Defining The Secretome: A Gene Model Extender For Secreted Proteins

ZhongQiang Chen[1], Amit Bahl[1], Axel Bernal[3], Qian Liu[3], Manami Nishi[2], Bo Wu[2], Fernando Pereira[3], David Roos[2]
[1]Genomics and Computational Biology, [2]Departments of Biology and [3]Computer & Information Science, Penn Genomics Institute, University of Pennsylvania, Philadelphia PA 19104 USA

Defining the complete complement of secreted proteins encoded in a genome (the secretome) is important for understanding intracellular signaling, intercellular communication and development, and many other aspects of organismal biology. The secretome is also critical for both host defense against pathogens, and pathogen manipulation of the host. This is particularly true for intracellular parasites, such as *Plasmodium* (which causes malaria) and *Toxoplasma* (a prominent congenital pathogen and opportunistic infection associated with immunodeficiency). Secreted proteins are therefore of great interest as biomarkers, drug targets and vaccine candidates. The increasing availability of genome sequences suggests a comparative genomic strategy for characterizing the predicted secretome, as orthologous protein groups include many cases where signal peptides (or other features) are missed due to inaccurate gene models, particularly involving the first exon. This is a common problem: comparing multiple prediction algorithms for the human genome with an extensively curated dataset reveals <70% accuracy in first coding exons vs >80% for internal exons. In order to recover the secretome, we have implemented a strategy for correcting gene predictions by identifying coding exons containing features (such as secretion signals) when comparison with orthologous proteins suggests that this is appropriate. Various sources of evidence (alternative splice sites, gene model probabilities, translational potential, EST evidence, etc) were combined into a probabilistic graphical model representing potential new gene structures. For *Toxoplasma gondii*, this approach successfully recovers ~1.5% of the genome as secretome, such as the missing first exon of a particularly challenging protein with multiple differentially-spliced isoforms. The same pipeline can readily be applied to other features, and other organisms, microbial or metazoan.