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

Summary

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, MIST$_2$ VIDA, Viral ORFeome, miRBase) and organism-specific pathway information (BioCyc, KEGG, NCBI BioSystems) relevant to infectious diseases. Under 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 patterns 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.

Methodology

- Protein separation from infected human serum via LC-MS/MS (LTQ Orbitrap Velos, Thermo Fisher).
- Peptide spectra analysis to identify microbial peptides, and scoring of sequence matches against pathogenic microbe sequences (ABOid fasta).
- Automated spectrum processing via pre-configured tools (PTM-enhanced SEQUEST 3G, Scaffold, Trans-Proteomic Pipeline [TPP] on the Sorcerer™ analysis platform (SAGE-N).
- A pre-loaded microbial knowledgebase (IO Informatics) enriched with harmonized public resources (total of 15 different) under a common application ontology was used for semantic integration of the experimental peptide lists (Sentient Knowledge Explorer™).
- Experimental datasets were enriched with functional peptide and gene annotations into a uniform systems biology network for microbial pathogens.
- Network analysis, complexity reduction and SPARQL queries to establish molecular marker patterns with common disease pathway relationships.
- Construction of an Applied Semantic Knowledgebase (ASK) containing collections of patterns used as actionable decision support in screening and biological thread identification.

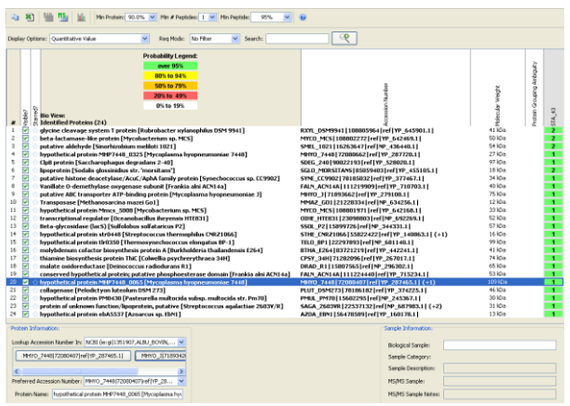

Fig. 1: ***Workflow from human serum sample to microbial pathogen identification***

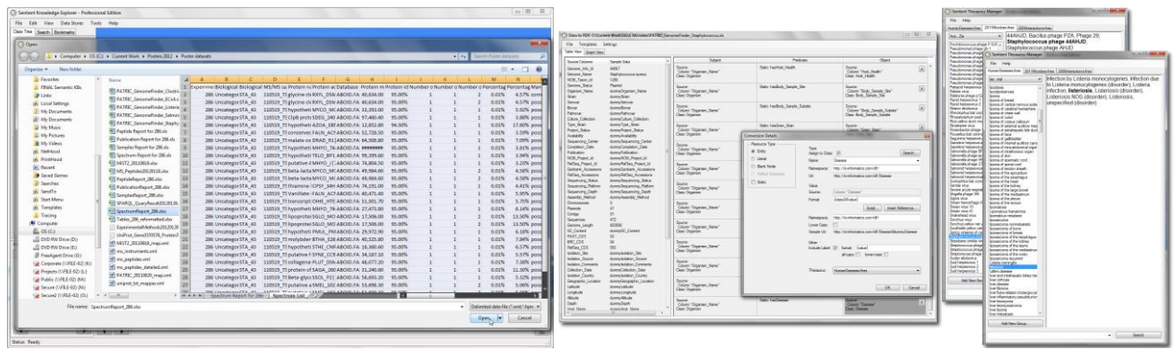

Fig. 1: ***Semantic Import Mapping of Peptides, Enrichment and Harmonization with Public Resources***

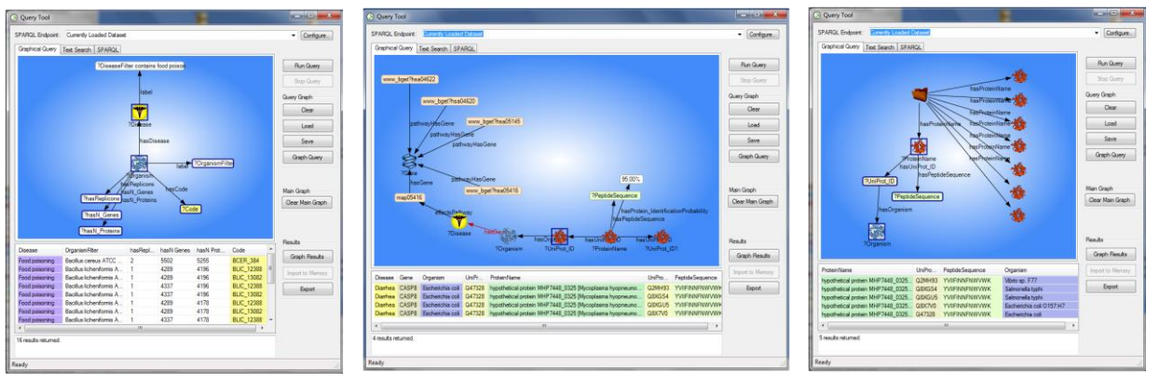

Fig. 2: ***Visual SPARQL Query: Peptide Patterns Sub-Networks as Biological Signatures***

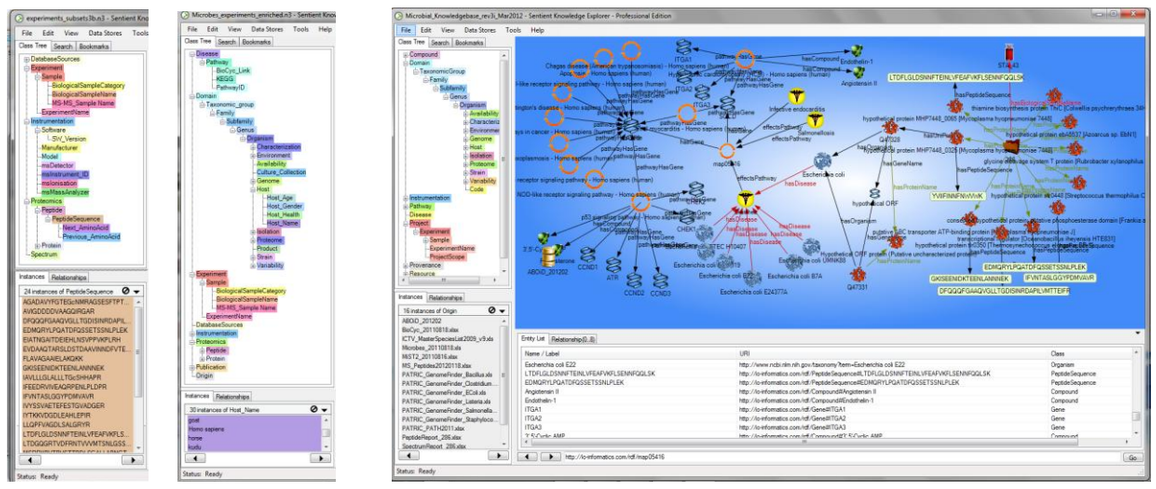

Fig. 3: ***Dynamic Ontology Management and Integrated Semantic Microbial Knowledgebase***

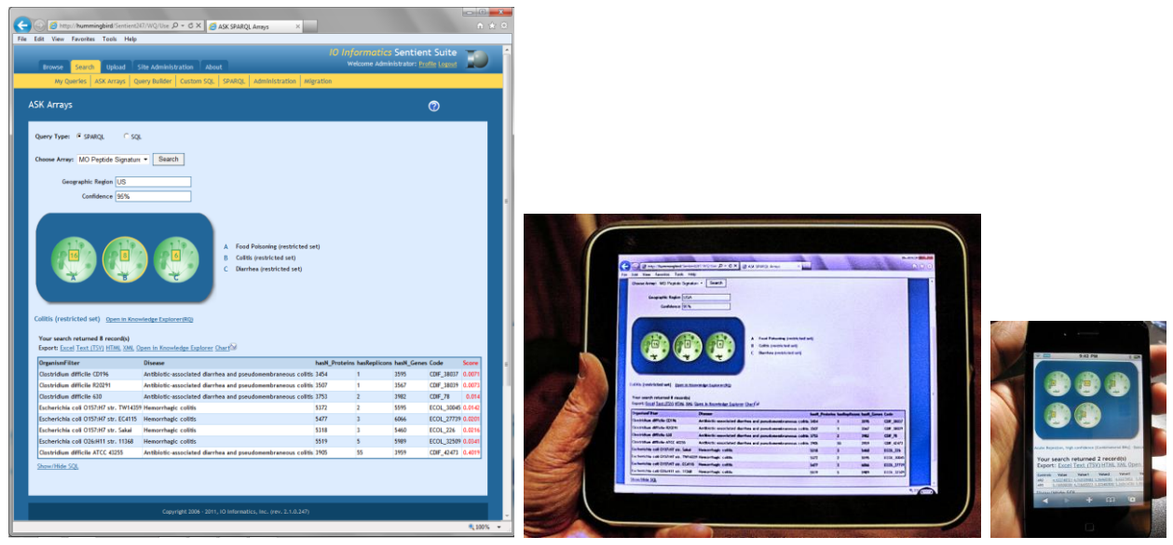

Fig. 4: ***ASK: Actionable Applied Semantic Knowledgebase for Microbial Pathogens***

Results

- Using a pre-configured analytical pipeline to identify peptides from LC-MS/MS of serum samples containing microbial pathogens, it was possible to identify microbial peptide sequences and distinguish them from host peptides at > 95% sequence scoring (Step 1, Sorcerer™ Platform). Exemplary peptides obtained this way are listed in Table 1 below.
- A semantic microbial knowledgebase was created using a broad list of known bacterial and viral peptides and output from public microbial databases under a common, dynamically established application ontology. Harmonization was achieved through use of thesauri for microorganisms and diseases during the semantic import mapping and merging. This knowledgebase was further enriched with disease-related pathway information relevant to the pathogen peptides and their genomic annotations, providing a richly annotated systems-biology network of microbes in their taxonomic categorization (Sentient Knowledge Explorer™, OpenLink Virtuoso Universal Server).
- The peptide list from Table 1 was imported into the pre-configured KB, the peptides visualized in the network and analyzed for intersections between disease pathways of pathogenic organisms.
- Visual SPARQL queries identified peptides with similar disease-causing functions from several pathogens, leading to discovery of key pathway intersections commonly affected in the disease.
- Iterative refinement of peptide patterns leads to molecular marker signatures, used in an Applied Semantic Knowledgebase - ASK™ which is directly applicable for microbial pathogen screening.

	Sample ID	Type	Protein Name	Protein Accession Numbers	Database Sources	Protein MolW (Da)	Protein Probability
1	STA_43	serum	beta-lactamase-like protein [Mycobacterium sp. MCS]	MYCO_MCS|108802272|ref|YP_642469.1|	ABOID.FASTA	49,984.60	95.00%
2	STA_43	serum	hypothetical protein MHP7448_0325 [Mycoplasma hyopneumoniae 7448]	MHYO_7448|72080662|ref|YP_287720.1|	ABOID.FASTA	27,473.80	95.00%
3	STA_43	serum	hypothetical protein str0448 [Streptococcus thermophilus CNRZ1066]	STHE_CNRZ1066|55822422|ref|YP_140863.1| STHE_LMG18311|55820531|ref|YP_138973.1|	ABOID.FASTA	16,360.60	95.00%
4	STA_43	serum	malate oxidoreductase [Deinococcus radiodurans R1]	DRAD_R1|15807565|ref|NP_296302.1|	ABOID.FASTA	64,508.80	95.00%
5	STA_43	serum	conserved hypothetical protein; putative phosphoesterase domain [Frankia alni ACN14a]	FALN_ACN14A|111224440|ref|YP_715234.1|	ABOID.FASTA	52,728.50	95.00%
6	STA_43	serum	Transposase [Methanosarcina mazei Go1]	MMAZ_GO1|21228334|ref|NP_634256.1|	ABOID.FASTA	12,437.70	95.00%

	Unique Peptides	Unique Spectra	Total Spectra	Sequence Coverage	Peptide Sequence	Prev. AA	Next AA	Peptide Probability	SEQUEST XCorr Score	SEQUEST DCn Score	Enzymatic Termini	Calc. +1H Mass (AMU)	Peptide Start IDX	Peptide Stop IDX
1	1	1	2	4.58%	IVYSSVAETEFESTGVADGER	K	Y	95.00%	3.14	0.224	2	2,246.05	76	96
2	1	1	1	6.14%	YVIIFINNFNWVWK	K	M	95.00%	4.29	0.111	2	1,855.99	6	19
3	1	1	1	6.57%	YYTYLEDNK	R	S	95.00%	2.75	0.283	2	1,208.55	23	31
4	1	1	1	7.09%	VEEMLDNVTVNDVRMIVATDSSAILGIGDQGFGGMAISIGK	R	L	95.00%	2.89	0.215	2	4,223.10	147	187
5	1	1	1	3.59%	IFVNTASLGGYPDMVAVR	R	E	95.00%	3.16	0.27	2	1,909.98	335	352
6	1	1	1	16.50%	IYTKKVDGDLEAHLEPIR	R	K	95.00%	3.57	0.319	2	2,097.13	89	106

Discussion

- This study focused on establishing initial sets of microbial peptide markers in conjunction with a pre-configured integrated semantic knowledgebase (Sorcerer, Sentient Knowledge Explorer) to check the validity to rapidly identify biological marker patterns applicable to pathogen screening.
- It should be emphasized that applying semantic technology was instrumental to the success. While this study represents promising steps towards marker-based rapid microbial pathogen detection, it also should be noted that additional work in validation will be required to assure broad applicability.
- Applying semantic technology to integration of experimental and public knowledge resources provides a rapid, cost-effective, extensible and biologically sound method towards understanding of disease mechanisms of microbial pathogens.

Future Developments

- Once fully validated across larger sample sets, peptide marker signatures for microbial pathogens will lead to development of low-cost multiplexed assays for rapid detection of biological threats, to characterize origin and type of disease outbreaks and to develop preventive measures (such as broadly applicable drugs or vaccines) effective for entire classes of pathogens.

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):4286-9