Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Efficiency of a hierarchical protocol for high throughput structure-based virtual screening on GRID5000 cluster grid

This article was published in the following Dove Press journal:
Open Access Bioinformatics
22 May 2010
Number of times this article has been viewed

Leo Ghemtio[1]
Emmanuel Jeannot[2,3]
Bernard Maigret[1]

[1]LORIA, Groupe ORPAILLEUR, [2]LORIA, Groupe ALGORILLE, Campus scientifique, Nancy Université, Vandoeuvre-lès-Nancy Cedex, France; [3]LABRI, Groupe Runtime, Cours de la Libération, Bordeaux Université, Talence Cedex, France

Abstract: Most modern computational techniques in the drug discovery areas put demands on large computer resources. Grid computing offers a powerful alternative way of running computationally intensive applications. One field of the drug innovation process that can benefit greatly from the use of grid resources is the high-throughput virtual screening approach for docking huge chemical compound libraries into known protein-binding sites. The use of computational grids is the combination of computer resources from multiple administrative domains, heterogeneous, and geographically dispersed applications to a common task that requires a great number of computer-processing cycles or the need to process large amounts of data. This study detailed a screening campaign, on Grid5000 cluster grid computing infrastructure, concerning the ZINC database, from which a subset of ~600,000 "drug-like" molecules was extracted, against three structures of the liver-X receptor β (LXR β). A funnel strategy was used for that purpose, starting from a fast but simple shape matching procedure and achieved with more complex molecular dynamics simulations. From a total of ~91 million three-dimensional conformations which were generated at the beginning of the funnel and after intermediate filtering steps, the process ended with 45 putative hits. The GRID5000 is a highly reconfigurable, controllable, and monitorable experimental cluster grid, connecting nine sites geographically distributed in France, and featuring more than 3,200 processors and 5,700 cores. To hide the complexity of the grid system from the user, the GRID5000 has been used through the virtual screening manager for grid computing (VSM-G) platform, dedicated to *in silico* screening and to provide maximum computing power by using grid resources efficiently. The whole screening process required around 82 days (78 days of pre-processing and 3.6 days for the docking funnel itself) and utilized 3,144 nodes over nine sites. The use of grid infrastructures and hierarchical filtering protocol enable us to perform evaluations of the binding capabilities of millions of compounds on several conformations of a given target and propose that, with a low cost, most promising compounds for *in vitro* tests.

Keywords: high-throughput virtual screening, molecular filtering, docking, liver-X receptors, grid computing, molecular dynamics simulation

Introduction

Drug discovery is a very expensive process. There are now millions of chemicals which can be tested by *in vitro* high-throughput screening (HTS) to determine their ability to bind to, inhibit, or activate biomolecular targets.[1–3] Because of the cost of large HTS campaigns, many researchers have directed their efforts to developing computational tools able to perform virtual screening (VS) prior to HTS.[3–5] The cost of performing *in silico* screens is less than that of HTS methods. Thus structure-based VS approaches, using molecular docking engines, may provide the key to limit the huge number of

Correspondence: Bernard Maigret
Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239 54506,Vandoeuvre-lès-Nancy Cedex, France
Email bernard.maigret@loria.fr

compounds to be evaluated by HTS to a smaller subset that is more likely to yield putative "hits".[5–7] Structure-based virtual screening uses knowledge of the target protein's structure to select candidate compounds with which it is likely to favourably interact. Even when the structure of the target is known, the design of a molecule capable of binding to it is a daunting challenge, as the complexity of the problem is in reality far greater than the simple lock-and-key picture. For example, the conformations of both the ligand and the receptor may change during the process of binding. In addition, the thermodynamics of the binding process need to be taken into account.[8,9] When applied at a high-throughput level, structure-based VS requires sufficient computing power to handle the docking of millions of potential ligands into hundreds of target conformations.[10–12] Despite the decreasing cost and increasing speed of computing hardware, so that a single docking calculation (one ligand conformation into one target) can take as little as two minutes on a personal computer, the total time needed to perform 100,000,000 calculations would be 381 years! Moreover, as screening millions of molecules on different geographical sites requires a high communication demand for data exchange, a computational data challenge is also emerging.

Access to very large computing resources is therefore needed for successful high-throughput virtual screening campaigns.[13–16] Basically, there are two approaches to speed up the structure-based high-throughput virtual screening (HTVS) process, and to significantly reduce the constraint of required time and resources addressing the performance bottleneck.[1,3,4,10–12,17,18] The first is to run the calculations on massively parallel supercomputers. This requires the docking programs to be modified; the code is parallelized and optimized for the particular machine configuration.[19,20] The second approach is to spread the calculations on an array of processors (a grid infrastructure). This allows the docking codes to be kept as they are.[14–16,21,22]

Since dedicated parallel computers are more difficult to access, grid systems are now a more popular way to cope with a growing requirement for high computational resources and to speed-up the drug development process.[21–25] Moreover, the flexibility of grid systems allows them to be configured as parallel computers. Parallel programming runtimes can be used to develop parallel applications that utilize the collection of processors simultaneously. This high performance computing or parallel processing approach can be used to substantially reduce the runtime of large applications such as molecular dynamics (MD) for which subtasks need to communicate many times per second.[26–28] Alternatively a batch-scheduling system can be employed to combine the computing power of hundreds or thousands of computing nodes together for individual sequential calculations such as fast rigid docking.[29] This high-throughput computing or embarrassingly parallel approach is done by running many independent jobs simultaneously on multiple processors for which subtasks never have to communicate.[30,31]

Our use of both types of grid tasks (sequential and parallel) is described in this paper in the framework of the GRID5000 project.[32,33] To identify successfully, among a large chemical library, a subset of compounds significantly enriched in hits, the challenge is therefore to develop a selection pipeline which uses the benefits of the two approaches above to speed-up and reduce constraint, time, and resources in the structure-based HTVS experience.

Thus the selection pipeline is built up with a structure-based program ranging from crude-but-fast matching procedures, able to consider millions of putative ligands and using an embarrassingly parallel approach with cluster grid computing, to accurate-but-slow methods working only on reduced subsets with parallel processing. We have used the sequential approach for the preliminary stages of structure-based virtual screening of the ZINC database against three structures of the liver-X receptor β (LXR β). These two stages were the fast geometrical matching molecular surface spherical harmonic (MSSH)[29]/spherical harmonic coefficient filter (SHEF)[34] procedure, followed by the GOLD[35] flexible docking program. We used the parallel approach in the final NAMD[36] MD simulation of the best protein and ligand structures.

Material and methods

Grid'5000 infrastructure

Grid'5000 (Grid5000) is a French scientific instrument devoted to the study of large scale parallel and distributed systems. It aims at providing a highly reconfigurable, controllable, and monitorable experimental platform to its users. The initial aim to reach 5,000 computing cores in the platform was reached during winter 2008–2009. The infrastructure of Grid5000 is geographically distributed on different sites hosting the instrument, initially nine in France. Porto Alegre, Brazil is now officially the 10th site. Figure 1 presents an overview of Grid5000. Every site hosts one or several clusters and all sites are connected by high speed network (10 Gbps since 2007). Figure 1 gives the number of CPUs for every cluster. Two-thirds of the nodes are dual CPU 1U racks equipped with two AMD Opteron (AMD, Paris, France) running at 2 GHz, 2 GG of memory and two 1 Gpbs ethernet

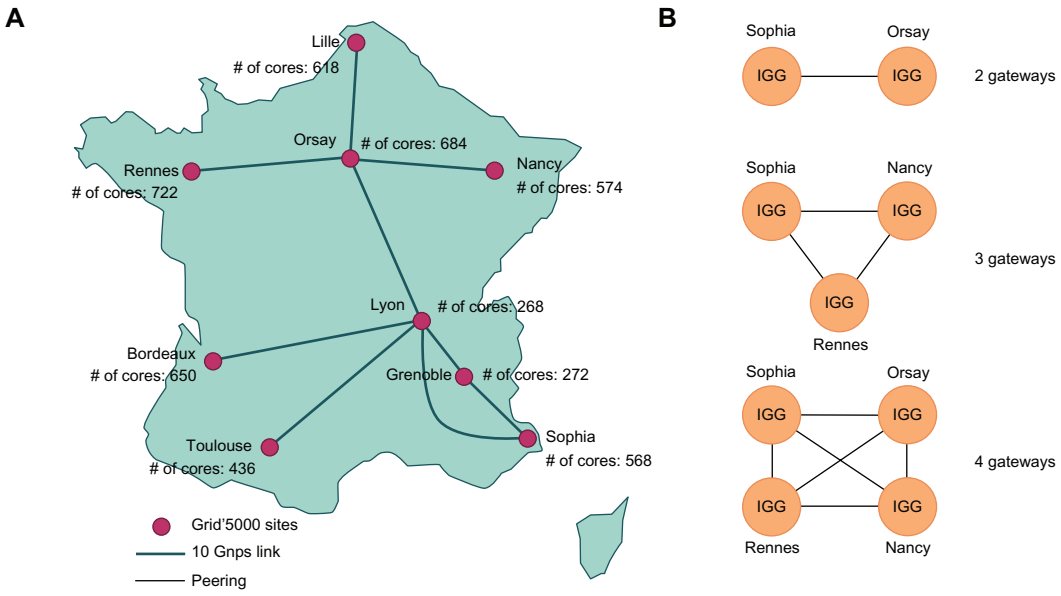


Figure 1 Overview of Grid5000, showing the Grid5000 sites as well as the gateway configurations.[68]

adapters (Xerox Corporation, Norwalk, CT). The associated clusters are also equipped with high-speed network connections (Myrinet, InfiniBand, etc). To hide the complexity of the implementation, a portal is layered above the grid infrastructure and the end-users. The portal is configured so that the end-user will see only the resources allocated plus information on the running of the jobs.

Implementation of VSM-G on the Grid5000 grid

The portal management was provided by a parameter sweep tool (APST).[37,38] The APST was embedded in our VSM platform dedicated to HTVS and was used to build a multi-cluster environment, connected through the high speed network, from which a user can allocate a number of processors and utilize them shortly before releasing them back to the resource pool.[37] Such a portal must interface with pieces of software that efficiently control the backend system.

The VSM platform consisted of pre-processing engines for screening both protein targets and small molecules, and of a funnel-based docking strategy (Figure 2). At each step of the funnel, depending on tuning, a proportion of inappropriate molecules were discarded. The system aimed to start with several million compounds along with several hundred target conformations, eventually yielding a small set of putative hit compounds.

Alongside this we have developed monitoring Java applets for checking the available Grid5000 resources, allowing a user to make the appropriate reservations, to submit jobs, to parse the APST logs, to select the files describing candidate molecules, and to provide information on the progress of the computations. Other applets applied check to avoid duplicate calculations and, in cases where calculations had failed, organize repeat calculations. An important feature of our system is fault tolerance. Thanks to APST, when a processor fails, is reclaimed by a higher priority task or is returned to the pool of accessible processors. Because its reservation has finished, tasks assigned to this processor are not lost and are automatically executed to other resources. When the screening of all the molecules was complete, a final applet retrieved the result files. An example of how APST/Grid5000 calculations are managed is presented in Figure 3.

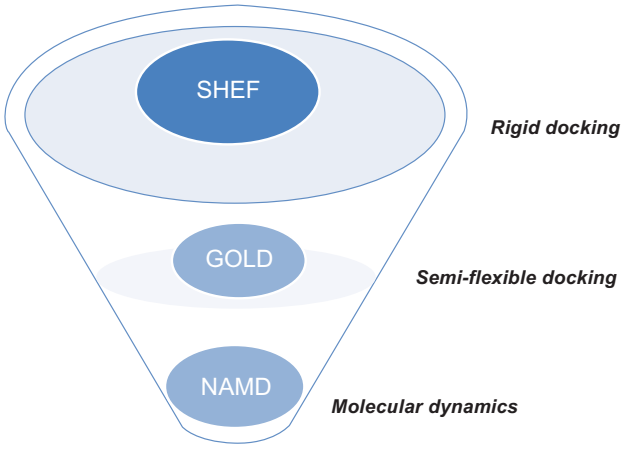


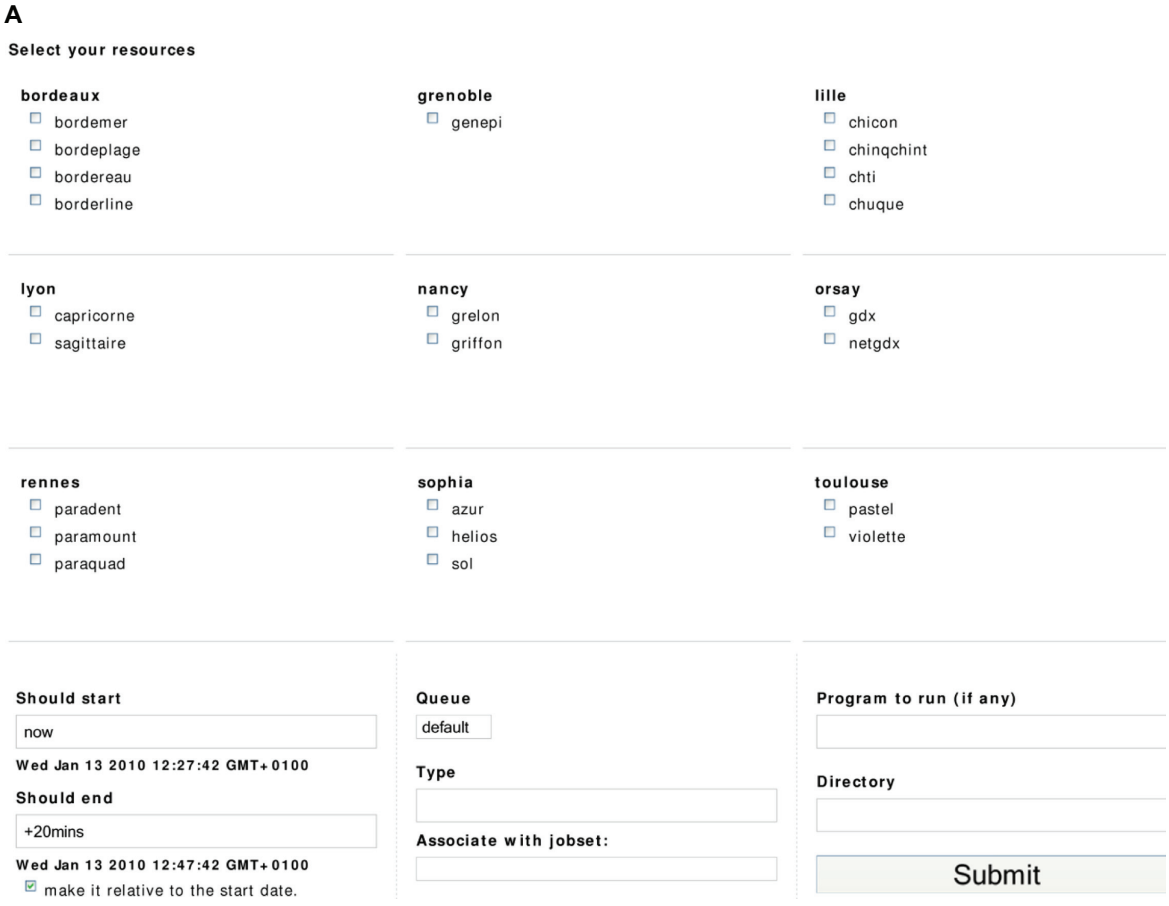
Figure 2 VSM multiple steps screening funnel

The APST daemon, SHEF, GOLD, and NAMD applications were installed on all nine Grid5000 sites.

Data preparation

Target structures

Liver-X receptors are interesting targets because their therapeutic interest is well acknowledged, especially in cardiovascular diseases. This work specifically studies the LXR β isoform of liver-X receptors because several X-ray structures are available in the Protein DataBank (PDB):[39] 1P8D,[40] 1PQ6,[41] and 1PQ9,[41] providing some highlight about the protein binding site flexibility which has been studied in detail.[40–42] The differences between these structures reveal a large plasticity of the ligand-binding pocket to accommodate



B

Job details:

Id	User	State	Queue	NbNodes	NbCores	Type	Properties	Reservation	Walltime	Submission Time	Start Time	Scheduled Start
252473	bguillon	Running	default	3	12	INTERACTIVE	deploy = 'YES'	Scheduled	5:59:57	2010-01-13 11:03:21	2010-01-13 11:03:23	2010-01-13 11:03:23
252762	probert	Running	default	175	1000	INTERACTIVE		None	2:0:0	2010-01-13 13:09:32	2010-01-13 13:09:35	2010-01-13 13:09:35
252766	probert	Running	default	6	20	PASSIVE		None	2:0:0	2010-01-13 13:13:28	2010-01-13 13:13:29	2010-01-13 13:13:29
252767	probert	Running	default	3	20	PASSIVE		None	2:0:0	2010-01-13 13:13:34	2010-01-13 13:13:35	2010-01-13 13:13:35
252768	probert	Running	default	4	20	PASSIVE		None	2:0:0	2010-01-13 13:13:37	2010-01-13 13:13:43	2010-01-13 13:13:43
252769	probert	Running	default	3	20	PASSIVE		None	2:0:0	2010-01-13 13:13:39	2010-01-13 13:13:43	2010-01-13 13:13:43
252770	probert	Running	default	4	20	PASSIVE		None	2:0:0	2010-01-13 13:13:41	2010-01-13 13:13:43	2010-01-13 13:13:43
252771	probert	Running	default	3	20	PASSIVE		None	2:0:0	2010-01-13 13:13:43	2010-01-13 13:13:52	2010-01-13 13:13:52
252775	probert	Running	default	1	2	PASSIVE		None	2:0:0	2010-01-13 13:15:15	2010-01-13 13:32:31	2010-01-13 13:32:31
252776	probert	Running	default	2	2	PASSIVE		None	2:0:0	2010-01-13 13:16:26	2010-01-13 13:32:31	2010-01-13 13:32:31
252777	rpottier	Running	default	2	8	PASSIVE	deploy = 'YES'	None	4:0:0	2010-01-13 13:18:32	2010-01-13 13:36:28	2010-01-13 13:36:28
252778	rpottier	Running	default	1	4	PASSIVE	deploy = 'YES'	None	4:0:0	2010-01-13 13:18:33	2010-01-13 13:36:19	2010-01-13 13:36:19
252780	ejeannot	Waiting	default	10	0	INTERACTIVE	cluster="griffon"	None	10:0:0	2010-01-13 13:46:40		2010-01-13 15:09:37
252783	acherif	Waiting	default	1	0	INTERACTIVE		None	2:0:0	2010-01-13 14:09:25		2010-01-13 15:09:37

C

Grid5000 Nancy OAR nodes

Summary:

OAR node status	Free	Busy	Total
Nodes	2	198	259
Cores	8	1148	1310

Reservations:

grelon-1	Down				grelon-2	252762	252762	252762	252762	grelon-3	252473	252473
grelon-4	252762	252762	252762	252762	grelon-5	252762	252762	252762	252762	grelon-6	252762	252762
grelon-7	252762	252762	252762	252762	grelon-8	252762	252762	252762	252762	grelon-9	252762	252762
grelon-10	252762	252762	252762	252762	grelon-11	252762	252762	252762	252762	grelon-12	252762	252762
grelon-13	252762	252762	252762	252762	grelon-14	252762	252762	252762	252762	grelon-15	252762	252762
grelon-16	252762	252762	252762	252762	grelon-17	252762	252762	252762	252762	grelon-18	252762	252762
grelon-19	252762	252762	252762	252762	grelon-20	252762	252762	252762	252762	grelon-21	252762	252762
grelon-22	252762	252762	252762	252762	grelon-23	252473	252473	252473	252473	grelon-24	252473	252473

Figure 3 Management of Grid5000 calculations.[33] **A)** Interface to select the grid computing resources for job submission. **B)** Details of job executed on reserved Grid5000 resources. **C)** Details of nodes reserved on Grid5000 resources.

compounds with noticeably different shapes and sizes.[41] For each of these X-ray structures the most complete chain was retained (chain A for 1P8D, and chain B for 1PQ6 and 1PQ9) and eventually completed in order to be used in the docking calculation (this preparative procedure has already been described).[43]

Ligand databases

The ligands used in this study are composed of molecules present in the ZINC[44] database that were commercially available in March 2006 from three suppliers, ChemDiv,[45] Enamine,[46] and Comgenex.[47] We used this molecular database as it has been already used as a reference set in proof-of-concept studies of virtual screening techniques.[17] A preliminary filtering procedure was performed on the whole ZINC library, using Lipinski's rule-of-five[48] allowing a single violation for each structure. This gave a total of 598,375 unique molecules that was stored into the VSM-G working database. In order to take ligand flexibility into account, conformational sampling was performed for each compound in this working database using Omega software (version 2.2.1; OpenEye Scientific Software Inc., Santa Fe, NM). A maximum of 400 conformers were allowed for each compound giving an average of about 150 conformers by compound when considering the whole set of compounds. This produced the VSM-G conformer database containing around 91 million three-dimensional (3D) structures.

Screening process

For each protein LXR β structure, a virtual screening experiment was performed according to the three steps constituting our selection funnel (Figure 4). A first screen of the database containing the 91 million conformer 3D structures as selected above was done with the SHEF docking method. Prior to this step, the molecular surface of the compounds and the cavity surface of the protein active sites were calculated with the MSSH program. Next, the SHEF docking results were collected, ranked according to their score, and a new, much smaller, database was built consisting of a user-defined top percentage of the ranked compounds. Briefly, each 3D structure of the conformer database was submitted to the MSSH procedure in order to describe its shape with the spherical harmonics expansion coefficients. The surface integration for calculating the expansion coefficients is computationally expensive. Representation of the molecular surfaces of a target binding site and ligand by their expansion coefficients allows a shape comparison between the two surfaces to be achieved. For this purpose, considering the surface of the target as rigid and fixed, the coefficients of the ligand molecule are rotated in order to obtain the minimal root-mean-square distance of these coefficients and those of the target by SHEF. This was achieved within the VSM-G platform. According to our experience, 8,000 SHEF-score top-ranked molecules were therefore retained for each 1PQ6, 1PQ9, and 1P8D protein structure to be passed to the next filter. This number of 8,000 was retained according to previous studies as containing most of the molecules of interest.[17]

This new set of putative ligands was screened using the semi-flexible GOLD docking method which performs more accurate, but more computationally expensive dockings. The molecules and the LXR β target binding sites were

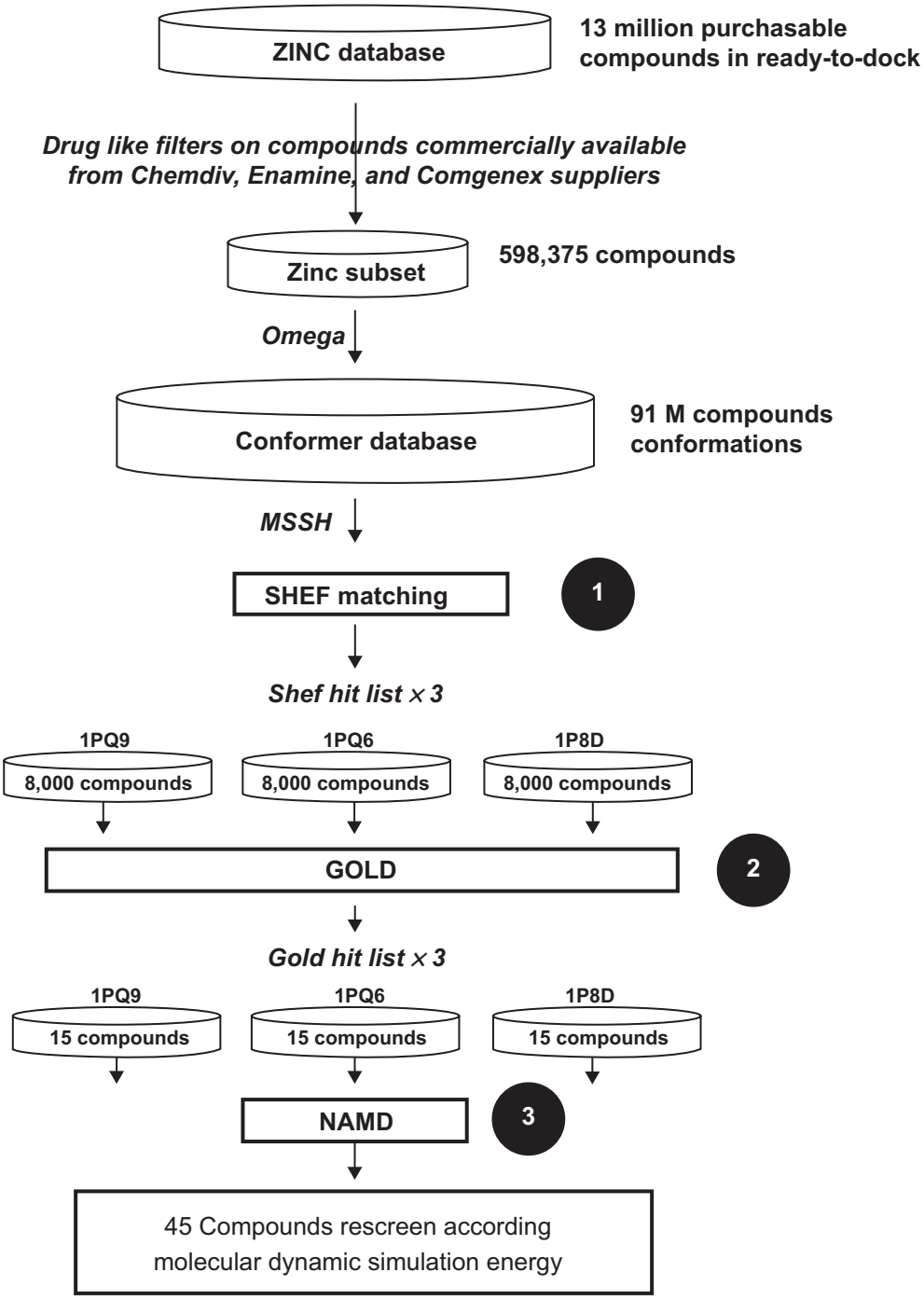


Figure 4 Overview of VSM-G implementation coupled with APST on Grid5000 platform.

prepared with the VSM routines, 50 docking trials were performed on each protein target for each compound, and the one giving the best score was retained. The docking results were collected from each of these additional GOLD screenings and used to build a new database consisting of the GoldScore scoring function ranked compounds. From them, the top 15 structures for each protein target were selected for the next filter.[35]

Finally, MD simulations were used to re-rank these 15 molecules on each target. At this stage each protein/ligand complex was embedded within an explicit water box of 80Å^3 size. The NAMD inputs were similar to those already used in our paper concerning induced fit phenomena in LXR β complexes[43] and were prepared using an interface with the visual VMD[49] program. For each one of these 45 complexes (15 molecules × 3 protein structures), 10 ns of MD

simulation were recorded after the necessary minimization (6,400 conjugate gradients) and equilibration (100 ps) steps. Each set of 10,000 frames recorded from the MD runs for each complex was further analyzed to check the intermolecular interactions energy using the "pair-interactions" feature of NAMD.[50]

From this point, a consensus scoring of the results from all three filters ranked all candidates according to the average predicted value given by each filters or "rank–by-number" was used to determine which compounds had the predicted highest binding affinity for the protein's binding site. This method, which combines multiple scoring functions in binding affinity estimation, leads to higher hit-rates in virtual screening.[51] These compounds will be tested in the laboratory.

Results

Our results for the three stages are summarized in Table 1 and detailed below.

MSSH preliminary calculations for SHEF

The average time needed to calculate the spherical harmonic surface for a single conformer was two seconds. Hence the total processor time for 91,063,822 conformers would be ~6 years. It is necessary to find the optimal number of conformers to be sent to each of the 1,200 nodes allocated in the Grid5000 cluster. This was found to be 200 after several trials (Figure 5).

The first step before starting the Grid5000 experiment was to set the best number of conformers to be sent in a single task on a given cluster in order to achieve the best balance between the IO and CPU times. For that purpose, several sizes of IO were tested and the optimal value was found to be around 200 conformers per MSSH run (Figures 5a and 5b). The packages (200 conformers) that present the best ratio were used to screen the three targets with all 91,063,822 ligand conformers. Overall computing time and nodes used are described in Table 1. For all three LXR β targets the total Grid5000 time was 78 days, with 73 days lost in data transfer, compared to 93,000 days without Grid5000.

These results highlight different problems during the MSSH simulation such as the saturation of the bandwidth, the too-long time used for file transfer, and the poor distribution of tasks. But this task has to be done only one time as the spherical harmonic surfaces for all these compounds can be reused later for another search on other targets.

SHEF filter

A single SHEF run for one conformer versus one LXR β cavity was only 0.014 seconds. Bearing this in mind, together with the file transfer problems encountered with the MSSH runs on the whole grid, we have run all the 91,063,822 SHEF calculations with one LXR β target calculation on only one cluster of the Grid5000 grid using 64 nodes where all the data was stored locally. The results are shown in Table 1. For all three LXR β targets the total Grid5000 time was ~17 hours with no time lost in data transfer, compared to 45 days without Grid5000.

GOLD filter

The next step was to run the GOLD program on the limited subset of 8,000 molecules appearing as putative suitable candidates after the SHEF calculations on one LXR β cavity. The molecules and the LXR β binding site for one target were prepared with the VSM routine. For one molecule, using 50 possible attempts, the average processor time of GOLD execution was around 12 minutes, thus the time for running the total subset of molecules would be 67 days on a single processor. For the 8 hours reservation we had, the experiment used six sites and 600 nodes. The results obtained using 50 molecules/run are detailed in Table 1. All calcula-

Table 1 Computing time for each layer of VSM-G platform with three LXR β as the target

	Number of compounds	Granularity of program	Optimal number of compounds in one task	Total time for one task	Total time to transfer input for one task	Number of nodes used	Estimated sequential time	Total time used with Grid5000	Total time lost in transfer files
MSSH	91,063,822	2 s	200	29.4 s	27.4 s	1,200	93,000 days	78 days	73 days
SHEF	91,063,822	0.014 s	200	0.014 s	0	64	45 days	17 h	0
GOLD	24,000	12.58 min	1	13 min	2 s	600	217 days	9 h	80 sec
NAMD	45 (9 × 5)	360 min	1	362 min	2 min	1,280 (5 clusters of 256 nodes each)	12 days (1 cluster of 256 nodes)	60 h	90 min

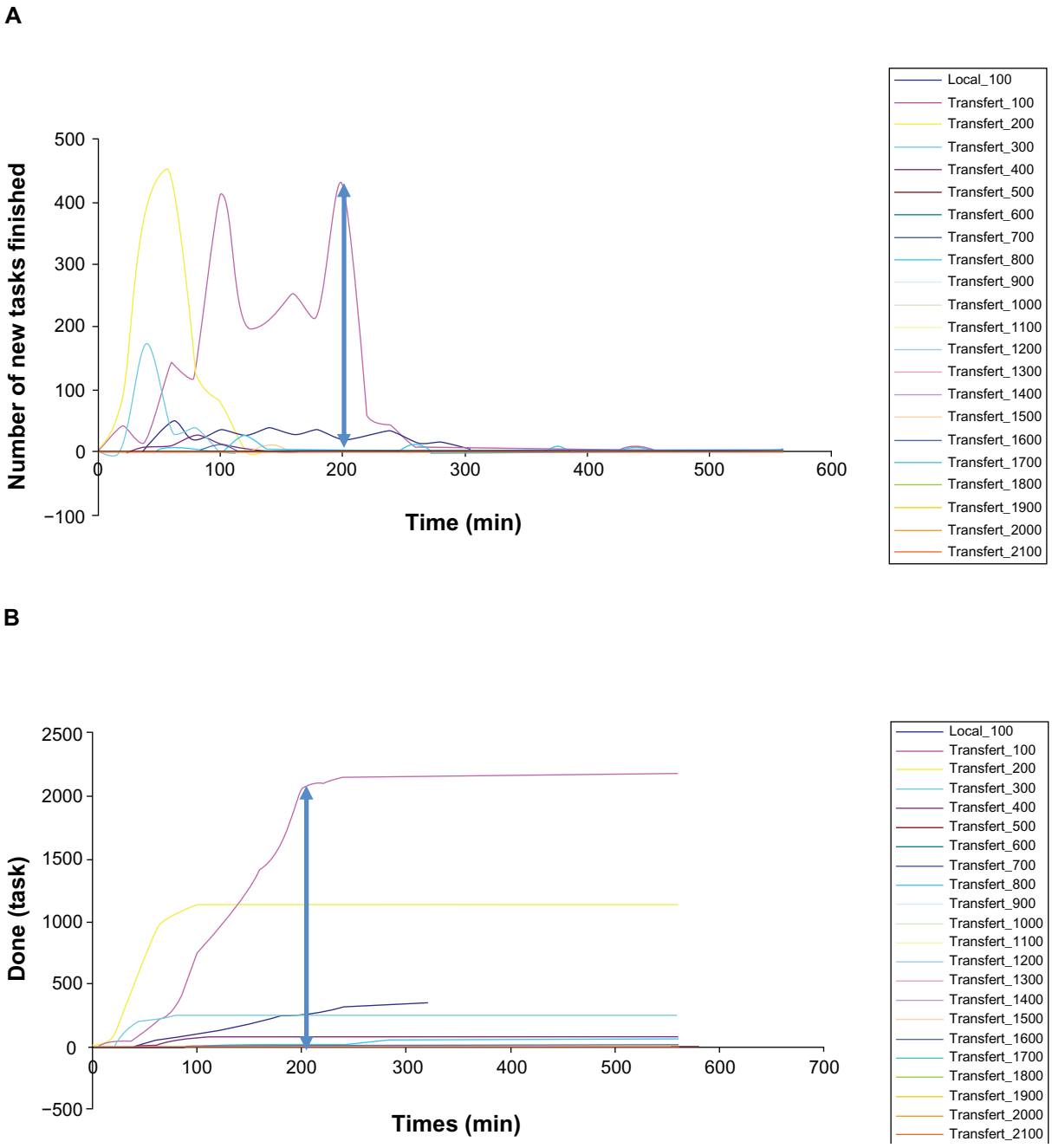


Figure 5 Identification of the best sized input files for MSSH computing on grid. **A)** Throughput of task for each size of input files by computing time. **B)** Number of tasks carried out for each size of the input files according to time.

tions on the three LXR β targets give a total Grid5000 time of 9 hours, compared to 216 days on a single processor. Only 80 seconds were lost in data transfer.

NAMD final refinements

From the 8000 molecules docked with GOLD, we selected the 45 best ranked compounds and their corresponding poses within the LXR β binding site (15 best molecules for each PQ6, PQ9, and P8D protein conformations) to be submitted to a MD refinement. For each one of these 45 complexes, 10 ns MD simulations were recorded after the necessary minimization (6,400 conjugate gradients) and equilibration (100 ps) steps. These 45 MD simulations were spread across five Grid5000 sites with nine tasks/clusters using 1,280 nodes (256 nodes on each cluster). The procedure is similar to that used by other workers.[15,52–59] We took advantage of the InfiniBand (InfiniBand Trade Association, Beaverton, OR) capabilities of the Grid5000 clusters providing 100%

efficiency for each node on which NAMD was running. The results of these calculations are shown in Table 1. The entire MD calculations for all three LXR β targets was achieved on the Grid5000 cluster grid in a total time of 2.5 days with 90 minutes lost in data transfer. The same 45 calculations, using only one Grid5000 cluster with 256 nodes, would have taken 12 days and, on a smaller cluster of 32 nodes without InfiniBand, 244 days.

From all these funnel calculations, and according to the protein/ligand intermolecular interaction energy plots obtained after the MD calculations (giving an estimate of the binding capabilities of the considered ligands), five compounds gave very good results, similar or better to the ones obtained for known effective ligands (such as the tularic compound) in previous calculations on the same system.[43] The best molecule was obtained from the Enamine provider under ID T5540452 and its protein/ligand interaction plot is presented in Figure 6. The superposition of T5540452 compound with the tularic (44B) crystallographic ligand from LXR β RX structure is shown in Figure 7. It shows clearly that the T5540452 compound binds in the same place as the crystallographic ligand (44B) and fits better within the binding pocket of protein. These molecules will be used now, at a moderate cost, for *in vitro* tests.

Discussion

The use of grid infrastructures is now widely used for performing high-throughput "classical" docking calculations, making possible the evaluation of the binding capabilities of millions of compounds on a given target.[15,56] But, as the number of chemicals to be tested continuously increases, these "classical" approaches, even using larger and larger grids, will be faced by several bottlenecks. These are due to the total CPU time needed to handle so many docking calculations, and the large amount of data that needs to be transferred to and from the computer, and stored. Our approach, using both a hierarchy of filters to reduce step by step the number of the chemicals to be considered, and a hierarchy of machines in a grid infrastructure, adapted to each screening step in the filter funnel, should help to overcome these bottlenecks. Thanks to Grid5000 all levels of the hierarchy can be executed on the same platform because in Grid5000 each node can be reserved individually (for small independent tasks) and each Grid5000 site hosts a cluster that can be reserved (for high-performance parallel application).

Nevertheless, with the workflow used here and with the results presented in this paper, several problems can be solved. This applies especially to the capability of middleware to handle the enormous number of very short CPU

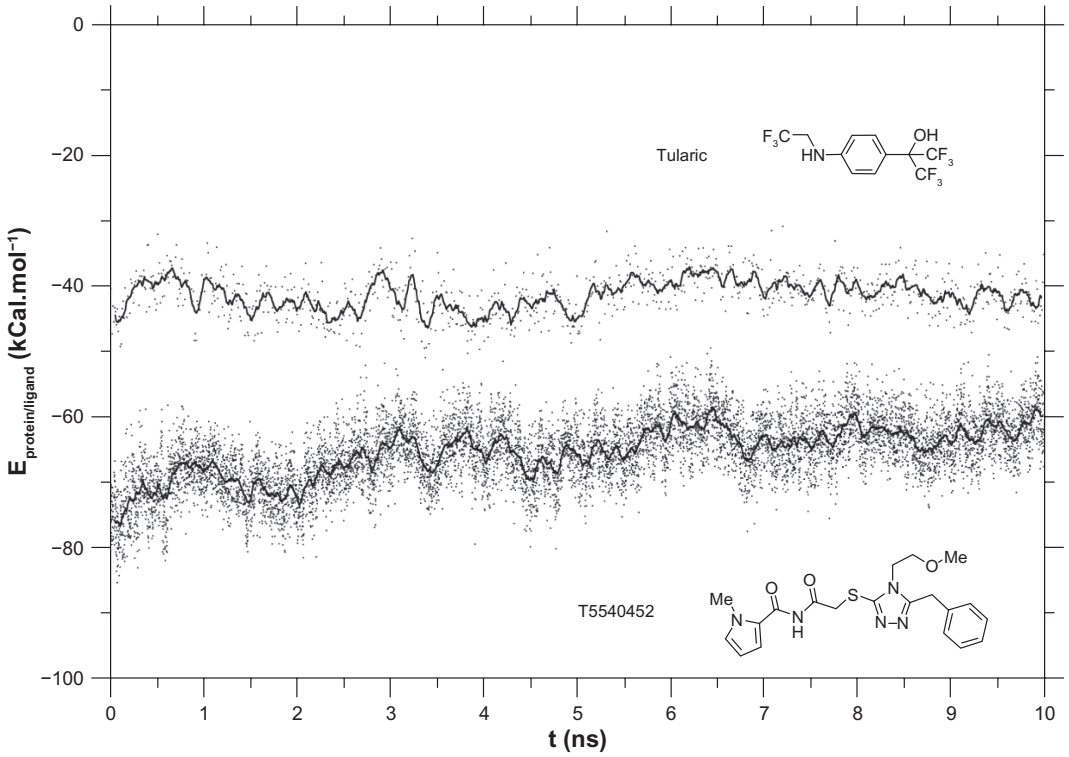


Figure 6 Variation of protein/ligand interaction energies for T5540452 versus LXR β crystallographic ligand tularic (44B) in 1PQ6 RX structure during the molecular dynamic trajectories using NAMD.

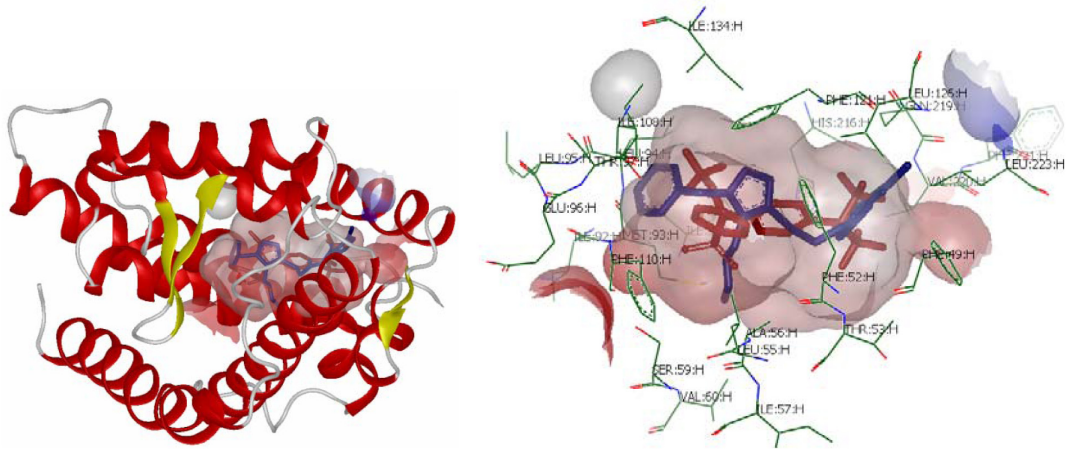


Figure 7 LXR β structure complexes with Enamine T5540452 compound. The Enamine T5540452 compound (blue) binds in the same place than the crystallographic ligand (44B, also called Tularic) (red) of the LXR β structure (1pq6) and occupies a larger volume in the binding pocket.

consuming jobs. In fact, as seen in Table 1, an equilibrated balance between the input/output and the calculations to be performed on a grid node seems a necessary requirement to achieve good performances on a very large scale simulation. Works are in progress in several laboratories to handle this problem on grids.[60–63]

It is clear that the MSSH preliminary calculations are the lengthy ones (78 days). Of course if we compare the times required for one molecule using GOLD (~13 minutes) and using MSSH (2 seconds/conformer and therefore ~13 minutes for the 400 conformations generated for one molecule), it is obvious that the use of MSSH could be questionable. But several important points need to be made. Firstly, although an average of 150 conformers/molecule was produced, only an average of 5 minutes was used by MSSH/molecule meaning that MSSH remains at least three times faster than GOLD. Secondly, MSSH is able to handle a larger conformational diversity/molecule than GOLD as it will consider many more conformations for each ligand an average of 150 highly different conformers for MSSH compared to an average of 25 weakly different conformers for GOLD (this point is highlighted in Figure 8 for the T5540452 molecule). And thirdly, MSSH has to be done only one time and its results are reusable without any additional computing time for another docking of the same ligand database on another target or another conformation of the same target. This MSSH

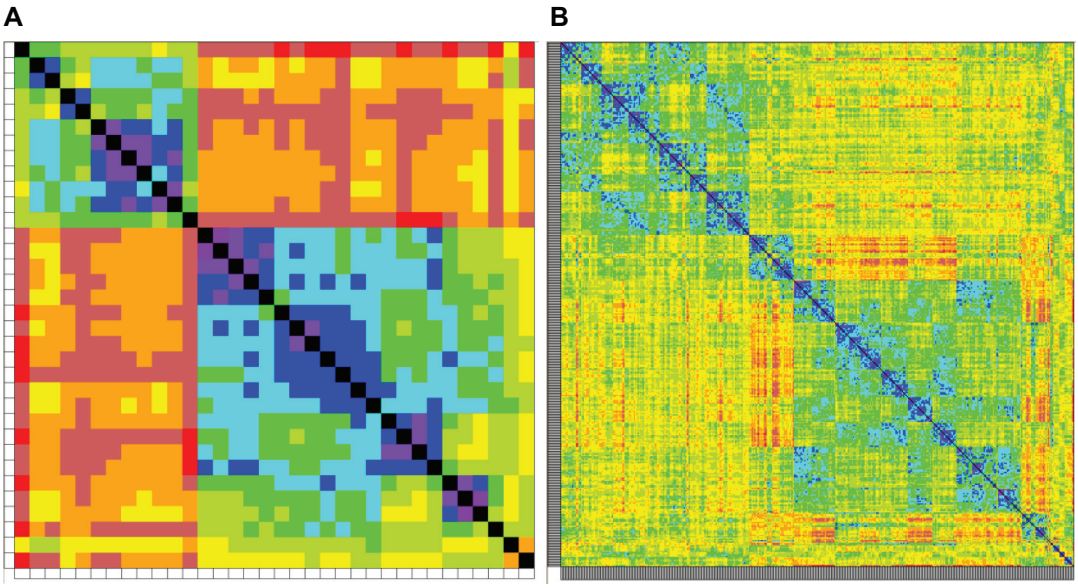


Figure 8 Conformational diversity of GOLD (34 conformers) (**A**) and MSSH (400 conformers generated by Omega) (**B**) for T5540452 compound. Atomic root mean square displacement distance criteria between pairs of structures was used to represented the conformational diversity with MacroModel XCluster, version 9.6 (Schrödinger, New York, NY).

preliminary step could also be more efficiently performed on a user's workstation using GPU resources (work in progress) and not on the cluster grid itself.

Nevertheless, if, in this experience, we consider only the timings of the SHEF, GOLD and NAMD calculations (Table 1), it appears that spreading calculations on a cluster grid clearly provides an efficient benefit for high-throughput virtual screening (total time for the whole docking part itself of 3.6 days compared to 274 days without Grid5000). This result confirms the interest for developing such technology in several research groups.[59,64,65] Nevertheless, the APST middleware which was used here is clearly not adapted to processing several million very short jobs, but is satisfactory for thousands of medium CPU time-consuming jobs. Because the distribution of tasks was not made in an optimal way, all the resources available were not used in an optimal way. This situation was mainly due to the transfer time of the input files for different jobs and due to the fact that APST uses only one daemon (apstd), which takes care of the distribution of all the tasks. To optimize MSSH on the grid, we plan to test the Athapascan[66] or BOINC[67] tools, which, while similar to APST, have multiple threads for the distribution of tasks.

Conclusions

From the present work, it appears clearly that with the VSM-G platform, coupled with a dedicated middleware, a user on the local cluster can execute an entire virtual screening experiment, ranging from large-scale fast matching between candidate chemicals and protein binding sites to more elaborate molecular dynamics refinements, with only a few commands since the processes of job distribution, monitoring progress, data manipulation, and retrieval of results has been automated by the platform. The large pool of resources offered by cluster grid computing enables to take advantage of a thorough, yet computationally expensive, docking strategy employing both sequential and parallel algorithms when screening a very large database of compounds.

Acknowledgments

The authors gratefully acknowledge P Bladon for advice given and for correcting the manuscript. We thank Y Asses for helpful discussions and V Leroux for assistance in the pair-interaction calculations. The simulations presented in this paper were carried out using the Grid5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER, and several universities as well as other funding bodies (see http://www.grid5000.fr/). L Ghemtio was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), CNRS (Centre National pour la Recherche Scientifique), and the Bill and Melinda Gates Foundation. We thank Openeye and Chemaxon for providing free academic licenses for their software. This project was supported by Region Lorraine within the framework of the PRST MISN (MBI operation).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Good AC, Krystek SR, Mason JS. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov Today*. 2000;5:61–69.
2. Carnero A. High throughput screening in drug discovery. *Clin Transl Oncol*. 2006;8:482–490.
3. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov*. 2002;1:882–894.
4. Abagyan R, Totrov M. High-throughput docking for lead generation. *Curr Opin Chem Biol*. 2001;5:375–382.
5. Lyne PD. Structure-based virtual screening: an overview. *Drug Discov Today*. 2002;7:1047–1055.
6. Dias R, de Azevedo WF Jr. Molecular docking algorithms. *Curr Drug Targets*. 2008;9:1040–1047.
7. Kontoyianni M, Madhav P, Suchanek E, Seibel W. Theoretical and practical considerations in virtual screening: a beaten field? *Curr Med Chem*. 2008;15:107–116.
8. Zhang LY, Gallicchio E, Friesner RA, Levy RM. Solvent models for protein-ligand binding: comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *J Comp Chem*. 2001;22:591–607.
9. Kaya H, Chan HS. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J Mol Biol*. 2003;326:911–931.
10. Seifert MH, Kraus J, Kramer B. Virtual high-throughput screening of molecular databases. *Curr Opin Drug Discov Devel*. 2007;10:298–307.
11. Toledo-Sherman LM, Chen D. High-throughput virtual screening for drug discovery in parallel. *Curr Opin Drug Discov Devel*. 2002;5:414–421.
12. Zoete V, Grosdidier A, Michielin O. Docking, virtual high throughput screening, and in silico fragment-based drug design. *J Cell Mol Med*. 2009;13:238–248.
13. Davies EK, Richards WG. The potential of Internet computing for drug discovery. *Drug Discov Today*. 2002;7:S99–S103.
14. Claus BL, Johnson SR. Grid computing in large pharmaceutical molecular modeling. *Drug Discov Today*. 2008;13:578–583.
15. Wolf A, Shahid M, Kasam V, Ziegler W, Hofmann-Apitius M. In silico drug discovery approaches on grid computing infrastructures. *Curr Clin Pharmacol*. 2009 Jan 1. [Epub ahead of print].
16. Woods CJ, Ng MH, Johnston S, et al. Grid computing and biomolecular simulation. *Philos Transact A Math Phys Eng Sci*. 2005;363:2017–2035.
17. Beautrait A, Leroux V, Chavent M, et al. Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment. *J Mol Model*. 2008;14:135–148.

18. Mestres J. Virtual screening: a real screening complement to high-throughput screening. *Biochem Soc Trans*. 2002;30:797–799.
19. Peters A, Lundberg ME, Lang PT, Sosa CP. High throughput computing validation for drug discovery using the DOCK program on a massively parallel system. *RedPaper*. 2008;REDP-4410–00.
20. Shave SR, Taylor P, Walkinshaw M, Smith L, Hardy J, Trew A. Ligand discovery on massively parallel systems. *IBM J Res Dev*. 2008;52: 57–67.
21. Clery D. Infrastructure. Can grid computing help us work together? *Science*. 2006;313:433–434.
22. Coveney PV. Scientific grid computing. *Philos Transact A Math Phys Eng Sci*. 2005;363:1707–1713.
23. Sild S, Maran U, Lomaka A, Karelson M. Open computing grid for molecular science and engineering. *J Chem Inf Model*. 2006;46: 953–959.
24. Koh JT. Making virtual screening a reality. *Proc Natl Acad Sci U S A*. 2003;100:6902–6903.
25. Kasam V, Salzemann J, Botha M, et al. WISDOM-II: screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar J*. 2009;8:88.
26. Catlow CR. New science from high-performance computing: an introduction. *Philos Transact A Math Phys Eng Sci*. 2002;360:1075–1078.
27. Cuticchia AJ. High performance computing and medical research. *CMAJ*. 2000;162:1148–1149.
28. Pitera JW. Current developments in and importance of high-performance computing in drug discovery. *Curr Opin Drug Discov Devel*. 2009;12:388–396.
29. Cai W, Shao X, Maigret B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J Mol Graph Model*. 2002;20:313–328.
30. Brown SP, Muchmore SW. High-throughput calculation of protein-ligand binding affinities: modification and adaptation of the MM-PBSA protocol to enterprise grid computing. *J Chem Inf Model*. 2006;46: 999–1005.
31. Swindells M, Rae M, Pearce M, Moodie S, Miller R, Leach P. Application of high-throughput computing in bioinformatics. *Philos Transact A Math Phys Eng Sci*. 2002;360:1179–1189.
32. Cappello F, Caron E, Dayde M, et al. Grid'5000: a large scale and highly reconfigurable grid experimental testbed. In: Grid 2005: Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing; 2005 Nov 13–14; Seattle, USA. Washington, D.C: IEEE Computer Society; 2005:99–106.
33. Yiannis G, Richard O, Capit N. Evaluations of the lightweight grid CIGRI upon the Grid'5000 platform. In: Fox G, Chiu K, Buyya R, editors. eScience 2007: Proceedings of the 3rd IEEE International Conference on e-Science and Grid Computing; 2007 Dec 10–13; Bangalore, India. Los Alamitos: IEEE Computer Society; 2007. P. 279–286.
34. Cai W, Xu J, Shao X, Leroux V, Beautrait A, Maigret B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. *J Mol Model*. 2008;14:393–401.
35. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;267:727–748.
36. Phillips JC, Braun R, Wang W, et al. Scalable molecular dynamics with NAMD. *J Comput Chem*. 2005;26:1781–1802.
37. Casanova H, Berman F. Parameter sweeps on the grid with APST. *Concurrency: Pract Exper*. 2002;1:1–15.
38. Casanova H, Obertelli G, Berman F, Wolski R. The AppLeS parameter sweep template: user-level middleware for the grid. *Sci Program*. 2000;8:111–126.
39. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28:235–242.
40. Williams S, Bledsoe RK, Collins JL, et al. X-ray crystal structure of the liver X receptor beta ligand binding domain: regulation by a histidine-tryptophan switch. *J Biol Chem*. 2003;278:27138–27143.
41. Farnegardh M, Bonn T, Sun S, et al. The three-dimensional structure of the liver X receptor beta reveals a flexible ligand-binding pocket that can accommodate fundamentally different ligands. *J Biol Chem*. 2003;278:38821–38828.
42. Lala DS. The liver X receptors. *Curr Opin Investig Drugs*. 2005;6: 934–943.
43. Beautrait A, Karaboga AS, Souchet M, Maigret B. Induced fit in liver X receptor beta: a molecular dynamics-based investigation. *Proteins*. 2008;72:873–882.
44. Irwin JJ. Using ZINC to acquire a virtual screening library. *Curr Protoc Bioinformatics*. 2008;Chapter 14:Unit 14.6.
45. Chemdiv. The chemistry of cures [online]. [cited 2009, Nov 30]. Available from: http://chemdiv.emolecules.com.
46. Enamine. [online]. [cited 2009, Nov 30]. Available from: http://www.enamine.net.
47. Amri Direct. Chemical compound database [online]. [cited 2009, Nov 30]. Available from: http://www.amridirect.com.
48. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Delivery Rev*. 2001;46:3–26.
49. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14:33–38, 27–28.
50. Patane S, Pietrancosta N, Hassani H, et al. A new Met inhibitory-scaffold identified by a focused forward chemical biological screen. *Biochem Biophys Res Commun*. 2008;375:184–189.
51. Wang R, Wang S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci*. 2001;41:1422–1426.
52. Kluszczyński R, Bała P. Supporting NAMD Application on the Grid Using GPE. In: Wyrzykowski R, Dongarra J, Meyer N, Wasniewski J, editors. Post-Proceedings of the 6th International Conference on Parallel Processing and Applied Mathematics; 2005 Sept 11–14; Poznan, Poland. Berlin: Springer; 2008. p. 762–769.
53. Tantar AA, Conilleau S, Parent B, et al. Docking and biomolecular simulations on computer grids: status and trends. *Curr Comput Aided Drug Des*. 2008;4:235–249.
54. Lee HC, Salzemann J, Jacq N, et al. Grid-enabled high-throughput in silico screening against influenza A neuraminidase. *IEEE Trans Nanobioscience*. 2006;5:288–295.
55. Salzemann J, Kasam V, Jacq N, Maass A, Schwichtenberg H, Breton V. Grid enabled high throughput virtual screening against four different targets implicated in malaria. *Stud Health Technol Inform*. 2007;126: 47–54.
56. Jacq N, Salzemann J, Legre Y, et al. Demonstration of in silico docking at a large scale on grid infrastructure. *Stud Health Technol Inform*. 2006;120:155–157.
57. Bullard D, Gobbi A, Lardy MA, Perkins C, Little Z. Hydra: a self regenerating high performance computing grid for drug discovery. *J Chem Inf Model*. 2008;48:811–816.
58. Chien A, Foster I, Goddette D. Grid technologies empowering drug discovery. *Drug Discov Today*. 2002;7:S176–S180.
59. Jiang X, Kumar K, Hu X, Wallqvist A, Reifman J. DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0. *Chem Cent J*. 2008;2:18.
60. Pérez MS, Carretero J, García F, Peña JM, Robles V. MAPFS-Grid: A Flexible Architecture for Data-Intensive Grid Applications. In: Rivera FF, Bubak M, Gómez A, Doallo R, editors. Grid Computing: First European Across Grids Conference; 2003 Feb 13–14; Santiago de Compostela, Spain. New York, NY: Springer; 2004. p. 111–118.
61. Kukla T, Kiss T, Kacsuk P, Terstyanszky G. Integrating open grid services architecture data access and integration with computational grid workflows. *Philos Transact A Math Phys Eng Sci*. 2009;367: 2521–2532.

62. Anglano C, Canonico M. A comparative evaluation of high-performance file transfer systems for data-intensive grid applications. In: Fredriksson M, Gustavsson R, Ricci A, Omicini A, editors. WETICE 2004: Proceedings of the 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises; 2004 June 14–16; Modena, Italy. Washington D.C: IEEE Computer Society; 2005. p. 283–288.
63. Tristan G, Johan M, Diane L, Xavier P. Flexible and efficient workflow deployment of data-intensive applications on grids with MOTEUR. *Int J High Perform Comput Appl*. 2008;22:347–360.
64. Kasam V, Zimmermann M, Maass A, et al. Design of new plasmepsin inhibitors: a virtual high throughput screening approach on the EGEE grid. *J Chem Inf Model*. 2007;47:1818–1828.
65. Levesque MJ, Ichikawa K, Date S, Haga JH. Design of a grid service-based platform for in silico protein-ligand screenings. *Comput Methods Programs Biomed*. 2009;93:73–82.
66. Athapascan. Athapascan is a high level application programming interface [homepage on the Internet]. [cited 2009, Nov 30]. Available from: http://www-id.imag.fr/Logiciels/ath1/.
67. Anderson DP. BOINC: a system for public-resource computing and storage. In: Buyya R, editor. Grid 2004: Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing; 2004 Nov 8; Pittsburgh, USA. Los Alamitos: IEEE Computer Society; 2005. p. 4–10.
68. Costanzo A di, Assuncao MD de, Buyya R. Harnessing cloud technologies for a virtualized distributed computing infrastructure. *IEEE Internet Computing*. 2009;13:24–33.