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

Katerina C. Nastou[#], Georgios N. Petichakis[#], Zoi I. Litou and Vassiliki A. Iconomidou*

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

[#]Equally contributing first authors

*To whom correspondence should be addressed

Assistant 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. However, special topological characteristics of these proteins have not been utilized to date, which results in weaknesses regarding the correct class categorization of predicted proteins. Here we report the development of LiGIoNs, a profile Hidden Markov Model (pHMM) method for the prediction and ligand-based classification of LGICs, utilizing their special topological characteristics. The method consists of a library of 35 pHMMs, built from the alignment of transmembrane segments of representative LGIC sequences. In addition,14 Pfam pHMMs are used to further annotate and correctly 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

Typically, living cells exhibit a membrane potential at their plasma membrane [1]. However, the lipid bilayer that forms the plasma membrane poses an immense energy barrier for charged particles. To overcome this obstacle, cells use specialized transmembrane proteins that carry the ion current, known as ion channels [2]. 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 a result of either changes to the membrane potential or binding of specific ligands. Thus, channels are classified, according to their gating trigger, into Voltage-Gated Ion Channels (VGICs) and Ligand-Gated Ion Channels (LGICs) [4]. They both are extremely diverse and are composed of numerous members further classified into various subfamilies.

Many genes encode LGICs' subunits and most of those form heteropolymers. The variety of 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 emerge as attractive targets for the development of new therapeutic agents [6]. By convention, LGICs comprise the excitatory, cation-selective, nicotinic acetylcholine receptors [7, 8], 5-HT_3 receptors [9], ionotropic glutamate receptors [10], IP_3 receptors [11], P2X receptors [12], epithelial sodium channels [13] and acid-sensing (proton-gated) ion channels [14] and the inhibitory, anion-selective, $GABA_A$ receptors [15] and glycine receptors [16]. The nicotinic acetylcholine, 5-HT_3, $GABA_A$ and glycine receptors (and an additional zinc-activated channel) form the family of Cys-loop receptors [17]. The special structural characteristics of all LGIC subfamilies 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 underline 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 designated role as drug targets [18, 19], several methods have been developed for the prediction of these proteins using information encoded exclusively in their amino acid sequence [20-27]. Ion channel prediction is mainly based on SVMs, and several feature selection techniques have been used to train the machine learning classifiers, mainly for the prediction and classification of VGICs. The main drawback of all methods is that the informative parameters used for training are amino acid, dipeptide and tripeptide compositions, while the special topological features of ion channels are not taken into consideration during method development. On top of that, while IonChanPred [21], PSIONplus [22] and two recently developed machine learning classifiers by Tiwari and Srivastava [27] and by Han *et al.* [26] are the only methods that can detect the class of LGICs, neither those, nor any other method classifies LGICs into subfamilies.

As suggested in the expert review by Lin and Chen [28] new ion channel predictors should utilize physicochemical characteristics, overrepresented motifs or functional domains of these proteins during their training. Taking this suggestion under consideration 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). These pHMMs are created by utilizing the special topological characteristics of LGICs, and specifically the amino acid sequence of the transmembrane segments of these proteins. LiGIoNs is the first method that on top of detecting LGICs, performs a ligand-based classification into the 10 known 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 algorithm consists of two levels: a prediction (detection and classification) and an annotation level. For the prediction level of the method, a dataset of LGICs was collected and classified into subfamilies following the IUPHAR classification scheme [29] already presented in **Table 1**. For the creation of the training set, proteins were originally collected from the IUPHAR database [29]. 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 reviewed UniProt [30] entries for LGICs, and incorporated those in the final training dataset. In addition, we cross-checked our dataset with entries in LGICdb [31]. However, this database has not been updated since 2007 and the only records that we hadn't previously detected belonged to the unreviewed subset of UniProt entries (UniProt/TrEMBL), which we had already decided not to use to train our method. Thus, we opted to not use this resource further. The final dataset of protein sequences used for the LiGIoNs training dataset is shown in **Supplementary Table 1**.

The boundaries of all transmembrane regions of LGICs used in this study were extracted from information documented in UniProt. These boundaries were used to extract transmembrane segments of proteins belonging to the same subfamily, which were aligned according to their arrangement in the sequence, i.e., the first transmembrane segment of a sequence was aligned with the first transmembrane segment of the remaining proteins of the same subfamily, the second to the second, and so on. The procedure was applied to all transmembrane segments of each subfamily. HMMER [32] was then used to construct the respective pHMMs corresponding to each transmembrane segment. Each LGIC subfamily has as many pHMMs as its transmembrane segments, as shown in **Table 2**.

Table 2. pHMMs constructed for each LGIC subfamily. A pHMM library (LGICslib) containing all these profiles was constructed and was incorporated in the prediction level of our method.

LGIC subfamily	*pHMMs*
Epithelial sodium channels (ENaCs)	Epithelial_1, Epithelial_2
P2X Receptors	P2X_1, P2X_2
Acid-sensing ion channels (ASICs)	Acid_1, Acid_2
Ionotropic Glutamate Receptors	Ionotropic_1, Ionotropic_2, Ionotropic_3

LGIC subfamily	*pHMMs*
Nicotinic Acetylcholine Receptors	Nicotinic_1, Nicotinic_2, Nicotinic_3, Nicotinic_4
5-HT_3 Receptors	5HT3_1, 5HT3_2, 5HT3_3, 5HT3_4
$GABA_A$ Receptors	GABAA_1, GABAA_2, GABAA_3, GABAA_4
Glycine Receptors	Glycine_1, Glycine_2, Glycine_3, Glycine_4
Zinc-activated channels (ZACs)	ZAC_1, ZAC_2, ZAC_3, ZAC_4
IP_3 Receptors	IP3_1, IP3_2, IP3_3, IP3_4, IP3_5, IP3_6

The following procedure is used to characterize an unknown protein sequence as an LGIC. Initially, the unknown protein sequence is scanned against the library of all pHMMs (LGICslib) presented in **Table 2**. This is followed by recording the number of pHMMs that align with the unknown protein sequence. In order for an unknown protein sequence to be characterized as a member of a specific LGIC subfamily, at least half of the pHMMs corresponding to this subfamily ($n \geq$(number of TM segments in subfamily)/2) must be aligned with the sequence, with a score higher than the threshold set for each pHMM. The thresholds for each pHMM were set manually in order to maximize sensitivity and specificity, following the protocol introduced by Ioannidou *et al.* [33].

For the creation of the annotation level of LiGIoNs, all characteristic pHMMs that are found on LGICs were identified and isolated from the Pfam protein family database [34]. This procedure preceded chronologically the construction of the pHMMs in LGICslib, in an attempt to identify known 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 3**) in combination with those in LGICslib (**Table 2**), allowed both an additional validation of the results obtained using LiGIoNs in proteomes and the creation of the annotation level of the method.

The characteristic pHMMs 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 3** and were used to create a second library, named PfamLGICslib. HMMER was used once again to scan sequences – that have been previously characterized as LGICs by

LiGIoNs – against PfamLGICslib, and sequences that had a "hit" from pHMMs in the Pfam library were annotated with these domains.

Table 3. 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 MSA. Then we measured the correct classification ability of the newly created pHMM, to the removed sequence and to randomly selected 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 soluble proteins. The number of sequences that were selected from each negative dataset at each run was equal to the number of sequences used to train the corresponding pHMM every time. This was done to better balance the difference between the number of sequences in the positive and the negative datasets, while not compromising the validation

procedure. At this point, it should be noted that it was not possible to perform this screening for ZACs, as only two protein members for this subfamily have been recorded to date.

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, which were subjected to homology reduction using the CD-HIT clustering method [35, 36]. A 30% similarity threshold homology was used, and a representative set of 1500 transmembrane proteins was isolated (**Supplementary Table 2**). The other negative dataset comprises of 300 globular proteins and is the same one used for the evaluation of performance for IonChanPred 2.0 [21] (**Supplementary Table 2**).

In addition to the above evaluation, the method was compared with the IonChanPred 2.0 method to test its overall performance. It should be emphasized at this point that a comparison with the PSIONPlus method [22] and the classifiers developed by Tiwari and Srivastava [27] and by Han *et al.* [26] was practically impossible, as these methods are not available online. It should also be mentioned that none of the above methods are capable of classifying LGICs into subfamilies.

For the prediction performance of LiGIoNs five measures were used, namely Accuracy, Sensitivity, Specificity, Balanced Accuracy and Matthew's Correlation Coefficient. True/false positives (TP, FP) and true/false negatives (TN, FN) were counted on a per protein basis.

Accuracy is the proximity of measurement results to the true value and is calculated as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1).$$

Sensitivity, or true positive rate is:

$$Sn = \frac{TP}{(TP + FN)} \quad (2.2),$$

and Specificity, or true negative rate is:

$$Sp = \frac{TN}{(TN + FP)} \text{ (2.3).}$$

Besides these measures, the balanced accuracy and Matthew's Correlation Coefficient (MCC) were used to evaluate the performance of LiGiONs. Balanced accuracy is the average of sensitivity and specificity and, together with MCC, is considered a better measure [37] when the data sizes of the positive and negative datasets are not balanced. MCC is calculated as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{((TN + FN) \cdot (TN + FP) \cdot (TP + FN) \cdot (TP + FP))}} \text{ (2.4).}$$

Moreover, LiGIoNs was applied to 30 selected reference eukaryotic proteomes (**Supplementary Table 3**) retrieved from UniProt **(**release: 2019_07**)** in order to further assess the method's ability to detect LGICs in proteomes from various eukaryotic kingdoms.

3. RESULTS AND DISCUSSION

3.1. LiGIoNs algorithm

The prediction level of LiGIoNs consists of a library of 35 pHMMs (LGICslib), created from the MSAs of the protein's transmembrane segments belonging to the 10 LGIC subfamilies (**Table 2**). For a protein to be characterized as a subunit of a specific LGIC subfamily, at least half of the pHMMs in LGICslib corresponding to this subfamily must score higher than the profile's threshold. For example, for a channel to qualify as a $GABA_A$ receptor, at least two of the four subfamily profiles (**Table 2**) must be detected in the sequence and score higher than the threshold set. If multiple subfamilies meet the aforementioned conditions, the one with the highest overall score is chosen to characterize the unknown sequence. For the annotation level, a library of 14 pHMMs containing characteristic LGIC domains recorded in Pfam was created (PfamLGICslib, **Table 3**). 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.

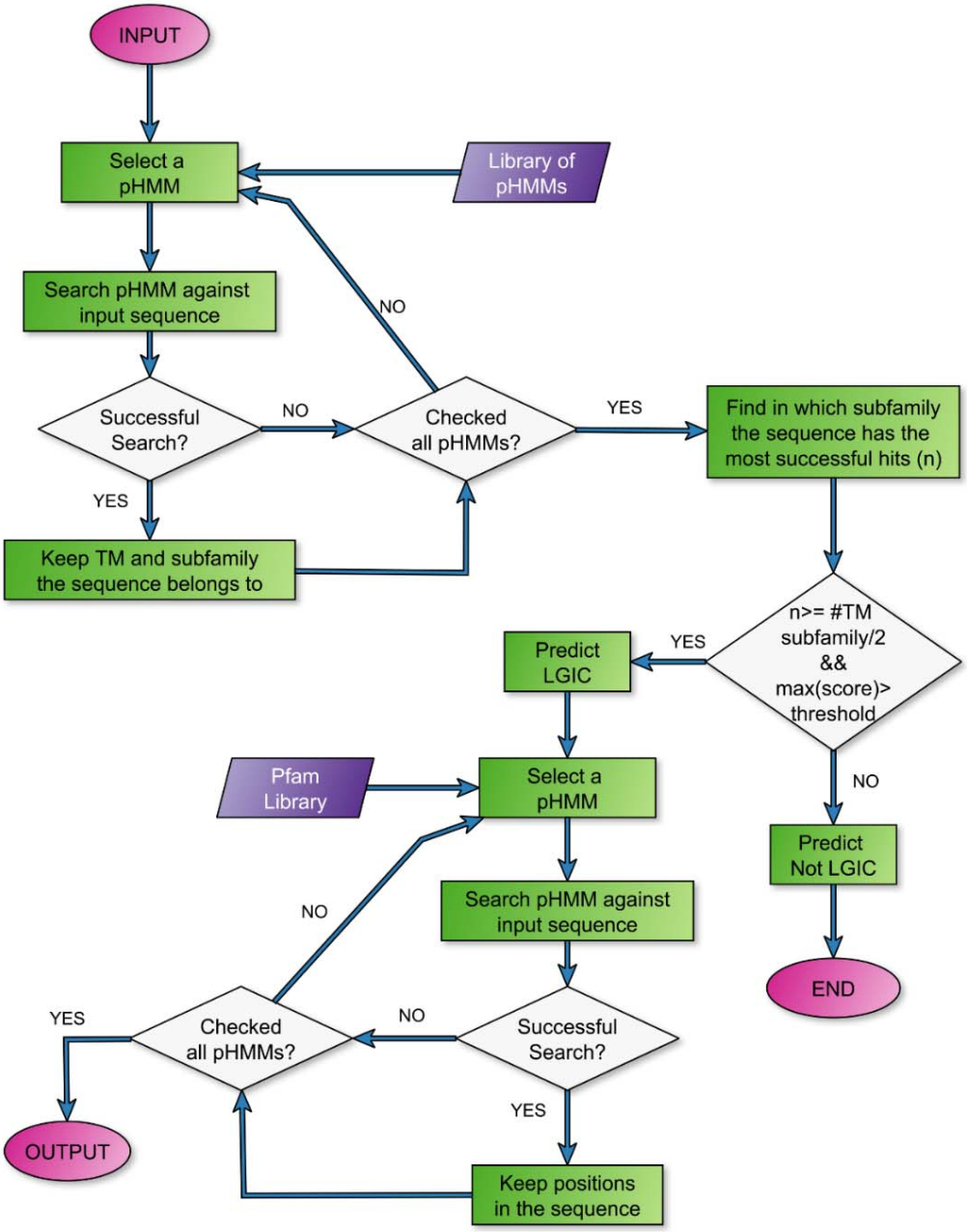


Figure 1. Flowchart of the LiGIoNs algorithm. A fasta formatted sequence is used as input. A pHMM from the LGICslib is selected and is searched against the input sequence. If the search is successful, the transmembrane segment and the LGIC subfamily to which the selected pHMM corresponds to is recorded and regardless of the result, the sequence is searched until all 35 pHMMs have been examined. Afterwards, the subfamily where the sequence presents most successful hits (n) is selected and an additional test is performed to ensure that the number of successful hits is greater or equal than the number of transmembrane segments of the subfamily and that the maximum value of the score is over the profiles' thresholds. 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 an LGIC and is later 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 either by submitting a single or a set of protein sequences in the textbox provided, or by uploading a file with fasta formatted 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 perform a new query and download the results in a text file.

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**.

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) is used in this example. The four pHMMs of LGICslib that correspond to $GABA_A$ receptors (Table 2), as well as two pHMMs

from PfamLGICslib, characteristic of LGICs that belong to the Cys-loop family (Table 3) are detected in the sequence. Two lines under the sequence show the positions where the different 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 – the position and score of the transmembrane segments, Pfam domain(s) present in the protein and the protein sequence. 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.

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**.

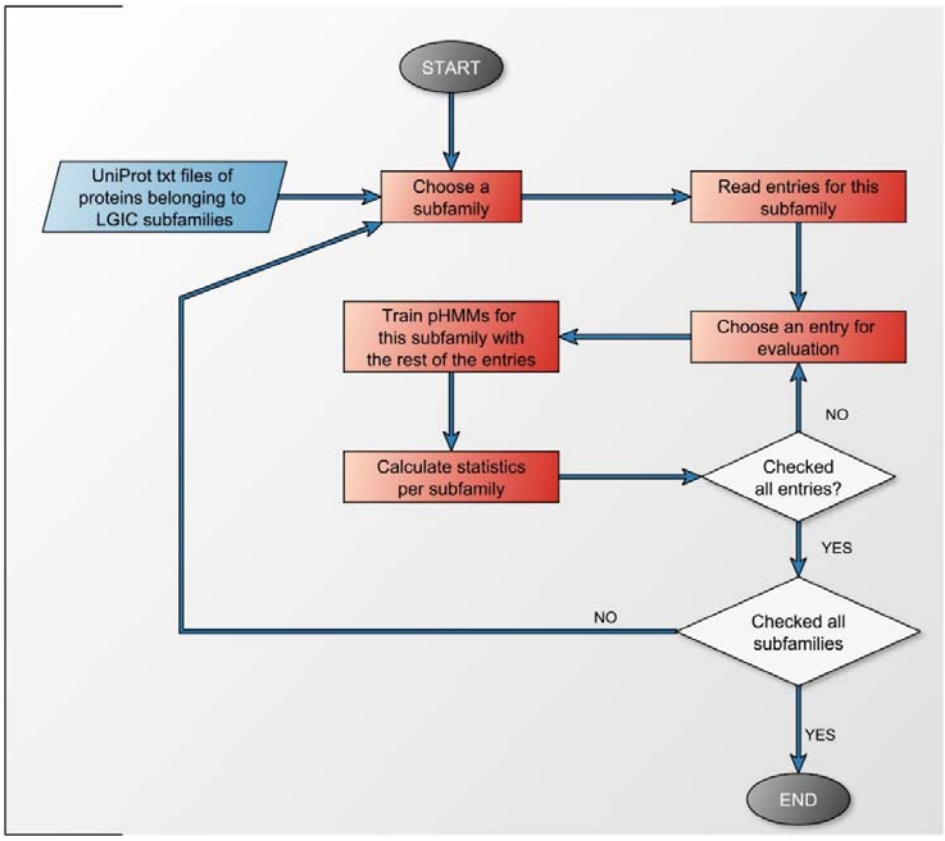


Figure 3. The workflow of the jackknife cross-validation experiment to assess the performance of LiGIoNs. Initially a subfamily of LGICs is chosen and all entries used for the training of the specif-

ic 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 sequence, as well as, an equal randomly selected number of sequences of the three negative datasets (**Supplementary File 2**). The method's ability to correctly identify the sequences is calculated 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.

LiGIoNs performs very well during cross-validation, with high overall performance metrics against all negative test datasets (**Table 4**, **Table 5** and **Table 6**). The results from the jackknife test showed the method's ability to correctly identify pseudo-novel sequences as LGICs (Sensitivity over 90% in all cases), while not erroneously detecting non-LGICs as such (Specificity over 99% in all cases). These values are indicative of the fact that pHMMs have a good ability to differentiate between LGICs and non-LGICs, however they have a slightly worse performance when classifying LGICs into the right subfamily. For that reason, the method was executed normally following the workflow described in **Section 3.1**, and proteins that have hits against multiple subfamilies are classified based on the highest overall score (see **Figure 1**). The method's accuracy and MCC are 100% in all cases when this type of validation is performed.

Table 4. Results from the cross-validation of LiGIoNs using the jackknife technique against the 300 globular proteins negative dataset.

LGIC subfamily	***Accuracy (%)***	***Sensitivity (%)***	***Specificity (%)***	***MCC (%)***
Epithelial Sodium Channels (ENaCs)	99.92	97.22	100.00	98.56
P2X Receptors	99.69	92.00	100.00	95.76
Acid-sensing Ion Channels (ASICs)	99.82	95.65	100.00	97.71
Ionotropic Glutamate Receptors	99.96	96.67	100.00	98.30
Nicotinic Acetylcholine Receptors	99.99	99.14	100.00	99.56
5-HT3 Receptors	100.00	100.00	100.00	100.00
GABAA Receptors	100.00	100.00	100.00	100.00
Glycine Receptors	100.00	100.00	100.00	100.00
Zinc-activated Channels (ZACs)	-	-	-	-
IP3 Receptors	99.52	92.86	100.00	96.12

Table 5. Results from the cross-validation of LiGIoNs using the jackknife technique against the 1500 transmembrane proteins of the negative dataset.

LGIC subfamily	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*MCC (%)*
Epithelial Sodium Channels (ENaCs)	99.85	97.22	99.92	97.15
P2X Receptors	99.69	92.00	100.00	95.76
Acid-sensing Ion Channels (ASICs)	99.64	95.65	99.81	95.46
Ionotropic Glutamate Receptors	99.96	96.67	100.00	98.30
Nicotinic Acetylcholine Receptors	99.99	99.14	100.00	99.56
5-HT3 Receptors	100.00	100.00	100.00	100.00
GABAA Receptors	99.98	100.00	99.98	99.35
Glycine Receptors	100.00	100.00	100.00	100.00
Zinc-activated Channels (ZACs)	-	-	-	-
IP3 Receptors	99.52	92.86	100.00	96.12

Table 6. Results from the cross-validation of LiGIoNs using the jackknife technique against LGICs belonging to other subfamilies than the one validated.

LGIC subfamily	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*MCC (%)*
Epithelial Sodium Channels (ENaCs)	99.92	97.22	100.00	98.56
P2X Receptors	99.69	92.00	100.00	95.76
Acid-sensing Ion Channels (ASICs)	99.82	95.65	100.00	97.71
Ionotropic Glutamate Receptors	99.96	96.67	100.00	98.30
Nicotinic Acetylcholine Receptors	99.99	99.14	100.00	99.56
5-HT3 Receptors	100.00	100.00	100.00	100.00
GABAA Receptors	94.62	100.00	94.55	42.66
Glycine Receptors	84.31	100.00	83.39	46.70
Zinc-activated Channels (ZACs)	-	-	-	-
IP3 Receptors	99.52	92.86	100.00	96.12

Even though all results presented above appear excellent, it should be emphasized that they are probably products of overfitting, especially, since the datasets for some LGIC subfamilies are extremely small (**Supplementary File 1**). Moreover, results could not be produced for Zinc-activated Channels due to the small number of protein sequences belonging to this subfamily (only two sequences). The improvement of the method's ability to identify and characterize more proteins belonging to these subfamilies will be realized 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 [21]. Based on the results presented in **Table 7**, it is obvious that our method outperforms IonChanPred 2.0 in the detection of LGICs. The datasets that were used to compare the two methods are the same as those presented in the original IonChanPred 2.0 publication [21]. 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 7. Comparison of LiGIoNs with IonChanPred 2.0 in their ability to detect LGICs from their amino acid sequence.

Method	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Balanced Accuracy (%)*	*MCC (%)*
LiGIoNs	**99.85**	**1**	**99.75**	**99.87**	**99.71**
IonChanPred 2.0	98.72	**1**	97.08	98.54	97.43

3.4. Application in eukaryotic reference proteomes

The application of LiGIoNs in 30 eukaryotic reference proteomes showed that, ca. up to 0.5% of proteins in these proteomes are potential LGICs (**Supplementary Table 3**). The percentages vary significantly based on the kingdom and phylum in which these organisms belong. Specifically, in all proteomes of organisms from non-metazoan kingdoms – except for *Dictyostelium discoideum* and *Capsaspora owczarzacki* – no protein belonging to the ten LGIC subfamilies is detected by LiGIoNs. In addition, in both these proteomes, the proteins designated as LGICs are very few and account for 0.02-0.05% of the total number of proteins. On the other hand, results are quite different for Metazoa, where LGICs of these subfamilies account for 0.2-0.5% of the proteomes. There is no statistically significant difference regarding the percentage of LGICs between the different metazoan phylums, with Arthropoda and Cephalochordata presenting lower absolute values of protein representatives in their proteomes, a finding that could be easily attributed to their overall smaller proteome size.

The inability of the method to detect LGICs in other kingdoms beyond Metazoa, further establishes our fears regarding the overfitting of the method to currently available data. It is thus extremely difficult to apply the method to other proteomes, as is. The performance of good manual annotation of proteomes belonging to

other kingdoms, like plants or fungi, would allow the retraining of the pHMMs we have created to contain data derived from all eukaryotic kingdoms and phylums. For this purpose, we have developed programmatic scripts that recreate the 35 pHMMs when a new training dataset is provided, thus allowing for the constant training of the method with new data, when those become available. The scripts are available through the home page of LiGIoNs at http://bioinformatics.biol.uoa.gr/ligions/. Currently, due to the method's inability to detect and annotate proteins in the majority of the 30 eukaryotic reference proteomes, it was considered untimely to apply LiGIoNs to all eukaryotic reference proteomes or to prokaryotic proteomes, as any attempt to comment on the results could lead to biased conclusions.

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, using information encoded in their special topological features. Moreover, LiGIoNs annotates predicted LGICs with information from Pfam, providing a full description of each sequence's characteristics to the method's users.

LiGIoNs exhibits very high specificity and sensitivity rendering it a prototype for the detection of homologous multi-pass transmembrane proteins belonging to any of the several known classes and families of the same protein type. The small number of proteins used to train the method is an additional feature that allows the extension of its application to many other families of transmembrane proteins, both of prokaryotic and eukaryotic origin. The programmatic scripts we provide through our webpage can be used to train predictors for other families – if appropriate changes are applied – where multiple "domains" can be used to detect proteins belonging to the same group.

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 overcome the issues we have faced when applying the method to evolutionary distant proteomes than those used in the creation of the 35 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: KCN, GNP, ZIL; Method design and development: GNP, KCN; Web Application Design and Development: GNP; Web Application Quality Assurance: KCN, GNP, ZIL, VAI; Writing – original draft: KCN; Writing – review and editing: KCN, GNP, ZIL, VAI; Supervision: VAI.

REFERENCES

[1] F. Hucho, C. Weise, Ligand-Gated Ion Channels, Angewandte Chemie International Edition, 40 (2001) 3100-3116.

[2] M.J. Ackerman, D.E. Clapham, Ion channels--basic science and clinical disease, N Engl J Med, 336 (1997) 1575-1586.

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

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

[5] S.P. Alexander, A. Christopoulos, A.P. Davenport, E. Kelly, N.V. Marrion, J.A. Peters, E. Faccenda, S.D. Harding, A.J. Pawson, J.L. Sharman, C. Southan, J.A. Davies, C. Collaborators, THE CONCISE GUIDE TO PHARMACOLOGY 2017/18: G protein-coupled receptors, Br J Pharmacol, 174 Suppl 1 (2017) S17-S129.

[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.

[7] J.P. Changeux, Allosteric receptors: from electric organ to cognition, Annu Rev Pharmacol Toxicol, 50 (2010) 1-38.

[8] N.S. Millar, C. Gotti, Diversity of vertebrate nicotinic acetylcholine receptors, Neuropharmacology, 56 (2009) 237-246.

[9] N.M. Barnes, T.G. Hales, S.C. Lummis, J.A. Peters, The 5-HT3 receptor--the relationship between structure and function, Neuropharmacology, 56 (2009) 273-284.

[10] D. Lodge, The history of the pharmacology and cloning of ionotropic glutamate receptors and the development of idiosyncratic nomenclature, Neuropharmacology, 56 (2009) 6-21.

[11] M.J. Berridge, The Inositol Trisphosphate/Calcium Signaling Pathway in Health and Disease, Physiol Rev, 96 (2016) 1261-1296.

[12] M.F. Jarvis, B.S. Khakh, ATP-gated P2X cation-channels, Neuropharmacology, 56 (2009) 208-215.

[13] 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.

[14] 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.

[15] 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.

[16] J.W. Lynch, Native glycine receptor subtypes and their physiological roles, Neuropharmacology, 56 (2009) 303-309.

[17] P.S. Miller, T.G. Smart, Binding, activation and modulation of Cys-loop receptors, Trends Pharmacol Sci, 31 (2010) 161-174.

[18] Z. Jiang, Y. Zhou, Using bioinformatics for drug target identification from the genome, Am J Pharmacogenomics, 5 (2005) 387-396.

[19] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there?, Nat Rev Drug Discov, 5 (2006) 993-996.

[20] 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.

[21] 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).

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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, Frontiers in Genetics, 10 (2019).

[27] A.K. Tiwari, R. Srivastava, An efficient approach for the prediction of ion channels and their subfamilies, Comput Biol Chem, 58 (2015) 205-221.

[28] H. Lin, W. Chen, Briefing in Application of Machine Learning Methods in Ion Channel Prediction, The Scientific World Journal, 2015 (2015) 7.

[29] S.D. Harding, J.L. Sharman, E. Faccenda, C. Southan, A.J. Pawson, S. Ireland, A.J.G. Gray, L. Bruce, S.P.H. Alexander, S. Anderton, C. Bryant, A.P. Davenport, C. Doerig, D. Fabbro, F. Levi-Schaffer, M. Spedding, J.A. Davies, I. Nc, The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY, Nucleic Acids Res, 46 (2018) D1091-D1106.

[30] UniProt_Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res, 46 (2018) 2699.

[31] M. Donizelli, M.A. Djite, N. Le Novere, LGICdb: a manually curated sequence database after the genomes, Nucleic Acids Res, 34 (2006) D267-269.

[32] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res, 39 (2011) W29-37.

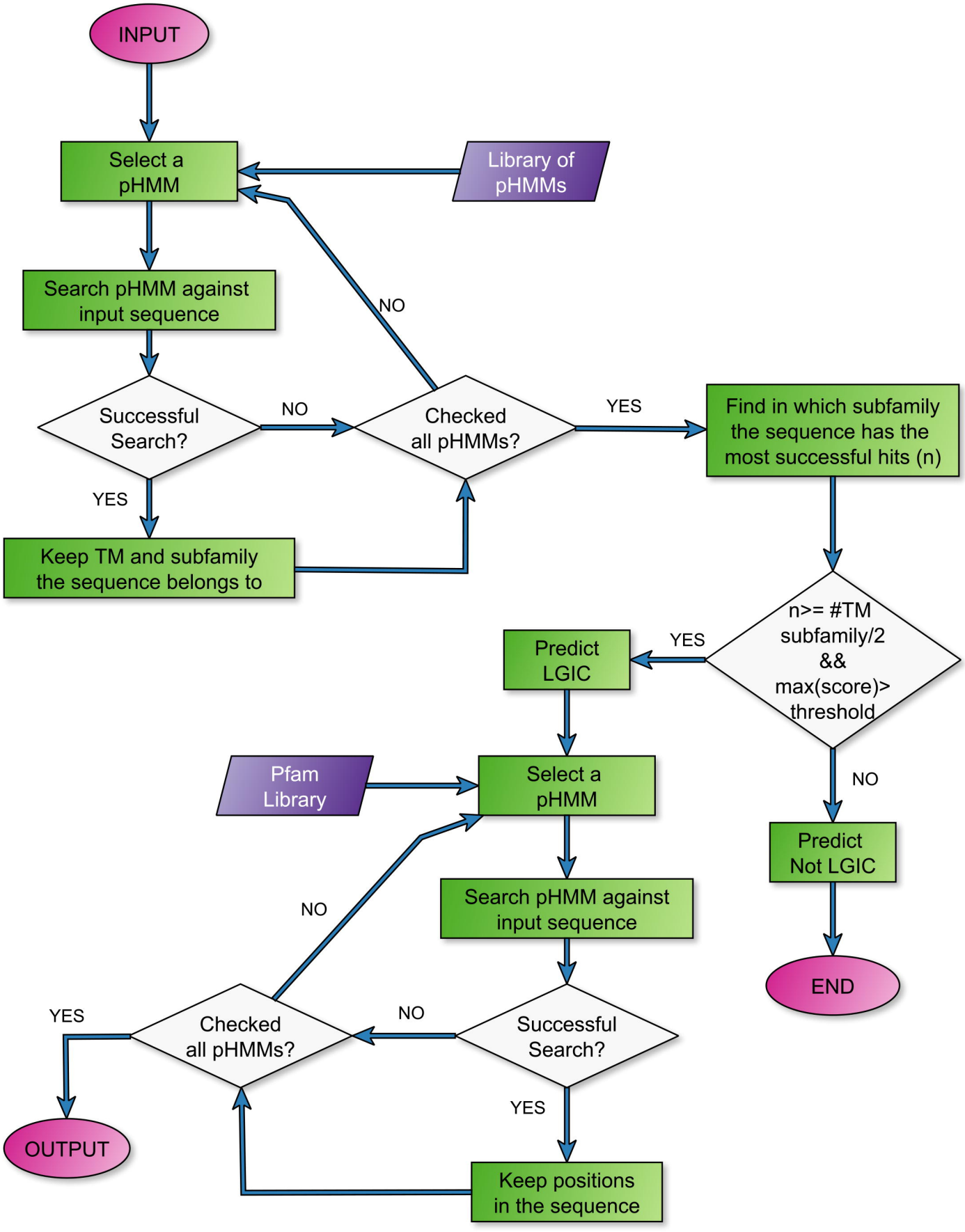
[33] Z.S. Ioannidou, M.C. Theodoropoulou, N.C. Papandreou, J.H. Willis, S.J. Hamodrakas, CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models, Insect Biochem Mol Biol, 52 (2014) 51-59.

[34] 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. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Res, 42 (2014) D222-230.

[35] 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.

[36] 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.

[37] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The Balanced Accuracy and Its Posterior Distribution, in: Proceedings of the 2010 20th International Conference on Pattern Recognition, IEEE Computer Society, 2010, pp. 3121-3124.


INPUT
Select a pHMM
Library of pHMMs
Search pHMM against input sequence
NO
Successful Search?
NO
Checked all pHMMs?
YES
Find in which subfamily the sequence has the most successful hits (n)
YES
Keep TM and subfamily the sequence belongs to
n>= #TM subfamily/2 && max(score)> threshold
YES
Predict LGIC
NO
Predict Not LGIC
END
Pfam Library
Select a pHMM
Search pHMM against input sequence
NO
YES
Checked all pHMMs?
NO
Successful Search?
YES
OUTPUT
Keep positions in the sequence

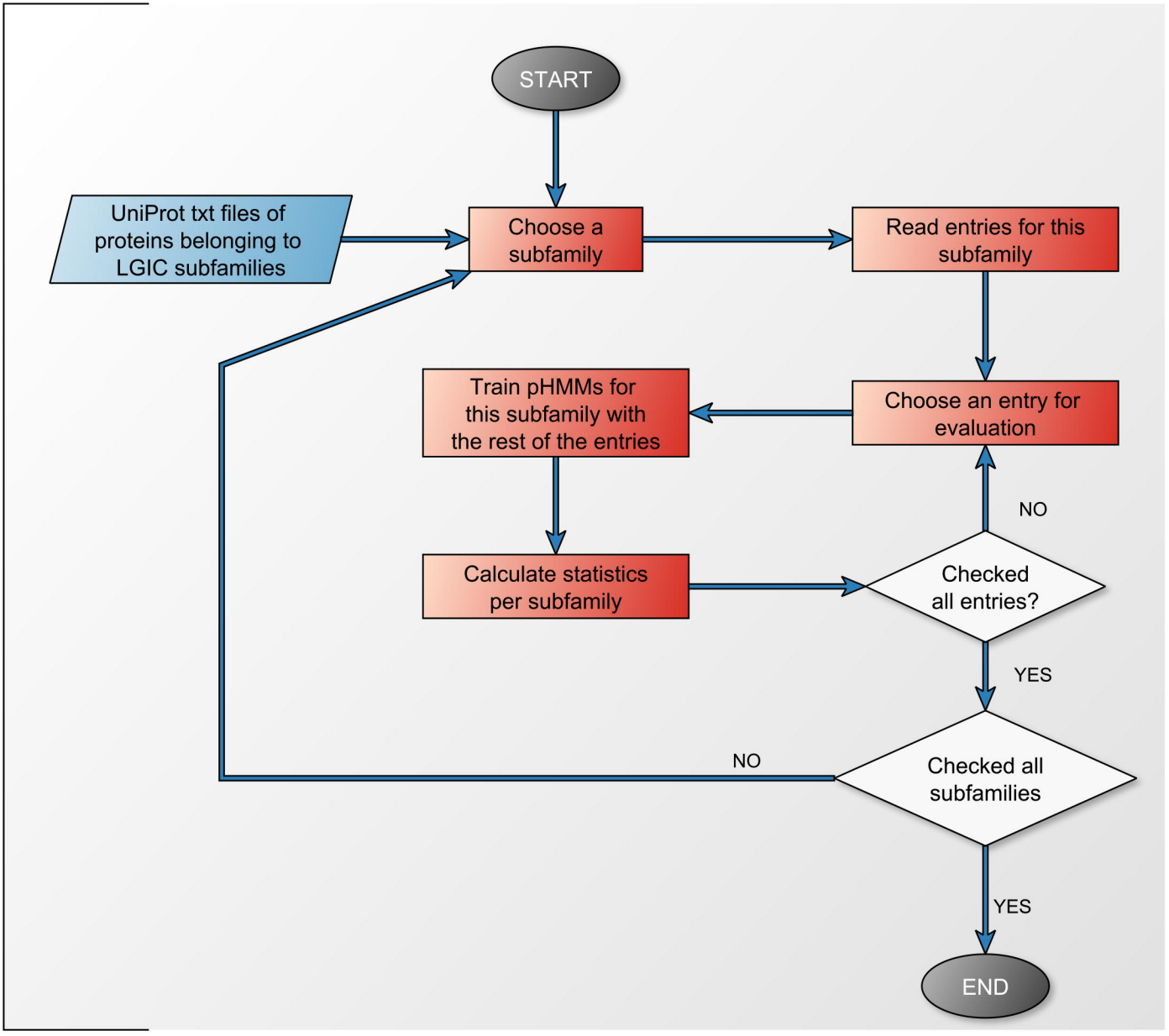
Name	LGIC	Family	Domains
sp|P63138|GBRB2_RAT	LGIC	GABAA	245-270,265-291,304-330,447-473

RESULTS TABLE

Full name			
sp|P63138|GBRB2_RAT			
Domains Name	**Domains Ranges**	**Symbol**	**SCORE**
GABAA_1	245-270	1	53.4
GABAA_2	265-291	2	55.1
GABAA_3	304-330	3	47.0
GABAA_4	447-473	4	46.4
Pfam Domains Name	**Pfam Domains Ranges**	**Pfam Symbol**	**Pfam Domains Score**
Neur_chan_memb	249-469	b	160.9
Neur_chan_LBD	38-241	a	169.7

SEQUENCE ANNOTATION

```
  1 MWRVRKRGYFGIWSFPLIIAAVCAQSVNDPSNMSLVKETVDRLLKGYDIRLRPDFGGPPV  60
HMMp ------------------------------------------------------------
Pfam ------------------------------------aaaaaaaaaaaaaaaaaaaaaa
 61 AVGMNIDIASIDMVSEVNMDYTLTMYFQQAWRDKRLSYNVIPLNLTLDNRVADQLWVPDT 120
HMMp ------------------------------------------------------------
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
121 YFLNDKKSFVHGVTVKNRMIRLHPDGTVLYGLRITTTAACMMDLRRYPLDEQNCTLEIES 180
HMMp ------------------------------------------------------------
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
181 YGYTTDDIEFYWRGDDNAVTGVTKIELPQFSIVDYKLITKKVVFSTGSYPRLSLSFKLKR 240
HMMp ------------------------------------------------------------
Pfam aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
241 NIGYFILQTYMPSILITILSWVSFWINYDASAARVALGITTVLTMTTINTHLRETLPKIP 300
HMMp ----1111111111111111111******2222222222222222222---------
Pfam a-------bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
301 YVKAIDMYLMGCFVFVFMALLEYALVNYIFFGRGPQRQKKAAEKAANANNEKMRLDVNKM 360
HMMp ---333333333333333333333333----------------------------
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
361 DPHENILLSTLEIKNEMATSEAVMGLGDPRSTMLAYDASSIQYRKAGLPRHSFGRNALER 420
HMMp ------------------------------------------------------------
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
421 HVAQKKSRLRRRASQLKITIPDLTDVNAIDRWSRIFFPVVFSFFNIVYWLYYVN  474
HMMp -------------------------444444444444444444444444-
Pfam bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb-----
```


START
UniProt txt files of proteins belonging to LGIC subfamilies
Choose a subfamily
Read entries for this subfamily
Choose an entry for evaluation
Train pHMMs for this subfamily with the rest of the entries
Calculate statistics per subfamily
Checked all entries?
NO
YES
Checked all subfamilies
NO
YES
END