平成 11 年度～平成 13 年度科学研究費補助金（基盤研究（ｃ）（２））

研究成果報告書

タンパク質立体構造における局所構造モチーフの分類と解析

課題番号　１１６８０６６６

平成 14 年 3 月

研究代表者　　輪湖　博

早稲田大学社会科学部・教授

平成 11 年度～平成 13 年度

科学研究費補助金（基盤研究（ｃ）（２））

研究成果報告書

研究課題名

タンパク質立体構造における局所構造モチーフの分類と解析

課題番号　１１６８０６６６

研究組織

研究代表者　　輪湖　博
早稲田大学社会科学部・教授

研究経費

平成 11 年度	１，１００千円
平成 12 年度	１，０００千円
平成 13 年度	１，１００千円
計	３，２００千円

研究発表

（１）学会誌等

・Jianghong An, Takao Nakama, Yasushi Kubota, Hiroshi Wako, and Akinori Sarai, Construction of an integrated environment for sequence, structure, property and function analysis of proteins. *Genome Informatics*, **10**, 229-230 (1999)

・Jianghong An, Hiroshi Wako, and Akinori Sarai, Analysis of structural motifs of proteins in terms of sets of codes representing local structures. *Mol. Biol.* **35,** 1056-1062 (2001)

・Hiroshi Wako, Jianghong An, and Akinori Sarai, Environment-dependent and position-specific frequencies of amino acid occurrences in α-helices. submitted.

（２）口頭発表

・輪湖　博、Delaunay四面体分割を利用した局所構造モチーフの解析：疎水コア、蛋白合同年会・横浜 99（第 11 回日本蛋白工学会年会）1999 年 6 月、横浜。

・輪湖　博、Delaunay四面体分割を利用した局所構造モチーフの解析：疎水コア、日本生物物理学会第 37 回年会 1999 年 10 月、和光。

・安　江虹、輪湖　博、皿井明倫、蛋白質立体構造モチーフ解析、蛋白合同年会・東京 2000（第 12 回日本蛋白工学会年会）2000 年 6 月、東京。

・安　江虹、輪湖　博、皿井明倫、立体構造モチーフの表現と探索、日本生物物理学会第 38 回年会、2000 年 9 月、仙台。

・輪湖　博、安　江虹、皿井明倫、PROSITEパターンを含むrigidな局所構造の同定、第１回日本蛋白質科学会年会、2001 年 6 月、大阪。

・Hiroshi Wako, Jianghong An, and Akinori Sarai, *Rigid local structure motifs in proteins detected by the Delaunay tessellation method.* 4-th International Conference on Biological Physics, 2001 年 7-8 月、京都。

・Yoshinori Nagai, Hiroshi Wako, and Stephen T. Hyde, *Geometrical analysis of protein structures,* 4-th International Conference on Biological Physics, 2001 年 7-8 月、京都。

・輪湖　博、安　江虹、皿井明倫、PROSITEパターン近傍の局所構造の解析、日本生物物理学会第 39 回年会、2001 年 10 月、大阪。

・Hiroshi Wako, *Detection of local structure motifs with respect to inter-residue networks obtained from Delaunay tessellation,* 2001 年 12 月、大阪。

（３）出版物

なし

研究の概要

(1) 研究の目的

タンパク質立体構造の構築原理を明らかにするために、タンパク質の立体構造データバンク（ＰＤＢ）に登録された大量の立体構造データを総合的に解析する手法を開発し、それに基づいた解析を行うことが、本研究の目的である。具体的には、タンパク質の立体構造をそれぞれ Delaunay 四面体で分割、各四面体に局所構造を反映したコードを付け、局所構造をそれらコードで表現することによって、異なるタンパク質の立体構造の間に共通に見出される部分構造（局所構造モチーフ）を同定するという申請者らが提案した方法（Wako & Yamato, 1998）を改良し、その方法によって得られたさまざまな局所構造モチーフを分類、整理する。その上で、モチーフごとに、各サイトにおけるアミノ酸の出現頻度や、構造間の違いなどを調べ、局所構造モチーフにおけるアミノ酸が不変な部位と多様な部位、構造が限定されている部位と構造の違いが大きい部位などの特徴を明らかにし、その構築原理に関する知見を得ることを目的として研究を行った。

(2) 研究の特色

ここで用いられた方法の概略は以下の通りである（添付資料⑥）。

a) 与えられたタンパク質の立体構造をＣα原子を頂点にもつ Delaunay 四面体で空間的に分割する。なお、この分割は一意的に定まる。
b) 各四面体に、局所構造の特徴を反映するような規則に従ってコード付けを行う。
c) 高い相同性をもつものを除いたＰＤＢデータのセットについて、Delaunay 四面体分割とコード付けを行い、コード付けされた四面体のデータベースを構築する。

同じコードを持つ四面体をさまざまなタンパク質から収集することによって、同じタイプの局所構造を収集することができる。

この方法は以下のような利点をもつ。

a) それぞれの局所構造がコード(文字列)で記述されるため、コンピュータで局所構造データの検索や統計的処理を行うのに都合がよい。
b) 局所構造が、Delaunay 四面体の辺（空間的に近接する残基間の相互作用を表すと考えられる)のネットワークに注目して定義されるため、孤立した二次構造ではなく、その周囲にあるアミノ酸残基を含んだ構造要素として定義される。

これまで局所構造を解析するために提案されてきた方法のほとんどが、ペプチド鎖に沿って連続した領域によって作られる構造を対象としていたのに対して、この方法では、ペプチド鎖に沿っていることは陽には考慮せず（実際には暗に考慮されている)、あくまでも空間的に隣接するアミノ酸残基によって作られる局所構造を対象としているた

め、アミノ酸の相互作用に注目した局所構造の同定法となっているところが特徴である。

(3) 研究成果

a) 局所構造コードの改良

Delaunay 四面体に付けるコードは、当初アラビア数字１－８が用いられたが（添付資料⑥)、本研究を進めて行くうちに、アルファベットＡ－Ｈおよびａ－ｈを使用することでより的確な、そして情報量もより豊富なコードが付けられることを見出し、コード付けのルールを改訂した（添付資料③)。これによって、

1) 複数のサブユニットに属するアミノ酸残基で構成される局所構造にも適用できるようになった。
2) 局所構造を構成するアミノ酸が、ペプチド鎖に沿っていくつの領域から構成されているかが表現でき、また、その各領域のペプチド鎖に沿っての順番に無関係にコードが付けられることから、コードと局所構造の対応がより明確なものとなった。

ｂ）より大きな局所構造モチーフの同定法の開発

一つの局所構造コードだけで表すことのできる局所構造の大きさは限定されている。そこで、より大きな局所構造も同定できるように、新たな方法を開発した。

1) PROSITE のようなアミノ酸配列のモチーフとして与えられた局所構造の集合について、そこに含まれる Delaunay 四面体のコードを互いに比較することで、立体構造の類似性を調べた。このとき、コードの類似度という考え方を初めて導入した。(添付資料②)
2) 同じコードをもつ一つの四面体をデータベースから検索するだけでなく、それをスタートに、それに隣接し、しかも同じコード番号をもつ四面体をできる限り探し出すことで、より大きな局所構造モチーフを同定するアルゴリズムを新たに考えた。（添付資料⑤)

ｃ）応用　１．データベース構築の検討

理研筑波研究所の皿井らによって構築されたタンパク質立体構造およびその機能に関する総合的データベース 3DinSight に本研究の Delaunay 四面体を利用した局所構造のデータベースを組みこむための検討を行った（添付資料①)。残念ながら、現在のところ実現にまでは至っていない。

ｄ）応用　２．αヘリックスにおける位置依存及び環境依存のアミノ酸出現頻度

αヘリックス上のアミノ酸の出現頻度について、αヘリックスの周囲にあるアミノ酸残基への依存性（環境依存性）とそのときのαヘリックス上の位置への依存性（位置依存）を考慮した解析を行った。(添付資料③)

ｅ）応用　３．疎水コア

4 つの頂点がすべて疎水基で占められるようなDelaunay四面体を含む局所構造を疎水コアと定義し、局所構造のデータベースから検索、そこからモチーフと呼べるような何か特徴的構造が浮かび上がってくるか調べた。(添付資料④)

ｆ）応用　４．相同タンパク質の分類

PROSITEのようなアミノ酸配列のモチーフとして与えられた局所構造の集合について、コードの類似度を考慮しながら、そこに含まれる Delaunay 四面体のコードを互いに比較することで、立体構造の類似性を調べた。(添付資料②)

ｇ）応用　５．相同、非相同タンパク質間で見出される共通の構造

PROSITE のようなペプチド鎖に沿って定義された配列モチーフを、空間的に隣接する残基からなる局所構造モチーフとして再定義し、解析した。そこで、PROSITE で定義されたいくつかの配列モチーフについて、ペプチド鎖に沿って離れたアミノ酸で、モチーフ構造に重要な役割を果たしているアミノ酸を検出できるかどうかを調べた結果、いくつかの重要な役割を果たすアミノ酸を新たに同定することができた。(添付資料⑤)

添付資料

① Jianghpng An, Takao Nakama, Yasushi Kubota, Hiroshi Wako, and Akinori Sarai, Construction of an integrated environment for sequence, structure, property and function analysis of proteins. *Genome Informatics*, **10**, 229-230 (1999)

② J. An, H. Wako, and A. Sarai, Analysis of structural motifs of proteins in terms of sets of codes representing local structures. *Mol. Biol.* **35,** 1056-1062 (2001)

③ Hiroshi Wako, Jianghong An, and Akinori Sarai, Environment-dependent and position-specific frequencies of amino acid occurrences in α-helices. submitted.

④ 疎水コア

⑤ より大きな局所構造の同定

⑥ Hiroshi Wako and Takahisa Yamato, Novel method to detect a motif of local structures in different protein conformations. *Prot. Eng*. **11**, 981–990 (1998).

Construction of an Integrated Environment for Sequence, Structure, Property and Function Analysis of Proteins

Jianghong An [1] ajh@rtc.riken.go.jp
Takao Nakama [2] nakama@rtc.riken.go.jp
Yasushi Kubota [2] kubota@rtc.riken.go.jp
Hiroshi Wako [3] wako@mn.waseda.ac.jp
Akinori Sarai [1] sarai@rtc.riken.go.jp

[1] Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan
[2] Advanced Technology Institute Inc., 3-23-15 Jinbo, Kanda, Tokyo 101, Japan.
[3] School of Social Science, Waseda University, Shinjuku-ku, Tokyo 169-8050, Japan

1 Introduction

One of the most important goals in molecular biology is to elucidate the relationship among sequence, structure, function and properties of biomolecules. Such knowledge would enable us to design the modifications of biomolecules for particular functions, and drugs to modify the function and property of biomolecules. Now, the number of entries in the Protein Data Bank (PDB) is over 10,000. These prized structural data should be used to understand the molecular mechanism of structural integrity and stability of biomolecules. The functionally important sites such as active sites in enzyme and ligand binding sites tend to be conserved among a family of proteins. The conserved amino acid sequences are called motif, and many motifs have been known so far. The physico-chemical properties of biomolecules are studied by various biophysical and biochemical methods. The structure, function and property of biomolecules are often closely related, but it is usually difficult to infer the relation from individual data. Thus, if researchers are interested in the structure of particular molecules and its relationship with function and physico-chemical properties, they usually need to examine several databases and literatures to obtain the information of their interest. On the other hand, structure comparison is one of the most important and interesting subject of bioinformatics because it plays a key role in structure classification, structure search, motif detection, function prediction, and so on.

It would be useful to have an integrated environment where one can examine the relationship among sequence, structure, function and property of biomolecules based on databases and a lot of search, analysis and visualization tools.

2 An Integrated Environment for Biomolecule Information

2.1 3DinSight: an integrated database

We have developed an integrated database of structure, function and property of biomolecules, called 3DinSight [1, 2], by focusing on the following points:

(1) Integrate PDB's structure, PROSITE's motif, PMD's Mutations and a lot of data of amino acid property into a relational database. The motifs and mutations are mapped to the 3D structures;
(2) Provide strong and flexible programs to search the database by keywords, sequences, motifs, and so on, which are very difficult for the original flat-formatted databases;
(3) Provide a World-Wide-Web (WWW) interface so that researchers around the world can access to the data and can carry out searches easily;

(4) Visualize the relationship among structure, functional sites and property automatically in 3D space, together with the link to associated document information.

2.2 Towards an integrated environment: new databases and tools

We are developing following new databases and tools, which are integrated to 3DinSight:

1) Thermodynamic Database for Proteins and Mutants(ProTherm)[3], which is a collection of various thermodynamic data such as Gibbs free energy, enthalpy, heat capacity, and so on of proteins and mutants. now over 5,000 entries are loaded in the database. This database and search tool will help researchers in studying the mechanism of stability of proteins and mutants.
2) Protein-Nucleic Acid Recognition Database, it consists of

   2.1) Protein-Nucleic Acid Complex Database, which is a collection of structural data of protein-nucleic acid complex. The database enables users to examine sequence-dependent DNA conformation in a form of table or graph. We plan to implement information on the conformation changes of protein upon complexation, and create interface to extract detailed information about base-amino acid interactions from the complex structure.
   2.2) Database of Base-Amino Acid Interactions, which collects pairs of atoms between bases and amino acids within 4 angstrom into a database table. User can specify residue names (base and amino acid), atom types (they can be checked by clicking "Atom Name") and side-chain/backbone to search the database. After the search, all the atom pairs with distance values will be displayed and all the atom pairs will be highlighted in the complex structure if you want to show the image. Thus, users can examine the specific interactions between base and amino acids in each structure.
   2.3) Thermodynamic Database for Protein-Nucleic Acid Interactions, which collects various thermodynamic data on interaction between proteins and nucleic acids;
   2.4) Tools for the predictions of binding sites and target genes of transcription factors.

   These databases and search tool will provide users with insight into the mechanism of protein-nucleic acid recognition from various aspects.
3) Protein-Ligand Database, which collects all ligands and the binding information with proteins from PDB. The ligand can be search by name, formula, structure and binding conditions.
4) Structure analysis tools of proteins, which are based on a novel method called Delaunay tessellation[4]. The interior space of the protein can be uniquely divided into Delaunay tetrahedra whose vertices are the $C\alpha$ atom positions. Then one unique code can be assigned to each tetrahedra by the vertex residues and four surrounding tetrahedron. Because the structure is represented in a string of digits, more easily and rapidly programs of structure analysis tools can be developed. The tools include structure classification, 3D structure searching of proteins, motif detection, and so on.

References

[1] An, J., Nakama, T., Kubota, Y., and Sarai, A., 3DinSight: an integrated relational database and search tool for structure, function and property of biomolecules, *Bioinformatics*, 14:188–195, 1998.

[2] Nakama, T., An, J., Kubota, Y., and Sarai, A., Visualization of functional sites on protein structures by virtual reality modeling language, *Bioimages*, 5:59–64, 1997.

[3] Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., and Sarai, A., ProTherm: thermodynamic database for proteins and mutants, *Nucleic Acids Res.*, 27:286–288, 1999.

[4] Wako, H. and Yamato, T., Novel method to detect a motif of local structures in different protein conformations, *Protein Eng.*, 11:981–990, 1998.

Molecular Biology, Vol. 35, No. 6, 2001, pp. 905–910. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 6, 2001, pp. 1056–1062.

UDC 577.32.53

Analysis of Protein Structural Motifs in Terms of Sets of Codes Representing Local Structures

J. An[1], H. Wako[2], and A. Sarai[1]

[1] *Tsukuba Institute, The Institute of Physical & Chemical Research (RIKEN), Tsukuba, Ibaraki 305-0074, Japan; E-mail: sarai@rtc.riken.go.jp*

[2] *School of Social Sciences, Waseda University, Tokyo 169-8050, Japan*

Received July 19, 2001

Abstract—An amino acid sequence pattern conserved among a family of proteins is called motif. It is usually related to the specific function of the family. On the other hand, functions of proteins are realized through their 3D structures. Specific local structures, called structural motifs, are considered as related to their functions. However, searching for common structural motifs in different proteins is much more difficult than for common sequence motifs. We are attempting in this study to convert the information about the structural motifs into a set of one-dimensional digital strings, i.e., a set of codes, to compare them more easily by computer and to investigate their relationship to functions more quantitatively. By applying the Delaunay tessellation to a 3D structure of a protein, we can assign each local structure to a unique code that is defined so as to reflect its structural feature. Since a structural motif is defined as a set of the local structures in this paper, the structural motif is represented by a set of the codes. In order to examine the ability of the set of the codes to distinguish differences among the sets of local structures with a given PROSITE pattern that contain both true and false positives, we clustered them by introducing a similarity measure among the set of the codes. The obtained clustering shows a good agreement with other results by direct structural comparison methods such as a superposition method. The structural motifs in homologous proteins are also properly clustered according to their sources. These results suggest that the structural motifs can be well characterized by these sets of the codes, and that the method can be utilized in comparing structural motifs and relating them with function.

Key words: motif, three-dimensional structure, Delaunay tessellation

INTRODUCTION

Functions of proteins are usually derived from some specific sites, such as active sites in enzyme and ligand binding sites. The three-dimensional (3D) structures of such sites (structural motifs) are one of the most important factors for understanding the function of the proteins. However, most databases, such as PROSITE [1], BLOCKS [2], and PRINT [3], are collecting the motifs by analyses of amino acid sequences, i.e., define the motifs as sequence patterns or profiles. The sequence pattern can be readily used for searching the same motif in other protein sequences. However, if the sequence pattern is short or ambiguous, many false positives, i.e., incorrect proteins with different functions are picked up together with true positives, because a significance of differences in the sequences is underestimated. On the other hand, if the sequence pattern is too specific or strict, many false negatives occur, i.e., many proteins with the desired function are missing in detected proteins, because even slight differences in the sequence from the given sequence pattern are not allowed. As far as the sequence information alone is used for finding motifs, we cannot overcome the limitations in minimizing these false positives and false negatives. One of the reasons for such limitations are that a functional site in a protein usually consists of several regions which are separated along the polypeptide chain, but close to each other in space. In this sense the structural information must help to improve this situation.

Nowadays, a large number of 3D structures of proteins are known. The list of entries in the Protein Data Bank (PDB) [4, 5]; exceeds 15,000 and is expected to increase drastically owing to the progress of structural genomics in near future. Although it is well known that the structural information is useful and required, it is more difficult to derive common features in the 3D structures from different proteins than those in the sequences. The difficulty mainly comes from the dimensionalities of the sequence and structure (one and three dimensions, respectively). Therefore, it is useful to develop a simple way to characterize specific 3D local structures, i.e., structural motifs, in a one-dimensional description. Previously, one of the authors has developed the method to represent protein local structures by a digital string, i.e., code [6] using the Delaunay tessellation. It has been shown that the method can represent characteristic features of the

0026-8933/01/3506-0905$25.00

local structures properly; for example, it can classify the secondary structures more closely with respect to the residues surrounding them.

In this study we are interested in structural motifs that are larger in size than those considered by Wako and Yamato (local structure represented by one code) and usually consist of more than one code. Consequently, it is necessary to introduce some similarity measure to compare the structural motifs in terms of those codes. We introduced a simple measure of distance between the set of codes to evaluate similarity between structural motifs. The similarity measure was, then, applied to the structural motifs to cluster them (in this paper, the term *local structure* is used for the local structure represented by one code defined by Wako and Yamato; the term *structural motif* is used for a set of the local structures containing a given PROSITE sequence pattern, irrespective of whether it is true or false positive of the pattern). Two points were examined. Firstly, can the false positives be distinguished from the true ones? The structural motifs of false positives should be different from those of the true positive proteins. Secondly, can the structural motifs of the true positives be clustered in some meaningful manner? Their structures should resemble each other, but may differ to some extent. If the clustering provides some meaningful results with respect to functions, the sets of codes assigned to the structural motifs can be said to bear useful information about their 3D structures to distinguish protein functions.

ALGORITHM FOR CODING LOCAL STRUCTURES

The program to code 3D local structures is based on the method called Delaunay tessellation [6], which uniquely divides the interior space of a protein into nonoverlapping volume elements named Delaunay tetrahedrons. Since each residue is represented only by a Cα atom in that method, the vertices of Delaunay tetrahedrons are the Cα atom positions. One unique code (string of digits) is assigned to each tetrahedron according to the four vertex residues on the relevant tetrahedron and at most four more vertex residues on the tetrahedrons surrounding it. Finally, a set of the codes assigned to these five tetrahedrons is assigned to the relevant tetrahedron as its own code. The coding rules are briefly summarized as follows (see [6] for more detail);

Step 1: Represent each residue in a given protein only by its Cα atom.

Step 2: Apply the Delaunay tessellation to the 3D structure of the protein.

Step 3: Consider a Delaunay tetrahedron, say, T_0.

Step 3.1: Number four vertices of T_0 with the digit 1 to 4 according to the increasing order of their residue numbers.

Step 3.2: Number the vertex residues of the surrounding tetrahedrons of T_0 that share one facet with T_0, say, T_1, T_2, T_3, and T_4, with the digits 5 to 8, respectively, irrespective of their residue number (the residues on the shared facet with T_0 are excluded, because they are numbered in Step 3.1).

Step 3.3: Rearrange these digits in increasing order of the corresponding residue numbers.

Step 3.4: Assign these four to eight digits (the number of digits depends on the number of tetrahedrons surrounding T_0) to T_0 as the code.

Step 4: Assign the code to every tetrahedron in the given protein.

Step 5: Re-assign a set of five codes of T_0, T_1, T_2, T_3, and T_4 to T_0 as its own code.

In this paper a local structure is defined as a set of residues related to such five tetrahedrons. These residues are not necessarily consecutive along the polypeptide chain. In most cases they are localized in two or three separate regions of the chain. But, of course, they are close to each other in space.

ANALYSIS OF STRUCTURAL MOTIFS

In order to examine the ability of our codes for characterizing 3D structures of proteins, we coded all the protein structures in PDB and analyzed them. It has been already shown that these codes are useful to represent local structure features, that is, the local structures with the same code have 3D structures similar to each other [6]. We are now studying local structures larger in size, i.e., structural motifs, than those studied in that paper. Since we will focus our attention on the structural motif related to the PROSITE sequence pattern in this paper, the term *structural motif* is used in the more restricted meaning as follows.

For a given protein, consider the residues consecutively connected along the polypeptide chain and containing a given PROSITE sequence pattern completely. The first and last residues must be the ones given in the PROSITE pattern. Some of the tetrahedrons obtained by the Delaunay tessellation of the protein contain these residues. We define the structural motif as a set of residues that are contained in such tetrahedrons. Therefore, it should be noted that many residues separated from the given PROSITE pattern along the chain, but near to it in space are included in the structural motif defined here.

The structural motif is represented by a set of codes for these tetrahedrons. Our interest is whether or not the differences between the structural motifs

can be distinguished in terms of these codes. For this purpose, we introduce at first the similarity measure between the sets of codes.

Measure of Distance between Code Sets

In order to evaluate similarity between structural motifs we define a distance between two code sets corresponding to the structural motifs of two proteins in the following manner.

Consider two code sets $A = \{A_1, A_2, \ldots, A_m\}$ and $B = \{B_1, B_2, \ldots, B_n\}$, where $A_i = \{a_{i0}, a_{i1}, a_{i2}, a_{i3}, a_{i4}\}$ $(i = 1, \ldots, m)$ and $B_i = \{b_{i0}, b_{i1}, b_{i2}, b_{i3}, b_{i4}\}$ $(i = 1, \ldots, n)$. n and m are the total numbers of tetrahedrons in the structural motifs, respectively. The distance between the two code sets A and B is defined as follows:

$$Distance(A, B) = 1 - \frac{(Similarity(A \longrightarrow B) + Similarity(B \longrightarrow A))}{m + n},$$

$$Similarity(A \longrightarrow B) = \sum_{i=1}^{m}\left(\max_{j=1}^{n}\left(\sum_{k=0}^{4}(S_{ij,k} * V_k)\right)\right),$$

$$Similarity(B \longrightarrow A) = \sum_{i=1}^{n}\left(\max_{j=1}^{m}\left(\sum_{k=0}^{4}(S_{ij,k} * V_k)\right)\right);$$

$S_{ij,k} = 1$ if $a_{ik} = b_{jk}$ or $b_{ik} = a_{jk}$, otherwise 0. V_k $(0 < V_k < 1, \sum_{k=0}^{4} V_k = 1)$ is a weighting factor. They can be set arbitrarily, but we set them so as to emphasize the significance of the code a_{i0} or b_{i0} for the central tetrahedron; $V_0 = 0.6$ and $V_1 = V_2 = V_3 = V_4 = 0.1$.

Clustering Structural Motifs

The distance between two code sets, i.e., between two structural motifs in different proteins, defined above was applied to protein sets with a given PROSITE motif to cluster them. These structures are expected to contain some pairs resembling with each other very much or moderately (e.g., pairs of the true positives of the PROSITE motif) and others different from each other (e.g., pairs of the true and false positives), even if they have the same PROSITE sequence motif. We examined the ability of the code sets and the similarity measure among them defined above to distinguish such similarity and difference.

Figure 1 shows an example of the clustering analysis for the structural motifs with 19-residue pattern, Lactalbumin_lysozyme, C–x(3)–C–x(2)–[LMF]–x(3)–[DEN]–[LI]–x(5)–C (PS00128), which is involved in sugar metabolism (the structural motifs are denoted by the PDB entry codes of the proteins that contain the structural motifs). In this example, since all the structural motifs are true positive, their 3D structures are very similar to each other, except for 1hwa (lysozyme; determined by NMR; minimized average structure), which has remarkably different from the others. Kasuya and Thornton [7] compared these structures with respect to the root-mean-square differences and obtained the same result.

The structural motifs except for 1hwa are further divided into the two clusters, α-lactalbumin and lysozyme, in Fig. 1. Although few proteins belong to wrong clusters, the clustering seems to be carried out properly in general. In the lysozyme cluster the two and one major clusters from human and chicken, respectively, are well recognized. The lysozymes from other species of birds, such as quail, guinea fowl, turkey, and pheasant, seem also to be clustered properly. As for the lysozymes from mammalian except for human, lysozyme from Australian echidna is contained in the α-lactalbumin cluster and that from horse is located in the most outer group of the lysozyme cluster.

Figure 2 shows another example of Peroxidase_1 with a 11-residue pattern [DET]–[LIVMTA]–x(2)–[LIVM]–[LIVMSTAG]–[SAG]–[LIVMSTAG]–H–[STA]–[LIVMFY] (PS00435), which is involved in various biosynthetic and degradation functions. The clustering divided the proteins with this sequence motif into the two major groups: Group 1 includes the false positives for this PROSITE motif, cathepsin B from rat (1mirA, 1mirB, 1theA, 1theB, 1cpjA, 1cpjB, 1cteA, 1cteB). Group 2, the cluster of the remaining proteins, can be divided further into several subgroups. The groups of CO dehydrogenase (1qj2l and 1qj2C) and glucose oxidase (1gpeB and 1gpeA) are false positive. The groups of myeloperoxidase from dog (1mypD, 1mypC,) and human (1mhlD, 1mhlC) located in the outer cluster are true positives, but have different 3D structures from other peroxidases, as shown by Kasuya and Thornton [7]. The remaining are peroxidases, and true positives for this PROSITE motif. Their major subgroups, cytochrome *c* peroxidase from baker's yeast, lignin peroxidase from white rot basidiomycete, and peroxidase from *Arthromyces ramosus* and from horseradish, are well clustered.

DISCUSSION

Although motifs of proteins are usually described by sequence patterns, they are only parts of the structural motifs. It is well recognized that the local structures should be taken into consideration to understand the function of the protein. The coding method applying the Delaunay tessellation has been originally developed for characterizing local structures, and shown that it is useful for that purpose. On the other hand, in this paper, we consider local 3D structures larger in size than those in the original paper and examine if the sets of codes representing them are use-

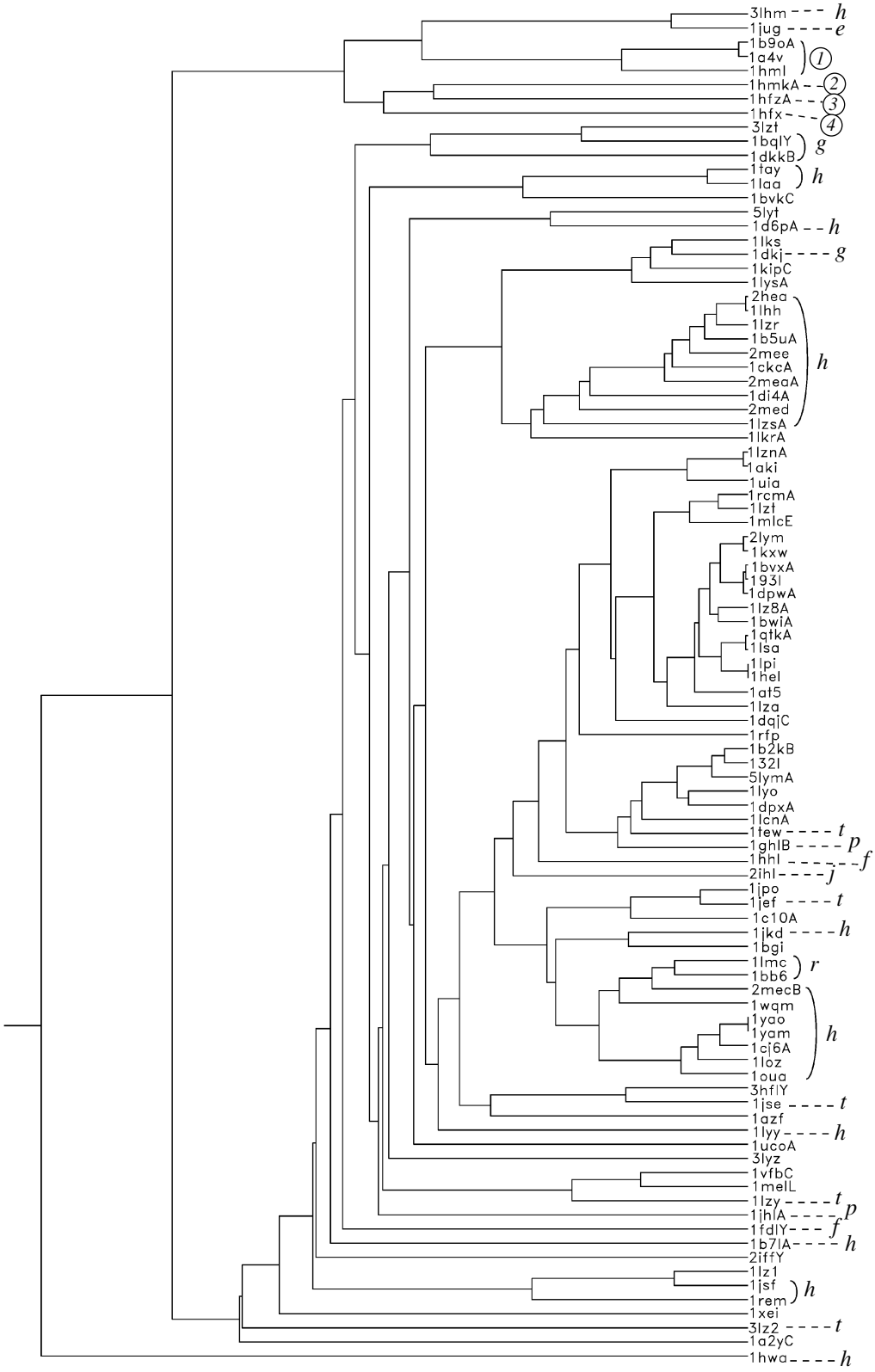


Fig. 1. Clustering of the proteins with Lactalbumin_lysozyme PROSITE motif, C–x(3)–C–x(2)–[LMF]–x(3)–[DEN]–[LI]–x(5)–C (PS00128). ①, ②, ③, and ④ are human, goat, bovine, and Guinea pig α-lactalbumins, respectively. The remaining are lysozyme from human (*h*), chicken (without symbols), Japanese quail (*j*), quail (*q*), Guinea fowl (*f*), turkey (*t*), pheasant (*p*), Australian echidna (*e*), and rainbow trout (*r*).

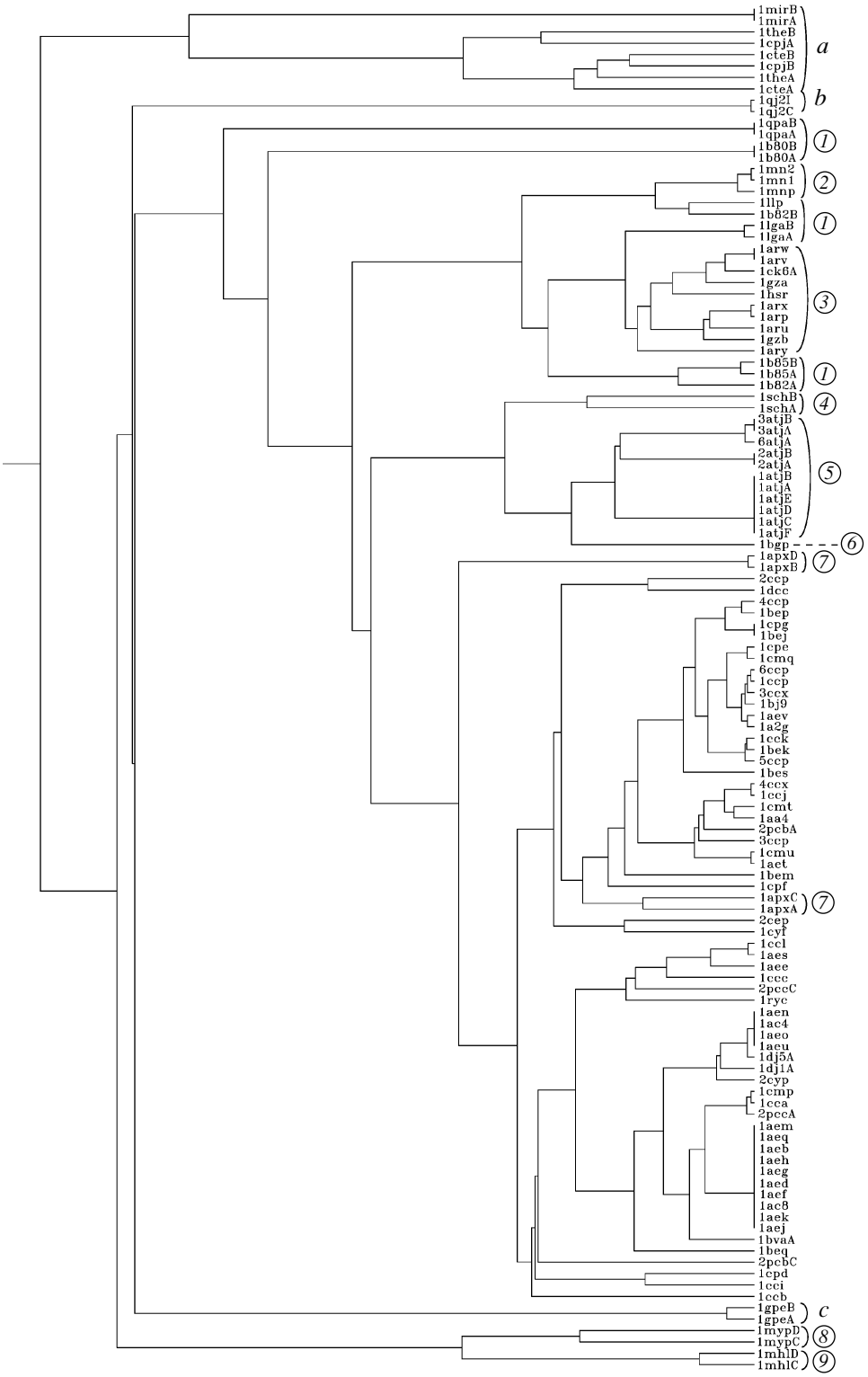


Fig. 2. Clustering of the proteins with Peroxidase_1 PROSITE motif, [DET]–[LIVMTA]–x(2)–[LIVM]–[LIVMSTAG]–[SAG]–[LIVMSTAG]–H–[STA]–[LIVMFY] (PS00435). Lignin peroxidase from white rot basidiomycete (①); manganese peroxidase from a basidiomycete (②); peroxidase from *Arthromyces ramosus* (③), from peanut (④), and from horseradish (⑤); peroxidase 1 from barley (⑥); ascorbate peroxidase from pea (⑦); cytochrome *c* peroxidase from baker's yeast (without symbols); myeloperoxidase from dog (⑧) and from human (⑨); cathepsin B from rat (*a*); CO dehydrogenase from *Pseudomonas carboxydovarans* (*b*); glucose oxidase from *Penicillium amagasakiense* (*c*).

ful to distinguish the differences between the structures. The above results indicate that the method presented here have the good ability to properly cluster the mixed set of structures with the same PROSITE sequence pattern, i.e., not only true and false positives, but also the homologous proteins. This is because the sets of the codes include the information not only about the secondary structure contents, but also about the network (interactions) among them.

One of the problems in this method is that there are an enormous number of codes. It means that the codes are sensitive to the local structure differences. The sensitivity of the codes, however, has both merit and demerit. The merit is that the codes can detect the small differences in the structures. The demerit is that similar structures (structures that seem similar for us) cannot be detected as a similar structure in some cases, because they are assigned to different codes. The similarity measure between the code sets were introduced in this context, and seems to work well not only to distinguish the differences between the larger local 3D structures, but also to detect the similarity, according to the above results.

It is also important to represent local structures that should be called structural motifs by these codes properly, because the codes are a very easy way to treat in computer. In this sense the two local structures consisting of exactly the same code sets are important, because they must be very similar to each other (in [6], the local structures consisting of only one code were considered). If such local structures are found, it means that the structural alignment of the two local structures are performed simultaneously, because the correspondence between the residues can be obtained from the correspondence between the vertices of the corresponding tetrahedrons. However, the sensitivity of the code to the local structure differences discussed above is a problem at present to use this coding method more efficiently. The study is now in progress.

The coding method presented here can detect structural motifs in new proteins, if the structural motifs of all the proteins in PDB are more properly coded. This will provide significant implication to structural genomics and functional genomics, where functions of the new proteins required to be annotated.

REFERENCES

1. Bairoch, A., Bucher, P., and Hofmann, K., *Nucl. Acid Res.*, 1997, vol. 25, pp. 217–221.
2. Henikoff, S. and Henikoff, J.G., *Genomics*, 1994, vol. 19, pp. 97–107.
3. Attwood, T.K., Beck, M.E., Flower, D.R., Scordis, P., and Selley, J.N., *Nucl. Acid Res.*, 1998, vol. 26, pp. 304–308.
4. Bernstein, F.C., Koetzle, T.F., Williams, G.J., *et al.*, *J. Mol. Biol.*, 1977, vol. 112, pp. 535–542.
5. Berman, H.M., Westbrook, J., Feng, Z., *et al.*, *Nucl. Acid Res.*, 2000, vol. 28, pp. 235–242.
6. Wako, H. and Yamato, T., *Protein Eng.*, 1998, vol. 11, pp. 981–990.
7. Kasuya, A. and Thornton, J.M., *J. Mol. Biol.*, 1999, vol. 286, pp. 1673–1691.

Running title: Frequency of amino acid occurrence in α-helix.

Environment-dependent and position-specific frequencies of amino acid occurrences in α-helices

Hiroshi Wako[1], Jianghong An[2] and Akinori Sarai[2]

[1]School of Social Sciences, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan

[2]Tsukuba Institute, The Institute of Physical & Chemical Research (RIKEN)

3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Keywords:

Delaunay tessellation / local structure motif / amino acid substitution / principal component analysis / helix capping

Abstract

Using the method to define local structure motifs of proteins by the Delaunay tessellation proposed by Wako and Yamato (Protein Eng. (1998) 11:981-990) we analyzed environment-dependent and position-specific frequencies of amino-acid occurrences in α-helices. In that method the three-dimensional structure of a protein molecule is uniquely divided up into non-overlapping Delaunay tetrahedrons, each vertex of which is occupied by one of the comprising residues. Then a code is assigned to each tetrahedron so as to characterize a local structure containing it. To the tetrahedrons located in the interior of α-helices 36 kinds of codes are assigned. The differences in the codes reflect the existence and absence of four surrounding residues around the relevant region of the α-helix. In other words the environment of the α-helix can be differentiated by these codes. Accordingly we analyzed the frequencies of amino acid occurrences on each vertex of the tetrahedrons for each of these codes. Such data provide information about possible amino acid substitutions specific to a vertex position (i.e., a position in the α-helix) for a given code (i.e.,

environment around the α-helix). Furthermore, the principal component analysis was carried out to reveal general features of the amino acid occurrences in the α-helices. Relating to these results such frequencies at N- and C-terminals of α-helix are also discussed.

INTRODUCTION

An α-helix is the most easily recognizable local structure in protein structures owing to its regularity, and an important structural element for protein folding. Many efforts have been made for predicting the locations of the α-helices in a protein from its amino acid sequence. For such a purpose many experimental and theoretical data on determination of propensities of amino acid residues to occur in the α-helices have been provided (Chou and Fasman, 1974; Levitt, 1978; Wako et al., 1983; Williams, et al., 1987; Altmann et al., 1990; Blundell and Zhu, 1995; Chakrabartty and Baldwin, 1995; Kumar and Bansal, 1998a, b). Although some amino acid residues do demonstrate preference for the α-helix, it is only marginal. For example, the most helix-preferring amino acid Glu occurs in α-helices only 59% more frequently than random. Even Gly and Pro residues, which are not stereo chemically compatible with the α-helical conformation, are found in α-helices about 40% as often as random (Creighton, 1993).

The α-helix is characterized by consecutive main chain (*i*, *i*-4) hydrogen bonds between an amide hydrogen and a carbonyl oxygen. This pattern is, however, interrupted at N- and C-termini because, upon termination, no turn of helix follows to provide additional hydrogen bond patterns. Such end effects are substantial for the amino acid preference (Presta and Rose, 1988; Aurora et al., 1994; Blundell and Zhu, 1995; Aurora and Rose, 1998).

The position-specific preferences of particular amino acid residues are found especially at the terminal regions of the α-helices. For example, acidic Asp and Glu residues predominate at the N-terminus, and basic Lys, Arg, and His residues at the C-terminus, as a result of favorable interactions of their charges with the helix dipole. Although Pro residues rarely occur in the interior of α-helices, where their unusual backbones interrupt the α-helix and cause it to kink, they occur frequently at the N-terminal first turn of the α-helix, where their particular geometry fits well. Asn, Asp, Ser, and Thr residues often occur in the first turn, where their side chains tend to form a hydrogen bond to the backbone of the third residue father along. On the other hand, Gly residues occur at the carboxyl end of about a third of all α-helices, where the more flexible backbone of this residue tends to disrupt the α-helix by tending toward 3_{10} type conformation (Creighton, 1993).

For the interior of the α-helices the amino acid preference is provided by the normalized frequency (the fraction of an amino acid residue occurring in the α-helices

divided by its fraction in all of the proteins). This property indicates the propensity of each amino acid for forming the α-helix. Since the α-helix has a very regular structure in the interior, it is usual not to pay much attention to position specificity of the amino acid preference in such a region.

It is well known, however, that the amino acid occurrences vary with the locations in protein folding, the geometries, and the lengths of the α-helices, even in the interior region (Blundell and Zhu, 1995; Zhu and Blundell, 1996; Kumar and Bansal, 1998a, b). As a matter of fact, many α-helices are amphipathic, in which they have predominantly non-polar residues along one side of the α-helical cylinder and polar residues along another (Schiffer and Edmundson, 1968). The two sides are usually referred to as hydrophobic and hydrophilic ones, respectively. However, this definition is not adequate to analyze the position-specific amino acid preferences in the interior of the α-helices, because the two sides are defined owing to the richness in hydrophobic and hydrophilic residues, respectively. Alternatively the position-specific amino acid propensities in the interior of the α-helices were obtained by using solvent inaccessibility of the residues in them (Blundell and Zhu, 1995; Zhu and Blundell, 1996).

In this paper we are also interested in the environment-dependent and position-specific frequencies of amino acid occurrence in the α-helices, but our approach to such a problem is quite different from those described above. We used the method to define local structure motifs of proteins by the Delaunay tessellation proposed by Wako and Yamato (1998). In that method the three-dimensional structure of a protein molecule is divided up into non-overlapping Delaunay tetrahedrons, on each vertex of which one of the comprising residues is located. The Delaunay tessellation can be performed uniquely for a given protein. Then, in order to characterize a local structure constructed by the residues on the vertices of several tetrahedrons spatially neighboring each other, a code (Delaunay code) is assigned to each tetrahedron according to the rules proposed by Wako and Yamato, while they were slightly modified from the original ones in this paper.

We focus our attention on the interior of the α-helix, because we want to demonstrate the ability of the Delaunay code to differentiate the interior positions of the α-helix in spite of its regularity. Although the N- and C-terminal regions of the α-helix are also interesting, we will discuss them briefly, because many researches have been carried out on the terminal regions. It should be also emphasized here that the Delaunay code was devised to analyze not only the α-helix, but also various local motifs. Although we confined ourselves to analyzing the Delaunay codes just related to the α-helix in this paper, the same approach is applicable to other codes.

As described below, the Delaunay code assigned to the tetrahedron located in the interior of an α-helix is given as FHABCDEG: FHABCDEG: x: y: FHABCDEG, where 36 kinds of codes are possible for x: y. The differences in the codes for the α-helices arising

from x: y reflect the situation whether or not some residues are located around the relevant α-helices. In other words the environment of the α-helix can be differentiated by these 36 kinds of codes. Accordingly, we examined the frequency of amino acid occurrence on a given position of a given code (i.e., a given vertex of the Delaunay tetrahedron in a given environment) in the interior of the α-helix. It should be noted that our interest is the analysis of the amino acid frequencies of occurrences with respect to its environment rather than their preferences for the formation of the α-helix to other conformational states such as an extended structure, a turn, and random coil.

METHODS

Local structure code

We review the Delaunay tessellation and code assignment to the tetrahedrons briefly at first. In this paper the code assignment rules are slightly changed from those defined in Wako and Yamato (1998).

The three-dimensional structure of a protein molecules is represented as a set of Cα atoms. In the previous paper, if a protein has more than one chain, they have to be treated independently. In this study, however, we changed the code assignment rules so that two or more chains can be treated together. This modification makes it possible to assign a code to the tetrahedron consisting of residues in the interface regions of the two subunits.

By the Delaunay tessellation the interior space of the protein is divided up into non-overlapping Delaunay tetrahedrons whose vertices are the Cα atoms. Some edges of the tetrahedrons are virtual bonds connecting adjacent Cα atoms along the polypeptide chain and others connect two non-adjacent Cα atoms near each other in space.

Consider a Delaunay tetrahedron T_0. The amino acid residue number at the four vertices of T_0 are denoted $v_1(T_0)$, $v_2(T_0)$, $v_3(T_0)$, and $v_4(T_0)$. Here we can require the suffixes to satisfy $v_1(T_0)<v_2(T_0)<v_3(T_0)<v_4(T_0)$ without loosing generality.

We also consider the tetrahedrons neighboring T_0, which share one of the facets (triangle faces) of T_0. At most four tetrahedrons, T_5, T_6, T_7, and T_8, are possible to exist, although they do not always do so. If any, the four tetrahedrons, T_5, T_6, T_7, and T_8, are defined as sets of vertex residues, $\{v_2, v_3, v_4, v_5\}$, $\{v_1, v_3, v_4, v_6\}$, $\{v_1, v_2, v_4, v_7\}$, and $\{v_1, v_2, v_3, v_8\}$, respectively.

Furthermore, we take into account the tetrahedrons neighboring T_5 to T_8. At most 12 more tetrahedrons are possible to exist. They are defined as $T_9=\{v_3, v_4, v_5, v_9\}$, $T_{10}=\{v_2, v_4, v_5, v_{10}\}$, $T_{11}=\{v_2, v_3, v_5, v_{11}\}$, $T_{12}=\{v_3, v_4, v_6, v_{12}\}$, $T_{13}=\{v_1, v_4, v_6, v_{13}\}$, $T_{14}=\{v_1, v_3, v_6, v_{14}\}$, $T_{15}=\{v_2, v_4, v_7, v_{15}\}$, $T_{16}=\{v_1, v_4, v_7, v_{16}\}$, $T_{17}=\{v_1, v_2, v_7, v_{17}\}$, $T_{18}=\{v_2, v_3, v_8, v_{18}\}$, $T_{19}=\{v_1, v_3, v_8, v_{19}\}$, and $T_{20}=\{v_1, v_2, v_8, v_{20}\}$. In this way we can assign v_1 to v_{20} to some residues in the given protein uniquely for each tetrahedron.

We consider local structures consisting of these 20 residues. In actual fact, however,

most local structures consist of less than 20 residues, because (1) some tetrahedrons do not exist, and (2) some vertices coincide with other vertices.

Then, we assign the two kinds of codes, called ST and NNT codes in the previous paper, to each tetrahedron.

The ST code is defined by arranging the vertex residue number v_1 to v_8 in increasing order. The ST code is a string of the suffices of v's in such order, and assigned to the tetrahedron T_0. For example, if $v_8<v_7<v_1<v_2<v_5<v_6<v_3<v_4$, then the ST code of the tetrahedron T_0 is 87125634. In this paper, however, we modified this code assignment rule with respect to the three points. For the modification, first of all, the figures 1 to 8 are converted to alphabets A to H, respectively. Both the upper and lower cases are allowed in following the rules described below.

The first point of the modification is to distinguish the residues in separate regions along the polypeptide chain(s). For the above example, assume that the eight residues are localized into three regions, (v_8, v_7, v_1, v_2), (v_5, v_6), and (v_3, v_4), in the chain. We can also assume another case consisting of two regions such as (v_8, v_7, v_1, v_2,v_5) and (v_6, v_3, v_4). Here, the two residues are regarded to be separated, if they are separated by more than 3 residues along the chain (or they are in different chains). In the examples the differences between v_5 and v_2 and between v_3 and v_6 in the first case, and between v_5 and v_6 in the second case, should be greater than 3. The two cases cannot be distinguished by the original code assignment rules. By using the lower and upper cases of the alphabets alternately for the neighboring regions, 87125634 is, for example, converted into HGABefCD and HGABEfcd for the first and second cases, respectively. We use the alphabets rather than the figures, just because we can use the upper and lower cases.

In addition to the above example, we introduce two more examples to explain the another new rule introduced into the code assignment. One is $v_5<v_6<v_8<v_7<v_1<v_2<v_3<v_4$ comprising three regions (v_5, v_6), (v_8, v_7, v_1, v_2), and (v_3, v_4). Another is $v_8<v_7<v_1<v_2<v_3<v_4<v_5<v_6$ comprising three regions (v_8, v_7, v_1, v_2), (v_3, v_4), and (v_5, v_6). The codes EFhgabCD and HGABcdEF, are assigned to these tetrahedrons, respectively, according to the rules described so far. While in the previous paper we regarded these three codes HGABefCD, EFhgabCD, and HGABcdEF as different local structures, we intended to regard them as the same local structure in this paper. For this purpose we add a new rule: if there are more than one region in a code, they are arranged in the order of their sizes (the numbers of residues contained in the regions); if the sizes are equal to each other, they are arranged in the alphabetical order. In the examples any of the three codes are represented by HGABcdEF. Owing to this modification we can not only unify several codes into one code, but also assign the code to a tetrahedron lying across the two chains (such assignment is impossible according to the original rules, because a comparison of residue numbers in different chains makes no sense).

The third point is that if the residue is missing at any of the vertices v_5 to v_8, alphabet x is used for such vacant vertices. In the previous paper such vertices are not included in the code. As a result any ST code consists of eight alphabets, while four to eight figures in the previous paper. This rule is introduced just for computational convenience.

After the ST codes are assigned all the tetrahedrons in the protein, then the NNT codes are re-assigned to them taking into account the surrounding tetrahedrons. That is, the NNT code for T_0 is defined as $c(T_0)$: $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$, where $c(T_i)$ is the ST code for the tetrahedron T_i. This procedure is the same as the previous one. Hereinafter, the NNT code is simply referred to as code.

Codes for α-helix

One of the typical codes of the interior of the α-helix is FHABCDEG: FHABCDEG: $c(T_6)$: $c(T_7)$: FHABCDEG (which corresponds to the code 68123457: 68123457: $c(T_6)$: $c(T_7)$: 68123457 in the previous paper). The correspondence between residues and vertex positions are given in Table 1 and Fig. 1. This type of codes is the most abundant in the proteins. There are 36 possible codes for $c(T_6)$: $c(T_7)$ as shown in Table 2. Both $c(T_6)$ and $c(T_7)$ have the string ABEHCD, followed by F/f, G/g, and/or x in various combinations.

Statistical analysis of amino acid occurrence

Let $f_{j,c,b}$ be the frequency of occurrence of amino acid j ($j = 1, 2, \cdots, 20$) at a given site b (a given vertex of the tetrahedron) of a given code c, and define $\mathbf{f}_{c,b} = (f_{1,c,b}, \cdots, f_{20,c,b})$. This is a key property in this study. For the interior of the α-helix $c = 1, 2, \cdots, 36$. As for the vertices we confine ourselves to the four vertices v_1 to v_4 in the central tetrahedron T_0 in the following statistics (i.e., $b = 1, 2, 3, 4$).

Accordingly there are 36×4=144 points for $\mathbf{f}_{c,b}$ in the 20-dimensional space. Their distribution can be characterized by the principal component analysis as follows.

Mean $< f_j >$ and standard deviation s_j for a given amino acid *j* are calculated over the 4 vertices of the 36 codes. Then normalized frequency $x_{j,c,b}$ is introduced as

$$x_{j,c,b} = \frac{f_{j,c,b} - < f_j >}{s_j} \qquad (1)$$

For the sake of convenience, these values are gathered up in a matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1,1} & \cdots & x_{20,1,1} \\ x_{1,1,2} & \cdots & x_{20,1,2} \\ \vdots & & \vdots \\ x_{1,m,4} & \cdots & x_{20,m,4} \end{bmatrix} \qquad (2)$$

A correlation between amino acid i and j, $C_{ij} = \frac{1}{4m}\sum_{b=1}^{4}\sum_{c=1}^{m} x_{i,c,n} x_{j,c,b}$, is expressed in a matrix form:

$$\mathbf{C} = \frac{1}{4m}\mathbf{X}^t\mathbf{X} \qquad (3)$$

where $\mathbf{X}^t$ is a transpose matrix of $\mathbf{X}$. In order to make a principal component analysis the eigenvalues and eigenvectors of matrix $\mathbf{C}$ are calculated:

$$\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k \qquad (4)$$

where eigenvector $\mathbf{u}_k$ satisfies the orthonormal condition such that $\mathbf{u}_i\mathbf{u}_j = 1$ if $i = j$, and $u_i u_j = 0$ if $i \neq j$. We assume $\lambda_1 > \lambda_2 > \cdots > \lambda_{20}$. Then $\mathbf{u}_k$ is a unit vector along the kth principal component axis. Since the data projected on the kth axis have the standard deviation $\sigma_{kj} = \sqrt{\lambda_k} u_{kj} s_j$, the following property is convenient to see the frequency of amino acid occurrence along the kth axis:

$$f_j^* = < f_j > \pm \sigma_{kj} \qquad (5)$$

where u_{kj} is the jth component of $\mathbf{u}_k$.

RESULTS

Structure data set

In this study we used the structure data set of 682 representative protein chains which have less than 25% homology with each other selected from Protein Data Bank (Bernstein, 1977; Hobohm, et al., 1994). If a entry of PDB contains two or more chains, the Delaunay tessellation and code assignment were carried out for the system including all the chains in the entry, even if only one of the chains is included in the representative structural data set. The tetrahedrons are classified into three categories: intra-chain (vertices v_1 to v_8 are in a representative chain), contact surface (v_1 to v_4 are in a representative chain, but at least one of v_5 to v_8 in another chain), and inter-chain (some of v_1 to v_4 are in a representative chain,

but at least one of them in another chain).

The numbers of tetrahedrons categorized into intra-chain and contact surface observed in the 682 representative protein chains for the codes related to the interior of the α-helix are shown in Table 2. No tetrahedron categorized into inter-chain is observed for these codes as a matter of course. In the statistical analyses described below only the intra-chain tetrahedrons contained in the representative chains were used.

The contact-surface tetrahedrons were analyzed separately. The results are also given below briefly, since the number of the tetrahedrons which belong to this category is too small to make a reliable statistical analysis.

Principal component analysis

At first we examine the position-independent properties. Mean frequency of amino acid occurrence and its standard deviation over vertices v_1 to v_4 of the 36 codes related to the interior of the α-helix are given in Fig. 2(a). Ala (12.8%) and Leu (12.2%) are the most abundant, and followed by Glu, Val, Lys, Ile, and Arg (6 to 8 %). On the other hand, the percentages for His, Trp, Cys, and Pro are less than 2 %. These results are well correlated generally with the other analyses about amino acid preferences for formation of the α-helix (Chou and Fasman, 1974; Williams et al., 1987), although the preference is usually defined as the fraction of residues of each amino acid that occur in α-helix, divided by the fraction of its random occurrence.

The standard deviations for Leu, Glu, Ala, and Lys (5.2, 4.6, 4.0, and 3.7 %, respectively) are larger than 3.0 %. The fact that the standard deviation of Leu is much larger than Ala in spite of their comparative mean values indicates that the occurrence of Leu is more biased (i.e., more dependent on environment of vertices) than Ala (i.e., more independent). In the same context, the occurrence of Glu and Lys is more biased than Val, Ile, and Arg.

The top four largest eigenvalues, λ_1 to λ_4, are 11.1, 2.7, 1.3, and 0.9, respectively. Their contribution to variance, i.e., $\lambda_k / 20$, are 55.6 %, 13.5 %, 6.7 %, and 4.3 %, respectively. The top two eigenvalues λ_1 and λ_2 contribute to variance as much as about 70 %.

f_j^* defined by Eq. 5 are plotted for the first and second principal components in Figs. 2b and 2c, respectively. The large negative values of σ_{kj} for Glu (-4.2 %) and Lys (-3.3 %) and the large positive one for Leu (4.7 %) in the first principal component (the line with solid square symbols in Fig. 2b; Although the negative and positive signs are interchangeable as shown by the lines with the solid square and open triangle symbols in

Fig. 2b, we refer them in this manner for convenience sake) indicate that they are characteristic amino acids to differentiate the two type of positions in the interior of the α-helix, i.e., hydrophilic ($\sigma_{kj} < 0$) and hydrophobic ($\sigma_{kj} > 0$) ones. Asp (-2.5 %), Arg (-1.9 %), and Gln (-1.8 %) have also relatively larger negative σ_{kj} values, and Ile (2.6 %), Val (2.5 %), Ala (1.9 %), and Phe (1.6 %) have relatively larger positive σ_{kj} values.

In the second principal component the large negative values of σ_{kj} are observed for Leu (-1.7 %), Lys (-1.1 %), and Arg (-0.8 %), and the large positive ones for Ala (3.0 %), Gly (1.3 %), and Ser (0.7 %) (the line with solid square symbols in Fig. 2c; The negative and positive signs are also interchangeable as described above). The second principal component indicates another kind of characteristics to differentiate the positions in the interior of the α-helix, i.e., the preference for the amino acids with larger ($\sigma_{kj} < 0$) and smaller ($\sigma_{kj} > 0$) sidechains. In actual fact those values for the amino acids with aromatic rings His, Tyr, Phe, and Trp are negative, and those for Pro, Val, Thr and Cys are positive, although their absolute values are much smaller than the above residues.

The correlation coefficient C_{ij} between the normalized variables x_i and x_j are given in Fig. 3. The large positive (negative) C_{ij} values indicate that if an amino acid *i* is favorable at a given vertex of a given code, another amino acid *j* is favorable (unfavorable) at this vertex. The amino acids can be divided into three groups; hydrophilic (Arg, Asp, Glu, Asn, Lys, and Gln), hydrophobic (Tyr, Cys, Trp, Met, Phe, Val, Ile, and Leu), and others (His, Ser, Thr, Gly, Pro, and Ala). The amino acids belonging to the first and second groups have very large positive C_{ij} values with other amino acids in the same group, and very large negative C_{ij} values with the amino acids in another group. The amino acids belonging to the last group can be divided further into two groups; (His, Ser, Thr, and Pro) and (Gly and Ala). Most of their $|C_{ij}|$ values with the other amino acids are, however, less than 0.5, differently from the first and second groups. It should be remarked that this group except His contains the amino acids with relatively smaller sidechains, compared with the first two groups.

Environment-dependent and position-specific frequency

Each of eight vertices corresponding to i-2 to i+5 for the 36 codes for the interior of the α-helix has a characteristic frequency patterns of amino acid occurrence $\mathbf{f}_{c,b}$ (see Table 1 for the correspondence between residue and vertex positions). In Fig. 4 the results are shown for the three kinds of codes, $c(T_6)$: $c(T_7)$ = ABEHCDfG: ABEHCDfG (a), ABEHCDfG: ABEHCDxx (b), ABEHCDxx: ABEHCDgx (c), as an illustration. As for the four surrounding residues around the α-helix, v_{12}, v_{13}, v_{16}, and v_{17}, all of them exist in Fig. 4a, v_{12} and v_{13} in Fig. 4b, and v_{17} in Fig. 4c.

In Fig. 4a hydrophobic amino acids are preferable at the vertices v_1 to v_6, since all the four surrounding residues exist. At the vertices v_9 and v_{18} both hydrophobic and hydrophilic amino acids occur with nearly equal frequencies. These vertices are relatively independent from the surrounding residues, and located on the opposite side of the α-helix cylinder against v_1 and v_4 (Fig. 1). It frequently occurs that when hydrophobic amino acids are preferable on one side, hydrophilic amino acids are preferable on another side. This fact may be reflected in the statistics for these vertices. It holds also in Figs. 4b and 4c.

For the interpretation of the results shown in Figs. 4b and 4c, we have to take into consideration the vertices composing the tetrahedrons T_{12}, T_{13}, T_{16}, and T_{17}; i.e., $T_{12}=\{v_3, v_4, v_6, v_{12}\}$, $T_{13}=\{v_1, v_4, v_6, v_{13}\}$, $T_{16}=\{v_1, v_4, v_7, v_{16}\}$, and $T_{17}=\{v_1, v_2, v_7, v_{17}\}$ (see also Fig. 1). In Fig. 4b v_3, v_4, and v_6 prefer hydrophobic amino acids owing to the existence of v_{12} and v_{13}, while v_1, v_2, and v_5 prefer hydrophilic ones owing to the absence of v_{16} and v_{17}. On the other hand, in Fig. 4c v_2 and v_5 (= v_7) prefer hydrophobic amino acids owing to the existence of v_{17}, while v_4 and v_6 prefer hydrophilic ones owing to the absence of v_{12}, v_{13}, and v_{16}. v_1 is ambivalent.

Gly at v_9 shows remarkable high frequency. Since Gly is frequently found C-cap region of the α-helix, it may indicate that the situation represented by this code appears more often near the C-terminal of the α-helix.

In order to illustrate that the frequencies $\mathbf{f}_{c,b}$ differ depending on the codes (i.e., environment) even for the same vertex position, some examples are given in Figs. 5a and b. For the four codes, (1) $c(T_6)$: $c(T_7)$ = ABEHCDfG: ABEHCDfG, (2) ABEHCDfG: ABEHCDxx, (3) ABEHCDxx: ABEHCDfg, (4) ABEHCDxx: ABEHCDxx., the frequencies $\mathbf{f}_{c,b}$ on the vertices v_1 and v_4 are shown. As for the four surrounding residues around the α-helix, v_{12}, v_{13}, v_{16}, and v_{17}, all of them exist in (1), only v_{12} and v_{13} in (2), only v_{16} and v_{17} in (3), and none of them in (4). v_1 is affected by the existence of a pair of v_{16} and v_{17}, and v_4 by a pair of v_{12} and v_{13}. As a result, v_1 of (1) and (3) prefers hydrophobic amino

acids (Fig. 5a), and so does v_4 of (1) and (2) (Fig. 5b). Both v_1 and v_4 of (4) prefer hydrophilic amino acids.

In Fig. 5c the frequencies $\mathbf{f}_{c,b}$ for the surrounding vertices v_{13} and v_{16} are shown. In any case the hydrophobic amino acids are strongly preferable, because they interact with the residues in the α-helix.

Through Figs. 4 and 5 Ala residue is remarkable. While Ala behaves essentially like hydrophobic amino acids, it frequently appear at hydrophilic amino acid preferable verteiecs. The small sidechain and strong preference for forming the α-helix are considered to make it possible.

The results shown so far can be examined from the viewpoints of principal components. The first and second principal components for the frequencies $\mathbf{f}_{c,b}$, i.e., the projection on the first and second principal axes were calculated for each of the 36 codes. Only the data for the vertices (a) v_1 and (b) v_4 are plotted in Fig. 6. In order to clarify the results the codes were divided into six and five groups for v_1 and v_4, respectively, taking into account the existence of residues at the four vertices, v_{12}, v_{13}, v_{16}, and v_{17} (see the caption of Fig. 6). In the first principal axis (horizontal one) the large positive (negative) value indicates the preference for the hydrophobic (hydrophilic) amino acids. In the second principal axis (vertical one) the large positive (negative) values indicates the preference for amino acids with smaller (larger) sidechains.

Since $T_{13}=\{v_1, v_4, v_6, v_{13}\}$ and $T_{16}=\{v_1, v_4, v_7, v_{16}\}$, both v_1 and v_4 are affected by v_{13} and v_{16}. On the other hand, $T_{12}=\{v_3, v_4, v_6, v_{12}\}$ and $T_{17}=\{v_1, v_2, v_7, v_{17}\}$ indicate that v_{12} can affect v_4 but not v_1, while v_{17} can affect v_1 but not v_4. Generally speaking, Fig. 6 shows that in the existence of such influential residues the hydrophobic amino acids with smaller sidechains (open symbols in Fig. 6) are preferable, and that in the absence of them the hydrophilic amino acids with larger sidechains (closed symbols) are preferable. However in some cases (Group 5 for v_1 and Group 4 for v_4; symbols + and * in Fig. 6), where some of the influential residues exist but others do not, hydrophobic amino acids with larger sidechains are preferable.

Tetrahedron on the contact surface

The tetrahedrons classified as contact surface consist of the vertices v_1 to v_4 in the same chain and at least one of the vertices v_5 to v_8 in a different chain. Although such tetrahedrons assigned to the α-helix related codes are found in the protein structure set considered here, the number of such tetrahedrons is too small to obtain reliable statistical analysis results (see Table 2). For the cases of $c(T_6)$: $c(T_7)$= ABEHCDfG: ABEHCDfG and ABEHCDfG: ABEHCDfg, where the numbers of the data are relatively larger than the

others, the differences of $\mathbf{f}_{c,b}$ between tetrahedrons classified as intra-chain and contact surface are plotted for some vertices in Fig. 7. In Fig. 7a the differences for the vertices v_1, v_4, and v_6 facing surrounding residues v_{12}, v_{13}, v_{16}, and v_{17}, some of which is in a different chain, are plotted. On the other hand, in Fig. 7b, the differences for the vertices v_2 and v_9 facing the interior residues in their own chain are plotted. The frequencies of the hydrophilic amino acids increase in Fig. 7a and those of the hydrophobic amino acids do so in Fig. 7b, although such changes are too subtle to assert it.

Cap regions

The cap regions are very interesting in the analysis of the α-helix. In this paper, however, we analyzed only the tetrahedrons characterized by the codes $c(T_1) = c(T_5) =$ FHABCDEG but $c(T_8) \neq$ FHABCDEG, and $c(T_1) = c(T_8) =$ FHABCDEG but $c(T_5) \neq$ FHABCDEG, corresponding to the N- and C-caps, respectively, while $c(T_1) = c(T_5) = c(T_8) =$ FHABCDEG in the interior of the α-helix. Either $c(T_8)$ or $c(T_5)$ not equal to FHABCDEG is an indication that the N- or C-terminal of the α-helix is distorted, respectively.

In Table 3 the codes, for which more than 60 data are found in the structural data set used in this study, are shown. The correspondence between the residue and vertex positions are given in Table 4. In Table 4 the helix positions for the helix capping defined by Aurora and Rose (1998) are also given. The frequency $\mathbf{f}_{c,b}$ at some vertices for (a) the N-cap with the code $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$ = FHABCDEG: FGABEHCD: ABEHCDxx: ABCDEGxx and (b) the C-cap with the code FHABCDxx: ABEHCDgx: ABEHCDFG: FHABCDEG are plotted in Fig. 8.

Aurora and Rose pointed out the significance of the interactions of N3 and N4 (i.e., v_3 and v_4) with the N' or N" (v_{12} and v_{13}) and the significance of Nc (v_6) for the N-cap. In Fig. 8a hydrophobic amino acids occur more frequently at vertices v_3, v_4, and v_{13}, while Asp, Glu, and Gln are also frequent at v_3 and Lys and Gln at v_4. The occurrence of Asp, Ser, Thr and Pro are remarkably higher at v_6.

For the C-cap Aurora and Rose revealed that Gly is rich at C' (v_5) and interactions of C3 (v_1) with the residues following C' (not included in the codes related to the C-cap considered here). In Fig. 8b Gly is exceptionally abundant at v_5, and hydrophobic residues, especially Leu, is found much more often at v_1.

These results for the N-cap and C-cap well agree with the analysis by Aurora and Rose in general.

DISCUSSION

The amino acid preference for forming the α-helix is usually defined as the ratio of the frequency of a given amino acid found in the α-helix state to that found in all the states. In this paper we examined frequencies of occurrence of 20 amino acids for a given position of the α-helix in a given environment, which is specified with respect to the vertex position of Delaunay tetrahedron and the Delaunay codes assigned to it according to the rules proposed by Wako and Yamato (1998). The former parameters are useful to predict secondary structure locations in amino acid sequences. In other words, they are useful, when the conformational state for a give residues to take in a given sequence is interest. On the contrary, when we adopt the latter parameters, we can get information about preferable amino acids at a given position of the α-helix in a given environment. Since the codes are assigned so as to reflect the environment of the relevant position, we can regard them as the environment-dependent and position-specific amino acid frequencies.

Assume that we have a protein whose three-dimensional structure is known. We can assign the code to every Delaunay tetrahedron after the Delaunay tessellation. If we want to substitute a different amino acid for some amino acid residue in the protein, we can examine the local structures containing the tetrahedrons with the same code in other protein structures. The statistical analyses described here can provide helpful information about the candidate amino acids for the substitution, which is environment dependent and position specific.

As for the environment-dependent amino acid substitution tables, they were constructed by structural alignment data of homologous proteins by Overington et al. (1990, 1992). In constructing the tables the conformational states (defined as combination of secondary structures, buried/exposed, hydrogen bond formation, and so on) are taken into account. These tables show that the substitution pattern depend on the conformational states. Wako and Blundell (1994a, b) used these tables in the prediction of the secondary-structure and the solvent-accessibility classes, and emphasized the significance of the position-dependent (or conformational state dependent) information about amino acid substition patterns.

In connection with the environment-dependent amino acid substition tables, it should be emphasized for the method discussed in this paper that the conformational states and environment can be taken into account through the Delaunay codes. The classification reflects more precise situation where the local structure is located, and does not require the collection and alignment of homologous proteins. Although we confine ourselves into the α-helix in this paper, the same method can be apply any local structure motifs defined by Delaunay codes. In other words the present method makes it possible to analyze the amino acid frequencies of occurrence for the structures without being restricted to conventional conformational states such as α-helix, β-structure, turn, and so on.

Finally, we summarize the results obtained in this paper for the α-helix.

(1) The interior region of α-helix represented by the Delaunay code FHABCDEG: FHABCDEG: x: y: FHABCDEG is examined. There are 36 codes of this type. The differences between the codes reflect the existence or absence of surrounding residues around the α-helix.

(2) The major factors to the occurrence of the 20 amino acid in the interior of the α-helix are (a) preference for α-helix formation, (b) hydrophobicity and hydrophilicity, and (c) sizes of amino acid sidechains.

(3) The above factors (b) and (c) are environment-dependent and position-specific. The codes used in this paper can represent the environment of the α-helix, and the statistics based on the codes can provide the position-specific frequencies of amino acid occurrence.

(4) Ala is a notable amino acid in the α-helix. Its behavior is essentially for hydrophobic amino acid. However, it occurs at the vertices preferable for hydrophilic amino acids to some extent.

(5) The frequencies of amino acid occurrences on the α-helix surface in contact with the residues in a different chain are examined by analyzing tetrahedrons lying across the two chains. Although the difference in the frequencies from those in the one chain is very subtle and the data is not enough for the reliable statistical analyses, it indicates that the hydrophilic amino acids are slightly more preferable on the surface in connecting with another chain, while the hydrophobic amino acids are slightly more preferable on the surface facing the interior of the own chain.

(6) The N-cap and C-cap of the α-helix represented by the codes $c(T_1) = c(T_5) =$ FHABCDEG but $c(T_8) \neq$ FHABCDEG, and $c(T_1) = c(T_8) =$ FHABCDEG but $c(T_5) \neq$ FHABCDEG, respectively, are examined. The frequencies of amino acid occurrences have much more particular patterns compared with the interior of the α-helix. Unfortunately, the number of data is too small to perform the statistical analyses with confidence.

REFERENCES

Altmann, K.-H., J. Wójcik, M. Vásuez, and H.A. Scheraga, 1990. Helix-coil stability for the naturally occurring amino acid in water. XXIII. Proline parameters from random poly(hydroxybutylgultamine-co-L-prorine). *Biopolymers*. 30:107-120.

Aurora, R., G.D. Rose. 1998. Helix capping. *Protein Sci.* 7:21-38.

Aurora, R., R. Srinivasan, G.D. Rose. 1994. Rules for α-helix termination by glycine. *Science* 264:1126-1130.

Bernstein, F.C., T.F. Koetzle, G.J. Williams, E.E. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*. 112:535-542.

Blundell, T.L. and Z.-Y. Zhu. 1995. The α-helix as seen from the protein tertiary structure: a 3-D structureal classication. *Biophys. Chem*. 55:167-184.

Chakrabartty, A., and R.L. Baldwin. 1995. Stability of α-helices. *Adv. Protein Chem.* 46:141-176

Chou, P.Y. and G.D. Fasman. 1974. Conformational parameters for amino acid in helical, β -sheet, and random coil regions calculated from proteins. *Biochemistry.* 13:211-221.

Creighton, T.E. 1983. Proteins: Structures and Molecular Properties. 2nd ed. W.H. Freeman and Company, New York.

Hobohm, U., M. Scharf, R. Schneider, and C. Sander. 1992. Selection of representative protein data set. *Protein Sci*. 1:409-417.

Hobohm, U., and C. Sander. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3:522-524.

Kumar, S., and M. Bansal. 1998a. Dissecting α-helices: Position-specific analysis of α-helices in globular proteins. *Proteins* 31:460-476.

Kumar, S., and M. Bansal. 1998b. Geometrical and sequence characteristics of α-helices in globular proteins. *Biophys. J*. 75:1935-1944.

Levitt, M. 1978. Conformational preferences of amino acids in globular proteins. *Biochemistry*. 17:4277-4285.

Overington, J., M.S. Johnson, A. Sali, and T.L. Blundell. 1990. Tertiary structural constraints on proteins evolutionary diversity: templates key, residues and structure prediction. *Proc. Roy. Soc. London*, Ser B. 241:132-145.

Overington, J., D. Donnelly, M.S. Johnson, A. Sali, and T.L. Blundell. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci*. 1:216-226.

Presta, L.G. and G.D. Rose. 1988. Helix signals in proteins. *Science* 240:1632-1641.

Schiffer, M. and A.B. Edmundson. 1967. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* 7:121-135.

Wako, H. and T.L. Blundell. 1994a. Use of amino acid environment-dependent substitution

tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. Mol. Biol*. 238:682-692.

Wako, H. and T.L. Blundell. 1994b. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol*. 238:693-708.

Wako, H., N. Saito, and H.A. Scheraga. 1983. Statistical mechanical treatment of α-helices and extended structures in proteins with inclusion of short- and medium-range interactions. *J. Protein Chem.* 2:221-249.

Wako, H. and T. Yamato. 1998. Novel method to detect a motif of local structures in different protein conformations. *Protein Eng*. 11:981-990.

Williams, R.W., A. Chang, D. Juretic, S. Loughran. 1987. Secondary-structure predictions and medium-range interactions. *Biochim. Biophys. Acta.* 916:200-204.

Zhu, Z.-Y. and T.L. Blundell. 1996. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J. Mol. Biol.* 260:261-276.

TABLE 1. Codes for the interior of α-helix
and correspondence between residues and vertex positions

(FHABCDEG: FHABCDEG: c(T_6): c(T_7): FHABCDEG)

Residue	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	others *			
Vertex	v_{18} v_{20}	v_6 v_8	v_1 v_{10}	v_2 v_{14}	v_3 v_{15}	v_4 v_{19}	v_5 v_7	v_9 v_{11}	v_{12}	v_{13}	v_{16}	v_{17}
c(T_0)		F,H	A	B	C	D	E,G					
c(T_5)			F,H	A	B	C	D	E,G				
c(T_6)		A	B	E,H	C	D			F/f,x	G/g,x		
c(T_7)			A	B	E,H	C	D				F/f,x	G/g,x
c(T_8)	F,H	A	B	C	D	E,G						

* The residues separated by more than 3 residues from the residues i-2 to i+5 along the chain are classified as others.

TABLE 2. Codes of $c(T_6)$ and $c(T_7)$ for the interior of α-helix

$c(T_6)$	$c(T_7)$	No. of data observed		Surrounding
		intra-chain	contact surface	residues *
ABEHCDfG	ABEHCDfG	2346	414	1111
ABEHCDfG	ABEHCDfg	1425	157	1111
ABEHCDfG	ABEHCDfx	1987	160	1110
ABEHCDfG	ABEHCDgf	370	39	1111
ABEHCDfG	ABEHCDgx	381	82	1101
ABEHCDfG	ABEHCDxx	1000	60	1100
ABEHCDfg	ABEHCDfG	1175	159	1111
ABEHCDfg	ABEHCDfg	732	93	1111
ABEHCDfg	ABEHCDfx	869	44	1110
ABEHCDfg	ABEHCDgf	189	23	1111
ABEHCDfg	ABEHCDgx	582	78	1101
ABEHCDfg	ABEHCDxx	1292	40	1100
ABEHCDfx	ABEHCDfG	503	94	1011
ABEHCDfx	ABEHCDfg	562	76	1011
ABEHCDfx	ABEHCDfx	376	33	1010
ABEHCDfx	ABEHCDgf	163	21	1011
ABEHCDfx	ABEHCDgx	1203	108	1001
ABEHCDfx	ABEHCDxx	2208	66	1000
ABEHCDgf	ABEHCDfG	414	41	1111
ABEHCDgf	ABEHCDfg	157	15	1111
ABEHCDgf	ABEHCDfx	320	14	1110
ABEHCDgf	ABEHCDgf	74	8	1111
ABEHCDgf	ABEHCDgx	127	16	1101
ABEHCDgf	ABEHCDxx	273	14	1100
ABEHCDgx	ABEHCDfG	1895	122	0111
ABEHCDgx	ABEHCDfg	786	46	0111
ABEHCDgx	ABEHCDfx	802	54	0110
ABEHCDgx	ABEHCDgf	253	14	0111
ABEHCDgx	ABEHCDgx	211	14	0101
ABEHCDgx	ABEHCDxx	226	13	0100
ABEHCDxx	ABEHCDfG	1034	73	0011
ABEHCDxx	ABEHCDfg	985	37	0011
ABEHCDxx	ABEHCDfx	315	19	0010
ABEHCDxx	ABEHCDgf	253	16	0011

ABEHCDxx	ABEHCDgx	2185	70	0001
ABEHCDxx	ABEHCDxx	1889	0	0000

* Existence and absence of residues on the vertices v_{12}, v_{13}, v_{16}, and v_{17} are indicated by 1 and 0, respectively. For example, 1001 means that residues exist on v_{12} and v_{17}, but not on v_{13} and v_{16}.

TABLE 3. Codes for N-cap and C-cap of α-helix

$c(T_5)$	$c(T_6)$	$c(T_7)$	$c(T_8)$	No. of data observed
N-cap				
FHABCDEG	FABEHCDg	ABEHCDfx	ABCDEGxx	94
FHABCDEG	FGABEHCD	ABEHCDfG	ABCDEGhx	92
FHABCDEG	FGABEHCD	ABEHCDfG	ABCDEGxx	77
FHABCDEG	FGABEHCD	ABEHCDfg	ABCDEGxx	127
FHABCDEG	FGABEHCD	ABEHCDfx	ABCDEGxx	196
FHABCDEG	FGABEHCD	ABEHCDxx	ABCDEGfx	124
FHABCDEG	FGABEHCD	ABEHCDxx	ABCDEGxx	216
C-cap				
FHABCDEg	ABEHCDxx	ABEHCDgx	FHABCDEG	62
FHABCDxx	ABEHCDfG	ABEHCDFG	FHABCDEG	87
FHABCDxx	ABEHCDfg	ABEHCDFG	FHABCDEG	82
FHABCDxx	ABEHCDgx	ABEHCDFG	FHABCDEG	158
FHABCDxx	ABEHCDxx	ABEHCDFG	FHABCDEG	122
HABCDFxx	ABEHCDgx	ABHCEDGf	FHABCDEG	61

TABLE 4. Codes for N-cap and C-cap
and correspondence between residue and vertex positions

N-cap (FHABCDEG: FHABCDEG: FGABEHCD: ABEHCD??: ABCDEG??) *

Residue	i-4, i-3, i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	others			
Vertex	v_{12} v_{13}	v_6 v_8	v_1 v_{10}	v_2 v_{14}	v_3 v_{15}	v_4 v_{19}	v_5 v_7	v_9 v_{11}	v_{16}	v_{17}	v_{18}	v_{20}
c(T_0)		F,H	A	B	C	D	E,G					
c(T_5)			F,H	A	B	C	D	E,G				
c(T_6)	F,G	A	B	E,H	C	D						
c(T_7)			A	B	E,H	C	D		F/f,x	G/g,x		
c(T_8)		A	B	C	D	E,G					F/f,x	H/h,x
Helix position	N’ N”	Nc	N1	N2	N3	N4	N5					

C-cap (FHABCDEG: FHABCDxx: ABEHCD??: ABEHCDFG: FHABCDEG) *

Residue	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	others		none	
Vertex	v_{18} v_{20}	v_6 v_8	v_1 v_{10}	v_2 v_{14}	v_3 v_{15}	v_4 v_{19}	v_5 v_7	v_{16} v_{17}	v_{12}	v_{13}	v_9	v_{11}
c(T_0)		F,H	A	B	C	D	E,G					
c(T_5)			F,H	A	B	C	D				E/e,x	G/g,x
c(T_6)		A	B	E,H	C	D			F/f,x	G/g,x		
c(T_7)			A	B	E,H	C	D	F,G				
c(T_8)	F,H	A	B	C	D	E,G						
Helix position	C5	C4	C3	C2	C1	Cc	C’	C”				

* Symbol ? represents either F/f, G/g, or x.

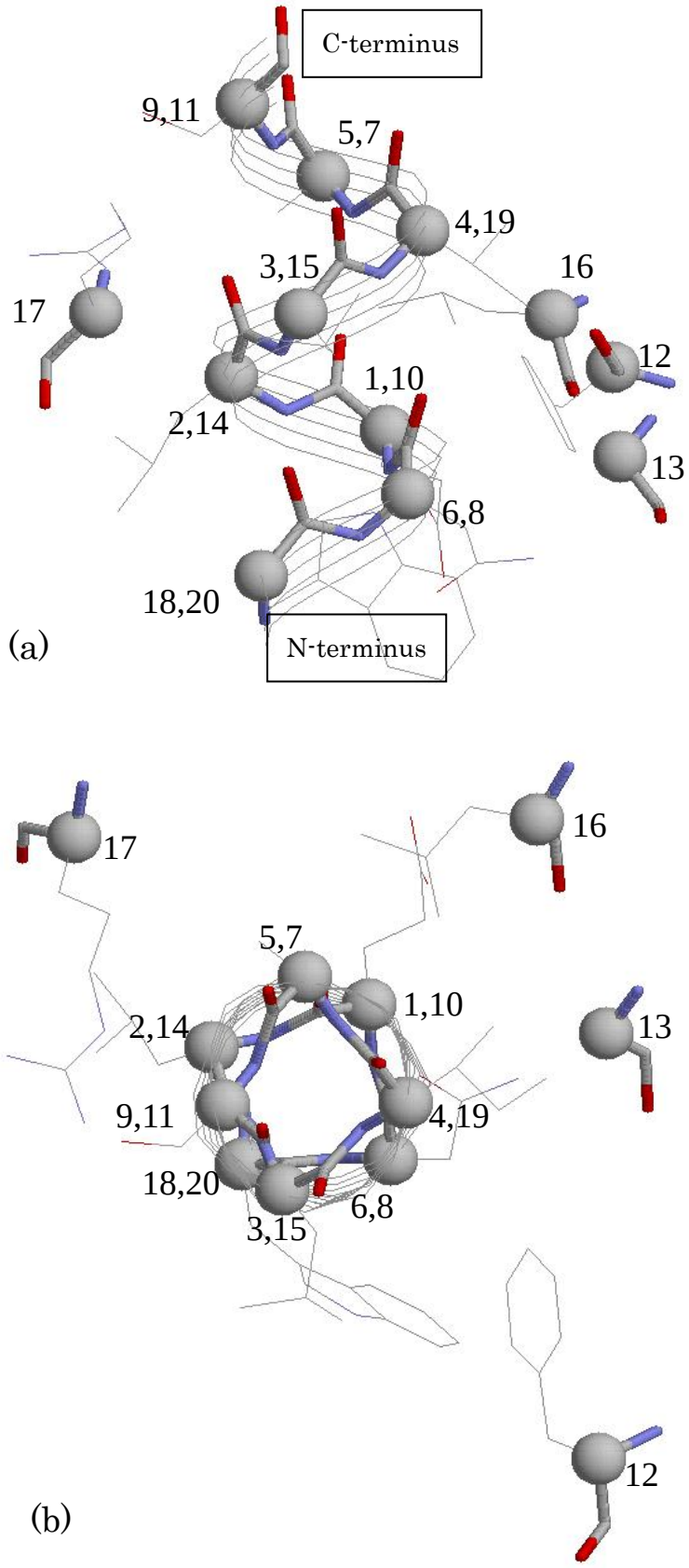


FIGURE 1. Vertex numbers in the interior of α-helix with the code FHABCDEG: FHABCDEG: c(T_6): c(T_7): FHABCDEG. Cα atoms are shown by the balls. Whether or not the vertex residues 12, 13, 16, and 17 exist is reflected in c(T_6) and c(T_7) (see Table 1). Two views are shown: (a) parallel and (b) perpendicular to the helix axis.

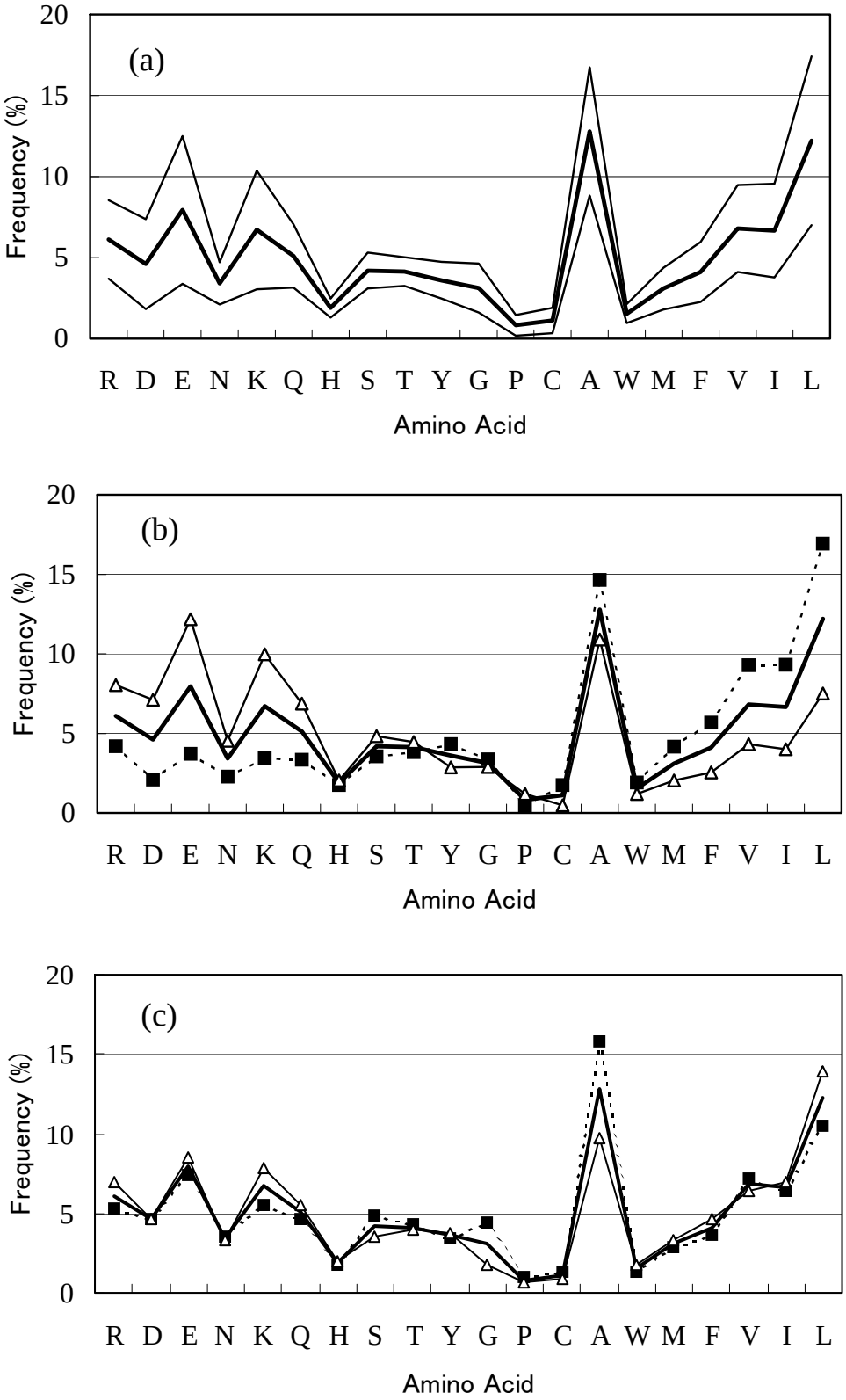


FIGURE 2. Frequency of amino acid occurrence in the interior of the α-helix. (a) $< f_j >$ (thick line) and $< f_j > \pm s_j$ (thin lines); (b) and (c) f_j^* (Eq. 5) for the first and second principal axes, respectively (thin lines) and $< f_j >$ (thick line).

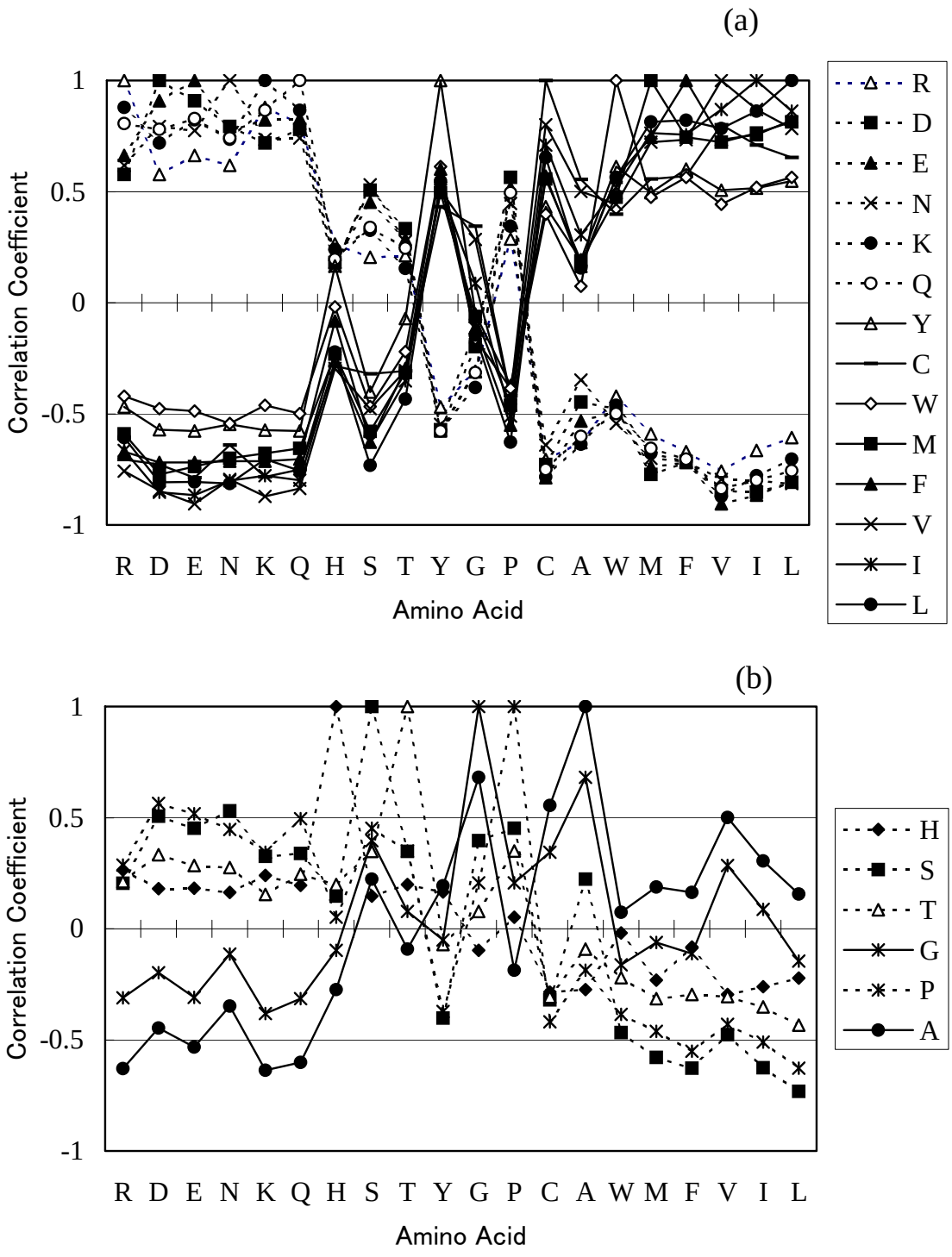


FIGURE 3. Correlation coefficients C_{ij} of 20 amino acids with other amino acids.

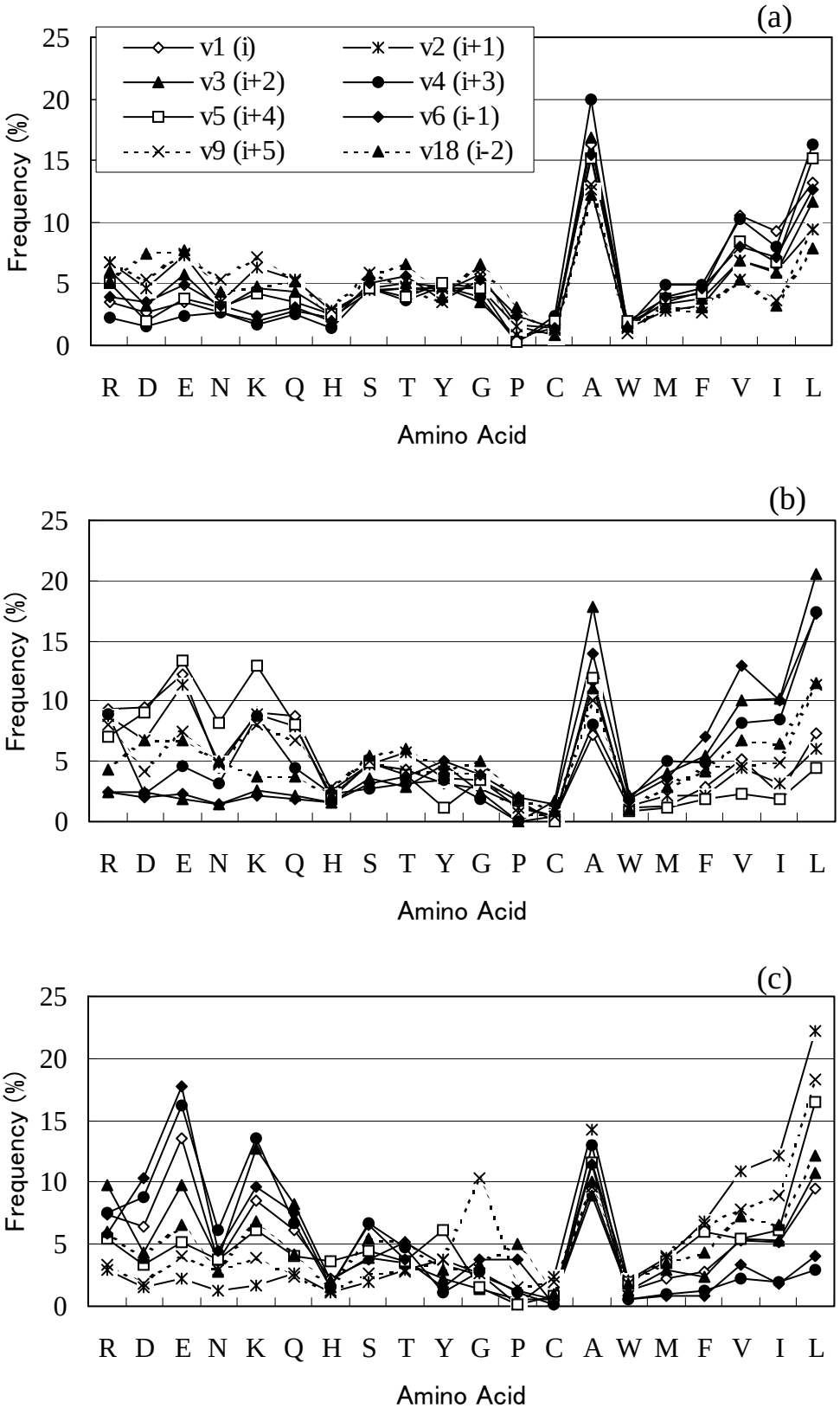


FIGURE 4. Frequency of amino acid occurrence $\mathbf{f}_{c,b}$ at eight vertices corresponding to residues *i*-2 to *i*+5 in the interior of the α-helix. The data are shown for the three kinds of codes: (a) c(T_6): c(T_7)=ABEHCDfG: ABEHCDfG, (b) ABEHCDfG: ABEHCDxx, (c) ABEHCDxx: ABEHCDgx.

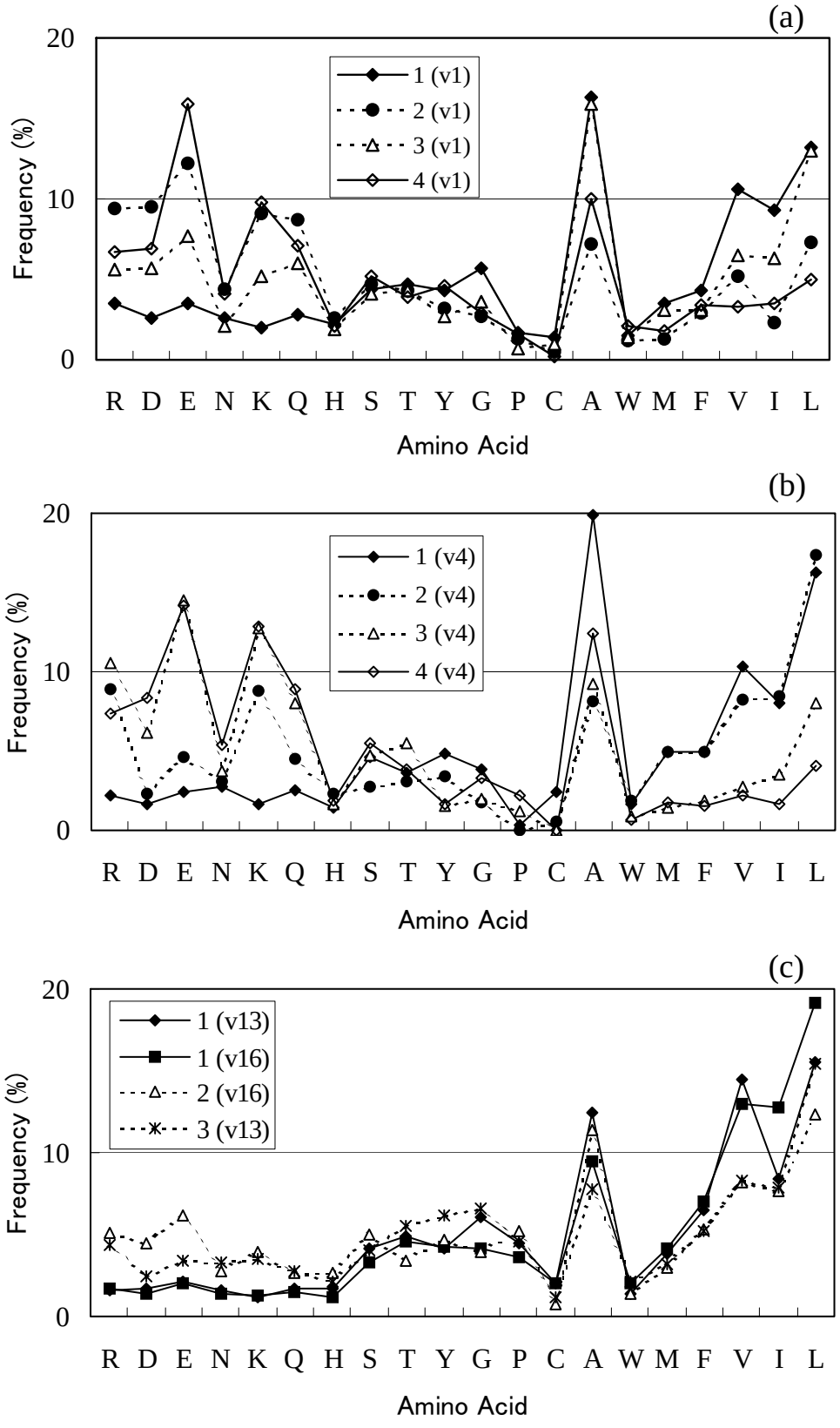


FIGURE 5. Frequency of amino acid occurrence $\mathbf{f}_{c,b}$ at the vertices, v_1, v_4, v_{13}, and v_{16} in the interior of the α-helix. The data are shown for the four kinds of codes: (1) $c(T_6)$: $c(T_7)$ = ABEHCDfG: ABEHCDfG, (2) ABEHCDfG: ABEHCDxx, (3) ABEHCDxx: ABEHCDfg, and (4) ABEHCDxx: ABEHCDxx. $\mathbf{f}_{c,b}$ for (a) v_1 and (b) v_4 of the four codes, and (c) for v_{13} and v_{16} of the three codes are given.

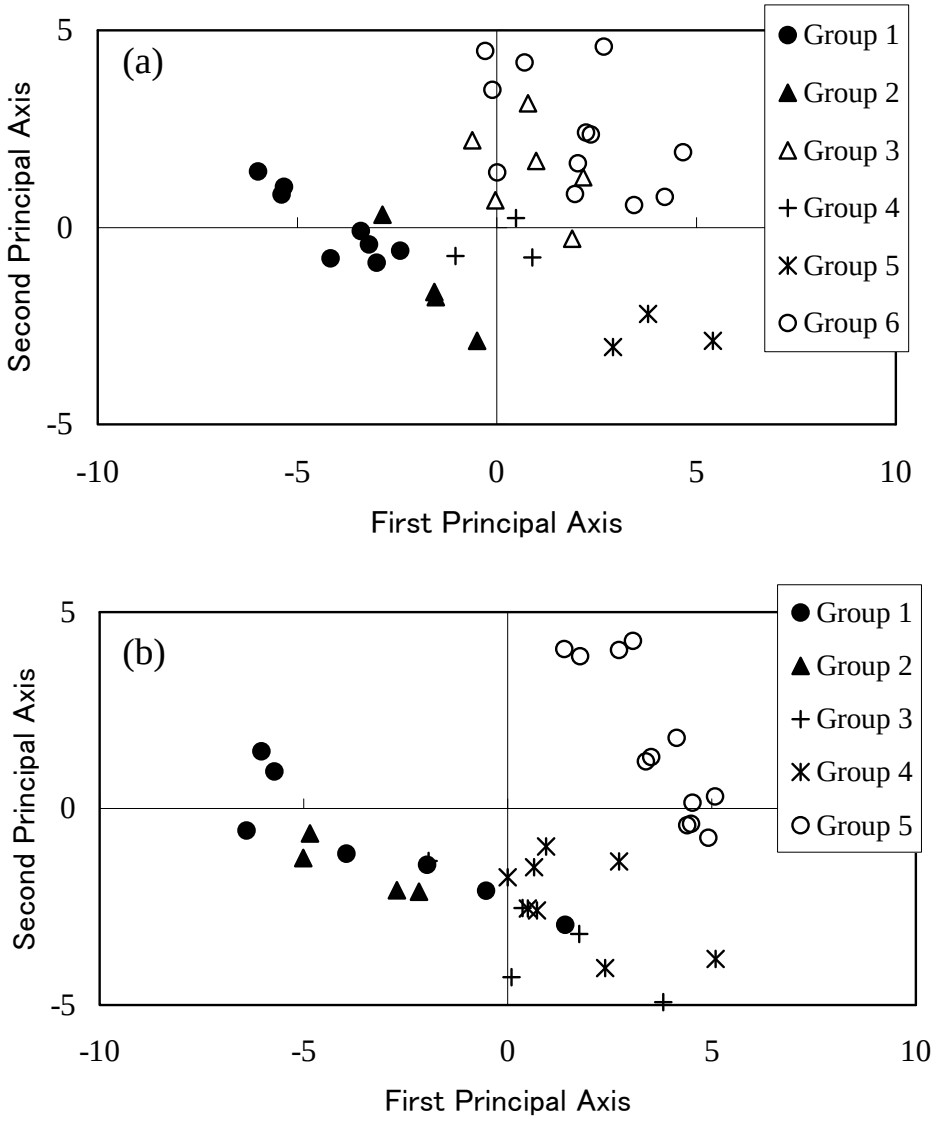


FIGURE 6. The first and second principal components for the frequencies $\mathbf{f}_{c,b}$ at vetices v_1 (a) and v_4 (b). Grouping of the 36 codes given in Table 2 is as follows: (a) Group 1 (0000, 0001, 0100, 1000, 1001, 1100), Group 2 (1010, 1110), Group 3 (0011, 1011), Group 4 (1101), Group 5 (0010, 0101, 0110) and Group 6 (0111, 1111); (b) Group 1 (000, 0001, 1001, 1000, 1100), Group 2 (0011, 0010), Group 3 (1101, 0101, 0100), Group 4 (0111, 0110, 1011, 1010), and Group 5 (1110, 1111) (see the column Surrounding Residues in Table 2 for the four-figure notation).

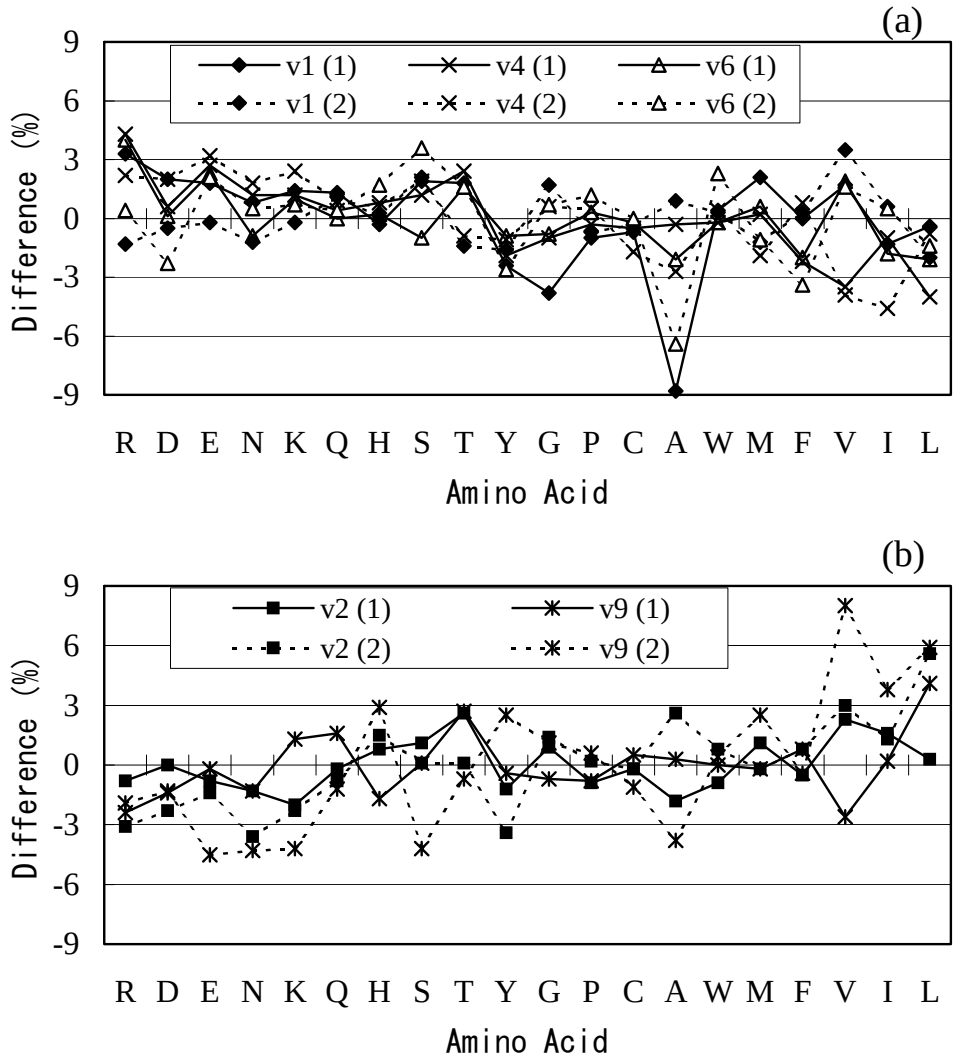


FIGURE 7. Difference in $\mathbf{f}_{c,b}$ for contact-surface tetrahedrons from ones for inta-chain tetrahedrons. The results are shown for the six vertices v_1, v_2, v_4, v_6, and v_9 of the two codes, (1) $c(T_6)$: $c(T_7)$ = ABEHCDfG: ABEHCDfG and (2) ABEHCDfG: ABEHCDfg.

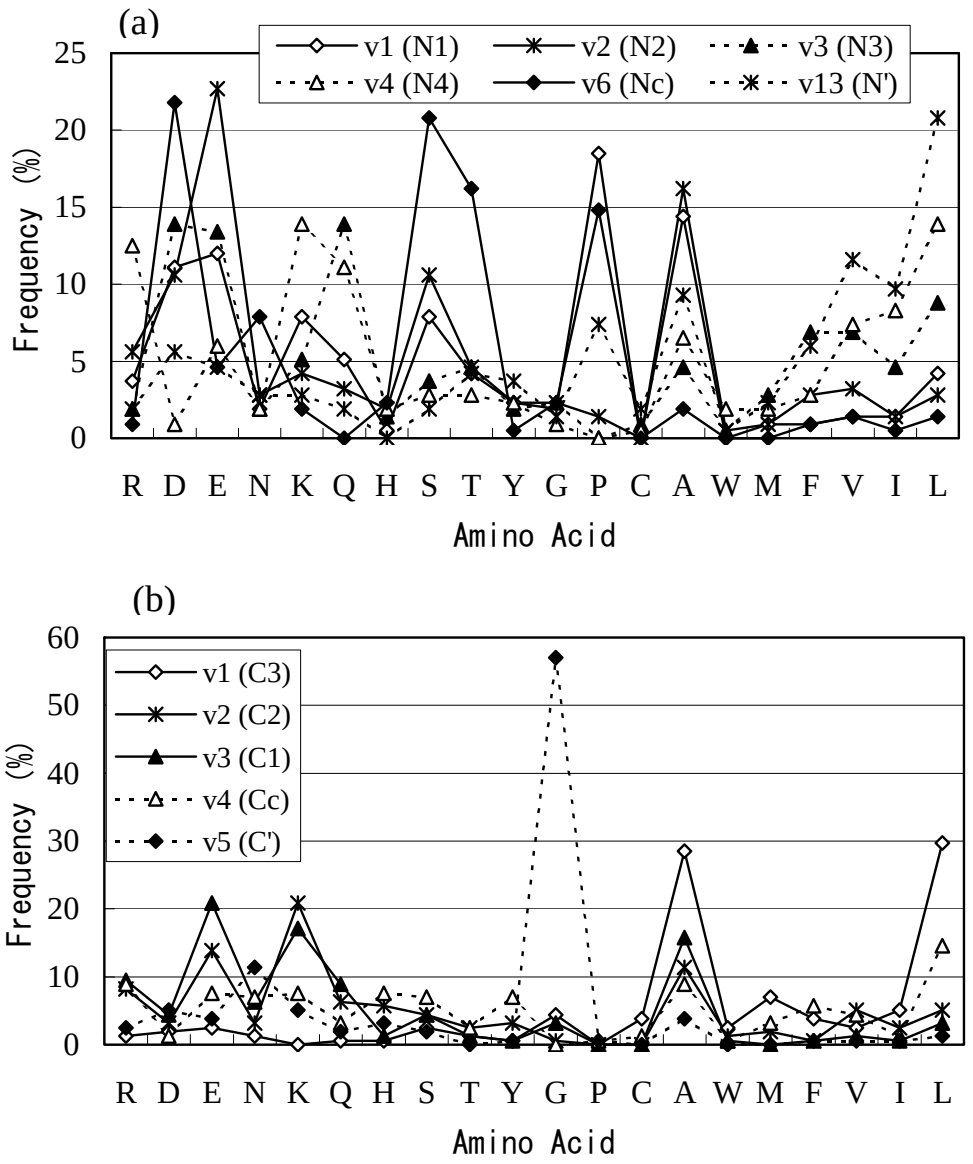


FIGURE 8. Amino acid frequencies $\mathbf{f}_{c,b}$ for N-cap and C-cap. (a) $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$ = FHABCDEG: FGABEHCD: ABEHCDxx: ABCDEGxx (N-cap) and (b) FHABCDxx: ABEHCDgx: ABEHCDFG: FHABCDEG (C-cap).

疎水コア

Delaunay 四面体分割を利用して局所構造を考えることの利点の一つは、四面体の辺が作り出すネットワークが、ペプチド鎖の折りたたみと、残基間の相互作用をともに表していることである。したがって、「4 つの頂点がすべて疎水基(Ala, Val, Leu, Ile, Met, Phe, Trp)で占められ、しかもその四面体に隣接することのできる4つの四面体がすべて存在するDelaunay四面体を含む局所構造」といった定義を与えたとき、それは単純に疎水性アミノ酸残基がペプチド鎖に沿って局在していることを意味しているわけではなく、互いに相互作用する疎水性残基が空間的に局在し、しかも Native な立体構造の表面にはないといったことを意味することになる。こうして定義された局所構造は、したがって、疎水コアと呼ぶことができよう。

Delaunay コードが付けられた局所構造データベースからこうした四面体を検索した結果を以下に紹介する。

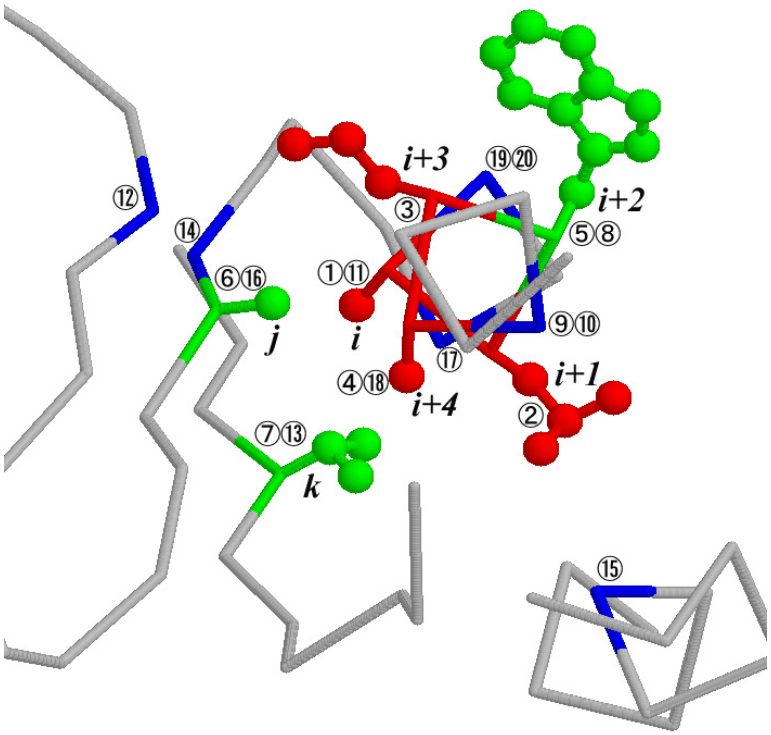
例１：　出現頻度の高い疎水コア

出現頻度の高い疎水コアには、次のようなものがあった。

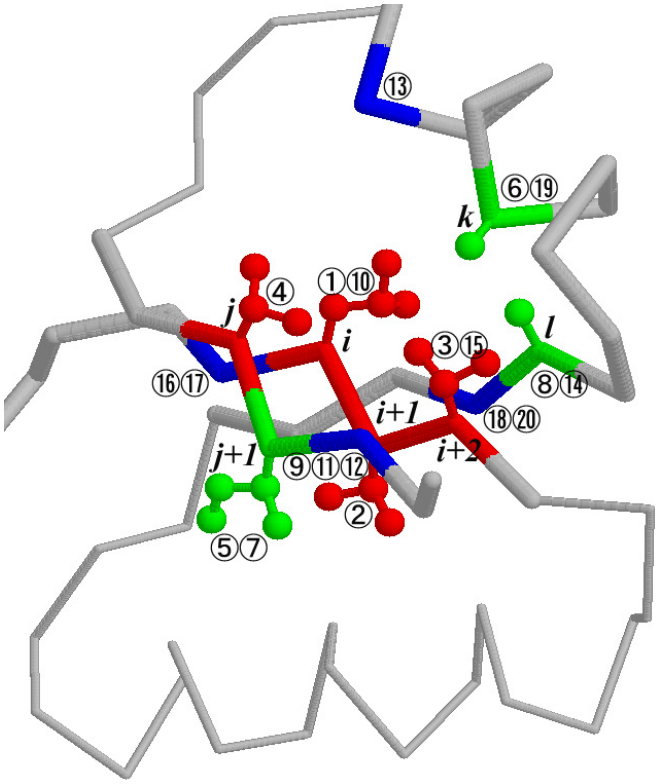
a. α ヘリックスの疎水面を形成する残基からなる四面体。4つの頂点の残基番号パターンは *(i, i+1, i+3, i+4)* で、もっとも多く検出された。これらの疎水基は、ペプチド鎖に沿って離れた、しかし空間的には α ヘリックスに隣接するアミノ酸とともに疎水コアを形成している。
b. 次に多かったのは、α / β 型構造の一部を構成する局所構造であった。四面体頂点の残基番号パターンは、(*i, i+1, i+2, j*) で、これらは平行 β シートを形成する 2 本の β ストランド上にある。周囲の構造までみると、ほとんどの場合、α ヘリックスと少なくとも 3 本の平行 β シートからなる α / β 型構造の一部であった。残基 *i, i+2, j* が α ヘリックスと向かいあう面で側鎖を突き出し、α ヘリックスの β シート側の残基とともに疎水コアを形成している。
c. 2 本の α ヘリックスの接触面にある残基からなる四面体。頂点の残基番号パターンは (*i, i+1, i+4, j*) で、これらが疎水コアそのものを形成している。2 本の α ヘリックスは、互いに垂直に交わるものと、鋭角に交わるものとがあった。
d. b と同様 α / β 型であるが、四面体頂点の残基番号パターンは(*i, j, j+3, j+10*)である。残基 *i* と *j+10* が β 構造で相対しており、*j* と *j+3* は β シートに相対する α ヘリックス上の残基である。

具体的な構造、および残基番号パターンと四面体の頂点の番号との対応関係を次ページの図で示した。

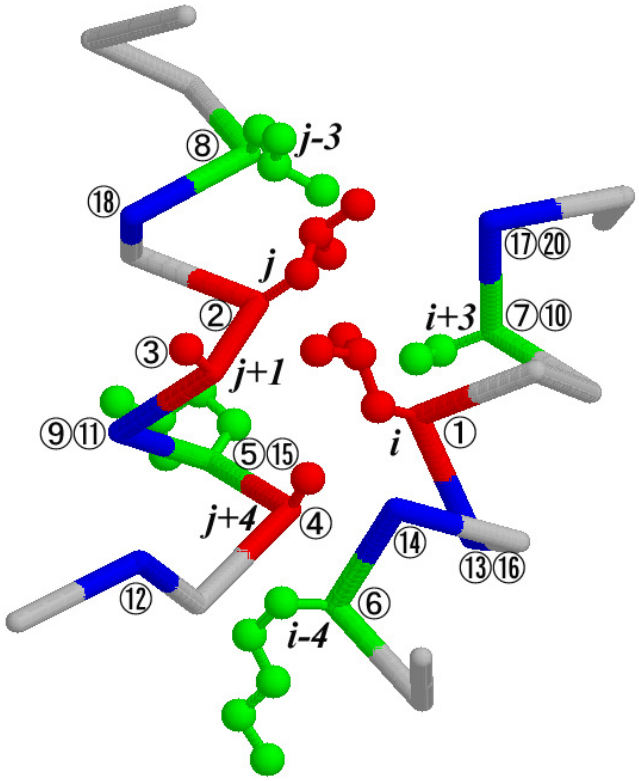
a


i+3
⑲⑳
i+2
⑤⑧
⑫
⑭
③
⑥⑯
①⑪
⑨⑩
j
i
⑰
④⑱
i+1
i+4
②
⑦⑬
k
⑮

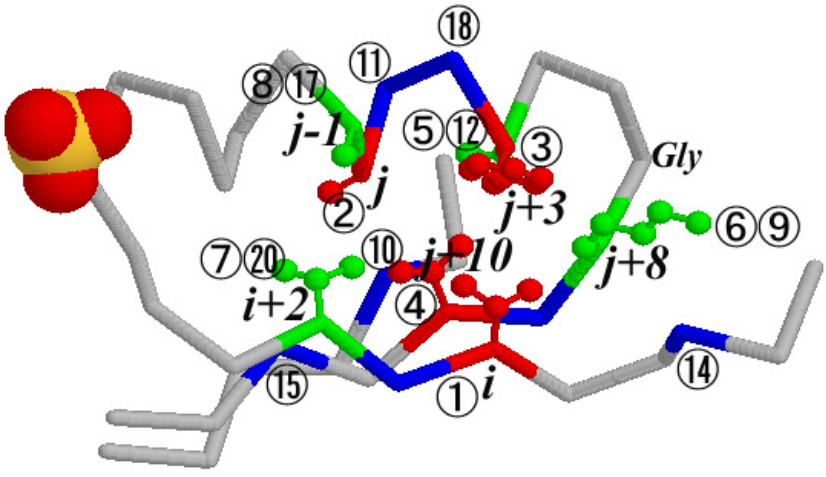
b


⑬
⑥⑲
k
①⑩
④
j
i
l
③⑮
⑯⑰
⑧⑭
i+1
⑱⑳
j+1
⑨⑪⑫
i+2
②
⑤⑦

c


⑧
j-3
⑱
j
②
③
j+1
⑨⑪
⑤⑮
j+4
④
⑫
i-4
⑥
⑭
⑬⑯
i
①
i+3
⑦⑩
⑰⑳

d


⑱
⑪
⑧⑰
j-1
⑤⑫
③
Gly
②
j
j+3
⑥⑨
⑦⑳
⑩
i+10
j+8
i+2
④
⑮
①
i
⑭

例２：　疎水コアとリガンド結合部位

疎水コアを調べていく中で、コア部分だけでなく、コア近傍に共通性をもった局所構造も見出された。

以下に示す例では６つの局所構造を掲げているが、コードの種類としては３つあり、互いによく似ている。実際、局所構造も互いによく似ている。また、コードの違いが、局所構造間の微妙な違いを反映しているのもわかる。これらの例で興味ある点は、疎水コアと同定された部分から少し離れたところにリガンド結合部位があること、しかもそのリガンドの種類は互いに異なっていることである。

局所構造の研究といったとき、これまで多くの研究者の関心はこうしたリガンドの結合部位、あるいは活性部位に集中してきた。しかし、そうした部位を支える部分で、さまざまなタンパク質の間で共通にみられる構造といった視点はあまりなかったようである。このように局所構造の同定に新たな視点を与えることができるのが、本研究で利用してきた Delaunay 四面体コードによる局所構造研究の利点であろう。

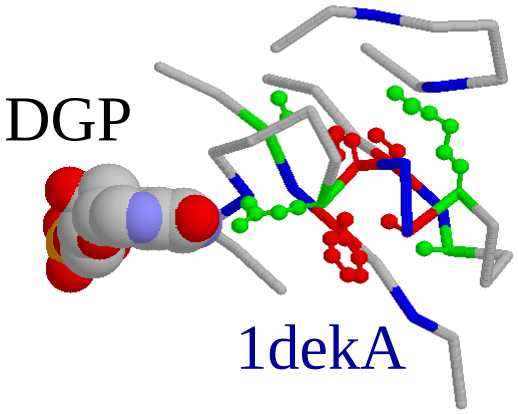
	ECDFbhGA	ABHCdEgx	AHBCedFg	ECHagbFD	AGBcfDeH	1
1dekA	192A V	5A F	182A M	185A A	181A E	186A R 190A A 7A S

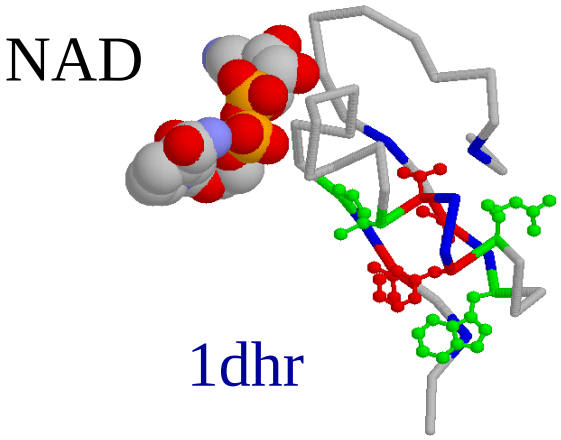
	ECDFbhGA	ABHCdgEx	AHBCedFg	ECHagbDf	AGBhcfDE	2
1ofgA	125A V	102A V	115A A	118A A	114A F	119A F 123A K 104A I
1gtmA	239A V	215A I	228A A	231A M	227A L	232A S 237A M 217A I

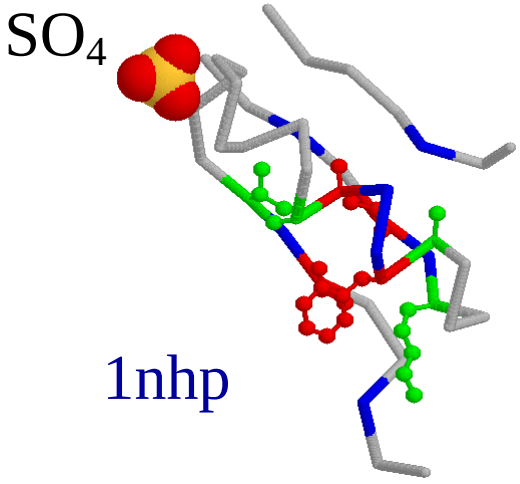
	ECDFbhGA		ABHCdgEx		AHBCedFg		ECHagbFD		AGBhcfDE		3			
1tfr	149	I	15	I	139	V	142	F	138	L	143	S	147	H 17 L
1nhp	175	V	152	V	165	A	168	F	164	A	169	A	173	K 154 V
1dhr	33	V	9	V	23	V	26	F	22	C	27	R	31	W 11 V

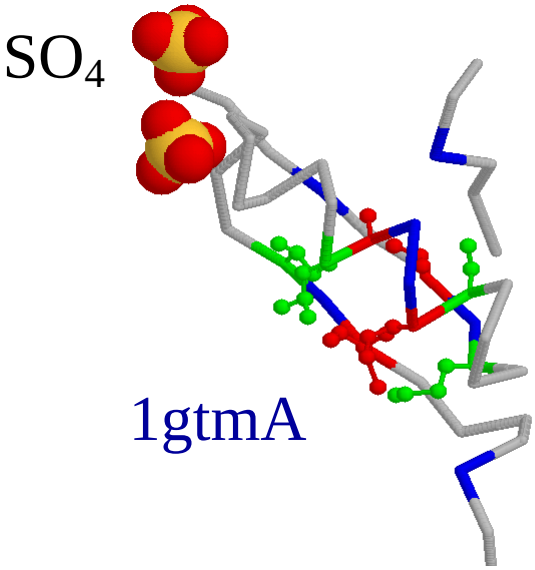
６種類のタンパク質で見出された疎水コア部分のアミノ酸

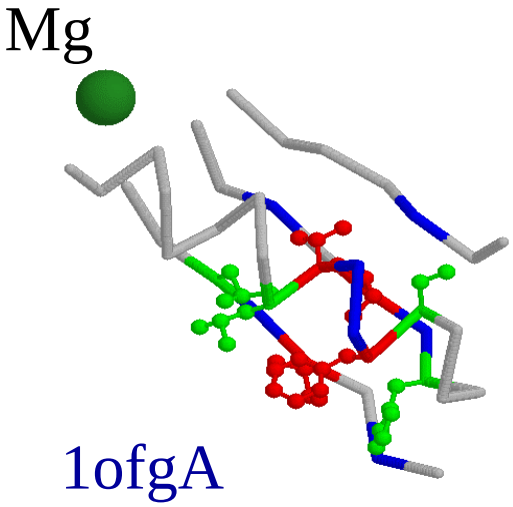
1dek	Phosphotransferase	*P-loop containing nucleotide triphosphate hydrolases*	
1ofg	Oxidoresuctase	*NAD(P)-binding Rossmann-fold domains*	
1gtm	Oxidoresuctase	*NAD(P)-binding Rossmann-fold domains*	(SCOP)
1tfr	Hydrolase	*Resolvase-like*	
1nhp	Oxidoresuctase	*FAD/NAD(P)-binding domain*	
1dhr	Oxidoreductase	*NAD(P)-binding Rossmann-fold domains*	

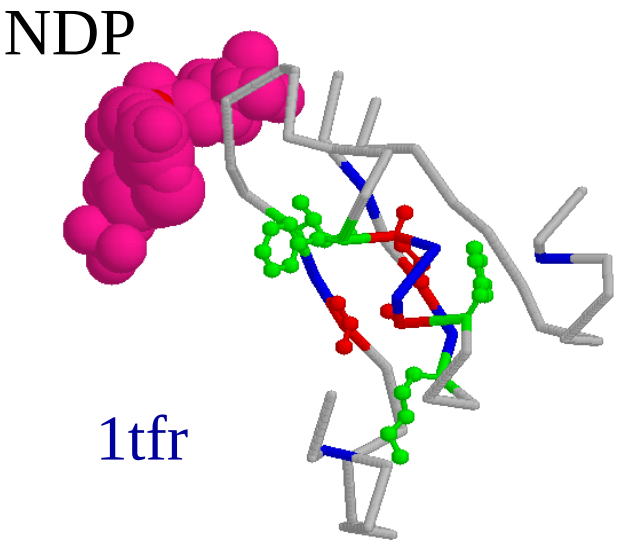

DGP
1dekA


NAD
1dhr


SO4
1nhp


SO4
1gtmA


Mg
1ofgA


NDP
1tfr

より大きな局所構造の同定

方法（1）

われわれは、Delaunay 四面体分割法を利用して、異なるタンパク質の立体構造中に現れる共通の局所構造を見出す方法を提案した。方法の概略は以下の通りである。[*Wako & Yamato, Protein Eng.* (1998) **11**:981-990]

① タンパク質立体構造を、各アミノ酸残基を C_α 原子で代表して表現する。そしてその立体構造を、C_α 原子を頂点とする Delaunay 四面体で分割する。このとき分割は一意的に定まる。四面体の各辺は、ペプチド鎖に沿って隣接する残基を結ぶ線分、あるいはペプチド鎖に沿っては離れているが空間的には隣接している残基を結ぶ線分となる。

② 各四面体に、予め定められたルールによってコードを付与する。コードは、アルファベットの文字列で、局所構造の特徴を反映するように考えられたものである。

③ コード付け作業を、与えられたタンパク質セットについて行う。今回の解析では、互いの相同性が 50%以下のタンパク質からなるタンパク質セットを用いた。

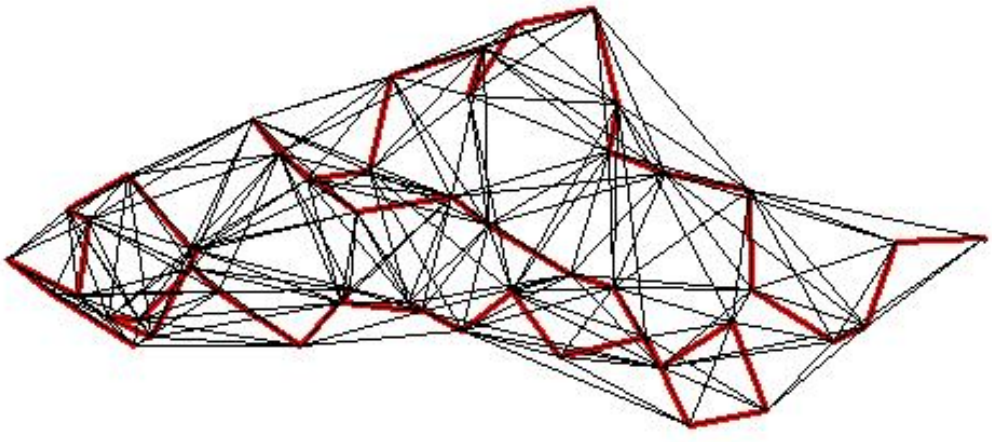


図１．タンパク質立体構造の Delaunay 四面体分割の例
Human transforming growth factor α (PDB 4tgf; 残基数 50)
太線は主鎖を表す

四面体へのコードの付け方

四面体$\mathbf{T}_0$を考える。４つの頂点にあるアミノ酸の残基番号を$v_1(\mathbf{T}_0)$，$v_2(\mathbf{T}_0)$，$v_3(\mathbf{T}_0)$，$v_4(\mathbf{T}_0)$とする。ここで添字１から４は、$v_1(\mathbf{T}_0)<v_2(\mathbf{T}_0)<v_3(\mathbf{T}_0)<v_4(\mathbf{T}_0)$となるように付けるものとする。

四面体$\mathbf{T}_0$と面を共有する周囲の四面体を考える。必ずしも存在するとは限らないが、存在する場合には、それらを$\mathbf{T}_5$，$\mathbf{T}_6$，$\mathbf{T}_7$，$\mathbf{T}_8$とする。それら四面体の頂点にあるアミノ酸で、$\mathbf{T}_0$と共有しない残基の番号を$v_5(\mathbf{T}_0)$，$v_6(\mathbf{T}_0)$，$v_7(\mathbf{T}_0)$，$v_8(\mathbf{T}_0)$とすると、4 つの四面体$\mathbf{T}_5$，$\mathbf{T}_6$，$\mathbf{T}_7$，$\mathbf{T}_8$はそれぞれ{v_2，v_3，v_4，v_5}，{v_1，v_3，v_4，v_6}，{v_1，v_2，v_4，v_7}，{v_1，v_2，v_3，v_8}という残基で形成されるものとする。このルールによって、１から８まで、一意的に頂点の番号が付けられる。

＊ v_5 から v_8 のうち、ある２つの頂点が同一の頂点であることがよく起きる。

以下のルールにしたがって、中央の四面体$\mathbf{T}_0$にコード$C(\mathbf{T}_0)$を付ける。

(1) 残基番号の大きさ順に v_1 ～ v_8 を並べる。

(1-a) もし $a,b\geq 5$ に対して、v_a=v_b かつ $a<b$ ならば、（すなわち、四面体 $\mathbf{T}_a$ と$\mathbf{T}_b$ の頂点が一致しているならば），v_aとv_bはこの順に並べる。

(1-b) もし$a\geq 5$に対して v_aが存在しないならば、(すなわち、四面体$\mathbf{T}_a$が存在しないならば)，v_aはこの並べ替えに含めない。

(2) もし v_1～v_8がv_p，v_q，v_r，v_s，v_t，v_u，v_v，v_w と並んだ場合(添字 p～w は数字 1～8 のどれかを表す)、添字の列‘pqrstuvw’をコードとする。

(3) 1 から 8 までの数字を A から H までのアルファベットに置き換える。アルファベットは大文字、小文字を使い、ペプチド鎖に沿って近いものはクラスター化する。アミノ酸残基が存在しない頂点は x で表記する

(4) 大きなクラスター順に並べ替える。

＊（３），（４）は新たな変更点で、以下の例を参照

(5) すべての四面体にコードを付け、最終的に$\mathbf{T}_0$、$\mathbf{T}_5$、$\mathbf{T}_6$、$\mathbf{T}_7$、$\mathbf{T}_8$に付けられた５つのコードをセットにして$\mathbf{T}_0$のコードとする。

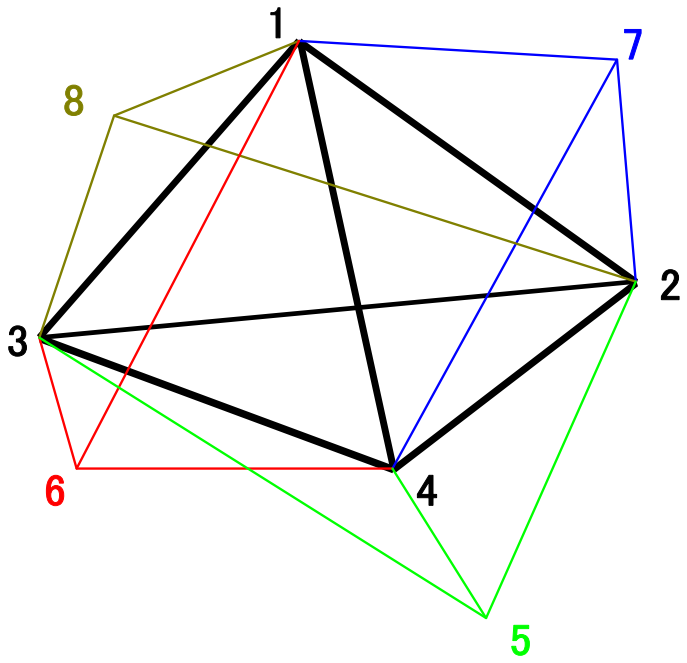


図２．四面体の頂点の番号付け

中央の四面体 1234 とそれを取り囲む４つの四面体 2345，1346，1247，1238

Example:

$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}=\{32, 34, 35, 36, 45, 59, 23, 22\}$

-> 87123456 -> HGabcdEf -> ABCDhgEf

$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}=\{32, 34, 35, 36, 45, 59, 83, 82\}$

-> 12345687 -> ABCDeFhg -> ABCDhgEf

$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}=\{13, 14, 16, 17, 15, 82, 82, 15\}$

-> 12583467 -> ABEHCDfg

$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}=\{51, 60, 62, 63, , 12, , 11\}$

-> 861234 -> HFaBCDxx -> BCDhfAxx -> ABCegDxx

$\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}=\{60, 62, 63, 81, 12, , 11, \}$

-> 751234 -> EGabcDxx -> ABCegDxx

方法（２）

より多くの四面体を含む局所構造への発展

①　あるタンパク質Aについて、何か局所構造（何らかのモチーフと考えられる局所構造で、ここではPROSITEの配列パターンを使った）から、その構成アミノ酸残基をいくつか指定する。

②　そのアミノ酸残基を含む四面体（T_1(A)と呼ぶことにする）について、その四面体のコードと同じコードをもつ四面体をデータベースで検索する。

③　検索の結果、タンパク質Bにそうした四面体が見出されたとすると（その四面体をT_1(B)と呼ぶことにする)、次に、それぞれの四面体に隣接する四面体について、同じコードをもつものがあるかどうかを調べる。なお、このとき、一つの四面体には面を共有する最大で４つの四面体が存在するが、コード付けの規則により、T_1(A)とT_1(B)に隣接する四面体の間には一意的な対応関係があることが重要なポイントとなる。

④　T_1(A)とT_1(B)を出発点として、対応する四面体が同じコードをもつ限りこの操作を繰り返し、隣接する四面体のネットワークを広げていく（図３）。

四面体の対応関係から、その頂点にある残基間の対応関係は自然に定まる。すなわち、立体構造を考慮したアラインメントが一意的に定まることになる。

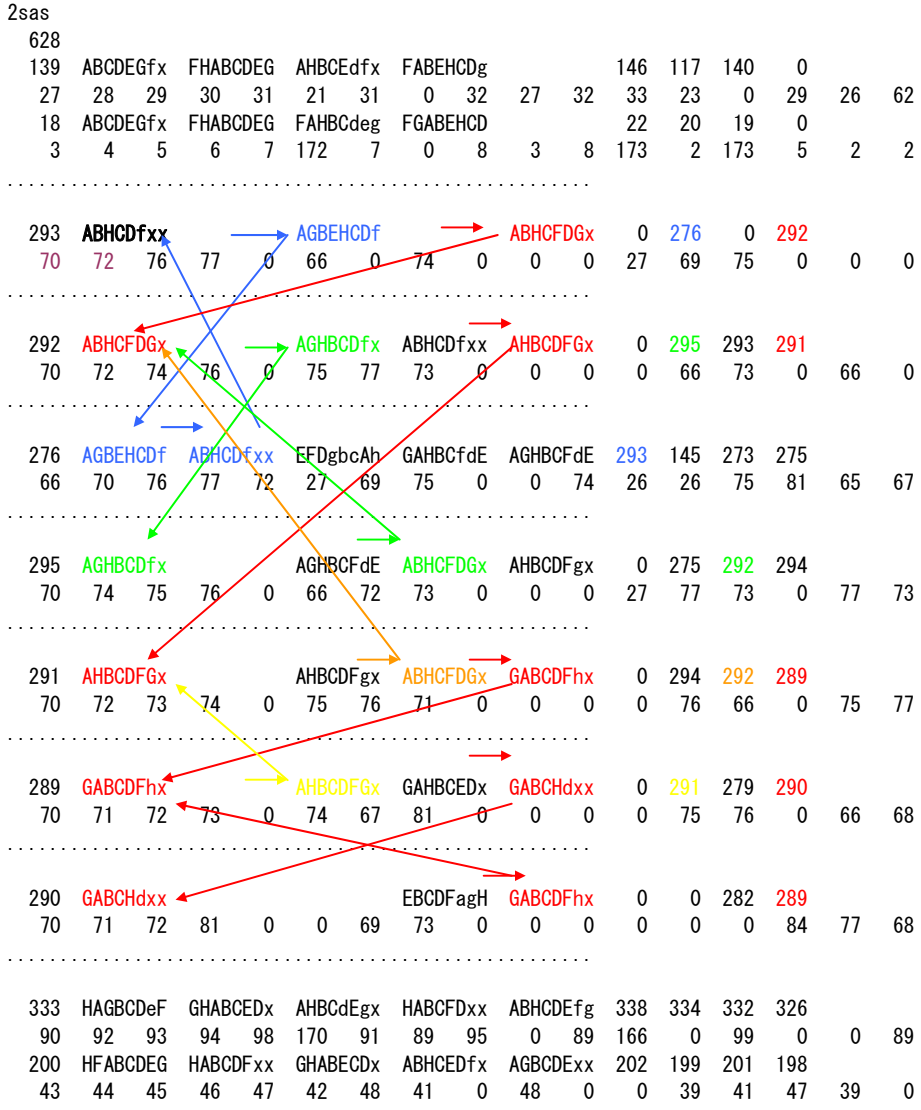
今回の研究では、２０個以上の四面体が共通のコードをもつことを条件にいくつかのPROSITEモチーフについて調べた。

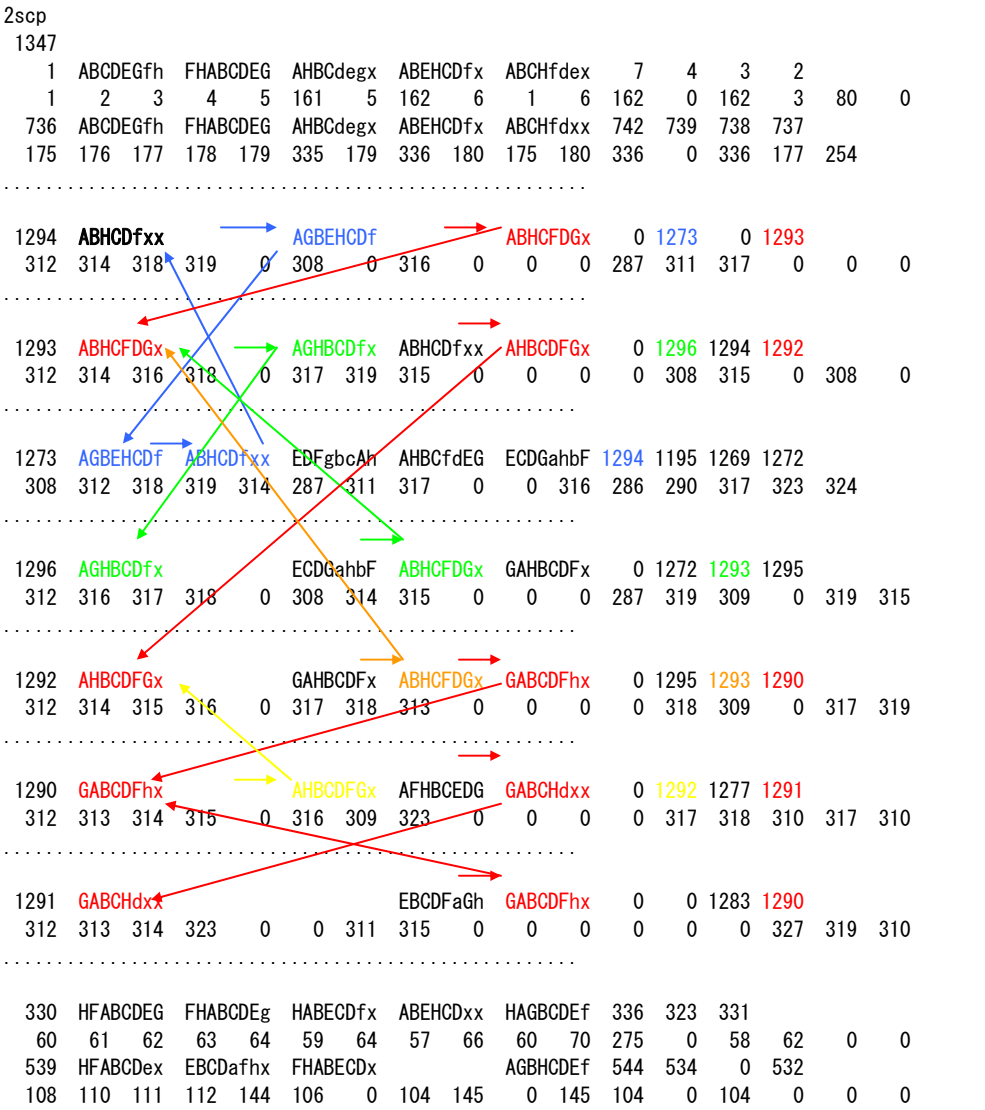
例としてFr-1 protein (1FRB)の結果を示す。このタンパク質は３つのPROSITE モチーフ (PS00062, PS00063, PS00798)　をもっている。

図３．　方法の概念図

２つのタンパク質 2SAS と 2SCP を例として

2SAS、2SCP はそれぞれ　628 個、1,347 個の四面体からなる。
各四面体には、通し番号、コード、隣接する四面体の通し番号、構成するアミノ酸残基のデータが与えられている





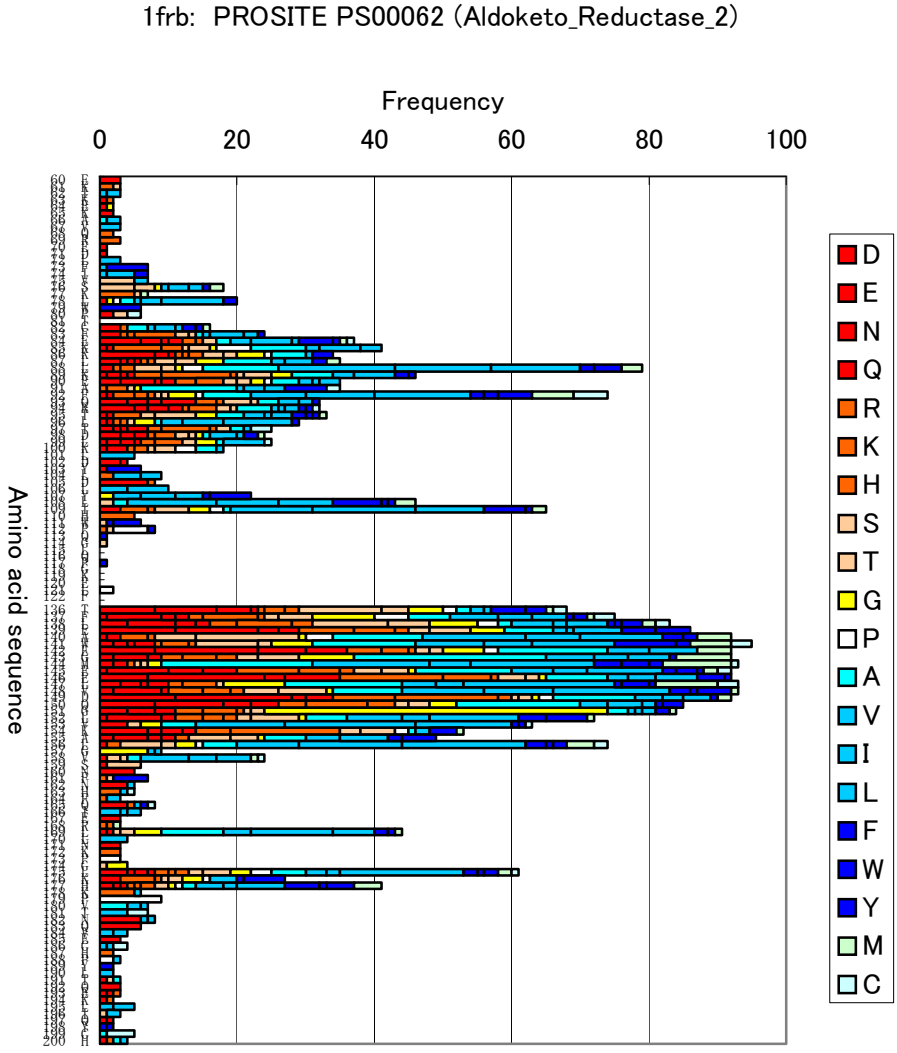


図４. PROSITE モチーフ近傍の局所構造

以下の配列モチーフのアミノ酸を含む四面体（複数ある）をスタートに、同じコードをもつ四面体を他のタンパク質について探索し、見出された頻度をヒストグラムで示した。縦軸が参照タンパク質のアミノ酸配列（部分）。横軸が出現頻度。

PROSITE　PS00063　ALDOKETO_REDUCTASE_3

[LIVM]-[PAIV]-[KR]-[ST]-x(4)-R-x(2)-[GSTAEQK]-[NSL]-x(2)-[LIVMFA]

260　268　275

I P K S V T P S *R* I Q *E N* I Q *V*

図５

図４に対応するアラインメント


PROSITE PS00062 ALDOKETO_REDUCTASE_2 [LIVMFY]-x(9)-[KREQ]-x-[LIVM]-G-[LIVM]-[SC]-N-[FY]
144 154 161
M E E L V D Q G L V K A L G V S N F

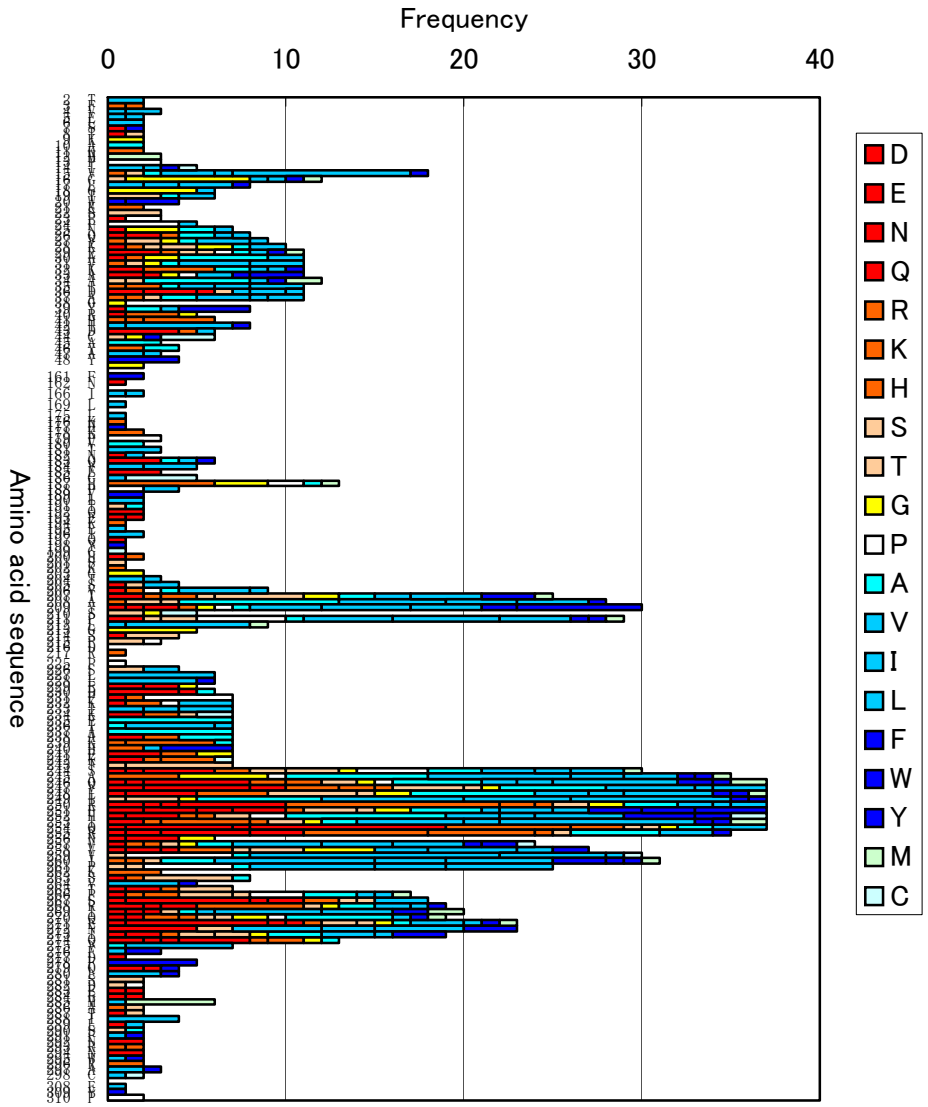


図６．　PROSITE モチーフ近傍の局所構造

図４の説明を参照。

PROSITE　PS00798　ALDOKETO_REDUCTASE_1

G-[FY]-R-[HSAL]-[LIVMF]-D-[STAGC]-[AS]-x(5)-E-x(2)-[LIVM]-G

38　45　51　55

G Y R H I D C A Y A Y C N *E* N E *V G*

図７

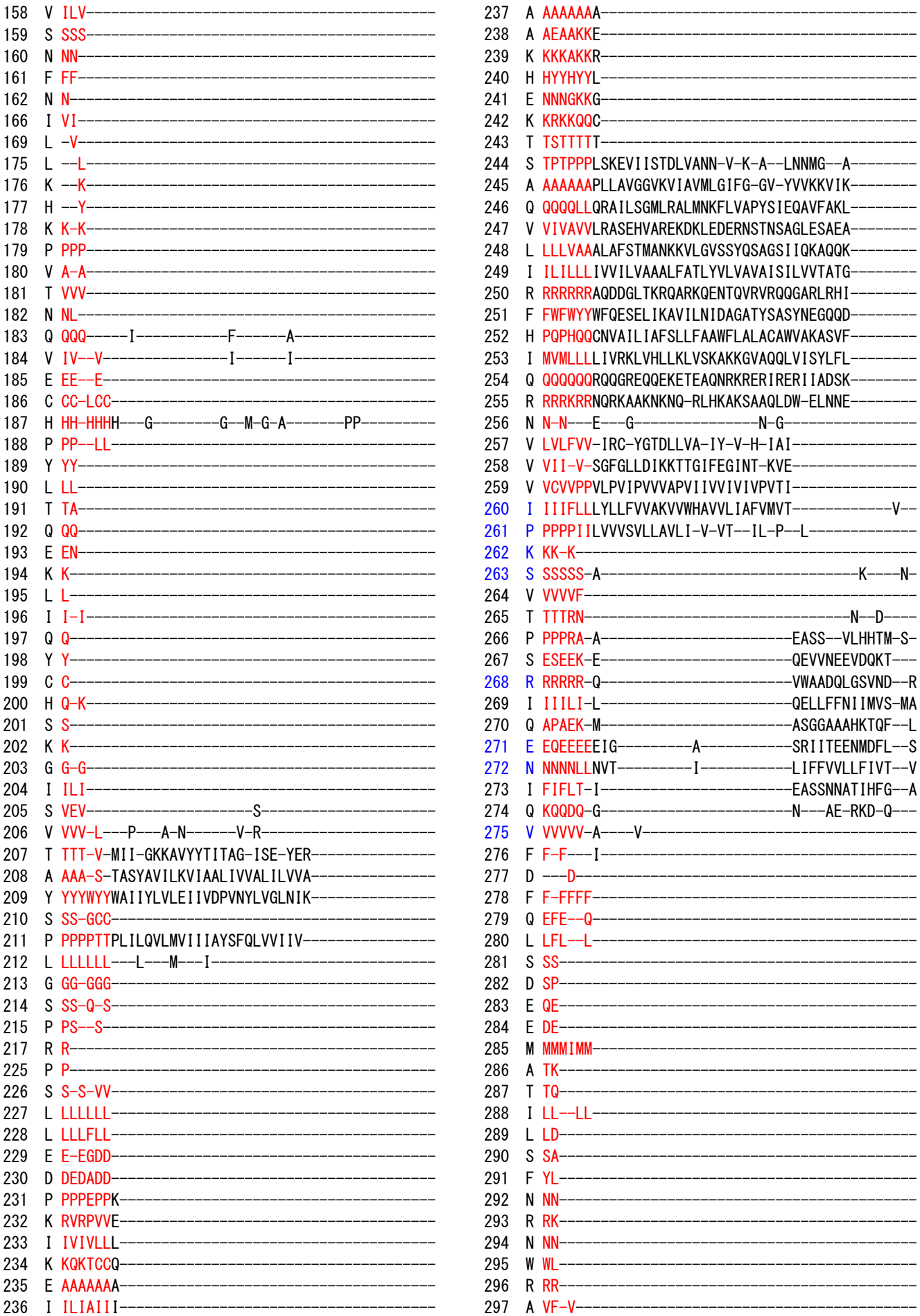
図６に対応するアラインメント

PROSITE PS00063 ALDOKETO_REDUCTASE_3

[LIVM]-[PAIV]-[KR]-[ST]-x(4)-R-x(2)-[GSTAEQK]-[NSL]-x(2)-[LIVMFA]

260 268 275

I P K S V T P S *R* I Q *E N* I Q *V*



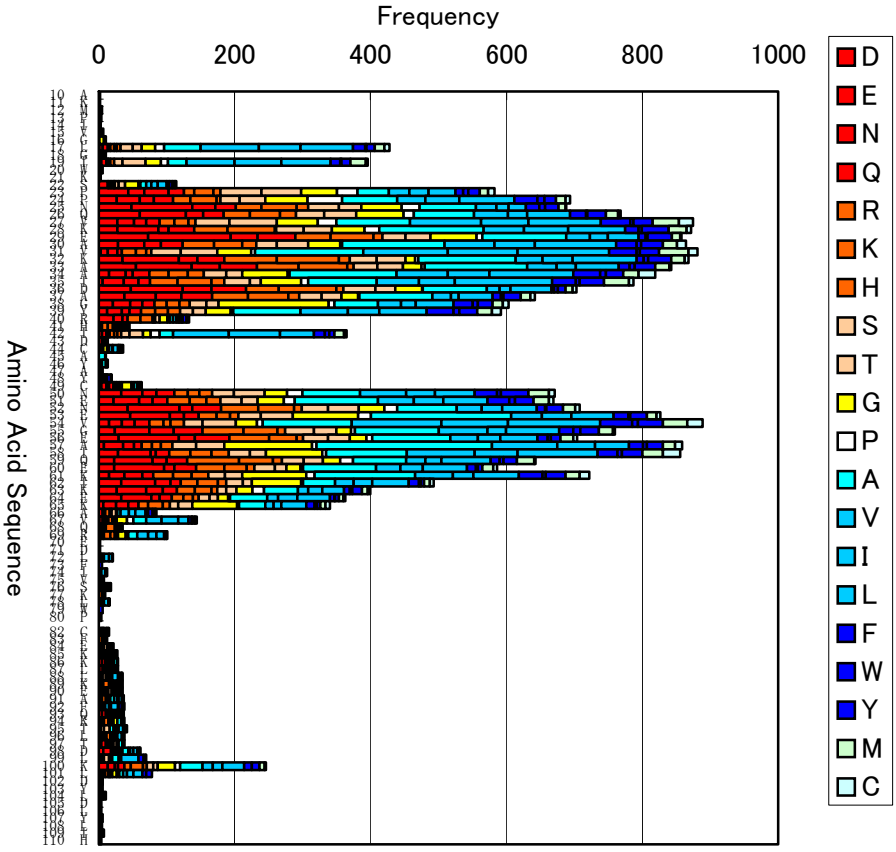


図８. PROSITE モチーフ近傍の局所構造

図４の説明を参照。

PROSITE　PS00062　ALDOKETO_REDUCTASE_2

[LIVMFY]-x(9)-[KREQ]-x-[LIVM]-G-[LIVM]-[SC]-N-[FY]

144　　154　　161

M E E L V D Q G L V *K* A *L* *G* *V* *S* *N* *F*

図９

図８に対応するアラインメント

PROSITE PS00798 ALDOKETO_REDUCTASE_1

G-[FY]-R-[HSAL]-[LIVMF]-D-[STAGC]-[AS]-x(5)-E-x(2)-[LIVM]-G

38 45 51 55

G Y R H I D C A Y A Y C N *E* N E *V G*

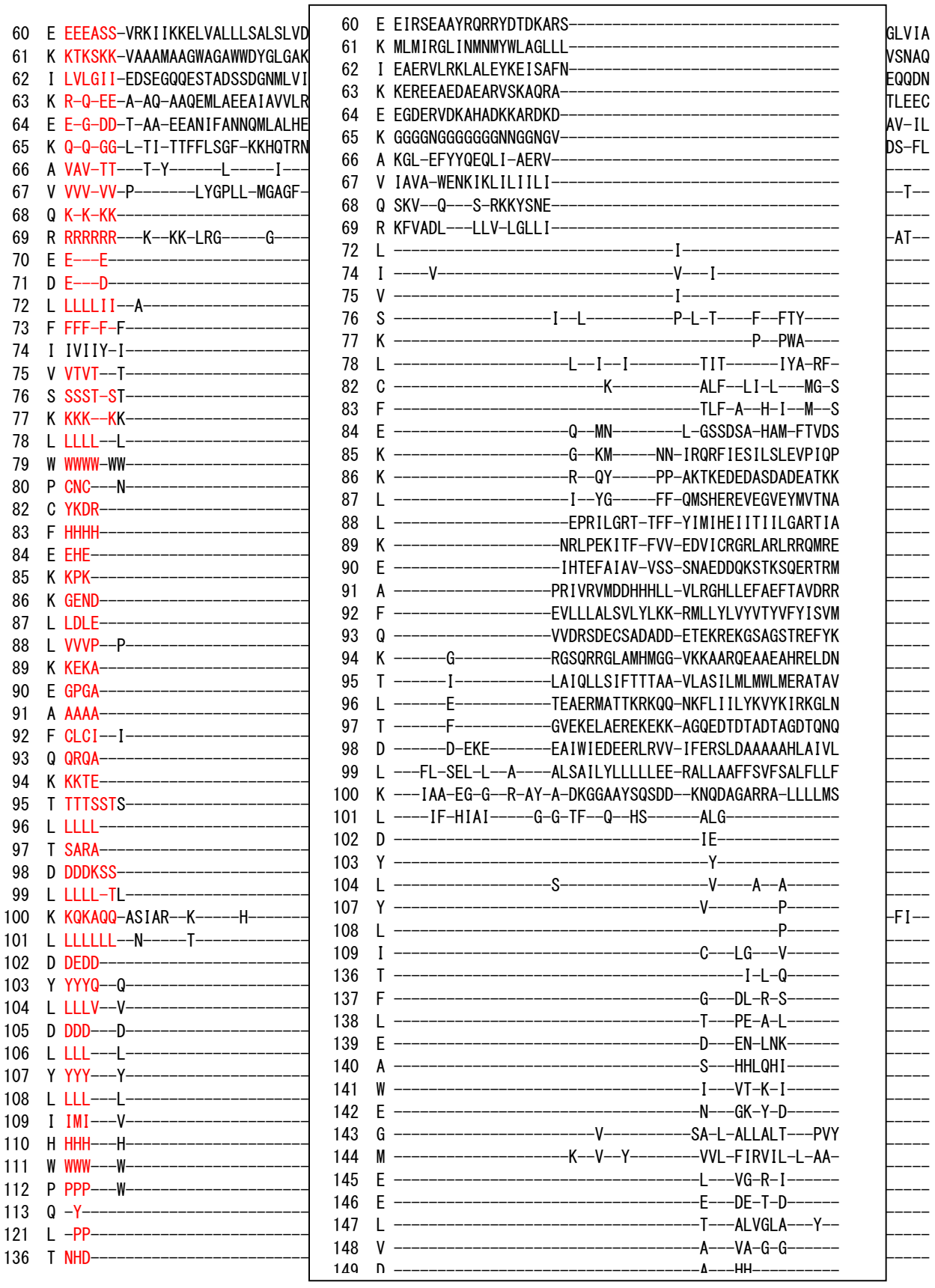
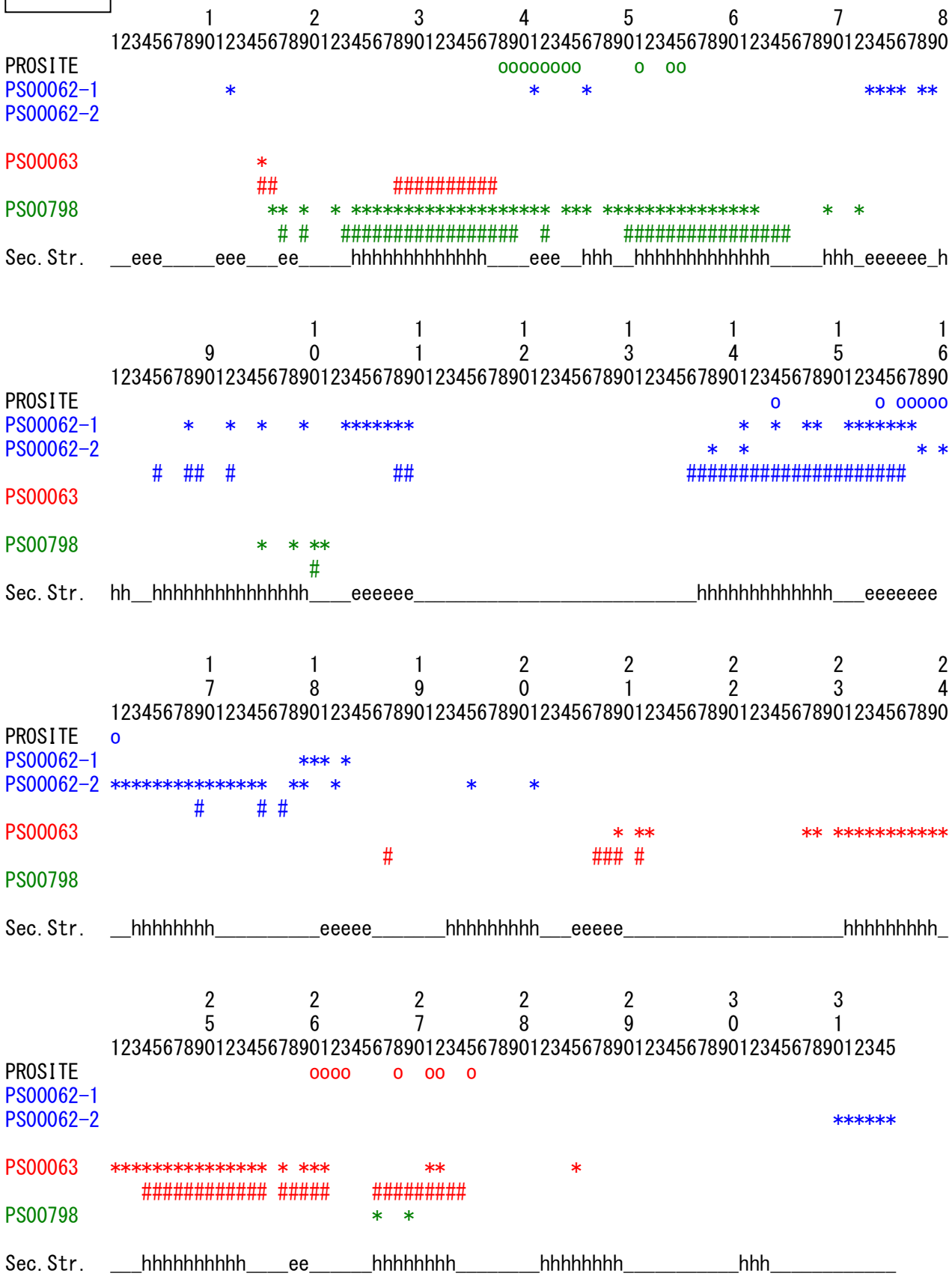


図１０

Commonly occurring local structures in Fr-1 protein (1frb)



o o o　*PROSITE Patterns:　PS00062(blue), PS00063(red), PS00798(green).*
* * *　*common in homologous proteins.*
#　*common in non-homologous proteins.*

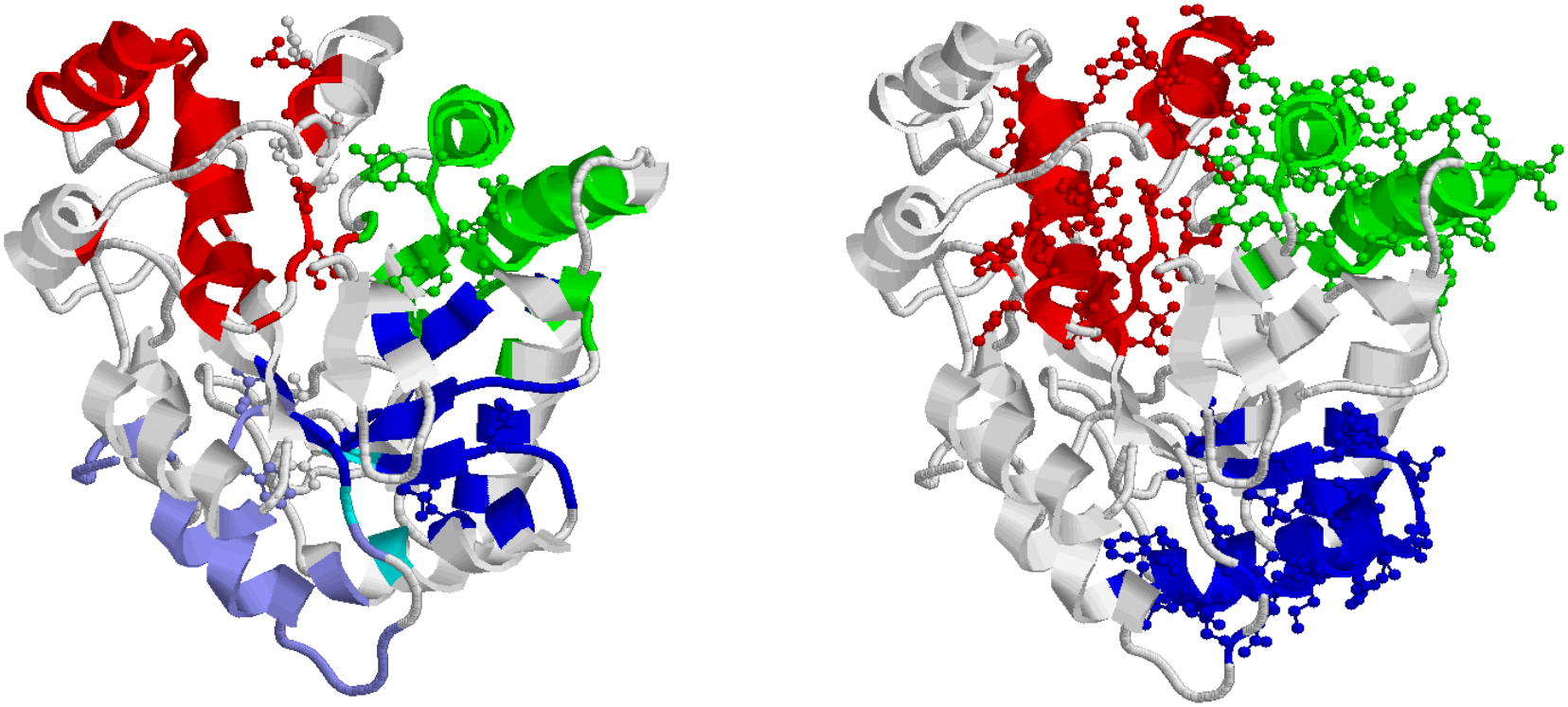


図 11. 相同タンパク質間で共通に見られる局所構造（左）と非相同タンパク質間で共通に見られる局所構造（右）
図 4-10 で示した 3 つの PROSITE モチーフに対応する。

この方法で何が得られるか

１．相同タンパク質間で共通に見出される局所構造

1.1　Rigid な局所構造

いくつか相同タンパク質間で共通に同定される局所構造は、互いに非常によく似ており、その意味で Rigid な局所構造と言える。しかも、もし、これらの局所構造が非相同タンパク質には見出されないならば、そのファミリーに特異的な、重要な局所構造モチーフの可能性が高い。

1.2　自動的な立体構造アラインメント

四面体で定義された２つの等しい局所構造の各頂点の対応関係は、一意的に定まっている。したがって、局所構造のアラインメントが同時に実行されたことになる。

２．非相同タンパク質間で共通に見出される局所構造

2.1　局所構造モチーフ

異なるファミリーに属するタンパク質に比較的頻繁に、共通に見出される局所構造は、モチーフと呼ぶことができる。見出されるファミリーが多ければ多いほど、タンパク質立体構造構築に基本的な、あるいは重要な局所構造とみなすことができよう。

2.2　各サイトでのアミノ酸の出現頻度

局所構造モチーフはそれぞれコードを与えられた四面体で定義されるため、各四面体の頂点は一義的に定義され、各頂点におけるアミノ酸の出現頻度分布によって、その構造を特徴づけることができる。

Protein Engineering vol.11 no.11 pp.981–990, 1998

Novel method to detect a motif of local structures in different protein conformations

Hiroshi Wako[1] and Takahisa Yamato[2]

School of Social Sciences, Waseda University, Shinjuku-ku, Tokyo 169-8050 and [2]Faculty of Science, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan

[1]To whom correspondence should be addressed

In order to detect a motif of local structures in different protein conformations, the Delaunay tessellation is applied to protein structures represented by C_α atoms only. By the Delaunay tessellation the interior space of the protein is uniquely divided up into Delaunay tetrahedra whose vertices are the C_α atom positions. Some edges of the tetrahedra are virtual bonds connecting adjacent residues' C_α atoms along the polypeptide chain and others indicate interactions between residues nearest neighbouring in space. The rules are proposed to assign a code, i.e., a string of digits, to each tetrahedron to characterize the local structure constructed by the vertex residues of one relevant tetrahedron and four surrounding it. Many sets comprised of the local structures with the same code are obtained from 293 proteins, each of which has relatively low sequence similarity with the others. The local structures in each set are similar enough to each other to represent a motif. Some of them are parts of secondary or supersecondary structures, and others are irregular, but definite structures. The method proposed here can find motifs of local structures in the Protein Data Bank much more easily and rapidly than other conventional methods, because they are represented by codes. The motifs detected in this method can provide more detailed information about specific interactions between residues in the local structures, because the edges of the Delaunay tetrahedra are regarded to express interactions between residues nearest neighbouring in space.

Keywords: Delaunay tessellation/residue–residue interaction/structural classification/structural codes/structural motifs

Introduction

In order to understand protein structure, a study of commonly occurring local structures in different proteins, which we refer to as a motif, is very important. It is well known that the three-dimensional (3D) structure of proteins has been better conserved during evolution than amino acid sequence. A simple relationship between sequence identity and structural similarity has been found among homologous proteins (Chothia and Lesk, 1986; Sander and Schneider, 1991; Flores *et al.*, 1993; Chelvanayagam *et al.*, 1994). Even if it is hard to show statistically significant sequence similarity, evolutionary relationships between distantly related proteins can be inferred based on 3D structural similarity (Johnson *et al.*, 1990a,b). Several structural alignment methods, which take into account several structural properties besides the amino acid sequence, have been also proposed to improve accuracy of the sequence alignment (Sali and Blundell, 1990; Flores *et al.*, 1993; Holm and Sander, 1993; Chelvanayagam *et al.*, 1994; Russell and Barton, 1994). To go a step further, methods have been developed for systematic structure comparison among not only homologous but also non-homologous proteins. They are suitable for databank searches and clustering (Taylor and Orengo, 1989; Alexandrov *et al.*, 1992; Holm and Sander, 1993, 1996; Alexandrov and Go, 1994; Mizuguchi and Go, 1995; Orengo and Taylor, 1996).

Since most motifs play a functionally or structurally important role, the motif searches may be expected not only to give insight into the relationship between proteins and their possible evolutionary origins, but also to deepen our understanding of the relationships between the amino acid sequence and the 3D structure. Accumulation of many motifs can serve as a database that is helpful for homology modelling, *ab initio* prediction of structure from sequence, and *de novo* design of proteins. Our purpose in this paper is to propose a novel method to detect a motif of local structures in different protein conformations.

It is usually more comprehensive to assume that protein structures are arranged in hierarchical fashion, i.e., amino acid sequence, secondary structure, supersecondary structure, domain and tertiary structure (Schurtz and Schirmer, 1979; Richardson, 1981). Accordingly, it is reasonable to define motifs at each level of this hierarchy. At secondary structure level α-helix, β-strand and various kinds of turns are regarded as motifs. At supersecondary structure level there are various motifs assembling a couple of secondary structure elements such as β-α-β, α-turn-α, parallel- and antiparallel-β-sheet and α-helix bundle. At domain level the motif is a more ambiguous concept. Their classification has not been established yet. In fact, even in FSSP (Holm and Sander, 1994), SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1993, 1996), well-known domain classification databases, their classifications do not coincide with each other for some proteins at present. The motifs in which we are interested are at secondary and supersecondary structure levels, or at a structure level classified in between. As a matter of fact, some of the motifs defined in this paper are parts of secondary and supersecondary structures, and others are structures of similar size not directly related to the secondary structures. We refer to these structure as local structures. The method presented here can provide more detailed information about the motifs at such a level.

A view of protein structures as an assembly of secondary structures is the most popular. Such a picture is considered reasonable, but not trivial. In fact, a module proposed by Go (1985) is defined based on the compactness of a contiguous local region of protein irrespective of the secondary structures. In the method proposed here we do not presuppose the existence of the secondary structures either. We think that the method to detect motifs without the presupposition of secondary structures is significant, even if it may be shown after the analyses that secondary structures are essential elements to describe the protein structure. This attitude is also useful to

analyse the regions with ambiguously defined structures such as N- and C-terminals of secondary structures and irregular (but possibly definite) loop structures.

It is also optional in the analysis of the motifs whether or not they are defined as a contiguous region along the polypeptide chain. In this study we do not give such a restriction to the detection algorithm explicitly. As described below, we will focus our attention on the network of spatially nearest neighbouring residues in protein structures. However, information about sequential connectivity of the amino acid residues is included implicitly through the unique numbering method of vertex residues on the Delaunay tetrahedron proposed below.

In this paper we apply the Delaunay tessellation to protein structures to detect a motif of local structures. Delaunay tessellation has been used for structural analyses of various disordered systems. In most such cases it has served as a valuable tool for structure description (Voloshin *et al.*, 1989; Vaisman *et al.*, 1994). As for protein structures, Yamato *et al.* (1994) employed the Delaunay tessellation to analyse a thermally fluctuating protein structure in molecular dynamics simulation. It was also applied to the analyses of pressure-induced deformations of proteins derived from normal mode analysis (Yamato, 1996; Kobayashi *et al.*, 1997). In these studies the Delaunay tetrahedra are used to define topographical structures and metric of the protein molecule at atomic level.

Recently, Singh *et al.* (1996) made statistical analysis of residue composition of Delaunay tetrahedra and revealed that non-random preferences for certain quadruplets of amino acids to be clustered together. This work was developed further by the same group (Munson and Singh, 1997) to derive empirical multi-body potentials and to apply the obtained potentials to sequence–structure alignment.

The Delaunay tetrahedron is composed of four vertices (C_α atom positions in this study) and six edges connecting them. Since the edges connect essentially the nearest neighbouring C_α atoms in space, the characterization of the networks of the edges can be used to define types of local structures with respect to residue–residue interactions. For such a characterization we propose a method to assign a code, i.e., a string of digits, to each Delaunay tetrahedron, and then show that the local structures with the same code are similar enough to each other to represent a motif.

Materials and methods

Delaunay tessellation

In this work the Delaunay tessellation is applied to the analysis of protein structures. By representing amino acid residues in the protein molecule only by C_α atoms, the protein structure is described as a set of points (C_α atoms) in 3D space. The Delaunay tessellation of the protein structure generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices (we refer to them as Delaunay tetrahedra or simply as tetrahedra in this paper). To explain the Delaunay tessellation we first describe a related geometric construct, the Voronoi polyhedron. That is, an entire 3D volume is divided into non-overlapping regions called Voronoi polyhedra, each of which is defined as a set of points closest to one particular particle (C_α atom in this case) of the protein. The boundary points of the Voronoi polyhedra are thus equally distant from two particles. Particles whose Voronoi polyhedra share a common boundary are said to be in contact or nearest neighbours of each other. If we connect particles in contact with a line segment, there appear the Delaunay simplices composed of the vertices (i.e., the C_α atoms) and the line segments connecting them. It is well known that the Delaunay simplices in 3D space are always tetrahedra. The complete set of tetrahedra divides up the interior space of the protein into non-overlapping volume elements. It is called Delaunay tessellation. This tessellation uniquely defines all the Delaunay tetrahedra for a given protein structure.

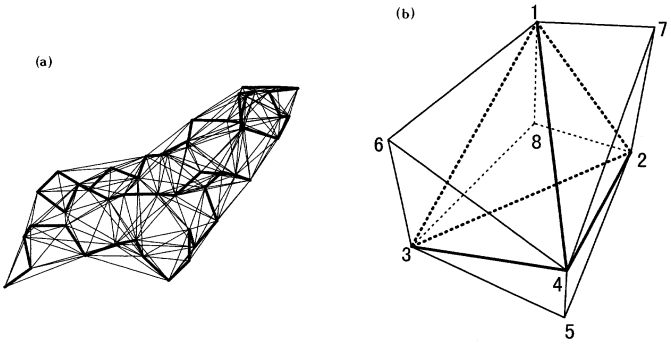


Fig. 1. An example of the Delaunay tessellation of protein and vertex numbering. (**a**) A thick line is a C_α trace of human transforming growth factor α, which consists of 50 amino acid residues (PDB entry ID is 4tgf; Kline *et al.*, 1990). A thin line is an edge of the Delaunay tetrahedron generated by the Delaunay tesselation. (**b**) A central tetrahedron 1234 and four surrounding tetrahedra, 2345, 1346, 1247 and 1238. To the four vertices of the central tetrahedron, digits, 1, 2, 3 and 4, are assigned in increasing order of the corresponding vertex residues' numbers. To the four vertices of the surrounding tetrahedra besides the common vertices with the central one, digits, 5, 6, 7 and 8, are assigned just as shown in this figure. The vertex residues 5, 6, 7 and 8 do not always exist.

In order to calculate the Delaunay tessellation we use the program Qhull (Barber *et al.*, 1995) obtained by anonymous ftp (geom.umn.edu/pub/software). In the calculated tessellation for a given protein structure, however, we found that some pairs of C_α atoms too distant from each other are connected in some geometrically irregular regions, such as on the surface, active site crevasses and N- and C- terminals. Since we assume that the edges of the Delaunay tetrahedra connecting two amino acid residues reflect some interactions between these residues in this work, we took into account only the Delaunay tetrahedra in which all of the four edges are shorter than a given cut-off distance (10 Å in this study). An example of the Delaunay tessellation of a protein molecule is shown in Figure 1a.

The Delaunay tessellation is just a geometrical operation without any explicit consideration of properties specific to protein structures. It should be emphasized that the chain connectivity and secondary structures are not taken into computation of the tessellation, even though they are taken into account implicitly, because the residues in close contact are very likely to be connected as the edges of the Delaunay tetrahedra.

Code assignment to tetrahedron

Consider a Delaunay tetrahedron T_0. The amino acid residue numbers at the four vertices of the tetrahedron are denoted $v_1(T_0)$, $v_2(T_0)$, $v_3(T_0)$ and $v_4(T_0)$. Here we require the following rule for putting the suffixes 1 to 4, i.e., $v_1(T_0) < v_2(T_0) < v_3(T_0) < v_4(T_0)$. Accordingly, we can uniquely number four vertices of any Delaunay tetrahedron with the digits 1 to 4.

We also consider the surrounding tetrahedra which share one of the facets (triangles) of T_0. At most it is possible for four tetrahedra to exist, T_5, T_6, T_7 and T_8, although they do

not always do so. If they exist, we denote the residue numbers at the vertices of these tetrahedra not contained by T_0, $v_5(T_0)$, $v_6(T_0)$, $v_7(T_0)$ and $v_8(T_0)$, respectively. The four tetrahedra, T_5, T_6, T_7 and T_8, are defined as sets of vertex residues, $\{v_2, v_3, v_4, v_5\}$, $\{v_1, v_3, v_4, v_6\}$, $\{v_1, v_2, v_4, v_7\}$ and $\{v_1, v_2, v_3, v_8\}$, respectively. It should be noticed that the suffixes 5 to 8 are also uniquely assigned to the vertex residues in relation to suffixes 1 to 4 (see Figure 1b). It should be also noticed that it frequently occurs that some of the vertices v_5 to v_8 share the same residue.

Then we assign a code, a string of digits, $c(T_0)$, to the tetrahedron T_0 according to the following rules:

(1) Arrange v_1 to v_8 in increasing order.

(1a) If $v_a = v_b$ and $a < b$ for a, $b \geqslant 5$ (i.e., some vertex residues of T_a and T_b are coincident with each other), arrange v_a, v_b in this order.

(1b) If v_a for $a \geqslant 5$ does not exist (i.e., tetrahedron T_a does not exist), v_a is not included in the arrangement.

(2) If v_1 to v_8 are arranged as v_a, v_b, v_c, v_d, v_e, v_f, v_g, v_h (the digits 1 to 8 may be assigned to suffixes a to h in various permutation), code 'abcdefgh' is assigned to this tetrahedron. If the number of the vertices is less than 8, the number of digits in the code is also less than 8 according to the above rule (1b). Since the digits 1 to 4 of the suffixes always appear in this order, the code has at least four digits.

Let us show some examples. If $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\} = \{32, 34, 35, 36, 45, 59, 23, 22\}$, then the code is 87123456 ($v_8 < v_7 < v_1 < v_2 < v_3 < v_4 < v_5 < v_6$). For {13, 14, 16, 17, 15, 82, 82, 15}, the code is 12583467 ($v_1 < v_2 < v_5 = v_8 < v_3 < v_4 < v_6 = v_7$). This is the case where $v_5 = v_8$ and $v_6 = v_7$. For {51, 60, 62, 63,, 12,, 11}, the code is 861234 ($v_8 < v_6 < v_1 < v_2 < v_3 < v_4$). This is the case where v_5 and v_7 are missing. Some other examples can be found in Results.

There are 1680 possible codes of 8 digits. There are also 840, 180, 20 and 1 possible codes of 7, 6, 5 and 4 digits, respectively. The total number of possible codes is 2721.

As shown in Results, however, this code $c(T_0)$ is not enough to detect a motif of local structures. We found that it is better to take into account the tetrahedra T_5, T_6, T_7 and T_8, neighbouring T_0. The codes are assigned to these tetrahedra in the same manner as T_0. Including these four codes we assign a set of the codes $c(T_0)$: $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$ to the tetrahedron T_0. We refer to the former code as a single tetrahedron (ST) code, and to the latter as a nearest neighbour tetrahedra (NNT) code, in this paper.

In the NNT code, at most 20 residues are taken into account, since each of the four tetrahedra, T_5, T_6, T_7 and T_8, has three nearest neighbouring tetrahedra, T_9 to T_{11}, T_{12} to T_{14}, T_{15} to T_{17} and T_{18} to T_{20}, respectively, except for T_0. The residue numbers at the vertices of T_9 to T_{20} not included in T_5 to T_8 are referred to as v_9 to v_{20}, respectively. Since some of the residues v_9 to v_{20} do not exist, or share the same vertices similarly to v_5 to v_8, the net number of residues related to the NNT code is not necessarily equal to 20, but usually less than 20. This is the size of the local structures considered in this paper.

(The vertex residues for the tetrahedra T_9 to T_{20} are explicitly given in the followings: $T_9 = \{v_3, v_4, v_5, v_9\}$, $T_{10} = \{v_2, v_4, v_5, v_{10}\}$, $T_{11} = \{v_2, v_3, v_5, v_{11}\}$, $T_{12} = \{v_3, v_4, v_6, v_{12}\}$, $T_{13} = \{v_1, v_4, v_6, v_{13}\}$, $T_{14} = \{v_1, v_3, v_6, v_{14}\}$, $T_{15} = \{v_2, v_4, v_7, v_{15}\}$, $T_{16} = \{v_1, v_4, v_7, v_{16}\}$, $T_{17} = \{v_1, v_2, v_7, v_{17}\}$, $T_{18} = \{v_2, v_3, v_8, v_{18}\}$, $T_{19} = \{v_1, v_3, v_8, v_{19}\}$, $T_{20} = \{v_1, v_2, v_8, v_{20}\}$).

Table I. The most abundant ST codes and their residue number patterns

Code	Number of tetrahedra	Residue number pattern							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
68123457	18 094 (total)								
	14 493	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	573	i	i+1	j	k	k+1	i–1	k+1	i–1
	482	i	i+1	i+2	i+3	i+4	i–1	j	i–1
	475	i	i+1	i+2	j	k	l	m	n
	296	i	i+1	i+2	i+3	i+4	j	i+4	i–1
	1775	other patterns							
67125834	5584 (total)								
	1586	i	i+1	i+3	i+4	i+2	j	j	i+2
	1272	i	i+1	j	j+1	j–1	i–1	i–1	j–1
	1178	i	i+1	i+3	i+4	i+2	j	k	i+2
	1548	other patterns							
12583467	4762 (total)								
	924	i	i+1	j	j+1	j–1	k	m	j–1
	894	i	i+1	j	j+1	j–1	k	k	j–1
	847	i	i+1	j	j+1	k	j+2	m	i+2
	561	i	i+1	j	j+1	j–1	j+2	j+2	j–1
	419	i	i+1	i+3	i+4	i+2	i+5	i+5	i+2
	1117	other patterns							
125834	4614 (total)								
	3683	i	i+1	i+3	i+4	i+2	–	–	i+2
	931	other patterns							

Proteins analyzed

The proteins analysed in this paper are given in Appendix 1.

Results

By the Delaunay tessellation 208 434 tetrahedra are obtained from 293 proteins. The two kinds of codes, ST and NNT, are assigned to each tetrahedron according to the rules described above. Of the 2721 possible ST codes, 405 are not found, 206 are assigned only to one tetrahedron and 131 to two tetrahedra.

On the contrary, the most abundant ST code is 68123457, which is assigned to 18 094 tetrahedra. The next most abundant ST codes are 67125834, 12583467, 125834, 1258346 and 12683457, which are assigned to 5584, 4762, 4614, 1782 and 1771 tetrahedra, respectively.

In Table I the residue number patterns at the eight vertices v_1 to v_8, $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, denoted by $\{v_1$–$v_8\}$ hereinafter, are shown for the top four most abundant ST codes. The residue number patterns are used as implication of 3D structures at the first glance (the structures with the same residue number patterns are similar to each other in general; however, it is necessary to make certain of the structural similarity by their superposition). Residue number v_1 is set to i, and then the relative residue numbers are shown for v_2 to v_8. If the residue number is regarded to have no relation with v_1, different characters j, k, ... are used.

Table I shows that there are various residue number patterns for one code. As shown below, the first code 68123457 corresponds essentially to part of the α-helix structure. However, the second and third codes, 67125834 and 12583467, correspond to either part of the α-helix or β-sheet. This means that different local structures are assigned to the common ST code. Consequently, the ST code alone is not enough to distinguish the local structures.

Table II. The most abundant NNT codes and their residue number patterns

Code	Number of tetrahedra	Residue number pattern							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
125834:68123457:0:0:68123457									
	3355 (total)								
	3271	i	i+1	i+3	i+4	i+2	–	–	i+2
	78	i	i+1	j	j+2	k	–	–	j+1
	6	other patterns							
68123457:68123457:125834:125834:68123457									
	905 (total)								
	902	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	i	i+1	i+2	i+3	i+4	i–3	i+4	i–3
68123457:68123457:1258346:7125834:68123457									
	574 (total)								
	574	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
t68123457:68123457:125834:1258347:68123457									
	536 (total)								
	536	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
68123457:68123457:125834:67125834:68123457									
	378 (total)								
	375	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	other patterns							
68123457:68123457:125834:125834:68123457									
	360 (total)								
	357	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	other patterns							
123457:68123457:0:67125834:0									
	355 (total)								
	341	i	i+1	i+2	i+3	i+4	–	i+4	–
	14	other patterns							

In Table II the residue number patterns $\{v_1–v_8\}$ are given for the top seven most abundant NNT codes. Although each NNT code corresponds to only one residue number pattern in general, some codes correspond to more than one pattern. Even in such a case, however, the exceptional patterns are only small variations of the major one. Completely different patterns are very few. (The residue number patterns for the second to seventh codes shown in Table II are incidentally identical. There are various residue number patterns for other codes not shown here). Residue numbers v_9 to v_{20} are not shown in Table II, because it is too lengthy to show them and because they have more varieties of patterns in some cases.

In the following we demonstrate with some examples that motifs of local structures are well detected with the NNT codes. In the analyses we consulted the database SCOP (Murzin *et al.*, 1995) and examined the results with the help of the program, PROMOTIF (Hutchinson and Thornton, 1996).

α-Helix

The top 20 most abundant ST codes can be classified into two groups except for four codes. In one group the codes include a sequence 125834; for example, 67125834 (5584), 12583467 (4762), 125834 (4614), 1258346 (1782), 7125834 (1612), 61258347 (1532), 6125834 (1416), 1258347 (1413) etc. (the numbers of tetrahedra found with these codes, N_{ST}, are shown in the parentheses). In another group the codes include a sequence 1234: for example, 68123457 (18 094), 681234 (1458), 123457 (1377), 6812345 (1227), 8123457 (1045) etc. The four exceptional codes are 12683457 (1771), 68125734 (1427), 68123574 (1131) and 16823457 (989), which are discussed in the β-sheet subsection below.

The typical residue number pattern $\{v_1–v_8\}$ for the codes with 125834 is {i, i+1, i+3, i+4, i+2, j, k, i+2}. The pattern

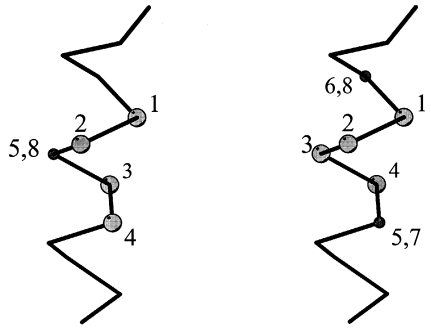


Fig. 2. Vertex residues on α-helix. (**a**) Code 125834; (**b**) code 68123457. Figures 2–8 were drawn by MOLSCRIPT (Kraulis, 1991).

for the code 68123457 is {i, i+1, i+2, i+3, i+4, i–1, i+4, i–1}. In dual expressions, the vertices v_1, v_2, $v_5 = v_8$, v_3 and v_4 correspond to the residue numbers, i to i+4, respectively, for the former code (symbol = is used to indicate that the two vertex residues, v_5 and v_8 in this case, are identical). v_6 and v_7 do not have definite residue numbers relative to v_1. For the latter code, the vertices $v_6 = v_8$, v_1, v_2, v_3, v_4, and $v_5 = v_7$ correspond to the residue numbers, i–1 to i+4, respectively. The local structures corresponding to both the codes are one to two turns of α-helix, as shown in Figure 2. In other words, the two abundant ST codes, 68123457 and 125834, are principal codes for the α-helix.

The most abundant NNT code,

NNT code = 125834: 68123457: 0: 0: 68123457
$\{v_1–v_8\}$ = {i, i+1, i+3, i+4, i+2,,, i+2}

contains the above two ST codes (see Table II). It represents part of the α-helix structure composed of tetrahedra T_0, T_5 and T_8. Tetrahedra T_6 and T_7 are missing. The residue numbers constructing T_0, T_5 and T_8 are (i, i+1, i+3, i+4), (i+1, i+2, i+3, i+4) and (i–1, i, i+1, i+2), respectively. $c(T_6) = c(T_7) = 0$ means that there are no residues near this part of the α-helix.

For comparison we give an example that has a code with the same $c(T_0)$, $c(T_5)$ and $c(T_8)$ as the above NNT code, but with nonzero $c(T_6)$ and $c(T_7)$ [$c(T_0)$ = 12583467 rather than 125834, because T_6 and T_7 exist]. The NNT code, the residue number pattern $\{v_1–v_8\}$, the number of tetrahedra with this NNT code found in the set of 293 proteins, N_{NNT}, and some examples are shown in Table III(a). Fifteen out of 40 structures, which are superposed with respect to the vertex residues v_1 to v_8, are shown in Figure 3. While residues $v_1–v_5$ and v_8 are on the same α-helix, $v_6 = v_7$ are remote from them along the chain ($v_6 = v_7 > v_1$). The residues v_9 to v_{20} and those neighbouring v_1 to v_{20} along the polypeptide chain are also included in Figure 3. Although all the residues are not well fitted to each other, because the superposition is performed with respect only to v_1 to v_8, these local structures seem to have a common feature with each other to represent a motif. This is supported by the fact that favorable amino acid types are limited at some vertices. For example, hydrophobic amino acid residues are favorable at $v_6 = v_7$ (out of 40 structures with this code, 13, 4, 4, 2, 3 and 2 are Ala, Val, Leu, Ile, Gly and Pro, respectively) and hydrophilic amino acid residues or those with a small side chain are favorable at v_2 (11, 5, 4, 2 and 3 are Glu, Asp, Lys, Arg and Asn, and 5 and 3 are Ala and Gly, respectively).

The residue number patterns of the vertex residues $v_6 = v_7$, v_{12} and v_{14} in the segments located on the right of the α-

Table III. Examples of motifs: NNT codes, residue number patterns and vertex residues[a]

Protein[b]	Vertex residue[c]																			
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{16}	v_{17}	v_{18}	v_{19}	v_{20}
(a) 12583467: 68123457: 16823745: 126834: 68123457, {v_1–v_8} = {i, i+1, i+3, i+4, i+2, j, j, i+2}, N_{NNT} = 40																				
3chy	66 L	67 E	69 L	70 K	68 L	101 A	101 A	68 L	71 T	71 T	66 L	102 G	67 E	98 A	–	69 L	–	70 K	65 G	65 G
1gox	235 A	236 E	238 A	239 R	237 D	275 A	275 A	237 D	240 L	240 L	235 A	276 A	236 E	272 V	–	238 A	–	239 R	234 T	234 T
1ads	28 T	29 E	31 V	32 K	30 A	57 A	57 A	30 A	33 V	33 V	28 T	58 I	29 E	54 V	–	31 V	–	32 K	27 V	27 V
5p21	69 D	70 Q	72 M	73 R	71 Y	103 V	103 V	71 Y	74 T	74 T	69 D	104 K	70 Q	100 I	–	72 M	–	73 R	68 R	68 R
1wsyA	82 F	83 E	85 L	86 A	84 M	121 V	121 V	84 M	87 L	87 L	82 F	122 G	83 E	118 C	–	85 L	–	86 A	81 C	81 C
(b) 12576834: 67125834: 16823457: 68123574: 12583467, {v_1–v_8} = {i, i+1, j+1, j+2, i+2, j, i+2, j}, N_{NNT} = 24																				
2mnr	266 L	267 A	294 M	295 S	268 M	293 P	268 M	293 P	269 P	266 L	269 P	314 H	314 H	267 A	294 M	244 Q	244 Q	292 I	295 S	292 I
5timA	124 I	125 A	164 I	165 A	126 C	163 V	126 C	163 V	127 I	124 I	127 I	209 L	209 L	125 A	164 I	92 V	92 V	162 V	165 A	162 V
4enl	318 V	319 A	342 L	343 L	320 D	341 A	320 D	341 A	323 T	318 V	323 T	370 M	370 M	319 A	342 L	295 E	295 E	339 A	343 L	339 A
1chrA	243 M	244 A	266 F	267 S	245 D	265 V	245 D	265 V	248 L	243 M	248 L	294 Y	294 Y	244 A	266 F	220 E	220 E	263 V	267 S	263 V
1btc	338 I	339 L	377 V	378 A	340 N	376 R	340 N	376 R	341 F	338 I	341 F	414 F	414 F	339 L	377 V	293 A	293 A	375 I	378 A	375 I
(c) 12357468: 67125834: 1782354: 12576834: 12368475, {v_1–v_8} = {i, i+1, i+2, j, j–1, k, j–1, k}, N_{NNT} = 6																				
5p21	5 K	6 L	7 V	56 L	55 I	75 G	55 I	75 G	8 V	5 K	8 V	71 Y	–	6 L	7 V	54 D	54 D	77 G	56 L	76 E
2cmd	2 K	3 V	4 A	31 S	30 L	70 A	30 L	70 A	5 V	2 K	5 V	66 A	–	3 V	4 A	29 E	29 E	72 V	31 S	71 D
1s01	27 K	28 V	29 A	91 Y	90 L	119 M	90 L	119 M	30 V	27 K	30 V	114 A	–	28 V	29 A	89 S	89 S	121 V	91 Y	120 D
3chy	7 K	8 F	9 L	34 E	33 V	51 Y	33 V	51 Y	10 V	7 K	10 V	45 K	–	8 F	9 L	32 N	32 N	53 F	34 E	52 G
1glt	3 K	4 L	5 G	36 H	35 L	80 L	35 L	80 L	6 I	3 K	6 I	77 L	–	4 L	5 G	34 E	34 E	82 V	36 H	81 D
(d) 81236457: 68172345: 68172345: 12583467: 12368475, {v_1–v_8} = {i, i+1, j, j+2, j+3, j+1, j+3, i–1}, N_{NNT} = 6																				
2reb	90 C	91 A	138 V	140 V	141 I	139 D	141 I	89 T	187 T	90 C	132 L	187 T	88 K	89 T	138 V	142 V	142 V	115 L	139 D	115 L
1glt	4 L	5 G	80 L	82 V	83 I	81 D	83 I	3 K	110 T	4 L	77 L	110 T	2 I	3 K	80 L	84 L	84 L	36 H	81 D	36 H
8atcA	157 V	158 A	222 V	224 I	225 L	223 D	225 L	156 H	261 M	157 V	219 M	261 M	155 L	156 H	222 V	226 Y	226 Y	185 Y	223 D	185 Y
1sbp	7 N	8 V	58 A	60 T	61 V	59 D	61 V	6 L	226 V	7 N	52 V	226 V	5 L	6 L	58 A	62 T	62 T	41 S	59 D	41 S
2cmd	3 V	4 A	70 A	72 V	73 V	71 D	73 V	2 K	112 A	3 V	67 L	112 A	1 M	2 K	70 A	74 L	74 L	31 S	71 D	31 S
(e) 68123457: 68123457: 61258347: 125834: 68123457, {v_1–v_8} = {i, i+1, i+2, i+3, i+4, i–1, i+4, i–1}, N_{NNT} = 41																				
4fxn	97 R	98 D	99 F	100 E	101 E	96 M	101 E	96 M	102 R	97 R	102 R	84 L	114 P	98 D	99 F	–	–	95 W	100 E	95 W
2cyp	89 F	90 K	91 F	92 L	93 E	88 G	93 E	88 G	94 P	89 F	94 P	46 L	108 F	90 K	91 F	–	–	87 N	92 L	87 N
4ts1A	99 A	100 R	101 I	102 K	103 E	98 S	103 E	98 S	104 Q	99 A	104 Q	69 V	122 N	100 R	101 I	–	–	97 W	102 K	97 W
2ctc	224 K	225 S	226 A	227 V	228 E	223 A	228 E	223 A	229 A	224 K	229 A	202 L	240 Y	225 S	226 A	–	–	222 V	227 V	222 V
3rubS	84 L	85 A	86 E	87 V	88 E	83 V	88 E	83 V	89 E	84 L	89 E	42 L	101 I	85 A	86 E	–	–	82 Q	87 V	82 Q
(f) 81623574: 67125834: 81276345: 16825734: 81276345, {v_1–v_8} = {i, j, j+1, j+4, j+3, i+3, j+3, i–1}, N_{NNT} = 9																				
3grs	35 S	346 G	347 R	350 A	349 L	38 R	349 L	34 A	348 K	35 S	348 K	351 H	39 A	34 A	347 R	153 L	153 L	343 I	38 R	31 G
1pda	234 M	287 G	288 I	291 A	290 L	237 R	290 L	233 A	289 S	234 M	289 S	292 E	238 L	233 A	288 I	261 A	261 A	284 E	237 R	230 A
1mioC	58 C	185 G	186 H	189 A	188 I	62 V	188 I	57 G	187 H	58 C	187 H	190 N	63 M	57 G	186 H	174 C	174 C	183 S	62 V	55 Y
1pfkA	19 A	264 A	265 S	268 G	267 M	22 G	267 M	18 A	266 R	19 A	266 R	269 A	23 V	18 A	265 S	122 L	122 L	261 R	22 G	15 G
2dri	19 L	237 P	238 D	241 G	240 I	22 G	240 I	18 S	239 Q	19 L	239 Q	242 A	23 A	18 S	238 D	88 L	89 D	236 L	22 G	15 F
(g) 681234: 0: 71258346: 0: 123457, {v_1–v_8} = {i, i+1, i+2, i+3,, i–1,, i–1}, N_{NNT} = 18																				
5fbpA	188 P	189 A	190 I	191 G	–	187 D	–	187 D	–	–	–	192 E	186 L	189 A	–	–	–	–	191 G	–
2gb1	47 D	48 A	49 T	50 K	–	46 D	–	46 D	–	–	–	51 T	45 Y	48 A	–	–	–	–	50 K	–
1aapA	25 V	26 T	27 E	28 G	–	24 D	–	24 D	–	–	–	29 K	23 F	26 T	–	–	–	–	28 G	–
8catA	171 P	172 Q	173 T	174 H	–	170 N	–	170 N	–	–	–	175 L	169 R	172 Q	–	–	–	–	174 H	–
1cseI	58 P	59 G	60 T	61 N	–	57 N	–	57 N	–	–	–	62 V	56 Y	59 G	–	–	–	–	61 N	–
(h) 571234: 61857234: 0: 18723564: 0, {v_1–v_8} = {i, i+1, i+2, i+3, j,, j, }, N_{NNT} = 10																				
2cp4	313 K	314 K	315 G	316 D	301 L	–	301 L	–	300 I	313 K	302 T	–	–	–	–	315 G	312 L	304 D	–	–
2hpdA	348 E	349 K	350 G	351 D	335 A	–	335 A	–	334 Y	348 E	336 K	–	–	–	–	350 G	347 L	338 D	–	–
1glaF	114 K	115 V	116 G	117 D	61 A	–	61 A	–	60 V	114 K	62 P	–	–	–	–	116 G	113 V	64 D	–	–
1hdxA	84 K	85 P	86 G	87 D	73 V	–	73 V	–	72 I	84 K	74 E	–	–	–	–	86 G	83 V	76 V	–	–
1plc	65 A	66 K	67 G	68 E	31 N	–	31 N	–	30 K	65 A	32 N	–	–	–	–	67 G	63 L	35 F	–	–
(i) 8123574: 67125834: 0: 86125734: 6182345, {v_1–v_8} = {i, i+1, i+2, j, j–1,, j–1, k}, N_{NNT} = 7																				
2sn3	40 Y	41 C	42 Y	47 W	46 C	–	46 C	20 G	45 A	40 Y	45 A	–	–	–	42 Y	39 G	25 C	19 L	–	21 E
1ixa	61 S	62 C	63 K	72 W	71 C	–	71 C	51 C	70 E	61 S	70 E	–	–	–	63 K	60 G	56 C	50 Q	–	54 N
1pnh	18 L	19 G	20 K	29 V	28 C	–	28 C	9 Q	27 E	18 L	27 E	–	–	–	20 K	17 L	12 C	8 C	–	13 R
2tgi	77 C	78 C	79 V	112 S	111 C	–	111 C	45 A	110 K	77 C	110 K	–	–	–	79 V	76 P	48 C	44 C	–	46 G
1hcc	46 K	47 C	48 L	53 S	52 W	–	52 W	26 G	51 K	46 K	51 K	–	–	–	48 L	45 A	8 P	25 Y	–	27 E
1avdA	82 Q	83 C	84 F	94 K	93 L	–	93 L	61 Q	92 V	82 Q	92 V	–	–	–	84 F	81 G	64 F	4 C	–	62 P
1ppn	109 G	110 V	111 R	210 V	209 P	–	209 P	73 Q	208 Y	109 G	208 Y	–	–	–	111 R	107 T	72 L	69 W	–	76 A

[a]Only five examples are given for each motif except for (i).
[b]Protein names are given by PDB entry codes.
[c]Underlined residues are discussed in the text.

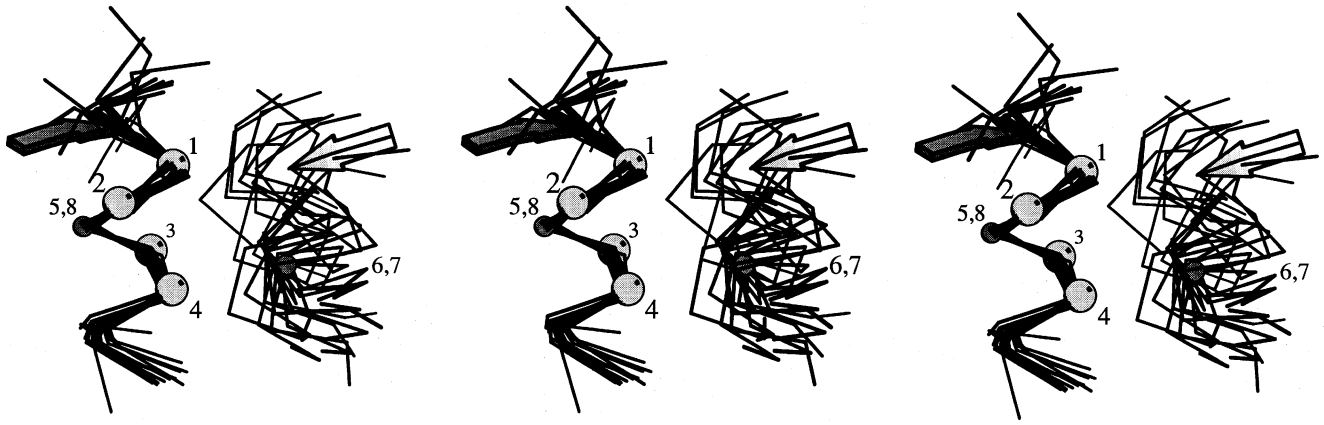


Fig. 3. Stereoscopic view of the local structures with NNT code 12583467: 68123457: 16823745: 126834: 68123457. Fifteen structures with this code are superposed with respect to v_1 to v_8. The residues corresponding to v_1 to v_{20} [given in Table III (a)] and additional residues neighbouring them along the polypeptide chain are included in the structures drawn here. The four large balls indicate the residues v_1 to v_4. The small balls indicate the residues v_5 to v_8. In this case, however, v_5 and v_8, and v_6 and v_7 are the same residues, respectively. The balls indicate the locations of the corresponding residues for the first protein in the list of Table III (a). The arrows do not represent β-sheet, but indicate the direction of the polypeptide chains.

helices in Figure 3 are ($v_6 = v_7$, v_{12}, v_{14}) = (j, j+1, j–3). This pattern implies that the structures of the segments are also α-helix. In actual fact, they are part of the α-helix. However, some are C-terminal and others are in the middle part of the α-helix. In addition, their relative spatial positions to the α-helices on the left differ (remember that the superposition of the structures was performed with respect only to the vertex residues v_1 to v_8). Nonetheless, the fact that favorable amino acid types are limited at $v_6 = v_7$, as described above, indicates an importance of some particular interactions of the residues at $v_6 = v_7$ in the segment on the right with the residues in the α-helix on the left in Figure 3.

The second to seventh NNT codes in Table II commonly include the ST code 68123457 for tetrahedra T_0, T_5 and T_8, while either $c(T_6)$ or $c(T_7)$ is slightly different from each other. The varieties in $c(T_6)$ and $c(T_7)$ reflect the differences in the environment of the α-helices. This fact indicates that the α-helices can be classified according to the NNT codes that reflect the residues surrounding them.

The codes related to N- and C-terminal structures of the α-helix are also found. An example for the C-terminal one is:

NNT code = 12583467: 681234: 57168234: 12683457: 68123457,
$\{v_1–v_8\}$ = {i, i+1, i+3, i+4, i+2, i+5, i+5, i+2}.

(data is not shown here). $N_{NNT} = 26$. Residues v_1, v_2, v_3 and v_5 take part in forming α-helix, and residue v_4 terminates it. v_4 is a special residue. Out of 26, 21 are Gly, three are Asn, and the remaining are Asp and His. These are a typical example of Gly-based motifs that cap the C-terminal end of α-helices, so-called C-cap of α-helices (Harper and Rose, 1993; Aurora *et al.*, 1994).

Residues v_{12} and v_{14} are on another segment, similar to Figure 3. The structures of these segments are much more different from each other than those in Figure 3, while the structures of C-terminal regions, where residues v_1, v_2, v_3 and v_5 are located, are well fitted to each other. The residue number patterns of the segments have also more varieties; (v_{12}, v_{14}) = (i, i), (i, i+2), (i, i+3), (i, i+4), (i, i+5) etc.

β-Sheet

There are many codes to represent parallel and antiparallel β-sheet. Similar to the α-helix, these codes correspond to parts of the β-sheets. The most abundant ST codes related to the β-sheet are 67125834, 12583467, 12683457, 68125734, 68123574 and 12576834 (N_{ST}= 5584, 4762, 1771, 1427, 1131 and 813, respectively). Some of them have already appeared above in the top 20 most abundant ST codes. As shown in Table I, the first two codes correspond to part of the α-helix, too. The residue number patterns for these six codes that correspond to the β-sheet (other patterns that do not correspond to the β-sheet for these codes are omitted here) are $\{v_1–v_8\}$ = {i, i+1, j, j+1, j–1, i–1, i–1, j–1}, {i, i+1, j, j+1, i+2, j+2, j+2, i+2}, {i, i+1, j, j+1, j+2, j–1, j+2, j–1}, {i, i+1, j, j+1, i+2, i–1, i+2, i–1}, {i, i+1, i+2, j, j–1, k, j–1, k}, and {i, i+1, j, j+1, i+2, j–1, i+2, j–1}, respectively. Here, the residues, i and j, are on different β-strands.

An example for parallel β-sheet is given in Figure 4 and Table III (b). The residues on the two parallel β-strands, (i, i+1, i+2) and (j, j+1, j+2), are (v_1, v_2, $v_5 = v_7$) and ($v_6 = v_8$, v_3, v_4), respectively. There are four strands in Figure 4. Residues v_1 to v_8 are on the two central strands, while residues $v_{12} = v_{13}$ and $v_{16} = v_{17}$ are on the left and right strands, respectively. The folding types of proteins are mainly TIM barrels. α/β and all-β proteins are also included.

For local structures related to the parallel β-sheet, we found the cases where limited amino acid types are favoured by some particular vertices. Let us show two examples. [The structures are not shown here. Only the residues corresponding to v_1 to v_{20} are shown in Table III (c) and (d)]. In Table III (c) five of six v_1's are Lys, and the remaining one is Thr. The residues for v_2 and v_5 are hydrophobic; out of six, two, two, one and one are Leu, Val, Ile and Phe for v_2, and four, one and one are Leu, Val and Ile for v_5, respectively. In Table III (d) five of six v_6's are Asp, and the remaining one is Gly. Incidentally the folding types of all the proteins containing these motifs are α/β.

Next we give an example for anti-parallel β-sheet:

NNT code = 67125834: 6812345: 68125734: 8123574: 12683457,
$\{v_1–v_8\}$ = {i, i+1, j, j+1, j–1, i–1, i–1, j–1}

(the data is not shown here). $N_{NNT} = 39$. The residues on the two anti-parallel β-strands, (i–1, i, i+1) and (j+1, j, j–1) are ($v_5 = v_6$, v_1, v_2) and (v_4, v_3, $v_5 = v_8$), respectively. This motif consists of four strands. Residues v_1 to v_8 are on the two central strands, while residues v_{17} and v_9 are on the outside

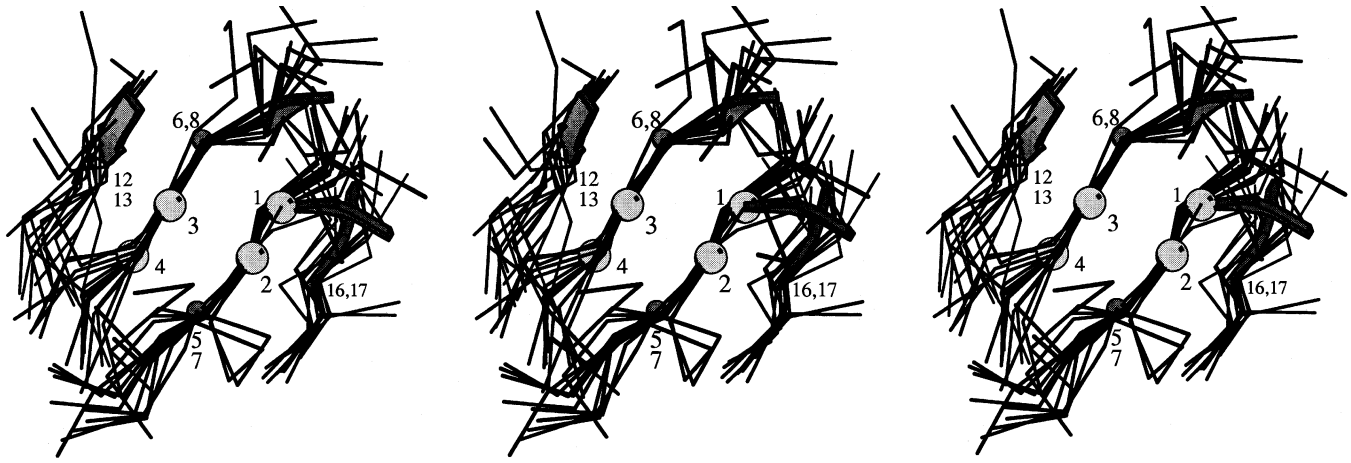


Fig. 4. Stereoscopic view of the local structures with NNT code 12576834: 67125834: 16823457: 68123574: 12583467. The amino acid residues are shown in Table III (b). Fourteen structures are superposed. See also the caption of Figure 3.

Fig. 5. Stereoscopic view of the local structures with NNT code 68123457: 68123457: 61258347: 125834: 68123457. The amino acid residues are shown in Table III (e). Thirteen structures are superposed. See also the caption of Figure 3.

strands. As for folding classes of the proteins, α+β, all-β and multi-domain are dominant.

Supersecondary structures

By the term supersecondary structure we mean a local structure composed of more than one secondary structure element, α-helix and/or β-strand in this paper. Three examples are given here. (Although the β-sheets are a supersecondary structure in this sense, they have been shown above according to the conventional classification).

The β-α-β is the most typical supersecondary structure (Richardson, 1981). Figure 5 and Table III (e) show an example related to this motif. 68123457 is a principal code of α-helix as described above. The two groups of β-strands lay under the α-helix in Figure 5. The vertex residues v_{12} and v_{13} are on the left and right groups of the β-strands, respectively. The relatively wide distribution of the β-strands reflect the fact that their relative spatial positions to the α-helices are not determined strictly in the β-α-β motif. As a matter of course, folding classes of most proteins are α/β.

Two anti-parallel α-helices is found with the following code:

NNT code = 7128354: 71823546: 0: 1682354: 1283457,
$\{v_1–v_8\}$ = {i, i+1, j, j+4, j+1,, i–3, i+4},

N_{NNT} = 12. There is one exceptional case of $\{v_1–v_8\}$ = {i, i+1, j, j+4, j+1,, i–4, i+4}. The folding types of the whole proteins are all-α.

In Figure 6 and Table III (f) an example related to two parallel α-helices over a β-strand is shown. There is one exceptional case of $\{v_1–v_8\}$ = {i, j, j+1, j+4, j+3, i+4, j+3, i–1} for 1mioC. A one-residue insertion caused this difference. It is observed in the lower part of the right α-helix in Figure 6. While most of the vertex residues v_1 to v_{20} are on either of the two α-helices, residues v_{16} = v_{17} are on the β-strand under the α-helices in Figure 6. It is remarkable that the amino acid type of v_2 is Gly for seven of nine, and those of the remaining two are Ala and Pro. The amino acid residues with a small side chain are favorable for v_4 and v_8. The amino acid type of v_4 is Ala for five of nine and those of the remaining four are Gly, Val and Ile. The amino acid type of v_8 is Ala and Gly for four and three of nine, respectively. Those of the remaining two are Ser and Thr. The folding classes of the proteins are either α/β or α+β.

Miscellaneous motifs

There are many interesting motifs not related to secondary structures, each of which is a set of local structures with the same code. Only four examples are shown here.

Table III (g) shows an example related to a turn (the structure is not shown here). Residues v_1 to v_4 form a β-turn of type I (type IV in few cases). The β-strands preceding and succeeding this turn form an antiparallel β-sheet with each other. The residues Asp and Asn are favorable at v_6 = v_8.

Figure 7 and Table III (h) show an example related to a

Fig. 6. Stereoscopic view of the local structures with NNT code 81623574: 67125834: 81276345: 16825734: 81276345. The amino acid residues are shown in Table III (f). Nine structures are superposed. See also the caption of Figure 3.

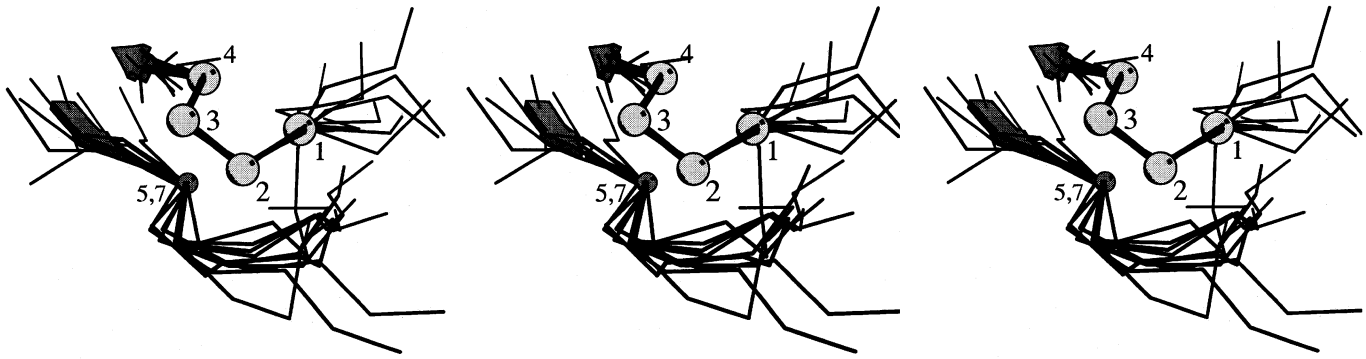


Fig. 7. Stereoscopic view of the local structures with NNT code 571234: 61857234: 0: 18723564: 0. The amino acid residues are shown in Table III (h). Ten structures are superposed. See also the caption of Figure 3.

β-bulge, which is an irregular region in a β-sheet lacking the normal pattern of hydrogen bonding. The type of the β-bulge shown here is called antiparallel (A) and G1 (G) according to Chan *et al.* (1993) (the type was defined by the program PROMOTIF [Hutchinson and Thornton (1996) in this study]). The three key residues, called X, 1 and 2, are $v_5 = v_7$, $v_3 = v_{15}$ and v_4, respectively. The fact that residue 1, i.e. $v_3 = v_{15}$, is Gly is a characteristic point for the β-bulge of type AG. It is also remarkable that the amino acid type of v_4 is Asp for seven of ten, and the remaining are Glu and Gln.

Ca^{2+} binding sites of the EF hand, i.e., loop structures connecting the E and F α-helices in Ca^{2+} binding proteins (Kretsinger, 1980) are found;

NNT code = 7182346: 0: 6178234: 7128354: 1823467,
$\{v_1–v_8\} = \{i, i+3, i+4, i+5,, i+6, i–3, i+2\}$

(the data is not shown here). $N_{NNT} = 6$. The residues at vertices $v_{17} = v_{20}$, $v_8 = v_{14}$, v_3 and v_{19}, usually called X, Y, Z and –Y, respectively, provide oxygen atoms to chelate the Ca^{2+} ion. It is known that the Gly residue at v_4 is highly conserved in different Ca^{2+} binding proteins. The Asp and Asn residues at v_1, v_3 and $v_8 = v_{14}$, and the hydrophobic residues at $v_{13} = v_{16}$ are also well conserved.

Figure 8 and Table III (i) are an example related to structures including disulfide bridges. The fold of 2sn3 and 1pnh is classified as knottins (disulfide-bound fold containing β-hairpin with two adjacent disulfides), and that of 2tgi as cystine-knot cytokines (three disulfide links arranged in a knot-like topology) according to SCOP. 1ixa and 1hcc are also disulfide-rich small proteins. However, 1avdA and 1ppn are classified as rather different fold types than the others, although they also have disulfide bonds.

In 2sn3, 1ixa, 1pnh and 2tgi the Cys residues at v_5 and v_{17} form a disulfide bond, while the residues at these two site are not Cys in the other proteins. There are varieties of the formation of the disulfide bonds of the Cys residues at v_2. In the disulfide bonds 16–41 of 2sn3, 78–15 of 2tgi and 47–5 of 1hcc the counterpart Cys residues do not belong to this local structure of v_1 to v_{20}, while those in the disulfide bonds 62–51 of 1ixa and 83–4 of 1avdA are included in this local structure. There are other disulfide bonds: 8–26 of 1pnh and 44–109 of 2tgi. The local structure of 1ppn has no Cys residues, although this protein has three disulfide bonds in the other region. The fitting of the structures in Figure 8 is not good when compared with other examples shown in this paper. Nonetheless, these structures seem to have common characteristic features.

Discussion

In this paper we have proposed a novel method to detect a motif of local structures in protein conformations using the Delaunay tessellation. There are motifs at various levels, such as amino acid sequences, secondary structures, supersecondary structures, domains and tertiary structures. The motifs detected in this paper are parts of secondary and supersecondary structures, and other kinds of structures of similar size not directly related to the secondary structures. In this sense we call them motifs of local structures.

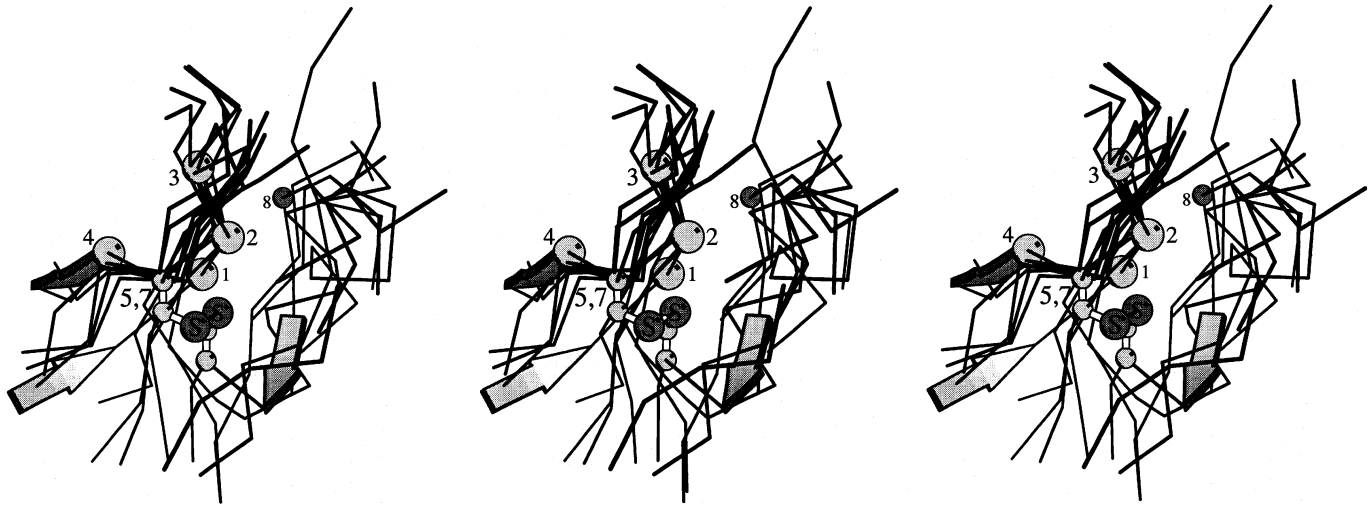


Fig. 8. Stereoscopic view of the local structures with NNT code 8123574: 67125834: 0: 86125734: 6182345. The amino acid residues are shown in Table III (i). Seven structures are superposed. See also the caption of Figure 3.

The Delaunay tessellation divides the protein interior into the Delaunay tetrahedra whose vertices are C_α atom positions uniquely. The edges of the tetrahedra can be regarded as an expression of the network of interactions between residues in 3D space. We intended to translate this 3D information about the network of the interactions into a one-dimensional sequence of digits, i.e., code, in this paper. We have achieved this by the unique numbering method of the vertices so that the code has information about the spatial arrangement of C_α atoms and the sequential connectivity of the chain.

The local structures with the same NNT code can be collected easily by computer. The structures with the same code resemble each other for most cases. Although the NNT code can include information about topographical structure of the residues, it cannot contain their 3D structure explicitly. Therefore, the fact that the local structures with the same code are similar to each other is not trivial. In fact, there are the cases where few exceptional structures are included even for the NNT codes. Consequently, superposition of the structures is necessary to assess the degree of resemblance. Nonetheless, the method of structural comparison based on topographical properties is important, if we consider the structural flexibility of protein molecules (Mizuguchi and Go, 1995).

The structures with the same code have a large variety in some cases (e.g., Figures 3, 5 and 8). Such a variety comes from the variety of relative spatial positions of the constituent segments and/or from the variety of structures of the segments in themselves. Nonetheless, the structures with the same code appear to have some common features. For example, favorable amino acid residues are limited at some particular vertices in Figure 3 and Table III (c) and (d), and the locations of disulfide bridges are common in Figure 8. β-α-β in Figure 5 and α-α on a β-strand in Figure 6 are typical supersecondary structures. The present method detects a motif of local structures from a topographical point of view rather than a superposition of the structures. This is the reason why such motifs can be detected by this method and one of the characteristic points of the method.

Although the final assessment of the degree of structural similarity should be carried out by their superpositions, a residue number pattern of the eight vertices v_1 to v_8 is useful for the first assessment. In most cases a given NNT code has only one residue number pattern, although there are some codes which correspond to more than one residue number patterns. A gap or insertion is usually responsible for this difference, but there are few cases that completely different structures corresponds to the same NNT code. Although there are some exceptions, the NNT code has enough ability to distinguish the local structures in general.

Since there are a huge number of NNT codes, it is impossible to see all local structures corresponding to every code on a graphic display at present. Currently we directed our attention to the most abundant codes and to the codes corresponding to the structures in which the amino acid types are likely to be limited at some particular vertices.

The most abundant codes are, as expected, related to α-helix and β-sheet. It should be emphasized, however, that the secondary structures are not taken into account explicitly in the Delaunay tessellation and the code assignment procedures. As far as the interactions between nearest neighbouring residues in space are taken into account, it is natural that these structures are identified as a motif. However, the motifs defined in this study differ from other motifs proposed up to now, even if our discussion is confined to the secondary structures. First, the local structures corresponding to the NNT codes are neither one whole α-helix nor one whole β-strand, but parts of the α-helix and the β-sheet. In addition some of the residues (i.e., structures they form) surrounding these parts are taken into account in the code. Consequently there are many codes related to the α-helix and the β-sheet reflecting the differences of structures located near the relevant secondary structures. More detailed classification of the α-helix and β-sheet based on such differences may be possible so that we can understand interactions between residues to stabilize these structures.

The local structures not related directly to the secondary structures are also detected as shown in the Miscellaneous motifs subsection of Results. As emphasized above, these structures can be detected, because we did not put any presupposition about secondary structures into the method. Again we want to emphasize that various kinds of motifs, such as those shown in Results, are detected in the same procedures, i.e., the Delaunay tessellation and code assignment to the Delaunay tetrahedron.

Representation of the local structure by the code is convenient in finding similar local structures by computer. At first the

Delaunay tessellation is applied to all proteins in the database, and then codes are assigned to the Delaunay tetrahedra obtained. The whole tetrahedra obtained for all the proteins in the database can be sorted easily with respect to the codes assigned, because the code is expressed as a number. As a consequence of the sorting, the tetrahedra, i.e., local structures, with the same code are collected automatically. We can use this set of sorted local structures as a new secondary database of protein structures.

This database can be used as follows. Suppose that we have a protein we wish to study. At first we can apply the Delaunay tessellation to this protein, and assign a code to every Delaunay tetrahedron obtained. Then, for each tetrahedron, i.e., for each local structure of this protein, we can easily derive the local structures with the same code from the database. Information about similar structures found in different proteins is helpful to understand and design them. However, it may be more helpful if we can know the different codes corresponding to similar local structures. In other words, clustering of the codes may be useful in such a database.

One of the difficulties of expressing local structures with a code is the problem of visualizing the local structures from that code. At present we have no idea about more comprehensive expression of the network of edges of the Delaunay tetrahedra. We may get more familiar with the codes in the future as the analyses of the codes progress.

There are huge number of codes. The analyses of local structures for whole codes have not been accomplished yet. Such analyses are now in progress.

References

Alexandrov,N.N. and Go,N. (1994) *Protein Sci.*, **3**, 866–875.

Alexandrov,N.N., Takahashi,K. and Go,N. (1992) *J. Mol. Biol.*, **225**, 5–9.

Aurora,R., Srinivasan,R. and Rose,G.D. (1994) *Science*, **264**, 1126–1130.

Barber,C.B., Dobkin,D.P. and Huhdanpaa,H.T. (1995) *ACM: Trans on Mathematical Software*, **22**, 469–483.

Chan,A.W.E., Hutchinson,E.G., Harris,D. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1574–1590.

Chelvanayagam,G., Roy,G. and Argos,P. (1994) *Protein Engng*, **7**, 173–184.

Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.

Flores,T.P., Orengo,C.A., Moss,D.S. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1811–1826.

Go,M. (1985) *Adv. Biophys*, **19**, 91–131.

Harper,E.T. and Rose,G.D. (1993) *Biochemistry*, **32**, 7605–7609.

Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138.

Holm,L. and Sander,C. (1994) *Nucleic Acids Res.*, **22**, 3600–3609.

Holm,L. and Sander,C. (1996) *Methods Enzymol.*, **266**, 653–662.

Hutchinson,E.G. and Thornton,J.M. (1996) *Protein Sci.*, **5**, 212–220.

Johnson,M., Sutcliffe,M. and Blundell,T.L. (1990a) *J. Mol. Evol.*, **30**, 43–59.

Johnson,M., Sali,A. and Blundell,T.L. (1990b) *Methods Ezymol.*, **183**, 670–690.

Kline,T.P., Brown,K., Brown,S.C., Jeffs,P.W., Kopple,K.D. and Mueller,L. (1990) *Biochemistry*, **29**, 7805–7813.

Kobayashi,N., Yamato,T. and Go,N. (1997) *Proteins Struct. Funct. Genet.*, **28**, 109–116.

Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.

Kretsinger,R.H. (1980) *Crit. Rev. Biochem.*, **8**, 119–174.

Mizuguchi,K. and Go,N. (1995) *Protein Engng*, **8**, 353–362.

Munson,P.J. and Singh,R.K. (1997) *Protein Sci.*, **6**, 1467–1481.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,T. (1995) *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.

Orengo,C.A., Mitchie,A., Jones,S., Jones,D., Swindells,M. and Thornton,J. (1996) *PDB Quarterly Newsletter*, **78**, 8–9.

Orengo,C.A. and Taylor,W.R. (1996) *Methods Enzymol.*, **266**, 617–635.

Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 167–339.

Russell,R. and Barton,G. (1994) *J. Mol. Biol.*, **244**, 332–350.

Sali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.

Sander,C. and Schneider,R. (1991) *Proteins Struct. Funct. Genet.*, **9**, 56–68.

Schulz,G.E. and Schirmer,R.H. (1979) *Principles of Protein Structure*. Springer-Verlag, Berlin.

Singh,R.K., Tropsha,A. and Vaisman,I. (1996) *J. Comp. Biol.*, **3**, 213–221.

Taylor,W.R. and Orengo,C.A. (1989) *J. Mol. Biol.*, **208**, 1–22.

Vaisman,I.I., Brown,F.K. and Tropsha,A. (1994) *J. Phys. Chem.*, **89**, 5559–5564.

Voloshin,V.P., Naberukhin,Y.I. and Medvedev,N.N. (1989) *Mol. Simulation*, **4**, 209–227.

Yamato,T. (1996) *J. Mol. Graph.*, **14**, 105–107.

Yamato,T., Saito,M. and Higo,J. (1994) *Chem. Phys. Lett.*, **219**, 155–159.

Received April 3, 1998; revised May 26, 1998; accepted June 15, 1998

Appendix 1.

The following 293 proteins are used for the analysis in this paper. They have less than 25% sequence homology with each other:

129l, 1aaf, 1aaj, 1aapA, 1aba, 1abk, 1abmA, 1add, 1ads, 1aep, 1alkA, 1aozA, 1apa, 1apmE, 1aps, 1arb, 1atnA, 1atx, 1avdA, 1avhA, 1baa, 1babB, 1bbl, 1bbpA, 1bbt1, 1bbt2, 1bgeB, 1bgh, 1bllE, 1bovA, 1brd, 1bsaA, 1btc, 1bw3, 1c5a, 1caj, 1cauB, 1ccr, 1cdb, 1cde, 1cdtA, 1chrA, 1cid, 1cmbA, 1cobA, 1colA, 1cpcA, 1cpcL, 1cpt, 1crl, 1cseI, 1ctaA, 1d66A, 1dfnA, 1dhr, 1dog, 1dsbA, 1eaf, 1eco, 1ede, 1erp, 1ezm, 1faiL, 1fas, 1fbaA, 1fc1A, 1fc2C, 1fdd, 1fha, 1fiaB, 1fod4, 1fxiA, 1gal, 1gatA, 1gdhA, 1gky, 1glaF, 1glaG, 1glt, 1gmfA, 1gox, 1gps, 1gsrA, 1hbq, 1hcc, 1hddC, 1hdxA, 1hgeB, 1hivA, 1hleA, 1hleB, 1hmy, 1hra, 1hsbA, 1huw, 1ipd, 1isuA, 1ixa, 1lcdA, 1le4, 1lenA, 1lgaA, 1lis, 1ltsA, 1ltsC, 1ltsD, 1mdaA, 1mdc, 1mioC, 1mrt, 1mup, 1mypC, 1nar, 1ndk, 1nipB, 1nrcA, 1nxb, 1ofv, 1omb, 1omf, 1omp, 1onc, 1osa, 1pda, 1pdc, 1pdgB, 1pfkA, 1phh, 1plc, 1pnh, 1poa, 1poc, 1poxA, 1ppfE, 1ppn, 1ppt, 1prcC, 1prcM, 1pyp, 1r094, 1r1a2, 1rcb, 1rec, 1rhd, 1ribA, 1rinB, 1rnd, 1rprA, 1rveA, 1s01, 1sbp, 1sgt, 1shaA, 1shfA, 1sltA, 1smrA, 1snc, 1spa, 1sryA, 1tabI, 1tbpA, 1ten, 1tfd, 1tfi, 1tgsI, 1tie, 1tlk, 1tml, 1tnfA, 1tplA, 1tpm, 1trb, 1troA, 1ttbA, 1ula, 1utg, 1vaaB, 1vil, 1vsgA, 1wsyA, 1wsyB, 1zaaC, 2aaiB, 2achB, 2atcB, 2ayh, 2azaA, 2bbkH, 2bds, 2bopA, 2bpa1, 2bpa2, 2bpa3, 2cas, 2cbh, 2ccyA, 2cdv, 2cmd, 2cp4, 2cpl, 2crd, 2cro, 2ctc, 2cts, 2cyp, 2dnjA, 2dri, 2ech, 2end, 2er7E, 2gb1, 2hipA, 2hpdA, 2ihl, 2lh2, 2liv, 2mev1, 2mhr, 2mnr, 2mrb, 2ms2A, 2msbA, 2mtaC, 2pcdA, 2pf1, 2pgd, 2pia, 2plv1, 2pmgA, 2por, 2reb, 2rn2, 2sas, 2scpA, 2sga, 2sim, 2sn3, 2snv, 2spo, 2stv, 2tbvA, 2tgi, 2tmvP, 2tscA, 2ztaA, 3adk, 3b5c, 3chy, 3cla, 3cox, 3dfr, 3egf, 3gapA, 3gbp, 3grs, 3inkC, 3monA, 3pgk, 3rubS, 3sdhA, 3sgbI, 3tgl, 4blmA, 4cpaI, 4enl, 4fgf, 4fxn, 4gcr, 4htcI, 4insB, 4rcrH, 4sbvA, 4sgbI, 4tgf, 4ts1A, 4xis, 4znf, 5fbpA, 5nn9, 5p21, 5timA, 5znf, 6taa, 7apiB, 8abp, 8atcA, 8catA, 8i1b, 8rxnA, 9ldtA, 9rnt, 9rubB, 9wgaA.