

統計的手法を用いた逆平行および平行 β sheet中でのstrandペア形成予測法の開発

誌名	明治大学農学部研究報告 = Bulletin of the Faculty of Agriculture, Meiji University
ISSN	04656083
著者名	鈴木,博実 竹本,周二
発行元	明治大学農学部
巻/号	56巻4号
巻号補足	
掲載ページ	p. 209-227
発行年月	2007年3月

農林水産省 農林水産技術会議事務局筑波産学連携支援センター

Tsukuba Business-Academia Cooperation Support Center, Agriculture, Forestry and Fisheries Research Council
Secretariat



〔研究論文〕

統計的手法を用いた逆平行および平行 β sheet 中での strand ペア形成予測法の開発

鈴木博実・竹本周二

（2007年2月2日受理）

Development of Prediction Method for β Strand Pairing in Antiparallel and Parallel Sheets using Statistical Theory

Hiromi Suzuki and Shuji Takemoto

Abstract

Development of a protein tertiary structure predicting method based on amino acid sequences is one of the most attractive research topics from the viewpoint of fulfilling probing intellectual curiosity as well as from a practical use side. The prediction accuracy of secondary structures has already reached 75%, and the packing prediction method of these secondary structures has been further developed. In this packing process, β sheet formation is one of the most important steps. In this study, the statistical scores were computed from the frequency of the inter-strand amino acid residue pairs based on experimental protein structure data, taking the existence of hydrogen bond formation into consideration. These scores were used for predicting the inter-strand register of interacting β strands from 8 alternative offset positions either side of the experimentally observed β sheet register and prediction accuracies were 38% and 35% of the antiparallel and parallel pairs, respectively. When the score obtained from the secondary structure prediction method was added to these statistic scores, accuracies went up to 40% and 38% for antiparallel and parallel strand pairs, respectively. Prediction accuracies for the strand triplets by moving the central strand became 50% and 52% when adding the scores obtained from secondary structure prediction. In any case, the percentage predicted as offset positions of 1 and 2 residues from the exact position was 5% or more, and the accuracies regarding the secondary structure prediction to all residues in β strand regions reached 68% and 80% for the strand pairs and for the strand triplets, respectively, exceeding the 58% accuracy from the secondary structure predicting method. Thus, the method developed in this study would be effective not only for prediction of the β strand register but also for β strand region prediction, and would make a valuable contribution to protein tertiary structure prediction.

Key words: inter-strand residue pairs, hydrogen bond, secondary structure prediction, global environment, strand triplet

明治大学農学部農芸化学科生物物理学研究室

〒214-8571 神奈川県川崎市多摩区東三田 1-1-1

E-mail: hsuzuki@isc.meiji.ac.jp

Tel & Fax: 044-934-7104

Laboratory of Biophysics, School of Agriculture, Meiji University

1-1-1 Higashimita, Tama-ku, Kawasaki, Kanagawa, 214-8571, Japan

要 旨 アミノ酸配列に基づくタンパク質の立体構造予測法の開発は、実用面からも知的探究心を満たすという面からも大いに注目されている研究課題の1つである。既に2次構造領域の予測精度は75%に達しており、この2次構造のパッキング予測方法の開発が進められている。このパッキング過程における重要なステップの1つとして β strandの会合による β sheet形成がある。本研究では、既知のタンパク質立体構造データに基づいて、水素結合形成の有無を考慮した β strand間での残基ペアの出現頻度より統計的スコアを算出した。このスコアを用いてstrandペアに対して正確な位置から前後8残基の範囲で移動させながらstrand配置関係予測を行ったところ、その検出精度は逆平行ペアに対しては38%、平行ペアに対しては35%であった。統計的に算出したスコアに2次構造予測から得られたスコアを加えた場合には、それぞれ40%および38%に上昇した。Strandトリプレットに対して、左右のstrandの位置を固定して中央のstrandを正確な位置から前後8残基の範囲で移動させた場合の検出精度は50%、2次構造予測から得られたスコアを加えた場合には52%となった。正確な位置から前後1および2残基のずれとして予測された割合はいずれの場合も5%以上存在し、 β strand領域の全残基に対する2次構造予測としてみた精度はstrandペアでは68%、strandトリプレットでは80%に達し、用いた2次構造予測法による精度58%を大幅に上回った。このように今回開発した方法は、 β strandの会合予測だけでなく、 β strand領域予測に対しても有効であり、タンパク質の立体構造予測への貢献が十分期待できる。

キーワード : strand間残基ペア, 水素結合, 2次構造予測, グローバル環境, strandトリプレット

緒 言

近年の様々なゲノムプロジェクト、プロテオームプロジェクト等や測定技術の開発・向上により多くのタンパク質の立体構造が決定され、タンパク質立体構造データベースPDB（Bermanら 2000）等に登録された立体構造のエントリー数は急増してきている。しかしながら、まだ配列決定されたタンパク質に比べ立体構造が決定されたタンパク質数ははるかに少ない。このような状況に加え、アミノ酸配列と立体構造との関連性の解明という知的探究心を満たすという点からも、立体構造未知のタンパク質に対してアミノ酸配列に基づいて立体構造を予測する方法の開発が続けられてきている。

立体構造予測法の1つとして、大域的エネルギー関数の最小化というアプローチが挙げられる（Leeら 1999；Liwoら 1999）。この方法は、アミノ酸残基が占めることができる構造空間全体を探索し、最小エネルギーとなる構造を見つけようとするものであり、長さが60～100残基程度の小さなタンパク質断片に対しては成功している（Liwoら 1999）。しかしこの方法ではより大きなタンパク質に対しては膨大な計算時間が必要となってしまう。

2つのタンパク質が高いアミノ酸配列類似性を示す場合には、ほとんどの場合その立体構造も類似している。このため立体構造未知のタンパク質が、構造既知のタンパク質と高いアミノ酸配列類似性を示す場合には、この類似タンパク質の立体構造を鋳型として立体構造を比較的容易に予測することができる。隠れマルコフモデル（Karplusら 1998）や配列プロファイル検索（Altschulら 1997；Schafferら 1999）を用いることにより、アミノ酸配列の類似度が20%程度まで減少してもほぼこの方法が可能である。

構造未知のタンパク質の配列が、他のタンパク質との間に検出可能な配列類似性を持たない場合の方法の1つとしてthreading法がある（Jones 1999；Kellyら 2000；Shiら 2001）。この方法は、タンパク質の立体構造には疎水性環境中に2次構造が互いに近接して詰め込まれたコア領域が存在することを利用して、問い合わせタンパク質のアミノ酸配列と既知のタンパク質立体構造の構造コアとの適合性を調べる。配列を立体構造データベース中のタンパク質コアと比較し、適合するコアを探していく。その結果ある構造コアとの間に十分な適合性が見つかれば、そのタンパク質はそのコアと類似した立体構造に折りたたまれると予測される。

この threading 法において、適合したコアが見つからない場合には、2 次構造領域をまず予測して、それがどのように空間的にパッキングされるかを予測することが必要となる。2 次構造の予測精度については既に 75% に達しており (Frishman と Argos 1997)、この 2 次構造のパッキング予測法の開発が重要になってきている。2 次構造のパッキング過程の 1 つとして、複数の β strand が平行 strand か逆平行 strand に配向して β sheet を形成する過程が挙げられる。これまでの研究により、隣接した β strand 間では、ランダムに全てのアミノ酸ペアが現れるのではなく、その出現傾向に偏りがあることが明らかになっている (Mandel-Gutfreund ら 2001 ; Fooks ら 2006)。この出現傾向に基づいてペアとなる領域を検索する方法が開発されている (Zhu と Braun 1999 ; Steward と Thornton 2002) が、その精度はまだ十分なものではない。

本研究では、アミノ酸配列を用いながら、ローカルなアミノ酸配列のみだけではなく、strand 間で水素結合を形成する残基ペアと形成しない残基ペアを区別したアミノ酸残基ペアの出現傾向を用いることにより、1 つの β strand がその領域とペアを形成する領域を予測する方法の開発を行った。またこの際に、2 次構造予測の結果から算出したスコアを併用することによる予測精度向上についても検証した。さらに、今回開発したペア形成領域予測法と従来の 2 次構造予測法における β strand 領域に対する予測精度の比較を行った。

方 法

HB 残基と nHB 残基

本論文で開発した方法は、一方の strand 領域を他方の strand に対して不正確なアライメントの位置よりスライドさせることにより、正確な位置関係を予測させるものである。この方法の本質の理解のためにそれぞれ 3 本の β strand からなる逆平行 sheet および平行 sheet の様相を図 1 に示した。

逆平行 sheet の場合、2 本の strand 間での水素結合は 1 残基おきに形成される。ここで 2 本の strand 間

の水素結合に関わっているアミノ酸残基を水素結合 (以下 HB とする) 残基、関わっていない残基を非水素結合 (以下 nHB とする) 残基と呼ぶことにする。図 1(a) の strand ペア 1-2 に着目した場合、残基 i 、および j は HB 残基、残基 $i+1$ および $j-1$ は nHB 残基である。また strand 1 から見た残基ペア (i, j) および strand 2 から見た残基ペア (j, i) は共に HB 残基ペアであるが、残基ペア ($i+1, j-1$) あるいは ($j-1, i+1$) は nHB 残基ペアとなる。nHB 残基ペアのおのおの残基は反対側に別の strand が存在する場合にこの strand 上の残基と水素結合を形成し、この strand ペアにおいて HB 残基ペアとなることができる。図 1(a) の strand ペア 2-3 において残基 j は nHB 残基、残基 $j-1$ は HB 残基であり、残基ペア (j, k) および (k, j) は nHB 残基ペア、($j-1, k+1$) および ($k+1, j-1$) は HB 残基ペアである。

平行 sheet の場合、strand 1 上の残基 i は strand 2 上の残基 $j-1$ および $j+1$ と水素結合を形成しており (図 1(b))、strand ペア 1-2 に着目した場合、残基 i は HB 残基となる。これに対して、この残基 i と向かい合う strand 2 上の残基 j は strand 3 上の残基 $k-1$ および $k+1$ と水素結合を形成しており、strand ペア 1-2 に着目した場合には、nHB 残基となる。ゆえに残基ペア (i, j) は HB-nHB 残基ペアであり、(j, i) は nHB-HB 残基ペアとなる。同様に strand ペア 2-3 に関して、(j, k) は HB-nHB 残基ペア、(k, j) は nHB-HB 残基ペアとなる。一方 ($i, j+1$) や ($j-1, i$) は HB 残基ペアであり、($j, i+1$) や ($i-1, j$) は nHB 残基ペアとなる。

スコアの計算

strand ペアを形成するためには、1) その領域がどの程度 β strand を形成しやすいか、2) 逆平行、あるいは平行のどちらに配向しやすいか、3) ペアとして存在しやすいか、という点が重要な意味を持つ。この観点から、1 組の strand ペアの形成しやすさを表すスコアを strand 領域の自己スコア、残基間スコアおよび 2 次構造スコアから構成した。2 本の strand 上のそれぞれ i 番目と j 番目に位置して残基ペアを形成

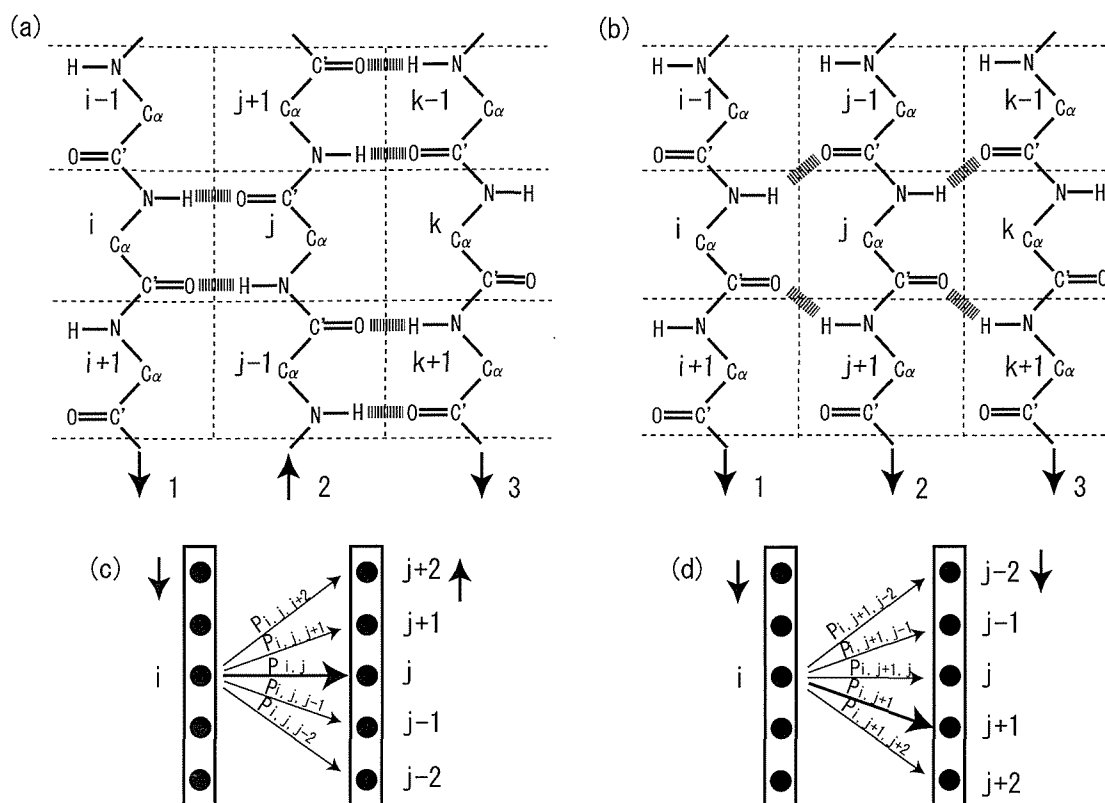


Fig. 1. Schematic view of β -triplets and residue pairs. Antiparallel (a) and parallel (b) β -triplets are shown with hydrogen-bonding pattern. Inter-residue scores, $P_{i,j,j+m}$ and $P_{i,j+1,j+m}$, are derived with $(i \sim j)$, $(i \sim j \pm 1)$, and $(i \sim j \pm 2)$ residue pairs for antiparallel (c) and parallel (d) strand pairs.

する残基 R_i , R_j に関する自己スコア (S_i , および S_j) は、それらの残基が逆平行と平行を区別した β strand 内に存在しやすいかどうかを表し、また残基間スコア ($P_{i,j,j+m}$ および $P_{i,j,i+m}$) は隣接する strand 間において、出現頻度の高い残基ペアかどうかを示す。さらに 2 次構造スコアは 1 本の strand 上において、そのアミノ酸配列に基づいて個々のアミノ酸残基がどの程度 β strand として出現するかを示している。

(i) 自己スコア

逆平行 strand 中における自己スコアは、式(1)によって表すことができる。

$$S_{A,i} = \log(f_{A,R_i}/f_A) - \log(f_{R_i}/f_{all}) \quad (1)$$

ここで R_i は逆平行 strand の i 番目に位置する残基の種類、 f_{A,R_i} は逆平行 strand 中の R_i 残基の合計数、 f_A は逆平行 strand 中の全残基数、 f_{R_i} はタンパク質中の R_i 残基の合計数、 f_{all} はタンパク質中の全残基数を表している。つまりこの式はタンパク質に存在する全残基中での R_i 残基の存在比に対する逆平行 strand に存

在する全残基中での R_i 残基の存在比の比率を示しており、逆平行 strand 中での存在比が全タンパク質中での存在比より高ければ $S_{A,i} > 0$ 、逆に低ければ $S_{A,i} < 0$ となる。同様に平行 strand における自己スコアも計算する。なお、1 本の strand が 1 つの strand と逆平行に、別の strand と平行に配向している場合、該当する残基数の 1/2 をそれぞれ逆平行および平行 strand 領域としてカウントした。

(ii) 残基間スコア

β strand が 2 本以上平行にあるいは逆平行に配向して β sheet を形成する場合、1 本の strand 上のアミノ酸残基は交互に sheet の上面あるいは下面方向に飛び出す。このため、1 本 strand 上の位置 i に存在する残基 R_i が他の strand 上の位置 j に存在する残基 R_j とペアを形成するかどうかについて、 $j \pm 2$ の位置に存在する残基の影響が十分に予想される。実際 Steward と Thornton (2002) の結果により、残基 R_i に着目した場合、位置 j だけではなく、 $j-2, j-1, j+1, j+2$ も

情報を持つ、つまりこの位置に存在する残基に明確な偏りが存在する、ことが示された。そこで本方法においては、 R_i と R_j のペアについてのスコアに加えて、 R_i に対する $j-2, j-1, j+1, j+2$ のそれぞれの位置における各アミノ酸残基の存在傾向もスコア化した。この場合、HB 残基と nHB 残基を区別してそれぞれスコアを求めた。

まず逆平行 strand ペア上の残基ペア (R_i, R_j) に対しては次の式によりスコア化する。

$$P_{i,j} = \log(f_{AX, R_i \setminus R_j} / f_{AX, R_i}) - \log(f_{AX, R_j} / f_{AX}) \quad (2)$$

ここで、 $f_{AX, R_i \setminus R_j}$ は観察される HB あるいは nHB 残基ペア (R_i, R_j) 数、 f_{AX, R_i} および f_{AX, R_j} は HB あるいは nHB 残基中での残基 R_i および R_j の合計数、 f_{AX} は HB あるいは nHB 残基の合計数を表している。つまりこの式は逆平行 strand の全 HB あるいは nHB 残基中での R_j 残基の存在比に対する逆平行 strand の HB あるいは nHB R_i 残基中で R_j 残基とペアを形成している割合の比率を示している。 R_i 残基のペアの相手としての R_j 残基の比率が逆平行 strand の HB あるいは nHB 残基の比率より高ければ、つまりペア形成の相手として選択される傾向が高ければ $P_{i,j} > 0$ となり、逆であれば $P_{i,j} < 0$ となる。 $f_{AX, R_i \setminus R_j}$ が 0 の場合にそのまま計算するとスコアは $-\infty$ となってしまうので、この場合には全てのスコア中での最小値よりやや小さい値 (-2.50) を割り当てた。

平行 strand においては R_i 残基の水素結合の相手は R_{j-1} 残基と R_{j+1} 残基となるが、本方法では R_{j+1} 残基との組み合わせ、つまり (R_i, R_{j+1}) をカウントして $P_{i,j+1}$ を求める (R_i 残基と R_{j-1} 残基の関係については (R_{j-1}, R_i) としてカウントする)。よって残基ペア (R_i, R_j) が与えられた場合には $P_{i,j}$ や $P_{j,i}$ の代わりに $P_{i,j+1}$ および $P_{j,i+1}$ を計算する。

次に逆平行 strand ペアにおいて R_i とペアとなっている R_j より、 m ($m = -2, -1, 1, 2$) だけずれた位置にあるアミノ酸残基に対するスコア (図 1(c)) は次の式により求める。

$$P_{i,j,j+m} = \log(f_{AX, R_i \setminus R_j \setminus R_{j+m}} / f_{AX, R_i \setminus R_j}) - \log(f_{R_{j+m}} / f_{all}) \quad (3)$$

ここで $f_{AX, R_i \setminus R_j \setminus R_{j+m}}$ は R_i と R_j がペアの位置にある際に、 R_j から m 残基離れた位置にある残基 R_{j+m} の合計数、 $f_{R_{j+m}}$ は $j+m$ の位置に存在する残基 R_{j+m} の全タンパク質中の合計数を示している。なお、位置 $j+m$ は必ずしも β strand 中である必要はないとした。 $P_{i,j,m}$ の全ての可能な組み合わせは、全部で $20 \times 20 \times 20 = 8,000$ 通り存在するが、全ての場合に対して十分な数のサンプルが得られないので、今回の解析においては、 R_j をアミノ酸の性質に基づいて i) Cys, ii) 小さいアミノ酸 (Gly, Pro, Ser, Thr, Ala), iii) 負電荷を有する側鎖かその類似側鎖を持つアミノ酸 (Gln, Glu, Asn, Asp), iv) 正電荷を有するアミノ酸 (Arg, Lys, His), v) 疎水性アミノ酸 (Leu, Met, Ile, Val), vi) 芳香族アミノ酸 (Tyr, Phe, Trp) の 6 グループ (Dayhoff 1978) にまとめ、 $20 \times 6 \times 20 = 2,400$ 通りに分類した。なお、 $f_{AX, R_i \setminus R_j \setminus R_{j+m}}$ が 0 の場合には $f_{AX, R_i \setminus R_j}$ の場合と同様に -2.50 を割り当てた。

平行 strand ペアの場合には、 $P_{i,j}$ や $P_{j,i}$ の代わりに $P_{i,j+1}$ および $P_{j,i+1}$ を計算したのと同様に、 $P_{i,j,j+m}$ の代わりに $P_{i,j+1,j+m}$ を計算する。ただし、位置 j を基準にした場合、 $m = -2, -1, 0, 2$ となる (図 1(d))。

(iii) 2 次構造予測スコア

今回開発した方法は、立体構造既知のタンパク質とアミノ酸類似性をほとんど示さないタンパク質に適応することを目指したものである。このため、立体構造に関する情報を直接利用する方法を避けている。タンパク質 2 次構造予測プログラム PREDATOR (Frishman と Argos 1995, 1996) は立体構造既知のタンパク質に対して、そのアミノ酸配列と 2 次構造の関係を統計的に処理し、得られた傾向に基づいて与えられたアミノ酸配列に対する 2 次構造を予測する。この方法による予測結果は、各アミノ酸に対して 3 種類の 2 次構造 (α ヘリックス, β strand, ランダムコイル) のうちのどの構造になるかだけでなく、その期待値も同時に与える (最大値は 1 であり、もし 3 つの 2 次構造の間に優位性の差が見られない場合には 0 となる)。この PREDATOR により β strand 領域を含むタンパク質の全領域を用いて 2 次構造予測を行い、得られた予測結果を用いて実際の β strand

に該当する領域の前後10残基を含めた範囲の各残基に対して、

予測された2次構造	期待値	2次構造スコア(SS)
β strand	>0	期待値
β strand	$=0$	0.4
β strand 以外	>0	$(1 - \text{期待値})/2$
β strand 以外	$=0$	0.3

というスコアを割り当てた。

以上のスコアを用いると、ある strand 上の i 番目の残基と別の strand 上の j 番目の残基がペアとなる場合のスコア S は逆平行 strand ペアの場合は

$$\begin{aligned}
 S = & 10 \times (S_j + SS_j) + P_{i,j} \\
 & + \sum_{m=-2}^{-1} P_{i,j,j+m} + \sum_{m=1}^2 P_{i,j,j+m} + P_{j,i} \\
 & + \sum_{m=-2}^{-1} P_{j,i,i+m} + \sum_{m=1}^2 P_{j,i,i+m} \quad (4)
 \end{aligned}$$

となり、平行 strand ペアの場合には

$$\begin{aligned}
 S = & 10 \times (S_j + SS_j) + P_{i,j+1} + P_{i,j+1,2} \\
 & + \sum_{m=-2}^0 P_{i,j+1,j+m} + P_{j,i+1} + P_{j,i+1,2} \\
 & + \sum_{m=-2}^0 P_{j,i+1,i+m} \quad (5)
 \end{aligned}$$

となる。1組の残基ペアに対し、10個の残基間スコアの計算を行っているが自己スコアおよび2次構造スコアについては1回の計算しか行っていないので、そのまま加えてしまうと残基間スコアの方が全体のスコアに対して10倍の影響を与えてしまう。そこで、本方法では自己スコアおよび2次構造スコアを10倍して残基間スコアに加えた。この残基ペアに対するスコアを strand の全領域について加えた結果を配列の相対位置でのスコアとした。一方の strand（残基 i が存在する strand）を固定し、他方の strand（残基 j が存在する strand）の位置を1残基ずつスライドさせながら、おのおのの相対位置におけるスコアを計算し、最高値となる位置を予測位置とした。

立体構造データ

使用するタンパク質の立体構造データは Hobohm

ら（1992）によって作成された配列的に類似性を持たないタンパク質のリストの2001年版に基づいて選択し、PDB データは PDBj (<http://www.pdbj.org>) より入手した。全ての配列は40残基以上の長さであり、どの配列ペアも25%以上の類似性を示さない。なお α 炭素の座標しか含まない構造データは除外した。

β strand 領域の決定

タンパク質の2次構造の決定は Kabsh と Sander（1983）による DSSP 法を用いた。この際、採択した最小の strand 長を逆平行のペアの場合は4残基、平行のペアの場合は3残基とした。また、一方の strand の連続した2残基が他方の strand の1残基とペアを形成することにより生じる β バルジ構造を持つペアについては、解析から除外した。このようにして得られた逆平行、平行 strand 数はそれぞれ5736本および2192本であり、それぞれの strand 領域の残基数は37744残基および11523残基であった。また片側が逆平行ペアに、反対側が平行ペアに配向している strand 数は480本であり、その領域の残基数は2130残基であった。これらの strand 領域に対して、逆平行 strand ペアか平行 strand ペアかという区分に、HB 残基か nHB 残基かの区分を加えて残基ペア数を集計した。3本以上の β strand からなる β sheet においては、sheet の両端に位置する strand 以外の strand は、それぞれ2個の strand ペアを形成する。そのためこれらの strand 上の残基の多くを、1回ずつ HB 残基ペアおよび nHB 残基ペアの片方としてカウントした。

予測精度の検証

この方法の精度は、入手した既存の立体構造データを用いてジャックナイフ法で検証した。ジャックナイフ法とは、与えられたデータをほぼ同等の n 個のグループに分割し、 $n-1$ 個のグループのデータを予測方法の調整用に、残り1つのグループのデータを予測用に使用するルーチンを、予測用グループを変えながら n 回繰り返す方法である。本研究では、抜き出

した strand ペアを20グループに配分し、これに対応するように PDB データを分割した。

結 果

自己スコア

図2に各アミノ酸残基の逆平行および平行 strand に対する自己スコアの分布の様子を示した。疎水性アミノ酸や芳香族アミノ酸のスコアが高く、電荷を帯びたアミノ酸や Pro のスコアが低いなど、これまでに得られているアミノ酸の β strand 領域での出現傾向と同様の傾向を示した (Chou と Fasman 1974(a); 1974(b); Levitt 1978; Jiang ら 1998)。また逆平行 strand の自己スコアと平行 strand の自己スコアに差があまり見られないアミノ酸残基も多いが, Cys, Gln, Glu, Arg, Lys, Ile, Trp ではその差が0.4以上となっているが, これも Steward と Thornton (2002) の結果と同様の傾向を示した。

残基間スコア

逆平行 strand ペアについては残基ペア (i, j) に対して, 平行 strand ペアについては残基ペア (i, j+1) に対して, それぞれ HB 残基ペアか nHB ペアかを区別して残基間スコアを計算し, その結果を表1に示した。この表から明らかなように, HB 残基ペアと

nHB 残基ペアのスコアに明確な差が見られる。典型的な例としては, Pro は HB 残基ペアとしてはまったく存在しなかったが, nHB 残基ペアとしては存在した。またスコアが1以上となった組み合わせは, HB 残基ペアでは逆平行 strand ペアの Glu と Lys の場合だけであり, 平行 strand ペアの場合では Arg→Cys の0.83が最高であった。これに対して, nHB 残基ペアにおいては平行 strand ペアの Pro→Pro の1.7をはじめとして10数個の組み合わせが存在した。このことは, HB 残基ペアよりも nHB 残基ペアにおいて, 特定の残基がペアの位置に存在する可能性が高い, つまりアミノ酸残基の選択性が高いことを示している。言い換えるならば, HB 残基ペアにおいては, nHB 残基ペアの場合よりも多くの種類の残基と残基ペアを形成することを示している。一方スコアが-1以下となる組み合わせは, HB 残基ペアにおいても nHB 残基ペアにおいて多数存在した。このことは, HB 残基ペアにおいても nHB 残基ペアにおいても, ペアの相手として相応しくない残基の種類が多く存在し, その数はペアの相手として相応しい残基の種類よりも多い, つまり正の選択性よりも負の選択性が強いことを示している。

逆平行および平行 strand ペアにおける残基ペア (i, j), および (i, j+1) 以外の残基ペアについての残基間スコアを, R_i が Ala で R_j が疎水性残基の場合の例を表2に示した。この場合も, 逆平行 strand ペアと平行 strand ペア間, および HB ペアと nHB ペア間に差が見られ, さらに m の位置による違いも見られる。またここに表示した条件では, 全ての状態の全ての m の位置において正の値を示した Val をはじめ, Ile, Leu などの疎水性残基や Tyr など芳香族残基などが多く正の値を示しているのに対し, Asn, Asp, Glu, Gln などでは多く負の値を示している。

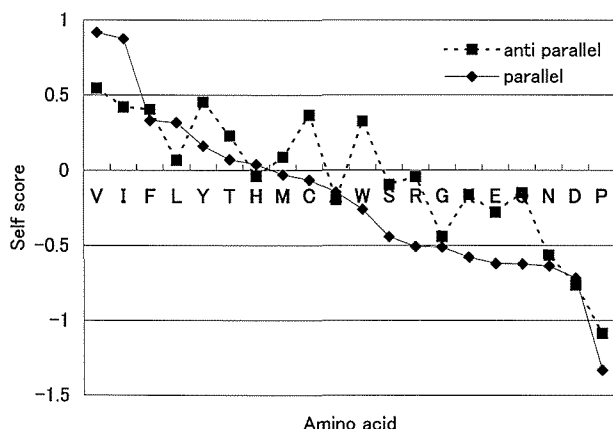


Fig. 2. Self-score values for each type of amino acid calculated for antiparallel and parallel strands. Positive self-scores indicate a residue favored in β -strands whereas a negative scores indicate a residue disfavored in β -strands. Closed square (■) and diamond (◆) indicate self scores for antiparallel and parallel strands, respectively.

Strand ペアにおける正確な位置関係の予測

strand ペアによって形成される正確な位置関係を予測できるかをチェックするために, 一方の strand を固定し, 他方の位置を正確な位置から前後8残基の範囲で移動させながら, 最高の得点を与える位置を

Table 1 Inter-residue scores for (i, j) and (i, j + 1) pairs.

i\j	antiparallel HB pairs																			
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	-0.53	-0.12	-0.54	0.08	0.41	-0.34	-0.27	0.06	-0.46	0.12	0.12	-0.32	0.05	0.00	-2.50	-0.21	-0.06	0.30	-0.08	0.34
Arg	-0.12	-0.66	0.11	0.38	-0.06	0.58	0.95	-0.48	0.03	-0.21	-0.31	-0.27	-0.24	-0.20	-2.50	0.04	0.39	-0.44	0.08	-0.13
Asn	-0.54	0.11	-0.14	0.66	-0.66	0.53	0.53	0.08	0.22	-0.41	-0.78	-0.08	-0.15	-0.37	-2.50	0.64	0.64	0.06	-0.11	-0.36
Asp	0.05	0.43	0.66	0.12	-1.14	0.30	0.20	0.52	0.33	-0.81	-0.58	0.68	-0.31	-1.18	-2.50	0.36	0.47	-0.33	-0.41	-0.37
Cys	0.41	-0.06	-0.66	-1.14	0.49	-0.42	-0.43	0.01	0.07	-0.30	0.03	0.07	0.39	0.14	-2.50	0.33	-0.60	0.14	0.07	0.00
Gln	-0.34	0.58	0.53	0.30	-0.42	-0.22	-0.23	-0.19	-0.37	-0.33	-0.36	0.61	0.32	-0.32	-2.50	0.30	0.21	0.01	-0.15	-0.02
Glu	-0.25	0.95	0.56	0.20	-0.43	-0.23	-1.11	-0.49	-0.18	-0.11	-0.36	1.05	-0.46	-0.64	-2.50	0.10	0.39	-0.14	-0.52	-0.25
Gly	0.06	-0.48	0.08	0.52	0.01	-0.19	-0.49	-0.39	0.21	-0.09	-0.14	-1.07	-0.44	0.37	-2.50	-0.24	0.13	0.33	0.28	0.23
His	-0.46	0.03	0.22	0.33	0.07	-0.37	-0.18	0.21	-0.05	-0.19	-0.35	0.03	-0.02	0.03	-2.50	0.31	0.08	0.10	-0.06	0.20
Ile	0.12	-0.21	-0.41	-0.81	-0.30	-0.33	-0.11	-0.09	-0.19	-0.33	0.44	-0.24	0.23	0.15	-2.50	-0.33	-0.23	-0.03	0.32	0.13
Leu	0.12	-0.31	-0.78	-0.58	0.03	-0.36	-0.36	-0.14	-0.35	0.44	-0.28	-0.16	0.21	0.26	-2.50	-0.29	-0.21	0.02	0.02	0.26
Lys	-0.32	-0.27	-0.08	0.68	0.07	0.61	1.05	-1.07	0.03	-0.24	-0.16	-0.35	0.21	-0.75	-2.50	0.45	0.32	-0.07	-0.12	-0.29
Met	0.05	-0.24	-0.15	-0.31	0.39	0.32	-0.46	-0.44	-0.02	0.23	0.21	0.21	-0.34	-0.10	-2.50	-0.03	-0.02	0.20	-0.24	-0.02
Phe	0.00	-0.20	-0.37	-1.18	0.14	-0.32	-0.64	0.37	0.03	0.15	0.26	-0.75	-0.10	-0.31	-2.50	-0.26	-0.46	0.23	0.24	0.31
Pro	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50
Ser	-0.21	0.03	0.64	0.36	0.33	0.28	0.10	-0.24	0.36	-0.33	-0.29	0.45	-0.03	-0.26	-2.50	-0.11	0.42	0.27	-0.19	-0.21
Thr	-0.06	0.39	0.64	0.47	-0.65	0.21	0.39	0.13	0.08	-0.23	-0.21	0.32	-0.02	-0.46	-2.50	0.42	-0.42	-0.72	-0.07	-0.22
Trp	0.30	-0.44	0.06	-0.33	0.14	0.01	-0.14	0.33	0.10	-0.03	0.02	-0.07	0.20	0.23	-2.50	0.27	-0.72	-1.56	-0.11	-0.05
Tyr	-0.08	0.08	-0.11	-0.41	0.07	-0.15	-0.52	0.28	-0.06	0.31	0.02	-0.12	-0.24	0.24	-2.50	-0.20	-0.07	-0.11	-0.64	0.13
Val	0.34	-0.13	-0.36	-0.37	0.00	-0.02	-0.25	0.23	0.20	0.13	0.25	-0.29	-0.05	0.31	-2.50	-0.20	-0.22	-0.05	0.13	-0.55

i\j	antiparallel nHB pairs																			
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	-0.53	-0.53	-0.31	-0.24	-0.47	-0.67	-0.61	-0.01	-0.15	0.43	0.43	-0.53	-0.14	0.27	-0.30	-0.21	-0.03	-0.14	0.03	0.29
Arg	-0.53	-0.82	0.28	0.84	-0.71	0.16	0.85	-0.45	-0.26	-0.40	-0.20	-0.40	-0.58	-0.28	-0.16	-0.07	0.32	0.11	0.37	0.03
Asn	-0.31	0.28	0.28	0.20	-0.85	0.20	0.17	0.24	0.55	-0.42	-0.21	0.26	-0.03	-0.19	-0.30	0.51	0.22	-0.13	0.05	-0.55
Asp	-0.21	0.81	0.20	-0.58	-1.03	0.56	-0.95	0.14	0.57	-0.56	-0.43	0.87	-0.57	-0.40	-0.14	0.48	0.12	-0.18	-0.04	-0.53
Cys	-0.47	-0.71	-0.85	-1.03	1.64	-0.12	-0.86	0.28	-0.31	-0.06	-0.02	-0.77	0.07	0.14	0.31	-0.31	-0.60	1.33	-0.22	0.07
Gln	-0.67	0.16	0.20	0.56	-0.12	-0.66	0.17	-0.33	0.24	-0.58	-0.14	0.18	-0.44	-0.02	0.16	0.43	0.71	-0.36	-0.01	-0.46
Glu	-0.64	0.85	0.15	-0.95	-0.86	0.17	-0.95	-0.77	0.08	-0.21	-0.21	1.00	-0.29	-0.16	-0.37	0.27	0.38	-0.64	-0.05	-0.33
Gly	-0.01	-0.45	0.24	0.14	0.28	-0.33	-0.77	-0.66	-0.37	0.24	0.29	-0.77	0.19	0.17	0.14	0.18	-0.12	0.04	0.11	0.04
His	-0.15	-0.26	0.55	0.57	-0.31	0.24	0.08	-0.37	-1.43	0.07	-0.27	0.16	-0.05	-0.16	0.00	0.18	0.45	-0.21	0.19	-0.50
Ile	0.43	-0.40	-0.42	-0.56	-0.06	-0.58	-0.21	0.24	0.07	-0.20	0.24	-0.41	0.39	0.01	-0.18	-0.32	-0.47	-0.31	0.00	0.46
Leu	0.43	-0.20	-0.21	-0.43	-0.02	-0.14	-0.21	0.29	-0.27	0.24	-0.16	-0.28	0.10	0.29	-0.10	-0.28	-0.63	0.18	-0.09	0.22
Lys	-0.53	-0.40	0.26	0.87	-0.77	0.18	1.00	-0.77	0.16	-0.41	-0.28	-1.11	-0.50	-0.45	-0.08	0.05	0.42	0.32	0.42	-0.28
Met	-0.14	-0.58	-0.03	-0.57	0.07	-0.44	-0.29	0.19	-0.05	0.39	0.10	-0.50	0.54	0.42	0.41	0.16	-0.32	-0.38	-0.15	-0.07
Phe	0.27	-0.28	-0.19	-0.40	0.14	-0.02	-0.16	0.17	-0.16	0.01	0.29	-0.45	0.42	-0.10	0.16	-0.23	-0.53	0.09	0.00	0.16
Pro	-0.30	-0.16	-0.30	-0.14	0.31	0.16	-0.37	0.14	0.00	-0.18	-0.10	-0.08	0.41	0.16	-0.68	-0.63	0.17	0.83	0.44	-0.05
Ser	-0.21	-0.06	0.51	0.48	-0.31	0.45	0.27	0.18	0.13	-0.32	-0.28	0.05	0.16	-0.23	-0.63	-0.23	0.53	-0.08	-0.15	-0.29
Thr	-0.03	0.32	0.22	0.12	-0.57	0.71	0.38	-0.12	0.45	-0.47	-0.63	0.42	-0.32	-0.53	0.17	0.53	0.09	-1.04	-0.41	-0.15
Trp	-0.14	0.11	-0.13	-0.18	1.33	-0.36	-0.64	0.04	-0.21	-0.31	0.18	0.32	-0.38	0.09	0.83	-0.08	-1.04	-0.06	0.28	-0.33
Tyr	0.03	0.37	0.05	-0.04	-0.22	-0.01	-0.05	0.11	0.19	0.01	-0.09	0.42	-0.15	0.00	0.44	-0.15	-0.41	0.28	-0.58	-0.05
Val	0.29	0.04	-0.55	-0.53	0.07	-0.46	-0.33	0.04	-0.50	0.46	0.23	-0.28	-0.05	0.15	-0.05	-0.30	-0.15	-0.33	-0.05	-0.05

予測位置とした。入手した PDB データに基づいて得られた 2 本の β strand ペアは 5040 組であり（逆平行 strand ペア 3467 組，平行 strand ペア 1573 組），ジャックナイフ法に用いた 1 つのグループ当たりのペア数は 252 組であった。2 本の strand が，平行か反平行に配向するかは決まっているとして，ジャックナイフ法によって予測した結果の平均と標準偏差を表 3 にまとめた（なお逆平行 strand ペア数および平行 strand ペア数はグループ毎に異なっているので，ペアの種類に対する予測結果は全組数に対する正しく予測した組数の割合として示した）。自己スコア，つまりその領域が β strand として逆平行，あるいは平行に配向しやすいかどうかの傾向のみに基づいた予測で

は，平均で 25% と低い精度となった。この原因としては，1) 個々の残基が示す strand 形成傾向性の和に基づいた β strand 領域の予測精度がそれほど高くない，2) 用いた領域に 2 つ以上の β strand 領域が存在し，実際の水素結合を形成する領域以外の部分がより高い β strand 傾向を持つ，ということが考えられる。1) については，自己スコアでは単にその領域にどのようなアミノ酸残基が何個存在するかを示しているだけであり，その配列パターンについてはまったく考慮されていないことが主な理由である。 β strand 形成傾向が高いアミノ酸残基としては疎水性残基や芳香族残基が挙げられるが，両親媒性の β sheet を形成している strand では 1 残基おきに親水性残基と疎水性残基

Table 1 (continue).

i \ j + 1	parallel HB pairs																			
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0.06	-0.12	-0.24	0.49	0.12	-0.30	0.18	0.32	-0.06	0.01	-0.07	-0.21	-0.26	-0.02	-2.50	-0.02	0.04	0.50	-0.08	-0.10
Arg	0.36	-1.41	-1.15	-0.50	0.83	-0.60	-2.16	-0.11	-0.80	0.21	0.17	-0.86	0.55	0.03	-2.50	0.14	-0.55	-1.19	-0.90	0.24
Asn	-0.15	-0.05	0.03	0.68	0.17	-1.43	-0.84	-0.05	-0.53	0.24	0.33	-0.37	0.49	-0.28	-2.50	0.08	0.01	-0.24	-0.15	-0.31
Asp	0.09	-0.79	0.46	-0.07	-0.62	-1.48	-0.38	0.53	-0.59	-0.14	0.46	-0.43	-0.03	0.30	-2.50	0.35	-0.58	-2.50	0.12	-0.22
Cys	-0.08	-0.38	0.73	0.34	0.74	-0.39	0.21	0.30	-0.18	-0.43	-0.01	-0.53	-0.56	0.36	-2.50	0.76	0.71	-0.61	-1.20	-0.48
Gln	-0.89	-0.60	-1.43	-0.09	-1.09	0.32	-0.70	0.29	-0.10	0.15	0.25	-0.24	0.16	-0.26	-2.50	0.34	-0.55	0.32	-0.14	0.22
Glu	0.30	-0.58	0.15	-0.60	-1.57	-0.70	-1.62	-0.05	-0.39	0.20	0.31	-1.66	-0.01	-0.01	-2.50	-0.29	-0.07	-0.19	-0.03	0.11
Gly	0.15	-0.57	0.59	0.34	0.41	-0.73	-0.05	0.40	-0.52	-0.03	0.11	-1.57	0.37	0.13	-2.50	0.36	-0.02	-0.92	-0.02	-0.28
His	-0.13	-0.80	-0.53	0.52	0.11	-0.41	-0.05	0.26	0.10	-0.05	0.21	-1.36	0.14	0.41	-2.50	0.02	-0.46	0.10	-0.21	0.02
Ile	-0.02	-0.05	0.47	0.36	0.02	-0.06	0.15	0.30	0.28	-0.14	-0.17	-0.31	0.11	-0.10	-2.50	0.24	0.00	-0.21	-0.02	-0.03
Leu	0.20	-0.26	0.30	0.72	0.32	0.07	0.21	0.38	-0.10	-0.12	-0.04	-0.90	-0.57	0.02	-2.50	0.36	0.09	-0.12	-0.13	-0.24
Lys	-0.21	-0.86	-0.38	-0.65	0.32	-0.05	-0.74	-0.47	-0.25	-0.12	0.43	-0.79	0.08	0.11	-2.50	-0.63	-0.18	0.28	0.07	0.30
Met	0.11	-0.07	0.02	-0.26	-0.20	-0.77	0.44	0.56	0.29	-0.35	-0.19	-0.40	-0.01	0.14	-2.50	-0.23	-0.30	-0.26	0.38	0.11
Phe	0.04	-0.26	0.23	0.11	0.18	0.23	0.25	0.10	0.30	-0.13	-0.02	-0.18	0.14	0.14	-2.50	0.46	0.08	-0.04	0.15	-0.36
Pro	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50	-2.50
Ser	-0.18	-0.47	0.57	0.52	-0.26	-0.77	-0.75	0.03	0.02	-0.12	0.19	-0.47	-0.42	0.22	-2.50	0.13	-0.42	0.60	-0.47	0.18
Thr	0.04	-0.34	-0.40	-0.05	0.59	-0.39	-0.20	0.28	-0.46	-0.04	0.10	-0.25	0.07	-0.05	-2.50	0.50	-0.21	0.48	-0.65	0.06
Trp	-0.04	-0.09	-0.24	0.11	0.53	-0.81	-2.50	0.70	0.40	0.17	-0.12	-0.65	-0.97	-0.04	-2.50	0.03	-0.03	-2.50	0.21	-0.01
Tyr	0.29	-0.19	-0.28	0.28	0.20	-0.14	-0.11	0.46	-0.50	-0.08	-0.13	-0.70	0.46	0.15	-2.50	0.11	-0.14	-0.49	0.10	-0.06
Val	0.05	-0.06	0.22	0.58	0.31	0.15	0.02	0.20	0.36	-0.14	-0.08	-0.19	0.16	-0.49	-2.50	0.24	0.12	-0.25	0.02	-0.11

i \ j + 1	parallel nHB pairs																			
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0.35	0.44	0.12	0.69	0.07	-0.18	0.12	0.38	0.30	-0.25	-0.16	-0.16	-0.12	-0.26	0.70	0.38	-0.07	-0.58	-0.14	-0.28
Arg	-0.13	-0.58	0.37	0.26	-0.09	0.26	0.27	0.54	0.15	-0.57	-0.09	-0.39	0.11	0.04	0.01	0.74	-0.30	0.45	-0.60	0.09
Asn	0.30	0.22	0.83	0.23	-0.69	0.54	-0.10	0.96	-0.14	-0.37	0.05	-1.66	-0.02	-0.20	0.52	0.14	-0.11	-0.87	-0.03	-0.37
Asp	-0.07	0.38	1.18	0.52	1.02	-1.50	-0.01	0.84	-0.16	-0.33	-0.59	0.16	-0.26	-0.26	0.97	0.53	0.25	-0.41	-0.56	-0.55
Cys	-0.19	0.47	0.24	0.68	0.33	1.05	1.03	0.96	-0.33	-0.38	-0.48	0.22	0.08	-0.21	-0.08	0.35	0.48	-2.50	-1.98	-0.61
Gln	-0.43	-0.67	0.13	0.99	0.46	0.76	-0.39	0.73	0.49	0.02	-0.12	-0.77	-1.20	0.20	1.02	-0.04	0.02	0.46	-0.79	-0.59
Glu	0.35	-0.20	0.86	0.53	0.00	-1.48	0.36	0.96	-0.15	-0.25	-0.23	-0.12	-0.54	-0.66	0.69	-0.09	-0.10	0.30	-0.55	-0.17
Gly	-0.12	0.33	0.64	0.71	0.04	-0.31	0.03	0.50	0.76	-0.32	-0.49	-0.55	-0.99	0.23	0.90	-0.15	0.09	0.44	0.00	-0.18
His	0.45	-0.53	-0.14	0.43	0.18	-1.31	0.19	0.95	0.03	-0.25	0.03	-1.05	0.15	-0.25	0.47	-0.03	-0.25	0.90	-0.55	-0.25
Ile	-0.15	-0.02	0.47	0.51	0.01	0.30	0.22	0.43	-0.07	-0.17	-0.27	-0.10	0.09	-0.06	0.18	0.37	0.08	0.47	0.01	-0.16
Leu	0.05	0.14	0.45	0.64	-0.25	0.27	0.22	0.44	0.19	-0.28	-0.31	0.12	0.08	-0.21	0.50	0.44	0.13	-0.87	0.10	-0.22
Lys	0.09	0.02	0.41	0.85	-1.58	0.34	0.17	-0.26	-0.65	-0.07	-0.16	-0.27	-1.60	0.07	0.31	-0.18	0.29	-0.36	-0.11	0.01
Met	0.19	-1.50	0.21	0.92	-0.21	0.61	-0.03	0.65	-0.07	-0.40	0.15	0.19	0.61	-0.35	-1.21	0.10	0.13	0.62	-0.25	-0.41
Phe	-0.16	0.04	0.68	0.55	0.37	-0.08	0.63	0.55	0.01	-0.18	-0.31	-0.14	-0.23	0.27	0.40	0.21	0.05	0.21	0.27	-0.53
Pro	-0.20	-0.28	1.22	1.12	-0.08	0.33	0.89	0.63	0.62	-0.51	-0.20	-0.10	-0.52	-0.23	1.70	0.09	-0.23	-0.68	-0.30	-1.20
Ser	0.16	0.05	0.15	0.84	0.12	-0.22	-0.35	0.50	-0.16	0.11	-0.04	0.31	0.22	-0.14	1.09	-0.24	-0.69	0.19	-0.12	-0.52
Thr	0.16	-0.04	0.14	0.65	0.30	0.02	-0.95	0.23	-0.25	-0.02	-0.08	-0.31	-0.13	-0.01	0.34	0.35	-0.44	0.43	-0.36	-0.04
Trp	-1.28	1.14	0.77	0.30	-0.06	-2.50	0.31	0.44	0.20	-0.59	-0.17	0.88	-0.80	-0.26	0.04	0.78	0.33	-2.50	0.31	-0.36
Tyr	0.18	0.50	0.29	-0.16	0.30	0.19	0.23	0.63	0.24	-0.25	-0.16	-0.52	-0.09	0.04	0.58	-0.12	-0.05	-1.47	0.09	-0.18
Val	0.00	0.01	0.17	0.40	-0.02	0.08	0.16	0.32	0.06	-0.15	0.04	-0.08	0.13	-0.09	-0.01	0.21	0.10	0.19	-0.15	-0.18

が並んでおり、個々の残基としては β strand 形成傾向が低くてもある配列パターン中では出現傾向が高い残基も存在する。しかしながら、自己スコアはあくまで平均的に見た個々の残基が示す傾向のみを反映したものであるから、予測精度が低くなった。さらに β strand の形成には、strand 領域のアミノ酸残基パターンのみならず、立体構造的な周囲の環境も影響するのであろう。2) については特に逆平行 strand ペアを形成する strand 領域の前後10残基の範囲に別の β strand 領域（それが実際に逆平行に配向する相手である場合も含む）を含むことが多く観察されるという点で裏づけられる。

残基間スコアのみの場合の予測精度は30%となり、自己スコアのみの場合と比較して5ポイントほど向上した。逆平行 strand ペアにおいては、自己スコアのみの場合と比較して10ポイント上昇したが、平行 strand ペアの場合には逆に5ポイント減少した。この差は β strand 領域の長さの分布の違いと大いに関連していると思われる。多くのタンパク質中では平行に配向するよりも、逆平行に配向しているペアの方が圧倒的に多く、今回採用したタンパク質中においても同様であった。そのため、逆平行 strand ペアの場合には strand 領域の長さが4残基以上のペアのみを採用したのに対し、平行 strand ペアの場合には3残基以上のペアを採用した。それぞれペア領域の長

Table 2 Inter-residue scores (positions i=Ala, j=hydrophobic residues).

antiparallel	HB pairs §				nHB pairs			
	j-2	j-1	j+1	j+2	j-2	j-1	j+1	j+2
Ala	0.01	0.01	-0.16	0.44	-0.22	0.01	-0.05	-0.01
Arg	-0.26	0.09	-0.06	-0.63	-0.92	-0.02	0.13	-0.59
Asn	-0.55	-0.61	-0.31	-0.31	-0.41	-0.26	-0.19	-0.06
Asp	-0.55	-0.80	-0.44	-0.14	-0.47	-0.25	-0.31	-0.34
Cys	0.14	-0.11	0.06	-0.43	0.48	-0.13	-0.31	0.38
Gln	0.06	0.16	-0.19	-0.57	-0.24	0.23	0.02	-0.17
Glu	-0.74	-0.35	0.00	-0.95	-1.18	-0.06	-0.18	-0.43
Gly	0.13	-0.76	0.16	0.03	0.24	-0.28	0.12	0.13
His	-0.14	0.16	-0.22	-0.59	-0.81	0.16	0.29	0.00
Ile	0.20	0.21	0.15	0.33	0.51	0.22	0.20	0.29
Leu	0.26	0.33	-0.57	0.18	0.35	-0.40	-0.25	0.05
Lys	-0.27	-0.10	0.15	-0.34	-0.41	0.00	0.00	-0.24
Met	0.44	0.30	-0.21	0.48	-0.03	0.09	-0.03	-0.61
Phe	0.22	-0.01	-0.25	0.55	0.27	0.32	0.45	0.25
Pro	-1.06	-0.42	0.00	-1.17	-0.20	-1.63	-2.50	-0.36
Ser	-0.20	0.14	0.11	-0.15	-0.77	0.16	-0.08	-0.08
Thr	-0.11	0.40	0.54	-0.37	0.13	0.20	0.27	0.17
Trp	0.14	-0.27	0.05	0.29	-0.51	-0.05	-0.37	0.49
Tyr	0.10	0.06	0.16	0.37	0.09	0.37	0.68	0.21
Val	0.69	0.33	0.24	0.46	0.80	0.32	0.03	0.33

parallel	HB pairs				nHB pairs			
	j-2	j-1	j	j+2	j-2	j-1	j	j+2
Ala	-0.61	-0.53	0.16	-0.18	0.01	0.15	-0.23	-0.16
Arg	0.21	0.38	-0.96	-0.50	-0.19	-0.42	-0.42	-2.44
Asn	-0.08	-0.71	-0.55	-0.30	0.36	-2.37	-0.47	-0.02
Asp	0.77	-0.26	-0.67	0.30	0.17	-0.37	-0.59	-0.03
Cys	-0.87	-0.46	0.23	-0.88	-0.80	-0.80	-0.80	-0.39
Gln	-0.94	-0.72	-1.23	-0.39	-0.64	-2.21	-0.46	-1.57
Glu	-0.34	-0.89	-0.53	-0.44	-0.36	-0.68	-1.15	-0.81
Gly	0.16	-0.04	-0.79	-0.04	0.43	-0.71	-0.83	0.31
His	-0.03	0.25	-0.75	0.11	0.04	0.19	0.18	0.33
Ile	0.23	0.71	0.96	0.64	-0.25	0.72	0.29	0.69
Leu	-0.19	0.60	0.55	0.07	-0.02	0.67	0.52	0.07
Lys	0.48	-0.15	-1.41	-1.19	0.23	-0.15	-0.52	-1.62
Met	-1.06	0.45	-0.14	0.19	-0.28	0.41	0.27	0.41
Phe	-0.30	-0.09	0.32	-0.19	-0.36	-0.70	-0.01	0.78
Pro	-0.90	-1.13	-1.42	0.37	0.20	-0.84	-2.50	-0.65
Ser	-0.31	-1.20	-1.01	-0.41	-0.16	-0.65	-0.17	-0.24
Thr	-0.25	-0.04	-0.18	-0.11	-0.72	0.10	0.16	-0.03
Trp	-1.39	-0.72	-0.01	-0.02	-0.64	-2.50	0.29	-0.64
Tyr	-0.49	0.35	0.02	-0.10	0.09	-0.61	0.35	0.50
Val	0.66	0.46	1.02	0.65	0.36	1.03	1.11	0.44

§ HB and nHB pairs are definition for residue pairs (i, j) and (i, j+1) in case of antiparallel and parallel pairs, respectively.

さに対する分布を図3に表示した。平行 strand ペアの場合には strand 長3ないし4のペアが圧倒的に多く、それ以上の長さになると急激に減少し、strand 長9以上のペアはほとんど存在しない。これに対し

逆平行 strand ペアの場合には、strand 長が長くなるにつれて徐々に減少しているが、strand 長10以上のペアも100組以上存在している。残基間スコアは、アミノ酸残基 i に対して、j-2 から j+2 の位置における

Table 3 Prediction accuracy for strand pairs and triplets.

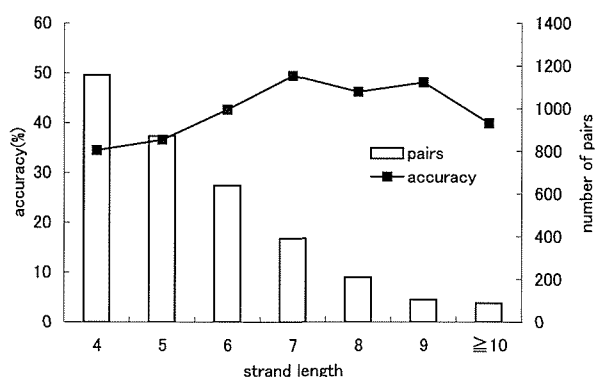
	strand pair						strand triplet			
	self §	pair [†]	2nd [†]	self, pair	self, pair, 2nd	pair, 2nd	pair	self, pair	self, pair, 2nd	pair, 2nd
accuracy										
total	0.250 $\pm$ 0.022	0.307 $\pm$ 0.036	0.232 $\pm$ 0.023	0.367 $\pm$ 0.022	0.390 $\pm$ 0.029	0.361 $\pm$ 0.042	0.425 $\pm$ 0.029	0.496 $\pm$ 0.035	0.516 $\pm$ 0.040	0.485 $\pm$ 0.032
anti-parallel	0.226	0.334	0.212	0.376	0.397	0.375	—	—	—	—
parallel	0.303	0.248	0.276	0.348	0.379	0.330	—	—	—	—

§ self score

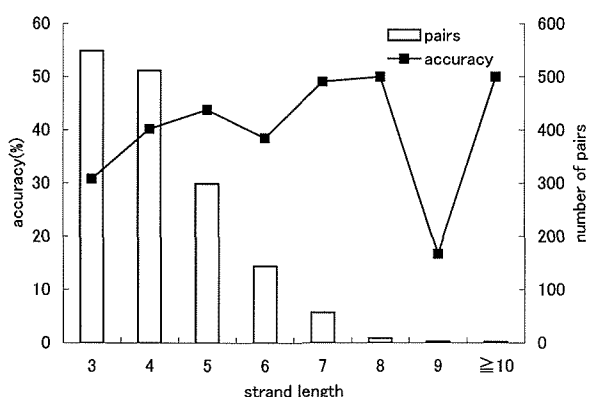
† pair score

† secondary structure score estimated from PREDATOR's result

(a)



(b)

**Fig. 3.** Prediction accuracy for strand pairs and distribution of number of strands against a strand length. Accuracy is shown by a closed square and number of strands is shown by an open bar. (a) antiparallel and (b) parallel pairs.

アミノ酸残基の存在傾向に基づいて計算している。位置 j は β strand 中に存在しなければならないが、それ以外の位置は必ずしも β strand 中とは限らない。Strand 領域が長くなるにつれて、これらの位置が β strand である割合が上昇してくる。strand 長 3 の場合、 $j \pm 1$ の位置が β strand である割合は共に 0.67、 $j \pm 2$ の位置では 0.33 であるのに対し、strand 長 4 の場合には、 $j \pm 1$ の位置では 0.75、 $j \pm 2$ の位置では 0.5 と

なる。逆平行 strand ペアの場合、平行 strand ペアと比較して strand 領域が長いために、 j 以外の位置も β strand 中である割合が平行 strand ペアよりも高くなっている。つまり、逆平行 strand ペアの方がより β strand 中に見られる傾向を反映しやすくなり、結果として自己スコアのみの場合よりも高い予測精度となったと思われる。これに対し、平行 strand ペアの場合には、 β strand 領域外の情報をより多く含むこととなり、この結果自己スコアのみの場合よりも低い予測精度となったのであろう。

自己スコアと残基間スコアを用いた場合、平均で 37%、逆平行 strand ペアで 38%、平行 strand ペアで 35% と自己スコアと残基間スコアを単独で用いた場合よりも予測精度が改善された。これは、自己スコアと残基間スコアが共に strand 間の正確な配置関係に欠かせないことを示している。また今回得られた数値は Steward と Thornton (2002) らの結果（逆平行 strand ペアで 31%、平行 strand ペアで 37%）とほぼ互角かそれ以上の値となった。彼らの方法との違いとしては、HB 残基と nHB 残基を区別してそれぞれスコアを求めた点にあり、このように HB 残基と nHB 残基を区別することが有効であることが今回の結果から明らかになった。

残基間スコアに自己スコアを加えることにより予測精度の向上が見られることから、ペアの相手となりうる領域の 2 次構造予測の精度を上げれば、より精度の高い配置予測を行えることが期待できる。自己スコアによってもある程度 β strand 形成能を判断できるが、これとは別に 2 次構造予測で得られた情報を加えることの効果を検証した。2 次構造予測プログラム PREDATOR を用いてタンパク質の全領域に対する 2

次構造予測を行い、該当するアミノ酸残基の2次構造（ α ヘリックス、 β strand、あるいはコイル）の割り当て及びその信頼度を得た。このPREDATORによる2次構造予測の情報に基づいて、方法に示した算出方法に従って求めた2次構造予測スコアを、自己スコアと残基間スコアの和に加えた場合の配置予測精度は全体で39%、逆平行 strand ペアに対して40%、平行 strand ペアに対して38%となり、全体と逆平行 strand ペアについて2ポイント、平行 strand ペアに対して3ポイントほど上昇した。この精度の上昇が2次構造スコアそのものによって達成されたものなのかどうかを判断するために、2次構造スコアのみの場合および2次構造スコアと残基間スコアの場合について検討した。この結果、2次構造スコアのみを用いた場合の予測精度は全体で23%、逆平行 strand ペアで21%、平行 strand ペアで28%となり、2次構造スコアと残基間スコアの場合、全体で36%、逆平行 strand ペアの場合で38%、平行 strand ペアの場合で33%となった。これらの結果はいずれも2次構造スコアを自己スコアに置き換えた場合の結果よりも低くなった。この結果より、自己スコアと残基間スコアの和に2次構造スコアを加えたことにより予測精度が向上したのは、自己スコアと2次構造スコアは一方が他方を完全に包括するような関係ではなく、重なり合う部分はあるがお互いに補完する関係のためであることが示された。これは、2次構造スコアの場合には、ある程度連続したアミノ酸配列が持つ領域としての性質を反映しているが逆平行 strand と平行 strand ペアの違いが反映されていないのに対し、自己スコアの場合には、個々のアミノ酸残基として逆平行 strand と平行 strand ペアの違いを反映しているが、アミノ酸配列としての性質を反映していないためであろう。以上の結果より、ペア形成における正確な位置を予測する上で、自己スコア、残基間スコアおよび2次構造スコアの和を用いるのは妥当であることが示された。

ペア領域の長さとの予測精度の関係

図3に strand の長さとの予測精度の関係を示した。

逆平行 strand ペアも平行 strand ペアも共に strand 長が伸びるにつれて、精度が向上している。これは、strand 領域が長くなることにより、判断に使用される情報が増加するためである。一方で8残基以上の長さの逆平行ペアについては精度の低下が見られる。これは strand 領域が長いために、多少正確な位置関係からずれたとしてもその配置でも実際に存在しやすい残基ペアが多く形成されるためであると考えられる。

予測位置の分布状況

正確な配置からの予測位置のずれを offset と呼ぶことにし、N 端側へのずれを負の値に、C 端側へのずれを正の値とした。図4はこの offset に対する予測結果の分布状況を示したものである。逆平行 strand ペア、平行 strand ペアとも offset が0、つまり正確な位置関係を予測した場合が最も多く分布し、offset が

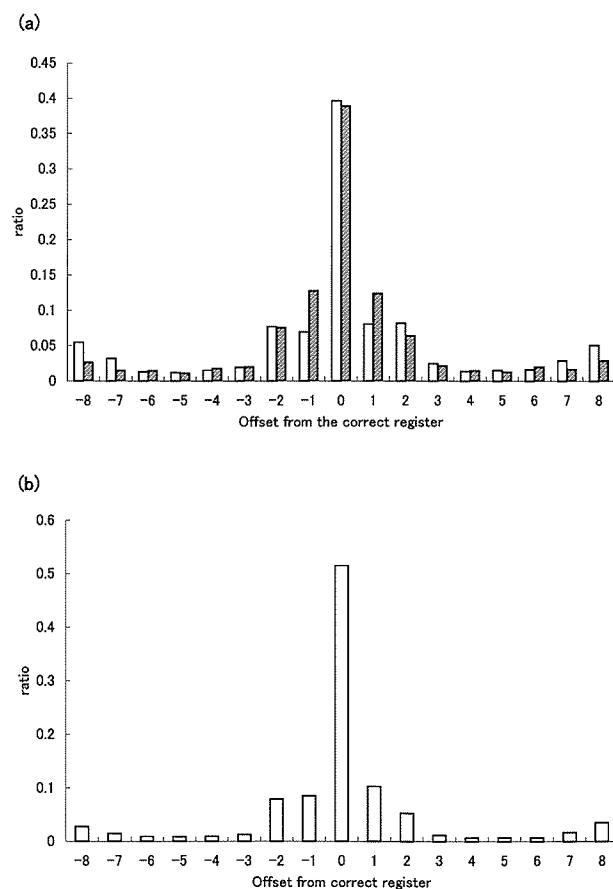


Fig. 4. Distribution of number of strands against a predicted position. Predicted position is shown by an offset. (a) Open and hatched bars denote antiparallel and parallel pairs. (b) strand triplets.

1 あるいは 2 となる場合がこれに続く。逆平行 strand ペアの場合には、offset 1 の場合よりも、2 の場合の方がやや多くなっているのに対し、平行 strand ペアの場合には offset 1 の場合が offset 2 の場合よりも多くなっている。これはおそらく水素結合様式の違いを反映していると考えられる。逆平行 strand ペアの場合には、向き合った残基が 1 残基おきに水素結合を形成するために、HB 残基ペアと nHB 残基ペアが交互に並ぶ。今回の解析においては HB 残基ペアと nHB 残基ペアを明確に区別している。多くの残基ペアには HB 残基か nHB 残基かの指向性があるので、奇数残基分ずれるとこの HB 残基か nHB 残基かの識別が全てずれることになり、そのような位置関係は好ましいものではない。偶数残基分ずれる場合には、この HB 残基と nHB 残基の関係は維持されるので、1 残基分のずれよりも 2 残基分のずれの方が起こりやすいのであろう。これに対し、平行 strand の場合には、全体的に strand 領域が短く、斜め向かいの残基との間で水素結合を形成する。また向き合った相手の残基を中心に見れば、HB 残基と nHB 残基の位置が逆になる。このために、逆平行 strand ペアの場合に比べると 1 残基位置が前後にずれてもそのような位置関係が許されるのかもしれない。3 残基から 6 残基分のずれについては徐々に減少あるいはほぼ横ばいであるのに対し、7 残基、8 残基分のずれではやや増加している。これは、逆平行 strand ペアの場合には、strand 間を結ぶループの長さが短いものも多く存在し、今回設定した範囲内にこれら別の strand 領域が含まれることも多々存在する。このような実際とは別の領域をペアの相手として認識されるために、7 残基、8 残基分ずれた領域を予測したものと考えられる。

Strand トリプレットに対する位置予測

3 本以上の strand からなる β sheet においては、両端の strand を除いた strand では、その両側に strand が位置し、典型的な水素結合を形成する場合には、中央に位置する strand 領域の各残基は両側に存在する 2 本の strand のうちのどちらか 1 本の strand の残基と水素結合を形成する。このため 3 本の strand

(strand トリプレット) が並ぶ場合、2 つの strand ペアが独立して形成されるのではなく、一方のペア形成が、他方のペア形成に制限を与える。この制限が 1 つの限定条件となり、strand 配置予測精度の向上につながる事が十分予想される。そこで PDB から得られた 2714 組の strand トリプレットに対し、両側の 2 本の strand を実際の位置関係に固定し、中央の strand をスライドさせることにより正確な配置関係を予測できるかを検証した。

予測精度の結果を表 3 にまとめた。残基間スコアのみの場合の予測精度は 43% であり、strand ペアの場合と比較すると 18 ポイント向上し 2 倍近くの精度となった。この結果は strand トリプレットによる制約が予測精度の向上に十分効果的であることを示している。自己スコアと残基間スコアを用いた場合は 50% であり、Steward と Thornton (2002) らの結果 (逆平行トリプレットで 45.2%，平行トリプレットで 48.5%) を strand ペアの場合と同様に上回った。さらに 2 次構造スコアを加えた場合には 52% まで上昇した。残基間スコアと 2 次構造スコアの場合は 49% となり、このトリプレットの場合も、strand ペアの場合と同様に自己スコアと PREDATOR による 2 次構造予測の結果は補完的な関係であった。

β スtrand 存在比と予測精度

逆平行 strand ペアと平行 strand ペアの予測精度の差や strand の長さや予測精度の関係より、予測精度は strand 長に依存していた。これは strand 間での残基ペア数の増加が予測精度の向上をもたらしたと見なせる。タンパク質立体構造中で β strand 領域の占める割合はヘモグロビンなどのようにまったく含まないものから、膜タンパク質の一種であるポリンのように 50% を超えるものまで様々である。 β strand 領域がタンパク質中で占める割合が高ければ、個々の strand が長くなったり、3 本以上の strand から構成される β sheet の占める割合が高くなると予想できる。既に表示したように、strand の長さが伸びるに従って予測精度の向上が見られ (図 3)、また strand ペアよりも strand トリプレットにおいて予測精度が大幅に向上

することから、 β strand 領域の占める割合（タンパク質中の全アミノ酸残基数に対する strand 領域に含まれる残基数の割合）と strand ペアおよびトリプレットの予測精度の関係を検討し、その結果を図 5 に示した。Strand ペアに対する予測精度（図 5(a)）は 37% から 43% という狭い範囲で変動しており、 β strand 領域の占める割合の違いによる明確な差は見られなかった。しかしながら、予測精度が 43% となるのはいずれも β strand 領域の存在比が 0.35 以上の場合であることから、弱いながらも β strand 領域の存在比に対する依存性があるように見える。Strand トリプレットの場合（図 5(b)）は、予測精度の範囲は 39% から 61% と strand ペアの場合よりも差が広がった。 β strand 領域の存在比が 0.05 以下の場合、予測精度は 0.39 と他の場合に比べて低くなったが、これは strand

トリプレットが 23 組とデータ数が極端に少ないことが影響しているのかもしれない。しかしながら、これを除いても変動幅は 14 ポイントとなり、strand ペアの場合の 2 倍の変動幅であった。 β strand 領域の存在比が 0.05 以下の場合を除外した場合、存在比 0.25 から 0.35 付近の予測精度が最も低く、それより β strand 領域の存在比が低下するあるいは増加するにつれて予測精度が上昇する傾向が見られた。Strand ペアに対しても、この存在比 0.25 から 0.35 付近では他の領域よりもやや予測精度が低下していた。また存在比 0.4 から 0.45 の範囲に対する予測精度も strand ペアの場合も strand トリプレットのいずれの場合でも低下していた。

次に正確な位置を予測できなかった strand ペアおよび strand トリプレットに対し、正確な位置からのずれの分布パターンを図 6 に示した。Strand ペアにおいて（図 6(a)）、offset の絶対値が 1 または 2 の割合は、どの β strand 領域存在比においても 45% 以上となっているが、存在比が高まるにつれて、offset 1 の割合が減少する一方で offset 2 の割合が増加し、存在比 0.35 から 0.4 の範囲を境にしてその比率の大小関係が逆転した。さらに、offset 7 や offset 8 の割合も β strand 存在比と共に増加する傾向を示したが、strand 存在比が 0.45 以上の範囲ではやや減少した。Strand トリプレットの場合（図 6(b)）も strand ペアの場合と同様に offset 1 および offset 2 の割合が高く、どの β strand 領域存在比においても 50% 以上となった。特に offset 2 の割合は strand ペアの場合よりも高い傾向を示した。一方 offset 8 の存在比は strand 領域存在比 0.25 から 0.3 付近が最高であり、strand 領域存在比が減少する方向に対しても増加する方向に対しても、offset 8 の存在比は減少した。この offset 8 が示す傾向と予測精度との間には関連性が見られる。すなわち、offset 8 の存在比が低下するにつれて予測精度が向上していた（図 5(b) と図 6(b)）。strand 領域の存在比が 0.25 より低下する方向において offset 8 が減少し予測精度が向上したのは、正確な配置関係から前後 8 残基の範囲に他の β strand 領域が存在する割合が低下したためと思われる。Strand 領

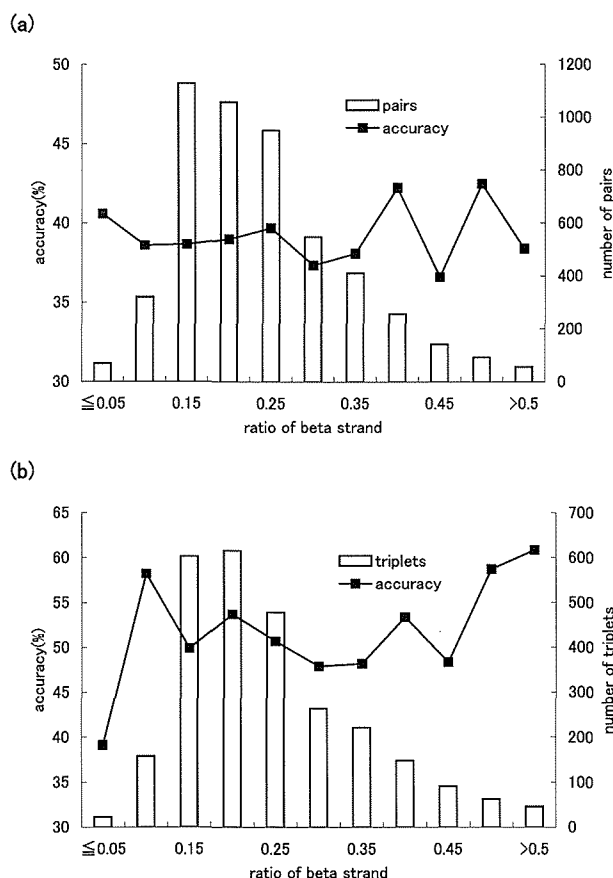


Fig. 5. Prediction accuracy against ratio of β strand. Ratio of β strand is a value that number of residues in β strand regions is divided by number of residues in a protein. Accuracy is shown by a closed square and number of strand pairs or triplets is shown by an open bar. (a) strand pairs and (b) strand triplets.

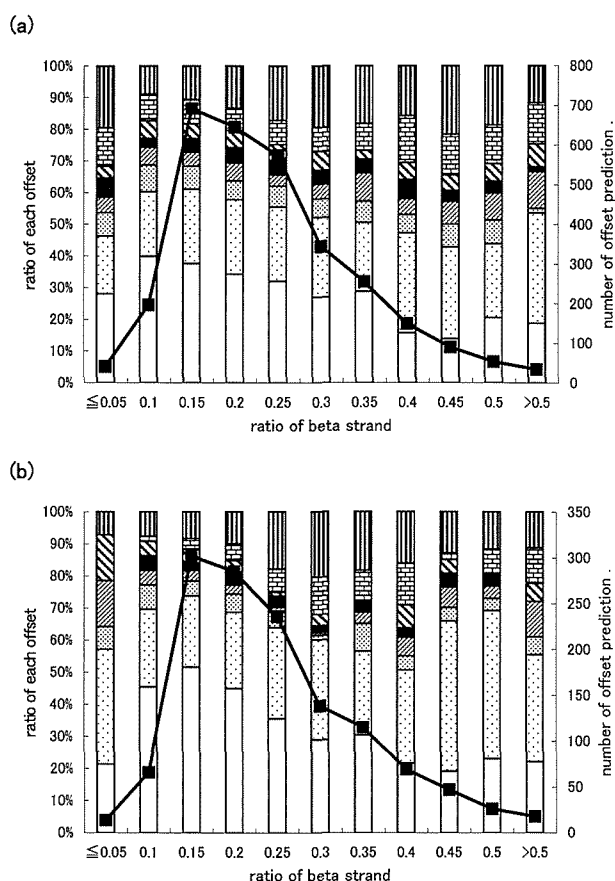


Fig. 6. Offset pattern against ratio of β strand. Number of offset i is a sum of pairs or triplets that predicted strand position is i -residue away from the correct position. Ratio of number of offset is denoted by the height of a bar (offset 1: , offset 2: , offset 3: , offset 4: , offset 5: , offset 6: , offset 7: , offset 8: ). Total number of offset prediction is denoted by a closed square. (a) strand pairs and (b) strand triplets.

域の存在比が低下すれば、スコアを計算する際に検索する領域中に他の β ストランドが存在する確率が低下し、そのような strand 領域を選択する確率も低下するので、結果として正確な位置関係を選択する可能性が高まるのであろう。一方、strand 領域の存在比が0.3以上では、存在比が上昇するにつれて検索範囲に他の β strand が存在する確率は高まるが、3本以上の strand からなる sheet が存在する割合も増加する。Strand トリプレットの場合には、両側に存在する2本の strand との関係と同時に満足する配置関係になればならず、offset 8 となるような位置関係は適さない場合が多いのであろう。また、strand 領域の存在比が0.35以上になると offset 1 よりも offset 2 の割合が高くなるのも、3本以上の strand からなる sheet

が増加し、水素結合と非水素結合のパターンを維持する必要が高まっているからと思われる。

2 次構造の予測精度の向上

strand 間の配置予測精度の向上に対して2次構造予測の結果が有効であることが上のように示された。この配置予測精度の向上ということは、 β strand 領域の予測精度の向上と見ることも可能である。そこでどの程度 β strand 領域の予測精度が改善されたかの検討を行った。ここで PREDATOR による β strand 領域の予測精度は、実際の β strand 領域全域において β strand として予測された残基が占める割合とした。これに対し、本方法による β strand 領域予測の精度は、実際の strand 領域中の残基がペアとして予測された strand 領域に占める割合として示した。Strand 長4の領域に対して予測された配置関係が offset 1 の場合、実際の β strand 領域中の3残基は strand 領域として予測されるので、この場合の予測精度は0.75であり、offset の絶対値が strand 長よりも大きくなれば、その領域に対する β strand 予測精度は0となる。

β strand ペアとトリプレットに対する結果を表4に示した。PREDATOR による実際の β strand 領域に対する予測精度は58%であるが、自己スコア、残基間スコア及び2次構造スコアを用いた予測では68%と10ポイントの精度向上が見られた。この方法では正確な配置関係の予測精度は39%であったが、offset の絶対値が1および2であった予測結果がそれぞれ5%以上存在するので、合計60%の strand に対しては、2残基以内のずれで strand 領域を予測したことになる。このため、実際の strand 領域のうち、 β strand と予測された残基の割合が68%となった。strand トリプレットにおける strand 領域の予測ではさらなる精度向上が見られた。配置関係予測において正確な予測が50%を超えていたことに加え、この場合でも ± 2 以下のずれとして予想された strand の割合はそれぞれ10%弱存在するので、実際の strand 領域のうち β strand と予測された残基の割合が80%となり、PREDATOR のみによる予測精度に対して23ポイントの上昇となった。実際の2次構造は、ローカルな環境

Table 4 Prediction accuracy for β strand regions.

	strand pair					strand triplet	
	pair [†]	self §, pair	pair, 2nd [†]	self, pair, 2nd	2nd	self, pair, 2nd	2nd
total residues			51,492			18,154	
correct residues	30,678	33,873	33,502	34,820	29,670	14,602	10,442
accuracy	0.596	0.658	0.651	0.676	0.576	0.804	0.575

§ self score

† pair score

† score estimated from PREDATOR's result

のみならずいわゆるカメレオン配列のように空間的な環境に影響される場合があることが知られている (Minor と Kim 1996)。タンパク質の配列全体の類似性に基づいて鋳型となる既存の立体構造を類推し、類推した立体構造に基づいて2次構造を推定するのではなく、PREDATORのようにローカルなアミノ酸配列パターンと2次構造の相関性から2次構造を予測する方法では、特に β strand領域の予測精度が α ヘリックス領域よりも低くなってしまふ。今回の方法では、残基ペアに関する傾向というグローバルな情報を加えることにより、PREDATORが持つ弱点を補うことができたために、このような大幅な予測精度の向上という結果になったといえる。

考 察

本研究で用いた方法では、HB残基とnHB残基を区別して自己スコアや残基間スコアを求めることにより、従来の方法よりも高い精度でstrandペアやstrandトリプレットの配置関係を予測することが出来た。特に逆平行strandペアにおいて予測精度が向上した。これは逆平行strandに配向する場合には、HBペアとnHBペアが交互に並ぶためである。表1に示したように、隣接する2本のstrand上の残基ペア (i, j) がHBペアかnHBペアであるかによって、多くの残基ペアで差が生じている。つまり実際のタンパク質中では、隣接するstrand間で、向き合う位置に存在するアミノ酸ペアはHBペア部分だけではなく、nHBペア部分についても任意の残基が位置しているわけではないことを示している。2本のstrandだけで形成されるsheet (β ブリッジ) については、その2本のstrand間でのペア形成のみが重要であり、

strand上のアミノ酸残基はHB残基かnHB残基のどちらか一方の役割を満たせばよい。極端な場合、出現頻度が高いHBペアが十分形成されるようであればnHBペアについてはstrandペア形成を妨げるようなアミノ酸ペアでなければよい場合もあるであろう。このような場合には、nHBペアを形成するアミノ酸残基の選択は自由度が高くなる。しかしながら、3本以上のstrandがsheetを形成する場合、sheetの両端に位置するstrandを除く残りのstrand (central strand) はその両側に位置するstrandとの間で水素結合を形成する必要がある。典型的な逆平行strandペアのみでsheetが形成されている場合には、central strand上の各アミノ酸残基は一方のstrandとHBペアを、他方のstrandとnHBペアを形成し、しかもその向きが1残基ごとに入れ替わることになる。この場合には、central strand上の各アミノ酸残基はHBペアの相手としてもnHBペアの相手としても制限を受けることになる。これは、隣接する2本のstrand間でのペア形成が、このstrandを含むsheet内の他のペア形成に影響を持つことを意味する。4本のstrandからなるsheetにおいて横に並んだ部分を考える場合、1番のstrandとその隣の2番のstrand間でのHBペアは、一見、残りの2本のstrand間でのHBペアとは無関係のように見える。しかしながら、2番のstrandと3番のstrand間ではnHBペアが形成されることになるので、このnHBペアが好ましいアミノ酸残基の組み合わせであることが要望される。このため、strand 1と2の間でのHBペアに関与するアミノ酸残基の組み合わせはstrand 3と4の間でのHBペアに関与するアミノ酸残基のペアに影響を持つことになる。このことは隣接するstrand間の残基ペアの影響

は単にその strand 間にのみ影響を持つのではなく、sheet 全体について影響を持つことを意味する。また逆の見方をすれば、4 本以上の strand から形成される sheet の部分構造である strand トリプレットにおいても、sheet 全体からの制約を受けたペア形成が求められることになり、それがこの部分のアミノ酸配列に反映され可能な配列パターンが限定される。strand トリプレットの配置予測を行う際、この限定的なパターンのいずれかを持つ領域はそれほど多くは存在しないために、配置予測精度の大幅な向上に結びついたのであろう。Sheet を構成する strand の本数が増加するほど、また strand 領域が長くなるほど、sheet 全体での制約は強くなる。このため、strand 領域の存在比が0.35以上の場合、図 5(b)における予測精度の向上や図 6(b)における offset 1 や 8 の減少および offset 2 の増加という結果をもたらしたのであろう。一方、strand ペアまで分解してしまうと、strand ペアを形成するという条件において実際の位置よりも適した部分が存在する確率が高くなり、strand トリプレットほどは、strand 領域の存在比の効果は現れないのであろう。

今回の方法は従来の方法よりもデータを詳細に整理したことにより高い精度で strand ペアや strand トリプレットの配置関係を予測することが出来た。この傾向を延長させる、つまりよりデータを細分化していけば、より予測精度が高まることが期待できる。しかしながら、このために必要となるデータ数も増加してしまう。データ数が少ない場合にデータを細分化してしまうと、使用したアミノ酸配列全体が持つ様々な傾向を統計的に示すというよりもそのまま記憶してしまうことになる。この結果、スコアを計算するために利用したタンパク質から取り出した strand ペアや strand トリプレットに対する正確な位置関係の検出精度は高まるが、テスト用に残しておいた strand ペアや strand トリプレットに対する予測精度は低下してしまう。本研究においては、残基間スコア $P_{i,j,j+m}$ や $P_{i,j+1,j+m}$ を算出する際に、位置 j の残基を20種類の残基に区分する代わりに、6 グループに区分したのはこのためである。20種類の残基に区分した場合につ

いても計算した（データ未表示）が、その予測精度は6グループに区分した場合よりも低下してしまった。この結果から、残基間スコアについて位置 j の残基を20種類に区分して処理するためには、単純に見積もって最低でも3.3 (20/6) 倍のデータ数が必要と思われる。このような状況から判断して、今回用いたタンパク質数に対するデータの細分化については本研究で用いた方法が最善であり、区分方法を変化させることによる精度向上はあまり期待できないであろう。

本研究において、自己スコアと残基間スコアに PREDATOR による2次構造予測の結果も加えることにより、strand ペアおよび strand トリプレットの配置関係の予測精度がさらに向上することも示された。この結果はある程度予想できるものであったが、残基間スコアと PREDATOR の結果のみを使用した場合の予測精度が、自己スコアと残基間スコアを用いた場合の予測精度と同程度かそれ以下であったことは注目すべきであろう。自己スコアではどのアミノ酸残基が目的とする領域に含まれているかは反映されるが、どのように配列されているかは反映されていない。これに対し、PREDATOR による2次構造予測では、配列まで考慮して2次構造予測を行っている。この点において PREDATOR による結果を利用した場合に、より正確に strand 領域を予測することが期待される。それにも関わらず、自己スコアと残基間スコアの組み合わせのほうが残基間スコアと PREDATOR の予測結果の組み合わせよりも予測精度がわずかに勝っていた理由としては、残基間スコアが果たした役割の違いが考えられる。PREDATOR による2次構造予測結果では、アミノ酸配列を反映していることから、strand 領域とそれ以外の領域との区別がより明確にされるのに対し、自己スコアの場合には配列パターンが反映されないために平均化されたスコアとなり、strand 領域とそれ以外の領域との差が PREDATOR の結果よりも小さいことが十分考えられる。このため、自己スコアと残基間スコアの組み合わせの場合には相対的に残基間スコアの重みが増す結果となり、strand 間のグローバルな相互作用がより反映され予測精度が高まったのではないと思われる。

る。これに対し、PREDATORの結果と残基間スコアの組み合わせの場合にはPREDATORの結果により左右され、グローバルな相互作用が反映されている残基間スコアの影響が相対的に低くなったのであろう。事実、PREDATORにより実際のstrandペア形成の相手となる領域以外のstrand領域が予想された場合では、後者の領域を予測領域として選択する場合が多く観測されている（図4のoffset ± 7 , ± 8 など）。自己スコア、残基間スコアおよびPREDATORの予測結果に基づくスコアの3種類を使用した場合には、2種類のスコアを用いた場合に比べてそれぞれのスコアが占めるウェイトが低下し、これら3種類のスコアがお互いに補完的に作用することで、より高い予測精度が得られたのであろう。

今回の方法において、2次構造予測の情報を加えることにより予測精度の向上が確認されたがその予測精度は十分に高いとは言えず、まだ改善の余地が残されている。第1の改善方法は、アミノ酸配列からstrand領域を予測する精度を高めるために、より精度の高い2次構造予測の結果などを利用することであり、第2の方法は、より多くの残基間情報を用いて残基間スコアの改善を図ることである。2次構造の予測精度の向上には、類似タンパク質がある程度の種類存在する場合にはマルチアライメントを作成して配列プロファイルを作成でき、この配列プロファイルを用いることが有効であることが知られている（RostとSander 1993；1994）。よってstrandペアの配置関係を予測するタンパク質に対して、配列が類似したタンパク質を用いて得られた配列プロファイルに基づく2次構造予測を用いれば、この研究で用いた方法の予測精度の向上が期待できる。またこの配列プロファイルを作成する際に利用したマルチプルアライメントより、保存性や揺らぎパターンを反映させた残基間スコアを求めることが可能となる。このようにして求めたスコアを用いることにより、配列ペアの多様性についてもある程度対応することが可能となるので、未知のアミノ酸配列に対する予測精度の向上が見込まれる。この類似配列によるマルチプルアライメントを用いる方法については今後さらに検討していく予定である。

今回開発した方法は、一方の領域を固定し、他方のstrandの配向が判っているという制約は設けているが、他の立体構造的な制約は含めていない。このため、多くの逆平行ペアにおいて、検索領域中に別の β strand領域や自分自身を含んだ例に対して、これらの領域をペア形成領域として選択した場合が見られた。またstrand間を結ぶ領域が2残基以下となるような予測結果も含まれていた。このような例は、実際のsheet構造予測の段階では除去されることを考慮すると、この方法の予測精度はさらに向上する。これらの点を踏まえると、今回開発した方法は単独で使用するよりも他の立体構造予測法と併用することにより、その機能をより発揮することは明らかである。特にある範囲内でのstrandペアを形成する領域の検索や複数の候補領域から最適なペアの相手を決定する場合に適している。また、HB残基とnHB残基を区別して計算を行っているので、このHB残基とnHB残基の識別が重要になる3本以上のstrandからなるsheetにおける配置関係予測の場合が最適であると思われる。さらにこのような立体構造予測に対する直接的な関与に加えて、2次構造、特に β strand領域、の予測の精度向上による間接的な関与も十分期待できる。多くの2次構造予測法では、ローカルなアミノ酸配列情報を用いて2次構造の割り当てを行っているが、実際に β strandを形成するかどうかについては空間的な環境にも左右されている（MinorとKim 1996）。今回開発した方法では、残基間スコアという形で空間的な環境を含ませることにより、PREDATORによる β strand領域に対する予測精度を大幅に補正することができた。

以上のように、この方法は今後の立体構造予測法の精度改善に大いに貢献することが期待できる。

引用文献

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25: 3389–3402. 1997.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. The protein data bank. *Nucleic Acids Research*. 28: 235–242. 2000.

- Chou PY and Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*. 13: 211–222. 1974(a).
- Chou PY and Fasman GD. Prediction of protein conformation. *Biochemistry*. 13: 222–245. 1974(b).
- Dayhoff MO. Survey of new data and computer methods of analysis. In *Atlas of Protein Sequence and Structure*. vol. 5, suppl. 3. National Biomedical Research Foundation. Washington DC. 1978.
- Fooks HM, Martin ACR, Woolfson DN, Sessions RB and Hutchinson EG. Amino acid pairing preferences in parallel β -sheets in proteins. *Journal of Molecular Biology* 356: 32–44. 2006.
- Frishman D and Argos P. Knowledge-based protein secondary structure assignment. *Proteins*. 23: 566–579. 1995.
- Frishman D and Argos P. Incorporation of non-local interaction in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*. 9: 133–142. 1996.
- Frishman D and Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*. 27: 329–335. 1997.
- Hobohm, U, Scharf M, Schneider R and Sander C. Selection of representative protein data sets. *Protein Science*. 1: 409–417. 1992.
- Jiang B, Guo T, Peng LW and Sun ZR. Folding type-specific secondary structure propensities of amino acids, derived from alpha-helical, beta-sheet, alpha/beta, and alpha + beta proteins of known structures. *Biopolymers*. 45: 35–49. 1988.
- Jones DT. Gen THREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*. 287: 797–815. 1999.
- Kabsch W and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22: 2577–2637. 1983.
- Karplus K, Barrett C and Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 14: 846–856. 1998.
- Kelly LA, MacCallum RM and Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*. 299: 499–520. 2000.
- Lee J, Liwo A, Ripoll DR, Pillardy J and Scheraga HA. Calculation of protein conformation by global optimization of potential energy function. *Proteins (suppl.)*, 3: 204–208. 1999.
- Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry*. 17: 4277–4285. 1978.
- Liwo A, Lee J, Ripoll DR, Pillardy J and Scheraga HA. Protein structure prediction by global optimization of a potential energy function. *Proceedings of National Academic Sciences of U.S.A.* 96: 5482–5485. 1999.
- Mandel-Gutfreund Y, Zaremba SM and Gregoret LM. Contributions of residue pairing to β -sheet formation: Conservation and covariation of amino acid residue pairs on antiparallel β -strands. *Journal of Molecular Biology*. 305: 1145–1159. 2001.
- Minor DL Jr. and Kim PS. Context-dependent secondary structure formation of a designed protein sequence. *Nature*. 380: 730–734. 1996.
- Rost B and Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of National Academic Sciences of U.S.A.* 90: 7558–7562. 1993.
- Rost B and Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*. 19: 55–72. 1994.
- Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L and Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*. 15: 1000–1011. 1999.
- Shi J, Blundell TL and Mizuguchi K. Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*. 310: 243–257. 2001.
- Steward RE and Thornton JM. Prediction of strand pairing in antiparallel and parallel β -sheets using information theory. *Proteins: Structure, Function, and Genetics*. 48: 178–191. 2002.
- Zhu H and Braun W. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β -sheet formation in proteins. *Protein Science*. 8: 326–342. 1999.