解説・総説

天然変性タンパク質データベース IDEAL†

福地佐斗志，安保勲人

IDEAL, a database for intrinsically disordered proteins†

Satoshi Fukuchi* and Hiroto Anbo**

Intrinsically disordered proteins are the ones that have intrinsically disordered regions adopting no tertiary structures under the physiological condition. We have been developing a database of intrinsically disordered proteins, IDEAL (https://www.ideal-db.org), for the last decade. In this article, we will look back our activity and introduce IDEAL as the 10th anniversary of IDEAL.

Key words : ***Intrinsically disordered proteins, Biological database, Bioinformatics***
（***Key words 天然変性タンパク質，生物学データベース，バイオインフォーマティックス***）

1　はじめに

IDEAL(Intrinsically Disordered proteins with Extensive Annotations and Literature)[1]は，天然変性タンパク質のデータベースであり，前橋工科大学・福地研究室は名古屋大学・太田研究室と共同で 2011 年からこのデータベースを開発・維持してきた．本年，開発から 10 年目の節目を迎えたのを機に，IDEAL のこれまでを振り返るとともに，提供する情報や機能を紹介したい．

2　天然変性タンパク質とは

タンパク質はアミノ酸が多く結合した数珠（じゅず）のような分子である．この数珠が立体的に折り畳まれ、独特の構造を形成し機能を発揮する．タンパク質に関するこのような描像は普遍的だと思われていたが、2000 年頃からこの描像から逸脱したタンパク質，天然変性タンパク質（英語では intrinsically disordered protein, natively unfolded protein 等と表記される）が注目されだした．天然変性タンパク質は，立体的に折りたたまれない領域，天然変性領域を持っており，さらにこの構造を取らない領域に機能に重要な断片を持っている[2)3)]．天然変性タンパク質にはアミノ酸配列の全域にわたり立体構造を形成しないものもあるが、一般的には数十から数百残基にも及ぶ天然変性領域と立体構造を持つ構造ドメインの組み合わせからなっている．

天然変性領域中の機能部位は多くの場合，長大な天然変性領域中の十～数十残基くらいの領域である．この部位は相互作用相手のタンパク質と出会うと局所的に立体構造を形成して結合する場合がある．この現象は結合と共役した構造形成（coupled folding and binding）として知られている[4]．天然変性領域はアミノ酸組成に特徴があり、計算機プログラムを使って予測できる．このような予測を用いた解析から、天然変性領域は真核生物のタンパク質の 30%程度を占めること，核タンパク質に特に多く、それらは転写調節・シグナル伝達といった現象に関与すること，リン酸化・アセチル化といった翻訳後修飾を受ける残基も，天然変性領域に多く分布すること等が示唆されている[5]．

3 天然変性タンパク質データベース IDEAL

3・1　生命情報学におけるデータベース

生命情報学は生物学から得られる情報を計算機を用い解析する分野であり，情報が存在しなければ成立しない．この情報を提供しているのが公共データベースである．タンパク質の世界で最もよく使われているデータベース UniProt は，1980 年中頃に始まり，現在では人の手で情報が付加されたタンパク質 56 万件，それ以外のタンパク質は 2 億件を超えている．また，タンパク質の立体構造を収録している Protein Data Bank (PDB)は 1970 年代 からデータの収集・配布を開始し，現在では 17 万件以上の構造が収録されている．これらの大量のデータがデータベースに集約され誰でも利用できることが情報インフラとなり，生命情報学が成立していると言って良いだろう．一方，天然変性タンパク質の研究は，今世紀に入り研究が開始された．天然変性領域の存在は以前から認識されていたが，重要な機能を持つことが発見されたのは 20 世紀末だったからだ．このため，立体構造をとる領域が PDB に集約されてきた一方，構造を取らない領域は存在が確認されていたとしても論文の中に埋もれたままであった．現状，PDB に登録時に機械的に取得できる天然変性領域も存在するが，生物学的に意味

† 原稿受理 令和３年２月２６日 Received February 26, 2021
* 生命情報学科 (Department of Life Science and Informatics)
** 生命情報学専攻(Graduate school of Engineering, Division of Life Science and Informatics)

のある天然変性領域を知るためには，論文を読む必要があるのが現状である．このような状況を受け，IDEAL は実験的に確認された天然変性領域を論文を精読して見つけ出しデータベース化してきた．同様なデータベースとして DisProt[6]があるが，DisProt は天変変性領域こそ収集しているものの，その中の機能領域の注釈付は行っていない．我々が IDEAL の開発を始めた大きな理由はこの，天然変性領域中の機能部位データの不足であった．

3・2　天然変性タンパク質情報の収集

IDEAL では天然変性領域を，機械的に取得する方法と論文を精読する方法の二つで取得している．X 線結晶構造解析において，座標が特定できない領域が存在する．結晶中には大量のタンパク質分子が規則正しく並んでいるが，このように座標が特定できない部分は各分子の中で位置がマチマチであり，構造中でフラフラとしている領域と考えられる．このような領域は伝統的に missing residue と呼ばれ，PDB の情報から機械的に取得できる．一般に missing residue も天然変性領域と解釈されている．また，核磁気共鳴法(NMR)で決定された構造では一つの PDB エントリの中に 10～20 個程度の構造モデルを収録するのが常である．これらの構造モデルを重ね合わせたときに，各モデル間で座標の分散の大きい領域も機械的に判断可能 [7]で，天然変性領域として登録している．

一方，同じ NMR で決定された構造中で，NMR のデータに残基の帰属ができない領域も存在する．このような領域は PDB のエントリには記述されておらず，論文を精読することでしか収集できない．また，タンパク質の特定の領域が天然変性領域か否かを判定するために，NMR を用いて構造解析が行われることもある．このような情報も PDB からは知ることができない．特定の領域が天然変性か否かを判定する実験手法でこの他によく用いられるのは，円偏光二色性(circular dichroism，以下 CD)スペクトル測定が挙げられる．αヘリックス，βシート，不規則構造では紫外部領域での CD スペクトルパターンと強度が異なり，このことを利用し二次構造が推定できる．すなわち，不規則構造様のスペクトルを示せば天然変性領域と判断するわけである．また，X 線結晶解析の論文でも，結晶を作る際に取り除いた領域の情報も天然変性領域の指標となる．この様な領域はふらふらとしており，結晶の生成を難しくしていると考えられる．一般に，PDB から得られる missing residue は比較的短い領域（数残基から 10 残基程度）が多く，典型的な天然変性領域と見なされる領域は，NMR や CD といった手法により確認されていることが多い．　これらの情報は PDB や UniProt には記述されていない情報である．

天然変性領域中には他のタンパク質や DNA といった分子と相互作用する領域が存在する．IDEAL ではこの様な領域を Protean Segment (ProS)と呼び収集している．Protean とはギリシャ神話で変幻自在に形を変えることが出来る神・プロメテウスに由来する言葉で，文字通り変幻自在といった意味がある．ProS は単独では紐状だがパートナー分子と結合する際，二次構造を形成するものがあったり，また，結合するパートナーにより二次構造が異なることもある．この様に ProS のパートナーや構造に対する柔軟さから protean という語を用いた．IDEAL の ProS の判定基準は，相互作用相手が居ない時に天然変性である実験的証拠があり，かつ，PDB にパートナーとの結合した構造が登録されていることである．ProS はほとんどの場合，球状構造領域と結合するため，結合した状態ではほとんどの残基は位置に揺らぎがなく座標を決定できる．このため，PDB には結合状態での構造が登録されている例が多い．また，他の生物種の同じタンパク質で ProS と認定されているなど，状況からほぼ ProS と認定して良い場合もある．前者の PDB に結合状態のデータが存在する場合を verified ProS，後者の状況証拠から判断した場合を possible ProS と区別して登録している．

天然変性タンパク質は特に真核生物で多く，転写調節などに深く関わるため，IDEAL ではヒトの核タンパク質を出発点として収集を始めた．ただし，相互作用相手のタンパク質も IDEAL に登録するルールのため，必ずしも全てが核タンパク質というわけではない．核と細胞質を行き来するタンパク質は天然変性タンパク質が多く[8]このようなタンパク質は細胞質に局在するタンパク質とも相互作用を行う．また実験系の構築上，相互作用相手のタンパク質もヒト由来とは限らず，マウス，ラットといった他のモデル生物やウイルス由来のタンパク質も収録されている．

3・2　IDEAL のインターフェース

IDEAL のトップページを図 1 に示す．[1]に現在公開中のバージョンが示されている．IDEAL では，新たな天然変性領域の実験的証拠を収集すると共に，現在収録されているタンパク質に新たに加わった情報もアップデートし追加している．この作業により，可能な限り各タンパク質に最新の情報が付加できるよう配慮している．[2]の Browse ページには収録されたタンパク質の一覧，[3]の Search ページには検索機能，[4]の Download ページにはダウンロード可能な情報がまとめられている．検索

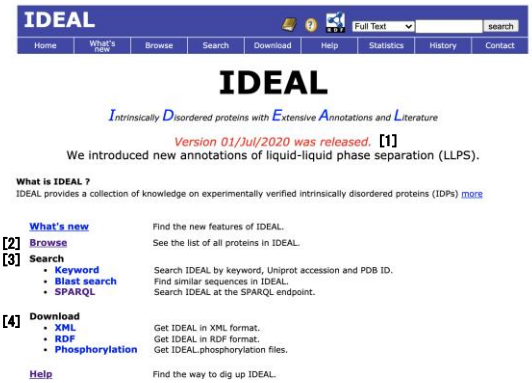

図 1. IDEAL トップページ

機能では，通常のキーワード検索や blast を用いた配列検索に加え，SPARQL エンドポイントも提供している．[4]の取得可能な情報にも記されているように，IDEAL では RDF フォーマットの情報を用意している．RDF とはデータを記述する様式の一つであり，RDF 化されたデータを用いれば，インターネット上で公開されているバラバラの情報を繋ぎ合わせ，統合的に検索などを行うことが出来る．SPARQL 検索はこのような検索を可能とする仕組みである．IDEAL の RDF 化は長年課題であったが，2019 年にライフサイエンス統合データベースセンターの協力で実現した．

IDEAL では収集した天然変性領域および ProS を直感的にわかりやすい図にして提供している．図 2 に例を示す．この例はヒトのカテニンβ1 の例である．図中[1]

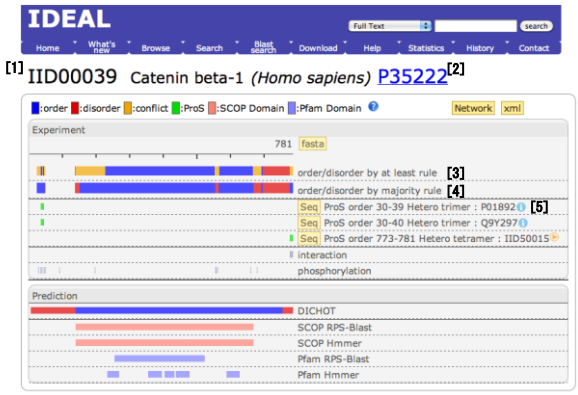

図 2. IDEAL エントリーの例.

はこのエントリの識別子で IDEAL では IID で始まる番号を用いている．[2]は uniprot へのリンク，[3][4]は天然変性領域，構造領域の表示である．本図は白黒だが，実際のページではカラーで表示され，天然変性領域は赤，構造領域は青，コンフリクト領域は黄色で示される．ここでコンフリクト領域とは，実験条件により天然変性領域となったり構造領域となったりする場合をこの様に定義している．PDB には同じタンパク質の実験条件の異なる構造が複数登録されていることが多く，この様なコンフリクト領域が出てくるのは珍しいことではない．また，多くの PDB 構造に関して上記 3 つの領域の定義を決める際，多数決を取る方法(majority rule)と一例でも天然変性の報告がある場合は天然変性と表示する方法(at least rule)の 2 通りで表示している．[5]の部分には ProS の領域が示されている．[3][4][5]の領域のバーはクリックすることが可能で，さらに詳細な情報を見ることが出来る．ProS の定義として PDB に結合状態の構造が存在することが挙げられる．この構造も表示することが出来る(図 3)．また，IDEAL に収録されたタンパク質間の相互作用で形成されるタンパク質間相互作用ネットワークの図も用意されている．天然変性タンパク質は，タンパク質間相互作用ネットワークでハブとして機能することが示唆されており，マップを見ることでハブ性なども確認できる．IDEAL にはこの他に様々な機能がある．詳しい操作法については，トップページからヘルプページへ行けば，詳しい記述が得られる．

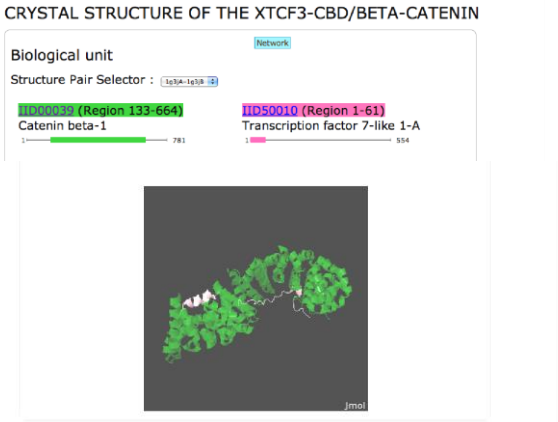

図 3. ProS 構造表示の例. 図中，明るい領域が ProS であり. 構造ドメインに結合している様子が窺える.

4 IDEAL のこれまでと今後

以上のように，IDEAL は紆余曲折しながらも 10 年に渡りデータベースを維持してきた．当初，収録したタンパク質は 153 件であったが現在は 995 件となり，ProS を有するタンパク質数も 72 件から 317 件へと情報を蓄積してきた．天然変性タンパク質の情報は論文を査読して収集する必要があり，欧州で運営され国際連携で情報収集しているデータベース DisProt の登録件数が 1700 であることを考えると，IDEAL の収録数もそれなりの成果と言って良いのではないかと思う．また，DisProt の招きで欧州の情報共有プラットフォームである全欧バイオ情報基盤整備計画(ELIXIR)にも参画し，DisProt との情報共有化も検討中である．

一方で，天然変性タンパク質を取り巻く環境も変化してきている．最近になり液液相分離という現象が注目を集めている．細胞中で多数の分子が液滴と呼ばれる油を水に落としたような集合体を形成し，この膜のないオルガネラは化学反応の反応場を提供するとされる [9]．また，この液滴は生成・解離が制御されており，この制御が破綻するとアミロイド形成により神経疾患を引き起こすことも知られている．天然変性領域はこの液滴形成の駆動力となっている事例が数多く報告されている [10]．天然変性領域の中でも特に，低複雑性領域(Low Complexity Domain; LC ドメイン)と呼ばれる少数の種類のアミノ酸で構成される領域の相互作用が，液滴形成に大きな役割を果たしているとされる．LC ドメインの多くは繰り返し配列を含んでいるが，同じ天然変性領域中の相互作用領域である ProS では，このような繰り返し配列はあまり見られない．このことから，LC ドメインの相互作用とこれまで収集してきた ProS では相互作用の様式が異なっているもかもしれない．このように新しい発見・研究の進展に伴い，IDEAL の情報収集や注釈

付の定義など，見直しを行う必要が出てきたのかもしれない.

1) Fukuchi S, et al., Nucleic Acids Res. 42, D320-D325, 2014.
2) Dunker AK, et al., Biochemistry, 41, 6573-6582, 2002.
3) Wright PE, et al., J. Mol. Biol., 293, 321-331, 1999.
4) Dyson K, et al., Nat. Rev. Mol. Cell Biol., 6, 197-208, 2005.
5) Ward JJ, et al., J. Mol. Biol., 337, 635-645, 2004.
6) Hatos A, et al., Nucleic Acids Res. 48, D269-D276, 2020.
7) Ota M, et al., J. Str. Biol. 181, 29-36, 2013.
8) Ota M, et al., PLOS One, 11, e0156455, 2016.
9) Banani SF, et al., Nat. Rev. Mol. Cell Biol., 18, 185, 2017.
10) Uversky VN, Curr. Opin. Str. Biol., 44, 18-30, 2017.