Rapid Development of Nucleic Acid Diagnostics

J. PATRICK FITCH, SENIOR MEMBER, IEEE, SHEA N. GARDNER, THOMAS A. KUCZMARSKI, STEFAN KURTZ, RICH MYERS, LINDA L. OTT, THOMAS R. SLEZAK, ELIZABETH A. VITALIS, ADAM T. ZEMLA, AND PAULA M. MCCREADY

Invited Paper

There has been a significant increase, fueled by technologies from the human genome project, in the availability of nucleic acid sequence information for viruses and bacteria. This paper presents a computer-assisted process that begins with nucleic acid sequence information and produces highly specific pathogen signatures. When combined with instrumentation using the polymerase chain reaction, the resulting diagnostics are both specific and sensitive. The computational and engineering aspects of converting raw sequence data into pathogen-specific and instrument-ready assays are presented. Examples and data are presented for specific pathogens, including foot-and-mouth disease virus and the human immunodeficiency virus.

Keywords*—*Deoxyribonucleic acid (DNA), diagnostics, genomics, nucleic acid, ribonucleic acid (RNA), suffix tree.

I. INTRODUCTION

Life is the property that makes it possible to reproduce, to adapt to surroundings, and to take in food and create energy from it. These three attributes are also known as reproduction, self-regulation, and metabolism. Biologists have devised several methods for categorizing living organisms. Despite centuries of evaluation, the categories for labeling life continue to be updated as organisms are reclassified, new organisms are discovered, and new approaches to classification are proposed [1]. The traditional categories are based on similar phenotypes (observable attributes) and follow a sophisticated "if it looks like a duck and quacks like a duck, it is a duck" method. The past decade has introduced many new approaches to categorize life including similar genotypes (genetically similar at the nucleic acid level). One approach exploits certain genes that seem to occur in all living organisms. The 16S ribosomal gene is one such ubiquitous genetic marker. A phylogeny can be defined based on nucleic acid differences in the 16S gene and measured [2]. We mention this phylogeny because our pathogen detection approach is also based on nucleic acid differences. The target of our detectors is not the evolutionary changes in a gene like 16S, but rather association with specific symptoms in a host. All living organisms are potential hosts ranging from bacteria to humans. We focus on human, animal, and plant hosts. The threat pathogens we focus on are bacteria and viruses.

By most definitions, viruses are not alive. A virus is a parasite that replicates only within the cells of a living host and does not self-regulate or metabolize. A virus is a small (15 to 300 nm) infectious agent that is a complex combination of proteins and nucleic acids. Because all viruses have either ribonucleic acid (RNA) or deoxyribonucleic acid (DNA), we can apply to viruses the same nucleic acid approach that we use for pathogenic bacteria.

The goals for pathogen signature development include a high degree of specificity, compatible with several available detection instruments, validated through interlaboratory exercises, and available for dissemination through a network that can impact public health—the Centers for Disease Control (CDC) Laboratory Response Network, for example. The conventional approach to developing pathogen signatures is to focus on a limited number of candidates that are generated through biological experiments and/or existing knowledge. The next step is to do focused screening of the candidates in a limited number of tests for both detection and specificity. Our approach (see Fig. 1) differs in several respects including computer-assisted generation of a large number of candidate signatures, computer-based preliminary screening and verification of instrument compatibility, and a high-throughput

Manuscript received March 15, 2002; revised July 15, 2002. This work was supported in part by the Lawrence Livermore National Laboratory (LLNL) under the Laboratory Directed Research and Development Program and the Chemical and Biological National Security Program of the Nuclear National Security Agency, and in part by the U.S. Department of Energy under Contract W-7405-Eng-48.

J. P. Fitch, S. N. Gardner, T. A. Kuczmarski, L. L. Ott, T. R. Slezak, E. A. Vitalis, A. T. Zemla, and P. M. McCready are with the Chemical and Biological National Security Program, LLNL, University of California, Livermore, CA 94550 USA (e-mail: jpfitch@llnl.gov).

S. Kurtz is with the Center for Bioinformatics, University of Hamburg, Hamburg, Germany.

R. Myers is with the Centers for Disease Control and Prevention, Atlanta, GA 30303 USA.

Digital Object Identifier 10.1109/JPROC.2002.804680

approach to doing the wet biochemistry of testing the agents for sensitivity and specificity. This approach exploits much of the high-throughput automation and bioinformatics developed as part of the Human Genome Project [3], [4].

II. Background Biology

Nucleic acids (DNA and RNA) are the parts list and assembly instructions for cells and for parasites like viruses that invade them. The basic units of inheritance (genes) are composed of nucleic acids. Even though nucleic acids represent less than 5% of the dry mass of a typical cell, nucleic acids control the production of proteins that make up about 75% of the dry mass. The information contained in the nucleic acids influences when and how cells respond to environmental conditions through the production of proteins. Nucleic acids store the "parts list" for protein structure and function, and they dynamically interact with proteins and other molecules to regulate the timing and amount of protein production. Nucleic acids are therefore a logical place to look for signatures for identification of specific bacteria and viruses.

We provide a brief primer on the biology necessary to present nucleic acid diagnostic development. For more thorough and increasingly complex coverage of the subject, see [5]–[7]. DNA is a macromolecule built from repeating subunits. The subunits are composed of a nitrogenous base, a sugar, and a phosphate group generically denoted dNTP for deoxyribonucleotide triphosphate. The nitrogenous base is one of adenine (A), cytosine (C), guanine (G), or thymine (T) with the associated deoxynucleotides denoted dATP, dCTP, dGTP, and dTTP. The dNTPs can be joined along a sugar-phosphate backbone to form a single strand of DNA with the bases occurring in any order. "N" in dNTP denotes any of the four bases. The dNTPs and the strand have an orientation based on that of the carbon atoms in the sugar. One end of the strand is designated five prime and the other three prime, 5′ and 3′, respectively. The list of bases in a strand of DNA is known as the DNA sequence and might appear as "5′-CGCGCTCCCTGAACC-3′." Single-stranded DNA (ssDNA) is somewhat fragile. DNA usually occurs as a double strand (dsDNA) with the strands joined at each base by hydrogen bonds to a complementary base on the opposite strand (see Fig. 2). The base pairs in double-stranded DNA must occur as A-to-T or C-to-G. The strands are also antiparallel—e.g., "3′-GCGCGAGGGACTTGG-5′," for the earlier example. The two strands tend to twist into the familiar double helix shape associated with DNA.

As with DNA, RNA is a macromolecule built from repeating subunits. DNA has one less oxygen atom (deoxy) in the ribose sugar than RNA. In RNA, the base uracil (U) replaces thymine, and the uracil complement is adenine (A). The subunits for RNA are composed of a nitrogenous base, a ribose sugar, and a phosphate group generically denoted NTP (note that dNTP was used for DNA). The NTPs are ATP, CTP, GTP, and UTP. RNA often occurs as a single strand. The single-stranded structure is less stable than the double-stranded DNA structure, and it may fold back onto itself (see Fig. 3), forming a hairpin loop [8]. For analysis and manipulation, RNA is often copied into a complementary strand of DNA known as cDNA. The single-stranded cDNA can be paired into a double strand, producing a relatively stable record of the RNA sequence.

The nucleic acid sequence for a specific target pathogen is derived using a process where numerous overlapping fragments are chemically "read" and computationally sorted and combined based on their similarity [9]–[11]. A complete list of nucleic acid bases that meet an agreed-upon quality metric is referred to as finished sequence. Preliminary sequence data are called draft or raw sequence. There are many databases

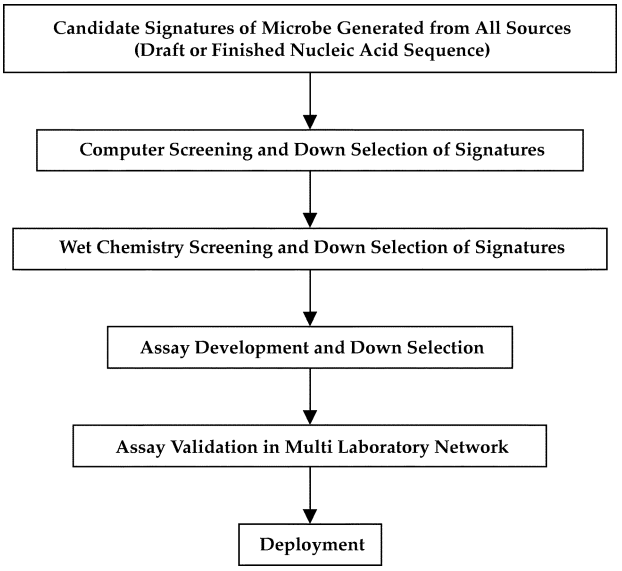


Fig. 1. Computer-assisted design of nucleic acid diagnostics for infectious diseases. Computer algorithms are used to identify the nucleic acid sequences that represent the target pathogen or a consensus composite of many strains of the target but are not found anywhere else in nature. Increases in the size and accuracy of nucleic acid sequence databases allow better computational signature recommendations. Assay development follows a similar approach with testing of a candidate assay for inclusion in target pathogen strain panels, exclusion from near neighbors, and no false alarms from complex environmental samples.

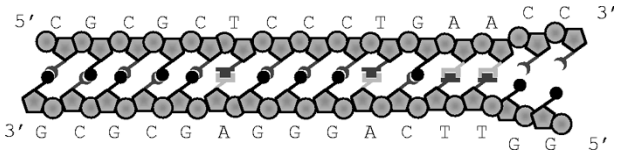


Fig. 2. DNA with complementary base pairs associated with the 5′-CGCGCTCCCTGAACC-3′ top strand (by convention 5′-left to 3′-right on top). The final two base pairs (both C-G) are not completely hybridized between the two strands.

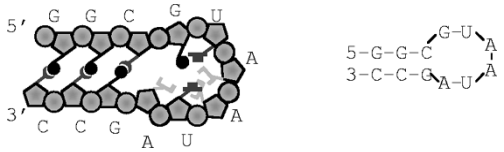


Fig. 3. RNA can have hairpin loops, especially in ribosomal RNA. The hairpin loop GUAAUA is embedded in 5′-GGCGUAAUAGCC-3′ as was modeled and measured using nuclear magnetic resonance by Foutain, *et al.* [8].

```
7621 gacgacatcg tggtggcaag cgattatgat ctggactttg aggccctcaa gcctcacttc
7681 aaatctcttg gccaaaccat tactccagct gacaaaagcg acaaaggttt tgttcttggt
7741 cactccatta ctgacgtcac tttcctcaaa agacacttcc acatggatta tggcactggg
7801 ttttacaaac ctgtgatggc ctcgaagacc ctcgaggcta tcctctcctt tgcacgccgt
7861 gggaccatcc aggagaagtt gatttccgtg gcaggactcg ccgtccactc cggaccagac
7921 gagtaccggc gtctctttga gcccttccag ggcctctttg agattccaag ctacagatca
7981 ctttacctgc gttgggtgaa cgccgtgtgc ggcgacgcat aatcccccaa acatcacaat
8041 tggcacaatg attttggggc gcgcgacgcc gtgggagtga aaagcccgaa agggctttcc
```

Fig. 4. Nucleic acid sequence data available on-line from Genbank [12] with accession number AJ133 357. These data are from the foot-and-mouth disease virus—a linear strand of RNA. The sequence data begins at base 7621 from the 5′ end selected for reference. The circled thymine (t) base is at position 7700 (see Fig. 9). The three sets of underlined bases are part of the diagnostic signature we developed for this virus.

that make sequence data available on-line, including Genbank [12], The Institute for Genomics Research (TIGR) [13], the Sanger Centre [14], and the Joint Genome Institute [15]. For example, a portion of the foot-and-mouth disease virus (FMDV) genome can be downloaded by going to Genbank at the National Center for Biotechnology Information (NCBI) [12] and using a nucleotide search with accession number AJ133 357 in the search window. The sequence data have 8115 RNA bases from a linear strand of FMDV, strain C, isolate c-s8c1 [16]. A subset of the sequence ranging from base 7621 to 8100 is given in Fig. 4.

III. Detection Using the Polymerase Chain Reaction

In nature, the complementary base pairing of two strands of nucleic acids helps create a more stable molecule. We can also exploit the complementary base pairing in diagnostics. Basically, a strand of nucleic acids can be designed to fish a complementary target strand of nucleic acids out of a complex background. The process of joining two complementary strands of nucleic acids together is called hybridization. The complementary base pairing of nucleic acids makes the hybridization specific for the sequence and forms the basis of DNA chips and microarrays.

The polymerase chain reaction (PCR) is a biochemical process for making copies of DNA using various enzymes and temperature cycling [17]. A template of DNA is denatured (separated into two single strands) by raising the temperature, two pieces of complementary DNA fragments (primers) hybridize to the denatured strand(s), the temperature is lowered, complementary bases beginning at the 3′ end of the primers fill in the rest of the double strand, and the process is repeated. The number of copies of DNA between the two primer sites, known as the amplicons or amplified strands, roughly doubles with each cycle (Fig. 5). Because G–C base pairs have three hydrogen bonds and A–T base pairs have only two, the denaturing temperature for dsDNA increases with G–C content. When multiple PCR primers are used in one reaction, differences in melting temperature can make the reactions incompatible.

PCR can also be used to detect specific fragments of DNA. As the PCR reaction makes copies between the two primers, a third DNA probe can be designed to hybridize to the template strand between the primers. As the rest of the second strand of DNA is being generated in the PCR replication process, the enzymes interact with the third probe. A specific approach to doing this is known as TaqMan [18]. For the TaqMan assay, the third probe has a fluorescent label attached at the 5′ end of the probe. At the 3′ end of the probe is a quenching molecule that suppresses the fluorescent reporter. As replication of the DNA occurs in the PCR process, the fluorescent label is released from the quenching molecule by the *Thermus aquaticus* (Taq) enzyme, and the fluorescent signal increases (see Fig. 6). The quencher and the fluorescent molecules are separated only if DNA replication occurs. The design of Taq assays therefore requires the design of three nucleic acid probes—two PCR primers and one fluorescent probe located between the primers.

IV. A New Paradigm for DNA Signature Construction

The historic approach to creating DNA signatures has been a wet-lab process. Knowledge about gene functions, a unique protein or toxin for instance, was used to search for signature candidates in the sequence databases for the species and genes of interest. Some alternative techniques have been developed that biochemically compare DNA from different organisms. Techniques like suppressive subtractive hybridization (SSH) [19] can yield unique fragments from pairwise comparisons of different genomes that could then be sequenced. As with the gene function approach, the SSH sequence data were manipulated with simple computational tools and manually compared with other target strains and to other DNA that might interfere. The candidate signatures were then tested against the target pathogen and other available strains.

In the summer of 2000, the Department of Energy (DOE) Chemical and Biological National Security Program (CBNP) initiated a project at Lawrence Livermore (LLNL) and Los Alamos National Laboratories to create a Biological Aerosol Sentry Information System (BASIS) to be deployed for biodefense at the 2002 Winter Olympic Games in Salt Lake City, UT. Fig. 7 shows part of the system as deployed in Salt Lake City. Environmental samples were collected on dry filters, the filters were segmented and biochemically processed to release nucleic acids from the bacteria and viruses on the filter segment, and then PCR was used to determine if specific pathogens were present. Some of LLNL's responsibilities for BASIS at the Olympics were to implement a field laboratory for processing environmental samples and establishing validated assays for the laboratory. This system was designed to detect a number of pathogens

Fig. 5. PCR is a biochemical method to copy or detect DNA. The orientation of the right end of the strand is given by the 3 or 5 in the strand name. The primers are given in lower case, T is for template, RP is for reverse primer, FP is for forward primer, and AS is for amplified strand.

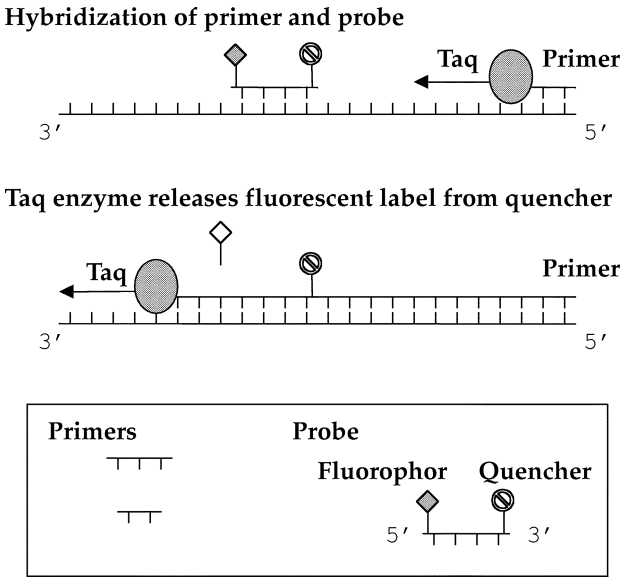


Fig. 6. The TaqMan approach for using PCR to detect amplification of a specific piece of DNA. In addition to the two primers for bracketing the PCR reaction, a third probe is designed to attach to the PCR amplicons. Hybridization of the third probe to the PCR amplicons is detected when the *Thermus aquaticus* (Taq) enzyme cleaves the fluorescent label away from the quencher on the probe.

on the CDC disease and agent lists [20], including several from the high-priority "Category A" list in Table 1.

In August 2000, a variety of traditional wet-lab signature development methods were failing to yield signatures specific to the most important pathogen for the BASIS deployment. In testing more than 1000 candidates, all cross-reacted with neighbor species commonly found in background environments within the continental United States. The leader of the signature development project challenged the informatics support team to help. This prompted a quick and dirty test, using the Basic Local Alignment Search Tool (BLAST) [21] algorithm to compare the entire genome of the target pathogen against all the microbial genomes then in Genbank (about 85). Several gigabytes of raw BLAST output were parsed, and all exact match regions of the pathogen genome with another genome were removed or "masked out" of the pathogen genome. All remaining pathogen sequence was then considered "potentially unique" to the target pathogen.

The MIT Primer3 program [22] was used to design potential signature primer pairs on each fragment of sufficient length. Nearly 4000 potential signature candidates on the ∼5 Mbase pathogen genome resulted from this crude BLAST parsing effort. We chose 400 at random, ordered the PCR primer oligos, and screened against the panel of near neighbors that had wiped out all 1000 of the wet-lab candidates. Several dozen of the candidate signatures passed this screening, and four survived rigorous testing of the complete set of DNAs used to ensure a high degree of specificity and have been successfully fielded. We realized that a threshold had been crossed: it was now feasible to apply whole-genome analysis techniques to the field of DNA signature development.

V. Computational Design of DNA Signature Candidates

We were encouraged by the initial success of the crude BLAST parsing demonstration. However, we realized that a lot of work needed to be done to make even a modest prototype of an automated DNA signature generation pipeline. Fig. 8 shows a simplified diagram of the system that we developed over the 18 months that followed the initial success.

Important distinctions must be made between the signature development of bacterial and viral pathogens, and what techniques are appropriate depending on whether or not at least one whole genome sequence is available. Viral pathogens are small: FMDV is about eight kilobases (kb) in length. Variola (the causative agent of smallpox) is one of the largest viruses, at over 185 kb in length. Bacterial pathogens such as *Bacillus anthracis* (the causative agent of anthrax) and *Yersinia pestis* (the causative agent of plague) are often in the range of three to five million bases (3–5 Mb) in length. A very important difference between bacterial and viral pathogens (for the purposes of signature creation) is that viral pathogens lack DNA repair mechanisms. Hence, viruses mutate at rates that can exceed 1000 times that of bacteria. An example of virus mutation is the annual progression of the influenza virus, creating a different genome each year and therefore usually requiring a different vaccine.

The time and cost to sequence a genome is directly proportional to the size (number of bases) of the genome. Hence, it is no surprise that there are often many whole virus genomes sequenced and comparatively few whole bacterial pathogen genomes sequenced. It is fortunate that virus genomes are the

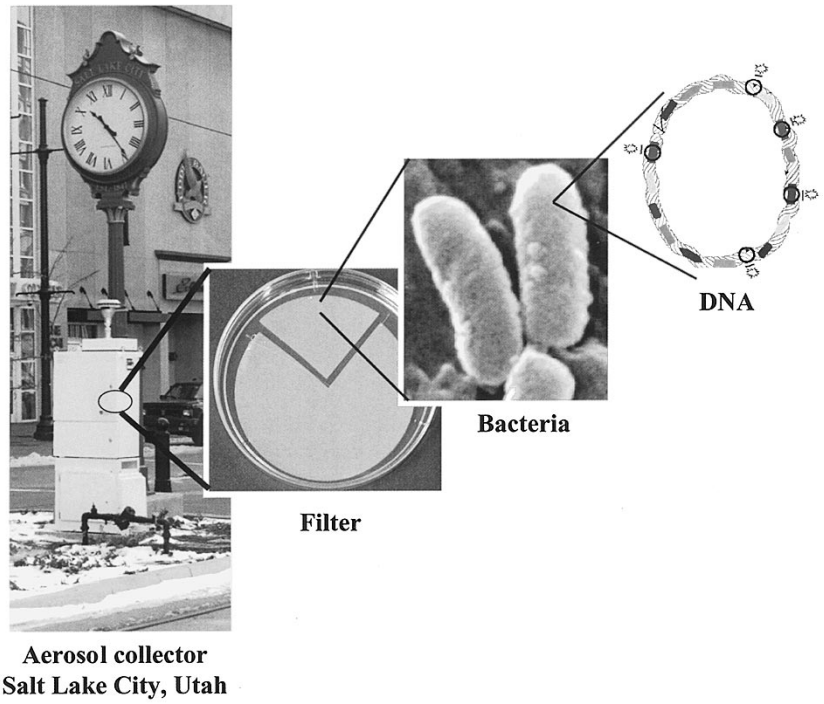


Fig. 7. The system from aerosol collector through to multilocus DNA signatures. The BASIS aerosol collector is a slight modification of a commercially available EPA air quality monitoring system. Material courtesy of the BASIS project leaders: Wiley Davidson (Los Alamos National Laboratory) and Dennis Imbro (LLNL). Bacilli bacteria photograph of *Pseudomonas aeruginosa* is courtesy of Janice Carr, CDC, PHIL #232.

Table 1
Category A Biological Diseases and Agents from the CDC List [20]. The CDC also maintains lists for Category B and C diseases and agents. The A list diseases represent special public health challenges or potential impact

Name	Causative Agent	Comments
Anthrax	Bacillus anthracis	Sporulating bacterium
Botulism	Clostridium botulinum toxin	Toxin from bacteria
Plague	Yersinia pestis	Bacterium
Smallpox	Variola major	Virus
Tularemia	Francisella tularensis	Bacterium
Viral hemorrhagic fevers	Filoviruses and Arenaviruses	Examples include Ebola and Marburg for Filoviruses and Lassa and Machupo for Arenaviruses

least expensive to sequence, because the fast mutation rates require multiple strains to be sequenced. Length of DNA sequence also affects the scalability of DNA signature development algorithms. Techniques that may be computationally feasible for small viral genomes often fail on the genomes of bacteria that may be 1000 times larger. These economic and scientific realities of DNA signature construction influenced our signature development process (see Fig. 8). The pipeline has two entry points, nominally one for viral and one for bacterial pathogens. The viral pathogen entry point can also be used as a last resort for bacterial genomes where only a few gene fragment sequences are available, but we will not discuss these sorts of degenerate (but real) cases further.

A. Overview of DNA Signature Pipeline

Virus strain genomes are preprocessed to extract a single "consensus gestalt" genome. Bacterial genomes are assumed to have (at best) one completed genome. In either case, we employ an efficient algorithm to compare the pathogen target against all other sequenced bacterial and viral genomes, "masking out" in the pathogen all DNA sequence over K bases in length that "match" (exactly or close enough for PCR) one or more of the other nontarget genomes. We then mine the remaining "candidate-unique" DNA to search for suitable signature primers. These are checked electronically and stored in a database, and a subset is ordered using a computer-generated procurement process.

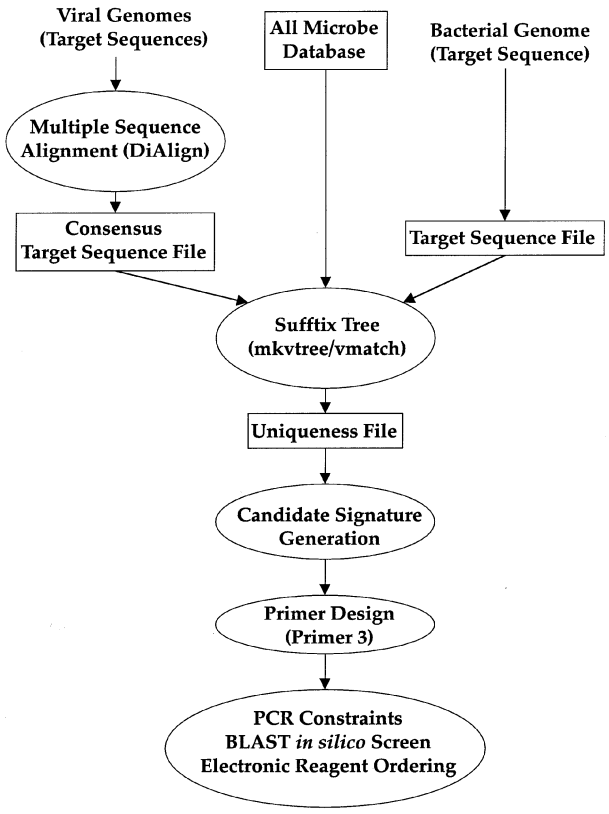


Fig. 8. Automation of the development process creates an efficient pipeline for DNA signature design. Nucleic acid sequence data through to ordering of reagents for experimental testing are performed electronically. Software is from a variety of sources: DiAlign [26], mkvtree/vmatch [28], and Primer3 [22].

Upon delivery from an external vendor, they are screened in the wet laboratory against the target DNA and against a panel of near neighbors and environmental backgrounds. Successful ones are sent to the CDC (or other collaborator) for additional testing and validation. Validated assays are used in both public health and biodefense applications.

B. Viral Pathogen Front-End Processing

Viral pathogens typically have at least one complete genome sequence available, as well as patchwork gene-sequencing entries. For example, go to http://www.ncbi.nlm.nih.gov and type "West Nile Virus." A summary of all West Nile Virus Genbank entries is returned. There will also be summaries of everything that mentions West Nile Virus—some relevant to signature development and some not relevant. The general technique of multi-sequence alignment (MSA) is needed to "line up" all the viral sequences with each other. An example MSA for six strains of FMDV is given in Fig. 9. Note that normal viral mutations are probably indistinguishable from sequencing errors or even some classes of assembly errors. There are many MSA algorithms available. See [23] for a brief comparison, or use your favorite Web search engine to do your own up-to-date review. Unfortunately, many of the MSA algorithms and codes do not scale well for even modest genome lengths or for enough strains in the alignment. We located a very good MSA algorithm called DiAlign [24]–[26] that scales better than any other we tested.

Available through the Web and running on a cluster of eight Alpha computers, DiAlign also has some scaling limitations. In an attempt to align 93 instances of West Nile Virus genomes and gene fragments, DiAlign failed on memory limitations. The alignment worked after trimming the request from 93 down to 57 instances. The trimmed request worked but took nearly five days. System load information was not available, so these runs might have worked better at another time with a lighter load. In a different request that ran for one week, DiAlign did not complete alignment of six Variola (smallpox) genomes. The Variola request was manually divided into three overlapping sections and then run successfully. It is clear that better MSA algorithms would greatly benefit the design of viral pathogen DNA signatures.

Assuming that MSA can be completed successfully, we then process the aligned genomes into a "consensus gestalt" sequence. Most simply stated, this places a "dot" (period) in all positions of the output alignment if the bases in that position are not in agreement. The base is capitalized if all the bases in that position match and the position is part of at least an 18-base consecutive run of matches. Otherwise, the base is lowercase (i.e., there is a match at that position but no run of 18 matches). An 18-position consecutive run of bases is used because this is the minimum length for a forward or reverse PCR primer.

C. Bacterial Genome Processing, and Continuation of Viral Processing

Fig. 8 shows that bacterial pathogens take a different route through the candidate signature pipeline. Even in the (rare today) case where more than one full-length bacterial genome is available, current MSA algorithms cannot handle input of that length. In addition, bacterial genomes violate the subtle assumptions of colinearity that underlie most MSA algorithms. Colinearity is violated because of real biological issues, including genome rearrangements, gene duplication, and other events.

There is no magic involved in determining candidate unique DNA signatures. Indeed, if the exact sequence of all organisms was somehow made available, it would be a rather trivial exercise to compare a pathogen target's DNA against that of all other organisms and determine what DNA (if any) was truly unique. Any such unique stretches could then have appropriate signature assay reagents designed and the constraints of a given assay format applied to the output. Today we have only a tiny fraction of the earth's organisms sequenced. Thus, the best we can do for a target pathogen is to calculate what stretches of DNA are *not yet known* to match that of any other organism. We do this by finding all the currently known matches and masking them out from the pathogen target genome.

Clearly, signatures designed under these conditions of very minimal knowledge are doomed to "erode" over time as either more organisms become sequenced or the signatures are tested in more environments. Techniques to update and verify

signatures are needed as well as new designs. Fortunately, we have repeatedly demonstrated that the computational derivation of signature candidates is far more effective than previous processes at delivering an enriched pool of highly specific DNA pathogen signatures that perform robustly in complex environmental samples. This enrichment has allowed the bench chemistry screening to focus on a smaller subset of candidate signatures that have a higher probability of success in the field.

D. The Use of Suffix-Trees to Determine Potentially Unique Sequence

The most efficient algorithms for finding substrings that match between two inputs are known as suffix-tree algorithms [27]. We have used an extremely efficient suffix-tree implementation known as *mkvtree/vmatch* [28]. The author, S. Kurtz, added several features that met application-specific needs for signature development and supplied the project with numerous binary executable updates and performance enhancements. As designed, a suffix-tree expects to compare two strings. We realized that we could compare the pathogen target (either a bacterial genome or a consensus gestalt sequence from MSA of virus strains) against all other bacterial/viral genomes. The "all other" microbe genomes are concatenated together into a huge "synthetic genome" for the purposes of the suffix-tree run. For instance, in a recent run we compared a 5 Mb bacterial pathogen genome against more than 700 Mb of other available sequence genomes. This required five hours of processing time. From the output of *mkvtree/vmatch* we create a "uniqueness gestalt" file that is the substrate for our further analysis.

We currently use two screening databases ("all_microbes" and "all_virus") for comparison with target genomes. As the name implies, all_microbes contains the best available completed or draft sequence for all microbes that we can acquire publicly or by collaboration [29]. The all_virus database is a compromise that includes a selection of the single-reference viral genomes from NCBI [30] and other strains/isolates as appropriate to our work. At the time of this writing, all_virus contains 630 entries and is 127 Mbytes in size. The all_microbes database contains 162 entries and is more than 675 Mbytes. We typically screen all pathogens against both databases. A typical pathogenic bacterium completes in four to six hours running in parallel on more than 20 CPUs. Maintenance of these databases can be nontrivial.

It is important to note the value of having sequence for appropriate near neighbors to the pathogen of interest. For example, *Yersinia pestis* evolved recently from another bacterium known as *Yersinia pseudotuberculosis*. When *Y. pestis* is compared against the 700-Mb all_microbes database without *Y. pseudotuberculosis*, about 33% of *Y. pestis* can be masked as nonunique signatures. When comparing *Y. pestis* directly with the sequence of the single genome for *Y. pseudotuberculosis*, more than 90% of the sequence is masked as nonunique signatures. It is also worth noting that it can be nontrivial to determine exactly what is an "appropriate near neighbor" for a species. We were once told that *Legionella pneumophilia* [31] was believed to be a near

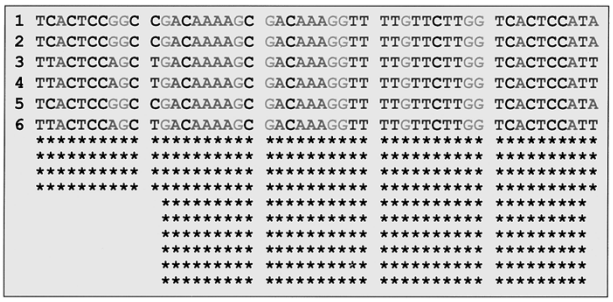


Fig. 9. MSA of strains of FMDV. The bottom sequence (6) is from accession number AJ133357, strain C, isolate c-s8c1, [16]. The first base is position 7700 in the source sequence (see Fig. 4). MSA helps identify highly conserved sequence regions that may be good nucleic acid signature candidates. The number of "*" in the display is proportional to the quality of the alignment match.

neighbor to *Coxiella burnetti* [32]. This prediction was based on 16S RNA phylogeny [33]. However, *mkvtree/vmatch* showed that only 1% of the sequence was in common, which was no better than a random pick. Caution is advised when making decisions about what are appropriate near neighbors.

A fortunate side effect of the use of a Suffix-Tree algorithm to determine candidate uniqueness is that we are not limited to using only completed genomes or finished DNA sequence. Our method works well on draft genomes and even raw, unassembled sequencer reads of good quality. Candidates that span assembly gaps will be missed, but that is typically a tiny fraction of the potential signatures for a bacterial genome. Consequently, for the same sequencing cost it may make more sense to draft three strains at $3\times$ quality instead of drafting a single strain to $10\times$ quality.

E. Mining the Uniqueness Gestalt for Candidate Signatures

Our unique DNA signatures are designed for public health and other applications. Most of the end users are requesting assay formats supported by commercial sources—TaqMan is our principal focus [18]. TaqMan assays require the design of three DNA fragments (oligos): Forward and Reverse PCR primers and an internal hybridization oligo for positive confirmation of the amplified fragment. Thus, the primers might not themselves have to be unique across all genomes, but the internal oligo would need to be specific to the target genome.

We designed custom software to mine the uniqueness gestalt output of the *mkvtree/vmatch* program and search for regions of apparently unique DNA large enough to land a Forward [F] primer, Reverse [R] primer, and internal Oligo [O] probe subject to a number of parameterized constraints, including:

1) total length of amplified product (including F, O, R) not to exceed 300 base-pairs;
2) minimum length of F and R is 18 base-pairs;
3) minimum length of O is 22 base-pairs;
4) at least three base-pairs must separate F from O and O from R.

Note that a single contiguous run of 64 apparently unique bases could, in theory, support all three components. More commonly, the algorithm will find two or three smaller runs of apparently unique bases within the maximum window

size. The gestalt fragment is prepared for the next step, searching for suitability of PCR primers, by replacing all "dots" in the gestalt fragment with "N"s, which will be ignored.

There are literally dozens of PCR primer prediction algorithms, all based on predictions of DNA melting temperatures. Unfortunately, even when presented with identical inputs, the algorithms do not always provide consistent output. After some experimentation, we chose to use MIT's Primer3 [22] because it had both a Web interface and available source code allowing us to batch it into our pipeline. Many other primer design programs we considered lacked any capability of batch operation, rendering them unsuitable for our purposes. Primer3 has dozens of control parameters, and considerable experimentation was required to get the parameters set properly for our needs. Our application had stricter melting temperature constraints than for ordinary PCR applications that could use the Primer3 defaults.

Many potential candidates for unique DNA signatures fail in the primer-picking step because the apparently unique run(s) of DNA do not contain the right base-pair composition to meet the strict parameters of the assay to be used (TaqMan in our case). We note that unsuitability for TaqMan does not necessarily preclude that DNA target from being suitable for other diagnostics applications with different parameters. We have not yet had an opportunity to demonstrate this in practice. However, the unique DNA signatures have been used to evaluate what makes an organism unique—and this is a valuable research tool.

F. PCR Rules of Thumb: The Final Hurdle

Signature candidates that pass the Primer3 step result in a {F, O, R} tuple that needs a final screening pass of what we termed "PCR Rules of Thumb." This is a collection of folklore based on the physics of the PCR process and includes such things as:

1) the internal Oligo F cannot begin with a G;
2) the internal Oligo F cannot have more Gs than Cs;
3) the internal Oligo F cannot have more than two consecutive Gs;
4) the F and R primers cannot have more than two Gs and/or Cs in the last five nucleotides (the 3′ end).

These rules knock out a rather surprising number of otherwise promising unique signature candidates. We note that these rules of thumb should not be applied too rigidly, and we confess to having loosened these (and other) rules when there were no other options.

G. A Final Check of Signature Candidate Survivors

Signature candidates that survive the above automated pipeline process are then checked using BLAST [21] to ensure that they do hit the target genome and do not cross-react with other bacterial/viral genomes. It is possible for the Forward primer of a signature or genome T to hit on one strand of genome X and the Reverse primer to hit on the other strand of Genome X. We would consider accepting the signature for T, since a valid fragment would not be generated for genome X. Other complex situations are possible in cases where unique signature candidates are scarce, as is often the case with viral pathogens. We discovered that one promising signature candidate for *Yersinia pestis* cross-reacted with cat DNA in wet-lab screening, which suggests that electronic screening against DNA from more complex organisms might be increasingly warranted in the future as more genomes of animals and plants are obtained.

H. Selection, Ordering, and Screening of DNA Signature Candidates

For most bacterial pathogen genomes we have worked on, there is a surplus of DNA signature candidates. This is most often due to the fact that we do not have sequence data for a close-enough near neighbor to mask out most of the target genome. *Yersinia pestis* and *Yersinia psuedotuberculosis* are the exception rather than the rule. We apply the following general rules for selecting candidates to order and screen (up to some user-set threshold):

1) select any candidates that lie on genes known to be involved in virulence;
2) select any candidates that lie on genes with known function;
3) randomly select candidates that lie on genes;
4) randomly select candidates that do not lie on genes.

We do not have enough data to prove that selections based on known function or association with genes necessarily deliver better detection assays than ones chosen from anonymous "junk" DNA.

Our system generates oligo order files for the selected candidates, in formats suitable for various oligo vendors. We note that we have ordered oligos from a variety of sources with varying consistency and stability. In some cases we have even received our oligos in incompatible formats to our processing queue. In some ways, the absence of reliable oligo sources was the most painful part of the entire system. Assuming that oligos can actually be delivered as ordered, in the specified format and the agreed-upon price, it is time for wet-bench screening of the candidate signatures. We have constructed a significant infrastructure for tracking the results of signature screening, including a complex database and numerous large and cumbersome Web interfaces. The details are outside the scope of this paper.

Once received, the oligos undergo PCR screening, in an effort to understand how a given primer set will behave when exposed to different kinds of samples in a diagnostic environment. Our current schema includes an initial testing phase of 500–1000 primer pairs against one target strain that the oligos were designed to recognize and one near neighbor strain that we want to ensure is not detected, as this would indicate a "false positive" of the diagnostic. This first phase routinely triages the initial data set down to less than 100 primer sets, and is considered a manageable number to do in-depth screening. The 100-primer set is then tested against as many near neighbor strains as is possible to identify, usually three to six, to further rule out nonproductive reagents.

This smaller primer set is reacted with a target strain, ensuring that the final diagnostics can recognize the diversity of the organism that has been found in nature. The primer sets are again rearrayed or reorganized so that only the highly specific primers are continued in the testing. The next series of bench screens evaluates the primer sets with

1) general microbial diversity seen in nature;
2) organisms that exhibit similar clinical symptoms;
3) DNAs from higher organisms that may be present in samples collected from the field;
4) complex environmental backgrounds, the composition of which is most uncharacterized.

The complex environmental backgrounds have been considered the most discriminating sample set for screening. The rate of genetic exchange in the environment is great, implying that a given piece of DNA may be found in an organism distantly related to our target organism. Therefore, it is frequently the case that many primer pairs that did not cross-react with well-characterized pure DNAs will cross-react against field samples that do not contain the target organism. The environmental background sample set also illustrates the reason why there have been many environmental false positives when using assays that have not been screened in this manner. The final primer pairs that successfully pass all of the above screening are frequently less than 20 of the original 1000. Our false positive rate for screening assays derived from the above process against real field samples is currently 1 in 10 000. Because there is still some probability that there is an organism in the environment we have not represented in our screening process, we employ an assay panel or multiple different assays to confirm the identity of a given target organism. The current process for screening field samples challenges the sample with a single signature or screening. If this preliminary screening is positive, the sample is reacted with two to five additional signatures, to ensure that the rest of the microbial genome is present for the given target organism. If known, the virulence genes are also examined. Associating the diagnostic with virulence provides the ability to discriminate between a virulent strain and a "hoax" strain.

VI. Turning Signatures Into Validated Assays

Our CDC collaborators and the BASIS laboratory use the final DNA signatures to synthesize assays and optimize their use on different instruments used by the Laboratory Response Networks component laboratories and DOE deployments. They conduct formal validation tests for the assays that will eventually be made available for public health applications. These validation exercises involve multiple labs that represent the final user community. Each lab is sent kits of identical reagents, protocols, and blind samples. The originating laboratory collects and analyzes the results. Ironically, our lab at LLNL was preparing our *Bacillus anthracis* signatures for shipment in a validation test on September 11, 2001 when the terrorist attacks affected everybody's plans. Our signatures underwent more than six months of continuous testing on environmental samples before the validation exercise could be rescheduled.

The reader may have noticed that we make a distinction between a signature and an assay. To us, an assay is a signature that has been tuned for the particular requirements and constraints of a particular instrument used under a specific protocol. Thus, a TaqMan assay for signature 1 of *B. anthracis* on a Cepheid SmartCycler [34] will be different from the TaqMan assay for the same signature on an Idaho Technology Ruggedized Advanced Pathogen Identification Device (R.A.P.I.D.) [35]. It is an important point that tuning a signature for a particular instrument may involve as much as or more work than was initially involved in the signature creation itself. The BASIS laboratory has successfully used the Cepheid SmartCycler.

A. Some Realities of DNA Signature Development

The above description of computational development of DNA signatures for bacterial and viral pathogens may sound as if all that is involved is a computer and some ideas for new algorithms. A few additional obstacles need to be overcome:

1) achieving legitimacy in a rather closed field;
2) obtaining information about signatures already developed on pathogens of interest;
3) obtaining access to adequate DNA sequence, if not in public databases;
4) obtaining access to actual DNA/RNA template materials, for sequencing or screening;
5) obtaining collaborations with appropriate agencies to get theoretical signatures tested.

Some of these challenges are technical and some are sociopolitical. Even with a good track record in genomics and national security applications, some of the nontechnical challenges caused significant delays in the early stages of our research.

Many pathogens are Biosafety Level (BSL) 3 or 4 agents, or are even further restricted by U.S. law. For example, a signature for FMDV can be tested only at the U.S. Department of Agriculture (USDA) facility on Plum Island, NY [36]. A similar situation exists for many human pathogens—e.g., smallpox virus signatures must be tested in Atlanta at the CDC. Even if you are at a facility that has an accredited BSL-3 or BSL-4 lab, there are strict permit processes that must be adhered to in order to obtain materials.

Researchers wishing to engage in this field should start with pathogens that have a relative abundance of data available in Genbank. This increases the probability of having your work checked by experts on those pathogens. For some pathogens it is possible to compare new signature candidates against signatures already proven to work in extensive testing—information that usually requires biological expertise to acquire. Additionally, local plant researchers in need of diagnostics for various plant pathogens may be good collaborators for signature development. Getting adequate DNA sequence of the pathogens will be the primary obstacle to progress, but success with a plant pathogen would be a good first step toward proving the worth of your algorithm or system.

The issue of security also arises when working in pathogen detection. On one hand, signatures for nearly all pathogens

sequenced to date are publicly available in Genbank for all to see. On the other hand, some would argue that detection assays for public health diseases (plague, anthrax, etc.) should be strictly classified and available only to the military. In practice, for bacterial pathogens, there are sufficient signature candidates to allow public health to be released a reliable set while various military agencies can maintain their own set(s) in confidence. It is a clear intent of all involved in biodefense to obtain sufficient signatures of required virulence mechanisms for each pathogen that any attempts to engineer around all of the detection points would of necessity render the organism harmless to humans.

VII. Discussion of Limitations of the Current System and Plans for Future Work

Our current system is a usable prototype for highly automated development of candidate DNA signatures for bacterial and viral pathogens. To date, we have used it to process more than two dozen microbes. A number of our assays have been in continuous use in several national defense deployments since fall 2001, and others have entered the public health system. We have established collaborations with the CDC, the USDA, the Food and Drug Administration, the U.S. Army Medical Research Institute of Infectious Diseases, the Defense Threat Reduction Agency, the Department of Transportation, the Department of Justice, and many other agencies to work on detection assays and systems to meet their needs.

Our focus on deployable assays for complex environmental monitoring has exposed several limitations of our current system.

1) Our signatures are developed at a snapshot in time, based on the current state of our bacterial and viral screening databases and our algorithms. We do not automatically check our signatures for "erosion" as new or updated sequences enter into our system.
2) Acquiring new or updated sequences into our system is a manual process that can be rather time-consuming. Although we get most of our relevant sequence data from Genbank [12], a significant amount of bacterial sequence data (especially unfinished genomes) is found at TIGR [13], the Sanger Centre [14], and the Joint Genome Institute [15]. Still other genome projects in progress are reported only at the university where the sequencing is being performed.
3) Our PCR "rules of thumb" appear to be overly harsh, especially as applied to viral genomes. They frequently reject all signature candidates that survive to that point.
4) Our system is not integrated into a database, but relies on the Unix file system. In some cases this leads to problems. For instance, with some bacteria that do not have good near neighbors, tens of thousands of little files accumulate in one Unix directory.
5) Our pipeline runs all of the obviously parallel portions across multiple available central processing units. However, there is no overall scheduling mechanism to deal with running multiple pipelines at once on different organisms.
6) We have not yet automated the search for structural homology models for all proteins in a pathogen genome.

Our work in progress is addressing all of these limitations.

Future work plans involve extending our capabilities to generate protein detection signatures (antibodies and ligands) on a whole-genome basis and exploiting our consensus and uniqueness gestalt information as appropriate. Preliminary evidence of the feasibility of this approach is shown in Fig. 12 with a computer model of FMDV protein that highlights signature and conserved loci. We also plan to integrate available relevant information about all genes in these pathogens, to help in future host/pathogen interaction studies that could lead to improved techniques for pathogen detection within hosts.

VIII. Examples

A. Foot-and-Mouth Disease Virus

We were developing our system when a major outbreak of FMDV occurred in the United Kingdom [37]. The cattle industry in California, with its crowded feed lots along the I-99 corridor, would be especially hard hit by an outbreak of FMDV. This prompted us to look at using our system on a viral pathogen for the first time, which led to our locating the Dialign MSA tool.

Fig. 10 shows a segment of the consensus gestalt file for 16 Genbank full genomes of FMDV. We were readily able to show that the indicated region is the only region of the FMDV genome where there is sufficient conservation to detect all seven serotypes known to date [38]. Note that we had to use two degenerate bases, one each in the internal hybridization probe and reverse primer, in order to satisfy the Primer3 and PCR rules of thumb. These were determined by manual inspection of the gestalt file. We noted that the degenerate base in the reverse primer indicated an A/T potential single-nucleotide polymorphism that distinguishes the Taiwanese strain of FMDV from all others. FMDV is considered a "foreign animal disease" by the USDA; as such, it can be kept only at their Plum Island, NY, facility. We gave our assay to them in the summer of 2001 for testing, but the September 11 aftermath delayed testing until April 2002. At this time, our assay was tested and confirmed to detect all seven serotypes as predicted [39].

B. The Human Immunodeficiency Virus: Why Commercial Primers Failed to Work on All Sub-Types

When we examined an MSA of 27 human immunodeficiency virus (HIV) samples from Genbank (chosen to represent all major subtypes and geographical regions), we concluded that there was insufficient conservation to have any hope of detecting them all with a single DNA signature. Initially, we were surprised that a patented DNA signature for HIV existed [40]. The authors noted that there was difficulty detecting several subtypes with this assay. We mapped their assay onto our consensus gestalt file of the 27 HIV genomes, (see Fig. 11) and noted that their primers had been chosen on a region with a very low degree of conservation. We could examine the 27 input genomes and see that indeed, the subtypes

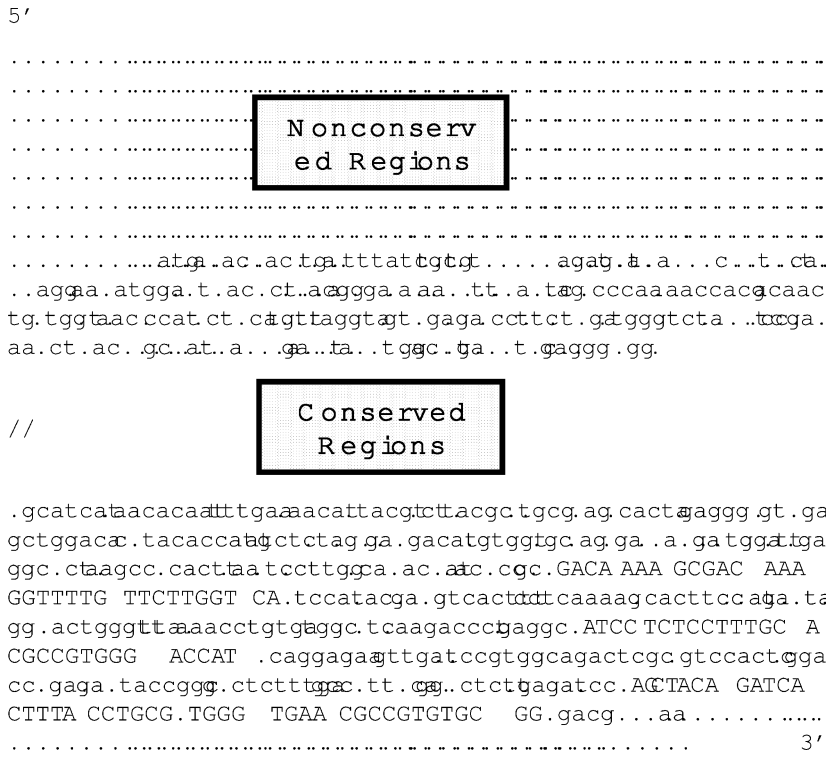


Fig. 10. A "consensus gestalt" view of 16 FMDV genomes. The forward primer, reverse primer, and probe TaqMan oligo hybridization locations are identified in capital letters.

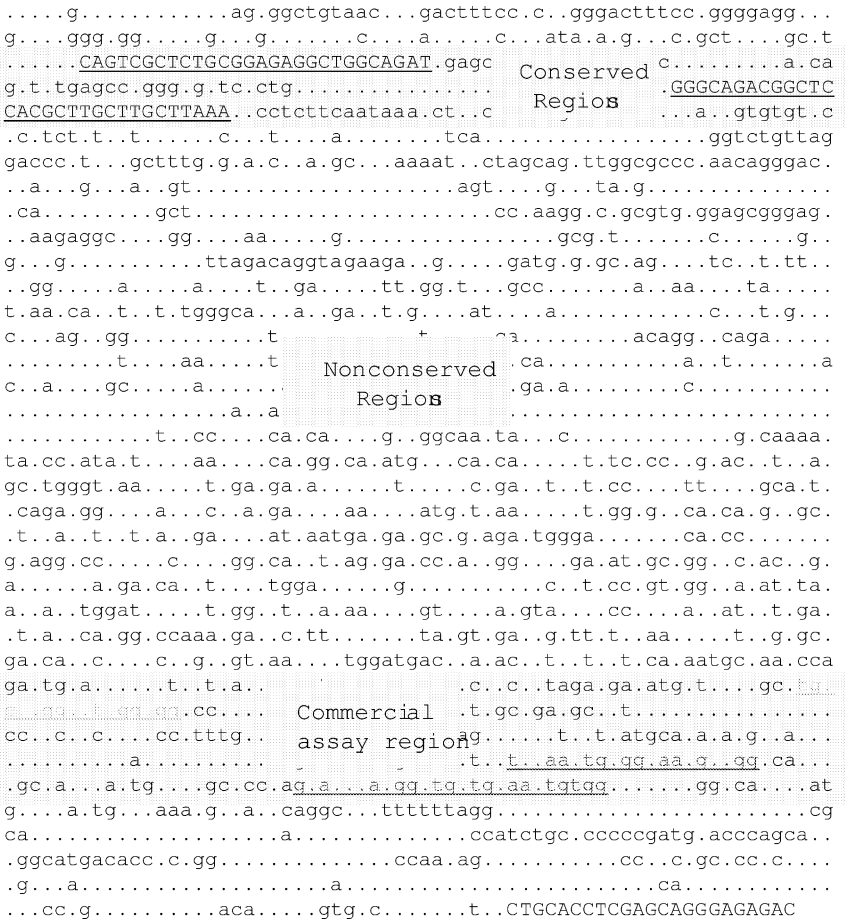


Fig. 11. A commercial HIV PCR primer pair mapped onto a "consensus gestalt" of 27 HIV genomes. Fast mutation rates for many viruses make design of an enduring nucleic acid signature difficult.

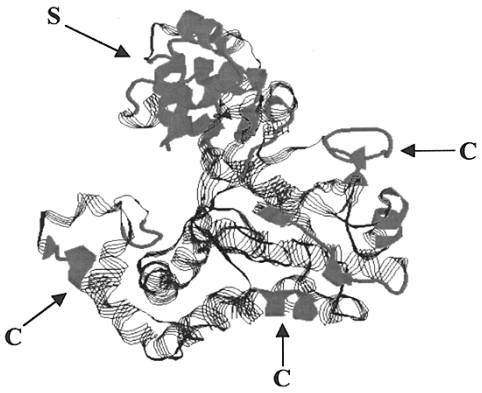


Fig. 12. Computational model of FMDV virus protein. The signature (S) and conserved (C) regions of the protein are labeled. These sites are potential protein signatures and vaccine targets.

that could not be detected had too many bases disagreeing with the primers for hybridization to occur.

We use this example to illustrate how the incredible mutation rate of viruses can over time render any DNA signature ineffective in detecting all variants. It also illustrates how we can use our consensus gestalt information to quickly determine any such "signature erosion."

IX. Summary

For the last several years we have used computer and automation tools to accelerate the design of highly specific nucleic acid diagnostics. Because our goal is to support public health objectives, the signatures are targeted for widely available instruments and use the well-understood TaqMan PCR assay. The various engineering and computer science methods used have greatly accelerated the diagnostic design process and improved the product. Many of our signatures have been used to monitor complex environments and have performed above expectations.

Our initial focus was bacteria, and we have recently expanded to viruses as well. Viruses represent a special challenge because of the often rapidly evolving nature of their genomes. As shown, there are virus targets that do not meet our initial diagnostic design strategy. We are currently expanding our process to address these limitations and to automatically incorporate new nucleic acid sequence data from all sources. We plan to apply a similar approach to the design of nonnucleic acid assays—e.g., antibody and ligand assays. Improvements in "postgenomics" tools for both measurement and computation in structural biology are enabling a similar pipeline [41], [42], [4].

Acknowledgment

The authors wish to thank members of the BASIS team, including the project leads W. Davidson (Los Alamos) and D. Imbro (LLNL). They would also like to acknowledge the talented pathogen team at LLNL that has contributed to their projects, including K. Montgomery, L. Danganan, J. Avila, and C. Strout. They also acknowledge their many collaborators around the world in the pathogen diagnostics community for continued teamwork in these important projects.

References

[1] C. R. Woese, "Interpreting the universal phylogenetic tree," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 15, pp. 8392–8396, July 18, 2000.

[2] K. H. Wilson, W. J. Wilson, J. L. Radosevich, T. Z. DeSantis, V. S. Viswanathan, T. A. Kuczmarski, and G. L. Andersen, "High density microarray of small subunit ribosomal DNA probes," *Appl. Env. Micro.*, vol. 68, no. 5, pp. 2535–2541, May 2002.

[3] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 15, 2001.

[4] J. P. Fitch and B. Sokhansanj, "Genomic engineering: Moving beyond DNA sequence to function," *Proc. IEEE*, vol. 88, pp. 1949–1971, Dec. 2000.

[5] J. P. Fitch, *An Engineering Introduction to Biotechnology*. Bellingham, WA: SPIE Press, 2002.

[6] B. R. Glick and J. J. Pasternak, *Molecular Biotechnology: Principles and Applications of Recombinant DNA*. Washington, DC: American Soc. of Microbiology, 1998.

[7] J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner, *Molecular Biology of the Gene*, 4th ed. Menlo Park, CA: Benjamin Cummings, 1987.

[8] M. A. Fountain, M. J. Serra, T. K. Krugh, and D. H. Turner, "Structural features of a six-nucleotide RNA hairpin loop found in ribosomal RNA," *Biochemistry*, vol. 35, no. 21, pp. 6539–6548, May 28, 1996.

[9] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Nat. Acad. Sci. USA*, vol. 74, pp. 5463–5467, Dec. 1977.

[10] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, "Large-scale and automated DNA sequence determination," *Science*, vol. 254, no. 5028, pp. 59–67, Oct. 4, 1991.

[11] J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller, "Shotgun sequencing of the human genome," *Science*, vol. 280, no. 5369, pp. 1540–1542, June 5, 1998.

[12] National Center for Biotechnology Information.. [Online]. Available: http://www.ncbi.nlm.nih.gov/.

[13] The Institute for Genomics Research (TIGR).. [Online]. Available: http://www.tigr.org.

[14] The Sanger Centre.. [Online]. Available: http://www.sanger.ad.uk.

[15] The Joint Genome Institute.. [Online]. Available: http://www.jgi.doe.gov.

[16] M. Toja, C. Escarmís, and E. Domingo, "Genomic nucleotide sequence of a foot-and-mouth disease virus clone and its persistent derivatives: Implications for the evolution of viral quasispecies during a persistent infection," *Virus Research*, vol. 64, no. 2, pp. 161–171, Nov. 1999.

[17] K. B. Mullis, F. A. Faloona, S. J. Scharf, R. K. Saiki, G. T. Horn, and H. A. Erlich, "Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction," in *Cold Spring Harbor Symp. Quant. Biol.*, vol. 51, 1986, pp. 263–273.

[18] Primer Express tutorial for real time quantitative PCR primer and probe design. *Applied Biosystems* [Online]. Available: http://www.appliedbiosystems.com/support/tutorials/taqman/.

[19] P. G. Agron, R. L. Walker, H. Kinde, S. J. Sawyer, D. C. Hayes, J. Wollard, and G. L. Andersen, "Identification by subtractive hybridization of sequences specific for *Salmonella enterica* serovar Enteritidis," *Appl. Environ. Microbiol.*, vol. 67, no. 11, pp. 4984–4991, Nov. 2001.

[20] Centers for Disease Control and Prevention. Biological Disease/Agents List. [Online]. Available: http://www.bt.cdc.gov/Agent/Agentlist.asp.

[21] National Center for Biotechnology Information.. Basic Local Alignment Search Tool (BLAST). [Online]. Available: http://www.ncbi.nlm.nih.gov/BLAST/.

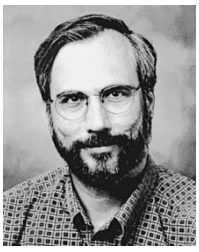
[22] S. Rozen and H. J. Skaletsky. (1996 and 1997) Primer3 software distribution. [Online]. Available: http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

[23] H. B. Nicholas, Jr., A. J. Ropelewski, and D. W. Deerfield, II, "Strategies for multiple sequence alignment," *BioTechniques*, vol. 32, no. 3, pp. 572–591, Mar. 2002.

[24] DiAlign 2: A novel algorithm for pairwise as well as multiple alignment of nucleic acid and protein sequences. [Online]. Available: http://www.gsf.de/biodv/dialign.html.

[25] B. Morgenstern, A. Dress, and T. Werner, "Multiple DNA and protein sequence alignment based on segment-to-segment comparison," *Proc. Nat. Acad. Sci. USA*, vol. 93, no. 22, pp. 12 098–12 103, Oct. 29, 1996.

[26] B. Morgenstern, K. Frech, A. Dress, and T. Werner, "DIALIGN: Finding local similarities by multiple sequence alignment," *Bioinformatics*, vol. 14, no. 3, pp. 290–294, Apr. 1998.
[27] P. Weiner, "Linear pattern matching algorithms," in *Proc. 14th IEEE Annual Symp. Switching and Automata Theory*, Iowa City, 1973, pp. 1–11.
[28] S. Kurtz, "Construction and application of virtual suffix trees," Ph.D. dissertation, Technische Fakultöten, Universitat Bielefeld, Bielefeld, Germany, 2002.
[29] Entrez genome (bacterial genomes). [Online]. Available: http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html.
[30] Entrez genome (viral genomes). [Online]. Available: http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html.
[31] *Legionella* genome project. [Online]. Available: http://genome3.cpmc.columbia.edu/~legion/leg_info.html.
[32] Centers for Disease Control. Q Fever. [Online]. Available: http://www.cdc.gov/ncidod/dvrd/qfever/.
[33] W. Ludwig and K.-H. Schleifer. Phylogeny of bacteria beyond the 16S rRNA standard: Despite expanding data sets and alternative markers for inferring phylogenetic relationships, nothing beats the 16S rRNA gene. [Online]. Available: http://www.vermicon.de/english/news/Science/khs99111.htm.
[34] Cepheid SmartCycler. [Online]. Available: http://www.smartcycler.com.
[35] Idaho Technology Inc. R.A.P.I.D. [Online]. Available: http://www.idahotech.com.
[36] Plum Island Animal Disease Center.. U.S. Department of Agriculture. [Online]. Available: http://www.ars.usda.gov/plum/.
[37] United States Animal Health Association and National Animal Health Emergency Management System. Foot-and-mouth disease outbreak in Great Britain. [Online]. Available: http://www.usaha.org/issues/fmd2001.html.
[38] S. M. Reid, N. P. Ferris, G. H. Hutchings, A. R. Samuel, and N. J. Knowles, "Primary diagnosis of foot-and-mouth disease by reverse transcription polymerase chain reaction," *J. Virological Methods*, vol. 89, no. 1–2, pp. 167–176, Sept. 2000.
[39] P. M. McCready *et al.*, manuscript in preparation.
[40] P. Kasper, "Primers and probes for the detection of HIV," US Patent 5 985 544, Nov. 16, 1999.
[41] S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, F. W. Studier, and S. Swaminathan, "Structural genomics: Beyond the human genome project," *Nature Genetics*, vol. 23, no. 2, pp. 151–157, Oct. 1, 1999.
[42] R. Sánchez and A. Sali, "Comparative protein structure modeling in genomics," *J. Comput. Phys.*, vol. 151, pp. 388–401, May 1999.



J. Patrick Fitch (Senior Member, IEEE) received B.S. degrees in physics and in engineering science from Loyola College, Baltimore, MD, in 1981 and the Ph.D. degree in electrical engineering from Purdue University, W. Lafayette, IN, in 1984.

Prior to life science applications, he was the principal investigator for a variety of imaging and computing projects applied to astronomy, nondestructive evaluation, and national security. He is currently Chemical and Biological National Security Program Leader at the Lawrence Livermore National Laboratory (LLNL) in Livermore, CA. In the past decade, his LLNL responsibilities have included genomics, bioengineering, and engineering research. He has led teams with more than 250 scientific and technical staff members. He has also successfully developed and marketed a medical device business strategy to venture investors. He is the author of *An Engineering Introduction to Biotechnology* (Bellingham, WA: SPIE Press, 2002) and *Synthetic Aperture Radar* (New York: Springer-Verlag, 1988). He is an Editorial Board Member of *Biomolecular Engineering*, Elsevier. His research interests include bioinformatics, bioinstrumentation, mechanisms of pathogenicity, and medical devices.

Dr. Fitch is a Fellow of the American Society for Laser Medicine and Surgery, and a member of the SPIE. He received an IEEE best paper award in 1988 and national FLC awards for medical devices in both 1998 and 1999.

Shea N. Gardner received the B.A. degree from Princeton University, Princeton, NJ, in 1991 in ecology, evolution, and behavior, graduating with highest honors, and the Ph.D. degree in population biology from the University of California, Davis, in 1997. She did postdoctoral research in applied mathematical ecology at the National Environment Research Council Centre for Population Biology, Imperial College, London, England.

She was a biology professor at the 1997 Research Science Institute at MIT, Boston, MA, and the 1999 Pacific Institute for the Mathematical Sciences in Vancouver, BC, Canada. She is currently a Lawrence Postdoctoral Fellow at the Lawrence Livermore National Laboratory, Livermore, CA. She is part of the bioinformatics team for the Chemical and Biological National Security Program developing signatures for pathogen detection. She has also developed computational models of multidrug cancer chemotherapy to tailor treatments for individuals, for which she holds a provisional patent. Her research interests include bioinformatics and mathematical modeling in evolution, medicine, and population dynamics.

Dr. Gardner is a member of the American Association for Cancer Research, the International Society for Computational Biology, Who's Who International, and Phi Sigma and Sigma Xi Honor Societies.

Thomas A. Kuczmarski received the B.S. degree in engineering from Case-Western Reserve University, Cleveland, OH, in 1967 and the M.S. degree in computer science from the University of Wisconsin, Madison, WI, in 1969.

He is currently a computer scientist at the Lawrence Livermore National Laboratory, Livermore, CA. His responsibilities include the design and implementation of computational techniques for the development of DNA signatures, which are used for the rapid detection of microbial and viral pathogens. During his career he has engaged in a wide variety of computer science activities, including operating system internals, compiler implementation, graphical user interfaces, and system administration of complex systems. In addition to applying his computer skills to the biological sciences, he has also designed systems for the computational and graphical analysis of atmospheric data. His research interests include effective man-machine interactions.

Stefan Kurtz received the M.S. degree in computer science from the University of Dortmund, Dortmund, Germany, in 1990 and the Ph.D. degree in computer science from the University of Bielefeld, Bielefeld, Germany, in 1995.

From 1995 to 2002 he was an Assistant Professor in the Department of Computer Science of the University of Bielefeld. In 1996 and 1997 he was a postdoc at the University of Arizona, Tucson. He is currently an Associate Professor for Bioinformatics at the University of Hamburg, Hamburg, Germany. His current research interest is focused on bioinformatics, in particular on index structures for large biosequence databases, and efficient algorithms for matching complex patterns in biosequences.

Rich Myers, photograph and biography not available at time of publication.

Linda L. Ott received the B.S. degree in chemistry from Indiana University, Bloomington, IN.

She is a Computer Scientist in the Chemical and Biological National Security Program at the Lawrence Livermore National Laboratory (LLNL), Livermore, CA. She has worked on many different projects in her career at LLNL and was recently part of a team that designed and deployed a "keyboard-less" biological field lab tracking system.