Biochimica et Biophysica Acta 1824 (2012) 488–492

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Identification and classification of conopeptides using profile Hidden Markov Models

Silja Laht [a,*], Dominique Koua [b,c], Lauris Kaplinski [a], Frédérique Lisacek [c], Reto Stöcklin [b], Maido Remm [a]

[a] Estonian Biocentre, Riia 23, 51010, Tartu, Estonia
[b] Atheris Laboratories, Case Postale 314, CH-1233 Bernex-Geneva, Switzerland
[c] Swiss Institute of Bioinformatics, Proteome Informatics Group, Rue Michel-Servet 1, 1211, Geneva, Switzerland

ARTICLE INFO

Article history:
Received 15 June 2011
Received in revised form 13 December 2011
Accepted 19 December 2011
Available online 30 December 2011

Keywords:
Conotoxin
Conopeptide
Hidden Markov Model
Conopeptide superfamilies
Protein prediction

ABSTRACT

Conopeptides are small toxins produced by predatory marine snails of the genus *Conus*. They are studied with increasing intensity due to their potential in neurosciences and pharmacology. The number of existing conopeptides is estimated to be 1 million, but only about 1000 have been described to date. Thanks to new high-throughput sequencing technologies the number of known conopeptides is likely to increase exponentially in the near future. There is therefore a need for a fast and accurate computational method for identification and classification of the novel conopeptides in large data sets. 62 profile Hidden Markov Models (pHMMs) were built for prediction and classification of all described conopeptide superfamilies and families, based on the different parts of the corresponding protein sequences. These models showed very high specificity in detection of new peptides. 56 out of 62 models do not give a single false positive in a test with the entire UniProtKB/Swiss-Prot protein sequence database. Our study demonstrates the usefulness of mature peptide models for automatic classification with accuracy of 96% for the mature peptide models and 100% for the pro- and signal peptide models. Our conopeptide profile HMMs can be used for finding and annotation of new conopeptides from large datasets generated by transcriptome or genome sequencing. To our knowledge this is the first time this kind of computational method has been applied to predict all known conopeptide superfamilies and some conopeptide families.

1. Introduction

Conopeptides are small, usually cysteine-rich, peptides that are found in the venom of the marine snails from genus *Conus*. Cone snails are predator mollusks, hunting for either worms, snails or fish, with a few species being harmful to humans. Conopeptides are used as valuable probes in neurophysiological studies due to their exceptional specificity for different isoforms of ion channels, receptors and transporters [1] and provide lead compounds for drug discovery [2,3].

Each conopeptide precursor (with a few exceptions) consists of three parts: a signal peptide at the N-terminus (typically 20–25 amino acids in length), a pro-peptide (for most conopeptides 30–60 amino acids in length) and a mature peptide at the C-terminus (8 to>40 amino acids, usually 12–30 amino acids) (Fig. 1). During maturation in the venom gland, the signal peptides and the pro-peptides are cleaved, correct disulphide crosslinks are formed and often some amino acids are modified. The mature peptides act as toxins when they are injected into a prey [4].

Conopeptides have been classified into 16 superfamilies defined by a common signal sequence (one will refer to the A, D, M, O or T superfamily for example). A superfamily does not reflect the biological activity. Indeed, a single mutation in a mature sequence can drastically change its pharmacological properties and members of different superfamilies can reveal similar biological activities [5]. Additionally, some 30 conopeptide families have been described based on a typical structural pattern (such as the cysteine motif) coupled to a given biological activity on a specific subtype of ion channel or receptor. For example, one will refer to the alpha-A, delta, mu, mu-O, conantokin or conopressin family. In this paper superfamilies and families together will be referred to as classes.

The most recent studies have estimated that the number of different conopeptides detected in the venom of a single species can exceed 1,000 [6,7]. The number of *Conus* species is currently estimated to reach 800, thus suggesting a putative natural library in the range of 1 million biomolecules, mostly bioactive peptides and mini-proteins. However, despite of such a huge molecular diversity, only approximately 1000 conopeptides have been described so far [8,9]. Taking into account the huge potential that remains to be offered by conopeptides and given the fact that novel sequencing techniques provide a vast amount of sequence data, there is a need for an automated process for identification and annotation of new conopeptide sequences from large datasets. For this purpose we have selected profile Hidden Markov Models (pHMMs) built from pre-existing data. Hidden Markov Models (HMMs) are a class of probabilistic models that are generally applicable to time series or linear sequences. Profile HMMs (pHMMs) are a type of HMMs that are designed to represent profiles of multiple sequence alignments [10]. pHMMs are widely used to predict and find members

* Corresponding author. Tel.: +372 737 5001, +372 527 6487 (mobile); fax: +372 742 0286.
E-mail address: siljalaht@ebc.ee (S. Laht).

1570-9639/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2011.12.004

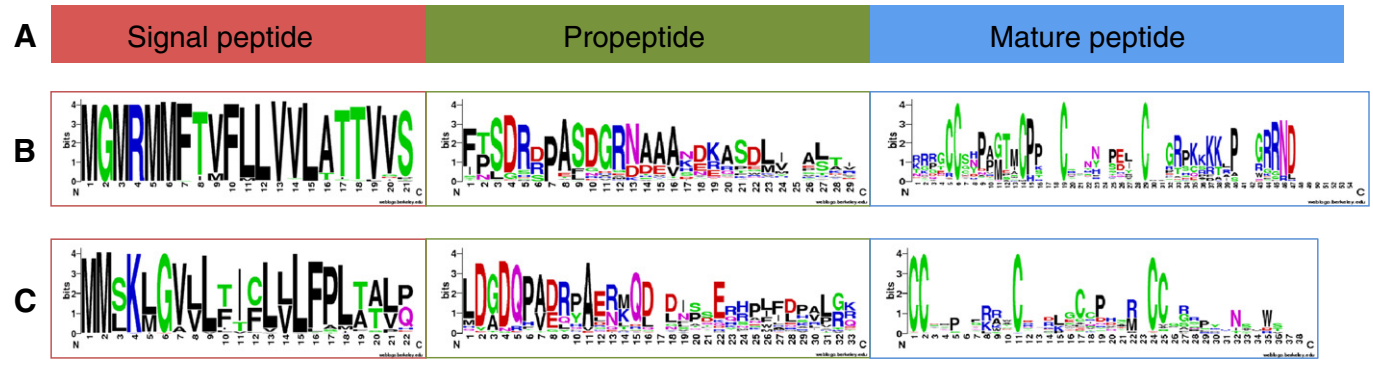

Fig. 1. Conopeptide precursor structure. Panel A—a schematic represenation of conopeptide precursor sequence. Panels B and C present sequence logos [25] for the A superfamily and the M superfamily sequences, respectively, to illustrate the level of sequence conservation of different precursor parts within given superfamilies.

of protein families; for example they set the basis of the Pfam database [11].

Several approaches for conopeptide superfamily prediction have been published over the last years [12–14]. Their main focus has been the prediction of the conopeptide superfamily based on a mature peptide sequence only, excluding superfamilies where only a few sequences had been described. We aimed at building a set of models that can be used to annotate all conopeptide superfamilies and families that have been described so far, even from partial sequences (mass-spec data, next generation sequencing data, etc.).

2. Materials and methods

2.1. Building of pHMMs for all conopeptide superfamilies and families

Previously described conopeptide sequences were downloaded from ConoServer [8] that has become the reference database for conopeptides. The sequences were grouped into 24 classes: 16 superfamilies (defined by signal region) and 8 families (defined by other patterns) by classification provided in the ConoServer (Table 1). Data redundancy was removed within each class using the CD-HIT program with 100% identity cutoff. With that step identical sequences and sequences contained within other sequences were removed but similar sequences, even with just one amino acid difference, were kept. CD-HIT is a program for clustering large sequence database at high sequence identity thresholds [15]. Only full-length precursor sequences consisting of signal, pro- and mature peptides were used for 8 superfamilies that contained at least 10 sequences. For smaller classes all available sequences were used. Sequences of each class were further subdivided into their signal, pro- and mature peptide parts according to the positions available in ConoServer. Each part was aligned with MAFFT version 6.707b using the L-INS-i method. MAFFT L-INS-i is one of the most accurate multiple sequence alignment methods currently available. L-INS-i is in particular suitable for alignment of 10–100 protein sequences [16,17].

A pHMM was built for each subset using *hmmbuild* from the HMMER 3.0 package [18]. *Hmmpress* from the same package was used to construct binary compressed data files for *hmmscan*.

Table 1
Conopeptide superfamilies and families that were modeled and the number of sequences used for pHMM training. Only full-length precursor sequences consisting of signal, pro- and mature peptide were used, if not otherwise stated.

No.	Superfamily	No of sequences in training set	No	Superfamily or family	No of sequences in training set
1	A	142	**13**	S[a]	8
2	D	18	**14**	T	129
3	I1	9	**15**	V	2
4	I2	35	**16**	Y	1
5	I3	7	**17**	Conantokin	7
6	J	6	**18**	Conkunitzin[b]	2
7	L	7	**19**	Conolysin[b]	2
8	M	75	**20**	Conophan[b]	2
9	O1	396	**21**	Conopressin[b]	6
10	O2	44	**22**	Conorfamide[b]	2
11	O3	21	**23**	Contryphan	9
12	P	6	**24**	Contulakin	3

[a] 7 full-length precursors and one mature peptide in the training set.
[b] Only a mature peptide sequence has been described for this conopeptide class.

2.2. Determination of how the number of sequences used for pHMM training affects sensitivity and specificity of classification

The 3 largest conopeptide superfamilies (A, O1, T; each containing at least 130 sequences) were randomly divided into one test set (50 sequences) and several training sets consisting of 2, 3, 5, 10, 20, 30, 40, 50, 60, 70 or 80 sequences. The training sequences were aligned with MAFFT L-INS-i program, and pHMMs were built for each set using *hmmbuild*. These pHMMs were formatted with *hmmpress* and then used to scan for full-length precursor sequences from the test sets with *hmmscan* (HMMER 3.0 package) with the default settings. The number of matches found with each model within the test set of the same class was recorded as the true positives. The same models were also scanned against a negative test set that contained all other conopeptide classes, except for the one that was used for training, with the same default parameters. The number of matches found from the negative test set was recorded as the false positives for each model.

The random division, sequence alignment, model building and testing were repeated 10 times, average number of matches and the standard deviation were calculated based on those iterations.

2.3. Determination of specificity of conopeptide pHMMs on UniProtKB/Swiss-Prot protein database

In order to determine the ability of conopeptide pHMMs to distinguish between conopeptides and other proteins all conopeptide pHMMs were scanned against the UniProtKB/Swiss-Prot protein database (downloaded on 17.08.2011, containing 531,473 protein sequences) [19,20] using *hmmscan* from the HMMER 3.0 package with different *E*-value cutoffs. HMMER3 only does local alignment, so there was no need to divide the protein sequences tested into different domains, when looking for matches with the pHMMs. All matches from *Conus* sp. were considered true positive. The true positives were manually revised for non-conopeptides, but all sequences retrieved with pHMMs that were from *Conus* sp. were indeed conopeptides. All sequences from other organisms were considered false positives.