Cambridge Healthtech Institute's Eleventh Annual
Bio·IT World
CONFERENCE & EXPO '12

Semantically Enhancing Protein Identification: Systems Biology Knowledgebase for Infectious Disease Screening

Erich Gombocz[1)], James Candlin[2)], Robert Stanley[1)], David Chiang[2)]

1) IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, USA
2) SAGE-N Research, 1525 McCarthy Blvd, Suite 1000, Milpitas, CA 95035, USA

Correspondence: egombocz@io-informatics.com

ABSTRACT:

Bacterial and viral-caused infectious diseases account for major health threats, yet rapid detection, discovery of causal relationships and prevention remains challenging. While significant progress has been made in genomics, protein identification in serum samples has been difficult due to overlap between host proteins and those of the microbial pathogens.

This poster describes a novel integrated systems-biology approach using a semantic knowledgebase to identify peptides from different microorganisms with common disease mechanisms. Those peptide patterns are categorized, refined and qualified as potential biomarker signatures for decision support in screening for microbial threats prior to outbreak of diseases.

Step 1 of the workflow to accomplish this consists of LC-MS/MS analysis of peptides from human serum (depleted of abundant proteins), and the identification of those peptides by scoring matches in a database of pathogenic microbial protein sequences (ABOID.fasta). Spectrum processing is performed automatically through a set of analytical tools (Sorcerer-SEQUEST, Scaffold and the Trans-Proteomic Pipeline[TPP]) pre-configured on the Sorcerer™ analysis platform. The result of this first step is a list of peptides, their sequences, probability scores and assigned microorganisms. Sorcerer also includes a pre-configured systems biology based microbial knowledgebase. This knowledgebase has been created through semantic integration (Sentient Knowledge Explorer™) of lists of microbial peptides with associated genes and their genomic sequences via public microbial resources (PATRIC, ICTV, VIDA, Viral ORFeome, miRBase) and organism-specific pathway information (BioCyc, KEGG) relevant to infectious diseases. Under a common application ontology it also includes similarity-based clustered sequences for homologous protein families (HPFs), functional classification, links to related protein structures, boundaries of conserved regions and bacterial or virus-specific genes. To harmonize the different resources, thesauri for microorganisms and diseases have been applied during import and merging of public resources (Sentient Thesaurus Manager™). This integrated knowledgebase provides a network with functional peptides annotations and their relationships to diseases.

The second step consists of importing the experimental peptide list obtained in Step 1 into this knowledgebase. Visual network analysis and semantic querying (SPARQL) for key relationships provides the complexity reduction needed to identify those peptides which have similar disease-causing functions and appear in several pathogens. Further interrogation of the sub-network results in discovery of key pathway intersections commonly involved in the disease. Iterative refinement of the resulting pattern leads to molecular marker signatures contained in Applied Semantic Knowledgebases (ASK). These signatures can be used for decision support, assisting in outbreak detection and providing mechanistic insights into microbial threats.

References

(1) J. E. Elias, S. P. Gygi: "Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics" in: "Proteome Bioinformatics: Methods in Molecular Biology", S. J. Hubbard, A. R. Jones (Eds.) Vol. 604, 55-71, DOI: 10.1007/978-1-60761-444-9_5 (2010).

(2) E. Gombocz: "Semantic cross-domain integration: The intersection of research, public, and clinical data; creating applicable knowledge for decision support in patient-centric healthcare", *NCBO Webinar Series* Stanford, CA, May 4, 2011.
(3) T.N. Plasterer, R. Stanley, E. Gombocz: "Correlation Network Analysis and Knowledge Integration" In: "Applied Statistics for Network Biology: Methods in Systems Biology", M. Dehmer, F. Emmert-Streib, A. Graber, A. Salvador (Eds.), Wiley-VCH, Weinheim, ISBN: 978-3-527-32750-8 (2011).
(4) S. V. Deshpande, R. E. Jabbour, P. A. Snyder, M. Stanford, C. H. Wick et al.: "ABOid: A Software for Automated Identification and Phyloproteomics Classification of Tandem Mass Spectrometric Data", J Chromatograph Separat Techniq S5:001. doi:10.4172/2157-7064.S5-001 (2011).
(5) M. Mar Albà, M. L. D. Pearl, F.M.G. Shepherd, A.J. Martin, N. Orengo, C.A. P. Kellam, P.: "VIDA: a virus database system for the organisation of virus genome open reading frames". Nucleic Acids Research 2001: 29(**1**), 133-136.
(6) J. Pellet, L.Tafforeau, M. Lucas-Hourani, V. Navratil, L. Meyniel, G. Achaz, A. Guironnet-Paquet, A. Aubl in-Gex, G. Caignard, P. Cassonnet, A. Chaboud, T. Chantier, A. Deloire; C. Demeret, M. Le Breton, G. Neveu, L. Jacotot, P. Vaglio, S. Delmotte, C. Gautier, C. Combet; G. Deleage; M. Favre, F. Tangy, Y. Jacob, P. Andre, V. Lotteau, C. Rabourdin-Combe, P.O. Vidalain: ViralORFeome: an integrated database to generate a versatile collection of viral ORFs. Nucleic Acids Research 2010: 38(1) (Database issue): D371-8.
(7) J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. S. Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon, B. W. Sobral:"PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species", Infect Immun. 2011: 79(11):4286-98.