Vol.2019-MPS-123 No.56
Vol.2019-BIO-58 No.56
2019/6/19

配列類似性ネットワークに基づく高感度な遠縁タンパク質検索

牧垣 秀一朗[1,a)]　石田 貴士[1,b)]

概要：相同タンパク質検索は，その構造や機能の予測・進化の解明に重要であり，タンパク質構造・機能研究の中心技術となっている．そのなかで，配列類似度は低いが構造や機能は類似しているタンパク質が多く存在していることが知られるようになり，そのような遠縁のタンパク質をデータベース中から発見する研究が進められている．本研究では，アミノ酸配列データベース内に配列類似性に基づくネットワーク構造を定義し，そのネットワークを検索することで，遠縁のタンパク質を高感度に検出することを試みた．ネットワーク構造を定義することにより，進化の関係を検索に導入することができ，共通の祖先を経由した遠縁のタンパク質を発見できると考えられる．また，アミノ酸配列データベースとタンパク質立体構造データベースを本手法で結合することにより，アミノ酸配列から遠縁の構造既知のタンパク質を発見することも期待できる．

1. 序論

タンパク質は生物学や生化学，および製薬科学における重要な高分子であり，タンパク質の機能を明らかにするためには，タンパク質の構造と機能の関係を理解することが不可欠である．類似した機能を持つタンパク質はしばしば進化的な関連があり，これらのタンパク質はホモログと呼ばれている．タンパク質の機能を推定するために，これらの関連を明らかにし，その構造を研究することが分子生物学において重要である．タンパク質構造はX線結晶学または核磁気共鳴法などの実験的手段によって決定することができ，得られたタンパク質立体構造はProtein Databank（PDB）に登録され，誰もが閲覧することができる．しかし，タンパク質構造を決定するための実験方法は改良されているにもかかわらず，アミノ酸配列を明らかにすることができる速度は，対応するタンパク質の構造を確かめる我々の能力を上回っている．したがって，タンパク質構造予測法，すなわち所与のアミノ酸配列の3次元構造モデルを生成するための計算技術の使用は，依然として不可欠である．

これまでタンパク質構造を予測するための様々な手法が提案されており，物理化学的シミュレーション（*de novo*とも呼ばれる）を用いた方法と，ホモロジーモデリングと呼ばれる方法がある．後者は，テンプレートおよび予測対象アミノ酸配列とのアライメントに基づいて構造を予測する．テンプレートとなる構造は，相同性検索で見いだされたホモログタンパク質の構造を用いる．予測対象タンパク質と構造的に類似した良いテンプレートと，予測対象配列とのアライメントができれば，*de novo*法と比べて予測モデルがはるかに正確である．そのため，現在，テンプレートベースのモデリング方法が最も実用的である．

ホモロジーモデリングでは，テンプレートタンパク質なしで予測構造モデルを生成されることができない．相同なタンパク質を発見できれば，ホモロジーモデリングの対象とできるタンパク質を増やすことができる．そのため，相同なタンパク質をデータベース中から検出することが重要であると考えられてきた．そのなかで，配列類似度は低いが構造や機能は類似しているタンパク質が多く存在していることが知られるようになり，そのような遠縁のタンパク質をデータベース中から発見する研究が進められている．古くはFASTA[1]およびBLAST[2]による相同性検索研究が知られており，PSI-BLAST[3]およびDELTA-BLAST[4]のような，複数の配列アライメントに基づくプロファイルを用いた手法が，高い精度で相同性を検出できている．また，配列プロファイルを用いた他の手法として，隠れマルコフモデルを用いた遠縁のホモログを検出する手法が存在し，HHpred[5]は構造予測ベンチマークで優れた性能を発揮している[6], [7].

配列類似度が低くても，相同なタンパク質を複数発見す

[1] School of Computing, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, 152–8550, Japan
[a)] makigaki@cb.cs.titech.ac.jp
[b)] ishida@c.titech.ac.jp

ⓒ 2019 Information Processing Society of Japan

Vol.2019-MPS-123 No.56
Vol.2019-BIO-58 No.56
2019/6/19

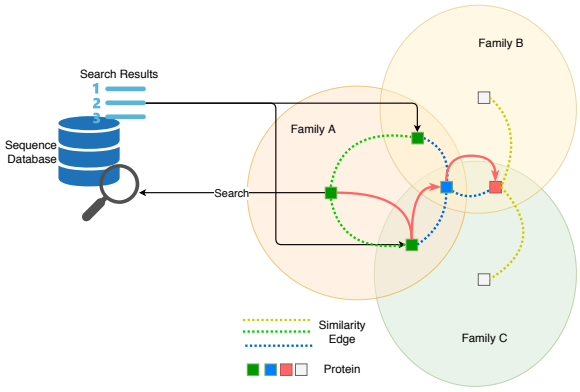

図 1　データベース中の各配列を検索し，検索結果に対してのみ検索配列からのエッジを定義することでネットワークを構築する（点線）．目的は，配列類似度が低い同ファミリー、または相同タンパク質（赤と青で示されるタンパク質）を検出することである．これらのタンパク質は，赤い線で示されるような，共通して類似するノードを複数（2 または 3）経由することで発見できる．

ることができれば，より精度の高い構造予測ができる場合もある．しかしそのためには，配列類似度を直接比較するだけでは検出できない遠縁のタンパク質を検出することが必要となる．本稿では，アミノ酸配列データベース内に配列類似性に基づくネットワーク構造を定義し，そのネットワークを検索することで，遠縁のタンパク質を高感度に検出する手法を示す．その検出性能を HHsearch，PSI-BLAST，DELTA-BLAST と比較した．最後に，提案手法の問題点と今後の展望を示す．

2. 手法

手法の概要を**図 1** に示す．アミノ酸配列データベース（DB）に含まれる各配列をノード，配列類似度をエッジの重みとするネットワーク構造を定義し，これを配列類似性ネットワークと呼ぶ．配列類似度は配列相同性検索手法により出力される値やアミノ酸一致度など，任意の評価値を用いることができる．配列類似度はすべてのノード間で定義することが可能であるため，本来は配列類似性ネットワークはフルメッシュのネットワークとなる．しかし，DB サイズが大きくなった場合に類似度の計算量が爆発してしまう．また，相同性があるとは言えないほど配列類似度が低いエッジには検索に含まれるべきではない．以上のことから，DB 内で配列相同性検索手法を用いて配列を検索し，相同であると検出されたアミノ酸配列のみを連結する．

本稿では DB に SCOP[8], [9] のサブセットである SCOP40 を用いた．SCOP とは，機能と構造分類に基づいて，タンパク質をクラス，フォールド，スーパーファミリー，ファミリー，およびドメインに分類した DB である．また，SCOP40 は配列一致度が 40 % 未満のドメインのみを含むサブセットである．配列類似度には DELTA-BLAST の Evalue を用いた．SCOP40 の各ドメインをクエリとし，DELTA-BLAST を用いて SCOP40 内を検索し，検索結果に現れるドメイン間にのみエッジを定義する．定義されたネットワーク構造は GraphML 型式 [10] で出力される．本稿ではグラフ指向データベースソフトウェアとして Neo4j[11] を用い，出力された GraphML ファイルをインポートすることで，検索可能な DB を構築した．

次に検索手法を示す．まず，検索を開始するノード（開始ノード）を決定する必要がある．本稿では DELTA-BLAST で DB を検索することで，この開始ノードを決定した．疑陽性を減らすために，Evalue 閾値を 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000 の順で変化させ，1 つ以上の検索結果が出力された時点での検索結果を開始ノードとする．この開始ノードから 3 ホップで到達できるノードを検索する．経路上の Evalue を合計し，昇順に並び替えて最終出力とする．これらの検索手法の実装には，Neo4j 独自のグラフ検索言語である Cypher を用いた．

3. 結果

テストデータとして，SCOP からドメインを多く含むスーパーファミリーを選び，そのスーパーファミリーに分類されているすべてのドメインを検索クエリとした．同じスーパーファミリー内のドメインを検出できた場合を正例，そうでない場合を負例とし，経路上の合計 Evalue をスコアとした．性能評価には ROC_{50} 曲線下面積（AUC_{50}）の平均を用いた．ROC_{50} とは，偽陽性の結果が 50 個出力された時点で打ち切った ROC であり，AUC_{50} とは，その曲線下面積を全真陽性の個数と偽陽性の打ち切り閾値で正規化した値である．評価に用いたテストデータと，検索性能評価の結果を**表 1** に示した．

ネットワーク構築時に使用した DELTA-BLAST と比べると，`b.1.2` と `d.58.7` を除くテストデータにおいて性能が向上した．また，PSI-BLAST と比較すると，すべてのデータセットにおいて性能が向上した．State-of-the-art と言われる HHsearch との比較では半数のテストデータで性能が低下し，`b.1.1` や `g.39.1` で特に大きな性能差が出る結果となった．

4. 結論

本稿では，アミノ酸配列データベース内に配列類似性に基づくネットワーク構造を定義し，そのネットワークを検索することで，遠縁のタンパク質を高感度に検出する手法を示した．その性能を PSI-BLAST，DELTA-BLAST，HHsearch と比較した．ネットワーク構築時に使用した DELTA-BLAST と比べると，一部を除くテストデータで性能が向上した．一方，State-of-the-art と言われる HHsearch との比較では，半数のテストデータで性能が低下し，一部のテストデータでは大きく性能を落とす結果となった．

ⓒ 2019 Information Processing Society of Japan

表 1 #はスーパーファミリーに含まれるドメインの数を示す．小数値はすべて AUC_{50} であり，太字は最も高い AUC_{50} を示している．（PSI と DELTA は，それぞれ PSI-BLAST と DELTA-BLAST を示す．）

SCSS	Description	#	HHsearch	Proposed	PSI	DELTA
a.4.1	Homeodomain-like	57	0.57	**0.65**	0.45	0.54
a.4.5	“Winged helix” DNA-binding domain	55	0.35	**0.53**	0.26	0.28
b.1.1	Immunoglobulin	46	**0.70**	0.58	0.52	0.65
b.1.2	Fibronectin type III	53	**0.79**	0.74	0.70	**0.79**
b.36.1	PDZ domain-like	44	0.69	**0.71**	0.69	0.70
c.37.1	P-loop containing nucleotide triphosphate hydrolases	37	0.61	**0.86**	0.38	0.55
c.47.1	Thioredoxin-like	44	0.60	**0.70**	0.30	0.45
d.58.7	RNA-binding domain	74	**0.60**	0.59	0.53	**0.60**
g.37.1	β-β-α zinc fingers	59	**0.73**	0.62	0.44	0.48
g.39.1	Glucocorticoid receptor-like (DNA-binding domain)	47	**0.72**	0.48	0.20	0.28

多くの類似配列を持つ “ハブ” となるようなドメインが経路上に存在している場合，このノードを経由することで多くの偽陽性を発生させる可能性がある．しかし，この “ハブ” ノードを経由することで，遠縁の相同タンパク質を発見できる可能性もあるため，扱いには注意を要する．この問題に関しては，検索結果の経路を調査することで解決できると考えている．今後の展望として，アミノ酸配列 DB とタンパク質立体構造 DB を本手法で結合することにより，アミノ酸配列から遠縁の構造既知のタンパク質を発見する手法を構築することを考えている．

参考文献

[1] Pearson, W. R. and Lipman, D. J.: Improved tools for biological sequence comparison., *Proceedings of the National Academy of Sciences*, (online), DOI: 10.1073/pnas.85.8.2444 (1988).
[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: Basic local alignment search tool., *Journal of molecular biology*, Vol. 215, No. 3, pp. 403–10 (online), DOI: 10.1016/S0022-2836(05)80360-2 (1990).
[3] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J.: Gapped BLAST and PSI-BLAST:a new generation of protein database search programs, *Nucleic Acids Research*, Vol. 25, No. 17, pp. 3389–3402 (online), DOI: 10.1093/nar/25.17.3389 (1997).
[4] Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J. and Madden, T. L.: Domain enhanced lookup time accelerated BLAST, *Biology Direct*, Vol. 7, pp. 1–14 (online), DOI: 10.1186/1745-6150-7-12 (2012).
[5] Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N. and Alva, V.: A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core, *Journal of Molecular Biology*, No. Table 1, pp. 1–7 (online), DOI: 10.1016/j.jmb.2017.12.007 (2017).
[6] Hildebrand, A., Remmert, M., Biegert, A. and Söding, J.: Fast and accurate automatic structure prediction with HHpred, *Proteins: Structure, Function and Bioinformatics*, Vol. 77, No. SUPPL. 9, pp. 128–132 (online), DOI: 10.1002/prot.22499 (2009).
[7] Meier, A. and Söding, J.: Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling, *PLoS Computational Biology*, Vol. 11, No. 10, pp. 1–20 (online), DOI: 10.1371/journal.pcbi.1004343 (2015).
[8] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, (online), DOI: 10.1016/S0022-2836(05)80134-2 (1995).
[9] Fox, N. K., Brenner, S. E. and Chandonia, J. M.: SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures, *Nucleic Acids Research*, Vol. 42, No. D1, pp. 1–6 (online), DOI: 10.1093/nar/gkt1240 (2014).
[10] GraphML Team: The GraphML File Format, GraphML Team (online), available from ⟨http://graphml.graphdrawing.org/⟩ (accessed 2019-05-14).
[11] Neo4j, Inc.: Neo4j Graph Platform – The Leader in Graph Databases, Neo4j, Inc. (online), available from ⟨https://neo4j.com/⟩ (accessed 2019-05-14).

ⓒ 2019 Information Processing Society of Japan