BioMedBridges

Deliverable D9.3

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-service for segmentation and identification of components in volume data	
WP No.	9	
Lead Beneficiary:	1: EMBL	
WP Title	Use case: From cells to molecules- integrating structural data	
Contractual delivery date:	31 December 2015	
Actual delivery date:	16 December 2015	
WP leader:	Martyn Winn	4: STFC
Partner(s) contributing to this deliverable:	1: EMBL 4: STFC	

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

Contents

1 Executive summary

We have developed a software pipeline, named SMaSB, to perform automated volume/shape matching. Together with a database of over 1000 volumes from the Electron Microscopy Data Bank (EMDB) and Protein Data Bank (PDB), the pipeline underpins a web service PDBeShape. These tools are novel in providing access to a growing class of structural biology data viz. volume data. The development of SMaSB was reported in Deliverable 9.1, and the development of PDBeShape in Deliverable 9.2.

Deliverable 9.3 implements automated and manual segmentation into the PDBeShape web service. Automated segmentation with Chimera-Segger is now an integral part of volume pre-processing, with the results output as Chimera-Segger .seg files. Automated segmentation is useful as a guide to features in a volume, but does not reliably give segments corresponding to biological components (such as protein or RNA chains). Therefore, PDBeShape also supports annotation with the results of manual segmentation. As a proof of principle, 86 entries from the PDBeShape volume database were manually segmented and annotated, and the results are available from the PDBeShape service. Biological annotation links volume segments to entries in Pfam / Rfam and Uniprot.

Segmentation of structural volume data is an area of intense discussion in the international community. Issues include the reliability of algorithms, the representation and annotation of results, and the practicalities of sharing the results. The EBI hosted a workshop on “3D segmentations and transformations - building bridges between cellular and molecular structural biology” on 7th / 8th December 2015. Work from the current deliverable was presented as an example user case, and will lead to further work in this area.

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	X	
2	Develop software to search this database using atomic or volume data	X	
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, e.g. EMDB). It will be delivered via a software stack covering the underlying matching algorithms, a web-based front end, and database operations.

Deliverable D9.1 was the first from this work package, and covered the underlying software pipeline (SMaSB) for matching volumes and capturing appropriate metadata. D9.2 delivered the first public version of PDBeShape, 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. Here we describe deliverable D9.3 which extends SMaSB and PDBeShape to include automated and manual segmentation of the volume data.

The aim of 3D segmentation of molecular volume data is usually to identify individual components in a larger complex (Patwardhan *et al*., 2012 & 2014). This is distinct from segmentation of 2D images or 3D tomograms, where the aim is to delimit regions of interest (for example membranes or whole complexes). Identification of individual macromolecular components facilitates more precise annotation with molecular identifiers. That is, rather than annotating a volume with a list of contents, we can annotate each segment with a single identifier for the molecule which is located there. This is turn reveals details of specific protein-protein interactions, or dissociation pathways.

In the context of volume matching, 3D segmentation also allows us to define subvolumes. Searching against a subvolume is generally more robust than searching for a small region of a larger volume.

3.2 Segmentation user survey

We sought to understand the state-of-the-art of 3D segmentation in the field of cryoEM. We have carried out a survey of what tools and methods structural scientists use (Milestone 23). The survey covered single particle analysis, but also segmentation of tomograms, which will be useful for a future extended version of the volume-matching service. The survey is available at https://www.surveymonkey.com/s/3dseg and consists of 13 questions.

The survey covered both segmentation of volumes from single particle reconstruction and segmentation of tomograms. The former is more relevant to the current deliverable, for which the Segger plugin for Chimera is the most popular (Pintilie G *et al*., 2010; Pettersen *et al*., 2004). Segger uses a watershed algorithm to perform an initial segmentation of the volume. This tends to oversegment the volume, and so a scale-space filtering method is used to group segments. Other software mentioned for segmenting individual volumes included JUST and Bsoft. Most scientists supplement automated methods with manual segmentation or checking in a graphics program.

3.3 Automated segmentation in SMaSB / PDBeShape

For segmentation of single particle volume data, Chimera-Segger remains our preferred choice (Pintilie *et al.*, 2010). It is the most commonly used tool in the

community, and thus well tested, and has a convenient Python interface for scripting.

We are currently using Chimera-Segger (in non-graphical mode) in the volume pre-processing pipeline of SMaSB. The aim is to obtain a rough measure of the number of features in the map, which is then used as a guide for gmconvert (the initial step of the volume matching program Gmfit, Kawabata, 2008). The pre-processing pipeline is applied in a completely automated fashion to all volumes in the database, and the segmentation step results in a .seg file containing details of the segments found. This file can be loaded and viewed in Chimera:

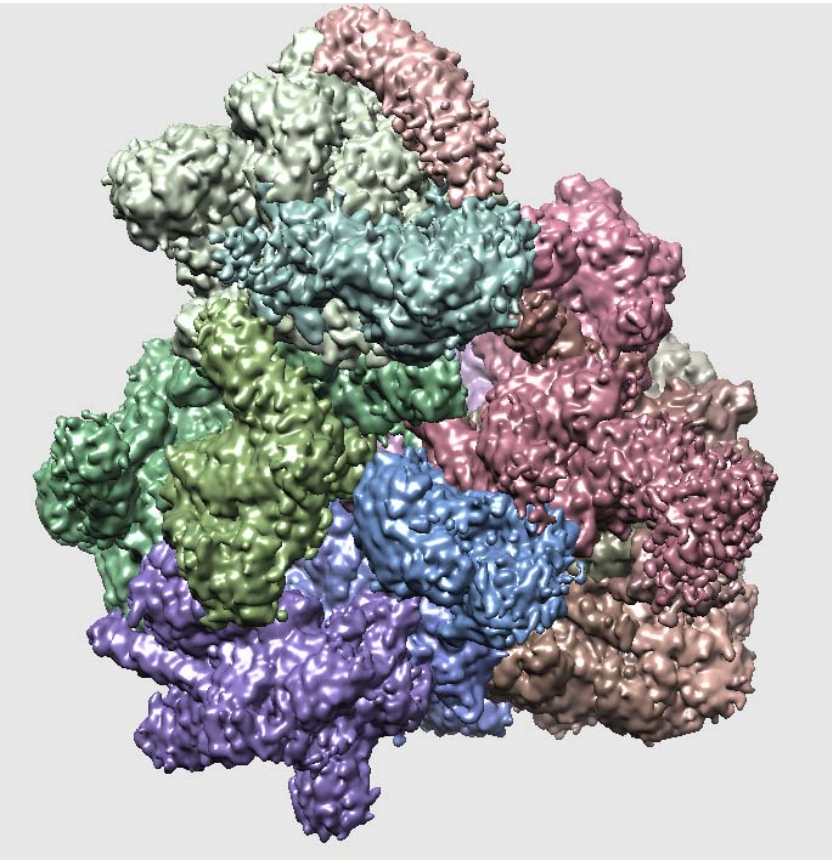

Figure 1 Automated segmentation of EMD-5591 (Drosophila melanogaster EF2- and Vig2-bound 80S ribosome). The figure shows the output .seg file from SMaSB, as viewed in Chimera

3.4 Annotation with user segmentation

Automated segmentation often gives a good indication of what components or domains are present in a structural volume, but it is rarely completely accurate. A scientist with in-depth knowledge of the biological system will usually do a better job (although there may still be ambiguity or different interpretations). It is

therefore useful to annotate volume entries with one (or more) user-created segmentations. A user of a particular volume in PDBeShape will have access to multiple automated and manual segmentations, reflecting the inherent uncertainty.

To illustrate this approach, we have manually segmented a selection of entries from the PDBeShape volume database. To-date these are 82 chaperonins, and 4 ribosomes with cryoEM volumes in the EMDB (as well as 14 helicases not currently in the volume database). The aim of segmentation was to have one segment for each protein or RNA chain. The segmentation was carried out in Chimera, with guidance from several sources of information. In around half of cases, there was a fitted model deposited in the PDB, and that indicates immediately where the segments are located. If there was no fitted model, then the associated publication was checked for information (e.g. in figures) on component locations inferred for example from biochemical data. If the publication was not clear, then the volume was compared to related structures where component location is known.

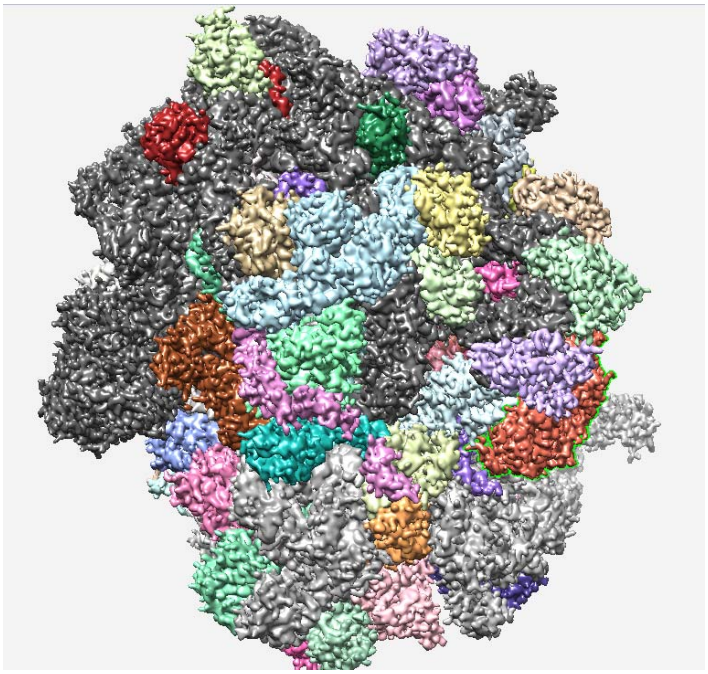

Figure 2 Manual segmentation of EMD-2847 (E. coli ribosome-EF-Tu complex). The selected segment (bordered in green) is the elongation factor

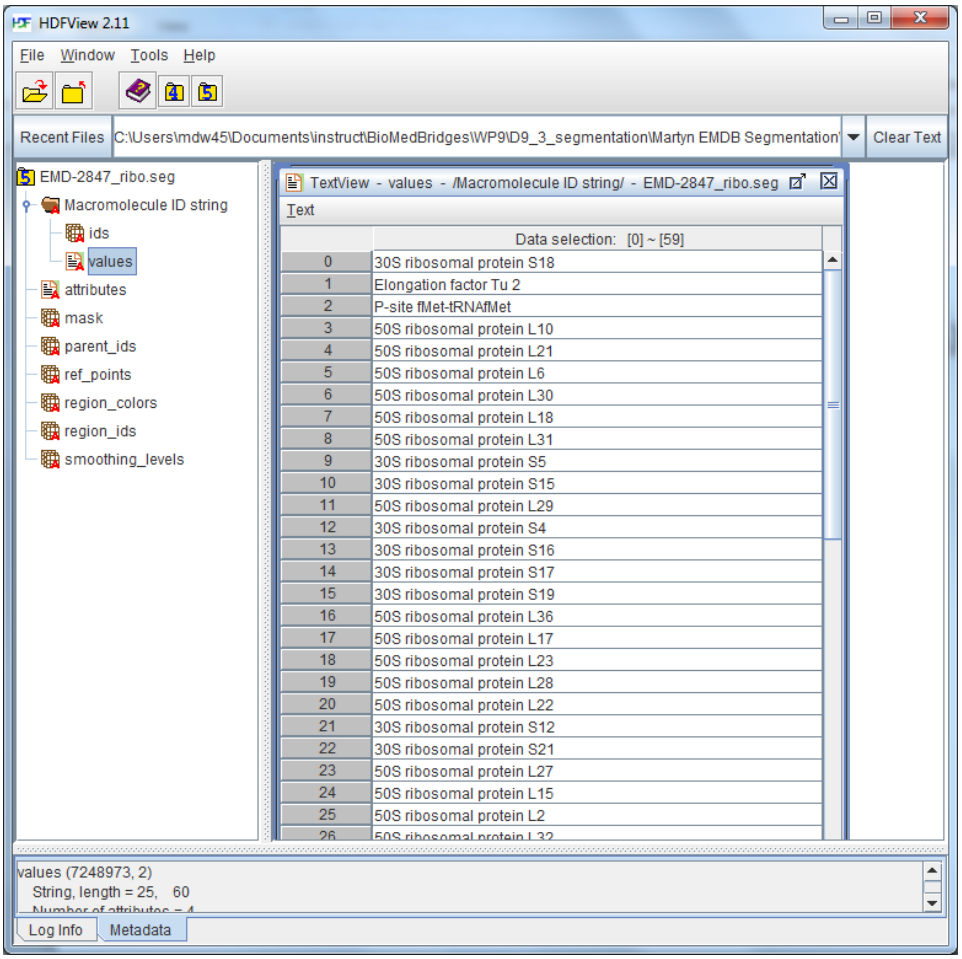

Figure 3 Corresponding annotation, as recorded in the Chimera .seg file

The manual segmentation is recorded in a Chimera-Segger .seg file. A separate Excel spreadsheet holds further details on each manual segmentation. For each segment, there is the name as used in the .seg file, the length and type of the polymer chain (protein, RNA, etc.), the Pfam/Rfam or Uniprot identifier, the expected molecular weight, the copy number, and the fitted model (if any).

3.5 Incorporation into PDBeShape

The database schema has been updated to include pointers to external files holding manual segmentations. Where such files exists (in the first instance, the 86 manual annotations described above), a link is provided on the “Details” page of the volume in question. The public version of PDBeShape is accessible at http://www.ebi.ac.uk/pdbe/emdb/pdbeshape/welcome/

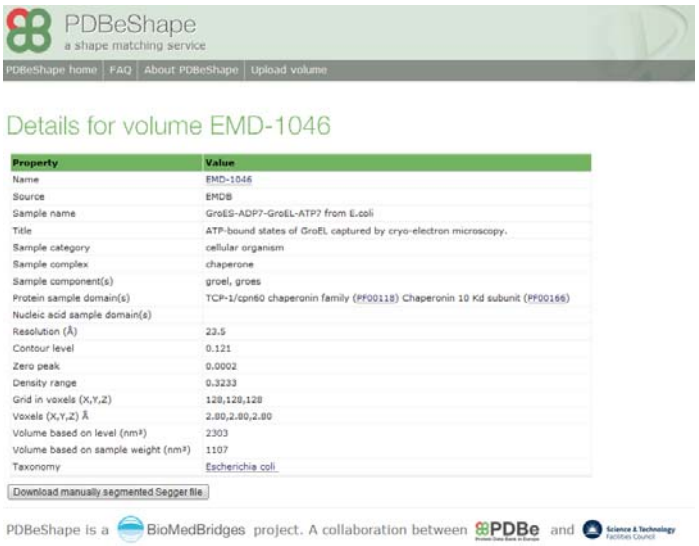

Figure 4 Example volume entry for a chaperonin, with a link to the manual segmentation file at the bottom

3.6 Community workshop on segmentation

The EBI hosted a workshop on “3D segmentations and transformations - building bridges between cellular and molecular structural biology” on 7th / 8th December 2015. The workshop was attended by leading international scientists and software developers in electron microscopy and electron tomography. One of the aims was to develop an agreed data model for volume segmentation and the associated biological annotation, together with a file format representation of the data model. A draft model and a reference HDF5 implementation (segmentation file format, SFF) were agreed.

The PDBeShape service including the volume database was presented as a use case for segmentation and annotation. As part of the sustainability plan for BioMedBridges, the agreed data model will be adopted by the PDBeShape service. Segmentations that are currently held in Chimera .seg format will be converted to SFF. These files will also hold biological annotations that are currently held separately.

The development of a consensus on volume segmentation is an important step forward for the international cryoEM community (Patwardhan *et al.*, 2014), but is taking place on a longer timescale than the work in BioMedBridges WP9. Nevertheless, the PDBeShape service with the volume database and the set of manually annotations have played a central role in this process.

4 References

[1] Patwardhan A, Carazo JM, Carragher B, Henderson R, Heymann JB, Hill E, Jensen GJ, Lagerstedt I, Lawson CL, Ludtke SJ, Mastronarde D, Moore WJ, Roseman A, Rosenthal P, Sorzano CO, Sanz-García E, Scheres SH, Subramaniam S, Westbrook J, Winn M, Swedlow JR, Kleywegt GJ. (2012) “Data management challenges in three-dimensional EM.” *Nat Struct Mol Biol* **19**, 1203-1207. DOI: 10.1038/nsmb.2426

[2] Patwardhan A, Ashton A, Brandt R, Butcher S, Carzaniga R, Chiu W, Collinson L, Doux P, Duke E, Ellisman MH, Franken E, Grünewald K, Heriche JK, Koster A, Kühlbrandt W, Lagerstedt I, Larabell C, Lawson CL, Saibil HR, Sanz-García E, Subramaniam S, Verkade P, Swedlow JR, Kleywegt GJ. (2014) “A 3D cellular context for the macromolecular world.” *Nat Struct Mol Biol* **21**, 841-845. DOI: 10.1038/nsmb.2897

[3] 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

[4] Pintilie, G et al. (2010). “Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions.” *J Struct Biol*. **170**, 427-38

[5] 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 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.

For automated segmentation, Chimera-Segger is run in non-graphical mode via a custom python script, and requires three parameters to be set. As an example, the following segments an 80S ribosome from D. melanogaster:

```
chimera --nogui --script "chimera_run_segger.py emd_5591.map 0.285 6.0 4428.21"
```

The first parameter is the minimum density threshold which we set as the contour level giving an enclosed volume appropriate to the molecular weight. The second is the estimated resolution of the map. The third is the estimated volume of the map. The latter two parameters are used to estimate a target number of segments, based on a largest segment being approximately 50 residues for a 3Å resolution map and proportionately larger for lower resolutions. The scale-space filtering and segment grouping is repeated until the number of segments is less than this target.

Chimera-Segger outputs a .seg file in HDF5 format. The file contains a number of tables describing the volume segments. In particular, there is a mask covering the same 3D grid as the input map, with grid points containing a pointer to a specific segment.

For manual segmentation, Chimera-Segger was run in graphical mode with the following protocol:

1. Load volume, and set the contour level of the displayed map.
2. Run Segger to automatically segment the entire map.
3. Group / Ungroup segments as necessary to get one segment per chain. This is guided by fitted models, by reference to the associated article, or by comparison with related volumes.
4. In the Attributes table of the segmentation, add column "Segment name" and fill with a descriptive name for each segment.

The final segmentation is saved in a .seg file. Manual annotation of each segment with Pfam/Rfam or Uniprot identifiers is recorded in a separate Excel spreadsheet.

6 Delivery and schedule

The delivery is delayed: ☐ Yes X No

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.