UCRL-ID-150580

FY02 CBNP Annual Report Input: Bioinformatics Support for CBNP Research and Deployments

T. Slezak

October 31, 2002

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

FY02 CBNP Annual Report Input

Bioinformatics Support for CBNP Research and Deployments

Tom Slezak
Lawrence Livermore National Library
Slezak@llnl.gov

Murray Wolinsky
Los Alamos National Laboratory
murray@lanl.gov

Co-Investigators:

Shea Gardner, Linda Ott, Tom Kuczmarski, Ed Miller, Mark Wagner, Elizabeth Vitalis, Adam Zemla, and Carol Zhou
Lawrence Livermore National Laboratory

Karla Atkins, Tom Brettin, Rob Leach, Jian Song, Charlie Strauss, David Torney and Electra Sutton
Los Alamos National Laboratory

Project Objectives

The events of FY01 dynamically reprogrammed the objectives of the CBNP bioinformatics support team, to meet rapidly-changing Homeland Defense needs and requests from other agencies for assistance:

- Use computational techniques to determine potential unique DNA signature candidates for microbial and viral pathogens of interest to CBNP researcher and to our collaborating partner agencies such as the Centers for Disease Control and Prevention (CDC), U.S. Department of Agriculture (USDA), Department of Defense (DOD), and Food and Drug Administration (FDA)
- Develop effective electronic screening measures for DNA signatures to reduce the cost and time of wet-bench screening
- Build a comprehensive system for tracking the development and testing of DNA signatures
- Build a chain-of-custody sample tracking system for field deployment of the DNA signatures as part of the BASIS project.
- Provide computational tools for use by CBNP Biological Foundations researchers

FY02 Progress

Several bioinformatics developers were part of 3 separate BASIS deployments in FY02, spending nearly 3 calendar months running intensive pathogen-detection surveillance. During this time they added advanced error detection/prevention capabilities to the sample tracking software that drastically reduced user input errors and the need for

developers to be on call at all times. Some deployments involved intensive training of local staff, and subsequent continuing on-call remote support.

Urgent requests for new assays throughout the year kept the team busy running the DNA signature pipeline and analyzing results. Collaborators at USAMRIID and CDC requested additional orthopox detection assays. A set of 53 computationally-derived assays was screened at USAMRIID and 45 proved positive against their orthopox panel. Subsequent screening at the CDC's BSL-4 facility yielded a large increase in the nation's supply of solid assays for variola, vaccinia, and other orthopox viruses that can affect humans. The whole-genome assay development methods highlighted unique gene regions that were unknown to the disease experts, who requested further information about the signatures. A comprehensive signature-annotation effort was launched, including advanced protein structure modeling based on homology templates. This information has been delivered to the collaborators for further study. As an example of the speed of the CBNP signature development system, signature candidates for varicella zoster (chickenpox, a rule-out assay) were delivered in less than 36 hours from when a request was made. Details on assay development and screening statistics can be found in Dr. Paula McCready's section on Assay Development. Signature candidates have been generated computationally for all major threat list pathogens for which adequate genomic sequence is available. The work on rapid development of nucleic acid diagnostics received one of only two Science and Technology Awards given at LLNL for 2002.

Late in the fiscal year CBNP received supplemental funding that was used to purchase computational and storage infrastructure at both LANL and LLNL. The funding was also used to prototype a more automated version of the DNA signature pipeline. Named *KPATH* (as in a radio station: "All pathogens, all the time") it can automatically download new or updated pathogen sequence data from a variety of public databases, and maintain signature sets for pathogens of interest. At year-end, the system was being verified in parallel with the original semi-automated pipeline. Utilizing the new powerful hardware, the time to create a signature set for a large bacterial pathogen has been reduced to less than 4 hours, compared to 24 hours in our original implementation. Many other improvements have been made in selection of unique signature candidates, automated primer design, and electronic screening of potential candidates. A significant breakthrough this year has been obtaining early-release access to a new multi-genome alignment tool, MGA, from Stefan Kurtz, an algorithms collaborator in Germany (http://www.techfak.uni-bielefeld.de/~kurtz/) This tool allows us to align multiple genomes from large 5Mb bacterial genomes, a 100x improvement in genome size over prior algorithms. MGA has been integrated into the *KPATH* system and can now take full advantage of situations where multiple completed genomes are available for one species, or multiple genomes for one family-level signature set. Other improvements this year include the ability to deal with consensus-degenerate primers (necessary for some single-stranded viruses where mutation rates are high) and specialized tools for determining potential signatures in pathological cases where standard techniques come up empty.

The widespread use of the BASIS system allowed all potential federal agencies to see CBNP's capabilities first hand during FY02. The success of computationally-derived

assays in continual field use prompted numerous collaborations and gave the CBNP program a reputation as leading the field of DNA assay development. During FY02 it was decided to leverage the DNA signature work and determine if a similar speed-up could be achieved for protein signature development. Team member Adam Zemla's protein structural homology modeling was tested on proteins that contained DNA-signature targets for Foot and Mouth Disease Virus (FMDV), West Nile Virus (WNV), and variola. WNV was chosen for our ability to screen signature candidates locally (via collaborators at UC Davis) Unique DNA and protein sequence regions can be modeled onto the protein structure model. Working with Rod Balhorn, a set of antibodies to the WNV envelope glycoprotein have been chosen from this computational analysis and will be tested in early FY03. Protein annotation specifically relevant to automated signature determination has been prototyped by Carol Zhou and will be an integral part of all work in this area.

Our effort continued development of several novel bioinformatics tools, and successfully employed these tools to assist CBNP researchers. The genome viewing tools (BugBrowser and Comparative BugBrowser) were substantially enhanced and used to help analyze the *B. anthracis* Ames chromosome. Additional query features were added in response to researcher request. A new tool called Comparative BugBrowser has been developed. It visualizes up to four genomes simultaneously on its circular map and up to six genomes on its linear map. An architecture using a loosely-coupled data server was developed and employed for both genome viewer applications: this architecture allows us to use data from a variety of internal and external sources without modifying the source code. In particular, the tool can now import GenBank data, Distributed Annotation System (DAS) data, and Tandem Repeat Finder data. The ability to accept data from multiple sources is especially critical for the comparative tool, because of the knowledge that can be gained through comparing genomes from different repositories. The data server is integrated with BLAST: if the server gets a request for specific alignment data, and if that data has not been pre-computed, the server automatically generates FASTA files, builds BLAST databases, executes BLAST, and archives the results. BLAST parameters are selected by the user and tracked by the software such that multiple versions of a set of alignments are preserved and reused according to specified combinations of parameters.

The phylogenetic tree visualization tool was substantially enhanced and used to support signature development. In particular, the tool now displays apomorphy data for parsimony trees to facilitate a polymorphism-based approach to signature development. The apomorphy data tells the user which characters in a sequence distinguish an organism from its nearest relatives. Scott White and Rich Okinaka view this as an important contribution to the signature development process. As new sequence data for threat pathogens starts to arrive, interest in use of this capability is increasing. An algorithm to compute inter-taxa distances was designed and implemented in PhyloVis in early FY02. Recently, it has been used to generate data for statistical work being performed by David Torney and Scott White. Another large data set will soon be available for analysis.

Software for an automated pipeline for identifying VNTR candidates has developed: During the process of identifying VNTR candidates for both *B. anthracis* and *Y. pestis,* we developed a pipeline for VNTR identification for any microbial genomes. We can use both contig and complete genome sequences for VNTR analysis. This pipeline allows us to identify all possible VNTR candidates in 24 hours and also provide primer design and biological significance analysis for all identified VNTR candidates.

A similar capability was developed for SNP detection on any closely related species using complete or contig sesquences. Through the SNP analysis performed on *Burkholderia* spp. and *Bacillus* spp., we developed an automated pipeline that allow us to perform high-throughput SNP analysis on any newly sequenced genome (either contig or finished genome sequences) within a week. Currently the rate-limiting step is still the initial Blast. We hope the new Linux cluster purchased through CBNP supplemental funds will allow us to perform a complete SNPs analysis on a whole genome within 2 days.

Immediately after the September 11 terrorist action, the external web sites at LANL were taken down and replaced with a behind-the-firewall threat pathogen web site to support internal CBNP researchers. Currently the internal web site includes *B. anthracis* (pXO1 and pXO2 plasmids), *Y. pestis* (CO92 chromosome and pCD1, pMT1 and pPCP1 plasmids), and *Y. enterocolitica* (pVVe8081 plasmid).

In addition to the above tool development, a substantial volume of annotation and analysis for signature development was performed. Full annotation of *Yersina pestis* has been completed although final quality check is yet to be done. Currently we have 4090 ORFs in our database while Sanger database has 4012 ORFs. For preliminary analyses, we created various specialized tables (e.g., DNA repeat, functional catergories, whole genome comparison, surface proteins, etc.) to summarize some of the features of *Y. pestis*. More specialized comparative analyses are still needed. However, we have shifted our annotation efforts towards complete annotation of *B. thringiensis* 97-27 and *B. thringiensis* Al Hakam.

The *B. anthracis* Ames strain annotation has continued: 3647 new genes have been added, for a total of 5344 gene records. Currently 3406 gene records (out of this 5344) have been fully annotated. Annotation of *B. thringiensis* 97-27 and *B. thringiensis* *Al Hakam* has been added and will be complete by the end of calendar year 2002.

Signature development activities included an analysis of tandem repeats in both *Y. pestis* and *B. anthracis* genomes (chromosome and plasmids). This repeat analysis identified:

- For *Y. pestis*: (1) 43 MLVA markers used by Paul Keim's laboratory for *Y. pestis* strains/isolates identification onto specific locations in our *Y. pestis* genome sequence database. This allows them to better understand the biological significance of the 43 MLVA markers used in their assays. The results of this analysis were made available to them through a password protected web page

- 615 new VNTR candidates for experimental screening. All of the 615 VNTR candidates were mapped onto specific genes or/and intergenic locations.
- For *B. anthracis*: (1) 25 VNTR candidates were identified on the two large virulence plasmids (11 on pXO1 and 14 on pXO2). All of the 25 VNTR candidates were mapped to specific genes or/and intergenic locations in our sequence databases. The results of this analysis are available on our internal site.
- 186 VNTR candidates were identified on the chromosome.

For all of these signature candidates, primers to amplify all the identified repeats designed: Primers for PCR screening of the 615 identified VNTR candidates in *Y. pestis* are designed and the expected sequences and amplicon sizes are also predicted. Results of this analysis are available at our internal site, on request.

Similar analysis for SNPs was performed on *Burkholderia pseudomallei* and *Bacillus ahthracis* and their related species/strains. We performed SNPs analysis on *Burkholderia mallei and B. pseudomallei* in collaboration with Rich Okinaka and on *B. anhracis* Ames and *B. thringiensis* 97-27 in collaboration with Scott White. Numerous SNPs candidates have been identified and made them available at our internal site.

We developed a relational database to store, access, and manipulate a massive amount of 16S rDNA sequence data accumulated from backgrounds studies. This database is ready to be used by Kuske's group.

Finally, a 252-node Linux cluster was acquired and our software is in the process of being ported to improve performance and develop new capabilities, particularly in the area of protein structure prediction.

Future Outlook

- Continued development of the *KPATH* system for automated development and maintenance of pathogen DNA signatures will result add numerous features including better handling of incomplete genomes, automatic notification of signature erosion, and improved utilization of consensus-degenerate primers. The team is working with collaborator Stefan Kurtz to define a new version of MGA that can align both complete genomes and sequence fragments from Genbank, to take full advantage of available strain diversity information. Many high-priority genomes still lack adequate sequence, which sometimes makes manual signature determination necessary. As collaborators begin requesting signatures from fungal pathogen genomes in FY03 further challenging scaling issues are anticipated.
- Initial work on automating protein signature development using protein structure homology modeling will be scaled up to a full proof-of-principle prototype for antibody development. Additional work will be done in conjunction with CBNP ligand design experts to determine if the current approach for DNA signatures can be

leveraged to dramatically reduce the time to locate good targets for high affinity ligands.

- Collaborations with other agencies for pathogen detection assays will be pursued at high priority to ensure that CBNP remains the collaborator of choice for rapid, cost-effective development of reliable DNA and protein detection assays.
- As the BASIS effort evolves into its next phase, continued development will be required for improved sample tracking with flexibility to handle multiple simultaneous deployments with arbitrary mixes of technologies, instruments, and protocols. The team is ready for proposals to scale up for a nation-wide system of pathogen detectors (air, water, food, agriculture).
- Current support for microbial forensics will begin to mature as the new Department of Homeland Security is established. CBNP has a real opportunity to help establish the information architecture for integrating all the data required for a comprehensive forensics data system for law enforcement.
- The tools needed to support a polymorphism-based approach to signature development warrant further development. This work will continue and extend the initial work on VNTRs and SNPs and allow more general polymorphisms to be used as signature elements as well. This computation-intensive work will take advantage of our new Linux cluster.
- Annotation is crucial to developing Phase III signatures. CBNP annotation processes are becoming increasingly efficient and the contents of this annotation are vital in establishing the functional properties of signature candidates.
- We also believe that protein structure prediction is necessary to support annotating the large number of hypothetical conserved proteins that annotation reveals:

Sidebar

CBNP Bioinformatics staff used their unique computational tools to design DNA signatures on very short notice for multiple bacterial and viral pathogens. Their signature development database tracked all the local wet-lab screening, and their BASIS field lab system tracked the completed signature on multiple deployments throughout the year. Validated signatures are now entering the public health system through our close collaboration with the CDC. Several of the staff spent many months on extended deployments to meet the extraordinary demands of FY02, a year in which they were involved in every step from genomic sequence analysis to providing bio-security for multiple cities.

Demonstrations/Exercises/Field Tests

The bioinformatics team was involved in the BASIS 2002 Winter Olympics and the NBDI testbed scheduled demonstrations as well as 2 unscheduled field deployments related to Homeland Security following the September 11 attacks.

Project Citations

J.P. Fitch., S.N. Gardner, T.A. Kuczmarski, S. Kurtz, R. Meyers, L.L. Ott, T.R. Slezak, E.A. Vitalis, A.T. Zemla, and P.M. McCready, "Rapid Development of Nucleic Acid Diagnostics", IEEE Proceedings (in press for Nov. 2002)

J.P. Fitch, B.A. Chromy, C.E. Forde, E. Garcia, S.N. Gardner, P. Gu, T.A. Kuczmarksi, C. Melius, S.L. McCutchen-Maloney, F.M. Milanovich, V. L. Motin, L.L. Ott, A. Quong, J. Quong, J.M. Rocco, T. R. Slezak, B.A. Sokhansanj, E. A. Vitalis, A. T. Zemla, and P. M. McCready, "Biosignatures of pathogen and host," *Proc. IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, NC, Oct. 12-13, 2002. (UCRL-JC-149741)

DoD Genome Sequence for Defense – Poster presentation "Microbial Pathogen Sequence Databases", December, 2001

2nd ASM/TIGR Conference on Microbial Genome – Poster presentation "Annotation and Analysis of *Bacillus anthracis* Virulence Plasmids", January, 2002

Graphics Captions

(1. Stylized photo of some of the *KPATH* developers with computer server. A hard copy of this photo is hanging on Beth George's wall. She may want the electronic copy even if it isn't determined to be useful for the annual report.)

A team of 6 summer students at LLNL (back for their 2nd, 3rd, or 4th year) worked with Tom Kuczmarski (in photo) and others under the direction of Tom Slezak to build the *KPATH* system for automated DNA signature development. The new 24-CPU compute server in the photo was acquired with mid-year supplemental funding and is the heart of the current computational work on DNA and protein signature development.

(2. GIF image of protein structure with DNA signature highlighted.)

Several unique DNA signatures for variola (smallpox) were determined using the *KPATH* algorithms in FY02. The photo shows a structural model of one protein containing a DNA signature region on a gene not known by smallpox experts to contain something unique to that disease. The model was determined by Adam Zemla's structural homology codes, which were developed under LDRD funding in prototype form at LLNL in FY02. The DNA unique signature region is highlighted in blue and red, and was analyzed at the protein sequence level. The amino acid regions shown in blue were found to be common to all orthopox family members (monkeypox, cowpox, camelpox, variola, etc.) while the central red region containing a hair-pin turn is specific to variola. The green region is the most highly conserved sequence motif of the protein family. Immediately adjacent to our DNA signature region, it is critical for protein function. This example shows our ability to combine analyses of both DNA and protein sequence and project them onto structural models that we can create using unique structural homology approaches. It also illustrates how research focused on finding unique pathogen detection signatures can augment the fundamental knowledge of much-studied pathogens. This is one of several variola

signature regions that are now being examined by subject experts at the CDC and USAMRIID.