PROTEINS: Structure, Function, and Genetics 17:138–151 (1993)

Structural Relationships of Homologous Proteins as a Fundamental Principle in Homology Modeling

Martina Hilbert, Gerald Böhm, and Rainer Jaenicke
Institute for Biophysics and Physical Biochemistry, University of Regensburg, 93040 Regensburg, Germany

ABSTRACT **Protein structure prediction is based mainly on the modeling of proteins by homology to known structures; this knowledge-based approach is the most promising method to date. Although it is used in the whole area of protein research, no general rules concerning the quality and applicability of concepts and procedures used in homology modeling have been put forward yet. Therefore, the main goal of the present work is to provide tools for the assessment of accuracy of modeling at a given level of sequence homology. A large set of known structures from different conformational and functional classes, but various degrees of homology was selected. Pairwise structure superpositions were performed. Starting with the definition of the structurally conserved regions and determination of topologically correct sequence alignments, we correlated geometrical properties with sequence homology (defined by the 250 PAM Dayhoff Matrix) and identity. It is shown that both the topological differences of the protein backbones and the relative positions of corresponding side chains diverge with decreasing sequence identity. Below 50% identity, the deviation in regions that are structurally not conserved continually increases, thus implying that with decreasing sequence identity modeling has to take into account more and more structurally diverging loop regions that are difficult to predict.**

Key words: structure prediction, computer modeling, structure comparison, sequence identity, protein homology

INTRODUCTION

The correlation between the sequence of a protein and its three-dimensional structure has been a central issue in structural biology for more than 30 years.[1–3] Irrespective of the evidence for the one-to-one relationship, presently, there is still no algorithm at hand that would correlate the sequence information with the corresponding native three-dimensional structure; it is even doubtful whether a *unique folding code* can be developed within the framework of the available database.[4] Proof for the existence of an abstract folding code in nature is that obviously all sequences that occur in vivo fold into highly specific, unique spatial structures in a given environment mimicking physiological conditions. This argument holds in spite of the fact that a given protein may exist in highly populated substates of one conformation, and that accessory proteins seem to be essential for the in vivo folding of proteins.[5,6] Structure prediction methods currently used with general applicability and reasonable success may be summarized under the equivalent terms *knowledge-based* modeling, *homology* modeling, or *comparative* modeling.[7–9] The basis for these concepts is the availability of at least one well-defined structure of a protein family analyzed either by high-resolution X-ray crystallography or nuclear magnetic resonance (NMR).

A known structure is used as a template or parent structure for the unknown protein to be modeled. Identical amino acid positions are depicted by a pairwise or multiple sequence alignment and are kept as a *framework*. Several methods and principles have been put forward for the modeling of side chain conformations of amino acid exchanges in the corresponding sequence and structure positions.[10,11] Insertions and deletions in homologous sequences generally occur in positions of loops and turns, since otherwise the regular motif of the secondary structural elements would be disrupted; again, several methods are in use of this modeling step. These are commonly database searches for appropriate loops and turns.

Some examples for homology modeling of protein structure have been reported in the literature.[12–15] However, what is missing are rules for the general applicability of methods known so far, their advantages and disadvantages, limits and pitfalls, depen-

Received September 11, 1992; revision accepted May 28, 1993.
Address reprint requests to Dr. Martina Hilbert, Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Universitätsstraße 31, 93040 Regensburg, Germany.

dent on the degree of sequence homology or identity. The aim of the present work is to investigate aspects of these fundamental questions concerning homology modeling.

As most of the errors in a model originate from the differences between the real three-dimensional structures of unknown and homologous parent proteins, the superposition of homologous structures, and studies on the relationship between sequence homology, on one hand, and structural differences, on the other, can provide useful information regarding the accuracy of modeling. Therefore, in the present study pairwise superpositions of a large set of known structures from different conformational and functional classes, but various degrees of homology, are performed. A structural and sequence-based analysis of the superimposed structures reveals rules and correlations which may be useful for the modeling of protein structures by homology.

MATERIALS AND METHODS

Protein Structures

The Brookhaven Protein Data Bank, release 53 (July 1990), updated according to release 62 (October 1992), was used to select structures for the comparative analysis.[16] The following rules are applied to the chosen structures: (1) at least *two distinct* proteins from the class have to be included into the sample, except for the tryptophan repressor structures 1WRP and 2WRP, which present two different crystal forms, trigonal and orthorhombic, respectively (see Table I); (2) functionally related proteins are grouped within one class; (3) only one subunit (monomer) is used in the case of *homomeric* proteins; (4) in the case of *heteromeric* proteins, the subunits are treated as different proteins; (5) the structures have to be reasonably *complete* in terms of missing atoms or residues; (6) when several equivalent structures are available, the entry with the highest resolution is chosen; generally the resolution has to be better than 3.0 Å; (7) immunoglobins are excluded from the current approach; they will be dealt with in a separate investigation. Table I lists the atomic coordinate entries selected for this work, which represent a sufficient database covering the whole range of protein homology.

Common Core and Structurally Correct Sequence Alignments

In order to define the structurally conserved regions of the protein pairs, and to determine the respective structurally correct alignments, the protein structures have to be superimposed. The program applied for this purpose is MNYFIT Version 5.0 from the program package COMPOSER which was kindly provided by Prof. T.L. Blundell.[17] It performs a rigid body superposition of the α-carbon atoms of protein structures.[18] The initial superposition is performed using at least three predefined atoms from each molecule which occupy topologically equivalent positions. In order to determine such pairs of α-carbon atoms without having the structures superimposed in advance, sequence alignments are generated using the program SEQCOMP by D. Bacon; it implements the dynamic programming method proposed by Needleman and Wunsch,[19] and was used in the extended version suggested by Lesk et al.[20] which applies variable gap penalties to explicitly shift insertions and deletions into regions of loops and turns. Gap penalties are −15 for regions of α-helix or β-sheet sections, −9 for the border between these and loop/turn regions, and −8 for loop/turn ranges. The underlying point exchange matrix is that proposed by M.O. Dayhoff for an evolutionary distance of 250 expected point mutations related to a sequence length of 100 amino acids (250-PAM table) with a matrix bias of 2.[21]

For the initial superposition, α-carbon atoms of pairs of identical residues in major elements of secondary structure from each molecule, according to the generated alignments, are used; they are supposed to occupy topologically equivalent positions. The structurally conserved α-carbon atoms in the superimposed structures are defined as those which are separated by a distance less than a preset cutoff value of 3.8 Å. We decided for this value because—in contrast to previously published studies[22]—it has a structural meaning: it is the mean distance of adjacent α-carbon atoms in an extended polypeptide chain, i.e., the shift of an α-carbon atom of 3.8 Å completely reassigns the spatial position. Thus, for a pair of proteins a common core is constituted which comprises structurally conserved regions not including diverging loop and turn regions. As a control, all the superpositions and structurally correct alignments generated by MNYFIT are inspected by computer graphics for their accuracy. Where needed, manual corrections are performed. Only those protein pairs are taken into account whose superposition promises reasonable modeling of the structure. For graphic display, the program INSIGHTII Version 2.0 (Biosym Technologies Inc., San Diego, CA) is used. All programs applied are installed on a Silicon Graphics IRIS 4D/70 GTB workstation running IRIX 3.2.

Analysis of Superpositions and Structurally Correct Alignments

The analysis of the superpositions and structurally correct alignments requires a number of terms to be defined: *Identity* measures the relative amount of identical amino acids in corresponding positions in the structurally correct alignment, with respect to the sum of residue pairs in the structurally correct alignment. *Homology* is defined as the sum of Dayhoff scores (250 PAM-table, matrix bias 2) from the structurally correct alignment, with respect to the *larger* value of the intrinsic sum of Dayhoff

TABLE I. List of Entries From the Brookhaven Protein Data Bank Used in This Work

PDB code	Functional group / Protein	Organism, tissue	Resolution (Å)	*R*-factor	Number of residues
Calcium binding protein					
1ALC*	α-Lactalbumin	*Papio cynocephalus*, milk	1.7	0.22	123
3CLN	Calmodulin	*Rattus rattus*, testis	2.2	0.175	148
Troponin C (contractile system protein)					
4TNC	Troponin C	*Gallus gallus*, skeletal muscle	2.0	0.172	162
5TNC	Troponin C	*Meleagris gallopavo*, skeletal muscle	2.0	0.155	162
Tryptophan repressor (DNA binding protein)					
1WRP	Tryptophan repressor (trigonal form)	*Escherichia coli*	2.2	0.204	107
2WRP	Tryptophan repressor (orthorhombic form)	*Escherichia coli*	1.65	0.180	107
Cytochrome (electron transfer)					
1CCR	Ferricytochrome *c*	*Oryza sativa*, rice embryos	1.5	0.19	111
1CYC	Ferrocytochrome *c*	*Katsuwonus pelamis*	2.3		103
2CDV	Cytochrome c_3 (reduced)	*Desulfovibrio vulgaris miyazaki* IAM 12604	1.8	0.176	107
3C2C	Cytochrome c_2 (reduced)	*Rhodospirillum rubrum*	1.68	0.175	112
451C	Cytochrome c_{551} (reduced)	*Pseudomonas aeruginosa*	1.6	0.187	82
5CYT	Cytochrome *c* (reduced)	*Thunnus alalunga*, heart	1.5	0.159	103
Ferredoxin (electron transfer)					
1FDX	Ferredoxin	*Peptococcus aerogenes*	2.0		54
2FXB	Ferredoxin	*Bacillus thermoproteolyticus*	2.3	0.204	81
3FXC	Ferredoxin	*Spirulina platensis*	2.5	0.31	98
4FD1	Ferredoxin	*Azotobacter vinelandii* op, ATCC 13705	1.9	0.212	106
Rubredoxin (electron transfer)					
1RDG	Rubredoxin	*Desulfovibrio gigas*	1.4	0.136	52
7RXN	Rubredoxin	*Desulfovibrio vulgaris*	1.5	0.098	52
4RXN	Rubredoxin	*Clostridium pasteurianum*	1.20	0.128	54
Insulin (hormone)					
2INS(A,B)	Insulin	*Bos taurus*	2.5	0.18	21/29
4INS(A,B)	Insulin	*Sus scrofa*	1.5	0.153	21/30
Acid proteinase (hydrolase)					
3APP	Penicillopepsin (EC 3.4.23.7)	*Penicillium janthinellum*	1.8	0.126	323
2APR	Rhizopuspepsin (EC 3.4.23.6)	*Rhizopus chinensis*	1.8	0.143	325
4APE	Endothiapepsin (EC 3.4.23.10)	*Endothia parasitica*	2.1	0.158	330
Phospholipase A_2 (hydrolase)					
1BP2	Phospholipase A_2 (EC 3.1.1.4)	*Bos taurus*, pancreas	1.7	0.171	123
1P2P	Phospholipase A_2 (EC 3.1.1.4)	*Sus scrofa*, pancreas	2.6	0.241	124
1PP2(R)	Phospholipase A_2 (EC 3.1.1.4)	*Crotalus atrox*	2.5	0.178	122

Serine proteinase (hydrolase)					
1SGT	Trypsin (EC 3.4.21.4)	*Streptomyces griseus*, strain K1	1.7	0.161	223
1TPO	β-Trypsin (EC 3.4.21.4)	*Bos taurus*, pancreas	1.7	0.180	223
2GCH	γ-Chymotrypsin A (EC 3.4.21.1)	*Bos taurus*, pancreas	1.9	0.181	241
2PRK	Proteinase K (EC 3.4.21.14)	*Tritirachium album limber*	1.5	0.167	279
2SGA	Proteinase A (EC number not assigned)	*Streptomyces griseus*, strain K1	1.5	0.126	181
3EST	Native elastase (EC 3.4.21.11)	*Sus scrofa*, pancreas	1.65	0.169	240
4CHA(A)	α-Chymotrypsin (EC 3.4.21.1)	*Bos taurus*	1.68	0.234	241
Sulfhydryl proteinase (hydrolase)					
2ACT	Actinidin (EC number not assigned)	*Actinidia chinensis*	1.7	0.171	220
9PAP	Papain (EC 3.4.22.2)	*Carica papaya*, fruit latex	1.68	0.161	212
Oxidoreductase [aldehyde(d)-NAD(a)]					
1GD1(O)	D-Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	*Bacillus stearothermophilus*	1.8	0.177	334
1GPD(G)	D-Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	*Homarus americanus*	2.9		333
Oxidoreductase [CH-NH(d)-NAD(+) or NADP(+)(a)]					
3DFR	Dihydrofolate reductase (EC 1.5.1.3)	*Lactobacillus casei*	1.7	0.152	162
4DFR(B)	Dihydrofolate reductase (EC 1.5.1.3)	*Escherichia coli* MB1428	1.7	0.155	159
8DFR	Dihydrofolate reductase (EC 1.5.1.3)	*Gallus gallus*, liver	1.7	0.188	189
Oxidoreductase [CHOH(d)-NAD(a)]					
1LDM	Lactate dehydrogenase M_4 (EC 1.1.1.27)	*Squalus acanthias*, muscle	2.1	0.173	329
4MDH(A)	Cytoplasmic malate dehydrogenase (EC 1.1.1.37)	*Sus scrofa*, heart	2.5	0.167	333
5LDH	Lactate dehydrogenase H_4 (EC 1.1.1.27)	*Sus scrofa*, heart	2.7	0.196	333
8ADH	Apo-alcohol dehydrogenase (EC 1.1.99.8)	*Equus caballus*, liver	2.4	0.19	374
Myglobin (oxygen storage)					
1MBA	Myoglobin (met)	*Aplysia limacina*	1.6	0.193	146
1MBD	Myoglobin (deoxy)	*Physeter catodon*, sperm	1.4		153
1MBS	Myoglobin (met)	*Phoca vitulina*	2.5		153
1PMB(A)	Myoglobin (aquomet)	*Sus scrofa*, recombinant from *E. coli*	2.5	0.185	153
Oxygen transporting proteins					
1ECD	Erythrocruorin (reduced, deoxy)	*Chironomus thummi thummi*	1.4		136
2HMQ(B)	Hemerythrin (met)	*Themiste dyscritum*	1.66	0.189	113
1LH4	Leghemoglobin (deoxy)	*Lupinus luteus*, root nodules	2.0		153
2DHB(B)	Hemoglobin (deoxy)	*Equus caballus*	2.8		146
2HHB(B)	Hemoglobin (deoxy)	*Homo sapiens*	1.74	0.16	146
Periplasmic binding protein					
1ABP	L-Arabinose-binding protein	*Escherichia coli*	2.4		306
2LBP	Leucine-binding protein (LBP)	*Escherichia coli* K12	2.4	0.213	346
2LIV	Leucine/isoleucine/valine-binding protein (LIVBP)	*Escherichia coli*	2.4	0.179	344
Phosphotransferase (transferase)					
2PFK(A)	Phosphofructokinase (EC 2.7.1.11)	*Escherichia coli*	2.4	0.166	320
3PFK	Phosphofructokinase (EC 2.7.1.11)	*Bacillus stearothermophilus*	2.4	0.171	319

*Column 1 contains the entry code from the Brookhaven Protein Data Bank supplemented by the specific identifier for the selected subunit where needed.

scores.* *Homology, gap penalty 8* is defined equivalently to Homology, core, except that for each insertion a gap penalty of 8 is subtracted from the sum of Dayhoff scores. *Residues in core* measures the relative amount of residue pairs in the structurally correct alignment with respect to the *longer* protein sequence.

The structural deviation between two superimposed proteins is commonly defined as the root mean square (RMS) difference (in Å) of intermolecular atom distances.[23] *RMS difference* (*Cα*), *core* measures the deviation of α-carbon atoms in the structurally conserved regions of two proteins, i.e., with distances of corresponding α-carbon atoms not greater than 3.8 Å. *RMS difference* (*Cα*), *total* additionally takes into account α-carbon atoms in structurally divergent regions. For this purpose, the structurally correct alignment is extended into these regions using again the criterion of minimum α-carbon atom distance. Thus, a global sequence alignment is generated. The number of residue pairs which constitute the core regions on one hand, and the global alignment on the other, are indicated in Table II.

Apart from considering α-carbon atoms alone, the RMS difference is also calculated from the distances between corresponding *main chain atoms* and α- and β-*carbon atoms*. Furthermore, the distances between *normalized* β-*carbon atoms* are calculated by individually superimposing each pair of structurally equivalent α-carbon atoms, accompanied by a corresponding spatial shift of their side chains. The newly arranged β-carbon atoms are called normalized β-carbon atoms; the RMS difference of their pairwise distances gives insight into the relative topological positions of corresponding side chains. Finally, the number of insertions that are necessary for a structural alignment is calculated.

The results of the comparison of the 51 pairs of homologous proteins whose structure superpositions promise reasonable modeling are summarized in Table II.

RESULTS AND DISCUSSION

Relative Size of Common Core

The structurally conserved regions in which the general fold of the polypeptides is very similar are defined by the preset cutoff value of 3.8 Å. The fraction of residues in the common core region drops with decreasing sequence identity, as indicated in Figure 1. Pairs whose identity exceeds 50% have 90% or more of their individual residues in structurally conserved regions. If sequence identity is below 20%, the common cores still include 65% or more of the amino acids of the protein structures.

Structural Divergence of the Common Core

As mentioned, only those protein pairs were taken into account whose superimposed structures allow reasonable modeling, i.e., whose tertiary structures retain a common topology regarding major elements of secondary structures. Thus, although the pairwise superimposed proteins have similar biological functions, this criterion does not necessarily imply structural homology. For example, the ferredoxin family consisting of four members (see Table I) was excluded from further investigation due to significant structural differences. Hereby, both the small size of the molecules (sequences ranging between 54 and 106 residues) and a low level of secondary structure prohibit a well-defined structural superposition.

Differences in length between two protein sequences exceeding 10% (with reference to the shorter sequence) usually cause a degree of structural divergence, which does not allow reasonable modeling (data not shown). However, there are exceptions to this rule, as shown by the three pairs marked with double daggers in Table II.

Figure 2 shows the relationship between sequence identity in the common core region and the RMS difference of the corresponding α-carbon atoms. Starting from about 0.25 Å, the structural differences increase exponentially to a value of 2.12 Å as identity drops from 100 to 14%.

A similar result with a smaller and more selective database has been reported previously by Chothia and Lesk.[22] However, a few protein pairs deviate significantly from the smooth curve (see Fig. 2). In the case of the cytochrome and myoglobin pairs [1CYC:5CYT (number 2), 1CCR:1CYC (number 6), 1CYC:3C2C (number 9), 1MBS:1PMB (number 3), 1MBD:1MBS (number 4)] the deviations can most probably be attributed to an insufficient refinement of the crystallographic data of the protein structures 1CYC and 1MBS, respectively.

The quality of the glyceraldehyde-3-phosphate dehydrogenase database entry 1GPD (lobster) and, hence, the superposition 1GD1:1GDP (Fig. 2, number 7) may suffer from the low resolution of the 1GPD structure: with 2.9 Å it is the lowest in the database chosen.

At this point, one may argue that a stricter selection rule might be preferable in terms of higher resolution. However, in this case some of the protein classes (e.g., glyceraldehyde-3-phosphate dehydrogenases or lactate/malate dehydrogenases) would lack a sufficient number of structures. Furthermore, in a modeling process one may depend on protein structures whose crystallographic data suffer from low resolution or insufficient refinement. Therefore, it is useful to understand the problems resulting

*The intrinsic sum of Dayhoff scores is calculated by self-alignments of the sequence parts constituting the common core region of each protein structure.

TABLE II. Results of the Pairwise Superpositions*

Protein A	Protein B	Number of residues A	Number of residues B	Difference of length (%)	Residues in core (%)	Identity in core (%)	Homology in core (%)	Homology (gp 8)† in core (%)	Insertions	Core Atom pairs	Core RMS (Å) Cα	Core RMS (Å) Cβ†	Total Atom pairs	Total RMS (Å) Cα	Total RMS (Å) Cβ†
1WRP	2WRP	107	107	0.0	93.3	100.0	100.0	100.0	0	97	0.92	0.45	101	2.81	0.65
2INS(B)	4INS(B)	29	30	3.4	96.7	100.0	100.0	100.0	0	29	0.38	0.42	29	0.38	0.42
2GCH	4CHA	241	241	0.0	97.5	100.0	100.0	100.0	0	233	0.44	0.31	236	0.51	0.40
4TNC	5TNC	162	162	0.0	98.8	98.1	98.6	98.6	0	159	0.24	0.41	160	0.45	0.42
1CYC	5CYT	103	103	0.0	98.1	98.0	99.6	99.6	0	101	1.37	0.79	103	1.61	0.83
2INS(A)	4INS(A)	21	21	0.0	100.0	90.5	98.2	98.2	0	21	0.32	0.30	21	0.32	0.30
1MBS	1PMB	153	153	0.0	95.4	89.0	95.3	95.3	0	146	1.47	0.78	153	1.84	0.84
1MBD	1PMB	153	153	0.0	100.0	86.3	94.9	94.9	0	153	0.49	0.29	153	0.49	0.29
1BP2	1P2P	123	124	0.8	93.5	84.5	93.5	92.6	1	116	0.59	0.44	123	1.51	0.63
1MBD	1PMS	153	153	0.0	96.1	84.4	94.6	94.6	0	147	1.50	0.76	153	1.90	0.81
2DHB(B)	2HHB(B)	146	146	0.0	99.3	82.1	90.5	90.5	0	145	0.71	0.44	146	0.83	0.47
2LBP	2LIV	346	344	0.6	98.3	79.7	88.8	88.1	2	340	0.75	0.39	344	0.77	0.41
1RDG	7RXN	52	52	0.0	100.0	71.2	87.2	87.2	0	52	0.63	0.54	52	0.63	0.54
1LDM	5LDH	329	333	1.2	89.5	71.1	83.8	82.6	3	298	1.63	0.90	329	2.13	1.02
4RXN	7RXN	52	54	3.8	94.4	68.6	85.0	85.0	0	51	0.51	0.45	52	0.51	0.48
1RDG	4RXN	52	54	3.8	96.3	63.5	81.0	81.0	0	52	0.53	0.37	52	0.53	0.37
1CCR	1CYC	111	103	7.8	89.2	61.6	79.3	79.3	0	99	1.34	0.76	103	1.64	0.84
1CCR	5CYT	111	103	7.8	92.8	59.2	79.0	79.0	0	103	0.57	0.35	103	0.57	0.35
2PFK	3PFK	320	319	0.0	94.0	57.3	73.3	73.3	0	300	0.87	0.32	300	0.87	0.32
1GD1	1GPD	334	333	0.3	92.2	55.8	71.2	68.8	6	308	1.31	0.80	330	1.58	0.90
3APP	4APE	323	330	2.2	92.4	55.7	77.7	75.3	6	305	1.23	0.47	321	1.53	0.55
2ACT	9PAP	220	212	3.8	92.7	49.5	65.9	63.7	4	202	0.67	0.48	211	0.97	0.56
1TPO	2GCH	223	241	8.1	86.4	47.1	62.9	59.6	6	204	0.93	0.62	217	1.26	0.71
1TPO	4CHA	223	241	8.1	87.0	47.1	62.5	59.2	6	208	0.94	0.61	219	1.37	0.75
1P2P	1PP2	124	122	1.6	82.3	46.1	60.3	55.5	5	102	1.06	0.60	113	1.60	0.77
1BP2	1PP2	123	122	0.8	85.4	45.7	59.5	55.7	4	105	1.09	0.53	113	1.50	0.69
2APR	3APP	323	325	0.6	87.7	42.5	60.6	54.9	14	285	1.24	0.51	313	1.94	0.65
2APR	4APE	325	330	1.5	86.7	41.6	58.8	53.1	14	286	1.27	0.56	316	1.80	0.71
3EST	4CHA	240	241	0.0	89.2	41.6	61.5	57.3	8	214	1.11	0.60	228	1.54	0.75
1CCR	3C2C	111	112	0.9	83.9	41.5	64.5	59.7	4	94	1.09	0.64	101	1.18	0.71
2GCH	3EST	241	240	0.0	89.6	40.9	60.9	57.2	7	215	1.09	0.58	226	1.37	0.73
1CYC	3C2C	103	112	8.7	77.7	39.1	65.0	59.8	4	87	1.71	0.86	100	2.07	1.01
1TPO	3EST	223	240	7.6	89.6	39.1	57.3	52.6	9	215	1.02	0.54	221	1.14	0.59
3C2C	5CYT	112	103	8.7	82.1	38.0	66.1	61.2	4	92	1.00	0.48	100	1.11	0.59
1SGT	1TPO	223	223	0.0	86.5	36.8	54.3	49.0	9	193	1.04	0.55	211	1.48	0.72
1SGT	4CHA	223	241	8.1	78.7	35.6	54.8	48.8	10	188	1.08	0.61	213	1.75	0.88
1SGT	2GCH	223	236	5.8	79.2	35.3	54.7	48.6	10	187	1.03	0.66	211	1.68	0.86
1SGT	3EST	223	240	7.6	83.8	34.3	48.1	41.9	11	201	1.16	0.64	216	1.82	0.78
4DFR	8DFR	159	189	18.9‡	76.3	32.4	50.6	44.5	8	142	1.42	0.87	156	1.87	1.01
3DFR	8DFR	162	189	16.7‡	76.3	31.7	50.5	44.3	8	142	1.27	0.79	160	1.84	1.02
3DFR	4DFR	162	159	1.9	88.9	27.8	49.1	46.1	4	144	1.12	0.61	157	1.78	0.82
1MBA	1MBD	146	153	4.8	85.6	26.7	45.3	41.7	4	131	1.72	0.52	143	1.91	0.58
1MBA	1PMB	146	153	4.8	88.2	24.4	45.0	41.5	4	135	1.82	0.57	143	1.95	0.59
1MBA	1MBS	146	153	4.8	83.7	23.4	42.7	39.1	4	128	2.05	0.76	143	2.51	0.93
4MDH	5LDH	333	333	0.0	64.6	21.9	43.8	36.7	13	215	2.05	0.95	313	3.66	1.30
1LDM	4MDH	329	333	1.2	76.0	19.8	38.7	31.4	16	253	1.69	0.77	307	2.55	1.06
1ECD	2HHB(B)	136	146	7.4	82.9	19.8	40.4	36.6	4	121	2.04	0.76	136	2.28	0.82
1LH4	2DHB(B)	153	146	4.8	69.9	17.8	39.6	36.4	3	107	2.12	0.71	144	2.99	1.09
1LH4	2HHB(B)	153	146	4.8	64.7	17.2	37.9	34.4	3	99	2.03	0.59	144	3.17	1.04
1ECD	1LH4	136	153	12.5‡	66.0	16.8	40.1	35.5	4	101	1.97	0.61	131	2.58	0.98
1ECD	2DHB(B)	136	146	7.4	83.6	13.9	34.9	31.1	4	122	2.11	0.76	136	2.32	0.97

*The results are listed according to the decreasing sequence identity in the core region.
†gp8 and Cβ stand for gap penalty 8 and normalized β-carbon atoms, respectively (see Materials and Methods).
‡The difference in length between the two proteins exceeds 10%.

from these common difficulties. Obviously, the given correlation function may serve as a quality control for X-ray structures.

In further examining outliers in Figure 2, the occurrence of negative temperature factors in the crystallographic data of the endothiapepsin structure 4APE may also indicate errors in the refinement procedure (cf. comment to the respective database entry). These may also contribute to the high value of the structural deviation of the superimposed aspartyl proteinases 4APE (endothiapepsin) and 3APP (penicillopepsin) (Fig. 2, number 8); however, functional differences of the proteins may also be involved.

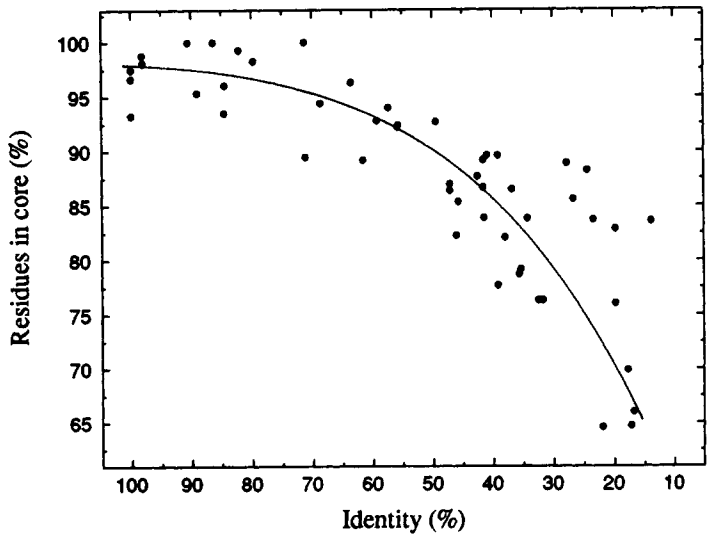

Fig. 1. Relative size of the common core (number of residue pairs in the core region related to the longer sequence length) as a function of sequence identity.

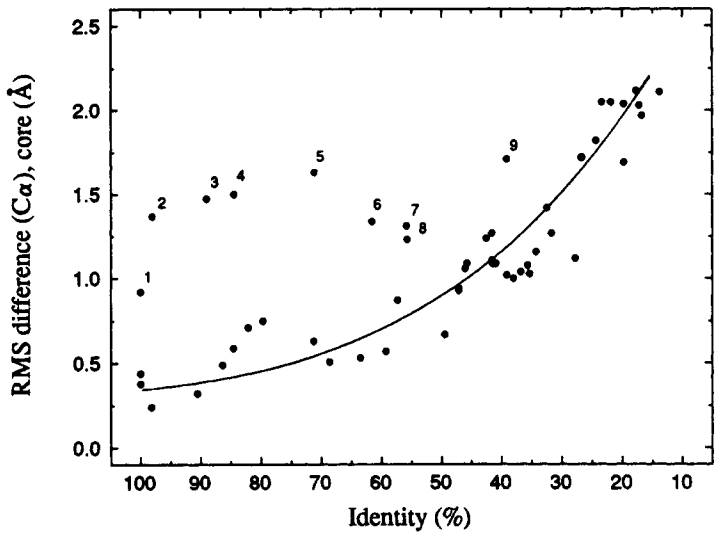

Fig. 2. Relationship between the RMS difference of the corresponding α-carbon atoms and sequence identity in the common core region. Proteins showing significant deviations are numbered: (1) 1WRP:2WRP, (2) 1CYC:5CYT, (3) 1MBS:1PMB, (4) 1MBD:1MBS, (5) 1LDM:5LDH, (6) 1CCR:1CYC, (7) 1GD1:1GPD, (8) 2APP:4APE, (9) 1CYC:3C2C; see Table II.

Functional arguments may serve as a possible explanation for the following examples where high sequence identity of protein pairs is in remarkable contrast to the calculated large structural differences: (1) The superposition of the structures of the liganded trigonal (1WRP) and orthorhombic (2WRP) crystal forms of tryptophan repressor from *Escherichia coli* (Fig. 2, number 1) reveals that there is a large structural difference (0.92 Å) in spite of the 100% sequence identity. Obviously, this is attributable to a locally well-defined structural region involving the DNA-binding domain (Fig. 3). Helices D and E connected by a short turn (helix–turn–helix motif) constitute one of the two "reading heads" of the dimeric regulatory protein.[24] The comparison of the superimposed crystal structures indicates a movement of the two helices in the reading head motif relative to each other, accompanied by conformational adjustments in the main chain of both the interdomain hinge turns (C–D and E–F) and the interhelical turn (D–E). This intradomain flexibility of the reading heads may be functionally required by the various binding modes proposed by tryptophan repressor in its search for and adherence to

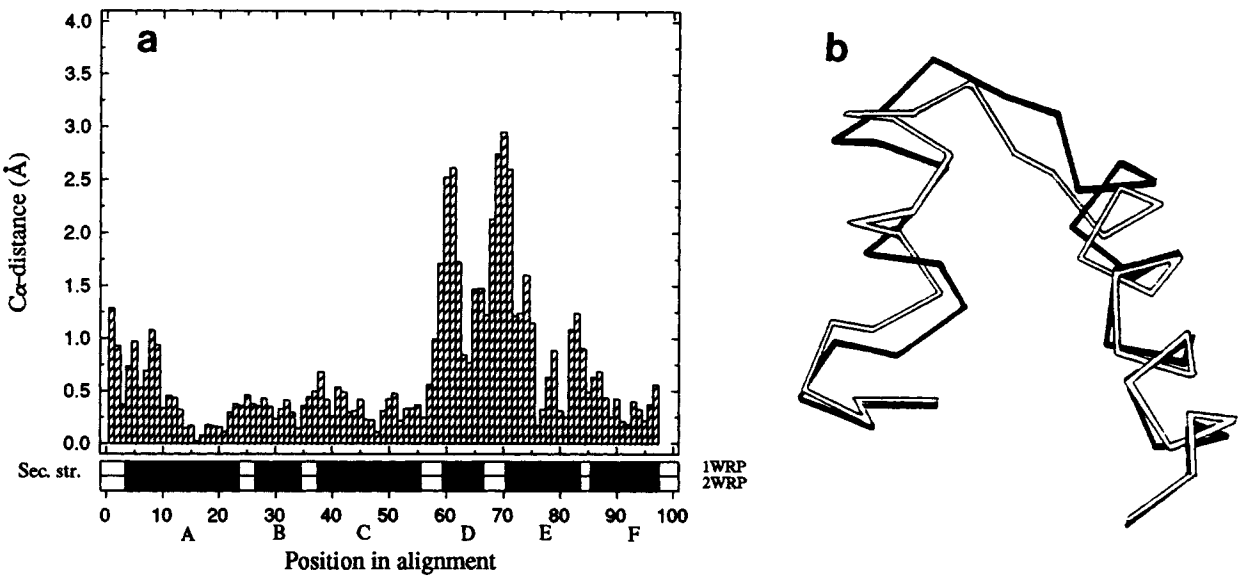

Fig. 3. The structures of the tryptophan repressor (1WRP: 2WRP). **(a)** Distances of α-carbon atoms at the corresponding positions in the structural correct alignment. The filled boxes indicate elements of secondary structure (Sec.str.), i.e., α-helices and β-sheets. The tryptophan repressor structure consists of six α-helices denoted A, B, C, D, E, F. Helices D and E, connected by a short turn (helix–turn–helix motif), constitute the DNA-binding domain. **(b)** Superposition of the DNA-binding domain of one subunit of the dimeric tryptophan repressor 1WRP:2WRP (2WRP: black line). Positions of α-carbon atoms are shown for the D-helix (on the right side), D–E-turn, and E-helix.

its three different operator sites.[24] (2) In the case of the superposition of the monomers of the lactate dehydrogenases from *Squalus acanthias* muscle (1LDM) and *Sus scrofa* heart (5LDH) a similar anomaly (structural deviation 1.63 Å vs. sequence identity 71.1%; cf. Fig. 2, number 5) may be explained by the regulation mechanisms of the two isoenzymes in their respective tissues.[25]

Figure 4 demonstrates the relationship between the RMS difference of corresponding α-carbon atoms, on the one hand, and sequence *homology* instead of *identity*, on the other, where homology is based on the 250 PAM table with a matrix bias of 2 (Fig. 4a) and, additionally, a gap penalty of 8 (Fig. 4b); see Materials and Methods. The exponential form of the correlation, previously shown in Figure 2, is confirmed in both cases. The values of sequence homology vary between 100 and 31% with, or 35% without considering the gap penalty. Evidently, at least 60% sequence homology is needed in order to keep the structural difference below 1 Å (not considering the nine significantly differing protein pairs mentioned before).

No significant changes in the results are observed when the structural differences are calculated on the basis of all corresponding *main chain atoms* instead of the α-carbon atoms alone (data not shown). This can be easily explained by the low variability of the rigid and plane peptid units between adjacent α-carbon atoms.

Furthermore, the overall relationship between sequence identity or homology, on the one hand, and structural differences, on the other, is not biased by the preset cutoff distance—as one may object. We determined the effect of varying the cutoff values in the range between 2.0 and 3.8 Å (data not shown). The result is that the qualitative information (e.g., with respect to proteins deviating from the normal behavior) remains unaffected.

Total Structural Difference

As a measure for the total structural difference, the structurally correct alignment was extended into diverging loop regions, as described in the Materials and Methods section. The relationship between the overall structural divergence and sequence identity is given in Figure 5. The deviations now vary between 0.32 and 3.66 Å. Apart from the previously mentioned anomalies (Fig. 2), in this case one more protein pair, pancreatic phospholipases A_2 from *Bos taurus* (1BP2) and *Sus scrofa* (1P2P) (Fig. 5, number 10), differs distinctly from the exponential relationship. The total structural divergence yields 1.51 Å, compared with a sequence identity of 84.5%, normally correlating with a much lower divergence. The inspection of the structural superposition reveals an appreciable local conformational difference in the loop 59 to 70 caused by the single substitution of Val-63 (bovine, at the surface of the molecule) against Phe (porcine, in the interior) (Fig. 6a, b).[26] Therefore, correct modeling of this loop region on the basis of one of the two coordinates would fail, as the identity of the rest of the loop implies an unambiguous sequence alignment and successful

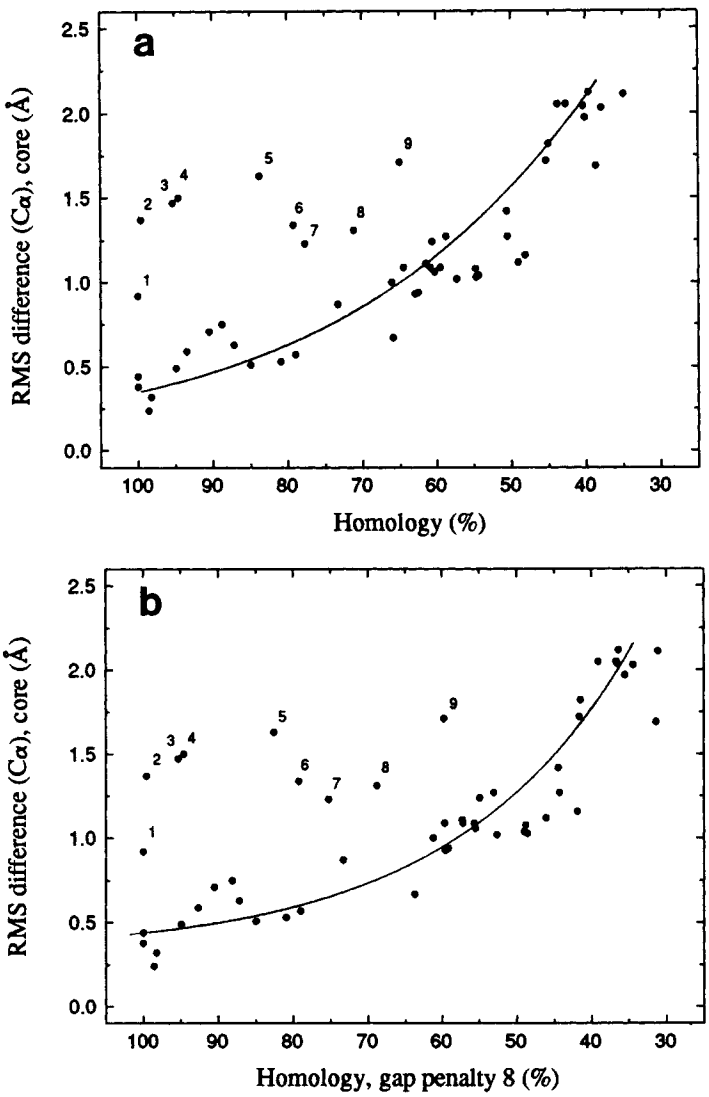

Fig. 4. Correlation of the RMS difference of corresponding α-carbon atoms and sequence homology based on the 250 PAM table **(a)** and, additionally, a gap penalty of 8 **(b)**. Numbering of the deviating protein pairs as in Figure 2.

modeling. However, ^{1}H NMR studies of bovine and porcine phospholipases A_2 suggest that the solution-state conformations of the two enzymes are very similar in this region.[27] The deviation may arise from crystal packing effects or, more likely in this case, errors in the X-ray analysis.[28]

The Relevance of Structurally Divergent Regions

The relevance of the structurally nonconserved regions for protein modeling may be investigated by plotting the correlation between the structural divergences resulting from the superposition of the common core regions and the complete molecules (Fig. 7). The broken line defines the limiting case where RMS differences in the core and in the complete molecule are identical. If the RMS deviation in conserved regions is less than about 1.0 Å (corresponding to a sequence identity of at least 50%, see Fig. 2), the overall structural difference yields comparable values. This implies that the structurally divergent regions (loops, turns) have similar conformational arrangements; therefore, modeling will be successful even in loop regions. However, at a lower degree of sequence identity, the total RMS difference steadily increases, with the deviations in the conserved regions of the common core as a reference. As a consequence, modeling of homologous loop regions at low degrees of sequence identity does not yield unequivocal results. Numbers 1, 2, and 3 in Figure 7 refer to three anomalous protein pairs: in the case of 4TNC:5TNC (number 1) and 1WRP:

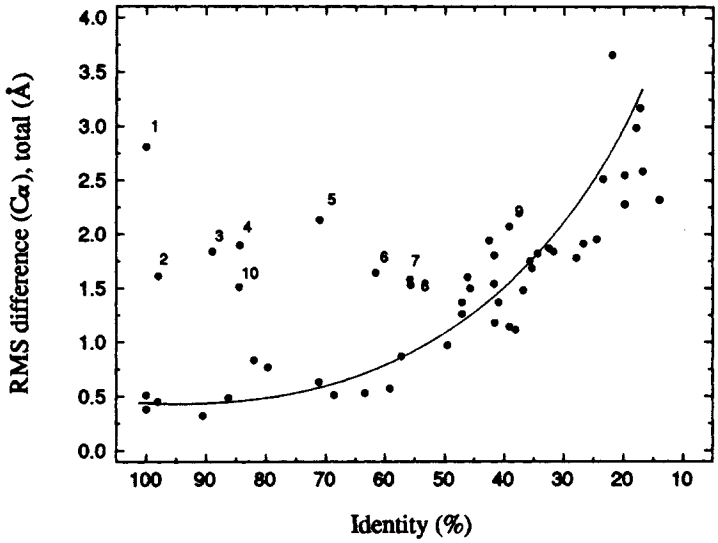

Fig. 5. Relationship between the RMS difference of corresponding α-carbon atoms considering the complete superposition and sequence identity. Numbering of the deviating protein pairs as in Figure 2; number 10 denotes the superposition of the phospholipases A_2 (1BP2 and 1P2P).

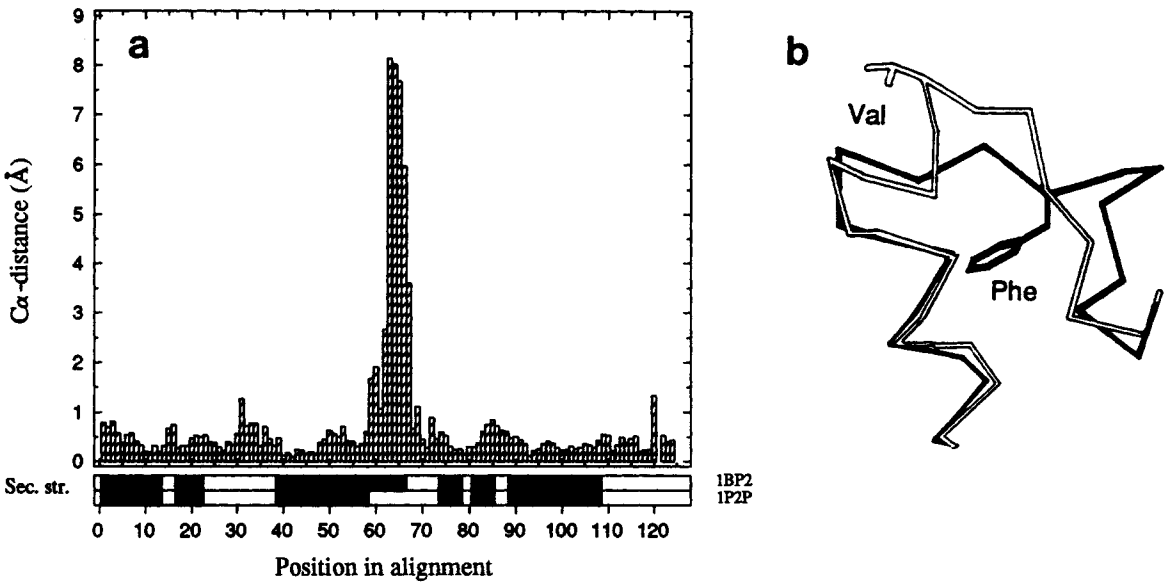

Fig. 6. The phospholipases A_2 (1BP2:1P2P). **(a)** Distances of α-carbon atoms at the corresponding positions in the structurally correct alignment. The filled boxes indicate elements of secondary structure (Sec.str.), i.e., α-helices and β-sheets. **(b)** Structurally divergent region within the superposition of the phospholipases A_2, apparently caused by the substitution of Val-63 (1BP2, in the surface of the molecule) against Phe (1P2P, in the interior of the molecule) (1P2P: black line). The deviation may equally well be due to an error in X-ray analysis.

2WRP (number 3), flexible C- or N-terminal arms are responsible for the high values of the total RMS difference, whereas number 2 refers to the phospholipases A_2 (1BP2 and 1P2P) discussed before.

Considering β-carbon atoms in addition to the α-carbons discussed so far, information with respect to the relative topological positions of corresponding side chains may be gained. In this case normalization was accomplished by superimposing corresponding α-carbon atoms (see Materials and Methods). In Figure 8, the RMS differences of normalized β-carbon atoms with respect to the common core region, on the one hand (Fig. 8a), and the complete molecules, on the other (Fig. 8b), are related to sequence identity. Given a sequence identity of about 50% or more, both correlation functions show simi-

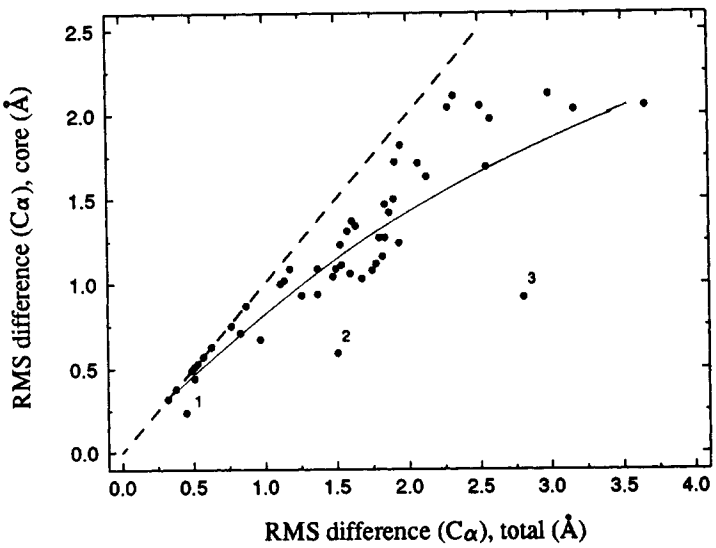

Fig. 7. Correlation between the structural divergence of the core region and the complete superposition based on corresponding α-carbon atom distances. Numbers 1, 2, and 3 denote the protein pairs 4TNC:5TNC, 1BP2:1P2P, and 1WRP:2WRP. The broken line defines the limiting case where the RMS difference in the cores and in the complete molecules are identical.

lar results, with the average structural divergence of normalized β-carbon atoms ranging between 0.31 and 0.55 Å. If the sequence identity drops below 50%, the overall structural divergence increases relative to the common core region as shown in Figure 8c. Assuming a total structural divergence of about 1.0 Å (corresponding to a sequence identity between 20 and 30%, cf. Fig. 8b), the average deviation of the two RMS differences (regarding the common core and the complete molecules, respectively) yields 0.25 Å. Thus, as sequence identity decreases below 50%, not only the protein backbone but also the topological orientation of corresponding side chains steadily diverge in the structurally nonconserved regions, compared to the common core.

Insertions

Intuition suggests that decreasing sequence identity correlates with increasing numbers of insertions and deletions. To quantify this assumption, the total number of insertions necessary for a correct structural alignment is calculated. Figure 9 corroborates the hypothesis by showing that increased numbers of insertions clearly correlate with decreasing sequence identity. A maximum of 16 is observed for 1LDM:4MHD. At a high level of identity, beyond 60%, virtually no insertions occur, with the exception of three protein pairs, 1BP2:1P2P (84.5% identity, one insertion), 2LBP:2LIV (79.7% identity, two insertions), and 1LDM:5LDH (71.1% identity, three insertions). However, the insertions in the two latter cases comprise no more than two additional amino acids, given a total length of the sequences of more than 320 amino acids. The obvious plateau at a level of four insertions within a range of sequence identities between 14 and 28% cannot be explained by the selection of medium-sized monomers in the database, nor can it be correlated with the number of loop regions of the corresponding protein structures (data not shown). It may be fortuitous and merely a consequence of the selection of protein structures from the Brookhaven Protein Data Bank.

CONCLUDING REMARKS

To set up homology modeling as a reliable tool in structure prediction, limitations have to be evaluated in an objective, standardized way. The database is constituted of well-established X-ray and NMR data of comparably high resolution, and modeling of a sufficiently large set of structures. As has been demonstrated, the analysis of pairwise superpositions of known structures from different conformational and functional classes, but various degrees of homology, provides useful information regarding the accuracy of modeling.

The correlations between several structural properties and sequence identity deduced from the correct alignments demonstrate that a known protein structure may provide a close model for other proteins if the sequence identity amounts to 50% or more. At a lower degree of identity the deviation in structurally not conserved regions continually increases with respect to both the protein backbone and the relative positions of side chains; this is induced by loop regions which are difficult to predict with certainty. Therefore, homology modeling studies of structures with less than 50% sequence iden-

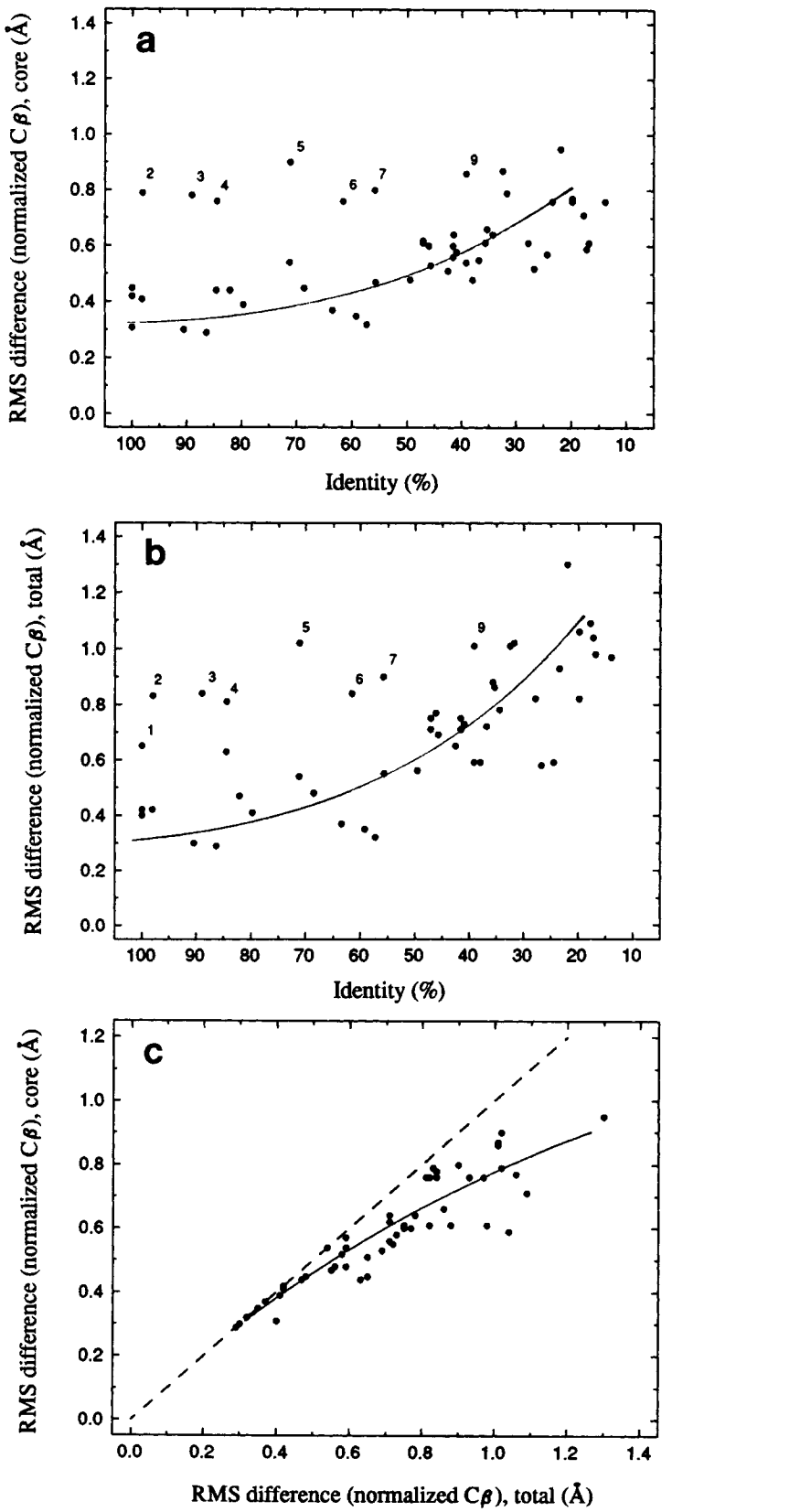

Fig. 8. The RMS difference of normalized β-carbon atoms regarding **(a)** the common core region and **(b)** the complete molecules as a function of sequence identity. Numbering of the deviating protein pairs as in Figure 2. **(c)** Correlation between the structural divergence of the core region and the complete superposition based on normalized β-carbon atoms. The broken line defines the limiting case where the RMS differences in the cores and in the complete molecules are identical.

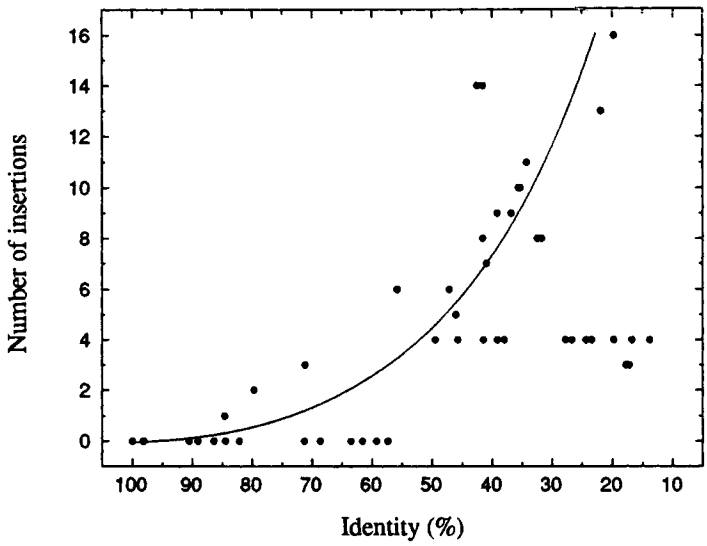

Fig. 9. The number of insertions necessary for a structural correct alignment as a function of sequence identity.

tity may provide useful information. Computer calculations along this line are in progress.

The overall relationship between structural homology and sequence identity is not valid in cases where functional requirements cause homologous structures to diverge to an unexpected extent; this holds especially for local regions involved in catalysis or regulation. Thus, without knowledge regarding, e.g., the mode of ligand binding etc., homology modeling may lead to large deviations from the real structure of a given protein. Crystallographic data of low quality, i.e., with resolution values greater than 2.5 Å or insufficient refinement, may cause similar problems. Therefore, if modeling requires such protein structures to be used as templates, the degree of structural deviation may yield an unexpected high value. In the case of low sequence identity, insertions have to be carefully taken into account; given identities beyond 60%, they may generally be ignored for the structural alignment.

The main goal of further studies will be attempts to develop criteria for knowledge-based modeling at various degrees of homology, especially low degrees, and to define a standard test protocol to estimate the quality of the various methods applied.

ACKNOWLEDGMENTS

Work was supported by the Deutsche Forschungsgemeinschaft (Grant Ja 78/29-2) and the Fonds der Chemischen Industrie.

REFERENCES

1. Levinthal, C. Are there pathways of protein folding? J. Chim. Phys. 65:44–45, 1968.
2. Jaenicke, R. Is there a code for protein folding? In: "Protein Structure and Protein Engineering, 39. Colloquium—Mosbach 1988." Winnacker, E.-L., Huber, R. (eds). Berlin: Springer-Verlag, 1988: 16–36.
3. Suhai, S. Modeling of protein structures on the basis of sequence data. In "Modern Methods in Protein- and Nucleic Acid Research." Tschesche, H. (ed.). Berlin: Walter de Gruyter, 1990: 395–422.
4. Rooman, M.J., Wodak, S. Identification of predictive sequence motifs limited by protein structure database size. Nature (London) 355:45–49, 1988.
5. Frauenfelder, H., Parak, F., Young, R.D. Conformational substates in proteins. Annu. Rev. Biophys. Biophys. Chem. 17:451–479, 1988.
6. Jaenicke, R. What does protein refolding in vitro tell us about protein folding in the cell? Phil. Trans. Roy. Soc. B, 339:287–295, 1993.
7. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature (London) 326: 347–352, 1987.
8. Moult, J. Comparative modeling of protein structure—progress and prospects. J. Res. Natl. Inst. Stand. Technol. 94:79–84, 1989.
9. Sali, A., Overington, J.P., Johnson, M.S., Blundell, T.L. From comparisons of protein sequences and structures to protein modeling and design. Trends Biochem. Sci. 15: 235–240, 1990.
10. Moult, J., James, M.N.G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1:146–163, 1986.
11. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
12. Svensson, B., Vass, I., Cedergren, E., Strying, S. Structure of donor side components in photosystem II predicted by computer modelling. EMBO J. 9:2051–2060, 1990.
13. Robson, B., Platt, E. Comparison of the X-ray structure of baboon α-lactalbumin and the tertiary predicted computer models of human α-lactalbumin. J. Comp. Aid. Mol. Design 4:369–379, 1990.
14. Weber, I.T. Evaluation of homology modelling of HIV protease. Proteins 7:172–184, 1990.
15. Scully, J.L., Evans, D.R. Comparative modeling of mammalian aspartate transcarbamylase. Proteins 9:191–206, 1991.
16. 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. The protein data bank: A computer-

based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
17. Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L., Sutcliffe, M. Knowledge-based protein modeling and design. Eur. J. Biochem. 172:513–520, 1988.
18. Sutcliffe, M.J., Haneef, I., Carney, D., Blundell, T.L. Knowledge based modeling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng. 1:377–384, 1987.
19. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
20. Lesk, A.M., Levitt, M., Chothia, C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng. 1:77–78, 1986.
21. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure," Vol. 5, Supplement 3. Dayhoff, M.O. (ed.). Washington, D.C.: National Biomedical Research Foundation, 1978: 345–352.
22. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
23. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: The significance of the root-mean-square deviation. J. Mol. Biol. 138:321–333, 1980.
24. Lawson, C.L., Zhang, R., Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Sigler, P.B. Flexibility of the DNA-binding domains of trp repressor. Proteins 3:18–31, 1988.
25. Grau, U.M., Trommer, W.E., Rossmann, M.G. Structure of the active ternary complex of pig heart lactate dehydrogenase with S-lac-NAD at 2.7 Å resolution. J. Mol. Biol. 151: 289–307, 1981.
26. Dijstraka, B.W., Renetseder, R., Kalk, K.H., Hol, W.G.J., Drenth, J. Structure of porcine pancreatic phospholipase A2 at 2.6 Å resolution and comparison with bovine phospholipase A2. J. Mol. Biol. 168:163–179, 1983.
27. Fisher, J., Primrose, W.U., Roberts G.C.K., Dekker, N., Boelens, R., Kaptein, R., Slotboom, A.J. ^{1}H NMR studies of bovine and porcine phospholipase A_2: Assignment of aromatic resonances and evidence for a conformational equilibrium in solution. Biochemistry 28:5939–5946, 1989.
28. Topham, C. M., Thomas, P., Overington, J. P., Johnson, M. S., Eisenmenger, F., Blundell, T. L. An assessment of composer: a rule-based approach to modelling protein structure. In: "Protein Structure, Prediction and Design." Kay, J., Lunt, G. G., Osguthorpe, D. J. (eds.). London: Biochemical Society Symposium, 1990: 1–9.