BioMedBridges

Deliverable D9.1

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 shape-matching software	
WP No.	9	
Lead Beneficiary:	4: STFC	
WP Title	Use case: From cells to molecules- integrating structural data	
Contractual delivery date:	31 December 2014	
Actual delivery date:	23 December 2014	
WP leader:	Martyn Winn	4: STFC
Partner(s) contributing to this deliverable:	1: EMBL 4: STFC 20: CIRMMP	

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

Contents

1 Executive summary

A software pipeline named SMaSB ("Shape MAtching service for Structural Biology") has been developed to perform the volume/shape matching which will underpin the WP9 service. The service is 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 software has been developed following an exploration of a range of techniques, and tested against a variety of volume datasets (see the design document in MS20, and the description of the prototype in MS21).

We have now released the first full version of SMaSB. The software can be downloaded and run locally by a scientist to compare a volume against another volume or a list of volumes. The same software is being used in the prototype web-based service[1]. The development work has clarified the data items required for the WP9 service. Where possible, we have used existing data models, such as that underpinning EMDB (and serialised via an XML file which the pipeline downloads). We are also linking to data models being developed, such as that for molecular complexes.

SMaSB is available as a compressed tarball from the downloads page of the CCP-EM project[2]. In this document, we give details on how to obtain and install the software, and an overview of how to run it. More extensive documentation is provided in the download itself.

2 Project objectives

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

[1] BioMedBridges project deliverable 9.2 Final version of the web-based shape-matching service
[2] http://www.ccpem.ac.uk/download.php

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		X
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		X

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, viz EMDB). It will be delivered via a software stack covering the underlying matching algorithms, a web-based front end, and database operations. D9.1 is the first deliverable from this work package, and covers the underlying software for matching volumes and capturing appropriate metadata. This software, named SMaSB, will underpin the web service PDBeShape which is under development.

SMaSB has been made available from the CCP-EM website as a downloadable software package, which can be run locally to perform specific

queries. The functionality was described in detail in MS21 “Prototype of the shape-matching software”, but briefly covers the following steps:

1) Map/Structure pre-processing
   a) Fetch EMDB/PDB files or user upload
   b) Process PDB file
   c) Process Map file
   d) Segmentation
   e) Feature extraction
   f) Store map details in XML files

2) Map-Map alignment
   a) Map filtering
   b) Alignment
   c) Map transformation and scoring
   d) Store alignment details in XML files

The same software will handle queries submitted via the web service. When the volume database is present, the details stored in XML files are uploaded to MySQL tables.

3.2 SMaSB: Shape MAtching service for Structural Biology

The tool searches for similarities between 3-dimensional shapes or volumes which represent proteins, nucleic acids, or complexes of several such components. These volumes are obtained from structural biology techniques such as electron cryo-microscopy (cryoEM), electron tomography, small angle X-ray scattering or X-ray tomography. A similarity between volumes may imply a similarity in function or an evolutionary relationship.

Such shape matching has been done for a long time for cases where the atomic structure is known, i.e. for coordinate data from X-ray crystallography or NMR spectroscopy. These techniques rely on matching individual atoms. Nowadays, especially with improvements in cryoEM, there is an increasing amount of structural data at lower resolutions where the individual atomic positions are not known. Thus a new set of tools needs to be developed.

Some programs existed before, for manual comparison of a specific pair of volumes. The SMaSB pipeline makes use of these programs, but allows a scientist to do broader searches of known volume data. As well as automating the process, SMaSB also implements a number of scoring functions for assessing the match (since the similarity of volume data is perhaps harder to quantify than the similarity of discrete data such as sequences).

D9.1 delivers the underlying software pipeline. This includes the core functionality required for identifying matches between molecular volumes. As a command line implementation, it is probably most suitable for structural biologists who are used to manipulating structural files. SMaSB underpins the PDBeShape web service which is currently under development and will provide a more appropriate interface for the general biologist.

SMaSB will be maintained and further developed by the CCP-EM and EMDB projects. 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. The Electron Microscopy Data Bank (EMDB) at PDBe (www.ebi.ac.uk/pdbe/emdb/) is developing a range of services for the cryoEM community. The PDBeShape web service will be part of this portfolio, and the EMDB will help to maintain the underlying SMaSB pipeline.

3.4 Running the software

The software is available as a compressed tarball, which is intended to be downloaded and run locally. Further details on obtaining and installing the software are given in Supplement 1.

In addition to the software itself, installation creates a data directory containing the following subdirectories:

em/ - Volume files downloaded from EMDB. Known alignments of this volume are stored in subdirectories.

pdb/ - Structure files downloaded from PDB. Known alignments of this structure are stored in subdirectories.

protocols/ - A set of alternative protocols for volume matching (formatted as XML following the SMaSB schema), including program and parameters

score/ - A set of alternative scoring functions for assessing matches (formatted as XML following the SMaSB schema)

taxonomy/ - Taxonomic classifications for species relevant to the volumes examined, as specified by the NCBI identifier (e.g. 9606 for Homo sapiens)

sample/ - Sample descriptions for the volumes examined, specified as a category/complex/component/domain hierarchy. The domain level for proteins and non-coding RNA is described using Pfam/Rfam annotations. We expect to use descriptions from EBI's complex portal for the complex/components layers.

A typical command line is:

```
python SMaSB.py -c all -i Vol1,Vol2 -m 4
```

The option “-c all” indicates that both pre-processing and alignment steps should be run. “Vol1” and “Vol2” are the two volumes to be matched, identified through their EMDB or PDB ID codes. “Vol2” can also point to a file with a list of EMDB or PDB IDs, allowing a search over many references volumes. The option “-m 4” means use the 4th protocol, as defined in the protocols subdirectory of the data directory (see above).

4 References

[1] Farabella I et al. (2014). “Tempy: A python library for assessment of 3D electron microscopy density fits”, *Submitted*.

[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.

[3] DOI: 10.1093/bioinformatics/btl625

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

[5] DOI: 10.1023/A:1023813504011

[6] 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.

[7] DOI: 10.1529/biophysj.108.137125

[8] 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).

[9] DOI: 10.1016/j.jsb.2013.09.021

[10] 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.

[11] DOI: 10.1002/jcc.20084

5 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

6 Adjustments made

No adjustments were made to the deliverable.

7 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.

Supplement 1: Details of the software package

SMaSB (the shape matching software) is available as a compressed tarball from the downloads page of the CCP-EM project[3]. In addition to a set of Python modules written specifically for WP9, it includes a version of Open Astex Viewer[4] that incorporates additional functionalities for volume processing. The SMaSB package includes a README file which gives full details on required dependencies and installation instructions. A setup script install_setup.py is provided, which allows paths to data directories and 3rd-party software to be set.

The package requires installation of the TEMPy sofware library for core tasks, and one or more of Gmfit, Chimera, ADP-EM for volume matching. These are all free of cost for non-commercial use.

- TEMPY (Farabella *et al*., 2014) is a suite of python scipy/numpy based modules for coordinate/map transformation and validation/scoring of alignments[5]
- GMFIT (Kawabata, 2008) is used as the default alignment method as it gives fast and reasonably good alignments[6]. 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)[7]
- ADP-EM (Garzon et al. 2007) is relatively slower but follows a more exhaustive spherical harmonic based search[8]
- lxml is a Python module for XML and HTML processing[9]

[3] http://www.ccpem.ac.uk/download.php
[4] OAV; http://openastexviewer.net/web/; Hartshorn, 2002; Lagerstedt et al., 2013
[5] available from http://tempy.ismb.lon.ac.uk/
[6] available from http://strcomp.protein.osaka-u.ac.jp/gmfit/
[7] available from https://www.cgl.ucsf.edu/chimera/
[8] available from http://chaconlab.org/methods/fitting/adpem
[9] If not already installed on users' systems, this is available from http://lxml.de/