doi:10.1016/j.jmb.2004.11.005 J. Mol. Biol. (2005) 345, 759–771

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

SPI-EM: Towards a Tool For Predicting CATH Superfamilies in 3D-EM Maps

Javier A. Velázquez-Muriel[1,2], Carlos O. S. Sorzano[1] Sjors H. W. Scheres[1,2] and José-María Carazo[1,2]*

[1]*National Center for Biotechnology, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

[2]*Escuela Politecnica Superior Universidad Autonoma de Madrid, 28049 Madrid, Spain*

**Corresponding author*

In this paper the theoretical framework used to build a superfamily probability in electron microscopy (SPI-EM) is presented. SPI-EM is a new tool for determining the homologous superfamily to which a protein domain belongs looking at its three-dimensional electron microscopy map. The homologous superfamily is assigned according to the domain-architecture database CATH. Our method follows a probabilistic approach applied to the results of fitting protein domains into maps of proteins and the computation of local cross-correlation coefficient measures. The method has been tested and its usefulness proven with isolated domains at a resolution of 8 Å and 12 Å. Results obtained with simulated and experimental data at 10 Å suggest that it is also feasible to detect the correct superfamily of the domains when dealing with electron microscopy maps containing multi-domain proteins. The inherent difficulties and limitations that multi-domain proteins impose are discussed. Our procedure is complementary to other techniques existing in the field to detect structural elements in electron microscopy maps like α-helices and β-sheets. Based on the proposed methodology, a database of relevant distributions is being built to serve the community.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: structural biology; three-dimensional electron microscopy; fold recognition; fitting; cross-correlation

Introduction

Three-dimensional electron microscopy (3D-EM) is an emerging technique in structural biology able to obtain a 3D map of the electron density of the specimen under study combining different projection images taken with an electron microscope.[1] One of the advantages of this technique is the wide range of specimen sizes that can be studied: from cellular organelles like the ribosome, to viruses, protein complexes, or individual proteins. Another advantage is that it does not require large amounts of purified specimens, as X-ray diffraction and NMR spectroscopy does, avoiding potential bottleneck in protein expression and purification (a keystone problem in many of the running structural proteomics projects). This makes the technique suitable for studies where the quaternary structure of protein complexes has to be elucidated under a variety of conditions. The resolution that can be obtained in the study depends on the quality, quantity and angular coverage of the projection images, together with the structural homogeneity of the specimen. Low-resolution 3D-EM maps should be expected if the specimen shows a high structural heterogeneity. Generally speaking, it is very difficult to achieve atomic resolution experimentally using EM. Most often, 3D-EM maps are in the resolution range between 6 Å and 30 Å.[1–5] The best resolutions, 6–9 Å, can only be achieved when the specimen is highly symmetrical (like icosahedral viruses[6,7]) or witty very well defined conformations (like the ribosome[8] or GroEL[9,10]). When the single-particle approach is employed, resolutions in the range of 8–12 Å are considered of good quality. At these resolutions it is impossible to trace the backbone and residual groups of the amino acids making up the protein, as done in X-ray crystallography. It is also difficult to see the secondary structure elements, α-helices and β-sheets. The homologous superfamily of the protein domains

Abbreviations used: SPI-EM, superfamily probability in electron microscopy; 3D-EM, three-dimensional electron microscopy; LCCC, local cross correlation coefficients.

E-mail address of the corresponding author: carazo@cnb.uam.es

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.

can be understood as the next level within the protein structural hierarchy. The aim of this paper is to search for it in medium-resolution 3D-EM maps. This allows searching for bigger and more function-defined pieces of information in the 3D-EM maps than secondary structure elements alone, and does not depend on their type. However, if secondary-structure information is available, it can be used in combination with the method presented here. For instance, two recent tools have been developed to find individual secondary structure elements in 3D-EM maps at resolutions better than 8 Å: HELIX-HUNTER[11] and SHEETHUNTER.[12]

The amino acid sequence of the domain can be a valuable source of information as well. There are many computational techniques to deal with the problem of predicting the fold from the amino acid sequence: homology modeling, threading, and *ab initio* methods, but a method to locate the predicted folds into the 3D-EM maps is still needed. Hence, these methods and the one proposed here are complementary, too.

Different manual or computational techniques for finding protein domains into 3D-EM maps have been reported. In the manual approach, the researcher employs visualization programs like O[13] or VMD[14] to fit a high resolution structure, typically obtained by X-ray crystallography, into a medium resolution 3D-EM map. This approach has the disadvantages of being time consuming and subject to the personal appreciation of the researcher. Most computational approaches are based on finding the orientation of the high-resolution structure in the 3D-EM map by maximizing a measure of fitness. The most popular measures of fitness are the cross-correlation coefficient (CCC) (used by COAN,[15] SITUS-COLORES[16] and FOLDHUNTER[11]) and its variants: local cross-correlation coefficient (LCCC) (used by EMfit[17]) and rotational correlation function (used by FRM[18]). All computational techniques tend to fail when only a part of the structure is fitted, because of blurring of surrounding domains that occurs when the resolution is lowered. As the resolution decreases, every atom blurs into a bigger region of space, and this region can be situated inside the space occupied by the domain of interest. The maximum of the CCC can place the high-resolution structure in a wrong position due to the presence of surrounding electron densities interfering with the density at the correct position. This is especially evident when fitting remote homologues to a given map.[11] To decrease this effect, the Laplacian of the electron density is added to the measure of fitness by Chacon *et al.*,[16] thus taking into account surface information for the fitting. The local cross-correlation coefficient minimizes the influence of the surrounding density, because the CCC is only computed in the voxels occupied by the fitted domain. EMfit[17] maximizes LCCC at the cost of increased computation time. The problem with the surrounding electron density can also be treated manually. If the researcher knows in which zone of the map the high-resolution structure is located, the rest of the volume can be masked out.

In both approaches, the researcher assumes that the high-resolution structure that he or she is fitting is actually present in the 3D-EM map and in the same structural conformation. Here, a first step towards eliminating these assumptions is presented. Because a given domain may not be present, or appear in a slightly different conformation, the aim is not to find that particular domain, but to determine with a high probability to which homologous superfamily (if any) the domain belongs. Superfamilies are defined by domain databases like CATH,[19] SCOP[20] or FSSP.[21] Here the classification given by CATH is used. A homologous superfamily, or superfamily for short, is defined by CATH as the group of protein fold realizations (domains) sharing a common ancestor. Two domains belong to the same superfamily if 60% of the larger structure is similar to the smaller one.

In the following, the SPI-EM methodology is presented. It is based on fitting many high-resolution domains into the 3D-EM map of the specimen and the application of a probability framework to determine its superfamily. Its ability to deal with 3D-EM maps containing isolated protein domains is proven. The influence of the resolution in the confidence of the results, and the possibility of not being able to distinguish two superfamilies at the working resolution are explored. Finally, the methodology is applied to simulated and experimental multi-domain maps demonstrating its usefulness. The inherent difficulties observed are discussed.

Theory

Related work

In the field of sequence homology search, there exists a robust statistical theory[22] to compute p-values, the probability of having by chance a given score or better when aligning two amino acid sequences. The p-value is calculated by relating the observed alignment score, s, to the expected distribution of the alignment scores obtained with random sequences of the same length and composition as the query. It can be proven[22] that the limit distribution for maximal segment scores depends on two parameters K and λ, and is defined by equation (1):

$$\mathrm{Prob}(S > s) = 1 - \exp(-Ke^{-\lambda s}) \qquad (1)$$

where K and λ are computed from the set of scores obtained from comparisons of random sequences. The p-value indicates if the level of homology between two sequences is statistically significant (the lower the p-value, the more significant it is).

It is possible to combine the p-values of various sequence alignments to estimate the probability of a sequence being a member of a CATH protein superfamily. In the work of Bailey *et al.*,[23] the

product of the p-values of various independent sequence alignments is used, and the probability of observing a value smaller than or equal to x for the product of p-values is given by the distribution:

$$x = \prod_{i=1}^{n} p_i$$
$$\text{Prob}(X \leq x) = x \sum_{i=0}^{n-1} \frac{(-\ln x)^i}{i!} \tag{2}$$

Equation (2) is valid for any product of independent values, not necessarily p-values. In later work,[24] the case of correlated p-values is treated, which is important when the p-values combined in the product belong to alignments of proteins of the same superfamily, because they are correlated.

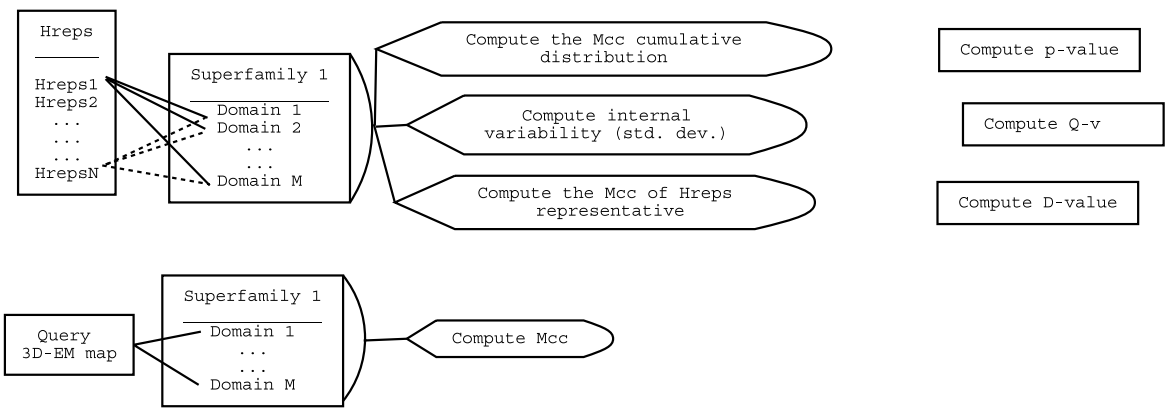
SPI-EM probabilistic approach to superfamily assignment

We follow a conceptually similar approach to determine the superfamily to which a given 3D-EM map belongs: first, the local cross correlation coefficients (LCCCs, this measure is formally introduced below) of fitting the domains of a CATH superfamily into the 3D-EM map are combined, as Bailey *et al.*[23,24] combine the p-values. Note that, due to the fact that the fitting is done *via* LCCCs (−1,1), higher values indicate better fits. This is in opposition to the sequence homology search, where low p-values correspond to correct alignments.

Second, we compare M_{CC}, the arithmetic mean of the LCCCs, with a background distribution of arithmetic means for the CATH superfamily that we are testing. Ideally, this background distribution would be determined by fitting all superfamily domains against a set of all existing domain structures. Since the total number of possible folds in nature is not known, this is not possible. Instead, we use a set of all H-level representatives (Hreps) given by CATH v2.4 to generate the background distribution. It is obtained by fitting each member of the superfamily against all Hreps, and computing the mean of LCCCs for each Hrep. All proteins in the same H-level (homologous superfamily level) share the same overall shape, connectivity of the secondary structures, and a common ancestor. The Hreps set contains a representative domain of every CATH superfamily, forming a complete and non-redundant set covering all known feasible cases in the space of protein structures. Therefore, it is a reasonable set for computing the background distribution.

The difficulties to transfer the statistical methodology employed with protein sequences to 3D-EM are not only related with the election of the appropriate set. The LCCC values of fitting an individual superfamily domain against Hreps do not follow the distribution given by equation (1) for the case of sequence data (data not shown). Neither does the distribution of the product of LCCC values for each Hreps representative follow the formula given in equation (2). This is because equation (2) is derived assuming that each of the terms (in that case the p-values) entering into the product are independent, while this is not the case with LCCC values for different members of a superfamily. We have tried to adjust the distribution of the product of LCCCs to the model for non-independent results proposed in Bailey *et al.*,[24] but without success. Instead, here we show that the arithmetic mean of LCCCs, and the cumulative background distribution of the means can be employed in the classification task.

Finally, the probability that the mean M_{CC} value is significant is computed by reporting the fraction of means in the background distribution with a poorer value. This fraction is our P-value. It is computed directly as the corresponding value for M_{CC} in the cumulative background distribution. We interpret this P-value as the probability that the



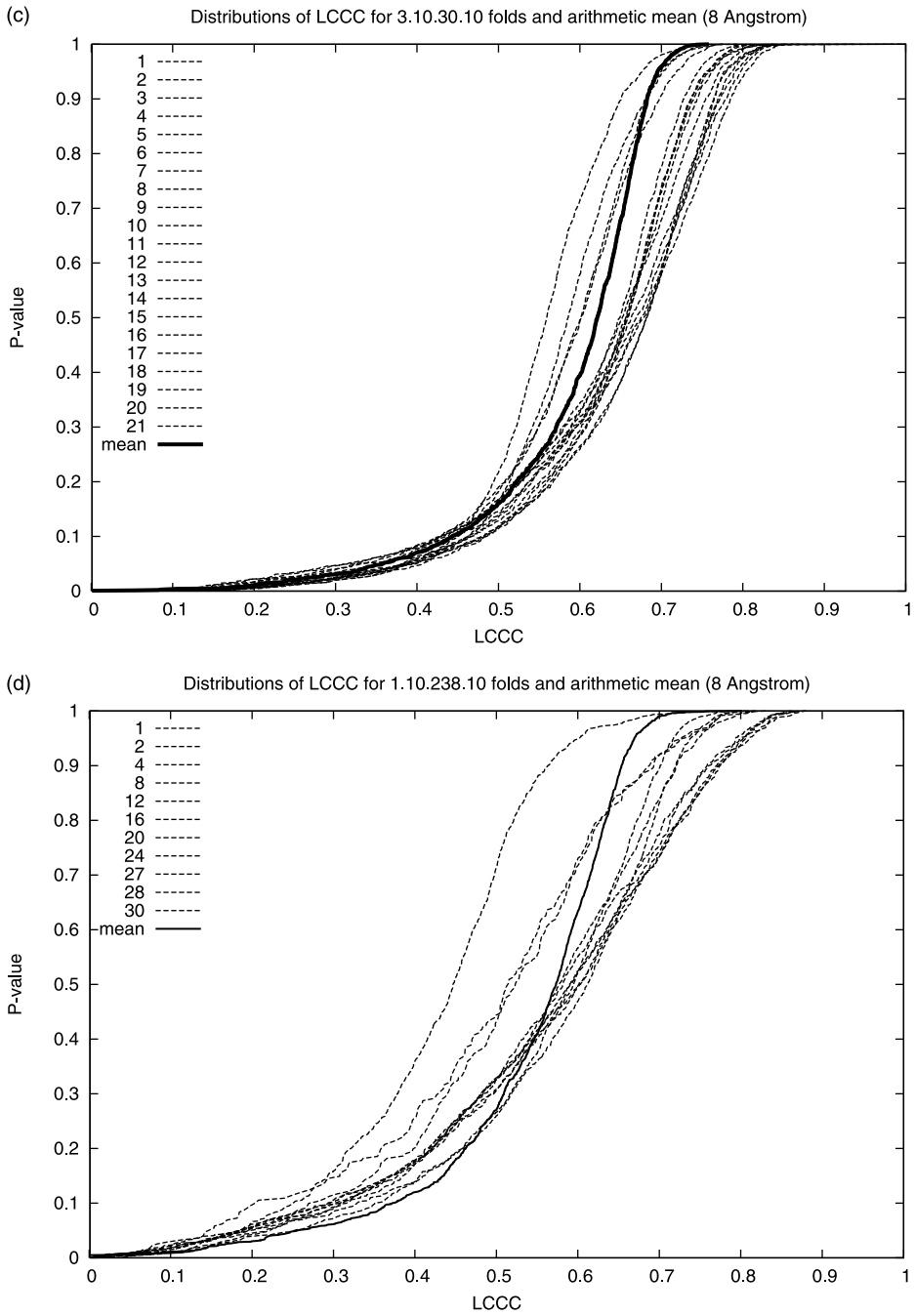


Figure 3. Cumulative distributions of the mean of LCCC for different CATH superfamilies at 8 Å resolution. LCCC distributions for some individual domains of each superfamily fitted into Hreps are also represented. (a) 1.20.1060.10; (b) 2.30.100.10; (c) 3.10.30.10; (d) 1.10.238.10.

imply that the underlying atomic structures are indeed similar. If the study is repeated at a resolution of 8 Å, the two superfamilies are clearly separated (data not shown).

To show that superfamilies from very different classes can also aggregate, a similar experiment was performed with superfamilies 1.20.1060.10 (mainly alpha) and 2.60.11.10 (mainly beta). At 12 Å these families are aggregated, with P-value > 0.99, D-value > -0.1 and Q-value < 0.2, but at a resolution of 8 Å they can be separated (Q-value between 2 and 14). Figure 4(b) shows an example of a domain from superfamily 1.20.1060.10 fitted into a member of superfamily 2.60.120.40 at 12 Å. Although the backbone structure of the two domains is very different, a high LCCC value of 0.879 was obtained. This is explained by fact that many of the voxels of one domain are contained into the other.

A third example illustrates that the aggregation effect may be asymmetric. At 12 Å 2.60.120.40 domains fit well into superfamily 1.20.1060.10 (P-value > 0.95, D-value > -0.09, Q-value < 0.8), but 1.20.1060.10 domains do not fit into 2.60.120.40

Table 1. Detection of the superfamily for 28 domains at 8 Å

Position	1					2					3					4				
Domain	superfamily	mean	P	D	Q	superfamily	mean	P	D	Q	superfamily	mean	P	D	Q	superfamily	mean	P	D	Q
1AXN03	1.10.220.10	0.91	1.000	-0.02	0.42	1.10.1250.10	0.78	0.998	0.03	0.48	2.30.30.140	0.82	0.994	-0.04	1.13	1.10.150.30	0.84	0.993	-0.09	3.00
1AYM40	4.10.81.10	0.59	0.982	-0.40	153.32	1.20.5.300	0.59	0.971	-0.38	31.99	4.10.170.10	0.63	0.971	-0.30	11.67	4.10.80.10	0.46	0.960	-0.23	0.75
1B2PA0	2.90.10.10	0.91	1.000	-0.05	1.60	2.10.150.10	0.72	0.995	-0.26	4.89	2.60.40.550	0.66	0.990	-0.28	5.29	1.10.540.10	0.74	0.983	-0.19	6.80
1BD800	1.25.40.20	0.81	1.000	0.08	0.77	1.10.286.10	0.67	0.999	-0.32	46.47	1.20.120.80	0.73	0.999	-0.24	5.04	1.20.140.10	0.76	0.999	-0.21	15.64
1C9SA0	2.60.40.50	0.98	1.000	0.01	0.40	3.30.46.10	0.71	0.994	-0.21	2.05	3.90.20.10	0.54	0.987	0.05	0.33	4.10.170.10	0.64	0.985	-0.28	10.95
1CZFA0	2.160.20.10	0.78	1.000	0.03	0.26	3.40.367.20	0.68	0.998	-0.24	3.67	3.40.670.10	0.70	0.998	-0.29	19.06	1.10.560.10	0.69	0.997	-0.28	8.37
1D0BA0	3.80.20.10	0.67	1.000	0.01	0.04	1.20.120.80	0.73	0.999	-0.24	5.09	1.20.200.10	0.68	0.999	-0.09	0.47	3.40.367.20	0.71	0.999	-0.21	3.17
1D6ZA3	1.25.30.10	0.41	0.982	-0.34	2.32	1.10.640.10	0.53	0.982	-0.11	0.62	1.25.30.10	0.41	0.982	-0.34	2.32	3.90.269.10	0.61	0.976	-0.33	17.07
1DIM00	2.120.10.10	0.76	0.999	-0.16	1.66	3.90.550.30	0.68	0.998	0.12	1.21	2.40.180.10	0.52	0.997	-0.24	1.06	3.40.800.10	0.76	0.997	-0.19	4.98
1EG9B0	1.10.1030.10	0.61	0.980	-0.37	41.01	3.40.50.940	0.61	0.959	-0.35	3.76	1.10.599.10	0.59	0.922	-0.33	8.64	3.40.1050.10	0.59	0.902	-0.15	0.91
1EXTA1	4.10.280.10	0.62	0.998	-0.03	0.25	1.20.15.40	0.71	0.992	-0.25	5.49	2.10.90.10	0.63	0.992	0.08	0.81	1.20.89.10	0.67	0.988	-0.31	14.11
1EY0A0	2.40.50.90	0.94	1.000	0.02	0.24	1.10.1130.10	0.75	0.999	-0.17	3.83	3.40.50.9100	0.81	0.991	-0.18	6.13	3.90.440.10	0.72	0.985	-0.26	20.57
1EYNA2	3.40.50.1010	0.69	1.000	0.00	0.03	3.65.10.10	0.92	1.000	-0.01	0.21	3.30.9.10	0.71	0.999	-0.04	0.30	3.40.50.880	0.77	0.999	-0.02	0.23
1FQTA0	3.30.465.20	0.81	0.994	-0.18	76.98	3.90.440.10	0.74	0.994	-0.24	18.92	3.50.11.20	0.75	0.992	-0.21	2.64	2.102.10.10	0.77	0.990	-0.03	0.33
1GEN00	2.110.10.10	0.88	1.000	0.00	0.06	3.40.50.700	0.77	0.998	-0.21	8.48	3.40.50.1270	0.75	0.998	-0.14	0.78	1.10.443.10	0.69	0.997	0.04	0.24
1GPB02	3.40.670.10	0.98	1.000	0.00	0.19	3.40.367.20	0.70	0.999	-0.22	3.40	2.160.20.10	0.67	0.998	-0.07	0.56	3.30.590.10	0.68	0.998	-0.28	8.49
1H70A0	3.40.50.200	0.80	0.999	-0.12	2.39	3.20.19.10	0.77	0.998	-0.23	112.70	3.40.630.20	0.78	0.998	-0.16	4.86	3.40.50.970	0.71	0.997	-0.02	0.14
1HJ5A2	2.140.10.20	0.99	1.000	0.00	0.08	2.130.10.20	0.71	0.999	-0.07	0.45	1.10.630.10	0.67	0.998	-0.20	1.71	2.140.10.10	0.68	0.998	-0.27	4.52
1I0HB0	3.90.149.10	0.91	1.000	0.00	0.04	2.60.120.10	0.64	0.999	-0.18	1.82	3.30.559.10	0.68	0.999	-0.01	0.05	3.30.565.20	0.76	0.999	-0.20	10.06
1I8FA0	2.30.30.140	0.94	1.000	0.08	2.13	3.30.210.10	0.81	0.980	-0.12	4.95	1.10.1200.20	0.79	0.960	-0.15	3.37	4.10.109.10	0.83	0.945	-0.15	3.96
1IJQA0	3.90.550.30	0.71	1.000	0.14	1.51	1.10.580.10	0.70	0.999	-0.26	2.75	2.70.98.10	0.73	0.999	-0.26	56.51	2.130.10.20	0.70	0.999	-0.08	0.48
1JGMA0	3.20.20.140	0.76	1.000	-0.02	0.17	3.75.10.10	0.74	0.998	-0.23	3.43	3.40.800.10	0.75	0.996	-0.21	5.36	2.60.130.10	0.65	0.995	-0.11	0.49
1KID00	3.50.7.10	0.87	1.000	0.18	1.52	2.60.90.10	0.76	0.998	-0.16	4.83	3.30.530.10	0.65	0.990	-0.29	6.84	3.40.47.10	0.63	0.988	-0.04	0.40
1QH5A0	2.60.120.260	0.73	0.999	0.03	0.54	3.40.50.1240	0.70	0.999	-0.07	0.33	3.40.91.20	0.73	0.999	-0.08	0.58	3.40.225.10	0.77	0.999	-0.23	37.29
1QSJA0	1.50.10.40	0.78	1.000	0.00	0.00	2.60.130.10	0.67	0.999	-0.09	0.40	3.20.19.10	0.77	0.998	-0.23	115.26	3.20.20.280	0.74	0.998	-0.05	0.38
4BDPA2	1.20.1060.10	0.87	1.000	0.09	0.82	2.60.11.10	0.71	0.999	-0.26	14.49	4.10.280.10	0.58	0.994	-0.07	0.52	4.10.49.10	0.59	0.990	-0.40	108.75
7FD1A0	1.10.472.10	0.81	1.000	0.02	0.16	1.10.1250.10	0.83	1.000	0.08	1.11	3.30.70.20	0.90	1.000	0.18	1.47	1.10.150.40	0.86	0.999	-0.11	19.22
9ATCB2	2.20.30.60	0.97	1.000	0.01	1.03	3.30.210.10	0.82	0.987	-0.12	4.63	3.10.20.10	0.83	0.984	-0.08	0.94	4.10.310.10	0.79	0.982	-0.04	0.41

Each domain has been tested against 477 superfamilies, but only the four first results are shown, ordered by *P*-value. The shadowed results are the correct superfamilies for each domain. Where no superfamily has been shadowed for a given domain, the correct superfamily was not among the four first results.

Table 2. Domain fitting: domains of 1.20.1060.10 fitted into 1.10.530.40 at 12 Å; domains of 1.10.530.40 fitted into 1.20.1060.10 at 12 Å

Domain	Superfamily	M_{CC}	P	D	Q	Domain	Superfamily	M_{CC}	P	D	Q
3BDPA2	1.10.530.40	0.836	0.964	-0.078	0.436	150LD0	1.20.1060.10	0.831	0.996	-0.046	0.379
2BDPA2	1.10.530.40	0.831	0.955	-0.084	0.464	137LB0	1.20.1060.10	0.825	0.993	-0.053	0.432
4BDPA2	1.10.530.40	0.830	0.954	-0.084	0.468	150LC0	1.20.1060.10	0.823	0.992	-0.054	0.445
1XWL02	1.10.530.40	0.828	0.948	-0.087	0.484	169LE0	1.20.1060.10	0.823	0.992	-0.055	0.450
1KRPA2	1.10.530.40	0.818	0.920	-0.097	0.538	116LC0	1.20.1060.10	0.821	0.992	-0.056	0.461
2KZMA2	1.10.530.40	0.816	0.914	-0.099	0.548	216LB0	1.20.1060.10	0.821	0.992	-0.056	0.461
5KTQA2	1.10.530.40	0.816	0.913	-0.099	0.549	173L00	1.20.1060.10	0.821	0.991	-0.057	0.463
1KSPA2	1.10.530.40	0.815	0.913	-0.099	0.552	116LB0	1.20.1060.10	0.820	0.991	-0.057	0.470
1QSLA2	1.10.530.40	0.815	0.911	-0.100	0.557	1QUHA0	1.20.1060.10	0.820	0.991	-0.058	0.472
1KFSA2	1.10.530.40	0.814	0.910	-0.101	0.560	217L00	1.20.1060.10	0.820	0.991	-0.058	0.474
1BDP02	1.10.530.40	0.813	0.906	-0.102	0.567	1G0QA0	1.20.1060.10	0.818	0.991	-0.059	0.484
1D9FA2	1.10.530.40	0.811	0.899	-0.103	0.575	168LD0	1.20.1060.10	0.817	0.990	-0.060	0.493
2KFNA2	1.10.530.40	0.811	0.899	-0.103	0.575	216LA0	1.20.1060.10	0.817	0.990	-0.061	0.496
2KFZA2	1.10.530.40	0.811	0.898	-0.104	0.576	189L00	1.20.1060.10	0.816	0.990	-0.061	0.500
4KTQA2	1.10.530.40	0.811	0.894	-0.104	0.579	1L97A0	1.20.1060.10	0.815	0.989	-0.063	0.512
2KZZA2	1.10.530.40	0.808	0.884	-0.106	0.591	1TLA00	1.20.1060.10	0.814	0.988	-0.064	0.520
2KTQA2	1.10.530.40	0.808	0.883	-0.107	0.592	1L1000	1.20.1060.10	0.814	0.987	-0.064	0.523
1KTQ02	1.10.530.40	0.807	0.880	-0.107	0.597	116LA0	1.20.1060.10	0.813	0.987	-0.064	0.526
1D9DA2	1.10.530.40	0.807	0.880	-0.107	0.597	1QUOA0	1.20.1060.10	0.813	0.987	-0.065	0.528
1D8YA2	1.10.530.40	0.805	0.875	-0.110	0.609	1QT3A0	1.20.1060.10	0.813	0.987	-0.065	0.531
1QSSA2	1.10.530.40	0.801	0.866	-0.113	0.630	1DYF00	1.20.1060.10	0.812	0.986	-0.065	0.534
3KTQA2	1.10.530.40	0.800	0.860	-0.114	0.636	169LD0	1.20.1060.10	0.812	0.986	-0.065	0.534
1QSYA2	1.10.530.40	0.800	0.859	-0.115	0.637	246L00	1.20.1060.10	0.812	0.986	-0.065	0.534
1QTMA2	1.10.530.40	0.800	0.856	-0.115	0.640	1L0500	1.20.1060.10	0.812	0.986	-0.065	0.534
1JXE02	1.10.530.40	0.796	0.833	-0.119	0.662	168LE0	1.20.1060.10	0.812	0.986	-0.066	0.537
1TAUA4	1.10.530.40	0.790	0.815	-0.125	0.694	242L00	1.20.1060.10	0.812	0.986	-0.066	0.540
1TAQ04	1.10.530.40	0.780	0.769	-0.135	0.749	107L00	1.20.1060.10	0.812	0.986	-0.066	0.540
1CMWA4	1.10.530.40	0.758	0.645	-0.157	0.872	1NHB00	1.20.1060.10	0.812	0.986	-0.066	0.540
1BGXT4	1.10.530.40	0.748	0.582	-0.167	0.926	1L1800	1.20.1060.10	0.811	0.986	-0.066	0.542
1T7PA2	1.10.530.40	0.582	0.102	-0.333	1.849	1L3300	1.20.1060.10	0.811	0.986	-0.066	0.542

Only the best 30 are shown. Shadowed results indicate than the domain meets the recognition criteria (P-value>0.90, D-value>-0.1 and Q-value<1).

(P-value<0.86, D-value>-0.1, Q-value<3). This can be explained by the fact that the electron density of 2.60.120.40 is contained in that of 1.20.1060.10. At 8 Å, the two superfamilies are separated (P-value<0.86, D-value>-0.2 and Q-value between 1.8 and 6).

Application to multi-domain maps with SITUS-COLORES

The statistical approach introduced in this paper can be extended to multi-domain 3D-EM maps, provided that individual domains are correctly located within the multi-domain map. We used the program COLORES[16] integrated in the SITUS suite for this task. Since the LCCC is used as a similarity measure in our methodology, surrounding voxels of other domains should not interfere in the CCC computation. Therefore, our methodology, which was developed for single-domain 3D-EM maps, is still valid for detection of superfamilies within multi-domain 3D-EM maps. As an example, Figure 5 shows the result of the correct fitting of domain 4BDPA2 (filled blue) into a simulated 8 Å map of 4BDP (crystal structure of *Bacillus stearothermophilus* DNA polymerase I, fragment 297–876[25]) using SITUS-COLORES. The results of superfamily detection for this multi-domain structure at 8 Å are shown in Table 3. This structure comprises four different domains: 4BDPA1 (amino acid residues 297–490, superfamily 3.30.420.10, mixed alpha–beta), 4BDPA2 (amino acid residues 491–605, 1.20.1060.10, mainly alpha), 4BDPA3 (amino acid residues 606–656 and 801–875, 3.30.70.370, mixed alpha–beta), and 4BDPA4 (amino acid residues 657–800, 1.10.473.10, mainly alpha). We only show the results corresponding to the superfamily assignment of domain 4BDPA2, since this isolated domain was previously studied (see Table 1). Also as part of a larger structure, the correct superfamily for this domain can still be detected based on P-value ranking (see Table 3). However, in this case the P-value is 0.833, while when considered as an isolated domain it was 1.00. The D and Q-values deteriorate in a similar way. We explain this deterioration by the blurring effect caused by surrounding domains: as the resolution decreases every atom blurs into a wider space, affecting the region of the studied domain. This imposes limits on the methodology, which are inherent to all fitting methods. Consequently, our criteria to reject superfamilies adopted for the isolated domains case should be relaxed.

This experiment further illustrates the advantage of the SPI-EM approach over only taking M_{CC}, the LCCC mean, as shown in the results of Table 3. If only M_{CC} were considered, an incorrect superfamily would have been selected, while the P-values of statistical approach readily distinguish the correct solution.

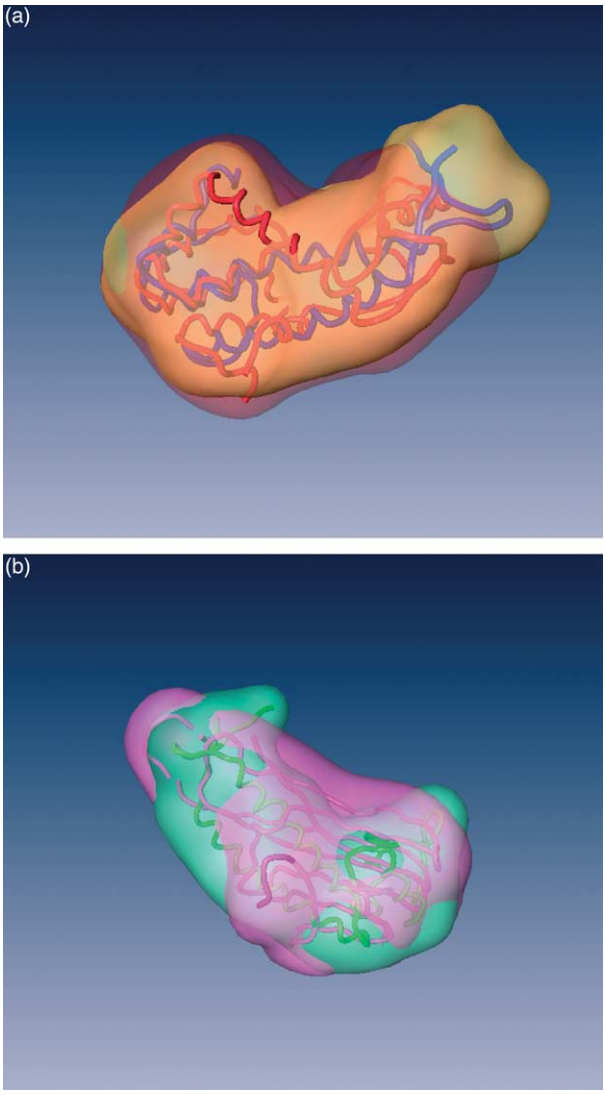


Figure 4. (a) Fitting of 2KZZA2 (1.20.1060.10, mainly-alpha, yellow map and blue skeleton) into 150LD0 (1.10.530.10, mainly-alpha, red map and skeleton) at 12 Å. (b) Fitting of domain 3BDPA2 (1.20.1060.10, mainly-alpha, green) into 4TSVA0 (2.60.120.40, mainly-beta, purple) at 12 Å.

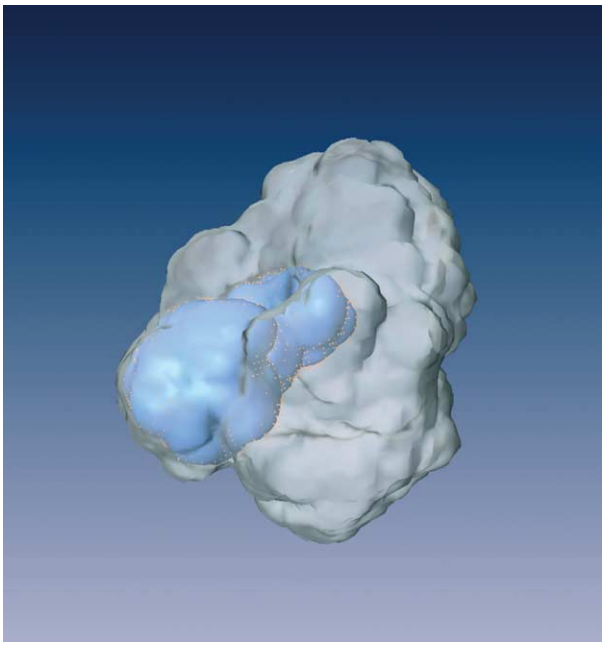


Figure 5. Domain CATH 4BDPA2 (filled blue) fitted into 4BDP (light blue) at 8 Å by SITUS-COLORES. The small yellow spheres represent the voxels used to compute the LCCC (voxels located into the inner part of 4BDPA2 not shown).

Application to experimental multi-domain maps

We studied the 10.3 Å EM-structure of unliganded, *Escherichia coli* GroEL chaperonin (EMD database[26] code: EMD-1042) to illustrate the applicability of SPI-EM to experimental 3D-EM maps. GroEL contains three different domains: GroEL domain 1, (1.10.560.10), GroEL domain 2 (3.30.260.10) and the apical domain (3.50.7.10), which are repeated 14 times throughout the structure. The results of their superfamily detection are presented in Table 4. Superfamily 1.10.560.10 was readily detected, and superfamily 3.50.7.10 emerged as the second candidate, despite the fact that its *P*-value is relatively low. Superfamily 3.30.260.10 was not detected. To compare these results with those obtained if the density map were of still better quality, the experiment was repeated with a 10 Å simulated EM-map, which was based on the corresponding PDB structure (1GR5) for EMD-1042 (Table 5). Now superfamilies 1.10.560.10 and 3.50.7.10 are both correctly detected with *P*-values of 0.962, and 0.847, respectively. The

Table 3. Superfamily detection for domain 4BDPA2 in PDB 4BDP file at 8 Å

PDB code	Superfamily	M_{CC}	P-value	D-value	Q-value
4BDP	1.20.1060.10	0.508	0.833	-0.362	3.412
4BDP	1.10.220.10	0.623	0.515	-0.275	2.399
4BDP	1.10.530.40	0.504	0.408	-0.366	1.488
4BDP	1.10.1250.10	0.556	0.368	-0.229	2.142
4BDP	1.10.238.10	0.514	0.320	-0.288	1.834
4BDP	2.60.11.10	0.408	0.195	-0.567	31.194
4BDP	2.60.120.40	0.509	0.112	-0.401	6.026

The shadowed result is the correct superfamily.

Table 4. Superfamily detection for domains of experimental 3D-EM map EMD-1042 (GroEL chaperonin, unliganded from *E. coli*)

3D-EM map	Superfamily	M_{CC}	P-value	D-value	Q-value
EMD-1042	1.10.560.10	0.642	0.663	-0.285	1.987
EMD-1042	3.50.7.10	0.584	0.206	-0.231	2.298
EMD-1042	1.10.220.10	0.521	0.134	-0.417	4.858
EMD-1042	3.30.260.10	0.553	0.118	-0.413	6.077
EMD-1042	1.10.1250.10	0.495	0.108	-0.374	3.639
EMD-1042	1.20.1060.10	0.448	0.105	-0.409	4.715

Resolution, 10.3 Å. Shadowed results are the three superfamilies present in the map.

P-value for 3.50.7.10 is similar to the one observed for the correct superfamily in the 8 Å experiment of the previous section (Table 3). For superfamily 3.30.260.10 a *P*-value of 0.431 was obtained, which is relatively low compared to the values for the other two superfamilies. This indicates that this superfamily is more difficult to detect, and suggests that this fold lies at the limits of the method ability.

As in the simulated multi-domain case, the superfamily rejection criteria based on *P*, *D* and *Q*-values need to be relaxed when the method is applied to the experimental multi-domain case. In this case, the noise present in the EM micrographs and the reconstruction procedure, as well as a potentially lower signal power at frequencies close to 10 Å, form additional sources of error lowering the LCCC values.

Discussion

Here a new statistical approach for recognizing the superfamily of a given 3D-EM domain map is presented. The method fits all superfamily members into the 3D-EM map and computes the arithmetic mean of the LCCC values (M_{CC}) within each superfamily. The fraction of lower mean values in the background distribution for the superfamily (built fitting the Hreps from CATH against all members of the superfamily) gives the significance of M_{CC}, the *P*-value. This *P*-value is interpreted as the probability that the domain present in the 3D-EM map belongs to the CATH superfamily tested. The *P*-value is complemented computing the difference *D*-value between the M_{CC} obtained with the query volume and the M_{CC} obtained with the superfamily representative in Hreps. The difference is compared with the internal superfamily variability through the *Q*-value. Detection of the most likely superfamily is based on a combined analysis of these three measures.

This approach has been shown to work with simulated 3D-EM maps of isolated domains. Analysis of LCCC distributions showed that the statistical approach of SPI-EM has clear advantages over the fitting of single domains in low-resolution electron density maps. In a general applicability test comprising 28 highly different domain structures, the correct superfamily (among 477) could be readily detected in 80% of the cases.

As the resolution of the EM map decreases, different superfamilies become indistinguishable. Since this effect, which we termed superfamily aggregation, imposes a fundamental restriction to the methodology, we studied it in more detail. As expected, the power of the method depends strongly on the level of superfamily aggregation at the working resolution. Best results are expected at resolutions higher than 10 Å, since we observed that for these resolutions the aggregation effects may be significantly decreased.

Multi-domain 3D-EM maps represent a further challenge to superfamily detection. Since, especially for lower resolutions, domain boundaries cannot be recognized and the blurring of neighboring domains in the relevant one lowers the sensitivity of superfamily detection. The SPI-EM approach was tested on both simulated and experimental multi-domain maps. Although the rejection criteria should be greatly relaxed in these cases, our results suggest that superfamily detection may still be possible in such maps.

One of the main drawbacks of the method is that it requires large amounts of CPU. Especially when dealing with multi-domain maps, most of the fitting programs are very time consuming.

Table 5. Superfamily detection for domains of simulated 3D-EM map from 1GR5

3D-EM map	Superfamily	M_{CC}	P-value	D-value	Q-value
1GR5	1.10.560.10	0.736	0.961	-0.191	1.33
1GR5	3.50.7.10	0.754	0.846	-0.019	0.19
1GR5	3.30.260.10	0.739	0.431	-0.227	3.34
1GR5	1.10.220.10	0.530	0.144	-0.407	4.75
1GR5	1.10.1250.10	0.483	0.102	-0.341	3.31
1GR5	1.20.1060.10	0.319	0.012	-0.539	6.20

Resolution, 10 Å. The shadowed superfamilies are the three ones present in the map.

SITUS-COLORES takes about 30 minutes on a 1 GHz Alpha CPU to complete a single fitting. In this way, one medium-sized superfamily of 30 elements takes about 15 hours. These computing times are expected to be greatly reduced once SITUS-COLORES with Fast Rotational Matching is released (W. Wriggers, personal communication). Another possibility to reduce the required computing time is a combination of SPI-EM with complementary methods to search structural elements in low-resolution maps. In particular, we mention the possibility to combine our approach with HELIXHUNTER.[11] HELIXHUNTER can be used as a discriminative tool to establish the minimum number of α-helices (the program is built to avoid false positives). Therefore, after running HELIXHUNTER on a query 3D-EM map, all superfamilies with fewer α-helices can be discarded with relative confidence. In this way the superfamily search space can be reduced, saving large amounts of computation time (results not shown). Another time-consuming aspect is the computation of background distributions of M_{CC} for each CATH superfamily, since it implies the fitting of all members of the superfamily against all Hreps at the relevant resolution. We are currently building a database with these distributions at different resolutions that will be made accessible *via* the WWW to serve the community.

Finally, it is worth noticing that the presented statistical approach is independent on the fitting program used, provided that this program allows fitting of single domains (of any architecture) in multi-domain maps. This implies that SPI-EM may benefit from any advances in the field of domain fitting in low-resolution maps.

Materials and Methods

The SITUS suite of programs v2.0 was used to perform most of the operations implied in the fittings. The fittings of isolated domains were performed with FRM.[18] For the multi-domain case, SITUS-COLORES[27] was employed with an angular step for the search of 30 degrees and the Powell optimization option set to off. Python scripts were developed to perform sequentially the required fittings, compute LCCC means, build the M_{CC} distributions, and derive the *P*, *D* and *Q*-values. The PDBLUR utility from the SITUS suite was used to lower the resolution of the PDB files and CATH domains. The CATH domains files for the fittings were generated directly from the corresponding PDB files using the CATH v2.4 description of domains. The visualization software AMIRA was employed to test the results of the fittings and develop the Figures of 3D-EM maps shown here.

Acknowledgements

Partial support is acknowledged from the Comisión Interministerial de Ciencia y Tecnología of Spain through project BIO2001-1237, special action BIO2002-10855-E and current project BFU2004-00217; the Fondo de Investigaciones Sanitarias through project FIS-IM3; the BBVA through project Towards high resolution structures of biological machines at work: computational tools to push the limits of three-dimensional electron microscopy; the EU through projects TEMBLOR (QLRI-CT-2001-00015) and the Network of Excellence on New electron microscopy approaches for studying protein complexes and cellular supramolecular architecture (FP6-502828); and from the NIH through grant 1R01HL67465-01. J.A.V. is supported by the Spanish Ministry of Science and Education through a FPU grant. The authors also thank Willy Wriggers at the University of Texas Health Science Center at Houston, who very kindly provided the FRM software to perform the fittings, and Matthew Baker, Wen Jiang and Wah Chiu at the National Center for Macromolecular Imaging at Houston, for many fruitful discussions.

References

1. van Heel, M., Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T. *et al*. (2000). Single-particle electron cryo-microscopy: towards atomic resolution. *Quart. Rev. Biophys.* **33**, 307–369.
2. Bottcher, B., Wynne, S. A. & Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, **386**, 88–91.
3. Jiang, W., Li, Z., Zhang, Z., Baker, M. L., Prevelige, P. E., Jr & Chiu, W. (2003). Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nature Struct. Biol.* **10**, 131–135.
4. Modis, Y., Trus, B. L. & Harrison, S. C. (2002). Atomic model of the papillomavirus capsid. *EMBO J.* **21**, 4754–4762.
5. Zhou, Z. H., Dougherty, M., Jakana, J., He, J., Rixon, F. J. & Chiu, W. (2000). Seeing the herpesvirus capsid at 8.5 Å. *Science*, **288**, 877–880.
6. Zhou, Z. H. & Chiu, W. (2003). Determination of icosahedral virus structures by electron cryomicroscopy at subnanometer resolution. *Advan. Protein Chem.* **64**, 93–124.
7. Baker, T. S., Olson, N. H. & Fuller, S. D. (1999). Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiol. Mol. Biol. Rev.* **63**, 862–922.
8. Valle, M., Zavialov, A., Li, W., Stagg, S. M., Sengupta, J., Nielsen, R. C. *et al*. (2003). Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. *Nature Struct. Biol.* **10**, 899–906.
9. Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L. & Saibil, H. R. (2001). ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell*, **107**, 869–879.
10. Ludtke, S. J., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2004). Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure (Camb)*, **12**, 1129–1136.
11. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001).

Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033–1044.
12. Kong, Y. & Ma, J. (2003). A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *J. Mol. Biol.* **332**, 399–413.
13. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta. Crystallogr. sect. A*, **47 (Pt 2)**, 110–119.
14. Hymphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38.
15. Volkmann, N. & Hanein, D. (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J. Struct. Biol.* **125**, 176–184.
16. Chacon, P. & Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.* **317**, 375–384.
17. Rossmann, M. G. (2000). Fitting atomic models into electron-microscopy maps. *Acta. Crystallogr. D*, **56 (Pt 10)**, 1341–1349.
18. Kovacs, J. A., Chacon, P., Cong, Y., Metwally, E. & Wriggers, W. (2003). Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta. Crystallogr. D*, **59**, 1371–1376.
19. Orengo, C. A., Pearl, F. M. & Thornton, J. M. (2003). The CATH domain structure database. *Methods. Biochem. Anal.* **44**, 249–271.
20. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
21. Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231–234.
22. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
23. Bailey, T. L. & Gribskov, M. (1998). Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
24. Bailey, T. L. & Grundy, W. N. (1999). *Proceedings of the 3rd Int. Conference on Comp. Mol. Biol. (RECOMB 99)*, ACM Press, Lyon, France.
25. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. (1998). DNA Visualizing replication in a catalytically active *Bacillus* DNA polymerase crystal. *Nature*, **391**, 304–307.
26. Tagari, M., Newman, R., Chagoyen, M., Carazo, J. M. & Henrick, K. (2002). New electron microscopy database and deposition system. *Trends Biochem. Sci.* **27**, 589.
27. Chacon, P., Tama, F. & Wriggers, W. (2003). Mega-Dalton biomolecular motion captured from electron microscopy reconstructions. *J. Mol. Biol.* **326**, 485–492.

Edited by W. Baumeister

(Received 21 June 2004; received in revised form 29 October 2004; accepted 2 November 2004)