David J. Ecker,[1,4] Ph.D.; Rangarajan Sampath,[1] Ph.D.; Paul Willett,[1] B.S.; Vivek Samant,[1] Ph.D.; Christian Massire,[1] Ph.D.; Thomas A. Hall,[1] Ph.D.; Kumar Hari,[1] Ph.D.; John A. McNeil,[1] B.S.; Cornelia Büchen-Osmond,[2] Ph.D.; and Bruce Budowle,[3] Ph.D.

The Microbial Rosetta Stone Database: A Common Structure for Microbial Biosecurity Threat Agents

ABSTRACT: Infectious microorganisms are important to multiple communities engaged in biodefense and biosecurity, including the agencies responsible for health, defense, law enforcement, agriculture, and drug and food safety. Many agencies have created lists of high priority infectious microorganisms to prioritize research efforts or to formally control the possession and distribution of specific organisms or toxins. However, the biological classification of infectious microorganisms is often complex and ambiguous, leading to uncertainty and confusion for scientists involved in biosecurity work. To address this problem, we created a database, known as the Microbial Rosetta Stone, which resolves many of these ambiguities and includes links to additional information on the microbes, such as gene sequence data and scientific literature. Here we discuss the efforts to coordinate organism names from pathogen lists from various governmental agencies according to biological relatedness and show the overlap of high-priority organisms from multiple agencies. To our knowledge, this is the first comprehensive coordination of pathogens, synonyms, and correct taxonomic names. The organized tables and visual aids are freely available at http://www.microbialrosettastone.com. This website provides a single location where access to information on a broad range of disease-causing organisms and toxins is available to members of the biosecurity community.

KEYWORDS: forensic science, microbial forensics, database, pathogen, threat list, regulated pathogens, biological weapons

The delivery of *B. anthracis* spores through the United States mail in 2001 highlighted the need for techniques to support criminal forensic investigations and prosecution using data obtained from microorganisms. This growing field of study is called microbial forensics (1). A number of government agencies have created lists of the infectious microorganisms most relevant to their missions. Unfortunately, these “threat lists” are not standardized in intent or substance, and are frequently modified. The result is that a scientist, for example, may not be able to readily obtain taxonomic and gene sequence data on an organism from the U.S. Department of Agriculture’s (USDA) high consequence pathogens list because the pathogen name used may differ from the currently accepted taxonomic standard. Standardized taxonomic names are needed to connect an organism of interest with information such as the genomic sequence data accessible through the National Center for Biotechnology Information (NCBI) Taxonomy Browser (2). While the NCBI does not claim to be an authoritative source for information on taxonomy or phylogeny, at this time it is the most practical gateway to additional information, particularly sequence data, about an organism. Complicating matters, the classification of microorganisms is the subject of much disagreement within the biological community, and the names of some biological agents have been changed over time. Infectious agents are often referred to using common names or by the disease they cause (e.g., respiratory infection or typhoid fever) rather than by the accepted taxonomic nomenclature. This creates ambiguity when multiple species are capable of causing the disease. Not surprisingly, naming ambiguities pose a significant hurdle to communication among the diverse communities that must deal with biosecurity issues, especially for those without formal training in phylogenetic classification. While a number of public repositories for microbial data exist, including those maintained by the NCBI (3,4), The Institute for Genomics Research (5), the Tree-of-Life Web Project (6), and the Western Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research (7), each of these projects specializes in a particular type of data and none attempts to compile or link to all available data.

To facilitate coordinated access to information on disease-causing organisms and toxins, we have developed a database known as “The Microbial Rosetta Stone” that uses a new data storage structure designed to manage the complexities of microbiological data. Details of the data model and computational methods used to create and maintain this database will be described in a separate manuscript. This article serves as an introduction and visual accompaniment to the database and provides a basic explanation of the issues that must be addressed to avoid confusion in the community. We provide a compilation of internet sources for lists of important and/or regulated biological agents from the Department of Health and Human Services (HHS), the USDA, the National Institute of Allergy and Infectious Diseases (NIAID), and other literature references and sources (Table 1). We have organized and curated the names on these government lists to be consistent with the NCBI nomenclature and organisms were linked to genetic sequence data

[1] Ibis Division of Isis Pharmaceuticals, 1891 Rutherford Rd., Carlsbad, CA 92008.

[2] International Committee on Taxonomy of Viruses, Department of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032.

[3] Laboratory Division, Federal Bureau of Investigation, Washington, DC 20535.

Received 30 Oct. 2004; and in revised form 23 Mar. and 20 May 2005; accepted 4 June 2005; published 14 Sept. 2005.

Copyright © 2005 by ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959.

TABLE 1—*Pathogen lists and symbols used in figures.*

Threat List		Symbol	Microbial Rosetta Stone Table	Agency Source
Select Agents	HHS Select Agent	⚠	http://www.microbialrosettastone.com/ThreatListTables/HHS_Select.htm	http://www.cdc.gov/od/sap/index.htm
	USDA High Consequence Pathogen	⚠ animal ⚠ plant	http://www.microbialrosettastone.com/ThreatListTables/USDA.htm	http://www.aphis.usda.gov/vs/ncie/bta.html
NIAID Priority Pathogens	Priority A	🔴	http://www.microbialrosettastone.com/ThreatListTables/NIAID_A.htm	http://www2.niaid.nih.gov/biodefense/bandc_priority.htm
	Priority B	🟠	http://www.microbialrosettastone.com/ThreatListTables/NIAID_B-C.htm	
	Priority C	🔵		
Validated Biological Weapon	Cellular Life Toxins	☣ ☠	http://www.microbialrosettastone.com/ThreatListTables/Biowarfare_Cellular.htm	http://www.bt.cdc.gov/HealthProfessionals/index.asp
Globally Important Pathogens	Infectious Pathogens	⚕	http://www.microbialrosettastone.com/ThreatListTables/Global.htm	http://www.who.int/en/

in NCBI's GenBank database (if available). Important synonyms, previously used names or common names that identify the organisms have been gathered and stored. These tables are available at http://www.microbialrosettastone.com/ThreatListIndex.html and are organized according to correct taxonomic hierarchy. We have also organized the organisms according to biological relatedness and have provided graphic tree structure representations (Figs. 1–3) so that these relationships can be easily visualized. The Microbial Rosetta Stone database project aims to coordinate key information about microorganisms that can be used as biological weapons in a fashion that allows users of all backgrounds to readily retrieve relevant data. The merger of these lists of critical organisms from various sources and correlation with available phylogenetic and genomic sequence data will be of value to scientists across many disciplines as well as to governmental agencies charged with protecting the public.

Methods

Government agency lists were either taken directly from the specified agencies (sources indicated in Table 1) or compiled from primary literature references and other public sources as described below. Organism names from all lists were converted to NCBI-designated species names. This included changes from species synonyms to NCBI-accepted nomenclature and the expansion of names from Genus to Genus species designations. Where possible, a disease name was associated with the name of the most predominant organism known to cause that disease. Common synonyms of each agent and the accession number(s) for the complete genome of each agent (if available) were included in the tables of the database available on-line. For Fig. 1, cellular life forms (bacteria, fungi, and protozoa) were organized according to currently accepted phylogenetic tree structures. The viral trees in Figs. 2 and 3 were created by using standard methods or taken from literature references (8). The color-coding in the figures match the color-coding in the tables available on-line. Symbols used in the figures link pathogens to the threat lists as shown in Table 1.

Select Agents with Severe Human and Animal Health Consequences

Microorganisms that pose severe threats to human and animal health are regulated by two main government agencies, HHS and the USDA. HHS regulates the possession of biological agents and toxins that have the potential to pose a severe threat to public health and safety. HHS's Select Agent Program (9) regularly reviews and updates the list of select agents. The Select Agent Program currently requires registration by government agencies, universities, research institutions, and commercial entities that possess organisms or toxins on the select agent list. Similarly, the USDA is required by federal law to protect animal and plant health. High consequence livestock pathogens and toxins are agents that the USDA considers to have the potential to pose a severe threat to animal or plant health, or to animal or plant products (10). The organisms of greatest concern for these two agencies span all domains of life, including 20 bacteria, 48 viruses, four fungi, two protozoans, one prion, and 11 toxins. A significant number of these agents appear on both the HHS and USDA lists.

NIAID Priority Pathogens

The National Institute of Allergy and Infectious Diseases (NIAID) has undertaken a plan to increase national preparedness in the event of a bioterrorist attack. The *NIAID Strategic Plan For Biodefense Research* (11) lists specific characteristics that make organisms or toxins potential bioterror agents. These are: "high morbidity and mortality; potential for person-to-person transmission, directly or by vector; low infective dose and high infectivity by aerosol, with commensurate ability to cause large outbreaks; ability to contaminate food and water supplies; lack of a specific diagnostic test and/or effective treatment; lack of a safe and effective vaccine; potential to cause anxiety in the public and in health care workers; and the potential to be "weaponized" (12).

Agents possessing some or all of these characteristics are placed into one of three categories based on the magnitude of threat posed by the agent. The agents offering the highest potential danger from a bioterrorist attack are Category A Priority Pathogens. These agents are easily disseminated and transmitted person-to-person, with high mortality rates. They could easily cause a public panic and would require special action for public health preparedness (13). Category B Priority Pathogens are those agents offering moderate potential for danger. These agents are only moderately easy to disseminate and would cause low mortality and moderate morbidity. Quick responsiveness would require improvements in diagnostics and disease surveillance. Category C Priority Pathogens are emerging or reemerging pathogens that could pose a future danger. These agents are easy to obtain, produce, and disseminate, and potentially have high morbidity and mortality rates (14).

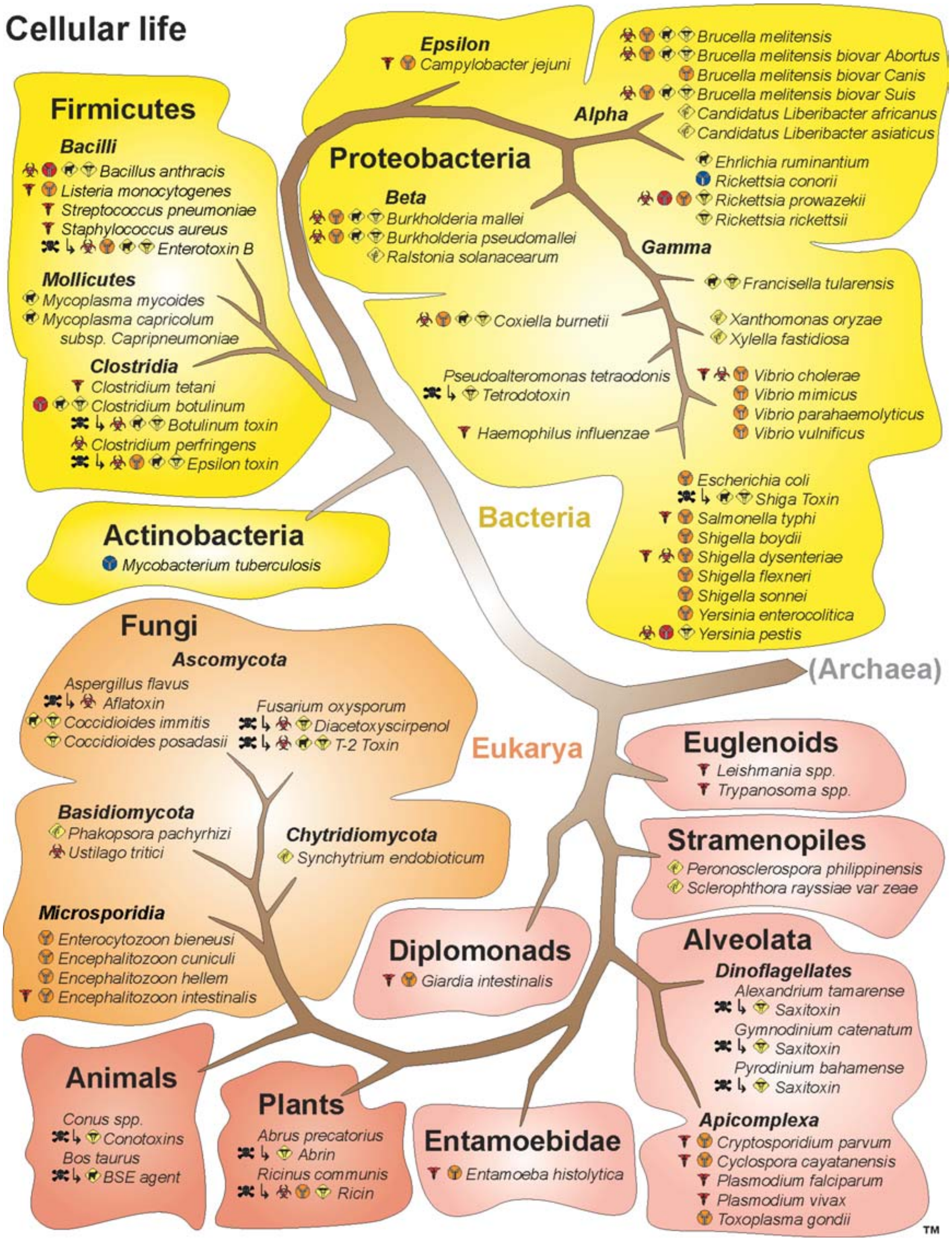

FIG. 1—*Phylogenetic relationships among pathogenic cellular life forms including bacteria, fungi, protozoa, and the plants and animals that produce toxins. Each branch of the tree is color-coded to correspond to colors in the tables in the database. The symbols are explained in Table 1. The versions of figures and tables available on-line are hyperlinked to additional information on each pathogen available through public resources such as the NCBI. The overlap in organisms on the priority pathogen lists from various agencies is apparent in this figure. For example, certain Bacilli, Proteobacteria, and Brucella strains appear on HHS, USDA, NIAID lists and are validated bioweapons.*

Bioweapons and Bioterrorism Agents

Any microorganism that could cause disruption to society can be considered a potential terrorist agent and thus the list of potential agents could be vast. Not all organisms or toxins make useful biological weapons, however. The properties that make organisms amenable for use as biological weapons have been discussed extensively (15,16). The most important features include: 1) availability/culturability, 2) stability, 3) disseminability, 4) resistance to environment, 5) infectivity, 6) target immunicity, 7) casualty effectiveness, 8) therapy, 9) epidemicity, and 10) retroactivity (15). In this work, we have defined two classes of bioweapons to aid in future studies: *Validated Bioweapons* and *Potential Bioweapons*. Microorganisms or toxins were categorized as *Validated Bioweapons* if they

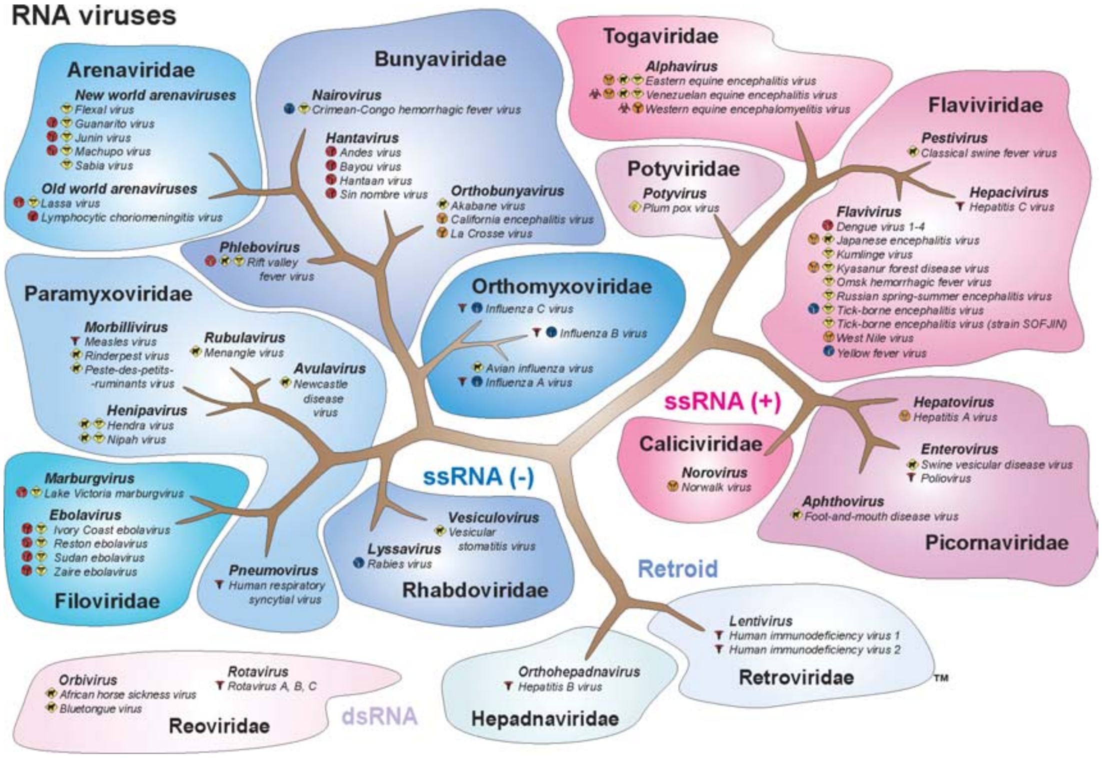

FIG. 2—*Phylogenetic relationships among pathogenic RNA viruses. The hyperlinked tables available in the database provide some of the functionality of the full database. For example, when one selects Influenza C virus, the NCBI taxonomy browser is immediately opened. This browser provides links to additional information to genomic sequence and disease facts at the CDC.*

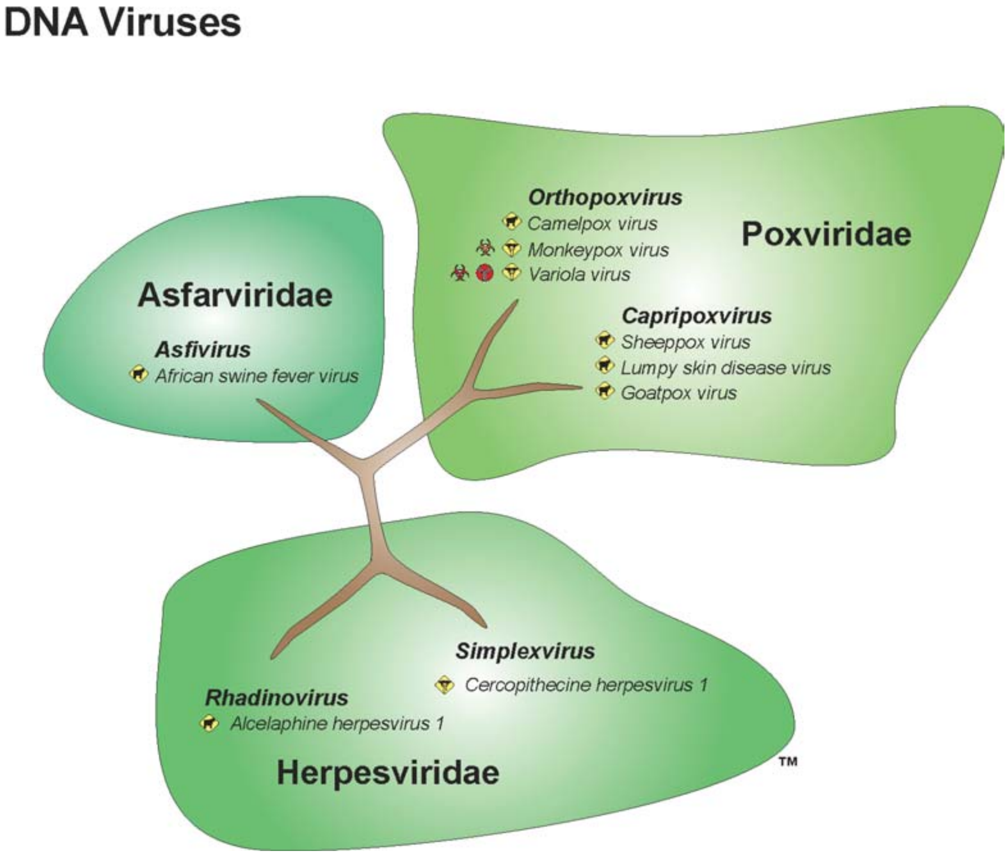

FIG. 3—*Phylogenetic relationships of pathogenic DNA viruses. Although there are many fewer pathogenic DNA viruses than RNA viruses, these viruses readily accept the insertion of foreign genes, and thus have a high potential for bioengineering.*

have been used or prepared for use as biological weapons in the past. *Potenial Bioweapons* were taken from the hierarchical ranking of infectious agents and toxins of biological origin prepared by Greenwood et al. (15). Fifty agents were numerically ranked according to ten criteria. The 23 highest-scoring *Potential Bioweapons* overlapped substantially with our list of *Validated Bioweapons*. Only validated biological weapons are displayed on the tree drawings. The display of biological weapons on the phylogenetic trees allows a comparison of these organisms to other important human and agricultural pathogens.

Globally Important Pathogens

The list of globally important pathogens was taken from the list maintained by the World Health Organization (WHO) found at (17). The leading causes of death worldwide are described by disease categories, where the disease can be caused by multiple diverse organisms or co-infections of different organisms, as is the case for acute respiratory disease, or by specific organisms, such as acquired immune deficiency syndrome (AIDS) or tuberculosis. In the former case, we identified the most prominent organisms responsible for the disease for placement in the curated list and on the phylogenetic charts.

Results and Discussion

Compiling the Data

The first step in creation of the Microbial Rosetta Stone Database was to compile a list of organisms that currently or potentially threaten public health as described in the Methods section. The names used for organisms on each list were converted to NCBI-accepted nomenclature and, if necessary, were expanded from Genus to Genus species (e.g., Shigella to Shigella dysenteriae or Shigella sonnei). In cases where only a disease name was given the disease was associated with the name of the predominant organism known to cause the disease. Organisms were organized by taxonomic hierarchy. Common synonyms of each agent and the accession number(s) for the complete genome of each agent (if available) were determined.

Phylogenetic Relationships of Important Infectious Microorganisms

The hyperlinked tables available at http://www.microbialrosettastone.com/ThreatListIndex.html provide some of the functionality of the full Microbial Rosetta Stone database. For example, when one selects Alphaproteobacteria from the table of NIAID Priority Pathogens Category B, the NCBI taxonomy browser is immediately opened. This browser provides links to additional information to genomic sequence and disease facts at the CDC. The accession numbers in the tables are also linked to the appropriate records in NCBI's Entrez Genome database.

There are three major domains of cellular life: bacteria, Eukarya, and Archaea. Organisms of the Archaeal domain are distinguished from bacteria and Eukaryotes by their ability to live in extreme environments. So far, no members of the Archaea are known to cause disease. Since most known archaea thrive in environments incompatible with human life there has presumably been little chance for infection (this assumption may change in the near future as free-living archaea that grow at moderate temperatures are being characterized). Bacterial *phyla* (the second highest level of the taxonomic classification hierarchy) are shown in the upper part of Fig. 1, with the Firmicutes (gram-stain positive organisms) and the Proteobacteria (gram-stain negative organisms) subdivided to the *Class* level (the third level of the taxonomic classification hierarchy, shown in italics). The Firmicutes and Proteobacteria account for roughly three quarters of the infectious bacterial species. The symbols used in the figures link pathogens to particular government threat lists as shown in Table 1.

Of the 20 phyla currently recognized in the NCBI taxonomy, 13 are not present in Fig. 1 due to the absence of any noteworthy pathogens. Some of these missing phyla are relatively important like the Cyanobacteria responsible for many Proterozoic oil deposits, but most unrepresented phyla are restricted to a handful of species and/or environmental niches. Also of interest is the relative weight of the infectious agents within their respective phylum: While virtually all Spirochaetes and Chlamydiae constitute potential infectious agents due to their parasitic lifestyle, the phylum Actinobacteria has few pathogens relative to the overall diversity of the phylum. It should be cautioned that our current view of bacterial diversity is biased towards cultivatable organisms, which represent a small fraction of all bacteria. Therefore, this compilation represents the high-visibility infectious bacterial agents, but should in no way be considered exhaustive or complete.

The lower part of Fig. 1 summarizes the eukaryotic agents that are of importance to public health. Eukaryotic microbial pathogens are clearly dominated by the Fungi and Protozoa. Within the Fungi, the phylum Ascomycota has many human and animal pathogens and is a major source of toxins. Protozoans are unicellular eukaryotic organisms that are responsible for globally important diseases such as the malaria-causing *Plasmodium* species, and the *Leishmania* and *Trypanosoma* species that cause significant mortality in the developing countries.

The tree of pathogenic RNA viruses (Fig. 2) is divided into three major branches based upon the virus genome: a) single-stranded negative-strand viruses (ssRNA(−)), b) single-stranded positive-strand viruses (ssRNA(+)), and c) retroid viruses. The common origin of RNA viruses and their tentative relationships as indicated on this chart are based on an extensive analysis of genetic loci involved in nucleic acid synthesis (manuscript in preparation). The double-stranded RNA viruses are not part of this tree, because they are likely to have arisen independently of each other and independently of the single-stranded RNA viruses (18). Particularly noteworthy negative-strand viruses are the deadly Ebola and Influenza viruses, the latter has claimed tens of millions of human lives in the past century. Retroid viruses include the two HIV viruses and the Hepatitis B virus. The main family of double-stranded RNA viruses, the Reoviridae, can be deadly to both humans and animals.

DNA viruses (Fig. 3) were organized based on the work of Iyer, Aravind, and Koonin, who showed the common ancestry of four large DNA virus families (8). The most noteworthy DNA virus family for the biosecurity community is the Poxviridae, which includes the etiologic agent of smallpox, *Variola virus*, and other important animal pathogens. These viruses have large genomes that can readily accept the insertion of foreign genes, and thus have a high potential for bioengineering.

Perhaps not surprisingly, a significant overlap was observed amongst the organisms in the lists from various agencies. For instance, over 60% of all the NIAID Category A priority pathogens are also listed in the HHS select agent list; 60–70% of all the biowarfare agents are found on the various government agency lists. The organization of these organisms according to biological relatedness in Figs. 1–3, with the symbols showing the multiple listing agencies (Table 1) provides a convenient illustration of the overlap of high-priority organisms.

In general, emerging diseases are not widely recognized as biosecurity threats and are therefore not well represented on official government lists. Of course, engineered pathogens fall outside the realm of known and therefore listed pathogens. However, engineered pathogens are created by using components of known organisms, thus understanding the relatedness of organisms will aid in determining what genetic exchanges might occur naturally vs. unnaturally. Updates to the database tables will contain a comprehensive list of emerging infectious agents and medically important pathogens.

Summary and Conclusions

This manuscript provides a description of the population of data tables in the Microbial Rosetta Stone Database and a visualization of the relatedness of important infectious microorganisms. Organism names from government high-priority pathogen lists were converted to NCBI species names, facilitating computational analysis and linkage to public genomic databases. The infectious organisms have been placed in tree structures that illustrate biological relatedness. The Microbial Rosetta Stone Database project includes additional links to disease information, seminal publications, genomic sequence data, and other external data sources. Frequent updates to the database and the addition of new tables are critical to maintaining the utility of this resource. The database is currently supported by government agencies involved in biosecurity, and we envision that a version of the full system able to be queried will be publicly available soon at www.microbialrosettastone.com. The curated data used to populate the database can already be found at this website as described in the text. In these tables, infectious agents of concern for biosecurity are organized visually by biological relatedness. We plan to further emphasize emerging infectious disease in future data releases. This resource should facilitate access to information on disease-causing organisms and toxins to members of the biosecurity community.

Acknowledgments

Development of the database was initiated with seed funding from DARPA and it has been augmented and maintained with support from the FBI and The Department of Homeland Security. The authors acknowledge editorial assistance from Dr. Jackie Wyatt.

References

1. Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JA, et al. Public health. Building microbial forensics as a response to bioterrorism. Science Sep 26 2003;301(5641):1852–3.
2. The NCBI Taxonomy Homepage. Available at: http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html.
3. NCBI. Available at: http://www.ncbi.nlm.nih.gov/.
4. Wheeler DL, Church DM, Federhen S, Bryant SH, Canese K, Church DM, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res Jan 1 2003;31(1):28–3.
5. The Institute for Genomics Research. Available at: http://www.tigr.org.
6. Tree of Life Web Project. Available at: http://tolweb.org/tree/.
7. Western Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research. Available at: http://rce.swmed.edu/biot/biot.htm.
8. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 2001;75:11720–34. [PubMed]
9. CDC Select Agent Program. Available at: http://www.cdc.gov/od/sap/.
10. APHIS Agricultural Select Agent Program. Available at: http://www.aphis.usda.gov/program/ag_selectagent/index.html.
11. NIAID Biodefense Research. Available at: http://www2.niaid.nih.gov/biodefense/research/strat_plan.htm.
12. Fauci AS. NIAID strategic plan for biodefense research: Bethesda, MD: National Institute of Allergy and Infectious Diseases, February 2002; NIH Publication No. 03-5306.
13. National Institute of Allergy and Infectious Diseases. NIAID biodefense research agenda for CDC category A agents. Bethesda, MD: National Institute of Allergy and Infectious Diseases, Feb 2002; NIH Publication No. 03-5308.
14. National Institute of Allergy and Infectious Diseases. NIAID biodefense research agenda for CDC category B and C priority pathogens. Bethesda, MD: National Institute of Allergy and Infectious Diseases, 2003.
15. Greenwood DP. A relative assessment of putative biological-warfare agents. Lexington, MA: Massachusetts Institute of Technology; 17 July 1997; ESC-TR-97-054.
16. Kortepeter MG, Parker GW. Potential biological weapons threats. Emerging Infect Dis 1999;5(4):523–7. [PubMed]
17. WHO: Communicable Disease Surveillance & Response (CSR). Available at: http://www.who.int/csr/disease/en/.
18. Gibbs MJ, Koga R, Moriyama H, Pfeiffer P, Fukuhara T. Phylogenetic analysis of some large double-stranded RNA replicons from plants suggests they evolved from a defective single-stranded RNA virus. J Gen Virol 2000;81:227–33. [PubMed]

Additional information and reprint requests:
David J. Ecker, Ph.D.
Ibis Division of Isis Pharmaceuticals
1891 Rutherford Road
Carlsbad, CA 92008