From sequence and structure to function: Odorant-binding proteins

A THESIS TO BE SUBMITTED TO

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

BHAVIKA MAM

UNDER THE GUIDANCE OF

PROF. RAMANATHAN SOWDHAMINI (GUIDE)

NCBS, BANGALORE, INDIA

PROF. BERNARD OFFMANN (CO-GUIDE)

UNIVERSITY OF NANTES, NANTES FRANCE

NCBS, BANGALORE

JUNE, 2021

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013

BENGALURU - 560064

CERTIFICATE

This is to certify that the work incorporated in this thesis **"From sequence and structure to function: Odorant-binding proteins"** submitted by **Bhavika Mam** was carried out under my supervision. No part of this thesis has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Research Supervisor

Prof. Ramanathan Sowdhamini
Professor, Computational Approaches to Protein Sciences
National Center for Biological Sciences
Tata Institute for Fundamental Research
GKVK, Bangalore - 560065

Research co-supervisor

Prof. Bernard Offmann
Professor, Unité Fonctionnalité et Ingénierie des Protéines (UFIP)
UMR CNRS 6286, Université de Nantes
June, 2021

Abstract

Title: From structure to function: odorant binding proteins

Abstract:

The olfactory system can detect and discriminate among environmental odorants that play a crucial role in various insect behaviours. Among them, odorant-binding proteins (OBPs) have been considered important components of the olfactory apparatus in recognising odorants. OBPs undergo conformational changes due to pH changes and ligand binding altering their interactions with odorants.

My thesis aims to understand the function of odorant-binding proteins in the context of the sequence, structure and evolution across insect orders and mammals. A pipeline was constructed to integrate genomics and machine-learning approaches to predict whether a protein sequence belongs to the OBP subfamily or other classes. In the hemolymph of malaria-causing female *Anopheles gambiae*, OBP1 and OBP4 may form heterodimers to facilitate perception of indole, different from those of the individual proteins, thus increasing the number of detectable odours. Our work showed that a decrease in pH is associated with a concomitant reduction in the inter and intra-monomeric dynamic fluctuations of the indole-bound heterodimer. Further work in my thesis has identified a set of naturally derived compounds as inhibitors to heterodimeric OBP-indole and monomeric OBP-indole complexes in mosquitoes through computational virtual screening and molecular dynamic simulations. Furthermore, protein-ligand docking, molecular dynamic (MD) simulations and thermal binding affinity analysis of OBP isoforms from the buffalo nasal epithelium revealed four residues (Phe69, Phe104, Asn118 and Phe134) from OBP1a, contributed strong binding affinities towards two sex pheromones, specifically oleic acid and p-cresol. The predicted protein sequences further helped in understanding evolutionary relationships among various subfamilies of OBPs.

Chapter 1

Chapter 1 provides an introduction and background to the modality of olfaction in insects and mammals. It introduces the overall mechanism and the players involved. It then elaborates on odorant-binding proteins in both systems and their structural differences, existing techniques and, known and unanswered questions.

Chapter 2

Chapter 2 of the thesis describes the pipelines built using machine learning and AI - SoCCer and DEELIG. SoCCer identifies and discriminates among subfamilies in odorant-binding family and other families in insects involved in insect chemical communication. DEELIG is an approach to predict protein-ligand binding affinity using a deep learning methodology involving convolutional neural networks.

Chapter 3

Chapter 3 of the thesis involves identifying orthologs of OBPs across insect species to understand evolutionary relationships among OBP subfamilies and across insect orders through phylogeny. It also involves domain analysis of odorant-binding proteins using bioinformatic approaches.

Chapter 4

Chapter 4 of the thesis addresses the influence of pH and indole on the heterodimerization of OBP1 and OBP4 from female ***Anopheles gambiae*** through protein-protein docking, protein-ligand docking and molecular dynamic simulations. AgamOBPs may form heterodimers with novel binding properties, different from those of the individual proteins, thus increasing the number of detectable odours. Specifically, AgamOBP1 and AgamOBP4 hypothesized to facilitate perception of indole through heterodimerization. In order to understand how these two OBPs interact in the presence of indole and investigate the possibility of pH-dependent heterodimerization, crystal structures of AgamOBP4 at varying pH (provided by collaborators) were applied for modelling protein-protein interaction and molecular dynamic simulation studies.

Chapter 5

Chapter 5 of the thesis identifies natural repellents and metabolites against the heterodimeric odorant-binding proteins of female ***Anopheles gambiae*** using docking, virtual screening and molecular dynamic simulations.

Chapter 6

Chapter 6 of the thesis addresses the characterization of ligand-binding to isoforms of nasal odorant-binding protein in ***Bubalus bubalus***, Indian water buffalo, using modelling, protein-ligand docking, and molecular dynamics approaches.

Chapter 7

Chapter 7 of the thesis summarizes all the previous chapters and future directions of the thesis to facilitate understanding the role of OBPs in a broader context.

Chapter 8

Chapter 8 contains the bibliography for Chapters 1 to 7. Research articles, literature and other references cited in previous chapters have been listed here in alphabetical order.

List of Publications/Projects Related to this Thesis

1. Chidhambaram Manikkaraja*, **Mam Bhavika***, Randhir Singh, Balasubramanian Nagarathnam, Geen George, Akash Gulyani, Govindaraju Archunan & Ramanathan Sowdhamini (2020) Molecular and functional characterization of buffalo nasal epithelial odorant binding proteins and their structural insights by *in silico* and biochemical approaches, Journal of Biomolecular Structure and Dynamics,
   DOI: 10.1080/07391102.2020.1854117

2. Ahmed, A., **Mam, B.***, & Sowdhamini, R. (2021). DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. Bioinformatics and Biology Insights. DOI: 10.1177/11779322211030364

3. **Bhavika Mam**, Ramanathan Sowdhamini (2020). SoCCer: A pipeline to identify classes of soluble proteins in chemical communication in insect genomes. DOI: 10.21203/rs.3.pex-1095/v1

4. **Bhavika Mam***, Katerina E. Tsitsanou*, Panagiota G.V. Liggri Jarjapu Mahita, George Leonis, Christina E. Drakou, Manthos Papadopoulos, Philippe Arnaud, Bernard Offmann, Dimitrios Fessas, Ramanathan Sowdhamini and Spyros E. Zographos (2021). Influence of pH on indole-dependent heterodimeric interactions between mosquito odorant-binding proteins AgamOBP1 and AgamOBP4. Manuscript in preparation.

5. **Bhavika Mam**, Snehal Karpe and Ramanathan Sowdhamini (2021). Genome-wide survey of odorant binding proteins in Hymenopteran *Apis florea*. Manuscript under review.

6. **Bhavika Mam**, Shailya Verma, Ramanathan Sowdhamini.Virtual screening of natural repellants against AgamOBP heterodimer. Manuscript in preparation.

7. a. Integrated genome-wide, machine learning and phylogenetic approach towards identifying soluble proteins in insect chemical communication. Manuscript in preparation. **Bhavika Mam** with Prof. R. Sowdhamini's group.
   b. Homology-based approaches to identify OBPs across insect taxa. **Bhavika Mam** with Prof. R. Sowdhamini's and Prof B. Offmann's groups.

List of other publications/projects from lab

Mahantesha B.N. Naika, Sathyanarayanan Nitish, Radha Sivarajan Sajeevan, Bhattacharyya Teerna, Ghosh Pritha, Iyer Meenakshi, Jarjapu Mahita, Joshi Adwait G., K. Harini, K. Mohamed Shafi, Kalmankar Neha, Karpe Snehal D., **Mam Bhavika**, Shaik Naseer Pasha, Ramanathan Sowdhamini. Exploring the Medicinally Important Scondary Metabolites Landscape Through the Lens of Transcriptome Data in Fenugreek (Trigonella Foenum Graecum L.) Research Square; 2022. DOI: 10.21203/rs.3.rs-1441723/v1.

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

DECLARATION BY THE CANDIDATE

I declare that this thesis entitled "**From sequence and structure to function: Odorant-binding proteins**" submitted for the award of Doctor of Philosophy to THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY, Bengaluru, is my original work, conducted under the supervision of my guide **Prof. R. Sowdhamini** (and co-guide, **Prof. Bernard Offmann**). I also wish to inform that no part of the research has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged

I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Place: Bengaluru **Signature of the Candidate**

Date: 14th June, 2021

Name of candidate: Bhavika Mam
Reg. No.: 20916030149
June, 2021

Acknowledgement

I express utmost gratitude to my PhD advisor, guide and mentor, Prof. Ramanathan Sowdhamini. In the JRF, rotation and PhD periods, Ma'am has been supportive and shown concern even in more challenging times. Her simplicity, high thinking, knowledge, stoicism and strength helped me immensely towards constant betterment and in my professional goals. Her mentoring, motivation, faith and tremendous patience with her students enabled me to develop and explore ideas creatively. It resulted in the fruition of several projects. Scientific comments and constructive discussions with my advisor have been insightful and valuable to my work and improved me to critically work and improve paper writing and presentation skills. I was given the opportunity by my advisor to train and work with several students in my PhD and collaborate with scientific groups. I was also given opportunities to volunteer and participate in meaningful conferences and workshops, and perform some of my PhD work at the University of Nantes, France. Ma'am also encouraged me to present my work to her postdoc advisor, Professor Tom Blundell (Cambridge University, UK), during his visit to NCBS and interact at length with Professor Shoba Ranganathan (Macquarie University, Australia) during her visit and stay at NCBS. I am grateful to be given the opportunity to work on projects across bioinformatics in genomics, transcriptomics, protein sequence and structural bioinformatics, phylogeny and machine learning, and AI. I thank Ma'am for taking the initiative towards implementing chalk-talks, and for encouraging me to organize and moderate the weekly journal club for bioinformatics for the academic year 2018-19.

Ma'am's personal and professional journey in science, as shared with lab members in the innovative Career Talks sessions, has been very inspiring and motivating. Herself an accomplished Carnatic vocalist and instrumentalist (Veena), Ma'am's belief that students should engage meaningfully in creative pursuits to complement research has been immensely encouraging during my years at NCBS. I admire the immense discipline and organizational skills of Ma'am. I thank Ma'am for recommending me to the team of Avesthagen Pvt. Ltd., Bangalore, for my first research-based industry job in December 2020. Personally and professionally, I am ever grateful for her able and compassionate leadership.

I am also grateful to my thesis committee members (TCM) at NCBS. In addition to my advisor, my TCM comprises Prof. Shannon Olsson and Prof. Axel Brockmann. My thesis committee members have always been supportive, encouraging, positive and full of valuable feedback. Right from my initial years as a PhD student, I have had the opportunity for numerous scientific discussions on OBPs, ecology, evolution and insect neurobiology with my TCM that increased my appreciation and curiosity towards my own work. I am very thankful to them for devoting time, patience and constructive feedback at my poster presentations, insightful one-on-one discussions, and committee meetings, and in sharing/ suggesting scientific literature with me that has been helpful in my journey.

I am sincerely thankful to Professor Bernard Offmann (University of Nantes, France) for agreeing to host and mentor me for some of my thesis work in his lab, agreeing to co-guide me, and for his comments and discussions that have been useful. I am grateful to him, and all his lab members at UFIP, France for the support and kind hospitality extended towards me during my time in Nantes.

I am thankful for the wonderful collaborative teams that I had an opportunity to work with. Detailed manuscript discussions with Professor Spyros Zographos and his team at NHRF, Greece, have proved valuable to the progress of the exciting work on heterodimeric OBPs. I am also thankful to Professor G. Archunan at BDU (Tamil Nadu) and his student, Mr Manikkaraja C for their support and positive comments on the mammalian OBP work. I thank Dr Alexander G. de Brevern, INSERM, Paris, and Dr. Rajas Rao for providing valuable inputs and feedback on work I had done so far.

My family has been there for me genuinely and unconditionally all my life, especially during my PhD. Their support, care and selfless sacrifices have been invaluable, and they have never let me worry about anything. I thank my father, Mr Jawahar Lal Mam, a marine mechanical engineer retired from the Indian Navy for his inculcation of discipline, professionalism and sincerity in me, and for motivating scientific discussions, his child-like curiosity and enthusiasm and guidance. I thank my mother, Mrs Pawanlata Kaul Mam, senior executive banker, for her inputs, exemplary balance of professional and family life, and in inculcating discipline, optimism, faith, strength, integrity and tranquility in me.I thank them both for their utmost patience, selfless dedication, knowledge and wisdom. I thank my younger sister, Ambika, a dedicated computer engineer, for her continued love and sincerity in encouraging and genuinely supporting me through the whole

journey. I thank her for her unconditional support, patience and wits. I remain humbled and grateful beyond words to my loving and caring grandparents, parents, and Ambika for all of this and more. I am thankful to my family and teachers during schooling who taught me science, math, programming, and other subjects. I got married to a wonderful person in November, 2021. I thank my husband Anshu who has been an amazing pillar of love, care and support for me during my postdoc and the several months leading to my PhD defense and related formalities. I'm thankful for his presence in our new life together and for his prudence, positivity and good humor. I thank the Almighty for everything.

There are many people whose efforts I would like to sincerely acknowledge in the journey of my PhD. I express my gratitude to them all.

I am sincerely thankful to Mrs K. Visalakshi, NCBS and Mr Ravikumar G, Assistant Registrar - Academics, TDU to enable an efficient administrative process throughout. I sincerely thank Ravikumar Sir for being genuinely supportive in all matters related to our admission, funding, submissions, review meetings, coursework, workshops and all related queries throughout the PhD program. I am thankful to the Tata Trusts Foundation for funding my research work at NCBS. I thank the office of the Vice-Chancellor, Dean, Registrar, staff at TDU and the excellent committee reviewing Tata Trusts fellowship for their positivity and support. I also thank the offices of the Dean and the Academic Dean at NCBS for their support throughout. I thank Upi and other faculty for their constructive feedback and questions during my Annual Work seminars at NCBS. I also thank NCBS IT help, Chakrapani Sir, staff in electrical, administrative, accounts, instrumentation and all other departments for their support. I am thankful to NCBS and the BLiSc community that provides excellent scientific and cultural exposure and a great support system including world-class infrastructural, administrative, technical, food, medical and lodging facilities to its staff and students. I thank TDU, NCBS, JNCASR, ICTS, IISc, IISER, ACTREC and Schrodinger, Inc. teams for conducting scientific lectures, and workshops or popular outreach lectures. I am blessed to have pursued PhD in the culturally vibrant and wonderful city of Bengaluru in the beautiful state of Karnataka.

I thank Professor Upinder Bhalla, Professor Vatsala Thirumalai, late Professor K.S. Krishnan, and all the members of my GS-interview panels at NCBS. I have had the privilege of also being inspired by the presence of late Professor Obaid Siddique on campus at NCBS. I thank late Professor N. Srinivasan for inspiring us students with his kindness and scientific curiosity. I thank Prof Mukund Thattai and Dr Minhaj Sirajuddin for their good wishes. I thank the journal clubs (JC) at NCBS and I enjoyed actively participating in many- Bioinformatics, EEJC (evolution and ecology), DBJC (developmental biology) and NBJC (neurobiology). I thank the course instructors and evaluators at IBB, Pune University (SPPU), NCBS and TDU. I enjoyed several courses such as TDK (at TDU), courses at NCBS such as Advanced Bioinformatics, Neurobiology, Evolution, Cell signalling, Cell biology, Biophysical chemistry, Stochastic processes (randomness in biology), and Public health. I thank my teachers at IBB, who taught me bioinformatics, biochemistry, statistics, genomics and microarray data analysis, among other beautiful courses. I am grateful for all the projects I have worked on since my Bachelor's and the people I have worked with within these projects.

I am grateful for the opportunity to work with many talented lab members on various projects while in lab. They include Dr Sowmya I, Dr Snehal K, Dr Mahita J, Dr Meenakshi I, Ms Shailya V, Mr. Asad A and Dr Nagarathnam B. I thank Ma'am for guiding me.

I thank past and present system administrators for their support and sector coordinators for their efforts during the lockdown. I thank all my seniors (Oommen, Adwait, Pritha, Anshul, Atul, Harini, Murugavel, Mahantesha, Rithvik, Malini, Sheetal, Sajeevan). I thank my lab members and colleagues for enabling remote work smoothly and for feedback during lab meets. I thank Shafi and Pankaj for help with logistics.

I thank all those I have worked with and provided me with helpful inputs.

I also thank friends in Paris, France and Basel, Switzerland for their congenial hospitality.

I would like to sincerely thank Dr Villoo Patell and the wonderful and resilient team at Avesthagen Ltd., Bangalore, including Dr Renuka Jain and Dr Naseer Pasha, for being supportive, positive and professional.

I thank my batchmates in NCBS and TDU. I thank my former and current lab members for their inclusivity, support and professionalism and sincerely wish them well in their endeavours ahead. I thank my friends and housemates such as Bhakti, Srishti and Vinaya (and many others) for their support and positivity. I thank the Cultural Club at NCBS for outstanding opportunities to engage

in creative expression. I thank my wonderful teachers of Yoga, classical music and classical dance. I thank my former school principal, late Mr Mark David, his respected wife, and all my teachers at Bombay Scottish School (P) and beyond such as Mrs Alka Gusain, Mr George W, Mrs Visalakshi L, Mrs Jincy J and many more.

I thank my parents, Ambika, Anshu, all my well-wishers, and community members, and Bala Vihar, Chinmaya Mission (Mumbai), for their blessings and goodwill.

I am, once again, grateful for this opportunity to pursue research here and for the guidance received. I thank Ma'am for providing me with a golden opportunity to pursue research in her lab and for accepting me as her PhD student.

Dedication

To,

Ma’am and my family

Table of Contents

List of Tables

List of Acronyms

3D – Three dimensional

AfloOBP- *Apis florea* odorant-binding protein

AgamOBP – *Anopheles gambiae* odorant-binding protein

BLAST – Basic Local Alignment Search Tool

DP - Dynamic Programming

GLIDE – Grid-based LIgand Docking with Energetics

GROMACS – GROningen Machine for Chemical Simulations

Gscore – Glide score

GWS - Genome-wide Survey

H-bond - Hydrogen bond

HMM – Hidden Markov Model

InsectOR – Insect Olfactory Receptor gene predictor

MD - Molecular Dynamics

ML - Maximum-Likelihood

MSA – Multiple Sequence Alignment

NCBI – National Centre for Biotechnology Information

OBP – Odorant Binding Protein

OPLS-aa - Optimized Potentials for Liquid Simulations all-atom

PAM250 – Point Accepted Mutation 250 matrix

PCA- Principal Component Analysis

PDB - Protein Data Bank

PDF - Probability Density Function

pdf - Portable Document Format

RAxML - Randomized Axelerated Maximum Likelihood

RMSD – Root Mean Square Deviation

RMSF – Root Mean Square Fluctuation

RNA - Ribonucleic Acid

SP –Standard Precision

XP – Extra Precision

x2

Chapter 1

Introduction

1.1 Chemosensory System

Chemosensation is the ability of an organism to perceive chemicals in its external environment. It is carried out by the chemosensory system of an organism and is essential for its survival and reproduction. Each sensory system receives stimuli from the external environment in cues that typically include visual, auditory, tactile, acoustic or/ and chemical. Chemosensory systems are one of the oldest sensory systems, ranging from bacteria (Berg 1975; Parkinson 1993; Grebe & Stock 1998) to mammals. In several phyla of vertebrates and invertebrates, chemoreception occurs when semiochemicals activate chemoreceptors directly or indirectly and elicit a downstream signalling cascade leading to an adapted behaviour. Chemosensation may also promote speciation (Smadja and Butlin, 2009). Various modalities like olfaction, gustation, vision play a role in the survival and mating of the species. Olfaction is mediated by olfactory stimuli-based activation of odorant receptors. In insects and mammalian systems, olfactory stimuli include small non-polar volatile odorants, semiochemicals and pheromones. Olfaction is essential in regulating various insect behaviours such as resource gathering, locating prey, avoiding predators, kin recognition, mate selection and oviposition (Jacquin-Joly & Merlin, 2004).

Behaviours based on recognising chemical stimuli have been widely recognised in insects and rodents (reviewed in Wyatt, 2003). The modalities of olfaction and gustation are the most ubiquitous sensory systems (Ache and Young, 2005) as seen in members in several taxa such as fish (sticklebacks) (McLennan, 2004), hermit crabs (Gherardi and Tiedemann, 2004), Antarctic prion birds (Bonadonna and Nevitt, 2004) and rotifers (Kotani *et al.*, 2001).

1.2 Odours as chemical cues

Olfaction is measured in several quantitative and qualitative ways, with the perceived intensity of the stimulus being a standard measure. Molecular features such as molecular weight and partial charge on the most negative atom correlate with perceived intensity (Keller and Voshall, 2016). In addition to molecular features such as size, volatility and lipophilicity also contribute towards an ‘odorous’ property (Boelens, 1983). Odorants may be sapid instead of volatile, depending on the ecological niche (Ache and Young, 2005).

An odour plume typically comprises multiple small-sized odorants (Blank et al., 1991). Odour molecules typically weigh less than 300 g/mol (Dunkel et al., 2009) and may be classified as

functional groups. Thus, deciphering signals provided by odorous substances is a common challenge among diverse phyla.

1.3 Biological components of the olfactory sensory systems

The olfactory sensory system in various organisms consists of the olfactory sensory/ receptor neurons (OSNs/ ORNs) on the membrane of the dendrites projected in the sensillary lymph of insects and the nasal epithelia in mammals. Present in the lymph are various soluble proteins, including the odorant-binding proteins (OBPs) and odorant degrading enzymes (ODEs). An incoming odorant stimulus that can activate the olfactory receptor leads to the depolarisation of its corresponding Olfactory Receptor Neurons (ORNs). Depolarised neurons relay these signals to centers known as glomeruli. These signals are relayed to the brain/ CNS, where encoding and integration corresponding to a specific smell occurs. Further, the usefulness of the perceived smell to the organism is assessed.

1.3.1 Insect olfactory system

The olfactory system primarily consists of two pairs of appendages, antenna and palps, in an adult insect. Fine, hair-like sensillae present on the antennae have multiple cuticular pores that facilitate the entry of diverse odorants. Based on the morphology, sensillum may be classified into various types such as ascoidea, auricillica, ampullacea, basiconica, Bohm's bristles, campaniforma, chaetica, coeloconica, placodea, styloconica, squamiforma, trichodea and vesiculoclada (Hallberg and Hansson, 2003; Keil, 1999). Most olfactory activity occurs in maxillary palp and the third segment of the antenna (Vosshall and Stocker, 2007; Kaupp, 2010). Each sensillum may house 1 to 4 ORN that relay signals to the olfactory glomeruli in the antennal lobe. The signal is further processed in the mushroom body. Among the various sensillae, basiconic, coeloconic and trichoid have been identified in insects such as ***Drosophila*** and mosquito (Order: Diptera) and honeybee (Order: Hymenoptera) (de Bruyne and Baker, 2008).

1.3.2 Vertebrate olfactory system

In the vertebrate model organism, various components are involved with olfactory/chemosensory roles (Su *et al*., 2009; Kaupp 2010). Odorant receptors (ORs) in the main olfactory epithelium

(MOE) and septal organ (SO) are known to respond to generic odours. Briefly, neurons from the MOE and SO project their axons to the main olfactory bulb glomeruli (MOB). Parallelly, neurons in the Grünenberg ganglion (GG) and guanylate cyclase D-containing cells in MOE project neurons to the necklace glomeruli (NG). Neurons in the vomeronasal organ (VNO) extend axons to the accessory olfactory bulb (AOB). From MOE, cilia from ORNs extend into the nasal epithelial mucus, thus being directly exposed to odorants. Olfactory signals from AOB is relayed to regions of the brain, including the anterior olfactory nucleus (AON), cortical amygdala (CoAMG), entorhinal cortex (ENT) and piriform cortex (PIR).

1.4 Odorant-binding protein

The discovery of OBPs in an invertebrate organism (giant moth, ***Antheraea polyphemus***) (Vogt *et al.*, 1981) and a vertebrate organism (cow) (Pelosi et al., 1981, 1982; Pevsner et al., 1990) has historically overlapped. Genomic sequences identifying OBPs have been subsequently identified in more than a hundred insect species, including ***Bombyx mori*** (silk moth) (Maida and Pelosi, 1990), ***Anopheles gambiae*** (Manoharan et al., 2013), and several vertebrate species. Research interest has increased dramatically in identifying OBPs and understanding the typical and atypical functions of OBPs in both invertebrates and vertebrates.

Insect odorant-binding protein

Insect OBPs are soluble, alpha-helix rich compactly folded globular proteins in the sensillary lymph of antennae (Vogt, 1981), other chemosensory organs such as reproductive appendages, and non-chemosensory organs. Insect OBPs are typically 14-20 kDa in molecular weight and have been identified and annotated across most phylogenetic orders in Class Insecta.

It has been suggested that insect OBPs help solubilise ligands and transport them across the aqueous sensillary lymph leading to OR activation and downstream signalling events. It facilitates rapid timescales at which insects respond to odorant stimuli.

1.4.1 Classification of insect odorant-binding protein- based on ligand-binding

Odorant-binding proteins in insects have been previously classified based on the binding as general odorant-binding proteins (GOBP), antennal binding protein (ABP) and pheromone binding proteins (PBP) (Leal, 2005).

1.4.2 Classification of insect odorant-binding protein- based on conserved cysteine signature

Aligning OBPs based on overall sequence information is challenging due to divergence in the amino acid composition. It has been suggested that *OBP* genes are rapidly evolving through duplication (Vogt et al., 2002; Zhou et al., 2008). However, a conserved cysteine signature across insect Orders is an important feature of OBPs and is used for classification into subfamilies (Breer et al., 1990; Raming et al. 1990; Krieger et al. 1993; Krieger et al. 1996). The presence of disulphide bonds due to the cysteines presents leads to the structural stability of the three-dimensional fold of OBP (Sandler et al., 2000; Tegoni et al., 2004; Zhou et al., 2008). Based on conserved cysteine signature and disulphide connectivity, OBPs have also been previously classified into Classic (six conserved cysteines), Minus-C (Foret and Maleszka, 2006), Plus-C (Zhou et al, 2008), Atypical (Hekmat-Scafe et al, 2002; Manoharan et al, 2013) and dimer OBPs (Manoharan et al, 2013). Plus-C OBPs show two additional cysteines and one proline, dimer OBPs show two cysteine signatures, Minus-C OBPs show loss of two of the six conserved cysteines, and atypical OBPs show the presence of 9-10 cysteines and a long C-terminus. Work done previously from our lab (Manoharan et al, 2013) identified 110 new members.

Vertebrate odorant-binding protein

Vertebrate OBPs are members of the extracellular lipocalin superfamily (Flower, 1996). They have been isolated as soluble proteins in the nasal mucus and nasal epithelia of various organisms such as rat, pig, cow, buffalo, humans. They are structurally characterised by a centrally located β-barrel flanked by terminal helices. The central cavity of the β-barrel is lined by non-polar residues that interact with putative ligands through various interactions such as hydrophobic interactions and pi-pi stacking.

Members of the lipocalin superfamily play key roles in the transport of delivering retinol and fatty acids to aid in the growth and development of the vertebrate organism (Monaco et al., 1995). As OBPs have been found primarily in olfactory tissue, it has is suggested that their chief role maybe direct interaction with odorants (Bignetti et al., 1985; Dal Monte, Andreini, Revoltella & Pelosi, 1991; Ohno, Kawasaki, Kubo & Tohyama, 1996; Pes et al., 1998; Scaloni et al., 2001; Guiraudie

et al., 2003). However, few OBPs are expressed in specific body fluids to modulate pheromone-specific behaviours (Böcskei et al., 1992; Marchese et al., 1998; Mastrogiacomo et al., 2014; Černá et al., 2017). This modulation may be done via transport of pheromones in biological glands and bodily fluids (Cavaggioni et al., 2000; Vincent et al., 2000; Spinelli et al., 2002; Hurst et al., 2004; Rajkumar et al., 2010; Ilayaraja et al., 2014), and other volatiles (Loebel et al., 2000; Spinelli et al., 2002; Muthukumar et al., 2018).

1.5 Binding properties of insect OBPs

The binding properties of insect OBPs have been characterised experimentally using competitive fluorescent binding assays and isothermal titration calorimetry. Various techniques have been used to characterise the binding properties of insect OBPs, with competitive fluorescent binding assays being most preferred. OBPs in *Drosophila melanogaster* such as LUSH, OBP28a and OBP19b are known to bind to dibutyl phthalate & pheromones, floral-like compounds and amino acids, respectively (Zhou et al, 2008; Rihani et al, 2021). The use of fluorescent binding assays showed binding of several long-chain alcohols, pheromonal compounds and fatty acids (FAs) to moth PBPs (Campanacci et al, 2001).

Studies show that binding of OBPs to specific ligands induce pH-specific conformational changes of amino acids at the C-terminus, disrupting salt bridge interactions (Laughlin et al, 2008; Manoharan et al, 2013).

This **thesis** aims to investigate various functional and evolutionary aspects of odorant-binding proteins (OBPs) in insects and mammals by leveraging diverse computational approaches on genomic and protein sequence data as well as three-dimensional structural data.

1.6 Computational Approaches Used In This Thesis

1.6.1 Sequence searches

Homology detection is based on how likely two sequences have evolved from a common ancestor. Homology is observed by identifying similar and dissimilar patterns of sequence and conserved functionally important residues (FIRs) during evolution aided by various dynamic programming methods. These patterns are represented as probabilities in substitution score matrices used to trace

alignment path with maximal summation of substitution score. The BLOSUM substitution matrix, for example, is obtained from local alignments of related proteins, and the BLOSUM62 matrix performs well in the case of distantly related proteins (Henikoff and Henikoff, 1993).

1.6.2 PSI-BLAST

Position-Specific Iterative Basic Local Alignment Search Tool, popularly known as PSI-BLAST, is a rapid and highly sensitive BLAST program used for detecting weak relationships, such as for divergent families of proteins with specific signatures. A series of iterations against a database using a scoring matrix indicates how each aligned position varies from the alignment of high-scoring sequence matches. This process iterates till convergence is attained (Altschul et al, 1997).

1.6.3 Exonerate

Exonerate is a technique to perform rapid and accurate sequence alignments using the Bounded Sparse Dynamic Programming (BSDP) approach (Slater and Birney, 2005). It combines both heuristic and exhaustive dynamic programming techniques. It is helpful for genome-wide searches with introns represented as gaps and has in-built models such as cDNA-to-DNA and protein-to-DNA alignments with flexible intron length thresholds and splice-site predictions. From previous observations and personal analysis of annotation of insect OBP gene models, the mean intronic size is around 2,000 bases in insects. Input consisting of multiple protein queries can be rapidly parsed for less than a GB genome. Closely and distantly related homologous genes can be aligned to the genome of interest-based on gene structure. It is helpful for genome-wide surveys.

1.6.4 Phylogeny

Phylogenetic reconstruction of biological entities is used to discover their evolutionary relatedness. Biological sequences of DNA, RNA or proteins can be used for phylogenetic reconstruction. Amongst various methods available for this (distance-based, parsimony-based, statistical-model-based), few are more suited for distantly related sequences. Currently, statistical methods such as Maximum-Likelihood (ML) are known to recover true trees most frequently. Likelihood estimates probabilities of sequences given a model of evolution, with more probable the sequence for the

given tree – more preferred the tree. Ideally, all possible trees for a given number of sequences need to be generated, and all sequences must be fit into it to evaluate these trees. Hence these are usually highly computationally intensive. However, as users can choose the model of evolution, it can be used for highly divergent sequences and get more accurate trees than other methods of phylogenetic reconstruction.

1.6.5 RAxML

RAxML (Randomized Accelerated Maximum Likelihood) is one of the most widely used techniques for the fast computation of large phylogenetic trees (Stamatakis 2006, 2014). It can compute 1000 taxon trees on a single PC processor in less than a day, which is highly useful for the phylogenetic reconstruction of sequences belonging to large protein families. As vertebrates and insects possess huge repertoires of OR sequences that vary a lot from species to species, RAxML is best suited for cross-species phylogenetic reconstruction. It is commonly used for phylogenetic analysis of insect genes by other laboratories (Zhou *et al*., 2010; Brand & Ramírez 2017; Brand *et al*. 2018).

1.6.6 Machine learning

Machine learning (ML) is a technique in artificial intelligence (AI) that creates a learning model based on given features and, depending on the nature of the feedback, trains the model in a supervised, unsupervised, semi-supervised way or through reinforcement. Depending on the dataset and the nature of the question, the dataset is split into training, testing, and validation sets to achieve reasonable accuracy, sensitivity and specificity. The predicted output may be binary (logistic) or a regression variable and is typically chosen as per the question to be answered. ML has vast applications across biology and other fields and may be used to derive meaningful predictions.

1.6.7 CNN

A deep neural network such as Convolutional Neural Network (CNN) uses input, output and hidden convolutional layers. LeCunn incorporated the first CNN in 1998, in which the connectivity pattern was based on the mammalian visual cortex. CNNs have been widely proven helpful in

biology with advancements in technology and data processing. The input layer contains many neurons that is equal to the total number of features in data. Data from the input layer is subsequently fed into hidden layer(s) depending upon data size and model. Output at each stage is calculated using layer-specific learnable weights and matrix multiplication of the current layer. The non-linearity of the network arises due to the addition of learnable biases and an activation function. Output from the hidden layer is relayed to a logistic function such as softmax or sigmoid where a probability score for each class is obtained. Subsequently, error in the model is calculated using functions such as square loss error, and loss is minimised by calculating derivatives through a procedure called backpropagation.

1.6.8 ADMET

ADMET is commonly used in pharmacology and drug development and refers to the "absorption, distribution, metabolism, excretion and toxicity" properties of a drug within an organism. These properties are known to affect the kinetics, performance and pharmacological activity of the drug in the system.

1.6.9 Docking

In bioinformatics and computational biology, molecular docking is a computational process that predicts how likely two chemical entities are likely to bind to one another. Docking may be typically carried out between two proteins, protein and ligand or a protein-ligand complex with another entity. In order to predict the set of most likely docked orientations or 'poses' between the two entities, docking score is calculated for various candidate poses. The docking score for a given pose is calculated using a scoring function that may approximate the binding free energy of the complex, such as in the Glide algorithm (Schrödinger suite). Here, force field-based interactions are considered, with favourable interactions being rewarded and unfavourable ones being penalised. Rapid, large-scale docking is the basis of virtual screening of a protein against a library of ligands.

1.6.10 Glide

Glide is a module from the Schrödinger suite that is used for docking and virtual screening (Friesner *et al*. 2006; Halgren 2007; Shelley *et al*. 2007). Quick and exhaustive search is carried out within the binding cavity grid to dock the ligand. It estimates the quality of a pose using GlideScore and Docking score. Both scores are equivalent in the absence of specific penalties. The scoring is achieved using hierarchical filters, including site-point search, diameter search, subset test, greedy scoring and refinement, grid minimisation and Monte Carlo sampling. Various energy terms are part of the equation that represent hydrogen bond, metal-ligand, Coulomb and vdW (van der Waals) and lipophilic interactions between the protein and ligand.

Virtual screening is a computational technique to identify the most probable binding site(s) of a given protein-ligand pair from a library of ligands. The ligands are screened stepwise with increasing stringency in a high throughput method (HTVS), followed by more stringent filtering criteria (Standard Precision; SP) and finally ending with Extra Precision (XP) docking. The penalties for unfavourable interactions get harsher, along with an increase in screening time per docked pose with each successive docking step. Hence, for docking a ligand to a protein moel with an imperfect binding pocket, the SP mode of Glide docking would be more suitable.

1.6.11 Binding affinity

The degree of protein-ligand binding measured in terms of binding affinity can determine its functional roles. It helps predict reaction mechanisms and reaction kinetics in the fields of drug development and pharmacology. This approach is practical when experimental approaches are not feasible. Binding affinity is quantified in various measures such as dissociation constant (Kd), inhibition constant (Ki), and changes in free energy measures (delta G, delta H and IC50). The docked complex may further be tested for binding affinity prediction. In order to predict the binding affinity, calculations such as molecular mechanics/generalised Born surface area (MM/GBSA) are used. The MM/GBSA approach consists of the molecular mechanics, the generalised Born model and solvent accessibility method to calculate binding affinity of the protein-ligand binding complex.

Prediction techniques of scoring binding affinity such as MM/GBSA and MM/PBSA are classical in approach and predict free binding energies of small ligands bound to a macromolecule. These

scores complement in-silico docking, virtual screening and experimental approaches in drug design.

1.6.12 Molecular Dynamics simulations

The need for molecular dynamic simulations arises because biological macromolecules such as proteins are dynamic in solution and typically conformationally sensitive to pH and ligand. Prediction of its dynamics and related properties is essential in optimising protein-ligand interaction studies and drug discovery. Due to the size of a typical protein, quantum mechanical calculations to calculate molecular movement are currently resource-inefficient. They are based on the usage of the Schrödinger equation to describe atomic wave motions. Hence, classical Newtonian equations of motion are solved for each atom in the system by factoring in interatomic potentials in the form of a force field. This is the basis of classical all-atom molecular dynamics simulations and is supported by timescales, including motion of protein domains (10^{-7} seconds) and rotation of sidechains in a protein (10^{-11} seconds). A force field is chosen based on the nature of the solvent and the overall system. The steps involved in a typical MD simulation are as follows-

1. Defining the system

The system comprises protein, ligand, solvent and added ions if any. Depending on the context, ions may be added appropriately to neutralise the system or mimic ionic conditions in the biological system. The forcefield is defined. The boundaries and their geometry is also defined here. The system may be energy minimised before proceeding further.

2. Equilibration

The system is heated, and initial velocities are assigned. The ensemble system is allowed to reach thermodynamic equilibrium using pressure, volume and temperature measures.

3. Production run

The simulation of all atoms in a system begins once the system is appropriately energy minimised, with constant temperature and pressure conditions and the release of selective restraints.

4. Analysis

The simulated trajectory is corrected for periodic boundary conditions to ensure that the system is still within the boundaries of the box within which it was initially assigned. Depending on the biological question, the corrected trajectory is used for obtaining various properties, including root mean standard deviation (RMSD) of the C-alpha atom of each residue in the protein, root mean square fluctuation (RMSF) of the residues, energy of the system, chemical nature of protein-ligand interactions and their respective duration. These and various other measures provide an understanding towards predicting phenomena such as stability of protein-ligand complex or conformational transitions in the protein.

1.6.13 EDA

EDA is often supplemented with cluster analysis. The patterns obtained through EDA explain maximum variance in the data and thus are most likely to determine essential conformational changes across structures (Amadei et al., 1993, Wolf & Kirschner, 2013). These "essential motions" or "essential dynamics" are explained by the first few principal components (PCs) and are necessary for protein function (Hayward & de Groot, 2008). The central assumption to PC analysis is that the input data is obtained by the linear combination of its basis vectors. The PCs are obtained by diagonalising the covariance matrix, C.

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^{\mathbf{T}}$$

The resultant diagonal matrix Λ contains eigenvalues, whereas the matrix **V** contains corresponding eigenvectors. Sorting eigenvectors according to their eigenvalues in descending order gives the first PC, i.e., the largest eigenvalue and so on. The first PC explains the highest proportion of variance in the given data, followed by that of the second PC, orthogonal to the first PC.

Analysis of trajectories makes it possible to calculate how all atoms within the molecule move as a system. The dynamic cross-correlation map typically comprises NxN heatmap, where N is the number of residues (Cα atoms) in the system. DCC between each i,j atom is calculated, and the correlation values lie between -1 to +1, where 1, -1 and 0 indicate total correlation, total anti-

correlation and no correlation, respectively. Cluster representative models were calculated based on minimal RMSD to the cluster average.

1.6.14 Clustering

Clustering is an unsupervised learning method to cluster data by minimising intra-cluster while maximising inter-cluster differences using similarity or dissimilarity measures (Tan et al., 2005), e.g. k-means (Ding and He, 2004; Forgy, 1965).

1.6.15 Average silhouette width

Silhouette coefficient (Kaufman, 1990; Rousseeuw, 1987) is a measure to determine the suitability of a data-point in a cluster with respect to neighbouring clusters. It helps in evaluating the optimal number of clusters for projected data.
For a given sample, it is given by

$$\mathbf{(b - a) / max(a, b)}$$

where a is the mean intra-cluster distance, b is the mean nearest-cluster distance for each sample. Its value ranges from -1 to 1, where a coefficient close to or equal to 1 would indicate that the data point is at maximum distance from other clusters, and -1 would likely mean incorrect cluster assignment of samples or poor clustering.

Figures

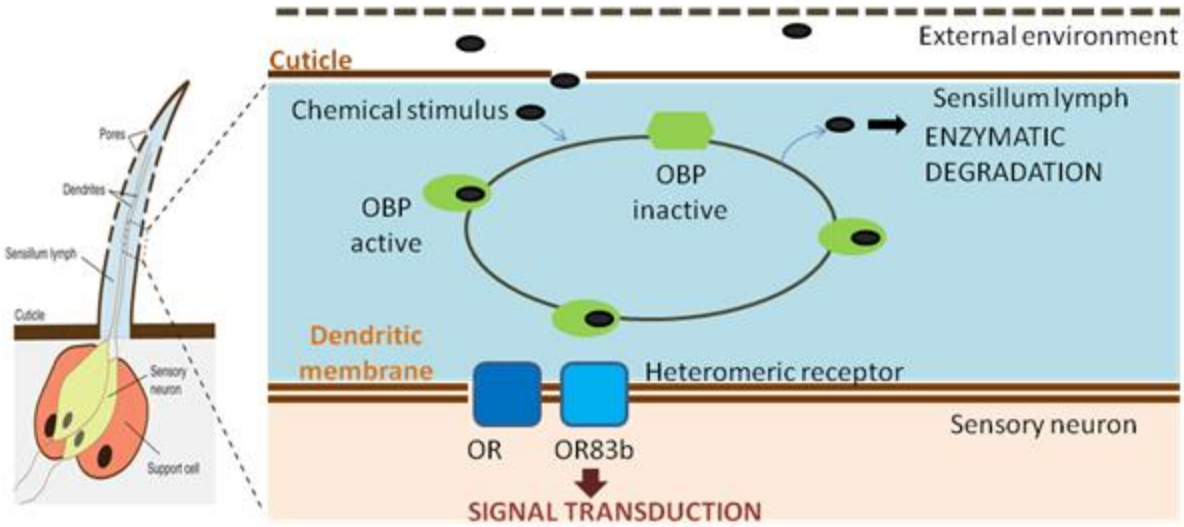
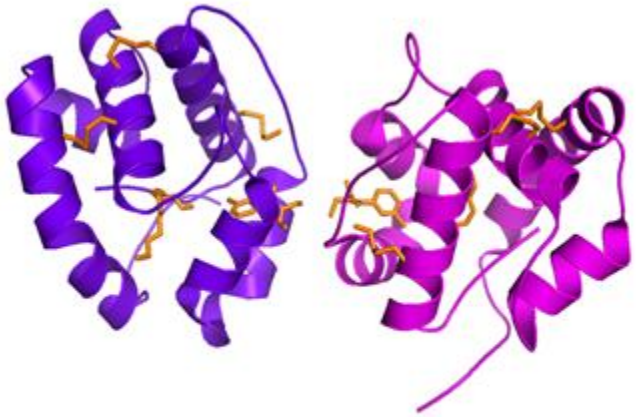


Figure 1.1: Overview of olfaction in insects- the proposed mechanism in the figure has been adapted from Sanchez et al, 2009.



Classic AgamOBP dimer bound to ligand

PDB ID: 3N7H

Figure 1.2: Structure of an insect OBP

Chapter 2

Pipelines

2.1 Introduction

The recent shift towards machine learning and deep-learning-based methods in gene annotation and structural biology has helped make biologically significant predictions. Techniques like regression and 'learning' intrinsic patterns in a complex plane of available data have led to predictions while optimising accuracy and computational resources.

I have worked on two pipelines towards different yet complementary objectives.

1. The first pipeline is a machine learning pipeline to identify, discriminate and annotate OBPs in insect genomes. This pipeline has been named SOCCER.
2. The second pipeline is a deep learning pipeline to predict a non-peptide ligand binding affinity with a high-resolution protein structure. This pipeline has been named DEELIG.

2.2 SoCCer

2.2.1 Introduction

Insects perceive olfactory cues to help them navigate their environment for survival and reproduction. As an increasing number of insect genomes get realised, an automatic method to classify proteins involved in olfaction is needed. It is helpful to enable their cognate odorant assignment and eliminate errors due to misclassification into protein families with related functions, biochemical or structural properties. Thus, a near-automatic computational protocol identifies and classifies 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 OBPs in large genomes but could be applied to many other globular proteins.

2.2.2 Materials and Methods

Workflow

The workflow of our approach is as described below (**Figure 2.1**). 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 NPC-2. A control dataset was prepared with 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 minimise 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. In case of SoCCer, many protein sequences are isoforms of one another and maybe closely related. Hence, hold out/ random sampling was used as our dataset was large enough and to overcome time constraints. MinMax scaler was used to normalise the data. Performance was evaluated using accuracy, precision, recall, f1-support, and Mathews' correlation coefficient (MCC). The performance was also visualised 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) – Two distinct features
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. The isoelectric point of sequence (pI)
l. Molecular coefficient of reduction extinction

m. Molecular coefficient of disulphide
n. The entropy of sequence
o. Number of globular domains in the sequence
p. Residue Adjacency Matrix of cysteines (RAM)

Insect genomes were collected from NCBI UniProtKB/Swiss-Prot, Ensembl and VectorBase. Genomic alignments were obtained using protein datasets representative of each of the seven classes as the 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 a list of predicted proteins on the presence of cysteine topology, PBP/GOBP domain, length cut-off, and the presence of signal peptide using predictive and in-house scripts. Due to the nature of predictions, scoring for only a 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 was tested using optimised 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-occurring domains identified yielded unique insights into the evolution and possible functional significance of OBP and other protein families involved in communication in insects.

SoCCer was benchmarked using OBPs from Apis florea. As Minus-C is underrepresented overall, currently, and due to the evolutionary close relationship of Minus-C and Classic, the algorithm classifies OBPs (Classic and Minus-C) with 57.8% accuracy and only Classic OBPs with 53.8%. This will be improved further by testing the stratified sampling method. The algorithm will also be tested on other datasets in other organisms.

User-end inputs-

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

1. 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 up to the first twenty characters.

2. An input folder with a genome of the organism of interest in fna format, genomic alignment output, 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 the Github repository https://github.com/bhavikamam/SoCCer/featurizer.py after appropriate changes in the 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'.

2.2.3 Results and Discussion

Advantages

Due to the similarities across some protein families in a 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. It identifies whether a given protein sequence is one of the major families in insect chemical communication and classifies it further.

Comparison with other methods

Generally, computational pipelines may combine gene prediction tools, starting from handling raw sequencing data, gathering transcript evidence and so on. Sequence search techniques like BLAST are known to 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.

We wanted to overcome constraints due to missing secondary structure and connectivity information. Thus, SoCCer is trained to discriminate among OBP subfamilies and some other soluble chemosensory protein families.

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 concerning the secondary structure and re-submit. Similarly, N-terminal and C-terminal overhangs can be submitted to Pfam to recognise proper domain boundaries.

Time taken

The estimated time taken for the pipeline depends on the genome size 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 are done using in-house Python scripts and take approximately 5 hours for ~1000 protein sequences. Training neural networks took 10-12 hours, whereas testing protein sequences takes less than 30 minutes.

Results and Discussion

An in-house study (Manoharan, Sankar et al, 2013) from the lab that sub-classifies family of mosquito OBPs into known subfamilies has been insightful work. To the best of my knowledge, this is a novel attempt to classify a test protein sequence into subclasses of OBPs or other soluble proteins present in the same chemical milieu of the insect sensillum.

The pipeline analyses 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 is further subjected to feature matrix generation used as input to the optimised and trained neural network model (**Figure 2.1**). The network's output is a regression value indicating the likelihood of the sequence belonging to each of the classes. Thus implementation of this pipeline streamlines gene and protein sequence prediction of classes important for insect chemical communication. It is done taking into account size, conserved cysteine pattern,

disulphide spacing, and secondary structure. The use of ANN, LR, RF and Gradient Boost machine learning models has successfully helped in classifying soluble proteins in chemical communication in insects with high accuracy (>85%) (**Table 2.1**).

Acknowledgement: I thank Dr Shannon Olsson and Dr Axel Brockmann for their suggestions to further improve the pipeline, ie, training and testing on a control dataset. I acknowledge technical inputs by Asad Ahmed in the neural network modules and some technical comments in the machine learning module.

2.3 DEELIG

Various approaches have been previously implemented to predict binding affinity using deep learning. These include using feature learning along with gradient boosting algorithms to predict binding affinity (He et al, 2017), using protein sequence level features in a CNN-based framework for prediction (Öztürk et al, 2018), and using atomic-level features of complexes in a CNN-based framework for binding affinity prediction (Li et al, 2019).

The aim was to apply deep learning to predict the binding affinity of high-resolution protein with non-peptide ligands without the need for a docked pose as input. CNNs have been used to quantify protein-ligand binding affinity using features corresponding to protein and ligand through spatial relationships amongst the data **without using docked poses as input from the user end**.

Detailed information on the network architecture and the training is found in the related publication (Ahmed*, **Mam***, Sowdhamini, 2021):

https://doi.org/10.1177%2F11779322211030364

2.3.1 Materials and Methods

Raw Data

The raw data for our novel dataset was obtained from the RCSB PDB dataset with the input queries consisting of 'chain type' as 'protein only', 'binding affinity' as 'Kd or Ki value', 'chemical components' as the presence of 'ligand(s)', and 'X-ray crystallography method' with resolution up to 2.5 Å. The search results include the structures deposited in the PDB database, PDBBind, Binding MOAD (Ahmed, Smith et al, 2014; Hu et al, 2005) and scPDB (Benson et al, 2008). The raw dataset contained protein structures in PDB format, protein sequences in FASTA format,

ligands in SDF format, and binding affinity values of 5464 protein-ligand pairs corresponding to 29,650 complex **unique** chain-ligand pairs.

Refinement of Dataset

The PDB, FASTA and SDF files obtained were further processed to refine the novel dataset as described. Protein chain-ligand pairs with missing or zero affinity values were removed to reduce noise. It filtered our dataset to 4,750 protein-ligand pair complexes. Cavity information was extracted from the protein using Ghecom (Kawabata, 2010) and converted to MOL2 format using Chimera (Pettersen et al, 2004). It filtered our dataset to 4286 pocket-ligand pair complexes.
Protein-ligand pairs with missing dihedral angle information, PSSM profiles, or secondary structures were discarded. The final dataset contained 4041 pocket-ligand pairs, corresponding to 7414 pocket-ligand pairs with unique chains.

Extraction of features

The deep learning network was trained on raw information resulting in increased time to converge with lesser accuracy. A conventional methodology was used for feature extraction using a deep learning framework to learn the interactions between the protein-pocket (cavity) and ligand for their affinity prediction.

Features of Protein Cavity

A detailed multi-level methodology for feature extraction was used. It utilised information at the atomic, sequence and structural levels.

1. The 'Atomic level' methodology resulted in a vector size of 19 bits. Here, the encoding was as described below.
    a. 9 bits 1 hot or all null hot encoded for boron, carbon, nitrogen, oxygen, phosphorus, sulphur, selenium, halogen and metal atoms
    b. 1 integer for hybridization
    c. 1 integer representing the number of bonds with heavy atoms
    d. 1 integer representing the number of bonds with hetero atoms

e. 5 bits encoding SMARTS pattern of acceptor, aromatic, donor, hydrophobic, and ring properties
f. 1 float value for partial charges
g. 1 integer to distinguish between the protein as +1 and ligand as -1

2. The 'Amino Acid level' methodology resulted in a vector size of 25 bits. We utilised the sequence information of protein to get more features about the protein pocket-ligand interaction. Here, the encoding was as described.
    a. Position-Specific Scoring Matrix (PSSM): PSSM is a matrix that represents the probability of mutation at each point of the sequence. It gives a 20-bit probability matrix for each amino acid at each location. PSSM profiles were obtained using PSI-BLAST against the SwissProt database with an E-value cut-off of 0.001. Chains with more than 50 amino acids were retained.
    b. Relative Solvent Accessibility (RSA): It is encoded by one bit of information for each amino acid indicating exposure to solvent. RSA was obtained using NACCESS with a threshold of 25%.
    c. Secondary Structure: It is encoded by 1 bit of information about the structure as coil, helix or plate and was predicted using the DSSP (Kabsch and Sander, 1983).
    d. Dihedral Angles: It is encoded by two bits of information with phi/psi angles of each amino acid and was predicted using DSSP to obtain dihedral angles.

Features of Ligand

Standard features of ligand were calculated for ligands in our dataset using PADEL (Yap, 2011) and fingerprints (1D, 2D and chemical fingerprints), including hybridisation, atom pair interaction, and counts of various functional groups. We also used QikProp (Schrödinger Release **2020-3**) to derive ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties, including the physical properties, solubility and partition coefficients. A 1D array was obtained with 14,716 dimensions containing ligand features. It was further used as a feature vector representing the ligand represented in MOL2 format.

Preprocessing

Please refer to the published manuscript for further details.

Calculation of binding affinity

The predicted value of our regression-based approach is the negative natural logarithmic value of Kd or Ki. It is then converted to its antilog to obtain Kd or Ki value in nanoMolar quantity.

Performance

The accuracy is measured in terms of scores Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Standard Deviation (SD).

Case studies of specific protein families

Recently deposited complexes of COVID-19 main protease with various inhibitors deposited in the PDB were used for our study (**Table 2.2**). The crystal structure complexes (PDB IDs: 5R7Y, 5R7Z, 5R82, 5R84) of the COVID-19 main protease with inhibitors ((Z45617795: N-[(5-methylisoxazol-3-yl)carbonyl]alanyl-L-valyl-N~1~- ((1R,2Z)-4- (benzyloxy)-4-oxo-1-{[(3R)-2-oxopyrrolidin-3-yl]methyl)but-2-enyl)-L-leucinamide); Z1220452176: (~{N)-[2- (5-fluoranyl-1~{H)-indol-3-yl)ethyl]ethanamide); Z219104216: 6- (ethylamino)pyridine-3-carbonitrile; Z31792168: 2-cyclohexyl-~{N)-pyridin-3-yl-ethanamide)) respectively has been recently deposited in PDB (2020; unpublished).

Another study has deposited the complex of the COVID-19 main protease with a broad-spectrum inhibitor X77 (N- (4-tert-butylphenyl)-N-[(1R)-2- (cyclohexylamino)-2-oxo-1- (pyridin-3-yl) ethyl]-1H-imidazole-4-carboxamide (2020; unpublished).

In order to compare affinity of deoxycholate with homologous proteins of the periplasmic C-type cytochrome, Ppc homologs PpcA (PDB: 1OS6), PpcB (PDB: 3BXU), PpcC (PDB: 3H33), PpcD (PDB: 3H4N) and PpcE (PDB: 3H34) and ligand deoxycholic acid (Pubchem CID: 222528) were gathered. These were processed, and DEELIG was used to predict the binding affinity of each homolog with the ligand.

OBP-ligand binding affinities are currently being analysed and will be updated accordingly.

Case studies of odorant-binding proteins

Protein-ligand complexes from PDB were obtained with the protein family annotation as 'PBP/GOBP family' or 'Lipocalin / cytosolic fatty-acid binding protein family'. Complexes with the domain identifier 'PBP_GOBP' and 'lipocalin' were retained to predict OBP-ligand affinity. Domain identifiers' PBP_GOBP' (**PF01395**) and 'lipocalin' (**PF00061**) correspond to insect and vertebrate odorant-binding classes, respectively.

2.3.2 Results and Discussion

Two modules were trained. The first module was trained using a small set of features for protein and ligand, which were represented together in a 3D grid space. This approach has also been part of a previous study (Stepniewska-Dziubinska et al, 2018). However, the previous study uses a restricted ligand set that does not involve larger ligands. Here we have used a diverse set of ligands as one of our inputs. While we have used ~68% of the PDBbind core set in our training dataset, this subset of PDBbind core set incorporated constitutes only ~2.9% of our entire training dataset. With the training of the Atomic Model for *35 epochs*, ***and MAE score of 2.84 was achieved***.

Another module was constructed that enabled us to improve on the ligand and protein-based information. To this purpose, we used an increased feature vector size which amounted to 14716 bits in size for ligand and 44 bits for each atom of protein. With the training of the Composite Model for only *four epochs*, **MAE score of 2.27** was achieved.

The performance of our model was further evaluated using ligand-bound complexes from the kinase superfamily from PDB. The composite model outperformed the atomic model significantly and with a lower standard deviation.

DEELIG algorithm relies on certain inputs, including sensitive binding cavity detection by the Ghecom algorithm (Kawabata, 2010) that uses mathematical morphology to find deep and shallow pockets (if any) in a given protein. The coordinates of the predicted binding cavity of the protein (grid) are rotated to various combinations. They are placed around the centroid of the ligand, and the resultant 4-D tensor is processed further for features along with the CNN (**Materials and Method**). Hence, ligand-bound poses are not used as input. Our dataset has ~5k+ complexes and

includes complexes that were not part of PDBBind (usually used to benchmark and derived from PDB). The ligand set we have used also represents a diverse set and is one of the highlights of our approach. The predictions from DEELIG can help existing databases like RSCB PDB, PDBMoad and PDBBind in filling missing binding affinity data for complexes.

Care was taken to reduce noise in the data during training and testing but due to the stochastic nature of the algorithm as well as pooling and processing features resulting from varying algorithms, there are variations present as expected. Care was also taken to train the model on atomic-level features, sequence and structural-level features. These features were shortlisted from an even larger pool of features available on the basis of biological relevance, relevance to the question and context and the information content that would be applicable for most protein classes and ligand types.

DEELIG was additionally tested on the PDBbind core set (**see publication**), and the percent correct classification (PCC) score of 0.89. The PCC score was higher than Autodock Vina (core set v13, PCC 0.6) and four other methods that have been previously tested on the same PDBbind core set. DEELIG also outperformed TOPBP (Complex) using the PDBbind core set v16.

DEELIG has eliminated **the need of providing ligands in a complex pose with protein at the user end**. Thus a given protein pocket can be tested for the degree of binding for any given ligand. It can be extended to predicting potential binding partners for proteins in other superfamilies as well. It is also important to consider that docking score and pose is not a reliable correlation with MM/GBSA poses (Pantsar, Poso, 2018). **DEELIG can be used to predict binding affinity for a high-resolution structure of a member of any protein superfamily and a non-peptide ligand, the docking pose of which may or may not be known.** Although our dataset contains 5464 complexes compared to 16,151 complexes found in PDBbind, the ligands used as part of our training include 452 unique ligands absent in PDBbind. This helps in **achieving ligand diversity (see publication)** during training the CNN model. The similarity matrix constructed from the binary fingerprints of ligands used in the dataset supports our claim of improved ligand diversity in our dataset.

Binding affinity predictions through DEELIG can be extended to protein-ligand complexes of protein superfamilies where the affinity is quantitatively unknown due to experimental limitations or where the potential for binding is yet to be explored *in vitro*.

DEELIG applied on protein-ligand complexes from the coronavirus (CoV) family in light of the coronavirus pandemic. The COVID-19 main protease is a key enzyme for the novel strain of coronavirus implicated in the pandemic. A recent study involved testing of in-vitro binding efficacy of coronavirus COVID-19 virus main protease (Mpro) with a potent irreversible synthetic inhibitor, N3 (Jin et al, 2020). However, the highly potent inhibition by N3 rendered the experimental determination of binding affinity not achievable. Using the structure of Mpro at high resolution (7BQY: 1.7 Å), we have been able to predict the binding affinity of N3 to 3.1e+4 nanomolar. This value agrees with the observed high affinity in the course of recent experiments (Jin et al, 2020).

We used complexes of COVID-19 main protease with various inhibitors to predict their respective binding affinities as their experimental values have not been made available. Based on our model-based predictions, broad-spectrum inhibitor X77 scores for highest affinity followed by ligands Z45617795, N3, Z31792168, Z1220452176 and Z219104216 in the order of decreasing binding affinity (**Table 2.2**), strengthening the suitability of X77 as a potential candidate against COVID-19 virus protease.

A triheme cytochrome from the sulfur-, metal- and radionuclide-reducing bacteria, *Geobacter sulfurreducens*, PpcA binds strongly to deoxycholate. However, its triheme paralogous counterparts PpcB, PpcC, PpcD and PpcE do not bind to deoxycholate. Our results also predict that ligand deoxycholate binds with high affinity to periplasmic C-type cytochrome A (PpcA) but not to its homologs PpcB, PpcC, PpcD and PpcE (**Table 2.3**).

One of the main objectives in the construction of DEELIG

is the **rapid** and **accurate** estimation of binding affinity between protein chain-ligand pairs for which docking pose information may not be available, where binding affinity has not been quantified (as in the case of some COVID protease-drugs in **Table 2.2**), or where the specificity of protein-ligand interactions do not correlate strongly with binding affinity. A good example of a class representing the latter conundrum is found in Odorant binding proteins (Murphy *et al*, 2013). Towards addressing this, the predictive model in DEELIG was applied on OBP-ligand complexes obtained from PDB. As is expected in the analysis, OBP chain- ligand

complexes from PDB were not more than 100 in either insect OBPs (PBP/GOBP members= 28) or mammalian OBPs (part of lipocalin family= 103). While DEELIG predictions on lipocalins have a relatively higher RMSE and PCC than that of members of the PBP/GOBP family (**Table 2.4**), RMSE and PCC values are comparable to the general performance of DEELIG (RMSE=2.71, PCC=0.77) (analysis in manuscript). However, statistical metrics like PCC are also affected by the limited sample of ligand-bound complexes available in PDB, as in OBPs, CSPs (PF03392; A10/OS-D) and NPC-2 proteins (PF02221; ML domain). Across all domains examined, especially in OBPs, DEELIG was sensitive to inter-chain differences and differences in the protein atom in contact with the ligand in the 4-D tensor grid. It reflects in the different binding affinities of protein chain-ligand pairs for the same protein and position of binding atom on protein chain (**Table 2.5**), such as in cases of (a) pheromone-binding proteins in *Apis mellifera* that bind to 9-oxodec-2-enoic acid, a queen substance pheromone (PDB: 3CYZ) and n-butyl-benzenesulfonamide, a neurotoxic plasticiser (PDB: 3CZ1), (b) human lipocalin (PDB: 1L6M) with 2,3-Dihydroxybenzoic acid, a naturally-occurring phenol, (c) bovine OBP (PDB: 1OPB) with retinol (vitamin A), and (d) mouse lipocalin (PDB: 2QM9) with troglitazone, an anti-diabetic and anti-inflammatory drug.

DEELIG was additionally tested on the PDBbind core set (**Table 2.6; Ahmed, Mam, Sowdhamini, 2021**), and the percent correct classification (PCC) score of 0.89. The PCC score was higher than Autodock Vina (Chen, Ke, Lu, et al, 2019) (core set v13, PCC 0.6) and four other methods (Li et al, 2015; Wang, Zhang, 2017; Cang et al, 2018; Nguyen and Wei, 2019) that have been previously tested on the same PDBbind core set. DEELIG also outperformed TOPBP (Complex) using the PDBbind core set v16 (**Table 2.6**).

Combining approaches in (**Section 2.2) SoCCer**, **DEELIG (Section 2.3) and reciprocal homology approach (Section 3.2)** promises further novel insights into the understanding of OBP-ligand interactions. Specifically, OBPs and other proteins involved in insect chemical communication identified in insect genomes through the SoCCer pipeline are currently being screened against a ligand library of common odorants, metabolites and volatiles to identify combinatorial OBP-ligand binding to give further context to the odorant receptor-odour coding that is known to occur in insects and mammals.

Tables and Figures

Figure 2.1: SoCCer: a combined method using genomic, protein-based and ML approaches

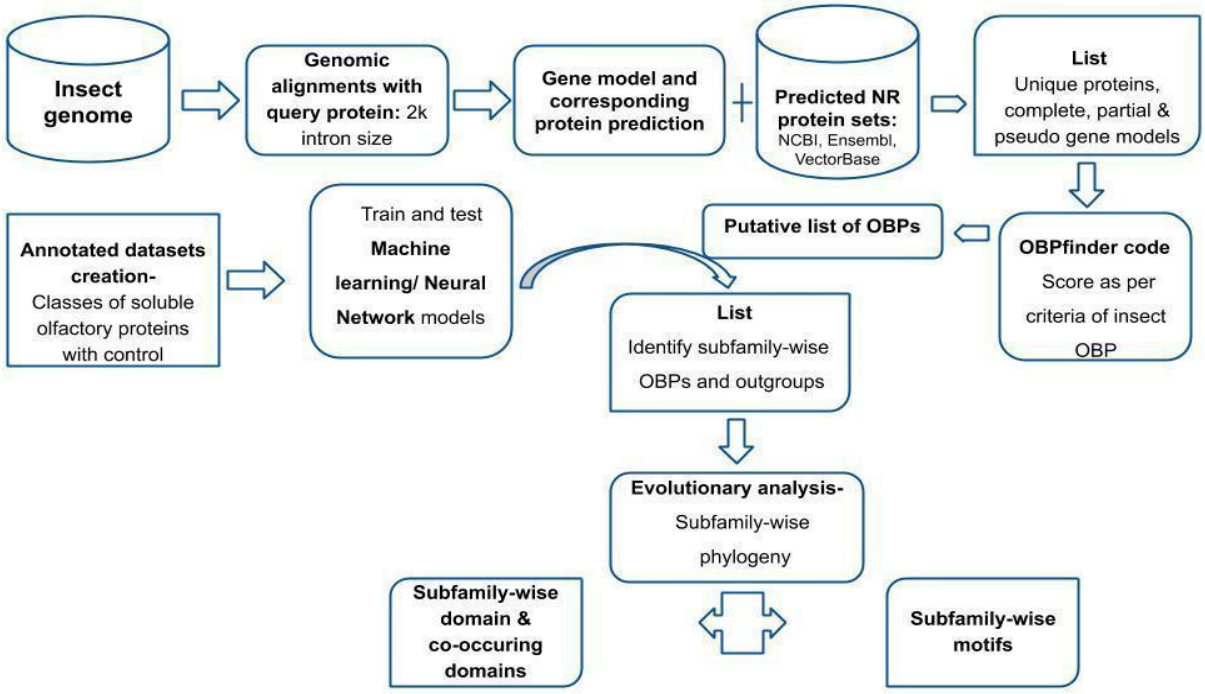


Table 2.1: Accuracy of ML and NN models to standardize the pipeline

ML models	**Accuracy (%)**
Support Vector Machine (SVM)	**83**
Naive Bayes Gaussian	**57.19**
Naive Bayes Complement	**54.59**

Logistic Regression	**84.57**
AdaBoost	**40.72**
Gradient Boost	**90.64**
Xtreme Gradient Boost	**89.77**
Random Forest Gaussian	**88.90**
Artificial Neural Network (ANN)	**90**
Convolution Neural Network (CNN)	**83**

Table 2.2: DEELIG predictions of binding affinity on COVID-19 protease-drug complexes

PDB	**Ligand**	**-Log (Kd/Ki)**	**[Kd] or [Ki] (nM)**
5R7Y	Z45617795	11.96	6.39e+3
5R7Z	Z1220452176	7.69	4.57e+5
5R82	Z219104216	6.12	2.18e+6
5R84	Z31792168	8.32	2.43e+5
6W63	X77	15.34	2.17e+2
7BQY	N3	10.38	3.10e+4

Table 2.3: DEELIG predictions of binding affinity on homologs of Periplasmic C-type cytochrome (Ppc) family

Homolog	PDB ID	Prediction Kd or Ki (uM)
PpcA	1OS6	4.512
PpcB	3BXU	416.042
PpcC	3H33	835.232
PpcD	3H4N	483.678
PpcE	3H34	187.157

Table 2.4: DEELIG prediction accuracy on odorant-binding proteins

Domain	MAE	RMSE	PCC
PBP/GOBP	2.38	2.78	0.51
Lipocalin	1.93	2.52	0.67

Table 2.5: DEELIG performance on protein-ligand complexes from the insect OBP and lipocalin families in PDB

Domain identifier	Complex (PDB ID)	Real value (-log (kd/ki)	Predictive value (-log (kd/ki)
PBP_GOBP	2PQL_TSS_A_145	14.46	10.15
PBP_GOBP	2QEH_SRO_A_145	17.55	13.63
PBP_GOBP	3CYZ_9OD_A_124	14.33	10.47
PBP_GOBP	3CYZ_9OD_B_121	14.33	9.61

PBP_GOBP	3CYZ_9OD_B_122	14.33	15.15
PBP_GOBP	3CZ0_9OD_B_122	14.33	10.05
PBP_GOBP	3CZ1_NBB_A_122	14.42	12.44
PBP_GOBP	3CZ1_NBB_B_122	14.42	12.51
PBP_GOBP	3CZ1_NBB_B_123	14.42	12.86
PBP_GOBP	3D76_NBB_A_121	15.29	12.43
PBP_GOBP	3D78_NBB_A_120	15.29	15.37
PBP_GOBP	3D78_NBB_B_120	15.29	17.20
PBP_GOBP	3D78_NBB_B_121	15.29	10.40
PBP_GOBP	3DZT_EAH_A_305	14.33	10.85
PBP_GOBP	3N7H_DE3_A_129	8.07	7.45
PBP_GOBP	3N7H_DE3_B_132	8.07	8.64
PBP_GOBP	3NHI_EAH_A_1	17.12	13.96
PBP_GOBP	3NHT_U46_A_1	13.84	10.03
PBP_GOBP	3R1V_AZB_A_128	11.34	10.59

PBP_GOBP	3R1V_AZB_B_128	11.34	12.89
PBP_GOBP	3S0B_FNA_A_120	12.66	15.58
PBP_GOBP	3S0D_CTV_A_120	12.79	12.81
PBP_GOBP	3S0E_EOL_A_120	13.93	11.31
PBP_GOBP	5V13_JH3_A_301	15.28	12.39
PBP_GOBP	5V13_JH3_B_301	15.28	16.83
PBP_GOBP	5V13_JH3_C_301	15.28	12.11
PBP_GOBP	3CZ0_9OD_A_124	14.33	12.28
PBP_GOBP	3CZ0_9OD_B_121	14.33	14.02
Lipocalin	1ADL_ACD_A_135	13.23	11.44
Lipocalin	1CBS_REA_A_200	13.13	16.37
Lipocalin	1DZK_PRZ_A_700	11.74	11.26
Lipocalin	1ERB_ETR_A_184	16.35	13.72
Lipocalin	1FDQ_HXA_A_133	14.45	11.85
Lipocalin	1FDQ_HXA_B_633	14.45	12.93

Lipocalin	1G74_OLA_A_132	11.11	13.36
Lipocalin	1G85_3OL_A_1001	10.32	9.70
Lipocalin	1G85_3OL_A_1002	10.32	9.93
Lipocalin	1G85_3OL_B_1003	10.32	10.12
Lipocalin	1G85_3OL_B_1004	10.32	10.09
Lipocalin	1GT1_3OM_B_1159	11.34	11.94
Lipocalin	1GT1_ANC_A_1161	11.52	9.45
Lipocalin	1GT1_ANC_B_1158	11.52	9.60
Lipocalin	1GT1_PRZ_A_1160	10.32	7.72
Lipocalin	1GT3_3OM_A_1161	11.34	11.09
Lipocalin	1GT3_3OM_B_1158	11.34	13.44
Lipocalin	1GT3_DHM_A_1160	12.57	12.01

Lipocalin	1GT3_DHM_B_1159	12.57	11.25
Lipocalin	1GT4_UNA_A_1160	12.72	12.36
Lipocalin	1GT4_UNA_B_1158	12.72	10.45
Lipocalin	1GT5_BZQ_A_1160	11.74	9.97
Lipocalin	1GT5_BZQ_B_1158	11.74	8.78
Lipocalin	1GX8_RTL_A_1163	14.84	13.69
Lipocalin	1HMR_ELA_A_133	12.79	15.89
Lipocalin	1HMS_OLA_A_133	12.36	13.43
Lipocalin	1HMT_STE_A_133	11.02	12.51
Lipocalin	1HN2_3OL_B_3003	10.32	7.11
Lipocalin	1HN2_ANC_A_3001	11.52	11.53
Lipocalin	1HN2_ANC_B_3002	11.52	11.73

Lipocalin	1L6M_DBH_A_201	16.36	16.41
Lipocalin	1L6M_DBH_A_202	16.36	18.16
Lipocalin	1L6M_DBH_B_302	16.36	11.26
Lipocalin	1L6M_DBH_C_401	16.36	12.88
Lipocalin	1LFO_OLA_A_128	10.45	9.27
Lipocalin	1LFO_OLA_A_129	10.45	13.12
Lipocalin	1LIF_STE_A_132	7.14	11.31
Lipocalin	1LKE_DOG_A_500	15.02	12.41
Lipocalin	1LNM_DTX_A_700	17.73	14.71
Lipocalin	1N0S_FLU_A_500	14.86	11.77
Lipocalin	1N0S_FLU_B_501	14.86	13.87
Lipocalin	1OPB_RET_A_134	14.51	18.90
Lipocalin	1OPB_RET_B_134	14.51	16.68
Lipocalin	1OPB_RET_C_134	14.51	14.96
Lipocalin	1OPB_RET_D_134	14.51	15.64

Lipocalin	1VYG_ACD_A_1134	16.12	11.39
Lipocalin	2ACO_VCA_B_501	10.6	10.79
Lipocalin	2ANS_2AN_A_201	11.52	9.55
Lipocalin	2FR3_REA_A_300	13.13	18.42
Lipocalin	2G78_REA_A_200	13.13	12.05
Lipocalin	2G79_RET_A_200	13.64	13.10
Lipocalin	2GJ5_VD3_A_163	16.87	16.55
Lipocalin	2HNX_PLM_A_135	14.01	13.40
Lipocalin	2HZQ_STR_A_300	11.52	15.90
Lipocalin	2NNQ_T4B_A_293	13.03	14.64
Lipocalin	2QM9_TDZ_A_201	15.59	15.44
Lipocalin	2QM9_TDZ_B_202	15.59	19.50
Lipocalin	2RA6_ETY_A_604	8.12	5.09

Lipocalin	2RA6_ETY_B_601	8.12	5.56
Lipocalin	2RA6_ETY_C_603	8.12	6.50
Lipocalin	2RA6_ETY_D_602	8.12	6.46
Lipocalin	3CWK_REA_A_300	13.13	15.59
Lipocalin	3DSZ_LIZ_A_187	15.59	13.81
Lipocalin	3DSZ_LIZ_B_187	15.59	15.04
Lipocalin	3FW4_CAQ_A_180	19.34	10.13
Lipocalin	3FW4_CAQ_C_180	19.34	12.84
Lipocalin	3NQ3_DKA_A_3053	6.41	10.64
Lipocalin	3NQ9_OCA_A_1111	6.99	5.12
Lipocalin	3UEX_STE_A_163	13.64	13.79
Lipocalin	4GNY_SDS_A_201	7.57	11.68
Lipocalin	4QZT_RTL_A_201	17.73	17.69

Lipocalin	4QZT_RTL_C_201	17.73	16.72
Lipocalin	4QZU_RTL_C_201	17.73	15.53
Lipocalin	4ZHF_4OL_A_202	15.33	16.10
Lipocalin	4ZHF_4OL_B_202	15.33	11.41
Lipocalin	4ZHF_4OL_C_202	15.33	13.70
Lipocalin	4ZHF_4OL_D_202	15.33	14.17
Lipocalin	4ZHF_4OL_E_202	15.33	14.29
Lipocalin	4ZHF_4OL_F_202	15.33	14.72
Lipocalin	4ZHG_4OL_A_202	15.06	12.83
Lipocalin	4ZHG_4OL_B_202	15.06	14.01
Lipocalin	4ZHG_4OL_C_202	15.06	15.39
Lipocalin	4ZHG_4OL_D_202	15.06	12.98
Lipocalin	4ZHG_4OL_E_202	15.06	17.44
Lipocalin	4ZHG_4OL_F_202	15.06	14.79
Lipocalin	5H8T_RTL_A_201	15.51	12.85

Lipocalin	5HA1_RNE_A_201	14.27	14.54
Lipocalin	5HBS_RTL_A_201	15.51	14.60
Lipocalin	5L8N_6RQ_A_1001	8.08	8.10
Lipocalin	5L8N_6RQ_C_1001	8.08	10.09
Lipocalin	5LJB_RTL_A_201	16.92	16.40
Lipocalin	5LJC_RTL_A_201	14.25	14.65
Lipocalin	5LJD_RTL_A_201	14.18	14.42
Lipocalin	1DZK_PRZ_B_700	11.74	8.01
Lipocalin	1RBP_RTL_A_183	13.18	17.11
Lipocalin	1VYF_OLA_A_1134	16.23	11.91
Lipocalin	2ANS_2AN_B_202	11.52	10.26
Lipocalin	2GJ5_VD3_A_164	16.87	17.78
Lipocalin	4QYN_RTL_B_201	17.73	14.14
Lipocalin	4QZU_RTL_A_201	17.73	17.31

Lipocalin	4QZU_RTL_B_201	17.73	16.18
Lipocalin	5LJE_RTL_A_201	12.9	13.43
Lipocalin	1L6M_DBH_B_301	16.36	11.41

Table 2.6: Pearson Correlations Coefficient on PDBbind core set (Ahmed, Mam, Sowdhamini, 2021)

Method	**PDBbind v2013**	**PDBbind v2016**
Autodock Vina [10]	0.6	-
RF::VinaElem [11]	0.752	-
Wang et al RF20 [12]	0.732	-
TOPBP (Complex) [13]	0.808	0.861
AGL Score [14]	0.792	-
DEELIG	0.894	0.889

Chapter 3

Genome-wide survey of odorant-binding proteins in insects

3.1 Introduction

Genome-wide surveys (GWS) have been previously performed across insect genomes to identify odorant-binding proteins in the existing literature. Odorant-binding proteins have been predicted in various species, including *Apis mellifera* (Order: Hymenoptera) (Forêt and Maleszka, 2006), *Drosophila melanogaster* (Order: Diptera) (Hekmat and Scafe, 2002; Graham and Davies, 2002), *Anopheles gambiae* (Order: Diptera) (Manoharan *et al*, 2013), *Periplaneta americana* (Order: Blattodea) (He *et al*, 2017) using sequence searches with homologues in other closely related organisms as a typical start-point.

I have worked on the GWS of OBPs in eusocial honeybee (*Apis florea*), mosquitoes and other insects.

3.2 GWS of OBPs in mosquito

3.2.1 Background

While there exist documented species-specific differences in Hymenopteran OBPs, including 21 OBPs in eusocial *Apis mellifera* (Foret and Maleszka, 2006), 7 OBPs in fig wasp *Ceratosolen solmsi* (Wang et al, 2014) that living in closed spaces and 90 in *P. xylostella* (Vieira et al, 2012) that lives in open spaces, reference Dipteran fruit fly, *Drosophila melanogaster* and Japanese encephalitis vector *Culex quinquefasciatus* have been found to have 51 and 110 putative OBPs respectively (Hekmat-Scafe et al., 2002; Manoharan et al, 2013). Previous work on a genome-wide survey of OBPs in mosquito species *Anopheles gambiae, Aedes aegypti,* and *Culex quinquefasciatus* (Manoharan *et al*, 2013) identified more than 100 OBPs, through sensitive computational searches, in each of the mosquito species with subfamily annotation. Here, I have extended the work to identify and annotate OBP and their subfamilies in other mosquito species.

3.2.2 Materials and Methods

In order to identify OBPs across mosquito genomes, mOBP pipeline (**Figure 3.1 A**), initially developed in Python by Ms Charlotte Berthelier and Prof. Bernard Offmann at the University of Nantes, France, was implemented and extended with modification (**Figure 3.1 B**), where the Blastp algorithm was incorporated in place of PSI-BLAST in order to overcome challenges due to infinite looping encountered in case of certain genomes. This was also further modified to perform

and automate the analysis for species across insect orders and compute the resulting analysis to generate reciprocal orthology data across species in a tabular format.

Additionally, an iterative scanning algorithm to detect each of the OBP subfamilies in a given amino acid sequence was written and implemented (**Section 2.2**). The cysteine signatures specific to the Classic, MinusC, PlusC, double domain and atypical OBP subfamilies were considered.

All the analysis above was implemented in Python3.

Two query sets were used parallelly- (a) 'Q_d': OBPs from *Drosophila melanogaster* (Hekmat-Scafe, 2000), and (b) 'Q_combined': dataset in (a) and mosquito OBPs derived from an in-house study (Manoharan *et al*, 2013) on three model genomes - *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. The pipeline was applied to 21 predicted protein datasets (**Table 3.1**) derived from VectorBase as the subject dataset. For genus *Anopheles*, the mosquito species were *albimanus* (AALB), *atroparvus* (AATE), *christyi* (ACHR), *coluzzii, culicifacies*(ACUA), *darling,epiroticus, farauti* (AFAF), *funestus* (AFUN), *gambiae* (AGAP), *maculatus* (AMAM), *melas* (AMEC), *merus* (AMEM), *minimus, punctulatus, quadriannulatus* (AQUA), *sinensis* (ASIS) and *stephensi* (ASTEI). Additional members of order Diptera, *Aedes aegypti*, *Aedes albopictus* and *Culex quinquefasciatus* (CPIJ) were also chosen.

Best reciprocal hits (BRH) were obtained for each query and subject set using the mOBP pipeline described above (**Figure 3.1 B**). For a given subject genome, the ratio of BRH using query 'Q_combined' and BRH using query 'Q_d' yielded OEF, a fold estimate of orthologs unique to the family Culicidae (mosquitoes) within the Order Diptera. This is expressed as the equation below.

$$OEF = BRH(Q\_combined) \,/\, BRH(Q\_d)$$ **..... (Equation 1)**

Here,

OEF = Ortholog fold estimate unique to a family (here, *Culicidae,* Order *Diptera*)

BRH(Q_combined) = Total number of best reciprocal hits (BRH) for a combined query set that consists of queries from a reference organism (here, *Drosophila melanogaster*, family *Drosophilidae,* Order *Diptera*) as well as queries from organism(s) within the family of interest (here, *Culicidae,* Order *Diptera*).

BRH(Q_d) = Total number of best reciprocal hits (BRH) for a query set that consists of queries from a reference organism
(here, *Drosophila melanogaster*, family *Drosophilidae*, Order *Diptera*).

Linear regression was carried out using the 'polyfit' method using the NumPy library in Python v3.6 to determine if the dependent variable (number of orthologous OBPs) is a function of genome size. The binomial equation (y = mx +c) of the best-fit regression line was used to define how variation in genome size explains variation in orthologs of OBPs. Pearson correlation coefficient (PCC) and R-squared values were calculated for each best fit using the 'linregress' method using the SciPy library in Python v3.6.

3.2.3 Results and Discussion

As expected, the hits in mosquito genomes obtained using mosquito OBPs in query set (Q_combined) were greater than those obtained using *Drosophila* OBPs as query set (Q_d).
Among *Anopheles* species, ASTE, AARA, AGAP and ACUA are important vectors of malaria across different ecological niches. AGAP and AARA are part of the *A. gambiae* complex and are widespread vectors in Sub-Saharan Africa. ASTE and ACUA constitute the most successful malaria vectors in Southeast Asia, especially India.
In malaria vectors ACHR, ASTEI, ACUA with a genome size of <300 Mb each and query 'Q_combined', the highest number of amino acid sequences with PBP/GOBP domains have been detected (> 130 sequences each genome). In contrast, query 'Q_d' yielded only 29 sequences for ACHR, not a known disease vector but close outgroup to *A. gambiae* complex, with an OEF score of 5.38, indicating five-fold more orthologs count unique to family *Culicidae* within Diptera with reference to *Drosophila melanogaster*. High OEF scores (>1.5) in ASTEI (OEF score 3.58), an urban malaria vector in India, AALB (OEF score 2.18), a pan-American malarial vector, and AMAM (OEF score 2.09), a malarial vector in hilly S. Asian regions contrasted with low OEF scores (<1.5) in species tending towards a varying degree of specialist behaviour such as AGAP, ASIC, ACUA, AMEM, AMEC. The association of OEF scores with the degree of feeding behaviour is currently being further statistically corroborated through scaled-up analysis. Using queries from a phylogenetic relative, *Drosophila melanogaster*, a number of orthologous culicid

OBPs is weakly positively correlated (PCC=0.53, r2=0.28) with a genome size of mosquito species (**Figure 3.2**).

Detection of cysteine signatures specific to insect OBP revealed that the frequency of Minus-C OBPs is higher than Classic and Plus-C OBPs (**Figure 3.3**), suggesting that OBP subfamilies may be higher across certain insect orders based on ecological niche, functionality and ligand specificity.

This work suggests that the presence of cysteine signatures thought specific to insect OBPs may not exclusively co-occur with PBP/GOBP domain as previously believed, and ongoing work is statistically corroborating the same.

This work done through the mOBP pipeline is currently being analyzed in conjunction with the SoCCer pipeline (**Chapter 2: Section 2.2**) with scaled-up data of various insects that are disease vectors.

3.3 GWS of OBPs in honeybee

3.3.1 Introduction

Previous work in our laboratory (Karpe *et al*, 2016) has identified odorant receptors (ORs) in *Apis florea* using an exhaustive genomic pipeline (Karpe *et al*, 2016; 2019; 2020). In order to complement the search of ORs towards a better understanding of odour coding, this study investigated odorant-binding proteins (OBPs) in *Apis florea*.

Apis florea, or the red dwarf honey bee, exhibits the complex behavior of eusociality, where there is reproductive division of labour within a colony that comprises a female queen, male drones and female worker bees. While worker bees perform important tasks such as foraging, guarding the colony hive, maintenance and other diverse tasks for the colony, the queen and drone perform reproductive roles (Page and Robinson, 1991).

Members of the species exhibit haplodiploidy (Halling *et al*, 2001) system of genetic inheritance, where the males in this species are haploid, possessing half the number of chromosomes as diploid females. *Apis florea* is geographically distributed with a preference for warm climate (Otis, 1991) in regions such as mainland Asia, southern border of the Himalayas, plateau of Iran, Oman and in Vietnam, southeast China and peninsular Malaysia (Hepburn *et al*, 2005; Oldroyd and Nanork, 2009; Moritz et al, 2010) and display open nesting typically on low-lying tree branches in shaded regions (Wongsiri *et al*, 1997; Hepburn *et al*, 2005). *Apis florea* is an important pollinator of tropical and ornamental plants as well as agricultural crops. They primarily feed on pollen and nectar from flowering plants. Like other honey bees, the body of *Apis florea* is studded with various types of sensilla, among which olfactory sensilla (sensilla basiconica and sensilla chaetica) are prominent structures (Gupta, 1992). The antenna of the insect is typically the main site for olfactory receptors (Wigglesworth, 1965). The antennae of *Apis florea* harbor hair-like sensillae trichodea types I, II, III, IV, sensilla basiconica, sensilla placodea and sensilla ampullaceal (Gupta, 1992; Kumar *et al*, 2014).

3.3.2 Materials and Methods

Obtaining genome of honeybee *Apis florea*

Aflo_1.1 genome was obtained from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov).

Preparing query dataset from *Apis mellifera*

AmelOBPs were pooled from the NCBI non-redundant protein database (29 putative AmelOBPs) and a previous study (Foret and Malesczka, 2007; 21 AmelOBPs) to obtain a filtered set of query protein sequences. Reciprocal homology was performed using the query set obtained and AmelOBPs from a recent study (Vieira *et al*, 2011; 21 AmelOBPs). An e-value cutoff of e^-10 was used. The resultant matches as well as unmatched OBPs (putative OBPs with no reciprocal hit; 10 protein sequences) resulted in a final dataset of annotated and unique AmelOBPs.

Query protein to subject genome alignments

Genomic alignments were obtained using Exonerate (Slater & Birney, 2005) with intron sizes of 500, 2000, 5000 and 10000, respectively, with BLOSUM62 (Henikoff and Henikoff, 1992) as the substitution matrix.

The genomic alignments were processed as per the methodology in a previous in-house study from the lab (Karpe *et al*, 2016; 2017; 2020). The pipeline involves thoroughly scanning and scoring alignments to the genome based on length, degree of similarity and the best match of the scaffold location in the subject genome to the query sequence. The unique set of genomic alignments was then processed further to translate amino acids from corresponding in-frame codons. The resultant set of gene models and protein sequences were also manually corrected for missing start and stop codons, missing N-terminal and C-terminal amino acids and annotated as Complete, Partial or pseudogene.

Homology-based validation & nomenclature

The predicted *Apis florea* OBPs (AfloOBP) were subjected to reciprocal homology with our manually curated AmelOBP dataset, as explained above. The final dataset of predicted AfloOBPs comprised resultant matches as well as unique sequences with no corresponding hits found in the AmelOBP dataset. The AfloOBP predicted protein sequence dataset was thus annotated with respect to AmelOBP homolog if present, as well as its status as '**Complete**' or '**Partial**'.

Secondary structure prediction

The secondary structure of the protein sequences was predicted using neural network-based PSIPRED v3.2 (Conesa *et al*, 2005; Buchan *et al*, 2013).

Signal peptide detection

N-terminal signal peptide was detected using SignalP4.1 (Nielsen et al, 1997; Petersen et al, 2011). This algorithm uses neural networks and Hidden Markov Models to determine signal peptides in a given protein sequence. The predicted signal peptide for a given sequence was cleaved off, and the "mature" sequence was used for multiple sequence alignment and phylogeny.

Preparing dataset of insect OBPs for rooted and unrooted phylogeny

In order to prepare an outgroup for the rooted phylogeny, annotated chemosensory proteins of *Apis mellifera* (AmelCSPs) were obtained from a previous study (Forêt, Wanner and Maleszka, 2015), namely, AmelCSP1, AmelCSP2, AmelCSP3, AmelCSP4, AmelCSP5 and AmelCSP6.

In order to construct the phylogeny (Vogt, Große-Wilde and Zhou, 2015; Missbach, Vogel, Hansson and Große-Wilde, 2015), protein sequences of OBPs from 11 insect orders from representative insect species were obtained from previous literature and UniProt (The UniProt Consortium, 2019) database. The insect orders, corresponding species, and the number of species-specific OBPs have been tabulated as in **Table 3.2**.

Structure-based sequence alignment and phylogenetic analysis

A structure-based seed template was obtained from the PASS2.5 database (Gandhimathi *et al*, 2012) with the SCOP ID of the fold as 47565. The dataset of "mature" AfloOBP sequences was aligned against the seed template using MAFFT (Katoh *et al* 2002; 2013). A phylogenetic tree was constructed using RaxML (Stamatakis *et al* 2006; 2014) with the maximum likelihood method with 100 bootstraps and the WAG evolutionary model (Whelan and Goldman, 2001). The phylogenetic tree was annotated and visualized using iTOL (Letunic and Bork, 2006; 2016).

3.3.3 Results and Discussion

We filtered and re-annotated OBPs from closely related reference genome *Apis mellifera* using a homology-based approach (See Methods). The final dataset (SI Table 1 in BioRxiv DOI) comprised 25 AmelOBP protein sequences.

Our genome-wide survey of *Apis florea* revealed 22 novel OBP protein sequences with 15 complete and 7 partial sequences either towards the N-terminus, C-terminus, or both with an average exon number of 5 (Table 2 BioRxiv DOI). Secondary structure analysis revealed OBPs with high confidence. Typically, 6-7 alpha-helices per complete AfloOBP sequence was predicted.

Out of 15 *AfloOBP* genes predicted as complete, manually corrected and annotated, 16 translated protein sequences were predicted to have signal peptide sequences. The average length of signal peptides predicted in our AfloOBP dataset was 19 amino acids. Cleavage position ranged from 16th to 24th amino acids in the sequence.

Sequences AfloOBP1-AfloOBP13 were found to display the conserved cysteine signature of Classic and Minus-C subfamilies, like their orthologs in *Apis mellifera,* and comparable to that of AgamOBP (**Figure 3.4)**. Multiple sequence alignment revealed conserved cysteine profiles specific to Classic and MinusC subfamilies in the ***Apis florea*** genome (**Figure 3.5**). Sequences AfloOBP14-AfloOBP21 were found to show the conserved Minus-C cysteine signature where cysteine residues in the conserved second and fifth positions are missing. Our analysis shows that the conserved cysteine signature for both subfamilies in ***Apis florea*** is similar to the representative signature observed in a previous study (Xu *et al*, 2009). The conserved cysteine signature for the Classic subfamily for the Hymenopteran insect order was determined as C1-X **23:35**-C2-X**3**-C3-X 27:45-C4-X 7:14-C5-X**8**-C6 (Xu *et al*, 2009). Our study has identified 13 Classic and 9 Minus-

C OBPs in ***Apis florea***. We observe the Classic cysteine signature to be conserved similarly as C1-X **27:37**-C2-X**3:4**-C3-X 33:43-C4-X 9:13-C5-X**8-9**-C6.

Phylogenetic inference revealed the clustering of Minus-C OBPs as a sub-clade of the Classic OBP subfamily comprising members of both ***Apis mellifera*** and ***Apis florea*** OBPs (**Figure 3.6 B**). Moreover, a conserved cysteine signature specific to the chemosensory protein (CSP) family was observed in the outgroup chemosensory proteins (AmelCSP) (**Figure 3.5**). AmelCSPs used as outgroup clustered distinctly (**Figure 3.6 A**) from the odorant-binding proteins input to the phylogeny with 100% bootstrap value. Minus-C OBPs were found to cluster together with 60% bootstrap value closest to AfloOBP9, annotated as a Classic OBP. OBPs of the Minus-C subfamily, AfloOBP 14-20 emerge closest to AfloOBP13, a Classic OBP with an observed six cysteine signature. Interestingly, all the other Classic OBPs cluster distinctly in a clade corresponding to the insect Classic subfamily. However, AfloOBP13 clusters closely with the Minus-C group in a distinct sub-clade suggesting an evolutionary ancestral link (**Figure 3.6 A)**. Interestingly, antennal OBP (MsexABP1) from Lepidopteran insect *Manduca sexta* clustered close to the Minus-C clade along with other bee species (Hymenopteran) with high bootstrap support of 97%. It is also observed that Classic OBPs in ***Apis florea*** are phylogenetically distant from Minus-C (bee OBPs) than clades representing Atypical OBPs in Dipterans and Plus-C insect OBPs. This suggests that Minus-C OBPs in honey bees may have evolved from a single ancestral Classic OBP (similar to AfloOBP13, AmelOBP13) of its species by deletion of second and fifth cysteines. The evolution and insect order-specific occurrence of Minus-C, Plus-C, and Atypical subfamilies of insect OBPs may have functional roles and would be interesting to investigate.

The above observations from a comprehensive bioinformatic analysis strongly suggest that Minus-C OBPs are likely to have evolved from a Classic OBP subfamily member in insects. It is possible that the evolution of a subfamily could be an adaptation to the local niche of the insect species for functional specificity (Zhou *et al*, 2020).

Tables and Figures

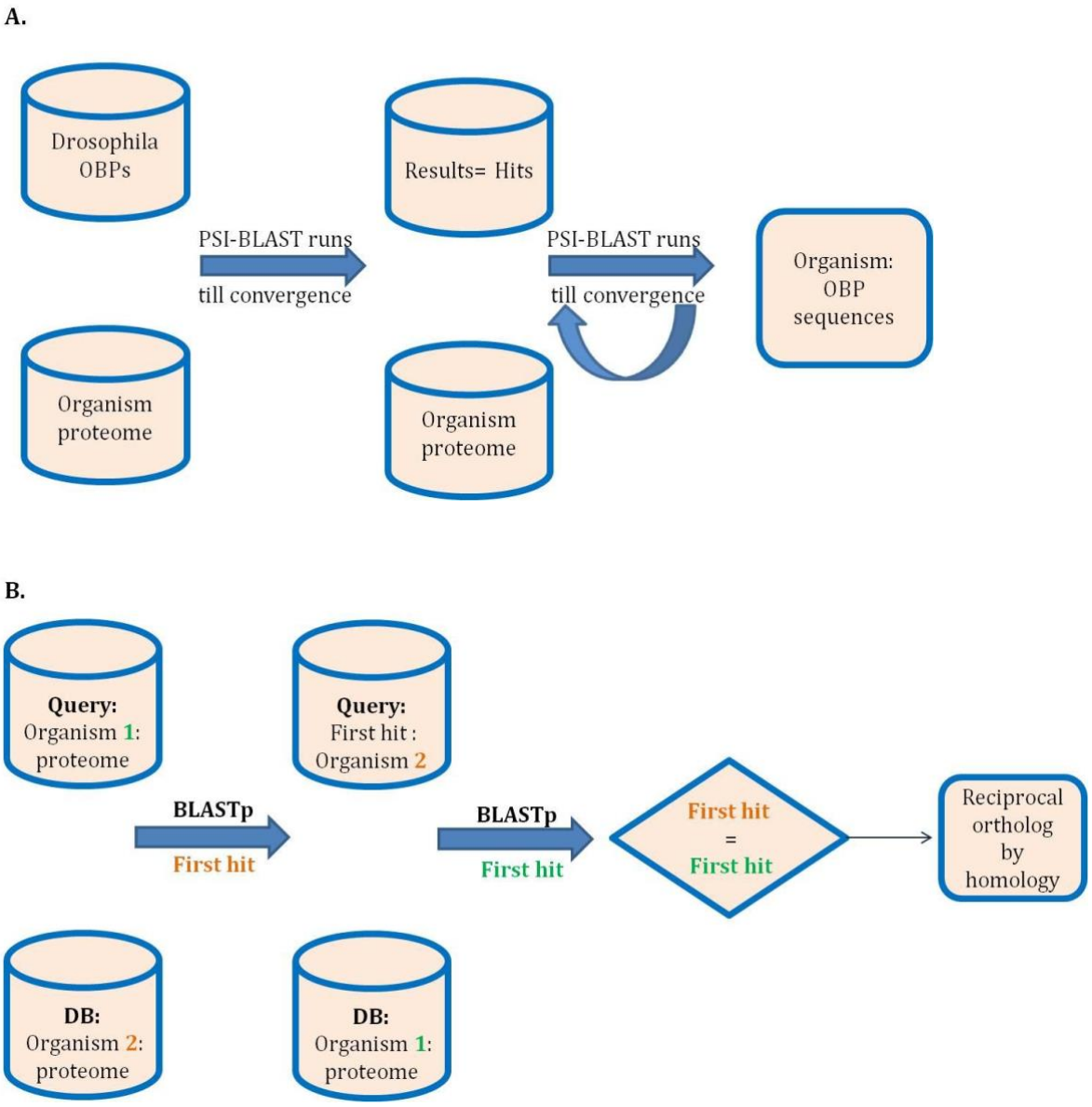


Figure 3.1: Workflow of mOBP pipeline (A) originally developed by Charlotte Berthelier and Prof. Bernard Offmann, and (B) extended and modified version as part of my thesis work

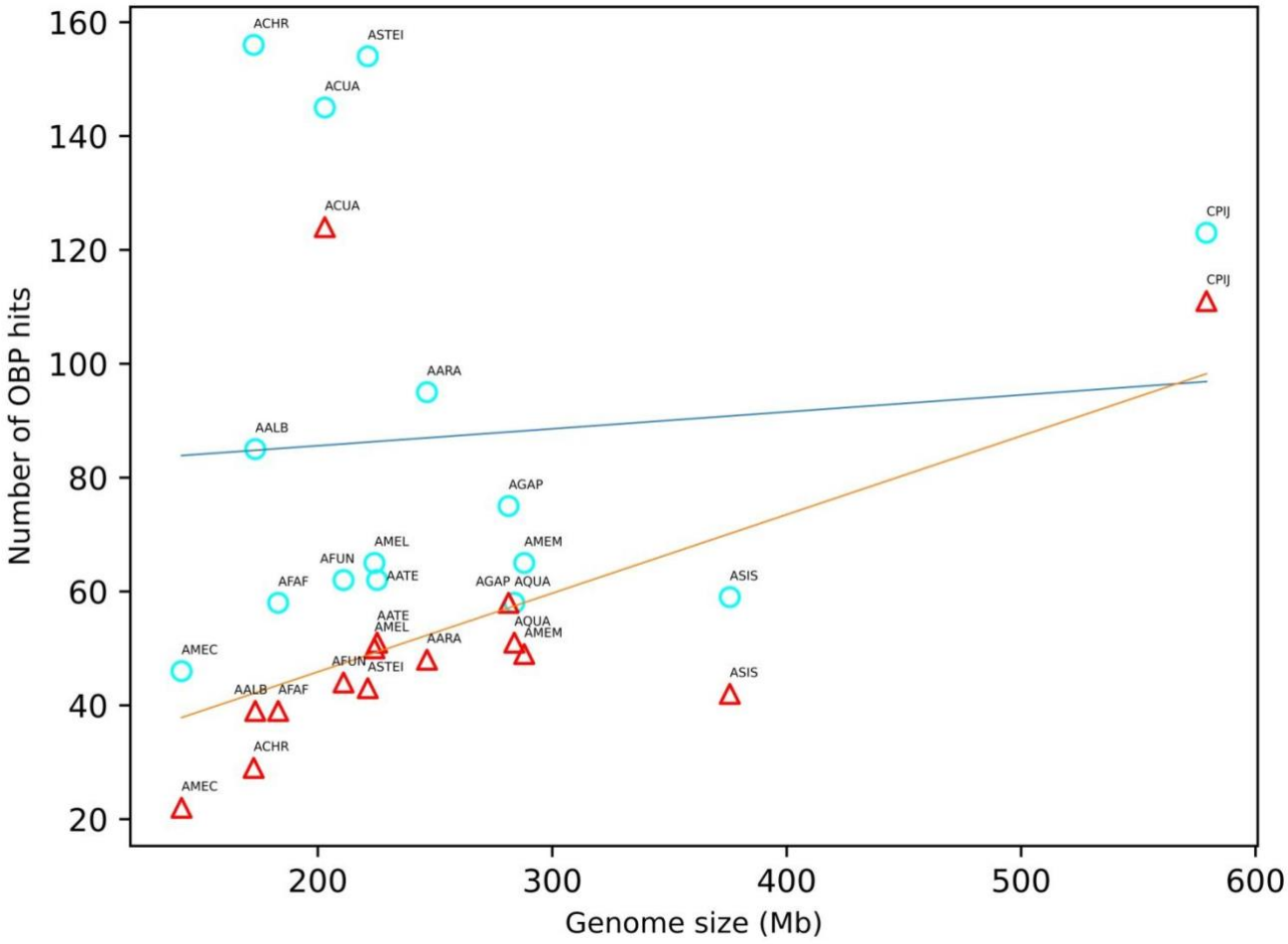


Figure 3.2: Comparison of total number of orthologous OBPs and genome size for mosquito species: number of OBPs that were obtained as best reciprocal hits and genome size in Mb is plotted for 15 mosquito species, Order Diptera. The number of orthologous hits with query sets as OBPs from Drosophila melanogaster (Q_d; triangles), and mosquito OBPs (Manoharan et al*, 2013) along with Q_d (Q_combined; circles) respectively have been represented. Pearson correlation coefficient (PCC) values corresponding to first-order regression for the two datasets is indicated in orange (Q_d; PCC=0.54) and blue (Q_combined; PCC=0.08). Mosquito species have been annotated as per VectorBase identifiers.*

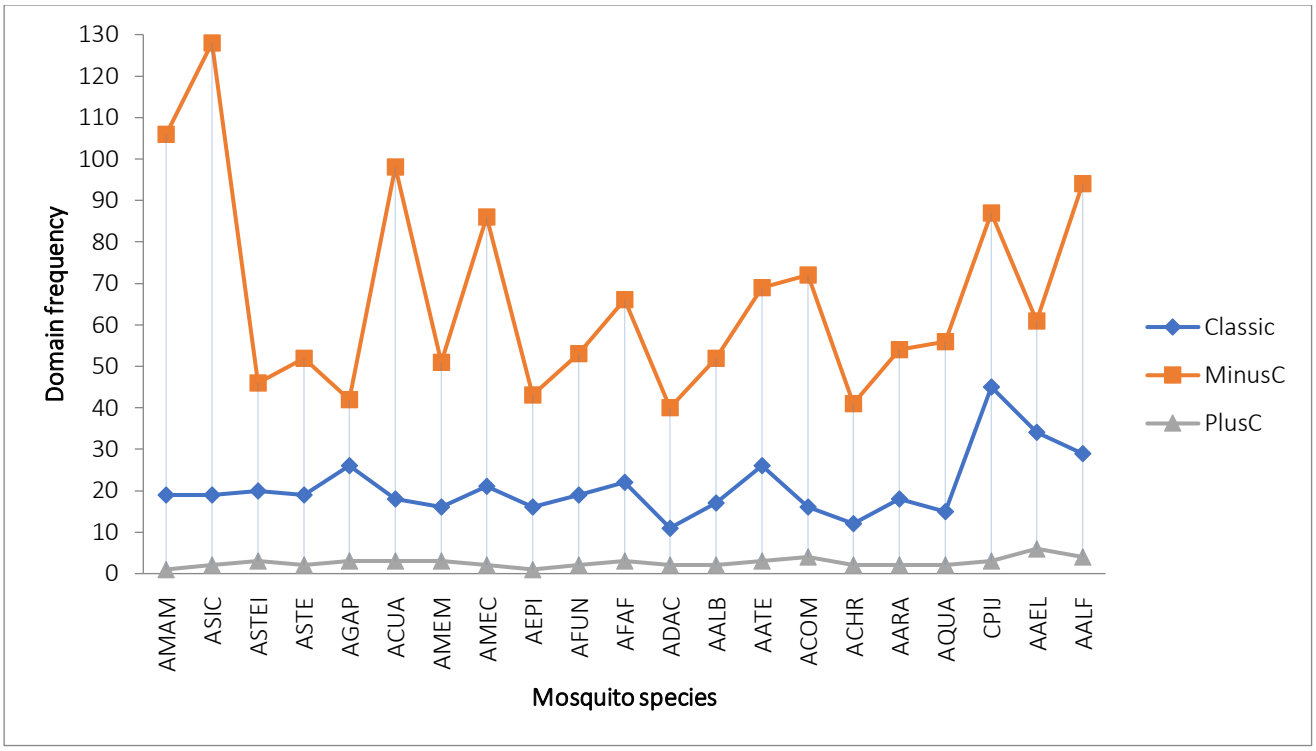


Figure 3.3: Cysteine signatures specific to OBPs across mosquito species show a consistent trend in distribution of subfamilies

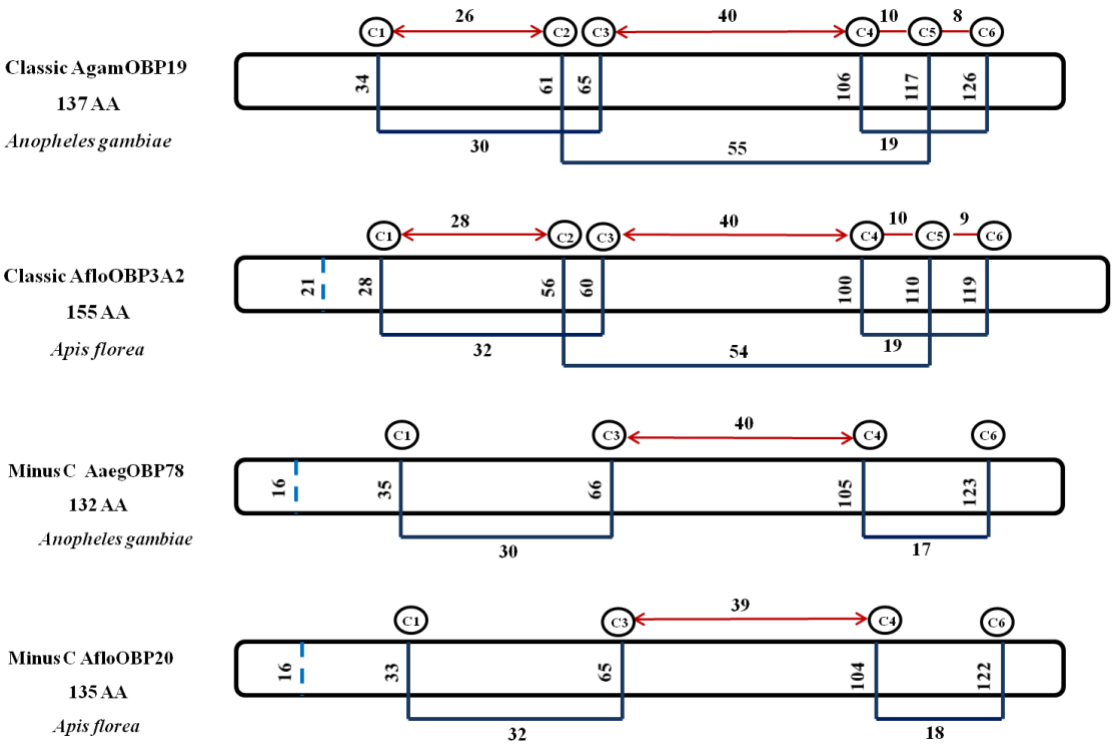


Figure 3.4: cysteine signatures across insect orders are comparable. cysteine residue positioning and inter-disulphide spacing is conserved in Apis florea (Hymenoptera) genome across Classic and Minus-C subfamilies. The cysteine signature of OBPs from Anopheles gambiae is given for reference.

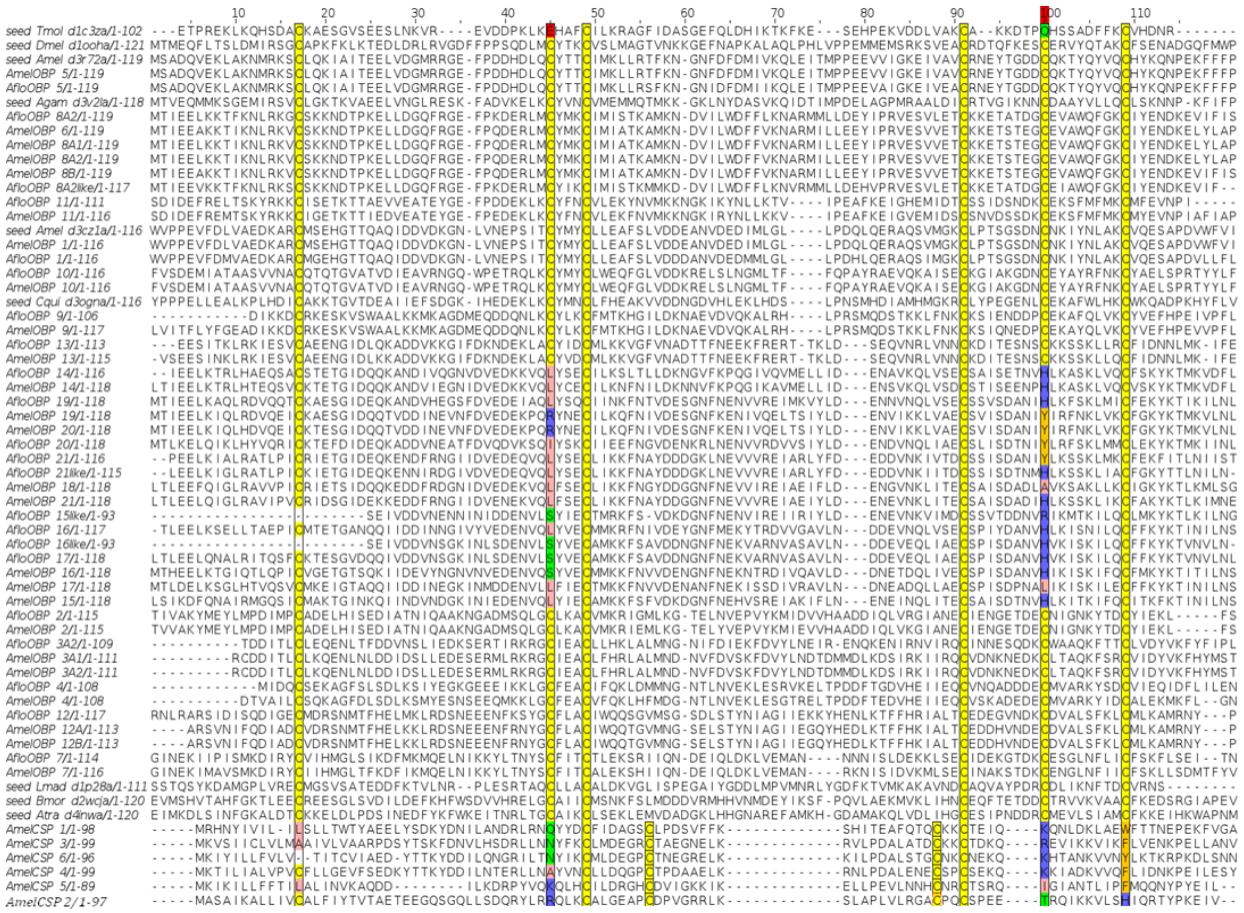
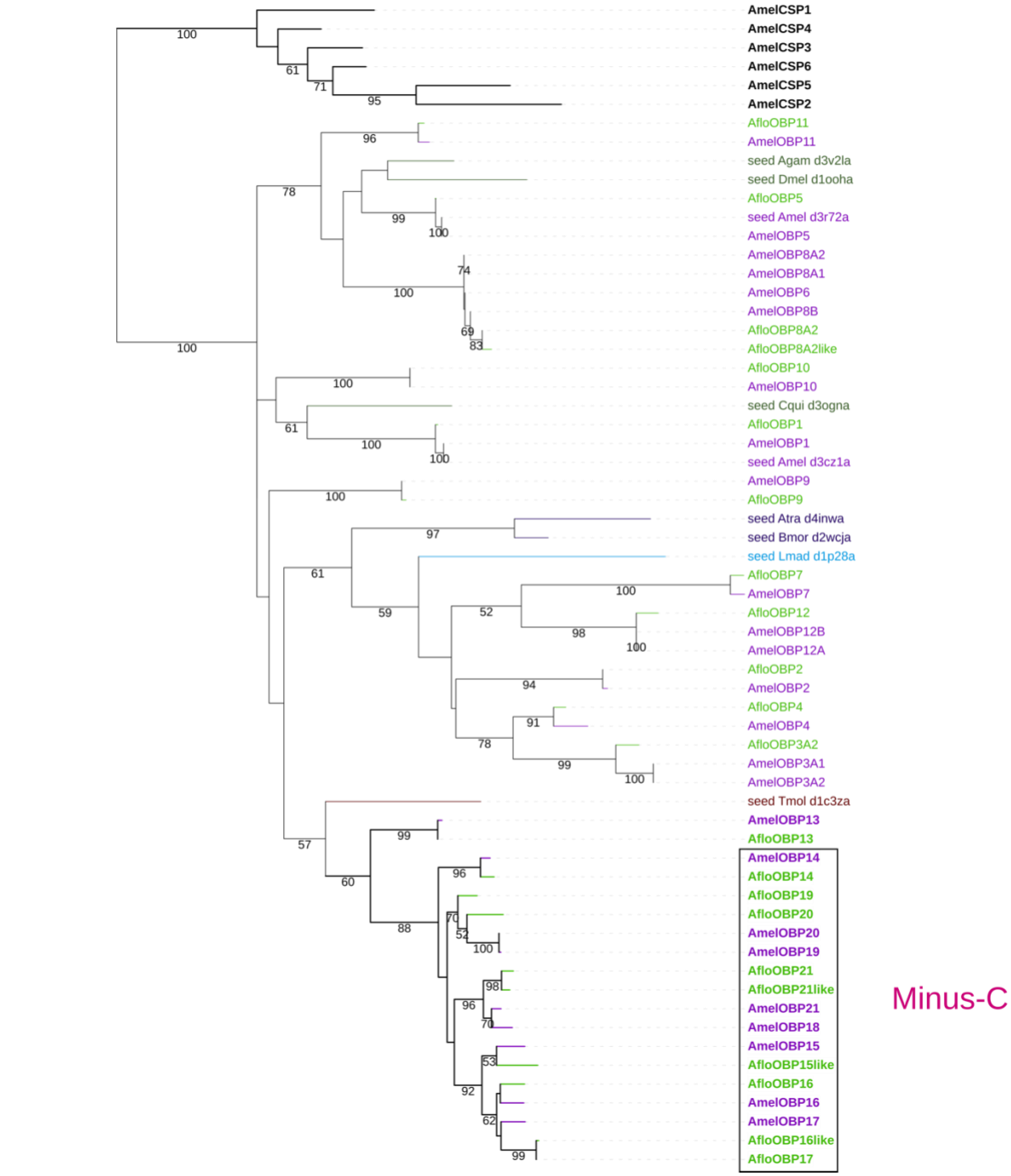


Figure 3.5: Multiple sequence alignment of OBPs from Apis florea. AfloOBPs. The alignment also contains OBPs from its phylogenetic neighbour Apis mellifera (AmelOBP). Chemosensory proteins from Apis mellifera (AmelCSPs) are present as an outgroup.

A.


Tree scale: 1
AmelCSP1
AmelCSP4
AmelCSP3
AmelCSP6
AmelCSP5
AmelCSP2
AfloOBP11
AmelOBP11
seed Agam d3v2la
seed Dmel d1ooha
AfloOBP5
seed Amel d3r72a
AmelOBP5
AmelOBP8A2
AmelOBP8A1
AmelOBP6
AmelOBP8B
AfloOBP8A2
AfloOBP8A2like
AfloOBP10
AmelOBP10
seed Cqui d3ogna
AfloOBP1
AmelOBP1
seed Amel d3cz1a
AmelOBP9
AfloOBP9
seed Atra d4inwa
seed Bmor d2wcja
seed Lmad d1p28a
AfloOBP7
AmelOBP7
AfloOBP12
AmelOBP12B
AmelOBP12A
AfloOBP2
AmelOBP2
AfloOBP4
AmelOBP4
AfloOBP3A2
AmelOBP3A1
AmelOBP3A2
seed Tmol d1c3za
AmelOBP13
AfloOBP13
AmelOBP14
AfloOBP14
AfloOBP19
AfloOBP20
AmelOBP20
AmelOBP19
AfloOBP21
AfloOBP21like
AmelOBP21
AmelOBP18
AmelOBP15
AfloOBP15like
AfloOBP16
AmelOBP16
AmelOBP17
AfloOBP16like
AfloOBP17
Minus-C

B.

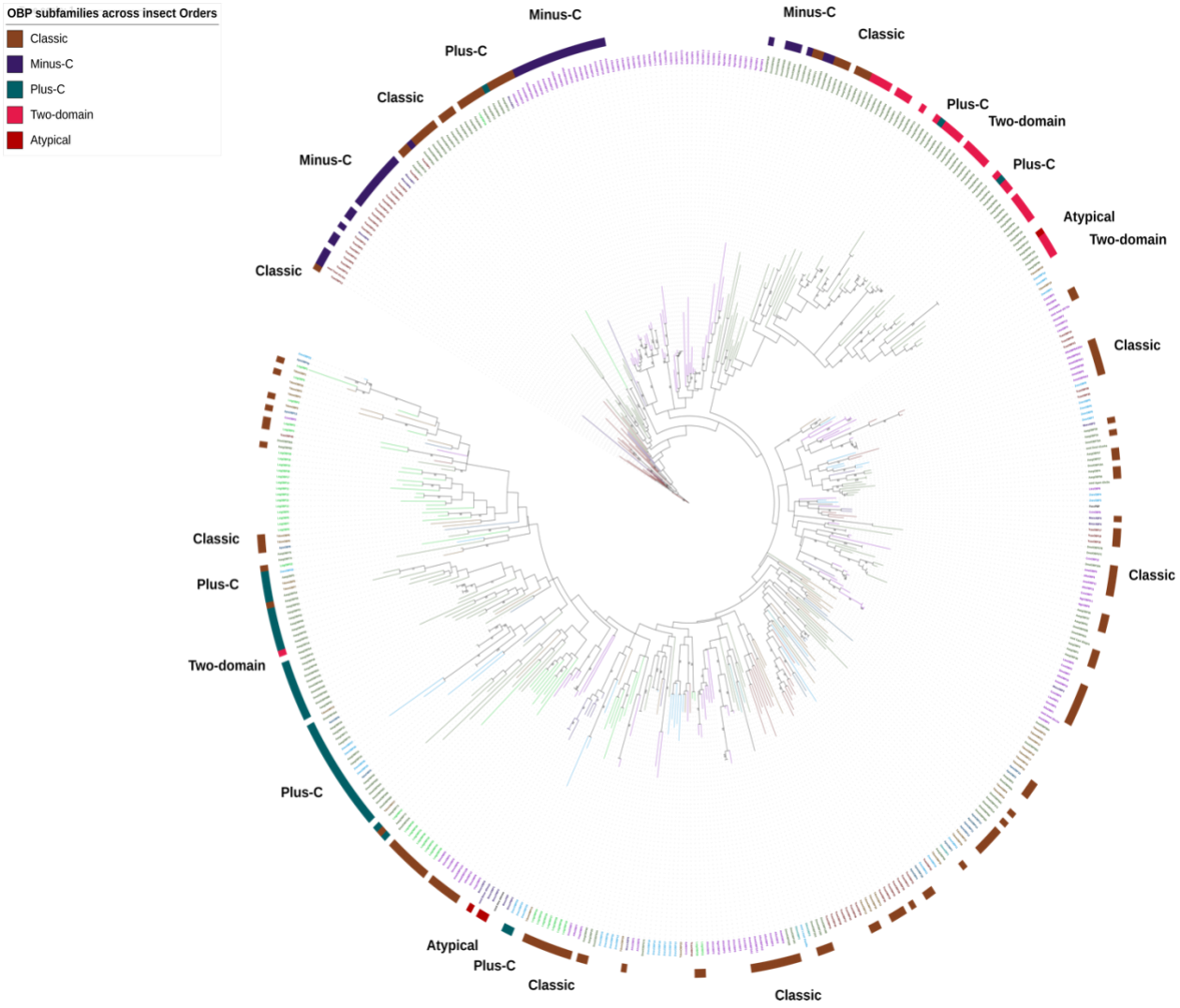


Figure 3.6: Phylogenetic tree of ***Apis florea*** *OBPs with other insect OBPs*

Rooted phylogeny of **(A)** OBPs in sister species, ***Apis florea*** *(*Aflo; in pink*)* and ***Apis mellifera*** (Amel; in purple). Members of the alignment template are colored in green, whereas the outgroup consisting of ***Apis mellifera*** chemosensory proteins (AmelCSP) is colored in red. The bootstrap values of the branches are indicated on the nodes in percentage values. **Unrooted phylogeny (B)**

of OBPs from representative members of **11 insect orders** represent clade denoting Classic subfamily is colored in **brown,** Atypical in **red**, Minus-C in **violet** and Plus-C in **cyan**. The outer circle denotes member clades. The inner branch colors and label colors are colored as per order. Hymenoptera is denoted in **violet**. The bootstrap values of the branches are indicated on the nodes in percentage values.

Table 3.1: OBPs with PBP/GOBP domains identified in mosquito species studied through the mOBP pipeline. Additional information on species annotation, habitat and vectorial transmission have been obtained from VectorBase and scientific literature.

Mosquito species	**ID**	**Region**	**Remarks**	**PBP/GOBP domains detected**	**OEF**
Anopheles maculatus	AMAM	SE Asia NE India, hilly regions, Indian subcontinent	Malaria vector	45	2.09
Anopheles culifacies	ACUA	N-S India, Central,rural-peri-urban	Major malaria vector; drug resistance	54	1.17
Anopheles sinensis	ASIC	SE Asia	Important vector of vivax malaria & filiarisis	52	1.40
Anopheles steph. India	ASTEI	Urban areas, NW India.	Malaria vector	60	3.58
Anopeheles steph. SDA	ASTE	Strain from Pakistan	Malaria vector	49	--
Anopheles gambiae	AGAM	Sub-Saharan Africa	Efficient malaria vector.	70	1.29

Anpheles merus	AMEM	East & S Africa-coastline	Agam complex, malaria vector	58	1.33
Anopheles melas	AMEC	Coastal W Africa	Agam complex, malaria vector	58	1.30
Anopheles epiroticus	AEPI	Thailand, Cambodia	Malaria vector	52	--
Anopheles funestus	AFUN	Africa	Malaria vector	58	1.41
Anopheles farauti	AFAF	Papua N. Guinea, Cambodia & Australia	Malaria vector	52	1.49
Anopheles darlingi	ADAC	Neotropics (Mexico, Central and South America)	Malaria vector	37	--
Anopheles albimanus	AALB	Central, N & S America	Malaria vector	45	2.18
Anopheles atroparvus	AATE	Europe	Was main malaria vector in Europe	60	1.22
Anopheles coluzzi	ACOM	Mali	Agam complex, Malaria vector	57	--
Anopheles christyi	ACHR	Kenya	Closely related to Agam complex.	37	5.38
	AARA	Sudan	Agam complex, Malaria vector	63	1.98

Anopheles arabiensis					
Anopheles quadriannulatus	AQUA	South Africa	Agam complex Nearly entirely zoophilic	56	1.14
Culex quinquefasciatus	CJIP	Worldwide, Tropical & sub-tropical.	Vector for encephalitis, filiaris	127	1.11
Aedes aegypti	AAEL	Tropical, sub-tropical & temperate regions	Vector for yellow fever, dengue, Chikunngunya	109	--
Aedes albopictus	AALF	China, Asia and ~ 30 other countries.	Vector for encephalitis, West Nile fever, dengue, Chikungunya	120	--

Table 3.2: List of insect orders represented in the phylogenetic tree of insect OBPs: The protein sequences of OBPs were obtained from an exhaustive literature survey referenced alongside.

Sr.no.	**Order**	**Reference**
1.	Archaeognatha	Missbach *et al*, 2015
2.	Blattaria	Xu *et al*, 2009
3.	Coleoptera	Xu *et al*, 2009; Gu *et al*, 2015
4.	Diptera	Manoharan *et al*, 2013; Hekmat-Scafe *et al*, 2002, NCBI
5.	Hemiptera	Xu *et al*, 2009; Zhou *et al*, 2010

6.	Hymenoptera	Xu *et al*, 2009; Donnell *et al*, 2013; Gress *et al*, 2014; Li *et al*, 2015
7.	Isoptera	Terrapon *et al*, 2014; NCBI
8.	Lepidoptera	Gong *et al*, 2009; Xu *et al*, 2009; Li *et al*, 2015
9.	Orthoptera	Xu *et al*, 2009
10.	Thysanoptera	NCBI
11.	Zygentoma	Missbach *et al*, 2015

Additional Files:

The following files have been hosted on bioRxiv linked here-

DOI: 10.1101/2021.02.25.432941

a. **Table -** OBPs from ***Apis florea*** annotated in the study have been listed with the scaffold identity, coding exons, complete or partial status of predicted protein sequence, length of protein sequence and its ortholog in ***Apis mellifera*** (Table 2 in bioRxiv paper).

b. Supplementary File 1- SI_Table1
Table of OBPs from reference organism *Apis mellifera* derived from various literature sources and re-annotated to obtain a standard dataset

c. Supplementary File 2- SI_Table2
Table of predicted OBPs from *Apis florea* with annotation

d. Supplementary_File 3- SI_Table3
Table of blastp alignment hits obtained with query as our annotated set of AfloOBPs against Non-Redundant database.

Chapter 4

Heterodimeric interactions in mosquito OBPs

4.1 Introduction

There have been two models proposed to describe the interplay between OBPs and ORs. The first model, observed in moth and mosquito OBPs, proposes that an OR can be activated by binding an odorant (Pelosi et al., 2006). This pattern supports that odorant uptake and release by the OBPs is highly pH-dependent. OBPs bind odorants at a basic pH, transport them, and finally release them in the acidic environment (pH 4.5) of ORs in the dendrite region (Di Luccio, Ishida et al., 2013, Lautenschlager, Leal et al., 2005). Experimental findings showing that ORs are activated by their specific ligands in the absence of OBPs (Nakagawa, Sakurai et al., 2005) are consistent with this model. Furthermore, studies performed on *Bombyx mori* PBP1 have shown that change of pH from neutral to acidic promotes conformational changes such as the folding of the C-terminus into a seventh α-helix which enters the binding cavity, leading to the displacement of the bound pheromone molecule (Damberger, Nikonova et al., 2000, Sandler, Nikonova et al., 2000). Analogous pH-dependent conformational changes have been observed in other insect OBPs, including the orthologous OBP1 proteins of *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus* mosquitoes (Han, Zhang et al., 2014, Leite, Krogh et al., 2009, Mao, Xu et al., 2010, Wogulis, Morgan et al., 2006, Zubkov, Gronenborn et al., 2005). However, in these proteins, the short C-terminal loop does not form an extended α-helix but a wall over the binding pocket, where it is held in place by hydrogen bonds and polar interactions. Disruption of these contacts at low pH may lead to the opening of the C-terminal "lid", thus exposing the binding site and the bound odorant to the solvent.

The OBP-odorant complex is transported as a unit through the lymph as per another model, activating the OR. Following this model, *Drosophila melanogaster* OBP LUSH was proposed to mediate the activation of the receptor OR67d through conformational changes that occur only when the pheromone 11-cis vaccenyl acetate (cVA) is bound to the protein, and which are being recognised by the receptor (Laughlin, Ha et al., 2008). However, this hypothesis has been recently questioned study showing that a mutant version of LUSH which mimics the cVA-induced conformational changes does not activate OR67d neurons (Gomez-Diaz, Reina et al., 2013). The authors consequently suggested that cVA molecules alone activate the OR, thus supporting that the ligand must be released from the OBP before receptor activation.

Insect OBPs undergo conformational changes due to pH changes and ligand binding (Han et al., 2014, Ziemba, Murphy et al., 2013). The structural modifications of OBPs caused by a change in pH were found to affect their specific binding properties, altering their interactions with ligands (Brito et al., 2016). In addition, ligand-induced conformational changes in OBPs have been suggested to be implicated in the regulation of OBPs heterodimeric interactions (Davrazou, Dong et al., 2011, Wei, Duan et al., 2020).

The heterodimerisation of OBPs has been demonstrated in mosquitoes using co-immunoprecipitation and cross-linking studies (Andronopoulou, Labropoulou et al., 2006) and other insect species (Wang, Guan et al., 2013, Wei et al., 2020). The formed OBP heterodimers exhibit novel binding properties, different from those of the individual monomers, thereby recognising a more comprehensive range of odorants. Specifically, it was found that *A. gambiae* AgamOBP1 and AgamOBP4 can interact directly, forming heterodimers (Andronopoulou et al., 2006). Further competitive binding studies showed that these proteins preferentially bind several structurally related compounds (Qiao, He et al., 2011). AgamOBP1 was reported to be requisite for olfactory responses to two major components of human sweat, indole, and 3-methyl-indole
(Biessmann et al., 2010) while its heterodimeric interaction with AgamOBP4 plays a crucial role in indole perception (Davrazou et al., 2011). Moreover, whole-mount fluorescence in situ hybridisation (WM-FISH) experiments indicated that AgamOBP1, AgamOBP4, and one of the most narrowly tuned olfactory receptors which is strongly activated by indole (Carey, Wang et al., 2010), the AgOR2, are co-localised in the same sensillum (Schultze, Pregitzer et al., 2013). It supports the notion that these two "ligand-matched" OBPs may functionally cooperate with AgOR2 to mediate the specific recognition of indole.

As was proposed by Davrazou et al (2011), the apo-AgamOBP4 exists in a molten globule state in solution. However, the binding of indole to the protein promotes conformational changes that alter its structure to a well-ordered state, leading to forming a binding site for AgamOBP1. In the absence of indole, no significant interactions were observed between AgamOBP4 and AgamOBP1 (Davrazou et al., 2011).

Consequently, in order to investigate how these two OBPs interact in the presence of indole and to explore the possibility of a pH-dependent mechanism for their heterodimer formation/dissociation, we solved the crystal structures of AgamOBP4 at pH 4.6 and 8.5 and undertook comparative thermodynamic studies of AgamOBP4 at pH 5.0 and 8.0. Furthermore, we performed protein-protein docking and MD simulations using the crystal structures of AgamOBP1 (Wogulis et al., 2006), the indole-bound form of AgamOBP4 (pH 6.85) (Davrazou et al., 2011), and our apo- AgamOBP4 structures at two different pH values (4.6 and 8.5). Analysis of the models of AgamOBP1-AgamOBP4 heterodimer, generated by guided docking at acidic, neutral and alkaline pH in this work shows that the decrease of pH is associated with a concomitant reduction in the internal dynamics of the heterodimer. Our results support a pH-dependent mechanism of AgamOBP1-AgamOBP4 heterodimerisation promoted by indole wherein a neutral or basic environment would be conducive for heterodimerisation. Our findings highlight the collective impact of pH and indole on the interactions at the interface, intra- and inter-monomeric motions, local conformational changes, thus further supporting the hypothesis of a pH-dependent odorant release mechanism in the sensilla.

4.2 Materials and Methods

AgamOBP4 monomers for protein-ligand dynamics

Monomers of AgamOBP4 were obtained directly from the structural coordinates provided at pH 8.5 (PDB: 3N8C) and pH 4.5 (PDB: 3N8A). Indole was docked to the binding cavity of OBP4 as mentioned in Davarazou et al, 2011 with a previously deposited model (PDB ID: 3Q8I) as reference for the binding site coordinates. Side-chain protonation states of residues were equilibrated to pH 4.5, 6.5 and 8.5 using PROPKA in GROMACS v5.14 (Abraham, Murtola et al., 2015; Bekker, Berendsen et al., 1993; Berendsen, Vanderspoel et al., 1995; Van der Spoel, Lindahl et al., 2005) wherever applicable such as for apo-OBP4 at pH 6.5.

Heterodimer modelling

The AgamOBP1-AgamOBP4 heterodimer model at neutral pH was constructed using the crystal structures of AgamOBP1 and AgamOBP4 monomers. Structural coordinates of AgamOBP1 at pH 8.0 were retrieved from the RCSB PDB database (PDB: 2ERB; Wogulis et al., 2006). The structural coordinates of apo-AgamOBP4 at pH 8.5 (PDB: 3N8C) were used.

Side chain protonation states of residues were equilibrated to pH 6.5 using PROPKA in GROMACS v5.14 (Abraham, Murtola et al., 2015; Bekker, Berendsen et al., 1993; Berendsen, Vanderspoel et al., 1995; Van der Spoel, Lindahl et al., 2005).

Protein-protein docking was performed using the Haddock software (van Zundert, Rodrigues et al., 2016). The residues of both AgamOBP1 and AgamOBP4 that were found to display chemical shift changes upon mutual interaction in the presence of indole (**Table 4.1**) in previous NMR experiments (Davrazou et al., 2011), were used to guide the docking of AgamOBP1 and AgamOBP4. The top clusters were identified, and docked poses obtained were provided as input to the PPCheck algorithm to identify the top-ranked poses according to their energetics at the interface. The top clusters were selected based on their Z-score. Cluster members of the top three clusters were identified and analysed for interface stability, the orientation of helices and intact disulfide bonds.

Accordingly, the protonation states of amino acid side-chains of the selected heterodimeric model at neutral pH were assigned to acidic, neutral and alkaline pH using PROPKA in GROMACS v5.14 (Abraham et al., 2015, Bekker et al., 1993, Berendsen et al., 1995, Van der Spoel et al., 2005), in order to generate the AgamOBP1-AgamOBP4 models at pH 4.5, 6.5 and 8.5.

Validation of heterodimer models

The generated models of AgamOBP1-AgamOBP4 heterodimer were then compared concerning their interacting interface residues, interface energy, and stability of the complex. Furthermore, energetic analysis of the interacting interfaces of the models was carried out using the PPCheck platform (Sukhwal & Sowdhamini, 2015).

MD simulations of the AgamOBP4 monomer and AgamOBP1-AgamOBP4 heterodimeric models

MD was carried out for a production run period of 200 nanoseconds and in replicates (n=3). GROMACSv5.1.4 (Abraham et al., 2015, Bekker et al., 1993, Berendsen et al., 1995, Van der

Spoel et al., 2005) was used for simulations. For each simulation, an octahedron box filled with SPCE water model was generated with dimensions of 1.5 nm. For atomic representation, GROMOS54a7 force-field parameters were selected. Periodic boundary conditions (PBC) were applied in all directions to mimic the infinite system, and a box type and dimension were appropriately selected for the globular protein system. The Particle Mesh Ewald approach was employed for long-range electrostatics using a unit cutoff distance for electrostatic and van der Waals interactions. The steepest gradient descent algorithm was used to minimise the system while generating ions. Constraints for bond lengths were applied using the LINCS algorithm with a 2 femtosecond time-step in each simulation. Before the production run, all systems were neutralised with sodium and chloride ions to mimic physiological conditions. Conjugate gradient minimisation with a maximum force of 100 kJ/mol/nm was performed to remove unfavourable interactions. A two-stage equilibration protocol was performed to prepare for the production run for which position restraints were applied to all atoms. In the first stage, the system was simulated under a constant volume (NVT) ensemble to achieve 300 K by the V-rescale method for 200 picoseconds. The equilibrated output from the NVT ensemble was treated to constant pressure (NPT) equilibration using the Parinello-Rahman barostat under an isotropic pressure of 1 bar. Restraints were released post-equilibration for production runs for which the V-rescale thermostat and Parinello-Rahman barostat were applied to maintain the temperature and pressure at 300K and 1 bar, respectively.

Analysis of MD

Trajectories obtained through MD simulations were analysed for various measures of system equilibration and structural dynamics. The root mean square displacement (RMSD) measure of the Cα atoms, root mean square fluctuation (RMSF) of the residue atoms, and radius of gyration (Rg) were calculated using the gmx RMS, gmx rmsf and gmx gyr modules of the GROMACS package. The total energy of the system was calculated using gmx energy. The inter-monomeric distance of the heterodimer model was calculated using gmx mindist. The distance between protein residues and ligand molecules, termini atoms and hydrogen bond distances were calculated using gmx dist module. Secondary structure (data not shown) was determined using the DSSP v2.2.1 (Kabsch & Sander, 1983) algorithm through the mkdssp module. Snapshots of the trajectory were extracted

every 1 ns using gmx trjconv for interface analysis. Trajectories were visualised and analysed using VMD software. (Humphrey et al.,1996).

Essential dynamics analysis

Essential dynamics analysis (EDA) is a valuable unsupervised technique that statistically determines patterns in high-dimensional data such as Cartesian coordinates containing trajectory snapshots of MD simulation data (Amadei, Linssen et al., 1993) or experimental structures (vanAalten, Conn et al., 1997). EDA is based on Principal Component Analysis (PCA) (Hotelling, 1933, Jolliffe, 2002, Pearson, 1901). Trajectory replicates (n=3) were pooled as input coordinates to analyse principal components in the conformational subspace. PCA analysis was carried using the Bio3D package (Grant, Rodrigues et al., 2006, Skjærven, Yao et al., 2014) in the RStudio (RStudio Team, 2020) environment using the R (R Core Team, 2013) programming language. The pooled trajectory file was converted to dcd format and analysed using the covariance matrix. The first two principal components obtained were further analysed for the number of optimal clusters to partition the data. Naive Bayes fviz function from factoextra package in R was used to determine silhouette coefficient (Rousseeuw, 1987) and the sum of squared error values (data not shown) against the number of clusters, respectively.

Conformations captured by the first two principal components were extracted as movies. A dynamic cross-correlation map (DCCM) was calculated for each residue in the heterodimer model. Further details of the methodology will be found in the Supplementary Information file.

Analysis of protein-protein interactions at the dimer interface

The PPCheck webserver (Sukhwal & Sowdhamini, 2015) was used to find the best native-like docking pose among several poses generated by the docking program used (Haddock). Briefly, the algorithm employed by PPCheck calculates the total stabilising energy and normalised energy per residue for every interacting pair of protein chains. The total stabilising energy value is calculated by summating the hydrogen bond energy, van der Waals' energy and electrostatic energy for each interacting pair of chains.

PPCheck was used to analyse the interface energies of the AgamOBP1 homodimer (PDB code: 2ERB) and the acquired heterodimeric models. Additionally, the snapshots extracted from the MD trajectory, at every 1 ns, were analysed for their interface interactions by the same algorithm. Plots

showing the variations in the energetics and number of interface residues calculated by PPCheck were generated using the igraph tool provided with the R software using custom in-house scripts.

4.3 Results and Discussion

In the case of the AgamOBP4 monomer, the apo- and indole-bound systems equilibrate well within 200 ns of the simulation with models at pH 4.5 and pH 6.5 equilibrating by 100 ns of the MD simulation (Figure 4.1 A). Overall, there is a gradual decrease in RMSF as pH increases from 4.5 to 8.0 (Figure 4.1 B), suggesting that flexibility of AgamOBP4 reduces with increasing pH.

Transient secondary structure elements such as 310-helices and pi-helices in inter-helical loops and alpha-helical regions are observed and prolonged with an increase in pH from acidic to basic. Coils, bends, turns were also observed in inter-helical regions. Beta-sheets and beta-bridges, observed in the loop connecting helices α3 and α4, are absent in apo-AgamOBP4 at acidic pH and sparse in indole-bound AgamOBP4 at neutral pH. At acidic pH, helix α4 gets unstructured during the simulation, but it gets stabilised by indole. At neutral pH, however, α4 helix remains stable with and without indole-binding.

Moreover, residue fluctuations are relatively higher for helices α4, α5 and the α4-α5 connecting loop in apo- and indole-bound AgamOBP4 at acidic pH concerning those at other pH values. It agrees with observations of the crystallographic B-factors for the α4-loop-α5 region that are clearly higher at acidic pH (**Figure 4.1 D**). The loop between helices α3 and α4 fluctuates highest for apo-AgamOBP4 models at neutral and alkaline pH. Thus, taking together crystal structure and *in silico* analyses, the significant differences between AgamOBP4 structures at acidic and neutral pH are in the α4 & α5 helices, α4-α5 connecting loop as well as in the α3-α4 loop.

The C-terminus fluctuates least for indole-bound AgamOBP4 at neutral pH. At neutral pH (**Figure 4.1 (B, D)**), alpha-helical regions show correlated motions with one another, and the effect is more pronounced in the presence of indole (**Figure 4.1 D**). The N-termini of apo OBP4 at acidic pH show high fluctuations suggesting a likely disordered state.

Summarising, our MD calculations showed that the α4-loop-α5 region is more flexible at acidic pH in the absence or the presence of indole. Both, the presence of indole and increase of pH influence stabilisation of α4 helix. In addition, both, pH incensement and indole result in concerted, rather than random motions of the α-helices. These results agree with the crystallographic and

thermodynamic results in this work and further support the combined stabilisation effect of increased pH and binding of indole.

Generating structural models of apo and indole-bound AgamOBP1-AgamOBP4 heterodimers formed under different pH conditions

To investigate the combined influence of pH and indole on the AgamOBP1-AgamOBP4 heterodimer, we attempted to generate structural models that likely mimic the heterodimer complexes possibly formed under neutral, acidic and alkaline conditions of pH, with and without indole bound to AgamOBP4. We devised a protocol to model the heterodimeric structures assumed to form under different pH conditions, using a heterodimer model formed under neutral pH as a template.

Previous NMR studies have suggested several AgamOBP1 and AgamOBP4 residues present at the heterodimeric interface **(Table 4.1)** and those that potentially contribute towards indole binding AgamOBP4 (Davrazou et al., 2011). In the second step, these residues were used to guide the docking between the neutral pH-adjusted AgamOBP1 and AgamOBP4 monomers. The final model was chosen based on an intensive evaluation of the lowest-ranked docked models (for details, please refer to Materials and Methods section) and could be considered the representative model of the heterodimeric complex at neutral pH. The rationale behind this choice was based on the suggestion in Davrazou et al, 2011 (Davrazou et al., 2011) that the binding of indole primarily aids the formation of the heterodimer model at neutral pH to it. Since we used the indole-bound form of AgamOBP4 to construct the heterodimer model, it considers subtle conformational changes in AgamOBP4 associated with indole-binding. Therefore, the resulting heterodimer model will be more relevant to the actual indole-induced heterodimer complex formed under physiological conditions. The model is illustrated in **Figure 4.3**.

In the third step, in order to assess the effect of pH change on the interactions across the heterodimer interface, we adjusted the ionisation states of the titrable amino acids of the heterodimer model at neutral pH (pH 6.5), to acidic (pH 4.5) and alkaline pH (8.5). This process was followed by energy minimisation and led to the generation of two heterodimeric models representing the hypothetical AgamOBP1-AgamOBP4 complexes formed under acidic and alkaline conditions. The models generated do not contain indole in the AgamOBP4 subunit and are referred to as the apo forms of the heterodimer.

To create the indole-bound forms of the heterodimer at the three pH values of 4.5, 6.5 and 8.5, the indole molecule was added to the binding cavity of the AgamOBP4 subunit using the indole-bound AgamOBP4 crystal structure (Davrazou et al., 2011) as a reference. Thus, six heterodimeric models were generated in this study - three apo and three indole-bound forms.

Evaluating the dynamics of the AgamOBP1-AgamOBP4 heterodimer

To investigate differences not apparent in the static structures of the heterodimer models, long-scale MD simulations of 200 ns were carried out on each of the models. Stability and interactions present in the heterodimer interfaces across all the six trajectories were evaluated. The RMSD plots of Cα atoms of the models reveal that each system attains its equilibrium state at different time points and remains stable for the rest of the simulation time at less than 1 nm. The radius of gyration (Rg) of all models lies between 1.9 to 2.1 nm with a slight upward trend in the case of indole-bound heterodimer at neutral pH and the apo-model at acidic pH suggesting changes in the models structure from 100 ns onwards, thereby resulting in a general decrease in compactness. In contrast, Rg for the apo-model at pH 6.5 does not fluctuate much throughout 200 ns. The mean total energy of each system lies within the range of -800000 kJ/mol to -805000 kJ/mol. The RMSF of atoms for AgamOBP1 (Chain A) and AgamOBP4 (Chain B) monomers of each heterodimeric model has been depicted in**,** respectively. The indole-bound model at neutral pH exhibits the highest fluctuation throughout in both chains A and B. Interestingly, the residues of AgamOBP1 show changes in the intensity of the fluctuations during the change of pH. . In the AgamOBP4 chain, the residues that show high fluctuations vary with pH suggesting that the pH change contributes to residue-wise fluctuations. These residues belong to different helices of AgamOBP4. The indole-bound model at neutral pH and the apo-model at acidic pH maintain their higher fluctuations (>0.5 nm) in comparison to the apo-model at neutral pH and the indole-bound model at acidic pH, thus indicating that both indole binding and pH value may play a significant role in the dynamic changes of the heterodimer. Following the verification of the validity of the system subjected to MD simulations, we checked whether the generated models remained intact throughout the simulations. The minimum distance between the chains of AgamOBP1 and AgamOBP4 in each heterodimer model was found to be below 0.3 nm, implying that each dimer remains intact throughout the 200 ns simulation. In addition, end-to-end distances of the heterodimer models were measured during the simulation course between the N-terminals of

AgamOBP1 and AgamOBP4 chains and the C-terminals of those monomers. The C-terminal distances in all models remain relatively conserved within 3 to 4 nm compared to the N-terminal distances, which increase from 100 ns onwards for the apo-heterodimer at acidic pH and the indole-bound model at neutral pH. Analysis of the energetics by PPCheck (**Materials & Methods**) from the trajectory shows that the normalised stabilising energy per residue at the interface varies between -1.7 and -4.25 kJ/mol (**Figure 4.4**). The interface energies were analysed from 80 ns onwards based on the RMSD analysis of trajectories. Time-series evolution of unfavourable electrostatic interactions is depicted at the interfaces of heterodimer models with respect to pH and ligand-binding. For each case, unfavourable interactions have been reported. It is observed that at 100 ns, there is an overlap in the interface strength (-2.3 kJ/mol) of apo- models at acidic and neutral pH. To learn more about the individual contributions of pH and indole, as well as their combined effect on the heterodimer stability and dynamics, we carried out a comparative analysis of the interaction energies (to assess the differences instability) and essential dynamics followed by the identification of cross-correlated motions in order to determine the regions of the heterodimer contributing to the majority of its dynamics under different conditions. Unfavourable electrostatic interactions between C-beta atoms were observed between various residue pairs, including Glu17-Glu88, Glu14-Glu75, Asp24-Asp95 and His77-Lys61 in apo heterodimer at acidic pH. In indole-bound heterodimer at acidic pH, Glu75 exhibits unfavourable interactions with Glu14 and Glu17 throughout the course of the simulation. In apo-heterodimer at neutral pH, residue pairs Glu14-Glu75, Glu17-Glu75, Lys63-Lys69, Lys75-Lys61 and His77-Lys61 maintain unfavourable electrostatic interactions throughout, whereas on indole binding, Asp24-Glu88 and Lys29-Lys94 are observed.

i. Investigating the influence of pH alone on heterodimer stability and dynamics

Comparison of the interactions at the heterodimeric interface, calculated through PPCheck, between the apo models at acidic (pH 4.5), neutral (pH 6.5) and alkaline pH (pH 8.5) at the end of the 200 ns simulations shows a trend of an increase in the number of intermolecular hydrogen bonds at the interface along with the increase in pH (2 at pH 4.5, 4 at pH 6.5 and 6 at pH 8.5). However, the number of electrostatic and hydrophobic interactions remains roughly similar between the acidic and neutral forms even at 200 ns, there are considerably more electrostatic interactions observed in the alkaline form with interface strength of -2.1 kJ/mol (**Figure 4.4 A**). In addition, in order to assess the interface strength, we carried out an essential dynamics analysis

(EDA) of trajectories obtained through MD simulations which uncovers linear relationships between atomic motions and protein conformation(s) by reducing the dimensionality of the pooled trajectory dataset and maximising variance using uncorrelated variables. Hierarchical clustering, an unsupervised learning technique, was further used to partition the transformed data into optimal clusters. In apo and indole-bound heterodimers, with increasing pH (**Figure 4.5**), inter-cluster separation increases, indicating distinct conformational states. Binding of indole indicates increased conformational homogeneity with an increase in pH, as observed by increasing cluster separation and size. Dynamic cross-correlation analysis of the sampled Cα atoms revealed a difference in correlated motions of AgamOBP1 and AgamOBP4 chains of the heterodimer at acidic and neutral pH (**Figure 4.6)** . A correlation score of 1 indicates maximum correlation between Cα atoms of two residues, whereas -1 indicates maximum anti-correlation. In case of AgamOBP1 in apo-heterodimer model at pH 4.5, inter-monomeric motions are highly concerted in directionality, but they are anti-correlated in the apo-heterodimer at pH 8.5.

ii. ***Influence of indole on heterodimer stability and dynamics***

To assess the influence of indole on the stability of the heterodimer under varying pH conditions, we compared the stability and intermolecular interactions between the apo models and their corresponding indole-bound forms at the three pH values, as described below.

a. At acidic pH (pH 4.5)

The interface strength at the end of the 200 ns simulation of the apo and indole-bound models does not vary much (-2.1 kJ/mol in apo versus -2.4 kJ/mol in indole-bound form). Despite the absence of a change in the interface strength, an analysis of their secondary structures (data not shown) revealed pH-specific conformational differences during the MD simulations. The secondary structure of the model at acidic pH remains helical throughout with the presence of bends, coils, and turns (150-200 ns). However, the indole-bound heterodimer at acidic pH remains alpha-helical for helices α1-α3 and α6 while turns and coils are observed in helices α4 and α5 of AgamOBP4 monomer. These α4 and α5 helices are present at the interface of the AgamOBP1 and AgamOBP4 complex. Subtle differences are also noted through a comparative essential dynamics analysis between the two forms. In the case of the apo-heterodimer **(Figure 4.5 A; Movies M1, N1, N2),** the first two principal components explain 83.78% of the motions in the trajectory data. Optimal cluster size of two with an average silhouette width of 0.7 indicated two major

conformations representing the trajectories. Unlike in its corresponding apo form, the variance captured through the data is only 58.7% for the indole-bound heterodimer. An optimal cluster size of four with an average silhouette width of 0.65 for the two models indicated four major conformations representing its trajectory. A dynamic cross-correlation analysis of the apo-heterodimers (**Figure 4.5 A-C**) shows that motions of helix α2 and the loop connecting helices α2 and α3 of AgamOBP1 are highly correlated with helices α2, α3, α4 +and α5 of AgamOBP4. Strong anti-correlation motions are present in AgamOBP4 among non-identical residues except the correlated motions between the helices α2, α3, the loop between α1 and α2 with the segment spanned by helix α6, the C-terminus tail and the α5-α6 connecting loop.

iii. Combined influence of pH and indole on heterodimer stability and dynamics

The interface energy of both apo and indole-bound heterodimer models extracted at the end of the 200 ns simulations is comparatively higher at pH 8.5 than in acidic and neutral pH. As described in the preceding sections, the interface energy does not vary considerably between the apo and indole-bound forms of the heterodimer model at both acidic and neutral pH. This lesser variance (of motions in the space) captured for the indole-bound heterodimeric model at pH 4.5 compared with others models- the indole-bound heterodimer at pH 6.5 and their apo-heterodimeric models suggests that the dynamics represented by its first two principal components cannot be adequately linearly represented. The intrinsic assumption that each frame of the simulation trajectory sampled represents a protein conformation from a system at equilibrium does not hold in the case of the ligand-bound heterodimer at pH 4.5. The observation further supports this that the RMSD of the system alone did not converge within 200 ns. Monomeric AgamOBP4 shows reduced flexibility with an increase in environmental pH. Our results suggest that the extent of helicity increases with pH and that helicity may be important for binding to indole. Taken together with correlated motions of residues in AgamOBP4 at neutral pH, indole likely favourably binds to AgamOBP4 at neutral pH. Interestingly, ordered apo-AgamOBP4 at neutral pH shows a high degree of correlated motions among its helices and termini. Indole-binding was further observed to enhance these correlated motions. Conversely, the heterodimeric residue-specific flexibility, the distance measures as well as the correlated inter-monomeric interactions suggest that indole-bound heterodimer at neutral pH is a favourable model, in which helices α1, α3, α4, and α5 of OBP4 seem to play an important role through correlated motions. Beta-bridges and beta-sheets, which were observed in the monomeric AgamOBP4 are absent in the heterodimer, indicating that the

heterodimerisation further strengthens the conformational stability of the helices. The presence of transient 310 helices suggests a supporting intermediate alpha-helix formation or helix to coil transition (Armen, Alonso et al., 2003) during the simulation, depending on the local pH. Frequent and transient pi-helices indicate a high likelihood of a proximal functional site (Weaver, 2000) in the OBP, which could serve as an evolutionary adaptation in ligand binding and function.

4.4 Conclusion

Current understanding of the model of OBP-ligand binding

It has been previously proposed that OBPs may interact with other OBPs in the lymph to form dimers with novel binding properties that further participate in recognising odorants (Andronopoulou et al., 2006; Qiao et al., 2011). Moreover, pH-induced conformational changes can be related to the odorant release mechanism, as shown for some OBPs (Leal, Chen et al., 2005a, Leal, Chen et al., 2005b). Several theoretical models and *in vitro* studies have been carried out to propose a pH-dependent mechanism of OBP-OBP and OBP-ligand interactions (Kaissling, 2001, Pelosi et al., 2006).

Herein, we compared the crystal structures of the apo-forms of AgamOBP4 at acidic (4.6) and alkaline pH (8.5) with its indole bound form at neutral pH (6.85). We found that the apo- structure of the protein at alkaline pH resembles the indole-bound form. Also, we observed a remarkable destabilisation of the N-terminus at acidic pH, which was confirmed by MD calculations. Based on these observations, we performed thermodynamic studies, which suggested that the shift from acidic to neutral and alkaline pH results in a significant stabilisation of AgamOBP4 protein structure. In contrast, the stabilising effect of indole is modest.

In order to examine the possible effect of pH on the formation of AgamOBP1-AgamOBP4 heterodimer, we generated many AgamOBP1-AgamOBP4 heterodimeric models at varying pH values (4.5, 6.5 and 8.5) using guided docking to represent the structural variations in the local pH across the sensillum- from the physiological pH of the sensillar lymph to the localised acidic environment near the ORs (Damberger, Michel et al., 2013, Keil, 1984, Wojtasek & Leal, 1999). Our results reinforce the idea that the binding of indole to AgamOBP4 plays a vital role in promoting the formation of a functional heterodimeric interface. AgamOBP4 may exist in a molten globule state at neutral pH (Davrazou et al., 2011). Our MD results suggest that AgamOBP4 is

stabilised by indole prior to its heterodimerisation with AgamOBP1, which is also monomerised from its dimeric state. At neutral pH, the presence of indole appears to enhance correlated motions in the AgamOBP4 monomer but may be causal towards heterodimeric stability. These structural movements may lead to forming a binding site for AgamOBP1 monomer, thus promoting the generation of the heterodimer as proposed by Davrazou et al. (2011). The indole molecule, bearing an aromatic ring is known to get protonated at acidic pH. The lower predicted stability of the heterodimeric complex near the negatively-charged membrane of the sensory neuron suggests the dissociation of the AgamOBP1-AgamOBP4 heterodimer in this acidic environment and the subsequent release of the protonated ligand near the ORs.

As indicated by the structural data, the binding cavity of AgamOBP4 possesses two openings to the solvent. Importantly, our crystallographic studies showed that as pH value decreases from 8.5 to 4.6, the central opening becomes significantly wider. In addition, the comparison of the interface residues and interface energies of the generated models combined with analysis of their structures and dynamics across varying pH, demonstrates that the indole-bound AgamOBP1-AgamOBP4 heterodimer appears to be stable between pH 6.5-8.5. All the above findings suggest an additional dependence of the heterodimerisation process on pH, directly correlated with the stabilising mechanism of monomeric AgamOBP4.

It is possible that the observed flexibility of the N-terminus of AgamOBP4 at acidic pH as well as the structural changes that may take place in the regions of helices α1, α3, α4, and α5, and the loop connecting α3 and α4 helices during the transition from neutral/alkaline to acidic pH, promote the disruption of critical interactions at the heterodimer interface of AgamOBP1-AgamOBP4 dimer. Consequently, the decrease of pH may lead to the dissociation of indole from the complex, allowing its release and the subsequent activation of the receptor.

Based on our findings, we propose an AgamOBP1-AgamOBP4 heterodimer formation/dissociation model depended on the pH for the transportation and release of indole (**Figure 4.7**) through the aqueous lymph of sensilla.

This model is consistent with the proposed mechanism of pH-induced odour-release observed in mosquito OBPs, according to which the binding of the odorant alone can activate the OR. Furthermore, Davrazou et al. (Davrazou et al., 2011) suggested that indole, besides the induced conformational changes to AgamOBP4 to form a binding site for AgamOBP1 may also cause the

dissociation of AgamOBP1 dimers to allow monomer to bind to AgamOBP4. It further supports the hypothesis that indole may promote the heterodimerisation of the two proteins by *shifting* the balance towards *forming a more stable complex at neutral pH, which then destabilises in the* acidic environment near the ORs. *This proposed mechanism of ligand- and pH-induced dimer formation and pH-induced ligand release might be adopted among insect OBP-heterodimers in general.*

Tables and Figures

Table 4.1: Residues selected to guide docking of AgamOBP1-AgamOBP4 heterodimer. Residues associated with chemical shifts in NMR spectra of AgamOBP1 and AgamOBP4 in the presence of indole (Davrazou et al., 2011). For AgamOBP1, the chemical shifts were measured in the presence of indole and AgamOBP4, whereas for AgamOBP4 they were obtained in the presence of indole and AgamOBP1. The chemical shift signals vary in intensity and residues have been sorted accordingly (greater than mean, greater than mean plus 1 standard deviation, greater than mean plus 1.5 standard deviation). Residues of AgamOBP4 that interact with indole are indicated in bold fonts while those displaying chemical shifts but are distant from the indole binding site are underlined.

Chain	**Residues with chemical shift changes specific to the presence of indole (Davrazou et al., 2011)**
A	**>mean + 1.5 S.D**- E14 (a1), **A18 (a1)**, **K20(a1)**, H23(a1), **F59(a3),** H60(a3), C95 (a5-a6 loop), **Y97 (a5-a6 loop)**, L110 (a6), Y122 (C-tail) **> mean + 1 S.D**- D42 (a2), V64(a3-a4 loop)**,** L96 (a5-a6 loop), K112 (a6)

(AgamOBP1)	**> mean-** D7 (a1), L19 (a1), L58 (a3), **K63 (a3-a4 loop)**, D66 (a3-a4 loop), H72 (a3-a4 loop), H85 (a5), D86 (a5), C113 (a6), W114 (a6), D118 (C-tail)
B **(AgamOBP4)**	**>mean + 1.5 S.D**- M55 (a3-a4 loop), A56 (a3-a4 loop), G57 (a3-a4 loop), T60 (a3-a4 loop), Q73 (a4), I74 (a4), M77 (a4-a5 loop). **> mean + 1 S.D**- S10 (a1), M11 (a1), D12 (a1), **T58 (a3-a4 loop),** M59(a3-a4), L78 (a4-a5 loop), F122 (a6) **> mean-** Q6 (a1), M14 (a1), Q16 (a1), C18 (a1-a2 loop), K46 (a3), Y48 (a3), A49 (a3), M50 (a3), **A53** (a3), E64 (a3-a4 loop), I65 (a3-a4 loop), K69 (a4), **T70 (a4)**, M71 (a4), E75 (a4), K87 (a5), L90 (a5), T91(a5), T98 (a5), A111 (a6), K112 (a6), A115 (a6)

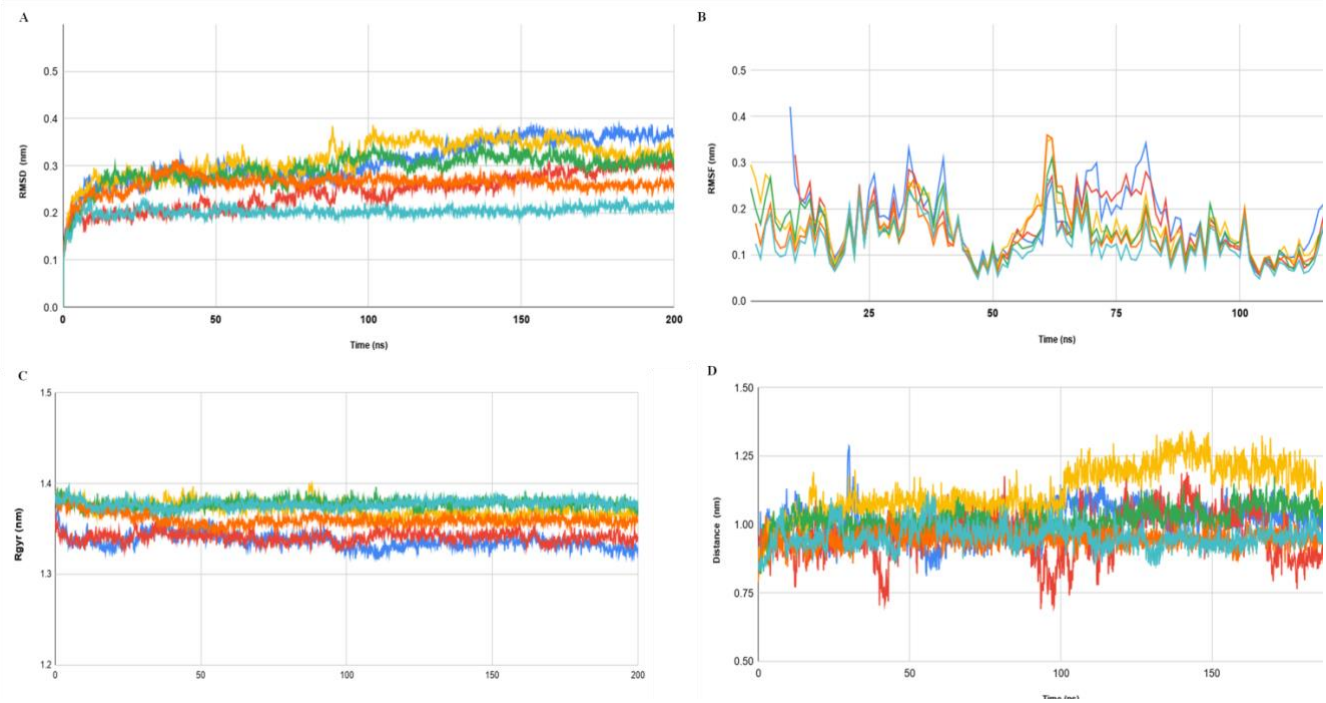


Figure 4.1: *Analysis of trajectories obtained through MD simulations of AgamOBP4 (A) Root mean square deviation (RMSD) of Cα atoms (B) Root mean squared fluctuations (RMSF) of protein atoms in monomeric AgamOBP4 (C) radius of gyration (Rg) (D) Distance from N-terminus to C-terminus in monomeric AgamOBP4. Models are represented by dark blue (apo, pH 4.5), yellow (apo, pH 6.5), orange (apo, pH 8.5), red (indole-bound, pH 4.5), green (indole-bound, pH 6.5) and cyan (indole-bound, pH 8.5).*

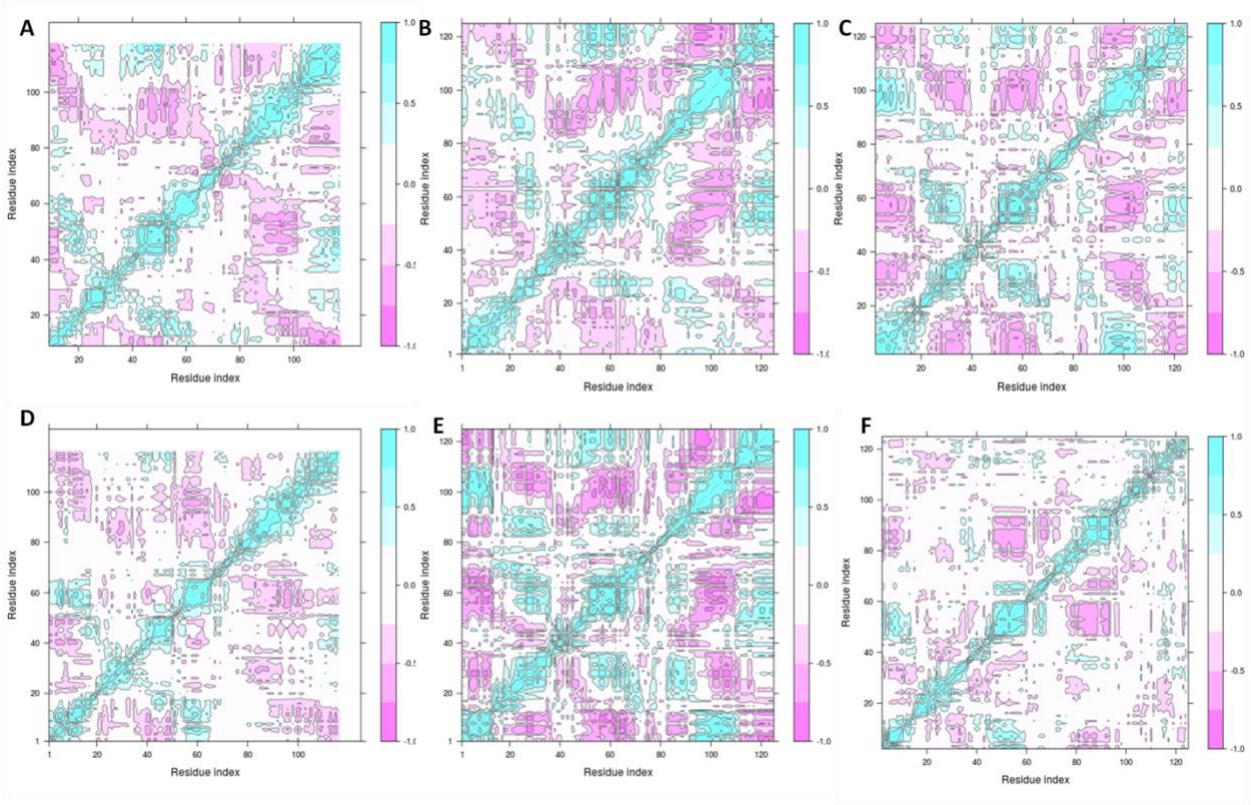

A
B
C
D
E
F
Residue index
Residue index

Figure 4.2: Dynamic cross-correlation matrix of AgamOBP4 monomer in (A) apo form at pH 4.5, (B) apo form at pH 6.5, (C) apo form at pH 8.5, (D) indole-bound at pH 4.5, (E) indole-bound at pH 6.5 and (F) indole-bound at pH 8.5. Bar relating degree of cross-correlation score is shown on the side of each plot. Each plot represents pooled replicate trajectories (n=3).

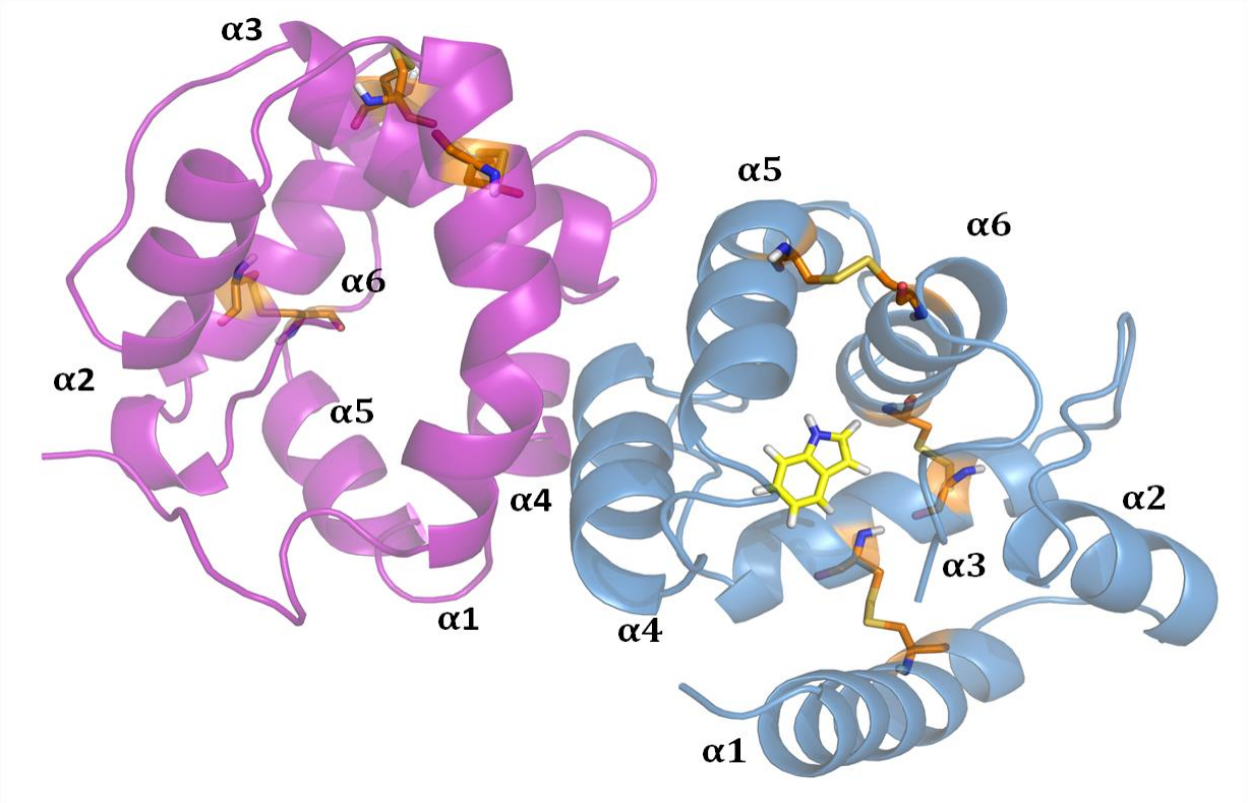


Figure 4.3: The AgamOBP1-AgamOBP4 model at pH 6.5 was constructed using the guided docking method. Chains A and B of the docked model are represented by magenta (AgamOBP1) and blue (AgamOBP4), respectively. Disulfide bonds are in yellow. Indole (yellow sticks) has been added to the model to indicate the location of the indole binding-site of AgamOBP4

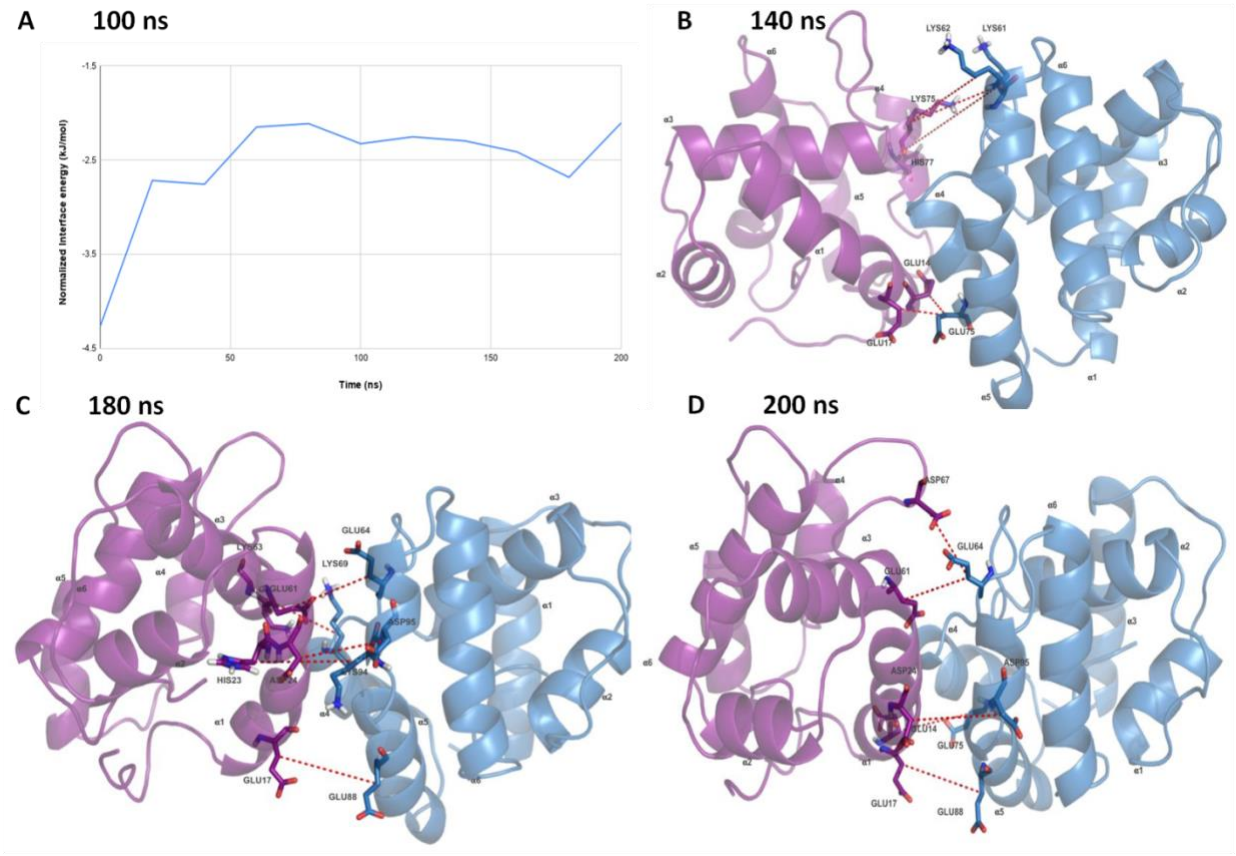


Figure 4.4: Heterodimer models at pH 4.5 at timesteps A. 100 ns B. 140 ns C. 180 ns and D. 200ns. The unfavourable interactions at the interface are depicted as red dots and the residues involved are shown as sticks. . Time-based evolution of PPCheck interface strength in the AgamOBP1-AgamOBP4 models at pH of 4.5 and 6.5.

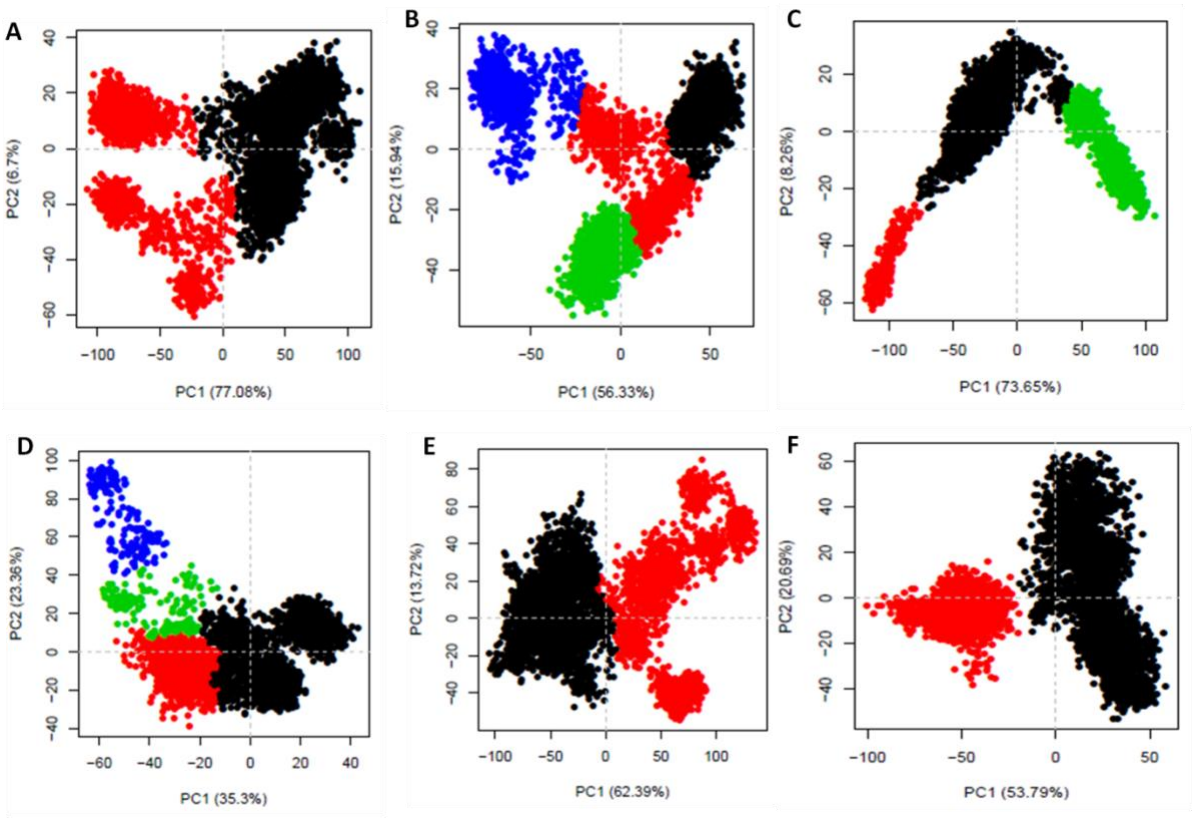


Figure 4.5: First two principal components from essential dynamics analysis of heterodimer in (A) apo model at pH 4.5, (B) apo model at pH 6.5, (C) apo model at pH 8.5, (D) indole-bound at pH 4.5, (E) indole-bound at pH 6.5 and (F) indole-bound at pH 8.5. Each plot represents pooled replicate trajectories (n=3) of 200 ns each.

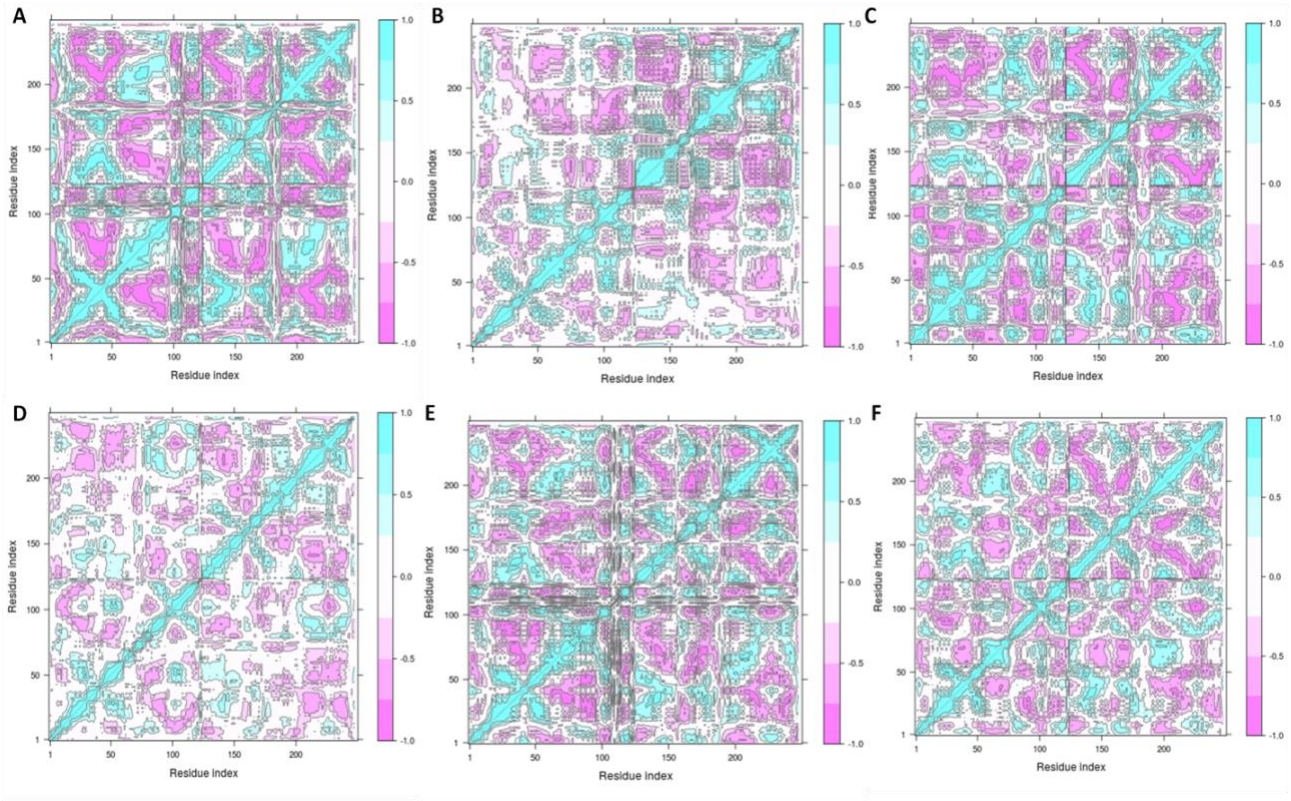


Figure 4.6: Dynamic cross-correlation matrix of heterodimer in (A) apo model at pH 4.5, (B) apo model at pH 6.5, (C) apo model at pH 8.5, (D) indole-bound at pH 4.5, (E) indole-bound at pH 6.5 and (F) indole-bound at pH 8.5. Bar relating degree of cross-correlation (color intensity) and cross-correlation score is shown on the side of each plot. Each heterodimer has a 247 Cα-residue index 1 to 125 (midpoint) that corresponds to AgamOBP1 (Chain A)

and a residue index 126 to 247 for AgamOBP4 (Chain B) of the heterodimer model. Each plot represents pooled replicate trajectories (n=3).

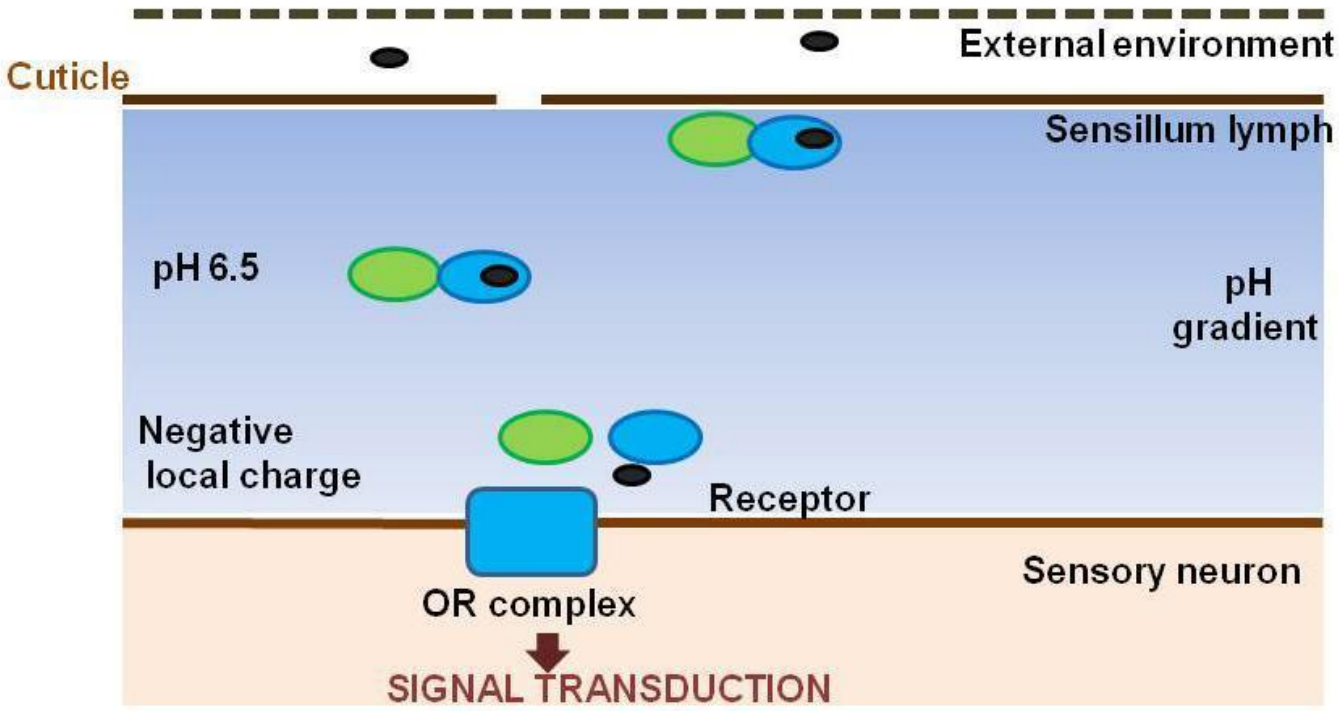


Figure 4.7: Proposed model of formation of AgamOBP1-AgamOBP4 heterodimer and subsequent transportation of odor molecules. According to the proposed model, the heterodimeric formation tendency is predicted to be higher between pH 6.5 and pH 8.5, as indicated by the energetics of the heterodimer interface. At low pH, this tendency reduces, thereby promoting the release of the odorant molecules near the odorant receptor

Chapter 5

Virtual Screening in OBPs to identify natural repellants and metabolite partners

5.1 Introduction

It has been hypothesized that each insect pheromone component could bind to a unique pheromone-binding protein (PBP). A previous study (Du and Prestwich, 1995) showed that two pheromone binding proteins in the moth species, *Antheraea pernyi*, have differing affinities for distinct pheromone components. It was strongly indicative of odorant specificity encoded into an odorant-binding protein.

For insect OBPs, OBP-ligand binding may not always occur with high binding affinity and high specificity together, i.e., affinity and specificity need not follow one another (Murphy et al, 2013; Venthur et al, 2018). For binding with high affinity and good shape complementarity, optimal interactions may be attained by optimizing electrostatic and hydrophobic interactions between the drug candidate and the target. High affinity can be achieved only with requisite shape and interaction complementarity over the few interaction sites available. Specificity helps discriminate functionally between two molecules that compete for the same ligand. Differences in the free energy of interaction for the molecules of interest- thermodynamically as the coupled equilibria between entities in solution (von Hippel and Berg, 1986; Eaton, Gold and Zichi, 1995).

Monomers OBP1 and OBP4 require indole for heterodimerization (**Chapter 4; Davarazou et al, 2011**).

The compound diethyltoluamide (DEET) is a well-studied mosquito repellent that is also toxic to the human nervous system (Ditzen et al, 2008). However, sulcatone (6-methyl-5-heptene-2-one; 6-MH) is a known insect repellent that is a naturally occurring volatile in over 400 plant organisms (Pherobase, El-Sayed, 2018) and released by essential oils and fruits (Anet, 1972). A previous study (Murphy et al, 2013) showed that repellents 6-MH and DEET separately bind at the same site to AgamOBP1 to potentially **disrupt OBP1-OBP4** interactions.

Ligand 6-MH induces conformational ordering of **OBP4** but not heterodimerization. DEET stabilizes **OBP1 homodimer**. OBP1 is monomeric in the presence of 6-MH and DEET. Given that 6-MH and DEET bind in an opening between helix 4 and helix 5 of AgamOBP1, it is unknown if ligands 6-MH and DEET block heterodimeric interactions through different mechanisms. Helices 1, 3 and 4 of OBP4 provided evidence of interaction site, but in NMR studies of residues in helices 4 and 5 of OBP1 couldn’t be observed in the OBP1-indole complex.

5.2 Specific questions

a) Can a naturally occurring volatile from a known insect (or mosquito) repellent plant organism be used to preferentially bind to OBP heterodimer and monomers over indole, a known attractant?

b) **Based on additional literature survey, I have identified other unexplored questions that I explored through work done in this project :**

c) Does 6-MH and DEET binding disrupt interaction with OBP4 by preventing helix4 and helix5 to interact at the interface of OBP1 and OBP4?

d) Does 6-MH and DEET that is bound to OBP1 induce conformational changes in OBP4 to prevent heterodimerization?

e) Can OBP4 form a complex with 6-MH?

f) It is known that 6-MH competes for OBP1-binding with indole, i.e., (OBP1-indole) or (OBP1-6-MH). However, 6-MH and DEET do not compete with 1-NPN but bind together, i.e., 6-MH (or DEET) -- OBP1-- 1-NPN. Can this be also be observed *in silico*?

g) Can multiple ligands bind to a few OBPs simultaneously? Could a combination of different ligands be bound to a single OBP to cause conformational reordering? Could certain OBP-ligand complexes be required to form heterodimers to elicit a response?

h) Recently, unconventional roles of OBP have been uncovered with potential roles in immunity, hematopoiesis and detection of humidity outside olfaction (**Chapter 1; 1.4.1**). Can metabolites found in ***Anopheles gambiae*** and other mosquito species bind to OBP?

5.3 Materials and Methods

Ligand curation and ligand library preparation

In order to identify suitable ligands to bind to monomeric and dimeric models of mosquito OBP, two ligand libraries were constructed through mining existing literature. The first ligand library consisted of active compounds of various plant organisms (data not shown). For this purpose, plant organisms that were known to have mosquito repellent or insect repellent properties through literature were mined for their active components with volatile properties. The second ligand library consisted of non-peptide metabolites across mosquitoes and other insects (data not shown).

This approach yielded **112 volatiles** and **433 non-peptide metabolites** (data not shown). Possible conformers were prepared for each ligand library at neutral pH (pH 6.5) using the LigPrep (Schrödinger Release 2019-1: LLC, 2019) module in Maestro. The ligands were desalted and tautomers were generated, and chiralities were determined from the 3D structure.

Protein structure models

AgamOBP1 homodimer at pH 8.5 was obtained from PDB (PDB ID: 2ERB).

AgamOBP4 structures at pH 4.6, 6.5 and pH 8 -experimental collaborator, Prof. Spyros Zographos' group at NHRF, Greece (**Chapter 4**)

Heterodimer models were constructed in-house as explained in **Chapter 4**.

The crystal structure data for OBP1 complexed with sulcatone (6-MH) was obtained from PDB (PDB ID: 4FQT).

Protein preparation

Each of the models was prepared using the Protein Preparation Wizard (Sastry et al., 2013) in MAESTRO version 11.9.011, Schrodinger Inc. (Schrödinger LLC, 2018 & 2019) on the Linux platform. Each model was corrected for bond orders, bond types, missing atoms and charge at pH 6.5. Waters beyond 5 Å were deleted from heteroatom groups. Water molecules with less than three hydrogen bonds to non-water molecules were removed. Protonation and orientation states were corrected accordingly, and the hydrogen bond network was optimized to aid protein preparation following which restrained minimization was carried out to allow hydrogen atoms to be minimized using OPLS3e (Jorgensen & Tirado-Rives, 1988; Jorgensen, Maxwell & Tirado-Rives, 1996) force field and to relax strained bonds, angles and clashes.

Site Analysis

The indole-binding sites for OBP4 and the heterodimer OBP1-OBP4 models were obtained from work done in **Chapter 4**; **4.2**. Additionally, for these models, OBP1 monomer and OBP1 homodimer models, minimization was performed on each of them to screen for potential ligand-binding and druggable sites. These sites were identified using a restrictive definition of hydrophobicity and a standard grid with crop site maps at 4 Å cut off from site point in Sitemap

(Halgren, 2007 & 2009; Schrödinger LLC, 2018). This was done to identify top-ranked potential receptor-binding sites. The ligand-binding score (Site Score) and druggability (D-score) were used for filtering predicted binding sites with a cut-off of 0.8 each. A score greater than or equal to 1 indicates a strong probability of being a potential binding site whereas a score between 0.8 and 1 indicates a partial binding site. The site corresponding to a **potential binding site** was chosen further for preparation of receptor grid and docking against ligand. A receptor grid was generated corresponding to each potential binding site using Van der Waals radius scaling factor of 1 and a partial charge cut-off of 0.25. The docked ligand was confined to the enclosing box and the centroid was taken into account. The force field used was OPLS3e.

Virtual Screening

Virtual screening (VS) was performed by screening the ligand library against the receptor grid generated around a potential binding site. In the procedure of VS, steps of docking were performed as high throughput (HTVS), Standard Precision (SP), and Extra Precision (XP). At each step, 10% of the ligands were filtered as input for the successive step. Finally, ligands were shortlisted on the basis of XP docking score and protein-ligand interactions. These were further clustered hierarchically in CANVAS using binary fingerprinting and flexible beta-cluster linkage method. Cluster centroids or a cluster member with the most favourable interactions with the receptor were chosen as cluster representatives. These ligands constituted the final set used to determine binding free energy using MM/GBSA in Maestro Prime (Schrodinger Inc.). Molecular Dynamic (MD) simulation was performed for each of the docked complex poses that consisted of ligands present in our newly reduced set after cluster analysis. In order to perform MD, the system was built using OPLS3e force field and an orthorhombic geometry of 10 Å on each side with TIP4P solvation model and buffered with a salt concentration of 0.15 M. The production run was performed for a simulation time of 100 ns, at 300 K temperature and 1.01325 bar pressure conditions.

5.4 Results and Discussion

Preliminary results from OBP-ligand MD simulations

OBP4-ligand interactions

a. Apo-OBP4

The crystal structure of apo- OBP4 when simulated in an aqueous environment remains in equilibrium (0.2 nm) throughout a period of 100 ns with intact alpha-helical secondary structure.

b. Complex OBP4-indole

In indole-bound OBP4, residue Ser110 interacts with OBP4 with 37% strength through hydrogen bonding and water bridge networks in the first 50 ns. Fluctuation of Pro79 is observed (0.185 nm).

c. Complex Indole-OBP4-(6-MH)

In the presence of 6-MH within the same binding cavity, slightly increased interactions of indole with OBP4 are observed through hydrogen bonding by Ser110 (49% strength) and pi-pi stacking interactions of Phe122 (31% strength). Partial helix to coil transition for helix 5 in OBP4 is observed. However, in the same complex in the presence of indole, Met123 interacts with 6-MH with 32% contact strength through hydrogen bonding and water bridges.

d. Complex 1NPN-OBP4-(6-MH)

In the presence of 1NPN, hydrogen bonding is observed between Thr58 of OBP4 and 6-MH with 45% contact strength, and 6-MH equilibrates only after 60 ns of simulation time. However, 1-NPN interacts with the C-terminus of OBP4 through hydrogen bonding, pi-pi stacking and other side-chain hydrophobic interactions. Phe122 interacts with the binding cavity with ~68% contact strength.

a. Complex indole-OBP4-icaridin

In the presence of icaridin, indole does not equilibrate well within 50 ns and does not have any significant interactions with OBP4. However, helical regions 3, 4 and 5 and coil region between helix 3 to helix 4 show interactions with icaridin, where residues such as Thr58 effectively interacts with icaridin with 62% strength.

b. Complex (indole-OBP4-(6-MH))-icaridin

I further tested two poses where the indole-OBP4-(6-MH) complex was docked with icaridin in the same binding cavity with weak and strong binding affinities, respectively (data not shown). Interestingly, in both docked poses in the presence of icaridin, the interactions of OBP4 with both indole and 6-MH reduce drastically and do not account for more than 20% of overall simulation time separately. The C-terminal tail of OBP4 makes strong contacts with icaridin through Met123 by hydrogen bonding, hydrophobic interactions and water bridges. In the tightly bound complex, hydrogen bonding and water bridge network between Thr58 and icaridin are also observed.

OBP1 monomer-ligand interactions

a. Complex 6MH-OBP1-indole

6-MH takes 40 ns to equilibrate with OBP1 in the presence of indole within the same cavity. Val125 forms hydrogen bonds with ketone O atom of 6-MH with ~92% contact strength. Indole bound to it interacts with Phe59 with 40% contact strength and Ser79 with 50% strength. Indole interactions with helices 3 and 4 and the coiled regions between them have also been observed.

OBP1 homodimer - ligand interactions

Results of OBP1 homodimer interactions with ligands from in-house ligand libraries (**Section 5.3**) are being investigated as future work.

OBP1-OBP4 heterodimer - ligand interactions

a. Complex Heterodimer-indole

In the presence of indole as the only ligand, C-terminal Phe122 (45%), Ser110 (30%) and Met77 (30%) of OBP4 subunit of the heterodimer interact with indole.

b. Complex Indole-heterodimer-Oleuropein

In the presence of phenolic metabolite oleuropein, the native interactions of indole with heterodimeric OBP4 reduce dramatically [Phe122 (19%), Met77 (negligible), Ser110 (5%)] although few additional interactions are observed through Ala107 (30%) and C-terminal Phe124 (30%).

5.5 Conclusion and Future Direction

An *in vivo* study (Beismann et al, 2010) revealed that OBP1 is required for the perception of indole in the female *Anopheles* mosquito. It is also observed *in vitro* that indole induces interactions between AgamOBP1 and AgamOBP4 (**Chapter 4**; Davarazou et al, 2011; Qiao et al, 2011; Ziemba et al, 2012; Murphy et al, 2013). Several of the questions outlined in **Section 5.2** have been attempted to probe into the nature of OBP-ligand interactions and to search for an effective mosquito repellant that would inhibit the binding of indole with OBP1 and other OBPs in female *Anopheles gambiae.*

Although indole is known to interact with both OBP1 and OBP4 *in vitro,* the addition of 6MH displaces indole and breaks heterodimeric interactions (Murphy et al, 2013). Indeed, 6MH (sulcatone) is effectively observed to bind to OBP1 with stronger interactions (92% contact strength) even in the presence of indole. However, in indole-bound OBP4, 6MH competes with it for interactions with the C-terminal tail. While 6MH can interact with indole-bound OBP1 and OBP4 separately, it is unable to displace fluorophore 1-NPN from OBP4 even in our simulations, in parallel with NMR experiments on 1-NPN and 6-MH binding with OBP1-OBP4 heterodimer (Murphy et al, 2013). Thus, our *in silico* findings are in agreement with experimental evidence of the mechanism of 6MH-OBP interactions.

Icaridin, a broad-spectrum synthetic insect repellent, appears to mask volatile odorants when presented together but does not directly interact with ORs in *Anopheles gambiae* (Afifiy et al, 2019). *In silico* analysis on **OBP4-ligand complexes docked with icaridin** suggests that both single-ligand and two-ligand complexes drastically reduce their interactions with indole and 6-MH (sulcatone) each irrespective of the binding affinity between icaridin and the complex. Instead, a strengthening of interactions between icaridin and C-terminal tail and mid-helical regions of OBP4 is observed through a complex interplay of water bridges, hydrogen bonding, pi-pi and other hydrophobic interactions. It would be interesting to determine how icaridin interacts with indole

and sulcatone in the course of MD simulations in the context of “masking” volatile odorants, as has been recently reported.

Furthermore, **oleuropein**, a naturally occurring phenolic secondary metabolite found in olive oil and in olive plant and fruit, was found to be a strong candidate to be tested *in vitro* as an inhibitor of OBP1-OBP4 heterodimeric interactions. Virtual screening of 112 naturally occurring volatiles (**Section 5.2**) helped identify oleuropein as the best repellant candidate against OBP-indole interactions through this work. This has implications in designing a natural and safe repellant against mosquitoes (or/and insects) by targeting their olfactory pathway.

The current work has also been extended to understand interactions of OBP1 with molecules in human sweat such as indole and will be investigated as future work. Additional analysis is currently underway regarding metabolites and volatiles screened against OBP model structures.

Furthermore, work done on **OBP1 homodimer** using modified **Supernatural DB** as ligand library (**Shailya V; not included in the thesis Chapter**) will aid the identification of candidate repellant compounds against mosquito OBPs, thus enhancing our understanding of olfaction in insects and the role of OBPs in the same.

Chapter 6

Mammalian OBP-pheromone interactions

6.1 Summary

It has been proposed that OBPs expressed in both nasal mucus and saliva of male buffaloes may be important in odour perception for communication of reproductive cues in the female buffalo (Rajkumar et al., 2010; Muthukumar et al., 2018a). However, the interactions between its OBPs and sex pheromones are unknown. In contrast to the alpha-helix rich secondary structure in insects (invertebrates), mammalian OBP is characterized by a beta-barrel lipocalin fold. A centrally located cavity enclosed by the beta-barrel has been considered the classical binding pocket. However, additional binding sites, if present, were not reported so far.

6.2 Specific questions

Here, I aimed to understand the structure and function of the nasal epithelium wild-type OBPs by *in-silico* approaches to

i. identify and predict probable binding sites for pheromone binding
ii. to investigate their molecular interactions and binding with buffalo estrus-specific pheromones (p-cresol and oleic acid) using docking, protein-ligand MD simulations of wild-type and mutant protein-ligand complexes and affinity analysis of simulation data.

6.3 Outline of the approach used

a. Multiple Sequence Alignment of bnOPB isoforms
b. Domain analysis
c. Protein structure optimization and binding cavity prediction
d. Ligand structure optimization
e. Protein-ligand docking for wild-type and mutant bnOBP isoforms
f. Protein-ligand binding affinity for wild-type and mutant protein-ligand complexes
g. Molecular Dynamic (MD) simulations for wild-type and mutant protein-ligand complexes
h. The binding affinity of wild-type and mutant protein-ligand complex derived from MD simulation trajectories
i. Analysis and prediction of key OBP – pheromone interactions in wild-type and mutant models

6.4 Published research article related to the study

The published manuscript detailing the work above refers to –

Chidhambaram Manikkaraja*, **Mam Bhavika***, Randhir Singh, Balasubramanian Nagarathnam, Geen George, Akash Gulyani, Govindaraju Archunan & Ramanathan Sowdhamini (2020) Molecular and functional characterization of buffalo nasal epithelial odorant binding proteins and their structural insights by *in silico* and biochemical approaches Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1854117

In this article, background, methodology adopted for the binding studies are provided along with related results, discussion and corresponding supplementary files have been provided.

Movie files M1 to M20 referenced throughout the manuscript have been provided as supplementary files in the following link.

https://docs.google.com/document/d/1vzTZZzglZnpKaGqY2rCUlxxUoNC1AqyUdoUQQnjMYzQ/edit?usp=sharing

Chapter 7
Conclusion

7.1 Summary

An insect relies typically on the modalities of olfaction, vision and gustation to various degrees depending on the ecological niche and sensory adaptations to its environment for survival and reproduction. Olfaction is a complex chemosensory process that requires a rapid relay of signals present as environmental chemical cues to the brain to elicit a response that would preserve or enhance the fitness of an organism.

Odorant-binding proteins have been studied for the last forty years to understand their localization, binding properties, and functional roles, conventional and unconventional (**Chapter 1**). Genome-wide surveys and subfamily-based investigations were performed by various groups in the scientific community to facilitate this. Experiments performed to identify OBPs and analyze their significance typically involve *in silico*, *in vitro* and *in vivo* approaches.

My doctoral thesis is a sincere attempt towards the functional understanding of OBP using sequence and structure of insect and mammalian OBPs as a start-point. Towards this, I devised and implemented a machine learning pipeline that integrates information derived from genomic and protein sequence data to accurately predict (>85%) whether a user-annotated OBP has been correctly annotated as an OBP or is more likely to belong to other classes of soluble proteins with similar features that are also involved in insect chemical communication and found dissolved in sensillar lymph (**Chapter 2, 2.2; Chapter 3**). This classification pipeline aims to help towards the large-scale annotation of OBP subfamilies- Classic, Atypical, Minus-C, Plus-C, as well as NPC2, CSP and D7 protein families across insect orders.

An increasing number of studies reveal the presence of OBPs in various chemosensory and non-chemosensory organs of the insect body and having previously unknown roles in taste, humidity sensing and immunity response of the insect. In order to make rapid and accurate predictions of diverse ligands with classical and atypical but likely binding pockets, I was heavily involved in the synthesis and application of a deep learning pipeline (**Chapter 2, 2.3**) that rapidly predicts the binding affinity of a given protein (resolution > 2.5 Å) with a non-peptide ligand using only sequence homology and standard protein and ligand features. The docked pose is not required to be provided by the user. Predictions made by DEELIG were observed as per existing biologically relevant observations (e.g., Ppc homologs) and were used in predicting affinities for COVID-19

protein–drug complexes using case studies. As was the original objective behind DEELIG, I have extended this work on odorant-binding proteins and their binding predictions with a diverse set of ligands.

Previous work from our lab (Karpe et al., 2016; 2019; 2020) resulted in a homology-based genomic pipeline to detect insect odorant receptors. My work standardized this pipeline to be applied to insect OBPs. This modified approach coupled with phylogeny and other validation measures when applied on eusocial honeybee ***Apis florea*** led to the identification of 22 OBPs. This work also added to the growing hypothesis that the Minus-C OBP subfamily has diverged from a member of the Classic OBP subfamily. It could be perhaps an ecological or functional adaptation. As in **Chapter 2, 2.2** on infectious disease mosquito vectors, the application of the approach revealed that unlike in Hymenoptera, Dipteran OBPs appear to have a higher number of Minus-C OBPs (**Chapter 3**). This opens more interesting questions, such as whether the frequency of OBP subfamilies varies across insect orders or ecological niche.

Focusing on specific OBPs led us to hypothesis-driven modelling of heterodimeric interactions between two OBPs of female ***Anopheles gambiae***. Through all-atom MD simulations and unsupervised learning, it was observed that both pH and indole influence heterodimeric interactions of OBP1 and OBP4 (**Chapter 4**). Furthermore, virtual screening and MD of monomeric and heterodimeric AgamOBP models with small ligands helped identify naturally occurring biomolecules that served as active volatiles and metabolites as suitable targets to the indole-binding site in OBP4 (**Chapter 5**). The ligands shortlisted from this study and a previous study (unpublished thesis work of Dr. Malini Manoharan) can be used *in vitro* and further through behavioural assays as candidates to determine competitive inhibition of OBP-indole binding.

Finally, *in silico* work performed in collaboration with Prof G Archunan (Centre for Pheromone Research, Bharathidasan University), on mammalian OBP (bnOBP; buffalo OBP) isoform models (wildtype and mutant) identified an additional putative binding site that is laterally located and residues (Phe69, Phe104, Asn118 and Phe134) potentially involved in interacting with estrous cycle-regulated pheromones p-cresol and oleic acid (**Chapter 6**). Our predictions were successfully validated using fluorescence binding experiments by our collaborative group. The *in silico* work I performed in this project using docking, MD simulations, and affinity analysis further opened up questions such as experimentally validating the predicted lateral binding cavity. It would also be worthwhile to extend the *in silico* work to prepare a residue contact map of

mammalian OBP (bnOBP) isoforms with each volatile that is perceived as a reproductive and mating cue in the organism.

It is an exciting time for unravelling the typical and atypical functions of OBPs across vertebrates and invertebrates, given rapid advancements in sequencing, computing and molecular biology.

Summary

The olfactory system is capable of detecting and discriminating among environmental odorants that play a key role in various key behaviours. Among them, odorant-binding proteins (OBPs) have been highlighted as important components of the olfactory apparatus, as they participate in the recognition of odorants. OBPs undergo conformational changes as a result of pH changes and/or ligand binding altering their interactions with odorants.

The work in the course of my thesis has aimed to understand the function of odorant-binding proteins in the context of sequence, structure and evolution across insect orders and mammals. There were several questions that were asked, investigated and answered along these lines.

In a step towards identifying OBPs across **insect species**, an integrated genomics and machine-learning pipeline was constructed to predict whether a protein sequence belongs to the OBP subfamily among other classes. The predicted protein sequences further helped in understanding evolutionary relationships among various subfamilies of OBPs. A deep learning framework was also constructed to predict protein-ligand binding affinity without the need of an input docking pose. Genome-wide survey of OBPs across insect genomes including eusocial honeybee, *Apis florea*, indicated the evolutionary relationships between MinusC and the Classic subfamilies. In the **hemolymph of malaria-causing female *Anopheles gambiae,*** OBP1 and OBP4 may form heterodimers to facilitate perception of indole, different from those of the individual proteins, thus increasing the number of detectable odors. The current thesis work shows that decrease in pH is associated with a concomitant reduction in the inter- and intra-monomeric dynamic fluctuations of the indole-bound heterodimer. Further work in my thesis has identified a set of naturally derived compounds as inhibitors to heterodimeric OBP-indole and monomeric OBP-indole complexes in mosquito through computational virtual screening and molecular dynamic simulations. Protein-ligand docking, molecular dynamic (MD) simulations and thermal binding affinity analysis of OBP isoforms from the **buffalo nasal epithelium** revealed and was further experimentally verified that four residues (Phe69, Phe104, Asn118 and Phe134) from OBP1a, contributed to strong binding affinities towards two sex pheromones, specifically oleic acid and p-cresol.

Taken together, these studies lead us towards the identification, and understanding of the mechanistic and functional aspects of odorant-binding proteins.

References

Ahmed, A., Mam, B., & Sowdhamini, R. (2021). DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinformatics and Biology Insights*. DOI:10.1177/11779322211030364

Ahmed Aqeel, Smith Richard D, Clark Jordan J, Dunbar Jr. James B, and Carlson Heather. Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucl. Acids Res.*, **2014**, *43 (D)*: 465–469. DOI: 10.1093/nar/gku1088

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. DOI:10.1016/S0022-2836(05)80360-2

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1: 19-25

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66: 213-21

Ahmed Aqeel, Smith Richard D, Clark Jordan J, Dunbar Jr. James B, and Carlson Heather A. Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucl. Acids Res.*, 2014, *43 (D)*: 465–469. DOI: 10.1093/nar/gku1088

Afify A, Betz JF, Riabinina O, Lahondère C, Potter CJ. (2019). Commonly Used Insect Repellents Hide Human Odors from *Anopheles* Mosquitoes. *Curr Biol*. 29(21):3669-3680.e5. DOI:10.1016/j.cub.2019.09.007

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton (2017). Imagenet classification with deep convolutional neural networks. Commun. ACM, 1097–1105. DOI: 10.1145/3065386

Amadei A, Linssen AB, Berendsen HJ (1993) Essential dynamics of proteins. *Proteins* 17: 412-25

Anet, E.F.L.J. (1972). Superficial scald, a functional disorder of stored apples VIII. Volatile products from the autoxidation of α-farnesen. *J. Sci. Food Agric.*, 23: 605-608. DOI:10.1002/jsfa.2740230508

Andronopoulou E, Labropoulou V, Douris V, Woods DF, Biessmann H, Iatrou K (2006) Specific interactions among odorant-binding proteins of the African malaria vector *Anopheles gambiae*. *Insect Mol Biol* 15: 797-811

Angeli, S., Ceron, F., Scaloni, A., Monti, M., Monteforti, G., Minnocci, A., Petacchi, R., & Pelosi, P. (1999). Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *European Journal of biochemistry*, 262(3), 745–754. http://hdl.handle.net/11382/305335

Archunan G. (2018). Odorant Binding Proteins: a key player in the sense of smell. *Bioinformation*, 14(1), 36–37. DOI: 10.6026/97320630014036

Archunan, G., (2020). Reproductive enhancement in buffalo: Looking at urinary pheromones and hormones. *Iranian Journal of Veterinary Research.* DOI:10.22099/ijvr.2020.34903.5149.

Archunan, G., Rajanarayanan, S., & Karthikeyan, K. (2014). Cattle Pheromones. In C. Mucignat-Caretta (Ed.), *Neurobiology of Chemical Communication*. CRC Press/Taylor & Francis.

Archunan, G., Rajanarayanan, S. (2010) Composition for enhancing bull sex libido. Indian Patent No. 244991

Armen R, Alonso DO, Daggett V (2003) The role of alpha-, 3(10)-, and pi-helix in helix--coil transitions. *Protein Sci* 12: 1145-57

Ausili A, Pennacchio A, Staiano M, Dattelbaum JD, Fessas D, Schiraldi A, D'Auria S (2013) Amino acid transport in thermophiles: Characterization of an arginine-binding protein from *Thermotoga maritima*. 3. Conformational dynamics and stability. *Journal of Photochemistry and Photobiology B-Biology* 118: 66-73. DOI: 10.1016/j.jphotobiol.2012.11.004.

Bacchini, A., Gaetani, E., & Cavaggioni, A. (1992). Pheromone binding proteins of the mouse, *Mus musculus*. *Experientia*, 48(4), 419–421. DOI: 10.1007/BF01923448.

Barbiroli A, Bonomi F, Ferranti P, Fessas D, Nasi A, Rasmussen P, Iametti S (2011) Bound Fatty Acids Modulate the Sensitivity of Bovine beta-Lactoglobulin to Chemical and Physical Denaturation. *Journal of Agricultural and Food Chemistry* 59: 5729-5737

Barone G, Delvecchio P, Fessas D, Giancola C, Graziano G (1992) Theseus - a New Software Package for the Handling and Analysis of Thermal-Denaturation Data of Biological Macromolecules. *Journal of Thermal Analysis* 38: 2779-2790

Ache Barry W., Young Janet M. (2005). Olfaction: Diverse Species, Conserved Principles, Neuron, 48 (3), 417-430. DOI:10.1016/j.neuron.2005.10.022.

Bekker H, Berendsen HJC, Dijkstra EJ, Achterop S, Vondrumen R, Vanderspoel D, Sijbers A, Keegstra H, Reitsma B, Renardus MKR (1993) Gromacs - a Parallel Computer for Molecular-Dynamics Simulations. *Physics Computing '92*: 252-256

Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, and Carlson HA. (2008). Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.*, 36 (D): 674–678. DOI: 10.1093/nar/gkm911

Bentley MD, Day JF (1989) Chemical Ecology and Behavioral-Aspects of Mosquito Oviposition. *Annual Review of Entomology* 34: 401-421

Berg HC. (1975). CHEMOTAXIS IN BACTERIA. *Annu. Rev. Biophys. Bioeng.* 4:119–136.

Berendsen HJC, Vanderspoel D, Vandrunen R (1995) Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation. *Comput Phys Commun* 91: 43-56

Bhavika Mam, Ramanathan Sowdhamini (2020). SoCCer: A pipeline to identify classes of soluble proteins in chemical communication in insect genomes, *PREPRINT (Version 1) available at Research Square*. DOI: 10.21203/rs.3.pex-1095/v1

Bhavika Mam, Snehal Karpe, Ramanathan Sowdhamini (2021). Genome-wide survey of odorant-binding proteins in the dwarf honey bee *Apis florea*. *bioRxiv* 2021.02.25.432941; doi: DOI:10.1101/2021.02.25.432941

Bianchet, M. A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S. H., Monaco, H. L., & Amzel, L. M. (1996). The three-dimensional structure of bovine odorant binding protein and its mechanism of odor recognition. *Nature Structural Biology*, 3(11), 934–939. DOI:10.1038/nsb1196-934

Biessmann H, Andronopoulou E, Biessmann MR, Douris V, Dimitratos SD, Eliopoulos E, Guerin PM, Iatrou K, Justice RW, Krober T, Marinotti O, Tsitoura P, Woods DF, Walter MF (2010) The *Anopheles gambiae* odorant binding protein 1 (AgamOBP1) mediates indole recognition in the antennae of female mosquitoes. *PLoS One* 5: e9471

Biessmann H, Walter MF, Dimitratos S, Woods D (2002) Isolation of cDNA clones encoding putative odourant binding proteins from the antennae of the malaria-transmitting mosquito, Anopheles gambiae. *Insect Molecular Biology* 11: 123-132

Bignetti, E., Cavaggioni, A., Pelosi, P., Persaud, K. C., Sorbi, R. T., & Tirindelli, R. (1985). Purification and characterisation of an odorant-binding protein from cow nasal tissue. *European Journal of Biochemistry*, 149(2), 227–231. DOI:10.1111/j.1432-1033.1985.tb08916.x

Blank I, Sen A, Grosch W. (1991). Aroma impact compounds of *Arabica* and *Robusta* coffee. Qualitative and quantitative investigations. In: *ASIC*, 14e Colloque, San Francisc. 117–129.

Böcskei, Z., Groom, C. R., Flower, D. R., Wright, C. E., Phillips, S. E., Cavaggioni, A., Findlay, J. B., & North, A. C. (1992). Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature*, 360(6400), 186–188. DOI:10.1038/360186a0

Boelens H. (1983). Structure activity relationships in chemoreception by human olfaction. *Trends Pharmacol Sci.* 4(10):421–6

Boix E (2001) Eosinophil cationic protein. *Methods Enzymol* 341: 287-305

Bonadonna F, Nevitt GA (2004) Partner-specific odor recognition in an Antarctic seabird. Science 306:835

Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., ... & Salmon, J. K. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, IEEE.

Brand P *et al*. (2015). Rapid evolution of chemosensory receptor genes in a pair of sibling species of orchid bees (Apidae: Euglossini). *BMC Evol. Biol.*, 15:176.

Brand P, Ramírez SR. (2017). The Evolutionary Dynamics Of The Odorant Receptor Gene Family In Corbiculate Bees. *Genome Biol. Evol.* 9:2023–2036.

Brockmann A, Brückner D. (2001). Structural differences in the drone olfactory system of two phylogenetically distant *Apis* species, *A. florea* and *A. mellifera*. *Naturwissenschaften.* 88:78–81.

Brockmann A, Dietz D, Spaethe J, Tautz J. (2006). Beyond 9-ODA: sex pheromone communication in the European honey bee *Apis mellifera* L. *J. Chem. Ecol.* 32:657–67.

Buck L, Axel R. (1991). A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. Cell. 65:175–187.

Breer, H., Krieger, J., Raming, K. (1990). A novel class of binding proteins in the antennae of the silk moth *Antheraea pernyi*. Insect Biochemistry 20, 735–740.

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

Briand, L., Huet, J., Perez, V., Lenoir, G., Nespoulous, C., Boucher, Y., Trotier, D., & Pernollet, J. C. (2000). Odorant and pheromone binding by aphrodisin, a hamster aphrodisiac protein. *FEBS letters*, 476(3), 179–185. DOI:10.1016/S0014-5793(00)01719-1.

Brito NF, Moreira MF, Melo ACA (2016) A look inside odorant-binding proteins in insect chemoreception. *Journal of Insect Physiology* 95: 51-65

Brown, R. E., & Macdonald, D. W. (1985). *Social odours in mammals*. Oxford University Press, Ox- ford, United Kingdom 1:1-50.

Burova, T. V., Choiset, Y., Jankowski, C. K., & Haertlé, T. (1999). Conformational stability and binding properties of porcine odorant binding protein. *Biochemistry*, 38(45), 15043–15051. DOI:10.1021/bi990769s

Campanacci, V., Krieger, J., Bette, S., Sturgis, J.N., Lartigue, A., Cambillau, C., Breer, H., Tegoni, M. (2001). Revisiting the specificity of *Mamestra brassicae* and *Antheraea polyphemus* pheromone-binding proteins with a fluorescence binding assay. *Journal of Biological Chemistry* 276, 20078–20084.

Carey AF, Wang GR, Su CY, Zwiebel LJ, Carlson JR (2010) Odorant reception in the malaria mosquito *Anopheles gambiae*. *Nature* 464: 66-U77

Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26: 1668-88

Case DA, Darden TA, Cheatham I, T. E., Simmerling C, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts BP, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Liu J et al. (2010) *AMBER 11, University of California, San Francisco, CA*

Cavaggioni, A., & Mucignat-Caretta, C. (2000). Major urinary proteins, alpha(2U)-globulins and aphrodisin. *Biochimica et biophysica acta*, *1482*(1-2), 218–228. DOI: 10.1016/s0167-4838(00)00149-7

Cavaggioni, A., Sorbi, R. T., Keen, J. N., Pappin, D. J., & Findlay, J. B. (1987). Homology between the pyrazine-binding protein from nasal mucosa and major urinary proteins. *FEBS letters*, 212(2), 225–228. DOI:10.1016/0014-5793(87)81349-2

Cavalieri, J., Eagles, V., Ryan, M., & Macmillan, K. L. (2003). Comparison of four methods for detection of oestrus in dairy cows with resynchronised oestrous cycles. *Australian Veterinary Journal*, 81(7), 422–425. DOI:10.1111/j.1751-0813.2003.tb11553.x

Černá, M., Kuntová, B., Talacko, P., Stopková, R., & Stopka, P. (2017). Differential regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the house mouse. *Scientific Reports*, 7(1), 11674. DOI:10.1038/s41598-017-12021-2

Chakrabarti S, Sowdhamini R (2004) Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modelling using distant relationships. *FEBS Letters* 569: 31-36

Chidhambaram Manikkaraja, Mam Bhavika, Randhir Singh, Balasubramanian Nagarathnam, Geen George, Akash Gulyani, Govindaraju Archunan & Ramanathan Sowdhamini (2020). Molecular and functional characterization of buffalo nasal epithelial odorant binding proteins and their structural insights by in silico and biochemical approaches , *Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1854117*

Chun Wei Yap (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem.*, *32 (7)*:1466–1474. DOI: 10.1002/jcc.21707.

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* 20: 426-7 COLLABORATIVE COMPUTATIONAL PROJECT N (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50: 760-3

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J Am Chem Soc* 117: 5179-5197

Dal Monte, M., Andreini, I., Revoltella, R., & Pelosi, P. (1991). Purification and characterization of two odorant-binding proteins from nasal tissue of rabbit and pig. Comparative biochemistry and physiology. B, Comparative biochemistry, 99(2), 445–451. DOI:10.1016/0305-0491(91)90068-O

Damberger F, Nikonova L, Horst R, Peng G, Leal WS, Wuthrich K (2000) NMR characterization of a pH-dependent equilibrium between two folded solution conformations of the pheromone-binding protein from Bombyx mori. *Protein Sci* 9: 1038-41

Damberger FF, Michel E, Ishida Y, Leal WS, Wuthrich K (2013) Pheromone discrimination by a pH-tuned polymorphism of the Bombyx mori pheromone-binding protein. *Proceedings of the National Academy of Sciences of the United States of America* 110: 18680-18685

David H Hubel and Torsten N Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.*, 160 (1):106–154. DOI: 10.1113/jphysiol.1962.sp006837

Davrazou F, Dong E, Murphy EJ, Johnson HT, Jones DN (2011) New insights into the mechanism of odorant detection by the malaria-transmitting mosquito Anopheles gambiae. *J Biol Chem* 286: 34175-83

de Bruyne, M., Baker, T.C. (2008). Odor Detection in Insects: Volatile Codes. J Chem Ecol 34, 882–897. DOI:10.1007/s10886-008-9485-4

DeLano WL (2002) The PyMOL Molecular Graphics System, DeLano Scientific. In Palo Alto, CA, USA:

Deng, T., Pang, C., Lu, X., Zhu, P., Duan, A., Tan, Z., Huang, J., Li, H., Chen, M., & Liang, X. (2016). De Novo Transcriptome Assembly of the Chinese Swamp Buffalo by RNA Sequencing and SSR Marker Discovery. PloS one, 11(1), e0147132. DOI:10.1371/journal.pone.0147132

Desaphy Jérémy, Bret Guillaume, Rognan Didier, and Kellenberger Esther (2015). sc-PDB: a 3D-database of ligandable binding sites—10 years on. Nucl. Acids Res., 43 (D1):399–404. DOI: 10.1093/nar/gku928

Di Luccio E, Ishida Y, Leal WS, Wilson DK (2013) Crystallographic Observation of pH-Induced Conformational Changes in the *Amyelois transitella* Pheromone-Binding Protein AtraPBP1. *Plos One* 8(2): e53840. DOI: 10.1371/journal.pone.0053840.

Diederik P Kingma and Jimmy Ba (2014). Adam: A method for stochastic optimization, *arXiv*, Preprint arXiv: 1412.6980.

Ding C, He XF (2004) Cluster structure of K-means clustering via principal component analysis. *Lect Notes Artif Int* 3056: 414-418

Ditzen M, Pellegrino M, Vosshall LB (2008). Insect odorant receptors are molecular targets of the insect repellent DEET. Science. 319(5871):1838-42. DOI: 10.1126/science.1153121.

Du, G., & Prestwich, G. D. (1995). Protein Structure Encodes the Ligand Binding Specificity in Pheromone Binding Proteins. Biochemistry, 34(27), 8726–8733

Eaton BE, Gold L, Zichi DA. (1995). Let's get specific: the relationship between specificity and affinity. Chem Biol., 2(10):633-8. DOI: 10.1016/1074-5521(95)90023-3.

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography* 60: 2126-2132

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin (2004). UCSF Chimera -- a visualization system for exploratory research and analysis. *J Comp Chem., 25 (13)*:1605–1612. DOI: 10.1002/jcc.20084.

Fleischer, J., Breer, H., & Strotmann, J. (2009). Mammalian olfactory receptors. Frontiers in Cellular Neuroscience, 3, 9. . DOI:10.3389/neuro.03.009.2009

Flower D. R. (1996). The lipocalin protein family: structure and function. The Biochemical journal, 318 (Pt 1), 1–14. DOI:10.1042/bj3180001

Flower, D. R., North, A. C., & Sansom, C. E. (2000). The lipocalin protein family: structural and sequence overview. *Biochimica et biophysica acta*, *1482*(1-2), 9–24.

Forêt S, Maleszka R. (2006). Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). *Genome Res*. 16(11):1404-1413. DOI:10.1101/gr.5075706

Forêt S, Wanner KW, Maleszka R (2007). Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. *Insect Biochem Mol Biol*. 37(1):19-28. DOI: 10.1016/j.ibmb.2006.09.009. Erratum in: Insect Biochem Mol Biol. 2015 Oct; 65:107. PMID: 17175443.

Forgy EW (1965) Cluster Analysis of Multivariate Data - Efficiency Vs Interpretability of Classifications. *Biometrics* 21: 768-&

French S, Wilson K (1978) Treatment of Negative Intensity Observations. *Acta Crystallographica Section A* 34: 517-525

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739–1749. DOI:10.1021/jm0306430

G. E. Dahl, T. N. Sainath and G. E. Hinton (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, *8609-8613*, DOI: 10.1109/ICASSP.2013.6639346.

Gandhimathi A, Nair AG, Sowdhamini R (2012). PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic Acids Res*: D531-4. DOI: 10.1093/nar/gkr1096.

Gandhimathi, A., & Sowdhamini, R. (2016). Molecular modelling of human 5-hydroxytryptamine receptor (5-HT2A) and virtual screening studies towards the identification of agonist and antagonist molecules. *Journal of Biomolecular Structure & Dynamics*, *34*(5), 952–970.

Garibotti M, Navarrini A, Pisanelli AM, Pelosi P. (1997). Three odorant-binding proteins from rabbit nasal mucosa. *Chem Senses* 22(4):383-390. DOI:10.1093/chemse/22.4.383

Gherardi, F., Tricarico, E. & Atema, J. (2005).Unraveling the Nature of Individual Recognition by Odor in Hermit Crabs. *J Chem Ecol* **31,** 2877–2896. DOI:10.1007/s10886-005-8400-5

Gomez-Diaz C, Reina JH, Cambillau C, Benton R (2013) Ligands for pheromone-sensing neurons are not conformationally activated odorant binding proteins. *PLoS Biol* 11: e1001546

Gouet P, Courcelle E, Stuart DI, Metoz F (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* 15: 305-8

Graham LA, Davies PL. (2002). The odorant-binding proteins of *Drosophila melanogaster*: annotation and characterization of a divergent gene family. *Gene*. 292(1-2):43-55. DOI: 10.1016/s0378-1119(02)00672-8.

Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22: 2695-2696

Grebe TW, Stock J. (1998). Bacterial chemotaxis: The five sensors of a bacterium. *Curr. Biol.* 8:R154–R157.

Greenwood, J. R., Calkins, D., Sullivan, A. P., & Shelley, J. C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-aided Molecular Design*, 24(6-7), 591–604. DOI:10.1007/s10822-010-9349-1

Guariento M, Assfalg M, Zanzoni S, Fessas D, Longgu R, Molinari H (2010) Chicken ileal bile-acid-binding protein: a promising target of investigation to understand binding co-operativity across the protein family. *Biochemical Journal* 425: 413-424

Guiraudie, G., Pageat, P., Cain, A. H., Madec, I., & Nagnan-Le Meillour, P. (2003). Functional characterization of olfactory binding proteins for appeasing compounds and molecular cloning in the vomeronasal organ of pre-pubertal pigs. *Chemical senses*, 28(7), 609–619. DOI:10.1093/chemse/bjg052

Gupta M (1986). A quantitative study and ultrastructure of flagellar sensillae of *Apis florea* F. (Hymenoptera: Apidae). *Proc. Indian Acad. Sci. (Anim. Sci.)*, 95: 5: 595-603.

Gupta M (1992). Scanning electron microscopic studies of antennal sensilla of adult worker *Apis florea* F (Hymenoptera: Apidae). *Apidologie*, Springer Verlag, 23 (1), 47-56.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34 (17):i821–i829. DOI: 10.1093/bioinformatics/bty593

Halgren T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2), 377–389. DOI:10.1111/j.1747-0285.2007.00483.x

Hallberg E, Hansson BS, Löfstedt C (2003) Sensilla and proprioreceptors. In: Kristensen NP (ed.), Lepidoptera, Moths and Butterflies: morphology, physiology and development, 2. WdG, New York, Berlin, 267–288

Halling L, Oldroyd B, Wattanachaiyingcharoen, W et al. (2001). Worker policing in the bee *Apis florea*. Behav Ecol Sociobiol 49, 509–513. DOI: 10.1007/s002650100325

Han L, Zhang YJ, Zhang L, Cui X, Yu J, Zhang Z, Liu MS (2014) Operating mechanism and molecular dynamics of pheromone-binding protein ASP1 as influenced by pH. *PLoS One* 9: e110565

Han, L., Zhang, Y. J., Zhang, L., Cui, X., Yu, J., Zhang, Z., & Liu, M. S. (2014). Operating mechanism and molecular dynamics of pheromone-binding protein ASP1 as influenced by pH. PloS one, 9(10), e110565. DOI:10.1371/journal.pone.0110565

Hansson BS (1999) *Insect Olfaction*. Springer-Verlag Berlin Heidelberg,

Hansson BS, Stensmyr MC (2011) Evolution of Insect Olfaction. *Neuron* 72: 698-711

Hayward S, de Groot BL (2008) Normal modes and essential dynamics. *Methods Mol Biol* 443: 89-106

He P, Li ZQ, Zhang YF, Chen L, Wang J, Xu L, Zhang YN and He M. (2017). Identification of odorant-binding and chemosensory protein genes and the ligand affinity of two of the encoded proteins suggest a complex olfactory perception system in *Periplaneta americana*. *Insect Mol Biol*, 26: 687-701. DOI: 10.1111/imb.12328

He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform. 2017;9:24. doi:10.1186/s13321-017-0209-z.

Hekmat-Scafe DS, Scafe CR, McKinney AJ, Tanouye MA (2002). Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Res*. 12(9):1357-1369. DOI:10.1101/gr.239402

Henikoff S, Henikoff, J.G (1992). Amino Acid Substitution Matrices from Protein Blocks. *PNAS*. 89 (22): 10915–1091910919. DOI:10.1073/pnas.89.22.10915

Hepburn, RH, Radloff S, Otis G, Fuchs S, Verma L, Tan K, *et al*. (2005). *Apis florea*: morphometrics, classification and biogeography. *Apidologie*, 36(3), 359-376.

Hildebrand JG, and Shepherd, GM (1997). Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annu. Rev. Neurosci*. 20, 595–631. DOI: 10.1146/annurev.neuro.20.1.595

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 24: 498-520

Hu L and RD Smith MG Lerner HA Carlson ML, Benson (2005). Binding MOAD (Mother Of All Databases. *Proteins*, 60:333–40. DOI : 10.1002/prot.20512

Hubbard, S.J. and Thornton, J.M. (1993). 'NACCESS', computer program, Department of Biochemistry and Molecular Biology, University College, London.

Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* 14: 33-38

Hurst, J. L., & Beynon, R. J. (2004). Scent wars: the chemobiology of competitive signalling in mice. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *26*(12), 1288–1298.

Ilayaraja, R., Rajkumar, R., Rajesh, D., Muralidharan, A. R., Padmanabhan, P., & Archunan, G. (2014). Evaluating the binding efficiency of pheromone binding protein with its natural ligand using molecular docking and fluorescence analysis. *Scientific Reports*, 4, 5201. DOI:10.1038/srep05201

Ishida Y, Tsuchiya W, Fujii T, Fujimoto Z, Miyazawa M, Ishibashi J, Matsuyama S, Ishikawa Y, Yamazaki T. (2014). Niemann-Pick type C2 protein mediating chemical communication in the worker ant. *Proc Natl Acad Sci U S A.* 11;111(10):3847-52. DOI: 10.1073/pnas.1323928111.

Izaguirre JA, Catarello DP, Wozniak JM, Skeel RD (2001) Langevin stabilization of molecular dynamics. *J Chem Phys* 114: 2090-2098

Jacobson, M. P., Friesner, R. A., Xiang, Z., & Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology*, 320(3), 597–608. DOI:10.1016/S0022-2836(02)00470-9

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., & Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2), 351–367. DOI:10.1002/prot.10613

Jacquin-Joly E, Vogt RG, François MC, Nagnan-Le Meillour P. (2001). Functional and expression pattern analysis of chemosensory proteins expressed in antennae and pheromonal gland of Mamestra brassicae. *Chem. Senses.* 26, 833–844. DOI:10.1093/chemse/26.7.833

Jin, Z., Du, X., Xu,Y. et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature, 582, 289–293. DOI: 10.1038/s41586-020-2223-y

Jolliffe IT (2002) Principal component analysis. Springer, New York

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6), 1657–1666. DOI:10.1021/ja00214a001

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45), 11225-11236. DOI:10.1021/ja9621760

Jurcik A, Bednar D, Byska J, Marques SM, Furmanova K, Daniel L, Kokkonen P, Brezovsky J, Strnad O, Stourac J, Pavelka A, Manak M, Damborsky J, Kozlikova B (2018) CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics* 34: 3586-3588

Kabsch W, Sander C (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22: 2577-2637

Kaissling KE (2001) Olfactory perireceptor and receptor events in moths: A kinetic model. *Chemical Senses* 26: 125-150

Karen Simonyan and Andrew Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*, preprint arXiv:1409.1556.

Karpe S.D., Jain R., Brockmann A. and Sowdhamini R. (2016). Identification of Complete Repertoire of *Apis florea* Odorant Receptors Reveals Complex Orthologous Relationships with *Apis mellifera*. Genome Biology and Evolution, 8, 2879–2895.

Karpe SD, Dhingra S, Brockmann A. and Sowdhamini R. (2017). Computational genome-wide survey of odorant receptors from two solitary bees *Dufourea novaeangliae* (Hymenoptera: Halictidae) and *Habropoda laboriosa* (Hymenoptera: Apidae). *Sci. Rep,* 7, 10823.

Karpe SD, Tiwari V, Ramanathan S. (2021). InsectOR—Webserver for sensitive identification of insect olfactory receptor genes from non-model genomes. *PLOS ONE* 16(1): e0245324. DOI: 10.1371/journal.pone.0245324

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–80.

Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research, 33(2), 511–518. DOI:10.1093/nar/gki198

Kaufman, L. and Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York. DOI:10.1002/9780470316801

Kaupp U. B. (2010). Olfactory signalling in vertebrates and insects: differences and commonalities. Nature Reviews Neuroscience, 11(3), 188–200. DOI:10.1038/nrn2789

Takeshi Kawabata. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins*, **2010**, *78 (5)*:1195–1211. DOI: 10.1002/prot.22639

Keil TA (1984) Surface coats of pore tubules and olfactory sensory dendrites of a silkmoth revealed by cationic markers. *Tissue Cell* 16: 705-17

Keller, A., Vosshall, L.B. (2016). Olfactory perception of chemically diverse molecules. *BMC Neurosci* 17, 55. DOI:10.1186/s12868-016-0287-2).

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic acids research*, *47*(D1), D1102–D1109.

Kotani T, Ozaki M, Matsuoka K, Snell TW, Hagiwara A (2001). Reproductive isolation among geographically and temporally isolated marine *Brachionus strains*. *Hydrobiologia* **446**: 283–290.

Kotlowski, C., Larisika, M., Guerin, P. M., Kleber, C., Kröber, T., Mastrogiacomo, R., ... & Knoll, W. (2018). Fine discrimination of volatile compounds by graphene-immobilized odorant-binding proteins. *Sensors and Actuators B: Chemical*, 256, 564-572. DOI:10.1016/j.snb.2017.10.093

Krieger, J., Ganble, H., Raming, K., Breer, H. (1993). Odorant binding proteins of Heliothis virescens. *Insect Biochemistry and Molecular Biology* 23, 449–456.

Krieger, J., von Nickisch-Rosenegk, E., Mameli, M., Pelosi, P., Breer, H. (1996). Binding proteins from the antennae of *Bombyx mori*. Insect Biochemistry and Molecular Biology 26, 297–307

Kumar N, Nayyar K, Sharma R and Anudeep (2014). Ultramorphology of antennal sensilla of open-nesting honey bees *Apis florea F. and Apis dorsata F.* (Hymenoptera: Apidae*) J. Appl. & Nat. Sci.* 6 (1): 315-319.

Larisika, M., Kotlowski, C., Steininger, C., Mastrogiacomo, R., Pelosi, P., Schütz, S., Peteu, S. F., Kleber, C., Reiner-Rozman, C., Nowak, C., & Knoll, W. (2015). Electronic Olfactory Sensor Based on A. mellifera Odorant-Binding Protein 14 on a Reduced Graphene Oxide Field-Effect Transistor. *Angewandte Chemie* (International ed. in English), 54(45), 13245–13248. DOI:10.1002/anie.201505712

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* (Oxford, England), 23(21), 2947–2948. DOI:10.1093/bioinformatics/btm404

Larter NK, Sun JS, Carlson JR (2016). Organization and function of *Drosophila* odorant binding proteins. *Elife* 5:e20242. DOI: 10.7554/eLife.20242

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

Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography* 26: 283-291

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), 283-291. DOI:10.1107/S0021889892009944

Laughlin JD, Ha TS, Jones DN, Smith DP (2008) Activation of pheromone-sensitive neurons is mediated by conformational activation of pheromone-binding protein. *Cell* 133: 1255-65

Lautenschlager C, Leal WS, Clardy J (2005) Coil-to-helix transition and ligand release of *Bombyx mori* pheromone-binding protein. *Biochem Biophys Res Commun* 335: 1044-50

Leal WS, Chen AM, Erickson ML (2005a) Selective and pH-dependent binding of a moth pheromone to a pheromone-binding protein. *Journal of Chemical Ecology* 31: 2493-2499

Leal WS, Chen AM, Ishida Y, Chiang VP, Erickson ML, Morgan TI, Tsuruda JM (2005b) Kinetics and molecular properties of pheromone binding and release. *Proceedings of the National Academy of Sciences of the United States of America* 102: 5386-5391

Leite NR, Krogh R, Xu W, Ishida Y, Iulek J, Leal WS, Oliva G (2009) Structure of an odorant-binding protein from the mosquito *Aedes aegypti* suggests a binding pocket covered by a pH-sensitive "Lid". *PLoS One* 4: e8006

Leslie AGW (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Jnt CCP4/ESF-EACBM Newsl Protein Crystallogr* No. 26

Letunic I and Bork P (2016). Interactive Tree Of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* DOI: 10.1093/nar/gkw290

Letunic I and Bork P (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127-8

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research*, 47(W1), W256–W259. DOI:10.1093/nar/gkz239

Li Y, Rezaei MA, Li C, Li X, Wu D. DeepAtom: a framework for protein-ligand binding affinity prediction. arXiv. 2019, preprintarXiv:1912.00318. https://arxiv.org/pdf/1912.00318.pdf

Liu Zhihai, Li Yan, Han Li, Li Jie, Liu Jie, Zhao Zhixiong, Nie Wei, Liu Yuchen, and Wang Renxiao (2014). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31 (3):405–12. DOI: 10.1093/bioinformatics/btu626

Loebel, D., Scaloni, A., Paolini, S., Fini, C., Ferrara, L., Breer, H., & Pelosi, P. (2000). Cloning, post-translational modifications, heterologous expression and ligand-binding of boar salivary lipocalin. *The Biochemical Journal*, 350 Pt 2 (Pt 2), 369–379. DOI:10.1042/bj3500369

Manoharan M, Sankar K, Offmann B, Ramanathan S. Association of Putative Members to Family of Mosquito Odorant Binding Proteins: Scoring Scheme Using Fuzzy Functional Templates and Cys Residue Positions. Bioinformatics and Biology Insights. January 2013. doi:10.4137/BBI.S11096

Manoharan M, Ng Fuk Chong M, Vaïtinadapoulé A, Frumence E, Sowdhamini R, Offmann B. (2013). Comparative genomics of odorant binding proteins in *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*. *Genome Biol Evol*. 5(1):163-180. DOI:10.1093/gbe/evs131

Mao Y, Xu X, Xu W, Ishida Y, Leal WS, Ames JB, Clardy J (2010) Crystal and solution structures of an odorant-binding protein from the southern house mosquito complexed with an oviposition pheromone. *Proc Natl Acad Sci U S A* 107: 19102-7

Marchese, S., Pes, D., Scaloni, A., Carbone, V., & Pelosi, P. (1998). Lipocalins of boar salivary glands binding odours and pheromones. *European Journal of Biochemistry*, 252(3), 563–568. DOI:10.1046/j.1432-1327.1998.2520563.x

Stepniewska-Dziubinska Marta M, Zielenkiewicz Piotr, and Siedlecki Pawel (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34 (21):3666–3674, 05. DOI: /10.1093/bioinformatics/bty374

Mastrogiacomo, R., D'Ambrosio, C., Niccolini, A., Serra, A., Gazzano, A., Scaloni, A., & Pelosi, P. (2014). An odorant-binding protein is abundantly expressed in the nose and in the seminal fluid of the rabbit. *PloS One*, 9(11), e111932. DOI:10.1371/journal.pone.0111932

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

Monaco, H. L., Rizzi, M., & Coda, A. (1995). Structure of a complex of two plasma proteins: transthyretin and retinol-binding protein. Science (New York, N.Y.), 268(5213), 1039–1041. DOI:10.1016/S0167-4838(00)00148-5

Moritz, RF et al. (2010). Invasion of the Dwarf Honeybee *Apis florea* into the Near East. *Biological Invasions* 12.5: 1093-9.

Motohiro Tomizawa and Izuru Yamamoto (1992). Binding of nicotinoids and the related compounds to the insect nicotinic acetylcholine receptor. *J PESTIC SCI., 17(4)*: 231–236. DOI:10.1584/jpestics.17.4_231

Mulla, M. Y., Tuccori, E., Magliulo, M., Lattanzi, G., Palazzo, G., Persaud, K., & Torsi, L. (2015). Capacitance-modulated transistor detects odorant binding protein chiral interactions. *Nature Communications*, 6, 6010. DOI:10.1038/ncomms7010

Murphy EJ, Booth JC, Davrazou F, Port AM, Jones DN (2013). Interactions of Anopheles gambiae odorant-binding proteins with a human-derived repellent: implications for the mode of action of n,n-diethyl-3-methylbenzamide (DEET). *J Biol Chem*. 288(6):4475-4485. doi:10.1074/jbc.M112.436386

Muthukumar, S, Muniasamy, S, Srinivasan, M, et al. (2018). Evaluation of pheromone-based kit: A noninvasive approach of estrus detection in buffalo. *Reprod Dom Anim*. 53: 1466– 1472. DOI: 10.1111/rda.13281

Muthukumar, S., Rajesh, D., Selvam, R. M., Saibaba, G., Suvaithenamudhan, S., Akbarsha, M. A., Padmanabhan, P., Gulyas, B., & Archunan, G. (2018a). Buffalo nasal odorant-binding protein (bunOBP) and its structural evaluation with putative pheromones. *Scientific Reports*, 8(1), 9323. DOI:10.1038/s41598-018-27550-7

Nakagawa T, Sakurai T, Nishioka T, Touhara K (2005) Insect sex-pheromone signals mediated by specific combinations of olfactory receptors. *Science* 307: 1638-42

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1-6.

Northey, T., Venthur, H., De Biasio, F., Chauviac, F. X., Cole, A., Ribeiro, K. A., Junior, Grossi, G., Falabella, P., Field, L. M., Keep, N. H., & Zhou, J. J. (2016). Crystal Structures and Binding Dynamics of Odorant-Binding Protein 3 from two aphid species *Megoura viciae* and *Nasonovia ribisnigri*. *Scientific Reports*, 6, 24739. DOI:10.1038/srep24739

Ohno, K., Kawasaki, Y., Kubo, T., & Tohyama, M. (1996). Differential expression of odorant-binding protein genes in rat nasal glands: implications for odorant-binding proteinII as a possible pheromone transporter. *Neuroscience*, 71(2), 355–366. DOI:10.1016/0306-4522(95)00454-8

Oldroyd B, and Nanork P (2009). Conservation of Asian Honey Bees. *Apidologie* 40.3: 296–312.

Olsson SB, Hansson BS (2013). Electroantennogram and single sensillum recording in insect antennae. *Methods Mol Biol.* 1068:157-77. DOI: 10.1007/978-1-62703-619-1_11.

Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* 55: 383-394

Otis GW (1991). Revised distribution of three recently recognised species of honey bees Asia. *Honeybee Sci* 15:167-170.

Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34:i821-i829. doi:10.1093/bioinformatics/bty593.

Page RE, and Robinson G (1991). The genetics of division of labour in honey bee colonies. *Adv insect physiol* 23: 117-169.

Painter J, Merritt EA (2006) TLSMD web server for the generation of multi-group TLS models. *Journal of Applied Crystallography* 39: 109-111

Panagiotis L Kastritis and Alexandre MJJ Bonvin (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface.*, *10 (79)*:20120835. DOI: 10.1098/rsif.2012.0835

Pantsar, Tatu; Poso, Antti. Binding Affinity via Docking: Fact and Fiction. *Molecules*, **2018**, *23, 8:* 1899. DOI: 10.3390/molecules23081899

Parkinson JS. (1993). Signal transduction schemes of bacteria. Cell. 73:857–871.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst*, *32*, 8024–35.

Paul, S. S., Mandal, A. B., & Pathak, N. N. (2002). Feeding standards for lactating riverine buffaloes in tropical conditions. *The Journal of dairy research*, *69*(2), 173–180.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2: 559-572

Pelletier J, Guidolin A, Syed Z, Cornel AJ, Leal WS (2010) Knockdown of a mosquito odorant-binding protein involved in the sensitive detection of oviposition attractants. *J Chem Ecol* 36: 245-8

Pelosi C, Saitta F, Wurm FR, Fessas D, Tine MR, Duce C (2019) Thermodynamic stability of myoglobin-poly(ethylene glycol) bioconjugates: A calorimetric study. *Thermochimica Acta* 671: 26-31

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

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

Pelosi P, Maida R (1990) Odorant-Binding Proteins in Vertebrates and Insects - Similarities and Possible Common Function. *Chemical Senses* 15: 205-215

Pelosi P, Zhou JJ, Ban LP, Calvello M (2006) Soluble proteins in insect chemical communication. *Cell Mol Life Sci* 63: 1658-76

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

Pelosi, P. (1994). Odorant-binding proteins. *Critical reviews in biochemistry and molecular biology*, *29*(3), 199-228.

Pelosi, P., Mastrogiacomo, R., Iovinella, I., Tuccori, E., & Persaud, K. C. (2014). Structure and biotechnological applications of odorant-binding proteins. *Applied Microbiology and Biotechnology*, 98(1), 61–70. DOI:10.1007/s00253-013-5383-y

Pelosi, P., Zhou, J. J., Ban, L. P., & Calvello, M. (2006). Soluble proteins in insect chemical communication. *Cellular and molecular life sciences: CMLS*, 63(14), 1658–1676. DOI:10.1007/s00018-005-5607-0

Pelosi, P., Zhu, J., & Knoll, W. (2018). Odorant-Binding Proteins as Sensing Elements for Odour Monitoring. *Sensors (Basel, Switzerland)*, 18(10), 3248. DOI:10.3390/s18103248

Pelosi P, Pisanelli AM, Baldaccini NE, Gagliardo A (1981) Binding of 3H-2-isobutyl-3-methoxypyrazine to cow olfactory mucosa. *Chem Senses* 6: 77–85.

Pelosi P, Baldaccini NE, Pisanelli AM (1982) Identification of a specific olfactory receptor for 2-isobutyl-3-methoxypyrazine. *Biochem J* 201: 245–248.

Pelosi, P. & Maida, R. **(**1990**)** Odorant-binding proteins in vertebrates and insects: similarities and possible common function. *Chemical Senses*, 15, 205 – 215.

Pevsner J, Hou V, Snowman AM, Snyder SH (1990) Odorant-binding protein characterization of ligand binding. *J Biol Chem* 265: 6118–6125.

Peralta, O. A., Pearson, R. E., & Nebel, R. L. (2005). Comparison of three estrus detection systems during summer in a large commercial dairy herd. *Animal reproduction science*, 87(1-2), 59–72. DOI:10.1016/j.anireprosci.2004.10.003

Pes, D., Mameli, M., Andreini, I., Krieger, J., Weber, M., Breer, H., & Pelosi, P. (1998). Cloning and expression of odorant-binding proteins Ia and Ib from mouse nasal tissue. *Gene*, 212(1), 49–55. DOI:10.1016/S0378-1119(98)00131-0

Peter W Rose, Bojan Beran, Chunxiao Bi, Wolfgang F Bluhm, Dimitris Dimitropoulos, David S Goodsell, Andreas Prlic, Martha Quesada, Gregory B Quinn, John D Westbrook, et al. (2010). The RCSB protein data bank: redesigned website and web services. *Nucleic acids Res.*, 39: D392–D401. DOI:10.1093/nar/gkq1021

Petersen TN, Brunak S, von Heijne G and Nielsen H (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8:785:786. DOI: 10.1038/nmeth.1701

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF Chimera -- a visualization system for exploratory research and analysis. *J Comp Chem.*, **2004**, *25 (13)*:1605–1612. DOI: 10.1002/jcc.20084.

Petrulis A. (2013). Chemosignals, hormones and mammalian reproduction. *Hormones and Behavior*, 63(5), 723–741. DOI:10.1016/j.yhbeh.2013.03.011

Pevsner, J., & Snyder, S. H. (1990). Odorant-binding protein: odorant transport function in the vertebrate nasal epithelium. *Chemical senses*, 15(2), 217-222. DOI:10.1093/chemse/15.2.217

Pokkuluri PR, Londer YY, Duke NE, Long C, and Schiffer M., (2004). Family of Cytochrome c7-Type Proteins from *Geobacter sulfurreducens*: Structure of One Cytochrome c7 at 1.45 Å Resolution. *Biochemistry*, *43* (4), 849-859. DOI: 10.1021/bi0301439

Pokkuluri PR, Londer YY, Duke NE, Pessanha M, Yang X, Orshonsky V, Orshonsky L, Erickson J, Zagyanskiy Y, Salgueiro CA, Schiffer M. (2011). Structure of a novel dodecaheme cytochrome c from *Geobacter sulfurreducens* reveals an extended 12 nm protein with interacting hemes. *J Struct Biol.* 174(1):223-33. DOI: 10.1016/j.jsb.2010.11.022.

Pokkuluri PR, Londer YY, Yang X, et al. (2010). Structural characterization of a family of cytochromes c(7) involved in Fe(III) respiration by *Geobacter sulfurreducens*. *Biochimica et Biophysica Acta.* 1797(2):222-232. DOI: 10.1016/j.bbabio.2009.10.007.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1), W200–W204. DOI:10.1093/nar/gky448

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) In Numerical recipes: The art of scientific computing (Press, C. U., ed.). 521-538, Cambridge, UK.

Qiao H, He X, Schymura D, Ban L, Field L, Dani FR, Michelucci E, Caputo B, della Torre A, Iatrou K, Zhou JJ, Krieger J, Pelosi P. (2011). Cooperative interactions between odorant-binding proteins of Anopheles gambiae. *Cell Mol Life Sci.* 68(10):1799-813. DOI: 10.1007/s00018-010-0539-8.

Rajanarayanan, S., & Archunan, G. (2004). Occurrence of flehmen in male buffaloes (Bubalus bubalis) with special reference to estrus. *Theriogenology*, 61(5), 861–866. DOI:10.1016/j.theriogenology.2003.07.004

Rajkumar, R., Karthikeyan, K., Archunan, G., Huang, P. H., Chen, Y. W., Ng, W. V., & Liao, C. C. (2010). Using mass spectrometry to detect buffalo salivary odorant-binding protein and its post-translational modifications. *Rapid Communications in Mass Spectrometry: RCM*, 24(22), 3248–3254. DOI:10.1002/rcm.4766

Raming, K., Krieger, J. & Breer, H. (1990). Primary structure of a pheromone-binding protein from *Antheraea pernyi*: homologies with other ligand-carrying proteins. *J Comp Physiol B* 160, 503–509. DOI:10.1007/BF00258977

Rastelli G, Del Rio A, Degliesposti G, Sgobba M. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J Comput Chem.*, *31(4)*:797-810. DOI: 10.1002/jcc.21372.

Rekwot, P. I., Ogwu, D., Oyedipe, E. O., & Sekoni, V. O. (2001). The role of pheromones and biostimulation in animal reproduction. *Animal Reproduction Science*, 65(3-4), 157–170. DOI:10.1016/S0378-4320(00)00223-2

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang (2004). The PDBbind Database: collection of binding affinities for Protein-Ligand complexes with Known Three-Dimensional Structures. *J. Med. Chem., 47 (12)*: 2977–80. DOI: 10.1021/jm030580l

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang (2005). The PDBbind Database: Methodologies and Updates. *J. Med. Chem., 48 (12)*:4111–9. DOI: 10.1021/jm048957q

Ricatti, J., Acquasaliente, L., Ribaudo, G., De Filippis, V., Bellini, M., Llovera, R. E., Barollo, S., Pezzani, R., Zagotto, G., Persaud, K. C., & Mucignat-Caretta, C. (2019). Effects of point mutations in the binding pocket of the mouse major urinary protein MUP20 on

ligand affinity and specificity. *Scientific Reports*, 9(1), 300. DOI:10.1038/s41598-018-36391-3.

Rihani, Karen; Ferveur, Jean-François; Briand, Loïc (2021). The 40-Year Mystery of Insect Odorant-Binding Proteins *Biomolecules* 11, 4: 509. DOI:10.3390/biom11040509

Robert, X., & Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research*, 42(Web Server issue), W320–W324. DOI:10.1093/nar/gku316

Roberts E, Eargle J, Wright D, Luthey-Schulten Z (2006) MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 7, 382. DOI:10.1186/1471-2105-7-382

Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J Comp Phys* 23: 327-341

Šali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.

Salomon-Ferrer R, Case DA, Walker RC (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3: 198-210

Samuel Genheden and Ulf Ryde (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities *Expert Opin Drug Discov.*, 2015, 10 (5): 449–461. DOI: 10.1517/17460441.2015.1032936

Sánchez-Gracia A, Vieira FG, Rozas J. (2009). Molecular evolution of the major chemosensory gene families in insects. Heredity (Edinb) 103(3):208-16. DOI: 10.1038/hdy.2009.55.

Sandler BH, Nikonova L, Leal WS, Clardy J (2000). Sexual attraction in the silkworm moth: structure of the pheromone-binding-protein-bombykol complex. *Chem Biol* 7: 143-51

Sankaran, S., Panigrahi, S., & Mallik, S. (2011). Odorant binding protein based biomimetic sensors for detection of alcohols associated with Salmonella contamination in packaged beef. Biosensors & bioelectronics, 26(7), 3103–3109. DOI:10.1016/j.bios.2010.07.122

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. Journal of computer-aided molecular design, 27(3), 221–234. DOI:10.1007/s10822-013-9644-8

Scaloni, A., Paolini, S., Brandazza, A., Fantacci, M., Bottiglieri, C., Marchese, S., Navarrini, A., Fini, C., Ferrara, L., & Pelosi, P. (2001). Purification, cloning and characterisation of odorant- and pheromone-binding proteins from pig nasal epithelium. Cellular and molecular life sciences: CMLS, 58(5-6), 823–834. DOI:10.1007/PL00000903

Schrödinger Release 2019-1: Schrödinger Suite 2019-1 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2019; Impact, Schrödinger, LLC, New York, NY, 2018; Prime, Schrödinger, LLC, New York, NY, 2018.

Schrödinger Release 2019-1: Schrödinger Suite 2019-1 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2019; Impact, Schrödinger, LLC, New York, NY, 2018; Prime, Schrödinger, LLC, New York, NY, 2018.

Schrödinger Release 2019-1: SiteMap, Schrödinger, LLC, New York, NY, 2019.

Schrödinger Release 2019-3**:** Glide, Schrödinger, LLC, New York, NY, 2019.

Schrödinger Release 2019-3: Maestro, Schrödinger, LLC, New York, NY, 2019.

Schrödinger Release 2019-3: Maestro, Schrödinger, LLC, New York, NY, 2019.

Schrödinger Release 2019-3**:** Desmond Molecular Dynamics System, D. E**.** Shaw Research, New York, NY, 2019. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2019.

Schrödinger Release 2019-3**:** SiteMap, Schrödinger, LLC, New York, NY, 2019.

Schrödinger Release 2020-3: QikProp, Schrödinger, LLC, New York, NY, 2020.

Schultze A, Pregitzer P, Walter MF, Woods DF, Marinotti O, Breer H, Krieger J (2013) The co-expression pattern of odorant binding proteins and olfactory receptors identify distinct trichoid sensilla on the antenna of the malaria mosquito *Anopheles gambiae*. *PLoS One* 8: e69412

Seonwoo Min, Byunghan Lee, and Sungroh Yoon (2017). Deep learning in bioinformatics. Brief. Bioinformatics, 18 (5):851–869. DOI: 10.1093/bib/bbw068

Sergey Ioffe and Christian Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* preprint, arXiv:1502.03167.

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., & Uchimaya, M. (2007). Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. Journal of computer-aided molecular design, 21(12), 681–691. DOI:10.1007/s10822-007-9133-z

Shen, Y., Maupetit, J., Derreumaux, P., & Tufféry, P. (2014). Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. Journal of chemical theory and computation, 10(10), 4745–4758. DOI:10.1021/ct500592m

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 11;7: 539. DOI: 10.1038/msb.2011.75

Skjærven L, Yao XQ, Scarabelli G, Grant BJ (2014) Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics* 10; 15(1):399. DOI: 10.1186/s12859-014-0399-6.

Slater GSC, Birney E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.

Sleiman MH, Csonka R, Arbez-Gindre C, Heropoulos GA, Calogeropoulou T, Signorelli M, Schiraldi A, Steele BR, Fessas D, Micha-Screttas M (2015) Binding and stabilisation effects of glycodendritic compounds with peanut agglutinin. *Int J Biol Macromol* 80: 692-701

Smadja, C., Butlin, R. (2009). On the scent of speciation: the chemosensory system and its role in premating isolation.*Heredity*, 102, 77–97. DOI:10.1038/hdy.2008.55

Spinelli, S., Vincent, F., Pelosi, P., Tegoni, M., & Cambillau, C. (2002). Boar salivary lipocalin. Three-dimensional X-ray structure and androsterol/androstenone docking simulations. *European journal of biochemistry*, 269(10), 2449–2456. DOI:10.1046/j.1432-1033.2002.02901.x

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–90.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–3.

Altschul Stephen F, Madden Thomas L, Schäffer Alejandro A, Zhang Jinghui, Zhang Zheng, Miller Webb, and Lipman David J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, 25 (17): 3389–3402. DOI:10.1093/nar/25.17.3389

Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics. 2018;34:3666-3674. doi:10.1093/bioinformatics/bty374

Stone, J, Walker, J. (1995). Plant protein kinase families and signal transduction. *Plant Physiology*, 108(2):451–457. DOI:10.1104/pp.108.2.451

Stopková, R., Zdráhal, Z., Ryba, S., Sedo, O., Sandera, M., & Stopka, P. (2010). Novel OBP genes similar to hamster Aphrodisin in the bank vole, Myodes glareolus. BMC genomics, 11, 45. DOI:10.1186/1471-2164-11-45

Su CY, Menuz K, Carlson JR. (2009). Olfactory Perception: Receptors, Cells, and Circuits. Cell. 139:45–59.

Sukhwal A, Sowdhamini R (2015) PPCheck: A Webserver for the Quantitative Analysis of Protein-Protein Interfaces and Prediction of Residue Hotspots. *Bioinformatics and Biology Insights* 9: 141-151

Suwannapong G, Seanbualuang P & Benbow ME (2011). Using sensillum potential analysis to quantify pheromone sensing of the antennal sensilla of ***Apis florea Fabricius*** (1787), foragers and guards. *Journal of Asia-Pacific Entomology*, 14(1), 7–10.

Takeshi Kawabata (2010). Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* 78 (5):1195–1211. DOI: 10.1002/prot.22639

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

Tegoni, M., Pelosi, P., Vincent, F., Spinelli, S., Campanacci, V., Grolli, S., Ramoni, R., & Cambillau, C. (2000). Mammalian odorant binding proteins. *Biochimica et biophysica acta*, 1482(1-2), 229–240. DOI:10.1016/S0167-4838(00)00167-9

Tegoni, M., Ramoni, R., Bignetti, E., Spinelli, S., & Cambillau, C. (1996). Domain swapping creates a third putative combining site in bovine odorant binding protein dimer. *Nature Structural Biology*, 3(10), 863-867. DOI:10.1038/nsb1096-863

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47: D506-515.

Thévenet, P., Shen, Y., Maupetit, J., Guyon, F., Derreumaux, P., & Tufféry, P. (2012). PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Research*, 40(Web Server issue), W288–W293. DOI:10.1093/nar/gks419

Thode AB, Kruse SW, Nix JC, Jones DN (2008) The role of multiple hydrogen-bonding groups in specific alcohol binding sites in proteins: insights from structural studies of LUSH. *J Mol Biol* 376: 1360-76

Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester (2017). Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal Cheminfom.*, 9 (1):1– 14. DOI: 10.1186/s13321-017-0209-z

Torre A, Iatrou K, Zhou JJ, Krieger J, Pelosi P (2011) Cooperative interactions between odorant-binding proteins of Anopheles gambiae. *Cell Mol Life Sci* 68: 1799-813

Tsitsanou KE, Thireou T, Drakou CE, Koussis K, Keramioti MV, Leonidas DD, Eliopoulos E, Iatrou K, Zographos SE (2012) Anopheles gambiae odorant binding protein crystal complex with the synthetic repellent DEET: implications for structure-based design of novel mosquito repellents. *Cell Mol Life Sci* 69: 283-97

Tsui V, Case DA (2000) Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* 56: 275-291

Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. *Journal of Applied Crystallography* 30: 1022-1025

Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26: 1701-1718

van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, van Dijk M, de Vries SJ, Bonvin AMJJ (2016) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* 428: 720-725

vanAalten DMF, Conn DA, deGroot BL, Berendsen HJC, Findlay JBC, Amadei A (1997) Protein dynamics derived from clusters of crystal structures. *Biophys J* 73: 2891-2896

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

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

Vincent, F., Spinelli, S., Ramoni, R., Grolli, S., Pelosi, P., Cambillau, C., & Tegoni, M. (2000). Complexes of porcine odorant binding protein with odorant molecules belonging to different chemical classes. *Journal of Molecular Biology*, 300(1), 127-139. DOI:10.1006/jmbi.2000.3820

Vogt GR, Große-Wilde E, Zhou JJ (2015). The Lepidoptera Odorant Binding Protein gene family: Gene gain and loss within the GOBP/PBP complex of moths and butterflies, *Insect Biochemistry and Molecular Biology* 62: 142-153. DOI:10.1016/j.ibmb.2015.03.003.

Vogt, R., Riddiford, L. (1981). Pheromone binding and inactivation by moth antennae. *Nature* **293,** 161–163. DOI:10.1038/293161a0

Vogt R, Rogers M, Franco M, Sun M (2002). 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*; 205: 719-744

Von Hippel, P., & Berg, O. (1986). On the Specificity of DNA--Protein Interactions. *Proceedings of the National Academy of Sciences of the United States of America, 83*(6), 1608-1612.

Vosshall LB, Stocker RF. (2007). Molecular Architecture of Smell and Taste in *Drosophila*. *Annu. Rev. Neurosci.* 30:505–533.

Wolf, A., Kirschner, K.N. (2013). Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain. *J Mol Model* **19,** 539–549. DOI:10.1007/s00894-012-1563-4

Zhou, J.-J., He, X.-L., Pickett, J.A. and Field, L.M. (2008), Identification of odorant-binding proteins of the yellow fever mosquito *Aedes aegypti*: genome annotation and comparative analyses. *Insect Molecular Biology*, 17: 147-163. DOI:10.1111/j.1365-2583.2007.00789.x

Ziemba, B.P., Murphy, E.J., Edlin, H.T. and Jones, D.N.M. (2013), A novel mechanism of ligand binding and release in the odorant binding protein 20 from the malaria mosquito *Anopheles gambiae*. *Protein Science*, 22: 11-21. DOI:10.1002/pro.2179

Zhou, J.J., (2010). Odorant-binding proteins in insects. In: Gerald, L. (Ed.), Vitamins & Hormones. Academic Press, Burlington, 241–272.

Appendices

Appendix A

List of Ligand features calculated for DEELIG

Following properties were calculated using PADEL:

- Basic Group Count
- Carbon Type
- Hybridization Ratio
- Manhold LogP (The Ratio of carbon to hetero atoms)
- Number of Aromatic bonds
- MACCS Key
- Klehotaroth fingerprints (Types and Counts)
- AtomPair2D fingerprints (Types and Counts)

Following ADMET properties were calculated using PADEL:

- *donorHB*
- *accptHB*
- *Constitutional (Electronegativity)*
- *rotatableBondCounts (#ringatoms)*
- *RuleofFive*
- *VABC (Volume)*
- *Weight (mol_MW)*

Following ADMET properties were calculated using QikProp:

- Amine
- Amidine
- Acid
- Amide
- Rotor
- rtvFG (reactive functional groups)
- mol_MW, dipole
- Volume

- donorHB
- accptHB
- QPpolrz (polarizability)
- SASA
    - SASA (probe of 1.4A)
    - FOSA (hydrophobic component of SASA)
    - FISA (hydrophilic component of SASA)
    - PISA (pi of SASA)
    - WPSA (polar of SASA)
    - SAFluorine
    - SAamideO
- **Partition coefficients =>**
    - QPlogPC16
    - QPlogPoct
    - QPlogPw
    - QPlogPo/w
- CIQPlogS (**Conformation indie aqueous solubility**)
- IP (ev) (ionization potential)
- EA (eV) (electron affinity)
- #metab (likely metabolic reactions)
- PSA (van der waals SA of polar N and O atoms)
- #NandO, #ringatoms (number of atoms in rings)
- #in34 (number of atoms in 3 or 4 membered rings)
- #in56 (number of atoms in 5 or 6 membered rings)
- #noncon (ring atoms cannot form conjugated aromatic bonds)
- #nonHatm (heavy atoms- nonhydrogen atoms)
- RuleOfThree
- RuleOfFive (lipinski violations)
- QPlogKhsa (binding to human serum albumin)
- PercentHuman-OralAbsorption

- Globular nature index

Appendix B

Clustering as a measure of ligand diversity in DEELIG

Ligand clusters as a measure of ligand diversity in DEELIG. Hierarchical clustering of unique ligands used in training DEELIG was performed on the basis of physicochemical properties. Each of the 125 clusters is represented by a cluster ID and corresponding ligands as members. These ligands are part of training DEELIG, and have been represented by ligand ID from RCSB PDB. Cluster centroid (in **bold**) obtained was assigned as the cluster representative.

Table B_1: Ligand clusters as a measure of ligand diversity in DEELIG

Cluster ID	Cluster members
1	COA,4CO,MYA,RFC,CAO,**ACO**,MC4,0ET,1HA,UCA,CAA,MLC,WCA,5F9,IVC
2	**NAD**,NAP,A3D,NHD,DND,ZID,NAD,8ID
3	NDP, **NAI**
4	ADP,ATP,A4P,ACP,APC,APR,PAX,ANP,A3P,AMP,ODP,A2P,FAO,TXP,AQP,TXD, TM1,M33,12D,CMP,AGS,T5A,V3L,A1R,2BA,A22,ADX,BTX,**AP5**,DTP,2AM,Z5A
5	SAM,SAH,ADN,SON,NOS,SFG,KAA,52H,53H,54H,3BH,M24,A4D,A5D,A6D,A7D,3F5,8OX,A3S,A3T,A3G,DSH,J1A,J1C,J1D,0UM,SAH,3NZ,VRT,SO8,A2H,67D,6CE,**ADN**,75G,KB1
6	NIA,3BK,3FD,3GO,AIR,3P1,TO1,N5P,AW1,0QK,AW2,7CH,SGV,GB8,**7DT**
7	GDP,GNH,2GP,GTP,GNP,MGP,GMP,IMP,**5GP**,DGP,GNG,GSP,01G,C2E,MGO,MGQ,MGV,GTA,GTG,G4P,2GP,GFB,1SY,PCG,1YC,1YD,PCG,4UR,5O8,6CK,C2E,6SW,6SX,6SZ
8	TPG,**MGT**,M7M,M7G
9	DCM,TMP,UFP,8HG,8DA,1AA,2AA,8OG,F01,**CAR**,829,T3P,AO9,NOH,CTN,5UD,5AE,CTP,CC7
10	UMP,UDP,DUP,DUD,UC5,**U5P**,UDM,DUT,UPG,44P,DUR,2TU,UD1,URI,2KH,U3P,Y6W,DUS,6SY
11	G6P,G3P,**A5P**,TA6,G16,X1P,RP5,FDP,FBP,PRP,G1P,F1X,FRU,F1P,1FT,S7P,ALX,T6P,LAO
12	13P,S3P,PAI,137,DXP,MOD,**6PG**,CAP,HPF,SC1,F6F,F19,F9F,ICR,3PG,GG9,R5P,PGH,TPO,PEQ,P6T,OPE,TD4,PGA,D5X,D8X
13	4IP,I3P,**I5P**,ITP,IHP
14	PIB,**DB4**,PIZ,PBU
15	A2G,CTO,NLC,BDZ,NGA,**NLC**,16G,8VZ
16	A24,1DM,LEC,**147**,HNW,HNV,FVQ,FYZ,J73,CWX,3XO,3X8,4J0,4J9,6Y2

17	BLM,MGS,233,233,BDD,IFM,DAN,SIA,OXZ,SMD,MLD,ODI,XUL,NOY,GOX,TTZ,PAF,DIH,NOK,OAN,NHT,STZ,NGT,GOL,MBG,GC2,LTM,PAU,TRS,MU2,**TDG**,DQT,MQT,MN0,PGZ,GS9,IFL,NGW,70B,YO5,GRX,PK9,AO3,ASC,ETF,423,425,DCL,SLT,X6X,MO0,SNW,7K2,KIF,8MP
18	TD2,5**KS**,5KT,TGZ
19	FCA,ACR,RIP,SMS,ACR,GET,RIP,NOJ,RNS,MFU,CTS,CGB,AB0,B19,AZF,VDM,QPU,KDB,KDR,UNG,GCU,SFU,FHN,XBK,DFU,ENE,END,ZWZ,KDA,TOR,KDO,KO2,TXT,SCR,IHS,INS,H1S,3S6,SIS,**5GO**,EDG,9QZ
20	GIM,PGI,NTZ,GI1,GI2,GI3,GI4,**MVL**
21	DEG,BOG,AFO,FK9,XNS,**KGM**
22	**SUC**,RAF,DQR,FNY
23	GAL,BCD,GAL,LAT,MAN,MAL,GYP,XLM,ARW,MLR,GZL,ACX,MLB,LAT,RGG,**MTT**,LAT,MLR,4CQ,RCD,BCD,LAT
24	PUT,ACA,AMH,GUA,**SIN**,TSA,DMG,HXA,FA1,KAI,SPD,BET,PBE,DHK,CXP,CIT,XAP,H01,TF4,4CS,6CS,LY5,LMR,7HE,GA4,GPF,ABU,SF8,SAR,GIX,5AB,4U7,FB6,03H,BUA,MZT,RUJ,G8M,QIC,164,GA3,XA2,X7A,011,4XW,522,A9H,IPM,QZQ,21G
25	BES,140,MCO,GSH,HAL,DOR,GTX,**PCA**,LEH,YBY,GDS,MHI,F42,QEG,GHE,N8P,LZ6,2VE,2VD
26	**FOL**,MTX,D16,MS1,F89,DHF,C2F,GHW,0J2,0J4,0J5,Z25,04J,MT1
27	FLA,ABH,PSQ,S2C,ORN,CBH,QUS,ABH,4TB,DGL,MLZ,MLY,DSN,6HN,NNH,HAR,LN7,LN6,**API**,3ZA,2W2,DAL,RPI,6DB,TIH
28	AMQ,T44,T3,**IYR**,4ZK,E42,DPN
29	ACD,OLA,STE,VCA,PLM,**DAO**,PAM,DKA,OCA,MYR,DCR
30	**BTN**,BTN,IMI,IMI,KYT,DTB,ZOF,B1R
31	**BS5**,44K,44M,590,44L,594,44N,44O,44Q,44R
32	5B3,**1TM**,KTJ,KTK,1CQ,69M,69W,69S,69R,69U
33	BPY,BP3,**IPH**,3CH,3FA,ANF,43M,BZX,RCO,EMO,CAQ,IOL,BML,4CH,FPN,03V,1NP,BP7,PV2,PV8,PV0,ERH,HQE,6X8,6JD
34	STL,BHQ,2OH,TOH,OHT,ETY,G50,NAR,2DL,C0Y,AEF,AEH,AED,PFL,DIE,IP0,2MY,RE2,**JZ0**,JZ1,JZ2,JZ4,261,JTH,JWS,AZB,MLO,EOL,LDP,N7I,HC4,2LP,1G4,1OM,MCT,EKZ,DFL,XX9,LNR,NVS
35	QUE,MYC,MYU,GEN,SL0,AGI,57D,KMP,**LU2**,QCT,FSE,7LU,6B5
36	STI,5MP,ZY5,ZY7,XDK,3AP,NCT,NIL,706,3QO,0QA,3D4,**3D5**,3D2,RH2,3D8,TQL,LKB,7M2,8HH
37	6XC,78J,**YEJ**,YD3
38	LY2,7NI,PY1,H52,M77,KSA,L08,FRZ,LY4,BI1,WAI,CL3,IQB,IYZ,SB2,7NH,3,P34,SS3,SS4,P83,VGH,646,GD9,XJ1,985,4F8,4FY,YKI,DEQ,CCK,SX8,HC7,HC6,STU,G97,1SQ,3YY,Q6W,M61,3RH,3RU,DIW,03M,B29,913,Z5P,TB7,**LZ1**,8DQ,C1I,0J0,IRG,GG5,LM3,TDH,YS4,1ER,AJ4,AJ5,AJ6,AJ7,18F,1M1,1M0,17X,1RQ,X6K,SB4,2GQ,40B,31V,30M,38G,3C5,X6W,BK9,3WH,4L2,92O,CWC,M1W,BMF,539,5D1,5RO,5WS,O82,O83,O80,RX8,68M,68J,68K,6BS,6B3,6DP,6U5,7W4,7VP,88J,86J,8DV,DQX,BZE
39	TK4,2AN,ANC,DSO,MNS,C4H,11D,**5NS**,FNA
40	**COU**,8MO,8CM,XCG,EVD,03T,SE2,AKH,BZF
41	FIP,PRL,IOP,3IB,TSS,SRO,7L4,11S,5H1,G3E,ES1,ES5,ET,0QY,MQN,39R,FK1,AO,DTR,5SG,ITW,A7W,6VD,14O,7W1,**IND**
42	1PP,1AU,**1AW**,1BU,R48,SS6,R39,Z84,Z83,B96,DG7,K9Y,KAO

43	BZI,HAN,VGF,**AX7**,ES7,ES9,ES0,VGD,ET2,1M2,CPZ,6RQ,7WA
44	CKR,2WU,**2YM**,2YO,SAQ,SQO
45	PH2,LG3,7IG,LGA,A13,V15,LZJ,LZK,TOP,C0P,B2K,B2L,B2T,B2X,B56,EV0,3RP,VHA,VHC,VHE,WHA,MUJ,CP6,TXW,43U,43S,XOJ,6GH,**6GF**,6GM,7WG
46	CK2,18K,**19K**,16K,20K
47	1N9,R78,627,P4T,GUI,AM0,**AM9**,HT1,585,27A,HI0,NP7,RLI,IA9,R78,2TA,38O,C5N,5XQ,8X7,P48,7ZF,7ZL,7ZO,7Z6, 1E9
48	JIN,3FN,FLS,**FLW**,P17,96M
49	5VY,5VZ,5W0,**5W1**,5W2
50	BER,AVX,833,BBY,SAU,**CTI**,SGE,1GH,1H2,2K7,2M2,2LO,36Z,37N,3P2,4LD,53W,5SW,5U6
51	FMM,AQZ,PFH,CX4,**DXQ**,E67,E72,DB8,IRE,LEV,DTQ,0N5,1K0,TZN,3ZC,5GD,88Z,XZN,93J,H8H,9HJ,9HG
52	M0Y,**M0Z**,1YZ,1E0A
53	541,ML1,ML2,V5X,IMN,**MYI**,BZ3,A8D,FG2,FG3,FG4,FG5,FG6,2B5,MS0,QI9,5WD
54	DRJ,XCX,**ABV**,36R,94W,6KT,7FU
55	FD1,FD2,FD3,FD4,TL3,TL4,**BEN**,PNT,BRN,PBZ,JLZ,AZ3,NTN
56	50U,45U,46U,49U,12U,10U,**11U**,13U
57	6MG,7AF,6NN,**6XO**,79F,79E,7GR,7K0
58	MK1,XN3,XN1,**XN2**
59	K57,JE2,075,O33,**006**,JZP,JZQ
60	FRA,SPB,Y27,SHH,1CM,CM6,CMW,CM9,RT8,NBB,N4E,AJ1,VAB,WVH,T9H,0S2,0S3,**20N**,1EB,25S,4YO,5QY,5QX,IQ5,N91,7H1,6DK,8NB

61	88S,88W,**88V**,88N,W7E
62	1B0,**1ED**,1G7,1G9,1GB,1B0,7HG,78E,78X,838
63	ROC,FAF,RIT,FRG,BLN,HAE,ACM,DR7,RO1,EGD,MGX,AB1,D1R,1UN,E64,2NC,ARX,C1P,REM,44U,BB2,3N6,RSZ,613,NH5,08Y,091,0UE,15U,FCE,KI1,1RD,7AW,OP1,38Z,3G7,3D0,N6L,MRZ,FY8,58W,5T0,5T4,841,5TD,5TO,5WZ,6C5,6C4,**EBU**,8T5,7HC,92A
64	**HCI**,PRA,PEA,PBN,AHC,MPP,PPT,5PV,AEG,B40,B41,MBN,PXY,IBP,I2E,88R,WJ1,1WE,PYJ,3H0,N4B,I4B,OXE
65	**ABN**,PNZ,PZM,C2B,G00,VXU,VXQ,SW2,91B,7Z3,K82,54E,86L,OQC
66	1DK,12G,1DL,11Y,1D5,**1D6**,1D8,1D9,1DE
67	3HI,4HI,5HI,6HI,7HI,**882**,8HI,9HI
68	0L9,LF2,LF0,**4BI**,LF8
69	3C4,DHY,4HP,5ON,5UH,5V5,5V6,**IO2**,YI6
70	DIF,570,AIC,EOT,CCR,YPA,NLA,DPD,307,GAS,388,393,2C2,LK2,T4B,TOL,LDT,LKM,GUS,BYH,0CR,LIZ,28P,D30,EAA,B72,4D9,**G24**,DOY,3RM,GW0,EK8,2B4,6LX,0JD,H1L,1E5,1EK,8PR,1YE,29V,2YB,RF4,XTK,UET,4TX,42O,IEE,L6I,WY1,SBK,70R,VT1

71	700,ZEN,434,1N1,5CL,YAN,230,NZA,CRP,LZE,XBV,WSH,EAM,KSF,RO6,BDO,**BDJ**,JQ1,08H,08K,0C7,SE5,52C,4VA,1LH,OCZ,0LM,0LL,0LK,0R3,0S6,0Y7,1C1,I29,I18,ZHK,1CY,1WQ,20J,29Q,2KD,2LW,2SW,2TZ,2ZW,36U,390,33R,ZKW,19H,3UD,4B7,4NJ,4NX,4T4,NUT,CKA,5D2,CZE,5OD,5XL,5XK,YIE,XMS,XYY,0PL,NH6,DL3,RRQ,9GD,7AV
72	5O5,J4K,6J5,**YIH**
73	YH5,X6L,**YH1**,YH2,YH6,YH3,YH4,5RU,5RV,5RW,5RX,5RY,5RZ
74	HPA,IMU,IRP,9HX,URC,DIH,IMH,MTI,HHR,HA1,PIR,9D9,**229**,22A,B55,GSY,IM5,5JL
75	M1A,9PP,**ADE**,TDI,ZEA,DF9,BIG,ES4,3MA,4CT,QMQ,ADK,U85,EOI,DX4,YH7,NBR
76	FMN,FAD,D4T,ORO,**RBF**,CFF,PNX,TP6,TSF,JCL,47D,POY,BAU,4WF,HNR,N6A,1VY,39O,48I,F49,TDR,HMU,FYU,5NM,5XF,L66,6YA,7M8
77	**UBA**,UBC,UB1,UBF,UBE
78	PPG,ORX,PE1,PMP,**PDD**,PLP,PLS,UEG
79	TDP,VIB,**TPP**,TPS,TPU,D7K,8PA,NFM,26G,0YN,XX8,2VY
80	FPP,DED,IPE,DP6,GRG,FM0,3PO,H6P,2JA,**GPP**,4LR
81	GER,YFE,CTV,YFD,YFB,2TX,**SQL**,JH3
82	**1O8**,2VR,2VP,2VZ,2W0,3RB
83	**MAE**,ISC,PEP,CCU,OAA,COI,PYR,2KT,XI7,AKG,OGA,P47,A8S,D1X,KIV,OXM,2H8,KMT,3SW,3T2,5O6,5OO,6ZX,OXD,OTA
84	NOV,TFP,BZT,BT3,BPP,BEE,FLU,BRL,RWF,RNP,CBN,DMP,EQU,NH1,OPB,OI1,667,RRC,26D,T1D,HQL,LG4,FL1,UN9,37A,LIX,JJ3,TPV,1PC,TDZ,D26,6AN,WOE,849,YIS,YIW,YIX,575,3F1,I5I,IM4,TH4,LN1,25P,XNC,S45,IX6,J0Z,PP0,D8Y,3TO,3KR,Z80,MS7,LX8,MKY,RSV,09H,34M,0B3,0B4,N1Y,N2Y,0MB,V6F,PW1,NIH,4DC,10K,09K,TZM,0LU,4HV,1EP,1JR,1L7,KKN,1QK,KLV,04G,1MV,YL3,LOC,2CK,24W,2J5,2J6,2J7,2L1,2L2,31W,09L,**3A8**,3CR,9X3,HXY,WEF,WF4,3VR,3W1,3XV,43L,4OL,XOC,UL7,ETU,K6K,XEB,JH5,536,XZ8,5FF,5FK,UO1,5KH,73K,GWM,VQ7,6F9,I74,77X,XHA,2AP,8HZ,8NN,2OW,79A,7Q1,7QV,7QS,7R7,7R4,7R1,8FD,9AA
85	**1QE**,H49,1QF,6E5A,H36
86	TPF,ZST,LS5,BMU,BNZ,1AN,DMZ,ATC,FID,PYM,DFN,TSN,444,4TZ,4TR,198,FNZ,NQ,FLV,KLN,1MR,SBI,A51,RO5,II2,II4,MIY,B49,XIN,G93,BSU,4BM,4DM,BAX,53A,469,AVR,0LI,HIZ,3TR,Z89,CEL,VOR,OJ8,SU9,QID,919,TDU,42Q,CVI,VWS,54F,93G,L5V,CQE,3PF,0NS,60K,62K,G9L,TQU,TZF,0X3,0XB,4GU,12R,IQK,X2N,1HR,1L5,HRC,2C9,2CQ,2QJ,2RR,Q7M,RV1,VJP,3M6,3R7,3VQ,3VW,3VX,41O,ZBD,3IT,1YN,SFQ,4OQ,4OR,4QF,LQ5,UFV,**EB0**,BEA,5TQ,5WC,LVY,7FF,68L,GV0,6HO,6Q8,P4P,6UY,78Y,PKU,LS8,2BG,9EW,JR5,7OV,927,921,0BP,DEN
87	P86,NXG,P74,**P51**,P96,KMN
88	TCL,4BC,43P,PCQ,226,2GG,2GJ,**YI0**,YI5,6QM,BZ8,FT1,FT2,9UN,50Q,XKL,PZB,PCI,T6C,B1T,5JC,FGZ
89	NPE,TAA,9AR,AAH,TNF,PGG,PNF,4NM,CLM,N11,BFF,O73,C56,**NPO**,CNI,2ZM,36V,4JT,4NC,6NT,6J1,80R
90	REA,ETR,RTL,**RET**,RNE

91	**TLM**,TL6,TLG,7RD,TLH
92	GDM,FKA,FK5,RAP,HYF,TMH,S1A,TG1,SMA,BID,VD5,BC6,BC2,**NRB**,1D4,1D2,GYN,GDM,CW1,CY9,BGV,ZGA,Z18,QLE,SM8,0V4,BLD,2UO,SNE,3V4,MY8,6K9
93	AZI,FMS,1MC,BCT,BR,DSS,BUQ,NCH,PRZ,R18,PBR,TBP,MTD,MSF,MTG,MMQ,DHM,2HP,PPF,CHT,HBR,MPJ,MOI,PP2,IPZ,NO3,CCE,715,KAR,MSM,TP5,MLK,CEJ,AA7,BVC,VX6,LIA,DI9,HI6,OBI,CHH,SCK,AT3,ETM,2PN,89I,QIN,MR3,TSZ,NMU,ACH,486,CMO,MES,EMH,FLC,MXD,B2J,BJM,DRL,SX3,SX5,**SX6**,9OD,MI1,ICF,79Z,B10,32A,18,20,JZ9,J1Z,PO4,717,AJD,AKD,KCS,XAV,3MC,614,EML,JKZ,ZG1,MPV,3M2,94M,PC,SW1,K9S,513,3MQ,TR7,VMY,7SB,MEU,L2D,L2M,9LI,0OC,ZMA,2N6,0TM,0XT,0RT,10N,POP,SO4,1C6,CE3,KOV,CFX,CLS,CNL,1WG,2C1,TM7,2FQ,EJ4,KK3,39U,RXT,K25,K17,FQI,ACT,3WG,25T,TMO,49J,0A9,RHU,RZH,TAU,50V,56H,5FJ,5JA,5TY,G45,XDI,6C8,6E8,6EB,6FG,6FR,6K5,PFB,7PF,7DF,8MY,A4G,A1Y,YAM
94	RDC,COC,CLW,BCG,OIN,LOB,P2N,RDA,M1S,7CS,TKT,CWB,MS9,KDH,4ME,96P,**MT6**,T21,6F3,6FX,RHQ,7MX,7MW,4I8
95	PAE,6PR,G88,FOM,GPJ,**34F**,P7I,FM6,FM7,FM8,P2H,476,NI9,PAL
96	HSM,DMI,**DHI**,1MZ,CL6,EKO,ES3,X89,ECN,9PL,AVI,HIC
97	3OL,UNA,MJI,NCW,9HO,AMF,PJ2,U51,EAH,HHG,FEE,S1P,1O2,DPV,U46,LTC,NOO,QUY,QUV,LP5,**DB6**,SDS,TMY,RHL,OHN,LTD,RJW
98	EY5,ERB,ZIT,CTY,**ERY**
99	HW8,HW9,QJ8,Q13,XFK,**XFJ**,1EV,HW1,1EW,Q05,QJ7,1Q6,Q14,S5D,4V4,SKO,7S9,S90,EXI
100	**8ZH**,8ZZ,92Q,90E,90W,90Z,91H,8ZT,90T,91K,8ZK,8ZQ,8ZW
101	HOP,NM1,78P,GK4,410,SNX,VNI,DF3,AGY,**STJ**,KXO,925,N4D,MJG,DE3,746,3P3,085,TJF,9VU,0LN,0RY,20Y,CG4,3E7,3OO,66T,3YX,VFV,3VS,45E,41N,SVR,4KC,4S4,FSU,D7N,SJR,KRX,4QJ,5MK,5MV,5TU,5X0,6P9,6PD,6SR,2LX,8FN,8FK,912,9TW,71T
102	FS8,FJI,HB9,78L,**79Q**
103	BA1,BZQ,AZN,**R69**,R68,R94,R96,PQA,DBE,3M1,TCW,RHN,WL3,537
104	DM2,DM5,**DM1**,DM2
105	EB4,ECA,VBN,5LC,PXJ,**8LC**,EHS
106	**VHD**,XJG,L81,XJX,1RC,PV1,PFT
107	**21Q**,L46,57G,L33,57F,L40,57E,579,56Y,L26,56M,L28,XDM,71X,71Y,99E
108	DXC,DOG,DTX,CHC,DGX,**CHD**,TCH,4OA,DOG,6SB
109	**EST**,ESM,ESL,EZT,3WF
110	DHT,VDX,STR,PDN,ANB,CAM,NOG,DEX,MC9,NDR,MOF,TES,HC3,CLR,TH2,VD3,ANO,C3S,KEM,K2B,3T5,**PLO**,DLC,VDP,C0R
111	X6C,3JG,3JK,3JU,3JF,3JS,3JH,3JO,**3JJ**,6Z3,9BT,9BW,4YY,9BK,9BH,9B8
112	PYC,FBC,26C,205,NCN,ZZA,FUN,NTM,AVD,4P5,1FL,52P,6PC,BZ2,4BY,HJZ,BFH,NJQ,HTJ ,3M3,INO,NNO,P1Y,P4Y,GRI,3MI,4A2,H1I,H0Y,X0B,4WL,ICO,5MI,QNZ,0HO,1F1,0W1,TTY,16V,**6P3**,1E4,1E6,1E7,1EJ,1DZ,1XS,1XT,8HC,1XU,22Y,UL4,ZXM,Y0V,5

	DN,5KR,5WU,61O,60P,65V,MMK,6.00E+07,6ED,6EN,6EO,6EP,QJS,M7O,9FH,9FE,9TZ
113	PHB,BE2,PAB,DBH,DBH,AIN,MHB,DMB,ANN,3HA,SAL,DHB,**BEZ**,AI7,3B4,PHT,GAB,JKE,3HB,0HN,TWO,54G,VJJ
114	394,184,**564**,156,254,TTB,9RA,EQN
115	FLF,OFL,794,968,NFL,P2C,ZZL,ID8,0C0,0C3,0C8,73B,9S3,54S,AU6,68Q,S16,**68W**,A5B,6EC,6J3,C98
116	**PX3**,O2N,O4N,N44,OM3,OM2,ZT0,OM1,ZT2,ZT4,ZSV,NZL,X0P,S3G,4VW,VXO,FYM, K1T
117	**FSB**,FFB,BSB,F2B,F6B,INV,INW,IOA,IOC,IOE,IOF
118	EG1,EG2,EG3,SBR,STB,B22,B30,RYJ,RYV,RYX,**RYY**,RYZ,RZ0,RZ1,RZ5,RZ7,RZ8
119	FB1,BL1,E36,E38,E02,E27,E59,**E49**,E2I,E50,E65,E90,E1E,E1F,E1G,EWW,EWY,EWZ
120	VIA,VDN,5CO,5BO,5IO,**5EO**
121	478,**017**,DJR,KGQ
122	NGH,HS7,**HS1**,HS3,HS4,HS5,HS6,Z79,NHK,JFK,1KY,1KZ
123	FBS,M25,M28,M29,ZYX,FBV,**FB2**,FBW,SKE,M4S,JS7,JDR,J45,J90,J71,J75,J43,9TH,WW3,KR5,3UF,VD9,4J8,4JC,4JE,S2O,5RD,TG4,TG5,949
124	V50,V13,V90,V14,WWX,V1F,**WWO**,3TV,5DU,5EF
125	AZM,AL5,AL6,AL9,AL1,AL4,AL2,AL3,AL7,TPD,ETS,EZL,EVE,EVF,**EVG**,EVH,EVI,EVJ,1YY, BZ1,2EU,XCZ

Appendix C

Taxonomic and domain distribution of dataset

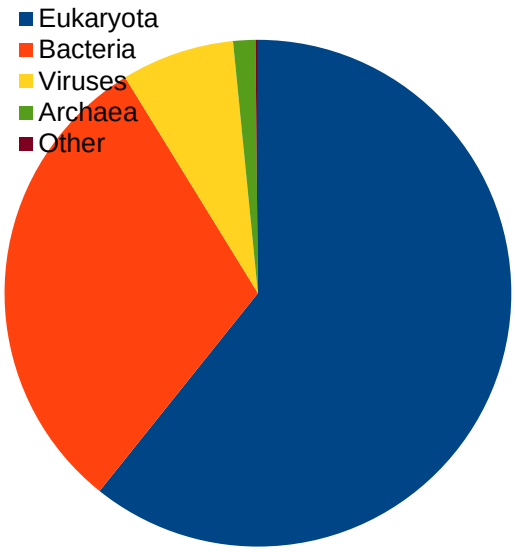


Figure C_1: Pie-chart representing taxonomic distribution of protein structures used in DEELIG

Taxonomy and Species

Table C_1: Taxonomic and species-specific distribution of protein structures used in training DEELIG

Organism	**Number**
Homo sapiens	*1613*
Escherichia coli	*244*
Mycobacterium tuberculosis	*158*
Human immunodeficiency virus spp.	*152*
Bos taurus	*150*
Mus musculus	*139*
Rattus norvegicus	*94*
Saccharomyces cerevisiae	*64*
Pseudomonas aeruginosa	*53*
Bacillus subtilis	*48*
Escherichia virus T4	*40*
Pseudomonas putida	*34*
Streptomyces avidinii	*32*
Thermotoga maritime	*30*
Arabidopsis thaliana	*25*
Staphylococcus aureus	*25*

Oryctolagus cuniculus	*24*
Bacillus megaterium	*21*
Streptococcus pneumoniae	*19*
Lactococcus lactis	*18*
Hepacivirus C	*17*
Human immunodeficiency virus	*16*
Vibrio cholerae	*15*
Helicobacter pylori	*15*
Salmonella enterica	*15*
Gallus gallus	*14*
Lactobacillus casei	*13*
Vibrio harveyi	*13*
Eukaryota	*2400*
Bacteria	*1204*
Viruses	*285*
Archaea	*56*
Other	*6*

Domain Distribution

Table C_2: Frequency of domain distribution of protein structures used for training DEELIG

Domain Name	**Frequency**
p450	*255*
Pkinase	*237*
Carb_anhydrase	*237*
Pkinase_Tyr	*201*
Bromodomain	*191*
Hormone_recep	*141*
RVP	*137*
Neur_chan_LBD	*134*
Avidin	*127*
Ldh_1_N	*92*
Transthyretin	*89*
Lipocalin	*88*
PNP_UDP_1	*81*
HATPase_c	*76*
Gal-bind_lectin	*68*
Trypsin	*68*
VHL	*68*
rve	*67*
HSP70	*65*
DHquinase_II	*56*
NO_synthase	*55*
Lig_chan-Glu_bd	*52*
Pantoate_ligase	*51*
Phage_lysozyme	*48*

V-set	*44*
Glyco_hydro_1	*44*
FimH_man-bind	*44*
cNMP_binding	*41*
HPPK	*39*
SBP_bac_8	*38*
Asp	*38*
Aldo_ket_red	*37*
C1-set	*35*
SWIB	*35*
DAO	*34*
GST_C_3	*34*
adh_short_C2	*34*
DHFR_1	*33*
Ras	*32*
P-II	*32*
show	*31*
GST_N	*31*
PARP	*30*
TetR_N	*29*
P53	*29*
FabA	*29*
ketoacyl-synt	*28*
RdRP_3	*28*
SBP_bac_1	*28*
Flavodoxin_1	*26*
Aminotran_1_2	*26*
PA-IL	*26*
Aldedh	*25*
Jacalin	*25*
2OG-FeII_Oxy_3	*24*
PBP_GOBP	*24*
NNMT_PNMT_TEMT	*24*
Flavodoxin_2	*24*
YgbB	*23*
PH	*23*
DAHP_synth_1	*23*
tRNA-Thr_ED	*23*
EPSP_synthase	*23*
Thymidylat_synt	*22*
Pribosyltran	*22*
Flavin_Reduct	*21*
IF4E	*21*
GLF	*21*
FBPase	*21*
peroxidase	*20*
Pentaxin	*20*

SnoaL_2	*20*
Lactamase_B	*19*
FKBP_C	*19*
Dioxygenase_C	*19*
An_peroxidase	*19*
NDK	*19*
Hist_deacetyl	*18*
TMEM173	*18*
dUTPase	*18*
ECH_1	*18*
AMPK1_CBM	*18*
Glyco_hydro_18	*18*
Aminotran_3	*18*
Filo_VP35	*18*
Peptidase_M10	*18*
LysR_substrate	*17*
PDEase_I	*17*
Peripla_BP_4	*17*
BCDHK_Adom3	*17*
PALP	*17*
Nitroreductase	*17*
DXP_reductoisom	*16*
BPL_C	*16*
DIOX_N	*16*
HMG-CoA_red	*16*
Thiolase_N	*16*
Pyr_redox_2	*16*
NTP_transferase	*16*
E1-E2_ATPase	*16*
4HBT	*16*
F420H2_quin_red	*15*
RbsD_FucU	*15*
NAD_binding_1	*15*
Pterin_bind	*15*
Sugar-bind	*15*
Oxidored_FMN	*15*
Phospholip_A2_1	*15*
S_100	*15*
Peripla_BP_2	*15*
Arginase	*14*
DctP	*14*
polyprenyl_synt	*14*
Acetyltransf_3	*14*
SET	*14*
Iso_dh	*14*
ABC_tran	*14*
SBP_bac_3	*13*
FAD_binding_3	*13*

Y_phosphatase	*13*
Peptidase_M4_C	*13*
LRR_8	*13*
Globin	*12*
COesterase	*12*
ArgoMid	*12*
Pan_kinase	*12*
CBS	*12*
NUDIX	*11*
DSBA	*11*
Enterotoxin_b	*11*
GP120	*11*
Thy1	*11*
Putative_PNPOx	*11*
Glyco_hydro_20	*11*
2-Hacid_dh_C	*11*
DHO_dh	*11*
ADH_zinc_N	*11*
Trp_syntA	*11*
NAGidase	*11*
Amidase_2	*11*
GTP_EFTU	*11*
Phosphorylase	*11*
Macro	*11*
Glyco_transf_6	*11*
Thymidylate_kin	*10*
Synapsin_C	*10*
Choline_kinase	*10*
NAPRTase	*10*
HpcH_HpaI	*10*
Exotox-A_cataly	*10*
FeoB_N	*10*
Redoxin	*10*
Pep_deformylase	*10*
Adenine_glyco	*10*
Antibiotic_NAT	*10*
CTP_transf_like	*10*
AlaDh_PNT_C	*10*
TPP_enzyme_N	*10*
Thia_YuaJ	*10*
Usp	*10*
Mab-21	*10*
PHZA_PHZB	*10*
MoaC	*9*
Rad51	*9*
PGK	*9*
Fungal_lectin	*9*
JmjC	*9*

Ribonuclease	*9*
Peptidase_M1	*9*
Serum_albumin	*9*
Orn_Arg_deC_N	*8*
SQS_PSY	*8*
Cytochrome_B	*8*
Ferritin	*8*
OmpA	*8*
ILEI	*8*
TGT	*8*
GDP_Man_Dehyd	*8*
Semialdhyde_dhC	*8*
RnaseA	*8*
Amino_oxidase	*8*
CPSase_L_D2	*8*
adh_short	*8*
7tm_1	*8*
Bcl-2	*8*
Rep-A_N	*8*
PFK	*8*
F_bP_aldolase	*8*
OTCace_N	*8*
Alpha_L_fucos	*8*
CoA_transf_3	*8*
Adeno_knob	*8*
Citrate_synt	*8*
GMP_PDE_delta	*8*
HTH_Crp_2	*8*
SBP_bac_5	*8*
WD40	*8*
Flu_PB2	*8*
LigT_PEase	*8*
EF-hand_1	*7*
Gag_p24	*7*
Ricin_B_lectin	*7*
PilZ	*7*
ADPrib_exo_Tox	*7*
Laminin_G_1	*7*
Abhydrolase_6	*7*
Bet_v_1	*7*
Peptidase_C14	*7*
Prenyltransf	*7*
PAS_3	*7*
Lectin_legB	*7*
Pro_isomerase	*7*
AMP-binding	*7*
RicinB_lectin_2	*7*
APH	*7*

Polyketide_cyc2	*7*
PrmA	*7*
Sec7	*7*
Peptidase_C1	*7*
MIF	*7*
SBP_bac_6	*7*
F5_F8_type_C	*7*
Glyco_hydro_7	*7*
PdxJ	*6*
Xlink	*6*
Hydrolase	*6*
Cu-oxidase	*6*
DHquinase_I	*6*
Hyd_WA	*6*
MHC_I_3	*6*
Rotamase	*6*
Alpha_kinase	*6*
AA_kinase	*6*
Septin	*6*
SBP_bac_11	*6*
AIRC	*6*
ADH_N	*6*
Glyco_hydro_47	*6*
Rubis-subs-bind	*6*
VWA	*6*
Serpin	*6*
DapB_C	*6*
Peripla_BP_1	*6*
cobW	*6*
Abhydrolase_3	*6*
Peptidase_M17	*6*
PfkB	*6*
Chromo	*6*
Arm	*6*
SusD_RagB	*6*
QRPTase_C	*5*
Lectin_C	*5*
OKR_DC_1	*5*
EAL	*5*
Flavoprotein	*5*
MeaB	*5*
AAA	*5*
PI3Ka	*5*
His_binding	*5*
AAA_26	*5*
ApbA_C	*5*
DMRL_synthase	*5*
Me-amine-dh_L	*5*

BTB	*5*
Tubulin	*5*
CRAL_TRIO	*5*
tRNA-synt_1	*5*
Glyoxalase	*5*
Nitrophorin	*4*
Peptidase_C80	*4*
Tyrosinase	*4*
Neur	*4*
Autoind_bind	*4*
NAD_synthase	*4*
OpuAC	*4*
Flavi_NS5	*4*
HTH_27	*4*
Auxin_BP	*4*
TFR_dimer	*4*
Chorismate_synt	*4*
Enoyl_reductase	*4*
CIMR	*4*
rhaM	*4*
PAF-AH_p_II	*4*
O-FucT	*4*
Hfq	*4*
Spermine_synth	*4*
RNase_H	*4*
WWE	*4*
UDPG_MGDP_dh_N	*4*
ACPS	*4*
PARP_regulatory	*4*
Spin-Ssty	*4*
03/DD/YY	*4*
DPPIV_N	*4*
SIR2	*4*
DOT1	*4*
FAD_binding_4	*4*
GBP	*4*
MBT	*4*
TIM	*4*
Glyco_transf_41	*4*
Methyltransf_4	*4*
RdRP_1	*4*
CoA_binding_2	*4*
DNA_pol3_beta_2	*4*
Abhydrolase_2	*4*
DHH	*4*
Kringle	*4*
Piwi	*4*
UbiA	*4*

tRNA-synt_2	*4*
tRNA-synt_1b	*4*
CBM_48	*4*
Amidinotransf	*4*
DAHP_synth_2	*4*
AraC_binding	*4*
Pept_tRNA_hydro	*4*
L_protein_N	*4*
UPRTase	*4*
RIP	*4*
TrkA_N	*4*
Peripla_BP_3	*4*
Methyltransf_3	*4*
Alk_phosphatase	*4*
Methyltransf_25	*4*
FMN_dh	*4*
Glycos_transf_2	*4*
OCD_Mu_crystall	*4*
IU_nuc_hydro	*4*
ATP-grasp	*4*
Peptidase_S29	*4*
CdAMP_rec	*4*
RuBisCO_large	*4*
NMT	*3*
Hexokinase_2	*3*
LamB	*3*
MMR_HSR1	*3*
OS-D	*3*
Ribonuc_L-PSP	*3*
FAD_binding_7	*3*
MarR	*3*
OTCace	*3*
DNA_ligase_aden	*3*
PadR	*3*
SHMT	*3*
RIO1	*3*
BPL_LplA_LipB	*3*
LpxC	*3*
Alpha-amyl_C2	*3*
Ribonuc_red_lgC	*3*
Bac_luciferase	*3*
APS_kinase	*3*
FtsJ	*3*
Sulfotransfer_1	*3*
Acetyltransf_14	*3*
PanZ	*3*
EIIC-GAT	*3*
Bmp	*3*

Folate_rec	*3*
TetR_C_1	*3*
HAD_2	*3*
IIGP	*3*
Lipocalin_7	*3*
TetR_C_24	*3*
Metalloenzyme	*3*
2-oxoacid_dh	*3*
AnmK	*3*
Prenyltrans	*3*
BNR_2	*3*
AIG1	*3*
BioY	*3*
Dynamin_N	*2*
IL2	*2*
HD	*2*
YBD	*2*
Guanylate_cyc	*2*
PI-PLC-X	*2*
S-methyl_trans	*2*
IlvN	*2*
Gp_dh_C	*2*
Kinesin	*2*
HIT	*2*
DHDPS	*2*
Ubie_methyltran	*2*
WH1	*2*
Dak1	*2*
Glucodextran_N	*2*
Glyco_hydro_76	*2*
CYTH	*2*
Glyco_hydro_14	*2*
Reprolysin_5	*2*
Sigma54_activat	*2*
ParA	*2*
Me-amine-dh_H	*2*
Methyltransf_33	*2*
NeuB	*2*
NAGLU	*2*
CoA_binding	*2*
Transferrin	*2*
HBM	*2*
HSP90	*2*
CM_2	*2*
ACP_syn_III_C	*2*
Epimerase	*2*
Sial-lect-inser	*2*
Amidase	*2*

Myosin_head	*2*
MHC_I	*2*
Glyco_hydro_3	*2*
PAP_assoc	*2*
ROK	*2*
YcgR_2	*2*
NAD_binding_4	*2*
AICARFT_IMPCHas	*2*
MoeA_N	*2*
TNF	*2*
Lipocalin_2	*2*
PAS_4	*2*
RasGEF	*2*
SpoU_methylase	*2*
Mur_ligase_M	*2*
ADK	*2*
DcpS	*2*
TctC	*2*
SNF	*2*
Trehalase	*2*
CBM_6	*2*
ATP-synt_ab	*2*
Ald_Xan_dh_C2	*2*
Arf	*2*
Nucleos_tra2_C	*2*
3HCDH_N	*2*
NEAT	*2*
Actin	*2*
Peripla_BP_6	*2*
Aldose_epim	*2*
G-alpha	*2*
Lon_C	*2*
Methyltransf_2	*2*
Glycolytic	*2*
ZipA_C	*2*
Pneumo_ncap	*2*
NMT1	*2*
Beta-lactamase2	*2*
Oxysterol_BP	*2*
HipA_C	*2*
SKI	*2*
Asp_Glu_race	*2*
Peptidase_S21	*2*
Fic	*2*
LON_substr_bdg	*2*
6PF2K	*2*
Leukocidin	*2*

PhzC-PhzF	*2*
GST_C_2	*2*
UTRA	*2*
CVNH	*2*
Glyco_hydro_43	*2*
GGDEF	*2*
ATP-gua_Ptrans	*2*
Diphtheria_C	*2*
PilM_2	*2*
Fusion_gly	*2*
Abhydrolase_1	*2*
HI0933_like	*2*
6PGD	*2*
Carboxyl_trans	*2*
Glyco_hydro_15	*2*
FGGY_N	*2*
Cupin_5	*2*
VP4_haemagglut	*2*
NmrA	*2*
XylR_N	*2*
tRNA_m1G_MT	*2*
PGM_PMM_I	*2*
tRNA-synt_1g	*2*
Glucosamine_iso	*2*
Fumble	*2*
ATP-synt_DE_N	*2*
HN	*2*
FYVE	*2*
dCache_1	*2*
HMGL-like	*2*
Mannosidase_ig	*2*
TPP_enzyme_M_2	*2*
IGPS	*2*
GMC_oxred_N	*2*
Peptidase_C10	*2*
MDD_C	*2*
PRMT5_TIM	*1*
NAD_kinase	*1*
Phosphonate-bd	*1*
PX	*1*
Aminotran_4	*1*
Sua5_yciO_yrdC	*1*
Ferrochelatase	*1*
PBP_like_2	*1*
Glyco_hydro2_C5	*1*
SGL	*1*
Transketolase_N	*1*
FAD_binding_2	*1*

Regulator_TrmB	*1*
PEPCK_GTP	*1*
PH_9	*1*
BioW	*1*
Glyco_hydro_85	*1*
Peptidase_M27	*1*
Pirin_C	*1*
DNA_topoisoIV	*1*
Inhibitor_I9	*1*
RVT_connect	*1*
FAD_binding_6	*1*
Scytalone_dh	*1*
JHD	*1*
Calici_coat_C	*1*
Autoind_synth	*1*
DEAD	*1*
PH_12	*1*
Pro_dh	*1*
N6_N4_Mtase	*1*
PAE	*1*
Amidohydro_1	*1*
Clp_N	*1*
ABC_ATPase	*1*
Isochorismatase	*1*
PI3_PI4_kinase	*1*
7tm_2	*1*
Peroxidase_2	*1*
DrrA_P4M	*1*
EF-hand_7	*1*
DUF5115	*1*
FBPase_3	*1*
HRM	*1*
ACBP	*1*
TcdA_TcdB	*1*
CAT	*1*
Arabino_trans_C	*1*
IMPDH	*1*
FmrO	*1*
Tachylectin	*1*
HisG	*1*
Glyco_trans_1_4	*1*
Peptidase_C3	*1*
Kelch_1	*1*
KIX	*1*
53-BP1_Tudor	*1*
SusF_SusE	*1*
TP_methylase	*1*
DNA_gyraseB	*1*

OMPdecase	*1*
AKAP7_NLS	*1*
2OG-FeII_Oxy_2	*1*
Chal_sti_synt_N	*1*
GDI	*1*
PARG_cat	*1*
CAP	*1*
NTF2	*1*
UDPGP	*1*
Neocarzinostat	*1*
GHMP_kinases_C	*1*
SRPRB	*1*
ENTH	*1*
Exotox-A_bind	*1*
Transpeptidase	*1*
DUF1080	*1*
4HBT_2	*1*
HhH-GPD	*1*
DNA_pol3_delta2	*1*
His_Phos_2	*1*
MFS_1	*1*
PPV_E2_N	*1*
Menin	*1*
DAGK_cat	*1*
MIR	*1*
Ins_P5_2-kin	*1*
His_biosynth	*1*
RabGAP-TBC	*1*
SPRY	*1*
DUF4154	*1*
C2	*1*
Hexapep	*1*
PHD	*1*
Glycos_transf_1	*1*
Carb_kinase	*1*
FTO_NTD	*1*
RsgA_GTPase	*1*
PNTB	*1*
Triabin	*1*
AAA_lid_9	*1*
GTP1_OBG	*1*
DUF3372	*1*
UQ_con	*1*
YCII	*1*
Haspin_kinase	*1*
T3SS_TC	*1*
HsbA	*1*

MaoC_dehydrat_N	*1*

Appendix D

DEELIG dataset

Details of PDB complexes used in the **training**, **testing** and **validation** for DEELIG. Protein-ligand complexes used have been presented in the format '$PDB-ID_$Ligand-ID_$Chain-ID_$Atom-number' in comma separated text. Here, the ligand ID (Ligand-ID) is as obtained from PDB and 'Atom-number' indicates the atom number at which ligand is bound to the protein structure.

Table D_1: Details of PDB complexes used in the training, testing and validation for DEELIG

Training set
182L_BZF_A_401,183L_DEN_A_400,184L_I4B_A_401,185L_IND_A_400,187L_PXY_A_400,1A0T_SUC_P_1,1A0T_SUC_P_2,1A0T_SUC_Q_2,1A0T_SUC_R_2,1A4H_GDM_A_300,1A4K_FRA_B_3083,1A4K_FRA_H_3083,1A4R_GDP_A_200,1A4R_GNH_B_400,1A50_FIP_A_270,1A7X_FKA_B_201,1A99_PUT_B_371,1A99_PUT_C_371,1A99_PUT_D_371,1A9Q_HPA_A_290,1A9R_HPA_A_290,1A9T_HPA_A_290,1ABF_FCA_A_307,1ABF_FCB_A_308,1ADL_ACD_A_135,1ADO_13P_A_1104,1AG9_FMN_A_177,1AG9_FMN_B_177,1AGM_ACR_A_495,1AGM_ACR_A_496,1AHX_HCI_A_411,1AHX_HCI_B_411,1AHY_MAE_A_411,1AHY_MAE_B_411,1AJ6_NOV_A_1,1AJ7_NPE_H_217,1AKT_FMN_A_149,1AKU_FMN_A_150,1AKV_FMN_A_152,1AKW_FMN_A_149,1AM1_ADP_A_300,1AMW_ADP_A_300,1APB_FCA_A_307,1APB_FCB_A_308,1ATP_ATP_E_355,1ATR_ADP_A_486,1ATS_ADP_A_486,1AVN_AZI_A_265,1AVN_HSM_A_264,1AX0_A2G_A_401,1AXZ_GAL_A_402,1AXZ_GLA_A_401,1AZL_FMN_A_149,1AZM_AZM_A_262,1B38_ATP_A_381,1B39_ATP_A_381,1B42_M1A_A_600,1B55_4IP_A_171,1BA0_ADP_A_486,1BA1_ADP_A_486,1BAP_ARA_A_307,1BAP_ARB_A_308,1BCD_FMS_A_500,1BCJ_NGA_1_1,1BCJ_NGA_2_1,1BCU_PRL_H_280,1BF3_PHB_A_396,1BGQ_RDC_A_300,1BIR_2GP_A_106,1BIR_2GP_B_107,1BKF_FK5_A_108,1BKJ_FMN_A_241,1BKJ_FMN_B_241,1BKY_1MC_A_600,1BM7_FLF_A_501,1BN1_AL5_A_555,1BN3_AL6_A_555,1BN4_AL9_A_555,1BNN_AL1_A_555,1BNQ_AL4_A_555,1BNT_AL2_A_555,1BNU_AL3_A_555,1BNV_AL7_A_555,1BNW_TPD_A_555,1BP0_UMP_A_317,1BTN_I3P_A_107,1BWN_4IP_A_1,1BWN_4IP_A_172,1BWN_4IP_B_1,1BX6_BA1_A_351,1BZY_IMU_B_300,1BZY_IMU_C_300,1BZY_IMU_D_300,1C0I_BE2_A_1364,1C0I_BE2_A_1365,1C0L_FLA_A_1364,1C3V_NDP_A_801,1C3V_NDP_B_1301,1C5C_TK4_L_1001,1C6Y_MK1_B_524,1C6Z_ROC_B_505,1C7E_FMN_A_1149,1C7E_FMN_B_2149,1C7F_FMN_A_1149,1C7F_FMN_B_2149,1C9H_RAP_A_108,1CAM_BCT_A_500,1CBS_REA_A_200,1CC4_PHB_A_396,1CC6_PHB_A_396,1CEA_ACA_A_90,1CEA_ACA_B_90,1CEB_AMH_A_90,1CEB_AMH_B_90,1CJB_IRP_A_300,1CJB_IRP_B_300,1CJB_IRP_C_300,1CJB_IRP_D_300,1CMP_DMI_A_295,1CNW_EG1_A_555,1CNX_EG2_A_555,1CNY_EG3_A_555,1CT8_TAA_D_551,1CTR_TFP_A_153,1CVU_ACD_A_701,1CVU_ACD_B_2701,1CZC_GUA_A_414,1CZE_SIN_A_414,1CZK_FMN_A_170,1CZL_FMN_A_170,1CZO_FMN_A_170,1CZR_FMN_A_170,1D04_FMN_A_170,1D2E_GDP_A_1301,1D2E_GDP_B_1302,1D2E_GDP_C_1303,1D2S_DHT_A_301,1D2V_BR_A_601,1D2V_BR_A_758,1D2V_BR_A_889,1D2V_BR_B_601,1D2V_BR_B_758,1D2V_BR_B_843,1D2V_BR_B_889,1D3D_BZT_B_400,1D3P_BT3_B_400,1D3V_ABH_A_551,1D3V_ABH_B_552,1D4P_BPP_B_400,1D6H_COA_A_390,1D7I_DSS_A_301,1D7I_DSS_B_302,1D7J_BUQ_B_321,1DAR_GDP_A_692,1DB1_VDX_A_428,1DEH_NAD_A_377,1DET_2GP_A_105,1DF8_BTN_A_201,1DF8_BTN_B_202,1DHF_FOL_A_187,1DHF_FOL_B_187,1DHI_MTX_A_161,1DHI_MTX_B_361,1DHJ_MTX_A_161,1DHJ_MTX_B_361,1DJX_I3P_A_1,1DL7_NCH_L_999,1DMB_BCD_A_371,1DRF_FOL_A_187,1DRJ_RIP_A_272,1DRK_RIP_A_272,1DRU_NAD_A_301,1DRV_A3D_A_301,1DRW_NHD_A_301,1DUV_PSQ_G_401,1DUV_PSQ_H_402,1DUV_PSQ_I_403,1DVX_DIF_B_125,1DVZ_OFL_A_1

25,1DZK_PRZ_A_700,1E3G_R18_A_1000,1E3V_DXC_A_801,1E4H_PBR_A_998,1E5A_TBP_A_999,1E5A_TBP_B_999,1E7V_LY2_A_3095,1E8G_FAD_B_600,1E8W_QUE_A_2095,1E96_GTP_A_200,1EA1_TPF_A_470,1EC1_BEE_A_501,1ECM_TSA_A_500,1ECM_TSA_B_501,1EL7_MTD_A_801,1EL8_MSF_A_801,1EL8_MSF_B_811,1EL9_MTG_A_801,1EL9_MTG_B_811,1ELI_PYC_A_801,1ELI_PYC_B_811,1ELI_PYC_B_815,1EOU_SMS_A_301,1ERB_ETR_A_184,1EW8_PAE_A_556,1EW8_PAE_B_556,1EW9_MMQ_A_559,1EW9_MMQ_B_559,1EWJ_BLM_C_1004,1EWJ_BLM_D_1003,1EWJ_BLM_E_2001,1EWJ_BLM_F_2002,1EWJ_BLM_H_2003,1EX8_A4P_A_171,1EXA_394_A_450,1F3F_D4T_A_160,1F3F_D4T_B_161,1F3F_D4T_C_164,1F5F_DHT_A_253,1FAO_4IP_A_274,1FBY_9CR_A_500,1FBY_9CR_B_1500,1FCX_184_A_450,1FCY_564_A_450,1FCZ_156_A_450,1FD0_254_A_450,1FDQ_HXA_A_133,1FDQ_HXA_B_633,1FGY_4IP_A_1101,1FHW_I5P_B_1002,1FHX_4IP_A_1001,1FHX_4IP_B_1002,1FJ4_TLM_A_500,1FJ4_TLM_B_501,1FJ4_TLM_C_502,1FJ4_TLM_D_503,1FKB_RAP_A_108,1FKF_FK5_A_108,1FKJ_FK5_A_108,1FKL_RAP_A_108,1FL3_SPB_L_225,1FLM_FMN_A_123,1FLM_FMN_B_123,1FLR_FLU_L_600,1FM6_9CR_A_501,1FM6_9CR_U_502,1FM6_BRL_D_503,1FM6_BRL_X_504,1FM9_570_D_200,1FM9_9CR_A_201,1FRB_ZST_A_351,1FTM_AMQ_A_428,1FV0_9AR_A_401,1FZQ_GDP_A_301,1G1D_FSB_A_555,1G2K_NM1_B_501,1G45_FSB_A_555,1G46_F2B_A_555,1G48_F6B_A_555,1G4J_FFB_A_555,1G4O_BSB_A_555,1G52_F2B_A_555,1G53_F6B_A_555,1G54_FFB_A_555,1G5Y_9CR_B_501,1G5Y_9CR_C_502,1G6S_S3P_A_601,1G6T_S3P_A_601,1G74_OLA_A_132,1G85_3OL_A_1001,1G85_3OL_A_1002,1G85_3OL_B_1003,1G85_3OL_B_1004,1GAH_ACR_A_497,1GCA_GAL_A_350,1GNI_OLA_A_1001,1GNI_OLA_A_1003,1GNI_OLA_A_1005,1GT1_3OM_B_1159,1GT1_ANC_A_1161,1GT1_ANC_B_1158,1GT1_PRZ_A_1160,1GT3_3OM_A_1161,1GT3_DHM_A_1160,1GT3_DHM_B_1159,1GT4_UNA_A_1160,1GT5_BZQ_A_1160,1GU1_FA1_A_201,1GU1_FA1_B_201,1GU1_FA1_C_201,1GU1_FA1_D_201,1GU1_FA1_E_201,1GU1_FA1_F_201,1GU1_FA1_G_201,1GU1_FA1_H_201,1GU1_FA1_I_201,1GU1_FA1_J_201,1GU1_FA1_K_201,1GU1_FA1_L_201,1GUA_GNP_A_170,1GW6_BES_A_1611,1GWR_EST_A_600,1GWR_EST_B_600,1GWV_LAT_A_1371,1GWV_UDP_A_1369,1GWV_UDP_B_1369,1GWW_GLC_B_1371,1GX4_UDP_A_401,1GX4_UDP_B_401,1GX8_RTL_A_1163,1GZC_LAT_A_1559,1GZF_NAD_A_1248,1GZF_NAD_B_1252,1H0A_I3P_A_1164,1H0R_FA1_A_200,1H60_STR_A_500,1H61_PDN_A_1366,1H62_ANB_A_501,1H8S_AIC_A_1000,1HA2_SWF_A_3001,1HDR_NAD_A_244,1HJO_ADP_A_383,1HMR_ELA_A_133,1HMS_OLA_A_133,1HMT_STE_A_133,1HN2_3OL_B_3003,1HN2_ANC_A_3001,1HN2_ANC_B_3002,1HPM_ADP_A_486,1HPV_478_B_200,1HQ5_S2C_B_552,1HS6_BES_A_901,1HSG_MK1_B_902,1HTB_NAD_A_377,1HTB_NAD_B_377,1HXB_ROC_A_100,1HXW_RIT_B_301,1HZZ_NAD_A_400,1I00_D16_A_315,1I00_D16_B_409,1I1E_DM2_A_3001,1I4D_GDP_D_200,1I7Z_COC_C_302,1I80_9HX_A_301,1I80_9HX_B_302,1I80_9HX_C_303,1I9L_INV_A_555,1I9L_INV_A_666,1I9M_INW_A_555,1I9M_INW_A_666,1I9N_IOA_A_555,1I9O_IOC_A_555,1I9O_IOC_A_666,1I9P_IOE_A_555,1I9Q_IOF_A_555,1I9Q_IOF_A_666,1IE9_VDX_A_500,1IEP_STI_A_201,1IEP_STI_B_202,1IF7_SBR_A_555,1IF8_SBS_A_555,1IPC_MGP_A_1000,1IUT_PAB_A_396,1IUU_PAB_A_396,1IUW_PHB_A_396,1IUX_PHB_A_396,1IWI_CAM_A_418,1IWJ_CAM_A_418,1J0X_NAD_R_2336,1JD0_AZM_B_2401,1JGL_EST_L_911,1JIF_BLM_B_401,1JLR_GTP_A_303,1JLR_GTP_B_301,1JLR_GTP_C_300,1JLR_GTP_D_302,1JMF_UMP_A_317,1JMG_UMP_A_317,1JOC_ITP_A_101,1JOC_ITP_B_102,1JPZ_140_A_1470,1JQV_ORO_B_2313,1JQX_ORO_A_1313,1JQX_ORO_B_2313,1JR0_A24_E_104,1JR0_A24_H_104,1JRB_ORO_A_1313,1JRB_ORO_B_2313,1JRC_ORO_A_1313,1JRC_ORO_B_2313,1JT1_MCO_A_700,1K1I_FD1_A_999,1K1J_FD2_A_999,1K1L_FD3_A_999,1K1M_FD4_A_999,1K1N_CCR_A_999,1K6C_MK1_B_902,1K6P_XN3_B_807,1K6V_XN2_B_252,1K74_9CR_A_463,1KDK_DHT_A_301,1KDM_DHT_A_301,1KEL_AAH_H_219,1KF0_ACP_A_418,1KNR_FAD_A_800,1KV1_BMU_A_391,1KXM_BZI_A_2,1KY8_NAP_A_999,1KYV_RBF_A_501,1KYV_RBF_B_502,1KYV_RBF_C_503,1KYV_RBF_D_504,1KYV_RBF_E_505,1KZN_CBN_A_1,1KZO_FPP_B_1002,1L2T_ATP_A_1301,1L2T_ATP_B_1302,1L5Q_700_A_862,1L5Q_700_B_1862,1L5Q_CFF_A_863,1L5Q_CFF_A_864,1L5Q_CFF_B_1863,1L5R_700_A_862,1L5R_RBF_A_859,1L5S_700_A_862,1L5S_700_B_1862,1L5S_URC_A_863,1L5S_URC_B_1863,1L6M_DBH_A_201,1L6M_DBH_A_202,1L6M_DBH_B_301,1L6M_DBH_B_302,1L6M_DBH

_C_401,1L7X_700_B_1862,1L7X_CFF_A_863,1L7X_CFF_A_864,1L7X_CFF_B_1863,1L83_BNZ_A_400,1L8B_MGP_A_1000,1L8B_MGP_B_2000,1LBF_137_A_300,1LBK_GSH_A_504,1LBK_GSH_B_503,1LF9_ACR_A_700,1LF9_ACR_B_701,1LFO_OLA_A_128,1LFO_OLA_A_129,1LGT_BP3_A_300,1LGT_BP3_A_301,1LGW_1AN_A_401,1LHU_EST_A_301,1LHW_ESM_A_301,1LI2_IPH_A_400,1LI3_3CH_A_401,1LI6_5MP_A_400,1LIF_STE_A_132,1LIN_TFP_A_153,1LIN_TFP_A_154,1LIN_TFP_A_155,1LIN_TFP_A_156,1LNM_DTX_A_700,1LO7_4CO_A_170,1LRH_NLA_C_7190,1LT8_CBH_A_601,1LT8_CBH_B_602,1LUQ_BTN_B_5200,1LV8_9PP_A_3001,1LV8_9PP_D_3006,1LV8_9PP_E_3004,1LV8_9PP_F_3005,1LVU_9PP_A_6100,1LVU_9PP_B_6200,1LVU_9PP_C_6300,1LVU_9PP_D_6400,1LVU_9PP_E_6500,1LVU_9PP_F_6600,1LWL_DSO_A_900,1M01_NAG_A_507,1M13_HYF_A_435,1M2Z_DEX_A_301,1M2Z_DEX_D_401,1M48_FRG_A_301,1M48_FRG_B_302,1M5W_DXP_A_1001,1M5W_DXP_B_1002,1M5W_DXP_E_1005,1M5W_DXP_F_1006,1M5W_DXP_G_1007,1M6P_M6P_A_200,1M6P_M6P_B_200,1M7Y_PPG_A_600,1M83_ATP_A_400,1M8D_CLW_B_907,1M9J_CLW_A_906,1M9K_7NI_A_906,1MAU_ATP_A_400,1MD2_233_D_2201,1MD2_233_D_2203,1MD2_233_E_1204,1MD2_233_E_1206,1MD2_233_F_1207,1MEP_BTN_A_5100,1MEP_BTN_B_5200,1MEP_BTN_C_5300,1MEP_BTN_D_5400,1MES_DMP_B_323,1MET_DMP_A_323,1MEU_DMP_A_323,1MH5_HAL_B_601,1MH5_HAL_H_602,1MJ7_HAL_H_301,1MJJ_HAL_A_1001,1MK5_BTN_A_5100,1MK5_BTN_B_5200,1MPW_TMH_B_450,1MS6_BLN_A_281,1MSM_JE2_B_1001,1N0S_FLU_A_500,1N3I_DIH_A_401,1N3I_DIH_B_402,1N3I_DIH_C_403,1N43_BTN_B_5200,1N43_BTN_C_5300,1N43_BTN_D_5400,1N49_RIT_B_301,1N49_RIT_D_401,1N4H_REA_A_500,1N4K_I3P_A_1000,1N6B_DMZ_A_501,1N8U_BDD_A_411,1N8U_BDD_A_412,1N8V_BDD_A_502,1N8V_BDD_A_503,1N8V_BDD_B_511,1N8V_BDD_B_513,1N9M_BTN_A_5100,1N9M_BTN_B_5200,1N9M_BTN_C_5300,1N9M_BTN_D_5400,1NDJ_BTN_A_5100,1NDJ_BTN_B_5200,1NDJ_BTN_C_5300,1NDJ_BTN_D_5400,1NF3_GNP_A_200,1NF8_ISC_A_220,1NFS_DED_B_302,1NGA_ADP_A_486,1NGC_ADP_A_486,1NGD_ADP_A_486,1NJA_DCM_A_317,1NJD_UMP_A_317,1NJE_DCM_A_317,1NJF_ADP_C_302,1NJJ_GET_A_601,1NJJ_GET_B_602,1NJJ_GET_C_603,1NJJ_ORX_A_602,1NJJ_ORX_C_604,1NJJ_ORX_D_426,1NKI_PPF_B_5002,1NL5_MAL_A_371,1NLI_ADE_A_248,1NM5_NAD_A_400,1NO6_794_A_322,1NOX_FMN_A_300,1NPL_MAN_A_110,1NPL_MAN_A_112,1NPL_MAN_A_120,1NPL_MAN_A_130,1NQM_BTN_A_150,1NQM_BTN_B_350,1NQM_BTN_C_550,1NQM_BTN_D_750,1NR6_DIF_A_501,1NRJ_GTP_B_1001,1NW4_IMH_A_301,1NW4_IMH_B_302,1NW4_IMH_C_303,1NW4_IMH_D_304,1NW4_IMH_E_305,1NW4_IMH_F_306,1NW5_SAM_A_401,1NW7_SAH_A_401,1O7O_LAT_A_1369,1O7O_LAT_B_2369,1O7O_UDP_A_1371,1O7Q_UDP_A_1374,1O7Q_UDP_B_1376,1O7Q_UDP_B_1377,1OAR_AZN_I_500,1OAR_AZN_J_500,1OAR_AZN_K_500,1OBA_CHT_A_1341,1OBO_FMN_A_1170,1OBV_FMN_A_1170,1ODI_ADN_A_1237,1ODI_ADN_C_1238,1ODI_ADN_D_1237,1ODI_ADN_E_1237,1ODI_ADN_F_1238,1ODJ_GMP_A_1237,1ODJ_GMP_B_1237,1ODJ_GMP_D_1237,1ODJ_GMP_E_1237,1ODJ_GMP_F_1237,1OE7_GSH_B_301,1OE8_GSH_A_301,1OE8_GSH_B_301,1OFZ_FUL_A_1313,1OFZ_FUL_A_1317,1OFZ_FUL_B_1313,1OFZ_FUL_B_1315,1OFZ_FUL_B_1317,1OFZ_FUL_B_1318,1OGD_RIP_A_1134,1OGD_RIP_B_1132,1OGD_RIP_C_1132,1OGD_RIP_D_1132,1OGD_RIP_E_1132,1OGX_EQU_A_1128,1OGX_EQU_B_1328,1OGZ_EQU_A_1126,1OIF_IFM_B_1446,1OIM_NOJ_A_1447,1OIM_NOJ_B_1447,1OKN_STB_A_555,1OKO_GAL_A_901,1OKO_GAL_B_902,1OKO_GAL_D_902,1OKO_GLA_C_901,1OKO_GLA_D_901,1OLS_TDP_A_601,1OLU_TDP_A_601,1OLX_TDP_A_601,1ONZ_968_A_322,1OPB_RET_A_134,1OPB_RET_B_134,1OPB_RET_C_134,1OPJ_STI_A_3,1OPJ_STI_B_4,1OS5_NH1_A_901,1OSV_CHC_A_202,1OSV_CHC_B_201,1OVD_ORO_A_900,1OVD_ORO_B_901,1OW4_2AN_B_1003,1OWB_NAD_B_3001,1OXR_AIN_A_141,1OZ0_MS1_A_1002,1OZ0_MS1_B_1001,1P19_IMP_A_1300,1P19_IMP_B_2300,1P19_IMP_D_4300,1P1N_KAI_A_999,1P1O_QUS_A_301,1P1Q_AMQ_A_602,1P1Q_AMQ_B_603,1P1Q_AMQ_C_601,1P28_HBS_B_1003,1P9P_SAH_A_300,1PFU_MPJ_A_553,1PG0_MOD_A_553,1PGP_6PG_A_502,1PK8_ATP_A_800,1PK8_ATP_B_801,1PK8_ATP_C_802,1PK8_ATP_D_803,1PK8_ATP_E_804,1PK8_ATP_F_805,1PK8_ATP_G_806,1POY_SPD_1_400,1POY_SPD_2_400,1POY_SPD_3_400,1PWM_FID_A_1320,1PX2_ATP_A_800,1PXJ_CK2_A_500,1PZI_1DM_F_106,1PZI_1DM_G_107,1Q0Y_MOI_H_401,1Q1G_MTI_A_301,1Q1G_

MTI_B_302,1Q1G_MTI_C_303,1Q1G_MTI_D_304,1Q1G_MTI_E_305,1Q1G_MTI_F_306,1Q72_COC_H_401,1Q7B_NAP_A_1901,1Q7B_NAP_B_2901,1Q7B_NAP_C_3901,1Q7B_NAP_D_4901,1Q8T_Y27_A_930,1Q8U_H52_A_961,1Q8U_H52_A_962,1Q8W_M77_A_960,1Q9D_OI1_A_416,1Q9D_OI1_B_406,1QBS_DMP_A_323,1QFT_HSM_A_176,1QFT_HSM_A_177,1QFT_HSM_B_176,1QFT_HSM_B_177,1QFY_NAP_A_310,1QFY_NAP_B_810,1QG6_TCL_A_601,1QG6_TCL_B_602,1QG6_TCL_C_603,1QG6_TCL_D_604,1QGA_NAP_A_310,1QGA_NAP_B_810,1QIW_DPD_A_154,1QIW_DPD_B_154,1QK3_5GP_A_300,1QK3_5GP_B_300,1QK3_5GP_C_300,1QK3_5GP_D_300,1QK4_IMP_A_300,1QK4_IMP_B_300,1QK4_IMP_C_300,1QK4_IMP_D_300,1QL7_ZEN_A_999,1QLT_FAD_A_600,1QLT_FAD_B_600,1QPB_PYM_A_602,1QPB_PYM_B_602,1QPE_PP2_A_1904,1QQM_ADP_A_486,1QQN_ADP_A_486,1QQO_ADP_A_486,1QXO_FMN_A_4001,1QXO_FMN_B_4002,1QXO_FMN_C_4003,1QXO_FMN_C_4005,1QXO_FMN_D_4004,1QXO_FMN_D_4006,1QY1_PRZ_A_300,1QY2_IPZ_A_300,1R0E_DFN_A_702,1R0E_DFN_B_501,1R0P_KSA_A_0,1R0X_ATP_A_1,1R0X_ATP_B_2,1R0X_ATP_C_3,1R37_NAD_A_403,1R37_NAD_B_403,1R4W_GSH_B_401,1R4W_GSH_D_901,1R6N_434_A_1111,1R9L_BET_A_1001,1RBO_CAP_H_476,1RBO_CAP_L_476,1RD4_L08_A_328,1RD4_L08_B_1328,1RD4_L08_C_2328,1RD4_L08_D_3328,1RG7_MTX_A_161,1RGL_2GP_A_105,1RK3_VDX_A_500,1RMZ_NGH_A_269,1RN8_DUP_A_777,1RPJ_ALL_A_291,1RR6_IMH_A_300,1RU1_PH2_A_181,1RU1_PH2_B_381,1RU2_APC_A_171,1RU2_HHR_A_181,1RX1_NAP_A_164,1RYF_GDP_A_538,1RYF_GDP_B_1538,1RYH_GNP_A_538,1RZ0_FAD_A_1200,1RZ0_FAD_B_2200,1RZ0_FAD_C_3200,1RZ0_FAD_D_4200,1RZ0_FAD_E_5200,1RZ0_FAD_F_6200,1RZ0_FAD_H_8200,1RZ1_FAD_C_3200,1RZ1_FAD_D_4200,1RZ1_FAD_E_5200,1RZ1_FAD_F_6200,1RZ1_FAD_G_7200,1RZ1_FAD_H_8200,1S19_MC9_A_500,1S3X_ADP_A_383,1S8J_NO3_A_801,1SBR_VIB_A_501,1SBR_VIB_B_503,1SBR_VIB_B_504,1SDU_MK1_B_902,1SDV_MK1_B_902,1SG0_STL_A_501,1SG0_STL_B_502,1SGU_MK1_B_2632,1SH9_RIT_B_301,1SQN_NDR_A_1001,1SR7_MOF_A_301,1SR7_MOF_B_302,1SRE_HAB_A_300,1SRE_HAB_B_300,1SRG_MHB_A_300,1SRG_MHB_B_300,1SRI_DMB_A_300,1SRI_DMB_B_300,1SV3_ANN_A_347,1SVS_GNP_A_355,1SW1_PBE_A_301,1SW2_BET_A_301,1SWD_BTN_D_600,1SWE_BTN_A_300,1SWE_BTN_D_600,1SWG_BTN_A_600,1SWG_BTN_C_800,1SWK_BTN_A_5100,1SWK_BTN_D_5400,1SWN_BTN_A_300,1SWN_BTN_D_600,1SWP_BTN_A_300,1SWP_BTN_D_600,1SWP_BTQ_C_500,1SWR_BTN_B_400,1SWR_BTN_C_500,1SWR_BTN_D_600,1SWT_BTN_A_500,1SWT_BTN_B_600,1SYN_F89_A_301,1SYN_F89_B_301,1SYO_M6P_A_500,1SYO_M6P_B_1500,1SZ0_M6P_B_1500,1SZD_APR_A_1001,1T3R_017_A_1200,1T5Z_DHT_A_931,1T63_DHT_A_931,1T64_TSN_A_386,1T64_TSN_B_1387,1T79_DHT_A_200,1T7F_DHT_A_200,1T7I_017_A_200,1T7J_478_A_200,1T7R_DHT_A_300,1TCO_FK5_C_509,1TD7_NFL_A_2001,1TL2_NDG_A_237,1TL2_NDG_A_238,1TL2_NDG_A_239,1TL2_NDG_A_240,1TL2_NDG_A_241,1TLC_DGP_B_265,1TLG_GAL_A_126,1TLG_GAL_B_126,1TMM_APC_A_171,1TMM_APC_B_371,1TMM_HHR_A_181,1TMM_HHR_B_381,1TOG_HCI_A_410,1TOG_HCI_B_410,1TOI_HCI_A_410,1TOJ_HCI_A_410,1TOK_MAE_A_410,1TOK_MAE_B_410,1TOX_NAD_A_536,1TPZ_GDP_A_1500,1TPZ_GDP_B_2500,1TR7_DEG_A_501,1TR7_DEG_B_502,1TSD_UMP_A_267,1TSD_UMP_B_265,1TSY_UMP_A_317,1TT1_KAI_A_998,1TT1_KAI_B_999,1TVO_FRZ_A_1001,1TYS_DHF_A_566,1TYS_TMP_A_565,1U1B_PAX_A_901,1U1B_PAX_B_902,1U1W_3HA_A_700,1U1W_3HA_B_701,1U21_P2C_A_211,1U21_P2C_B_221,1U3D_ANP_A_511,1U5B_TDP_A_601,1UAD_GNP_A_200,1UAD_GNP_B_1200,1UCN_ADP_A_1160,1UCN_ADP_B_2160,1UCN_ADP_C_3160,1UDT_VIA_A_1000,1UGW_GAL_A_200,1UGW_GAL_C_1200,1UHO_VDN_A_1000,1UNL_RRC_A_1293,1UOD_G3H_A_1369,1UOD_G3H_B_1369,1UPV_444_A_462,1UPW_444_A_1462,1UTJ_ABN_A_246,1UTL_PRA_M_246,1UTN_ABN_A_246,1UTN_ABN_A_247,1UTO_PEA_A_1246,1UTP_PBN_A_1246,1UU3_LY4_A_1374,1UV6_CCE_D_1206,1UW1_ADP_A_1074,1UWF_DEG_A_1159,1UZV_FUC_B_999,1V11_TDP_A_601,1V16_TDP_A_601,1V1M_TDP_A_1401,1V48_HA1_A_290,1VC9_ATP_A_1001,1VC9_ATP_B_1002,1VFN_HPA_A_300,1VG0_GER_A_952,1VJC_ATP_A_417,1VJD_ATP_A_417,1VKJ_A3P_A_604,1VPO_TES_H_1010,1VYF_OLA_A_1134,1VYG_ACD_A_1134,1VYS_TNF_X_500,1VZC_UFP_A_317,1VZD_UFP_A_317,1W0O_SIA_A_1783,1W0P_SIA_A_1784,1W3J_OXZ_A_1447,1W6O_LAT_A_4000,1W6O_LAT_B_4000,1W6Y_EQU_

A_201,1W96_S1A_B_1567,1W96_S1A_C_1550,1WCB_PE1_L_1216,1WDQ_MAL_A_498,1WDR_MAL_A_496,1WDR_MAL_A_498,1WE2_DHK_A_181,1WEI_ADE_A_1428,1WPG_TG1_A_1003,1WPG_TG1_B_1103,1WPG_TG1_C_1203,1WS4_GYP_A_501,1WS4_GYP_G_504,1WS5_MMA_A_501,1WS5_MMA_C_502,1WS5_MMA_E_503,1WS5_MMA_G_504,1WVA_S2C_A_1317,1WVA_S2C_B_1316,1X07_IPE_A_901,1X09_IPE_A_902,1X70_715_A_801,1X70_715_B_801,1X7W_TDP_A_801,1X7X_TDP_A_801,1X7Z_TDP_A_801,1X80_TDP_A_801,1X8D_RNS_A_1105,1X8D_RNS_B_1106,1X8D_RNS_C_1107,1X8D_RNS_D_1108,1X8R_SC1_A_601,1X8V_ESL_A_471,1X8V_ESL_A_472,1X9D_SMD_A_1003,1X9Q_FLU_A_1341,1XA5_KAR_A_153,1XGE_DOR_A_1410,1XH4_R69_A_351,1XH5_R68_A_1001,1XJB_NAP_A_1001,1XJB_NAP_B_5001,1XK5_TPG_A_400,1XKK_FMM_A_91,1XPZ_4TZ_A_270,1XQ0_4TR_A_270,1XQ3_R18_A_1001,1XQD_DND_A_502,1XW6_GSH_A_750,1XW6_GSH_C_752,1XW6_GSH_D_753,1XWS_BI1_A_1001,1XZX_T3_X_500,1Y0L_HAN_A_601,1Y0L_HAN_D_701,1Y0L_HAN_F_801,1Y2F_WAI_A_300,1Y2G_CL3_A_301,1Y52_BTN_X_801,1Y52_BTN_Y_802,1Y5B_TL3_T_720,1Y5U_TL4_T_360,1Y6Q_TDI_A_233,1Y6Q_TDI_B_234,1Y7I_SAL_A_501,1Y7I_SAL_A_502,1Y7I_SAL_B_503,1Y8O_ADP_A_504,1Y93_HAE_A_301,1YA3_STR_A_1001,1YA3_STR_C_3001,
1YC1_4BC_A_401,1YC4_43P_A_301,1YDB_AZM_A_264,1YDD_AZM_A_264,1YDK_GTX_A_5100,1YDK_GTX_B_5200,1YDT_IQB_E_351,1YEI_PGG_L_551,1YEJ_PNF_L_551,1YET_GDM_A_275,1YID_ATP_B_3000,1YJQ_NAP_A_501,1YKJ_PHB_A_1396,1YKJ_PHB_B_2396,1YKL_DHB_D_1550,1YKL_DHB_F_2550,1YKL_DHB_H_3550,1YKL_DHB_J_4550,1YKL_DHB_L_5550,1YKP_DHB_D_1550,1YKP_DHB_H_3550,1YKP_DHB_J_4550,1YKP_DHB_L_5550,1YP6_PRZ_A_203,1YYZ_AMP_A_435,1YZ0_AMP_A_1844,1YZ0_AMP_B_1841,1Z10_COU_B_501,1Z10_COU_D_501,1Z11_8MO_A_501,1Z11_8MO_B_501,1Z11_8MO_C_501,1Z11_8MO_D_501,1Z3V_LAT_A_500,1ZC9_PMP_A_502,1ZDQ_MSM_B_1508,1ZDS_ACM_A_2500,1ZDS_ACM_A_2503,1ZDS_ACM_B_2501,1ZDS_ACM_B_2504,1ZDS_ACM_C_2502,1ZDS_ACM_C_2505,1ZHA_PEP_A_1268,1ZHA_PEP_B_2268,1ZHX_HC3_A_1001,1ZHY_CLR_A_1001,1ZOA_140_A_1470,1ZSB_AZM_A_264,1ZZ1_SHH_A_2452,1ZZ1_SHH_C_2652,1ZZ1_SHH_D_2752,220L_BNZ_A_169,223L_BNZ_A_169,225L_PXY_A_169,2A06_SMA_P_3001,2A2I_A5P_A_1269,2A3C_PNX_A_1434,2A3C_PNX_A_1435,2A3C_PNX_B_2433,2A3C_PNX_B_2434,2A4W_BLM_A_750,2A5B_8HG_A_255,2A5C_8DA_A_303,2A5C_8DA_B_201,2A8G_GNG_B_201,2AA5_STR_A_301,2AA6_STR_A_401,2AA6_STR_B_402,2AAC_FCB_B_179,2ABA_STR_A_1500,2ACO_VCA_B_501,2AEB_ABH_A_551,2AEB_ABH_B_552,2AGT_FID_A_320,2AGV_BHQ_A_1004,2AGV_BHQ_B_1104,2AGV_TG1_A_1003,2AGV_TG1_B_1103,2AMT_1AA_A_1901,2AMT_1AA_C_2901,2AMT_1AA_C_3901,2AMT_1AA_E_4901,2AMT_1AA_E_5901,2ANS_2AN_A_201,2ANS_2AN_B_202,2AO6_R18_A_1001,2AQB_TL6_A_600,2AQB_TL6_D_602,2AQU_DR7_B_300,2ARC_ARA_A_171,2ARM_OIN_A_401,2AS1_TP5_A_500,2AS3_IPH_A_500,2AS4_3FA_A_500,2AS4_3FA_A_700,2ATE_NIA_A_400,2AVO_MK1_B_902,2AVS_MK1_B_902,2AVV_MK1_A_901,2AVV_MK1_E_902,2AY1_AHC_B_414,2AY3_MPP_A_414,2AY4_PPT_B_414,2AY5_IOP_B_414,2AY6_3IB_B_414,2AY8_4TB_A_414,2AY8_4TB_B_414,2AY9_5PV_A_414,2AY9_5PV_B_414,2AZ5_307_A_1,2B3B_BGC_C_2607,2B3B_BGC_D_2608,2B3B_GLC_A_2605,2B3B_GLC_B_2606,2B3B_GLC_E_2609,2B3B_GLC_F_2610,2B3F_GAL_C_3312,2B3F_GAL_D_3313,2B3F_GAL_E_3314,2B3F_GAL_F_3315,2B4L_BET_A_273,2B60_RIT_B_100,2B7Z_MK1_B_200,2B9A_FBC_A_301,2B9A_FBC_B_302,2BAK_AQZ_A_401,2BFR_ADP_A_1193,2BIK_BI1_B_1306,2BJM_ANF_H_500,2BMK_PDD_L_1215,2BMV_BEN_A_1165,2BMZ_XLM_A_1142,2BMZ_XLM_B_1142,2BMZ_XLM_B_1146,2BOJ_ARW_A_1115,2BOJ_ARW_C_1115,2BOJ_ARW_D_1115,2BSA_NAP_A_1305,2BSL_DHB_B_1313,2BT9_MFU_A_1091,2BT9_MFU_A_1092,2BT9_MFU_B_1089,2BT9_MFU_B_1090,2BT9_MFU_C_1090,2BT9_MFU_C_1091,2BUP_ADP_A_486,2BUV_DHB_B_1542,2BYR_MLK_A_301,2BYR_MLK_B_301,2BYR_MLK_C_301,2BYR_MLK_D_301,2BYR_MLK_E_301,2BYR_MLK_F_301,2BYR_MLK_H_301,2BYR_MLK_I_301,2BYS_LOB_A_301,2BYS_LOB_B_301,2BYS_LOB_C_301,2BYS_LOB_D_301,2BYS_LOB_G_301,2BYS_LOB_H_301,2BYS_LOB_I_301,2BZD_GAL_A_1649,2BZD_GAL_B_1649,2BZD_GAL_C_1649,2C1P_FNZ_B_1215,2C1P_FNZ_H_1214,2C1Q_BTN_A_1126,2C3I_IYZ_B_1306,2C5O_CK2_A_1297,2C5O_CK2_C_1298,2C80_GTX_A_1212,2C80_

GTX_B_1213,2C8H_NAD_A_1248,2C8H_NAD_B_1250,2C92_TP6_A_701,2C92_TP6_B_701,2C92_TP6_C_701,2C92_TP6_D_701,2C94_TSF_B_701,2C94_TSF_C_701,2C94_TSF_D_701,2C97_JCL_B_701,2C97_JCL_C_701,2C97_JCL_D_701,2C97_JCL_E_701,2CA8_GSH_A_1212,2CB3_MLD_A_1345,2CB3_MLD_B_1344,2CB3_MLD_C_1344,2CBO_TH2_A_120,2CBU_CTS_A_1447,2CBU_CTS_B_1445,2CBV_CGB_A_1447,2CBV_CGB_B_1445,2CBZ_ATP_A_1873,2CES_GIM_A_1447,2CES_GIM_B_1446,2CET_PGI_A_1447,2CEX_DAN_B_1307,2CGF_P2N_A_1214,2CGR_GAS_H_215,2CI0_1CM_A_1450,2CIB_CM6_A_1450,2CIR_BG6_A_501,2CIX_CEJ_A_1320,2CJW_GDP_A_999,2CJW_GDP_B_999,2CKM_AA7_A_1536,2CLE_F6F_A_1268,2CLF_F6F_A_1269,2CLF_F6F_B_1396,2CLH_F19_A_1268,2CLI_F9F_A_1269,2CLK_G3H_A_1268,2CLO_F19_A_1268,2CVD_HQL_A_2201,2CVD_HQL_B_2401,2CVD_HQL_D_2801,2D06_EST_B_305,2D09_FLV_A_431,2D09_FLV_A_432,2D0E_NQ_A_409,2D0E_NQ_A_410,2D0K_FOL_B_2161,2DM5_ODI_A_312,2DOO_C4H_A_601,2DOO_C4H_B_602,2DQS_TG1_A_1003,2DRC_MTX_A_161,2DRC_MTX_B_161,2DRI_RIP_A_272,2DU8_BEZ_A_352,2DU8_BEZ_G_2352,2DU8_BEZ_J_3352,2DU8_FAD_B_1351,2DU8_FAD_G_2351,2DU8_FAD_J_3351,2E27_AB0_H_6001,2E2R_2OH_A_1401,2E5Y_ATP_B_1002,2E7F_C2F_B_4000,2EF9_A5P_A_1269,2EF9_A5P_B_2269,2EUQ_BVC_A_401,2EWA_SB2_A_361,2EWS_ANP_A_1001,2EWS_ANP_B_1002,2EXX_NAP_A_3698,2EZ7_DHI_A_301,2F01_BTN_A_5100,2F01_BTN_B_5200,2F01_BTQ_A_5101,2F01_BTQ_B_5201,2F14_FL1_A_1270,2F2U_M77_A_501,2F2U_M77_B_1501,2F32_EGD_A_300,2F34_UBA_A_801,2F34_UBA_B_802,2F35_UBC_A_801,2F35_UBC_B_802,2F3K_RO1_A_506,2F47_MGX_A_300,2F4J_VX6_A_514,2F5T_MAL_X_343,2F78_BEZ_A_1001,2F78_BEZ_A_1003,2F78_BEZ_B_1002,2F78_BEZ_B_1005,2F7A_BEZ_A_1003,2F7A_CCU_B_1001,2F7I_26C_A_326,2F7I_26C_B_325,2F80_017_B_301,2F81_017_A_302,2F8F_GSH_A_212,2F8F_GSH_B_301,2F8I_205_B_1001,2FF7_ADP_A_708,2FKE_FK5_A_108,2FKF_G16_A_465,2FLH_ZEA_A_701,2FLH_ZEA_A_702,2FLH_ZEA_A_708,2FLH_ZEA_B_703,2FLH_ZEA_B_704,2FLH_ZEA_C_709,2FLH_ZEA_D_706,2FOU_B22_A_303,2FOV_B30_A_301,2FOY_B30_A_311,2FOY_B30_A_313,2FOY_B30_A_315,2FOY_B30_B_312,2FOY_B30_B_314,2FQW_NOS_A_400,2FQX_GMP_A_400,2FSV_NAD_A_600,2FV5_541_A_1,2FV5_541_B_2,2FWP_CIT_B_1002,2FXD_DR7_A_102,2FXS_RDA_A_1001,2FXU_BID_A_402,2FXV_5GP_A_200,2FZW_NAD_A_1377,2FZW_NAD_B_2377,2G5U_PCQ_A_240,2G5U_PCQ_B_239,2G79_RET_A_200,2G86_UMP_A_317,2GBP_BGC_A_310,2GCE_RFC_A_2751,2GCE_RFC_B_2752,2GCE_RFC_C_2753,2GCE_RFC_D_2754,2GCE_SFC_A_1751,2GCE_SFC_B_1752,2GCE_SFC_C_1753,2GEK_GDP_A_2567,2GFD_RDA_B_1002,2GH7_BTN_A_5100,2GH7_BTN_B_5200,2GH7_BTQ_A_5101,2GH7_BTQ_B_5201,2GH9_MLR_A_2913,2GJ5_VD3_A_164,2GL0_ADN_A_901,2GL0_ADN_B_902,2GL0_ADN_C_903,2GL0_ADN_D_904,2GL0_ADN_E_905,2GL0_ADN_F_906,2GQG_1N1_A_501,2GQG_1N1_B_502,2GQX_5CL_A_1450,2GVV_DI9_A_502,2GYU_HI6_A_702,2GYW_OBI_A_952,2GZ1_NAP_A_702,2GZ2_A2P_B_501,2GZ3_NAP_A_367,2GZ3_NAP_B_367,2GZ3_NAP_C_367,2GZ3_NAP_D_367,2H12_OAA_A_5007,2H12_OAA_B_5008,2H12_OAA_C_5009,2H12_OAA_D_5010,2H12_OAA_E_5011,2H12_OAA_F_5012,2H15_B19_A_300,2H21_SAM_A_801,2H21_SAM_B_802,2H21_SAM_C_803,2H23_SAH_A_800,2H23_SAH_B_801,2H23_SAH_C_802,2H2J_SFG_B_801,2H2J_SFG_C_802,2H3E_6PR_A_331,2H3E_6PR_C_332,2H4N_AZM_A_264,2H5A_X1P_X_600,2H6B_3C4_A_302,2H6B_3C4_B_301,2H77_T3_A_1,2H79_T3_A_1,2H7R_1MZ_A_416,2H8Z_8CM_A_801,2H8Z_8CM_A_802,2HA0_CHH_A_1602,2HA0_CHH_B_2602,2HA2_SCK_A_901,2HA2_SCK_B_951,2HA3_CHT_A_951,2HA3_CHT_B_961,2HA3_CHT_B_962,2HA5_AT3_A_1901,2HA5_AT3_A_1902,2HA5_AT3_B_1951,2HA5_AT3_B_1952,2HA5_AT3_B_1961,2HA5_ETM_A_1905,2HA5_ETM_A_1909,2HA5_ETM_B_1955,2HA5_ETM_B_1959,2HA6_SCK_A_902,2HA6_SCK_B_951,2HA6_SCK_B_952,2HAW_2PN_A_2001,2HAW_2PN_B_2002,2HAW_MG_A_1001,2HAW_MG_A_1002,2HAW_MG_A_1003,2HAW_MG_A_1004,2HAW_MG_B_1005,2HAW_MG_B_1006,2HAW_MG_B_1007,2HAW_MG_B_1008,2HBT_UN9_A_1,2HBU_UN9_A_1,2HF8_GSP_A_300,2HF9_GSP_A_300,2HJ4_PNZ_A_1203,2HJ4_PNZ_D_1201,2HJ4_PNZ_D_1202,2HJ4_PNZ_H_1204,2HJB_PZM_D_1102,2HMU_ATP_B_602,2HMV_ADP_A_601,2HMV_ADP_B_602,2HQU_DUP_A_777,2HQU_DUP_B_777,2HQU_DUP_C_777,2HRM_UC5_A_777,2HS1_017_A_201,2HS1_017_B_203,2HS2_017_A_201,2HYY_STI_A_600,2HYY_STI_B_600,2H

YY_STI_C_600,2HYY_STI_D_600,2HZI_JIN_B_600,2HZL_PYR_A_601,2HZL_PYR_B_602,2HZQ_STR_A_300,2I4J_DRJ_A_999,2I4P_DRH_A_999,2I4V_DJR_B_300,2I4X_KGQ_B_300,2I4Z_DRH_A_999,2I6B_89I_A_500,2IDW_017_B_401,2IDZ_ZID_A_300,2IEN_017_B_402,2IEO_017_A_402,2II3_CAO_A_500,2II3_CAO_G_500,2IJ7_TPF_A_2472,2IJ7_TPF_B_2470,2IJ7_TPF_C_2471,2IJ7_TPF_D_2473,2IKI_388_A_600,2IKJ_393_A_600,2IKO_7IG_A_601,2IL2_LIX_A_401,2IL2_LIX_B_402,2ILZ_GTP_A_1001,2IMD_GSH_A_301,2IMD_TOH_A_304,2IME_GSH_A_301,2IME_TOH_A_304,2IMF_GSH_A_301,2ITM_XUL_A_4930,2ITM_XUL_B_4931,2IW4_2PN_B_1318,2IW4_MN_A_1313,2IW4_MN_A_1314,2IW4_MN_B_1317,2IWX_M1S_A_1215,2IZF_BTN_B_300,2IZF_BTN_D_301,2IZG_BTN_B_300,2IZG_BTN_D_301,2IZH_BTN_D_300,2IZI_BTN_A_300,2IZL_IMI_B_300,2IZL_IMI_D_300,2J2I_LY4_B_1307,2J4K_U5P_A_227,2J4K_U5P_B_227,2J4K_U5P_C_227,2J4K_U5P_D_227,2J4K_U5P_F_227,2J75_NOY_A_1446,2J75_NOY_B_1445,2J77_NOJ_A_1446,2J77_NOJ_B_1446,2J78_GOX_A_1451,2J78_GOX_B_1446,2J79_GTL_A_1446,2J79_GTL_B_1446,2J7B_NTZ_B_1446,2J7E_GI2_A_1446,2J7E_GI2_B_1446,2J7F_GI3_A_1446,2J7F_GI3_B_1446,2J7G_GI4_A_1446,2J7H_AZF_A_1446,2J7H_AZF_B_1445,2J9L_ATP_A_1752,2J9L_ATP_C_1755,2J9L_ATP_D_1753,2J9L_ATP_F_1750,2J9N_BEN_A_1241,2JBJ_G88_A_1768,2JBV_FAO_A_1528,2JBV_FAO_B_1531,2JBZ_COA_A_1126,2JCV_FOM_A_1390,2JCX_FOM_A_1390,2JDM_MFU_B_1117,2JDM_MFU_C_1117,2JDM_MFU_D_1117,2JDN_MMA_B_700,2JDN_MMA_C_700,2JDN_MMA_D_700,2JDP_MFU_A_1115,2JDP_MFU_C_1115,2JDU_MFU_C_1117,2JDU_MFU_D_1117,2JDY_MMA_A_1115,2JDY_MMA_B_1115,2JDY_MMA_C_1115,2JF4_VDM_A_1548,2JFF_LK2_A_1440,2JFZ_003_B_1256,2JFZ_DGL_B_1257,2JG0_TTZ_A_1548,2JGS_BTN_A_1001,2JGS_BTN_B_1001,2JGS_BTN_C_1001,2JGS_BTN_D_1001,2JHF_NAD_A_403,2JHF_NAD_B_403,2JJ3_JJ3_B_1501,2JJC_LGA_A_1225,2JJO_EY5_A_413,2JJP_KLN_A_413,2JLB_UDM_A_1569,2JLB_UDM_B_1569,2MAS_PIR_A_400,2MAS_PIR_B_400,2MAS_PIR_C_400,2MAS_PIR_D_400,2NMX_M25_A_311,2NMX_M25_B_312,2NMY_ROC_A_401,2NN1_M28_B_312,2NN7_M29_A_311,2NN7_M29_B_312,2NND_PRZ_A_300,2NNG_ZYX_A_301,2NNK_ROC_A_401,2NNO_M28_A_301,2NNO_M28_A_302,2NNQ_T4B_A_293,2NNS_CL_A_321,2NNS_M25_A_301,2NNS_M25_A_302,2NNV_M29_A_302,2NPQ_BOG_A_1000,2NPQ_BOG_A_2000,2NSL_C2R_A_300,2NWS_A5P_A_1269,2NWS_A5P_B_2269,2NX1_RP5_A_1269,2NX1_RP5_B_2269,2NX3_A5P_A_269,2NX3_A5P_B_269,2NX3_A5P_C_269,2NX3_A5P_E_269,2NX3_A5P_F_269,2NX3_A5P_G_269,2NX3_A5P_H_269,2NX3_A5P_J_269,2NX3_A5P_K_269,2NX3_A5P_L_269,2NXE_SAM_A_302,2NXE_SAM_B_303,2NZ5_226_A_431,2NZ5_226_A_432,2NZ5_226_B_431,2NZ5_226_B_432,2O1V_ADP_A_301,2O1V_ADP_B_302,2O3P_QUE_A_501,2O3Z_AI7_A_301,2O3Z_AI7_B_301,2O4K_DR7_A_301,2O4L_TPV_A_403,2O4P_TPV_A_300,2O4S_AB1_A_400,2O5Z_ANO_H_501,2O63_MYC_A_501,2O64_MYU_A_501,2O9A_PYR_A_905,2O9A_PYR_B_906,2O9A_PYR_C_907,2O9A_PYR_D_908,2OAH_QIN_A_902,2OFP_NAP_A_304,2OFP_NAP_B_802,2OFP_PAF_A_602,2OFP_PAF_B_601,2OFU_1N9_A_1,2OGY_C2F_A_3000,2OGY_C2F_B_4000,2OI2_DP6_A_300,2OIQ_STI_A_1001,2OKG_G3H_B_501,2OO0_PLP_A_600,2OO0_PLP_B_600,2OO0_XAP_A_601,2OOR_NAD_A_500,2OOR_TXP_C_400,2OTY_YAN_X_305,2OU0_MR3_X_304,2OV4_AQP_A_400,2OVV_PFH_A_999,2OZ2_D1R_A_301,2OZ2_D1R_C_300,2OZL_TPP_C_1330,2P09_ATP_A_500,2P0D_I3P_A_800,2P15_EZT_B_601,2P4S_DIH_A_401,2P4S_DIH_B_403,2P7A_43M_A_1,2P7G_2OH_A_1,2P7Z_OHT_A_201,2P8N_ADE_A_271,2P9T_3PG_A_500,2PCP_1PC_B_227,2PCP_1PC_C_212,2PD5_ZST_A_600,2PDF_ZST_A_600,2PDH_47D_A_600,2PDJ_393_A_600,2PDK_SBI_A_600,2PDL_TOL_A_600,2PDM_ZST_A_600,2PDN_47D_A_600,2PDP_393_A_600,2PDQ_47D_A_600,2PDU_393_A_600,2PDW_FID_A_600,2PDX_ZST_A_600,2PDY_FID_A_600,2PEV_FID_A_321,2PEV_LDT_A_320,2PF8_FID_A_321,2PF8_LDT_A_320,2PFH_FID_A_321,2PFH_LDT_A_320,2PFY_PCA_A_302,2PFY_PCA_B_302,2PFY_PCA_C_302,2PGZ_COC_D_401,2PHB_230_A_9001,2PHO_TSZ_A_501,2PJO_NMU_A_303,2PK5_075_A_900,2PK6_O33_B_301,2PLL_ABH_A_555,2PLL_ABH_B_556,2PQL_TSS_A_145,2PVW_G88_A_1768,2PY4_DUP_A_777,2PYM_1UN_A_1001,2PYN_1UN_A_1001,2PZE_ATP_B_2,2PZN_393_A_320,2PZV_IPH_A_401,2PZV_IPH_C_403,2PZV_IPH_D_404,2Q0I_BEZ_A_990,2Q0J_BEZ_A_500,2Q0J_BEZ_B_500,2Q2N_H01_A_9001,2Q5S_NZA_A_7001,2Q5S_NZA_B_5001,2Q63_1UN_A_1001,2Q64_1UN_B_1001,2Q6M_P34_A_100

1,2Q6M_P34_A_1002,2Q7Q_C2B_D_2002,2Q7Q_C2B_H_2001,2Q88_4CS_A_501,2Q8H_TF4_A_438,2Q9F_C3S_A_600,2QBU_SAH_B_501,2QBU_SAH_B_502,2QEH_SRO_A_145,2QFO_A13_A_900,2QFO_A51_A_999,2QFQ_S3P_A_701,2QFS_S3P_A_701,2QFT_GPJ_A_801,2QFT_S3P_A_701,2QFU_GPJ_A_801,2QFU_S3P_A_701,2QHD_DAO_B_502,2QHM_7CS_A_500,2QM7_GDP_A_601,2QM9_TDZ_A_201,2QNQ_QN3_A_2501,2QPQ_CIT_A_302,2QPQ_CIT_C_302,2QPU_QPU_A_3000,2QPU_QPU_C_3000,2QRY_TPS_A_400,2QRY_TPS_B_400,2QRY_TPS_C_400,2QS1_UB1_A_603,2QS2_UBF_A_603,2QS2_UBF_B_602,2QS3_UBE_A_603,2QS3_UBE_B_602,2QS4_LY5_C_259,2QTA_TDP_A_887,2QTA_TDP_B_887,2QTN_NCN_B_191,2QTR_NXX_A_191,2QTR_NXX_B_192,2QTR_NXX_C_191,2QVD_BER_A_1811,2QVU_FDP_A_338,2QWL_ADP_B_486,2QWM_ADP_A_486,2QWM_ADP_B_486,2QWO_ADP_A_487,2QWP_ADP_A_487,2QWQ_ADP_A_487,2QWX_ML1_A_233,2QWX_ML1_B_233,2QX4_ML1_A_233,2QX6_ML1_A_233,2QX8_ML2_A_235,2QX8_ML2_B_233,2QX9_ML2_A_233,2QX9_ML2_A_234,2R09_4IP_A_405,2R09_4IP_B_400,2R0D_4IP_A_402,2R0D_4IP_B_400,2R0H_CTO_A_165,2R0H_CTO_C_165,2R0H_CTO_D_165,2R14_TXD_A_600,2R3D_ACM_A_303,2R58_MLY_A_436,2R5A_MLZ_A_436,2R5N_R5P_A_670,2R5P_MK1_B_902,2R5P_MK1_D_902,2R75_01G_1_614,2RA6_ETY_A_604,2RA6_ETY_B_601,2RA6_ETY_C_603,2RCB_DSN_A_901,2RCN_GDP_A_600,2RD6_78P_A_900,2RDE_C2E_A_301,2REG_CHT_A_1,2REG_CHT_B_1,2RFY_CBI_A_431,2RFY_CBI_B_431,2RIN_ACH_A_1,2RIO_ADP_A_1101,2RIO_ADP_B_2101,2RKU_R78_A_500,2RTD_BTN_B_300,2RTE_BTN_B_300,2RTF_BTN_B_300,2RTF_BTN_D_301,2RTG_BTN_B_300,2RTG_BTN_D_301,2STD_CRP_A_175,2UVI_UNG_A_1431,2UWD_2GG_A_1228,2UXI_G50_A_1211,2UXI_G50_B_1211,2UXI_G50_B_1212,2UXU_NAR_A_1211,2UXU_NAR_B_1211,2UYN_2KT_A_1129,2UYN_2KT_B_1129,2UYN_2KT_C_1129,2UYW_BTN_A_1123,2UYW_BTN_D_1124,2UZJ_E64_B_1254,2UZT_SS3_A_1351,2V00_V15_A_1336,2V3U_DSN_A_1262,2V57_PRL_A_1188,2V58_LZJ_A_1448,2V58_LZJ_B_1447,2V59_LZK_A_1446,2V59_LZK_B_1446,2V5V_FMN_A_170,2V5X_V5X_A_1380,2V5X_V5X_B_1380,2V7A_627_A_1504,2V7A_627_B_1504,2V8Q_AMP_E_1327,2V8Q_AMP_E_1328,2V8Q_AMP_E_1329,2V8W_MGO_A_1218,2V8X_MGQ_A_1218,2V8X_MGQ_E_1218,2V8Y_MGV_A_1218,2V8Y_MGV_E_1218,2V92_AMP_E_1329,2V92_ATP_E_1327,2V92_ATP_E_1328,2VB8_TLM_A_1405,2VB8_TLM_B_1405,2VB8_TLM_C_1405,2VB8_TLM_D_1405,2VBA_P4T_A_1405,2VC9_NOK_A_1909,2VCB_OAN_A_1911,2VCI_2GJ_A_1225,2VCW_ZZA_B_1201,2VCX_D26_B_1201,2VCX_D26_C_1202,2VCX_D26_D_1201,2VE3_REA_A_1445,2VE3_REA_B_1445,2VFK_AMP_A_1293,2VHJ_ADP_A_1301,2VHJ_ADP_B_1301,2VHJ_ADP_C_1301,2VHQ_ATP_A_1327,2VHQ_ATP_B_1327,2VHQ_ATP_C_1327,2VHW_NAI_A_1373,2VHW_NAI_B_1372,2VHW_NAI_C_1372,2VHW_NAI_D_1371,2VHW_NAI_E_1374,2VHW_NAI_F_1374,2VK2_GZL_A_1298,2VKU_DBE_A_1446,2VKU_DBE_A_1447,2VKU_DBE_A_1449,2VMC_A2G_A_501,2VMC_NGA_A_505,2VMD_MBG_A_501,2VMF_MVL_A_1865,2VMF_MVL_B_1868,2VNH_NAP_A_274,2VNI_A2P_A_274,2VNJ_NAP_A_274,2VNV_MMA_A_201,2VNV_MMA_B_201,2VNV_MMA_C_201,2VNV_MMA_D_201,2VNV_MMA_E_201,2VO4_4NM_A_1221,2VPN_4CS_A_1311,2VPN_4CS_B_1311,2VPO_6CS_A_1312,2VQZ_MGT_A_1483,2VQZ_MGT_B_1483,2VQZ_MGT_D_1483,2VQZ_MGT_E_1483,2VQZ_MGT_F_1483,2VRJ_NCW_A_1446,2VSR_9HO_A_1477,2VSR_9HO_A_1478,2VSR_9HO_B_1476,2VT3_ATP_A_302,2VT3_ATP_B_302,2VTD_LKM_A_1441,2VTZ_COA_A_1393,2VTZ_COA_B_1393,2VTZ_COA_C_1393,2VTZ_COA_D_1393,2VU3_LZE_A_1299,2VUK_P83_B_1291,2VVN_NHT_A_1718,2VVN_NHT_B_1716,2VW5_BC6_A_500,2VW5_BC6_C_500,2VW5_BC6_D_500,2VXN_PGH_A_1251,2VZM_NRB_A_1408,2VZM_NRB_B_1408,2VZR_GCU_A_1129,2VZR_GCU_B_1129,2W08_TPO_A_500,2W08_TPO_B_500,2W08_TPO_C_500,2W08_TPO_D_500,2W08_TPO_E_500,2W0B_CMW_A_1450,2W4X_STZ_A_1591,2W87_GCU_A_1150,2W8J_PLS_A_1420,2W8W_PLS_A_1420,2W8Y_486_A_1000,2W8Y_NDR_B_1000,2W95_NGA_A_254,2W95_NGA_C_254,2W97_GOL_A_1222,2W97_GOL_B_1220,2W9G_NDP_A_1160,2W9H_TOP_A_1160,2W9S_NDP_B_1159,2W9S_NDP_C_1159,2W9S_TOP_B_1160,2W9S_TOP_D_1158,2W9S_TOP_E_1160,2W9S_TOP_F_1159,2WBD_RO5_B_1337,2WBD_RO5_C_1337,2WBD_RO5_D_1337,2WBD_RO5_E_1337,2WBD_RO5_F_1337,2WBD_RO5_G_1337,2WBD_RO5_H_1337,2WC4_AMF_A_1446,2WC4_AMF_C_1446,2WCV_FUC_B_1141,2WCV_FUC_D_1141,2WCV_F

UC_E_1141,2WCV_FUC_F_1141,2WCV_FUC_G_1141,2WCV_FUC_I_1141,2WCV_FUC_J_1141,2WD8_VGF_A_1270,2WD8_VGF_B_1269,2WD8_VGF_C_1270,2WD8_VGF_D_1269,2WDO_ACO_A_1125,2WDY_COA_A_1128,2WE0_UMP_A_1257,2WE3_DUT_A_1257,2WEG_FBV_A_1263,2WEH_FB1_A_1263,2WEJ_FB2_A_1263,2WEQ_GDM_A_1215,2WER_RDC_A_1215,2WER_RDC_B_1215,2WF5_BG6_A_1831,2WGE_TLM_A_1424,2WGG_TLM_A_1417,2WGG_TLM_C_1417,2WGG_TLM_E_1417,2WGG_TLM_F_1417,2WGG_TLM_G_1419,2WGG_TLM_H_1417,2WH8_II2_A_460,2WH8_II2_D_460,2WHF_II4_A_1405,2WHF_II4_A_1406,2WHW_1D4_A_1410,2WHW_1D4_B_1410,2WI9_1D2_A_1409,2WIC_GNP_A_1265,2WJG_GDP_A_200,2WJG_GDP_B_200,2WJI_GNP_A_1166,2WJI_GNP_B_1166,2WKT_COA_A_1397,2WKT_COA_B_1398,2WKT_COA_C_1394,2WKT_COA_D_1395,2WKV_COA_A_1397,2WKV_COA_B_1399,2WKV_COA_C_1393,2WKV_COA_D_1395,2WL4_COA_A_1402,2WL4_COA_B_1401,2WL5_COA_A_1399,2WL5_COA_B_1399,2WN9_ZY5_B_301,2WN9_ZY5_C_301,2WN9_ZY5_D_301,2WN9_ZY5_E_301,2WNC_TKT_A_300,2WNC_TKT_C_300,2WNC_TKT_D_300,2WNC_TKT_E_300,2WNJ_ZY7_A_301,2WNJ_ZY7_B_301,2WNJ_ZY7_C_301,2WNJ_ZY7_D_301,2WOR_2AN_A_1098,2WOS_6AN_A_1098,2WQ4_SFU_B_200,2WQ4_SFU_C_200,2WQ5_MIY_A_1120,2WSA_646_A_1423,2WSV_LBT_A_1685,2WTV_ZZL_A_1390,2WTV_ZZL_B_1390,2WUF_KEM_B_1289,2WUZ_TPF_A_1460,2WVA_TPU_A_600,2WVA_TPU_B_600,2WVA_TPU_E_600,2WVA_TPU_F_600,2WVA_TPU_V_600,2WVA_TPU_Y_600,2WVA_TPU_Z_600,2WX2_TPF_A_1460,2WX2_TPF_B_1460,2WXP_GD9_A_1500,2WYF_MBG_A_202,2WYF_MBG_C_202,2WYF_MBG_D_202,2WYF_MBG_F_202,2WYF_MBG_G_202,2WYF_MBG_H_202,2WZG_UPG_A_1525,2X00_GYN_A_301,2X00_GYN_B_301,2X00_GYN_E_301,2XAB_VHD_A_1224,2XAB_VHD_B_1224,2XB7_GUI_A_1501,2XBK_XBK_A_1399,2XBP_ATP_A_1114,2XBV_XBV_A_1245,2XC3_RT8_A_1434,2XCG_XCG_B_602,2XDE_1B0_A_1148,2XDE_1B0_B_1147,2XDK_XDK_A_1226,2XDL_2DL_A_1225,2XDU_LGA_A_1227,2XDX_WOE_A_1224,2XE6_3PG_A_1417,2XFH_CL6_A_1413,2XFH_CL6_A_1414,2XGS_44P_A_1569,2XGS_44P_B_1569,2XHR_C0P_A_1001,2XHT_C0Y_A_1224,2XI7_XI7_A_1186,2XI7_XI7_B_1186,2XIB_DFU_A_1472,2XIB_DFU_C_1472,2XIB_DFU_D_1472,2XJ2_985_A_1001,2XJ7_GC2_A_1000,2XJ7_GC2_B_1000,2XJG_XJG_A_1224,2XJJ_L81_B_1225,2XJX_XJX_A_1232,2XK2_ADP_A_1226,2XM1_LTM_A_1717,2XM1_LTM_B_1716,2XN3_ID8_A_1356,2XN5_FUN_A_1356,2XNJ_FAD_A_1257,2XNJ_FAD_B_1257,2XNJ_NAP_A_1258,2XNJ_NAP_B_1258,2XP2_VGH_A_9000,2XP8_4FY_A_1165,2XTM_GDP_A_1228,2XTM_GDP_B_1227,2XTN_GTP_A_1231,2XUZ_EB4_A_1300,2XV1_ECA_A_1301,2XXT_KAI_A_900,2XXT_KAI_B_900,2XXV_KAI_A_900,2XXV_KAI_B_900,2XYA_7L4_A_1181,2XZW_AKG_A_201,2XZW_AKG_B_201,2XZW_AKG_F_201,2XZW_AKG_H_201,2XZW_AKG_I_201,2Y1D_34F_A_1390,2Y1W_849_A_1005,2Y1W_849_B_1005,2Y1W_849_C_1005,2Y1W_849_D_1005,2Y33_UN9_A_901,2Y34_UN9_A_601,2Y3F_BTN_A_1136,2Y4K_GDP_A_401,2Y4K_GDP_B_401,2Y4S_BCD_A_1895,2Y5A_3AP_A_1295,2Y6O_1N1_A_1892,2Y7J_B49_A_1294,2Y7J_B49_B_1294,2Y7J_B49_C_1294,2Y7J_B49_D_1294,2Y7K_SAL_A_1302,2Y7K_SAL_A_1303,2Y7K_SAL_B_1304,2Y7K_SAL_B_1305,2Y7K_SAL_C_1302,2Y7K_SAL_D_1300,2Y7P_SAL_A_1000,2Y8L_ADP_E_1327,2Y8L_ADP_E_1328,2Y9G_LAT_A_1151,2Y9G_LAT_A_1153,2YAY_DUP_A_1265,2YAZ_UMP_A_1265,2YAZ_UMP_D_1265,2YAZ_UMP_E_1266,2YB0_DUR_A_1266,2YB0_DUR_B_1266,2YB0_DUR_D_1266,2YB0_DUR_E_1266,2YCH_ATP_A_502,2YDW_WSH_B_1182,2YDW_WSH_C_1183,2YEK_EAM_A_1188,2YEK_EAM_B_1183,2YEL_WSH_A_1170,2YEM_WSH_A_1459,2YEM_WSH_B_1460,2YFA_LMR_B_501,2YFE_YFE_A_999,2YFX_VGH_A_9000,2YGA_GDM_A_1215,2YGE_GDM_A_1215,2YGF_GDM_A_1215,2YHW_BM3_A_1718,2YHY_BM3_A_1000,2YI5_YI5_A_1226,2YI6_6QM_A_1225,2YI7_BZ8_A_1225,2YIM_MC4_A_1363,2YIM_MC4_B_1363,2YIM_MC4_D_1363,2YIS_YIS_A_1353,2YIW_YIW_A_1353,2YIX_YIX_A_1355,2YK1_NCT_H_300,2YLD_CMO_A_129,2YLG_CMO_A_129,2YMD_SRO_A_1213,2YMD_SRO_B_1212,2YMD_SRO_C_1213,2YMD_SRO_D_1214,2YMD_SRO_F_1214,2YMD_SRO_G_1213,2YMD_SRO_I_1213,2YME_CWB_A_1207,2YME_CWB_B_1207,2YME_CWB_C_1207,2YME_CWB_D_1207,2YME_CWB_E_1207,2YME_CWB_F_1207,2YME_CWB_H_1207,2YME_CWB_J_1207,2YOO_K2B_A_1404,2YOO_K2B_B_1404,2YOO_K2B_C_1404,2YOO_K2B_D_1404,2YXS_LAT_A_401,2YY1_NLC_A_161,2Z8E_MES_

A_11,2Z8E_MES_B_12,2Z92_ENE_A_1,2Z93_END_C_1,2Z97_7HE_A_417,2Z97_CAM_A_422,2ZB1_GK4_A_361,2ZBF_TG1_A_1003,2ZDK_50U_A_501,2ZDM_46U_A_501,2ZDN_49U_A_501,2ZFT_10U_A_501,2ZHD_12U_A_501,2ZIF_SAM_B_298,2ZKJ_ADP_B_501,2ZM1_KSF_A_513,2ZM3_575_A_1,2ZM3_575_C_1,2ZQ1_11U_A_501,2ZQ2_13U_A_501,2ZRU_FMN_A_669,2ZRU_FMN_B_669,2ZRU_FMN_C_669,2ZRU_FMN_D_669,2ZSC_BTN_A_301,2ZSC_BTN_B_301,2ZWZ_ZWZ_A_901,2ZWZ_ZWZ_B_902,2ZYN_BCD_A_901,3A1C_ACP_B_997,3A1D_ADP_A_997,3A1E_ACP_A_997,3A1E_ACP_B_997,3A1S_GDP_A_1,3A1S_GDP_B_2,3A1T_GDP_A_1,3A20_FMN_A_123,3A20_FMN_B_123,3A5Y_KAA_A_1990,3A5Y_KAA_B_1991,3A6Q_FMN_A_123,3A6Q_FMN_B_123,3A6R_FMN_B_123,3A6R_FMN_D_123,3A6T_8OG_A_134,3A73_PJ2_A_1012,3A73_PJ2_B_1011,3AAS_GUS_A_1001,3ACL_3F1_A_302,3AD8_PYC_B_801,3ADS_IMN_B_3,3ADV_SRO_A_1,3ADV_SRO_A_3,3ADV_SRO_B_4,3ADW_MYI_A_3,3AJ5_NGA_A_2000,3AJ5_NGA_B_2000,3AKM_11D_B_132,3AKM_11D_C_132,3ALT_MLB_A_301,3ALT_MLB_B_301,3ALT_MLB_D_301,3AM7_MGP_A_1000,3AMA_SKE_A_351,3AMB_VX6_A_351,3AO1_BZX_A_213,3AO1_BZX_A_214,3AO1_BZX_B_1,3AO2_AVX_A_1,3AO2_AVX_B_1,3AO4_833_A_1,3AO4_833_B_1,3AO5_BBY_A_213,3AOX_EMH_A_901,3AP4_LBT_A_155,3AP4_LBT_B_155,3AP4_LBT_C_155,3AP6_LAT_A_155,3AP6_LAT_B_155,3AP6_LAT_C_155,3AP7_LAT_A_155,3AQA_BYH_A_201,3AQT_RCO_A_230,3AQT_RCO_B_230,3AR3_TG1_A_1003,3AR4_ATP_A_1002,3AR5_TG1_A_1003,3AR5_TM1_A_1002,3AR6_TG1_A_1003,3AR7_TG1_A_1003,3ARP_DEQ_A_606,3ARQ_DM5_A_606,3ARR_PNX_A_606,3ARR_PNX_A_607,3ART_DEQ_A_606,3ART_DEQ_A_607,3ART_DEQ_A_608,3ARU_PNX_A_606,3ARU_PNX_A_607,3ARU_PNX_A_608,3ARV_SAU_A_607,3ARX_POY_A_606,3ARX_POY_A_607,3ARY_I5I_A_606,3ARY_I5I_A_607,3ARZ_I5I_A_606,3ARZ_I5I_A_607,3AS0_SAU_A_606,3AS0_SAU_A_607,3AT4_CCK_A_336,3ATU_ADP_A_1452,3ATV_ADP_A_1452,3AXZ_ADN_A_401,3AY9_ADP_A_1452,3AYE_LAT_A_0,3B0W_DGX_A_1,3B0W_DGX_B_1,3B24_B2J_A_237,3B25_B2K_A_1,3B27_B2T_A_1,3B28_B2X_A_237,3B2Q_ATP_A_461,3B2T_M33_A_1996,3B2T_M33_B_1996,3B4P_3B4_A_500,3B4P_3B4_B_500,3B4Y_FLC_B_337,3B50_SLB_A_311,3B5J_12D_A_708,3B6H_MXD_A_551,3B78_NAD_B_700,3B78_NAD_D_701,3B78_NAD_F_702,3B82_NAD_B_700,3B82_NAD_D_701,3B82_NAD_F_702,3B8H_NAD_B_700,3B8H_NAD_D_701,3B99_U51_A_700,3B99_U51_A_701,3B99_U51_A_702,3B99_U51_B_700,3B99_U51_B_701,3BBF_GDP_B_160,3BBF_GDP_C_160,3BBF_GDP_D_160,3BBF_GDP_E_160,3BBF_GDP_F_160,3BEN_LEH_A_500,3BEN_LEH_B_500,3BEX_PAU_A_248,3BEX_PAU_B_248,3BEX_PAU_C_248,3BEX_PAU_D_248,3BEX_PAU_E_248,3BF1_ADP_A_247,3BF1_ADP_B_247,3BF1_ADP_C_247,3BF1_ADP_D_247,3BF1_ADP_E_247,3BF1_ADP_F_247,3BF1_PAU_B_248,3BF1_PAU_C_248,3BF1_PAU_E_248,3BF1_PAU_F_248,3BIR_2GP_A_105,3BJM_BJM_A_1,3BJM_BJM_B_2,3BL1_BL1_A_300,3BPX_SAL_A_257,3BRA_AEF_A_394,3BRN_SRO_A_158,3BRN_SRO_B_158,3BU1_HSM_A_301,3BUF_AEG_A_394,3BUG_AEH_A_394,3BUH_AED_A_394,3BVA_2NC_B_200,3BVB_017_B_401,3BWJ_ARX_A_352,3BWK_C1P_A_300,3BWK_C1P_B_300,3BWK_C1P_C_300,3BWK_C1P_D_300,3BXE_13P_B_401,3BXF_13P_B_401,3BXF_FBP_A_401,3BXG_BG6_A_401,3BXG_BG6_B_401,3BXH_F6P_A_401,3BXH_F6P_B_401,3BYO_AM9_A_1,3C2R_PHT_A_300,3C2R_PHT_B_301,3C2V_PRP_A_300,3C4H_DRL_A_601,3C79_IM4_B_302,3C79_IM4_C_303,3C7Q_XIN_A_1172,3C84_TH4_A_301,3C84_TH4_B_302,3C84_TH4_C_303,3C84_TH4_D_304,3C8E_GSH_A_289,3C8E_GSH_A_290,3C8E_GSH_B_290,3C9E_E64_A_308,3CAJ_CL_A_263,3CAJ_EZL_A_265,3CAQ_NDP_B_327,3CB0_FMN_A_200,3CCT_3HI_B_1,3CCT_3HI_B_2,3CCT_3HI_C_4,3CCW_4HI_B_1,3CCW_4HI_B_2,3CCW_4HI_D_3,3CCZ_5HI_A_876,3CCZ_5HI_B_876,3CCZ_5HI_C_876,3CCZ_5HI_D_876,3CD0_6HI_B_1,3CD0_6HI_B_2,3CD0_6HI_C_4,3CD0_6HI_D_3,3CD5_7HI_B_1,3CD5_7HI_D_3,3CD5_7HI_D_876,3CD7_882_A_2,3CD7_882_C_4,3CD7_882_D_3,3CDA_8HI_A_2,3CDA_8HI_B_1,3CDA_8HI_C_4,3CDA_8HI_D_3,3CDB_9HI_A_1,3CDB_9HI_A_2,3CDB_9HI_C_4,3CDB_9HI_D_3,3CEM_AVD_A_833,3CEM_AVD_B_833,3CFN_2AN_A_1,3CFN_2AN_B_1,3CFQ_DIF_A_1,3CFQ_DIF_B_1,3CFT_5NS_A_1,3CJ2_SX3_A_571,3CJ2_SX3_B_571,3CJ4_SX5_A_571,3CJ4_SX5_B_571,3CJ5_SX6_A_571,3CJ5_SX6_B_571,3CK7_ACX_B_688,3CK7_ACX_C_689,3CK8_BCD_A_800,3CK8_BCD_B_900,3CKB_MLR_A_903,3CKB_MLR_B_904,3CLA_CLM_A_221,3CLP_CMP_A_1,3CLP_CMP_C_2,

3COW_52H_B_402,3COY_53H_A_301,3COY_53H_B_302,3COZ_54H_A_401,3COZ_54H_B_402,3CR4_PNT_X_102,3CR5_PNT_X_94,3CRZ_NAP_A_305,3CS9_NIL_A_600,3CS9_NIL_B_600,3CS9_NIL_C_600,3CUK_4P5_A_402,3CUK_4P5_B_402,3CUK_4P5_C_402,3CUK_4P5_D_402,3CWK_REA_A_300,3CYU_0CR_A_263,3CYW_017_A_201,3CYX_ROC_A_201,3CYZ_9OD_A_124,3CYZ_9OD_B_121,3CZ0_9OD_A_124,3CZ0_9OD_B_121,3CZ0_9OD_B_122,3CZ1_NBB_A_122,3CZ1_NBB_B_122,3CZ1_NBB_B_123,3D0B_SNX_A_233,3D0E_G93_A_1,3D0E_G93_B_2,3D1Y_ROC_A_201,3D1Z_017_B_201,3D2E_ATP_C_1001,3D2R_ADP_A_501,3D2R_ADP_B_501,3D2T_1FL_A_502,3D2T_1FL_B_500,3D4F_LN1_A_601,3D76_NBB_A_121,3D78_NBB_A_120,3D78_NBB_B_120,3D78_NBB_B_121,3D7F_YBY_A_1,3D7K_D7K_A_572,3D7K_D7K_B_572,3D91_REM_A_350,3D91_REM_B_350,3DA9_44U_B_1,3DC3_AZM_A_263,3DCW_EZL_A_301,3DD0_EZL_A_301,3DDQ_RRC_A_299,3DDQ_RRC_C_299,3DF9_DF9_A_233,3DF9_DF9_B_233,3DGL_ATP_A_81,3DGO_ATP_A_82,3DKF_SX8_A_1,3DKG_SX8_A_1,3DKJ_PRP_A_501,3DLA_NXX_A_680,3DLA_NXX_B_680,3DLA_NXX_C_680,3DMX_BNZ_A_900,3DNT_ATP_A_500,3DNT_ATP_B_501,3DP4_AMQ_A_374,3DP9_BIG_A_301,3DP9_BIG_C_301,3DRC_MTX_A_161,3DRC_MTX_B_361,3DST_GRG_B_334,3DSZ_LIZ_A_187,3DZT_EAH_A_305,3E3C_HHG_A_150,3E3C_HHG_B_150,3E5A_VX6_A_500,3E5U_3C4_A_501,3E5U_3C4_A_505,3E5U_3C4_B_503,3E5U_3C4_C_504,3E5U_3C4_D_502,3E64_5B3_A_1,3E6V_ABH_A_552,3E6Y_CW1_A_261,3E6Y_CW1_B_261,3E7F_6PG_A_843,3E7F_6PG_B_843,3E85_BSU_A_158,3E85_BSU_A_159,3E85_BSU_A_160,3E85_BSU_A_161,3EBL_GA4_A_401,3EBL_GA4_B_401,3EBL_GA4_C_401,3EBL_GA4_D_401,3EBL_GA4_E_401,3EBL_GA4_F_401,3EBS_N4E_A_1,3EBS_N4E_B_1,3EBS_N4E_D_1,3ED0_EMO_A_163,3ED0_EMO_C_161,3EEB_IHP_A_210,3EFL_706_A_501,3EFL_706_B_502,3EFS_ATP_A_1001,3EFS_BTN_B_2001,3EKD_PAM_A_471,3EKP_478_C_200,3EKT_017_B_200,3EKT_017_D_200,3EKV_478_A_200,3EKW_DR7_B_100,3EKX_1UN_B_201,3EKY_DR7_A_100,3EL0_1UN_A_201,3EL4_ROC_A_100,3EL5_1UN_B_201,3EL9_DR7_A_100,3ELC_F01_A_901,3ELC_F01_B_902,3ELC_F01_C_903,3EM3_478_B_200,3EM4_DR7_A_100,3EM4_DR7_V_100,3EQG_4BM_A_1,3ERN_CAR_B_900,3ERN_CAR_C_900,3ERN_CAR_D_900,3ERN_CAR_E_900,3EUF_BAU_A_400,3EUF_BAU_B_400,3EUF_BAU_C_400,3EVD_GTP_A_901,3EW2_BTN_A_400,3EW2_BTN_C_600,3EW2_BTN_D_700,3EW2_BTN_F_900,3EW2_BTN_G_350,3EXE_TPP_E_1011,3EXE_TPP_G_1008,3EXH_TPP_A_1502,3EXH_TPP_C_1502,3EXH_TPP_E_1502,3EXH_TPP_G_1502,3EYG_MI1_A_1,3F15_HS1_A_0,3F16_HS3_A_0,3F17_HS4_A_0,3F18_HS5_A_0,3F1A_HS7_A_0,3F33_PFL_A_2001,3F34_DIE_A_2001,3F35_DIE_A_2001,3F36_IP0_A_2001,3F37_2MY_A_2001,3F38_2MY_A_2001,3F3T_1AU_A_1,3F3T_1AU_A_534,3F3U_1AW_A_1,3F3U_1AW_A_534,3F6E_8PA_X_602,3F78_ICF_A_1,3F78_ICF_B_2,3F80_6HN_A_551,3F80_6HN_A_552,3F8F_DM1_A_127,3FAT_AMQ_A_427,3FAT_AMQ_C_429,3FC1_52P_X_362,3FEG_HC7_A_1,3FFA_GSP_A_1,3FGC_FMN_A_3402,3FHB_GAB_A_601,3FHE_28P_A_612,3FJ7_PEQ_A_301,3FJ7_PEQ_B_301,3FJG_3PG_A_301,3FJG_3PG_B_301,3FJG_3PG_C_301,3FJG_3PG_D_301,3FJX_S3P_A_428,3FK1_GPF_A_429,3FK1_S3P_A_428,3FK7_4DM_A_210,3FK7_4DM_B_210,3FLS_FLS_A_361,3FLW_FLW_A_361,3FQS_585_A_1,3FSJ_D7K_X_601,3FSK_RO6_A_450,3FSM_2NC_A_204,3FTX_BES_A_720,3FTX_RE2_A_710,3FUC_9D9_B_301,3FUE_11S_A_710,3FUE_BES_A_720,3FUH_5H1_A_710,3FUH_BES_A_720,3FUP_MI1_A_1,3FUP_MI1_B_1,3FW4_CAQ_A_180,3FW4_CAQ_C_180,3FZL_3FD_A_401,3FZN_D7K_B_602,3FZN_D7K_C_604,3FZY_IHP_A_8000,3FZY_IHP_B_8001,3G08_FEE_A_286,3G15_HC6_A_603,3G15_HC6_B_603,3G3E_G3E_B_353,3G3E_G3E_C_353,3G3E_G3E_D_353,3G5D_1N1_A_1,3G5D_1N1_B_1,3GA5_RGG_A_400,3GAY_P6T_A_327,3GAY_P6T_B_329,3GCP_SB2_A_361,3GCS_BAX_A_401,3GCU_R48_B_401,3GCV_SS6_A_361,3GDQ_ADP_A_388,3GF2_SAL_A_147,3GFD_IYR_B_302,3GHW_GHW_A_187,3GK1_32A_A_94,3GK2_27A_A_94,3GK4_53A_X_94,3GKZ_B40_A_500,3GM0_B41_A_600,3GNI_ATP_B_1,3GNW_XNC_A_579,3GNW_XNC_B_579,3GP0_NIL_A_1,3GPB_G1P_A_900,3GPB_G1P_A_901,3GPO_APR_A_161,3GPO_APR_B_161,3GPO_APR_C_161,3GPO_APR_D_161,3GR6_NAP_A_361,3GR6_NAP_D_257,3GR6_NAP_G_257,3GR6_NAP_J_257,3GR6_TCL_A_371,3GR6_TCL_D_258,3GR6_TCL_G_258,3GUK_MBN_A_501,3GUK_MBN_B_501,3GUM_PXY_A_501,3GUM_PXY_B_501,3GUO_IPH_A_501,3GUO_IPH_B_501,3GUS_N11_B_211,3GUZ_PAF_A_178,3GUZ_PAF_B

_177,3GVU_STI_A_1001,3GVU_STI_A_1002,3GW9_VNI_A_490,3GW9_VNI_C_490,3GWS_T3_X_500,3GX0_GDS_A_301,3GY2_BRN_A_1,3GY3_PNT_A_246,3GY4_PBZ_A_1,3GY5_BRN_A_246,3GY6_BRN_A_1,3GY7_BEN_A_1,3GY8_BRN_A_1,3GZ8_APR_A_302,3GZ8_APR_B_302,3GZ8_APR_C_302,3H0A_9RA_A_500,3H0A_D30_D_500,3H1X_IMN_A_301,3H26_B56_B_902,3H2K_BOG_A_398,3H2K_BOG_A_399,3H78_BE2_A_350,3H78_BE2_B_350,3H78_BE2_B_351,3H7R_NAP_A_401,3H7U_NAP_A_401,3H7W_018_A_1,3H82_020_A_351,3HAU_2NC_A_0,3HAW_2NC_B_0,3HBO_2NC_A_0,3HCM_S45_A_94,3HDK_2NC_B_0,3HDQ_GDU_A_500,3HDQ_GDU_B_500,3HDQ_GDU_C_500,3HDQ_GDU_D_500,3HDQ_GDU_F_500,3HDQ_GDU_G_500,3HDQ_GDU_H_500,3HDQ_GDU_I_500,3HDQ_GDU_J_500,3HDY_GDU_A_500,3HDY_GDU_B_500,3HDY_GDU_C_500,3HDY_GDU_D_500,3HDY_GDU_E_500,3HDY_GDU_F_500,3HDY_GDU_H_500,3HDY_GDU_I_500,3HDY_GDU_J_500,3HEC_STI_A_1,3HEG_BAX_A_1,3HGM_ATP_A_148,3HGM_ATP_B_148,3HGM_ATP_C_148,3HGM_ATP_D_148,3HJO_EAA_A_211,3HKU_TOR_A_300,3HKW_IX6_A_579,3HKW_IX6_B_579,3HKW_IX6_C_579,3HKY_IX6_B_579,3HMO_STU_A_1,3HP5_52P_A_401,3HP8_SUC_A_108,3HP8_SUC_A_109,3HP8_SUC_B_109,3HPI_SUC_A_371,3HPI_SUC_B_375,3HQR_OGA_A_600,3HS4_AZM_A_701,3HS4_AZM_A_702,3HS4_AZM_A_703,3HS5_ACD_A_1,3HT6_JZ0_A_167,3HT8_JZ2_A_167,3HTB_JZ4_A_167,3HU3_AGS_B_800,3HU8_261_A_167,3HUA_JZ9_A_165,3HUK_J0Z_A_166,3HUQ_J1Z_A_164,3HV6_R39_A_361,3HV8_C2E_A_900,3HVG_EV0_A_441,3HVG_EV0_B_441,3HVG_EV0_B_442,3HWW_AKG_A_557,3HX3_RET_A_400,3HXG_GTA_A_1,3HXI_GTG_A_501,3HY8_FMN_A_390,3HZF_B72_A_1,3HZL_6PC_A_399,3HZM_KDO_B_303,3I33_ADP_A_388,3I73_ADP_A_589,3I7S_PYR_B_302,3I9G_S1P_L_401,3IAI_AZM_B_263,3IAI_AZM_C_263,3IAI_AZM_D_263,3IAI_PO4_A_600,3IAI_PO4_A_601,3IAI_PO4_B_600,3IAI_PO4_C_600,3IAI_PO4_D_600,3IAW_2NC_A_204,3IAZ_AIN_A_1202,3IB1_IMN_A_701,3IB2_IBP_A_3960,3IE3_N11_A_211,3IE3_N11_B_211,3IJG_AVR_C_200,3IJH_KO2_B_214,3IJH_KO2_C_214,3IK3_0LI_A_1,3IK3_0LI_B_2,3IKF_717_A_163,3IKF_717_A_202,3IKF_717_B_202,3ILZ_B72_A_1,3IMC_BZ3_A_702,3IMC_BZ3_B_701,3IME_BZ2_A_301,3IME_BZ2_B_302,3IMG_BZ2_A_301,3IMG_BZ2_B_302,3IMG_BZ3_A_401,3IOB_A4D_A_301,3IOB_A4D_B_302,3IOC_A5D_A_301,3IOC_A5D_B_302,3IOD_A6D_A_302,3IOD_A6D_B_301,3IOE_A7D_A_901,3IOE_A7D_B_902,3IP9_ABU_A_357,3ISJ_A8D_B_305,3IUB_FG2_A_303,3IUB_FG2_B_302,3IUC_ADP_A_1,3IUC_ADP_C_1,3IUE_FG3_A_302,3IUE_FG3_A_303,3IUE_FG3_B_302,3IVC_FG4_B_302,3IVG_FG5_B_305,3IVX_FG6_A_302,3IVX_FG6_A_303,3IVX_FG6_B_302,3IW2_EKO_A_665,3IW5_DF3_A_362,3IW6_PP0_A_361,3IW8_HIZ_A_361,3IX1_NFM_A_401,3IX1_NFM_B_401,3JQM_GTP_A_158,3JQM_GTP_B_158,3JQM_GTP_C_158,3JQM_GTP_D_158,3JQM_GTP_E_158,3JQM_GTP_F_158,3JQM_GTP_G_158,3JQM_GTP_H_158,3JQM_GTP_I_158,3JRS_A8S_A_1,3JRS_A8S_B_2,3JRX_S1A_A_1000,3JUK_UPG_A_282,3JUK_UPG_C_283,3JUN_AJD_A_500,3JUN_AJD_B_500,3JUN_AKD_A_600,3JUN_AKD_B_600,3JUO_AJD_A_500,3JUP_AKD_A_500,3JUP_AKD_B_500,3JUQ_AJD_A_500,3JUQ_AJD_B_500,3JUQ_AKD_B_600,3JVS_AGY_A_900,3JVW_DMP_A_323,3JVY_017_B_401,3JW2_017_A_401,3JXU_ADP_A_486,3K00_MTT_A_405,3K02_TXT_A_405,3K1J_ADP_B_642,3K2P_JTH_B_1001,3K4V_ROC_B_201,3K4V_ROC_D_201,3K54_1N1_A_1,3K5I_AIR_A_402,3K5I_AIR_C_402,3K5I_AIR_D_402,3K5K_DXQ_A_2000,3K5K_DXQ_B_2001,3K5V_STI_A_2,3K5V_STJ_A_1,3K7K_KCS_A_301,3K87_FAD_B_500,3K8O_229_E_9006,3K8O_229_Q_9002,3K8O_229_S_9005,3K8O_229_T_9004,3K8O_229_U_9003,3K8O_229_Y_9001,3K98_1RC_A_901,3K98_1RC_B_901,3KA2_2NC_A_204,3KAF_4D9_A_164,3KDC_JZP_A_100,3KDD_JZQ_A_100,3KDM_TES_H_226,3KE1_829_A_164,3KE1_829_B_164,3KGT_GEN_A_3071,3KGU_GEN_A_3071,3KGU_GEN_B_3071,3KJD_78P_A_1,3KMR_EQN_A_1,3KMX_G00_A_501,3KMX_G00_B_501,3KMY_D8Y_A_501,3KMY_D8Y_B_501,3KN0_3TO_A_501,3KN0_3TO_B_501,3KPU_ES1_A_290,3KPU_SAH_A_2001,3KPU_SAH_B_2002,3KPV_ADE_A_290,3KPV_ADE_B_290,3KPV_SAH_A_2001,3KPV_SAH_B_2002,3KPW_1SQ_A_290,3KPW_SAH_A_2001,3KPW_SAH_B_2002,3KQM_ES3_A_290,3KQM_ES3_B_290,3KQM_SAH_A_2001,3KQM_SAH_B_2002,3KQO_ES4_A_290,3KQO_ES4_B_290,3KQO_SAH_A_2001,3KQO_SAH_B_2002,3KQP_ES5_B_290,3KQP_SAH_A_2001,3KQR_OPE_A_700,3KQR_OPE_C_2700,3KQR_OPE_D_3700,3KQR_OPE_E_4700,3KQS_AX7_A_290,3KQS_AX7_B_290,3KQS_SAH_A_20

01,3KQT_ES7_A_290,3KQT_ES7_B_290,3KQT_SAH_A_2001,3KQT_SAH_B_2002,3KQW_ES9_A_290,3KQW_ES9_B_290,3KQW_SAH_A_2001,3KQW_SAH_B_2002,3KQY_ES0_A_290,3KQY_SAH_A_3001,3KQY_SAH_B_3002,3KR1_SAH_A_3001,3KR1_SAH_B_3002,3KR1_VGD_A_290,3KR1_VGD_B_290,3KR2_ET2_A_1001,3KR2_ET2_B_1002,3KR2_SAH_A_3001,3KR2_SAH_B_3002,3KRK_ACD_A_700,3KRK_ACD_B_700,3KRQ_3TR_A_607,3KRQ_3TR_A_608,3KRY_3KR_A_1800,3KRY_3KR_B_2800,3KRY_3KR_C_3800,3KV2_NNH_A_901,3KV2_NNH_B_902,3KV4_OGA_A_5798,3KX3_140_A_1470,3KX3_140_B_1470,3KXO_KXO_A_202,3KXO_KXO_B_203,3KYQ_DPV_A_201,3KZL_ACO_A_266,3KZL_ACO_C_266,3L1N_PLM_A_161,3L59_BDJ_A_1,3L5B_BDO_A_1,3L5B_BDO_B_455,3L5M_COU_A_801,3L5M_COU_A_802,3L66_COU_A_803,3L68_COU_A_801,3L68_COU_A_803,3L6H_BET_A_1,3L8S_BFF_A_361,3L8X_N4D_A_361,3LB1_IOL_A_191,3LB1_IOL_B_192,3LB2_BML_A_191,3LB2_BML_B_192,3LB3_4CH_B_192,3LB4_FPN_A_191,3LBD_9CR_A_424,3LBZ_Z89_A_201,3LCU_SAH_A_6732,3LCV_SAM_B_301,3LDP_3P1_A_408,3LDP_3P1_B_502,3LDQ_3P1_A_401,3LE8_2B5_A_301,3LE8_2B5_B_301,3LEN_ZST_A_600,3LEP_388_A_600,3LF0_ATP_A_701,3LF0_ATP_C_701,