W238–W241 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 31 May 2012*
doi:10.1093/nar/gks337

ConoDictor: a tool for prediction of conopeptide superfamilies

Dominique Koua[1,2,*], Age Brauer[3], Silja Laht[3], Lauris Kaplinski[3], Philippe Favreau[1], Maido Remm[3], Frédérique Lisacek[2] and Reto Stöcklin[1]

[1]Atheris Laboratories, Case postale 314, CH-1233 Bernex-Geneva, Switzerland, [2]Proteome Informatics Group, Swiss Institute of Bioinformatics, CH-1211 Geneva, Switzerland and [3]Bioinformatics Workgroup, Estonian Biocentre, EE-51010 Tartu, Estonia

Received January 27, 2012; Revised March 27, 2012; Accepted April 4, 2012

ABSTRACT

***ConoDictor* is a tool that enables fast and accurate classification of conopeptides into superfamilies based on their amino acid sequence. *ConoDictor* combines predictions from two complementary approaches—profile hidden Markov models and generalized profiles. Results appear in a browser as tables that can be downloaded in various formats. This application is particularly valuable in view of the exponentially increasing number of conopeptides that are being identified. *ConoDictor* was written in Perl using the common gateway interface module with a php submission page. Sequence matching is performed with hmmsearch from HMMER 3 and ps_scan.pl from the pftools 2.3 package. ConoDictor is freely accessible at http://conco.ebc.ee.**

INTRODUCTION

Conopeptides are the main bioactive component of cone snail venom. These marine animals produce complex venoms that contain hundreds of peptides and proteins. Recently, conopeptides have attracted a great deal of interest as a result of their selectivity for, and potent effects on, ion channels and receptors (1,2). Most are cysteine-knotted peptides that have been classified into superfamilies and families based on their structural or functional features (3,4). To date, >1500 non-redundant conopeptide sequences are stored in public databases and this number is increasing exponentially. Conopeptides are classified into 'gene superfamilies' based on their signal sequence. Currently, there are 16 major superfamilies, namely: A, D, I1, I2, I3, J, L, M, O1, O2, O3, P, S, T, V and Y. The precursors generally contain an N-terminal signal sequence, a central propeptide region and a C-terminal hypervariable mature toxin (4,5). The Conoserver database (http://www.conoserver.org/) is a repository of nucleic acid and protein sequences, and of structural information on conopeptides (4).

The naming and classification of new conopeptide protein sequences has become an important issue because of the sharp increase in the number of new conopeptides being identified, and because studies to determine the peptide's functional characteristics are based on this classification. The Conoserver prosequence analyzer (ConoPrec) is the most specific web tool available for elucidation of conopeptide class. It provides hints based on the signal peptide sequence of the submitted precursor (6). However, this tool does not work when the signal sequence is missing, which is often the case with conopeptides identified by proteomic and mass spectrometry studies of toxins identified as mature bioactive peptides in venom or dissected venom gland. As data generated by spreading venom high-throughput omics is notoriously incomplete, the classification of new sequences into conotoxin superfamilies should not be restricted to the signal peptide sequence. There is indisputable evidence for the relevance of consensus sequences of propeptides and cysteine frameworks in conopeptide sequences. Thus, the inclusion of these criteria should also be considered for conopeptide classification. We recently demonstrated the reliability of conopeptide family prediction and classification based on profile hidden Markov models (pHMM) of propeptides and mature peptides (7).

ConoDictor has been developed in the context of the CONCO project (www.conco.eu) and is a web-based tool that exploits pHMMs and position-specific scoring matrix (PSSM, also known as generalized profiles) to classify conopeptide into superfamilies based on their amino acid sequence. *ConoDictor* is a user-friendly tool that meets users' demands for an easy-to-use environment for sequence classification and superfamily prediction. As a fully automated tool, ConoDictor provides classification results that must be checked by users.

*To whom correspondence should be addressed. Tel: +41 228500585; Fax: +41 228500586; Email: dominique.koua@atheris.ch

MATERIALS AND METHODS

Preparation of the model data set

Sequences used for generating the models were obtained from Conoserver. Only precursor sequences with gene superfamily annotation were considered. The training set consisted of 933 sequences. Each sequence was manually annotated with the gene superfamily classification after checking the classification provided by Conoserver. Each sequence was divided into three parts, which were stored separately: signal, propeptide and mature peptide. Separate files were also created for each of the 16 superfamilies. The sequences were then aligned using the MAFFT version 6.707b software. The alignments were manually refined when necessary using the JALVIEW 2.5 software, and the resulting 48 alignments were used to build the models.

Hidden Markov models

We previously described pHMM ability for conopeptide classification (7). We constructed pHMMs for each of the 48 alignments using the hmmbuild script from the HMMER 3.0 package (8,9). Matches between pHMMs and the sequence data set were searched using the hmmsearch script with an *e*-value significance level set to 0.1.

Generalized profiles (PSSM)

Generalized profiles were constructed using the pftool package, version 2.3. The most recent methodology based on annotated multiple sequence alignment (AMSA) was used. The generalized profiles were generated using apsimake in a semi-global mode after weighing of alignments. The resulting models were calibrated against randomized sequences and cut-off values tuned manually. These approaches have already been validated for classification of other proteins (10,11).

Testing of models on known conopeptides

The test set was constructed from publicly available conopeptide sequences extracted from the NCBI Protein database and UniProtKB (release 2010_11). The test set contained 1225 manually curated sequences. Sequences were manually annotated and assigned to the relevant superfamily according to UniProtKB annotations, cysteine frameworks and sequence similarity. Sequences not belonging to any superfamily were added to the test set as negative controls.

ConoDictor implementation

Input sequences are first classified using pHMMs and PSSMs separately. pHMM models of signal (X_sig), propeptide (X_pro) and mature peptide (X_mat) are used in parallel and corresponding predictions are combined. The same process is applied with PSSM models. Resulting pHMM and PSSM classifications are merged to produce a global combined classification.

(i) For pHMM-based classification, we adopted the product of *E*-values as final score:

$$\begin{aligned}\text{pHMM\_Score}(\text{sequence } i, \text{superfamily}X) = {}& E\text{−value}(i,\text{pHMM\_X\_sig})\\ & * E\text{−value}(i,\text{pHMM\_X\_pro})\\ & * E\text{−value}(i,\text{pHMM\_X\_mat}),\end{aligned}$$

provided that the corresponding *E*-value exists. A sequence was predicted to belong to the superfamily with the smallest pHMM score when this score was at least one hundred times lower than that of any other superfamily. If the difference was smaller, the sequence was predicted as ‘CONFLICT’ for the pHMM. When no score was generated for any superfamily, the sequence was tagged ‘UNKNOWN’.

(ii) For generalized profile predictions, it is not possible to compare and merge scores obtained from separate profiles. The PSSM prediction score for a sequence is the number of models of one superfamily (1–3) that match the sequence:

$$\begin{aligned}\text{PSSM\_Score (sequence } i, \text{superfamily}X) = {}& \text{HasMatch}(i,\text{PSSM\_X\_sig})\\ & +\text{HasMatch}(i,\text{PSSM\_X\_pro})\\ & +\text{HasMatch}(i,\text{PSSM\_X\_mat}),\end{aligned}$$

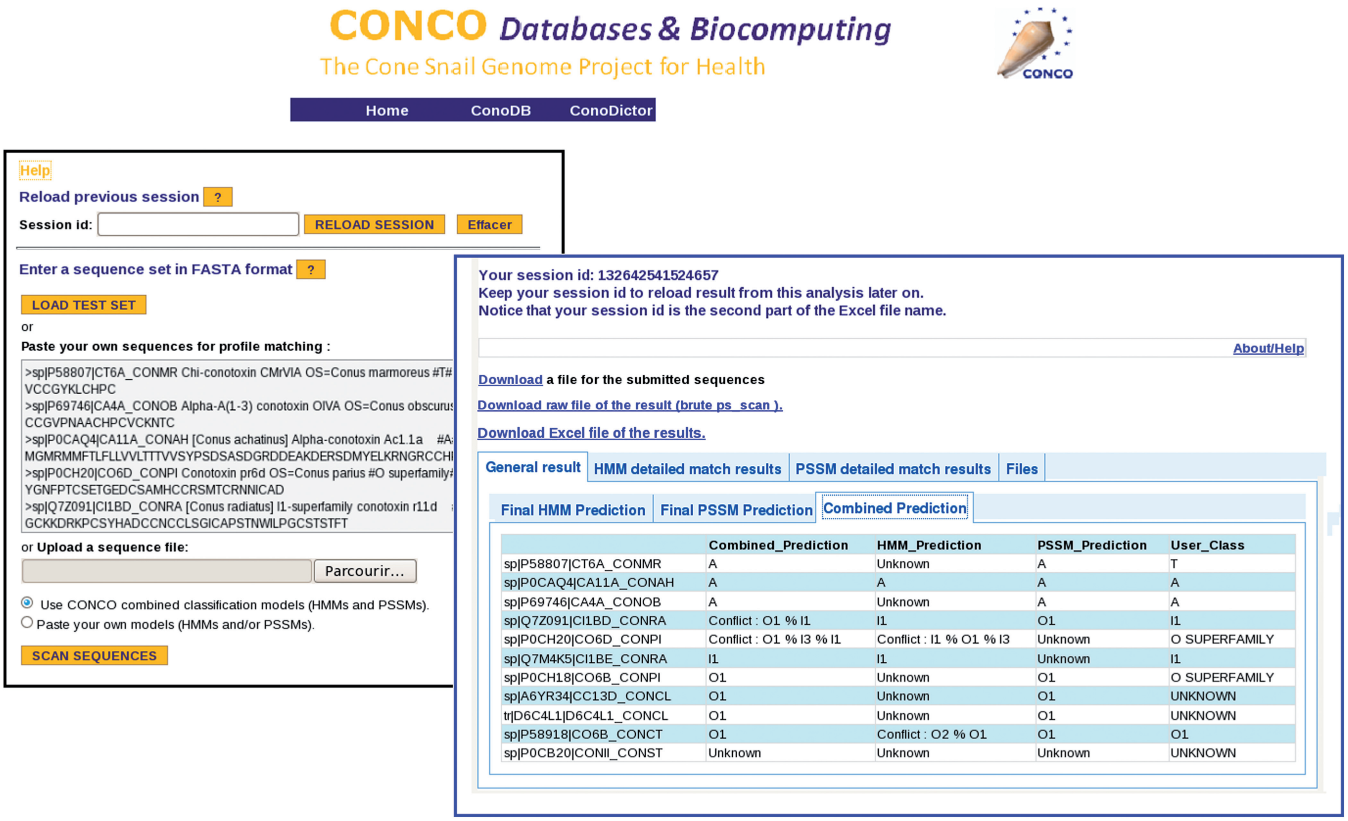
where the boolean function HasMatch(sequence, model) returns 1 if the sequence matched the considered model, or 0 otherwise. The sequence is predicted to belong to the superfamily with the highest score. If two or more superfamilies have the same score, the sequence is tagged as ‘CONFLICT’, and the list of conflicting families is returned. When no match is reported for a given sequence, the sequence is tagged ‘UNKNOWN’.

Match lists of pHMMs and PSSMs are merged, and each prediction is weighted according to its frequency. The combined prediction is the superfamily with the highest frequency. When the highest frequency is linked to more than one superfamily, the sequence is tagged ‘CONFLICT’. When no match is reported for either method, the sequence is tagged ‘UNKNOWN’. Even if HMM and PSSM are very robust classification approaches, the reduced size of learning set in some families and/or the underlying scoring system can justify rare cases of misclassification. The ‘CONFLICT’ and ‘UNKNOWN’ tag can therefore represent not modelled families (may be new ones) or divergent sequences from an existing family. In any case, all classifications have to be validated by users before being used for further studies.

RESULTS

Conopeptide models

For each of the 16 known conotoxin superfamilies, three separate models based on signal, pro- and mature peptides were built, providing a total of 48 hidden Markov models and 48 generalized profiles. The models were named according to the superfamily and the region of the precursor that they targeted. Each model demonstrated very



Reload previous session

Session id:

Enter a sequence set in FASTA format

or

Paste your own sequences for profile matching :

```
>sp|P58807|CT6A_CONMR Chi-conotoxin CMrVIA OS=Conus marmoreus #T#
VCCGYKLCHPC
>sp|P69746|CA4A_CONOB Alpha-A(1-3) conotoxin OIVA OS=Conus obscurus
CCGVPNAACHPCVCKNTC
>sp|P0CAQ4|CA11A_CONAH [Conus achatinus] Alpha-conotoxin Ac1.1a    #A
MGMRMMFTLFLLVVLTTTVVSYPSDSASDGRDDEAKDERSDMYELKRNGRCCH
>sp|P0CH20|CO6D_CONPI Conotoxin pr6d OS=Conus parius #O superfamily#
YGNFPTCSETGEDCSAMHCCRSMTCRNNICAD
>sp|Q7Z091|CI1BD_CONRA [Conus radiatus] I1-superfamily conotoxin r11d
GCKKDRKPCSYHADCCNCCLSGICAPSTNWILPGCSTSTFT
```

or Upload a sequence file:

● Use CONCO combined classification models (HMMs and PSSMs).
○ Paste your own models (HMMs and/or PSSMs).

Your session id: 132642541524657
Keep your session id to reload result from this analysis later on.
Notice that your session id is the second part of the Excel file name.

Download a file for the submitted sequences

Download raw file of the result (brute ps_scan).

Download Excel file of the results.

	Combined_Prediction	HMM_Prediction	PSSM_Prediction	User_Class
sp|P58807|CT6A_CONMR	A	Unknown	A	T
sp|P0CAQ4|CA11A_CONAH	A	A	A	A
sp|P69746|CA4A_CONOB	A	Unknown	A	A
sp|Q7Z091|CI1BD_CONRA	Conflict : O1 % I1	I1	O1	I1
sp|P0CH20|CO6D_CONPI	Conflict : O1 % I3 % I1	Conflict : I1 % O1 % I3	Unknown	O SUPERFAMILY
sp|Q7M4K5|CI1BE_CONRA	I1	I1	Unknown	I1
sp|P0CH18|CO6B_CONPI	O1	Unknown	O1	O SUPERFAMILY
sp|A6YR34|CC13D_CONCL	O1	Unknown	O1	UNKNOWN
tr|D6C4L1|D6C4L1_CONCL	O1	Unknown	O1	UNKNOWN
sp|P58918|CO6B_CONCT	O1	Conflict : O2 % O1	O1	O1
sp|P0CB20|CONII_CONST	Unknown	Unknown	Unknown	UNKNOWN

Figure 1. *ConoDictor* input (background) and output (foreground) interfaces. The input interface provides a text area for amino acid sequence in FASTA format and areas for users to upload their own models. A test set is also provided and can be loaded via a simple click. The output interface provides detailed, self-explanatory tables grouped by analysis type. The combined prediction/classification is summarized under the 'General result' tab.

good discriminative abilities, with high sensitivity (~95%) and selectivity (~99%) [(7) and Koua *et al.*, unpublished data] . When tested using known conopeptide sequences, these models enabled extensive and reliable classification even between superfamilies containing mature peptides with high sequence similarities. The models provided good evidence of complementarity between signal, pro- and mature peptide sequences for superfamily determination, as well as complementarity between pHMMs and generalized profiles (Koua *et al.*, submitted).

ConoDictor input interface

ConoDictor accepts amino acid sequences in FASTA format as input. The sequences can be pasted in the prepared field or uploaded as a file from the user's computer (Figure 1). Sequences can be annotated with a predicted superfamily in the header between sharps (#), otherwise they are considered as 'UNKNOWN'. By default, the models built in the framework of the CONCO project are used to analyse the input sequences. However, users can also upload their own PSSMs and/or pHMMs. An annotated testing set (attached to a 'LOAD TEST DATA' button) is also available from the input interface.

Visualization interface

The *ConoDictor* output interface offers user-friendly tab views of matching outputs and predictions (Figure 1). The main tab provides combined prediction, as well as a summary of pHMM- and PSSM-based prediction. Detailed result tabs for pHMM- and PSSM-based predictions provide the number of sequence matches for each model, the position for each sequence/model match, and the related *e*-value and score of individual model match. Tab headers and table column names explain the results displayed. The result page is automatically updated until analysis results are available. An Excel file (.xls) and raw text versions (.txt) of all results can be downloaded. A session identifier is also provided, and the results can be accessed and visualized on the server for up to 3 weeks after submission or last viewing. A detailed help page provides clear explanations and screen shots of the most important tables of the analysis (http://conco.ebc.ee/ConoDictor_help.html).

CONCLUSION

ConoDictor is a web-based application, based on preliminary studies that established PSSM and pHMM complement each other for conopeptide identification and classification. Thanks to a user-friendly interface, *ConoDictor* provides an easy-to-use environment for classification of conopeptides into superfamilies based on their amino acid sequence. In view of the rapidly increasing number of new conopeptides being discovered by next-generation transcriptomic platforms, *ConoDictor* is a valuable bioinformatics tool for their classification and serves as a starting point for investigation of their functional characteristics.

ACKNOWLEDGEMENTS

We thank Estelle Bianchi, Daniel Biass, Nicolas Hulo and Christian Sigrist for expert assistance.

FUNDING

CONCO project [LSHB-CT-2007-037592, in part]: www.conco.eu funded by EU 6th Framework Programme (LIFESCIHEALTH); the European Regional Development Fund (Estonian Center of Excellence in Genomics); the Atheris Laboratories. Funding for open access charge: LIFESCIHEALTH [LSHB-CT-2007-037592].

Conflict of interest statement. None declared.

REFERENCES

1. Norton,R. and Olivera,B. (2006) Conotoxins down under. *Toxicon*, **48**, 780–798.
2. Stöcklin,R. and Vorherr,T. (2010) Venoms—a natural source for mini-protein drugs. *Pharmanufacturing Int. Peptide Rev.*, **(Sept. 2010)**, 44–46.
3. Olivera,B. and Cruz,L. (2001) Conotoxins, in retrospect. *Toxicon*, **39**, 7–14.
4. Kaas,Q., Westermann,J. and Craik,D. (2010) Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon*, **55**, 1491–1509.
5. Jones,R. and Bulaj,G. (2000) Conotoxins—new vistas for peptide therapeutics. *Curr. Pharm. Des.*, **6**, 1249–1285.
6. Kaas,Q., Yu,R., Jin,A.-H., Dutertre,S. and Craik,D.J. (2012) Conoserver: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.*, **40**, D325–30.
7. Laht,S., Koua,D., Kaplinski,L., Remm,M. and Stöcklin,R. (2012) Identification and classification of conopeptides using hidden Markov models. *Biochim. Biophys. Acta.*, **1824**, 488–49.
8. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, UK.
9. Johnson,L.S., Eddy,S.R. and Portugaly,E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.
10. Koua,D., Cerutti,L., Falquet,L., Sigrist,C.J.A., Theiler,G., Hulo,N. and Dunand,C. (2009) PeroxiBase: a database with new tools for peroxidase family classification. *Nucleic Acids Res.*, **37**, D261–D266.
11. Oliva,M., Theiler,G., Zamocky,M., Koua,D., Margis-Pinheiro,M., Passardi,F. and Dunand,C. (2009) PeroxiBase: a powerful tool to collect and analyse peroxidase sequences from Viridiplantae. *J. Exp. Bot.*, **60**, 453–459.