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/© 2016 Elsevier B.V. All rights reserved.

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.