AI を用いたタンパク質間複合体予測から、

機能未知遺伝子の機能を推察する

河口 真一
大阪大学 大学院生命機能研究科

1. はじめに

細胞内で生命現象の物理化学的反応を担っているのは、多くの場合、タンパク質である。タンパク質は、20 種類のアミノ酸が連なったポリマーであり、アミノ酸配列に応じた立体構造に折り畳まれる。生物のゲノム DNA にコードされているタンパク質の種類は、ヒトで約 20,000 程度、モデル生物のショウジョウバエでも、約 12,000 種類も存在する。世界中の研究者によって、各遺伝子がコードするタンパク質の機能を調べる研究が続けられており、多くの遺伝子について分子機能が明らかにされてきた。しかし、未だに機能が解明されていない遺伝子が 20%以上も残されており、分子生物学上の大きな謎となっている。これら機能未知な遺伝子の分子機能に関する手がかりを得る上で、直接的に相互作用する分子を同定することは、極めて有用である。なぜなら、機能がわかっているタンパク質との相互作用が明らかになれば、機能未知遺伝子も、同じ経路、あるいは関連する経路で働いている可能性が高いからである。相互作用するタンパク質を実験的に同定する研究も広く行われているが[1,2]、時間と労力が多くかかるため、迅速なスクリーニング手法が望まれている。タンパク質間の相互作用は、各タンパク質の立体構造表面の特性によって決まる。そのため、タンパク質の立体構造情報を用いて、ドッキング法により結合を予測する試みも行われてきたが、未だ信頼できる予測法は確立されていない。

最近、DeepMind 社によって開発された AlphaFold2 という AI プログラムは、アミノ酸配列の共進化情報を利用して、タンパク質の立体構造を高精度で予測できるとして注目されている[3]。さらに、タンパク質の複合体構造も予測できることが期待されている[4]。そこで本研究では、AlphaFold2 を利用して、迅速なタンパク質複合体のスクリーニングを行い、機能予測の一助とすることを試みた。

2. ALphaFold2 プログラムを用いたタンパク質複合体構造予測

AlphaFold2 による立体構造予測は、主に２つのステップに分けることができる。

１）アミノ酸配列の多重アライメント作成

２）立体構造、及び複合体構造の予測

予測したいタンパク質のアミノ酸配列と類似したアミノ酸配列を、数多くの生物種から収集し、アライメントすることによって、互いに依存した変化（共進化）が推測される部分を抽出する。共進化している部分は、立体構造上、距離的に近接すると仮定して、立体構造のモデルを構築する。本研究では、ショウジョウバエのほぼ全てのタンパク質についての多重アライメントを作成・保存し、それらを再利用することによって、大規模な複合体予測に必要な時間を短縮させている。

AlphaFold2 は、数多くの立体構造を学習し、予測に用いるモデルパラメータを５パターン備えている。オリジナルの計算フローでは、５つのパターンに対応する５つの構造を、１つの GPU で逐一計算する仕様であったが、大阪大学サイバーメディアセンターの SQUID では、GPU を５つ用いて並列に計算するように変更し、予測時間をさ

らに短縮している（NEC ソリューションイノベータ株式会社との共同研究）。これらの変更によって、１日、１アカウントあたり、約 150 ペアのタンパク質複合体予測ができるようになり、通常のコンピュータと比べて、約１００倍速く計算することが可能になった。

3. ２０種類のタンパク質間の複合体構造予測

3.1 信頼性スコアの分布

本研究では、ショウジョウバエの生殖細胞内に存在するタンパク質集積体の１つであるヌアージュ構造に着目した [5]。ヌアージュでは、piRNA という小分子 RNA が産生されており、piRNA と結合した PIWI ファミリータンパク質が、ゲノム DNA の損傷を引き起こすトランスポゾンの発現を抑制している。ヌアージュに局在する１８種類のタンパク質と、機能的に関連するミトコンドリア局在タンパク質、２種類を合わせた合計２０種類のタンパク質について、１：１の総当たりで４００ペアの結合予測を行った。

表１：本研究で用いた２０種類のタンパク質

タンパク質	残基数	局在	ドメイン
vas	661	Nuage	DEAD
spn-E	1434	Nuage	DEAD
tej	559	Nuage	Lotus, Tudor
tapas	1222	Nuage	Lotus, Tudor
qin	1857	Nuage	RING, Tudor
Kots	892	Nuage	Tudor
krimp	746	Nuage	Tudor
squ	241	Nuage	
mael	462	Nuage	Meal
aub	866	Nuage	PAZ, Piwi
AGO3	867	Nuage	PAZ, Piwi
papi	576	Mitochondria	Tudor, KH
vret	691	Nuage	Tudor
bel	801	Nuage	DEAD
zuc	253	Mitochondria	PLD-like
cup	1117	Nuage	
tral	657	Nuage	Lsm
me31B	459	Nuage	DEAD
shu	455	Nuage	PPIase
BoYb	1059	Nuage	DEAD, Tudor

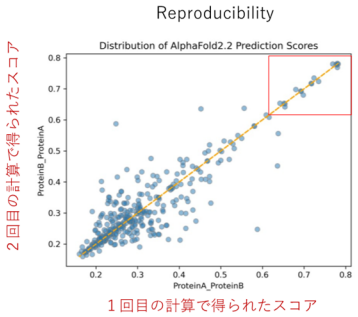

図１：AlphaFold2 の予測スコア分布

AlphaFold2 は、各複合体の予測構造に対して、タンパク質単体での構造と、複合体界面の構造を評価し、信頼性スコアを出力する。スコアは、０から１の範囲であり数値が大きいほど、信頼性が高いことを意味する。総当たりで計算すると、同じペアのタンパク質複合体を２回計算することになる。１回目と２回目の信頼性スコアをプロットしたところ、スコアが低い（0.5 以下）場合には、スコアのばらつきが見られた。スコアが高い（0.6 以上）場合には、比較的再現性の良い結果が得られたことから、スコア 0.6 以上のペアを複合体形成の候補とした。

3.2 実験的検証

このスクリーニングにおいて、スコアが 0.6 以上のペアは、１３種類見られた。そのうち、７つについては、既に結合が報告されているペアであった。このことは、AlphaFold2 が複合体ペアを予測できることを示唆している。残りの６ペアについては、これまでに報告がない新規な結合候補であった。その中で、Spn-E タンパク質と Squ タンパク質のペアに注目した。Spn-E は、RNA ヘリケースであり、ヌアージュにおける piRNA の産生に重要であることが知られている [5]。一方で、Squ タンパク質は、機能を推測できるドメインもなく、詳しい機能解析もされていない、機能未知タンパク質である。両者が複合体を形成するかどうか、実験的検証を行った（図２）。

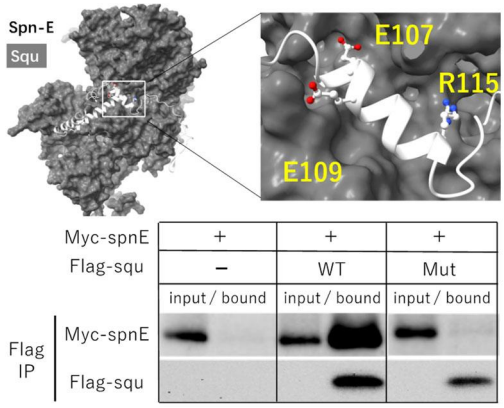

図２：AlphaFold2 が予測した SpnE-Squ 複合体と共免疫沈降法による実験的検証

Myc タグ、Flag タグをそれぞれ付加した Spn-E と Squ を昆虫培養細胞内で発現させた。Flag タグに特異的な抗体で免疫沈降を行い、Myc-Spn-E の結合を Myc 特異的抗体で検出した。その結果、Squ タンパク質が共発現している場合にのみ、Spn-E が結合画分に検出され、２つのタンパク質が結合し、複合体を形成することが明らかになった（図２下中央）。さらに、Squ のα-ヘリックスの１つが、SpnE 表面の溝に埋まっている構造が予測されたことから、この部分が結合に寄与していると考えられた。実際に、このα-ヘリックスと Spn-E との間で塩橋を形成すると予測された残基を変異させたところ、その変異 Squ は、Spn-E との結合が損なわれていた（図２下右）。したがって、AlphaFold2 による予測構造は正しいことが示唆された。これらの結果から AlphaFold2 を用いた複合体のスクリーニングが有効であると考えられる。機能未知である Squ の役割は、Spn-E に結合することで、Spn-E の安定性や活性を制御しているか、あるいは、Spn-E が他の因子と相互作用することを制御していると推察される。

4. 細胞内の全タンパク質をスクリーニング

AlphaFold2 による複合体スクリーニングを、細胞内の全タンパク質に対して適用することを試みた。ここでは、生殖細胞と体細胞の両方で、多様な機能が知られている Piwi タンパク質を用いた。Piwi タンパク質は、piRNA と結合し、それと相補的な配列をもつトランスポゾンの mRNA を切断することによって、有害なトランスポゾンの発現を抑制している。さらに、核内では染色体のヘテロクロマチン化を誘導し、転写レベルで発現を抑制することも知られている[6]。

ショウジョウバエのゲノムには、約 12,000 種類のタンパク質がコードされている。Piwi タンパク質と全タンパク質との１：１結合スクリーニングを行った。大阪大学サイバーメディアセンターの SQUID を用いて、１日に約１００ペアの構造予測を行い、４ヶ月ほどで計算を終えることができた。その結果、大部分のペアについては、複合体に対する信頼性スコアが 0.3 未満と低く、結合していないと予測された。スコアが 0.6 以上のものは、全体の約１％に相当する 177 個見つかっている（図３）。

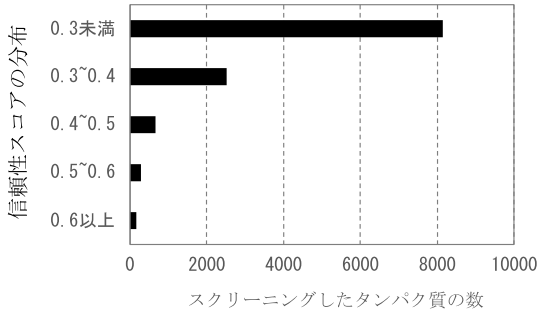

図３：Piwi との複合体予測におけるスコア分布

この中には、これまでに Piwi と結合することが報告されている Arx タンパク質に加えて、いくつかの機能未知タンパク質も含まれていた。特に、Piwi と CG33703 という機能未知タンパク質の複合体については、高い信頼性スコア(0.82)をもって予測された。CG33703 遺伝子は、ゲノム上に多数のコピーがあることが知られているが、その機能については全く不明である。

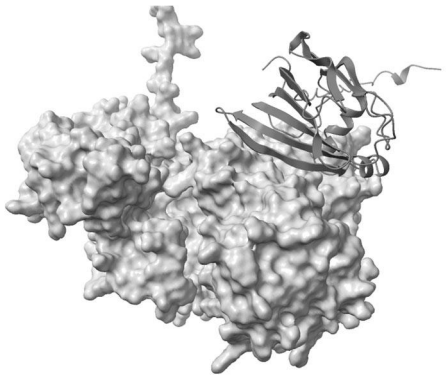

図４：予測された Piwi と CG33703 との複合体
Piwi: 空間充填モデル
CG33703: リボンモデル

今後、実験的検証を行い、新規な結合ペアを見出すことによって、機能解析を展開できると期待される。

5. おわりに

タンパク質のアミノ酸配列から立体構造を予測する試みは、古くから行われてきた。その過程で、2000 年代に日本で行われたタンパク３０００プロジェクトが、タンパク質の構造パターンの多くを明らかにしたことによって、ドメイン単位での立体構造予測が可能となった。現在の AlphaFold2 では、ドメイン間の配置もかなりの精度で予測できるようになっている。

現在、AlphaFold2 のような AI プログラムの開発が盛んに行われており、生命科学の発展に大きく貢献することが期待されている。しかし、そのような計算機科学と、それを利用する実験科学の間には、まだ隔たりがあるのが実情である。両者の隔たりを埋めて、最新の計算機科学を、実験科学が迅速に取り込むことができる環境が望まれる。今後は、癌化した細胞などで過剰発現が見られるタンパク質や、実験的に扱うのが困難な毒性タンパク質の相互作用を in silico でスクリーニングし、新規な結合候補を示すことによって、タンパク質の機能解析を進展できると期待される。

このプロジェクトを進めるにあたり、サイバーメディアセンター応用情報システム研究部門の伊達進教授、高性能計算・データ分析融合基盤協働研究所の曽我隆特任准教授（常勤）に大変お世話になりました。感謝の意を表します。

参考文献

1. Giot, L. *et al.* A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727–1736 (2003).
2. Guruharsha, K. G. *et al.* A protein complex network of Drosophila melanogaster. *Cell* **147**, 690–703 (2011).
3. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034 Preprint at https://doi.org/10.1101/2021.10.04.463034 (2021).
5. Lin, Y., Suyama, R., Kawaguchi, S., Iki, T. & Kai, T. Tejas functions as a core component in nuage assembly and precursor processing in Drosophila piRNA biogenesis. *J Cell Biol* **222**, e202303125 (2023).
6. Huang, X. & Wong, G. An old weapon with a new function: PIWI-interacting RNAs in neurodegenerative diseases. *Translational Neurodegeneration* **10**, 9 (2021).