Data Quality Problems in PDB Virus Models

Tom Goddard
October 11, 2003
Workshop on Visualization of Biological Complexes (20 minutes).

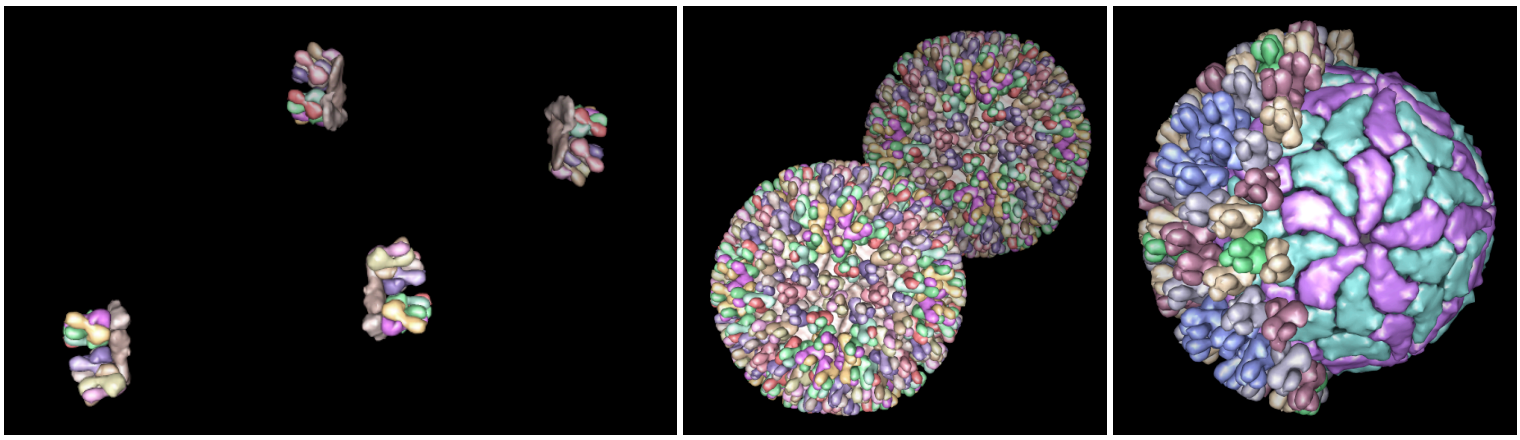
Viruses and Matrices

- 206 virus capsid entries in PDB: 163 crystallography, 23 fiber diffraction, 19 electron microscopy.
- Particles are icosahedral (60-fold symmetry) or helical.
- PDB model provides atomic coordinates for asymmetric unit.
- Matrices position asymmetric units to make "biological unit".
- For crystal structures, matrices give unit cell.
- Crystal unit cell is obtained by multiplying crystal symmetry matrices times non-crystal symmetry matrices.
- About half the virus models the matrices are missing, or not machine-readable, or are incorrect.
- Will show examples of problems and discuss how to fix them.

Examples of Problems in PDB Virus Entries

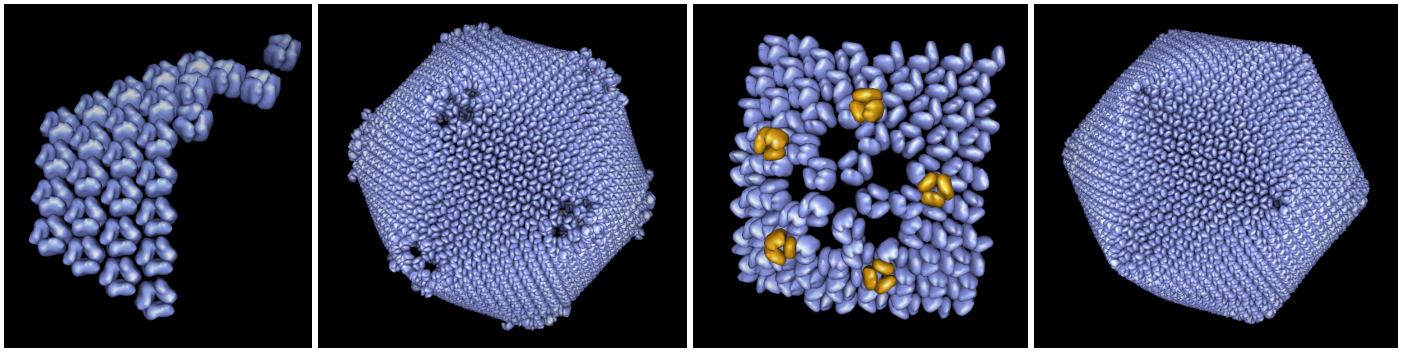
Bluetongue Virus (2btv, sept 1998)

- Matrices for biological unit not provided.
- Crystal unit cell contains 2 virus particles.
- Non-crystallographic symmetry matrices have incorrect "given" flag.
- Sub-structures (trimer, inner and outer layers) not described.
- Chimera demo used a Python script to specify matrices.



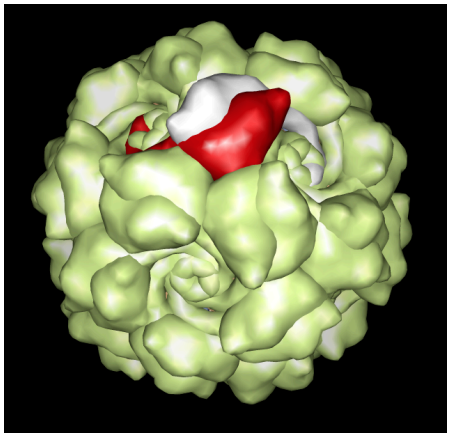
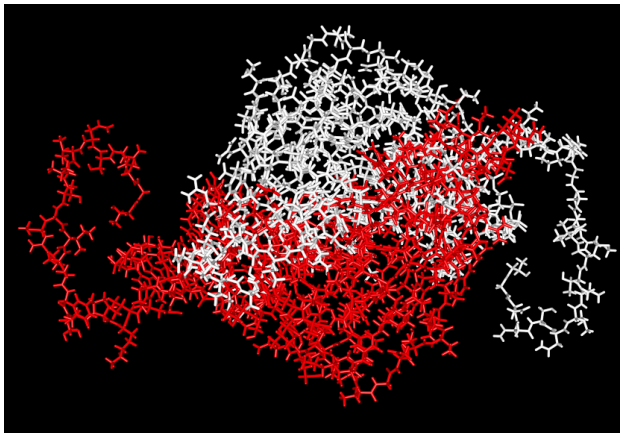
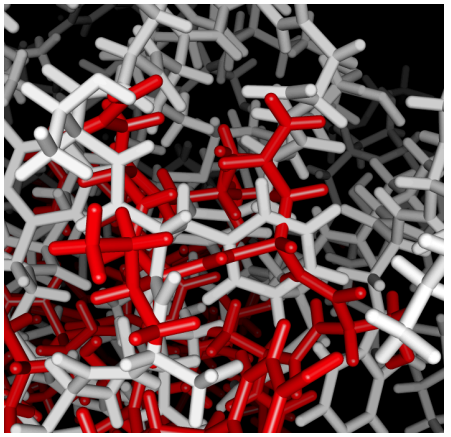
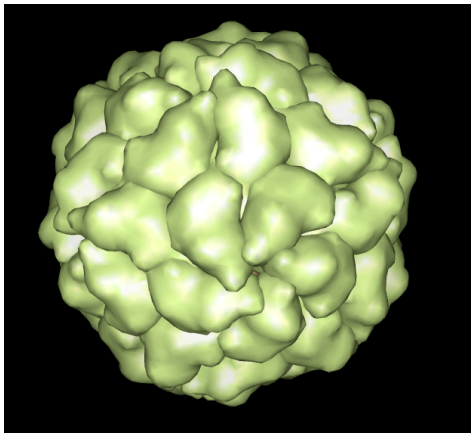
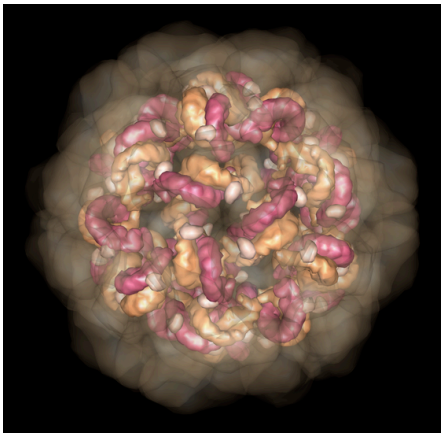
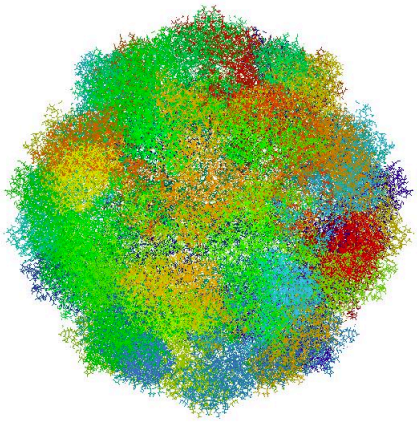
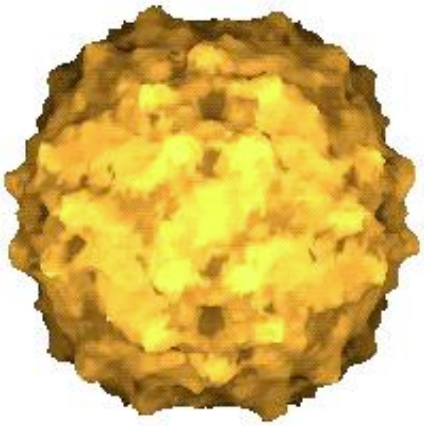
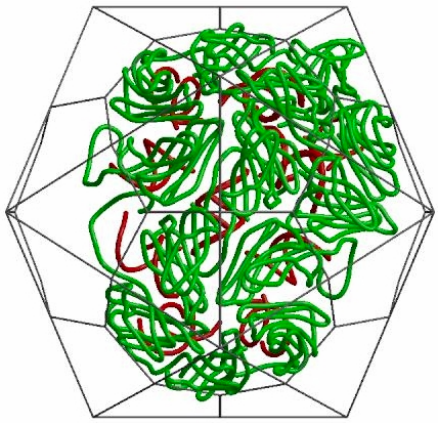
Paramecium Bursaria Chlorella (PBC) Virus (1m4x, july 2002)

- This entry fits a crystal structure to a cryo-EM density map.
- Provides human-readable instructions for multiplying together 28 explicit matrices and 2-fold, 3-fold, 5-fold and 2-fold icosahedral symmetry matrices to produce 1680 matrices.
- 60 subunits are in the wrong place because of a matrix sign error.



Satellite Tobacco Mosaic Virus (1a34, jan, 1998)

- Provides biological unit and crystal unit cell matrices.
- Biological unit is wrong, having chains on top of each other.
- Crystal unit cell gives correct capsid protein, but doubles RNA.
- RCSB staff believed they were ok based on 240,000 atom image.
- Virus Particle Explorer (VIPER) web site gives correct model.

Nucleic Acids Database | Virus Particle Explorer | VIPER

Summary of examples

- Above 3 examples chosen from looking at only 8 virus models.
- Matrix problems appear to be common for virus models.
- Problems may go undetected for a long time because of inadequate visualization software.

Survey for all virus entries

- 80 of 163 crystal structures provide non-crystal symmetry.
- 122 of 206 provide matrices for biological unit.
- No entries provide definitions of substructures, although mmCIF files can represent those definitions.

What to fix

- Priorities: 1) crystal unit cell, 2) biological unit, 3) biological substructures in machine readable format.
- PDB provides files containing all atomic coordinates for biological unit. These come from the Protein Quaternary Structure (PQS) server. Doesn't work well for viruses because of 100,000 atom limit in PDB file format. Doesn't work for above 3 examples. Visualization software needs to know subunit structure to work efficiently.
- Almost all virus entries contain enough human-readable information to figure out matrices. Use redundancy of biological unit and crystal unit cell matrices.
- mmCIF file format can completely describe the biological unit, and sub-structures, and the crystal.
- PDB file format relies on REMARK records for biological unit, and can't handle unusual crystal unit cell situations (like STMV rna fit twice).
- Most visualization software, including Chimera, reads PDB file format but not mmCIF file format.
- Might take one person 4 months to correct 200 PDB virus entries.

Conclusions

- Half of the PDB virus particles cannot be easily displayed.
- Lack of visualization software leads to poor data quality.
- Poor data quality leads to lack of visualization software.
- Virus structures are the best available data for guiding the development of software to analyze large complexes.