Vol.2010-BIO-22 No.7
2010/7/29

タンパク質構造予測関数と ホモロジーモデリングシステムの構築

荒井まみ† 加納和彦†† 寺師玄記††
梅山秀明†† 岩舘満雄†††

タンパク質モデリングのアライメント情報から，モデリングの正確性を調べるための構造予測関数を考案した．また，その予測関数を組み込んだホモロジーモデリングシステムを使用し，タンパク質モデリングの国際コンテスト CASP のサーバー部門に参加している．

Construction of the Prediction Function of Protein Structure and the Homology Modeling System

Mami Arai†, Kazuhiko Kanou††, Genki Terashi††,
Hideaki Umeyama†† and Mitsuo Iwadate†††

We hatched out the prediction function of protein structure calculating by alignment imformation in protein modeling to see about modeling accuracy. And We attend CASP which international competition of protein modeling using the homology modeling system with the prediction function.

1. SUMMARY

タンパク質モデリングの国際コンテストである CASP(The Critical Assessment of protein Structure Prediction)の過去の経験を踏まえ, 構造予測関数によるスコアの scoreA とホモロジーモデリングシステムの構築を行った．モデリング前の情報であるアミノ酸配列のアライメントからターゲットにどの位近いモデルができるかを調べるため，構造類似性を予測するスコア scoreA を既存の PF_score を元に考案した．またホモロジーモデリングを行うシステム FAMS を基本とし, その scoreA や主成分分析を組み込んだホモロジーモデリングシステムを用いて５月から７月末まで開催されるモデリングの国際コンテストの CASP9 に参加している．

2. INTRODUCTION

タンパク質モデリングの国際コンテストである CASP の過去の経験を踏まえ, 構造予測関数によるスコアの scoreA とホモロジーモデリングシステムの構築を行った. モデリングにはホモロジーモデリングソフト FAMS[1]を使用した.

データベースとアライメントによりモデリングの元となるテンプレートタンパク質を探し出すホモロジーモデリングでは，テンプレートの精度がモデリングの結果に大きな影響を及ぼす．そのため，モデリングの元となるアライメントの正確性を見極めることが必要とされている．今研究ではそういったモデリング前の情報から，モデリングのターゲットにどの程度近いモデルができるかを調べるために構造類似性を予測するスコア scoreA を考案した. scoreA はホモロジーモデリングにおけるアライメント情報とそのテンプレート構造からモデリングの正確性を予測するスコアであり，モデリング正確性を予測する既存のスコア PF score[2]を元に最適化を行った.

scoreA の計算には以下の３つのパラメーターが使用される．length はアライメントから得たテンプレートタンパク質のアミノ酸配列の長さである．align_score はアミノ酸置換行列 BLOSUM62 とアフィンギャップペナルティを用いて計算したアライメントの相同性を表すスコアである．ss_score はモデリングのターゲット配列から PSIPRED[3]で予測した二次構造配列とテンプレートから STRIDE[4]で識別した二次構造配列を比較し，テンプレートの二次構造の類似性を表すスコアである．この scoreA をモデリングシステムに組み込むことで正確性を向上させている．

† 中央大学大学院理工学研究科物理学専攻
Physics Course, Graduate School of Science and Engineering, Chuo University
†† 北里大学薬学部生物分子設計学教室
Department of Biomolecular Design, School of Pharmaceutical Sciences, Kitasato University
††† 中央大学理工学部生命科学科
Department of Biological Sciences, Faculty of Science and Engineering, Chuo University

ホモロジーモデリングシステム全体は次の手順で行われる．モデル構造の物理化学的な評価には CIRCLE[5]を使用した．(i) モデリングのターゲット配列を複数のアライメントツール(PSI-BLAST[6], HHsearch[7], SPARKS2[8], SP3[9], HMMER[10], HHM_BLAST）にかけて複数のホモロジー検索結果とそのアライメント候補を得る．(ii）各候補のアライメントを元にＦＡＭＳモデリングを行い，複数のモデルを得る．（iii）モデル群に対してアライメントとモデルから得られる情報を要素とした主成分分析を行い，結果をプロットする．(iv）プロットされたモデルを階層クラスタリングにかけてクラスターを作る．(v) クラスター内の平均 CIRCLE 値によって代表クラスターを選択する．(vi) CIRCLE 値と scoreA の合成スコアを使って代表クラスターから代表モデルを選択する．(vii）モデルがターゲットに対して末端から１００残基以上短い場合，代表モデルとそれを補うモデルとのドッキングを行い，それを新たな代表モデルとする．(viii）代表モデルをフルモデリングにかけ，最終的なモデルを得る．

構築したホモロジーモデリングシステムを使用して，2010 年 5 月 3 日から開催された CASP9 へ参加した．

3. METHOD

3.1 PDB, CASP の学習セットとホモロジーモデリングシステムのデータベース

scoreA 最適化のラーニングデータセットとして，２００８年５月２日の PDB(Protein Data Bank)に登録されている 110556 個のタンパク質のアミノ酸配列データベースを使用する. PDB サイトでは blastclust により互いにアミノ酸同一性が９５％以上であるタンパク質を集めるクラスタリングが行われており，18512 個のクラスターが得られた．PDB からダウンロードできる PDB のクラスターデータは quality factor 順にソートされている．quality factor は以下の式によって計算される．

$$quality\_factor = \frac{1}{\text{結晶回折の分解能}} - R\text{値}$$

$$R\text{値} = \frac{\text{回折実験の測定値} - \text{構造からの計算値}}{\text{回折実験の測定値}}$$

ラーニングデータセットには，各クラスターから最も quality factor が高いタンパク質を選び出した．

また CASP の過去の経験を学習するためのラーニングデータセットとして，前大会である CASP8 の 127 配列のターゲットと 165 ドメインの正解構造を使用した. ターゲットよりも正解構造が多いのは，ターゲットでは配列がドメインごとに分けられていないためである．その各ターゲットに対し，アライメントをモデリングシステムと同様の 6 種類のアライメントツールによって行い，アライメント群を作成した．また，その全てのアライメントを FAMS によりモデリングし，モデル群を作成した．これらの各ターゲットに対するアライメント群, モデル群を学習セットとして使用している．また，CASP8 ではコンテスト後，ターゲットを正解構造ドメインごとに４種類の難易度カテゴリー分け（**TBM-HA** (Template Based Modeling - High Accuracy), **TBM** (Template Based Modeling), **TBM/FM** (overlap between TBM & FM categories), **FM** (Free Modeling)）が発表されている．

CASP9 の参加には PDB サイトから最新のクラスタリングされたデータを得ることができなかったため, PDB に登録された配列データにローカルで blastclust を実行しクラスタリングデータを得た. また, クラスターごとに PDB データの分解能（resolution）がクラスター内の最高値を取り，かつその中でもアミノ酸配列長が 1 番長いタンパク質を１つずつ選び出しホモロジーサーチのデータベースとした. CASP9 の参加中はこのデータベースを毎週更新してアライメントツールなどに使用した．

3.2 モデルの評価

モデルの評価には GDT_TS(Global Distance Test Total Score)を使用する．GDT_TS はｘÅ 以内にあるＣα原子ペアが最大になるようにフィッティングする．ｘ＝１もしくは２，４，８Å すべての割合を算出し，その平均をとったものが GDT_TS となる．

$$GDT\_TS = \frac{GDT\_P1 + GDT\_P2 + GDT\_P4 + GDT\_P8}{4} \times 100$$

$$GDT\_Px = \frac{x\text{Å以内にある残基ペア数}}{\text{全アミノ酸残基数}}$$

また最適化の際に使用する MAX GDT_TS と rate GDT_TS を定める．MAX GDT_TS は１つのターゲットに対するアライメント群の GDT_TS の中で１番高い GDT_TS である．また 1 つのモデルの GDT_TS の MAX GDT_TS に対する割合（%）を rateGDT_TS とする．例えばターゲット i から構築したモデル群中で一番 GDT_TS が高いモデルの GDT_TS を MAX GDT_TS(i)とし，モデル群のあるモデル j の GDT_TS を GDT_TS(i,j) とすると，

$$rateGDT\_TS(i,j) = \frac{GDT\_TS(i,j)}{MAX\_GDT\_TS(i)} \times 100$$

と表わされる．この値は難易度が異なる，つまり MAX GDT_TS が異なるターゲット同士のモデルの質を同等に扱うために使用している．特に最適化などで全ターゲットの合計を考える際は, 簡単なターゲットほど topGDT_TS が高くなり重みが増してしまうため rateGDT_TS 合計を使用している．

また, CASP ではターゲットに対する各チームの提出モデルの比較に Z_score が使用されている. Z_score は以下の式のように各成分の平均 $\bar{x}$ からのずれを標本標準偏差 v で割ることで求められる．

Vol.2010-BIO-22 No.7
2010/7/29

$$Z\_score = \frac{x_i - \bar{x}}{v} \qquad (i = 1,2,\cdots,n)$$

CASP の場合 x はモデルの GDT_TS，n はモデル数である．

3.3 モデル評価システム CIRCLE の使用

Verify3D[14]を元に考案された, 側鎖パッキングの経験的ポテンシャルをベースとするモデル評価プログラム CIRCLE をモデル評価スコアとしてモデリングシステムなどに用いる．モデルクオリティーは各残基の側鎖環境（極性環境，埋まっている領域，２次構造）と，ターゲット配列からの予測２次構造とモデルの２次構造の類似性により計算されている．

この CIRCLE では CM（Comparative Modeling）と FRorNF（FoldRecognition or New Fold）の２種類にターゲットの難易度を識別し, それに対応した計算が行われている．その判別には SVM（Support Vector Machine）が使用されている．CIRCLE の論文で使用されている学習セット CASP6 は現在の CASP とは識別の方法が違っているため，CIRCLE を計算する際の CM と FRorNF の判断基準は新しく調べる必要がある．そこで２種類の CIRCLE 値（CIRCLE-CM と CIRCLE-FR）の振舞いを調べるために CASP8 に対して得られた各 CIRCLE 値の topGDT_TS を比較した（図 1）．

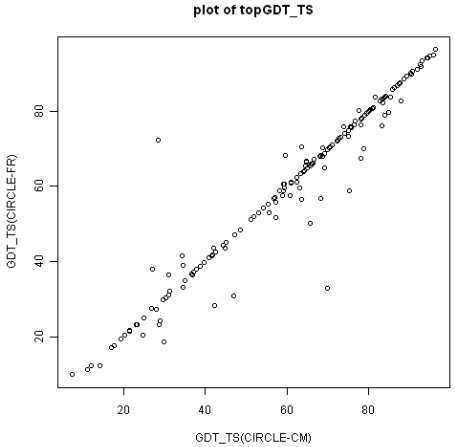


図 1　topGDT_TS(CIRCLE-FR)と topGDT_TS(CIRCLE-CM)のプロット

CIRCLE-CM と CIRCLE-FR はほとんどの場合で同じくらいの topGDT_TS，もしくは CIRCLE-CM の方が高い値となっている．CIRCLE-FR の topGDT_TS の方が 0.5Å以上高いターゲットは 165 ターゲット中 23 ターゲット，１Å以上は全 165 ドメイン中 18 ターゲットだった. また, その 23 ターゲットの topGDT_TS（CIRCLE-CM）と topGDT_TS（CIRCLE-FR）はそれぞれ 7.37 から 81.56, 9.85 から 83.69 の広い範囲に渡っており，CASP8 でのカテゴリーでも 23 ターゲット中 21 ターゲットが TBM と TBM-HA だった．

以上から，CM と FR の最適な難易度判別は基準を設定することは難しいと考えられる. よってモデリングシステムでは CM であるべきターゲットに誤って CIRCLE-FR を使用しないことを優先し，主成分分析を除く全ての CIRCLE 計算で CIRCLE-CM を用いることとした．以下，本論文での CIRCLE 値は CIRCLE-CM のことである．

3.4 PF_score

PF_score（Power Function score）はタンパク質３次元構造モデルの正確性を予測するアミノ酸配列アライメントだけから行うスコアである．PF_score は length（モデルの長さ），homology（アミノ酸同一性（%）），ss_homology（二次構造同一性（%））によって構成され，以下の式で求められる．

$$PF\_score = k_i \times (\mathrm{hom}\,o\log y)^m \times (ss\_\mathrm{hom}\,o\log y)^n$$

$$\mathrm{hom}\,o\log y = \frac{\text{一致したアミノ酸残基ペアの数}}{\text{残基ペアの総数}}$$

$$ss\_\mathrm{hom}\,o\log y = \frac{\text{一致した二次構造ペアの数}}{\text{二次構造ペアの総数}}$$

係数 *k*i と乗数 *m,n* はアライメントツール（PSI-BLAST, BLAST, RPS-BLAST, IMPALA, FASTA, Pfam-BLAST）と homology の大きさにより最適化されている．scoreA は PF_score の式を元に予測関数を考案している．

3.5 scoreA の最適化

scoreA はホモロジーモデリングに1おけるアライメント情報からモデリング正確性を予測するスコアである．モデリング正確性を予測する既存のスコア PF score を元に考案した．scoreA の計算にはモデルのアミノ酸配列長の length，アライメントの質を評価するために用いられるアライメントスコアの align_score(alignment score)，二次構造類似性スコアの ss_score(Secondary Structure score)の３つのパラメーターが用いられる．length はアライメントから得られるモデルのアミノ酸配列の長さである．align_score はアミノ酸置換行列 BLOSUM62 とギャップペナルティを用いて計算したアライメントの相同性を表すスコアである．ギャップペナルティはリニアギャップペナルティを-10，アフィンギャップペナルティを-1 とした．ss_score はテンプレートの二次構造が，ターゲット配列から予測される二次構造にどのくらい近いかを示す二次

構造類似性スコアである．つまり選ばれたテンプレートのふさわしさを二次構造の視点から評価するスコアと言える．ターゲット配列からの二次構造予測には二次構造予測ソフト PSIPRED が使用された. テンプレートの二次構造は STRIDE という二次構造認識ソフトから得た．この２つのソフトからはどちらも二次構造配列が得られるが，二次構造の表記に差異が存在する（表 1）.

・PSIPRED	・STRIDE
H = AlphaHelix	H = Alpha helix
E = Strand	E = Extended conformation
C = Coil	C = Coil (none of the above)
	G = 3-10 helix
	I = PI-helix
	B or b = Isolated bridge
	T = Turn

表 1　PSIPRED と STRIDE の二次構造表記

PSIPRED は３種類，STRIDE は７種類の二次構造を表記することができる．その表記内容には重複が見られるが，STRIDE にしかない表記（ G, I, B, C ）が PSIPRED でのどれに対応するのか，PSIPRED の H が本当に STRIDE の H と対応しているか，などが分からないために正確な比較は難しい. そこで上記のラーニング用の PDB データセットを使用し，PSIPRED と STRIDE の構造表記の比較を行った.

まずデータセットの全タンパク質に対して，アミノ酸配列に PSIPRED を，構造に STRIDE をそれぞれかけて，１つのタンパク質につき PSIPRED と STRIDE による２種類の二次構造配列を作成した．この２つの配列は全く同じターゲットに対するものであるため, これらを比較することで PSIPRED と STRIDE の関係を調べることができる. このような二次構造表記の比較をデータセットの全タンパク質で行った（表 2）.

PSI＼STR	H	B	E	G	I	T	C
H	1160919	3442	27293	64761	365	99898	64416
E	16407	9592	656081	7116	29	55821	85508
C	142723	31447	190978	68741	221	679678	571550

表 2　PSIPRED と STRIDE の二次構造対応表

この表から STRIDE の T は PSIPRED では C と予測されることが多いが，約１２％が H とも予測されている．更に，STRIDE の G は PSIPRED では H と C がどちらも同じくらいの割合で予測されていることが分かる. これらを STRIDE の B は PSIPRED の C といった様に１つに定めるのは難しい．そこで，この様な関係を表す値として次の式を用いてペア必然性を計算した．これは，例えば STRIDE の T を PSIPRED が C と予測することを，PSIPRED の C と STRIDE の T がペアを組むと考えて，偶然ペアを組む確率で実際にペアを組んでいる確率を割り，２つがペアを組む必然性を調べたものである．そして必然性を式により（表 2）を下記の式によって変換した（表 3）.

$$\frac{P(C:T)}{Pp(C)\times Ps(T)}$$

P(C:T) : 全ペア中の C（PSIPRED）と T（STRIDE）のペアの出現確率
Pp(C)　: PSIPRED 配列中の C の出現確率
Ps(T)　: STRIDE 配列中の T の出現確率

PSI＼STR	H	B	E	G	I	T	C
H	2.436	0.214	0.086	1.276	1.644	0.331	0.247
E	0.059	1.022	3.557	0.240	0.224	0.317	0.562
C	0.253	1.652	0.510	1.142	0.839	1.901	1.851

表 3　PSIPRED と STRIDE のペア必然性

このペア必然性表を BLOSUM のようなスコア行列として使用し，二次構造配列の類似性を計算したものを ss_score とした. 片方にギャップがあるペアは計算していない.

以上の３つのパラメーターlength，align_score，ss_score を使用していくつかの式の候補を試した結果，scoreA の式を以下のように定めた．length の係数を a，ss_score の係数を b とする.

$$\mathrm{scoreA} = a \times \mathrm{length} + \mathrm{align\_score} + b \times \mathrm{ss\_score}$$

係数 a, b の最適化は 2008 年 4 月 25 日のアミノ酸同一性 30%でクラスタリングされた PDB データセットで，各クラスターの quality factor が一番高いタンパク質 6498 個に対して以下の手順で行った．まずＰＤＢデータセットの各アミノ酸配列をクエリーとして同じくＰＤＢデータセットを検索対象とした６種類のアライメントツール（PSI-BLAST, HHsearch, SPARKS2, SP3, HMMER, HHM_BLAST）を実行し，アライメントを得る．ホモロジーが低い，いわゆる難易度が高いアライメントに焦点を合わせて最適化を行うため，アライメントの homology が 50%という閾値を設けて，それ以下の homology を持つアライメントのみを選び出し，541611 個のアライメントとなった．次に，ホモロジーがそれぞれ 50%，40%，30%，20%，10%以下のアライメント全てに FAMS の Cα原子構造のみのモデリングを行いモデルを構築した．Cα のみのモデリングは大量のモデリングに要する計算時間の短縮のためである．その次に，構築

された全てのモデルと解析構造（PDB に登録されている構造）との GDT_TS を計算し，各ターゲットの MAX GDT_TS を探し出す．最後に，式の係数 a と b の最適化をそれぞれ-20 から 20 まで 0.1 刻みの全組み合わせで行う．各係数 a,b の組み合わせに対し，scoreA 最大のモデルの rateGDT_TS の全ターゲットでの合計値をそれぞれ計算し，最大化することで係数のセットを定めた（表 4）．

homology閾値	a	b	平均rateGDT_TS	rateGDT_TS合計	ターゲット数
50	−0.5	1.7	85.94	554924.1	6457
40	−0.5	1.7	84.17	543296.2	6455
30	−0.5	1.7	80.33	518380.0	6453
20	−0.5	1.2	72.28	465452.0	6440
10	−1.6	2.1	47.24	287540.4	6087

表 4　各ホモロジー閾値での係数最適化

（図 2）は横軸が a，縦軸が b の rateGDT_TS 合計の等高線プロットである．

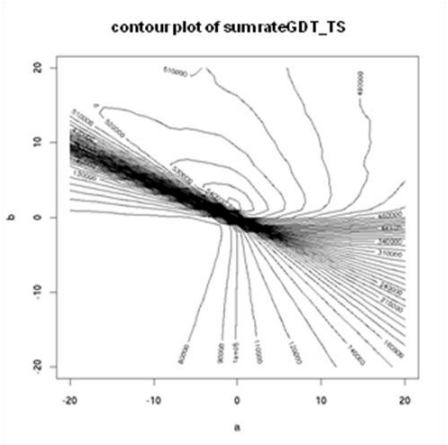


図 2　rateGDT_TS 合計の等高線プロット(ホモロジー閾値 50%)

length の係数が-0.5 という負の値をとったのは, align_score と ss_score の値がともにモデル配列長に大きく依存しているためである．length の係数である a がマイナスを取り，その余分な依存性を打ち消す働きをしていると考えられる．また，ss_score の係数が 1.7 と大きいのは，align_score に比べて値のオーダーが平均的に小さいためである．

また使用した PDB データベースの確かさを調べるために scoeA 係数最適化の交差検定を行った．最適化に使用したデータを１０等分し，１つをテストセット，残りの９つで係数最適化を１０セット全てで行いその振舞いを調べた（表 5）．

	a	b	rateGDT_TS平均（テスト）	rateGDT_TS合計（テスト）	rateGDT_TS合計（全体）
all	−0.5	1.7	85.94	554924.1	554924.1
set_1	−0.4	1.6	86.13	55556.4	554913.6
set_2	−0.4	1.6	86.17	55577.1	554913.6
set_3	−0.5	1.7	86.35	55694.9	554924.1
set_4	−0.6	1.8	86.18	55588.2	554914.5
set_5	−0.5	1.7	84.28	54360.1	554924.1
set_6	−0.6	1.8	87.00	56115.1	554914.5
set_7	−0.4	1.6	87.33	56329.2	554913.6
set_8	−0.3	1.5	84.31	54382.1	554739.8
set_9	−0.6	1.8	84.34	54398.2	554914.5
set_10	−0.4	1.6	85.68	55262.8	554913.6

表 5　scoreA 係数最適化の交差検定

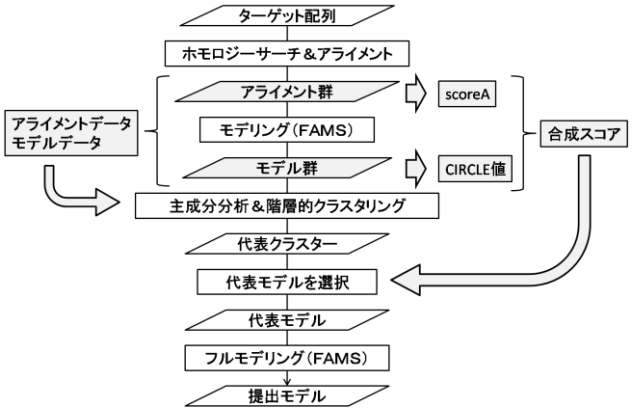
交差検定で得られた scoreA の係数の組み合わせは４種類となったが，係数 a と b に大きなばらつきは見られず安定している．またその４種類の soreA の全体での rateGDT_TS 合計を調べたところ，最適化の中で rateGDT_TS 合計が最も高い上から４つの係数の組み合わせであった．以上から，データセットは最適化に十分な大きさと種類を持っていると言える．

3.6　ホモロジーモデリングシステムの構築

scoreA と主成分分析を組み込んだモデリングシステムの構築を行った．モデリングシステムは以下の手順で動く（フローチャート 1）．

（i）ターゲットを学習セットと同様にアライメントにかけて複数のホモロジー検索結果とそのアライメント候補を得る．(ii）各候補のアライメントを元にＦＡＭＳモデリングを行い，複数のモデルを得る．（iii）モデル群に対してアライメントとモデルから得られるパラメーターを要素とした主成分分析を行い，結果をプロットする．（iv）プロットされたモデルを階層クラスタリングにかけてクラスターを作る．（v）各クラスター内の CIRCLE 値平均の高さによって代表クラスターを選択する．（vi）CIRCLE 値と scoreA の合成スコアから代表クラスターから代表モデルを選択．（vii）モデルがターゲットに対して末端から１００残基以上短い場合，代表モデルとそれを補うモデルとのドッキングを行い，それを新たな代表モデルとする．（viii）代表モデルをフルモデリングにかけ，最終的なモデルを得る．

Vol.2010-BIO-22 No.7
2010/7/29



フローチャート 1　モデリングシステムのフローチャート

このシステムの特徴は，作成されたモデル群からモデルを選択する手段として，主成分分析によるプロットと階層的クラスタリングを使用していることである．主成分分析は多くの変数によるデータの変数間の相関を排除して，情報の損失をできるだけ抑えた少数個の合成変数に集約する手法である．この手法によりモデルのデータをプロットすることで，特徴別にモデルを識別できると考えた．モデルのデータとして scoreA とその計算で用いたデータ（length, align_score, ss_score, scoreA）と２つの CIRCLE 値（CIRCLE-CM, CIRCLE-FR）に加え，PF_score とそのデータ（homology, ss_homology, PF_score）を加えた．また，階層的クラスタリングでは１つのクラスター内のファクター数（モデルの数）を CASP8 の学習セットに対して Z_score 合計の最大化による最適化を行い 12 と定めた. クラスタリングの手法の１つである非階層的クラスタリングは初期中心座標をランダムに定めることから結果にもばらつきが生じるため，最適化に不向きと判断して階層的クラスタリングを使用している．

主成分分析の利点は合成変数により相関の少ない，つまりランダム性が高いデータをノイズとして排除できることにある．主成分分析には寄与率という値があり，これは元のデータをどの程度保っているかを示している．そのため，この寄与率の合計である累積寄与率によって元のデータをどのくらい使用するかをある程度指定できる．そこで捨てるノイズの範囲を定めるため, 累積寄与率の閾値を 0.05 から 1.00 まで 0.05 きざみで設定して各閾値での主成分分析を使用したモデリングシステムを実行し，Z_score 合計を計算した. その結果 Z_score 合計の１番高い閾値は 0.85 であることが分かった．もしノイズを捨てるという手法が意味をなさない場合は閾値 1.00 が１番よい結果となる筈であるが，閾値 1.00 は１番低い結果となった．モデリングシステムの主成分分析は元のデータの約 15%をノイズとして切り捨てて使用することに効果があり，データにはノイズが若干含まれているということになる．以上からクラスタリングシステムでは累積寄与率の閾値を 0.85 と定めて使用する．

次の特徴として, 代表モデルを選択する際に CIRCLE 値と scoreA の合成スコアを使用している. この時の scoreA の計算には, 合成スコア最適化の複雑さを回避するため，ホモロジー閾値が 50%（40%，30%）で最適化された係数に固定して使用している．この scoreA 係数選択の方法については現在検討中である．CIRCLE 値は疎水性残基のパッキングといったモデル構造の振舞いの自然さを調べるスコアであり，ターゲットへの類似性は考慮されなていない．よってモデルの選択には CIRCLE 値に加えてアミノ酸配列, 二次構造の両方の類似性を比較している scoreA も考慮すべきと考えられる．そこで２つのスコアの合成スコアを使用することにした．合成スコアは以下の式の scoreA の weight を CASP8 のデータで最適化することで定める．

$$\mathrm{CIRCLE} + \mathrm{weight} \times \mathrm{scoreA}$$

weight を 0.01 から 1.00 まで 0.01 刻みで変化させ, それぞれの Z_score 合計を最大化することで最適化を行った（図 3）.

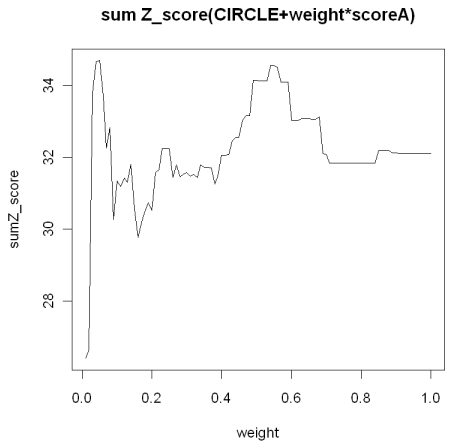


図 3　各 weigth の合成スコアでの Z_score 合計のグラフ

この最適化では weigth=0.05 と weigth=0.54 の２ヵ所でほぼ同じ値のピークがある．これは scoreA のオーダーが CIRCLE 値に比べ１桁ほど高いことからそれぞれ，weight

が 0.05 は CIRCLE 値主体で scoreA を若干考慮した値，0.54 が scoreA 主体で CIRCLE 値を若干考慮した値であると考えられる．また 0.05 の方がわずかに最高値が高く，これは CIRCLE 値の方が高い topGDT_TS 合計を持つためである．この合成スコアはよいモデルを選ぶ能力の高い CIRCLE 値を主体とするスコアにしたいため，weight を 0.05 として合成スコアを導く関数を定めた．

$$\mathrm{CIRCLE} + 0.05 \times \mathrm{scoreA}$$

この合成スコアの目的は CIRCLE 値のモデル選択の精度を保ちながら，類似性を表す scoreA の要素を加えることで，総合的に精度を引き上げることである．

4. RESULT and DISCUSSION

4.1 scoreA と合成スコアについて

scoreA の精度を調べるために，学習セットである CASP8 の全ターゲットに対するアライメント群とモデル群の scoreA や CIRCLE 値, 合成スコアを計算して実際のモデルの GDT_TS との比較を行った．まず topGDT_TS 平均（表 6），rateGDT_TS 平均（表 7），GDT_TS（表 8）との相関係数の平均を CASP8 のカテゴリーごとに求めた．

まず scoreA に関して CIRCLE 値と比較する. scoreA は良いモデルを選ぶという点では CIRCLE 値より劣るが，一方で GDT_TS との高い相関を示しており，難易度の高いターゲットでは CIRCLE 値にわずかに勝る結果を出しているということが分かった．

次に合成スコアについて，（表 7）の rateGDT_TS 平均では TBM が下がってしまっているが全体的には CIRCLE 値よりも精度を上げている. この結果からは scoreA の要素を追加することで FM/TBM や FM の難易度の高いターゲットのモデリングの精度が高まることが分かった．今大会である CASP9 では PSI-BLAST では良い E 値のアライメントが得られない難しいターゲットが多いため，さらに scoreA の適用方法について検討を重ねてゆきこの合成関数がより役に立つのではないかと考えられる．

category	MAX_GDT_TS	CIRCLE	scoreA	合成スコア
TBM-HA	87.820	83.532	81.871	84.308
TBM	63.135	55.834	53.763	55.270
TBM/FM	37.127	25.190	27.820	26.957
FM	33.790	22.838	22.960	24.069
ALL	68.396	61.706	59.980	61.702

表 6　カテゴリー別の MAX_GDT_TS と topGDT_TS 平均

category	CIRCLE	scoreA	合成スコア
TBM-HA	95.088	93.181	96.002
TBM	87.423	83.577	86.578
TBM/FM	68.504	74.724	72.359
FM	67.734	68.571	72.057
ALL	88.213	85.428	88.306

表 7　カテゴリー別の rateGDT_TS 平均

category	CIRCLE	scoreA	合成スコア
TBM	0.664	0.771	0.725
TBM-HA	0.758	0.841	0.810
TBM/FM	0.866	0.953	0.924
FM	0.383	0.648	0.515
ALL	0.711	0.829	0.775

表 8　カテゴリー別の GDT_TS との相関係数平均

4.2 主成分分析と階層的クラスタリングについて

主成分分析では実際にプロットされたものが GDT_TS と相関があるのかを調べるため，CASP8 のターゲットである T0388 を主成分分析による座標で GDT_TS をプロットした（図 4）.

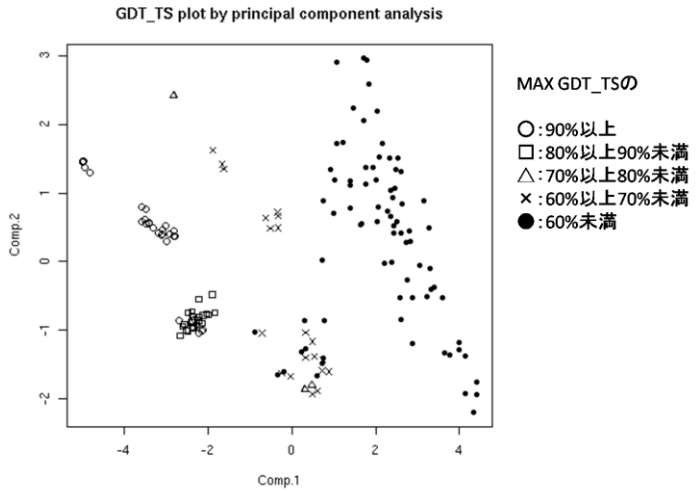


図 4　主成分分析による GDT_TS のプロット(T0388)

x 軸が第 1 主成分，y 軸が第 2 主成分によるプロットである．第 1，2 主成分の累積寄与率はそれぞれ 0.698, 0.856 である．（図 4）では MAX GDT_TS に対する GDT_TS の大きさ（%）で○□△×●のマークでプロットを行った．するとプロットが明確にマークごとに分かれてクラスターを形成することが分かる．また，このプロットと GDT_TS との相関係数は Comp.1（第 1 主成分）が-0.957 で Comp.2 が-0.248 である．特に Comp.1 はとても大きな負の相関があり Comp.1 の値が低いほど GDT_TS が高い傾向にあると分かる．しかし Comp.1 のみでは高い相関を持つとしても GDT_TS が高いモデルと低いモデルのプロットは混ざってしまっている．その混ざっているプロットを分けているのが Comp.2 である．

他ターゲットでも同様に調べて全体の平均を求めた（表 9）．相関係数は負の相関となることもあるが，相関があるという意味に変わりはないため平均を計算する時は絶対値をとっている．モデリングシステムでは累積寄与率の閾値によって使用する主成分の数を決めている．そのため（表 9）では使用する主成分の数（num=2~4）ごとに分けて，第 x 主成分（Comp.x）ごとに GDT_TS との相関係数平均を求めている（x=2~4）．

num	Comp.1	Comp.2	Comp.3	Comp.4
2	0.921785	0.233226		
3	0.758314	0.248535	0.153167	
4	0.648774	0.500832	0.246305	0.029328
ALL	0.797356	0.248913	0.155237	0.029328

表 9　使用する主成分の数ごとの相関係数平均

例えば num=2 では Comp.1 と Comp.2 が使用され，Comp.3 と Comp.4 は使用されていないので空欄となっている．表を見ると，num=2 の時は Comp.1 の相関係数はとても高い．（図 4）も num=2 のターゲットであり，平均が約 0.922 という相関係数はこれだけでも GDT_TS をよく予測していると言える．これは第 1 と第 2 主成分でデータの 85%をカバーしているため，1 つ 1 つが大きな割合をカバーできているためである．その証拠に num=3 では約 0.758，num=4 では約 0.649 とプロットに使用する主成分の数が増えるほど Comp.1 の相関は下がってきている．反対に Comp.2 以下は num が増えるごとに増大する傾向にある．num の値がそのままプロットする次元の数に相当することから，複雑で元のデータを上手く反映できない時に次元を上げることで反映度を高めているととれる．同時に，判別の難しいモデルを次元を上げることでクラスタリングしやすくなっていると考えられる．

以上のことからモデルのアライメントと構造についてのデータを用いた主成分分析による特徴づけとプロットは GDT_TS と高い相関を持ち，クラスタリングの元データとして有用であることが分かった．

参考文献

1 Ogata K, and Umeyama H: An automatic homology modeling method consisting of database searches and simulated annealing. J Mol Graph Model. 2000 Jun;18(3):258-72, 305-6.
2 Iwadate Mitsuo, Kanou Kazuhiko, Terashi Genki, Umeyama Hideaki, and Takeda-Shitaka Mayuko: Method for Predicting Homology Modeling Accuracy from Amino Acid Sequence Alignment: the Power Function. CHEMICAL & PHARMACEUTICAL BULLETIN 58(1), 1-10 (2010).
3 Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292:195-202 (1999).
4 Heinig, M., and Frishman, D.: STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. Nucl. Acids Res. , 32, W500-2 (2004).
5 Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., and Umeyama H.: Fams-ace: A combined method to select the best model after remodeling all server models. Proteins, 69, Suppl 8, 98-107 (2007).
6 Altschul, SF, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. J Mol Biol 215(3):403-10 (1990).
7 Söding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics, 21, 951–960 (2005).
8 Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins (2004) 55:1005–1013.
9 Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins (2005) 58:321–328.
10 R. Durbin, S. Eddy, A. Krogh, and G. Mitchison: *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press (1998).