Structural volume data

Martyn Winn [1]

- Proteins
- Structures
- Intermediate
- 1 hour

This course provides a basic introduction of volume matching techniques in the field of structural bioinformatics. It explains the use of electron microscopy (EM) and electron tomography (ET) data in creating 3D volume maps of proteins, and the approaches that can be used to compare volumes. It describes how volume matching can be used to find similar structures, or to highlight specific differences between related structures.

Learning objectives:

- Gain an understanding of what structural volume data is
- Appreciate the value of volume data even in the absence of atomic coordinates
- Comprehend the variety of methods available to compare volumes
- Appreciate the limitations of available scores

What are volume data?

Macromolecules inside living cells include DNA, RNA, proteins, lipids, and polysaccharides; as well as hybrid macromolecules such as the lipopolysaccharides found in the outer membrane of Gram-negative bacteria. These macromolecules can form short-lived complexes, such as kinases or phosphatases binding their substrate during a signalling cascade; or can form large assemblies such as the proteins and rRNA [2] which make up the ribosome. Structural volume data describe the extent in three-dimensional space of these macromolecules and their complexes. The shape of macromolecules can provide us with important information on their function and interactions.

Volume data are generated by a number of experimental methods, including X-ray crystallography [3], electron microscopy [4], electron or soft X-ray tomography [5], and small angle X-ray or neutron scattering. In electron microscopy, volumes representing the electron potential of the sample molecule are reconstructed from many 2D projections extracted from micrographs. In crystallography, solution of the phase problem gives electron density maps; which are then invariably interpreted in terms of atomic models.
Volume data is typically represented in software or files as a set of density values on a rectangular grid, referred to as a map (see panel A below). That is, the continuous volume is sampled at regular intervals, with the sampling frequency typically higher than the intrinsic resolution of the data.

There are also a number of graphical representations of volume data, such as that shown in panel B above. Rather than show all density values held in a map file, graphical programs typically show a surface, derived from a suitable contour level of the density values.

**Segmentation**

Volume data can be segmented in three-dimensional space. Different segments may represent different component molecules in a complex, or may represent visually distinct regions in a tomogram. Segments are usually supposed to have clear boundaries, though in reality there may be some uncertainty as to where the border between two segments lies. There may also be a hierarchy of segments.

**What is volume matching?**
Volume matching is the procedure of superimposing one volume on another. This allows us to look for similarities or differences between the two volumes. Whole volumes can be matched with each other, or subvolumes associated with particular biological components may be mapped. The aim is to compare volume data from different samples, or collected under different conditions.

Superposition of two proteins based on their atomic coordinates is already a common operation, and there are many existing programs and services available. As structural biology moves to considering larger complexes and molecular machines, atomic coordinate models are not always available. Volume matching addresses this scenario.

**Figure 3** Schematic of matching a subvolume (red) to a larger volume (purple)

The above schematic illustrates what we are trying to achieve. There are a number of methods for aligning two volumes against each other - described later in the methodologies section. Although it may be time-consuming, alignment is relatively straightforward. A harder question to answer is finding out what is the best alignment from several possibilities; this issue is described in the section on scoring functions.

**Difference maps**

Once two volumes have been matched, and the transformation required to superimpose them determined, a difference map can be calculated. Features occurring in the first map but not the second will appear as positive peaks, while features occurring in the second map but not the first will appear as negative peaks. This means an additional macromolecular component present in one volume will be highlighted as a region of positive peaks. Conformational changes are also highlighted as a pair of positive and negative peaks, representing the new and old positions.

In practice, to get clean difference maps, the range of grid values have to be brought onto the same scale at all spatial frequencies (power spectrum matching).
What biological questions can we answer with volume matching?

The central tenet of structural biology is that macromolecular structure determines biological function. While function is often inferred from sequence information, it is the three dimensional structure of the protein or RNA that actually performs the function. This idea still holds even if the atomic structure (and hence chemistry) of the macromolecule is unknown.

Volume data can reveal details of biological function, and matching volumes between two datasets will detail similarities and differences in these functions. Specific examples include:

- Stoichiometry of complexes - the number of each molecule type contributing to the full complex. This can vary during the lifecycle of the complex.
- Binding of ligands - whether a particular ligand binds, and then the location of the binding site.
- Conformational changes - macromolecules are flexible, and comparisons can reveal changes under certain conditions, or a range of conformations for one condition.
- Similarity of function or binding partners - conserved shape may imply conserved function.
Volume pre-processing

When preparing to match two volumes from different sources, there are a number of pre-processing steps that help the comparison:

Bandpass filtering [14]

The resolution range of the higher resolution volume will typically be restricted to that of the lower resolution volume, since the higher resolution information cannot be matched.

Background peak

The distribution of density values in the map typically has a peak representing the large number of voxels having a background value. The map values are all shifted by a constant amount so that this peak appears at zero.

![Figure 5 Distribution of map values, illustrating the background peak at zero. (Plots are copied from the Advanced Validation pages of the corresponding EMDB entries.)](image)

Contour level

The map contains a continuous range of values at its grid positions. While high values are expected to represent the macromolecule and low values the surrounding solvent, there is no explicit division between the two. When interpreting a map, software will often set a contour level that represents the boundary between molecule and solvent, but this choice is somewhat subjective. Typically, it is chosen so that the enclosed volume corresponds to the known molecular weight at a typical molecular density.
Figure 6 A volume map [15] displayed at different contour levels. At higher contours, the volume is restricted to smaller regions of high density. It is not necessarily clear which is the "correct" contour level. (Figure prepared with UCSF Chimera.)

**Noise removal**

Maps often contain many small peaks arising from noise in the data, which can confuse interpretation. These can be removed by "dusting", i.e. by removing disconnected volumes smaller than a given cutoff. It may also be useful to hide dust in difference maps, in order to focus on major changes in the volume data.
Volume-matching methodologies

Methodologies for matching volume data can be loosely divided into two categories:

Direct comparison

Volumes are represented as a grid of density values. One volume is rotated and translated with respect to the other, and the overlapping density values are compared, typically via a correlation coefficient. As the grids do not, in general, lie on top of each other, some interpolation of density values is required. The comparison may be done within a smaller masked region representing the overlap between the volumes (e.g. local correlation coefficient).

Pros:

- Works with original volume data as derived from experiment
- Full resolution of volume data is retained

Cons:

- Computational intensive, as map values at a large number of grid points (potentially a few million) need to be compared, for each trial rotation/translation.

Reduced representation

Alternatively, volume data can be coarse-grained in some way to make calculations faster. One method represents the volume as a set of anisotropic overlapping Gaussian functions. Another expands the volume as a series of orthogonal functions, for example spherical harmonics, and compares the coefficients of the expansion. The first step of volume matching involves finding the optimum representation in the reduced form, while the second step involves fast matching of the individual terms.

Pros:
Structural volume data
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

- Matching the reduced representations of two volumes is considerably faster than matching the original maps.

**Cons:**

- Need to decide the level of detail required for each volume.
- Generating the optimum reduced representation may become the rate limiting step.

**Scoring functions of volume matching**

Volume matching algorithms yield a set of putative alignments between a pair of volumes. These are typically scored according to some criterion internal to the software used. However, there are a range of scoring functions which can be applied to get external validation, and these can be sensitive to different characteristics of the volumes. If there is a consensus between several different scores, then one can have some confidence in the proposed alignment.

Several scoring functions have been designed for quantifying the quality of alignments. A comparison is available of these functions in the Vasishtan and Topf (2011) review paper:

1. Local Cross-Correlation Coefficient (Roseman AM. 2000) - a correlation coefficient between the density values of the two maps, calculated for the overlap region only.
2. Laplacian-filtered CCF (Chacon and Wriggers, 2002) - the Laplacian filter helps to pick out the surfaces of the volumes.
3. Overlap score (Chimera, Pettersen et al. 2004) - fraction of overlapping voxels with respect to the smaller of the two volumes.
4. Core-weighted envelope score (Wu et al. 2003) - a cumulative measure based on the number of overlapping voxels, with a penalty for non-overlapping voxels.
5. Mutual information score (Shatsky et al. 2009) - a measure of how similar the joint distribution of density values in the superimposed maps is to the individual density distributions.
6. Chamfer distance (Knossow et al. 2008) - an average distance between nearest neighbour points on the surfaces of the two superimposed maps.
7. Normal Vector score (Ceulemans and Russell. 2004) - an average angle between vectors normal to the two surfaces.

Many of these scoring functions depend on the surface of the volume being identified as a particular contour level of the map. The choice of contour level is not necessarily straightforward.

**Volume matching software**

A complete procedure for volume matching requires several steps:

1. **Pre-processing** [13] of volume data, including bandpass filtering [14], dust removal, determination of optimum contour level, etc.
2. Alignment of volumes against each other, using a search over relative rotations and translations.
3. Re-scoring of alignments.
4. Presentation and download of results, including postprocessing steps such as generation of difference maps.
We have written such a pipeline, consisting of the underlying software SMaSB and a web service PDBeShape. The pipeline makes use of 3rd party programs for individual steps, and is written in such a way that it can be updated as new algorithms and software are developed. Programs currently used include:

**TEMPy**

TEMPy (Template and Electron Microscopy comparison using Python) is a toolkit designed for assessing density fits in intermediate-to-low resolution maps, both globally and locally (Farabella et al 2015).

**Chimera**

Chimera is a molecular graphics program with many analytical tools included. We make use of the map fit method, available as a command line python module, which implements a direct comparison of two maps. A 6D search is carried out by random initial placements of the search map, followed by steepest ascent search to maximize the alignment score. We typically use 200 trial placements (Pettersen et al 2004).

**gmfit**

Gmfit relies upon a reduced representation of the map density in terms of a Gaussian Mixture Model (GMM), which is a linear combination of several Gaussian Distribution Functions (GDFs). The GMM enables a quick fit of the density map and the subunit models, based on the overlap of two GMMs. The number of GDFs controls the description of the map; a larger number generates a more detailed density function at the expense of computational time. Overlap between two GMMs is quantified by the fitness energy. The 6D search to align two GMMs is carried out by random sampling, followed by a steepest descent search using the fitness energy gradient (Kawabata 2008).

**Limitations**

Volume matching is still computer intensive, especially if one wishes to search over a large set of volumes (e.g. all those contained in the [EMDB](https://www.ebi.ac.uk/emdb/)). The PDBeShape server provides access to a set of pre-calculated volume alignments which can be browsed or searched. The first release contains high quality volume data for prokaryotic and eukaryotic ribosomes (823 volumes), and for class I and II chaperonins (233 volumes), taken from EMDB and [PDB](https://www.rcsb.org/)

**Volume matching use case**

**Aligning bacterial ribosomes**

The ribosome [18] is a large and complex molecular machine that serves as the site of biological protein synthesis. The bacterial ribosome is considered to consist of three rRNA [2] molecules and around 50 proteins. However, it is also the site for a large number of additional proteins that have specific roles in the life cycle of the ribosome. Aligning volume data for ribosomes can help to identify these additional proteins, and any structural changes they may trigger.

Tetracycline resistance protein TetO protects the bacterial ribosome from binding the antibiotic tetracycline. It is a translational GTPase with significant similarity to the elongation factor EF-G. We can compare the 9.8 Å volume of a TetO-bound bacterial ribosome (EMD-5562) with a 9.9 Å volume of a bacterial ribosome with EF-G (and RRF) bound (EMD-1917).
Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

You can learn more about structural volume tools and resources at the following websites:

Electron Microscopy Data Bank (EMDB) [19]

PDBShape [20]

Collaborative Computational Project for Electron cryo-Microscopy [21]

References


Funding

This e-learning module was developed as part of the BioMedBridges [33] project, funded by FP7 Capacities
Specific Programme, grant agreement no. 284209.

Contributors

Martyn Winn [1]

STFC
Group Leader Computational Biology

Chemist by training, spent two years in a theoretical physics department, and now group leader for computational biology. I've worked on software for macromolecular crystallography (CCP4), and am now expanding to other structural biology techniques through the new Instruct Centre and the new CCP in electron cryo-microscopy (CCP-EM). I am also working on modelling and simulation techniques, with a particular connection to high performance computing (CCPBioSim). As well as traditional HPC, I am also interested in data-intensive applications in biology, particular in the context of the Hartree Centre at Daresbury. A current example is algorithms for transcriptome assembly.
Ingvar Lagerstedt [34]
EMBL-EBI
Scientific Programmer, BioMedBridges

Tom Hancocks [35]
EMBL-EBI
Scientific Training Officer - BioMedBridges

Tom Hancocks works as a Scientific Training Officer for the Training Team at EMBL-EBI.

He studied Human Genetics at the University of Leeds and McMaster University in Hamilton, Ontario; before completing an MSc in Analytical Genomics at the University of Birmingham.

Tom has worked for the NHS in diagnostic genetics and as a bioinformatics trainer for healthcare scientists and clinicians.

Source URL: https://www.ebi.ac.uk/training/online/course/structural-volume-data

Links
[1] https://www.ebi.ac.uk/training/online/trainers/martyn.winn
[2] https://www.ebi.ac.uk/training/online/glossary/rrna
[3] https://www.ebi.ac.uk/training/online/glossary/term/730
[4] https://www.ebi.ac.uk/training/online/glossary/term/278
[5] https://www.ebi.ac.uk/training/online/glossary/term/279
[7] https://www.ebi.ac.uk/training/online/glossary/mesh
[8] https://www.ebi.ac.uk/training/online/course/structural-volume-data/what-biological-questions-can-we-answer-volume-mat
[9] https://www.ebi.ac.uk/training/online/glossary/atomic-coordinates
[10] https://www.ebi.ac.uk/training/online/course/structural-volume-data/volume-matching-methodologies
[12] https://www.ebi.ac.uk/training/online/glossary/transformation
[13] https://www.ebi.ac.uk/training/online/glossary/pre-processing
[14] https://www.ebi.ac.uk/training/online/glossary/filtering
[15] https://www.ebi.ac.uk/training/online/glossary/volume-map
[16] https://www.ebi.ac.uk/training/online/glossary/emdb
[17] https://www.ebi.ac.uk/training/online/glossary/pdb
[33] http://www.biomedbridges.eu/
[34] https://www.ebi.ac.uk/training/online/trainers/ingvar
[35] https://www.ebi.ac.uk/training/online/trainers/Tom