Bio-Informatics for Biodefense

Past, Present, and Future Perspective

Tom Slezak
Lawrence Livermore National Lab
slezak@LLNL.GOV

UCRL-PRES-235435

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

- Founded in 1952 by Nobel Laureate E.O. Lawrence
- Located in Livermore, California
- Managed by the University of California for the Department of Energy 1955-2007
- Now managed by LLNS, LLC
- Annual budget about $1.5B
- 1.2 square mile main site
- 10 square mile remote site
- ~8,000 employees

http://www.llnl.gov
The Chemical & Biological National Security Program (CBNP) leverages a broad range of LLNL capabilities.
We cover a range from basic science to deployed operational systems.
Tom Slezak: Brief Bio Sketch

- Univ. San Francisco Computer Science, 1975
- MS in Computer Science, Univ. California Davis 1977 (LLNL campus)
- Worked in industry 1977-78

- At LLNL since November 1978
  - supporting bio research
  - Unix support 1979-1986
  - Bio-defense program, 1999-present

2002 Team Photo
Our primary mission: design pathogen DNA identification signatures with extremely high sensitivity and specificity

• We do applied research and development for the US Department of Homeland Security (DHS) and other federal agencies
• The Centers for Disease Control (CDC) is the primary customer for our signatures
• We developed the BASIS system for wide-area biosecurity for the 2002 Winter Olympics (Salt Lake City)
• This system was used after the anthrax letter attacks in 2001, and evolved into the US-wide BioWatch aerosol pathogen monitoring system in 2003
• LLNL is involved in all aspects of biosecurity in the US
We pioneered a comparative genomics approach to DNA signature design

• Determine consensus over all target genomes
• Subtract out any exact matches to non-targets (18+bp)
• Use Primer3 (MIT) to mine for good TaqMan® signatures
• Lab test against 2,200+ backgrounds, extensive strain & near-neighbor panels
• Select multiple conserved/unique loci to ensure specificity
• Validate with CDC or other customers
• Multiplex using Luminex bead-based protocol
• Constantly monitor new sequence to adjust signatures as required

10/19/2007
Our Human Genome Program background gave us a unique perspective

- We had years of experience in scaling up infrastructure supporting biological research
  - Whole-chromosome physical mapping and sequencing for human ch19 (60Mb)
- We were humble enough to let the data itself tell us which portions of a pathogen’s genome were important in terms of uniqueness
- We were desperate enough to conceive and risk a paradigm-busting change in how DNA signatures could be designed
- We were focused on achieving specific deliverables
We had significant experience in large-scale comparative genomics

Physical map of 2 Mbp of the P-Telomeric end of human chromosome 19

Comparison of human chromosome 19 with corresponding regions of various mouse chromosomes
LLNL pioneered whole-genome signature discovery and enabled rapid development of low false positive assays

- **Computationally accelerate signature discovery via comparative genomics**
  - Create consensus target sequence
  - Discover conserved potential signatures
  - Reject unacceptable signatures
    - Based on sequence data
    - Instrumentation and assay constraints
  - Direct experimental screening from reagent procurement to analysis
  - Nucleic acid operational; protein targets in development

- **Computationally monitor for erosion of deployed signatures and assays**
  - Automatic weekly erosion checking
  - Detected erosion of deployed assays allowed corrective action

**Subtractive Suppressive Hybridization (SSH) computed in parallel over all organisms**
We have had a fully-automated DNA signature pipeline for over 4 years

- We let the genome itself tell us what is unique
  - Prior approaches focused on specific genes of interest only
- Scalable algorithms for efficient processing
  - Fast anchored (MGA) and BLAST-based multiple genome alignment/consensus to deal with a range of conditions
  - Suffix-array sub-string comparison (Vmatch)
  - Minimal Set Clustering for RNA viral organisms
- Rigorous selection and in silico testing of candidate signatures; automatic evaluation of candidates as new sequence data acquired
- Results depend on quantity and quality of target and near-neighbor genomic sequence
- Extensive bench screening (at LLNL and with collaborators) and formal validation

We have designed signatures for nearly all threat-list organisms.
Extensive bench screening is the only way to achieve reliable signatures

• We screen each signature against 2,400+ backgrounds
  – Soils, aerosols
  – “Zoo panel” of potentially confounding DNAs (human, flea, tick, mosquito, rat, etc.)

• Extensive target DNA panel
  – DNA strain collection at LLNL
  – Additional collections available at collaborator sites

• Near-Neighbor DNA panel
  – As many near neighbors as can be obtained through genomic or phylogenetic/taxonomic comparison
  – Avirulent and vaccine strains when available

This rigorous screening occurs before any end-user formal validation.
Our signatures have processed over 6M field assays with zero false positives since Oct. 2001
The BioWatch program uses our assays in dozens of US cities for daily monitoring.

Photos from New York, 2002-2003
Our methods have determined signatures that appear to have clear links to function.

**Blue** - conserved portions of DNA signature (conserved at family level)

**Red** - portion of signature unique to target organism (species level)

**Green** - most highly conserved sequence motif of protein family adjacent to signature region: critical for protein function

*LLNL protein structure modeling used to determine structure for proteins not solved in PDB.*
With DOE, DHS, and other sponsorship, we have worked closely with the CDC since 2000

- A broad CDC – LLNL MOU is in place
- LLNL provides advanced analyses and DNA signature candidates to be tested by the CDC
- Since 2003, the CDC has requested our help on smallpox, monkeypox, cowpox, avian influenza, brucella, Rift Valley fever, Crimean Congo hemorrhagic fever, Marburg virus, and SARS
- Assays provided by us have been tested and used by the CDC in field operations

We have become a major source for CDC pathogen genomic analysis assistance
We were requested to help CDC with SARS in 2003

- We provided ~100 signature candidates
- 33 were tested and 1 was used in a primate study
- Graphical signature viewer shows location of amplicons and signature components
- Manuscript published in 2006 by US Army Infectious Disease researchers
SARS near-neighbor sequence affects signature results and provides biological insight

- Recent sequencing of four BAT SARS genomes provided near neighbor to help point to Human SARS unique region and clue to host specificity

116 SARS genomes: 94 signatures

116 SARS genomes with 4 BAT SARS genomes as rule-outs:

6 of 9 signatures cluster on glycoprotein surface protein involved in host response
Multiplexing offers the ability to combine many assays and controls into one test tube.

Informatics developed to determine potential primer cross-hybridization potential, greatly reducing testing time.
With end-users, we have developed and transitioned several DNA and protein multiplex assays

<table>
<thead>
<tr>
<th>Assay Panel</th>
<th>Customer/End-users</th>
<th>Operational Use</th>
<th># Assays Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-plex biothreat agent nucleic acid panel</td>
<td>CDC/LRN. Currently in validation</td>
<td>BioWatch testing</td>
<td>&gt;25,000 for testing and evaluation</td>
</tr>
<tr>
<td>9-plex biothreat agent protein panel</td>
<td>CDC/LRN. Validated, commercially available</td>
<td>BioWatch</td>
<td>&gt;40,000 at events and in LRN labs</td>
</tr>
<tr>
<td>21-plex FMD/FAD pathogen nucleic acid panel</td>
<td>USDA, NAHLN, PIADC. Has undergone multi-lab comparison testing (14 NAHLN Labs)</td>
<td>Pending USDA approval process</td>
<td>&gt;10,000 for testing and evaluation</td>
</tr>
<tr>
<td>18-plex respiratory pathogen immunoassay panel</td>
<td>UCD Medical Center, clinical diagnostic labs, LRN. In testing</td>
<td>UCD Medical Center, clinical diagnostic labs</td>
<td>&gt;5,000 for testing and evaluation</td>
</tr>
<tr>
<td>11-plex respiratory pathogen nucleic acid panel</td>
<td>UCD Medical Center, clinical diagnostic labs, LRN. In development</td>
<td>UCD Medical Center, clinical diagnostic labs</td>
<td>&gt;1,000 for testing and evaluation</td>
</tr>
</tbody>
</table>
Sequence- and Structure-based analyses guide the development of high-affinity ligands and antibodies for protein-based detection assays

We have devised quantitative means of identifying conserved/unique residues in proteins.

Structure-based analysis: LGA structure comparison enables identification of regions that are conserved among target proteins but deviate with respect to proteins that are structurally related.

Sequence-based analysis: pScore allows identification of residue infrequency in local sequence context.

Powerful protein *structure* comparison using our LGA software reveals conserved/unique regions.

Structure comparison of reference toxin sequence against each structural homolog:

- N-terminus
- Related toxins
- near-neighbor proteins
- C-terminus

10/19/2007

Close-up of sample comparison ➔
We are carrying out structural analysis of orthopox variants of host-response and potential virulence proteins, working with the CDC

We built a 3D model to understand the structural basis of sequence deviations between two clades of monkeypox that differ in virulence potential and host range (spinoff of the 2003 outbreak in the US)

Single residue divergences between vaccinia - variola - monkeypox highlighted on 3D model of C3L: immune evasion protein that varies between orthopox species and may be key to differences in virulence

Diagnostics, vaccines, and therapeutics are potential outcomes of these sorts of analyses.
We were asked to determine “How many genomes are needed to make good DNA signatures?”

- Many US agencies are involved in pathogen sequencing
- Effective coordination of resources was desired
- Nobody knew how many genomes were required
- We leveraged our automated DNA signature pipeline:
  - Monte Carlo wrapper, run DNA signature pipeline over hundreds or thousands of random combinations of pathogen target and near-neighbor genomes
  - Determine number of signature candidates as we increased the number of target and near-neighbor genomes
  - Run this on all branches of the taxonomic tree where threat pathogens occur. Use non-threat organisms if insufficient data available for threat pathogens
How many genomes needed for signatures?
(HPV type 16: ds DNA)

The number of target and near neighbor sequences are both important: need ~4-5 of each

Clear evidence of diminishing returns after ~4 or 5 target and near-neighbor genomes

~18 signatures with 5 targets & near-neighbor genomes

~8 signatures when all 36 targets and 20+ near-neighbors used
How many genomes needed for signatures?
(SARS Virus: ss RNA +)

Additional target sequences not needed at this time. Need phylogenetically close near neighbors.

Prior to the 4 Bat SARS genomes, this plot demonstrated that there were NO sufficiently-close neighbors to SARS yet sequenced.
How many genomes needed for signatures?
(Human Poliovirus: ss RNA +)

Only the number of target sequences is important: We have shown that 12-18 genomes, chosen for diversity across time/space/virulence, are needed.

Most single-strand RNA viruses are too diverse to create even one pan-species TaqMan® PCR signature without extensive use of degenerate bases. We use Minimal Set Clustering in these cases.
Similar analyses have been performed for many major threat list pathogens

• All major bacterial & viral pathogen families where sufficient full-genome sequence data was available
  – In some cases where threat pathogen data was insufficient, appropriate near-neighbor non-threat organisms were used as surrogates

• Results supplied in 2004 to a national coordinating committee, which used them to guide multi-agency funding for pathogen and near-neighbor sequencing

• In 2006 we re-analyzed using the newly-sequenced data and confirmed our predictions

These analyses would not have been feasible without our fully-automated DNA signature pipeline infrastructure.
We generate conserved sets of signatures to detect all strains of divergent RNA viruses

- Minimal set clustering (MSC) software design and application for signatures for highly divergent viruses, which includes most RNA viruses
  - Rift Valley Fever, Crimean-Congo hemorrhagic fever, Venezuelan equine encephalitis, Bluetongue, Hanta, Marburg, Caliciviruses (Norwalk), Influenza serotypes (including Avian)
- CDC and FDA collaborations
  - We generated FDA’s gold standard signatures for Hepatitis A, and Calicivirus signature screening is ongoing by FDA researchers
  - RVF, CCHV, BT, and VEE signatures performing well in ongoing screening

Avian influenza MSC

CCHF Minimal Set clustering for TaqMan Signatures

H5 signatures divided the targets into lineages: Eurasian and North American. To date, all avian strains of influenza infecting humans have been of Eurasian lineage.
Faster Virulence Analysis Would Better Prepare Us for Emerging and Engineered Threats

- Speed response time to a novel pathogen
- Allow us to make use of the enormous amount of genomic sequence available (see NCBI chart below)
- Position us to take advantage of new technologies that lower the cost and time barriers to genomic sequencing (e.g. 454 see below)
- Allow us to quickly compare genes across many genomes

Reproduced from http://www.454.com

Detection techniques need to be carefully matched to mission needs, cost constraints, and desired level of performance.

Mission Space: First Responders; Research; Forensics; Cost-sensitive BioWatch; Clinical Use

Multiplex Level == information content

High

Low

Sensitivity + Specificity == Consequence/Regret

10/19/2007
We created recognition models for known virulence and antibiotic-resistance genes

Create gene lists and capture gene related data in relational tables

Create probabilistic models of gene sequence (hidden Markov Models)

unstructured  structured

754 gene families

Run all models against all Genomes on LLNL Supercomputers

Analyze alignments for sets of virulence genes and infer possible virulence function

810 models
- 328 publicly available
- 482 custom built

Computationally align models to genomic DNA sequence
All Mechanisms versus all Organisms:
Computational predictions to be verified by chips

Note the potential for a new type of “functional forensics” via this approach
Use of NimbleGen microarrays to detect mechanisms of virulence, antibiotic resistance, and potential bacterial genetic engineering

Probe design using various thermodynamic parameters for virulence, A/R genes and bacterial vectors

Fragment sample DNA, label with Cy-5

Probe synthesis in situ

Hybridization, wash, scan

Each chip is a “custom” chip

Data Analysis

We are leveraging the bioinformatics work to make possible rapid determination of the potential of an organism.
Microarray results: BioWatch aerosol sample spiked with Staph Aureus, showing sensitive ID of virulence and antibiotic-resistance mechanisms.
We have shown Detection of Virulence Signatures within hours, speeding time to Actionable Results

• Detection of probes specific to *E. faecalis* virulence signatures in 4 hr hybridization time

• Same day results are possible for high priority samples
Species and Strain Discrimination has been shown

**KEY**
- BA discriminating genes
- BC discriminating genes
- BT discriminating genes
- BA virulence genes
- BA strain discrimination
- BT strain discrimination

<table>
<thead>
<tr>
<th>Virulence group</th>
<th>Name</th>
<th>BA 0442</th>
<th>BA Sterne</th>
<th>BA 0193</th>
<th>BT 97_0027</th>
<th>BT Hakem</th>
<th>BC 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>toxin</td>
<td>L1 component of the B. cereus HBL tripartite haemolytic toxin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>toxin-anthrax</td>
<td>lef, zinc metalloprotease that cleaves MAPKK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>toxin-anthrax</td>
<td>cya, calcium/calmodulin-dependent adenyllycyclase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>multi-drug efflux pump</td>
<td>CII/pagA provides entry of B.anthracis lef and cya into host cell</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>capsular polysaccharide synthesis</td>
<td>putative major facilitator superfamily export protein</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>iron acquisition-siderophore synthesis</td>
<td>wzt2, bexA, putative ABC transporter for capsular polysaccharide</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LPS biosynthesis</td>
<td>pchE, irp2, HMWP2 pyochelin/yersiniabactin synthetase</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>iron acquisition-siderophore synthesis</td>
<td>wzt, ABC transporter, ATP binding component</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>pchC, lipase similar to pyochelin biosynthetic protein from P. aeruginosa</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>cpmF carbapenem resistance</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>BPSL2708, putative exported metallo-beta-lactamase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>borB bacitracin transport permease protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>mphA like, macrolide 2-phosphotransferase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>nitrogen metabolism</td>
<td>aliphatic amidase (EC 3.5.1.1) (H. pylori amiE)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>nitrogen metabolism</td>
<td>formamidase (EC 3.5.1.49) (H. pylori amiE)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>bsr, blastcidin-S deaminase (EC 3.5.4.23)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>capsule biosynthesis</td>
<td>capD, required for capsule biosynthesis in Bacillus anthracis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>capsule biosynthesis</td>
<td>capB, capsule synthesis B. anthracis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>capsule biosynthesis</td>
<td>capC, capsule synthesis B. anthracis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>boc, bacitracin transport permease protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>ksgA, erm(1.A,B,C) Dimethyladenosine transferase (EC 2.1.2.1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>bla(2.B) Beta-lactamase IMP-1 precursor (EC 3.5.2.6)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TAT secretion</td>
<td>tatD, TAT secretion, sec independent protein translucose, desaminase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>transcriptional control</td>
<td>agrC, quorum-sensing sensor kinase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>mcG, resistance to the peptide antibiotic microcins</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>antibiotic resistance</td>
<td>bls, blastcidin S-acetyltransferase (EC 2.3.1.-)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>surface antigen synthesis</td>
<td>glycosyltransferase 2(F. tularensis FTT0707, FTT0708)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Detecting and Characterizing Complex Threats requires a spectrum of techniques & resolution.
Numerous opportunities exist for Comparative Genomics in the near future

- Evolutionary predictions (influenza, SARS, Ebola, etc.)
  - Flexible evolutionary diagnostics/forensics
- Enhanced knowledge of virulence mechanisms
- Improved understanding of host/pathogen interactions
- Faster/better/cheaper/smaller systems for accurate identification of pathogen threats across a wide spectrum of missions, customers, and constraints

Many practical and important problems remain to be solved. Both public health and bio-defense will be improved by continuing advances in comparative genomics. Both summer and permanent job postings can be found at jobs.llnl.gov
Acknowledgements

• Former Team Members:
  • Linda Ott, Tom Kuczmarski, Ed Miller, Chitra Manohar

• Key Collaborators:
  • Rich Meyer, CDC
  • Inger Damon, CDC
  • Stefan Kurtz, U. Hamburg
  • Stephen Salzberg, U. Maryland
  • Gary Hartman, FDA
  • Joe DeRisi, UCSF
  • Matt Dyer, Gordon Lemmon (former BYU graduates and summer students)

• Current Team:
  • Elizabeth Vitalis, Shea Gardner, Carol Zhou, Adam Zemla, Marisa Lam, Clint Torres, Jason Smith, Mark Wagner, Mimi Yeh, Pauline Gu, Crystal Jaing, Peter Williams, Nisha Mulakken, Kevin McLoughlin
Using HMM to Predict the Degree of Virulence: Influenza virus

- Influenza virus: 8 segments, 10 genes
- Hemagglutinin (HA) is the Viral receptor

Hemagglutinin Subtypes:

- H1
- H2
- H3
- H4
- H5 → Avian Flu
- H6
- H7 → Avian Flu
- H8
- H9 → Avian Flu
- H10
- H11
- H12
- H13
- H14
- H15
Several Influenza Pandemics Have Occurred Through the Centuries

**Fig. 2** History of influenza pandemics 1700–2000. Not to exact scale

C.W. Potter, A History of Influenza, Journal of Applied Microbiology Volume 91 Issue 4 Page 572, October 2001
Scores for Alignment to H5 HMM Suggest a Single Most Virulent Conformation for the HA Gene

H5 HMM Aligned to Human Specific H1, H2 and H3

Mundy, N. et al. Structure of the Uncleaved Human H1 Hemagglutinin from the Extinct 1918 Influenza Virus. 2004 Vol 303 p1866-1870