

DNA Microbial and Viral Identification using Ultraspecific Probes “Blind” to Host and Background DNA

Catherine Putonti
Dept. of Computer Science
University of Houston
Houston, TX 77204

George Fox
Dept. of Biology and Biochemistry
Dept. of Chemical Engineering
University of Houston
Houston, TX 77204

Richard C. Willson
Dept. of Biology and Biochemistry
Dept. of Chemical Engineering
University of Houston
Houston, TX 77204

B. Montgomery Pettitt
Dept. of Computer Science
Dept. of Biology and Biochemistry
Dept. of Chemistry
University of Houston
Houston, TX 77204

Yuriy Fofanov
Dept. of Computer Science
University of Houston
Houston, TX 77204

ABSTRACT

The reliable detection and identification of microbes and viruses in complex samples without separation of DNA of the organism of interest from the sample background is a challenging and important problem. We have developed a set of novel algorithms that make it feasible to analyze the occurrence of all possible short sequences of length 10 to 25 nucleotides in complete genome sequences of any size. As a result, we can identify all unique sequences present in each of a large set of pathogen genomes and absent in (and not within up to three mismatches) the human genome. We found that it is unusual to find a single, unique genomic sequence present simultaneously in all genomes of interest and absent in all other genomes, including the host organism, even for groups of closely related organisms (e.g., the West Nile virus). This result leads us to suggest using a set of probes that are absent in the host genome, likely to be found in the pathogen genome, and expressed in a unique pattern for each pathogen for pathogen identification. Herein we use an evolutionary programming approach to design microarrays so as to minimize the number of probes required, to avoid false positives and to achieve maximal sensitivity. Supporting the proposed approach, initial *in silico* and *in vitro* microarray experimental results are provided.

INTRODUCTION

Despite growing concerns about the use of biological agents in acts of terrorism, it remains a difficult task to identify pathogens, e.g. viruses, particularly when they are present in a complex sample. Specific tests for identification are increasingly based on the use of nucleic acid technologies [1] including polymerase chain reaction (PCR) [2] and microarrays of cDNA [3] and oligonucleotides [4]. Identification is possible by these technologies through the use of unique “genetic signatures” as probes/primers for each target.

These technologies typically necessitate processing of the complex sample prior to testing, further complicating the task at hand. In the more traditional method, the DNA of interest is extracted from the complex sample (natural waters, rumen contents, food, blood, tissue, etc.) for analysis [5-8]. In the case of viruses, extraction is necessary because one cannot cultivate viruses *in vitro*, however the viral DNA/RNA purification is a difficult and expensive process. As an alternative, PCR primers can be designed such that they recognize sequences that are uniquely present in the organism of interest and absent in other genomes [9-11]. Because the concentration of the background DNA is usually orders of magnitude higher than that of the pathogen, PCR will exponentially amplify and effectively “purify” DNA of the organism of interest but not other DNA that may be present. It is also clear that only nucleic acid sequences present in the organism of interest and absent in the background DNA can be used as potential specific identification signatures. This implies either extensive experimental effort (the customary approach) or the extremely difficult computational task of finding such organism-unique and “background-blind” subsequences in the genome of each organism to be detected.

Several problems stem from the inherent principals of these technologies. First, to design such unique sequences one must know the pathogen’s genomic sequence. Secondly, some viruses, like flu, dengue virus and West Nile, show very high mutation rates; as a result it is a serious concern that any “unique” sequence designed based on an already known genomic sequence will not work for the modified pathogen. Third, the host genome may also be susceptible to mutations, e.g. SNPs within the human genome. Thus, it is preferable for the unique sequence to not be close (within relatively few mismatches) to any host sequences, reducing the probability of a false identification.

Literature debates whether “unique” sequences can be used to identify viruses at all because: (1) while there are hundreds of strains of a particular species, relatively few

have been sequenced; (2) strains may vary in pathogenicity and the virulence of their near-neighbors, or lack thereof, may affect different hosts differently; and (3) because of the high mutation rate of viruses, it just might be that there is no unique sequence shared by all strains of a species [12]. Thus, a more reliable approach for identification is needed. Making use of microarray technology, we propose that identification/detection can be achieved by *unique patterns* of probe hybridization with the target pathogen's genomic sequence. Using known genomic sequences, a set of "host-blind" sequences can be selected for probe design such that when a particular strain in a complex sample is tested, a unique pattern appears. (We refer to the probes used for identification as being ultraspecific probes.) In the case that a new strain is present in the tested sample, a different pattern will appear. By comparing this pattern to the patterns generated by known strains, deductions can be made regarding the evolution and potential relatives of this new genomic sequence.

As a demonstration of our proposed method, we chose to develop a set of ultraspecific probes for the identification of the West Nile virus (WNV) in the presence of a human host sample. All sequences (n -mers) of size up to 22 nucleotides long are examined in order to select sequences that are present in the pathogen and absent in the human genome. Moreover, to exclude possible false positive calls due to variations in the human genome, we concentrate our attention only on n -mers that do not appear in the human genome with all possible changes of 1 or 2 nucleotides. We say that such n -mers are t -tolerant or tolerant of t mismatches while still not matching any human sequence. We found several hundred "human-blind" sequences including: sequences present in each individual viral genome, sequences present in all considered viruses, and sequences unique for each individual genome. An evolutionary programming approach was used to create an optimal set of probes for WNV detection.

MATERIAL AND METHODS

Data

For our calculations we use the human genome build 34 version 3 available from GenBank [13] (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606). In all of our calculations, we use both the original and complementary human genome sequences. This version contains 794,007 unknown/unidentified bases. For simplicity, all n -mers containing such characters were excluded from our analysis. Because the file structure of the available human genome assembly does not allow assembling each chromosome without gaps, we decided to exclude from our analysis all n -mers that simultaneously contain nucleotides from two files.

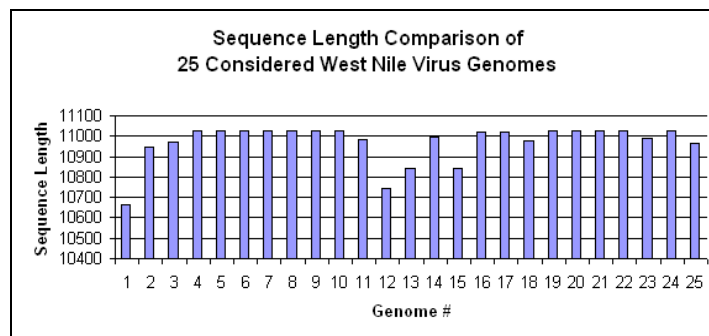


Figure 1. Comparison of the lengths of the 25 WNV sequences used in our calculations.

Table 1 lists the 25 complete sequences of the WNV as of May 25, 2004 (one more WNV genome was added on June 1, AY289214, but is not included in our analysis). WNV belongs to the family *Flaviviridae*. Like all flaviviruses, WNV is a single-stranded RNA virus with a positive polarity RNA genome of approximately 11kb [8]. Minor variation in the length of the considered WNV genomes is represented in Figure 1. All of the considered WNV sequences are available from GenBank (<http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/VIAL/11082/11082.html>).

1	Kunjin virus gene for polyprotein (D00246)
2	WNV strain HNY1999 polyprotein (AF202541)
3	WNV VLG-4 polyprotein precursor (AF317203)
4	WNV strain IS-98 STD (AF481864)
5	WNV isolate WN MD 2000-crow265 (AF404753)
6	WNV isolate WN NJ 2000 MQ5488 (AF404754)
7	WNV isolate WN NY 2000-grouse3282 (AF404755)
8	WNV isolate WN NY 2000-crow3356 (AF404756)
9	WNV isolate WN Italy 1998-equine (AF404757)
10	WNV polyprotein precursor (AF533540)
11	WNV isolate KN3829 polyprotein gene (AY262283)
12	WNV isolate LEIV-Krnd88-190 (AY277251)
13	WNV isolate LEIV-Vlg99-27889 (AY27752)
14	WNV isolate Ast99-901 (AY278441)
15	WNV isolate LEIV-Vlg00-27924 (AY278442)
16	Kunjin virus clone FLSDX polyprotein mRNA (AY274504)
17	Kunjin virus clone pAKUN polyprotein mRNA (AY274505)
18	WNV isolate 2741 (AF206518)
19	WNV strain NY99-eqhs (AF260967)
20	WNV strain Eg101 (AF260968)
21	WNV strain RO97-50 (AF260969)
22	WNV strain NY99-flamingo382-99 (AF196835)
23	WNV strain PaAn001 polyprotein (pol) gene (AY268132)
24	WNV strain Chin-01 (AY490240)
25	WNV (NC_001563)

Table 1. List of complete genomes of West Nile virus (WNV) strains and neighbors used for calculations taken from GenBank listing.

Calculations

For each of the 25 WNV genomes we identify all subsequences (n -mers) of size 16-22 present in each of the genomes in addition to those n -mers absent from the human genome and tolerant of any 1 or 2 mismatches while still not matching any human sequence. These calculations were made using novel algorithms developed in a previous work [13]. By including 1-tolerant and 2-tolerant n -mers, the likelihood of false positives due to single nucleotide variations (SNPs) in the human genome is greatly decreased. As seen in Table 2, it is only when $n \geq 16$ that the human genome does not include all possible 2-tolerant n -mers. For n -mers of size less than or equal to 14 nucleotides, sequences present in the human genome cover all possible combinations of 14 nucleotides with one mismatch. It is also necessary to kept in mind that the number of sequences considered to be “human-blind” must be large enough to have a reasonable probability for such sequences to be found in a viral genome.

<i>n</i> -mer size	Number of <i>n</i> -mers absent in human genome		
	0-tolerant	1-tolerant	2-tolerant
11	94	0	0
12	44,266	0	0
13	2,382,362	0	0
14	41,294,237	0	0
15	408,251,176	2,527,338	0
16	2,748,145,040	134,605,886	2,359
17	14,533,356,512	2,431,439,270	3,520,008

Table 2. Number of 11- to 17-mers present in the human genome with 0, 1, and 2 tolerances.

The number of *n*-mers present in each particular WNV genome was approximately the length of the particular genome minus *n* plus 1 as is expected if there is relatively few to no *n*-mers present more than once. Table 3 lists the *n*-mers, for *n*=16 to *n*=22, that are found in each of the 25 WNV genomes collectively (union) and the *n*-mers shared by all genomes (intersection). It is important to note that both the union and intersection are taken from the *n*-mers sets, which consider both the original and complementary strands. In the last column of Table 3, a trend appears; as *n* increases, the number of *n*-mers found in all genomes relative to the number found in these virus genomes collectively decreases.

<i>n</i> -mers size	Number of <i>n</i> -mers		
	Union (U)	Intersection (I)	I/ U %
16	92030	108	0.11735%
17	94510	92	0.09734%
18	96960	78	0.08045%
19	99206	68	0.06854%
20*	89982	60	0.06668%
21	92360	54	0.05847%
22	93762	50	0.05333%

Table 3. Number of 16- to 22-mers in the 25 WNV listed in Table 1.
*For 20-mers, WN_19 & WN_20 are excluded from this count due to file problems.

Next calculations were made for 0, 1, and 2-tolerant *n*-mers found in the WNV genomes (pathogen) and absent or “human-blind” to the human genome (host). Table 4 lists some of these results. While no 2-tolerant 19-mers are present in all 25 WNV, quite a few are present in the genomes. Figure 2 shows the distribution of the presence of these subsequences. It is striking to see that genome 12, WNV isolate LEIV-Krnd88-190 (AY277251), has nearly twice as many 19-mers than the other 24 genomes.

<i>n</i> -mer size	tolerance		
	0	1	2
16	41942	768	0
17	68622	5138	0
18	87186	18524	68
19	96128	43812	1252

<i>n</i> -mer	tolerance		
	0	1	2
16	42	0	0
17	58	0	0
18	64	10	0
19	64	22	0

Table 4. Number of 16- to 19-mers present in: (left) the 25 WNV genomes collectively (union) and (right) all 25 WNV genomes (intersection) for tolerances of 0, 1, or 2 mismatches to the human genome.

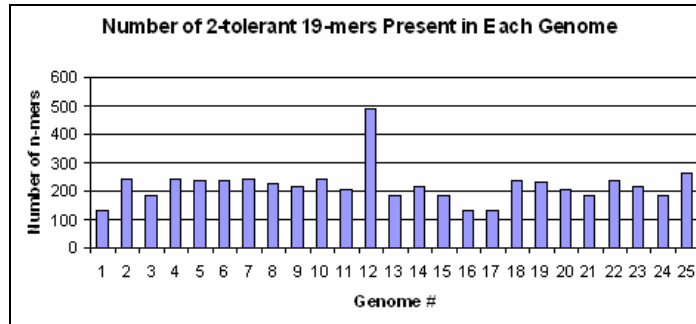


Figure 2. Comparison of the distribution of presence of 2-tolerant 19-mers in the 25 WNV genomes.

Computational Approach

An evolutionary programming approach was taken to design microarrays so as to minimize the number of probes required, to avoid false positives and to achieve maximum sensitivity. Based upon the computations performed, a set of sequences can be selected for a particular size and tolerance to be used as probes for such a microarray. It is our intent to define the minimum optimal set of subsequences, s_{min} , that can both identify the presence of a particular pathogen and distinguish between different strains of the pathogen. Consider a particular strain's genome, genome i , and the set s_{min} defined for the pathogen of which genome i is a strain. A subset of the subsequences included in s_{min} , $s_{min}(i)$, will be present in genome i . A pathogen strain not yet sequenced, genome x , will also have a subset of s_{min} 's sequences, $s_{min}(x)$.

To insure the sensitivity needed to properly identify a genomic sequence, s_{min} must meet the following criteria:

1. Each genome under consideration must be identifiable by at least α subsequences in s_{min} .
2. Each genome must be distinguishable from each other genome by at least γ subsequences.

If s_{min} meets these criteria, each sequenced pathogen strain will have a distinct $s_{min}(i)$. Moreover, genome x 's subset, $s_{min}(x)$ should not match the $s_{min}(i)$ of any genome i .

The value of α is dependent upon the size or number of sequences in s_{min} . Because viral genomes mutate rapidly and for a particular mutation there is roughly a 25% probability for each of the four nucleotides, the worst-case scenario is that 25% of the sequences in s_{min} could become present in a new strain in the future just at random. This however is not desirable since a microarray using this set of probes could no longer accurately distinguish between two particular genomes. Thus $\alpha > 25\%$ of the size of s_{min} is necessary. Furthermore, the greater the value of γ , the less likely it is that a newly sequenced pathogen genome's $s_{min}(x)$ will match $s_{min}(i)$ for genome i .

While many sets, s , may meet the criteria above, a fitness function is needed to measure how "good" a particular set s is in order to determine if it is in fact the optimal solution. Three factors contribute to this measurement:

- **A.** Each genome contains a subset of s 's n -mers, $s(i)$. A is the set size of the $s(i)$ with the least number of sequences.
- **G.** Each genome's $s(i)$ may share a subset of n -mers with some other genome's $s(i)$. G is the minimum number of distinct n -mers that are a member of a particular genome's $s(i)$ that are not shared with some other one genome's $s(i)$.
- **set size.** The number of n -mers in s .

Because of the criteria imposed on s_{min} , A and G must be greater than or equal to α for further consideration of set s . The fitness of a particular set can be evaluated by:

$$f(s) = \mathbf{A} + \mathbf{G} / \text{set size}.$$

There are numerous sets s with varying values of A and G for different set sizes whose fitness must be calculated in the search for the optimal minimum set. Figure 3 illustrates this three-dimensional search space. If $f(s_1) > f(s_2)$, s_1 is more fit than s_2 . Thus, for the optimal set s_{min} , there exists no other set with a greater fitness value and $f(s_{min}) \geq (\alpha + \gamma)$. We refer to the subsequences included in s_{min} as being ultraspecific.

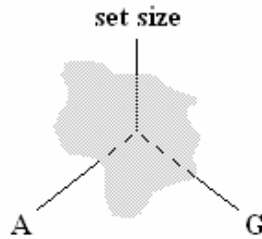


Figure 3. Three-dimensional search space of sets s for which one may be s_{min} .

The pseudo-code below describes how this search space is explored to find the optimal set s_{min} for a given n and t . The parameter mutate is added; mutate indicates how many elements in s will be replaced if s is not optimal.

Given: $n, t, \alpha, \gamma, \text{set_size}, \text{mutate}$

initialize set_size to some value

populate s with set_size t-tolerant n-mers (no duplicates or complements).

old_fitness = 0; //fitness of present "best" solution set

old_set = NULL; //set with "best" fitness

f = 0;

while (f \leq old_fitness)

 while *s is not the optimal solution*

 A = **A** value of s;

 G = **G** value of s;

 if(A \geq α && G \geq γ)

 f = (A+G)/set_size;

 if *f is optimal*

 old_fitness = f

 old_set = s

shrink s and exit inner while loop

 for(i=0; i<(set_size * mutate); i++)

randomly select element in s and replace it with a new t-tolerant n-mer that is not (nor is its complement) already a member of s.

 end while

end while

s_{min} = old_set;

RESULTS

Implementing the pseudo-code just presented, we were able to identify many s_{min} for various sizes of n and tolerances. Here we present the results for the s_{min} of 2-tolerant 19-mers. Looking at the results presented in Table 4, it appears that there are several 2-tolerant 19-mers that are present in the 25 WNV genomes but not any shared by all of the genomes. When taking a closer look at the pair-wise intersection of 2-tolerant 19-mers, a different scenario appears. Genome 12 (WNV isolate LEIV-Krmd88-190 (AY277251)) and Genome 25 (WNV (NC_001563)), coincidentally the same genome with nearly twice the number of 2-tolerant 19-mers as the other genomes and the one WNV genome with a RefSeq accession number, share practically no 2-tolerant 19-mers with the other 23 genomes. Removing these two genomes from consideration, the remaining genomes share 12 2-tolerant 19-mers (actually 6 and their complementary sequences). In some cases 236 2-tolerant 19-mers are shared between sequences, in particular genomes 4 (WNV strain IS-98 STD (AF481864)), 10 (WNV polyprotein precursor (AF533540)) and 22 (WNV strain NY99-flamingo382-99 (AF196835)). Each one of these genomes has only one or two 2-tolerant n -mers not shared with the others.

Before beginning the search for a set of sequences to distinguish and identify WNV genomes, the complementary sequences were removed from each genome's n -mer set. The following threshold values were used: $\alpha = 30\%$ of the set size and $\gamma = 1$. The value of γ cannot be any greater because genomes 4 and 22 only have one unique n -mer each that is not included in the 236 shared between 4, 10, and 22. Consequently, two of the n -mers chosen for a set must be the unique n -mer for 4 and the unique n -mer for 22 in order to meet the γ threshold. The search was started with a set size of 200 sequences. Because of the number of wells in our potential microarray, 96 is the smallest set size to be considered. A set s_{min} was identified for a set size of 96 (thus $\alpha=28$). Figure 4 presents two of the sample patterns of hybridization that would occur when a WNV/human sample were to be introduced to a microarray containing 96 wells in which each well contained one of the ultraspecific n -mers belonging to s_{min} . The blue squares indicate where hybridization of the complementary sequences of the probe and pathogen genome would occur. Each of the 25 WNV has a unique pattern appearing similar to those presented in Figure 4. Deriving the optimal set of 2-tolerant 19-mers from precomputed n -mer sets took approximately 45 minutes on a Pentium 4 2.39GHz processor.

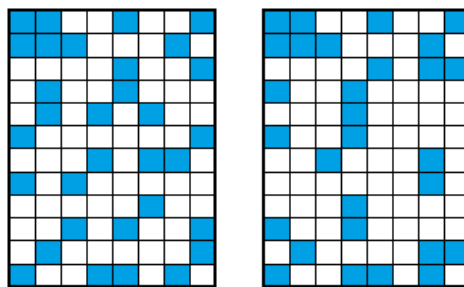


Figure 4. On the left appears the distinguishing pattern for a genome 7 (WNV isolate WN NY 2000-grouse3282 (AF404755))/ human mixed sample. For this genome, 32 members of s_{min} also occur in the genome, thus 32 spots distinguish this pattern. The right presents the pattern that would appear for genome 21 (WNV strain RO97-50 (AF260969)). Genome 7 and genome 21 share 42 n -mers (complementary excluded) of which 19 of those shared n -mers are included in s_{min} .

CONCLUSIONS

Based upon our *in silico* experiments we believe that it is possible to both identify and detect the presence of a particular WNV genome in a WNV/human host mixed sample. Due to high mutation rates, it is difficult if not impossible to identify a particular strain of a virus using a unique sequence. As the virus further mutates and more strains and more viruses become sequenced, this “unique” sequence may no longer be a reliable identifier. Our alternative approach, using a *unique pattern* to identify a particular genome, decreases the possibility of false positives while achieving a greater level of sensitivity. Although we have not tested a microarray using these probes with a WNV/human sample due to availability and access to WNV samples, we have tested our methodology using *E. coli*. An *E. coli*/ human mixed sample was added to a microarray. Half of the probes selected for this microarray were *E. coli*-blind and human-positive; while the other half were human-blind and *E. coli*-positive. This experiment produced the pattern expected from the *in silico* experiments.

There are many extensions to the work presented here that are currently under development. First, further analysis has begun on the *n*-mers shared between particular WNV genomes and perhaps more importantly the two genomes showing very little similarity to the other strains. Moreover, we are extending “host-blind” to include not only human but also mouse, rat, chicken and chimpanzee. We have generated probe sets for the WNV flavivirus neighbor the dengue virus with similar results. Here, however, the fitness function was modified to incorporate the distinction of types within the dengue virus classification. It is our intent to next generate a set of ultraspecific probes that can identify and distinguish between all known flaviviruses. When designing this new set of probes, we intend to also include additional criteria for the sequence selection such as melting temperature, GC content, etc.

REFERENCES

1. Relman, D.A. (1998) Detection and identification of previously unrecognized microbial pathogens. *Emerg Infect Dis*, **4**, 382-9.
2. Markham, A.F. (1993) The polymerase chain reaction: a tool for molecular medicine. *BMJ*, **306**, 441-6.
3. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-70.
4. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat Genet*, **21**, 20-4.
5. Diehn, M. and Relman, D.A. (2001) Comparing functional genomic datasets: lessons from DNA microarray analyses of host-pathogen interactions. *Curr Opin Microbiol*, **4**, 95-101.
6. Dozois, C.M., Daigle, F. and Curtiss, R. III. (2003) Identification of pathogen-specific and conserved genes expressed *in vivo* by an avian pathogenic *Escherichia coli* strain. *Proc Natl Acad Sci (US)*, **100**, 247-52.
7. Wang, D., Urisman, A., Liu, Y.-T., Springer, M., Ksiazek, T.G., Erdman, D.D., Mardis, E.R., Hickenbotham, M., Magrini, V., Eldred, J., Laterille, J.P., Wilson, R.K., Ganem, D. and DeRisi, J.L. (2003) Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biol*, **1**, 257-60.
8. Shi, P.-Y., Kauffman, E., Ren, P., Felton, A., Tai, J.H., Dupuis, A.P. II, Jones, S.A., Ngo, K.A., Nicholas, D.C., Maffei, J., Ebel, G.D., Bernard, K.A. and Kramer, L.D. (2001) High-Throughput Detection of West Nile Virus RNA. *J Clin Microbiol*, **39**, 1264-71.

9. Relman, D.A., Loutit, J.S., Schmidt, T.M., Falkow, S. and Tompkins, L.S. (1990) The agent of bacillary angiomatosis: An approach to the identification of uncultured pathogens. *N Engl J Med*, **323**, 1573-80.
10. Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker, C.T., Saxman, P.R., Farris, R.J., Garrity, G.M., Olsen, G.J., Schmidt, T.M. and Tiedje, J.M. (2001) The RDP-II (Ribosomal Database Project). *Nucl Acids Res*, **29**, 173-4.
11. Nichol, S.T., Spiropoulou, C.F., Morzunov, S., Rollin, P.E., Ksiazek, T.G., Feldmann, H., Sanchez, A., Childs, J., Zaki, S. and Peters, C.J. (1993) Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. *Science*, **262**, 832,834-6.
12. Heller, Arnie. (2004) On the Front Lines of Biodefense. *Science & Technology Review*, 4-9, http://www.llnl.gov/str/April04/pdfs/04_04.pdf.
13. Belapurkar, C., Li, T.-B., Pettitt, B.M., Fox, G.E., Willson, R.C. and Fofanov, Y. (2003) Improved R-Q Set operations facilitate subsequence analysis of genomes. The 20th Annual Houston Conference on Biomedical Engineering Research. Houston, TX: The Houston Society for Engineering in Medicine and Biology, 72.