“Faster Drug Discovery Using FPGA Accelerated Protein Docking and Screening”

A White Paper

By

Dr. Mathew Wortman
Dr. Eric Stahlberg
Daryl Popig (MSCS)

EDDR Informatics LLC

EDDR Informatics, LLC

Accelerating Change ^{tm}

Introduction

Structured based drug design is a long and expensive process due to the fact that it depends upon identifying biologically active small molecules from large chemical collections (libraries) that often number into the millions. Biological screening of large libraries is an extremely expensive undertaking which has prompted all pharmaceutical companies, and an increasing number of academic labs, to 'pre-filter' screening candidates to significantly smaller subsets of molecules that have an increased likelihood of having the desired biological effect. This 'pre-filtering' consists of using computational algorithms to match structural and chemical characteristics of the target biological molecule (i.e., protein or DNA) and/or known active small molecules to each member of the chemical screening library. Each compound in the chemical library is given a score based upon parameters such as shape and electrostatic complementarily, logP, number of rotatable bonds, polar surface area, etc. This drug design process can be dramatically improved in both time and cost if promising molecules can be found earlier in the selection process.

Pharmaceutical companies and Computational Chemists will screen selected molecules against very large databases to narrow their focus so their efforts are applied to only those molecules that meet the criteria above. These molecules have a dynamic 3-dimensional atomic substructure (called a pharmacophore) that can bind against the target protein. This drug molecule further complicates the docking process because it is highly flexible. This means the drug molecule exists as an ensemble of several hundred low-energy states, each of which must be tested for binding to a biological target. Since the dynamics of this biological process depends on the target protein structure and flexibility of the interacting molecules involved. These interactions in a three-dimensional space are determined by the charge density distribution of the interacting molecules and directly affect the chemical binding when trying to predict drug to DNA or protein binding at the atomic level.

Current molecular docking methods used in drug discovery algorithms use a fixed charge to keep computational overhead to a minimum. We propose developing and experimenting with a new innovative algorithm where these charges are adjusted over a small range to allow the drug molecule to move with the changes in these charges. This then becomes the basis for an optimization problem. To speedup these calculations we will build upon prior success in FPGA algorithm acceleration and develop this new drug docking algorithm in hardware and run it at hardware speeds using proven FPGA accelerator technology.

Scientific Advancement Potential

Traditional techniques to identify candidate drugs among the combined datasets employ large numbers of general purpose processors (clusters), using off-the-shelf docking programs to identify candidates. As increases in processor speeds level off and the economically viable end of Moore's law is reached, demands for faster turnaround grow and a new approach is needed.

Field Programmable Gate Array (**FPGA**) technologies, combined with new screening algorithms, offer an exciting new opportunity. Using the FPGA solution with parallel performance and reduced power consumption, Drug discovery ventures can take full advantage of the increasingly large drug and protein libraries.

Scientific advantages will be realized by providing a seamless integration path between the application and the FPGA where the scientist can focus on the science and drug discovery process rather than on the software and hardware interfacing and FPGA tool development. Currently, there is no “off the shelf” high-speed drug screening solution employing FPGA technology where a researcher can simply plug in and start analyzing data sets. Our product will advance clinical and biomedical research by providing the foundation for developing desktop FPGA solutions that can be adapted to real time analysis, as well as analysis of large very data sets (millions of compounds) such as those produced and NIH screening centers. Furthermore, this type of hardware implementation is significantly less expensive than cluster-based technologies in both initial cost and cost of maintenance, and does not require technical staff to support. These lower cost desktop systems will provide and advantage to small research groups that are interested in finding biologically active small molecules for their research programs but that do not otherwise have the computational staff to support current complex in silico screening solutions. By enabling the scientist to do more with less overhead, further discoveries will be made that will enable America’s competitive edge in the market place through technology and meet further market needs in areas such as nanotechnology, supercomputing, and new drug development.

Commercial Market Potential

EDDR-Informatics can realize considerable commercial potential realized in both the drug discovery IP development as a licensed product and as a valuable service to provide FPGA integration to commercial database systems such as Oracle and to open source systems such as MySQL. This project ultimately will lead to a high-speed chemical database search product based on FPGAs, thus addressing the current challenges of searching complex chemical libraries as well as provide a solid foundation to begin addressing concepts such as chemical and drug-like space which may contain as many as 10$_{26}$ chemical species. While the aggregate research investment in drug development increases, the pharmaceutical industry is undergoing transformation. As large pharmaceutical companies increasingly outsource their efforts, smaller companies are picking up the slack. The larger number of small companies greatly increases the market for the low power FPGA solution for high performance drug screening.

The present market is occupied commercially by vendors including Tripos, Accelrys, Schrodinger, and MDL delivering fully integrated solutions for drug discovery, including components for protein-ligand docking and scoring. These established vendors while potentially competitive, will also be developed as delivery channels for deploying the FPGA-based docking solution. Our FPGA-based product is delivered as a component to be integrated into solutions, both commercial and open source, at both the desktop and server level. The combination of component licensing with existing vendors, sales at the desktop level and sales of server integration components are expected to generate at least $1M within 18 months and $3M annually within 3 years. A university program will allow researchers access to the product at a reduced price but in return, EDDR-Informatics will receive a royalty on the drugs that are developed with our product.

Scientific Background

Drug docking is the most computationally intensive step in drug design. Although it is conventionally thought of as a method used simply to identify potentially active “hits,” almost all success stories report using a multiple-step process to identify lead molecules. In each of these cases, an initial library of drug-like small molecules is evaluated using a fast docking method, which screens out some of the top-scoring candidates. These candidates are subjected to a more detailed screening step involving manual inspection, matching to a pharmacophore, or scoring using another, more detailed, program. A multiple step process is necessary because of the computational intractability of the docking process. A relatively simple forcefield-based binding energy calculation for a receptor-ligand complex with known structure involves multiple energy calculations in order to find the best, or most likely, binding energy and can take tens of seconds on a high end workstation. Obviously this time will increase dramatically when the docked structure must be found – a global energy minimization problem. Current reviews in this area point out that even if docking programs can find a reasonable docked position, computing an accurate activity is next to impossible because of the smoothed energy functions usually employed.

From the software’s perspective, the problem is that it has been created to solve either a problem that is too generic, or too specific. Many docking programs, e.g. AutoDock, attempt to dock a molecule to the whole receptor structure. This approach necessitates the use of oversimplified scoring functions in order to find the global minimum. On the other hand, highly detailed calculations can be performed on small systems. Many docking programs take this approach and prune the structure back to just the binding pocket, or even a simple pharmacophore description of it. Although a highly complicated scoring function can be defined for these cases, the resulting score is still based only on an approximation to the actual system. The affinity of a ligand for its target receptor is the critical factor that determines its activity and it is directly related to the necessary patient dosage. What makes the docking problem attractive is that this is observable and can theoretically be computed exactly from computer simulations. The caveat is that this requires quantum mechanics and a sample size approaching infinity, since the binding free energy is computed as an average over all the possible states of the system. In real-world applications, a very large sample can be drawn from the most relevant, docked, states of the receptor-ligand-solvent system, and approximations to quantum mechanics have to be made in order to make energy evaluation computationally feasible. These approximations are collectively termed “force-field methods,” and have proven to be consistently successful. The average deviation from quantum-mechanical energies is only 0.38 kcal/mol. for the MMFF94 force field, which corresponds to a deviation of 0.64 in computed log(K) values. Even using forcefield methods, computing an accurate binding affinity for receptor-ligand complexes is heavily taxing to the point where it is generally not included in conventional docking programs, which search only for the minimum energy structure and do not account for entropic effects by sampling. If a minimum-energy structure is all that is desired, a smoothed energy function can be arbitrarily defined by eliminating non-polar hydrogen atoms and decreasing the repulsive force of molecule overlap; and this is the approach normally taken in forcefield docking programs. It is also common in docking programs to use an empirical scoring function, which is derived by fitting a set of parameters in an arbitrary function of receptor and ligand atom positions to measured binding affinities.

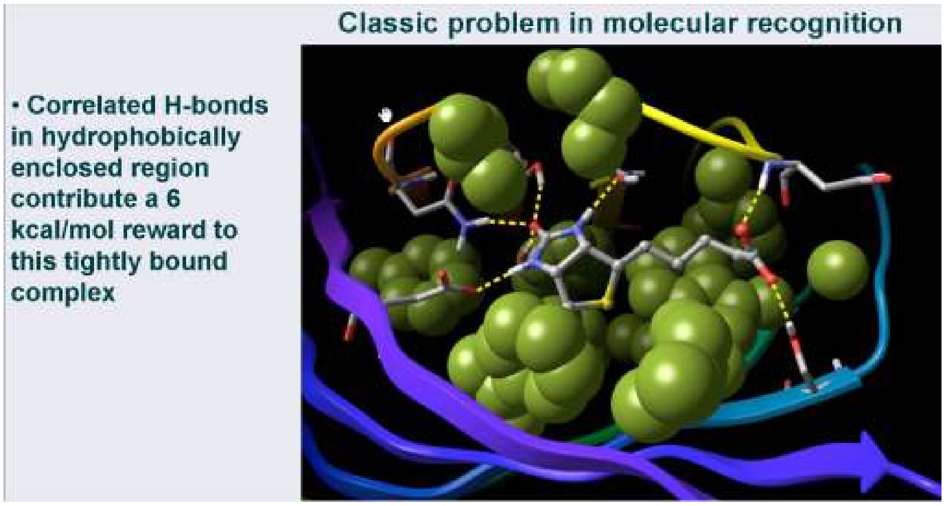
All of the successful docking programs developed to-date, have a simple 2-part strategy: generate likely candidate structures or “poses” of the ligand molecule inside the receptor’s binding pocket, and then score them. Since the posing is not directly based on the scoring function, the positions of the ligand atoms are usually moved until the energy score is at a minimum (energy minimization). Docking methodologies can thus be distinguished by the methods by which they do both of these tasks.

Molecule placement:
Genetic algorithms: AutoDock, GOLD
Random walk (e.g. Monte-Carlo): ICM, SEA
Surface complimentarity: FlexX, GOLD
Random empty-space filling: Dock, Fred
Clique-detection: UNITY, SEA
Binding Energy Function:
Forcefield-based: Dock, Fred, ICM, SEA
Empirical Function: AutoDock, Flexx, ICM
Pharmacophore-based: UNITY, Fred, SEA

Each docking methodology is optimized to the prediction of either the measured biological binding energy or the forcefield-based binding energy, which is close enough to exact. Unfortunately, by assuming that all the docking can be carried out in a single-step, the programs end up generating and then throwing away valuable information about the system. Accurate binding affinity calculations are essential in assessing any modifications made to the drug during development. Usually, this is the point where a computational chemist gets involved in the project and runs molecular dynamics simulations of the ligand and its receptor in order to achieve the high accuracy of forcefield methods. These accurate calculations are required in order to discriminate between the screened-out hits from the docking step, but could greatly benefit from the positional information generated there. This, of course, takes time and money, since the computational chemist must now set up a simulation for each high-ranking docked structure for each hit.

One of the earliest “docking” methods is based on the pharmacophore hypothesis. This hypothesis states that for a given active site on a target protein, there exists a set of chemical features (hydrogen bond donor/acceptor, aromatic stacking, hydrophobic cavity, etc.) that the ligand molecule has to compliment in order to bind to the receptor. This hypothesis has also been consistently shown to be true and even to be selective for a single target protein response. The current ubiquitously accepted “docking” program using pharmacophore information to find ligand poses is UNITY, and is used in many drug-design processes as a first step due to the speed of its highly efficient pose-generation method, clique-detection. The other docking programs on the list do not take pharmacophore information into account in their molecule placement stage because in order to become more general, however in every serious drug development program the researcher has or can generate a pharmacophore for the target of interest.

Our proprietary program (SEA) starts by generating the set of all possible ligand positions that satisfy the pharmacophore constraints and then applying the standard minimization approach using a forcefield energy minimization function. This is what would be expected in order to reproduce the successes of all the other docking programs available. However, the job of the researcher does not stop once hits have been identified. During the process of hit refinement, modifications to the molecules have to be made in order to reduce toxicity and improve the ADME profile before the hit can be sent to trial. SEA takes a unique “building-up” approach to the docking problem by filtering out the unfavorable poses of the ligand using their minimized score and then sampling the energy landscape around those minima in order to calculate an actual binding free energy. This step parallels the molecular dynamics simulation, except that here it has been combined into the docking program. This combination allows the pharmacophore information to be re-used as well in the form of a biasing potential. This makes the sampling much more efficient and is close to the approach of ICM. Since the energy cutoffs can be tuned and the minimization and sampling steps can be turned on or off, the complete program can be used in hit searching (strong cutoff, no sampling) as well as lead refinement (weak cutoff, high sampling). These considerations make SEA able to act as a multipurpose binding program.



Classic problem in molecular recognition

- Correlated H-bonds in hydrophobically enclosed region contribute a 6 kcal/mol reward to this tightly bound complex

In conclusion, EDDR Informatics presented a software algorithm re-written to run in hardware inside of a FPGA that demonstrated a 13X speedup over its software counterpart. The key to achieving this performance gain is to write the algorithm to be pipelined and to run in parallel in hardware. It is likely that this energy minimization function can be further optimized to run in hardware to get greater speedup. The API call to the energy minimization function has been written so it can take advantage of additional FPGAs as the hardware is scaled up. A 26X speedup is obtainable by splitting the work between two FPGAs. Just one parameter is changed in the function call to enable additional FPGAs. It is demonstrated that the approach we have taken is both scalable and re-configurable so other algorithms can be programmed in the hardware when needed by the application. The energy minimization algorithm proved to be a good candidate for hardware acceleration because it did a large amount of calculations in a nested loop on one dataset. On the application side the CPU just had to do a DMA burst of the atom data to the FPGA hardware which took some transfer time and the hardware crunched through the calculation loops until it was finished.

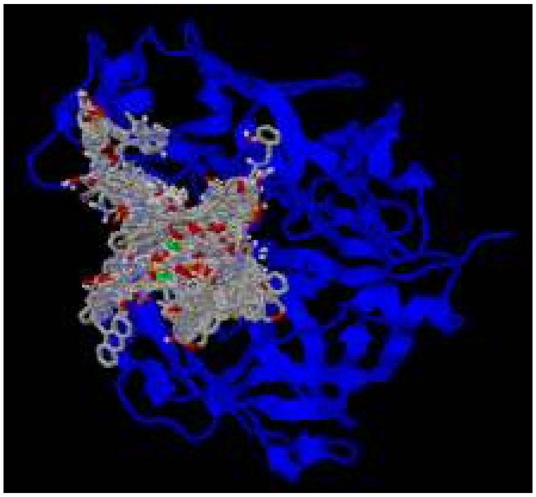


Figure 2 (From: Wolf, Antje)
Small molecules (potential drugs) are shown here positioned into the binding site of a target protein which is associated with a specific disease.

References

[Ursu et al 2006] Oleg Ursu, Mircea Diudea, Shin-ichi Nakayama, “3D Molecular Similarity: Method and Algorithms”, J. Comp Chem. Jpn, Vol. 5, No. 1, 2006, pp 39-46.
Web reference http://www.sccj.net/publications/JCCJ/v5n1/a16/document.pdf

[Halgren 1999] Halgren, Thomas A., “MMFF VI. MMFF94s Option for Energy Minimization Studies”, J. Comp Chem., Vol 20, No 7, April 9, 1999, pp. 720-729

[Casey 2006] Casey, Richard M., “Bioinformatics in Structure-Based Drug Design”
Published: 2006, J. Comput. Chem Jpn, Vol. 5, No. 1, pp. 39-46
Web Reference http://www.b-eye-network.com/view/2593

[Schlick 2006] Tamar Schlick, Molecular Modeling and Simulation, Springer Science 2006, pp. 464-492.

[Popig et al 2006] Daryl Popig, Debra Ryle, David Pellerin, Eric Stahlberg, “Applying Field Programmable Logic Arrays to Biological Problems”, Presented June 2006, OCCBIO Conference, Athens Ohio Web Reference http://www.acceleratedata.com/FPGA_biometrics.pdf

[Wang et al 2006] Mingliang Wang, Xiangqian Hu, David N. Beratan, and Weitao Yang, “Designing Molecules by Optimizing Potentials”, JACS, 2006, Vol 128, No. 10, pp 3228-3232

[Wolf 2006] Antje Wolf, “WISDOM The first biomedical application on EGEE Grid” , Presentation, 2006,Web Reference: http://gks06.fzk.de/slides/Wisdom_Antje_Wolf.pdf

[Alam, Agarwal et al – 2007] Alam, S.R., P.K. Agarwal, M.C. Smith, J.S. Vetter, and D. Caliga, “Using FPGA Devices to Accelerate Biomolecular Simulations,” IEEE Computer, 40/3, 66-73, (2007).