

修士論文

Multi-discriminator
generative adversarial networks を用いた
タンパク質主鎖構造デザイン

令和 2 年度修了

三重大学大学院 工学研究科 情報工学専攻
ヒューマンインタフェース研究室

清水 一生

はじめに

生物はタンパク質の情報を遺伝子に記録しており、遺伝情報に従って 20 種類のアミノ酸を結合させる事によりタンパク質を発現する。タンパク質は生体内でのあらゆる機能に関連している。近年研究が進んでいる天然変性タンパク質などの構造を持たないタンパク質が存在する一方で、タンパク質の多くは分子内の相互作用により折り畳まれた 3 次元構造を持つ。そして、この構造がタンパク質それぞれの機能を果たす上で重要な役割を担っている。Silva らの研究 [1] ではタンパク質設計において抗がん剤 IL-2 と IL15 の強力かつ選択的な模造品を設計しており、タンパク質設計技術を医学分野へ応用することが期待されている。さらにタンパク質の構造を一からデザインして活用するというタンパク質デノボデザインが行われており、古賀らの研究 [2] ではタンパク質の二次構造とそれらをつなぐループ構造に着目したタンパク質の構造構築原理を提唱しており、タンパク質構造及びアミノ酸配列をデザインしている。しかし、デノボデザインで設計の基礎となる主鎖構造は熟練した研究者の手によって作成されるためデノボデザインのハードルは高く、また構造のバリエーションを増やすことは難しい。そこでデノボデザインで作成可能な構造の選択肢を広げ、新たな構造を用いたデザインを容易に実現することを目指して主鎖構造を自動で生成する手法を提案する。

深層学習を利用したタンパク質主鎖構造の生成を試みている研究として、Anand らによる生成的深層学習モデルの一つである Deep Convolutional Generative Adversarial Networks (DCGANs) [9] を用いたタンパク質主鎖の部分生成の研究 [3] がある。Anand らは学習データとしてタンパク質の 3 次元構造を 2 次元の行列として扱える距離マップを用いたが、畳み込み特徴抽出のみを利用した手法での生成は難しいことを示している。また生み出されたタンパク質を様々な分野に応用するためには、実現可能性と多様性のあるデザイン技術が必要となる。我々は Anand らが提案した DCGANs を利用した生成手法を踏襲し、新たな識別器を導入することにより生成される距離マップの質の向上を試みた。

本研究では、Multi Discriminator Generative Adversarial Networks (MDGANs) を用いたタンパク質主鎖構造の生成手法を提案する。タンパク質の主鎖を構成する各アミノ酸に含まれる炭素原子の 1 種である C_{α} 間の距離を総当たりで求めることにより作成した 2 次元

行列（距離マップ）を用いて MDGANs を学習させることで、学習データに類似した距離マップの生成を試みる。生成された距離マップは Multidimensional Scaling (MDS) によって 2 次元行列から 3 次元座標に変換される。MDS は、距離マップとの平均誤差（ストレス値）を最小化するように各 C_α 座標を反復的に計算する手法として用いる。ストレス値が小さいほど距離マップを正確に 3 次元座標に復元できる。また、GANs の学習度合いを測る指標としては、十分な最適化を行った後の MDS のストレス値を用いる。

本手法では DCGANs に新たな識別器を追加したモデルを MDGANs と名付け、2 種類の MDGANs (MDGANs1, MDGANs2) の実装と学習を行う。そして、Anand らが提案した DCGANs と提案手法である MDGANs を比較することにより性能を評価する。DCGANs では、距離マップの畳み込みで抽出された特徴を用いて、識別器が生成器によって生成された距離マップとデータセットから生成された距離マップを識別し、学習を繰り返す。MDGANs1 は DCGANs に加え、距離マップの対角要素の特徴を捉えるための識別器 (対角要素識別器) と距離マップ行列の対称性を捉えるための識別器 (非対角要素識別器) について識別と学習を行う。MDGANs2 は DCGANs に加え、距離マップから予測されるアミノ酸配列情報について識別と学習を行う。

学習の結果、MDGANs1 で最小の MDS ストレス値 ($15460 \pm 197\text{\AA}$) が得られた。この結果は DCGANs よりもストレス値が小さく、主鎖生成において MDGANs が有効であることが示された。また、距離マップの対角要素の特徴と距離マップ行列の対称性についての検証も行い、MDGANs1 は十分にこれらの特徴を学習できていることを確認した。

本論文では、1 章に研究背景と目的、2 章に本研究に関連する技術、3 章に本研究の提案手法と評価手法、4 章に実験に使用したデータセットの詳細、5 章に計算機による生成実験と考察、最後に 6 章で全体のまとめと今後の課題について述べる。

目次

はじめに	i
第 1 章 緒言	1
1.1 タンパク質主鎖デザインとは	1
1.2 本研究の基礎知識	2
1.3 先行研究	4
1.4 研究目的	4
第 2 章 本研究に関連する技術	5
2.1 Neural Networks: ニューラルネットワーク	5
2.2 Convolutional Neural Network: CNN	10
2.3 Residual Neural Network: 残差ネットワーク	13
2.4 Long short term memory: LSTM	14
2.5 Generative adversarial networks: GANs	14
2.6 Multidimensional Scaling: MDS	15
第 3 章 提案手法	16
3.1 MDGANs1	16
3.2 MDGANs2	20
3.3 評価手法	22
第 4 章 データセット	23
4.1 GANs 用の学習データ	23
4.2 アミノ酸配列データ	24
第 5 章 計算機による実験	25
5.1 実験条件	25
5.2 実験結果	26

5.3	実験考察	28
第 6 章	結言	31
6.1	まとめ	31
6.2	今後の展望	31
付録 A	ソースプログラム等のデータ	32
A.1	プログラム	32
A.2	実験データ	32
A.3	環境構築情報	32
A.4	プログラムの詳細	33
付録 B	発表資料	34
B.1	修士論文発表資料	34
	謝辞	37

第 1 章

緒言

1.1 タンパク質主鎖デザインとは

生物はタンパク質の情報を遺伝子に記録しており，遺伝情報に従って 20 種類のアミノ酸を結合させる事によりタンパク質を発現する．タンパク質は生物の形を構成する主要な要素であり，かつ生体内でのあらゆる機能に関連している．近年研究が進んでいる天然変性タンパク質などの構造を持たないタンパク質も存在する一方で，タンパク質の多くは分子内の相互作用により折り畳まれた 3 次元構造を持つ．そしてこの構造がタンパク質それぞれの機能を果たす上で重要な役割を担っている．Silva らはタンパク質設計において抗がん剤 IL-2 と IL15 の強力かつ選択的な模造品を設計しており [1]，タンパク質設計技術の医学分野への応用が期待されている．さらにタンパク質の構造を一からデザインして活用するタンパク質デノボデザインが行われており，古賀ら [2] はタンパク質の二次構造とそれらをつなぐループ構造に着目したタンパク質の構造構築原理を提唱し，タンパク質構造及びアミノ酸配列をデザインしている．しかし，デノボデザインにおいて設計の基礎となる主鎖構造は熟練した研究者の手によって作成されるため，デノボデザインのハードルは高く，また構造のバリエーションを増やすことは難しい．そこでデノボデザインで作成できる構造の選択肢を広げ，新たな構造を用いたデザインを容易に実現することを目指して，主鎖構造を自動で生成する手法を提案する．

1.2 本研究の基礎知識

1.2.1 タンパク質構造

タンパク質は 20 種類のアミノ酸が枝分かれすることなく、1 本の鎖状につながった高分子である。ペプチド結合は図 1.1 のように、アミノ酸に含まれるカルボキシル基とアミノ酸が脱水縮合することで形成され、複数のアミノ酸が繋がることでタンパク質を構成している。アミノ酸の中でアミノ基とカルボキシル基を共有結合する炭素原子はカルボキシル基に含まれる C 原子と区別するために C_{α} 原子と呼ばれる。

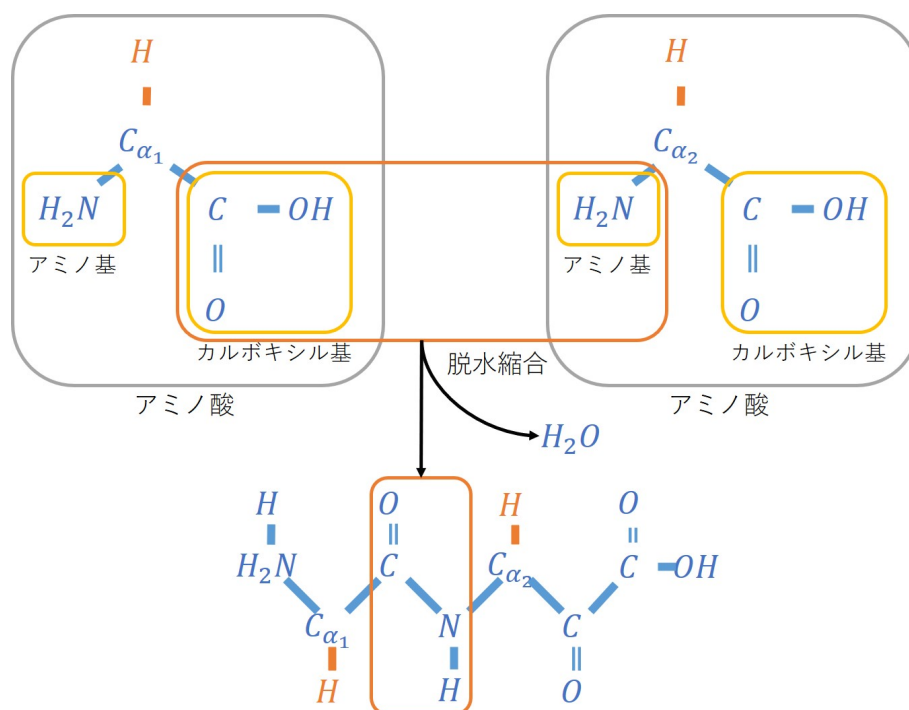


図 1.1: 2 つのアミノ酸の脱水縮合

図 1.2 に示すように複数のアミノ酸がペプチド結合して繋がった際に現れる N , C_{α} , C の原子列を主鎖と呼ぶ。主鎖はタンパク質構造の骨格であり、本研究では C_{α} の座標を基に作成された距離行列（距離マップ）を深層学習モデルを学習するためのデータとして扱った。

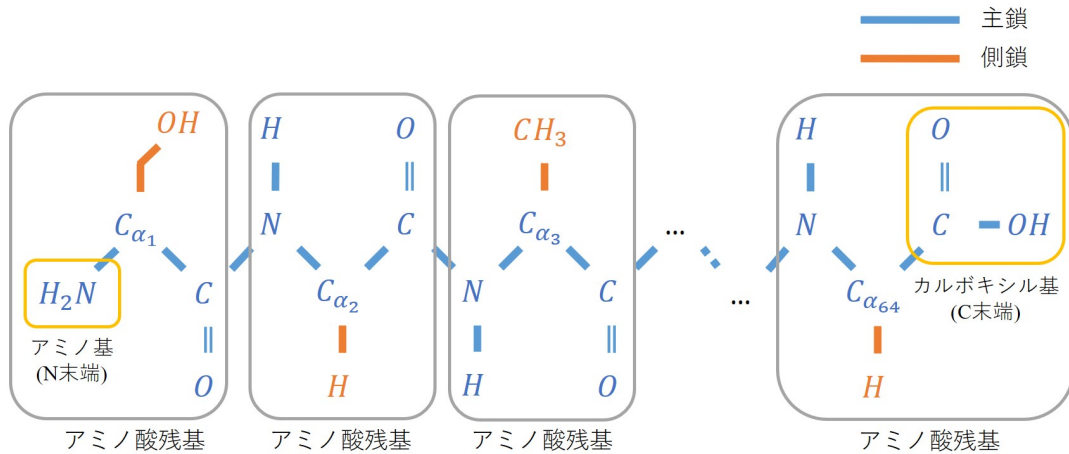


図 1.2: タンパク質構造の一部

1.2.2 距離マップ

Distance map (距離マップ) とは 2 つの C_α 原子間の 3 次元空間上の距離をすべての C_α 原子の組み合わせに対して計算した行列のことである。距離マップはタンパク質の 3 次元構造を 2 次元の行列として扱えるため、画像データを入力に持つ Convolutional Neural Network (CNN) などの画像の扱いが得意な深層学習モデルと相性が良い。また、距離マップは 3 次元構造の回転や並進に対しても不変であるという特性を持つ。図 1.3 はタンパク質の 3 次元構造と距離マップが相互変換可能であることを示す。本論文では、距離マップの距離値を輝度に変換し、可視化して表示する。

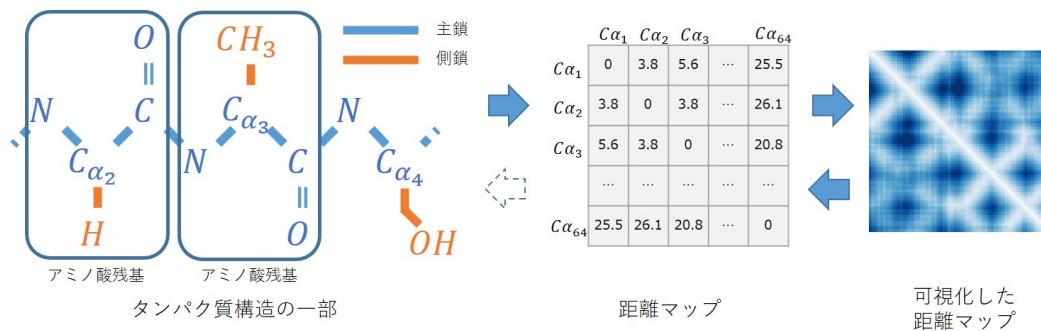


図 1.3: タンパク質構造と距離マップの相互変換

1.3 先行研究

1.3.1 DCGANs による主鎖生成に関する研究

深層学習を利用することでタンパク質主鎖構造の生成を試みている研究として、Anand らによる生成的深層学習モデルの一つである Deep Convolutional Generative Adversarial Networks (DCGANs) [9] を用いたタンパク質主鎖の部分生成の研究 [3] がある。Anand らは、距離マップのみからのタンパク質生成が難しいことを示している。

1.3.2 アミノ酸配列予測に関する研究

Chen らは、距離マップからアミノ酸配列を予測する手法を提案している [14]。ResNet [5] と LSTM を用いたネットワークモデルを用いて、距離マップから特徴抽出を行い、得られた特徴を LSTM の入力としてアミノ酸配列を予測する。本研究の生成モデル中で配列予測を行う部分では、Chen らの研究を参考にモデルを設計している。

1.4 研究目的

DCGANs で生成されたタンパク質を様々な分野に応用するためには、実現可能性と多様性のあるデザインを行う技術が重要となる。我々は Anand らが提案した DCGANs を用いた生成手法を踏襲しつつ、新たな識別器を導入することにより生成される距離マップの質の向上を試みる。

第 2 章

本研究に関連する技術

本章では，提案手法で用いる主鎖自動生成手法を構成する要素技術について述べる．

2.1 Neural Networks: ニューラルネットワーク

ニューラルネットワークは脳の神経細胞を模したアルゴリズムであり，ニューロンモデルと呼ばれるユニットとノードを多層的に結合したモデルである．ニューラルネットワークは図 2.1 で示すように入力層，中間層，出力層から構成される．

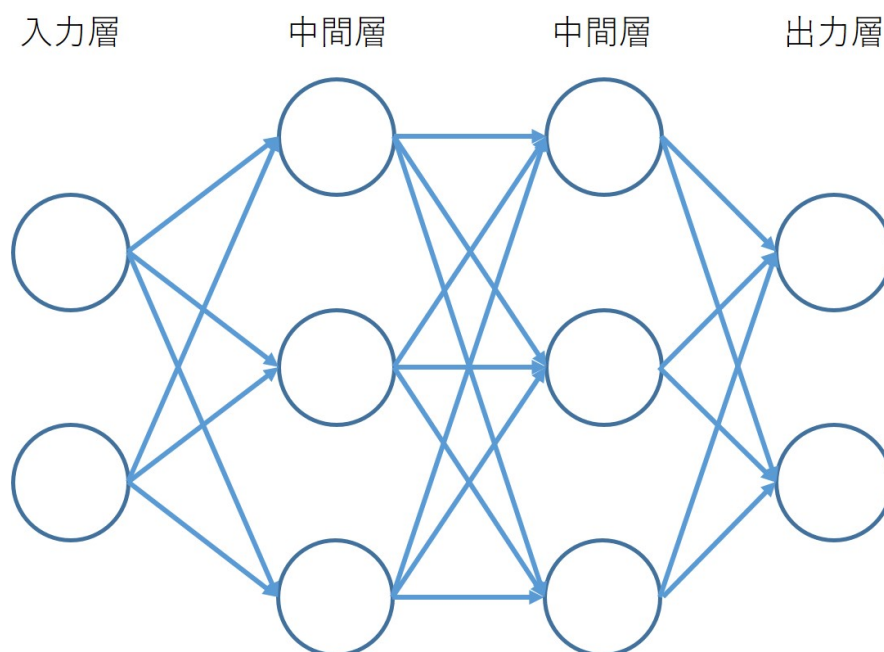


図 2.1: ニューラルネットワークの例

2.1.1 ニューロン: 神経細胞

ニューロンは複数の受信器（樹状突起：dendrite）と一つの送信器（軸索：axon）で構成され、軸索上を伝わる電気パルスによって他のニューロンへと情報が伝達される。軸索は、シナプスと呼ばれるインターフェースを介して、電気パルスの到来をニューロンに伝達する。ニューロンは電子パルスを受け取ることで、細胞内の電気レベル（膜電位）が上下する。この変動は、入力を受け取るシナプスの状態（シナプス伝達強度）に依存する。そして、膜電位の値がある一定値を超えると、その電子パルスは発信され、軸索を通して他のニューロンへと伝達される。

2.1.2 ニューロンモデル

複数の受信器と一つの送信器で構成されるニューロンを、単純な数理モデルで表したものをニューロンモデルと呼ぶ。 x_1, x_2, \dots, x_n をニューロンへの入力、 $\omega_1, \omega_2, \dots, \omega_n$ をニューロンへの伝達強度（重み）、 b をニューロンの発火のしやすさをコントロールするバイアス、 z をニューロンの出力とする。出力 z は接続された次のニューロンの入力となる。ニューロンの入出力の関係は式 (2.1), (2.2) で示される。

$$y = \sum_{i=1}^n \omega_i x_i + b \quad (2.1)$$

$$z = f(y) \quad (2.2)$$

式 (2.2) の $f()$ は非線形関数であり、活性化関数と呼ばれる。活性化関数はニューロンの応答に非線形性を与える役割がある。以下に 4 種の活性化関数を示す。

- ReLU 関数

ReLU(Rectified Linear Unit) 関数は式 (2.3) で定義される。

$$f(x) = \begin{cases} x & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad (2.3)$$

入力値が 0 以上の場合は、入力値がそのまま出力値となり、0 未満の場合は 0 となる。ReLU 関数は Convolutional Neural Network; CNN [4] の畳み込みフィルターや全結合層の後に置かれ、抽出された特徴を強調する働きがある。

- Leaky ReLU 関数

Leaky ReLU (Rectified Linear Unit) 関数は式 (2.4) で定義される。

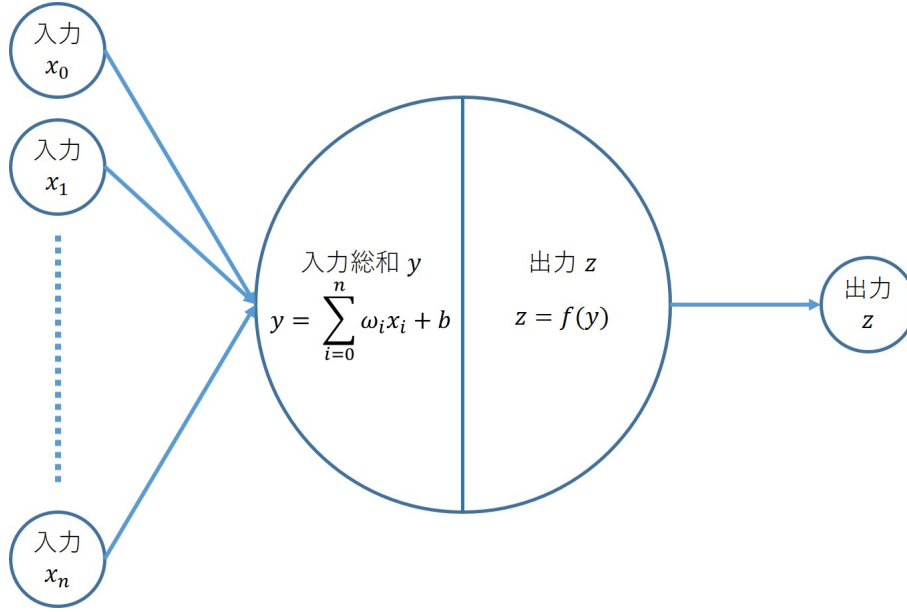


図 2.2: ニューロンモデル

$$f(x) = \begin{cases} x & (x \geq 0) \\ \alpha x & (x < 0) \end{cases} \quad (2.4)$$

α は 0.01 ほどの小さな値を取る. 0 未満の場合, 入力値が 0 より下であっても 0 にはならず, 勾配が発生する. ReLU 関数では入力値が 0 以下になると勾配が消失し, 学習が進みにくくなることがあり, Leaky ReLU はこの問題を解消している. 本研究では GAN [8] の生成器内部で使用する.

- シグモイド関数

シグモイド関数は式 (2.5) で定義される.

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (2.5)$$

入力値 x が小さいほど出力値は 0 に近づき, 大きいほど出力値は 1 に近づく. シグモイド関数 $f(x)$ の出力範囲は $0 < f(x) < 1$ であり, 2 クラスの識別問題の場合によく用いられる活性化関数である.

- softmax 関数

softmax 関数は式 (2.6) で定義される.

$$f(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}} \quad (2.6)$$

softmax 関数は i 個存在する入力値 x_i をそれぞれ確率とし、入力値の総和が 1 となるように正規化する活性化関数である。

2.1.3 Feedforward Neural Networks: 順伝播型ニューラルネットワーク

順伝播型ニューラルネットワークは単純パーセプトロンを並べたものを一つの層とし、複数の層を並べて結合したネットワークモデルであり、多層パーセプトロン (multilayer perceptrons: MLPs) と呼ばれる。多層パーセプトロンは入力層、中間層、出力層の 3 種類の層から構成される。図 2.3 は入力層が 1 層、中間層が 2 層、出力層が 1 層の合計 4 層で構成されており、入力データに線形変換と活性化関数による非線形変換を繰り返すことで、任意の関数を近似することが可能である。複数の中間層を持つネットワークは表現力が大きいことが分かっているが、入力に高次元の特徴ベクトルを入力する場合、膨大な数の中間層と複雑な活性化関数によって、ネットワークのパラメータ数が急増する。多数のネットワークパラメータの学習には、膨大な時間と計算資源を必要とするため、学習が困難になる。

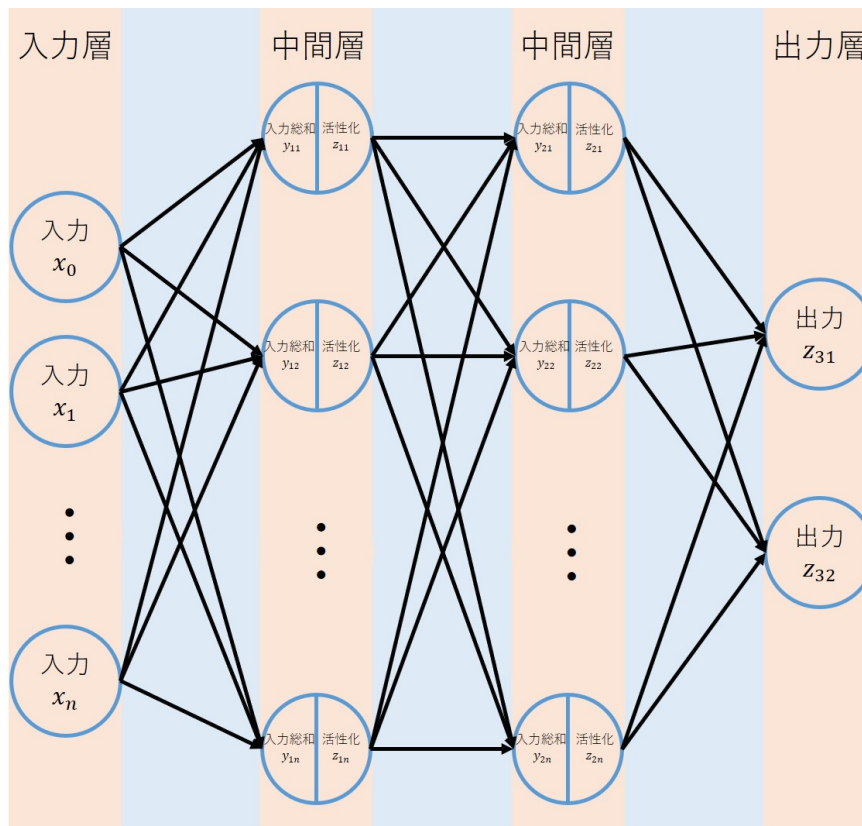


図 2.3: 中間層を 2 つを持つ多層パーセプトロンの例

2.1.4 損失関数

損失関数はニューラルネットワークの出力と正解との差異を表す損失値を求める関数であり、損失値が小さくなるように最適なパラメータを探索する。クラス分類の問題を解く際、入力データをニューラルネットワークに通して得られた出力と、予め付けられている正解ラベルとの誤差を損失関数によって計算することができる。この損失関数には、任意の関数を使用することができるが、一般的に 2 乗和誤差 (2.8), 交差エントロピー誤差 (2.8) などが用いられる。

$$E = \frac{1}{2} \sum_k (y_k - t_k)^2 \quad (2.7)$$

$$E = - \sum_k t_k \log y_k \quad (2.8)$$

ここで、 E は出力される損失 (Loss) 値、 y_k はネットワークの出力から得られる事後確率の値、 t_k は one-hot 表現された正解ラベルである。

2.1.5 誤差逆伝播法

ニューラルネットワークの学習は損失関数で得られた値を最小化するように重みパラメータを更新することでモデルの最適化を行う。誤差逆伝播法 (Backpropagation) [7] は、ニューラルネットワークの学習に用いられるアルゴリズムであり、損失関数の傾斜を計算し、ネットワーク内の重みを最適化する。損失関数の傾斜である勾配から、損失関数の値を減らすように重みの更新を行う。逆伝播では入力層から順伝播することで得られる出力値を出力ノードから入力ノードへ伝播することで、ニューラルネットワーク内の重みが更新される。

2.1.6 ミニバッチ学習

ミニバッチ学習とはニューラルネットワークの学習方法の一つである。膨大なデータを学習に使用する場合、すべてのデータに対して損失関数の値を求めるには時間がかかる。そこで、ミニバッチと呼ばれる無作為あるいは順に選んだデータの一部分を用いて学習するミニバッチ学習が使われる。本研究ではこのミニバッチ学習を行っており、ミニバッチ学習を繰り返して全データが学習される単位を epoch とする。epoch ごとにミニバッチ内のデータはランダムに抽出される。

2.2 Convolutional Neural Network: CNN

畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) は、画像から特徴を抽出する手法の一種である。CNN は主に畳み込み層、活性化層、プーリング層で構成されたモデルを繰り返し、入力から出力まで伝播させることで、特徴を得ることができる。画像認識問題では得られた特徴を全結合層に受け渡し、出力層でモデルが推定した事後確率ベクトルを出力する。LeCun らの研究 [4] では CNN における誤差逆伝播法を用いた学習法が確立され、文字認識において高い精度を収めている。

2.2.1 畳み込み層

畳み込み (convolution) は入力された画像の一部と任意の大きさのフィルタの積和演算をフィルタをスライドさせながら画像全体に対して行う処理である。

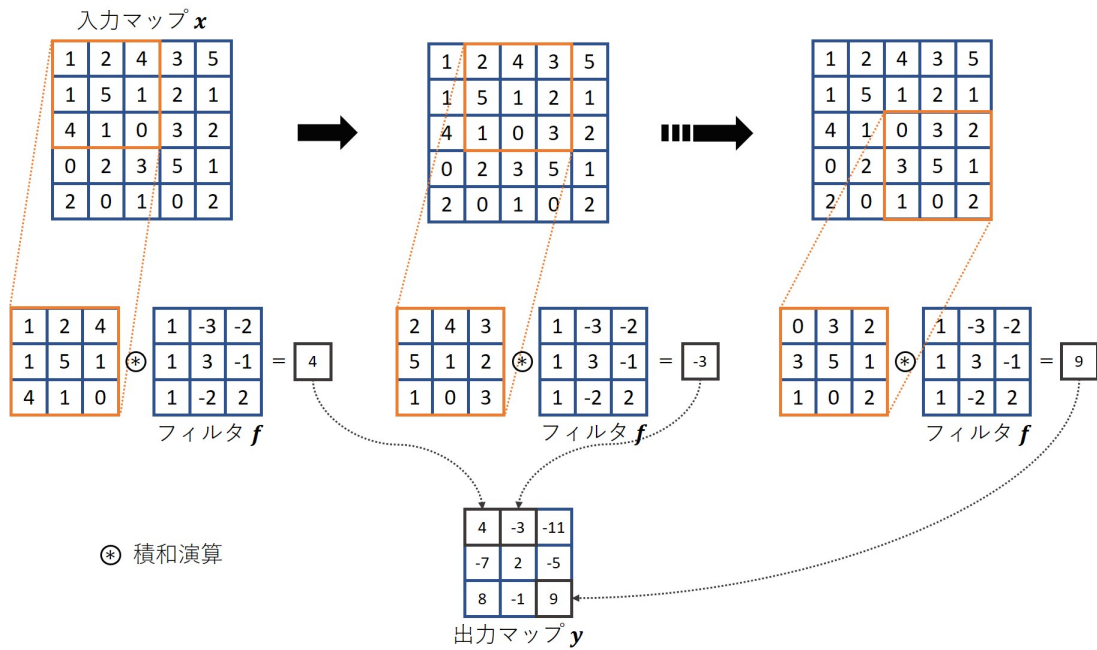


図 2.4: 畳み込み処理の例

図 2.4 は入力マップ x に対してサイズが 3×3 のフィルタでストライド値 1 の畳み込みを行う様子を示していて、出力マップ y は式 (2.9) で定義される。

$$y_{(i,j)} = \sum_{p=1}^3 \sum_{q=1}^3 f_{(p,q)} \times x_{(i+p,j+q)} \quad (2.9)$$

$f_{(p,q)}$ はフィルタの p 行 q 列目の要素, $y_{(i,j)}$ は出力マップの i 行 j 列目の要素を表している.

2.2.2 活性化層

畳み込み層で得られた特徴に対しては, 一般的に活性化処理が行われる. 活性化層への入力マップは活性化関数を通すことで, 非線形変換が行われて出力される. 図 2.5 に一般的に用いられる代表的な活性化関数である ReLU 関数を用いた, 活性化処理が行われたマップを示す. ReLU 関数は式 (2.10) で示され, 負値を 0 に, 正值はその値を返す関数である.

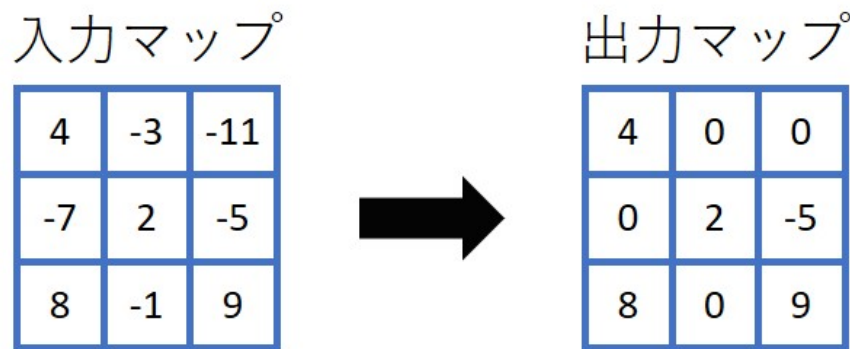


図 2.5: ReLU 関数を用いた活性化処理の例

$$f(x) = \begin{cases} x & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad (2.10)$$

2.2.3 プーリング層

プーリング層で行われるプーリングは畳み込みによる特徴抽出を行った後, 活性化関数を通して最適化されたマップを, 位置ずれに対して頑健にするために行われる処理である. 一般的に任意の大きさの局所領域内の最大値を取る max pooling 処理が行われ, マップのサイズを圧縮する. 特徴マップサイズが 3, 局所領域サイズが 2×2 の場合の例を図 2.6 に示す.

2.2.4 Deconvolutional Neural Network: 逆畳み込み層

逆畳み込みはマップに対して画素間に値を追加することでマップサイズを拡大し, 拡大したマップにフィルタを適用して畳み込みを行う処理である. 通常の畳み込み処理では

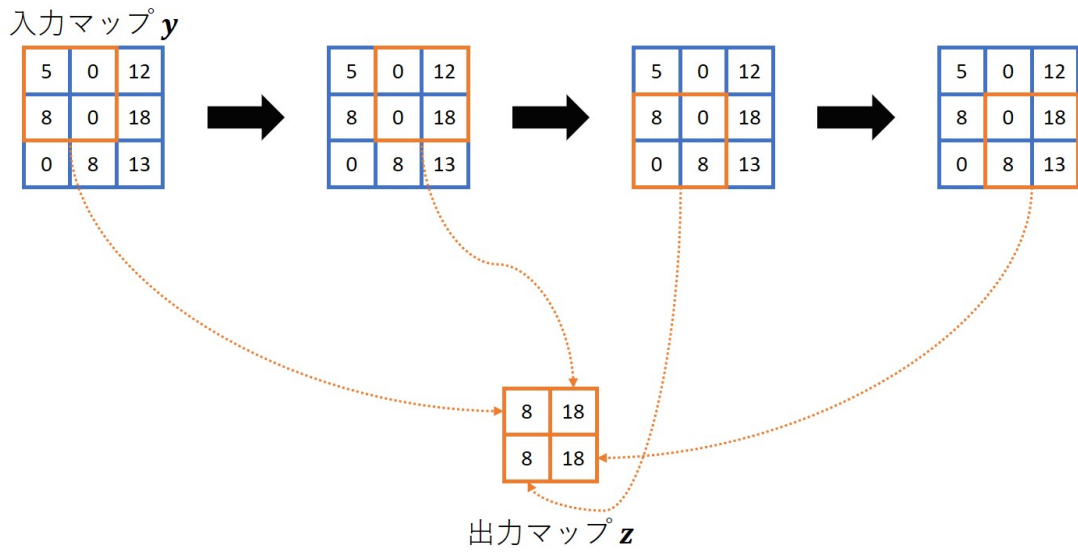


図 2.6: max pooling によるプーリング処理の例

マップサイズが縮小されるのに対し、逆畳み込みは任意のサイズの拡大を行うことができ、入力より大きいサイズのマップを出力できる。入力マップサイズ 2×2 、フィルタサイズが 2×2 の場合の例を図 2.7 に示す。

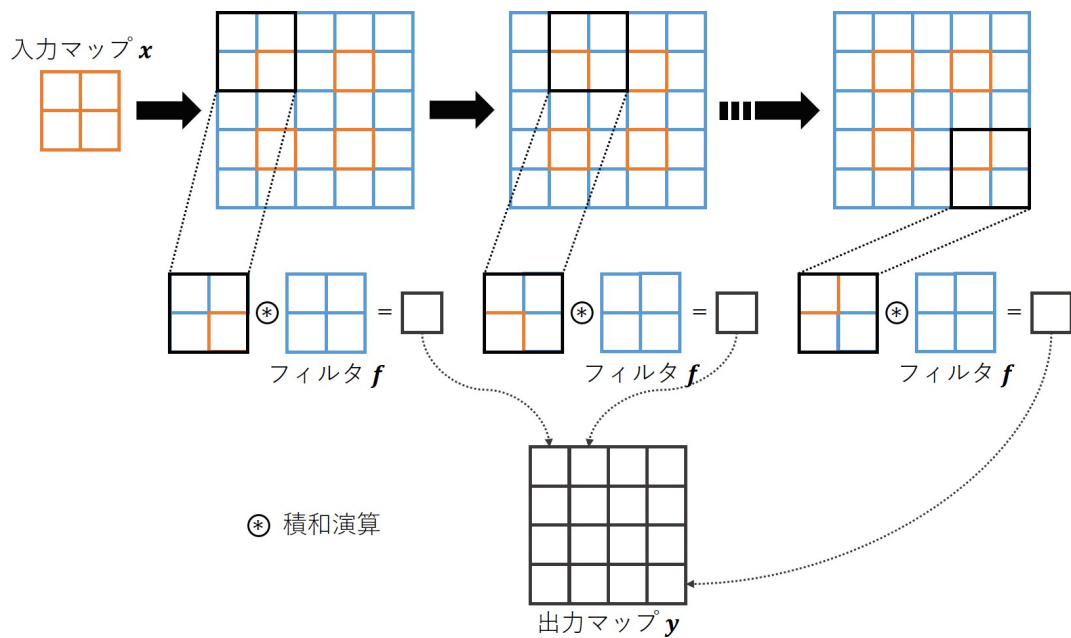


図 2.7: 逆畳み込み処理の例

2.3 Residual Neural Network: 残差ネットワーク

残差ネットワーク (ResNet) は層ごとの入出力の差分を学習することで、深層学習における学習を容易にするネットワークモデルである。多層ニューラルネットワークは層から層への伝播は主に積和演算で構成されることから、入出力の差は層の増加によって指数関数的に減少する。このため、多層ニューラルネットワークは表現力が高い一方、深層の学習では層への入出力の差が小さくなり、勾配が消失することがある。また、多層では重みの掛け算が膨大となり、入出力の差が大きくなりすぎて、勾配発散が発生することがある。勾配消失と勾配発散が発生すると学習ができなくなり、想定した精度も期待できなくなる。そこで ResNet では、層における出力から入力引いた残差を学習し、勾配消失を抑えることができる。残差の学習では、図 2.8 で示すようなショートカットコネクションが行われ、式 (2.11) で表される伝播が行われる。

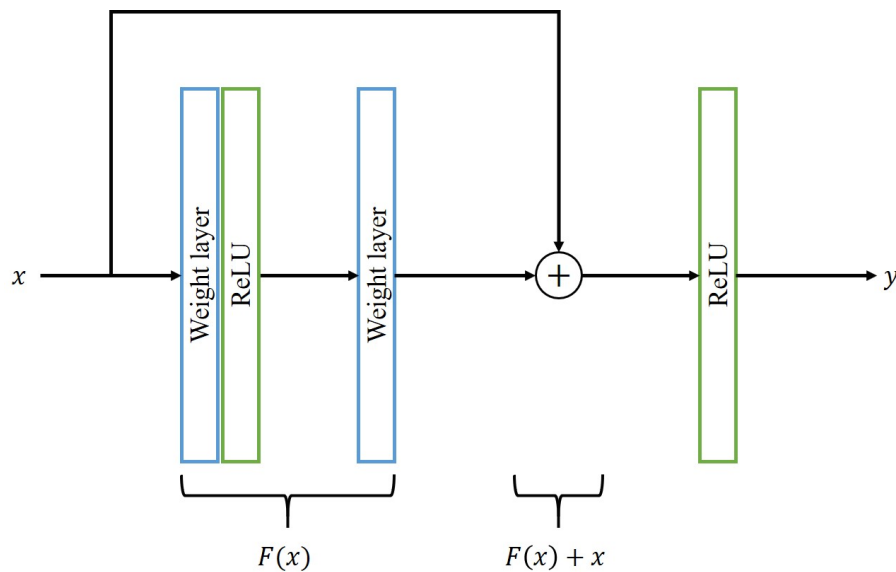


図 2.8: ショートカットコネクションの例

$$y = F(x) + x \quad (2.11)$$

このモデルを用いることで非常に表現力を持つ深層のネットワークモデルの学習により、物体認識問題において高い精度が得られている [5]。本研究ではアミノ酸配列予測のためのネットワーク内で使用している。

2.4 Long short term memory: LSTM

Long short term memory (LSTM) は自然言語などの時系列データを扱うニューラルネットワークのモデルである。連続するデータの予測に対して有効な手法であり、Chen らの研究 [14] でアミノ酸配列予測を ResNet と LSTM を用いて行っており、本研究でも Chen らのネットワークを参考にアミノ酸配予測器を作成した。

2.5 Generative adversarial networks: GANs

Generative adversarial networks (GANs) は Goodfellow らによって提唱された教師なし学習モデル [8] の一種であり、データから特徴を学習することで実在しないデータの生成が可能である。GANs は主に生成器と識別器という 2 種類のネットワークを持ち、生成器と識別器の損失値 (loss) は式 (2.12), (2.13) で表される。 m はバッチサイズ、 z は生成器へ入力するノイズ、 G は生成器の出力を示す関数、 D は識別器の出力を示す関数である。この 2 つのネットワークが競い合うように訓練されるため、教師なし学習が可能である。

$$\text{生成器}_{\text{loss}} = \frac{1}{m} \sum_{i=1}^m \log (1 - D(G(z^{(i)}))) \quad (2.12)$$

$$\text{識別器}_{\text{loss}} = \frac{1}{m} \sum_{i=1}^m \left[\log D(x^{(i)}) + \log (1 - D(G(z^{(i)}))) \right] \quad (2.13)$$

2.5.1 Deep Convolutional Generative adversarial networks: DCGANs

Deep Convolutional Generative adversarial networks (DCGANs) は Radford らのによって研究 [9] によって提案された GANs の生成器と識別器に対して CNN を利用する手法である。Anand らの部分的な主鎖生成の研究 [3] でも利用されており、本提案手法でも主鎖構造全体を生成する一手法として図 2.9 のようなモデルを利用している。

2.5.2 Multi Discriminator Generative Adversarial Networks: MDGANs

Multi Discriminator Generative Adversarial Networks (MDGANs) は Hardy らの研究 [10] で示された複数の識別器を持つ GANs である。複数の識別器を持つことで、学習が難しい GANs において性能の向上しうることが示されている。本研究では MDGANs に新たな識別器を追加した手法を提案する。

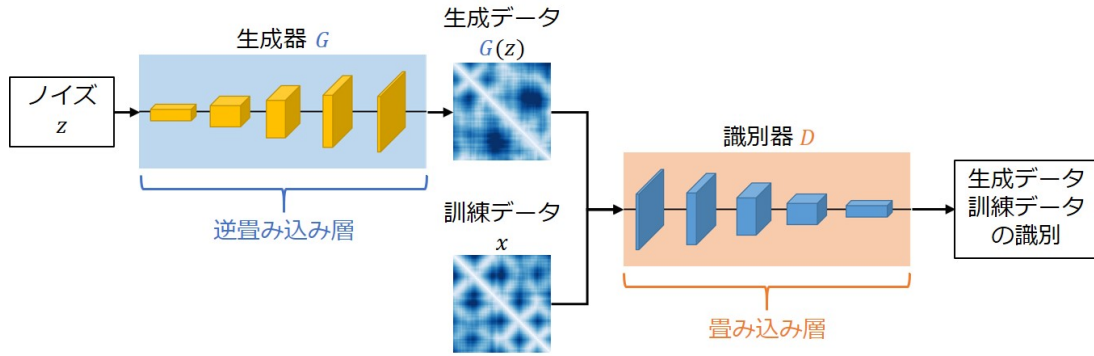


図 2.9: DCGANs を用いる主鎖生成手法

2.6 Multidimensional Scaling: MDS

Multidimensional Scaling (MDS) [11] は多変量解析手法の一種である．この解析ではユークリッド空間上に距離データを配置した距離構造データを扱うことができる．MDS では式 (2.14) で定義されるストレス値 σ を反復計算により最小化することで i 番目の 3 次元座標 r_i を求める． $d_{(i,j)}$ は i 番目と j 番目の距離を示している．最適な 3 次元座標を求めるには SMACOF アルゴリズムを利用する．

$$\sigma = \sum_{i < j} (d_{(i,j)} - \|r_i - r_j\|)^2 \quad (2.14)$$

第 3 章

提案手法

本章では, Multi Discriminator Generative Adversarial Networks (MDGANs) を用いるタンパク質主鎖構造生成手法の詳細について示す. 提案手法では 2 種類の生成手法として MDGANs (MDGANs1, MDGANs2) を作成したため, それぞれの生成手法のネットワークモデルを示す.

3.1 MDGANs1

MDGANs1 は図 3.1 に示すように生成器を 1 つ, 識別器を 3 つ持つネットワークである. 生成器は図 3.2 で示すモデルを通して, ノイズから距離マップを生成する. ここではタンパク質の残基数を $n = 64$ としている. 生成された距離マップはデータセットの距離マップとともに各識別器へ入力される. 畳み込み識別器 D_1 は CNN を用いた距離マップからの特徴抽出を行い, 距離マップが生成器で作られたものか, データセットから作成されたものかを分類する. 図 3.3 に畳み込み識別器のネットワークモデルを示す.

距離マップは 2 つの C_α 間の距離をすべての C_α の組み合わせで求めることによって生成される. そのため, データセットから作成された距離マップの対角要素は自分自身との距離を示しているため, 0 となる. 対角要素識別器 D_3 は距離マップの対角要素を取り出し, 3 層のニューラルネットワークを用いて対角要素が 0 となるように学習させる. 図 3.4 に対角要素識別器のネットワークモデルを示す.

データセットから作成された距離マップは対称行列であるため, 距離マップの n 行と n 列の要素が一致する. この性質を用いて, n 次元の行と列を連結した $2n$ 次元の入力データを n 個作成し, 1 層目のニューラルネットワークに入力する. 1 層目は入力データ $n \times n$ 次元データに対して n 次元の出力が得られ, 次の層への入力とする. 2 層目は n 次元のデータを入力として受け取り, 判定結果を 1 次元で出力する. 非対角要素識別器 D_2 は図 3.5 に示す.

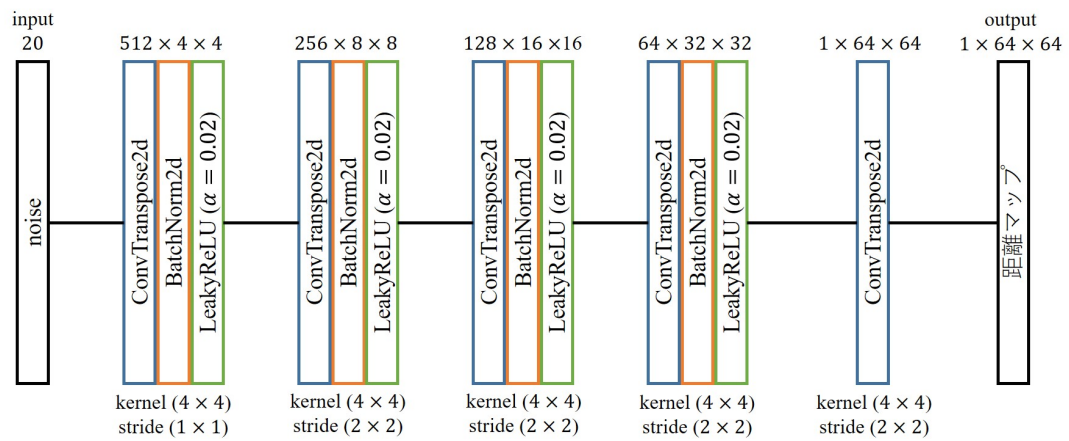
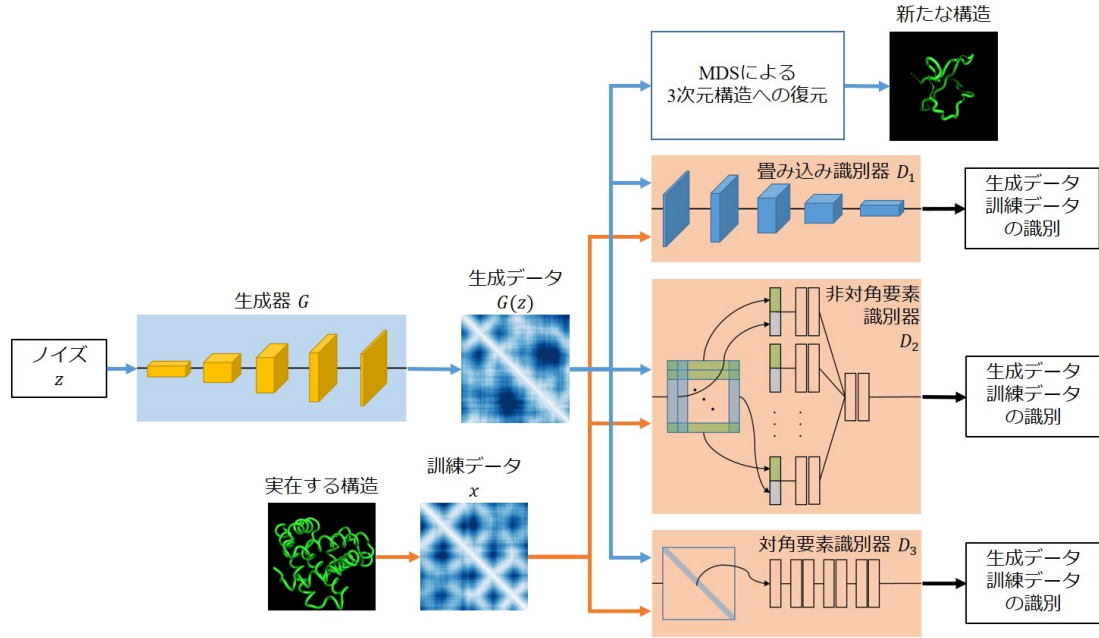


図 3.2: 生成器のモデル

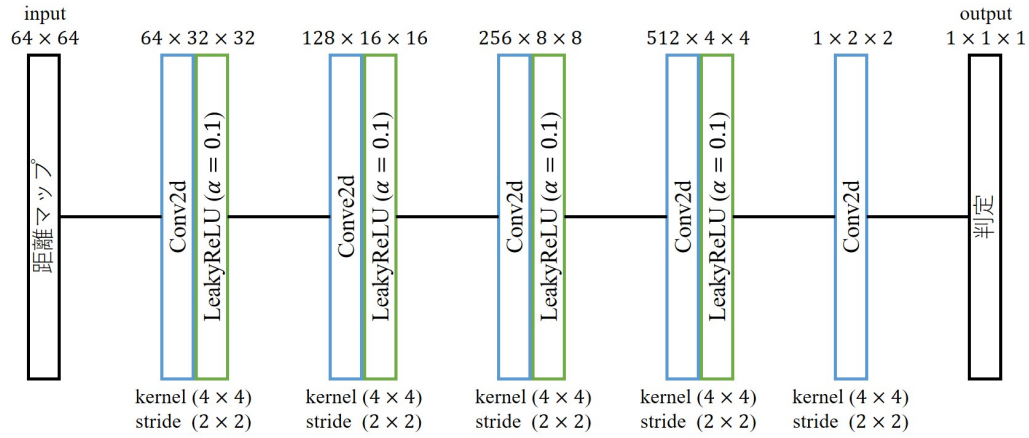


図 3.3: 畳み込み識別器のモデル: 64 残基のタンパク質主鎖構造を距離マップに変換した 64×64 次元のデータを入力とする. 畳み込み層の $A \times B \times C$ は大きさ $B \times C$ のフィルタを A 個持つ層であることを示す.

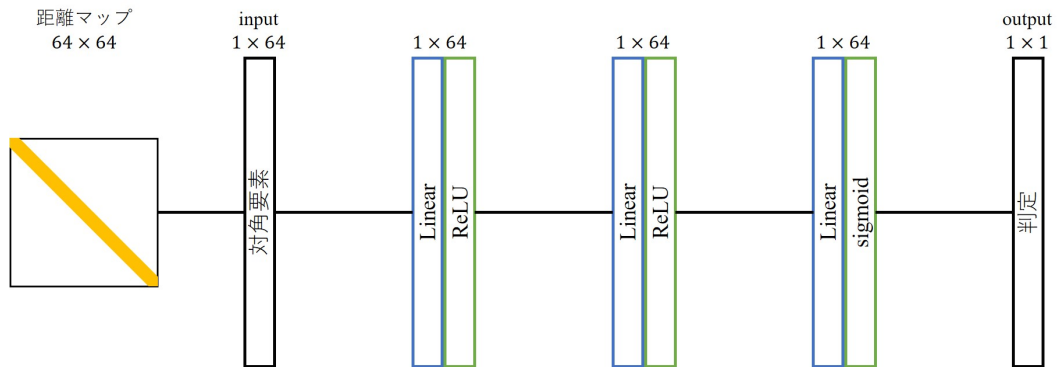


図 3.4: 対角要素識別器のモデル

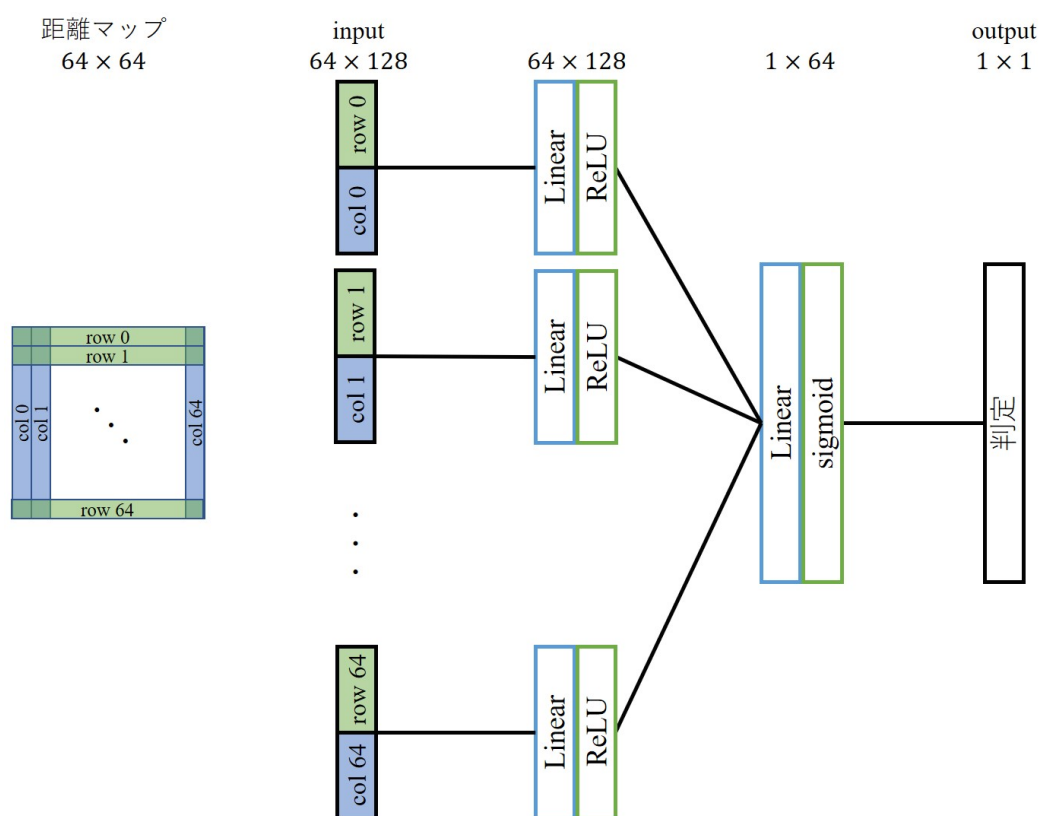


図 3.5: 非対角要素識別器のモデル: 本実験では 64 残基のタンパク質構造について, 生成モデルの構築を行った.

3.2 MDGANs2

MDGANs2 は図 3.6 に示すように生成器を 1 つ、識別器を 2 つ持つネットワークモデルとした。生成器は MDGANs1 と同じく図 3.2 で示すモデルを通して、ノイズから距離マップを生成する。

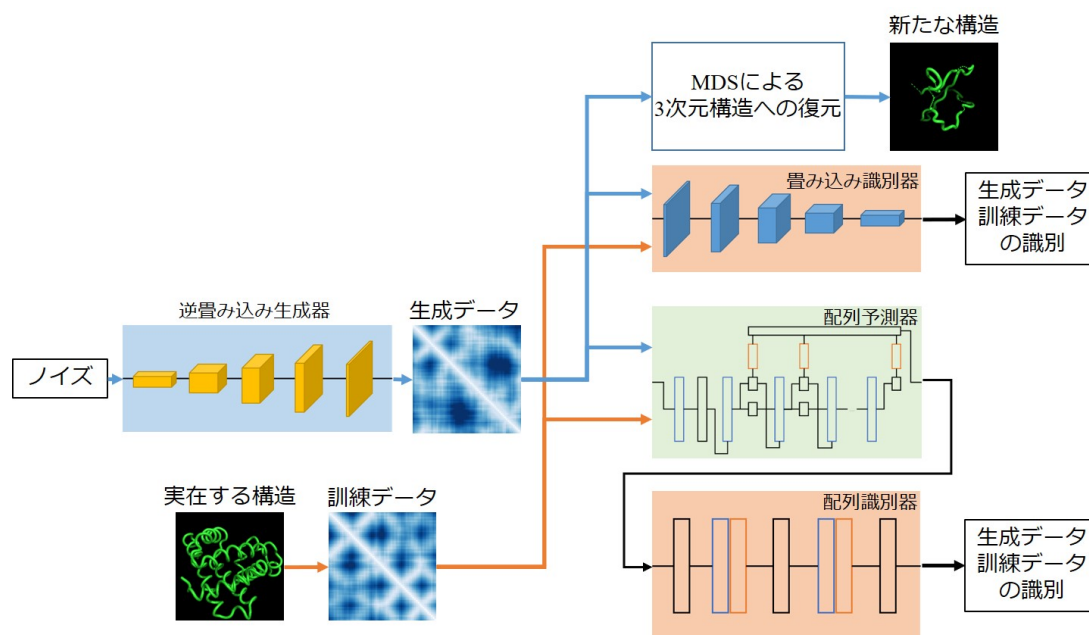


図 3.6: MDGANs2 のデータの流れ

配列予測器は距離マップを入力とし、アミノ酸配列を出力する (図 3.7). MDGANs1 と同様に残基数は $n = 64$ としている。距離マップを入力とし、18 層の残差ネットワーク [5] を用いて特徴抽出を行う。抽出された特徴を LSTMCell に入力し、64 残基分のアミノ酸を予測させる。予測するアミノ酸クラスは入力記号、終了記号、基本的なアミノ酸以外と 20 種類のアミノ酸の計 23 個であり、LSTM はそれぞれのクラスの確率を出力する。したがって、距離マップの入力に対して 23 クラスの確率が 64 残基分出力される。

アミノ酸配列識別器を図 3.8 に示す。配列予測器からの出力がネットワークへの入力となり、2 層のネットワークを経て識別器として機能する。1 層目では 1 残基ごとの入力 23 次元を入力として 1 次元の特徴を出力する。1 つのアミノ酸データ (64 残基) の入力に対して、64 次元の出力が得られる。2 層目では 1 層目から得られた 64 次元の特徴を入力として、距離マップから予測されたアミノ酸配列が、生成器で作られたものか、データセットから作成されたものかを分類する。

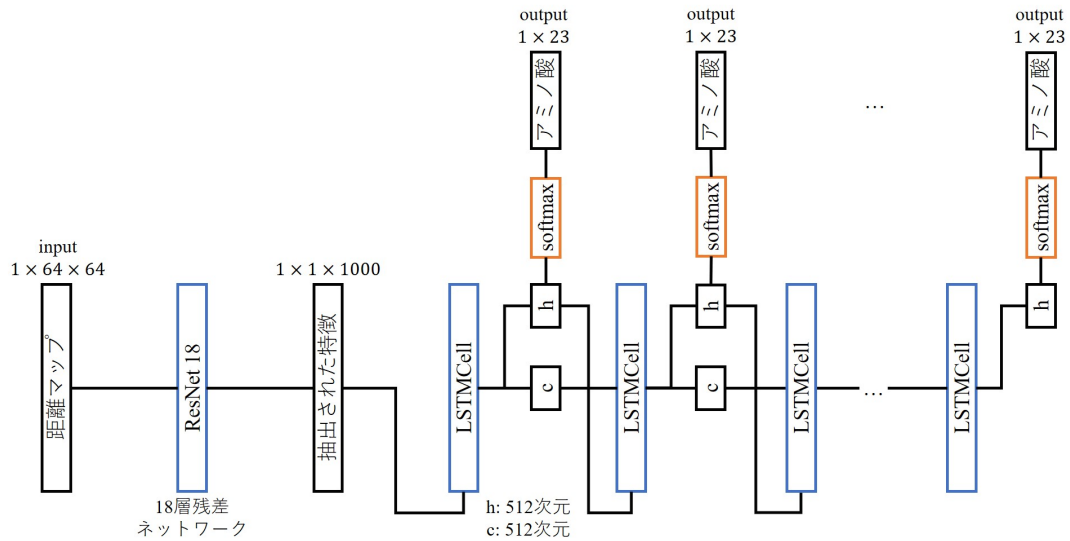


図 3.7: 配列予測器のデータの流れ: ResNet は He らの論文 [5] で提案されている 18 層の残差ネットワークである. LSTM の出力である c はセル状態, h は隠れ層であり, h を softmax 関数で各アミノ酸の確率を算出する.

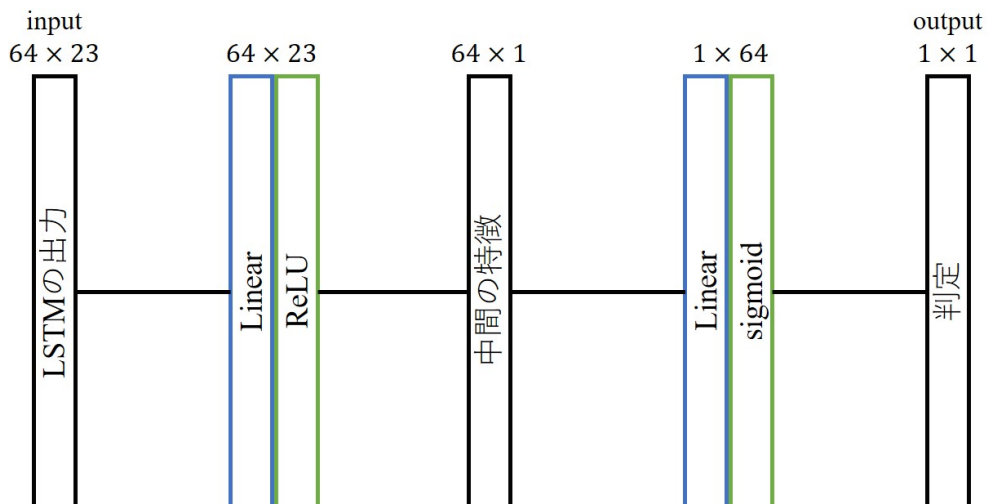


図 3.8: 配列識別器のデータの流れ

3.3 評価手法

生成された距離マップの2次元配列は Multidimensional Scaling (MDS) によって C_α の3次元構造に変換される。MDS は変換された C_α 座標間の距離と距離マップの値との誤差（ストレス値） σ を最小化するための反復的計算手法として用いる。

$$\sigma = \sum_{i < j} (d_{(i,j)} - \|r_i - r_j\|)^2 \quad (3.1)$$

ストレス値が小さいほど距離マップから正確に3次元構造を復元が可能できている。本研究の評価には3次元復元における精度として SMACOF アルゴリズムの最適化を最大30,000回実行することで得られた最小のストレス値を用いる。学習過程での復元精度を図るため、MDGANs が500 epoch 学習するごとに1,000個の距離マップを生成し、この距離マップの平均ストレス値を求める。

第 4 章

データセット

本章では、研究で使用するデータセットの詳細を示す。

4.1 GANs 用の学習データ

本研究では Evolutionary Classification of Protein Domains: ECOD (develop210) [12, 13] の配列一致度 99 % 代表構造セットをもとに作成されたデータセットを学習して生成を行う。残基数が 64 以上のタンパク質構造において、主鎖片側の末端のアミノ基である N 末端から 64 残基を切り取った構造 39,514 個を GAN 訓練データとし、 C_{α} 原子間の距離マップ (図 4.1) を作成した。本研究では画像生成モデルを元に作成したモデルを使用し、十分なデータセット数を必要とするため、データ数が十分に用意できる 64 残基のデータセットを作成した。

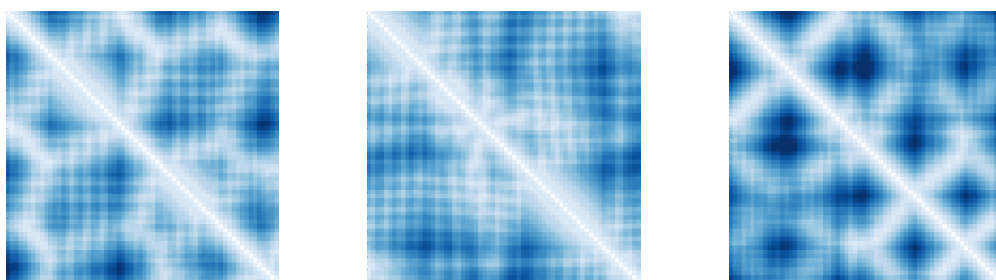


図 4.1: データセットから生成された距離マップの一部

4.2 アミノ酸配列データ

また、アミノ酸配列データも GANs で使用する 64 残基からデータの作成を行った。図 4.2 にアミノ酸配列データの作成手順を示す。1 残基に対して 23 クラスの分類がなされており、20 種類のアミノ酸、20 種類に含まれないその他のアミノ酸が 1 種類、LSTM による生成のための開始フラグと終了フラグの 2 種類で構成される。1 残基ごとに one-hot エンコードされた正解データとして作成されるため、アミノ酸配列データは 1 つのタンパク質あたり (64×23) 次元のデータとなる。

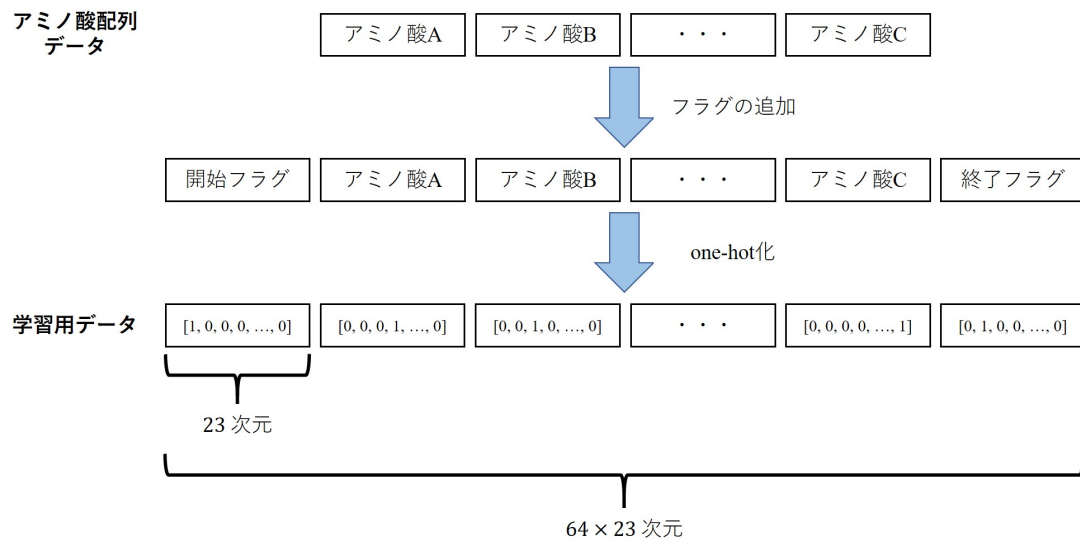


図 4.2: アミノ酸配列データの作成手順

第 5 章

計算機による実験

5.1 実験条件

実験での細かなパラメータ設定について以下に記述する.

5.1.1 学習に使用した計算機

学習には表 5.1 で示す構成の計算機を用いて学習を行った. 学習プログラムは python 言語によって pytorch プラットフォーム上で記述する.

表 5.1: 学習に使用した計算機の構成

OS	Ubuntu 20.04.1 LTS	
カーネル	5.4.0-60-generic	
CPU	AMD Ryzen Threadripper 3970X	32-Core Processor
GPU	GeForce RTX 3090	CUDA Version: 11.1
VRAM	24GB	
メモリ	256GB	

5.1.2 配列予測器の最適化

提案手法の MDGAN2 で使用した配列予測器は Chen らの提案した ResNet と LSTM を用いた予測手法 [14] を参考に, 本手法の一部として取り入れた. 距離マップからの特徴抽出のために ResNet18 層を用い, 抽出した特徴から LSTM を用いて配列を予測した. バッチサイズは 256, LSTM 内の隠れ層 512 で 3000 epoch 学習済みのネットワークを配列予測器として用いた. 本ネットワークの学習には GANs の学習データを用いており, 学

習データと検証データの割合は 7:3 である。

図 5.1 は配列予測器における epoch ごとの精度の推移を示している。図 5.2 は配列予測器における epoch ごとの損失の推移を示している。図 5.1, 図 5.2 の結果をみると学習データの精度が向上 (損失が減少) しているのに対し、テストデータの精度向上は頭打ちになっていることから明らかに過学習を起こしており、この予測器の精度を向上させる余地があると考えられる。

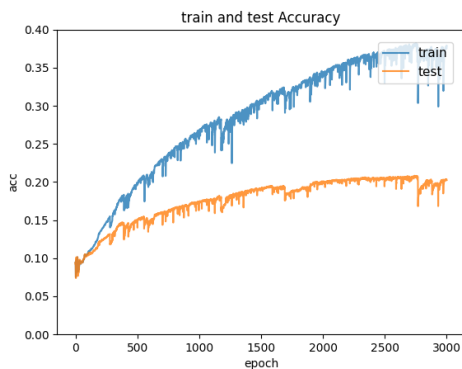


図 5.1: 配列予測器の精度推移

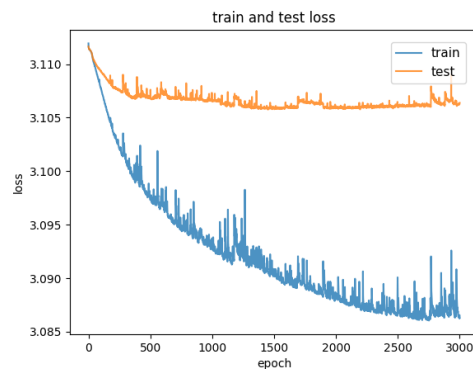


図 5.2: 配列予測器の損失推移

5.1.3 GANs の学習における条件

本研究では GANs によるタンパク質主鎖生成手法について、DCGANs のみを使用したもの、提案手法 MDGANs1, MDGANs2 の 3 種類の比較実験を行った。各ネットワークは 500epoch ごとにネットワークモデルを保存しつつ、5000epoch まで学習を行った。学習時のバッチサイズは 256 とし、39,514 個の 64 残基アミノ酸を用いた。MDGANs2 において使用された配列予測器には、バッチサイズ 256, LSTM 内の隠れ層 512 で 3000 epoch 学習済みのネットワークモデルを用い、GANs の学習過程では予測器内の重み更新は行わないものとする。

5.2 実験結果

図 5.3 に DCGANs によって生成された距離マップを示す。図 5.4 に MDGANs1 によって生成された距離マップを示す。図 5.5 に MDGANs2 によって生成された距離マップを示す。

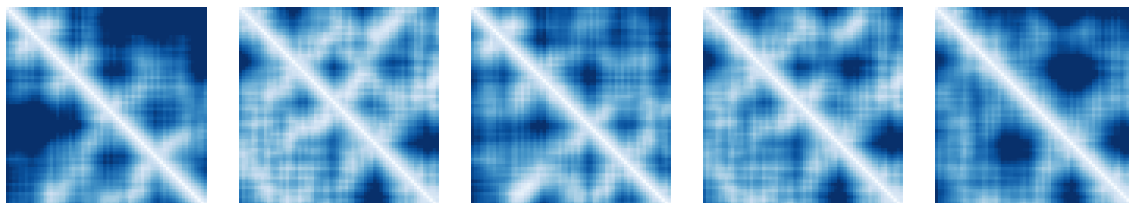


図 5.3: 学習済みの DCGANs モデルを用いて生成した距離マップ

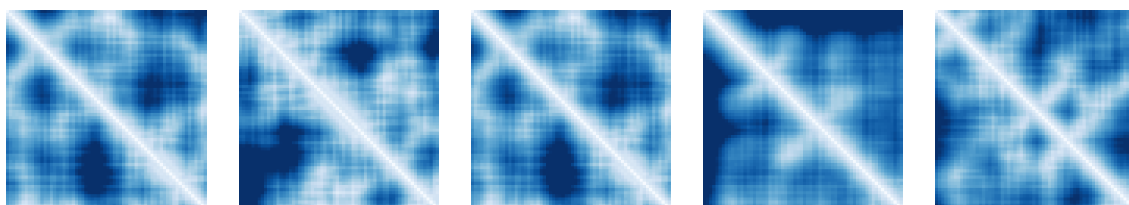


図 5.4: 学習済みの MDGANs1 モデルを用いて生成した距離マップ

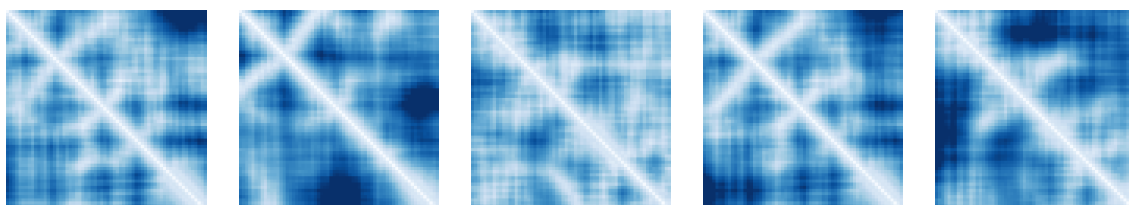


図 5.5: 学習済みの MDGANs2 モデルを用いて生成した距離マップ

5.3 実験考察

実験の結果から，学習時の MDS のストレス平均値の推移，距離マップの対角要素誤差の推移，距離マップ対称性の推移について考察を行った．5000 epoch までの学習過程において各 500 epoch ごとの生成モデルに対して距離マップを 1000 個生成した．各距離マップに対して MDS による最適化を行った後に，最終的なストレス値を求め，平均値を計算した．

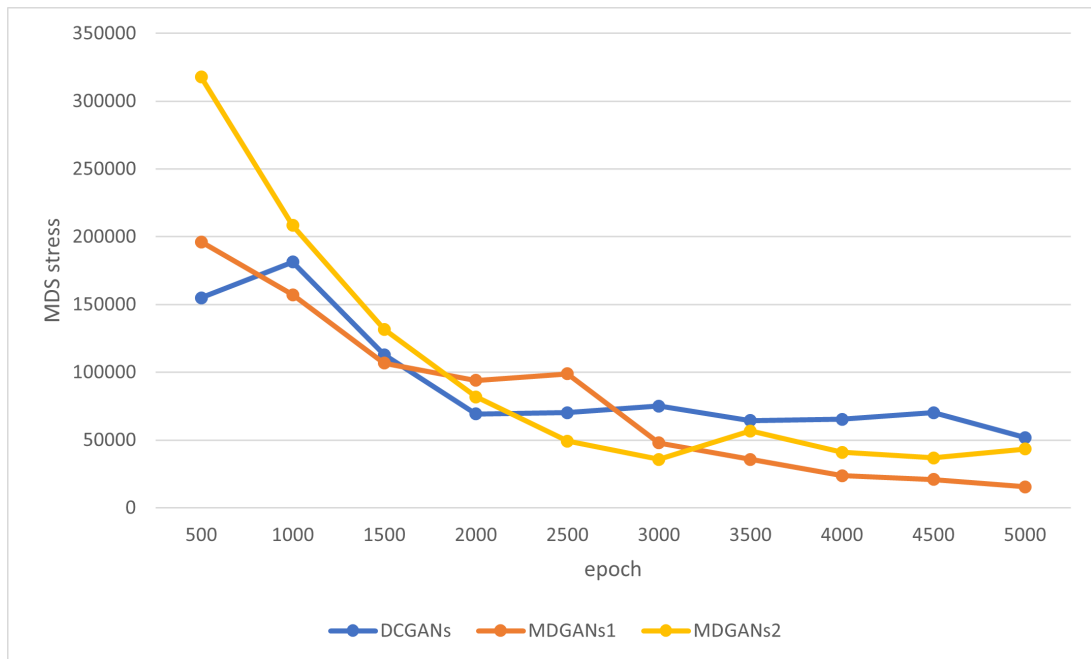


図 5.6: 各ネットワークモデルにおける学習時の MDS のストレス平均値の推移

図 5.6 に学習時の MDS のストレス平均値の推移を示す．最小のストレス値を得たのは MDGANs1 であり，その値は $15460 \pm 197\text{\AA}$ であった．すべてのモデルでストレス値の減少があり，epoch によってはやや上下することもあることがわかる．

図 5.7 は距離マップ対角要素誤差の推移について示したものである．誤差 E_1 は式 (5.1) で与えられ， d_{ii} は距離マップにおける i 行 i 列の要素を指す．

$$E_1 = \sum_{i=1}^{64} d_{ii}^2 \quad (5.1)$$

5000 epoch 学習後の MDGANs1 は対角要素誤差が平均で 4.35 となった．DCGANs は学習過程の中で誤差が上下しており，畳み込み識別器のみでは，行列の特徴の一部を学習

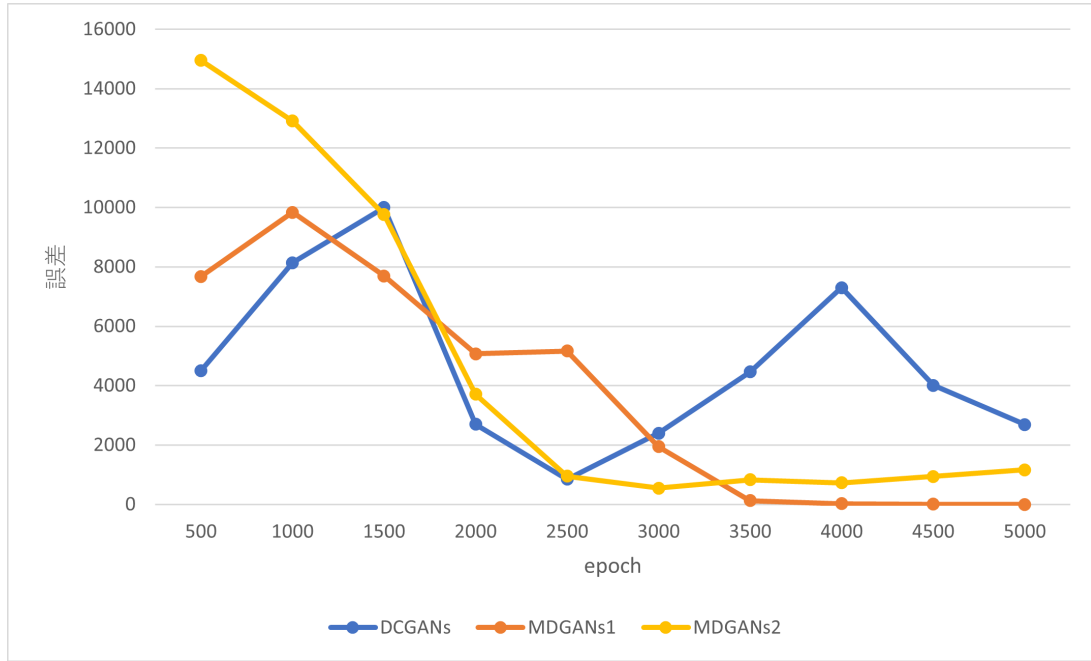


図 5.7: 各ネットワークモデルにおける学習時の距離マップ対角要素誤差の推移

に十分反映できていないと考えられる．MDGANs1 では対角線上の値が 0 となることを学習していると考えられる．

図 5.8 は学習時の距離マップの対称性の推移である．誤差 E_2 は式 (5.2) で与えられ， d_{ij} は距離マップにおける i 行 j 列の要素を指す．

$$E_2 = \sum_{i=1}^{64} \sum_{j=1}^{64} (d_{ij} - d_{ji})^2 \quad (5.2)$$

どのネットワークモデルも誤差が減少しているが，大きく差は出なかった．MDGANs1 は非対角要素識別器をマップの対称性を学習させるべく追加したが，十分な学習精度は得られなかった．原因はネットワークモデルが小さく，過学習を引き起こしたのではないかと考えられる．

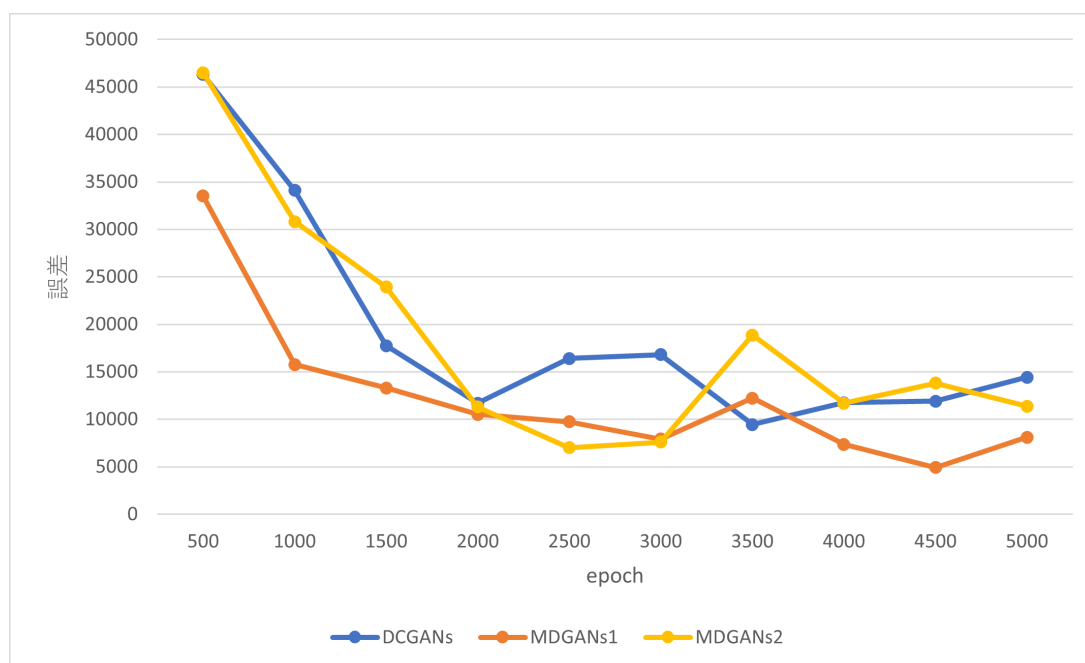


図 5.8: 各ネットワークモデルにおける学習時の距離マップ対称性の推移

第 6 章

結言

6.1 まとめ

本研究ではタンパク質主鎖生成器として、畳み込み識別器，対角要素識別器，非対角要素識別器を組み込んだ MDGANs1，および畳み込み識別器，配列予測器を組み込んだ MDGANs2 を提案した．先行研究の手法をもとに作成した DCGANs と MDGANs1，MDGANs2 を実装し，学習過程における MDS のストレス値，距離マップ対角要素並びに対称性について考察を行った．実験の結果から MDGANs1 で最小の MDS ストレス値 ($15460 \pm 197\text{\AA}$) が得られた．主鎖自動生成について MDGANs の有効性が示された．

6.2 今後の展望

本研究で実装した MDGANs1 に配列識別器を加えた新たなネットワークを構成することで生成される距離マップの精度向上につながることが期待される．

MDGANs による手法によって生成器の学習に条件付けを行うことができたことから，距離マップとして満たすべき要件についてそれぞれ識別器を用意することで，正確に距離マップを学習できると考える．GANs の学習においてデータ数を増やすにより，安定した学習を得られるとも考えられる．

また，十分な精度を持つ距離マップの生成が可能になることで，主鎖生成技術を活用したデノボデザインによる新たなタンパク質の設計が可能となり，医薬品の開発に活用できると考えられる．

付録 A

ソースプログラム等のデータ

A.1 プログラム

本研究に関するプログラムはすべて以下のディレクトリ

`/home/Shimizu/ImageCaptioning/GenerateProtein3`

以下のリンクはリモートリポジトリ

<http://portal.hi.info.mie-u.ac.jp/gitbucket/shimizu/GenerateProtein3>

これまでの研究のリモートリポジトリ

<http://portal.hi.info.mie-u.ac.jp/gitbucket/shimizu/GenerateProtein2>

<http://portal.hi.info.mie-u.ac.jp/gitbucket/shimizu/MakeProteinMovie>

A.2 実験データ

実験データは PDB 形式のテキストフォーマットで以下のディレクトリに保存されている。

`/home/Shimizu/ImageCaptioning/GenerateProtein3/data`

データは約 9.2GB あるためリモートリポジトリには置かれていません。

A.3 環境構築情報

プログラムの環境構築は

`/home/Shimizu/ImageCaptioning/GenerateProtein3`

以下に存在する README.md に記載されている方法で行う。

python3.7 の venv 環境を利用している。

A.4 プログラムの詳細

`/home/Shimizu/ImageCaptioning/GenerateProtein3/README.md`

詳しい実行方法は README.md を参照してください.

付録 B

発表資料

本付録では 2021 年 2 月 15 - 16 日に実施された修士論文発表の資料を載せる。

B.1 修士論文発表資料

Human Interface Lab. @Me University



Multi-discriminator generative adversarial networks を用いた
タンパク質主鎖構造デザイン

三重大学大学院 工学研究科 情報工学専攻
ヒューマンインターフェース研究室
419MS10 清水一希

2021/02/15 修士論文発表

●○○○ Introduction Human Interface Lab. @Me University

研究背景 Introduction

●○○○ Introduction Human Interface Lab. @Me University

タンパク質の機能と構造

▶生命におけるタンパク質

- タンパク質は生命現象を担う「素子」であり生体内でのあらゆる機能に関連する。
- 機能発現にはタンパク質の形状が重要な役割を果たす。



ミオグロビンの3次元構造

▶タンパク質の化学的構造

- タンパク質は20種類のアミノ酸残基が繋がったもの。
- アミノ酸を繋げる順番 (アミノ酸配列) が決まれば人工的にタンパク質を製造できる。
- 1本の鎖状に繋がった高分子が折り畳まれ、3次元構造を持つ。
- 主鎖と側鎖によって構成される。



タンパク質構造の一部

●○○○ Introduction Human Interface Lab. @Me University

タンパク質構造の数値表現 (距離マップ)

▶距離マップ

- 各 C_{α} 原子間に対して3次元空間上の距離を計算した行列。
- 行列として構造を扱うことができる。
- 3次元構造の回転や並進に対して不変。



	$C_{\alpha 1}$	$C_{\alpha 2}$	$C_{\alpha 3}$	$C_{\alpha 4}$
$C_{\alpha 1}$	0	3.8	5.6	25.5
$C_{\alpha 2}$	3.8	0	3.8	26.1
$C_{\alpha 3}$	5.6	3.8	0	20.8
$C_{\alpha 4}$	25.5	26.1	20.8	0

距離マップ



可視化した距離マップ

●○○○ Introduction Human Interface Lab. @Mei University

タンパク質構造のデザイン

▶タンパク質構造をデザインする研究分野

- タンパク質主鎖の3次元構造のみからアミノ酸配列を設計する技術が確立しつつある^{1,2}.

▶タンパク質主鎖構造のデザイン

- Anandらの生成的深層学習モデルを利用した手法³
 - タンパク質の距離マップを画像と見立てて、画像生成モデルを用いてタンパク質主鎖を部分的に生成する研究。

実在するタンパク質

¹Kathleen et al. "Design of a novel globular protein fold with atom-level accuracy" Science 353, 294 (2015) 1384-1388.

²Koga et al. "Predicting the changing chiral protein structure" Nature 453, 742 (2015) 323-327.

³Anand et al. "Generative Modeling For Protein Structure", Proceedings of the 34th International Conference on Neural Information Processing Systems 2016.

5

●○○○ Introduction Human Interface Lab. @Mei University

画像生成モデルのタンパク質主鎖構造生成への応用

▶Deep Convolutional Generative Adversarial Networks (DCGANs)⁴

- 生成器と識別器を持つ教師なし学習モデル。
- 学習は生成器と識別器を相互に学習する。(mはバッチサイズ)

$$\text{生成器}_{loss} = -\frac{1}{m} \sum_{i=1}^m \log(1 - D(G(x^{(i)})))$$

$$\text{識別器}_{loss} = -\frac{1}{m} \sum_{i=1}^m [\log D(x^{(i)}) + \log(1 - D(G(x^{(i)})))]$$

- loss関数の推移で学習度を測ることは一般的に難しい。

⁴Radford et al. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks" Proc. ICML, pp. 1-16, 2016.

6

●○○○ Introduction Human Interface Lab. @Mei University

先行研究の問題点と本研究の目的

▶Anandらの手法の問題点

- 正しく修復されたものもあったが、復元した構造が元の構造と異なるものも出現していた。
- 距離マップからの畳み込み特徴のみを利用した生成は難しい。

▶本研究の目的

- 先行研究の手法を改良し、距離マップ生成の性能を向上させる。
- 距離マップ生成モデルの学習度を測る方法及び生成された距離マップの評価手法を確立する。

▶改良の道筋

- HardyらのMulti-Discriminator Generative Adversarial Networks (MDGANs) についての研究⁵
 - 識別器を追加することによりGANの性能を向上させることが示された。
- 本研究ではMDGANsの識別器を追加するという考え方を応用する。

⁵Hardy et al. "3D-gan: Multi-discriminator generative adversarial networks for distributed datasets", 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 1225, 2019.

7

●○○○ Methods Human Interface Lab. @Mei University

提案手法

Methods

MDGANs1 と MDGANs2
学習度の評価手法

8

●○○○ Methods Human Interface Lab. @Mei University

MDGANs1: 行列要素に対する識別器を追加した手法

▶2つの識別器を追加したモデル

- 生成される距離マップが数学的な制約を満たしているかを新たに導入した2つの識別器で判別する。

$$\text{生成器}_{loss} = -\frac{1}{m} \sum_{i=1}^m \log(1 - D_1(G(x^{(i)})))$$

$$\text{識別器}_{loss} = -\frac{1}{m} \sum_{i=1}^m [\log D_1(x^{(i)}) + \log(1 - D_2(G(x^{(i)})))]$$

実在する構造

9

●○○○ Methods Human Interface Lab. @Mei University

行列要素に対する2つの識別器

制約1: 距離マップは対称行列

▶非対角要素識別器

- 生成される距離マップはそのままでは制約1を満たさない。
- 距離マップで一致すべき行と列の組を繋げて入力する。

制約2: 距離マップの対角要素は0

▶対角要素識別器

- 生成される距離マップはそのままでは制約2を満たさない。
- 距離マップの対角要素を入力する。

10

●○○○ Methods Human Interface Lab. @Mei University

MDGANs2: アミノ酸配列識別器を追加した手法

▶アミノ酸配列予測器とアミノ酸配列識別器を追加したモデル

- 距離マップのデータをアミノ酸配列のデータに変換した後識別する方法。
- アミノ酸配列予測器は事前に学習済みのモデルを用いる。

実在する構造

11

●○○○ Methods Human Interface Lab. @Mei University

距離マップの評価手法

▶生成された距離マップを3次元構造へ復元する手法として Multidimensional Scaling (MDS)⁶を用いる。

距離マップ

MDSによる3次元構造への復元

3次元構造

- MDSでは以下の式で定義されるストレス値 σ を反復計算により最小化することで3次元座標 r_i ($0 \leq i, j < 64$)を求める。

$$\sigma = \sum_{i,j} (d_{ij} - \|r_i - r_j\|)^2$$

(d_{ij} は距離マップにおける(行)列の要素)

- ストレス値 σ は距離マップの3次元座標への復元可能性の指標となる。

▶ストレス値 σ は距離マップ生成モデルの学習度を測る指標になり、生成された距離マップの評価手法として利用できる。

⁶"Nonlinear Multidimensional Scaling: Theory and Applications" Beng. L. Geman P. Springer Series in Statistics (1997)

12

実験結果

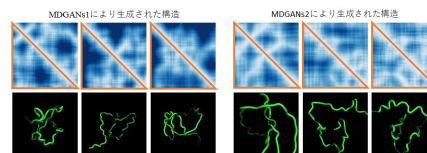
Result

13

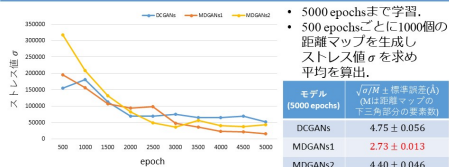
生成された距離マップと3次元構造

➤ 5000 epochs 学習済みのMDGANs1, MDGANs2を用いて距離マップ生成し, MDSにより3次元座標に復元を行った.

- 各距離マップの下三角要素(オレンジで囲まれた部分)を用いてMDSにより3次元構造に復元した構造.



14

学習時のストレス値 σ の推移

- 5000 epochsまで学習.
 - 500 epochsごとに1000個の距離マップを生成しストレス値 σ を求め平均を算出.
- 最小のストレス値 σ を得たのはMDGANs1でありその値は $15460 \pm 197(\text{\AA}^2)$ であった.
- 距離マップの1要素に換算し平方根をとると $\sqrt{\sigma/M} = 2.73 \pm 0.013(\text{\AA})$.
- ストレス値 σ が距離マップ生成モデルの学習度を測る良い指標となっている.

15

まとめ

➤ 提案手法

- 畳み込み識別器, 対角要素識別器, 非対角要素識別器を組み込んだMDGANs1を提案した.
- アミノ酸畳み込み識別器, アミノ酸配列予測器を組み込んだMDGANs2を提案した.

➤ 実験結果

- MDGANs1において既存の手法DCGANsよりも小さいストレス値 $\sigma = 15460 \pm 197(\text{\AA}^2)$ を得た.
- 識別器を新たに追加することで距離マップ生成の性能が向上した.
- ストレス値 σ が距離マップ生成モデルの学習度を測る指標として利用できることが示された.

➤ 今後の展望

- 学習データセットの拡充により, 精度向上を目指す.

16

謝辞

本論文の執筆及び，研究生活において多くの方々からご支援を頂きました．多くの御助言や御指導をしていただいた若林哲史教授，研究の細かな御指導とや多くの相談に乗っていただいた白井伸宙助教，専門的な最新の情報技術について詳しい御指導していただいた盛田健人助教，タンパク質の研究分野に関して多くのアドバイスやアイデアを頂いた佐久間航也氏，研究の基礎知識について御指導していただいた三宅康二名誉教授，また，日々の研究活動において様々な場面でお世話をしていただいた吉永みゆき事務官に深く感謝いたします．

参考文献

- [1] Silva, D. A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labao-Almeida, C., Ali, L. R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., Biary, T., Crowley, S. J., Marcos, E., Walkey, C. D., Weitzner, B. D., Pardo-Avila, F., Castellanos, J., Carter, L., Stewart, L., Riddell, S. R., Pepper, M., Bernardes, G. J. L., Dougan, M., Garcia, K. C., Baker, D. "De novo design of potent and selective mimics of IL-2 and IL-15." *Nature* 565.7738 (2019): 186-191.
- [2] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, David Baker. "Principles for designing ideal protein structures." *Nature* 491.7423 (2012): 222-227.
- [3] Namrata Anand, Possu Huang. "Generative Modeling For Protein Structures" , Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018.
- [4] LeCun Yann, Leon Bottou, Yoshua Bengio, Patrick Haffner. "Gradient-based learning applied to document recognition", *Proceedings of the IEEE* 86.11 (1998): 2278-2324.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [6] Andrew L. Maas, Awni Y. Hannun, Andrew Y. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models", *Proc. icml*. Vol. 30. No. 1. 2013.
- [7] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams. "Learning representations by back-propagating errors", *Nature* 323,pp.533-536, (1986)
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. "Generative Adversarial Nets" *Proc. NIPS2014*, pp. 2672–2680 2014.
- [9] Alec Radford, Luke Metz, Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks" *Proc. ICLR*, pp. 1-16,

- 2016.
- [10] Hardy, Corentin, Erwan Le Merrer, and Bruno Sericola. "Md-gan: Multi-discriminator generative adversarial networks for distributed datasets." 2019 IEEE international parallel and distributed processing symposium (IPDPS). IEEE, 2019.
 - [11] Borg, I.; Groenen P., Modern Multidimensional Scaling - Theory and Applications Springer Series in Statistics (1997).
 - [12] Hua Cheng ,R. Dustin Schaeffer ,Yuxing Liao ,Lisa N. Kinch,Jimin Pei,Shuoyong Shi,Bong-Hyun Kim,Nick V. Grishin. "ECOD: an evolutionary classification of protein domains." PLoS Comput Biol 10.12 (2014): e1003926.
 - [13] Hua Cheng Yuxing Liao R. Dustin Schaeffer Nick V. Grishin. "Manual classification strategies in the ECOD database." Proteins: Structure, Function, and Bioinformatics 83.7 (2015): 1238-1251.
 - [14] Sheng Chen, Zhe Sun, Lihua Lin, Zifeng Liu, Xun Liu, Yutian Chong, Yutong Lu, Huiying Zhao, and Yuedong Yang. "To improve protein sequence profile prediction through image captioning on pairwise residue distance map." Journal of chemical information and modeling 60.1 (2019): 391-399.
 - [15] 藤 博幸 (2015). タンパク質の立体構造入門講談社
 - [16] 斎藤 康記 (2018). ゼロから作る Deep Learning - Python で学ぶディープラーニングの理論と実装株式会社オライリー・ジャパン
 - [17] 斎藤 康記 (2018). ゼロから作る Deep Learning 2 - 自然言語処理編株式会社オライリー・ジャパン