Proposal for the Bioeconomy Blueprint   December 6, 2011

Division of Foodborne, Waterborne and Environmental Diseases
National Center for Emerging and Zoonotic Infectious Diseases
Centers for Disease Control and Prevention

Contact:
Christopher R. Braden, M.D.
Director, Div. Foodborne, Waterborne and Environmental Diseases
National Center for Emerging and Zoonotic Infectious Diseases
Centers for Disease Control and Prevention

Executive Summary

Safer Food and Improved Patient Diagnosis through an Advanced Nation-wide Genomic-based Enteric Disease Detection and Investigation Network. Foodborne disease is a large, costly, and under-recognized health challenge, but one that has the potential to be significantly addressed by biotechnological innovation. PulseNet and the Nation's foodborne disease surveillance system have been among the most successful new drivers of food safety in a generation, and play a pivotal role in our increasingly globalized food production and distribution system. Current successes likely represent only a small fraction of the true potential of these networks. This system depends on the nation's clinical laboratories for samples from ill individuals, which are compared by "DNA fingerprinting" techniques to find patterns of disease that may signify an unrecognized problem in our food or water supplies. However, new clinical diagnostic technology is being introduced which will make private and commercial testing activities incompatible with public sector disease-tracking, thus threatening our ability to detect and respond to outbreaks. Development of next-generation genomic and metagenomic technologies and analytic capacity will give the nation an opportunity to sync private sector and public health disease detection efforts, improve the predictive value of diagnostics, and will likely allow the nation's food safety system to more fully exploit the rich resource that foodborne disease surveillance represents. Furthermore, software and hardware developed to address this issue, coupled with publicly-available data from a large-scale metagenomic infrastructure will leverage development across a range of biotechnology, pharmaceutical and diagnostic manufacturers.

1) **Grand Challenge:** Safer Food and Improved Patient Diagnosis through an Advanced Nation-wide Metagenomic-based Enteric Disease Detection and Investigation Network

   Background
   National foodborne disease tracking networks have a large impact on disease: Each year approximately 48,000,000 people in the U.S., or 1 in 6 Americans, contract a foodborne disease, 128,000 are hospitalized and 3,000 die.   Unlike many other conditions, foodborne

disease is largely preventable, and for that reason has been designated as a "winnable battle" by the CDC. Disease surveillance has proven to be one of the most robust mechanisms available for detecting problems in the food and water supplies that are not recognized through the normal regulatory processes, and one of the most straightforward and cost-effective means of improving food safety. New tools, such as PulseNet, have significantly improved our ability to address the problem.

PulseNet, the nation's molecular subtyping network for enteric disease, has played a key role in virtually every high-profile foodborne disease reported in the national media since its inception in 1996. PulseNet obtains a "DNA fingerprint" from pathogens isolated from individual patients and matches the fingerprints to detect local, national, or international clusters of disease that may represent wide-spread outbreaks. It is comprised of 87 federal, state, and local laboratories in the U.S., and 82 countries around the world. Together with OutbreakNet, the national network of state and local officials who investigate foodborne outbreaks, over one half billion pounds of contaminated food have been recalled from the U.S. marketplace in PulseNet's first 15 years. More importantly, PulseNet-triggered investigations have stimulated improved safety practices in a wide range of food industries, such as the beef, poultry, shell egg, ready-to-eat meat and prepared meal, leafy green and vine vegetable, sprout, melon, flour, spice, tree nut, and peanut industries. In addition to prompting self-improvement by industry, these investigations have provided FDA and USDA the information they need to focus their activities on the products most likely to cause disease, both in domestic and imported products. Outbreaks can now be detected more quickly than ever before, making it less likely that American's will experience the type of widespread *E. coli* outbreak that devastated Germany and the rest of Europe in 2011. While it may seem counter intuitive, improved outbreak detection and investigation improves confidence in U.S. industries in the long run, and improves the quality-of-life for every American.

Need for new technology:
The current National Foodborne Outbreak Detection System is not in sync with private development, which threatens our Country's ability to rapidly detect food-related outbreaks: The ultimate source of all information for foodborne disease surveillance is obtained from diagnostic testing and exposure assessment of ill patients. A new generation diagnostic testing that does not depend on culture and isolation of pathogenic bacteria is being developed by the private sector at a rapid pace, with several kits already in the medical device approval pipeline. While these "non-culture" tests may improve service to individual patients, they do not provide the isolates which are needed for public health surveillance programs such as PulseNet. Furthermore, some non-culture tests require specimens that are incompatible with culture, making reflex culture (culture of positive specimens) impossible. This deficit will weaken critical foodborne disease surveillance programs that depend on

isolates, including PulseNet, OutbreakNet, FoodCORE, the Foodborne Diseases Active Surveillance Network (FoodNet) and the National Antimicrobial Resistance Monitoring System (NARMS), which will result in vulnerabilities and gaps in the food safety system. Programs at FDA and USDA which depend on CDC data will also be negatively affected, such as FDA CORE, and the USDA's predictive analytics program. If diagnostic development eventually moves to completely preclude primary culture, major unintended consequences could include hundreds of thousands of unnecessary illnesses resulting from the inability of our surveillance systems to rapidly and reliably detect outbreaks, and reduced information needed for industry to produce safe food. New "culture-independent" molecular strain typing and characterization methods need to be developed for the public sector that are compatible with, or complement, tests developed in the private sector so that the needs of both patients and the general public are met.

Foodborne disease surveillance system has vast largely untapped potential: As successful as our foodborne disease surveillance and investigation programs have been, they are likely operating at only a small fraction of their true potential. A relatively small percentage of PulseNet clusters are solved in most states, leading to a loss of many prevention opportunities.. The basic elements needed to make the system more effective both for patients and for the general public are well known. They involve getting better and faster information about what made individual cases sick, and more effectively detecting and investigating identified clusters of disease.

The current system relies on many steps that occur in a more-or-less sequential manner, from the patient presenting to their physician; the physician ordering a test; reporting of the positive test result to public health authorities; shipping the isolate to a public health laboratory; performing molecular subtyping tests; uploading of "DNA fingerprints" to local and national databases; detecting and investigating clusters of illness; interviewing patients for food and other exposures; and finally follow-up by health, regulatory, and industry partners. The longer the process takes, the lower the probability that the investigations will be successful because patients quickly forget their activities and what they ate, contaminated foods may be completely consumed or disposed of by the time they are identified as a probable vehicle, and the trail of investigation clues quickly grows cold.

PulseNet addresses a small but important fraction of foodborne disease cases. Currently tracked microorganisms represent less than three percent of all foodborne disease infections, meaning that approximately ninety-seven percent of this valuable information remains to be exploited. Over eighty percent of infections are thought to be due to unknown etiologic agents. Identifying the bacteria, viruses, and parasites that are currently unknown but nevertheless making individuals sick is a particularly productive means toward devising new prevention strategies.

The solutions to making the system dramatically more effective are technological (e.g. getting molecular epidemiology information closer to patient diagnosis) and process-dependent (e.g. obtaining high-quality standardized exposure information).   These needs were recognized in the "FDA Food Safety Modernization Act of 2010" (S 510) and the "Improving Food-borne Illness Surveillance and Response Act of 2008" (S. 3358).

Harnessing advanced biological technology to meet the challenge

Genomic methods: Whole microbial genomes can now be sequenced in days or hours, instead of weeks or months that were required just a few years ago. Bioinformatic analysis is now the rate-limiting step (see "critical challenges" below). Currently, the most effective means of sequencing is by the use of microbial "isolates." Genomic data can be used for:

(1) Tracking microbes in the population. Any level or multiple levels of strain resolution needed for outbreak detection or investigation can be generated with whole genome sequence data. Genomic methods may be backwards compatible with current methods through the use of *in silico* analyses. If done rapidly, the routine use of genomic data to compare human disease with environmental, food, and animal testing can help quickly identify what foods or other exposures are making people ill, without many of the limitations of current fragment-based molecular surveillance techniques.
(2) Identifying new targets for next-generation PulseNet methods, which will allow the identification of specific microbial strains and important virulence factors in complex samples, such as human stool. When combined with industry-developed platforms, this has the potential of revolutionizing the commercial diagnostic market.

Metagenomic methods: It is now theoretically possible to sequence all members of a microbial community, such as human stool. This is currently the "holy grail" of enteric diagnostics. Unlike the" holy grail", building national capacity to make metagenomics a routine diagnostic and public health tool is an attainable goal. However, it is not yet practical to analyze metagenomics data in a timely manner (see "critical challenges" below).

Application of genomic and metagenomics sequencing methods: These methods have the potential to:

- Rapidly identify any known microbial pathogen, or combination of pathogens. This "Star-Trek" type technology was only science fiction only a few years ago, but is rapidly becoming technically feasible. Preliminary data suggests that many enteric illnesses are due to combinations of pathogens, or pathogens in the presence of other bacteria or specific virulence factors. There is currently no practical way to identify these effects, either for patient diagnosis or public health action.

- Produce sequences for unknown microbial pathogens in any patient specimen, which can be linked and evaluated through outbreak investigations or large sporadic case control studies
- Preserve and improve PulseNet, OutbreakNet, FoodCORE, and our National Foodborne Disease Surveillance System. Bringing microbial strain tracking closer to the ill patient has the potential to greatly improve the speed and effectiveness of outbreak detection and investigation, especially when coupled with electronic cluster detection and national standardized exposure assessments. The improved resolution of sequencing compared to current subtyping methods will also enable us to identify microbial pathogens associated with particular food commodities and thereby to predict which food commodities are associated with sporadic foodborne illness (attribution). This information may also be used to target food safety intervention at the production level
- Identify food consumed by ill individuals. This may be very useful for illness with short incubation times, such as foodborne intoxications or toxin-mediated infections
- Identify microbial factors that contribute to the development of disease.
- Identify host factors contributing to disease.
- Increase the precision of strain assignments.

2) **High impact research, innovative opportunities, and Federal priorities**
   a. **Direct impact:** For the same reasons that foodborne disease has been designated as a "winnable battle" by the CDC, this research will likely have a relatively rapid impact on a very large problem. An estimate by the Pew Trust places the total cost of foodborne disease to the U.S. economy at $152 billion dollars per year, with a high proportion of expenditures required for healthcare. As described above, foodborne disease is largely preventable, but prevention requires that we better understand the reasons causing 48,000,000 Americans to become sick each year. Improved outbreak detection represents one of our best opportunities for reducing the burden of disease by stimulating change in industry and focusing regulatory activities. It also is expected to reduce impact of outbreaks on industry, as advanced technology and epidemiology should reduce the time needed to identify single farms or producers responsible for contaminated products, thus protecting safe producers of the same commodity.
   b. **Broad application:** Stool represents one of the most complex possible matrices, with over 10^{11} microorganisms/ml and more than 500 microbial species. PulseNet currently requires 50,000 – 60,000 analyses per year. Extending metagenomics to clinical diagnostics would be orders of magnitude more extensive. Development of the data pipeline and analytical tools needed to conduct large-scale metagenomics analyses is a bit akin to development of space technology. The tools needed to complete the task are broadly applicable to other national priorities

c. **International competitiveness:** A number of foreign countries including Canada and Denmark have made significant investment in bioinformatics infrastructure, at least in part to address enteric disease issues. Bioinformatics appears to be the rate-limiting step in the next phase of biotechnology development.

3) **Technical challenges**

a. **Genomic sequencing technology development for metagenomics:** In current sequencing technologies all DNA in a sample is sequenced in small bits and then assembled using specialized software. However, many genomic targets are present outside the chromosome, e.g., on extrachromosomal elements like plasmids, or are found in many different bacteria, both pathogens and commensals. Such targets may encode important virulence factors and antimicrobial resistance. In order to use metagenomics analysis for diagnostic and surveillance purposes it is critical to know exactly to which organism such genomic targets belong. It is therefore critical to develop high throughput technology to sequence all diverse members of the microbial population in a complex sample, e.g., stool, on cell at a time to ensure that all sequencing information from potential pathogens are captured reliably.

b. **Computing hardware, data pipeline, and software:** A whole genome sequence of a stool pathogen may contain 4,000,000 – 6,000,000 DNA base pairs. Depending on the sequencing assay, a single metagenomic analysis of stool may involve 500 times as much information. If one factors in depth of coverage, the data burden is more than an order of magnitude higher. Currently, analysis of a single stool sample can take weeks or months, and requires dedicated, high performance computing equipment that is not widely available. Even the networking infrastructure for transferring such large volumes of data within and between institutions is inadequate, and the available software is still under active development and not amenable to routine use.

c. **Research priorities to address challenges: Since our experience with whole genome and metagenomic sequence analysis is relatively new, it is still largely necessary to manipulate huge amounts of data to identify the information of interest.** Research involving industry, academia, and government could focus on streamlining analytical algorithms to reduce the amount of data that must be manipulated. Considerable research will be required to interpret metagenomic data. It will be necessary to examine known and unknown pathogens and pathogen combinations as epidemiological risk factors for disease, and to develop strategies to discount or discard data from organisms that are likely to be part of normal stool flora.

d. **Simple, concrete goals:**
   i. **Next-generation PulseNet method**: Development of a new method for detecting and high-resolution characterizing single pathogens such as

Salmonella spp, Shiga toxin-producing *E. coli*, and *Listeria monocytogenes* directly from a clinical stool sample in one day for $50.

   ii. **Stool metagenomics:** Stool metagenomic analysis in one day for $100. While ambitious, this is achievable in a few years if advances continue at their current pace, and if bioinformatics deficits are addressed.

4) **Prediction of gene product function through genomics:** Virulence factors for most bacteria are poorly understood, which makes for diagnostic and regulatory uncertainty. For example, the USDA has a zero tolerance for *Listeria monocytogenes* in ready-to-eat meats, but it not clear that all *Listeria spp* identified as *L. monocytogenes* are equally dangerous. Similarly, Shiga-toxin producing *E. coli* (STEC) are a diverse group of organisms found frequently in cattle and beef, but STECs are not equally likely to produce disease. Identification of additional virulence factors and their coding regions will be facilitated by analysis of large numbers of genomes coupled with epidemiological and medical information gathered through routine surveillance activities.

5) **Barriers preventing discoveries from moving from the lab to commercial markets:** The most effective surveillance system is one where analysis occurs as close to patient diagnosis as possible. That necessitates collaboration with private industry, which has primary responsibility for diagnostic test development. Components of an advanced analytical platform will need to be produced by private industry. Barriers should be minimal.

6) **Changes to SBIR and STTR programs required for accelerated commercialization:** not addressed in this document.

7) **High value data:** Large databases of genomic and metagenomic data will be an invaluable resource for commercial development. Currently, plans are underway to make all genomic data from enteric bacteria accessible to interested entities. Software developed as part of government-funded research will model development for applications in the private sector.

8) **Challenges with existing private-sector models**: Despite its importance, foodborne disease surveillance is a small market for commercial development, which has been a contributing factor to other data handling issues in the public sector. Furthermore, testing is not reimbursable as it benefits the general public but not the individual patient whose sample is tested. Development of metagenomic tools and capacity is a large project without immediate prospects of financial reward. As with NASA activities, commercial spin-offs utilizing traditional models should be plentiful.

9 - 17   Not applicable to this RFI