Immunome Research

BioMed Central

Database

Open Access

Protein sequence database for pathogenic arenaviruses

Huynh-Hoa Bui[1], Jason Botten[2], Nicolas Fusseder[1], Valerie Pasquetto[1], Bianca Mothe[3], Michael J Buchmeier[2] and Alessandro Sette*[1]

Address: [1]La Jolla Institute for Allergy and Immunology, Division of Vaccine Discovery, 9420 Athena Circle, La Jolla, CA 92037, USA, [2]The Scripps Research Institute, Molecular & Integrative Neurosciences Department, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA and [3]California State University, Department of Biology, San Marcos, CA 92096, USA

Email: Huynh-Hoa Bui - hbui@liai.org; Jason Botten - jbotten@scripps.edu; Nicolas Fusseder - fusseder@liai.org; Valerie Pasquetto - valerie@liai.org; Bianca Mothe - bmothe@csusm.edu; Michael J Buchmeier - buchm@scripps.edu; Alessandro Sette* - alex@liai.org

* Corresponding author

Published: 8 February 2007

Immunome Research 2007, **3**:1 doi:10.1186/1745-7580-3-1

Received: 20 October 2006
Accepted: 8 February 2007

This article is available from: http://www.immunome-research.com/content/3/1/1

© 2007 Bui et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Arenaviruses are a family of rodent-borne viruses that cause several hemorrhagic fevers. These diseases can be devastating and are often lethal. Herein, to aid in the design and development of diagnostics, treatments and vaccines for arenavirus infections, we have developed a database containing protein sequences from the seven pathogenic arenaviruses (Junin, Guanarito, Sabia, Machupo, Whitewater Arroyo, Lassa and LCMV).

Results: The database currently contains a non-redundant set of 333 protein sequences which were manually annotated. All entries were linked to NCBI and cited PubMed references. The database has a convenient query interface including BLAST search. Sequence variability analyses were also performed and the results are hosted in the database.

Conclusion: The database is available at http://epitope.liai.org:8080/projects/arena and can be used to aid in studies that require proteomic information from pathogenic arenaviruses.

Background

Arenaviridae are a family of viruses whose members are associated with rodent-transmitted disease in humans. Each virus usually is associated with a particular rodent host species in which it is maintained. Arenavirus infections, occur when an individual comes into contact with the excretions of an infected rodent, are relatively common in humans in some area of the world and primarily cause hemorrhagic fevers, including Lassa fever (LF; Lassa virus), Argentine hemorrhagic fever (AHF; Junin virus), Bolivian hemorrhagic fever (BHF; Machupo virus), Venezuelan hemorrhagic fever (VHF; Guanarito virus) and Brazillian hemorrhagic fever (BrHF; Sabia virus) [1-6]. These diseases can be devastating and often lethal. Lymphocytic choriomeningitis virus (LCMV), a known human teratogen, can cause aseptic meningitis [7-9], and Whitewater Arroyo Virus (WWA) was recently attributed to two deaths in California [10,11].

The arenaviruses can be classified phylogenetically into Old World (which includes LCMV and Lassa virus) and New World; this latter group has been further divided into three lineages, A-C [12,13]. Except for WWA virus which belongs to lineage A, the four most pathogenic New World agents (Junin, Machupo, Guanarito and Sabia viruses) all belong to lineage B, suggesting that the highly

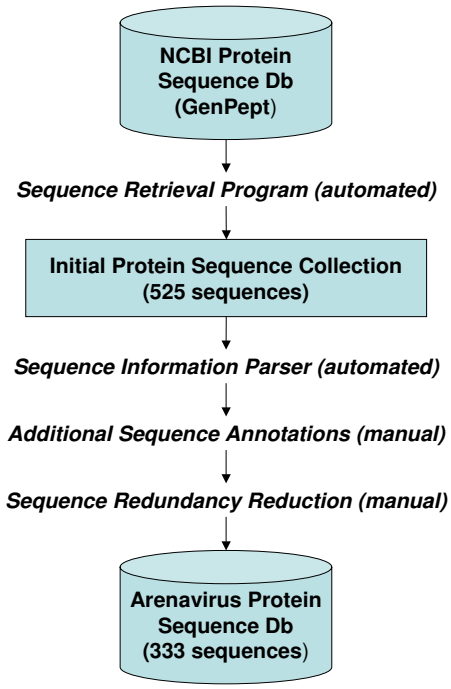


Figure 1
Arenavirus protein sequence database construction flowchart.

pathogenic phenotype may derive from a common ancestral virus [12,14]. All of these viruses cause significant morbidity and mortality. Lassa virus and other hemorrhagic fever arenaviruses (Junin, Machupo, Guanarito and Sabia) are included in category A of potential bioterrorism microbial weapons [15].

Currently, there are no virus-specific treatments approved for use against arenavirus hemorrhagic fevers. Ribavirin is the only compound that has shown partial efficacy against some arenavirus infections [16] (successful against human Lassa infections only if given within the first week following disease onset [17]), and to date only one vaccine (against AHF) has been evaluated in humans [2]. Because of its severe morbidity and high mortality together with lack of immunization or effective treatment, scientists and researchers are challenged with developing containment, treatment, and vaccine strategies for arenavirus infection. For the purpose of developing diagnostic reagents and designing novel vaccine constructs, our group has been conducting active studies in identifying MHC class I and II restricted T cell epitopes from pathogenic arenaviruses. As a component of the studies, we have compiled and developed a database of protein sequences for the seven arenaviruses (Lassa, LCMV, Junin, Guanarito, Sabia, Machupo and WWA) known to cause disease in humans. Herein, we make this database available as a public resource to aid in studies that require proteomic information from pathogenic arenaviruses.

Construction and content

Arenaviridae are enveloped viruses with a genome consisting of two single-stranded RNA, the small (S) and the large (L), segments. Each segment encodes two different proteins. The S RNA encodes the nucleocapsid protein (NP) and the glycoprotein precursor (GPC) which undergoes post-translational processing to yield two mature proteins (GP1 and GP2) [18]. The L RNA encodes the viral RNA-dependent RNA polymerase (L) and a zinc-binding matrix protein (Z) [19,20]. These four proteins (GPC, L, NP and Z) are the collection targets of our database.

The process of compiling arenaviral protein sequences for the database includes 1) retrieving published sequences from NCBI, 2) parsing for sequence information, 3) manually verifying the information and performing additional annotations, and 4) removing duplicated entries. A schematic flow chart of this process is shown in Figure 1. MySQL was used as the storage database engine, Tomcat as the webserver, and Java servlets were used to develop the web interface.

To obtain protein sequences encoded by the seven pathogenic arenaviruses, we first searched the NCBI database through the use of an automated computer program developed in our laboratory. This program was written to: (1) search for protein sequence IDs, (2) retrieve protein sequence records and (3) parse the records into annotated fields. First, protein sequence IDs (GI numbers) were retrieved via an NCBI Esearch programming utility using NCBI taxonomy IDs as search parameters. Next, GI numbers were used to retrieve protein sequence records in the GenPept format. "NCBISequenceDB" java class from biojava 1.4 package was used to programmatically retrieve the GenPept sequence records. Finally, a customized java class was written to parse each record into annotated fields which include protein sequence data, source references, virus, strain and gene name.

Using the automated program, 525 protein sequences were retrieved from NCBI. Next, verification of sequence authenticity was performed via supporting publications and/or written summaries. Finally, we conducted manual protein sequence alignments to identify and remove duplications. Of the 525 protein sequences from the initial cohort, a total of 333 unique protein sequences from one or more strains of the seven pathogenic arenavirus

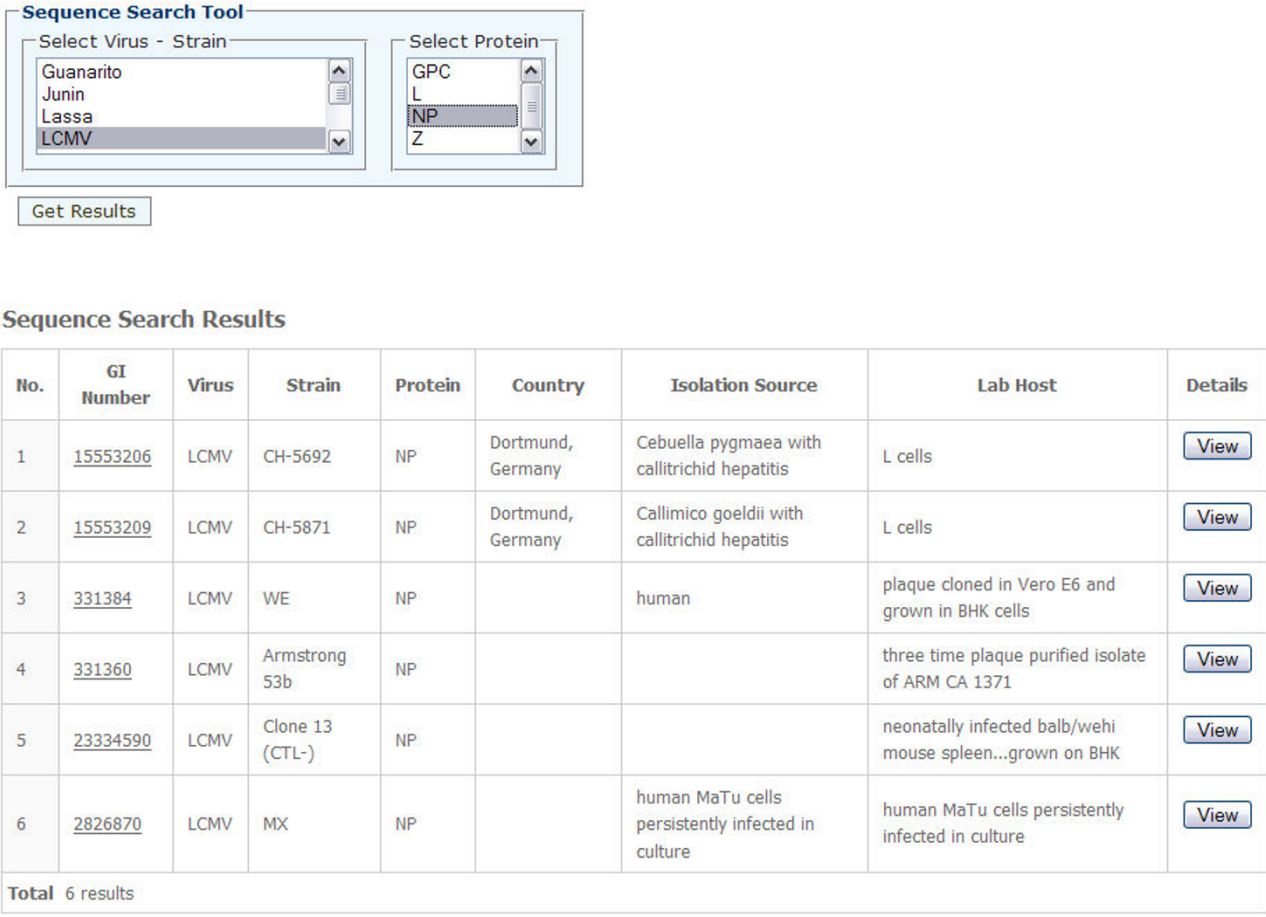


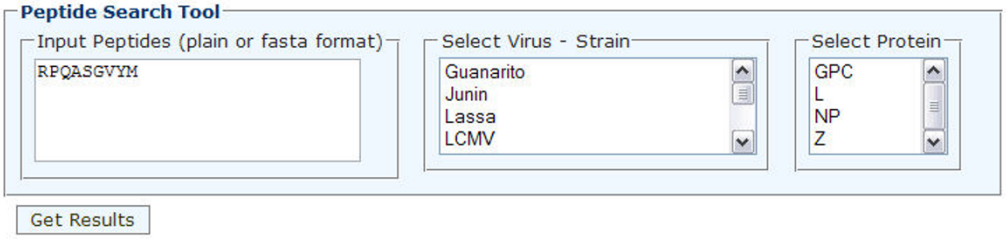
Figure 2
Sequence search interface.

viruses were obtained (Table 1). At present, the Z protein sequence has not been published for WWA virus. As a result of renewed interests in arenaviruses, we anticipate that more protein sequences will become available in the near future. We plan to periodically monitor the scientific literature and the NCBI database for new sequence depositions, and update our database accordingly.

Utility and discussion

Arenavirus protein sequence annotation

To maximize the usefulness of the arenavirus protein sequence database to the scientific community, each record was annotated with specific information, including the host and geographical region from which each protein sequence was isolated and the passage history of each viral strains between its original isolation from its natural host and the time it was sequenced. Inclusion of the host that each protein sequence was isolated from is of potential relevance in studies examining specific host-derived immune pressure or host-specific viral adaptations. The inclusion, if available, of the geographical region is relevant in determining whether the available viral strains are represented in endemic locations. Finally, the passage history of each strain is relevant in the context of the high mutation rates associated with these RNA viruses and the potential for genetic changes to accumulate as a result of *in vitro* passage. Mutations generated as a result of viral passage in non-reservoir animals or cell lines would not be representative of the natural variation present in field or clinical strains. All annotated information was obtained via collected publications and/or via direct correspondence with the authors of a given protein sequence. Most protein sequences were derived from human infections, while the remaining samples came from naturally



Peptide Search Results

No.	Peptide Description	Peptide Sequence	Peptide Length	Peptide Position(s)	GI Number	Virus - Strain	Protein	Country	Isolation Source	Lab Host	Details
1		RPQASGVYM	9	118 - 126	15553206	LCMV - CH-5692	NP	Dortmund, Germany	Cebuella pygmaea with callitrichid hepatitis	L cells	View
2		RPQASGVYM	9	118 - 126	15553209	LCMV - CH-5871	NP	Dortmund, Germany	Callimico goeldii with callitrichid hepatitis	L cells	View
3		RPQASGVYM	9	118 - 126	331384	LCMV - WE	NP		human	plaque cloned in Vero E6 and grown in BHK cells	View

Figure 3
Peptide search interface.

infected reservoir rodents. Universally, each of the viruses sequenced prior to 2002 was propagated in Vero E6 or BHK cell lines prior to sequencing of the viral genome.

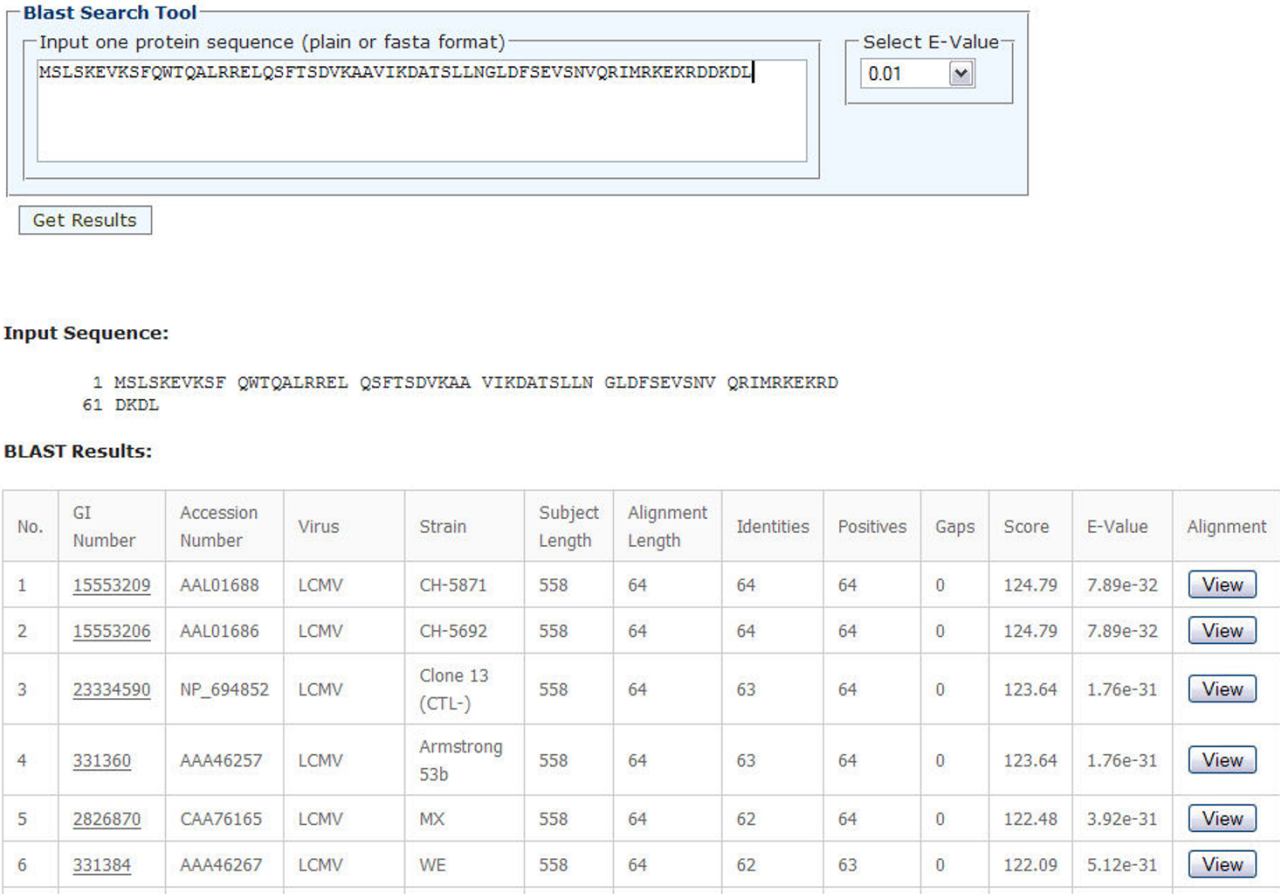
Search Interface

The arenavirus protein sequence database has a convenient search interface allowing querying by virus, strain and protein names (Figure 2). All entries in the database are linked to the original NCBI records and cited PubMed references (if available). In addition, a tool utility is also provided that allows searching for arenavirus sequences containing specific peptides or epitope sequences (Figure 3). For example, this would allow researchers to quickly determine whether known epitopes are expressed by various arenavirus strains and species. This information could therefore be used to develop arenaviral epitope-based diagnostics and/or vaccine constructs. A BLAST search was also implemented allowing users to search for similar sequences contained within the database (Figure 4).

Arenavirus protein sequence variability analysis

Using the sequences in this database, we further investigated the arenaviral protein sequence conservancy/variability. Our goal was to identify conserved or variable regions that could be targeted for development of a universal arenaviral vaccine or diagnostics, respectively. To do this, we performed multiple sequence alignments and entropy analyses between different strains of a virus and between different viruses.

Multiple sequence alignments were performed using CLUSTAL W program [21] using default parameters. To



No.	GI Number	Accession Number	Virus	Strain	Subject Length	Alignment Length	Identities	Positives	Gaps	Score	E-Value	Alignment
1	15553209	AAL01688	LCMV	CH-5871	558	64	64	64	0	124.79	7.89e-32	View
2	15553206	AAL01686	LCMV	CH-5692	558	64	64	64	0	124.79	7.89e-32	View
3	23334590	NP_694852	LCMV	Clone 13 (CTL-)	558	64	63	64	0	123.64	1.76e-31	View
4	331360	AAA46257	LCMV	Armstrong 53b	558	64	63	64	0	123.64	1.76e-31	View
5	2826870	CAA76165	LCMV	MX	558	64	62	64	0	122.48	3.92e-31	View
6	331384	AAA46267	LCMV	WE	558	64	62	63	0	122.09	5.12e-31	View

Figure 4
BLAST search interface.

estimate the diversity a multiple protein sequence alignment, Shannon entropy (H) was calculated using equation 1 [22]:

$$H = -\sum_{i=1}^{M} P_i \log_2 Pi \qquad (1)$$

where P_i is the fraction of residues of amino acid type *i*, and M is the number of amino acid types (20). H ranges from 0 (only one residue in present at that position) to 4.322 (all 20 residues are equally represented in that position). Typically, positions with $H \geq 2.0$ are considered variable, whereas those with $H \leq 2$ are consider conserved. Highly conserved positions are those with $H \leq 1.0$ [23].

Shannon entropy analyses of protein sequences contained in our database indicated that arenavirus protein sequences are fairly conserved between different strains of the same virus, but less so between different viruses. This is consistent with the view that arenaviruses are relatively stable genetically with amino acid sequence homologies of 90–95% among different strains of the same virus species and of 44–63% for homologous proteins of different arenavirus species [24]. As a result, to develop a universal vaccine against different arenaviruses, a construct that contains epitopes conserved within each virus should be used. For the purpose of developing diagnostics, however, epitopes derived from non-conserved regions would be excellent candidates.

The most important arenaviral proteins are NP and GPC, and the NP proteins have been known as being the most conserved among arenaviruses. Our entropy analysis also revealed that the NP protein has a highly distinct inter-virus conserved region between residues 1–310 with average $H \approx 0.5$ (Figure 5). Another distinct conserved region

Table 1: Arenavirus protein sequence distribution

	Protein				
Virus	**GPC**	**L**	**NP**	**Z**	***Total***
Guanarito	1	1	32	1	***35***
Junin	43	3	45	2	***93***
Lassa	12	7	64	6	***89***
LCMV	10	9	6	5	***30***
Machupo	28	4	30	3	***65***
Sabia	1	1	1	1	***4***
Whitewater Arroyo	2	1	14	0	***17***
Total	***97***	***26***	***192***	***18***	***333***

is in the GPC protein between residues 290–500 (Figure 6). As a result, epitopes derived from these regions have high probabilities to be cross-reactive between different arenaviruses. In contrast to long conserved regions observed in the NP and GPC proteins, the L and Z proteins have much shorter inter-virus conserved regions which could be related to the proteins' shared functional homology.

As curated in the Immune Epitope Database (IEDB) [25,26] and at the time of this analysis, epitopes derived from arenaviruses (mainly Lassa and LCMV), were exclusively from NP and GPC proteins. The majority of these epitopes were mouse MHC class I restricted and located in the conserved regions of NP and GPC proteins (data not shown). This indicates the identified T cell epitopes from NP and GPC proteins may cross-react among different arenavirus species. Nevertheless, whether these mouse MHC restricted epitopes would also be reactive in humans remains to be experimentally validated. It should be noted here that the lack of L and Z derived epitopes, as reported in IEDB, may imply that the curation is incomplete or more likely that no studies have yet been done to look for epitopes in these proteins.

Conclusion

In conclusion, the database developed here, to our knowledge, is the only public resource that provides a non-redundant complete set of viral protein sequences for the seven highly pathogenic arenaviruses. These protein sequences can be used for epitope discovery studies, and their specific annotations are highly relevant for consideration in the complex task of developing diagnostics and/or vaccines. In another aspect, this database would also be a useful resource for scientists to investigate function-sequence conservation relationships among the arenaviruses.

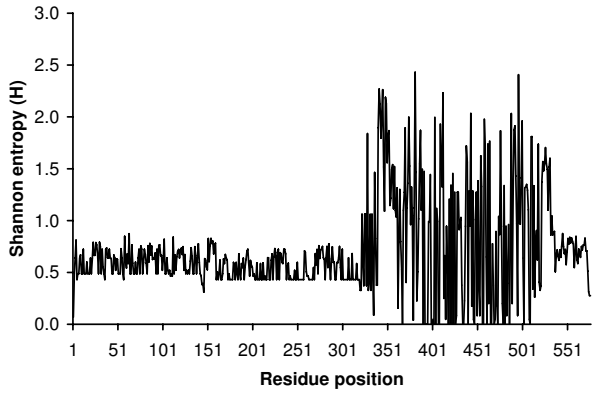


Figure 5
Inter-virus sequence diversity of NP protein.

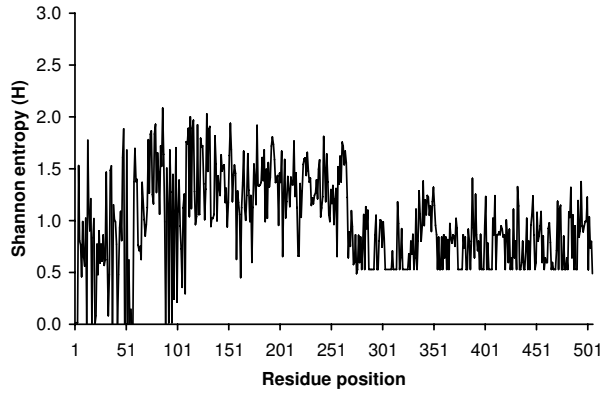


Figure 6
Inter-virus sequence diversity of GPC protein.

Availability and requirements

Project name: Arenavirus protein sequence database

Project homepage: http://epitope.liai.org:8080/projects/arena

Programming language: Java

Operating system: Fedora Linux

Other requirements: Apache Tomcat 5.5.12, MySQL 4.1, Firefox version 1.5 or higher

License: None

List of abbreviations used

AHF: Argentine hemorrhagic fever

BHF: Bolivian hemorrhagic fever

BrHF: Brazillian hemorrhagic fever

GPC: Glycoprotein

IEDB: Immune Epitope Database and Analysis Resources

LCMV: Lymphocytic choriomeningitis virus

LF: Lassa fever

MHC: Major Histocompatibility complex

NCBI: National Center for Biotechnology Information

NP: Nucleoprotein

VHF: Venezuelan hemorrhagic fever

WWA: Whitewater Arroyo

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

HHB developed the database and performed sequence variability analyses. JB manually annotated of the sequences. NF programmed the web interface. HHB, JB and AS wrote the manuscript. All authors participated in discussions, reviewed and approved the final manuscript version.

Acknowledgements

This work was supported by the National Institutes of Health's contract HHSN266200400023C (Class I and Class II Restricted Epitopes from a Representative Sample of the Different Arenavirus Species Pathogenic in Humans), NIH grants AI50840 to MB, T32 AI07354 and F32 AI056827 to JB, and Kirin pharmaceutical division. This is LIAI publication number 820.

References

1. McCormick JB: **Epidemiology and control of Lassa fever.** *Curr Top Microbiol Immunol* 1987, **134:**69-78.
2. Maiztegui JI, McKee KT Jr., Barrera Oro JG, Harrison LH, Gibbs PH, Feuillade MR, Enria DA, Briggiler AM, Levis SC, Ambrosio AM, Halsey NA, Peters CJ: **Protective efficacy of a live attenuated vaccine against Argentine hemorrhagic fever. AHF Study Group.** *J Infect Dis* 1998, **177(2):**277-283.
3. Maiztegui JI: **Clinical and epidemiological patterns of Argentine haemorrhagic fever.** *Bull World Health Organ* 1975, **52(4-6):**567-575.
4. Salas R, de Manzione N, Tesh RB, Rico-Hesse R, Shope RE, Betancourt A, Godoy O, Bruzual R, Pacheco ME, Ramos B, *et al.*: **Venezuelan haemorrhagic fever.** *Lancet* 1991, **338(8774):**1033-1036.
5. de Manzione N, Salas RA, Paredes H, Godoy O, Rojas L, Araoz F, Fulhorst CF, Ksiazek TG, Mills JN, Ellis BA, Peters CJ, Tesh RB: **Venezuelan hemorrhagic fever: clinical and epidemiological studies of 165 cases.** *Clin Infect Dis* 1998, **26(2):**308-313.
6. Lisieux T, Coimbra M, Nassar ES, Burattini MN, de Souza LT, Ferreira I, Rocco IM, da Rosa AP, Vasconcelos PF, Pinheiro FP, *et al.*: **New arenavirus isolated in Brazil.** *Lancet* 1994, **343(8894):**391-392.
7. Barton LL, Budd SC, Morfitt WS, Peters CJ, Ksiazek TG, Schindler RF, Yoshino MT: **Congenital lymphocytic choriomeningitis virus infection in twins.** *Pediatr Infect Dis J* 1993, **12(11):**942-946.
8. Larsen PD, Chartrand SA, Tomashek KM, Hauser LG, Ksiazek TG: **Hydrocephalus complicating lymphocytic choriomeningitis virus infection.** *Pediatr Infect Dis J* 1993, **12(6):**528-531.
9. Wright R, Johnson D, Neumann M, Ksiazek TG, Rollin P, Keech RV, Bonthius DJ, Hitchon P, Grose CF, Bell WE, Bale JF Jr.: **Congenital lymphocytic choriomeningitis virus syndrome: a disease that mimics congenital toxoplasmosis or Cytomegalovirus infection.** *Pediatrics* 1997, **100(1):**E9.
10. Fulhorst CF, Bowen MD, Ksiazek TG, Rollin PE, Nichol ST, Kosoy MY, Peters CJ: **Isolation and Characterization of Whitewater Arroyo Virus, a novel North American Arenavirus.** *Virol* 1996, **224:**114-120.
11. Byrd RG, Cone LA, Commess BC, Williams-Herman D, Rowland JM, Lee B, Fitzgibbons MW, Glaser CA, Jay MT, Fritz CI, Ascher MS, Cheung M, Kramer VL, Reilly K, Vugia DJ, Fulhorst CF, Milazzo ML, R.N. C: **Fatal Illness Associated with a new world arenavirus-California, 1999-2000.** *Morbidity and Mortality Weekly Report* 2000, **49(August 11):**709-711.
12. Bowen MD, Peters CJ, Nichol ST: **Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between arenaviruses and their rodent hosts.** *Mol Phylogenet Evol* 1997, **8(3):**301-316.
13. Emonet S, Lemasson JJ, Gonzalez JP, de Lamballerie X, Charrel RN: **Phylogeny and evolution of old world arenaviruses.** *Virology* 2006, **350(2):**251-257.
14. Bowen MD, Peters CJ, Nichol ST: **The phylogeny of New World (Tacaribe complex) arenaviruses.** *Virol* 1996, **219(1):**285-290.
15. Borio L, Inglesby T, Peters CJ, Schmaljohn AL, Hughes JM, Jahrling PB, Ksiazek T, Johnson KM, Meyerhoff A, O'Toole T, Ascher MS, Bartlett J, Breman JG, Eitzen EM Jr., Hamburg M, Hauer J, Henderson DA, Johnson RT, Kwik G, Layton M, Lillibridge S, Nabel GJ, Osterholm MT, Perl TM, Russell P, Tonat K: **Hemorrhagic fever viruses as biological weapons: medical and public health management.** *Jama* 2002, **287(18):**2391-2405.
16. Enria DA, Maiztegui JI: **Antiviral treatment of Argentine hemorrhagic fever.** *Antiviral Res* 1994, **23(1):**23-31.
17. McCormick JB, King IJ, Webb PA, Scribner CL, Craven RB, Johnson KM, Elliott LH, Belmont-Williams R: **Lassa fever. Effective therapy with ribavirin.** *N Engl J Med* 1986, **314(1):**20-26.
18. Buchmeier MJ, Oldstone MB: **Protein structure of lymphocytic choriomeningitis virus: evidence for a cell-associated precursor of the virion glycopeptides.** *Virol* 1979, **99(1):**111-120.
19. Salvato MS, Shimomaye EM: **The completed sequence of lymphocytic choriomeningitis virus reveals a unique RNA structure and a gene for a zinc finger protein.** *Virol* 1989, **173(1):**1-10.
20. Buchmeier MJ, Bowen MD, Peters CJ: **Arenaviridae: The Viruses and Their Replication.** *Fields Virology* 2001, **2:**1635-1668.

21. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673-4680.
22. Shannon CE: **The mathematical theory of communication.** *The Bell system Technical Journal* 1948, **27:**379-423 & 623-656.
23. Litwin S, Jores R: **.** In In *theoretical and experimental insights into immunology* Edited by: Perelson, A. S., Weisbuch G. Berlin , Springer-Verlag; 1992.
24. Sevilla N, Domingo E, de la Torre JC: **Contribution of LCMV towards deciphering biology of quasispecies in vivo.** *Curr Top Microbiol Immunol* 2002, **263:**197-220.
25. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A: **The immune epitope database and analysis resource: from vision to blueprint.** *PLoS Biol* 2005, **3(3):**e91.
26. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger SP, Stewart S, Surko P, Way S, Wilson S, Sette A: **The design and implementation of the immune epitope database and analysis resource.** *Immunogenetics* 2005, **57(5):**326-336.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral