

Integration of Instance-based learning & Text Mining for Identification of Potential Virus / Bacterium as Bioterrorism Weapons*

Xiaohua Hu¹, Xiaodan Zhang¹, Daniel Wu¹, Peter Rumm²,

¹College of Information Science and Technology, Drexel University,
Philadelphia, PA 19104

{thu@cis, xzhang@cis, daniel.wu}@drexel.edu

²School of Public Health, Drexel University, Philadelphia, PA 19104
pdr26@drexel.edu

Abstract. There are some viruses and bacteria that have been identified as bioterrorism weapons. However, there are a lot other viruses and bacteria that can be potential bioterrorism weapons. A system that can automatically suggest potential bioterrorism weapons will help laypeople to discover these suspicious viruses and bacteria. In this paper we apply instance-based learning & text mining approach to identify candidate viruses and bacteria as potential bioterrorism weapons from biomedical literature. We first take text mining approach to identify topical terms of existed viruses (bacteria) from PubMed separately. Then, we use the term lists as instances to build matrices with the remaining viruses (bacteria) to discover how much the term lists describe the remaining viruses (bacteria). Next, we build a algorithm to rank all remaining viruses (bacteria). We suspect that the higher the ranking of the virus (bacterium) is, the more suspicious they will be potential bio-terrorism weapon. Our findings are intended as a guide to the virus and bacterium literature to support further studies that might then lead to appropriate defense and public health measures.

1 Introduction

Terrorist attack concerns many people in the world. Biological agent is one of five categories of terrorist weapons. For certain biological agents, the potential for devastating casualties is very high. The anthrax mail attack in October, 2001 terrorism caused 23 cases of anthrax-related illness and 5 deaths. Due to the widespread availability of agents, widespread knowledge of production methodologies, and potential dissemination devices, bioterrorism can be very cute for now and future. Because it is very difficult for laypeople diagnose and recognize most of the diseases caused by biological weapons, we need surveillance systems to keep an eye on potential uses of such biological weapons [1]. In this paper, we propose an instance based learning

* This work is supported partially by the NSF Career grant IIS 0448023 and NSF 0514679 and PA Dept of Health Tobacco Formula Grants.

method to discover biological agents as potential Bioterrorism Weapons (BW). Before discovering potential BW, it's reasonable to study the characteristics of biological agents identified by human experts as BW. Some human experts have generalized some criteria for identifying virus and bacteria. The more detail is in section 3. However, it's hard for human being to map all the viruses and bacteria one by one to these criteria. Moreover, the list is compiled manually, requiring extensive specialized human resources and time. Because the biological agents such as viruses are evolving through mutations, biological or chemical change, some biological substances have the potential to turn into deadly virus through chemical/genetic/biological reaction, there should be an automatic approach to keep track of existing suspicious viruses and to discover new viruses as potential weapons. We expect that it would be very useful to identify those biological substances and take precaution actions or measurements. For better studying the characteristics of existed biological agents as BW, we use a text mining approach to extract topical MeSH terms from them. This is an exhaustive approach, so we believe that the topical MeSH terms we extract are very representative of the particular BW collection. Then, we use this discovered terms to build a term biological agent matrix from which we check how much these terms can be topical terms for the remaining biological agents. Later, we use the combination of these terms to rank each remaining biological agent. In the end, we get a top ranked term list that can be used as key words for human experts to examine the remaining biological agents. The most important is that we generate a biological agent as potential BW ranked by the extracted terms from the existed biological agents. We suspect that the higher rank the biological agent, the more it can become potential BW.

2 Related Works.

The problem of mining implicit knowledge/information from biomedical literature was exemplified by Dr. Swanson's pioneering work on Raynaud disease/fish-oil discovery in 1986 [5]. Back then, the Raynaud disease had no known cause or cure, and the goal of his literature-based discovery was to uncover novel suggestions for how Raynaud disease might be caused, and how it might be treated. He found from biomedical literature that Raynaud disease is a peripheral circulatory disorder aggravated by high platelet aggregation, high blood viscosity and vasoconstriction. In another separate set of literature on fish oils, he found out the ingestion of fish oil can reduce these phenomena. But no single article from both sets in the biomedical literature mentions Raynaud and fish oil together in 1986. Putting these two separate literatures together, Swanson hypothesized that fish oil may be beneficial to people suffering from Raynaud disease [5] [6]. This novel hypothesis was later clinically confirmed by DiGiacomo in 1989 [2]. Later on [4] Dr. Swanson extended his methods to search literature for potential virus. But the biggest limitation of his methods is that, only 3 properties/criteria of a virus are used as search key word and the semantic information is ignored in the search procedure. In this paper, we present a novel biomedical literature mining algorithms based on this philosophy with significant extensions. Our objective is to extend the existing known virus list compiled by CDC or bacterium recognized by domain experts as BTW to other viruses/ bacteria that might

have similar characteristics. We thus hypothesize that viruses/bacteria that have been researched with respect to the characteristics possessed by existing viruses are leading candidates for extending the virus/bacterium lists. Our findings are intended as a guide to the according literature to support further studies that might then lead to appropriate defense and public health measures.

3 Background of Virus and Bacterium

Before initiating suspicious viruses and bacteria mining systems, we should identify what biological agents could be used as biological weapons.

3.1 Virus

Geissler identified and summarized 13 criteria (shown in Table 1) to identify biological warfare agents as viruses [3]. Based on the criteria, he compiled 21 viruses. Figure 1 lists the 21 virus names in MeSH terms. The viruses in Figure 1 meet some of the criteria described in Table 1.

▪ Hemorrhagic Fever Virus, Crimean-Congo	▪ Encephalitis Virus, Eastern Equine	
▪ Lymphocytic choriomeningitis virus	▪ Encephalitis Virus, Japanese	
▪ Encephalitis Virus, Venezuelan Equine	▪ Encephalitis Viruses, Tick-Borne	
▪ Encephalitis Virus, Western Equine	▪ Encephalitis Virus, St. Louis	
▪ Arenaviruses, New World	▪ Chikungunya virus	▪ Hepatitis A virus
▪ Marburg-like Viruses	▪ Dengue Virus	▪ Orthomyxoviridae
▪ Rift Valley fever virus	▪ Ebola-like Viruses	▪ Junin virus
▪ Yellow fever virus	▪ Hantaan virus	▪ Lassa virus
		▪ Variola virus

Fig. 1. Geissler's 21 Viruses

Based on the criteria, government agencies such as CDC and the Department of Homeland Security compile and monitor viruses which are known to be dangerous in bio-terrorism.

3.2 Bacterium

There are known 13 bacteria that can cause deadly disease. For example, anthrax is an acute infectious disease caused by the spore-forming bacterium *Bacillus anthracis*. Anthrax most commonly occurs in wild and domestic lower vertebrates (cattle, sheep, goats, camels, antelopes, and other herbivores), but it can also occur in humans when they are exposed to infected animals or to tissue from infected animals or when anthrax spores are used as a bioterrorist weapon. Q fever is a zoonotic disease caused

by *Coxiella burnetii*, a species of bacteria that is distributed globally. *Coxiella burnetii* is a highly infectious agent that is rather resistant to heat and drying. It can become airborne and inhaled by humans. A single *C. burnetii* organism may cause disease in a susceptible person. This agent could be developed for use in biological warfare and is considered a potential terrorist threat. For other deadly diseases caused by bacteria, please refer table 1.

Table 1. Bacteria used in biological warfare

Bacteria name (caused disease)	Bacteria name (caused disease)
Bacillus anthracis (anthrax)	Francisella tularensis (tularemia)
Clostridium botulinum (botulism)	Burkholderia mallei, Burkholderia pseudomallei (glanders)
Brucella melitensis, Brucella abortus, Brucella suis (brucellosis)	Coxiella burnetii (Q fever)
Vibrio cholerae (cholera)	Salmonella (Salmonellosis, typhoid fever)
Yersinia pestis (plague)	Shigella dysenteriae (shigellosis)

4. Method

MedMeSH Summarizer [8] summarizes a group of genes by filtering the biomedical literature and assigning relevant keywords describing the functionality of a group of genes. Each Gene cluster contains N genes, while each gene has a set of terms associated with it. A co-occurrence matrix is thus built, with number of citations associated with the gene and containing the mesh term as the cell value. Based on this matrix and some statistical information, they made overall relevance ranking for all the terms describing the topic of certain cluster of genes. There are 630 bacteria defined in PubMed database. We found it quite reasonable to extract topical terms for known 13 bacteria and then use these terms to look for suspicious remaining bacteria.

- **Normalized term bacterium matrix:** $\tilde{f}_{ij} = F_{ij} / (\sum_{i=1}^M F_{ij})^\alpha$ ($0 \leq \alpha \leq 1$) ①

where F_{ij} is a term by bacteria matrix. It means that how many PUBMED documents retrieved by bacterium j contains term i .

- **Relevance Ranking:**

1. Cluster Topics (Major): Terms occur in most bacteria with high frequency. Criterion $R1$: Rank the MeSH terms by decreasing order of the means μ_j .
2. Cluster Topics (Minor): Terms occur in most bacteria with low frequency. Compute $\sigma_i = \sqrt{(\sum_{j=1}^N (\tilde{f}_{ij} - \mu_i)^2) / N}$. Criterion $R2$: Rank the MeSH terms by decreasing order of the ratios μ_j / σ_i 's.
3. Particular Topics: Terms occur in a few bacteria with high frequency. Criterion $R3$: Rank the MeSH terms by decreasing order of the ratios σ_i^2 / μ_j 's.
4. Each MeSH term in Ω is ranked based on each of the above three criteria. The terms were then given an overall relevance rank R where:

$R = wR_1 + ((1-w)/2)R_2 + ((1-w)/2)R_3$ ② The weight parameter in formula

② has been assigned so that the major topics are given weight w being the most important set of terms in providing a summary of the cluster. The remaining weight $1 - w$ is divided equally between the minor topics and the particular topics. In our system, w is set to 0.5 because we look for more topical terms of the whole bacteria cluster.

▪ **Procedure of algorithm**

1. Submit query “bacteria name [major]” to the pubmed and download Mesh term after applying stop word list for each biological agent. We download documents of 13 existing bacteria. Our stop word list is composed of MeSH terms extracted from PubMed documents (1994-2004) by their overall usage. For example, some MeSH term without biomedical meaning is used very frequently such as “Government Supported”.
2. Build a normalized matrix (①) of terms by bacterium (13 bacteria).
3. Rank all the terms according to formula② and pick top k terms.
4. Download the documents of the remaining 617 bacteria. And use terms above to build a matrix of terms by bacteria (617 bacteria) (①).
5. Let the rank value of term be R_i . We use formula $R^B = \sum_{i=1}^M \tilde{f}_{ij} \times R_i$ to rank each bacterium.

5 Experimental Results

We apply our method to two data sets: viruses and bacteria. As for space, we only list the result of bacteria. Table 2 displays the top ranked bacteria by R^B criteria.

Table 2. Ranked bacterium

	Top 1-10	weight		Top 11-20	weight
1	Clostridium tetani	38.8	11	Brucellaceae	23.68
2	Erysipelothrix	36.96	12	Campylobacter fetus	22.74
3	Coxiellaceae	31.57	13	Yersinia enterocolitica	21.95
4	Sarcina	31.27	14	Bacillus thuringiensis	21.24
5	Yersinia pseudotuberculosis	28.16	15	Pediococcus	21.2
6	Atypical Bacterial	26.41	16	Mycobacterium bovis	20.36
7	Corynebacterium diphtheriae	26.22	17	Proteus vulgaris	20.23
8	Photobacterium	26.13	18	Haemophilus influenzae-b	19.89
9	Brucella	24.9	19	Nocardia asteroides	19.88
10	Haemophilus ducreyi	24.69	20	Bacillus megaterium	19.69

6 Potential Significance for Public Health and Homeland Security

This work is critical to public health and homeland security. Our nation is spending alone this year just in disbursements to states, territories and local health over a billion dollars to prepare for terrorism including such efforts as building public health capacity, disease surveillance and laboratory notification. [9] However, without the ability to prioritize these resources which have improved public health capacity and laboratory capacity we cannot further improve both national and international preparedness efforts. [10] In 1999 the Department of Defense was involved in building a directory of known emerging infectious diseases and laboratory tests worldwide and identified approximately 40 high threat agents for bio-terrorism including many of the hemorrhagic viruses. [11] However since that time we have had the emergence of SARS, Avian Flu virus and many other threats to the public health. We must be prepared and without continued work such as this to identify additional threats, the preparedness efforts may fall short.

References

1. SdfsdfBüchen-Osmond C. Taxonomy and Classification of Viruses. In: Manual of Clinical Microbiology, ASM Press, Washington DC, 8th ed, Vol 2, p. 1217-1226, 2003
2. DiGiacome, R.A, Kremer, J.M. and Shah, D.M. Fish oil dietary supplementation is patients with Raynaud's phenomenon: A double-blind, controlled, prospective study, American Journal of Medicine, 158-164m, 8, 1989.
3. Geissler, E. (Ed.), Biological and toxin weapons today, Oxford, UK: SIPRI (1986)
4. Swanson, DR, Smalheiser NR, & Bookstein A. Information discovery from complementary literatures: categorizing viruses as potential weapons. JASIST 52(10): 797-812 , 2001
5. Swanson, DR., Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30(1), 7-18, 1986
6. Swanson, DR., Undiscovered public knowledge. Libr. Q. 56(2), pp. 103-118 1986
7. Hu X., I. Yoo. P. Rumm, M. Atwood., Mining Candidate Viruses as Potential Bio-Terrorism Weapons from Biomedical Literature, in 2005 IEEE International Conference on Intelligence and Security Informatics (IEEE ISI-2005), Atlanta, Georgia, May 19-20, 2005
8. P. Kankar, S. Adak, A. Sarkar, K. Murari, K. and G. Sharma. "MedMeSH Summarizer: Text Mining for Gene Clusters", in the Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, 2002
9. Guidance on cooperative agreements from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and the Human Resource Service Administration. Accessible at www.bt.cdc.gov
10. Rumm P.D. Bioterrorism preparedness: potential threats remain. Am J Public Health. 2005 Mar;95(3):372 (comment on previous article)
11. Rumm P, Gaydos J, Mansfield J and Kelley P, A Department of Defense (DOD) Virtual Public Health Laboratory Directory, *Mil Med*, 2000;Jul,165-Supp. 2):73.