LiGIoNs: A Computational Method for the Detection and Classification of Ligand-Gated Ion Channels

Avgi E. Apostolakou[#1], Katerina C. Nastou[#1,2], Georgios N. Petichakis[#1], Zoi I. Litou[1] and Vassiliki A. Iconomidou[1]*

[1]Section of Cell Biology and Biophysics, Department of Biology, National and Kapodistrian University of Athens, Panepistimiopolis, Athens 15701, Greece

[2]Present address: Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

[#]Equally contributing first authors

*To whom correspondence should be addressed

Associate Prof. Vassiliki A. Iconomidou

Section of Cell Biology and Biophysics, Department of Biology,

National and Kapodistrian University of Athens, Panepistimiopolis,

Athens 15701, Greece

Phone: +30 210 727 4871

Fax: +30 210 727-4254

e-mail: veconom@biol.uoa.gr

http://biophysics.biol.uoa.gr

ABSTRACT

Ligand-Gated Ion Channels (LGICs) are one of the largest groups of transmembrane proteins. Due to their major role in synaptic transmission, both in the nervous system and the somatic neuromuscular junction, LGICs present attractive therapeutic targets. During the last few years several computational methods for the detection of LGICs have been developed. These methods are based on machine learning approaches utilizing features extracted solely from amino acid composition. Here we report the development of LiGIoNs, a profile Hidden Markov Model (pHMM) method for the prediction and ligand-based classification of LGICs. The method consists of a library of 10 pHMMs, one per LGIC subfamily, built from the alignment of representative LGIC sequences. In addition, 14 Pfam pHMMs are used to further annotate and classify unknown protein sequences into one of the 10 LGIC subfamilies. Evaluation of the method showed that it outperforms existent methods in the detection of LGICs. On top of that, LiGIoNs is the only currently available method that classifies LGICs into subfamilies.

The method is available online at http://bioinformatics.biol.uoa.gr/ligions/.

KEYWORDS

membrane, prediction, ligand-gated ion channels, profile Hidden Markov Models

ABBREVIATIONS

LGIC(s): Ligand-Gated Ion Channel(s)

VGIC(s): Voltage-Gated Ion Channel(s)

pHMM(s): profile Hidden Markov Models

TM: Transmembrane

MSA: Multiple Sequence Alignment

MCC: Matthew's Correlation Coefficient

1. INTRODUCTION

Most living cells exhibit a membrane potential due to the membrane acting as a barrier for charged particles [1]. To overcome this obstacle, cells use specialized transmembrane proteins – known as ion channels [2] – that control the flow of ions in and out of the cell. These proteins are highly selective and can discriminate both between anions and cations as well as between monovalent and divalent ions [3]. Their gating is typically a result of either change to the membrane potential or binding of specific ligands. Thus, channels are typically classified in two main classes, according to their gating trigger, Voltage-Gated Ion Channels (VGICs) and Ligand-Gated Ion Channels (LGICs) [4]. Both classes are extremely diverse and composed of numerous members further classified into various families and subfamilies.

Most LGICs consist of multiple subunits forming heteropolymers; these subunits are encoded by many genes. The variety of possible subunit combinations within each subfamily of LGICs leads to a wide range of receptors with different pharmacological and biophysical properties and diverse expression patterns both within the nervous system and in other tissues [5]. Thus, LGICs have emerged as complex yet attractive targets for the development of new therapeutic agents [6]. LGICs comprise the anion-selective $GABA_A$ receptors [7] and glycine receptors [8], and the cation-selective nicotinic acetylcholine receptors [9,10], 5-HT_3 receptors [11], ionotropic glutamate receptors [12], IP_3 receptors [13], P2X receptors [14], epithelial sodium channels [15], acid-sensing (proton-gated) ion channels [16] and zinc-activated channels [17]. The nicotinic acetylcholine, 5-HT_3, $GABA_A$ and glycine receptors (and an additional member, the zinc-activated channels) form the family of Cys-loop receptors [18]. The distinct structural characteristics of each LGIC subfamily are shown in **Table 1**.

Table 1. Characteristics of the 10 LGIC subfamilies. The number of subunits, transmembrane segments per subunit and average length of each subunit in amino acid residues is shown for each subfamily. Members of the Cys-loop family are marked with *italics* and <u>underline</u> in the table.

LGIC subfamily	***Subunit Count***	***Transmembrane (TM) segments per subunit***	***Subunit Length***
Epithelial Sodium Channels (ENaCs)	3	2	~650
P2X Receptors	3	2	~435
Acid-sensing Ion Channels (ASICs)	3	2	~550
Ionotropic Glutamate Receptors	4	3	~950

LGIC subfamily	*Subunit Count*	*Transmembrane (TM) segments per subunit*	*Subunit Length*
Nicotinic Acetylcholine Receptors	5	4	~500
5-HT_3 Receptors	5	4	~460
$GABA_A$ Receptors	5	4	~480
Glycine Receptors	5	4	~470
Zinc-activated Channels (ZACs)	5	4	~410
IP_3 Receptors	4	6	~2750

Considering the importance of ion channels for normal cellular function and their role as drug targets [19,20], several prediction methods have been developed for the identification of ion channels based on their amino acid sequence [21–29]. In a recent survey [30] ion channel predictors were compared and evaluated and only two of those, IonChanPred 2.0 [27] and PSIONplus [26], were both available for use and capable of classifying ion channels into LGICs. IonChanPred 2.0 is built on an SVM-based model that employs pseudo-dipeptide composition and physicochemical properties. PSION-plus combines SVM and PSI-BLAST and relies on multiple features; it performed better than IonChanPred 2.0, but still modestly so. Both methods were designed to work in parallel in order to detect ion channels, then classify them into LGICs or VGICs and in the case of VGICs to further classify them into four subtypes according to ion selectivity. Since the publication of the survey a new predictor by Han *et al.* [28] came out, that is however not available, and PSIONplus^m [29], built on PSIONplus, that addressed some issues pointed out by the survey, including now classifying LGICs into four subtypes, same as VGICs. However, there is no available method that classifies LGICs into subfamilies.

In order to fill this gap, we decided to design and develop LiGIoNs, a sequence-based predictor, that identifies LGICs in proteomes with the use of profile Hidden Markov Models (pHMMs). Each pHMM describes a different LGIC subfamily and can be used to detect novel members. As such, LiGIoNs is the first method that in addition to detecting LGICs, performs classification of LGICs into the 10 known ligand-based subfamilies (**Table 1**). We have also developed a web server to host the method, available at http://bioinformatics.biol.uoa.gr/ligions.

2. METHODS

The LiGIoNs method consists of two parts: the prediction part (detection and classification) and the annotation part. For the prediction part of the method, a dataset of LGICs was collected for each subfamily following the IUPHAR classification scheme [31] already presented in **Table 1**. For the training set, proteins were originally collected from the IUPHAR database [31]. However, since IUPHAR contains data only for human, mouse and rat proteins, the exclusive use of this source would greatly limit the diversity of the training dataset. For this reason, we isolated all UniProt [32] entries with evidence at protein or transcript level (an exception was made for ZACs) that belong to an LGIC subfamily (UniProt queries found in **Supplementary Table 1**). These proteins were cross-checked with the IUPHAR database and only two proteins in IUPHAR were excluded due to their existence evidence in UniProt. The final dataset of protein sequences used for the LiGIoNs training dataset is shown in **Supplementary Table 1**. The CD-HIT clustering algorithm [33,34] was used to reduce sequence similarity to 40% within each subfamily. The remaining members in each subfamily were aligned using ClustalO [35], resulting in 10 multiple sequence alignments (MSA). HMMER 3.3.2 [36] was then used to construct the 10 respective pHMMs corresponding to the MSA of each LGIC subfamily, which all together constitute the LGICslib pHMM library.

The following algorithm is used to characterize an unknown protein sequence as an LGIC. Initially, the unknown protein sequence is scanned against LGICslib. This is followed by recording the number of pHMMs that align with the unknown protein sequence. In order to characterize an unknown protein sequence as a member of a specific LGIC subfamily the alignment score must be higher than the threshold that is set for that pHMM. The thresholds for each pHMM were set manually in order to maximize sensitivity and specificity, based on the protocol introduced by Ioannidou *et al*. [37].

For the annotation part of LiGIoNs, all characteristic pHMMs for the LGICs were identified and isolated from the Pfam protein family database [38]. This procedure preceded the construction of the LGICslib pHMMs, in an initial attempt to identify existing pHMMs that could be used to uniquely describe LGIC subfamilies. However,

this was not possible using data extracted exclusively from Pfam, as there is no combination of pHMMs deposited in the database that allows the successful classification of LGICs into subfamilies. Nevertheless, pHMMs from Pfam (**Table 2**) in combination with those in LGICslib, allowed both an additional validation of the results obtained using LiGIoNs in proteomes and the creation of the annotation part of the method.

The pHMMs that are characteristic of LGICs were collected from the Pfam cross-references provided in all UniProt entries of the LGICs training set (**Supplementary Table 1**). As a result, 14 pHMMs were isolated from Pfam, which are presented in **Table 2** and were used to create a second library, named PfamLGICslib. HMMER is used again to scan sequences – previously characterized as LGICs by LiGIoNs – against PfamLGICslib, and sequences that have a "hit" from pHMMs in the Pfam library are annotated with these domains.

Table 2. Correlation between pHMMs deposited in Pfam and LGIC subfamilies.

LGIC subfamily	***pHMMs from Pfam***
Epithelial sodium channels (ENaCs) Acid-sensing ion channels (ASICs)	PF00858 (ASC)
P2X Receptors	PF00864 (P2X_receptor)
Ionotropic Glutamate Receptors	PF01094 (ANF_receptor) PF10562 (CaM_bdg_C0) PF10613 (Lig_chan-Glu_bd) PF00060 (Lig_chan) PF10565 (NMDAR2_C)
Nicotinic Acetylcholine Receptors 5-HT_3 Receptors $GABA_A$ Receptors Glycine Receptors Zinc-activated channels (ZACs)	PF02931 (Neur_chan_LBD) PF02932 (Neur_chan_memb)
IP_3 Receptors	PF01365 (RYDR_ITPR) PF08454 (RIH_assoc) PF02815 (MIR) PF08709 (Ins145_P3_rec) PF00520 (Ion_trans)

A jackknife cross-validation experiment was conducted to assess the performance of LiGIoNs, by evaluating the performance of pHMMs in correctly classifying LGICs

belonging to different subfamilies. For each LGIC subfamily, one sequence was removed from the multiple sequence alignment (MSA) of the pHMM's seed set, and a new pHMM was constructed from the remaining sequences of the new MSA. Then we measured the newly created pHMM's ability to correctly classify the removed sequence and the sequences from three negative datasets. The three negative datasets consisted of (1), a set of all LGICs of the other subfamilies, (2) a set of non-LGIC transmembrane proteins and (3) a set of globular proteins. The set of all LGICs of other families includes the protein members of each subfamily after the sequence similarity reduction. The negative test set of non-LGIC transmembrane proteins was isolated by searching UniProt/SwissProt for transmembrane proteins with many transmembrane segments (subcellular location: "Multi-pass membrane protein") not containing the keyword "Ligand Gated Ion Channel" in the entry's text file. This search returned 52581 entries. An additional dataset of globular proteins was collected from UniProt/SwissProt (excluding subcellular location: "Multi-pass membrane protein" or "Single-pass membrane protein") of which 1000 were selected at random. Both datasets were subjected to sequence similarity reduction using CD-HIT to below 40% similarity resulting in a representative set of 1508 transmembrane proteins and a set of 953 globular proteins (**Supplementary Table 2**).

In addition to the above evaluation, the method was compared to the IonChanPred 2.0 method to test its overall performance. The negative dataset used by IonChanPred 2.0 [27] for performance evaluation, comprised of 300 non-LGIC transmembrane proteins, was used to compare the two methods, in addition to the aforementioned datasets. A comparison to the PSIONPlus^{m} method [29] was not feasible, as the online tool is prohibitively slow. It should also be mentioned that none of the above methods are capable of classifying LGICs into subfamilies.

True/false positives (TP, FP) and true/false negatives (TN, FN) were counted on a per protein basis. Specifically, for the jackknife cross-validation a protein in the negative dataset was counted as a FP if it is incorrectly classified as LGIC in at least one iteration and as a TN if it is always correctly classified as non-LGIC. For the prediction performance of LiGIoNs three measures were used, namely Recall, Precision and F-score. These measures are best suited for binary classification and are not affected by the number of TN [39], necessary for the imbalanced datasets used in this study.

Recall, or Sensitivity is:

$$Recall = \frac{TP}{(TP + FN)} \text{ (2.1)},$$

Precision is:

$$Precision = \frac{TP}{(TP + FP)} \text{ (2.2)},$$

and F-score is:

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \text{ (2.3)}.$$

Moreover, LiGIoNs was applied to all reference eukaryotic, bacterial and archaean proteomes (**Supplementary Table 3**) retrieved from UniProt **(**release: 2021_02**)** in order to further assess the method's ability to detect LGICs in proteomes. An additional measure was used for the proteome analysis, the False Positive Rate (FPR), to assess the number of proteins expected to be predicted as LGICs by chance given the size of each respective proteome.

FPR is calculated as:

$$FPR = \frac{FP}{FP + TN} \text{ (2.4)}.$$

3. RESULTS AND DISCUSSION

3.1. LiGIoNs algorithm

The prediction part of LiGIoNs consists of a library of 10 pHMMs (LGICslib), created from the MSAs of the proteins belonging to the 10 LGIC subfamilies (Table 3), after sequence similarity reduction. For a protein to be characterized as a subunit of a specific LGIC subfamily, the corresponding pHMMs in LGICslib must score higher than the profile's threshold. If multiple subfamilies meet the aforementioned condition, the one with the highest overall score is chosen to characterize the unknown sequence, but all hits are made available to the user for further consideration. For the annotation level, a library of 14 pHMMs containing characteristic LGIC domains recorded in Pfam was created (PfamLGICslib, **Table 2**). This library is scanned in positive cases only – i.e. an unknown sequence is characterized as an LGIC in the previous step – in order to provide the user with more information regarding the protein being studied. The flowchart in **Figure 1** depicts in detail how the LiGIoNs method works.

Table 3. Dataset of LGICs per subfamily. LGICs were collected from UniProt and the training datasets were created after a reduction of sequence similarity at 40%. (More details in **Supplementary Table 1**)

LGIC subfamily	*Number of entries found in UniProt*	*Number of entries used for pHMM training*
Epithelial Sodium Channels (ENaCs)	57	11
P2X Receptors	175	12
Acid-sensing Ion Channels (ASICs)	54	12
Ionotropic Glutamate Receptors	535	67
Nicotinic Acetylcholine Receptors	799	97
5-HT_3 Receptors	19	5
$GABA_A$ Receptors	236	18
Glycine Receptors	74	17
Zinc-activated Channels (ZACs)	52	6
IP_3 Receptors	66	8

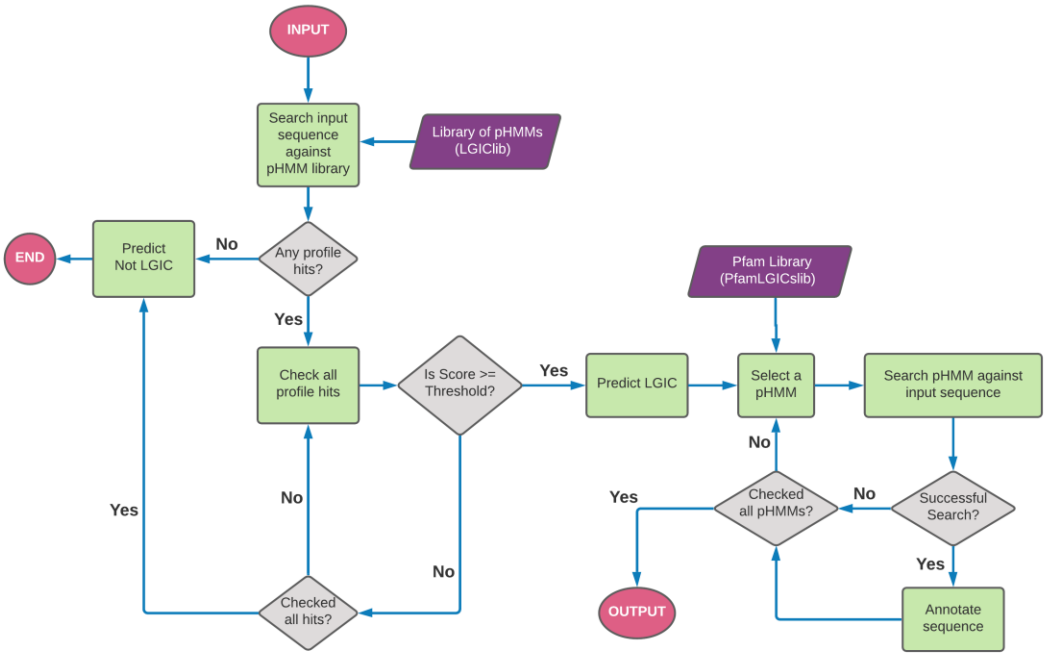
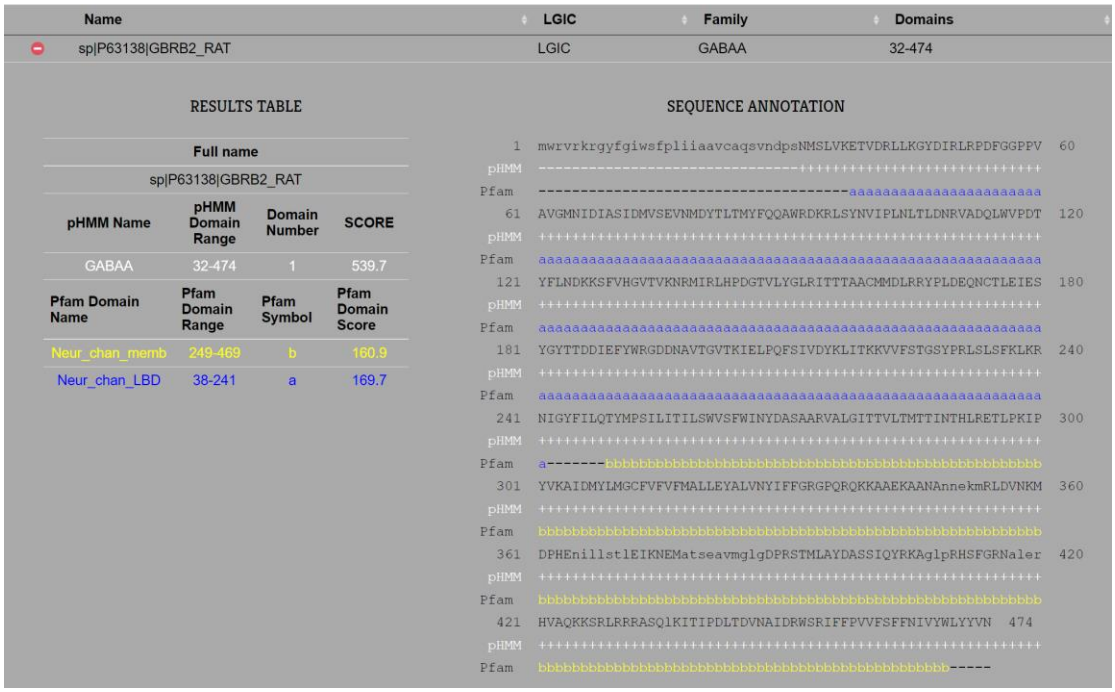


Figure 1. Flowchart of the LiGIoNs algorithm. A FASTA formatted sequence is used as input. The input sequence is used to search against the pHMM library LGICslib. If the search is successful, the profile hits are checked for the highest scoring hit that has a score above the corresponding profile threshold. If these conditions are not met, then the sequence is characterized as a non-LGIC and the program exits. On the other hand, if the conditions are met the sequence is characterized as LGIC and is then scanned against the library of Pfam profiles (PfamLGICslib). After all profiles have been searched against the sequence, the annotation procedure is over, and the results are presented to the user.

3.2. User interface and website features

A web interface has been created for LiGIoNs and the method is publicly available through http://bioinformatics.biol.uoa.gr/ligions. Via the "Home" page, the user can access the query submission page ("Run") and the "Manual" page. Query submission can be performed using a single or a set of protein sequences in the textbox provided, or by uploading a file with FASTA formatted sequences (up to 3000 sequences). After a successful query submission, users are transferred to the results page where they can gather information about their submission and download a file with the results. In the same page users can download the results in a text file and perform a new query.

At the "Results" page, all results are presented in a table format, where each line corresponds to a single protein and each column contains the basic characteristics of the proteins and the prediction of LiGIoNs (LGIC or no LGIC). More details about each protein are provided, as shown in **Figure 2**.



Name	LGIC	Family	Domains
sp|P63138|GBRB2_RAT	LGIC	GABAA	32-474

RESULTS TABLE

Full name
sp|P63138|GBRB2_RAT

pHMM Name	pHMM Domain Range	Domain Number	SCORE
GABAA	32-474	1	539.7

Pfam Domain Name	Pfam Domain Range	Pfam Symbol	Pfam Domain Score
Neur_chan_memb	249-469	b	160.9
Neur_chan_LBD	38-241	a	169.7

SEQUENCE ANNOTATION

```
  1 mwrvrkrgyfgiwsfpliiaavcaqsvndpsNMSLVKETVDRLLKGYDIRLRPDFGGPPV  60
pHMM ------------------------------++++++++++++++++++++++++++++
Pfam ------------------------------------aaaaaaaaaaaaaaaaaaaaaa
 61 AVGMNIDIASIDMVSEVNMDYTLTMYFQQAWRDKRLSYNVIPLNLTLDNRVADQLWVPDT 120
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
121 YFLNDKKSFVHGVTVKNRMIRLHPDGTVLYGLRITTTAACMMDLRRYPLDEQNCTLEIES 180
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
181 YGYTTDDIEFYWRGDDNAVTGVTKIELPQFSIVDYKLITKKVVFSTGSYPRLSLSFKLKR 240
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
241 NIGYFILQTYMPSILITILSWVSFWINYDASAARVALGITTVLTMTTINTHLRETLPKIP 300
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam a-------bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
301 YVKAIDMYLMGCFVFVFMALLEYALVNYIFFGRGPQRQKKAAEKAANAnnekmRLDVNKM 360
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
361 DPHEnillstlEIKNEMatseavmglgDPRSTMLAYDASSIQYRKAglpRHSFGRNaler 420
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
421 HVAQKKSRLRRRASQlKITIPDLTDVNAIDRWSRIFFPVVFSFFNIVYWLYYVN  474
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb-----
```

Figure 2. Webpage results after a successful run of LiGIoNs against an unknown protein sequence. The sequence of the $GABA_A$ receptor from *Rattus norvegicus* (Rat) (UniProtAC: P63138) is used in this example. The pHMM of LGICslib that corresponds to $GABA_A$ receptors, as well as two pHMMs from PfamLGICslib, characteristic of LGICs that belong to the Cys-loop family (Table 2) are detected in the sequence. The first line under the sequence displays the positions where the input sequence and the $GABA_A$ pHMM are aligned (+ symbol). The second line under the sequence shows the positions where the different Pfam domains have been detected in the sequence, using a different color and character for each one of them. Positions where multiple domains have been detected, are marked with asterisks (*).

The results text file contains a protein identifier, the protein subfamily that the protein belongs to – if it is a positive hit, as well as other subfamily hits and their scores, the position and score of each domain, Pfam domain(s) present in the protein, the protein sequence and the alignment annotation from HMMER. Users are provided with a JobID for each submission, which can be used for up to two weeks to retrieve results after a prediction has been performed. LiGIoNs is fast, since for a query length the size of the human proteome the method produces results in approximately two hours, which makes it well-suited for proteomic scale applications. Lastly, in addition to the web interface all scripts and associated files needed to run LiGIoNs or recreate the pHMMs are available here https://github.com/DawnBio/LiGIoNs.

3.3. Method Evaluation

As mentioned above, LiGIoNs was evaluated using jackknife cross-validation. The workflow of the evaluation procedure is shown in **Figure 3**. The method's ability to correctly identify the sequences is assessed each time and the procedure is repeated until all entries belonging to all subfamilies have been checked. The overall performance statistics of LiGIoNs are calculated afterwards.

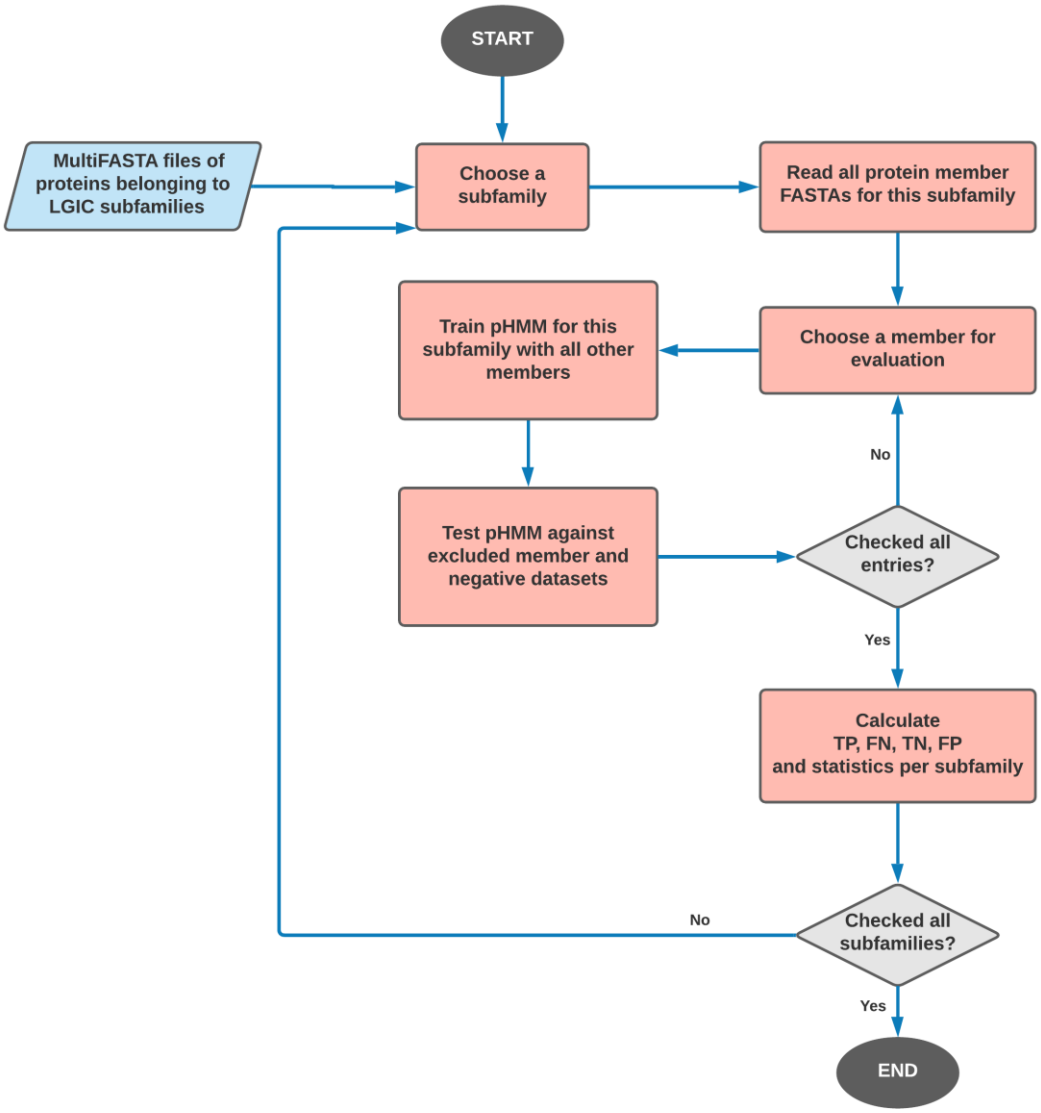


Figure 3. The workflow of the jackknife cross-validation experiment to assess the performance of LiGIoNs. Initially a subfamily of LGICs is chosen and all protein member FASTAs used for the training of the specific subfamily are selected. One protein sequence is used each time, which is left out, and the rest are used to create an MSA, train a pHMM and scan it against this one sequence, as well as, against the three negative datasets (**Supplementary Table 2**).

LiGIoNs performed very well during cross-validation, with high overall performance metrics against the globular and transmembrane negative test datasets (**Table 3** and **Table 4**). The results from the jackknife test showed the method's ability to correctly identify pseudo-novel sequences as LGICs (Recall over 87% in most cases), except for the two smallest subfamilies (5-HT_3 Receptors and ZACs). Additionally, no globular proteins and very few non-LGIC transmembrane proteins were erroneously classified as LGICs, with Precision being 100% and over 78% for each negative dataset respectively. Overall, the performance of LiGIoNs against the transmembrane negative dataset was good as the F-score was over 80% for most families, again with the two smallest subfamilies having a lower score but still over 72%. These values are indicative of the fact that pHMMs have a good ability to discriminate LGICs and non-LGICs, however they appear to underperform when classifying LGICs into the right subfamily (**Table 5**). For that reason, the method was also executed normally following the workflow described in **Section 3.1**, and proteins that have hits against multiple subfamilies were classified based on the highest overall score (see **Figure 1**). In that more realistic scenario, it became evident that LiGIoNs correctly classifies each known LGIC in the appropriate subfamily, as only a single LGIC classified in the wrong subfamily.

Table 3. Results from the cross-validation of LiGIoNs using the jackknife technique against the 953 globular proteins negative dataset.

LGIC subfamily	*Recall (%)*	*Precision (%)*	*F-score (%)*
Epithelial Sodium Channels (ENaCs)	100.00	100.00	100.00
P2X Receptors	91.67	100.00	95.65
Acid-sensing Ion Channels (ASICs)	100.00	100.00	100.00
Ionotropic Glutamate Receptors	97.01	100.00	98.48
Nicotinic Acetylcholine Receptors	98.97	100.00	99.48
5-HT_3 Receptors	60.00	100.00	75.00
$GABA_A$ Receptors	88.89	100.00	94.12
Glycine Receptors	88.24	100.00	93.75
Zinc-activated Channels (ZACs)	66.67	100.00	80.00
IP3 Receptors	87.50	100.00	93.33

Table 4. Results from the cross-validation of LiGIoNs using the jackknife technique against the 1508 transmembrane proteins of the negative dataset.

LGIC subfamily	*Recall (%)*	*Precision (%)*	*F-score (%)*
Epithelial Sodium Channels (ENaCs)	100.00	78.57	88.00
P2X Receptors	91.67	100.00	95.65
Acid-sensing Ion Channels (ASICs)	100.00	80.00	88.89

Ionotropic Glutamate Receptors	97.01	100.00	98.48
Nicotinic Acetylcholine Receptors	98.97	98.97	98.97
5-HT_3 Receptors	60.00	100.00	75.00
$GABA_A$ Receptors	88.89	94.12	91.43
Glycine Receptors	88.24	100.00	93.75
Zinc-activated Channels (ZACs)	66.67	80.00	72.73
IP3 Receptors	87.50	100.00	93.33

Table 5. Results from the cross-validation of LiGIoNs using the jackknife technique against LGICs belonging to other subfamilies than the one validated.

LGIC subfamily	*Recall (%)*	*Precision (%)*	*F-score (%)*
Epithelial Sodium Channels (ENaCs)	100.00	50.00	66.67
P2X Receptors	91.67	100.00	95.65
Acid-sensing Ion Channels (ASICs)	100.00	52.17	68.57
Ionotropic Glutamate Receptors	97.01	100.00	98.48
Nicotinic Acetylcholine Receptors	98.97	70.07	82.05
5-HT_3 Receptors	60.00	5.17	9.52
$GABA_A$ Receptors	88.89	30.77	45.71
Glycine Receptors	88.24	45.45	60.00
Zinc-activated Channels (ZACs)	66.67	15.38	25.00
IP_3 Receptors	87.50	100.00	93.33

Even though most results presented above appear excellent, the pHMMs for 5-HT_3 Receptors and ZACs perform moderately because the datasets for those LGIC subfamilies are extremely small (**Supplementary Table 1**). An additional issue is that these subfamilies belong to the Cys-loop family of LGICs, so the ability to distinguish them from other members of the same family, primarily Nicotinic Acetylcholine Receptors, becomes inherently difficult (**Table 5**). The method's ability to identify and characterize more proteins belonging to these subfamilies can be improved in the future, if LGICs belonging to proteomes that are evolutionary distant to mammals are annotated as such and are subsequently used during the creation of pHMMs. This is discussed further in Section 3.4.

LiGIoNs was also compared with IonChanPred 2.0 **[27]**. Based on the results presented in **Table 6**, it is obvious that our method outperforms IonChanPred 2.0 in the detection of LGICs. The positive dataset that was used to compare the two methods is the same as that presented in the original IonChanPred 2.0 publication **[27]**. For the negative set the original negative dataset presented in IonChanPred 2.0 was added to the two negative datasets presented earlier in this work (globular proteins and transmembrane proteins other than LGICs). It should be noted that the two methods have

different abilities, since, while IonChanPred 2.0 can detect LGICs, it lacks the ability to classify them into subfamilies. For this reason, the two methods are only compared for their ability to predict if a protein is an LGIC or not.

Table 6. Comparison of LiGIoNs with IonChanPred 2.0 in their ability to detect LGICs from their amino acid sequence.

Method	*Recall (%)*	*Precision (%)*	*F-score (%)*
LiGIoNs	93.33	**97.22**	**95.24**
IonChanPred 2.0	**100.00**	40.43	57.58

This comparison shows that IonChanPred 2.0 has a recall of 100%, detecting all LGICs in the test set, but at the same time has very low precision, indicating that the high recall of the method is probably a result of overfitting to the positive set. In contrast, LiGIoNs can most often correctly identify LGICs, while not falsely detecting non-LGICs as LGICs. Overall, this demonstrates that LiGIoNs is a more balanced method, that avoids the issue of overfitting.

3.4. Application in reference proteomes

LiGIoNs was applied in all reference proteomes in UniProt for the detection of possible LGICs (**Supplementary Table 3**). Across the three superkingdoms, at least one potential LGIC was detected in almost half Eukaryotic proteomes, 304 Bacterial proteomes and only a single Archaeal proteome (**Table 7**). The FPR was calculated based on the transmembrane negative dataset and is 0.265%, which is higher than the percentage of LGICs found in Bacteria and Archaea, and with only 438 Eukaryotic proteomes surpassing this number. However, it should be noted that a) the true value of FPR is likely lower as transmembrane proteins are 30% or less of a whole proteome [40] and the FPR for globular proteins is 0%, and b) the threshold values for the pHMM are relatively high, possibly excluding more distant not yet annotated members.

Table 7. Results of the application of LiGIoNs in the reference proteomes of each superkingdom.

Superkingdom	*Number of reference proteomes*	*Proteomes with at least one potential LGIC*	*Average number of LGICs*	*Average percentage of LGICs*	*Highest percentage of LGICs*
Eukaryota	1586	726	87.23*	0.33*	1.09*
Bacteria	7999	304	1.05	0.03	0.08
Archaea	330	1	1	0.04	0.04

*excluding incomplete proteomes UP000265894 and UP000297026 with less than 500 proteins

Overall, the percentage of LGICs in Eukaryota is 0.33, but this number can vary significantly based on the kingdom and phylum. While LGICs are detected in all kingdoms and phyla, only some proteomes of Metazoan phyla have a higher percentage of LGICs than the calculated FPR. The results also showed that each LGIC subfamily varies in representation across the taxa, with Chordata being the only taxon with representatives of all subfamilies. 5-HT_3 Receptors, important for serotonin neurotransmission, were found in Chordata and in Ecdysozoa, specifically in nematodes. ZACs, first discovered in humans and dogs [41], were found in Chordata including several mammals, bony fish and birds. Also, no ZACs were found in the rat and mouse proteomes, which is in accordance with the literature. Members of the subfamilies ASICs and ENaCs – both part of the ENaC/Degenerin family – were found in almost all Metazoan phyla, both vertebrate and invertebrate, as expected [42]. $GABA_A$ receptors similarly were primarily limited in Metazoa – no members of this family were detected in plants despite the presence of GABA regulated ion channels in plants [43]. Additionally, 57 $GABA_A$ receptors were predicted by LiGioNs in some bacterial proteomes. While this type of receptors has been found in bacteria [44], the percentage of predictions is less than the FPR. Next, it is known that plants have ionotropic glutamate receptors [45], and indeed most predicted LGICs in plants where assigned to the Ionotropic Glutamate Receptors subfamily. Glutamate receptors were also found in Metazoa and in several bacterial proteomes. Their presence in both eukaryotes and bacteria suggests that they emerged before the split from the common ancestor [46]. Another ancient subfamily is the IP3 Receptors that rose in eukaryotes before multicellularity according to previously published studies [47]. Possible LGICs belonging to this family were found in most eukaryotic kingdoms and phyla. Most predicted Nicotinic Acetylcholine Receptors belonged to Metazoa with fewer possible members found in other eukaryotic taxa and some bacteria. Even though it is thought that Acetylcholine has been synthesized since the beginning of life, receptors for this molecule have not been identified in any taxon apart from Metazoa [48]. Lastly, P2X Receptors were detected with LiGioNs, amongst other taxa, in Fungi [49] and Amoebozoa [50].

Approximately 61% of all detected LGICs are found in Chordata (253 proteomes) and all organisms with a percentage of LGICs higher than the FPR belonged to Metazoa. This is indicative of such organisms being better annotated and supports the

notion that LiGIoNs could be overfitted to this type of data. Available evidence supports the existence of LGICs in several other taxa, such as bacteria [51,52], archaea [52] and fungi [49], however these receptors are not as of yet well studied and annotated, and are therefore underrepresented in protein sequence databases. LiGIoNs can easily be updated once new training data become available in order to improve its performance on evolutionary distant organisms. However, there is still significance in the results obtained from the current version of LiGioNs for organisms that are evolutionary distant to Metazoa. Specifically, despite the percentage of predicted prokaryotic LGICs not surpassing the FPR, a closer look at these 318 proteins (**Supplementary Table 3)** shows that the vast majority are potential LGICs or transporters, a closely related category of transmembrane proteins [53]. Many of them are also assigned to an LGIC-related Pfam entry (**Table 2**). Therefore, a more detailed examination of the results showed that they can serve as a steppingstone for further experimental validation of the function of prokaryotic proteins and could shed light in the evolution of this protein class of valuable pharmacological targets.

4. CONCLUSIONS

LiGIoNs is a fast and accurate method, which can detect Ligand-Gated Ion Channels from sequence alone and is therefore applicable to entire proteomes. LiGIoNs is the only publicly available method that classifies LGICs into one of the ten known sub-families of these proteins. Moreover, LiGIoNs annotates predicted LGICs with information from Pfam, providing a full description of each sequence's characteristics to the method's users. LiGIoNs was applied to all UniProt reference proteomes. LGICs were detected in Eukaryota, Bacteria and Archaea, however the vast majority of receptors were found in Chordata.

In addition, the method we have developed is retrainable if more LGIC sequences become available in sequence databases. Retraining LiGIoNs when more LGICs are annotated, will allow us to increase the detection of LGICs in evolutionary distant proteomes than those used in the creation of the 10 core pHMMs. We plan to update the LGIClib pHMM library of LiGIoNs when new sequences are available, and this will allow us to build more descriptive profiles in the future and render the method more broadly applicable. Moreover, if new domains describing LGICs are added in Pfam, we plan to incorporate them in our method, as well. LiGIoNs is available at http://bioinformatics.biol.uoa.gr/ligions/.

ACKNOWLEDGEMENTS

The authors thank the National and Kapodistrian University of Athens for granting access to university premises and equipment.

Conflict of Interest: none declared.

Author Contributions

Study design: KCN, ZIL, VAI; Conceptualization: AEA, KCN, GNP, ZIL; Method design and development: AEA, GNP, KCN; Web Application Design and Development: AEA, GNP; Web Application Quality Assurance: AEA, KCN, GNP, ZIL, VAI; Writing – original draft: AEA, KCN; Writing – review and editing: AEA, KCN, GNP, ZIL, VAI; Supervision: VAI.

REFERENCES

[1] F. Hucho, C. Weise, Ligand-Gated Ion Channels, Angewandte Chemie International Edition. 40 (2001) 3100–3116. https://doi.org/10.1002/1521-3773(20010903)40:17<3100::AID-ANIE3100>3.0.CO;2-A.

[2] M.J. Ackerman, D.E. Clapham, Ion channels--basic science and clinical disease, N Engl J Med. 336 (1997) 1575–1586. https://doi.org/10.1056/NEJM199705293362207.

[3] B. Hille, Ion channels of excitable membranes, 3rd ed., Sinauer, Sunderland, Mass., 2001.

[4] J. Zheng, M.C. Trudeau, Handbook of ion channels., Crc Press, 2015.

[5] S.P.H. Alexander, A. Mathie, J.A. Peters, E.L. Veale, J. Striessnig, E. Kelly, J.F. Armstrong, E. Faccenda, S.D. Harding, A.J. Pawson, J.L. Sharman, C. Southan, J.A. Davies, CGTP Collaborators, THE CONCISE GUIDE TO PHARMACOLOGY 2019/20: Ion channels, Br J Pharmacol. 176 Suppl 1 (2019) S142–S228. https://doi.org/10.1111/bph.14749.

[6] J. Dunlop, M. Bowlby, R. Peri, D. Vasilyev, R. Arias, High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology, Nat Rev Drug Discov. 7 (2008) 358–368. https://doi.org/10.1038/nrd2552.

[7] D. Belelli, N.L. Harrison, J. Maguire, R.L. Macdonald, M.C. Walker, D.W. Cope, Extrasynaptic GABAA receptors: form, pharmacology, and function, J Neurosci. 29 (2009) 12757–12763. https://doi.org/10.1523/JNEUROSCI.3340-09.2009.

[8] J.W. Lynch, Native glycine receptor subtypes and their physiological roles, Neuropharmacology. 56 (2009) 303–309. https://doi.org/10.1016/j.neuropharm.2008.07.034.

[9] J.-P. Changeux, Allosteric receptors: from electric organ to cognition, Annu Rev Pharmacol Toxicol. 50 (2010) 1–38. https://doi.org/10.1146/annurev.pharmtox.010909.105741.

[10] N.S. Millar, C. Gotti, Diversity of vertebrate nicotinic acetylcholine receptors, Neuropharmacology. 56 (2009) 237–246. https://doi.org/10.1016/j.neuropharm.2008.07.041.

[11] N.M. Barnes, T.G. Hales, S.C.R. Lummis, J.A. Peters, The 5-HT3 receptor--the relationship between structure and function, Neuropharmacology. 56 (2009) 273–284. https://doi.org/10.1016/j.neuropharm.2008.08.003.

[12] D. Lodge, The history of the pharmacology and cloning of ionotropic glutamate receptors and the development of idiosyncratic nomenclature, Neuropharmacology. 56 (2009) 6–21. https://doi.org/10.1016/j.neuropharm.2008.08.006.

[13] M.J. Berridge, The Inositol Trisphosphate/Calcium Signaling Pathway in Health and Disease, Physiol Rev. 96 (2016) 1261–1296. https://doi.org/10.1152/physrev.00006.2016.

[14] M.F. Jarvis, B.S. Khakh, ATP-gated P2X cation-channels, Neuropharmacology. 56 (2009) 208–215. https://doi.org/10.1016/j.neuropharm.2008.06.067.

[15] C.M. Canessa, A.M. Merillat, B.C. Rossier, Membrane topology of the epithelial sodium channel in intact cells, Am J Physiol. 267 (1994) C1682-1690. https://doi.org/10.1152/ajpcell.1994.267.6.C1682.

[16] S. Kellenberger, L. Schild, International Union of Basic and Clinical Pharmacology. XCI. structure, function, and pharmacology of acid-sensing ion channels and the epithelial Na+ channel, Pharmacol Rev. 67 (2015) 1–35. https://doi.org/10.1124/pr.114.009225.

[17] S.M. Trattnig, A. Gasiorek, T.Z. Deeb, E.J.C. Ortiz, S.J. Moss, A.A. Jensen, P.A. Davies, Copper and protons directly activate the zinc-activated channel, Biochem Pharmacol. 103 (2016) 109–117. https://doi.org/10.1016/j.bcp.2016.02.004.

[18] P.S. Miller, T.G. Smart, Binding, activation and modulation of Cys-loop receptors, Trends Pharmacol Sci. 31 (2010) 161–174. https://doi.org/10.1016/j.tips.2009.12.005.

[19] Z. Jiang, Y. Zhou, Using bioinformatics for drug target identification from the genome, Am J Pharmacogenomics. 5 (2005) 387–396. https://doi.org/10.2165/00129785-200505060-00005.

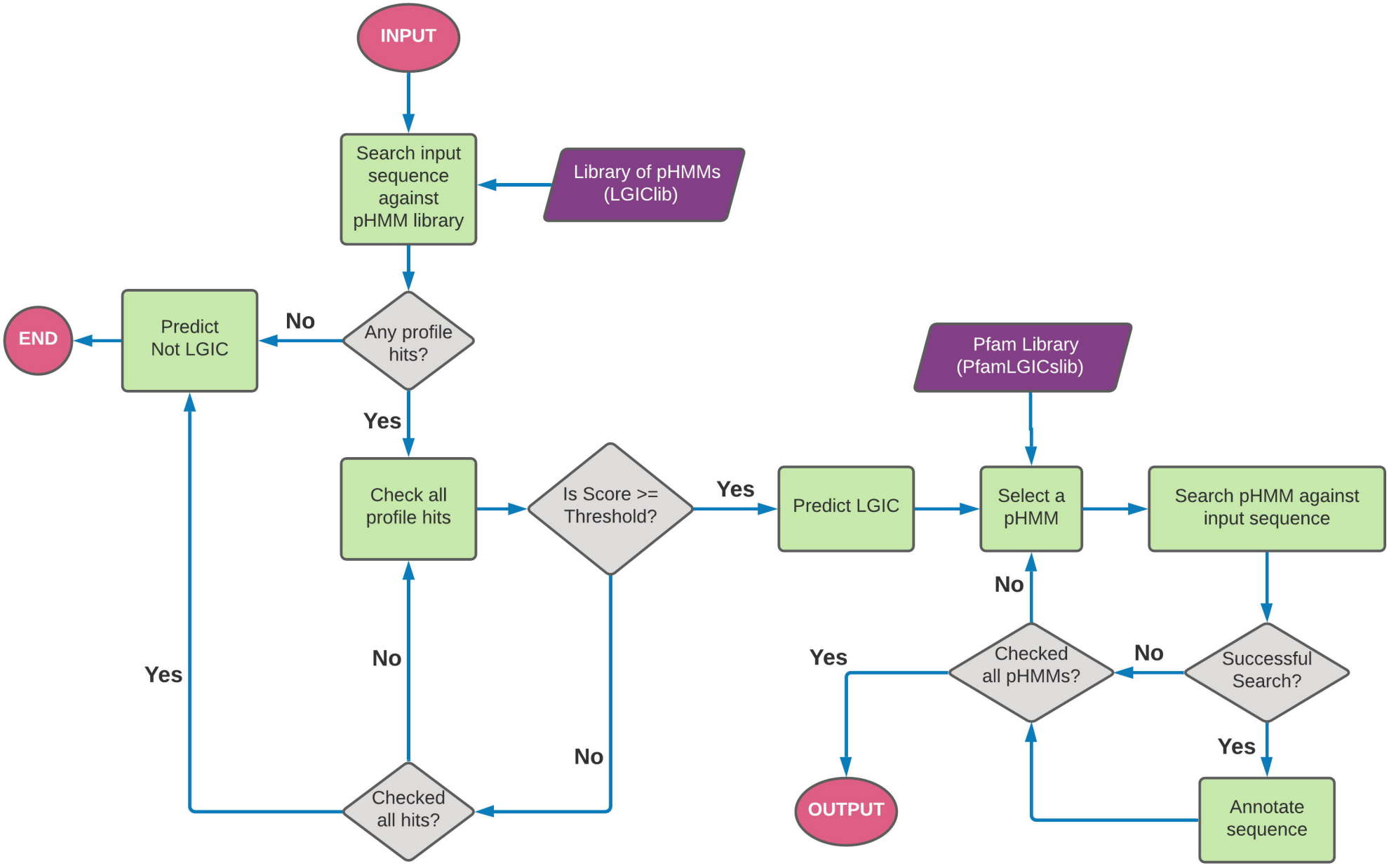
[20] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there?, Nat Rev Drug Discov. 5 (2006) 993–996. https://doi.org/10.1038/nrd2199.

[21] S. Saha, J. Zack, B. Singh, G.P.S. Raghava, VGIchan: prediction and classification of voltage-gated ion channels, Genomics Proteomics Bioinformatics. 4 (2006) 253–258. https://doi.org/10.1016/S1672-0229(07)60006-0.

[22] H. Lin, H. Ding, Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, J Theor Biol. 269 (2011) 64–69. https://doi.org/10.1016/j.jtbi.2010.10.019.

[23] W. Chen, H. Lin, Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine, Comput Biol Med. 42 (2012) 504–507. https://doi.org/10.1016/j.compbiomed.2012.01.003.
[24] W.-X. Liu, E.-Z. Deng, W. Chen, H. Lin, Identifying the subfamilies of voltage-gated potassium channels using feature selection technique, Int J Mol Sci. 15 (2014) 12940–12951. https://doi.org/10.3390/ijms150712940.
[25] A.K. Tiwari, R. Srivastava, An efficient approach for the prediction of ion channels and their subfamilies, Comput Biol Chem. 58 (2015) 205–221. https://doi.org/10.1016/j.compbiolchem.2015.07.002.
[26] J. Gao, W. Cui, Y. Sheng, J. Ruan, L. Kurgan, PSIONplus: Accurate Sequence-Based Predictor of Ion Channels and Their Types, PLoS One. 11 (2016) e0152964. https://doi.org/10.1371/journal.pone.0152964.
[27] Y.-W. Zhao, Z.-D. Su, W. Yang, H. Lin, W. Chen, H. Tang, IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types, Int J Mol Sci. 18 (2017) E1838. https://doi.org/10.3390/ijms18091838.
[28] K. Han, M. Wang, L. Zhang, Y. Wang, M. Guo, M. Zhao, Q. Zhao, Y. Zhang, N. Zeng, C. Wang, Predicting Ion Channels Genes and Their Types With Machine Learning Techniques, Front Genet. 10 (2019) 399. https://doi.org/10.3389/fgene.2019.00399.
[29] J. Gao, H. Wei, A. Cano, L. Kurgan, PSIONplusm Server for Accurate Multi-Label Prediction of Ion Channels and Their Types, Biomolecules. 10 (2020) E876. https://doi.org/10.3390/biom10060876.
[30] J. Gao, Z. Miao, Z. Zhang, H. Wei, L. Kurgan, Prediction of Ion Channels and their Types from Protein Sequences: Comprehensive Review and Comparative Assessment, Curr Drug Targets. 20 (2019) 579–592. https://doi.org/10.2174/1389450119666181022153942.
[31] J.F. Armstrong, E. Faccenda, S.D. Harding, A.J. Pawson, C. Southan, J.L. Sharman, B. Campo, D.R. Cavanagh, S.P.H. Alexander, A.P. Davenport, M. Spedding, J.A. Davies, NC-IUPHAR, The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY, Nucleic Acids Res. 48 (2020) D1006–D1021. https://doi.org/10.1093/nar/gkz951.
[32] T. UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46 (2018) 2699. https://doi.org/10.1093/nar/gky092.
[33] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics. 22 (2006) 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
[34] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics. 26 (2010) 680–682. https://doi.org/10.1093/bioinformatics/btq003.
[35] F. Sievers, D.G. Higgins, The Clustal Omega Multiple Alignment Package, Methods Mol Biol. 2231 (2021) 3–16. https://doi.org/10.1007/978-1-0716-1036-7_1.
[36] S.C. Potter, A. Luciani, S.R. Eddy, Y. Park, R. Lopez, R.D. Finn, HMMER web server: 2018 update, Nucleic Acids Research. 46 (2018) W200–W204. https://doi.org/10.1093/nar/gky448.
[37] Z.S. Ioannidou, M.C. Theodoropoulou, N.C. Papandreou, J.H. Willis, S.J. Hamodrakas, CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models, Insect Biochem Mol Biol. 52 (2014) 51–59. https://doi.org/10.1016/j.ibmb.2014.06.004.
[38] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230. https://doi.org/10.1093/nar/gkt1223.
[39] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management. 45 (2009) 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
[40] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J Mol Biol. 305 (2001) 567–580. https://doi.org/10.1006/jmbi.2000.4315.
[41] P.A. Davies, W. Wang, T.G. Hales, E.F. Kirkness, A novel class of ligand-gated ion channel is activated by Zn2+, J Biol Chem. 278 (2003) 712–717. https://doi.org/10.1074/jbc.M208814200.

[42] I. Hanukoglu, A. Hanukoglu, Epithelial sodium channel (ENaC) family: Phylogeny, structure-function, tissue distribution, and associated inherited diseases, Gene. 579 (2016) 95–132. https://doi.org/10.1016/j.gene.2015.12.061.
[43] A.M. Kinnersley, F.J. Turano, Gamma Aminobutyric Acid (GABA) and Plant Responses to Stress, Critical Reviews in Plant Sciences. 19 (2000) 479–509. https://doi.org/10.1080/07352680091139277.
[44] A.J. Thompson, M. Alqazzaz, C. Ulens, S.C.R. Lummis, The pharmacological profile of ELIC, a prokaryotic GABA-gated receptor, Neuropharmacology. 63 (2012) 761–767. https://doi.org/10.1016/j.neuropharm.2012.05.027.
[45] P. Dietrich, U. Anschütz, A. Kugler, D. Becker, Physiology and biophysics of plant ligand-gated ion channels, Plant Biology. 12 (2010) 80–93. https://doi.org/10.1111/j.1438-8677.2010.00362.x.
[46] G.Q. Chen, C. Cui, M.L. Mayer, E. Gouaux, Functional characterization of a potassium-selective prokaryotic glutamate receptor, Nature. 402 (1999) 817–821. https://doi.org/10.1038/45568.
[47] K.J. Alzayady, A. Sebé-Pedrós, R. Chandrasekhar, L. Wang, I. Ruiz-Trillo, D.I. Yule, Tracing the Evolutionary History of Inositol, 1, 4, 5-Trisphosphate Receptor: Insights from Analyses of Capsaspora owczarzaki Ca2+ Release Channel Orthologs, Mol Biol Evol. 32 (2015) 2236–2253. https://doi.org/10.1093/molbev/msv098.
[48] Y. Horiuchi, R. Kimura, N. Kato, T. Fujii, M. Seki, T. Endo, T. Kato, K. Kawashima, Evolutional study on acetylcholine expression, Life Sci. 72 (2003) 1745–1756. https://doi.org/10.1016/s0024-3205(02)02478-5.
[49] X. Cai, P2X receptor homologs in basal fungi, Purinergic Signal. 8 (2012) 11–13. https://doi.org/10.1007/s11302-011-9261-8.
[50] S.J. Fountain, K. Parkinson, M.T. Young, L. Cao, C.R.L. Thompson, R.A. North, An intracellular P2X receptor required for osmoregulation in Dictyostelium discoideum, Nature. 448 (2007) 200–203. https://doi.org/10.1038/nature05926.
[51] A. Tasneem, L.M. Iyer, E. Jakobsson, L. Aravind, Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels, Genome Biology. (2005). https://doi.org/10.1186/gb-2004-6-1-r4.
[52] G. Rendon, M.R. Kantorovitz, J.L. Tilson, E. Jakobsson, Identifying bacterial and archaeal homologs of pentameric ligand-gated ion channel (pLGIC) family using domain-based and alignment-based approaches, Channels. 5 (2011) 325–344. https://doi.org/10.4161/chan.5.4.16822.
[53] F. Ashcroft, D. Gadsby, C. Miller, Introduction. The blurred boundary between channels and transporters, Philosophical Transactions of the Royal Society B: Biological Sciences. 364 (2009) 145–147. https://doi.org/10.1098/rstb.2008.0245.


INPUT
Search input sequence against pHMM library
Library of pHMMs (LGIClib)
Any profile hits?
No
Predict Not LGIC
END
Yes
Check all profile hits
Is Score >= Threshold?
Yes
Predict LGIC
Pfam Library (PfamLGICslib)
Select a pHMM
Search pHMM against input sequence
Successful Search?
No
Yes
Checked all pHMMs?
No
Yes
OUTPUT
Annotate sequence
Checked all hits?
Yes
No
No

Name	LGIC	Family	Domains
sp|P63138|GBRB2_RAT	LGIC	GABAA	32-474

RESULTS TABLE

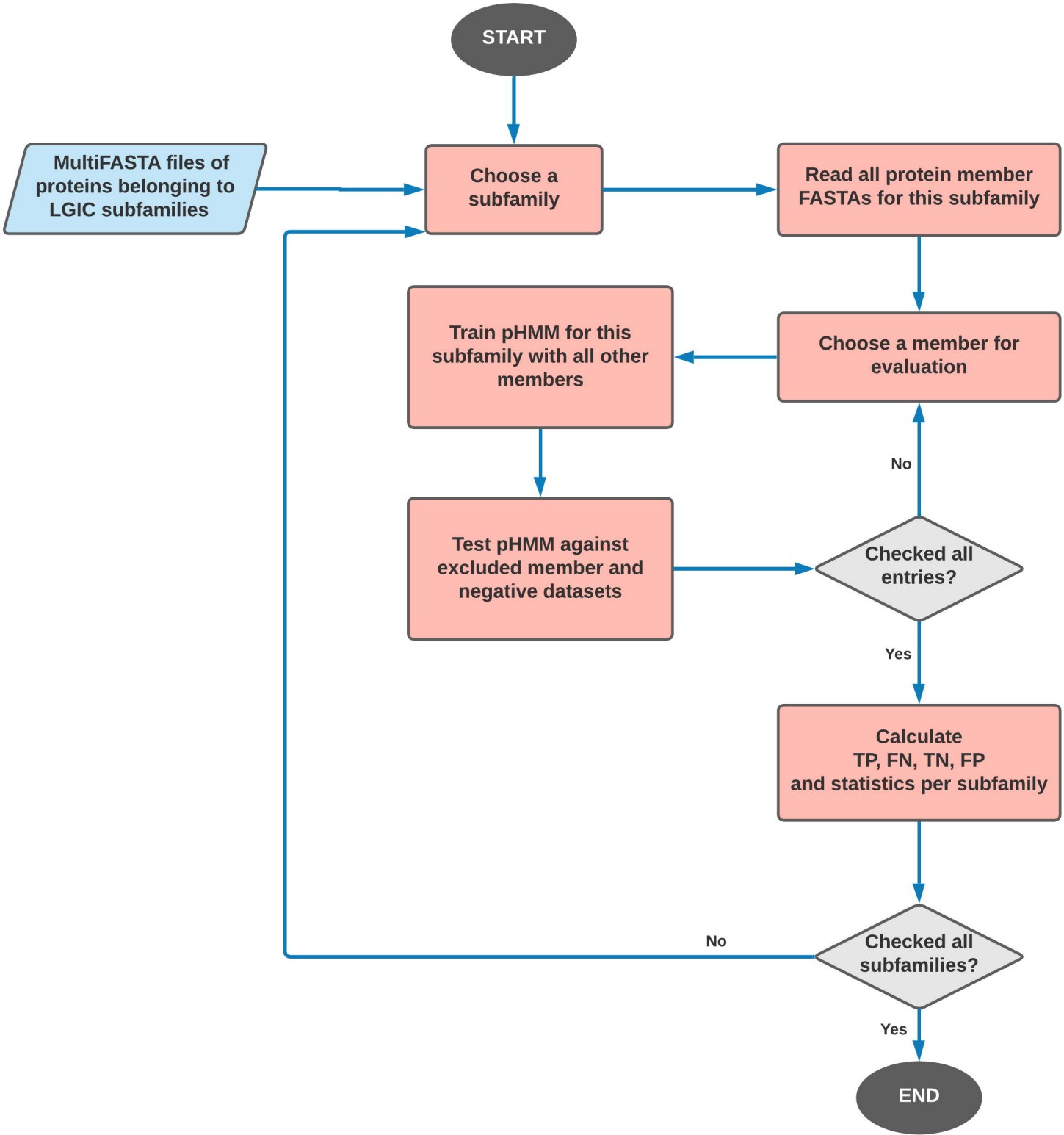
Full name
sp|P63138|GBRB2_RAT

pHMM Name	pHMM Domain Range	Domain Number	SCORE
GABAA	32-474	1	539.7

Pfam Domain Name	Pfam Domain Range	Pfam Symbol	Pfam Domain Score
Neur_chan_memb	249-469	b	160.9
Neur_chan_LBD	38-241	a	169.7

SEQUENCE ANNOTATION

```
  1  mwrvrkrgyfgiwsfpliiaavcaqsvndpsNMSLVKETVDRLLKGYDIRLRPDFGGPPV  60
pHMM ------------------------------++++++++++++++++++++++++++++
Pfam ------------------------------------aaaaaaaaaaaaaaaaaaaaaaa
 61  AVGMNIDIASIDMVSEVNMDYTLTMYFQQAWRDKRLSYNVIPLNLTLDNRVADQLWVPDT  120
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
121  YFLNDKKSFVHGVTVKNRMIRLHPDGTVLYGLRITTTAACMMDLRRYPLDEQNCTLEIES  180
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
181  YGYTTDDIEFYWRGDDNAVTGVTKIELPQFSIVDYKLITKKVVFSTGSYPRLSLSFKLKR  240
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
241  NIGYFILQTYMPSILITILSWVSFWINYDASAARVALGITTVLTMTTINTHLRETLPKIP  300
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam a-------bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
301  YVKAIDMYLMGCFVFVFMALLEYALVNYIFFGRGPQRQKKAAEKAANAnnekmRLDVNKM  360
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
361  DPHEnillstlEIKNEMatseavmglgDPRSTMLAYDASSIQYRKAglpRHSFGRNaler  420
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
421  HVAQKKSRLRRRASQlKITIPDLTDVNAIDRWSRIFFPVVFSFFNIVYWLYYVN  474
pHMM ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb-----
```


START
MultiFASTA files of proteins belonging to LGIC subfamilies
Choose a subfamily
Read all protein member FASTAs for this subfamily
Train pHMM for this subfamily with all other members
Choose a member for evaluation
No
Test pHMM against excluded member and negative datasets
Checked all entries?
Yes
Calculate TP, FN, TN, FP and statistics per subfamily
No
Checked all subfamilies?
Yes
END