

# Proceedings of The Institute of Food Technologists' First Annual Food Protection & Defense Research Conference

November 3-4, 2005  
Atlanta, Georgia

[Session: **Detection and Diagnostics**]

## DNA Detection Strategies: Sequences, Signatures and Significance

DR. THOMAS CEBULA

CENTER FOR FOOD SAFETY AND APPLIED NUTRITION, FOOD AND DRUG ADMINISTRATION

Well, I'll start off by saying that Jason Gans is ill, and I am fast fading. So I'm hoping to get through this talk with a voice. I want to thank the organizers for the opportunity to speak with you today, and I must say, I am looking forward to this session.

For Jason, of course, you now know that you'll have to wait and hear about his signatures sequences another time. When you look at the range of topics that is being covered in this short period—from avian flu to PCR detection using mass spec—I think we're in for a little bit of a ride this morning.

What I'd like to do with my time is share some of my thoughts provoked by interactions at this meeting, and then try to put them together in some sort of coherent fashion. I do believe Winnie the Pooh's instructions hold that — "Before beginning a Hunt, it is wise to ask someone what you are looking for before you begin looking for it" (A.A. Milne, 1926, Pooh's Little Instruction Book).

In the area of microbial forensics – my topic for today – we look to identify and recognize patterns in a disease outbreak; communicate those patterns to the public health community at large; determine the pathogen involved; contain the outbreak and trace the microbe to its source.

Now for the epidemiologists in the audience like Craig Hedburg — they are saying that this is what we've done for years, and I agree that there is a significant overlap. The difference between microbial forensics and epidemiology can be seen once we introduce Bhagwan Shree Rajneesh to you. He's the one on the left of that dynamic duo. Some of his followers laced salad bars in Oregon with *Salmonella*. And, epidemiologists indeed traced the outbreak of Salmonella cases to those salad bars. They wrote up their investigation, they published their work in peer review journals, and, 2 years later, when one of Bhagwan's followers confessed, we found a bio-crime had been committed. So, the difference between epidemiology and microbial forensics is the assumption, at the outset, that a bio-crime indeed occurred and distinguishes microbial forensics as an emerging discipline.

In a way of talking about the forensics continuum for strain identification – again I'm going to use Craig Hedberg – we had an interesting sidebar, discussing the last few days what is necessary for detection. And I think this fits another sidebar that I had with Mary Torrence. That is, if a technique works and gives you all of the information that you need, there is no need to look for another technique. To paraphrase Craig, "sometimes the method can be very blunt; it doesn't have to be sharp edged". If it captures everything you want then use it.

Again, I use this continuum to say if you wish to differentiate two

bacterial strains, then the "N" that you must consider is the number of biomarkers that says one strain is different from another. In all cases you would like the methods to be valid, and the biomarkers that you are using should be stable in order to have the science on your side. However, when you are saying that a strain is absolutely identical to another then the N value is very different. N is represented by the genomic diversity that exists for that species. And, when we talk of microbial diversity — there are greater than  $6 \times 10^{30}$  microorganisms sitting out there in our environment — now that's a large number. I trained as a chemist and Avogadro's number always fascinated me as being an extremely large number. A mole of anything is large; well you are talking about a million moles of microbes scattered about in our environment.

We also heard during the last few days the integration of food safety with food defense. For food safety, we were told by Mel Bernstein, that we have things under control. Perhaps, a more refined version of Mel's statement is more accurate, that is, we think that we are well on our way to having it under control. Again I point out that there are some fundamental differences between food safety and food defense. Usually when we are tracing an organism for food safety reasons, it usually stops at best at the serotype or serovar level. However, for attribution in food defense, detection is at the individual strain level.

So again to emphasize the sidebar that Craig and I had, when you are talking about microbial forensics, the techniques may have to be a bit more acute and sharp-edged than the blunter edged sabers that might be used in food safety.

Again a recurring theme is bacterial diversity. I say it's large and, when we begin to study diversity, the first question we have to raise is; whose strains define the universe of diversity that we study? That is, we usually go to someone's strain collection, use a subset, and say that we have captured diversity. I submit to you that we have just scratched the surface. I think there is a genuine lack of appreciation concerning the extent of diversity that exists among plant and animal pathogens. That is going to come back to haunt us if we don't pay attention to it quickly.

I really just want to go back a few years – 8 years ago now; it seems like yesterday when Gene LeClerc and I wrote somewhat of an indictment of the stage of microbial genomics in 1997. We started by saying, "at this writing with the complete sequence of 4 bacterial genomes already known and a 5th, that of *E. coli* K12 to be unveiled shortly, some still myopically question whether bacterial genomics will offer many surprises".

The Salmonella sequencing project has been impacted, – we were nice, it actually had been tabled – hampered by the belief that *Salmonella* was too much like *E. coli* to warrant intense effort. This apathy seems steeped in the naïve assumption that experiments conducted in the unnatural setting, the test tube, can be correlated directly with how bacteria behave in their natural environment.

Clearly, we didn't share in this belief and fortunately, a lot of people in this room didn't share in that belief either. Obviously things have changed. So now we flash forward to June 2nd, 2005; in planning a talk I went to [www.genomesonline.org](http://www.genomesonline.org) and found that there were 265 totally sequenced and annotated microbial genomes. A few weeks later, September 16th, there were about 300 microbial genomes sequenced and annotated. A few more days later, 303 and you can see the pace between October 20th and October 25th, ten genomes in about 6 d, completely sequenced and annotated. Now during this time, a new sequencing technology was introduced. 454 technology is a non-Sanger sequencing technology; it is based on real time pyrosequencing. I'll mention more about pyrosequencing a bit later.

Now I mention new technology for a couple of reasons. Recently, at a small genomes conference in Wisconsin, it was announced that a 2nd *E. coli* K12 genome had been published. A group from Japan sequenced strain W3110, and Fred Blattner and his group at the Univ. of Wisconsin had sequenced strain MG1655 originally. So this gave an opportunity to check for sequencing errors in the original MG1655 genome. Now the Wisconsin group used a series of technologies – radioactive gels, plate gels, and capillary gel electrophoresis – to determine the sequence of strain MG1655. The Japanese demonstrated nicely that there were sequencing errors introduced by Fred Blattner and crew. In fact there were 4 sequencing errors in the entire genome; that is, 4 sequencing errors in about 4.5 million base pairs of DNA. So it's a wonderful testimony of the carefulness that went into the original *E. coli* sequencing project.

Things changed, however, when in 1998 the Dept. of Energy announced that its sequencing program would only provide high-coverage draft sequences, rather than complete genomes with annotation, for organisms of interest. When you see the rapid progress that we are making in bacterial genomes and the new technologies being introduced, one must ask again about error rates in sequencing. Draft sequence coverage and 454 technology move sequencing forward quickly but introduce errors at rates of  $1 \times 10^{-3}$  to  $1 \times 10^{-4}$ , while in a completed genome with annotation, the error rate is more likely to be about  $1 \times 10^{-5}$ . So whereas the old technology gave sequencing error on the order of one in a million, we have given up some of the accuracy so that we can have more sequence available.

I also want to point out some interesting findings by Claire Fraser and her group at TIGR. A paper that published in PNAS at the end of September described eight different genomic sequences of strains of *Streptococcus agalactiae*. They found that every time a new strain of *Streptococcus* was sequenced, on average, 33 novel, new genes inserted into the chromosome. The genome stayed the same size so that meant, as new sequences inserted, other sequences were deleted in each strain. This gave rise to a concept of an open pan-genome for *Streptococcus agalactiae*. This means we will never arrive at the total genome; we will keep on adding to our knowledge but we will never get to the ultimate genome, hence the open pan-genome. They also state in the same article the sequencing of four genomes of *Bacillus anthracis* captures everything about that organism, and this they termed a closed pan-genome.

One must consider that the diversity of *Streptococcus agalactiae* and the diverse clinical collections that exist for *Streptococcus* gave rise to the open pan-genome concept while more limited bacterial collections of *B. anthracis* gave rise to the closed pan-genome concept. One has to wonder if the closed pan-genome for *B. anthracis* is due

to an evolutionary bottleneck or if it may reflect a more limited sampling of *B. anthracis* diversity. Again just to comment on sequencing errors, Claire Fraser also talked about the value of complete genome sequencing and emphasized that “you get what you pay for”. So if you look at draft consensus, you may have to accept a lot of sequencing errors. I keep on emphasizing this because as we develop new technology that relies on *in silico* approaches, that is, comparing what is in the current data base, it is our responsibility to ensure the laundering out of sequencing errors that exist within the data base.

So for example, though I won't have time to discuss in detail, when we examined *E. coli* O157:H7 over a very small stretch, that is, about one percent of the genome, we found roughly 45 sequencing errors; these data were sent to the Wisconsin group to help sanitize the published O157:H7 genome sequence. The reason we are so interested in using sequencing is that it is very useful in showing clandestine relationships that occur between bacterial strains.

My laboratory studies evolution of pathogens, and we also study how particular signatures can help discern between and among bacterial strains. I am showing this slide to emphasize a couple of things. When we examine *Salmonella*, there are basically eight subgroups that are comprised of over 2,300 serovars. The diversity wrapped in the cloud of group one. *Salmonella* is comparable to that of the diversity for the 150 or so serotypes of *E. coli*. Within the group one *Salmonella* are contained greater than 95% of all *Salmonella* strains that cause human infection. I also want to emphasize that just like in eukaryotes, the role of recombination is finally being recognized in bacteria. The time of the classical clonal theory for bacteria is now long passed. There is reason to suspect that homeologous recombination and homologous recombination are framing evolution of pathogens.

As you might suspect for the salmonellae this gave us the opportunity to look at very closely related strains where we could show that recombination occurred much more frequently than more distant organisms. So what this tells you is that for recombination to occur you need two bacteria sitting in the same niche, emphasizing that both physical (geographical) and genetic characteristics underpin recombination in bacteria.

Turning to whole genome arrays, I want to emphasize some work being done by Michael McClelland and colleagues at the Sidney Kimmel Cancer Institute. We helped with the construct of non-redundant genomic arrays of all open reading frames for *S. typhimurium*, *S. typhi*, and *S. enteritidis*, but this now has been extended by McClelland and coworkers to include five different serovars of *Salmonella*. The unique open reading frames have been placed on one array as a genome typing method to ask how many genomovars and genovars are contained within the salmonellae. I told you that there are roughly 2300 serovars, but there are probably in excess of 6500 genovars. That is salmonellae are out there – some wearing the same coat with entirely different genomes contained within, and some wearing different coats but with nearly the same genome inside them. So the techniques developed in yesteryear, for example, serotyping, may not be the techniques that we wish to use when we try to attribute particular strains of bacteria to a bio-crime. Now again picking up on Mary Torrence's theme for her talk – “we have the toys but where is the playground or what playground” – I believe, like Craig and like Mary, that finding a use for a method is definitely not synonymous with finding a method that's useful. So if you have a method that's useful by all means use it; if it's microbiological, use it, if it's blunt, use it.

I would like to tell you about 3 techniques that we use in a triaging strategy to get information at the genome level for individual bacterial strains. I'm going to use *E. coli* O157:H7 as example. *E. coli* O157:H7, unlike *E. coli* K12, has roughly an extra million base pairs of information. Instead of having a 4.5 MB genome, its genome size is 5.5 MB, indicating that not all *E. coli* are created equally. We randomly sampled

that genome by dividing the chromosome into 60 segments and took 1000 base pair sequences from each of those segments. So we have snippets of the DNA and what we are going to do, based on a reference sequenced genome is make a series of targets. The targets are each 29 bases and are designed to tile each of the 1000 bp segments. The targets are contiguous and overlap each other by 24 base pairs, thus leaving a 5 base pair gap. This design allows one to interrogate 60,000 base pairs, and each base pair will be interrogated with 5 of the 29 base targets.

Working with NimbleGen (Madison, Wis.), we designed a semi-high throughput array, a 12 well array that allows interrogation of 11 test strains and a reference strain. In each well, measuring about 4 millimeters in diameter, there are 14,000 targets, this density allows ready detection of single nucleotide polymorphisms (SNPs).

After appropriate fluorescent labeling and hybridization, the ratios of fluorescent intensities of the test versus reference. We anticipated that very closely related strains, more than likely would show a one-to-one correspondence, that is, have a ratio of about one for the hybridization signal, signaling no difference in sequence.

There will be regions of new material, the so-called pathogenicity islands that we've all read about in O157 – insertions or deletions that will show up as discrepancies in the hybridization ratio. Most importantly, single base pair polymorphisms, SNPs, will be signaled by three to five targets on the array. So this is the principle of tiling arrays, and from a one-percent sampling, we ask if we can differentiate *E. coli* O157:H7 strains. I am just going to show you the data for five strains, but I can tell you from analyses of about 100 O157:H7 isolates, each up to now has shown a distinct pattern. Thus, without identifying the precise differences, by pattern recognition alone, we can distinguish one strain of O157:H7 from another.

I should have said all strains but one. As this analysis was done in a semi-blinded fashion, one strain did not show any differences. Upon breaking the code, we found that we were examining EDL933, the strain which served as our reference strain. Our isolate of EDL933 was obtained from CDC more than a decade ago, and, though it rested in our freezers over that period, it gave a very stable signature.

Though we are localizing the differences, we have not, as yet, identified all differences. When we wish to identify SNPs, pyrosequencing is utilized. I won't have time to describe the technique in detail, but basically it is a 4-enzyme reaction ultimately giving rise to flashes of light that are captured by a CCD camera. Whenever the correct deoxyribonucleotide is incorporated into the extending strand of DNA, a pyrophosphate molecule is split off.

That pyrophosphate is exchanged with a non-hydrolysable ATP molecule, now making it hydrolysable. In the presence of ATP, luciferin and luciferase, light is emitted. As light is emitted in quanta, this becomes a precise, quantitative assay. For example, if TTP yields a signal that is twice the peak height that means that the DNA strand has been extended by incorporation of two thymidines. This allows one to sequence small runs of about 100 or so bases in about 10 min. Thus, once you've localized the SNP to within five nucleotides by tiling, you then can use pyrosequencing to identify the SNP. This allows us to pick readily a number of SNP regions and begin to build bar codes for various strains of O157:H7, in essence, binning like O157s with each other and differentiating them from other isolates of O157:H7.

Ultimately, we use SNPs in a cladistics analysis, as we have done in our evolution studies, to build trees. We know, however, that some markers are not useful in building these trees (because of homoplasy). For a gene like *roi*, which is a phage gene moving between and among strains, if we were to incorporate it in building a tree, it would collapse the tree basically making it look like a shrub. Rather we use *roi* to decorate the tree. By decorating the tree then you can now separate

those bins still further such that, with roughly 3,000 base pairs of a 5.5 million base pair genome, we begin seeing the individual diversity among O157:H7 strains. By combining signals forged by mutation and recombination into your arsenal, you can, with very little sequence information, begin to differentiate bacteria at the individual strain level.

Again it's important to consider what you use as biomarkers. As I said, we study bacterial evolution and, by-and-large, we look for what is termed synapomorphies. Synapomorphies are base pair changes that define clades; the nodes defining the branches of those trees. And, of course, we are discouraged at looking for autapomorphies because homoplasy can distort the tree and confound tracing the lineage of a bacterial species. However when you're looking for resolution at the individual strain level, autapomorphies are essential. The *roi* gene, for example, becomes very important as a discriminator. So again it is all what you are looking for. If you wish to study evolution the hunt for synapomorphies is important; if you are trying to characterize individual strains, begin to step out and look for autapomorphs.

Finally, this was prompted by a Mel Bernstein comment; I wish I had his "spectrum of threats" slide here, but he said pathogens of food safety interest are a low threat, but a bioengineered strain is a high threat. How would you go about finding a bioengineered strain? Again, I think O157:H7 serves as a good example. I told you that O157:H7 has one Mb more DNA than does some of its siblings, and that this extra DNA is deposited at 200 different sites around the genome. So we don't really talk about path islands any longer, these are truly archipelagos. Again I would remind everyone that Mother Nature is the quintessential terrorist. She has been manipulating these strains for eons and we, at best, for a couple of decades. So by mapping the docking sites where Mother Nature has allowed new information to be incorporated, we can monitor whether a strain is engineered or not.

I emphasize the use of optical mapping as example. Everyone in this room has heard of pulsed field gel electrophoresis; optical mapping is pulsed field gel electrophoresis without the gel. Optical mapping data were developed in collaboration with OpGen (Madison, Wis.). Closed-circular DNA is spread onto a specially etched microscope slide, allowed to dry, and then restricted with your enzyme of choice. In our case, we use BamH1, which cuts at discrete sites within the DNA. The fragments are labeled with a fluorescent dye, YOYO1 in our case, and then scanned with a fluorescence microscope. Computer software is used to reconstruct all of the fragments to develop a consensus map for that genome.

So in this way a consensus map is derived. Iterations of about 50 reads or more may be used to develop a consensus map for that sequence and again, based on a known sequence strain you can begin to compare various strains whose sequence are not known. I don't know how this projects but I tried to use this cyan blue as a marker to say these are regions that are the same in all of these strains and the adjacent greens are also the same. However these white portions are now insertions that are unique for a particular strain. The software allows you to mark these as regions of difference. Here is a strain, for example, where 20% of the genome is inverted. Much like we do with our tiling arrays we use pattern recognition in our optical mapping to distinguish O157 strains. As you can see, the patterns are sufficiently different to distinguish one strain from another.

The message I leave you with – the conclusion is very simple. For forensics to work you are going to need more than one tool. As Ben Franklin said, "don't think to hunt two hares with but one dog". The techniques that are used must be validated, and they have to be in our arsenal long before an emergency. When an emergency occurs its going to be too late to begin to invent the technology and importantly, if the technology is going to work, it's going to have to work in a time-efficient manner because waiting for the right result years after the fact doesn't do anyone any good.