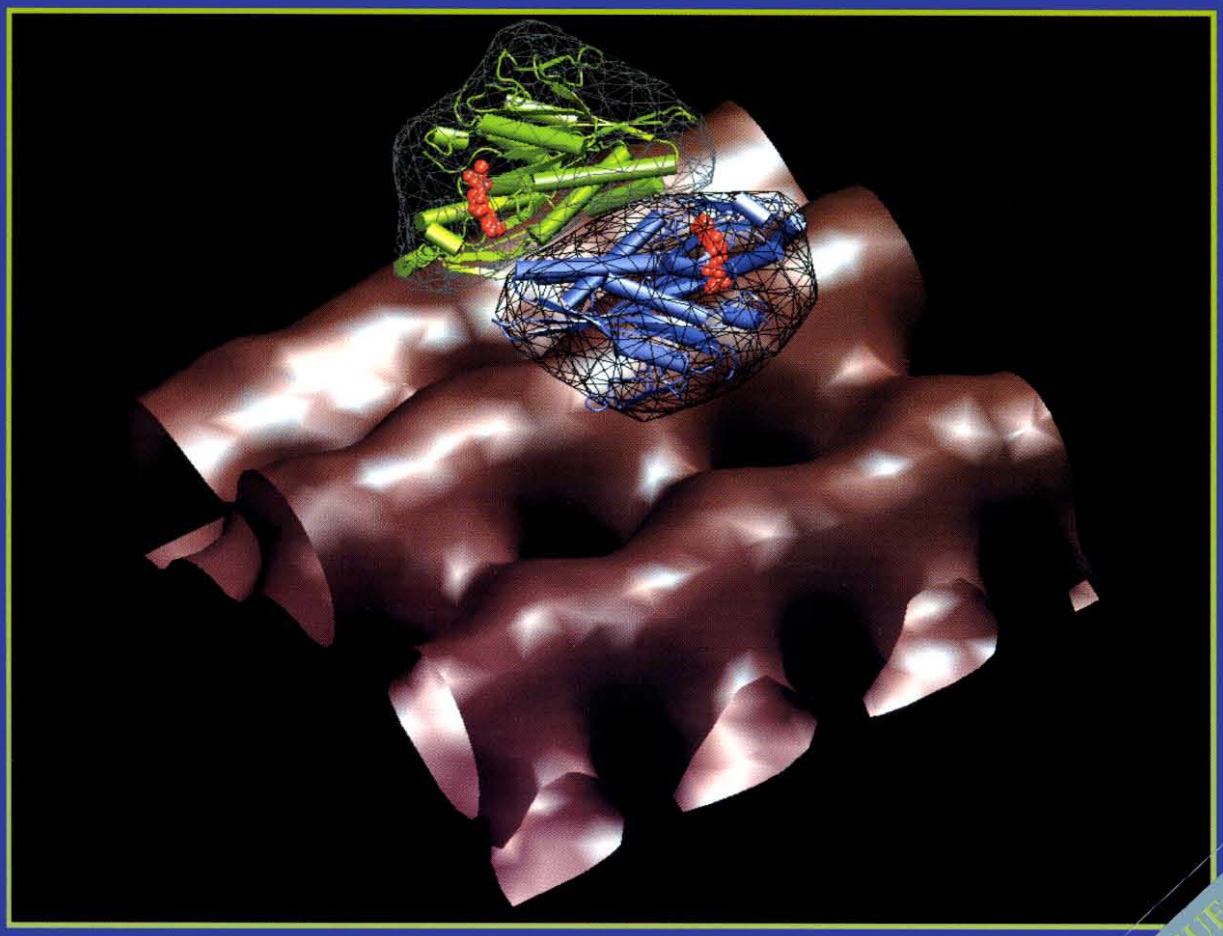
Volume 125, Numbers 2/3, April/May 1999
This Number Completes Volume 125

ISSN 1047-8477

Journal of Structural Biology



SPECIAL ISSUE ON MOLECULAR VISUALIZATION SOFTWARE

ACADEMIC PRESS

Journal of Structural Biology **125,** 185–195 (1999)
Article ID jsbi.1998.4080, available online at http://www.idealibrary.com on IDEAL®

Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy

Willy Wriggers,*,†,[1] Ronald A. Milligan,† and J. Andrew McCammon*

**Department of Chemistry and Biochemistry and Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0365; and †Department of Cell Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037*

Received October 26, 1998, and in revised form December 16, 1998

Three-dimensional image reconstructions of large-scale protein aggregates are routinely determined by electron microscopy (EM). We combine low-resolution EM data with high-resolution structures of proteins determined by x-ray crystallography. A set of visualization and analysis procedures, termed the Situs package, has been developed to provide an efficient and robust method for the localization of protein subunits in low-resolution data. Topology-representing neural networks are employed to vector-quantize and to correlate features within the structural data sets. Microtubules decorated with kinesin-related ncd motors are used as model aggregates to demonstrate the utility of this package of routines. The precision of the docking has allowed for the extraction of unique conformations of the macromolecules and is limited only by the reliability of the underlying structural data.

Key Words: **topology representing neural networks; multiresolution; visualization; macromolecular assemblies; kinesin; microtubules.**

INTRODUCTION

The role of visualization in structural biology is currently expanding, as novel modeling and graphics tools begin to integrate data from a variety of biophysical sources (Bajaj, 1998). The development of "multiresolution" structural visualization programs is stimulated in part by recent successful constructions of electron microscopy (EM)-based atomic-resolution models of viruses and cytoskeletal motor-filament complexes (Rayment *et al.*, 1993; Schröder *et al.*, 1993; Milligan, 1996; DeRosier and Harrison, 1997; Chiu *et al.*, 1997). Such synergistic applications require the development of advanced modeling techniques to bridge the scales in space and complexity of the underlying biological data (Mendelson and Morris, 1997; Hanein *et al.*, 1998; Wriggers *et al.*, 1998; Belnap *et al.*, 1999). In this work, we describe Situs, a set of routines for the quantitative docking of 3D data at variable resolution. The programs are deployed to support structural biologists in the construction of near-atomic resolution models of large-scale macromolecular assemblies.

In past applications, researchers employed mainly "visual docking" to determine the position of protein subunits in low-resolution envelopes. This approach is feasible in cases where the reliability of cryo-EM data provides a docking precision of four to five times the resolution of the experimental data (Baker and Johnson, 1996) or where there are additional constraints such as subunit localizations on gold labeling. However, in the case of ambiguous protein shapes, divergent models of complexes docked by eye have been reported (Sosa *et al.*, 1997; Kozielski *et al.*, 1998). Therefore, it is desirable to employ a quantitative method that would allow the researcher to assess the docking of multiple conformations of the complex by means of a score function. In this work, we describe a suite of programs that simplify this task of generating optimally docked macromolecules.

Uncertainties of the visual approach are overcome through the use of the 3D density distribution that describes the protein structure at low resolution (Frank, 1996). Key to the docking procedure are topology-representing neural networks (TRNs) that are used to correlate the high- and low-resolution data sets (Wriggers *et al.*, 1998). TRNs are one of several possible algorithmic realizations of vector quantization, a data compression technique frequently used in image and speech processing applications (Gray, 1984; Makhoul *et al.*, 1985). The representation of data by a small number of so-called codebook vectors reduces the combinatorial complexity of the structural comparison, as data sets are

[1] To whom correspondence should be addressed. Fax: (619) 534-7042. E-mail: wriggers@ucsd.edu.

1047-8477/99 $30.00

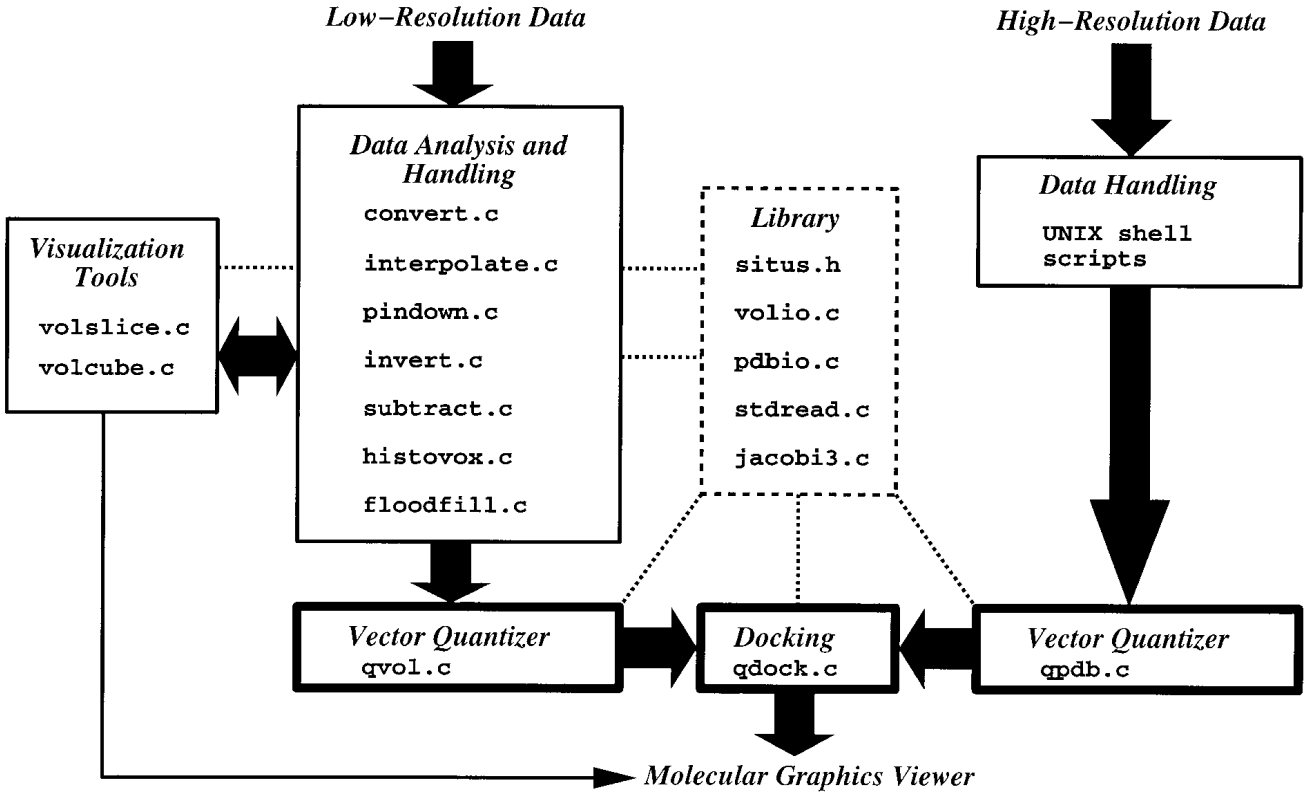


FIG. 1. Schematic diagram of the Situs package (version 1.0). Individual shell scripts and C program components are classified by their functionality. The data flow is indicated by the arrows. The procedures are discussed in the text. Additional documentation is available at URL http://chemcca10.ucsd.edu/~situs.

compared indirectly by the corresponding vectors. This reduced representation is crucial to the performance of the method. Situs performs a full exhaustive search of possible docking conformations within seconds, allowing the user to assess interactively the biological relevance of a list of ranked docking conformations.

The described docking makes use of single-molecule density maps that can be obtained by subtraction of maps from specimens of variable subunit composition. The more complex problem of localizing subunits in a single low-resolution data set of a biomolecular aggregate is described elsewhere (Wriggers *et al.*, 1998). We demonstrate the application of the Situs package using microtubules decorated with the kinesin-related motor protein ncd as a model specimen. The results are compared to published models of the assembly that were obtained by visual docking (Sosa *et al.*, 1997). Since the orientation of ncd relative to the microtubule is still controversial (Kozielski *et al.*, 1998), we discuss these preliminary results mainly in terms of the quality of the fit. We hope that the Situs package will become a standard tool that removes human bias from the docking of multiresolution data and thereby contributes to a consensus regarding the orientation of kinesin-related motors relative to the microtubule. We also discuss possible applications of the procedures to other specimens and to other areas of microscopy where 3D biological data sets are compared.

DESIGN OF THE SITUS PACKAGE

Overview. The software currently consists of individual, stand-alone programs for format conversion, analysis, visualization, manipulation, vector quantization, and docking of 3D data sets. Each program is self-explanatory as the user is asked to enter all relevant information at the shell prompt during interactive use on a UNIX workstation. The modular design of the package allows the user to maintain flexibility in its application. The documented source code written in ANSI C can be ported to other architectures. File names are passed as arguments to the programs of the distribution: "program [input file] [output file]."

The series of steps and the programs that are required to dock an atomic-resolution structure into low-resolution EM data are shown schematically in Fig. 1. Standard EM formats are supported and are converted into Situs format. Subsequently, the data are inspected and, if necessary, prepared for the vector quantization using a variety of visualization and analysis tools. Atomic coordinates in PDB format (Abola *et al.*, 1997) are filtered through UNIX shell scripts (e.g., to purge a crystal structure of

associated water molecules). After the vector quantization, the high-resolution structure is docked to the low-resolution density by the corresponding codebook vectors. The resulting docked complex can be inspected using a standard molecular graphics viewer. The user learns about the progress of each step from comments on results and parameters. All information provided by the user at the prompt is validated; error messages are returned if undefined program behavior would ensue.

The suite of programs is supported by a header file (situs.h) containing user-defined parameters (e.g., the floating point type used, the allowed range of the number of codebook vectors, the number of iterative steps of the TRN algorithm) and by auxiliary library programs. The library programs handle input and output of atomic coordinates in PDB format (pdbio.c), input and output of volumetric data (volio.c), input of data at the prompt (stdread.c), and Eigenvector computation for real symmetric 3×3 matrices (jacobi3.c). The program pdbio.c was adapted from the program DOWSER (Zhang and Hermans, 1996). The main C programs (Fig. 1) supported by the library programs will be described in some detail in the following paragraphs. The ".c" suffix will be dropped to denote the executables.

Conversion of file formats. Volumetric 3D density data are represented in a Situs-specific format that provides data compatibility between the individual programs to keep track of the relative positions of created maps. The convert utility reads file formats of standard EM application software. These include the MRC (Henderson *et al.*, 1990) and SPIDER (Frank *et al.*, 1996) formats, as well as similar 4-byte floating-point binary formats in which a header is followed by the sequence of density values. ASCII files that contain a sequence of density values in free format are also recognized. In the current version, a grid of cubic voxels is required, and the voxel size, as well as the number of x, y, and z increments (columns, rows, and sections), must be known in advance. The order of values in the sequence of densities is altered, if necessary, such that x increments change fastest and z increments change slowest. A short header holds the mentioned voxel size and numbers of x, y, and z increments, as well as the grid origin (i.e., the 3D coordinates of grid point $x = 1$, $y = 1$, $z = 1$). The header is followed by the sequence of data values. The converted Situs files are in ASCII format, allowing the user to verify the successful conversion of the data.

Preparation of volumetric data. The package contains a variety of routines for the manipulation of volumetric data in the Situs format. The subtract utility subtracts a set of density values from a second set exhibiting identical grid parameters. If necessary, the sign of the density values can be changed with the invert utility so that the protein corresponds to the maximum values of the density distribution. A smaller grid defining a region of interest can be extracted from a large data set with the pindown utility. The interpolate utility allows the user to alter the voxel size; density values will be converted to a new grid by linear interpolation. The floodfill utility finds a contiguous volume of density values above a threshold density level in the vicinity of a start grid point. Note that a special utility for the rescaling of density values is not necessary, because the data are automatically normalized by the vector quantization routines described below.

Volumetric density sets can be inspected in the UNIX shell without resorting to external graphics packages. Cross sections of the density data in the (x, y), (y, z), or (z, x) planes can be inspected with the visualization program volslice (Fig. 2). Those voxels whose corresponding density values exceed a threshold level are represented by text characters. The background is represented by a two-dimensional grid to help identify positions in the cross section. This simple rendering method is sufficient to locate individual voxels in the map.

The histovox utility prints the voxel histogram (Frank *et al.*, 1991; Frank, 1996) of the density values (Fig. 3). The histogram illustrates two general properties of EM density distributions. First, a pronounced peak at low densities (here: <12) is due to background scattering. The protein density corresponds to a second, broader peak at higher densities (here: >20). When integrating the histogram "from the top down," the known molecular volume of a protein can be used to compute its boundary density value (Frank, 1996). The histovox program also allows the user to add a constant value to the densities to shift the background density peak to the origin (Wriggers *et al.*, 1998).

Vector quantization. The TRN algorithm (Martinetz and Schulten, 1993; Martinetz and Schulten, 1994) offers a flexible way to develop a discrete representation of biological data including neighborhood relationships (Wriggers *et al.*, 1998). It can be shown that the algorithm minimizes the so-called distortion error (Makhoul *et al.*, 1985; Martinetz and Schulten, 1994) of the discrete representation. From a signal processing point of view this error measures the fidelity of the data encoding, i.e., the mean-squares deviation of the discrete set of vectors from the corresponding data (Makhoul *et al.*, 1985). If the global minimum of the error is found, the vector distribution is fully determined by the input density distribution. In this situation the codebook vectors form a set of control points that provide information about the shape and density distribution of a 3D biological object, similar

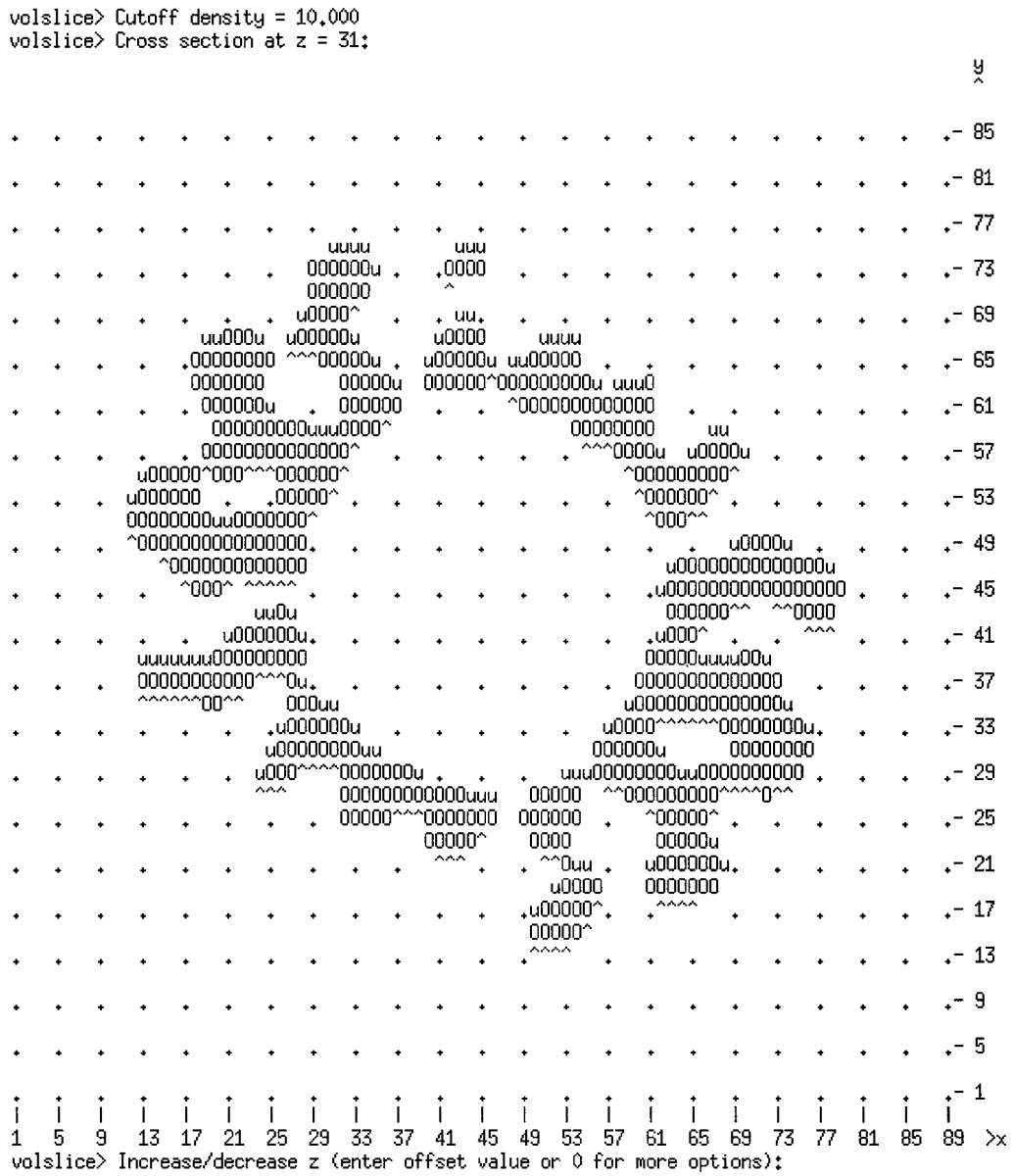


FIG. 2. Inspecting volumetric data with the volslice utility. Shown is the (x, y) cross section of a 534 Å × 534 Å × 366 Å data block of an ncd dimer-decorated microtubule (Sosa *et al.*, 1997). Note that pairs of voxels neighboring in the y-direction are represented by a single character: "^" if the upper voxel density exceeds a threshold (cut-off) level, "u," if the lower voxel density exceeds the threshold level, "0" if both upper and lower voxel densities exceed the threshold level, and "+," if the densities are below the threshold. This program requires the use of a fixed-width font in the shell window. The recommended font on a Silicon Graphics workstation is "fixed(misc) medium semicondensed 13 pt."

to the control points that determine the shape of Bézier spline curves in computer graphics (Foley *et al.*, 1990). Therefore, the TRN algorithm presents a good choice for learning the characteristics of 3D biological data sets at reduced complexity. During the iterative training, the distribution of codebook vectors responds to random input stimuli determined by the density distribution, lowering the distortion error.

Like other vector quantizers, the TRN algorithm is not guaranteed to find the global minimum of the distortion error. In the Situs implementation, a small number of calculations (eight by default) are repeated with different random number seeds. The averaged codebook vectors are returned in PDB-formatted atom coordinate files, and the rms variabilities, representing the precision of the codebook vectors, are written to the occupancy fields of the PDB-style atom entries. Both the statistical uncertainty and the existence of multiple minima of the distortion error contribute to the rms variabilities of the found codebook vectors. In general, the average variability depends on the shape of the 3D input density distribution and on the number of vectors. To optimally represent the input density distribution, the user should select the number of codebook vectors that minimizes this average rms variability of

```
histovox> Printing voxel histogram, 50 histogram bins
histovox> (density value; count):

  -32.000 | 1
  -30.367 | 7
  -28.735 |= 37
  -27.102 |= 46
  -25.469 |==== 113
  -23.837 |=== 93
  -22.204 |======== 210
  -20.571 |========== 247
  -18.939 |==== 109
  -17.306 |======= 192
  -15.673 |====== 170
  -14.041 |===== 134
  -12.408 |=========== 294
  -10.776 |============== 352
   -9.143 |========== 262
   -7.510 |============================== 762
   -5.878 |======================= 582
   -4.245 |========================================================================= 1723
   -2.612 |========================================================================= 1717
   -0.980 |============================ 696
    0.653 |================================================================ 1642
    2.286 |========================================= 1045
    3.918 |============ 318
    5.551 |============ 307
    7.184 |===== 140
    8.816 |========= 236
   10.449 |========= 226
   12.082 |==== 108
   13.714 |======== 204
   15.347 |======== 218
   16.980 |==== 111
   18.612 |======== 214
   20.245 |========== 247
   21.878 |==== 122
   23.510 |========= 227
   25.143 |==== 113
   26.776 |======== 217
   28.408 |========= 237
   30.041 |===== 136
   31.673 |========== 263
   33.306 |=========== 272
   34.939 |===== 142
   36.571 |=========== 294
   38.204 |=========== 276
   39.837 |==== 117
   41.469 |========= 227
   43.102 |=== 75
   44.735 |==== 101
   46.367 |= 40
   48.000 | 3
```

FIG. 3. Voxel histogram of volumetric data (Frank *et al.*, 1991; Frank, 1996) echoed to the screen by histovox. The shown density distribution corresponds to a $(150\ \text{Å})^3$ block of data from a microtubule surface (Sosa *et al.*, 1997), as shown in Fig. 6.

the vectors. Also, the convergence of the algorithm is improved by a neighborhood ranking of the codebook vectors to avoid getting trapped in local minima (Martinetz *et al.*, 1993). In combination with the averaging of the vectors from statistically independent runs, one can achieve near-optimal results that are sufficiently reliable for the docking problem.

Two vector quantization routines are provided by the Situs package: qvol, for the quantization of volumetric data, and qpdb, for the quantization of atomic resolution data. Only densities above a user-defined threshold value are considered by qvol to eliminate background noise in the EM data. The mass-weighted atomic coordinate vectors of the protein subunit form a discrete input vector distribution for qpdb. Each atom i of mass m_i is represented by a number $n_i = \text{round}(m_i/14)$ of equally weighted input vectors. As a result, chemical elements 1–3 (including H) are ignored (i.e., $n_i = 0$), chemical elements 4–10 (including C, N, and O) are represented by one input vector, chemical elements 11–16 (including P and S) are represented by two input vectors, etc. The vector quantization is not affected, within the precision of the codebook vectors, by this simple but efficient mass-weighting scheme (W.W., unpublished results).

Combining high- and low-resolution data. Let us assume that each of the data sets is represented by k codebook vectors $\mathbf{x}_i$, corresponding to high-resolution data, and by k codebook vectors $\mathbf{y}_j$, corresponding to low-resolution data ($i, j = 1, \ldots, k$). If the index map $I: j \rightarrow i$, defining the k pairs of corresponding vectors, is known, it is straightforward to dock the vectors $\mathbf{x}_{I(j)}$ to the vectors $\mathbf{y}_j$ by a least-squares fit (Kabsch, 1976, 1978). The resulting rigid-body transformation, applied to the original atomic structure, solves the docking problem. In practical situations, however, the index map I is not known *a priori*, and all $k! = k \cdot (k-1) \ldots 3 \cdot 2$ possible permutations $(I(1), \ldots, I(k))$ must be explored. The program qdock carries out an exhaustive search of the permutations and returns a list of best least-squares fits, ranked

```
qdock> Computing 7! = 5.040E+03 possible pairs of
  corresponding codebook vectors...
qdock> Printing 20 best least-squares fits (rmsd in Angstrom)
  and their correlation coefficients
qdock> Permutations indicate the order of qpdb vectors in
  file foo.pdb fitted to qvol vectors in file bar.pdb
qdock>
  1.  3.115   0.913  (7,5,1,6,4,2,3)
  2.  4.946   0.904  (2,3,5,7,4,6,1)
  3.  5.455   0.897  (6,1,3,2,4,7,5)
  4.  6.316   0.882  (5,7,4,3,1,2,6)
  5.  7.612   0.867  (5,7,1,4,6,3,2)
  6.  7.855   0.888  (3,2,4,1,5,6,7)
  7.  7.994   0.884  (1,6,4,5,3,7,2)
  8.  8.001   0.863  (6,1,4,3,5,2,7)
  9.  8.192   0.888  (2,6,4,3,1,7,5)
 10.  8.244   0.850  (7,5,6,2,1,3,4)
 11.  8.298   0.881  (2,6,7,5,1,3,4)
 12.  8.340   0.894  (6,2,4,1,3,5,7)
 13.  8.481   0.867  (3,4,6,2,1,5,7)
 14.  8.516   0.885  (2,3,4,5,1,7,6)
 15.  8.532   0.857  (7,5,4,1,3,6,2)
 16.  8.985   0.861  (6,1,5,7,4,3,2)
 17.  8.988   0.838  (3,4,5,7,1,2,6)
 18.  9.092   0.839  (3,2,5,4,7,1,6)
 19.  9.124   0.858  (7,5,3,2,4,1,6)
 20.  9.236   0.858  (1,6,5,7,4,2,3)
```

FIG. 4. List of best least-squares fits echoed to the screen by the qdock tool. Shown are the rms deviations of the codebook vectors corresponding to high- and low-resolution data, the correlation coefficients, and the vector permutations. The correlation coefficients C_{pq} between the grid-based EM density $p_{x,y,z}$ and the protein densities $q_{x,y,z}$, calculated by linear interpolation from the x-ray crystal structure in the trial orientations, were determined using the expression

$$C_{pq} = \frac{\sum_{x,y,z} p_{x,y,z} \cdot q_{x,y,z}}{\left(\sum_{x,y,z} p^2_{x,y,z}\right)^{1/2} \left(\sum_{x,y,z} q^2_{x,y,z}\right)^{1/2}}.$$

The results shown correspond to the docking of the ncd structure to the density of the attached motor domain, as described under Application of the Situs Package. The complex model resulting from the best fit is shown in Fig. 5.

by the remaining rms deviation after superposition of the vectors $\mathbf{x}_{I(j)}$ and $\mathbf{y}_j$ (Fig. 4). The ranking by codebook vector rms deviation typically produces a clear prediction of the optimum docking configuration. Figure 5 presents an example of optimally superimposed codebook vectors.

Volkmann and Hanein (1999) use a correlation coefficient that measures the overlap of the high- and low-resolution data as the criterion for the docking. The program qdock computes this coefficient for the best least-squares fits of the codebook vectors, as shown in Fig. 4. A lower rms deviation of the least-squares fit typically corresponds to a higher correlation coefficient. However, the coefficients all lie within a very narrow numeric range. Therefore, as mentioned by Hanein *et al.* (1998), fits based on the correlation coefficient alone are often ambiguous.

The program volcube in the Situs distribution produces wireframe meshes or solid surfaces of isocontours that can be displayed with atomic models of proteins using the free molecular graphics package VMD (Humphrey *et al.*, 1996), available at URL http://www.ks.uiuc.edu/Research/vmd. The volcube utility takes advantage of VMD's Tcl scripting capabilities (Dalke and Schulten, 1997) that allow the user to display geometric objects and surfaces consisting of shaded triangles. In particular, volcube generates the vertices and vertex norm vectors of triangles that can be used for solid rendering of surfaces. The VMD-specific part of the volcube code is concentrated in a single subroutine and can easily be modified to other standard graphics packages with surface or mesh rendering capabilities. The isofurfaces are generated with an improved version (Heiden *et al.*, 1993) of the "marching cubes" algorithm (Lorensen and Cline, 1987), adapted from the

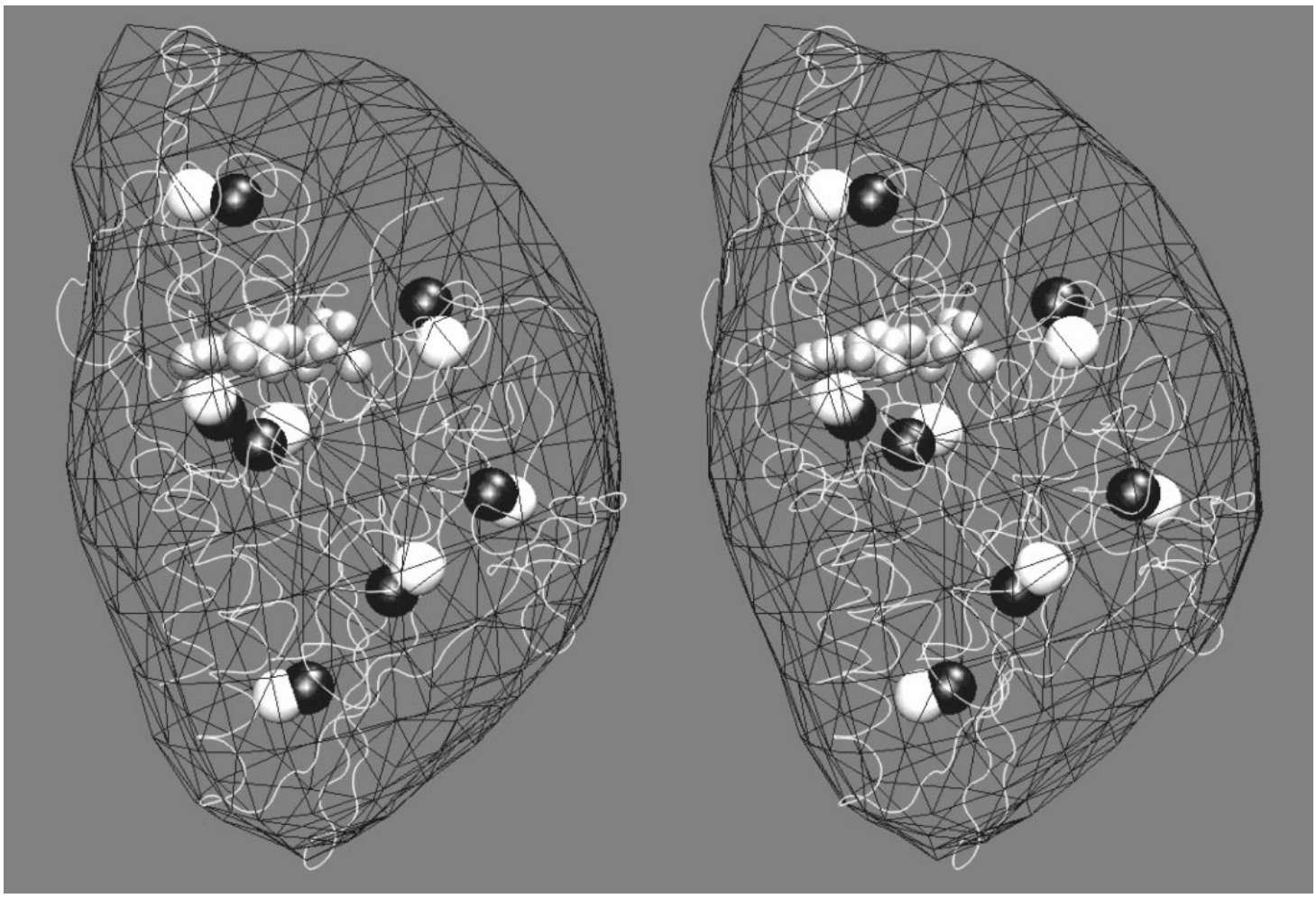


FIG. 5. Docking of high- and low-resolution data with the qdock tool (stereo view). Shown are the crystal structure of the ncd motor domain (Sablin *et al.*, 1998) and volumetric data (black wireframe) corresponding to the attached motor, as described in the text. A white backbone trace represents the protein, gray van-der-Waals spheres correspond to the ADP nucleotide and associated Mg^{2+} ion, white spheres indicate the position of the high-resolution codebook vectors, and black spheres identify the position of the low-resolution codebook vectors. The isocontour wireframe (isocontour level 20) was generated with volcube and the scene was rendered with the molecular graphics program VMD (Humphrey *et al.*, 1996) as viewed from the outside of the microtubule.

program VIEWMOL (Bleiber and Hill, 1993). In a practical application of volcube, the user first extracts from the volumetric data a region of interest (e.g., with pindown or floodfill) and then generates the VMD script using volcube. Information requested by volcube include the rendering style (wireframe or solid) and the new voxel size for the rendering of the isocontours, allowing the user to adjust the mesh size (the input grid is automatically interpolated).

APPLICATION OF THE SITUS PACKAGE

In practice a user would begin the docking with a vector quantization of the high-resolution structure with the program qpdb. The ncd monomer structure (residues 335–587 and 597–672) used in this work was obtained from the crystal structure of dimeric ncd (Sablin *et al.*, 1998). Nine residues of loop L11 are not defined in the crystal due to their disorder. We did not attempt to include these residues in the docking as was done by Sosa *et al.* (1997), since crystal B-factors and differences in two ncd crystal structures (Kull *et al.*, 1996; Sablin *et al.*, 1996, 1998), as well as the variability of the loop in molecular dynamics simulations of kinesin (Wriggers and Schulten, 1998), suggest that this is a flexible loop that changes conformation and position when the motor binds to the track. We also note that the crystal structure used in this work includes part of ncd's "neck" region not reported in the earlier crystal structure of this protein (Sablin *et al.*, 1996) and, therefore, corresponds better to the construct used for EM.

The rms variability of the codebook vectors indicates whether a unique minimum of the distortion error was found. Hence, to obtain more accurate results, the vector quantization of the high-resolution structure should be repeated for a range of the number of vectors. In the case of ncd, a minimum

TABLE I
Performance of the Programs qpdb and qdock as a Function of the Number *k* of Codebook Vectors

k	rmsf (Å)	Time (qpdb)	Time (qdock)
4	2.2	35 s	≪1 s
5	4.0	42 s	≪1 s
6	3.4	53 s	<1 s
7	1.4	63 s	1 s
8	1.7	74 s	4 s
9	2.5	85 s	46 s
10	2.7	95 s	8 min
11	1.4	110 s	85 min
12	1.8	122 s	17 h

Note. The rms fluctuations (rmsf) of the codebook vectors were obtained with the program qpdb from eight statistically independent TRN calculations involving 100 000 iterations each (Wriggers *et al.*, 1998). The variability values depend on the underlying structural data. Shown here are the values computed for the structure of an ncd motor domain from the dimer crystal structure (Sablin *et al.*, 1998) described in the text. Also shown are the times required to complete the qpdb calculations, as well as times for an exhaustive search of the *k*! least-squares fits (Kabsch, 1976, 1978) with the program qdock, on a Silicon Graphics Indigo-2 R4400 workstation.

average rms variability of 1.4 Å was obtained using 7 and 11 codebook vectors (Table 1).

The performance of the qpdb and qdock utilities depends on the number *k* of codebook vectors chosen (Table 1). The running time of the vector quantization, roughly on the order of a minute on a standard UNIX workstation, increases only linearly with *k*. However, the explosive growth of the *k*! combinations of the codebook vectors renders an exhaustive search impractical for comparisons of more than 12 vector pairs. It would be possible, in principle, to implement more efficient, nonexhaustive search methods in the qdock program for a large number of vectors. We note that the number of vectors necessary for the encoding of the underlying EM data is limited both by the experimental resolution and by the number of recognizable features in the data. Hence, for single molecules and a resolution below 10 Å, less than 10 vectors will usually be sufficient. In the present example, 7 vectors were chosen based on the vector rms variability criterion (Table 1).

As a next step, the user prepares aligned volumetric data sets of protein aggregates such that difference densities would correspond to single-molecule density distributions. The procedures presented here were tested on data sets of (1) undecorated microtubules and microtubules decorated with (2) monomeric ncd (residues 335–700) and (3) dimeric ncd proteins (residues 250–700) described by Sosa *et al.* (1997). The three density maps were converted to the Situs format using the convert utility, and difference maps (2-1) and (3-2) were computed with the subtract utility. The difference density (2-1) can be attributed to the first motor domain (residues 335–700) when attached to the microtubule. The difference density (3-2) corresponds to the density of the second motor domain (res. 335–700) plus the density of the coiled coil stalk (residues 250–334) of the dimer (Sablin *et al.*, 1998). The position of individual proteins in the density maps can be identified with the volslice program (cf. Fig. 2). As a final step in the preparation of the data, individual, contiguous protein densities are extracted with the floodfill utility. We note that for the ncd difference maps a (dimensionless) minimum threshold density of 12 was necessary to separate the individual protein densities.

Subsequently, the user carries out the vector quantization of the low-resolution density maps with the program qvol. A threshold value must be chosen to separate the protein density (above the threshold) from the background noise (below the threshold). This information can be obtained with the histovox utility (see above) or by consideration of the known molecular volume. For the data sets from decorated microtubules, an upper bound of 20 was estimated (cf. Fig. 3). A value of 12 was the lower bound based on the distribution of individual protein densities after application of floodfill. We chose a threshold value of 20 for the density corresponding to the first motor domain and a value of 12 for the density corresponding to the second motor domain to adjust for the lower densities of the latter data set (Sosa *et al.*, 1997). In the quantization of the EM data $k = 7$ codebook vectors were used, which were found to represent the crystal structure optimally (see above). The codebook vectors corresponding to high- and low-resolution data were then superimposed using the qdock utility (cf. Fig. 5). We note that the results were not sensitive to the choice of threshold value in the quantization. Similar to the docking of the crystal structure to the density of the attached motor domain (Fig. 4), the docking to the difference density between the dimer- and the monomer-decorated microtubule resulted in an unambiguous prediction using the rms deviation of the codebook vectors as a criterion (best fit 4.8 Å; second best fit 7.1 Å). The correlation coefficients (Eq. (1)) corresponding to the 20 best fits fell into the range of 0.82–0.90. The correlation coefficient criterion alone failed to identify a unique orientation of ncd: Both the 1st and the 11th fits (codebook vector rms deviation 9.4 Å) exhibited the maximum coefficient 0.90, but they are considerably different.

The resulting model of the two motor domains docked to the microtubule is shown in Fig. 6. The orientation of the first motor domain is close to that reported by Sosa *et al.* (1997). For the second, detached motor domain, the structural details in the EM map were insufficient to allow determination of a unique fit by eye (Sosa *et al.*, 1997). It is therefore surprising that the docking with Situs clearly pre-

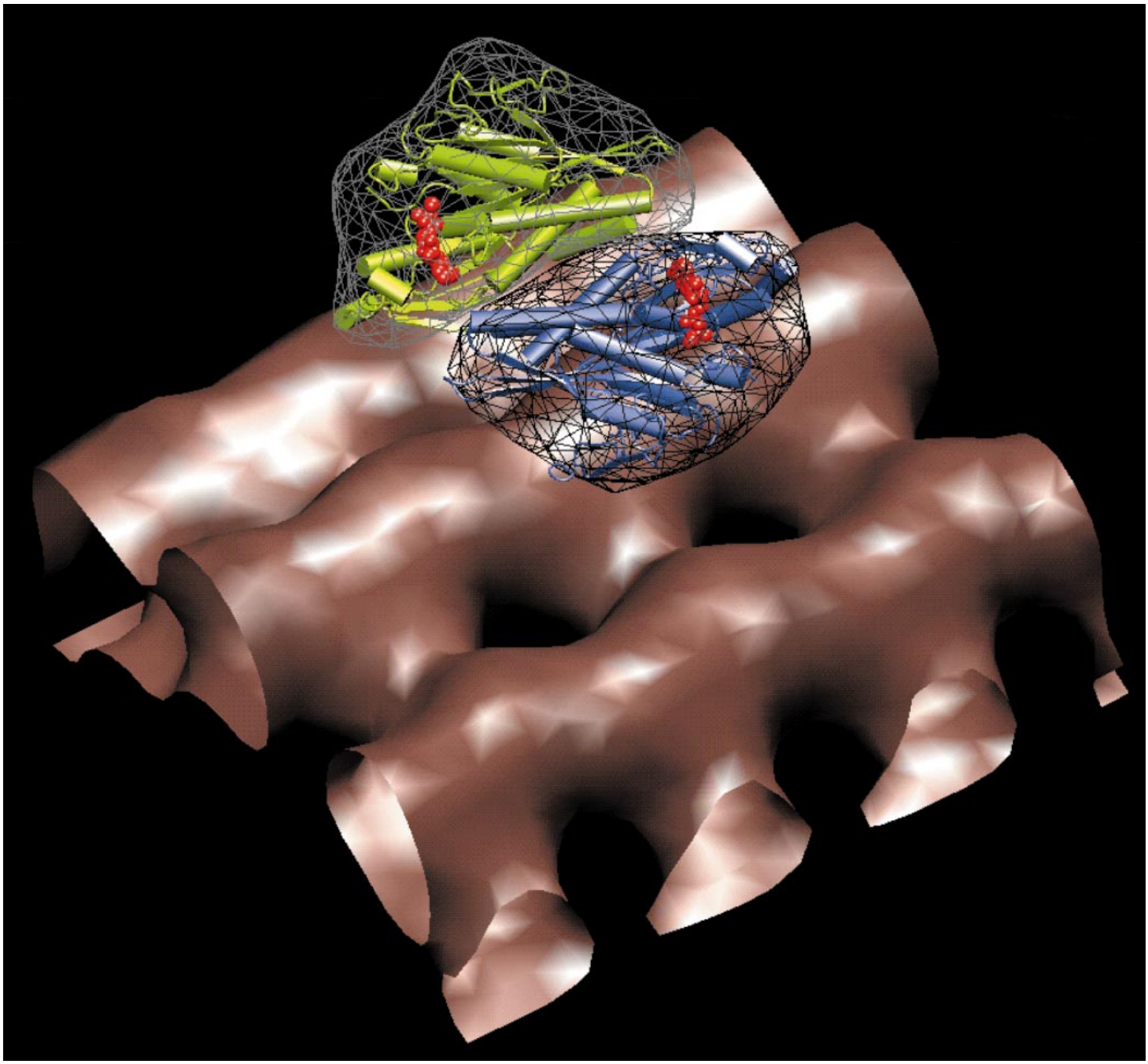


FIG. 6. Visualization of docked complexes with volcube. Shown is a solid surface representation of a (150 Å)^2 patch of the microtubule surface (Sosa *et al.*, 1997) at isocontour level 20. Also shown are wireframe representations of the attached (black; isocontour level 20) and detached (gray, isocontour level 12) motor-domain density maps and cartoon representations of the best fitting ncd structures (as described in the text), with associated ADP nucleotides and Mg^{2+} ions (red). The microtubule "+" end is oriented toward the lower left corner. The scene was rendered with the molecular graphics program VMD (Humphrey *et al.*, 1996).

dicted a single conformation of the motor domain that fits the EM data very well (Fig. 6). In this model, the two motor domains form a close-packed dimer: The heads are mutually attached by their neck regions, the nucleotide binding sites are oriented away from the microtubule, and the putative microtubule binding sites (Sosa *et al.*, 1997; Woehlke *et al.*, 1997) face the microtubule. The positions of the motor domains represented by their centers of mass suggest that the two motor domains move about 10 Å closer compared to the dimer crystal structure, where they are separated by the stalk (Sablin *et al.*, 1998). Also, the relative orientation of the two heads is significantly different from that in the dimer crystal structure, where the nucleotide-binding sites point in opposite directions (Sablin *et al.*, 1998).

The familiar, motor domain-like shape of the difference density distribution between the dimer- and the monomer-decorated microtubule motivated the docking of the second, detached motor domain. However,

the accuracy of the docking is limited by the reliability of the underlying structural data. Therefore, we enter two caveats against overly optimistic conclusions drawn form the unique fit of the second motor domain. First, the densities of the detached motor domain were significantly lower than those of the first motor domain (Sosa *et al.*, 1997), suggesting a higher degree of disorder of the detached head, which may affect the density distribution. Second, the model does not account for the density of the stalk of the ncd dimer (residues 250–334) and one must assume that this part of the protein is delocalized. Further experimental work and a comparison of data from different laboratories are necessary to determine with more confidence the orientation of the second motor domain.

DISCUSSION AND FUTURE PROSPECTS

We have developed and disseminated a set of procedures for the quantitative and reproducible docking of x-ray crystal structures to EM data. It is possible for a single operator to complete the docking within about an hour, a process that previously would have involved visual predictions by several individuals and a subsequent structural evaluation of the resulting fits (Sosa *et al.*, 1997). The package is freely available and designed to be user-friendly and portable. The modular design and the use of the ANSI C programming language should readily accommodate any modifications desired for applications other than rigid-body docking of crystallographic and EM data.

The major conceptual difference between Situs and other methods of combining 3D biological data (Mendelson and Morris, 1997; Volkmann and Hanein, 1999) is the discretization of the configurational search space by vector quantization. This discretization yields a list of best-scoring superpositions of the codebook vectors that represent the structural data sets. Our results suggest that this ranking constitutes a more stringent criterion for finding the optimum docking configuration than the correlation coefficient. The correlation coefficient, which measures the position-dependent overlap of the data sets, depends less sensitively on the relative orientation of the data sets and, therefore, may lead to ambiguous results unless it is combined with additional biochemical information (Hanein *et al.*, 1998) to eliminate false positives.

The vector quantization of 3D biological data has additional benefits that can be exploited in various practical applications: First, the pairs of corresponding codebook vectors could be used as constraints for flexible docking of protein structures in steered molecular dynamics simulations (Izrailev *et al.*, 1998). Second, the quantization paradigm can be employed in the construction of large-scale assemblages of protein subunits. As described recently (Wriggers *et al.*, 1998), atomic resolution models of actin filaments can be built from low-resolution EM data using the TRN technique. Third, the quantization simplifies the computational task of the search in configurational space. Comparisons of less than 10 pairs of codebook vectors can be carried out interactively within seconds. This speed would make it possible, in principle, to develop a World Wide Web-based version of the Situs package.

The mutual fitting of low-resolution density maps represents a problem often encountered in microscopy. A recent study of the formation of the gap junction intercellular channel indicated that two apposing connexons, imaged at low resolution, require a 30° rotation for the protrusions of one connexon to fit into the valleys of the other, which would make for an ionically tight interface necessary for a functional cell–cell channel (Perkins *et al.*, 1998). This model was substantiated by a docking of two connexons into a reconstruction of the whole channel. Future applications of the Situs package may be helpful to automate pure low-resolution modeling efforts such as described by Perkins *et al.* (1998). The described multiresolution method may also be useful in other microscopic applications where 3D biological data sets are compared, such as in the combination of electron- with light-microscopic images of cellular organelles.

The current implementation of the Situs package is available from URL http://chemcca10.ucsd.edu/~situs.

We thank Michael Zeller, Devin Lee Drew, and Gina Sosinsky for discussions. We also thank Jörg-Rüdiger Hill and Jan Hermans for their permission to include source code of the VIEWMOL and DOWSER programs in the Situs distribution. This work was supported in part by grants from NIH and the NRAC program of the NSF Supercomputer Centers. W.W. acknowledges the LJIS Interdisciplinary Training Program and The Burroughs Wellcome Fund for fellowship support.

REFERENCES

Abola, E., Sussman, J., Prilusky, J., and Manning, N. (1997) Protein data bank archives of three-dimensional macromolecular structures. *in* Carter, C., Jr., and Sweet, R. (Eds.), Methods in Enzymology Vol. 277, Academic Press, San Diego.

Bajaj, C. (1998) Data Visualization Techniques. Wiley, New York.

Baker, T., and Johnson, J. (1996) Low resolution meets high: Towards a resolution continuum from cells to atoms, *Curr. Opin. Struct. Biol.* **6,**585–594.

Belnap, D. M., Kumar, A., Folk, J. T., Smith, T. J., and Baler. T. S. (1999) Low-resolution density maps from atomic models: [How stepping "back" can be a step "forward"], *J. Struct. Biol.* **125,** 166–175.

Bleiber, A., and Hill, J.-R. (1993) VIEWMOL, Version 1.4. Arbeitsgruppe Quantenchemie an der Humboldt-Universität zu Berlin. Max-Planck-Gesellschaft.

Chiu, W., Burnett, R., and Garcea, R. (1997) Structural Biology of Viruses, Oxford Univ. Press, New York.

Dalke, A., and Schulten, K. (1997) Using Tcl for molecular visualization and analysis. *In* Proceedings of the Pacific Symposium on Biocomputing 97 on Interactive Molecular Visualization. World Scientific, Singapore.

DeRosier, D., and Harrison, S. (1997) Macromolecular assemblages: Sizing things up, *Curr. Opin. Struct. Biol.* **7,** 237–238.

Foley, J., van Dam, A., Feiner, S., and Hughes, J. (1990) Computer Graphics, Principles and Practices, Addison-Wesley, New York.

Frank, J. (1996) Three-Dimensional Electron Microscopy of Macromolecular Assemblies, Academic Press, San Diego.

Frank, J., Penczek, P., Grassucci, R., and Srivastava, S. (1991) Three-dimensional reconstruction of the 70S *E. coli* ribosome in ice: The distribution of ribosomal RNA, *J. Cell Biol.* **115,** 597–605.

Frank, J., Radermacher, J., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., and Leith, A. (1996) SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields, *J. Struct. Biol.* **116,** 190–199.

Gray, R. (1984) Vector quantization, *IEEE ASSP Mag.* **1,** 4–29.

Hanein, D., Volkmann, N., Goldsmith, S., Michon, A.-M., Lehman, W., Craig, R., DeRosier, D., Almo, S., and Matsudaira, P. (1998) An atomic model of fimbrin binding to F-actin and its implications for filament crosslinking and regulation, *Nat. Struct. Biol.* **5,** 787–792.

Heiden, W., Goetze, T., and Brickmann, J. (1993) Fast generation of molecular surfaces from 3D data fields with an enhanced "marching cube" algorithm, *J. Comput. Chem.* **14,** 246–250.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H. (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy, *J. Mol. Biol.* **213,** 899–929.

Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD—Visual molecular dynamics, *J. Mol. Graphics* **14,** 33–38.

Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F., Wriggers, W., and Schulten, K. (1998) Steered molecular dynamics. *in* Algorithms for Macromolecular Modelling, Lecture Notes in Computational Science and Engineering, Springer Verlag, Berlin.

Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors, *Acta Crystallogr. Sect. A* **32,** 922–923.

Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Crystallogr. Sect. A* **34,** 827–828.

Kozielski, F., Arnal, I., and Wade, R. (1998) A model of the microtubule–kinesin complex based on electron cryomicroscopy and X-ray crystallography, *Curr. Biol.* **8,** 191–198.

Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J., and Vale, R. D. (1996) Crystal structure of the kinesin motor domain reveals a structural similarity to myosin, *Nature* **380,** 550–555.

Lorensen, W., and Cline, H. (1987) Marching cubes: A high resolution 3D surface construction algorithm, *Comput. Graph.* **21,** 163–169.

Makhoul, J., Roucos, S., and Gish, H. (1985) Vector quantization in speech coding, *Proc. IEEE* **73,** 1551–1588.

Martinetz, T., and Schulten, K. (1994) Topology representing networks, *Neural Networks* **7,** 507–522.

Martinetz, T., and Schulten, K. (1993) A neural network for robot control: Cooperation between neural units as a requirement for learning, *Comput. Elect. Eng.* **19,** 315–332.

Martinetz, T. M., Berkovich, S. G., and Schulten, K. (1993) "Neural gas" for vector quantization and its application to time-series prediction, *IEEE Trans. Neural Networks* **4,** 558–569.

Mendelson, R., and Morris, E. (1997) The structure of acto-myosin subfragment 1 complex: Results of searches using data from electron microscopy and X-ray crystallography, *Proc. Natl. Acad. Sci. USA* **94,** 8533–8538.

Milligan, R. (1996) Protein–protein interactions in the rigor actomyosin complex, *Proc. Natl. Acad. Sci. USA* **93,** 21–26.

Perkins, G., Goodenough, D., and Sosinsky, G. (1998) Formation of the gap junction intercellular channel requires a 30° rotation for interdigitating two apposing connexons, *J. Mol. Biol.* **277,** 171–177.

Rayment, I., Holden, H., Whittaker, M., Yohn, C., Lorenz, M., Holmes, K., and Milligan, R. (1993) Structure of the actin–myosin complex and its implications for muscle contraction, *Science* **261,** 58–65.

Sablin, E. P., Kull, F. J., Cooke, R., Vale, R. D., and Fletterick, R. J. (1996) Crystal structure of the motor domain of the kinesin-related motor ncd, *Nature* **380,** 555–559.

Sablin, E., Case, R., Dai, S., Hart, C., Ruby, A., Vale, R., and Fletterick, R. (1998) Direction determination in the minus-end-directed kinesin motor ncd, *Nature* **395,** 813–816.

Schröder, R., Manstein, D., Jahn, W., Holden, H., Rayment, I., Holmes, K., and Spudich, J. (1993) Three-dimensional atomic model of F-actin decorated with *Dictyostelium* myosin S1, *Nature* **364,** 171–174.

Sosa, H., Dias, P., Hoenger, A., Whittaker, M., Wilson-Kubalek, E., Sablin, E., Fletterick, R., Vale, R., and Milligan, R. (1997) A model for the microtubule–ncd motor protein complex obtained by cryo-electron microscopy and image analysis, *Cell* **90,** 217–224.

Volkmann, N., and Hanein, D. (1999) Quantitative fitting of atomic models into observed densities derived by electron microscopy, *J. Struct. Biol.* **125,** 176–184.

Woehlke, G., Ruby, A., Hart, C., Ly, B., Hom-Boohrer, N., and Vale, R. (1997) Microtubule interaction site of the kinesin motor, *Cell* **90,** 207–216.

Wriggers, W., and Schulten, K. (1998) Nucleotide-dependent movements of the kinesin motor domain predicted by simulated annealing, *Biophys. J.* **75,**646–661.

Wriggers, W., Milligan, R., Schulten, K., and McCammon, J. (1998) Self-organizing neural networks bridge the biomolecular resolution gap, *J. Mol. Biol.* **284,** 1247–1254.

Zhang, L., and Hermans, J. (1996) Hydrophilicity of cavities in proteins, *Proteins: Struct. Function Genet.* **24,** 433–438.