LLNL-TR-406492

# DHS-STEM Internship at Lawrence Livermore National Laboratory

B. Feldman

August 21, 2008

## Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-406492

# DHS-STEM:
# Internship at Lawrence Livermore National Laboratories

Benjamin Feldman
Under mentorship of Tom Slezak
Summer of 2008

This summer I had the fortunate opportunity through the DHS-STEM program to attend Lawrence Livermore National Laboratories (LLNL) to work with Tom Slezak on the bioinformatics team. The bioinformatics team, among other things, helps to develop TaqMan and microarray probes for the identification of pathogens. My main project at the laboratory was to test such probe identification capabilities against metagenomic (unsequenced) data from around the world. Using various sequence analysis tools (Vmatch[1] and Blastall[2]) and several we developed ourselves, about 120 metagenomic sequencing projects were compared against a collection of all completely sequenced genomes[*] and Lawrence Livermore National Laboratory's (LLNL) current probe database. For the probes, the Blastall[2] algorithms compared each individual metagenomic project using various parameters allowing for the natural ambiguities of *in vitro* hybridization (mismatches, deletions, insertions, hairpinning, etc.). A low level cutoff was used to eliminate poor sequence matches, and to leave a large variety of higher quality matches for future research into the hybridization of sequences with mutations and variations. Any hits with at least 80% base pair conservation over 80% of the length of the match. Because of the size of our whole genome database, we utilized

[1] Vmatch is a sequence analysis tool developed by Stefan Kurtz. http://www.vmatch.de
[2] Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. (1997). 'Gapped BLAST and PSIBLAST: A new generation of protein database search programs', Nucleic Acids Res., Vol. 25(17), pp. 3389–3402
[*] This is a collection from several public genomic databases (NCBI, Sanger, JCVI/TIGR, JGI, etc.).

the exact match algorithm of Vmatch[1] to quickly search and compare genomes for exact matches with varying lower level limits on sequence length.

I also provided preliminary feasibility analyses to support a potential industry-funded project to develop a multiplex assay on several genera and species. Each genus and species was evaluated based on the amount of sequenced genomes, amount of near neighbor sequenced genomes, presence of identifying genes - metabolistic or antibiotic resistant genes - and the availability of research on the identification of the specific genera or species. Utilizing the bioinformatic team's software, I was able to develop and/or update several TaqMan probes for these and develop a plan of identification for the more difficult ones. One suggestion for a genus with low conservation was to separate species into several groups and look for probes within these and then use a combination of probes to identify a genus. This has the added benefit of also providing subgenus identification in larger genera.

During both projects I had developed a set of computer programs to simplify or consolidate several processes. These programs were constructed with the intent of being reused to either repeat these results, further this research, or to start a similar project. A big problem in the bioinformatic/sequencing field is the variability of data storage formats which make using data from various sources extremely difficult. Excluding for the moment the many errors present in online database genome sequences, there are still many difficulties in converting one data type into another successfully every time.  Dealing with hundreds of files, each hundreds of megabytes, requires automation which in turn requires good data mining software. The programs I developed will help ease this issue and make more genomic sources available for use. With these programs it is extremely easy to gather the data, cleanse it, convert it and run it through some analysis software and even analyze the output of this software. When dealing with vast amounts of data it is vital for the researcher to optimize the process – which became clear to me with only ten weeks to work with.

Due to the time constraint of the internship, I was unable to finish my metagenomic project; I did finish with success, my second project, discovering TaqMan identification for genera and species. Although I did not complete my first project I made significant findings along the way that suggest the need for further research on the subject. I found several instances of false positives in the metagenomic data from our microarrays which indicates the need to sequence more metagenomic samples. My initial research shows the importance of expanding our known metagenomic world; at this point there is always the likelihood of developing probes with unknown interactions because there is not enough sequencing. On the other hand my research did point out the sensitivity and quality of LLNL's microarrays when it identified a parvoviridae infection in a mosquito metagenomic sample from southern California. It also uniquely identified the presence of several species of the adenovirus which could mean that there was some archaic strain of the adenovirus present in the metagenomic sample or there was a contamination in the sample, requiring a further investigation to clarify.

This project could also provide preliminary research into the ability of using complex samples (earth, water, etc.) in the detection process. As of now, most samples are from humans – blood, saliva, excrement – or in an air sample but with more research the possibility exists of taking a sample of anything and testing to see if it contains a pathogen. Although my research indicates the need for more sequencing, with the development of some analysis software it could be possible to create some buffer between false positives and true positives. True, it is less exact and may not be put into practice any time soon, but the possibility does exist at this point for such a program. This program would be of invaluable assistance to the government to provide point-of-attack assessment of pathogens without wasting as much time in collection, preparation and purification of samples; results would be able to be extrapolated much quicker.

The second project has the potential to lead to a publication, but at this point it is still too early to say with certainty. There is a stronger possibility that in collaborations with LLNL and DHS a publication could be written about the research of my first project and the unique results I discovered. But a more important aspect of this paper would be the potential for future research into a program to analyze metagenomic samples to distinguish between false positives and false negatives. It would also work great as a collaborative paper with some of the wet laboratory scientists who work with purifying samples for probe identification. They are using techniques like buffering out common pathogens in samples to create a simpler sample. With in-lab data, parameterization of a diagnostic program would be more complete and accurate.

I believe that this internship experience has had profound influence on my future research career. I applied for and participated in this internship with the intent that it would not only introduce me into the field of bioinformatics but into the general profession of research. I have spent most of my life learning the basics in school, but it is an entirely different thing to be able to converse with other scientists on a daily basis, develop and implement ideas and actually utilize the information from my education. In fact this was a more fruitful educational experience than my undergraduate career thus far. It is easy to loose track of the implementation of my schoolwork when it’s class after class, test after test, but being able to visit out at LLNL and work with Tom Slezak, allowed for me to recognize the big picture and where I fit within it. Tom provided me a very flexible environment where I could work on any project with whomever I wished so that I could find my niche within the large field of bioinformatics. I met and discussed protein modeling with some of the best in the field, attended weekly lectures from scientists around the laboratory describing their current and future work, how students could get involved and the possibilities for someone in their field. Initially, I was overwhelmed with the warm environment at the laboratory, some of the laboratory's top employees would take time off to come and give lectures to the interns, people like Ben Santer, Ed Moses, and the director of the Laboratory, George Miller. I particularly took interest in several of the fields including nonproliferation and dynamic network analysis. The laboratory environment made it extremely easy for me to contact people in both departments and set up meetings to discuss their fields in more details and what I would have to do and learn to move into such a field. One such meeting with Tina Eliassi-Rad helped me change and redesign the rest of my undergraduate course load. Per her advice, I am now in discussions with several prominent professors at my university in order to set up some classes or research, these are people at my own university I never knew existed before she pointed them out to me. I am as unsure of my future as when I started college, but this laboratory experience showed me areas of research and science that I had never thought to look into. Now however, I have the connections with the laboratory and the awareness of these sciences to make well informed decisions about my future career.

Once Tom Slezak and I discussed and designed my project I really didn't see myself having any involvement with the actually laboratory work, in fact I saw myself interacting among all computer scientists. But I was pleasantly surprised with the tight interaction between the scientists who design the probes and do the computer-based research and the scientists who run the experiments and interpret the results. During weekly meetings I could see firsthand how experiments were run, what they meant and how our future work would be changed.

The way in which the bioinformatics team took me in as one of their own was almost shocking at first. Being an undergraduate in a sea of undergraduates at my university, professors and administrators can be kind of cold at times and standoffish. But at the laboratory I was accepted almost immediately as a colleague. There was a presence of respect between the team and I, that not only made working there easier but allowed me to be more open with them about and problems and questions I had. Often when I had a problem or a question my co-workers would put aside time in their busy schedules to help me out or point me to someone else that could. In any other environment I'm not sure how well I would have performed, the services my team provided was invaluable for my work and my enjoyment of the summer. At meetings they would ask for my opinion about any and all issues, even ones that were clearly above my head. At one meeting of the whole department, the administrator leading the discussion was calling on interns as much as full time scientists to give input. When I presented my research, they were genuinely interested in my work and asked questions and followed up on them.

It was an interesting environment at the laboratory this year with the recent change in ownership of LLNL; I would not say that this hurt my internship in anyway but to some degree added to it. I witnessed several individuals from the laboratory leave and form companies and organizations to further their work in the private sector. I rarely see such ventures since I attend a strongly research-based university where few go off and work in the industry. In talks with my mentor and others I realize that a career in research at a laboratory doesn't have to end there, it is easy enough, and to some degree common, for people to leave and start companies, or to do both at the same time. This was also my first time working in connection with the government and I got to see the intricacies of how a national laboratory must operate to do so. It is a difficult process to elicit funds, grants and other compensation and I have never really seen firsthand how it works. It is a part of research that they don't really teach you in school, but nevertheless is just as important. I wouldn't say I have the skills yet to go out and apply for grants, however I am beginning to see how it must be done and what skills I will have to learn and develop to make it in research.

Another great thing while working at the laboratory was not the actual networking but learning how to network. LLNL has so many domestic and foreign collaborators that all work together in a symbiotic relationship. There were so many times when in a discussion or meeting someone would bring up a foreign source of samples, or information, or research that they had been talking to. Before coming out to Livermore, I had just assumed that it being a government facility that all the research would be kept top secret and secluded from everyone else, but in reality that is far from the truth. There are so many people from around the world working on such diverse projects that could benefit from another's and it is nice to see Lawrence Laboratories partaking in this useful network. Being an intern, I was naturally tentative to go out and network with all these world renowned scientists at the laboratory, but just seeing all these scientists work together and interact really helped me develop my networking and social skills.

I am currently entering my third year of undergraduate and have two more summers before I would enter into graduate school, or some job. With these two summers there is a good chance I would want to come back to Lawrence Livermore National Laboratory and continue my work. This being my first experience with research, LLNL, and the bioinformatics group there was a huge learning curve I had to overcome before I could begin to be productive. This would make coming back to the lab that much more effective; I am already familiar with the laboratory protocols and processes, the bioinformatics team and resources. The DHS-Stem program should not only allow reapplying for the internship program but should encourage it. I would say that for any intern, myself especially, they would be twice as productive – twice the work for the money – if they returned for a subsequent year. In discussions with other interns I found that many of us hit time constraints in our research and small adjustments could really make all the difference to the experience of the intern.

On what the department of homeland security can do for domestic security I really encourage for DHS to look into financing more metagenomic sequencing of both pathogenic and nonpathogenic samples and to finance a program to assess the feasibility of using metagenomic samples with the pathogen probes that are in use today at Lawrence Livermore National Laboratory and other locations which use similar methods of probe synthesis and detection.