ApicoAP Complete Suite Client Software User Manual

ApicoAP-CS Client Software has the following capabilities:

- Prepares training sets with user assistance/supervision to be used in training classifiers that identifies apicoplast-targeted proteins with bipartite signal (ApicoTPs).
- Utilizes the prepared training sets to train ApicoTP classifiers that are customized to an apicomplexan species, which can then be used to identify ApicoTPs from the proteins of this species.

This manual will go over the steps involved in using these capabilities. A video demonstrating the use of the software can be found at <u>http://youtu.be/qvDDLFKUIcM</u>.

Generating training data

User is expected to set two options to start the train data generation process:

- Proteome option: Apicomplexan species whose proteome information was found through EupathDB are listed for user's convenience. User won't need to provide proteome information (all the proteins that are expressed by this species) for the listed species. User also had the option to upload his own proteome information by selecting the Other option. When this option is selected, user is expected to enter the species name (should be in two-word format, for example *Plasmodium Falciparum*) and upload a proteome file which should be in a special FASTA file format. Please see the Required file formats section for the details of this format.
- Seed set option: User should either use available seed sets or take the option to enter additional seed information in list file format. Please see the Required file formats section for the details of this format. Seed proteins should strictly be selected from proteins which are either experimentally confirmed to be an ApicoTP (as positive seed) or experimentally confirmed to localize to an organelle that rules out apicoplast-targeting (as negative seed).

Figure 1 shows the initial window on which user sets out these options. This window expands if user selects the *Other* option as the proteome option or if user selects to enter additional seed information. Once the required options are set, the **Generate training data** button will be visible. Clicking this button will initiate the train data generation process in the server and user will be kept informed with the status messages received from the server. Once the operation is complete, user will have the option to review/update the utilized seed training sets and the generated training sets. Figure 2 shows a snapshot of the software after an example run of train set generation.





2			
Proteor	ne option		
Babesia	bovis	10	e
Seed se	t option		
Use ava	ilable seed set	87	-
	Generate train	ning data	
Train D	ata Generatio	n	
Train D Download	ata Generatio	n DNE	
Train D Download Review	Data Generatio ling output files: DC v seed sets Re	n DNE view trainin	g sets
Train D Download Review	oata Generatio ling output files: DO v seed sets Re P Classifier tra	n DNE view trainin hining	g sets
Train D Download Review picoAl	Data Generatio ling output files: D(v seed sets Re P Classifier tra g parameters	n DNE view trainin hining	g sets
Review picoAl GA conv	Data Generatio Jing output files: DO v seed sets Re P Classifier training g parameters vergence parameters	n DNE view trainin hining er: 30	g sets
Review Review ApicoAl Trainin GA conv	Data Generatio ling output files: DO v seed sets Re P Classifier tra g parameters vergence paramet	n DNE view trainin nining er: 30	g sets

Figure 2: A snapshot of the software after train data generation is complete.

Review seed and training sets

After train data generation is complete, user is provided with the option to review, store and update the seed training and/or final training sets. Figure 3 shows an example data review window. Data generation logs and the details of the generated/used positive and negative seed/final training sets are displayed on this window.

ApicoAP-	CS Pipeline Client	
Log fil Log cre Started	<pre>extion date and time: 2013-04-02 17:53:17.332000 i parsing ApiLoc database to make up the POSITIVE seed training set Extracted 75 records parsing ApiLoc Extracted 18 additional records parsing available current literature set Now trying to resolve protein ids that are not recognized by EuPathDB</pre>	Save to file
Started	d parsing ApiLoc database to make up the NEGATIVE seed training set Extracted 421 records parsing ApiLoc	
Positiv	e set	Save to
1 2 3 4 5	TGME49_299810 T.gondii ApiLoc: localizes to apicoplast,cytoplasm cysteinyl-tRNA synthetase, PF3D7_1235600 P.falciparum ApiLoc: localizes to apicoplast,cytoplasm,mitochondrion serine hy PF3D7_1020800 P.falciparum ApiLoc: localizes to apicoplast dihydrolipoamide acyltransferase PF3D7_1467300 P.falciparum ApiLoc: localizes to apicoplast 1-deoxy-D-xylulose 5-phosphate re PF3D7_1333000 P.falciparum ApiLoc: localizes to apicoplast cochaperonin	tile Load from file Reload
6 7 4	PF3D7_0907900 P.falciparum ApiLoc: localizes to apicoplast peptidyl deformylase PF3D7_1440300 P.falciparum ApiLoc: localizes to apicoplast delta-aminolevulinic acid dehydra 🗸	original file
Negati	ve set	Save to
INDEX	IDENTIFICATION ORGANISM REASON DESCRIPTION	file
1	PBANKA 081990 P.Derghei Apiloc: localizes to nucleus nucleosome assembly protein	
3	TGME49 223920 T.gondi Abiloc: localizes to apical hypothetical protein	Load
4	TGME49 280680 T.gondii ApiLoc: localizes to nucleus hypothetical protein	Trom Tile
5	PKH_146980 P.knowlesi ApiLoc: localizes to apical reticulocyte binding protein, putative (fr	Reload
6	ETH_00010835 E.tenella ApiLoc: localizes to parasite plasma membrane gi 38602660 emb AJ58653	original
7	PF3D7_0628200 P.falciparum ApiLoc: localizes to apical protein kinase PK4	file
•		
	< Back to main window	
Note	e: You can edit the content of the boxes, or simply load your own files. Once you click the back button, you will be asked if you want to save	
whatev	er is displayed in the boxes as the seed/training data to be used in train data generation or classifier training. You can always go back to original	

Figure 3: An example data review window

User has the option to alter the positive and the negative sets, by either adding/deleting same formatted lines on the spot, or saving the original file, altering it using a preferred editor and loading the altered file back into the system. It's critical to comply with the existing format when altering these files. Make sure they are tab-delaminated and have the same column names as the original ones. **Save to file**, **Load from file**, **Reload original** buttons are meant to help the user in this altering process, if user prefers to perform altering. **Back to main window** button should be clicked in order to save the changes in data sets and/or in order to go back to the main window to proceed with the classifier training. Be aware that if you alter the seed training sets, train data generation should be repeated with the new seed sets, and therefore GUI will direct you to do that. If you alter the training sets, once you reach the main window and proceed with classifier training, the altered training sets will be used in training instead of the original ones.

ApicoAP Classifier Training

Once the train data generation process is completed successfully, user may proceed with the classifier training. 3 training parameters are expected to set by the user: GA convergence parameter, GA re-run amount, and GA seed for cross validation. ApicoAP training involves an optimization procedure conducted via a genetic algorithm (GA). Essentially, the purpose of this training process is to find the classifier that will perform the best job of identifying ApicoTPs which involves a search in the space of all possible classifiers according to ApicoAP model. Stopping condition for this search process is determined by the **GA convergence parameter**. Default value is set as 30. Higher values may increase the possibility to find better solutions, but this promise comes with the cost of increased execution time. Since each optimization run has the potential to miss the best possible solution, it is common practice to make a couple of re-runs. GA re-run amount determines this amount and the default value is set to 10. Higher values may lead to better solutions, again with the cost of increased execution time. Due to the randomized nature of this optimization and the performance assessment process we used (cross validation), starting with different seeds may result in slightly different results. Seeds are required by the pseudo-random number generation routines, and they are usually positive integers. Users are advised to note down the GA seed for cross validation for future reference, if they choose to change the default value, which is 1.

Once the required parameters for the classifier training are set, user may proceed using the **Train ApicoAP Classifier** button. Clicking this button will initiate the classifier training process in the server and user will be kept informed with the status messages received from the server. This operation may take up to an hour depending on the size of the training set. Once the operation is completed, the main window will expand further, like in the snapshot given as Figure 4, and user will have the chance to save the generated classifier for further use. This classifier can be integrated to ApicoAP software; please see the **Integration with ApicoAP** section for detailed instructions. The software automatically performs predictions with the generated classifier over the proteins of the given species that are predicted to contain signal peptide by SignalP. User is given the option to access to these predictions and save them to a file, if they desire. Figure 5 shows an example prediction display window.

🌠 ApicoAP-CS Pipeline Client 📃 💻 🗙					
Train data generate					
Proteome option					
Babesia bovis					
Seed set option					
Use available seed set 🔹					
Generate training data					
Train Data Generation					
Downloading output files: DONE					
Review seed sets Review training sets					
ApicoAP Classifier training					
Training parameters					
GA convergence parameter: 10					
GA re-run amount: 1					
GA seed for cross validation: 1					
Train ApicoAP classifier					
Training results					
Generated classifier is 100.00% accurate on the available training set.					
Save classifier View predictions					

Figure 4: A snapshot of the software after the classifier training is completed.

Predictions			470.00
ID	SP cleavage	Prediction result	
BOV_III000410	21	ApicoTP	
BOV_III000510	28	ApicoTP	
BOV III000530	36	ApicoTP	
BOV III000740	29	ApicoTP	
BOV III000830	22	ApicoTP	
BOV III001140	34	ApicoTP	
BOV III001170	34	ApicoTP	
BOV III001500	16	ApicoTP	
BOV III001660	16	ApicoTP	
BOV III001690	21	ApicoTP	
BOV 111002240	25	ApicoTP	
BOV 111002560	20	ApicoTP	
BOV 111002660	22	ApicoTP	
BOV 111002870	24	ApicoTP	
BOV 111002890	25	ApicoTP	
BOV 111003260	19	ApicoTP	
BOV 111003920	23	ApicoTP	
BOV 111003930	18	ApicoTP	
BOV 111004330	18	ApicoTP	
1) F
Save predictions	1		< Back to main window

Figure 5: An example prediction display window

Required file formats

Special FASTA format

We require the user to use a special FASTA format when inputting proteome information. This format is not much different than the well-known FASTA format, but restricts the description lines to only contain protein id information and nothing else. Here is an example protein sequence in this format:

>PF3D7_0815700 MIKKVHICLFIIFFYVIFLIHICKGIRLQNYKNERINNRHMLNTIRNNVSIYQNKHISNN NTKENKCNIMINYYDKSNIFCKSFLLSMEEKDNIKNIKKKIEEIYGIPLTLQEILYDNKK LENNITIQNIIKDKQIKILNLRLITILPHLFLQKDDDNNTNKRNDVSSSSSSSLYNNEY IKSNKRITYLKNKLTYYGYLTLLNEYKKLSHILENKKYILKNDDILESFKSFDQEFEKTL KNNNINLHKIKKEINKLKHIDKKKLLLRLEVDYPLMSNNLTKRIKQLFQYYMGDITTII KFSIFFYILYKYADYPNHIKKFFLYLSILFLISPFKPFYKFSHFLFFSVPNNILFSGFTN ILSASYQQILMCQ

List file format

We require the user to use list file format when inputting positive or negative seed training data. This format is nothing but a titled tab-delaminated text file format that is commonly used to list properties of a number of records where properties are delaminated by tab character and each records starts from a newline. Title line is required and following are column titles that the software expects to see:

- IDENTIFICATION (required): Id of a protein.
- ORGANISM: Organism of which this protein belongs to.
- DESCRIPTION: Description/annotation related with the protein.
- REASON: The reason why this protein is included as a seed training record.

Please make sure that these column titles appear exactly as they are listed here in your input file. Only the required field is enforced and any column named different than above will be ignored. Here is an example file in this format:

IDENTIFICATION ORGANISM DESCRIPTION REASON BBOV_III001660 Babesia bovis LytB protein Confirmed [Caballero11] JN114412 Babesia bovis acyl carrier protein Confirmed [Caballero11] TGME49_002440 Toxoplasma gondii hypothetical p. Confirmed [Sheiner]

Integration with ApicoAP

ApicoAP Version 3 software (which can be found at <u>https://code.google.com/p/apicoap/</u>) is an engine that interprets ApicoTP classifiers and provides the interface to run them on given protein sequences. If user places the classifier file generated by ApicoAP-CS Client Software under the Rules directory of ApicoAP V3 Software, next time ApicoAP is started, it will recognize the newly added classifier automatically.