令和２年度　博士論文

機械学習を用いた天然変性領域中の機能部位予測法の研究

環境・生命工学専攻

1856501 安保　勲人

目次

第１章　序論

1.1　研究背景

我々の体内では様々な生命活動が営まれており，タンパク質はそれらと密接に関わっている．タンパク質がうまく機能しなければ正常な生命活動を維持することができず疾患などを引き起こす．タンパク質はアミノ酸が重合し数珠状に連なった高分子化合物である．この数珠状の分子が規則的に折りたたまれ，特定の立体構造を作ることで生体内において機能を発揮する(図 1A)．長年タンパク質にとってこのような立体構造形成は必要不可欠のイベントであると考えられてきた．しかし近年，この考えに当てはまらないタンパク質の存在が明らかとなってきた[1]．このタンパク質は天然変性タンパク質と呼ばれ生理的条件下であっても単独では立体構造を形成しない天然変性領域を保持している[1，2](図 1B)．

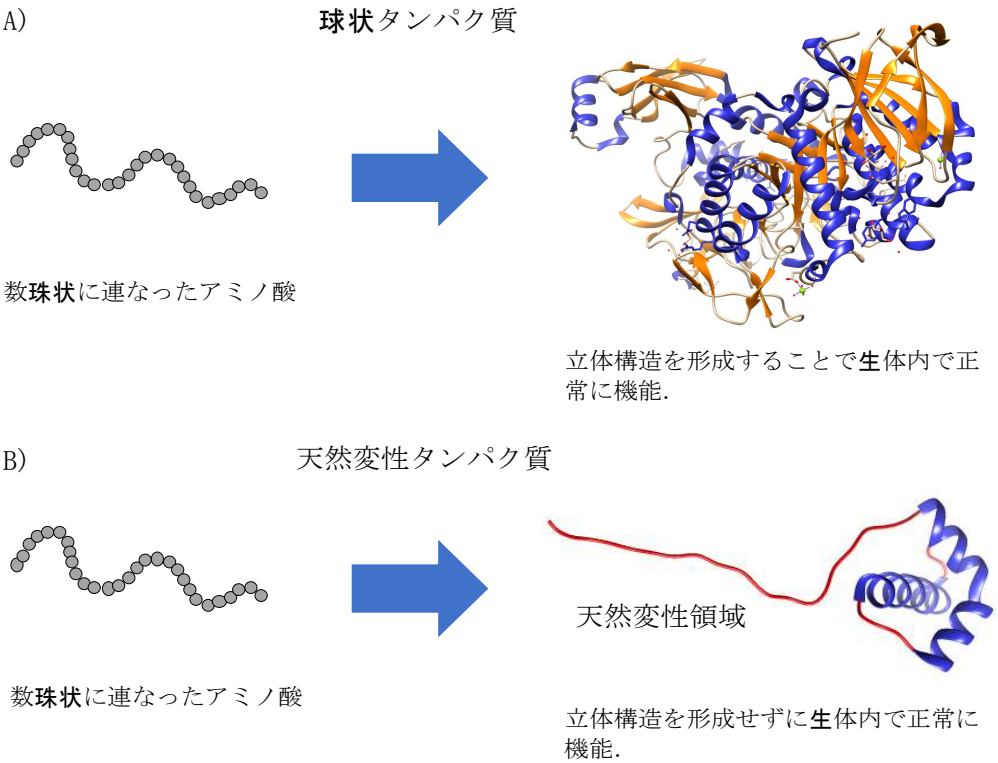


図 1．球状タンパク質(A)と天然変性タンパク質(B)．B の赤いひも状部分が天然変性領域を表している．

天然変性領域は複合体を形成するタンパク質を繋ぎ止めて反応させるテザリング機能，リン酸化およびユビキチン化などの翻訳語修飾部位の提示，および構造ドメイン間を繋ぐドメインリンカーなどの様々な機能を保持している(図 2)．

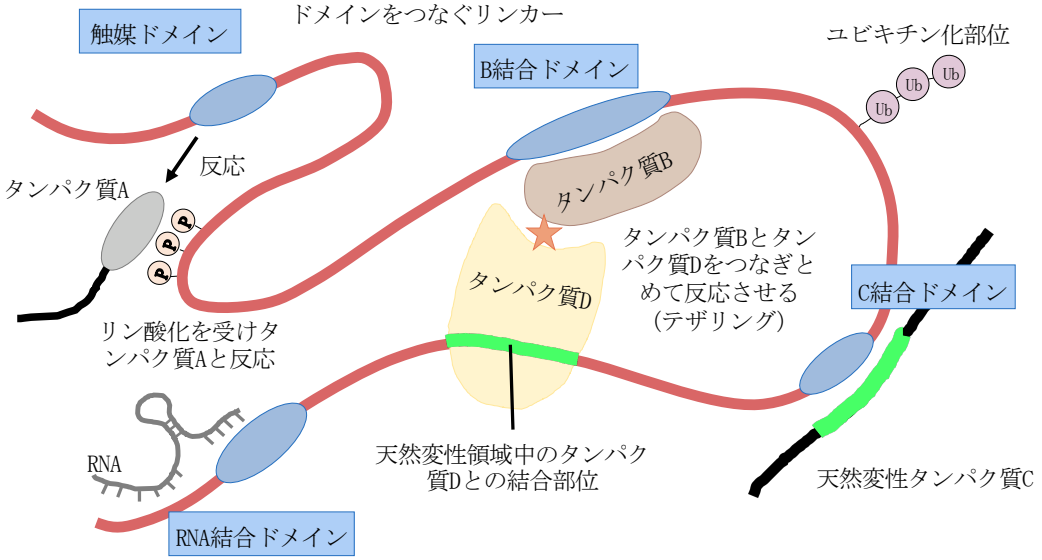


図 2. 天然変性領域の役割．緑のひも状部分は機能部位，赤いひも状部分は天然変性領域，青い楕円は構造領域を示している．実験医学 Vol. 37 No. 18(11 月号)2019 図 1 より改編．

天然変性領域はコンピュータを用いたバイオインフォマティクスの技術によってアミノ酸配列から予測が可能である[3-19]．コンピュータ技術の進歩やゲノム解析の進歩によるデータ数の増加によって，天然変性タンパク質および天然変性領域は予測精度の評価指数である Matthews Correlation Coefficient(第 3 章参照)において 0.5～0.6 という高い精度での予測が可能である．この精度は，長い研究の歴史のある 2 次構造予測と同程度であり，実験を行う研究者が予測結果をある程度信頼し実験を進めることができるだけの実用精度を達成していると言える．この実用精度を達成した予測モデルによって，天然変性領域は真核生物の核タンパク質に多く見られること[20-22]や，リン酸化などの翻訳語修飾を受けるアミノ酸が多く含まれることなど，天然変性タンパク質の性質が明らかになってきた[22]．またタンパク質間相互作用ネットワークにおいて多くの相互作用パートナーを持つハブタンパク質には天然変性タンパク質が多く含まれていることも報告されている[23, 24]．また，天然変性タンパク質は疾患に関わることが知られており，乳がんに関わる BRCA1[25]，アルツハ

イマー病に関わる Tau タンパク質[26]，パーキンソン病に関わるα-synuclein[27]および筋萎縮性側索硬化症に関わる Fus タンパク質[28]などは疾病に関わる天然変性タンパク質として知られている．さらに，天然変性タンパク質は液―液相分離(liquid-liquid phase separation: LLPS)と呼ばれる，膜のないオルガネラに深く関わっていることも報告されている[29-32]．LLPS は細胞内の反応の場として機能しており，多くの生物学的プロセスや疾患に関連している．Fus タンパク質は LLPS を形成する天然変性タンパク質として，近年さかんに研究が行われている天然変性タンパク質である．

天然変性タンパク質のユニークな特徴の１つとして，天然変性領域中に相互作用相手であるパートナータンパク質や，DNA および RNA などの生体分子と結合するための機能部位を保持することが挙げられる．機能部位は一般的に数残基から数十残基のとても短い領域であり，パートナーと出会うと局所的な立体構造を形成し結合する(図 3)．

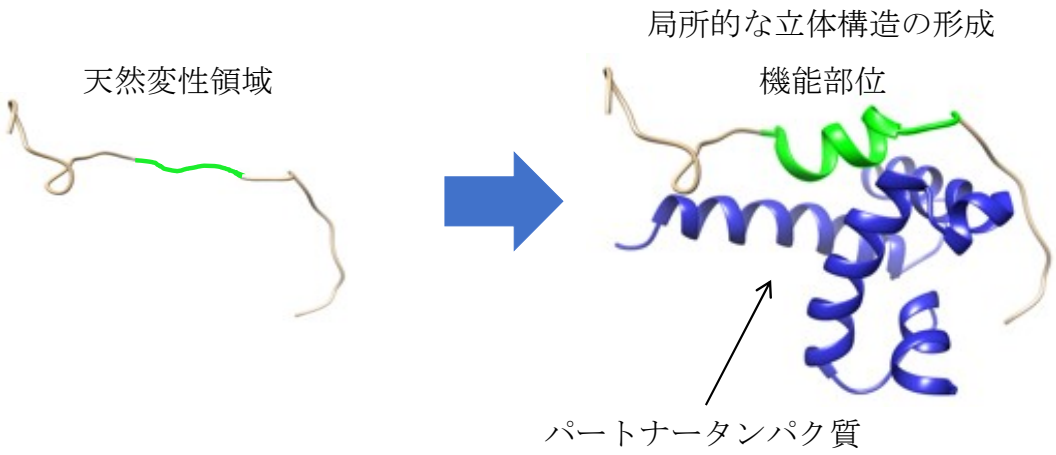


図 3. 天然変性領域中の機能部位.

この現象は結合と共役した構造形成(coupled folding and binding)と呼ばれており[33, 34], このメカニズムを介して天然変性タンパク質はシグナル伝達や転写調整などの様々な重要な生命現象に関与している[2, 35-37]．また，パートナーとの結合において局所的な立体構造の形成を必要としない機能部位[38, 39]や，相互作用パートナーも天然変性領域である例[39, 40]も報告されており機能部位は結合の際の構造や

パートナーの構造的特徴などにおいて多様性を持つ.

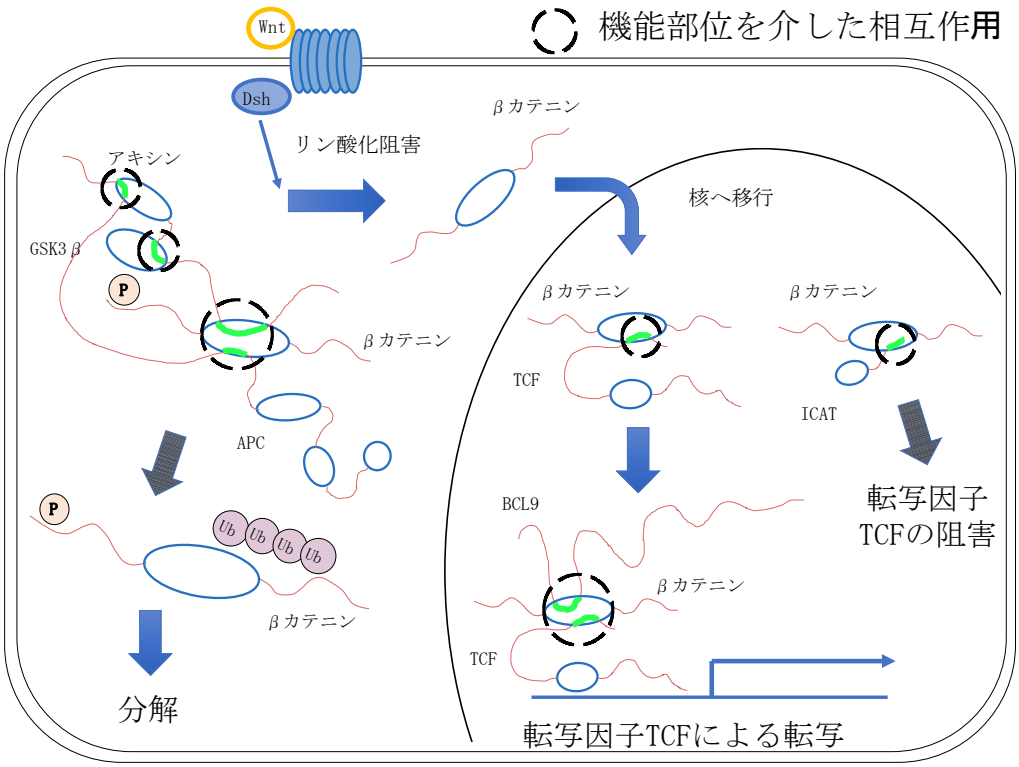


図 4. 機能部位を介した Wnt シグナル経路. 緑のひも状部分は機能部位, 赤いひも状部分は天然変性領域, 青い楕円は構造領域を示している. 細胞工学 Vol.33 No.7 2014 図 2 より改編.

図 4 は天然変性領域中の機能部位を介した生物学的プロセスである Wnt シグナル経路である. Wnt シグナル経路は発生や形態形成に関与するシグナル伝達経路であり, 分泌タンパク質 Wnt が細胞膜上の受容体に結合することで経路のスイッチがオンになる. βカテニンは Wnt 経路において中心的な役割を果たす天然変性タンパク質である. 経路のスイッチがオンの時βカテニンは核移行し, 転写因子 TCF および BCL9 の天然変性領域中の機能部位と相互作用することで TCF による転写が行われる. 天然変性タンパク質 ICAT はβカテニンの TCF との相互作用部位に結合することで TCF とβカテニンの相互作用を阻害する. Wnt が存在せず経路のスイッチがオフの時は, βカテニンはアキシン, GSK3βおよび APC と複合体を形成する. 複合体形成には APC およびアキシンの長い天然変性領域中の機能部位を介した相互作用が重要な働きを担っている. 複合体中でβカテニンは GSK3βによってリン酸化を受ける. リン酸化された

βカテニンはユビキチン化を受け分解される．複合体形成に関わる天然変性領域中の機能部位に変異などの異常が発生し，相互作用ができないとβカテニンは分解されず細胞内に蓄積され膵臓ガン，大腸ガン，胃ガンおよび前立腺ガンなどのガンを誘発する[41](図 5)．

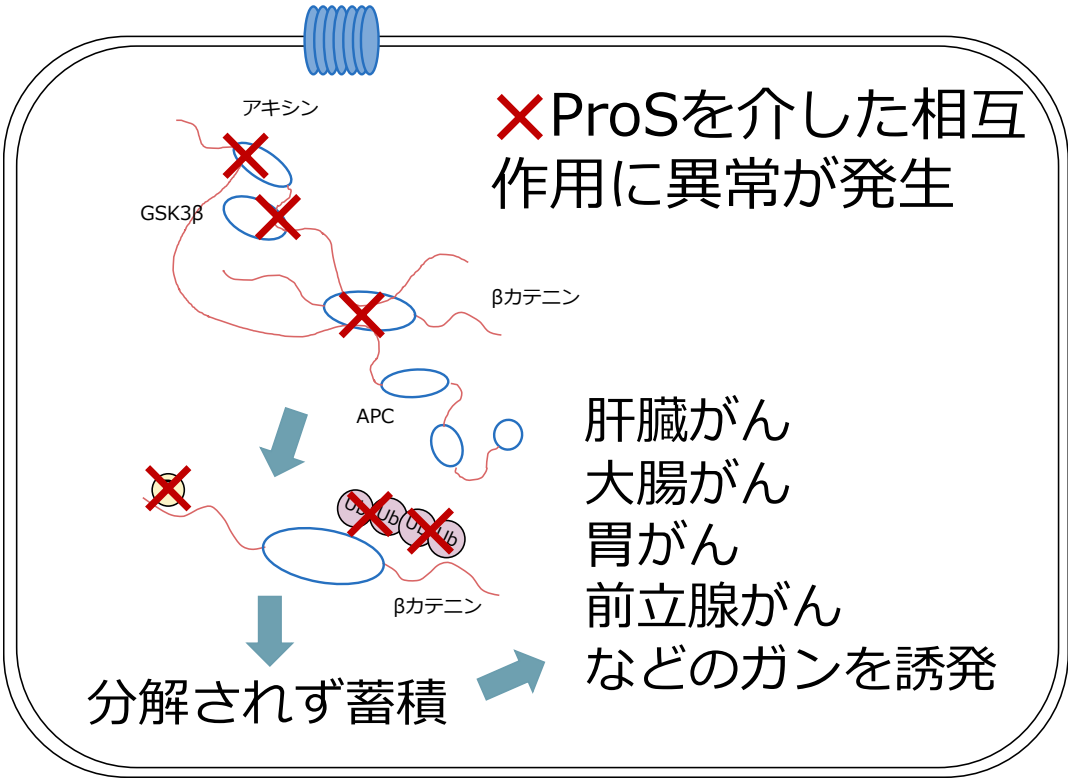


図 5. Wnt シグナル経路の異常と疾病．赤いひも状部分は天然変性領域，青い楕円は構造領域を示している．

上記のように機能部位は生物学的プロセスにおいて重要な役割を果たしており，変異などによる機能部位異常は疾病を引き起こす．現在，機能部位データを収録しているデータベースは存在しており，機能部位データを提供している．各データベースにおいて機能部位は”molecular recognition features” (MORFs) [42], “short linear motifs” (SLiMs) [43], および “disordered binding sites” (DIBSs) [44] などと呼ばれている．また，天然変性タンパク質データベース IDEAL[45, 46]では，実験的に検証された機能部位を”protean segments”(変幻自在な部位) (ProSs)とし，提供している．IDEAL には 146,276 残基の構造領域，33,053 残基の天然変性領域および 9,444 残基の ProS 残基が収録されているが，ProS のデー

タ数は構造領域や天然変性領域のデータ数と比較して少ない(図 6)．そのため計算機を用いて機能部位を予測する技術が必要とされている．

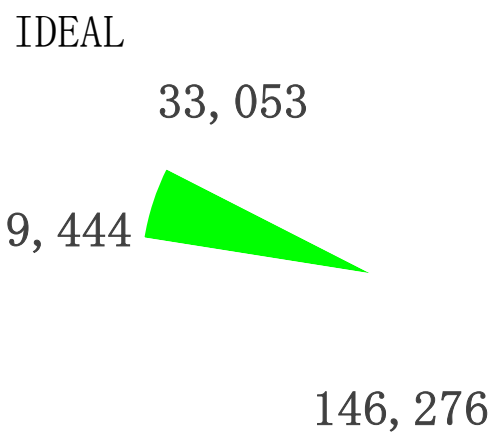


図 6. IDEAL データベースの内訳．緑は ProS(機能部位)，赤は天然変性領域，青は構造領域を示しており単位は残基数である．

近年，この機能部位を予測する複数のプログラムが開発されている[6, 15, 47-58]．しかし，天然変性領域予測が実用精度に達しているのに対して，機能部位予測はその域に達しておらず，複数の予測プログラムの予測結果には再現性が見られない．機能部位予測を困難にしている原因の１つとして，既存の機能部位のデータ数の不足が挙げられる．機能部位を予測する多くのプログラムが機械学習を用いて予測を行なっており，一般的に機械学習ではデータ数の不足は予測精度向上のボトルネックとなる．様々な生命現象に関与している天然変性タンパク質の重要性を考慮すると，機能部位のデータ数が不足している状況下であっても，天然変性領域中の機能部位を予測できるプログラムが期待される．高精度での機能部位予測は生命現象のさらなる理解に必須である．創薬分野では疾病に関わるタンパク質間相互作用を阻害する，経口投与が可能な分子量 1,000 以下の低分子化合物の生成が期待されている．天然変性領域中の機能部位は数残基から数十残基の短い領域であり低分子である．疾病に関わる天然変性領域の機能部位を予測することで，その領域を介したタンパク質間相互作用を阻害する化合物の生成が可能となる．また翻訳語修飾であるリン酸化においては，リン酸化を行うキナーゼは標的である基質のリン酸化部位を正確に見分けてい

る．近年，MAP キナーゼ基質にはリン酸化部位の他にキナーゼが結合する標的配列が天然変性領域中に存在することが明らかになっており[59]，他の基質においても同様の標的配列の存在が示唆されている．これらの標的配列の同定においても天然変性領域中の機能部位予測プログラムは応用できる．また LLPS においては，その形成に天然変性タンパク質が関わっているとういう報告がある．LLPS 内ではタンパク質や RNA などが比較的弱い相互作用をすることで LLPS を維持している．しかし，そのメカニズムは不明な部分が多く今後の解明が期待される．そのために，天然変性領域中の機能部位予測プログラムの精度向上は必須である．

1.2　研究目的

本研究では天然変性領域中の機能部位を予測するプログラム,NeProc(Next ProS classifier)の開発を目的とした．前節で述べたように機能部位のデータ数は不足しており，学習データとして使用することが難しい．この問題に対処するため，本研究では機能部位データを学習せずに，機能部位を予測するプログラムを作成した．天然変性領域中の機能部位は，天然変性領域中にありながら構造領域的性質を示すことが知られている．この傾向をターゲットとすることで予測を試みる．また，提案する予測法の予測精度を既存の機能部位予測プログラムと比較することで，機能部位を学習せずとも既存の機能部位を学習しているプログラムと同程度の精度で予測が可能であることを示す．結果として本研究の提案手法が機能部位のデータ不足を克服できる可能性があることを示す．

１.３　本論文の構成

図 7 に本論文の構成を示す．本論文では 2 章で関連研究として既存の天然変性領域予測プログラム，天然変性領域中の機能部位予測プログラムについてそれぞれの特徴を述べる．3 章では本研究で開発した NeProc について開発方法や予測精度を示す．4 章ではヒトプロテオームに対して NeProc を用いることで，ヒトプロテオーム中にどの程度機能部位が存在すると推定されるか,および NeProc の予測する機能部位の

傾向や問題点について焦点をあてる．5 章では本論文のまとめや NeProc の限界および今後の展望について話を展開する．

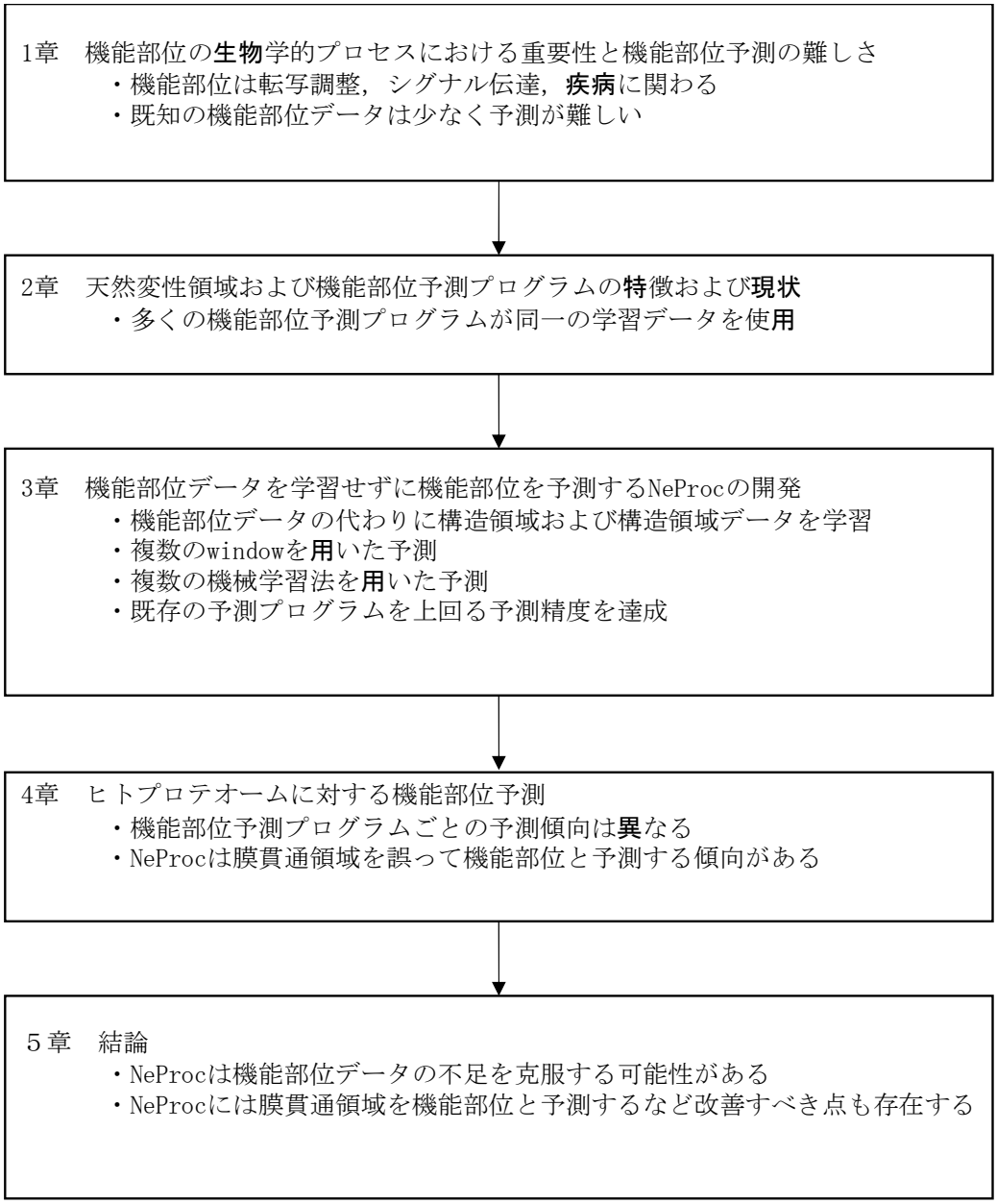


図 7. 本論文の構成.

第 2 章　関連研究

２．１　はじめに

近年，様々な天然変性領域中の機能部位を予測するプログラムが開発されてきた[6, 15, 48-58]．これらの予測プログラムの一部を表 1 に示した．表 1 には各プログラムの名称，作成年度，用いている手法，機能部位情報を学習しているかおよび入力している特徴量を示している．多くの予測プログラムが天然変性領域情報を特徴量としている．これは機能部位が天然変性領域中の領域であることを考慮すると違和感はなく，機能部位予測において天然変性領域情報は重要な特徴量であると言える．そこで本章では，機能部位予測において重要な特徴量である天然変性領域情報を提供する天然変性領域予測プログラムが，どのようにして天然変性領域を予測しているかを説明した後に，既存の機能部位予測プログラムがどのような戦略で機能部位を予測しているか述べる．

表 1．機能部位予測プログラム．

predictor name	year	method	learing functional region	input feature	reference
retro-MoRFs	2010	SF	+	IDR, AAS	[47]
ANCHOR2	2018	SF	+	IDR, pairwise energy	[6]
α-MoRFpred	2005	ML(NN)	+	IDR, PP, 2D	[48]
MoRFpred	2012	ML(SVM)	+	IDR, RSA, PP, βF, PSSM	[49]
MFSPSSMpred	2013	ML(SVM)	+	PSSM	[50]
PepBindPred	2013	ML(NN)	+	IDR, AAS, 2D, PP	[51]
DISOPRED3	2015	ML(NN, SVM)	+	IDR, PSSM	[15]
DisoRDBbind	2015	ML(LR)	+	IDR, RSA, PP, βF, PSSM	[52]
MoRFchibi-Web	2016	ML(SVM, NB)	+	IDR, AAS, PSSM	[53]
fMoRFpred	2016	ML(SVM)	+	IDR, RSA, PP, βF, PSSM	[54]
Predict-MoRFs	2016	ML(SVM)	+	PSSM	[55]
SPOT-MoRF	2020	ML(RNN)	+	IDR, PSSM, 2D, AAS	[56]
NeProc	2020	ML(NN, SVM)		IDR, PSSM	
OPAL+	2018	CN	+	2 programs	[57]
DEPICTER	2019	CN	+	10 programs	[58]

SF はスコア関数を用いた予測モデル，ML は機械学習を用いた予測モデル，CN はコンセンサスをとる予測モデルを示している．その他の略語は NN(neural network)，SVM(support vector machine),LR(logistic regression)，NB(naïve Bayes)，IDR(intrinsically disordered region)，AAS(amino acid sequence)，PP(physicochemical property),RSA(relative solvent accessibility)，2D(secondary structure),βF(β-factor),PSSM(position-specific scoring matrix)を示しており，“+”は機能部位の情報を学習していることを示している．

２．２　天然変性領域予測プログラム

これまでに天然変性領域を予測するために多くのプログラムが開発されてきた[3-7, 10-19, 60]．これらの予測プログラムを採用している予測法により大別して「スコア関数を用いた予測モデル」,「機械学習を用いた予測モデル」,「複数の予測モデルのコンセンサスを取る予測モデル」の３パターンに分類し，表２に示した．天然変性領域予測ではタンパク質のアミノ酸配列に特徴量を付加し，予測モデルへ入力することで予測を行う．使用する特徴量はプログラムごとに異なるが，アミノ酸の物理化学的性質，２次構造形成傾向および位置特異的スコア行列（Position Specific Scoring Matrix ： PSSM)などを特徴量とするプログラムが多い．PSSM は，一般的に相同性検索プログラム PSI-BLAST[61]を，データベースに対して複数回実行することで得られるマルチプルアライメントでのアミノ酸の出現頻度を示したスコア行列である．また天然変性領域予測では，予測したいアミノ酸残基の周辺の残基の特徴も含めて予測モデルへ入力することが定石となっている．このことを“window を用いた予測”といい，“window サイズ 15 で予測する”というのは“予測したい残基を中心に N および C 末端両方向に隣接する 7 残基の合計 15 残機残基の特徴量を用いて予測する”，ことを意味している(図 8)．多くの予測プログラムが window を用いた予測を行なっており，採用する window のサイズは各プログラムで異なる．これより，３つのタイプ別に予測プログラムを紹介する．

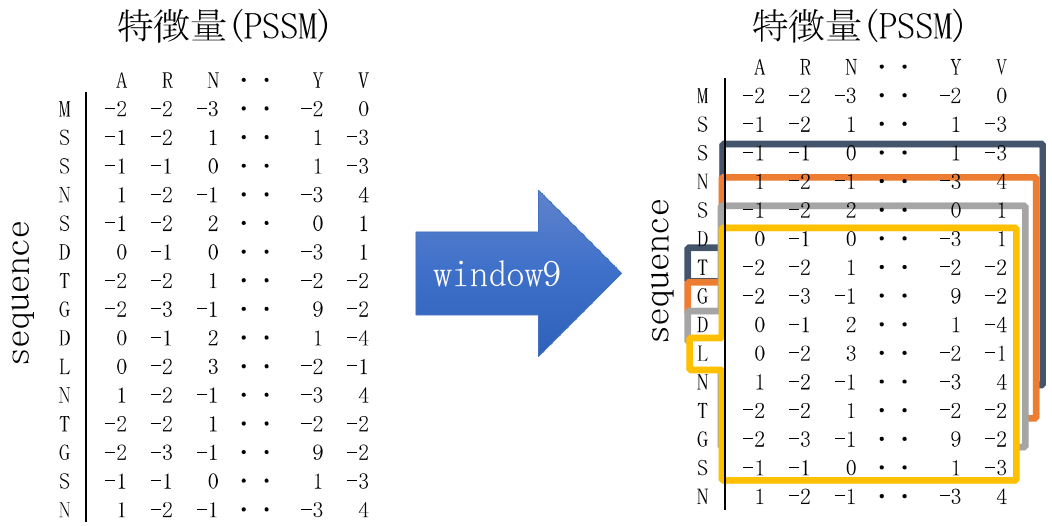


図 8. window の取り方.

表 2. 天然変性領域予測プログラム.

predictor name	year	method	input feature	reference
uversky plot	2000	SF	PP	[3]
Glob plot	2003	SF	AAC	[4]
FoldIndex	2005	SF	PP	[5]
IUPred2A	2018	SF	pairwise energy	[6]
PONDR	1999	ML(NN, SVM)	PSSM, PP, 2D	[7]
DISpro	2005	ML(RNN)	PSSM, 2D	[8]
PrDOS	2007	ML(SVM)	PSSM, AAS	[9]
ESpritz	2011	ML(RNN)	PSSM, AAS	[10]
SPINE-D	2012	ML(NN)	PP, PSSM, 2D	[11]
MFDp2	2013	ML(SVM)	PSSM, 2D, AAS	[12]
s2D	2014	ML(NN)	PSSM	[13]
SLIDER	2014	ML(NN, SVM)	PP	[14]
DISOPRED3	2015	ML(NN, SVM, KNN)	PSSM	[15]
DisPredict	2015	ML(SVM)	PSSM, AAS, 2D, ASA, PP	[16]
SPOT-Disorder	2017	ML(RNN)	PSSM, 2D, AAS	[17]
NeProc	2020	ML(NN, SVM)	PSSM	
MetaDisorder	2012	CN	13 programs	[18]
MobiDB-lite	2017	CN	8 programs	[19]

SF はスコア関数を用いた予測モデル, ML は機械学習を用いた予測モデル, CN はコンセンサスをとる予測モデルを示している.その他の略語はNN(neural network), SVM(support vector machine),RNN(recurrent neural network), KNN(k- nearest neighbor algorithm), PP(physicochemical property), AAC(amino acid composition), PSSM(position-specific scoring matrix), 2D(secondary structure), AAS(amino acid sequence), ASA(accessible surface area)を示している.

2.2.1 スコア関数を用いた予測モデル

はじめにスコア関数を用いた予測モデルを紹介する．スコア関数を用いた予測モデルでは，天然変性領域と構造領域を識別するためのスコア関数を人が定義し，最適な係数(パラメータ)を決定することで天然変性領域予測を行う予測モデルである．これらの予測プログラムは機械学習を用いた予測モデル，コンセンサスを取る予測モデルと比較して，予測を行う際の計算量が少なく，予測にかかる時間が短い傾向にある．また予測に用いられる特徴量が少なく，任意のスコア関数を用いているため，予測モデルがどのように天然変性領域を見分けているかなど，予測結果を直感的に理解しやすい.

(A) Uversky Plot

Uversky Plot は他の天然変性領域予測プログラムとは異なり，入力されたタンパク質が天然変性タンパク質であるか否かを予測する．Uversky Plot は天然変性タンパク質が球状タンパク質と比較して荷電残基が多い傾向があること[62, 63]，疎水性アミノ酸の含有率が少ない傾向があること[62, 63]，の 2 つの性質に着目し，タンパク質の疎水性と net charge を特徴量に用いることで天然変性タンパク質を予測する．図 9 は Uversky Plot を示しており，縦軸に平均の net charge<R>，横軸にタンパク質の平均の疎水性<H>をプロットしている．図中の境界線によって天然変性タンパク質と球状タンパク質を識別している．Uversky Plot に対して window を採用した FoldIndex[5]では，天然変性領域の予測を提供している．

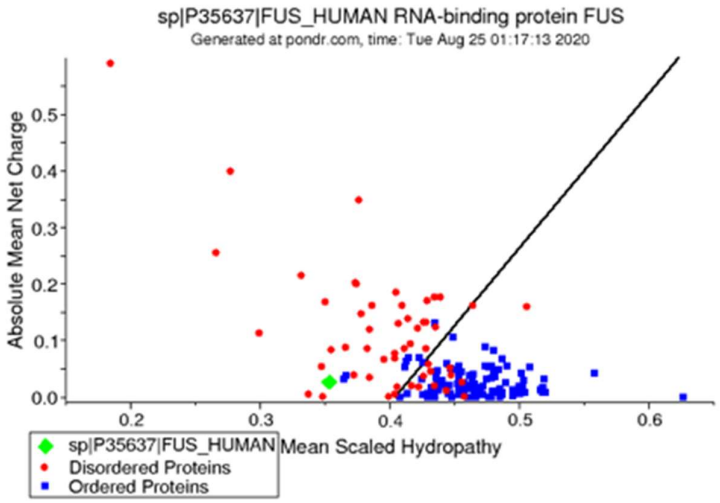


図 9. Uversky Plot.

(B) IUpred2A

IUpred2A は統計ポテンシャルを用いて，タンパク質中のアミノ酸残基のペアが接触を形成する傾向を指標として予測するモデルである．アミノ酸 i の位置 k におけるエネルギーは次のように定義される．

$$e_i^k = \sum_{j=1}^{20} P_{ij} c_j^k,$$

c_j^kは，位置 k の周辺（window サイズ 100）におけるアミノ酸 j の頻度を示している．P_{ij}

はアミノ酸 i とアミノ酸 j の相互作用エネルギーを示しており，既知の球状構造のエネルギーに適合するように，学習によって最適化されたエネルギースコアである．IUpred2A の興味深い点は，天然変性領域予測の核とも言える相互作用エネルギーを球状構造のみから最適化しているにも関わらず，天然変性領域を識別できることである．

2.2.2 機械学習を用いた予測モデル

次にニューラルネットワーク(NN)，サポートベクターマシン(SVM)，k 近傍法(KNN)などの機械学習法を用いた予測モデルを紹介する．「機械学習を用いた予測モデル」では「スコア関数を用いる予測モデル」とは異なり，天然変性領域を識別する関数は選択した機械学習モデルに依存している．機械学習モデルは特徴抽出の過程がブラックボックスであるために，予測結果を直感的に理解するのは難しい．しかし，機械学習モデルは多くの特徴量からでも学習によって天然変性領域の特徴を捉えることが可能である．そのため，複数の特徴を予測モデルへ入力するモデルが多い．これより，DISOPRED3，SPOT-Disorder，PONDR を順に紹介する．これらの予測プログラムは「機械学習を用いた予測モデル」の中でも高い予測精度を達成している予測プログラムである．

(A) DISOPRED3

DISOPRED3 [15]は NN，SVM，KNN を用いて，天然変性領域を予測するプログラムである．DISOPRED3 の前バージョンである DISOPRED2 [22]は長い天然変性領域の予測を苦手としていた．DISOPRED3 はこれを克服するために，初期バージョンの DISOPRED [64]の NN を，新たに作成した長い天然変性領域が多く含まれているデータセットを用いて学習し予測モデルを作成し直した．DISOPRED3 では，新たに作成した NN の予測モデルの予測結果，DISOPRED2 の予測結果に加えて，さらに KNN を用いた予測結果とタンパク質の末端を表す情報の 4 つを特徴量として，NN を用いて天然変性領域を予測している．DISOPRED3 では各アミノ酸残基に対して天然変性領域予測を行う際，window サイズを 15，特徴量を 4 つの 4×15 次元の特徴量で予測を行う．DISOPRED3 を構成している DISOPRED(NN)，DISOPRED2(SVM)，KNN の 3 つの機械学習モデルへの入力は，PSI-BLAST で生成した PSSM を特徴量として与えている．

(B) SPOT-Disorder

SPOT-Disorder [17]はリカレントニューラルネットワーク(RNN)の一種であるLSTM(long-short term memory)を用いて天然変性領域を予測するプログラムである. LSTM は予測したいアミノ酸の前後の情報を加味して最適化を行う. つまり LSTM では window を指定せずとも予測したいアミノ酸の周辺の情報を, 予測に反映する能力があると言える. SPOT-disorder では予測したアミノ酸残基の前後 100 残基分の情報を記憶するモデルを構築している. SPOT-disorder では特徴量として PSI-BLAST で生成した PSSM(20 次元), マルチプルアライメントの各サイトのシャノンエントロピー(2 次元), 2 次構造予測プログラムの SPIDER2 の出力(17 次元), 疎水性や電荷などアミノ酸の物理化学的性質(7 次元) の合計 46 次元に対してサイズが 200 の window を採用した非常に多い特徴量を用いて予測を行なっている[17].

(C) PONDR

PONDR では 5 種類の予測モデル, VLXR, XL1_XT, CAN_XT, VL3_BA, VSL2 を提供しているが, ここでは VSL2 について紹介する. VSL2 は 5 種類の予測モデルの中で最も精度よく天然変性領域を予測する. これは VSL2 が他の予測モデルが苦手としていた, 30 残基以下の IDR 予測を改善したことに起因している. VSL２は VSL2-L, VSL2-S, MetaPredictorM の 3 種類のモデルから構成されており, 各モデルはアミノ酸組成, 配列の位置情報(N 末 or C 末 or 中間), window 範囲内の K2 エントロピー, net charge, 平均疎水性, 柔軟性, PSSM, 2 次構造予測結果の 54 次元の特徴量で学習が行われている. VSL2-L は 30 残基以上の天然変性領域に特化した SVM を用いたモデルであり, VSL2-S は 30 残基未満の天然変性領域に特化した SVM を用いたモデルである. しかし, 未知のタンパク質が含んでいる天然変性領域の長さはわからないため, 予測したいアミノ酸残基を VSL2-L, VSL2-S のどちらのモデルを用いて予測を行うかは決めることができない. この問題を VSL2 では MetaPredictorM を用いて解決している. MetaPredictorM は VSL2-L, VSL2-S とは独立して学習が行われたモデルであり, 入力されたアミノ酸が 30 残基以上の天然変性領域であるか, 30 残基未満の天然変性領域であるか, を求めるモデルである. VSL2-L と VSL2-S の出力値に対して MetaPredictorM の出力を重みとして用いることで, VSL2-L と VSL2-S のどちらの出力を採用するかを決定している.

2.2.3 複数の予測モデルのコンセンサスを取る予測モデル

コンセンサスを取る予測モデルは既存の天然変性領域予測プログラムから任意の天然変性領域予測モデルを採用し，その予測結果に対してコンセンサスを取ることで天然変性領域を予測する．コンセンサスを取る際のルールは各予測プログラムで異なる．比較的厳しいルールを採用している予測プログラムが多く，予測される天然変性領域数は少ない傾向があるが，確度が高い予測が可能であり偽陽性が抑えられているプログラムが多い．

(A) MobiDB-lite

MobiDB-lite[19]は8種類の天然変性領域予測プログラムの予測結果に対してコンセンサスを取ることで天然変性領域を予測する．MobiDB-lite は ESprits-DisProt [10], ESpritz-NMR [10], ESpritz-X-ray [10], IUpred-long [65], IUpred-short [65], DisEmble-465 [66], DisEmble-HL [66], GlobPlot [4]の 8 種類の天然変性領域予測プログラムを用いている．MobiDB-lite では 8 種類のプログラムの内，少なくとも 5 つのプログラムで天然変性領域と予測された領域を天然変性領域としている．さらに予測された天然変性領域と構造領域に長さによるカットオフを設けている．3 残基以下の天然変性領域は構造領域へと，3 残基以下の構造領域は天然変性領域へと予測が変更される．また，10 残基以下の構造領域は，その領域の前後が 20 残基以上の天然変性領域である場合は予測が天然変性領域へ変更される．最後に 20 残基以上の天然変性領域が予測として採用される．このように MobiDB-lite は，非常に厳しいルールを設けているため，偽陽性率の低い予測を可能としている

(B) metaDisorder

metaDisorder[18]は 13 種類の天然変性領域予測プログラムの予測結果に対して重みを付加しコンセンサスを取ることで天然変性領域を予測する．metaDisorder では，DisEMBLE[66], DISOPRED2[22], DISpro[67], Globplot[4], iPDA[68], IUPred-long[65], IUpred-short[65], Pdisorder, POODLE-S[69], POODLE-L[70], PrDOS[9], Spritz[71], RONN[72]の 13 種類の予測結果を用いる．metaDisorder では各予測プログラムが出力したスコアに対して，その予測プログラムの予測精度をかけわせた値を足し合わせる．その値を正規化し metaDisorder の予測値とする．そして 10 回のクロ

スバリデーションで決定した閾値を用いて各残基の予測を行う．最後に，サイズ 5 の window 幅で平滑化を行うこと．スムージングされた予測値を用いて最終的な天然変性領域を決定する．

2.2.4 まとめ―天然変性領域予測プログラム

3 つの分類に従い代表的な天然変性領域予測プログラムを紹介してきた．予測精度に関しては「機械学習を用いた予測モデル」が高い傾向にあるが，どの予測プログラムも実用精度を達成している．これらの予測プログラムが予測する天然変性領域は機能部位予測の特徴量として用いられる．そのため，天然変性領域を正確に予測できることは機能部位予測において必要とされる．次項では機能部位予測モデルが，何をターゲットとしてどの様に機能部位を予測しているかを，既存の予測プログラムと共に論じる．

2.3 機能部位予測プログラム

表 1 では天然変性領域予測プログラムと同様に「スコア関数を用いた予測モデル」,「機械学習を用いた予測モデル」,「複数の予測モデルのコンセンサスを取る予測モデル」の 3 パターンに分類した．天然変性領域予測プログラムと同様に機械学習を採用しているプログラムが多い．前項で述べた各分類の特徴は機能部予測においても共通である．機能部位予測において頻繁に使用される特徴量は，天然変性領域情報，アミノ酸の物理化学的性質，および PSSM である．また，NeProc を除く全てのプログラムがデータ数の不足している機能部位データを学習している．これより代表的な機能部位予測プログラムを，学習データおよび予測手法に触れながら紹介する．

2.3.1 代表的なプログラム

(A) ANCHOR2

ANCHOR2[6]は天然変性領域予測プログラム IUpred と同様にアミノ酸残基間のエネルギーを計算することで機能部位を予測している．ANCHOR2 では“構造領域中の結合領域のような領域”かつ“天然変性領域中に存在する領域”を機能部位と定義し予測を行なっている．IUpred が相互作用エネルギー P を球状タンパクに適合するように最適化していたのに対して，ANCHOR2 では天然変性領域中の結合領域データベース

(Database of Disordered Binding Sites : DIBS) [44]から獲得した機能部位データに適合するように最適化している．ANCHOR2 では，エネルギー関数の結果と IUpred2A の天然変性領域予測の結果を特徴量として機能部位を予測している．ANCHOR2 では精度を評価する際に，天然変性領域と注釈がある領域に対して機能部位と予測した場合，未知の機能部位である可能性があるため評価しない．

(B) MoRFpred

MoRFpred[49]は，α-MoRFpred[48]の機能を拡張した機能部位予測プログラムである．α-MoRFpred がαヘリックス構造を形成する機能部位のみを予測するのに対して．MoRFpred は形成する２次構造に関わらず機能部位を予測する．MoRFpred では，学習データは Protein Data Bank(PDB) [73, 74]を元に作成している．70 残基以上の球状タンパク質との複合体構造が存在する 5 残基から 25 残基の短いペプチドを機能部位として抽出し，それらを UniProt データベースへマップすることで，タンパク質の配列を獲得する．このタンパク質の機能部位が単独状態において天然変性領域かを Gunasekaran らの評価法[75]を用いて評価し，天然変性領域ならば機能部位としている．その結果，10549 残基の機能部位を含んだ 840 タンパク質からなるデータセットを作成した．このデータセットは表１の ANCHOR, PepBindPred, DisoRDBbind および NeProc 以外全てのプログラムで用いられている．MoRFpred は特徴量に天然変性領域予測，アミノ酸の物理化学的性質，PSSM, Relative Solvent Accessibility(RSA), β-factor を特徴量に用いている．これらの特徴量に window サイズ 24 を採用し SVM へ入力する．そしてクエリ配列を上記のデータセットへアライメントした結果と統合することで，機能部位を予測している．予測結果は機能部位とその他の領域の２つに識別する．

(C) MoRFchibi-Web

MoRFchibi-Web[53]は機械学習の SVM とベイズ推定を用いて機能部位を予測する．学習データは MoRFpred と同様のデータを学習して予測モデルを構築している．MoRFchbi-Web では内部で MoRFchib[76]と $MoRF_{DC}$ の２つのモデルを使用している．MoRFchibi では２つの SVM を用いて，予測したい領域のアミノ酸組成のコントラストおよびアミノ酸配列の学習データ内の機能部位配列との類似度を数値化している．組

成のコントラストとは，機能部位と隣接領域の組成の不一致度を意味しており，隣接領域との差に着目している点がユニークである． $MoRF_{DC}$ は天然変性領域予測プログラム ESpritz-DisProt[10]を用いて 天然変性領域を予測している．さらに機能部位は配列保存性が高いことに着目し，PSI-BLAST を用いた保存性も評価している．MoRFchibi-Web ではこれらのアミノ酸組成の隣接領域とのコントラスト，配列類似度，天然変性領域予測および配列の保存度の４つの指標に対してベイズ推定を用いることで統合し，機能部位とその他の領域の２つに識別する．

(D) DISOPRED3

DISOPRED3[15]は天然変性領域と共に機能部位を SVM を用いて予測する．学習データは MoRFpred でのデータをベースに作成している．DISOPRED3 ではこの学習データにネガティブデータを追加している．ネガティブデータとして，立体構造が既知のタンパク質中の構造ドメインを繋ぐドメインリンカーを採用している．DISORED3 の SVM では天然変性領域を機能部位と結合には直接関与しない領域かを識別している．特徴量として PSSM, 天然変性領域情報および window 内のアミノ酸組成を採用している．最終的に DISOPRED3 では天然変性領域予測の結果も統合し構造領域，天然変性領域および機能部位を予測する．

2.3.2 まとめ―機能部位予測

本節では機能部位予測プログラムについて記述した．これらの全ての予測プログラムでは天然変性領域情報を特徴量として含んでいる．それに加えて PSSM やアミノ酸の物理化学的性質を特徴量として機能部位を予測している．これらの予測プログラムはある程度の予測精度を達成しているが，天然変性領域予測ほどの精度は達成できていない[77]．また，表１の ANCHOR, PepBindPred, DisoRDBbind および NeProc 以外全てのプログラムが MoRFpred の学習データを用いて予測モデルを構築している．同一のデータを用いている正確な理由は定かではないが，学習に利用できる機能部位データが不足していることが原因の１つであると考えられる．同一のデータでは予測モデルの精度向上に限界があり，異なる学習データでの予測プログラムを開発することが必要である．しかし，実験的に検証された機能部位データは不足している．また実験的に確認された機能部位データ数が爆発的に増加する可能性が低いことを考慮する

と，機能部位データを学習せずに機能部位を予測するプログラムが必要である．

2.4 まとめ

本章では機能部位予測において重要な特徴量である天然変性領域情報を提供する天然変性領域予測プログラムの予測法および既存の機能部位予測プログラムの予測法について記述した．全ての機能部位予測プログラムが天然変性領域予測の結果を特徴量としていることから，機能部位予測における天然変性領域予測プログラムの予測精度は重要であると言える．また，機能部位予測における機能部位を学習せずに予測するプログラムの必要性を述べた．次章からは機能部位データを学習せずに機能部位を予測する NeProc のモデル構築から予測精度などについて論じる．

第３章　天然変性領域中の機能部位予測プログラム NeProc の開発

３．１　はじめに

天然変性タンパク質は，天然変性領域中の機能部位を介した相互作用によって，転写調整やシグナル伝達などの生物学的に重要な役割を果たしている[2, 35-37]．2 章では既存の機能部予測プログラムについて触れたが，その多くが同一の学習データを用いて作成されていた．また，これらの予測プログラムの予測精度は実用精度には達していない．これは既知の機能部位データの数が限られていることが原因の１つとして考えられる．天然変性タンパク質データベース IDEAL[45, 46]では実験的に検証された機能部位を“protean segments” (ProSs)とし提供している．IDEAL では 146,276 残基の構造領域，33,053 残基の天然変性領域および 9,444 残基の ProS 残基が収録されているが，天然変性領域や構造領域と比較して ProS データの数は少ない．そのため機能部位データを用いずに機能部位を予測することが可能な予測プログラムが必要である．

天然変性領域中の機能部位は単独ではヒモ状であるが相互作用パートナーと出会うと単純な 2 次構造を形成し相互作用する(2 次構造を形成せずにヒモ状のまま相互作用する機能部位も存在する)．そのため，天然変性領域中の機能部位と，その他の天然変性領域では異なる物理化学的性質を保持している[78, 79]．また構造領域は数珠状に連なったアミノ酸配列が折りたたまれ複雑な構造を形成しているが，局所的な範囲に着目するとα-helix やβ-sheet などの単純な 2 次構造の組み合わせによって構成されている．そのため，天然変性領域中の機能部位と構造領域中の局所的な領域には単純な 2 次構造を形成できる点において共通であり，予備研究において 2 次構造予測プログラムによってα-helix を形成する機能部位は比較的予測できることが確認されている．さらに，天然変性領域中の特定の機能部位は配列の保存度が高く[80, 81]，

構造領域との類似性が天然変性領域との類似性より高いことが報告されている[82]. また，機能部位のアミノ酸組成は“構造領域のアミノ酸組成”および“天然変性領域のアミノ酸組成”，その双方の特徴を示すことも報告されている[79]．したがって，天然変性領域中の結合領域は長い天然変性領域中に存在する構造領域的性質を示す短い領域と定義できる．そこで本研究では，データ数に限りがある機能部位データは用いずに，天然変性領域および構造領域データを学習することで，機能部位の予測を可能とするプログラム，NeProc (Next ProSs Classifier) を開発した．本章では NeProc のモデル構築および精度などに焦点を当てる．

３．２　方法

NeProc では天然変性領域中の機能部位データを学習せずに，機能部位を予測することを目的とした．機能部位データを含まないデータを学習することによって，機能部位がどの程度予測できるかを検証するために，NeProc を比較対象プログラムと同様に単純な予測法を用いて開発した．比較対象プログラムとして ANCHOR2 [6]，DISOPRED3 [15]，および MoRFchibi-Web [76]を採用した．ANCHOR2 は統計ポテンシャルを用いて，機能部位を予測する．DISOPRED3 と MoRFchibe-Web は機能部位データを学習した単層のニューラルネットワークやサポートベクターマシンのシンプルな機械学習モデルを採用し機能部位を予測している．また，MoRFchibi-Web は最近の結合領域予測プログラムの比較において，高い精度を記録している[53, 77]．NeProc は DISOPRED3 と MoRFchibi-Web の２つの予測プログラムと同様にシンプルな機械学習モデルを採用した．したがって，DISOPRED３および MoRFchibe-Web が用いる予測法は NeProc に類似しているが，学習データに機能部位データを含む点が異なる．NeProc は長い天然変性領域中の構造領域的性質を示す比較的短い領域を特定することで，機能部位を予測する．この予測法を実現するために，NeProc では長い window サイズを採用した Lmodel と短い window サイズを採用した Smodel の２つのモデルを用いている．Lmodel を用いて天然変性領域を予測し，Smodel を用いて予測された天然変性領域内の

構造領域様な傾向を示す短い領域を識別する．2 章において述べたが，機能部位予測では天然変性領域予測の精度が機能部位予測の精度に大きく影響する.そこで,NeProcの天然変性領域予測の精度が既存の天然変性領域予測プログラムと比較してどの程度であるかを，検証するためにSPOT-disorder [17]，DISOPRED3 [15]，IUpred2A [6]およびMobiDB-lite [19]を比較対象プログラムとして比較を行う．機能部位予測の予測精度の比較にはANCHOR2 [6]，DISOPRED3 [15]およびMoRFchibi-Web [76]を比較対象プログラムとした.

３．２．１　データセット

NeProc はアミノ酸配列から天然変性領域を予測し，予測された領域内にある構造領域的性質を示す領域を予測するために，天然変性領域および構造領域を学習データとして用いる．NeProc では学習データとして DM4229 を用いている．DM4229 は天然変性領域予測プログラムである DPINE D[11]で使用されている学習データであり，PDB[73]と DisProt [83]から獲得した 4229 個のタンパク質で構成されているデータセットである.PDB からは解像度が 2Å以下の X 線結晶構造解析により構造が決定された 60 残基以上のタンパク質で，座標情報がないアミノ酸残基を含むタンパク質を配列相同性が 90%以下となるように選定されている.DisProt からは disorder と注釈があるタンパク質を獲得している．PDB と DisProt から得られたタンパク質を配列相同性が 25%以下となるように，Blastclust を用いてクラスター分析を行い，1)最も天然変性領域が長いタンパク質，2)最も天然変性領域が短いタンパク質，3)最も全長が長いタンパク質，の優先順位に従って各クラスターから代表タンパク質が選ばれている．それらのタンパク質に，DisProt から得られた全域が 天然変性領域であるタンパク質を含めて,再度 Blastclust を用いたクラスター分析により配列相同性を 25%まで減らし，4229 個の代表タンパク質を選定している.

本研究では Blastclust を用いて DM4229 と以下で説明するテストデータセットの配列相同性を 25%以下となるように DM4229 からタンパク質を取り除いた．結果と

して 925,412 残基の構造領域と 100,284 残基の天然変性領域からなる 4,189 個のタンパク質が得られた．それらのタンパク質のうち 842 個のタンパク質をハイパーパラメータの決定に，残りの 3,347 個のタンパク質をパラメータである重みとバイアスの最適化に用いた．

天然変性領域予測精度を計測するためのテストデータとして Critical Assessment of protein Structure Prediction 10 (CASP10)[84] の天然変性領域予測問題，IDEAL データベース[45, 46]および CheZOD データベース[85]の３つのデータを用いた．CASP10 は天然変性領域予測プログラムの精度を測るベンチマークとして用いられている．IDEAL と CheZOD は天然変性領域の情報を収録しているデータベースである．

IDEAL は，天然変性領域中の機能部位である ProS の情報を提供している．IDEAL の ProS データは論文から収集されており，i）単独状態での天然変性領域の証拠，および ii）PDB 内の 1 つ以上の結合パートナーを持つ 1 つ以上の構造を保持している，ことが実験的に検証されたデータである．したがって，IDEAL の ProS データを使用して結合領域予測の精度を測定することができる．

予測モデルの精度を評価するには，可能な限り多くのデータセットで評価することが望ましい．なぜならデータベースには収録するデータを収集する際の方法や基準によって，偏りが生じている可能性があるからである．IDEAL データベースの天然変性タンパク質は核タンパク質が多く含まれており，データに偏りがある可能性がある．そのため IDEAL データベースでの精度評価のみでは不十分であり，他の機能部位データを用いた精度評価を行う必要がある．

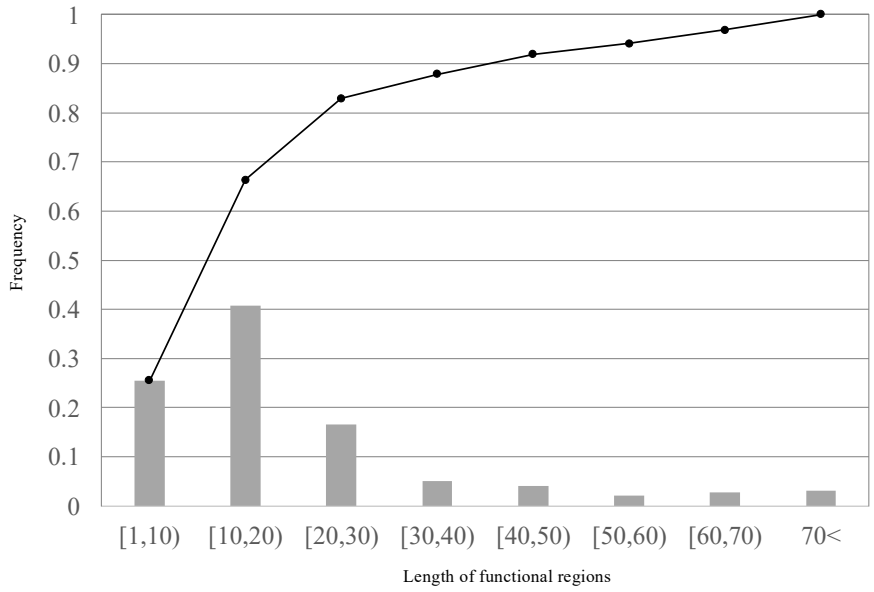


図 10.　IDEAL データベース内の機能部位の長さの分布.

そこで本章では UniProt データベース[86](リリース 2018_07)より天然変性領域中の機能部位データセットを作成した．UniProt データベース[86]はタンパク質の実験に基づいた確度の高い機能情報を提供している．また，天然変性領域予測プログラムの予測精度は実用精度に達している．つまり，予測プログラムによって予測された天然変性領域に UniProt においてタンパク質間相互作用に関わる情報が存在すれば，その領域は機能部位の可能性があるとみなすことができる．UniProt にはタンパク質データに Swiss-Prot と TrEMBL の２つのタイプがある．Swiss-Prot はマニュアルでアノテーションされレビューされた確度の高いデータであり，TrEMBL はコンピュータを用いて自動的にアノテーションされたデータである.本章では,確度の高い Swiss-Prot データを使用し，タンパク質の機能セクションの情報を抽出した．機能セクションには，その機能の説明および機能を有する領域の配列上の位置などの情報が含まれており，それら全ての情報を抽出した．IDEAL データベースが収録している ProS は比較的短い領域であり，80%以上の ProS が 30 残基より短い(図 10)．従って本章では 30 残基より短い領域の機能情報を採用した．抽出した機能の説明を分析したところ，

“region of interest”，“mutagenesis site”および“short sequence motif”の３つの機能に結合関連の単語が含まれていることが確認できた．そこで，この３つの機能を持つ領域をターゲットとした．また機能が“region of interest”の場合は，機能の説明の中に“interact”，“bind”および“motif”のいずれかのワードが含まれる領域を採用した．機能が“mutagenesis site”の場合にも，説明の中に“interact”または“bind”が含まれる領域を採用した．ただし，その条件に当てはまる場合も，機能の説明に“no”や“not”が含まれる領域は採用しなかった．天然変性領域は予測プログラムを用いて決定する．より正確な予測を行うために３つのプログラムを用いた天然変性領域の予測を行った．天然変性領域予測プログラム，MobiDB-lite [19]，DISOPRED3 [15]およびDICHOT [87]の３つの予測プログラムを採用した．MobiDB-liteは８つの天然変性領域予測プログラムのコンセンサスをとり天然変性領域を予測するプログラムであり，偽陽性を抑制しているプログラムである[19]．DISOPRED3 [15]はPSI-BLASTから生成されたPSSMを学習データとして機械学習アルゴリズムのニューラルネットワークを用いて予測するプログラムである．DISOPRED3はCASPコンテストに[84]においてトップクラスの予測精度を記録したプログラムである．DICHOTは，配列相同性をベースとした既知のドメインとの比較と，機械学習アルゴリズムのサポートベクターマシンを組み合わせた予測プログラムである．これら３つの予測プログラムのうち，少なくとも２つ以上が天然変性領域と予測した領域を採用した．この条件によって採用した天然変性領域にUniProtの機能情報が含まれていた場合，その領域をputative ProS(以降pProSと呼ぶ)と定義した．pProS抽出の一連の流れを図11に示した．

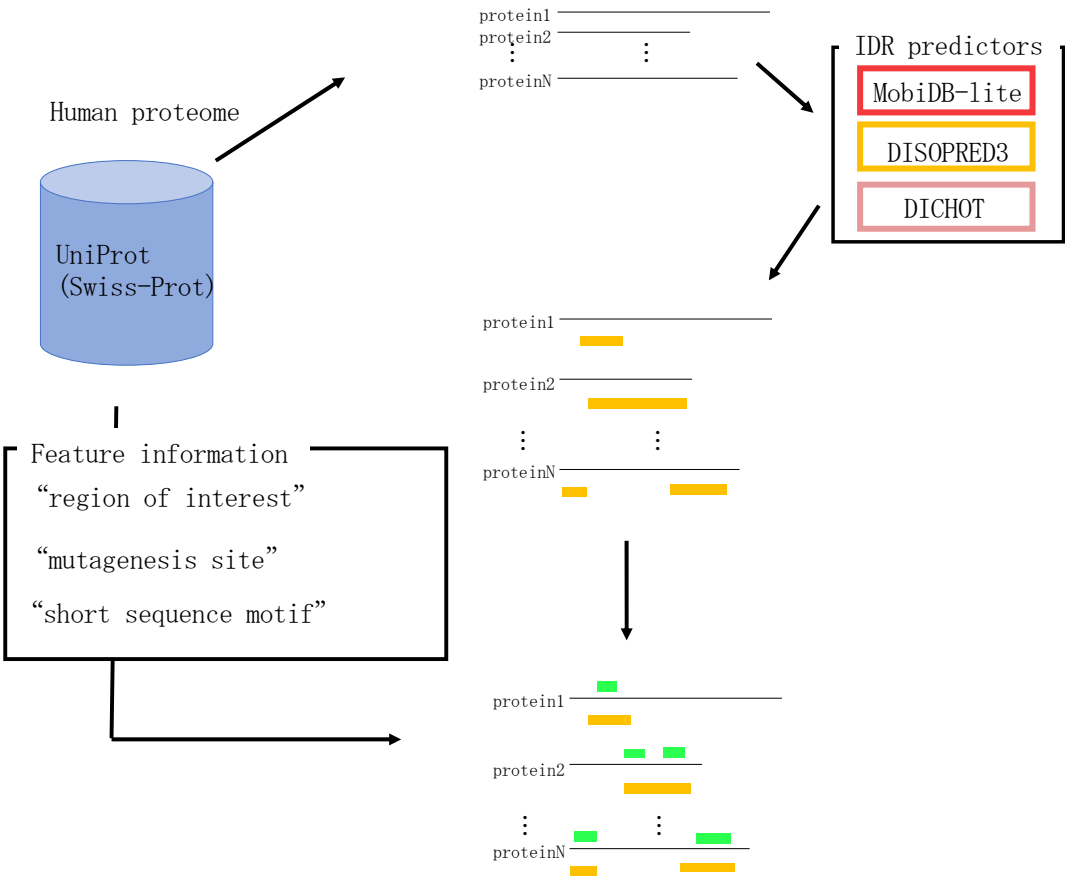


図 11.　pProS の抽出方法の概要．図中央の黒線はタンパク質を示しており，緑色の帯は pProS を示している．赤い帯は MobiDB-lite の予測結果，黄色い帯は DISOPRED3 の予測結果，ピンク色の帯は DICHOT の予測結果を示している．“region of interest”の場合は，機能の説明の中に“interact”，“bind”および“motif”のいずれかのワードが含まれる領域を採用した．機能が“mutagenesis site”の場合にも，説明の中に“interact”または“bind”が含まれる領域を採用した．ただし，その条件に当てはまる場合も，機能の説明に“no”や“not”が含まれる領域は採用しなかった．

pProS を含むタンパク質の構造領域は UniProt の“cross reference”セクションにある PDB データを用いて決定した．pProS と PDB 情報が重複した場合は，pProS を 30 残基以下の領域と定義したことを踏まえて，重複する PDB の構造が 31 残基以上である場合，その pProS はテストデータから除いた．表 3 は，これらのデータセットの統計を示している．

表 3. 各データセットの集計.

		sequences	residues	ordered	disordered	ProS	no anotations	ordered : disordered (ordered : ProS)
Training dataset	PDB	4,123	1,011,526	925,412	86,114	0	0	11 : 1
	DisProt	66	14,170	0	14,170	0	0	
Test dataset	CASP10	74	25,371	22,688	1,502	0	1,180	15 : 1
	CheZOD	117	13,069	4,082	8,987	0	0	1 : 2
	IDEAL	915	594,351	135,443	23,878	7,253	427,777	6 : 1 (18 : 1)
	pProS	1,518	1,162,230	170,302	0	12,418	979,510	(14 : 1)

ProS は天然変性領域中の機能部位を示している. no annotations は構造情報が不明なアミノ残基の数を示している.

３．２．２　NeProc のモデル構造

NeProc のモデルを作成するにあたり，DISOPRED3 の予測モデル（図 12）を参考とした. DISOPRED3 は 2 つのニューラルネットワークを用いており，最初のニューラルネットワークの出力を次のニューラルネットワークの入力とするモデル構造である. 各ニューラルネットワークは全て 1 層の入力層，隠れ層，出力層の 3 層から構成されている.

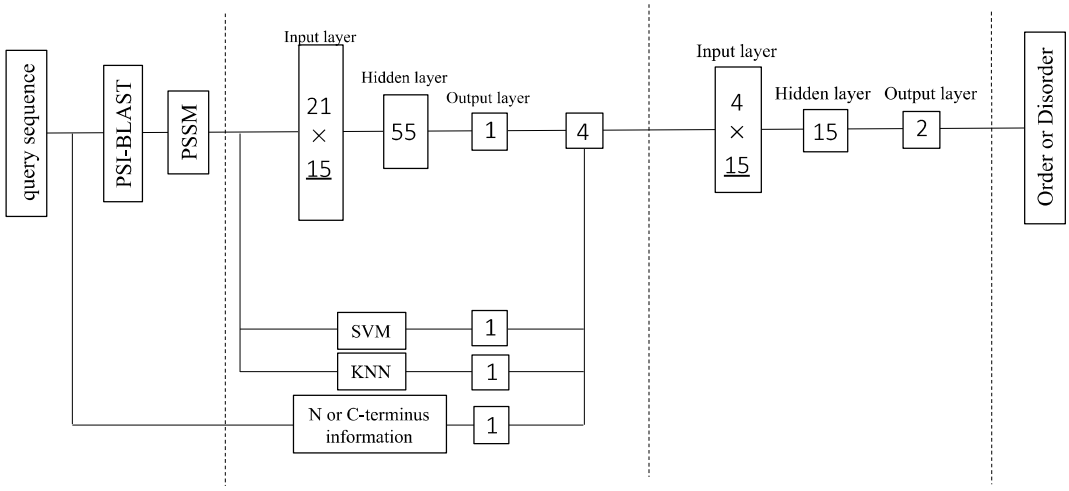


図 12.　DISOPRED3 の天然変性領域予測モデルの構造. 下線が付いている数字は window サイズを表している. その他の数字はノードの数を示している.

NeProc ではニューラルネットワークを 2 つ重ねて予測を行う DISOPRED3 の構造を採用し，Lmodel と Smodel の 2 つのモデルを作成した(NeProc のモデルを図 13 に示す). Lmodel の最初のニューラルネットワークでの window サイズは DISOPRED3 の 15

残基を採用した．Smodel ではLmodel との window サイズの差を大きくし，より局所的な構造領域的性質を捉えるため 3 残基の短い window サイズを採用した．各ニューラルネットワークの隠れ層および隠れノードの数は表 4 の組み合わせを用いてテストを行い決定した．２番目のニューラルネットワークの構築では，複数の window サイズをテストした．Lmodel では 15， 30， 40， 50 および 60 残基の 5 つの window サイズをテストした．15， 30 および 60 残基は 15 の倍数とし選択し，40 と 50 残基は 30 と 60 残基の間のギャップを埋める目的で選択した．Smodel の winodow サイズは Lmodel より小さい winodow を選択する必要があるため Lmodel よりも小さな 3， 5 および 10 残基を選択しテストを行った．2 番目のニューラルネットワークの並列に配置する最適なニューラルネットワークの組み合わせを決定するために，3 番目のユニットとしてニューラルネットワークとサポートベクターマシンを用いてテストを行った．2 番目の各ニューラルネットワークの出力を組み合わせは，可能な組み合わせを全てテストした．このテストは Lmodel と Smodel で個別に行い各々のモデル内での最適な組み合わせを決定した．隠れ層と隠れノードの数は表 4 にリストされている組み合わせを使用してテストした．

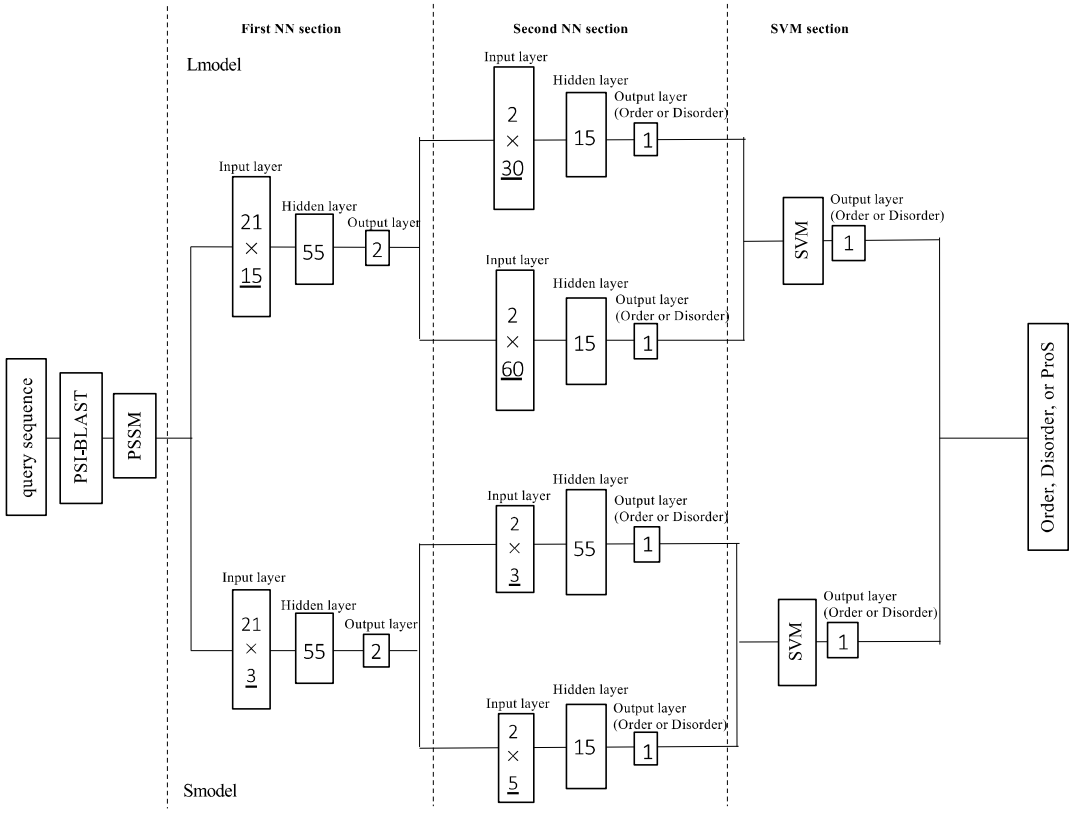


図 13. NeProc のモデル構造．下線が付いている数字は window サイズを表している．その他の数字はノードの数を示している．

Smodel と Lmodel は，確率ではなく構造領域(“ordered”)または天然変性領域(“disordered”)のバイナリー値を出力する．そのため最終的な機能部位の予測は，単純な決定ルールで実行される．Smodel と Lmodel からの出力の組み合わせは，次の 4 つの状態のいずれかである“disordered”/“disordered”(D / D)，“disordered”/“ordered”(D / O)，“ordered”/“disordered”(O / D)，または“ordered”/“ordered”(O / O)(Smodel の出力/ Lmodel の出力)． D / D の場合は天然変性領域であり，O / O は構造領域，O / D は機能部位領域であり，D / O は不明領域と予測する．天然変性領域予測の場合は予測された機能部位は天然変性領域として出力される．

表 4. 隠れ層および隠れノードの組み合わせ

A) Lmodel での隠れ層および隠れノードの組み合わせ.

	HIDDEN1	HIDDEN2	HIDDEN3
Set1	100	55	15
Set2	100	55	NA
Set3	55	15	NA
Set4	100	NA	NA
Set5	55	NA	NA
Set6	15	NA	NA

B) Smodel での隠れ層および隠れノードの組み合わせ.

	HIDDEN1	HIDDEN2
Set1	55	15
Set2	15	10
Set3	55	NA
Set4	15	NA
Set5	10	NA

３．２．３　NeProc のデータフローと学習の詳細

図 14 に Smodel の First NN section および Second NN section のデータフローを示す. NeProc の入力はタンパク質のアミノ酸配列である. 入力された配列から PSI-BLAST[61]を用いて位置特異的スコア行列（Position Specific Scoring Matrix：PSSM）が作成される. PSI-BLAST は E-value の閾値が 0.001 の条件下でクエリ配列を UniRef90 データベースへ 3 回相同性検索を実行し PSSM を作成する. 作成した PSSM は 44 次元のデータであるが, NeProc では各アミノ酸に対するスコア 20 次元と, 配列の各サイトの情報 1 次元の合計 21 次元（図 15）を特徴量とした.

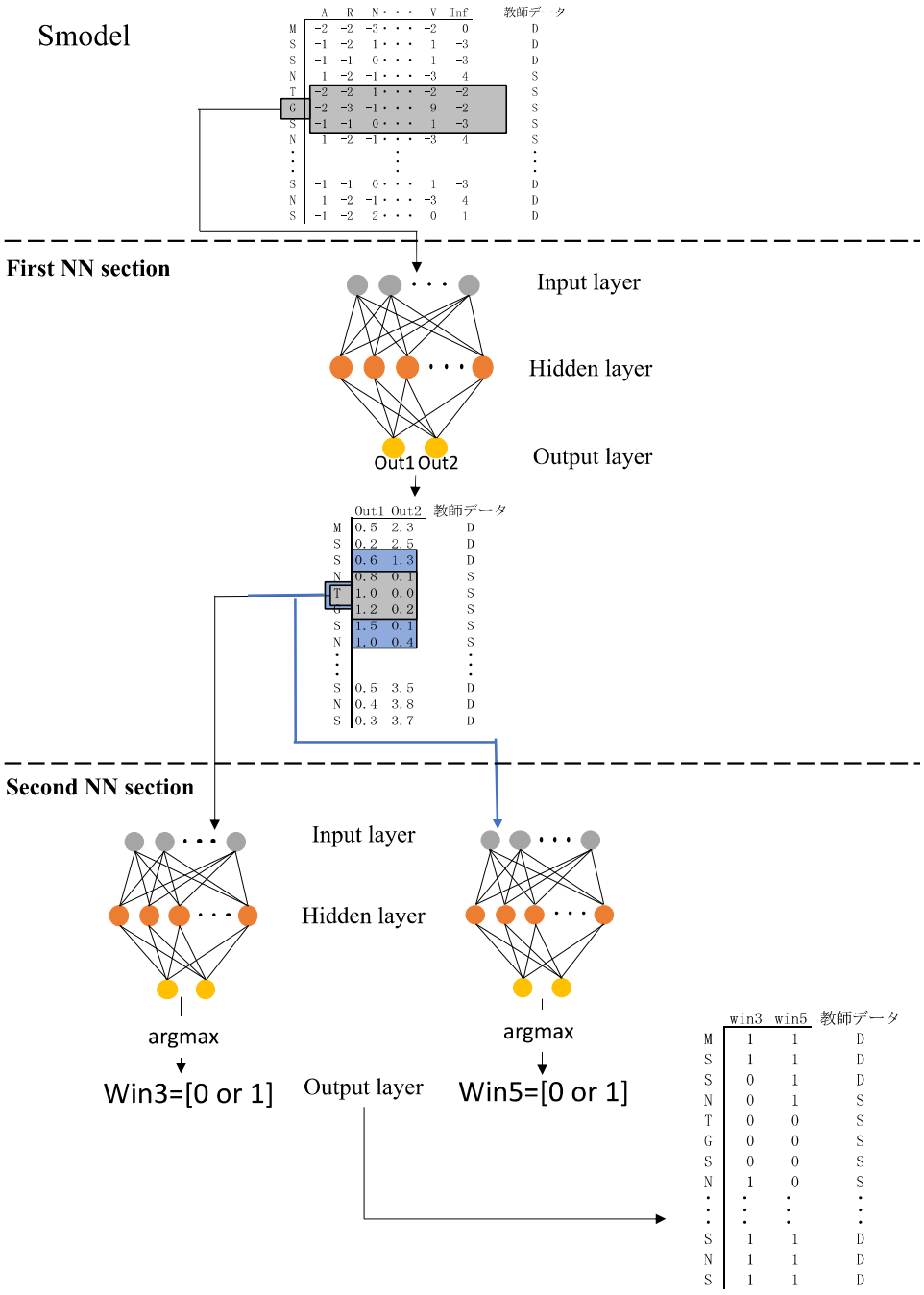


図 14. Smodel の First NN section および Second NN section でのデータフロー.

NeProc では全てのニューラルネットワークおよびサポートベクターマシンのハイパーパラメータ（隠れ層およびそのノードの数）の組み合わせとして，表 4 に示した組み合わせを試した．これらのモデルの weight と bias（パラメータ）の最適化

には「データセット」の項で述べた 3,347 個のタンパク質を用いた．このようにして得られた各モデルの性能評価を,残りの 824 個のタンパク質を用いて行なった.NeProc モデルは図 13 に示すように大きく分けて 3 つのセクションからなるが，まず，第一セクションである First NN section で最適なモデルを選ぶ．その後，この最適な First NN section の出力を入力値し，Second NN section の最適なモデルを選び，最後に最適な Second NN section 出力を用い SVM の最適化を行なった．この点で，First NN section，Second NN section および SVM section の学習は独立している．以下に各セクションでのモデル構築の詳細を述べる.

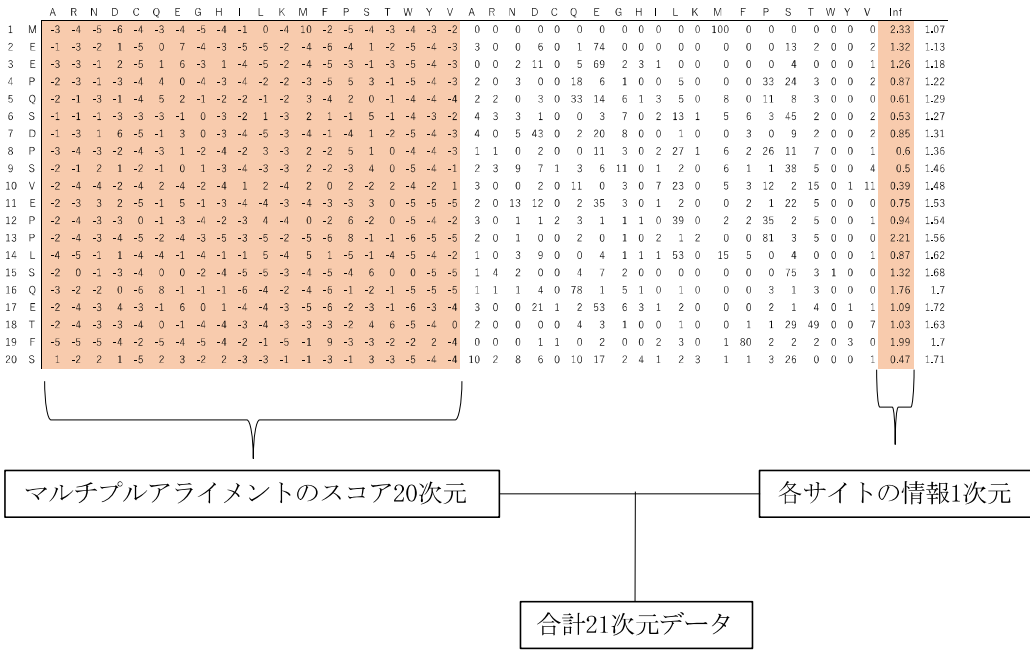


図 15. NeProc で入力する PSSM 情報.

図 14 の First NN section のニューラルネットワークの入力データは，図１５に示した 21 次元の PSSM データである．First NN の weight，bias，隠れ層の数およびそのノード数は図１６の示した手順で決定した．まず，上記の 3,347 個のタンパク質からなる学習データから作成した PSSM データを用いて，表 4B の 5 つのハイパーパラメータの Set からなる 5 つのニューラルネットワークのトレーニングを行う(図 16A)．次に 842 個のタンパク質から作成した PSSM データを用いて，5 つのニューラルネット

ワークをテストし最も精度が高いモデルを決定する(図 16B). このようにして得られた first NN の出力は, second NN のへ入力となる. 出力値はニューラルネットワークの output node の生の値である. 生の値とは ReLU 関数や Softmax 関数などの活性化関数へ入力する前の 2 次元の実数値である(図 14”output1”,”output2”).

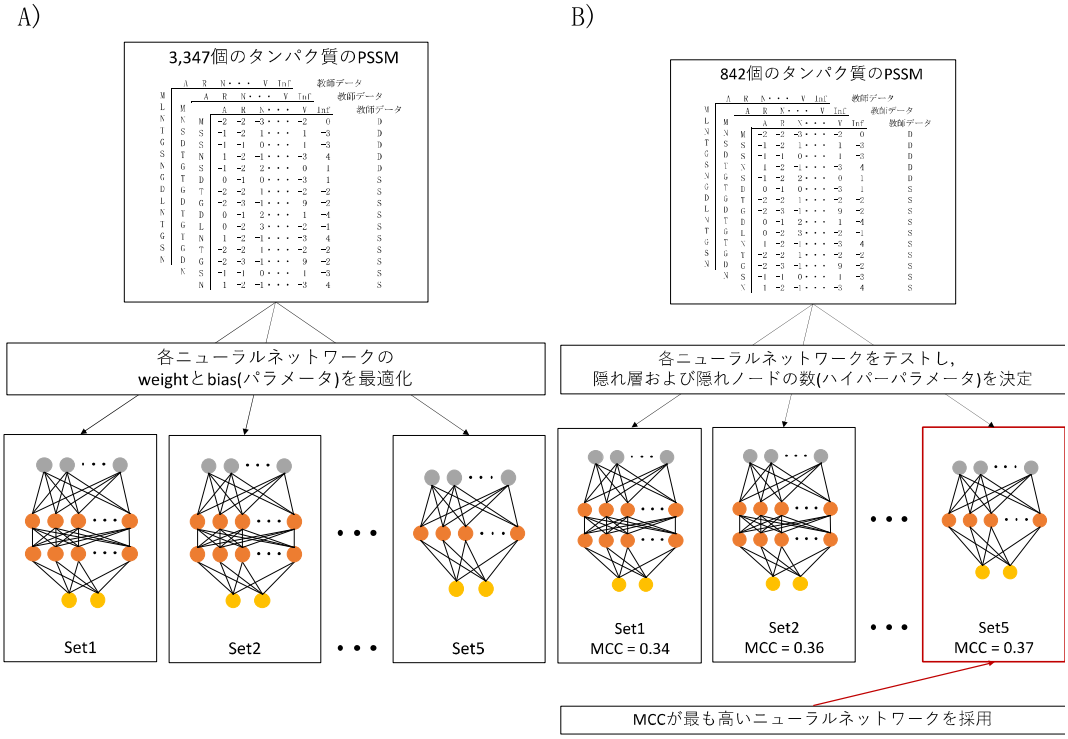


図 16. First NN section でのニューラルネットワークのパラメータの最適化とハイパーパラメータの決定. A がパラメータの最適化, B がハイパーパラメータの決定方法を示している.

Second NN section の構成を決定するにあたり, window サイズとして 3, 5, 10 を試した. これらのネットワークに関して, ハイパーパラメータとして表 4B の 5 つのハイパーパラメータを試した. すなわち, 3 つの window サイズ, 5 つのハイパーパラメータ, からなる各々異なる 15 個のニューラルネットワークを生成した(図 17A). この 15 個の bias と weight を決定するため, First NN のパラメータの最適化でも用いた 3,346 個のタンパク質を用いた. ただし, この場合の入力値は, 上記で決定した First NN の出力値である. すなわち, 3,347 個のタンパク質から生成した PSSM データを First NN に入力し得た 3,347 個の出力値である. 次にこの 15 個のネットワークの

精度を評価するため，842 個のタンパク質を用いて精度を評価した（図 17B）．この評価では，採用した window サイズの中での評価を行なった．すなわち，5 つある window サイズ 3 のモデルの中で最高精度のもの，同様に 5 個ずつある window サイズ 5，10 のモデルの中で最高精度のものを選んだ．この段階で，各 winodow サイズで最高精度のモデル 3 つが選ばれた．ここで決定した 3 つのニューラルネットワークから，Second NN section のニューラルネットワークとしてこれら 3 つのモデルの組み合わせを考えた．どの組み合わせを用いるかの評価は次の SVM section で行った．Second NN section の出力値は argmax 関数の出力値である 0 または 1 の 1 次元データである．argmax 関数は以下の式で表され，入力された output node のうち最大値である node の node 番号を出力する．

$$output\ node = [node0, node1]\text{の時}\quad argmax(output\ node) = \begin{cases} 0\ if\ (node0 \geq node1) \\ 1\ if\ (node0 < node1) \end{cases}$$

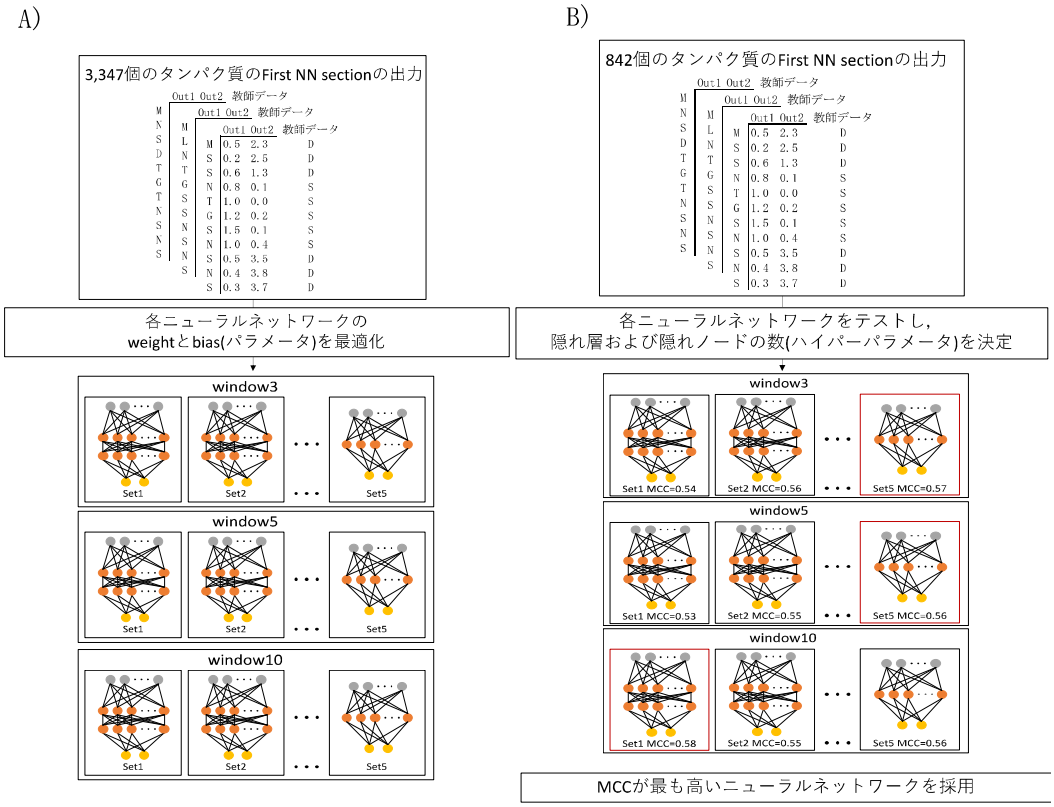


図 17. Second NN section でのニューラルネットワークのパラメータの最適化とハイパーパラメータの決定．A がパラメータの最適，B がハイパーパラメータの決定方法を示している．

SVM section では Second NN section の各 window サイズの出力値を組み合わせたデータを入力とする． 予備調査の結果，1 つのモデルのみを用いた結果より，2 つ以上のモデルを用いた場合の方が結果が良かったため，Second NN section で選択した 2 つ以上のモデルの組み合わせ，4 通りを試した．Second NN section の出力は 1 次元であるため，組み合わせるモデルにより SVM section への入力値は，3 つの出力を組み合わせた場合 3 次元，2 つの出力を組み合わせた場合は 2 次元となる．SVM モデルの構築には，3,347 個のタンパク質のデータを First NN および Second NN を通した結果として得られる Second NN の出力値を用いた(図 18A)．このようにして得られた 4 つの SVM モデルの性能を評価するため，残りの 842 個のタンパク質を用いて精度を評価した（図 18B)．この評価の結果，Second NN section の window サイズ 3，5 の

モデルを組みわせたものが最も高い精度を示した.

Lmodel においても同様のデータフローおよび学習手順である. NeProc では全てのニューラルネットワークのパラメータの初期値には He の初期値[88]を用いて初期化し，パラメータの更新のためのオプティマイザーには adaptive moment estimation (Adam) [89]を用いた. Adam のパラメータである学習率，1 次指数関数的減衰率，2 次指数関数的減衰率をそれぞれ 0.001，0.9，0.999 を用いた. 活性化関数には正規化線形活性化（ReLU）関数を採用し学習時の誤差の計測には Softmax 関数および交差エントロピー誤差を用いた. サポートベクターマシンは Sikit-learn ライブラリーの linearSVC を使用して, コストパラメータを 0.1 と 0.5 および 1 から 10 まで 1 刻みに変更し，他のパラメータはデフォルト値を用いた.

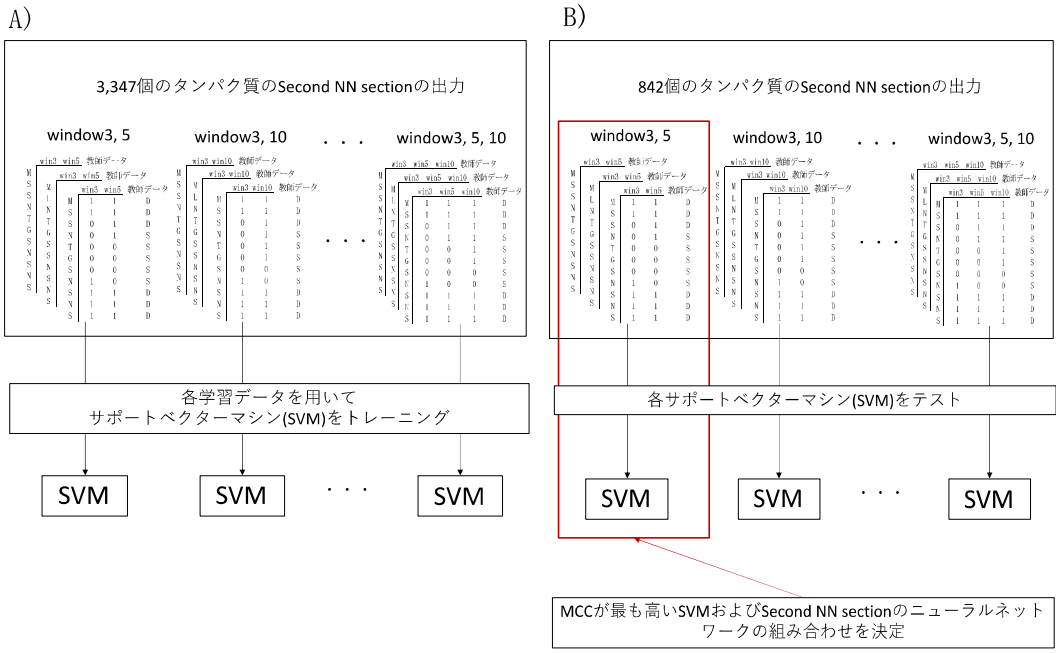


図 18. SVM section での SVM のトレーニング(A)と SVM section での SVM および Second NN section でのニューラルネットワークの組み合わせの決定(B).

3．2．4　性能評価

天然変性領域予測では天然変性領域を positive, 構造領域を negative として扱い，以下に示す 4 つの評価指数を用いて予測精度を評価する. それに対して天然変

性領域中の機能部位予測では予測の正誤を単純に判断できない場合がある．天然変性領域には，現時点で証拠が存在しないだけで未知の機能部位が含まれている可能性がある．つまり，正解ラベルが天然変性領域である残基を機能部位と予測した場合に，その予測を誤りと断定することが難しい．ANCHOR2 ではこの問題を回避するために，正解が天然変性領域中の機能部位とラベル付けされた残基と，構造領域とラベル付けされた残基を識別できるかを評価している[6]．IDEAL データセットに対する機能部位予測では本章においても ANCHOR2 の評価法を採用し，天然変性領域中の機能部位を正しく機能部位と予測できた場合 true positive,構造領域を正しく構造領域と予測できた場合 true negative,構造領域を誤って機能部位と予測した場合を false positive,機能部位を誤って構造領域と予測した場合を false negative，として予測精度の評価を行なった．天然変性領域を機能部位と予測した場合は正誤の判定を行なわず，以下に示す評価指数には反映されない．

本章では sensitivity，precision，F-score，マシューズ相関係数(MCC)の 4 つの評価指数を用いてモデルの性能を評価した．

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - score = 2\frac{Sensitivity \times Precision}{Sensitivity + Precision}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

式中の TP，TN，FP，および FN は，それぞれ真陽性，真陰性，偽陽性，および偽陰性を

表している.

また, NeProc の予測精度と比較対象のプログラムの予測精度に対して統計的有意性を次の手順で評価した.

Step1. テストデータより 80%のタンパク質をランダムにサンプリングし上記の４つの評価指数を計算する.

Step2. NeProc と比較対象のプログラムで 4 つの評価指数について差を求める.

Step3. Step1 および Step3 を 5,000 回実行し 4 つの評価指数について 5,000 個の差のデータを作成する.

Step4. 得られた 5,000 個の差について有意差を 0.05 としたシャピロ-ウィルク検定[90]を用いて, 正規分布に従うか確認する.

Step5. Step4 の結果が正規分布に従う場合は 2 標本 t 検定を, 正規分布に従わない場合はウィルコクソン符号順位検定[91]を用いて統計的有意性を評価する.

３．３　結果

３．３．１　天然変性領域予測の予測精度

NeProc は予測された天然変性領域中にある構造領域様な短い領域を機能部位として予測する. 従って天然変性領域予測の精度は, 機能部位の予測にとって重要な要素となる. これは比較対象プログラムにおいても, 機能部位の予測に天然変性領域予測の結果を用いていることから, 全ての機能部位予測プログラムにおいて同一のことがいえる. そこで, まず NeProc の天然変性領域予測精度の評価を行った. 表 5 は, 各テストデータセットにおける NeProc と比較対象プログラムの予測精度を示した表

である.

表 5. 天然変性領域予測結果.

	CASP10				IDEAL			
	MCC	Sensitivity	Precision	F-score	MCC	Sensitivity	Precision	F-score
NeProc	0.587	0.516	0.716	0.600	0.671	0.723	0.739	0.731
SPOT-disorder	0.542**	0.483**	0.664**	0.559**	0.665**	0.699**	0.753**	0.725**
DISOPRED3	0.536**	0.413**	0.748**	0.532**	0.630**	0.664**	0.728**	0.695**
IUpred2-short	0.270**	0.303**	0.325**	0.313**	0.557**	0.538**	0.730**	0.620**
IUpred2-long	0.165**	0.176**	0.247**	0.206**	0.538**	0.576**	0.663**	0.617**
Mobi-DB-lite	0.163**	0.037**	0.789**	0.071**	0.508**	0.364**	0.872**	0.514**

	CheZOD			
	MCC	Sensitivity	Precision	F-score
NeProc	0.536	0.712	0.921	0.803
SPOT-disorder	0.554**	0.726**	0.925**	0.813**
DISOPRED3	0.439**	0.525**	0.946**	0.675**
IUpred2-short	0.435**	0.639**	0.893**	0.745**
IUpred2-long	0.490**	0.715*	0.893**	0.794**
Mobi-DB-lite	0.398**	0.476**	0.94**	0.632**

MCC, Matthews correlation coefficient. * p-value < 1.0×10^{-3} and ** p-value < 1.0×10^{-5} in comparison with NeProc.

全体的に NeProc, SPOT-disorder および DISOPRED3 の予測精度は高い傾向がある. NeProc は CASP10 と IDEAL において最も高い MCC と F-score を記録している. IUpred2A と MobiDB-lite は比較的予測精度が低いが, 各々特徴がある. IUpred2A は２つの予測モデルがあり,CheZOD に対する予測では DISOPRED3 を上回り NeProc と SPOT-disorder に迫る予測精度を記録している. MobiDB-lite は CASP10, IDEAL および CheZOD 全てのデータセットにおいて高い Precision を記録している. これは MobiDB-lite が偽陽性を抑制していることを示している.

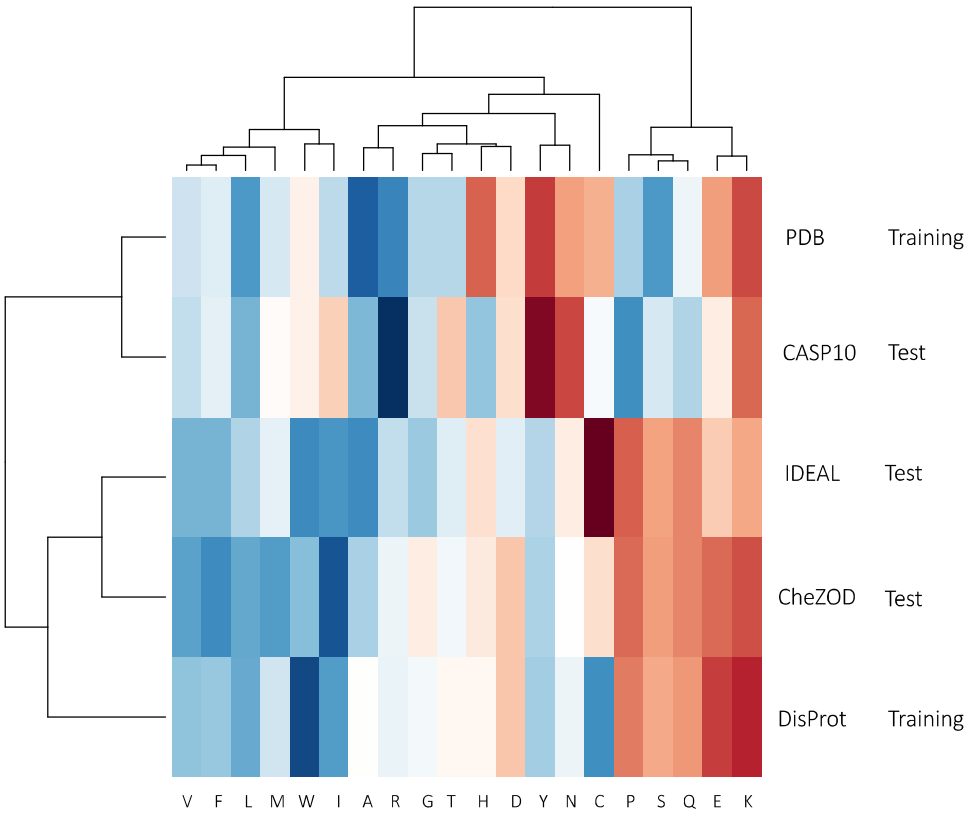


図 19. 各データセットの組成．データセット間の距離は、参照データセットである TrEMBL（UniProt リリース 2019_11）のアミノ酸組成に対する対数オッズによって算出した．赤い色は参照データよりも豊富なアミノ酸を表し，青い色は参照データよりも少ないアミノ酸を示している．色の明るさは参照データにどの程度，似ているまたは異なるかを表している．

各データセットでは含まれている天然変性領域に違いがある．CASP10 は X 線結晶構造解析によって構造が決定されたタンパク質が収録されており，CASP10 では構造内の欠落したアミノ酸残基を天然変性領域としている．IDEAL では X 線結晶構造解析，NMR，円偏光二色性測定などの様々な物理化学的手法によって実験的に決定された天然変性領域を収録している．CheZOD では NMR によって構造決定されたタンパク質が収録されている．既存の天然変性タンパク質データベース間では天然変性タンパク質のアミノ酸組成などが異なっていると報告されているが[92]，本章で用いたデータセット間でもアミノ酸組成はわずかではあるが異なっていた（図 19）．含まれる天然変性タンパク質に差がある 3 つのテストデータセットに対して NeProc， SPOT-disorder および DISOPRED3 は高い予測精度を達成した．これらの結果は NeProc の天然変性領域

予測を用いて機能部位予測を行うことに問題がないことを示している.

３．３．２　天然変性領域中の機能部位予測の精度

IDEAL データセットに対する NeProc の機能部位予測の精度を比較対象プログラムである，DISOPRED3，ANCHOR2，MoRFchibi-Web の結果とともに示した(表 6)．４つのプログラムの中で，NeProc は MCC，precision，F-score において最も高い値を達成したが，sensitivity では ANCHOR2 が最も高い値であった．NeProc の sensitivity は ANCHOR2 と比較して 0.01 低いが，precision では 0.013 高かった．これは NeProc が ANCHOR2 と比較して，わずかに偽陰性が多く，わずかに偽陽性を抑制していることを示唆している．ANCHOR2 は予測モデルを構築する際に DIBS データベース[44]より獲得したデータを学習している．DIBS データベースは天然変性領域中の結合領域を提供しているデータベースであり,IDEAL データと共通のタンパク質が 191 個含まれている．本研究では IDEAL データセットを用いて精度を測定しているため，ANCHOR2 は IDEAL データセットのタンパク質の一部を学習していると考えられる．それにもかかわらず，NeProc は天然変性領域中の機能部位のデータを学習せずに ANCHOR2 と同等の予測精度を達成した．

表 6. 機能部位予測結果.

	MCC	Sensitivity	Precision	F-score
NeProc	0.388	0.487	0.358	0.413
ANCHOR2	0.381**	0.497**	0.345**	0.408**
MoRFchibi-Web	0.196**	0.221**	0.249**	0.234**
DISOPRED3	0.175**	0.171**	0.233**	0.198**

MCC, Matthews correlation coefficient. * p-value $< 1.0 \times 10^{-3}$ and ** p-value $< 1.0 \times 10^{-5}$in comparison with NeProc.

３．３．３　UniProt データベースからの pProS 抽出と pProS データセットにおける機能部位予測精度

表 7. UniProt の機能情報の統計.

	All proteins	pProS	%pProS	pProS uniq
No. proteins	20,410	1,529	7.5%	1,529
No. annotations shorter than 30 residues	29,145	3,031	10.4%	2,942
"region of interest"	4,646	425	9.1%	411
"mutagenesis site"	21,269	1,198	5.6%	1,147
"short sequence motif"	3,230	1,408	43.6%	1,384

UniProt から獲得した 20,410 個のヒトタンパク質から 29,145 領域の 30 残基以下の機能情報を持つ領域が得られた(表 7). そのうち 4,646 領域が“region of interest”, 21,269 領域が“mutagenesis site”および 3,230 領域が“short sequence motif”であった. 20,410 個のタンパク質のうち 7.5%の 1,529 タンパク質が pProS を保持しており, 機能情報を持つ 29,145 領域のうち 10.4%の 3,031 領域が pProS として抽出された.機能情報ごとでは“region of interest”が 425 領域(9.1%),“mutagenesis site”が 1,198 領域(5.6%)および“short sequence motif”が 1,408 領域(43.6%)であった.

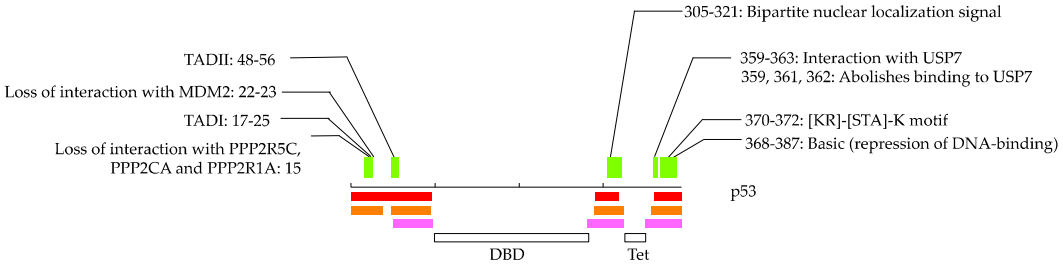


図 20. p53 における pProS の例. 中心の黒線はアミノ酸配列を表しており、天然変性領域の予測をその下に示した. ピンク, オレンジ, 赤は, それぞれ MobiDB-lite, DISOPRED3, DICHOT による予測結果を表している. 2 つの方法のいずれかが天然変性領域と予測する領域を天然変性領域として定義している. 緑の帯は pProS を表し, pProS の注釈を, 注釈の残基番号とともに上に示している. DBD:DNA 結合ドメイン, Tet:四量体化ドメイン.

本章で定義した pProS を持つタンパク質のうち一部は IDEAL データベースで

登録されているタンパク質もあり，その場合は参考として IDEAL データベースでのアクセッション番号を示している．p53 での pProS の例を図 20 に示した．p53 は典型的な天然変性タンパク質の１つであり DNA 結合ドメインと４量体形成ドメインの間，C 末端領域および N 末端領域が天然変性領域である（IDEAL：IID00015）．p53 には機能情報が残基番号 15～25 の領域に３つ，48～56 の領域に１つ，305～321 の領域に 1 つ，359～363 の領域に４つ，370～372 の領域に１つ，368～387 の領域に１つ，合計 11 個存在した．これらの機能情報のうち“TAD I”，“TAD II”，“Bipartite nuclear localization signal” および“[KR]-[STA]-K motif”は“short sequence motif”から得られ，“Interaction with UPS7”および“Basic”は“region of interest”から，“Loss of interactions with MDM2”，“Loss of interaction with PPP2R5C….”および“Abolishes binding to UPS7”は“mutagenesis site”から得られた．これらの機能情報の一部は互いに重複している場合がある．例えば，“Loss of interactions with MDM2”と“Loss of interaction with PPP2R5C….”は重複しているがこれらは異なる機能情報を示している．このような場合，表 7 では異なる pProS 領域として個別に集計している．一方，pProS 残基を集計する場合(表 3)は領域が重複していても，同一の残基として集計している．例えば，p53 において残基番号 370～372 の領域長が３の領域と 368～387 の領域長が 20 の領域は異なる機能情報を持つが pProS 残基としては 20 残基とカウントする．その結果，pProS データセットは 1,518 領域，12,148 残基の pProS からなるテストデータセットとなった(表 3)．

また，本章で抽出した機能情報の多くは，データセット内で１回のみ出現した．例えば p53 では“Interaction with protein A”や“Loss of interaction with protein B”などの特定のタンパク質との結合に関する情報であった．一方，核局在化シグナル（NLS），SH3 ドメイン，PDZ ドメインおよび核内受容体のコリプレッサーまたはコアクチベーターに存在する LxxLL モチーフなどの機能情報は pProS に頻繁に見られる（表 8）．網膜芽細胞腫関連タンパク質は，残基番号 858～881 の領域が NLS であり，この領域は単独では天然変性領域であるが[93]，インポーチンとの結合構造が解

明されている[94]（IDEAL: IID00017). SH3 ドメインは短い曖昧なモチーフを介してタンパク質間相互作用を仲介する機能ドメインであり[95], IDEAL データベースで ProS として登録されている（IDEAL：IID00256).PDZ ドメインはシグナル伝達や細胞輸送においてタンパク質複合体を形成する足場として機能するタンパク質に見られるドメインであり[96], 機能部位は単独で天然変性領域である[97]（IDEAL: IID90005). また，核内受容体の一つであるペルオキシソーム増殖因子活性化受容体γコアクチベーター1αの LxxLL モチーフは単独では天然変性領域であると報告されているが，ステロイドホルモン受容体との結合構造が解明されている[98, 99]（IDEAL: IID00103).

表 8. pPorS に多く見られる機能情報.

feature name	#counts	#proteins	description
Nuclear localization signal	346	303	sequence targeting nucleus
SH3-binding	74	51	SH3 domain binding
PDZ-binding	74	74	PDZ domain binding
Cell attachment site	63	30	cell adhesion related sequence
Prevents secretion from ER	49	49	short segments preventing secretion
LXXLL motif	47	24	essential for interaction with nuclear reveptors
Nuclear export signal	34	30	sequence exporting from nucleus
Bipartite nuclear localization signal	34	32	sequence targeting nucleus
ITIM motif	33	22	associates with the two phosphatases
PPxY motif	27	21	phosphorylation site
Microbody targeting signal	27	27	sequence targeting microbody
YXXM motif	23	3	Interacts via phosphorylated
SH2-binding	18	8	SH2 domain binding

UniProt データベースより抽出した pProS データセットに対する機能部位予測の結果を表 9 に示す. 全ての予測プログラムが IDEAL データを用いた機能部位予測(表 6)より予測精度が向上している． NeProc は MCC において 0.588 と非常に精度が高くい結果となった. また NeProc,ANCHOR2 および DISOPRED3 は sensitivity が高く，pProS を構造領域と予測する偽陰性を抑えられているかつ，多くの pProS を識別できたことを示している． Precision においては MoRFchibi-Web が高く偽陽性を抑えられている.

表 9. pProS データセットでの機能部位予測精度.

	MCC	Sensitivity	Precision	F-score
NeProc	0.588	0.896	0.421	0.572
ANCHOR2	0.569**	0.755**	0.481**	0.587**
DISOPRED3	0.418**	0.626**	0.307**	0.411**
MoRFchibi-Web	0.365**	0.282**	0.565**	0.376**

MCC, Matthews correlation coefficient. * p-value < 1.0×10^{-3} and ** p-value < 1.0×10^{-5}in comparison with NeProc.

３．４　考察

３．４．１　天然変性領域中の結合領域の長さと予測精度の関係

IDEAL データセットは 9,444 残基からなる 398 個の機能部位を含んでいる. この 398 領域について NeProc による予測精度と領域長に関係性があるか分析した. 図 21 は横軸に機能部位の長さを, 縦軸に sensitivity を示し sensitivity の中央値を三角で, 平均値をアスタリスクで示している. IDEAL データセット全体での sensitivity は 0.487(表 6)であったが, 10 残基から 50 残基の機能部位では sensitivity が概ね 0.6 と 10 残基未満の短い結合領域および 50 残基以上の比較的長めの機能部位と比較して高い値であった. また, 10 残基から 50 残基の機能部位では平均より中央値の方が高い値であった. これは, その長さの機能部位を正しく機能部位と予測できており, 誤って構造領域と予測される機能部位が少ないことを示している. Sensitivity の分布を示したバイオリンプロット(グレーの領域)においても同様の傾向を見ることができる. この結果は, NeProc が長い天然変性領域中の構造領域的性質が見られる短い領域をターゲットに機能部位を予測していることを反映しており, 10 残基から 50 残基の中央値が高いことは, NeProc が多くの短い機能部位を予測できたことを示している. これは IDEAL データセットでは 50 残基以下の機能部位が 90%以上を占めていることに起因すると考えられる(図 10). しかし, 機能部位が数残基から数十残基であることを考慮すると, 比較的短い機能部位の予測において NeProc は有効な予測プログラムである可能性がある.

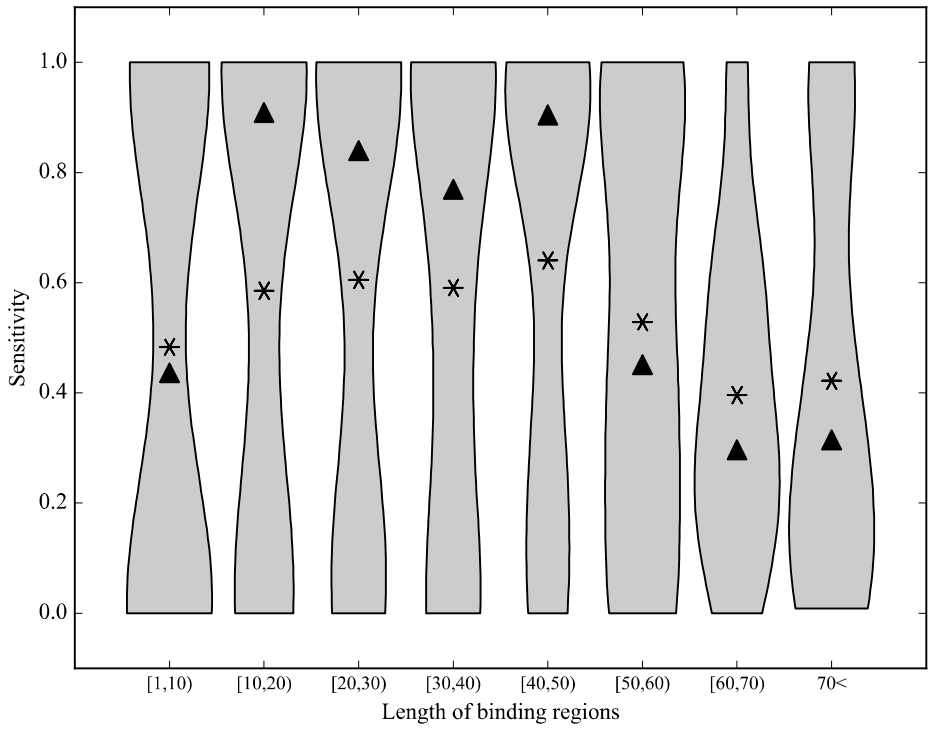


図 21. Sensitivity と機能部位の長さの関係．横軸は機能部位の長さを，縦軸は sensitivity を示している．アスタリスクは各長さの Sensitivity の平均値を，三角は中央値を示しており，グレーのエリアは sensitivity の分布を示している．

３．４．２　天然変性領域中の機能部位の２次構造と予測精度

一般的に機能部位はパートナーと出会うと局所的な 2 次構造を形成し相互作用する．その際の機能部位が形成する 2 次構造によって，予測精度が異なるかを解析した(表 10)．表 10A は機能部位中のα-helix，β-sheet および coil 構造を形成するアミノ酸残基についての sensitivity を示している．表 10A での sensitivity の計測方法を図 22A に示した．機能部位を構成するアミン酸残基の PDB における 2 次構造情報から，helix 残基，sheet 残基および coil 残基を決定する．そして各残基に対する NeProc の予測から各 2 次構造を形成する残基の sensitivity を計測する．その結果，coil 残基の sensitivity が最も高く，次いでα-helix 残基，β-sheet 残基と続く．表 10B は機能部位を形成する２次構造に基づき分類し，各クラスに含まれる機能部位

の sensitivity の平均を示している．図 22B には表 10B での sensitivity の計測方法を示した.H クラスは領域中に α-helix を含んでいる機能部位を,S クラスはβ-sheet を，H&S はα-helix とβ-sheet を，C クラスは α-helix とβ-sheet を全く含まない機能部位を示している．全ての機能部位の sensitivity を計測したのち，クラスごとに平均を求めた.

表 10. 機能部位予測の２次構造依存性.

A) 機能部位を構成する２次構造の残基ごとの精度.

	Helix	Sheet	Coil
Sensitivity	0.434	0.341	0.530

B) 機能部位を含んでいる 2 次構造で分類した場合の領域ごとの精度.

	H	S	C	H&S
Sensitivity	0.587	0.643	0.481	0.421

C) 機能部位を構造領域，天然変性領域，機能部位および不明な領域と予測した割合.

	Structured	Disorderd	Binding regions	Unknown	Average length
H	0.340	0.174	0.483	0.003	20.8
S	0.292	0.184	0.524	0.000	13.2
C	0.402	0.221	0.373	0.004	18.5
H&S	0.485	0.161	0.352	0.001	38.2

表 10A と表 10B を比較すると，α-helix を含む H クラスとβ-sheet を含む S クラスは α-helix 残基およびβ-sheet 残基と比較して sensitivity が高くなっている．一方，coil 残基の sensitivity は最も高い値であったが（表 10A），coil のみを含む C クラスでは値が低下した．これは，coil のみからなる機能部位の coil 残基よりα-helix と coil またはβ-sheet と coil から構成される機能部位中の coil 残基の予測精度が高いことを示している．近傍にα-helix などの２次構造が存在する場合，NeProc は window をとって予測を行なっているので近傍の coil 残基にも構造領域的傾向があると判断している可能性がある．つまり NeProc は相互作用時にα-helix またはβ-sheet を形成する機能部位予測に適している傾向がある．また，α-helix とβ-sheet の双方を含む機能部位は低い値であった（表 10B）． 表 10C は，２次構造クラスにおける予測の傾向と，各２次構造クラスに属する機能部位の平均の長さを示してい

る．H&S クラスは表 10B で最も sensitivity が低く，構造領域と予測される割合が高い(表 10C)．また H&S クラスは平均の配列長が４つのクラス中で最も長い．この機能部位の長さがH&Sクラスの低いsensitivityの原因の１つである可能性が考えられる．

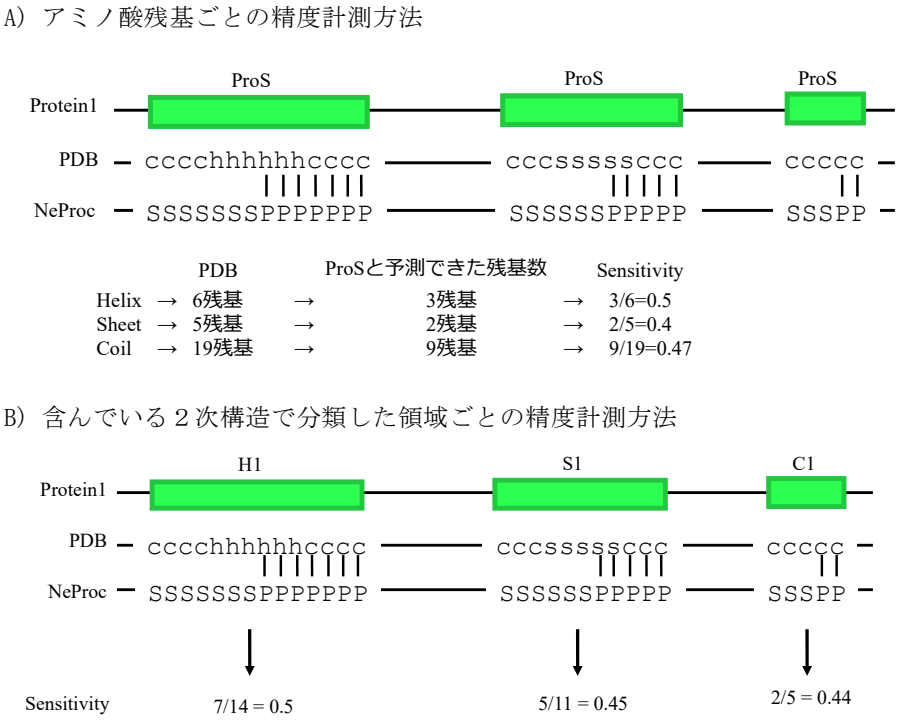


図 22. 表 10 での予測精度計測方法. ProS は機能部位を表している. H1 は H クラス，S1 は S クラス，C１は C クラスを表している.

３．４．３　天然変性領域中の機能部位と予測された領域のアミノ酸組成

NeProc は，機能部位の情報を学習せずに構造領域と天然変性領域の情報のみを学習することで，天然変性領域中の機能部位を予測している．この結果は，NeProc が天然変性領域中の機能部位と構造領域の双方で共通の特徴を識別していることを示唆している．そこで，機能部位と構造領域の配列間にどのような共通性があるか，または，機能部位と天然変性領域の配列間がどのように異なるかを分析した．図 23 には機能部位の組成と構造領域および天然変性領域の組成の類似度を示した（類似度の計算は補足説明 1 を参照)．赤い円は正しく機能部位と予測された機能部位を示してい

る．青い円は誤って構造領域と予測された機能部位を示している．水色の円は天然変性領域と予測された機能部位を示している．灰色の円は機能部位全体を示している．図中の対角線付近に円が来ている場合，その機能部位の組成は天然変性領域と構造領域の中間のアミノ酸組成を示している．また円が縦軸へ近づくと組成が構造領域と似ていることを，横軸と近づくと組成が天然変性領域と似ていることを示している．短い window サイズでは，機能部位全体の組成（灰色）は対角線付近に位置しており，window サイズが長くなるにつれて下部方向へ移動する．この傾向は，全てのクラスで同一であるが，機能部位が天然変性領域中に存在する短い領域であることから，window サイズを大きくすると機能部位周辺の天然変性領域が window の範囲に含まれるため，当然の傾向であると言える．４つのクラスのうち正しく予測できたクラスはこの傾向が顕著に見え，NeProc はこの傾向を識別することで機能部位を予測している可能性がある．天然変性領域と予測されたクラスは長い window サイズでは天然変性領域に近い組成を示しているが，短い window サイズの場合でも対角線付近に位置しないことから，Smodel において構造領域的性質を識別することが難しいと考えられる．一方，構造領域と予測される機能部位は長い window サイズにおいても対角線付近に位置していることから，Lmodel において天然変性領域と予測することが難しいと示唆される．この問題を解決するためには Smodel，Lmodel の予測精度の向上が必須となる．

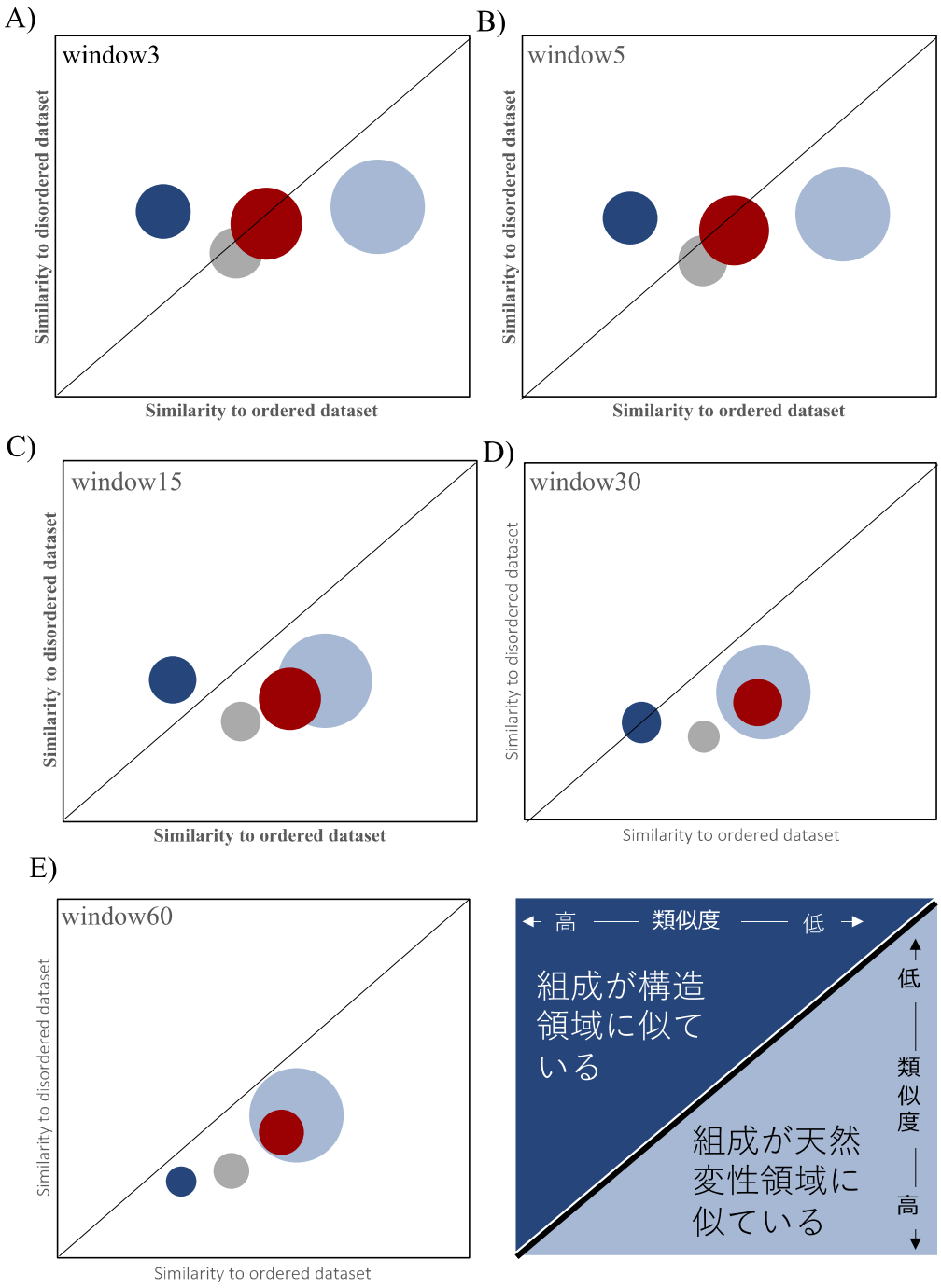


図 23. 機能部位のアミノ酸組成と天然変性領域および構造領域のアミノ酸組成の類似性. 横軸は学習データ内の構造領域との類似性を示し、縦軸は天然変性領域との類似性を示している. 赤い円は正しく機能部位と予測できた機能部位を, 青い円は誤って構造領域と予測された機能部位を, 水色の円は誤って天然変性領域と予測された機能部位を灰色の円はテストデータセットの全ての機能部位を表している. 円の中心は各アミノ酸組成の平均から構造領域の組成および天然変性領域の組成までの距離を示している. 円の面積は分散を表している. 組成は window ごとに示されており, 3, 5, 15, 30 および 60 はそれぞれ A), B), C), D) および E)に示されている.

３．４．４　NeProc と MoRFchibi-Web の予測精度の差について

MoRFchibi-Web は最近のテストにおいて高い予測精度を記録しているが[53, 77], IDEAL データセットに対しては，あまり精度が良くなかった．これは NeProc と MoRFchibi-Web で機能部位予測における戦略が異なることが原因の１つとしてあげられる． MoRFchibi-Web は機能部位とその他の領域の２つに識別する．その他の領域には天然変性領域も含まれており，未知の機能部位が含まれている可能性がある．実際に pProS の例では IDEAL データベースには存在していない領域を抽出した．このようにタンパク質中には未知の機能部位が含まれている可能性が高く本研究では未知の機能部位の存在を排除できないと判断した．そのため ANCHOR2 と同様の評価方法である，機能部位と構造領域の識別を重視した方法を採用した．その結果，NeProc と ANCHOR2 は pProS データセットのような未知の機能部位を含んだテストにおいても，機能部位を予測することが可能である．しかし，評価することができない予測機能部位も多数存在し一概に性能が良いかを判断することは難しい．一方，MoRFchibi-Web は未知の機能部位は考慮せずに，現時点において既知である機能部位を識別できるかを重視している．そのため予測する機能部位の数は少ないが他の予測プログラムと比較して偽陰性を抑えられている(表 9)．現状どちらの予測法が優れているかは判断することは難しい．

３．４．５　pProS データセットでの機能部位予測精度の向上

本章では UniProt データベースの機能情報と実用精度を達成している天然変性領域予測プログラムを用いて pProS を同定した．UniProt の機能情報は実験的に検証された情報であるため信用できると言える．また，天然変性領域の決定には３つの予測プログラム，DISOPRED3, DICHOT および MobiDB-lite を用いた．MobiDB-lite は 8 個の天然変性領域予測プログラムの予測結果に対してコンセンサスをとるため，実質 10 個の予測モデルを用いて天然変性領域を決定したことになる．これは非常に厳しい条件ではあるが，その分予測された天然変性領域の信頼性は高いと言える．このように確度の高い機能情報と厳しい条件での天然変性領域予測を用いて 1,500 領域を超え

る未同定の ProS 候補領域を抽出した．本章ではこの pProS データセットを用いて NeProc および比較対象プログラムの予測精度を測った．その結果，IDEAL データセットに対するテストと比較して全ての予測プログラムの予測精度が向上した．特に NeProc は MCC が 0.588 と非常に高く，未知の機能部位の予測において有効なプログラムである可能性がある．ただし，この予測精度の向上は pProS の同定方法に起因していると考えられる．上記で述べたが pProS は天然変性領域予測を用いて同定した．そして各機能部位予測プログラムは内部で天然変性領域予測を特徴量に用いている．つまり，pProS を含む領域の天然変性領域予測が容易であったと考えられる．その結果，予測精度が向上したと考えられる．つまり，先程述べた NeProc の有効性は正確ではなく“天然変性領域予測が容易な領域の未知の機能部位に対して有用なプログラムである”と言い換える必要がある．

３．４．６　pProS データセットの可能性

表 11. モデル生物における予測天然変性領域中の機能情報.

Organism	No. proteins	No. annotations shorter than 30 residues	pProS-containing proteins	pProS uniq	"region of interest"	"short sequence motif"	"mutagenesis site"
Arabidopsis thaliana	15,347	4,347	184	1,664	43	1,581	67
Bos taurus	5,999	1,619	140	1,427	364	1,131	2
Caenorhabditis elegans	3,917	687	23	136	7	115	14
Danio rerio	3,001	649	50	403	0	403	5
Dictyostelium discoideum	4,129	627	19	234	43	182	9
Drosophila melanogaster	3,407	805	54	465	45	408	24
Gallus gallus	2,287	734	66	539	51	479	14
Mus musculus	16,847	5,937	527	5,349	1,002	4,435	218
Oryza sativa subsp.japonica	3,825	1,054	68	424	10	409	5
Rattus norvegicus	7,966	2,874	250	2,703	441	2,255	90
Saccharomyces cerevisiae	6,721	1,562	66	601	73	495	60
Schizosaccharomyces pombe	5,141	683	12	78	0	75	3

pProS の抽出には DICHOT および MobiDB-lite の 2 つの天然変性予測プログラムを用いた. その際 2 つの予測プログラムが天然変性領域と予測した領域中に機能情報がある場合に pProS とした. No. proteins および pProS-containig proteins はタンパク質数を示しており, その他のカラムは残基数を示している. no annotations は構造情報が不明なアミノ残基の数を示している. “region of interest”, “mutagenesis site” および “short sequence motif” カラムは pProS として抽出されたアミノ酸残基の統計である.

本章ではヒトのタンパク質について解析を行ったが, 実験で検証された機能情報が豊富な生物であれば適用できる. 予備的な結果ではあるが, 代表的なモデル生物, *Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Danio rerio, Dictyostelium discoideum, Drosophila melanogaster, Gallus gallus, Mus musculus, Oryza sativa subsp.japonica, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe* の解析結果を表 11 に示した. ここでは, pProS の抽出方法がヒトプロテオームからの抽出方法(図 11)とは天然変性領域の決定方法が異なる. 表 11 では天然変性領域を DICHOT および MobiDB-lite を用いて決定しており, これら 2 つの予測プログラムが天然変性領域と予測した領域を天然変性領域とした. その結果, ヒトの pProS(表 7)と比較すると少ないが, *Mus musculus, Arabidopsis thaliana, Bos taurus, Rattus norvegicus,* において 1,000 残基を超える pProS を抽出することができた. 各モデル生物の pProS を保持しているタンパク質(表 11 “pProS-containing

protein”)についてUniProtの“cross reference”セクションにあるPDBデータを用いて構造領域を決定し，機能部位予測に用いるデータセットを作成した(表12).

表12. モデル生物における機能部位予測のデータセット.

Organism	pProS-containing proteins	Structured	Disordered	pProS	no annotations
Arabidopsis thaliana	184	491	128	1,664	95,185
Bos taurus	140	656	50	1,427	69,841
Caenorhabditis elegans	23	361	17	136	24,762
Danio rerio	50	30	0	403	28,941
Dictyostelium discoideum	19	215	8	234	15,609
Drosophila melanogaster	54	1,730	235	465	42,325
Gallus gallus	66	311	200	539	41,811
Mus musculus	527	13,788	2,266	5,349	383,328
Oryza sativa subsp.japonica	68	201	15	424	24,800
Rattus norvegicus	250	2,305	510	2,703	162,460
Saccharomyces cerevisiae	66	7,957	1,266	601	34,049
Schizosaccharomyces pombe	12	358	9	78	6,432

このようにして得られたデータセットに対してNeProcおよびANCHOR2を用いて機能部位予測を行なった．表13に結果を示す．ヒトのpProSを用いた機能部位予測結果(表9)と比較して，NeProcとANCHOR2双方が非常に高い予測精度を記録した．この結果はヒトのpProSを用いた予測と同様にpProSの同定方法に起因する可能性が高い．本項ではpProSを抽出する際の天然変性領域予測にDICHOTおよびMobiDB-liteを用いている．そして，双方の予測プログラムが天然変性領域と予測した領域に機能情報がある場合pProSとして抽出した．MobiDB-liteは偽陽性を抑制する傾向を強く示すプログラムである(表5)．そのためMobiDB-liteは天然変性領域である確率が低い領域は予測せず，より確度の高い領域を予測する．つまり本項のpProS抽出過程において予測された天然変性領域は数は少ないが，より確度が高い，天然変性領域と識別しやすい領域である可能性が高い．そのため，NeProcおよびANCHOR2においてもpProSを含む天然変性領域の予測が容易であった可能性が高い．

表 13. モデル生物における予測天然変性領域中の機能部位予測精度.

	NeProc				ANCHOR2			
Organism	MCC	Sensitivity	Precision	F-score	MCC	Sensitivity	Precision	F-score
Arabidopsis thaliana	0.875	0.947	0.961	0.954	0.892	0.986	0.951	0.968
Bos taurus	0.927	0.929	1.000	0.963	0.973	0.979	1.000	0.989
Caenorhabditis elegans	0.840	1.000	0.755	0.861	0.787	1.000	0.710	0.831
Danio rerio	0.899	0.971	1.000	0.985	-0.029	0.990	0.916	0.952
Dictyostelium discoideum	0.928	0.955	0.944	0.950	0.980	0.979	1.000	0.989
Drosophila melanogaster	0.792	0.913	0.726	0.809	0.792	0.988	0.699	0.819
Gallus gallus	0.935	0.953	0.978	0.966	0.939	0.997	0.952	0.974
Mus musculus	0.806	0.928	0.759	0.835	0.864	0.990	0.814	0.893
Oryza sativa subsp.japonica	0.873	0.896	0.992	0.942	0.970	0.975	1.000	0.987
Rattus norvegicus	0.870	0.930	0.906	0.918	0.899	0.990	0.909	0.948
Saccharomyces cerevisiae	0.680	0.981	0.490	0.653	0.742	0.987	0.579	0.730
Schizosaccharomyces pombe	0.632	1.000	0.478	0.647	1.000	1.000	1.000	1.000

モデル生物においても pProS を抽出することが可能であった．しかし，抽出できた pProS の数は少なく，最も多い *Mus musculus* で 5,349 残基であった．そのため，本項での機能部位予測は精度は高い結果となったがサンプル数が少ないため，参考程度の結果であると考えることが無難である．

本章で抽出できた pProS（表 7 および表 12)は学習データとして用いることも可能である．本章では非常に厳しい条件を用いて pProS を同定した．この条件を緩める事で信頼性と引き換えにさらに多くの pProS を同定することが可能である．ただし，pProS は予測された領域であるため，学習データとして用いるには細心の注意を払いその領域が真に機能部位かを精査する必要がある．

３．５　まとめ

既存の機能部予測プログラムの多くが，同一の学習データを用いて作成されている．また，これらの予測プログラムの予測精度は実用精度には達していない．これは既知の機能部位データの数が限られていることが原因の１つとして考えられる．そこで NeProc は機能部位のデータを用いず，天然変性領域と構造領域のデータのみを学習することで機能部位の予測を試みた．その結果，既存の機能部位を学習したプロ

グラムを上回る精度を達成した．特に NeProc は 10〜50 残基の比較的短い機能部位の予測精度が高く短い機能部位予測に有用であることが示された．この結果は，機能部位予測プログラムにおける，機能部位のデータ不足を克服する可能性を示している．

また，UniProt データベースと天然変性領域予測プログラムを用いることで pProS を抽出することが可能であることを示した．この結果は，タンパク質中には未知の機能部位が含まれていることを示唆している．さらに，NeProc は pProS データセットに対する機能部位予測においても既存の予測プログラムを上回る精度を達成した．これは pProS を含む領域の天然変性領域予測が容易な領域であれば NeProc は未知の機能部位予測においても有用なプログラムであることを示唆している．

第 4 章　NeProc によるヒトプロテオームへの機能部位予測

4.1 はじめに

これまでに多くの天然変性領域予測プログラムが作成されており，天然変性領域予測は実用精度を達成している．それらの予測プログラムはヒトプロテオームに約 30%～40%天然変性領域が含まれていることを示唆してきた[22]．また，天然変性領域の割合は細胞内局在ごと異なり核に局在するタンパク質は多くの天然変性領域を保持していることを明らかにしてきた[21, 22]．天然変性領域中には機能部位が存在し，この領域を介した相互作用によって天然変性タンパク質は多くの生物学的プロセスに関わっている．機能部位の割合は細胞内局在において天然変性領域と同様の傾向は見られるのだろうか？α-MoRF-Pred を用いた解析[48]では UniProt データベースの機能情報のうち“regulation”，“cell division”，“cytoskeleton”および“ribosomal proteins”の機能注釈を持つタンパク質には,平均的な真核生物のタンパク質よりも多くのα-ヘリックスを形成する機能部位が含まれていることが報告されている[48]が天然変性領域と機能部位の割合などには触れられていない．安易に想像すると，機能部位は天然変性領域中に存在することから，天然変性領域が長ければ機能部位も多く含んでいると想像できるが，それが真であるかを確認する必要がある．

また，ヒトプロテオーム中の機能部位の総数については 2014 年に Tompa らが ANCHOR[100]およびα-MoRF-Pred [48]を用いてヒトのタンパク質中に 10 万領域を超える機能部位が含まれていると推定した[101]．本研究では天然変性領域中の機能部位予測プログラム，NeProc を開発した．NeProc は IDEAL データベースを用いたテストおよび未知の機能部位である可能性がある pProS データセットにおいてある程度の予測精度を達成した．そこで，NeProc をヒトプロテオームに対して用いることで，NeProc がヒトプロテオームにどの程度の機能部位が存在すると推定するか確認する．また，2014 年以降に開発された ANCHOR2，DISOPRED3 および MoRFchibi-Web においてもヒトプロテオーム中にどの程度機能部位を推定するか確認する．さらにこれらの予測プログラムが予測する機能部位は，天然変性領域のように細胞内局在ごとで異なるかを検証する．本章を通して現在の機能部位予測プログラムが予測する機能部位の傾向および NeProc の機能部位予測における問題点を把握する．

4.2 方法

4.2.1 データセット

本章のデータセットは 3 章で UniProt より獲得した 20,410 個のヒトタンパク質を対象とした．ただし，予測プログラムごとに入力できるタンパク質のアミノ酸配列の長さが異なるため，DISOPRED3 および MoRFchibi-Web では 20,406 個のヒトタンパク質を対象に機能部位を予測した．

4.2.3 細胞内局在

ヒトプロテオームにおいて予測される機能部位が細胞内局在ごとに傾向があるかを解析する．タンパク質の細胞内局在は UniProt の“subcellular location”の情報を用いた．本章でもちいたヒトタンパク質では 3029 種類の膨大な細胞内局在情報が得られた．そこで結果を理解しやすいように本章では“核にのみに局在するタンパク質(N)”，“細胞質にのみに局在するタンパク質(C)”，“細胞膜にのみに局在するタンパク質(M)”，“核と細胞質に局在するタンパク質(CN)”，“核と細胞膜に局在するタンパク質(NM)”，“細胞質と細胞膜に局在するタンパク質(CM)”，“核と細胞質および細胞膜に局在するタンパク質(CMN)”および“その他に局在するタンパク質(others)”の 8 つの局在に分類し解析した．

4.3 結果と考察

4.3.1 ヒトプロテオームに対する機能部位の予測

表 14 にはヒトプロテオームに対する各プログラムが予測する機能部位の統計を示した．また，図 24 には予測される機能部位，天然変性領域性領域および構造領域のヒトプロテオーム中での割合を示した．MoRFchibi-Web は機能部位とその他の領域の予測を行うため天然変性領域と構造領域の割合は示していない．表 14 中の%residues は図 24 の機能部位の割合(青で示された割合)と一致する．NeProc，ANCHOR２および DISOPRED3 の天然変性領域の割合は 30%～35%と大きな差はない．しかし，機能部位に関しては様子が異なる．もっとも割合が高いのは ANCHOR2 であり 18.5%の約 210 万残基を機能部位と予測している．次に NeProc が 17.5%の約 200 万残基と予測している．一方 DISOPRED3 と MoRFchibi-Web は 6.7%の約 76 万残基および

3.3%の約 37 万残基と比較的控え目な予測数であった．NeProc は機能部位を約 11 万領域予測しており Tompa らの推定値[101]とほぼ同程度であった．NeProc と ANCHOR2 は共に 200 万残基程の推定を行っているが，領域数には大きな差がある．NeProc は機能部位を約 11 万領域予測しているのに対し，ANCHOR2 は約 5 万領域を機能部位と予測しており，NeProc の半数程度の予測結果であった．これに関連して，予測機能部位の平均の長さは NeProc が 17.7 残基，ANCHOR2 が 42.2 残基と差があった．この結果は採用している window サイズに起因すると考えられる．NeProc が Smodel において 3 および 5 との短い window サイズを採用しているのに対して，ANCHOR2 は window サイズを 41 と比較的長い．DISOPRED3 は予測する残基数は少ないが領域数は約 10 万領域と NeProc と同程度の領域数を予測している．予測領域の平均長は 8 残基と短い．MoRFchibi-Web は予測する残基数および領域数の双方が少ないが平均の領域長は 12.3 残基と NeProc と DISOPRED3 の中間的な長さであった．ヒトプロテオームでの機能部位予測では NeProc，DISOPRED3 および MoRFchibi-Web は 10〜20 残基程度の領域を予測しているが，予測する数に差がある．NeProc は積極的に機能部位を予測し，DISOPRED3 および MoRFchibi-Web は比較的消極的に予測する．NeProc および ANCHOR2 は予測する数は近い値であるが，予測機能部位の平均の長さは異なる結果であった．

表 14. ヒトプロテオームでの機能部位予測.

	Regions	Residues	%residues	average length
NeProc	112,925	1,994,012	17.5%	17.7
ANCHOR2	49,760	2,101,407	18.5%	42.2
DISOPRED3	95,104	757,865	6.7%	8.0
MoRFchibi-Web	30,332	372,803	3.3%	12.3

%residues はヒトプロテオーム中での機能部位残基の割合を示している

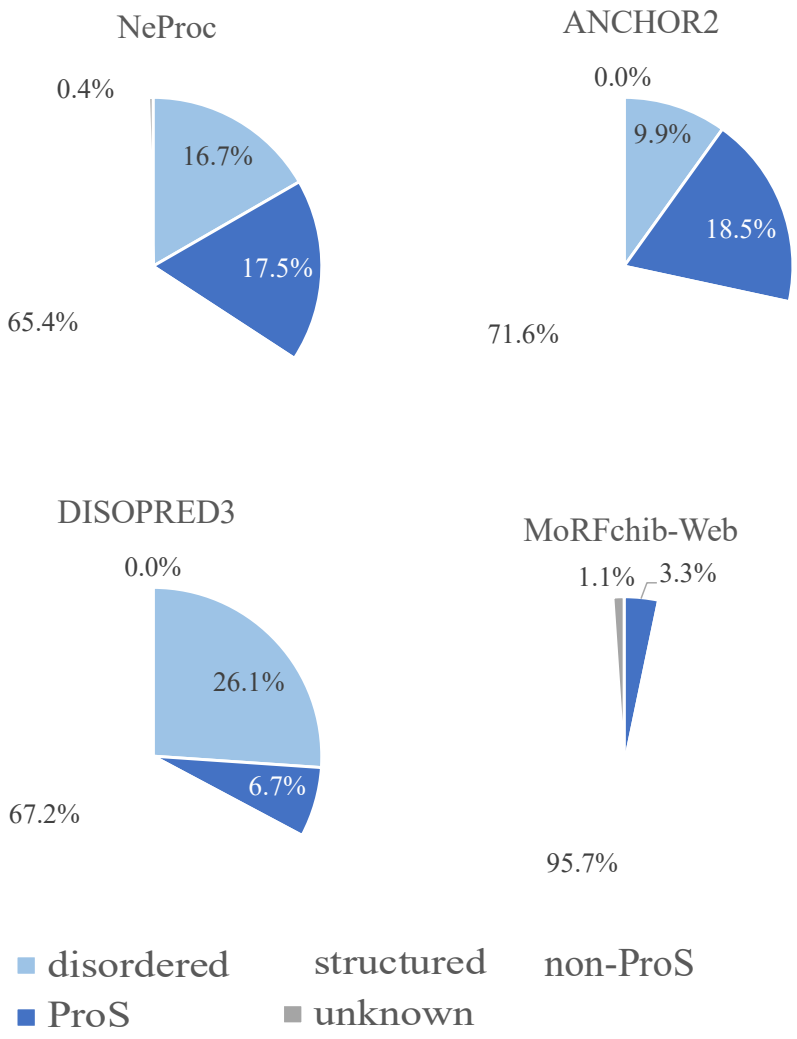


図 24. 各予測プラグラムのヒトプロテオームでの機能部位予測. ProS は天然変性領域中の機能部位を表している. non-ProS は天然変性領域および構造領域が含まれている.

4.3.2 細胞内局在ごとの予測機部位の割合

細胞内局在ごとのタンパク質数を図 25 に示した. タンパク質の数は others を除くと核が最も多く，膜タンパク質，細胞質，核および細胞質の順に続く．NeProc，DISOPRED3 および ANCHOR2 の 3 つの予測プログラムの機能部位および天然変性領域予測の結果を細胞内局在ごとでの統計を図 26 に示した．図 26 における天然変性領域率(水色)は天然変性領域または機能部位と予測された領域の割合を示しており，機能部位率(青)は機能部位と予測された領域の割合を示している．局在ごとの天然変性領域率は全ての予測プログラムにおいて同様の傾向が見られる．核に局在するタンパク質および核と細胞質に局在するタンパク質の天然変変性率が高く，これまでの報告[22, 102]と一致している. 一方, 機能部位率は各プログラムで異なる傾向を示した. NeProc では核での機能部位率が約 20%ともっとも高いが，全ての局在において概ね 11%～20%の機能部位率であった．天然変性領域に対する機能部位の割合(灰色折線グラフ)はもっとも天然変性領域率の高い核タンパク質(N)と次に高い核と細胞質タンパク質(CN)の機能部位の割合が低く, もっとも天然変性領域率が低い膜タンパク質(M)が機能部位の割合が高い. DISOPRED3 でも同様の傾向が見られ, DISOPRED3 の折線グラフの形状は NeProc の折線グラフの形状とほぼ一致する．しかし，ANCHOR2 では異なる傾向が見られ概ね天然変性領域率が高いと機能部位率も高くなる傾向が見られた．これらの結果から，NeProc および DISOPRED3 では天然変性領域率にかかわらず，ある程度一定の量の機能部位を予測している. 一方, ANCHOR2 は天然変性領域率に対して，ある程度一定の割合の機能部位を予測した．現状どちらの傾向が正しいかは判断できない.

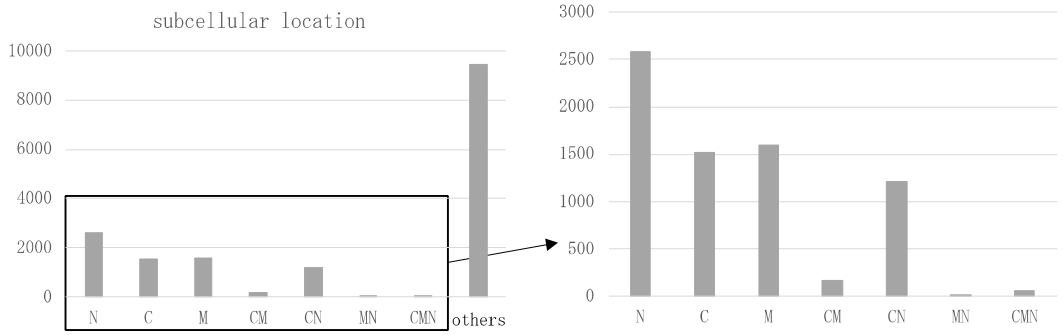


図 25. 細胞内局在ごとのタンパク質数．横軸は細胞内局在を表しており N： 核，C： 細胞質， M ： 膜， CN ： 細胞質と核，NM ： 核および膜，CM ： 細胞質および膜，CMN ： 核，細胞質および膜， others ：その他を表している.

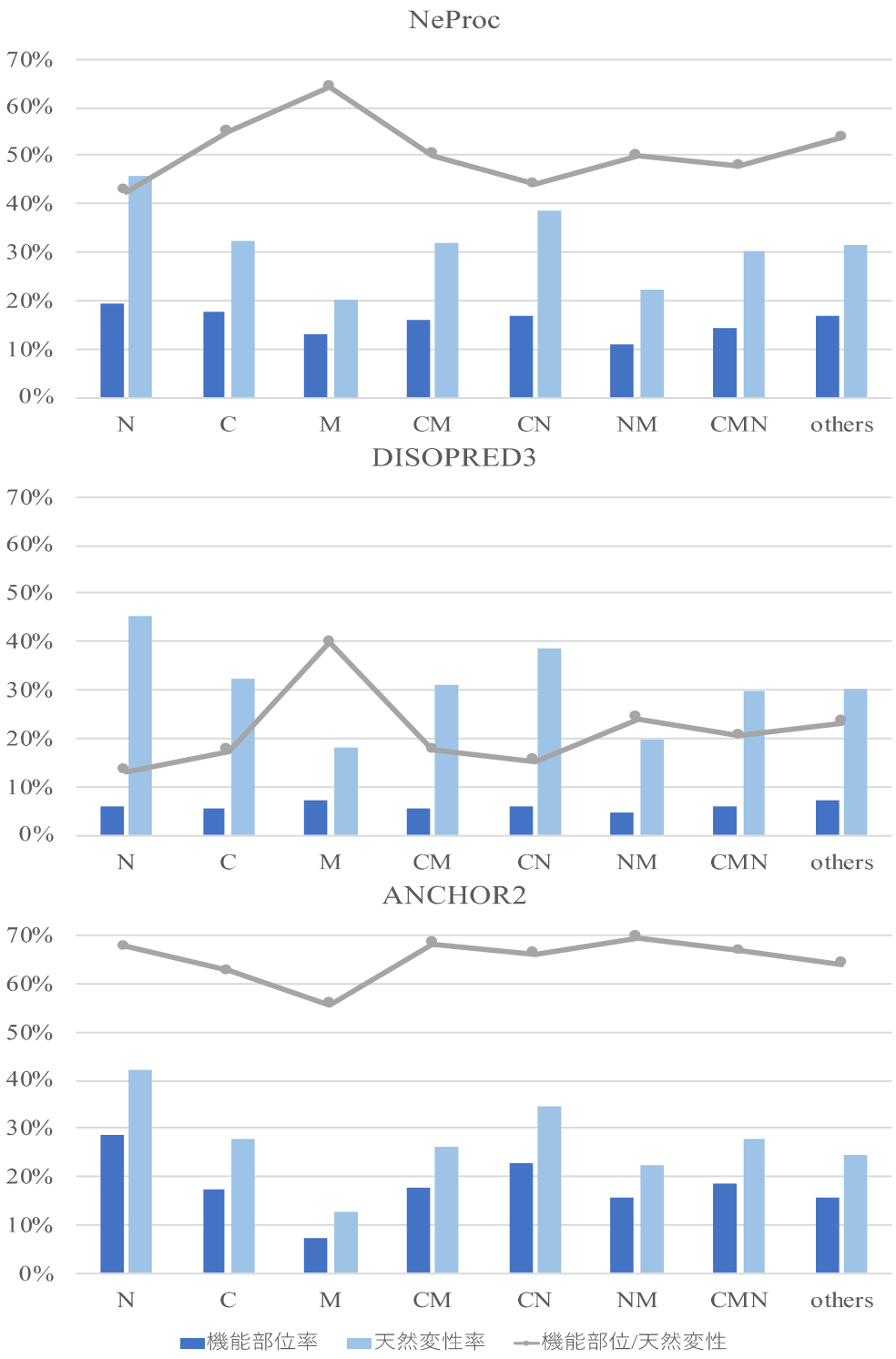
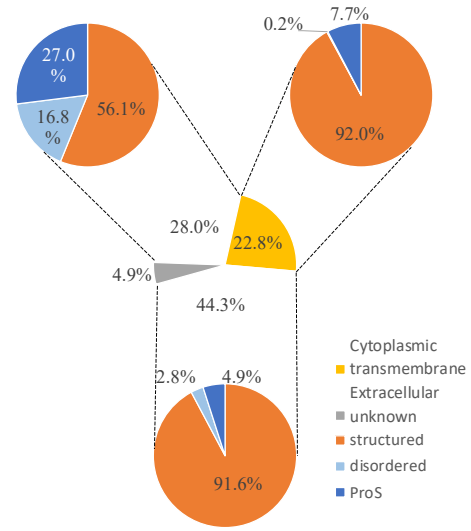


図 26. 細胞内局在ごとの予測される機能部位および天然変性領域の割合. 横軸は細胞内局在を表しており N ： 核, C ： 細胞質, M ： 膜, CN ： 細胞質と核, NM ： 核および膜, CM ： 細胞質および膜, CMN ： 核, 細胞質および膜, others ：その他を表している. 縦軸は割合を示している. 折れ線グラフは天然変性領域に対する機能部位の割合を示している

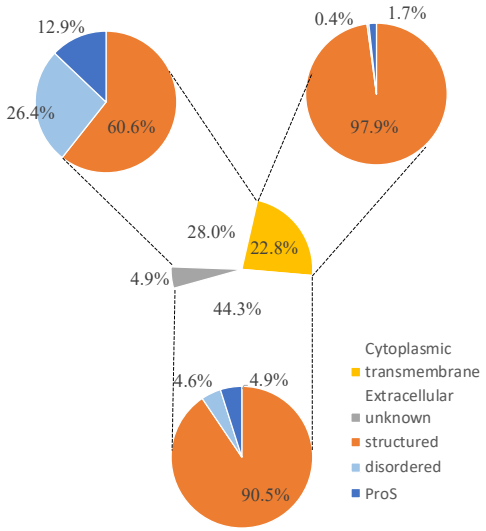
NeProc および DISOPRED3 は膜タンパク質での天然変性領域に対する機能部位の割合が高い．膜貫通型タンパク質は細胞外領域，細胞質側領域および膜貫通領域の 3 領域から成るタンパク質であり，細胞外からの刺激を細胞内へと伝達する働きが知られている．膜貫通領域はαヘリックスやβバレル構造をとり安定していることが知られている．つまり膜貫通領域は天然変性領域中の機能部位である可能性は極めて低い．一方，細胞外領域および細胞質側領域は比較的長い天然変性領域を含むことが知られている[103-108].G タンパク質共役型受容体は 7 回膜貫通するタンパク質であり，細胞質側領域を介してヘテロ三量体 G タンパク質，G タンパク質受容体キナーゼ，および β-アレスチンと相互作用することでシグナル伝達に関与する[109-111]．つまり細胞外領域または細胞質側領域に対して機能部位と予測している場合は，予測が正しいと断定することはできないが機能部位である可能性はある．しかし，膜貫通領域を機能部位と予測した場合は，積極的に予測が誤っていると言える．そこで各予測プログラムが膜タンパク質の 3 領域のうち，どの領域に機能部位を予測する傾向が高いかを解析した．図 27 には NeProc，DISOPRED3，ANCHOR2 および MoRFchibi-Web の膜タンパク質に対しての予測結果を示した．予測される機能部位の割合は予測プログラムごとで異なるが，全ての予測プログラムにおいて細胞質側領域の割合が高かった．この傾向は MoRFpred[49],DISOPRED2[22]および ANCHOR[100]を用いた解析結果[107] と一致する.しかし,予測された領域が真に天然変性領域中の機能部位であると断定することはできない．ただし，予測された領域に関する機能情報やタンパク質ごとの相互作用数などの既知のデータを組み合わせて解析をすることで，3 章において抽出した pProS のような領域を同定できる可能性があり今後の課題である．

膜貫通領域の機能部位予測結果から NeProc の問題点が見えた．DISOPRED3，ANCHOR2 および MoRFchibi-Web は膜貫通領域を機能部位と予測する割合が 1.7%，0%および 0.7%と低く抑えられている．一方，それらの予測プログラムと比較して，NeProc は機能部位を 7.7%予測してしまった．膜貫通領域はαヘリックスなどの安定した構造をとっているため，天然変性領域中の機能部位ではなく，NeProc の予測には問題がある．NeProc は Lmodel を用いて天然変性領域を予測し，Smodel を用いて予測された天然変性領域中に存在する構造領域的傾向を示す領域を機能部位と予測する．膜貫通領域は安定した 2 次構造を形成しており，Smodel はその傾向を正しく捉えていると言え

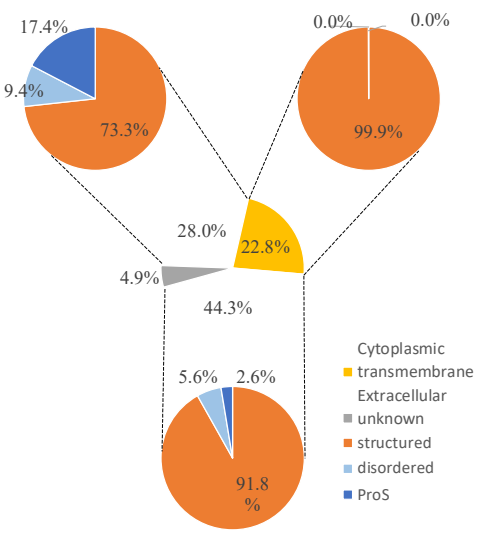
る. つまり Lmodel が誤って膜貫通領域を天然変性領域予測と予測してしまったことが原因である. 図 27 ではそのことが反映されており, 7.7%を機能部位と予測したのに対して, 天然変性領域と予測された領域は 0.2%であった. 機能部位は Lmodel によって予測された天然変性領域に存在する領域である. つまり Lmodel が膜貫通領域の 7.9%(7.7%+0.2%)を天然変性領域と予測したことを示しており, Lmodel によって予測された天然変性領域のほぼ全てを Smodel が機能部位と予測したことを示している. DISOPRED3, ANCHOR2 および MoRFchibi-Web 膜貫通領域を機能部位または天然変性領域と予測しないために, 入力されたタンパク質の膜貫通領域をマスクするなどの対策を施している[6, 10, 22]. NeProc ではこのような処理を行っておらず今後改善すべき点である.



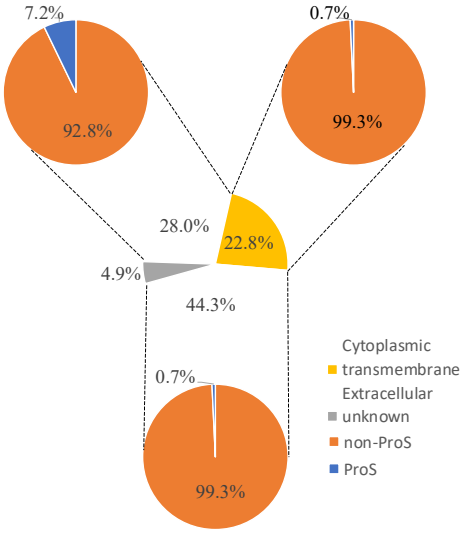
A) NeProcが予測する機能部位の割合.



B) DISOPRED3が予測する機能部位の割合.



C) ANCHOR2が予測する機能部位の割合.



C) MoRFchib-Webが予測する機能部位の割合.

図 27. 膜タンパク質の予測される機能部位の割合. ProS は天然変性領域中の機能部位を表している.

4.4 まとめ

本章ではヒトプロテオームに対して機能部位予測を行なった. その結果, 各機能部位予測プログラムではヒトプロテオームにおいて予測する機能部位の傾向が異なることが示された. NeProc と ANCHOR2 は約 200 万残基の機能部位を予測したが,

予測された機能部位の長さが異なり ANCHOR2 の予測する機能部位は NeProc の予測する機能部位の平均の長さは 2 倍程度であった．DISOPRED3 および MoRFchibi-Web が予測する機能部位の平均の長さは 10 残基程度と NeProc と同様に比較的短いが予測される機能部位の残基数は NeProc と比較して少なかった．細胞内局在ごとの統計では，NeProc および DISOPRED3 では天然変性領域率にかかわらず，ある程度一定の量の機能部位を予測した．一方，ANCHOR2 は天然変性領域率に対して，ある程度一定の割合の機能部位を予測した．しかし，どちらの傾向が正しいかを判断することはできず，今後の実験的に検証された機能部位データの増加が望まれる．

本章でのヒトプロテオームに対する機能部位予測では，NeProc は膜タンパク質の膜貫通領域を機能部位と予測してしまった．膜貫通領域は安定した 2 次構造を形成する領域であるため，NeProc の予測は誤りであると認めざるを得ない．この点は NeProc の課題であり今後改善していく必要がある．また，本章では予測される機能部位の割合を単純に解析したに過ぎず，機能情報やパートナータンパク質との相互作用数などを含めた詳細な解析をすることで NeProc の機能部位予測の傾向や問題点を明らかにできる可能性が考えられる．

第 5 章　結論

5.1 本研究の総括

天然変性タンパク質は天然変性領域中の機能部位を介した相互作用によって，転写調整やシグナル伝達などの生物学的に重要なプロセスに関わっている．機能部位の実験による同定には時間的，金銭的コストが莫大にかかる．そこで，予測によって機能部位を決定しようと様々な機能部位予測プログラムが開発されてきた．しかし，機能部位予測の予測精度は実用精度に達していない．この原因の１つとして機能部位データが少ないことが挙げられる．そこで本研究では機能部位データを学習せずに機能部位を予測するプログラム NeProc を開発した．

本研究では天然変性領域中の機能部位が，アミノ酸組成や保存度において構造領域的性質を示すこと，および機能部位が数残基から数十残基の比較的短い領域であることに着目し，機能部位を長い天然変性領域中の構造領域的性質を示す短い領域と定義した．この機能部位を予測するために NeProc は長い window サイズを用いてタンパク質のアミノ酸配列から長い天然変性領域を予測する．そして短い window サイズを用いて予測された天然変性領域中に存在する構造領域的性質を示す領域を識別し，機能部位として予測する．この予測法により NeProc は機能部位を学習することなく，構造領域と天然変性領域のみを学習することで機能部位を予測できることを示した．NeProc は IDEAL データを用いたテストにおいて，既存の機能部位を学習しているプログラムを上回る予測精度を達成した．このテストでは，NeProc が 10 残基から 50 残基の比較的短い機能部位に対して有効な予測プログラムであることが示唆された．しかし，この結果は天然変性領域中に存在する可能性がある未知の機能部位を排除せずに予測精度を評価したもので，他の予測プログラムに対する NeProc の予測性能はこれに依存している可能性も示された．

そこで UniProt データベースを用いてヒトプロテオームに未知の機能部位が存在しているかを，確度の高い機能情報と実用精度を達成している天然変性領域予測プログラムを用いることで分析した．その結果，未知の機能部位の可能性がある pProS を 1,500 領域以上，抽出することができることを示した．そして pProS データセットを用いて機能部位予測を行なったところ，NeProc は pProS を高い精度で予測することが可能であった．これは pProS を含む領域を天然変性領域と予測することが容易であることに

起因していると考えられるが，NeProc が pProS を高い精度で予測できたことを踏まえると，NeProc は天然変性領域予測が容易な領域中の未知の機能部位に対しても有効な予測プログラムであることを示した．

以上のことを踏まえて本研究の成果をまとめると，NeProc は機能部位を学習せずに機能部位を予測できることを示し，これは天然変性領域中に機能部位予測を困難にしている原因の１つである機能部位データの不足を克服する可能性があることを示している．

5.2 展望

NeProc は機能部位を学習せずに，既存の予測プログラムを上回る精度で機能部位を予測できることを示したが，実用精度は達成していない．また，機能部位予測プログラムでは未知の機能部位の扱い方においてプログラムごとに異なる対応をしている．NeProc や ANCHOR2 は未知の機能部位の存在を排除せずにモデルを構築している．そのため機能部位を多く予測する傾向があり，多くの機能部位を抽出できるが偽陽性も多い．一方，MoRFchibi-Web は未知の機能部位の存在を排除してモデルを構築している．そのため機能部位を少なく予測する傾向があり，偽陰性を抑制できるが未知の機能部位の推定には消極的である．どちらの手法が優れているかを判断することは難しい．またどちらの予測プログラムも実用精度に達してはいないため，より確度の高い予測を得るためには，NeProc の予測のみを用いるのではなく，他の予測プログラムの予測結果やデータベースの情報と組み合わせて用いることが最善であると考えられる．

現状の NeProc には上記した以外にも問題を含んでいる．NeProc は膜貫通領域を機能部位と予測してしまう傾向が他の予測プログラムより高い．この予測は誤りである可能性が高く今後の改善する必要がある．そのために既存の膜貫通領域予測プログラムなどを用いて膜貫通領域をマスクする方法が効果的であると考えられる．また，本研究では機能部位を一括りにして予測を行なったが，パートナータンパク質と相互作用する際の 2 次構造や細胞内局在など機能部位にも多様性がある．これらを一括りに予測を行うことは良いのだろうか？この疑問は実験的に確証が取れたデータが増えないことには解決できない事柄ではあるが，常に留意しておく必要がある．

NeProc は機能部位予測においてある程度の精度を達成した．さらなる予測精度の向上を達成するためには様々な視点から予測を試みる必要がある．上記で述べた

他の予測プログラムと組み合わせることは様々な視点から予測を試みることを意味している．さらに NeProc においても他の視点を含めていく必要がある．現状の NeProc では機能部位データが不足しているために機能部位を学習せずに予測モデルを作成した．しかし，本研究において UniProt より一定量の未知の機能部位を抽出することができた．さらに天然変性領域の決定条件を緩めればさらに多くの機能部位データを抽出することが可能である．ある程度機能部位のデータを確保できたならば，これらの機能部位データを用いて予測プログラムを新たに作成することが可能である．現状の機能部位を学習しない NeProc と新たに作成する機能部位を学習した予測モデルを組み合わせることで予測精度の向上を目指す．新たなモデルを作成するにあたり，現状の NeProc の予測の傾向を知ることは重要である．そこで，ヒトプロテオームでの NeProc の機能部位予測の解析を進めることで NeProc の機能部位予測における傾向および問題を捉え，その結果を踏まえて新たな予測モデル作成を試みる．

天然変性タンパク質が関わる生物学的プロセスは未だ解明されていないことが多く存在する．最近では天然変性タンパク質が液―液相分離(LLPS)に関与していることが多く報告されている．適時，形成したり分解したりすることで生体内での様々な反応を誘導している LLPS において，天然変性タンパク質は反応場を形成する役割が報告されているが例は少ない．LLPS の内部では強く結合してしまうと分解することができずにアミロイド繊維化し凝集してしまう．そのため，LLPS 内部では，結合しては離れ，結合しては離れを流動的に行う必要がある．これは天然変性領域中の機能部位の特徴に合致する．つまり，LLPS では天然変性領域中の機能部位が重要な役割を担っている可能性が考えられる．NeProc を含めた機能部位予測プログラムは実用精度こそ達成していないがある程度の傾向を掴むことは可能である．そこで NeProc や他のプログラムを組み合わせることで LLPS に関わるタンパク質の重要な領域の傾向を捉えることが期待される．

今後は本研究で作成した NeProc を用いた LLPS 関連の解析および予測される機能部位の傾向の解析を行いつつ，本研究で抽出した機能部位を用いた予測モデルの作成を進めていく．その上で，現状の NeProc を用いた解析では予測精度が実用精度に達していないことを踏まえ，他の予測プログラムとの組み合わせや UniProt などの確度の高い情報を組み合わせて解析を行っていく．また，新たな予測モデル作成では NeProc の問題点および機能部位の多様性に留意しつつ開発を行い，予測精度の向上お

よび天然変性領域中の機能部位の知見を深めていく.

参考文献

1. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-31.
2. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
3. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics.* Protein Sci, 2002. **11**(4): p. 739-56.
4. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.
5. Prilusky, J., et al., *FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded.* Bioinformatics, 2005. **21**(16): p. 3435-8.
6. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.
7. Garner, E., et al., *Predicting Binding Regions within Disordered Proteins.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 41-50.
8. Cheng, J., M.J. Sweredoski, and P. Baldi, *Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data.* Data Mining and Knowledge Discovery, 2005.
9. Ishida, T. and K. Kinoshita, *PrDOS: prediction of disordered protein regions from amino acid sequence.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W460-4.
10. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-9.
11. Zhang, T., et al., *SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method.* J Biomol Struct Dyn, 2012. **29**(4): p. 799-813.
12. Mizianty, M.J., Z. Peng, and L. Kurgan, *MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles.* Intrinsically Disord Proteins, 2013. **1**(1): p. e24428.
13. Sormanni, P., et al., *The s2D method: simultaneous sequence-based prediction of the*

statistical populations of ordered and disordered regions in proteins. J Mol Biol, 2015. **427**(4): p. 982-996.

14. Peng, Z., M.J. Mizianty, and L. Kurgan, *Genome-scale prediction of proteins with long intrinsically disordered regions.* Proteins, 2014. **82**(1): p. 145-58.
15. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
16. Iqbal, S. and M.T. Hoque, *DisPredict: A Predictor of Disordered Protein Using Optimized RBF Kernel.* PLoS One, 2015. **10**(10): p. e0141551.
17. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks.* Bioinformatics, 2017. **33**(5): p. 685-692.
18. Kozlowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins.* BMC Bioinformatics, 2012. **13**: p. 111.
19. Necci, M., et al., *MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins.* Bioinformatics, 2017. **33**(9): p. 1402-1404.
20. Fukuchi, S., et al., *Binary classification of protein molecules into intrinsically disordered and ordered segments.* BMC Struct Biol, 2011. **11**: p. 29.
21. Minezaki, Y., et al., *Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation.* J Mol Biol, 2006. **359**(4): p. 1137-49.
22. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
23. Haynes, C., et al., *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.* PLoS Comput Biol, 2006. **2**(8): p. e100.
24. Patil, A. and H. Nakamura, *Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.* FEBS Lett, 2006. **580**(8): p. 2041-5.
25. Romero, P.R., et al., *Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8390-5.
26. Zhu, S., et al., *Hyperphosphorylation of intrinsically disordered tau protein induces an*

amyloidogenic shift in its conformational ensemble. PLoS One, 2015. **10**(3): p. e0120416.

27. Auluck, P.K., G. Caraveo, and S. Lindquist, *α-Synuclein: membrane interactions and toxicity in Parkinson's disease.* Annu Rev Cell Dev Biol, 2010. **26**: p. 211-33.
28. Patel, A., et al., *A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation.* Cell, 2015. **162**(5): p. 1066-77.
29. Elbaum-Garfinkle, S., et al., *The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics.* Proc Natl Acad Sci U S A, 2015. **112**(23): p. 7189-94.
30. Nott, T.J., et al., *Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles.* Mol Cell, 2015. **57**(5): p. 936-947.
31. Kato, M., et al., *Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels.* Cell, 2012. **149**(4): p. 753-67.
32. Kwon, I., et al., *Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains.* Cell, 2013. **155**(5): p. 1049-1060.
33. Dyson, H.J. and P.E. Wright, *Coupling of folding and binding for unstructured proteins.* Curr Opin Struct Biol, 2002. **12**(1): p. 54-60.
34. Sugase, K., H.J. Dyson, and P.E. Wright, *Mechanism of coupled folding and binding of an intrinsically disordered protein.* Nature, 2007. **447**(7147): p. 1021-5.
35. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
36. Tompa, P., *The interplay between structure and function in intrinsically unstructured proteins.* FEBS Lett, 2005. **579**(15): p. 3346-54.
37. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
38. Demarest, S.J., et al., *Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators.* Nature, 2002. **415**(6871): p. 549-53.
39. Borgia, A., et al., *Extreme disorder in an ultrahigh-affinity protein complex.* Nature, 2018. **555**(7694): p. 61-66.

40. Mittag, T., et al., *Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor.* Proc Natl Acad Sci U S A, 2008. **105**(46): p. 17772-7.

41. Polakis, P., *Wnt signaling and cancer.* Genes Dev, 2000. **14**(15): p. 1837-51.

42. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.

43. Ren, S., et al., *Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions.* BMC Genomics, 2008. **9 Suppl 2**(Suppl 2): p. S26.

44. Schad, E., et al., *DIBS: a repository of disordered binding sites mediating interactions with ordered proteins.* Bioinformatics, 2018. **34**(3): p. 535-537.

45. Fukuchi, S., et al., *IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature.* Nucleic Acids Res, 2012. **40**(Database issue): p. D507-11.

46. Fukuchi, S., et al., *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.* Nucleic Acids Res, 2014. **42**(Database issue): p. D320-5.

47. Xue, B., A.K. Dunker, and V.N. Uversky, *Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction.* Int J Mol Sci, 2010. **11**(10): p. 3725-47.

48. Cheng, Y., et al., *Mining alpha-helix-forming molecular recognition features with cross species sequence alignments.* Biochemistry, 2007. **46**(47): p. 13468-77.

49. Disfani, F.M., et al., *MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins.* Bioinformatics, 2012. **28**(12): p. i75-83.

50. Fang, C., et al., *MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation.* BMC Bioinformatics, 2013. **14**: p. 300.

51. Khan, W., et al., *Predicting binding within disordered protein regions to structurally characterised peptide-binding domains.* PLoS One, 2013. **8**(9): p. e72838.

52. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.

53. Malhis, N., et al., *Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule.* PLoS One, 2015. **10**(10): p. e0141603.

54. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.

55. Sharma, R., et al., *Predicting MoRFs in protein sequences using HMM profiles.* BMC Bioinformatics, 2016. **17**(Suppl 19): p. 504.

56. Hanson, J., et al., *Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning.* Bioinformatics, 2020. **36**(4): p. 1107-1113.

57. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences.* Proteomics, 2019. **19**(6): p. e1800058.

58. Barik, A., et al., *DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server.* J Mol Biol, 2020. **432**(11): p. 3379-3387.

59. Peti, W. and R. Page, *Molecular basis of MAP kinase regulation.* Protein Sci, 2013. **22**(12): p. 1698-710.

60. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**: p. 208.

61. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

62. Hemmings, H.C., Jr., et al., *DARPP-32, a dopamine- and adenosine 3':5'-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. II. Purification and characterization of the phosphoprotein from bovine caudate nucleus.* J Neurosci, 1984. **4**(1): p. 99-110.

63. Weinreb, P.H., et al., *NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded.* Biochemistry, 1996. **35**(43): p. 13709-15.

64. Jones, D.T. and J.J. Ward, *Prediction of disordered regions in proteins from position specific score matrices.* Proteins, 2003. **53 Suppl 6**: p. 573-8.

65. Dosztanyi, Z., et al., *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* J Mol Biol, 2005. **347**(4): p. 827-39.

66. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.*

Structure, 2003. **11**(11): p. 1453-9.

67. Li, M., S.B. Cho, and K.H. Ryu, *A novel approach for predicting disordered regions in a protein sequence.* Osong Public Health Res Perspect, 2014. **5**(4): p. 211-8.

68. Su, C.T., C.Y. Chen, and C.M. Hsu, *iPDA: integrated protein disorder analyzer.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W465-72.

69. Shimizu, K., S. Hirose, and T. Noguchi, *POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix.* Bioinformatics, 2007. **23**(17): p. 2337-8.

70. Hirose, S., et al., *POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions.* Bioinformatics, 2007. **23**(16): p. 2046-53.

71. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W164-8.

72. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.* Bioinformatics, 2005. **21**(16): p. 3369-76.

73. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

74. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D301-3.

75. Gunasekaran, K., C.J. Tsai, and R. Nussinov, *Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers.* J Mol Biol, 2004. **341**(5): p. 1327-41.

76. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016. **44**(W1): p. W488-93.

77. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.

78. Davey, N.E., et al., *Attributes of short linear motifs.* Mol Biosyst, 2012. **8**(1): p. 268-81.

79. Fuxreiter, M., P. Tompa, and I. Simon, *Local structural disorder imparts plasticity on linear motifs.* Bioinformatics, 2007. **23**(8): p. 950-6.

80. Meszaros, B., et al., *Molecular principles of the interactions of disordered proteins.* J Mol Biol, 2007. **372**(2): p. 549-61.

81. Trudeau, T., et al., *Structure and intrinsic disorder in protein autoinhibition.* Structure, 2013. **21**(3): p. 332-41.

82. Ota, H. and S. Fukuchi, *Sequence conservation of protein binding segments in intrinsically disordered regions.* Biochem Biophys Res Commun, 2017. **494**(3-4): p. 602-607.

83.

84. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10.* Proteins, 2014. **82 Suppl 2**: p. 127-37.

85. Nielsen, J.T. and F.A.A. Mulder, *Quality and bias of protein disorder predictors.* Sci Rep, 2019. **9**(1): p. 5137.

86. The UniProt, C., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.

87. Fukuchi, S., et al., *Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains: its application to human transcription factors.* BMC Struct Biol, 2009. **9**: p. 26.

88. He, K., et al., *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.* Proceedings of the IEEE International Conference on Computer Vision, 2015: p. 1026-1034.

89. Kingma., D.P. and J.L. Ba., *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.* International Conference for Learning Representations(ICLR), 2015.

90. Shapiro, S.S. and M.B. Wilk, *An analysis of variance test for normality (complete samples).* Biometrika, 1965. **52**: p. 591-611.

91. Wilcoxon, F., *Individual comparisons of grouped data by ranking methods.* J Econ Entomol, 1946. **39**: p. 269.

92. Necci, M., D. Piovesan, and S.C.E. Tosatto, *Where differences resemble: sequence-feature analysis in curated databases of intrinsically disordered proteins.* Database

(Oxford), 2018. **2018**.

93. Rubin, S.M., et al., *Structure of the Rb C-terminal domain bound to E2F1-DP1: a mechanism for phosphorylation-induced E2F release.* Cell, 2005. **123**(6): p. 1093-106.

94. Fontes, M.R., et al., *Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha.* J Biol Chem, 2003. **278**(30): p. 27981-7.

95. Pawson, T. and P. Nash, *Assembly of cell regulatory systems through protein interaction domains.* Science, 2003. **300**(5618): p. 445-52.

96. Lee, H.J. and J.J. Zheng, *PDZ domains and their binding partners: structure, specificity, and modification.* Cell Commun Signal, 2010. **8**: p. 8.

97. Nomine, Y., et al., *Structural and functional analysis of E6 oncoprotein: insights in the molecular pathways of human papillomavirus-mediated pathogenesis.* Mol Cell, 2006. **21**(5): p. 665-78.

98. Greschik, H., et al., *Communication between the ERRalpha homodimer interface and the PGC-1alpha binding surface via the helix 8-9 loop.* J Biol Chem, 2008. **283**(29): p. 20220-30.

99. Kallen, J., et al., *Evidence for ligand-independent transcriptional activation of the human estrogen-related receptor alpha (ERRalpha): crystal structure of ERRalpha ligand binding domain in complex with peroxisome proliferator-activated receptor coactivator-1alpha.* J Biol Chem, 2004. **279**(47): p. 49330-7.

100. Meszaros, B., I. Simon, and Z. Dosztanyi, *Prediction of protein binding regions in disordered proteins.* PLoS Comput Biol, 2009. **5**(5): p. e1000376.

101. Tompa, P., et al., *A million peptide motifs for the molecular biologist.* Mol Cell, 2014. **55**(2): p. 161-9.

102. Ota, M., et al., *Multiple-Localization and Hub Proteins.* PLoS One, 2016. **11**(6): p. e0156455.

103. Minezaki, Y., K. Homma, and K. Nishikawa, *Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment.* J Mol Biol, 2007. **368**(3): p. 902-13.

104. De Biasio, A., et al., *Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand Delta-4.* J Proteome Res, 2008.

7(6): p. 2496-506.

105. Xue, B., et al., *Analysis of structured and intrinsically disordered regions of transmembrane proteins.* Mol Biosyst, 2009. **5**(12): p. 1688-1702.

106. Tusnády, G.E., L. Dobson, and P. Tompa, *Disordered regions in transmembrane proteins.* Biochim Biophys Acta, 2015. **1848**(11 Pt A): p. 2839-48.

107. Bürgi, J., et al., *Intrinsic Disorder in Transmembrane Proteins: Roles in Signaling and Topology Prediction.* PLoS One, 2016. **11**(7): p. e0158594.

108. Xue, B. and V.N. Uversky, *Structural characterizations of phosphorylatable residues in transmembrane proteins from Arabidopsis thaliana.* Intrinsically Disord Proteins, 2013. **1**(1): p. e25713.

109. Bellot, G., et al., *Structure of the third intracellular loop of the vasopressin V2 receptor and conformational changes upon binding to gC1qR.* J Mol Biol, 2009. **388**(3): p. 491-507.

110. Boguth, C.A., et al., *Molecular basis for activation of G protein-coupled receptor kinases.* Embo j, 2010. **29**(19): p. 3249-59.

111. Ostermaier, M.K., et al., *Functional map of arrestin-1 at single amino acid resolution.* Proc Natl Acad Sci U S A, 2014. **111**(5): p. 1825-30.

謝辞

本論文を審査して頂いた学外審査委員の太田元規教授，学内審査委員の，本間桂一教授，福地佐斗志教授，中村建介教授，佐川考広准教授より，貴重なご指導とご助言を賜りました．感謝申し上げます．

主指導教員である福地佐斗志教授には，大学学部学生時代から現在に至るまで私に，研究の楽しさや難しさ，研究の進め方から論文執筆まで多くのご指導をいただきました．心から感謝申し上げます．

また，福地研究室所属の細田和男博士ならびに元福地研究室所属の天貝宏樹さんにはNeProcの開発環境の構築からアルゴリズムのご指導まで多大なご協力を頂きました．感謝申し上げます．

最後に，所属する福地研究室のみなさまには作成したNeProcの動作確認を協力していただき感謝しております．また研究における議論から日常会話まで大変多くの刺激を得ることができました．お礼申し上げます．

補足資料

補足説明１　天然変性領域中の機能部位と構造領域および天然変性領域のアミノ酸組成の比較方法

3.3.4 において正しく予測できた機能部位および予測できなかった機能部位のアミノ酸組成を，学習データに含まれる構造領域および天然変性領域のアミノ酸組成と比較した．比較方法を以下に示す．

Step1. 正しく予測できた機能部位のアミノ酸組成を window をスライドさせて求める．

Step2. Step1 で求めた組成を学習データに含まれる構造領域の組成と
以下の式を用いてアミノ酸ごとの類似度を計算する．

$$S_i^O = \sqrt{(C_i^P - C_i^O)^2},$$

C_i^Pは機能部位のアミノ酸 i の頻度を表している．C_i^Oは学習データの構造領域のアミノ酸 i の頻度を表している．

Step3. Step2 と同様に学習データに含まれる天然変性領域の組成と
の類似度を計算する．

Step4. Step2 および Step3 で求めた各アミノ酸の類似度を用いて円の
中心を以下の式で求める．

$$O = (O^O, O^D) = \left(\frac{\sum^{all\ amino\ acid} S_i^O}{20}, \frac{\sum^{all\ amino\ acid} S_i^D}{20}\right),$$

また，円のサイズは各アミノ酸の類似度の分散を表している．

Step5. Step1 から Step4 までの手順で構造領域と予測してしまった機
能部位，天然変性領域と予測してしまった機能部位および機能部位
全体について計算する．