SoCCer: A pipeline to identify classes of soluble proteins in chemical communication in insect genomes

Bhavika Mam
National Centre for Biological Sciences
Ramanathan Sowdhamini (mini@ncbs.res.in)
National Centre for Biological Sciences https://orcid.org/0000-0002-6642-2367

Method Article

Keywords: Insect olfaction, Genome Analysis, Sequence Searches, Olfactory binding proteins

DOI: https://doi.org/10.21203/rs.3.pex-1095/v1

Abstract

Insects require olfactory cues to perform a number of processes. With an increasing number of insect genomes that are getting realized, there is a need for an automatic method to classify proteins involved in olfaction into various types so that their cognate odorant assignment becomes possible and to eliminate errors due to misclassification into protein families with related functions and/ biochemical or structural properties. We propose a near-automatic computational protocol to identify and classify protein sequences of a given genome into odorant binding protein subfamilies and other major soluble protein families involved in insect chemical communication. Such a protocol can be used to identify and classify odorant binding proteins (OBPs) in large genomes but could also be applied to large number of other globular proteins as well.

Introduction

Insect genomes contain hundreds of OBPs important for olfactory processes (Venthur and Zhou, 2018). They are generally of helical fold and could be classified into subfamilies (like Minus-C, Plus-C, Classic and Atypical) depending on the length of OBP domain, cysteine topology and disulphide connectivity. Soluble proteins in chemical communication remain dissolved in the chemosensory organs in an insect. An extensive literature survey revealed that these typically include Odorant-binding proteins (OBPs), D7-proteins, Chemosensory proteins (CSPs) and Niemann-Pick C2 type proteins (NPCs). Protein families of OBPs, CSPs and NPC2 are secretory in nature with an N-terminal signal peptide (Pelosi et al, 2018).

NPC2 proteins share some structural and functional characteristics with odorant-binding proteins (OBPs) and chemosensory proteins (CSPs), soluble polypeptides involved in detection and delivery of semiochemicals in vertebrates and in insects (Vogt and Riddiford, 1981; Angeli et al., 1999; Pelosi et al., 2014, 2018;). As OBPs and CSPs, also NPC2 proteins present a signal peptide revealing their secretory nature. The OBPs of insects are made of α-helical domains connected by short unstructured loops and held together in a compact structure by three conserved disulphide bridges (Vogt et al., 2002). Proteins of both classes are extremely compact and stable, and this characteristic is also shared with NPC2 proteins.

The vertebrate OBPs are markedly different from those of insects in terms of amino acid sequence and three-dimensional folding (Pelosi, 1994; Tegoni et al., 2000), despite their common name and biological function. They belong to the family of lipocalins and have the typical β-barrel structure, made of eight β-sheets and a short α-helical segment (Tegoni et al., 1996; Briand et al., 2000). Moreover, the three-dimensional folding of NPC2 (Vanier et al., 2004; Ishida et al., 2014) proteins resembles the β-barrel of lipocalins, while the presence of six conserved cysteines paired in three disulphide bridges is reminscent of insect OBPs.

Reagents

Equipment

Procedure

Workflow

The workflow of our approach is as described below (**Figure 1**). It may be broadly divided into genomics, proteomics, machine-learning and evolutionary sub-approaches.

Datasets were manually constructed and curated for seven protein families majorly involved in insect chemical communication. In order to build in-house datasets, well-annotated and curated protein sequences were mined from literature, UniProt and SwissProt searches. The protein 'classes' considered were Classic, Minus-C, Plus-C, Atypical, D7, CSP and NPC2. A control dataset was prepared that consisted of insect proteins that did not belong to any of the above classes. Unique and non-redundant protein sequences were retained.

Features were computed to contribute to bit-wise information about each class-specific sequence. A feature matrix was derived that comprised 36 bits per sequence in a given class and necessary steps were taken to minimize or nullify bias due to imbalanced datasets wherever applicable.

Machine learning models were built using sklearn library in Python v3.5 environment.

Various classifiers were used on the data. Data was split into training and testing using the 80:20 ratio. MinMax scaler was used to normalize the data. Data was split into training and testing for X and Y each. Performance was evaluated using measures of accuracy, precision, recall, f1-support, and Mathews' correlation coefficient (MCC). The performance was also visualized using a confusion matrix plotted using Python libraries.

A total of 17 features were selected to be computed per sequence in each class. These are as follows-

a. Position-Specific Score Matrix (PSSM)

b. Accessible Surface Area (ASA)

c. Phi and Psi torsion angles (dihedral angles)

d. Secondary structure

e. Disorder scores of protein (two scores)

f. Number of Cysteines

g. Length of protein

h. Molecular weight

i. Aromaticity

j. Stability score

k. Isoelectric point of sequence (pI)

l. Molecular coefficient of reduction extinction

m. Molecular coefficient of disulphide

n. Entropy of sequence

o. Number of globular domains in the sequence

p. Residue Adjacency Matrix of Cysteines (RAM)

Insect genomes were collected from NCBI, Ensembl and VectorBase. Genomic alignments were obtained using protein datasets representative of each of the seven classes as query.

These were then used to obtain unique gene models and non-redundant protein sequences using a combined approach. A four-pronged approach was then used to score list of predicted proteins on the presence of cysteine topology, PBP/GOBP domain, length cut-off as well as the presence of signal peptide using predictive and in-house scripts. Due to the nature of predictions, scoring for only few of these criteria does not guarantee the sequence to be an OBP due to biochemical similarities across various classes of insect proteins. Hence, sequence passing the cut-off score were tested using optimized machine learning models trained to discriminate amongst seven major types of protein families/sub-families mediating insect chemical communication.

Phylogenetic analysis of predicted classes/sub-families coupled with analyses of motifs, domains and co-occuring domains identified yielded unique insights into the evolution and possible functional significance of OBP and other protein families involved in communication in insects.

User-end inputs-

i.**First stage.** Two modes in input are possible- either A or B, or both A and B can be provided as input.

A. An **input file** with a list of protein sequences in FASTA format with special amino acids 'B', 'J' and 'X' removed. The header of every fasta sequence will be retained upto the first twenty characters.

B. An input folder with genome of the organism of interest in fna format, genomic alignment output file and a file with query proteins of interest from any organism.

ii.**Second stage.** An **input folder** with feature files corresponding to each protein sequence in the output fasta file of the first stage (i). These feature files are to be correctly named with the prefix same as that of the unique header ID of the protein sequence.

iii.The features can be computed using the code provided in Github repository https://github.com/bhavikamam/SoCCer/featurizer.py after appropriate changes in path have been made as per the requirement of the user. The 'readme' file will contain instructions for the software to be installed locally. The output file at this step has to be provided as input to the Python script for further identification and classification.

Software, scripts

The entire pipeline and codes for the same have been uploaded in a repository on Github titled 'SoCCer' by the username 'bhavikamam'.

Advantages

Due to the similarities across some protein families in combination of i. type of domain, ii. sequence, iii. structural and/ iv. functional properties, identifying a protein correctly as an OBP subtype is quite challenging. Our methodology derives a non-linear relationship from all the essential feature information provided and identifies whether a given protein sequence as one of the major families in insect chemical communication and accordingly classifies it further.

Comparison with other methods

There are a number of computational pipelines that combines gene prediction tools, starting from handling raw sequencing data, gathering transcript evidence and so on. However, such pipelines are not specialised for particular protein families. There are also sequence search techniques, but they mostly use amino acid exchange information to recognise homologous sequences and do not absorb or

incorporate features such as secondary structures or cysteine connectivities in the identification or classification of proteins into subfamilies.

Currently to our knowledge, there is no other existing computational technique that classifies protein into one of the OBP subfamilies or other families involved in insect chemical communication using an optimized ML-based approach.

Application

The technique is useful in identifying and classifying the protein family/ or sub-family of insect chemical communication to which a novel and/or existing protein sequence predicted would belong by using a combinatorial genomics, machine learning and evolutionary-based approach.

Target audience

The major target audience of interest for this methodology would comprise entomologists and evolutionary biologists. Olfaction being a major sensory modality for insects, entomologists would be highly interested to learn more about the olfactory system in insects across orders as well as the types of proteins participating in the process of olfaction. Near-automatic classification of different types of OBPs not only provides them an advantage in terms of functional annotation but it would also help them to recognize subfamilies for designing experiments. For example, pheromone receptors could be taken up for detailed biochemical studies by cloning putative pheromone-binding proteins predicted by this methodological pipeline.

Use of technique

This technical is useful because the scientific community has to otherwise typically engage in laborious manual process of aligning widely different types of OBPs since they differ in their length and disulphide connectivity and hence have to be examined manually to then be classified. Instead, we hereby, present a computational scheme by which we apply a novel combined approach to perform these classifications computationally. This can be used for few other non-OBP families as well.

Limitations

The limitation of the technique is that as and when we come across newer OBP sub-types as we mine more insect genomes, we may come across unexpected or surprising deviations from these disulphide bonds patterns by which automatic recognition of these subfamily types would be challenging, so we expect that the protocol also must be open to adapt itself to increasing knowledge that comes from merely looking at more number of insect genomes as we go by.

Troubleshooting

Troubleshooting

The presence of large insert regions could hamper accurate identification and classification. While handling protein sequences of insect genomes, the user can remove or trim huge indel regions with respect to secondary structure and re-submit. Similarly, N-terminal and C-terminal overhangs can be submitted to Pfam so that proper domain boundaries may be recognized.

Time Taken

Time taken

The estimated time taken for the pipeline depends on the size of the genome and that of the available predicted protein sets from curated public repositories. For a file size of 1Gb, obtaining genomic alignments with query proteins and gene predictions can be accomplished within two hours using a GPU-based cluster. Generating feature files and input feature matrices is done using in-house Python scripts and takes approximately 5 hours for ~1000 protein sequences. Training neural networks took 10-12 hours whereas testing protein sequences takes less than 30 minutes.

Anticipated Results

The pipeline analyzes diverse insect genomes to identify and discriminate among soluble proteins belonging to known classes of chemical communication, namely, OBPs, CSPs, NPC2 and D7.

The stepwise output of the pipeline first results in query-specific genomic alignments and predicted genes which are filtered along with proteins obtained from public repositories. The shortlisted set of predicted proteins are further subjected to feature matrix generation (**Figure 2A**) that is used as input to the optimized and trained neural network model (**Figure 2B**). The output of the network is a regression value indicating the likelihood of the sequence belonging to each of the classes.

As an example, in case of Hymenopteran *Apis florea* genome, the pipeline revealed 22 novel OBPs with 9 partial sequences either at the N-terminus, C-terminus or both with an average exon number of 5. Out of the 22 complete sequences annotated, 13 were predicted as Classic and 9 as Minus-C. The Classic OBP subfamily clade had expanded to Minus-C OBPs in honeybee and few other insect orders (**Figure 2C**).

Thus implementation of this pipeline streamlines gene and protein sequence prediction of classes important for insect chemical communication. This is done taking in account size, conserved cysteine pattern, disulphide spacing as well as secondary structure.

References

References

Angeli S, Ceron F, Pelosi P. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from Schistocerca gregaria. European Journal of Biochemistry, 1999; 262 (3) : 0014-2956

Briand L, Nespoulous C, Perez V, Rémy JJ, Huet JC, Pernollet JC. Ligand-binding properties and structural characterization of a novel rat odorant-binding protein variant. European Journal of Biochemistry, 2000; 267(10):3079-3089.

Ishida Y, Tsuchiya W, Fujii T, Fujimoto Z, Miyazawa M, Ishibashi J,

Matsuyama S, Ishikawa Y, Yamazaki T. Ant NPC2 mediating worker chemical communication. Proceedings of the National Academy of Sciences, 2014; 201323928.

Larter NK, Sun JS, Carlson JR. Organization and function of Drosophila odorant binding proteins. Elife, 2016; 5: e20242.

Pelosi P. Odorant-binding proteins. Critical Reviews in Biochemistry and Molecular Biology. 1994; 29(3):199-228.

Pelosi P, Calvello M, Ban L. Diversity of Odorant-binding Proteins and Chemosensory Proteins in Insects, *Chemical Senses*, 2005, 30 (1): 291–292

Pelosi P, Iovinella I, Felicioli A, Dani FR. Soluble proteins of chemical communication: an overview across arthropods. Frontiers in Physiology, 2014; 5:320

Pelosi P, Zhu J, Knoll W. Odorant-Binding Proteins as Sensing Elements for Odour Monitoring. Sensors (Basel), 2018; 18(10): 3248.

Tegoni M, Pelosi P, Vincent F, et al. Mammalian odorant binding proteins. Biochimica et Biophysica Acta, 2000; 1482 (1-2): 229-240.

Vanier MT, Millat G. Structure and function of the NPC2 protein. Biochimica et Biophysica Acta. 2004; 1685(1-3): 14-21.

Venthur H, Zhou J. Odorant Receptors and Odorant-Binding Proteins as Insect Pest Control Targets: A Comparative Analysis. Frontiers in Physiology 2018; 9, 1664-042.

Vogt R, Rogers M, Franco M, Sun M. A comparative study of odorant binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in Manduca sexta (Lepidoptera) and the organization of OBP genes in Drosophila melanogaster (Diptera). Journal of Experimental Biology 2002; 205: 719-744

Vogt R., Riddiford L. Pheromone binding and inactivation by moth antennae. Nature 1981: 293, 161–163.

Zhu J, Guo M, Ban L, et al. Niemann-Pick C2 Proteins: A New Function for an Old Family. Frontiers in Physiology 2018;9:52.

Acknowledgements

BM has been supported by Tata Trust Fellowship. RS would like to thank her JC Bose Fellowship (SB/S2/JC-071/2015) from the Science and Engineering Research Board, India. The authors would like to thank NCBS (TIFR) for infrastructural support.

Figures

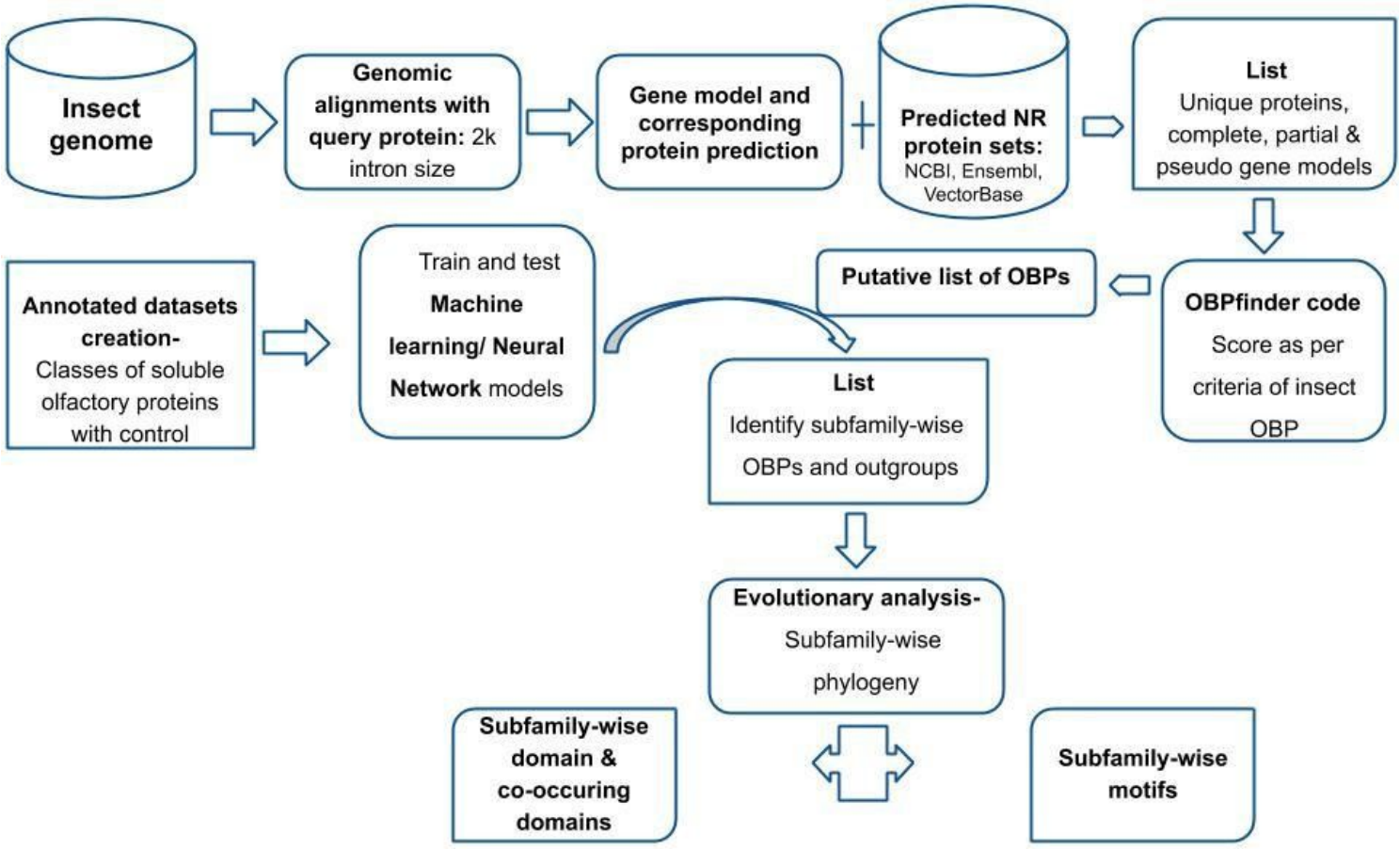

Figure 1

SoCCer: a combined method using genomic, protein-based, ML and evolutionary approaches.

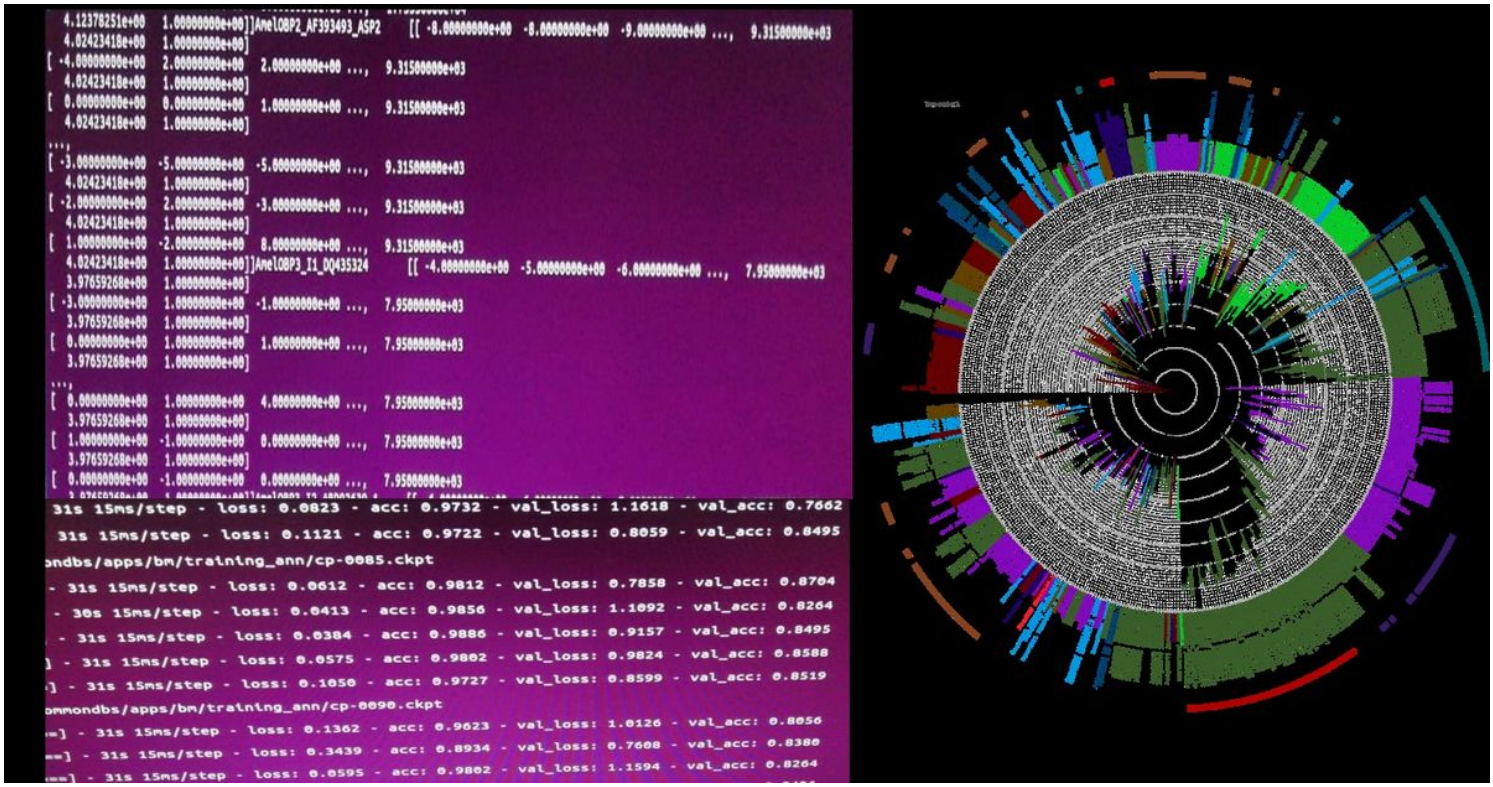

Figure 2

Sample results for SoCCer pipeline A. Feature matrices for protein sequences of Hymenopteran genome B. Performance of neural network during training C. Phylogeny of predicted odorant-binding proteins across 11 insect orders.