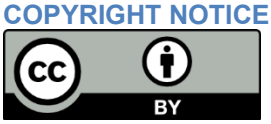
Deliverable D7.1

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	**675858**	
Deliverable title:	Multi-core implementation of PDB_REDO server	
WP No.	7	
Lead Beneficiary:	2: NKI-AVL	
WP Title	Joint research	
Contractual delivery date:	6	
Actual delivery date:	6	
WP leader:	name	Partner CSIC
Contributing partners:	NKI	

COPYRIGHT NOTICE



DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	R.P. Joosten & A. Perrakis	NKI/WP7	30.03.2016
Approved by:			

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	30.03.2016	First version	R.P. Joosten & A. Perrakis / NKI

Contents

1 Executive summary

— PDB_REDO is one of the services in the West-Life framework. It provides an automated procedure to optimise macromolecular structure models from X-ray crystallography experiments, by combining crystallographic refinement, automated model rebuilding, and extensive validation. Many decision-making algorithms in the PDB_REDO pipeline use comparative, statistical analysis of alternative models of the crystallographic data. These algorithms contribute strongly to PDB_REDO's performance in terms of model quality, but they also make the procedure time consuming. This hampers the incorporation of PDB_REDO in other services in the West-Life VRE. Here we describe the parallelisation of the PDB_REDO procedure, which gives an average 4-fold speedup at no expense in terms of final model quality.

2 Project objectives

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

No.	Objective	Yes	No
1	**Provide analysis solutions for the different Structural Biology approaches**	X	
2	**Provide automated pipelines to handle multi-technique datasets in an integrative manner**	X	
3	**Provide integrated data management for single and multi-technique projects, based on existing e-infrastructure**		X
4	**Foster best practices, collaboration and training of end users**	X	

3 Detailed report on the deliverable

3.1 Background

The Protein Data Bank (PDB)[1] is the international repository for experimentally determined macromolecular structure models with atomic detail. Most entries are derived from diffraction experiments (89%), another important experimental source being solution NMR (10%). The PDB strictly serves as a historic archive with oldest entries deposited more than 40 years ago. The entries are never updated, to ensure they represent the structure models as they were published. This means that existing entries do not benefit from the substantial improvements in (computational) structure determination methods. Fortunately, it has been common practice in protein crystallography to not only deposit the structure model in the PDB, but also the experimental data from which it was derived. This practice has become mandatory in 2008. As a result 90% of all crystallographic PDB entries have their experimental data available and this figure is increasing. Using this deposited experimental data it is possible to bring PDB entries up-to-date with current best practices and generate new, often improved, structure models for a wide scope of structural biology applications.[2] The PDB_REDO databank was created to provide these alternative models for all crystallographic PDB entries.

Filling the PDB_REDO databank required enormous computation resources that were provided through the life-science grid of the European Commission FP6 project EMBRACE.[3] As the number of PDB entries exceeded the number of available CPUs, the process was treated as ‘embarrassingly parallel’ and no effort was needed to parallelise the underlying PDB_REDO computational pipeline to achieve maximal throughput. The launch of the PDB_REDO webserver[4] to optimise work-in-progress structure models of crystallographic data (before the authors submit them to the PDB) changed the computational requirements for the PDB_REDO pipeline. Here, the number of available CPUs typically exceeds the number of models to be ‘redone’ and the speed of individual jobs is more important (for user experience) than overall throughput. This required speed-up was achieved by parallelisation of the PDB_REDO pipeline described below.

3.2 Multi-core implementation of PDB_REDO

3.2.1 PDB_REDO pipeline

The PDB_REDO procedure for structure model optimisation is a multi-step process that can be summarised as:

1. Initial model analysis
2. Refinement parameterisation
3. Model refinement
4. Model rebuilding
5. Additional model refinement
6. Model validation
7. Annotation of results

Figure 1 shows a flowchart representation of these steps. The bulk of the procedure can be regarded as an inherently serial problem in which the results of one process are used as input for the next. This severely limits the potential for parallelisation. There are however some exceptions. The refinement parameterisation (Step 2) and the model refinement (Steps 3 and 5) involve decision-making algorithms (described previously)[5] that select the optimal way of treating the model from a number of trial calculations in the crystallographic model refinement program REFMAC.[6,7] Additionally, in selected cases thorough model validation (Step 6) requires *k*-fold cross validation which uses *k* model refinements in REFMAC with different subsets of the experimental data. The required calculations with REFMAC in these steps are computationally demanding, but at the same time completely independent. Therefore, these steps were parallelised. The decision-making algorithms, the PDB_REDO step at which they occur, and the number of independent calculations required are summarised in Table 1.

Decision-making algorithm	Step	Computations required (range)	Computations required (typical)
(A) Select atom movement description detail	2	0-2	2
(B) Select domain movement model	2	0-101	2
(C) Select restraint weight for atom movement description (B-factor weight)	2	0-7	7
(D) Select relative weight of the diffraction data versus geometric restraints	3	5-7	7
(E) Fine-tune X-ray weight	5	1-3	3
(F) *k*-fold cross validation	6	0-100	10

Table 1. *Summary of PDB_REDO decision making algorithms. Algorithms are marked in blue in Figure 1.*

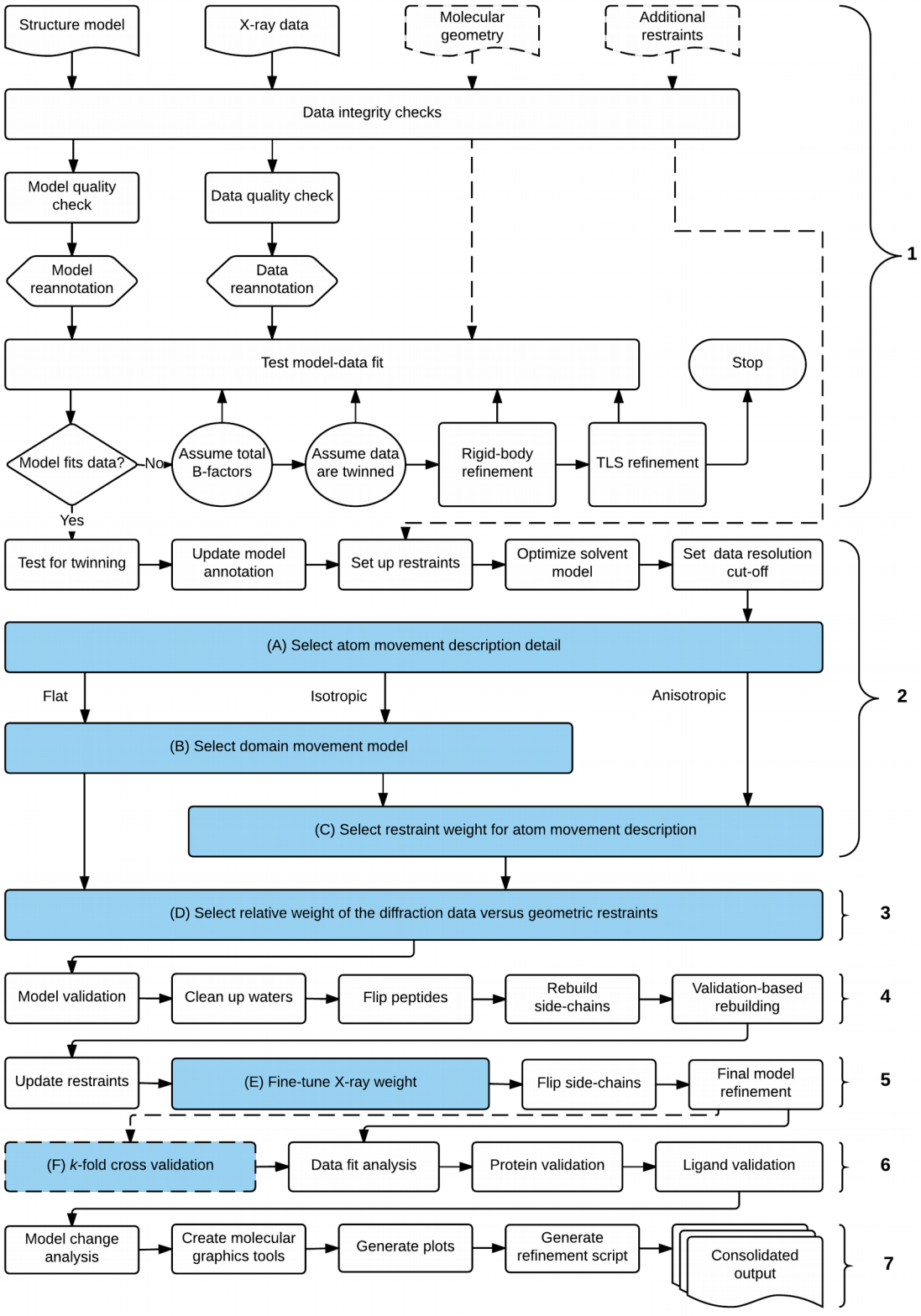


Figure 1. *Simplified flow chart of the PDB_REDO procedure. Parallelised algorithms are marked in blue.*

The parallelisation of the algorithms in Table 1 uses a simple mechanism that starts calculations as long as a user-defined maximum number of parallel calculations (P_{max}) is not exceeded. If more calculations are queued, they are launched as soon as another job finishes. By default, P_{max} = 1 (no parallel computations) which is suitable for high throughput calculations on many structure models, the overhead of the parallelisation code is negligible. For typical calculations not involving *k*-fold cross validation, P_{max} = 7 gives the highest job speed. This typically gives 4-fold speedup, reducing the running time of a PDB_REDO job from an average of seven hours to less than two. The exact speedup for any particular case depends on the input model and the experimental dataset. Model rebuilding (Step 4) is computationally demanding and cannot readily be parallelised. Because rebuilding is currently limited to protein structures with high or medium resolution crystallographic data, more speedup is achieved for low-resolution datasets and models of DNA/RNA structures or protein-DNA/RNA complexes. With maximum parallelisation (P_{max} = 10) some PDB entries can now be 'redone' in less than 15 minutes.

3.2.2 PDB_REDO server

In the context of the PDB_REDO server P_{max} is decided by the server's job dispatcher. Based on the number of free CPU cores at start of a PDB_REDO run, P_{max} is set to a value ranging from 1 to 7. Because the number of CPU cores utilised during the PDB_REDO run fluctuates between 1 and P_{max}, P_{max} CPU cores are reserved and cannot be used by other PDB_REDO runs. This is done for the sake of simplicity, but has an adverse effect on server throughput when it becomes very busy; but at the moment however, server capacity is sufficient for the current group of more than 800 active users who run ~350 PDB_REDO jobs per month. The average duration of a PDB_REDO run is just under two hours. In the context of the West-Life VRE, local server calculations may be moved to the Cloud, when it is needed to maintain sufficiently high server throughput.

References cited

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.
2. Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. (2009). *Acta Cryst.* **D65**, 176-185.
3. Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A.-C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A. L., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I. J., Vriend, G. (2009). *J. Appl. Cryst.* **42**, 376-384.
4. Joosten, R. P., Long, F., Murshudov, G. N., Perrakis, A. (2014). *IUCrJ.* **1**, 213-220.
5. Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Cryst.* **D68**, 484–496.
6. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240-255.
7. Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355-367.

Background information

This deliverable relates to WP7; background information on this WP as originally indicated in the description of work (DOW) is included below.

WP7 Title: Joint Research
Lead: CSIC
Participants: STFC, NKI AVL, EMBL, MU, CIRMMP, Instruct, UU

Work package number	7		**Start date or starting event:**		0
Work package title	Joint research				
Activity Type	COORD				
Participant number	1	2	3	4	5
Person-months per participant		2			
Participant number	6	7	8	9	10
Person-months per participant					

Objectives
This Joint Research Activity is aimed at exploring new ways to use existing or close to existing services so that broader user communities will be reached. This overarching goal will be accomplished through the following Objectives:

- **O7.1:** Extending and benchmarking existing services, such as ProteinCCD, PDB_REDO, REFMAC and HADDOCK
- **O7.2**: Combining services into new workflows
- **O7.3:** Studying large sets of output data using Big Data approaches
- **O7.4**: Evaluating current and developing metadata standards for workflow definition

Description of work and role of participants
The objectives above will be addressed through the following tasks:

Task 7.1. Extending and benchmarking existing web services
Lead partner: EMBL-HA, Participants: CSIC, STFC, NKI, MU, CIRMMP, UU

Four exemplary scientific cases will be addressed here:

a) ProteinCCD: User-driven selection of protein domains for heterologous expression
b) PDB_REDO: automated refinement, rebuilding and validation of atomic models
c) REFMAC: refinement of an atomic model against data from multiple experiment
d) HADDOCK: Protein-protein docking

ProteinCCD, developed by NKI, is an application that reads in the DNA sequence of a (synthetic) gene encoding a specific protein, using specific servers to analyse it for domain boundaries, secondary structure, disorder, membrane segments, etc., and it allows the user to interactively choose the desired regions they want to use in heterologous expression and are likely to give highly soluble protein (important for NMR) and with a good likelihood for crystallization. Then, ProteinCCD automatically designs optimized PCR primers for cloning reactions to facilitate creation of these constructs, bridging protein sequence analysis with gene cloning laboratory work. As ProteinCCD partially functions as a metaserver, we first want to move all sequence analysis server dependencies within the VRE infrastructure. Second, we want to extend ProteinCCD to include new sequence analysis servers. Third, we will add solubility and crystallisability scoring and ranking algorithms from available servers, to analyse the user-designed constructs. Finally, we will provide for a complete collection of cloning strategies including vectors and methods widely used in partner laboratories.

The PDB_REDO procedure, developed by NKI, uses atomic models from X-ray crystallography (either already deposited in the PDB or provided by the users directly) and automatically optimizes their refinement using the popular REFMAC software (see below for more information about REFMAC). In this task we want to first benchmark the success of PDB_REDO in already deposited versus user-supplied models. We then want to explore methods for deploying PDB_REDO in multiple CPU cores in parallel to achieve 3- to 5-fold speed-up. Currently, we run a PDB_REDO databank (optimized PDB models) and a PDB_REDO web server (for user models). We want to develop a system that will allow users to not only be able to examine PDB models that are already optimized by PDB_REDO, but also to request a PDB_REDO job on existing PDB entries that might have not been optimized in the recent past (the renewal cycle of the PDB_REDO databank is several months, while the development and improvement pace is faster). We aim for PDB models requested for optimization by users to be automatically deposited in the databank in a transparent manner for public use.

REFMAC represents a very successful approach developed by Garib Murshudov within the CCP4 project that is distributed to 1000s of users worldwide. REFMAC is also currently being used by a number of X-ray crystallography web servers, such as those based on ARP/WARP, Balbes, MrBUMP, and PDB_REDO as mentioned above. However, REFMAC has recently been used in a different way to refine models against Electron Microscopy maps as well as some types of NMR (orientation) data, with very promising scientific perspectives, but without a standardized or otherwise easily accessible way to apply these methods. Consequently, in this Task we aim at developing and benchmarking new Web Services, starting with the REFMAC engine, which will address both the Electron Microscopy and the NMR communities, enlarging the user base of already very successful approaches. The EMBL-HA will lead this part of the Task, as REFMAC experts, counting on the EM experience of both the CSIC and the STFC, on the NMR experience of CIRMMP and UU, and on the corresponding one in X-ray crystallography and high throughput approaches of MU.

Finally, In the case of biomolecular docking, the information driven docking software HADDOCK supporting protein-protein, protein-nucleic acids and protein-peptide/small ligand docking, developed at UU, will be extended and benchmarked to allow direct refinement against cryo-EM data, in a close collaboration with CSIC and STFC. Protocols and performance will be optimized and offered through a new web portal. In this task, we will also explore the possibility of integrating REFMAC as the computational engine in HADDOCK instead of CNS.

Task 7.2. Combining existing services into new workflows

Lead partner: CSIC, Participants: EMBL-EBI, STFC, NKI, MU, CIRMMP, UU

In this case, the exemplary case will be the automatic application of a number of quality/validation measures for structural data. The first new development will address the incorporation at the CSIC, STFC and EMBL-EBI of quality/validity information to Electron Microscopy maps using existing computational methods, such as ResMap and others. These quality assurance steps will form a workflow leading to submission of the validated structures to the PDB/EMDB, enlarging the value of public EM data in a moment in which data harvest is critical for this rapidly expanding discipline. The second development will address the quality analysis of predicted complexes. Indeed, comparative analysis and evaluation of predicted macromolecular complexes is crucial in supporting method development and will support the CAPRI community in conjunction with the work carried out in WP5 for model deposition and query system, a work to be carried on by the EMBL-EBI, NKI, MU and UU. We will combine the evaluation methods and make these available for analysis of predicted models.

Task 7.3. Handling and Mining Big Data
Lead partner: STFC, Participants: CSIC, MU

Integrated structural biology can be classed as a Big Data problem in a number of contexts. With the latest generation of detectors for crystallography and EM, datasets are now typically in the TBs (volume) and generated in hours (normally referred to in the field as “velocity”). In moving to interdisciplinary scientific studies, researchers increasingly have to integrate datasets from completely separate experiments (variety), which are sometimes of unknown provenance (veracity). In order to explore how Big Data approaches could be effectively applied to Structural Biology data, we will investigate the suitability of some of the latest software tools. In particular, building on the collaboration with IBM as stated in its strong Letter of Support, the IBM BigInsights package will be made available to the consortium in order to support a number of pilot investigations. This package includes an IBM implementation of Hadoop for problems that can be cast into a MapReduce framework. This could be applied, for example, to analytics on large numbers of images or large numbers of structural models. For specific data intensive jobs, the IBM BGAS system uses flash memory as non-volatile RAM to allow faster handling of large data. This could be applied to single particle reconstruction in EM, which often requires large memory or frequent disk I/O. In turn, Infosphere Streams provides a programming environment for analysis of streamed data, which could be applied to on-the-fly processing of large structural biology datasets. Finally, IBM’s Watson software may be suitable for pattern discovery when the VRE has amassed a large and disparate set of data. These approaches use state-of-the-art computational techniques which are not widely available, but which could underpin services provided through the VRE. While the IBM software mentioned above is in product form, the IBM team will also be developing enhancements to these and other prototypes on the evolving Data Centric Systems. This Task will be coordinated by STFC, who already has a close collaboration with IBM, and prototypes will be developed at STFC, MU, and CSIC and made available to the consortium to test their impact on the problems of interest.

Task 7.4. Assessing and Extending metadata standards
Lead partner: STFC, Participants: EMBL-EBI, NKI, MU, CSIC, CIRMMP, Instruct, UU
Current workflows in Structural Biology may not be properly described in an unambiguous manner due to the lack of appropriate metadata standards specifying them. For example, while some standards exist for models that are built using structural data, there is no agreed

ontology for the primary data processing, either at the level of integrated studies combining different technologies or even at the single technique level. In some cases, relevant ontologies may exist, but not be applied in the structural biology context. In other cases, relevant ontologies are still under development, for example for describing complexes and their components. The aim of this Task is to survey available metadata standards relevant to the set of services offered by the West-Life VRE. Where there are existing standards that could be used, these will be documented and communicated to WP6 for incorporation in the data management layer. Where there are gaps for particular techniques or for communicating between techniques, we will document what is needed and scope out a controlled vocabulary specifying the way data could be represented. Although it might be difficult to obtain community agreement on the timescale of the project, we will aim to engage the relevant stakeholders (structural biologists, methods developers, database providers) and begin the discussion. All partners have experience in standardization efforts in the various experimental and computational fields, with Instruct-Hub providing a widely accepted “reference” position, as is so much needed to focus community efforts. Several partners (EMBL-EBI, STFC, CIRMMP, Instruct) are involved in the cluster project BioMedBridges, which is developing a standards registry for biomedical sciences, and which will be used in the current project.

Deliverables

No.	Name	Due month
7.1	Multi-core implementation of PDB_REDO server	6
7.2	ProteinCCD with new analysis options	9
7.3	ProteinCCD with construct scoring and ranking	24
7.4	A REFMAC server for EM and NMR	24
7.5	A HADDOCK server for EM	24
7.6	EM quality assurance workflow	24
7.7	Quality analysis workflow for predicted complexes	30
7.8	Report on prototypes constructed using Big Data approaches	30
7.9	Report on existing metadata standards, and proposals for new vocabularies	30