BioMedBridges

Deliverable D9.2

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Final version of the web-based shape-matching service	
WP No.	9	
Lead Beneficiary:	1: EMBL	
WP Title	Use case: From cells to molecules- integrating structural data	
Contractual delivery date:	31 March 2015	
Actual delivery date:	10 June 2015	
WP leader:	Martyn Winn	4: STFC
Partner(s) contributing to this deliverable:	1: EMBL 4: STFC 20: CIRMMP	

Authors: Martyn Winn, Ardan Patwardhan, Agnel Praveen Joseph, Ingvar Lagerstedt

Contents

1 Executive summary

A software pipeline, named SMaSB, was previously developed to perform the volume/shape matching that underpins the set of tools being developed in WP9. These tools are novel in providing access to a growing class of structural biology data viz. volume data. The SMaSB software is primarily a set of Python codes which organise the metadata, control the data flow during volume/shape matching, and record the results. Third-party software is called to perform the compute-intensive steps in the pipeline. The first full version of SMaSB was released in December 2014, and reported in Deliverable 9.1.

We now report on the associated web service (PDBeShape) which provides a user-friendly front-end to SMaSB results. It provides web access to a curated database of structural volumes, derived from deposited structures in the Electron Microscopy Data Bank (EMDB) and Protein Data Bank (PDB), together with pre-processed alignments and scores. We have made the first version of the PDBeShape service publically available at http://wwwdev.ebi.ac.uk/pdbe/emdb/shape/welcome/.

In this document, we give details on the current functionality of PDBeShape, and the technical underpinning. More extensive documentation is provided on the web page.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Develop a database of annotated biomacromolecular volume data	√	
2	Develop software to search this database using atomic or volume data	√	
3	Methods for routine updates developed	√	

4	methods to identify components (“segments”) and annotate them implemented		X
5	Integration of SAXS and NMR data on flexible proteins in solution		X
6	Tools available via webserver	√	

3 Detailed report on the deliverable

3.1 Background

Work Package 9 aims to increase the availability and utility of volume data obtained from structural biology techniques working at the molecular or supra-molecular level, by providing search and analysis tools analogous to those available for atomic structures. This Use Case links Instruct (as the generator of volume data) with Elixir (as the curator of volume databases, e.g. EMDB). It will be delivered via a software stack covering the underlying matching algorithms, a web-based front end, and database operations. D9.1 was the first deliverable from this work package, and covers the underlying software pipeline (SMaSB) for matching volumes and capturing appropriate metadata.

Here we report on D9.2 which delivers the first public version of PDBeShape. The latter is a web portal for exploring and searching the results of volume matching performed by SMaSB on a database of high quality volumes taken from the EMDB and PDB. Sets of similar volumes can be discovered, and downloaded for further analysis. The sensitivity of results to the alignment algorithm and scoring function can be tested. SMaSB and PDBeShape provide the infrastructure for later deliverables in WP9, and for on-going development after the end of BioMedBridges.

3.2 PDBeShape

PDBeShape is designed to hold curated structural volume data from a variety of sources such as electron microscopy, subtomogram averaging, crystallography, and potentially other experimental techniques such as small angle scattering. Structural alignments between the volumes have been pre-

calculated, and these can be searched in PDBeShape to find volumes with similar morphology. Differences between otherwise similar volumes are highlighted, and may reveal conformational changes or location of additional components. This first release contains structural data for prokaryotic and eukaryotic ribosomes, with 823 volumes taken from the EMDB and PDB (EMDB volumes of resolution < 30Å and atomic models in PDB with molecular weight > 250KDa). In later releases, we will include a wider range of volume types.

This first public release of PDBeShape includes two protocols for volume matching: the density-based search out of Chimera (Pettersen et al., 2004), and the coarse-grained approach of Gmfit (Kawabata, 2008). The spherical harmonic based search implemented in ADP_EM (Garzon et al., 2007) will be included in the next release. Alignments can be re-scored according to 8 different metrics, as implemented in TEMPy (Farabella et al., 2015). By default, results are displayed based on alignment with Chimera with 200 trials, and with a weighted score consisting of the local correlation coefficient (80%) combined with the fractional overlap score (20%). Test of accuracies of different scores in Tempy (Farabella et al. 2015) on a dataset of 35 volume alignments (derived based on superposition of associated fitted coordinates), encompassing ribosomes(17), chaperones(5) and viruses(13), has shown that this combination has better discrimination.

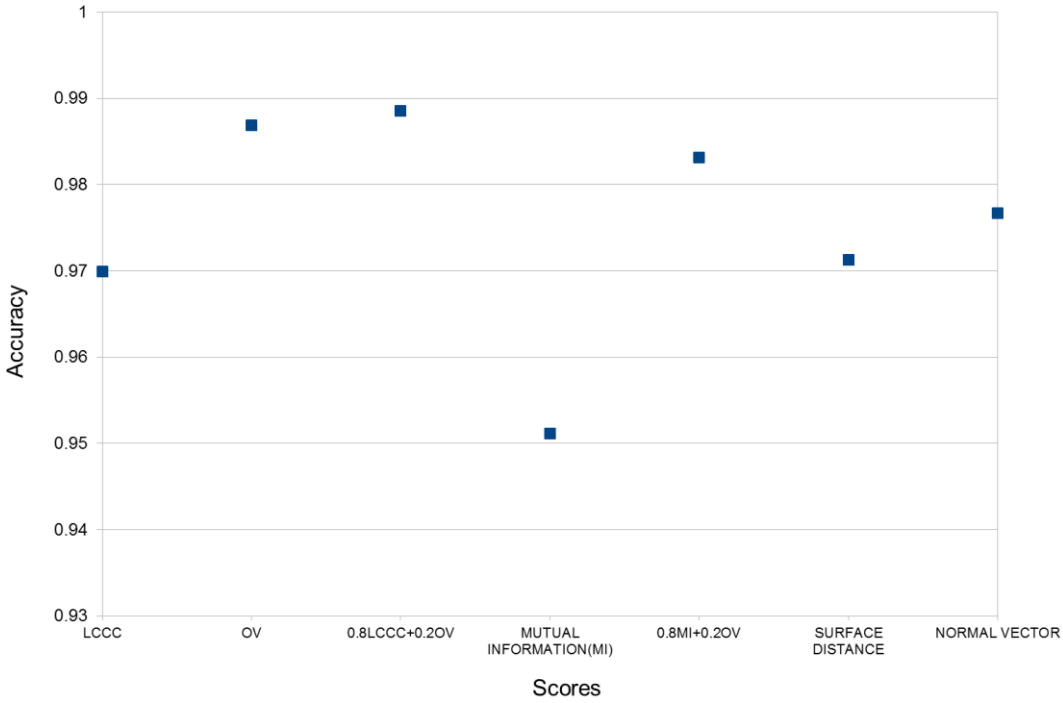

Figure 1 Accuracy of selected scoring metrics on a test dataset of 35 volume alignments. The third score ("0.8LCCC+0.2OV") has been chosen as the default in PDBeShape

SMaSB will be maintained and further developed by the CCP-EM project and PDBe. The CCP-EM project (www.ccpem.ac.uk) is a UK-funded Collaborative Computational Project, whose purpose is to provide long-term support for cryoEM analysis software. The sister project CCP4 has provided such support for X-ray crystallography over 35 years. PDBe manages the PDB and the Electron Microscopy Data Bank (EMDB; www.ebi.ac.uk/pdbe/emdb/) archives with its wwPDB and EMDataBank partners respectively and is developing a range of services for the cryoEM community. The PDBeShape web service will be part of this portfolio, and PDBe will help to maintain the underlying SMaSB pipeline.

3.3 Volume database

Category	Count
Volumes	823
Taxonomies	84
Samples	322
Alignment protocols	2
Alignment scores	8
Pair-wise alignments	1290386
Total number of fits	25331531

Figure 2 Contents of the volume database, as accessed on 3/6/2015

The volume database contains entries on volumes, alignments, samples and taxonomies. A summary of the current holdings can be accessed on the About page of the service.

The sample description consists of 4 layers:

- Category – e.g. virus/cellular component/synthetic
- Complex – e.g. ribosome/chaperone
- Component – e.g. rRNA/Elongation factor
- Domain – from pFam/rFam
  - pFam from Uniprot for PDB
  - rFam from flat files for PDB

Some components can be extracted from the EMDB sample description

Sample details for each volume can be accessed from the Details button on the Volume page of the service.

PDBeShape currently holds 84 taxonomies, named according to species, sub-species or strain. For each included taxonomy, details of the main ranks are given, with links to the NCBI Taxonomy Browser.

3.4 Using the web service

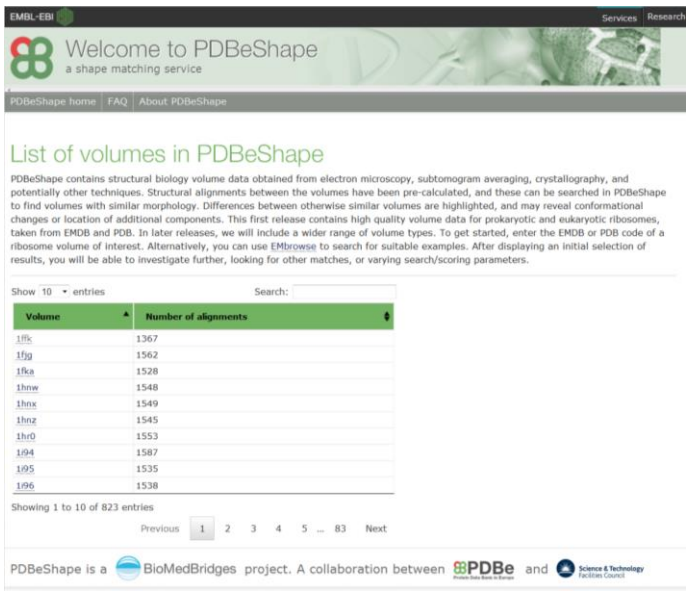

Figure 3 Home page of the PDBeShape web service

On the Home page of the service, the user enters the EMDB or PDB code of a ribosome volume of interest (for example “6306” for the cryo-EM structure of

the Bacillus subtilis MifM-stalled ribosome complex). Alternatively, they can use EMbrowse (http://www.ebi.ac.uk/pdbe/emdb/browse.html) to search for suitable examples, or simply browse the displayed list.

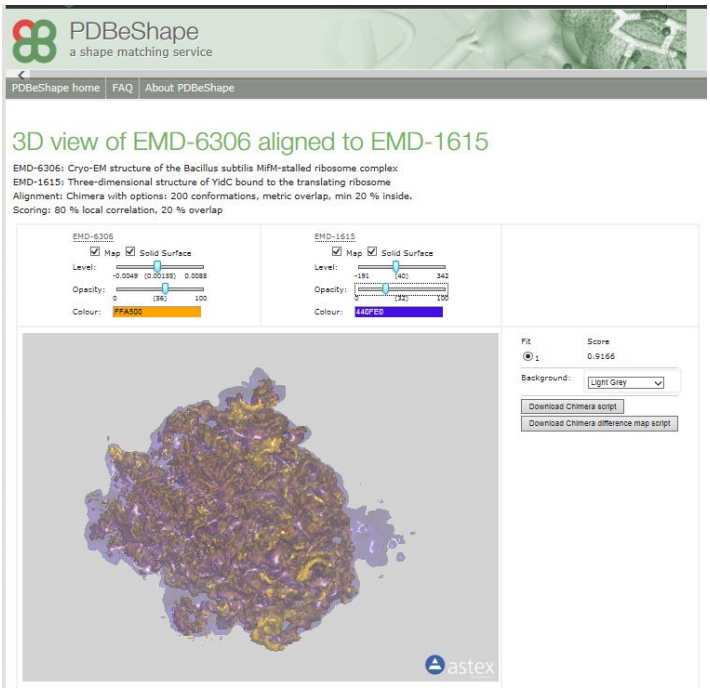

Figure 4 An example of an Alignment page, showing details of a particular volume-volume alignment

Clicking on the chosen volume takes the user to the specific Volume page. This displays further details about the volume (for example sample information and taxomony), and a link to the EMDB or PDB entry page. It also gives a list of pre-calculated alignments to other volumes, which can be sorted according to volume ID or score (for example, the best alignment for volume 6306 is EMD-1615 for an E.Coli ribosome). Clicking on View for an alignment takes the user to the specific Alignment page, which displays the calculated alignment in the Java-based molecular graphics viewer Open Astex Viewer (Hartshorn, 2002; Lagerstedt et al., 2013). From this screen, scripts can be downloaded for viewing the alignment and/or a difference map in a local installation of Chimera. If there is more than one predicted alignment between a given pair of volumes, then the user can select from a ranked list. The solutions proposed are filtered to avoid similar orientations with difference in rotations less than 6° and shift less than 3Å. Also, the aligned volumes should have at least 20% overlap.

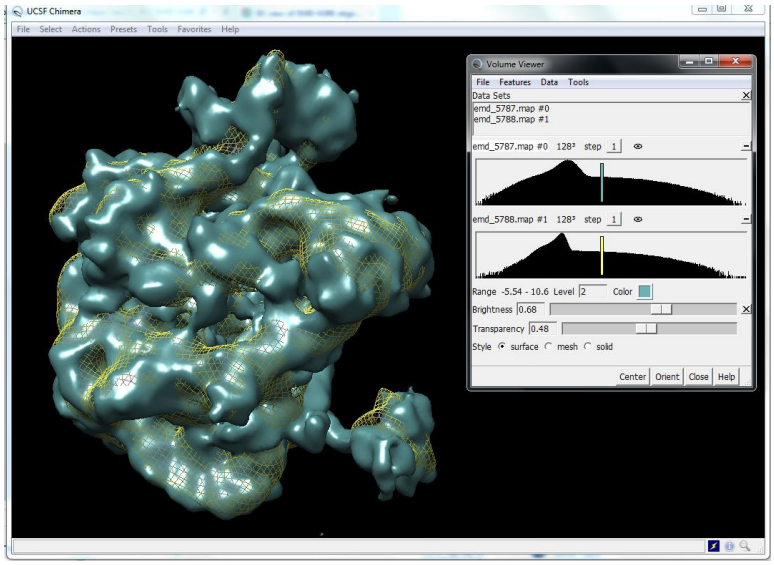

Figure 5 A volume alignment, as viewed in Chimera after downloaded of the results from PDBeShape

On the Volume page, the user can investigate alignments in more detail using the Advanced Search facility. Currently, alignment protocols from Chimera and gmfit are available, together with a choice of 8 scores (these are accessed via the SMaSB pipeline, and are described in more detail in the Milestone 20 report). The search can be restricted so that results match at a particular sample or taxonomy level, for example a search with volume 6306 can be restricted to volumes obtained from *Bacillus subtilis*.

4 References

[1] Farabella I et al. (2015). “Tempy: A python library for assessment of 3D electron microscopy density fits”, *in press*.

[2] Garzon JI, Kovacs J, Abagyan, R and Chacon P. (2007) “ADP_EM: Fast exhaustive multi-resolution docking for high-throughput coverage.” *Bioinformatics*. **23**(4):427-33.
DOI: 10.1093/bioinformatics/btl625

[3] Hartshorn, MJ. (2002) “AstexViewer: a visualisation aid for structure-based drug design.” *J Comput Aided Mol Des*. **16**(12):871-81.
DOI: 10.1023/A:1023813504011

[4] Kawabata T (2008) “Multiple Subunit Fitting into a Low-Resolution Density Map of a Macromolecular Complex Using a Gaussian Mixture Model.” *Biophys J.*, **95**(10): 4643–4658.
DOI: 10.1529/biophysj.108.137125

[5] Lagerstedt I., Moore W.J., Patwardhan A., Sanz-Garcia E., Best C., Swedlow J.R. and Kleywegt G.J. "Web-based visualisation and

analysis of 3D electron-microscopy data from EMDB and PDB". *J Struct Biol*, **184**, 173-181 (2013).
DOI: 10.1016/j.jsb.2013.09.021

[6] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC and Ferrin TE. (2004) “UCSF Chimera - a visualization system for exploratory research and analysis”. *J Comput Chem*. **25**(13):1605-12.
DOI: 10.1002/jcc.20084

5 Supplementary information

Supplement 1: Technical details

The PDBeShape web service has been developed within the Django framework. The metadata for the volume database is held in an SQL database, while the volume data itself is held in MRC-format files. An XML schema is used to represent the datamodel.

Alignment and scoring is done using SMaSB (the shape matching software) which is separately available as a compressed tarball from the downloads page of the CCP-EM project (http://www.ccpem.ac.uk/download.php). In addition to a set of Python modules written specifically for WP9, it includes a version of Open Astex Viewer (OAV; http://openastexviewer.net/web/; Hartshorn, 2002; Lagerstedt et al., 2013) that incorporates additional functionalities for volume processing. The SMaSB package makes use of the following 3rd party software (all free of cost for non-commercial use):

- TEMPY (Farabella et al., 2015) is a suite of python scipy/numpy based modules for coordinate/map transformation and validation/scoring of alignments, available from http://tempy.ismb.lon.ac.uk/.
- GMFIT (Kawabata, 2008) is used as the default alignment method as it gives fast and reasonably good alignments. It is available from http://strcomp.protein.osaka-u.ac.jp/gmfit/. The use of Gmfit for alignment requires Chimera to be installed as well.
- UCSF CHIMERA (Pettersen et al. 2004) generates volume alignments by random sampling and hence can be quick depending on the number of sampling steps used (200 by default). It is available from https://www.cgl.ucsf.edu/chimera/.

— ADP-EM (Garzon et al. 2007) is relatively slower but follows a more exhaustive spherical harmonic based search. It is available from http://chaconlab.org/methods/fitting/adpem.

— lxml is a Python module for XML and HTML processing. It may already be installed on users' systems, but otherwise is available from http://lxml.de/

6 Delivery and schedule

The delivery has been delayed by 2 months to allow for final testing. Work on other WP9 deliverables has been progressing in parallel, and there should be no knock-on effects on these deliverables.

7 Adjustments made

None

8 Background information

This deliverable relates to WP 9; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 9 Title: Use case: From cells to molecules- integrating structural data
Lead: Martyn Winn (STFC)
Participants: EMBL, STFC, CIRMMP

Work package number	WP9	**Start date or starting event:**	month 13
Work package title	From cells to molecules- integrating structural data		
Activity Type	RTD		
Participant number	1:EMBL	4: STFC	20: CIRMMP

Person-months per participant	32	33	8

Objectives

We will develop tools (software, database, web-based services) to bridge the resolution ranges encountered in atomic, molecular and cellular structural biology. Specifically, we will:

1. Develop a database of annotated biomacromolecular volume data (derived from PDB and EMDB and annotated by UniProt and other relevant database identifiers) and software to search this database using atomic or volume data that result from experimental structure determinations. These tools will be made available through a webserver. Methods will be developed to routinely update the database with every new release of PDB and EMDB.

2. Implement methods to identify components ("segments") and annotate them (using UniProt and other relevant database identifiers) in experimentally determined volume data (e.g., tomograms). This functionality will be made available as a webserver and will possibly be integrated in the deposition procedures for EMDB/PDB.

3. Integration of SAXS and NMR data on flexible proteins in solution in order to evaluate the average shapes, as well as the shapes of the various conformations sampled in solution.

Task 1. (STFC, EMBL)

Structural biology is producing unprecedented amounts of structural data that increase not only in number, but also in size and complexity and that span an ever-wider range of resolutions. Whereas X-ray crystallography and NMR spectroscopy produce structural models with atomic detail, techniques such as 3D cryo-Electron Microscopy and Tomography as well as Small-Angle Scattering (X-ray and neutrons) produce lower-resolution volume and shape data. Moreover, a deluge of hybrid techniques currently being developed is expected to produce complex mixtures of high-resolution and low-resolution structural information about ever more complex molecular machines. Whereas there are very good bioinformatics tools available for the analysis, validation and comparison of atomic structures, at present there are very few tools available that deal with low-resolution data (i.e., volume or shape data). In this task, we will address this by developing tools (software, database, web-based services) for searching the structural archive, not at the level of atoms or secondary structure elements (for which good tools are available, some of which were developed jointly by partners now involved in INSTRUCT and ELIXIR), but based on shape (volume data). The shape database will be derived from the holdings of PDB and EMDB and will contain annotated shape data at various level of resolution. The shape-matching software will be able to take structural data (be it an atomic model or volume data itself) and compare it to the contents of the shape database in order to identify known structures with similar shape or with a component of similar shape. Such software will be invaluable to assist in annotation of, for instance, whole-cell tomograms and for identification of components of known structure or shape in large multi-

molecule complexes. The software will be made available both stand-alone and as a web-server. Methods will be developed to routinely update the shape database with every new release of PDB and EMDB.

Task 2. (EMBL, STFC)
The second task focuses on delineation, identification and annotation of segments in experimentally determined volume data (single-particle reconstructions, tomograms, possibly small-angle scattering). At present, volume data can deposited in EMDB without any link to atomic structures, either because the structures are not yet known of because the authors of the study choose not to fit existing structures or to deposit them. The value of the EMDB archive would be enhanced substantially if volume data would be decomposed into its constituent biomacromolecular components (various proteins, possibly RNA or DNA, etc.) and identified through annotation using UniProt and other relevant database identifiers. We will examine and adapt existing segmentation software
so that it can be incorporated into the annotation tool. The annotation tool itself will be developed initially as a stand-alone web-server. It will also be considered for integration in the EMDB/PDB deposition pipelines, in consultation with the international partners in those two organisations. The two tasks together will result in significant new functionality that will aid:

- (structural) biologists who want to find out if a certain biomacromolecular structure has the same shape as a known structure (which may be known at atomic level or as part of an experimentally determined volume, such as an EM map or tomogram);
- (structural) biologists who want to interpret complex volume data in terms of possible and plausible structures of components of that data (e.g., when annotating particles in a tomogram);
- PDB/EMDB in the sense that previously deposited volumes for which no atomic data was available can be scanned regularly for fits of newly determined structures. Moreover, once segmentation and identification information is available, whenever an atomic structure becomes available for a component that was previously only known at the level of its shape, this information can be exploited automatically and the structure can be fit into the volume data. This will transform EMDB from a static archive of volume data, to a dynamic archive whose content will continue to develop and become richer as time goes by and new atomic structures become available.

Task 3. (CIRMMP, STFC, EMBL)
The third task relates to proteins which experience some kind of mobility in solution, and to how this mobility can become a descriptor in structural databases. The task consists of finalizing programs available and partly developed by CIRMMP to determine the shape of the various protein conformations sampled in solution and, according to their estimated statistical weight, to determine selected measurable properties. The programs will take advantage of experimental parameters mainly from NMR and SAXS. Once finalized, the programs will be integrated with the shape-matching software and service of Task 1.