LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-221584

MannDB: A microbial annotation database for protein characterization

Carol Zhou, Marisa Lam, Jason Smith, Adam Zemla, Matthew Dyer, Tom Kuczmarski, Tom Slezak

May 23, 2006

BMC Bioinformatics

MannDB – A microbial annotation database for protein characterization

Carol L Ecale Zhou[1,3], Marisa W Lam[1], Jason R Smith[1], Adam T Zemla[1], Matthew D Dyer[2], Thomas A Kuczmarski[1], Elizabeth A Vitalis[1], Thomas R Slezak[1]

[1]Lawrence Livermore National Laboratory, Pathogen Bio-informatics, Livermore, CA, USA
[2]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[3]Corresponding author

Author e-mail addresses:
CLEZ: zhou4@llnl.gov
MWL: lam9@llnl.gov
JRS: smith250@llnl.gov
ATZ: zemla1@llnl.gov
MDD: dyer20@llnl.gov
TAK: tomk45@gmail.com
EAV: vitalis1@llnl.gov
TRS: slezak1@llnl.gov

BMC Bioinformatics 2006 7:459

UCRL-JRNL-221584

Abstract

Background: MannDB was created to meet a need for rapid, comprehensive automated protein sequence analyses to support selection of proteins suitable as targets for driving the development of reagents for pathogen or protein toxin detection. Because a large number of open-source tools were needed, it was necessary to produce a software system to scale the computations for whole-proteome analysis. Thus, we built a fully automated system for executing software tools and for storage, integration, and display of automated protein sequence analysis and annotation data.

Description: MannDB is a relational database that organizes data resulting from fully automated, high-throughput protein-sequence analyses using open-source tools. Types of analyses provided include predictions of cleavage, chemical properties, classification, features, functional assignment, post-translational modifications, motifs, antigenicity, and secondary structure. Proteomes (lists of hypothetical and known proteins) are downloaded and parsed from Genbank and then inserted into MannDB, and annotations from SwissProt are downloaded when identifiers are found in the Genbank entry or when identical sequences are identified. Currently 36 open-source tools are run against MannDB protein sequences either on local systems or by means of batch submission to external servers. In addition, BLAST against protein entries in MvirDB, our database of microbial virulence factors, is performed. A web client browser enables viewing of computational results and downloaded annotations, and a query tool enables structured and free-text search capabilities. When available, links to external databases, including MvirDB, are provided. MannDB contains whole-proteome analyses for at least one representative organism from each category of biological threat organism listed by APHIS, CDC, HHS, NIAID, USDA, USFDA, and WHO.

Conclusions: MannDB comprises a large number of genomes and comprehensive protein sequence analyses representing organisms listed as high-priority agents on the websites of several governmental organizations concerned with bio-terrorism. MannDB provides the user with a BLAST interface for comparison of native and non-native sequences and a query tool for conveniently selecting proteins of interest. In addition, the user has access to a web-based browser that compiles comprehensive and extensive reports. Access to MannDB is freely available at http://manndb.llnl.gov/.

Background

MannDB was created to meet a need for rapid, comprehensive sequence analysis with an emphasis on protein processing, surface characteristics, and functional classification to support selection of pathogen or virulence-associated proteins suitable as targets for driving the development of protein-based reagents (e.g., antibodies, non-natural amino-acid ligands, synthetic high-affinity ligands) for pathogen detection [1,2]. Because comprehensive analyses of this type required using a large number of open-source tools, and because it was necessary to scale the computations for analysis of whole proteomes, we built a fully automated system for executing sequence analysis tools and for storage, integration, and display of protein sequence analysis and annotation data. In order to be

able to rapidly examine and compare whole bacterial and viral proteomes for selection of suitable target proteins for bio-defense applications, we compiled data for whole proteomes from representative organisms from all categories of biological threat agents listed by several governmental agencies: APHIS, CDC, HHS, USDA, USFDA, NIAID, and WHO [3-9] as well as taxonomic near-neighbor species as appropriate. Therefore, the scope of MannDB is automated sequence analysis and evidence integration for proteins from all currently recognized bio-threat pathogens. Emphasis is placed upon analyses that are most useful in characterizing potential protein targets and surface motifs that could be exploited for development of detection reagents. The content of MannDB is updated on a regular basis.

In recent years several software systems and accompanying databases have been developed for microbial genome annotation, each with a particular emphasis [10-19]. Some databases place an emphasis on gene prediction and DNA-based analyses vs. protein sequence-based analyses, or provide automated (primary) vs. curated (secondary) annotations. Although microbial annotation databases frequently include predictions of biological, chemical, structural, and physical properties of proteins (e.g., antigenicity, post-translational modifications, hydrophobicity, membrane helices), none currently offers the comprehensive suite of analyses (see MannDB website for complete list of tools) contained within MannDB for characterizing viral as well as bacterial proteins from human and agricultural/veterinary pathogens of interest to the bio-defense community and for rapidly identifying putative virulence-associated proteins for development of functional assays. The MannDB database was built and linked to MvirDB [20] in order to meet these requirements. In addition, we focus on sequence analyses that assist in selection of protein features (e.g., surface characteristics) most suited for targeting detection reagent development.

Construction and content

MannDB is implemented as an Oracle 10g relational database. The schema for MannDB data organization is available on the website. MannDB captures results from our fully automated, high-throughput, whole-proteome sequence analysis process pipeline, depicted in Fig. 1. Proteomes (lists of hypothetical and known proteins) representing human bacterial and viral pathogens and near-neighbor species are downloaded from GenBank and parsed into MannDB. Whenever possible, we begin with gene calls on finished genomes. However, the system can be used to predict genes on draft genomes, and can be used to analyze arbitrary lists of protein sequences. Reference genomes are updated on a quarterly basis to ensure that the software tools are being run on current sequence data. Annotations from SwissProt are downloaded when GenBank entries contain SwissProt identifiers, or when identical sequences are detected by blasting MannDB entries against the SwissProt protein fasta database. MannDB contains at least one reference genome for each category of pathogen listed as a bio-threat organism on websites maintained by APHIS, CDC, HHS, USDA, USFDA, NIAID, and WHO. Open-source tools are run either on local systems or by means of batch submission to external servers. As of this writing the system executes 36 tools, which are listed on the MannDB web site. Automated sequence analyses include predictions of post-translational modifications, structural conformation, chemical properties, functional assignment, and

antigenicity, as well as motif detection and pre-computed BLAST against protein and nucleic acid sequences in MvirDB, our database of microbial virulence factors, protein toxins, and antibiotic resistance genes [20]. Tools that are run in-house are updated periodically to ensure that the system is running the most recent software versions against the most recent data sets. Tools are selected and input parameters are set according to the taxon of the organism from which the protein set is constructed. For example, some tools (e.g., NetPicoRNA; [21]) are run only on specific organisms, whereas others (e.g., SignalP; [22]) have taxon-specific settings. In some cases we run more than one tool for a similar prediction. TMHMM and TopPred both predict membrane helices, but results may differ, for example, in the start and end residues for a given segment. Our strategy is to employ more than one tool, when available, so that conflicting results can be noted and evaluated by the user. In parsing results from each tool, data are inserted into one of nine tables (see schema on web site) depending on the type of prediction (e.g., protein chemistry); tools that make similar predictions tend to produce similarly structured output (although formatting differs considerably), which facilitates data storage and retrieval.

A web client browser enables viewing of automated analysis results, annotations, and links to MvirDB (Fig. 2). The user first selects a proteome, then a specific protein for which to view summary results, and finally selects the specific categories of analysis to be viewed. Only analyses returning results are displayed. Hyperlinks to external data sources are provided for additional information whenever external database identifiers are returned. The MannDB toolset includes a BLAST interface, which can be used to quickly identify an entry of interest by its sequence, when the gene name or locus tag is unknown, or to identify protein sequences related to a sequence of interest. A query tool allows the user to construct 3 types of searches: 1) free-text searches against all database fields that contain descriptive information, including fields containing gene names or external database identifiers, 2) structured searches against specific analysis types, and 3) a search for proteins linked to entries in MvirDB either by common unique identifier or by pre-computed blast homology. Reports and results sets from the query tool can be downloaded into Excel.

Utility and discussion

MannDB provides users with pre-computed sequence analyses for complete proteomes of bacterial and viral pathogens from several governmental agencies' lists of bio-threat agents. The genomes and tools are maintained up to date, with predictions being re-run every 3 months. The user can browse proteomes, or can blast sequences against MannDB to pull up related entries and associated data. MannDB provides a convenient source of automated sequence analyses and downloaded annotation information for whole proteomes of human pathogenic bacteria and viruses and has a high degree of integration with external databases.

MannDB provides sequence analysis information of primary interest to researchers in the bio-defense community. We have been using MannDB for several years to "annotate" DNA signatures [1] and more recently to assist collaborators in efforts to down-select from whole bacterial and viral genomes to identify suitable protein targets and protein features for driving the development of detection reagents [2]. For example, a common requirement for a detection assay is that it be performed with

minimal sample disruption. Therefore, an initial down selection for proteins expected to be on the surface of a bacterial particle might entail identification of proteins that are predicted to be secreted or membrane bound by using tools such as PSORT [23,24], TMHMM [25], SignalP, TargetP [26], TopPred [27], and HMMTOP [28]. Having results from several tools that provide similar predictions but using different algorithms or slightly different approaches allows us to compare predictions and make selections with greater confidence. Identification of surface features for targeting of detection reagents is done primarily by means of additional sequence- and structure-based analyses [2], although predictions pertaining to post-translational modifications (e.g., glycosylation, cleavage) are taken into consideration as they may affect protein recognition.

Conclusions

MannDB is a genome-centric database containing comprehensive automated sequence analysis predictions for protein sequences from organisms of interest to the bio-defense research community. Computational tools for the MannDB automated pipeline were selected based on customer needs in providing down selections from large sets of proteins (e.g., whole proteomes) to short lists of proteins most suitable for developing reagents to be used in field assays for detection of pathogens. For that reason we have focused our efforts on applying tools that would enable selection of proteins that meet assay requirements, such as cellular localization, that would assist in determining the value of a surface feature for targeting ligand binding, or that would identify antigenic sub-sequences of particular value in antibody development. As the goals of some of these assays have been to detect toxins or proteins associated with virulence, we constructed hard links between protein sequences in MannDB with entries in MvirDB in order to conveniently identify and characterize protein targets and features for these applications. We believe that MannDB will be of general use to the bio-defense and medical research communities as a resource for predictive sequence analyses and virulence information.

Availability and requirements

MannDB is freely accessible at http://manndb.llnl.gov/. Although the software that populates and updates MannDB is not open-source, the user may request collaborative sequence analysis services by contacting ppi_group@kpath.llnl.gov.

List of abbreviations

BLAST. Basic local alignment search tool.
APHIS. Animal and Plant Health Inspection Service.
CDC. Centers for Disease Control and Prevention.
HHS. Health and Human Services.
USDA. United States Department of Agriculture.
USFDA. United States Food and Drug Administration.
NIAID. National Institute of Allergies and Infectious Diseases.
WHO. World Health Organization.

Authors' contributions

CEZ and ML designed the database with input from TK. CEZ and ML designed and ML and TK built the pipeline and web interface tools. MDD and JS built software modules to facilitate porting of the data and interfaces beyond the LLNL firewall. AZ and BV contributed to design concepts for applications to protein-based studies. TS initiated and CEZ designed and managed the project. All authors read and approved the final manuscript.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48 and was supported by funding from the Department of Homeland Security.

References

1. Slezak T, Kuczmarski T, Ott L, Torres C, Mederos D, Smith J, Truitt B, Mulakken N, Lam M, Vitalis E, Zemla A, Zhou C, Gardner S: **Comparative genomics tools applied to bioterrorism defense.** *Briefings in Bioinformatics* 2003, **4**:133-149.
2. Zhou CEZ, Zemla A, Roe D, Young M, Lam M, Schoeinger J, Balhorn R: **Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin.** *Bioinformatics* 2005, **21**:3085-3096.
3. **APHIS Agricultural Select Agent Program select agent and toxin list** [http://www.aphis.usda.gov/programs/ag_selectagent/ag_bioterr_toxinslist.html].
4. **CDC bioterrorism agents/diseases list** [http://www.bt.cdc.gov/agent/agentlist-category.asp].
5. **HHS and USDA select agents and toxins list** [http://www.cdc.gov/od/sap/docs/salist.pdf].
6. **USFDA Bad Bug Book** [http://www.cfsan.fda.gov/~mow/intro.html].
7. **NIAID category A, B and C priority pathogens** [http://www3.niaid.nih.gov/biodefense/bandc_priority.htm].
8. **WHO list of major zoonotic diseases** [http://www.who.int/zoonoses/diseases/en/].
9. **WHO list of diseases covered by the Epidemic and Pandemic Alert and Response (EPR)** [http://www.who.int/csr/disease/en/].
10. Andrade MA, Brown NP, Leroy C, Hoersh S, de Daruvar A, Reigh C, Franchini A, Tamames J, Valencia A, Ousounis C, Sander C: **Automated genome sequence analysis and annotation.** *Bioinformatics* 1999, **15**:391-412.
11. Frishman D, Albermann K, Hari J, Heumann K, Metanomski A, Zollner A, Mewes H-W: **Functional and structural genomics using PEDANT**. *Bioinformatics* 2001, **17**:44-57.
12. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJA, Lachaize C, Veuthey A-L, Gasteiger E, Bairoch A: **Automated annotation of microbial proteomes in SWISS-PROT.** *Computational Biology and Chemistry* 2003, **27**:49-58.

13. Goesmann A, Linke B, Bartels D, Dondrup M, Drause L, Neuweger H, Oehm S, Paczian T, Wilke A, Meyer F: **BRIGEP—the BRIDGE-based genome-transcriptome-proteome browser.** *Nucleic Acids Research* 2005, **33**:W710-W716.
14. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto P, Ivanova N, Kyrpides NC: **The integrated microbial genomes (IMG) system: a case study in biological data management.** In *Proceedings of the 31st VLDB Conference: 2005; Trondheim, Norway*. 2005:1067-1078.
15. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Puhler A: **GenDB—an open source genome annotation system for prokaryote genomes.** *Nucleic Acids Research* 2003, **31**:2187-2195.
16. Peterson JD, Umayam LA, Dickinson TM, Hickey EK, White O: **The comprehensive microbial resource.** *Nucleic Acids Research* 2001, **29**:123-125.
17. Pruitt KD, Tatusova T, and Maglott DR: **NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Research* 2005, **33**:D501-D504.
18. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, pascal G, Scarpelli C, Medigue C: **MaGe: a microbial genome annotation system supported by synteny results.** *Nucleic Acids Research* 2006, **34**:53-65.
19. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS: **BASys: a web server for automated bacterial genome annotation.** *Nucleic Acids Research* 2005, **33**:W455-W459.
20. **MvirDB microbial virulence database** [http://mvirdb.llnl.gov].
21. Blom N, Hansen J, Blaas D, and Brunak S: **Cleavage site analysis in picornaviral polyproteins: Discovering cellular targets by neural networks**. *Protein Science* 1996, **5**:2203-2216.
22. Bendtsen JD, Nielsen H, von Heijne G, and Brunak S: **Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology** 2004, **340**:783-795.
23. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL: **PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.** *Bioinformatics* 2005, **21**:617-623.
24. Nakai K, Horton P: **PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization.** *Trends in Biochemical Science* 1999, **24**:34-35.
25. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *Journal of Molecular Biology* 2001, **305**:567-580.
26. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *Journal of Molecular Biology* 2000, **300**:1005-1016.
27. Claros MG, von Heijne G: **TopPred II: An improved software for membrane protein structure predictions.** *CABIOS* 1994, **10**:685-686.
28. Tusnady GE, Simon I: **Principles governing amino acid composition of integral membrane proteins: applications to topology prediction.** *Journal of Molecular Biology* 1998, **283**:489-506.

Figure legends

Figure 1. Data flow diagram for MannDB sequence analysis pipeline. External data sources (yellow) are downloaded into MannDB. Software systems (lavender boxes) process and enable display of data. MannDB pipeline manager controls execution of open-source tools (ovals) and blast against MvirDB (green oval).

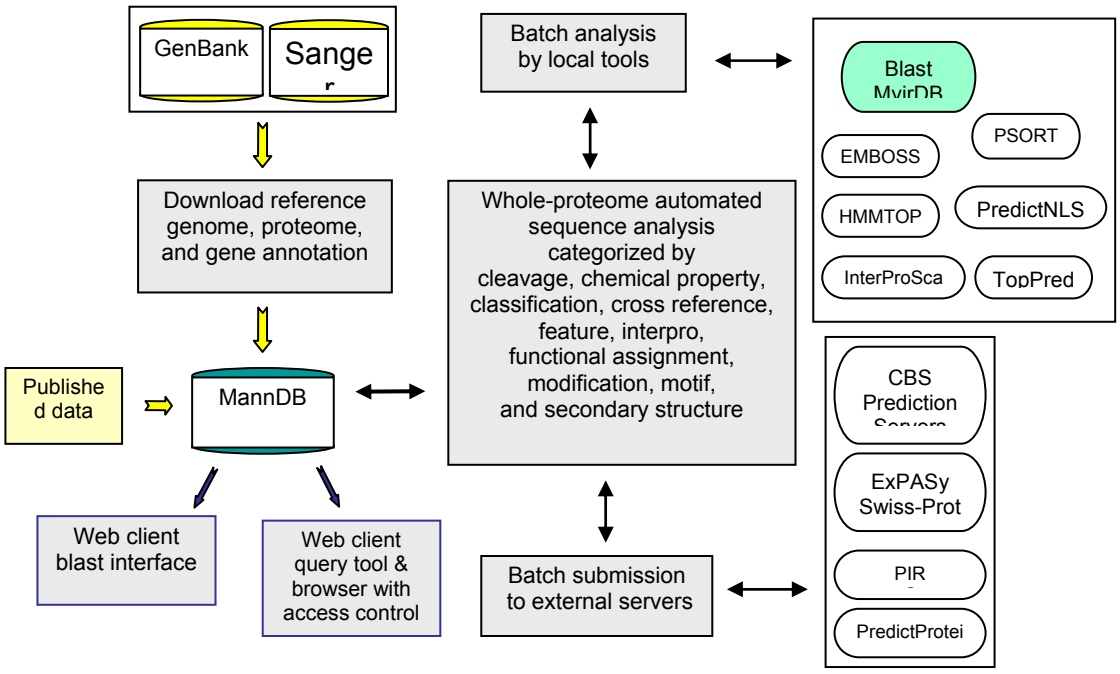

Figure 2. MannDB database query and browser sample web pages. In this example, user has selected the *Campylobacter jejuni* proteome (left), entered free text "toxin" (top oval), and checked the MvirDB homology checkbox (bottom oval), resulting in 3 database hits (top right). Selecting single chain protein id 64721 (top right, oval), followed by the "cross-reference" checkbox (middle right, oval) brings up a report page (bottom right) displaying the MvirDB cross reference link (oval).