

Human and Ecological Risk Assessment, 13: 1–15, 2007
Copyright © Taylor & Francis Group, LLC
ISSN: 1080-7039 print / 1549-7680 online
DOI: 10.1080/10807030701226855

The Importance of Virulence Prediction and Gene Networks in Microbial Risk Assessment

Trudy M. Wassenaar,¹ Junaid Gamielidien,¹ JoAnne Shatkin,² Petra Luber,³
Nelson Moyer,² Tom Carpenter,⁴ and David W. Ussery⁵

5 ¹Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany; ²The
Cadmus Group, Watertown, Massachusetts, USA; ³The Federal Office of Consumer
Protection and Food Safety, Berlin, Germany; ⁴Environmental Protection Agency,
Washington, DC, USA; ⁵Center for Biological Sequence Analysis, DTU, Lyngby,
Copenhagen, Denmark

10 ABSTRACT

For microbial risk assessment, it is necessary to recognize and predict virulence
of bacterial pathogens, including their ability to contaminate foods. Hazard char-
acterization requires data on strain variability regarding virulence and survival during
15 food processing. Moreover, information on virulence is important for qualitative or
quantitative description of public health outcomes following infection. The quest
to understand bacterial disease started with the isolation of bacterial pathogens and
continued with elucidating the mechanisms of bacterial pathogenicity. Now the goal
is to predict virulence genes from total genome sequences. The deterministic ap-
proach of considering gene function relating to an organism's pathogenicity has its
20 limits. Gene function prediction based on sequence similarity is also not without
flaws. Bioinformatic analysis can reveal virulence potential of a genome-sequenced
strain. However, a gene's contribution to phenotype is determined by the context
of other genes present in the genome and this should be considered. Quantita-
tive effects of gene expression should also be taken into account. Thus, if the gene
25 networks essential for bacterial pathogenesis are understood, we can better predict
genes coding for virulence. It may even become possible to identify species that are
not yet pathogenic, but have the correct genetic repertoire to become so if partic-
ular genes were acquired. Gene network identification may become an important
component for identification and characterization of microbial hazards, including
30 emerging pathogens, in the context of microbial risk assessment.

Key Words: virulence gene recognition, gene networks, pathogenicity, emerging
pathogens.

Received ; accepted .

The views expressed are solely those of the authors and are not representative of policies or
opinions of the U.S. Environmental Protection Agency.

Address correspondence to Dr. Trudy M. Wassenaar, Molecular Microbiology and Genomics
Consultants, Tannenstrasse 7, 55576 Zotzenheim, Germany. E-mail: mmgc.de@t-online.de

Q1

T. M. Wassenaar *et al.*

INTRODUCTION

The technique of Microbial Risk Assessment (MRA) has evolved tremendously over the last 15 years. Risk assessment has been used effectively in the past to assess and manage risks in chemistry and engineering, but its application to food safety has been more challenging when dealing with living food-borne pathogens. The field of MRA is rapidly developing, directed to assessing risks of microbial pathogens by analyzing molecular and genetic information. 35

Bacteria are defined as pathogenic when they cause damage to their host during colonization. Robert Koch postulated, in 1890, three criteria determining whether bacteria isolated from a host should be considered pathogenic. These criteria, crucial to the development of clinical bacteriology, still hold, but with the discovery of genes and the development of molecular biological tools, a deterministic approach has become fashionable. The challenge is to determine which genes are responsible for the pathogenicity of bacteria. By analogy, the molecular Koch's postulates describe three criteria for defining virulence genes: (i) a virulence phenotype should be found in pathogenic members of a genus or species; (ii) inactivation of the virulence gene(s) should result in a measurable loss in pathogenicity or virulence; and (iii) reversion or allelic replacement of the mutated gene should restore pathogenicity (Falkow 1988). For simplicity, the difference between virulence and pathogenicity is ignored in this work. Furthermore, disease is the outcome of an intricate interplay between host and pathogen (and sometimes interference with benign microorganisms present at the site of colonization) but these other players are also excluded here. Finally, we concentrate on bacterial pathogens in this Perspective, although we realize that by excluding viral and protozoan pathogens we ignore one important focus in the field of microbial risk assessment. 40 45 50 55

A new era of bioinformatics and genomics started with the sequencing of the first bacterial genome, which was a pathogen with a relatively short genome: *Haemophilus influenzae* (Fleischmann *et al.* 1995). The challenge shifted from identifying virulence genes by deterministic approaches to filter out virulence genes from complete genome sequences. This challenge proved more difficult than anticipated. As of March 2006, there were approximately 300 bacterial genomes publicly available, representing 220 species of which at least 120 can be considered pathogenic to humans. This fraction of pathogenic genomes is based on the known pathogenicity of members of the sequenced species, not taking into account if the actual genome sequence was derived from a pathogenic strain or not. That is not to say that we can recognize the genes essential or accessory to pathogenicity of each of these pathogens. And it would be difficult, with current knowledge, to recognize an unknown pathogen based on its genome sequence only, or to predict an emerging pathogen from genomic characterization. This view may seem pessimistic, but the field of genomics is rapidly evolving and soon we may be able to better predict virulence. 60 65 70

Recognition and prediction of virulence would be a most valuable asset for various aspects of microbial risk assessments. Although we know for a number of bacterial species that particular strains or clones are more pathogenic or virulent than others, currently for risk assessment all members of a species are considered as equal—for simplicity, lack of hard data, or out of precaution. Obviously, the results of such 75

Microbial Risk Assessment and Genomics

analyses do not have the desired accuracy and the limitations of this approach have
 80 been discussed (Schlundt 2000). For the future we wish hazard characterization
 to include more precise data on variability in virulence between strains; exposure
 assessment would take into account variation in survival properties during food
 processing and in infectious dose. We have not reached this level of precision yet.
 Obviously, determining the presence of *E. coli* in beef, for example, is not precise
 85 enough for correct risk assessment without differentiation between pathogenic and
 non-pathogenic strains; concentrating on *E. coli* O:157 (Cassin *et al.* 1998), however,
 is too narrow as it ignores other harmful serotypes, including the emerging shiga-
 toxin producing O:103 (Hussein and Sakuma 2005). Detection of virulence genes
 rather than their bacterial carriers would overcome this limitation, but then the ge-
 90 nomic content required for a virulence gene to confer pathogenicity also needs to be
 taken into account. The presence of a virulence gene per se is not sufficient to make
 that strain pathogenic, as exemplified by the observation that apathogenic *E. coli*
 isolates can contain the gene for hemolysin A, a known virulence gene for various
 pathotypes of *E. coli*, or shiga-toxin producing strains can be isolated from healthy in-
 95 dividuals (Boerlin *et al.* 1999). The limitation of gene identification due to technical
 limitations (the theoretical sensitivity of PCR detection is hardly ever reached when
 dealing with food carriers) is further limiting precision of data; enrichment pro-
 cedures partly overcome this, and also prevent detection of dead micro-organisms,
 but at the cost of speed. Technical limitations also apply to bacterial detection when
 100 dealing with viable, non-culturable forms (Wong *et al.* 2004). Due to data gaps, cur-
 rent models typically assume that virulence potential and persistence capacities are
 constant within species, and that these properties remain constant within the farm-
 to-fork chain. Exposure assessments could improve significantly if knowledge about
 relevant bacterial abilities of typical bacterial strains (*e.g.*, to survive processing steps
 105 or to persist in kitchens environments), was available. Dose response models do not
 generally incorporate strain-to-strain variation in their potential to cause disease.
 What needs to be known in order to quantify risks is how frequent genetic determi-
 nants for these properties are present in particular bacterial populations, and how
 important their impact is. Better information on virulence and persistence would
 110 enable us to obtain a more precise output for qualitative or quantitative description
 of public health consequences following infection. This Perspective summarizes the
 status quo, strategies and approaches that may or may not be able to predict more
 accurately virulence from genomic sequences.

PROBLEMS IN IDENTIFICATION OF GENES INVOLVED IN VIRULENCE

115 Limitations of a Deterministic Approach

When Falkow proposed his molecular Koch's postulates (Falkow 1988), it seemed
 by all means possible to establish, in an unbiased way, the role of bacterial genes in
 virulence. When a gene candidate would result in attenuation (weakened virulence)
 after inactivation, and when this attenuation could be restored by complementation
 120 of an intact copy of the gene, the gene in question could safely be called a viru-
 lence gene. Many hazard characterizations within MRA approaches are based on
 this definition of virulence genes. Nevertheless, this approach of identifying and
 defining virulence has its limitations, as has been outlined (Wassenaar and Gastra

T. M. Wassenaar *et al.*

2001). Basically, a pathogen is well adapted to the niche it colonizes. Destruction of any gene involved in this adaptation will result in (partial) loss of fitness, and thus decreased virulence. Large-scale mutant screening, backed up by genomic sequencing, is now feasible (Benton *et al.* 2004). However, the identified genes may not directly be involved in causing damage to the host, and only genes with that function are included here in the narrow definition of virulence genes. It is appreciated, however, that virulence genes need a whole array of accessory proteins in order to be functional and these are essential for the virulence phenotype. These may include regulation of expression (transcription factors and others regulating expression of virulence genes), protein folding (specific chaperonins), secretion (specified secretion systems such as type III secretion), *etc.* These would be more correctly grouped as "virulence-associated," although when inactivated, such genes result in attenuation. Even the inactivation of housekeeping genes (the term is used here to describe genes involved in cellular metabolism, maintenance and division) can result in decreased virulence. For instance, inactivation of the biosynthesis of aromatic amino acids by inactivation of *aroA* results in attenuation of *Salmonella* (Dougan *et al.* 1987). Including *aroA* as a virulence factor based on this attenuation can be considered a "false positive." The opposite, "false negatives" can result from complementary bacterial strategies: absence of phenotypic effects after inactivation of a true virulence gene can occur when other factors take over the missing function. That published examples of such negatives are limited might be caused by the bias for "positive" results for publication. Nevertheless, a well-adapted pathogen may have evolved multiple strategies resulting in a degree of redundancy or robustness for essential processes; virulence strategies may be considered essential for pathogens and indeed a robust virulence phenotype is encoded by redundant gene sets within a species as illustrated for *Salmonella* (Becker *et al.* 2006). Virulence can even be enhanced when certain genes are absent; such virulence enhancing deletions have been termed "black holes" and obviously these missing genes would be difficult to detect unless one knows how to look for them (Maurelli *et al.* 1998). The deterministic approach fails altogether when there is no animal model available to mimic disease. This applies even to common pathogens such as *Campylobacter jejuni* (Newell 2001). Finally, we might be missing some important players in the field, risking an underestimation of virulence. The presence of non-coding RNA (ncRNA) has been shown to be involved in pathogenesis. As an example, Pichon and Felden (2005) have found that several ncRNAs are produced only in pathogenic strains of *S. aureus*. It is probably only a matter of time before ncRNA genes are included in gene annotation or virulence determination.

To sum up, currently all published MRAs have been struggling to define virulence of bacterial pathogens, but recent progress in defining and characterizing virulence will enable future incorporation of virulence data based on molecular information of genes. The increase of knowledge in this field will significantly decrease uncertainties in future MRAs.

Limitations of Gene Comparison

Maybe we should forget about phenotypic screening and let the genes tell their story independent of genetic intervention. Because pathogens frequently use

Microbial Risk Assessment and Genomics

170 common strategies to cause disease (Finlay and Falkow 1989, 1997), we should
be able to recognize virulence genes by comparing gene and genome sequences.
Genome sequence comparison is facilitated by algorithms that are becoming ever
175 smarter (Chain *et al.* 2003; Azad and Borodovsky 2004). Nevertheless, drawing con-
clusions about virulence from genome comparison is not without flaws. The inter-
pretation depends heavily on correct annotation of gene function. Originally, the
function of a gene was experimentally assessed before it was published. Nowadays,
180 in most cases gene function is predicted from sequence similarity, and genome an-
notation is based on this process. Pitfalls of functional gene annotation based on
sequence homology are noted by Wassenaar and Gaastra (2001) and Wassenaar
(2004). Notably, gene diversity among bacterial populations is much larger than the
185 textbook view. Genes are stolen and misused by bacteria. Horizontal gene transfer
may result in genes that still bear the sequence characteristics of their function in the
species from which they originated, but may have adapted to novel functions that
are not obvious from alignment comparison. This, together with database pollution
by wrongly annotated genes and “putative” functions, hampers predictions on gene

185 function based on similarity scores only.
On a more positive side, there are many excellent models and databases available,
such as Pfam, although wisdom is needed in choosing which database is best to use
for a given situation, and what to make of the results. A few examples are listed here
that illustrate the limitations of gene similarity interpretation, even if they do not all
190 relate to virulence.

The first example deals with the algorithm applied for gene similarity and recog-
nition searches. The most widely applied algorithm is BLAST, an acronym for Basic
Local Alignment Search Tool (Altschul *et al.* 1990). This algorithm was designed to
identify gene similarity in coding region, but it is less suitable to identify similarity
195 in RNA genes. Several original genome publications missed the annotation of rRNA
genes. As an example, the bacterial genome sequence of *Agrobacterium tumefaciens*
(Goodner *et al.* 2001) had no annotation of the ribosomal 16S and 23S genes; this
omission was only partly corrected in the latest update, as on one chromosome the
5S and 23S rRNA genes are still missing. This not only illustrates the limitations
200 of BLAST (there are alternative search programs that are better equipped to iden-
tify rRNA and tRNA genes, such as tRNAscan (Schattner *et al.* 2005)) but it also
illustrates that there is no quality standard for genome annotation. In the published
genome of *Borrelia burgdorferi*, the 16S rRNA gene was annotated on the wrong strand
(Fraser *et al.* 1997), a mistake that was corrected in a later update of the GenBank
205 file. It should be obvious that ribosomal genes, essential to every living organism,
should be annotated correctly in a published genome. Even the genome sequences
of well-studied *E. coli* are not without flaws. The GenBank genome file of *E. coli*
CFT073 contains ambiguous sequences in the rRNA genes (Welch *et al.* 2002) and
the 2006 update version of an *E. coli* O:157 GenBank file still contains a large gap
210 of unsequenced DNA (Perna *et al.* 2001). Recognition of false or incomplete an-
notation is, unfortunately, still largely limited to a few specialists (bioinformaticists
with a good biological background, or microbiologists with a good understanding
of bioinformatics).

215 A second example demonstrates how confusing horizontal gene transfer in a
bacterial genomes can be. Selenocysteine (Sel) is a recently discovered amino acid.

T. M. Wassenaar *et al.*

It is ubiquitous in eukaryotes, and can replace cysteine in particular bacterial proteins as well, notably in formate dehydrogenase, where it uses the codon TGA (normally reserved for a stop). For correct incorporation into formate dehydrogenase, 4 genes are required: an enzyme to produce selenophosphate from Se^{2-} (SelD); one to encode the selenocysteine tRNA (*selC*); the aminoacyl transferase that loads this tRNA (encoded by *selA*); and an elongation factor to help the ribosome recognize the TGA codon to encode Sel and not stop (SelB) (Wassenaar and Meinersmann 2003). An analysis of bacterial genomes for the presence of these four genes, with the best tools available, resulted in confusing findings: one or more genes of the *selABCD* gene set could be missing, in which case formate dehydrogenase would use cysteine and not Sel. The reason why an incomplete gene set was retained on bacterial genomes could not be explained. Even more confusingly, *selABCD* could be present but Sel was not incorporated in formate dehydrogenase, nor in any other gene that could be identified, as in *Clostridium perfringens* (Wassenaar and Meinersmann 2003). Would the machinery of selenocysteine incorporation fulfil a different function in such a case? Or are genes maintained that are not functional? Either explanation is hard to accept. Bacterial genomes can have puzzling contents.

A third example is the presence of secretory mechanisms in Gram-negative bacteria. There are at least 5 secretory mechanisms described, with type III secretion systems (T3SS) frequently involved in the secretion and injection into target cells of virulence factors. This identifies T3SS components as virulence-associated factors (though frequently described as “virulence factors”). T3SS form a complex appendage structure consisting of multiple proteins. Although the components of these structures are redundant to some extent, one can expect that a minimally required number of genes need to be present in order for the secretory system to be functional. Nevertheless, if a component of a particular secretory system is identified in a genome, this will be stated in its annotation, irrespective of the presence of necessary further components. Thus, identifying a single T3SS component will not tell much, when it has not been assessed whether or not that T3SS is complete. Conversely, identifying a T3SS signal present in a particular virulence factor is useless if that cell does not even have a T3SS. Examples of such “orphaned” genes are plentiful in the bacterial world, and a first attempt is underway to assess the complete secretion repertoire of completely sequenced bacteria (Bendtsen *et al.* 2005). Such approaches should prevent misleading annotation and illustrate that the presence of individual genes or signals needs to be assessed in relation to the genomic background before a functional prediction can be accurately applied. A survey of more than 200 sequenced prokaryotic genomes found that, on average, most sequenced bacterial genomes are over-annotated by about 20%—that is, for every 120 genes, 100 are likely to be “real,” and the other 20 genes could be artefacts due to poor gene-finding and annotation (Ussery and Hallin 2005). Furthermore, in one of the few examples where the same team that sequenced a genome also investigated all the proteins expressed, they found that 5% of the proteins that were expressed had not been predicted (Jaffe *et al.* 2004). Thus genome annotation is inaccurate in that both too many genes are predicted as well as “real genes” are missing. In addition to the annotation problems, information on the biological background of a sequenced strain is frequently inaccurate and incomplete. Fortunately, there is a coordinated effort to obtain additional biological information about

Microbial Risk Assessment and Genomics

- the environment, growth conditions, *etc.*, of the host organism (Field and Hughes 2005).
- 265 An initiative to improve data collection and mining was the development of a database storing, for known pathogens: virulence genes, outbreak information, serotypes, genotypes and taxonomic information of waterborne pathogenic species, to evaluate virulence factor activity relationships (VFAR) (Jenkins *et al.* 2004). This VFAR concept could also be applied to food-borne pathogens. When this database
- 270 is linked to databases containing genome sequences, with improved annotation and correct biological information on the sequenced strain, data mining for risk assessment purposes will greatly improve.
- Current variation in quality of published information on virulence requires un-
- 275 recognised efforts to avoid inclusion of unreliable or inaccurate data in MRA. Although we can hope for powerful information from molecular analysis of pathogens in the future, data now available should be reviewed carefully when used for hazard characterizations or exposure assessments. Databases storing genetic and molecular information should be designed more carefully for MRA purposes.

NOVEL STRATEGIES OF VIRULENCE GENES

280 Novel Approaches Using Comparative Genomics

- Determining a genome sequence of a pathogen doesn't automatically reveal its virulence potential. For example, two genome sequences of *Campylobacter jejuni* have not been sufficient to identify the strategies of how this species causes diarrhea (Parkhill *et al.* 2000; Fouts *et al.* 2005). Similarly, the genome sequences from two
- 285 *Helicobacter pylori* isolates (not a food pathogen) have not elucidated novel insights into its pathogenicity (Tomb *et al.* 1997; Alm *et al.* 1999). Despite the listed limitations, comparative genomics can perform where deterministic biology fails. In future, comparative genomics combined with proteomics (identifying the complete protein content of a species) and metabolomics (defining all metabolic pathways in a cell), for instance, is expected to help to predict the survival rate for a bacterial
- 290 population in case of the application of a novel food processing technology.
- The elegance of genome comparison is that genes with unknown function can still be compared. In a typical genome, for 30 to 40% of the identified open reading frames (orfs) we do not know the function, and for perhaps half of these there is not
- 295 even a match in the sequence database. When a species comprises both pathogenic and non-pathogenic strains, genome comparison of both can reveal the genetic basis for this difference in lifestyle, including genes with "unknown function." Intra-species comparison has been done for *E. coli*, where K12 represents a non-pathogenic strain and O157:H7 a pathogenic strain. The comparison of both genome sequences
- 300 revealed the importance of bacteriophages as vehicles for horizontal gene transfer (Hayashi *et al.* 2001). Over 1600 orfs (30%) were identified as present in the pathogenic but not in the benign strain, and of these at least 131 were assumed to be virulence related. When comparing the gene content of two sequenced O157:H7 and uropathogenic *E. coli* CFT073 with a K12 strain, this number decreased to 44
- 305 pathogen-specific genes (J. Gamieldien and T.M. Wassenaar, unpublished observations). Naturally, plasmid-encoded genes should be included in such analysis, when

T. M. Wassenaar *et al.*

they are part of the genome of a given strain. Unfortunately, plasmid carriage can be strain-dependent. In the NCBI genome database plasmid sequences are listed without information from which strain these were isolated. Sometimes it is hard to know whether a particular plasmid is part of a genome of a particular strain with described virulence properties. This is one of the consequences of database structure that could not have been foreseen. A further drawback is that for some sequenced strains their virulence potential is unclear. In the case of *C. jejuni* 11168, the published paper reports the strain to be of clinical origin, but that only applied to the strain when it was isolated nearly 20 years earlier (Palmer *et al.* 1983). Since then, the strain has been cultured and stored for so long, that it has changed its phenotype and morphology, and considerable differences between early and late isolates can be detected by microarrays (Gaynor *et al.* 2004). It remains to be seen what consequences this has had, if any, to the genome sequence. Similarly, the first *H. influenzae* strain that was used for genome sequencing was of a non-pathogenic serotype, and only recently a pathogenic strain was sequenced that contained 280 orfs absent in the non-pathogenic strain (Harrison *et al.* 2005). Multiple genome sequences within one species are most informative when they represent a wide range of pathogenic and non-pathogenic representatives, provided sufficient information is available for the given strain.

The power of genome comparison was tested with a daring question we asked ourselves: can we recognize genes that are specifically present in bacterial human pathogens that can spread via contaminated water? The hypothesis leading to this question was that such waterborne pathogens would have particular phenotypes in common. For instance, they would have to be able to survive, if not multiply, in surface water and they would have to be pathogenic to humans via the oral-fecal route. These requirements would predict restrictions to growth temperatures and other conditions, which would be reflected in gene content. If we could identify some of these genes, we would be able to predict and detect emerging waterborne pathogens from sequenced genomes in the future, even before these species would be recognized as such. The approach followed was to select bacterial genomes from the public domain, one genome per genus, and group them according to pathogenic or non-pathogenic, waterborne or non-waterborne. The complete gene content of these genomes was then compared within and between these groups. The result was disappointing: there were no common genes conserved in the group of interest, waterborne pathogens. Apparently, various strategies have evolved in various pathogens to fulfil similar function, each with their own genetic make-up. In addition, genes can be shared between bacterial species that have been adapted to a species-specific function, so that genes are conserved between species with different life styles. Either way prevents a group of genes being filtered out that is conserved to a particular life-style.

We then focussed on those (related) species of which multiple genome sequences were available that represented different degrees of pathogenicity. An analysis of 3 pathogenic and one non-pathogenic *E. coli* combined with two pathogenic *Shigella flexneri* (which is closely related to *E. coli*) strains identified 30 conserved genes present in the pathogenic organisms only. Fifteen of these were hypothetical proteins or proteins of unknown function. Many of the identified genes were associated with mobile elements such as transposons or prophages, which exemplifies the importance of such moieties in the evolution of pathogens. In theory, the identified genes

Microbial Risk Assessment and Genomics

do not have to be involved in virulence themselves, but can be markers for virulence
 355 instead. When a BLAST search was performed for each gene against the complete
 GenBank database, 13 of them only found homologues in pathogenic organisms.
 Interestingly, in many cases the only pathogens with homologues were also food or
 waterborne pathogens. A similar analysis for *Vibrio* spp. combined with the closely
 related, apathogenic *Photobacterium profundum* resulted in 31 pathogen-associated
 360 candidate genes (Table 1). For this analysis, the genome sequences of pathogenic
Vibrio vulnificus CMCP6, *V. vulnificus* YJ016, *V. parahaemolyticus* RIMD, and *V. cholerae*
 El Tor N1696 were compared to that of closely related, non-pathogenic *P. profundum*.
 Genes with a homolog in all three pathogenic species but not in *P. profundum* were
 then checked for homologs in the complete GenBank database. The table lists only
 365 those genes (the accession numbers for *V. cholerae* are given) that either found no
 homologs in the database, or found homologs in other pathogens only. In a few
 cases homologs were identified in other pathogens that are apathogenic to humans,
 as indicated in the last column. With more genome sequences becoming available,
 such analyses will become possible for other pathogens too, and will gain accuracy
 370 with each additional genome.

We dare to predict that approaches using comparative genomics will lead to valu-
 able information on virulence bacterial strains within the species in the near future.
 Such data will enable better quantification of public health effect of infectious bac-
 terial agents and reduce the total uncertainty in MRAs. Moreover, knowledge from
 375 comparative genomics might allow predictions of future risks arising from newly
 emerging pathogens, possibly even before these actually become virulent or alter
 their resistance to better survive microbial reduction steps during food processing.

The Value of Gene Regulatory Networks

Thus far, the emphasis was on presence or absence of genes, and correct pre-
 380 diction of their role in virulence. But expression of genes is tightly regulated, and
 regulatory factors belong to the “virulence-associated” class. Two related questions
 are relevant in the context of virulence gene regulation. First, what is being regu-
 lated? And second, how is this regulation carried out? Over the course of evolution,
 bacteria have evolved highly efficient mechanisms for maintaining robust functional
 385 regulatory networks of genes (RNA and protein) which can withstand major assaults
 and still be functional. As a consequence, few “essential” genes are identified by
 means of “knock-out” experiments, because the network can still function (though
 perhaps not as well) when a given gene is lost. Redundancy is a successful strategy
 for adaptation. The key issue is to assess regulation of functional networks rather
 390 than regulation of individual genes. From this perspective there could be many dif-
 ferent ways of obtaining the same (or similar) phenotypes. The answer to the first
 question posed here, “what is being regulated” should therefore not be genes, but
 functional networks of gene products, such as “invasion” or “macrophage survival.”
 These functional networks regulate bacterial abilities that need to be addressed in
 395 MRA, *e.g.* the ability of bacteria to survive food processing, to multiply in particular
 food matrixes, or to persist in an environment before a human host can be colo-
 nized, or to successfully escape a host’s immune-defense strategy. In contrast to the
 identification of virulence genes described earlier, in gene regulatory networks it

T. M. Wassenaar *et al.***Table 1.** Genes identified from pathogenic *Vibrio* spp. with homologues in other species.

Tigr,GenBank Accession	Homologues found in human pathogens (a) (b)	Homologues found in other pathogens
VC0549	None	None
VC0666	None	None
VC1305	None	None
VC1371	None	None
VC1380	None	None
VC1506	None	<i>Bdellovibrio bacteriovorus</i> (a scavenger of other bacteria) <i>Edwardsiella ictaluri</i> (fish pathogen)
VC1600	None	None
VC1610	None	None
VC1644	None	None
VC1689	None	None
VC1748	<i>Salmonella typhi</i>	<i>Edwardsiella ictaluri</i> <i>Photobacterium luminescens</i> (insect pathogen)
VC1749	<i>Salmonella typhi</i>	<i>Edwardsiella ictaluri</i> <i>Photobacterium luminescens</i> <i>Bdellovibrio bacteriovorus</i> <i>Bacteroides thetaiotaomicron</i>
VC1750	<i>Salmonella typhi</i>	<i>Edwardsiella ictaluri</i> <i>Photobacterium luminescens</i>
VC2038	None	None
VC2662	<i>Aeromonas caviae</i>	None
VCA0048	<i>E. coli</i> <i>S. flexneri</i>	None
VCA0050	None	None
VCA0051	None	None
VCA0078	None	<i>Desulfovibrio vulgaris</i>
VCA0485	None	None
VCA0559	None	None
VCA0568	<i>Salmonella spp.</i> <i>E. coli</i> <i>S. flexneri</i>	<i>Vibrio (Listonella) anguillarum</i> (fish pathogen) <i>Erwinia carotovora</i> (plant pathogen)
VCA0595	None	None
VCA0695	None	None
VCA0735	None	<i>Vibrio (Listonella) anguillarum</i> (fish pathogen)
VCA0738	<i>E. coli, Escherichia fergusonii</i> <i>Salmonella spp.</i> <i>S. flexneri</i>	<i>Erwinia carotovora</i>
VCA0893	None	None
VCA0892	None	None
VCA0942	None	None
VCA1024	None	None
VCA1050	None	<i>Desulfotalea psychotrophila</i>

(a) Expect value (E value) < 0.0001.

(b) In all cases, the gene was identified in *V. cholerae*, *V. vulnificus*, and *V. parahaemolyticus*.

Microbial Risk Assessment and Genomics

is not presence or absence, but more or lesser expression of genes, that makes the
400 difference.

To focus on how these networks are being regulated, we can recognize three
fundamental levels of regulation, all of which can play a role in virulence. The first
is regulation at a global level, through chromatin proteins such as H-NS (Dorman
2004) that regulate DNA folding. This regulation is not very specific, and a very
405 large fraction (more than half) of the genome can be regulated by a relatively small
handful of abundant (~100,000 molecules/cell) chromatin proteins (Fis, IHF, H-NS,
HU). H-NS normally represses many genes, which are up-regulated in *hns* mutants.
Many of these genes are environmentally sensitive, and virulence-associated. The
second level of regulation is through sigma factors, which are essential in initia-
410 tion of transcription. The main sigma factor, RpoD in *E. coli*, exists at roughly 1000
molecules/cell during log phase growth. Other sigma factors, which exist at lower
concentrations, regulate specific sets of genes and gene networks, and their nature
and number vary per species. One of these families of Sigma factors is plasmid-borne
and regulates toxin gene expression (Kill *et al.* 2005). Finally, the third level of regu-
415 lation is transcription factors, which are quite abundant in number of different types
(several hundred in many bacterial genomes), although the copy number for most
is in the range of 10 or fewer molecules per cell. These regulate very specifically a
small set of genes and include many virulence-associated regulatory factors.

That virulence must be tightly regulated is not always appreciated. It is well known
420 that under-expression of virulence genes leads to attenuation, such as non-hemolytic
Streptococcus strains, but less accepted is that over-expression can also result in atten-
uation, as has been demonstrated for *Listeria monocytogenes* (Roberts *et al.* 2005). In
particular cases such quantitative effects might be relevant to address in MRAs.

Regulation of gene expression is studied at a whole-cell level by use of microarray
425 techniques. Currently most microarray experiments focus on differential expres-
sion: is the expression of individual genes expressed higher or lower relative to a
given gene? However, one could get much more information from an experiment,
in particular from high density Affymetrix or NimbleExpress chips, when the levels
of transcripts were (roughly) quantified. This requires proper normalisation, which
430 might be able to deliver information on how abundant a given transcript is. For
example, two genes might both be up-regulated 5-fold under particular conditions,
but if one gene changes from 2 copies per cell to 10, this might not have the same
effect as that of another, more abundant transcript, changing from approximately
10,000 molecules to 50,000 molecules per cell. The current practice of differential
435 expression ignores such quantitative effects, which is a missed chance to get more
insights into regulatory networks. Again, making the most of gene expression mi-
croarray techniques can provide very useful new inputs for hazard characterizations
and exposure assessments in the frame of MRAs.

440 VIRULENCE PREDICTION AND GENE NETWORKS IN MICROBIAL RISK ASSESSMENT

Although established and of proven value in other areas, risk assessment of mi-
croorganisms is still in its infancy. Most approaches so far have been made for
pathogen-commodity combinations and tried to address the whole food chain (*e.g.*

T. M. Wassenaar *et al.*

Rosenquist *et al.* 2003). The Achilles' heel of such farm-to-fork approaches is a notorious lack of data. When available, data are often of poor quality and assessors have to rely on surrogates and guesses. This is most acute for factors determining survival in the food chain, infectious dose, and virulence regulation. In case of data gaps, smaller risk assessments tackling only a part of the food chain or concentrating on a clear decision question can be very useful. 445

By using risk analysis models reflecting plausible interpretations of the realities of nature, observations (the facts we know) can be used to make predictions about uncertainties (things we do not know). In case of microbial pathogens, risk analysis can deliver insights into mechanisms of pathogenicity and increase our understanding of host-pathogen relationships, including the crucial factors that lead to infection and immunity (Kirschner *et al.* 2005). Molecular and cellular data are combined with epidemiological data in mathematical models. Because few virulence genes are shared by all members of a species, data on prevalence of genes within natural populations can be used to weigh the importance of individual genes. Once a system is described mathematically, this then serves as a starting point from which hypotheses can be generated and tested. For example, mathematical modelling can help us to get a better understanding of those gene networks that are essential for bacterial pathogenesis. The use of mathematical models to predict virulence can even help in reducing the number of challenge tests and animal experiments required. In due course, identification and prediction of genes responsible for virulence will improve. Gene network identification can then become an important component for identification and characterization of microbial hazards. Our skills to predict emerging pathogens are still limited, but knowing that pandemic *Salmonella* strains have replaced each other in the food chain in the past (Davis *et al.* 2002), we can predict that a next strain will eventually replace the declining *Salmonella typhimurium* DT104. And with a better understanding of virulence and survival strategies and regulation thereof, it may become possible to predict the next pandemic strain even before epidemiologists start to note such a possibility. 450 455 460 465 470

In the long run, we can expect a general improvement for microbial risk assessments. If successful in relating molecular mechanisms, attention will then shift from virulence to survival and persistence in the environment and in the food chain. A better quantification of survival of a given pathogen during food processing, or enhanced recognition and prediction of virulence will make microbial risk assessments more reliable. Future databases should store information on individual genes with respect to their role in virulence (for instance regulatory, secretion, accessory factor, toxin), whether the required genetic context is present in order to be functional, whether strain-to-strain variation within the species is likely, and if so, at what frequency this virulence factor occurs. The more accurate description of genes as virulence tell-tales for particular isolates, the higher quality data can be used as input in mathematical modeling. Based on assessments making use of high quality data, risk managers will then be able to better identify risk management options to tackle the burden of food borne diseases. 475 480 485

OUTLOOK

Microbial risk assessment in the field of food-borne pathogens is rapidly evolving. Clearly, our understanding of virulence-determining factors is still incomplete. Novel

Microbial Risk Assessment and Genomics

490 strategies and limitations of existing strategies to fill the gaps have been outlined.
Complete genome sequences can be a valuable tool in the process of virulence
recognition and description. Multiple genomes per species will add valuable in-
495 formation, provided these include, when available, both more and less pathogenic
strains. Databases storing biological information about the sequenced strain in ques-
tion, linking presence of individual genes with necessary genomic background and
gene networks, and storing information on quantitative aspects of bacterial pop-
ulations, should help close the gap between genotype and phenotype. A specific
database to fulfil the needs of microbial risk assessment should be considered. When
500 we can predict virulence components and their gene networks more accurately, our
understanding of pathogens will have increased so much that we may be able to
predict emerging pathogens even before they evolve. By incorporating all available
relevant knowledge on genes, genomes, proteomes and gene networks, microbial
risk assessment will become more accurate.

ACKNOWLEDGMENTS

505 DWU is funded by grants from the Danish Research Foundation.

REFERENCES

- Alm RA, Ling LS, Moir DT, *et al.* 1999. Genomic-sequence comparison of two unrelated isolates
of the human gastric pathogen *Helicobacter pylori*. *Nature* 397:176–80
- Altschul SF, Gish W, Miller W, *et al.* 1990. Basic local alignment search tool. *J Mol Biol* 215:403–
510 10
- Azad RK and Borodovsky M. 2004. Probabilistic methods of identifying genes in prokaryotic
genomes: Connections to the HMM theory. *Brief Bioinform* 5:118–30
- Becker D, Selbach M, Rollenhagen C, *et al.* 2006. Robust *Salmonella* metabolism limits possi-
bilities for new antimicrobials. *Nature* 440:303–7
- 515 Bendtsen JD, Binnewies TT, Hallin PF, *et al.* 2005. Genome update: Prediction of secreted
proteins in 225 bacterial proteomes. *Microbiology* 151:1725–7
- Benton BM, Zhang JP, Bond S, *et al.* 2004. Large-scale identification of genes required for full
virulence of *Staphylococcus aureus*. *J Bacteriol* 186:8478–89
- Boerlin P, McEwen SA, Boerlin-Petzold F, *et al.* 1999. Associations between virulence factors of
520 Shiga toxin-producing *Escherichia coli* and disease in humans. *J Clin Microbiol* 37:497–503
- Cassin MH, Lammerding AM, Todd ECD, *et al.* 1998. Quantitative risk assessment for *Es-*
cherichia coli O157:H7 in ground beef hamburgers. *Int J Food Microbiol* 41:21–44
- Chain P, Kurtz S, Ohlebusch E, *et al.* 2003. An applications-focused review of comparative
genomics tools: Capabilities, limitations and future challenges. *Brief Bioinform* 4:105–23
- 525 Davis MA, Hancock DD, and Besser TE. 2002. Multiresistant clones of *Salmonella enterica*: The
importance of dissemination. *J Lab Clin Med* 140:135–41
- Dorman CJ. 2004. H-NS: A universal regulator for a dynamic genome. *Nat Rev Microbiol*
2:391–400
- Dougan G, Hormaeche CE, and Maskell DJ. 1987. Live oral *Salmonella* vaccines: Potential use
530 of attenuated strains as carriers of heterologous antigens to the immune system. *Parasite*
Immunol 9:151–60
- Falkow S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. *Rev Infect*
Dis 10:S274–6
- Field D and Hughes J. 2005. Cataloguing our current genome collection. *Microbiology*
535 151:1016–9

T. M. Wassenaar et al.

- Finlay BB and Falkow S. 1989. Common themes in microbial pathogenicity. *Microbiol Rev* 53:210–30
- Finlay BB and Falkow S. 1997. Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev* 61:136–69
- Fleischmann RD, Adams MD, White O, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512 **540**
- Fouts DE, Mongodin EF, Mandrell RE, et al. 2005. Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species *PLoS Biol* 3:e15
- Fraser CM, Casjens S, Huang WM, et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390:580–6 **545**
- Gaynor EC, Cawthraw S, Manning G, et al. 2004. The genome-sequenced variant of *Campylobacter jejuni* NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. *J Bacteriol* 186:503–17; Erratum in: *J Bacteriol* 2004 186:8159 **550**
- Goodner B, Hinkle G, Gattung S, et al. 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* 294:2323–8
- Jaffe JD, Stange-Thomann N, Smith C, et al. 2004. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* 14:1447–61
- Harrison A, Dyer DW, Gillaspay A, et al. 2005. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: Comparative study with *H. influenzae* serotyped, strain KW20. *J Bacteriol* 187:4627–36 **555**
- Hayashi T, Makino K, Ohnishi M, et al. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8:11–22 **560**
- Hussein HS and Sakuma T. 2005. Prevalence of shiga toxin-producing *Escherichia coli* in dairy cattle and their products. *J Dairy Sci* 88:450–5
- Jenkins TM, Scott TM, Cole JR, et al. 2004. Assessment of virulence-factor activity relationships (VFARs) for waterborne diseases. *Water Sci Technol* 50:309–14
- Kill K, Binnewies TT, Sicheritz-Ponten T, Willenbrock H, et al. 2005. Genome update: Sigma factors in 240 bacterial genomes. *Microbiology* 151:3147–50 **565**
- Kirschner D, DeRita V, and Flynn J. 2005. Overcoming math anxiety: Matthus meets Koch. *ASM News* 71:357–62
- Maurelli AT, Fernández RE, Block CA, et al. 1998. “Black holes” and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci USA* 95:3943–8 **570**
- Newell DG. 2001. Animal models of *Campylobacter jejuni* colonization and disease and the lessons to be learned from similar *Helicobacter pylori* models. *Symp Ser Soc Appl Microbiol* 30:57S–67S
- Palmer SR, Gully PR, White JM, et al. 1983. Water-borne outbreak of *Campylobacter* gastroenteritis. *Lancet* (8319):287–90 **575**
- Parkhill J, Wren BW, Mungall K, et al. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* 403:665–8
- Perna NT, Plunkett G, Burland V, et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–33. Erratum in: *Nature* 2001, 410:240 **580**
- Pichon C and Felden B. 2005. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc Natl Acad Sci USA* 102:14249–54
- Roberts A, Chan Y, and Wiedmann M. 2005. Definition of genetically distinct attenuation mechanisms in naturally virulence-attenuated *Listeria monocytogenes* by comparative cell culture and molecular characterization. *Appl Environ Microbiol* 71:3900–10 **585**

Microbial Risk Assessment and Genomics

- Rosenquist H, Nielsen NL, Sommer HM, *et al.* 2003. Quantitative risk assessment of human campylobacteriosis associated with thermophilic *Campylobacter* species in chickens. *Int J Food Microbiol* 83:87–103
- 590 Schattner P, Brooks AN, and Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 1(33)[Web Server issue]:W686–9
- Schlundt J. 2000. Comparison of microbial risk assessment studies published. *Int J Food Microbiol* 58:197–202
- 595 Tomb J-F, White O, Kerlavage AR, *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388:539–47
- Ussery DW and Hallin PF. 2004. Genome update: Annotation quality in sequenced microbial genomes. *Microbiology* 150:2015–7
- Wassenaar TM. 2004. Risk assessment prediction from genome sequences, promises and dreams. *J Food Protect* 67:2053–7
- 600 Wassenaar TM and Gaastra W. 2001. Bacterial virulence, where to draw a line? *FEMS Microbiol Lett* 201:1–7
- Wassenaar TM and Meinersmann RJ. 2003. The TGA stop codon and the phylogeny of the selenocysteine pathway. *Genome Lett* 2:127–38
- 605 Welch RA, Burland V, Plunkett G, *et al.* 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 99:17020–4
- Wong HC, Shen CT, Chang CN, *et al.* 2004. Biochemical and virulence characterization of viable but nonculturable cells of *Vibrio parahaemolyticus*. *J Food Prot* 67:2430–5