Int J Legal Med (2009) 123:65–69
DOI 10.1007/s00414-008-0247-1

TECHNICAL NOTE

The Microbial Rosetta Stone: a database system for tracking infectious microorganisms

Kumar L. Hari · Alan T. Goates · Ravi Jain · Aaron Towers · Vanessa S. Harpin · James M. Robertson · Mark R. Wilson · Vivek S. Samant · David J. Ecker · John A. McNeil · Bruce Budowle

Received: 30 August 2007 / Accepted: 18 March 2008 / Published online: 8 July 2008

Abstract The Microbial Rosetta Stone (MRS) database system was developed to support the law enforcement community by providing a comprehensive and connected microbial pathogen data–information repository. To handle the myriad types of pathogen information required to support law enforcement and intelligence community investigations, a data model previously developed for medical and epidemiological information was enhanced. The data contained in MRS are a broad collection of expert-curated microbial pathogen information, but given the multitude of potential microbes and toxins that may be used in a biocrime or bioterrorism act continual information collection and updating are required. The MRS currently relates governmental community-specific pathogen priority lists, sequence metadata, taxonomic classifications, and diseases to strain collections, specific detection and treatment protocols, and experimental results for biothreat agents. The system contains software tools that help to load, curate, and connect the data. A shared MRS database can be populated in real time by multiple users in multiple locations. Querying tools also provide simple and powerful means to access the data in any part of the database.

Keywords Forensic science · Microbial forensics · Database · Pathogen · Threat list

Introduction

The emerging field of microbial forensics has many challenges. One is the sheer magnitude of addressing all potential microorganisms and toxins that may be used in an attack [12]. An added complexity stems from organisms that are closely related to known infectious agents. These, too, may serve as sources for infectious agents [5, 7, 8]. Threats to the food supply and to the environment greatly enlarge the pool of pathogens. Lastly, even seemingly innocuous microorganisms may be used to perpetrate hoaxes (also considered a criminal act) to cause fear and disruption [3, 4].

Electronic supplementary material The online version of this article (doi:10.1007/s00414-008-0247-1) contains supplementary material, which is available to authorized users.

This work was supported under FBI Contract Nos. J-FBI-02-127 and J-FBI-04-189.

K. L. Hari · A. T. Goates · V. S. Harpin · V. S. Samant ·
D. J. Ecker · J. A. McNeil
Ibis Biosciences,
1891 Rutherford Rd.,
Carlsbad, CA 92008, USA

K. L. Hari · R. Jain
cBio, Inc.,
37896 Abraham St,
Fremont, CA 94536, USA

A. Towers · J. M. Robertson · M. R. Wilson · B. Budowle
Laboratory Division, Federal Bureau of Investigation,
FBI Academy,
Quantico, VA 22135, USA

Present address:
J. A. McNeil
John McNeil and Co.,
La Jolla, CA, USA

Present address:
B. Budowle (✉)
FBI Laboratory,
2501 Investigation Parkway,
Quantico, VA 22135, USA
e-mail: bruce.budowle@ic.fbi.gov

The vastness of microbial space makes it difficult for individuals to take on the effort required to gather all supporting information necessary for scientists and investigators involved in a microbial forensics investigation. Support from a robust informatics infrastructure is required for the collection and maintenance of pathogen priority lists and the dissemination of data related to threat attribution [3, 4, 10, 11]. This paper describes the Microbial Rosetta Stone (MRS) system, which is (1) a database containing information on pathogen information gathered from disparate sources and (2) an architecture of query tools and interfaces to access and extract that information. The purpose of this publication is to inform the greater forensic science community of this new tool for use in forensic attribution and combating bioterrorism. While it is intended to support microbial forensics investigations and attribution needs, the architecture, infrastructure, and software, which are commercially available, can be populated with any forensic data desired.

The MRS structure

The MRS is designed to provide information to help identify a microorganism or toxin, the possible origin of the sample, and whether or not the pathogen has been manipulated or bioengineered. The MRS system is not intended to replace currently available tools; rather, it is meant to serve as a connector and synchronizer to resolve ambiguities and data conflicts and, when unable to resolve these, provide sufficient context for users to make the most informed decisions possible. In a similar sense, convicted offender DNA databases, such as the Combined DNA Index System (http://www.fbi.gov/hq/lab/codis/index1.htm) [1], are an elegant model demonstrating the effectiveness of a readily accessed, shared database for assisting human forensic investigations. But such database concepts must be expanded for microbial forensics purposes. Currently, the MRS system is scalable, flexible, and capable of storing annotations to define the source for the data. The MRS consists of the table structure shown in Fig. 1 (in supplementary material). It is hosted on a WebObjects server and the content resides in a MySQL database and can be queried through multiple web-based interfaces. Development and deployments can be served from machines running either the Apple OS X or RedHat Linux operating systems. All client–server interactions occur behind a secure firewall to protect the integrity and security of the data. Configuration of the Linux servers requires several manual modifications, including a recommended switch from the CGI WebObjects adaptor to an Apache adaptor. The MRS is designed such that it also can function as a stand-alone system and be secured by placing the database and linked documentation on an encrypted PGP disk.

MRS content

The deployed database provides storage space for microorganism properties (pathogen, taxonomy, isolates, genomic characteristics, and molecular–biochemical signatures), diseases (disease, symptom, epidemiology, transmission), threat lists, genome and sequence metadata (genome, gene, sequence, sequence feature), and documents (author, publication, organization, capability). In order to support ongoing investigations, storage space also exists for experimental protocols, prior forensic investigations, addresses and links to organizations with access to high-priority microorganisms, and contact detail of reference laboratories for various diagnostic procedures. The connections among laboratory materials, equipment, and genetic markers are tracked so that the origins of a bioengineered threat event (i.e., isolate “heritage”) may be determined.

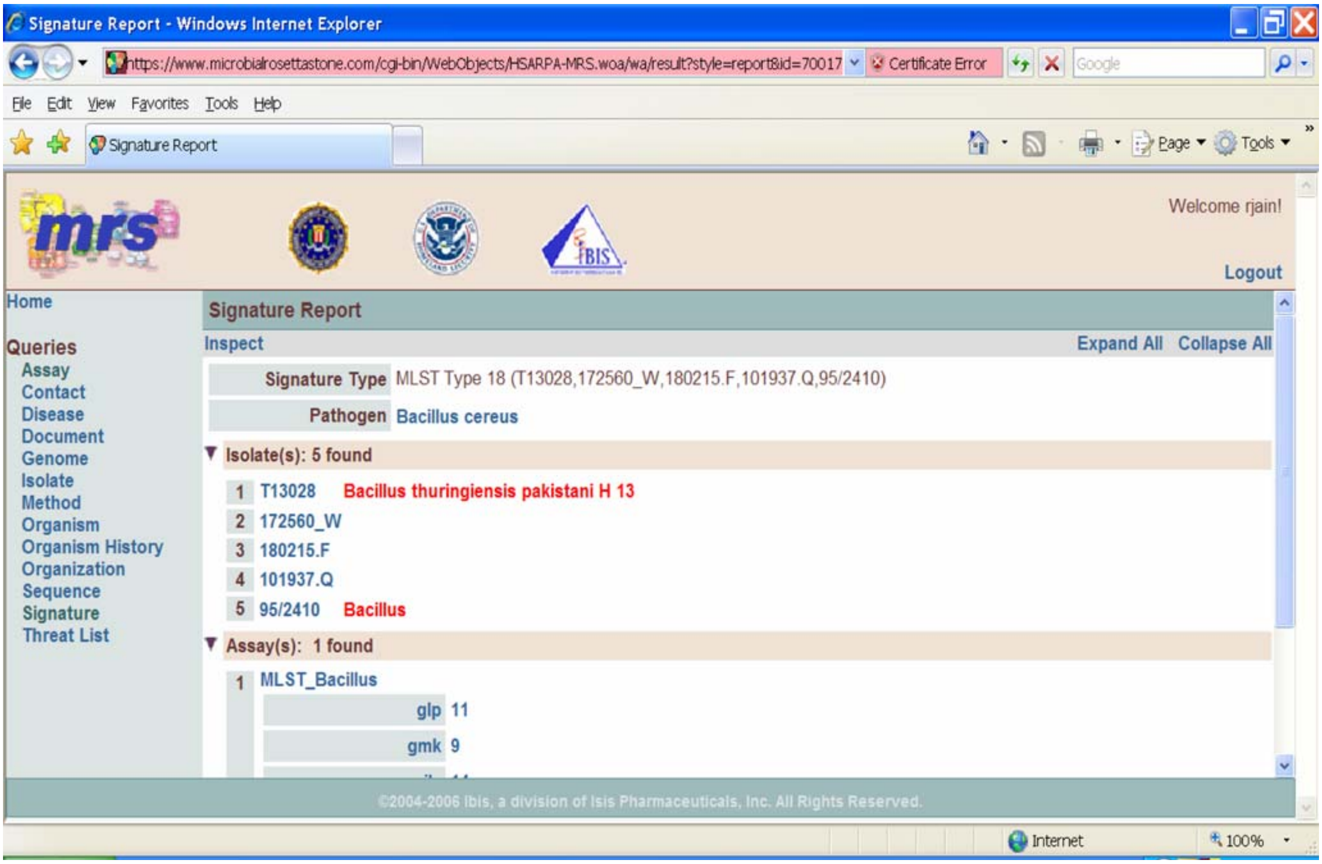
As part of the effort to populate the database, collaborators from the International Consortium on the Taxonomy of Viruses (ICTV) assisted in curation to link detailed genomic and DNA sequence data for viral species, strains, and isolates [2, 14]. Taxonomy data for viruses, bacteria, fungi, and other pathogenic microorganisms were also collected from National Center for Biotechnology Information (NCBI), and the viral taxonomies were coordinated with taxonomy data from ICTV. Species names and taxonomy designations often differed when comparing the NCBI and ICTV viral taxonomies. The MRS system addresses such data ambiguities by creating the concept of “equivalent pathogens.” In the MRS data model, equivalent pathogens allow the same organism to be represented in different ways, since the intent of MRS is not to define one taxonomic hypothesis as more correct; this is an important step in data curation. The MRS system maintains both hypotheses, or phylogenetic structures, and alerts the user to the alternatives. This key feature of the system makes it possible to represent multiple phylogenetic classification schemes for the same collection of isolates.

Organisms are categorized upon import into host or pathogen, depending on a flexible set of rules that can be overridden by a database curator at a later time. Sequence metadata are only retrieved from NCBI for organisms that appear in the curated threat agent lists or for organisms that have had disease information associated with them. This ameliorates the dataset from becoming overwhelming and difficult to search. Data also can be tied into sequence analysis tools (e.g., Basic Local Alignment Search Tool, etc.) in order to facilitate further research.

Curation is an essential part of development of a reliable and authoritative resource. This requires substantial interaction with subject matter experts and is the most demanding aspect of any data archive and resource. Our curation efforts for the data in the MRS repository include the manual, expert curation of a comprehensive list of threat organisms and collection of isolate and signature data for high-priority pathogens from public sources and primary literature. To date, experimental results have been uploaded from signatures that are biochemical, amplified fragment length polymorphisms, pulse field gel electrophoresis, multilocus sequence typing (MLST), multilocus variable number repeats, single-nucleotide polymorphisms, and small-nucleotide repeats. The links in MRS are maintained dynamically by a script and the repository is updated continuously with typing information as it becomes available. Upon request, the central MRS system is populated with customers' data as well. While during development the MRS is not accessible, the aim is to make available to collaborators and customers public domain content updates via a push–pull mechanism on a quarterly basis. The "MRS deployment model" section describes these interactions in greater detail.

MRS interfaces

An explicit goal for development of the MRS was to create facile methods for users to access the data. We built use-case-driven interfaces to render intuitive interactions [6]. The design is based on the seven high-level groupings of the MRS data: Assay, contact, disease, document, organism, sequence, and threat list (Fig. 1, supplementary material). By using these core groupings, customized search pages and reports were created to deliver query results as a hyperlinked summary of the most relevant data. Flexible search options exist for any data found in a report, including the ability to search based on synonyms for pathogens and diseases. To illustrate the report interfaces and the power of the signature algorithm, MLST data for *Bacillus* were loaded into MRS (Fig. 1) [9]. Five of the isolates tested in this assay grouped into a single cluster



Signature Report

Signature Type MLST Type 18 (T13028,172560_W,180215.F,101937.Q,95/2410)

Pathogen Bacillus cereus

▼ Isolate(s): 5 found

1	T13028	Bacillus thuringiensis pakistani H 13
2	172560_W	
3	180215.F	
4	101937.Q	
5	95/2410	Bacillus

▼ Assay(s): 1 found

1 MLST_Bacillus

glp	11
gmk	9

©2004-2006 Ibis, a division of Isis Pharmaceuticals, Inc. All Rights Reserved.

Fig. 1 Signature report highlighting conflicts. When the MRS generates signatures for all isolates run in the MLST assay for *B. cereus* [10], *B. thuringiensis pakistani H 13* strain clusters with other *B. cereus* strains. In the report for the signature for this MLST type, the difference in species is highlighted by displaying the alternate designation in *red*. An isolate with only a genus designation also is shown

denoted as MLST type 18. Three of the isolates belong to *Bacillus cereus*, while one has only been described as *Bacillus*. A fifth isolate, designated by strain ID T13028, has previously been described as *Bacillus thuringiensis pakistani H 13*. Because the MLST data for this isolate are identical to the *B. cereus* isolates, the signature algorithm groups them, while MRS highlights the species designation conflict. For the scientist, two possible interpretations are that either the species assignment of T13028 is incorrect or that the MLST test cannot distinguish between strains of *B. cereus* and *B. thuringiensis*. Finally, although summarized data are displayed in the example herein, users can access all the details of the data and relationships directly through inspect pages.

MRS deployment model

Thus far, a few MRS systems have been deployed and a central system is maintained. The central MRS repository contains only public domain information, while MRS systems deployed at customer sites (government and academic) contain client proprietary data. The current development of the MRS system aims to allow customers to query the central repository through a server, while loading and maintaining proprietary data locally. Upon release of the data into the public, customers will submit such data for incorporation into the central repository. The submitted data are reviewed for editing issues by an expert curator and then loaded into the central MRS. Data collection, curation, and loading from public sources occur continuously, where the curation steps and formats depend upon the assay type that is being curated. If desired by customers, content updates could be made available for download quarterly.

The MRS system will become commercially available in the near future. Stand-alone deployments will have phone and email support and, if desired, in-person support will be available as well. The overarching goal for developing the MRS system is to provide researchers and practitioners worldwide with an infrastructure that allows for aggregation of relevant forensics and diagnostics information from disparate sources and to support such efforts as best as is possible.

Future MRS developments

Thus far, the MRS system has been developed in collaboration with various US governmental agencies. The problems addressed by this system, such as fast access to pathogen strain typing information and data aggregation, however, are by no means limited to North American needs. Greater utility is premised on interaction internationally. For example, the Karolinska Institute and Pasteur Institute are well-established organizations for epidemiological and bioterrorism preparedness. Indeed, the European Union is constructing a system called the European Center for Infectious Diseases, which is similar to that of the Center for Disease Control and Prevention. Such groups would have additional data to populate the public access database and collaborations are necessary [13]. Going forward, we intend to solicit interaction with such sites worldwide for input and refinement of the current MRS system.

Such collaborations are necessary because a limitation and long-term challenge for any database are that current datasets for microbes are far from complete. As much information should be collated and curated to facilitate investigators, researchers, and practitioners. It is impossible to capture all desired information, such as strain diversity and endemicity, because they do not exist. However, just because all desired data do not exist is not justification to reject proceeding forward. A sophisticated system is necessary replete with extant data to combat bioterrorism and epidemics.

Future development of the MRS system may also include a feasibility study on the use of the MetaGraph data storage environment (http://www.metagraph.org/). MetaGraph supports the collection and analysis of complex biological datasets. The fundamental advantage of a MetaGraph schema is greater flexibility in annotating the data as well as the relationships, where relationships can be annotated with contextual information such as confidence, and multiple pieces of data can be related to form a hypothesis.

Acknowledgments This work was initiated with support from the Defense Advanced Research Projects Agency and also received funding from the Homeland Security Advanced Research projects Agency of the Department of Homeland Security. The authors thank Cornelia Buchen-Osmond for assistance with data modeling and curation and Jackie Wyatt and Paul Willett for critical reading of the manuscript.

The curated collection of pathogen threat lists is freely available at http://www.microbialrosettastone.com.

This is publication number 07-09 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

References

1. Baechtel FS, Monson KL, Forsen GE, Budowle B, Kearney JJ (1991) Tracking the violent criminal offender through DNA typing profiles—a national database system concept. Exs 58:356–360

2. Büchen-Osmond C (2003) The universal virus database ICTVdB. Comput Sci Eng 5:16–25
3. Budowle B, Murch R, Chakraborty R (2005) Microbial forensics: the next forensic challenge. Int J Legal Med 119:317–330
4. Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JA, Marrone BL, Robertson J, Campos J (2003) Public health. Building microbial forensics as a response to bioterrorism. Science 301:1852–1853
5. Cello J, Paul AV, Wimmer E (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science 297:1016–1018
6. Cockburn A (2001) Writing effective use cases. Addison-Wesley, Boston
7. Ferber D (2004) Synthetic biology. Microbes made to order. Science 303:158–161
8. Hoffmaster AR, Ravel J, Rasko DA, Chapman GD, Chute MD, Marston CK, De BK, Sacchi CT, Fitzgerald C, Mayer LW et al (2004) Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. Proc Natl Acad Sci U S A 101:8449–8454
9. Jolley KA, Chan MS, Maiden MC (2004) mlstdbNet—distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 5:86
10. Murch RS (2003) Microbial forensics: building a national capacity to investigate bioterrorism. Biosecur Bioterror 1:117–122
11. Schutzer SE, Budowle B, Atlas RM (2005) Biocrimes, microbial forensics, and the physician. PLoS Med 2:e337
12. Taylor LH, Latham SM, Woolhouse ME (2001) Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci 356:983–989
13. Tibayrenc M (2001) A European centre to respond to threats of bioterrorism and major epidemics. Bull WHO 79:12
14. Tidona CA, Darai G (eds) (2001) In: The Springer index of viruses. Springer, New York