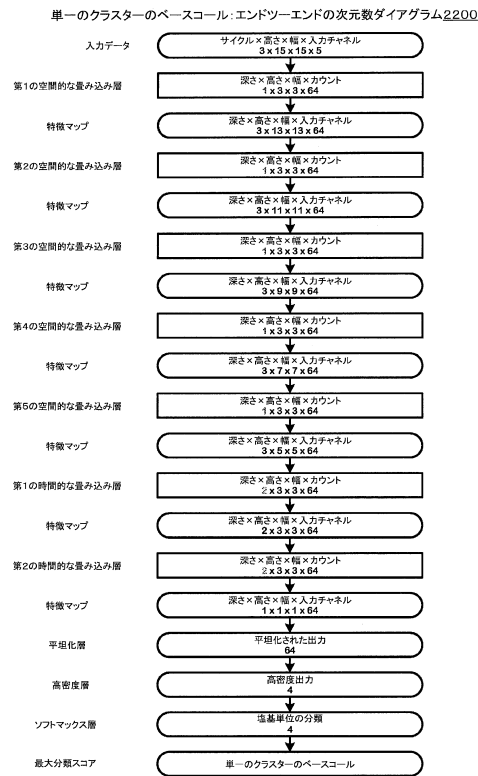
20

30

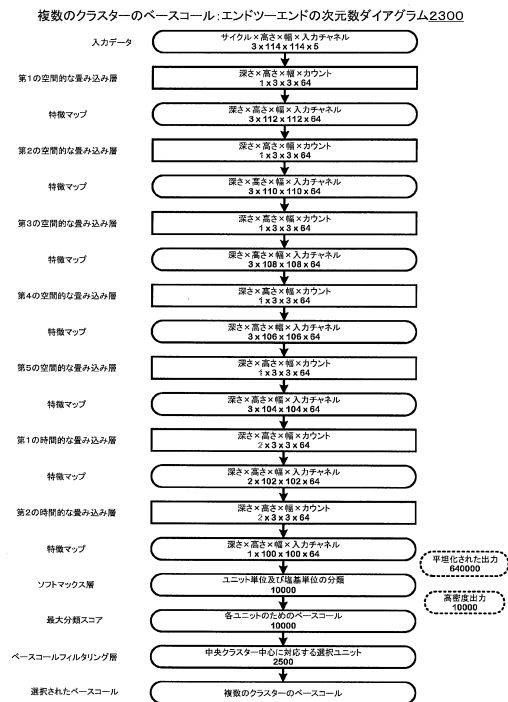
40

50

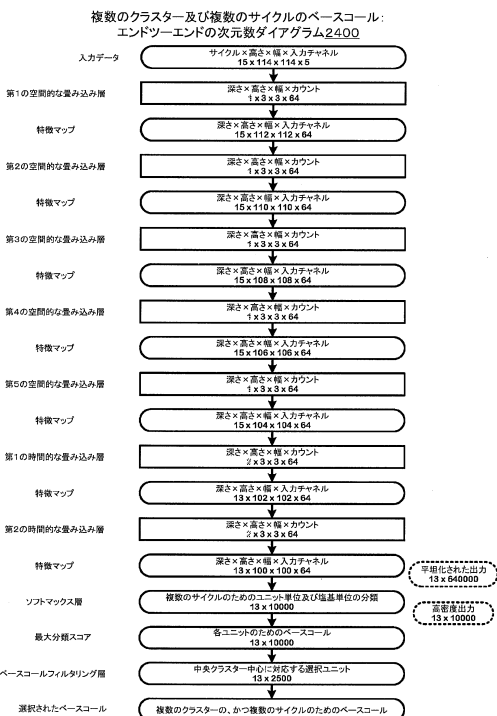
【図 2 2】



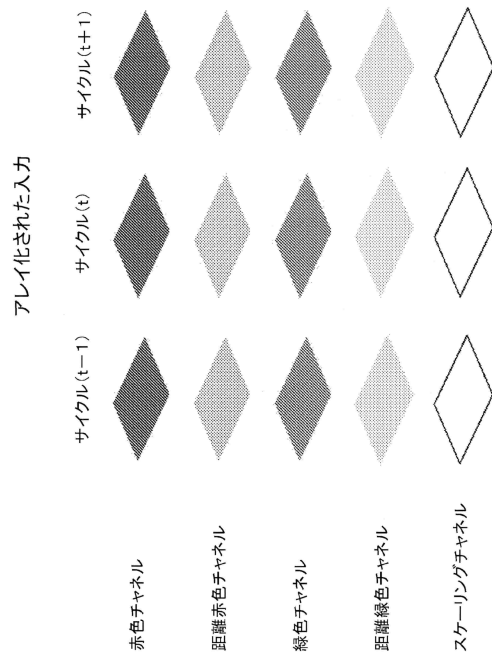
【図 2 3】



【図 2 4】



【図 2 5 a】



10

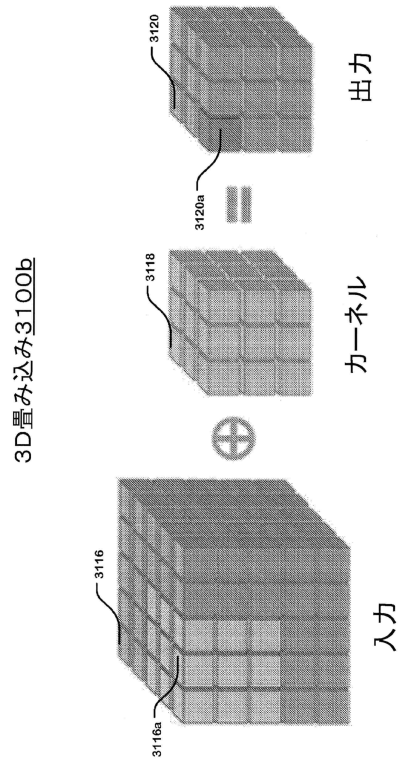
20

30

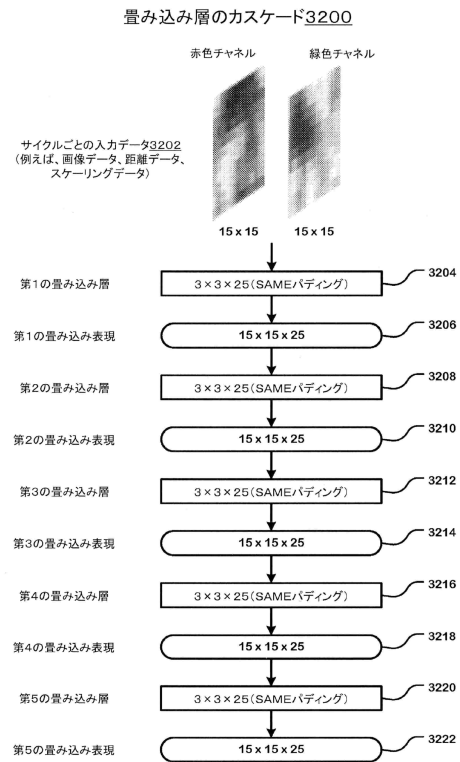
40

50

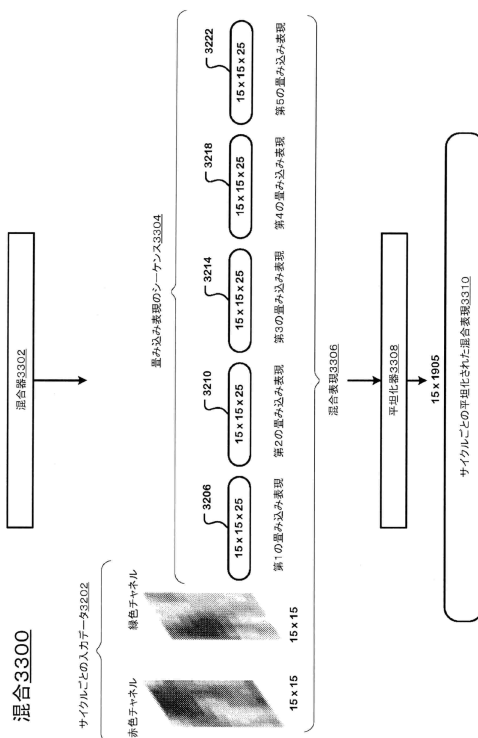
【図 3 1 b】



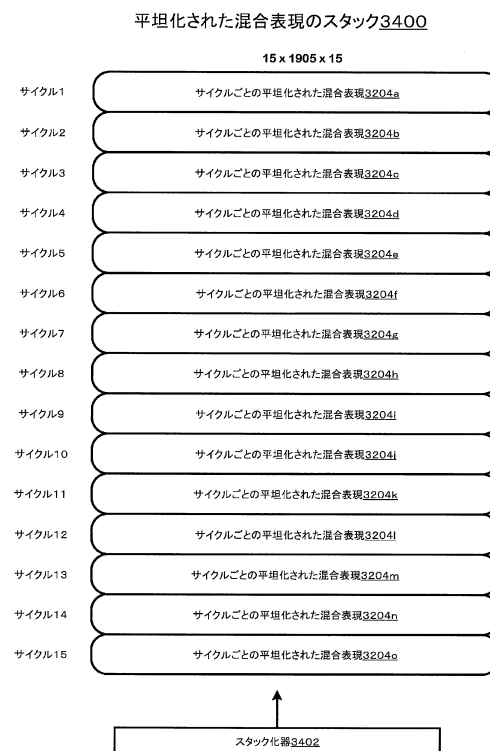
【図 3 2】



【図 3 3】



【図 3 4】



10

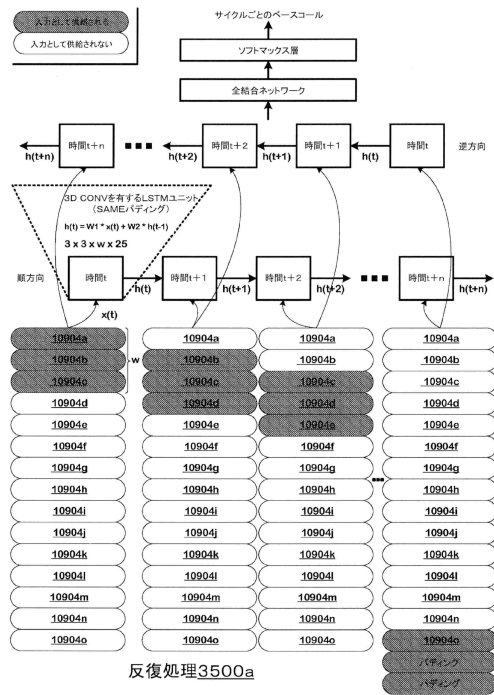
20

30

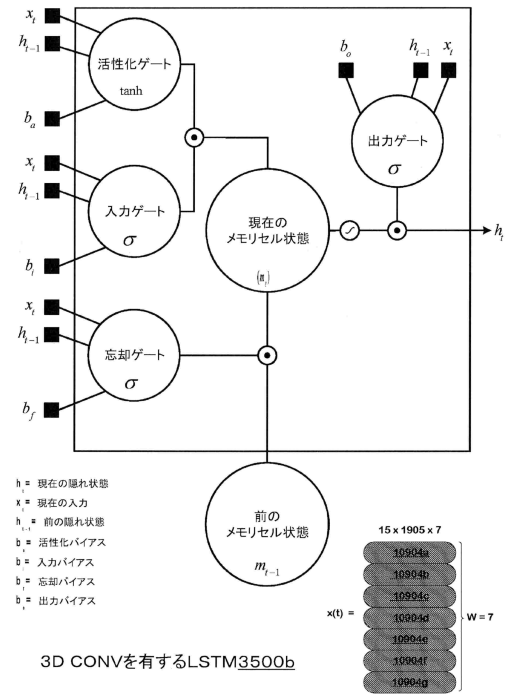
40

50

【図 3 5 a】



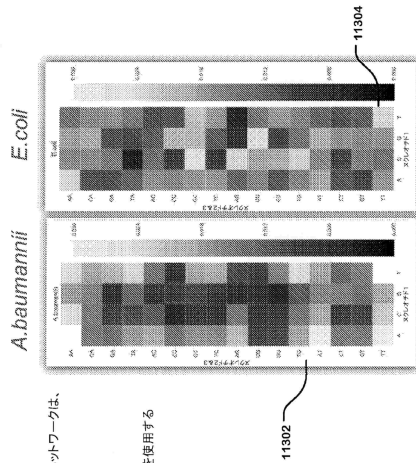
【図 3 5 b】



3D CONVを有するLSTM3500b

【図 3 6】

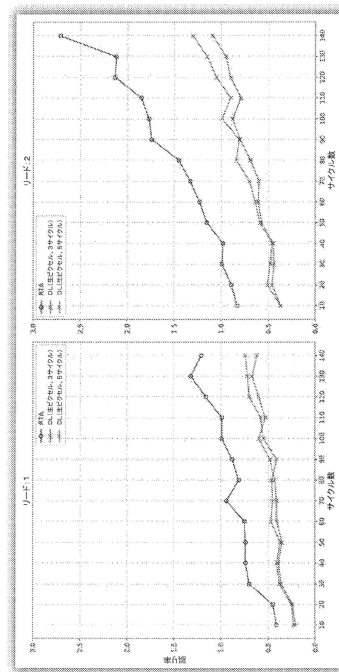
訓練データ内の3量体平衡化



- 300K語を超える訓練可能なパラメータを用いて、ネットワークは、ケラムに関する統計を学習することができる
- ネットワークに供給された3量体の数を均等化させて、ケラム統計をマスキングする
- 高マッピングスコアを有するリードをコールしたRTAを使用する

【図 3 7】

深層学習を使用したベースコール

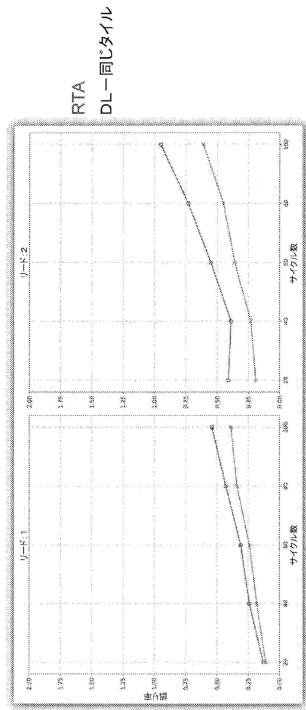


- NextSeq, 2チャネル, ランダムなフローセル
- 10モデルのアンサンブル
- 同じタイル上での訓練及び試験
- ~50%の誤り低減

【図 38】

深層学習を使用したベースコール

タイル間一般化

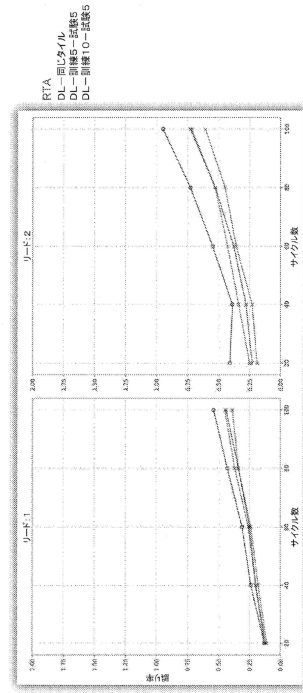


- NextSeq、2チャネル、ランダムなフローセル
- 単一モデル

【図 39】

深層学習を使用したベースコール

タイル間一般化

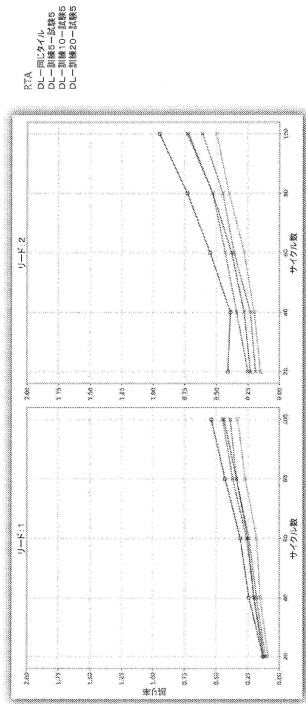


- NextSeq、2チャネル、ランダムなフローセル
- 単一モデル

【図 40】

深層学習を使用したベースコール

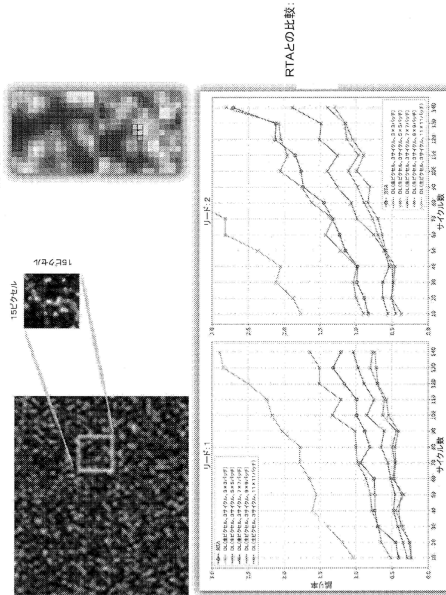
タイル間一般化



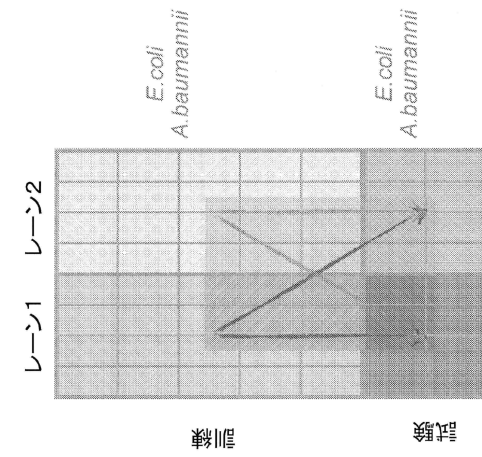
- NextSeq、2チャネル、ランダムなフローセル
- アンサンブルなし単一モデル

【図 41】

パッチサイズの影響

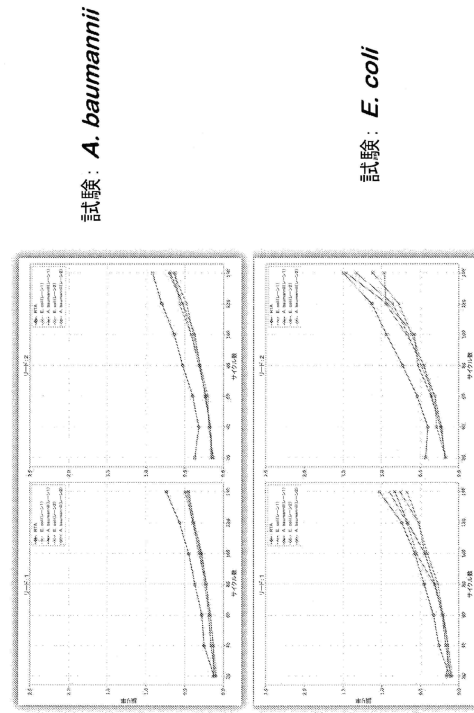


【図 4 2】

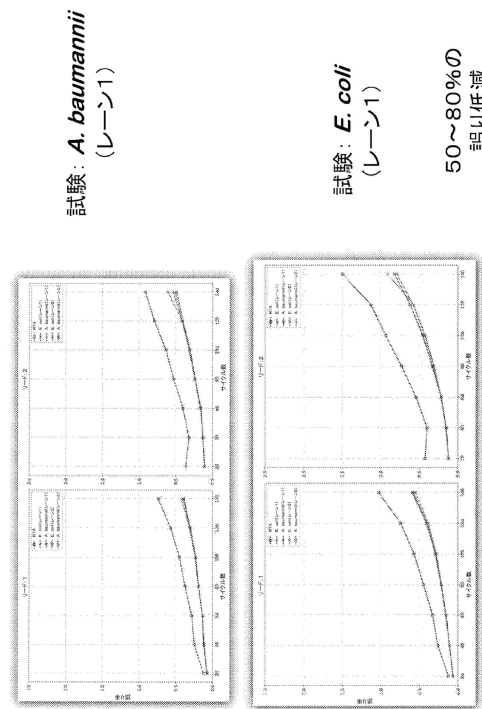


- ▶ レーン間一般化
- ▶ 有機体横断的チェック
 - ・ 複数の有機体一重化

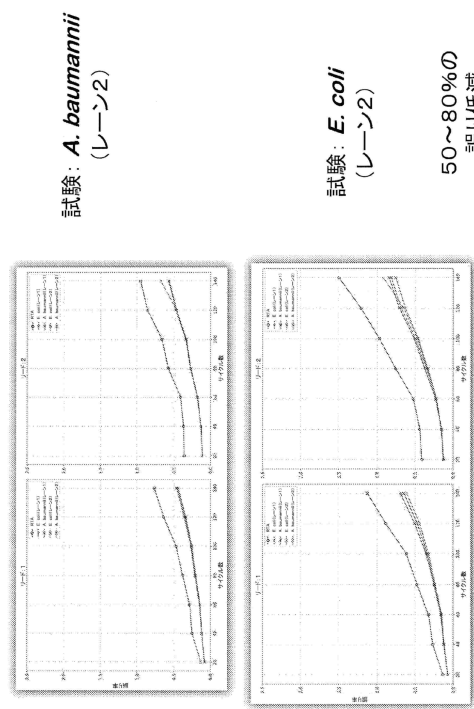
【図 4 3】



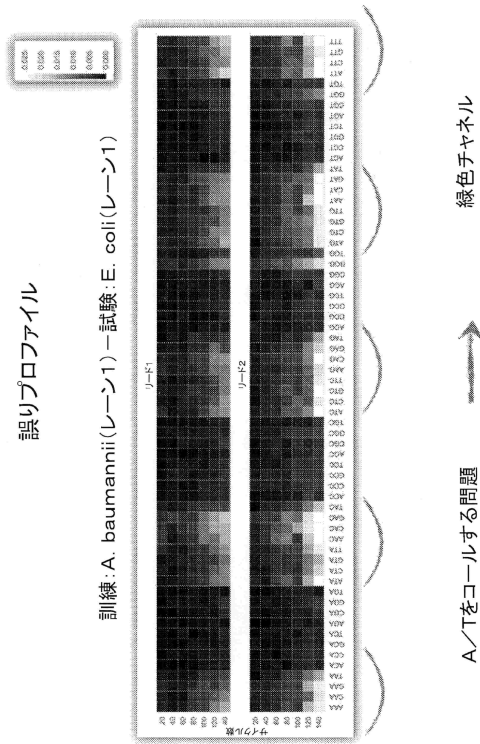
【図 4 4】



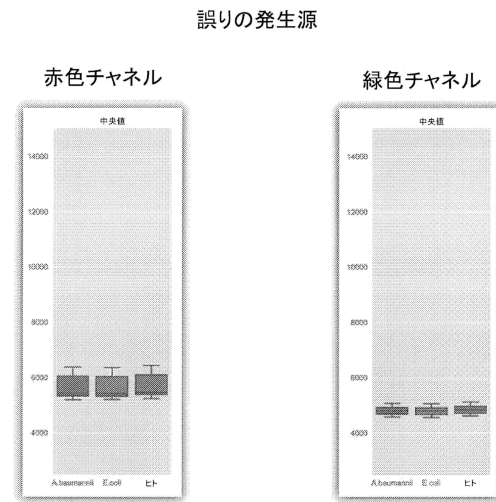
【図 4 5】



【図 4 6】



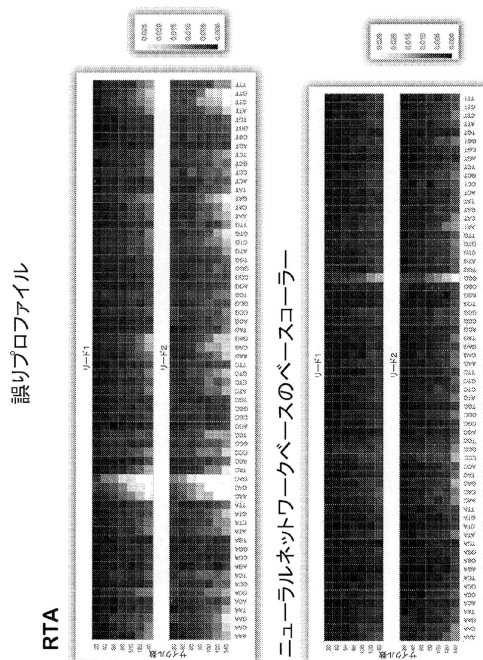
【図 4 7】



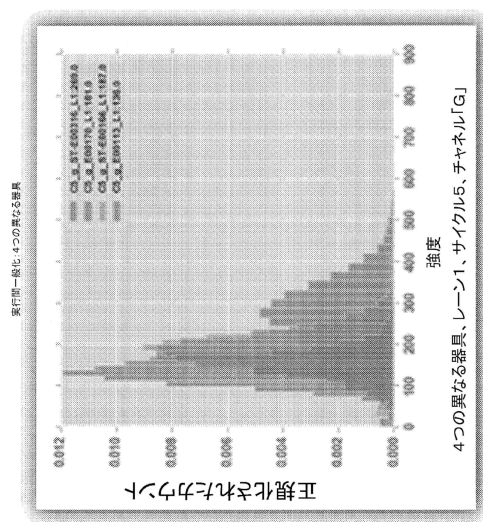
10

20

【図 4 8】



【図 4 9 a】



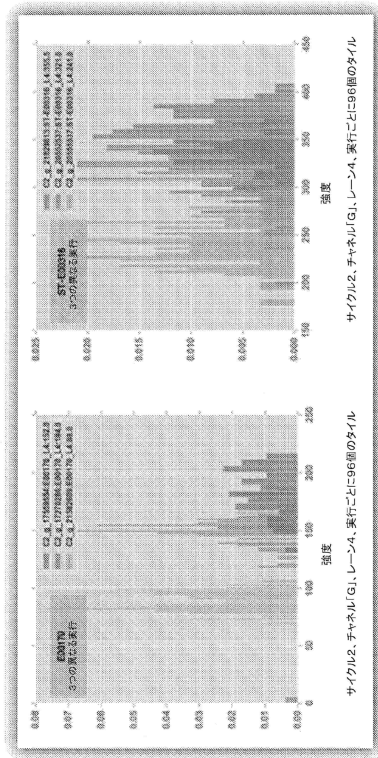
30

40

50

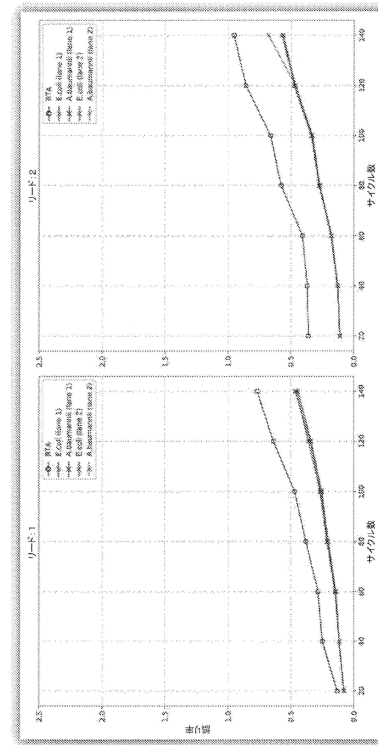
【図 49 b】

実行間一般化: 4つの異なる実行に対して同じ器具



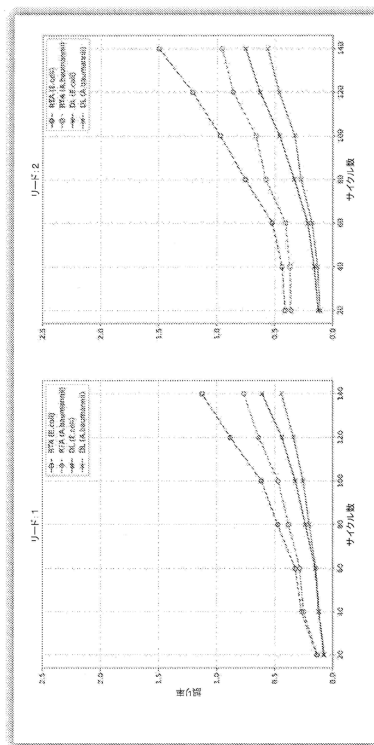
【図 50】

ゲノム統計



【図 51】

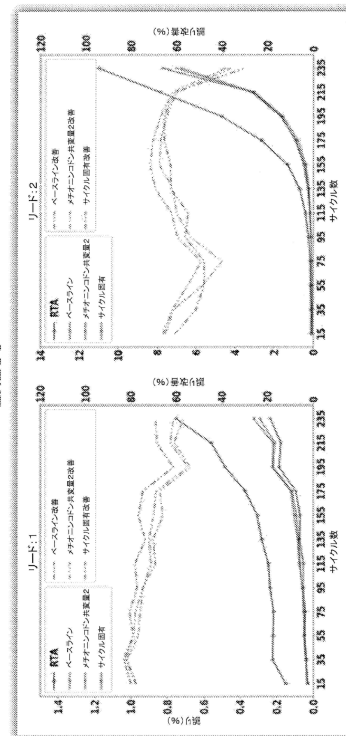
ゲノムコンテキスト



【図 52】

ロングリード

MiniSeqラン:
2x250



10

20

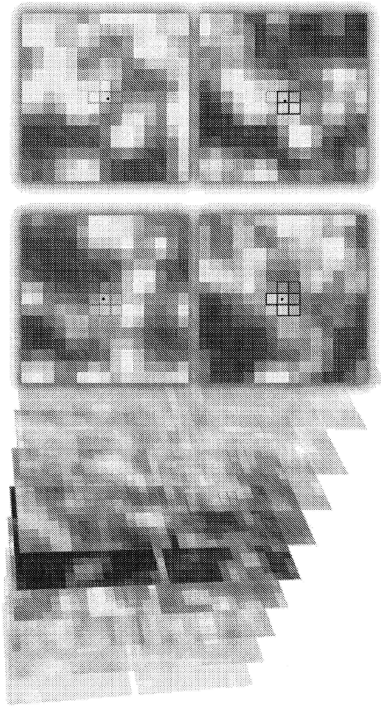
30

40

50

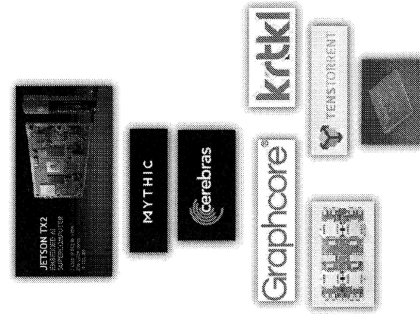
【図 5 3】

中央クラスターピクセル及びその隣接ピクセルへのアテンション



【図 5 4】

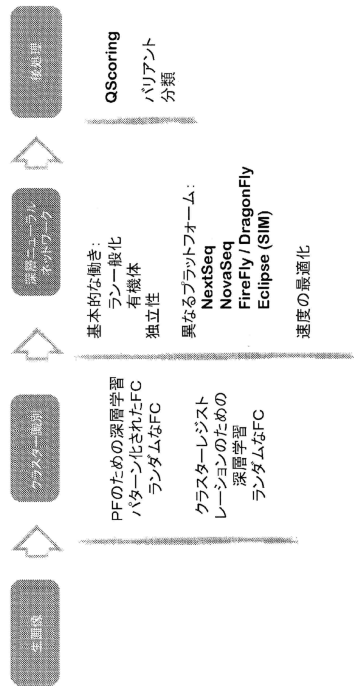
ハードウェア



- 以下のような推論エンジン:
 - Jetson (Nvidia)
 - Mythic
- ソフトウェアベースの強化:
 - FPGAを含む全てのプラットフォームとの協働
 - UCSDとのコラボレーション
- 器具上の利用可能なFPGA容量
- クラウドベースのソリューション
 - Firefly / Dragonfly
 - メモリ

【図 5 5】

ベースコールド



【図 5 6】

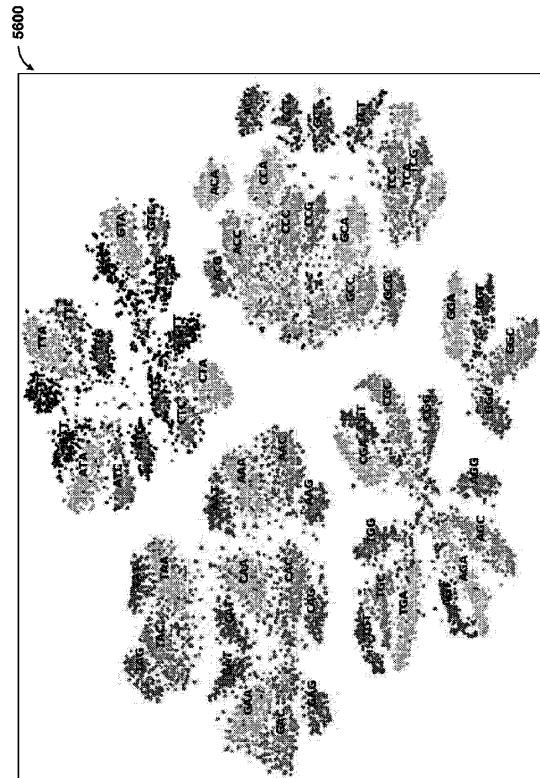
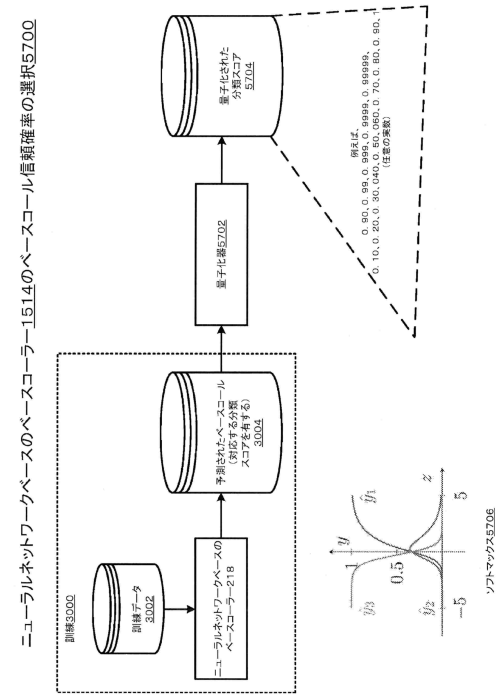
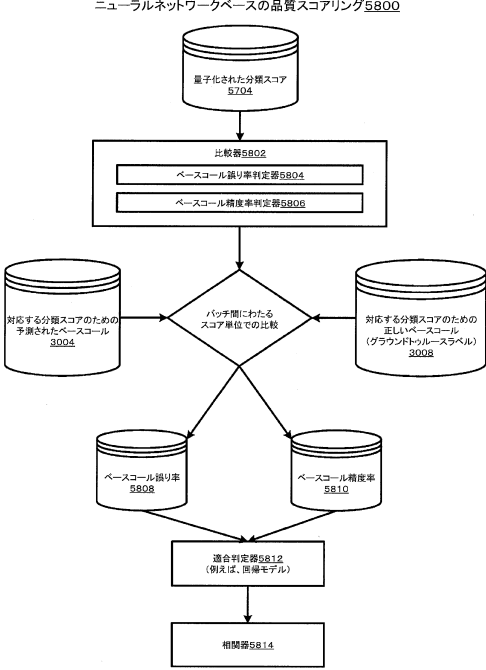


Figure 56

【図 5 7】



【図 5 8】



【図 5 9 a】

品質スコア対応5900

品質スコア	品質スコア対応スキーム5900a	ソフトマックススコア
Q10		0.90
Q20		0.99
Q30		0.999
Q40		0.9999
Q50		0.99999
Q60		0.999999

【図 5 9 b】

品質スコア対応スキーム5900b

品質スコア	品質スコア対応スキーム5900b	ソフトマックススコア
Q2-Q9 → Q06		0.80
Q10-Q19 → Q15		0.95
Q20-Q24 → Q22		0.993
Q25-Q29 → Q27		0.997
Q30-Q34 → Q33		0.9991
Q35-Q39 → Q37		0.9995
Q40 or > Q40 → Q40		0.9999

10

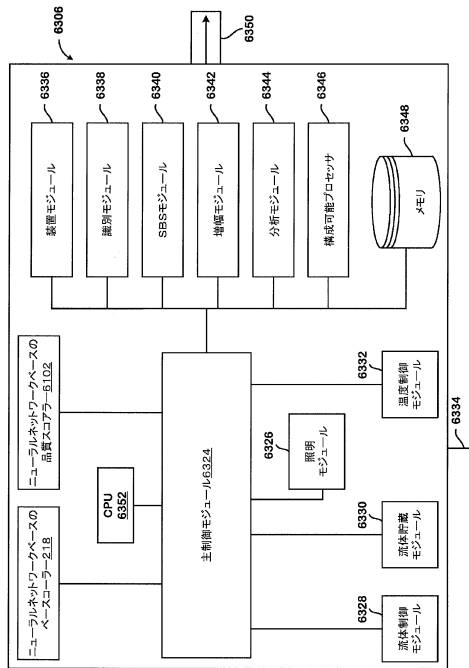
20

30

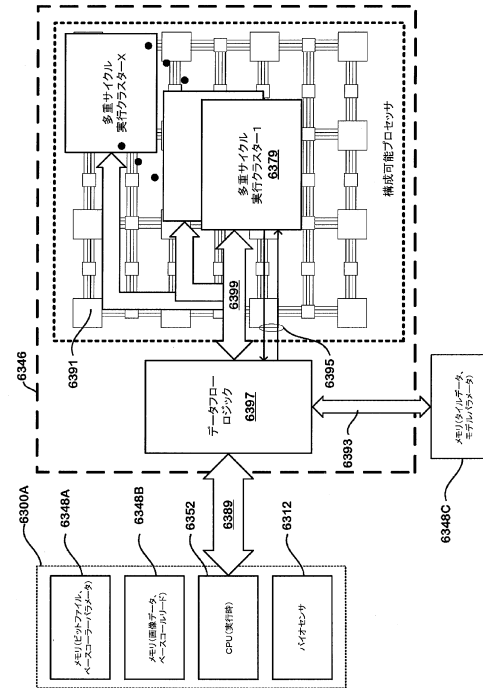
40

50

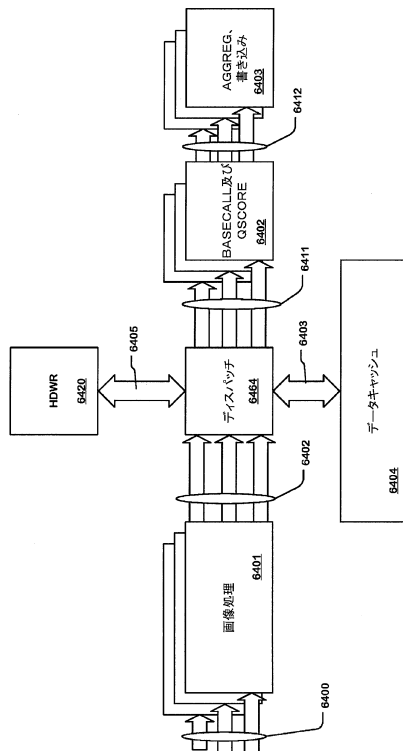
【図 6 3 B】



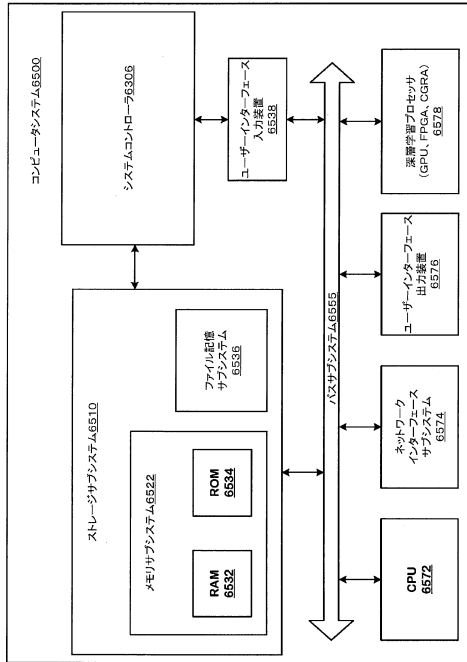
【図 6 3 C】



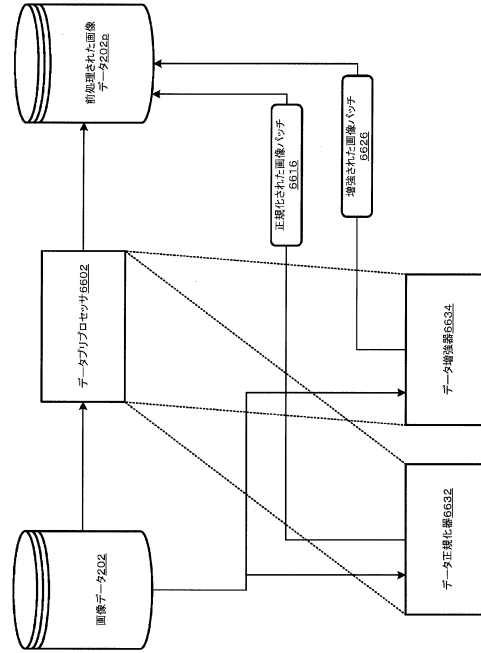
【図 6 4 A】



【図 6 5】

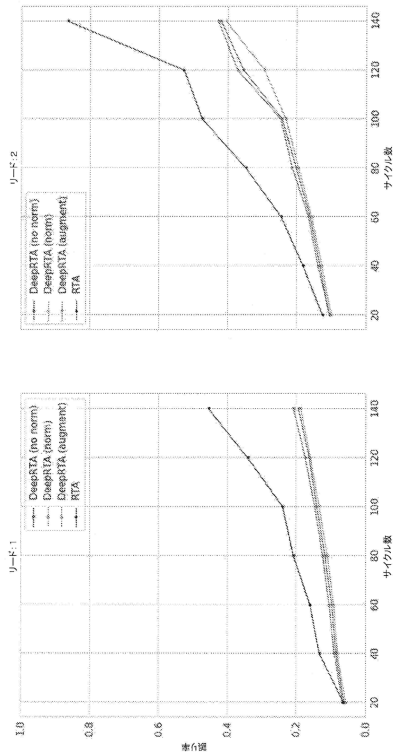


【図 6 6】



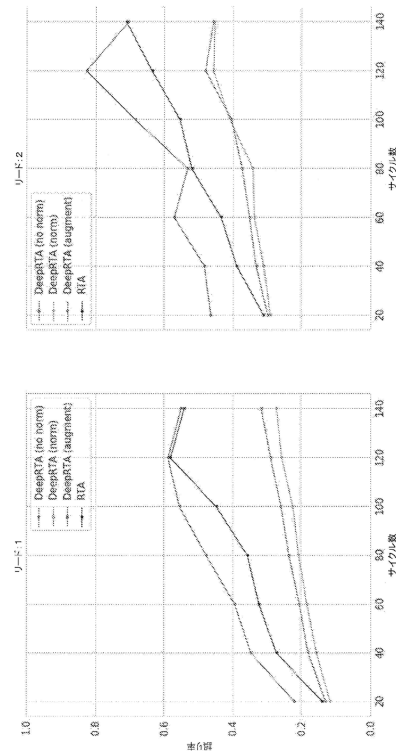
【図 6 7】

細菌上で訓練、ヒト上で試験(同じアッセイ)



【図 6 8】

アッセイ一般化-エクソーム配列決定



フロントページの続き

- (33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 62/821,681
(32)優先日 平成31年3月21日(2019.3.21)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 62/821,724
(32)優先日 平成31年3月21日(2019.3.21)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 62/821,766
(32)優先日 平成31年3月21日(2019.3.21)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 2023310
(32)優先日 令和1年6月14日(2019.6.14)
(33)優先権主張国・地域又は機関
オランダ(NL)
- (31)優先権主張番号 2023311
(32)優先日 令和1年6月14日(2019.6.14)
(33)優先権主張国・地域又は機関
オランダ(NL)
- (31)優先権主張番号 2023312
(32)優先日 令和1年6月14日(2019.6.14)
(33)優先権主張国・地域又は機関
オランダ(NL)
- (31)優先権主張番号 2023314
(32)優先日 令和1年6月14日(2019.6.14)
(33)優先権主張国・地域又は機関
オランダ(NL)
- (31)優先権主張番号 2023316
(32)優先日 令和1年6月14日(2019.6.14)
(33)優先権主張国・地域又は機関
オランダ(NL)
- (31)優先権主張番号 16/825,987
(32)優先日 令和2年3月20日(2020.3.20)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 16/825,991
(32)優先日 令和2年3月20日(2020.3.20)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 16/826,126
(32)優先日 令和2年3月20日(2020.3.20)
(33)優先権主張国・地域又は機関
米国(US)
- (31)優先権主張番号 16/826,134
(32)優先日 令和2年3月20日(2020.3.20)
(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 16/826,168

(32)優先日 令和2年3月21日(2020.3.21)

(33)優先権主張国・地域又は機関

米国(US)

(72)発明者 ジョン・ランドール・ゴベル

アメリカ合衆国・カリフォルニア・92122・サン・ディエゴ・イルミナ・ウェイ・5200

(72)発明者 アミラリ・キア

アメリカ合衆国・カリフォルニア・92122・サン・ディエゴ・イルミナ・ウェイ・5200

審査官 渡邊 加寿磨

(56)参考文献 米国特許出願公開第2009/0024331 (US, A1)

米国特許出願公開第2015/0169824 (US, A1)

米国特許出願公開第2006/0224529 (US, A1)

米国特許出願公開第2012/0020537 (US, A1)

米国特許出願公開第2018/0276333 (US, A1)

(58)調査した分野 (Int.Cl., DB名)

G16B 5/00-99/00

G16Z 99/00

C12Q 1/6869