Separating Drivers from Passengers in Whole Genome Analysis: Identification of Combinatorial Effects of Genes by Mining Knowledge Sources

Stephen Anthony
Centre for Health Informatics
University of New South Wales
Sydney, Australia
e-mail: s.anthony@unsw.edu.au

Enrico W. Coiera
Centre for Health Informatics
University of New South Wales
Sydney, Australia
e-mail: e.coiera@unsw.edu.au

Vitali Sintchenko
Centre for Infectious Diseases and Microbiology
Sydney Medical School, University of Sydney
Sydney, Australia
e-mail: vitali.sintchenko@sydney.edu.au

***Abstract*—This study aimed to develop a new informatics platform for the discovery, recovery and multi-level analysis of the effects of individual genes and multiple gene combinations on pathophenotypes of bacteria. Natural language processing algorithms were employed to extract gene-disease associations from PubMed literature and annotated genomes of bacteria with epidemic potential. From these associations gene virulence profiles were generated enabling the comparison of gene signatures within and across genomes. It allowed the identification of virulence genes and construction of their association networks as well as the detection of knowledge gaps. This proof-of-concept study confirmed the feasibility of our original approach for integrating bacterial genome level knowledge with published observations from clinical settings.**

Keywords: structural bioinformatics; whole genome analysis; text mining; infectious diseases; knowledge discovery

I. INTRODUCTION

The exponential growth in whole genome sequencing has created new challenges for data integration and analysis. It has crucially increased the rate of discovery of associations between genes and diseases [1]. The mapping of relationships between genes and disease phenotypes has become possible due to synergistic advances in text mining and the availability of quality data and indexed text in the public domain. For example, online catalogs of human genome-wide association studies exceeded 700 publications linking genetic variations and diseases.

Knowledge about the etiology and pathogenesis of diseases has increasingly been stored in literature and in databases, including sets of fully or partially annotated bacterial genomes [2]. Text mining approaches are gaining importance in the extraction and collation of data and text mining with bioinformatics databases [3-5]. In particular, significant progress has been made in building applications for the knowledge-based profiling of individual genes [6, 7], gene mining and mapping to diseases [8, 9] and mining complex features for predicting drug resistance [10]. These developments have drawn attention to the problem of diminishing returns of existing analytic approaches. The challenge of potential non-linearity in the mapping of genotypes to phenotypes and our ability to address it has been emphasized, with calls for an analytical retooling to address the combinatorial nature of gene-disease effects [11]. Bioinformatics techniques such as collapsing or binning have been borrowed from SNP-based genome-wide studies and applied to study human diseases that can be affected by multiple genes. However, microbial genome-wide association studies have not received due attention so far and bioinformatics applications for microbial genome analyses remain relatively underdeveloped.

The infectious disease (ID) domain presents a new frontier of high-throughput sequence analysis adding another, pathogen-specific, dimension to genome and disease association studies. Early findings indicate that disease-defining properties of pathogens are both multi-factorial and combinatorial [12] but further progress has been limited by the lack of analytical tools. The integrated microbial genome initiatives [13, 14] so far have focused attention on the alignment and comparison of genome sequences without linking sequencing data to ID attributes or clinical outcomes [15].

Whole genome sequence analysis has promised to improve the accuracy of ID risk assessment and the discriminatory power of tracking of outbreaks. Several new tools have been proposed to leverage these resources and to assist researchers with identification of genomic targets for vaccine and drug discovery [16] and the study of pathogen evolution [17]. Until recently, each gene or protein was studied as a single entity. However, new 'omics' technologies have allowed the analysis of large numbers of genes simultaneously and the generation of complex networks [18].

The aim of this study was to develop a new informatics platform for the discovery, recovery and multi-level analysis of the effects of individual genes and multiple gene combinations on properties of pathogenic bacteria.

 ISBN: 978-1-61208-137-3

II. METHODS

A. Approach and Definitions

The assessment of the impact of individual genes was based on searches for literature-based associations of genes with ID syndromes in order to separate key genes responsible for pathogenic phenotype of bacteria ('*drivers*') from non-pathogenic genes of little consequences to bacterial virulence ('*passengers*'). The working hypothesis was that virulence genes can be identified through the combination of virulence profiles, within and across bacterial genomes. Virulence profiles were generated from gene-disease associations.

The concepts of 'core' and 'dispensable' (or unique) genomes were considered [19]. The core genome consisted of the set of common genes conserved across all strains, encoding those functions necessary for the basic biology of the species. The dispensable genes contribute to the diversity within the species, including virulence, transmissibility, antibiotic resistance and niche adaptation [19]. Core genes were identified and the combined virulence profiles across species were compared to the virulence profiles of individual species. Core genes were also contrasted against genes that are unique to individual microorganisms.

B. Data Sources

The 2011 MEDLINE/PubMed Baseline Distribution was employed as the primary source of literature for generating gene virulence profiles. The 2011 baseline comprises 10,891,200 citations that contain text from article abstracts. The citations were loaded into a Postgres Database Management System (DBMS) with associated full text indexes generated for the title and abstract. The text vectors were generated by the DBMS using a template for English that is based on the Porter stemming algorithm [20].

Fully sequenced genomes available through the National Center for Biotechnology Information (NCBI) GenBank [21] were employed for gene symbol and gene sequence location information for a number of bacterial genomes with pathogenic potential representing both Gram positive and Gram negative obligatory and opportunistic pathogens with the genome size of 3-4Mb and a range of core genome sizes (Table 1).

A list of syndromes associated with infectious diseases (108 items) was constructed as reported previously [22]. Names of syndromes included *sepsis*, *pneumonia*, *meningitis*, *encephalitis*, *cellulitis*, *wound infection*, and *urinary tract infection*, among others. However, ID syndromes uniquely associated with specific pathogens such as tuberculosis, malaria, dengue, etc. were excluded from the list to minimize the detection of trivial associations.

C. Gene Symbol Classification

Each PubMed abstract was indexed for mentions of a syndrome or species related gene symbol. The SPECIALIST lexicon [23] supported the identification of variants in nomenclature, spelling, and clinical abbreviations.

TABLE I. BACTERIAL GENOMES UTILISED IN THE STUDY

Bacterial Genome	GenBank Accession	Proteins
Listeria monocytogenes 4b F2365	AE017262	2821
Mycobacterium tuberculosis H37Rv	AL123456	3988
Neisseria meningitidis 053442	CP000381	2020
Pseudomonas aeruginosa PA01	CP000744	5566
Salmonella typhi CT18	AL513382	4391
Staphylococcus aureus MRSA252	BX571856	2650
Streptococcus pyogenes MGAS6180	CP000056	1894

The identification of gene symbols was performed using a combined search and classification strategy. A gene symbol classification model was constructed by acquiring a set of 5,003 unique gene symbols for training sourced from nine fully sequenced NCBI genomes. Inclusion in the training set comprised the following criteria. Gene symbols had to contain at least two out of three characters from the classes: upper case, lower case, and numeric. Gene symbols could not contain underscores or hyphens or were otherwise excluded. A full text search was performed for each of the symbols in the training set. As the full text search configuration collapses case, a further constraint was applied to ensure each gene mention in the abstract appeared with identical orthography.

The rule extraction from each symbol's contextual window utilized morphosyntactic and lexical features. The features were generated from a set of base templates. Base template classes employed to generate features included tokens either side of the symbol, tokens from both sides of the symbol, a function to determine whether the symbol was enclosed in parentheses, and a function to determine whether mixed case tokens were present in the context. Tokens used in feature generation through templates were lemmatized and had their orthographic case folded.

An entropy maximization approach was employed to rank each contextual feature. The measure was calculated by determining the number of times each pattern co-occurred with gene symbols in relation to the total number of occurrences of the contextual pattern in the training corpus. The higher the entropy of the feature, the greater its contribution to the classification of a symbol. An evaluation of the approach is presented in the results section.

D. Representation of Multiple Gene-Disease Associations

A virulence profile was generated for each gene based on the co-occurrence between syndromes and individual pathogen-specific gene symbols. The latter was defined as a gene symbol that co-occurs with a bacterial species name in the same document (e.g., *Staphylococcus aureus* AND *dnaA*). Co-occurrence was calculated using pointwise mutual information (PMI) [24]. The PMI formula (1) provides a measure of association where p represents probability, and both x and y represent terms.

$$\mathrm{PMI}(x,y) = \log_2 (p(x, y) / (p(x) * p(y))) \qquad (1)$$

The association measures for pathogen-specific gene symbols were employed in the generation of vectors intended to express virulence potential of a gene. The summation of individual vector components represented a gene *virulence factor*. Each gene's vector was compared to every other vector in the genome, and across genomes, and similarities between expression vectors were estimated using a Euclidean distance measure [25].

III. RESULTS

A. Gene Classification

Our text mining strategy successfully extracted gene names and syndrome entities. For example, it differentiated the 1,670,132 abstracts that contain the word stem *year* from the three abstracts that contained the gene symbol *yeaR*.

Let us illustrate the performance of our approach using the fully sequenced genome for *Neisseria meningitidis* 053442 as an example. Specifically, it encodes the gene product *porA* (porin, class I outer membrane protein). A PubMed-wide search was initiated for all article titles or abstracts that contained the pathogen-specific symbol *porA* in association with any one of the 108 ID-related syndromes. The symbol *porA* was found to occur in 312 distinct articles, and the pathogen-specific form appeared in 175 distinct articles. A total of 2,079,834 distinct articles were found to contain any one or more of the syndrome terms. The terms co-occured in a total of 41 documents. Unsurprisingly in this case, over half of the associations could be attributed to co-occurrence with a mention of some form of the syndrome *meningitis*. The PMI association scores between the pathogen-specific gene symbol *porA* and individual ID syndromes constituted a virulence profile.

The gene classification approach was evaluated using an independent test set. The test set was sourced from an additional three NCBI genomes. Gene symbols from these three genomes that were present in the training set were excluded. The test set contained 3,649 gene symbols in total, reduced to 1,271 once common genes were excluded. The search with exact match criteria produced 4,128 contexts from documents that contained a test set gene symbol. The approach correctly classified 4,032 of the 4,128 test set instances, achieving and overall accuracy of 97.67%. The largest source of errors originated from instances where the context contained never before seen tokens, a high proportion of punctuation characters, or no mixed case tokens that are often indicative of gene symbols. For example, the gene *lys1* that is found exclusively in the test set and embedded in the following context was misclassified:

(UGA), lys1-1' (UGA) [PMID: 782552]

The contextual tokenization resulted in the following 4-token window either size of the target gene *lys1*:

| (| UGA |) | , | lys1 | -1 | ' | (| UGA |

A number of suggestions for the remediation of these types of errors are presented in the discussion section.

B. Genome-wide Virulence Profiling

The first representation generated from the expression vectors resulted in a genome-wide virulence factor. Table II illustrates overall virulence factors for bacterial genomes that have been calculated by summing the PMI ID-association scores for each pathogen-specific gene within a genome.

TABLE II. TOTAL VIRULENCE FACTOR PER MICROORGANISM

Bacterial Genome	Total No. of Genes	Virulence Factor (bits)
Listeria monocytogenes 4b F2365	756	78.95
Mycobacterium tuberculosis H37Rv	1672	371.04
Neisseria meningitidis 053442	869	183.16
Pseudomonas aeruginosa PA01	1736	810.92
Salmonella typhi CT18	2558	332.07
Staphylococcus aureus MRSA252	782	369.39
Streptococcus pyogenes MGAS6180	776	340.81

Genome-wide virulence profiles for individual genes can be presented in a circularized form (Figure 1) to reflect the exact position of genes in a genome and identify most commonly reported genes of an individual bacteria (*Neisseria meningitidis* in this case) that have been associated with clinical presentations or adverse outcomes of ID. These profiles were generated by plotting each gene's frequency of co-occurrence with ID-syndromes in the literature.

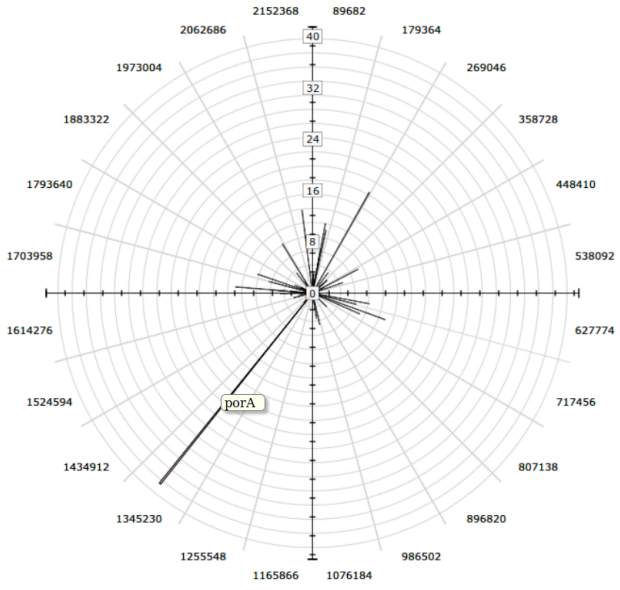

Figure 1. Genome-wide virulence profile for *Neisseria meningitidis*. Individual gene frequency of co-occurence with ID syndromes in the literature are plotted according to each gene's location in the genome.

1) Potential Knowledge Gap Identification

Core genes were determined and visualized by plotting the virulence profile for each of the genes that were common across multiple genomes (Figure 2). Such virulence profiling

contrasts core genes that have been associated with pathogenicity in one (e.g., *leusS, miaA*) or several (e.g., *murE, mutL*) bacterial species.

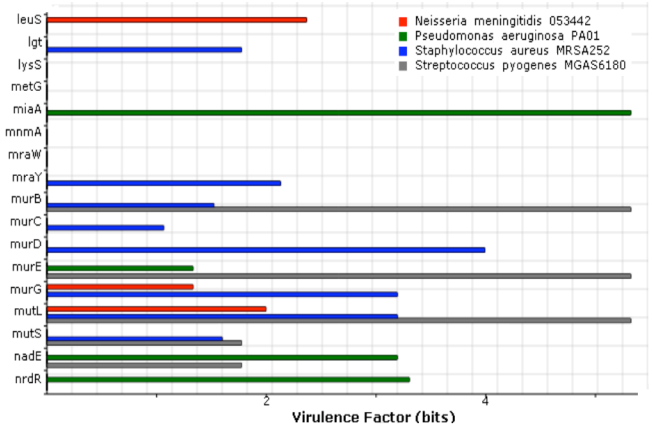

Figure 2. Core Gene Virulence Factor Comparison. A side-by-side comaprison of gene virulence factors for a fragment of alphabetically sorted genes common across selected genomes.

Gene virulence profiles were subsequently employed to identify potential gaps in the knowledge regarding a gene's impact on ID outcomes. The process was initiated by identifying common genes across all selected genomes. This set of genes was then restricted to those that contained at least one positive expression value in any genome. Each of the residual genes was plotted against its respective total (across-genome) virulence profile. For example, a striking discrepancy was identified between a high frequency of associations of many genes of bacteria with ID syndromes and the lack of those in the *Staphylococcus aureus* genome (Figure 3). Specifically, this observation suggested two knowledge gaps may exist for our understanding of *ftsZ* and *glyA* gene functions in the genome of *S. aureaus* (Figure 3). Interestingly, *ftsZ* gene has been associated with virulence properties of other bacteria due to its role in the synthesis of the protein tubulin that participates in replication of toxin-encoding plasmids [26]. Glycinecin gene (*GlyA*) has been also identified as a putative virulence gene in other bacteria because of its involvement in the synthesis of bacteriocins [27].

2) Identification of key virulence genes

The first approach to identifying drivers involved the identification of common genes. The simplest determination of a driver gene could be defined as a gene that is common across genomes and expresses an above-average overall virulence factor. This approach implicates the genes *dnaK*, *eno*, *folD*, *ftsZ*, and *glyA* amongst others as can be seen in Figure 3 as their virulence factors fall above the overall across-genome average indicated by the horizontal line.

The next approach to driver gene identification resulted in a network-based linkage analysis. The analysis was performed to detect combinations of genes that are typically associated with syndromes. The visualization of relationships between individual genes in Figure 4 shows links that are formed between genes within the *Listeria monocytogenes* 4b F2365 genome. Edge weights reflect the strength of association between gene virulence profiles. This network representation (Figure 4) also highlights indirect relationships between some genes (e.g., *plcB* and *prfA* through *actA*) and identifies highly connected virulence genes of particular relevance for the virulence assessment (e.g., *actA* and *prfA*). The *actA* gene, which is involved in the synthesis of bacterial protein actine, presents a compelling example of a 'driver' gene. A literature review, conducted following our experiments, confirmed the *actA* gene of *Listeria monocytogenes* has been a key player in several biological mechanisms relevant to virulence, such as in escaping from vacuoles, undergoing intracellular growth, and spreading to neighboring cells in cell cultures [28].

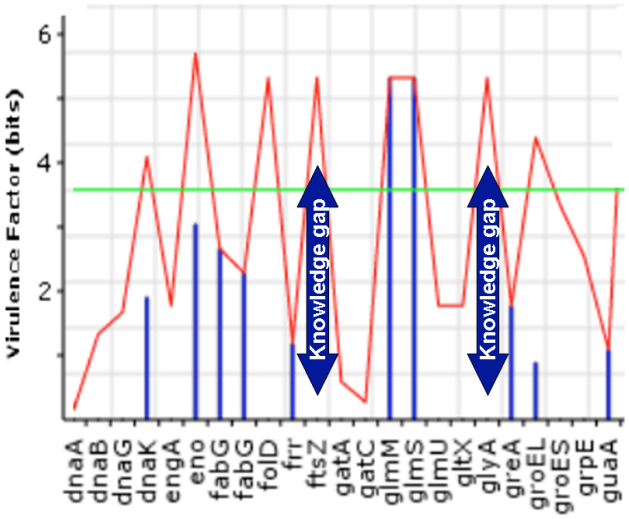

Figure 3. Gene Virulence Knowledge Gap Identification. Represents the virulence profiles for a fragment of core genes for the species *S. aureus* (vertical bars) in relation to the combined virulence profile across all genomes (line plot). The horizontal line represents the average virulence score of the combined virulence profiles.

3) Combinatorial effects of virulence genes

Similarities between virulence profiles were also compared both within and across microbial genomes in order to link individual genes with each other when they were found to be associated with ID syndromes. Figure 5 illustrates our approach to matching virulence profiles of individual genes within and between bacterial genomes. It seems to make more explicit the potential indirect links between genes of different function and origin that could be clustered according to their virulence. The links across the top of Figure 5 represent within-genome gene virulence similarity and the links across the bottom reflect between genomes similarity. Common or core genes were color coded in blue and genes unique to the genome of *S. pyogenes* were coded in red. Gene names in black represented house-keeping genes that could be found in several different bacterial species from our dataset.

 ISBN: 978-1-61208-137-3

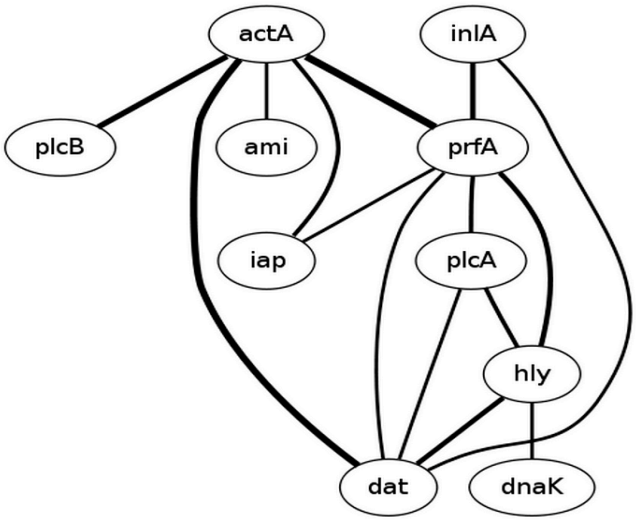

Figure 4. Gene virulence networks for *L. monocytogenes*. All pairs of gene expression vectors within each genome are compared and connected if their virulence profiles are similar.

4) Relative Impact of Unique Genes

The relative impact of unique or disposable versus core genes significantly differed between bacterial genomes in our dataset. This difference could be explained by their different propensity for lateral gene transfer. Table III lists the virulence factors for genes that are unique to each genome when compared against the remainder of genomes listed in the table. The table quantifies the aggregated total virulence factors for unique versus core genes for each species. There are a total of 202 core genes across all species. The number of core and unique gene occurrences per species are listed separately.

TABLE III. COMMON AND UNIQUE GENE VIRULENCE FACTORS PER MICROORGANISM

Bacterial Genome	Genes Virulence Factor (bits)		Number of Genes	
	Core	Unique	Core	Unique
L. monocytogenes 4b F2365	15.97	11.65	210	131
M. tuberculosis H37Rv	53.86	203.55	217	942
N. meningitidis 053442	14.81	54.84	205	153
P. aeruginosa PA01	102.13	367.60	208	719
S. typhi CT18	19.63	173.03	206	1408
S. aureus MRSA252	112.31	153.47	208	232
S. pyogenes MGAS6180	68.82	124.60	210	261

IV. DISCUSSION

Omics-based medicine demands significant re-tooling for continuous re-assessment of evidence for genome-phenome associations. The methods presented in this study can facilitate identification of combinations of genes within and across annotated bacterial genomes, differentiation of key virulence genes from genes of limited clinical relevance, detection of potential knowledge gaps, and measurement of the relative impact of individual genes and gene combinations. The methods are based on a set of tools and resources that comprise a large text corpus with full text search capabilities, a gene symbol classifier, and techniques for gene virulence profiling through literature-based association mining.

A key innovation in this work resulted from the establishment of literature-based pathogen-specific ID association measures spurred by a novel approach to increasing the specificity of gene symbol retrieval. For example, our text mining strategy differentiated the 171,203 PubMed mentions of the different forms of the indexed token *mode* from the 14 mentions of the gene symbol *modE*.

The transformation of our text mining measures into vectors that expressed gene virulence profiles led to a number of applications. The virulence profiles were applied both genome wide and across genomes. This resulted in the ability to identify combinations of genes that share virulence profiles. Comparing genome-wide gene profiles with overall virulence profiles across genomes also identified potential knowledge gaps. The next application identified potentially hidden driver genes indirectly linked by other genes that expressed similar virulence profiles.

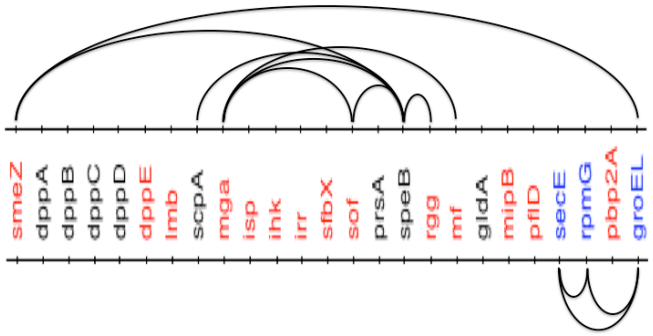

Figure 5. Within and between genome expression profile macthing. Links across the top represent genes that have similar expression profiles within the genome *S. pyogenes*. Links across the bottom represent genes with similar expression profiles across all genomes.

The novel approach taken to gene classification, as opposed to recognition, greatly improves search specificity. A number of further modifications have been discussed previously including a corresponding version of the feature generation templates that preserve case, and experiments with varying context-window sizes. Other potentially useful feature types include unlemmatized tokens, morphological analysis to inspect affixes, and semantic type analysis.

However, some potential limitations of the study have to be acknowledged. One of the major limitations to the gene classification evaluation and approach in general is the absence of negative examples. A possible solution to this problem could come in the form of leveraging examples from a text genre outside of the biomedical domain. Importantly, the aforementioned strategy does not eliminate false negatives that are collected by the search for gene symbols that are presented using a single orthographic case, particularly problematic are those that overlap with common English words such as *era*, *lip*, *map*, and *trap*. In order to combat this issue the abstract text for each article found to contain a gene symbol mention was tokenized and a four-token window either side of the symbol was extracted.

 ISBN: 978-1-61208-137-3

Training set gene symbols were found to occur in their exact form in 102,399 articles. Although not applied to this work the exact matching constraint employed to extract context could be relaxed to allow variants of gene symbols and capture other gene products. A number of adjustments to the classification algorithm could be made to redress potential limitations. For example, a parallel context could be constructed that excludes punctuation characters. Composite approaches are conceivable given the relatively high incidence of parenthesized gene symbols (11,296 out of 102,398 training instances). Another parameter that could be adjusted in future experiments is the context window size.

V. CONCLUDING REMARKS

This proof-of-concept study confirmed the feasibility of our original approach for integrating bacterial genome level knowledge with published observations from clinical settings. It opens a new opportunity for real-time assessment of virulence of bacterial genomes and for identification of high-impact genes and their combinations. Further 'wet lab' experiments are required to validate the utility of the knowledge gap detection function.

This work has culminated in an online toolset that enables researchers to explore, recover, and potentially discover new insights into microbiological mechanisms that contribute to infection. An online implementation of this work can be accessed from http://purl.org/infectious/genome.

REFERENCES

[1] K.E. Ormond, M.T. Wheeler, L. Hudgins, T.E. Klein, A.J. Butte, R.B. Altman, et al., "Challenges in the clinical application of whole-genome sequencing", Lancet, vol. 375, May. 2010, pp. 1749-1751, doi:10.1016/S0140-6736(10)60599-5.

[2] T. Korves, and M.E. Colosimo, "Controlled vocabularies for microbial virulence factors", Trends Microbiol, vol. 17, Jul. 2009, pp. 279-285, doi:10.1016/j.tim.2009.04.002.

[3] S. Ananiadou, D.B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology", Trends Biotechnol, vol. 24, Dec. 2006, pp. 571-579, doi:10.1016/j.tibtech.2006.10.002.

[4] Y. Kano, P. Dobson, M. Nakanishi, J. Tsujii, and S. Ananiadou, "Text mining meets workflow: linking U-Compare with Taverna", Bioinformatics, vol. 26, Oct. 2010, pp. 2486-2487, doi:10.1093/bioinformatics/btq464.

[5] J.O. Korbel, T. Doerks, L.J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, et al., "Systematic association of genes to phenotypes by genome and literature mining", PLoS Biol, vol. 3, May. 2005, pp. e134, doi:10.1371/journal.pbio.0030134.

[6] H. Xu, J.W. Fan, G. Hripcsak, E.A. Mendonca, M. Markatou, and C. Friedman, "Gene symbol disambiguation using knowledge-based profiles", Bioinformatics, vol. 23, Apr. 2007, pp. 1015-1022, doi:10.1093/bioinformatics/btm056.

[7] W. Xuan, P. Wang, S.J. Watson, and F. Meng, "Medline search engine for finding genetic markers with biological significance", Bioinformatics, vol. 23, Sep. 2007, pp. 2477-2484, doi:10.1093/bioinformatics/btm375.

[8] M. Garcia-Remesal, A. Cuevas, D. Perez-Rey, L. Martin, A. Anguita, D. de la Iglesia, et al., "PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids", Bioinformatics, vol. 26, Nov. 2010, pp. 2801-2802, doi:10.1093/bioinformatics/btq520.

[9] T. Matsunaga, and M. Muramatsu, "Disease-related concept mining by knowledge-based two-dimensional gene mapping", J Bioinform Comput Biol, vol. 5, Oct. 2007, pp. 1047-1067, doi:10.1142/S0219720007003077.

[10] H. Saigo, T. Uno, and K. Tsuda, "Mining complex genotypic features for predicting HIV-1 drug resistance", Bioinformatics, vol. 23, Sep. 2007, pp. 2455-2462, doi:10.1093/bioinformatics/btm353.

[11] J.H. Moore, F.W. Asselbergs, and S.M. Williams, "Bioinformatics challenges for genome-wide association studies", Bioinformatics, vol. 26, Feb. 2010, pp. 445-455, doi:10.1093/bioinformatics/btp713.

[12] D.G. Lee, J.M. Urbach, G. Wu, N.T. Liberati, R.L. Feinbaum, S. Miyata, et al., "Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial", Genome Biol, vol. 7, Oct. 2006, pp. R90, doi:10.1186/gb-2006-7-10-r90.

[13] T. Davidsen, E. Beck, A. Ganapathy, R. Montgomery, N. Zafar, Q. Yang, et al., "The comprehensive microbial resource", Nucleic Acids Res, vol. 38, Jan. 2010, pp. 340-345, doi:10.1093/nar/gkp912.

[14] V.M. Markowitz, I.M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, et al., "The integrated microbial genomes system: an expanding comparative analysis resource", Nucleic Acids Res, vol. 38, Jan. 2010, pp. 382-390, doi:10.1093/nar/gkp887.

[15] N. Mulder, H. Rabiu, G. Jamieson, and V. Vuppu, "Comparative analysis of microbial genomes to study unique and expanded gene families in Mycobacterium tuberculosis", Infect Genet Evol, vol. 9, May. 2009, pp. 314-321, doi:10.1016/j.meegid.2007.12.006.

[16] A. Muzzi, V. Masignani, and R. Rappuoli, "The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials", Drug Discov Today, vol. 12, Jun. 2007, pp. 429-439, doi:10.1016/j.drudis.2007.04.008.

[17] G.S. Vernikos, and J. Parkhill, "Resolving the structural features of genomic islands: a machine learning approach", Genome Res, vol. 18, Feb. 2008, pp. 331-342, doi:10.1101/gr.7004508.

[18] S.C. De Keersmaecker, I.M. Thijs, J. Vanderleyden, and K. Marchal, "Integration of omics data: how well does it work for bacteria?", Mol Microbiol, vol. 62, Dec. 2006, pp. 1239-1250, doi:10.1111/j.1365-2958.2006.05453.x.

[19] S. Bentley, "Sequencing the species pan-genome", Nat Rev Microbiol, vol. 7, Apr. 2009, pp. 258-9, doi:10.1038/nrmicro2123.

[20] M.F. Porter, "An Algorithm for Suffix Stripping", Program-Automated Library and Information Systems, vol. 14, 1980, pp. 130-137.

[21] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, "GenBank", Nucleic Acids Res, vol. 36, Jan. 2008, pp. D25-30, doi:10.1093/nar/gkm929.

[22] V. Sintchenko, S. Anthony, X.H. Phan, F. Lin, and E.W. Coiera, "A PubMed-wide associational study of infectious diseases", PLoS One, vol. 5, Mar. 2010, pp. e9535, doi:10.1371/journal.pone.0009535.

[23] A.C. Browne, A.T. McCray, and S. Srinivasan, The SPECIALIST Lexcon, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, 2000.

[24] K.W. Church, and P. Hanks, "Word association norms, mutual information, and lexicography", Computational Linguistics, vol. 16, March. 1990, pp. 22-29, doi:10.3115/981623.981633.

[25] E. Deza, M.M. Deza, and SpringerLink (Online service), Encyclopedia of Distances, Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2009.

[26] Y.J. Pan, T.L. Lin, C.R. Hsu, and J.T. Wang, "Isolation of genetic loci associated with phagocytosis and virulence in Klebsiella pneumoniae using a Dictyostelium model", Infect Immun, vol. Dec. 2010, pp. doi:10.1128/IAI.00906-10.

[27] W. Oswald, D.V. Konine, J. Rohde, and G.F. Gerlach, "First chromosomal restriction map of Actinobacillus pleuropneumoniae and localization of putative virulence-associated genes", J Bacteriol, vol. 181, Jul. 1999, pp. 4161-9.

[28] M. Conter, A. Vergara, P. Di Ciccio, E. Zanardi, S. Ghidini, and A. Ianieri, "Polymorphism of actA gene is not related to in vitro virulence of Listeria monocytogenes", Int J Food Microbiol, vol. 137, Jan. 2010, pp. 100-5, doi:10.1016/j.ijfoodmicro.2009.10.019.

 ISBN: 978-1-61208-137-3