SHEETSPAIR: A DATABASE OF AMINO ACID PAIRS IN PROTEIN SHEET STRUCTURES

Ning Zhang[1]*, Jishou Ruan*[2]*, Jie Wu*[1]*, Tao Zhang*[1*]

[1]*College of Life Science, Nankai University, Tianjin, 300071, China*
Email: zhangtao@nankai.edn.cn; zhni@eyou.com; sohutu@mail.nankai.edu.cn
[2]*Chern Institute of Mathematics and College of Mathematical Sciences and LPKM, Nankai University, Tianjin 300071, China*
Email: jsruan@nankai.edu.cn

ABSTRACT

Within folded strands of a protein, amino acids (AAs) on every adjacent two strands form a pair of AAs. To explore the interactions between strands in a protein sheet structure, we have established an Internet-accessible relational database named SheetsPairs based on SQL Server 2000. The database has collected AAs pairs in proteins with detailed information. Furthermore, it utilizes a non-freetext database structure to store protein sequences and a specific database table with a unique number to store strands, which provides more searching options and rapid and accurate access to data queries. An IIS web server has been set up for data retrieval through a custom web interface, which enables complex data queries. Also searchable are parallel or anti-parallel folded strands and the list of strands in a specified protein.

Keywords: Protein structure; Protein sheet; Amino acids pairs; Parallel/anti-parallel folds

1 INTRODUCTION

Protein tertiary structure plays an essential role in deciphering protein functions, and knowledge of the structure can greatly assist biologists in the generation and testing of hypotheses, as well as in the design of drugs. However, only a small fraction of all proteins have a known three-dimensional structure published in the Protein Data Bank (PDB) (Kloczkowski, Ting, Jernigan, et al., 2002). Prediction of the tertiary structure of a protein from its amino acid sequence remains an important and difficult task. Moreover, the main mechanisms responsible for protein folding pathway have not yet been identified (Galzitskaya, Ivankov, & Finkelstein, 2001; Onuchic, & Wolynes 2004).

The three-dimensional tertiary structure of a protein consists of repeating units of secondary structure elements. The two major types of secondary structures are the alpha helix and beta sheet. During the folding process of a protein, a certain fragment might first adopt a secondary structure (e.g., an a-helix or a b-sheet), and later different secondary structures interact with each other and adopt tertiary folds. In this instance, the protein secondary structure has turned out to be important and more and more critical in solving these problems (Dayalan, Gooneratne, Bevinakoppa, et al., 2006). Today, many approaches for predicting secondary structures from amino acid sequence have been developed (Jones, 1999; Qiu, Liang, & Zou, 2003; Wang, Liu, Li, et al., 2004; Cheng, Sen, & Kloczkowski, 2005). However, there is still a long way to go to reach the tertiary structure and the protein folding mechanism after getting secondary structures by those prediction methods. Therefore, a great deal of fundamental information should emerge from known secondary structural elements because methods of prediction of protein tertiary structure and understanding of protein folding mechanisms may use as a starting point some information on the protein secondary structures (Rohl, Strauss, Misura, et al. 2004; Lee, Kim, & Lee, 2005; Lee, Kim, Joo, et al., 2004). Along with providing new insight in the known protein secondary structure, the information gained will help us see the world around us in a different way.

In the Beta-Sheet structure, the backbone of the polypeptide chain, which extends in a zigzag, can be arranged side by side to form a structure resembling a series of pleats, interacting with each other by hydrogen bonds. The individual segments, single lengths of the polypeptide chain that form part of a Beta-Sheet [MU1], called strands,can be quite distant from each other in the linear sequence of the polypeptide or even on different polypeptide chains. In addition, the strands can be arranged to form antiparallel or parallel sheets, having the opposite or the same amino-to-carboxyl orientations respectively. In such a structure, amino acids (AAs) on every two adjacent strands pair by twos. Between antiparallel or parallel strands, there is a different hydrogen bond interaction style between the carbonyl oxygen atom of one residue and the amino hydrogen atom of another in each paired AA. Thus AA pairs can be divided into three types according to the hydrogen bond interaction styles: Types I, II, and III (Figure 1). Types I and II are on antiparallel strands, having hydrogen bonds or not, respectively. Type III pairs are on parallel strands.

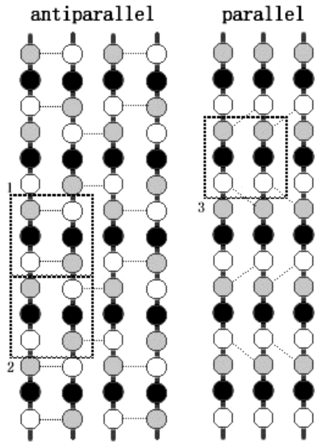

Figure 1. Antiparallel and parallel strands.
The three boxes show the three types of
amino acids pairs respectively.

This paper reports the development of a database that explores the interactions between the AA pairs on adjacent folded strands.

2 DESIGN AND IMPLEMENTATION

A database called SheetsPair has been developed and implemented based on SQL Server 2000 relational database system. An IIS web server has been set up for data retrieval through a custom web interface utilizing ASP language. The server runs the Windows 2000 Advanced Server operating system. The protein dataset utilized to populate the database was obtained from the PDB, excluding those that have no sheet structure and those with modified or uncertain residues or uncertain structures. The database is rigorous, precise, and accurate, with little uncertain or ambiguous factors.

3 DATA CONTENT

3.1 Amino acids pairs on strands

The database analyzes the raw data downloaded from PDB, reconstructing the sheet structures of each protein with the amino acids hydrogen bond interactions as in Figure 1. All amino acids pairs are registered in the database, which include detailed information of each pair such as PDB ID, chain identifier, position in chain, sheet identifier, index of strand, and pair type. Complex data queries on all the AA pairs can be made based on the detailed information. The database now contains a total of 756,897 amino acid pairs in sheet structures of 10,704 proteins.

When a strand bonds to another to form a sheet, usually the two strands do not have the same length and the ends of the strands do not align. That is to say, the splice segment is part of the strand but not the whole. All the sequences of those splice segments have been summarized in the database. Furthermore, the number of appearances of splice segments has been counted in all proteins.

3.2 Chain sequences

Each protein may have one or more chains. The database also houses the amino acid sequences of protein chains. Usually the sequence is stored in a free-text style as a character string, e.g. "ABCDEFG…," while the database utilizes a relational database structure to store the sequences instead of free text.

SheetsPair never uses a text character to identify an amino acid. Instead, it assigns an ID (a number) to the 20 amino acids. A database table "AAChar" stores the amino acid IDs. There is another database table called "AADesp" for the "description" of the 20 amino acids, including the three-letter code, the standard amino acid name, and even other properties such as the Eisenberg hydrophobic index (Eisenberg, Wilcox, & McLachlan, 1986). The hydrophobic index could be useful because hydrophobic interactions provide a major contribution to stabilizing protein structures, following Dill's view (Dill, 1990). Chain sequences are stored in another table named "Sequences." The table has four fields (PDBID, ChainID, Position, and AminoAcidID), with each record indicating one position in the sequence.

This database structure provides more searching options than the free-text style. For example, users are able to obtain the sequence of every segment part of a chain by simply giving the position range, while when using the free-text style, one gets the whole chain sequence and then must separate the segments manually. Furthermore, this structure avoids any potential sequencing errors or uncertainty factors caused by invalid characters in such long free-text strings because under this database structure, each amino acid ID indicates only one amino acid and any other invalid characters indicate nothing because there is no ID registered for them.

3.3 Secondary structures

The database marks the position and range of each Helix or Sheet segment in a chain. The PDB database uses SheetID and SheetNum to indicate a strand, e.g. "A 1", "A 2", etc. This may produce ambiguity in some specific structures, such as a barrel structure. In a barrel structure, anti-parallel (or parallel) sheets can roll up completely to join edges and form a cylinder or closed 'barrel,' in which the first strand is hydrogen bonded to the last. The strands form the "staves" of the barrel. The problem is that PDB uses a new SheetID and SheetNum to indicate the first strand when the last strand bonds to it. Because of this, the first strand has two sets of SheetID and SheetNum. That is to say, SheetID and SheetNum cannot be unique identifiers of a strand.

In contrast, the SheetsPair database assigns a unique number, called StrandID, to identify every strand in all proteins. It enables one to easily obtain a list of all strands in a certain chain as well as a chain list of a certain protein, without ambiguity. Each StrandID has been associated with the SheetID and SheetNum as a link to retrieving data from the PDB database.

Strands can be arranged to form antiparallel or parallel sheets; in other words, given a certain strand, it may participate in an antiparallel or parallel sheet. Given a certain strand, how can one know whether it participates in an antiparallel or parallel sheet? The database has used statistics and marks every strand with its participation style. Note that some strands may participate in both antiparallel and parallel sheets. Furthermore, because chain sequences are not stored in free-text, one can readily obtain the strand sequences in a list of the strands.

4 SIMPLE STATISTICAL ANALYSIS

To show the significance of the database, specific statistical analysis has been done. The number of appearances of the 20 amino acids and of 210 different pairs in all AA pairs collected in the database has been summarized. The results show that the amino acids V, L, I, A and pairs I:V, L:V, I:L, V:V, A:V, F:V, A:I, T:V, A:L, G:V, I:I, V:Y, F:L, L:L, F:I, G:L, L:Y have more probability in amino acids pairs in sheet structures.

The probability of 210 AA pairs (*p(Ai:Aj)*) has also been calculated. Also calculated were the 20 amino acids probability in whole protein sequences *(p(Ai)*) . Given a AA pair: Ai:Aj, if Ai and Aj pair randomly, *p(Ai:Aj)* will be equal to *p(Ai)p(Aj)*; if Ai and Aj have a tendency to interact with each other to form a pair, *p(Ai:Aj)* will bigger than *p(Ai)p(Aj)*. If *p(Ai:Aj)* is divided by *p(Ai)p(Aj)*, $\log_2 \frac{p(A_i : A_j)}{p(A_i)P(A_j)}$ can indicate the affinity of the two amino acids (Ai and Aj) in folding sheet structures.

We studied whether 20 amino acids paired randomly or not as a whole in all proteins by using the relative entropy ($D(p \| q)$) as the measurement. If amino acids pair randomly, the relative entropy will be zero.

$$D(p \| q) \overset{def}{=} \sum_{i=1}^{20} \sum_{j=1}^{20} p(A_i : A_j) \log_2 \frac{p(A_i : A_j)}{p(A_i)P(A_j)} = 1.1738 > 0$$

The result reveals that there are non-random propensities in amino acids pairs. So the AA pairs with detailed information collected in the database are useful, and the database can be a valuable tool to explore the interactions between strands in protein folding.

5 QUERY INTERFACE

A query interface to the data in the database has been implemented using an IIS web server and ASP language. It has been designed for ease of use with the primary interface consisting of pull down menus and simple text boxes. Query results are shown as a table, with each row containing element identifiers and the attribute data requested on the front page. All data within the SheetsPair database is freely available to the scientific community. Simple searches may be performed on all of the data.

5.1 Amino acid pairs

The database provides an interface for users to retrieve the details of amino acids pairs. Very complex queries can be made. Users can list all pairs of a given protein or chain or even a strand, with optional setting of the position range. Users are also able to customize their own samples to form certain AA pairs dataset for further research. Even a specific dataset can be customized; for instance, one can readily get the AA pairs in which the two amino acids are at the same chain or at different chains. Simple analytical parameters are provided in the form, such as the number count and the probabilities of 20 amino acids and of 210 AA pairs. Also searchable are the sequences and the appearance frequencies of the splice segments of bonded strand pairs in all proteins in the database.

Recently, a database of interchain β-sheet interactions has been created (Morrissey, Ahmed, & Shakhnovich, 2004). However, SheetsPair enables users to retrieve amino acids pairs on both inter-chain and intra-chain strands. SheetsPair has more functions and search options than the Morrissey database.

5.2 Strands and Sequences

SheetsPair implements a dynamic web interface that allows researchers to browse all strands and their sequences of proteins collected in the database. Because a standardized relational database structure has been utilized, such as creating amino acid IDs and StrandIDs as unique identifiers, the strands query result with its sequence is precise and accurate, without ambiguity. To obtain a strand list of a certain protein, one can select a protein, and a precise strand list will be returned. When a strand is selected in the summary page, all the amino acids pairs on the strand will be displayed in another single page. Also searchable are the lists of strands involved in parallel or antiparallel pairing.

Because non-freetext structure is implemented for amino acid sequences, all sequences as well as amino acids can be displayed in a user's chosen style, e.g. single-letter code, three-letter code, standard name, or even hydrophobic index, which provide an easy way to sequence converting or sequence coding.

5.3 Links

Links are provided to other information sources, e.g. the PDB database. From the main table, clicking on the protein name will link to the PDB database, and the detailed information on the protein structure from the PDB database will be returned.

5.4 Advanced SQL search

The user interface also provides a SQL search form for advanced searching. With a basic knowledge of SQL language, users are able to type their own SQL language into the web page form to retrieve data from the database.

6 CONCLUSIONS

The SheetsPair database is an Internet-accessible relational database containing pairs of amino acids on adjacent strands in protein sheet structure. As mentioned above, the database is rigorous, precise, and accurate, with little uncertain factors or ambiguity.

Although we can determine or predict more and more secondary structures, there is still a long way to go to reach the tertiary structure and the protein folding mechanism. Fundamental information should emerge from known protein secondary structural elements. Because non-random propensities of amino acids pairs in sheet structure are revealed in this database, there can be much useful information discovered about these pairs. Therefore, the SheetsPair database can be a valuable tool to explore interactions between amino acids on adjacent strands and additionally explore interactions between different strands or even predict a partner of a strand in a protein sheet structure. This information can be a valuable tool for scientists in predicting protein tertiary structure after the secondary structure is known or has been predicted. We anticipate that the statistics and insights gained from this database will contribute to the development of protein structure prediction and to the understanding of the mechanisms of sheet structure folding.

7 ACKNOWLEDGEMENTS

This work was supported by grants from the National Basic Research Program of China (2004CB720300) and the NSFC (30370386, 30470453).

8 REFERENCES

Cheng, H., Sen, T., & Kloczkowski, A. (2005) Prediction of protein secondary structure by mining structural fragment database. *Polymer 46(12):* 4314-4321.

Dayalan, S., Gooneratne, N., Bevinakoppa, S., et al. (2006) Dihedral angle and secondary structure database of short amino acid fragments. *Bioinformation 1(3*): 78-80.

Dill, K. (1990) Dominant Forces in Protein Folding. *Biochemistry 29(31*): 7133-7155.

Dou, Y., Baisnée, P. Pollastri, G., et al. (2004) ICBS: a database of interactions between protein chains mediated by β-sheet formation. *Bioinformatic, 20(16):* 2767-2777.

Eisenberg, D., Wilcox, W., & McLachlan, A. (1986) Hydrophobicity and amphiphilicity in protein structure. *Journal of Cell Biochemistry 31(1):* 11-7.

Galzitskaya, O., Ivankov, D., & Finkelstein, A. (2001) Folding nuclei in proteins. *FEBS Letters 489*: 113-118.

Hvidsten, T., Kryshtafovych, A., Komorowski, J., et al. (2003) A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics 19 Suppl. 2*: ii81–ii91.

Kloczkowski, A, Ting, K., Jernigan, R., et al. (2002) Protein secondary structure prediction based on the GOR algorithm incorporating multiple sequence alignment information. *Polymer 43*: 441-449.

Jones, D. (1999) Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology 292*: 195-202.

Lee, J., Kim, S., Joo, K., et al. (2004) Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins 56(4*): 704-14.

Lee, J., Kim, S., & Lee, J. (2005) Protein structure prediction based on fragment assembly and parameter optimization. *Biophys Chem. 115(2-3*): 209-214.

Morrissey, M., Ahmed, Z., & Shakhnovich, E. (2004) The role of cotranslation in protein folding: a lattice model study. *Polymer 45(2*): 557-571.

Qiu, J., Liang, R., Zou, X. (2003) Prediction of protein secondary structure based on continuous wavelet transform. *Talanta, 61*: 285-293.

Onuchic, J. & Wolynes, P. (2004) Theory of protein folding. *Current Opinion in Structural Biology 14*: 70–75.

Rohl, C., Strauss, C., Misura, K., et al. (2004) Protein structure prediction using Rosetta. *Methods Enzymol. 383*: 66-93.

Wang, L., Liu, J., Li, Y., et al. (2004) Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme. *Genome Informatics 15(2*): 181-190.

Ward, J., McGuffin, L., & Buxton, B. (2003) Secondary structure prediction with support vector machines. *Bioinformatics 19(13):* 1650-1655.