EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

Original article:

NEW INSIGHTS REGARDING PROTEIN FOLDING AS LEARNED FROM BETA-SHEETS

Ning Zhang[1,4], Yuanming Feng[1], Shan Gao[2,4], Jishou Ruan[2,3*], Tao Zhang[4*]

[1] Department of Biomedical Engineering, Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin, 300072, PR China
[2] College of Mathematical Science, Nankai University, Tianjin 300071, PR China
[3] State Key Laboratory for Medical Chemical and Biology at Nankai University, Tianjin, PR China, 300071
[4] College of Life Sciences, Nankai University, Tianjin, PR China, 300071
* corresponding authors: jsruan@nankai.edu.cn, Tel: +86 022 23501449
zhangtao@nankai.edu.cn, Tel: +86 022 23500237

ABSTRACT

The folding of denatured proteins into their native conformations is called Anfinsen's dogma, and is the rationale for predicting protein structures based on primary sequences. Through the last 40 years of study, all available algorithms which either predict 3D or 2D protein structures, or predict the rate of protein folding based on the amino acid sequence alone, are limited in accuracy (80 %). This fact has led some researchers to look for the lost information, from mRNA to protein sequences, and it encourages us to rethink the rationale of Anfinsen's dogma. In this study, we focus on the relationship between the strand and its partners. We find two rules based on a non-redundant dataset taken from the PDB database. We refer to these two rules as the "first coming first pairing" rule and the "loveless" rule. The *first coming first pairing rule* indicates that a given strand prefers to pair with the next strand, if the connected region is flexible enough. The *loveless rule* means that the affinities between a given strand and another strand are comparable to the affinity between the given strand and its partner. Of course, the affinities between the given strand and a helix/coil peptide are significantly less than the affinity between the given strand and its partner. These two rules suggest that in protein folding, we have folding taking place during translation, and suggest also that a denatured protein is not the same as its primary sequence. Rechecking the original Anfinsen experiments, we find that the method used to denature protein in the experiment simply breaks the disulfide bonds, while the helices and sheets remain intact. In other words, denatured proteins still retain all helices and beta sheets, while the primary sequence does not. Although further verification via biological experiments is needed, our results as shown in this study may reveal a new insight for studying protein folding.

Keywords: β-sheet, helix, near-neighbor pairing, strand-level, protein folding

INTRODUCTION

Anfinsen's dogma ensures that protein 3D-structure is perfectly determined by the amino acid sequence (Anfinsen, 1973). As the rationale to support the protein folding problem and the *de novo* structure prediction to obtain tremendous progresses. For example, the fragment assembly method (Bradley et al., 2005; Lee et al., 2005; Fujitsuka et al., 2006) and TASSER method (Wu et al., 2007; Zhou et al., 2007; Zhou and Skolnick, 2007). However, all available algorithms to describe amino acid sequences folding into their native structures

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

(Fooks et al., 2006; Parisien and Major, 2007; Dorn and Souza, 2010) have not arrived at the ideal accuracy. The protein folding kinetics and design are still challenge problems (Huang and Gromiha, 2010; Bowman et al., 2011). Why protein de novo structure prediction obstacles? We should rethink the Anfinsen dogma. Does the Anfinsen dogma have flaws? The denatured protein is really the same as the primary sequence?

The functional areas on protein tertiary structures often involve secondary structure elements (i.e., α-helices, β-sheets). The study of secondary structures is very important for recognizing protein folding and solving structure prediction problems (Steward and Thornton, 2002; Zhang et al., 2005a). Therefore, it is valid to mine this knowledge from secondary structures. Regarding α-helices and β-sheets, the α-helix has been understood in much detail, while comparatively little is known about the β-sheet (Jäger et al., 2007). The tertiary structures of β-sheet-containing proteins are especially difficult to simulate (Steward and Thornton, 2002; Kuhn et al., 2004; Wathen and Jia, 2010). Unlike α-helices folded by one peptide, β-sheets are folded by two or more disjoint peptides (strands). In this structure, adjacent β-strands bring distant residues into close contact with one another, and constitute a specific mode of amino acid pairing (like DNA base pairing) (Fooks et al., 2006; Ashkenazy et al., 2011; Zhang et al., 2009, 2010).

Studies on β-sheets have become interesting problems in bioinformatics. There is a growing recognition of the importance of strand-to-strand interactions among β-sheets (Nowick, 2008). Several studies, including statistical studies examining the frequencies of nearest-neighbor amino acids, found significantly different preferences for certain inter-strand amino acid pairs (Russell and Cochran, 2001; Fooks et al., 2006; Ashkenazy et al., 2011). Dou and his colleagues created a comprehensive database for interchain β-sheet (ICBS) interactions (Dou et al., 2004). In our previous studies, we also constructed the SheetsPair database (Zhang et al., 2007) to compile both interchain and intrachain amino acid pairs.

The known efforts on β-sheets focus mainly on the inter-residue contacts or amino acid partners (Baldi et al., 2000; Zhang et al., 2005b; Grana et al., 2005; Halperin et al., 2006; Kundrotas and Alexov, 2006; Cheng and Baldi, 2007). Although predictions of inter-residue contacts are interesting and useful for an understanding of protein folding (Zhang et al., 2005a; Cheng and Baldi, 2007), those studies should be viewed as the initial steps of β-sheet studies (Baldi et al., 2000). BETAPRO, a method to assemble β-strands to predict β-sheets, was introduced by Cheng and Baldi (2005). However, BETAPRO was based on prediction results of residue contacts, in which a single mis-prediction of one amino acid pair from the first stage could be amplified through subsequent stages and results in seriously incorrect strand pairs. Kato et al. (2009) stated that the prediction of planar β-sheet structures belongs to the NP-hard class of complexity in our present state of knowledge. Our previous studies showed that the interstrand amino acid pairs played a significant role in determining the parallel or antiparallel orientation of β-strands (Zhang et al., 2009), and the statistical results could possibly be used to predict β-strand orientation (Zhang et al., 2010). In our present study, we attempted to take further steps in the investigation of β-sheets in strand-level, in hopes of gaining insight.

Dataset

All protein structures used in this study were taken from a PISCES (Wang and Dunbrack, 2003, 2005) dataset, generated on May 16, 2009. In this dataset, the percentage identity cutoff is 25 %, the resolution cutoff is 2.0 angstroms, and the R-factor cutoff is 0.25. Besides removing the proteins containing disordered regions (Ferron et al., 2006; Linding et al., 2003; Liu et al., 2009), all data were further preprocessed according to the following criteria: (1) Protein chains having no β-sheet are removed; (2) Protein chains containing non-

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

standard residues (i.e., DPN, EFC, ABA, C5C, PLP, et al.) are removed because these protein chains have covalently-bounded ligands or modified residues; (3) Protein chains having no uncertain structures or incorrect data are removed. Finally, 2,298 protein chains are kept, and 6,740 parallel β-strand pairs and 12,474 antiparallel β-strand pairs are obtained from these 2,298 protein chains.

RESULTS AND DISCUSSION *of the statistical analysis of BSD*

The β-strand Distance (BSD)

A β-sheet is folded by two or more extended strands. We select the protein 1HZT (PDB code) as an illustrative example, shown in Figure 1(a). Protein 1HZT has three β-sheets, called A, B and C respectively. A, B and C are folded by 10 different β-strands numbered 1 to 10 from N-terminal to C-terminal, respectively. The locations and the amino acids corresponding to the 10 β-strands are shown in Figure 1(b). In each β-sheet folded by multiple strands, the subunit folded by two strands is referred to as a β-strand pair. Thus, each β-sheet has at least one β-strand pair. All β-strand pairs (SR) may be classified into parallel and antiparallel pairs, according to the directions of the two strands in the β-strand pair. Typically, the protein 1HZT has 2 parallel and 5 antiparallel pairs. The strands 'B3' and 'B4' are folded to an antiparallel pair, shown in Figure 1(d).

The β-sheet topology or architecture (i.e. the pairing assignments of all the forming β-strands) is essential for understanding a protein's tertiary structure (Zhang and Kim, 2000). In this study, the β-strand distance (BSD) of a strand pair is defined as the number of β-strands along the primary sequence between the two paired strands. No matter the number of residues between them, the BSD only considers the number of strands (Figure 1(c)). It is obvious that the BSD is 1 in the case where there are no other β-strands between the two paired strands along the primary sequence, and we refer to this as 'nearest pairing' in this study.

1(b): In each β-sheet folded by multiple strands, the subunit folded by two strands is referred to as a β-strand pair. Thus, each β-sheet has at least one β-strand pair. All β-strand pairs (SR) may be classified into parallel and antiparallel pairs, according to the directions of the two strands in the β-strand pair. Typically, the protein 1HZT has 2 parallel and 5 antiparallel pairs. The strands 'B3' and 'B4' are folded to an antiparallel pair, shown in Figure 1(d).

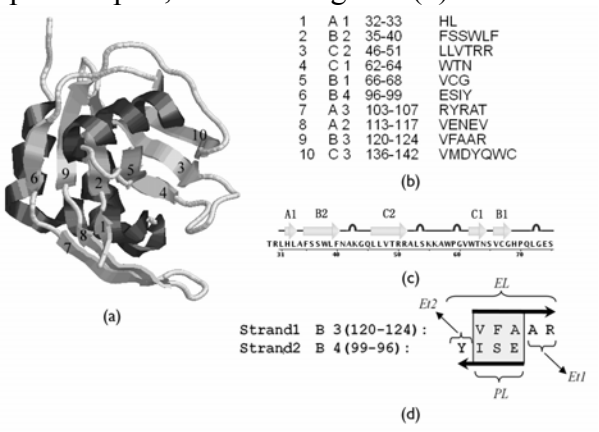

Figure 1: Illustration of β-strand pairing in a β-sheet (1HZT) (**a**) The sketch of the tertiary structure of the protein produced using RASMOL. Protein 1HZT is an α/β protein having 10 β-strands, numbered from 1 to 10 from N-terminal to C-terminal. These 10 β-strands fold into three β-sheets, and we can produce 7 strand pairs. (**b**) The sequences of the 10 β-strands with their initial and ending residue numbers. (**c**) The 10 β-strands in the linear primary sequence. The BSD of A1 and B2 is 1, while the BSD of A1 and C2 is 2 and the BSD of A1 and C1 is 3. (**d**) An example of a β-strand pair formed by strand "B3" and "B4", with the light gray box representing the common region of the pair.

The "First Coming First Pairing" rule

Based on the benchmark dataset consisting of the 6,740 parallel pairs and the 12,474 antiparallel pairs, we compute the β-strand Distance (BSD) for all strand pairs in our dataset. The rates of the number of strand pairs having different BSDs based on the set of 6,740 parallel pairs, the set of 12,474 antiparallel pairs and the entire benchmark dataset are shown in Table 1 and Figure 2. From Table 1, we note that

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

the maximal BSD within the parallel pairs is 30, and the maximal BSD within antiparallel pairs is 54. Typically, the occurrence rate of strand pairs with BSD=1 is about 60 %, the rate of these strand pairs with BSD less than 3 is about 80 %, and the rate of these strand pairs with BSD less than 10 is about 97 %. The cumulative percents, according to BSDs, are shown in Figure 2(a). It is obvious that the curve increases sharply when the BSD is small, and it seems to be constant as BSD increases to larger than 10. Moreover, this rule does not depend on the selection of the sets of parallel and antiparallel pairs. Figure 2(b) shows that the occurrence rate of strand pairs with BSD=1 is major, while the occurrence rates of these strand pairs having either BSD=2 or BSD=3 are both minor. Notably, strand pairs with BSD>3 are rare. This suggests that a β-strand most often prefers to choose its nearest strands to partner with. We term this propensity the "First Coming First Pairing" rule.

Among all non-nearest β-strand pairs (BSD>1), we mainly consider those pairs with BSD=2 and BSD=3. In other words, 1- or 2-interval strands are sandwiched by the two paired strands. Then the 1- or 2-interval strands may join to the same β-sheet of the given non-nearest β-strand pair, or join to another β-sheet. In the former case, we call the 1- or 2-interval strands "national strands". In the latter case, we call the 1- or 2-interval strands "foreign strands". Based on all strand pairs with BSD=2 or BSD=3, we compute the rates of national interval and foreign interval strands, respectively, with the statistical results shown in Table 2.

When the BSD=2, Table 2 shows that the rate of national interval strands is much greater than that of foreign interval strands. This suggests that the initial strand must wait for the next nearest strand to become its partner if the connection region between the initial strand and its nearest strand are not flexible. Then, the nearest strand will most often pair with another strand within the same sheet of the initial strand, and is only infrequently paired with a strand of another sheet.

Table 1: The maximal BSD and the cumulative rates of the strand pairs having different BSDs

Orientation	Number of strand pairs	BSD Range		Maximum percent		Cumulative percent of BSD≤3	Cumulative percent of BSD≤10
		Min.	Max.	BSD	percent		
parallel	6,740	1	30	1	56.19 %	81.45 %	98.15 %
antiparallel	12,474	1	54	1	61.43 %	79.34 %	96.99 %
overall	19,214	1	54	1	59.59 %	80.08 %	97.39 %

Table 2: Percentages of national interval and foreign interval strands, based on all β-strand pairs with BSD=2 or BSD=3 respectively

BSD	Orientation	Interval strand(s) "National" or "Foreign"	Percent
BSD=2 (1-interval strand)	parallel	National	64.95 %
		Foreign	35.05 %
	antiparallel	National	62.96 %
		Foreign	37.04 %
	overall	National	63.87 %
		Foreign	36.13 %
BSD=3 (2-interval strands)	parallel	Both are National	54.57 %
		One is National, the other is Foreign	4.81 %
		Both are Foreign	40.62 %
	antiparallel	Both are National	30.45 %
		One is National, the other is Foreign	17.05 %
		Both are Foreign	52.50 %
	overall	Both are National	40.27 %
		One is National, the other is Foreign	12.07 %
		Both are Foreign	47.66 %

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

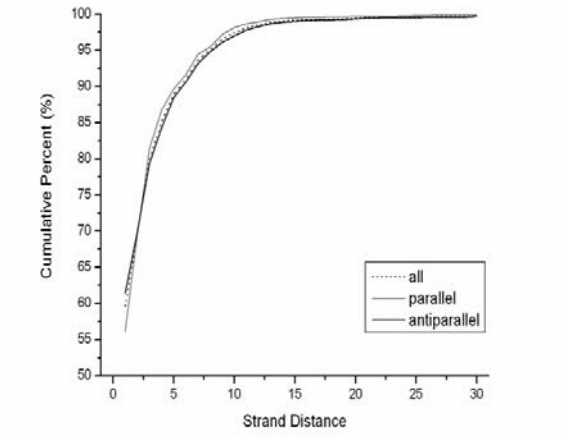

Figure 2(a)

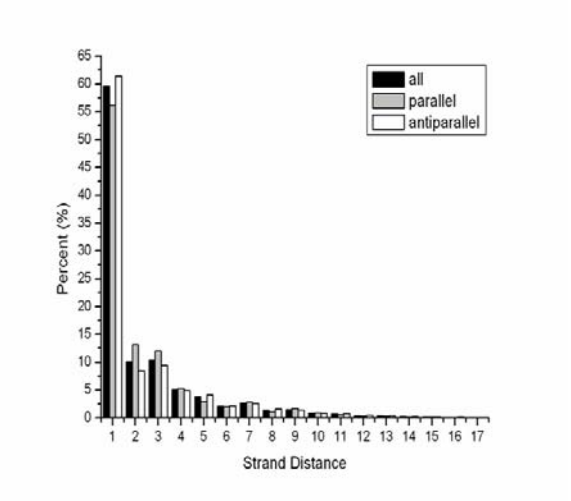

Figure 2(b)

Figure 2: (**a**) Cumulative percent of β-strand pairs as BSD increases. (**b**) Distribution of β-strand pairs as BSD changes (truncated to 17 since percents are almost 0 when BSD>17). Both pictures mention the sets of parallel, antiparallel and all strand pairs.

When the BSD=3, Table 2 shows that the pairing states of the two interval strands for parallel and antiparallel are different, although they would like to be paired with each other in both cases. For the parallel case, two interval strands will overwhelmingly prefer to be paired with each other, either remaining in the same sheet of the initial strand or going outside of the sheet. For the parallel case, it is rare that two interval strands do not pair with each other and become separated into two sheets (4.81 %). For the antiparallel case, however, this is not rare, as seen by the rate of 17.05 %. In each case, a similar explanation might be given as that for the case when the BSD=2. The nearest partner and the next-nearest strand of the initial strand are both blocked, and the initial strand must await the third nearest strand to partner with. In the former case (the same β-sheet), a possible blocking factor might be the fact that a strand cannot partner if it has already paired with others, since one strand can have no more than 2 partners. In the latter case (a different β-sheet), the different β-sheet formed by the two-interval strands could be a stronger blocker, obstructing the potential strands in closing with each other in 3-D space. As a matter of fact, the two interval strands are also nearest-pairing in most cases (see below). It is from this perspective that we offer the possible explanation that one nearest-pairing blocks another, with the result that the second BSD is 3.

For BSD=3 pairs (as shown in Figure 3), we further investigate all possible pairing styles, with results shown in Table 3. From Table 3, it is interesting to note that the case where the two interval strands ‘f’ and ‘g’ pair together accounts for the majority (overall 74.03 %). Note that the f-g pairing is also a nearest-pairing, obeying the “First Come First Pair” rule (BSD=1). One possible explanation could be that the rule is first obeyed between strands ‘f’ and ‘g’, which pair together in the first stage. Due to the blocking factor initiated by the f-g pairing, strand ‘a’ can neither choose its nearest neighbor ‘f’ as its partner, nor the next nearest ‘g’. Thus, it must choose the next-next-nearest, ‘b’, resulting in a BSD=3 pair. Collectively, although the BSD=3 case does not outwardly obey the “First Come First Pair” rule, it could indeed be a consequence of such a rule. Another observable fact supporting this assumption could be found in the case of the first style in Table 3, in which ‘f’ and ‘g’ cannot pair together. This could be due to blocking, caused by the a-f pair (BSD=1) and g-b pair (BSD=1), in which two nearest-pairings block another nearest-pairing ‘f-g’.

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

Table 3: Percentage of occurrences and cases of each of all possible pairing styles of a BSD=3 pair*

	Percentage of occurrence	Percent of cases in which f and g pair	percent of cases in which f and g do not pair
'a' pairs with 'f', and 'b' pairs with 'g'	0.70 %	0.00 %	100.00 %
'a' pairs with 'g', and 'b' pairs with 'f'	0.19 %	0.00 %	100.00 %
only 'a' pairs with 'f'	17.99 %	69.04 %	30.96 %
only 'b' pairs with 'g'	11.04 %	63.71 %	36.29 %
only 'a' pairs with 'g'	8.75 %	85.19 %	14.81 %
only 'b' pairs with 'f'	14.27 %	66.76 %	33.24 %
neither 'f' nor 'g' pairs with 'a' or 'b'	47.07 %	79.88 %	20.12 %
sum	100.00 %	74.03 %	25.97 %

*In all cases, 'strand a' and 'strand b' formed a pair. 'Strand f' and 'strand g' are the two interval strands between them ('strand f' is near 'strand a', 'strand g' is near 'strand b')

Figure 3: Strands along the primary sequence of a BSD=3 pair
The four dark gray lines represent the four β-strands in the primary sequence, while strand 'a' and 'b' are the given β-strand pair with BSD=3. Strand 'f' and 'g' are the 2 interval strands.

In summary, the "First Come First Pair" rule is encountered widely in β-strand pairing, but does not occur in all strand pairs. One possible reason could be that already-paired strands may hinder others from pairing with the nearest neighbors, considering the fact that one strand can only have 1 or 2 partners. There could be other reasons, in view of the complexity of protein folding. Regardless, the "First Come First Pair" rule remains important in β-strand pairing, which could eventually lead to protein folding pathways.

RESULTS AND DISCUSSION *of the features of real vs. pseudo strand pairs*

Terminal extensions of β-strand pairs

For the two strands in a pair, the N or C terminals of one strand do not always align with the N or C terminals of the other, giving rise to terminal extensions besides the common pairing region (Figure 1(d)). Let *PL* stand for the length of the common region, *Et1* and *Et2* stand for the length of the two terminal extensions, respectively, and let *EL* represent the total pair length (i.e. *EL* = *PL+Et1+Et2*). Then, the common paring ratio *R* could be calculated by:

$R = PL/EL \times 100\ \% =$
$PL/(PL + Et1 + Et2) \times 100\ \%$

If the lengths of two strands are represented by *SL1* and *SL2*, respectively, the ratio of the common pairing region to the length of each strand could be calculated by:

$Rt_i = PL/SL_i \times 100\%,\ i = 1,2$

The *R*, *Rt1* and *Rt2* of every strand pair in our dataset have been calculated in the present study. Results are shown in Figure 4. It can be seen from Figure 4 that when $Rt1 \geq 40\ \%$ and $Rt2 \geq 40\ \%$, the cumulative percentages of the two strands reach 94.26 % and 95.98 %, respectively; and when $R \geq 25\ \%$, the cumulative percentage rises to 96.97 %. Therefore, a rule of β-strand pairing could be as follows:

$R \geq 25\ \%$ and $Rti \geq 40\ \%$

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

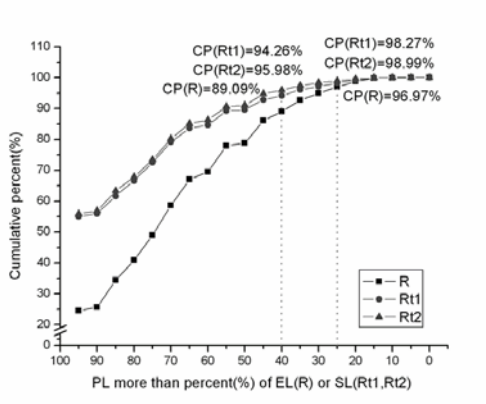

Figure 4: Cumulative percentages (CP) of R, Rt1 and Rt2 calculated from the present dataset

The horizontal axis denotes the percentage of common paired region PL to EL (for curve R) or to SL (for curves Rt1 and Rt2). Points on the R curve denote the cumulative percentages of samples whose R=PL/EL equals or exceeds the corresponding abscissa value. Points on the Rt1 and Rt2 curves denote the cumulative percentages of samples whose Rt1=PL/Rt1 or Rt2=PL/Rt2 equals or exceeds the corresponding abscissa value, respectively.

To reduce computational searching space, we will use this rule in subsequent steps when we traverse all possible relative positions of two specific strands to pair in the present study.

Real β-strand pairs and pseudo strand pairs

In order to investigate the assignments of β-strand pairs, we analyzed another 3 types of 'pseudo' strand pairs, as well as the 'real' β-strand pairs. The pseudo pairs are generated from primary sequences by randomly selecting stretches of different secondary structures as alternative partners of a β-strand. Since such pairs never occur in a functional protein, these types of pairs are called "Pseudo Strand Pairs".

The real β-strand pairs are denoted as SR (a β-Strand with its Real partner β-strand). The other three pseudo strand pairs are denoted as: (i) SS (a β-Strand with a no-real-partner β-Strand, i.e. the partner is randomly selected from other β-strand stretches from the primary sequence); (ii) SH (a β-Strand with a randomly selected α-Helix stretch from the primary sequence); (iii) SC (a β-Strand with a randomly selected Coil stretch from the primary sequence). The random procedure was repeated iteratively 5 times.

Ultimately, four types of pairs were obtained. In the next step, features of these pairs were extracted and compared.

Feature extraction from the four types of pairs

Many studies (Asogawa, 1997; Steward and Thornton, 2002; Fooks, et al., 2006) suggest that amino acid pairing in β-sheets involves implicit information which was not only helpful for the β-sheet structure prediction, but also significant to disclose the potential mechanisms and rules of β-sheet assembly. In this study, to extract features of the four types of strand pairs above, we used the Average Amino Acid Pairing Encoding Matrix (APEM) which was generated in our previous study (Zhang et al., 2009, 2010). The matrix compiled information regarding the amino acid pairs. The APEM matrix was an upper triangular matrix, since only 210 possible amino acid pairs were considered, regardless of the order of the two amino acids within one pair. An element in the matrix was defined as follows:

$$m(A_i : A_j) = P(A_i : A_j)/(P(A_i)P(A_j)), 1 \le i \le 20, 1 \le j \le 20, i \le j$$

in which A_i and A_j are the two amino acids forming an inter-strand pair, and $P(A_i : A_j)$ represents the observed frequency of the amino acid pair $A_i : A_j$. The terms $P(A_i), P(A_j)$ are the background probability generated by counting single amino acid frequencies of A_i, A_j respectively across all protein sequences in the dataset, which was similar to the previous work by Bryan et al. (2009).

The feature extraction steps were as follows:

Firstly, each element $m(A_i : A_j)$ in APEM was transformed by:

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

$$r(A_i : A_j) = \log m(A_i : A_j)$$
$$r(A_i : A_j) = -1, \quad if\ r(A_i : A_j) < -1,$$
$$r(A_i : A_j) = 1, \quad if\ r(A_i : A_j) > 1.$$

The average value of all *r(Ai:Aj)* was calculated by:

$$avgR = 1/210(\sum_{i=1}^{20}\sum_{j=1}^{i} r(A_i : A_j))$$

Then, we defined the feature score *f* and feature score *d* as follows:

$$f = \prod_{A_i:A_j}^{PL}(1 - Rpos(A_i : A_j))$$

$$d = \prod_{A_i:A_j}^{PL}(1 - Rneg(A_i : A_j))$$

where *PL* represents the length of the common pairing region of the two strands; $Rpos(A_i : A_j)$ and $Rneg(A_i : A_j)$ were calculated by:

$$Rpos(A_i : A_j) = \begin{cases} r(A_i : A_j) - avgR & if \quad r(A_i : A_j) \geq avgR \\ 0 & else \end{cases}$$

$$Rneg(A_i : A_j) = \begin{cases} abs(r(A_i : A_j) - avgR) & if \quad r(A_i : A_j) \leq avgR \\ 0 & else \end{cases}$$

For each strand pair (for both real and pseudo ones), all possible relative pairing positions were traversed according to the rule (R≥25 % and Rti≥40 %), running in both parallel and antiparallel fashions. The relative pairing position and the orientation fashion were determined for the maximum *f* value. The *f* value was one of the features used. At this position, the corresponding d value was calculated, which became another feature. For each pair of strands, one set of the two features was calculated, and then used in the next step.

The non-conservative (loveness) propensity of β-strand partner

We investigated and compared the extracted features of the real and the three pseudo strand pairs. A scatter plot of *d* value (y) versus *f* value (x) of the four types of pairs are given in Figure 5. It is obvious from Figure 5 that the distributions of SR and SS features are similar (Figure 5 (a) and (b)), while they differ for SH and SC pairs (Figure 5 (c) and (d)). It can also be seen that the distributions of SH are slightly more similar to SR than SC (d).

Since in a scatter plot the similarities between the four types cannot be clearly demonstrated, we adopt the famous pattern recognition method -- support vector machine (SVM) -- to attempt to distinguish the features of the four types of pairs. Here, SVM was used only for distinguishing features and not for prediction, as it was used in (Zhang et al., 2009). The distinguishing results obtained via SVM are shown in Table 4.

Table 4: Results of feature distinguishing between the four types of pairs, using SVM. (7-fold cross-validation test. RBF kernel function, with c and gamma set to the default value in LibSVM 2.83.)*

	Average accuracy	Average MCC
SR-SS	57.74 %	/
SR-SH	76.79 %	0.2989
SR-SC	82.60 %	0.5192
SS-SH	70.55 %	0.2500
SS-SC	77.05 %	0.4483
SH-SC	58.87 %	0.1776

*The results are the average of 5 times' random procedure for generating SS, SH and SC pairs.

From Table 4, it can be seen that the classification efficiency of SR-SH, SR-SC, SS-SH and SS-SC can be accepted, while the SH-SC result is poor, and SR-SS is very poor. This is consistent with observations from the scatter plot. The features of SR and SS are so similar that even SVM can not distinguish them. The poor efficiency of SH-SC indicates that helix and coil stretches both have similar characteristics, different from β-strand pairs. In view of strand pair formation, there is no significant difference between helix and coil segments. The much better distinguishing results of SR-SH and SR-SC indicate that a strand has the ability to distinguish its real partner

EXCLI Journal 2012;11:543-555 – ISSN 1611-2156
Received: June 12, 2012, accepted: August 22, 2012, published: August 27, 2012

from helix or coil segments. The moderately good distinguishing of SS-SH and SS-SC indicates that other non-real-partner strands have this ability as well. However, the very poor SR-SH distinguishing suggests that a strand cannot distinguish its real partner from other non-real-partner β-strands.

From these results, it can be concluded that the partner is loveness for a single β-strand. Although a single β-strand has the ability to distinguish its partner from a helix or a coil, it lacks the ability to distinguish it from other non-partner β-strands. Similar results were obtained in earlier studies. Ren et al. (2006) reported that pairs of residues on neighboring strands were neither more strongly conserved nor more strongly covariant than pairs of the same type in non-interacting positions. Mandel-Gutfreund et al. (2001) found that residue pairs in anti-parallel β-sheets were equally conserved and covaried as much as non-interacting residue pairs. Steward and Thornton (2002) also indicated that a single β-strand was able to recognize a non-interacting β-strand with greater accuracy than in the case of recognition between two random sequences. However, these studies did not consider partners among random selected stretches of different secondary structures. It could be suggested that the loveness nature of β-strand partners could also be employed to explain why the β-sheet structures are so especially difficult to simulate in protein tertiary structure predictions (Steward and Thornton, 2002; Kuhn et al., 2004).

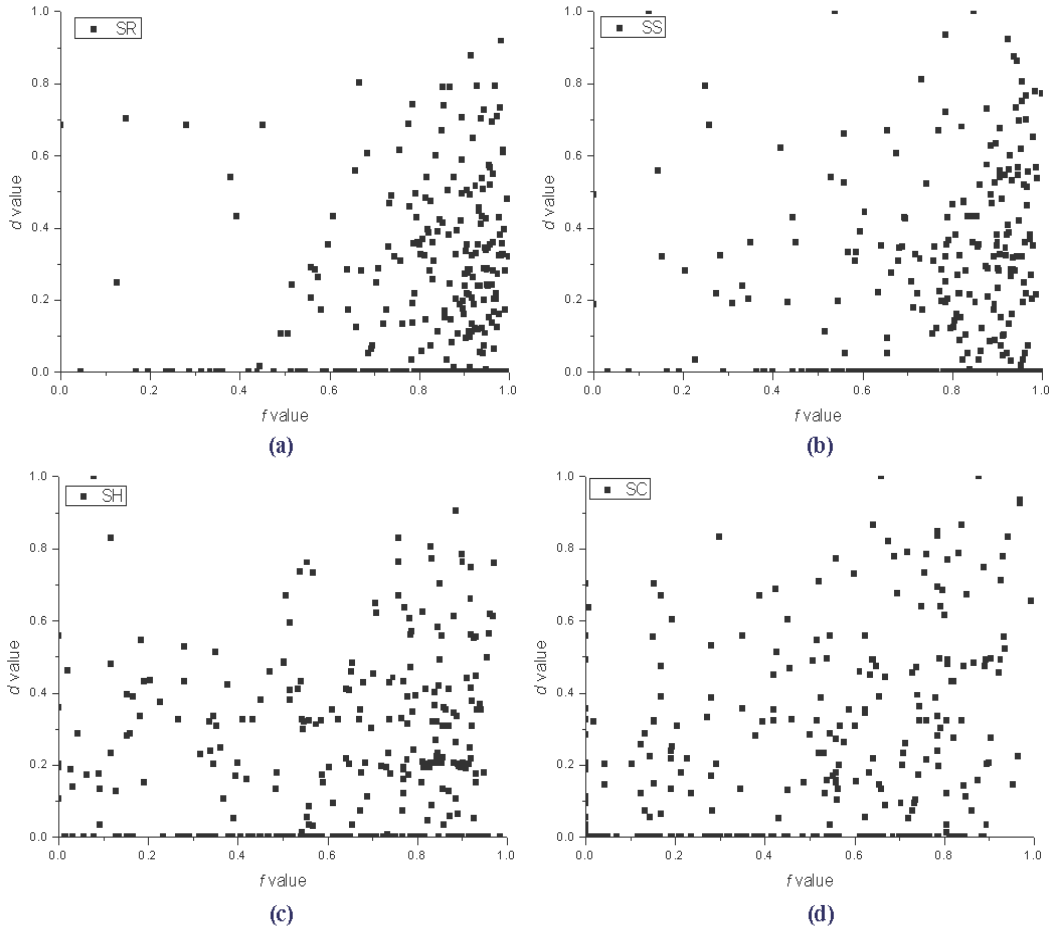

Figure 5: Scattered plot of *d* value (y) versus ***f*** value (x) of real β-strand pairs and pseudo strand pairs. (a) SR: real pairs (a β-Strand with its Real partner β-strand); (b) SS: pseudo pairs (a β-Strand with a no-real-partner β-Strand, i.e. the partner is randomly selected from other β-strand stretches from the primary sequence); (c) SH pseudo pairs (a β-strand with a randomly selected α-Helix stretch from the primary sequence); (d) SC pseudo pairs (a β-strand with a randomly selected Coil stretch from the primary sequence)

CONCLUSION

The "First Come First Pair" rule implies that one β-strand is inclined to pair with its nearest neighbor strands, or strands not far from it along the primary linear sequence. Analysis of pseudo strand pairs indicates that partner recognition is not conservative. Combining these two findings above, it can be concluded that in the process of β-sheet formation, the pairing of β-strands is not exclusively driven by specific residues or interacting amino acid pairs. Instead, in most cases, a single β-strand may follow the "First Come First Pair" rule to choose its partner. It prefers first to choose the nearest neighbor (with the smallest BSD value). However, if the first nearest neighbor is blocked, it must choose the next nearest neighbor, and if the next nearest is also blocked, it must then choose the next-next one, which still has a smaller BSD value. These results are in agreement with earlier studies by Wathen and Jia (2010), in which they investigated the initial nucleation step of β-sheet formation and indicated that nucleation was not primarily driven by specific interacting residue pairs; instead, β-nucleation was a local phenomenon resulting either from sequential or topological proximity. However, note that not all β-strand pairs obey this rule; although the reason is currently unclear. The chaperones may be one of the reasons, the circumstances may be another, but there definitely must be other reasons (Meiler and Baker, 2003).

The findings in this study complement Anfinsen's discovery. Anfinsen discovered three decades ago that denatured proteins can spontaneously self-assemble into their native conformations (Anfinsen, 1973). In various studies, Anfinsen further showed that denatured ribonuclease could be completely reversed by removing denaturing chemicals or by lowering the temperature. The ribonuclease could fold back to its natural functional state on its own. Therefore, Anfinsen concluded that the amino-acid sequence determines the structure of a protein. However, it still remains a challenge to explain how proteins fold into their native structures directly from their primary sequences (Bowman et al., 2011). It is worth pointing out that secondary structures were not known in Anfinsen's time, and secondary structures may not be totally collapsed during his denaturation process (Li et al., 1998). How do amino acids located far apart in the primary sequence find one another to interact in the 3D space? Studies shown that the degree of specificity between side-chain/side-chain interactions between residues on neighboring strands seem to be very weak (Wouters and Curmi, 1995). As a consequence, the interactions between amino acids and the 3D structures could not always be predicted if only the primary sequence is given. Since most proteins' folding processes are carried out simultaneous with translation, rather than after translation, the near-neighbor pairing propensity can be regarded in terms of the earliest-translated strands participating in pairing first. It is conceivable that this assumption is a consequence of the above fact, where the region of the translated primary sequence could partially determine the secondary structures and their neighbor interactions.

In conclusion, we imply that the 3D structure of a protein could be determined not only by the primary sequence, but also by the nearest-neighbor interacting secondary structure elements, which in turn may indeed be determined by local primary sequences. Although further verification must be done via biological experiments, the statistical results in the present study may point towards the notion that the nearest pairing propensity of secondary structure elements could be a potential rule among so many unknown protein-folding determining factors. This in turn could contribute to protein structure prediction, and the mechanisms of protein folding.

ACKNOWLEDGEMENT

We would like to thank Michelle Hanlon from Cross Cancer Institute, Edmonton, Alberta, Canada for her kindly

help. This work was supported by grants from the National Natural Science Foundation of China (31171053, 11232005, 68075049, 10671100, 31150110577, 31050110432 and 81171342), Tianjin research program of application foundation and advanced technology (12JCZDJC22300).

REFERENCES

Anfinsen CB. Principles that govern the folding of protein chains. Science 1973; 181(96):223-30.

Ashkenazy H, Unger R, Kliger Y. Hidden conformations in protein structures. Bioinformatics 2011;27:1941-7.

Asogawa M. Beta-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network. Proc Int Conf Intell Syst Mol Biol 1997;5:48-51.

Baldi P, Pollastri G, Andersen CA, Brunak S. Matching protein b-sheet partners by feedforward and recurrent neural networks. Proc Int Conf Intell Syst Mol Biol 2000;8: 25-36.

Bowman GR, Voelz VA, Pande VS. Taming the complexity of protein folding. Curr Opin Struct Biol 2011;21:4-11.

Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868-71.

Bryan AW Jr, Menke M, Cowen LJ, Lindquist SL, Berger B. BETASCAN: Probable b-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 2009; 5(3):e1000333.

Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks. alignments and graph algorithms. Bioinformatics 2005;21(Suppl 1):i75-i84.

Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 2007;8:113-21.

Dorn M, Souza OND. A3N: An artificial neural network n-gram-based method to approximate 3-D polypeptides structure prediction. Exp Syst Appl 2010;37:7497-508.

Dou Y, Baisnée PF, Pollastri G, Pécout Y, Nowick J, Baldi P. ICBS: a database of interactions between protein chains mediated by β-sheet formation. Bioinformatics 2004; 20:2767-77.

Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins-Structure Function and Bioinformatics 2006;65:1-14.

Fooks HM, Martin ACR, Woolfson DN, Sessions RB, Hutchinson EG. Amino acid pairing preferences in parallel β-sheets in proteins. J Mol Biol 2006;356:32-44.

Fujitsuka Y, Chikenji G, Takada S. Simfold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins 2006;62:381-98.

Grana O, Baker D, MacCallum R, Meiler J, Punta M, Rost B et al. CASP6 assessment of contact prediction. Proteins 2005;61:214-24.

Halperin I, Wolfson HJ, Nussinov R. Correlated mutations: advances and limitations. a study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006; 63: 832-45.

Huang L, Gromiha M. First insight into the prediction of protein folding rate change upon point mutation. Bioinformatics 2010; 26:2121-7.

Jäger M, Dendle M, Fuller AA, Kelly JW. A cross-strand Trp-Trp pair stabilizes the hPin1 WW domain at the expense of function. Protein Sci 2007;16:2306-13.

Kato Y, Akuts T, Seki H. Dynamic programming algorithms and grammatical modeling for protein beta-sheet prediction. J Comput Biol 2009;16:945-57.

Kuhn M, Meiler J, Baker D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins 2004; 54:282-8.

Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503:

Lee J, Kim S, Lee J. Protein structure prediction based on fragment assembly and parameter optimization. Biophys Chem 2005;115:209-14.

Li A, Fenselau C, Kaltashov IA. Stability of secondary structure elements in a solvent-free environment, II: The β-pleated sheets. Proteins 1998;Suppl 2:22-7.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453-9.

Liu B, Lin L, Wang XL, Wang X, Shen Y. Protein long disordered region prediction based on profile-level disorder propensities and position-specific scoring matrixes, In: IEEE International Conference on Bioinformatics and Biomedicine 2009;66-9.

Mandel-Gutfreund Y, Zaremba SM, Gregoret LM. Contributions of residue pairing to b-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel b-strands. J Mol Biol 2001;305: 1145-59.

Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci 2003;100:12105-10.

Nowick JS. Exploring β-sheet structure and interactions with chemical model systems. Acc Chem Res 2008;41:1319-30.

Parisien M, Major F. Ranking the factors that contribute to protein b-sheet folding. Proteins 2007;68:824-9.

Ren Y, Liu H, Xue C, Yao X, Liu M, Fan B. Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. Anal Chim Acta 2006;572:272-82.

Russell S J, Cochran A. Designing stable b-hairpins: Energetic contributions from cross-strand residues. J Am Chem Soc 2001;122:12600-1.

Steward RE, Thornton JM. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. Proteins 2002;48:178-91.

Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589-91.

Wang GL, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. Nucl Acids Res 2005;33:W94-8.

Wathen B, Jia ZC. Protein beta-sheet nucleation is driven by local modular formation. J Biol Chem 2010;285:18376-84.

Wouters MA, Curmi PMG. An analysis of side-chain interactions and pair correlations within antiparallel β-sheets: the differences between backbone hydrogen-bonded and non-hydrogenbonded residue pairs. Proteins 1995;22:119-31.

Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.

Zhang C, Kim S. The anatomy of protein beta-sheet topology. J Mol Biol 2000;2: 1075-89.

Zhang G, Huang DS, Quan ZH. Combining a binary input encoding scheme with RBFNN for globulin protein inter-residue contact map prediction. Patt Recogn Lett 2005a;26:1543-53.

Zhang Q, Yoon S, Welsh WJ. Improved method for predicting beta turn using support vector machine. Bioinformatics 2005b; 21:2370-4.

Zhang N, Ruan JS, Wu J, Zhang T. SheetsPair: a database of amino acids pairs in protein sheet structures. Data Sci J 2007;6:s589-95.

Zhang N, Ruan J, Duan G, Gao S, Zhang T. The interstrand amino acid pairs play a significant role in determining the parallel or antiparallel orientation of b-strands. Biochem Biophys Res Commun 2009;386:537-43.

Zhang N, Duan G, Gao S, Ruan J, Zhang T. Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines. J Theo Biol 2010;263:360-8.

Zhou H, Skolnick J. Ab initio protein structure prediction using Chunk-TASSER. Biophys J 2007;93:1510-8.

Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L et al. Analysis of TASSER-based CASP7 protein structure prediction results. Proteins 2007;69 (Suppl 8):90-7.