Assessment of virulence-factor activity relationships (VFARs) for waterborne diseases

* Dept of Crop and Soil Sciences
** Center for Microbial Ecology
*** Dept of Civil and Environmental Engineering
**** Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA
(E-mail: rosejo@msu.edu)

Abstract
Virulence-factor activity relationship (VFAR) is a concept that was developed as a way to relate the architectural and biochemical components of a microorganism to its potential to cause human disease. Development of these relationships requires specialised bioinformatics databases that do not exist at present. A pilot-scale VFAR database was designed for three different waterborne organisms: *Escherichia coli*, *Norovirus* and *Cryptosporidium*, to evaluate VFAR relationships. For the web-based database, each organism has separate pages containing virulence genes, occurrence genes, primer sets and probes, taxonomy, outbreaks, and serotype/species/genogroup/genotype. As the database continues to grow, it will be possible to relate the occurrence and prevalence of certain genes in various microorganisms to outbreak data and, subsequently, to establish the utility of using a combination of specific genes as markers of virulence and in establishing virulence-factor activity relationships (VFARs). The database and the VFARs established will be of use to the regulatory community as a way to assist with prioritising those organisms, which need to be regulated.

Keywords *Cryptosporidium; Escherichia coli; noroviruses; virulence-factor activity relationship (VFAR); waterborne disease*

Introduction
Virulence-factor activity relationships (VFARs) were first developed by The National Research Council’s Committee on Drinking Water Contaminants (NRC, 2001). The conceptual approach described was to (a) use microbial genomic databases to identify new microorganisms, (b) examine the potential human risks of microorganisms found in water, (c) determine the occurrence (persistence, prevalence and magnitude) of the microbial contaminants in water, and (d) then relate these factors to their potential to cause waterborne disease. The VFARs can be used to assess the pathogenicity and virulence of a microorganism similar to the use of quantitative structure–activity relationships (QSARs), which are used to assess the toxicity of chemicals (Cronin et al., 2003). QSARs use the chemical structure of known chemicals to predict the toxicity and solubility of newly identified or produced chemicals. VFARs can, therefore, be established by identifying the risk-associated genetic elements from known waterborne pathogens and then comparing these with new and emerging environmental and waterborne microorganisms to help assess their pathogenicity and virulence.

The development of VFARs for waterborne organisms would be of special interest to those working in the field, because the bulk of the research has been focused on traditional indicator organisms, despite their inability to predict the presence of pathogens (Griffin et al., 2001). This work also relied heavily on the ability to culture these organisms in the laboratory. VFARs will allow the researcher to characterise a water sample, based not on cultivability but on whether or not the sample contains a combination of genetic markers that
indicate that the public is at greater risk of contracting a waterborne disease. The linkages between the ability of an organism to cause harm in humans and the occurrence of these elements in environmentally important organisms can be established through the development of a database and by using a bioinformatics approach to the VFAR concept. The pilot project described in this research was focused on developing this database. The main goals were (a) to build a database for waterborne pathogens and their genetic information, (b) to build a framework which will allow for the development and evaluation of VFARs, and (c) to enable the development of detection methods based on VFARs.

**Materials and methods**

*Selection of preliminary waterborne pathogens*

Three representative organisms were chosen for the pilot project, based on the following criteria: (a) the organism should be known to be transmitted by water, (b) it should be known to cause disease in humans, and (c) a significant amount of genetic information must be available for the selected organism. Based on the above criteria, the bacterial, viral and protozoan candidate organisms chosen for initial analysis were *Escherichia coli*, norovirus, and *Cryptosporidium* respectively.

*Data collection and classification*

The information gathered included both gene sequence data and data connecting the organism to disease. The gene sequence data were classified into different categories, including genes related to virulence and those related to occurrence. Any gene sequence that had been shown at the laboratory level to be involved in or to enhance the disease process (e.g. adhesins, toxins, secretion systems) was classified into the virulence group. Any gene that had been used in the detection of the organism regardless of source (water, clinical or other) was classified into the occurrence group.

For each of the chosen organisms, the primary literature was searched and the available genetic information was divided into two main categories: those genes involved in virulence and those used in detecting the occurrence of these three organisms. Information about the serotype/species/genotype/genogroup and the corresponding reference for each gene sequence was collected. For each gene sequence, a literature search to locate primer sets and probes that had been used to detect the target genes was performed. Relevant data, including the position of the primer within the target gene and the type of molecular method used, were collected along with the primer sets and probes.

The primary literature was also searched to find waterborne outbreaks with an aetiological agent known to be one of the three chosen organisms. The outbreaks were divided into those that had used molecular detection in the identification of the causative agent (which were subsequently included in the database) and those that had not. For each of the outbreaks, there was a series of information that was deemed relevant to the VFARs and included the following: (a) the target genes and the primers/probes used for detection of the aetiological agent; (b) the location and date of the outbreak; (c) whether the aetiological agent was found in clinical and/or water samples collected at the time of the outbreak; (d) if there were other organisms (e.g. coliforms) present in the water samples; (e) the source of the contamination; and (f) data more specific to epidemiology, including symptoms, duration, and number of ill persons.

*Relating genetics to virulence*

Dose-response data have traditionally been used to assess the potency or propensity of an organism to cause disease. There is very little information relating dose-response data to the genetics of an organism. Therefore, it was determined that the genetic information
collected during outbreak situations would serve as the basis in relating an organism to disease. Outbreak organisms, that were identified and characterised at the genetic level, were included in the constructed database, and particular genetic components (sequences) identified within these organisms were labelled and included as putative virulence factors.

Construction of the web-based database

Following initial data collection, the data was formatted in a way that would be useful in building the database structure. This portion of the pilot project was centred on the implementation of how the final database would appear to the end-user and consisted of several phases. The initial phase was the requirements phase, in which “use case scenarios” (detailed descriptions of what a user would be doing when using the database) were developed. These were useful in determining whether all of the data types needed were included. The establishment of use cases also allowed the informaticists to estimate how long it would take to implement the use cases in terms of building the necessary modules. Estimation of time for each use case allowed their ranking and determination of which would be implemented within the pilot project and what would be set aside until further expansion of the database takes place.

The second phase of implementation was the design phase, beginning with the design of the interface. This consisted of designing web-based pages of how the data would fit together and be seen visually by the end-user. After the interface was developed, work began on bringing together the data types and the interactions that they had with one another. This commenced with the development of an entity-relationship diagram (ER-diagram) that captured all of the data types and their interactions that would be present in the implementation. After ER-diagram development, a table design was created and used as the overall template for entering the information into the VFAR database.

Results

Figure 1 depicts the interface design for virulence genes within *E. coli*, which was developed at the beginning of the implementation phase of the database. This allowed the pilot project to be reviewed by selected persons who represented the general users, both govern-
ment and academic, for both content and overall ease of use. This review led to further changes in both the data types being collected and the interface design. It also served as a template for how the occurrence genes are displayed.

Similarly, separate pages were specifically devoted to the categorisation of taxonomy, primer sets and probes, outbreaks, and serotypes associated with each type of organism. The primer sets and probes page consisted of (a) the target gene, (b) the classification group of the gene, (c) the name and sequence of the primer/probe, (d) the method the primer/probe was tested with, and (e) the references. The information found within the outbreak page is listed above. The serotype/species/genotype/genogroup (depending upon the organism) page consisted of the known virulence and known occurrence genes found within that subtype, the circumstances under which the subtype was isolated, and the relevant references.

Figure 2 is the table design that the informaticists and biologists created to link the data types and the database structure. It outlines how the different pieces of data are linked to each other allowing an initial view of the VFARs.

Currently, the pilot project is in the implementation phase. The informaticists are working on the user interface and translating the interface design into validated page code that is then being translated into JavaServer Pages code. This flow of work will continue until the selected use cases for the pilot project have been implemented and the necessary data entered. Data collection will continue until the end of implementation phase. At that time, final checks and corrections will be performed with a finished pilot-scale VFAR database for the three target organisms.

**Discussion and conclusions**

The next steps in developing the VFARs will bring in additional organisms to the database. Potentially, some of the organisms on the current contaminant candidate list (CCL), such as *Aeromonas hydrophila* or Coxsackievirus, should be the next organisms to be added. This would allow comparison of the virulence properties of an organism with its potential to cause waterborne disease. The known outbreaks will also play an interesting role in determining this potential. For example, in the case of *A. hydrophila*, which is ubiquitous in the

---

**Figure 2** Module design for data interactions
environment and contains virulence factors, no single-source outbreaks have been reported (Embrey et al., 2002).

In addition to adding new organisms, there is also a need to bring in additional entry points and bioinformatic tools. Initially, the user would only be able to access the database by entering in through one of the organisms. Once inside, the user would be able to move through the data by clicking on the preset headings. This way of searching is not conducive to the types of usages imaged for the database. Therefore, additional ways of searching for data (by adding search boxes and entry points via virulence, occurrence or outbreaks) is a necessity for efficient utilisation of the database. In order to develop and evaluate VFARs, common bioinformatic tools such as those listed in Table 1 will need to be used. The addition of these tools would be the next step once the basic structure of the database had been completed.

In addition to expanding the database vertically, there is also a need to expand the database horizontally. This pilot project focused only on the genes involved in virulence and occurrence. There are also genes that are thought to be involved in the ability of the organism to persist within the water environment. These factors have not been fully characterised in the primary literature, but may include, for example, the ability to form spores or cysts. As further scientific studies are undertaken, the specific roles of these persistence factors will be better understood and will enhance the ability to determine the VFAR of an organism.

The database must also be expanded horizontally in areas that are not directly linked with available genetic data. These areas include disinfection kinetics, dose-response data and other environmental data (such as associations among indicator organisms). This non-genetic data would be essential in fully developing the virulence-factor activity relationship of an organism. Therefore, determining the VFAR of an organism would not be dependent solely on its genetic make-up, but also on the presence of the organism.

A final piece of expansion that needs to occur, in order to be able to fully understand the VFAR of an organism, would be the ability to compare organisms in a numeric fashion. One way in which this ranking could be done would be by taking all of the genetic and non-genetic information in the database and then ranking the three outcomes (virulence, potency and persistence) individually. However, if a ranking system was developed, the scores would vary depending upon user knowledge of the organisms and their own experiences. A ranking system would give users a tool to compare two very different organisms on the same scale.

The water microbiology community needs to continue to contribute both expertise and innovative data to strengthen the links between organism genetic material and its ability to cause disease. With this assistance, the VFAR database will continue to grow and expand in its ability to determine in a more informative and reliable fashion if new and emerging pathogens present a health risk to the public.

<table>
<thead>
<tr>
<th>Table 1 Bioinformatic tools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tool</strong></td>
</tr>
<tr>
<td>BLAST</td>
</tr>
<tr>
<td>PSI-BLAST</td>
</tr>
<tr>
<td>Phylogenetic tree</td>
</tr>
<tr>
<td>Sequence alignment</td>
</tr>
</tbody>
</table>
Acknowledgements

We would like to thank the members of our laboratories as well as the members of the Ribosomal Database Project for all of their support and help. This research was supported by the US Environmental Protection Agency.

References


