Battelle Approach to Fun GCAT

Battelle Memorial Institute Columbus, Ohio 43201
Applied Genomics and Biology
Craig Bartling (bartlingc@battelle.org), Gene Godbold (godboldg@battelle.org)

Capabilities and Qualifications/ Areas of Interest

Biocuration

Virulence Factor = Threatening Sequence? Getting the biology right. We've spent eight years collecting threat sequences from the literature and classifying them according to their biological properties, including how they interact with host factors. Just because a researcher calls something a virulence factor, doesn't mean it is necessarily something that one has to worry about. Many call bacterial siderophores virulence factors. Threat sequences can be found in all sorts of organisms, not all of which cause disease.

Virulopedia. We have documented more than 850 sequences of concern, these include mammalian bioregulators as well as protein toxins. These have been collected from more than 95 virus types, 75 species of bacteria, 12 eukaryotic pathogens, and snakes, spiders, jellyfish, etc. These are categorized according to enzymatic activity, host interaction partners, mode of cell entry, and structure. All assertions are linked to citations from the professional literature. More than 4000 (full) texts currently support the dataset.

Criteriome™ and our biocuration pipeline. Expanding an existing ontology for antibiotic resistance, we used three independent sets of antibiotic resistant sequences and custom software tools to annotate the set of ~260,000 antibiotic resistant sequences contained in GenBank, assigning them to one of 3,400 different sequence types which are defined using citations from the literature. The use case is screening genomic sequence data for public health biosurveillance to assign drug resistance potential.

Risk Assessment

We use probabilistic risk assessment (PRA) to define relative risks. Consequence models are based on data describing the characteristics of biological agents and their effect on exposed individuals. The data are derived from USG reports and the professional literature. When data are not available, values can be assumed based on nearest neighbor and surrogate agents. Additionally, we have developed PRIA™ (Probabilistic Risk Informed Analysis), a state-of-the-art system developed specifically to address the food-safety concerns of poultry and other meat processors. Rather than relying on inefficient and limited spreadsheet-based methods, PRIA arms the user with intuitive software-based tools to evaluate risk at each stage of processing.

Data Analytics

We use data analytics to solve a variety of problems. For example, using data analytics and deep learning tools, we have developed the following to solve complex problems: our SmartVision™ approach combines human subject matter expertise and advanced analytical methods developed for national security and defense to detect patterns in vast amounts of unstructured data; our Sematrix™ system rapidly collects, processes and accurately identifies granular scientific and technical knowledge, and stores data as linked axioms in a semantic knowledgebase to identify and track targets (e.g. scientists, domains, topics), knowledge gaps, biases and tendencies and discover tacit or hidden patterns of relationships between scientists, organizations, and research activities. Advancing our data analytics platform to meet challenges in predicting outcomes in toxicology, we recently applied multiple regression random forest algorithms and related machine learning techniques to predict the cytotoxic potential of environmental toxicants and associated gene expression and mutational biomarkers of cellular exposure.

Our potential approach

- Update threat sequence dataset from literature
- Generate threat sequence ontology
- Annotate sequence types with ontology categories
- Classify sequence threat matrix with BTRA-type approach
- Analyze sequence types statistically to produce threat signature
- Recursive testing of threat signatures against existing public sequence sets to assess viability of models for avoidance of false positives

Looking for expertise…

- Protein folding—predicting higher orders of structure from primary sequence
- Advanced bioinformatics—alignment, sequence matching, k-mer analysis of motifs, domains, SNPs, etc.
- Computationally efficient searches.
- Large functional genomics datasets—relating gene sequences to function and interactions

About Battelle

Battelle is the world's largest nonprofit research and development organization, with over 22,000 employees at more than 60 locations globally. A 501(c)(3) charitable trust, Battelle was founded on industrialist Gordon Battelle's vision that business and scientific interests can go hand-in-hand as forces for positive change. Our interests broadly range from energy and environment to health and analytics to national security. www.battelle.org

Contact

Gene Godbold, Ph.D.
godboldg@battelle.org
434-951-2116

Craig Bartling, Ph.D.
bartlingc@battelle.org
614-424-5377