

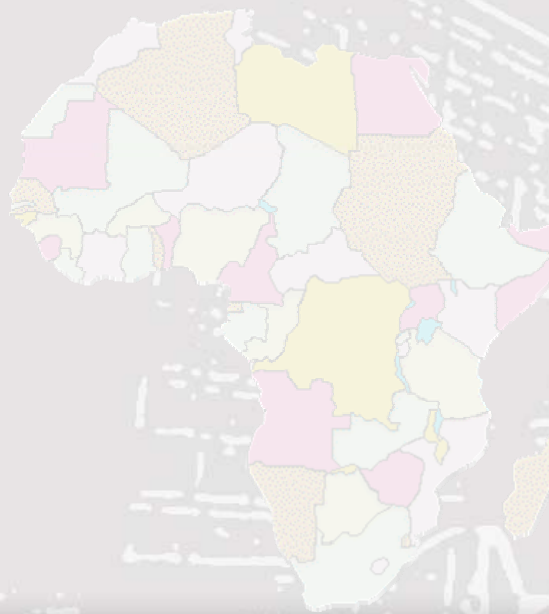


# Large-scale distributed *in silico* drug discovery using VSM-G

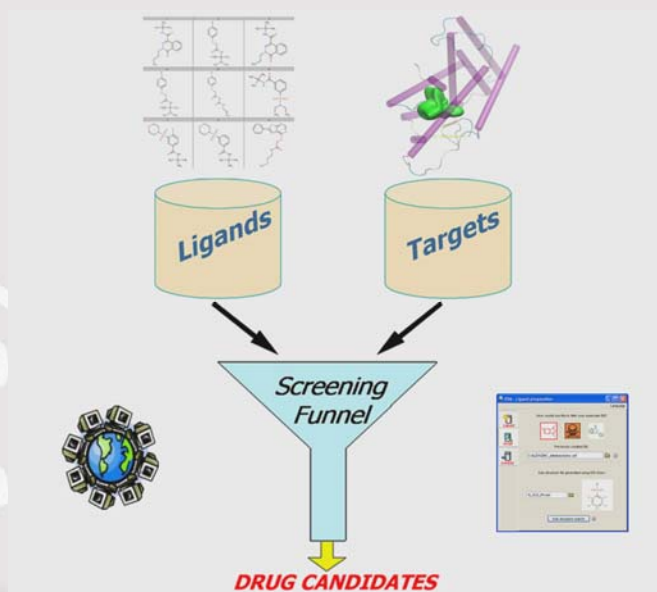
Leo GHEMTIO : Phd Student ([lghemtio@loria.fr](mailto:lghemtio@loria.fr))

Bernard MAIGRET : Advisor ([maigret@loria.fr](mailto:maigret@loria.fr))

INRIA LORRAINE, <http://www.loria.fr>

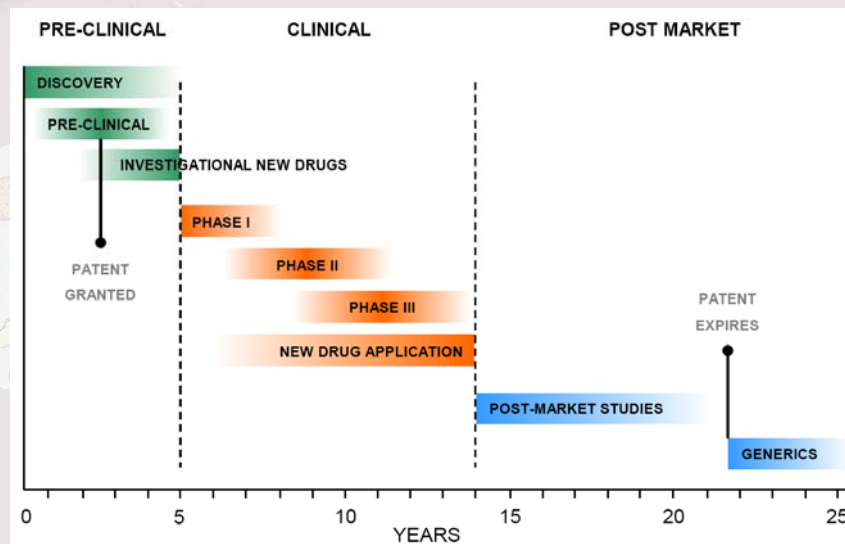


**V**irtual  
**S**creening  
**M**anager for  
**G**rids

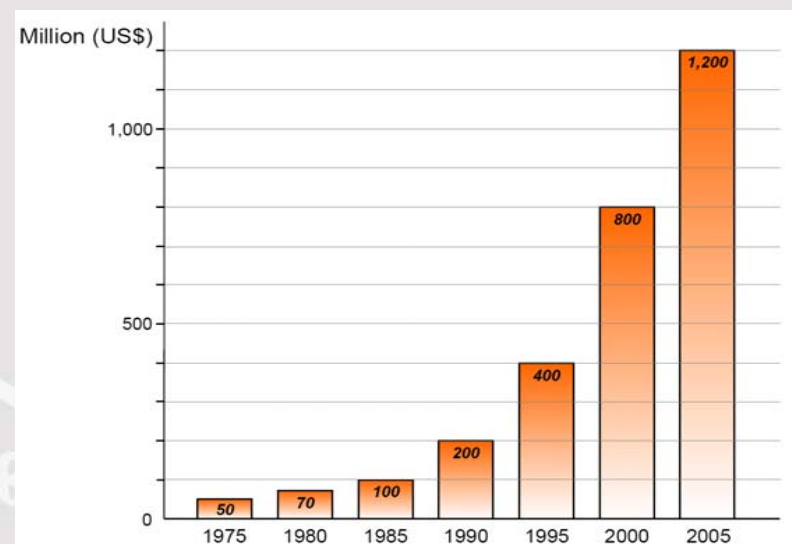




## Cost and development time of a new drug



*Time scale of new drug development.*

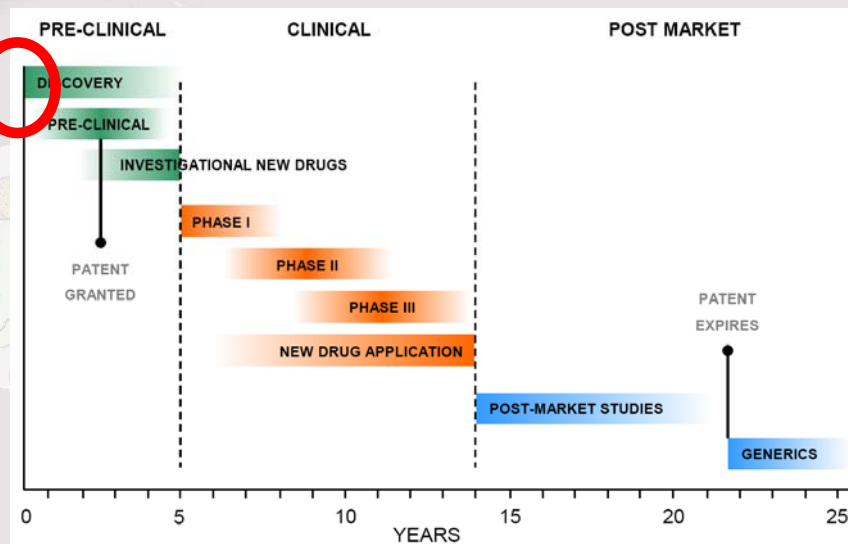


*Cost evolution of new drug development.*

**Nowadays it takes ~14 years and costs more than 1.2 billion US\$.**



## Cost and development time of a new drug



*Time scale of new drug development.*

## New solutions :

- Drug Design
- Virtual Screening

## Profusion of data...

Genome: 30,000 genes  
 Transcriptome: 40-100,000 mRNAs  
 Proteome: 100-400,000 proteins  
 Interactome: >1,000,000 interactions

## ...but also technology advances:

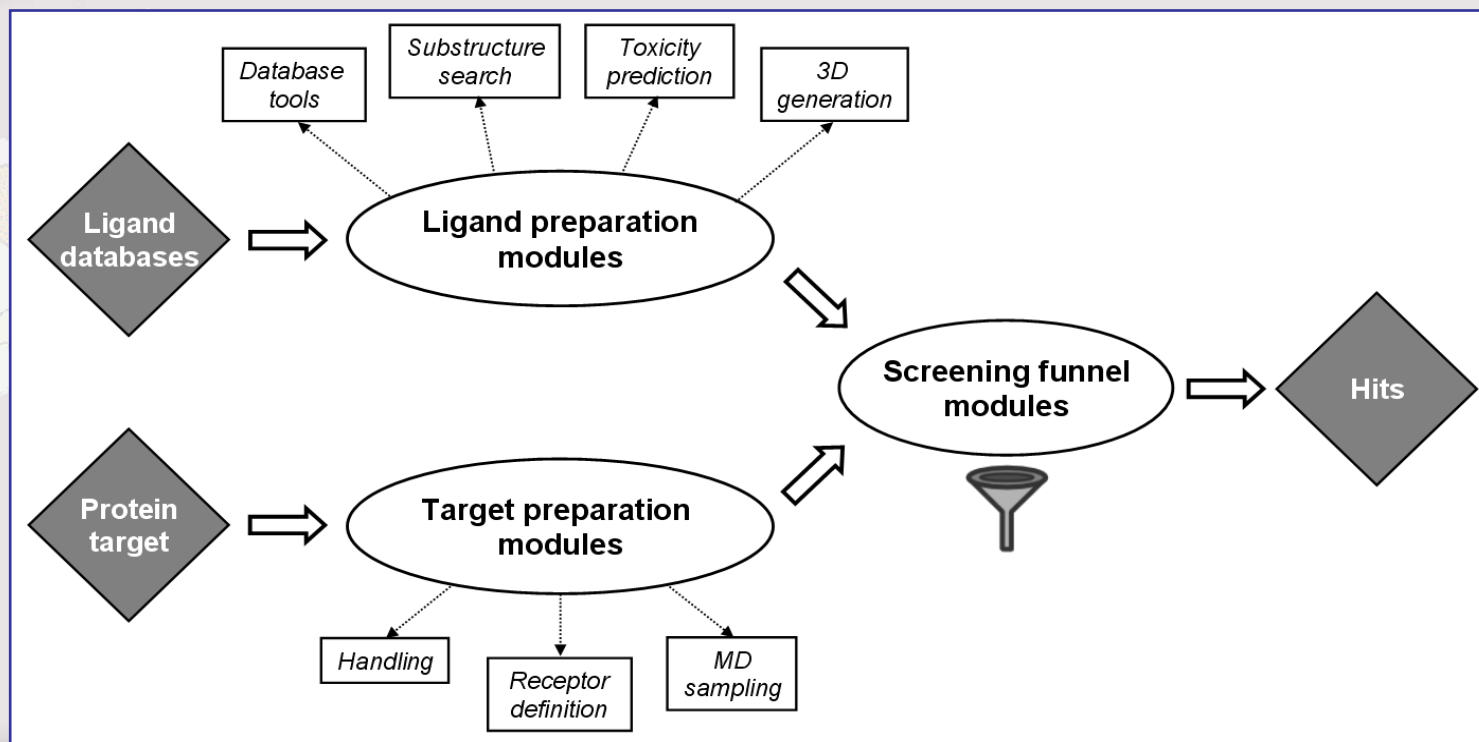
Computing power increase, fast networks, robots,...

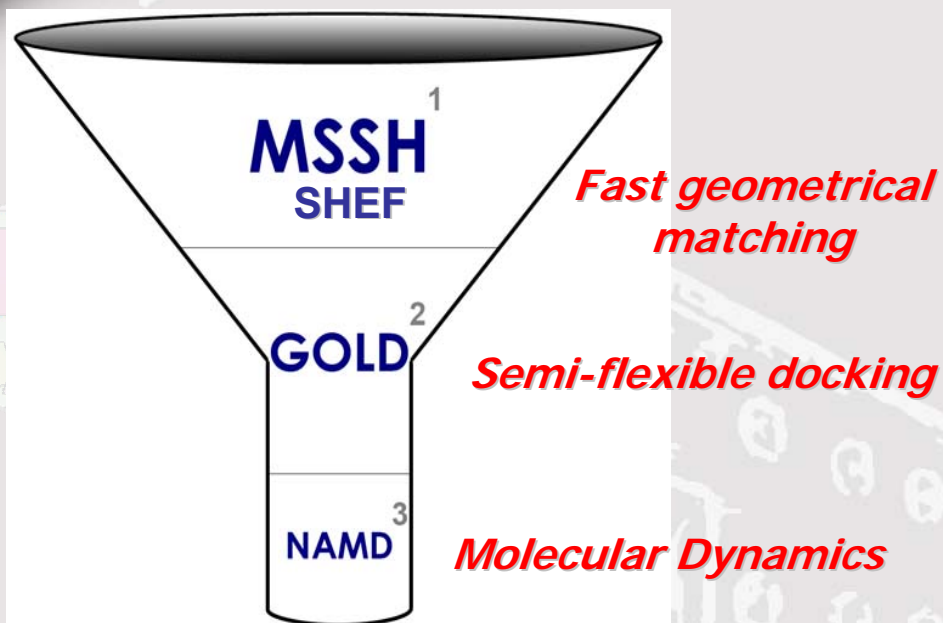






**VSM-G goal: ranks ligands in order to establish a limited set of hit compounds**





*Multi-step screening funnel*

### VSM-G funnel:

- Combination of several methods in the funnel allows both computational efficiency and high detection rate
- decrease the number of molecules after each step.

## Screening input preparation

**VSM - Ligand preparation**

Language

How would you like to create your molecular DB?

**CREATE**

**FILTER**

**HANDLE**

Scaffold file to use:

scaffold3.mol

List of fragment files:

R1.mol  
R2.mol  
R3.mol

Start

**VSM - Ligand preparation**

Language

How would you like to filter your molecular DB?

**CREATE**

**FILTER**

**HANDLE**

Previously created DB:

C:\VALEX\ZINC\_database\zinc.sdf

Sub-structure file generated using ISIS Draw:

N\_SO2\_Ph.mol

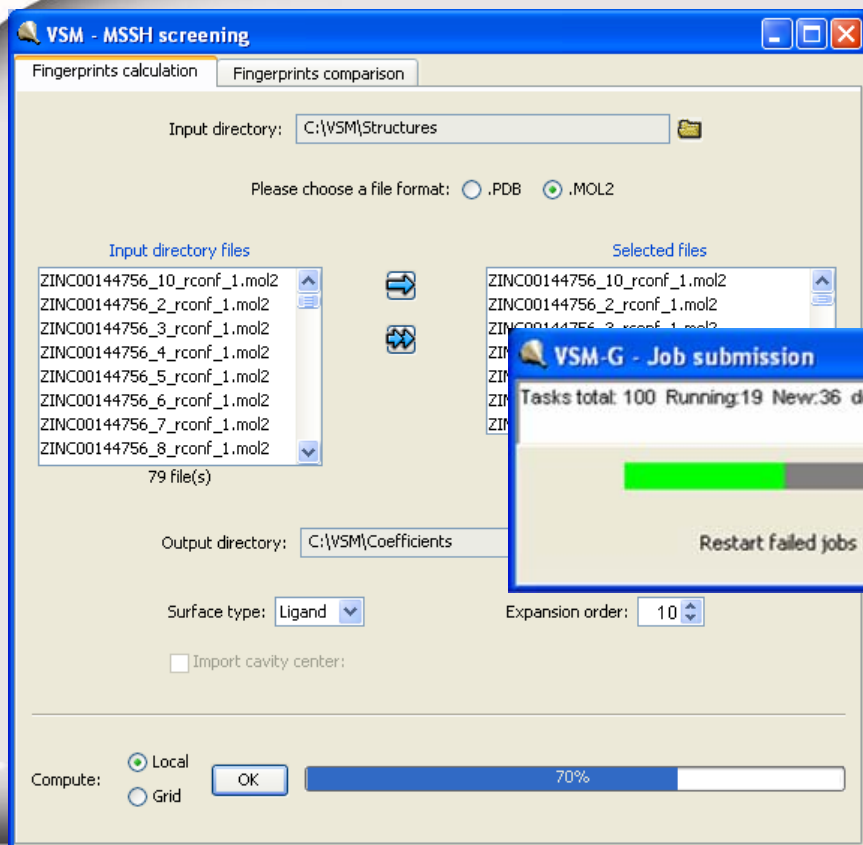
Sub-structure search

**Marvin: Base\_ZINC\_2D.sdf**

File Edit View

1	2	3
4	5	6
7	8	9

## Funnel interface Job supervision Filtering from results



**VSM - MSSH screening**

Fingerprints calculation | Fingerprints comparison

Input directory: C:\VSM\Structures

Please choose a file format:  .PDB  .MOL2

Input directory files: 79 file(s)

Selected files:

Output directory: C:\VSM\Coefficients

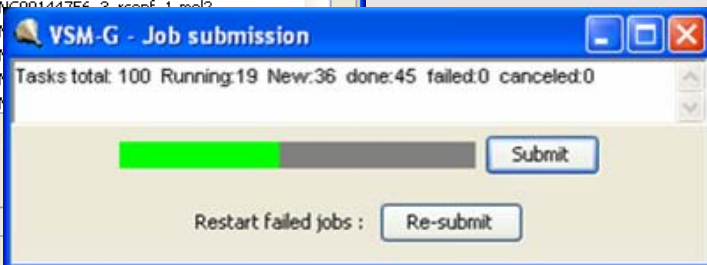
Surface type: Ligand

Expansion order: 10

Import cavity center:

Compute:  Local  Grid

OK [Progress bar: 70%]

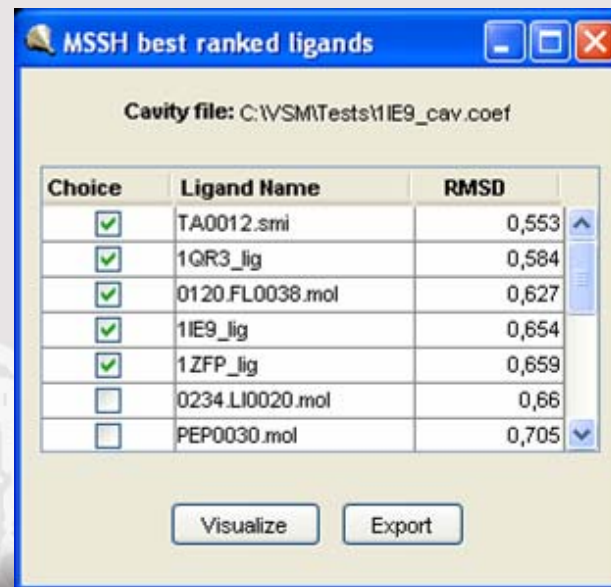


**VSM-G - Job submission**

Tasks total: 100 Running:19 New:36 done:45 failed:0 canceled:0

Submit

Restart failed jobs: Re-submit



**MSSH best ranked ligands**

Cavity file: C:\VSM\Tests\1IE9\_cav.coef

Choice	Ligand Name	RMSD
<input checked="" type="checkbox"/>	TA0012.smi	0,553
<input checked="" type="checkbox"/>	1QR3_lig	0,584
<input checked="" type="checkbox"/>	0120.FL0038.mol	0,627
<input checked="" type="checkbox"/>	1IE9_lig	0,654
<input checked="" type="checkbox"/>	1ZFP_lig	0,659
<input type="checkbox"/>	0234.LI0020.mol	0,66
<input type="checkbox"/>	PEP0030.mol	0,705

Visualize Export

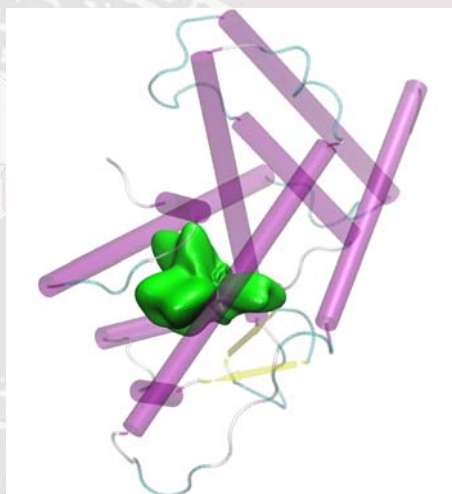
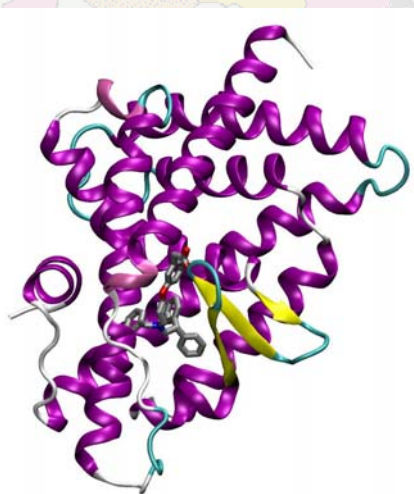




## Protein: LXR $\beta$ receptor

MD sampling ↓ Clustering

Set of representative target conformations



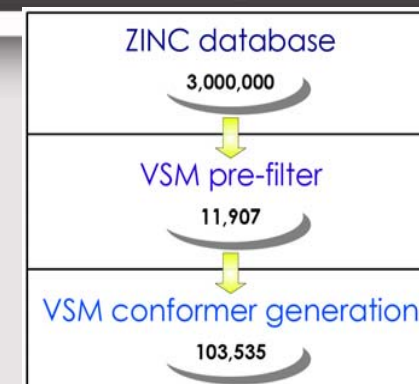
## Ligands:

➤ Starting database: ZINC

➤ Substructure-based pre-filter:  
*Phenyl-SO<sub>2</sub>-N* motif.

➤ Introduce a known active compound  
(GW3965) within the ligand database to check  
the validity of the approach

➤ Starting data :  
11,907 structures; 103,535 conformers.

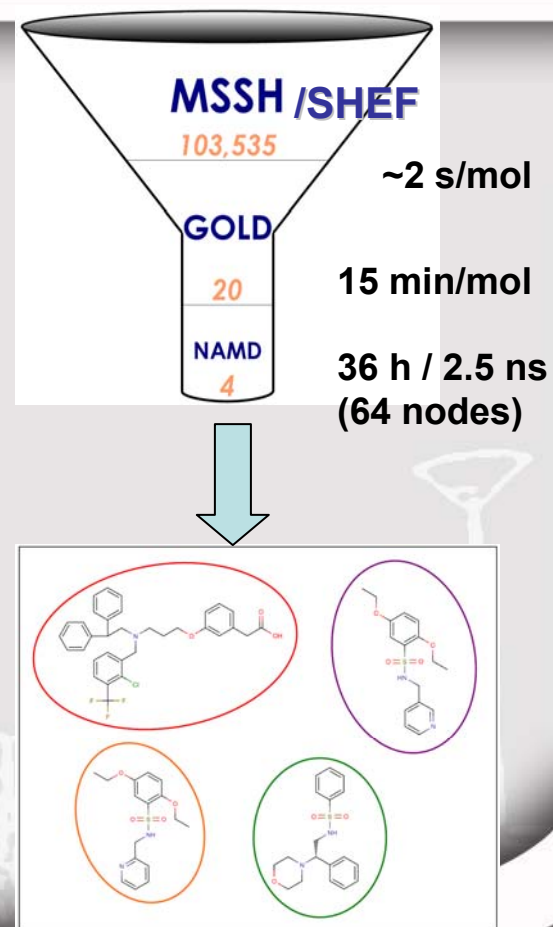




Reference compound **GW3965** has been extracted as a putative candidate with good rank at each level of the funnel:

- Top 10 after 1<sup>st</sup> filter (*shape complementarities*)
- #1 after 2<sup>nd</sup> filter (*semi-flexible docking*)
- Highest interaction energy after *MD refinements*

This rank would propose the **GW3965** molecule as the best candidate ligand for the LXR $\beta$  protein, with (computed) affinity far above other top compounds.



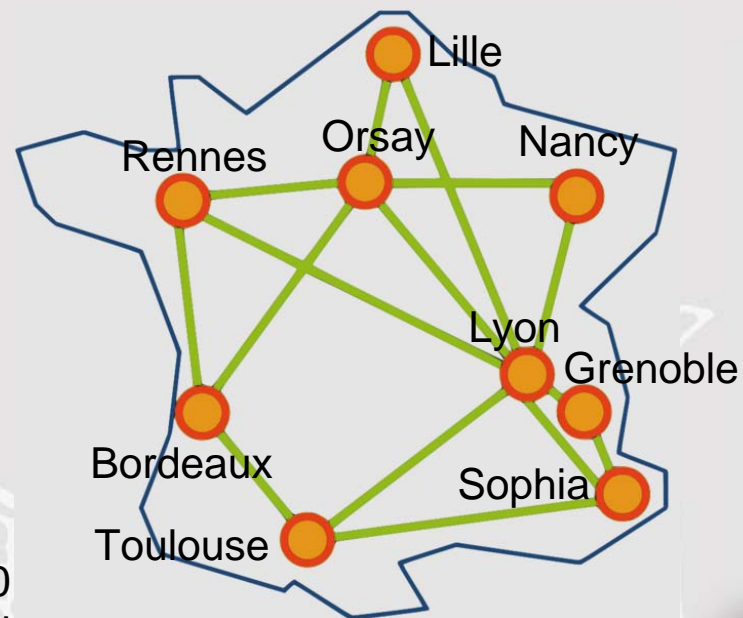


Use of grid computing allows large DB screening within a reasonable time.

**Grid'5000** ([www.grid5000.fr](http://www.grid5000.fr)) : 5000 CPUs for research in Grid Computing, eScience and Cyber-infrastructures

- We have used APST (A Parameter Sweep Tool) to schedule, distribute and manage calculations on this national-scale grid.

**Current objective:** Screening databases of ~1,000,000 molecules against ~10 protein conformations (experimental and from MD sampling)



France map with Grid'5000 sites.





1,000,000 molecules



Flexible docking  
(e.g. GOLD)

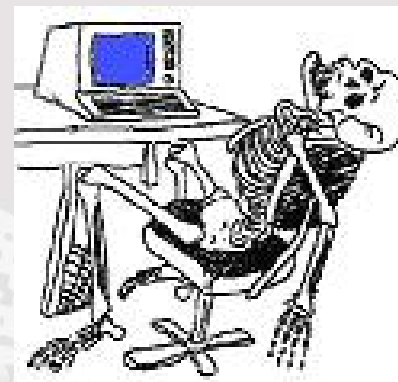


1,000,000 molecules



Flexible docking  
(e.g. GOLD)

**~ 1,000 years !**







1,000,000 molecules



**Screening funnel**  
(e.g. MSSH/SHEF  
then GOLD)

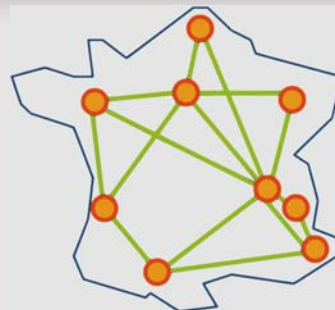
**5 years**





1,000,000 molecules

2 days



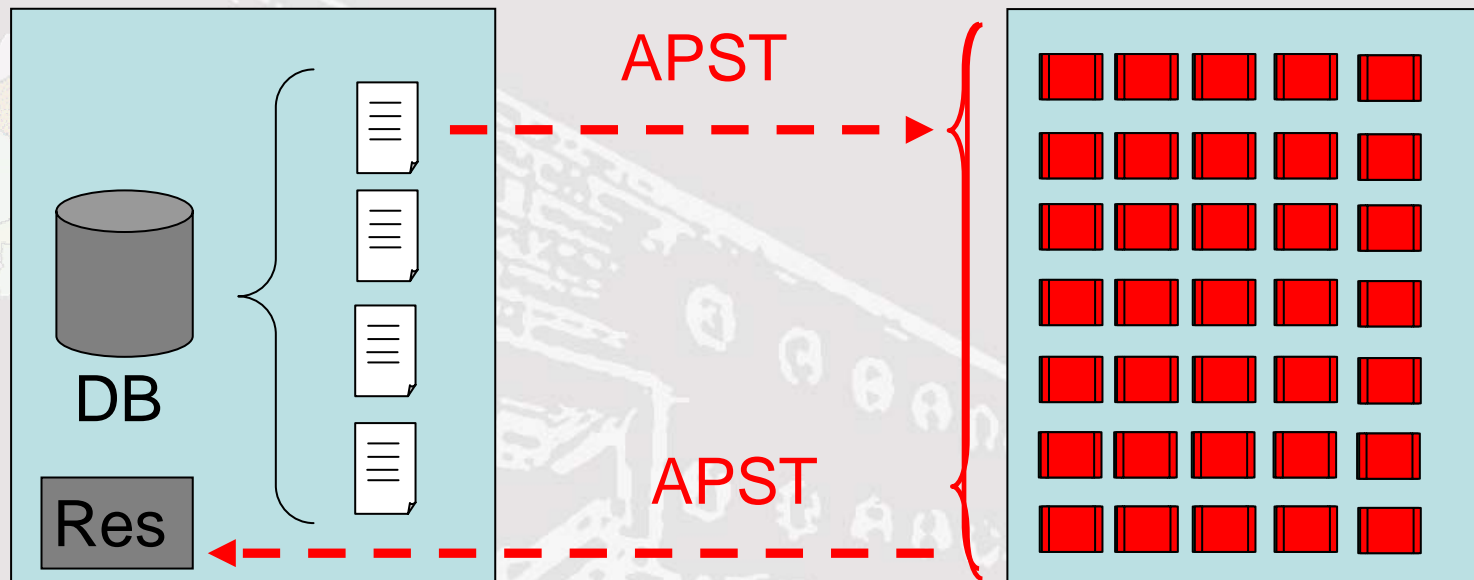
**Screening funnel**  
(e.g. MSSH/SHEF  
then GOLD)





## Distribution of job on Grid 5000 :

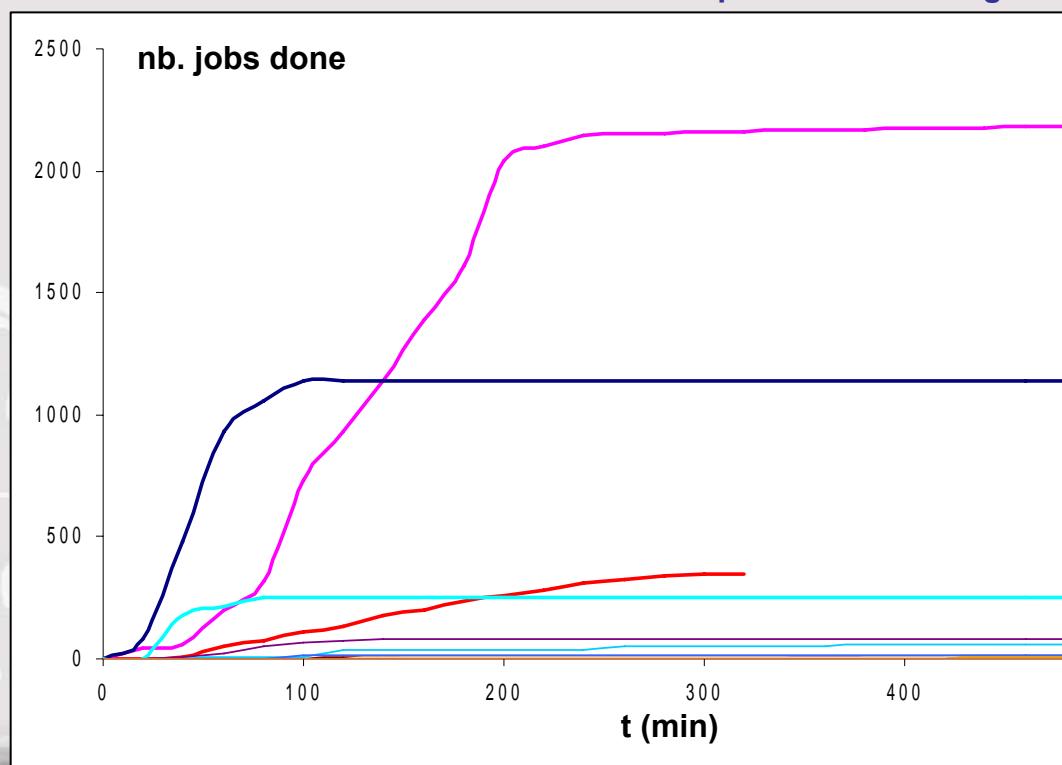
Local host:



Numbers of task carried out for each size of the input files according to time

By using these results for the jobs that present the best times we can compute:

- Average duration of each task : 29.4 s
- Total time to transfer one task input : 1.50 s
- Time to run program in one task: 27.9 s
- After the MSSH step we have run a SHEF filter on all molecules, selecting 10,000 for each cavity for the next screening funnel step (GOLD)





### VSM-G future developments:

- run GOLD on grid'5000 platform (at the present time we can run 20,000 GOLD in 14 hours on 1,200 processors)
- run NAMD on grid'5000 (10 NAMD jobs of 2.5 ns in one day on 32 processors)
- add possibility of QSAR modules in the funnel process
- replace commercial software by in-house or open-source ones
- enhance SHEF scoring function with physico-chemical features
- development of a relational database to store input data and VSM-G result
- *etc...*



## VHTS on TC80, a target against T. Cruzi trypanozoma

Biochem. J. (2005) 388, 29–38 (Printed in Great Britain)

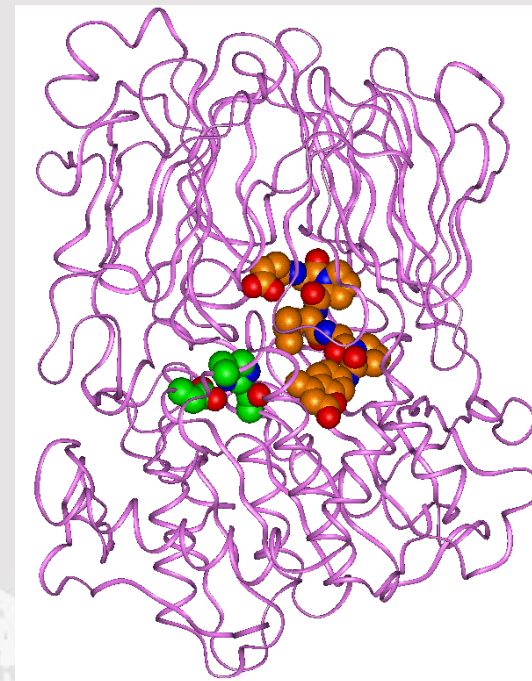


29

### Molecular, functional and structural properties of the prolyl oligopeptidase of *Trypanosoma cruzi* (POP Tc80), which is required for parasite entry into mammalian cells

Izabela M. D. BASTOS\*, Philippe GRELLIER†, Natalia F. MARTINS‡, Gloria CADAVID-RESTREPO\*, Marian R. DE SOUZA-AULT\*, Koen AUGUSTYNS§, Antonio R. L. TEIXEIRA\*, Joseph SCHRÉVEL†, Bernard MAIGRET||, José F. DA SILVEIRA¶ and Jaime M. SANTANA\*<sup>1</sup>

\*Laboratório Multidisciplinar de Pesquisa em Doença de Chagas (CP 04536), Universidade de Brasília, 70919-970, Brasília, DF, Brazil, †USM 0504, Département Régulations, Développement, Diversité Moléculaire, Muséum National d'Histoire Naturelle, 61 rue Buffon, 75231, Paris Cedex 05, France, ‡Embrapa, Genetic Resources and Biotechnology, CP 02372, Brasília, DF, Brazil, §Department of Medicinal Chemistry, The University of Antwerp, Belgium, ||Laboratoire de Chimie Théorique, Université de Nancy, 54506 Vandoeuvre-les-Nancy, France, and ¶Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina, R. Botucatu 862, CEP 04023-062, São Paulo, SP, Brazil



**In project :**

**Cooperation with Mali (Modibo Coulibaly et Seydou Doumbia) on Malaria kinases**



*Beutraït A., Leroux V., Chavent M., Souchet M., Cai W., Shao X., Moreau G., Bladon P., Devignes M.D., Smail-Tabone M., Yang G., Cavin X., Ray N., Elkhayar A., Suter F., Yao J., Liao Q., Yu F.*

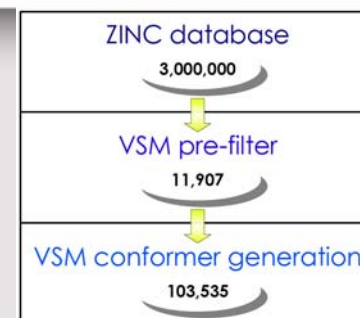
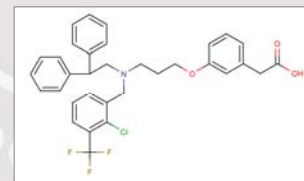
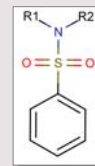
- Grid'5000 for providing access to grid computing resources.
- LORIA and Région Lorraine for financial support.



1. Cai, W.; Shao, X.; Maigret B. (2002). Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graph. Model.*, 20(4): p. 313-28.
2. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267: p. 727-748.
3. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16): p. 1781-1802.
4. Lala, D.S. (2005). The liver X receptors. *Curr. Opin. Investig. Drugs.*, 6(9): p. 934-43.
5. Irwin, J.J.; Shoichet, B.K. (2005). ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.*, 45(1): p. 177–182.

## Ligands:

- **Starting database: ZINC.**
  - 3 million commercially available compounds.
- **Pre-filter:** only compounds showing *Phenyl-SO<sub>2</sub>-N* motif.
  - 11,907 hits.
- **Introduce a known active compound** (*GW3965*) within the ligand database to check the validity of the approach.
- **Conformational analysis:** ~10 confs./compound.
  - 103,535 conformers.



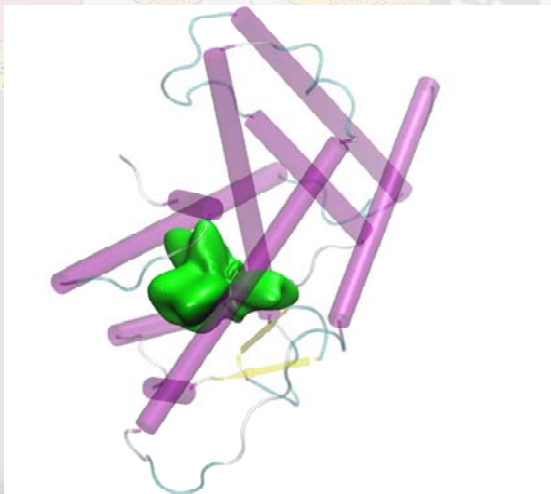


## Protein:

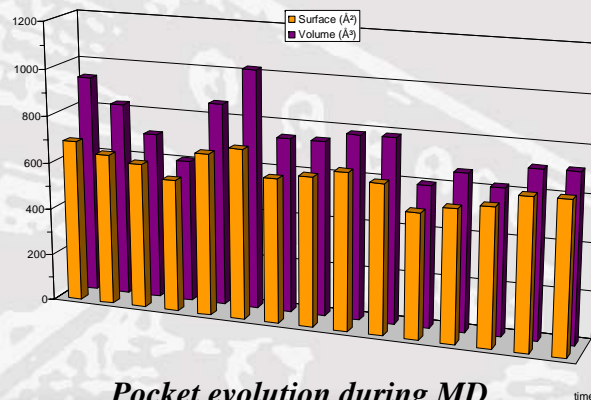
### ➤ MD sampling:

- Duration: 6 ns, CHARMM27 force field, NPT ensemble, 300 K, 1 bar, explicit TIP3P water.
- Snapshot extraction each 0.5 ns.

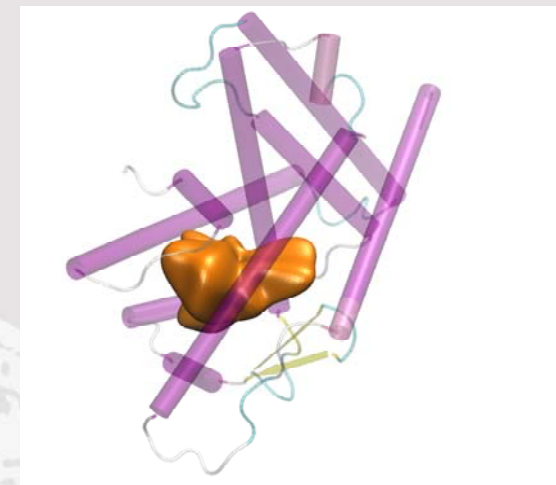
### ➤ Snapshot clustering to limit the targets number.



*Binding pocket at t=0.*



*Pocket evolution during MD sampling.*



*Binding pocket at t=6ns.*