Biochimica et Biophysica Acta 1864 (2016) 435–440

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

# GprotPRED: Annotation of Gα, Gβ and Gγ subunits of G-proteins using profile Hidden Markov Models (pHMMs) and application to proteomes

Vasiliki D. Kostiou [1,2], Margarita C. Theodoropoulou [2], Stavros J. Hamodrakas *

*Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 157 01, Greece*

ARTICLE INFO

*Article history:*
Received 21 October 2015
Received in revised form 20 January 2016
Accepted 3 February 2016
Available online 5 February 2016

*Keywords:*
Heterotrimeric G-proteins
Profile Hidden Markov Models (pHMMs)
Signal transduction
Proteome annotation

ABSTRACT

Heterotrimeric G-proteins form a major protein family, which participates in signal transduction. They are composed of three subunits, Gα, Gβ and Gγ. The Gα subunit is further divided in four distinct families $G_s$, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$. The goal of this work was to detect and classify members of the four distinct families, plus the Gβ and the Gγ subunits of G-proteins from sequence alone. To achieve this purpose, six specific profile Hidden Markov Models (pHMMs) were built and checked for their credibility. These models were then applied to ten (10) proteomes and were able to identify all known G-protein and classify them into the distinct families. In a separate case study, the models were applied to twenty seven (27) arthropod proteomes and were able to give more credible classification in proteins with uncertain annotation and in some cases to detect novel proteins. An online tool, GprotPRED, was developed that uses these six pHMMs. The sensitivity and specificity for all pHMMs were equal to 100% with the exception of the Gβ case, where sensitivity equals to 100%, while specificity is 99.993%. In contrast to Pfam's pHMM which detects Gα subunits in general, our method not only detects Gα subunits but also classifies them into the appropriate Gα-protein family and thus could become a useful tool for the annotation of G-proteins in newly discovered proteomes. GprotPRED online tool is publicly available for non-commercial use at http://bioinformatics.biol.uoa.gr/GprotPRED and, also, a standalone version of the tool at https://github.com/vkostiou/GprotPRED.



## 1. Introduction

Heterotrimeric G-proteins form a major protein family that is involved in signal transduction. They act as switches, triggering intracellular signalling mechanisms once the G-protein coupled receptors (GPCRs) are activated by a variety of extracellular stimuli. Heterotrimeric G-proteins consist of three subunits: Gα, Gβ and Gγ. Their nomenclature is determined by their α-subunit and they are classified in four families depending on the structural and functional similarity of their Gα subunits: $G_s$, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$. The key feature in their role as molecular switches is Gα subunit's ability to alternate between an inactive GDP-bound conformation and an active GTP-bound conformation [1]. In its inactive GDP-bound state, Gα subunit associates with the Gβγ heterodimer and the cytoplasmic tail and transmembrane loops of the receptor. When activated by a ligand, the receptor undergoes a conformational change which promotes the exchange of GDP for GTP, resulting in G-protein complex dissociation by the realignment of three flexible loops, named switches I, II and III. The activated Gα subunit and the free Gβγ heterodimer interact with downstream effectors, promoting cellular changes. The intrinsic GTPase activity of the Gα subunit hydrolyzes GTP to GDP which leads to the heterotrimer re-association and signaling termination [1–3].

The fact that G-proteins and more specifically Gα subunits interact with different proteins forced them to be highly constrained in order to preserve their functionality. Despite the large number of different interacting partners, heterotrimeric Gα subunits have diversified. Thus, heterotrimeric G-proteins constitute a highly conserved superfamily with some unique features among the distinct families [4]. It is obvious that signal transduction through heterotrimeric G-proteins is a mechanism of particular importance which controls the intracellular transfer of messages and ensures the proper function of organisms [1].

In mammalian systems, more than 20 Gα subunits have been described, belonging to the previously mentioned Gα families ($G_s$, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$) [1,5,6]. Additional G-proteins have been identified in many species based on sequence homology with the four distinct Gα families [6–8]. Moreover, several remotely related Gα genes which cannot be grouped in any of the four known families have been identified in invertebrates.

Studies on G-proteins from different families have been conducted in several invertebrate species and a considerable number of members

* Corresponding author at: Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, 157 01 Athens, Greece.
*E-mail address:* shamodr@biol.uoa.gr (S.J. Hamodrakas).
[1] Present address: Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK.
[2] Equally contributing authors.

http://dx.doi.org/10.1016/j.bbapap.2016.02.005
1570-9639/

of the known Gα families has been cloned [5,6,9,10]. All human Gα-subunit subgroups are represented in *Drosophila* [11,12]. An additional Gα subunit (Gαf) has been identified in Drosophila genome which suggests that it constitutes an insect-specific Gα subfamily [11,13]. In *Ceanorhabditis elegans* there are 21 Gα, 2 Gβ and 2 Gγ subunits [14–16]. There is one representative from each mammalian Gα family: GSA-1 ($G_s$), GOA-1 ($G_{i/o}$), EGL-30 ($G_q$) and GPA-12 ($G_{12}$) [14]. According to Bastiani *et al.*, the remaining *Ceanorhabditis elegans* Gα subunits (GPA-1, GPA-2, GPA-3, GPA-4, GPA-5, GPA-6, GPA-7, GPA-8, GPA-9, GPA-10, GPA-11, GPA-13, GPA-14, GPA-15, GPA-16, GPA-17 and ODR-3) are most similar to the G i/o family, but do not share sufficient homology to allow classification [14]. The *Dictyostelium discoideum* genome contains 8 Gα subunits: Gα-1, Gα-2, Gα-3, Gα-4, Gα-5, Gα-6, Gα-7 and Gα-8 [17,18], which share overall homology of 35–50%, compared to those from higher eukaryotes [17]. The Gα-1 and Gα-2 subunits, are almost identical within functional important regions, like GTPase activity and guanine nucleotide binding sites [17,19]. In *Arabidopsis* genome there are 1 canonical Gα (AtGPA1), 1 Gβ (AGB1) and 3 Gγ (AGG1, AGG2 and AGG3) genes encoded and they have roughly the same pattern for most diploid plants [20]. Several studies have shown that there is a possibility that signal transduction via G-proteins in plants can be performed in an alternative way. This is supported by the presence of unconventional plant specific Gα proteins, such as extra-large GTP-binding proteins (XLGs), composed of a C-terminal Gα-like domain and an N-terminal extension containing a nuclear localization signal and a cysteine-rich region [20,21]. Finally, *S. cerevisiae* genome contains 2 Gα-subunits (GPA1 and GPA2) [22,23].

Even though, heterotrimeric G-proteins constitute only a very small fragment of eukaryotic proteomes, their critical role in numerous signal transduction pathways makes their detection and efficient classification in newly identified proteomes very important. Pfam database [24] includes three pHMMs, which are commonly used for the detection of Gα (PF00503), Gβ (PF00400) and Gγ (PF00631) subunits of heterotrimeric G-proteins. Also, BLAST [25] is often used for the detection of G-proteins in proteomes using well annotated protein sequences. Neither of these two methods is fully automated nor can classify G-proteins into the three families. In 2008, a multi-modular SVM based method for the G-protein prediction, Vector-G, was introduced [26], but this method is no longer available and neither is any other method, apart from BLAST and the pHMMs offered in Pfam database. Here, we present the development of GprotPRED, a new, simple and fast method for the accurate detection and classification into families of G-proteins in newly identified proteomes, based on six especially designed G-protein specific profile Hidden Markov Models (pHMMs) [27].

## 2. Methods

The main features of our method, GprotPRED, are the six G-protein specific pHMMs. The Galpha (PF00503) profile of Pfam database [24] is able to detect α-subunits, but it may not classify them into the four known families. Therefore, we designed and built four distinct pHMMs, one for each known family of Gα-proteins. Regarding the two Pfam pHMMs for the detection of Gβ and Gγ subunits (PF00400 and PF00631 respectively), neither is exclusively specific and, as a result, two additional pHMMs were designed and built, one for each subunit.

### 2.1. Data collection

Initially, we collected all G-protein sequences of Gα, Gβ and Gγ subunits from the UniProt/Swiss-Prot database release 2010_09 [28]. 190 Gα subunits were retrieved, from which, 112 are classified into one of the four known heterotrimeric Gα-protein families (23 $G_s$, 55 $G_{i/o}$, 27 $G_{q/11}$, 7 $G_{12/13}$) while the remaining 78 are unclassified (i.e. they don't belong to any of the four known families), based on the annotation of the UniProt database (more specifically in the description field (DE)). Also, 77 Gβ and 59 Gγ subunits were retrieved (Table S1 of Supplementary File 1).

### 2.2. Selection and preparation of training sets

In order to build more accurate and specific pHMMs we used both positive and negative training sets (HMM-ModE) [29]. Each pHMM was created using the multiple alignments of the sequences that belong to the specific family (positive training set) and the sequences that display high similarity, but do not belong to this particular family (negative training set). For the Gα families, the positive training set contains only one representative from each organism and each subfamily, in order for all positive training sets to be as non-redundant and as balanced as possible. Apart from the $G_{i/o}$ family, all available sequences from each family were included in the positive training set. Using the positive and negative training sequences for each pHMM, we implemented multiple sequence alignments using ClustalW [30] which were then used as input in the *hmmbuild* program of the HMMER v2.3.2 package [27]. Our pHMMs were then modified by the HMM-ModE protocol [29], which has the ability to maximize the contributions of discriminating residues. After the build process, the six pHMMs were converted to HMMER v3.0 format using the *hmmconvert* program [27].

More specifically, each pHMM was created as follows:

1. $G_s$ family: This model was constructed from a positive protein multiple alignment set (23 sequences) that belong to this family and from a negative protein set (89 sequences) that belong to the other three families ($G_{i/o}$, $G_{q/11}$, $G_{12/13}$).
2. $G_{i/o}$ family: Since $G_{i/o}$ family is the most abundant one, only one representative from each organism and each subfamily was included in the positive training set, in order for all positive training sets to be as non-redundant and as balanced as possible. This model was constructed from a positive protein set multiple alignment (41 sequences) that belong to this family and from a negative protein set (57 sequences) that belong to the other three families ($G_s$, $G_{q/11}$, $G_{12/13}$).
3. $G_{q/11}$ family: This model was constructed from a positive protein multiple alignment set (27 sequences) that belong to this family and from a negative protein set (85 sequences) that belong to the other three families ($G_{i/o}$, $G_s$, $G_{12/13}$).
4. $G_{12/13}$ family: This model was constructed from a positive protein multiple alignment set (7 sequences) that belong to this family and from a negative protein set (105 sequences) that belong to the other three families ($G_{i/o}$, $G_{q/11}$, $G_s$).
5. Gβ subunit: This model was constructed to model the Gβ subunit in its full length, unlike the Pfam model (PF00400, name WD40) [24] that describes only the WD40 domain of this subunit. It was constructed from a positive protein multiple alignment set (50 sequences), with one representative from each organism and at least one representative from each type, and from a negative protein set (89 sequences). The negative training set was derived as follows:
   i. We ran the general Pfam model (PF00400) against Uniprot/SwissProt database. The program returned 2195 sequences.
   ii. Then, using a Perl Script we isolated sequences that had 7 WD40 repeats. The set were reduced to 582 sequences.
   iii. From those 582 sequences, we removed the Gβ subunits. The number of the remaining sequences was 505.
   iv. With the use of the CD-HIT web server [31], we derived a non-redundant set of 89 sequences. CD-HIT is a widely used program for clustering and comparing protein or nucleotide sequences. The 505 redundant sequences were clustered according to their sequence similarity by the CD-HIT program and each cluster's representative sequence was then included in the negative training set that was used to train the Gβ pHMM.

6. Gγ subunit: The same procedure was followed for the Gγ subunit. The existing Pfam model (PF00631) [24] describes a domain (GGL domain) that exists not only in Gγ subunits but also in Regulators of G-proteins (RGS). Our pHMM was constructed from a positive protein multiple alignment set (26 sequences), with one representative from each organism and at least one representative from each type, and a negative protein set (14 sequences). To create the negative training set we ran the general Pfam (PF00631) model against Uniprot/SwissProt database. From the 71 resulting sequences, we removed the Gγ subunits and the remaining 14 sequences were the final negative training set.
   All proteins included in the positive and negative training sets are available in Table S1 of Supplementary File 1.

### 2.3. Evaluation method for the models

The probability parameters in a profile HMM are converted to additive log–odds scores before aligning and scoring a query sequence [32]. The scores for aligning a residue to a profile match state are, therefore, comparable to the derivation of BLAST or FASTA scores [27].

Each pHMM was applied against the UniProt/SwissProt database (Release 2010_09), using the *hmmsearch* program of HMMER v3.0 <http://hmmer.janelia.org/> [27]. As proposed by Ioannidou et al. [33] in order to estimate the cutoff score for each model, the standard statistical measures for the performance of binary (a protein either belongs to a family or not) classification tests, specificity and sensitivity, were calculated for a range of 50 units of cutoff score for each model: Specificity = TN/(TN + FP) and Sensitivity = TP/(TP + FN), where TP is the number of True Positive predictive values, TN the number of True Negatives, FP the number of False Positives and FN the number of False Negatives. Then, a plot of specificity and sensitivity against the different scores was designed for each pHMM. Specificity curves were plotted against the different cutoffs to identify the cutoff range where sensitivity and specificity meet (Fig. 1).The cutoff score for each model was estimated as the middle value of the range where specificity meets sensitivity. In all cases using the corresponding cutoff score both sensitivity and specificity were equal to 1. Intuitively, this was chosen so that the cutoff score will represent the largest separation between protein sequences that belong to the family described by the model, and those that do not. We define the final cutoff score as:

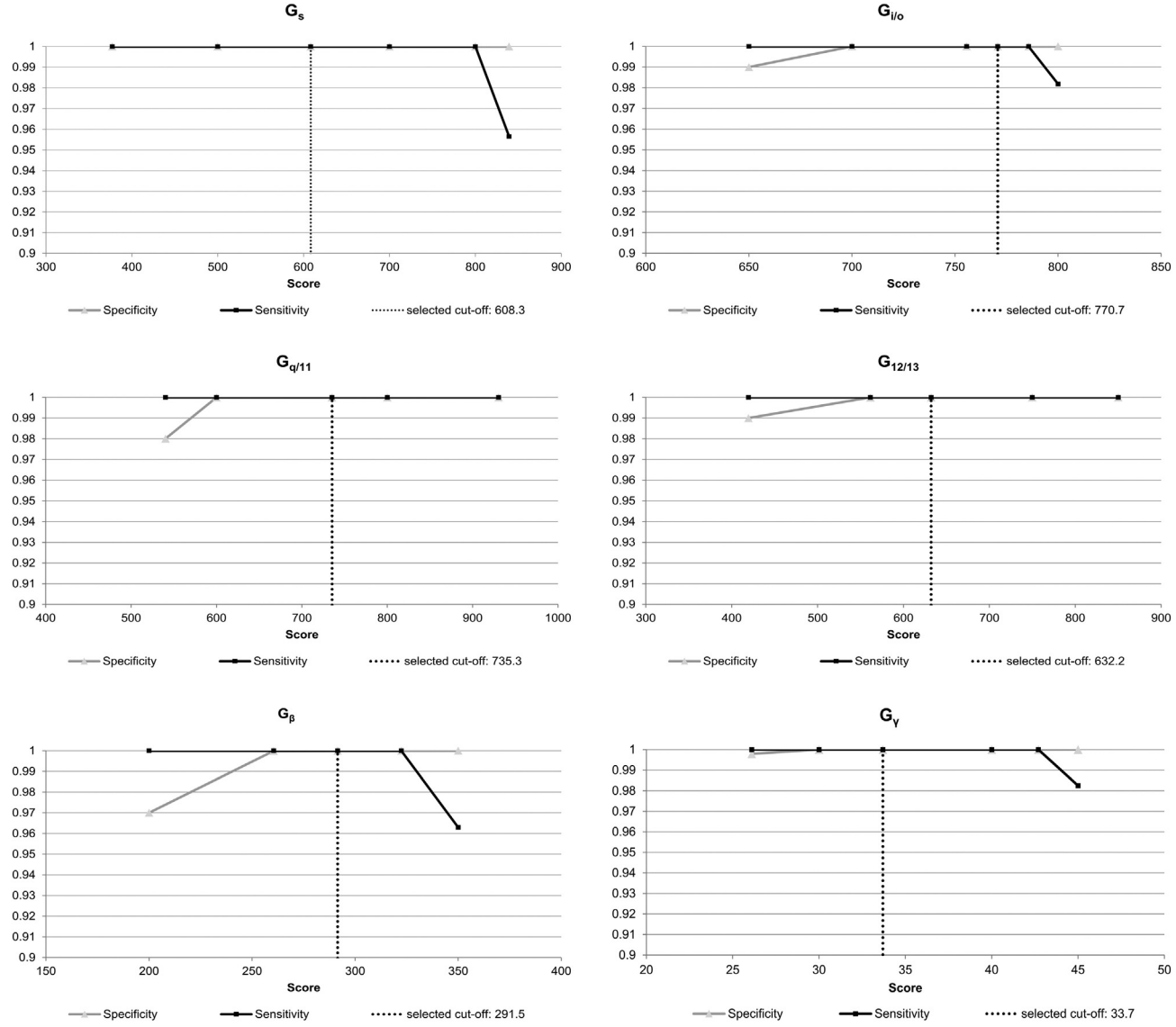


**Fig. 1.** Plots of sensitivity and specificity against different cutoff scores and the selected thresholds for each profile HMM. These values clearly separate the positive from the negative results.

(TC + NC)/2, where Trusted cutoff (TC) is the score of the lowest scoring true positive hit and Noise cutoff (NC) is the score of the highest scoring true negative hit [33].

With the cut off scores defined, all six models were used against UniProt/SwissProt database (Release 2014_11) [34], and a comparison with the three pre-existing Pfam's pHMMs was made.

### 2.4. Application of the pHMMs to proteomes

After the evaluation, our models plus the Pfam model for the Gα subunit (PF00503) [24] were used to detect potential G-proteins within ten (10) different proteomes. The organisms were 2 chordates (*Homo sapiens*, *Mus musculus*), 1 arthropod (*Drosophila melanogaster*), 1 nematode (*Caenorhabditis elegans*), 2 fungi (*Aspergillus niger*, *Saccharomyces cerevisiae*), 2 plants (*Arabidopsis thaliana*, *Selaginella moellendorffii*), 1 amoeboid protist (*Dictyostelium discoideum*) and 1 protozoan (*Trichomonas vaginalis*). The models were also applied to twenty seven (27) model organisms proteomes belonging to different orders of the Arthropod phylum (*Acromyrmex echinatior*, *Acyrthosiphon pisum*, *Aedes aegypti*, *Anopheles darlingi*, *Anopheles gambiae*, *Apis mellifera*, *Atta cephalotes*, *Bombyx mori*, *Camponotus floridanus*, *Culex quinquefasciatus*, *Danaus plexippus*, *Drosophila ananassae*, *Drosophila grimshawi*, *Drosophila mojavensis*, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila sechellia*, *Drosophila simulans*, *Drosophila virilis*, *Drosophila willistoni*, *Harpegnathos saltator*, *Megaselia scalaris*, *Nasonia vitripennis*, *Pediculus humanus*, *Rhodnius prolixus*, *Solenopsis invicta*, *Tribolium castaneum*) as a separate case study. All reference proteomes were retrieved from the UniProt database (Release 2014_11) [34].

### 2.5. Website implementation

The web page was implemented using the following technologies: the HTML markup language and the CSS style sheet language for the page layout and design, the Perl scripting language (CGI) for the server side functions and processing of the results and finally the HMMER v3.0 suite <http://hmmer.janelia.org/> which runs the searches on the server.

## 3. Results and discussion

### 3.1. Evaluation

#### 3.1.1. Estimation of cutoffs

The cutoff score for each model was estimated as the middle value of the range where specificity meets sensitivity. This was preferred based on the hypothesis that with larger separation between protein sequences that belong to the family type described by the model and protein sequences that do not, it is more likely to avoid misclassifications. The scores of true positives and false negatives of each pHMM did not overlap. In all cases apart from the Gβ profile, both specificity and sensitivity were equal to 1, in the score that was assigned as a cutoff, as shown in Fig. 1. In the Gβ case, sensitivity was equal to 1, while specificity was 0.99993. All cutoffs are listed in Table 1.

**Table 1**
Cutoff scores.

| | Galpha (PF00503) | $G_s$ | $G_{i/o}$ | $G_{q/11}$ | $G_{12/13}$ | Gβ | Gγ |
|---|---|---|---|---|---|---|---|
| Cutoff | 48.8 | 608.3 | 770.7 | 735.3 | 632.2 | 291.5 | 33.7 |

**Table 2**
Evaluation against Uniprot/Swissprot (Release 2014_11, 547,085 sequences) and comparison with the pre-existing Pfam's pHMMs.

| pHMM | Total sequences detected | Sequences belonging to family | Sequences not belonging to family |
|---|---|---|---|
| Galpha (PF00503) | 200 | 195 | 5 |
| $G_s$ | 23 | 23 | - |
| $G_{i/o}$ | 56 | 56 | - |
| $G_{q/11}$ | 27 | 27 | - |
| $G_{12/13}$ | 9 | 9 | - |
| WD40 (PF00400) | 1795 | 85 | 1710 |
| Gβ | 85 | 81 | 4 |
| G-gamma (PF00631) | 72 | 58 | 14 |
| Gγ | 58 | 58 | – |

#### 3.1.2. Evaluation against UniProt/SwissProt and comparison with the pre-existing Pfam's pHMMs

As shown in Table 2 the pre-existing Pfam's pHMMs are not specific enough in order to be used for G-protein detection and classification. PF00503 can be used for G-alpha subunit detection but it's not family specific. PF00400 and PF00631 describe domains that exist not only in Gβ and Gγ subunits but in other proteins as well. On the contrary, our Gβ and Gγ profiles have been trained to detect Gβ and Gγ subunits exclusively.

### 3.2. Application to proteomes

#### 3.2.1. Application to 10 proteomes

The six pHMMs plus the Galpha pHMM of Pfam database (PF00503) were applied to 10 available proteomes. A few protein sequences, which either had not been annotated (proteins with unknown function) or were not classified into a specific family, were found by the models. We were able to detect representatives for all Gα families and all Gβ and Gγ subunits in *H.sapiens*, *M. musculus*, *D. melanogaster* and *C. elegans* proteomes, as perhaps expected, showing that all components of G-protein mediated signal transduction are present in metazoans (Fig. 2). However, we were not able to classify the Gα subunits detected in *A. niger*, *S. cerevisiae*, *A. thaliana*, *S. moellendorffii*, *D. discoideum* and *T. vaginalis* proteomes in any of the four known Gα families. The number of the detected Gα subunits was significantly lower in *A. niger*, *S. cerevisiae*, *A. thaliana*, *S. moellendorffii*, *D. discoideum* and *T. vaginalis* proteomes, supporting the fact that in striking contrast to metazoans their repertoire is simpler. The results are summarized in Table 3. The result files produced by GprotPRED for all proteomes are available on <http://aias.biol.uoa.gr/GprotPRED/Proteomes_Results/>.

#### 3.2.2. Case study: application to 27 arthropod proteomes

The six pHMMs plus the Galpha pHMM of Pfam database (PF00503) [24] were applied to 27 available arthropod proteomes: 8 Hymenoptera, 14 Diptera, 2 Hemiptera, 2 Lepidoptera, 1 Phthiraptera and 1 Coleoptera. The 27 arthropod proteomes contain several representatives of almost all four Gα families and all Gβ and Gγ subunits, supporting the fact that G-protein mediated signal transduction pathway is present in insect cells (Fig. S1). The number of the detected Gα subunits was the highest in all proteomes, followed by the Gβ and Gγ respectively (Fig. S1, Fig. S2). $G_{i/o}$ and $G_{q/11}$ are the most highly represented families in all of the 27 arthropod proteomes apart from Lepidoptera where no $G_{i/o}$ subunits were detected and Phthiraptera where no $G_{q/11}$ subunits were detected. In Lepidoptera, a large number of unclassified Gα subunits (i.e. they don't belong to any of the four known families) was observed (57.14% of the total number of detected Gα subunits) whereas only 10% of the total number of Gα subunits was unclassified in Hemiptera (Fig. S2). The results are summarized in Table 4. The result

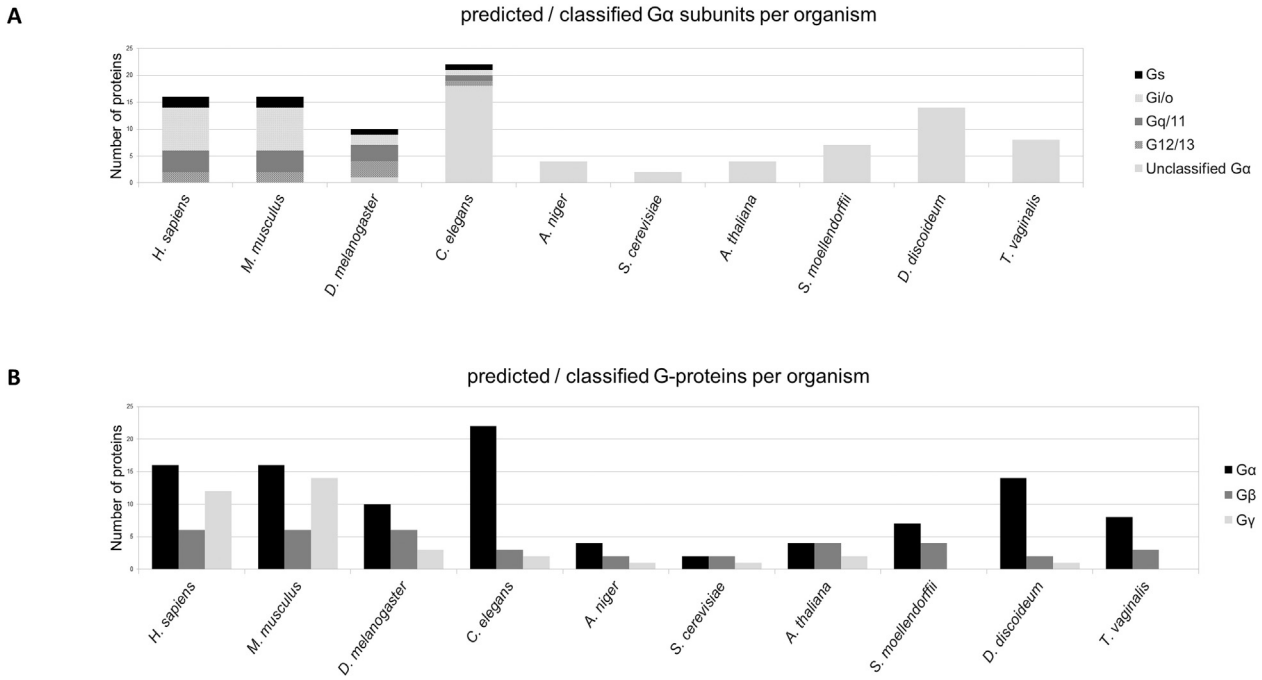


**Fig. 2.** A. Classification of the total number of detected Gα subunits into the four families in each organism. Proteins that do not belong to any of the four families are considered unclassified. B. Number of Gα, Gβ and Gγ subunits detected in each of the ten proteomes.

files produced by GprotPRED for all proteomes are available on <http://aias.biol.uoa.gr/GprotPRED/Proteomes_Results/>.

In general, the six pHMMs can detect with high accuracy the G-proteins in all the above proteomes and our results are in agreement with the proteomes' annotation when available. On the contrary, the Pfam pHMM for the Gα subunit (PF00503) tends to overpredict. Supplementary files 2 and 3 include an extensive list with the accession numbers of all predicted proteins and the comparison with the annotation provided in UniProt. After an extensive study of the annotation in UniProt, it is important to emphasize that regarding the unreviewed entries of the database one must be extremely cautious, since, to our experience, the annotation may be misleading. It is worth taking into consideration that the tool's performance could be improved as more, well annotated proteomes become available.

### 3.3. GprotPRED online tool

GprotPRED, available on <http://bioinformatics.biol.uoa.gr/GprotPRED/>, is an online tool with a user friendly interface that utilizes our profile Hidden Markov Models (pHMMs) for the four known heterotrimeric Gα protein families, plus the Gβ and the Gγ subunit in order to classify a set of query sequences into the appropriate G-protein family. The user may insert one or more protein sequences or a whole proteome in FASTA format. The results are presented to the user in the form of a table and can also be accessed through a text format file.

A standalone version of the tool for off line use is available on https://github.com/vkostiou/GprotPRED.

**Table 3**
Detection and classification of G-proteins within the proteomes of ten organisms using our pHMMs plus the PF00503 pHMM.

| Proteome | Size | Galpha PF00503 | $G_s$ | $G_{i/o}$ | $G_{q/11}$ | $G_{12/13}$ | Gβ | Gγ |
|---|---|---|---|---|---|---|---|---|
| *H. sapiens* | 20,861 | 16 | 2 | 8 | 4 | 2 | 6 | 12 |
| *M. musculus* | 22,136 | 16 | 2 | 8 | 4 | 2 | 6 | 14 |
| *D. melanogaster* | 19,447 | 10 | 1 | 2 | 3 | 3 | 6 | 3 |
| *C. elegans* | 20,275 | 22 | 1 | 1 | 1 | 1 | 3 | 2 |
| *A. niger* | 10,944 | 4 | – | – | – | – | 2 | 1 |
| *S. cerevisiae* | 6718 | 2 | – | – | – | – | 2 | 1 |
| *A. thaliana* | 27,222 | 4 | – | – | – | – | 4 | 2 |
| *S. moellendorffii* | 33,112 | 7 | – | – | – | – | 4 | – |
| *D. discoideum* | 12,732 | 14 | – | – | – | – | 2 | 1 |
| *T. vaginalis* | 50,189 | 8 | – | – | – | – | 3 | – |

**Table 4**
Detection and classification of G-proteins within the proteomes of twenty seven arthropod organisms using our pHMMs plus the PF00503 pHMM.

| Proteome | Size | Galpha PF00503 | $G_s$ | $G_{i/o}$ | $G_{q/11}$ | $G_{12/13}$ | Gβ | Gγ |
|---|---|---|---|---|---|---|---|---|
| *A. echinator* | 13,962 | 6 | 1 | 2 | 1 | 1 | 4 | 1 |
| *A. pisum* | 35,809 | 5 | 1 | 2 | 1 | 1 | 4 | 2 |
| *A. aegypti* | 16,554 | 9 | 1 | 2 | 2 | 1 | 6 | 2 |
| *A. darlingi* | 10,453 | 7 | 1 | 1 | 1 | 1 | 5 | 2 |
| *A. gambiae* | 13,072 | 14 | 1 | 3 | 8 | 1 | 6 | 2 |
| *A. melifera* | 10,910 | 7 | 1 | 1 | 1 | 1 | 4 | 2 |
| *A. cephalotes* | 18,079 | 6 | – | 1 | – | 1 | 4 | 2 |
| *B. mori* | 14,767 | 8 | 2 | – | 1 | – | 4 | – |
| *C. floridanus* | 14,787 | 5 | – | 2 | 1 | 1 | 4 | 2 |
| *C. quinquefasciatus* | 18,703 | 8 | 1 | 2 | 1 | 1 | 11 | 2 |
| *D. plexippus* | 16,253 | 6 | 1 | – | 1 | 1 | 4 | 2 |
| *D. ananassae* | 14,968 | 6 | 1 | 2 | 1 | 1 | 4 | 3 |
| *D. grimshawi* | 14,754 | 7 | 1 | 2 | 2 | 1 | 6 | 3 |
| *D. mojavensis* | 14,525 | 7 | 1 | 2 | 2 | 1 | 5 | 3 |
| *D. persimilis* | 16,754 | 7 | 1 | 1 | 1 | 1 | 4 | 3 |
| *D. pseudoobscura* | 16,756 | 8 | 1 | 2 | 3 | 1 | 4 | 3 |
| *D. sechelia* | 16,134 | 8 | 1 | 2 | 3 | 1 | 4 | 3 |
| *D. simulans* | 15,354 | 8 | 1 | 2 | 4 | 1 | 4 | 3 |
| *D. virilis* | 14,456 | 7 | 1 | 2 | 2 | 1 | 5 | 3 |
| *D. willistoni* | 15,447 | 6 | 1 | 2 | 1 | 1 | 4 | 3 |
| *H. saltator* | 15,029 | 6 | – | 2 | 1 | 1 | 4 | 3 |
| *M. scalaris* | 11,463 | 5 | – | – | – | – | 2 | 1 |
| *N. vitripennis* | 17,040 | 5 | 1 | 1 | 1 | 1 | 4 | 2 |
| *P. humanus* | 10,763 | 6 | 1 | 2 | – | 1 | 6 | 1 |
| *R. prolixus* | 15,181 | 5 | 1 | 1 | 1 | 1 | 3 | 1 |
| *S. invicta* | 14,193 | 5 | – | 2 | 1 | – | 4 | 1 |
| *T. castaneum* | 16,502 | 7 | 1 | 2 | 1 | – | 4 | 2 |

## 4. Conclusions

In this paper we introduce GprotPRED, an on-line tool for the detection and classification of G-proteins, from sequence alone. We hope that implementation of these pHMMs, will be useful in the functional annotation of newly discovered proteomes.

### 4.1. Availability and requirements

The GprotPRED is freely available at <http://bioinformatics.biol.uoa.gr/GprotPRED/>. The website has been tested with Internet Explorer, Firefox, Chrome, Opera and Safari browsers. A standalone version of the tool for off line use is available on https://github.com/vkostiou/GprotPRED. It is free of charge for non-commercial use.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbapap.2016.02.005.

## Conflict of interest

The authors declare no conflicts of interest.

## Transparency document

The transparency document associated with this article can be found, in online version.

## Acknowledgements

This research has been co-financed by the European Union (European Regional Development Fund — ERDF) and Greek national funds through the Operational Program "Competitiveness and Entrepreneurship" of the National Strategic Reference Framework (NSRF) (Project code 09SYN-13-999). Finally, we should like to thank the handling editor and the reviewers of this manuscript for their very useful and constructive criticism.

## References

[1] W.M. Oldham, H.E. Hamm, Heterotrimeric G protein activation by G-protein-coupled receptors, Nat. Rev. Mol. Cell Biol. 9 (2008) 60–71.
[2] B.R. Temple, A.M. Jones, The plant heterotrimeric G–protein complex, Annu. Rev. Plant Biol. 58 (2007) 249–266.
[3] A. de Mendoza, A. Sebe-Pedros, I. Ruiz-Trillo, The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity, Genome Biol. Evol. 6 (2014) 606–619.
[4] B.R. Temple, C.D. Jones, A.M. Jones, Evolution of a signaling nexus constrained by protein interfaces and conformational states, PLoS Comput. Biol. 6 (2010), e1000962.
[5] P.J. Knight, T.A. Grigliatti, Diversity of G proteins in Lepidopteran cell lines: partial sequences of six G protein alpha subunits, Arch. Insect Biochem. Physiol. 57 (2004) 142–150.
[6] G.J. Kang, Z.J. Gong, J.A. Cheng, Z.R. Zhu, C.G. Mao, Cloning and expression analysis of a G-protein alpha subunit—Galphao in the rice water weevil *Lissorhoptrus oryzophilus* Kuschel, Arch. Insect Biochem. Physiol. 76 (2011) 43–54.
[7] M. Jiang, N.S. Bajpayee, Molecular mechanisms of go signaling, Neurosignals 17 (2009) 23–41.
[8] Y. Oka, S.I. Korsching, The fifth element in animal Galpha protein evolution, Commun. Integr. Biol. 2 (2009) 227–229.
[9] E. Jacquin-Joly, M.C. Francois, M. Burnet, P. Lucas, F. Bourrat, R. Maida, Expression pattern in the antennae of a newly isolated lepidopteran Gq protein alpha subunit cDNA, Eur. J. Biochem. 269 (2002) 2133–2142.
[10] M. Rutzler, T. Lu, L.J. Zwiebel, Galpha encoding gene family of the malaria vector mosquito *Anopheles gambiae*: expression analysis and immunolocalization of AGalphaq and AGalphao in female antennae, J. Comp. Neurol. 499 (2006) 533–545.
[11] N. Katanayeva, D. Kopein, R. Portmann, D. Hess, V.L. Katanaev, Competing activities of heterotrimeric G proteins in *Drosophila* wing maturation, PLoS One 5 (2010), e12331.
[12] T. Boto, C. Gomez-Diaz, E. Alcorta, Expression analysis of the 3 G-protein subunits, Galpha, Gbeta, and Ggamma, in the olfactory receptor organs of adult *Drosophila melanogaster*, Chem. Senses 35 (2010) 183–193.
[13] F. Quan, W.J. Wolfgang, M. Forte, A *Drosophila* G-protein alpha subunit, Gf alpha, expressed in a spatially and temporally restricted pattern during *Drosophila* development, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 4236–4240.
[14] C. Bastiani, J. Mendel, Heterotrimeric G proteins in *C. elegans*, WormBook: The Online Review of *C. elegans* Biology 2006, pp. 1–25.
[15] E. Cuppen, A.M. van der Linden, G. Jansen, R.H. Plasterk, Proteins interacting with *Caenorhabditis elegans* Galpha subunits, Comp. Funct. Genomics 4 (2003) 479–491.
[16] G. Jansen, K.L. Thijssen, P. Werner, M. van der Horst, E. Hazendonk, R.H. Plasterk, The complete family of genes encoding G proteins of *Caenorhabditis elegans*, Nat. Genet. 21 (1999) 414–419.
[17] D.C. New, J.T. Wong, The evidence for G-protein-coupled receptors and heterotrimeric G proteins in protozoa and ancestral metazoa, Biol. Signals Recept. 7 (1998) 98–108.
[18] L.J. Wu, P.N. Devreotes, Dictyostelium transiently expresses eight distinct G-protein alpha-subunits during its developmental program, Biochem. Biophys. Res. Commun. 179 (1991) 1141–1147.
[19] A. Kumagai, J.A. Hadwiger, M. Pupillo, R.A. Firtel, Molecular genetic analysis of two G alpha protein subunits in *Dictyostelium*, J. Biol. Chem. 266 (1991) 1220–1228.
[20] D. Urano, J.G. Chen, J.R. Botella, A.M. Jones, Heterotrimeric G protein signalling in the plant kingdom, Open Biol. 3 (2013) 120186.
[21] H. Zhu, G.J. Li, L. Ding, X. Cui, H. Berg, S.M. Assmann, Y. Xia, Arabidopsis extra large G-protein 2 (XLG2) interacts with the Gbeta subunit of heterotrimeric G protein and functions in disease resistance, Mol. Plant 2 (2009) 513–525.
[22] T. Harashima, J. Heitman, The Galpha protein Gpa2 controls yeast differentiation by interacting with kelch repeat proteins that mimic Gbeta subunits, Mol. Cell 10 (2002) 163–173.
[23] H.F. Stratton, J. Zhou, S.I. Reed, D.E. Stone, The mating-specific G(alpha) protein of *Saccharomyces cerevisiae* downregulates the mating signal by a mechanism that is dependent on pheromone and independent of G(beta)(gamma) sequestration, Mol. Cell. Biol. 16 (1996) 6325–6337.
[24] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families database, Nucleic Acids Res. 40 (2012) D290–D301.
[25] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[26] P. Jain, P. Wadhwa, R. Aygun, G. Podila, Vector-G: multi-modular SVM-based heterotrimeric G protein prediction, In Silico Biol. 8 (2008) 141–155.
[27] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[28] C. UniProt, The Universal Protein Resource (UniProt) in 2010, Nucleic Acids Res. 38 (2010) D142–D148.
[29] P.K. Srivastava, D.K. Desai, S. Nandi, A.M. Lynn, HMM-ModE—improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences, BMC Bioinf. 8 (2007) 104.
[30] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.
[31] 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.
[32] C. Barrett, R. Hughey, K. Karplus, Scoring hidden Markov models, Comput. Appl. Biosci. 13 (1997) 191–199.
[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] C. UniProt, Activities at the Universal Protein Resource (UniProt), Nucleic Acids Res. 42 (2014) D191–D198.