(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7526210号
(P7526210)

(45)発行日　令和6年7月31日(2024.7.31)

(24)登録日　令和6年7月23日(2024.7.23)

(51)国際特許分類

Ｇ０６Ｆ　16/53　(2019.01)		Ｇ０６Ｆ	16/53
Ｇ１６Ｈ　30/40　(2018.01)		Ｇ１６Ｈ	30/40

ＦＩ

請求項の数　12　（全34頁）

(21)出願番号	特願2021-572430(P2021-572430)
(86)(22)出願日	令和1年6月7日(2019.6.7)
(65)公表番号	特表2022-542751(P2022-542751A)
(43)公表日	令和4年10月7日(2022.10.7)
(86)国際出願番号	PCT/EP2019/064967
(87)国際公開番号	WO2020/244775
(87)国際公開日	令和2年12月10日(2020.12.10)
審査請求日	令和4年6月7日(2022.6.7)

(73)特許権者　511079735
ライカ マイクロシステムズ シーエムエス ゲゼルシャフト ミット ベシュレンクテル ハフツング
Ｌｅｉｃａ Ｍｉｃｒｏｓｙｓｔｅｍｓ ＣＭＳ ＧｍｂＨ
ドイツ連邦共和国 ヴェッツラー エルンスト－ライツ－シュトラーセ １７－３７
Ｅｒｎｓｔ－Ｌｅｉｔｚ－Ｓｔｒａｓｓｅ １７－３７，Ｄ－３５５７８ Ｗｅｔｚｌａｒ，Ｇｅｒｍａｎｙ

(74)代理人　100114890
弁理士　アインゼル・フェリックス＝ラインハルト

(74)代理人　100098501

最終頁に続く

(54)【発明の名称】　生物学関連のデータを処理するためのシステムおよび方法、顕微鏡を制御するためのシステムおよび方法ならびに顕微鏡

(57)【特許請求の範囲】

【請求項１】

１つまたは複数のプロセッサ（１１０）と、１つまたは複数のストレージデバイス（１２０）と、を含むシステムであって、前記システムは、

生物学関連の言語ベースの検索データ（１０１）を受信し、前記生物学関連の言語ベースの検索データ（１０１）は、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述、または、生物学的機能もしくは生物学的活動の記述のうちの少なくとも１つであり、

前記１つまたは複数のプロセッサ（１１０）によって実行されるトレーニングされた言語認識機械学習アルゴリズム（２３０）によって、前記生物学関連の言語ベースの検索データ（１０１）の第１の高次元表現（２６０）を生成し、前記第１の高次元表現（２６０）は、それぞれ異なる値を有する少なくとも３つの次元を含み、

複数の生物学関連の画像ベースの入力データセットの複数の第２の高次元表現（１０５，２５０）を取得し、

前記第１の高次元表現（２６０）を、前記複数の第２の高次元表現（１０５，２５０）のうちのそれぞれの第２の高次元表現と比較する、

ように構成されており、

前記システムは、生物学的標本の画像を撮影することによって前記複数の生物学関連の画像ベースの入力データセットを取得するように構成された顕微鏡（５１０）をさらに含み、

前記システムは、トレーニングされた視覚認識機械学習アルゴリズム（２２０）により、前記複数の生物学関連の画像ベースの入力データセットの前記複数の第２の高次元表現（１０５，２５０）の前記第２の高次元表現（１０５，２５０）を生成することによって、前記第２の高次元表現（１０５，２５０）を取得するように構成されており、
前記複数の第２の高次元表現（１０５，２５０）のうちのそれぞれの第２の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含む、
システム。
【請求項２】
前記第１の高次元表現（２６０）の１つまたは複数の次元の値は、特定の生物学的機能または特定の生物学的活動が存在する尤度に比例する、
請求項１記載のシステム。
【請求項３】
前記第２の高次元表現（１０５，２５０）の１つまたは複数の次元の値は、特定の生物学的機能または特定の生物学的活動が存在する尤度に比例する、
請求項１または２記載のシステム。
【請求項４】
前記システムは、前記比較に基づいて、前記複数の第２の高次元表現（１０５，２５０）のうち、前記第１の高次元表現（２６０）に最も近い第２の高次元表現を選択するように構成されている、
請求項１から３までのいずれか１項記載のシステム。
【請求項５】
前記第１の高次元表現（２６０）の５つを超える次元の値は、前記第１の高次元表現（２６０）の次元の最大絶対値の１０％よりも大きく、
前記複数の第２の高次元表現（１０５，２５０）のうちのそれぞれの第２の高次元表現の５つを超える次元の値は、前記第２の高次元表現（１０５，２５０）の次元のそれぞれの最大絶対値の１０％よりも大きい、
請求項１から４までのいずれか１項記載のシステム。
【請求項６】
前記生物学関連の言語ベースの検索データ（１０１）は、２０文字を超える長さを含む、
請求項１から５までのいずれか１項記載のシステム。
【請求項７】
前記システムは、
第２の生物学関連の言語ベースの検索データと、論理演算子に関する情報と、を受信し、
前記１つまたは複数のプロセッサ（１１０）によって実行される前記トレーニングされた言語認識機械学習アルゴリズム（２３０）によって、前記第２の生物学関連の言語ベースの検索データ（１０２）の第１の高次元表現を生成し、
前記論理演算子に従って、第１の生物学関連の言語ベースの検索データの前記第１の高次元表現（２６０）と、第２の生物学関連の言語ベースの検索データの前記第１の高次元表現と、を組み合わせることに基づいて、１つの結合された高次元表現を決定し、
前記結合された高次元表現を、前記複数の第２の高次元表現（１０５，２５０）のうちのそれぞれの第２の高次元表現と比較する、
ように構成されている、
請求項１から６までのいずれか１項記載のシステム。
【請求項８】
１つまたは複数のプロセッサ（１１０）と、１つまたは複数のストレージデバイス（１２０）と、を含むシステムであって、前記システムは、顕微鏡（５１０）内に実装されているか、前記顕微鏡（５１０）に接続されているか、または、前記顕微鏡（５１０）を含み、前記システムは、
言語ベースの検索データ（４０１）を受信し、
前記１つまたは複数のプロセッサ（１１０）によって実行されるトレーニングされた言

語認識機械学習アルゴリズムによって、前記言語ベースの検索データ（４０１）の第１の高次元表現を生成し、前記第１の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含み、
視覚認識機械学習アルゴリズムによって、複数の画像ベースの入力データセットの複数の第２の高次元表現（４０５）を生成し、前記複数の第２の高次元表現（４０５）のうちのそれぞれの第２の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含み、
前記第１の高次元表現と、前記複数の第２の高次元表現のうちのそれぞれの第２の高次元表現（４０５）と、の比較に基づいて、前記複数の第２の高次元表現から１つの第２の高次元表現（４０５）を選択し、
選択された前記第２の高次元表現（４０５）に基づいて、前記顕微鏡（５１０）の動作を制御するための制御信号（４１１）を供給する、
ように構成されているシステム。
【請求項９】
前記システムは、選択された前記第２の高次元表現（４０５）に基づいて、顕微鏡目標位置を決定するように構成されており、
前記顕微鏡目標位置は、選択された前記第２の高次元表現（４０５）に対応する、前記画像ベースの入力データセットによって表現されている画像が撮像された位置であり、
前記制御信号（４１１）は、前記顕微鏡目標位置へと駆動するように前記顕微鏡（５１０）をトリガするように構成されている、
請求項８記載のシステム。
【請求項１０】
１つまたは複数のプロセッサ（１１０）と、１つまたは複数のストレージデバイス（１２０）と、を含むシステムが、生物学関連の言語ベースの検索データ（１０１）を処理する方法（６００）であって、前記システムは、生物学的標本の画像を撮影することによって複数の生物学関連の画像ベースの入力データセットを取得するように構成された顕微鏡（５１０）をさらに含み、前記方法は、
生物学関連の言語ベースの検索データ（１０１）を前記システムが受信するステップ（６１０）であって、前記生物学関連の言語ベースの検索データ（１０１）は、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述、または、生物学的機能もしくは生物学的活動の記述のうちの少なくとも１つであるステップ（６１０）と、
トレーニングされた言語認識機械学習アルゴリズムによって、前記生物学関連の言語ベースの検索データ（１０１）の第１の高次元表現を前記システムが生成するステップ（６２０）であって、前記第１の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含むステップ（６２０）と、
視覚認識機械学習アルゴリズムによって、前記複数の生物学関連の画像ベースの入力データセットの複数の第２の高次元表現（４０５）を前記システムが生成するステップ（６３０）であって、前記複数の第２の高次元表現（４０５）のうちのそれぞれの第２の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含むステップ（６３０）と、
前記第１の高次元表現を、前記複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と前記システムが比較するステップ（６４０）と、
を含む方法（６００）。
【請求項１１】
１つまたは複数のプロセッサ（１１０）と、１つまたは複数のストレージデバイス（１２０）と、を含むシステムが、顕微鏡（５１０）を制御する方法（７００）であって、前記システムは、前記顕微鏡（５１０）内に実装されているか、前記顕微鏡（５１０）に接続されているか、または、前記顕微鏡（５１０）を含み、前記方法は、
言語ベースの検索データを前記システムが受信するステップ（７１０）と、
トレーニングされた言語認識機械学習アルゴリズムによって、前記言語ベースの検索デ

ータの第１の高次元表現を前記システムが生成するステップ（７２０）であって、前記第１の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含むステップ（７２０）と、

視覚認識機械学習アルゴリズムによって、複数の画像ベースの入力データセットの複数の第２の高次元表現（４０５）を前記システムが生成するステップ（７３０）であって、前記複数の第２の高次元表現（４０５）のうちのそれぞれの第２の高次元表現は、それぞれ異なる値を有する少なくとも３つの次元を含むステップ（７３０）と、

前記第１の高次元表現と、前記複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と、の比較に基づいて、前記複数の第２の高次元表現から１つの第２の高次元表現を前記システムが選択するステップ（７４０）と、

選択された前記第２の高次元表現に基づいて、前記顕微鏡（５１０）の動作を前記システムが制御するステップ（７５０）と、

を含む方法（７００）。

【請求項１２】

プロセッサに請求項１０または１１記載の方法を実行させるためのコンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

各例は、生物学関連のデータの処理および／または顕微鏡の制御に関する。

【背景技術】

【０００２】

多くの生物学的用途において、膨大な量のデータが生成される。例えば、莫大な量の生物学的構造から画像が撮影され、データベース内に格納される。生物学的データを手動で分析するのは、非常に時間および費用がかかる。

【発明の概要】

【発明が解決しようとする課題】

【０００３】

したがって、生物学関連のデータを処理するため、かつ／または顕微鏡を制御するための改善されたコンセプトが必要とされている。

【課題を解決するための手段】

【０００４】

この要求は、請求項の主題によって満たすことができる。

【０００５】

いくつかの実施形態は、１つまたは複数のストレージデバイスに結合された１つまたは複数のプロセッサを含むシステムに関する。当該システムは、生物学関連の言語ベースの検索データを受信し、１つまたは複数のプロセッサによって実行されるトレーニングされた言語認識機械学習アルゴリズムによって、生物学関連の言語ベースの検索データの第１の高次元表現を生成するように構成されている。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、当該システムは、複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットの複数の第２の高次元表現を取得するように構成されている。さらに、当該システムは、第１の高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と比較するように構成されている。

【０００６】

言語認識機械学習アルゴリズムを使用することによって、生物学的テキスト検索用語を高次元表現にマッピングすることができる。高次元表現が（ワンホットエンコーディングされた表現とは対照的に）種々の異なる値を有するエントリを有することを可能にすることによって、意味論的に類似した生物学的検索用語を、類似した高次元表現にマッピングすることができる。複数の生物学関連の画像ベースの入力データセットまたは複数の生物

学関連の言語ベースの入力データセットの高次元表現を取得することにより、高次元表現が、検索用語の高次元表現に等しいことまたは類似することを検出することができる。このようにして、検索用語に対応する画像またはテキストを検出することが可能となり得る。このようにして、トレーニングされた言語認識機械学習アルゴリズムは、言語ベースの検索入力に基づいて、複数の生物学的画像（例えば、生物学的画像のデータベース）の中から生物学関連の画像を検索すること、または複数の生物学関連のテキスト（例えば、科学論文のコレクションまたはライブラリ）の中から生物学関連のテキストを検索することを可能にすることができる。既存のデータベース内での検索および／または実行中の実験によって生成された画像内での検索を、たとえこれらの画像が事前にラベル付けまたはタグ付けされていなかったとしても可能にすることができる。

【０００７】

いくつかの実施形態は、１つまたは複数のプロセッサと、１つまたは複数のストレージデバイスと、を含むシステムに関する。当該システムは、言語ベースの検索データを受信し、１つまたは複数のプロセッサによって実行されるトレーニングされた言語認識機械学習アルゴリズムによって、言語ベースの検索データの第１の高次元表現を生成するように構成されている。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、当該システムは、複数の画像ベースの入力データセットの複数の第２の高次元表現を取得し、第１の高次元表現と、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と、の比較に基づいて、複数の第２の高次元表現から１つの第２の高次元表現を選択するように構成されている。さらに、当該システムは、選択された第２の高次元表現に基づいて、顕微鏡の動作を制御するための制御信号を供給するように構成されている。

【０００８】

言語認識機械学習アルゴリズムを使用することによって、テキスト検索用語を高次元表現にマッピングすることができる。高次元表現が（ワンホットエンコーディングされた表現とは対照的に）種々の異なる値を有するエントリを有することを可能にすることによって、意味論的に類似した検索用語を、類似した高次元表現にマッピングすることができる。複数の画像ベースの入力データセットの高次元表現を取得することによって、高次元表現が、検索用語の高次元表現に等しいことまたは類似することを検出することができる。このようにして、検索用語に対応する画像を検出することが可能となり得る。この情報を用いて顕微鏡を、画像が撮影されたそれぞれの位置へと駆動して、その関心位置のさらなる画像を（例えば、より高倍率で、異なる光で、または異なるフィルタで）撮影することを可能にすることができる。このようにして、検索用語に対応する位置を検出するためにまず始めに標本（例えば、生物学的標本または集積回路）を低倍率で撮像して、その後、その関心位置をより詳細に分析することができる。

【０００９】

以下では、装置および／または方法のいくつかの例を、単なる例として、添付の図面を参照しながら説明する。

【図面の簡単な説明】

【００１０】

【図１】生物学関連のデータを処理するためのシステムの概略図である。
【図２】生物学関連のデータを処理するための別のシステムの概略図である。
【図３】生物学関連のデータを処理するための別のシステムの概略図である。
【図４】顕微鏡を制御するためのシステムの概略図である。
【図５】データを処理するためのシステムの概略図である。
【図６】生物学関連のデータを処理するための方法のフローチャートである。
【図７】顕微鏡を制御するための方法のフローチャートである。

【発明を実施するための形態】

【００１１】

次に、いくつかの例が示されている添付の図面を参照しながら、種々の例をより完全に

説明する。図面において、線、層および／または領域の厚さは、見やすくするために誇張されている場合がある。

【００１２】

したがって、さらなる例によって種々の修正形態および代替形態を実現することが可能であるが、そのうちのいくつかの特定の例が図面に示されており、続いて詳細に説明される。しかしながら、この詳細な説明は、さらなる例を、説明されている特定の形態に限定するものではない。さらなる例は、本開示の範囲内に含まれる全ての修正形態、均等形態および代替形態を網羅することができる。同一または同様の参照符号は、図面の説明全体にわたり同様または類似の要素を指しており、それらの要素を、互いに比較したとき、同一または類似の機能を提供しながら、同一または変更された形態で実現することができる。

【００１３】

ある要素が別の要素と「接続されている」または「結合されている」と記載されている場合、これらの要素は、直接的に接続または結合されていてもよいし、または１つまたは複数の介在要素を介して接続または結合されていてもよいと理解される。２つの要素ＡおよびＢが「または」を使用して組み合わされている場合、このことは、明示的または暗示的に別様に定義されていない限り、全ての可能な組み合わせ、すなわちＡのみ、Ｂのみならびにおよびが開示されているものと理解されたい。同じ組み合わせについての代替的な表現は、「およびＢのうちの少なくとも１つ」または「Ａおよび／またはＢ」である。同じことは、必要な変更を加えて、３つ以上の要素の組み合わせにも当てはまる。

【００１４】

特定の例を説明する目的で本明細書において用いられる用語は、さらなる例を限定することを意図するものではない。“ａ”，“ａｎ”および“ｔｈｅ”のような単数形が用いられ、単一の要素のみを用いることが必須であると明示的または暗示的に定義されていないときはいつでも、さらなる例が、複数の要素を用いて同じ機能を実現してもよい。同様に、ある機能が複数の要素を用いて実装されるものとして後で説明されている場合、さらなる例が、単一の要素または処理エンティティを用いて同じ機能を実現してもよい。さらに、“ｃｏｍｐｒｉｓｅｓ（含む）”、“ｃｏｍｐｒｉｓｉｎｇ（含んでいる）”、“ｉｎｃｌｕｄｅｓ（含む）”および／または“ｉｎｃｌｕｄｉｎｇ（含んでいる）”という用語は、使用される場合、記載された特徴、整数、ステップ、操作、プロセス、動作、要素および／または構成要素の存在を指定するが、１つまたは複数の他の特徴、整数、ステップ、操作、プロセス、動作、要素、構成要素および／またはそれらの任意のグループの存在または追加を排除しないと理解される。

【００１５】

別様に定義されていない限り、全ての用語（技術用語および科学用語を含む）は、本明細書において、各例が属する分野の通常の意味で使用されている。

【００１６】

図１は、１つの実施形態による、生物学関連のデータを処理するためのシステム１００の概略図を示す。システム１００は、１つまたは複数のストレージデバイス１２０に結合された１つまたは複数のプロセッサ１１０を含む。システム１００は、（第１の）生物学関連の言語ベースの検索データ１０１を受信し、１つまたは複数のプロセッサ１１０によって実行されるトレーニングされた言語認識機械学習アルゴリズムによって、（第１の）生物学関連の言語ベースの検索データ１０１の第１の高次元表現を生成するように構成されている。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリ（または互いに異なる値を有する少なくとも２０個のエントリ、少なくとも５０個のエントリ、または少なくとも１００個のエントリ）を含む。さらに、システム１００は、複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットの複数の第２の高次元表現１０５を取得するように構成されている。さらに、システム１００は、１つまたは複数のプロセッサ１１０によって、第１の高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現１０５と比較するように構成されている。

【００１７】

生物学関連の言語ベースの検索データ１０１は、生物学的構造、生物学的機能、生物学的挙動、または生物学的活動に関連するテキスト入力であってもよい。例えば、生物学関連の言語ベースの検索データ１０１は、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述および／または生物学的機能もしくは生物学的活動の記述であってもよい。テキスト入力は、実験またはデータセットの文脈での、生物学的分子（例えば、多糖類、ポリ／オリゴヌクレオチド、タンパク質、または脂質）またはその挙動を記述する自然言語であってもよい。例えば、生物学関連の言語ベースの検索データ１０１は、ヌクレオチド配列、タンパク質配列、または生物学的用語のグループの粗視化された検索用語であってもよい。

【００１８】

生物学的用語のグループは、同じ生物学的トピックに属する複数の粗視化された検索用語（またはいわゆる分子生物学的主題の見出し用語）を含むことができる。生物学的用語のグループは、触媒活性（例えば、抽出物および生成物を表す単語を使用する何らかの反応方程式として）、触媒経路（例えば、どの経路が関与しているか、例えば、糖分解）、触媒部位および／または触媒領域（例えば、結合部位、活性部位、ヌクレオチド結合部位）であってもよく、ＧＯ（遺伝子オントロジー）（例えば、分子機能、例えば、ニコチンアミドアデニンジヌクレオチドＮＡＤ結合、微小管結合）、ＧＯ生物学的機能（例えば、アポトーシス、グルコネオゲネシス）、酵素および／または経路データベース（例えば、ＢＲＥＮＤＡ／ＥＣ番号またはＵｎｉＰａｔｈｗａｙｓにおける、例えば、ｓｉｃ機能のための一意の識別子）であってもよく、細胞内局在（例えば、サイトゾル、核、細胞骨格）、ファミリーおよび／またはドメイン（例えば、翻訳後修飾のための、例えば、結合部位、モチーフ）であってもよく、オープンリーディングフレーム、一塩基多型、制限部位（例えば、制限酵素によって認識されるオリゴヌクレオチド）および／または生合成経路（例えば、脂質、多糖類、ヌクレオチド、またはタンパク質の生合成）であってもよい。この例において、生物学関連の言語ベースの検索データ１０１としてサイトゾル、核、または細胞骨格が使用される場合には、サイトゾル、核、または細胞骨格を有する画像を検索することができる。

【００１９】

生物学関連の言語ベースの検索データ１０１は、生物学関連の言語ベースの検索データ１０１として粗視化された検索用語が使用される場合には、５０文字未満（または３０文字未満もしくは２０文字未満）の長さを含むことができ、かつ／または生物学関連の言語ベースの検索データ１０１としてヌクレオチド配列またはタンパク質配列が使用される場合には、２０文字を超える（または４０文字を超える、６０文字を超える、もしくは８０文字を超える）長さを含むことができる。例えば、ヌクレオチド配列（ＤＮＡ／ＲＮＡ）は、アミノ酸に関して３つの塩基対が指定されているので、ポリペプチド配列（例えば、ペプチド、タンパク質）よりも約３倍長いことが多い。例えば、生物学関連の言語ベースの検索データ１０１は、生物学関連の言語ベースの検索データ１０１がタンパク質配列またはアミノ酸である場合には、２０文字を超える長さを含むことができる。生物学関連の言語ベースの検索データ１０１は、生物学関連の言語ベースの検索データ１０１がヌクレオチド配列または自然言語での記述テキストである場合には、６０文字を超える長さを含むことができる。例えば、生物学関連の言語ベースの検索データ１０１は、少なくとも１つの非数値文字（例えば、アルファベット文字）を含むことができる。生物学関連の言語ベースの検索データ１０１は、クエリテキスト、クエリ、入力テキスト、またはユーザ入力とも称され得る。生物学関連の言語ベースの検索データ１０１は、システム１００の入力インターフェース（例えば、キーボード）を介してユーザによって入力可能である。

【００２０】

高次元表現（例えば、第１および第２の高次元表現）は、隠れ表現、潜在ベクトル、埋め込み、意味論的埋め込みおよび／またはトークン埋め込みであってもよく、かつ／または隠れ表現、潜在ベクトル、埋め込み、意味論的埋め込みおよび／またはトークン埋め込みとも称され得る。

【００２１】
第１の高次元表現および／または第２の高次元表現は、（例えば、数値のみを含む）数値表現であってもよい。対照的に、生物学関連の言語ベースの検索データ１０１は、アルファベット文字または他の非数字文字のみを含むことができるか、またはアルファベット文字、他の非数字文字および／または数字の混合物を含むことができる。第１の高次元表現および／または第２の高次元表現は、１００を超える次元（または３００もしくは５００を超える次元）を含むことができ、かつ／または１００００未満の次元（または３０００未満または１０００未満の次元）を含むことができる。高次元表現のそれぞれのエントリは、高次元表現の１つの次元であってもよい（例えば、１００次元を有する高次元表現は、１００個のエントリを含む）。例えば、３００を超えて１０００未満の次元を有する高次元表現を使用することにより、意味論的相関を有する生物学関連のデータのために適した表現が可能となり得る。第１の高次元表現は、第１のベクトルであってもよく、それぞれの第２の高次元表現は、それぞれの第２のベクトルであってもよい。第１の高次元表現のエントリおよび第２の高次元表現のエントリのためにベクトル表現が使用される場合には、効率的な比較および／または他の計算（例えば、正規化）を実施することができるが、他の表現（例えば、行列として）も可能であり得る。例えば、第１の高次元表現および／または第２の高次元表現は、正規化されたベクトルであってもよい。第１の高次元表現および第２の高次元表現は、同じ値（例えば、１）に正規化可能である。例えば、トレーニングされた言語認識機械学習アルゴリズムの最後の層は、追加的に正規化を実行することができる非線形演算を表現することができる。例えば、トレーニングされた言語認識機械学習アルゴリズムが交差エントロピー損失関数によってトレーニングされている場合には、いわゆるソフトマックス演算：
【数１】

$$softmax = \frac{e^{\hat{y}_i}}{\sum_i^K e^{\hat{y}_i}}$$

を実施することができ、ここで、y_iは、入力値に対応するモデルの予測であり、Ｋは、全ての入力値の数である。このようにして、トレーニングされた言語認識機械学習アルゴリズムは、正規化された高次元表現を出力することができる。第２の高次元表現は、トレーニングされた視覚認識機械学習アルゴリズムによって生成可能であり、このトレーニングされた視覚認識機械学習アルゴリズムは、損失関数によってトレーニングされたものであってもよく、この損失関数は、トレーニングされた視覚認識機械学習アルゴリズムに、正規化された高次元表現を出力させる。しかしながら、第１の高次元表現および第２の高次元表現の正規化のための他のアプローチを適用することも可能であり得る。
【００２２】
例えば、第１の高次元表現および／または第２の高次元表現は、ワンホットエンコーディングされた表現とは対照的に、０に等しくない値を有する複数の異なるエントリ（少なくとも３つ）を含むことができる。第１の高次元表現に相応して、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリ（または互いに異なる値を有する少なくとも２０個のエントリ、少なくとも５０個のエントリ、または少なくとも１００個のエントリ）を含むことができる。０に等しくない値を有する複数の異なるエントリを有することができる高次元表現を使用することにより、高次元表現同士の間の意味論的な関係性に関する情報を再現することができる。例えば、第１の高次元表現のエントリの値の５０％超（または７０％超もしくは９０％超）および／または第２の高次元表現のエントリの値の５０％超（または７０％超もしくは９０％超）は、０に等しくなくてもよい。ワンホットエンコーディングされた表現が、０に等しくない２つ以上のエントリを有することも時にはあるが、高い値を有するエントリは１つだけであり、その他の全てのエントリは、ノイズレベル（例えば、その１つの高い値

の１０％未満）の値を有する。対照的に、第１の高次元表現の５つを超えるエントリ（または２０を超えるエントリもしくは５０を超えるエントリ）の値を、例えば、第１の高次元表現のエントリの最大絶対値の１０％よりも大きく（または２０％よりも大きく、もしくは３０％よりも大きく）することができる。さらに、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現の５つを超えるエントリ（または２０を超えるエントリもしくは５０を超えるエントリ）の値を、第２の高次元表現のエントリのそれぞれの最大絶対値の１０％よりも大きく（または２０％よりも大きく、もしくは３０％よりも大きく）することができる。例えば、複数の第２の高次元表現のうちの１つの第２の高次元表現の５つを超えるエントリ（または２０を超えるエントリもしくは５０を超えるエントリ）の値を、この１つの第２の高次元表現のエントリの最大絶対値の１０％よりも大きく（または２０％よりも大きく、もしくは３０％よりも大きく）することができる。例えば、第１の高次元表現および／または第２の高次元表現のそれぞれのエントリは、－１～１の間の値を含むことができる。

【００２３】

トレーニングされたパラメータのセットを有するトレーニングされた言語認識機械学習アルゴリズムの少なくとも一部（例えば、エンコーダ）を生物学関連の言語ベースの検索データ１０１に適用することによって、第１の高次元表現を決定することができる。例えば、トレーニングされた言語認識機械学習アルゴリズムによって第１の高次元表現を生成することは、トレーニングされた言語認識機械学習アルゴリズムのエンコーダによって第１の高次元表現を生成することを意味していてもよい。トレーニングされた言語認識機械学習アルゴリズムの、トレーニングされたパラメータのセットは、以下に説明するように、言語認識機械学習アルゴリズムのトレーニング中に取得可能である。

【００２４】

第１の高次元表現の１つまたは複数のエントリの値および／または第２の高次元表現の１つまたは複数のエントリの値は、特定の生物学的機能または特定の生物学的活動が存在する尤度に比例することができる。入力データセットの意味論的な類似性を維持する高次元表現を生成するマッピングを使用することにより、意味論的に類似した高次元表現同士は、意味論的に類似性の低い高次元表現同士よりも互いにより近い距離を有することができる。さらに、２つの高次元表現が、同じまたは類似した特定の生物学的機能または特定の生物学的活動を有する入力データセットを表現している場合には、これらの２つの高次元表現の１つまたは複数のエントリは、同じまたは類似した値を有することができる。意味論の維持により、高次元表現の１つまたは複数のエントリは、特定の生物学的機能または特定の生物学的活動の発生または存在を示すことができる。例えば、高次元表現の１つまたは複数のエントリの値が高くなればなるほど、これらの１つまたは複数のエントリと相関する生物学的機能または生物学的活動が存在する尤度がより高くなり得る。

【００２５】

トレーニングされた言語認識機械学習アルゴリズムは、テクスチュアルモデル、テキストモデル、または言語モデルとも称され得る。言語認識機械学習アルゴリズムは、トレーニングされた言語認識ニューラルネットワークであってもよいし、またはトレーニングされた言語認識ニューラルネットワークを含んでいてもよい。トレーニングされた言語認識ニューラルネットワークは、３０を超える層（または５０もしくは８０を超える層）および／または５００未満の層（または３００もしくは２００未満の層）を含むことができる。トレーニングされた言語認識ニューラルネットワークは、リカレントニューラルネットワーク、例えば長短期記憶ネットワークであってもよい。リカレントニューラルネットワーク、例えば長短期記憶ネットワークを使用することにより、生物学関連の言語ベースのデータのための高精度の言語認識機械学習アルゴリズムを提供することができる。しかしながら、他の言語認識アルゴリズムを適用することも可能であり得る。例えば、トレーニングされた言語認識機械学習アルゴリズムは、可変長の入力データを扱うことができるアルゴリズム（例えば、Ｔｒａｎｓｆｏｒｍｅｒ－ＸＬアルゴリズム）であってもよい。例えば、第１の生物学関連の言語ベースの検索データの長さは、第２の生物学関連の言語ベ

ースの検索データの長さとは異なる。例えば、タンパク質配列は、典型的に、数十から数百のアミノ酸の長さである（１つのアミノ酸は、タンパク質配列における１文字として表現される）。「意味論」、例えば（生物学ではポリペプチド、モチーフ、またはドメインと称される）配列からの部分文字列の生物学的機能は、長さに関して種々異なり得る。したがって、可変長の入力を受信することができるアーキテクチャを使用することができる。

【００２６】

複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットの複数の第２の高次元表現１０５は、（例えば、１つまたは複数のストレージデバイスによって格納された）データベースから第２の高次元表現１０５を受信することによって取得可能であるか、あるいは複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットに基づいて、複数の第２の高次元表現１０５を生成することによって取得可能である。例えば、システム１００は、複数の第２の高次元表現が複数の生物学関連の画像ベースの入力データセットに基づいている場合には、１つまたは複数のプロセッサによって実行されるトレーニングされた視覚認識機械学習アルゴリズムにより、複数の第２の高次元表現の第２の高次元表現を生成することによって、第２の高次元表現を取得するように構成可能である。例えば、トレーニングされた視覚モデルは、（例えば、第２の高次元表現として）意味論的埋め込み空間における画像を表現することが可能であり得る。択一的に、システム１００は、複数の第２の高次元表現が複数の生物学関連の言語ベースの入力データセットに基づいている場合には、１つまたは複数のプロセッサによって実行されるトレーニングされた言語認識機械学習アルゴリズムにより、複数の第２の高次元表現の第２の高次元表現を生成することによって、第２の高次元表現を取得するように構成可能である。

【００２７】

複数の生物学関連の画像ベースの入力データセットのうちのそれぞれの生物学関連の画像ベースの入力データセットは、ヌクレオチドもしくはヌクレオチド配列を含む生物学的構造、タンパク質もしくはタンパク質配列を含む生物学的構造、生物学的分子、生物学的組織、特定の挙動を有する生物学的構造および／または特定の生物学的機能もしくは特定の生物学的活動を有する生物学的構造の画像の画像データ（例えば、画像のピクセルデータ）であってもよい。生物学的構造は、分子、ウイロイドもしくはウイルス、人工もしくは天然の膜で包囲された小胞、（細胞小器官のような）細胞内構造、細胞、スフェロイド、オルガノイド、三次元細胞培養、生物学的組織、臓器スライス、または生体内もしくは生体外の臓器の一部であってもよい。例えば、生物学的構造の画像は、細胞内または組織内のタンパク質の位置の画像であってもよいし、または標識されたヌクレオチドプローブが結合する内因性ヌクレオチド（例えば、ＤＮＡ）を有する細胞または組織の画像（例えば、ｉｎ　ｓｉｔｕハイブリダイゼーション）であってもよい。画像データは、画像のそれぞれの色次元（例えば、ＲＧＢ表現の場合には３つの色次元）について、画像のそれぞれのピクセルに対するピクセル値を含むことができる。例えば、画像診断法に応じて、励起波長または発光波長、蛍光寿命、偏光、三空間次元でのステージ位置、種々異なる撮像角度に関連して、他のチャネルを適用してもよい。生物学関連の画像ベースの入力データセットは、ＸＹピクセルマップ、体積測定データ（ＸＹＺ）、時系列データ（ＸＹ＋Ｔ）、またはそれらの組み合わせ（ＸＹＺＴ）であってもよい。さらに、画像ソースの種類に応じて、チャネル（例えば、スペクトル発光帯）、励起波長、ステージ位置、マルチウェルプレートまたはマルチポジショニング実験および／またはミラーでのような論理的位置および／またはライトシート撮像でのような対物レンズ位置のような、追加的な次元を含むことができる。例えば、ピクセルマップまたは高次元のピクチャーとして画像をユーザが入力してもよいし、またはデータベースが提供してもよい。トレーニングされた視覚認識機械学習アルゴリズムは、この画像を意味論的埋め込み（例えば、第２の高次元表現）に変換することができる。複数の生物学関連の画像ベースの入力データセットは、１つまたは複数のストレージデバイスから、もしくはストレージデバイスによって格納されたデータベースから受信可能である。

【００２８】
生物学関連の言語ベースの検索データ１０１と同様に、複数の生物学関連の言語ベースの入力データセットのうちのそれぞれの生物学関連の言語ベースの入力データセットは、生物学的構造、生物学的機能、生物学的挙動、または生物学的活動に関連するテキスト入力であってもよい。例えば、複数の生物学関連の言語ベースの入力データセットのうちのそれぞれの生物学関連の言語ベースの入力データセットは、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述および／または生物学的機能もしくは生物学的活動の記述であってもよい。テキスト入力は、実験またはデータセットの文脈での、生物学的分子（例えば、多糖類、ポリ／オリゴヌクレオチド、タンパク質、または脂質）またはその挙動を記述する自然言語であってもよい。例えば、複数の生物学関連の言語ベースの入力データセットのうちのそれぞれの生物学関連の言語ベースの入力データセットは、ヌクレオチド配列、タンパク質配列、または生物学的用語のグループの粗視化された検索用語であってもよい。
【００２９】
トレーニングされた視覚認識機械学習アルゴリズムは、画像認識モデルまたは視覚モデルとも称され得る。トレーニングされた視覚認識機械学習アルゴリズムは、トレーニングされた視覚認識ニューラルネットワークであってもよいし、またはトレーニングされた視覚認識ニューラルネットワークを含んでいてもよい。トレーニングされた視覚認識ニューラルネットワークは、２０を超える層（または４０もしくは８０を超える層）および／または４００未満の層（または２００もしくは１５０未満の層）を含むことができる。トレーニングされた視覚認識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またはカプセルネットワークであってもよい。畳み込みニューラルネットワークまたはカプセルネットワークを使用することにより、生物学関連の画像ベースのデータのための高精度のトレーニングされた視覚認識機械学習アルゴリズムを提供することができる。しかしながら、他の視覚認識アルゴリズムを適用することも可能であり得る。例えば、トレーニングされた視覚認識ニューラルネットワークは、複数の畳み込み層および複数のプーリング層を含むことができる。しかしながら、カプセルネットワークが使用される場合、かつ／または例えば、畳み込みのためにストライド＝１の代わりにストライド＝２が使用される場合には、プーリング層を回避することができる。トレーニングされた視覚認識ニューラルネットワークは、正規化線形ユニット活性化関数を使用することができる。正規化線形ユニット活性化関数を使用することにより、生物学関連の画像ベースの入力データのための高精度のトレーニングされた視覚認識機械学習アルゴリズムを提供することができるが、他の活性化関数（例えば、ハードタン活性化関数、シグモイド活性化関数、またはタン活性化関数）を適用することも可能であり得る。例えば、トレーニングされた視覚認識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を含んでいてもよく、かつ／または入力画像のサイズに応じた深さのＲｅｓＮｅｔまたはＤｅｎｓｅＮｅｔであってもよい。
【００３０】
１つまたは複数のプロセッサ１１０は、第１の高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と比較するように構成可能である。第１の高次元表現と第２の高次元表現との間の距離を計算することによって、第１の高次元表現を、第２の高次元表現と比較することができる。第１の高次元表現および第２の高次元表現がベクトル（例えば、正規化されたベクトル）によって表現されている場合には、第１の高次元表現と第２の高次元表現との間の距離（例えば、ユークリッド距離またはアースムーバー距離（Earth mover’s distance））を、わずかな労力で計算することができる。複数の第２の高次元表現のうちのそれぞれの第２の高次元表現ごとに、距離の計算を繰り返すことができる。例えば、第１の高次元表現と、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と、の比較は、ユークリッド距離関数またはアースムーバー距離関数に基づく。計算された距離に基づいて、システム１００は、選択基準（例えば、最も近い距離を有するか、または距離しきい値の範囲内にある１つまたは複数の第２の高次元表現）

に基づいて、１つまたは複数の第２の高次元表現を選択することができる。例えば、システム１００は、比較に基づいて、複数の第２の高次元表現のうち、第１の高次元表現に最も近い第２の高次元表現を選択するように構成可能である。システム１００は、選択基準を満たす１つまたは複数の第２の高次元表現、複数の生物学関連の画像ベースの入力データセットのうちの、上記１つまたは複数の第２の高次元表現に対応する１つまたは複数の生物学関連の画像ベースの入力データセットおよび／または複数の生物学関連の言語ベースの入力データセットのうちの、上記１つまたは複数の第２の高次元表現に対応する１つまたは複数の生物学関連の言語ベースの入力データセットを出力または格納することができる。例えば、システム１００は、最も近い第２の高次元表現、複数の生物学関連の画像ベースの入力データセットのうちの、上記最も近い第２の高次元表現に対応する生物学関連の画像ベースの入力データセットおよび／または複数の生物学関連の言語ベースの入力データセットのうちの、上記最も近い第２の高次元表現に対応する生物学関連の言語ベースの入力データセットを出力または格納することができる。

【００３１】

０に等しくない複数のエントリを有する高次元表現を使用することに起因して、２つ以上の検索用語の論理的な組み合わせを検索するために、２つ以上の高次元表現を組み合わせることが可能となり得る。例えば、ユーザは、２つ以上の検索用語と、１つまたは複数の論理演算子（例えば、ＡＮＤ演算子またはＮＯＴ演算子）とを入力することができ、この論理演算子に基づいて、対応する生成された第１の高次元表現を組み合わせることができる。例えば、当該システムは、第２の生物学関連の言語ベースの検索データと、論理演算子に関する情報とを受信するように構成可能である。さらに、システム１００は、１つまたは複数のプロセッサによって実行されるトレーニングされた言語認識機械学習アルゴリズムによって、第２の生物学関連の言語ベースの検索データの第１の高次元表現を生成することができる。さらに、システム１００は、論理演算子に従って、第１の生物学関連の言語ベースの検索データの第１の高次元表現と、第２の生物学関連の言語ベースの検索データの第１の高次元表現と、を組み合わせることに基づいて、１つの結合された高次元表現を決定することができる。結合された高次元表現は、正規化された高次元表現（例えば、正規化されたベクトル）であってもよい。

【００３２】

さらに、システム１００は、結合された高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と比較することができる。結合された高次元表現と、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と、の比較に基づいて、選択基準（例えば、最も近い距離を有するか、または距離しきい値の範囲内にある１つまたは複数の第２の高次元表現）に基づいて、１つまたは複数の第２の高次元表現を選択することができる。

【００３３】

システム１００は、選択基準を満たす１つまたは複数の第２の高次元表現、複数の生物学関連の画像ベースの入力データセットのうちの、上記１つまたは複数の第２の高次元表現に対応する１つまたは複数の生物学関連の画像ベースの入力データセットおよび／または複数の生物学関連の言語ベースの入力データセットのうちの、上記１つまたは複数の第２の高次元表現に対応する１つまたは複数の生物学関連の言語ベースの入力データセットを出力または格納することができる。選択された１つまたは複数の生物学関連の画像ベースの入力データセット（例えば、生物学的画像）、または選択された１つまたは複数の生物学関連の言語ベースの入力データセット（例えば、生物学的テキスト）は、第１の生物学関連の言語ベースの検索データと、第２の生物学関連の言語ベースの検索データと、論理演算子に関する情報とによって表現されるような検索用語の論理的な組み合わせを含む生物学的構造を示すまたは記述することができる。このようにして、２つ以上の検索用語の論理的な組み合わせの検索が可能となり得る。論理演算子は、ＡＮＤ演算子、ＯＲ演算子、またはＮＯＴ演算子であってもよい。ＮＯＴ演算子は、望ましくないヒットを抑制することができる。ＮＯＴ演算は、否定された検索用語の検索によって決定可能である。例

えば、否定された検索用語の埋め込み（例えば、第１の高次元表現）を生成および反転することができる。次いで、否定された検索用語の埋め込みに最も近いｋ個の埋め込みを、画像に関連付けられた複数の埋め込み（複数の第２の高次元表現）の中から決定して、これらの複数の埋め込みから除去することができる。オプションとして、残りの複数の埋め込みの平均（例えば、メドイド平均または算術平均）を求めることができる。この新たに計算された第２の高次元表現は、より精確なヒットを得るための、埋め込み空間における新たなクエリのために機能することができる。ＯＲ演算は、それぞれの検索用語に対して最も近い要素またはｋ個の最も近い要素（第２の高次元表現）を決定することによって実施可能であり、なお、ｋは、２～Ｎの整数である。例えば、全てのＯＲ結合された検索用語は、ループで検索可能であり、最も近いヒットまたはｋ個の最も近いヒットを出力することができる。さらに、式を解析して検索を順番にまたは内側から外側へと処理することによって、複数の論理演算子を組み合わせることが可能となり得る。

【００３４】

例えば、論理演算子は、ＡＮＤ演算子であり、結合された高次元表現は、第１の生物学関連の言語ベースの検索データの第１の高次元表現と、第２の生物学関連の言語ベースの検索データの第１の高次元表現と、を加算および／または平均することによって決定される。例えば、第１の生物学関連の言語ベースの検索データの第１の高次元表現と、第２の生物学関連の言語ベースの検索データの第１の高次元表現との算術平均を求めることができる。例えば、算術平均は、

【数２】

$$\frac{1}{N}\sum_{i}^{N}\hat{y_i}$$

によって求めることができ、ここで、ｙ ｉは、第１の高次元表現であり、Ｎは、平均されるべきベクトルの数（例えば、論理的に組み合わされる検索用語の数）である。算術平均を求めると、結果として、正規化された高次元表現を得ることができる。択一的に、幾何平均、調和平均、二乗平均、またはメドイドを使用してもよい。メドイドは、穴（例えば、データ点のない囲まれた領域）を有する分布に対する大きな誤差を回避するために使用可能である。メドイドは、平均に最も近い要素を検出することができる。メドイドｍは、

【数３】

$$m = argmin_{y_i \in Y} \sum_{i=1}^{N} d(\hat{y}, y_i)$$

のように定義可能であり、ここで、Ｙは、埋め込み全体（複数の第２の高次元表現）であり、ｙ ｉは、第２の高次元表現のうちの１つであり、

【数４】

$\hat{y}$

は、検索用語に対応する埋め込み（第１の高次元表現）であり、ｄは、距離メトリック（例えば、ユークリッド距離またはＬ２－ノルム）である。例えば、平均に最も近いＹの要素を検出することができ、その後、メドイドに最も近いｋ個の要素を（例えば、クイックソートアルゴリズムによって）決定することができる。

【００３５】

上述のように、生物学関連の言語ベースの検索データ１０１は、種々異なる種類のもの（例えば、ヌクレオチド配列、タンパク質配列、または生物学的用語のグループの粗視化された検索用語）であってもよい。単一の言語認識機械学習アルゴリズムは、１つの種類の入力のみを扱うようにトレーニング可能である。したがって、システム１００は、生物

学関連の言語ベースの検索データ１０１に基づいて、複数のトレーニングされた言語認識機械学習アルゴリズムから、トレーニングされた言語認識機械学習アルゴリズムを選択するように構成可能である。例えば、複数のトレーニングされた言語認識機械学習アルゴリズムは、１つまたは複数のストレージデバイス１２０によって格納可能であり、システム１００は、生物学関連の言語ベースの検索データ１０１として受信した入力の種類に応じて、複数のトレーニングされた言語認識機械学習アルゴリズムのうちの１つを選択することができる。例えば、トレーニングされた言語認識機械学習アルゴリズムは、生物学関連の言語ベースの検索データ１０１を分類するように構成された分類アルゴリズム（例えば、言語認識機械学習アルゴリズム）によって、複数のトレーニングされた言語認識機械学習アルゴリズムから選択可能である。

【００３６】

システム１００は、顕微鏡内に実装されていてもよいし、顕微鏡に接続されていてもよいし、または顕微鏡を含んでいてもよい。顕微鏡は、１つまたは複数の生物学的標本の画像を撮影することによって複数の生物学関連の画像ベースの入力データセットを取得するように構成可能である。複数の生物学関連の画像ベースの入力データセットは、１つまたは複数のストレージデバイス１２０によって格納可能であり、かつ／または複数の第２の高次元表現を生成するために提供可能である。

【００３７】

システム１００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図２～図７）の１つまたは複数の例に関連して言及される。システム１００は、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００３８】

図２は、１つの実施形態による、生物学関連のデータを処理するためのシステム２００の概略図を示す。ユーザは、タンパク質配列、ヌクレオチド配列、または自然言語の形態でのようなテキスト２０１（例えば、生物学関連の言語ベースの検索データ）を使用してクエリを開始することができる。例えば、システム２００は、テキストモデルの意味論的埋め込みに対してトレーニングされた視覚モデル２２０（例えば、ＣＮＮ）を含み、この視覚モデル２２０は、大量のタンパク質配列（例えば、タンパク質配列データベース）、ヌクレオチド配列（例えば、ヌクレオチド配列データベース）、科学出版物（例えば、生物学関連の出版物のデータベース）、または関心オブジェクトの役割および／または生物学的機能を記述する、ブログ投稿、研究グループのホームページ、オンライン記事、ディスカッションフォーラム、もしくはソーシャルメディア投稿のような他のテキストに対してトレーニングされている。例えば、視覚モデル２２０は、以下に説明するようにトレーニング中にこれらの意味論的埋め込みを予測することを学習しているが、モデルをトレーニングするための他の手法も可能であり得る。ユーザ入力２０１（例えば、クエリテキスト）を、まず始めにテキストモデル２１０によってそれぞれのクラス（例えば、タンパク質配列、ヌクレオチド配列、または自然言語）に分類することができ、システム２００は、入力テキストのクラスを処理するために必要な１つまたは複数のテキストモデルを含んでいるこのようなモデルのリポジトリから、そのクラスのための正しい第２のテキストモデル２３０を検出することができる。次いで、クエリテキスト２０１は、それぞれの事前にトレーニングされた言語モデル２３０（トレーニングされた言語認識機械学習アルゴリズム）を通過する順方向パスを使用して、各自のそれぞれの埋め込み２６０（第１の高次元表現）に変換される。（例えば、１つまたは複数のストレージデバイスによって格納された）データベース２４０内の画像データ、または顕微鏡での実行中の実験の一部としての画像データは、事前にトレーニングされた視覚モデル２２０（トレーニングされた視覚認識機械学習アルゴリズム）を通過する順方向パスを介して、各自のそれぞれの埋め込み２５０（複数の第２の高次元表現）に変換可能である。例えば、パフォーマンス上の理由から、この部分を、ユーザクエリの前に実施して、（例えば、１つまたは複数のストレー

ジデバイスによって格納された）適切なデータベース２５５内に格納してもよいし、または例えば、画像データと共に格納してもよい。データベース２４０とデータベース２５５とは、同一または同じであってもよいが、それぞれ異なるデータベースであってもよい。しかしながら、実行中の実験でのように画像が単数または少数である場合には、画像の順方向パスをオンザフライで実施して、これにより、視覚的な埋め込み２５０の中間ストレージ２５５をバイパス２５７することができる。例えば、画像リポジトリ２４０は、パブリックまたはプライベートのデータベースを表すことができるか、または実行中の実験中における顕微鏡の記録媒体を表すことができる。生成された２種類の埋め込み、すなわち、クエリテキスト２６０のための１つの埋め込みと、画像２５０のための埋め込みとを、埋め込み空間において比較２７０することができる（例えば、これらの相対距離を計算することができる）。この比較のために、ユークリッド距離またはアースムーバー距離のような種々の距離メトリックを使用することができる。他の距離メトリックを使用することも可能である（例えば、クラスタリングにおいて使用される距離メトリック）。例えば、最も近い埋め込み２８０を決定することができ、それぞれの画像２９０をリポジトリ２４０内でルックアップして、ユーザに返送することができる。返送すべき画像の数は、ユーザによって事前に決定可能であるか、または距離しきい値もしくは他の基準に従って計算可能である。例えば、１つまたは複数の最も近い埋め込みを検索することにより、複数の埋め込み２５０（複数の第２の高次元表現）の中からｋ個の最も近い要素を提供することができ、なお、ｋは、整数である。例えば、検索クエリの埋め込みと、複数の埋め込み２５０の全ての要素との間のユークリッド距離（Ｌ２ノルム）を求めることができる。結果として得られた距離（例えば、複数の埋め込みにおける要素と同数）をソートすることができ、最短距離を有する要素、または最短距離を有するｋ個の要素を出力することができる。

【００３９】

例えば、第１段階の（テキスト）モデル２３０は、上述の粗視化された検索用語のうちの１つまたは複数を予測するようにトレーニングされている。ユーザは、クエリテキスト２０１を入力することができ、ドロップダウンリストから、またはクエリ言語を使用して、それぞれのフィールド（例えば、触媒活性、経路、またはその他）を選択することができる。入力クエリは、粗視化された検索用語に即した統制語彙を使用することができる。択一的に、ユーザは、クエリテキスト２０１を入力することができ、機械インテリジェンス２１０は、それぞれのフィールドを自動的に決定することができる。それぞれの検索フィールドは、クエリテキストを埋め込み２６０に変換するために、どのテキストモデル２３０を使用すべきかを決定することができる。例えば、以下に説明するように、それぞれのテキストモデルの埋め込みを使用して事前にトレーニングされた適切な視覚モデル２２０（例えば、ＣＮＮ）を用いて、画像埋め込みを、各自のそれぞれの埋め込みに変換することができる。

【００４０】

システム２００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１、図３～図７）の１つまたは複数の例に関連して言及される。システム２００は、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００４１】

図３は、１つの実施形態による、生物学関連のデータを処理するためのシステム３００の概略図を示す。ユーザは、タンパク質配列、ヌクレオチド配列、または自然言語の形態でのようなテキスト２０１（例えば、生物学関連の言語ベースの検索データ）、粗視化された検索用語、または画像を使用してクエリを開始することができる。オプションとして、適切な分類器２１０（例えば、入力の種類に応じてニューラルネットワーク、統計的機械学習アルゴリズム）を使用して、クエリ２０１の事前分類を実行してもよい。いくつかの実施形態では、事前分類をスキップ３１５することができる。事前分類の結果を使用し

て、適切なモデル２３０を選択することができ、この適切なモデル２３０は、特徴抽出器として機能する事前にトレーニングされたモデル２３０によって、ユーザクエリ２０１を、その関連する意味論的埋め込み２６０に変換することができる。

【００４２】

この意味論的埋め込み空間において、データソース２４０から到来したユーザ入力と画像とが結合および処理される。データソース２４０は、プライベートまたはパブリックのデータリポジトリであってもよいし、または顕微鏡のような撮像装置であってもよい。データは、画像、テキスト、粗視化された検索用語、またはデータソースによって記録された機器固有のデータのような種類であってもよい。例えば、テキストモデルの意味論的埋め込みに対してトレーニングされた視覚モデル２２０（例えば、ＣＮＮ）を含むことができ、この視覚モデル２２０は、大量のタンパク質配列（例えば、タンパク質配列データベース）、ヌクレオチド配列（例えば、ヌクレオチド配列データベース）、科学出版物（例えば、生物学関連の出版物のデータベース）、または関心オブジェクトの役割および／または生物学的機能を記述する、ブログ投稿、研究グループのホームページ、オンライン記事、ディスカッションフォーラム、もしくはソーシャルメディア投稿のような他のテキストに対してトレーニングされている。視覚モデル２２０は、トレーニング中にこれらの意味論的埋め込みを予測するように事前にトレーニング可能である。視覚モデル２２０と、入力特徴抽出器２３０（例えば、テキストモデル）との両方が、例えば、同じ埋め込み空間においてトレーニングされる。別のアプローチでは、視覚モデル２２０と、事前にトレーニングされたモデル２３０とは、同一であってもよい（例えば、検索入力が画像ベースである場合）。次いで、クエリ２０１は、入力特徴抽出器２３０を通過する順方向パスを使用して、各自のそれぞれの埋め込み２６０に変換される。データベースであるか、または顕微鏡での実行中の実験の一部であるデータソース２４０からのデータは、事前にトレーニングされたモデル２２０（視覚モデル）を通過する順方向パスを介して、各自のそれぞれの埋め込み２５０に変換可能である。例えば、パフォーマンス上の理由から、この手順を、ユーザクエリの前に実施して、意味論的埋め込みを、適切なデータベース２５５内に格納してもよいし、または例えば、画像データと共に格納してもよい。データベース２４０とデータベース２５５とは、同一または同じであってもよいが、それぞれ異なるデータベースであってもよい。しかしながら、実行中の実験でのように画像が単数または少数である場合には、画像の順方向パスをオンザフライで実施して、これにより、視覚的な埋め込みの中間ストレージ２５５をバイパス２５７することができる。ここで、生成された２種類の埋め込み、すなわち、クエリ２６０のための１つの埋め込みと、画像２５０のための埋め込みとを、埋め込み空間において比較２７０することができる（例えば、これらの相対距離を計算することができる）。この比較のために、ユークリッド距離またはアースムーバー距離のような種々の距離メトリックを使用することができる。他の距離メトリックを使用することも可能である。例えば、クラスタリングにおいて使用される距離メトリックが機能することができる。

【００４３】

システム３００は、最も近い埋め込み２８０を決定することができ、それぞれのデータ（例えば、画像）をリポジトリ２４０内で、または実行中の実験においてルックアップして、それらのデータを返送３８１することができる。最後のステップは、実施形態の正確な目的に応じて異なる下流のプロセスステップをもたらすことができる。いくつかのケースでは、サンプル座標およびステージ座標に関して発見されたオブジェクトの座標のようなデータを、実行中の実験の過程を変化させることができる画像ソース（例えば、顕微鏡）に供給する必要がある場合がある（３８３）。いくつかの実施形態では、それぞれのデータをユーザ３８５に出力することができ、このユーザ３８５が、実行中の実験を調整すること、またはデータをさらに処理することを決定することができる。他の実施形態は、将来の検索のためにそれぞれのデータをデータベース３８７にアーカイブすることができる。択一的に、依然として意味論的埋め込み空間にあるそれぞれのデータを、任意の入力データ型に変換し戻すことができ、また、依然として意味論的埋め込み空間にあるそれぞ

れのデータを使用して、パブリックのデータベース３８９にクエリして、科学出版物、ソーシャルメディアエントリまたはブログ投稿３９０、配列アライメント３９５を介して識別されたものと同じ生物学的分子の画像３９３または生物学的配列の画像を探索することができる。検出された全ての情報を、ユーザ３８５に返送することができ、かつ／または現在実行中の実験中に記録された画像、または探索されたデータが由来するリポジトリ内に記録された画像の機能的な注釈として、データベース３８７に書き込むことができる。

【００４４】

図３は、テキストクエリを使用したテキスト・ツー・イメージ検索の例を示す。画像リポジトリ２４０は、１つの実施形態では、パブリックまたはプライベートのデータベースを表すことができ、別の実施形態では、実行中の実験中における顕微鏡の記録媒体を表すことができる。

【００４５】

システム３００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図２および図４～図７）の１つまたは複数の例に関連して言及される。システム３００は、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００４６】

図４は、１つの実施形態による、顕微鏡を制御するためのシステム４００の概略図を示す。システム４００は、１つまたは複数のプロセッサ１１０と、１つまたは複数のストレージデバイス１２０と、を含む。システム４００は、言語ベースの検索データ４０１を受信し、１つまたは複数のプロセッサ１１０によって実行されるトレーニングされた言語認識機械学習アルゴリズムによって、言語ベースの検索データ４０１の第１の高次元表現を生成するように構成されている。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリ（または互いに異なる値を有する少なくとも２０個のエントリ、少なくとも５０個のエントリ、または少なくとも１００個のエントリ）を含む。さらに、システム４００は、複数の画像ベースの入力データセットの複数の第２の高次元表現４０５を取得し、１つまたは複数のプロセッサ１１０によって実行される、第１の高次元表現と、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現４０５と、の比較に基づいて、複数の第２の高次元表現から１つの第２の高次元表現４０５を選択するように構成されている。さらに、システム４００は、選択された第２の高次元表現４０５に基づいて、顕微鏡の動作を制御するための制御信号４１１を供給するように構成されている。

【００４７】

言語ベースの検索データ４０１は、テキスト入力であってもよく、検出または分析されるべき標本の特徴を記述してもよい。分析されるべき標本は、生物学的標本、集積回路、または顕微鏡によって撮像可能な任意の他の標本であってもよい。例えば、標本が生物学的標本である場合には、言語ベースの検索データ４０１は、生物学的関連の言語ベースの検索データであってもよく、例えば、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述および／または生物学的機能もしくは生物学的活動の記述であってもよい。テキスト入力は、実験またはデータセットの文脈での、生物学的分子（例えば、多糖類、ポリ／オリゴヌクレオチド、タンパク質、または脂質）またはその挙動を記述する自然言語であってもよい。例えば、言語ベースの検索データ４０１は、ヌクレオチド配列、タンパク質配列、または生物学的用語のグループの粗視化された検索用語であってもよい。例えば、標本が集積回路である場合には、言語ベースの検索データ４０１は、分岐回路（例えば、メモリセル、コンバータセル、ＥＳＤ保護回路）、回路素子（例えば、トランジスタ、コンデンサ、またはコイル）、または構造素子（例えば、ゲート、ビア、パッド、またはスペーサ）の記述、定義、または表現であってもよい。

【００４８】

複数の第２の高次元表現４０５は、データベースから取得可能であるか、または視覚認

識機械学習アルゴリズムによって生成可能である。例えば、システム４００は、１つまたは複数のプロセッサ１１０によって実行される視覚認識機械学習アルゴリズムによって、複数の画像ベースの入力データセットの複数の第２の高次元表現４０５を生成するように構成可能である。

【００４９】

顕微鏡は、標本の複数の画像を撮影するように構成可能である。複数の画像ベースの入力データセットは、標本の複数の画像を表すことができる。複数の画像ベースの入力データセットは、顕微鏡によって標本から撮影された画像の画像データであってもよい。例えば、所望の倍率で単一の画像によって撮影するには大き過ぎる標本全体または標本の関心領域をカバーするために、標本から種々異なる位置において複数の画像を撮影することができる。複数の画像のうちのそれぞれの画像の画像データは、複数の画像ベースの入力データセットのうちの１つの画像ベースの入力データセットを表すことができる。システム４００は、画像が撮影された位置を格納するように構成可能である。これらの位置を、対応する画像と共に、または対応する第２の高次元表現４０５と共に格納することができる。システム４００は、顕微鏡を含んでいてもよいし、または顕微鏡が、システム４００に接続されてもよいし、もしくはシステム４００を含んでいてもよい。

【００５０】

システム４００は、複数の第２の高次元表現のうち、選択基準を満たす第２の高次元表現（例えば、第１の高次元表現に最も近い第２の高次元表現）を選択することができる。第１の高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と比較することにより、第１の高次元表現に最も近い１つまたは複数の第２の高次元表現を提供することができる。システム４００は、比較に基づいて、複数の第２の高次元表現のうち、第１の高次元表現に最も近い１つまたは複数の第２の高次元表現を選択するように構成可能である。

【００５１】

システム４００は、選択された第２の高次元表現に基づいて、顕微鏡目標位置を決定するように構成可能である。顕微鏡目標位置は、選択された第２の高次元表現に対応する画像が撮影された位置であってもよい。例えば、顕微鏡目標位置は、選択された第２の高次元表現と共に、または選択された第２の高次元表現に対応する画像と共に格納された位置であってもよい。顕微鏡目標位置は、選択された第２の高次元表現に対応する、画像ベースの入力データによって表現されている画像が撮影された位置であってもよい。

【００５２】

システム４００は、決定された顕微鏡目標位置に基づいて、顕微鏡の動作を制御するための制御信号を供給するように構成可能である。制御信号４１１は、動き、倍率、光源選択、フィルタ選択および／または他の顕微鏡機能を制御するために顕微鏡に供給される電気信号であってもよい。例えば、制御信号４１１は、顕微鏡目標位置へと駆動するように顕微鏡をトリガするように構成可能である。例えば、顕微鏡の光学系および／または標本台は、制御信号４１１に応答して顕微鏡目標位置へと移動可能である。このようにして、検索の結果である位置において、標本からさらなる画像を撮影することができる。例えば、より高倍率、異なる光源および／または異なるフィルタを用いた画像を、関心領域から撮影することができる。例えば、言語ベースの検索データ４０１は、大きな生物学的標本における細胞核の検索を表すことができ、システム４００は、顕微鏡を細胞核の位置へと駆動するための制御信号４１１を供給することができる。複数の細胞核が検出され得る場合には、システム４００は、顕微鏡が、これらの位置でより多くの画像を撮影するために順次に異なる位置へと駆動されるように、制御信号４１１を供給するように構成可能である。

【００５３】

システム４００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図３および図５～図７）の１つまたは複数の例に関連して言及される。システム４００は、提案されているコンセプトおよび／または上記また

は下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００５４】

図１～図４のうちの１つに関連して説明したシステムは、コンピュータデバイス内に配置されている１つまたは複数のプロセッサおよび１つまたは複数のストレージデバイスを備えるコンピュータデバイス（例えば、パーソナルコンピュータ、ラップトップ、タブレットコンピュータ、または携帯電話）であってもよいし、またはこれを含んでいてもよい。あるいは、システムは、分散コンピューティングシステム（例えば、ローカルクライアントおよび１つまたは複数のリモートサーバファームおよび／またはデータセンター等の様々な場所に分散されている１つまたは複数のプロセッサおよび１つまたは複数のストレージデバイスを備えるクラウドコンピューティングシステム）であってもよい。システムは、システムの種々の構成要素を結合するためのシステムバスを含むデータ処理システムを含むことができる。システムバスは、システムの種々の構成要素間の通信リンクを提供することができ、シングルバスとして、複数のバスの組み合わせとして、または任意の他の適切な手法で実装可能である。システムバスには、電子アセンブリを結合することができる。電子アセンブリは、任意の回路または回路の組み合わせを含んでいてもよい。１つの実施形態では、電子アセンブリは、任意の種類のものとすることができる、プロセッサを含んでいる。本明細書で使用されるように、プロセッサは、例えば、顕微鏡または顕微鏡部品（例えば、カメラ）のマイクロプロセッサ、マイクロコントローラ、複合命令セットコンピューティング（ＣＩＳＣ）マイクロプロセッサ、縮小命令セットコンピューティング（ＲＩＳＣ）マイクロプロセッサ、超長命令語（ＶＬＩＷ）マイクロプロセッサ、グラフィックプロセッサ、デジタル信号プロセッサ（ＤＳＰ）、マルチコアプロセッサ、フィールド・プログラマブル・ゲート・アレイ（ＦＰＧＡ）、または任意の他の種類のプロセッサまたは処理回路等のあらゆる種類の計算回路を意図していてもよいが、これらに限定されない。電子アセンブリに含まれ得る他の種類の回路は、カスタム回路、特定用途向け集積回路（ＡＳｌＣ）等であってもよく、例えばこれは、携帯電話、タブレットコンピュータ、ラップトップコンピュータ、双方向無線機および類似の電子システム等の無線装置において使用される１つまたは複数の回路（通信回路等）である。システムは、ランダムアクセスメモリ（ＲＡＭ）の形態のメインメモリ等の特定の用途に適した１つまたは複数の記憶素子を含み得る１つまたは複数のストレージデバイス、１つまたは複数のハードドライブおよび／またはコンパクトディスク（ＣＤ）、フラッシュメモリカード、デジタルビデオディスク（ＤＶＤ）等のリムーバブルメディアを扱う１つまたは複数のドライブ等を含んでいる。システムは、ディスプレイ装置、１つまたは複数のスピーカおよびキーボードおよび／またはマウス、トラックボール、タッチスクリーン、音声認識装置を含み得るコントローラ、またはシステムのユーザがシステムに情報を入力することおよびシステムから情報を受け取ることを可能にする任意の他の装置も含んでいてもよい。

【００５５】

さらに、システムは、コンピュータデバイスまたは分散コンピューティングシステムに接続された顕微鏡を含むことができる。顕微鏡は、１つまたは複数の標本から画像を撮影することによって生物学関連の画像ベースの入力データセットを生成するように構成可能である。

【００５６】

顕微鏡は、光学顕微鏡（例えば、超解像顕微鏡またはナノ顕微鏡のような、回折限界顕微鏡またはサブ回折限界顕微鏡）であってもよい。顕微鏡は、スタンドアロン顕微鏡であってもよいし、または付属の構成要素（例えば、共焦点スキャナ、追加的なカメラ、レーザ、気候室、自動装填機構、液体搬送システム、付属の光学構成要素、例えば、追加的な多光子光路、光ピンセットなど）を有する顕微鏡システムであってもよい。例えば、生物学的配列（例えば、タンパク質、核酸、脂質）または他の標本に関連するオブジェクトの画像を撮影することができる限り、他の画像ソースを使用することも可能である。例えば、上記または下記の実施形態による顕微鏡により、深部を発見する顕微鏡法が可能となり

得る。

【００５７】

システムのさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図７）の１つまたは複数の例に関連して言及される。システムは、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００５８】

いくつかの実施形態は、図１～図４のうちの１つまたは複数に関連して説明したようなシステムを含む顕微鏡に関する。択一的に、顕微鏡は、図１～図４のうちの１つまたは複数に関連して説明したようなシステムの一部であってもよいし、またはこれに接続されていてもよい。図５は、１つの実施形態による、データを処理するためのシステム５００の概略図を示す。１つまたは複数の標本（例えば、生物学的標本または集積回路）の画像を撮影するように構成された顕微鏡５１０は、データを処理するように構成されたコンピュータデバイス５２０（例えば、パーソナルコンピュータ、ラップトップ、タブレットコンピュータ、または携帯電話）に接続されている。顕微鏡５１０およびコンピュータデバイス５２０は、図１～図４のうちの１つまたは複数に関連して説明したように実装可能である。

【００５９】

図６は、１つの実施形態による、生物学関連の言語ベースの検索データを処理するための方法のフローチャートである。方法６００は、生物学関連の言語ベースの検索データを受信すること６１０と、トレーニングされた言語認識機械学習アルゴリズムによって、生物学関連の言語ベースの検索データの第１の高次元表現を生成すること６２０と、を含む。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、方法６００は、複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットの複数の第２の高次元表現を取得すること６３０を含む。さらに、方法は、第１の高次元表現を、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と比較すること６４０を含む。

【００６０】

言語認識機械学習アルゴリズムを使用することによって、生物学的テキスト検索用語を高次元表現にマッピングすることができる。高次元表現が（ワンホットエンコーディングされた表現とは対照的に）種々の異なる値を有するエントリを有することを可能にすることによって、意味論的に類似した生物学的検索用語を、類似した高次元表現にマッピングすることができる。複数の生物学関連の画像ベースの入力データセットまたは複数の生物学関連の言語ベースの入力データセットの高次元表現を取得することにより、高次元表現が、検索用語の高次元表現に等しいことまたは類似することを検出することができる。このようにして、検索用語に対応する画像またはテキストを検出することが可能となり得る。このようにして、トレーニングされた言語認識機械学習アルゴリズムは、言語ベースの検索入力に基づいて、複数の生物学的画像（例えば、生物学的画像のデータベース）の中から生物学関連の画像を検索すること、または複数の生物学関連のテキスト（例えば、科学論文のコレクションまたはライブラリ）の中から生物学関連のテキストを検索することを可能にすることができる。既存のデータベース内での検索および／または実行中の実験によって生成された画像内での検索を、たとえこれらの画像が事前にラベル付けまたはタグ付けされていなかったとしても可能にすることができる。

【００６１】

方法６００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図５）の１つまたは複数の例に関連して言及される。方法６００は、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。

【００６２】
図７は、１つの実施形態による、顕微鏡を制御するための方法のフローチャートである。方法７００は、言語ベースの検索データを受信すること７１０と、トレーニングされた言語認識機械学習アルゴリズムによって、言語ベースの検索データの第１の高次元表現を生成すること７２０と、を含む。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、方法７００は、複数の画像ベースの入力データセットの複数の第２の高次元表現を取得すること７３０と、第１の高次元表現と、複数の第２の高次元表現のうちのそれぞれの第２の高次元表現と、の比較に基づいて、複数の第２の高次元表現から１つの第２の高次元表現を選択すること７４０と、を含む。さらに、方法７００は、選択された第２の高次元表現に基づいて、顕微鏡の動作を制御すること７５０を含む。
【００６３】
言語認識機械学習アルゴリズムを使用することによって、テキスト検索用語を高次元表現にマッピングすることができる。高次元表現が（ワンホットエンコーディングされた表現とは対照的に）種々の異なる値を有するエントリを有することを可能にすることによって、意味論的に類似した検索用語を、類似した高次元表現にマッピングすることができる。複数の画像ベースの入力データセットの高次元表現を取得することによって、高次元表現が、検索用語の高次元表現に等しいことまたは類似することを検出することができる。このようにして、検索用語に対応する画像を検出することが可能となり得る。この情報を用いて顕微鏡を、画像が撮影されたそれぞれの位置へと駆動して、その関心位置のさらなる画像を（例えば、より高倍率で、異なる光で、または異なるフィルタで）撮影することを可能にすることができる。このようにして、検索用語に対応する位置を発見するためにまず始めに試料（例えば、生物学的試料または集積回路）を低倍率で撮像して、その後、その関心位置をより詳細に分析することができる。
【００６４】
方法７００のさらなる詳細および態様は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図５）の１つまたは複数の例に関連して言及される。方法７００は、提案されているコンセプトおよび／または上記または下記の１つまたは複数の例の１つまたは複数の態様に対応する１つまたは複数の追加的なオプションの特徴を含むことができる。
【００６５】
以下では、（例えば図１～図７のうちの１つまたは複数に関連して）上記の実施形態のうちの１つまたは複数に関する用途および／または実装の詳細のいくつかの例について説明する。
【００６６】
１つの態様によれば、データベース内での、または実行中の顕微鏡実験でのテキスト・ツー・イメージ検索機能が提案されている。テキスト・ツー・イメージ検索の種類は、第１段階のテキストモデルによって作成されたクエリテキストの意味論的埋め込みに基づくことができる。第２段階の画像モデルは、これらの意味論的埋め込みを画像に関連付けることができ、これによって画像ドメインをテキストドメインに結び付ける。ヒットの関連性は、意味論的埋め込み空間における距離メトリックに従ってスコア化可能である。これにより、完全一致の検索だけでなく、関連する意味論を有する類似の画像の検索も可能にすることができる。生物学の文脈において、関連する意味論は、類似の生物学的機能を意味していてもよい。
【００６７】
一般的に生物学および顕微鏡法は、特に膨大な量のデータを生成することができるが、これらのデータには注釈が乏しいか、または全く注釈が付されないことが多い。例えば、振り返ってみて、どの注釈が有用であったか、または実験の時点でどの新しい生物学的発見が知られていないかが明らかになるだけであり得る。画像データに重きを置いてもよいが、提案されているコンセプトは、必ずしも画像データに限定されているわけではない。

画像は、２Ｄピクセルマップの範囲を超えてもよいが、むしろ、例えば、使用されている蛍光色素の物理的特性または撮像システムの特性に関連する三空間次元、時間次元およびさらなる次元を有する多次元の画像テンソルを包含してもよい。１つの態様によれば、データベース内に格納された大量の画像データ、または顕微鏡での実行中の実験の一部としての大量の画像データを意味論的に検索することを可能することによって、このようなデータにアクセスすることが可能となり得る。実験は、１回限りの実験であってもよいし、またはスクリーニングキャンペーンのような長期的な実験の一部であってもよい。

【００６８】

実行中の実験の文脈において、提案されているコンセプトは、単一細胞内、細胞小器官内、または組織内で発現されるタンパク質のような標本だけでなく、臓器または発達状態のようなより一般的な構造の一部でもある生物学的構造の検索を自動化するために役立つことができる。このようにして、標本中の関連する部分を検出するための時間のかかるステップを自動化することが可能となり得る。

【００６９】

他のコンセプトの場合には、テキストを介した画像のクエリは、人間の専門家によって作成された注釈に依拠し得る。それぞれの注釈は、検索を絞り込むための次元として機能することができる。しかしながら、このような注釈は、利用できないことが多いか、または作成に時間がかかることが多い。１つの態様によれば、意味論的埋め込み空間における距離メトリックを使用することができる。このアプローチは、人間の注釈付けの厳密な必要性を排除することができるのみならず、画像間の機能的な（例えば、生物学的な）類似性についての尺度を提供することもできる。

【００７０】

実行中の顕微鏡実験の文脈において、提案されているコンセプトは、顕微鏡標本中の関連する構造を検出するための時間のかかるステップを省略することができる。ユーザは、標本中を手動で検索する代わりに、検索クエリを入力することができ、提案されている顕微鏡は、標本中の関連するオブジェクトを検出することができる。このアプローチは、手動での検索によってしばしばもたらされる個々のバイアスを除去することもできる。

【００７１】

さらに、意味論的埋め込みを使用することによって、複数のクエリ要素を論理的または算術的に組み合わせることが可能となり得る。例えば、（細胞質に斑状の外観を有する）“clathrin（クラスリン）”および“nucleus（核）”を検索すると、斑状の外観を有する（が、必ずしもクラスリンの相互作用体として知られているわけではない）核内の多数のタンパク質が示される可能性がある。逆に、ミトコンドリアの酸化リン酸化経路の一部である２つの既知の相互作用体である“ATP Synthase（ＡＴＰ合成酵素）”および“Cytochrome c oxidase（チトクロムｃオキシダーゼ）”をより絞り込んで検索すると、より多くの、酸化リン酸化に関連するタンパク質と、ミトコンドリアに位置するタンパク質とがもたらされる可能性がある。したがって、既知の生物学的な相互作用体の検索と、可能性のある（不可能といってもいいほどの、すなわち、ありそうもない）新規の相互作用体の検索との両方を発見することができる。

【００７２】

１つの態様によれば、テキスト・ツー・イメージ検索が可能となり得る。顕微鏡画像の検索および探索のために現在使用されている画像データベースは、画像を検索可能にするための古典的なタグ付けおよび注釈付けを使用することができる。提案されているコンセプトは、良好な注釈が利用できないところでも画像を検索可能にすることができる。例えば、明示的に注釈付けされていない生物学的機能を検索することが可能となり得る。そうでなければこの注釈付けは、専門家による時間および作業を必要とするであろうが、本明細書では、これを省略することができる。

【００７３】

例えば、検索用語に意味論的に近いが同一ではない近似的なヒットを得ることができる。ゼロショット学習（例えば、トレーニングセットに含まれていなかったクラスを識別す

ること）としても知られているこの特性は、他の技術を使用していれば利用できないであろう。

【００７４】

さらに、クエリ用語に対する論理的な演算は、結果的に、意味論的な算術をもたらすことができる。例えば、複数の検索用語を組み合わせて、画像におけるそれらの検索用語の意味論的な組み合わせを得ることができ、例えば、“clathrin（クラスリン）”および“nucleus（核）”を検索すると、核スペックルが探索されるであろう。この特性は、意味論が計算可能なエンティティとして表現されている場合にのみ、利用可能であり得る。１つの態様によれば、意味論は、潜在ベクトルとして表現可能である。

【００７５】

提案されているコンセプトを、（顕微鏡のような）撮像システムにおいて実現することができる。このようにして、同じ（または類似の）生物学的機能を有する関連するオブジェクトを、例えば、生物学的配列、自然言語、または粗視化された検索用語を使用したテキスト検索に基づいて、標本中で識別することができる。

【００７６】

１つの態様によるテキスト・ツー・イメージ検索の種類は、第１段階のテキストモデルによるクエリテキストからの意味論的埋め込みの作成と、意味論的埋め込みを画像に関連付け、これによって画像ドメインをテキストドメインに結び付ける第２段階の画像モデルとに基づくことができる。ヒットの関連性は、意味論的埋め込み空間における距離メトリックに従ってスコア化可能である。第１段階および第２段階のモデルは、以下に説明するようにトレーニング可能であるが、これとは異なる手法でもトレーニング可能である。

【００７７】

テキスト・ツー・イメージ検索は、適切な検索用語を用いて注釈付けされていない画像を、テキストクエリによって検索可能にすることができる。これにより、専門家による手作業を省略し、バイアスを除去することができる。用途分野は、既存の画像データベースの検索であってもよいが、顕微鏡またはバイオ撮像装置における実行中の実験の検索であってもよい。生物学は、特に、一方では、生物学的配列、記述テキストおよび画像の形態での大量のデータを生成する傾向にあり得るが、多くの場合、データと共に標準化されていないか、またはいずれの注釈も有していない（例えば、後者は、主に画像に当てはまる）。写真と比較される生物学的画像の特徴空間が制限される場合があり、これにより、従来のアプローチではこれらを見分けることが困難になる。１つの態様によれば、画像認識のための深層学習ベースの画像モデル（例えば、ＣＮＮまたはＣａｐｓＮｅｔｓ）の能力を、生物学的配列または記述テキストに含まれている機能情報と組み合わせることができる。したがって、配列およびテキストは、生物学的機能のプロキシとして機能することができる。

【００７８】

さらに、検索用語に意味論的に関連するが検索用語と同一ではない近似的なヒットを得ることが可能となり得る。このことは、データベースまたは標本から、ユーザが存在を知らなかった関連する画像を探索するのに役立つことができる。データ量の増大および以前の実験の利用可能性の増大に鑑みて、このことは、既存の画像リポジトリを、実験を実行した研究者以外にとっても有用なデータ資産へと変えるために役立つことができる。別の用途は、組み合わされた検索用語の検索であってもよい。ユーザレベルでの組み合わせは、複数の検索用語の連結または入力のように見え得る。内部的には、それらの用語を、算術演算を受けることができるベクトルである意味論的埋め込みに変換することができる。例えば、データリポジトリ内での、または実行中の実験中での新たな関係性を発見することができる。専門家による注釈付けと比較すると、専門家は、自分が知っているか予想しているものに注釈付けすることしかできないが、新たな発見は、新たな研究課題につながることができる。したがって、提案されているコンセプトによるテキスト・ツー・イメージ検索は、既存のデータに新しい質問をすることを可能にすることができ、これによって、既存のデータの価値を高めることができる。提案されているテキスト・ツー・イメージ

検索の用途は、関連データの検出および実験記録時間の短縮を助ける、基本的な生物学的研究であってもよく、かつ／または例えば、創薬におけるヒット検証および毒物学的検査であってもよい。

【００７９】

トレーニングされた言語認識機械学習アルゴリズムおよび／またはトレーニングされた視覚認識機械学習アルゴリズムは、以下に説明するトレーニングによって取得可能である。生物学関連のデータを処理するための機械学習アルゴリズムをトレーニングするためのシステムは、１つまたは複数のプロセッサと、１つまたは複数のストレージデバイスと、を含むことができる。当該システムは、生物学関連の言語ベースの入力トレーニングデータを受信するように構成可能である。さらに、当該システムは、１つまたは複数のプロセッサによって実行される言語認識機械学習アルゴリズムによって、生物学関連の言語ベースの入力トレーニングデータの第１の高次元表現を生成するように構成可能である。第１の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、当該システムは、１つまたは複数のプロセッサによって実行される言語認識機械学習アルゴリズムによって、第１の高次元表現に基づいて、生物学関連の言語ベースの出力トレーニングデータを生成するように構成可能である。さらに、当該システムは、生物学関連の言語ベースの入力トレーニングデータと、生物学関連の言語ベースの出力トレーニングデータと、の比較に基づいて、言語認識機械学習アルゴリズムを調整するように構成可能である。さらに、当該システムは、生物学関連の言語ベースの入力トレーニングデータに関連付けられた生物学関連の画像ベースの入力トレーニングデータを受信するように構成可能である。さらに、当該システムは、１つまたは複数のプロセッサによって実行される視覚認識機械学習アルゴリズムによって、生物学関連の画像ベースの入力トレーニングデータの第２の高次元表現を生成するように構成可能である。第２の高次元表現は、それぞれ異なる値を有する少なくとも３つのエントリを含む。さらに、当該システムは、第１の高次元表現と第２の高次元表現との比較に基づいて、視覚認識機械学習アルゴリズムを調整するように構成可能である。

【００８０】

生物学関連の言語ベースの入力トレーニングデータは、生物学的構造、生物学的機能、生物学的挙動、または生物学的活動に関連するテキスト入力であってもよい。例えば、生物学関連の言語ベースの入力トレーニングデータは、ヌクレオチド配列、タンパク質配列、生物学的分子もしくは生物学的構造の記述、生物学的分子もしくは生物学的構造の挙動の記述および／または生物学的機能もしくは生物学的活動の記述であってもよい。生物学関連の言語ベースの入力トレーニングデータは、トレーニンググループのうちの第１の生物学関連の言語ベースの入力トレーニングデータセット（例えば、入力文字のシーケンス、例えば、ヌクレオチド配列またはタンパク質配列）であってもよい。トレーニンググループは、複数の生物学関連の言語ベースの入力トレーニングデータセットを含むことができる。

【００８１】

生物学関連の言語ベースの出力トレーニングデータは、オプションとして次の要素の予測を含む、生物学関連の言語ベースの入力トレーニングデータと同じ種類のものであってもよい。例えば、生物学関連の言語ベースの入力トレーニングデータは、生物学的配列（例えば、ヌクレオチド配列またはタンパク質配列）であってもよく、生物学関連の言語ベースの出力トレーニングデータも、生物学的配列（例えば、ヌクレオチド配列またはタンパク質配列）であってもよい。言語認識機械学習アルゴリズムは、生物学関連の言語ベースの出力トレーニングデータが、オプションとして生物学的配列の次の要素の予測を含む、生物学関連の言語ベースの入力トレーニングデータと等しくなるようにトレーニング可能である。別の例では、生物学関連の言語ベースの入力トレーニングデータは、粗視化された検索用語の生物学的クラスであってもよく、生物学関連の言語ベースの出力トレーニングデータも、粗視化された検索用語の生物学的クラスであってもよい。

【００８２】

生物学関連の画像ベースの入力トレーニングデータは、ヌクレオチドもしくはヌクレオチド配列を含む生物学的構造、タンパク質もしくはタンパク質配列を含む生物学的構造、生物学的分子、生物学的組織、特定の挙動を有する生物学的構造および／または特定の生物学的機能もしくは特定の生物学的活動を有する生物学的構造の画像の画像トレーニングデータ（例えば、トレーニング画像のピクセルデータ）であってもよい。生物学関連の画像ベースの入力トレーニングデータは、トレーニンググループのうちの第１の生物学関連の画像ベースの入力トレーニングデータセットであってもよい。トレーニンググループは、複数の生物学関連の画像ベースの入力トレーニングデータセットを含むことができる。

【００８３】

生物学関連の言語ベースの入力トレーニングデータは、トレーニンググループのうちの１つの生物学関連の言語ベースの入力トレーニングデータセット（例えば、入力文字のシーケンス、例えば、ヌクレオチド配列またはタンパク質配列）であってもよい。トレーニンググループは、複数の生物学関連の言語ベースの入力トレーニングデータセットを含むことができる。システムは、トレーニンググループのうちの複数の生物学関連の言語ベースの入力トレーニングデータセットの各々のために、第１の高次元表現を生成することを繰り返すことができる。さらに、システムは、それぞれの生成された第１の高次元表現ごとに、生物学関連の言語ベースの出力トレーニングデータを生成することができる。システムは、トレーニンググループのうちの複数の生物学関連の言語ベースの入力トレーニングデータセットの、生物学関連の言語ベースの入力トレーニングデータと、対応する生物学関連の言語ベースの出力トレーニングデータとのそれぞれの比較に基づいて、言語認識機械学習アルゴリズムを調整することができる。換言すれば、システムは、生物学関連の言語ベースの入力トレーニングデータセットのトレーニンググループのうちのそれぞれの生物学関連の言語ベースの入力トレーニングデータごとに、第１の高次元表現を生成することと、生物学関連の言語ベースの出力トレーニングデータを生成することと、言語認識機械学習アルゴリズムを調整することとを繰り返すように構成可能である。トレーニンググループは、トレーニング目標（例えば、しきい値を下回るように損失関数の出力を変化させること）を満たすことができるように十分な生物学関連の言語ベースの入力トレーニングデータセットを含むことができる。

【００８４】

言語認識機械学習アルゴリズムのトレーニング中に生成された複数の全ての第１の高次元表現は、潜在空間または意味論的空間と称され得る。

【００８５】

システムは、トレーニンググループのうちの複数の生物学関連の画像ベースの入力トレーニングデータセットの各々のために、第２の高次元表現を生成することを繰り返すことができる。さらに、システムは、第１の高次元表現と、対応する第２の高次元表現とのそれぞれの比較に基づいて、視覚認識機械学習アルゴリズムを調整することができる。換言すれば、システムは、生物学関連の画像ベースの入力トレーニングデータセットのトレーニンググループのうちのそれぞれの生物学関連の画像ベースの入力トレーニングデータごとに、第２の高次元表現を生成することと、視覚認識機械学習アルゴリズムを調整することとを繰り返すことができる。トレーニンググループは、トレーニング目標（例えば、しきい値を下回るように損失関数の出力を変化させること）を満たすことができるように十分な生物学関連の画像ベースの入力トレーニングデータセットを含むことができる。

【００８６】

例えば、システム１００は、言語認識機械学習アルゴリズムと、視覚認識機械学習アルゴリズム（例えば、視覚意味論的モデルとも称される）と、の組み合わせを使用する。言語認識機械学習アルゴリズムおよび／または視覚認識機械学習アルゴリズムは、深層学習アルゴリズムおよび／または人工知能アルゴリズムであってもよい。

【００８７】

言語認識機械学習アルゴリズムをトレーニングするために交差エントロピー損失関数を使用することにより、トレーニングを高速に収束させることができ、かつ／または生物学

関連のデータのために十分にトレーニングされたアルゴリズムを提供することができるが、他の損失関数を使用することも可能である。

【００８８】

視覚認識機械学習アルゴリズムは、対応する入力トレーニングデータの、言語認識機械学習アルゴリズムによって生成された高次元表現と、視覚認識機械学習アルゴリズムによって生成された高次元表現と、の比較に基づいて、視覚認識機械学習アルゴリズムのパラメータを調整することによってトレーニング可能である。例えば、この比較に基づいて、視覚認識ニューラルネットワークのネットワーク重みを調整することができる。視覚認識機械学習アルゴリズムのパラメータ（例えば、ネットワーク重み）の調整は、損失関数を考慮して実施可能である。例えば、視覚認識機械学習アルゴリズムの調整のための、第１の高次元表現と第２の高次元表現との比較は、コサイン類似性損失関数に基づくことができる。視覚認識機械学習アルゴリズムをトレーニングするためにコサイン類似性損失関数を使用することにより、トレーニングを高速に収束させることができ、かつ／または生物学関連のデータのために十分にトレーニングされたアルゴリズムを提供することができるが、他の損失関数を使用することも可能である。

【００８９】

例えば、視覚モデルは、意味論的埋め込み空間において（例えば、ベクトルとして）どのようにして画像を表現すべきかを学習することができる。したがって、予測Ａ（第２の高次元表現）と、グラウンドトゥルースＢ（第１の高次元表現）とを表現することができる、２つのベクトルの距離に対する尺度を使用することができる。例えば、１つの尺度は、

【数５】

$$\text{類似性} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

の形態で定義されるようなコサイン類似性であり、予測Ａとグラウンドトゥルースとのドット積を、（例えば、Ｌ２－ノルムまたはユークリッドノルムのように）各自のそれぞれの絶対値のドット積で除算したものである。

【００９０】

機械学習アルゴリズムをトレーニングするためのシステムの非トレーニング特有の態様に関するさらなる詳細は、提案されているコンセプトおよび／または上記または下記（例えば、図１～図７）の１つまたは複数の例に関連して言及される。

【００９１】

実施形態は、機械学習モデルまたは機械学習アルゴリズムの使用に基づいていてもよい。機械学習は、モデルおよび推論に依存する代わりに、コンピュータシステムが、明示的な命令を使用することなく、特定のタスクを実行するために使用し得るアルゴリズムおよび統計モデルを参照してもよい。例えば、機械学習では、ルールに基づくデータ変換の代わりに、過去のデータおよび／またはトレーニングデータの分析から推論されるデータ変換が使用されてもよい。例えば、画像コンテンツは、機械学習モデルを用いて、または機械学習アルゴリズムを用いて分析されてもよい。機械学習モデルが画像コンテンツを分析するために、機械学習モデルは、入力としてのトレーニング画像と出力としてのトレーニングコンテンツ情報を用いてトレーニングされてもよい。多数のトレーニング画像および／またはトレーニングシーケンス（例えば単語または文）および関連するトレーニングコンテンツ情報（例えばラベルまたは注釈）によって機械学習モデルをトレーニングすることによって、機械学習モデルは、画像コンテンツを認識することを「学習」するので、トレーニングデータに含まれていない画像コンテンツが機械学習モデルを用いて認識可能になる。同じ原理が、同じように他の種類のセンサデータに対して使用されてもよい：トレーニングセンサデータと所望の出力を用いて機械学習モデルをトレーニングすることによって、機械学習モデルは、センサデータと出力との間の変換を「学習し」、これは、機械

学習モデルに提供された非トレーニングセンサデータに基づいて出力を提供するために使用可能である。

【００９２】

機械学習モデルは、トレーニング入力データを用いてトレーニングされてもよい。上記の例は、「教師あり学習」と称されるトレーニング方法を使用する。教師あり学習では、機械学習モデルは、複数のトレーニングサンプルを用いてトレーニングされ、ここで各サンプルは複数の入力データ値と複数の所望の出力値を含んでいてもよく、すなわち各トレーニングサンプルは、所望の出力値と関連付けされている。トレーニングサンプルと所望の出力値の両方を指定することによって、機械学習モデルは、トレーニング中に、提供されたサンプルに類似する入力サンプルに基づいてどの出力値を提供するのかを「学習」する。教師あり学習の他に、半教師あり学習が使用されてもよい。半教師あり学習では、トレーニングサンプルの一部は、対応する所望の出力値を欠いている。教師あり学習は、教師あり学習アルゴリズム、例えば分類アルゴリズム、回帰アルゴリズムまたは類似度学習アルゴリズムに基づいていてもよい。出力が、値の限られたセットに制限される場合、すなわち入力が値の限られたセットのうちの１つに分類される場合、分類アルゴリズムが使用されてもよい。出力が（範囲内の）任意の数値を有していてもよい場合、回帰アルゴリズムが使用されてもよい。類似度学習アルゴリズムは、分類アルゴリズムと回帰アルゴリズムの両方に類似していてもよいが、２つのオブジェクトがどの程度類似しているかまたは関係しているかを測定する類似度関数を用いた例からの学習に基づいている。教師あり学習または半教師あり学習の他に、機械学習モデルをトレーニングするために教師なし学習が使用されてもよい。教師なし学習では、入力データ（だけ）が供給される可能性があり、教師なし学習アルゴリズムは、例えば、入力データをグループ化またはクラスタリングすること、データに共通性を見出すことによって入力データにおいて構造を見出すために使用されてもよい。クラスタリングは、複数の入力値を含んでいる入力データを複数のサブセット（クラスター）に割り当てることであるので、同じクラスター内の入力値は１つまたは複数の（事前に定められた）類似度判断基準に従って類似しているが、別のクラスターに含まれている入力値と類似していない。

【００９３】

強化学習は機械学習アルゴリズムの第３のグループである。換言すれば、強化学習は機械学習モデルをトレーニングするために使用されてもよい。強化学習では、１つまたは複数のソフトウェアアクター（「ソフトウェアエージェント」と称される）が、周囲において行動を取るようにトレーニングされる。取られた行動に基づいて、報酬が計算される。強化学習は、（報酬の増加によって明らかにされるように）累積報酬が増加し、与えられたタスクでより良くなるソフトウェアエージェントが得られるように行動を選択するように、１つまたは複数のソフトウェアエージェントをトレーニングすることに基づいている。

【００９４】

さらに、いくつかの技術が、機械学習アルゴリズムの一部に適用されてもよい。例えば、特徴表現学習が使用されてもよい。換言すれば、機械学習モデルは、少なくとも部分的に特徴表現学習を用いてトレーニングされてもよい、かつ／または機械学習アルゴリズムは、特徴表現学習構成要素を含んでいてもよい。表現学習アルゴリズムと称され得る特徴表現学習アルゴリズムは、自身の入力に情報を保存するだけでなく、多くの場合、分類または予測を実行する前の前処理ステップとして、有用にするように情報の変換も行ってもよい。特徴表現学習は、例えば、主成分分析またはクラスター分析に基づいていてもよい。

【００９５】

いくつかの例では、異常検知（すなわち、外れ値検知）が使用されてもよく、これは、入力またはトレーニングデータの大部分と著しく異なることによって疑念を引き起こしている入力値の識別を提供することを目的としている。換言すれば、機械学習モデルは、少なくとも部分的に異常検知を用いてトレーニングされてもよく、かつ／または機械学習アルゴリズムは、異常検知構成要素を含んでいてもよい。

【００９６】

いくつかの例では、機械学習アルゴリズムは、予測モデルとして決定木を使用してもよい。換言すれば、機械学習モデルは、決定木に基づいていてもよい。決定木において、項目（例えば、入力値のセット）に関する観察は、決定木のブランチによって表されてもよく、この項目に対応する出力値は、決定木のリーフによって表されてもよい。決定木は、出力値として離散値と連続値の両方をサポートしてもよい。離散値が使用される場合、決定木は、分類木として表されてもよく、連続値が使用される場合、決定木は、回帰木として表されてもよい。

【００９７】

相関ルールは、機械学習アルゴリズムにおいて使用され得る別の技術である。換言すれば、機械学習モデルは、１つまたは複数の相関ルールに基づいていてもよい。相関ルールは、大量のデータにおける変数間の関係を識別することによって作成される。機械学習アルゴリズムは、データから導出された知識を表す１つまたは複数の相関的なルールを識別してもよい、かつ／または利用してもよい。これらのルールは、例えば、知識を格納する、操作するまたは適用するために使用されてもよい。

【００９８】

機械学習アルゴリズムは通常、機械学習モデルに基づいている。換言すれば、用語「機械学習アルゴリズム」は、機械学習モデルを作成する、トレーニングするまたは使用するために使用され得る命令のセットを表していてもよい。用語「機械学習モデル」は、例えば、機械学習アルゴリズムによって実行されるトレーニングに基づいて学習した知識を表すデータ構造および／またはルールのセットを表していてもよい。実施形態では、機械学習アルゴリズムの用法は、基礎となる１つの機械学習モデル（または基礎となる複数の機械学習モデル）の用法を意味していてもよい。機械学習モデルの用法は、機械学習モデルおよび／または機械学習モデルであるデータ構造／ルールのセットが機械学習アルゴリズムによってトレーニングされることを意味していてもよい。

【００９９】

例えば、機械学習モデルは、人工ニューラルネットワーク（ＡＮＮ）であってもよい。ＡＮＮは、網膜または脳において見出されるような、生物学的ニューラルネットワークによって影響を与えられるシステムである。ＡＮＮは、相互接続された複数のノードと、ノード間の、複数の接合部分、いわゆるエッジを含んでいる。通常、３種類のノードが存在しており、すなわち入力値を受け取る入力ノード、他のノードに接続されている（だけの）隠れノードおよび出力値を提供する出力ノードが存在している。各ノードは、人工ニューロンを表していてもよい。各エッジは、１つのノードから別のノードに、情報を伝達してもよい。ノードの出力は、その入力の和の（非線形）関数として定義されてもよい。ノードの入力は、入力を提供するエッジまたはノードの「重み」に基づく関数において使用されてもよい。ノードおよび／またはエッジの重みは、学習過程において調整されてもよい。換言すれば、人工ニューラルネットワークのトレーニングは、与えられた入力に対して所望の出力を得るために、人工ニューラルネットワークのノードおよび／またはエッジの重みを調整することを含んでいてもよい。

【０１００】

択一的に、機械学習モデルは、サポートベクターマシン、ランダムフォレストモデルまたは勾配ブースティングモデルであってもよい。サポートベクターマシン（すなわち、サポートベクターネットワーク）は、例えば、分類または回帰分析においてデータを分析するために使用され得る、関連する学習アルゴリズムを伴う、教師あり学習モデルである。サポートベクターマシンは、２つのカテゴリのいずれかに属する複数のトレーニング入力値を伴う入力を提供することによってトレーニングされてもよい。サポートベクターマシンは、２つのカテゴリのいずれかに新しい入力値を割り当てるようにトレーニングされてもよい。択一的に、機械学習モデルは、確率有向非巡回グラフィカルモデルであるベイジアンネットワークであってもよい。ベイジアンネットワークは、有向非巡回グラフを用いて、確率変数とその条件付き依存性のセットを表していてもよい。択一的に、機械学習モデルは、検索アルゴリズムと自然淘汰の過程を模倣した発見的方法である遺伝的アルゴリ

ズムに基づいていてもよい。

【０１０１】

本明細書で使用されるように、用語「および／または（かつ／または）」は、関連する記載項目のうちの１つまたは複数の項目のあらゆる全ての組み合わせを含んでおり、「／」として略記されることがある。

【０１０２】

いくつかの態様を装置の文脈において説明してきたが、これらの態様が、対応する方法の説明も表していることが明らかであり、ここではブロックまたは装置がステップまたはステップの特徴に対応している。同様に、ステップの文脈において説明された態様は、対応する装置の対応するブロックまたは項目または特徴の説明も表している。ステップの一部または全部は、例えば、プロセッサ、マイクロプロセッサ、プログラマブルコンピュータまたは電子回路等のハードウェア装置（またはハードウェア装置を使用すること）によって実行されてもよい。いくつかの実施形態では、極めて重要なステップのいずれか１つまたは複数が、そのような装置によって実行されてもよい。

【０１０３】

一定の実装要件に応じて、本発明の実施形態は、ハードウェアまたはソフトウェアで実装され得る。この実装は、非一過性の記録媒体によって実行可能であり、非一過性の記録媒体は、各方法を実施するために、プログラマブルコンピュータシステムと協働する（または協働することが可能である）、電子的に読取可能な制御信号が格納されている、デジタル記録媒体等であり、これは例えば、フロッピーディスク、ＤＶＤ、ブルーレイ、ＣＤ、ＲＯＭ、ＰＲＯＭおよびＥＰＲＯＭ、ＥＥＰＲＯＭまたはＦＬＡＳＨメモリである。したがって、デジタル記録媒体は、コンピュータ読取可能であってもよい。

【０１０４】

本発明のいくつかの実施形態は、本明細書に記載のいずれかの方法が実施されるように、プログラマブルコンピュータシステムと協働することができる、電子的に読取可能な制御信号を有するデータ担体を含んでいる。

【０１０５】

一般的に、本発明の実施形態は、プログラムコードを備えるコンピュータプログラム製品として実装可能であり、このプログラムコードは、コンピュータプログラム製品がコンピュータ上で実行されるときにいずれかの方法を実施するように作動する。このプログラムコードは、例えば、機械可読担体に格納されていてもよい。例えば、コンピュータプログラムは、非一過性の記録媒体に格納されていてもよい。いくつかの実施形態は、実行されるときに提案されているコンセプトまたは上述した１つもしくは複数の例による方法を実施するための機械可読命令を含む、非一過性の記録媒体に関する。

【０１０６】

別の実施形態は、機械可読担体に格納されている、本明細書に記載のいずれかの方法を実施するためのコンピュータプログラムを含んでいる。

【０１０７】

したがって、換言すれば、本発明の実施形態は、コンピュータプログラムがコンピュータ上で実行されるときに本明細書に記載のいずれかの方法を実施するためのプログラムコードを有するコンピュータプログラムである。

【０１０８】

したがって、本発明の別の実施形態は、プロセッサによって実行されるときに本明細書に記載のいずれかの方法を実施するために、格納されているコンピュータプログラムを含んでいる記録媒体（またはデータ担体またはコンピュータ読取可能な媒体）である。データ担体、デジタル記録媒体または被記録媒体は、典型的に、有形である、かつ／または非一過性である。本発明の別の実施形態は、プロセッサと記録媒体とを含んでいる、本明細書に記載されたような装置である。

【０１０９】

したがって、本発明の別の実施形態は、本明細書に記載のいずれかの方法を実施するた

めのコンピュータプログラムを表すデータストリームまたは信号シーケンスである。データストリームまたは信号シーケンスは例えば、データ通信接続、例えばインターネットを介して転送されるように構成されていてもよい。

【０１１０】

別の実施形態は、処理手段、例えば、本明細書に記載のいずれかの方法を実施するように構成または適合されているコンピュータまたはプログラマブルロジックデバイスを含んでいる。

【０１１１】

別の実施形態は、本明細書に記載のいずれかの方法を実施するために、インストールされたコンピュータプログラムを有しているコンピュータを含んでいる。

【０１１２】

本発明の別の実施形態は、本明細書に記載のいずれかの方法を実施するためのコンピュータプログラムを（例えば、電子的にまたは光学的に）受信機に転送するように構成されている装置またはシステムを含んでいる。受信機は、例えば、コンピュータ、モバイル機器、記憶装置等であってもよい。装置またはシステムは、例えば、コンピュータプログラムを受信機に転送するために、ファイルサーバを含んでいてもよい。

【０１１３】

いくつかの実施形態では、プログラマブルロジックデバイス（例えば、フィールド・プログラマブル・ゲート・アレイ）が、本明細書に記載された方法の機能の一部または全部を実行するために使用されてもよい。いくつかの実施形態では、フィールド・プログラマブル・ゲート・アレイは、本明細書に記載のいずれかの方法を実施するためにマイクロプロセッサと協働してもよい。一般的に、有利には、任意のハードウェア装置によって方法が実施される。

【符号の説明】

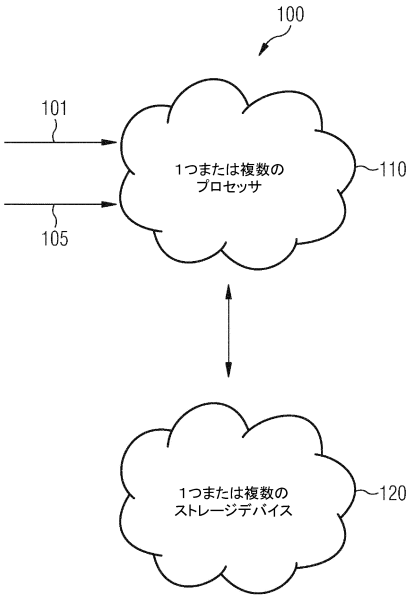
【０１１４】

１００　生物学関連のデータを処理するためのシステム
１０１　生物学的関連の言語ベースの検索データ
１０５　第２の高次元表現
１１０　１つまたは複数のプロセッサ
１２０　１つまたは複数のストレージデバイス
２００　生物学関連のデータを処理するためのシステム
２０１　クエリ、検索クエリ、生物学的関連の言語ベースの検索データ
２１０　テキストモデル、分類器
２２０　トレーニングされた視覚認識機械学習アルゴリズム、視覚モデル
２３０　トレーニングされた言語認識機械学習アルゴリズム、テキストモデル、言語モデル
２４０　データベース
２５０　埋め込み、複数の第２の高次元表現
２５５　データベース、中間ストレージ
２５７　バイパス
２６０　埋め込み、第１の高次元表現
２７０　埋め込み空間における比較
２８０　最も近い埋め込み
２９０　それぞれの画像
３００　生物学関連のデータを処理するためのシステム
３１５　スキップされた事前分類
３８１　最も近い埋め込みに対応する画像を返送する
３８３　画像ソースにデータを供給する
３８５　ユーザ
３８７　データベース

３８９　パブリックのデータベース
３９０　科学出版物、ソーシャルメディアのエントリ、またはブログ投稿
３９３　生物学的分子の画像
３９５　生物学的配列
４００　顕微鏡を制御するためのシステム
４０１　言語ベースの検索データ
４０５　第２の高次元表現
４１１　制御信号
５００　データを処理するためのシステム
５１０　顕微鏡
５２０　コンピュータデバイス
６００　生物学関連の言語ベースの検索データを処理するための方法
６１０　生物学関連の言語ベースの検索データを受信すること
６２０　第１の高次元表現を生成すること
６３０　複数の第２の高次元表現を取得すること
６４０　第１の高次元表現をそれぞれの第２の高次元表現と比較すること
７００　顕微鏡を制御するための方法
７１０　言語ベースの検索データを受信すること
７２０　第１の高次元表現を生成すること
７３０　複数の第２の高次元表現を取得すること
７４０　第２の高次元表現を選択すること
７５０　顕微鏡の動作を制御すること

【図面】

【図１】



【図２】

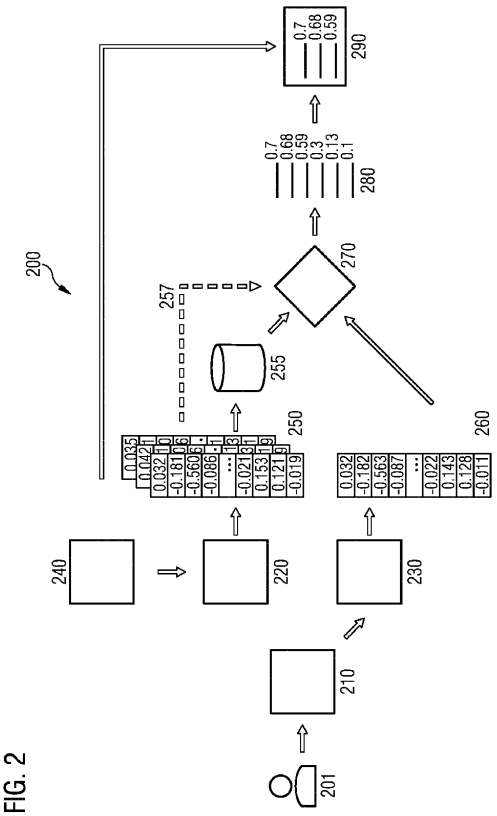
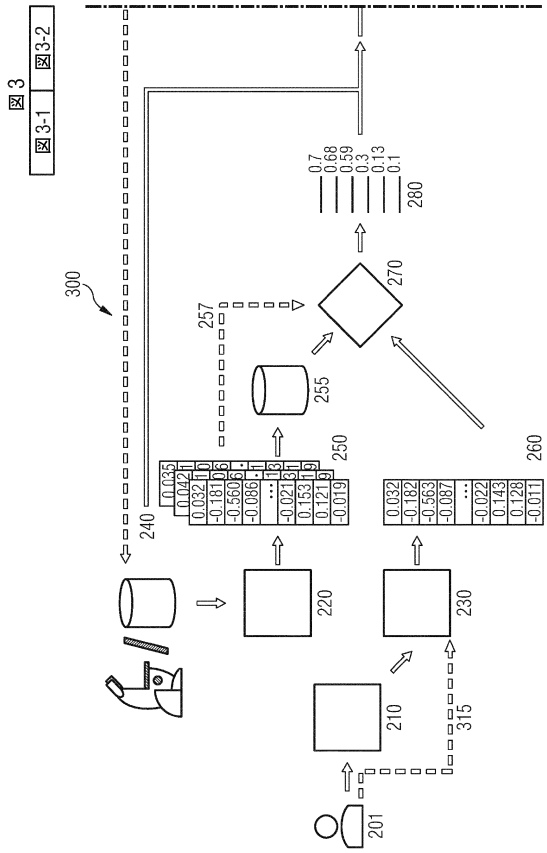
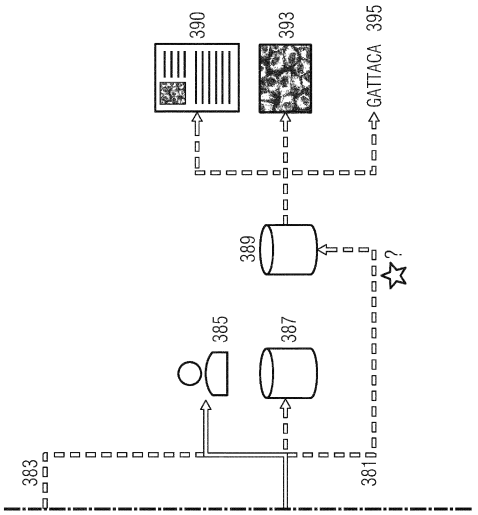


FIG. 2

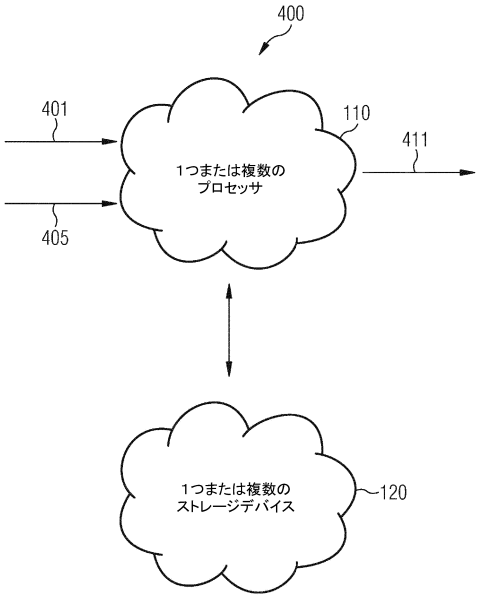
【図３－１】



【図３－２】

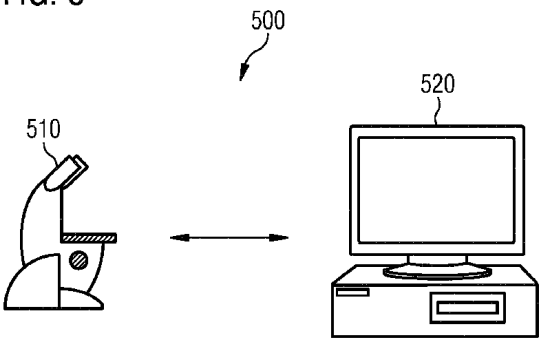


【図４】



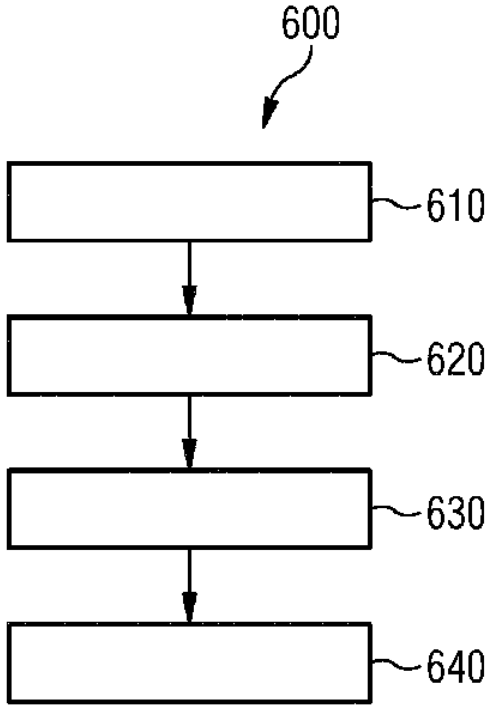
【図５】

FIG. 5



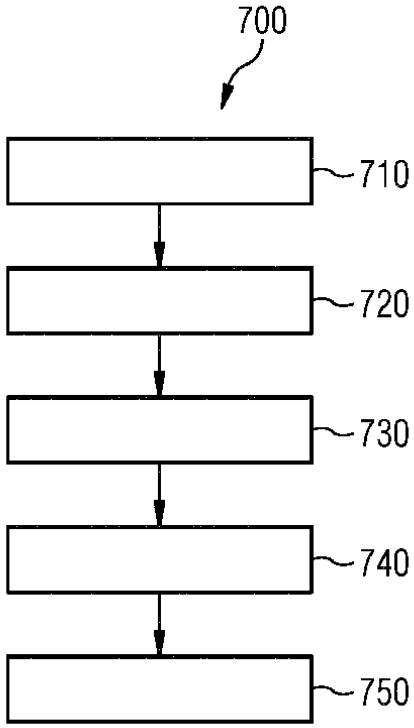
【図６】

FIG. 6



【図７】

FIG. 7



フロントページの続き

弁理士　森田 拓
(74)代理人　100116403
弁理士　前川 純一
(74)代理人　100134315
弁理士　永島 秀郎
(74)代理人　100162880
弁理士　上島 類
(72)発明者　コンスタンティン カッペル
ドイツ連邦共和国 シュリースハイム ウツェスリング ８０

審査官　酒井 恭信
(56)参考文献　特開２０１８－０７７８０６（ＪＰ，Ａ）
米国特許出願公開第２０１８／０２３２４５１（ＵＳ，Ａ１）
特表２０１０－５２９５１８（ＪＰ，Ａ）
特開２０１４－０２９７３２（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００ － １６／９５８
Ｇ１６Ｈ　３０／４０