Identifying GPCRs in the Genome of the Sand Fly *P. papatasi* using Ensemble*

RJ Nowling[1], M Wadsworth[2,3], JL Abrudan[2,3], DA Shoue[2,3], B Abdul-Wahid[1], GM Stayback[2,3], FH Collins[2,1,3], MA McDowell[2,3], and JA Izaguirre[1,3]

[1]Computer Science & Engineering, University of Notre Dame [2]Biological Sciences, University of Notre Dame [3]Eck Institute for Global Health, University of Notre Dame

Abstract

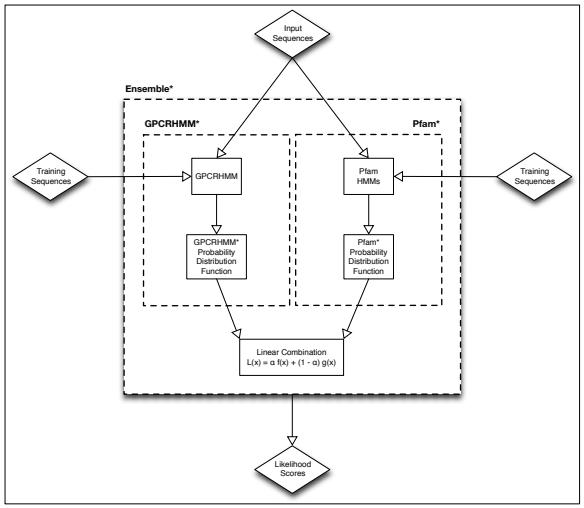
G-protein Coupled Receptors (GPCRs) are a class of seven transmembrane (7TM) proteins involved in signal transduction [1, 2] that respond to a diverse range of stimuli. A sign of their importance in regulating many physiological processes, GPCRs are relatively abundant in metazoan genomes (1% of the *Drosophila melanogaster* and 1.6% of the *Anopheles gambiae* genomes [3, 4]). Due to their physiological importance, abundance, and specificity, GPCRs are attractive targets for the development insecticides, repellents, and other products for the control of vector populations [5-7].

In our previous work [8], we evaluated existing GPCR classifiers on vector peptide sequences, showing that their accuracy and sensitivity are less than desired. In response, we developed Ensemble*, a novel GPCR classifier tuned for arthropod genomes. Ensemble* was validated on test sets of known GPCRs and applied to the vector species *Aedes aegypti*, *An. gambiae*, and *Pediculus humanus*, resulting in 52 novel hits. Validation of the hits confirmed 19 of the predictions as GPCRs and gave evidence that another 11 hits were putative GPCRs.

The genome of the sand fly *Phlebotomus papatasi,* a vector of leishmaniasis and pappataci fever, has recently been sequenced and assembled [9]. Ensemble* was run on the *P. papatasi* genome peptide translations, resulting in 142 hits. Subsequent validation with BLAST against the NCBI nr-database [10] and ScanPROSITE [11] resulted in the identification of 97 confirmed and 7 hypothetical GPCRs.

Ensemble* GPCR Classifier

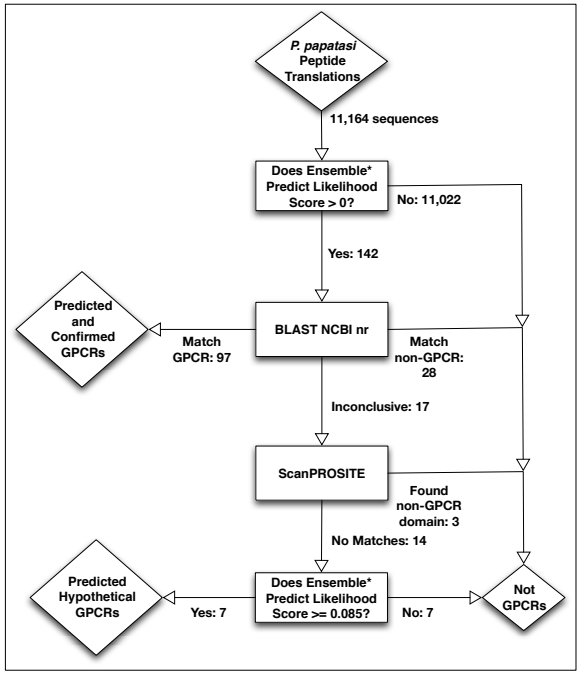
Ensemble* combines the prediction capabilities of GPCRHMM [12, 13] and the Pfam Clan A GPCR Hidden Markov Models [14]. Discrete functions are used to map scores to likelihood values between 0 and 1, which are combined via a linear weighting to produce an overall likelihood score between 0 (not a GPCR) and 1 (a GPCR) for each input sequence.



References

[1] Pierce, et al. *Nat Rev: Mol Cell Bio* 2002, **3**(9):639-650.
[2] Broeck *JV. Arch Insect Biochem Physiol* 2001, **48**:1-12.
[3] Adams, et al. *Science* 2000 **287**(5461):2185-2195.
[4] Hill, et al. *Science* 2002 **298**(5591):176-178.
[5] Grimmelikhuijzen, et al. *Pest Manag Sci* 2007, **416**(2006):413-417.
[6] Justice, et al. *Bioassays* 2003, **25**(10):1011-1020.
[7] Meyer, et al. *PLoS Negl Trop Dis* 2012, **6**:e1478.
[8] Nowling, et al. *Parasites & Vectors* 2013, **6**:150.
[9] McDowell, et al. Whitepaper 2006.
[10] Pruitt, et al. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.
[11] Gattiker, et al. *Appl Bioinformatics* 2002, **1**(2):107-108.
[12] Sgourakis N, et al. *Bioinformatics* 2005, **21**(22):4101-4106.
[32] Wistrand M, et al. *Protein Sci* 2006, **15**(3): 509-21.
[14] Punta, et al. *Nucleic Acids Res* 2012, **40**(Database issue): D290-D301.
[15] Krogh, et al. *J Mol Bio* 2001, **305**(3):567-580.

Identification and Validation of *P. papatasi* GPCRs



Conclusions and Future Work

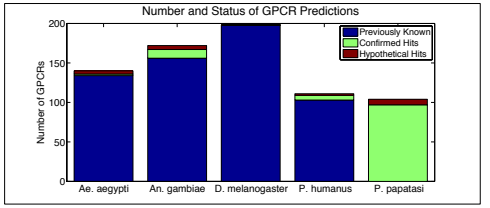
Statistics for the *P. papatasi* GPCRs were compared with those for *Ae. aegypti*, *An. gambiae*, *D. melanogaster*, and *P. humanus*. The total number (known, confirmed hits, and hypothetical hits) of GPCRs and relative abundance in the respective genomes agree, suggesting that we have likely found a majority of the *P. papatasi* GPCRs.

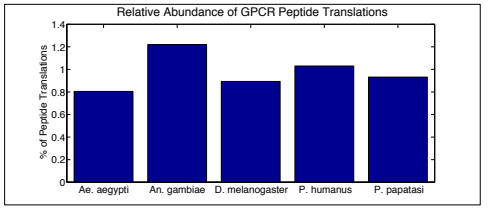
Analysis of the GPCR peptide translation sequence length distributions indicates that the *P. papatasi* GPCRs (287 aa on average) sequenes are significantly shorter than those of the other organisms (*Ae. aegypti* – 443 aa, *An. gambiae* – 565 aa, *D. melanogaster* – 692 aa, and *P. humanus* – 481 aa). TMHMM [15] was used to predict the number of TM helices in the GPCRs for each of the organisms. Comparison of the resulting distributions indicate that many of the *P. papatasi* sequences have fewer TM helices than the expected seven and the GPCRs in the other organisms. Consequently, the *P. papatasi* sequences may be incomplete due to incorrect assembly or gene prediction, requiring additional curation before further analysis can be done.

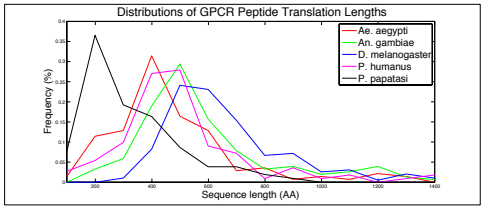
Despite the potential issues, Ensemble* was able to successfully identify the GPCRs in the novel genome.

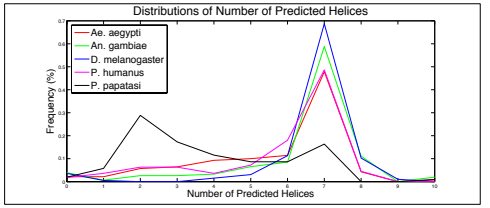
In the future, we will use Ensemble* to identify GPCRs in the *Lutzomyia longipalpis* sand fly genome. We will compare the GPCRs from *P. papatasi* and *L. longipalpis* with those of the *Glossina morsitans* tetse flies and *Rhodnius prolixus* kissing bugs with the goal of relating differences to variations in hosts, parasites, and physiological mechanisms. Furthermore, species-specific GPCRs could prove to be attractive targets for the development of insecticides for population control.

Comparison of *P. papatasi* Predicted GPCR Distributions with Other Arthropod GPCRs









Acknowledgements

This project was supported by an award from the Department of Defense Telemedicine and Advanced Technology Research Center (#W81XWH-10-1-0085). RJN gratefully acknowledges funding from a Department of Education GAANN Fellowship.