VFAR Activity Group

Update for the CCL Work Group

Plenary Meeting

February 5-6, 2003

Jeff Griffiths, Graciela Ramirez-Toro, and Colin Stine
VFAR Activity Group

Today's update contains…

- Some Initial Findings
- Future Investigations / What's new
- A Proposal to Look at Known Genomes (both viral and bacterial) which we recommend
- Next Meeting Feb 6

VFAR Discovery Phases

- Phase I September-October, 2002
  - ✓ Limited virulence factor keyword search of GenBank
  - ✓ Basic Local Alignment Search Tool (BLAST) alignments
- Phase II October-November, 2002
  - ✓ Comprehensive keyword search of GenBank
  - ✓ Comparison of other available genomic databases
- Phase III November-December 2002
  - ✓ Keyword search of whole genomes
  - ✓ Selective virulence factor sequence alignments against other whole genomes
  - ✓ BLAST alignment of complete virus genomes

Phase I Findings

- 17 keywords and 50 pathogens produced 3,074,532 ‘hits’ in GenBank
- Organism keyword search produced 1 to 1,493 ‘hits’
- 30 of 42 organisms produced virulence factor keyword ‘hits’
- Many ‘hits’ not related to virulence, e.g. structural genes or literature citations

Phase I Challenges

- GenBank contains an overwhelming ("large") amount of data
- GenBank data are not uniformly indexed in a readily searchable format- databases were not set up for our purposes
- Keyword search yields uncertainty about structure/function relationships for some molecules

Phase II Activities

- 64 virulence factor keywords and 63 organisms used in GenBank search
- Representative pathogens selected for BLAST sequence searches
  - ✓ Enterovirus (virus)
  - ✓ *Shigella flexneri* (bacterium)
  - ✓ *Cryptosporidium parvum* (protozoan parasite)

Phase II Findings

- No results with several pathogens while *E. coli* produced 5,869 ‘hits’ – GenBank database reflects prior scientific interests, not necessarily our current or future interests
- Many sequences identified were unrelated to virulence, i.e. structural genes

So: 909 sequences identified from the pathogen test set

- 258 sequences were subjected to BLAST search to seek common sequences
- Keyword searches of TIGR-CMR were compared to those conducted in GenBank

P. II - GenBank vs TIGR-CMR

- TIGR-CMR sequences are better indexed than GenBank sequences
- TIGR-CMR produces fewer spurious results
- TIGR-CMR results were directly related to keyword virulence sequences

Phase III Activities

- GenBank searched for whole genomes of 63 pathogens
- *Yersinia pestis* (agent of classic plague) selected as model bacterial pathogen for BLAST sequence search
- Nine whole virus genomes were chosen for pair-wise BLAST sequence alignment

Phase III Findings

- Whole genome searches yielded 10 viruses, 8 bacteria, and no protozoan parasites (see later codicil on slide 13)
- 33 of 64 keywords produced ‘hits’ in the genome of *Y. pestis*
- 22 *Y. pestis* sequences were found in *E. coli* O157:H7 (many bacteria have 3-5,000 genes)
- 3/9 virus genomes demonstrated significant alignment, in closely related viruses

Phase III Results

- Simply aligning sequences based on bacterial sequences ("BLAST results") sometimes yields ambiguous data about virulence, since virulence of the gene product may only be putative based on similarity to other genes
- Shared sequences among virulence genes was seen between related but not unrelated viruses
- The group has tentatively concluded that the VFAR approach appears feasible for bacteria, and additional work is necessary to assess the VFAR concept in viruses

Initial Findings

- GenBank is an enormous database, expanding exponentially; annotations are sometimes incomplete, ambiguous, and putative; not set up to facilitate keyword searches, primarily functions as a gene sequence repository; not designed for our purposes…
- TIGR-CMR is better organized and indexed than GenBank, but contains fewer entries – thus initial efforts may need to focus on selected databases
- New databases in the works (e.g. D of D)

Initial Findings

(continued)

- The VFAR concept appears feasible for bacteria, and additional work is warranted for assessing VFAR potential in viruses.
- No whole protozoan parasite sequences were found in GenBank, so VFAR approach was not explored with these microorganisms – but *Cryptosporidium* and malaria (*P. falciparum*) genomes are now done and in the wings…

Islands & Sex (bacterial)

- Pathogenicity Islands ("PAIs")
  - ✓ Bacterial genes controlling related functions frequently cluster in contiguous regions on the chromosome
  - ✓ PAIs contain genes controlling production, elaboration, and function of virulence factors
  - ✓ PAIs are frequently exchanged among bacteria, conferring virulence on previously avirulent strains

Islands & Sex (bacterial)

- Extra-chromosomal Nucleic Acids
  - ✓ Bacteria exchange genetic material by conjugation, transduction, and transformation
  - ✓ Extra-chromosomal nucleic acids may be present in the form of plasmids (bits of DNA) or bacteriophage (a virus); they frequently mediate virulence in bacteria
  - ✓ Extra-chromosomal nucleic acids have not been considered a potential VFAR

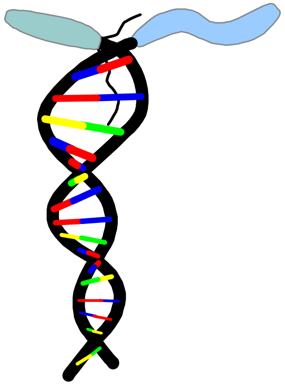
Wassenaar Proposal

Identification of Gene Combinations Associated with Water-borne Pathogen Virulence

T. Wassenaar and J. Gamieldien

Molecular Microbiology and Genomic Consultants

Zotzenheim, Germany



Pilot Project Goal

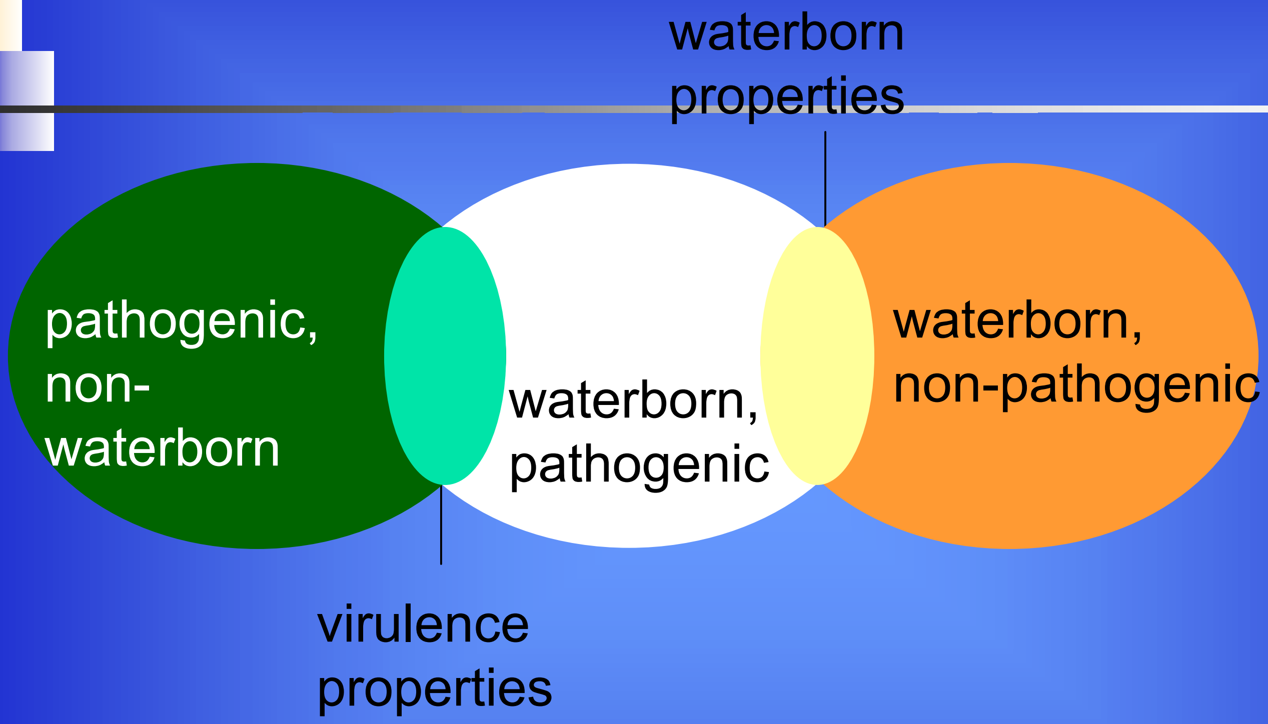
Identify genes and gene combinations unique to waterborne pathogens that may be associated with virulence, by searching available genomic databases using an automated, high-throughput gene content comparison to produce a listing of relevant genes, gene combinations, including their annotations, and phylogenetic relationship.

Research Approach

- High-throughput computational comparison of genome of four groups
  - ✓ Water-borne pathogens
  - ✓ Water-borne non-pathogens
  - ✓ Pathogens that are not water-borne
  - ✓ Non-pathogens that are not water-borne

- Concept will be applied to viruses first, then bacteria

Novel approaches



Subtract: all genes from non-pathogenic, non-waterborn organisms (minimal gene set)

Deliverables

- Detailed report containing procedures, results, and recommendations
- Multi-sequence files of coding sequences, and proteins for virulence factors unique to water-borne pathogens
- Annotations for each orthologous data set
- Phylogenetic distance and E-value matrices for each orthologous protein

Project Milestones

Time	Accomplishment
At 3 months	Genome comparison for viruses completed
At 6 months	System optimized for bacterial genomic data, data freeze
At 9 months	Genome comparison for bacteria completed
At 12 months	Matrix interpretation for identification of water-borne pathogen-specific genes completed
At 14 months	Final report and electronic datasets compiled and delivered