

Computer-assistant Drug Discovery with the Netezza Architecture

Li Bai

Laboratory of Cell and Molecular Immunology
College of Medicine, Henan University
Kaifeng, Henan 475001, China

Qitai Xu

Department of Pharmacology
Pharmacy College of Henan University
Kaifeng, Henan 475001, China

Guy M Lingani

Department of Biochemistry and Molecular Biology
Howard University College of Medicine
Washington, DC 20059, USA

Clay S Gloster

Department of Electrical Engineering
Howard University
Washington, DC 20059, USA

William Southerland

Department of Biochemistry and Molecular Biology
Howard University College of Medicine
Washington, DC 20059, USA

Zengjian Hu

Department of Biochemistry and Molecular Biology
Howard University College of Medicine
Washington, DC 20059, USA
zhu@howard.edu

Abstract—while relational databases have become critically important in business applications and web services, they have played a relatively minor role in scientific computing, especially in computer-assistant drug discovery, which has generally been concerned with modeling and simulation activities. However, massively parallel database architectures are beginning to offer the ability to quickly search through terabytes of data with hundred-fold or even thousand-fold speedup over server-based architectures. These new machines may enable an entirely new class of algorithms for scientific applications. The drug discovery and development community is able now to make good use of these new database machines.

Keywords—Netezza architecture; drug discovery; database; scientific computing; modeling

I. INTRODUCTION

Modern drug discovery is a complex, risky, time consuming, and costly process. Despite increased investment in research and development, the numbers of drugs launched has declined in recent years [1]. In the so-called post-genomic era, rapid progress is being made in structural genomics, functional genomics, proteomics, and pharmacokinetics and drug metabolism. This has provided information about protein structure and function, cellular profiles of proteins, molecular distributions, and metabolism. The volume of public sequence databases, target databases, and compound database are increasing exponentially these days. Drug discovery process, therefore, is steadily becoming more information driven and a data-centric problem where new drug discovery is based on analysis and data mining to unveil the information hidden behind the large genomic, proteomic, and small molecular databases.

To improve productivity, knowledgebase-guided decisions must be incorporated into the drug discovery and development process.

Computational methodologies have become a crucial component of many drug discovery processes, from hit identification to lead optimization and beyond [2], and approaches such as structure- or ligand-based virtual screening techniques and information-based method are widely used in many drug discovery efforts [3]. It is well known that relational databases have become critically important in business applications, but they have played a relatively minor role in drug discovery and development process, which has generally been concerned with modeling and simulation activities. In recent years, massively parallel database architectures are beginning to offer the ability to quickly search through terabytes of data with hundred-fold or even thousand-fold speedup over server-based architectures, and the Netezza Performance Server (NPS®) system [4], a massively parallel database machine originally designed for such analytic searches, provides a powerful tool that integrates multiple computational algorithms for meeting the data mining needs of biologists, chemists, and pharmacologists to speed up the new drug discovery. In this paper, the various uses of Netezza in the drug discovery process have been investigated.

II. THE NETEZZA ARCHITECTURE

The Netezza Performance Server (NPS®) system's architecture, depicted in Figure 1, is a two-tiered system designed to handle very large queries from multiple users, which consists of a closely coupled server with many parallel Snippet Processing Units, each with their own disk and streaming database logic chip to perform fast pattern matching.

The second tier consists of dozens to hundreds or thousands of Snippet Processing Units (SPUs) operating in parallel which are connected by Gigabit Ethernet to both the host server and to the other SPUs. Each SPU is an intelligent query processing and storage node, and consists of a powerful commodity processor, dedicated memory, a disk drive and a field-programmable disk controller with hard-wired logic to manage data flows and process queries at the disk level, as depicted in Figure 2. The massively parallel, shared-nothing SPU blades provide the performance advantages of massively parallel processors.

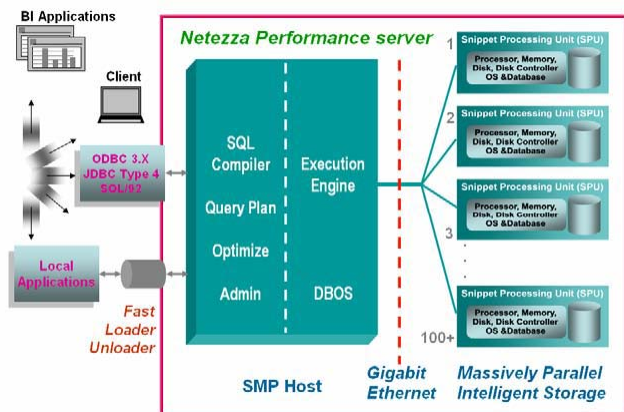


Figure 1. The Netezza Performance Server system

Nearly all query processing is done at the SPU level, with each SPU operating on its portion of the database. All operations that easily lend themselves to parallel processing (including record operations, parsing, filtering, projecting, interlocking and logging) are performed by the SPU nodes, which significantly reduces the amount of data moved within the system. Intelligent Query Streaming™ is performed on each SPU by a Field-Programmable Gate Array (FPGA) chip that functions as the disk controller, but which is also capable of basic processing as data is read from the disk.

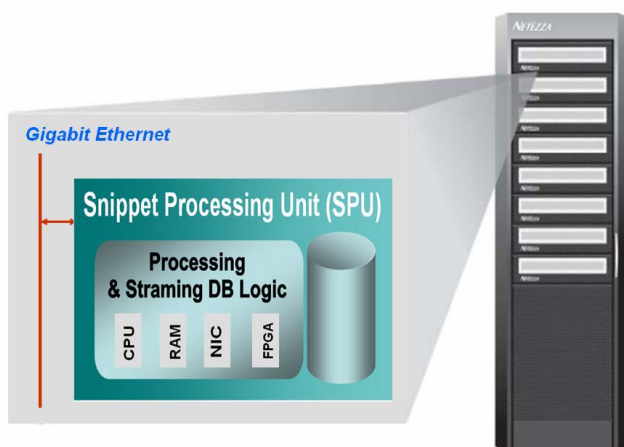


Figure 2. Snippet Processing Units

The system is able to run critical database query functions such as parsing, filtering and projecting at full disk reading speed, while maintaining full ACID (Atomicity, Consistency,

Isolation, and Durability) transactional operations of the database.

To achieve high performance, the storage interconnection, which is a bottleneck with traditional systems, is eliminated by directly attaching the disks so that data can stream straight into the FPGA for initial query filtering. Then, to further reduce the workload on the central server, the intermediate query tasks are performed in parallel on the SPUs.

III. APPLICATIONS IN COMPUTER-ASSISTANT DRUG DISCOVERY

Computational methodologies have become an integrate and crucial component of many drug discovery processes. Netezza provides a powerful tool that integrates multiple computational algorithms for meeting the data mining needs of drug discovery process and could make new drug discovery faster, cheaper and smarter.

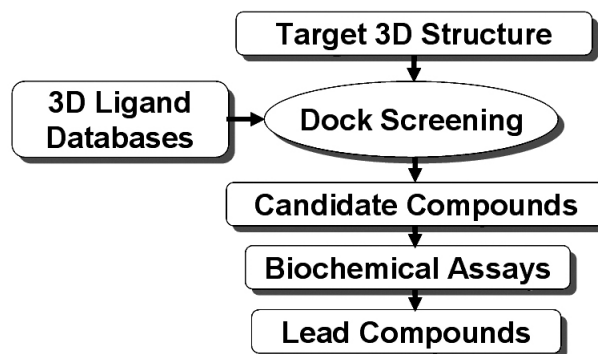
A. Structure-based drug discovery

Structure-based virtual screening (VS) method, as depicted in Figure 3, for drug discovery is typically carried out by computationally docking a large amount of compounds into the active site of protein target [5].

Figure 3. Structure-based drug discovery

The completion of the Human Genome Project and recent advances in structural genomics and proteomics have identified a large number of human proteins as drug targets [6-8], which were estimated about 5000 [9]. However, the explosion in the protein target available is yet to create a commensurate increase in the efficiency of the drug discovery process. The success of structure-based method strongly depends on the amount and quality of information available about the system under investigation. It can be argued that obtaining protein target is no longer the key issue facing early drug discovery, and that challenge has now becomes using proper tools and suitable methods to represent, store, retrieve, and analyze the protein targets and apply the information to select right target to screening.

Furthermore, during the past few years, a great deal of



effort has gone into the development of computational methods for filtering screening databases. More and more people are skeptical of the need to engage in ultra-VS, where many

hundreds of thousands of compounds are screened. Using Netezza's massively parallel database architecture, we can filter the databases and screen a small but highly diverse collection, such as the compounds that comprise drug-like synthetics, natural products, and FDA-approved drugs.

B. Information-based drug discovery

Since 1990, the United States National Cancer Institute (NCI) has conducted an anticancer drug discovery program in which approximately 10 000 compounds are screened every year in vitro against a panel of 60 human cancer cell lines from different organs[10-12]. Available are screening results of compounds that are not covered by a confidentiality agreement, and the compound list is updated at least once a year (<http://dtp.nci.nih.gov>), which provides us with a valuable source for computer-based virtual screening of anticancer drugs using a bioinformatics-based approach.

A number of studies have shown that although growth inhibitory activity for a single cell line is not informative, the activity patterns across the 60 cell lines provide incisive information on the mechanism of action of screened compounds and also on molecular targets and modulators within the cancer cells. Several algorithms have been introduced to use the activity information for discovery of anticancer drugs and for understanding of the molecular pharmacology of cancer, which have proven to be very useful for finding agents with activity patterns similar to that of a "seed" compound and for finding compounds with activity patterns that correlate well across the 60 cell lines with the expression levels of particular molecular targets [13-16]. An "information-intensive" approach has been developed to use this anticancer database for studies of molecular pharmacology of cancer and for the identification of potential protein targets of an anticancer drug [17-23]. The application of Netezza would be of great help in speeding up this "information-intensive" approach.

C. Ligand-based drug discovery

Ligand-based approach for drug discovery begins typically with a collection of molecules known to bind to a set of related target. This collection of compounds is then used to perform similarity searching, pharmacophore searches or property profiling against one or more databases which might contain several million chemicals.

In recent years, the number of chemicals generated by traditional and contemporary approaches has increased dramatically. In principle, there could be as many as 10^{47} quadrillion chemicals that can be made to interact with human protein targets [9].

Although screening methods and scoring contribute to a successful screen, no factor has a larger role than the compounds used for the screen. It is being recognized that increasing the quality of screening databases, rather than their quantity, is likely to be an important determinant for the identification of active compounds that have a chance to make it through the drug discovery pipeline [2,24]. There is a current trend to 're-rationalize' drug discovery research, that is,

departing from a mere 'numbers game' and carrying out fewer, but 'smarter' screening.

Factors such as physical properties, target classes and 'drug-likeness' are all important considerations, and computational approaches using Netezza can be valuable in addressing several of these issues. During the past few years, a great deal of effort has gone into the development of computational methods for filtering screening databases.

IV. CONCLUSION

A large number of drug targets, drug candidates, and a paucity of suitable computer tools to represent, store, retrieve, and analyze these information in the drug discovery process have created a 'target-rich and lead-poor' imbalance. The application of Netezza could help resolve the imbalance between target-rich and lead-poor since cheminformatics methods can be applied to extract knowledge from large-scale molecule databases using Netezza in a shorter time periods in order to assure that good properties are achieved before screening.

ACKNOWLEDGMENT

This work is supported by grant RCMI-NIH 2G12RR03048, and NSF 07-510 Major Research Instrumentation (MRI) grant CNS-0723060.

REFERENCES

- [1] Ghose AK, Herbertz T, Salvino JM, Mallamo JP. Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discov Today*. 2006; 11: 1107-1114.
- [2] Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov*. 2002; 1: 882-894.
- [3] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*. 2004; 3: 935-949.
- [4] Netezza 2006, Netezza Inc., (www.netezza.com)
- [5] Walters WP, Namchuk M. Designing screens: how to make your hits a hit. *Nat Rev Drug Discov*. 2003; 2: 259-266.
- [6] Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. 2001; 291: 1304-1351.
- [7] Banks RE, Dunn MJ, Hochstrasser DF, Sanchez JC, Blackstock W, Pappin DJ, Selby PJ. Proteomics: new perspectives, new biomedical opportunities. *Lancet*. 2000; 356: 1749-1756.
- [8] Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. *Nat Genet*. 1999; 23: 151-157.
- [9] Pang YP. In silico drug discovery: solving the "target-rich and lead-poor" imbalance using the genome-to-drug-lead paradigm. *Clin Pharmacol Ther*. 2007; 81:30-34.
- [10] Alley, M. C.; Scudiero, D. A.; Monks, A.; Hursey, M. L.; Czerwinski, M. J.; Fine, D. L.; Abbott, B. J.; Mayo, J. G.; Shoemaker, R. H.; Boyd, M. R. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Res*. 1988; 48: 589-601.
- [11] Monks, A.; Scudiero, D. A.; Shoemaker, R. H.; Paull, K. D.; Vistica, D.; Hose, C.; Langley, J.; Cronise, P.; Vaigro-Wolff, A.; Gray-Goodrich, M.; Campell, H.; Mayo, J.; Boyd, M. R. Feasibility of a high-flux anticancer screen using a diverse panel of cultured human tumor lines. *J. Natl. Cancer Inst*. 1991; 83: 757-766.

- [12] Boyd, M. R.; Paull, K. D. Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. *Drug Dev. Res.* 1995; 34: 91-109.
- [13] Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* 1989; 81: 1088-1092.
- [14] Koo, H.-M.; Monks, A.; Mikheev, A.; Rubinstein, L. V.; Gray-Goodrich, M.; McWilliams, M. J.; Alvord, W. G.; Oie, H. K.; Gazdar, A. F.; Paull, K. D.; Zarbl, H.; Vande Woude, G. F. Enhanced sensitivity to 1-beta-D-arabinofuranosylecytosine and topoisomerase II inhibitors in tumor cell lines harboring activated ras oncogenes. *Cancer Res.* 1996; 56: 5211-5216.
- [15] Wosikowski, K.; Schuurhuis, D.; Johnson, K.; Paull, K. D.; Myers, T. G.; Weinstein, J.; Bates, S. E. Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns. *J. Natl. Cancer Inst.* 1997; 89: 1505-1513.
- [16] Zaharevitz, D. W.; Gussio, R.; Leost, M.; Senderowicz, A. M.; Lahusen, T.; Kunick, C.; Meijer, L.; Sausville, E. A. Discovery and initial characterization of the paullones, a novel class of small-molecule inhibitors of cyclin-dependent kinases. *Cancer Res.* 1999; 59: 2566-2569.
- [17] Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural computing in cancer drug development: Predicting mechanism of action. *Science* 1992; 258: 447-451.
- [18] Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997; 275: 343-349.
- [19] Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery database: cluster analysis of ellipticine analogs with p53-inverse and CNS-selective patterns of activity. *Mol. Pharmacol.* 1998; 53: 241-251.
- [20] Shi, L. M.; Fan, Y.; Lee, J. K.; Waltham, M.; Andrews, D. T.; Scherf, U.; Paull, K. D.; Weinstein, J. N. Mining and Visualizing Large Anticancer Drug Discovery Databases. *J. Chem. Inf. Comput. Sci.* 2000; 40: 367-379.
- [21] Marner, F.-J. Iridals and cycloiridals, products of an unusual squalene metabolism in sword lilies (Iridaceae). *Curr. Org. Chem.* 1997; 1: 153-186.
- [22] Shao L, Lewin NE, Lorenzo PS, Hu Z, Enyedy IJ, Garfield SH, Stone JC, Marner FJ, Blumberg PM, Wang S. Iridals are a novel class of ligands for phorbol ester receptors with modest selectivity for the RasGRP receptor subfamily. *J Med Chem.* 2001; 44: 3872-3880.
- [23] Zhang M, Fang X, Liu H, Guo R, Wu X, Li B, Zhu F, Ling Y, Griffith BN, Wang S, Yang D. Bioinformatics-based discovery and characterization of an AKT-selective inhibitor 9-chloro-2-methylellectinium acetate (CMEP) in breast cancer cells. *Cancer Lett.* 2007; 252: 244-258.
- [24] Smith A. Screening for drug discovery: the leading question. *Nature.* 2002; 418: 453-459.