Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 1075–1078, February 1984
Biochemistry

On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations

(cooperativity/protein folding/amino acid sequence homology)

WOLFGANG KABSCH AND CHRISTIAN SANDER

Department of Biophysics, Max Planck Institute of Medical Research, 6900 Heidelberg, Federal Republic of Germany

Communicated by Sir John Kendrew, November 9, 1983

ABSTRACT **The search for amino acid sequence homologies can be a powerful tool for predicting protein structure. Discovered sequence homologies are currently used in predicting the function of oncogene proteins. To sharpen this tool, we investigated the structural significance of short sequence homologies by searching proteins of known three-dimensional structure for subsequence identities. In 62 proteins with 10,000 residues, we found that the longest isolated homologies between unrelated proteins are five residues long. In 6 (out of 25) cases we saw surprising structural adaptability: the same five residues are part of an α-helix in one protein and part of a β-strand in another protein. These examples show quantitatively that pentapeptide structure within a protein is strongly dependent on sequence context, a fact essentially ignored in most protein structure prediction methods: just considering the local sequence of five residues is not sufficient to predict correctly the local conformation (secondary structure). Cooperativity of length six or longer must be taken into account. Also, we are warned that in the growing practice of comparing a new protein sequence with a data base of known sequences, finding an identical pentapeptide sequence between two proteins is not a significant indication of structural similarity or of evolutionary kinship.**

The folding process of a globular protein is highly selective: a long amino acid chain ends up in one or a few out of a huge number of possible three-dimensional conformations. In contrast, conformational preference of single amino acid residues is weak. So the high selectivity of protein folding is only possible through the interaction of many residues. What is the minimum number of residues for folding into a unique structure? Five residues can form more than one turn of an α-helix or an entire β-strand. Cooperativity of length five could therefore be deemed sufficient for stabilization of secondary structure. We were thus surprised by the discovery of amino acid sequences of length five that in one protein are in α-helical and in another, quite different protein, in β-sheet conformation.

Search for Identical Oligopeptides

Detailed knowledge of protein structure comes mainly from the more than 100 three-dimensional structures solved by x-ray crystallography (1). Of these, at least 62 are known to sufficient resolution to allow assignment of details of hydrogen-bonded (secondary) structure (2). A systematic search among these 62 proteins revealed 25 subsequence identities of length five, not counting identities flanked by more extensive sequence homology. For example, the sequence Lys-Val-Leu-Asp-Ala occurs both in prealbumin and in carbonic anhydrase, two proteins otherwise unrelated in sequence, function, or structural type.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Probability of Occurrence. A systematic search for identical pentapeptides seems not to have been undertaken previously, perhaps because one intuitively, but wrongly, does not expect to find any. An old party joke is analogous: one intuitively does not expect to find two identical birthdays in a group of 30 randomly chosen people, yet one can, statistically, be 70% certain of finding them by chance (3). What matters is not that the number of people is small compared with the number of days in a year but that the number of all possible pair comparisons among 30 people, 435, is larger than 365.

By analogy, the occurrence of identical pentapeptide sequences in unrelated proteins in the data base is actually highly probable: although the number of residues in the data base of known protein structures (10^4) is very small compared with the number of possible pentapeptide sequences ($20^5 = 3.2 \times 10^6$), the number of pair comparisons (5×10^7) is larger. More precisely, assume we have n amino acid residues arranged in random order with, for simplicity, frequency of occurrence of $p = 1/20$ each. The probability that all possible pairs of peptides of length l are different is

$$P = (1 - p^l)^N, \qquad [1]$$

where $(1 - p^l)$ is the probability for two peptides of length l to be different and $N = n(n - 1)/2$ is the number of pair comparisons. The formula holds for $n << 1/p^l$ and neglects the complication of overlap of successive pentapeptides. For pentapeptides ($l = 5$) in the data base of known protein structures ($n = 10{,}000$), we calculate $P = 2 \times 10^{-7}$. The probability to find at least one pair is $1 - P = 0.9999998$. Thus the odds are more than 1,000,000:1 for finding one or more such pairs, and indeed we find 25 pairs.

For hexapeptides, the odds for finding one or more pairs in unrelated proteins are 1:1 ($P = 0.5$, $1 - P = 0.5$). Consistent with that, within the data base the systematic search yields none. Outside the data base, without a systematic search, we have so far found two: (*i*) Cys-Arg-Asp-Lys-Ala-Ser are residues 63–68 in rhodanese (data set 1RHD) and residues 151–156 in a gonococcal pilus protein (Thomas Meyer, personal communication) and (*ii*) Gly-Tyr-Ile-Thr-Asp-Gly are residues 92–97 in actinidin (data set 2ACT) and residues 68–73 in phage T4 DNA ligase (John Armstrong, personal communication).

Conformation of Identical Pentapeptides

How similar in conformation are the discovered identical pentapeptides? A good objective measure of similarity of three-dimensional structure is the minimum rms distance [d(rms)] between the C(α) atom positions of equivalent residues obtainable by rotation and translation (4). Table 1 gives the pentapeptide pairs in their sequence context ordered in terms of decreasing d(rms)—i.e., increasing structural similarity. Considering that five residues can be more than one turn of an α-helix and that many β-strands are not longer

Table 1. Conformation of identical pentapeptides occurring in two different proteins

Different conformation in different proteins ($d > 2.5$ Å)					Similar conformation in different proteins ($d < 2.1$ Å)				
d, Å	Protein subsequence (conformation)	Protein (secondary structure of pentapeptide)	Data set	Residues	d, Å	Protein subsequence (conformation)	Protein (secondary structure of pentapeptide)	Data set	Residues
4.5	NIEAD VNTFV ASHKP hhh hhhhh hhhgg	erythrocruorin (hemogl.) alpha-helix	1ECD	80-84	2.0	CCQEA YGVSV IVGVP tb tt t eee e	alcohol dehydrogenase	4ADH	286-290
	DRCKP VNTFV HESLA ss s eeeee s hh	ribonuclease-s beta-strand	1RNS	44-48		QGGTH YGVSV VGIGR t ees ss ee s	thermolysin (protease)	2TLN	251-255
4.3	CPLMV KVLDA VRGSP eee eeeet tttee	prealbumin beta-strand	2PAB	6-10	2.0	AHTDF AGAEA AWGAT hhs g ggghh hhhhh	erythrocruorin (hemogl.)	1ECD	115-119
	NPKLQ KVLDA LQAIK gggh hhhht gggt	carbonic anhydrase B alpha-helix	1CAB	155-159		KFAGQ AGAEA ELAQR hhtt tthhh hhhhh	cytochrome C551	251C	38-42
4.1	DNGIR LAPVA tteee bb	acid protease beta-strand	1APR	320-324	1.7	AKKIV SDGDG MNAWV hhhhh hsssg gggsh	lysozyme (hen)	7LYZ	100-104
	CTTNC LAPVA KVLHE hhhhh hhhhh hhhhh	GP dehydrogenase alpha-helix	1GPD	153-157		LKAGD SDGDG KIGVD hhht tt ss eeehh	Ca binding parvalbumin	1CPV	91-95
4.1	LIGQK VAHAL AEGLG hhhhh hhhhh htt	triose phosp. isomerase alpha-helix	1TIM	112-116	1.6	LMVKV LDAVR GSPAI eeeee eettt tee	prealbumin	2PAB	8-12
	VSVNG VAHAL TAGHC eeett eeeee e hhh	strep.gris. proteinase A beta-strand	1SGA	25-29		PVYDS LDAVR RCALI hhhhh s hhh hhhhh	lysozyme (phage)	1LZM	91-95
4.1	LSGEE KAAVL ALWDK hhh hhhhh hhhtt	hemoglobin (horse) alpha-helix	2MHB	150-154	1.4	LLILP DEAAV GNLVG see ssttt gggtt	acid protease	1APR	229-233
	KVIKC KAAVL WEEKK s eee eeeeb stts	alcohol dehydrogenase beta-strand	4ADH	10-14		QQKRW DEAAV NLAKS htt tttss sstts	lysozyme (phage)	1LZM	127-131
3.9	GFFSK IIGEL PNIEA hhhhh hhhtt t hh	erythrocruorin (hemogl.) alpha-helix	1ECD	69-73	1.3	SLKPL SVSYD QATSL ts e eee t t	carbonic anhydrase C beta-strand	1CAC	45-49
	LSKTF IIGEL HPDDR hgggg eeeee gggg	cytochrome B5 beta-strand	2B5C	73-77		NADAT SVSYD VDLND ts e eeeee tt	concanavalin A beta-strand	3CNA	74-78
3.3	KITVL GVRQV GMACG eeee tthh hhhhh	lactate dehydrogenase alpha-helix start	1LDX	27-31	1.2	TLFPP SSEEL QANKA e tttt ttt	immunoglobulin helix	1FAB	117-121
	AAKTD GVRQV QPYNQ sb s eeee ss h	papain (protease) beta-strand	8PAP	109-113		AIHPT SSEEL VTLR sss sgggg ss	glutathione reductase 3-10-helix	2GRS	453-457
3.3	GNEGS TGSSS TVGYP s stts bt	subtilisin BPN'	1SBT	159-163	1.1	ELPGR SVIVG AGYIA s ss eeee shhh	glutathione reductase beta-strand	2GRS	173-177
	VTISC TGSSS NIGAG eeee e tt tttss	immunoglobulin	1FAB	23-27		EAYGV SVIVG VPPDS ttt e eee tt	alcohol dehydrogenase beta-strand	4ADH	289-293
3.1	FIPLS GGIDV VAHEL b gg g hhh hhhhh	thermolysin	2TLN	135-139	0.6	GNWVC AAKFE SNFNT hhhhh hhhhh tssbt	lysozyme (hen) alpha-helix	7LYZ	31-35
	DTVQV GGIDV TGGPQ b sss b s tt	acid protease	1APR	95-99		KETA AAKFE RQHMD h hhhhh hhhb	ribonulease-s alpha-helix	1RNS	5-9
3.1	KAGIQ LSKTF VKVVS stt e eette eeee	GP dehydrogenase	1GPD	299-303	0.5	LTESQ AALVK SSWEE t hhh hhhhh hhhhh	leghemoglobin alpha-helix	1HBL	8-12
	TDARE LSKTF IIGEL hhhhh hgggg eeeee	cytochrome B5	2B5C	68-72		ZKAND AALVK MRAAA tttth hhhhh hhhhh	cytochrome B562 alpha-helix	156B	28-32
3.0	LVKKM TDDKG AKTKM ttttt s s s	cytochrome C550	155C	90-94	0.3	VFSTE LPASQ QSGHS e bss s hhh htts	acid proteinase alpha-helix start	1APP	46-50
	KLGIQ TDDKG HIIVD ttt b tts b	glutathione reductase	2GRS	290-294		RYGFL LPASQ IIPTA sstt ggg hhhhh	carboxypeptidase 3-10-helix	1CPA	280-284
2.9	GRPIY VLKFS TGGSN s eee eeee s ss	carboxypeptidase	1CPA	48-52					
	NAGVE VLKFS QVKEV tss eette eeeee	glutathione reductase	2GRS	228-232					
2.8	DFPIA KGERQ SPVDI s ggg g ss ss ee	carbonic anhydrase C	1CAC	21-25					
	AGIKK KGERQ DLVAY hhhhh hhhhh	cytochrome C	1CYT	88-92					
2.6	LDNLK GTFAA LSELH ggghh hhshh hhhhh	hemoglobin (horse)	2MHB	225-229					
	GGAVV GTFAA RVFPG tteee eeeee ee s	alpha lytic protease	1ALP	52-56					

Each pair of pentapeptide conformations is preceded by the structural dissimilarity d, the minimum possible rms distance between equivalent C(α) positions. Pairs with the strongest structural discrepancy are listed first; those with the best structural agreement are listed last. Each occurrence of the pentapeptide is given with its (first line) local sequence context (± five residues), protein name, Protein Data Bank identifier, residue number range, and (second line) conformation/secondary structure (h, α-helix; g, 3- to 10-helix; e, β-strand; b, β-bridge; t, three-, four-, or five-turn; s, bend; space, straight piece not in β-structure). Residue numbers and structure notation are from the Dictionary of Protein Secondary Structure (2) and may differ from those given elsewhere. The amino acid one-letter code is A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. Structural dissimilarity per atom,

$$d = \left[\frac{1}{n-2}\sum_{i=1}^{n}(\mathbf{r}_i - \mathbf{r}_i')^2\right]^{1/2},$$

is normalized by $n - 2$, the number of nontrivial degrees of freedom. This is because the superposition of sets of two C(α) vectors trivially has $d = 0$ due to the peptide bond geometry.

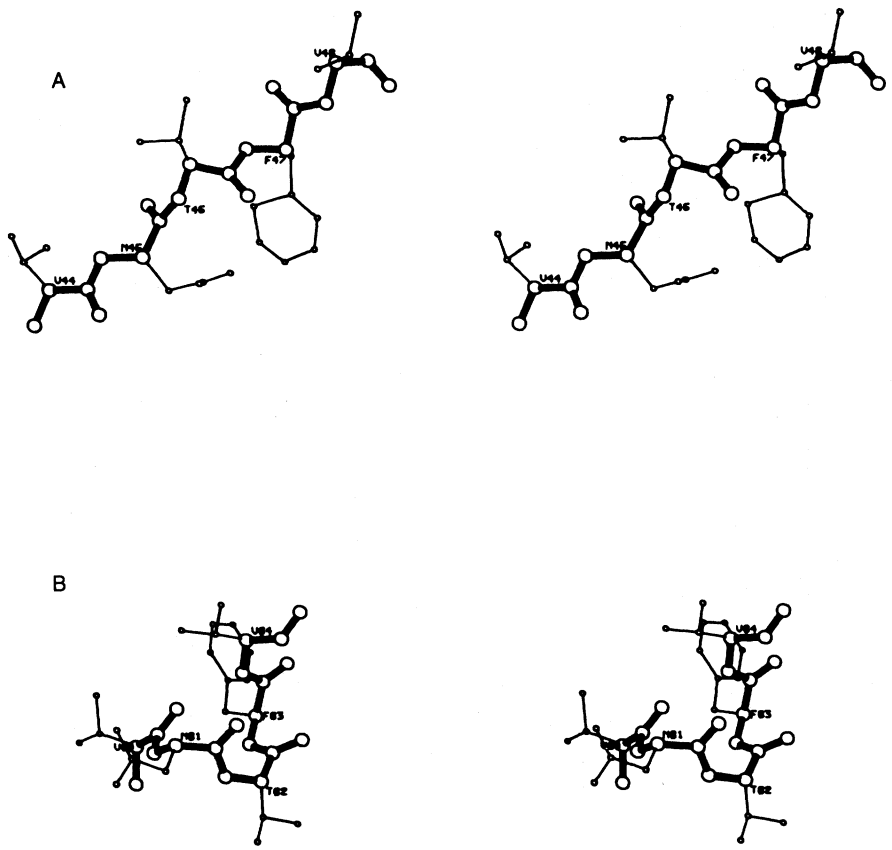

FIG. 1. Stereo view of variability of pentapeptide structure: the sequence Val-Asn-Thr-Phe-Val is part of a β-strand in ribonuclease (5) (*A*) and part of an α-helix in the hemoglobin erythrocruorin (6) (*B*). The figure was drawn by using the program PLUTO (W. D. S. Motherwell, Crystallographic Data Centre, University Chemical Laboratory, Cambridge, England, personal communication).

than five residues (2), the observed differences in structure are surprising.

The most striking examples of "same sequence-different structure" are the seven pentapeptides Val-Asn-Thr-Phe-Val (Fig. 1), Lys-Val-Leu-Asp-Ala, Leu-Ala-Pro-Val-Ala, Val-Ala-His-Ala-Leu, Lys-Ala-Ala-Val-Leu, Ile-Ile-Gly-Glu-Leu, and Gly-Val-Arg-Gln-Val: each occurs once as part of an α-helix and once as part of a β-strand, two very different types of hydrogen-bonded secondary structure. This is possible only if their structure is determined by sequence context—that is, by their interaction with other parts of the protein.

At the other extreme, "same sequence-same structure," each of the six protein subsequences Leu-Pro-Ala-Ser-Gln, Ala-Ala-Leu-Val-Lys, Ala-Ala-Lys-Phe-Glu, Ser-Val-Ile-Val-Gly, Ser-Ser-Glu-Glu-Leu, and Ser-Val-Ser-Tyr-Asp have very similar conformations in one protein and in another unrelated protein. As the sequence context for each occurrence is different, either the local conformational preference of these pentapeptides dominates over the interaction with neighboring segments or the influence of the neighboring segments is similar in each protein.

Conclusions

Interpretation of Amino Acid Sequence Comparisons. In attempts to predict protein function, it has become routine to compare the sequence of a new protein with the data base of proteins whose structure or function is known (7). Significant homology is interpreted to imply similarity of structure or function. For example, the (very) probable function of the *sis* oncogene protein was dramatically elucidated by the discovery of a 100-residue homology with platelet-derived growth factor (8, 9). Often, however, the discovered sequence homologies are marginal and the predictions, speculative. For example, weak homologies have been reported between the p21 protein of the *ras* oncogene family and, separately, an ATPase (10), some dinucleotide-binding enzymes (11), and the guanine nucleotide binding site of elongation factor Tu (R. Leberman and U. Schneider, personal communication). From these, partly conflicting (11) predictions of *ras* oncogene structure were derived. Our survey puts a definite lower bound on the length of an uninterrupted stretch of identical residues deemed statistically significant or structurally meaningful.

Statistical Significance of Identical Pentapeptides. The survey shows that an isolated uninterrupted homology of length five between a new protein and the data base is in itself not a safe indication of evolutionary kinship. The statement depends, of course, on the total length of the sequences compared. For example, suppose the new protein sequence has length $m = 300$ and you find one pentapeptide identity between it and a protein in the data base of $n = 10{,}000$ residues. Substituting in Eq. **1** $N = n \times m$, the number of comparisons, we get $P = 0.39$ for the probability not to find this identity by chance. Appropriate values for n and m must be used when comparing a given length of protein with a data base.

The danger of overinterpretation of temptingly strong but too short sequence homologies is, of course, a particular case of the general danger of overzealous homology searches. A good example of temptingly long but too weak homologies is the 16-residue homology dispersed among 129 residues between ribonuclease and lysozyme postulated by Manwell in 1967 (12) and shown by Haber and Koshland (13) as well as by others to be statistically insignificant. Inciden-

tally, the identical pentapeptide Ala-Ala-Lys-Phe-Glu (Table 1) was not part of Manwell's alignment.

Structural Meaning of Identical Pentapeptides. Table 1 shows that it is wrong to deduce similarity of conformation from a single pentapeptide homology between otherwise not homologous sequences. This result is independent of the size of the data base or any statistical estimate. It is, however, likely that particular pentapeptide sequences vary in structural adaptability.

Message to Protein Folders. Why is the accuracy of current widely used secondary structure prediction methods not better than 56% [correctly predicted residues in a three-state model (14)]? The structural discrepancies in Table 1 show that even complete statistics on the conformational preferences of pentapeptides (which would require at least 300 times the current data base) in themselves are not sufficient to correctly predict protein structure. The key to improved protein structure prediction is not better statistics but new approaches. Some are appearing (15, 16). We can be sure that the strictly hierarchical approach of first predicting secondary structure segments, then tertiary structure by interaction of those segments, cannot succeed unless cooperativity of sequence range larger than five is taken into account in the first step.

We thank the referees for suggestions and the Deutsche Forschungsgemeinschaft for support to the Protein Structure Theory project.

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
2. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
3. Spiegel, M. R. (1975) *Schaum's Outline of Theory and Problems of Probability and Statistics* (McGraw-Hill, New York), pp. 30–31.
4. Kabsch, W. (1978) *Acta Crystallogr. Sect. A* **34,** 827–828.
5. Fletterick, R. J. & Wyckoff, H. W. (1975) *Acta. Crystallogr. Sect. A* **31,** 698–700.
6. Steigemann, W. & Weber, E. (1979) *J. Mol. Biol.* **127,** 309–338.
7. Doolittle, R. F. (1981) *Science* **214,** 149–159.
8. Waterfield, M. D., Scrace, G. T., Whittle, N., Stroobant, P., Johnsson, A., Wasteson, A., Westermark, B., Heldin, C. H., Huang, J. S. & Deuel, T. F. (1983) *Nature (London)* **304,** 35–39.
9. Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. & Antoniades, H. N. (1983) *Science* **221,** 275–277.
10. Gay, N. J. & Walker, J. E. (1983) *Nature (London)* **301,** 262–264.
11. Wierenga, R. & Hol, W. G. J. (1983) *Nature (London)* **302,** 842–844.
12. Manwell, C. (1967) *J. Comp. Biochem. Physiol.* **23,** 383–406.
13. Haber, J. E. & Koshland, D. E. (1970) *J. Mol. Biol.* **50,** 617–639.
14. Kabsch, W. & Sander, C. (1983) *FEBS Lett.* **155,** 179–182.
15. Ptitsyn, O. B. & Finkelstein, A. V. (1983) *Biopolymers* **22,** 15–25.
16. Taylor, W. R. & Thornton, J. M. (1983) *Nature (London)* **301,** 540–542.