Identification of Genomic Signatures for the Design of Assays for the Detection and Monitoring of Anthrax Threats

S. Draghici, P. Khatri, Y. Liu, K.J. Chase, E.A. Bode, D.A. Kulesh, L.P. Wasieloski, D.A. Norwood, and J. Reifman

Pacific Symposium on Biocomputing 10:248-259(2005)

IDENTIFICATION OF GENOMIC SIGNATURES FOR THE DESIGN OF ASSAYS FOR THE DETECTION AND MONITORING OF ANTHRAX THREATS

SORIN DRAGHICI[1,†], PURVESH KHATRI[1,2,†], YANHONG LIU[4], KITTY J CHASE[3], ELIZABETH A BODE[3], DAVID A KULESH[3], LEONARD P WASIELOSKI[3], DAVID A NORWOOD[3], JAQUES REIFMAN[2]

[1] *Dept. of Computer Science, Wayne State University, Detroit, MI 48202*
[2] *Bioinformatics Cell, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Ft. Detrick, MD 21701*
[3] *Diagnostic Systems Division, US Army Medical Research Institute of Infectitious Diseases, Ft. Detrick, MD 21701*
[4] *US Dept. of Agriculture, Agricultural Research Service, Eastern Regional Research Center, Wyndmoor, PA 19038*

Sequences that are present in a given species or strain while absent from or different in any other organisms can be used to distinguish the target organism from other related or un-related species. Such DNA signatures are particularly important for the identification of genetic source of drug resistance of a strain or for the detection of organisms that can be used as biological agents in warfare or terrorism. Most approaches used to find DNA signatures are laboratory based, require a great deal of effort and can only distinguish between two organisms at a time. We propose a more efficient and cost-effective bioinformatics approach that allows identification of genomic fingerprints for a target organism. We validated our approach using a custom microarray, using sequences identified as DNA fingerprints of *Bacillus anthracis*. Hybridization results showed that the sequences found using our algorithm were truly unique to *B. anthracis* and were able to distinguish *B. anthracis* from its close relatives *B. cereus* and *B. thuringiensis*.

1. Introduction

The area of organism identification using DNA sequences has many applications in various life science areas. However, there are also many challenges. For instance, sheep pox and goat pox viruses are so closely related that they cannot be distinguished using clinical signs, pathogenesis or sero-reactivity.[30] Furthermore, both cross-infectivity and cross-resistance have

† These authors should be considered joint first authors.

been reported[38] to the point that the two were thought to be caused by a single viral species. However, genetic analysis demonstrated that sheep pox and goat pox are actually caused by two related, but genetically distinct viruses. Furthermore, the identification of a few base pair differences in the sequence coding for the P32 protein allowed the design of a polymerase chain reaction (PCR) restriction fragment length polymorphism (PCR RFLP) assay able to distinguish between the two species. This assay involves a PCR amplification with a common primer, followed by a digestion with a *Hinf I* restriction enzyme that produces fragments of different sizes allowing the identification of the two species.

The issue of distinguishing between different species is somewhat academic if the two species exhibit both cross-infectivity and, most importantly, allow passive cross-protection as the sheep pox and goat pox do.[37] However, this is not always the case. Genes that are present in certain isolates of a given bacterial species and are substantially different or absent from others can determine important strain-specific traits such as drug resistance[13] and virulence.[51] As an example, *B. anthracis*, *B. cereus*, and *B. thuringiensis* are genetically so close that it has been proposed to consider them a single species.[27] At the same time, these bacteria are very different on a phenotypic level. *B. cereus* is a frequent food contaminant but only a mild opportunistic human pathogen;[16,28] *B. thuringiensis* is actually a useful bacterium being used as a pesticide[46] while *B. anthracis* is a virulent pathogen for mammals that has been used as a bio-terror and biological warfare agent.[12,53]

In such cases, the identification of an organism-specific DNA sequence gains an increased importance. Even if such sequences are not functionally active, they can still be extremely useful if used as genetic fingerprints. DNA sequences that are present in a given species while absent from any other organisms can be used to distinguish the target organism from other related or un-related species. If such genetic fingerprints were available for organisms that can be potentially used as biological or terrorist weapons, the task of rapid threat identification, characterization, and selection of appropriate medical countermeasures could be immensely facilitated. Genetic fingerprints can also aid identification of genetic source of drug resistance of a strain,[17] which can be useful to drug developers in pharmacogenomics.

2. Existing work

The existing work in the areas of organism identification using DNA signatures can be divided into two different categories. One approach uses a laboratory assay to identify the organism. Techniques used include amplified fragment length polymorphism (AFLP),[44,45] suppression subtractive hybridization (SSH)[3] and custom DNA microarrays.[36] A second approach uses a purely bioinformatics analysis of the characteristics of the genomes of various species and extracts those features that are characteristic to individual species.

The laboratory based approach does not necessarily require information about the entire genomes involved and is better suited for the development of assays for monitoring and identification of biological threats. For instance, SSH, a PCR-based DNA subtraction method, allows identification of genomic sequence differences in a "tester" DNA relative to a "driver" DNA. AFLP relies on the analysis of a fluorescence based signal proportional to the size of various DNA fragments.[49] SSH and AFLP have been successfully used to identify genomic sequence differences between various strains or species of bacteria.[4,5,10,31,44] The major drawback of this approach is that it permits identification of genomic differences only between two organisms. For instance, in order to differentiate two species, one needs to use an SSH assay to compare each strain of one species with each strain of the other species.[44] Clearly, this approach cannot be used to provide a genomic signature that would differentiate a given organism from all others.

The *in silico* approach to identifying genomic signatures is usually based on an analysis of the entire genomes involved and aims at extracting features such as species-specific codon usage.[1,2,23,32–34,52] While this type of genomic signature can be informative about the given organisms and the relationships among them, it may not be directly usable for detection and monitoring purposes.

Comparative sequence analysis has also been useful in detecting intronic and intergenic regions[25,40] as well as uncovering novel repeated structures.[18,26] Several genome scale alignment tools are available: MUMmer,[14,15,39] AVID,[11] MGA,[29] WABA,[35] and GLASS[7] among others. TaxPlot[22] provides visual representation of protein homologs in microbial and eukaryotic genomes. Most of these pair-wise[a] alignment tools assume that the input genomes are closely related. Therefore, there will be a mapping

[a]MGA is a multiple alignment tool but the alignment is still computed pair-wise.

of large subsequences between the two input genomes. In turn, they assume that these large subsequences, appearing in the same order in the closely related genomes, are very likely to be part of the final alignment. These regions are used as anchors for the alignment of the input genomes.

In general, anchor-based genome alignment programs first create a suffix tree from the two input genomes. A suffix tree is a compact representation of all suffixes in the input string.[41,54] A suffix of a string is a substring starting at any position in the string and extending up to the end of the string. Next, the suffix tree is searched for sequences that appear in both input genomes. These exact matching subsequences are known as maximal exact matches (MEMs). The anchors are chosen from these MEMs. Different programs apply different criteria for the selection of anchors. For instance, MUMmer uses the longest increasing subsequence (LIS)[24] for the selection of anchors.[14] MUMmer allows the selection of overlapping anchors whereas AVID and MGA only select non-overlapping anchors. Since MGA allows alignment of more than two genomes, it only selects MEMs that are present in all of the input genomes. AVID first finds the length of the longest MEM and discards all the MEMs that are less than half the length of the longest MEM. After selecting the anchors, MUMmer employs a variant of the Smith-Waterman algorithm[47] to close the gaps between the anchors. MGA and AVID close the gaps by recursively creating suffix trees for the non-anchored parts of the input genomes and hence, gradually reducing the gap sizes. Once the gaps are smaller than a threshold, MGA and AVID close them using the ClustalW[48] and Needleman-Wunsch algorithms,[42] respectively.

These large number of tools are all geared towards finding large-scale *similarities* between two or more genomes. Our focus here is different. While these algorithms were developed to find sequence similarities, our goal is to find sequence dissimilarities. These two problems are related but not reciprocal. Simply put, one cannot just take the complement of the sequences found in a similarity search and use them as genomic signatures. The main reason is related to the fact that a search aiming to find similarity will sometimes discard entire blocks after only a summary inspection because they are not sufficiently similar to the target sequence. On the other hand, a search aiming to find dissimilarities, i.e., unique signatures, has to actually focus on exactly those areas that are discarded without extensive analysis during the similarity search.

Here, we propose an algorithm for finding genomic fingerprints that distinguish an organism from all other organisms with known genomes.

As the number of sequenced organisms increases, this approach has the potential to substitute existing laboratory based approaches such as AFLP and SSH.

In this paper, we used this approach to find a genetic signature for *B. anthracis*. Identification of genomic regions unique to *B. anthracis* can provide clues to its genetic relationship to other highly similar organisms. Related work for the detection of *B. anthracis* used plasmid-encoded toxin genes for rapid DNA-based assays.[8] However, these failed to detect non-plasmid containing strains of *B. anthracis* isolated from the environment.[50] Also, there have been efforts to design real-time PCR assays. However, these assays only targeted a single locus and they yielded false-positive results with some strains of *B. cereus*.[20,43]

3. Analysis methods

Our goal is to find unique DNA sub-sequences for a given target genome across all available known genomes. An obvious approach is to compare (i.e., align) the genome of our interest against all available known genomes. These alignments will reveal the parts of the target genome that do not align with any other genome (i.e., are unique to the target genome). However, this seemingly simple approach is computationally very expensive. The GenBank database at NCBI contains nucleotide sequences from more than 140,000 organisms.[9] The length of these genomes vary from a few thousand base pairs to a few billion base pairs. Aligning the input genome with each of these genomes is computationally unfeasible.

The amount of computation can be considerably reduced by using the phylogenetic background of the target. Today biologists agree that various organisms have evolved from common ancestors. During evolution, functional genomic elements are conserved. Hence, two closely related genomes are expected to have many matching subsequences. If a subsequence that distinguishes the target from all organisms exists, this subsequence will also distinguish the target from its closest relative. Hence, a good initial set of potential genomic signatures can be obtained by comparing the target only with its closest relative and by retaining only those sequences that are different. Subsequently, each of these potential signatures is compared with all other known genomes. This approach drastically reduces both the number of comparisons required as well as the length of sequences to be compared (from a few million to a few thousand base pairs, at most).

In order to find the exact matching sequences between the target and its

closest relative, we start by using their concatenated sequences to create a suffix tree. We then use a suffix tree search algorithm as the one employed in MUMmer to find the exact matching sequences in both genomes. Since our goal is to determine a set of relatively short sequences to be used on a microarray type assay, we have to search both the forward and the reverse strands. Any sequences that match between the two organisms are removed from further consideration. The result is a set of short segments of the target genome that can be considered potential signatures. These are then compared with all sequences in the blast-nt[21] database from NCBI.[6] We consider a sequence is unique for the target genome if it does not align to any sequence from any other organism with an expected value (E-value) less than a threshold of 0.01. Fig. 1 provides an overview of this approach.

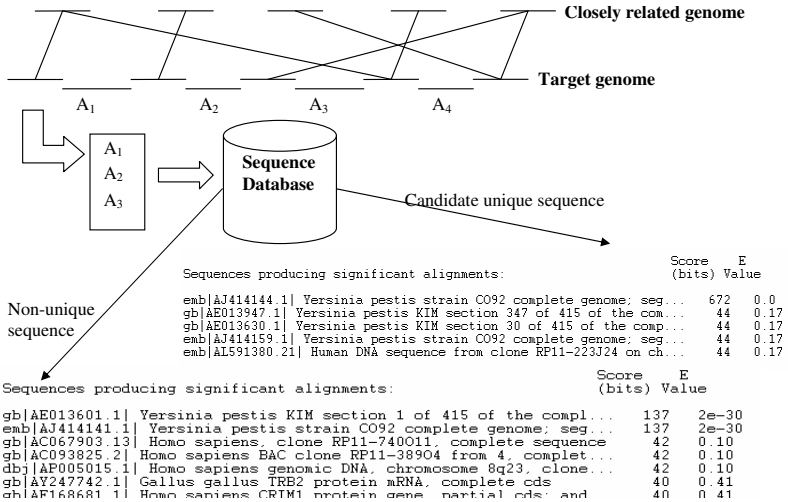

Figure 1. The genomic fingerprinting approach. Two genomes are searched for exact matching subsequences (MEMs). The MEMs are removed from the target genome and the remaining segments of the target genome (A_1, A_2, ..., A_n) are searched against the nt database. If the length of a segment is less than the user specified length, it is discarded and not searched in the nt database. As shown, if a sequence does not align with any sequence from another organism with E value less than the specified threshold it is considered as a sequence unique to the target genome.

4. Results and discussion

In order to validate our approach, we designed a custom microarray using sequences identified as genomic fingerprints for *B. anthracis*. This array

was then hybridized with *B. anthracis* and *B. cereus*.

In order to find a genomic signature for *B. anthracis* we proceeded as follows. We searched the *B. anthracis str. Ames* genome (GenBank contig accession number NC_03997) for subsequences of 30 base pairs or more matching anywhere (direct and reverse strand) with sequences from the genome of *B. cereus ATCC 14579* (GenBank contig accession number NC_004722). We chose *B. cereus ATCC 14579* genome as a closely related genome because it is considered to be a good representative of the *B. cereus* family.[19] Then, we removed all of matching sequences from the *B. anthracis* genome. This step produced over 6,000 sequences of length 50 or more. These sequences were then searched against the *nt* database using *blastn*. The sequences in the BLAST output that were not found in any other organism with E value less than 0.01 were retrieved and considered part of the genomic fingerprints of *B. anthracis*. There were 140 such sequences. Note that this analysis stage also removed sequences that matched the genomes of other close relatives of *B. anthracis*, such as *B. thuringiensis*, without ever directly comparing them. These 140 target sequences were provided to CombiMatrix (Mukilteo, WA) for the design of a custom microarray. CombiMatrix designed 2 probes for 80 target sequences and 1 probe for 22 target sequences (for a total of 182 probes for 102 target sequences) with melting temperature in the range of 70°C to 75°C and a length of 35 base pairs or more. Probes of the required length and melting temperatures could not be identified for the remaining 38 target sequences. The microarray was designed with three replicates of each of the 182 probes.

The custom microarray was then hybridized with samples of *B. anthracis* and *B. cereus*. The hybridization results showed that 18 probes only hybridized to the *B. anthracis* sequences indicating that they were true genomic fingerprints of *B. anthracis*. Table 1 provides the positions of the sequences on *B. anthracis* genome that were found to be unique in the microarray experiment.

Surprisingly, many of the initial 182 probes also hybridized with *B. cereus*. We further searched these cross-hybridizing probes against the blast-*nt* database. For the probes that hybridized to *B. cereus* the results of this comparison showed that although the target sequences of those probes are only present in *B. anthracis*, the part of the target sequence on which the probes were designed was not unique to *B. anthracis* and is present in other genomes. This shows that the probe design stage lost some specificity due to its unique added requirements: melting temperatures in a very narrow range, limited lengths, etc. In all cases, although the initial,

longer sequence was unique across the blast-*nt* database, by selecting a shorter subsequence, the probe became unspecific. Hence, another BLAST search is recommended before printing the assay, to check whether the subsequences selected as probes continue to be good signatures for the target organism.

Table 1. The following 18 probes identify 17 unique sequences of *B. anthracis (Ames)*. The first and second columns indicate the start and end, respectively, of the target sequences from *B. anthracis*. The third and the fourth column are the start and end positions, respectively, on the corresponding target sequences for which probes were designed.

Sequence start	Sequence end	Probe start	Probe end
175,231	175,455	6	44
175,567	175,677	36	71
488,976	489,620	130	166
945,569	946,596	151	190
1,629,522	1,630,538	489	523
1,629,522	1,630,538	529	568
1,845,001	1,845,363	111	145
2,021,535	2,022,919	491	529
2,098,619	2,099,274	591	625
2,783,190	2,783,405	17	54
2,918,788	2,920,251	977	1013
3,037,856	3,038,113	115	152
3,524,649	3,524,731	17	55
3,808,069	3,809,046	797	834
3,821,617	3,822,163	449	483
4,374,364	4,375,478	227	311
4,375,581	4,376,123	149	186
4,933,405	4,933,482	9	43

5. Conclusion

DNA sequences that are present in a given species or strain while absent from any other organism can be used to distinguish the target organism from other related or un-related species. The identification of such DNA signatures is particularly important for organisms that may be potentially used as biological warfare agents or terrorism threats.

Most approaches used to identify DNA signatures are laboratory based and require a significant effort and time. A bioinformatics approach can provide results faster and more efficiently. However, most tools built for

genome comparisons only allow alignment of two genomes at a time. Using this approach to find unique DNA signatures across all known organisms is unfeasible. In addition, all existing tools are limited to finding the similarity between two genomes. In contrast, looking for DNA signatures requires the development of tools that identify sequence dissimilarities. In this paper, we describe an approach to find the DNA fingerprints of an organism. We used this approach to find a set of unique sequences for *B. anthracis* which were then used to design probes for a DNA microarray. The hybridization results revealed that a subset of these probes were truly unique to *B. anthracis* and were able to distinguish between *B. anthracis* and *B. cereus*, which is a close genetic relative.

Acknowledgements

This work was supported by the research area directorates of the US Army Medical Research and Materiel Command and the Defense Threat Reduction Agency. The first two authors are also supported by: NSF DBI-0234806, NIH 1S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1R21 CA10074001, 1R21 EB00990-01 and 1R01 NS045207-01.

References

1. T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura. A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: Self-organizing map of oligonucleotide frequency. *Genome Informatics*, (13):12–20, 2002.
2. T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura. Informatics for unveiling hidden genome signatures. *Genome Research*, 13(4):693–702, 2003.
3. P. G. Agron, M. Macht, L. Radnedge, E. W. Skowronski, W. Miller, and G. L. Andersen. Use of subtractive hybridization for comprehensive surveys of prokaryotic genome differences. *FEMS Microbiology Letters*, 211(2):175–182, Jun 2002.
4. I. Ahmed, G. Manning, T. Wassenaar, S. Cawthraw, and D. Newell. Identification of genetic differences between two *Campylobacter jejuni* strains with different colonization potentials. *Microbiology*, 148:1203–1212, 2002.
5. N. Akopyants, A. Fradkov, L. Diatchenko, and et. al. PCR-based subtractive hybridization and differences in gene content among strains of *Helicobacter pylori*. *Proc. Natl. Acad. Sci.*, 95:13108–13113, 1998.
6. S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, M. W., and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sept 1997.
7. S. Batzoglu, L. Pachter, J. Mesirov, B. Berger, and E. Lander. Human and

mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 10:950–958, 2000.
8. C. Bell, J. Uhl, T. Hadfield, J. David, R. Meyer, T. Smith, and F. Cockerill III. Detection of *Bacillus anthracis* dna by light-cycle pcr. *J. Clin. Microbiol.*, 40:2897–2902, 2002.
9. D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostel, and D. Wheeler. GenBank: update. *Nucleic Acids Research*, 32(1):D23–D26, January 2004.
10. M. Bogush, T. Velikodvorskaya, Y. Lebedev, and et. al. Identification and localization of differences between *Escherichia coli* and *Salmonella typhimurium* genomes by suppressive sutractive hybridization. *Mol Gen Genet*, 262:721–729, 1999.
11. N. Bray, I. Dubchak, and L. Pachter. AVID: A global alignment program. *Genome Research*, 13(1):97–102, January 2003.
12. E. Check. Bioshield defence programme set to fund anthrax vaccine. *Nature*, 429(6987):4, May 2004.
13. J. Davies. Inactivation of antibiotics and the dissemination of resistance genes. *Science*, 264(5157):375–382, Apr 1994.
14. A. Delcher, S. Kasif, R. Fleischmann, J. Peterson, O. White, and L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–237, 1999.
15. A. Delcher, A. Phillippy, J. Carlton, and S. Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11):2478–2483, 2002.
16. F. Drobniewski. *Bacillus cereus* and related species. *Clin. Microbiol. Rev.*, 6:324–338, 1993.
17. S. Drăghici and B. Potter. Predicting HIV drug resistance with neural networks. *Bioinformatics*, 19(1):98–107, January 2003.
18. I. Dunham, N. Shimuzu, B. Roe, S. Chissoe, and et. al. The DNA sequences of chromosome 22. *Nature*, 402:489–495, 1999.
19. K. Dwyer, J. Lamonica, J. Schumacher, L. Williams, J. Bishara, A. Lewandowski, R. Redkar, G. Patra, and D. V.G. Identification of bacillus anthracis specific chromosomal sequences by suppressive subtractive hybridization. *BMC Genomics*, 5(1):15, Feb 2004.
20. H. Ellerbrok, H. Nattermann, M. Ozel, L. Beutin, B. Appel, and G. Pauli. Rapid and sensitive identification of pathogenic and apathogenic *Bacillus anthracis* by real-time pcr. *FEMS Microbiol. Lett.*, 214:51–59, 2002.
21. N. C. for Biotechnology Information. Blast nucleotide database. ftp://ftp.ncbi.nih.gov/blast/db/.
22. N. C. for Biotechnology Information. Taxplot. http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi?
23. R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8:r49–r62, 1980.
24. D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, 1997.
25. R. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal regulatory elements: a reason to sequence the mouse genome. *Genome research*, 7:959–966, 1997.

26. M. Hattori, A. Fujiyama, T. Taylor, H. Watanabe, and et. al. The DNA sequence of human chromosome 21. *Nature*, 405:311–319, 2000.
27. E. Helgason, D. Caugant, I. Olsen, and A. Kolsto. Genetic structure of population of *Bacillus cereus* and *B. thuringiensis* isolates associated with periodontitis and other human infections. *J. Clin. Microbiol.*, 38:1615–1622, 2000.
28. O. I. K. A. Helgason E, Caugant DA. Genetic structure of population of bacillus cereus and b-thuringiensis isolates associated with periodontitis and other human infections. *Journal Of Clinical Microbiology*, 38(4):1615–1622, Apr 2000.
29. M. Hohl, S. Kurtz, and E. Ohlebusch. Efficient multiple genome alignment. *Bioinformatics*, 18(Suppl. 1):S312–S320, 2002.
30. M. Hosamani, B. Mondal, P. A. Tembhurne, S. K. Bandyopadhyay, R. K. Singh, and T. J. Rasool. Differentiation of sheep pox and goat poxviruses by sequence analysis and pcr-rflp of p32 gene. *Virus Genes*, 29(1):73–80, Aug 2004.
31. B. Janke, U. Dobrindt, J. Hacker, and G. Blum-Oehler. A subtractive hybridization analysis of genomic differences between the uropathogenic *E. coli* strain 536 and the *E. coli* k-12 strain mg1655. *FEMS Microbial Lett.*, 199:61–66, 2001.
32. S. Kanaya, M. Kinouchi, T. Abe, Y. kudo, Y. Yamada, T. Nishi, H. Mori, and T. Ikemura. Analysis of codon usage diversity of bacterial genes with a self-organizing map (som): Characterization of horizontally transferred genes with emphasis on the e. coli o157 genome. *Gene*, 276:89–99, 2001.
33. S. Kanaya, Y. Kudo, T. Abe, T. Okazaki, D. Carlos, and T. Ikemura. Gene classification by self-organizing mapping of codon usage in bacteria with completely sequences genome. *Genome Informatics*, 9:369–371, 1998.
34. S. Kanaya, Y. Kudo, Y. Nakamura, and T. Ikemura. Detection of genes in escherichia coli sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage. *CABIOS*, 12:213–225, 1996.
35. W. Kent and A. Zahler. Conservation, regulation, synteny and introns in large-scale *C. briggsae-C. elegans* genomic alignment. *Genome research*, 10:1115–1125, 2000.
36. M. Kingsley, T. Straub, D. Call, D. Daly, S. Wunschel, and D. Chandler. Fingerprinting closely related xanthomonas pathovars with random nonamer oligonucleotide microarrays. *Appl. Environ. Microbiol.*, 68:6361–6370, 2002.
37. R. P. Kitching. Passive protection of sheep against capripoxvirus. *Res Vet Sci*, 41(2):247–250, Sep 1986.
38. R. P. Kitching and W. P. Taylor. Clinical and antigenic relationship between isolates of sheep and goat pox viruses. *Trop Anim Health Prod*, 17(2):64–74, May 1985.
39. S. Kurtz, A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5:R12, 2004.
40. G. Loots, R. Locksley, C. Blankespoor, Z. Wang, W. Miller, E. Rubin, and

K. Frazer. Identification of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons. *Science*, 288:136–140, 2000.
41. E. McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2):262–272, 1976.
42. S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, 1970.
43. Y. Qi, G. Patra, X. Liang, L. Williams, S. Rose, R. Redkar, and V. DelVecchio. Utilization of the *rpoB* gene as a specific chromosomal marker for real-time pcr detection of *Bacillus anthracis*. *Appl. Environ. Microbiol.*, 67:3720–3727, 2001.
44. L. Radnedge, P. G. Agron, K. Hill, P. Jackson, L. Ticknor, P. Keim, and A. G. L. Genome differences that distinguish bacillus anthracis from bacillus cereus and bacillus thuringiensis. *Applied And Environmental Microbiology*, 69(5):2755–2764, May 2003.
45. L. Radnedge, S. Gamez-Chin, P. McCready, P. Worsham, and G. Andersen. Identification of nucleotide sequences for the specific and rapid detection of yersinia pestis. *Applied And Environmental Microbiology*, 67(8):3759–3762, Aug 2001.
46. E. Schnepf, N. Crickmore, J. Van Rie, D. Lereclus, J. Baum, J. Feitelson, D. R. Zeigler, and D. H. Dean. Bacillus thuringiensis and its pesticidal crystal proteins. *Microbiology And Molecular Biology Reviews*, 62(3):775–806, Sep 1998.
47. T. Smith and M. Waterman. Identification of common molecular subsequences. *J. Molecular Biology*, 147(1):195–197, 1981.
48. J. Thompson, D. Higgins, and T. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
49. L. Ticknor, A. Kolsto, K. Hill, P. Keim, M. Laker, M. Tonks, and P. Jackson. Fluorescent amplified fragment length polymorphism analysis of norwegian bacillus cereus and bacillus thuringiensis soil isolates. *Appl Environ Microbiol.*, 67(10):4863–4873, 2001.
50. P. Turnbull, R. Hutson, M. Ward, M. Jones, C. Quinn, N. Finnie, C. Duggleby, J. Kramer, and J. Melling. *Bacillus anthracis* but not always anthrax. *J. Appl. Bacteriol.*, 72:21–28, 1992.
51. M. K. Waldor and M. J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, 272(5270):1910–1914, Jun 1996.
52. H. Wang, J. Badger, P. Kearney, and M. Li. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Molecular Biology and Evolution*, 18:792–800, 2001.
53. G. Webb. A silent bomb: the risk of anthrax as a weapon of mass destruction. *Proceedings of the National Academy of Sciences USA*, 100(7):4346–4351, 2003.
54. P. Weiner. Linear pattern matching algorithms. In *Proc. 14th IEEE Symp. Switching & Automata Theory*, pages 1–11, 1973.