Accelerating Biomedical Research in Designing Diagnostic Assays, Drugs, and Vaccines

The US Department of Defense Biotechnology High-Performance Computing Software Applications Institute for Force Health Protection develops state-of-the-art high-performance computing applications that accelerate biomedical research in the development of diagnostic assays, drugs, and vaccines. The BHSAI works together with DoD life scientists to develop and integrate HPC software applications into DoD biomedical research programs.

The US Department of Defense (DoD) Biotechnology High-Performance Computing Software Applications Institute for Force Health Protection (BHSAI) develops state-of-the art HPC applications to accelerate biomedical research and support the development of diagnostic assays, drugs, and vaccines. At BHSAI, we work with DoD life scientists to develop and integrate HPC software applications into tools that form an integral part of DoD biomedical research programs. We've assembled expert teams in systems biology, bioinformatics, computational chemistry, physiology, and computer science to develop and apply algorithms, tools, and techniques across a broad class of biomedical areas. These teams port and parallelize existing codes from workstations to HPC, develop novel computational chemistry and bioinformatics algorithms, and implement these algorithms to run efficiently on HPC platforms at DoD Supercomputing Resource Centers (DSRCs). BHSAI then transitions the developed software systems to DoD life scientists by providing graphical user interfaces (GUIs) to run applications, modify input and code parameters, and access and visualize generated data.

Our broad impact across the DoD biomedical community is reflected in the range of topics that we address: identification of genomic and proteomic biomarkers,[1,2] bioinformatics-based prediction of protein function,[3,4] virtual high-throughput screening of drug-like compounds,[5] engineering and design of proteins,[6] computational prediction of protein structure for medical countermeasures,[7–9] modeling of chemical reactions in biological systems,[10,11] and bio-inspired detection systems.[12]

Here, we highlight four specific problems and projects where our software systems have proven critical in military biomedical research by providing capabilities that just wouldn't be possible without HPC.

1521-9615/10/$26.00 © 2010 IEEE
Copublished by the IEEE CS and the AIP

Anders Wallqvist, Nela Zavaljevski, Ravi Vijaya Satya, Rajkumar Bondugula, Valmik Desai, Xin Hu, Kamal Kumar, Michael S. Lee, In-Chul Yeh, Chenggang Yu, and Jaques Reifman
US Army Medical Research and Materiel Command

Designing Pathogen Diagnostic Assays

Diagnostic assays let scientists detect and identify pathogens—including biological threat agents—in clinical and environmental samples. Advances in genome sequencing technology led to the availability of many pathogen genome sequences, making sequence-based diagnostic assays an attractive option. The availability of these genomic sequences has further opened opportunities to develop whole genome-based diagnostic assays, such as DNA microarrays and polymerase chain-reaction assays, which offer more flexibility than traditional methods based on a single gene or selected regions within a target genome. Microarray-based pathogen diagnostic assays can test for hundreds or even thousands of pathogens in a single diagnostic test; given this, scientists are widely using them for various diagnostic applications.

A microarray-based diagnostic assay consists of thousands of short DNA (oligonucleotide) sequences attached to a solid surface. These oligonucleotide sequences, or *probes*, are used as fingerprints for identifying pathogens and hence should be unique to the pathogen (or target) genome with respect to all other nontarget genomes. As a result, designing microarray-based pathogen diagnostic assays entails the computationally expensive comparison of target genomes with all available nontarget sequences. Using HPC bioinformatics tools is essential for performing these comparisons in a reasonable amount of time. Although many different methods have been developed to guide pathogen diagnostic assay design, none use HPC resources to design oligonucleotide probes suitable for microarray-based diagnostic assays.

To this end, we developed the GUI-driven HPC-based Tool for Oligonucleotide Fingerprint Identification (TOFI) pipeline, which designs microarray probes for multiple related bacterial and viral pathogens by identifying probes from the input target sequences that are unique with respect to all available nontarget sequences.[1,2] TOFI performs these computations efficiently by

- preprocessing the input sequences and identifying a small set of nonredundant target sequences from which to design fingerprints;
- parallelizing various steps in the probe design process; and
- using the parallel Blast implementation, mpiBlast (www.mpiblast.org), for performing the specificity analysis.

The pipeline scales well as target genomes increase and can potentially design fingerprints for hundreds of related target genomes in a single run. Given a set of target genomes, TOFI finds microarray fingerprints that are unique to any subset of the target genomes with respect to all sequenced nontarget genomes.

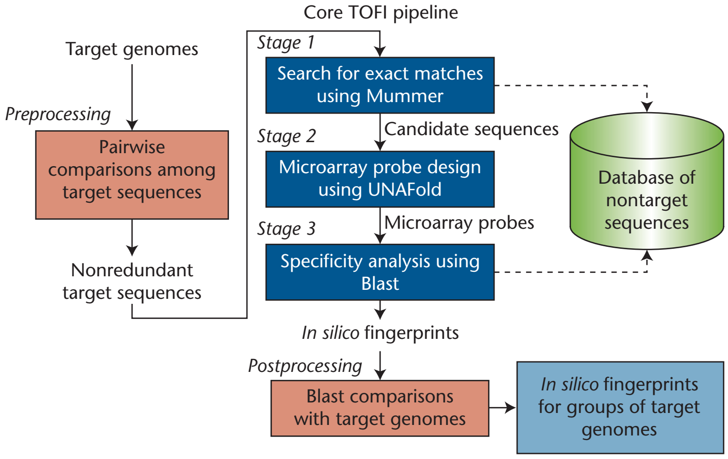
Core TOFI pipeline
Target genomes
Preprocessing
Pairwise comparisons among target sequences
Nonredundant target sequences
Stage 1
Search for exact matches using Mummer
Stage 2
Candidate sequences
Microarray probe design using UNAFold
Stage 3
Microarray probes
Specificity analysis using Blast
Database of nontarget sequences
In silico fingerprints
Postprocessing
Blast comparisons with target genomes
In silico fingerprints for groups of target genomes

Figure 1. Overview of the Tool for Oligonucleotide Fingerprint Identification (TOFI) pipeline. TOFI's preprocessing stage eliminates redundant sequences from the target genomes. The actual fingerprint design process, including the comparison with nontarget genomes, happens in the three core TOFI stages. The post-processing module identifies fingerprints that are common to groups of target genomes.

As Figure 1 shows, the TOFI pipeline consists of the three main stages. However, before the target genomes are submitted to the pipeline's first stage, we preprocess them to build a set of nonredundant target sequences. TOFI uses the suffix tree-based Mummer program to compare the target genomes with each other and eliminate any repeated occurrences of identical DNA segments. This preprocessing step reduces the input target genomes to a set of nonredundant target sequences.

In stage 1, TOFI uses the Mummer program to perform pairwise comparisons of nonredundant target sequences with each nontarget genome. The goal is to eliminate regions in the target sequences that have exact matches with any of the nontarget genomes. TOFI then passes the surviving regions, or *candidate sequences*, on to the pipeline's second stage.

In stage 2, TOFI identifies oligonucleotides of the desired length from the candidate sequences that satisfy DNA microarray experimental conditions, such as melting temperature and GC content. TOFI uses the open source UNAFold software to identify these oligonucleotides.

In stage 3, TOFI performs a Blast search for each probe against a comprehensive sequence database. The Blast comparisons are performed in parallel on multiple cores using the `blastn` program of mpiBlast. TOFI computes each probe's specificity based on multiple, user-selected specificity criteria.[2] Probes with significant alignments

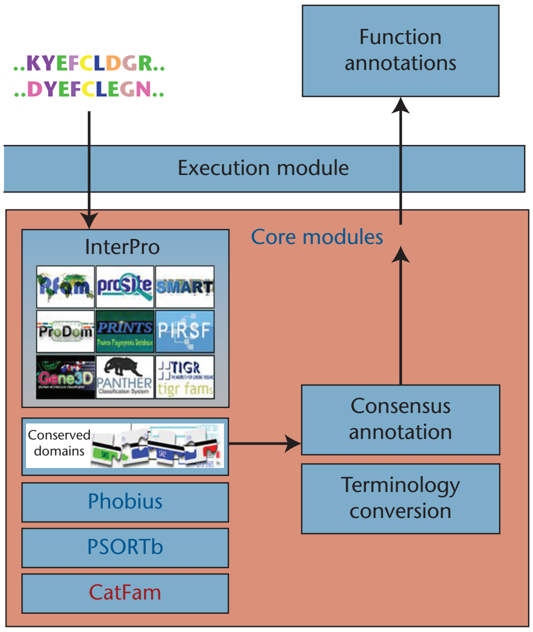

Figure 2. The key modules of Pipeline for Protein Annotation. PIPA's programs are organized into multiple modules. The pipeline execution module consists of programs that enable user access to and control of the pipeline's parallel execution of multiple programs. The execution module wraps the core modules, containing the integrated resources, the terminology conversion program, and the consensus annotation program.

to nontarget genomes are eliminated, and the surviving probes are reported as the *in silico* DNA fingerprints for the target genomes.

All three pipeline stages are implemented in parallel. The first two stages are highly parallelizable and provide linear speedup with increases in the number of cores. In stage 3, the mpiBlast program provides two levels of parallel execution: database fragmentation and query segmentation. With the database fragmentation option, the nontarget database is split into smaller fragments distributed among the computing cores; in the query segmentation, the input queries are distributed.

We installed the TOFI pipeline on a Linux cluster with a distributed memory architecture, on which each compute node consists of two 2.8 GHz quad-core Intel Nehalam processors and 24 Gbytes RAM. When run with a set of eight *Burkholderia* genomes (that is, eight target genomes), the TOFI pipeline took 9 hours using 74 cores and designed 5,015 fingerprints, which included fingerprints unique to each individual genome as well as fingerprints common to multiple *Burkholderia* genomes.[1] Life scientists at the US Army Medical Research Institute of Infectious Diseases (USAMRIID) at Ft. Detrick, Maryland, experimentally validated these fingerprints and found that more than 80 percent identified the intended targets. Importantly, in a one-way blinded test, fingerprints designed to identify common signatures of multiple bacterial strains of the *Burkholderia pseudomallei* species successfully identified a different, unsequenced strain of the same species.[1]

USAMRIID life scientists are using the TOFI pipeline to design diagnostic assays for various viral and bacterial pathogens. In addition, plant pathologists at the US Department of Agriculture, Ft. Detrick, are using it to design fingerprints for various plant pathogens.

Annotating Newly Sequenced Genomes

Identifying proteins and enzymes essential for microbial pathogens to sustain their life or maintain their virulence is the first step in developing effective countermeasures against new strains of pathogens or bioengineered pathogens. Protein and enzyme functions can be quite varied. Among a multitude of other functions, they're responsible for the biosynthesis of microbial cell walls (which form the first line of defense against the human immune system response) and are critical components of the bacterial secretion systems (used in host cell invasion and bacterial toxin production). Traditional experimental methods to determine the functions of proteins encoded in genomic sequences cannot keep pace with the avalanche of sequence data produced by new high-throughput sequencing technologies. This prompted researchers to develop numerous *in silico* approaches for protein function annotation. However, the different approaches' varied function classification terminologies precluded the integration of multisource predictions.

To address these issues, we developed the GUI-driven, HPC-based Pipeline for Protein Annotation. PIPA is a genome-wide protein function annotation pipeline that integrates different bioinformatics resources and uses Gene Ontology (GO), the *de facto* protein function annotation standard, to provide consistent annotation and resolve prediction conflicts.[3]

Figure 2 shows PIPA's modular configuration, which permits easy development of specialized databases and integration of various bioinformatics tools.[3] The first module, the pipeline execution module, consists of programs that let users access and control the pipeline's parallel execution of multiple jobs, each searching a particular database for a chunk of the input data. The execution module wraps the core pipeline modules, which include the integrated resources, the program for terminology conversion to GO, and the consensus annotation program. A special PIPA module provides customized generation of protein function databases. We used that module to construct a database for enzyme catalytic function prediction (CatFam).[4] The current PIPA implementation annotates protein functions by combining the results of CatFam and the results of multiple integrated resources, including the 11-member databases of InterPro and the Conserved Domain Database, into common GO terms.

We validated PIPA's catalytic function prediction based on the CatFam database on a testing set of nearly 20,000 proteins (both enzymes and nonenzymes) not included in the CatFam database generation process and compared them with those of Priam, a similar well-established database. We used precision as a measure of prediction accuracy and recall as a measure of prediction coverage. *Precision* is the fraction of function predictions of a particular method that agrees with the gold standard annotations, whereas *recall* is the fraction of the gold standard function annotations that are predicted by a particular method. For this test, CatFam achieved a precision of 95.9 percent and recall of 97.0 percent compared with Priam's precision of 82.6 percent and recall of 87.9 percent.[4]

To evaluate the performance of PIPA's consensus prediction, we used the 31,589 proteins with annotated GO terms from the Swiss-Prot database. Figure 3 shows a comparison of the performance of GO annotations with and without the consensus algorithm for the GO molecular function category. We obtained the data points corresponding to the consensus algorithm by changing the algorithm's parameters.[3]

For comparison, we created GO annotations without the consensus algorithm by changing the integrated databases' cut-off thresholds. Figure 3 shows the results, which suggest a significant trade-off between precision and recall. However, for a given recall, applying the consensus algorithm yielded higher precision than when consensus wasn't used. The precision of molecular function predictions was improved by up to 8.0 percent. The results suggest that the consensus algorithm effectively integrates different function inferences to improve the precision of GO annotations. The low recall, which indicates a low coverage of GO terms predicted by the pipeline, is likely due to the incompleteness of the GO mappings that link individual databases with GO terms.

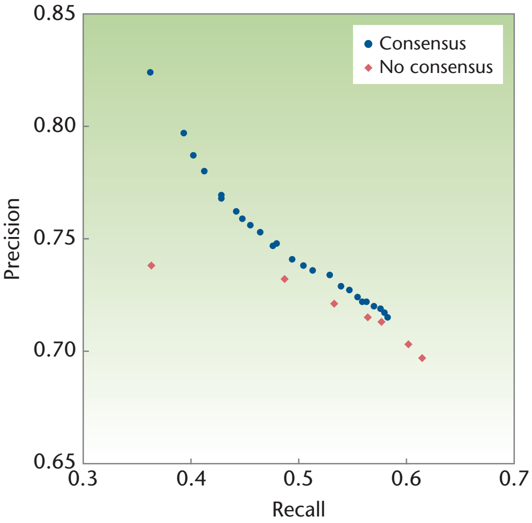

Figure 3. Gene Ontology (GO) consensus evaluation. A comparison of precision and recall, evaluated using GO's hierarchical structure, for GO molecular function annotations with and without consensus. The comparison is based on 31,589 manually annotated proteins.

PIPA is deployed at the DSRC's Army Research Laboratory and the Maui HPC Center. USAMRIID bacteriologists use PIPA to annotate protein functions of a number of pathogens with the ultimate goal of identifying common drug targets. Scientists at the Naval Medical Research Center in Rockville, Maryland, also use PIPA to predict protein functions for newly sequenced bacterial pathogens and their near-neighbor strains to understand phenotypic variations among bacterial species and strains.

Predicting Protein Structures

The number of sequenced genomes has increased dramatically over the past few years. This includes a growing number of genomes from harmful organisms, such as hemorrhagic fever viruses, intracellular pathogens, and parasites. While the number of known protein sequences in these genomes is large and rapidly growing, the total number of known protein structures for all genomes is on the order of tens of thousands, representing less that 1 percent of known protein sequences.

Proteins participate in almost all biological functions, and their 3D atomic structures are essential to understanding their functions at the molecular level. Access to 3D structures of all proteins associated with these potential threat organisms is crucial for rapid *in silico* drug screening and assessment of vaccine development and

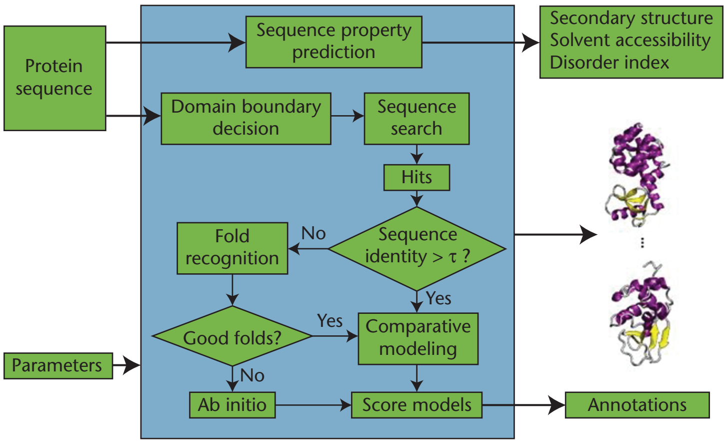

Figure 4. Overview of the Protein Structure Prediction Pipeline. PSPP first predicts the secondary structure, solvent accessibility, protein disorder, and putative domain boundaries of input sequences. At the domain level, it compares input sequences with sequences of proteins with known structures. If no known structures are found to be similar, it uses the predicted secondary structure and other properties of the proteins to find compatible folds in a library of known protein folds. The search results of these two approaches are used to build the models of the input protein using comparative modeling. If both of the previous search attempts are unsuccessful, then the computationally intensive ab initio Rosetta program is used. The ab initio models are annotated by structurally comparing against the library of protein folds.

design. In drug design, candidate small molecules are computationally assessed for their potential to bind to and interfere with the function of specific pathogen proteins. In vaccine design, a protein structure provides insight into potential surface characteristics that can serve as attack points for the immune system and regions that might be modifiable to increase thermal stability and decrease aggregation.

Our goal is to develop a high-throughput software pipeline that can predict 3D atomic structures based on user-supplied sequences at a genome-wide scale.[8] The structure prediction pipeline proceeds in two stages related to the whole protein and to each of the protein's separate structural domains. At the whole protein stage, it predicts the secondary structure, solvent accessibility, and structural disorder index of each amino acid in the sequence. Furthermore, the protein sequence is compared against the Structural Classification of Proteins (SCOP) database of known structural domains to identify sequence boundaries that delineate individual structural domains of the protein.

For each identified domain, we first compare its amino acid sequence against a library of sequences of known protein structures using the standard sequence comparison method PSI-Blast. Next, we compare the predicted sequences of secondary structure and other physicochemical properties against a corresponding library of known structural folds using Prospect II. Using the results of these searches, we build models of the query sequence using comparative modeling (see Figure 4). When the above two searches are unable to find reasonable matches, users can choose to build protein structures by ab initio folding using the Rosetta code, a computationally intensive method. We structurally compare the approach's derived models to a library of known protein structures and annotate them accordingly.

We designed this program to efficiently use HPC resources to both handle the numerous sequences contained in a genome (typically 1,000 to 10,000) as well as process computationally intensive ab initio folding. Harnessing the computational power of the DoD HPC machines, the work of predicting a protein using *de novo* techniques is reduced from months on a single-CPU workstation to less than a day in many cases. Using HPC resources, we can now process a whole genome in one month—a calculation that would have been infeasible using desktop computing.

PSPP has been applied to several problems relevant to USAMRIID DoD investigators. Investigators have used the pipeline in investigations on *Yersinia pestis* (plague) and *Escherichia coli* genomes. Researchers used the theoretical prediction results from the *Y. pestis* genome to determine which proteins to pursue for the first round of a high-throughput protein-protein interaction study. Specifically, they chose proteins with high sequence similarities to known protein structures, such that newly identified protein-protein interactions could be readily modeled.

A second study involved determining the protein structure of VP24, the smallest protein in the Ebola and Marburg virus genomes.[9] This protein was known to bind the human nuclear import protein, importin-α, which interferes with the nuclear transport of transcriptional activator proteins key to the innate cellular immune response. As Figure 5 shows, the results demonstrate that through multiple runs of the ab initio folding module, they were able to predict that the fold type of the central 150 residues of VP24 is in the same fold family as the nuclear import and export proteins.[9]

In another case, we worked with researchers at the US Army Medical Research Institute of Chemical Defense in Aberdeen, Maryland, to predict the structure of human paraoxonase (HuPON1). This protein is a serum enzyme with a broad spectrum of hydrolytic activities, including the hydrolysis of organophosphates (nerve gas), esters, and lactone substrates. Despite intensive site-directed mutagenesis and other biological studies, the structural basis for the specificity of HuPON1's substrate interactions remains elusive.

By applying homology modeling, docking, and molecular dynamic simulations, we obtained a theoretical model of substrate binding and specificity associated with wild-type and mutant forms of HuPON1.[7] We can now apply this knowledge in designing nerve gas bioscavengers based on HuPON1 variants.

Finding Lead Compounds for Drug Development

A key factor in developing small-molecule therapeutics against biological threats is the ability to identify initial lead compounds that can affect the function of proteins or other macromolecules critical to a pathogen. These compounds serve as the starting point for medicinal chemists to optimize their efficacies and pharmacological properties to turn them into drug candidates. Given that the explosive growth of commercial and publicly available chemical databases now provides access to tens of millions of compounds, efficiently evaluating these compounds as potential target inhibitors becomes imperative.

Docking provides a computational method to predict the interaction between a small molecule and a protein. Virtual high-throughput docking is an *in silico* screening method that searches large chemical databases and predicts a molecule's binding conformation and affinity with a protein target. Virtual screening has become an accepted tool in drug discovery. Researchers have successfully applied it in several therapeutic programs at the lead discovery stage, primarily to focus and reduce the number of potential experimental compounds, and thereby save time and resources. We've developed a Docking-based Virtual Screening (Dovis) pipeline[5] based on AutoDock that completely automates the docking process and removes the technical complexities and organizational problems associated with large-scale high-throughput virtual screening.

The Dovis application provides a scalable parallelization scheme for AutoDock that makes it possible to run large-scale virtual screening in parallel on Linux clusters. We made the computational scheme highly efficient by automating load balancing, significantly reducing the file I/O operations, providing outputs that conform to industry-standard file formats, and providing a general wrapper-script interface for rescoring docked ligands.

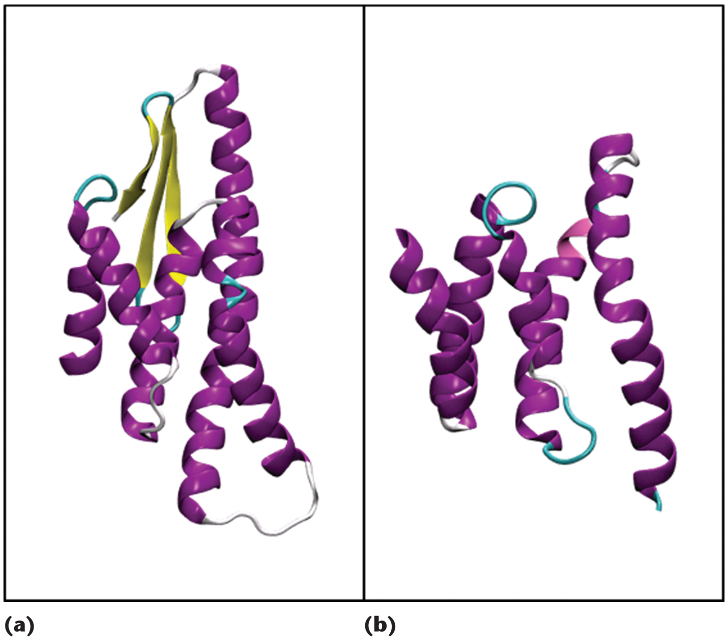

Figure 5. Ab initio prediction from the protein structure prediction pipeline. A side-by-side comparison of (a) the predicted structure of Ebola/Marburg protein VP24 (residues 50–200) and (b) a fragment of exportin (residues 84–190), which binds to importin-α (PDB ID: 1WA5).

Figure 6 shows an overview of the computational elements of Dovis together with an example of a docked ligand in a protein's binding pocket. Researchers can use the software system to screen arbitrary chemical databases, but it currently comes preloaded with a collection of 8.4 million preprocessed small-molecule compounds from the ZINC database. Dovis comes with a GUI for users to specify docking parameters and a protocol to retain the user-specified top percentage of docked ligands based on their docking scores. The GUI also lets users submit docking jobs and query and visualize ligands docked to the target protein.

A high-throughput screening campaign typically has many more ligands (millions) than the number of cores (hundreds) available. Thus, input ligands are partitioned into blocks of *N* ligands, where *N* is specified by the user. During parallel docking, each core copies the energy grids and other required files to its own temporary directory and requests a block of ligands through a file lock mechanism. This ensures that each core gets one unique set of ligands to work with at a time. After completing a block of ligands, each core registers the finished job and updates the top-ranking ligands to the project directory. The core

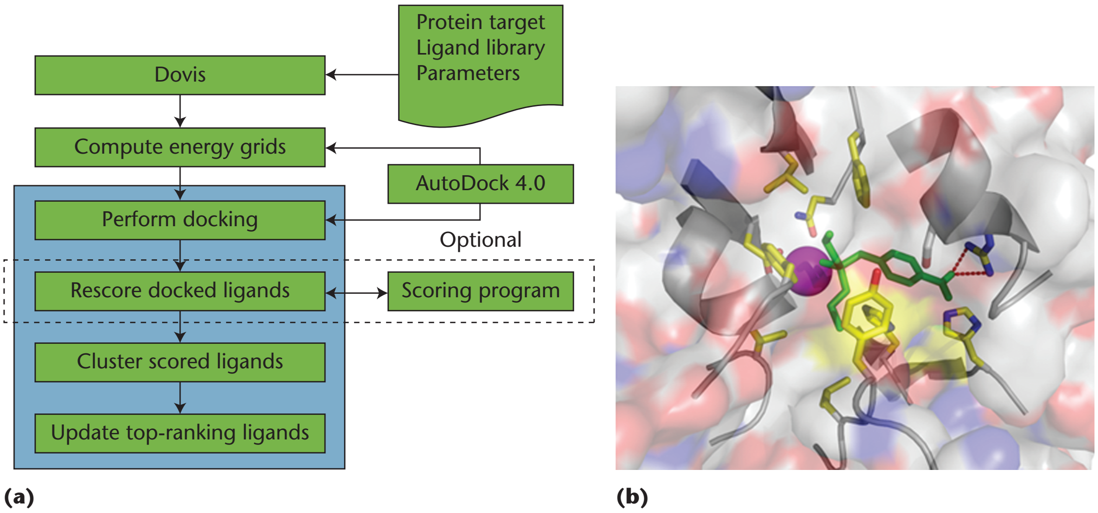

Figure 6. The Docking-based Virtual Screening (Dovis) pipeline. (a) Overview of the Dovis pipeline. In the pipeline's preprocessing steps, the protein target's information and docking parameters are used to set up the reusable energy grids that are passed to individual cores. The small-molecule ligand library is parsed into smaller chunks for processing by multiple instances of the AutoDock program. Alternatively, ligands can be rescored on the fly using auxiliary scoring programs; Dovis then clusters the scored results and updates the master list of top-ranking ligands after processing each chunk of ligands. Calculations in the blue box are run in parallel across the compute cluster. (b) The Dovis prediction of the binding conformation of paraoxon with human paraoxonase (HuPON1). The yellow sticks represent HuPON1's active side residue side chains; the magenta sphere the catalytic calcium ion; and the green stick the substrate paraoxon. The nitrophenyl moiety of paraoxon that's cleaved off by the enzyme forms hydrogen bonding interactions with the positively charged guanidine group of residue R192.

then requests another assignment. This process is iterated until all ligands in a block are exhausted.

Finally, an assessment script verifies whether all assignments were successfully completed. Any requested but unfinished ligand block is added back to the original list of blocks to be reprocessed. Each core works on one ligand block at a time and these blocks are continuously requested by the available cores during a Dovis run. Thus, by using small block sizes when the number of ligands are much larger than the total number of cores assigned to the job, this scheme can provide an effective mechanism for automated, dynamic load balancing.

In the AutoDock program, only one ligand is docked at a time in a docking run, and at each time the associated energy grid files corresponding to every atom type in the ligand are loaded into the corresponding core. Depending on the size of the energy grid and number of atom types, ~10 MB of data are loaded at every docking run for each ligand. Most of the energy grids, however, can be reused from ligand to ligand. Hence, to improve runtime efficiency and reduce file I/O operations, we modified the AutoDock 4.0 source code to load the energy grid files only once, while docking multiple ligands in a single run. In this mode, energy grids of all atom types are loaded and a block of *N* ligands is sequentially docked to a receptor.

By implementing the parallelization scheme to handle larger blocks of ligand data and introducing a new multiple-ligand docking mode of AutoDock, we can efficiently streamline and automate the virtual screening process. We installed the Dovis pipeline on several Linux clusters with distributed memory architecture. For example, on the *mana* cluster at the Maui HPC Center, each compute node consists of two 2.8 GHz quad-core Intel Nehalem processors and 24 GB RAM. The docking of 8.4 million drug-like compounds in the ZINC database to a protein target took a total of 31 days on 128 cores for a throughput rate of 2,116 molecules per core, per day.

The Dovis application has been downloaded more than 714 times and is used in drug discovery projects throughout the world. Dovis is actively used in several projects funded by the DoD Defense Threat Reduction Agency to initially screen for compounds that can bind to and inhibit proteins relevant to biodefense. For example, USAMRIID researchers are using Dovis to initiate drug discovery efforts against the ricin A-chain toxin and the Ebola virus.

Software Access and Availability

All of our pipelines are available to DSRC users through Web-based GUIs, which let users access

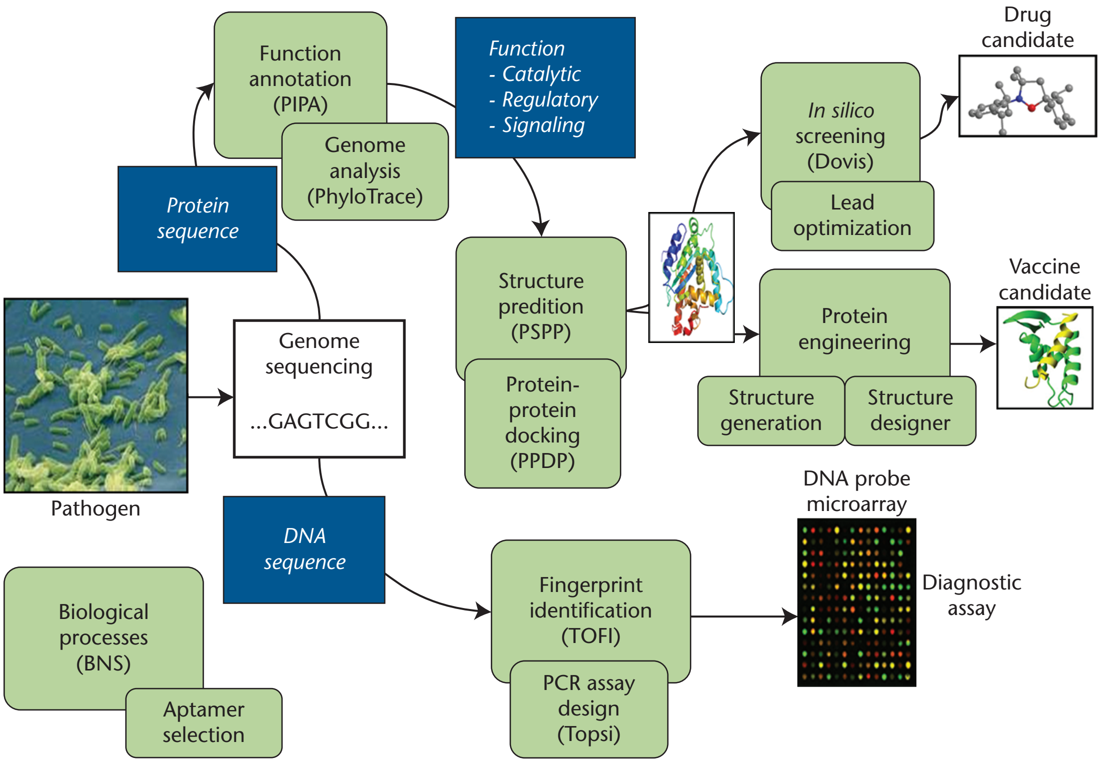

Figure 7. Software portfolio overview. The software developed at the Biotechnology High-Performance Computing Software Applications Institute (BHSAI) ranges from genome sequencing analysis to identifying genomic and proteomic biomarkers (the Tool for Oligonucleotide Fingerprint Identification, or TOFI, and the Tool for Polymerase chain reaction (PCR) Signature Identification, or Topsi, respectively); bioinformatics-based prediction of protein function (Pipeline for Protein Annotation, or PIPA, and the comparative genomics pipeline phylogenomic trace, or PhyloTrace, respectively); computational prediction of protein structure for medical countermeasures (the Protein Structure Prediction Pipeline, or PSPP, and the Protein-Protein Docking Pipeline, or PPDP, respectively); the engineering and design of proteins (the structure generator and the structure designer, respectively); virtual high-throughput screening of drug-like compounds (the Docking-based Virtual Screening pipeline, or Dovis, and lead optimization); and modeling chemical reactions in biological systems and bio-inspired detection systems (the Biomolecular Network Simulator, or BNS, and the Aptamer Selection, respectively).

HPC clusters and run HPC jobs from any Web browser. The GUIs use the user interface toolkit (UIT), which provides standardized functionalities across our software GUIs and secure access to HPC resources for communicating with DSRC HPC clusters. Users are authenticated using their Kerberos credentials via the UIT Web service.

The BHSAI serves as a resource to develop HPC applications to accelerate research and development of military-relevant medical products for Force Health Protection. Our software portfolio and projects in progress span the spectrum of biotechnology applications, ranging from pathogen detection to drug development. Figure 7 shows an overview of all our systems and indicates where they've been applied and integrated into biomedical research. Additional information about each of the BHSAI-developed HPC software applications is available on our website (www.bhsai.org).

DoD life scientists are using all of these software systems in research projects and product development. Our software systems' HPC capabilities now let DoD biomedical researchers systematically advance computational hypotheses that can be investigated using experimental techniques. Working closely with individual research groups and investigators, we've iteratively developed software solutions that support a range of diverse biomedical investigations in several areas:

- Developing diagnostic assays for biological warfare agent detection and identification. *Institute*: USAMRIID. *Software*: TOFI and Topsi.

- Large-scale annotation and comparative analysis of bacterial genomes to identify "universal" protein targets for drug and vaccine. *Institutes*: USAMRIID, Navy Medical Research Center, and Walter Reed Army Institute of Research. *Software*: PIPA and comparative genomics pipeline phylogenomic trace (PhyloTrace).
- Computational structural biology studies to elucidate the 3D structure of human, parasite, bacterial, and viral proteins. *Institutes*: USAMRIID, Walter Reed Army Institute of Research, US Army Medical Research Institute of Chemical Defense, Engineer Research and Development Center. *Software*: PSPP, protein-protein docking pipeline (PPDP), structure generator, and structure designer.
- *In silico* screening of large chemical databases to identify lead compounds against toxins, bacterial, and viral protein targets. *Institute*: USAMRIID. *Software*: Dovis and lead optimization.
- Studies aimed at both understanding how host-pathogen protein-protein interactions promote infection and finding commonalities among bacterial pathways active during infection. *Institutes*: USAMRIID and BHSAI. *Software*: PIPA, PhyloTrace, and PPDP.
- Developing computational pharmacological and toxicological platforms to accelerate drug development. *Institute*: USAMRIID, Walter Reed Army Institute of Research, and BHSAI. *Software*: PSPP, Dovis, and Lead Optimization.
- Assembling bio-inspired machines and toxicant detection systems. *Institutes*: Air Force Research Laboratory and Wright-Patterson Air Force Base. *Software*: BNS, Dovis, and Aptamer Selection.

HPC is transforming how DoD life scientists solve problems and conduct research. Instead of relying solely on laboratory experimentation, they now regularly use HPC simulations to rule out unfeasible solutions, generate testable hypotheses, guide research directions, and significantly increase research efficiency.

Acknowledgments

The US DoD HPC Modernization Program supports this work under the HPC Software Applications Institutes Initiative. The opinions and assertions contained herein are the authors' private views and aren't necessarily the views of the US Army or DoD. This article has been approved for public release.

References

1. R.V. Satya et al., "In Silico Microarray Probe Design for Diagnosis of Multiple Pathogens," *BMC Genomics*, vol. 9, no. 496, 2008; doi:10.1186/1471-2164-9-496.
2. R.V. Satya et al., "A High-Throughput Pipeline for Designing Microarray-Based Pathogen Diagnostic Assays," *BMC Bioinformatics*, vol. 9, no. 185, 2008; doi:10.1186/1471-2105-9-185.
3. C. Yu et al., "The Development of PIPA: An Integrated and Automated Pipeline for Genome-Wide Protein Function Annotation," *BMC Bioinformatics*, vol. 9, no. 52, 2008; doi:10.1186/1471-2105-9-52.
4. C. Yu et al., "Genome-Wide Enzyme Annotation with Precision Control: Catalytic Families (Catfam) Databases," *Proteins*, vol. 74, no. 2, 2009, pp. 449–460.
5. X. Jiang et al., "Dovis 2.0: An Efficient and Easy to Use Parallel Virtual Screening Tool Based on Autodock 4.0," *Chem Cent J.*, vol. 2, no. 18, 2008; doi:10.1186/1752-153X-2-18.
6. I.C. Yeh et al., "Free-Energy Profiles of Membrane Insertion of the M2 Transmembrane Peptide from Influenza A Virus," *Biophysical J.*, vol. 95, no. 11, 2008, pp. 5021–5029.
7. X. Hu et al., "In Silico Analyses of Substrate Interactions with Human Serum Paraoxonase 1," *Proteins*, vol. 75, no. 2, 2009, pp. 486–498.
8. M.S. Lee et al., "PSPP: A Protein Structure Prediction Pipeline for Computing Clusters," *PLoS ONE*, vol. 4, no. 7, 2009, e6254; doi:10.1371/journal.pone.0006254.
9. M.S. Lee, F.J. Lebeda, and M.A. Olson, "Fold Prediction of VP24 Protein of Ebola and Marburg Viruses Using de novo Fragment Assembly," *J. Structural Biology*, vol. 167, no. 2, 2009, pp. 136–144.
10. F.J. Lebeda et al., "Onset Dynamics of Type A Botulinum Neurotoxin-Induced Paralysis," *J. Pharmacokinet Pharmacodyn*, vol. 35, no. 3, 2008, pp. 251–267.
11. J.M. Frazier, Y. Chushak, and B. Foy, "Stochastic Simulation and Analysis of Biomolecular Reaction Networks," *BMC Systems Biololgy*, vol. 3, no. 64, 2009; doi:10.1186/1752-0509-3-64.
12. Y. Chushak and M.O. Stone, "In Silico Selection of RNA Aptamers," *Nucleic Acids Research*, vol. 37, no. 12, 2009; doi:10.1093/nar/gkp408.

***Anders Wallqvist** is deputy director of the US DoD Biotechnology HPC Software Applications Institute. His research interests are in computational chemistry, bioinformatics, systems biology, and computational drug design. Wallqvist has a PhD in chemical physics from Columbia University. He is a member of the American Chemical Society, the Biophysical Society, and the Protein Society. Contact him at awallqvist@bioanalysis.org.*

***Nela Zavaljevski** is a research scientist at US DoD Biotechnology HPC Software Applications Institute.*

Her research interests are in bioinformatics, data mining, and systems biology. Zavaljevski has a PhD in nuclear engineering from the University of Cincinnati. Contact her at nelaz@bioanalysis.org.

Ravi Vijaya Satya is a research scientist at US DoD Biotechnology HPC Software Applications Institute. His research interests are in combinatorial optimization, and computational biology. Satya has a PhD in computer science from the University of Central Florida, Orlando. He is a member of IEEE and the International Society for Computational Biology. Contact him at rsatya@bioanalysis.org.

Rajkumar Bondugula is a research scientist at US DoD Biotechnology HPC Software Applications Institute. His research interests are in systems biology, structural bioinformatics, and machine learning. Bondugula has a PhD in computer science from the University of Missouri-Columbia, Columbia, Missouri. He is a member of the IEEE Engineering in Medicine and Biology Society and the IEEE Computational Intelligence Society. Contact him at rbondugula@bioanalysis.org.

Valmik Desai is a software developer at US DoD Biotechnology HPC Software Applications Institute. His research interests are in bioinformatics, high-performance computing, Web technologies, and databases. Desai has an MS in computer science from Wayne State University in Detroit, Michigan. Contact him at valmik@bioanalysis.org.

Xin Hu is a research scientist at US DoD Biotechnology HPC Software Applications Institute. His research interests are in computational chemistry and computational drug design. Hu has a PhD in pharmaceutical science from North Dakota State University. He is a member of the American Chemical Society. Contact him at xhu@bioanalysis.org.

Kamal Kumar is a senior software developer at US DoD Biotechnology HPC Software Applications Institute. His research interests are in bioinformatics, computational chemistry, and computational drug design. Kumar has an MS degree in computer science from Purdue University. Contact him at kamal@bioanalysis.org.

Michael S. Lee is a research scientist at US Army Research Laboratory. His research interests are in protein structure prediction and computational chemistry. Lee has a PhD in theoretical chemistry from the University of California, Berkeley. Contact him at michael.scott.lee@us.army.mil.

In-Chul Yeh is a research scientist at US DoD Biotechnology HPC Software Applications Institute. His research interests are in computational biophysical chemistry. Yeh has a PhD in chemistry from the University of North Carolina, Chapel Hill. He is a member of the American Chemical Society and the Biophysical Society. Contact him at icy@bioanalysis.org.

Chenggang Yu is a research scientist at US DoD Biotechnology HPC Software Applications Institute. His research interests are in bioinformatics, data mining, and machine learning. Yu has a PhD in nuclear engineering from the University of Cincinnati. He is a member of the International Society for Computational Biology. Contact him at cyu@bioanalysis.org.

Jaques Reifman is scientific director of the US DoD Biotechnology HPC Software Applications Institute. His research interests are in physiological data modeling, systems biology, bioinformatics, genomics, and proteomics. Reifman has a PhD in nuclear engineering from the University of Michigan, Ann Arbor. He is a member of the American Nuclear Society, American Telemedicine Association, and the International Society for Computational Biology. Contact him at jaques.reifman@us.army.mil.