Protein Science (1992), *1*, 1269–1278. Cambridge University Press. Printed in the USA.

Correlation functions as a tool for protein modeling and structure analysis*

GERALD BÖHM AND RAINER JAENICKE

Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Universitätsstraße 31, D-8400 Regensburg, Germany

(RECEIVED April 23, 1992; ACCEPTED May 5, 1992)

Abstract

Proteins present unique folding structures whose conformations are determined primarily by their amino acid sequences. At present, there is no algorithm that would correlate the sequences with the structures determined by X-ray analysis or NMR. Comparative modeling of a new protein sequence based on the known structure of a functionally related protein promises to yield model structures that may provide relevant properties of the protein. To analyze the quality of a model structure, a set of correlation functions was derived from calculations on a subset of proteins from the structure database. Twenty-three highly resolved protein structures with resolutions of at least 1.7 Å from various protein families were used as the primary database. The purpose of this initial work was to find highly sensitive functions (including statistical error limits for the parameters) that describe properties of "real" proteins. Each correlation described is characterized by the correlation coefficient, the parameters for linear or nonlinear regression (coefficients of the equation), standard deviation and variance, and the confidence limits describing the statistical probability for values to occur within these limits, e.g., the natural variability of the property under examination. In addition, a method was developed for creating reasonably misfolded proteins. The ability of a correlation function to discriminate between the native structure and the misfolded conformations is expressed by the reliability index, which indicates the sensitivity of a correlation function. The term correlation functions thus summarizes a variety of efforts to find a mathematical description for the properties of protein structures, for their correlation, and for their significance.

Keywords: homology modeling; model structure verification; protein folding; protein model

The deduction of the structure of a protein from the amino acid sequence is still an unresolved problem because the protein folding code is not known (Jaenicke, 1987, 1988). The rational design of molecular structures of drugs, polymers, proteins, etc. is a central field of research; presently the most promising method of structure prediction is the design of molecules based on sequence homology to known structures by comparative modeling (Blundell et al., 1987; Moult, 1989).

Unfortunately, there are no unambiguous criteria at hand that would discriminate between useful and wrong model structures. In this context, a study of a number of parameters has been performed by Novòtny and coworkers (1988). Comparing proteins in their native and completely misfolded states, they arrived at the conclusion that there are only a few criteria that are suited to evaluate the quality of a model structure. The most significant ones are (1) the ratio of the solvent-accessible surfaces of apolar and polar side chains, (2) empirical free energy functions, and (3) the number of buried charged atoms. Further studies including three-dimensional profile scores allow localization of differences between correctly and wrongly folded or mistraced structures without providing an unambiguous assessment (Baumann et al., 1989; Lüthy et al., 1992).

The correlation between the size of typical globular proteins and their respective surfaces has been proposed by Miller and coworkers (1987). More recent studies revealed that there is a strong correlation between the size of a protein and the solvation free energy of its folding (Chiche et al., 1990); this correlation should therefore be a useful criterion for the detection of incorrect model structures.

In connection with attempts to apply homology modeling to an extremely halophilic dihydrofolate reductase from *Halobacterium volcanii* and a hyperthermophilic

* This paper is dedicated to Professor Peter L. Privalov on the occasion of his 60th birthday.

Reprint requests to: Gerald Böhm, Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Universitätsstraße 31, D-8400 Regensburg, Germany.

D-glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima* (Böhm, 1992), a large set of properties derived from known protein structures was calculated. The initial aim was to find an appropriate mathematical and physical description of known high-resolution crystal structures for the parameterized quality control of the above-mentioned model structures. For this purpose, a subset of 23 out of the ca. 300 highly resolved proteins from the Brookhaven Protein Data Bank was examined with regard to their molecular properties. From these data a large number of correlation functions, the respective correlation coefficients, confidence limits, as well as their variance and standard deviations were derived. Taken together, these characteristics may be assumed to describe the structure of proteins beyond the chosen subset. Models that fit all of the functions within a given error limit (e.g., one standard deviation, i.e., 65% confidence level, or two standard deviations, i.e., 95% confidence level) are considered to be reliable models with respect to the applied criteria.

The application of correlation functions is widespread and includes, for example, the determination of the reliability of model structures from homology modeling; the evaluation of the quality of crystal structures, especially at low resolution; the description of anomalous properties of proteins from extremophiles (e.g., extreme halophiles or thermophiles); and the automatic generation and analysis of site-directed mutants of enzymes (e.g., for thermostabilization purposes).

Results and discussion

Presently, a total of 335 properties (cf. Table 2) are included in the list of parameters, with 316 presently computed characteristics for all 23 proteins from the database (see Diskette Appendix for calculated properties for the 23 proteins). They lead to a matrix of $316 \times 315 = 99{,}540$ pairwise correlations. Correlation analysis was performed using (1) linear correlation analysis and (2) nonlinear analysis using polynomials up to the fifth degree.

The importance of hydrogen bonds and salt bridges for the stabilization of native structures has been discussed previously (Jaenicke, 1987, 1991a). In order to define the geometrical restraints for these two types of interactions, the proteins summarized in Table 1 were examined with respect to the distance constraints. Figure 1 shows the distances between positively and negatively charged atom centers for the proteins from the database. As can be seen, there is a significant positive interaction that ends at a distance of 5.9 Å, apart from the random (Gaussian) distribution of the distances. This value is therefore taken to be the maximal distance of salt bridges. Figure 2 displays the geometrical constraints observed for the hydrogen bonds in the protein database; hydrogen bond angles (Fig. 2a) are distributed between 60 and 180° without a significant preference, whereas the distance constraint (Fig. 2b) has a maximum at a donor–acceptor distance of about 2.9 Å (considering atomic centers).

Table 1. *List of protein structures included in this work for the database*

Code	Protein	E.C. type	Source organism	Resolution
1bp2	Phospholipase A_2	Hydrolase	*Bos taurus* (pancreas)	1.7 Å
1ccr	Cytochrome c	Electron transport	*Oryza sativa* (embryos)	1.5 Å
1crn	Crambin	Plant seed protein	*Crambe abyssinica* (seed)	1.5 Å
1ecd	Erythrocruorin (deoxy)	Oxygen transport	*Chironomus thummi thummi*	1.4 Å
1gcr	γ-Crystallin	Crystallin	*B. taurus* (eye lens)	1.6 Å
1ins	Insulin	Hormone	*Sus scrofa*	1.5 Å
1lz1	Lysozyme	Hydrolase (*O*-glycosyl)	*Homo sapiens*	1.5 Å
1mbd	Myoglobin (deoxy)	Oxygen storage	*Physeter catodon*	1.4 Å
1nxb	Neurotoxin B	Neurotoxin (postsynaptic)	*Laticauda semifasciata*	1.4 Å
1pcy	Plastocyanin	Electron transport	*Populus nigra* var. *italica*	1.6 Å
1ppt	APP	Pancreatic hormone	*Meleagris gallopavo*	1.4 Å
1tpp	β-Trypsin	Serine proteinase	*B. taurus* (pancreas)	1.4 Å
1ubq	Ubiquitin	Chromosomal protein	*H. sapiens* (erythrocytes)	1.8 Å
2act	Actinidin	Sulfhydryl hydrolase	*Actinidia chinensis*	1.7 Å
2mhr	Myohemerythrin	Oxygen binding	*Themiste zostericola*	1.7 Å
2tmn	Thermolysin	Hydrolase	*Bacillus thermoproteolyticus*	1.6 Å
3dfr	Dihydrofolate reductase	Oxidoreductase	*Lactobacillus casei*	1.7 Å
3est	Elastase	Serine proteinase	*S. scrofa* (pancreas)	1.65 Å
4dfr	Dihydrofolate reductase	Oxidoreductase	*Escherichia coli*	1.7 Å
4rxn	Rubredoxin	Electron transfer (iron-sulfur)	*Clostridium pasteurianum*	1.2 Å
5pti	Trypsin inhibitor	Proteinase inhibitor (trypsin)	*B. taurus* (pancreas)	1.0 Å
7rsa	Ribonuclease A	Phosphodiester hydrolase	*B. taurus* (pancreas)	1.3 Å
9pap	Papain	Sulfhydryl proteinase	*Carica papaya* (fruit latex)	1.65 Å

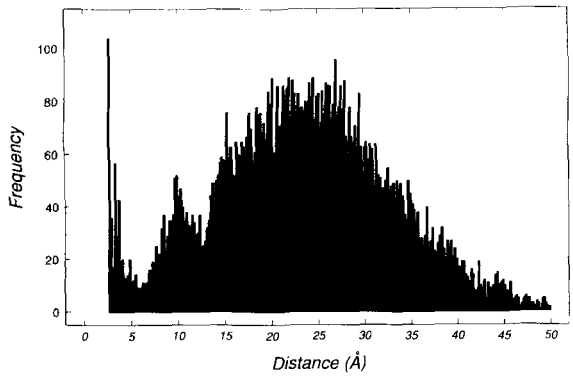

Fig. 1. Frequency distribution of distances between positively and negatively charged atom centers, i.e., salt bridges, in the proteins from the database.

Figure 3a–k shows the sample correlation functions that were selected for this work. Evidently, this selection is arbitrary, given a total number of pairwise correlations of 10^5. The common denominator is the molecular mass. As summarized in Table 2, the correlations could equally combine other parameters such as buried surface properties, chemical properties, etc. Functions with high correlation coefficients (cf. Fig. 3d) are useful for discriminating native from incorrectly folded structures, whereas functions with low correlation coefficients are relevant in discussing structure formation from a theoretical point of view.

For example, it is striking that buried hydrophobic surface areas exhibit large deviations from linearity with respect to the molecular mass. This may be attributed either to deviations from the additivity of the transfer energies per residue or to possible errors caused by the fact that the thermodynamic parameters represent small differences between large numbers (Jaenicke, 1991b).

Figure 3a shows a simple relationship between two statistical properties: the normalized hydrophobicity according to Sweet and Eisenberg (Cornette et al., 1987) and the

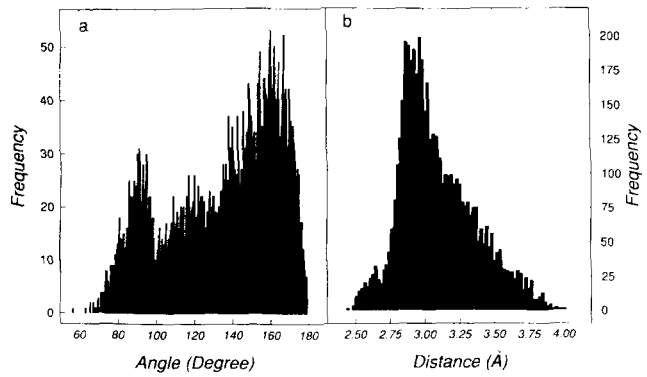

Fig. 2. Hydrogen bond geometries as derived from the proteins in the database. **a:** Distribution of dihedral angles between donor, hydrogen, and acceptor. **b:** Distance between donor and acceptor atomic center.

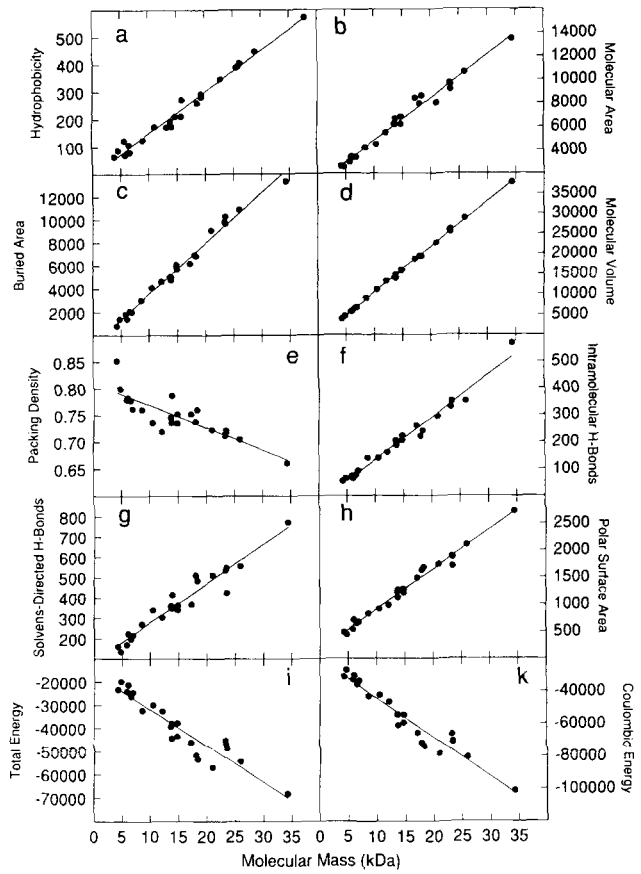

Fig. 3. Sample correlation functions. **a:** Hydrophobicity according to Sweet and Eisenberg (Cornette et al., 1987). **b:** The molecular surface area is a linear (!) function of the molecular mass. **c:** The buried molecular area. **d:** The native solvent-excluded volume is highly correlated to the molecular mass. **e:** The packing density drops with increasing size of the proteins. **f:** The number of intramolecular hydrogen bonds. **g:** The number of solvent-directed hydrogen bonds. **h:** Effectively charged surface area (polarity is defined by the MNDO/STO method). **i:** Total forcefield energy from the Discover (CVFF forcefield) program. **k:** Coulombic energy term from the Discover (CVFF forcefield) program.

molecular mass, both derived exclusively from the sequence of the proteins in the database. This correlation seems trivial, since no structural parameters are taken into account; however, not all of the normalized hydrophobicity properties summarized in Table 2 follow a linear relationship when correlated with the molecular mass.

Figure 3b,c describes geometrical properties of the surface and the core of the proteins. The calculated areas are similar for both the external surface (Fig. 3b) and the "internal surface" buried in the core of the protein (Fig. 3c). This does not hold for the small proteins where significant differences between the two surfaces are observed. Surprisingly, the surfaces may be approximated by a linear function; earlier work has shown a nonlinear relationship between the size and the surface of proteins (Miller et al., 1987). However, this holds only if hydrogen atoms are neglected and Pauling's dataset for the van der Waals

Table 2. *List of the properties available for each of the proteins from Table 1*

- Documentation
  - Protein Data Bank filename
  - Name of the protein
  - Enzyme classification number
  - Enzyme classification type
  - Source organism
  - Resolution of crystal data
  - *R* factor after refinement
  - Number of reliability datasets
- Hydrophobicity properties
  - Normalized hydrophobicity scale, Zimmermann (1968)
  - Normalized hydrophobicity scale, Jones (1975)
  - Normalized hydrophobicity scale, Levitt (1976)
  - Normalized hydrophobicity scale, Hopp and Wood (1981)
  - Normalized hydrophobicity scale, Fauchere and Pliska (1983)
  - Normalized hydrophobicity scale, Kuntz (1971)
  - Normalized hydrophobicity scale, Aboderin (1971)
  - Normalized hydrophobicity scale, Meek (1980)
  - Normalized hydrophobicity scale, Bull and Breese (1973)
  - Normalized hydrophobicity scale, Eisenberg et al. (1982)
  - Normalized hydrophobicity scale, Kyte and Doolittle (1982)
  - Normalized hydrophobicity scale, Chothia (1976)
  - Normalized hydrophobicity scale, Wertz and Scheraga (1978)
  - Normalized hydrophobicity scale, Janin (1979)
  - Normalized hydrophobicity scale, Olsen (1980)
  - Normalized hydrophobicity scale, Meirovitch et al. (1980)
  - Normalized hydrophobicity scale, Ponnuswamy et al. (1980)
  - Normalized hydrophobicity scale, Chothia (ΔG)
  - Normalized hydrophobicity scale, Wertz and Scheraga (ΔG)
  - Normalized hydrophobicity scale, Janin (ΔG)
  - Normalized hydrophobicity scale, Guy (1985)
  - Normalized hydrophobicity scale, Guy (mean) (1985)
  - Normalized hydrophobicity scale, Krigbaum and Komoriya interaction (1979)
  - Normalized hydrophobicity scale, Krigbaum and Komoriya transfer (1979)
  - Normalized hydrophobicity scale, Nishikawa and Ooi (1980)
  - Normalized hydrophobicity scale, Mijazawa and Jernigan (1985)
  - Normalized hydrophobicity scale, Rose (1985)
  - Normalized hydrophobicity scale, Sweet and Eisenberg (1983)
  - Normalized hydrophobicity scale, Dayhoff et al. (1978)
  - Normalized hydrophobicity scale, Heijne and Blomberg (1979)
  - Normalized hydrophobicity scale, Frömmel (1984)
  - Normalized hydrophobicity scale, Eisenberg and MacLachlan (1986)
  - Maximal continuous hydrophobic surface area
- Native surface properties
  - Total molecular area
  - Total van der Waals surface
  - Total contact surface
  - Total reentrant surface
  - Total negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total positive monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total positive and negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total noncharged monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total negatively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total positively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
- Native surface properties (*continued*)
  - Total positively and negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total positively minus negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total weighted negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total weighted positive surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total weighted positive plus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Total weighted positive minus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
- Unfolded surface properties
  - Unfolded molecular area
  - Unfolded van der Waals surface
  - Unfolded contact surface
  - Unfolded reentrant surface
  - Unfolded negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded positive monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded positive and negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded noncharged monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded negatively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded positively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded positively and negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded positively minus negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded weighted negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded weighted positive surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded weighted positive plus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Unfolded weighted positive minus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
- Buried surface properties
  - Buried molecular area
  - Buried van der Waals surface
  - Buried contact surface
  - Buried reentrant surface
  - Buried negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried positive monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried positive and negative monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried noncharged monopolar surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried negatively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried positively (partially) charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried positively and negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried positively minus negatively charged surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)

(*continued*)

Table 2. *Continued*

- Buried surface properties (*continued*)
  - Buried weighted negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried weighted positive surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried weighted positive plus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
  - Buried weighted positive minus negative surface (CVFF, Amber, Charge/2, ECEPP/2, MNDO)
- Statistical properties
  - Number of atoms
  - Number of residues
  - Number of subunits
  - Molecular mass
  - Calculated partial specific volume (Cohn & Edsall, 1943)
  - Absolute and fractional number of traditional polar residues
  - Absolute and fractional number of traditional apolar residues
  - Absolute and fractional number of traditional positive residues
  - Absolute and fractional number of traditional negative residues
  - Absolute and fractional number of potential hydrogen bond acceptors
  - Absolute and fractional number of potential hydrogen bond donors
  - Absolute and fractional number of methyl groups
  - Absolute and fractional number of sulfhydryl groups
  - Absolute and fractional number of amide groups
  - Absolute and fractional number of guanidine groups
  - Absolute and fractional number of carboxyl groups
  - Absolute and fractional number of imidazole groups
  - Absolute and fractional number of hydroxyl groups
  - Absolute and fractional number of aromatic rings
  - Ratio of polar versus apolar residues
  - Ratio of negative versus positive residues
- Geometrical properties
  - Absolute and fractional number of intramolecular hydrogen bonds
  - Absolute and fractional number of hydrogen bonds between backbone atoms
  - Absolute and fractional number of hydrogen bonds between backbone and side chain
  - Absolute and fractional number of hydrogen bonds between side chains
  - Absolute and fractional number of solvent-directed hydrogen bonds
  - Absolute and fractional number of solvent-directed hydrogen bonds from backbone
  - Absolute and fractional number of solvent-directed hydrogen bonds from side chains
  - Absolute and fractional number of disulfide bridges
  - Native solvent-excluded volume
- Geometrical properties (*continued*)
  - Unfolded solvent-excluded volume
  - Native molecular volume
  - Unfolded molecular volume
  - Ratio of solvent-excluded volumes
  - Packing density of protein core
  - Fractal index of protein surface
  - Number of alpha-helices in molecule
  - Number of beta-sheets in molecule
  - Number of turns in molecule
  - Absolute and fractional number of amino acids in helical conformation
  - Absolute and fractional number of amino acids in sheet conformation
  - Absolute and fractional number of amino acids in turn conformation
  - Absolute and fractional number of salt bridges
- Energetic properties
  - Molecular electrostatic energy (PB) in vacuum (kJ/mol)
  - Molecular electrostatic energy (PB) in water (kJ/mol)
  - Molecular solvation energy (PB) (kJ/mol)
  - Completely and partially minimized molecular dipole moment
  - Completely and partially minimized bond energy (CVFF) (kJ/mol)
  - Completely and partially minimized theta energy (CVFF) (kJ/mol)
  - Completely and partially minimized phi energy (CVFF) (kJ/mol)
  - Completely and partially minimized out-of-plane energy (CVFF) (kJ/mol)
  - Completely and partially minimized nonbonded energy (CVFF) (kJ/mol)
  - Completely and partially minimized nonbonded dispersion energy (CVFF) (kJ/mol)
  - Completely and partially minimized nonbonded repulsion energy (CVFF) (kJ/mol)
  - Completely and partially minimized coulombic energy (CVFF) (kJ/mol)
  - Completely and partially minimized forcefield energy (CVFF) (kJ/mol)
  - Transfer energy from core to water
- Experimental properties
  - Change in molal heat capacity
  - Midpoint of heat transition
  - Sedimentation value $s_{20,w}$
  - Partial specific volume
  - Cold denaturation temperature midpoint
  - Minimal growth temperature of source
  - Optimal growth temperature of source
  - Maximal growth temperature of source
  - pH optimum of protein stability

radii (Pauling, 1960) is used, together with a united atoms model. As demonstrated previously, large proteins deviate significantly from idealized spheres (Bryant et al., 1989). In this respect, a description of the surface of proteins by means of a fractal geometry instead of the traditional Euclidean one may be more appropriate (Böhm, 1991).

The volume of the proteins (calculated by M. Connolly's program PQMS) shows a virtually perfect linear correlation with the molecular mass (Fig. 3d). This is easy to explain, because the packing of globular proteins tends toward a maximum in order to exclude water from the (hydrophobic) core and to optimize intramolecular interactions. However, with increasing size it becomes more and more complicated to pack all amino acids into the core, still maintaining maximum packing density. It is therefore not surprising that packing decreases with increasing protein size (Fig. 3e). In this context, packing density is defined as the ratio of the solvent-excluded volumes of the unfolded and the folded structures. A note

of caution must be added at this point: some of the structures in the database show cavities large enough to contain one or more water molecules. A filled cavity would increase the local packing density of an otherwise poorly packed region of the protein. Internal cavity water, however, is difficult to quantify by crystallography in an unambiguous way. Therefore, data given in Figure 3e refer to estimates of the packing density ignoring internal cavity water.

Figure 3f,g shows the number of intramolecular and solvent-directed hydrogen bonds occurring in the database. The calculated numbers may be considered arbitrary because hydrogen atoms are not part of common crystal structures and therefore their atomic coordinates have to be modeled; however, because all proteins were treated by the same standard protocol, their comparison should be valid. It should be noted that a maximum hydrogen-acceptor distance of 3.0 Å was chosen, and all possible dihedral angles between donor hydrogen and hydrogen acceptor were allowed (for the respective donor–acceptor distances and dihedral angles, see Fig. 2a,b). Thus, the data represent an upper limit for the number of hydrogen bonds, since additional limitations (e.g., a distance constraint of 2.1 or 2.5 Å) are useless as long as no precise energy functions are available that would describe individual hydrogen bonds in their specific electronic environment. It should be noted that there are about twice as many solvent-directed hydrogen bonds (modeled by water layers; Fig. 3g) than intramolecular hydrogen bonds (Fig. 3f), stressing the fact that molecular modeling calculations should always include the explicit representation of water molecules instead of artificial continuum models.

The previously mentioned correlation is also reflected in Figure 3h: the "effectively charged surface area"[1] describes the effective "polarity" that a given protein surface exposes to the solvent. This stabilizing force must increase with increasing molecular mass in order to compete with destabilizing (mainly entropic) forces. A similar correlation is found with the respective polarity inside the proteins, i.e., the buried surface areas.

Forcefield energies (Fig. 3i,k), as a result of numerous programs devised for molecular interaction calculations, may be assumed to assess the correctness of a model structure in a quantitative way. With increasing size of the proteins, the number of interactions grows linearly; correspondingly, the enthalpies are found to be increased (Fig. 3i). This holds also in cases where subterms of the forcefield algorithm, e.g., coulombic interactions, are examined (Fig. 3k).

Some of the correlations described previously may be considered trivial. However, because all data were obtained by a common protocol, they gain relevance from the fact that standardized parameters were used. The results are comparable within the given set of proteins, parameters, and even computer equipment; they may differ slightly from similar calculations that were previously published. Structures created by homology modeling should be comparable to the proteins from the database; if their properties deviate significantly from the correlations, either errors in the model structure or an anomalous property of the protein under consideration may be the reason.

In the present study, some sample correlation functions were selected that allow the quality of model structures to be estimated. The computer program written in connection with this work enables the user to compare model structures with more than 90,000 correlation functions based on the properties summarized in Table 2. Many of them are mutually related by a linear dependence and provide, therefore, only limited additional information. However, a complete analysis might find irregularities in structures that would not be detected otherwise. The high level of redundancies in the chosen properties, and, as a result, in the correlation functions, circumvents many problems in parameterization and allows a sound analysis of relations in protein structures.

The present correlation functions have been shown to be valuable tools in molecular modeling. They are, first, an attempt to describe mathematically certain relationships of properties in protein structures. The significance of these correlations is of importance for the quality control of protein models. This may be used for conformations created by homology modeling methods, as well as structures derived from X-ray crystallography, especially at low resolution. Second, in connection with reliability indices, correlation functions are expected to reveal criteria for specific properties of proteins from extremophiles, either extremely halophilic or hyperthermophilic (Böhm, 1992). Properties significantly different in mesophilic and thermophilic proteins are possible candidates for thermal adaptation strategies. Third, correlation functions may serve to devise mutagenesis experiments or protein design and protein engineering. An automated procedure may be developed to generate site-directed mutants that are subsequently analyzed in terms of new properties, e.g., thermal stabilization. The corresponding algorithm would simulate random mutagenesis and evolutionary events by exchanging in turn every sequence position with each naturally occurring amino acid. Correlation analysis is then used to determine the quality of the respective model structures and to estimate their stability properties. However, this requires the integration of forcefield and geometry calculation programs into the correlation analysis package; at the moment, these tools are used separately. The next generation of correlation functions will include local instead of global properties, e.g., rotamer

[1] This is defined as the sum of positively and negatively charged surfaces, multiplied by the partial charge—taken from the MNDO/STO charge set—that is associated with the surface; see Materials and methods.

library considerations. This will improve the usefulness of the functions by assigning local quality and reliability indices, thus allowing localization of sources of errors more directly. Aside from these practical considerations, the information database on properties and correlations will provide a quantitative description of general principles of protein structure in terms of specific properties such as electrostatic interactions, hydrogen bonds, and hydropathicity, among others.

Table 3. *Pairwise correlation between the charge datasets used; the correlation indicates the relatedness between the charge datasets*

	CVFF	Amber	Charge/2	ECEPP/2	MNDO/STO
CVFF	—	0.9042	0.9050	0.9019	0.9555
Amber	0.9042	—	0.9109	0.9575	0.9195
Charge/2	0.9050	0.9109	—	0.9358	0.9166
ECEPP/2	0.9019	0.9575	0.9358	—	0.9364
MNDO/STO	0.9555	0.9195	0.9166	0.9364	—

Materials and methods

Computer equipment

All structure calculation programs were installed on a Silicon Graphics workstation (Iris 4D/70GTB) with Unix System V rel. 3.3.2. The programs Insight II version 2.1, Discover versions 2.6 and 2.7, and DelPhi versions 2.0 and 2.1 (Biosym, Inc., San Diego, Calif.) were used for visualization, energy calculations, and electrostatic/solvation energy calculations, respectively. Discover 2.7 was mainly used on a Cray Y-MP 4/432 vector processing facility at the Leibniz-Rechenzentrum in München, Germany. For surface and volume calculations, Dr. M.L. Connolly's molecular surface program PQMS, version 1.6 (October 1991) was used on the Iris and Cray computer. Van der Waals radii were taken from Iijima et al. (1987) (cf. Table 3). Hydrogen atoms were always treated explicitly; the probe radius for calculating the accessible surface area was taken to be 1.40 Å. To compare the results of different approaches, area and volume calculations were performed with a modified version of the program Access version 2 on a VAX 11/750 computer equipped with VAX/VMS 5.4; the program was kindly provided by Professor F.M. Richards.

If data files for properties of proteins not included in the selected subset of 23 standard structures are intended, the programs Insight, Discover, and PQMS are necessary to prepare structure files and to compute energies and surfaces. In addition, some special programs are required, which are described below.

Selection of proteins

The Brookhaven Protein Data Bank, release 50 (Bernstein et al., 1977), was scanned for complete structures with resolutions of 1.7 Å or better. A set of 20 proteins representing 20 functional and/or conformational protein families with a broad range of molecular masses, functions, and varying quaternary structures was selected. To check for variability of homologous proteins, two dihydrofolate reductases, two serine proteases, and two cysteine proteases were included, making up a final set of 23 structures. The following Protein Data Bank coordinate files (in alphabetical order) are part of this work (Table 1): 1bp2, 1ccr, 1crn, 1ecd, 1gcr, 1ins, 1lz1, 1mbd, 1nxb, 1pcy, 1ppt, 1tpp, 1ubq, 2act, 2mhr, 2tmn, 3dfr, 3est, 4dfr, 4rxn, 5pti, 7rsa, 9pap. Included on the Diskette Appendix are 23 plain ASCII files containing lists of the calculated properties for the 23 protein structures (see file Bohm.doc in the folder Bohm.DAT, which is in the SUPLEMNT folder).

Missing side chains, which are not reported in the crystal structures due to their high flexibility, were inserted in standard configurations. Instead of using the "united atoms" model, hydrogens (usually not resolved by X-ray crystallography) were added in geometries causing minimum perturbation of the structures. Water molecules (including internal cavity water), coenzymes, substrate analogs, and other ligands were removed from the structures. For the statistical and geometrical part of the work, the unmodified coordinates of the crystal structures were used, in contrast to the energy calculations, where small local structural alterations were allowed in order to avoid unreasonable strain in the molecules and to fit the structure to one set of forcefield parameters.

Simulation of the denatured state

To determine driving forces in protein folding, model structures for the denatured state of each of the proteins from the database subset were built. The backbone folding of the native structures was changed into an all-β-strand conformation for the complete polypeptide chain, except the proline residues. This results in very long, nearly linear chains with maximum solvent-accessible surface area. The difference between this surface area of the denatured state and the native surface area is considered to be the interior buried surface area of the protein. This is a reasonable model for the protein core.

Calculated surfaces

Following the definitions of molecular surfaces by Connolly (1983) and Richards (1985), there are two kinds of surfaces: (1) the contact surface, i.e., the surface of the protein where the probe is in contact with the van der Waals envelope of an atom; and (2) the reentrant surface created by the probe touching two or more atoms. The

latter is not part of the van der Waals surface of one of the protein atoms and therefore reflects the irregularity of a surface. Table 3 summarizes different sets of van der Waals radii taken from the literature. In the present work the dataset of Iijima et al. (1987) was used because it is derived from experimental data taken from protein structures. Therefore it may be considered the most appropriate choice.

Effective charged surface area

In the present study, the effective charged surface area is introduced as a new concept similar to the one described by Baumann et al. (1989). It describes the effective electrostatic interaction of bordering surfaces and is calculated for every atom by the point charge of an atom, multiplied by the surface area associated with the atom. This allows interactions of surfaces with different sizes and/or different partial charges to be compared and follows from the model that upon folding of polypeptide chains corresponding surfaces with complementary charges approach each other. The major disadvantage of this concept is that the partial charges on proteins are generally ill-defined. To diminish this problem, five different charge sets from the literature were used in parallel (cf. Table 4): CVFF (taken from the program Discover version 2.6), Amber (Amber 3.0, 1986), ECEPP/2 (Némethy et al., 1983), Charge/2 (Abraham et al., 1991), and MNDO/STO (Gruschus & Kuki, 1990). The first three charge sets represent forcefield parameters, whereas the latter two were derived from atomic structure theory; a comparison of the charge sets by characterization of the frequency distribution of partial charge values is presented in Figure 4.

Using the concept of partial charges, the results of the correlation function analysis show that there are highly correlated properties. Since Kauzmann's fundamental work on the importance of the hydrophobic effect on protein stability (Kauzmann, 1959), there has been an ongoing discussion on the definition of hydrophobic groups. One controversial issue is the parameterization of the hydrophobic effect. This is because the accurate determination of the dipole moment of the functional

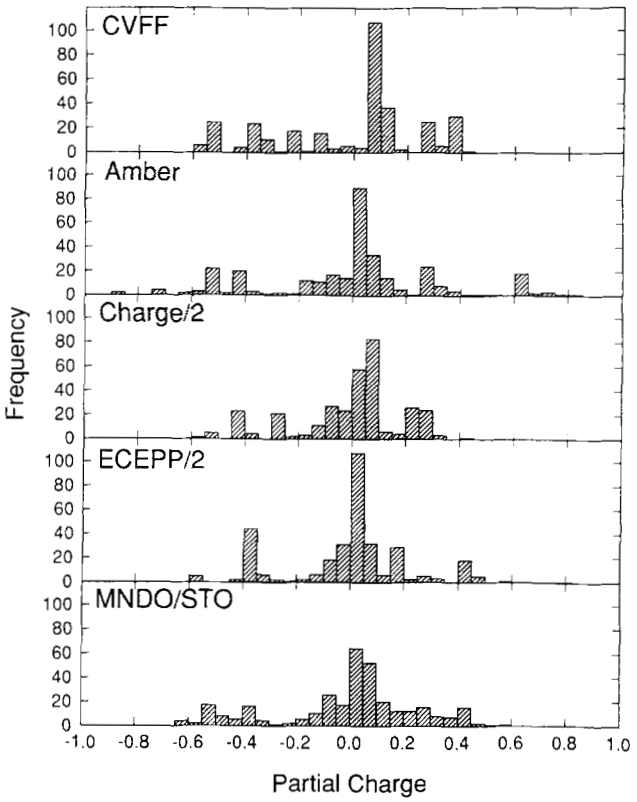

Fig. 4. Frequency of partial charge values from different charge sets. As indicated by the finite population at partial charges of ca. ±0.4, the boundary between polar and apolar groups is ill-defined.

groups in a given protein structure is not possible with current approximation methods. On the other hand, it is presently not feasible to do semi-empirical quantum mechanical calculations on protein structures, which would yield the distinct electron distribution. Because of the ill-defined boundary between polar and apolar groups (cf. Fig. 4), we did not use the concept of hydrophobic surfaces.

The concept of misfolded structures

To evaluate the quality of a correlation function, it has to be applied to sets of native structures and misfolded

Table 4. *Comparison of van der Waals radii as taken from the literature*[a]

	Carbon (C)	Nitrogen (N)	Oxygen (O)	Hydrogen (H)	Sulfur (S)
Pauling (1960)	1.85–1.9	1.65–1.75	1.6–1.7	1.2	1.9
Jorgensen (1981)	1.535	—	1.430	0.945	—
Iijima et al. (1987)	1.41	1.28	1.16	1.00	(1.5)

[a] The most commonly used dataset is that from Pauling; however, the dataset from Iijima et al. seems to be the most appropriate one for the present work because these data were derived from protein crystal structures. The value for sulfur is not defined in the work of Iijima et al. but is a linear estimate based on the other data.

conformations. Previous attempts to create misfolded structures were not satisfactory (Novòtny et al., 1988). Therefore, a new protocol was developed (Böhm, 1991): structures of native proteins are disrupted on the graphics screen by introducing random torsion angles in several regions. Starting from these conformations, molecular dynamics simulations "refold" the protein to a compact nativelike, but misfolded, structure. The degree and location of the disruptions, the simulation parameters, and the molecular dynamics temperature (van Gunsteren & Berendsen, 1990) allow a broad spectrum of more or less misfolded structures to be generated.

Correlation calculation

Linear and nonlinear correlations of data were calculated according to the linear regression equation

$$y = y_m + \frac{\sum(x_i \cdot y_i) - y_m \cdot \sum x_i}{\sum x_i^2 - x_m \cdot \sum x_i} \cdot (x - x_m), \qquad (1)$$

with the correlation coefficient r_{xy} for the correlation between values x_i and y_i defined by

$$r_{xy} = \frac{\sum(x_i - x_m) \cdot (y_i - y_m)}{\sqrt{\sum(x_i - x)^2 \cdot \sum(y_i - y)^2}}, \qquad (2)$$

where i ranges from the first to the last element (N), x_m and y_m are the respective averages of the data vectors ($x_m = N^{-1} \cdot \sum x_i$ and $y_m = N^{-1} \cdot \sum y_i$). The value for r_{xy} ranges between +1.0 and −1.0. The standard deviation σ for the xy pairs is defined by

$$\sigma_{xy} = \frac{1}{N-1} \cdot \left[\sum(x_i \cdot y_i) - \frac{\sum x_i \cdot \sum y_i}{N}\right], \qquad (3)$$

and the variance is σ_{xy}^2. For the nonlinear analysis, data were fitted by the Gaussian least-squares method to a polynomial instead of a linear equation:

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + a_4 \cdot x^4 + a_5 \cdot x^5, \qquad (4)$$

with $a_0 \ldots a_5$ as the parameters of the polynomial. The degree of the polynomial used in the calculations may vary between 2 and 5.

Reliability index

For each correlation function, a reliability index may be calculated. It ranges between 0 and 1 and describes the ability of the correlation function to discriminate between native structures and misfolded proteins. The calculation of reliability indices requires the availability of misfolded structures generated from the proteins in the database; to produce such data is extremely computer-time-consuming. The correlation function for the properties x and y is transformed into a Gauss–Laplace normal distribution:

$$\zeta = \left\{ \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot \int e - \frac{\left[y/x - \sum(y/x)/N\right]^2}{2 \cdot \sigma^2} \right\}. \qquad (5)$$

This is performed for the native database (ζ_n) as well as for the misfolded set of data (ζ_u). From the two resulting bell-shaped functions, the reliability index ξ is finally computed according to

$$\xi = 1 - \left[\frac{\zeta_{(u-n)}}{\zeta_u + \zeta_n}\right]^2, \qquad (6)$$

where ξ is reverse proportional to the overlap of the two functions.

Special programs

For the complete correlation analysis, as well as for the conversion of data formats, a series of special-purpose programs for the MS-DOS® environment was developed. Some are available by the anonymous/ftp mechanism, including complete files of calculated properties and correlation tables that are not suitable for printed publication because of their size.

The following steps are required: (1) Establish an internet connection to the local computer (rbisg1.biologie.uni-regensburg.de; internet number 132.199.1.42). (2) Log in as user "anonymous"; use your personal name instead of a password. (3) Set the mode of data transfer to binary (use the command "binary"). (4) Change to the directory "correlation_function" ("cd CORRELATION_FUNCTION"). (5) Transfer all files from the respective directory to your PC ("mget*"). (6) Move the files to the respective directory, and start the executable programs. They are self-extracting utilities that decompress all data files and programs upon execution. A complete version of the program will be distributed at the end of 1992; this version will use the Microsoft Windows® environment. Additional information for modifications to the programs will be found in the release notes distributed with the files.

Acknowledgments

Fruitful discussions with Drs. T. Kiefhaber, J. Buchner, and R. Rudolph are gratefully acknowledged. This work was supported by Deutsche Forschungsgemeinschaft grant Ja 78/29-1.

References

Abraham, R.J., Grant, G.H., Haworth, I.S., & Smith, P.E. (1991). Charge calculations in molecular mechanics. Part 8: Partial charges from classical calculations. *J. Comput. Aided Mol. Des. 5*, 21–39.

Amber 3.0 (1986). Written by Singh, U.C., Weiner, P.K., Caldwell, J.W., & Kollman, P.A. Department of Pharmaceutical Chemistry, University of California, San Francisco.

Baumann, G., Frömmel, C., & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng. 2*, 329–334.

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). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., & Thornton, J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature 326*, 347–352.

Böhm, G. (1991). Protein folding and deterministic chaos: Limits of protein folding simulations and calculations. *Chaos Solitons Fractals 1*, 375–382.

Böhm, G. (1992). Strukturmodellierung extremophiler Proteine: Analyse der Zuverlässigkeit von Strukturprognosen und Grenzen der Modellierbarkeit von Proteinen. Ph.D. Thesis, Universität Regensburg, Regensburg Germany.

Bryant, S.H., Islam, S.A., & Weaver, D.L. (1989). The surface area of monomeric proteins: Significance of power law behavior. *Proteins Struct. Funct. Genet. 6*, 418–423.

Chiche, L., Gregoret, L.M., Cohen, F.E., & Kollman, P.A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. USA 87*, 3240–3243.

Connolly, M.L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science 221*, 709–713.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol. 195*, 659–685.

Gruschus, J.M. & Kuki, A. (1990). Partial charges by multipole constraint. Application to the amino acids. *J. Comput. Chem. 11*, 978–993.

Iijima, H., Dunbar, J.B., & Marshall, G.R. (1987). Calibration of effective van der Waals atomic contact radii for proteins and peptides. *Proteins Struct. Funct. Genet. 2*, 330–339.

Jaenicke, R. (1987). Folding and association of proteins. *Prog. Biophys. Mol. Biol. 49*, 117–237.

Jaenicke, R. (1988). Is there a code for protein folding? In *Protein Structure and Protein Engineering, 39. Colloquium—Mosbach 1988* (Winnacker, E.-L. & Huber, R., Eds.), pp. 16–36. Springer-Verlag, Berlin, Heidelberg, New York.

Jaenicke, R. (1991a). Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem. 20*, 715–728.

Jaenicke, R. (1991b). Protein folding: Local structures, domains, subunits, and assemblies. *Biochemistry 30*, 3147–3161.

Jorgensen, W.L. (1981). Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *J. Am. Chem. Soc. 103*, 335–340.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. In *Advances in Protein Chemistry*, Vol. 14 (Anfinsen, C.B., Jr., Anson, M.L., Bailey, K., & Edsall, J.T., Eds.), pp. 1–63. Academic Press, New York, London.

Lüthy, R., Bowie, J.U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature 356*, 83–85.

Miller, S., Janin, J., Lesk, A.M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol. 196*, 641–656.

Moult, J. (1989). Comparative modeling of protein structure—Progress and prospects. *J. Res. Natl. Inst. Standards Technol. 94*, 79–84.

Némethy, G., Pottle, M.S., & Scheraga, H.A. (1983). Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem. 87*, 1883–1887.

Novòtny, J., Rashin, A.A., & Bruccoleri, R.E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct. Funct. Genet. 4*, 19–30.

Pauling, L. (1960). *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd Ed. Cornell University Press, Cornell, New York.

Richards, F.M. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol. 115*, 440–464.

van Gunsteren, W.F. & Berendsen, H.J.C. (1990). Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl. 29*, 992–1023.