Journal of
Data Mining in Genomics & Proteomics

Daefler, J Data Mining Genomics Proteomics 2013, 4:1
DOI: 10.4172/2153-0602.1000e103

Editorial Open Access

Using Datasets for Modeling of Infectious Diseases

Simon Daefler*

Mount Sinai School of Medicine, One Gustave Levy Place, New York, USA

Infections can be described as a complex interplay of microorganisms (bacteria, viruses, fungi) and a host (humans, animals). This interplay determines the outcome of the infection (clearance or latency, symbiosis, death of the invader or the host) and can be influenced by medical measures such as antimicrobial therapy or vaccination. Great strides have been made against infectious diseases by phenomenology and serendipity alone over the last 100 years, but increasing antibiotic resistance and emerging pathogens pose serious challenges at a time when we have high-throughput technologies available. Another challenge is that we still have not figured out any measures to modulate the immune system or metabolism of the host to tilt the host-pathogen interactions to the host's advantage.

At this point the field of infectious diseases is flooded with huge datasets that encompass many different individual compartments but hold little immediate insight. Complete genomic and proteomic datasets of the causative pathogens (bacteria, viruses, and fungi) can now be generated within a few days. Metabolomic datasets are somewhat lagging due to its complexity, but are gaining rapidly. Databases of human genetics that determine the interaction with the pathogen and thus the potential outcome of the infection are increasing rapidly. There is no shortage of computational approaches for the mining of these datasets, as illustrated in the following examples. Genomic analyses can identify novel pathogenicity factors [1], determine the origin of strains causing epidemics [2]or antibiotic resistance markers [3], whole genome constrained based modeling of microorganisms [4] allows to identify metabolic bottlenecks that may be used as targets for novel antimicrobials [5], and correlation of genetic markers with infections allows insight into pathways that are critical to the infectious process [6,7]. Another approach that attempts to integrate both microbe-specific data and host defense mechanisms is presented in this journal [8]. Infection with the emerging pathogen and potential bio-threat agent *Francisella tularensis* causes tularemia, a rapidly fatal disease without proper treatment. Several high-throughput datasets have become available for this organism, but there are limited experimental data and a lack of good model systems to study infections. The approach described in this journal attempts to use existing datasets and build a model that incorporates both pathogen and host. Such a model can then be used to calculate effectiveness of potential vaccine candidates and novel antimicrobial interventions.

Decisions such as the right choice of antibiotics could now be based on genomic and proteomic analysis in the context of determining the genome and proteome of the infecting host and pathogen rather than on traditional phenomenologic methods that resemble a black-box approach. To achieve this goal the mining of huge datasets in real time and the incorporation of such data into suitable models have to be developed. In addition, however, there still needs to be the laborious work of verifying the hypotheses generated and of testing the clinical efficacy of predictions. It seems clear from the strides being made in the respective fields that such approaches will become the standard in the future since they provide much more information than the currently used phenomenological methods. However, their development and implementation may take some time.

References

1. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, et al. (2011) Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 365: 709-717.
2. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, et al. (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364: 33-42.
3. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469-474.
4. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol **7**: 129-143.
5. Raghunathan A, Reed J, Shin S, Palsson B, Daefler S (2009) Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. BMC Syst Biol 3: 38.
6. Alcais A, Abel L, Casanova JL (2009) Human genetics of infectious diseases: between proof of principle and paradigm. J Clin Invest 119: 2506-2514.
7. Zhang SY, Jouanguy E, Ugolini S, Smahi A, Elain G, et al. (2007) TLR3 deficiency in patients with herpes simplex encephalitis. Science 317: 1522-1527.
8. Attie O, Daefler S (2012) An agent based model of tularemia. Data mining in genomics and proteomics.

***Corresponding author:** Simon Daefler, Mount Sinai School of Medicine, One Gustave Levy Place, New York, USA, Tel: 212-241-4690; Facsimile: 212-534-3240; E-mail: simon.daefler@mssm.edu

Received December 03, 2012; **Accepted** December 03, 2012; **Published** December 07, 2012

Citation: Daefler S (2013) Using Datasets for Modeling of Infectious Diseases. J Data Mining Genomics Proteomics 4: e103. doi:10.4172/2153-0602.1000e103