Chapter 9

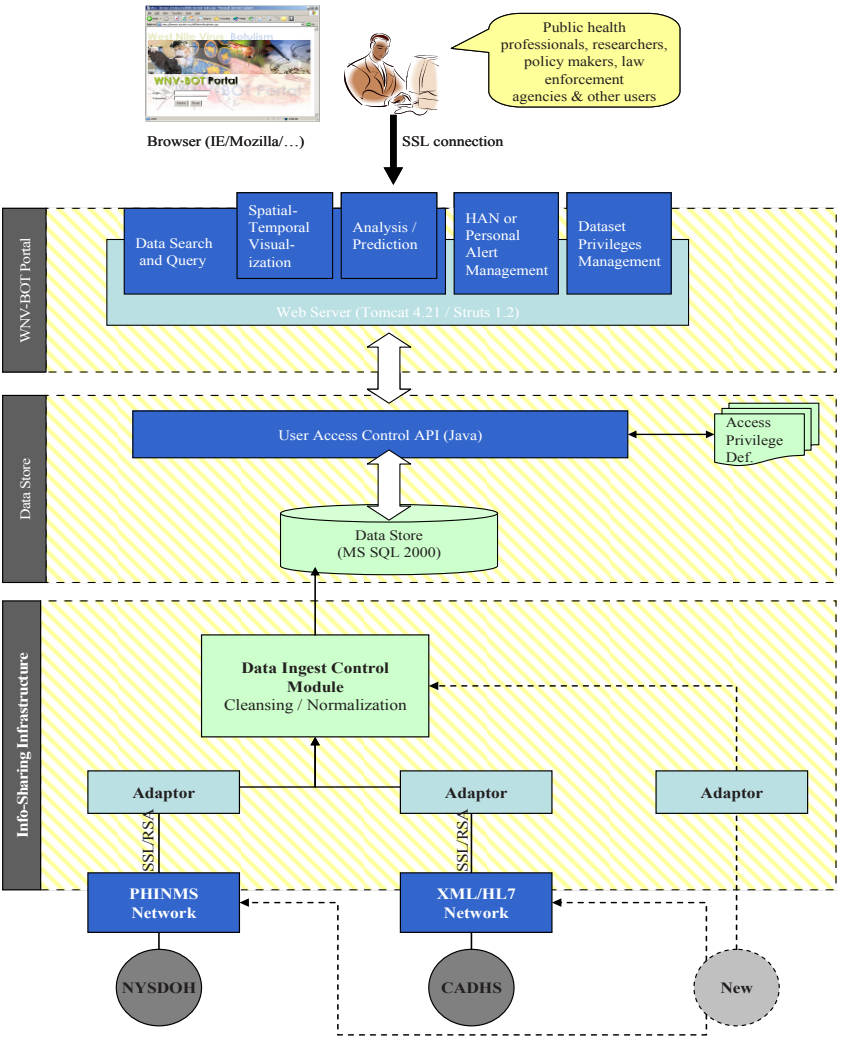
BIOPORTAL

The BioPortal project was initiated in 2003 by the University of Arizona Artificial Intelligence Lab and its collaborators in the New York State Department of Health and the California Department of Health Services to develop an infectious disease surveillance system. The project has been sponsored by NSF, DHS, DoD, Arizona Department of Health Services, and Kansas State University's BioSecurity Center, under the guidance of a federal inter-agency working group named the Infectious Disease Informatics Working Committee (IDIWC). Its partners include all the original collaborators as well as the USGS, University of California, Davis, University of Utah, the Arizona Department of Health Services, Kansas State University, and the National Taiwan University.

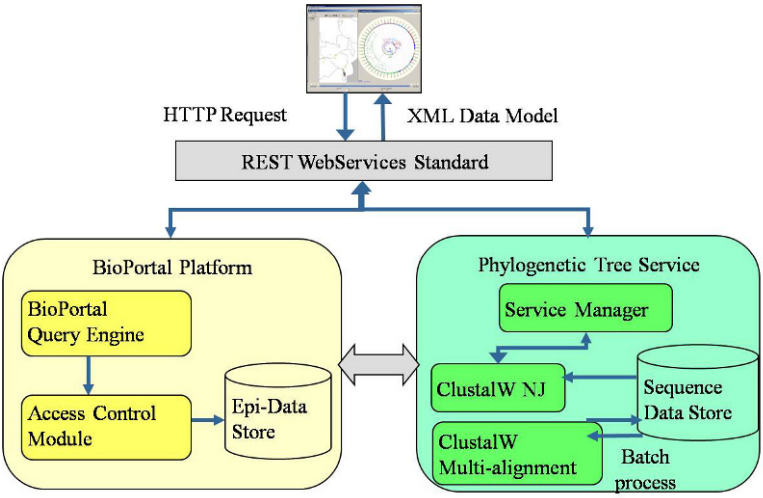
The BioPortal system provides distributed, cross-jurisdictional access to datasets concerning several major infectious diseases, including Botulism, West Nile Virus, foot-and-mouth disease, and live stock syndromes. Figure 9-1 shows the BioPortal system architecture. This portal system provides Web-based access to a variety of distributed infectious disease data sources including hospital ED free-text chief complaints (both in English and Chinese) as well as other epidemiological data. It features advanced spatial-temporal data analysis methods that include industry standard hotspot analysis algorithms and in-house developed innovative clustering-based techniques for retrospective and prospective data analysis. The analyses results are displayed via Spatio-Temporal Visualizer (STV). BioPortal also supports analysis and visualization of lab-generated gene sequence information. Its social network analysis module can be used to aid in the understanding of infectious disease transmission processes.

H. Chen et al., *Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense*, Integrated Series in Information Systems 21, DOI 10.1007/978-1-4419-1278-7_9,

The BioPortal system aims to improve the ability of public health practitioners to detect, and maintain situational awareness of outbreaks of emerging diseases and bioterrorist attacks, allowing for more timely and efficient deployment of resources for further investigation and response measures.



(a) BioPortal information sharing and data access infrastructure.



(b) BioPortal system architecture with epidemiological data and gene sequence data integrated.

Figure 9-1. BioPortal system architecture.

1. BIOPORTAL DATA COLLECTION

ED chief complaint data in the free-text format are provided by the Arizona Department of Health Services and several hospitals in a batch mode for syndrome classification. Various disease-specific case reports for both human and animal diseases are another source of data for BioPortal. It also makes use of surveillance datasets such as dead bird sightings and mosquito control information. The system's communication backbones, initially for data acquisition from New York or California disease datasets, consist of several messaging adaptors that can be customized to interoperate with various messaging systems. Participating syndromic data providers can link to the BioPortal data repository via the PHINMS and an XML/HL7 compatible network.

2. BIOPORTAL DATA ANALYSIS

BioPortal provides automatic syndrome classification capabilities based on free-text chief complaints. One method recently developed uses a concept ontology derived from the UMLS (Lu et al., 2008). For each chief complaint (CC), the method first standardizes the CC into one or more medical concepts in the UMLS. These concepts are then mapped into existing symptom groups

using a set of rules constructed from a symptom grouping table. For symptoms not in the table, a Weighted Semantic Similarity Score algorithm, which measures the semantic similarity between the target symptoms and existing symptom groups, is used to determine the best symptom group for the target symptom. The ontology-enhanced CC classification method has also been extended to handle CCs in Chinese.

BioPortal supports hotspot analysis using various methods for detecting unusual spatial and temporal clusters of events. A hotspot is a condition indicating some form of clustering in a spatial distribution. Hotspot analysis facilitates disease outbreak detection and predictive modeling based on historical spatial-temporal data and in turn uses them for predictive purposes.

SaTScan is made available as part of the BioPortal system through a simple Web interface and STV. BioPortal also supports the Nearest Neighbor Hierarchical Clustering method, and two new methods (Risk-Adjusted Support Vector Clustering, and Prospective Support Vector Clustering) developed in-house (discussed in Chapter 4) (Chang et al., 2005; Zeng et al., 2004a). The version of SaTScan that is incorporated in the BioPortal system uses the Bernoulli method. The distribution of baseline observations (or controls) and the distribution of new observations (or cases) are compared and circular clusters are identified where the proportion of new observations is significantly higher than the proportion of new observations outside the circle. RSVC is a clustering-based, spatio-temporal hotspot analysis algorithm developed at the Artificial Intelligence Laboratory of the University of Arizona. It combines the power of support vector machines (SVM) with the risk adjustment approach from CrimeStat®. It clusters points with consideration for baseline information (data under normal conditions) to find the emerging at risk area. In addition, BioPortal uses the RNNH algorithm provided by CrimeStat® III. The Nearest Neighbor Hierarchical clustering (NNH) routine in CrimeStat identifies groups of incidents that are spatially close. It clusters points together and then proceeds to group the clusters together. The Risk-adjusted Nearest Neighbor Hierarchical clustering routine (RNNH) combines the hierarchical clustering capabilities with kernel density interpolation techniques.

3. BIOPORTAL VISUALIZATION, INFORMATION DISSEMINATION, AND REPORTING

Figure 9-2 shows the screenshot of the interactive Web-based surveillance portal. This application allows the user to explore the incidence of infectious diseases. The portal allows the user to: (1) select a disease of concern and access-related databases; (2) narrow the scope by time-frame and geographic area of interest; (3) view a variety of data aggregations; and (4) perform hotspot analysis to focus attention on critical areas.

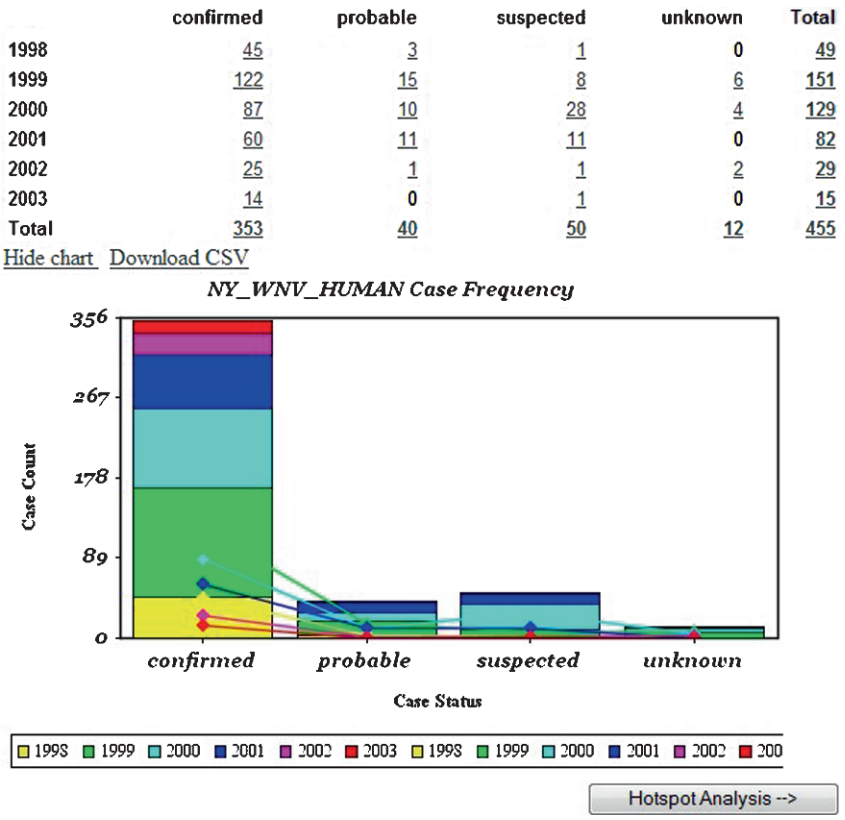


Figure 9-2. Interactive Web-based BioPortal surveillance portal.

Monitored disease incidence time series are shown on the surveillance dashboard for the participating hospitals and other healthcare organizations to view (Figure 9-3). The dashboard is integrated with time series detection capability and the BioPortal hotspot analysis and visualization tools. Detected abnormalities are alerted on the upper panel.

BioPortal makes available a visualization environment called the Spatial-Temporal Visualizer (STV), which allows users to interactively explore spatial and temporal patterns, based on an integrated tool set consisting of a GIS view, a timeline tool, and a periodic pattern tool (Hu et al., 2005).

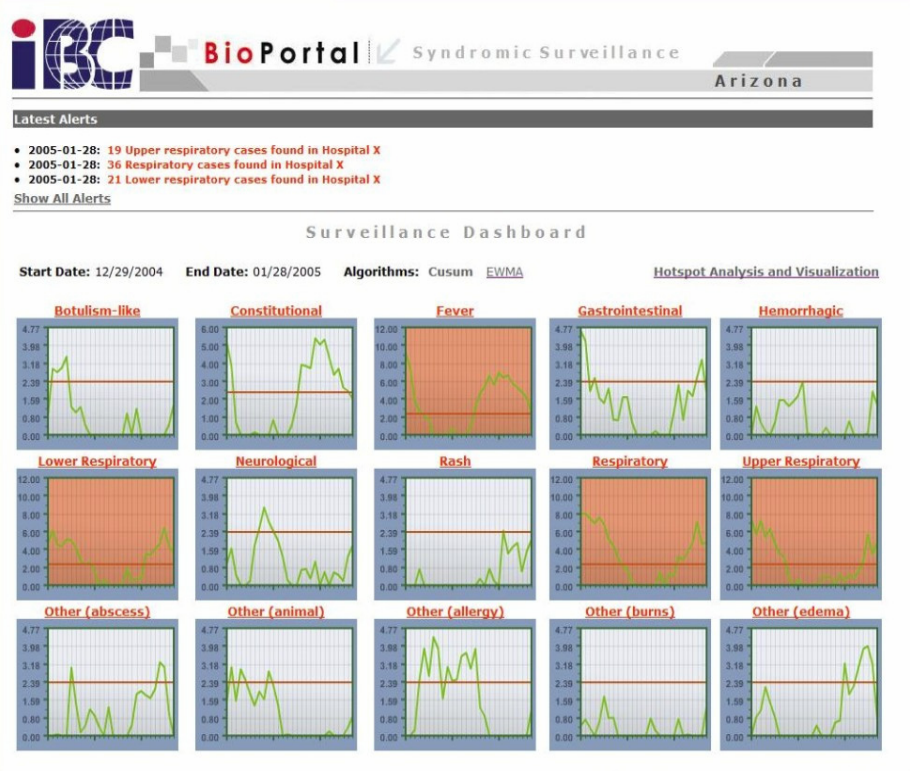


Figure 9-3. BioPortal syndromic surveillance dashboard integrated with time series detection capability and the hotspot analysis and visualization tools.

Figure 9-4 illustrates how these three views can be used to explore an infectious disease dataset. The GIS view displays cases and sightings on a map. The user can select multiple datasets to be shown on the map in different layers using the checkboxes (e.g., disease cases, natural land features, and land-use elements). Through the periodic view the user can identify periodic temporal patterns (e.g., which months or weeks have an unusually high number of cases). The unit of time for aggregation can also be set as days or hours. The timeline view provides a timeline along with a hierarchical display of the data elements, organized as a tree.

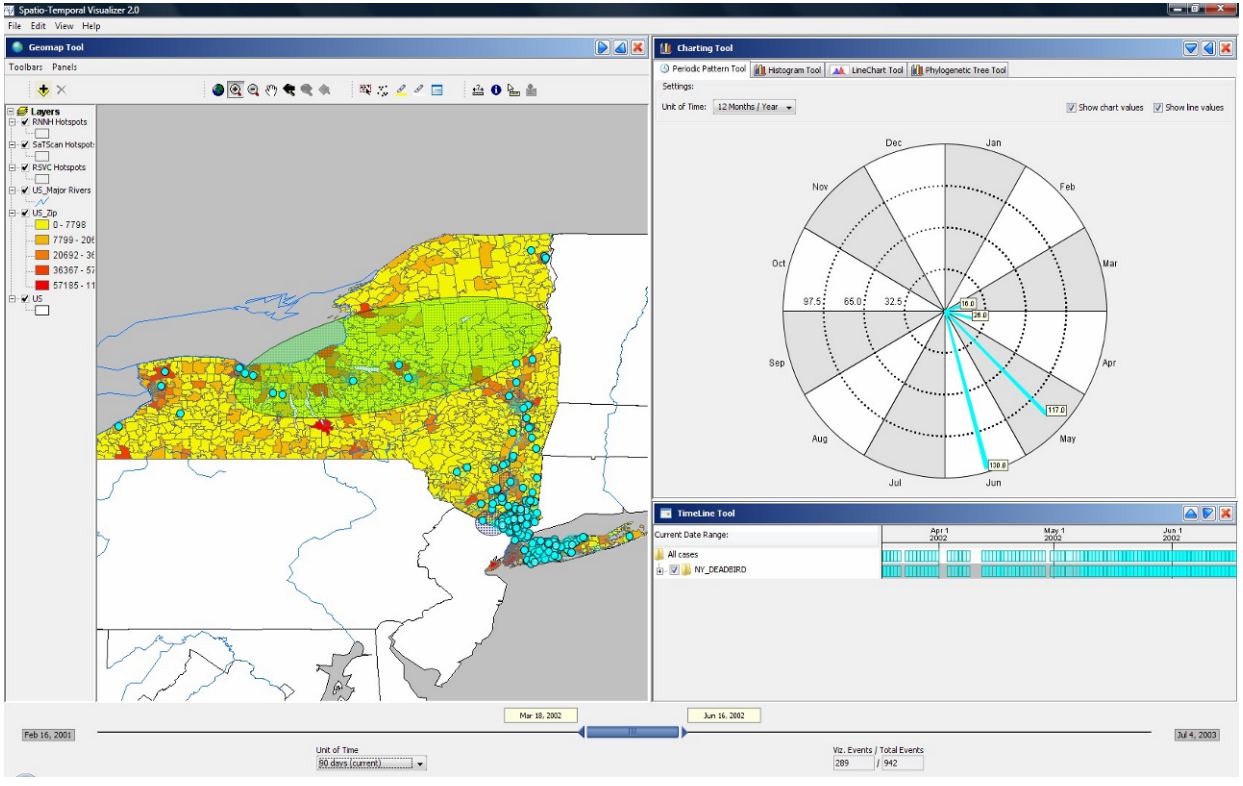


Figure 9-4. BioPortal Spatial-Temporal Visualizer.

A new sequence-based phylogenetic tree visualizer has been recently developed for diseases such as the foot-and-mouth disease, for which gene sequence information is available (Figure 9-5). Phylogenetic tree analysis examines the DNA of pathogens to determine the genetic relationship between various strains, and to identify possible sources or mutation. The results of an analysis can be drawn as a phylogenetic tree showing the hierarchical hypothesized evolutionary relationships (phylogeny) between organisms. Each member in a branch is assumed to be descended from a common ancestor. The module color-codes outbreak occurrences based on distance in genetic space to help predict distribution of virus strains, and aids in more efficient vaccine distribution (Thurmond et al., 2007).

The BioPortal system also provides Social Network Analysis (SNA) capability for epidemic transmission process investigations (Figure 9-6). Examining social networks is a useful epidemiological tool for understanding the progression of the spread of infectious diseases such as sexually transmitted diseases. The SNA module in the BioPortal system incorporates

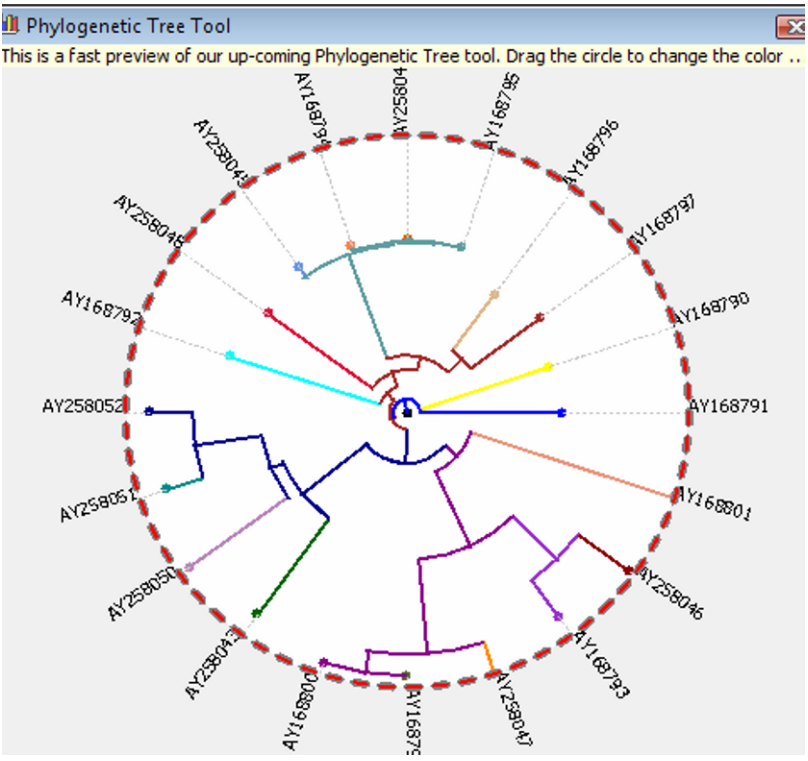


Figure 9-5. BioPortal phylogenetic tree analysis (source: BioPortal Web page).

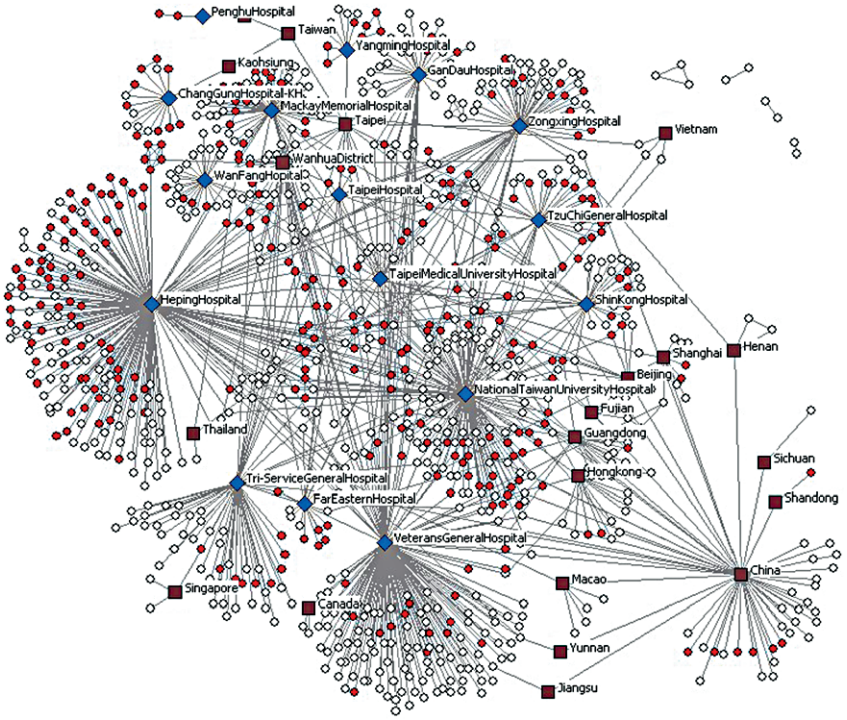


Figure 9-6. Social network analysis to analyze the SARS epidemic in Taiwan in 2003 (Chen et al., 2007).

geographical locations, which might be high risk areas such as hospitals, into social networks to examine the role of such locations in infectious disease transmission, and to identify potential bridges between locations. This helps to maintain situational awareness and target incident investigation and mitigation efforts more effectively. Social Network Analysis was also employed to analyze the SARS epidemic in Taiwan in 2003.

Data confidentiality, security, and access control are among the key research and development issues for the BioPortal project. An access control mechanism is implemented based on data confidentiality and user access privileges. For example, access privileges to the zip code and county level of individual patient records may be granted to selected public health epidemiologists. The project also developed various Memoranda of Understanding (MOUs) for data sharing among different local and state agencies.

4. CASE STUDY: FOOT-AND-MOUTH DISEASE SITUATIONAL AWARENESS

Foot-and-Mouth Disease (FMD) is considered to be one of the most contagious infectious animal diseases in the world. BioPortal plays an important role in the collaborative efforts with the FMD Laboratory at the University of California, Davis, for developing global real time surveillance for foot-and-mouth disease. The FMD BioPortal focuses on: (1) gathering global FMD data; (2) identifying surrogates of risks; (3) modeling and predicting FMD virus evolution; and (4) evaluating and testing FMD surveillance methodologies.

FMD BioPortal integrates information and data related to foot-and-mouth disease from public sources and collects proprietary or confidential data through secure specific routing structures. Major data sources include the World Reference Laboratory at Pirbright, animal surveillance data from FAO (Food and Agriculture Organization of the United Nations) and OIE (World Organisation for Animal Health), and GenBank sequence data.

Analytical and visualization tools for data summarization and trend detection can be selected and invoked through the FMD BioPortal Web-based platform as illustrated in Figure 9-7. The BioPortal infrastructure provides generic support for summarizing and visualizing FMD-related data with prominent spatial and temporal data elements through the Spatial-Temporal Visualizer (STV) (an example is shown in Figure 9-8).

A major enhancement to STV developed specifically for FMD BioPortal is the phylogenetic tree visualization that allows the incorporation of genomic information visualization in addition to the existing spatial and temporal data visualization capabilities (Figure 9-9). The phylogenetic tree visualization is used to display temporal-spatial genomic variation of FMD isolates and allows user-driven evaluation of differences in genomic variation over time and geographic location.

In addition, FMD News monitoring is an ongoing effort by the Artificial Intelligence Lab at the University of Arizona and the FMD Lab at UC Davis to collect open source FMD breaking news. A team of epidemiologists from different countries at the FMD Lab reviews more than 40 Web sites daily and sends out the selected news items in a summary format to a listserv. An automatic FMD related news collection and classification system was recently developed by the AI Lab at the University of Arizona.

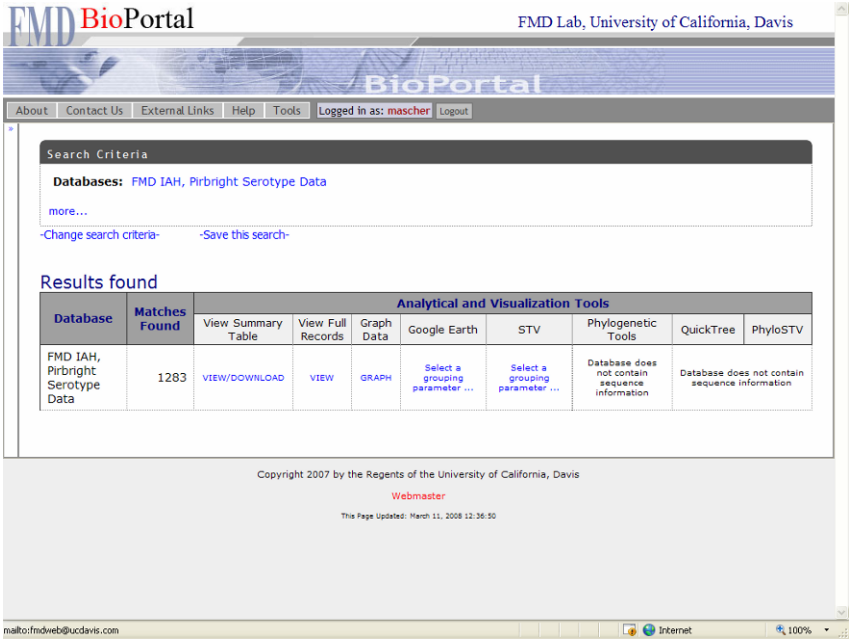


Figure 9-7. FMD BioPortal for accessing analytical and visualization tools (source: FMD BioPortal Web site).

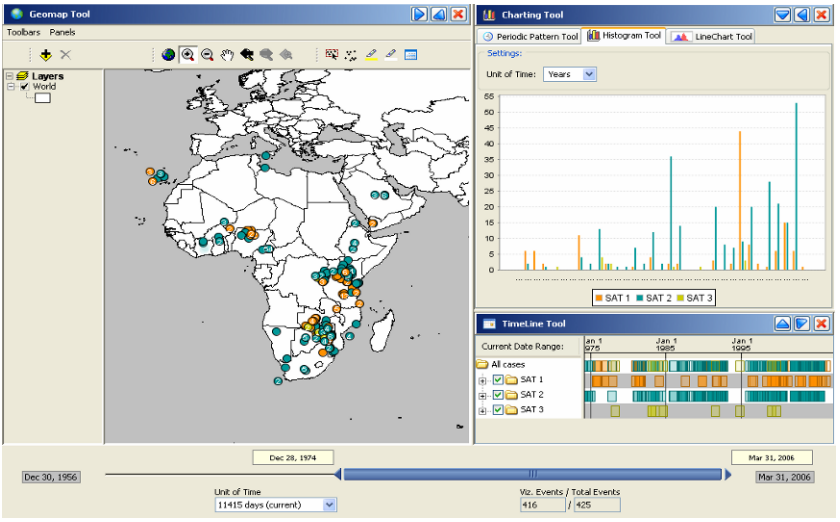


Figure 9-8. Visualization of FMD geographical distribution (source: FMD BioPortal Web site).

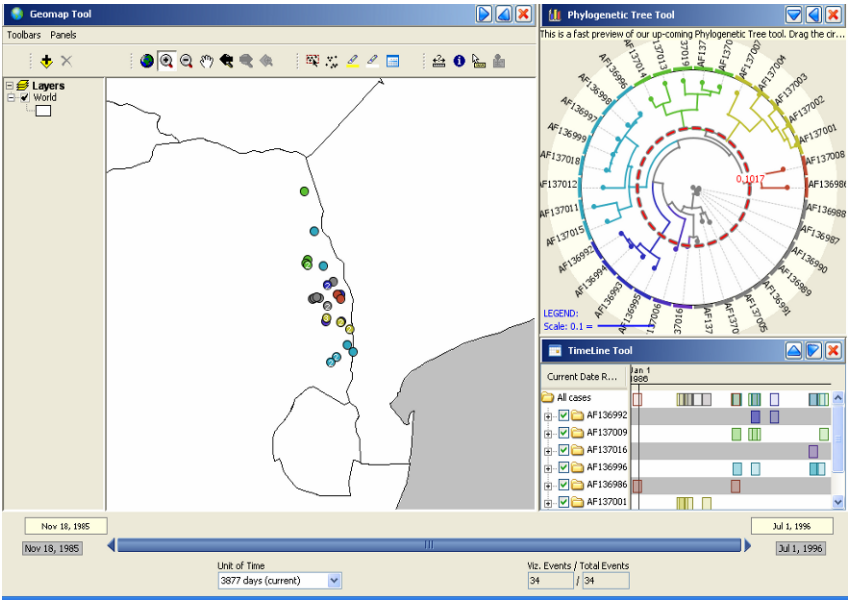


Figure 9-9. FMD phylogenetic tree visualization (source: FMD BioPortal Web site).

5. FURTHER READINGS

We provide the following project link and some key readings for the readers who might be interested in learning more details about the BioPortal project.

Project link:

http://biocomputingcorp.com/bphome.html
http://ai.arizona.edu/research/bioportal/index.htm

Important readings:

1. Hu, P., D. Zeng, H. Chen, C. Larson, W. Chang, C. Tseng, and J. Ma (2007). "System for Infectious Disease Information Sharing and Analysis: Design and Evaluation," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 11, No. 4.
2. Lu, H.-M., D. Zeng, L. Trujillo, K. Komatsu, and H. Chen (2008). "Ontology-Enhanced Automatic Chief Complaint Classification for Syndromic Surveillance," *Journal of Biomedical Informatics,* Vol. 41, No. 2, pp 340–356.

3. Chang, W., D. Zeng, and H. Chen (2008). "A Stack-Based Prospective Spatio-Temporal Data Analysis Approach," *Decision Support Systems*, Vol. 45, No. 4, pp 697–713.
4. Zhang, Y.L., Y. Dang, Y.-D. Chen, H. Chen, M. Thurmond, C.-C. King, D. Zeng, C. Larson (2008). "BioPortal Infectious Disease Informatics research: disease surveillance and situational awareness," *in proceedings of International Conference on Digital Government Research*, pp 393–394.