Inverse sequence similarity of proteins does not imply structural similarity

Stephan Lorenzen, Christoph Gille, Robert Preissner, Cornelius Frömmel*

Institute of Biochemistry, Charité Humboldt University, Monbijoustr. 2a, 10117 Berlin, Germany

Received 21 February 2003; revised 14 April 2003; accepted 15 April 2003

First published online 16 May 2003

Edited by Robert B. Russell

Abstract There is a debate on the folding of proteins with inverted sequences. Theoretical approaches and experiments give contradictory results. Many proteins in the Protein Data Bank (PDB) show conspicuous inverse sequence similarity (ISS) to each other. Here we analyze whether this ISS is related to structural similarity. For the first time, we performed a large scale three-dimensional (3-D) superposition of corresponding Cα atoms of forwardly and inversely aligned proteins and tested the degree of secondary structure identity between them. Comparing proteins of less than 50% pairwise sequence identity, only 0.5% of the inversely aligned pairs had similar folds (99 out of 19 073), whereas about 9% of forwardly aligned proteins in the same score and length range show similar 3-D structures (1731 out of 19 248). This observation strongly supports the view that the inversion of sequences in almost all cases leads to a different folding property of the protein. Inverted sequences are thus suitable as *protein-like* sequences for control purposes without relations to existing proteins.

Key words: Protein sequence; Inverse sequence similarity; Alignment; Superposition

1. Introduction

Proteins with sufficient sequence similarity can be expected to have a similar fold. We here raise the question whether also inverse sequence similarity (ISS) is an indicator of structural similarity. It has been shown before that many proteins in the Protein Data Bank (PDB) show ISS [1]. A key point in this issue is the question whether a protein with an inverted sequence would fold to a structure similar to the 'forward' protein. An inverted protein has the same amino acid composition, hydropathy profile and periodicity as the corresponding forward protein, and the periodicity of helices is maintained.

As early as 1986, Taylor [2] supposed to use inverted sequences as a control group resembling real proteins. Later, it was suggested that inverted proteins might have the mirror image structure of the original protein [3], but later computational lattice model studies on the inverted B domain of protein A by Olszewski et al. [4] revealed that it should adopt the same structure as the forward protein but be somewhat less stable. Olszewski also argued that a mirror structure is not feasible due to the occurrence of left-handed α-helices in the mirrored structure.

Different groups then started to synthesize inverted proteins. Lacroix et al. [5] synthesized the inverted B domain of protein A, the SH3 domain of α-spectrin and the B1 domain of protein G – proteins of three different folding types. They found all three inverted proteins to be unfolded. Inverted ubiquitin also does not fold, neither if synthesized synthetically (unpublished results) nor if expressed in *Escherichia coli* [6]. Witte et al. [7] again synthesized the inverted B domain of protein A and found it to be stable after addition of trifluoroethanol which stabilizes α-helical structures. Circular dichroism (CD) spectroscopy revealed that the protein adopts a similar structure to the forward B domain. Another protein showed to adopt a stable fold when inverted is the GCN4 leucine zipper (35 amino acids) [8]. GCN4 was the first retro-protein crystallized [9]. The structure of the dimer could be perfectly superposed with the structure of the native protein (root mean square deviation (rmsd) 0.37 Å), and the inverse protein forms stable tetramers.

Another functional protein with inverted parts is α-hemolysin (175 amino acids) with a retro-transmembrane domain of 25 amino acids [10]. The mutant protein formed functional pores with similar properties to the wild-type protein.

It was shown that secondary structure prediction methods in most cases give coinciding predictions for inverted proteins and their forward analogs, and that the prediction accuracy rises by 4% if forward and backward predictions coincide [11].

In contrast to proteins, much more is known about smaller peptides with inverted sequences. Retro-inverso-peptides (reversed peptides consisting of D instead of L amino acids) mostly have structures very similar to the original peptide and have a long history as peptidomimetics (for review see [12,13]). Retro-inverso-analogs of peptides can even bind to antibodies raised against the original protein and stimulate antibody formation themselves [14].

Many authors also found that peptides and even whole proteins consisting of D amino acids form the mirror image structure of the respective L amino acid protein (for review see [15]). A clear result was provided by the synthesis of the complete D-human immunodeficiency virus (HIV) protease which showed reciprocal chiral specificity on peptide substrates and inhibitors [16].

In general, it remains unclear whether proteins fold if their sequence is inverted and whether the resulting structure is related to the structure of the original protein. The question

*Corresponding author. Fax: (49)-30-450 528942.
E-mail address: cornelius.froemmel@charite.de (C. Frömmel).

Abbreviations: PDB, Protein Data Bank; ISS, inverse sequence similarity

0014-5793/03/$22.00
doi:10.1016/S0014-5793(03)00450-2

if and to what extent the direction of the protein chain is important for folding still remains.

2. Materials and methods

Coordinate files of the proteins were taken from the PDB [17,18]. Culled subsets of the PDB with several sequence identity thresholds were obtained from the internet site of the Dunbrack group [19]. Sequences and structures of the domains in the SCOP database [20] with a homology threshold of 40% were obtained from http://astral.stanford.edu/.

Sequence alignment was performed by using the gapped BLAST algorithm [21,22] with standard parameters (BLOSUM62 matrix [23–25], gap creation penalty 11, gap extension penalty 1 with filtering for low complexity). Hits with an expectation value of ≤ 10 were regarded as promising sequence alignments.

To make the forward alignments comparable to the inverse ones, both should have a similar score and length distribution. For each inverse alignment, we chose out of the pool of forward alignments the one which has the smallest deviation in length and score to the inverse alignment. Each forward alignment was allowed to appear only once in the data set. The new data sets contain equal numbers of alignments which have the same length and score distribution.

As a further control, each sequence was shuffled. The new sequences were aligned against the original sequences again. The shuffling conserves the length and amino acid composition of the proteins.

To estimate the relevance of the frequency of amino acid tuples in the query protein, we also generated new sequences with the same length distribution as in the original data set by using Markov models conserving the frequency of duplets, triples and quadruples of amino acids. These sequences were also aligned against the original forward sequences. To test the structural relevance of an alignment, we calculated the rmsd of the superimposed Cα atoms corresponding to each other in the alignment [26].

To test whether the optimal structural superposition of two structures without prior allocation of corresponding atom pairs leads to the same result as obtained by sequence alignment, we used the superposition algorithm described in [27] with a cut-off of 4 Å.

The application and all data can be obtained from the authors on request.

3. Results

Since many proteins and their strong relatives occur several times in the PDB and would lead to a biased statistic, we decided to use a culled PDB data set of 3904 proteins with a sequence identity threshold of 50% [19]. As indicated in Table 1, we found 1.25 times as much inverse than shuffled ('by chance') alignments. This shows that there exists considerable inverse similarity between proteins. Taking the number of shuffled alignments as background, 54% of the forward alignments and 20% of the inverse alignments are beyond the threshold of pure chance. Also if the shuffling procedure is done using Markov models conserving the frequency of amino acid duplets, triples and quadruples, the number of inverse alignments still exceeded the number of shuffled alignments (Table 1). The distribution of scores and lengths in the shuffled alignments was identical to the inverse alignments (Fig. 1). Since the number of inverse alignments was even higher than the number of alignments with proteins of quadruple frequencies equal to native proteins, we think that inverted sequences not only resemble native proteins concerning amino acid pairs, triples and quadruples, but also obey some more complex sequence rules. These might concern long range correlations or periodicities.

Table 1
Numbers of alignments against the PDB (culled with homology threshold of 50%) using different data sets

Data set	number of alignments
Shuffled	21 761
MM	20 791
MM (pairs maintained)	21 669
MM (triples maintained)	23 130
MM (quadruples maintained)	24 469
Inverse	27 295
Forward	47 303

'Shuffled': obtained by shuffling the amino acids of each single protein; MM: obtained using Markov models.

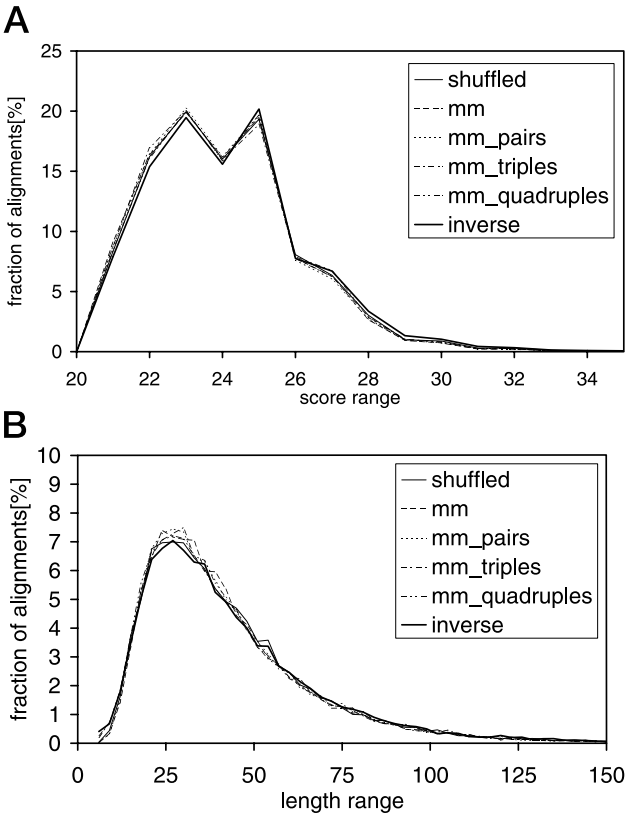

Fig. 1. Score (A) and length (B) distribution of inverse and shuffled alignments and alignments between proteins generated by Markov models conserving the amino acid pairs, duplets and quadruples, respectively. His tags have been removed. The diagrams show the fractions of alignments in windows of one score unit or three length units, respectively. The distributions are very much alike.

To avoid ambiguous inverse sequence alignments, we only consider sequence alignments which reach a better score in inverse than in forward direction. For this purpose, every pair of inversely aligned sequences was also aligned in the forward direction. We get 26 315 sequence pairs with better score in inverse direction.

We generated a subset of the forward sequence alignments (see Section 2) with an equal number of entries and showing score and length distribution equal to the inverse sequence alignments.

We searched for differences between the forward and inverse sequence alignments. For both, we compared the length of continuous patches of identical or similar amino acid residues (denoted with '+' by BLAST), identical, hydrophilic and hydrophobic patches, the distribution of gaps with different lengths, the frequencies of single amino acids and the matching frequency per amino acid and between different amino

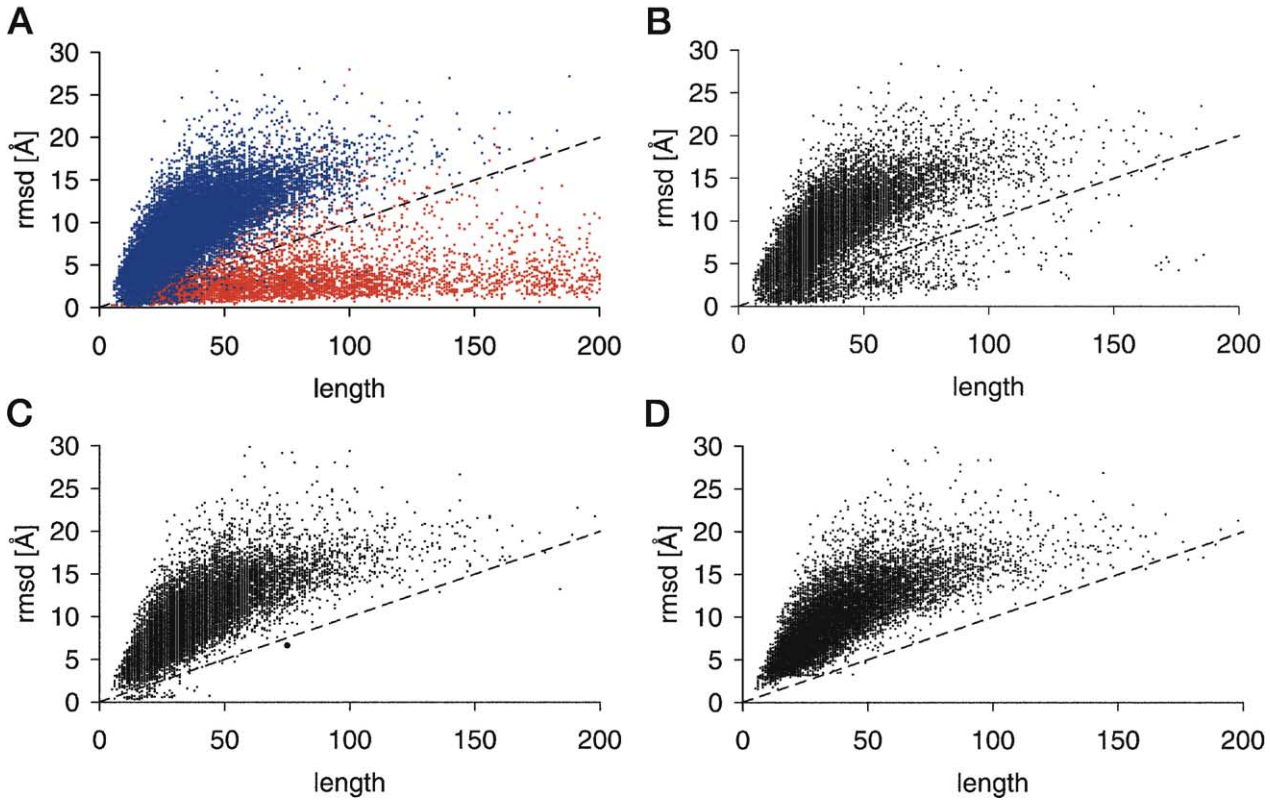

Fig. 2. Rmsd of aligned protein parts vs. length of the alignment without gaps. A: Domains from SCOP database. Red: belonging to the same superfamily; Blue: belonging to different superfamilies. B: Forward alignments from culled PDB. C: Inverse alignments from culled PDB. D: Mirrored structures of inversely aligned proteins. Line: upper threshold for 'good' structural superposition. Highlighted dot in C: example alignment of Fig. 3.

acids. No significant difference between forward and inverse sequence alignments could be found.

Obviously, there exists considerable ISS in proteins. The important question is whether this is coincided with structural similarity, or, in other words, whether this sequence similarity is *meaningful*. To evaluate this question, we superimposed the corresponding Cα atoms of the aligned regions of both proteins. To test the hypothesis that inverse sequences form the mirror image of the forward structure, we also superposed the mirror images of the inversely aligned proteins with the forward proteins.

To estimate which rmsd is indicative of a similar fold between both aligned parts, we aligned and superposed 3600 domains from the SCOP database [20] with less than 40% sequence identity by the same procedure (see Section 2). The SCOP database has the advantage that the comprising domains are divided in superfamilies, so one can tell whether two domains share the same fold or not. If the rmsd between these domains is plotted against the length of the alignment (not counting the gaps), the alignments can be clearly separated into two groups. A threshold of an rmsd (in Å) less than a tenth of the alignment length accurately separates proteins of the same superfamily from pairs of unrelated proteins (Fig. 2A). Not counting alignments of a domain with itself, the threshold line leads to an identification of 6996 out of 8369 (83.6%) alignments between proteins belonging to the same superfamily or identificates 28 109 of 28 549 (98.5%) of alignments between proteins of different superfamilies (Table 2) and can thus be considered as reasonable.

The same plot of the forward alignments of the PDB (Fig. 2B) also clearly shows two distinct groups of alignments. Applying the same threshold line, 1731 of the aligned protein pairs have a similar structure to each other. In contrast, the inverse alignments (Fig. 2C) only contain 99 pairs of proteins below the threshold line. 72 of them represent helical struc-

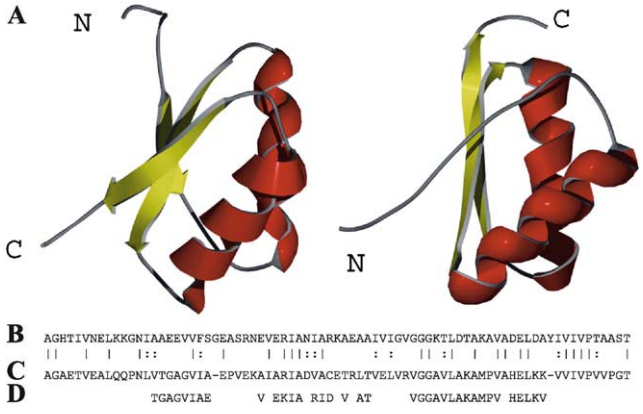

Fig. 3. Inverse similarity of aldolase (1EUA) and glycerol dehydrogenase (1JQ5). A: Both proteins show very similar structures with opposite backbone direction. B: Sequence of the aligned part of 1JQ5. C: Part of the inverted 1EUA aligned to 1JQ5 by BLAST. D: Amino acids of 1EUA associated to 1JQ5 by structural alignment. The superposition algorithm detects that the proteins fit above each other in opposite direction and leads to an allocation of amino acids similar to the one found by the inverse BLAST run (compare lines C and D).

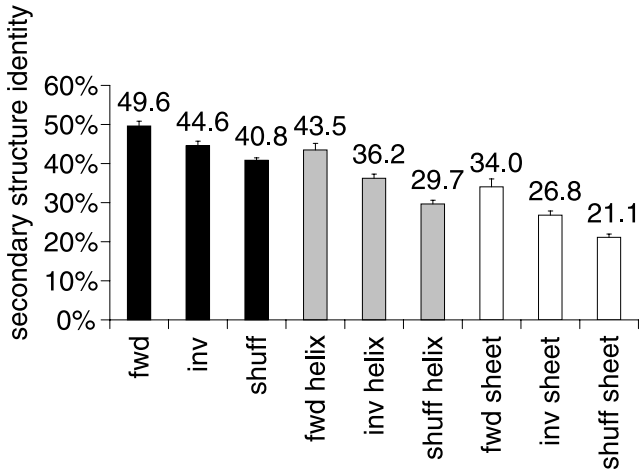

Fig. 4. Histograms showing the percentage of identical secondary structures between forwardly and inversely aligned proteins as well as shuffled proteins (the secondary structure classification of each amino acid was maintained). Black bars: overall percentage of identical secondary structure; gray bars: percentage of α-helical conformation in amino acids aligned to α-helices; white bars: percentage of β-sheet conformation in amino acids aligned to β-sheets.

tures up to a length of 49 amino acid residues. Superposition of mirrored structures (Fig. 2D) in only 17 cases leads to rmsd values below the threshold line.

Most of the alignments of SCOP domains below the threshold line (94%) are alignments between proteins of the same superfamily (Table 2). In the case of inverse sequence alignments, only seven examples showing sizes in the domain range (more than 50 amino acid residues) appear below the threshold line. Two of them represent pairs of protein structures which are similar only concerning their general shape: 1,4-β-D-xylan-xylanohydrolase with transcription factor malt (PDB codes 1XYZ and 1HZ4, respectively; alignment length 184) and alkaline protease with rhamnogalacturonase A (PDB codes 1KAP and 1RMG, respectively; alignment length 136). The first two proteins form horseshoe-like structures, the second pair represents longer cylindrical structures built from β-sheets. In both pairs the elements of secondary structures are not superimposed. In consequence, both examples should be considered as false positives.

As the best example, the structures of aligned parts of aldolase and glycerol dehydrogenase (PDB codes 1EUA and 1JQ5, respectively), the highlighted dot in Fig. 2C, are shown in Fig. 3. In the inverse BLAST run, the score was 29.6 while the score of the corresponding forward sequence alignment is only 13.9. A superposition of the corresponding 76 Cα atoms leads to an rmsd of 6.6 Å whereas a superposition in forward direction gives an rmsd of 11.4 Å. An automatic optimal superposition without allocating the corresponding atom pairs of both proteins overlays 41 of 76 Cα atoms with an rmsd of 1.73 Å. The algorithm automatically detects that the proteins have to be superimposed in backward and not in forward direction, and the found structural alignment resembles the inverse sequence alignment (Fig. 3).

Since the secondary structure prediction of inverted sequences yields nearly the same results as for the respective native proteins, we analyzed the fraction of secondary structure identity of inversely versus forwardly aligned and randomized proteins of the same score and length quality (Fig. 4). In all cases, i.e. summing all secondary structure elements as well as considering only helices or sheets, the inverse alignments lie between shuffled and forward alignments. Note that the fraction seems to be relatively high in the randomized alignments due to the fact that a high proportion of proteins adopt coiled, helical or sheet conformation.

Table 2
Numbers of 'true' (same superfamily) and 'false' (different superfamily) alignments between domains of the SCOP database with less than 40% sequence identity

	# Total	# True	# False
Scop	36 918	8 369	28 549
Scop uT	7 436	6 996	440
Scop aT	29 482	1 373	28 109

Alignments of a protein with itself are not counted. uT: under threshold line, rmsd ≤ length/10; aT: above threshold line, rmsd > length/10.

4. Discussion

In the PDB, one finds a surprising number of inverse sequence alignments showing a similarity in the so-called twilight zone of 20–35% sequence identity [28]. Amazingly, the number of inverse sequence alignments is clearly larger than one would expect by chance. The reason remains unknown.

In contrast to further studies which focused on sequence comparison [2], we here show a detailed large scale three-dimensional (3-D) superposition of all inversely sequence similar protein pairs within a set of proteins from the PDB. The resulting rmsd values are compared to rmsd values obtained by superposing forwardly aligned proteins. To be sure of not missing eventually structurally similar pairs with low ISS, we superposed all found pairs regardless of alignment score and length. As a control, we also studied alignments between shuffled sequences and sequences with pair correlations equal to real proteins. We further tested the hypothesis that proteins with inverted sequences might form the mirror image of the native protein.

Our data clearly show that the occurrence of ISS is not indicative for 3-D similarity. The fraction of meaningful forward sequence alignments (9.0%) was much larger than the fraction of meaningful inverse sequence alignments (0.5%). This is astonishing since all neighbor effects of the comprising amino acids as well as the amino acid composition are conserved in inverted sequences. Our results clearly exclude mirrored protein structures of inverted sequences. Furthermore, the very low number of similar 3-D structures between inversely similar proteins suggests that the inversion of protein sequences probably leads to unfolded proteins or structures with different folds than the original protein. Inversion of protein sequences can thus be considered as a shuffling method to get a reference group with conservation of the amino acid composition, all neighbor effects and periodicities but having folding characteristics of random proteins [2].

Interestingly, the GCN4 leucine zipper which was shown to keep its 3-D fold upon inversion of the sequence shows no ISS with itself: The GCN4 leucine zipper with the PDB code 1GCM was included in our data set but yielded no inverse BLAST hit with itself. This supports the finding that short helices keep their conformation when the sequence is inverted (note the 72 examples of short helices we found). The high

proportion of folded helices with inverted sequences suggests that the folding of helices is probably only weakly dependent on sequence order. Obviously, helices are much more stable to inversion than other structures. Therefore, we suppose that the folding capacity of a protein is not primarily determined by the helices – in contrast, turns seem to play a much more important role.

Acknowledgements: We thank Andrean Goede and Mathias Dunkel for supplying the superposition algorithm and Clemens Gröpl for fruitful discussions about shuffling sequences. The work was supported by the 'Berliner Centrum für Genombasierte Bioinformatik' (BCB).

References

[1] Preißner, R., Goede, A., Michalsky, E. and Frömmel, C. (1997) FEBS Lett. 414, 425–429.
[2] Taylor, W.R. (1986) J. Mol. Biol. 188, 233–258.
[3] Guptasarma, P. (1992) FEBS Lett. 310, 205–210.
[4] Olszewski, K.A., Kolinski, A. and Skolnick, J. (1996) Protein Eng. 9, 5–14.
[5] Lacroix, E., Viguera, A.R. and Serrano, L. (1997) Fold. Des. 3, 79–85.
[6] Blöcker, H., personal communication.
[7] Witte, K., Skolnick, J. and Wong, C.-H. (1998) J. Am. Chem. Soc. 120, 13042–13045.
[8] Mittl, P.R.E., Deillon, C., Sargent, D., Liu, N., Klauser, S., Thomas, R.M., Gutte, B. and Grütter, M.G. (2000) Proc. Natl. Acad. Sci. USA 97, 2562–2566.
[9] Liu, N., Deillon, C., Klauser, S., Gutte, B. and Thomas, R.M. (1998) Protein Sci. 7, 1214–1220.
[10] Cheley, S., Braha, O., Lu, X., Conlan, S. and Bayley, H. (1999) Protein Sci. 8, 1257–1267.
[11] Park, J., Dietmann, S., Heger, A. and Holm, L. (2000) Bioinformatics 16, 978–987.
[12] Chorev, M. and Goodman, M. (1993) Acc. Chem. Res. 26, 266–273.
[13] Fletcher, M.D. and Campbell, M.M. (1998) Chem. Rev. 98, 763–796.
[14] Phan-Chan-Du, A., Petit, M.C., Guichard, G., Briand, J.P., Muller, S. and Cung, M.T. (2001) Biochemistry 40, 5720–5727.
[15] Chorev, M. and Goodman, M. (1995) Trends Biotechnol. 13, 438–445.
[16] Milton, R.C., Milton, S.C. and Kent, S.B. (1992) Science 256, 1445–1448.
[17] Protein Data Bank, http://www.rcsb.org/pdb/.
[18] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucleic Acids Res. 28, 235–242.
[19] Culling the PDB, http://www.fccc.edu/research/labs/dunbrack/culledpdb.html.
[20] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[21] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[22] BLAST, http://www.ncbi.nlm.nih.gov/BLAST/.
[23] Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad. Sci. USA 89, 10915–10919.
[24] Henikoff, S. and Henikoff, J.G. (1993) Proteins 17, 49–61.
[25] Henikoff, S. and Henikoff, J.G. (2000) Adv. Protein Chem. 54, 73–97.
[26] Kearsley, S.K. (1989) Acta Cryst. A45, 208–210.
[27] Preissner, R., Goede, A. and Frömmel, C. (1999) Protein Eng. 12, 825–832.
[28] Rost, B. (1999) Protein Eng. 12, 85–94.