CHAPTER OVERVIEW

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 bio-terrorism weapons from biomedical literature. We first take text mining approach to identify topical terms of existed viruses (bacteria) from PubMed separately. Then, we apply a text mining method bridge these terms as instances with the remaining viruses (bacteria) and thus to discover how much these terms describe the remaining viruses (bacteria). In the end, we build an 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.
Chapter 18. Hu, et al.

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 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. The rest of the paper is organized as follows. Section 2 briefly discusses the relevant works. Section 3 describes the background information of virus and bacteria as biological agent. Section 4 discusses our method in detail. The experimental results are presented in Section 5.
Potential significance for public health and homeland security are discussed in Section 6.

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 [9]. 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 [9] [10]. This novel hypothesis was later clinically confirmed by DiGiacomo in 1989 [2]. Later on [11] 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 to other viruses that might have similar characteristics. We hypothesize, therefore, that viruses that have been researched with respect to the characteristics possessed by existing viruses are leading candidates for extending the virus lists. Our findings are intended as a guide to the virus literature to support further studies that might then lead to appropriate defense and public health measures. In our former work [5], we let human experts to define the key words that help find viruses that can be potential biological weapons. In this paper, we will provide a text data mining approach to target the terms that help identify potential weapons and to rank the viruses according these terms.

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 [3] identified and summarized 13 criteria (shown in Table 18-1) to identify biological warfare agents as viruses. Based on the criteria, he compiled 21 viruses. Table 18-2 lists the 21 virus names in MeSH terms. The viruses in Table 18-2 meet some of the criteria described in Table 18-1.

Table 18-1. Geissler’s 13 Criteria for Viruses

|   |                                                                                           |
|---|-------------------------------------------------------------------------------------------|
| 1 | The agent should consistently produce a given effect: death or disease.                    |
| 2 | The concentration of the agent needed to cause death or disease the infective dose should be low. |
| 3 | The agent should be highly contagious.                                                     |
| 4 | The agent should have a short and predictable incubation time from exposure to onset of the disease symptoms. |
| 5 | The target population should have little or no natural or acquired immunity or resistance to the agent. |
| 6 | Prophylaxis against the agent should not be available to the target population.            |
| 7 | The agent should be difficult to identify in the target population, and little or no treatment for the disease caused by the agent should be available. |
| 8 | The aggressor should have means to protect his own forces and population against the agent clandestinely. |
| 9 | The agent should be amenable to economical mass production.                                |
|10 | The agent should be reasonably robust and stable under production and storage conditions, in munitions and during transportation. Storage methods should be available that prevent gross decline of the agent’s activity. |
|11 | The agent should be capable of efficient dissemination. If it cannot be delivered via an aerosol, living vectors (e.g. fleas, mosquitoes or ticks) should be available for dispersal in some form of infected substrate. |
|12 | The agent should be stable during dissemination. If it is to be delivered via an aerosol, it must survive and remain stable in air until it reaches the target population. |
|13 | After delivery, the agent should have low persistence, surviving only for a short time, thereby allowing a prompt occupation of the attacked area by the aggressor’s troops. |

Table 18-2. Geissler’s 21 Viruses

| Hemorrhagic Fever Virus, Crimean-Congo | Encephalitis Virus, Eastern Equine |
|---------------------------------------|-----------------------------------|
| Lymphocytic choriomenigitis 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                 |
| Marburg-like Viruses                  | Dengue Virus                      |
| Rift Valley fever virus               | Ebola-like Viruses                |
| Yellow fever virus                    | Hantaan virus                     |
|                                       | Hepatitis A virus                 |
|                                       | Orthomyxoviridae                  |
|                                       | Junin virus                       |
|                                       | Lassa virus                       |
|                                       | Variola virus                     |
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 some bacteria (by the time we examine, there are 13) that 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 18-3.

Table 18-3. Bacteria used in biological warfare

| Bacteria name                      | Disease caused       |
|------------------------------------|----------------------|
| Bacillus anthracis                 | Anthrax              |
| Clostridium botulinum              | Botulism             |
| Brucella melitensis, Brucella abortus, Brucella suis | Brucellosis        |
| Vibrio cholerae                    | Cholera              |
| Yersinia pestis                    | Plague               |
| Francisella tularensis             | Tularemia            |
| Burkholderia mallei, Burkholderia pseudomallei | Glanders         |
| Coxiella burnetti                  | Q fever              |
| Salmonella                         | Salmonellosis, typhoid fever |

4. METHOD

MedMeSH Summarizer [6] 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, using the number of citations associated with the gene and containing the mesh term. Based on this matrix and some statistical
information, overall relevance rankings were made for all the terms describing the topic of certain cluster of genes. There are 487 viruses known to us in PubMed database. We found it is quite reasonable to take the 21 viruses (biological weapons) as a cluster of viruses and apply the method discussed above to discover and thereby rank the terms that describes these viruses. We then take the remaining 466 viruses as another cluster and then build a matrix of terms (from 21 known viruses) by viruses (466 viruses) and thus rank all the 466 viruses through a ranking formula. We suspect that the higher the virus rank, the more likely the virus will be bio-terrorism weapon. Similarly, there are 630 bacteria defined in PubMed database. As mention above, we apply the same methodology to the existed 13 bacteria and the remaining 617 bacteria. For clear statements, we only take virus as an example to introduce our algorithm. However, we will introduce the experiment results of both virus and bacteria.

- **Virus Cluster:**

  Let \( V = \{V_1, V_2, ..., V_N\} \) be the given cluster containing \( N \) viruses, where \( V_j \) will be used to denote the \( J^{th} \) virus in the cluster.

- **MeSH Term List:**

  Let \( \Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_N \), where \( \Omega_j \) is the set of MeSH terms associated with the virus \( V_j \) \((j = 1, 2, ..., N)\) (after MeSH stop word filtering). Moreover, let \( \Omega = \{T_1, T_2, ..., T_N\} \), where \( T_i \) \((i = 1, 2, ..., M)\) denote the MeSH terms associated with the virus in the cluster.

- **Matrix:**

  Let \( F = ((F_{ij}))_{M \times N} \) be the co-occurrence matrix, where \( F_{ij} = \) Number of citations that are associated with the virus \( V_j \) by the PubMed database and contain the MeSH term \( T_i \) \((i = 1, 2, ..., M; j = 1, 2, ..., N)\).

- **Normalization by Virus Relevance:**
There are two contradicting requirements for normalization: dominant viruses in cluster should not highly skew results in their favor; some weight should be given to the fact that the virus is well studied. To achieve this normalized frequency of the MeSH term, \( T_i \) for virus \( V_j \) is computed as

\[
\tilde{f}_{ij} = F_{ij} / \left( \sum_{i=1}^{M} F_{ij} \right) ^{\alpha} (0 \leq \alpha \leq 1)
\]

Equation 18-2

Based on experiment results of MedMeSH Summarizer, the default value of \( \alpha \) in our system is also 0.67. Now each MeSH term \( T_i \in \Omega \), is characterized by the MeSH feature vector \( \tilde{f}_i = (\tilde{f}_{i1}, \tilde{f}_{i2}, \ldots, \tilde{f}_{iN}) \), where \( \tilde{f}_{ij} (i = 1,2,\ldots,M; j = 1,2,\ldots,N) \) are the normalized frequencies described above.

Overall Relevance Ranking:
1. **Cluster Topics (Major):** These are MeSH terms that are “commonly” associated with almost all viruses in the cluster and hence likely to have a high total frequency of occurrence. For this, the MeSH terms are ranked by the mean of their virus distribution feature vectors as follows:
   - Compute
   \[
   \mu_j = \left( \sum_{j=1}^{N} \tilde{f}_{ij} \right) / N (i = 1,\ldots,M).
   \]
   - Ranking Criterion \( R1 \): Rank the MeSH terms by decreasing order of the means \( \mu_j \).
2. **Cluster Topics (Minor):** These are MeSH terms which had moderate-to-low total frequency but still appear with most of the genes. This type of terms is expected to have moderate mean and low variance. For this, the MeSH terms are ranked by the ratio of mean/standard deviation of their MeSH feature vectors as follows:
   - Compute
   \[
   \sigma_i = \sqrt{\left( \sum_{j=1}^{N} (\tilde{f}_{ij} - \mu_j) ^2 \right) / N}, (i = 1,\ldots,M).
   \]
   - Ranking Criterion \( R2 \): Rank the MeSH terms by decreasing order of the ratios
\[ \mu_j / \sigma_i \]’s.

3. **Particular Topics:**

These are MeSH terms that were not related to the whole cluster but were strongly associated with a subgroup of the cluster. This type of terms is expected to have high variance and moderate-to-low mean. For this, the MeSH terms are ranked by the ratio of variance/mean of their MeSH feature vectors as follows:

- **Ranking Criterion R3:** Rank the MeSH terms by decreasing order of the ratios

\[ \sigma_i^2 / \mu_j \]’s.

4. Each MeSH term in \( \Omega \) is ranked based on each of the above three criteria. The terms were then given an overall relevance rank \( R \) where:

\[
R = wR_1 + \frac{1-w}{2}R_2 + \frac{1-2}{2}R_3
\]

*Equation 18-3*

5. The weight parameter in Equation 18-3 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. The default weights in the system are: \( w = 0.50 \) for the first ranking criterion and 0.25 each for the second and third criteria.

- **Procedure of algorithm**

1. Submit query “virus name [MeSH]” to the pubmed and download Mesh term after applying stop word list for each biological agent. We download documents of 21 known viruses. (MeSH term is the subjective terms presented by human experts for each document) We take each virus as a category. Our stop word list is composed of MeSH terms extracted from PubMed documents (1994-2004) by their overall usage. For example, some MeSH terms are used very frequently such as “English Abstract”, “Government Supported”, “Non Government Supported” and so on, and these terms have nothing to do with our viruses and bacterium mining.

2. Build a matrix \( F \) (Equation 18-1) of terms by viruses (21 viruses) and then normalize it through Equation 18-2.
3. Rank all the terms according to Equation 18-3 and pick top k terms.
4. Download the documents of the remaining 466 viruses. And use terms above to build a matrix $F$ of terms by viruses (466 viruses) (Equation 18-1). Normalize the matrix by Equation 18-2.
5. Let the rank value of term be $R_i (i = 1,2,\ldots,M)$. $R_i$ is the rank value of term in the term by viruses (466 viruses) matrix. Eq.

$$
R^V = \sum_{i=1}^{M} \tilde{F}_{ij} \times R_i (i = 1,2,\ldots,M; j = 1,2,\ldots,N)
$$

Equation 18-4

is used to rank each remaining virus marked as Rank1. We also rank virus using $R_i$ from term by viruses (21 viruses) matrix marked as Rank2.
5. EXPERIMENTAL RESULTS

We apply our method to two data sets: viruses and bacteria. Section 5.1 lists the experiment results of virus, while section 5.2 is for bacteria. Table 18-6 to 18-9 displays the top ranked topical terms and suspicious viruses by $R^v$ criteria (rank1 and rank2 respectively). Accordingly, Table 18-12 to 18-15 show the top ranked topical terms and bacteria by $R^v$ criteria (rank1 and rank2 respectively). From the results, there is a big match between viruses/bacteria names and their associated diseases and topical terms. Take bacteria as an example, 12 out of 13 known bacteria names were ranked within top 50 terms in Table 18-12. Moreover, most of disease names caused by the 13 bacteria were also matched in the table. For the potential significance of suspicious viruses/bacteria that we detected, please refer to section 6.

5.1 Experiment Results of Suspicious Viruses Mining

Table 18-4. The search keywords (21 Virus names) and the according number of Documents downloaded

| Search Keywords                                      | # of Doc. |
|------------------------------------------------------|-----------|
| "Chikungunya virus"[MeSH]                            | 397       |
| "Hemorrhagic Fever Virus, Crimean-Congo"[MeSH]       | 202       |
| "Encephalitis Virus, Eastern Equine"[MeSH]           | 292       |
| ...                                                   | ...       |
| Total                                                | 31080     |

Table 18-5. The search keywords (466 Virus names) and the according number of Documents downloaded

| Search Keywords                                      | # of Doc. |
|------------------------------------------------------|-----------|
| "Abelson murine leukemia virus"[MeSH]                | 416       |
| "Dependovirus"[MeSH]                                 | 1874      |
| "Adenoviridae"[MeSH]                                 | 20178     |
| ...                                                   | ...       |

Table 18-6. Top ranked topical terms by rank1

| Rank1 | Top 1-25 terms               | Weight | Top 26-50 terms               | Weight |
|-------|------------------------------|--------|------------------------------|--------|
| 1     | blood-borne pathogens        | 1.47   | infectious anemia virus, equine | 1.03   |
| 2     | transmissible gastroenteritis virus | 1.45 | classical swine fever virus  | 1.02   |
| 3     | hepatitis e virus            | 1.42   | needlestick injuries         | 1.01   |
| Rank | Top 1-25 terms | Weight | Top 26-50 terms | Weight |
|------|----------------|--------|----------------|--------|
| 4    | herpesvirus 1, equid | 1.4    | visna-maedi virus | 1.01   |
| 5    | influenza a virus, porcine | 1.37   | rhinovirus | 1      |
| 6    | hepatitis e | 1.34 | african swine fever virus | 1      |
| 7    | muromegalovirus | 1.32 | ectromelia virus | 0.98 |
| 8    | bacteriophage mu | 1.32 | lactate dehydrogenase-elevating virus | 0.96 |
| 9    | rauscher virus | 1.24 | borna disease | 0.96 |
| 10   | mycobacterium | 1.2 | hepatitis a virus, human | 0.94 |
| 11   | staphylococcus aureus | 1.17 | staphylococcal infections | 0.94 |
| 12   | bacillus phages | 1.16 | encephalitis virus, california | 0.94 |
| 13   | rift valley fever virus | 1.15 | mammary tumor virus, mouse murine hepatitis virus | 0.92 |
| 14   | hemorrhagic fever, ebola viruses, unclassified | 1.11 | bluetongue virus | 0.87 |
| 15   | herpesvirus 1, cercopithecine endogenous retroviruses | 1.11 | bacteriophage phi x 174 | 0.86 |
| 16   | mengovirus | 1.1 | immunodeficiency virus, feline arboviruses | 0.86 |
| 17   | salmonella phages | 1.1 | staphylococcus | 0.86 |
| 18   | rift valley fever | 1.07 | mice minute virus | 0.85 |
| 19   | influenza virus c sarcoma virus, woolly monkey | 1.07 | phlebovirus | 0.84 |
| 20   | hepatitis delta virus | 1.06 | transfusion-transmitted virus | 0.84 |
| 21   | ebola-like viruses | 1.05 | norwalk virus | 0.84 |
| 22   | maus elberfeld virus | 1.04 | monkeypox virus | 0.83 |

Table 18-7. Top ranked viruses by rank1

| Rank | Top 1-25 viruses | Weight | Top 26-50 viruses | Weight |
|------|-----------------|--------|-----------------|--------|
| 1    | Blood-Borne Pathogens | 45.77 | Hepatitis Delta Virus | 17.52 |
| 2    | Filoviridae | 32.31 | Rubulavirus | 16.94 |
| 3    | Phlebovirus | 28.73 | Herpesvirus 1, Equid | 16.94 |
| 4    | Hepatitis E virus | 28.13 | Salmonella Phages | 16.75 |
| 5    | Hepatovirus | 27.31 | Visna-maedi virus | 16.62 |
### Table 18-8. Top ranked topical terms by rank2

| Rank2 | Top 1-25 terms                          | Weight | Top 26-50 terms                                                                 | Weight |
|-------|----------------------------------------|--------|--------------------------------------------------------------------------------|--------|
| 1     | variola virus                          | 2.06   | yellow fever                                                                   | 1.27   |
| 2     | lymphocytic choriomeningitis virus      | 1.84   | hepatitis antibodies                                                           | 1.27   |
| 3     | arenaviruses, new world                | 1.78   | dengue virus                                                                   | 1.26   |
| 4     | hepatitis a virus, human               | 1.73   | hepatitis a antibodies                                                          | 1.26   |
| 5     | hepatitis a                            | 1.72   | viral hepatitis                                                                 | 1.26   |

### Chapter 18. Hu, et al.

| Rank1 | Top 1-25 viruses | weight | Top 26-50 viruses | weight |
|-------|------------------|--------|-------------------|--------|
| 6     | Bunyaviridae     | 22.48  | Togaviridae       | 16.61  |
| 7     | Hantavirus       | 22.37  | Echoviridae       | 16.47  |
| 8     | Staphylococcus Phages | 22.13 | Alphavirus         | 16.32  |
| 9     | Influenza A Virus, Porcine | 21.96 | Distemper Virus, Canine | 16.19 |
| 10    | Arenaviridae     | 21.55  | Rhinovirus         | 16.13  |
| 11    | Hepatitis A Virus, Human | 21.53 | Rubella virus      | 16.08  |
| 12    | Orthobunyavirus  | 21.52  | Mammary Tumor Virus, Mouse Herpesvirus 3, Human | 16.03 |
| 13    | Arboviruses      | 21.52  | Rubivirus          | 15.86  |
| 14    | Arenavirus       | 21.3   | Classical swine fever virus         | 15.85  |
| 15    | Viruses, Unclassified | 20.62 | Picornaviridae     | 15.71  |
| 16    | Encephalitis Virus, California | 20.49 |       |        |
| 17    | Arenaviruses, Old World                   | 20.05 | Lyssavirus         | 15.61  |
| 18    | Herpesvirus 1, Suid | 18.67 | Mycobacteriophage s | 15.55  |
| 19    | Rauscher Virus | 18.65  | Muromegalovirus    | 15.49  |
| 20    | Transmissible gastroenteritis virus      | 18.58 | Poliovirus         | 15.44  |
| 21    | Encephalitis Viruses | 18.2  | Norwalk virus       | 15.42  |
| 22    | Influenzavirus A | 18.05 | Parainfluenza Virus 3, Human | 15.4   |
| 23    | Influenza A virus | 17.92 | Orbivirus          | 15.39  |
| 24    | Flavivirus       | 17.87  | Norovirus          | 15.38  |
| 25    | Mumps virus      | 17.59  | Rabies virus       | 15.36  |
| Rank | Top 1-25 terms                          | Weight | Top 26-50 terms                          | Weight |
|------|----------------------------------------|--------|-----------------------------------------|--------|
| 6    | chikungunya virus                      | 1.68   | vaccines                                | 1.25   |
|      | influenza a virus                      |        | encephalitis virus                      | 1.2    |
| 7    | encephalitis viruses, tick-borne       | 1.64   | lassa virus                             | 1.2    |
|      |                                        |        | encephalitis, tick-borne                | 1.19   |
| 8    | smallpox                               | 1.63   | hemorrhagic fever                       | 1.19   |
| 9    | encephalitis, tick-borne               | 1.63   | influenza, avian                        | 1.18   |
|      |                                        |        | hemorrhagic fever with renal syndrome   | 1.17   |
|      |                                        |        | hemorrhagic fever, crimean-congo        | 1.16   |
| 10   | hepatitis a virus                      | 1.62   | encephalitis virus                      | 1.14   |
| 11   | yellow fever virus                     | 1.6    | dengue                                  | 1.12   |
|      |                                        |        | encephalitis, eastern equine            | 1.11   |
|      |                                        |        | ixodes                                  | 1.11   |
| 12   | encephalitis virus, japanese           | 1.57   | hemorrhagic fever                       | 1.17   |
|      |                                        |        | influenza b virus                       | 1.16   |
| 13   | influenza                               | 1.54   |                                       |        |
| 14   | encephalitis virus, venezuelan equine  | 1.52   | dengue                                  | 1.12   |
|      |                                        |        | encephalitis, eastern equine            | 1.11   |
|      |                                        |        | ixodes                                  | 1.11   |
| 15   | hantaan virus                          | 1.49   |                                        |        |
| 16   | ebola-like viruses                     | 1.47   |                                        |        |
|      |                                        |        |                                        |        |
| 17   | hemorrhagic fever, american            | 1.47   |                                        |        |
|      | lymphocytic choriomeningitis           | 1.45   |                                        |        |
|      |                                        |        |                                        |        |
| 18   | rift valley fever virus                | 1.44   |                                        |        |
| 19   | hepatitis a vaccines                  | 1.4    |                                        |        |
|      |                                        |        |                                        |        |
| 20   | encephalitis, japanese                | 1.38   |                                        |        |
|      |                                        |        |                                        |        |
| 21   | rift valley fever                      | 1.37   |                                        |        |
| 22   | hepatitis a virus                     | 1.4    |                                        |        |
|      |                                        |        |                                        |        |
| 23   | smallpox vaccine                       | 1.37   |                                        | 1      |
|      |                                        |        | lassa fever                             | 0.93   |
| Rank2 | Top 1-25 terms                  | Weight | Top 26-50 terms          | Weight |
|-------|--------------------------------|--------|--------------------------|--------|
| 24    | hemorrhagic fever, ebola       | 1.27   | neuraminidase             | 0.9    |
| 25    | influenza a virus, avian       | 1.27   | arenaviridae              | 0.9    |

Table 18-9. Top ranked viruses by rank2

| Rank2 | Top 1-25 viruses | weight  | Top 26-50 viruses | weight |
|-------|-----------------|---------|-------------------|--------|
| 1     | Hepatovirus     | 62.38   | Hepatitis E virus | 33.34  |
| 2     | Arenaviridae    | 60.56   | Influenza A Virus, Porcine Poxviridae | 33.29  |
| 3     | Arenavirus      | 59.56   | Encephalitis Virus, California Flaviviridae Virus, Unclassified Picornaviridae | 32.56  |
| 4     | Arenaviruses, Old World | 58.77 | | 32.18  |
| 5     | Filoviridae     | 56.96   | | 32.03  |
| 6     | Flavivirus      | 51.46   | | 31.56  |
| 7     | Encephalitis Viruses | 49.86 | | 31.03  |
| 8     | Hepatitis A Virus, Human | 47.99 | | 30.01  |
| 9     | Blood-Borne Pathogens | 47.3  | | 29.7   |
| 10    | Influenza A virus | 46.95  | West Nile virus | 29.66  |
| 11    | Influenzavirus A | 46.81  | Vesicular stomatitis-Indiana virus | 29.59  |
| 12    | Phlebovirus     | 45.73   | Poliovirus          | 29     |
| 13    | Arboviruses     | 42.59   | Norovirus           | 28.94  |
| 14    | Bunyaviridae    | 39.7    | Polioviruses        | 28.92  |
| 15    | Alphavirus      | 39.1    | Gross Virus         | 28.55  |
| 16    | Hantavirus      | 38.68   | Adenoviridae        | 28.54  |
| 17    | Influenza A Virus, Human | 38.3  | Nairovirus | 28.4   |
| 18    | Togaviridae     | 37.31   | Respirovirus        | 28.07  |
| 19    | Orthopoxvirus   | 36.58   | Semliki forest virus | 27.87  |
| 20    | Encephalitis Viruses, Japanese | 36.21 | | 27.56  |
| 21    | Influenzavirus B | 35     | Norwalk virus       | 27.5   |
| 22    | Influenza A Virus, Avian | 34.68  | Caliciviridae       | 27.5   |
5.2 Experiment Results of Suspicious Bacteria Mining

Table 18-10. The search keywords (13 bacteria names) and the according number of downloaded documents.

| Search Keywords                        | # of Doc. |
|----------------------------------------|-----------|
| "Bacillus anthracis" [major]          | 1153      |
| "Clostridium botulinum" [major]       | 1191      |
| "Brucella melitensis" [major]         | 391       |
| "Brucella abortus" [major]            | 1415      |
| "Brucella suis" [major]               | 18        |
| "Vibrio cholerae" [major]             | 3503      |
| "Yersinia pestis" [major]             | 1323      |
| "Francisella tularensis" [major]      | 621       |
| "Burkholderia mallei" [major]         | 19        |
| "Burkholderia pseudomallei" [major]   | 443       |
| "Coxiella burnetti" (No major topic)  | 172       |
| "Salmonella" [major]                  | 21677     |
| "Shigella dysenteriae" [major]        | 687       |
| Total                                  | 32613     |

Table 18-11. The search keywords (617 bacteria name) and the according number of documents downloaded.

| Search Keywords                        | # of Doc. |
|----------------------------------------|-----------|
| "Acetobacter"[major]                   | 279       |
| Acetobacteraceae" [major]              | 543       |
| "Acetobacterium"[major]                | 4         |
| ...                                    | ...       |

Table 18-12. Top ranked topical terms by rank1

| Rank | Term                              | Weight | Rank | Term                              | Weight |
|------|-----------------------------------|--------|------|-----------------------------------|--------|
| 1    | erysipeloethrix                   | 1.69   | 26   | leuconostoc                       |        |
| 2    | sarcina                           | 1.65   | 27   | leptospira interrogans            |        |
| 3    | campylobacter fetus               | 1.6    | 28   | bacillus megaterium               |        |
| 4    | yersinia pseudotuberculosis       | 1.5    | 29   | nocardia asteroides              |        |
| Rank | Bacterium                          | Value | Rank | Bacterium                          | Value |
|------|-----------------------------------|-------|------|-----------------------------------|-------|
| 1    | **Clostridium tetani**            | 38.8  | 26   | **Mycobacterium avium**            | 18.69 |
| 2    | **Erysipelothrix**                | 36.96 | 27   | **Treponema pallidum**             | 18.58 |
| 3    | **Coxiellaceae**                  | 31.57 | 28   | **Vibrioaceae**                   | 18.43 |
| 4    | **Sarcina**                       | 31.27 | 29   | **Vibrio**                        | 18.41 |
| 5    | **Yersinia pseudotuberculosis**   | 28.16 | 30   | **Clostridium difficile**          | 18.26 |
| 6    | **Atypical Bacterial Forms**      | 26.41 | 31   | **Bacillus stearothermophilus**    | 18.18 |
| 7    | **Corynebacterium diphtheriae**   | 26.22 | 32   | **Escherichia coli O157**          | 18.01 |
| 8    | **Photobacterium**                | 26.13 | 33   | **Erwinia**                       | 18.01 |
| 9    | **Brucella**                      | 24.9  | 34   | **Propionibacterium acnes**        | 17.9  |
| 10   | **Haemophilus ducreyi**           | 24.69 | 35   | **Lactobacillus casei**            | 17.88 |
| 11   | **Brucellaceae**                  | 23.68 | 36   | **Chromobacterium**               | 17.83 |
| 12   | **Campylobacter fetus**           | 22.74 | 37   | **Bordetella pertussis**           | 17.79 |
| 13   | **Yersinia enterocolitica**       | 21.95 | 38   | **Lactobacillus acidophilus**      | 17.67 |
| 14   | **Bacillus thuringiensis**        | 21.24 | 39   | **Mannheimia haemolytica**         | 17.65 |
| 15   | **Pediococcus**                   | 21.2  | 40   | **Nocardia**                      | 17.63 |

**Table 18-13. Top ranked bacterium by rank1**
### Table 18-14. Top ranked topical terms by rank2

| Rank2 | Top 1-25 terms | weight | Top 26-50 terms | Weight |
|-------|----------------|--------|----------------|--------|
| 1     | vibrio cholerae | 3.69   | salmonella enteritis | 1.72   |
| 2     | brucella abortus | 3.36   | brucella vaccine | 1.65   |
| 3     | clostridium botulinum | 3.19   | brucellosis | 1.65   |
| 4     | bacillus anthracis | 3.01   | cholera vaccines | 1.63   |
| 5     | yersinia pestis | 3.01   | fleas | 1.5   |
| 6     | shigella dysenteriae | 2.81   | shigella sonnei | 1.42   |
| 7     | cholera | 2.72   | complement fixation tests | 1.36   |
| 8     | botulinum toxins | 2.42   | spores, bacterial | 1.31   |
| 9     | salmonella typhimurium | 2.4   | shigella flexneri | 1.28   |
| 10    | anthrax | 2.38   | salmonella food poisoning | 1.28   |
| 11    | francisella tularensis | 2.32   | shiga toxins | 1.27   |
| 12    | plague | 2.27   | mutagens | 1.27   |
| 13    | botulism | 2.13   | brucella | 1.2   |
| 14    | dysentery, bacillary | 2.1   | food microbiology | 1.2   |
| 15    | brucellosis, bovine | 2.1   | shigella boydii | 1.2   |
| 16    | cholera toxin | 2.09   | escherichia coli o157 | 1.18   |
| 17    | burkholderia | 2.09   | fimbriae | 1.18   |
| Rank2 | Top 1-25 terms | weight | Top 26-50 terms | Weight |
|-------|----------------|--------|----------------|--------|
| 18    | pseudomallei tularemia | 2.03 | proteins drug resistance, bacterial drug resistance, | 1.16 |
|       |                  | 43     | multiple, bacterial anthrax vaccines |        |
| 19    | salmonella      | 2.01 | drug resistance, drug resistance, | 1.15 |
|       |                  | 44     | multiple, bacterial anthrax vaccines |        |
| 20    | salmonella infections, animal | 1.99 | anthrax vaccines | 1.15 |
|       |                  | 45     |                |        |
| 21    | salmonella enterica | 1.95 | bioterrorism | 1.15 |
|       |                  | 46     |                |        |
| 22    | salmonella infections | 1.9 | plague vaccine bursa of fabricius neurotoxins | 1.12 |
|       |                  | 47     |                |        |
| 23    | salmonella melioidosis | 1.86 | | 1.12 |
|       |                  | 48     |                |        |
| 24    | mutagenicity tests | 1.73 | | 1.12 |
|       |                  | 49     |                |        |
| 25    | brucella melitensis | 1.72 | | 1.11 |
|       |                  | 50     |                |        |

Table 18-15. Top ranked bacterium by rank2

| Rank2 | Top 1-25 Bacterium | Weight | Top 26-50 Bacterium | weight |
|-------|-------------------|--------|---------------------|--------|
| 1     | Brucella          | 82.21  | Endospore-Forming Bacteria | 49.17 |
| 2     | Brucellaceae      | 79.08  | Gram-Positive Endospore-Forming Rods | 49.05 |
| 3     | Clostridium tetani | 71.33  | Bacillaceae | 48.31 |
| 4     | Vibrio            | 70.21  | Vibrio parahaemoliticus Photobacterium | 48.27 |
| 5     | Vibrionaceae      | 67.1   | Photobacterium Campylobacter | 47.9 |
| 6     | Clostridium       | 61.73  | Proteobacterium Bacillus thuringiensis | 46.71 |
| 7     | Escherichia coli O157 | 59.79  | | 46 |
| 8     | Sarcina           | 58.5   | Bacillus thuringiensis | 45.65 |
### 18. Text Mining Biomedical Literature for Bio-terrorism Weapons

| Rank2 | Top 1-25 Bacterium | Weight | Top 26-50 Bacterium | weight |
|-------|--------------------|--------|---------------------|--------|
| 9     | Yersinia pseu dotuber culosis | 57.81  | Bacteria            | 45.51  |
| 10    | Enterobacter iaceae      | 57.76  | Gram-Negative Bacteria | 45.46  |
| 11    | Spores, Bacterial       | 57.54  | Lactobacillus acidophilus | 45.33  |
| 12    | Listeria               | 56.71  | Erysipelothrix       | 44.73  |
| 13    | Listeria monocytogenes  | 55.04  | Escherichia          | 44.56  |
| 14    | Burkholderia ceae       | 54.79  | Campylobacter jejuni | 44.51  |
| 15    | Mycobacterium bovis     | 54.7   | Escherichia coli     | 44.51  |
| 16    | Gram-Negative Facultatively Anaerobic Rods | 53.79 | Lactobacillus casei | 44.31  |
| 17    | Clostridium perfringens | 53.58  | Alphaproteobacteria diaphtheriae | 44.13  |
| 18    | Atypical Bacterial Forms | 53.31 | Pasteurella multocida | 43.25  |
| 19    | Bacillus cereus         | 50.56  | Corynebacterium diphtheriae | 43.21  |
| 20    | Gammaproteobacteria     | 50.55  | Mannheimia haemolytica | 43.11  |
| 21    | Probiotics              | 50.51  | Propionibacterium    | 43.01  |
| 22    | Yersinia enterocolitica | 50.3   | Mannheimia           | 42.87  |
| 23    | Gram-Positive Endospore-Forming Bacteria | 49.81 | Bifidobacterium | 42.7   |
| 24    | Propionibacterium acnes | 49.6   | Micrococcus          | 42.48  |
| 25    | Coxiellaceae           | 49.22  | Propionibacteriaceae | 42.35  |
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, territory and local health over a billion dollars to prepare for terrorism including such efforts as building public health capacity, disease surveillance and laboratory notification [4]. 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 [7]. 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 [8]. 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.

7. ACKNOWLEDGEMENTS

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

REFERENCES

1. Büchen-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. 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
5. 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
6. Hu X., Zhang X., Yoo, I., Atwood M., Rumm, P., A Text Mining Approach for Identifying Candidate Viruses as Potential Bio-terrorism Weapons, GESTS International Transaction on Compute Science and Engineering, Vol. 9, NO 1., July 2005
7. 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

8. Rumm P.D. Bioterrorism preparedness: potential threats remain. Am J Public Health. 2005 Mar;95(3):372 (comment on previous article)

9. 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.

10. Swanson, DR., Fish-oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30(1), 7-18, 1986

11. Swanson, DR., Undiscovered public knowledge. *Libr. Q.* 56(2), pp. 103-118 1986

12. Swanson, DR., Smallheiser NR, & Bookstein A. Information discovery from complementary literatures: categorizing viruses as potential weapons. *JASIST* 52(10): 797-812 , 2001

**SUGGESTED READINGS**

- 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.

- Hu X., Zhang X., Yoo, I., Atwood M., Rumm, P., A Text Mining Approach for Identifying Candidate Viruses as Potential Bio-terrorism Weapons, *GESTS International Transaction on Compute Science and Engineering*, Vol. 9, No. 1, July 2005.

- Swanson, D.R., Undiscovered public knowledge. *Libr. Q.* 56(2), pp. 103-118, 1986.

**ONLINE RESOURCES**

- Taxonomy and Classification of Viruses: [http://www.ncbi.nlm.nih.gov/ICTVdb/MCM8.pdf](http://www.ncbi.nlm.nih.gov/ICTVdb/MCM8.pdf)

- 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 [http://www.bt.cdc.gov/](http://www.bt.cdc.gov/)

- PUBMED: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=books](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=books)

**DISCUSSION QUESTIONS**

1. In our presented problem, we summarize all existed viruses/bacteria as a whole and try to identify topical terms crossing all different
viruses/bacteria related documents. What other techniques might help to summarize existing viruses/bacteria? How do you balance those terms against viruses/bacteria that have very few documents?

2. Given the weight of topical terms, what other techniques do you think can help target the most suspicious virus/bacteria?

3. Can the terms used to describe disease symptoms caused by viruses/bacteria help identify potential viruses/bacteria? How can these terms be extracted?

4. Describe three other problems that can be solved using the method presented in this chapter.