PROTEINS: Structure, Function, and Genetics 41:47–57 (2000)

Discriminating Between Homodimeric and Monomeric Proteins in the Crystalline State

Hannes Ponstingl,[1] Kim Henrick,[1] and Janet M. Thornton[1,2]*

[1]*EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom*

[2]*Biomolecular Structure and Modelling Unit, Biochemistry and Molecular Biology Department, University College London and Crystallography Department, Birkbeck College, London, United Kingdom*

ABSTRACT **Scores calculated from intermolecular contacts of proteins in the crystalline state are used to differentiate monomeric and homodimeric proteins, by classification into two categories separated by a cut-off score value. The generalized classification error is estimated by using bootstrap re-sampling on a nonredundant set of 172 water-soluble proteins whose prevalent quaternary state in solution is known to be either monomeric or homodimeric. A statistical potential, based on atom-pair frequencies across interfaces observed with homodimers, is found to yield an error rate of 12.5%. This indicates a small but significant improvement over the measure of solvent accessible surface area buried in the contact interface, which achieves an error rate of 15.4%. A further modification of the latter parameter relating the two most extensive contacts of the crystal results in an even lower error rate of 11.1%. Proteins 2000;41:47–57.**

Key words: dimerization; crystal contacts; solvent accessible surface area; pair potential

INTRODUCTION

The number and association of subunits composing a multimeric protein can often not be derived without ambiguity from crystallographic studies of protein structure. Symmetry arguments are of limited help in selecting the functional assembly of subunits, not least because macromolecules frequently exhibit point-group symmetry coinciding with crystallographic symmetry. Structure entries deposited in the Protein Data Bank (PDB) often lack information on the quaternary structure of the protein in solution. A further complication arises from the lack of consistent information on symmetry operations that generate the macromolecule from the deposited atomic coordinates, which usually represent the asymmetric unit. However, this information is in many cases essential for a complete understanding of protein function.

The size of interchain contacts in protein crystals, usually measured by the solvent accessible surface area (ASA), can be exploited to discriminate functional subunit-subunit interfaces against unspecific contacts that are artifacts of the crystal packing.[1–3] This observation can be rationalized by the high specificity of subunit-subunit interactions requiring the formation of extensive contact area.

Subunit-subunit interfaces have been characterized also by various other parameters. Among these are hydrophobicity,[4–9] preference for certain amino acids[5,9,10] and geometric features like shape complementarity,[6,8] planarity, and circularity.[11] Although ranking schemes based on such parameters identified interaction patches on protein surfaces,[12] their power in discriminating functional from artificial crystal contacts remains to be scrutinized and compared with the relatively successful size criterion.

Based on contact area, a Protein Quarternary Structure (PQS) file server[13] is being maintained by the Macromolecular Structure Database (MSD) group at the European Bioinformatics Institute (http://pqs.ebi.ac.uk) from where for each entry of the PDB the most likely macromolecular assembly can be downloaded in PDB format. A preliminary estimation of the apparent error rate associated with the prediction of quaternary structure by contact area has been provided,[13] for which 19% mismatch with online annotations was found for the highly redundant set of 2,895 PDB entries classified as homodimers. This set also contained 190 entries of pseudodimeric lysozyme. For a further 18%, no online annotation was available.

The work presented here is an attempt to estimate this error on more rigorous statistical grounds and, based on equivalent measures, to propose criteria that improve on the contact area to be used for automatic quaternary structure assignment. As a first step toward solving this task, we focus again on the prediction of homodimer formation. To this end, we have compiled from the literature and from data bases a list of 172 nonhomologous structures of water-soluble proteins with a quaternary state known to be homodimeric or monomeric.

As an improvement over the parameter of contact area for the prediction of protein quaternary states, we use log-scores based on pair frequencies of residue or atom types observed across protein-protein interfaces.[14,15] The use of pair-frequency scores that originated in the field of protein tertiary structure recognition and prediction,[16–19] has been facilitated by the increasing amount of structural data available on protein complexes. Correlations with free energies and the generalization behavior seem to be

*Correspondence to: Janet M. Thornton, Biomolecular Structure and Modelling Unit, Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, United Kingdom. E-mail: thornton@biochem.ucl.ac.uk

Received 28 February 2000; Accepted 2 June 2000

TABLE I. The Data Set[†]

Monomers							
16PK	1A0K	1A19	1A6Q	1A8O	1AAY	1AF7	1AFK
1AH7	1AHQ	1AKO	1AKZ	1AM6	1AMJ	1AOH	1AUA
1AUN	1AVP	1AYI	1AYL	1BC2	1BE0	1BEA	1BG0
1BGC	1BKZ	1BMB	1BN8	1BP1	1BRY	1BU1	1BWZ
1C3D	1CKI	1CKM	1CTJ	1DFF	1DJX	1DMR	1EMA
1ESF	1ESO	1FDR	1FEH	1FLP	1FSU	1GCI	1IAE
1INP	1IPS	1KFS	1KPT	1KWA	1LRV	1MB1	1MDT
1MH1	1MPG	1NP4	1NUC	1OPS	1PDA	1PGS	1PJR
1PMI	1PPO	1PS1	1RGP	1RHS	1TON	1UCH	1URO
1VJW	1XGS	1YGE	1ZIN	232L	2ABX	2ACY	2ATJ
2BLS	2CY3	2END	2FGF	2GPR	2HEX	2IHL	2MBR
2MHR	2PTH	2RN2	3CMS	3DFR	3SIL	5CP4	8PAZ
Homodimers							
1A3C	1AD3	1AF5	1AFW	1AJS	1ALK	1ALO	1AMK
1AOM	1AOR	1AQ6	1AUO	1BAM	1BIF	1BSR	1BUO
1CG2	1CHM	1CMB	1CP2	1CSH	1CTT	1CZJ	1DAA
1FIP	1FRO	1GVP	1HJR	1HSS	1ICW	1IMB	1ISA
1ISO	1JHG	1JSG	1KBA	1KPF	1LYN	1MJL	1MKA
1MOQ	1NOX	1NSY	1OAC	1OPY	1OTP	1PGT	1PRE
1PUC	1RFB	1RPO	1SES	1SLT	1SMN	1SMT	1SOX
1TOX	1TRK	1TYS	1UBY	1UTG	1WGJ	1XSO	2CCY
2ILK	2RSP	2TCT	2TGI	3GRS	3PGH	3SDH	3SSI
4KBP	5CSM	5TMP	9WGA				

[†]PDB identification codes of 96 monomers and 76 dimers constituting the data set of water-soluble proteins used in this study. Protein chains within each subset exhibit sequence identities $< 25\%$ and are structurally nonhomologous.

generally satisfying, which is reflected in attempts to associate such pair scores with free energies by means of the Boltzmann statistic. However, improvements over simple considerations of hydrophobicity/polarity might be marginal,[20] and the theoretical basis for this approach has been questioned.[21]

Attracted by their simplicity, we test a distance-dependent atom-pair score as an automatic tool for the classification of protein homodimerization and estimate the classification error by using a bootstrap method[22] on the data set of 96 monomers and 76 homodimers. The classification error is compared with the error obtained when the simple measure of contact area is used.

The thermodynamic equilibrium between monomeric and dimeric forms of a protein is treated here as a two-state model with only two discrete categories for classification. This approach is reasonable because, in the vast majority of cases, the equilibrium is shifted far to the formation of either the monomer or the homodimer (see Discussion section).

MATERIALS AND METHODS

Data Set of Nonhomologous Protein Structures

The data set of 172 protein crystal structures used in this study, is divided into two macromolecular classes comprising 96 monomers and 76 homodimers. Members within one class are structurally nonhomologous and show less than 25% sequence identity. Atom coordinates of these molecules were obtained from the Protein Data Bank[23] (PDB) now operated by the Research Collaboratory for Structural Biology[24] (RCSB; http://www.rcsb.org) and the Macromolecular Structure Database group at the European Bioinformatics Institute (EBI-MSD; http://msd.ebi.ac.uk). Proteins were classified according to their prevalent multimeric state in solution as annotated in the PDB or in the protein sequence data base SWISS-PROT.[25] The classification was checked with the original literature when deemed necessary. Annotations of PDB and SWISS-PROT entries were extracted by using the Sequence Retrieval System[26] (SRS, version 5.0.3). Proteins known to dimerize upon ligand binding, for instance when binding to DNA or RNA, and membrane associated proteins were not included in the data set. PDB entries containing only fragments of the relevant polypeptide chains were also disregarded. Only those molecules were retained whose three-dimensional structures were determined by X-ray diffraction on single crystals with a resolution of better than 3.0 Å.

Structure and sequence redundancies of entire protein chains were removed from each macromolecular class consulting the FSSP[27] (Fold classification based on Structure-Structure alignment of Proteins) data base. The FSSP family tree was cut at a Z-score of 4 standard deviations above average (FSSP release, January 1999) for pair-wise structural comparisons to define structural families approximately at the fold-topology level. For each multimer category, at most, one representative chain per fold-topology family defined in this way was selected for the final set of macromolecular structures. The resulting PDB entries are listed by their identification codes in

TABLE II. Closely Homologous Monomer/Dimer Pairs*

	Category		Seq. id.		Correct cl.	
Family	mon	dim	all [%]	ifc	mon	dim
Ribonuclease	1AFK	1BSR*	81	31/45	√	√
Galectin (S-lectin)	1BKZ	1SLT	32	2/14	√	−
Cu,Zn superoxide dismutase	1ESO	1XSO	25	4/19	√	−
Hemoglobin	1FLP	3SDH	19	6/26	√	√
Sulphatase/phosphatase	1FSU	1ALK	8	7/100	√	√
Inositol poly-/monophosphatase	1INP	1IMB	13	5/43	√	√
Diphtheria toxin	1MDT	1TOX*	99	106/106	√	√
Aminopeptidase/creatinase	1XGS	1CHM	8	7/87	√	√
Cytochrome c_3	2CY3	1CZJ	25	3/18	√	√

*Homologies across the subsets of monomeric (mon) and homodimeric (dim) proteins. Protein chains with PDB identification codes listed horizontally belong to the same FSSP class defined by a Z-score > 15 standard deviations for pair-wise structural comparisons. The name of the protein class is listed in the leftmost column. Different family names for monomers and dimers are separated by a slash. Sequence identities (seq. id.) are given in percentage of mean chain length for entire chains (all) and as a ratio for the dimer interface (ifc). A ratio of 2/14 means that 2 of the 14 interface residues of the dimer are identical in the homologous monomer. Sequence identities were determined from the FSSP alignment, whereas interface residues were characterized by a loss of more than 1 Å^2 in solvent accessible surface area (ΔASA) upon complexation. Correct (√) and false (−) classifications, respectively, are indicated in the rightmost column (correct cl.). Dimeric structures marked with an asterisk exhibit domain rearrangements relative to the homologous monomer, a phenomenon known as domain-swapping.

Table I. Because monomers and homodimers were treated separately, homologies occurred across these two categories. Twenty-three monomer/dimer pairs belonged to the same topology family characterized by an FSSP Z-score > 4 standard deviations. Of those, Table II outlines the nine clearly related pairs as defined by a Z-score > 15.

Generation of Hypothetical Dimers

All crystal contacts were generated for each PDB entry of the data set applying crystallographic symmetry operations to the deposited atom coordinates. The two protein chains exhibiting the largest contact area were retained for evaluation. This procedure was applied to monomeric and dimeric proteins regardless of the number of chains deposited with the PDB entry, which usually represents the asymmetric unit of the crystal. When the maximum contact size was not unique, an arbitrary choice was made.

In the following, we refer to the generated dimeric structure as the hypothetical dimer of the molecule. Calculations of the symmetry-related atom coordinates were performed using the CCP4 (Collaborative Computational Project, Number 4, 1994) suite of programs and in-house software.[13] Before symmetry operations, the molecule was placed within one unit cell from the origin, to reduce the translational space searched for interchain contacts to $(-2, -1, 0, +1, +2)$ unit cells.

Calculation of Contact Area

The contact area between two protein chains A and B was measured as the solvent accessible surface area[28] (ASA) per isolated chain buried upon formation of the complex AB:, i.e., as ΔASA = [ASA(A) + ASA(B) − ASA(AB)]/2. ASA values were calculated by using the program NACCESS[29] (version 2.1.1), which implements an algorithm by Lee and Richards.[28] The radius of the probe sphere was 1.4 Å, and the z-slice thickness was 0.05 Å. Van-der-Waals radii were taken from Chothia.[30] All hydrogen atoms and atoms described in HETATM lines of the PDB file were suppressed, including oxygen atoms of water molecules. We also calculated the difference in ΔASA between the largest and the second largest interface in the crystal, which is denoted ΔΔASA.

Pair-Frequency Scoring Function

We based our scoring function on atom-pair frequencies observed across hypothetical dimer interfaces in the nonredundant data subset of 76 homodimers. Two atoms of a hypothetical dimer counted as a pair if they belonged to different protein chains and were less than 8.0 Å apart. Atoms were distinguished according to their covalent connectivity by using the SATIS scheme,[31] resulting in 17 different atom types for the 20 common amino acids. Hydrogen atom positions were suppressed in all calculations. Distances were sampled up to 8.0 Å by nonoverlapping bins of 0.5 Å width.

The way we calculated the distance dependent scoring scheme from the observed atom-pair frequencies is adopted from the construction of so-called statistical potentials, often also termed "knowledge-based potentials of mean force" or "pair potentials." We used a form proposed by Sippl,[18] which includes a weighting factor σ to the effect of a gradual suppression of small atom-pair frequencies observed. Let N_{ab} be the number of pairs with atom types a and b. $n_{ab}(r)$ represents the fraction of these pairs where the atoms of type a and b are separated by a distance r. With $n(r)$ denoting the fraction of all pairs with atoms separated by r and setting scaling factors to unity, the scoring function $s_{ab}(r)$ for an atom-pair of types a and b has the form:

$$s_{ab}(r) = \ln(1 + \sigma N_{ab}) - \ln\left(1 + \sigma N_{ab}\,\frac{n_{ab}(r)}{n(r)}\right)$$

σ was set to 0.02 in our calculations.

The log-odds score for a particular monomeric or homodimeric molecule is obtained as the sum of terms $s_{ab}(r)$ for all atom pairs across the interface of the hypothetical dimer, i.e., for all atom pairs at distance < 8.0 Å and involving one atom from each polypeptide chain. In the following, we refer to this score simply as the "pair score."

Estimation of the Generalized Classification Error

A naive way of assessing the classification error of the proposed method implies deriving the scoring function from all homodimers in the data set of Table I and applying it to the entire data set, including monomeric molecules. A score value is then chosen as a cut-off that divides the entire data set into two discrete categories such as to minimize the number of false classifications. We refer to this number of false classifications as the apparent classification error or the apparent discrepancy. Relating this number to the total number of molecules, monomers, and homodimers, yields the apparent misclassification rate or the apparent discrepancy rate. Note that the form of the scoring function $s_{ab}(r)$ is derived solely from homodimeric molecules, whereas both monomers and homodimers influence the choice of the cut-off value.

The problem remains of how to estimate the discrepancy rate obtained if other, new homodimers were included in the derivation of the scoring function, different monomeric structures, or both, were available for the classification problem. In particular, we have no a priori knowledge about the ratio of the number of monomeric versus dimeric molecules naturally occurring. Nevertheless, we would like to estimate how well the model generalizes when presented with new structures. The usual approach to this problem is to artificially decrease the size of the data set in a random manner for the deduction of the model (i.e., the scoring function together with the cut-off value) and to infer from the error observed with the rest of the data the generalized error that would most likely be obtained if data not yet known to us were included in the model.

We chose the bootstrap method[32] for the estimation of the discrepancy rate, which can be regarded as an extension of K-fold adjusted cross-validation with reduced variability.[22] A bootstrap sample of equal size to the original data set was obtained as a random draw with replacement from the pooled monomers and homodimers. This sample constituted the training set. It, thus, always consisted of 172 monomeric or homodimeric PDB entries chosen from Table I, but included multiple copies of some entries. Structures that remained unselected by this procedure constituted the test set. Because the sets were disjunct, homologous proteins could not occur in both the trainings and the test set unless they belonged to the homologous pairs listed in Table II.

The scoring function was derived from the fraction of homodimeric entries in the training set. The number of unique entries in this fraction fluctuated around 64 entries. The best discriminating cut-off value was determined from the scores calculated for the entire training set. This cut-off value was subsequently applied to obtain the discrepancy on the test set. The cycle was repeated 200 times computing at each step the averaged discrepancy[22]:

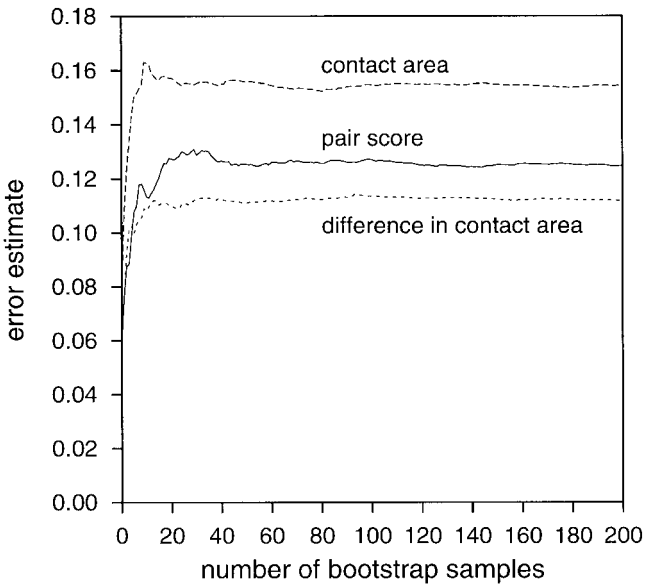

Fig. 1. Bootstrap estimates of the discrepancy for the contact area (ΔASA, dashed line), pair scores (continuous line), and the difference in contact area of the two largest contacts in the crystal (ΔΔASA, dotted line) versus the number of bootstrap samples. Convergence was reached after a number of approximately 50 samples.

$$\epsilon_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{Q_i}{B_i}.$$

Here, N is the size of the bootstrap sample, which is equal to the size of the data set ($N = 172$). B_i is the number of samples not containing protein i, i.e., where protein i was a member of the test set, and where for Q_i of these B_i samples protein i was misclassified. Convergence for ϵ_0 was reached after approximately 50 cycles (Fig. 1). The bootstrap estimate ϵ_{bs} of the classification error is calculated as a weighted average between the downwardly biased apparent error ϵ_{app} obtained from the entire data set without bootstrapping, and the upwardly biased ϵ_0[22]:

$$\epsilon_{bs} = (1 - v)\epsilon_{app} + v\epsilon_0$$

The weighting factor

$$v = \frac{0.632}{1 - 0.368r}$$

depends on the so-called overfitting rate

$$r = \frac{\epsilon_0 - \epsilon_{app}}{\gamma - \epsilon_{app}}$$

where $\gamma = p(1 - q) + (p - 1)q$ depicts the error rate that was obtained if the classification procedure provided no information at all, i.e., if predictors and responses were independent. This rate is calculated from fraction q of entries in the entire data set belonging to one category, for

instance to the monomeric molecules, and from the fraction p of the entire data set that are assigned that class by the classification procedure. The factors 0.632 and 0.368 arise from the fact that on average a fraction $1 - e^{-1}$ of the data set is selected for the bootstrap sample. The resulting bootstrap estimate has been shown to be nearly unbiased in simulations of a variety of classification problems.[22]

Applying this procedure to the ΔASA scores involved simply determining the best discriminating cut-off value from the training set and using it for classification of the test set. In the following, we quote two classification error rates for pair-frequency and ΔASA scores: the apparent discrepancy and the estimated generalized discrepancy obtained from the bootstrap procedure.

RESULTS

Results of the bootstrap analysis are depicted in Figure 1 for the three scores calculated from intermolecular contacts encountered in the protein crystal. The pair score based on atom-pair distances across the interface and the solvent accessible surface area per protein chain hidden in the contact (ΔASA) are determined for the most extensive contact of the crystal. Figure 2 shows their distributions and Figure 3 reveals their correlation explicitly. The inherent structure of the two parameters can be inferred from Figure 4.

Pair and ΔASA scores can be applied to complexes isolated from the crystal context, whereas the third parameter ΔΔASA relates the two most extensive contacts in the crystal. Because of their more general applicability, we focus on the results of ΔASA and pair score, and quote results for ΔΔASA in Table III, which summarizes the absolute numbers of true and false classifications.

Discrepancies of Pair Scores

For pair scores, the best-separating cut-off, i.e., the score value that minimizes the total number of false classifications, was −70.1 for the entire data set (see Fig. 2). With this cut-off, 5 of 96 monomers were misclassified as dimers and 7 of 76 dimers were misclassified as monomers, yielding an apparent error rate of 7.0%. Evaluation of 200 bootstrap samples resulted in a generalized error rate of 12.5% as shown in Figure 1. Here, the average cut-off score was −78 ± 13. The PDB identification codes of the misclassified entries are depicted in Figure 4.

Discrepancies of ΔASA Scores

We found 14 monomers and 9 homodimers classified wrongly by contact area by using an optimum cut-off of 856 Å^2 (Fig. 1). The apparent error rate was thus 13.4%. The bootstrap method estimated the actual discrepancy to 15.4% and the cut-off value to 874 Å^2 ± 107 Å^2. So, the application of our statistical pair-frequency based scoring scheme improved by only 3% on the simple parameter of contact area when classifying proteins that are known to be either monomers or homodimers in solution.

As summarized in Table III and Figure 4, 12 of 14 monomers and 4 of 9 homodimers that were misclassified by contact area were correctly classified by the pair score. On the other hand, the pair score misclassified two monomers and four dimers that were correctly assigned by contact area.

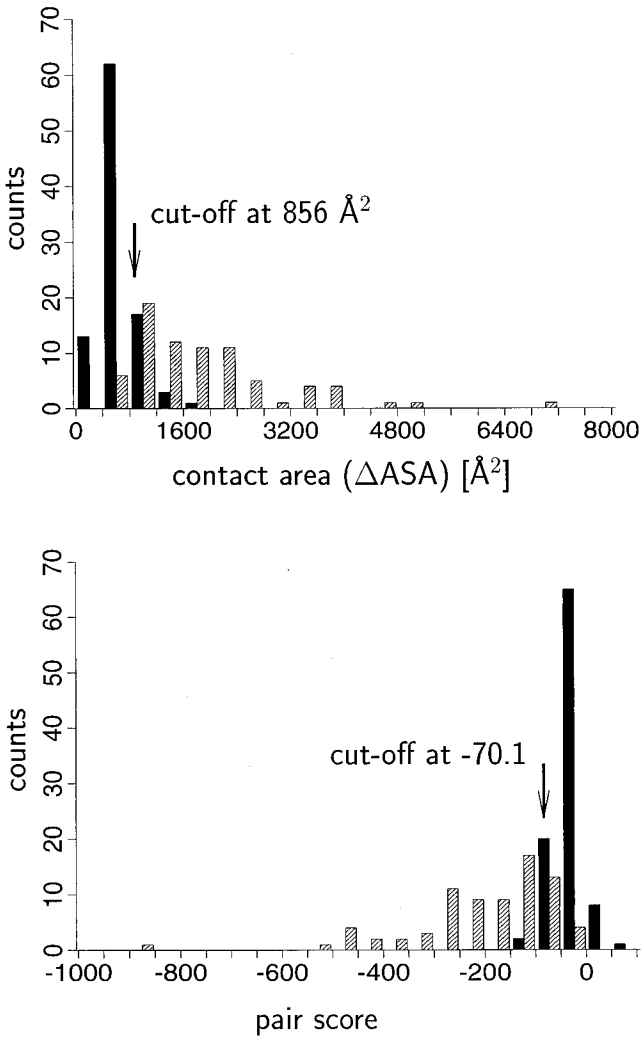

Fig. 2. Histograms of scores based on the distance-dependent statistical potential (**bottom**) and of contact area values (**top**) for the sets of monomeric and homodimeric proteins of Table I. Histograms show the absolute number of counts indicated by filled bars for monomers and hashed bars for homodimers. A bin width of 50 was chosen for the scores and 400 Å^2 for the contact area. Vertical arrows indicate the cut-off values used for classification.

Discrepancies of ΔΔASA

Having completed the comparison of ΔASA and pair scores for the largest intermolecular contact in the crystal, we also looked at parameters that introduce information about the other crystal contacts. We found that the ΔΔASA measure, the difference in ΔASA between the largest and the second largest contact encountered in the crystal, yields improved error rates. Eight dimers and nine monomers are misclassified by using this parameter (Table III) with a cut-off of 373 Å^2, resulting in an apparent error rate

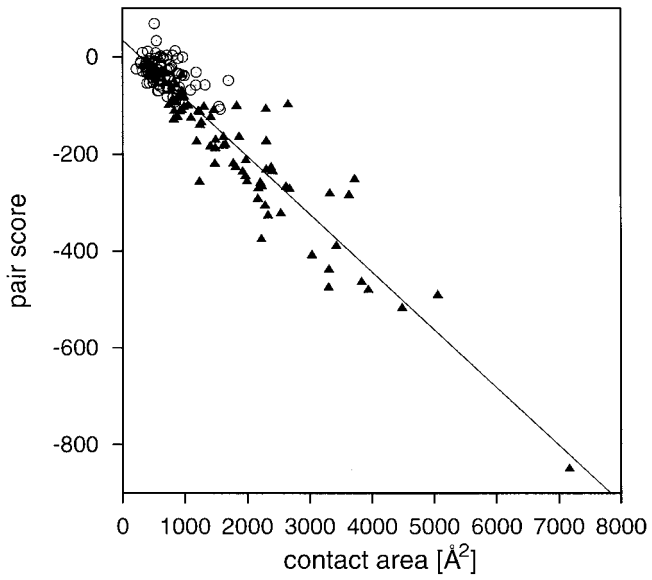

Fig. 3. Correlation between pair scores and contact area (ΔASA) for monomers (empty circles) and homodimers (filled triangles). A correlation coefficient of −0.93 was calculated. The line was obtained by linear regression resulting in the relation $s = -0.119 * a + 34$ between the score s and the contact area a in units of Å².

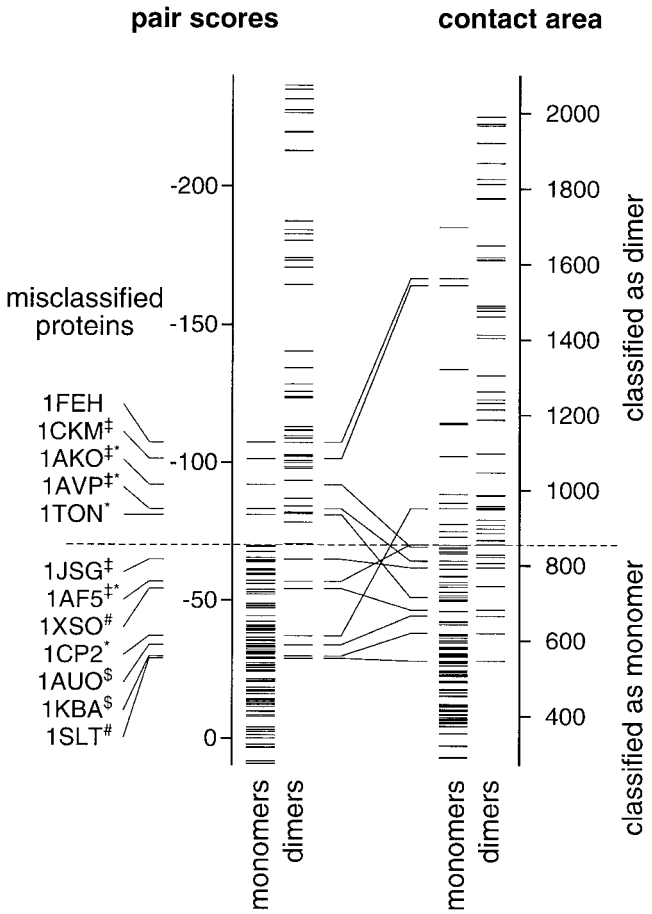

Fig. 4. Chart of the pair (left) and ΔASA (right) scores near the cut-off values. Small horizontal bars indicate the score values and are shown in vertical columns for monomeric and dimeric proteins separately. Cut-off values are depicted by the dashed horizontal line in the middle of the chart. PDB identification codes of proteins misclassified by pair scores are listed. Monomeric proteins misclassified as dimeric are listed above the dashed cut-off line, whereas dimeric proteins misclassified as monomeric are listed below that line. Connecting lines assign PDB codes their score values. Additional signs indicate correct assignments by ΔASA (asterisks), DNA or RNA binding proteins (double daggers), proteins with homologs across the monomer/dimer subsets (pound signs), and proteins belonging to families comprising monomers and homodimers (dollar signs).

of 9.9%. The boostrap method estimates the generalized error rate to 11.1% where the cut-off is 392 Å² ± 53 Å². Results of the bootstrap procedure are shown in Figure 1.

Monomers Classified as Homodimers by Pair Score

The five monomers wrongly assigned the dimeric state by pair score (Fig. 4) are monomeric Fe-only hydrogenase from *Clostridium pasteurianum* (1FEH[33]), mRNA capping enzyme from *Chlorella* virus PBCV-1 (1CKM[34]), *E. coli* exonuclease III (1AKO[35]), human adenovirus protease (1AVP[36]), and serine protease Tonin (1TON[37]) from rat submaxillary gland. The more dimeric ranking 1FEH and 1CKM are misclassified also by contact area, whereas 1AKO, 1AVP, and 1TON are correctly classified by this parameter.

1FEH is known as *CpI,* one of two hydrogenases from *C. pasteurianum* both of which have been characterized as monomeric, cytoplasmic molecules by a variety of methods.[38] There is also no evidence for dimerization for 1AVP, which adopts a different fold than viral proteases that form dimers[39] like cytomegalovirus protease.[40] 1AKO has been proposed to bind to DNA as a monomer.[35] In the hypothetical dimer constructed for our data set, the proposed DNA binding site of the monomer is located at opposite sides of the dimer model such that a single DNA molecule would have to be bent in order to bind to the two binding sites.

For these three of the misclassified monomers, the crystal structure provides no clues as to why they are classified as dimers. The other two monomeric structures, however, may be influenced by the crystal environment.

Sedimentation studies suggest that 1CKM functions as a monomer in solution,[41] although it crystallizes as a dimer in the assymmetric unit. The molecule consists of two domains that adopt different conformations in the two chains of the assymmetric unit. Only one conformation, the closed form, appears to be able to bind manganese ions as cofactors.[34] A divalent cation like magnesium or manganese is required to form the guanylated enzyme intermediate of the guanyl transfer reaction onto mRNA.[41] These are indications that the conformational arrangement of the domains that enables dimer formation is an artifact of the crystal environment.

The crystal structure of 1TON was solved in the presence of Zn^{2+}, which resulted in better diffracting crystals. The ion binds in the catalytic site distorting the native conformation and inhibiting catalytic activity.[37] The position of the Zn^{2+} atom is in the largest intermolecular

TABLE III. Classification Results*

	Monomer			Dimer		
Classfied as	Pair score	ΔASA	ΔΔASA	Pair score	ΔASA	ΔΔASA
Monomer	91	82	87	7	9	8
Dimer	5	14	9	69	67	68

*Overview of the classification results in absolute numbers for pair score, buried contact area ΔASA, and the difference in ΔASA (ΔΔASA).

contact site where it is coordinated by three histidines of one molecule and a glutamate of a symmetry related molecule. Thus, one might attribute dimer formation in this case to the high zinc concentration facilitating crystallization.

Dimers Classified as Monomers by Pair Score

The seven dimers misclassified as monomers by pair score can be split into two groups. 1AF5 and 1CP2 change their ranking order relative to the ΔASA score as depicted in Figure 4 where they were correctly classified as dimers. In contrast 1SLT, 1KBA, 1AUO, 1XSO, and 1JSG are also misclassified by ΔASA with essentially unchanged ranking order.

The nitrogenase iron protein from *Clostridium pasteurianum* (1CP2[42]) exhibits the largest difference in the ranking order between pair score and ΔASA. The molecule has a 4Fe:4S cluster bound in the dimer interface. In the homologous enzyme from *Azobacter vinelandii,*[43] which has almost twice the interface area of 1CP2, dimerization is lost when the 4Fe:4S cluster is removed. Because the presence of the 4Fe:4S cluster is ignored in our method, the pair-frequency score of the hypothetical dimer could well be in accordance with a situation in solution where the 4Fe:4S center is absent from the dimer interface.

1AF5[44] is the structure of I-*Cre*I, a mobile intron endonuclease that was reported to be a homodimer in solution[44] as revealed by dynamic light scattering studies. The molecule is classified as monomeric by pair score and as dimeric according to ΔASA, but the ΔASA value ranks close to the cut-off.

RuvC resolvase (1HJR[45]), a Holliday Junction-specific endonuclease, was described as a dimer by Ariyoshi et al.[45] The asymmetric unit contains a homodimer that has been used to model the complex with the DNA substrate. The prevalent multimeric state in substrate free solution, however, seems not to have been investigated by experiment and remains unclear. Although 1HJR is assigned different categories by both pair score and ΔASA, it ranks close to the cut-offs for both parameters (Fig. 4).

Among the entries with unaltered ranking order between ΔASA and pair score, Bovine spleen galectin-1 (1SLT[46]), which is isolated as a homodimer, and κ-bungarotoxin (1KBA[47]), which is the only dimeric member of otherwise monomeric κ-neurotoxins, are most strongly predicted as monomers. The crystal structure of 1KBA is remarkable in that the interface of the dimer in the asymmetric unit, which is thought to be the physiological relevant one[47] and has a ΔASA of 490.1 Å^2 and a pair score of −34.5, is not the largest contact in the crystal with a ΔASA of 621 Å^2 and a pair score of −29.7. Encouragingly, pair scores, although classifying both interfaces as crystal artifacts, indicate the smaller interface being more dimer-like, i.e., physiologically relevant than the largest interface in the crystal.

Pseudomonas fluorescens carboxylesterase (1AUO[48]) has been characterized as a strong homodimer in solution[49] but is clearly classified as a monomer by both ΔASA and pair score. The subunit-subunit interface contains the active site clefts from both subunits and one catalytic residue takes part in the dimerization interface.[48] This observation is in marked contrast to the homologous cholesterol esterase from *Candida cylindracea,*[50] which is also a homodimer but with the active sites far away from the subunit interaction site. On these grounds, it has been speculated that homodimerization is also not essential for catalytic activity in the case of 1AUO.[48]

Cu, Zn superoxide dismutase from (1XSO[51]) from *Xenopus laevis* comes in three isoforms *aa*, *bb*, and *ab* composed of two different subunits *a* and *b*. 1XSO represents the (*bb*) homodimer. There is evidence that the homodimeric forms are slightly more stable than the heterodimer, which nevertheless may occur in vivo.[52] The dimer interface appears highly conserved among superoxide dismutases as indicated by subunit exchange experiments.[53] Differences in thermostability between *aa* and *bb* types of 1XSO which, like in other superoxide dismutases, is very high[52] have been attributed to amino acid variations in the dimer interfaces.[51] The closely homologous enzyme from *Escherichia coli* is a monomer (see Table II and below).

The oncogene product p14^{TCL1} (1JSG[54]) was included in the set of homodimers because of the notion that 1JSG eluted as a dimer in gel filtration experiments under purification conditions and that the mass spectrum revealed a contaminant peak at twice the molecular weight of the monomer.[54] The authors themselves taken an ambivalent position regarding the multimeric state as they also state[54] "We therefore cannot exclude that p14^{TCL1} forms a dimer in solution". In retrospect, 1JSG may well be in a monomer-dimer equilibrium at physiological concentrations in accordance with the ranges of contact area as well as pair-frequency score.

Summarizing the list of false classifications from Figure 4, the multimeric state of one entry, 1JSG, is not well defined in the literature and might be a monomer-dimer equilibrium. 1CP2 from the set of homodimers has atoms bound in the dimer interface that may be crucial for dimerization behaviour but are not captured by our ap-

proach. Furthermore, there is reason to believe that the dimeric conformations observed with 1TON and 1CKM are artifacts of the crystallization conditions or the crystal environment. Four proteins (1CKM, 1AKO, 1AF5, and 1AVP) bind nucleic acids and four (1AUO, 1KBA, 1SLT, and 1XSO) belong to families comprising monomers as well as dimers. The dimers 1XSO and 1SLT have close homologs in the set of monomers as shown in Table II and discussed below.

Homologous Monomer-Dimer Pairs

There are sequential and structural homologies across the subsets of monomers and homodimers. Table II lists monomeric and dimeric proteins belonging to the same FSSP family defined by a Z-score > 15 standard deviations above average (FSSP,[27] January 1999 release) and some display more than 25% sequence identity.

All monomers and all but two dimers in this set of homologous pairs are classified correctly by using the score suggesting that the method is able to discriminate between different multimeric states of homologous proteins. However, as can be seen from Table II, the residues constituting the interface of the dimer are generally not conserved in the corresponding monomer with the exception of diphteria toxin and ribonucleases. This holds also for galectins (1BKZ, 1SLT) and aminopeptidase/creatinase (1XGS, 1CHM), which are the only structurally homologous pairs where both monomer and homodimer happen to crystallize in the same space group.

Diphtheria toxin forms long-lived metastable dimers upon freezing in phosphate buffer. The dimeric structure (1TOX[55]) revealed a rearrangement of two single-chain domains with respect to the monomeric form (1MDT[56]) to form an intertwined configuration of the two chains. Because one domain of the monomer is replaced by the equivalent domain from the second chain, this was also called a "domain swapped" conformation.[57] A similar phenomenon was also observed for bovine seminal ribonuclease (1BSR[58]) which occurs in two isomers.[59] The swapped N-terminal segments of two chains forming two interchain disulphide bonds resemble the major isomer in solution. Interestingly, homologous ribonuclease A from bovine pancreas (1AFK[60]) does not seem to exhibit domain swapping as it is observed as a monomer in solution.

The other homologous pairs of Table II exhibit the following characteristics. The majority of Cu, Zn superoxide dismutases form dimers in solution. The dimer interfaces are well conserved across species. The monomeric enzyme from *Escherichia coli* (1ESO[61]) is an exception. By structural superposition of the *E. coli* enzyme with the each monomer of the dimeric *Xenopus laevis* molecule (1XSO[51]), Pesce et al.[61] attributed the loss of dimerization of 1ESO largely to the introduction of charges and the loss of hydrogen bonds at positions conserved among dimeric Cu,Zn superoxide dismutases from eukaryotes. A reduced shape and apolar/polar residue complementarity was also observed. Both molecules were classified as monomers by ΔASA (1ESO: 390 Å^2; 1XSO: 682 Å^2) as well as by pair-frequency score (1ESO: -13.8; 1XSO: -54.2) with the ranking essentially unchanged between the two methods.

Human galectin-7 (1BKZ[62]) and bovine spleen galectin-1 (1SLT[46]) are members of carbohydrate recognizing molecules that are involved in mediating cell-cell or cell-matrix interaction. 1SLT is strongly assigned as a monomer but galectin-1 molecules are isolated as homodimers displaying "subunit multivalency" for carbohydrate recognition. However, human galectin-7 is correctly assigned as a monomer, and it does not exhibit multivalency, recognizing carbohydrates as a monomer.

The monomeric sulfide-reactive hemoglobin from *Lucina pectinata* (1FLP[63]) is clearly classified as monomeric by ΔASA (289 Å^2) and pair-frequency score (-13.2). Although the hemoglobin from *Scapharca inaequivalvis* (3SDH[64]) binds its ligand heme-group, which is not considered in our approach, cooperatively as a homodimer, the molecule is classified as a homodimer by pair frequency score (-123.4) and less clear-cut also by ΔASA (886 Å^2).

Human lysosomal sulfatase (1FSU[65]), a monomer, and the dimeric alkaline phosphatase from *E. coli* (1ALK[66]) are assigned correct quaternary categories by either method discussed here. The same holds for *Pyrococcus furiosus* methionine aminopeptidase (1XGS[67]) and the homodimeric *Pseudomonas putida* creatinase (1CHM[68]), which has been crystallized in monoclinic and trigonal crystal forms.

Human inositol monophosphatase (1IMB[69]) is assumed to exist as a dimer in solution in accordance with inositol monophosphatases from other sources. Bovine inositol polyphosphate 1-phosphatase (1INP[70]) is a monomer in solution as well as in the crystal. From the *Desulfovibrio desulfuricans* strain Norway, a monomeric (1CY3[71]) and a homodimeric (1CZJ[72]) cytochrome c_3 have been correctly assigned.

DISCUSSION

The differentiation of functional dimer interfaces from contacts that are artifacts of crystallization is a difficult task, because it is known that features such as hydrophobicity, hydrogen bonding, and amino acid composition vary widely in subunit interfaces and means differ only marginally from crystal contacts.[73] Indeed, the similarity of crystal contacts and functional interfaces has been used to design a score for the ranking of docked protein structures based solely on cystal contacts of monomeric proteins.[74]

Given this similarity, the measure of contact size differentiates remarkably well between monomers and homodimers with an estimated error of only 15.4%. This error is already quite small for "deductive" classification methods. So, it is not surprising that a pair score cannot improve dramatically on this error. The 3% improvement is statistically significant, supported by t-testing 10 error estimates for each score, the ΔASA score, and the pair score, obtained by 10-fold adjusted cross-validation as discribed by Davison and Hinkley[75] (data not shown). The hypothesis of the error of ΔASA being smaller than or equal to the error of the pair score was rejected at a level of 0.01 (a P value of less than 0.2% was obtained). As is apparent from Figure 4, the pair-frequency score improves

over the ΔASA score mainly by recognizing some large interfaces as crystal artifacts.

Remarkably, the ΔΔASA score, which relates the two largest interfaces in the crystal, performs even slightly better than the pair score according to the generalized error, although the apparent error is greater than for the pair score.

The pair score is much more prone to overfitting the data than are the ΔASA and ΔΔASA scores. This is evident from the differences between the apparent errors and the bootstrap estimate. The obvious reason is that the pair-frequency score uses the subset of homodimers twice, in the derivation of the scoring function and in the subsequent determination of the cut-off value, whereas only the latter step is required for classification by ΔASA score.

Despite taking great care in selecting entries from the PDB for our nonredundant data set, it is often difficult to be confident of the assignment of the multimeric state. Different experimental techniques with varying sensitivities are used under a range of different conditions to establish the prevalent multimeric state. Moreover, in the case of those enzymes, where the dimerization has the easily recognizable role of providing the active site, the dimeric state may be well characterized by experiment, whereas less care may have been taken in establishing the multimeric state in cases where dimerization is not obviously linked to function.

The effect of false a priori dimers on the form of the scoring function could, in principle, be minimized by a recursive procedure. Calculating the scoring function at the beginning from all dimers of the data set, one would discard those dimers that rank around the cut-off value for the calculation of the next round. One could hope that the scores converge resulting in a self-consistent scoring function. By using distance information in our approach, the size of the dimeric set from which the scoring function is calculated could be critical, however. A recursive procedure is currently being tested, but given the small data set of incorrectly assigned structures, we do not expect a huge improvement.

For pair ΔASA scores, we considered in each crystal structure only the interface with the highest ΔASA among all interchain contacts. We, thereby, anticipated for the derivation of the pair scoring function that pair and ΔASA scores are highly correlated as can be inferred from Figure 3. Still, for a small fraction of dimers, ΔASA and pair scores may indicate different interaction sites as functional. We found only one dimer (1KBA) in our dimer's data set, for which the largest interface was not the one regarded as the functional one.

In our methods, we disregard all small molecules attached to the protein chain whose atom coordinates are described in the HETATM lines of the PDB files. In particular many enzymes have cofactors bound, which, if the enzyme functions as a dimer, frequently are located at the dimer interface. These molecules can be important for dimerization as it is probably the case of nitrogenase iron protein (1CP2). Therefore, the neglect of such molecules represents a further obstacle for a correct classification of the multimeric state.

One could argue, that the introduction of a distance dependent scoring function diminishes statistics considerably compared with a scoring scheme based on contact pairs within a cut-off sphere. However, the bootstrap estimate of the classification error takes into account any noise that may have been introduced in that way. Also, as more protein structures are determined, the statistics will clearly improve, making the scoring scheme used here more accurate. We have also tested a residue-pair scoring scheme proposed by Moont et al.[15] where it performed best for their set of homodimers in a docking study. The scheme counts residue contacts whenever any side-chain atoms of two residues are within 4.0 Å distance across the intermolecular contact. For our data set, this scheme did not discriminate better than the ΔASA parameter.

Our results for the ΔΔASA parameter suggest that the improvement of the pair score over a contact size measure could be matched by simple contact size-based classifiers, like linear combinations of ΔASA values of the crystal. Distinguishing hydrophobic and hydrophilic surface areas may lead to further improvement here, which we are currently investigating. However, the ΔASA and pair scores are more universal in that they can be used for interfaces isolated from the crystal environment.

It is important to note that we took no cognizance of the energetics of the dimerization process and that we were including in our data set not only very stable dimers but undoubtedly some with marginal stability. Clearly, the multimeric state of some proteins is indeterminate and environment dependent—so that not only the concentration but also temperature, pH, salinity, etc., can all affect the observed status of the protein. Therefore, it is our view that near the cut-off values (of ΔASA or pair score) the proteins are most likely to be in an monomer-dimer equilibrium and their status will depend on the specific environment in the experiment.

We also note that in this study we have attempted to distinguish biological contacts from crystal contacts, although we know that both are physically possible. We argue that discriminating physiologically relevant contacts from crystal contacts is a harder test for the ability of the method to identify protein-protein interaction sites than is the ranking computationally docked structures, which may be completely nonphysical.

Where dimerization is important for the function of a molecule, we can expect that the interface residues may well have evolved to optimize this interaction and may be conserved during evolution. Thus, studying the conservation of interface residues may also be of value in assigning multimeric status where primary sequence data is abundant,[76,77] and could improve the reliability of the method.

CONCLUSION

In this search for criteria that would allow the automatic classification of monomeric and homodimeric proteins from their crystal structures, we compared the performance of an atom-pair potential, which we refer to as pair

score, with a size measure of intermolecular contacts in the crystal. The basis for this comparison was a nonredundant data set of 172 water-soluble proteins whose prevalent quaternary state in solution is known and well defined.

We found that a score, based on atom-pairs across interfaces, including distance information, gives rise to an apparent error of only 7.0%, which increases to 12.5% when cross-validated using a bootstrap procedure. This rate represents a small, but significant improvement over the contact area, measured as the loss of solvent accessible surface area (ΔASA) upon formation of the dimer from the free monomers. The ΔASA score results in an apparent error of 13.4% but is less prone to overfitting as indicated by a bootstrap estimated classification error of only 15.4%. A modified ΔASA score, the difference between the two largest contacts encountered in the crystal, differentiates the multimeric state with a generalized error of only 11.1% is even more reliable than the pair score, although the apparent error rate with 9.9% is larger than that of the pair score. However, the pair score has the advantage of being applicable to protein structures isolated from the crystal environment. Therefore, it would also be suitable for the prediction of dimerization sites of structures determined in solution by nuclear magnetic resonance spectroscopy.

ACKNOWLEDGMENTS

We thank Lorenz Wernisch for useful suggestions regarding the bootstrap. John B. Mitchell kindly provided a computer readable list of the SATIS classification scheme.

REFERENCES

1. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
2. Janin J, Rodier F. Protein-protein interaction at crystal contacts. Proteins 1995;23:580–587.
3. Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol 1997;4:973–974.
4. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101–113.
5. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
6. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
7. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
8. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
9. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
10. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63:31–65.
11. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
12. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
13. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. TIBS 1998;23:358–361.
14. Dasgupta S, Iyer GH, Bryant SH, Lawrence CE, Bell JA. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. Proteins 1997;28:494–514.
15. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
16. Tanaka S, Sheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three dimensional structures of proteins. Macromolecules 1976;9:945–950.
17. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
18. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
19. Torda AE. Perspectives in protein-fold recognition. Curr Opin Struct Biol 1997;7:200–205.
20. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
21. Ben-Naim A. Statistical potentials extracted from protein structures: are these meaningful potentials? J Chem Phys 1997;107:3698–3706.
22. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997;92:548–560.
23. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein databank: a computer-based archival file for macromolecular structure. J Mol Biol 1977;112:535–542.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
25. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res 1999;27:49–54.
26. Etzold T, Argos P. SRS- an indexing and retrieval tool for flat file data libraries. Comput Appl Biosci 1993;9:49–57.
27. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
28. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
29. Hubbard SJ. The analysis of protein-protein recognition. PhD thesis, University of London, London, England, 1992.
30. Chothia C. Hydrophobic bonding and accessible surface area in proteins. Nature 1974;248:338–339.
31. Mitchell JBO, Alex A, Snarey M. SATIS: atom typing from chemical connectivity. J Chem Inf Comput Sci 1999;39:751–757.
32. Efron B, Tibshirani R. An introduction to the Bootstrap. New York: Chapman & Hall; 1993.
33. Peters JW, Lanzilotta WN, Lemon BJ, Seefeldt LC. X-ray crystal structure of the Fe-only hydrogenase (CpI) from *Clostridium pasteurianum* to 1.8 Ångström resolution. Science 1998;282:1853–1858.
34. Håkansson K, Doherty AJ, Shuman S, Wigley DB. X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes. Cell 1997;89:545–553.
35. Mol CD, Kuo CF, Thayer MM, Cunningham RP, Tainer JA. Structure and function of the multifunctional DNA-repair enzyme exonuclease III. Nature 1995;374:381–386.
36. Ding J, McGrath WJ, Sweet RM, Mangel WF. Crystal structure of the human adenovirus proteinase with its 11 amino acid cofactor. EMBO J 1996;15:1778–1783.
37. Fujinaga M, James MN. Rat submaxillary gland serine protease, tonin; structure solution and refinement at 1.8 Å resolution. J Mol Biol 1987;195:373–396.
38. Adams MWW. The structure and mechanism of iron-hydrogenases. Biochim Biophys Acta 1990;1020:115–145.
39. Babé LM, Craik CS. Viral proteases: evolution of diverse structural motifs to optimize function. Cell 1997;91:427–430.
40. Qiu X, Culp JS, DiLella AG, Hellmig B, Hoog SS, Janson CA, Smith WW, Abdel-Meguid SS. Unique fold and active site in cytomegalovirus protease. Nature 1996;383:275–279.
41. Ho CK, van Etten JL, Shuman S. Expression and characterization of an RNA capping enzyme encoded by *Chlorella* virus PBCV-1. J Virol 1996;70:6658–6664.
42. Schlessman JL, Woo D, Joshua-Tor L, Howard JB, Rees DC. Conformational variability in structures of the nitrogenase iron proteins from *Azotobacter vinelandii* and *Clostridium pasteurianum*. J Mol Biol 1998;280:669–685.
43. Anderson GL, Howard JB. Reactions with the oxidized iron protein of *Azotobacter vinelandii* nitrogenase: formation of a 2Fe center. Biochemistry 1984;23:2118–2122.
44. Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL. The structure of I-*Cre*I, a group I intron-encoded homing endonuclease. Nat Struct Biol 1997;4:468–476.

45. Ariyoshi M, Vassylyev DG, Iwasaki H, Nakamura H, Shinagawa H, Morikawa K. Atomic structure of the RuvC resolvase: a Holliday junction-specific endonuclease. Cell 1994;78:1063–1072.
46. Liao DI, Kapadia G, Ahmet H, Vasta GR, Herzberg O. Structure of S-lectin, a developmentally regulated vertebrate β-galactoside-binding protein. Proc Natl Acad Sci USA 1994;91:1428–1432.
47. Dewan JC, Grant GA, Sacchettini JC. Crystal structure of κ-bungarotoxin at 2.3-Å resolution. Biochemistry 1994;33:13147–13154.
48. Kim KK, Song HK, Shin DH, Hwang KY, Choe S, Yoo OJ, Suh SW. Crystal structure of carboxylesterase from *Pseudomonas fluorescens,* an α/β hydrolase with broad substrate specificity. Structure 1997;5:1571–1584.
49. Hong KH, Jang WH, Choi KD, Yoo OJ. Characterization of *Pseudomonas fluorescens* carboxylesterase: cloning and expression of the esterase gene in *Escherichia coli.* Agric Biol Chem 1991;55:2839–2845.
50. Ghosh D, Wawrzak Z, Pletnev VZ, Li NY, Kaiser R, Pangborn W, Jörnvall H, Erman M, Duax WL. Structure of uncomplexed and linoleate-bound *Candida-cylindracea* cholesterol esterase. Structure 1995;3:279–288.
51. Carugo KD, Battistoni A, Carrì MT, Polticelli F, Desideri A, Rotilio G, Coda A, Wilson KS, Bolognesi M. Three-dimensional structure of *Xenopus laevis* Cu,Zn superoxide dismutase *b* determined by X-ray crystallography at 1.5 Å resolution. Acta Crystallogr 1996;D52:176–188.
52. Capo CR, Polticelli F, Calabrese L, Schininà ME, Carrì MT, Rotilio G. The Cu,Zn superoxide-dismutase isoenzymes of *Xenopus laevis:* purification, identification of a heterodimer and differential heat sensitivity. Biochem Biophys Res Commun 1990;173:1186–1193.
53. Tegelström H. Interspecific hybridisation in vitro of superoxide dismutase from various species. Hereditas 1975;81:185–198.
54. Hoh F, Yang YS, Guignard L, Padilla A, Stern MH, Lhoste JM, van Tilbeurgh H. Crystal structure of p14^{TCL1}, an oncogene product involved in T-cell prolymphocytic leukemia, reveals a novel β-barrel topology. Structure 1998;6:147–155.
55. Bell CE, Eisenberg D. Crystal structure of diphteria toxin bound to nicotinamide adenine dinucleotide. Biochemistry 1996;35:1137–1149.
56. Bennett MJ, Eisenberg D. Refined structure of monomeric diphteria toxin at 2.3 Å resolution. Protein Sci 1994;3:1464–1475.
57. Bennett MJ, Choe S, Eisenberg D. Domain swapping: entangling alliances between proteins. Proc Natl Acad Sci USA 1994;91:3127–3131.
58. Mazzarella L, Capasso S, Demasi D, Di Lorenzo G, Mattia CA, Zagari A. Bovine seminal ribonuclease: structure at 1.9 Å resolution. Acta Crystallogr 1993;D49:389–402.
59. Piccoli R, Tamburrini M, Piccialli G, Di Donato A, Parente A, D'Alessio G. The dual-mode quaternary structure of seminal RNase. Proc Natl Acad Sci USA 1992;89:1870–1874.
60. Leonidas DD, Shapiro R, Irons LI, Russo N, Acharya KR. Crystal structures of ribonuclease A complexes with 5′-diphosphoadenosine 3′-phosphate and 5′-diphosphoadenosine 2′-phosphate at 1.7 Å resolution. Biochemistry 1997;36:5578–5588.
61. Pesce A, Capasso C, Battistoni A, Folcarelli S, Rotilio G, Desideri A, Bolognesi M. Unique structural features of the monomeric Cu,Zn superoxide dismutase from *Escherichia coli,* revealed by X-ray crystallography. J Mol Biol 1997;274:408–420.
62. Leonidas DD, Vatzaki EH, Vorum H, Celis JE, Madsen P, Acharya KR. Structural basis for the recognition of carbohydrates by human galectin-7. Biochemistry 1998;37:13930–13940.
63. Rizzi M, Wittenberg JB, Coda A, Fasano M, Ascenzi P, Bolognesi M. Structure of the sulfide-reactive hemoglobin from the clam *Lucina pectinata.* J Mol Biol 1994;244:86–99.
64. Royer WE Jr. High-resolution crystallographic analysis of a co-operative dimeric hemoglobin. J Mol Biol 1994;235:657–681.
65. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, Hopwood JJ, Guss JM. Structure of human lysosomal sulfatase. Structure 1997;5:277–289.
66. Kim EE, Wyckoff HW. Reaction mechanism of alkaline phosphatase based on crystal structures. J Mol Biol 1991;218:449–464.
67. Tahirov TH, Oki H, Tsukihara T, Ogasahara K, Yutani K, Ogata K, Izu Y, Tsunasawa S, Kato I. Crystal structure of methionine aminopeptidase from hyperthermophile, *Pyrococcus furiosus.* J Mol Biol 1998;284:101–124.
68. Coll M, Knof SH, Ohga Y, Messerschmidt A, Huber R, Moellerig H, Rüssmann L, Schumacher G. Enzymatic mechanism of creatine amidinohydrolase as deduced from crystal structures. J Mol Biol 1990;214:597–610.
69. Bone R, Frank L, Springer JP, Pollak SJ, Osborne SA, Atack JR, Knowles MR, McAllister G, Ragan CI, Broughton HB, Baker R, Fletcher SR. Structural analysis of inositol monophophatase complexes with substrates. Biochemistry 1994;33:9460–9467.
70. York JD, Ponder JW, Chen ZW, Mathews FS, Majerus PW. Crystal structure of inositol polyphosphate 1-phosphatase at 2.3 Å resolution. Biochemistry 1994;33:13164–13171.
71. Czjzek FM, Payan F, Guerlesquin F, Bruschi M, Haser R. Crystal structure of cytochrome c_3 from *Desulfovibrio desulfuricans* Norway at 1.7 Å resolution. J Mol Biol 1994;243:653–667.
72. Czjzek M, Guerlesquin F, Bruschi M, Haser R. Crystal structure of a dimeric octaheme cytochrome c_3 (M_r 26000) from *Desulfovibrio desulfuricans* Norway. Structure 1996;4:395–404.
73. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
74. Robert CH, Janin J. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. J Mol Biol 1998;283:1037–1047.
75. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997.
76. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
77. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.