VKCDB: voltage-gated K$^+$ channel database updated and upgraded

Warren J. Gallin* and Patrick A. Boutet

Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

ABSTRACT

The Voltage-gated K$^+$ Channel DataBase (VKCDB) (http://vkcdb.biology.ualberta.ca) makes a comprehensive set of sequence data readily available for phylogenetic and comparative analysis. The current update contains 2063 entries for full-length or nearly full-length unique channel sequences from Bacteria (477), Archaea (18) and Eukaryotes (1568), an increase from 346 solely eukaryotic entries in the original release. In addition to protein sequences for channels, corresponding nucleotide sequences of the open reading frames corresponding to the amino acid sequences are now available and can be extracted in parallel with sets of protein sequences. Channels are categorized into subfamilies by phylogenetic analysis and by using hidden Markov model analyses. Although the raw database contains a number of fragmentary, duplicated, obsolete and non-channel sequences that were collected in early steps of data collection, the web interface will only return entries that have been validated as likely K$^+$ channels. The retrieval function of the web interface allows retrieval of entries that contain a substantial fraction of the core structural elements of VKCs, fragmentary entries, or both. The full database can be downloaded as either a MySQL dump or as an XML dump from the web site. We have now implemented automated updates at quarterly intervals.

INTRODUCTION

Voltage-gated K$^+$ channels (VKCs) constitute a structurally related family of intrinsic membrane proteins that respond to changes in transmembrane potential by opening and closing an ion-selective permeation pathway for K$^+$ ions (1). The sensitivity to the membrane potential and the kinetics of the response to changes in potential vary substantially between the different VKC proteins, which means that cells expressing different VKCs repolarize at different rates and at different parts of an action potential. As such, VKCs are the primary agents in shaping action potentials in excitable cells of the eumetazoa.

VKCs are tetramers, each subunit of which comprises a core structural domain consisting of six transmembrane helices and a re-entrant loop that forms the ion-selective channel (Figure 1) and highly variable C- and N-terminal domains that play a role in appropriate assembly of newly synthesized tetrameric channels (2), transport to the appropriate cell compartment (usually the plasma membrane) (3) and in modulating the core functionality of the channel (4). The core structural domain is sufficiently conserved that robust multiple alignments between VKCs of single subfamilies can be obtained based on sequence, and alignments between families can be obtained based on a combination of sequence and structural similarity. The N- and C-terminal domains are usually specific to a given subfamily, and thus give robust multiple alignments within subfamilies but not between subfamilies.

The existence of a diverse cohort of VKC paralogs within individual species of metazoa and of divergent orthologs between species indicates that the functional evolution of this family of proteins has been a significant factor in the evolution of electrophysiological excitability in the animal kingdom.

The original VKCDB (5) has been updated and expanded to provide support for evolutionary and comparative studies of the relationships between VKC sequence, electrophysiological characteristics of the individual channel proteins and ultimately the complex electrophysiological behavior of neurons and muscles in animals.

CONTENT UPDATE

Content update of VKCDB consists of two phases that are automated to run at quarterly intervals. The first is update of the information for existing records. For each record, the last update date of the record in VKCDB is compared

---

*To whom correspondence should be addressed. Tel: +780 492 1285; Fax: +780 492 9234; Email: wgallin@ualberta.ca

to the last update date of the corresponding GENBANK record. If the GENBANK record has been replaced, the replacement record is parsed to update the VKCDB entry. If the GENBANK record has only had new information added to it (for example, additional bibliographic references or new sequence annotations), the full record is parsed and the new information is written into the VKCDB record along with a new date-of-last-update. If the GENBANK record has been removed, the VKCDB record is flagged as no longer being a recognized channel sequence, but it is left in the database, and the replacement entry (if any) is added to the database.

The second phase is addition of new records for VKCs that were added to GENBANK after the last VKCDB update. Protein sequences for the bacterial VKCs, archaeal VKCs and each of the eukaryotic subfamilies are aligned (6) using MUSCLE v3.8 (http://www.drive5 .com/muscle/) used to create separate profiles (7) using hmmbuild from HMMER v3 (http://hmmer.janelia .org/). These profiles are then used to search the most recent GENBANK non-redundant protein database (downloaded in fasta format and searched locally) using hmmsearch (from HMMER) to identify potential new entries. The results of all the HMMER searches are pooled and redundant entries are removed. Each unique new record is tagged with name of the subfamily whose profile gives the highest full-sequence HMMER score and the results are then sorted based on this provisional subfamily designation followed by sorting on the basis of the HMMER score. This allows for fairly easy manual evaluation of cutoffs that distinguish between real VKC sequences and other, more distantly related sequences.

The GI numbers of the manually selected set of new entries from the HMMER search are then used to retrieve the relevant GENBANK records and create new entries in VKCDB.
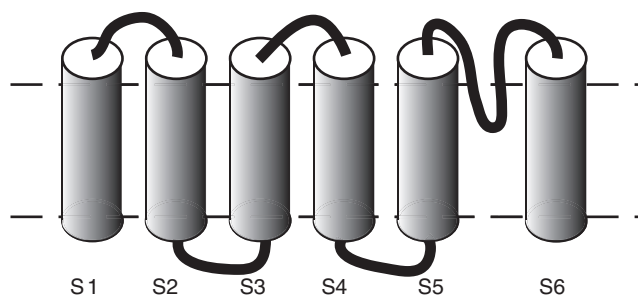
For the bacterial domain, the archaeal domain and each of the eukaryotic subfamilies, all previous entries of substantial length and all new entries are then subjected to a large multiple alignment to identify which of the new entries span a substantial portion of the core VKC structure (coded as 'fullish_length'). Small fragmentary records [those that are gapped in substantial areas of compact alignment of the core region, S1–S6 (Figure 1)] are flagged as not of sufficient length, thus providing the user with the option of selecting either nearly full length, fragmentary, or both in the standard web access. The remaining protein sequences are then realigned, the multiple alignment is pruned to remove any sites with >5% gaps, and the resulting data matrix is used for a Bayesian search (8) to generate a phylogenetic tree for each subfamily. This tree is used to check subfamily assignment made on the basis of the HMMER searches.

Table 1 gives the number of channel entries that have a substantial (fullish_length = Y) amount of protein sequence for channels in each distinct phylogenetic subfamily.

WEB ACCESS

Utilizing a customized LAMP (Linux, Apache, MySQL and PHP) setup to ensure stability, maintainability and future scaling needs, we have revised the web interface to VKCDB to provide a finer level of selection, detail and usability than the previous version. Corresponding sets of protein and nucleotide sequence data are now available for download from any viewed record on any page.

Figure 1. Schematic drawing of the conserved functional core of VKCs. The dotted lines represent the lipid bilayer of the membrane, with the extracellular compartment at the top of the drawing, intracellular at the bottom. The functional pore consists of the S5 and S6 transmembrane helices supporting the re-entrant loop that joins them and that forms the ion selective pore in the assembled tetramer. The voltage-sensing domain comprises S1 through S4. The array of positive charges every third residue in the S4 helix is the primary sensor of the voltage, with the other three helices forming a supporting scaffold around the S4 helix. The cytoplasmic N- and C-terminal domains (not shown in this schematic) are much more variable than the core structure.

Table 1. Table of different VKC families in current release of VKCDB

Channel family	Number of Entries (fullish length)
Kv1	270
Kv2	71
Kv3	180
Kv4	132
Kv5	21
Kv6	78
Kv7	191
Kv8	20
Kv8.2	18
Kv9	40
Kv10	99
Kv11	147
Kv12	107
Kv13	19
BK	151
Archaea	18
Bacteria	477
KCNH (unicellular)	24
Total	2063

These family designations are based on HMMer searches using profiles generated from well characterized, channel sequences. The KCNH family includes the Kv10, Kv11 and Kv12 families from Metazoa, but these families arose after the origin of Metazoa, so the unicellular KCNH channels do not fall into any of these families.

There are three options for searching VKCDB, accessible from links at the top of every web page.

The first, 'Search Database' allows for searching on the contents of various fields (Figure 2). This is designed provide users the ability to find entries for which they already have some identifying information, including the VKC ID number, the GENBANK ID number, accession number or authors of the paper reporting the sequence.

The second, 'Browse Database' is designed to allow comprehensive retrieval of sets of VKC sequences that are of general use in comparative and phylogenetic analyses. The 'Browse Database' section is divided into three subsections: 'By Family', 'By Organism' and 'By Electrophysiology'.

The 'By Organism' page allows all channels from a single organism to be retrieved with a single query (Figure 3). Note that this can return multiple alternatively spliced transcripts from a single gene, or occasionally different variant sequences. The only entries that are flagged as duplicates in VKCDB are those with identical amino acid sequences from the same organism.

Alternatively, all members of a particular family of channels can be selected and downloaded (Figure 4) from the 'By Subfamily' page, or all sequences for which quantitative functional parameters (half-activation voltage, half-inactivation voltage and activation threshold) are identified in the database.

In all cases the amino acid sequence or the nucleic acid sequence can be retrieved in the same order with the same identifiers, to facilitate a variety of evolutionary analyses by making it straightforward to align nucleic acid sequences against a pre-existing protein sequence alignment. All downloaded sequences are in plain text FASTA format, so combining multiple sets into a single file can be accomplished by simple concatenation of the individual output files.

The identifier line for the FASTA-formatted data always begins with the VKC ID number, which is the primary key for protein sequences, and can be configured to also include the GI number, the accession number, the definition line, and the channel subfamily (Kv1 through Kv13 and BK) (Figure 4).

The third search method is VKCDB-BLAST, which is simply the BLAST+ search tools BLASTP and BLASTN (9) implemented over all VKCDB entries (either protein or nucleic acid) that have been confirmed as potassium channels. This allows for rapid searches using query sequences that are potential potassium channels and a rapid first-order determination of possible affinities to other VKCs.

The full VKCDB database, including entries that have been annotated as invalid or obsolete, is available in XML or SQL format, downloaded from the 'Tools→Database Downloads' page.

FUTURE PLANS

The main enhancements planned over the next year are:

(1) Addition of pre-aligned sequences for each of S1, S2, S3, S4 and the S5-pore-S6 module, to allow easy access to a standardized data set for phylogenetic and comparative analysis, and
(2) Addition of more functional electrophysiological parameters to provide a standardized set of data for structure/function analyses and
(3) Refinement of subfamily designations.

ACKNOWLEDGEMENTS

Figure 2. Screen capture of results of a search for a particular channel based on its VKC ID number. To select any entry that has been returned by a search it is necessary to check the box at the left side of the entry and choose either protein sequence or nucleotide sequence before clicking on the download button.

Figure 3. Screen capture of top of results page for a search for all channels from *Callithrix jacchus*, the common marmoset. This search requires (**a**) prefiltering the search for Eukaryotes, (**b**) choosing long, fragmentary or both records for retrieval and then (**c**) selection of species from a pull-down menu.

Figure 4. (**a**) Screen capture of top of search page for BK family with download options set to retrieve protein sequence with the GENBANK GI number, GENBANK accession number, definition line from GENBANK record and the channel subfamily name as part of the FASTA header and with all the entries checked for retrieval. (**b**) Top of the text file that was retrieved with the above settings. Note that the sequence is on one unbroken line, so this figure has truncated the right-hand side of the image.

REFERENCES

1. Hille,B. (2001) *Ionic Channels of Excitable Membranes*, 3rd edn. Sinauer Associates Inc., Sunderland, Mass.
2. Li,M., Jan,Y.N. and Jan,L.Y. (1992) Specification of subunit assembly by the hydrophilic amino-terminal domain of the Shaker potassium channel. *Science*, **257**, 1225–1230.
3. Vacher,H., Mohapatra,D.P. and Trimmer,J.S. (2008) Localization and targeting of voltage-dependent ion channels in mammalian central neurons. *Physiol. Rev.*, **88**, 1407–1447.
4. Lvov,A., Greitzer,D., Berlin,S., Chikvashvili,D., Tsuk,S., Lotan,I. and Michaelevski,I. (2009) Rearrangements in the relative orientation of cytoplasmic domains induced by a membrane-anchored protein mediate modulations in Kv channel gating. *J. Biol. Chem.*, **284**, 28276–28291.
5. Li,B. and Gallin,W.J. (2004) VKCDB: voltage-gated potassium channel database. *BMC Bioinformatics*, **5**, 3.
6. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
9. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.