

H5N1 data challenge – status and results

Hurng-Chun Lee
Academia Sinica, Taiwan

www.eu-egee.org

- The H5N1 data challenge
- The data analysis
- The interactive virtual screening on the Grid

- **H5N1 is high pathogenic**
- **H5N1 virus has the potential to cause a large-scale pandemic**
 - K. S. Li *et al*, “Genesis of highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia”, *Nature*, Vol. 430, 2004
- **H5N1 may mutate and acquire the ability of drug resistance**
 - Menno D. de Jong *et al*, “Oseltamivir Resistance during Treatment of Influenza A (H5N1) Infection”, *N. Engl. J. Med.*, 353:2667-2672, 2005

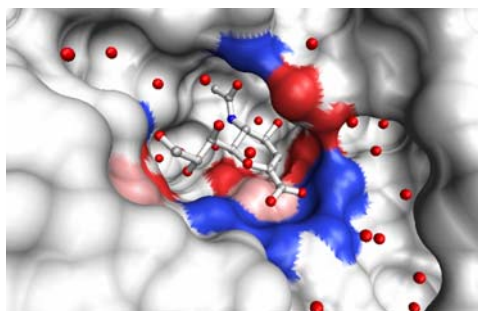
- Analyzing the efficiency of the known drugs to the possible mutations of the H5N1 virus
- Searching for new drugs
- Re-producing the Grid-enabled High Throughput Screening (HTS) following the first successful data challenge on Malaria
- Improving the reliability and efficiency of the Grid-enabled HTS service, moving toward an end-user Grid application

The challenge

Millions of chemical compounds available in laboratories



300,000 Chemical compounds:
ZINC
Chemical combinatorial library



Target (**PDB**) :
Neuraminidase (8 structures)



High Throughput Screening
2\$/compound, nearly impossible



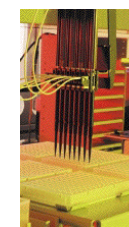
Molecular docking (**Autodock**)
~100 CPU years, 600 GB data



Data challenge on **EGEE**,
Auvergrid, **TWGrid**
~6 weeks on ~2000 computers

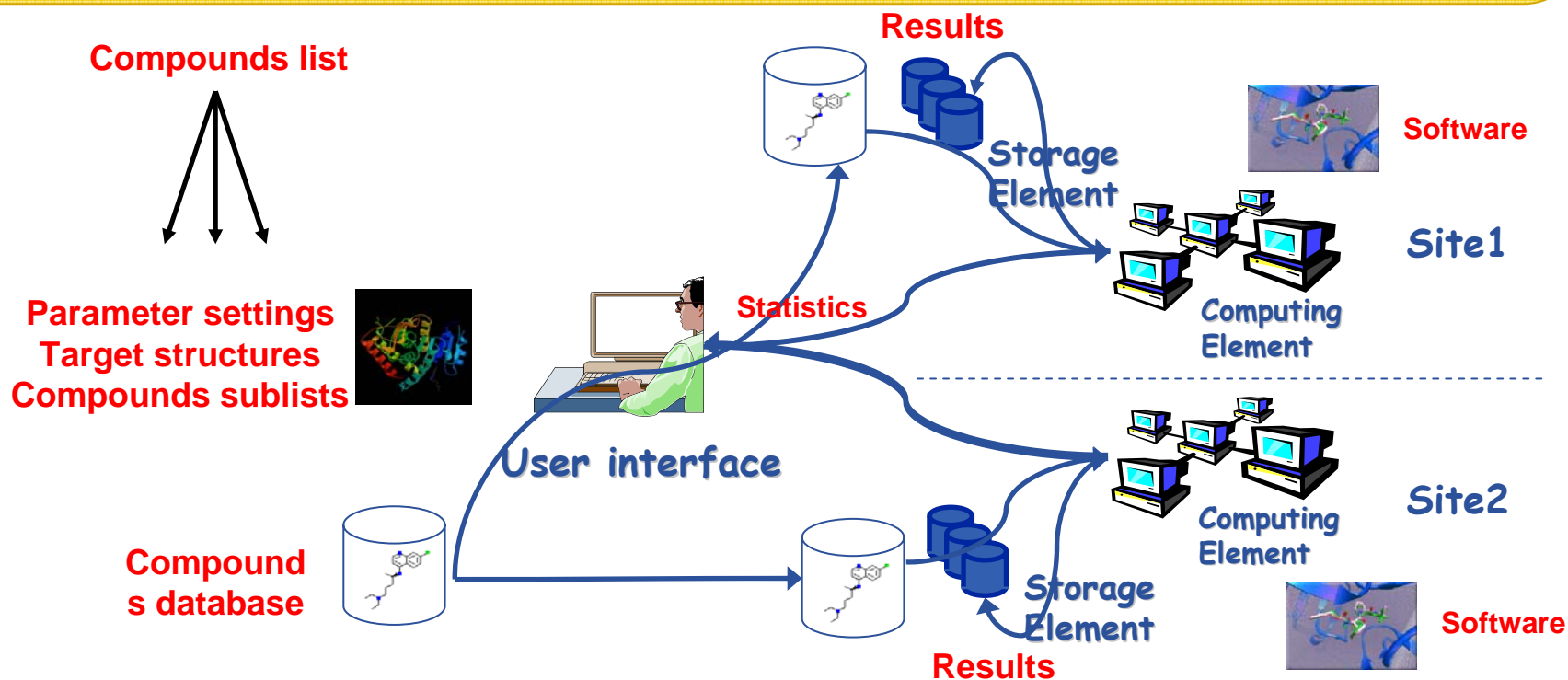


Hits sorting
and refining

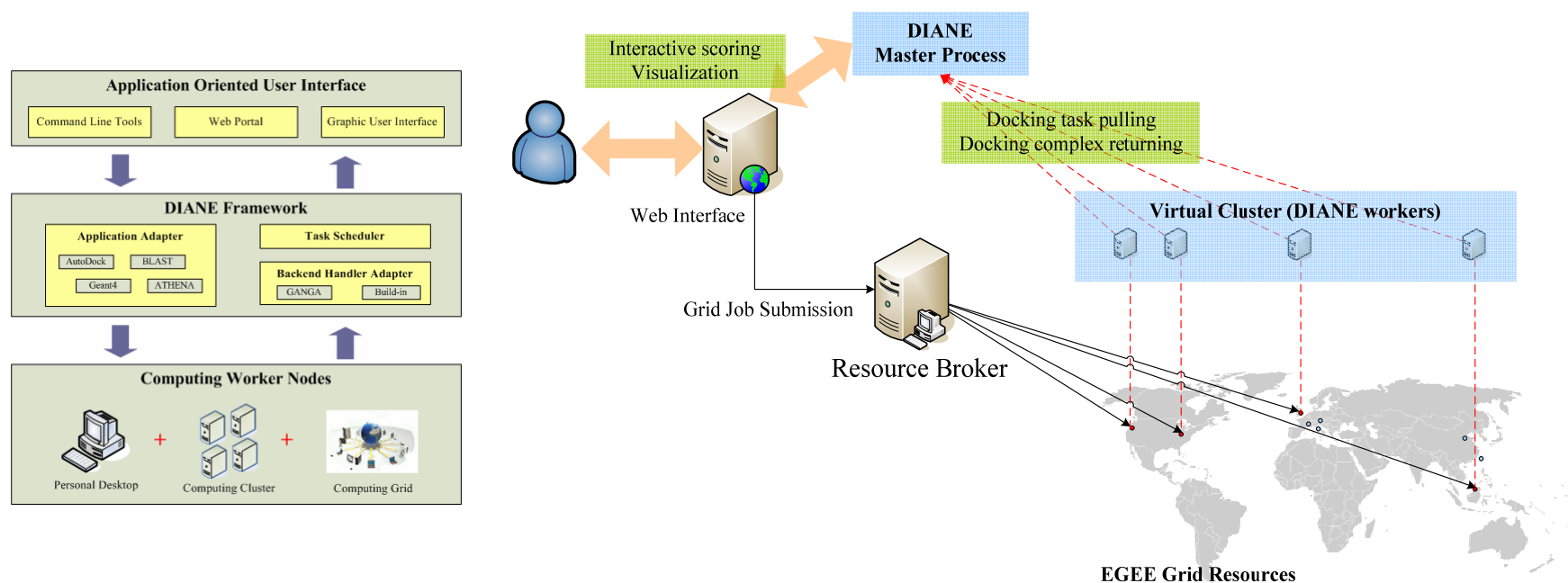


In vitro
screening
of 100 hits

- WISDOM: Wide In-Silico Docking On Malaria
- The platform has been successfully tested in previous challenge
- a workflow of Grid job handling: automatic job submission, status check and report, error recovery
- push model job scheduling + batch mode job handling



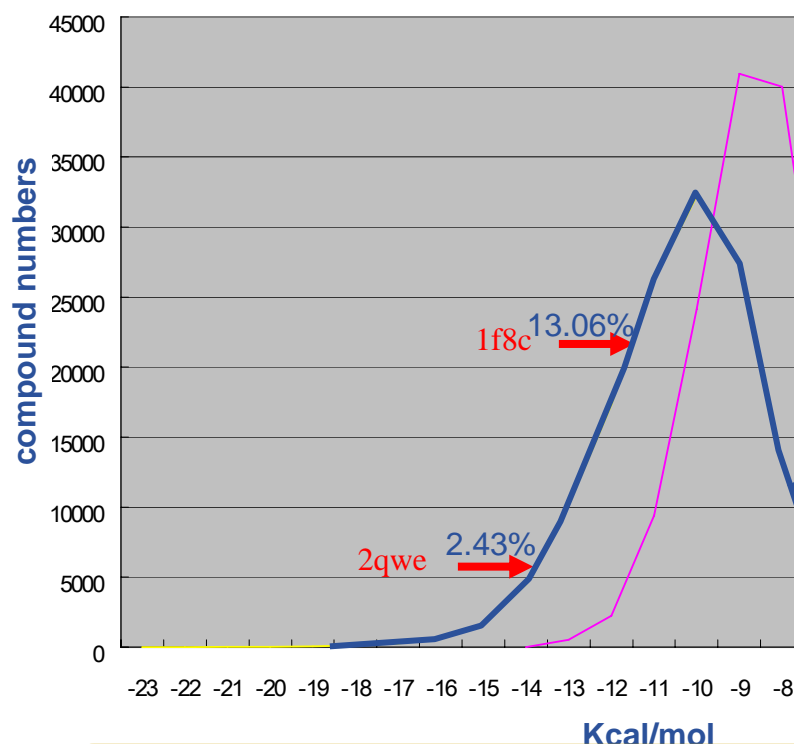
- DIANE: Distributed Analysis Environment
- An overlay system on top of a variety of distributed computing environment takes care of all synchronization, communication and workflow management details on behalf of application
- A lightweight framework for parallel scientific applications in master-worker model
- Pull model job scheduling + interactive mode job handling with flexible failure recovery mechanism



	WISDOM	DIANE
Total number of completed dockings	$2 * 10^6$	308,585
Estimated duration on 1 CPU	88.3 years	16.7 years
Duration of the experience	6 weeks	4 weeks
Cumulative number of the Grid jobs	54,000	2580
Max. number of concurrent CPUs	2,000	240
Crunching rate	912	203
Approximated distribution efficiency	46 %	84 %
Approximated throughput	2 sec/docking	10 sec./ docking

- ~600 GBytes of docking results are produced and archived on the Grid
- ~83% were successfully completed according to the Grid Logging and Bookkeeping; only ~70% of results were really produced on the Grid storage element

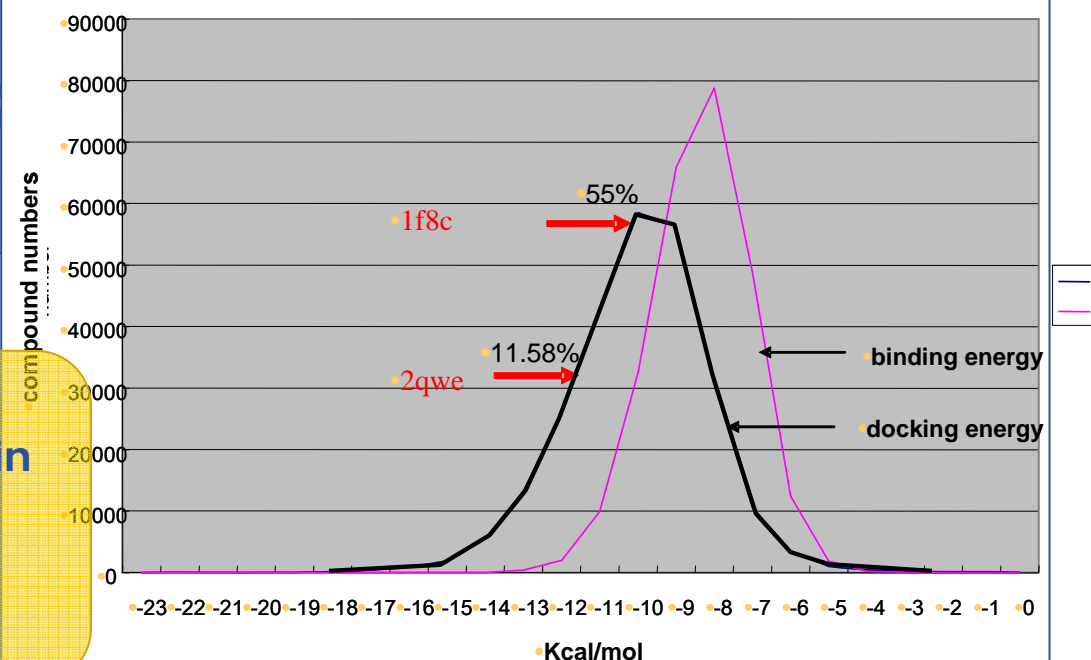
T06 (wild type)

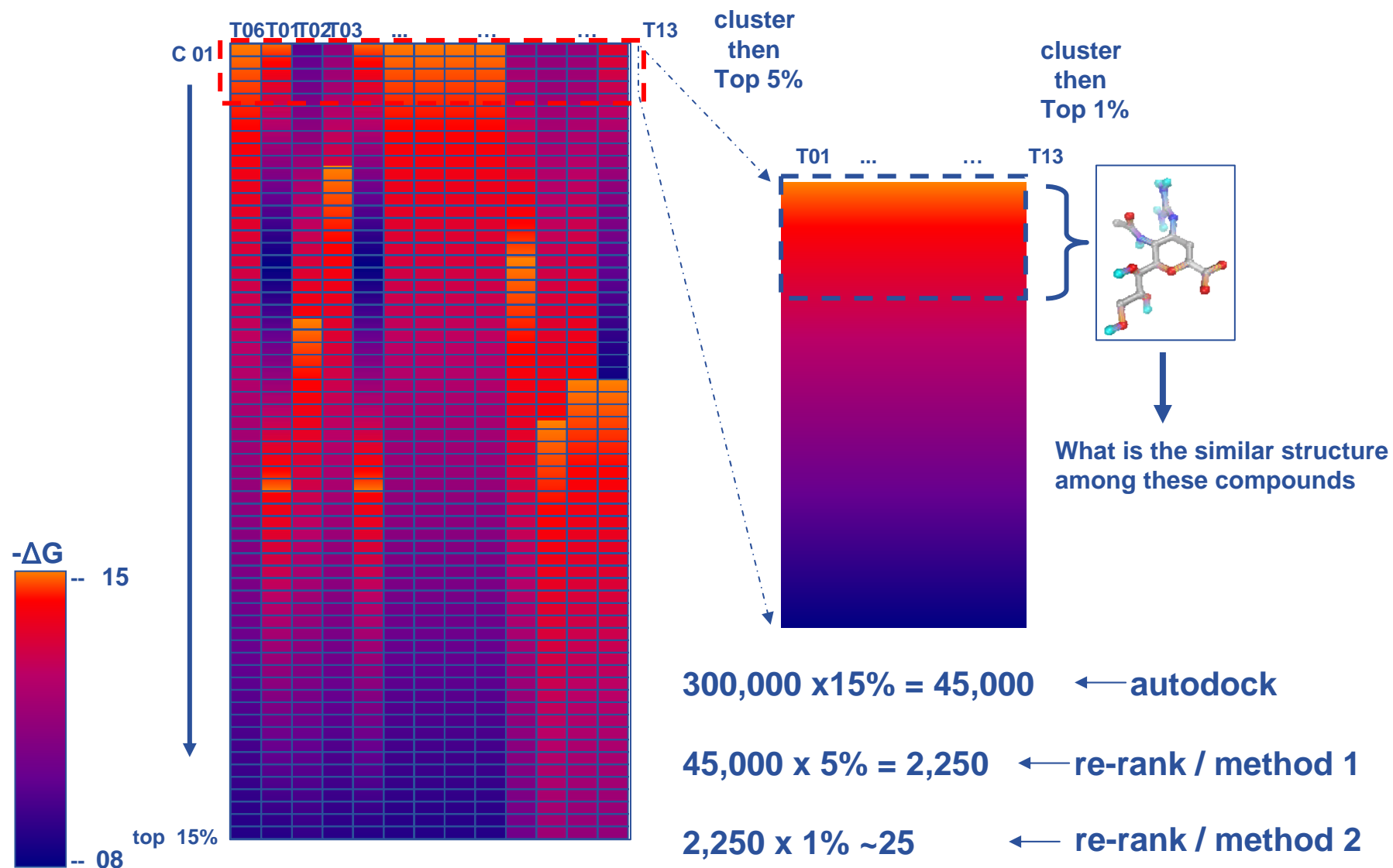


- 2qwe: Zanamivir (known drug)
- Two most effective drug compounds can be identified in the first 15% of the ranking

- The two most effective drug compounds lose the efficiency in binding with a mutated target

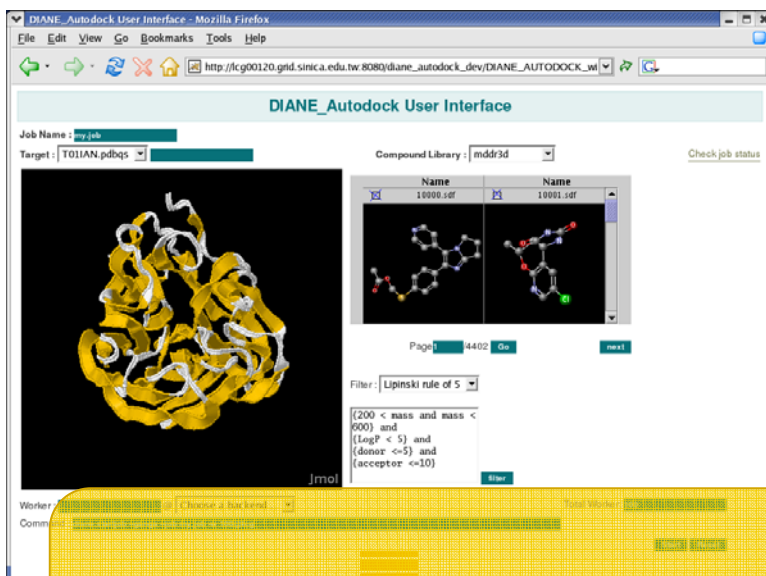
T01 (E119A)





- **Issues in the current computing model**
 - coordinative way of executing the data challenge is not feasible for normal end users
 - graphic interface allowing end users to intuitively configure docking parameters is not available
 - dealing a huge amount of produced docking results is still time-consuming
- **Leveraging on the DIANE framework, a web-based graphic interface was built to**
 - provide an intuitive interface for starting virtual screening on the Grid
 - monitor the progress of the virtual screening
 - visualize and summarize the completed dockings

The graphic user interface

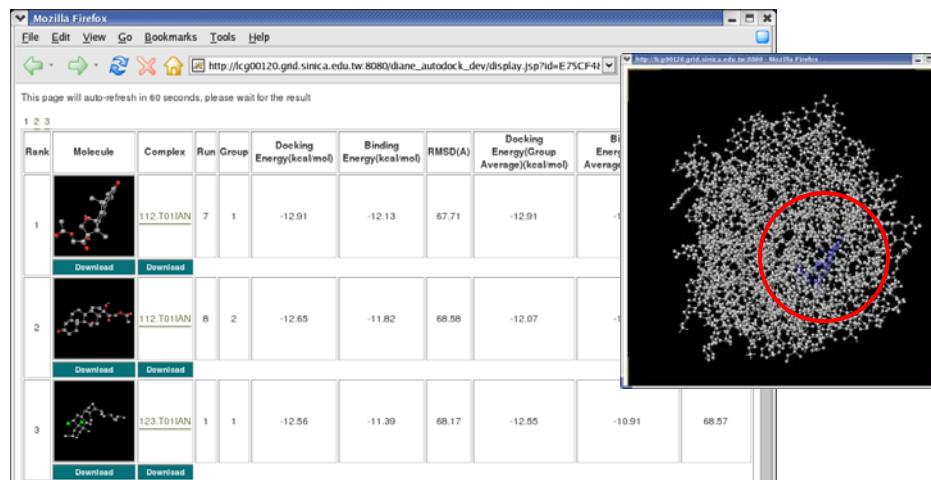


DIANE_Autodock User Interface




Job Name: my.job
Target: T01IAN.pdbqs
Compound Library: mddr3d

Filter: Lipinski rule of 5

(200 < mass and mass < 600) and (LogP < 5) and (donor <= 5) and (acceptor <= 10)



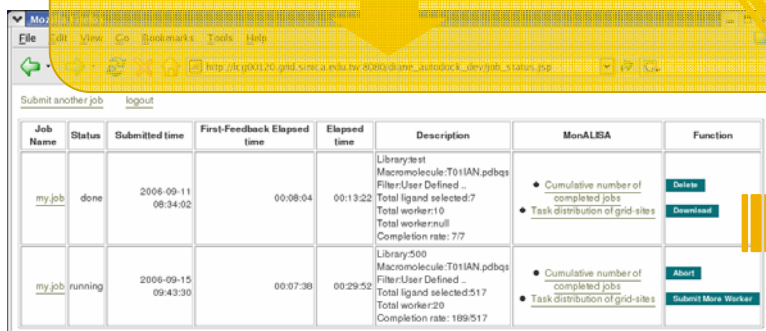
This page will auto-refresh in 60 seconds, please wait for the result

Rank	Molecule	Complex	Run	Group	Docking Energy(kcal/mol)	Binding Energy(kcal/mol)	RMSD(A)	Docking Energy(Group Average)(kcal/mol)	BI Energy Average
1		112.T01IAN	7	1	-12.91	-12.13	67.71	-12.91	
2		112.T01IAN	8	2	-12.65	-11.82	68.58	-12.07	
3		123.T01IAN	1	1	-12.56	-11.39	68.17	-12.55	-10.91 68.57

Download Download Download Download

3D molecular model showing a protein structure with a red circle highlighting a specific region.


Visit our booth in the demonstration session



Submit another job logout

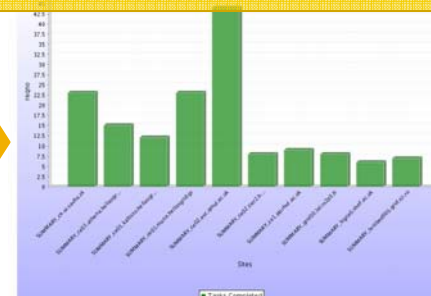
Job Name	Status	Submitted time	First-Feedback Elapsed time	Elapsed time	Description	MonALISA	Function
my.job	done	2006-09-11 09:34:02	00:08:04	00:13:22	Library:Test Macromolecule:T01IAN.pdbqs Filter:User Defined Total ligand selected:7 Total worker:10 Completion rate: 7/7	<ul style="list-style-type: none"> Cumulative number of completed jobs Task distribution of grid-sites 	Delete Download
my.job	running	2006-09-15 09:43:30	00:07:38	00:29:52	Library:200 Macromolecule:T01IAN.pdbqs Filter:User Defined Total ligand selected:517 Total worker:20 Completion rate: 169/517	<ul style="list-style-type: none"> Cumulative number of completed jobs Task distribution of grid-sites 	Alert Submit More Worker

Job History

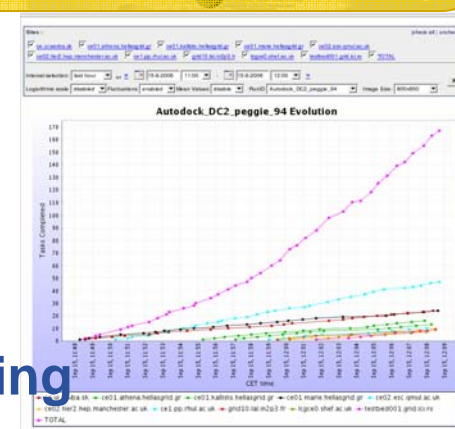


Submitted!
Number of submitted worker: 10
Type: .log

Close Window



Progress monitoring



- **We have reproduced a Grid-enabled high-throughput screening fighting against the H5N1 virus**
 - The 6-weeks activity has covered the computing requirement of over 100 CPU years
 - Two different computing models (WISDOM and DIANE) were adopted taking into account two different user aspects
- **We have prototyped a web-based graphic user interface aiming at providing an easy-to-use system for end users to do interactive screening on the Grid**
- **We are in the process of the data analysis trying to filter out over 99% of the compounds step-by-step**
 - the data challenge has helped to filter out 85%
 - the following steps are on-going

- **Docking workflow preparation**

- Contact point: Y.T. Wu
- E. Rovida
- P. D'Ursi
- N. Jacq

- **Grid resource management**

- Contact point: J. Salzemann
- TWGrid : H.C. Lee, H. Y. Chen
- AuverGrid : E. Medernach
- EGEE : Y. Legré

- **Platform deployment on the Grid**

- Contact point: H.C. Lee, J. Salzemann
- M. Reichstadt
- N. Jacq

- **Users (deputy)**

- J. Salzemann (N. Jacq)
- M. Reichstadt (E. Medernach)
- L. Y. Ho (H. C. Lee)
- I. Merelli, C. Arlandini (L. Milanesi)
- J. Montagnat (T. Glatard)
- R. Mollon (C. Blanchet)
- I. Blaque (D. Segrelles)
- D. Garcia



Academia Sinica
Genomics Research Center



IN2P3
INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE
ET DE PHYSIQUE DES PARTICULES

