Database of pore domain models from cationic channels

Ana Jerončić and Davor Juretić

Department of Physics, University of Split, N. Tesle 12,
21 000 Split, Croatia

The superfamily of P-region proteins is group of pore-forming subunits from functionally diverse cationic channels. These proteins represents a rich and relatively untapped vein of future therapeutic targets. Still, to realize their therapeutic potential, greater understanding of the underlying biology and physiology of channel function must be develop. Especially appealing in this context is the P-region pore domain, fragment of the protein that physically form the pore-wall and is central to the ion permeation and selectivity function.

We constructed a focused database of accurate models of P-region pore domain, with respect to precise sequence position of domain boundaries and main secondary structure elements, within. Database analysis was performed to deduce the future trends in ion channel research.

Database PPore can be interactively queried over the web at

http://split.pmfst.hr/PPore/index.html

Methods I

Homology transitivity and extensive literature browsing were used to collect the set of P-region superfamily members. If homology between (A,B) & (B,C) have been inferred from the literature then (A,C) were regarded as homologous. A, B, C represent (sub)families of cationic channels.

$$\left(A\xrightarrow{\text{homolog}}B\right).\text{and}.\left(B\xrightarrow{\text{homolog}}C\right)\Rightarrow\left(A\xrightarrow{\text{homolog}}C\right)$$

Data-mining was performed with filter:

{ { {“potassium channel” !MinK | { “calcium channel” !gamma !delta !alpha2delta} | “sodium channel”} !beta !amiloride !epithelial} | {{“glutamate receptor” !metabotropic} | {cyclic nucleotide | cyclic-nucleotide | CNG}}} !toxin !blocker !auxiliary !modulatory

constructed for full text search on SWISS-PROT/TREMBL databases.

Methods II

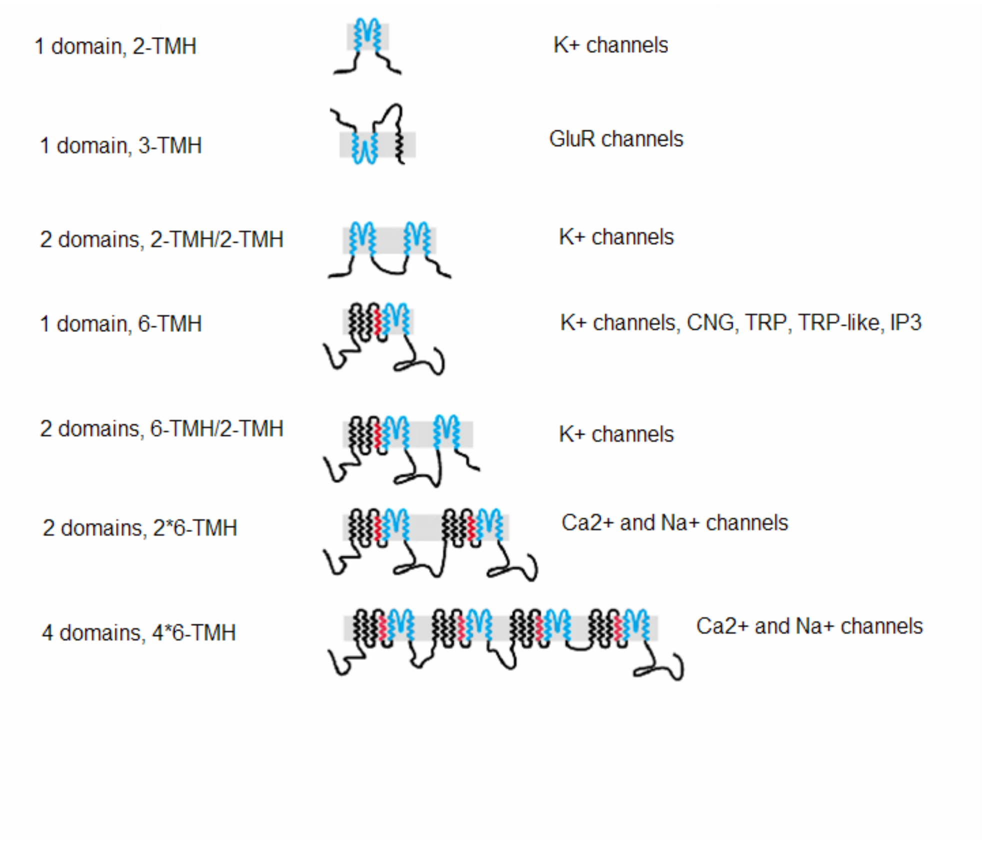
The modeling procedure aimed to identify the precise location of domain boundaries, TMHs and P-region in sequence. Combination of similarity searches and consensus results of several TMH prediction methods (SPLIT4.0, HMMTOP2, TMHMM 2.0, PHDhtm, MEMSAT and TMAP) served for assignation.

Perl script was written to execute batch processing on selected dataset with locally installed BLAST and set of representative P-region sequences as queries. E-value statistics was utilized in evaluation.

Data from publicly available database records and modeling procedures were fused and stored in MySQL database: **PPore**.

Perl CGI scripts were developed to query the database interactively

Total of 1539 protein/fragment entries related to P-region superfamily were isolated from protein databases, with tendency of this number to change on a monthly base.

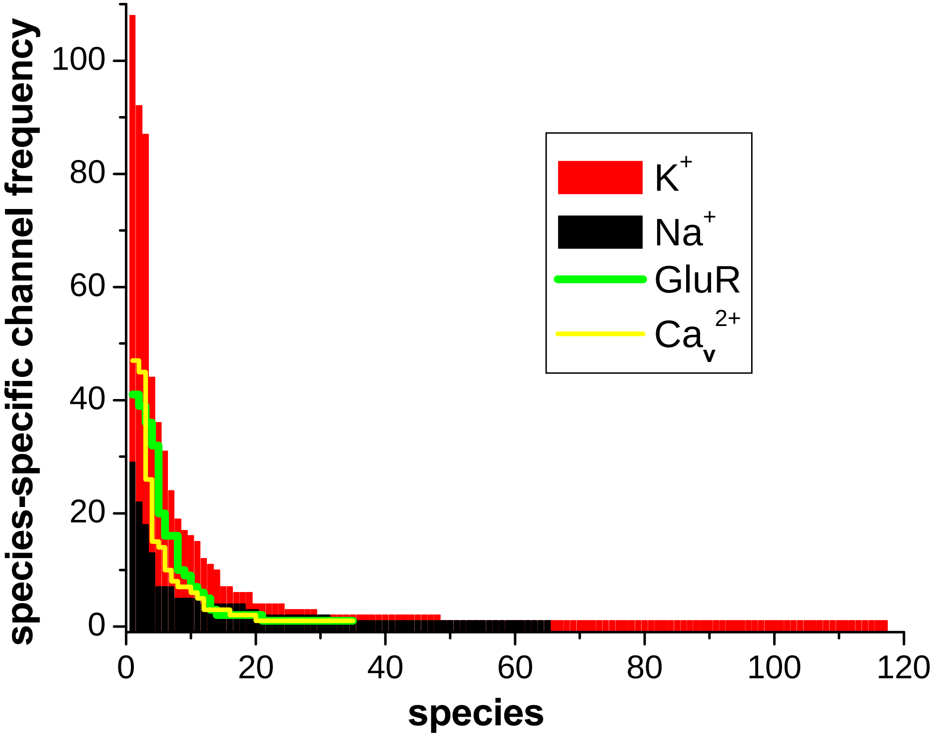


Identified set of functional families with membrane topologies

Traced proteins belonged to several functional families: Na$^+$, K$^+$ and Ca^{2+} specifec channels (Ca^{2+} with Ca^{2+} voltage-gated channels, ryanodine receptors (RYR) and IP3 family) and families with broader cationic specificity: glutamate receptors (GluR), cyclic-nucleotide gated (CNG), transient receptor potential (Trp) and Trp-like channels.

Observed channel diversity in databases
Functional divergency of ion currents in and among the species

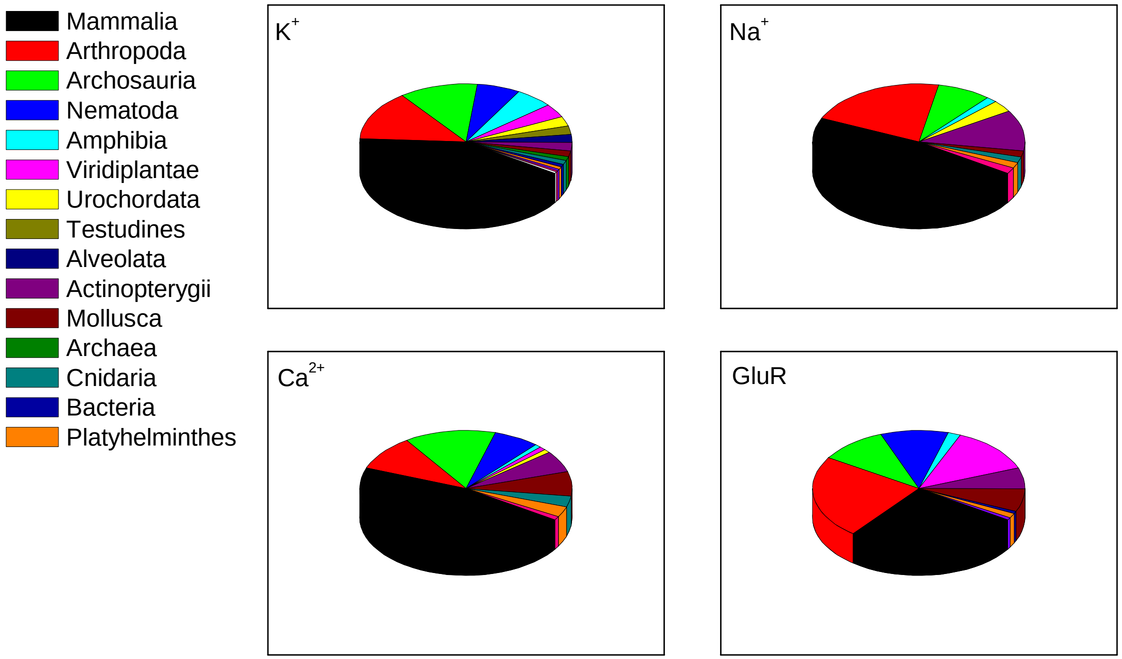
K$^+$ channels exhibit the largest iner and intra species diversity. Na$^+$ channels, with smallest intra-species divergency are also found in wide range of species, result of involvment in functions other then action-potential generation.



	total # channels	# species
K$^+$	**696**	**117**
GluR	**269**	**37**
Ca^{2+}	**226**	**38**
Na$^+$	**213**	**65**

Current functional family distribution in SWISS-PROT/TREMBL databases: For each family, species were sorted on X-axis in decreased order by frequency of isolated channels (Y-axis). Maximal value on X represents total number of species from which channels were isolated. GluR and Ca $^{2+}$ histograms are plotted as horizontal step lines.

Distribution over taxonomical classes – new functions to be found



Currently known distribution of channels in wide range of taxonomical classes, represented for different functional families. Each class is described by frequency of channels isolated from the most frequently used organism model, to minimize the effect of the extent of experimental work done on particular species.

Similar pattern is shown in distribution, indicating the observed channel frequency is result of class organism complexity (*Mammalia*, *Archosauria*), as well as usage of favorite organism models in ion channel research (*Arthropoda*, *Nematoda*). Higher abundance of channels in lower-complexity organisms (Arthropoda, Nematoda) can be related to experimental bias, or caused by presence of channels with function not related to neuronal-system (*Viridiplantae* (GluR), *Arthropoda* (Na^+), *Bacteria*, *Arhea*).

ARE we mislead?

When compared with other families, K^+ channels are prevalent. That is partly due to the least demanding processes of identification and functional expression of new K^+ channels.

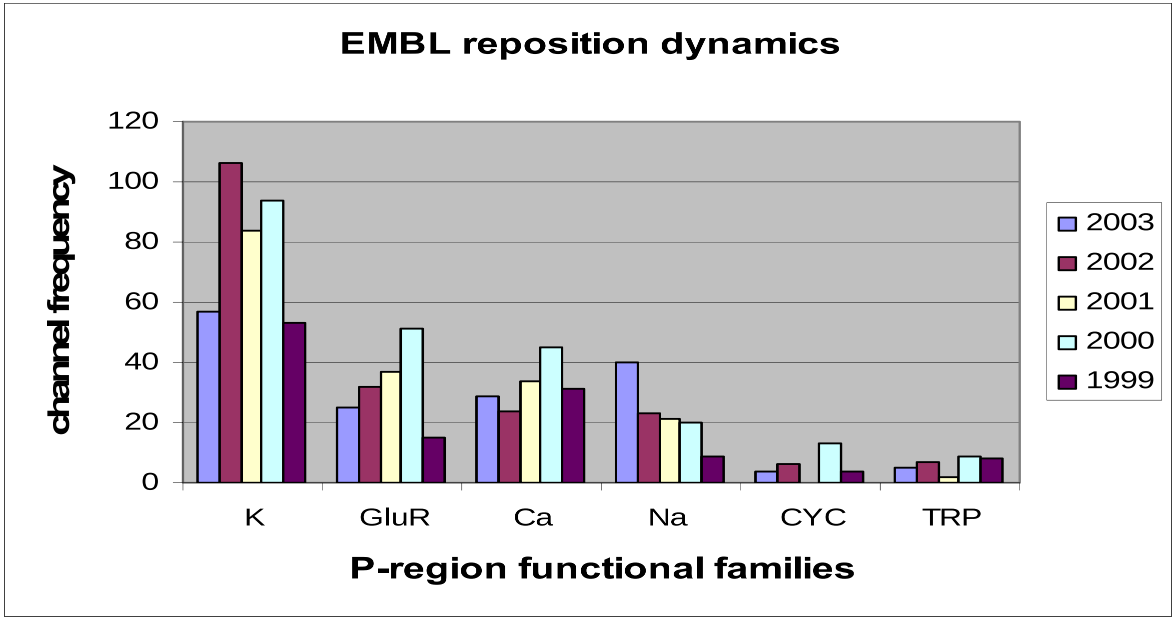
To probe the size of unidentified gene pools for families other then K^+, we investigated:

A) trends in reposition of new sequences to EMBL database during last 5 years - indicating relative (en/de)crease in number of isolated channels;

B) annotation-related rate of transferring newly identified channel entries from TREMBL to SWISS-PROT databases;

C) averaged divergence in MSA alignment positions from P-region, the most conserved fragment in these channels.

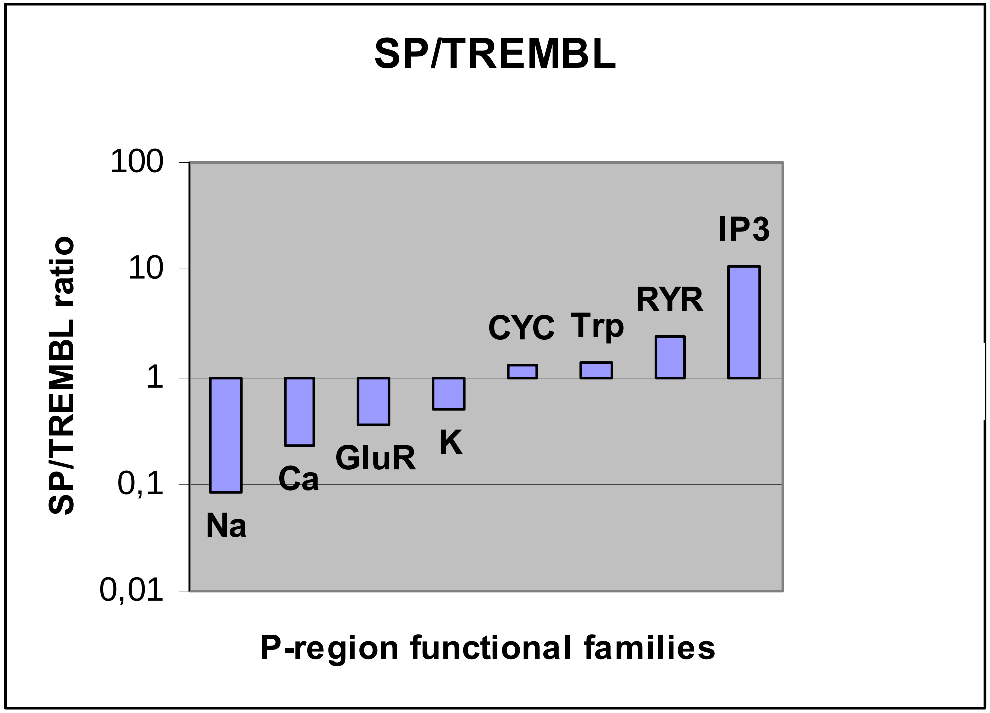
A) Is saturation in EMBL deposition reached?



Trends in EMBL reposition of new channel DNA sequences during last 5 years, shown for different families. Sharp changes are noted for K^+ (decrease for 44%) and Na^+ (increase for 45%) in 2003 year.

Sharp changes are noted for K^+ (decrease for 44%) and Na^+ (increase for 45%) in 2003 year. These tendencies show the great possibility of finding new Na^+ channels, and possible saturation in identification of new K^+ channels. GluR channels also constantly decrease since 2000.

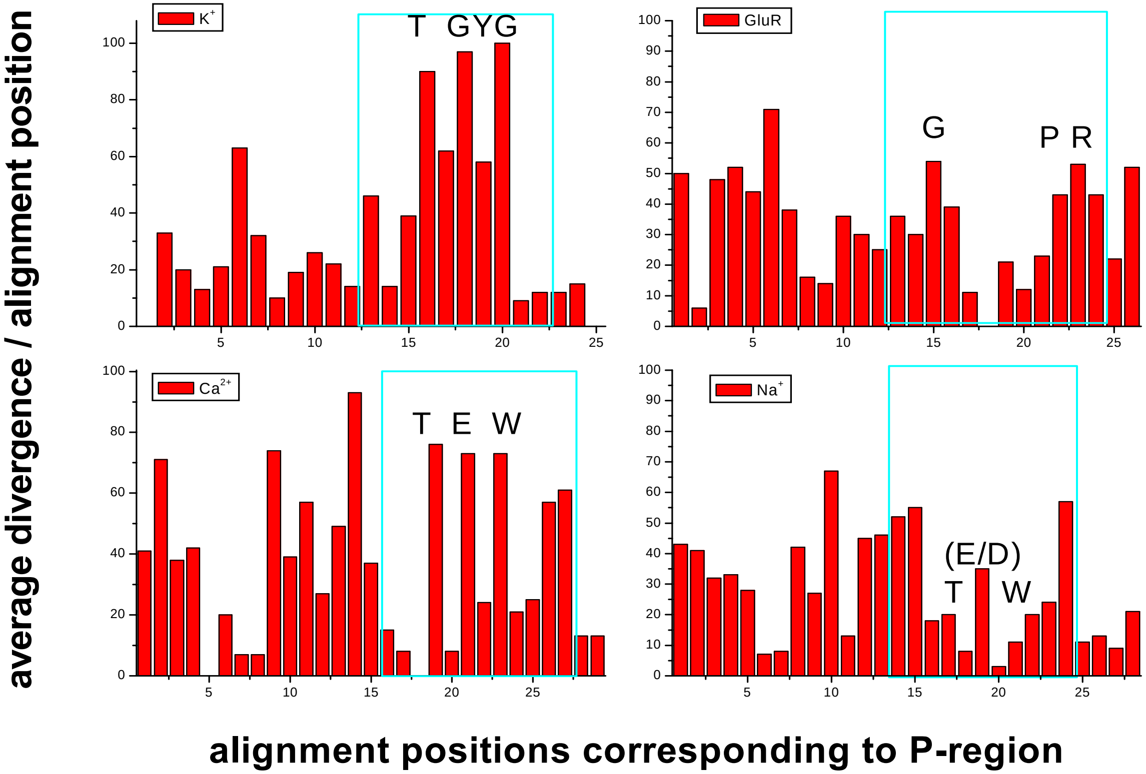
B) Is functional annotation troublesome?



Ratio of SWISS-PROT/TREMBL entries– is influenced by relative ease by which experts judge functional annotation of the TREMBL entry.

Channels with ratio <1 are prevalent in TREMBL and therefore more likely to have divergent sequences or to lack relevant experimental data.

C) How divergent the channel markers are?



Similarity search for new channels relays on conservation of the P-region. The most conserved part of this “channel marker” is selectivity filter region (cyan rectangle) where common motif reside. We showed average divergence in MSA alignment positions from P-region for different channel families.

The most abundant K^+ channels have the most conserved marker (P-region), with highest peaks in conserved amino acid positions of motif TXGYG.