Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies

Dominique Koua a,b, Silja Lahc c, Lauris Kaplinski c, Reto Stöcklin a, Maito Remm c, Philippe Favreau a, Frédérique Lisacek b,⁎

a Atheris Laboratories, Case Postale 314, CH-1233 Bernex-Geneva, Switzerland
b Swiss Institute of Bioinformatics, Proteome Informatics Group, Rue Michel Servet 1, CH-1206 Geneva, Switzerland
c Estonian Biocentre, Riiia Str. 23, Tartu 51010, Estonia

A R T I C L E   I N F O

Article history:
Received 27 March 2012
Received in revised form 1 December 2012
Accepted 26 December 2012
Available online 22 January 2013

Keywords:
Conopeptide
Position specific scoring matrix
Hidden Markov model
Combined prediction
Protein family prediction

A B S T R A C T

Classified into 16 superfamilies, conopeptides are the main component of cone snail venoms that attract growing interest in pharmacology and drug discovery. The conventional approach to assigning a conopeptide to a superfamily is based on a consensus signal peptide of the precursor sequence. While this information is available at the genomic or transcriptomic levels, it is not present in amino acid sequences of mature bioactives generated by proteomic studies. As the number of conopeptide sequences is increasing exponentially with the improvement in sequencing techniques, there is a growing need for automating superfamily elucidation. To face this challenge we have defined distinct models of the signal sequence, propeptide region and mature peptides for each of the superfamilies containing more than 5 members (14 out of 16). These models rely on two robust techniques namely, Position-Specific Scoring Matrices (PSSM, also named generalized profiles) and hidden Markov models (HMM). A total of 50 PSSMs and 47 HMM profiles were generated. We confirm that propeptide and mature regions can be used to efficiently classify conopeptides lacking a signal sequence. Furthermore, the combination of all three-region models demonstrated improvement in the classification rates and results emphasise how PSSM and HMM approaches complement each other for superfamily determination. The 97 models were validated and offer a straightforward method applicable to large sequence datasets.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Cone snails, their venoms and conopeptides

Cone snails are carnivorous marine gastropods that have evolved potent venoms to capture their prey, to defend themselves from predators and to deter competitors. Conopeptides, the main components of Conus venoms, represent a unique arsenal of neuropharmacologically active molecules that have been evolutionarily tailored to afford unprecedented action of the proteins de bond formation, amidation, hydroxyproline or pyroglutamate. The number and position of disulfide bonds refine the classification of conopeptides into structural families, and the mode of action of the proteins defines pharmacological or functional families [6,7].

C-terminal hypervariable mature toxin [7,8]. However, exceptions such as the I2-superfamily where the propeptide region is located down-stream of the mature toxin, may occur [9]. The precursor maturation, by the cleavage of the signal and propeptide from the mature region, leads to a biologically active conopeptide. The peptide maturation process also implies post-translational modifications of some residues, including disulfide bond formation, amidation, hydroxyproline or pyroglutamate. Consequently, protein sequence naming and classification have become a critical issue, as further studies towards function determination are based on this initial information. The Conoserver database [6,7,14] is a reference conopeptide repository, thanks to an impressive effort in update and annotation. But, due to the improvement of sequencing techniques, the number of published conopeptide sequences is continuously increasing. Crude
and milked venom diversity analysis [15–18] and venom gland transcriptome studies [19,20] lead to a challenging amount of precursors and mature peptide data. This justifies the design of automatic tools for assigning a new sequence to a conopeptide superfamily. The Conoserver prosence analyser [14] addresses this question by searching the signal peptide of a submitted precursor. But, since this prosence analyser is based on FASTA alignment of signal peptides, it fails to match truncated sequences missing the signal peptide and highly divergent precursors. More generally, classification tools are mostly based on the similarity of amino acid sequences though other approaches like ClanTox [21] attempt to merge many other criteria for characterising animal toxins as a broad category. In the present work, to gain finer insight in conotoxins we chose to rely on knowledge of currently known families and their corresponding multiple alignments.

A closer look at mature conopeptide alignment in each superfamily reveals both the wide evolutionary diversity of sequences and the occurrence of strongly conserved key residues, mostly cysteines defining a framework with characteristic spacing [1,5]. Consequently, a straightforward BLAST search with an individual mature sequence cannot pull out all sequences already known in the corresponding family. In many cases, conopeptide BLAST-based analysis requires manual processing of free text annotation and of weak matches with key residues in constrained positions. In fact, it has been long known that BLAST-processing of large amount of sequences generated in high throughput set-ups is not a sensitive approach. This issue was identified and addressed many years ago through the design of automatic classification tools of protein sequences matching functional and/or structural domains. Since then, the accuracy of automatic classification tools has gradually improved, especially in the field of model-to-sequence alignments [22–28].

Computational approaches applicable to the automatic annotation of large datasets commonly rely on protein signatures or profiles. A protein family profile is defined from the alignment of multiple family members. A range of techniques now considered classical are used to model family profiles. The two most popular are hidden Markov models (HMM) and position-specific scoring matrices (PSSMs). Thousands of protein profiles are publicly available in databases such as PROSITE [22], PRINTS [23] or Pfam [24] allowing protein sequence classification into known families. However, screening a sequence database with a family profile does not necessarily lead to unmistakably select all family members. Models are as accurate as possible and usually set to reach very high specificity and sensitivity thereby keeping false positives and false negatives to minimal values (less than 5%). Modelling of complete domains facilitates more biologically meaningful sequence annotation.

Hidden Markov models are particularly useful to represent sequence heterogeneity. HMM profiles assign a position-specific scoring system to residue substitutions, insertions, and deletions. Compared to most sequence alignment algorithms, the most popular profile builder HMMER is significantly more accurate and detects remote homologues thanks to the strength of the underlying probability models, improved in the latest version [25–27]. HMM profiles are particularly useful for fully automated sequence annotation not requiring human expertise.

A generalised profile is another method that evaluates sequence similarity. It is based on multiple alignments and requires building position-specific scoring matrices (PSSMs) where amino acid substitution is variably assessed depending on the position and observed frequencies. For example, a Tyr–Trp substitution at position x in a set of aligned sequences may be weighed differently than the same substitution at position y depending on observed frequencies at each respective positions. This approach contrasts with pre-computed and position-independent PAM and BLOSUM substitution matrices (http://www.ncbi.nlm.nih.gov/books/NBK21106/def-item/app15/) in which amino acid substitutions are assigned a score once and for all. The most recent PSSM methodology is based on AMSA (Annotated Multiple Sequence Alignment) [28] shown to enhance the separation of very close protein families. The discriminative ability of PSSMs is improved by a post-processing competition step in which the profile that produces the best matching score for each query sequence is selected [29]. Generalized profiles allow manual tuning by human experts prior to full automation.

This study aims at combining the advantages of HMMs (speed and reliable probability models) and generalized profiles (accuracy and fine-tunable models) for conopeptide superfamily classification based on different precursor regions and the potential application of this combined approach to large sequence datasets.

Public protein family databases do contain a few available conopeptide superfamily patterns and profiles but these are incomplete or ambiguous. Up to now, profiling conopeptides, was limited to the Conoserver prosence analyser but this option is not easily amenable to the treatment of large datasets and as pointed out earlier, may not work when the signal sequence is missing, which happens frequently in high throughput data. Furthermore, indisputable evidence for the relevance of consensus sequences of propeptides and cysteine frameworks supports the claim that the classification of new sequences in conopeptide superfamilies cannot be only based on the signal peptide [30]. Note that other bioinformatics approaches based on support vector machine (SVM) [31,32] and amino acid composition [33], have been proposed for conopeptide classification and were shown to improve classification results compared to standard evolutionary or sequence/alignment based approaches. However, the implementation of SVM methods is not suited to application and usage in high throughput set-ups and cannot realistically be undertaken for protein classification and family prediction in routine analysis. It should be noted that superfamily-specific profiles have not been defined so far neither for mature regions nor propeptides, even in the few available PROSITE patterns and profiles.

This work introduces a methodology which application demonstrates the added efficiencies of HMMs and generalized profiles to accurately and reliably predict conopeptide superfamily membership, not only from complete precursors but also from mature and propeptide regions. In addition, the combination of HMMs and generalized profiles also confirms the quality of model-based superfamily classification. Models were thus built starting from the 16 known conopeptide superfamilies. This approach can easily be included in a pipeline for processing large amount of data and saves long manual analysis of BLAST matches that incidentally also requires the input of experienced researchers.

2. Materials and methods

2.1. Preparation of model construction data set

Sequences used for generating the models were obtained from the Conoserver database in xml format and parsed with in-house perl scripts (conoserver_protein.xml, 2012, 06 28th) [6]. Only complete full-length precursor sequences with gene superfamily annotation were considered, amounting to 1364 sequences. An additional restriction to superfamilies containing at least 5 full precursor sequences was imposed, limiting the number of superfamilies to 14. The remaining V (2 sequences) and Y (1 sequence) superfamilies were left out. The training set consisted of 967 randomly picked sequences of the precursor dataset, spread in 14 superfamilies. The training set therefore represents roughly 2/3 of the sequence data.

Each sequence was divided into 3 parts stored separately:

1. signal sequence,
2. propeptide region, and
3. mature peptide.

Due to their distinct cysteine composition, the A, O1 and O2 superfamilies were split into two and corresponding subsets reflecting the number and/or patterns of cysteines in the mature peptide were...
defined. This led to the generation of A_4 and A_6 for sequences of the A superfamily bearing 4 and 6 cysteines respectively. Likewise, O1_6, O1_8, O2_6 and O2_8 for precursors of the O1 and O2 superfamilies were defined with 6 and 8 cysteines respectively.

Separate files were created for each superfamily and region, resulting in 51 subsets \((14 - 3)^3 + 6^3 = 51\). Sequences were then aligned using MAFFT, version 6.707b software [34]. The resulting alignments were manually refined with the JALVIEW 2.5 software [35] in order to reduce sequence redundancy. Table 1 summarises numbers of sequences in each training subset after redundancy reduction. The 51 alignments obtained were used to build the models: in FASTA format for PSSM construction and in STOCKHOLM format for HMMs. A FASTA file of the aligned sequences used for model training is available as additional file 1.

### 2.2. Model construction

In both cases of modelling techniques and for each of the 14 known conopeptide superfamilies, three separated models respectively based on signal sequences, propeptide regions and mature peptides were built.

#### 2.2.1. Hidden Markov models

Even though A and O superfamilies were split into A_4, A_6 and O1_6, O1_8, O2_6, O2_8 respectively, these subgroups mainly reflect differences in mature peptide regions and not necessarily in the signal peptide. Indeed, A_4 and A_6 share nearly the same signal peptide; likewise, O1_6 and O1_8 on the one hand, and O2_6, O2_8 on the other hand. The limited differences between the A_4 and A_6 propeptides also led to building the same model for both sets of sequences. Consequently only 51 – 4 = 47 HMM profiles needed to be constructed.

Profile HMMs were built for each of the 47 alignments with the hmmbuild script from HMMER 3.0 package using default parameters [26]. Models were named according to the superfamily and the region of the precursor they target. For example, the model built with the A superfamily signal region was named A_SIG, while A_4_MAT corresponds to the model of mature peptides of the A superfamily containing four cysteines (typically CC–C–C) and A_6_MAT to the model of mature regions of the A superfamily containing six cysteines. Then, the hmmsearch script was run to identify matches between newly built HMM profiles and sequences of the test sets, with the e-value significance level set to 0.1.

The quality of the 47 models was evaluated with sequences of the test set (see details in Table 1 and Section 2.3). The hmmsearch script outputs lists of potential family members with a significant score (i.e., e-value smaller than 0.1). Results are summarised in Table 2. For instance, the 43 signal peptide sequences of the A_4 test set were all recognised by the corresponding model, whereas only 41 of the 43 propeptide sequences significantly matched the model.

#### 2.2.2. Generalized profiles

In this case, 50 generalized profiles were built. Even with limited differences between the O1_6 and O1_8 signal peptides or A_6 propeptides, different models could be generated for these sets of sequences. Only models of A_4 and A_6 signal peptides remained undistinguishable.

Generalized profiles were constructed using the pfpool package version 2.3. Each model was built directly from the superfamily alignments as in the HMMs construction. The generalized profiles were generated using apsimake (from the pfpool package) in a semi-global mode after weighing the alignments. The profiles were then calibrated using a randomized version of the UniProtKB/Swiss-Prot database and cutoff values were tuned manually to avoid false positive matches. Finally, special "compete lines" were added in view of the post-processing competition. This competition step allows returning only the profile that produces the best score when more than one profile matched the same sequence. The pfsearch script was used to perform alignments between profiles and sequence sets. In-house Perl scripts were implemented to facilitate result visualization and exploitation.

The quality of the 50 models was evaluated with sequences of the test set (see details in Table 1 and in Section 2.3). The pfsearch script outputs lists of potential family members with a significant score. Results are summarised in Table 2. For instance, 39 of the 43 signal peptide sequences of the A_4 test set were recognised by the corresponding model, whereas only 42 of the 43 propeptide sequences significantly matched the model.

#### 2.3. Quality of model-based classification of known conopeptides

The initial test set was built with the 397 remaining full precursors of those 1364 extracted from Conoserver (approximately one third) and not used for model construction, thereby guaranteeing the independence of the training and test sets. These conopeptide precursor sequences were split into 3 parts resulting in a FASTA file containing 397 signal sequences, 397 propeptide regions and 397 mature peptides, leading to a final test set consisting of 1191 sequences. Each sequence was associated with both its corresponding superfamily and precursor region. A FASTA file of the split sequences used for model testing is available as additional file 2.

The selectivity and sensitivity of each model were evaluated using the following acknowledged formulas:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Selectivity} = \frac{TP}{TP + FP}
\]

where TP: true positive, FN: false Negative, and FP: false positive.

Note that the same test set is suitable to assess global selectivity and sensitivity. For example, since all mature sequences of all families are pooled together, the discriminatory power of a model built for a given family can be assessed with mature peptide sequences of close families. Moreover, it is also easy to verify that signal-based profiles will not match propeptide region or mature peptides.

Model matches as well as related sensitivity and selectivity are given in Supplementary file 3.
2.4. Merging predictions from HMM and generalized profiles

2.4.1. HMM-based classification

For HMM-based classification, we adopted the product of e-values for signal, propeptide region and mature peptide region models as the final score for each superfamily.

$$pHMMScore_{sequence~i,~superfamilyX} = e^{-value_i(\text{phHMMSig})} \times e^{-value_i(\text{phHMMPRO})} \times e^{-value_i(\text{phHMMMAT})}$$

To maximize the power of discriminating true from false homologues in a database search, statistical inference theory suggests that sequences should be scored by integrating over alignment uncertainty as opposed to scoring the single best alignment. This idea is implemented in HMMER 3.0 [26], where log-odds likelihood scores are summed over alignment uncertainty (forward scores), rather than optimal alignment (Viterbi scores). Furthermore, posterior probabilities of alignment confidence are reported, enabling detailed and intuitive assessments of alignments [26,27]. These embedded statistics refine the separation between correct and incorrect matches. The internal heuristic of HMMER is robust and has been used in this study to discriminate between superfamilies.

Best matches have lowest pHMM_Scores. The separation between scored matches was fixed to a factor of 100 according to results observed on the training sets. As hinted above, for each sequence analysed with the 47 built HMM profiles, a global family-linked pHMM score is evaluated by multiplying the three scores obtained in each modelled superfamily and corresponding to the signal-based, propeptide-based and mature-based models. For each sequence, a pHMM_Score is therefore produced for each conotoxin family. A sequence is then predicted to belong to the family with the lowest score, when the latter is at least one hundred times lower than the next lowest pHMM_Score.

### Table 2

Distribution of matches across the test set.

<table>
<thead>
<tr>
<th>Superfamilies</th>
<th>Signal sequence</th>
<th>Propeptide region</th>
<th>Mature peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HMM</td>
<td>PSSM</td>
<td>HMM</td>
</tr>
<tr>
<td>A_4</td>
<td>A_SIG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>A_PRO&lt;sup&gt;a&lt;/sup&gt;</td>
<td>34 A_4_MAT</td>
</tr>
<tr>
<td></td>
<td>10 A_6_SIG</td>
<td>6 A_6_SIG</td>
<td>39 A_4_MAT</td>
</tr>
<tr>
<td>A_6</td>
<td>1 M_6_SIG</td>
<td>10 D_10_SIG</td>
<td>8 A_6_MAT</td>
</tr>
<tr>
<td></td>
<td>11_8_SIG</td>
<td>7 D_10.PRO</td>
<td>10 D_10.PRO</td>
</tr>
<tr>
<td>I_1</td>
<td>4 I_1_8_SIG</td>
<td>4 I_1_8.PRO</td>
<td>34 I_1_8_MAT</td>
</tr>
<tr>
<td>I_2</td>
<td>15 I_2_8_SIG</td>
<td>15 I_2_8_PRO</td>
<td>4 I_2_8_PRO</td>
</tr>
<tr>
<td>I_3</td>
<td>11_8_SIG</td>
<td>2 I_3_8_PRO</td>
<td>2 I_3_8_PRO</td>
</tr>
<tr>
<td>J_1</td>
<td>4 J_1_4_SIG</td>
<td>4 J_1_4.PRO</td>
<td>4 J_1_4_MAT</td>
</tr>
<tr>
<td>L_1</td>
<td>11_8_SIG</td>
<td>1 L_4_PRO</td>
<td>1 L_4_PRO</td>
</tr>
<tr>
<td>M_1</td>
<td>99 M_6_SIG</td>
<td>99 M_6_PRO</td>
<td>72 M_6_MAT</td>
</tr>
<tr>
<td>O1_6</td>
<td>01_6_SIG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>133 O1_6_SIG</td>
<td>129 O1_6_MAT</td>
</tr>
<tr>
<td>O1_8</td>
<td>2 O1_8_SIG</td>
<td>2 O1_8_PRO</td>
<td>2 O1_8_PRO</td>
</tr>
<tr>
<td>O2_6</td>
<td>4 O2_8_SIG</td>
<td>7 O2_8_PRO</td>
<td>7 O2_8_PRO</td>
</tr>
<tr>
<td>O2_8</td>
<td>4 O2_8_SIG</td>
<td>4 O2_8_PRO</td>
<td>4 O2_8_PRO</td>
</tr>
<tr>
<td>O3</td>
<td>1 M_6_SIG</td>
<td>8 O3_6_PRO</td>
<td>7 O3_6_PRO</td>
</tr>
<tr>
<td>P_1</td>
<td>2 P_6_SIG</td>
<td>2 P_6_PRO</td>
<td>2 P_6_PRO</td>
</tr>
<tr>
<td>S_1</td>
<td>34 M_6_SIG</td>
<td>2 S_10_PRO</td>
<td>2 S_10_PRO</td>
</tr>
<tr>
<td>T_1</td>
<td>40 T_4_SIG</td>
<td>37 T_4_PRO</td>
<td>18 T_4_MAT</td>
</tr>
</tbody>
</table>

<sup>a</sup> A single model was built for the corresponding region.
2.4.2. PSSM-based classification

For generalized profile-based predictions, we evaluated a PSSM_Score for each superfamily. This score was correlated with the number of models of a given family that match the analysed sequence.

\[
\text{PSSM}_{\text{Score}}(\text{sequence i, superfamily}X) = \\
\text{HasMatch}(i, \text{PSSM}_{\text{sig}}) + \\
\text{HasMatch}(i, \text{PSSM}_{\text{pro}}) + \\
\text{HasMatch}(i, \text{PSSM}_{\text{mat}}),
\]

where the boolean function \(\text{HasMatch}(\text{sequence, model})\) returns 1 if the sequence matched the considered model, or 0 otherwise. In other words, for each superfamily of conopeptides, the existence of matches obtained with each of the 3 corresponding models (signal, propeptide or mature region-based model) defines the PSSM_Score.

Best matches have highest PSSM_Scores. As indicated in Section 2.2.2, an internal competition is performed during PSSM search in order to return only matches with the highest scores [28,29]. The strategy used to compute PSSM_Scores led to reliably determine the original superfamilies models matching a given sequence: PSMM_Scores indicate the number of models of a given superfamily matching the query sequence. Then, superfamily membership is defined as the highest number of matching models. In case of multiple highest values, the list of superfamilies with equivalent number of matching models is returned.

2.4.3. Combined classification

The final assignment of a sequence to a superfamily relies on the combination of PHMM_Scores and PSSM_Scores. The global prediction of a sequence membership to a superfamily follows the comparison of each classification outcome: the classification appearing twice is selected and attributed to the sequence. If only one of the two classification methods produces a result, then the sequence is assigned to the family predicted with a single method. If no classification is output, the sequence is tagged as “UNKNOWN”. If the two classification methods disagree and a sequence ends up in more than one superfamily, the sequence is tagged as “CONFLICT”.

Sequences tagged “UNKNOWN” or “CONFLICT”, were considered as incorrect predictions in the graphical representation summarising results in Fig. 2.

2.5. BLAST-based superfamily classification of mature peptides

The 397 mature peptides sequences of the test set were sent to a BLAST search against the NCBI non-redundant database. BLAST version 2.2.24 [Aug-08-2010] was run with options: -m 7 -T -P 3 -v 3 -b 3 -a 6. An XML file containing the first 3 hits for each sequence was returned (the first hit has a high probability of being the submitted sequence).

For each submitted sequence, a superfamily was manually attributed based on the hit description of the first BLAST match not being itself: hits with 100% sequence similarity were excluded since the goal is to classify new sequences, based on the description of the closest blast hit. Results obtained after manual BLAST output check were compared to the automatic attribution obtained with the combined HMM + PSSM approach (see 2.4.3).

3. Results

3.1. Extension of superfamily classification criteria

The 97 models described in Section 2.2 were applied to classify the 1191 sequences of the test set described in Section 2.3. The quality of the matches was assessed in terms of sensitivity (percentage of true positives in the whole data set) and selectivity (percentage of true positives among all sequences matched with a particular model). For each model, all matched sequences (see Table 2) are used to evaluate sensitivity and selectivity according to formulas given in Section 2.4.

For HMMs as well as for generalized profiles, in nearly all superfamilies, propeptide region and mature peptide based models show excellent classification abilities (Table 2 and Fig. 1).

As expected, the signal-based models are a better predictor than models based on mature peptides and propeptide regions. However, models based on propeptide and mature regions demonstrated good prediction ability. The propeptide-based models and mature-based models appear extremely useful for the classification of sequences missing the signal sequence. This was established previously for profiles HMM [30], the present study confirms the result for generalized profiles. As stated in [30], the classical classification paradigm for conopeptides can be enriched by the consideration of mature and propeptide regions. Results provided in the current study confirm this extension of the classification criteria with both types of models (HMM and PSSM).

For instance, in the test set and as reported in Table 2, A-superfamily is represented by 129 sequences; 43 with a signal peptide; 43 precursor fragments including a propeptide region and 43 only consisting of the mature peptide. The A.4 мат generalized profile correctly classified 39 of these sequences with only 2 false positives while the corresponding HMM correctly classified 34 sequences with no mistake. In the case of the M-superfamily, 73 out of the 99 signal-free amino acid sequences were successfully classified by both M.6 мат generalized profile and profile HMM. A last evidence of superfamily classification based on the mature region is the O1 superfamily: among the 133 sequences missing a signal sequence, 128 were correctly assigned by the O1.6 мат generalized profile. The HMM version correctly classified 129 mature sequences of O1-conopeptides along with some false positive hits.

Similar examples of the varying classification power of models based on the propeptide region are reported in Table 2. Globally, the selectivity and sensitivity of generalized profiles are respectively 99.42% and 92.81% that is, slightly higher than that of HMM profiles: 87.18% selectivity and 89.65% sensitivity. In superfamilies where sensitivity and/or selectivity were not equal, Fig. 1A and B represent the classification performance of the generalized profiles compared to that of HMMs. Matches obtained from the test set suggested that either the signal peptide, the propeptide or the mature peptide could be used to reliably classify conopeptide sequences into superfamilies (Supplementary file 3).

3.2. Classification based on combination of precursor region models

In the previous section, it was established that models built from the different parts of the precursor were individually suitable to classify superfamilies. It also appeared that for both HMMs and generalized profiles methods the combined use of the three models (signal, propeptide and mature region) of a given superfamily increased the superfamily assignment efficiency. For example, there were many cases where the propeptide-based model allowed fishing out and properly classifying sequences that were not matched by the mature peptide based model. This suggests that combining the two types of models improves the classification of conopeptides found in proteomic studies of venoms.

In the A-superfamily, when considering fragments missing a signal sequence, the combination of the propeptide and mature region HMM models could properly classify all of the 43 proteins of the test set (34 found by the mature-based model and 41 found by the propeptide-based model). The classification based on the combination of A-superfamily generalized profiles also reliably classified all sequences of the test set. Similarly interesting results are provided as supplementary material (file 3).

However, it must be noted that most of the complete precursors were matched simultaneously by the 3 models. This highlights the possibility of recovering some precursors with a divergent signal, propeptide or mature regions.
Some sequences of the test set remained unclassified or misclassified by either HMMs or generalized profiles. Most of them were mature sequences sharing the same cysteine framework (for instance, A-conotoxins classified as T). As expected, very close cysteine frameworks remained a decisive issue when performing a classification based on mature peptides.

3.3. Merging predictions from HMM and generalized profiles

Assessing the quality of conotoxin classification based on signal, propeptide and mature regions (Section 3.2) led us to consider the effect of merging separate predictions obtained by the region-based models (Section 3.3). This approach improved global HMM-based and global PSSM-based prediction/classification, respectively.

In addition, the study could show the efficiency of merging HMM and PSSM based classifications. Matches obtained from the separate combination of the three HMMs on one hand and, on the other hand, from the combination of the three generalized profiles for each superfamily, were processed together as indicated in the "Materials and methods" section. The merge of the two complementing profile generating techniques significantly increased the number of distinct proteins correctly predicted/classified for each superfamily (see Section 2.4.3).

Different situations occurred when combining the results from the two approaches. Fig. 2 illustrates the combined classification of mature peptides of the test set. In addition to the 3 sequences matched in common, the I1-superfamily HMM profile identified one sequence not classified by generalized profiles whereas the opposite was observed for A-superfamily. In I2-, O1- and O2-superfamilies, HMMs and generalized profiles detected some sequences in a mutually exclusive manner. After global merging, the number of distinct sequences correctly classified was considerably improved on the test set. In the M-superfamily, 89% of correct classification was achieved while individual prediction was 87.9% and 73.7% for generalized profile and for HMM-based classification, respectively. In the O2- and I2-superfamily, the combination of predictions resulted in a correct classification rate of 100% of sequences of the test set. The PSSM approach tended to be more predictive for highly variable motifs. For instance, in the T-superfamily, the global HMM only classified 45% of the test set while the global PSSM correctly predicted 85% of the submitted sequences. A similar phenomenon occurred in the hypervariate M-superfamily mature peptide region. The PSSM models failed to classify D-Superfamily sequences which were also detected with relatively similar scores by the M-Superfamily models, resulting in a conflict. The combination of HMM and PSSM allowed disambiguation since the HMM properly classified these sequences as D-Superfamily conopeptides.

Interestingly, a subset of models could allocate previously unclassified conopeptides to a known superfamily [6,7,14]. Conomarphin sequences were shed out by the M-superfamily signal model in a
very convincing manner. Contryphans were identified only by the M-superfamily propeptide model. The bromosleeper sequence was matched by the O3-superfamily mature model.

3.4. Comparison to BLAST-based superfamily prediction

As the BLAST tool is a standard approach often used for superfamily classification, a comparison was made with our models. A subset of 253 mature sequences of the test set was submitted to BLAST against the non-redundant NCBI database. Manual superfamily attribution based on the BLAST results was compared to the automatic prediction obtained with our model-based approach. It appeared that BLAST-based superfamily attribution rapidly became a fastidious task that could only be carried out by experienced users. The hit description section of the BLAST XML output is a free text section without any controlled vocabulary. It is therefore frequent to find lots of synonyms for a given superfamily. The hit annotation gives either gene superfamilies related information, or scaffold related information with roman or Arabic annotation as well as pharmacological family information. For example, “mu-O-conotoxin”, “scaffold VI/VII”, “G1”, “delta-conotoxin”, “omega-conotoxin”, “O-superfamily” and “superfamily O” were found to all stand for the O-superfamily. Some more complicated cases were found like “gi|12619395|gb| AAG60359.1|AF214931_1 conotoxin scaffold III/IV precursor” where scaffold III sequences belong to A-superfamily and scaffold IV precursors belong to M-superfamily.

It may be difficult to deduce the right hit except when the BLAST output description contained the precise assignment to the correct superfamily. In most of the cases, the superfamily assignment was imprecise (for instance, I-superfamily instead of I1-, I2- or I3- and O-superfamily instead of precise O1-, O2- or O3-superfamilies). However superfamilies were deduced from annotation not containing the precise superfamily. For instance “alpha-conotoxin” was interpreted as A superfamily, “mu-conotoxin” was accepted as M superfamily. In few cases, the BLAST hit annotation only allowed the detection of a conopeptide but with no indication on superfamily membership. Finally, in very rare cases, the best BLAST hit did not belong to a conopeptide sequence.

In the end, the BLAST-based classification does not appear easily applicable to automated annotation/prediction of large datasets. Automated BLAST result parsing would result in solving too many exceptional cases. Consequently, BLAST-based superfamily prediction seems less efficient than the model-based strategy. While the combined HMM/PSSM strategy allowed precise annotation of O1-, O2- and I1-superfamilies, the corresponding BLAST hits only allowed the deduction of O- or I-superfamilies. Our new approach also made it possible to classify some sequences for which no BLAST match were found with the default e-value (see Supplementary data file 4).

4. Conclusion

This study established that the conopeptide superfamily classification and identification can reliably be achieved based on the mature and propeptide regions and access to the signal peptide is not a prerequisite. The combination of hidden Markov models and generalized profiles appeared as an efficient approach to perform an extensive classification and/or prediction of conopeptide sequences into superfamilies or families. Specificity test on a large sequence dataset and a final comparison with a BLAST-based approach indicated the usefulness of the designed models. Each model built in this study demonstrated very high discriminative abilities, with high sensitivity and selectivity in superfamily classification. We obtained very high specificity when searching the whole UniProtKB database as well as excellent selectivity and sensitivity for closely related conopeptide superfamilies. For the first time, a method combining PSSM and HMM profiles built on signal, propeptides or mature sequences has been developed and validated for correct and precise superfamily prediction of conopeptides. Consequently, superfamily classification can be extended to truncated sequences missing a signal peptide, as is frequently the case in genomic, transcriptomic and proteomic studies. This combined prediction approach opens up new prospects for the annotation of conopeptides and others toxin families.

Authors’ contributions

DK prepared the sequence alignment including manual edition, built and calibrated PSSMs, wrote scripts for data analysis, interpreted the results, and drafted and corrected the manuscript. SL contributed to data
acquisition, designed HMM and analysed HMM matching result and was involved in the manuscript drafting. LK contributed to data acquisition and HMM results analysis, PF annotated sequence data, contributed to results interpretation and revised the manuscript. RS, MR and FL contributed to the analysis, conception and design, critical manuscript revisions and final approval.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbabap.2012.12.015.

Acknowledgements

We are most grateful to Estelle Bianchi, Daniel Bias, Nicolas Hulo and Christian Sigrist for expert assistance.

Funding: This work was supported by the European Commission: CONCO, the cone snail genome project for health (LSHB-CT-2007-037592; www.conco.eu) and the European Regional Development Fund (Estonian Centre of Excellence in Genomics).

References