



Project Number: **RI-312579**

Project Acronym: **ER-flow**

Project Full Title:

Building an European Research Community through Interoperable Workflows and Data

Theme: **Research Infrastructures**

Call Identifier: **FP7-Infrastructures-2012-1**

Funding Scheme: **Coordination and Support Action**

Annex D – Life Science Community

Due date of deliverable: 28/08/2013

Start date of project: 01/09/2012

Actual submission date: --/--/2013

Duration: 24 months

Lead Contractor: University of Westminster

Dissemination Level: PU

Version: 1.0





1 Table of Contents

Contents

1	Table of Contents.....	2
2	List of Figures	3
3	List of Tables	4
4	Status and Change History	5
5	Glossary	6
5.1	Workflows and Metaworkflows.....	7
5.1.1	DTI Pre-processing CVMFS.....	7
5.1.2	FSL BedpostX CVMFS	9
5.1.3	DTI Population Registration (with and without CVMFS)	11
5.1.4	Freesurfer Pial (with and without CVMFS)	14
5.1.5	Freesurfer Param (with and without CVMFS).....	16
5.1.6	Freesurfer recon-all CVMFS	18
5.1.7	Trac-all CVMFS	20
5.1.8	CQRS	21
5.1.9	AutoDock Vina with Sorting All Energies.....	23



2 List of Figures

Figure 1, DTI Pre-processing Workflow.....	7
Figure 2 FSL BedpostX Workflow	9
Figure 3, DTI Population Registration Workflow	12
Figure 4, Freesurfer pial Workflow	14
Figure 5, Freesurfer param Workflow	16
Figure 6, Freesurfer recon-all Workflow	18
Figure 7, Trac-all Workflow	20
Figure 8, Three Workflow Implementations of the CQRS algorithms.....	21
Figure 9, AutoDock Vina Workflow.....	23



3 List of Tables

Table 1, Deliverable Status	5
Table 2, Deliverable Change History	5
Table 3, Glossary	6
Table 4, DTI Pre-processing technical details	8
Table 5, FSL BedpostX technical details	10
Table 6, DTI Population Registration technical details.....	13
Table 7, Freesurfer Pial technical details.....	15
Table 8, Freesurfer Param technical details	17
Table 9, Freesurfer technical details	19
Table 10, TRAC-ALL technical details.....	20
Table 11, CQRS technical details.....	22
Table 12, AutoDock Vina technical details	24



4 Status and Change History

Status:	Name:	Date:	Signature:
Draft:	Gabriele Pierantoni	29/09/2014	n.n. electronically
Reviewed:			n.n. electronically
Approved:			n.n. electronically

Table 1, Deliverable Status

Version	Date	Pages	Author	Modification
0.1	29/09/14	All	GP	Created

Table 2, Deliverable Change History



5 Glossary

SSP	SHIWA Simulation Platform
WP	Work package

Table 3, Glossary

5.1 Workflows and Metaworkflows

5.1.1 DTI Pre-processing CVMFS

5.1.1.1 Nature and Relevance

Diffusion Tensor Imaging (DTI) has become a popular method to analyse the structure of the human brain, which is composed of grey and white matter tissue. The grey matter is known to facilitate signal processing, and the white matter comprises nerve bundles that connect brain regions to each other. These nerve bundles are macroscopic structures composed of microscopic fibers. DTI measures the local diffusion properties of water in the brain, and can detect these fibers due to their highly anisotropic diffusion properties. Water diffusion is high along the fibre orientation and low (restricted) in the perpendicular direction. This allows for reconstructing both the orientation and integrity of white matter, making it possible to study brain diseases affecting the white matter *in vivo*.

DTI data acquisition is performed in 12 to 60 three-dimensional (3D) orientations, resulting in a series of 3D volumes, in addition to an anatomical scan of the brain. These raw data cannot be directly interpreted, but need complex and computationally demanding analysis. Different toolboxes are available today for analysis of the DTI scans (after being pre-processed), for example FSL BedpostX (see section 10.4.3) and DTI Population Registration (see section 10.4.4).

The DTI pre-processing application implements a sequence of steps that prepare raw DTI data to be used in further analysis packages. The acquisition of a DTI dataset commonly takes 10 minutes, so patients might have moved during this time and motion correction is needed. In addition the data may contain artefacts and noise, which also have to be corrected or compensated for. Moreover, these packages normally require input data to be stored into a special file format and directory structure, as well as submitted to some additional pre-processing steps.

5.1.1.2 Workflow Details

This WS-PGRADE workflow has one component that runs the third-party application for DTI pre-processing. All output files are bundled together into one single tar file. This workflow has been redesigned in the second project year to take advantage of the CVMFS software repository. Once a new version of the third party application is available on the CVMFS repository, the same workflow can be easily instructed to take the desired version.



Figure 1, DTI Pre-processing Workflow



5.1.1.3 Software Details

Name	Description	Reference
DTI Pre-processing software	Pre-processing of DTI scan images made available via CVMFS	

5.1.1.4 Input/Output

Input:

- DTI scan image: It can be either in DICOM format or NIFTI format provided that extra metadata is provided next to it.

Output:

- A single tar file containing all the parameters calculated by the software.

5.1.1.5 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=5709
Application description template	
User documentation	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=5709
Contact details	m.jaghouri@amc.uva.nl

Table 4, DTI Pre-processing technical details

5.1.2 FSL BedpostX CVMFS

5.1.2.1 Nature and Relevance

DTI data provides information about connectivity of brain white matter (see introduction about DTI in Section **Error! Reference source not found.**). The FMRIB's Diffusion Toolbox of the FMRIB Software Library (FSL) developed at Oxford University is one of the toolboxes that are available today for analysis of the DTI scans. In particular, the FSL BedpostX tool is able to detect crossing fibers, which is crucial for correct reconstruction of the white matter structure. For example, this is important for neurosurgery planning.

The method uses a Markov Chain Monte Carlo-based model and is computationally demanding. The processing time of a typical patient's dataset is about 20 hours on a standard desktop computer. When studies are carried out, typically encompassing in the order of 100 subjects, the compute-time sums up to several weeks.

5.1.2.2 Workflow Details

This is a workflow based on split-and-merge pattern realized using generator and collector nodes in WS-PGRADE. It uses the FSL toolkit in all three components. The latest FSL toolkit is taken on demand from the CVMFS repository.

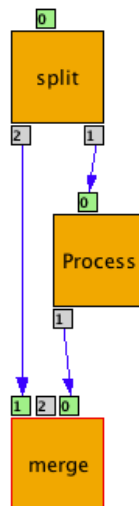


Figure 2 FSL BedpostX Workflow

5.1.2.3 Software Details

Name	Description	Reference
FSL toolkit	FMRIB's Diffusion Toolbox - FDT v2.0	http://fsl.fmrib.ox.ac.uk/fsl/fsl4.0/fdt/fdt_bedpostx.html

5.1.2.4 Input/Output

Input:

- The preprocessed DTI scan as calculated by the DTI preprocessing workflow.

**Output:**

- A single tar file containing all files generated by bedpostx.

5.1.2.5 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=5665
Application description template	
User documentation	http://fsl.fmrib.ox.ac.uk/fsl/fsl4.0/fdt/fdt_bedpostx.html
Contact details	m.jaghouri@amc.uva.nl

Table 5, FSL BedpostX technical details



5.1.3 DTI Population Registration (with and without CVMFS)

5.1.3.1 Nature and Relevance

In the study of the brain, it is necessary to use large collections of scans to obtain measurements that are representative for a particular population. A population contains a group of patients (e.g., displaying depression, schizophrenia, Amyotrophic Lateral Sclerosis, etc.) and control subjects. The scans in this case belong to different people; therefore, they need to be aligned to each other before the measurements (markers) can be compared.

The DTI Population Registration application determines spatial correspondence in DTI datasets of all subjects in a certain population. Typically 50 subjects are included, but larger populations are expected with the growth of study sizes. In this application the “mean” of all subjects serves as template, towards which all subjects are aligned (registered).

5.1.3.2 Workflow Details

In this workflow, the mean of all subjects serves as template, towards all subjects are registered. This mean template is iteratively updated until convergence (6 times is optimal). A Moteur implementation has been reported in the past period. In this year, the workflow has been ported to WS-PGRADE. Since WS-PGRADE does not directly support iterations, the main component has been repeated six times to obtain the optimal convergence. The main goal of this implementation is to enable this application from the new gateway at the AMC which is based on WS-PGRADE. We have made two implementations of this workflow, such that in one of them the software is shipped together with the workflow. In the other implementation, the software is downloaded automatically and on demand from the CVMFS repository. Both of these implementations are provided under the same abstract workflow in the SHIVA repository.

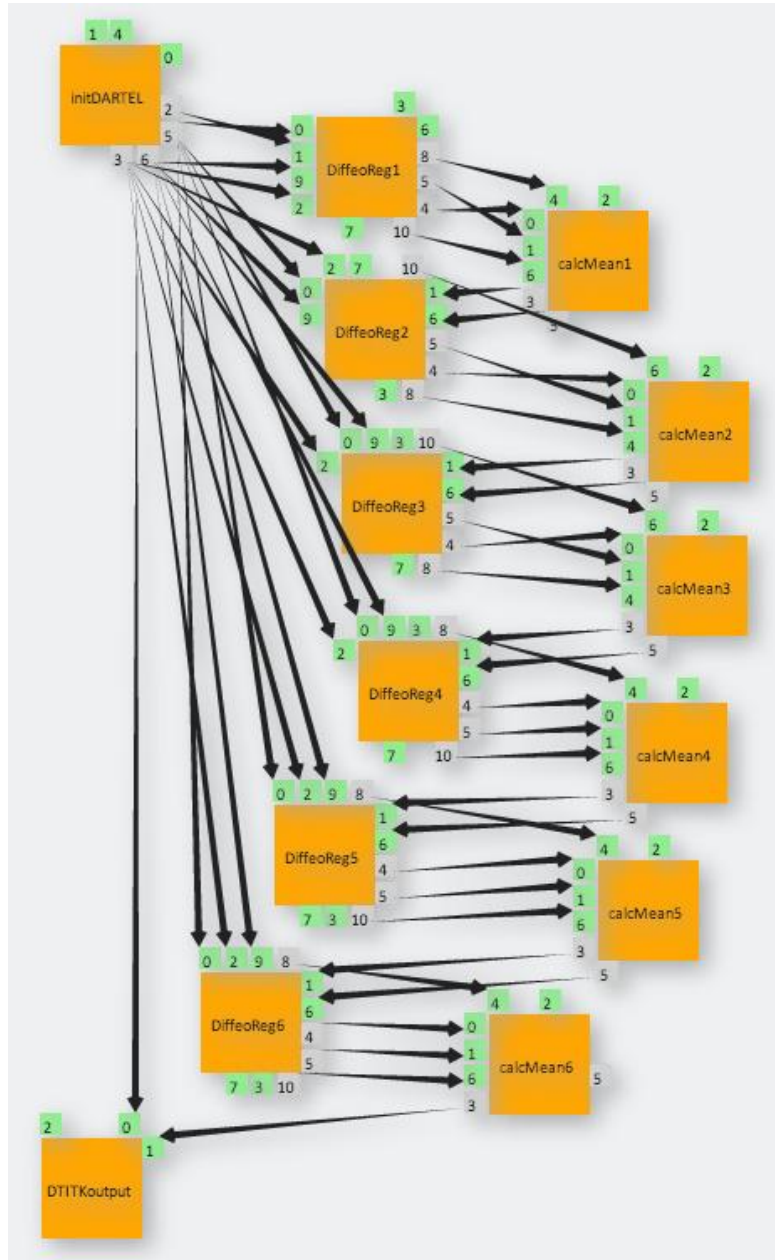


Figure 3, DTI Population Registration Workflow

5.1.3.3 Software Details

Name	Description	Reference

5.1.3.4 Input/Output

Input:

- The pre-processed DTI scan as calculated by the DTI pre-processing workflow.



Output:

-

5.1.3.5 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=5145
Application description template	
User documentation	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=5145
Contact details	v.korkhov@amc.uva.nl

Table 6, DTI Population Registration technical details

5.1.4 Freesurfer Pial (with and without CVMFS)

5.1.4.1 Nature and Relevance

Freesurfer is an automated tool to segment regions in the brain in MRI scans. It reconstructs the brain’s cortical surface from structural MRI data, and it can also overlay functional MRI data onto the reconstructed surface.

It is intensively used in neuroscience research to isolate regions of interest in which measurements are performed to characterize brain diseases. For example, the volume of the hippocampus is used as a marker of brain degeneration, or activation in the amygdala is used to differentiate between groups of persons with some kind of addiction and control subjects, i.e., healthy persons.

In some executions, the software is not able to detect all segments of the brain with a good precision. This requires human intervention after one execution round, and afterwards the pial option of the software can be used to improve the output quality.

5.1.4.2 Workflow Details

This WS-PGRADE workflow has one component that runs the Freesurfer with pial option. We have made two implementations of this workflow, such that in one of them the software is shipped together with the workflow. In the other implementation, the software is downloaded automatically and on demand from the CVMFS repository.

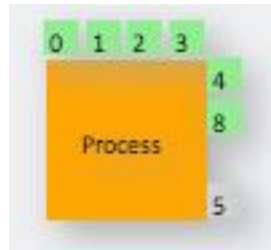


Figure 4, Freesurfer pial Workflow

5.1.4.3 Sub-Workflows / Software Details

Name	Description	Reference
Freesurfer	Brain Segmentation Tools	http://surfer.nmr.mgh.harvard.edu/fswiki

5.1.4.4 Input/Output

Input:

- One tgz file that was created by a previous run of Freesurfer.

Output:

- One single tar file containing all the outputs generated by the third party application.



5.1.4.5 Further Technical Details

Information	URL
SHIWA Repository (w/o CVMFS)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5954
SHIWA Repository (with CVMFS)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5955
Application description template	
User documentation	http://surfer.nmr.mgh.harvard.edu/fswiki
Contact details	m.jaghouri@amc.uva.nl

Table 7, Freesurfer PIAL technical details

5.1.5 Freesurfer Param (with and without CVMFS)

5.1.5.1 Nature and Relevance

Freesurfer is an automated tool to segment regions in the brain in MRI scans. It reconstructs the brain’s cortical surface from structural MRI data, and it can also overlay functional MRI data onto the reconstructed surface.

It is intensively used in neuroscience research to isolate regions of interest in which measurements are performed to characterize brain diseases. For example, the volume of the hippocampus is used as a marker of brain degeneration, or activation in the amygdala is used to differentiate between groups of persons with some kind of addiction and control subjects, i.e., healthy persons.

The Freesurfer toolkit offers a wide range of different options. Although a complete segmentation is the most used option, since it takes a very long time, researchers could opt to run the software with handpicked options in order to achieve much quicker results.

5.1.5.2 Workflow Details

This WS-PGRADE workflow has one component that runs the Freesurfer with user defined options. We have made two implementations of this workflow, such that in one of them the software is shipped together with the workflow. In the other implementation, the software is downloaded automatically and on demand from the CVMFS repository.



Figure 5, Freesurfer param Workflow

5.1.5.3 Sub-Workflows / Software Details

Name	Description	Reference
Freesurfer	Brain Segmentation Tools	http://surfer.nmr.mgh.harvard.edu/fswiki

5.1.5.4 Input/Output

Input:

- One DICOM or NIFTI file, or a tgz file already processed partially with freesurfer.

Output:

- One single tar file containing all the outputs generated by the third party application.



5.1.5.5 Further Technical Details

Information	URL
SHIWA Repository (w/o CVMFS)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5952
SHIWA Repository (with CVMFS)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5953
Application description template	
User documentation	http://surfer.nmr.mgh.harvard.edu/fswiki
Contact details	m.jaghouri@amc.uva.nl

Table 8, Freesurfer Param technical details

5.1.6 Freesurfer recon-all CVMFS

5.1.6.1 Nature and Relevance

Freesurfer is an automated tool to segment regions in the brain in MRI scans. It reconstructs the brain’s cortical surface from structural MRI data, and it can also overlay functional MRI data onto the reconstructed surface.

It is intensively used in neuroscience research to isolate regions of interest in which measurements are performed to characterize brain diseases. For example, the volume of the hippocampus is used as a marker of brain degeneration, or activation in the amygdala is used to differentiate between groups of persons with some kind of addiction and control subjects, i.e., healthy persons.

Freesurfer requires long computing times (24-36h) and is often executed for many (healthy and diseased) subjects in each study; therefore, its adoption for large neuroscience studies has become impractical. The execution of this application on the Dutch grid using MOTEUR has enabled a large number of new studies already. The goal in ER-Flow was to revise, optimize and publish the MOTEUR implementation and port it to WS-PGRADE. Moreover, integration of the science gateways with the data servers used by the neuroscientists is expected to further streamline the usage of this application.

5.1.6.2 Workflow Details

This WS-PGRADE workflow has one component that runs the Freesurfer with predefined recon-all option. This workflow has been redesigned in the second project year to take advantage of the CVMFS software repository. Once a new version of the third party application is available on the CVMFS repository, the same workflow can be easily instructed to take the desired version. The software is downloaded automatically and on demand from the CVMFS repository.

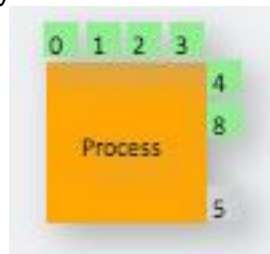


Figure 6, Freesurfer recon-all Workflow

5.1.6.3 Sub-Workflows / Software Details

Name	Description	Reference
Freesurfer	Brain Segmentation Tools	http://surfer.nmr.mgh.harvard.edu/fswiki

5.1.6.4 Input/Output

Input:

- One DICOM or NIFTI file.

Output:

- One single tar file containing all the outputs generated by the third party application.



5.1.6.5 Further Technical Details

Information	URL
SHIWA Repository	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5951
Application description template	
User documentation	http://surfer.nmr.mgh.harvard.edu/fswiki
Contact details	m.jaghouri@amc.uva.nl

Table 9, Freesurfer technical details

5.1.7 Trac-all CVMFS

5.1.7.1 Nature and Relevance

TRACULA is a tool for automatic reconstruction of a set of major white-matter pathways from diffusion-weighted MR images. It uses global probabilistic tractography with anatomical priors. Prior distributions on the neighboring anatomical structures of each pathway are derived from an atlas and combined with the FreeSurfer cortical parcellation and subcortical segmentation of the subject that is being analyzed to constrain the tractography solutions. This obviates the need for user interaction, e.g., to draw ROIs manually or to set thresholds on path angle and length, and thus automates the application of tractography to large datasets. (see <http://freesurfer.net/fswiki/Tracula> for more information).

5.1.7.2 Workflow Details

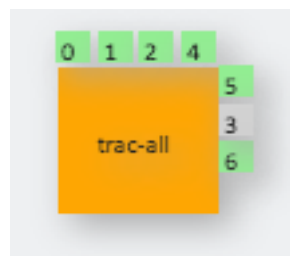


Figure 7, Trac-all Workflow

5.1.7.3 Software Details

Name	Description	Reference
Trac-all	TRActs Constrained by UnderLying Anatomy	http://freesurfer.net/fswiki/Tracula

5.1.7.4 Input/Output

Input:

- Output generated by DTI preprocessing
- Output generated by FSL BedpostX
- Output generated by DTI Freesurfer with recon-all option

Output:

- One tgz file containing all the outputs generated by the third party application

5.1.7.5 Further Technical Details

Information	URL
SHIWA Repository	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5711
Application description template	
User documentation	http://freesurfer.net/fswiki/Tracula
Contact details	m.jaghouri@amc.uva.nl

Table 10, TRAC-ALL technical details

5.1.8 CQRS

5.1.8.1 Nature and Relevance

CQRS is an algorithm to detect QRS-complexes in overnight ECG signals recorded within a polysomnography. It is a prerequisite for further analysis of the heart frequency variability. This application is developed as one of the important algorithms used in the SOMNO.NETZ project. Three workflows are created demonstrating three strategies in porting such an algorithm.

5.1.8.2 Workflow Details

Figure below shows the three strategies adopted in implementing this application as a WS-PGRADE workflow. The leftmost workflow, called CQRS-ONE, runs the application as one component and as input takes the identifier of the input data. This identifier is used to download the data from XNAT server; the output is also written directly to XNAT, so no output port is incorporated. This is a direct adoption of the traditional approach used in SOMNO.NETZ. The second workflow, called CQRS-DL, also transfers the input and output data directly to/from XNAT. However, it separates the two steps of the algorithm into two components. This allows for a more fine-grained control over execution of the workflow.

Finally, the rightmost workflow, called simply CQRS, abstracts from XNAT and only focuses on the implementation of the algorithm as a workflow. The input and output to the workflow are modelled explicitly as ports. Therefore, the communication with XNAT should be performed externally to the workflow execution. This approach allows for better interoperability and reusability.

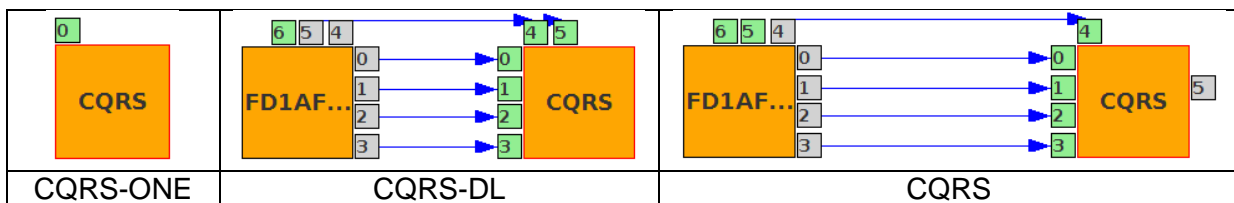


Figure 8, Three Workflow Implementations of the CQRS algorithms

5.1.8.3 Software Details

Name	Description	Reference
CQRS	QRS-complexes detection algorithm in Matlab	

You need the matlab_mcr 2013b installed. Download zip-File, unzip, run in Terminal:
`$PATH2MATLABMCR/install`

Program execution: First run FD1_AF2_DF2, then CQRS Run in Terminal
`$PATH2APPLICATION/run_FD1_AF2_DF2.sh`
`$PATH2MCR`
`$PATHTODATA/b000301_ecg`

If processed correctly, the following files are created

- b000301_ecg_fd1.dat
- b000301_ecg_af2.dat
- b000301_ecg_df2.dat



Run in the same directory

```
$PATH2APPLICATION/run_CQRS.sh  
$PATH2MCR  
$PATH2ODATA/b000301_ecg
```

If processed correctly, a file called b000301_ecg_cqrs.dat is created

5.1.8.4 Input/Output

Input:

- mit-Format (2 files: one header-file filebase.heg, one datafile filebase.dat)

Output:

- Text files (filename_ \$method.dat) for three methods, fd1, af2, and df2.

5.1.8.5 Further Technical Details

Information	URL
SHIWA Repository (CQRS)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5254
SHIWA Repository (CQRS-ONE)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5255
SHIWA Repository (CQRS-DL)	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5253
Application description template	
User documentation	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5253
Contact details	dagmar.krefting@htw-berlin.de

Table 11, CQRS technical details

5.1.9 AutoDock Vina with Sorting All Energies

5.1.9.1 Nature and Relevance

AutoDock Vina is one of the most advanced tools in virtual screening that is used in drug discovery research. Applying this application on upto millions of drug candidates requires huge computation and takes a few years of compute time. At AMC, we revised an existing workflow from the Shiwa repository in order to meet the requirements of our local users in terms of the outputs that should be returned from the workflow. Specifically, the workflow sorts the input ligands based on their binding energies. The workflow is additionally adapted for usage behind the recently launched Docking Gateway at the AMC.

5.1.9.2 Workflow Details

This is a workflow based on split-and-merge pattern realized using generator and collector nodes in WS-PGRADE. Additionally, to avoid collecting partial results, the generator passes the number of expected results to the collector.



Figure 9, AutoDock Vina Workflow

5.1.9.3 Software Details

Name	Description	Reference
AutoDock Vina	An open-source program for molecular docking	http://vina.scripps.edu/

For additional functionality, like sorting the binding energies, the workflow uses the available utilities in a standard Linux bash environment.

5.1.9.4 Input/Output

Input:

- Receptor: A file of type pdbqt describing the drug target



- Ligands to be docked: A list of drug candidates in pdbqt format combined into one zip file
- Configuration file: containing information like an area of interest for docking, exhaustiveness, energy range, and any other configuration that can be accepted by the AutoDock Vina software.

Output:

- Raw docking output: A tar-gzipped file containing all the raw outputs of the AutoDock Vina software
- Energy list: A comma separated and sorted list of energy values showing the affinity of each ligand

5.1.9.5 Further Technical Details

Information	URL
SHIWA Repository	http://repo.shiwa-workflow.eu/shiwa-repo/public/edit-application.xhtml?appid=5658
Application description template	
User documentation	http://vina.scripps.edu/tutorial.html
Contact details	m.jaghouri@amc.uva.nl

Table 12, AutoDock Vina technical details