Inverse sequence similarity in proteins and its relation to the three-dimensional fold

R. Preißner, A. Goede, E. Michalski, C. Frömmel*

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

Received 30 April 1997; revised version received 9 July 1997

Abstract Nowadays the most successful strategy for the prediction of the tertiary structure of proteins is the homology-based modelling using known structures. A real chance to predict the general fold of a protein arises only in cases with a sufficient sequence homology (e.g. 27% over 100 residues). In this analysis we examine the phenomenon of inverse sequence similarity (ISS) in proteins and its structural meaning. In sequence data bases we found a lot of examples for ISS up to 34% identity over 204 residues and a surprisingly large number of self-inverse protein sequences. By inspection of inverse similar sequence pairs with known tertiary structures we observe that inverse sequence alignments above the threshold indicating structural similarity generally do not imply comparable folds for both. From our analysis we conclude that the straightforward employment of ISS for protein structure prediction fails even above the known threshold for 'safe similarity'.

© 1997 Federation of European Biochemical Societies.

Key words: Protein sequence; Structure; Database; Homology

1. Introduction

Until now, molecular biologists have identified the complete amino acid sequences of more than 100000 proteins. But only less than 2000 3-D structures of proteins on level of atomic resolution are known [1]. The most reliable method to predict 3-D protein structure on the basis of the 1-D sequence is 'homology modelling' [2]. It is based on the detection of significant similarities of an amino acid sequence to such of a protein of known 3-D structure. The threshold of sequence similarity sufficient for structural similarity depends particularly on the length of aligned stretches in the sequence [3]. It is widely accepted that the amino acid composition in a distinct segment of the protein has a strong influence on the type of secondary structure of this segment gained in the folded structure [4,5]. In principle most methods of the secondary structure prediction are relying on this assumption. Recently by theoretical consideration it was conjectured that a protein with identical composition but with backward read primary structure should fold under native conditions to a similar structure compared with the original sequence [6]. This grid based analysis gives rise to the hope that structure prediction by homology modelling is possible on the basis of inverse sequence similarity (ISS). Although inverse peptide sequences are discussed by peptide chemists [7], they were not subjected to a detailed study in proteins.

Assuming that the threshold for structural similarity in sequence homology reflects only physico-chemical laws of protein folding it could be expected that inverse similar protein sequences fold in comparable manner.

In this paper, the analysis of sequence and structure databases of proteins shows that (self-)ISS frequently occurs and requires detailed consideration to evaluate its impact on model building studies.

2. Materials and methods

During computer modelling of inverted primary structures in the original 3-D topology, one is faced with several problems [6]. To adopt identical local structural elements the values of each pair of main chain torsion angles (φ,ψ-angles) had to be interchanged. Generally the exchange of φ and ψ corresponds to a φ,ψ-map which was mirrored at the diagonal from (−180°,−180°) to (180°,180°) (Fig. 1). Thus, values nearby the diagonal and in the other two corners will appear mirrored in allowed regions. In this respect, the exchange of φ and ψwould be suitable for α-helical conformations because typical φ,ψ-values might be around (−60°,−60°) [8] (see Fig. 2). For extended strands with φ,ψ-pairs of about (−170°, 170°) [9] an exchange is also possible but energetically less favourable because of some steric hindrance [10]. This difficulty could partly be overcome by turning the peptide bond (torsion angle ω) by few degrees out of planarity. Due to distinct φ,ψ-combinations in different types of loops their structure had to be changed dramatically [11]. According to steric hindrance by side chain atoms in loops and distorted H-bond patterns a simple interchange of main chain torsion angles is mostly impossible and particularly the fixation of φ-angle of prolyl residues to a value of about −60 ± 30° does often not allow the interchange of the main chain torsion angles. Furthermore the possibility of a flipped ω-torsion angle in prolyl residues (resulting in *cis*-conformation of the peptide bond) leads to further difficulties in model building for inverse sequences on the basis of the wild-type structure. During folding of inverse sequences the correct formation of secondary structure elements may be impaired due to wrong positions of special folding signals like helical caps, proline and glycine positions, respectively, in loops, and side chain charges in relation to both ends of the α-helices [12]. Also the packing of secondary structures to each other probably becomes difficult in inverted sequences due to the changed position of the Cβ-atom pointing in the direction of the former Hα-atom in the original structure. In helices this pseudo-rotation could be revoked by turning the helices around their axis. In β-sheets a translational shift of the partner is necessary to remain correct contacts. It is an open question whether an inversely oriented peptide chain is still able to fold.

We analysed inverse sequence similar pairs of structurally known proteins. For this purpose more than 4000 sequences of all protein chains were extracted from the Brookhaven Protein Data Bank (PDB) [13] and inverted. Then we searched for similarity of these inverted sequences in the PDB and Swissprot sequence database [14], respectively.

The results of the search were analysed with respect to the rate of identical amino acids at identical positions, quality score (Q) [15], the z-score which was found to be most selective and sensitive [16], and E(), the probability to find better alignments purely by chance [17]. For the alignment the PAM250 matrix was used. The influence of different scoring matrices (BLOSUM50) turned out to be low for the given examples as stated by Vogt et al. [18].

*Corresponding author. Fax: (49) 30-2802-6615.
E-mail: froemmel@rz.charite.hu-berlin.de

PII S0014-5793(97)00907-1

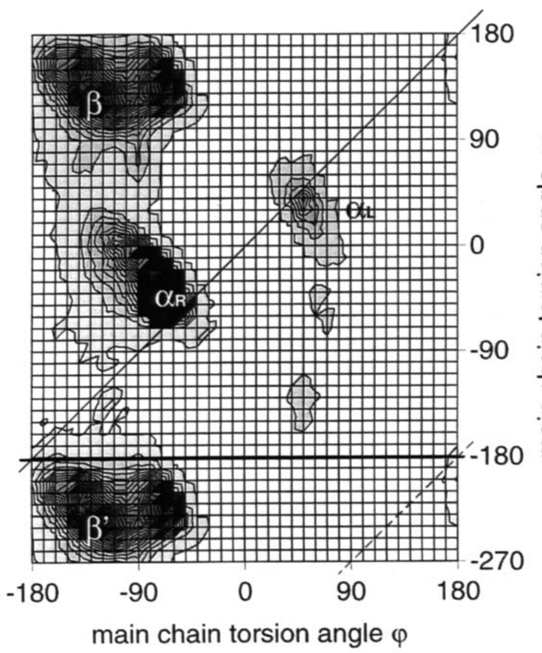

Fig. 1. Extended Ramachandran plot of 75 000 non-glycine residues from 400 non-redundant high-resolution structures [20] from PDB. The ψ-axis is elongated compared to usual plots to show that angles from the upper left corner appear near the diagonal. α_R denotes the α-helix, α_L the left-winded conformation, β marks the region typical for β-sheets, β′ represents the same data in the expanded part of the plot to show the distance to the diagonal.

3. Results and discussion

Surprisingly, for the (inverted) PDB sequences we found about 10^3 alignments in the Swissprot database above the threshold for 'safe structural homology' [3,19] indicating that inverse similarity is a widespread phenomenon (see Fig. 3). (The data are available via http://www.rz.charite.hu-berlin.de/ch/biochem/inverse.) If ISS could be used for structure prediction all of them would be candidates for homology based model building.

To check the structural significance of ISS we considered in detail those pairs with known 3-D structure for both. Totally 38 non-redundant pairs from PDB (see Tables 1 and 2) are found above the threshold indicative for similar 3-D structure whereof 21 are inverse self-similar. For none of the remaining 17 pairs (see Table 1) was a meaningful superposition of the protein backbone possible (e.g. see Fig. 4B). Neither sequence identity nor the quality score (Q) are in accordance with the observed structural similarity (see Fig. 4). Moreover, the secondary structure localisation and content, respectively are completely different. For ISS (without self-similarity) we find a mean of 40% secondary structure identity in pairs (22% SD). This value is close to the statistical expectation (37%; calculated from the secondary structure distribution in a non-redundant data set [20]). This gives evidence that model building based on ISS at the level of identity of sequen-

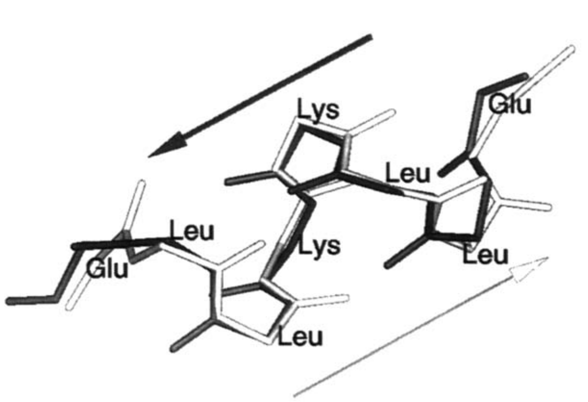

Fig. 2. X-ray structure from amphiphilic model helix (PDB code: 1all) superimposed with itself in opposite direction. The amino acids are designated (for both identical) near to their well-imposed Cα-atoms.

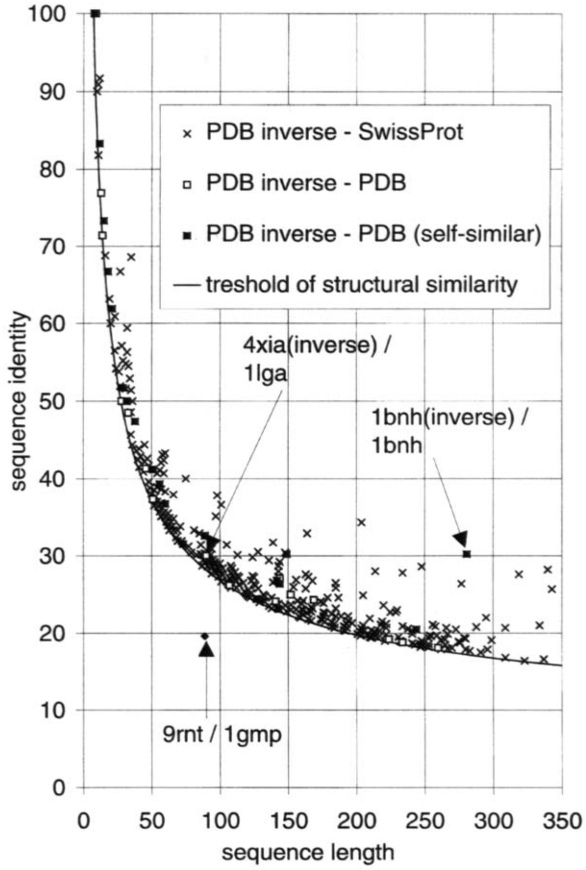

Fig. 3. Comparison of the sequence homology threshold (———) for structurally reliable alignments derived from known protein families [3,19] with the similarity between native and inverse sequences. Only values of sequence pairs showing sequence identity above the threshold are figured out. Redundant examples as well as examples with low sequence complexity like, e.g. cysteine-rich proteins, anti-freezing proteins (rich of alanine) or collagen (glycine, proline) were omitted. □ inverted sequence x from PDB aligned with sequence y from PDB [13], × inverse sequence x from PDB aligned with sequence y from Swissprot (release 33) [14], filled symbols (■) represent proteins for which inverse sequence shows similarity to the original primary structure itself, marked symbols represents the examples considered in detail in Fig. 2. The secondary structure identity was calculated according HSSP [3]. The up-to-date threshold t (%) calculated according $t = 290.15 * \text{length}^{-0.562} + 5$ gives the lowest value of sequence identity for which 'safe' 3-D structural homology can be assumed [19]. More than 4000 sequences were extracted from the Brookhaven Protein Data Bank (PDB) and inverted (each distinct peptide chain was considered separately). Homology search was done by the standard procedure FASTA [21]. In accordance with Pearson [19] and Landes [20] a gap penalty of 12 (gap extension of 4) and the widely used PAM250 matrix are the basis of the estimated quality scores (Q) for the alignments.

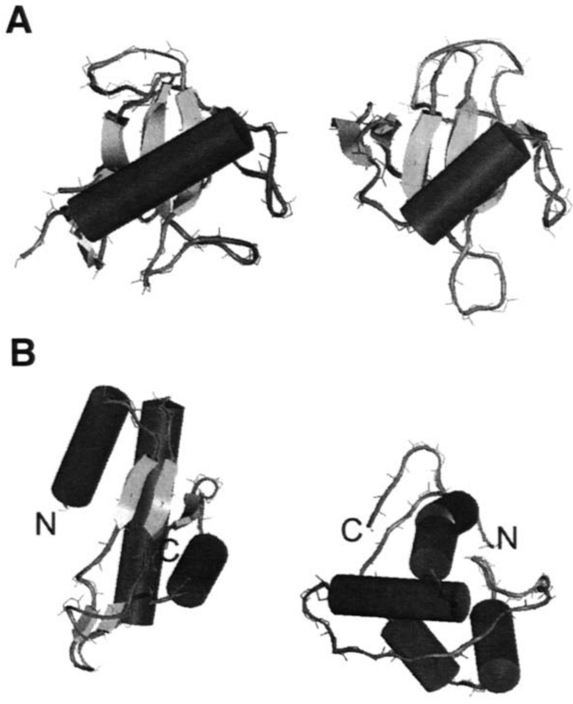

Fig. 4. The structural comparison of two pairs of protein structures illustrating the missing significance of inverse sequence similarity. a: The two structurally homologous proteins ribonuclease T1 (PDB code: 9rnt) and ribonuclease Sa (PDB code:1gmp) with sequence identity 20% and secondary structure identity of 85% (*Q*-score = −82). b: The two structurally unrelated proteins xylose isomerase (PDB code: 4xia) and lignin peroxidase (PDB code: 1lga). The sequence identity between both original sequences is low (≈10%) in the considered region. Inverting one of the sequences the percentage of identical amino acids rises to 30% (*Q*-score = 41) with low structural impact (only 32% secondary structure identity). c: The sequence alignment of the two pairs from (a) and (b). The secondary structural elements are noted above and below, respectively. Identical residues at equal positions are in bold:

```
Structural homologous pair of microbial ribonucleases:

              EEE     HHHHHHHHHHH             EEE              EEEEE                EEEE      EEEE        EEEE
1gmp      DVSGTVCLSALPPEATDTLNLIASDG-PFPYSQ-DGVVFQNRESVLPTQSYGYYHEYTVI----TPGARTRGTRRIICGEATQEDYYTGDH-YATFSLIDQTC
9rnt      CGSNCYSSSDVSTAQAAGYQLHE-DGETV-GSNSYPHKYNNYEGFDFSVSSPYY-EWPILSSGDVYSGGSPGADRVVFNENNQLAGVI-THTGASGNNFVECT
             EEE HHHHHHHHHHHHHHHH             EEE             EE EEE               EEEEEE    EEEEE E        EE

Structural unrelated pair of sequences aligned with FASTA (lignin peroxidase inverted):

           EE         EE        HH HHHHH   HHHHHHHHHHHH                    EE          HHHHHHHHHHHHHH   HHHHHHHH
1lga-inv  FPLGQVTPDVDNVAAVSHAS-LMW-VLELEDFEGADNVRAIIQDVTHFPEPVLGDPAPQTAPKRGTFFNMQPAGPCNSLAVAGAFAIFDGPTVGHKQVFPKQ
4xia      FALAKVLHNIDLAAEMGAETFVMWGGREGSEYDGSKDLAAALD---RMREGV--DTAAGYIKDKG--YNLRIALEPKPNEPRG--DIF-LPTVGHGLAFIEQ
          HHHHHHHHHHHHHHH    EEEE     EE       HHHHHH   HHHHHH  HHHHHHHHHH      EEEE           EE     HHHHHHHH
```

d: The sequence alignment of the native sequence of ribonuclease inhibitor (PDB code: 1bnh) and its inverted pendant. The alignment was carried out by FASTA [21] using the PAM250 matrix. Identical residues at equal positions are in bold. The similarity of the secondary structure is low (30.2% identity) although all scores for different scoring matrices indicate the homology of the sequences, e.g. the results using the BLOSUM50 matrix are similar: *Q*-score 254, *z*-score 189.7, *E*() 0.00055.

```
scores:   z-score: 182.3 E(): 0.0014
Smith-Waterman score (PAM250): 186;    30.2% identity in 281 aa overlap


          EE     HHHHHHHHHHHHH       EEE        HHHHHHHHHHHH      EEE     HHHHH  HHHHHHHH           E    E
1bnh-inv  LEKLTEKAQLVRCLDRCGSATIDCEWLWLTKLRSAPSLLGPCLEAIGADGLGNSGLDLERLSAQSAVIG--CLDKCNAPTLGC----NELR----L
1bnh      LLPLLQQYEVVR-LDDCGLTEEHCKDI-GSALRANPSLTELCLRTNELGDAG-VHLVLQGLQSPTCKIQKLSLQNCSLTEAGCGVLPSTLRSLPTL
          HHHHH    EEE EE     HHHHHHH HHHH       EEE      HHHH HHHHHHHH        EEE             HHHHHHH

          E       HHHHHHHHHHHHH      EEE      HHHHHHHHHHHHH       EEE       HHHHHHHHHHHHH      EEE     HHH
1bnh-inv  TELQCASDALGQGLVRAGAEGIDNNSVTLEKLART-ARLVSALPECSAATLRCYELQLKELHCQPDLLGECLLRLGADGLPNDSLHLERLTPLSRL
1bnh      RELHLSDNPLGDAGLRLLCEGLLDPQCHLEKLQLEYCRLTAASCEPLASVLRA-TRALKELTVSNNDIGEAGARVLGQGLADSACQLETL----RL
          EEE      HHHHHHHHHHHHH       EEE       HHHHHHHHHHHHH      EEE      HHHHHHHHHHHHH       EE    E

          HHHH             EEE        HHHHHHHH HHHH      EEE       HHHH HHHHHHH     EE EEE    HHHHHH
1bnh-inv  TSPLVGCGAETLSCNQLSLKQIKCTPSQLGQLVLHV-GADGLENTRLCLETLSPNARLASG-IDKCHEETLGCDDL-RVVEYQQLLPLLE
1bnh      EN----CGLTPANCKDLC--GIVASQASLRELDLGSNGLGDAGIAELCPGLLSPASRLKTLWLWECDITASGCRDLCRVLQAKETLKELS
                    HHHHHHHH  HHHHH     EEE      HHHHHHHHHHHH        EEE       HHHHHHHHHHHHH     EEE
```

Table 1
Examples of inverse sequence similar proteins with known structure

Protein 1				Protein 2				Alignment	
PDB code[a]	Chain identifier	Name	Length	PDB code[a]	Chain identifier	Name	Length	Identity[b]	Length
1ala		annexin V (chicken)	316	1avh	a	annexin V (human)	318	18.1	260
1avh	a	annexin V	318	1min	a2	nitrogenase	437	18.8	234
3pgk		phosphoglycerate kinase	415	1php		phosphoglycerate kinase	394	19.2	224
1lga	a	lignin peroxidase	343	1did	b	xylose isomerase	393	24.3	169
1lga	a	lignin peroxidase	343	5xia	b	xylose isomerase	393	25.0	152
2wgc	a	agglutinin	171	2cwg	a	agglutinin	171	27.1	144
1dri		ribose-binding protein	271	7abp		arabinose-binding protein	305	24.1	141
2phh		monooxygenase	391	1xyb	a	xylose isomerase	386	26.2	107
1tta	a	pre-albumin	127	1cax	a	canavalin	181	37.3	51
6taa		α-amylase	476	1min	a2	nitrogenase	437	37.3	51
1min	a2	nitrogenase	437	1btc		α-amylase	491	41.3	46
1atf		antifreeze protein	37	2mad	h	methylamine dehydrogenase	124	48.5	33
4mt2		metallothionein (black rat)	61	1mhu		metallothionein (human)	31	48.4	31
1tca		lipase	317	1maf	h	amine dehydrogenase	124	50.0	28
1gal		glucose oxidase	581	4cpa	a	carboxypeptidase A	307	71.4	14
1dgc	a	Gcn4 leucine zipper	55	1php		phosphoglycerate kinase	394	76.9	13
1rtp	a	α-parvalbumin	109	1scc		cytochrome p450	482	100.0	9

[a]PDB code of the proteins.
[b]Percentage over the length of the alignment.

ces considered here is not possible (20–30% identity; 100–200 amino acids). In consequence it can be concluded that the threshold must be different for correctly oriented and back read sequences. Either the folding to similar structure is observed only at higher degree of sequence identity (at least 5–10% higher, see Fig. 3) or the inverse sequences do not fold at all into any related structure compared to their originals.

Twenty-one cases of significant 'self-ISS' were found in the PDB (see Table 2). We have no general explanation for this observation. In complex sequences of globular proteins the occurrence of such inverse-native self-similarity by chance is very improbable. Interestingly in proteins showing self-ISS a mean secondary structure identity of 74% was found (25% SD), which may partly reflect a special content of secondary structure in the particular protein (e.g. dominating helix). A number of those proteins show symmetrical structural features as found by visual inspection (e.g. ribonuclease inhibitor, PDB code: 1bnh; for alignment see Fig. 4D).

The analysis given here leads to two new learning sets for protein design. Additional information why sequences do (not) fold into the expected structure may be achieved from inverse as well as from self-ISSs. Candidates for further detailed studies are the distribution of similar residues along the peptide chain, the positioning of key residues and the evaluation of alignment scores.

Nevertheless our analysis shows that protein homology (= similarity with evolutionary background; existence of a common ancestor) is more indicative for resembling structures

Table 2
Examples of self-ISS proteins with known structure

Protein				Alignment	
PDB code[a]	Chain identifier	Name	Length	Identity[b]	Length
1bnh		ribonuclease inhibitor	456	30.2	281
1llc		lactate dehydrogenase	320	20.5	244
2cwg	a	agglutinin	171	26.4	144
1deg		calmodulin	142	23.1	143
2wgc	a	agglutinin	171	26.8	142
1dpi	1	DNA polymerase	1065	24.4	127
2tma	a	tropomyosin	284	32.6	89
1tme	a	encephalomyelitis virus	256	31.5	73
1le2		apolipoprotein E2	144	36.7	60
4mt2		metallothionein	61	40.4	57
1bod		calbindin	74	39.3	56
1gd1	o	glycerald. dehydrogenase	334	41.2	51
1sha	a	tyrosine kinase transforming protein	103	47.4	38
1atf		antifreeze protein	37	50.0	32
1cpb	2	carboxypeptidase B	217	51.7	29
1efm	2	elongation factor Tu	393	61.9	21
1dgc	a	Gcn4 leucine zipper	55	66.7	18
1snw	a	sindbis virus capsid protein	151	73.3	15
2act		actinidain	218	83.3	12
1bbe	a	collagen	12	100.0	11
1al1		amphiphilic α-helix (synthetic)	12	100.0	8

[a,b]See Table 1.

than mere sequence similarity. The scores given in Fig. 4D illustrate that none of it is adequately predictive for structural similarity on the basis of ISS.

Although completely or partially inverted sequences will have to be examined experimentally, this study shows that an ISS does not necessarily result in similar protein 3-D structure and that a degree of ISS which normally would be highly significant for structurally related proteins is not sufficient to indicate structural resemblance. Therefore, the use of the inverse sequence space for straight-forward structure prediction of proteins is not practicable.

References

[1] Holmes, L. and Sander, C. (1996) Science 273, 595–602.

[2] Eisenhaber, F., Persson, B. and Argos, P. (1995) CRC Crit. Rev. Biochem. Mol. Biol. 30, 1–94.

[3] Sander, C. and Schneider, R. (1991) Prot. Struct. Funct. Genet. 9, 56–68.

[4] Eisenhaber, F., Imperiale, F., Argos, P. and Frömmel, C. (1996) Prot. Struct. Funct. Genet. 25, 157–168.

[5] Eisenhaber, F., Frömmel, C. and Argos, P. (1996) Prot. Struct. Funct. Genet. 25, 169–179.

[6] Olstewski, K.A., Kolinski, A. and Skolnick, J. (1996) Prot. Eng. 9, 5–14.

[7] Wermuth, J., Goodman, S. and Kessler, H., in: H.L.S. Maia (Ed.), Proc. 23rd Eur. Pept. Symp., ESCOM, Leiden, 1994, pp. 648–649.

[8] Pauling, L., Corey, R.B. and Branson, H.R. (1951) Proc. Natl. Acad. Sci. USA 37, 205–211.

[9] Salemme, F.R. and Weatherford, D.W. (1981) J. Mol. Biol. 146, 101–141.

[10] Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) J. Mol. Biol. 7, 95–99.

[11] Müller, G., Gurrath, M., Kurz, M. and Kessler, H. (1993) Prot. Struct. Funct. Genet. 15, 235–251.

[12] Sali, D., Bycroft, M. and Fersht, A.R. (1988) Nature 335, 740–743.

[13] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.

[14] Bairoch, A. and Boeckmann, B. (1994) Nucl. Acids Res. 22, 3578–3580.

[15] Pearson, W.R., in: R.F. Doolittle (Ed.), Methods in Enzymology, Vol. 183, Academic Press, San Diego, CA, 1990, pp. 63–98.

[16] Landes, C., Henaut, A. and Risler, J.-L. (1992) Nucleic Acids Res. 20, 3631–3637.

[17] Pearson, W.R. (1995) Prot. Sci. 4, 1145–1160.

[18] Vogt, G., Etzold, T. and Argos, P. (1995) J. Mol. Biol. 249, 816–831.

[19] HSSP-database, Schneider, R. and Sander, C.: http://www.sander.embl-heidelberg.de/hssp/

[20] Kleyweigt, G.J. and Jones, T.A. (1996) Structure 4, 1395–1400.

[21] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.