

タンパク質構造比較のための新規パラメータの開発とその評価

黒田研究室

05251508

矢島 龍

【背景・目的】 PDB [1]のタンパク質立体構造情報数が2006年を以って3万件を突破した現在に於いて、SCOP [2]、CATH [3]、FSSP [4]の様なタンパク質を構造毎に分類したデータベースの必要性は益々高まり、またそれらを構築するためのより迅速な構造比較ツールが求められている。さて、PDBに登録されているタンパク質立体構造情報に於けるタンパク質の位置、向きは類似の構造を有しているもの同士であっても先ず一致せず、従って原点からのベクトルでは比較不可能であるため、構造アライメントツールとして著名であるCATHに於けるSSAP [5]、FSSPに於けるDALI [6]は構造比較を行うための主パラメータとして α 炭素間距離(或いはベクトル)を使用している。この際、基準とする α 炭素を一つに絞っての正確なアライメントは不可能であり、半網羅的解析を採る必要が生じ、計算時間の増加に繋がっている。この問題は全タンパク質に普遍的パラメータを算出するための単一基準点を設ける事で解決出来ると考えられ、そこで本研究では重心を基準とするパラメータ、即ち重心- α 炭素間距離(Distance)、 α 炭素から重心へのベクトルと α 炭素から α 炭素に隣接する原子へのベクトルとの角度(Angle CB, C, N)を提案し、その評価を行う。重心はタンパク質進化の過程に於ける大きなDeletion、Insertionの影響を受け易い事は容易に想像出来るが、アライメントにてマッチした残基のみから算出した重心を用いた各パラメータの再計算によって補えられ、本研究ではその検証も行う。

【方法】 タンパク質立体構造情報はCATHのものを用いた。重心Pは主鎖原子及び β 炭素原子(Glycineでは仮想 β 炭素座標を算出し使用)(再計算ではマッチした残基の原子のみ使用)の座標の合計を原子数で割る事により算出し、Distanceは次式1、各Angleは次式2より算出した。パラメータの評価は、DSSP[7]にて算出される二面角PHI・PSIと溶媒露出面積ACC、 β 炭素を中心とする半径R球内にて β 炭素より目視可能である体積Visible Volume(VV)[8]、 α 炭素を中心とする半径R球内に存在する α 炭素数 C_{α} -Density(CAD)[9]

と比較する事で行った。パターンを目視に依る比較は曖昧であり、本研究では二タンパク質のアライメント残基対の相関係数による比較法を新たに開発し、これより比較・評価を行った。アライメントツールとしてはSSAPを用いた。

$$|C_{\alpha}P| = \sqrt{(C_{\alpha_x} - P_x)^2 + (C_{\alpha_y} - P_y)^2 + (C_{\alpha_z} - P_z)^2} \quad (1)$$

$$\angle PC_{\alpha}X = \cos^{-1} \left(\frac{C_{\alpha}P \cdot C_{\alpha}X}{|C_{\alpha}P| |C_{\alpha}X|} \right) \quad (X: C_{\beta}, C, N) \quad (2)$$

【結果・考察】 CATH各層に於ける各パラメータの結果をTable 1に、再計算したパラメータの結果をTable 2にそれぞれまとめた。CATHはClass, Architecture, Topology, Homologous Super Family, Sequential Familyの順に構造類似度が増加するDatabaseであり、従って全パラメータに於ける構造類似度の増加に伴う相関係数の増加が示され、今回の比較法の有効性は示された。相関係数の高さはアライメント能を示し、類似度が高いタンパク質間に於いて強い相関を示さないパラメータは構造アライメントに向かないと、また相関係数の標準偏差が小さい程、アライメント残基対の類似度を示すRMSDとの負の相関が強い程そのパラメータは正確であると言える。重心基準パラメータとして提案した三つのAngleは、CAD・Distanceに若干劣りはするが高いと言える相関係数を、また群を抜くRMSDとの相関を示したことから構造アライメントパラメータとしての有用性が示唆された。Distanceは高い相関係数を示すもののRMSDとの相関が非常に低いために少々使い難く思われる。しかしアライメントに基く再計算により一転し、非常に強いRMSDとの相関を示す事からその有用性が示唆された。重心基準パラメータが、比較的良好な結果を示したCAD・VVより短い時間で算出可能である事は原理上明らかであり、以上よりパラメータの再計算及び再アライメントを視野に含む重心基準パラメータを組み合わせたの正確且つ迅速な立体構造アライメントツールの可能性は大いに示された。

Table 1 CATH各層内二タンパク質対のアライメントにてマッチした残基対に於ける相関係数の平均値・標準偏差、及びCATH score・RMSD (R±SD) (n:母数)

Criteria	Distance	Angle CB	Angle C	Angle N	CAD	VV	PHI	PSI	ACC	RMSD
Class (n=120)	0.237±0.280	0.110±0.186	0.245±0.161	0.251±0.158	0.305±0.251	0.218±0.155	0.064±0.147	0.151±0.157	0.154±0.135	12.5±3.59
Architecture (n=120)	0.246±0.342	0.173±0.207	0.243±0.215	0.241±0.205	0.284±0.290	0.205±0.195	0.062±0.194	0.173±0.167	0.151±0.156	11.3±3.69
Topology (n=136)	0.380±0.393	0.341±0.223	0.427±0.211	0.414±0.211	0.411±0.296	0.344±0.232	0.096±0.205	0.193±0.206	0.262±0.145	6.65±3.81
Homolog. S.F. (n=126)	0.651±0.327	0.567±0.232	0.625±0.255	0.623±0.240	0.688±0.212	0.575±0.237	0.227±0.236	0.375±0.231	0.414±0.200	5.33±5.05
Seq. F. (n=135)	0.974±0.037	0.926±0.046	0.953±0.032	0.952±0.035	0.970±0.034	0.932±0.040	0.685±0.222	0.691±0.093	0.772±0.069	1.24±0.55
R of R-RMSD (n=637)	-0.560	-0.729	-0.729	-0.732	-0.608	-0.692	-0.520	-0.595	-0.729	

アライメントに基き
重心を再設定

↓

新重心を基準として
パラメータを再計算

Table 2 再計算パラメータに於ける相関係数 (R±SD) (n:母数)

Criteria	Distance	Angle CB	Angle C	Angle N
Class (n=120)	0.417±0.243	0.173±0.181	0.322±0.168	0.324±0.174
Arch. (n=120)	0.451±0.284	0.225±0.226	0.332±0.212	0.334±0.212
Topo. (n=136)	0.724±0.218	0.482±0.223	0.605±0.189	0.587±0.186
Homo. (n=126)	0.831±0.201	0.643±0.198	0.711±0.198	0.704±0.197
Seq. (n=135)	0.988±0.015	0.931±0.045	0.959±0.030	0.958±0.031
R of R-RMSD (n=637)	-0.843	-0.814	-0.859	-0.857

【参考文献】

- Bernstein, F.C., et al., & Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535-542.
- Murzin, A., Brenner, S.E., Hubbard, T.J.P. and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536-540.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) *Structure*, **5**, 1093-1108.
- Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G. (1992) *Protein Sci.*, **1**, 1691-1698.
- Taylor, W.R. and Orengo, C.A. (1989) *J. Mol. Biol.*, **208**, 1-22.
- Holm, L. and Sander, C. (1993) *J. Mol. Biol.*, **233**, 123-138.
- Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577-2637.
- Conte, L.L. and Smith, T. (1997) *J. Mol. Biol.*, **273**, 338-348.
- Higgins, D. and Taylor, W. (2000) In *Bioinformatics* (Higgins, D. and Taylor, W. eds), pp. 15-50, Oxford University Press, Oxford, United Kingdom