



3701 Market Street
5th Floor, Suite 119
Philadelphia, PA 19104
215.989.4880

Virtual Screening Using Amazon EC2 Cloud Computing June 18, 2010

Background

The primary mechanism for the identification of new lead compounds remains the physical screening of vast chemical libraries against biological targets. Virtual Screening (VS) provides an alternative approach in which the large libraries of chemical structures are computationally searched against a biological target of known structure. In these virtual or “docking” screens, large libraries of organic molecules are docked into receptor structures and ranked by the calculated affinity (1-6). Various scoring functions are used to rank the ‘hits’ coming from these screens (7-9).

For many researchers, the discovery process includes a cycle of high-throughput virtual screening, scoring and identifying promising compounds, and validation through biological screening. Maintaining a dedicated computational resource, updated chemical structure libraries, and software for virtual screening is not a practical or efficient use of resources. We therefore propose development of a high-throughput virtual screening platform using Amazon Web Services (AWS) (10) which will provide virtual screening capabilities to researchers on an as-needed basis. Access to these services will be provided through a web interface, decreasing the learning curve required for use of virtual screening software.

High-performance scientific computation has traditionally required dedicated and specialized hardware and associated infrastructure. Recent innovations in the internet and the availability of access to large, virtualized pools of compute resources have led to the development of what is known as “Cloud Computing.” In this paradigm, users access compute resources as needed, without the overhead of maintaining a dedicated infrastructure or support staff. This is known as Infrastructure as a Service (IaaS), for which Amazon Web Services (AWS) is a major provider.

By expanding access to resources for virtual screening, scientists which otherwise would be limited in their ability to apply advanced computational methods to their research. Combining this pool of newly-enabled researchers with the rapid increase in both crystallographic and homology models of proteins, and the availability of extensive, publically-available compound libraries, there exists a tremendous potential for the identification of novel therapeutic compounds for the treatment of human disease.

Preliminary Results

The impetus for this study was based the scarcity of computational resources for virtual screening for identification of novel compounds binding to a therapeutically-relevant target. Plans to expand existing computing infrastructure were evaluated and rejected due to maintenance and overhead costs on a defined-lifetime project. We identified the Amazon EC2 'cloud' computing environment as a potential alternative, due to the transformation of fixed costs into variable costs.

Experimental Design

We identified four activities associated with this evaluation.

1. Building Linux virtual machines with screening tools and associated software, for deployment on Amazon EC2.
2. Benchmarking performance and cost using a small test set of compounds that had been previously screened on our server. The test set was 24 compounds that our collaborators believe bind to HMG-CoA reductase. They varied in size from mw 221.2 to 1464.

- Benchmark using a larger set of 10,000 compounds derived from the ZINC database (11), subset #3, drug-like compounds. Assess the ability utilize multiple EC2 instances, and develop tools to provide job control across EC2 instances. Patches were contributed to PPSS, an open source job control software.
- Ligand and receptor preparation. Autodock (12) requires ligands to be in pdbqt format. The prepare_ligand4.py script from AutoDock Tools (ADT) (13) was used to add hydrogens, merge non-polar hydrogens, assign partial charges, and define rotatable bonds for ligands. When required, OpenBabel (14) was used to convert file formats. The receptor used was obtained from the PDB (1HWK), atorvastatin bound to HMG-CoA reductase (15). ADT was used to prepare the receptor coordinate files for AutoGrid and AutoDock. Grid dimensions were based on the identified regions of intermolecular interaction of the ligand within the receptor. Default AutoDock parameters were used. The resulting grid and configuration files were used for all subsequent screens.

Results:

- EC2 Environment. A CentOS 5.4 system image was created and uploaded to Amazon. This image was functionally equivalent to that running on our in-house server with identical operating system and software configuration. Security for data transfer was provided using SSH/SCP and public/private key pairs. Time to develop and test the environment was 2.5 days.
- Performance Benchmarking.

System configuration for test 1:

Internal server configuration: 2.13ghz quad-core, 4GB RAM rack mounted running CentOS 5.4, 500 GB disk storage.

Amazon EC2 environment (High-CPU Extra Large Instance): 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform. An EC2 compute unit is equivalent to a 1.0-1.2 ghz Xeon processor. Cost is \$0.68/hour, plus data transfer and storage fees.

Test 1 Summary of Results							
Server Type	# of Cores	# of Instances	# of Compounds	Total Time (hrs)	Time/Compound (hrs)	Total Cost	Cost/Compound
In-house	4	1	24	17.5	0.73	n/a	n/a
Amazon EC2	8	1	24	8.75	0.36	\$6.08	\$0.25

Table 1. Comparison of in-house server performance and Amazon EC2.

The majority of time was spent on the largest ligands. Note that these compounds are not representative of a traditional virtual screening library and represent a set defined by our collaborator. Performance is slightly faster than in-house server (Amazon instance has twice the number of cores).

System configuration for Test 2:

4 Amazon EC2 instances equivalent to that described above. We took advantage of spot pricing, a mechanism which allows bidding for unused instances at a reduced price. Cost averaged \$0.28 (per instance, plus data transfer and storage fees).

Test 2 Summary of Results							
Server Type	# of Cores	# of Instances	# of Compounds	Total Time (hrs)	Time/Compound (hrs)	Total Cost	Cost/Compound
Amazon EC2	32	4	10000	32.8	0.0033	\$38.33	\$0.0038

Table 2. Performance of a multiple-instance Amazon EC2 virtual screen.

Use of a compound library of molecular weight between 150 and 500 greatly improved performance. The use of spot pricing reduced cost. This test demonstrated the ability to use multiple instances for virtual screening.

In summary, we believe this pilot provided justification for continued use of the Amazon EC2 environment for virtual screening using Autodock and further development of the platform as a research tool. We have successfully used this existing environment in 4 other virtual screens. However, a high degree of expertise is required to maintain and utilize the existing software due to its complexity and lack of robustness typical of prototypic software.

Example of validation methodology

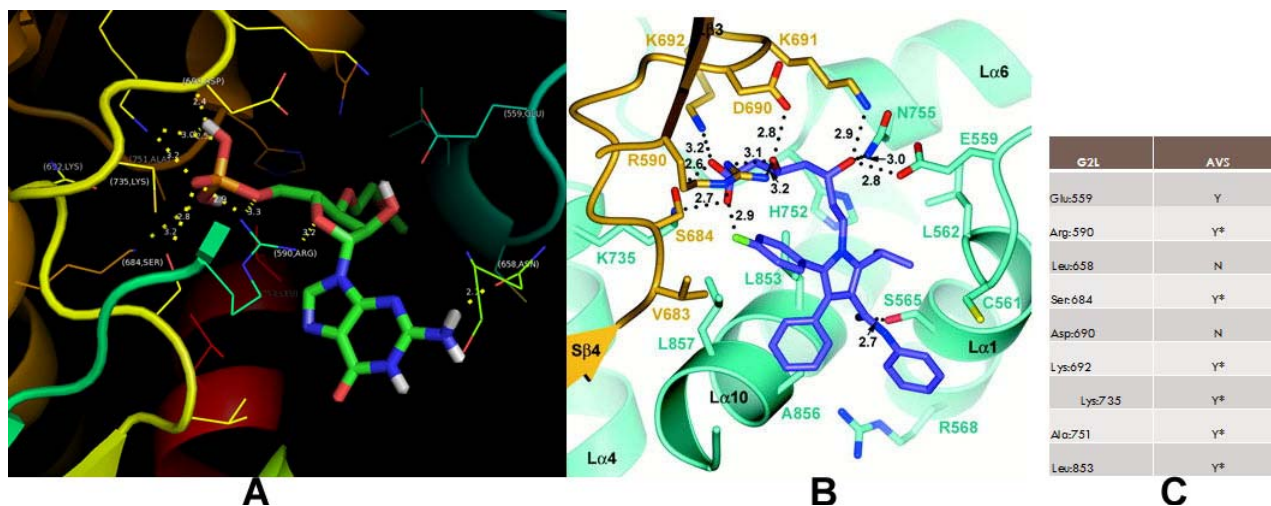


Fig 2. Analysis of binding of a compound, G2L (3'-o-methoxyethyl-guanosine-5'-monophosphate) to the Hmg-CoA reductase site. Panel A shows the molecular interactions of G2L with the HMG-CoA reductase binding site. Pymol was used for graphics, labels, and identifying H-bonds between ligand and residues in binding site. Panel B shows the molecular interactions of atorvastatin to the Hmg-CoA reductase site. Panel C shows a comparison table of interactions conserved between A and B. An asterisk indicated a critical bond found in the atorvastatin-HMG-CoA reductase interactions.

References

- Schoichet, B. Virtual screening of chemical libraries. 2004 Dec; *Nature* 432(7019): 862-865.
- Sousa SF, Cerqueira NM, Fernandes PA, Ramos MJ. Virtual screening in drug design and development. *Comb Chem High Throughput Screen*. 2010 Jun; 13(5):442-53.
- Bajorath, J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov*. 2002 Nov; 3(11):935-49.
- Jain AN. Virtual screening in lead discovery and optimization. 2004. *Curr Opin Drug Discov Devel*. Jul;7(4):396-403
- Ghosh S, Nie AH, An J, Huang Z: Structure-based virtual screening for drug discovery. 2006. *Curr Opin Chem Biol*, 10:194-202.
- Barril X, Hubbard RE, Morley SD. Virtual screening in structure-based drug discovery. *Mini Rev Med Chem*. 2004 Sep;4(7):779-91.
- de Azevedo WF Jr, Dias R. Computational methods for calculation of ligand-binding affinity. *Curr Drug Targets*. 2008 Dec;9(12):1031-9.
- Kitchen DB, Decornez H, Furr JR, Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*. 2004. Nov; 3(11):935-49.

9. Kroemer RT. Structure-based drug design: docking and scoring. *Curr Protein Pept Sci.* 2007 Aug;8(4):312-28.
10. Amazon Web Services (AWS). <http://www.aws.amazon.com>
11. Irwin JJ, Shoichet BK. ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* 2005; 45:177–182
12. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem.* 1998;19:1639–1662.
13. AutoDock Tools http://autodock.scripps.edu/resources/adt/index_html
14. OpenBabel <http://openbabel.sourceforge.net>
15. Istvan ES, Deisenhofer J. Structural mechanism for statin inhibition of HMG-CoA reductase. *Science.* 2001 May 11;292(5519):1160-4.