

GROCK

(GRid dOCK)

High Throughput Docking on the Grid

EMBnet/CNB

EGEE 4th Conference, Pisa October 2005

Long-term Goal

- ◆ The goal of GROCK is to develop a 3D structural complementarity screening tool:
 - ◆ find best matches between two molecular structures
 - ◆ for a probe molecule against all molecules in a database
 - ◆ drug against proteins
 - ◆ protein against proteins
 - ◆ protein against drugs

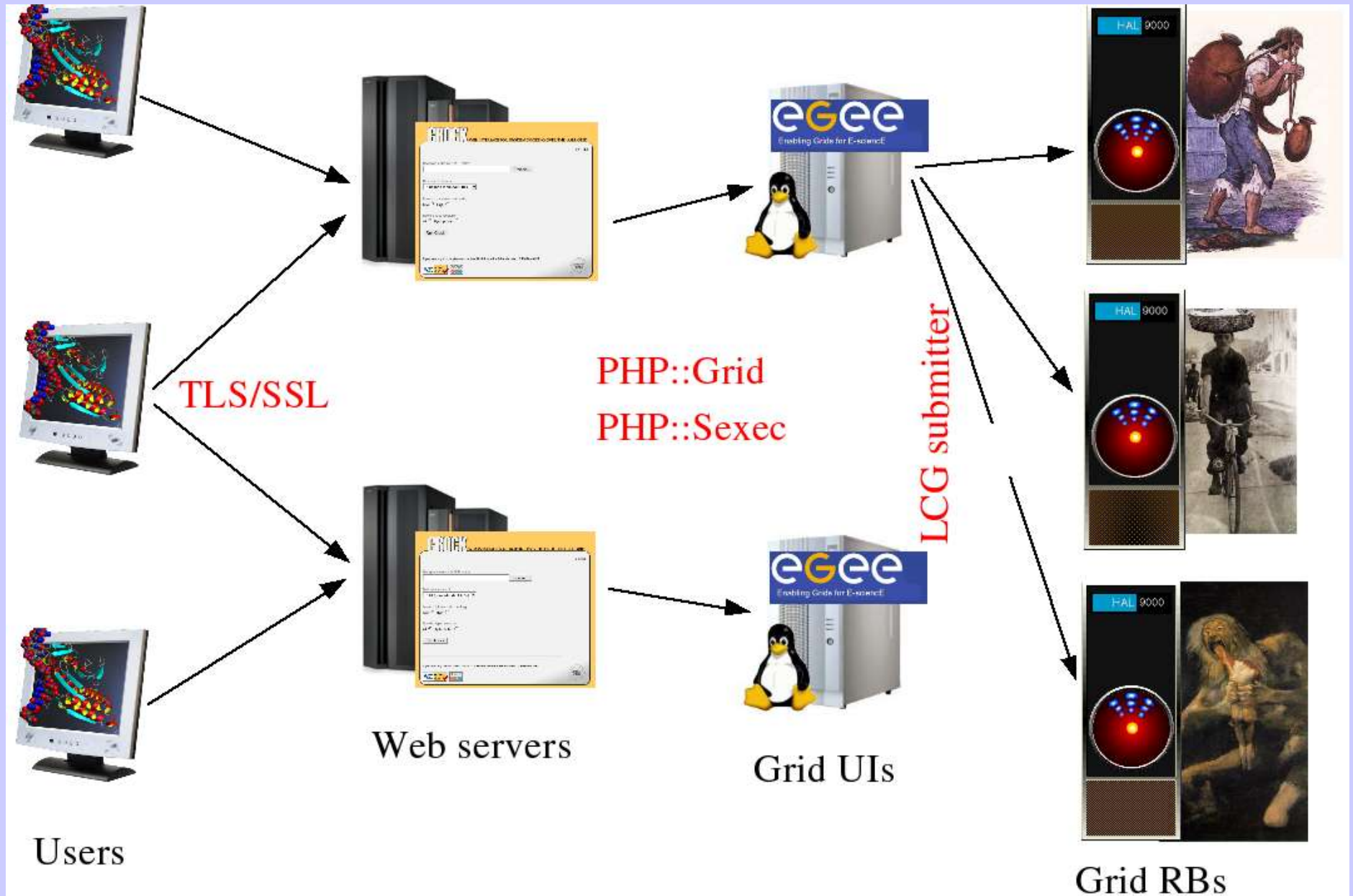
The Scientist Wishes

- ◆ To understand biomolecular interactions
 - ◆ to predict protein interactions
 - ◆ to detect putative drugs against target proteins
 - ◆ to screen drugs for putative effects
- ◆ In a way that is
 - ◆ easy to use
 - ◆ more reliable
 - ◆ feasible
 - ◆ efficient

Fulfilling Scientists Needs

- ◆ GROCK is a web-based tool that makes 3D molecular docking searches of a probe molecule against a database:
 - ◆ **Easy to use** thanks to an intuitive web interface
 - ◆ **More reliable** than pharmacophores thanks to use of well-known 3D docking methods
 - ◆ **Feasible** thanks to use of standard data and equipment
 - ◆ **Efficient** thanks to the Grid (EGEE)

GROCK Architecture



Cost Analysis

- ◆ Using GROCK allows scientists to explore high-throughput molecular docking using shared Grid resources (virtually free)
- ◆ GROCK expected use is initially low
- ◆ Without any need for expensive commercial software or databases
 - ◆ Yet, the latest would probably prove immensely useful (distilled data collections).
 - ◆ E.g. PDB contains much noise
- ◆ **GROCK is GPL**

Strengths and Advantages

- ◆ **Easy to use**
 - ◆ Simple interface
 - ◆ Powerful match explorer
- ◆ **Extensible** plugin mechanism for adding
 - ◆ databases
 - ◆ docking methods
- ◆ Exploits the **Grid**
 - ◆ Massive computing power
 - ◆ Massive storage
 - ◆ Efficient and resilient
 - ◆ Reduced (shared) cost

GROCK byproducts

- ◆ GridGRAMM
 - ◆ submit a single docking job to the Grid
- ◆ php::SExec
 - ◆ PHP class to manage SSH connections
- ◆ php::Grid
 - ◆ PHP class to manage Grid connections, sessions and jobs
- ◆ LCG-submitter for Biomed...

Submission and Tracking of large bunches of jobs is a common practice of all LCG communities

However this can become a difficult task

- too many submissions and jobs for monitoring

A new complete tool has been developed for large production

→ Developed originally for the Geant4 Collaboration

- Flexible enough to be used for any VO and any user application
- Adapted to the Biomed needs and presented in this talk
- Most of the improvements mostly relative to handling the output

Documentation: “*LCG2 User Guide*”

<http://grid-deployment.web.cern.ch/grid-deployment/cgi-bin/index.cgi?var=eis/docs>

Download:

http://goc.grid.sinica.edu.tw/gocwiki/User_tools

The framework was modified for the Biomed community:

It was decided to use all the available RBs available for Biomed

It was mandatory not to lose any job because of RB problems

All the available RBs for the Biomed VO were obtained from the IS
(11 RBs in total)

- Every job is sent to a randomly chosen RBs

- Homogeneous distribution of RB

In the case the chosen RB presents submission problems a new random number is calculated to reassign a new RB

- Results: 100% of submission efficiency

Next Steps of Action

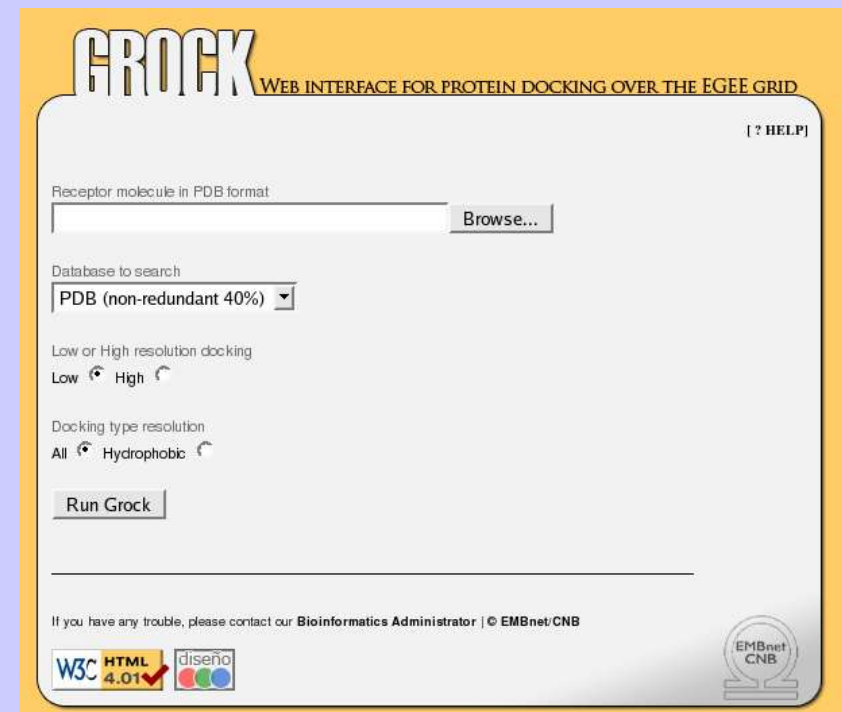
- ◆ Add support for additional docking methods
 - ◆ ftdock
 - ◆ autodock...
- ◆ Add support for other databases
 - ◆ HIC-Up
 - ◆ ZINC subsets
- ◆ Exploit Grid distributed storage system
 - ◆ Needed for truly massive jobs (e.g. drug screening)

GROCK in action

- ◆ Currently GROCK only supports
 - ◆ **Searching against PDB**
 - ◆ drug against proteins (drug effects)
 - ◆ protein against proteins (protein interactions)
 - ◆ PDB is noisy (sic)
 - ◆ **Using GRAMM**
 - ◆ general purpose method
 - ◆ very fast
 - ◆ less accurate (sic)
- ◆ But extensions are planned and on their way

A Real Time example

- ◆ Just for fun: we'll run a screening of aspirin against a small test database
 - ◆ Connect to **GROCK** server
 - ◆ Upload **aspirin**
 - ◆ Select options
 - ◆ Run



The screenshot shows the GROCK web interface, titled "GROCK WEB INTERFACE FOR PROTEIN DOCKING OVER THE EGEE GRID". The interface includes a text input field for "Receptor molecule in PDB format" with a "Browse..." button. Below this is a "Database to search" dropdown menu currently set to "PDB (non-redundant 40%)". There are two radio button options for "Low or High resolution docking": "Low" (selected) and "High". Another set of radio buttons for "Docking type resolution" shows "All" (selected) and "Hydrophobic". A "Run Grock" button is positioned below these options. At the bottom of the interface, there is a footer with the text "If you have any trouble, please contact our Bioinformatics Administrator | © EMBnet/CNB" and logos for W3C HTML 4.01 and diseño CNB.

- ◆ A full run takes longer than a demo session to run. We have various samples:
 - ◆ An incomplete run
 - ◆ Low resolution aspirin against test PDB
 - ◆ High resolution aspirin against test PDB
 - ◆ Low resolution aspirin against PDB 40% (non-redundant subset at 40% similarity)

Listing of top scores (1 - 51)



Match	Energy (-)	Description	Explore
aspirin vs. 1BRV_	700	1BRV_ mol:protein length:32 Protein G	Best 10 scores
aspirin vs. 1HCW_	528	1HCW_ mol:protein-het length:25 Bba1	Best 10 scores
aspirin vs. 1BZG_	524	1BZG_ mol:protein length:34 Parathyroid Hormone-Related Protein	Best 10 scores
aspirin vs. 1ANS_	496	1ANS_ mol:protein length:27 Neurotoxin III (Atx III) (NMR, 28 Structures)	Best 10 scores
aspirin vs. 1BAH_	494	1BAH_ mol:protein-het length:37 Charybdotoxin	Best 10 scores
aspirin vs. 1ACW_	486	1ACW_ mol:protein length:29 Natural Scorpion Peptide P01	Best 10 scores
aspirin vs. 1TFS_	476	1TFS_ mol:protein length:60 Toxin Fs2 (NMR, 20 Structures) - Chain _	Best 10 scores
aspirin vs. 1ERY_	464	1ERY_ mol:protein length:39 Pheromone Er-11	Best 10 scores
aspirin vs. 1AGG_	464	1AGG_ mol:protein length:48 Omega-Agatoxin-Ivb	Best 10 scores
aspirin vs. 1PFT_	453	1PFT_ mol:protein length:50 Tfilb	Best 10 scores
aspirin vs. 1B45_	452	1B45_ mol:protein-het length:15 Alpha-Cnia	Best 10 scores
aspirin vs. 1GAB_	450	1GAB_ mol:protein length:53 Protein Pab	Best 10 scores
aspirin vs. 1ERP_	449	1ERP_ mol:protein length:38 Pheromone Er-10 (NMR, 20 Structures) - Chain	Best 10 scores
aspirin vs. 1ERD_	445	1ERD_ mol:protein length:40 Pheromone Er-2 (NMR, 20 Structures) - Chain _	Best 10 scores
aspirin vs. 2MLP_	440	2MLP_ mol:protein-het length:27 Mcba Propeptide	Best 10 scores
aspirin vs. 1AFH_	440	1AFH_ mol:protein length:93 Maize Nonspecific Lipid Transfer Protein	Best 10 scores

Some noteworthy observations

- ◆ GROCK uses LCG submission system (thanks to Patricia Méndez @ CERN)
- ◆ GROCK detects and resubmits failed jobs
- ◆ Results may be saved for later analysis
- ◆ Matches may be explored individually
- ◆ Some matches may be unoptimal/misleading
 - ◆ unavailable or wrong data (e.g. RMN)
 - ◆ spurious matches on irrelevant organisms
 - ◆ real target being substituted by a remote relative
- ◆ **User must exercise cautious discretion**

Aspirin (acetylsalicylic acid)

- ◆ Induces its effect through phospholipase A2
 - ◆ Which is not on the search subset itself (sic)
- ◆ But has many other effects
 - ◆ on Protein G signalling
 - ◆ modulates hormone stimulated cyclic AMP production
 - ◆ protects against neurotoxicity
 - ◆ is used in dyslipidaemias
 - ◆ affects pulmonary surfactant
 - ◆ etc... (check PubMed).

Final comments

- ◆ GROCK to be presented at Grid/SC'05
- ◆ GROCK byproducts (mw) are available
- ◆ GROCK is fully automated
 - ◆ massive submission system (P. Méndez)
 - ◆ automatic error detection and recovery
 - ◆ dynamic resource allocation (sad but true)
- ◆ GROCK future
 - ◆ human control (comments?)
 - ◆ data distribution
 - ◆ more databases and dockers

Just a coincidence... to be proud of

After our first presentation of GROCK in Slovakia we were made aware of Google search results for Grock:



We wish to thank

- ◆ **YOU ALL**
 - ◆ for being here, your help, encouragement, feedback and support
- ◆ **Patricia Méndez (CERN)**
- ◆ **The TEAM at CNB**
 - ◆ **Bioinformatics**
 - ◆ *José R. Valverde, David J. García*
 - ◆ **Biocomputing**
 - ◆ José M. Carazo, Carlos Pérez-Roca, Enrique de Andrés, Natalia Jiménez, Sjors Schëres,...
- ◆ **THE E.U. for EGEE** (three hurrays for them!)