様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ４ 月 ２８ 日現在

機関番号：１１３０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１４３

研究課題名（和文）人工知能で目指すアミノ酸配列類似性検索法の高速化および高感度化研究

研究課題名（英文）Artificial intelligence for sequence similarity search

研究代表者
山田　和範（Yamada, Kazunori）

東北大学・情報科学研究科・准教授

研究者番号：２０７５６２１７

交付決定額（研究期間全体）：（直接経費）　　1,400,000円

研究成果の概要（和文）：アミノ酸の位置特異的置換行列（PSSM）はアミノ酸配列に対して計算可能な，アミノ酸の進化情報を有する行列形式のデータです．PSSMの構築をするためには，大きなデータベースに対して繰り返し配列の類似性検索をする必要があり，とても大きな時間がかかります．本研究では，PSSMの性質を維持したままその生成時間を短縮するための人工知能，SPBuildを開発しました．SPBuildの開発には再帰型ニューラルネットワーク（RNN）を利用しましたが，本研究ではさらにRNN自体の高度化研究に取り組みました．開発した新たなRNNであるYamRNNは既存のRNNと比較してより良い収束性能を示しました．

研究成果の学術的意義や社会的意義
生物学的文字列の類似性検索法は医学や生物学の解析をする際に最も基本的な情報科学的な解析法のひとつです．類似性検索法の利用によって，様々な発見がされてきましたし，この性能向上を達成することでさらなる発見が期待されます．今回の研究で開発した人工知能が出力する情報は，この配列類似性検索法を高性能に行うために必要な情報です．これまでにとても長い時間をかけて生成していたこの情報を高速に生成できるようにしました．また，この研究で人工知能を開発した方法は，それ自体がとても計算量が大きな方法でした．よって，この研究ではさらに発展的により計算量が少ない人工知能を開発する要素技術も新たに開発しました．

研究成果の概要（英文）：Position-specific substitution matrices (PSSMs) are matrices, which include evolutionary information about amino acids. PSSMs are fundamental information for sequence similarity search, evolutionary analysis of amino acids, etc. However, in order to generate PSSMs, it is necessary to perform repeated sequence similarity searches on a large database, which takes a lot of time. In the study, we have developed an artificial intelligence (AI), SPBuild, which could reduce the generation time of PSSMs, keeping information contents of the generated PSSMs. To develop SPbuild, we had utilized a recurrent neural network (RNN). Through the research, we realized that development of AI with existing RNNs would take a lot of time, due to its large time complexity. Thus, we had developed a novel RNN, YamRNN, which showed better convergence performance compared to existing RNNs. SPBuild and YamRNN is publicly available.

研究分野：人工知能

キーワード：人工知能　配列解析

１．研究開始当初の背景

アミノ酸プロファイルは位置特異的置換行列（position specific scoring matrix）とも呼ばれる，アミノ酸配列に対して計算可能な，アミノ酸の進化情報を有する行列データです．このアミノ酸プロファイルを用いることで，アミノ酸配列の類似性検索やアミノ酸配列の進化解析をすることが可能となる，配列解析をする際に最も基本となる情報のひとつです．しかし，興味のあるアミノ酸配列のプロファイルを構築するのにはとても大きな時間がかかります．研究開始当初においても，アミノ酸プロファイルの生成に莫大な時間がかかることは問題として認識されており，アミノ酸プロファイルを簡易的に構築するいくつかの方法（CSbuild，RPS-BLAST）が開発されていました．しかし，それらは検索するデータベースのサイズを小さくする，整理する，という，情報量を維持しながら情報量を削減しようする方法であり，研究開始当初にその活用が広まりつつあった人工知能，特に深層学習の方な洗練された人工知能を利用するようなものではありませんでした．そこで，本研究では文字列情報を処理することが可能なタイプの深層学習法の人工知能アルゴリズムである再帰型ニューラルネットワーク（RNN），特に，当時高性能であることが明らかにされ，様々な分野での応用がされつつあった超短期記憶（long short-term memory（LSTM））を利用して，アミノ酸プロファイルを高速に生成する人工知能を開発することにしました．

２．研究の目的

本研究では，アミノ酸プロファイルを高速に生成することが可能な人工知能を開発することを当初の目的としていましたが，3 年間という長い研究期間において，さらに発展的な課題に取り組みましたので，それらを併せて本研究の目的として列挙します．本研究の目的は以下のふたつです．

（１） 文字列を処理することが可能な人工知能である RNN の中でも，研究開始時点において特に性能が高いということが明らかにされ，様々な研究分野に適応されつつあった LSTM を用いて，アミノ酸プロファイルを高速に構築する人工知能の開発をすることとしました．

（２） LSTM は高性能な RNN ではあるものの，計算量が大きく，その学習に際してとても大きな時間が必要な人工知能です．アミノ酸プロファイルを高速に生成する人工知能の開発に際しても，学習に数か月の時間を要しました．よって，よりコンパクトで空間および時間計算量に優れる RNN を開発することとしました．

３．研究の方法

（１） 最初の目的を達成するために，具体的には，出力されるアミノ酸プロファイルが多くの情報を含み高性能であるものの，その生成にとても大きな時間がかかるアミノ酸プロファイル生成法を用いて，研究開始当時に得られる冗長性を除いたすべてのアミノ酸配列に対してアミノ酸プロファイルを生成しました．これらのアミノ酸プロファイルの情報は本研究で利用する教師データです．それらのアミノ酸プロファイル情報とそれに紐づくアミノ酸配列をそれぞれ教師データと入力データとして LSTM を成長させました．開発したアミノ酸プロファイルを高速に生成する人工知能 SPBuild と既存のアミノ酸プロファイル生成法との性能比較を，アミノ酸プロファイルの生成の速さと，それらのアミノ酸プロファイルを基にして行う配列類似性検索の感度（偽陽性率に対する感度）と受信者操作特性曲線（ROC 曲線）を描いた際の曲線化面積を評価指標として比較しました．

（２） 新たな RNN の開発に関しては，様々なトポロジーを持つ RNN を有するパラメータの個数以外に何の前提も設けずにランダムに大量に生成させました．それらを，RNN の記憶力を評価することに適したデータセットを用いてベンチマークにすることで，最も良い性能を示す RNN を同定しました．開発した RNN である，YamRNN と既存の RNN の性能を比較し，YamRNN の有効性を評価するとともに，YamRNN が高性能である理由の同定のための解析を行いました．

４．研究成果

（１） 新たなアミノ酸プロファイル生成法の開発

新たなアミノ酸プロファイル生成法である，SPBuild を開発しました．SPBuild の性能は，生成したアミノ酸プロファイルを利用した配列類似性検索の感度と偽陽性率，また，アミノ酸プロファイルの生成の速さを主な評価指標として評価しましたが，教師データを生成する際に利用したような厳密なアミノ酸プロファイルを構築可能な方法と比較した場合，偽陽性率に対する感度は遜色ない一方で，その計算の速さは数百倍になりました．また，それよりも少々性能が落ちるものの現在の配列類似性検索法としてデファクトスタンダートとなっている類似性検索法（PSI-BLAST）と比較した場合は，より高速により検索性能が高いアミノ酸プロファイルを生成できることが明らかになりました．また，人工知能を用いない簡易的なアミノ酸プロファイル生成法である CSBuild と RPS-BLAST と比較した場合，計算の速さは劣るものの，偽陽性率に対する感度は最大で約 2 倍でした（図 1）．また，この研究では LSTM の記憶力が今回のようなアミノ酸配列を処理する際にどれくらいの長さまで有効であるかについての解析を行いましたが，その結果として，少なくともアミノ酸配列の解析においては調べた範囲内においては記憶力が強ければ強いほど生成されるアミノ酸プロファイルの性能が良くなっていることがわかりました．また，既存のアミノ酸プロファイル生成法にも記憶力に似た機構を持つものがありましたが，それと比較しても LSTM は文脈を記憶し処理する能力が高いということがわかりました．本研究は，アミノ酸プロファイルの生成に LSTM を利用した世界で初めての研究です．開発した人工知能の性能は優れており，深層学習法の応用という観点からもその後の研究に与える影響は少なくないことが期待されます．また，アミノ酸プロファイルは医学や生物学における情報科学的な解析法をする際に最も基本的な情報です．配列類似性検索や進化解析を行う際に欠くことができない情報です．本研究の成果である SPBuild は今後そのような研究に応用されることが期待されます．SPBuild は，誰でも利用できるように公開しています．

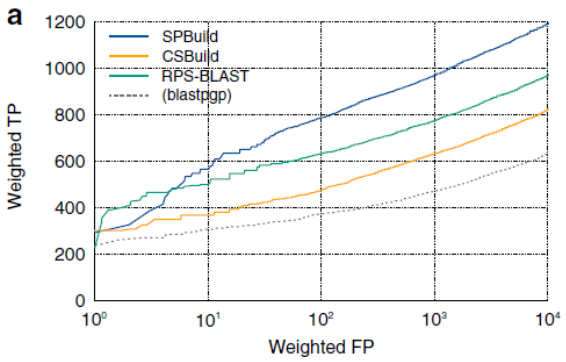

図 1．アミノ酸プロファイル生成法で生成したプロファイルを利用した配列類似性検索の性能比較

（２） 新たな RNN の開発

上述のアミノ酸プロファイル生成法の開発には多大なる時間がかかりました．教師データとして利用した配列の本数が数百万本にも達したことからデータセットが大きすぎたことがひとつの原因ですが，それよりも大きな問題として，LSTM の学習のための計算量が大きすぎるという問題がありました．そこで，本研究ではさらに発展的な研究として新たなる RNN の開発に取り組みました．開発した RNN はパラメータのサイズを既存の RNN と比較した場合，LSTM の約 50%であり，ゲート付き再帰型ユニット（GRU）と呼ばれる LSTM に次いでよく用いられている RNN の約 70%でした．パラメータサイズを減少させることは RNN の空間計算量を減少させること直結します．YamRNN は，既存の RNN と比較して空間計算量的にとても優れた RNN でした．また，その後の解析では様々な RNN との（ある一定の収束域に到達するまでの）計算時間の比較を

	Success count out of 10 trials				Number of epochs				Computation time			
The unit size of RNN layer	256	512	768	1024	256	512	768	1024	256	512	768	1024
EN	4	7	4	0	7460	12900	15000	—	893	2100	2800	—
LSTM	**10**	**10**	**10**	**10**	2900	1430	2970	4520	1250	925	2120	3490
GRU	**10**	**10**	**10**	**10**	2160	1950	2140	4090	729	983	1220	2530
S-LSTM	9	9	**10**	**10**	4350	4940	3480	5750	1070	1040	1400	2500
MGU	**10**	**10**	**10**	**10**	**1940**	**1170**	4230	7830	**491**	**421**	1750	3480
eGRU	**10**	**10**	**10**	**10**	4570	3070	2910	5840	1080	1040	1130	2480
SGU	5	**10**	**10**	**10**	27200	13200	10300	9440	7410	5100	4540	4440
IndRNN	0	0	0	0	—	—	—	—	—	—	—	—
YamRNN	**10**	**10**	**10**	**10**	4240	1850	**1690**	**1680**	795	461	**473**	**507**

図 2．YamRNN とその他の RNN の性能比較

行いましたが，YamRNN は LSTM，GRU に加えてその他のコンパクトな RNN と比較しても優れた計算時間を示しました（図 2）．さらに，各パラメータのユニットサイズを増やした際においてはその他の RNN が到達できなかった収束域まで到達しており，計算時間またその予測性能においても既存の RNN より概ね優れていました．なぜ YamRNN がそれほどまでに高性能を示したかについては現在までのところ不明です．既存の RNN の開発過程のような人の脳を模倣することをせずに開発した YamRNN が高性能を示したことから，これまでに発見されていない何らかの機構によってそれが達成されている可能性は否定できず，将来的な更なる研究の課題です．配列情報を処理可能なニューラルネットワークである RNN は様々な研究課題に適用可能な技術です．大量のデータを処理することが求められている研究分野において，計算量に優れた RNN は有用であり，YamRNN は今後様々な分野において応用されることが期待される．YamRNN は最もよく用いられている深層学習フレームワークである TensorFlow で利用可能な状態にて公開されている．

（3） 発表論文

- Yamada KD, Kinoshita K, De novo profile generation based on sequence context specificity with the long short-term memory network, BMC Bioinformatics, 19(1):272, 2018

- Yamada KD, Lin F, Nakamura T, Developing a novel recurrent neural network architecture with fewer parameters and good learning performance, Interdisciplinary Information Sciences, 27(1):25-40, 2021

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Yamada KD, Kinoshita K	４．巻 19
２．論文標題 De novo profile generation based on sequence context specificity with the long short-term memory network	５．発行年 2018年
３．雑誌名 BMC Bioinformatics	６．最初と最後の頁 272
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-018-2284-1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 YAMADA Kazunori D、LIN Fangzhou、NAKAMURA Tsukasa	４．巻 27
２．論文標題 Developing a Novel Recurrent Neural Network Architecture with Fewer Parameters and Good Learning Performance	５．発行年 2021年
３．雑誌名 Interdisciplinary Information Sciences	６．最初と最後の頁 25～40
掲載論文のDOI（デジタルオブジェクト識別子） 10.4036/iis.2020.R.01	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関