Journal of Structural Biology **136,** 214–225 (2001)
doi:10.1006/jsbi.2002.4439, available online at http://www.idealibrary.com on IDEAL®

Semi-automated Icosahedral Particle Reconstruction at Sub-nanometer Resolution

Wen Jiang,*,† Zongli Li,† Zhixian Zhang,† Christopher R. Booth,*,† Matthew L. Baker,*,† and Wah Chiu*,†

**Program in Structural and Computational Biology and Molecular Biophysics, †National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030*

Received October 15, 2001, and in revised form December 7, 2001; published online April 10, 2002

Electron cryomicroscopy of large macromolecular complexes is becoming an increasingly powerful tool for revealing three-dimensional structures without the need for crystallization. The execution of image processing, however, requires experience and is error-prone due to the need for a human operator to carry out interactive and repetitive processes. We have designed an approach which is intended to make image processing simple and rapid, both for experts and for novice users. We demonstrate this approach using the well-established reconstruction scheme for icosahedral particles. Finally, we implement semi-automated virus reconstruction (SAVR), an expert system that integrates the most CPU intensive and iterative steps using the scripting language Python. SAVR is portable across platforms and has been parallelized to run on both shared and distributed memory platforms. SAVR also allows the incorporation of new algorithms and facilitates the management of the increasingly large data sets needed to achieve higher resolution reconstructions. The package has been successfully applied to several data sets and shown capable of generating icosahedral reconstructions to sub-nanometer resolutions (7–10 Å). © 2002 Elsevier Science (USA)

Key Words: **3-D reconstruction; automation; electron cryomicroscopy; high throughput; icosahedral; python; virus.**

INTRODUCTION

Electron cryomicroscopy is a high-resolution structural technique from which several atomic models have been generated using 3–4 Å data from two-dimensional crystals (Henderson *et al.*, 1990; Kimura *et al.*, 1997; Kühlbrandt *et al.*, 1994; Murata *et al.*, 2000; Nogales *et al.*, 1998). Among noncrystalline objects, spherical viruses with icosahedral symmetry are particularly well suited for analysis by this imaging technique and are potentially capable of reaching equally high resolutions. Presently, icosahedral reconstructions are limited to a sub-nanometer resolution of 7–9 Å (Böttcher *et al.*, 1997; Conway *et al.*, 1997; Mancini *et al.*, 2000; Trus *et al.*, 1997; Zhou *et al.*, 2000, 2001).

The data processing of icosahedral single particle images consists of a series of steps: boxing of individual particles from micrographs, contrast transfer function (CTF) parameter determination, initial orientation determination, orientation refinement, correction of the CTF, 3-D reconstruction, visualization, segmentation, and structural interpretation (Baker *et al.*, 1999; Thuman-Commike and Chiu, 2000; van Heel *et al.*, 2000). Over the years, many programs and algorithms have been implemented which deal with each of these steps within a variety of software packages.

The current reconstruction process involves many programs working in tandem. The programs run interactively, requiring an experienced user to input appropriate parameters for each specific task. The user must repeat the interactive processing steps for each micrograph, and the whole process must be iterated through many cycles to achieve a refined structure. The interactive processing is not only error-prone, but also makes less than optimal use of computing resources due to operator delays. Currently, it takes months for a user to complete a structural determination of an icosahedral particle at 7–9 Å resolution. Needless to say, it will take much longer for a beginner to accomplish the same task, because the user must thoroughly understand all aspects of electron cryomicroscopy, image processing, the common lines principle, and Fourier Bessel transforms, as well as overcoming the steep learning curve for each program. To achieve an atomic-resolution structure, orders of magnitude more images will be needed (Saad *et al.*, 2001), necessitating even longer processing time if the current reconstruction process is used.

In order to make single particle cryomicroscopy an

1047-8477/01 $35.00
© 2001 Elsevier Science (USA)
All rights reserved.

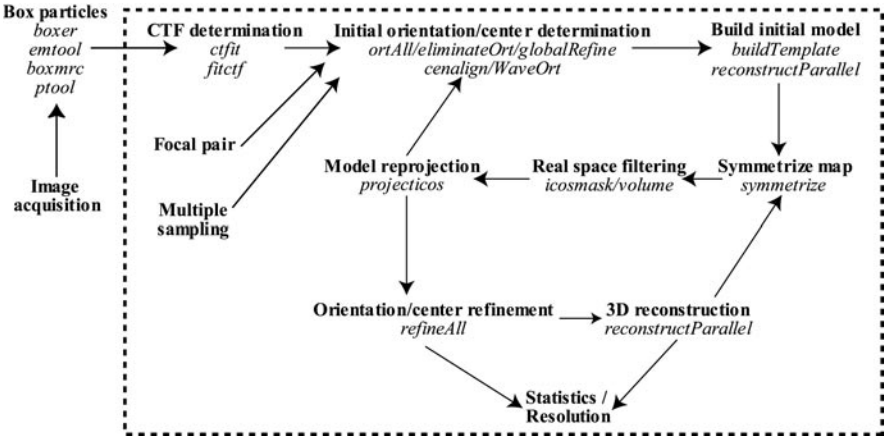


FIG. 1. Reconstruction process. A flow diagram of the commonly used processing scheme for icosahedral particle reconstruction is illustrated. Steps in the dotted box are included in SAVR.

attractive tool for biologists, the technology has to be not only capable of revealing reliable high-resolution features of the structures under study, but also easy to use. Therefore, we have developed a computational infrastructure, which will allow novice users to use the data processing software and experts to modify or add data processing steps easily. We have demonstrated this approach with a set of procedures for 3-D structural determination of icosahedral particles, called semi-automated virus reconstruction (SAVR). This package has been applied to several data sets for reconstructions spanning a broad range of resolutions from 30 to 8 Å.

METHODS

Icosahedral Reconstruction

Shown in Fig. 1 is a flow chart diagram of a commonly used process in icosahedral particle reconstruction. The experimental data consist of focal pair images with the first image taken close-to-focus and the second image of the same area far-from-focus (Fuller, 1987; Zhou *et al.*, 1994). The far-from-focus images are used for initial orientation determination of particles at low resolution. The initial orientation/center parameters are transferred to the corresponding close-to-focus particle images for refinement to higher resolution. The particles in both micrographs are boxed out by interactive manual selection and identical particles have to be correctly identified from the focal pair micrographs. The primary method for particle orientation/center parameter estimation and refinement is based on the principle of the Fourier common lines algorithm (DeRosier and Klug, 1968; Crowther, 1971; Thuman-Commike and Chiu, 1997). In addition, we have also developed a Wavelet transform-assisted projection matching algorithm for initial orientation determination when an existing model of the structure is available (Saad *et al.*, manuscript in preparation). The 3-D reconstruction employs the Fourier–Bessel synthesis algorithm (Crowther, 1971).

Software Design Principles

The design philosophy for SAVR is to maximize the use of existing proven programs rather than reinventing them. Our target is to integrate all these well-established individual interactive programs into a coherent, intervention-free package. This package should also make it easy for users to add new programs or modify the existing programs in the data processing scheme shown in Fig. 1. In contrast to the proprietary scripting languages implemented in many other image processing packages (comprehensively presented in several issues in *Journal of Structural Biology* Vols. 116, 125, and 133), we adopted the standard scripting language Python for the first time in electron cryomicroscopy image processing. The object-oriented scripting language Python has been used to glue the individual programs together (Huang *et al.*, 2000; Ramu *et al.*, 2000; Sanner, 1999). This choice was made because Python is an open source scripting language that is well designed and continuously improved by a large number of computer language experts and members of the user community (http://www.python.org). Python is feature rich, well documented, and easily accessible to any user.

Software Composition

There are five major components in SAVR, of which two are existing programs and three are newly implemented (Fig. 2). More detailed information about each of the programs is available at the SAVR web page (http://ncmi.bcm.tmc.edu/~wjiang/savr).

IMIRS. This software package consists of over 100 individual modular programs for icosahedral particle processing (Zhou *et al.*, 1998), including initial orientation estimation based on self-common lines, projection-based cross-common line orientation refinement, focal pair matching, and Fourier–Bessel reconstruction programs. The refinement and reconstruction programs have been parallelized under the shared memory programming paradigm (Zhou *et al.*, 1998) and distributed memory system with the message passing interface (MPI) standard (http://www-unix.mcs.anl.gov/mpi/). This package has been used to reconstruct a number of icosahedral particles between 7–9 Å resolutions (Zhou *et al.*, 2000, 2001). This software is written in C and FORTRAN.

EMAN. This package (Ludtke *et al.*, 1999) is implemented for processing single particles, with many useful programs for displaying micrographs and particle images, particle boxing, CTF parameter fitting, and Fourier transformation. It contains a C++ programming library, which is convenient for developing new programs, and a parallelization method, *runpar*. All programs in EMAN are implemented in C++.

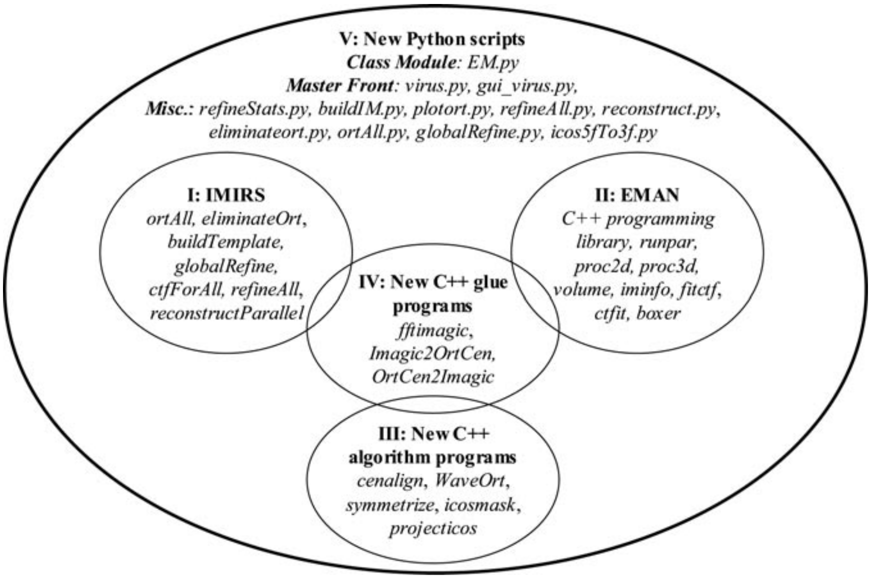

V: New Python scripts
Class Module: EM.py
Master Front: virus.py, gui_virus.py,
Misc.: refineStats.py, buildIM.py, plotort.py, refineAll.py, reconstruct.py, eliminateort.py, ortAll.py, globalRefine.py, icos5fTo3f.py
I: IMIRS
ortAll, eliminateOrt, buildTemplate, globalRefine, ctfForAll, refineAll, reconstructParallel
II: EMAN
C++ programming library, runpar, proc2d, proc3d, volume, iminfo, fitctf, ctfit, boxer
IV: New C++ glue programs
fftimagic, Imagic2OrtCen, OrtCen2Imagic
III: New C++ algorithm programs
cenalign, WaveOrt, symmetrize, icosmask, projecticos

FIG. 2. Composition of SAVR. The composition and relationship of the individual programs in SAVR are shown. More detailed information about each of the programs is available at the SAVR web page (http://ncmi.bcm.tmc.edu/~wjiang/savr).

New algorithmic programs. These programs include *WaveOrt*, a Wavelet transform-assisted projection matching algorithm for initial particle orientation determination; *symmetrize*, a real space symmetrizing program which can apply C_n, D_n and full icosahedral symmetries; *icosmask*, a masking program specially designed for icosahedral geometry; and *projecticos*, a real space projection program for computing-projections using multiple CPUs in parallel. These programs are implemented in C++ and use the EMAN library.

Glue programs. The glue programs allow the applicable programs in IMIRS, EMAN, and our newly implemented programs to share information and work together. These programs are essential because the above software packages use different file formats and conventions. These programs are written in C++ and use the EMAN programming library.

Top-level Python scripts. These scripts implement an expert system built on the knowledge and experience gained from past image processing and 3-D reconstruction of icosahedral objects. These scripts integrate the stand-alone programs into an automated, user intervention-free, icosahedral reconstruction method by looping through all the refinement cycles for all the micrographs and calling up the appropriate individual programs with appropriate parameters. More details on how the scripts work are presented in following sections.

Parallelization

SAVR has been parallelized and is capable of running on either distributed or shared memory platforms. Due to the large amount of data involved and the large computing resource required for the iterative refinements, parallel processing is essential in order to complete the task in the shortest possible time. Different parallelization strategies are used depending on the nature of the task. For instance, initial orientation and center determination and refinement belong to the "embarrassingly parallel" class of problems, where no inter-process communication is needed. The data set is naturally split into units based on individual micrographs and remains split during all the refinement iterations. The separate units are simply assigned to multiple processors using the *runpar* technique in EMAN (Ludtke *et al.*, 1999). Similarly, the symmetrization of the 3-D maps and the reprojection of the 3-D maps to generate 2-D images require no inter-process communication and can be efficiently parallelized using the *runpar* technique. Therefore, these three tasks are parallelized at the job submission level and not at the program level. In contrast, the 3-D reconstruction uses the Fourier–Bessel synthesis algorithm and requires inter-process communication. Thus, this task has been parallelized at the program level using MPI.

Portability

All of the five components in SAVR are portable across platforms. The individual programs from EMAN and the new algorithmic programs implemented using the EMAN library can run on different platforms. The icosahedral orientation refinement and Fourier–Bessel reconstruction programs in IMIRS have already been ported to several computer platforms (SGI IRIX, MS Windows NT/2000, and Linux). Since Python is designed to be a cross-platform portable language, the Python scripts in SAVR are naturally cross-platform portable.

SAVR has already been used on two dramatically different multiprocessor computer systems: a shared memory SGI Origin 2000 system and a distributed memory Linux cluster system. All the programs written in C++ or FORTRAN have been recompiled to run on the intended platform, while the Python scripts are readily used on any computer without any change.

Our main goal is to allow SAVR to run on different computer platforms so that the users are not limited by the computer system available to them. The recent Linux systems are especially attractive for the significantly high performance/price ratio. In addition to the speed of single CPU floating point operations and the memory bandwidths, the performance and time needed for processing using SAVR on multiprocessors are affected by many additional factors, such as file system performance and data transfer speed among different processors. Our tests on the P22 procapsid shell data set showed similar wall clock time on both an old SGI Onyx2 system (R10000 195 MHz) and a more recent Linux NetworX cluster system (PIII 800 MHz).

Graphical Interface

A graphic user interface was implemented to make it easier for the user to input the to-be-processed data and processing options into SAVR, as well as starting the automated reconstruction. The interface was implemented using PyQt (http://www.theKompany.com/projects/pykde), which is a Python binding of the multiplatform compatible Qt graphic programming library (http://www.trolltech.com). This graphic interface is a convenient add-on to the core SAVR programs and scripts, providing an alternative to text files for selecting data and choosing processing options in SAVR.

SAVR outputs several types of statistical information during processing. This information is presented in line or scatter plots using the popular plot program *xmgrace* (http://plasma-gate.weizmann.ac.il/Grace). The plots are formatted with appropriate titles, axis labels, and legends.

RESULTS

Overall Procedures

The individual programs (Fig. 1) which perform the actual data processing were implemented by different people in different programming languages, using different supporting libraries and with different conventions. In SAVR, Python is used to integrate these programs into an automated procedure for icosahedral particle reconstruction. The Python scripts automatically loop through each of the micrographs, call the appropriate programs to estimate the initial orientation, iteratively refine orientation parameters, correct CTF and overall experimental B-factor, compute the 3-D maps, symmetrize the 3-D map, perform the real-space 3-D filtering, and execute the reprojection of the 3-D model. The appropriate parameters for each of the programs at different refinement stages are different; SAVR automatically adapts and recomputes the appropriate parameters as the processing progresses. The whole process is simplified to a single run of a single Python script, *virus.py*, which executes from start to finish without any user intervention after initial setup. If the run is not completed due to any problem, such as a system crash, a simple rerun of *virus.py* will skip all the finished steps and restart from the point of interruption, thereby minimizing the waste of computing resources and time. This is done using *virus.py*, which checks for the stamping log files created after each of the finished processing steps. SAVR automatically identifies and skips the finished iterations and processed micrographs. This feature proved to be of extraordinary benefit during the development phase of SAVR, which took place on less robust computer systems.

There are several distinct features implemented in SAVR that can be enabled and disabled by the user.

Wavelet transform-assisted projection matching. Accurate initial orientation estimation is essential for convergent orientation refinement and to give a high yield of particles with common line phase residuals lower than the acceptable threshold. For high-resolution reconstruction, it is necessary to include close-to-focus micrographs (Saad *et al.*, 2001). However, the low-frequency signals in those micrographs are small due to the CTF modulation. This makes accurate initial orientation estimation problematic and results in a low yield of particles acceptable for inclusion in the 3-D reconstruction. To overcome this problem in SAVR, Wavelet transform-assisted projection matching is employed. The Wavelet transform is different from conventional Fourier low-pass filtering in that it not only suppresses noise but also causes less smearing of the details of the object (Starck *et al.*, 1998). Wavelet transform prefiltering combined with real-space projection matching provides an alternative to the conventional self-common lines method for estimating initial particle orientation, which is able to identify more particles and make a more accurate estimate of their orientation parameters (Saad *et al.*, manuscript in preparation). For the final refinement and reconstruction, however, the original images but not the Wavelet-transformed images are used. The Wavelet method has been found to be essential in estimating the initial orientation of particles in some data sets for which the conventional self-common line method has completely failed (e.g., Paredes *et al.*, manuscript in preparation).

Automatic tracking of corresponding particles in "focal pairs." The "focal pair" method is often used to estimate the initial orientations of particles in micrographs imaged under very close-to-focus conditions by first analyzing the corresponding particle images in the second far-from-focus micrograph. SAVR simplifies this procedure by automatically transferring the orientation estimations of particle images in the far-from-focus micrograph to the corresponding particle images in the close-to-focus one. This automation has eliminated the error-prone process of manual tracking. In SAVR, a data set could include a mixture of "focal pair" micrographs and single micrographs.

Automatic "multiple sampling" refinement. For a project targeted to high resolution, the sampling step size must be sufficiently fine. The common practice is that the targeted resolution should be about three times the image sampling resolution. In the early steps of the analysis, only the low-resolution (30–15 Å) data are important. Therefore, a coarser sampling can be used. It is advantageous if the images can be down-sampled for the early iteration steps and then switched back to the original fine sampling for the later refinement steps, as this greatly reduces the processing time. SAVR implements this "multiple sampling" refinement scheme and is able to automatically down-sample the particle images to the appropriate sampling levels at different stages, carrying out all the processing iterations at variable sampling step sizes. SAVR performs all the needed operations automatically if the "multiple sampling" option is enabled.

Real-space 3-D filtering. The density values for the solvent should ideally be uniform and featureless. Density modification by solvent flattening is a common and valuable method for improving the density maps during refinement for X-ray crystallography (Terwilliger, 1999). Solvent flattening has also been applied to electron cryomicroscopy of tubular crystals (Yonekura and Toyoshima, 2000). Solvent flattening improves the density map by defining the mask for the solvent and setting the density values for the solvent to a constant. For the 3-D reconstructions from each iteration, SAVR implements a real-space 3-D filtering that is equivalent conceptually to "solvent flattening" by masking out the solvent and background densities according to the icosahedral geometry and applying user-selected thresholds.

Truly independent refinement and reconstruction. The common criterion for evaluating the correctness and resolution of a reconstruction is to split the image set into two halves and compare the reconstructions from these two half image sets. Unfortunately, the split of data was often done at the stage just before final 3-D reconstruction step, meaning all images are refined together against the same 3-D model. The resulting reconstructions are therefore not truly independent. Truly independent refinement and reconstruction must split the data set at the very beginning and then proceed through the processing independently. The 3-D reconstructions from this truly independent refinement serve as the most stringent criterion to test the reproducibility of the reconstructions and the resolution. SAVR has implemented an option allowing even split of the data set at the very beginning and processes each of the data subsets completely independently. Alternatively, one may choose to split the data just before the 3-D reconstruction step as commonly practiced. However, the option available in SAVR will make it easy to practice a "truly independent" and theoretically a more proper procedure.

Information Input

There are two classes of information needed for SAVR to start the processing. The first concerns the intrinsic features of the data: size of the particle, microscope parameters, sampling stepsize, boxed-out particle image file paths for each micrograph, and CTF parameters. Since none of the automatic particle selection algorithms is robust enough, it is presently not included in SAVR. The particle images are preselected outside of SAVR and input into SAVR as boxed-out particles in a separate file for each micrograph. We normally use IMAGIC format for the input images, but the format could be any of the commonly used ones, including MRC, Spider, and IMAGIC, supported through the EMAN library. The second is the processing options that control various aspects of SAVR processing, including number of CPUs, starting model file path, choice of initial orientation determination method, etc. This information is optional, but provides the user with the option of modifying the default behaviors provided under SAVR.

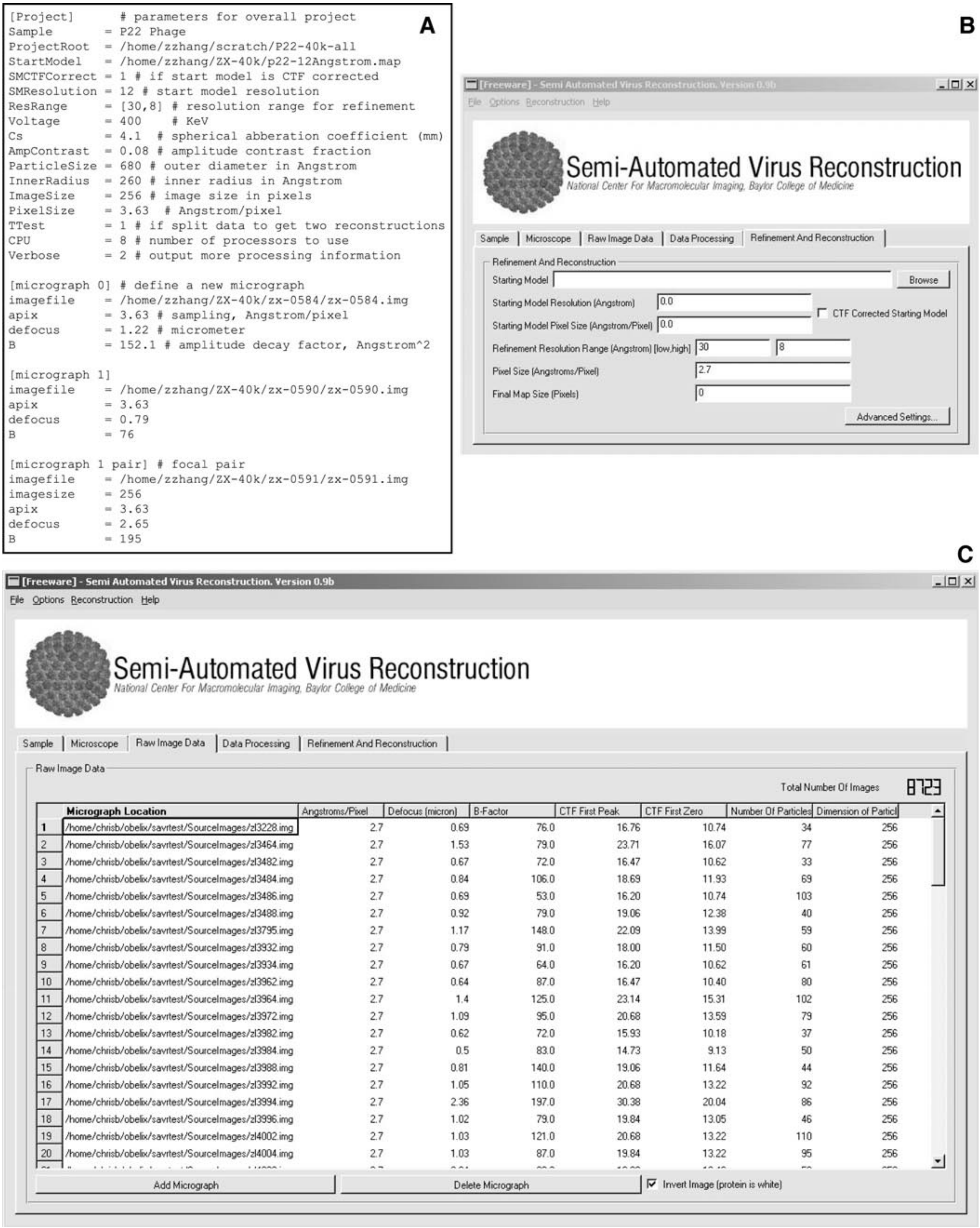
Information can be entered in two different ways. The first is through a simple text file. Figure 3A shows a template/example file with easily understandable comments for each parameter. Only the values on the right side of the equal signs should be modified for a new data set. It is worthy of note that while the CTF parameters (defocus and experimental B-factor) were shown as user input for each micrograph to SAVR, they could be fitted by SAVR automatically using the *fitctf* program provided by EMAN. Alternatively, data entry can be done manually through a graphic user interface program *gui_virus.py* (Figs. 3B and 3C).

SAVR Outputs

The main output of SAVR processing is 3-D maps. 3-D reconstruction is performed after each refinement cycle and serves as the starting model for the next iteration.

A preexisting model is not always available, making the building of an initial model before proceeding to the refinement iterations an essential first task. If no initial model is provided, SAVR builds one automatically using the self-common lines algorithm. Basically, SAVR emulates the way a user would build an initial model but has the advantages of being more patient, more quantitative, and exhaustively covering the candidate particles. SAVR sorts all micrographs according to the first peak frequency of the CTF and chooses those micrographs where the signal-to-noise ratio is higher at low resolution. SAVR then runs the self-common lines algorithm (*ortAll*), parses through all the top orientation candidates, and ranks the particles according to how well the different phase residual criteria agree. The best particles are used to build the first 3-D model, which is then used for the later refinement.

Python is also used to parse the outputs of the main processing programs in order to compute pro-


A
B
C
[Freeware] - Semi Automated Virus Reconstruction. Version 0.9b
File Options Reconstruction Help
Semi-Automated Virus Reconstruction
National Center For Macromolecular Imaging, Baylor College of Medicine
Sample
Microscope
Raw Image Data
Data Processing
Refinement And Reconstruction
Browse
Advanced Settings...
Add Micrograph
Delete Micrograph

```
[Project]      # parameters for overall project
Sample       = P22 Phage
ProjectRoot  = /home/zzhang/scratch/P22-40k-all
StartModel   = /home/zzhang/ZX-40k/p22-12Angstrom.map
SMCTFCorrect = 1 # if start model is CTF corrected
SMResolution = 12 # start model resolution
ResRange     = [30,8] # resolution range for refinement
Voltage      = 400    # KeV
Cs           = 4.1  # spherical abberation coefficient (mm)
AmpContrast  = 0.08 # amplitude contrast fraction
ParticleSize = 680 # outer diameter in Angstrom
InnerRadius  = 260 # inner radius in Angstrom
ImageSize    = 256 # image size in pixels
PixelSize    = 3.63  # Angstrom/pixel
TTest        = 1 # if split data to get two reconstructions
CPU          = 8 # number of processors to use
Verbose      = 2 # output more processing information

[micrograph 0] # define a new micrograph
imagefile    = /home/zzhang/ZX-40k/zx-0584/zx-0584.img
apix         = 3.63 # sampling, Angstrom/pixel
defocus      = 1.22 # micrometer
B            = 152.1 # amplitude decay factor, Angstrom^2

[micrograph 1]
imagefile    = /home/zzhang/ZX-40k/zx-0590/zx-0590.img
apix         = 3.63
defocus      = 0.79
B            = 76

[micrograph 1 pair] # focal pair
imagefile    = /home/zzhang/ZX-40k/zx-0591/zx-0591.img
imagesize    = 256
apix         = 3.63
defocus      = 2.65
B            = 195
```

Refinement And Reconstruction

Starting Model:

Starting Model Resolution (Angstrom): 0.0

Starting Model Pixel Size (Angstrom/Pixel): 0.0

☐ CTF Corrected Starting Model

Refinement Resolution Range (Angstrom) [low,high]: 30 8

Pixel Size (Angstroms/Pixel): 2.7

Final Map Size (Pixels): 0

Raw Image Data

Total Number Of Images 8723

	Micrograph Location	Angstroms/Pixel	Defocus (micron)	B-Factor	CTF First Peak	CTF First Zero	Number Of Particles	Dimension of Particl
1	/home/chrisb/obelix/savrtest/SourceImages/zl3228.img	2.7	0.69	76.0	16.76	10.74	34	256
2	/home/chrisb/obelix/savrtest/SourceImages/zl3464.img	2.7	1.53	79.0	23.71	16.07	77	256
3	/home/chrisb/obelix/savrtest/SourceImages/zl3482.img	2.7	0.67	72.0	16.47	10.62	33	256
4	/home/chrisb/obelix/savrtest/SourceImages/zl3484.img	2.7	0.84	106.0	18.69	11.93	69	256
5	/home/chrisb/obelix/savrtest/SourceImages/zl3486.img	2.7	0.69	53.0	16.20	10.74	103	256
6	/home/chrisb/obelix/savrtest/SourceImages/zl3488.img	2.7	0.92	79.0	19.06	12.38	40	256
7	/home/chrisb/obelix/savrtest/SourceImages/zl3795.img	2.7	1.17	148.0	22.09	13.99	59	256
8	/home/chrisb/obelix/savrtest/SourceImages/zl3932.img	2.7	0.79	91.0	18.00	11.50	60	256
9	/home/chrisb/obelix/savrtest/SourceImages/zl3934.img	2.7	0.67	64.0	16.20	10.62	61	256
10	/home/chrisb/obelix/savrtest/SourceImages/zl3962.img	2.7	0.64	87.0	16.47	10.40	80	256
11	/home/chrisb/obelix/savrtest/SourceImages/zl3964.img	2.7	1.4	125.0	23.14	15.31	102	256
12	/home/chrisb/obelix/savrtest/SourceImages/zl3972.img	2.7	1.09	95.0	20.68	13.59	79	256
13	/home/chrisb/obelix/savrtest/SourceImages/zl3982.img	2.7	0.62	72.0	15.93	10.18	37	256
14	/home/chrisb/obelix/savrtest/SourceImages/zl3984.img	2.7	0.5	83.0	14.73	9.13	50	256
15	/home/chrisb/obelix/savrtest/SourceImages/zl3988.img	2.7	0.81	140.0	19.06	11.64	44	256
16	/home/chrisb/obelix/savrtest/SourceImages/zl3992.img	2.7	1.05	110.0	20.68	13.22	92	256
17	/home/chrisb/obelix/savrtest/SourceImages/zl3994.img	2.7	2.36	197.0	30.38	20.04	86	256
18	/home/chrisb/obelix/savrtest/SourceImages/zl3996.img	2.7	1.02	79.0	19.84	13.05	46	256
19	/home/chrisb/obelix/savrtest/SourceImages/zl4002.img	2.7	1.03	121.0	20.68	13.22	110	256
20	/home/chrisb/obelix/savrtest/SourceImages/zl4004.img	2.7	1.03	87.0	19.84	13.22	95	256

☑ Invert Image (protein is white)

FIG. 3. SAVR input. Information can be entered into SAVR through (A) a text file or (B) a graphic user interface. The graphical interface also has a means to select multiple micrographs for processing (C).

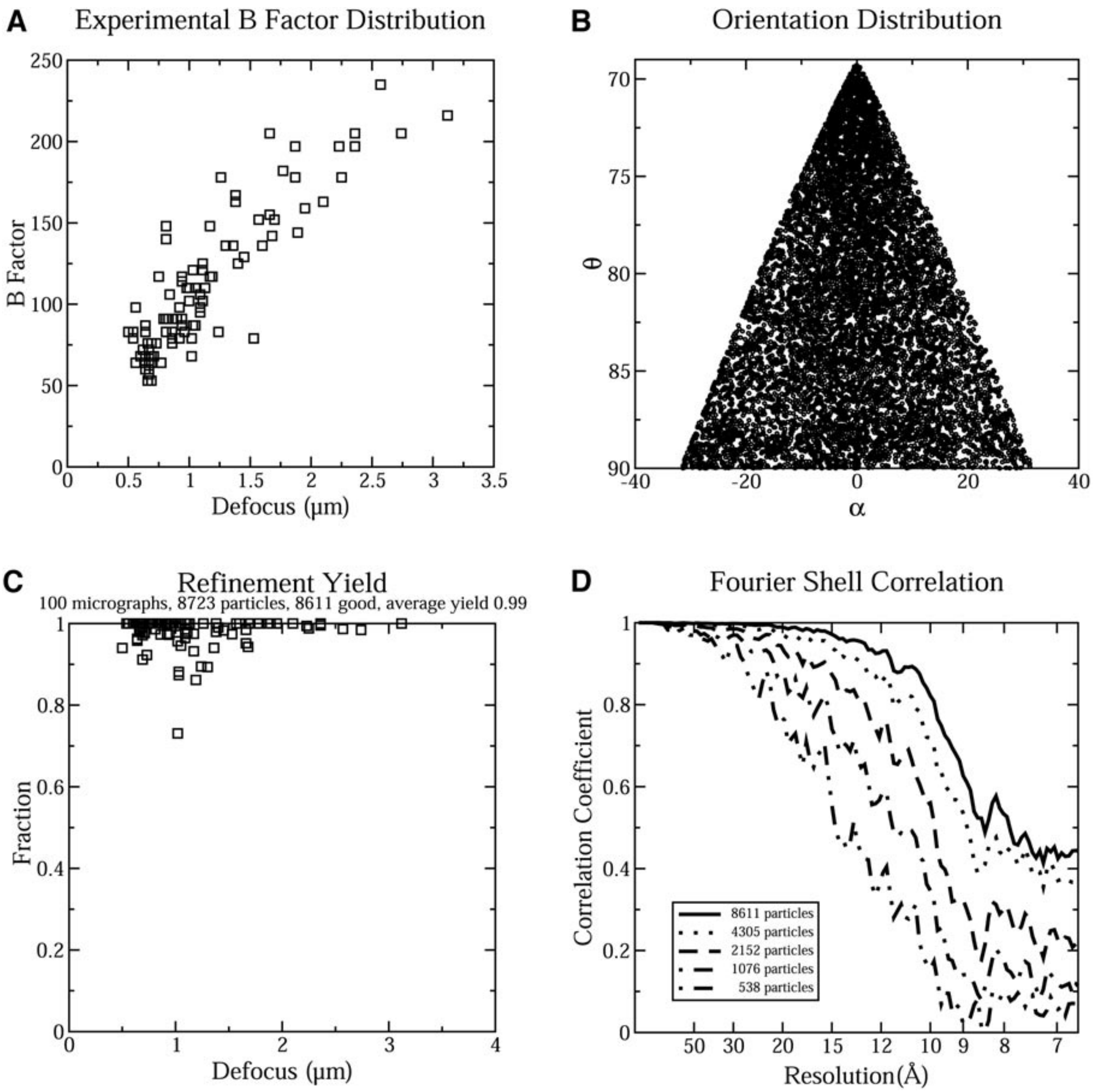
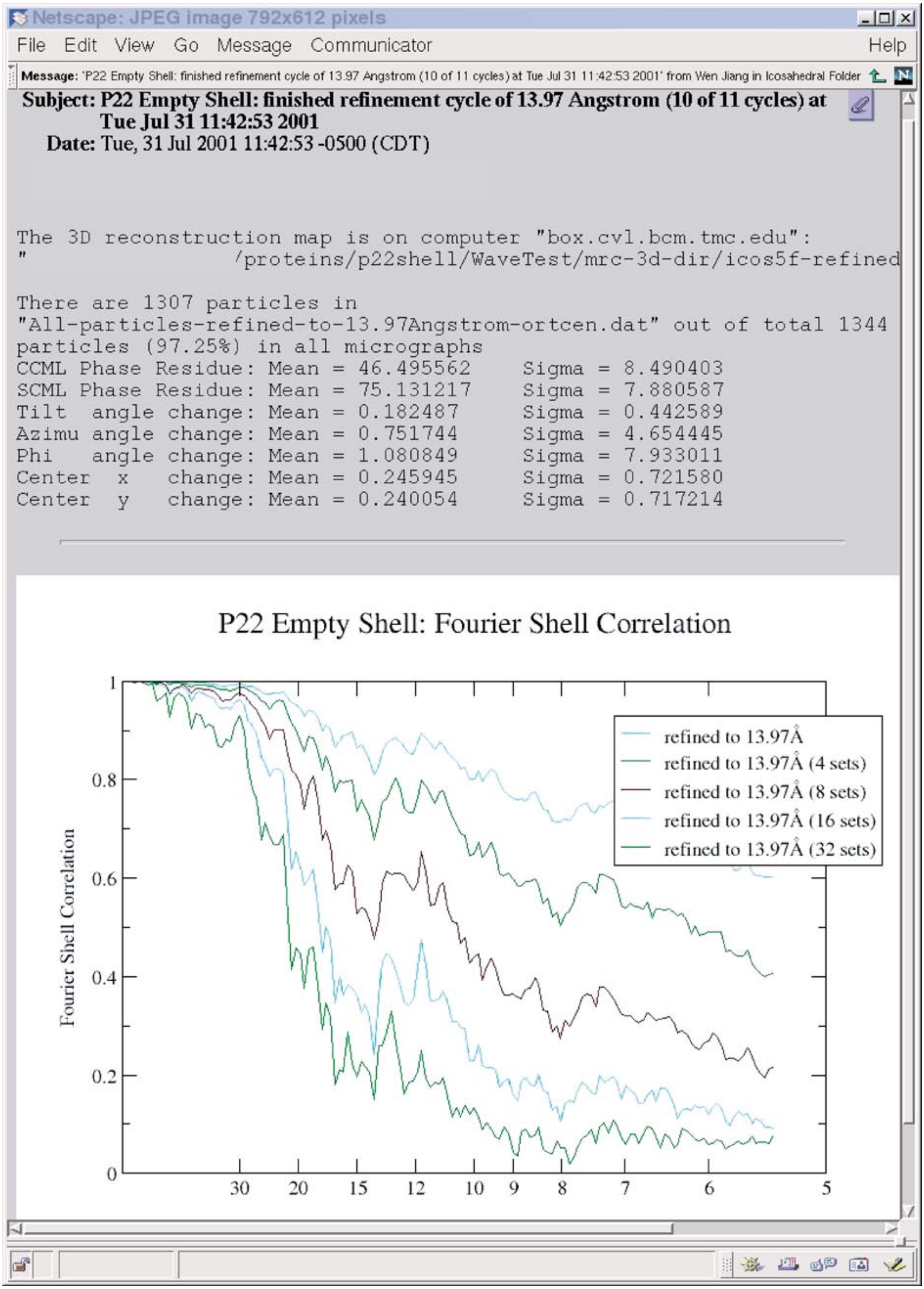


FIG. 4. Examples of SAVR-generated statistics. Several statistical measures are generated by SAVR automatically. Shown are experimental B-factor distributions at different defocuses (A), particle orientation distribution (B), "good" particle yield of micrographs at different defocuses (C), and Fourier shell correlation curve of two 3-D maps reconstructed from half data sets (D).

cessing statistics. For example, micrograph defocus and experimental B-factor histograms (Fig. 4A), particle orientation distributions (Fig. 4B), fraction of particle yield in each micrograph (Fig. 4C), and Fourier shell correlation of two independent reconstructions (Fig. 4D) are implemented. The example shown in Fig. 4C with the P22 procapsid shell shows a nearly 100% yield of boxed particles accepted to be included in the final 3-D reconstruction. It should be noted that the yield depends on particle structural homogeneity, image quality, particle selection, and image processing method. We have also seen much lower yields for other particle data sets.

Another very convenient SAVR output is the automatic e-mailing of the state of progress to a specified address (Fig. 5). Once started, SAVR processes all the data without user intervention, meaning that it might take days or weeks to finish the processing for large data sets. Therefore, it could prove difficult for a user to monitor the progress of the processing. However, an option is built into SAVR to e-mail the user after each stage of processing, the mean/sigma of phase residuals, the orientation and center parameter changes, and the Fourier shell correlation curve. This allows the user to keep track of processing progress just by checking e-mail regularly.


Netscape: JPEG image 792x612 pixels
File
Edit
View
Go
Message
Communicator
Help
P22 Empty Shell: Fourier Shell Correlation
Fourier Shell Correlation
0
0.2
0.4
0.6
0.8
1
30
20
15
12
10
9
8
7
6
5
refined to 13.97Å
refined to 13.97Å (4 sets)
refined to 13.97Å (8 sets)
refined to 13.97Å (16 sets)
refined to 13.97Å (32 sets)

Message: 'P22 Empty Shell: finished refinement cycle of 13.97 Angstrom (10 of 11 cycles) at Tue Jul 31 11:42:53 2001' from Wen Jiang in Icosahedral Folder

Subject: P22 Empty Shell: finished refinement cycle of 13.97 Angstrom (10 of 11 cycles) at Tue Jul 31 11:42:53 2001
Date: Tue, 31 Jul 2001 11:42:53 -0500 (CDT)

```
The 3D reconstruction map is on computer "box.cvl.bcm.tmc.edu":
"                 /proteins/p22shell/WaveTest/mrc-3d-dir/icos5f-refined

There are 1307 particles in
"All-particles-refined-to-13.97Angstrom-ortcen.dat" out of total 1344
particles (97.25%) in all micrographs
CCML Phase Residue: Mean = 46.495562    Sigma = 8.490403
SCML Phase Residue: Mean = 75.131217    Sigma = 7.880587
Tilt  angle change: Mean = 0.182487     Sigma = 0.442589
Azimu angle change: Mean = 0.751744     Sigma = 4.654445
Phi   angle change: Mean = 1.080849     Sigma = 7.933011
Center  x   change: Mean = 0.245945     Sigma = 0.721580
Center  y   change: Mean = 0.240054     Sigma = 0.717214
```

FIG. 5. Automated notification. As a convenience feature, SAVR notifies the user of the status of its processing. A snapshot of this e-mail report is shown. Personal account information has been masked from the original image.

Applications

SAVR has been applied to two large data sets of bacteriophage P22 (Zhang *et al.*, 2000) at two different maturation stages. Both image data sets were taken at 400 keV on a JEOL 4000 microscope and processed on a SGI Onyx2 supercomputer. The resolution criterion was 0.5 threshold (Böttcher *et al.*, 1997) for Fourier shell correlation of two independent reconstructions (Harauz and van Heel, 1986). The mature phage images were taken at × 40 000 magnification and scanned at 3.63 Å/pixel; 25 single micrographs (~5000 particles sampled at 256 × 256 pixels) were refined and reconstructed to 10.9 Å resolution within 3 days using eight R10000 processors. The P22 procapsid shell images were taken at × 50 000 magnification and scanned at 2.7 Å/pixel. The data set includes 100 micrographs (8723 particles sampled at 300 × 300 pixels; 53 micrographs had a defocus less than 1 μm and were processed as focal pairs, with the remaining 47 micrographs having a larger defocus and processed as single micrographs). These images were refined and reconstructed to 8.5 Å resolution in less than 2 weeks of processing time using eight processors. The resulting reconstructions showed clearly the secondary structure elements and significant conformational changes that occurred during the phage maturation process (Jiang *et al.*, manuscript in preparation).

DISCUSSION

Automation of Icosahedral Reconstruction and Refinement at Sub-nanometer Resolution

Over the years, many programs for icosahedral reconstruction have been developed and we have been able to obtain 3-D reconstructions in the subnanometer resolution range (6.8–10 Å). The data set for structure determination of an icosahedral particle of a moderate size (500–1200 Å in diameter) at this resolution typically needs 80–200 focal pairs of micrographs containing several thousand particle images. However, the image processing and 3-D reconstruction procedures are rather involved, requiring several months to a year of effort from an experienced investigator. Significantly more data will be needed if a higher resolution reconstruction is pursued, making the slow pace of existing methods of processing a serious bottleneck.

The current interactive processing mode requires the user to manually organize the particles into different file directories and then processes the images in the directories individually during each iteration, repeating the interactive processing for many iterative cycles to reach a refined reconstruction. Such repetitive interaction not only demands great patience and energy from the user, but, more seriously, is also very error-prone. Large data sets also demand optimal use of limited computing resources. This cannot be achieved with user-dependent processing, which invariably results in significant time waste during the idle intervals between the iterations and processing of different micrographs. By implementing SAVR, which completely automates the initial model building and iterative structural refinements, we have eliminated both user error and the problems arising from manual resource management.

SAVR also takes advantage of modern parallel computing to speed up the automated processing. Parallel processing with SAVR on eight processors of a SGI Onyx2 computer system (R10000 195 Mhz) has allowed us to generate two moderately high-resolution (~8.5 Å) reconstructions in about 2 weeks, where previously they would have taken months to a year to complete by inexperienced investigators. It is likely that the time needed would be further reduced if more processors and newer chips were used. While the two test data sets presented were both processed on an SGI Onyx2 shared memory platform, SAVR has been designed to be platform independent. We have also been able to run SAVR on a dramatically different multiple processor systems, namely, a distributed memory Linux cluster system.

The application of SAVR to two data sets of bacteriophage P22 showed that SAVR is indeed able to achieve the intended high throughput, automated processing of the critical steps in icosahedral reconstruction. The processing of both data sets yielded 3-D reconstructions to the practical resolution limits inherent in the image samplings. The time taken after data collection was much shorter than would have been the case using our traditional approach. Such significant progress in shortening the processing time for reconstruction means that the time-limiting factor in achieving a moderately high-resolution 3-D structure has shifted to the rate of data collection and time taken in structural interpretation. The automatic data collection methods now being developed for some electron microscopes will further enhance this progress and allow the entire project to be completed more easily and quickly.

It is always a serious concern that automation could result in overlooking potential problems or mistakes during the processing, especially for an inexperienced user. This is an unavoidable dilemma when any automation is attempted. The danger of overlooking mistakes could be effectively avoided by examining extensive statistical and other evaluations of the intermediate and final results. For example, a Fourier shell correlation test is performed after each cycle to monitor the quality of the reconstruction. Various statistics shown in Fig. 4 are also useful in detecting any problems at the earliest stage of analysis. Of course, the user should have enough knowledge of image processing to interpret these statistical results.

SAVR gives the user control of various aspects of the processing by providing means to override almost all the processing parameters and adapt to special needs when dealing with some less well behaved data sets. For example, the user could choose the algorithm for initial orientation determination, change the phase residue threshold, search step sizes for orientation refinement, select data to include in processing based on criteria of focal pair, defocus ranges, and experimental-B factor values, etc. For the advanced user, more controls could be added by changing the actual Python scripts controlling the automation. These extra controls provide convenient means for dealing with some less well behaved data sets for which the default processing might have failed.

Toward Complete Solution for Electron Cryomicroscopy

The current scope of SAVR integrates the image processing algorithms that we believe are mature enough to be free of user interaction. The relevant processing steps start with boxed particle images from each micrograph and end with a refined 3-D reconstruction (Fig. 1). Due to the design principles employed, it will not be hard to expand the scope of SAVR further and include even more related steps to maximally automate electron cryomicroscopy, image processing, and 3-D reconstruction.

A natural extension of SAVR would be to include automatic particle selection so that once the images are scanned from micrographs or acquired through a CCD, SAVR would be able to perform all the processing steps needed to generate a refined 3-D reconstruction to high resolution. Automatic boxing is not yet a robust enough technique even though many algorithms have been proposed (Boier Martin *et al.*, 1997; Harauz and Fong-Lochovsky, 1989; Kivioja *et al.*, 2000; Lata *et al.*, 1995; Ludtke *et al.*, 1999; Nicholson and Glaeser, 2001; Saad *et al.*, 1998; Thuman-Commike and Chiu, 1996). In practice, semiautomatic boxing is used for initial boxing followed by human interactive screening. However, the design of SAVR allows easy inclusion of any new program for automatic boxing when it becomes robust enough. An extremely attractive scenario would be automated data collection on a CCD (Carragher *et al.*, 2000; Zhang *et al.*, 2001) with the CCD images sent immediately to SAVR for processing. This would bring closer the realization of the ultimate goal of “sample in, structure out” electron cryomicroscopy. We are currently developing all the elements needed to achieve this goal, including automated data acquisition by CCD, a database for electron cryomicroscopy project management, automatic particle boxing, CTF parameter fitting, and the interfaces between these elements.

The extension of SAVR to post-reconstruction analysis would mean including some structural interpretation functionalities. While structural interpretation requires much more intelligence and is more subjective than data processing, there are some tangible extensions, which can be included in SAVR. SAVR has demonstrated the capability of achieving the resolution range (7–10 Å) in which secondary structures can be discerned. At these resolutions, the *helixhunter* program (Jiang *et al.*, 2001) can be used to identify helices present in the final reconstruction. Incorporation of *helixhunter* into SAVR would allow this to be done automatically, from which the helices’ coordinates could then be automatically used to probe the library of helices from all known structures to search for homologous or novel folds (Zhou *et al.*, 2001).

SAVR Provides a Platform for Further Development Toward Atomic Resolution

To approach high resolution at 3–4 Å, it is almost certain that the current algorithms used in refinement and 3-D reconstruction will need further improvements. For instance, independent CTF refinements for particles in different areas of the same micrograph (van Heel *et al.*, 2000) are likely to be needed. Testing new algorithms on such large data sets would be daunting if not impossible using interactive processing methods. SAVR provides a platform where further developments on each of the underlying algorithms can be easily tested, making the assessment of new ideas practical, even on a very large data set.

In the current implementation, we have already taken advantage of the development platform provided by SAVR. The user options in SAVR, such as choice of initial orientation determination method (the conventional self-common line algorithm or the newly developed Wavelet transform-assisted projection matching technique), the choice of single micrographs or focal pairs for initial orientation determination, defocus range selection, real-space 3-D filtering, and multiple sampling are just a few of the examples designed and developed under the SAVR umbrella.

Image Processing for Biologists

Electron cryomicroscopy has proven to be a very useful technology for solving structures of various macromolecular assemblies. Use of and demand for electron cryomicroscopy as a generally applicable tool has been increasing, but electron cryomicroscopy is still only practiced by dedicated electron microscopists and image processing experts. It has yet to be as broadly employed by nonspecialist biologists as are some other advanced and mature technologies. Although it is not the only barrier to wider use, the current methods of interactive reconstruction require too much effort for a novice user to become suitably proficient in their use. SAVR has been designed to automate image processing and 3-D reconstruction by integrating the knowledge and experience we have gained over many years. As a result, SAVR is easy for users to master and allows the novice user to carry out the critical and difficult steps in icosahedral reconstruction to a sub-nanometer resolution. Our aim is to make SAVR an image processing workbench in the effort to transform icosahedral reconstruction and electron cryomicroscopy from a specialist enterprise to a routine practice for molecular biologists. Such an approach can be taken for processing images of other macromolecules with and without symmetry.

We thank Dr. Z. Hong Zhou for the self-common lines, cross-common lines refinement, and icosahedral reconstruction programs in IMIRS, Dr. Ali Saad for the Wavelet-based projection matching algorithm, and Dr. Steven J. Ludtke for the EMAN library and programs. We also thank Dr. Peter E. Prevelige for providing the high-quality P22 bacteriophage samples and Drs. Steven J. Ludtke, Frazer Rixon, Michael F. Schmid, and Z. Hong Zhou for critical reading of the manuscript. This research has been supported by grants from the National Institutes of Health (R01A143656 and P41RR02250), the Robert Welch Foundation, the Agouron Foundation, and the National Partnership for Advanced Computational Infrastructure (NPACI), through the support of the National Science Foundation.

REFERENCES

Baker, T. S., Olson, N. H., and 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.

Boier Martin, I. M., Marinescu, D. C., Lynch, R. E., and Baker, T. S. (1997) Identification of spherical virus particles in digitized images of entire electron micrographs. *J. Struct. Biol.* **120,** 146–157.

Böttcher, B., Wynne, S. A., and Crowther, R. A. (1997) Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature* **386,** 88–91.

Carragher, B., Kisseberth, N., Kriegman, D., Milligan, R. A., Potter, C. S., Pulokas, J., and Reilein, A. (2000) Leginon: An automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* **132,** 33–45.

Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J., and Steven, A. C. (1997) Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature* **386,** 91–94.

Crowther, R. A. (1971) Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. *Philos. Trans. R. Soc. London, Ser. B Biol. Sci.* **261,** 221–230.

DeRosier, D. L., and Klug, A. (1968) Reconstruction of three-dimensional structures from electron micrographs. *Nature* **217,** 130–134.

Fuller, S. D. (1987) The T = 4 envelope of Sindbis virus is organized by interactions with a complementary T = 3 capsid. *Cell* **48,** 923–934.

Harauz, G., and Van Heel, M. (1986) Exact filters for general geometry three dimensional reconstruction. *Optik* **73,** 146–156.

Harauz, G., and Fong-Lochovsky, A. (1989) Automatic selection of macromolecules from electron micrographs by component labelling and symbolic processing. *Ultramicroscopy* **31,** 333–344.

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 cryo-microscopy. *J. Mol. Biol.* **213,** 899–929.

Huang, C. C., Novak, W. R., Babbitt, P. C., Jewett, A. I., Ferrin, T. E., and Klein, T. E. (2000) Integrated tools for structural and sequence alignment and analysis. *Pac. Symp. Biocomput.* 230–241.

Jiang, W., Baker, M. L., Ludtke, S. J., and Chiu, W. (2001) Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308,** 1033–1044.

Kimura, Y., Vassylyev, D. G., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T., and Fujiyoshi, Y. (1997) Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature* **389,** 206–211.

Kivioja, T., Ravantti, J., Verkhovsky, A., Ukkonen, E., and Bamford, D. (2000) Local average intensity-based method for identifying spherical particles in electron micrographs. *J. Struct. Biol.* **131,** 126–134.

Kühlbrandt, W., Wang, D. N., and Fujiyoshi, Y. (1994) Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367,** 614–621.

Lata, K. R., Penczek, P., and Frank, J. (1995) Automatic particle picking from electron micrographs. *Ultramicroscopy* **58,** 381–391.

Ludtke, S. J., Baldwin, P. R., and Chiu, W. (1999) EMAN: Semiautomated software for high resolution single particle reconstructions. *J. Struct. Biol.* **128,** 82–97.

Mancini, E. J., Clarke, M., Gowen, B. E., Rutten, T., and Fuller, S. D. (2000) Cryo-electron microscopy reveals the functional organization of an enveloped virus, Semliki Forest virus. *Mol. Cell* **5,** 255–266.

Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B., Engel, A., and Fujiyoshi, Y. (2000) Structural determinants of water permeation through aquaporin-1. *Nature* **407,** 599–605.

Nicholson, W. V., and Glaeser, R. M. (2001) Review: Automatic particle detection in electron microscopy. *J. Struct. Biol.* **133,** 90–101.

Nogales, E., Wolf, S. G., and Downing, K. H. (1998) Structure of the alpha beta tubulin dimer by electron crystallography. *Nature* **391,** 199–203.

Ramu, C., Gemund, C., and Gibson, T. J. (2000) Object-oriented parsing of biological databases with Python. *Bioinformatics* **16,** 628–638.

Saad, A., Chiu, W., and Thuman-Commike, P. (1998) Multiresolution approach to automatic detection of spherical particles from electron cryomicroscopy images. *In* IEEE Signal Processing Society International Conference on Image Processing, pp. 8647–8651, Chicago.

Saad, A., Ludtke, S. J., Jakana, J., Rixon, F. J., Tsuruta, H., and Chiu, W. (2001) Fourier amplitude decay of electron cryomicroscopic images of single particles and effects on structure determination. *J. Struct. Biol.* **133,** 32–42.

Sanner, M. F. (1999) Python: A programming language for software integration and development. *J. Mol. Graph. Model* **17,** 57–61.

Starck, J. L., Murtagh, F., and Bijaoui, A. (1998) Image Processing and Data Analysis: The Multiscale Approach, Cambridge Univ. Press, Cambridge, UK.

Terwilliger, T. C. (1999) Reciprocal-space solvent flattening. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1863–1871.

Thuman-Commike, P. A., and Chiu, W. (1996) PTOOL: A software package for the selection of particles from electron cryomicroscopy spot-scan images. *J. Struct. Biol.* **116,** 41–47.

Thuman-Commike, P. A., and Chiu, W. (1997) Improved common line-based icosahedral particle image orientation estimation algorithms. *Ultramicroscopy* **68,** 231–255.

Thuman-Commike, P. A., and Chiu, W. (2000) Reconstruction principles of icosahedral virus structure determination using electron cryomicroscopy. *Micron* **31,** 687–711.

Trus, B. L., Roden, R. B., Greenstone, H. L., Vrhel, M., Schiller, J. T., and Booy, F. P. (1997) Novel structural features of bovine papillomavirus capsid revealed by a three-dimensional reconstruction to 9 Å resolution. *Nat. Struct. Biol.* **4,** 413–420.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., and Patwardhan, A. (2000) Single-particle electron cryomicroscopy: Towards atomic resolution. *Q. Rev. Biophys.* **33,** 307–369.

Yonekura, K., and Toyoshima, C. (2000) Structure determination of tubular crystals of membrane proteins. III. Solvent flattening. *Ultramicroscopy* **84,** 29–45.

Zhang, P., Beatty, A., Milne, J. L., and Subramaniam, S. (2001) Automated data collection with a tecnai 12 electron microscope: Applications for molecular imaging by cryomicroscopy. *J. Struct. Biol.* **135,** 251–261.

Zhang, Z., Greene, B., Thuman-Commike, P. A., Jakana, J., Prevelige, P. E., Jr., King, J., and Chiu, W. (2000) Visualization of the maturation transition in bacteriophage P22 by electron cryomicroscopy. *J. Mol. Biol.* **297,** 615–626.

Zhou, Z. H., Prasad, B. V., Jakana, J., Rixon, F. J., and Chiu, W. (1994) Protein subunit structures in the herpes simplex virus A-capsid determined from 400 kV spot-scan electron cryomicroscopy. *J. Mol. Biol.* **242,** 456–469.

Zhou, Z. H., Chiu, W., Haskell, K., Spears, H., Jr., Jakana, J., Rixon, F. J., and Scott, L. R. (1998) Refinement of herpesvirus B-capsid structure on parallel supercomputers. *Biophys. J.* **74,** 576–588.

Zhou, Z. H., Dougherty, M., Jakana, J., He, J., Rixon, F. J., and Chiu, W. (2000) Seeing the herpesvirus capsid at 8.5 A. *Science* **288,** 877–880.

Zhou, Z. H., Baker, M. L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G., and Chiu, W. (2001) Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nat. Struct. Biol.* **8,** 868–873.