

# Comparative genomics tools applied to bioterrorism defence

**Tom Slezak**

has BS and MS degrees in computer science and has led the LLNL bioinformatics efforts since 1978.

**Tom Kuczmariski**

has an MS in computer science and built the initial DNA signature pipeline.

**Linda Ott**

has a BS in computer science and built the database infrastructure for signature screening and field applications of validated assays.

**Clinton Torres**

has a BS in computer science and built the second-generation DNA signature pipeline.

**Dan Medeiros**

has a BS in computer science and built the second-generation DNA signature pipeline.

**Jason Smith**

has a BS in biology and built the second-generation DNA signature pipeline.

**Brian Truitt**

has a BS in computer science and built the second-generation DNA signature pipeline.

**Nisha Mulakken**

has a BS in biology and built the second-generation DNA signature pipeline.

**Marisa Lam**

has a BS in neurobiology/computer science and built the second-generation DNA signature pipeline.

**Elizabeth Vitalis**

is the lead biologist on the pathogen bioinformatics team and has been involved with signature development, analysis, and screening.

**Adam Zemla**

is a protein modelling specialist and an organiser of the Critical Assessment of Structure Prediction (CASP) international contest.

**Carol Ecale Zhou**

has built a microbial annotation capability for the LLNL pathogen effort.

**Shea Gardner**

was a Lawrence Fellow at LLNL prior to joining the pathogen bioinformatics team and has contributed portions of the pipeline.

**Keywords:** *pathogens, comparative genomics, DNA detection signatures, protein-based detection signatures, sequence alignment algorithms, protein structure modelling, bioterrorism*

Tom Slezak,  
Lawrence Livermore National  
Laboratory,  
7000 East Ave. L-448,  
Livermore, CA 94550, USA  
E-mail: Slezak@LLNL.GOV

*Tom Slezak, Tom Kuczmariski, Linda Ott, Clinton Torres, Dan Medeiros, Jason Smith, Brian Truitt, Nisha Mulakken, Marisa Lam, Elizabeth Vitalis, Adam Zemla, Carol Ecale Zhou and Shea Gardner*

Date received (in revised form): 6th March 2003

## Abstract

Rapid advances in the genomic sequencing of bacteria and viruses over the past few years have made it possible to consider sequencing the genomes of all pathogens that affect humans and the crops and livestock upon which our lives depend. Recent events make it imperative that full genome sequencing be accomplished as soon as possible for pathogens that could be used as weapons of mass destruction or disruption. This sequence information must be exploited to provide rapid and accurate diagnostics to identify pathogens and distinguish them from harmless near-neighbours and hoaxes. The Chem-Bio Non-Proliferation (CBNP) programme of the US Department of Energy (DOE) began a large-scale effort of pathogen detection in early 2000 when it was announced that the DOE would be providing bio-security at the 2002 Winter Olympic Games in Salt Lake City, Utah. Our team at the Lawrence Livermore National Lab (LLNL) was given the task of developing reliable and validated assays for a number of the most likely bioterrorist agents.

The short timeline led us to devise a novel system that utilised whole-genome comparison methods to rapidly focus on parts of the pathogen genomes that had a high probability of being unique. Assays developed with this approach have been validated by the Centers for Disease Control (CDC). They were used at the 2002 Winter Olympics, have entered the public health system, and have been in continual use for non-publicised aspects of homeland defence since autumn 2001. Assays have been developed for all major threat list agents for which adequate genomic sequence is available, as well as for other pathogens requested by various government agencies.

Collaborations with comparative genomics algorithm developers have enabled our LLNL team to make major advances in pathogen detection, since many of the existing tools simply did not scale well enough to be of practical use for this application. It is hoped that a discussion of a real-life practical application of comparative genomics algorithms may help spur algorithm developers to tackle some of the many remaining problems that need to be addressed. Solutions to these problems will advance a wide range of biological disciplines, only one of which is pathogen detection. For example, exploration in evolution and phylogenetics, annotating gene coding regions, predicting and understanding gene function and regulation, and untangling gene networks all rely on tools for aligning multiple sequences, detecting gene rearrangements and duplications, and visualising genomic data. Two key problems currently needing improved solutions are: (1) aligning incomplete, fragmentary sequence (eg draft genome contigs or arbitrary genome regions) with both complete genomes and other fragmentary sequences; and (2) ordering, aligning and visualising non-colinear gene rearrangements and inversions in addition to the colinear alignments handled by current tools.

## INTRODUCTION

### DNA and protein signatures for pathogen detection

Most people have experienced 'food poisoning', nearly always a bacterial assault, as well as ailments such as colds or flu, caused by viruses. The majority of such cases do not fatally threaten those with sound health and immune systems. However, a smaller group of infectious agents do pose a threat, even to those with healthy immune systems. Many of these are discussed on a Centers for Disease Control (CDC) website<sup>1</sup> and in recent texts.<sup>2</sup>

**Pathogen detection can be accomplished by locating unique regions of the DNA or protein sequence**

Accurate and rapid diagnosis of pathogens is important, whether the context is determining the cause of an outbreak of food poisoning due to unsanitary food preparation/cooking/serving/storage or a deliberate dispersal of a pathogen not normally encountered in urban life (eg anthrax in the postal system.) There are two major approaches to pathogen detection: detecting a unique DNA (or RNA) sequence and detecting unique protein regions.

**Detection signatures need to be both conserved over all target strains and unique compared to all non-target genomes**

Recently, some of us published a paper<sup>3</sup> containing a tutorial primer on DNA detection via the polymerase chain reaction (PCR) technique that underlies most current DNA detection techniques. The assays currently produced by our system include the following:

**Protein detection mechanisms provide speed at the cost of sensitivity**

- A pair of PCR primers (forward and reverse) is used to *amplify* the region in question. This is a vital concept of PCR-based DNA diagnostics, since it allows even a single molecular copy of DNA to be amplified sufficiently for detection to take place. The PCR primers themselves do not necessarily have to be unique to the pathogen. Some researchers will use primers that are unique to a family of organisms, and rely on an internal hybridisation oligonucleotide (oligo) for specificity to a single species or strain of a pathogen. By default, our system will seek unique primers whenever possible to reduce the chances of cross-reaction.

- A hybridisation oligo is used to detect a section of the PCR-amplified fragment. A fluorescent tag attached to the oligo is released during the hybridisation process and can be detected via optical detection mechanisms, quantifying the amount of pathogen DNA that is present.

Hence, the problem of designing PCR-based DNA detection assays for pathogens can be simplified as requiring us to locate regions of the genome that are:

- unique (as far as can be determined given available sequence data);
- conserved (over all the different strains/isolates of the pathogen that can be tested);
- long enough to match the needs of a particular detection technology;
- meet any other special requirements of a particular detection technology (example at the website<sup>4</sup>).

The first step in pathogen detection via unique DNA is the mechanical or chemical disruption of a bacterial pathogen to access the DNA. This implies that preparation steps are needed before the DNA detection assay can be run. The systems used at the 2002 Winter Olympics required several minutes of mechanical disruption and several process steps before being run on a thermal cycler for about 30 minutes. Batch production runs thus required at least an hour to obtain results.

In some situations faster detection is desired that does not require extensive preparation or the amplification process of PCR. Protein signatures thus generally need to recognise surface proteins (eg the spore coat of anthrax). This type of detection is now commonly used for diagnostic kits, such as home pregnancy tests and tests for prostate specific antigen. One disadvantage of protein detection

assays is that, lacking amplification, they are less sensitive than PCR-based DNA assays. Many of the tools discussed below are also suitable for protein signature design, although we will not go into great detail on this ongoing research topic in this paper.

Our efforts have been primarily focused on the problem of detecting pathogens in the environment. Related problems include detection of pathogens within a host (human, plant or animal) and forensic determination of strain/isolate resolution.

**We have focused on detection of pathogens in the environment**

### COUNTER-BIOTERRORISM AND APPLIED GENOMICS

The anthrax attacks in autumn 2001 provided the first real awareness that most of the US general public had of bioterrorism. Numerous problems, including false positives and false negatives from different types of diagnostic assays and poor information being conveyed by officials to the public (eg anthrax being called a 'virus' when it is a bacterium), revealed how poorly prepared the USA was as a nation for this kind of event. Although the Department of Energy (DOE)-funded programme had been in operation for more than a year to produce pathogen diagnostics, a heightened sense of urgency required us to scale up our operations quickly. How this practical application of comparative genomics pushed the state of the art to meet the needs for fast results is described below, and the remaining challenges are also pointed out.

To date, DNA signature candidates have been produced for virtually every threat list agent<sup>5</sup> for which there is adequate genomic sequence available. The authors are involved in efforts to acquire sequence for those lacking such data. Many of our signature candidates that have been taken forward to assays have undergone rigorous CDC validation exercises and have begun to enter the public health system. Other signatures are available to, and are being used by, other agencies involved in aspects of national

**Pathogen DNA signatures are limited by sequence availability over strains and near neighbours**

**A whole-genome approach to DNA signature design yielded greatly improved efficiency over traditional approaches**

defence. For what should be obvious reasons, details are not given of the major threat agents the authors have worked on, the number of assays for each, or the actual diagnostic assay signatures. The same techniques have been applied to a number of human and livestock pathogens that are not considered to be potential bioterrorism agents, and these are used as examples to illustrate our capabilities.

### COMPARISONS WITH OTHER DNA SIGNATURE DEVELOPMENT EFFORTS

Traditional approaches to DNA signature development started with the assumption that a particular gene was vital to an organism's virulence, host range or other factors that might be considered 'unique'. Suitable primers and probe were designed for the detection system of choice without effective computational screening for uniqueness. The resulting assay would then be tested with the available strain(s) and success declared if they were detected while near-neighbours were not. This approach would sometimes yield good results, but failures frequently occurred owing to inadequate strain panel, near-neighbour and environmental testing. Failure could also occur because the chosen gene was not unique as assumed but was part of some mechanism that was common to other as-yet unsequenced organisms. Although commercial software and services exist to create DNA diagnostics to detect a specified DNA target, there is an assumption that the user has already determined the best target to use. The authors are not aware of any other software or service that purports to determine apparently unique DNA signature targets from whole genomes.

Our approach improves upon the traditional method by efficiently locating portions of the pathogen target genome that are definitely not unique, and eliminating them from further consideration. In some cases, appropriate near-neighbours have been sequenced to

**Extensive collaborations are necessary to test signatures on some live pathogens**

**Our whole-genome approach is a by-product of experiences gained on the Human Genome Programme**

**Pathogen sequence availability can vary widely**

**Pathogen sequence data is acquired from multiple sources and at several levels of completion**

provide maximum isolation of target regions related to virulence. Rigorous selection criteria are used on the remaining potentially unique portions of the genome to select candidate assays that should work against all strains of the target for which sequence is available. Finally, the authors have worked with appropriate collaborators to ensure that the assays are screened against large strain panels and robust environmental samples. Although our approach dramatically reduces failures both in initial laboratory screening and subsequent field-testing, the quality and quantity of target strain and near-neighbour sequence available are limiting. Thus, pathogen diagnostic design is viewed by the authors as a continuous process that is potentially affected by every new microbial genome sequenced.

In the past three years our team has interacted with every US government agency that is involved in bioterrorism and has had numerous personal interactions with DNA signature developers from the military, US Department of Agriculture (USDA), CDC, Food and Drug Administration (FDA), Justice Department and others. The team has also interacted with counterparts in the UK and Canada and with academic pathogen experts at multiple institutions. The team is not aware of any other DNA signature developers who have automated their signature development process beyond multiple sequence alignment, primer design software and simple similarity searching. For example, researchers at the USDA Plum Island facility have sequenced over 100 isolates of Foot and Mouth Disease Virus (FMDV) and are designing DNA diagnostics using that class of tools (D. Rock, personal communication). While these techniques may work well enough, especially for small viral pathogens, they do not scale well when the goal is to deliver and maintain reliable diagnostics for dozens of bacterial and viral pathogens.

## COMPUTATIONAL IMPROVEMENTS IN DNA SIGNATURE DEVELOPMENT

Members of our team spent many years working on the Human Genome Project prior to becoming involved with pathogen signature design. It was theorised that a whole-genome analysis approach, ie comparing a target pathogen genome against all other microbial genomes that had been sequenced, would allow the DNA to tell us what was unique (or, alternatively, what was definitely *not* unique and thus not worth our effort, unless it were part of a shared virulence mechanism). Beginning in August 2000, the team built a set of tools that would accomplish this goal. Several situations of different target pathogen sequence availability were quickly confronted:

1. We had a single, completed genome for the target pathogen.
2. We had multiple completed genomes for the target pathogen.
3. We had a single, draft genome (multiple contigs) for the target pathogen.
4. We had multiple draft genomes for the target pathogen.
5. We had a collection of less-than-genome sequence fragments for the target pathogen.
6. Combination of (1) and (5).
7. Combination of (2) and (5).
8. Combination of (3) and (5).
9. Combination of (4) and (5).

For our comparison database, all complete bacterial and viral genomes plus available draft bacterial genomes were downloaded from several sources. This included the NCBI<sup>6</sup> and major genome

**Identifying a non-unique sequence is the fastest way to identify the apparently-unique regions**

centres including TIGR,<sup>7</sup> DOE's Joint Genome Institute<sup>8</sup> and the Sanger Centre.<sup>9</sup> Our goal was to identify all portions of the target pathogen genome (larger than some minimum size threshold, 18 bp currently) that were *apparently unique* when compared against all other non-target genomic sequences. Note that this 'electronic subtraction' problem could be accomplished by solving the inverse problem: identifying all portions of the target pathogen genome that are shared with organisms in the comparison database.

**Few multiple-sequence alignment programs could meet our needs**

For target pathogen sequence availability cases 1 and 3 above, where only a single completed or draft sequence was available, the electronic subtraction problem stated above had to be solved. However, for all other cases it was first necessary to determine a consensus sequence of the available target pathogen sequence, and then perform the electronic subtraction using that consensus. In August 2000, several major problems were faced in trying to form consensus sequences for the availability cases listed above:

- There was no multiple genome alignment program available that scaled to handle full-sized bacterial genomes (3–5+ Mbp in length).
- There was no effective way to align less-than-genome fragmentary sequence with one or more completed full-sized bacterial genomes.
- There were no effective tools for aligning multiple draft genomes, either among themselves, and/or with completed genomes or fragmentary sequence. (Some sequence assembly tools might have had limited incremental assembly capabilities, but generally suffered from other limiting assumptions.)

**MGA and DIALIGN handle some but not all pathogen alignment needs**

The first problem was solved for most cases in 2002 with the release of MGA<sup>10</sup> (Multi-Genome Aligner), but the second

and third problems remain unsolved. This means that for bacterial pathogens, crude methods are still employed to handle cases 2 and 4–8 above, or else at least some of the available sequence data cannot be used, owing to lack of adequate alignment methods. This problem does not plague most viral genomes, however, which are small enough that an existing multisequence alignment tool (DIALIGN<sup>11</sup>) can handle cases 2, 5 and 7. Nevertheless, analysis of large viral genomes may require unacceptably long running times or may exhaust memory, and thus, like bacteria, may benefit from improved alignment tools.

The need to find efficient solutions for the electronic subtraction and consensus sequence problems led the team to examine the available choices for whole-genome comparison software.

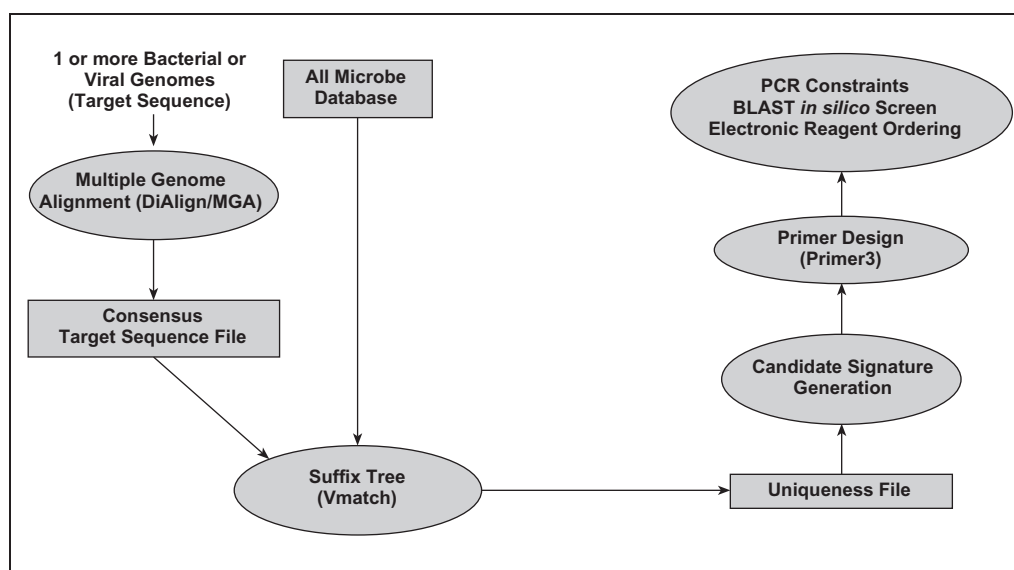
### **An automated system for predicting pathogen DNA signatures using comparative genomics tools**

What follows is a brief description of the system constructed to automate the task of predicting pathogen DNA signatures using some of the tools described above. Figure 1 gives a simplified diagram of the KPATH system.

#### ***Determination of target pathogen consensus sequence***

When the team began computational DNA signature development task in August 2000 there were no algorithms available that could efficiently align multiple bacterial genomes. After experimenting with programs that did not scale to handle even tiny viruses, DIALIGN was located, although it could take days to weeks to run on large viral genomes. In 2002, a new multiple-genome aligner program, MGA, became available from Michael Höhl, Stefan Kurtz and Enno Ohlebusch. This fast, anchor-based algorithm works well for a collection of whole genomes similar enough to have exact-match 'anchor' regions evenly distributed and present in

**Few tools are available to work with draft genomes**



**Figure 1:** The KPATH system flow. Inputs consist of one or more genomes. Small viral genomes are aligned with DIALIGN, larger genomes with MGA. A *consensus gestalt* (see Figure 2) is formed from these alignments and serves as input to Vmatch, which compares it against our *all\_microbes* database. This yields a *uniqueness gestalt* (see Figure 3) that is mined for potential signature candidates. Those candidates are processed by the Primer3 program and filtered by custom software to yield a final set of signature candidates. A final computational screening is done to verify that cross-reactions are not detected, before a subset of the candidates is ordered electronically for subsequent wet laboratory screening

**MGA assumes that there exist identical, colinear anchors across all input genomes**

each genome. As an example, aligning six variola (smallpox) genomes (~190 kbp in length) did not complete in a week on DIALIGN and required breaking the genomes each into three pieces. MGA can align these same six whole genomes in less than 30 minutes (much of this consumed by aligning gaps between the anchors). It was noted that all existing alignment tools assume colinearity and do not handle rearrangements or duplications

well. MGA currently cannot align genomes too distant to have sufficient exact anchors evenly distributed, nor can it align complete genomes along with incomplete ones (ie draft genomes or sets of sequence fragments). The team currently uses DIALIGN and/or MGA to align multiple whole-genome instances of any target genome that has more than one genome available. From this we determine a 'consensus gestalt' (Figure 2)

**A consensus gestalt indicates regions conserved across all inputs**

```

agtaatcgt.ATCATGTGACCCACTTGAGAAGTTAGTAAC.TTTTTTCTATTATAATC'TTGATACC
GTAAGATACAT'TACTACACATAGGAAT'CCCTGAT.GAGCAATGTTTAAATACATCTACATTTGGA
T.TGATGTAGTTGCGTATTTCTCTACAATATTAATACCATTTT'GCAACTATTTATTTCTAGACC
TTTTG.GATTAGTAATCTCAATAA'TTCTACGTCAATATTATCAGAT'TCTATATATTCGAATATATC
AAAGTCAT'TGATATTTTATAAT'TGGTAGAAGACAATAATGACACCACAACATCAGTTTTGATATT
CTTATTTTT.TTGGTAACGTATACATTTAATGAATTTTCATTACGTTCTACCAATGATTGTGCACT
GCAGGCATCAAAGTTTTACAAC'TATCATAAAGCATACTATCCTATCC
  
```

**Figure 2:** A *consensus gestalt*. MGA has been used to align several genomes of a pathogen. Capital letters indicate runs of conserved bases (eg identical in all inputs at that position) of 18 bases or longer. Lower-case letters indicate shorter runs of conserved bases. Dots indicate positions where all inputs do not agree. The consensus gestalt is necessary so that any signatures developed will detect all of the input genomes

that indicates regions of the genome that are conserved across the multiple strains/isolates. If our input consists of just a single draft or finished genome, the genome itself is treated as the consensus.

***Fast, scalable sequence comparison programs to locate unique sequence***

It is important to note that alignment is used only to achieve a consensus of the pathogen target genomes. Comparison to all non-target genomes is handled using a substring comparison program that quickly determines which portions of a bacterial or viral genome consensus are potentially unique, compared against other sequenced microbial genomes.

Suffix trees are the most efficient data structure for comparing two strings to determine matches. MUMmer<sup>12</sup> was used first, but the version available in late 2000 did not scale past 1 Mbp input (recent versions have removed the size limitations and added additional capabilities<sup>13</sup>). Fortunately, Stefan Kurtz's Vmatch<sup>14</sup> programs had just been developed and provided the most scaleable implementation of suffix trees to date. He graciously provided us with features unique to our application that allowed us to compare a target viral or bacterial genome against a 900+ Mb library of all publicly available microbial genomes in just a few minutes. This compares to the 2–4 days required for naive approaches (eg BLAST,<sup>15</sup> which could involve parsing enormous output files). The input is either one target genome, or the 'consensus gestalt' as described above.

Note that this approach clearly identifies what portions of the target genome are definitely *not* unique. Further effort is needed (described below) to determine the suitability of the remaining portions of the target genome, and to deal with the fact that the suffix-tree approach cannot detect all regions that potentially may cross-react when hybridised, even when these regions are not exact matches. It is also the case that we are often interested in non-unique signatures (eg for shared virulence mechanisms). However, it is the

power of being able to 'mask out' all non-unique portions of a target genome by comparing it with all other microbial genomes that has enabled us to achieve a high rate of signature success compared to traditional methods. As an example, when LLNL completed the sequence of the *Yersinia pseudotuberculosis* genome, it was found that 97 per cent of the *Y. pestis* genome was now 'masked out,' letting the focus be on the 3 per cent that was unique. The comparison library prior to the availability of this near-neighbour genome could mask less than 33 per cent of the *Y. pestis* genome out, thus illustrating the tremendous value of having a non-virulent close neighbour genome. Our method lets the genome itself define what is unique (and conserved, if appropriate), and presumably important.

In 2002 major improvements were made to our system that take advantage of multiple CPUs and allow for incremental additions to our sequence comparison database. Instead of comparing a target pathogen against a single suffix tree composed of the entire comparison database (minus the target), pairwise comparisons are now made of the target against each other genome. The output indicates which part(s) of the target over a defined length threshold exactly match the comparison genome and results are stored in a database. As new genomes are automatically acquired from various public databases, incremental pairwise Vmatch comparisons can be performed and the results added to our database. It should be noted that this is not a database of genome alignments, but rather a database of substrings that are in common between a pathogen target and all other non-pathogen genomes. At any time, we can take all the pairwise comparisons for any single target and rapidly compute the current 'uniqueness gestalt' file (Figure 3). This file is the input consensus gestalt genome, with '.' replacing any base positions that have been determined by Vmatch not to be unique. Runs of potentially unique nucleotides of at least

**Substring comparisons are used to locate non-unique sequence**

**Suffix trees enable fast comparisons**

**Vmatch can compare a pathogen genome against all sequenced microbial genomes in minutes**

**Near neighbour sequence helps isolate virulence-associated regions**



**Signature candidate annotation allows prioritisation for bench screening**

signature candidates. However, rather than selecting them randomly (when there are more candidates than are economically feasible to screen in the wet laboratory, usually about 500–1,000 per pathogen), a system has been built to help prioritise the candidate selection based on annotation. One genome is annotated from the input set of genomes used to generate signature candidates, or annotations are downloaded when available from public database, and our signatures mapped onto these annotations.

**Signatures from intergenic regions can help foil deliberate signature evasion for attacks or hoaxes**

Annotation allows us to scrutinise signatures in a biological context. Identifying genes responsible for rendering a pathogen virulent is one component of a good diagnostic signature set. Without virulence genes, an organism cannot cause potentially lethal illness unless virulence factors from other species have been inserted into the genome by genetic engineering. From the standpoint of public safety, we are more concerned about determining what an organism can *do* than about what an organism *is* or where it might have originated.

**Rigorous electronic screening and downselection reduces the time and cost of subsequent bench screening**

Our first goal in annotating a microbial genome is to identify the functional regions, including genes and regulatory sequences. Gene finding programs (eg *Critica*,<sup>18</sup> *Genmark*<sup>19</sup>) are applied and BLAST analyses run to define coding regions and to make tentative functional assignments to genes. DNA signature candidates are then aligned with known and/or hypothetical genes on a reference genome, and those that intersect genes associated with pathogenicity (eg virulence factors, host range determinants, antibiotic resistance) are given priority. Candidates associated with genes of interest are manually selected, and a random selection of candidates within intergenic regions is included, for wet laboratory screening. The random unique intergenic regions are selected as a guard against gene deletion or substitution engineering to evade DNA-based detection. There are few tools focused on viral gene finding, and none known to us

**Other assay formats can also be generated**

that can adequately predict genes in certain RNA virus families.

### **Computational and laboratory screening weeds out failures**

Further electronic screening is necessary for the surviving signature candidates. This includes the use of custom-tuned BLAST to ensure that most potential cross-reactions with degenerate sequence will be detected. In extreme cases (eg small single-stranded RNA viruses where few candidates are available and extensive mutation might evade BLAST screening), an electronic PCR program written by the authors will exhaustively search for potential hits.

At this point, our signatures are used for environmental sampling and not for detection of pathogens within infected humans. To guard against contamination during sample preparation and testing, our signature candidates are also screened against the human genome in addition to having human DNA in the wet laboratory screening. Not surprisingly, most of our primers have reasonable matches in the larger genomes now available (human, mouse, *Drosophila*), but the odds of having adequate matches of both primers and probe on the same chromosome/strand in the correct order and proximity are quite small.

The success of our method is constrained by the availability of genome sequence data. Extensive laboratory screening of the successful candidates against DNA panels including target strains, genetic near-neighbours, and environmental backgrounds is therefore performed before any candidate is considered ready to undergo official CDC validation or other field use. It should also be noted that we are currently only exploiting the development of TaqMan<sup>®</sup> signatures in our system. It would be quite easy to add additional back-ends to generate other types of DNA signatures (example: Molecular Beacons,<sup>20</sup> oligos for chip-based assays), should our end-users desire this capability.

**Signature erosion can be automatically detected**

***Automated updating of sequence information and automated signature maintenance***

Because of the dependence on the latest genomic sequence data, the process of checking several major data resources (Genbank, TIGR and the DOE's Joint Genome Institute) has been fully automated for new and updated genome sequences and installed in our local signature creation database. Our automated system now has the capability to check each existing signature candidate against the newest genome data, informing us electronically of any potential signature 'erosion' that may require action. Because of the way our system works, exactly which genome potentially could cross-react with a particular signature for a particular pathogen genome is known. This capability to automatically maintain our signature sets, as floods of new microbial genome sequencing hit the public databases, gives us a unique edge for DNA signature maintenance. As a new sequence comes in, potential problems with our signatures can be automatically detected and they can be re-designed or replaced with others that are held in reserve. Of course, in some lucky cases this may help us stumble on shared mechanisms of virulence, antibiotic resistance, host range selection or other evidence of pathogen evolution.

**Automatic maintenance of signature sets is performed as new sequence is required**

**A system for protein structural modelling is used to help understand important regions**

***Custom protein structural modelling of selected signature regions***

Our pipeline has generated numerous signatures for multiple bacterial and viral pathogens that have proven to work well in extensive field use. Many of these signatures were 'anonymous' when they were created, meaning that little, if anything, was known about the region upon which they landed. To find out more, a homology-based computational protein structure modelling system is being developed. The general 3D model building method is already implemented as an automated protein structure prediction server AS2TS<sup>21</sup> (Amino acid

**Homology to solved protein structures is exploited to create models**

Sequence to Tertiary Structure). Our high-throughput method for computational protein structure modelling is designed to identify the most suitable templates (structure-determined proteins from the PDB database<sup>22</sup>) for a given query sequence, to build sequence-template alignments, and to produce the protein structure via homology modelling procedures. 3D protein models are constructed automatically with some or all of the following steps:

- The set of preliminary sequence-structure alignments to known protein structures is generated using pairwise sequence alignment (Smith-Waterman,<sup>23</sup> FastA,<sup>24</sup> BLAST), and the multiple sequence alignment PSI-BLAST.<sup>25</sup>
- A selected set of alignment-based backbone models is created.
- The correctness of the alignment is verified by: (a) comparison and analysis of multiple sequence alignments and structure alignments between templates, (b) analysis of secondary structure prediction of modelled protein and comparison with secondary structure assignment from considered templates, and (c) structure comparison of all generated preliminary models.
- LGA<sup>26</sup> (Local-Global Alignment) software builds regions that cannot be directly 'copied' from the template structure: termini, insertions, deletions and loops.
- Side chains are added with the SCWRL<sup>27</sup> program. When there is more than 70 per cent sequence identity to the closest known structure, the coordinates for corresponding atoms are inserted directly to the model.
- The final model is constructed after an analysis and evaluation of all generated models when the completeness of the

prediction and the homology level of the templates used for model building are verified.

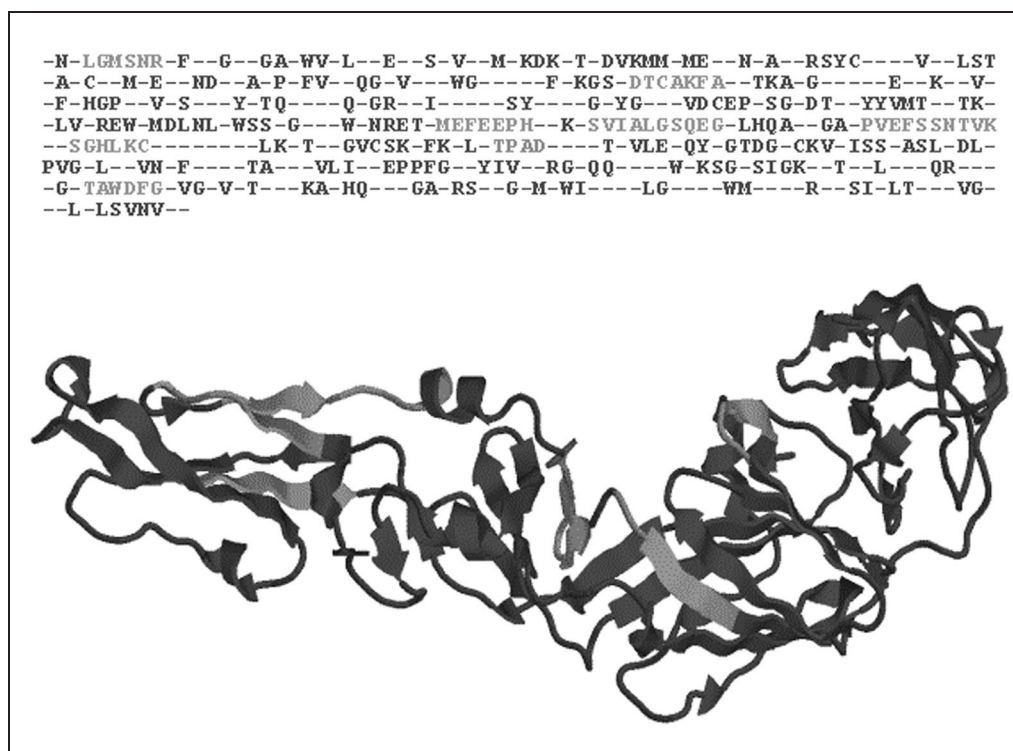
The strength of our AS2TS approach lies in combining sequence and structure searches of existing databases to identify structural homologues of proteins for which sequence similarity alone is insufficient to determine. This includes the study of sequence–structure patterns, secondary structure information and intermediate sequence search techniques. It significantly improves the process of the fold recognition and more distant homologue detection. It also enhances the quality and accuracy of the final 3D protein models produced, especially when

small selected fragments of proteins must be modelled.

Using our AS2TS protein structure modelling system it is possible to model some proteins of interest and indicate where conserved and unique target sequences are located. For example, our system was used to build a 3D model of the envelope glycoprotein of West Nile Virus (WNV) that lies on the surface of the virion envelope and is therefore accessible to antibody recognition and binding. Our DNA signature pipeline identified the conserved and unique sequence (candidate signature regions) in the gene of this protein (Figure 4). As discussed below, this capability is now being expanded, which is merely an

**Locating conserved and unique protein regions provides insight for different virulence characteristics**

**An example of mapping conserved and unique DNA sequence onto a protein structure model for West Nile Virus envelope glycoprotein**



**Figure 4:** The West Nile Virus envelope glycoprotein has been modelled using the AS2TS system under development at LLNL. The top section shows the protein sequence, where residues representing non-conserved DNA sequence have been replaced with a ‘-’. Conserved DNA fragments of length 18 or greater that appear to be unique when compared against our DNA database of microbial genomes have their corresponding residues shown in a lighter shade. These regions are highlighted on the AS2TS-derived model in the bottom section and have become targets for the generation of monoclonal antibodies; testing is underway. This illustrates our ability to map conserved and unique DNA sequence onto the resulting protein model.

**The pathogen signature pipeline has been scaled to process a bacterial genome in under two hours**

**The same process can be used to locate conserved and unique protein sequence signature targets**

**Protein signature targets can be used to design monoclonal antibodies for protein-based detection**

annotation for DNA signatures, as a vital tool for the creation of protein signatures.

***An expanded database tracking system to handle large volumes of sequence information, signature candidates and screening results***

Our initial implementation was a semi-automated pipeline that ran on a single Unix CPU. Early tests required up to two days to process a single bacterial pathogen. As mentioned, the team recently undertook a major project to fully automate the system and exploit simple parallelism on a new, 24 CPU Unix server. This system can now process a viral genome in a few minutes and a 5 Mbp bacterial genome in under two hours. A large database infrastructure has been built to support the generation, electronic screening and wet laboratory screening of DNA signature candidates. It also supports electronic ordering of oligos and probes from external vendors.

**Automation of protein signature development**

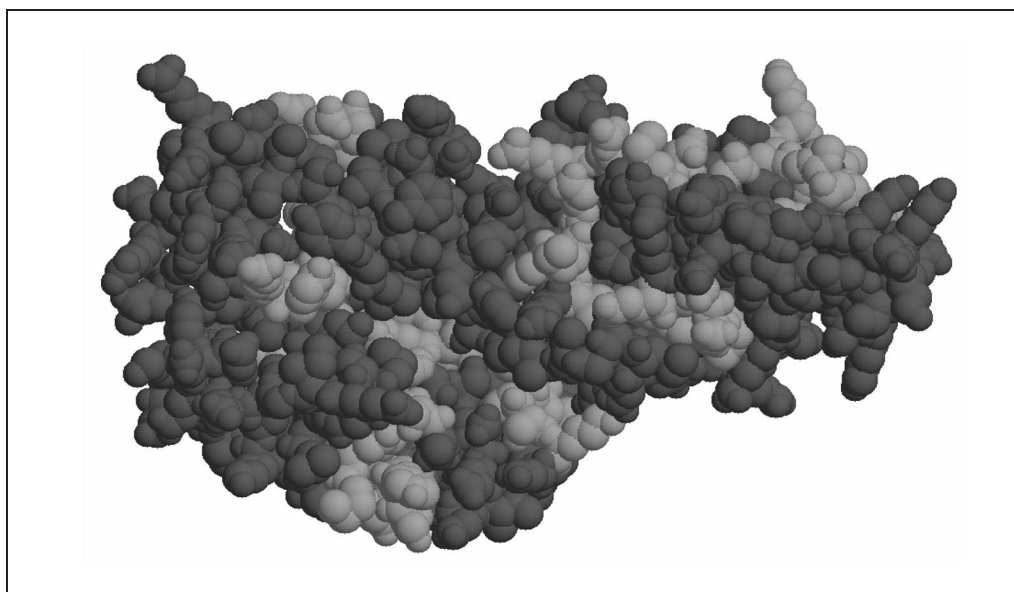
It should not be surprising that these same techniques easily extend to the realm of protein signatures. Protein signatures are desirable for situations where rapid detection of bio-threat agents is desired (eg no mechanically disruptive sample preparation steps to extract DNA) and there is an assumption of sufficient pathogen target available so that PCR amplification is not required. Protein-based signatures are key to identifying organisms that may have been genetically engineered to express proteins that enhance pathogenicity or that have been converted from an otherwise benign organism into a pathogen. In most cases, a positive detection from a protein signature would need to be confirmed with a DNA signature. The lower cost of protein detection makes this type of two-phase threat detection economically feasible for some pathogen detection scenarios, such as continual monitoring of air samples in a transit hub. Protein signatures can be characterised as target

binding sites for *monoclonal antibodies* (protein complexes, produced by an immune response, that specifically recognise and bind a portion of the protein) or *high-affinity ligands* (small molecules that bind to 'pockets' on the protein surface.) In either case, our analysis system can locate potential targets in the following way for a given pathogen protein:

- If more than one protein sequence is available, use DIALIGN and post-process to obtain the conserved regions of the protein.
- Use Vmatch to compare these conserved regions against the Genbank *nr* (non-redundant protein sequence) database to determine the regions of protein sequence of at least six amino acids that are both conserved and apparently unique.
- Determine which of the apparently conserved/unique protein sequence fragments are solvent-accessible (ie on the protein surface in normal conformation.) In the absence of a 3D model PredictProtein<sup>28</sup> can be used, whereas NACCESS<sup>29</sup> is used if we do have a structure model.
- If the protein structure has been deposited into the PDB protein database, or if it can be modelled as described earlier, the potential protein signature regions can be highlighted on the model, as shown in Figure 5.

Annotation is used to identify which proteins are most appropriate, based on biological characteristics, for antibody signature development. For example, annotation can be used to identify surface-exposed proteins for assays that detect whole microbial particles (eg aerosol collection with minimal sample preparation). In addition, computational predictions are applied of post-translational modifications based on predicted secondary structure, chemical

Conserved and unique protein fragments may also be targets for high-affinity ligands used for detection, vaccine, or therapeutic purposes



**Figure 5:** The system also has the ability to compute conserved and unique protein sequence that is on the surface (and thus presumably accessible to detection by antibodies or ligands) and map it onto a model of the protein determined by our AS2TS program. Comparison for uniqueness is done against the Genbank non-redundant protein database. Illustrated are several conserved, unique, and surface-accessible regions of a protein present in a major human pathogen. A 3D model of a pathogen protein, highlighting conserved and unique protein sequence peptides that are accessible on the protein surface. These indicate potential locations for protein sequences

properties, function and sequence motifs, so that unmodified surface-exposed regions of target proteins can be identified and therefore wise choices made from among unique candidate peptide regions generated by the protein signature pipeline.

Currently, manual inspection determines which potential protein targets should be exploited for antibody generation or ligand design and subsequent laboratory testing. Additional opportunities exist for further automation, including incorporation of external knowledge about the expression of the protein, location of the active site and rapid determination of protein surface topology to pre-screen for high-affinity ligand target sites that contain conserved and unique regions. Our long-term goal for protein-based pathogen detection is to be able to process a newly sequenced pathogen proteome in less than one day, automatically determining all high-

probability unique antibody target signatures as well as all potentially good high-affinity ligand-binding sites.

### Examples of DNA signature generation

These techniques described above have been used to develop nucleic acid detection signatures for more than three dozen bacterial and viral pathogens in the last two years. Many of these core assays have undergone rigid CDC validation assays and are now entering the public health network through the CDC's Laboratory Response Network<sup>30</sup> (LRN).

Our DNA signature design system is primarily sequence limited. As much strain variation sequence as possible is needed to ensure the design of robust assays that detect all known strains/serotypes. It is also necessary have as much relevant near-neighbour sequence as possible to whittle down all but the truly relevant sequence differences.

DNA signature targets produced by this system are now in regular use for public health in the USA

A long-term goal is to process an entire proteome in one day

**Insufficient strain and near-neighbour sequence for *Listeria* led to expensive bench screening**

**An assay using our signature target detected all 7 serotypes of Foot and Mouth Disease Virus**

**The Foot and Mouth Disease Virus genome has only one short conserved and unique region**

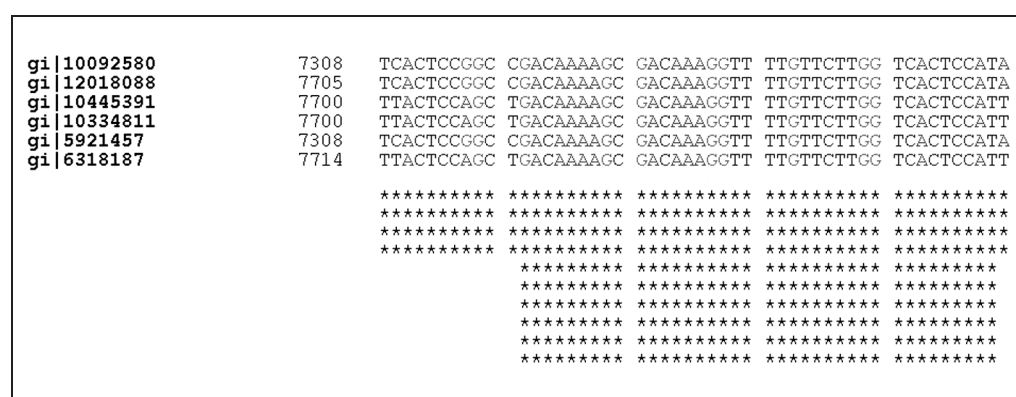
**Foot and mouth disease**

Figure 6 illustrates the use of the DIALIGN alignment algorithm to generate candidate signatures for FMDV. In this figure, FMDV genomes were examined to find regions of identical structure (with a specified minimum length). A scoring system, represented by the stars, evaluates the region, and those with ten stars are considered excellent candidates for signatures. For purposes of illustration, only 6 FMDV genomes are shown, but our analysis actually used 16 that are available from public sources and represent all recognised serotypes of FMDV. Our analysis showed that a TaqMan<sup>®</sup> PCR signature was possible to generate in exactly one small region of the genome (corresponding to a DNA polymerase III gene), owing to extensive variability throughout the remainder. Our signature was tested by the USDA at Plum Island,<sup>31</sup> NY, and was determined to detect all 7 serotypes of FMDV (unpublished). The upcoming release of nearly 100 genomes of additional strains of FMDV from the USDA is anticipated. Similarly, the TIGR is draft sequencing numerous anthrax genomes as part of the analysis of the 2001 attacks. This is a sign of what to expect in the future, as people exploit the low cost of sequencing to achieve detailed knowledge of the

diversity of important diseases. Alignment tools need to keep up with such massive data scaling.

**Listeria**

The extensive screening that was carried out to develop assays for the virulent food-borne pathogen *Listeria monocytogenes* highlights the necessity of sequence data for multiple strains of target and genetic near-neighbour species. Of the 869 computationally developed signatures that were bench screened, only 3 signatures emerged that were specific to *L. monocytogenes*, and also detected our large panel of target DNA samples. Fully 683 of the original signatures were derived from the sequence of a single strain, whereas 186 were extracted from a consensus file generated from an MSA of at least two target strains. The three survivors were all from the latter set, underscoring the value of alignments to identify reliable signatures. The lack of full-length sequence for near-neighbour species hindered our ability to electronically screen and thus ‘protect’ the signatures from cross-reaction on the bench, and ultimately in the field if sufficient biological material is not available in the laboratory for bench screening. Therefore, abundant sequence is necessary to ensure that signatures are



**Figure 6:** The DIALIGN program was used to align six genomes of Foot and Mouth Disease Virus (FMDV). A short region containing the forward primer from our diagnostic assay is shown. The number of stars at the bottom of the diagram indicates agreement of the alignment. Our assay components (forward and reverse primers and hybridisation probe) must land on highly conserved regions in order to reliably detect all inputs. Sixteen genomes were used as input for the actual development of the FMDV assay

both conserved across strains of the target species and unique to only that species.

#### **Human Immunodeficiency Virus (HIV)**

HIV and other rapidly mutating viruses present unique problems for TaqMan<sup>®</sup> identification. Based on a sequence alignment of 30 geographically widespread strains (a limiting input size to DIALIGN to achieve acceptable running time), it was computationally determined that at least nine different TaqMan<sup>®</sup> assays would be required to detect all the strains, owing to substantial sequence divergence.<sup>32</sup> Moreover, a phylogenetic relationship did not correlate closely with whether or not strains shared TaqMan<sup>®</sup> signatures. Approaches using degenerate bases also proved undesirable for this virus, because too many positions would need to contain degenerate bases for a single assay to detect all strains, thus decreasing sensitivity and selectivity of the detection assay. Therefore, rapidly mutating viruses will require using either alternate DNA detection methods (microarray, bead, highly multiplexed assays) or protein-based assays, as proteins are more conserved among strains than are the nucleotide sequences.

Genbank currently contains several hundred complete HIV genomes. None of the alignment tools that the authors are aware of can cope with input on this scale. Note that MGA was unable to align the 30 HIV genomes mentioned above because excessive HIV mutation rates did not allow MGA's anchor-based method to identify conserved 'anchors' present in all genomes. In contrast, anchorless DIALIGN could not align six 190 kb variola genomes in a week, but MGA could do it in less than 20 minutes. This illustrates the apparent fact that no single genome comparison tool will be suitable for all situations, and significant further scaling in comparative genomics tools is required.

#### **FUTURE DIRECTIONS**

Our efforts at LLNL are currently focused on completing pathogen DNA diagnostics as new sequence data become available

because of recently increased funding by several agencies. The team continues to be more aggressive in its processing of viral genomes (dealing with degeneracy in primers/probes), seeking out better methods of inexact matching, and adapting the system to deal with highly multiplexed assays using a variety of new technologies and platforms.

The focus is now on applying similar techniques to automate the selection of protein signature candidates (targets for monoclonal antibodies and high-affinity ligands) and leveraging our abilities in protein structure homology modelling. Our goal in this area is to be able to process an entire pathogen proteome within 24 hours, identifying high-yield unique protein signature candidates.

#### **CONCLUSIONS**

The methods developed in the interests of national security have direct benefits for improved public safety and human health as well as for agricultural safety and efficiency. Early detection and more rapid treatment of human, animal and plant diseases are often the key to prevention and cure, as illustrated by recent tragedies due to WNV transmission via blood transfusions.<sup>33</sup> Rapid, sensitive and accurate pathogen detection carries economic benefits as well. The 2002 outbreak of *L. monocytogenes* that triggered a recall of 27,000 tons of turkey<sup>34</sup> may likely have been minimised or prevented with better diagnostics. Continued computational exploitation of pathogen genomes and improvement on the approach the authors have pioneered will lead to better detection, understanding and management of the many pathogens sharing our environment.

#### **Acknowledgments**

The LLNL pathogen detection system was designed and implemented by the authors, with support from many others. Mimi Yeh gave database administration support, and Mark Wagner and Lisa Corsetti built and provided the essential computing infrastructure. Paula McCready's tireless leadership at LLNL enabled us to work with numerous agencies during the post-9/11 period,

**HIV is too variable for a single TaqMan<sup>®</sup> PCR signature to be able to detect all strains**

**Other techniques must be used for reliable detection of HIV**

**Existing alignment programs have various scaling issues**

**Better tools and more sequence will improve our detection of pathogens**

and her team got the assays screened and out into the real world in time to make a difference. J. Patrick Fitch at LLNL and Elizabeth George at DOE/NNSA headquarters enabled the extra funding that made it possible for us to rapidly expand the field of pathogen diagnostic development. A special thanks goes to TIGR and Steven Salzberg for making MUMmer and other software freely available and for pointing us to Stefan Kurtz. Many thanks to Stefan Kurtz, Enno Ohlebusch and Michael Höhl for giving us early access to, and valuable assistance with, Vmatch and MGA; and to Klaus May and colleagues at Genomatix for providing us with assistance for DIALIGN. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

### References

1. CDC National Center for Infectious Diseases (URL: <http://www.cdc.gov/ncidod/index.htm>).
2. Madigan, M. T., Martinko, J. M. and Parker, J. (2000), 'Brock Biology of Microorganisms', 9th Edition, Prentice Hall, Upper Saddle River, NJ.
3. Fitch, J. P., Gardner, S. N., Kuczmariski, T. A. et al. (2002), 'Rapid development of nucleic acid diagnostics', *Proc. IEEE*, Vol. 90(11), pp. 1708–1721.
4. URL: <http://www.appliedbiosystems.com/support/tutorials/pcropt/>
5. CDC Threat List (URL: <http://www.bt.cdc.gov/Agent/agentlist.asp>).
6. National Center for Biotechnology Information (URL: <http://www.ncbi.nlm.nih.gov>).
7. The Institute for Genomic Research (URL: <http://www.tigr.org>).
8. Joint Genome Institute (URL: <http://www.jgi.doe.gov>).
9. Sanger Centre (URL: <http://www.sanger.ac.uk/>)
10. Höhl, M., Kurtz, S. and Ohlebusch, E. (2002), 'Efficient multiple genome alignment', *Bioinformatics*, 18(Suppl. 1), pp. S312–S320.
11. Morgenstern, B., Dress, A. and Werner, T. (1996), 'Multiple DNA and protein sequence alignment based on segment-to-segment comparison', *Proc. Natl Acad. Sci. USA*, Vol. 93, pp. 12098–12103.
12. Delcher, A. L., Kasif, S., Fleischmann, R. D. et al. (1999), 'Alignment of whole genomes', *Nucleic Acids Res.*, Vol. 27(11), pp. 2369–2376.
13. Delcher, A., Phillippy, A., Carlton, J. and Salzberg, S. L. (2002), 'Fast algorithms for large-scale genome alignment and comparison', *Nucleic Acids Res.*, Vol. 30, pp. 2478–2483.
14. Kurtz, S. (2003), 'A Time and Space Efficient Algorithm for the Substring Matching Problem', Technical Report, Zentrum für Bioinformatik, Universität Hamburg.
15. Altschul, S., Madden, T., Schaffer, A. et al. (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.
16. Primer3 (URL: [http://www-genome.wi.mit.edu/genome\\_software/other/primer3.html](http://www-genome.wi.mit.edu/genome_software/other/primer3.html)).
17. Cepheid (URL: [http://www.cepheid.com/pages/td\\_system.html](http://www.cepheid.com/pages/td_system.html)).
18. Badger, J. and Olsen, G. (1999), 'CRITICA: Coding region identification tool invoking comparative analysis', *Mol. Biol. Evol.*, Vol. 4, pp. 512–524.
19. Borodovsky, M., McIninch, J., Koonin, E. et al. (1995), 'Detection of new genes in a bacterial genome using Markov models for three gene classes', *Nucleic Acids Res.*, Vol. 17, pp. 3554–3562.
20. Molecular Beacon (URL: [http://www.idtdna.com/program/techbulletins/Molecular\\_Beacons.asp](http://www.idtdna.com/program/techbulletins/Molecular_Beacons.asp)).
21. Zemla, A., Slezak, T., Barsky, D. et al. (2003), 'Automated 3D protein structure predictions based on sensitive identification of sequence homology' (unpublished computer program under development at LLNL).
22. Protein Data Bank (URL: <http://www.rcsb.org/pdb/>).
23. Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.*, Vol. 147(1), pp. 195–197.
24. Pearson, W. R. (1991). 'Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms', *Genomics*, Vol. 11, pp. 635–650.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. (1997). 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.
26. Zemla, A. (2002), 'LGA program: A method for finding 3–D similarities in protein structures' (paper in preparation, server accessed at <http://PredictionCenter.llnl.gov/local/lga>).
27. Bower, M., Cohen, F. E. and Dunbrack, R. L. Jr (1997), 'Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modeling', *J. Mol. Biol.*, Vol. 267, pp. 1268–1282.

28. Rost, B. and Sander, C. (1994), 'Conservation and prediction of solvent accessibility in protein families', *Proteins*, Vol. 20, pp. 216–226 (URL: <http://cubic.bioc.columbia.edu/predictprotein/predictprotein.html>).
29. Hubbard, S. J. and Thornton, J. M. (1993), 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London (URL: <http://wolf.bms.umist.ac.uk/naccess/>).
30. CDC Laboratory Response Network (URL: <http://www.phppo.cdc.gov/nltn/pdf/LRN99.pdf>).
31. USDA Plum Island (URL: <http://www.ars.usda.gov/plum/>).
32. Gardner, S., Kuczmarski, T. A., Vitalis, E. A. and Slezak, T. (2003), 'Limitations of TaqMan<sup>®</sup> PCR for detecting divergent viral pathogens illustrated by Hepatitis A, B, C, E and HIV viruses and Human Immunodeficiency Virus', in press.
33. West Nile Virus transmission via blood transfusion (URL: <http://www.cdc.gov/od/oc/media/pressrel/r020927d.htm>).
34. Listeria outbreak (URL: <http://www.fsis.usda.gov/OA/recalls/prelease/pr090-2002.htm>).