Shape Matching Service for Structural Biology (SMASB)

www.biomedbridges.eu

Agnel Praveen Joseph[1], Ingvar Lagerstedt[2], Gerard Kleywegt[2], Ardan Patwardhan[2], Maya Topf[3], Eduardo Sanz Garcia[2], Martyn Winn[1]

1. Research Complex, STFC Harwell , Didcot OX11 0QX, UK / Computational Science & Engineering Dept., STFC Daresbury Laboratory, Warrington WA4 4AD, UK
2. Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
3. Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

As part of the BioMedBridges project Work Package 9, PDBe (pdbe.org) and STFC (stfc.ac.uk) are setting up a service that will allow users to search and compare volumes derived from structures determined by X-ray crystallography, Electron Microscopy (EM), Electron Tomography (ET) and SAXS. Consequently we aim to arrive at a structural integration of PDB data with EMDB, provide structural annotation for EMDB maps and also link other databases like Uniprot (http://www.uniprot.org), IntAct (http://www.ebi.ac.uk/intact) and GO (http://geneontology.org). Here we discuss potential use cases, volume comparison strategies with some examples, scoring functions and the framework of the service.

Some use cases

Volume *alignments* :
- Derived from different species
- Generated by different experiment types
- Involving conformational variations
- Involving additional components (difference maps)

Structure/Component Annotation :
- *Automated segmentation* of volume (iteratively) into components
- *Fit models* into volumes/segments (PDB and EMDB)
- *Fold identification* by search with known folds

Function Annotation/Interactions:
- Get interaction data (*IntAct*) of components with sufficient linking data
- Get *UniProt* & *GO* annotation/ontology of components with sufficient linking data

Volume matching approaches considered

Density based 6D search :
- **Fast Translational Matching (COLORES)** : Fast Fourier Transform (FFT) driven translational search at each rotation (Chacon and Wriggers, 2002). Cross correlation scores with or without Laplacian filters, are used for matching
- **Fast Rotational Matching (ADP-EM, FRM5D)** : Spherical Harmonics accelerated rotational search at each translation (Garzon et al. 2007, Kovacs et al. 2003). Cross correlation scores with or without Laplacian filters, are used for matching
- **Random Sampling (CHIMERA)**: Randomly sample rotation and translation space (Goddard et al. 2007). An overlap score based on product moments is used for matching.

Reduced representation :
- **Gaussian Mixture Model (GMFIT)**: Linear combination of N 3D gaussian density (Kawabata, 2008). A gaussian overlap metric is used to score the alignment.
- **Point Cloud Matching (MatchPt)** : A self-organizing Topology Representing Network is used to represent volume features as N points (Birmann & Wriggers, 2007). Minimal RMSD or statistical variability measures are used for matching.

Volume pre-processing

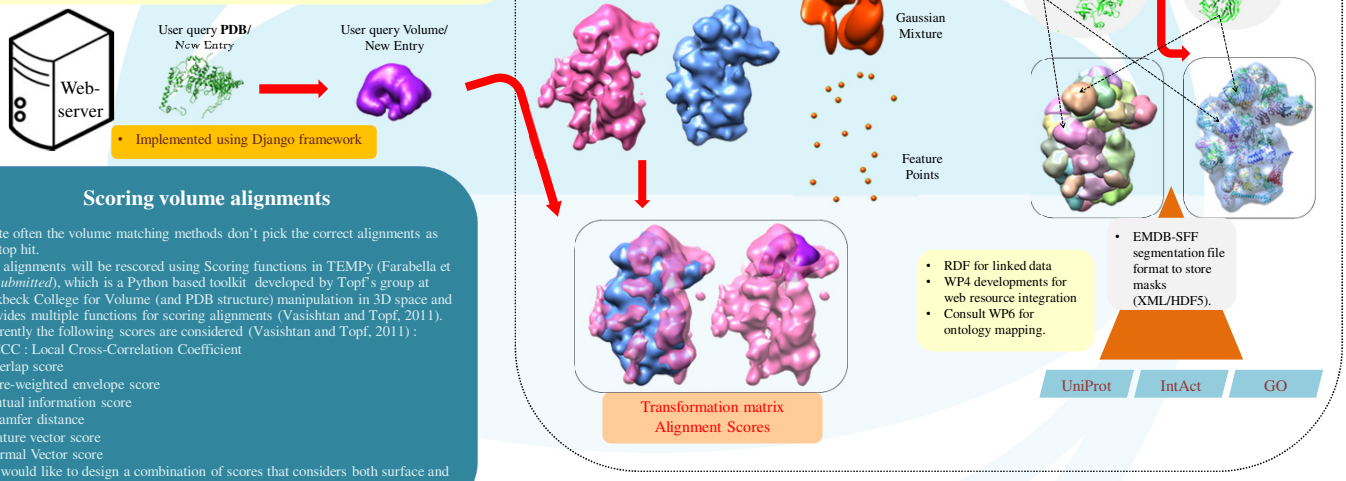
- Volumes of macromolecular complexes determined by EM/ET/SAXS include varying noise levels and the density distributions may differ significantly.
- Contour level determination for the volumes based on the molecular weight, standard deviation of density distribution and levels suggested by authors.
- The background peak is shifted to zero for cases with significant deviations.
- Feature representation (as Gaussian Mixtures and point clouds) is included as part of pre-processing.

Filtering

- The grid spacing and resolution of one map is filtered to match the other, prior to alignment.

- MySQL like relational database.
- XSD schema for representing meta-data
- OpenAstexViewer (http://openastexviewer.net/web) for volume visualization
- Can be provided in RDF format (http://www.ebi.ac.uk/rdf) using D2R (http://d2rq.org/d2r-server)

Database with shapes of EMDB and PDB entries and segmentations

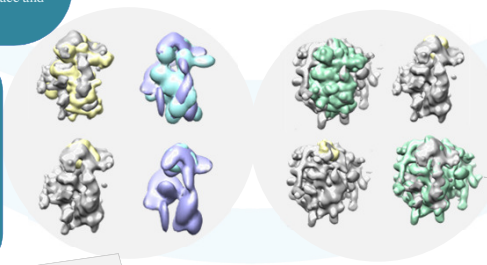


Scoring volume alignments

- Quite often the volume matching methods don't pick the correct alignments as the top hit.
- The alignments will be rescored using Scoring functions in TEMPy (Farabella et al. *submitted*), which is a Python based toolkit developed by Topf's group at Birkbeck College for Volume (and PDB structure) manipulation in 3D space and provides multiple functions for scoring alignments (Vasishtan and Topf, 2011).
- Currently the following scores are considered (Vasishtan and Topf, 2011) :
  - LCCC : Local Cross-Correlation Coefficient
  - Overlap score
  - Core-weighted envelope score
  - Mutual information score
  - Chamfer distance
  - Feature vector score
  - Normal Vector score
- We would like to design a combination of scores that considers both surface and density overlap.

Test set for volume matching methods

- About 50 volumes were selected from each of the four major sample categories from EMDB : Prokaryotic Ribosomes; Eukaryotic Ribosomes; Chaperones ; Viral structures
- The test set involves volumes of different grid sizes, resolutions, sample components and from different research groups.
- Simulated maps derived from PDB structures are also added to each category.
- The test set is publically available at: ftp://ftp.ebi.ac.uk/pub/databases/emtest/PDBeMatch_data/PDBeMatch Data.tar



Examples from sample set

- Viral RNA bound 80S ribosome (EMD-1138, 15Å)
- Viral RNA bound 40S ribosomal subunit (EMD-5527, 17.5 Å)
- 40S ribosomal subunit (EMD-5592, 5Å)
- Simulated 40S maps (PDB ID: 2XZM)
- Simulated maps of sub-complexes (PDB ID : 3J3A, 3 protein complex)

- **Random sampling** search (CHIMERA) is faster than exhaustive search methods. Each alignment takes about 7 mins on a single processor.
- The number of search steps depends on the ratio of volume sizes.
- Except for a 4Å 40S vs (15Å 80S, 17.5Å 40S), correct alignments were obtained in the top 10 hits.

- **Feature (coarse-graining)** based methods are extremely faster than the 6D density search methods. GMFIT takes a few seconds for matching two gaussian mixture models
- The gaussian overlap function used for scoring allows flexibility in matching different resolutions and sizes.
- The number of gaussians chosen to represent a volume is important and it depends on the resolution and size of the volume. Larger number is required to represent maximum features from a high resolution map.
- Search for a smaller component (subcomplex) failed with a constant number of 16 gaussians.
- We plan to automate feature identification and gaussian model generation as part of the pipeline.

- **Exhaustive search** methods (COLORES,ADP-EM) are much slower than Random sampling or Feature based methods. Each alignment takes about 1-2 hours on a single processor.
- Failed cases mainly involved subunit or subcomponent alignments.
- Overall, the exhaustive search methods do not seem to gain largely in the test set, considering the time taken for each run.

BioMedBridges is funded through a grant from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 284209. It brings together first and second generation biomedical sciences research infrastructures from the ESFRI roadmap.

Contact:
Agnel Praveen Joseph[1]
agnel-praveen.joseph@stfc.ac.uk
Ingvar Lagerstedt[2]
ingvar@ebi.ac.uk

1. Research Complex, STFC Harwell , Didcot OX11 0QX, UK

2. Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK