CancerGrid

gUSE in CancerGrid project

József Kovács
smith@sztaki.hu
MTA SZTAKI

Joint EGEE-EDGeS Summer School, 03/07/2009,
Budapest, Hungary

The CancerGrid infrastructure

CancerGrid

- The general framework components
  - WS-PGRADE/gUSE portal
  - SZTAKI Private/Local Desktop Grid
  - 3GBridge
- Customisation
  - algorithms
  - workflows
  - molecule database
  - resources
- Scenario
  - Handling molecules in the database
  - Workflow configuration/submission & monitoring

The CancerGrid project

CancerGrid

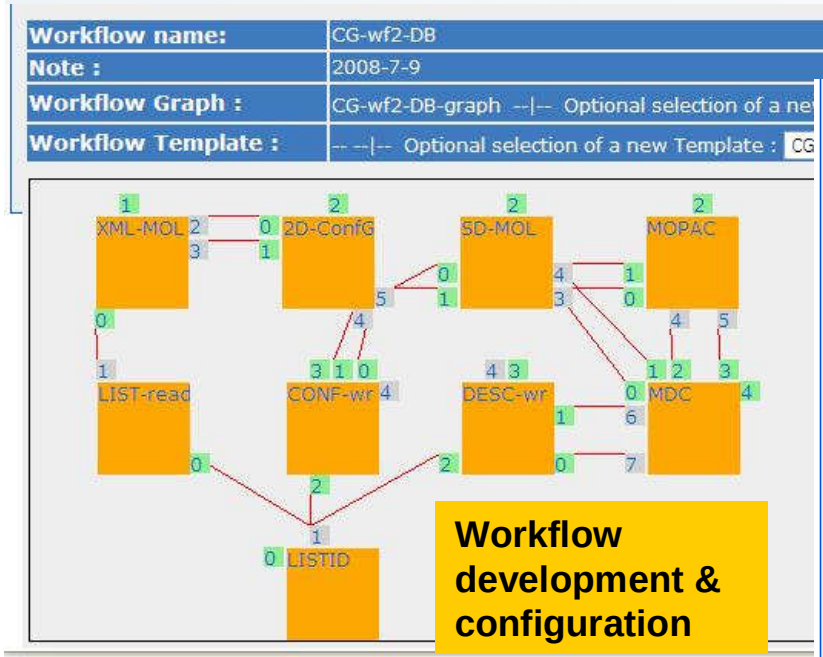
- EU Framework Program 6
- Title: Grid Aided Computer System For Rapid Anti-Cancer Drug Design
- Project period
  - January 1, 2007 – December 31, 2009 (June 30, 2010)
- Goals:
  - Developing *focused libraries* with a high content of anti-cancer leads, building *models* for predicting various molecule properties
  - Developing a *computer system* based on grid technology, which helps to accelerate and automate the *in silico design* of libraries for drug discovery processes

The CancerGrid consortium

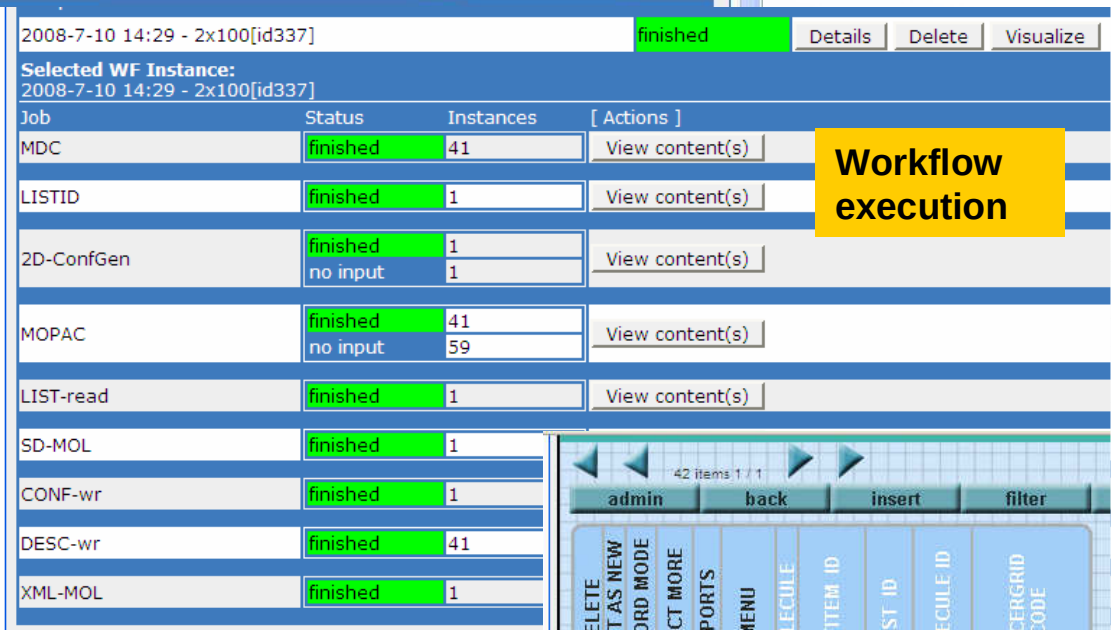
- TargetEx, Inc., Hungary (coordinator) [chemistry]
- AMRI Hungary, Inc., Hungary [chemistry]
- Inte:Ligand, Austria [chemoinformatics]
- Tallinn University of Technology, Estonia [chemoinformatics]
- University of Helsinki, Finland [biotechnology]
- GKI Economic Research, Hungary [economics]
- SZTAKI, Hungary [computer science]
- University of Jerusalem, Israel [chemoinformatics]
- DAC, Italy [biotechnology]
- University of Bari, Italy [chemistry]
- University of Pompeu Fabra, Spain [chemoinformatics]

The CancerGrid portal (gUSE & SZTAKI DG)

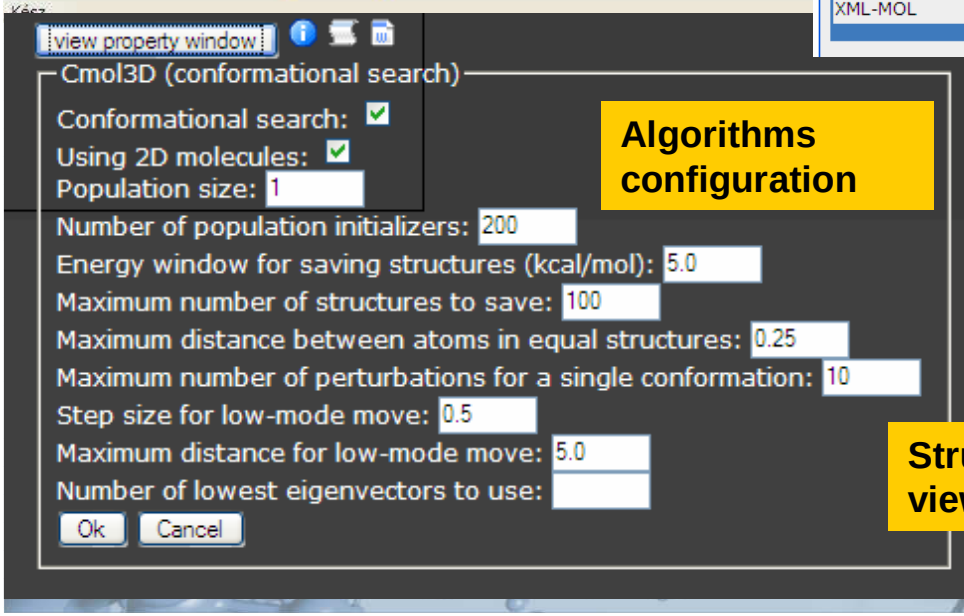
CancerGrid



Workflow development & configuration

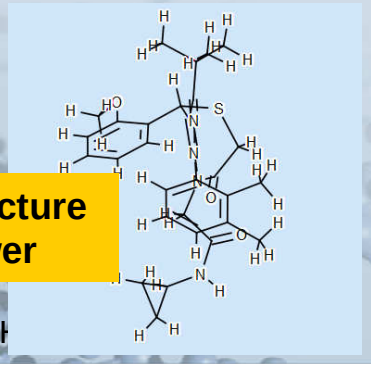


Workflow execution

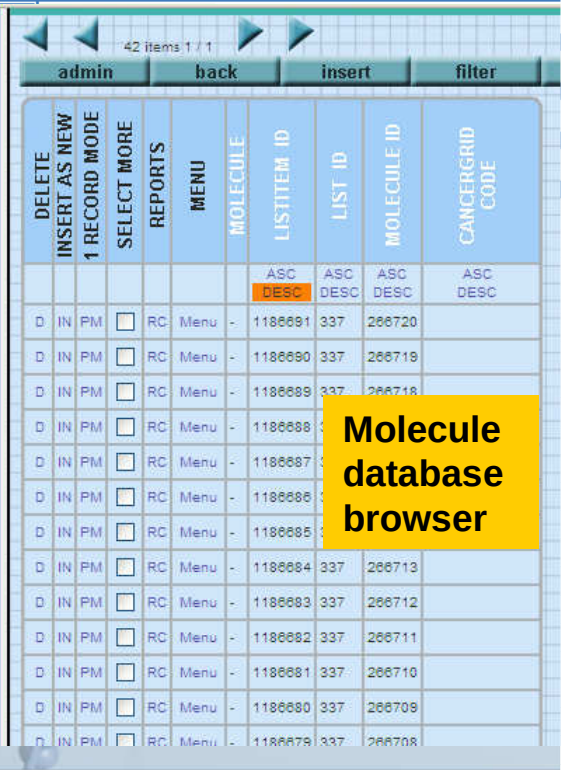


Algorithms configuration

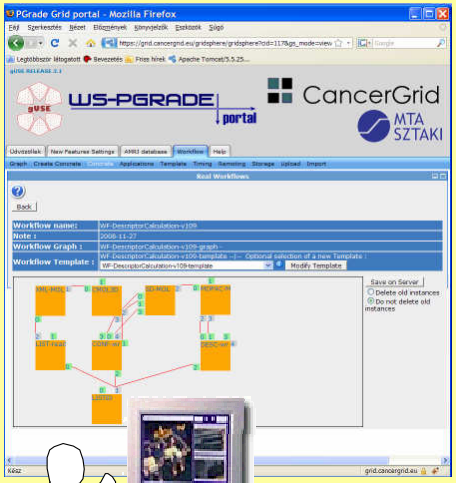
Integrated components of CancerGrid portal



Structure viewer



Molecule database browser



executing workflows

Portal

DG jobs

Local jobs

3G Bridge

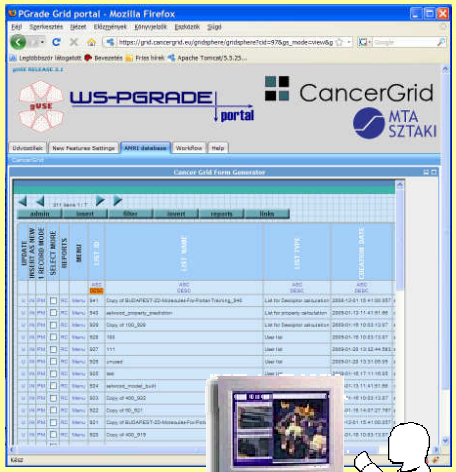
Job 1
Job 2
Job N

Local Resource

BOINC server

WU 1
WU 2
WU N

Portal Storage



browsing molecules

BOINC client

GenWrapper for batch execution

WU X

WU Y

Legacy Application

DG clients from all partners

Portal and DesktopGrid server

molecule database

Molecule database server

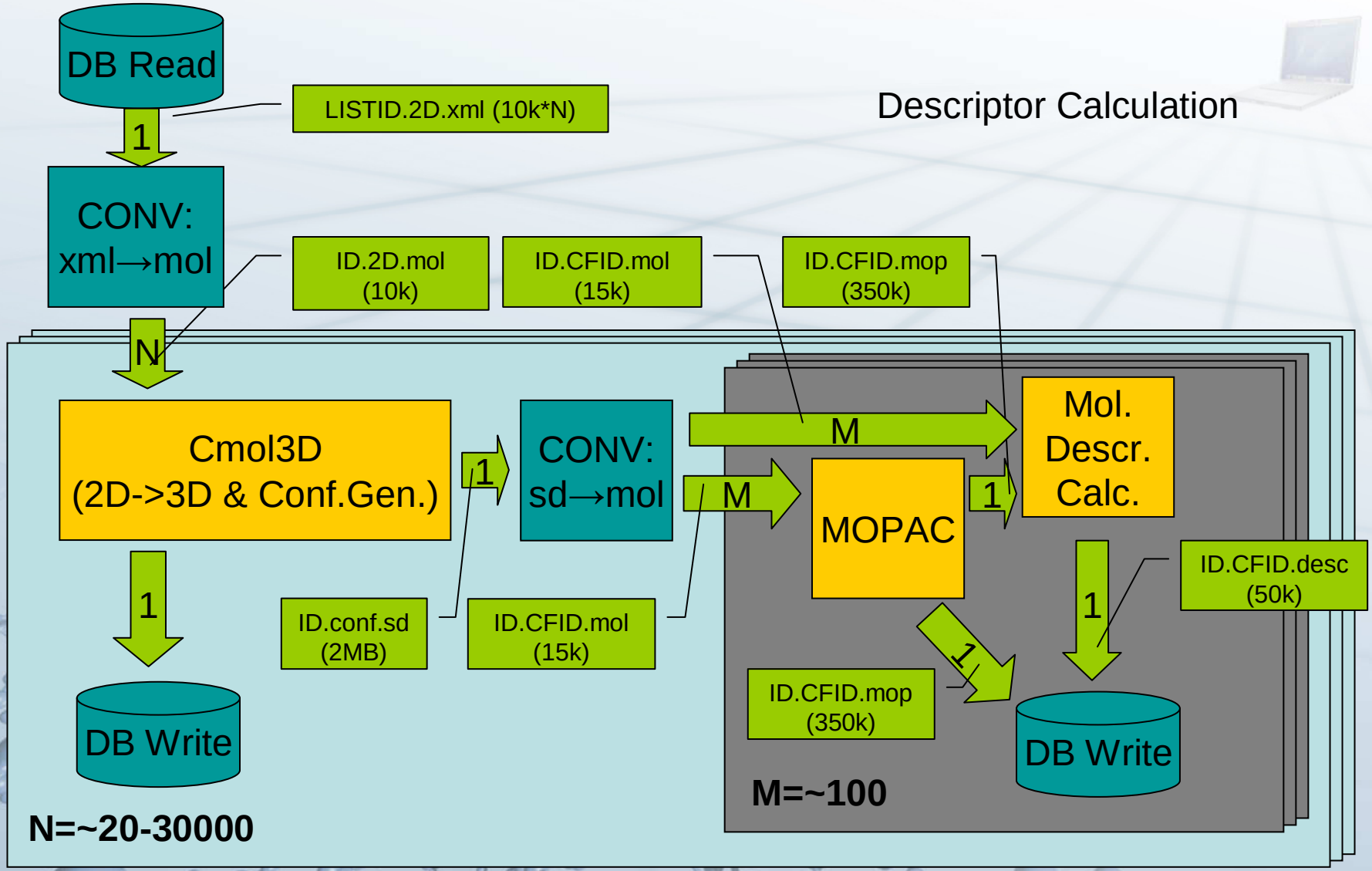
Applications in CancerGrid

CancerGrid

- Flexmol is an XML-based molecular language
- Molecule 2D/3D converter (Cmol3D)
- Molecule 3D conformation generator (Cmol3D)
- MOPAC (Molecular Orbital PACkage) is a semiempirical quantum chemistry program based on Dewar and Thiel's NDDO approximation
- **Codessa Pro** (**Co**mprehensive **De**scriptors for **S**tructural and **S**tatistical **A**nalysis) is a software suite for developing quantitative structure-activity/property relationships
- Matrix former
- QSAR Model builder **Quantitative structure-activity relationship (QSAR)** is the process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity.
- (Chemical) Property Predictor
- File format converters (to integrate the previous tools into a workflow)

Supported workflow 1/4

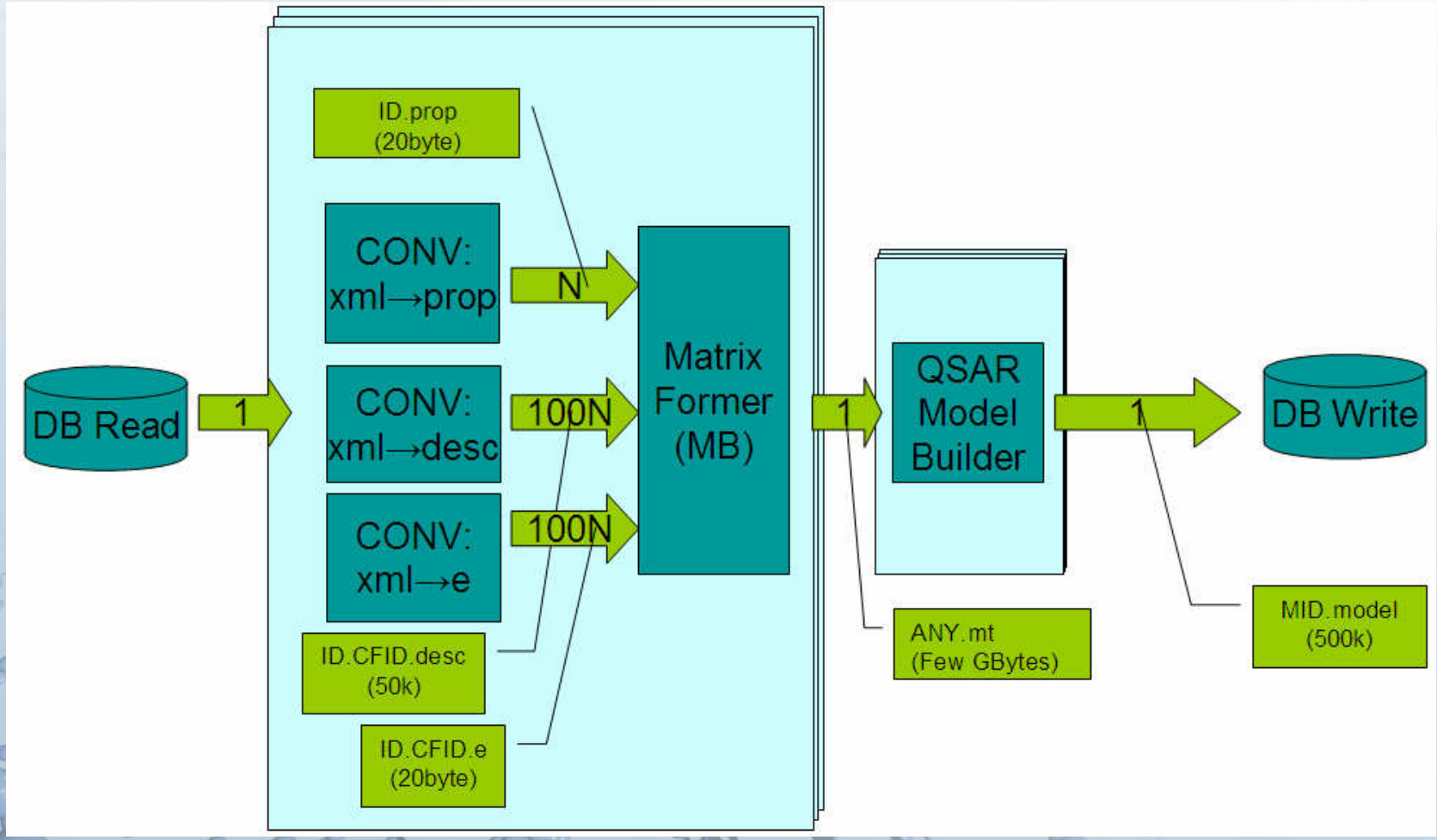
CancerGrid



Supported workflow 2/4

CancerGrid

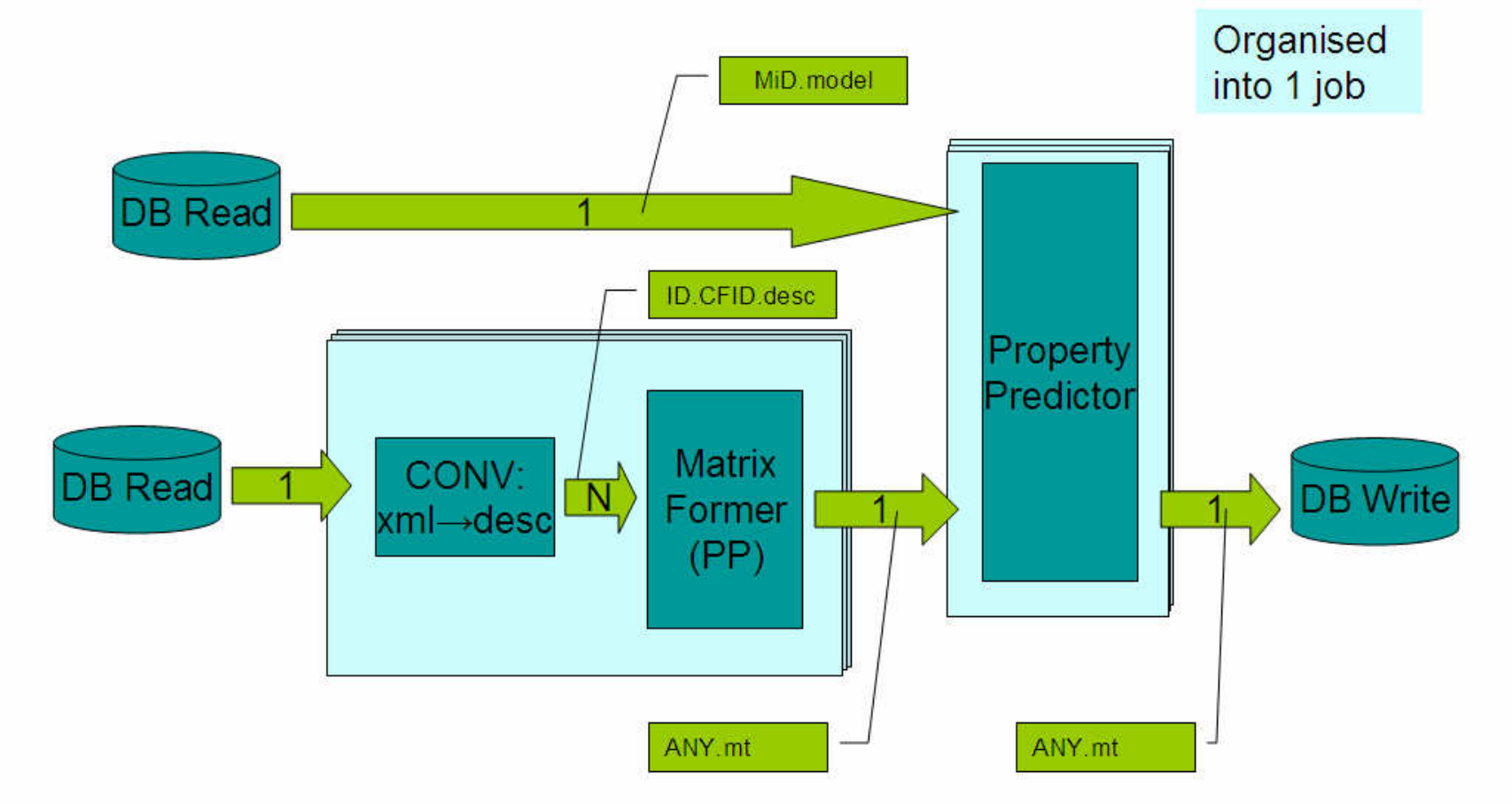
Model building



Supported workflow 3/4

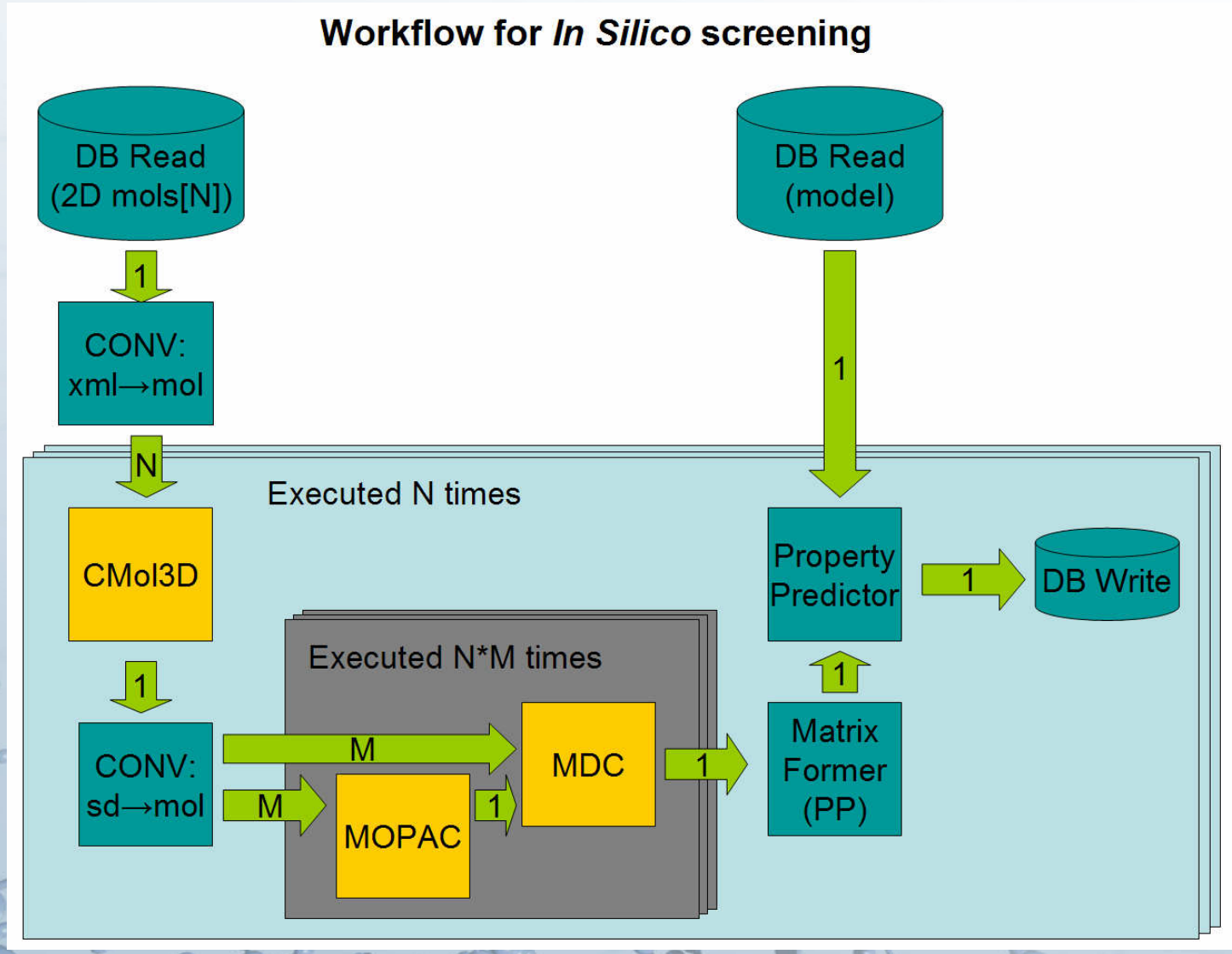
CancerGrid

Property Prediction



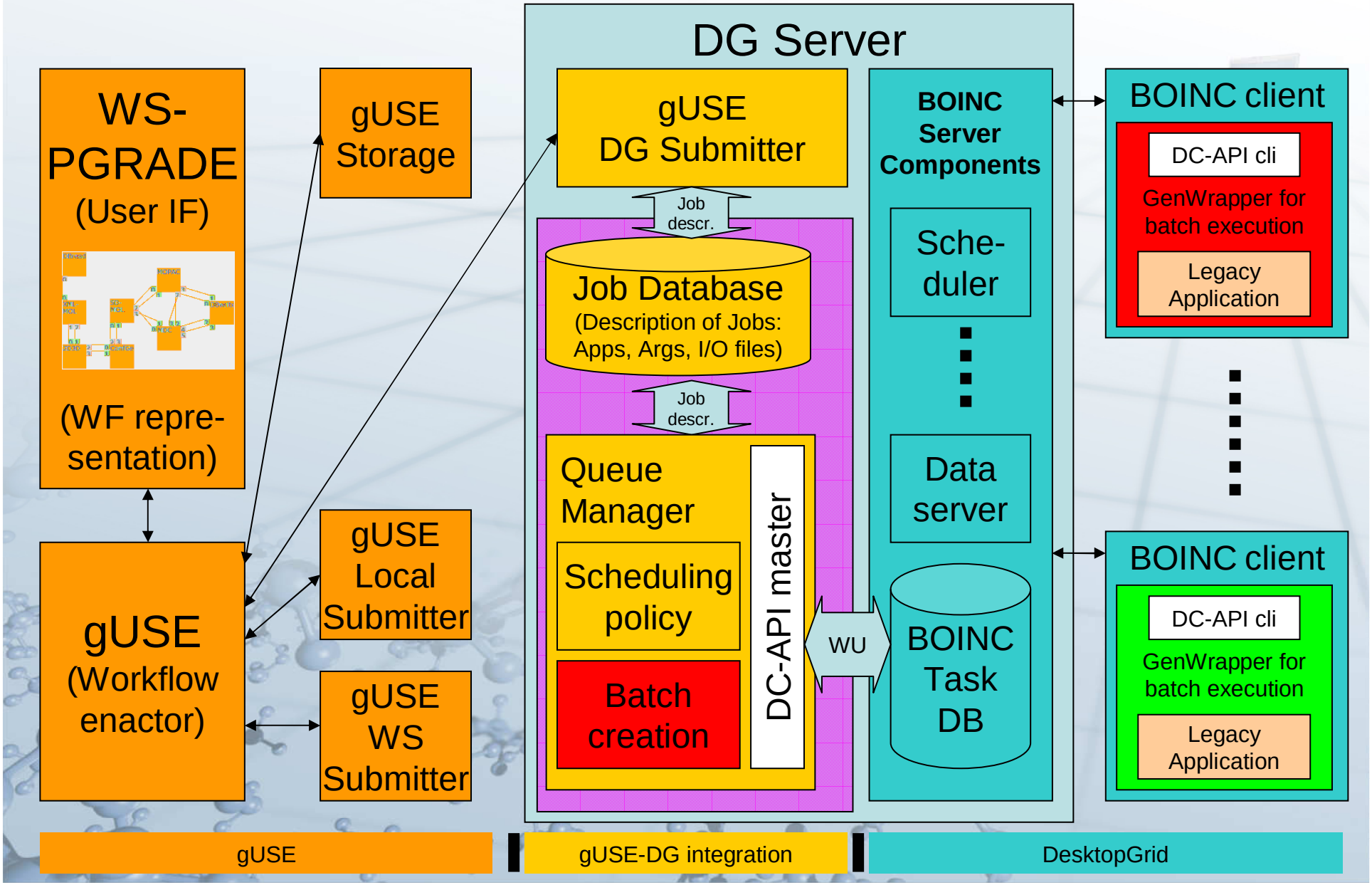
Supported workflow 4/4

CancerGrid



gUSE & SZTAKI integration through 3GBridge

CancerGrid



Job handling

CancerGrid

Applications: **cmol3d**, **mopac**, **mdc**, **fmt**, **fma**, etc.

Various requirements

- Fortran, C, C++
- processing/ memory requirements
- multi-binary applications (already contain some wrappers), libraries
- legacy binaries for Linux and Windows
- config file preparation before execution
- pure logging/ debugging information
- variable number of output files

GenWrapper tool
on the client machine

Low granularity

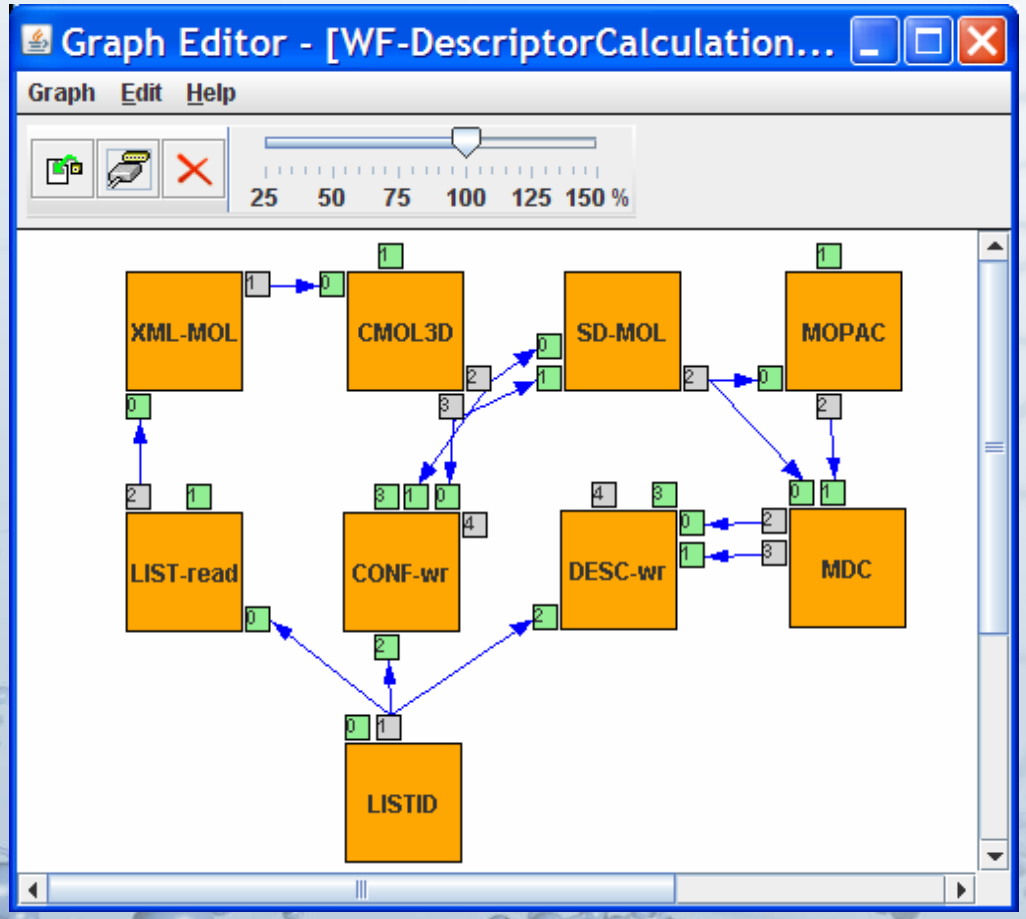
- Execution time is very small (simetimes only a few minutes)

Batch creation plugin in 3GBridge

- The workunit contains the shell script that is created to manage the execution of the batch
  - assembled from head, body and tail fragments
  - body part is repeated for each job in the batch
  - may contain macros like **`%{name}`**

Descriptor calculation workflow in gUSE

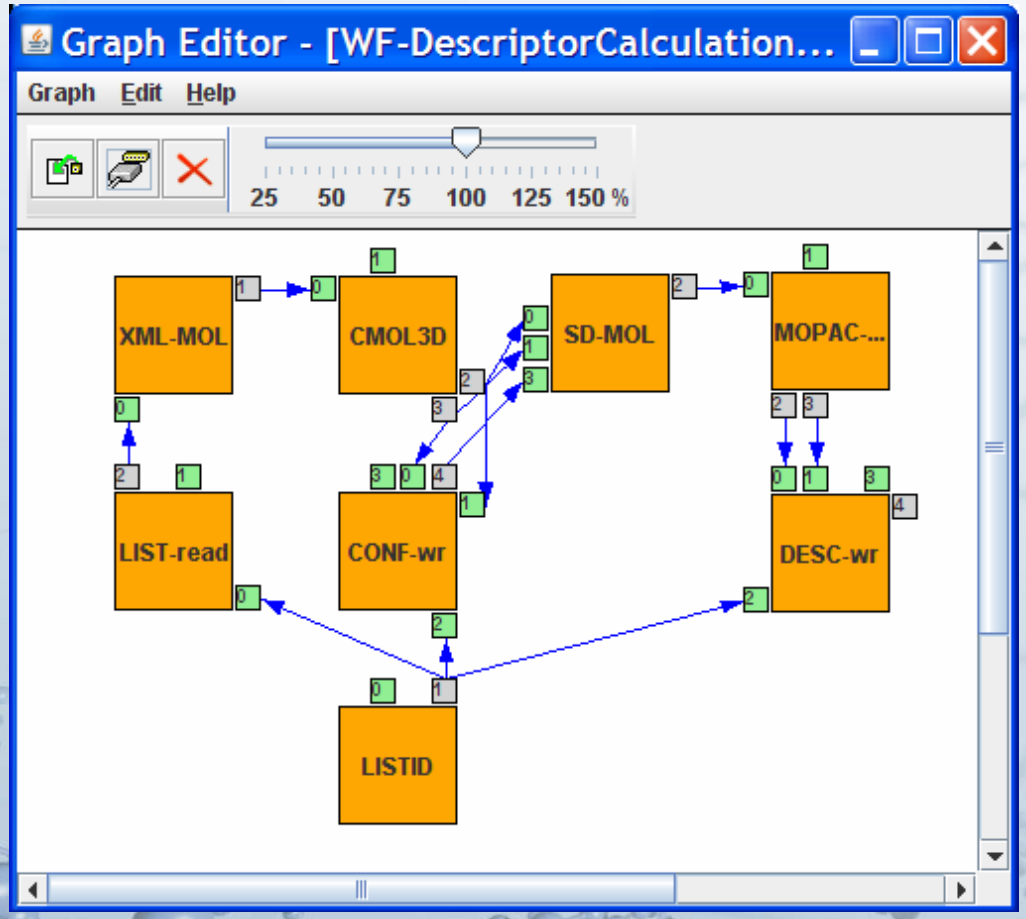
CancerGrid



- DG jobs:
  - CMOL3D
  - MOPAC
  - MDC
- WS jobs:
  - LIST-read
  - CONF-wr
  - DESC-wr
- Local jobs:
  - LISTID
  - XML-MOL
  - SD-MOL

Optimised descriptor calculation workflow in gUSE

CancerGrid



- DG jobs:
  - CMOL3D
  - MOPAC
  - MDC

  built into one job
- WS jobs:
  - LIST-read
  - CONF-wr
  - DESC-wr
- Local jobs:
  - LISTID
  - XML-MOL
  - SD-MOL

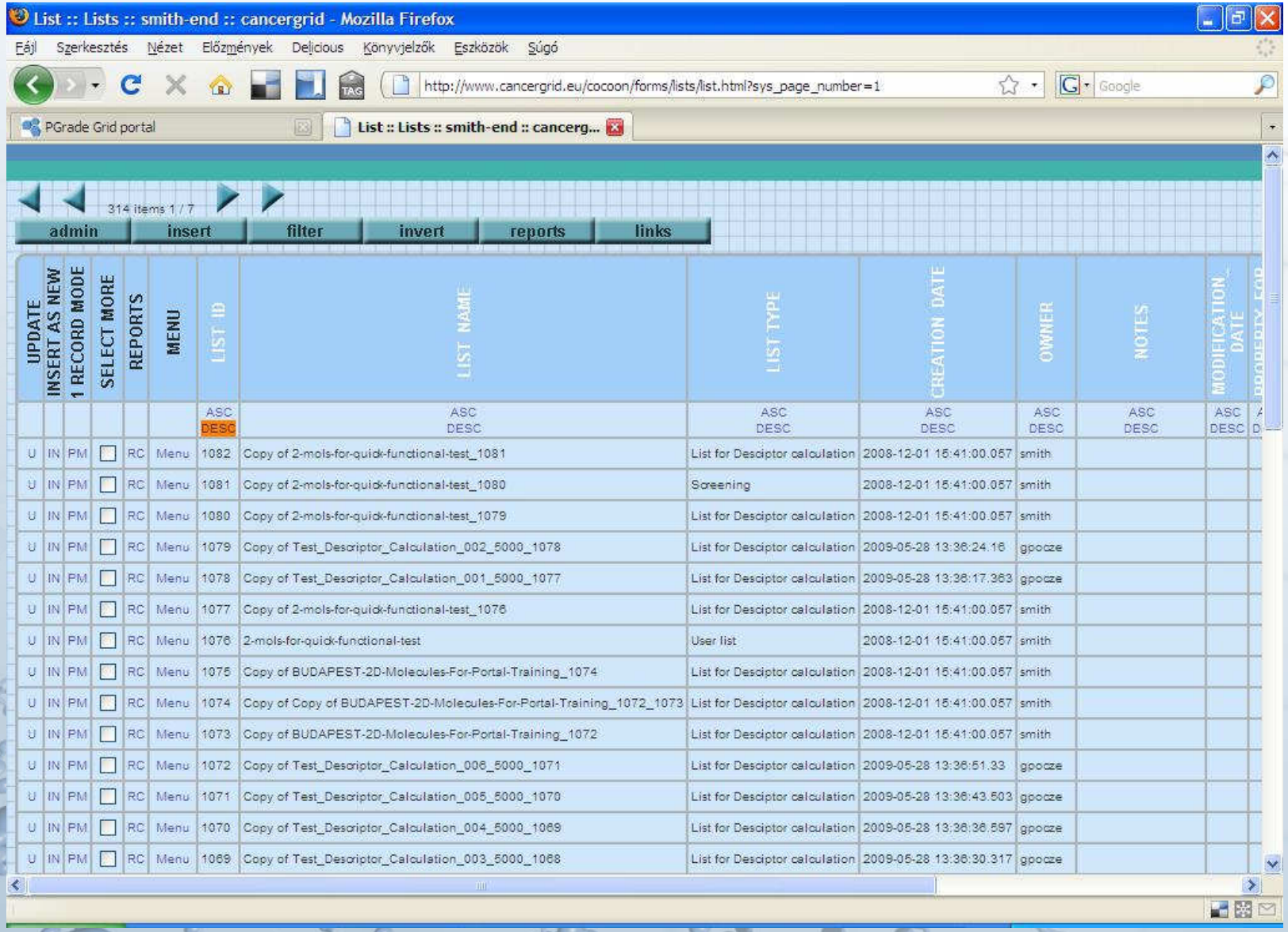
High-level scenario

CancerGrid

- Initial state: molecules/structures stored in DB, organised into lists
- User selects list of molecules/structures
- User selects/downloads a workflow from repository
- User configures the workflow to take the list as input
- User optionally updates parameters of the modules
- Submits workflow
- Optionally monitors the status
- When workflow finished, results are stored in the DB

Molecule lists

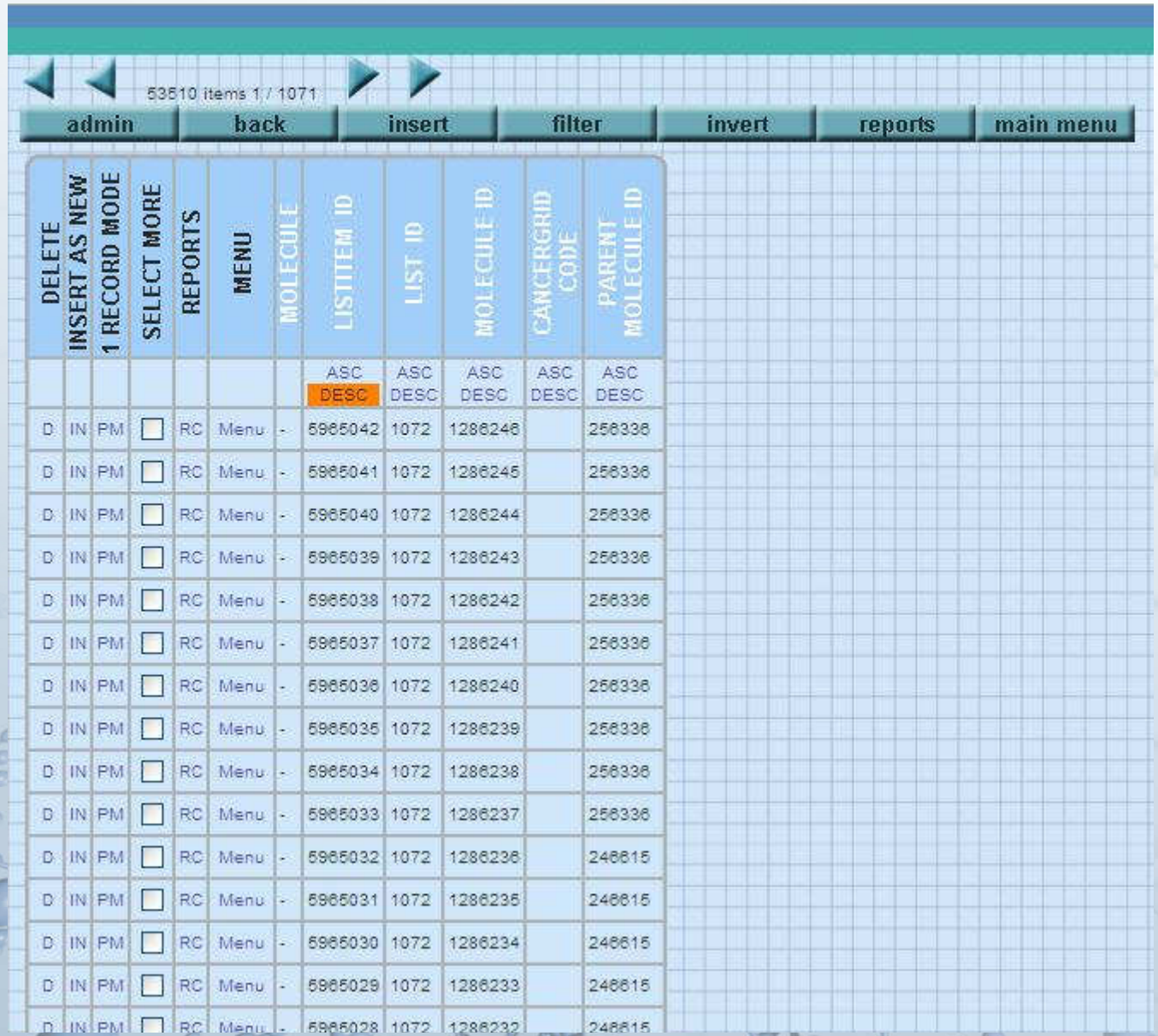
CancerGrid



UPDATE	INSERT AS NEW 1 RECORD MODE		SELECT MORE	REPORTS	MENU	LIST ID	LIST NAME	LIST TYPE	CREATION DATE	OWNER	NOTES	MODIFICATION DATE
						ASC DESC	ASC DESC	ASC DESC	ASC DESC	ASC DESC	ASC DESC	ASC DESC
U	IN	PM	☐	RC	Menu	1082	Copy of 2-mols-for-quick-functional-test_1081	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1081	Copy of 2-mols-for-quick-functional-test_1080	Screening	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1080	Copy of 2-mols-for-quick-functional-test_1079	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1079	Copy of Test_Descriptor_Calculation_002_5000_1078	List for Desciptor calculation	2009-05-28 13:36:24.16	gpocze		
U	IN	PM	☐	RC	Menu	1078	Copy of Test_Descriptor_Calculation_001_5000_1077	List for Desciptor calculation	2009-05-28 13:36:17.363	gpocze		
U	IN	PM	☐	RC	Menu	1077	Copy of 2-mols-for-quick-functional-test_1076	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1076	2-mols-for-quick-functional-test	User list	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1075	Copy of BUDAPEST-2D-Molecules-For-Portal-Training_1074	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1074	Copy of Copy of BUDAPEST-2D-Molecules-For-Portal-Training_1072_1073	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1073	Copy of BUDAPEST-2D-Molecules-For-Portal-Training_1072	List for Desciptor calculation	2008-12-01 15:41:00.057	smith		
U	IN	PM	☐	RC	Menu	1072	Copy of Test_Descriptor_Calculation_006_5000_1071	List for Desciptor calculation	2009-05-28 13:36:51.33	gpocze		
U	IN	PM	☐	RC	Menu	1071	Copy of Test_Descriptor_Calculation_005_5000_1070	List for Desciptor calculation	2009-05-28 13:36:43.503	gpocze		
U	IN	PM	☐	RC	Menu	1070	Copy of Test_Descriptor_Calculation_004_5000_1069	List for Desciptor calculation	2009-05-28 13:36:36.597	gpocze		
U	IN	PM	☐	RC	Menu	1069	Copy of Test_Descriptor_Calculation_003_5000_1068	List for Desciptor calculation	2009-05-28 13:36:30.317	gpocze		

Listitems view

CancerGrid



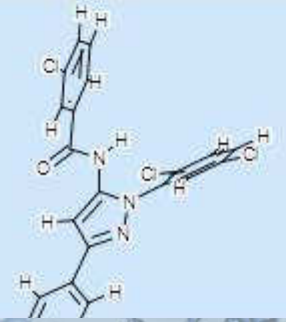
DELETE	INSERT AS NEW	1 RECORD MODE	SELECT MORE	REPORTS	MENU	MOLECULE	LISTITEM ID	LIST ID	MOLECULE ID	CANCERGRID CODE	PARENT MOLECULE ID
							ASC DESC	ASC DESC	ASC DESC	ASC DESC	ASC DESC
D	IN	PM	☐	RC	Menu	-	5965042	1072	1286246		256336
D	IN	PM	☐	RC	Menu	-	5965041	1072	1286245		256336
D	IN	PM	☐	RC	Menu	-	5965040	1072	1286244		256336
D	IN	PM	☐	RC	Menu	-	5965039	1072	1286243		256336
D	IN	PM	☐	RC	Menu	-	5965038	1072	1286242		256336
D	IN	PM	☐	RC	Menu	-	5965037	1072	1286241		256336
D	IN	PM	☐	RC	Menu	-	5965036	1072	1286240		256336
D	IN	PM	☐	RC	Menu	-	5965035	1072	1286239		256336
D	IN	PM	☐	RC	Menu	-	5965034	1072	1286238		256336
D	IN	PM	☐	RC	Menu	-	5965033	1072	1286237		256336
D	IN	PM	☐	RC	Menu	-	5965032	1072	1286236		246615
D	IN	PM	☐	RC	Menu	-	5965031	1072	1286235		246615
D	IN	PM	☐	RC	Menu	-	5965030	1072	1286234		246615
D	IN	PM	☐	RC	Menu	-	5965029	1072	1286233		246615
D	IN	PM	☐	RC	Menu	-	5965028	1072	1286232		246615

Molecule viewer

CancerGrid

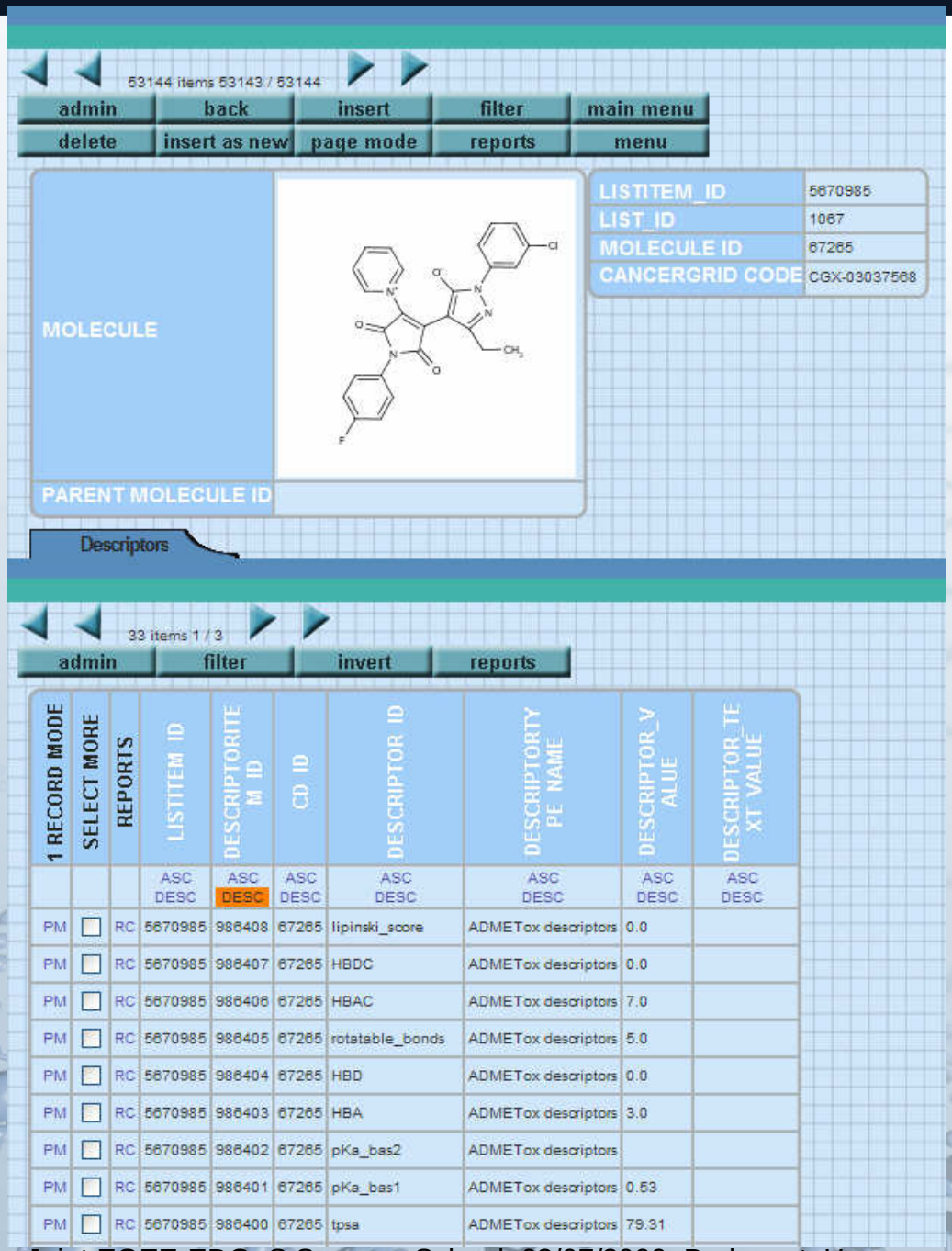
53510 items 1 / 1071

admin | back | insert | filter | invert | reports | main menu

DELETE	INSERT AS NEW	1 RECORD MODE	SELECT MORE	REPORTS	MENU	MOLECULE	LISTITEM ID	LIST ID	MOLECULE ID	CANCERGRID CODE	PARENT MOLECULE ID
							ASC DESC	ASC DESC	ASC DESC	ASC DESC	ASC DESC
D	IN	PM	☐	RC	Menu		5965042	1072	1286246		256336
D	IN	PM	☐	RC	Menu		5965041	1072	1286245		256336

CancerGrid

Inspecting molecules one by one…

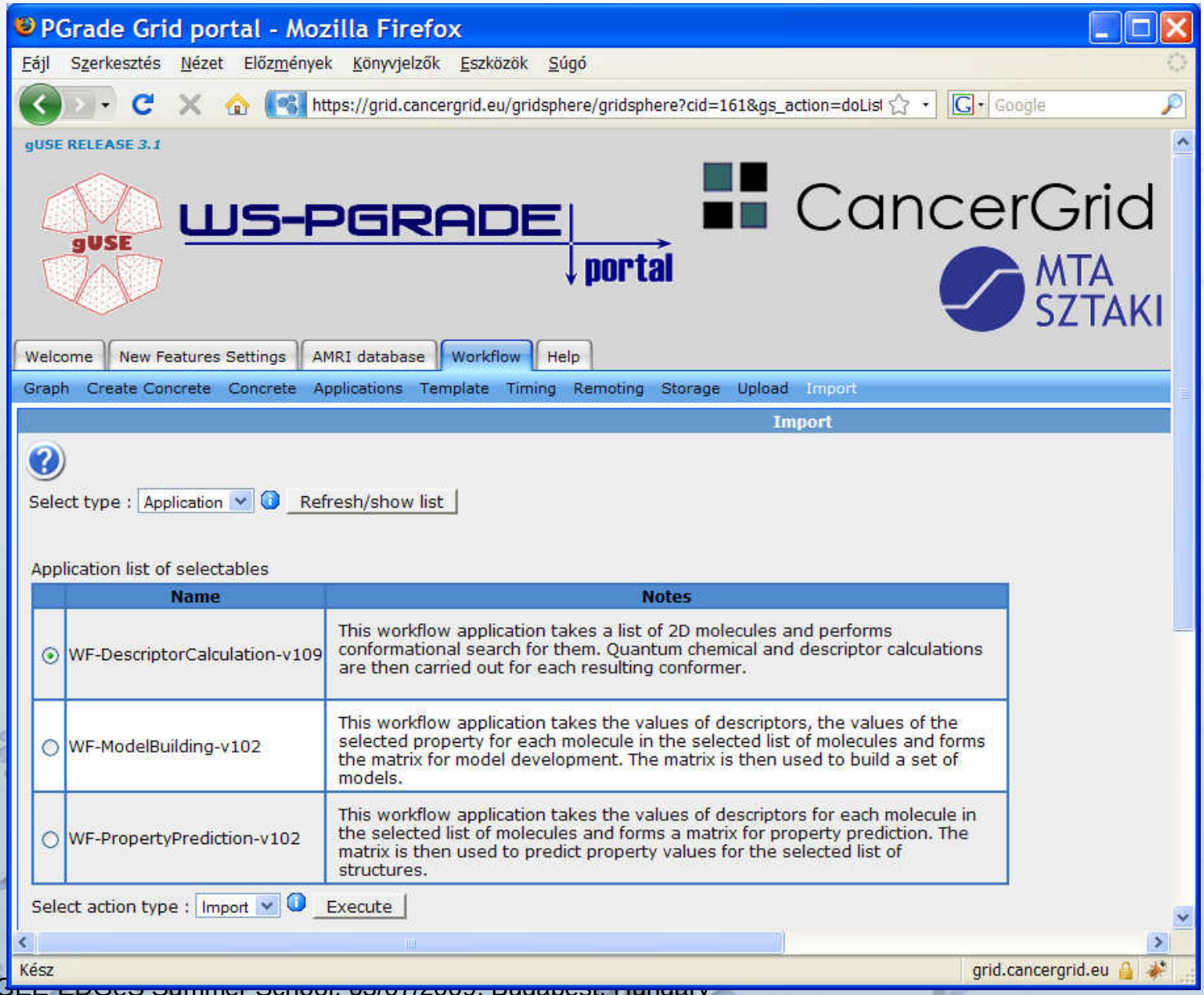


Manipulating the lists

CancerGrid

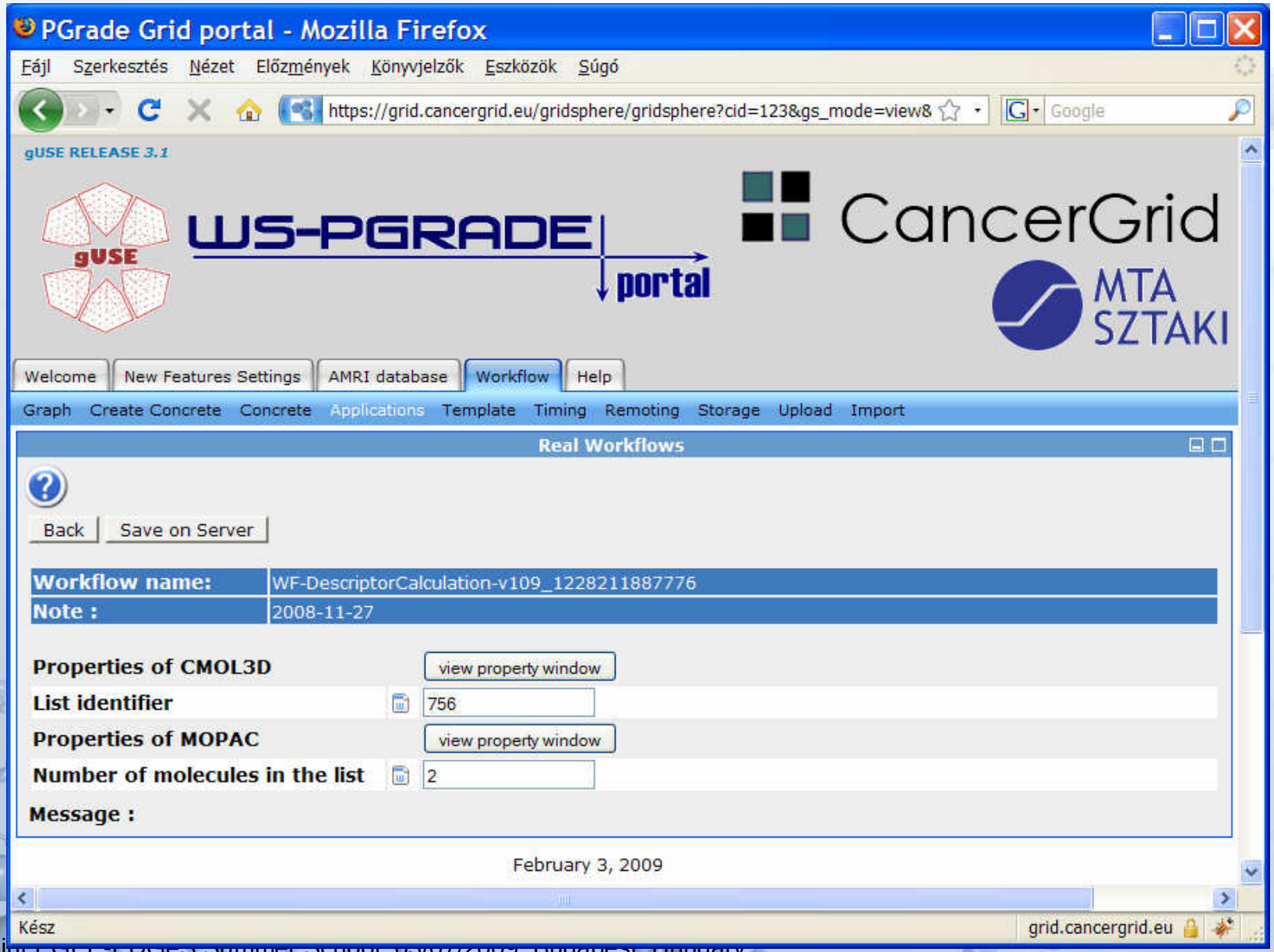
U	IN	PM	☐	RC	Menu	1074	Copy of Copy of BUDAPEST-2D-Molecules-For-Portal-Training_1072_1073	List for Desciptor calculation	2008-12-01 15:41:00.057	smith
U	IN	PM	☐	RC	Menu	1073	Copy of BUDAPEST-2D-Molecules-For-Portal-Training_1072	List for Desciptor calculation	2008-12-01 15:41:00.057	smith
U	IN	PM	☐	RC	Menu	1072	Copy of Test_Descriptor_Calculation_005_5000_1071	List for Desciptor calculation	2009-05-28 13:36:51.33	gpocze
U	IN	PM	☐	RC	Menu	1071	Copy of Test_Descriptor_Calculation_005_5000_1070	List for Desciptor calculation	2009-05-28 13:36:43.503	gpocze
U	IN	PM	☐	RC	Menu	1070	Copy of Test_Descriptor_Calculation_004_5000_1069	List for Desciptor calculation	2009-05-28 13:36:36.597	gpocze
U	IN	PM	☐	RC	Menu	1069	Copy of Test_Descriptor_Calculation_003_5000_1068	List for Desciptor calculation	2009-05-28 13:36:30.317	gpocze
U	IN	PM	☐	RC	Menu	1068	Copy of Test_Descriptor_Calculation_002_5000_1067	List for Desciptor calculation	2009-05-28 13:36:24.16	gpocze
U	IN	PM	☐	RC	Menu	1067	Copy of Test_Descriptor_Calculation_001_5000_1066	List for Desciptor calculation	2009-05-28 13:36:17.363	gpocze
U	IN	PM	☐	RC	Menu	1066	Test_Descriptor_Calculation_006_5000	User list	2009-05-28 13:36:51.33	gpocze
U	IN	PM	☐	RC	Menu	1065	Test_Descriptor_Calculation_005_5000	User list	2009-05-28 13:36:43.503	gpocze
U	IN	PM	☐	RC	Menu	1064	Test_Descriptor_Calculation_004_5000	User list	2009-05-28 13:36:36.597	gpocze
U	IN	PM	☐	RC	Menu	1063	Test_Descriptor_Calculation_003_5000	User list	2009-05-28 13:36:30.317	gpocze
U	IN	PM	☐	RC	Menu	1062	Test_Descriptor_Calculation_002_5000	User list	2009-05-28 13:36:24.16	gpocze
U	IN	PM	☐	RC	Menu	1061	Test_Descriptor_Calculation_001_5000	User list	2009-05-28 13:36:17.363	gpocze
U	IN	PM	☐	RC	Menu	1060	Test_Screening_003_6_50000_Duplicated_List_Not_work_property	User list	2009-05-28 10:03:15.177	gpocze
U	IN	PM	☐	RC	Menu	1059	Test_Screening_003_5_50000_Duplicated_List_Not_work_property	User list	2009-05-28 10:01:46.397	gpocze
U	IN	PM	☐	RC	Menu	1058	Test_Screening_003_5_50000_Duplicated_List_Not_work_property	Screening	2009-05-25 15:20:33.38	gpocze
U	IN	PM	☐	RC	Menu	1057	Copy of Test_Screening_003_5_50000_1056	Screening	2009-05-25 15:20:33.38	gpocze
U	IN	PM	☐	RC	Menu	1056	Test_Screening_003_4_50000_Duplicated_List_Not_work_property	Screening	2009-05-25 15:20:16.037	gpocze
U	IN	PM	☐	RC	Menu	1055	Copy of Test_Screening_003_4_50000_1053	Screening	2009-05-25 15:20:16.037	gpocze
U	IN	PM	☐	RC	Menu	1053	Copy of Test_Screening_003_3_50000_1051	Screening	2009-05-25 15:20:09.16	gpocze
U	IN	PM	☐	RC	Menu	1051	Copy of Test_Screening_003_2_50000_1044	Screening	2009-05-25 15:20:02.973	gpocze
U	IN	PM	☐	RC	Menu	1044	Copy of Test_Screening_003_1_50000_1041	Screening	2009-05-25 15:19:56.253	gpocze
U	IN	PM	☐	RC	Menu	1041	Copy of Test_Screening_002_10000_1040	Screening	2009-05-25 15:19:42.083	gpocze

Downloading workflow from repository



Workflow configuration

CancerGrid



Cmol3D property settings

CancerGrid

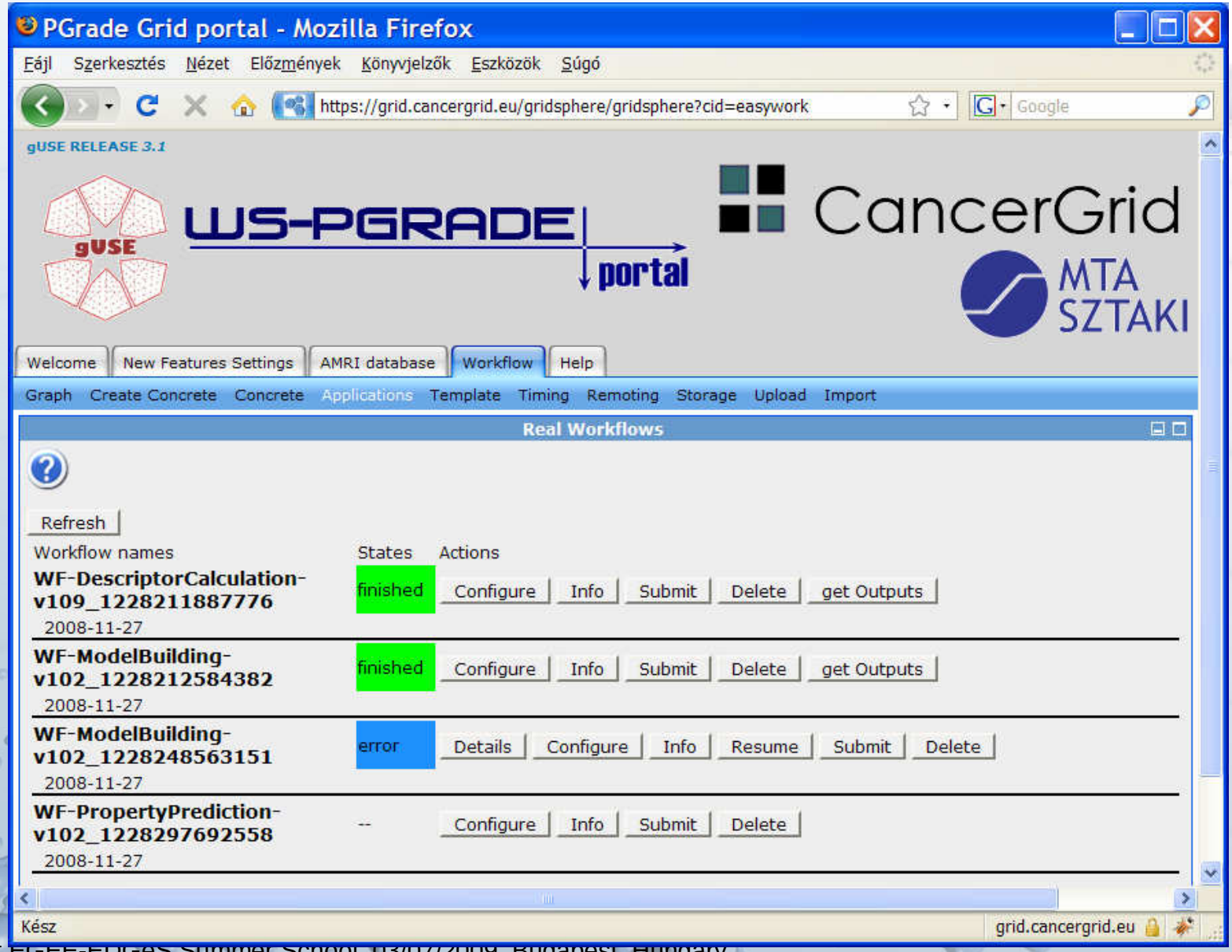
view property window

Cmol3D (conformational search)

Conformational search:	☑
Using 2D molecules:	☑
Population size:	100
Number of population initializers:	200
Energy window for saving structures (kcal/mol):	5.0
Maximum number of structures to save:	10
Maximum distance between atoms in equal structures:	0.25
Maximum number of perturbations for a single conformation:	10
Step size for low-mode move:	0.5
Maximum distance for low-mode move:	5.0
Number of lowest eigenvectors to use:	10

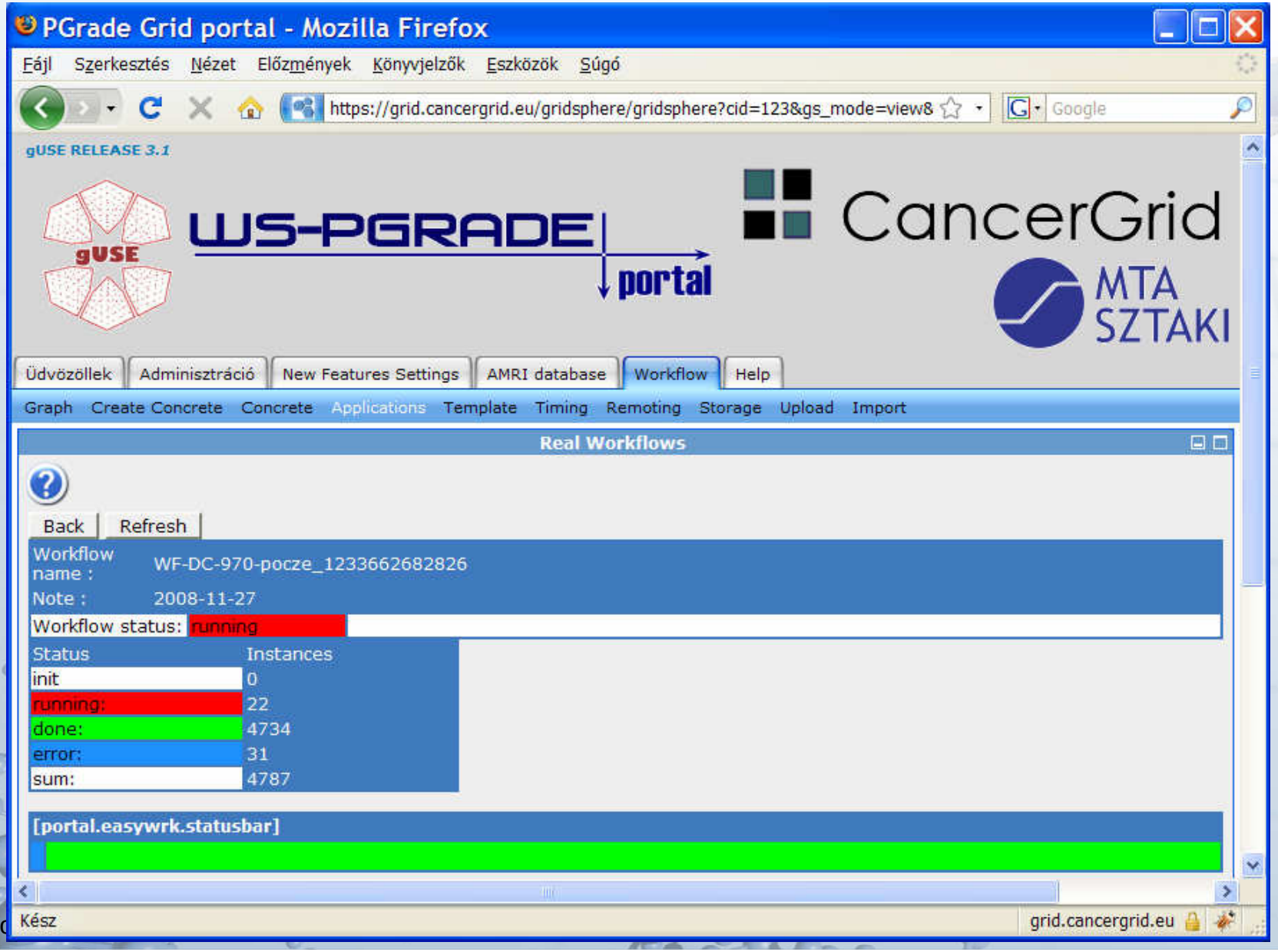
Ok Cancel Defaults

List of workflows (Novice user view)



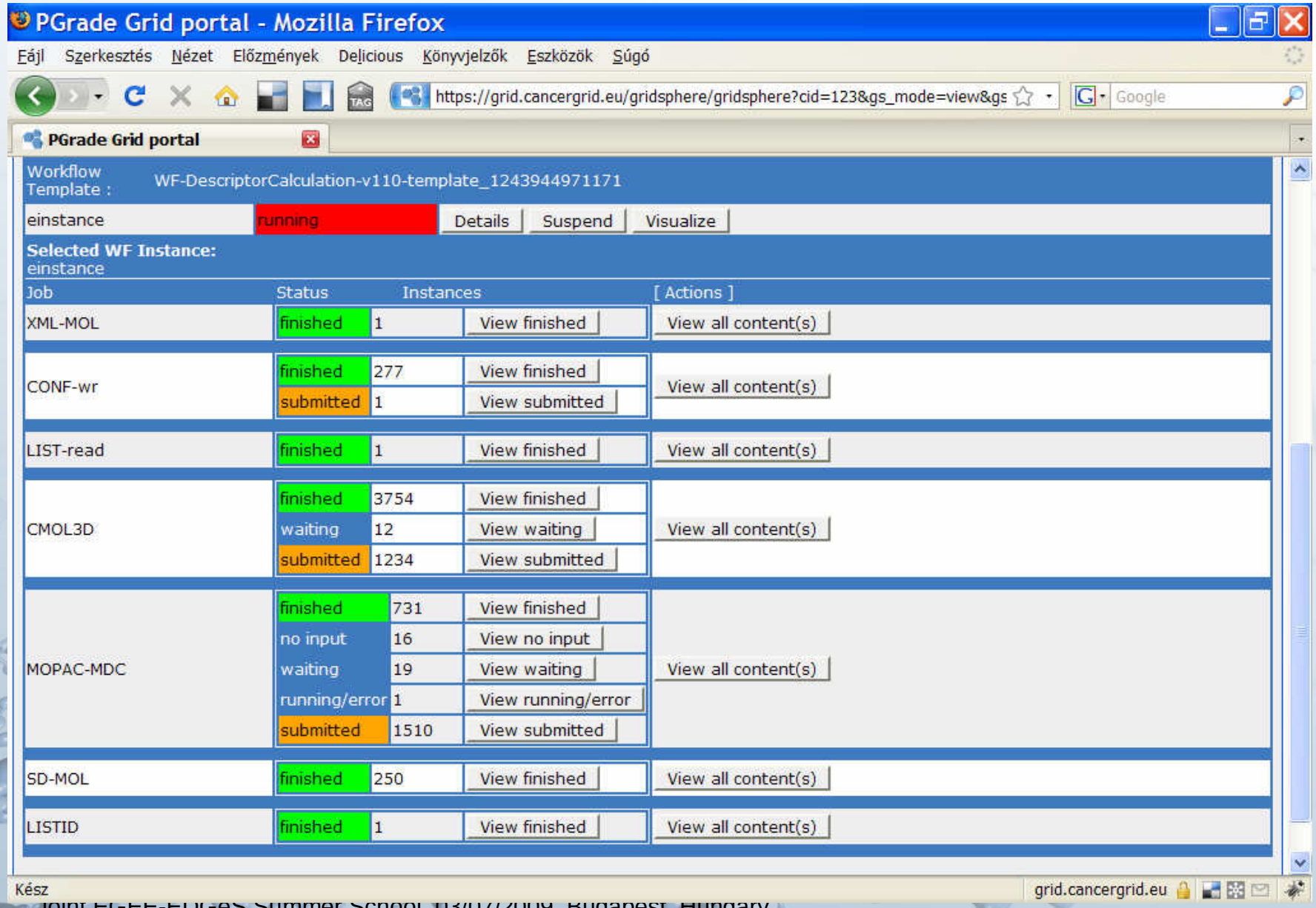
Status monitor

CancerGrid



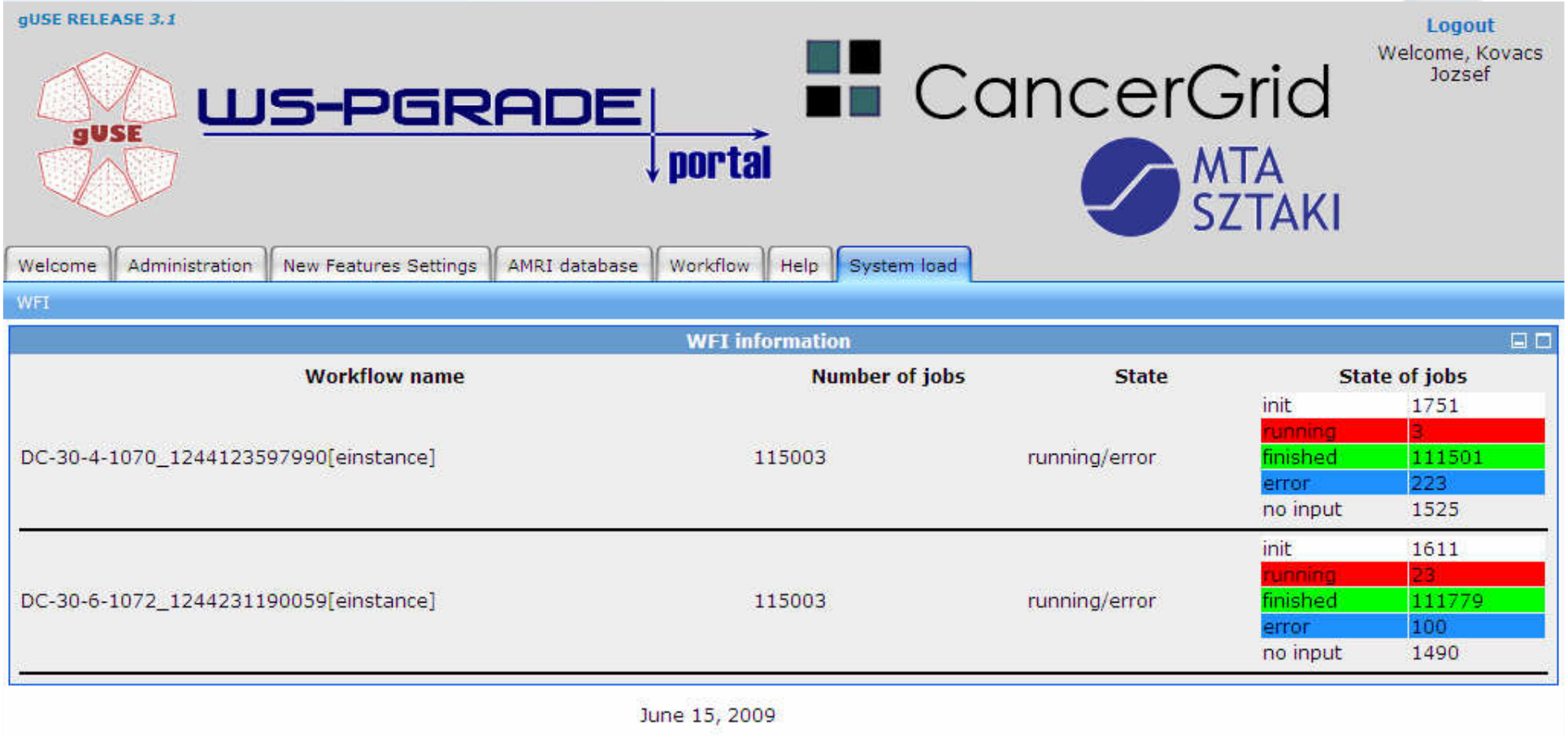
Job statuses of a workflow (Advanced view)

CancerGrid



Screenshots

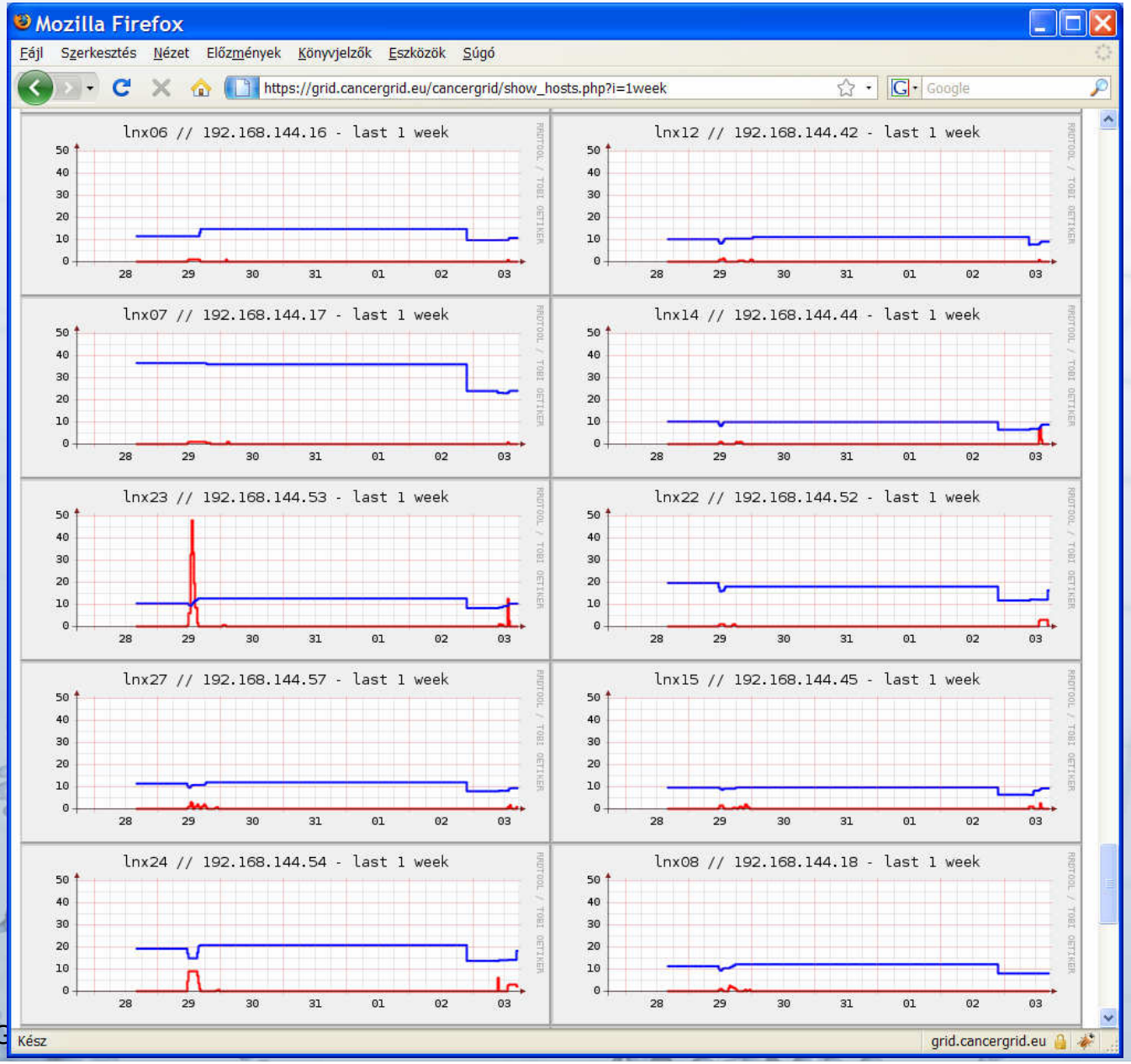
CancerGrid



June 15, 2009

Monitoring the clients

CancerGrid



The system in numbers…

CancerGrid

Statistics

- Currently 70 machines are integrated into the system, the goal is to have about 200 machines
- Requires only 1 central server to be maintained (clients needs no maintenance!)
- 13 algorithms has been integrated
- 4 workflows are available
- 5-6 partners from the consortium have used it for different purposes

Peak performance

- Description calculation for 30.000 compounds (with 10 conformers) takes 5 days (executing the combination of CMOL3D, MOPAC and MDC tools from CODESSA using 3D descriptors). With 1 PC it could be a year…

  Note: there are faster algorithms, but Grid is worth applying only for computationally intensive algorithms (like the ones above)
- By decreasing the number of conformers and by increasing the number of machines, the calculation can be even 100 times faster i.e. millions of compounds can be processed within 1 week. This is our current direction…

Summary & Future plan

CancerGrid

- Any community that has several computational intensive algorithms can easily execute them on a SZTAKI DesktopGrid (SZDG) system:
  - the community can create its own institutional SZDG project
  - can simply attach the available PC computers as resources
  - can insert the algorithms in their SZDG project
  - can easily combine these components into more complex workflow applications, if necessary
- Such a system
  - has been prototyped for the Cancer Research community within the CancerGrid projects
  - can be specialized for any community requirements
- SZTAKI is ready to support other communities with this technology
- In the future
  - we plan to continue extending the CancerGrid portal with more algorithms, workflows and PC resources
  - we plan to create an open portal where various algorithms from the field of computational chemistry and bioinformatics will be supported, which will be publicly available
  - we plan to integrate ZINC database

CancerGrid

If you need more detailed (technical) information,
email to desktopgrid@lpds.sztaki.hu or
visit www.desktopgrid.hu

MTA SZTAKI

Thank you for your attention!

Questions?

SZTAKI DESKTOP GRID

BOINC

Acknowledgement:

CancerGrid EU FP6 project (FP6-2005-LIFESCTHTALTH-7)

http://www.cancergrid.eu