PDB-descriptome: 構造生物学論文における構造言及の網羅的採集と分子立体構造データへの紐づけ

佐久間航也[1] 丹羽智美[2]
[1]名古屋大学 細胞生理学研究センター [2]大阪大学 蛋白質研究所
ksakuma@cespi.nagoya-u.ac.jp, echo s.niwa@protein.osaka-u.ac.jp | sed s/a@p/@p/g

概要

タンパク質を含む生体分子の立体構造データは、それを構成する原子の位置座標の集合であり、原子タイプなどのラベルが付随した点群とみなせる。立体構造決定を伴う構造生物学論文は、著者が機能的に重要とみなした部分構造について生物学的な解釈を与えることが多い。すなわち、構造生物学は、構造を表現した点群の部分集合を選択する「マスク」と、その部分に対する「キャプション」を与える学問分野であるとみなせる。しかし、これまでテキスト記述と立体構造データを紐づける仕組みは提案されていない。そこで、我々は PDB-descriptome プロジェクトと称し、分子立体構造データに対してテキスト記述を体系的に紐づける仕組み、および、それに基づいた構造–記述ペアデータの開発に取り組んでいる。今回、典型例として選んだ論文と立体構造ペアを用い、構造生物学者の視点で可能な限り詳細なアノテーションデータを作成した結果を報告する。

1 本研究の背景

構造生物学は、生体分子の三次元構造を明らかにし、生命現象に対して分子レベルの説明を与えることを目指す分野である。このため、この分野の論文は実験的に決定された立体構造データについての記述を必然的に多く含み、言及されている立体構造を同時に眺めながらでないと論文内容そのものの読解が難しい。論文とペアになる立体構造データは Protein Data Bank（PDB）で公開されている [1]。これは単に分子を構成する原子座標の集合であり、原子タイプなどのラベルが付随した点群データにすぎない。一方、構造生物学論文は、当該分子の物理化学的性質だけでなく分子機能や生物学的意義なども記述することで、この点群データに「意味」を与えるという重要な役割がある。

構造生物学テキストの特徴として、記述が機能的に重要な部位に限局された記述に富むことが挙げられる。この意味で、構造生物学は、立体構造のある部位に対する「キャプション」と、その部位を指定する「マスク」を与える営みと捉えることができる。

しかしながら、論文テキストの記述が、どの構造領域に関する説明を与えているかという対応関係はこれまで明示的にはデータ化されてこなかった。これは、個々の読者が記述と記述対象の対応を個々に見出す手間が生じるというのみならず、構造生物学的解釈と立体構造とペアデータとして取り扱うことを困難なものにしている。

本プロジェクトではこの課題を解決することを目指し、言語による記述データと立体構造データと明示的に紐づける新規枠組みの開発に取り組んでいる [2,3]。重要なこととして、本プロジェクトにおいては、対象の構造生物学論文からすべての構造への言及を拾い上げて立体構造データにマッピングし、何が重要視されるかについて可能な限り二次的なバイアスを持ち込まないことを理想形としている

本論文では、典型例として選んだ論文について、筆者（佐久間）がアノテーションを行った結果に基づき、記述を構造とあらわに紐づけることで、どのような知見が得られるか報告する。また、今後アノテーションを大規模化していくために重視すべき点を考察する。

2 本研究の方向性

構造生物学に縁遠い読者向けの例え話になるが、立体構造データを見ながら構造生物学論文を読解するという行為は、全身骨格模型を見ながら医学書を読むのと似ていると考えられる。この骨格模型には個々の骨や関節、もっと微妙で曖昧な領域などに名称や番号のラベルが貼られており、これをヒントにすることで「モノ」と記述の間で対応を取りながら医学的な文書を読むことができる。

しかし、構造生物学分野では骨格模型（＝立体構造モデル）の部分部分に貼り付けられているべきラ

ベルが剥がれ、一報の論文という箱に詰められた状態にある。テキストの記述と「かたち」の対応を理解するには、剥がれたラベルを適切な位置に貼り直す必要がある。

また、仮にヒト骨格だけを扱うのであれば、骨や領域の名称を網羅した辞書を用意したうえで、オントロジー的に知識を蓄積できるだろう。しかし、構造生物学分野では領域名称も領域定義も揺らぎがちであるうえ、対象となる生物種も可変である。ヒトと鳥類と魚類の「骨」に関する統一的辞書をつくるのは、あまり有効な方向性ではないように思われる。

こうした状況から、本プロジェクトではあくまで個々の具体的な立体構造と論文のペアリングを重視し、横断的な整理は任意の後処理に任せる。特に注力するのは、テキストで記述される領域名と、その実体である立体構造データ上の領域という異なるモダリティ間の紐づけをおこなう部分である[i]。

3 前提と目的

以下では本プロジェクトでのアノテーションにおける前提を確認し、そこで頻出する概念に呼称を与える。

3.1 前提：論文・構造ペア

出版された構造生物学論文は少なくとも一つの立体構造データと紐づいているとする。つまり論文の ID（たとえば DOI）と PDB 内のエントリ ID（PDB ID）のペアがすでに与えられているとする。

3.2 Entity と Referring Expression

構造生物学論文のテキストでは、すでに説明した通り、特定の構造領域を名指しで説明することが多い。この名指しされる部分構造のことを本研究では、Structural Biological Entity（SBE)と呼ぶことにする。SBE は例えば構造全体、ある鎖全体、任意の部分領域、アミノ酸残基の集合、原子などの階層がありえる。実際上は、当該構造を構成する原子の通し番号（原子 ID)の集合として SBE の定義を表現できる。

一方、これをテキスト側で言及するための表現を Structural Biological Referring Expression（SBRE）と呼ぶことにする。SBRE は分野的に定着した固有名詞的な名称でも、特定の論文著者が論文中で勝手につけたニックネームでも良く、記述としては残基名と残基番号のペア、また文脈を受けた代名詞や一般的な名詞・句などがありえる。

3.3 目的

本研究の目的は、構造生物学論文から SBRE 候補となる表現を網羅的に抽出し、それらに対応する SBE の定義を紐づけることである。

4 手法

4.1 対象とした論文と立体構造

本報告では、Toda らによるダイニン・微小管関連タンパク質複合体の構造決定論文[4]と、そこで報告されている立体構造（PDBID: 6L4P [5]）をアノテーションの対象とした。選定の理由は、論文がオープンアクセス（Creative Commons CC-BY）であること、および立体構造データが中サイズ程度のタンパク質 2 鎖からなるヘテロダイマー構造であり、試験的なアノテーションの対象として複雑さが適切であったことである。

4.2 テキスト側アノテーション

論文誌ウェブサイト上の HTML 版論文からアブストラクト、本文、図のキャプション、表内容をコピーしてプレーンテキスト化した。これを TextAE [6]で開き、目視で SBRE らしき表現を選択・保存した。冒頭から読み進めていき、その時点では実際に SBE と紐づくかどうか深く問わず表現をピックアップし、SBRE 候補とそのスパンのリストを得た。この時点で同一の SBE を指していると判断できた表現は同一のラベルを与え、一括して取り扱った。構造生物学者の視点から有効な SBRE 候補が取り尽くされたと判断できた時点でテキスト側のアノテーションを終了した。

4.3 立体構造側アノテーション

上で作成したリスト中の SBRE 候補で言及されている SBE の定義を立体構造専用の描画ソフトウェアである PyMOL [7]の Selection Syntax を用いて表現

[i] 将来的に、これらの Referring Expression Comprehension や Referring Expression Segmentation に相当するタスクは自動化されることが望ましいものの、現時点ではその基礎となるデータ自体が存在しないため、まずこれを人手で作成することを目指している。

した。具体的には、テキストのみから定義が推測できるもの（たとえば Arg-79 のような「残基名＋残基番号」の表現）は、どの鎖に属する残基か推定し、残基番号を用いて Selection Syntax を構成した。図表の読み取りに基づいて SBE 定義を推定する必要がある場合は、目視で原文の図（立体構造自体の描画や図示されたアミノ酸配列）および図中ラベルを読み取り、Selection Syntax を構成した。

いずれの場合も、実際に構成した Selection Syntax を用いて PyMOL で SBE を可視化し、結果がテキストの記述や図と整合していること確認した。今回対象とした立体構造の領域に割り当てられない SBRE 候補については、SBE 定義を与えなかった。

4.4 後処理

テキスト・構造アノテーションの結果を統合し、SBRE と SBE 定義のペアを作成した。次に SBRE ごとの SBE 定義に含まれる原子 ID の集合を比較し、一致した場合は同一の SBE を指示する同義 SBRE であるとみなしてグループ化した。このグループ内部にある SBRE たちを、同義 SBRE のバリエーションと呼ぶことにする。

5 結果と考察

5.1 アノテーション時間

網羅性と精度を重視したため、テキスト側のアノテーションに 10 時間程度、構造のアノテーションに 4 時間程度を要した。

5.2 統計的性質

SBE の出現頻度

全 SBRE の出現回数の総計は 445 回であり、異なる SBE は 41 個存在した[ii]。同義 SBRE グループ、つまり SBE ごとに言及数を数えると、出現数 top 5 までの SBE で約 80%（80.2%）、top10 で約 90%（89.9%）の言及をカバーしていた（表 1）。

SBRE のバリエーション

複数回出現する SBE の多くは複数の SBRE により言及されていた。固有名詞的な SBRE（たとえば MTBD や flap）を割り当てられる SBE は言及回数の割に SBRE のバリエーションが少なかった。一方、同時に複数の異なる基底的 SBE に言及する場合は SBRE のバリエーションが増加した（表 1）。

表 1 ：出現回数(#Obs)上位 12 の SBE における SBRE バリエーション数(#Var)と原子数(#Atom). 括弧内の数字は出現回数の内訳.

#Obs	#Var	#Atom	SBREs
126	4	1121	MTBD (112), OADγ-MTBD (8), the microtubule-binding domain (4), MTBDs (2)
122	4	1554	LC1 (118), axonemal dynein light chain-1 (2), dynein light chain-1 (1), the leucine-rich repeat (1)
48	4	155	flap (45), the other (1), a β-hairpin structure (1), a β-hairpin structure extended from the globular domain (1)
40	13	2675	LC1-MTBD (27), LC1-complexed MTBD (2), The complex of outer-arm dynein light chain-1 and the microtubule-binding domain of the γ heavy chain (1), MTBD-complexed LC1 structure (1), the new structure (1), the complex (1), LC1-complexed MTBD structure (1), the structure (1), the complex structure (1), the LC1-complexed MTBD from OADγ (1), the two proteins (1), our complex structure (1), LC-MTBD (1)
21	2	103	H5 (20), one site (1)
13	2	11	Arg-79 (11), R79Q (2)
10	8	258	two important sites (2), both regions of the MTBD (2), these sites (1), two regions (1), two interaction sites (1), these two sites (1), two sites of interaction (1), the H5 and flap interaction sites of the MTBD (1)
8	2	7	Thr-57 (6), T57A (2)
6	5	30	both regions of the MTBD (2), Three residues Thr-57, Arg-79, and Tyr-102 (1), the only residues (1), these three residues (1), both in MTBD (1)
6	1	12	Tyr-102 (6)
4	2	13	Met-182 (2), M182A (2)
3	3	151	20 in the MTBD (1), The residues of MTBD interacting with LC1 (1), the interaction sites (1)

たとえば「two important sites」として言及される SBE は、flap および H5 に対応する 2 つの頻出 SBE に同時に言及する際に用いられていた。これら flap と H5 は、いずれも LC1 と MTBD の相互作用界面を定義するという類似の役割を持つが、アミノ酸配列上は分離した領域である。このため、これらを役割から総称するには句として言及せざるを得ず、表現が多様化するものと考えられる。

SBE の出現頻度と大きさ

頻出の SBE のサイズは概ね 3 種類に分類できた（表 1）：(1) 数千原子程度のものは鎖単位（今回はドメイン単位と一致）やそれらの複合構造、(2) 数百原子程度のものは、二次構造数個からなる小さな部分構造やその複合構造、(3) 数十原子程度のものはアミノ酸残基やそれらの集合

[ii]複数の基底的 SBE からなる複合 SBE も冗長に数えた。

に相当する。このような記述の階層性の特徴は、「構造全体から徐々に詳細にフォーカスし、現象の原因を局在化していく」という構造生物学論文の還元主義的な傾向に矛盾しない。

5.3 統計的性質の解釈

個々の鎖やドメイン、アミノ酸残基単位という自明な階層だけでなく、中間的スケールの SBE も高頻度に言及されることは注目に値する。これは、鎖・ドメイン・アミノ酸残基といったタンパク質構造における階層性に対応した分節単位への言及を採集するだけでは、構造生物学論文中の記述を網羅的かつ体系的にデータ化するには不十分であることを示唆している。逆に、このような中間的な SBE がどのように認識・言及されるに至るか理解できれば、構造生物学者が無意識的に行っている立体構造の分節化の仕組みが明らかになり、SBE 定義の自動解決方策の発見につながると期待される。

5.4 SBE 定義推測の論拠

すでに述べた通り、出現数 top 5 までの SBE で全体の言及数の 80%を占めていたが、これらは全て論文中の図に基づいて SBE 定義を推測する必要があった。もし、例えば「残基名+残基番号」形の言及など、テキストのみから SBE 定義を推測可能な SBRE に絞ってアノテーションした場合、今回採取できた全言及の 20%以下しか得られなかったといえる。

理想的には、残基単位の SBRE に関しては構造データ中の残基名と ID を用いたパターンマッチングにより、鎖やドメイン単位の SBRE は既存の立体構造・配列の分類に基づいて自動で推測できる可能性もある[iii]。一方、flap や H5 に相当する SBE は対象の論文から定義を読み解くか、その SBE を初めて定義・言及した論文 [8]まで遡る必要がある[iv]。

ここからも、中間的スケールの SBE 定義を高精度に決定することが高品質の構造–言語ペアデータを得るためには重要であることがわかる。今後本プロジェクトを大規模化していくには、論文中の図からの直接的な SBE 定義抽出の効率化が強く求められる。

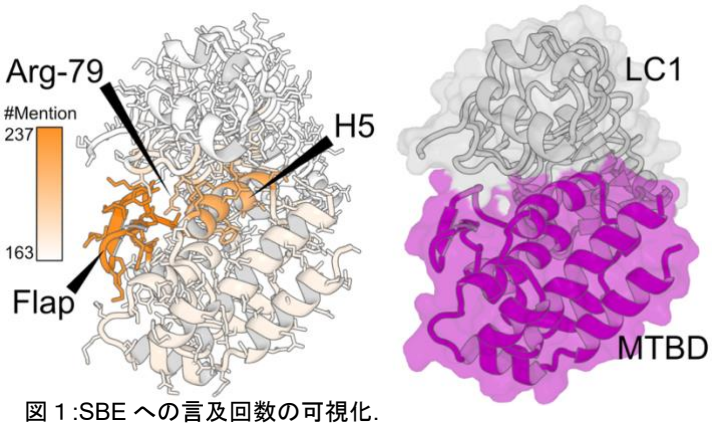

図 1 :SBE への言及回数の可視化.
（左）論文中の SBE 言及回数に基づいて立体構造データ（PDBID: 6L4P）を着色したもの. オレンジ色が濃いほど言及回数が多い.（右）LC1 を灰色, MTBD を紫で示した. 言及が LC1 と MTBD の界面に集中していることがわかる.

5.4 言及回数の立体構造へのマッピング

今回対象とした論文は「LC1 と MTBD がどのように相互作用するのか不明」[v]という疑問の解決が主題となっている。したがって、LC1 と MTBD の相互作用界面に言及が集中すると予想される。

実際に論文中の言及回数を SBE に割り当てると、界面部分、とくに H5 と flap への言及が集中していることが視覚的に把握できる（図 1）。逆に言えば、当該論文の著者らはこのような重み付けで立体構造を「見て」いたのだと考えられる。

このように構造生物学者が立体構造のどこに注目してどのような解釈を行うか、系統的かつ大規模にデータ化できれば、構造生物学という営み自体をモデル化するための基盤となることが期待される。

6 まとめ

具体的な構造生物学論文と立体構造データを用い、構造言及（SBRE）と言及される立体構造上の実体（SBE）を網羅的に紐づけるアノテーションに取り組んだ。構造言及と実体のペアデータ作成が、構造生物学者による注目領域を捉えるために有用であることが示唆された。今後は PDB-descriptome プロジェクトの大規模化に向け、アノテーション効率化手法の検討を進める。□

[iii] これら外部情報が論文の表現と一致する保証はない。
[iv]初出の論文図表を読み解く必要性は依然として残る。
[v] 原文では、要旨において—It has recently been shown that axonemal dynein light chain-1 (LC1) binds to the microtubule-binding domain (MTBD) of OADγ, leading to a decrease in its microtubule-binding affinity. However, it remains unclear how LC1 interacts with the MTBD and controls the microtubule-binding affinity of OADγ. —と課題定義されている。

謝辞

本プロジェクトの具体化に向けてご助言いただいたBLAH9参加者、およびTextAEについてご指導いただいた金進東博士にお礼申し上げます。本研究はROIS-DS-JOINTの支援、および(公財)村田学術振興・教育財団の助成を受けて行われました。

参考文献

[1] H. Berman, K. Henrick, and H. Nakamura, “Announcing the worldwide Protein Data Bank,” *Nat. Struct. Mol. Biol.*, vol. 10, no. 12, pp. 980–980, Dec. 2003,
doi: 10.1038/nsb1203-980.

[2] 佐久間航也, 丹羽智美, “タンパク質立体構造データと紐づけたコーパス作成の試み”, 言語処理学会第31回年次大会 発表論文集, 2025年3月.

[3] 佐久間航也, 丹羽智美, “タンパク質立体構造について語れる AI を将来的に実現するためのデータセット作成の試み”, 人工知能学会全国大会論文集, *2025, JSAI2025*, Jul. 2025,
doi: 10.11517/pjsai.JSAI2025.0_2Win526

[4] A. Toda, Y. Nishikawa, H. Tanaka, T. Yagi, and G. Kurisu, “The complex of outer-arm dynein light chain-1 and the microtubule-binding domain of the γ heavy chain shows how axonemal dynein tunes ciliary beating,” *J. Biol. Chem.*, vol. 295, no. 12, pp. 3982–3989, Mar. 2020,
doi: 10.1074/jbc.RA119.011541

[5] Toda, A., Nishikawa, Y., Tanaka, H., Yagi, T., Kurisu, G., “Crystal structure of the complex between the axonemal outer-arm dynein light chain-1 and microtubule binding domain of gamma heavy chain.”, Feb. 2020,
doi: 10.2210/pdb6l4p/pdb

[6] Kim, J.D. and Wang, Y., “PubAnnotation - a persistent and sharable corpus and annotation repository. In BioNLP,” *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp 202–205, Jun. 2012,
url: https://aclanthology.org/W12-2425/

[7] Schrodinger, “The PyMOL Molecular Graphics System, Version 1.8,” Nov. 2015.

[8] Kato YS, Yagi T, Harris SA, Ohki SY, Yura K, Shimizu Y, Honda S, Kamiya R, Burgess SA, Tanokura M. “Structure of the microtubule-binding domain of flagellar dynein.”, *Structure,* vol 22, no 4, pp. 1628-38, Nov. 2014,
doi: 10.1016/j.str.2014.08.021