Genomic and functional characterization of G protein-coupled receptors in the human pathogen Schistosoma mansoni and the model planarian Schmidtea mediterranea

Mostafa Zamanian

Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Recommended Citation

Genomic and functional characterization of G protein-coupled receptors in the human pathogen *Schistosoma mansoni* and model planarian *Schmidtea mediterranea*

by

Mostafa Zamanian

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Neuroscience

Program of Study Committee:
Timothy A. Day, Major Professor
Thomas J. Baum
Steve A. Carlson
M. Heather W. Greenlee
Michael J. Kimber

Iowa State University
Ames, Iowa
2011

Copyright © Mostafa Zamanian, 2011. All rights reserved.
DEDICATION

I would like to dedicate this thesis to my family. To my mom and dad, who sparked my interest in science as a child and provided a stable and nurturing environment for me to explore my own path. This could not have occurred without their unconditional love and timely wisdom. To my sister, Maryam, for the valuable emotional support she provided throughout.
# TABLE OF CONTENTS

## LIST OF TABLES

- vi

## LIST OF FIGURES

- vii

## ACKNOWLEDGEMENTS

- 1

## 1 GENERAL INTRODUCTION

1.1 Schistosomes: Significant Human Parasites  

1.2 Planarians: Robust Model Organisms  

1.3 Genomics  

1.4 Functional Genomics: RNA Interference  

1.4.1 Canonical RNAi Pathway  

1.4.2 RNAi in Schistosomes  

1.4.3 RNAi in Planaria  

1.5 G Protein-Coupled Receptors  

1.5.1 Signal Transduction  

1.5.2 Export and Regulation  

1.5.3 Phylogenetic Overview  

1.6 In silico GPCR Mining  

1.7 In silico GPCR Classification  

1.8 GPCR Deorphanization  

1.8.1 Difficulties and Pitfalls  

1.9 General Objectives  

- 2

- 4

- 5

- 7

- 9

- 10

- 16

- 17

- 20

- 21

- 22
The Repertoire of G Protein-Coupled Receptors in the Human Parasite Schistosoma mansoni and the Model Organism Schmidtea mediterranea  

Abstract .................................................. 23
2.1 Background ............................................. 24
2.2 Results and Discussion ................................. 27
2.3 Conclusions ............................................. 42
2.4 Acknowledgements ...................................... 44
2.5 Methods ................................................ 44
2.6 Figures .................................................. 47
2.7 Tables .................................................. 62

Novel RNAi-mediated Approach to Probing Platyhelminth G Protein-Coupled Receptors ........................................... 73

Abstract .................................................. 73
3.1 Introduction ............................................. 74
3.2 Results and Discussion ................................. 76
3.3 Conclusions ............................................. 80
3.4 Materials and Methods .................................. 81
3.5 Acknowledgements ...................................... 85
3.6 Figures and Tables ..................................... 85

PROF1 Localization in Planaria .................................... 93

4.1 Planarian In situ hybridization protocol .................. 93
4.2 PROF1 Localization Results ............................. 94

CONCLUSIONS ............................................. 95

APPENDIX A The genome of the blood fluke Schistosoma mansoni ........ 97
APPENDIX B Yeast GPCR Assay ................................ 104
APPENDIX C PROF1 Primers and RT-PCR ..................... 108
APPENDIX D RNAi vector .................................. 112
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Genome assembly statistics.</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Heterotrimeric G protein classification.</td>
<td>11</td>
</tr>
<tr>
<td>1.3</td>
<td>GRAFS species comparison.</td>
<td>15</td>
</tr>
<tr>
<td>2.1</td>
<td>Stage-specific GPCR tabulation.</td>
<td>64</td>
</tr>
<tr>
<td>2.2</td>
<td>GRAFS-based comparison of GPCR repertoires.</td>
<td>65</td>
</tr>
<tr>
<td>2.3</td>
<td>Planarian homologs of parasite GPCRs.</td>
<td>66</td>
</tr>
<tr>
<td>2.4</td>
<td>Planarian homologs of parasite GPCRs- continued.</td>
<td>67</td>
</tr>
<tr>
<td>2.5</td>
<td>Planarian homologs of parasite GPCRs- continued.</td>
<td>68</td>
</tr>
<tr>
<td>2.6</td>
<td>Comparison of PROF1 motifs and classical <em>Rhodopsin</em> motifs.</td>
<td>69</td>
</tr>
<tr>
<td>2.7</td>
<td><em>Rhodopsin</em> SVM training parameters and cross-validation accuracy.</td>
<td>70</td>
</tr>
<tr>
<td>2.8</td>
<td><em>Rhodopsin</em> SVM classifier results.</td>
<td>71</td>
</tr>
<tr>
<td>2.9</td>
<td>Amine SVM classifier results.</td>
<td>72</td>
</tr>
<tr>
<td>3.1</td>
<td>RNAi-based <em>GtNPR-1</em> deorphanization cAMP raw values</td>
<td>91</td>
</tr>
<tr>
<td>3.2</td>
<td>5-HT candidate receptor selection</td>
<td>92</td>
</tr>
<tr>
<td>A.1</td>
<td>Summary of Ligand-Gated Ion Channels (LGIC)</td>
<td>100</td>
</tr>
<tr>
<td>A.2</td>
<td>Summary of Ligand-Gated Ion Channels (LGIC)- Continued</td>
<td>101</td>
</tr>
<tr>
<td>A.3</td>
<td>Summary of Voltage-Gated and Other Ion Channels</td>
<td>102</td>
</tr>
<tr>
<td>A.4</td>
<td>Summary of Voltage-Gated and Other Ion Channels- Continued</td>
<td>103</td>
</tr>
<tr>
<td>C.1</td>
<td>PROF1 RT-PCR conditions</td>
<td>111</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Schistosome life cycle.</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>GPCR signal transduction.</td>
<td>12</td>
</tr>
<tr>
<td>1.3</td>
<td>GPCR classification tree.</td>
<td>17</td>
</tr>
<tr>
<td>1.4</td>
<td>SVM classification: the linear case.</td>
<td>19</td>
</tr>
<tr>
<td>1.5</td>
<td>GPCR deorphanization in mammalian culture.</td>
<td>21</td>
</tr>
<tr>
<td>2.1</td>
<td>Transmembrane domain-focused GPCR sequence mining strategy.</td>
<td>52</td>
</tr>
<tr>
<td>2.2</td>
<td>HMM-based identification of <em>S. mansoni</em> and <em>S. mediterranea</em> GPCRs.</td>
<td>53</td>
</tr>
<tr>
<td>2.3</td>
<td>Manual curation and expansion of <em>Schistosoma mansoni</em> GRAFS GPCRs.</td>
<td>54</td>
</tr>
<tr>
<td>2.4</td>
<td>GRAFS phylogenetic tree.</td>
<td>55</td>
</tr>
<tr>
<td>2.5</td>
<td><em>Rhodopsin</em> phylogenetic tree.</td>
<td>56</td>
</tr>
<tr>
<td>2.6</td>
<td>Aminergic receptors: <em>S. mediterranea</em> and <em>S. mansoni</em>.</td>
<td>57</td>
</tr>
<tr>
<td>2.7</td>
<td>Phylogenetic analysis of PROF1 GPCRs.</td>
<td>58</td>
</tr>
<tr>
<td>2.8</td>
<td>PROF1 multiple sequence alignment.</td>
<td>59</td>
</tr>
<tr>
<td>2.9</td>
<td>Phylogenetic analysis of <em>Glutamate</em> GPCRs.</td>
<td>60</td>
</tr>
<tr>
<td>2.10</td>
<td>Schematic of glutamate in association with LBD residues.</td>
<td>61</td>
</tr>
<tr>
<td>3.1</td>
<td>RNAi-based deorphanization approach overview.</td>
<td>85</td>
</tr>
<tr>
<td>3.2</td>
<td>cAMP ligand screen.</td>
<td>86</td>
</tr>
<tr>
<td>3.3</td>
<td>Semi-quantitative PCR reveals <em>GtNPR-1</em> knockdown.</td>
<td>87</td>
</tr>
<tr>
<td>3.4</td>
<td>RNAi-based GtNPR-1 deorphanization.</td>
<td>88</td>
</tr>
<tr>
<td>3.5</td>
<td>Pharmacological response profile of 5-HT receptor agonists.</td>
<td>89</td>
</tr>
<tr>
<td>3.6</td>
<td>Effects of <em>Smed-SER85</em> suppression on basal motility.</td>
<td>90</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>4.1</td>
<td>Localization of PROF1 transcripts in <em>S. mediterranea</em></td>
<td>94</td>
</tr>
<tr>
<td>B.1</td>
<td>Heterologous GPCR expression in yeast.</td>
<td>104</td>
</tr>
<tr>
<td>B.2</td>
<td>Smp_011940 sequence alignment.</td>
<td>105</td>
</tr>
<tr>
<td>B.3</td>
<td>Full-length gene transcript of Smp_011940.</td>
<td>107</td>
</tr>
<tr>
<td>C.1</td>
<td><em>S. mediterranea</em> PROF1 RT-PCR.</td>
<td>111</td>
</tr>
<tr>
<td>D.1</td>
<td>RNAi vector pPR244 (pDONRdT7).</td>
<td>112</td>
</tr>
<tr>
<td>E.1</td>
<td>Membrane preparation optimization with cAMP RIA.</td>
<td>113</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the many individuals who helped me over the course of my doctoral studies and research. First and foremost, I would like to thank my advisor Dr. Tim Day for his guidance and patience. Dr. Day’s refreshing approach to graduate education provided an environment that tremendously aided my development as a scientist. I will remain grateful for the opportunity he extended to me. I would like to thank Dr. Michael Kimber for the many hours he dedicated to my molecular biology training in his role as a postdoctoral research associate, and for his open door policy later on in his role as faculty. My research experience was greatly enriched by constructive advice from Drs. Baum, Carlson and Greenlee.

I would also like to take this opportunity to thank the many students who were involved with this work in various capacities. In particular, Dr. Ekaterina Novozhilova for her general helpfulness as the senior graduate student in the laboratory for much of the time I was a student. Valuable discussions and assistance were provided by Prince Agbedanu, Michael Song, and Nalee Xiong. To all of these individuals I am grateful.
1 GENERAL INTRODUCTION

This dissertation describes work towards the goal of characterizing an important superfamily of cell receptor proteins in two prominent organisms. This is of interest from an infectious disease and human health standpoint, as well as a purely biological standpoint. The protein superfamily investigated is the G protein-coupled receptor (GPCR) family. Extensive efforts were made to delineate the receptor complements of the human parasite *Schistosoma mansoni* and the model organism *Schmidtea mediterranea*. Further work primarily focuses on validation of a novel method for elucidating receptor function in the native cell membrane environment.

The first chapter of this dissertation describes these biomolecules and organisms in appropriate depth, justifies the significance of the research topic, and outlines a hypothesis-driven research plan, and serve as a general foundation for the research manuscripts that follow. These works are tied together and discussed in the final chapter, where future avenues of research are suggested. The appendices include various supplementary documents and my summarized contributions to related published work that are referenced within manuscripts.

1.1 Schistosomes: Significant Human Parasites

Schistosomes are trematodes of the phylum Platyhelminths, and etiological agents of schistosomiasis. In their parasitic capacity, schistosomes continue to pose a significant challenge to human health. Recent estimates place the number of infected humans at a staggering 207 million across 74 countries, with 280,000 deaths per annum attributed to schistosomiasis in sub-Saharan Africa alone [1]. In endemic areas, another 779 million people are thought to be at risk of infection [2]. It is calculated that 70 million disability-adjusted life years (DALYs)
are lost to schistosomiasis annually [2, 3]. This figure surpasses the global burdens posed by both malaria and tuberculosis, and is nearly equivalent to that of HIV/AIDS [4]. To further exacerbate the situation, there exists strong evidence that the prevalence of this disease helps account for the disproportionately high HIV-1 infection rates observed in afflicted areas [5].

The genus *Schistosoma* contains a number of species that infect humans, with varying geographic distributions. Among these, *Schistosoma mansoni*, *Schistosoma japonicum*, and *Schistosoma haematobium* account for a prevalence of infections. *S. mansoni* and *S. japonicum* cause intestinal schistosomiasis, while *S. haematobium* leads to urogenital schistosomiasis. The wide range of species-specific clinical manifestations associated with chronic schistosomiasis have been well described. These include persistent inflammation, and morbidity brought on by anemia, diarrhea, chronic abdominal pain, fatigue, malnutrition, and childhood stunting [6].

While there is great variance in the severity of pathologies, the overall health and economic impact of this disease is a crippling factor in many developing countries. The relative lack of attention schistosomiasis receives with respect to its human health burden has led to its classification as a neglected tropical disease (NTD).

At present, this overwhelming disease burden is met with a near exclusive reliance on treatment with the drug praziquantel. Significant advances in schistosomiasis control have been achieved in the past few years, due primarily to well-organized control programs that include aggressive chemotherapy [7, 8]. While these programs produce unquestionable public health benefit, their assertive use of praziquantel raises well-founded concerns regarding the long-term efficacy of this approach. Alarming reports of emerging drug resistance and the potential for resistance [9,10] have spurred recognition of the pressing need for new antischistosomals [1,11–15].

The schistosome life cycle is complex, requiring an intermediate fresh water host. We can begin description of this life cycle with the deposition of schistosome eggs into water. The eggs hatch and release motile miracidia, which parasitize snails as an intermediate host. These miracidia
then develop into sporocysts within the intermediate snail host. After successive generations, free-living cercariae emerge and are released into water. Cercariae penetrate human skin and transition into schistosomulae. Schistosomulae enter the venous circulatory system and travel to the lungs. From there, they migrate to the hepatic portal system where they undergo sexual maturation and develop into male and female adults. Adult mate pairs migrate to a final infection site in a species-dependent manner: either the mesenteric venules of the small intestine (intestinal schistosomiasis), or the venous plexus of the bladder (urinary schistosomiasis). The female parasites produce hundreds to thousands of eggs daily. These eggs pass through the lumen wall and are then expelled in feces or urine, allowing the life cycle to begin anew. The life cycle is displayed in Figure 1.1.

The schistosome life-cycle is depicted in linearized fashion. Life stages: egg (E), miracidium (M), sporocyst (S), cercaria (C), schistosomule (S), and adult (A). Intermediate host (snail) and host (human) stages are outlined.

1.2 Planarians: Robust Model Organisms

Planarians are free-living bilateral organisms of the phylum Platyhelminthes. The remarkable regenerative capabilities of these organisms and their evolutionary position as basal metazoans, have advanced their use as a model system to study stem cell biology and development [16–18]. The planarian body is seeded with stem cells exhibiting totipotency. These ‘neoblasts’ constitute > 20% of the total organismal cell count and exhibit the capacity to differentiate into
any of the estimated 40 planarian cell types [19]. Planarians are hermaphroditic and rely on either or both sexual or asexual modes of reproduction to propagate, in a species-dependent manner. Asexual reproduction occurs via transverse fission along the anteroposterior axis, and can be promoted by various environmental cues including population density and light-dark cycles [20,21]. A large number of planarian species exist with varying reproductive modes complicated by varying ploidy. Among these, Schmidtea mediterranea is the most widely studied.

S. mediterranea is a stable diploid \(2n = 8\) with a relatively small haploid genome. Interestingly, a Robertsonian chromosomal translocation involving the fusion of the whole arm of chromosome 1 to chromosome 3 has generated exclusively asexual strains of this species [22]. Both sexual and asexual strains are easily maintained and propagated in the laboratory. Clonal asexual lines derived by serial amputation severely limit genetic variation and are therefore ideal for experimental manipulation and examination of genome content. Other planarian species, such as Dugesia tigrina and Dugesia japonica, are also widely studied with respect to their strong regenerative abilities. However, their mixoploid genomes \(2n = 16\) and \(3n = 24\) coupled with the presence of large numbers of transposable elements [23], complicate genomic analyses. In the course of this work, we make use of S. mediterranea and D. tigrina.

1.3 Genomics

The most recent release of the S. mansoni genome is in the form of a 381 Mb assembly displaying approximately 6X coverage, accompanied by a set of 11,812 predicted proteins [24]. Schistosome genomic DNA was isolated and prepared using standard methods from mixed-sex cercariae of the Puerto Rico isolate. Whole-genome shotgun Sanger sequencing was performed on insert-containing plasmid, fosmid, and bacterial artificial chromosomes (BAC) libraries. The final predicted gene set resulted from the application of multiple ab initio gene prediction algorithms that were further refined with transcriptome data and information incorporated from the S. japonicum genome [25]. A manually curated subset of 402 schistosome gene models served as a training input for the predication algorithms employed.
The current *S. mediterranea* genomic assembly is estimated at \( \sim 865 \) Mb in length at 11X sequencing depth. A similar WGS (whole-genome shotgun) sequencing approach was used, whereby plasmid and fosmid libraries were constructed with genomic DNA purified from the clonally derived S2F2 strain. The predicted *S. mediterranea* proteome arises from 30,930 gene models that were produced by MAKER [26,27], an automated genome annotation pipeline that makes use of a number of gene prediction tools and available EST (expressed sequence tag) collections. However, there is evidence to suggest that this may be a vast overestimate of the true planarian gene count. In order to refine this figure, 31 *S. mediterranea* mRNAs were sequenced and aligned with the assembled genome and compared to corresponding MAKER predictions. This exercise generated estimates of transcripts represented in the MAKER gene set (90%), as well as transcripts extending across multiple contigs (30%) and transcripts incorrectly split into multiple gene models (15%). Extrapolation of these figures to the whole genome indicates a more reasonable protein-coding gene count of 15,570. Table 1.1 presents the current state of both nuclear genomic assemblies.

<table>
<thead>
<tr>
<th></th>
<th><em>S. mansoni</em></th>
<th><em>S. mediterranea</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>363 Mb</td>
<td>865 Mb</td>
</tr>
<tr>
<td>Assembly</td>
<td>v3.1</td>
<td>v3.1</td>
</tr>
<tr>
<td>Coverage</td>
<td>6X</td>
<td>11.6X</td>
</tr>
<tr>
<td>Repeats</td>
<td>45%</td>
<td>46%</td>
</tr>
<tr>
<td>A/T Richness</td>
<td>65%</td>
<td>69%</td>
</tr>
<tr>
<td>Contigs</td>
<td>31,407</td>
<td>94,682</td>
</tr>
<tr>
<td>C N50</td>
<td>17,677</td>
<td>19,025</td>
</tr>
<tr>
<td>Supercontigs</td>
<td>5,745</td>
<td>43,294</td>
</tr>
<tr>
<td>SC N50</td>
<td>879,876</td>
<td>40,862</td>
</tr>
<tr>
<td>Gene Prediction</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Genes</td>
<td>11,809</td>
<td>39,930* (15,570)</td>
</tr>
</tbody>
</table>

Table 1.1 Genome assembly statistics.
1.4 Functional Genomics: RNA Interference

The recent arrival of flatworm whole genome sequence data opens avenues for functional approaches to investigating the molecular physiology of both schistosomes and planarians. Careful enlistment of these genomic resources alongside improved helminth gene manipulation protocols, will aid in deciphering mechanisms of parasite survival and pathogenesis and may lead to new strategies for parasite control and elimination. Although significant strides have been made towards developing flatworm transgenic approaches \[28,29\], readily-adaptable and robust protocols are still lacking. In the case of schistosomes, this effort is hampered by the complex parasite life cycle which only allows for time-restricted maintenance of specific life stages \textit{in vitro} \[30\]. Reports of the successful introduction of transgenes in schistosomes have thus far been transient and stage-specific, involving electroporation as a means of delivery \[31,32\]. The potential for germ line transgenesis \[33\] could eventually yield parasite strains with heritable loss-of-function and gain-of-function genotypes. Advances in parasite culture techniques could conceivably catalyze this recognized and worthwhile goal.

The best-established functional genomics tool available for the study of both pathogenic and free-living flatworms is RNA interference (RNAi). RNAi was first discovered in the nematode \textit{Caenorhabditis elegans}, where introduction of exogenous double-stranded RNA (dsRNA) was shown to result in sequence-specific gene suppression \[34,35\]. This phenomenon has been shown to extend to protozoa and nearly all higher-order eukaryotes examined, and has rapidly become a standard tool for loss-of-function gene analysis. Long dsRNAs can elicit potent gene suppression when introduced into worms, flies, and plants \[36\]. In mammalian cells, much shorter silencing triggers are required to avoid a non-specific interferon response (IFN) \[37\]. What follows is an outline of the basic components of the canonical RNAi pathway.

1.4.1 Canonical RNAi Pathway

In general, the RNAi pathway involves the association of small non-coding RNAs (ncRNAs) with a ribonucleoprotein complex dubbed the RNA-induced silencing complex (RISC). RISC
uses sequence information from a guide strand to downregulate the translation of complementary mRNAs via transcript cleavage or translational block. The RISC-mediated silencing pathways triggered by small ncRNAs that originate from exogenously-applied long dsRNAs and endogenous microRNAs (miRNAs) essentially converge after a series of RNA processing events. Other pathways involving ncRNAs derived from single stranded RNA (ssRNA) precursors such as piwi-interacting RNAs (piRNAs) are set aside in this overview.

Exogenous dsRNAs are first processed into ∼ 21-25 nt dsRNAs by the RNase III family ribonuclease Dicer [38]. Dicer contains a PAZ domain which recognizes dsRNA helical ends, and two catalytic RNase III domains which cleave individual strands to produce siRNA duplexes with 2 nt 3’ overhangs [39]. The distance between the PAZ and RNase III domains corresponds to the length of the siRNA duplexes produced. The guide strand incorporated into RISC is determined by the relative thermodynamic stability of the 5’ ends of the two siRNA duplex strands [39–41].

In the microRNA pathway, genomically-encoded miRNAs are first transcribed by RNA polymerase II as long pri-miRNAs with a 5’ cap and a poly-A tail [42,43]. Pri-miRNAs are processed in the nucleus by a complex consisting of the RNase III enzyme Drosha and the dsRNA-binding protein Pasha, yielding ∼ 65-70 nt pre-miRNAs folded into stem-loop structures. Following export to the cytoplasm, pre-miRNAs are recognized and cleaved into imperfect miRNA:miRNA* duplexes by Dicer. As with siRNA duplexes, the mature miRNA guide strand is selected for incorporation into the RISC complex based on the relative thermodynamic stability of the 5’ ends of the two miRNA duplex strands.

The RISC complex contains members of the Argonaute (Ago) protein family [44]. Argonaute proteins play a role in guide strand selection and direct endonuclease activity against mRNA transcripts complementary to their bound siRNA or miRNA guide fragment. The strand that displays lower stability base pairing in its 5’ end preferentially associates with RISC, and its complementary passenger strand is degraded as the first RISC substrate. At this juncture in
the pathway, the origin of the silencing trigger (siRNA or miRNA) does not affect the mechanisms of downstream post-transcriptional silencing. This is instead determined by the degree of complementarity between the guide strand and a target transcript. miRNAs tend to exhibit partial complementarity to the 3’ UTR of one or more target transcripts, leading to translational repression [45]. This allows for cell-type and tissue specific gene regulation [46]. Perfect or near-perfect complementarity between siRNAs or miRNAs and their target transcripts leads to transcript cleavage.

In plants and nematodes, RNAi can be a systemic phenomenon. Cell-to-cell dispersion of silencing triggers in *C. elegans* involves a dsRNA channel (SID-1), which acts as a conduit for the rapid energy-independent import of dsRNA. There is evidence to suggest that alternative SID-1-independent pathways are responsible for the export of dsRNA from cells [47]. Most examined animals, except some insects, house at least a single SID-1 homolog. Less universally conserved mechanisms that improve RNAi efficiency include distinct pathways for the biogenesis of secondary siRNAs. In *C. elegans*, this form of signal amplification occurs via RNA-dependent RNA polymerase (RdRP) activity [48].

### 1.4.2 RNAi in Schistosomes

RNAi in schistosomes provides new opportunities for focused exploitation of genomic data. It is clear that schistosomes possess the cellular machinery that mediates RNAi, with a number of key molecular actors bioinformatically identified or characterized [49–53]. These include a single Dicer homolog (smDicer), two Drosha homologs (SmDrosha1 & 2), four Argonaute homologs (SmAgo1-4), and a SID-1 homolog. The latter suggests that these parasites retain cell-to-cell dsRNA transport mechanisms similar to *C. elegans*. miRNAs have also been identified *in silico* in both *S. mansoni* and *S. japonicum* [54–56], further establishing the presence of the canonical miRNA-processing and RNAi pathway in schistosomes.

Since the original publications documenting RNAi in *S. mansoni* [57, 58], there has been an
accumulation of published accounts of gene silencing in *S. mansoni* life stages including adults, schistosomules, sporocysts and eggs [51,59–70]. For the most part, silencing has been reported for abundantly expressed genes associated with the surface tissues or gut using dsRNA or siRNA triggers, such that the RNAi-susceptibility of genes associated with other tissue types is unclear. For example, the silencing of 32 different genes in the developing sporocyst revealed variable knockdown efficiencies and some off-target silencing, demonstrating the need to optimize and validate RNAi on a gene-by-gene basis [67].

1.4.3 RNAi in Planaria

Double-stranded RNA has been found to be useful for inducing acute gene silencing in planaria. The susceptibility of planarians such as *S. mediterranea* to dsRNA-mediated gene interference has been well demonstrated. RNAi was first shown to lead to both specific and near complete gene inhibition with a microinjection protocol [71]. As with schistosomes, miRNAs have been identified *in silico* in planaria [72–74]. With the required internal machinery established, other protocols have since been described to achieve the desired silencing effect in planarians, including soaking [75] and bacterial feeding [76, 77] as a means of dsRNA delivery. The latter involves transforming RNAi vector constructs into RNase III-deficient bacterial cells, induction of dsRNA synthesis, and ingestion of dsRNA-containing bacteria by planarians. This method has already been utilized in a high-throughput manner to examine the effects genes may have on regeneration and stem cell function by noting developmental and morphological defects, as well as other behavioral phenotypes [78].

1.5 G Protein-Coupled Receptors

G protein-coupled receptors (GPCRs) represent the largest known superfamily of proteins in the metazoa. Relevant to our aims, GPCRs are long-established as lucrative targets for therapeutic intervention, acting as targets for 50% of all prescription pharmaceuticals [79]. This is undoubtedly a function of their extensive role in eukaryotic signal transduction, and more precisely, their involvement in a myriad of consequential stimulus-response pathways. Sequence diversity
within the GPCR superfamily is matched appropriately by diversity in the panel of known exogenous and endogenous ligands [80]. Although GPCRs respond to a range of molecules that include biogenic amines, peptides, odorants and classical neurotransmitters, they are defined by a common structural motif: a core domain of well-conserved seven transmembrane-spanning (7TM) α-helices.

1.5.1 Signal Transduction

GPCRs transduce extracellular signals to intracellular signaling cascades by acting as guanine nucleotide exchange factors (GEFs). The activation of a GPCR by its cognate ligand promotes GDP-GTP exchange in associated heterotrimeric G proteins, comprised of an α subunit and a βγ dimer. The GTP-bound Gα subunit then dissociates from the dimer and activates a particular biochemical pathway, depending on its subtype. Gαs, Gαi/o, Gαq/11, and Gα12/13 constitute the larger G protein family groupings (Table 1.2) [81]. G proteins exert biochemical influence on effector molecules, leading to downstream accumulation or reduction of second-messengers. This can lead to changes in the activity of metabolic enzymes, ion channels, transporters, and the cellular transcriptional machinery, producing a biological effect. Although GPCR signaling is complex and can include G protein-independent pathways [82], the primary G protein-dependent pathways that result from receptor activation are the cAMP and phosphatidylinositol pathways [83].

<table>
<thead>
<tr>
<th>Family</th>
<th>Gα Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gt/o</td>
<td>Gi, Go, Gt, Ggust, Gz</td>
</tr>
<tr>
<td>Gα</td>
<td>Gs, Go</td>
</tr>
<tr>
<td>Gq/11</td>
<td>Gq, Gi1, G14, G15, G16</td>
</tr>
<tr>
<td>G12/13</td>
<td>G12, G13</td>
</tr>
</tbody>
</table>

Table 1.2 Heterotrimeric G protein classification.

Gαs and Gαi stimulate and inhibit the membrane-associated effector adenylate cyclase, which catalyzes the conversion of ATP to the second-messenger 3',5'-cyclic AMP (cAMP), respectively. Gαq/11 acts on the membrane-bound effector phospholipase C beta (PLC-β), which
cleaves phosphatidylinositol 4,5-bisphosphate (PIP2) to produce diacyl glycerol (DAG) and the second-messenger inositol 1,4,5-triphosphate (IP3). In turn, IP3 diffuses into the cytosol and agonizes IP3-gated ion channels on the ER surface, causing an increase in cytosolic Ca2+ concentration. Go12/13 is implicated mostly in such processes as actin remodeling and cellular migration. The Gβγ dimer lacks a catalytic domain and acts through regulated protein-protein interactions [84]. Although the list of Gβγ-interacting proteins continues to grow, Gβγ signaling is generally treated as a secondary signaling pathway with respect to Go-mediated signaling. Figure 1.2 presents an overview of the primary GPCR signaling pathways and the endpoints most frequently assayed as a function of GPCR activation.

![Figure 1.2 GPCR signal transduction.](image)

The three primary G protein subtypes (G-αs, G-αi and G-αq) are depicted along with their corresponding intracellular signaling pathways. The biochemical endpoints most commonly monitored in studies of GPCR activation are cAMP and Ca2+.

More recent findings have further nuanced our understanding of GPCR structure and signaling.

A growing body of biophysical studies indicate the presence of functional receptor dimers and higher order oligomers in Rhodopsin GPCRs with varying symmetries [85, 86]. While
examples of obligate homo- and hetero-dimerization have long existed for Glutamate family GPCRs [87], this represents a significant challenge to the classical view of receptor monomers as the functional signaling unit for the numerically dominant Rhodopsin family. Despite ambiguity as to the scale of occurrence, the possibility of widespread oligomeric signaling necessitates the development of more sophisticated pharmacological and regulatory models to describe receptor behavior [88].

1.5.2 Export and Regulation

GPCR synthesis, folding, and assembly occur at the ER. The underlying mechanisms of GPCR transport from the ER to the cell surface are not yet very well understood, at least in comparison to the extensive work performed on the endocytic pathway. Proper folding has been shown to involve the guidance of ER chaperones and accessory proteins [89]. Further, proteins such as receptor activity modifying proteins (RAMPs) can affect the pharmacology of eventual surface-exposed receptors [90]. Correctly-folded receptors which are able to pass the ER quality control mechanism are packaged into ER-derived vesicles for export. GPCR motifs have been identified that play a role in ER export regulation by selective interaction with components of the COPII transport system, influencing such parameters as cargo concentration and rate of exit [91]. GPCRs are transported in vesicles to the cell surface through the ER-Golgi intermediate complex (ERGIC), the Golgi, and the trans-Golgi network (TGN). Along the export pathway, the vast majority of GPCRs undergo some form of post-translational modification (e.g., palmitoylation and glycosylation) [92].

G protein export follows along similar lines, with both co- and post-translational modifications helping establish the physical association of G protein subunits with the cell membrane. Except in the case of Gαi, all Gα subunits are N-terminally palmitoylated [81]. Gαi subunits are also targets of N-terminal myristoylation. Although the Gβγ dimer exhibits greater hydrophobicity, this is insufficient for membrane association and Gγ subunits undergo C-terminal thio-ether-linked isoprenylation with either a farnesyl or geranylgeranyl moeity [93]. Once GPCRs and
G proteins have made their way to the cell surface, they are known to interact with molecules such as GRKs (GPCR kinases) [94, 95], arrestins [96, 97], and RGS (regulators of G protein signaling) proteins such as GAPs (GTPase-activating proteins) [98, 99]. These cellular actors regulate receptor internalization, coupling, ligand binding, recruitment and recycling.

The activation cycle of heterotrimeric G proteins can be described by the equilibrium of active G\(\alpha^*\)[GTP] and inactive G\(\alpha\)[GDP] subunits. GPCRs in their active conformation act as GEFs, promoting G\(\alpha\) GDP release and the rapid binding of GTP: [GTP]_{cytosol} \gg [GDP]_{cytosol}. There are two basic models that describe G protein-GPCR interactions [93]. In the ‘collision coupling’ model, G proteins associate with active GPCRs as a result of free stochastic movement within the membrane. In the ‘pre-coupling’ model, G proteins are coupled to GPCRs prior to their activation. Antagonistic to receptor-mediated GEF activity, G\(\alpha\) subunits have intrinsic GTPase activity. This reaction is catalyzed by GAPs, which significantly accelerate the rate of GTP hydrolysis and thus, signal termination.

\[
G_\alpha[GDP]G_{\beta\gamma} \xrightleftharpoons[GAP]{GEF} G_\alpha^*[GTP] + G_{\beta\gamma}
\]

To modulate the magnitude and specificity of intracellular biochemical responses to extracellular ligands, receptors are tightly regulated. The best-characterized pathway for receptor downregulation is the endocytic pathway via clathrin-coated pits (CCPs) [97, 100]. In this pathway, the intracellular domains of agonist-occupied GPCRs are phosphorylated by GRKs, followed by arrestin binding. Arrestin brings about receptor desensitization by inhibition of G protein coupling. Receptor/arrestin complexes can be recruited through an adapter complex into clathrin-coated pits, resulting in receptor internalization and the formation of the early endosome. Receptors are then sorted, either marked for lysosomal destruction or recycled back to the cell surface.

Another mechanism of homologous desensitization involves the activity of enzymes accumulated
downstream of G protein activation, such as protein kinase A (PKA) and protein kinase C (PKC) [101]. These enzymes can directly or indirectly (through GRKs) phosphorylate agonist-bound GPCRs. Alteration of the coupling profile of a receptor is another potential consequence of phosphorylation [102]. The molecular events governing heterologous desensitization are less well understood [103], however, it should be noted that receptor signaling pathways can exhibit extensive cross-talk leading to positive or negative forms of regulation [104].

1.5.3 Phylogenetic Overview

Early efforts to sub-classify and bioinformatically fingerprint the GPCR superfamily gave rise to the A-F classification system [105, 106]. Pharmacological organization of this receptor family was continued with the GPCRDB database [107]. More recent phylogenetic analysis of the entire known human GPCR complement revealed five primary GPCR groupings: Glutamate, Rhodopsin, Adhesion, Frizzled and Secretin [108]. Together, the families are referred to as ‘GRAFS’, and their lineages have been shown to extend throughout the metazoa [109, 110] (Table 1.3). Within-family phylogenetic analysis of Rhodopsin GPCRs revealed four primary clusters (α, β, γ, and δ), further sub-classified into 13 groups.

<table>
<thead>
<tr>
<th></th>
<th>H. sapiens</th>
<th>M. musculus</th>
<th>D. melanogaster</th>
<th>C. elegans</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>24</td>
<td>112</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>R</td>
<td>752</td>
<td>1106</td>
<td>76</td>
<td>124</td>
</tr>
<tr>
<td>A</td>
<td>27</td>
<td>13</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>S</td>
<td>20</td>
<td>28</td>
<td>13</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1.3 GRAFS species comparison.

Among these, the α amine and β peptide groupings are the most populated in most examined species, including the Ecdysozoa. Many of these Rhodopsin sub-groupings are only present in mammals, and examples of lineage-specific GPCR expansions can also be identified that fall outside this classification scheme. For example, nematode chemosensory receptors in C. elegans comprise over 85% of the total GPCR complement [111]. Flatworm whole genome sequence
data provides an important substrate for the potential identification of similar phylum or species-specific GPCR groupings.

1.6 In silico GPCR Mining

Growing genomic resources have provided a GPCR mining platform for a number of organisms, including *Homo sapien* [112], *Mus musculus* [113], *Gallus gallus* [114], *Rattus rattus* [115], *Tetraodon nigrovirdis* [116], *Anopheles gambiae* [117], *Drosophila melanogaster* [118], *Ciona intestinalis* [119], *Branchiostoma floridae* [120], *Xenopus tropicalis* [121] and *Canis familiaris* [122]. GPCR sequences have been accumulated in these genomes using a range of bioinformatics methods that include homology-based searching (BLAST), hidden Markov models (HMMs) and motif-driven queries [123]. The more successful GPCR mining protocols have involved the application of a combination of such methods and algorithms.

Known GPCR sequences are commonly used as queries to search against sequence assemblies with BLAST [124]. BLAST is a heuristic that approximates the optimal Smith Waterman alignment algorithm, allowing for efficient searches against large genomic datasets. This algorithm is ideal for finding sequences that share at least moderate sequence similarity, but is limited in its ability to detect highly-diverged or novel GPCR sequences. While BLAST identifies global sequence homology, motif-based approaches focus on highly-conserved and length-restricted aspects of GPCR primary structure. The ordered combination of motifs in the form of fingerprints can be used to filter false positives generated by individual motifs. While motifs can be declared with different levels of exactness, this approach suffers from similar disadvantages in the detection of atypical GPCR sequences.

Profile hidden Markov models (HMMs) represent a more sophisticated alignment-based approach to the modeling of protein families. In broad terms, a hidden Markov model consists of a system of ‘hidden’ states and observed variables. Conditional probabilities are ascribed to state transitions, as well as for the emission of variables within states. In the biological
application of HMMs, emitted variables are typically sequence alphabets (e.g. nucleotide or amino acid) and hidden state paths represent underlying biological units (e.g. introns or GpC islands) [125]. Profile HMMs have been introduced as a means of building consensus statistical models of protein families [126]. The underlying architecture of these HMMs consists of a ‘match’, ‘insert’, and ‘delete’ state, fit for linear sequence data. Profile HMMs have seen extensive use in GPCR discovery, and HMMs have also been developed for the prediction of G protein coupling profiles [127], TM domain boundaries [128, 129], and the identification of peptide ligands [130].

1.7 In silico GPCR Classification

![Figure 1.3 GPCR classification tree.](image)

GRAFS-based distinctions are presented at the family (level 1) plane, while consensus classification shared between the A-F and GRAFS systems are used at the subfamily (level 2) and sub-subfamily (level 3) plane. Profile HMMs can be used to effectively distinguish GPCRs from other transmembrane proteins and to place them within their primary families. SVM classifiers can be used for level 2 and 3 placement of Rhodopsin family GPCRs.

There is great overlap in the computational approaches to GPCR mining and GPCR classification, and the methods discussed above can be applied to both. While HMMs are adequate for GPCR identification and classification at the family plane, other algorithms outperform HMM and homology-based classifiers at deeper classification planes (Figure 1.3). Numerous support vector machine (SVM) classifiers have been shown to classify GPCRs with high accuracy at the subfamily and sub-subfamily levels. In reduced terms, SVMs represent a powerful
supervised-learning method for data classification. Given a combined set of positively and negatively labeled training instances, an SVM produces a binary classifier that can then be used to label unknown samples. Each instance is associated with a fixed-length numerical feature vector, containing certain attributes of the data to be classified. The SVM identifies a maximum-margin separating hyperplane to distinguish between vectors representing instances of opposite sign (Figure 1.4, [131]).

SVM training datasets are populated by pairings of feature vectors and classification groups:

\[ Dataset = \{ x_i, y_i \} \mid x_i \in \mathbb{R}^D, y_i \in \{-1, +1\} \]  

where \( i = 1, \ldots, n \) and \( n \) represents the number of training instances where \( D \) represents the dimensionality of feature vectors used in training.

In the linear case, the SVM identifies a hyperplane of the form:

\[ w^T x + b = 0 \]  

where \( w \) is the normal vector perpendicular to the hyperplane.

We can select normalization such that

\[ w^T x + b \geq +1 \]  

\[ w^T x + b \geq -1 \]

for positive (+1) and negative (−1) support vectors, respectively.

The margin between these is given by \( 2/\|w\| \), and an SVM can be trained to maximize this distance by minimizing \( \|w\| \). This is treated as a quadratic optimization problem with linear constraints [131]. Data that is not fully linearly separable is dealt with by the introduction of slack variables that introduce costs for misclassified instances. Additionally, non-linearly separable feature vectors must often be mapped to a higher dimensional space by the application
of kernel functions. The kernel trick allows for the construction of a separating hyperplane in the transformed feature space.

Recently, this approach has seen extensive use in the area of biosequence discrimination, and relevant to our goals, the particular problem of GPCR classification. In the first study on the matter, SVM-based classifiers were shown to drastically outperform their BLAST and HMM-based counterparts for level 1 and level 2 GPCR subclassification [132]. Subsequent studies further improved the predictive performance of SVMs with the introduction of dipeptide composition feature vectors [133, 134], achieving accuracies of 97.3% and 96.4% for level 1 (Rhodopsin) and level 2 (amine) classification, respectively. Alternative feature vectors have since been similarly validated [135, 136].

Although SVMs are binary classifiers, they can be applied to multi-class problems where the number of classes, $k$, is greater than 2. Such problems are typically solved using either the “one-versus-rest” (OvR) or “one-versus-one” (OvO) method. In the OvR scenario, $k$ binary classifiers are trained, such that each classifier separates one class from all others. The “winner-
A "takes-all" strategy is then commonly used to label unknown samples, whereby the classifier with the highest output decision function assigns the final class. In the OvO scenario, \( \frac{k(k-1)}{2} \) binary classifiers are constructed in a pair-wise manner. A voting strategy is then typically employed in classification, whereby each classifier accounts for one vote and the class with the maximum number of votes assigns the final label. Although the OvO method has been shown to perform better on a number of fronts [137], as far as the authors are aware, all previously described SVM-based GPCR classifiers available for public use rely on the simpler OvR method.

1.8 GPCR Deorphanization

Once identified, GPCRs require deorphanization, the process of pairing orphan receptors with their cognate ligands. The predominant approaches all require the transient or stable heterologous expression of receptors in a surrogate cell system and in most cases, this expression occurs in cells derived from other species and phyla. In the absence of a flatworm cell culture line, flatworm GPCRs have been expressed in such divergent cellular environments as CHO [138], HEK293 [139, 140], COS7 [139], yeast [140, 141], and *Xenopus* oocyte cells [142]. The traditional work-flow begins with full-length sequence cloning and introduction of the receptor coding sequence into an appropriate expression vector designed for the particular cell culture system.

Receptor activation assays incorporate different detection schemes (e.g. fluorescence, bioluminescence, and gene reporting) and are amenable to a host of receptor-specific adaptations. In mammalian cell culture, GPCR activation is commonly assayed as a function of Ca\(^{2+}\) mobilization from the ER (Figure 1.5). Agonist-evoked calcium flux can be detected using fluorescent calcium-sensitive dyes in a high throughput platform. In this system, co-transfected G protein chimeras [143, 144] can be used to divert GPCR signaling through the Go\(_q\) pathway, irrespective of the endogenous receptor coupling profile. The carboxyl-terminus of the Go subunit is primarily responsible for receptor specificity, and chimeric G proteins have been constructed via replacement of the five C-terminal amino acids of Go\(_q\) alpha with those of other G protein subtypes.
Yeast have also been shown to be attractive heterologous GPCR expression systems for a number of reasons. They provide a minimal eukaryotic host environment, are genetically robust and amenable to manipulation, and are low cost and relatively easy to maintain in the laboratory. Further, they lend themselves to high throughput-screening assays with simple outputs. A number of yeast strains have been genetically modified for the purpose of heterologous GPCR expression. In these engineered strains, modifications of the endogenous yeast GPCR-mediated pheromone response pathway direct the activation of a non-native GPCR to yeast growth as a phenotypic output [145]. Our laboratory and others have had recent success in employing yeast as a means of deorphanizing both platyhelminth [140] and nematode [146] GPCRs, and we have continued on this path with other flatworm receptors (Appendix B).

1.8.1 Difficulties and Pitfalls

Current approaches to GPCR deorphanization have severe limitations and are inefficient for large-scale projects. This has introduced a significant bottleneck in the way of both the pharma-
colological and structural characterization of GPCRs [147,148]. Although heterologous expression is not a theoretically challenging feat, individual targets routinely prove to be recalcitrant and consume inordinate effort. The time and effort required to establish the transient or stable expression of a receptor make traditional approaches undesirable for projects of larger scale. The processes that guide the proper folding and export of receptors are not necessarily well-conserved across cell lineages. Further, a significant set of potential concerns also present themselves post-transfection or transformation. In the event that a GPCR is successfully expressed on the surface of a host cell, the GPCR must operate in conjunction with a foreign complement of accessory and signaling proteins. The possibility of a receptor with altered structure or function is therefore a consequential worry. In view of these concerns, a receptor deorphanization method that could be applied in a native cell or membrane environment could side-step some of these concerns.

1.9 General Objectives

Objective 1. Whole-genome identification and classification of GPCRs in *S. mansoni* and *S. mediterranea* using a transmembrane-focused bioinformatics protocol. The primary hypothesis is that these platyhelminths possess an extensive and complex GPCR repertoire, and exhibit a host of lineage-specific features.

Objective 2. Development of a loss-of-function GPCR deorphanization protocol that occurs in the native cell membrane environment. The primary hypothesis is that joining RNAi and GPCR second messenger assays can lead to the elucidation of receptor agonists and G protein coupling pathways.

These objectives are pursued with the overarching goals of identifying schistosome drug targets, mapping platyhelminth receptor biology, promoting planarians as model organisms for parasite research, and establishing higher-throughput methods of receptor deorphanization and drug target validation. The selection of specific ‘rational’ targets is aided by these focused objectives.
2 The Repertoire of G Protein-Coupled Receptors in the Human Parasite Schistosoma mansoni and the Model Organism Schmidtea mediterranea

A paper submitted to the Journal *BMC Genomics*

Mostafa Zamanian $^{1,2,*}$, Michael J Kimber $^{1,2}$, Paul McVeigh $^3$, Steve A Carlson $^{1,2}$, Aaron G Maule $^3$, Tim A Day $^{1,2,*}$

Abstract

**Background:** G protein-coupled receptors (GPCRs) constitute one of the largest groupings of eukaryotic proteins, and represent a particularly lucrative set of pharmaceutical targets. They play an important role in eukaryotic signal transduction and physiology, mediating cellular responses to a diverse range of extracellular stimuli. The phylum Platyhelminthes is of considerable medical and biological importance, housing major pathogens as well as established model organisms. The recent availability of genomic data for the human blood fluke *Schistosoma mansoni* and the model planarian *Schmidtea mediterranea* paves the way for the first comprehensive effort to identify and analyze GPCRs in this important phylum.

**Results:** Application of a novel transmembrane-oriented approach to receptor mining led to the discovery of 116 *S. mansoni* GPCRs, representing all of the major families; 104 *Rhodopsin*, 2 *Glutamate*, 3 *Adhesion*, 2 *Secretin* and 5 *Frizzled*. Similarly, 291 *Rhodopsin*, 9 *Glutamate*, 21 *Adhesion*, 1 *Secretin* and 11 *Frizzled* *S. mediterranea* receptors were identified. Among

$^1$Department of Biomedical Sciences, Iowa State University, Ames, IA USA  
$^2$Interdepartmental Neuroscience Program, Iowa State University, Ames, IA USA  
$^3$Biomolecular Processes, School of Biological Sciences, Queen’s University Belfast, Belfast UK  
$^*$Corresponding Authors. Emails: mostafaz@iastate.edu, day@iastate.edu
these, we report the identification of novel receptor groupings, including a large and highly-diverged Platyhelminth-specific Rhodopsin subfamily, a planarian-specific Adhesion-like family, and atypical Glutamate-like receptors. Phylogenetic analysis was carried out following extensive gene curation. Support vector machines (SVMs) were trained and used for ligand-based classification of full-length Rhodopsin GPCRs, complementing phylogenetic and homology-based classification.

**Conclusions:** Genome-wide investigation of GPCRs in two platyhelminth genomes reveals an extensive and complex receptor signaling repertoire with many unique features. This work provides important sequence and functional leads for understanding basic flatworm receptor biology, and sheds light on a lucrative set of anthelmintic drug targets.

### 2.1 Background

The G protein-coupled receptor (GPCR) superfamily constitutes the most expansive family of membrane proteins in the metazoa. These cell-surface receptors play a central role in eukaryotic signal transduction, and conform to a structural archetype consisting of a core domain of seven transmembrane (TM)-spanning α-helices. GPCRs are also established drug targets, acting as sites of therapeutic intervention for an estimated 30-50% of marketed pharmaceuticals [79, 149]. This is undoubtedly a function of their extensive involvement in a wide range of important physiological processes. The diverse panel of known GPCR ligands includes biogenic amines, photons, peptides, odorants and classical neurotransmitters [80]. This diversity is mirrored by the significant degree of primary sequence variation displayed among GPCRs.

At present, there exists no comprehensive study of GPCRs for the phylum Platyhelminthes. This important phylum houses prominent endoparasites, both flukes and tapeworms, as well as free-living species that serve as established model organisms in the realm of developmental biology. Lack of sequence data and a reliance on techniques with a definably narrow expectation of success such as degenerate PCR have contributed to the very modest number of GPCRs thus
far identified or characterized [138–140, 142, 150, 151] in this phylum. The arrival of EST repositories [49, 152, 153] has only marginally contributed to this number, perhaps as a consequence of GPCR under-representation [154]. The recent availability of *Schistosoma mansoni* [24] and *Schmidtea mediterranea* [27] whole genome sequence data provides basis for the *in silico* accumulation and analysis of undiscovered and potentially novel receptors.

The blood fluke *Schistosoma mansoni* is the primary etiological agent of human schistosomiasis, a chronic and debilitating condition that afflicts a staggering 207 million people in 76 countries [2] and accounts for 280,000 deaths per annum in sub-Saharan Africa alone [155]. It is calculated that up to 70 million disability-adjusted life years (DALYs) are lost to schistosomiasis annually [4]. This figure surpasses the global burden posed by both malaria and tuberculosis, and is nearly equivalent to that of HIV/AIDS. At present, this overwhelming disease burden is met with a near exclusive reliance on treatment with the drug praziquantel. The threat of drug resistance [9, 10] has spurred recognition of the pressing need for new antischistosomal agents [1, 13, 15]. In this context, as modulators of a diverse range of critical biochemical and physiological pathways, GPCRs hold great promise as potential targets for disruption of crucial parasite survival and proliferation activities.

The free-living planarian *Schmidtea mediterranea* is an important platyhelminth studied extensively for its regenerative abilities [17, 156]. Like other planarians, it is abundantly seeded with totipotent stem cells with the ability to migrate and undergo division and differentiation at sites of injury. In addition to its current role as a powerful model organism for regeneration and stem cell biology, *S. mediterranea* presents itself as a potential parasite drug discovery model [157]. In the case of nematodes, the biology of the free-living model organism *Ceanhorhabditis elegans* features prominently in many anti-parasitic drug discovery efforts [158, 159]. Like *C. elegans*, *S. mediterranea* is significantly more tractable to modern genomic approaches compared to the parasitic members of its phyla. It is relatively easy to maintain and it is amenable to RNA interference (RNAi) [71]. Genome-wide analysis and comparison of the GPCR complements of
S. mansoni and S. mediterranea is a major step towards engaging this hypothesis.

The growing number of sequenced genomes has provided a GPCR mining platform for a number of organisms, including Homo sapien [112], Mus musculus [113], Gallus gallus [114], Rattus rattus [115], Tetraodon nigrovirdis [116], Anopheles gambiae [117], Drosophila melanogaster [118], Ciona intestinalis [119], Branchiostoma floridæ [120], Xenopus tropicalis [121] and Canis familiaris [122]. For these organisms, GPCR sequences have been accumulated with a range of bioinformatic methods that include homology-based searching (BLAST), hidden Markov models (HMMs) and motif-driven queries [123]. The more sophisticated GPCR mining protocols have involved the application of a combination of such methods and algorithms.

Phylogenetic studies of the GPCRs in a number of eukaryotic genomes have led to the introduction of the GRAFS classification system [108,109]. GRAFS outlines five major protein families thought to represent groupings of receptors with shared evolutionary ancestry present in the Bilateria: Glutamate, Rhodopsin, Adhesion, Frizzled, and Secretin. In addition to these primary families, some organisms are known to house groupings of lineage-specific receptors that constitute distinct GPCR families. Examples in the phylogenetic vicinity of the Platyhelminthes include the nematode chemosensory receptors [160] and insect gustatory receptors [161]. Any in silico protocol for genome wide GPCR identification should therefore cast a broad enough net to reveal any such highly-diverged receptor groupings, while also providing stringency to limit false positives.

Here, we apply an array of sensitive methods towards the goal of identifying, manually curating and classifying putative G protein-coupled receptor sequences in two prominent platyhelminths. Our hypothesis is that organisms in this phylum possess an extensive and complex complement of GPCRs, including phylum or species-specific GPCR groupings. We perform phylogenetic analysis of putative receptors with respect to the GRAFS classification system and employ a machine-learning approach for ligand-based classification of full-length Rhodopsin GPCRs.
2.2 Results and Discussion

In this study, we developed a robust transmembrane-focused strategy to identify, curate and classify putative platyhelminth GPCRs. TM-focused profile hidden Markov models (HMMs) were used to mine the predicted proteomes of *S. mansoni* and *S. mediterranea* in order to identify receptors at the GPCR family plane. Subsequent rounds of filtering were used to remove false positives, followed by homology-based searches against the original genome assemblies. Extensive manual curation of the final sequence dataset allowed for more refined phylogenetic analysis. Greater classification depth was achieved with a complementary transmembrane-focused support vector machine (SVM)-based classifier. An overview of this bioinformatics protocol is outlined in Figure 2.1.

Identification of GRAFS family receptors with TM-focused profile HMMs

Towards the goal of identifying members of the GRAFS GPCR families in our genomes of interest, we relied primarily on the use of family-specific profile HMMs. This alignment-rooted method has been successfully applied in other genomes and has been shown suitable for the identification and classification of GPCR sequences at the family level [123, 162]. In a departure from previously described protocols, we chose to focus HMM training exclusively on the most highly-conserved structural features that extend throughout the GPCR superfamily. The idea behind this measure was to dampen challenges posed by the inexact gene structures that underlie the flatworm predicted proteomes, as well as the sizable phylogenetic distance of this phylum from organisms with characterized GPCR complements.

In this framework, receptor transmembrane domains are convenient markers that can be identified with greater confidence than other GPCR stretches using sensitive prediction algorithms such as HMMTOP [128] and TMHMM [129]. Training sequences were procured from GPCRDB [107] and processed into what we will refer to as “transmembrane-only pseudosequences” (TOPs), representing the ordered concatenation of TM domains flanked bi-
directionally by 5 amino acids (Figure 2.1b). TM-focused HMMs were constructed for each of the major GPCR families, as well as for the nematode chemosensory and insect odorant families. The Adhesion and Secretin training sets were combined to build a single HMM, given that sequences belonging to these families are not easily distinguishable beyond the N-terminal ectodomain [154].

The predicted proteomes of S. mansoni and S. mediterranea were first filtered for the removal of globular proteins. Typical strategies limit investigation to proteins with 6-8 predicted TM domains, tolerating errors in the algorithmic prediction of these regions. We significantly relaxed this filter and broadened the search scope to include all proteins with 3-15 TM domains. The utility of this change then was to alert us to partial sequences or incorrectly predicted gene models that may be reconstructed with manual curation and that otherwise would have been screened from detection. Family-derived profile HMMs already provide an adequately stringent filter for distinguishing between GRAFS family receptors and other transmembrane proteins.

The proteins that survived this filter were processed into TOPs in the same manner as the training sequences. These sequences were searched against the set of profile HMMs, and the resulting hits for each GPCR family were ranked according to E-value. A primary cut-off was selected at the point where subsequent hits showed significant homology to other known proteins or GPCRs belonging to other families. This was accomplished with a BLASTp [124] search of all hits against the NCBI non-redundant (nr) database. Sequences that displayed GPCR-related homology, along with those that returned no significant BLAST results, were retained. As evidenced later, the requirement of statistically meaningful GPCR-related homology introduces an unnecessary selection bias that can mask the identification of unique receptors.

Application of the Rhodopsin HMM to the S. mansoni predicted proteome led to the examination of the 400 top-ranking hits (E-value < 0.007), 77 of which remained after removal of false
positives via homology-based searches. Similarly, 270 of the 450 top-ranked (E-value < 0.002) Rhodopsin HMM hits remained for S. mediterranea. Redundancy within the S. mediterranea genome assembly warranted the detection and removal of identical sequences. BLAT [163] was used to self-align the nucleotide sequences of the predicted proteins that survived the HMM filtering process. Redundant sequences were removed and if a choice was presented, the longest member of a set of identical sequences was retained. This led to the removal of 14 Rhodopsin sequences from the S. mediterranea dataset. Figure 2.2 displays the overall transmembrane distribution for both proteomes at these various stages of processing for the Rhodopsin family. These steps were likewise performed for the nematode chemosensory and insect odorant GPCR families, however no flatworm orthologs were identified. This is not unexpected, considering their lack of conservation among the Ecdysozoa.

**Manual editing of gene models**

Candidate GPCR sequences underwent manual inspection, and the corresponding gene models were edited. This labor-intensive step is crucial in improving the reliability of any further analysis on this gene family. Common manual edits included the merging or splitting of gene models, movement of intron-exon boundaries, and sequence extension or truncation in either or both directions. This process was aided by examination of open reading frames (ORFs) in the vicinity of a gene models. ORFs that housed common receptor motifs, displayed GPCR-related homology or contained transmembrane stretches were typically incorporated. In many cases, sequencing gaps prevented any meaningful improvement. S. mansoni GRAFS sequences and S. mediterranea GAFS sequences were curated in this manner. We avoided genome-wide manual curation of S. mediterranea Rhodopsin sequences in light of the dubious condition of the draft genome. The A/T rich (69%), highly repetitious (46%) and heterozygous nature of the genome has significantly complicated automated assembly efforts. However, as we elaborate later, we did construct and edit gene models for a particular grouping of Rhodopsin-like Schmidtea GPCRs. The significant level of improvement achieved by manual gene editing is shown in Figure 2.3.
Homology-based search and final gene editing

To account for the likelihood that the primary sets of gene models used do not provide perfect accounting of all gene-encoding regions within the assemblies, we exercised a translated nucleotide BLAST (tBLASTn). For each family, putative GPCRs were combined from both species and searched locally against the original nucleotide assemblies translated in all six frames. Hits with E-value < 0.1 were manually examined for GPCR-related homology. In cases where identified regions of homology overlapped with a given gene model, that gene model was added to the sequence pool. Conversely, if no gene model was found to be present at a particular genomic location, a simple preliminary gene model was built by connecting the high-scoring segment pairs (HSP) that contributed to the tBLASTn hit. In keeping with the HMM approach, only putative receptors with a TM count ≥ 3 were retained. This led to a further significant expansion of the total unique sequence count in both organisms (Table 2.1). This reported sequence count is not equivalent to a receptor count, as many of these sequences may represent fragments of a single protein or prove to be redundant sequences. To bridge this gap and to improve the general state of this additional sequence data, manual editing of gene models was again performed.

Overall phylogenetic view

Putative receptor sequences were tentatively divided into three sequence bins based on the number of predicted TM domains: full-length, near full-length and partial. Full-length sequences were those that likely had their entire 7TM domain intact as predicted by HMMTOP with user oversight. Alignments to homologous proteins were used to help make a final decision with respect to the potential algorithmic miscounting of TM domains. Near full-length sequences are predicted to contain ≥ 4 TM domains, while all other sequences (< 4 TM domains) were placed into the partial sequence bin. Phylogenetic analysis was carried out for full-length and many near full-length receptors. Figure 2.4 displays a topological overview of the primary flatworm GPCR groupings. This phylogenetic analysis confirms the distinct and analogous presence of the primary GRAFS families, and further reveals two novel flat-
worm GPCR families: Platyhelminth-specific *Rhodopsin*-like orphan family 1 (PROF1) and Planarian *Adhesion*-like receptor family 1 (PARF1).

**The *Rhodopsin* family**

The *Rhodopsin* family is divided into four main groups (α, β, δ, and γ) and further subdivided into 13 major sub-families via analysis of fully sequenced mammalian genomes [110]. The α and β subfamilies are well-populated in both *S. mansoni* and *S. mediterranea* (Figure 2.5), while the δ and γ subfamilies are absent. Table 2.2 provides a preliminary classification of receptors identified with respect to the GRAFS classification system from a comparative perspective.

**Alpha (α) receptors**

The α subfamily houses amine, opsin-like, and melatonin receptors. Among these, the amine grouping is typically largest. This metazoan trend holds true for *S. mansoni* and *S. mediterranea*, each possessing at least 25 and 60 putative aminergic receptors, respectively. These numbers are greater than those observed among ecdysozoans, and in the case of *S. mediterranea*, the figure surpasses even the human amine GPCR complement. Biogenic amines such as serotonin (5-hydroxytryptamine; 5HT), dopamine, and histamine have been shown to play a prominent role in the flatworm nervous system [164,165]. Although a small number of aminergic GPCRs have been characterized in this phylum, the majority of receptors that mediate aminergic signaling have thus far remained elusive. Phylogenetic analysis was carried out on the flatworm amine GPCR complement with respect to *C. elegans* aminergic receptors, as shown in Figure 2.6. Two diverged flatworm-specific groupings can be outlined, including one that signifies a major paralogous expansion in schistosomes. Other flatworm receptors are grouped and tentatively associated with ligands corresponding to their phylogenetic relationships with deorphanized *C. elegans* GPCRs.

Four melanopsin-like receptors were identified in *S. mansoni*. Five melanopsin-like receptors were identified in *S. mediterranea*, along with a single receptor that displays moderate homology
to various ciliary opsins. Along with the presence of cyclic nucleotide gated (CNG) ion channels in the planarian genome, this raises the possibility of ciliary phototransduction. Another noteworthy observation is the conspicuous absence of melatonin-like receptors in *S. mansoni*, while *S. mediterranea* houses a relatively large complement of 9 such receptors. Melatonin is endogenously synthesized in planaria in a circadian manner [166,167], and has been implicated in regeneration [168]. Identification of melatonin receptors is a requisite for a more complete mapping of the underlying signal transduction pathway(s) in these important processes.

**Beta β receptors**

The β subfamily contains the great majority of peptide and peptide hormone GPCRs in examined organisms. Neuropeptidergic signaling is known to play a fundamental role in flatworm locomotion, reproduction, feeding, host-finding and regeneration [169,170]. The known flatworm neuropeptide complement has recently undergone considerable expansion with the application of bioinformatics and mass spectrometry-based (proteomics) approaches [171,172]. This represents a significant advance from the original handful of FMRFamide-like peptides (FLPs) and neuropeptide Fs (NPFs) first identified in the phylum. Many of these newly-identified amidated peptides are planarian or flatworm-specific, while others exhibit relatedness to peptides in other phyla, including myomodulin-like, buccalin-like, pyrokinin-like, neuropeptide FF (NPFF)-like, and gonadotropin (or thyrotropin) releasing hormone-like peptides.

Our efforts yielded at least 81 and 35 putative peptide receptors in *S. mediterranea* and *S. mansoni*, respectively. These numbers further evidence the notion that peptidergic signaling is the predominant mode of neurotransmission in the Platyhelminthes. Flatworm peptide receptors do not cluster into a single phylogenetic group (Figure 4b). One groupings contains GPCRs that share homology with receptors responsive to peptide ligands present in the vertebrate lineage, while the other is mostly populated with receptors that share homology with invertebrate neuropeptide-responsive receptors including Gt-NPR1 homologs. It can be noted that the putative flatworm peptide receptor count greatly outnumbers the set of currently
known peptide ligands. Although this may be explained by peptide promiscuity and receptor redundancy, it is also very possible that many neuropeptides have yet to be uncovered. Ligands cannot be confidently assigned to the majority of identified receptors. While some show moderate homology to characterized FLP and NPF-like receptors, most receptors display weak or insignificant homology to an assortment of thyrotropin-releasing hormone, capa, sex peptide, growth hormone secretagogue, proctolin, pyrokinin, myokinin, tachykinin, galanin, and orexin receptors. These tentative BLAST-based annotations may be used with caution to help guide receptor deorphanization efforts.

**Other receptors**

A large number of Rhodopsin receptors could not be individually annotated with confidence, and were placed in the “Other Rhodopsin” receptor bin. Receptors in this category lack phylogenetic support to be clustered with known Rhodopsin groupings, and lack meaningful homology to receptors with known ligands. Many receptors in this bin exhibit some weak peptide or amine receptor-relatedness, but these require functional validation before they can be added to the α or β subfamily counts. Many of these receptors are likely unique to the phylum, and therefore obscure the Rhodopsin family subdivisions apparent in the vertebrate lineage.

**Planarian homologs of parasite GPCRs**

Given the relative tractability of planarians to experimental manipulation, we identified the nearest homologs of *S. mansoni* Rhodopsin receptors in the *S. mediterranea* pool (Tables 2.3 - 2.5). It is a reasonable expectation that there is significant conservation in the biological and pharmacological properties of receptors sharing high sequence identity between these species. The characterization of certain planarian receptors is likely to inform us about the function and druggability of parasite receptors. Each *S. mansoni* receptor was first matched to its most similar *S. mediterranea* sequelog, and sequence pairs were ranked according to amino acid percent identity (PID): 8 receptor pairs were identified sharing > 50% PID, 15 with 40-50% PID, 48 with 30-40% PID, and the remaining sequences with < 20% PID. The top grouping is
comprised exclusively of biogenic amine (GAR and 5HT) and peptide GPCRs. Among them is a receptor pair orthologous to Gt-NPR1 [139], the only neuropeptide receptor deorphanized in this phylum. This degree of sequence conservation promotes the use of planaria as a convenient heterologous system to study parasite receptors.

**Platyhelminth-specific Rhodopsin-like orphan family 1 (PROF1)**

A large and distinct sequence clade comprised of 19 *S. mansoni* and 40 *S. mediterranea* proteins was identified and labeled Platyhelminth *Rhodopsin* Orphan Family 1 (PROF1). Members of this novel and highly-diverged phylogenetic grouping are predicted to house a 7TM domain with an extracellular N-terminus and seem to be exclusively derived from intronless genes. Most PROF1 sequences were revealed with homology-based searches after a small number of bait sequences survived the *Rhodopsin* HMM filtering stage. In fact, 31 of the 40 *Schmidtea* PROF1-containing ORFs were identified via tBLASTn, and only one of these ORFs coincided with an existing gene model. Similarly, 13 of 19 *Schistosoma* PROF1 were identified in this manner and only four of these were represented in the predicted gene set.

These receptors display remnants of classical *Rhodopsin* motifs at corresponding positions (Table 2.6), yet show no significant overall homology to any previously discovered GPCRs. It is important to point out that the absolute requirement of GPCR-related BLAST homology as part of the post-HMM filtering stage would have masked the identification of PROF1 receptors. BLASTp searches of all PROF1 sequences against the NCBI nr database (E-value cutoff = 0.1) returned no hits for the majority of sequences. The small pool of hits that did result, exhibit both very poor homology and represent an incongruous range of receptors that include peptide, lipid and odorant GPCRs. This further highlights the unique nature of these receptors.

Maximum parsimony analysis led to the subdivision of PROF1 into three primary phylogenetic groupings with good bootstrap support (Figure 2.7). Group I is the largest among these with 27 and 13 members from *S. mediterranea* and *S. mansoni*, respectively. The lack of obvious one-
to-one orthologs between species suggests expansion or contraction of these receptors occurred after the splitting of planaria and trematodes in the flatworm lineage. Group II includes 6 S. mansoni and 7 S. mediterranea sequences, while group III houses 6 S. mediterranea sequences. It is likely that the closest related receptor to the ancestral gene for this family is contained in group I or II. A multiple sequence alignment of TM domains I-IV (used for phylogenetic analysis) is shown in Figure 2.8.

Of additional interest, short PROF1-like sequence fragments were identified in both genome assemblies that could not be incorporated into full-length gene structures. These may constitute pseudogenes, or be ascribed to errors in assembly. RT-PCR was used to confirm transcript expression for a selection of putative full-length PROF1 receptors in Schmidtea: 8 from group I, 2 from group II and 3 from group III (highlighted in Figure 2.7). Correct-sized amplicons were visualized for all 13 targets. Similarly, we selected a representative from each Schistosoma PROF1 grouping and confirmed transcription in the adult stage: SMP084270 from group I and SMP041880 from group II. It is not currently possible to assign functions or putative ligands for the PROF1 family. However, given that they constitute one of the largest Rhodopsin-like subfamilies conserved between these monophyletic species, we suspect that they play an important biological role in this phyla.

The Adhesion and Secretin Families

Adhesion and Secretin receptors show sequence similarity in their 7TM domains and are commonly grouped as Class II GPCRs. The phylogenetic separation of these families under the GRAFS paradigm is mirrored by noticeable structural differences in their N-terminal ectodomains. Archetypal Adhesion GPCRs have a long N-terminus containing a diverse arrangement of functional domains. In the vertebrate lineage, this family constitutes the second largest grouping of GPCRs after Rhodopsin and is further partitioned into 8 clusters (I-VIII). Secretin GPCRs usually display N-terminal hormone-binding domains (HBD) that confer responsivity to peptide hormones and are thought to descend from the group V Adhesion re-
ceptors [173]. Additional Adhesion-like proteins have been identified in various organisms that stake more dubious evolutionary positions. The insect Methuselah receptors are one such example that have become the subject of great investigation, attributable to their role in life-span regulation and stress resistance [174]. More recently, a cluster of Adhesion-like receptors (NvX) was identified in the basal Cnidarian *N. vectensis* which share some homology with Methuselah receptors [173,175].

We have identified a novel cluster of 12 *Schmidtea* GPCRs that show moderate (> 20% PID) homology to NvX receptors. We denote this cluster Planarian Adhesion-like receptor family 1 (PARF1). Like NvX, PARF1 receptors contain a single Somatomedin B domain, except in the case of SMDC005966C which is predicted to contain two. Interestingly, no PARF1 orthologs were identified in *S. mansoni*. A single Adhesion GPCR in *S. mansoni* (SMP099670) was found to house a Somatomedin B domain, but it otherwise shares no significant homology with PARF1. Two Adhesion-like *Schmidtea* GPCRs (SMD002396 and SMD002965) were identified that most resemble vertebrate group V orphan GPR133. Two *Schmidtea* GPR157 homologues (SMD002980 and SMD009091) were also identified via Adhesion/Secretin HMM, however, these receptors exhibit vague sequence similarity to more than one GPCR family [115].

Latrophilin-like receptors were found to be present in both flatworms. *Schmidtea* SMD011811 contains a GPS domain, and can be grouped with sequence fragments SMDC001354A and SMDC001354B. *Schistosoma* SMP176830 contains a Somatomedin B domain, but shares no significant sequence similarity with the identified latrophilin-like planarian receptors. Evidence of the potential druggability of these particular receptors comes from the parasitic nematode *Haemonchus contortus*, where a latrophilin-like receptor has been identified as a target of an anthelmintic cyclodepsipeptide [176]. One other Adhesion-like parasite GPCR was identified (SMP058380) that displays an N-terminal GPS domain, but with no clear planarian ortholog.

The Secretin flatworm complement is comparatively smaller. Two *S. mansoni* and one *S.
*mediterranea Secretin* GPCRs were identified. SMP125420 and its planarian ortholog SMD004009 show high sequence similarity to diuretic hormone receptors and contain an N-terminal hormone receptor domain (HRM). These receptors likely play a role in homeostatic regulation. *Schistosome* SMP170560 exhibits an HRM domain and parathyroid hormone receptor homology. This receptor may in fact have a planarian ortholog, but despite the recognition of a short, nearly identical *Schmidtea* sequence fragment, we were unable to identify the rest of the hypothetical gene within the assembly.

**The Glutamate Family**

*Glutamate* GPCRs respond to a wide range of signals, including glutamate, γ-aminobutyric acid (GABA), Ca$^{2+}$ and odorants. The mammalian complement of metabotropic glutamate receptors (mGluRs) consists of 8 proteins that fall into 3 groups. They universally possess a large extracellular domain that contains within it a ligand binding domain (LBD). The *Drosophila* mGluR-like complement consists of two receptors, DmGluRA and DmXR. DmGluRA shares the structural profile of mammalian mGluR2 and mGluR3. DmXR constitutes one member of a larger insect-specific clade, and displays an atypically-diverged LBD that responds to L-canavanine [177,178]. Outside of the metazoa, a group of 17 *Dictyostelium* GABA$_B$-like receptors (GrlA-GrlR) have been forwarded as potential evolutionary precursors to mGluRs [179,180].

We identified 2 *Schistosoma* and 9 *Schmidtea* Glutamate-like sequences. Phylogenetic analysis of these sequences was performed with respect to both mammalian and non-mammalian Glutamate receptors (Figure 2.9). The *S. mansoni* Glutamate-like receptors both have corresponding orthologs in the *Schmidtea* genome. GSMP052660 and its ortholog GSMD025402 group with DmGluRA, and most of the remaining planarian sequences fall in the phylogenetic vicinity of the major mGluR groupings. However, GSMP128940 and its ortholog GSMD001419, along with GSMD004608, seem to be significantly diverged from both GABA$_B$ and mGluR receptors. In the case of DmXR and Grl receptors, the examination of key LBD residues involved
in glutamate binding led to the eventually validated conclusion that glutamate was not the primary ligand. We perform similar analysis as depicted in Figure 2.10. Although the residues of GSMD001419 involved in $\alpha$-amino and $\alpha$-carboxylic groups of glutamate are conserved, the residues associated with the $\gamma$-carboxylic group are not. This runs parallel to the observations made for DmXR. GSMP128940 displays an even more atypical LBD and conserves only a single putative glutamate-interacting residue. We hypothesize that these particular receptors either bind other amino acid-derived ligands or possess unusual pharmacological profiles.

**The Frizzled Family**

Wnt-mediated *Frizzled* signaling plays a significant regulatory role in a number of crucial developmental processes, including cell fate determination, cell motility, cell polarity, and synaptic organization [181]. In planaria, the canonical Wnt/$\beta$-catenin pathway is implicated as a molecular switch for anteroposterior polarity in regeneration [156,182]. We identified four *S. mansoni* Frizzled sequences, along with the 10 *S. mediterranea* sequences previously identified. A single Smoothened-like sequence was found for each species.

In humans, 10 Frizzled receptors are grouped into four clusters based on sequence identity: Fzd1/Fzd2/Fzd7 (I), Fzd5/Fzd8 (II), Fzd4/Fzd9/Fzd10 (III), and Fzd3/Fzd6 (IV) [181]. Both flatworm genomes house a single receptor (FSMP118970 and FSMD000018) that groups in cluster IV, sharing $\sim$45% amino acid identity with Drosophila Fzd1 and $\sim$38% identity with human Fzd6. Four planarian (FSMD023435, FSMD010098 and FSMD000054) and two schistosome (FSMP139180 and FSMP155340) receptors appear to belong to cluster II. Other flatworm Frizzled receptors show less clear relationships with their vertebrate counterparts.

**Ligand-based support vector machine (SVM) Rhodopsin subclassification**

Support vector machines (SVMs) represent a powerful supervised-learning method for data classification. Given a combined set of positively and negatively labeled training instances, an SVM produces a binary classifier that can then be used to label unknown samples. Each in-
stance is associated with a fixed-length numerical feature vector, containing certain attributes of the data to be classified. The SVM identifies a maximum-margin separating hyperplane to distinguish between vectors representing instances of opposite sign. More often than not, training instances are not linearly separable in the feature space, and feature vectors must first be mapped to a higher dimensional space. Non-linear classification is then performed by application of kernel functions which allow for the construction of a hyperplane in the transformed feature space. Recently, this approach has seen extensive use in the area of biosequence discrimination, and relevant to our goals, the particular problem of GPCR classification.

In the first study on the matter, SVM-based classifiers were shown to drastically outperform their BLAST and HMM-based counterparts for level 1 and level 2 GPCR subclassification [132]. Subsequent studies further improved the predictive performance of SVMs with the introduction of dipeptide composition feature vectors [133,134], achieving accuracies of 97.3% and 96.4% for level 1 (Rhodopsin) and level 2 (amine) classification, respectively. Alternative feature vectors have since been similarly validated [135, 136]. Although these computational approaches are touted as among the most sensitive, to the best of our knowledge, they have seen no utilization in the realm of genome-wide GPCR mining studies.

Perhaps one reason for this is that even in the case of publicly available SVM classifiers, training and validation occurs exclusively with full-length sequence data. More suitable classifiers would be tailored to the general deficiencies of sequence data resulting from in silico methods, where inexact gene structures are an unavoidable phenomenon. In this respect, we developed a classifier to complement our particular GPCR identification approach. This involved focusing SVM training on transmembrane domains, as identification of these conserved blocks had been a primary aim of both our receptor mining and manual curation protocols.
Multi-class SVM

Multi-class SVMs refer to classification problems where the number of classes, $k$, is greater than 2. Such problems are typically solved using either the “one-versus-rest” (OvR) or “one-versus-one” (OvO) method. In the OvR scenario, $k$ binary classifiers are trained, such that each classifier separates one class from all others. The “winner-takes-all” strategy is then commonly used to label unknown samples, whereby the classifier with the highest output decision function assigns the final class. In the OvO scenario, $\frac{k(k-1)}{2}$ binary classifiers are constructed in a pair-wise manner. A voting strategy is then typically employed in classification, whereby each classifier accounts for one vote and the class with the maximum number of votes assigns the final label. Although the OvO method has been shown to perform better on a number of fronts [137], as far as the authors are aware, all previously described SVM-based GPCR classifiers available for online use rely on the simpler OvR method. We constructed OvO GPCR classifiers for two levels of Rhodopsin family sub-classification.

Building feature vectors for ligand-based receptor discrimination

The general fixed-length feature vector, $\vec{F}$, contains frequency information for the $20^2(400)$ possible dipeptides over a given stretch of sequence, $L$ amino acids in length. Dipeptides are counted in both possible frames and there are therefore $L - 1$ total amino acid pairs.

$$\vec{F} = \langle P_1, P_2, ..., P_{399}, P_{400} \rangle \tag{2.1}$$

$$P_i = \frac{f_i}{L - 1} \tag{2.2}$$

where $f_i$ represents the frequency of dipeptide $i$

To better associate an SVM-based classification approach with our gene-mining strategy, we explored the idea of again focusing our efforts exclusively on the transmembrane domains. Two options in the way of final feature vector construction were pursued: $\vec{X}_{T1}$ and $\vec{X}_{T7}$. $\vec{X}_{T1}$ represents the 400-element dipeptide frequency vector taken over the entire length of a TM-only pseudosequence, while $\vec{X}_{T7}$ represents the 2800-element dipeptide frequency vector generated
from the ordered concatenation of the dipeptide frequency vectors for the seven individual TM domains. The standard dipeptide frequency vector calculated for full-length proteins, \( \vec{X}_{FL} \), was used for comparison. We will refer to the corresponding SVM classifiers as SVM\(_{T1} \), SVM\(_{T7} \), and SVM\(_{FL} \).

\[
\vec{X}_{T1} = \{ \vec{F} \}_{TM1 - TM7} \tag{2.3}
\]

\[
\vec{X}_{T7} = \{ \vec{F} \}_{TM1} \oplus \{ \vec{F} \}_{TM2} \oplus \{ \vec{F} \}_{TM3} \oplus \{ \vec{F} \}_{TM4} \oplus \{ \vec{F} \}_{TM5} \oplus \{ \vec{F} \}_{TM6} \oplus \{ \vec{F} \}_{TM7} \tag{2.4}
\]

\[
\vec{X}_{FL} = \{ \vec{F} \}_{FL} \tag{2.5}
\]

**SVM training: cross-validation and grid search**

*Rhodopsin* training sequences were divided into 17 subfamilies using the GPCRDB classification system. Programs were written to process this training data into feature vector form (Appendix F). Training was performed with the radial basis function (RBF) kernel, \( K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \), and a grid search was used to tune parameters \( \gamma \) and \( C \) with 5-fold cross-validation. For each proposed feature vector construction, the best performing \((C, \gamma)\) pair was selected in domains \( C = 2^{-5}, 2^{-4}, \ldots, 2^{15} \) and \( \gamma = 2^{-15}, 2^{-14}, \ldots, 2^{15} \) and used to train a final classifier (Table 2.7).

Our original expectation was that SVM\(_{T1} \) would display lower accuracy than SVM\(_{FL} \), given that a smaller subset of sequence information would be used for training. We hoped that this presumed disparity would be compensated by SVM\(_{T7} \) with the addition of position-specific information. Instead, both SVM\(_{T1} \) and SVM\(_{T7} \) registered higher cross-validation accuracies than SVM\(_{FL} \) for *Rhodopsin* subfamily classification. \( \vec{X}_{T7} \) was the best-performing classifier with 99.47% accuracy. These results led us to conclude that for the *Rhodopsin* family, the exclusion of sequence information outside of the transmembrane bundle improves dipeptide-based SVM classification. Encouragingly, this is in agreement with structure and ligand interaction data for the *Rhodopsin* family [183]. The same procedure was carried out in constructing classifiers
for amine GPCRs. SVM$^T_1$ was the best performing classifier with a cross-validation accuracy of 96.44%.

**SVM classification results**

*Rhodopsin* sequences with seven TM domains as predicted by HMMTOP were classified by the two-tiered SVM. TOPs were aligned and manually examined to correct for erroneously predicted TM domains. Sequences were then subclassified with the *Rhodopsin* SVM$^T_7$ classifier, and those discerned as amine-responsive were further sub-classified with the amine classifier SVM$^T_1$. A total of 121 *S. mediterranea* and 58 *S. mansoni* sequences were classified via *Rhodopsin* SVM. The majority of these receptors were identified as peptide-responsive (Table 2.7). This grouping also contains all PROF1 receptors included in this classification stage, perhaps providing some clues as to their ligands. A subset of 22 *S. mediterranea* and 21 *S. mansoni* sequences were identified as amine-responsive, and classified via amine SVM$^T_1$. These classification outputs are detailed in Table 2.9. These results can inform receptor deorphanization efforts, alongside traditional homology-based approaches.

### 2.3 Conclusions

This is the first comprehensive genome-wide study of G protein-coupled receptors in the phylum Platyhelminthes. Our transmembrane-focused receptor mining approach yielded a lower-bound estimate of 116 *S. mansoni* and 333 *S. mediterranea* GPCRs. Phylogenetic analysis established the presence of the primary metazoan GRAFS families, along with well-populated $\alpha$ and $\beta$ *Rhodopsin* subfamilies in both examined genomes. The identification of these receptors complements previous and ongoing efforts to identify biogenic amine and neuropeptide-like ligands in flatworms, and will help identify specific receptors that mediate important aspects of flatworm biology associated with the aminergic and peptidergic signaling systems.

The flatworm GPCR repertoire is also shown to house entirely novel receptor groupings with large numerical representation, including a Platyhelminth-specific *Rhodopsin* subfamily (PROF1).
and a planarian-specific Adhesion-like family (PARF1). These particular lineage-specific expansions, along with the many other highly-diverged receptors identified, may reveal functional innovations specific to these organisms. Many of these receptors have enhanced appeal as selective pharmacological targets. While their diverged structures are an attractive feature in the parasite drug discovery paradigm, it presents a challenge in posing more exact hypotheses related to receptor function.

To further aid the process of functionally pairing receptors and ligands, we provide a preliminary classification of full-length receptors using SVMs. This represents the first effort to apply SVMs to the problem of GPCR classification in a whole-genome manner, a task made difficult by the evolutionary distance of flatworms from other species with well-characterized GPCR complements. SVM results may be used in conjunction with phylogenetic and homology-based approaches to receptor classification. As the quality of the underlying gene models improves, and as a greater number of full-length receptor transcripts are sequence characterized, these SVMs can be applied to an expanding subset of identified GPCRs. Functional characterization of flatworm GPCRs is also likely to improve SVM accuracy by providing better training examples.

The notion that schistosome GPCRs represent lucrative anthelmintic drug targets is strengthened by data on the crucial biological role of related receptor signaling molecules in nearly-related organisms [184,185], as well as that of predicted platyhelminth GPCR ligands [164,165, 169,171]. The receptors, ligands and downstream biochemical pathways associated with GPCR signaling have been identified as potential targets for parasite life-cycle interruption [15, 186]. Enlistment of schistosome reverse genetics approaches alongside receptor sequence data can lead to the validation of specific receptors as drug targets.

In this regard, RNAi in schistosomes [69,70] provides new opportunities for focused exploitation of this dataset. A simple medium-throughput phenotypic classification system has recently
been described for both schistosomula and adult schistosomes [14]. These endpoints could readily be used in an RNAi-mediated GPCR loss-of-function screen. Assaying the temporal expression profiles of parasite GPCRs can also be a worthwhile measure as a selection tool for receptors expressed in intra-host stages. On this front, we further the case for planarians as a convenient model organisms to interrogate the function of trematode receptors, and provide a list of inter-species receptor pairings ranked by sequence identity.

2.4 Acknowledgements

The authors wish to acknowledge the sequencing centers and laboratories responsible for the publicly-accessible genomic data that made this work possible. We would like to thank Dr. Alejandro Sanchez Alvarado and Eric Ross for providing us the MAKER predicted gene set. This work was funded by a grant from the National Institutes of Health (NIH R01 AI49162) to TAD and AGM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.5 Methods

Predicted proteomes and training sequences

The most recent release of the *S. mansoni* genomic assembly is accompanied with a set of 13,197 predicted proteins [24]. The *S. mediterranea* predicted proteome consists of 31,955 predicted proteins that were produced with MAKER, although this number may represent a significant overestimate of the true protein count [26,27]. HMM and SVM training sequences were downloaded from GPCRDB [107] in FASTA file format. In total, 268 Glutamate, 5025 Rhodopsin, 175 Adhesion, 354 Frizzled and 185 Secretin sequences were procured for HMM training. 20,920 GPCRDB sequences were used for Rhodopsin SVM training, and 2,105 sequences were used for amine SVM subclassification.
Nomenclature

Putative receptors retain their original GeneDB or MAKER IDs in slightly modified form. In cases where a gene model was created, receptors were given a label in similar form that includes genomic contig or scaffold information. Letters are appended to the ends of these labels where necessary to distinguish among multiple gene models associated with a single contig or scaffold. All putative flatworm GPCR sequences are provided in association with their tentative IDs.

Transmembrane domain prediction

We applied two common algorithms, TMHMM 2.0 [129] and HMMTOP 2.1 [128], to identify transmembrane domains in our GPCR training set. HMMTOP correctly predicted 7 TM domains for 93.8% (4712/5025) of *Rhodopsin* family receptors, compared to 81.9% (4119/5025) in the case of TMHMM. This disparity in sensitivity held for all GPCR families, and was the basis for our decision to employ HMMTOP for most subsequent work. A robust Perl script (Appendix F) was written to parse coordinate predictions output from HMMTOP, and to generate sequence files containing only regions of interest from the original protein sequences as required.

TM-focused Profile hidden Markov model (HMM) construction

Provided a multiple sequence alignment, HMMER-2.3.2 [126] builds a probabilistic model (profile HMM) that can be used to query sequence databases to find (or align) homologous sequences. To prepare each GPCR family training set, predicted TM domains flanked bi-directionally by 5 amino acids were extracted and concatenated using coordinates produced in the previous section. These sequences were aligned with Muscle 3.6 [187] and a profile HMM was constructed for each family with *hmmbuild*. All models underwent calibration using *hmmcalibrate*, with the default parameters.
HMM-based GPCR identification

All predicted proteins in the *S. mansoni* and *S. mediterranea* genomes with a predicted number of TM domains in the range of 3-15 were processed in a manner identical to the HMM training set. These TOP-converted protein sets were searched against our family-specific profile HMMs using *hmmpfam*. The resulting hits for each GPCR family were ranked according to e-value, and a cut-off was selected at the point where subsequent hits showed significant homology to other known proteins or GPCRs belonging to other families. This was accomplished with a BLASTp search of all hits against the NBCI nr database. The BLAST results were parsed with a script (Appendix F) and top results were examined for removal of false positives.

Manual curation of putative GPCR-encoding genes

A large number of GPCR sequences underwent manual inspection of gene structure, and the original predictions were edited where possible. Common manual edits included the merging or splitting of gene models, modification of intron-exon boundaries, and sequence extension or truncation in either or both directions. All editing was performed with Artemis [188]. Curation was primarily guided by homology-based searches and identification of TM domains and family-specific GPCR motifs in ORFs that occurred in the vicinity of a gene model. In the case of *S. mansoni*, this labor-intensive process was aided by the extraction of GeneDB annotations for scaffolds thought to contain one or more receptors. More specifically, a script was written to compile pertinent scaffold information stored in EMBL formatted files, including the orientation, the number of predicted transmembrane domains and the top BLAST hits for all proteins identified by our profile HMMs. This data was parsed into a spreadsheet and proved significant in helping identify instances where manual curation was appropriate. In the case of *S. mediterranea*, annotated genomic regions were loaded into Artemis in GFF3 format and edited in a similar manner.
Phylogenetic analysis

Near full-length (TM > 5) receptors were first processed for removal of the N- and C-termini. ClustalX 2.0 [189] was used to generate multiple sequence alignments of the GPCRs to be examined, with default parameters. PFAAT [190] was used to edit the resulting alignment with attention to key motifs and residues housed within transmembrane domains. Low-entropy sequence blocks present in all sequences were retained. The Phylip 3.6 [191] package was used to generate phylogenetic trees. Alignments were bootstrapped using seqboot. Maximum parsimony trees were calculated with protpars with input order randomized. Neighbor-joining trees were calculated with protdist and neighbor using the JTT (Jones-Taylor-Thornton) distance matrix and with input order randomized. Consensus trees were built with consense, and visualized and edited with FigTree.

PROF1 RT-PCR

Total RNA was extracted from flatworm (schistosome or planarian) tissue using the RNAqueous Kit (Ambion), and RNA was treated with Turbo DNAase (Ambion) per manufacturer’s instructions. A two-step RT-PCR was performed, where reverse transcription was first carried out with the Retroscript kit (Ambion). Primers were designed for two schistosome PROF1 sequences and 13 planarian PROF1 sequences using Primer 3.0 [192] (Appendix C). PCR products were visualized by agarose gel electrophoresis to confirm transcript expression.

SVM

Programs were written to process training sequences into feature vector form for the training of three SVM classifiers: SVM$_{T1}$, SVM$_{T2}$, and SVM$_{FL}$ (Appendix F). TM prediction was performed on training sequences with HMMTOP, and fixed-length dipeptide frequency vectors were calculated in correspondence with each model. SVMs were implemented with the the LIBSVM [193] package. The RBF kernel was chosen and a grid-search was performed with an available python script for selection of kernel parameters. $C$ and $\gamma$ were assayed in the domains $C = 2^{-5}, 2^{-4}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^{15}$ to identify the $C, \gamma$ pair that
maximizes 5-fold cross validation ACC. The classifiers were trained in accordance with the GPCRDB ligand-based groupings, and applied to a subset of flatworm *Rhodopsin* receptors with 7 predicted TM domains.

### 2.6 Figures

**Figure 2.1 - Transmembrane domain-focused GPCR sequence mining strategy.**

(A) Family-specific profile HMMs are built using TM-only pseudosequences (TOPs) extracted from the GPCRDB [107] sequence repository. The predicted proteomes of both *S. mansoni* and *S. mediterranea* are processed in a manner identical to that of the training sequences and are searched against a set of family-specific profile HMMs. Results are ranked statistically and sequences meeting a conservatively selected cutoff undergo an automated BLASTp campaign against the NCBI “nr” database. The output is parsed, and transmembrane proteins exhibiting significant homology to non-GPCR proteins are removed. Redundant sequences are removed with the BLAT utility. The surviving sequence pool is then manually assessed and curated, followed by tBLASTn of sequences against the whole genome assemblies. *Adhesion* and *Secretin* GPCR sequences are distinguished from one another by inspection of their N-terminal ectodomains. Putative full-length *Rhodopsin* GPCRs, defined by the presence of an intact 7TM domain, are sub-classified via SVM. (B) Construction of TOPs is a two-step process involving the prediction of TM boundary coordinates by HMMTOP, followed by the ordered concatenation of TM domains flanked bi-directionally by 5 amino acids.

**Figure 2.2 - HMM-based identification of *S. mansoni* and *S. mediterranea* GPCRs.**

The transmembrane frequency distribution of the *S. mansoni* (left) and *S. mediterranea* (right) predicted proteomes is shown as predicted by HMMTOP at various junctures of the bioinformatics protocol for the *Rhodopsin* family. The top graphs overlay the HMM-derived sequence pools (black, yellow outline) on top of the entire predicted proteomes (white, black outline) in the assayed TM domain range (3-15). The middle graphs overlay the BLASTp filtered sequence pools (black, yellow outline) on top of the HMM-derived pools (white, black outline). The bot-
tom graphs display the final distributions upon filtering, and after the removal of redundant sequences in the case of *S. mediterranea*.

**Figure 2.3 - Manual curation and expansion of *Schistosoma mansoni* GRAFS GPCRs.**

The transmembrane distributions for the filtered *S. mansoni* HMM pool is shown before (top) and after (middle) manual editing of the underlying gene models, as predicted by HMMTOP. The number of GPCRs with a predicted intact 7TM domain increases from 27 to 41, coupled with a significant contraction of the distribution spread. The mean TM count shifts from 6.00 to 6.41, which equates to the identification and addition of roughly 42 missing TM domains during the first round of curation. Homology-bases searches against the genome assembly increased the putative 7TM receptor count to 59 (bottom). Receptors in the 8 and 9 TM bin can be considered full-length for our purposes, as the erroneously-predicted additional TM domains can be excised for phylogenetic analysis. Inclusion of these receptors brings the total putative full-length (7TM) receptor tally to 68 (of 116 total sequences).

**Figure 2.4/ 2.5 - GRAFS and *Rhodopsin* phylogenetic trees.**

(A) Overall topological view of the combined *S. mansoni* and *S. mediterranea* GPCR complements. Maximum parsimony analysis (bootstrap value = 100) was carried out using putative full-length non-*Rhodopsin* GPCRs and a subset of full-length *Rhodopsin*-like GPCRs. In addition to the phylogenetic clustering of sequences into the primary GRAFS families, this analysis reveals the presence of two distinct phylum-specific groupings: PROF1 and PARF1. * Sequence family is present in *S. mediterranea*. ** Sequence family is present in both *S. mansoni* and *S. mediterranea*. (B) Neighbor-joining tree of flatworm *Rhodopsin*-like GPCRs. To maximize the number of sequences included in this analysis, a sequence block housing TM domains I-IV was extracted from the overall alignment. This allowed for inclusion of 312 *Rhodopsin*-like sequences: 90 *S. mansoni* and 224 *S. mediterranea* receptors (bootstrap value = 200). The α (amine and opsin), peptide, melatonin, and PROF1 groupings are highlighted. Branches terminating in *Schistosoma* receptors are shown in green, and branches terminating in *Schmidtea*
receptors are shown in blue. Original consensus tree with bootstrap values and sequence labels is available.

**Figure 2.6 - Aminergic receptors: S. mediterranea and S. mansoni.**

Neighbor-joining tree (bootstrap value = 500) of putative biogenic amine-responsive GPCRs. Included in this analysis are 21 (of 25) S. mansoni and 58 (of 65) S. mediterranea full-length and near full-length aminergic receptors, alongside 14 known C. elegans biogenic amine receptors. The latter grouping includes receptors that respond to tyramine, octopamine, dopamine, serotonin, and acetylcholine [194]. Branch lengths are scaled to bootstrap support, branches terminating in Schistosoma receptors are shown in green, and branches terminating in Schmidtea receptors are shown in blue. Flatworm receptors are outlined (solid lines) and classified by ligand and with respect to their nearest-related C. elegans homologs. Two diverged flatworm-specific receptor groupings are outlined in dashed lines.

**Figure 2.7 - Phylogenetic analysis of PROF1 GPCRs.**

Maximum parsimony tree for all identified PROF1 receptors. An alignment block that included TM domains I-IV was bootstrapped 1000 times for parsimony analysis. PROF1 can be subdivided into 3 families with good bootstrap support (> 50%; relevant values displayed): I, II and III. Schistosome sequences are shown in green and Schmidtea sequences are shown in blue. The tree is rooted with a Schistosoma opsin-like GPCR (AAF73286.1). Schmidtea PROF1 receptors with transcript expression confirmed by RT-PCR are marked with red asterisks.

**Figure 2.8 - PROF1 multiple sequence alignment.**

Multiple sequence alignment of all PROF1 receptors over a sequence range that includes TM domains I-IV (used for phylogenetic analysis). Residues are colored according to an identity threshold set at 80% within each group. The locations of individual TM domains were approximated by alignment to Rhodopsin and are depicted above the MSA. Red asterisks are used
to mark residue locations where the among-group PROF1 identity level threshold (≥ 80%) is met.

**Figure 2.9 - Phylogenetic analysis of Glutamate GPCRs.**

Maximum parsimony tree of *Glutamate* family GPCRs. TM domains I-VII were used for phylogenetic analysis with the alignment bootstrapped 1000 times (bootstrap support values are provided). *Schistosome* sequences are shown in green and *Schmidtea* sequences are shown in blue. GSMD007320 and GSMD015264 were excluded as they remain incomplete over the sequence range used. GABA<sub>B</sub> receptors are highlighted, along with the primary vertebrate mGluR groupings and the more recently discovered insect Group X receptors. A human Calcium-sensing receptor (AAA86503.1) was used as an outgroup. Putative flatworm GPCRs that are diverged from both the GABA<sub>B</sub> and glutamate-responsive receptors are outlined in red. The ligand-binding domains of these receptors are further analyzed in Figure 9.

**Figure 2.10 - Schematic of glutamate in association with LBD residues.**

Conserved mGluR LBD residues involved in glutamate binding are shown (underlined) in comparison with the corresponding residues for flatworm *Glutamate*-like receptors GSMP128940 and GSMD004608. Numbers represent residue location with respect to the mouse mGluR3 sequence. Disagreement at a given position is highlighted in red. GSMP128940 displays overall divergence with the canonical glutamate binding pocket, while GSMD004608 retains only key residues that interact with the glutamate α-carboxylic and α-amino groups.
Figure 2.1 Transmembrane domain-focused GPCR sequence mining strategy.
Figure 2.2  HMM-based identification of *S. mansoni* and *S. mediterranea* GPCRs.
Figure 2.3  Manual curation and expansion of *Schistosoma mansoni* GRAFS GPCRs.
Figure 2.4  GRAFS phylogenetic tree.
Figure 2.5  *Rhodopsin* phylogenetic tree.
Figure 2.6  Aminergic receptors: *S. mediterranea* and *S. mansoni*. 
Figure 2.7  Phylogenetic analysis of PROF1 GPCRs.
Figure 2.8  PROF1 multiple sequence alignment.
Figure 2.9  Phylogenetic analysis of *Glutamate* GPCRs.
Figure 2.10 Schematic of glutamate in association with LBD residues.
2.7 Tables

Table 2.1 - Tabulated GPCR sequence count at various stages of processing.

Sequence counts are provided for each GPCR family at different stages in the receptor mining protocol. The *S. mansoni* count is shown after application of the TM-focused HMM and filtering of the predicted proteome (HMM), extensive manual curation (MC), homology-based searches against the nucleotide assembly (tBLASTn) and a final round of manual curation (Final). This progression is similarly displayed for *S. mediterranea*, with additional stages for the shedding of redundant sequences using BLAT (R).

Table 2.2 - GRAFS-based comparison of GPCR repertoires.

The GPCR repertoires of *S. mansoni* and *S. mediterranea* are shown from a GRAFS-based perspective, alongside those of other organisms with characterized GPCR complements. For *Rhodopsin* sub-classification, BLAST searches were used to help tentatively assign putative ligands to receptors omitted from phylogenetic analysis.

Tables 2.3, 2.4, 2.5 - Identification of Planarian sequelogs of parasite *Rhodopsin* receptors.

The nearest *S. mediterranea* sequelog of each *S. mansoni Rhodopsin* receptor is shown, along with the length of the BLASTp overlap region and the corresponding E-value. Receptor pairs are ranked by percent identity (PID). Parasite receptors closest to top of the table are likely candidates for indirect characterization via investigation of their nearest-related planarian counterpart.

Table 2.6 - Comparison of PROF1 motifs and classical *Rhodopsin* motifs.

PROF1 motifs are compared to ubiquitous *Rhodopsin* family motifs. Motifs are displayed as regular expressions. The two most frequently occurring amino acids are shown for each position in order of frequency, except in cases where a particular residue is absolutely conserved or when there is no clear second in frequency rank. Red text is used to highlight positional agreement.
between PROF1 motifs and the corresponding \textit{Rhodopsin} motifs. More specifically, instances where the most frequently occurring residue is equivalent.

\textbf{Table 2.7 - Rhodopsin SVM training parameters and cross-validation accuracy.}

RBF grid-search parameters used to train SVM models, along with the corresponding 5-fold cross-validation accuracy (ACC) for the training set. The best performing model for level 2 (subfamily) classification is SVM\textsubscript{T7}, while the best performing model for level 3 (amine) classification is SVM\textsubscript{T1}. Both classifiers exclusively employ transmembrane sequence data, and outperform the classifiers trained with full-length sequence data.

\textbf{Table 2.7 - Rhodopsin SVM classifier results.}

Ligand-based classification of flatworm \textit{Rhodopsin} GPCRs with \textit{Rhodopsin} SVM\textsubscript{T7}. PROF1 receptors are labeled with `*'. A total of 178 receptors were classified, with the vast majority placed in the peptide and amine groupings. Interestingly, all 45 PROF1 receptors were classified as peptide-responsive.

\textbf{Table 2.9 - Amine SVM classifier results.}

Ligand-based classification of flatworm amine-responsive GPCRs with amine SVM\textsubscript{T1}. A total of 43 receptors were identified as aminergic via \textit{Rhodopsin} SVM classification. In cases of erroneous TM boundary prediction, the SVM\textsubscript{FL} classifier was used. The classifier results display correct predictions for the three schistosome receptors thus far deorphanized in this subfamily, including two histamine-responsive GPCRs and one dopamine-responsive GPCR (labeled with `*').
### Table 2.1 Stage-specific GPCR tabulation.

<table>
<thead>
<tr>
<th>Family</th>
<th>S. mansoni</th>
<th>S. mediterranea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HMM</td>
<td>MC</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td>AS</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

*Note:* HMM, MC, tBLASTn, R denote different stages of the analysis.
Table 2.2  GRAFS-based comparison of GPCR repertoires.

<table>
<thead>
<tr>
<th></th>
<th>H. sapiens</th>
<th>A. gambiae</th>
<th>D. melanogaster</th>
<th>C. elegans</th>
<th>S. mansoni</th>
<th>S. mediterranea</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>AMIN (α)</td>
<td>44</td>
<td>18</td>
<td>21</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>MEC (α)</td>
<td>22</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>MTN (α)</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OPN (α)</td>
<td>11</td>
<td>12</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>PTGER (α)</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PEP (β)</td>
<td>43</td>
<td>29</td>
<td>21</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>CHEM (γ)</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>MCHR (γ)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SOG (γ)</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LGR (δ)</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>MRG (δ)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OLF (δ)</td>
<td>494</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PUR (δ)</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PROF1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Unclassified</td>
<td>20</td>
<td>77</td>
<td>79</td>
<td>124</td>
<td>21</td>
</tr>
<tr>
<td>F</td>
<td>FZD/SMO</td>
<td>11</td>
<td>13</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>TAS2</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>GLR</td>
<td>24</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>A/S</td>
<td>ADH</td>
<td>27</td>
<td>13</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SEC</td>
<td>20</td>
<td>1</td>
<td>13</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>PARF1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>\textit{S. mansoni} Query</td>
<td>Length</td>
<td>\textit{S. mediterranea} Hit</td>
<td>Length</td>
<td>Overlap</td>
<td>E-value</td>
<td>PID</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------</td>
<td>--------------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>SMP145540</td>
<td>492</td>
<td>mk4.017782.00.01</td>
<td>278</td>
<td>206</td>
<td>6.00E-90</td>
<td>69%</td>
</tr>
<tr>
<td>SMP148210</td>
<td>279</td>
<td>mk4.007388.02.01</td>
<td>339</td>
<td>196</td>
<td>2.00E-65</td>
<td>62%</td>
</tr>
<tr>
<td>SMP169680</td>
<td>178</td>
<td>mk4.000219.07.01</td>
<td>424</td>
<td>82</td>
<td>3.00E-25</td>
<td>62%</td>
</tr>
<tr>
<td>SMP118040</td>
<td>354</td>
<td>mk4.000375.06.01</td>
<td>371</td>
<td>314</td>
<td>1.00E-103</td>
<td>60%</td>
</tr>
<tr>
<td>SMP126730</td>
<td>463</td>
<td>mk4.007388.02.01</td>
<td>339</td>
<td>192</td>
<td>7.00E-62</td>
<td>59%</td>
</tr>
<tr>
<td>SMP152540</td>
<td>231</td>
<td>mk4.009070.00.01</td>
<td>704</td>
<td>131</td>
<td>5.00E-43</td>
<td>59%</td>
</tr>
<tr>
<td>SMP160020</td>
<td>320</td>
<td>mk4.013690.00.01</td>
<td>458</td>
<td>157</td>
<td>3.00E-43</td>
<td>52%</td>
</tr>
<tr>
<td>SMP149770</td>
<td>318</td>
<td>mk4.013819.05.01</td>
<td>308</td>
<td>141</td>
<td>5.00E-39</td>
<td>52%</td>
</tr>
<tr>
<td>SMP134820</td>
<td>360</td>
<td>mk4.011160.01.01</td>
<td>484</td>
<td>197</td>
<td>2.00E-51</td>
<td>48%</td>
</tr>
<tr>
<td>SMP132410</td>
<td>383</td>
<td>mk4.000375.06.01</td>
<td>371</td>
<td>314</td>
<td>4.00E-78</td>
<td>47%</td>
</tr>
<tr>
<td>SMP133550</td>
<td>503</td>
<td>mk4.010211.00.01</td>
<td>370</td>
<td>236</td>
<td>2.00E-53</td>
<td>47%</td>
</tr>
<tr>
<td>SMP140620</td>
<td>170</td>
<td>mk4.018209.00.01</td>
<td>365</td>
<td>171</td>
<td>3.00E-39</td>
<td>47%</td>
</tr>
<tr>
<td>SMP123350</td>
<td>535</td>
<td>mk4.000301.14.01</td>
<td>386</td>
<td>87</td>
<td>7.00E-24</td>
<td>47%</td>
</tr>
<tr>
<td>SMP127720</td>
<td>518</td>
<td>mk4.012183.00.01</td>
<td>509</td>
<td>167</td>
<td>1.00E-39</td>
<td>44%</td>
</tr>
<tr>
<td>SMP120620</td>
<td>587</td>
<td>mk4.002108.03.01</td>
<td>224</td>
<td>112</td>
<td>3.00E-26</td>
<td>44%</td>
</tr>
<tr>
<td>SMP001070.1</td>
<td>197</td>
<td>SMDC2955.1</td>
<td>334</td>
<td>126</td>
<td>4.00E-26</td>
<td>44%</td>
</tr>
<tr>
<td>SMP000900</td>
<td>412</td>
<td>mk4.001208.04.01</td>
<td>152</td>
<td>141</td>
<td>3.00E-25</td>
<td>43%</td>
</tr>
<tr>
<td>SMP150180</td>
<td>483</td>
<td>mk4.012659.00.01</td>
<td>387</td>
<td>417</td>
<td>9.00E-80</td>
<td>42%</td>
</tr>
<tr>
<td>SMP043340</td>
<td>551</td>
<td>mk4.005650.02.01</td>
<td>421</td>
<td>189</td>
<td>2.00E-35</td>
<td>41%</td>
</tr>
<tr>
<td>SMP043290</td>
<td>497</td>
<td>mk4.005650.02.01</td>
<td>421</td>
<td>168</td>
<td>2.00E-34</td>
<td>41%</td>
</tr>
<tr>
<td>SMPSC31</td>
<td>241</td>
<td>mk4.002460.05.01</td>
<td>306</td>
<td>160</td>
<td>1.00E-27</td>
<td>41%</td>
</tr>
<tr>
<td>SMP172810</td>
<td>450</td>
<td>mk4.000557.00.01</td>
<td>337</td>
<td>75</td>
<td>2.00E-12</td>
<td>41%</td>
</tr>
<tr>
<td>SMP134350</td>
<td>421</td>
<td>mk4.002485.01.01</td>
<td>254</td>
<td>213</td>
<td>9.00E-49</td>
<td>40%</td>
</tr>
<tr>
<td>SMP011940</td>
<td>396</td>
<td>mk4.002418.01.01</td>
<td>447</td>
<td>321</td>
<td>2.00E-72</td>
<td>39%</td>
</tr>
<tr>
<td>SMP194740</td>
<td>265</td>
<td>mk4.012712.00.01</td>
<td>279</td>
<td>143</td>
<td>3.00E-18</td>
<td>39%</td>
</tr>
<tr>
<td>SMP149580</td>
<td>377</td>
<td>mk4.010306.00.01</td>
<td>320</td>
<td>342</td>
<td>1.00E-55</td>
<td>38%</td>
</tr>
<tr>
<td>SMP145240</td>
<td>581</td>
<td>mk4.015843.00.01</td>
<td>270</td>
<td>246</td>
<td>2.00E-40</td>
<td>38%</td>
</tr>
<tr>
<td>SMP058080</td>
<td>319</td>
<td>mk4.000152.09.01</td>
<td>261</td>
<td>218</td>
<td>6.00E-39</td>
<td>38%</td>
</tr>
<tr>
<td>SMPSC12B</td>
<td>244</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>198</td>
<td>6.00E-38</td>
<td>38%</td>
</tr>
<tr>
<td>SMP161500</td>
<td>302</td>
<td>mk4.025533.00.01</td>
<td>287</td>
<td>216</td>
<td>1.00E-37</td>
<td>38%</td>
</tr>
<tr>
<td>SMP043320</td>
<td>584</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>173</td>
<td>5.00E-36</td>
<td>38%</td>
</tr>
<tr>
<td>SMP043260</td>
<td>560</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>183</td>
<td>1.00E-34</td>
<td>38%</td>
</tr>
<tr>
<td>SMPSC103</td>
<td>526</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>171</td>
<td>3.00E-33</td>
<td>38%</td>
</tr>
<tr>
<td>SMP164730</td>
<td>542</td>
<td>mk4.025533.00.01</td>
<td>287</td>
<td>209</td>
<td>1.00E-31</td>
<td>38%</td>
</tr>
<tr>
<td>SMP157050</td>
<td>387</td>
<td>mk4.014274.00.01</td>
<td>409</td>
<td>167</td>
<td>3.00E-25</td>
<td>38%</td>
</tr>
</tbody>
</table>

Table 2.3 Planarian homologs of parasite GPCRs.
<table>
<thead>
<tr>
<th>S. mansoni Query</th>
<th>Length</th>
<th>S. mediterranea Hit</th>
<th>Length</th>
<th>Overlap</th>
<th>E-value</th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMPSC10893</td>
<td>251</td>
<td>SMDC2411.1</td>
<td>385</td>
<td>128</td>
<td>6.00E-19</td>
<td>38%</td>
</tr>
<tr>
<td>SMPSC433</td>
<td>371</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>367</td>
<td>4.00E-66</td>
<td>37%</td>
</tr>
<tr>
<td>SMP041880</td>
<td>342</td>
<td>SMDC2411.1</td>
<td>385</td>
<td>366</td>
<td>2.00E-62</td>
<td>37%</td>
</tr>
<tr>
<td>SMP134100</td>
<td>293</td>
<td>mk4.005562.01.01</td>
<td>324</td>
<td>295</td>
<td>1.00E-51</td>
<td>37%</td>
</tr>
<tr>
<td>SMP043300</td>
<td>522</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>172</td>
<td>5.00E-34</td>
<td>37%</td>
</tr>
<tr>
<td>SMP023710</td>
<td>465</td>
<td>SMDC2411.1</td>
<td>385</td>
<td>134</td>
<td>1.00E-22</td>
<td>37%</td>
</tr>
<tr>
<td>SMP140250</td>
<td>196</td>
<td>mk4.005799.01.01</td>
<td>323</td>
<td>140</td>
<td>1.00E-21</td>
<td>37%</td>
</tr>
<tr>
<td>SMP104210</td>
<td>329</td>
<td>mk4.011006.00.01</td>
<td>348</td>
<td>127</td>
<td>4.00E-20</td>
<td>37%</td>
</tr>
<tr>
<td>SMP146450</td>
<td>367</td>
<td>mk4.012231.01.01</td>
<td>343</td>
<td>127</td>
<td>1.00E-16</td>
<td>37%</td>
</tr>
<tr>
<td>SMP170020</td>
<td>407</td>
<td>mk4.033855.01.01</td>
<td>243</td>
<td>187</td>
<td>8.00E-30</td>
<td>36%</td>
</tr>
<tr>
<td>SMP137050</td>
<td>454</td>
<td>mk4.002108.03.01</td>
<td>224</td>
<td>217</td>
<td>5.00E-35</td>
<td>35%</td>
</tr>
<tr>
<td>SMP157640</td>
<td>482</td>
<td>mk4.022218.00.01</td>
<td>344</td>
<td>127</td>
<td>1.00E-12</td>
<td>35%</td>
</tr>
<tr>
<td>SMP129810</td>
<td>497</td>
<td>mk4.006543.02.01</td>
<td>440</td>
<td>404</td>
<td>2.00E-61</td>
<td>34%</td>
</tr>
<tr>
<td>SMP091950</td>
<td>366</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>363</td>
<td>6.00E-57</td>
<td>34%</td>
</tr>
<tr>
<td>SMP172170</td>
<td>243</td>
<td>mk4.000152.09.01</td>
<td>261</td>
<td>220</td>
<td>3.00E-28</td>
<td>34%</td>
</tr>
<tr>
<td>SMP043270</td>
<td>490</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>175</td>
<td>3.00E-29</td>
<td>33%</td>
</tr>
<tr>
<td>SMP117340</td>
<td>192</td>
<td>SMDC2411.1</td>
<td>385</td>
<td>133</td>
<td>2.00E-21</td>
<td>33%</td>
</tr>
<tr>
<td>SMP164140</td>
<td>272</td>
<td>mk4.004167.00.01</td>
<td>260</td>
<td>145</td>
<td>5.00E-18</td>
<td>33%</td>
</tr>
<tr>
<td>SMP135660</td>
<td>461</td>
<td>mk4.001407.02.01</td>
<td>330</td>
<td>75</td>
<td>0.016</td>
<td>33%</td>
</tr>
<tr>
<td>SMP167870</td>
<td>319</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>291</td>
<td>4.00E-42</td>
<td>32%</td>
</tr>
<tr>
<td>SMPSC1003</td>
<td>373</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>362</td>
<td>1.00E-40</td>
<td>32%</td>
</tr>
<tr>
<td>SMP173010</td>
<td>354</td>
<td>mk4.002460.05.01</td>
<td>306</td>
<td>259</td>
<td>9.00E-37</td>
<td>32%</td>
</tr>
<tr>
<td>SMP056080</td>
<td>532</td>
<td>mk4.012712.00.01</td>
<td>279</td>
<td>283</td>
<td>3.00E-35</td>
<td>32%</td>
</tr>
<tr>
<td>SMPSC12A</td>
<td>305</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>234</td>
<td>3.00E-35</td>
<td>32%</td>
</tr>
<tr>
<td>SMPSC74</td>
<td>411</td>
<td>mk4.001407.02.01</td>
<td>330</td>
<td>277</td>
<td>6.00E-29</td>
<td>32%</td>
</tr>
<tr>
<td>SMP141880</td>
<td>492</td>
<td>mk4.011130.02.01</td>
<td>209</td>
<td>138</td>
<td>6.00E-16</td>
<td>32%</td>
</tr>
<tr>
<td>SMP083880</td>
<td>247</td>
<td>mk4.008619.00.01</td>
<td>542</td>
<td>83</td>
<td>1.00E-07</td>
<td>32%</td>
</tr>
<tr>
<td>SMP132730</td>
<td>278</td>
<td>mk4.024092.00.01</td>
<td>427</td>
<td>67</td>
<td>4.00E-06</td>
<td>32%</td>
</tr>
<tr>
<td>SMP180030</td>
<td>369</td>
<td>mk4.007964.02.01</td>
<td>210</td>
<td>190</td>
<td>2.00E-20</td>
<td>31%</td>
</tr>
<tr>
<td>SMP162870</td>
<td>195</td>
<td>mk4.021981.00.01</td>
<td>316</td>
<td>141</td>
<td>8.00E-20</td>
<td>31%</td>
</tr>
<tr>
<td>SMP084280</td>
<td>355</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>307</td>
<td>1.00E-40</td>
<td>30%</td>
</tr>
<tr>
<td>SMPSC63</td>
<td>358</td>
<td>SMDC6472.1</td>
<td>368</td>
<td>340</td>
<td>4.00E-40</td>
<td>30%</td>
</tr>
<tr>
<td>SMP072450</td>
<td>460</td>
<td>mk4.013492.00.01</td>
<td>417</td>
<td>379</td>
<td>1.00E-36</td>
<td>30%</td>
</tr>
<tr>
<td>SMPSC331A</td>
<td>354</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>333</td>
<td>8.00E-35</td>
<td>30%</td>
</tr>
<tr>
<td>SMP149170</td>
<td>441</td>
<td>mk4.031060.00.01</td>
<td>384</td>
<td>345</td>
<td>2.00E-30</td>
<td>30%</td>
</tr>
</tbody>
</table>

Table 2.4 Planarian homologs of parasite GPCRs- continued.
<table>
<thead>
<tr>
<th>S. mansoni Query</th>
<th>Length</th>
<th>S. mediterranea Hit</th>
<th>Length</th>
<th>Overlap</th>
<th>E-value</th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP178420</td>
<td>328</td>
<td>mk4.004400.01.01</td>
<td>317</td>
<td>149</td>
<td>1.00E-15</td>
<td>30%</td>
</tr>
<tr>
<td>SMP134460</td>
<td>521</td>
<td>mk4.005650.02.01</td>
<td>421</td>
<td>444</td>
<td>3.00E-42</td>
<td>29%</td>
</tr>
<tr>
<td>SMP084290</td>
<td>364</td>
<td>SMDC6472.1</td>
<td>368</td>
<td>340</td>
<td>8.00E-37</td>
<td>29%</td>
</tr>
<tr>
<td>SMP007070</td>
<td>580</td>
<td>mk4.000219.07.01</td>
<td>424</td>
<td>318</td>
<td>1.00E-30</td>
<td>29%</td>
</tr>
<tr>
<td>SMP159860</td>
<td>468</td>
<td>mk4.004728.01.01</td>
<td>352</td>
<td>126</td>
<td>2.00E-13</td>
<td>29%</td>
</tr>
<tr>
<td>SMP132220</td>
<td>135</td>
<td>mk4.008535.00.01</td>
<td>402</td>
<td>78</td>
<td>8.00E-07</td>
<td>29%</td>
</tr>
<tr>
<td>SMP153210</td>
<td>392</td>
<td>mk4.014320.00.01</td>
<td>350</td>
<td>98</td>
<td>6.00E-06</td>
<td>29%</td>
</tr>
<tr>
<td>SMP134960</td>
<td>367</td>
<td>SMDC2411.1</td>
<td>385</td>
<td>345</td>
<td>5.00E-36</td>
<td>28%</td>
</tr>
<tr>
<td>SMP084270</td>
<td>369</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>302</td>
<td>1.00E-33</td>
<td>28%</td>
</tr>
<tr>
<td>SMP059400</td>
<td>291</td>
<td>mk4.005650.02.01</td>
<td>421</td>
<td>338</td>
<td>1.00E-26</td>
<td>28%</td>
</tr>
<tr>
<td>SMP127310</td>
<td>545</td>
<td>mk4.013690.00.01</td>
<td>458</td>
<td>466</td>
<td>1.00E-41</td>
<td>27%</td>
</tr>
<tr>
<td>SMP145520</td>
<td>379</td>
<td>mk4.000354.16.01</td>
<td>413</td>
<td>319</td>
<td>2.00E-36</td>
<td>27%</td>
</tr>
<tr>
<td>SMP083940</td>
<td>352</td>
<td>SMDC8510.2</td>
<td>380</td>
<td>350</td>
<td>8.00E-31</td>
<td>27%</td>
</tr>
<tr>
<td>SMPSC15</td>
<td>409</td>
<td>mk4.000596.05.01</td>
<td>315</td>
<td>310</td>
<td>5.00E-28</td>
<td>27%</td>
</tr>
<tr>
<td>SMP008850</td>
<td>515</td>
<td>mk4.011509.03.01</td>
<td>461</td>
<td>381</td>
<td>3.00E-26</td>
<td>27%</td>
</tr>
<tr>
<td>SMPSC34</td>
<td>365</td>
<td>mk4.002618.00.01</td>
<td>384</td>
<td>314</td>
<td>2.00E-21</td>
<td>27%</td>
</tr>
<tr>
<td>SMP180350</td>
<td>190</td>
<td>mk4.010158.01.01</td>
<td>325</td>
<td>152</td>
<td>2.00E-19</td>
<td>27%</td>
</tr>
<tr>
<td>SMP027940</td>
<td>324</td>
<td>mk4.004400.01.01</td>
<td>317</td>
<td>271</td>
<td>2.00E-18</td>
<td>27%</td>
</tr>
<tr>
<td>SMP180140</td>
<td>354</td>
<td>mk4.001491.09.01</td>
<td>305</td>
<td>190</td>
<td>7.00E-12</td>
<td>27%</td>
</tr>
<tr>
<td>SMP170610</td>
<td>388</td>
<td>mk4.014274.00.01</td>
<td>409</td>
<td>117</td>
<td>4.00E-07</td>
<td>27%</td>
</tr>
<tr>
<td>SMP153200</td>
<td>367</td>
<td>mk4.000600.00.01</td>
<td>372</td>
<td>110</td>
<td>0.003</td>
<td>27%</td>
</tr>
<tr>
<td>SMPSC331B</td>
<td>353</td>
<td>SMDC1889.2</td>
<td>392</td>
<td>334</td>
<td>2.00E-28</td>
<td>26%</td>
</tr>
<tr>
<td>SMP128710</td>
<td>551</td>
<td>mk4.031060.00.01</td>
<td>384</td>
<td>258</td>
<td>7.00E-15</td>
<td>26%</td>
</tr>
<tr>
<td>SMP126890</td>
<td>324</td>
<td>mk4.002418.01.01</td>
<td>447</td>
<td>93</td>
<td>8.00E-08</td>
<td>26%</td>
</tr>
<tr>
<td>SMP137300</td>
<td>328</td>
<td>mk4.006327.00.01</td>
<td>315</td>
<td>292</td>
<td>5.00E-18</td>
<td>25%</td>
</tr>
<tr>
<td>SMP137310</td>
<td>327</td>
<td>mk4.004400.01.01</td>
<td>317</td>
<td>278</td>
<td>6.00E-18</td>
<td>24%</td>
</tr>
<tr>
<td>SMP137320</td>
<td>324</td>
<td>mk4.004400.01.01</td>
<td>317</td>
<td>290</td>
<td>2.00E-17</td>
<td>24%</td>
</tr>
<tr>
<td>SMP041700</td>
<td>367</td>
<td>mk4.021981.00.01</td>
<td>316</td>
<td>228</td>
<td>7.00E-16</td>
<td>24%</td>
</tr>
<tr>
<td>SMP137980</td>
<td>292</td>
<td>mk4.000600.00.01</td>
<td>372</td>
<td>298</td>
<td>2.00E-11</td>
<td>23%</td>
</tr>
<tr>
<td>SMP056060</td>
<td>487</td>
<td>mk4.012712.00.01</td>
<td>279</td>
<td>227</td>
<td>4.00E-10</td>
<td>23%</td>
</tr>
<tr>
<td>SMPSC261</td>
<td>229</td>
<td>mk4.014320.00.01</td>
<td>350</td>
<td>176</td>
<td>0.002</td>
<td>23%</td>
</tr>
<tr>
<td>SMPSC58</td>
<td>323</td>
<td>mk4.006327.00.01</td>
<td>315</td>
<td>283</td>
<td>7.00E-21</td>
<td>22%</td>
</tr>
<tr>
<td>SMP128170</td>
<td>581</td>
<td>mk4.000038.11.01</td>
<td>391</td>
<td>321</td>
<td>2.00E-17</td>
<td>21%</td>
</tr>
<tr>
<td>SMP080820</td>
<td>209</td>
<td>mk4.002569.02.01</td>
<td>367</td>
<td>133</td>
<td>5.00E-05</td>
<td>21%</td>
</tr>
</tbody>
</table>

Table 2.5 Planarian homologs of parasite GPCRs- continued.
<table>
<thead>
<tr>
<th></th>
<th>TM</th>
<th>II</th>
<th>III</th>
<th>VI</th>
<th>VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhodopsin</td>
<td>L</td>
<td>A</td>
<td>D</td>
<td>[D/E]</td>
<td>[Y/H]</td>
</tr>
<tr>
<td>PROF1-I</td>
<td>[L/I]</td>
<td>[A/S]</td>
<td>. .</td>
<td>[D/E]</td>
<td></td>
</tr>
<tr>
<td>PROF1-II</td>
<td>[L/I]</td>
<td>[A/T]</td>
<td>. .</td>
<td>[H/N]</td>
<td>[R/K]</td>
</tr>
</tbody>
</table>

Table 2.6  Comparison of PROF1 motifs and classical Rhodopsin motifs.
<table>
<thead>
<tr>
<th>SVM model</th>
<th>Scoring scheme</th>
<th>$\gamma$</th>
<th>$C$</th>
<th>5-fold ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM$_{T_1}$</td>
<td>OvO</td>
<td>16.0</td>
<td>32.0</td>
<td>99.01%</td>
</tr>
<tr>
<td>SVM$_{T_7}$</td>
<td>OvO</td>
<td>$2^{-8}$</td>
<td>2048.0</td>
<td>99.47%</td>
</tr>
<tr>
<td>SVM$_{FL}$</td>
<td>OvO</td>
<td>256.0</td>
<td>32.0</td>
<td>98.65%</td>
</tr>
</tbody>
</table>

**Level 2: Rhodopsin**

<table>
<thead>
<tr>
<th>SVM model</th>
<th>Scoring scheme</th>
<th>$\gamma$</th>
<th>$C$</th>
<th>5-fold ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM$_{T_1}$</td>
<td>OvO</td>
<td>256.0</td>
<td>32.0</td>
<td>96.44%</td>
</tr>
<tr>
<td>SVM$_{T_7}$</td>
<td>OvO</td>
<td>32.0</td>
<td>$2^{4.5}$</td>
<td>95.0%</td>
</tr>
<tr>
<td>SVM$_{FL}$</td>
<td>OvO</td>
<td>256.0</td>
<td>32.0</td>
<td>94.77%</td>
</tr>
</tbody>
</table>

**Level 3: Amine**

Table 2.7 *Rhodopsin* SVM training parameters and cross-validation accuracy.
<table>
<thead>
<tr>
<th>Amine</th>
<th>S. man.</th>
<th>S. med.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP126730</td>
<td>SMP043290</td>
<td>mk4.0111371.00.01 mk4.050593.01.01 mk4.009528.00.01</td>
</tr>
<tr>
<td>SMP127310</td>
<td>SMP043300 SMP180140</td>
<td>mk4.000656.10.01 mk4.001585.00.01 mk4.009970.00.01</td>
</tr>
<tr>
<td>SMP134460</td>
<td>SMP043320 SMP160020</td>
<td>mk4.003202.01.01 mk4.005562.01.01 mk4.005766.00.01</td>
</tr>
<tr>
<td>SMP134820</td>
<td>SMP043340 SMPFSC103</td>
<td>mk4.000742.09.01 mk4.000883.05.01 mk4.007388.02.01</td>
</tr>
<tr>
<td>SMP027940</td>
<td>SMP145520 SMP19860</td>
<td>mk4.003002.02.01 mk4.002635.00.01 mk4.029325.00.01</td>
</tr>
<tr>
<td>SMP137300</td>
<td>SMP145540</td>
<td>mk4.001678.03.01 mk4.007538.01.01 mk4.012659.00.01</td>
</tr>
<tr>
<td>SMP043260</td>
<td>SMP148210</td>
<td>mk4.011160.01.01 mk4.000664.00.01</td>
</tr>
<tr>
<td>SMP043270</td>
<td>SMP150180</td>
<td>mk4.001569.04.01 mk4.005650.00.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peptide</th>
<th>S. man.</th>
<th>S. med.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP134960*</td>
<td>SMP180140</td>
<td>mk4.0008535.00.01 mk4.009526.00.01 mk4.000699.11.01 mk4.027005.00.01 SMDC21116.2*</td>
</tr>
<tr>
<td>SMP1373010</td>
<td>SMP164730</td>
<td>mk4.002418.01.01 mk4.022218.00.01 mk4.000596.05.01 mk4.001419.04.01 SMDC14497.4*</td>
</tr>
<tr>
<td>SMP135310</td>
<td>SMP056890</td>
<td>mk4.014127.00.01 mk4.008555.00.01 mk4.002457.00.01 mk4.001062.00.01 SMDC958.4*</td>
</tr>
<tr>
<td>SMP135340</td>
<td>SMP027940 SMP180140</td>
<td>mk4.021573.00.01 mk4.000660.00.01 mk4.010162.00.01 SMDC504.1*</td>
</tr>
<tr>
<td>SMP135358</td>
<td>SMP137305</td>
<td>mk4.000375.06.01 mk4.000780.05.01 mk4.022227.01.01 mk4.005939.01.01 SMDC18000.2*</td>
</tr>
<tr>
<td>SMP135315</td>
<td>SMP137300</td>
<td>mk4.002618.00.01 mk4.005682.02.01 mk4.004000.01.01 mk4.001585.00.01 SMDC12781.1*</td>
</tr>
<tr>
<td>SMP135312</td>
<td>SMP060700</td>
<td>mk4.002569.02.01 mk4.006973.03.01 mk4.001491.09.01 mk4.007388.02.01 SMDC15168.2*</td>
</tr>
<tr>
<td>SMP135331A*</td>
<td>SMP084280*</td>
<td>mk4.0008774.00.01 mk4.024092.00.01 mk4.014245.00.01 SMDC943.5* SMDC423.2*</td>
</tr>
<tr>
<td>SMP135331B*</td>
<td>SMP1991950*</td>
<td>mk4.002807.03.03 mk4.006698.00.01 mk4.000130.00.01 SMDC175.4* SMDC926.2*</td>
</tr>
<tr>
<td>SMP135343S*</td>
<td>SMP1673870*</td>
<td>mk4.003471.02.01 mk4.002150.00.01 mk4.010834.00.01 SMDC8510.2* SMDC674.2*</td>
</tr>
<tr>
<td>SMP1353100<em>S</em></td>
<td>SMP145240</td>
<td>mk4.001291.01.01 mk4.007776.00.01 mk4.000020.08.01 SMDC12554.3* SMDC1917.2*</td>
</tr>
<tr>
<td>SMP135363*</td>
<td>SMP172450</td>
<td>mk4.005599.03.01 mk4.000341.05.01 mk4.044245.00.01 SMDC9261.1* SMDC3589.1A*</td>
</tr>
<tr>
<td>SMP13531700</td>
<td>SMP157050</td>
<td>mk4.001235.01.01 mk4.028014.00.01 mk4.011988.00.01 SMDC14380.2* SMDC5981.3B*</td>
</tr>
<tr>
<td>SMP13531194</td>
<td>SMP153200</td>
<td>mk4.010211.00.01 mk4.010306.00.01 mk4.022607.00.01 SMDC8510.3A* SMDC1749.3A*</td>
</tr>
<tr>
<td>SMP135314880</td>
<td>SMP169420* SMP157050</td>
<td>mk4.001235.04.01 mk4.002514.01.01 mk4.000557.00.01 SMDC8510.3B* SMDC1749.3B*</td>
</tr>
<tr>
<td>SMP13534080</td>
<td>SMP084280* SMP157050</td>
<td>mk4.018209.00.01 mk4.013402.00.01 mk4.009948.03.01 SMDC15385.3* SMDC27911.2*</td>
</tr>
<tr>
<td>SMP13531701</td>
<td>SMP158390* SMP157050</td>
<td>mk4.003634.00.01 mk4.030476.00.01 mk4.000906.02.01 SMDC15273.1* SMDC907.1A*</td>
</tr>
<tr>
<td>SMP13531880</td>
<td>SMP008850</td>
<td>mk4.003634.00.01 mk4.000617.00.01 mk4.012831.03.01 SMDC4350.2* SMDC907.1B*</td>
</tr>
<tr>
<td>SMP135312340</td>
<td>SMP158390* SMP157050</td>
<td>mk4.013004.00.01 mk4.011509.03.01 mk4.020797.01.01 SMDC19909.1* SMDC1825.1*</td>
</tr>
<tr>
<td>SMP13531880*</td>
<td>SMP158390* SMP157050</td>
<td>mk4.018682.00.01 mk4.012590.01.01 mk4.000210.07.01 SMDC14736.1*</td>
</tr>
</tbody>
</table>

Table 2.8 *Rhodopsin* SVM classifier results.
<table>
<thead>
<tr>
<th><strong>S. mansoni</strong></th>
<th><strong>S. mediterranea</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP126730</td>
<td>4</td>
</tr>
<tr>
<td>SMP127310</td>
<td>3*</td>
</tr>
<tr>
<td>SMP134460</td>
<td>4</td>
</tr>
<tr>
<td>SMP134820</td>
<td>5</td>
</tr>
<tr>
<td>SMP027940</td>
<td>5</td>
</tr>
<tr>
<td>SMP137300</td>
<td>5</td>
</tr>
<tr>
<td>SMP043260</td>
<td>4*</td>
</tr>
<tr>
<td>SMP043270</td>
<td>3</td>
</tr>
<tr>
<td>SMP043290</td>
<td>5</td>
</tr>
<tr>
<td>SMP043300</td>
<td>6</td>
</tr>
<tr>
<td>SMP043320</td>
<td>5</td>
</tr>
<tr>
<td>SMP043340</td>
<td>4*</td>
</tr>
<tr>
<td>SMP145520</td>
<td>5</td>
</tr>
<tr>
<td>SMP145540</td>
<td>3</td>
</tr>
<tr>
<td>SMP148210</td>
<td>5</td>
</tr>
<tr>
<td>SMP150180</td>
<td>3</td>
</tr>
<tr>
<td>SMP120620</td>
<td>6</td>
</tr>
<tr>
<td>SMP180140</td>
<td>5</td>
</tr>
<tr>
<td>SMP160020</td>
<td>5</td>
</tr>
<tr>
<td>SMPSC103</td>
<td>2</td>
</tr>
<tr>
<td>SMP159860</td>
<td>3</td>
</tr>
</tbody>
</table>

1: Muscarinic acetylcholine
2: Adrenoceptors
3: Dopamine
4: Histamine
5: Serotonin
6: Octopamine
7: Trace amine

Table 2.9 Amine SVM classifier results.
Abstract

G protein-coupled receptors (GPCRs) represent the largest known superfamily of membrane proteins extending throughout the Metazoa. There exists ample motivation to elucidate the functional properties of GPCRs in the phylum Platyhelminthes, given the heavy health burden exacted by pathogenic flatworms and the role of free-living flatworms as model organisms for the study of developmental biology and their parasite counterparts. Efforts on this front have been hampered by the unreliable nature of heterologous receptor expression platforms. We validate and describe a loss-of-function approach for ascertaining the ligand and G protein coupling properties of GPCRs in their native cell membrane environment. RNA interference (RNAi) was used in conjunction with a GPCR biochemical endpoint assay to monitor cAMP modulation in response to the translational suppression of individual receptors. This was used to confirm GYIRFamide as the cognate ligand for Dugesia tigrina GtNPR-1, while revealing its endogenous coupling to Gαi/o. The method was then applied to deorphanize a Schmidtea mediterranea 5-HT receptor, Smed-SER-7. A bioinformatics protocol guided the selection of receptor candidates mediating 5-HT-evoked responses. While these results establish the poten-
tial of this approach, future work can help optimize and adapt this receptor deorphanization strategy to a higher-throughput platform.

3.1 Introduction

G protein-coupled receptors (GPCRs) have been the subject of intense research scrutiny due to their central role in eukaryotic signal transduction and their exploitability as drug targets [79, 80, 149]. The phylum Platyhelminthes houses prominent human pathogens as well as tractable model organisms. Flatworm GPCRs represent lucrative anthelmintic targets, as evidenced by the biological activities of their putative ligands [164, 169] and the crucial biological functions of these receptors in nearly-related organisms [184, 185]. Signaling pathways associated with the GPCR superfamily have been specifically identified as potential targets for life-cycle interruption of flatworm parasites [15, 186]. The identification and pharmacological characterization of GPCRs is likely to generate lucrative drug discovery leads, while enhancing our basic understanding of receptor biology in this important phylum. This process has been slowed by an exclusive reliance on traditional heterologous receptor expression approaches.

Once identified, GPCRs undergo deorphanization, the process of pairing orphan receptors with their cognate ligands. The predominant approaches all require the transient or stable heterologous expression of GPCRs in a surrogate cell system and in most cases, this expression occurs in cells derived from other species and phyla [148, 195, 196]. The complex regulatory processes that guide the correct folding and export of receptors to the cell membrane [89–92] are not necessarily well-conserved across cell lineages. Further, the structural and functional integrity of receptors can depend on the local membrane lipid environment, the composition of which can differ between native and heterologous cells [197, 198]. The exact post-translational requirements for proper receptor expression and function vary greatly among receptors, making the task of identifying a suitable heterologous system a receptor-specific process of trial-and-error [148].
Recalcitrance of individual flatworm GPCRs to heterologous expression platforms has introduced a significant bottleneck in implementing functional assays to identify receptor agonists. Similar issues have impeded the structural elucidation of mammalian receptors [147]. Only a handful of flatworm GPCRs have thus far been deorphanized, with receptors expressed in such divergent cellular environments as CHO [138], HEK293 [139,140], COS7 [139], yeast [140,141], and *Xenopus* oocyte cells [142]. We describe a relatively simple loss-of-function deorphanization approach that could be applied in a native cell or membrane environment. This alternative strategy could help catalyze a first-pass mapping of receptors and ligands in this phylum.

**Inversing the Paradigm: RNAi as a Deorphanization Tool**

We describe and validate an RNA interference (RNAi)-based method that allows receptors to undergo deorphanization without the need for full-length cloning or transport to a heterologous expression system. In principle, a collection of putative GPCR ligands are screened against membrane preparations to evaluate their effects on second-messengers downstream of GPCR activation. RNAi is then used to assay whether observed responses can be altered or abolished by the knockdown of individual receptors from the membrane preparations. A successful “hit” confirms expression of a given receptor, functionally pairs the receptor with a given ligand, and couples the receptor with a specific G protein signaling pathway.

Bioinformatics approaches can be used to help identify receptors as putative targets for a particular ligand, or conversely, to narrow the list of potential ligands for a given receptor. The recent availability of platyhelminth genomic data [24, 25, 27] has led to the accumulation of a wealth of receptor and ligand data. A comprehensive *in silico* protocol revealed over 116 *Schistosoma mansoni* and 333 *Schmidtea mediterranea* GPCRs, which were classified using phylogenetic, homology-based, and machine-learning approaches [199]. Bioinformatics and proteomics-based studies have also led to the expansion of the known set of putative GPCR ligands [169,171,172].

The primary biochemical endpoints of GPCR activation are typically assayed by recording
agonist-evoked changes in cAMP (Gαs and Gαi/o) or Ca²⁺ (Gαq) levels. A variety of established labeling and detection schemes (e.g. fluorescent, luminescent, and radioisotope) are available for these second messengers [200]. In this study, we focus our efforts on the Gαs and Gαi/o pathways and employ a radioimmunoassay (RIA) for cAMP detection. Monitoring adenylyl cyclase modulation of cAMP allows us to simultaneously examine two of the three major GPCR activation endpoints.

While this loss-of-function approach limits pharmacological analysis, it is adaptable to higher-throughput platforms and can serve as an efficient first-pass ligand-receptor mapping tool. It should be noted that ligands and receptors can display pharmacological promiscuity. Ligands can act through more than one receptor and receptors can respond to more than one ligand, with a range of affinities. Further, receptors responsive to a given ligand do not necessarily share the same G protein coupling profile and are likely to be expressed in different abundances. However, this approach only concerns itself with the contribution of individual receptors to differences between control and RNAi response profiles. The scale and directionality of the differences between these profiles provide information relevant to ligand responsivity and G protein coupling, respectively. The basic logic of this deorphanization strategy is outlined in Figure 3.1.

3.2 Results and Discussion

cAMP Assay Optimization and Ligand Screen

A cell membrane preparation protocol was adapted [201] and optimized for planaria, and used to generate samples for treatment with various GPCR ligands. The downstream effects of ligand incubation on cAMP levels were monitored using a cAMP RIA. A ligand screen was first carried out on Dugesia tigrina membrane preparations with a small number of peptides and biogenic amines. These ligand classes are prominent in platyhelminth biology [164, 169, 171, 172], and there is a strong likelihood that a subset signal through one or more receptors coupled to either the Gαs or Gαi/o pathways. This would presumably be made apparent by stimulation
of basal [cAMP] or inhibition of forskolin (Fk)-stimulated [cAMP] as measured by RIA, respectively.

Included in this initial screen were the only two ligands definitively coupled to planarian GPCRs: GYIRFamide and the biogenic amine serotonin (5-HT; 5-hydroxytryptamine). It was a reasonable assumption that both GYIRFamide and 5-HT would modulate cAMP levels in a whole organism membrane preparation. We previously deorphanized the D. tigrina receptor GtNPR-1, showing it responded potently to the neuropeptide GYIRFamide in mammalian cell culture [138]. Chimeric G proteins (Gαq5 and Gαo5) were used to divert downstream GtNPR-1 signaling through the Gαq pathway, suggesting this receptor is Gαi/o-coupled in its native environment. More recently, a Dugesia japonica 5-HT GPCR has been deorphanized using Xenopus laevis oocytes [142], and there is long-established evidence of 5-HT stimulation of cAMP in both S. mansoni [203,204] and other planarian species [205], suggesting that 5HT acts through one or more Gαs-coupled GPCRs.

Alongside GYIRFamide and 5-HT, we included neuropeptide F (NPF) and octopamine as putative ligands. NPF has been shown to inhibit Fk-stimulated cAMP production in membranes isolated from S. mansoni [201]. Given the identification of planarian NPF homologues [171,172], we hypothesized that this peptide would have a similar inhibitory effect on cAMP levels. The results of this primary screen show that 10⁻⁵M 5-HT drastically stimulates cAMP production, while 10⁻⁴M GYIRFamide, 10⁻⁴M NPF, and 10⁻⁴M octopamine inhibit Fk-stimulated cAMP accumulation in Dugesia membrane preparations (Figure 3.2) to varying degrees. These changes in [cAMP] can be viewed as the additive response profile of each ligand.

We chose to pursue the response profiles of GYIRFamide and 5-HT. Provided that GtNPR-1 is a known target of GYIRFamide in D. tigrina, we first examined whether or not this would be apparent using this novel loss-of-function approach. Given that the inhibition of adenylate cyclase by GYIRFamide is less potent than that brought on by NPF, this also serves as validation
of assay sensitivity. In addition to this proof or principle, we chose to pursue deorphanization of an orphan S. mediterranea 5-HT receptor, aided by the availability of receptor sequence data for this species.

**Coupling Second-Messenger Assay with RNAi: GtNPR-1 Proof of Principle**

**Establishing RNAi-mediated Receptor Suppression**

Double-stranded (ds) RNA was introduced to isolated D. tigrina colonies using a bacterial-mediated feeding protocol. Planaria were randomly selected, isolated into treatment groups, and fed either non-flatworm control dsRNA or GtNPR-1 dsRNA. A 10 day RNAi feeding cycle consisted of four evenly-spaced feedings, followed by a four-day starvation period. Semi-quantitative RT-PCR was used to confirm gene knockdown. A small number of planarians were randomly selected from both experimental and control groups to assay GtNPR-1 suppression, and the remaining planarians were used for membrane assays. Significant GtNPR-1 knockdown (> 80%) is consistent and apparent in the experimental group, while GtNPR-1 expression remains robust in the control group (Figure 3.3) relative to endogenous standard.

**Deorphanization via Comparison of Response Profiles**

Membranes were prepared from both control and GtNPR-1 dsRNA-fed planarians, and treated with Fk (10⁻⁴M), GYIRFamide (10⁻⁴M), and Fk (10⁻⁴M) + GYIRFamide (10⁻⁴M). RIA was used to assay cAMP levels corresponding to these treatments. Comparison of the response profiles of control and RNAi planaria reveals near-complete abolishment of GYIRFamide-evoked inhibition of Fk-stimulated cAMP in the GtNPR-1 knockdown group (Figure 3.4, Table 3.1). Overall, GYIRFamide reduces Fk-stimulated cAMP production by an average of ~ 30% in the control group, and this inhibition was completely abolished by the suppression of GtNPR-1 expression in the RNAi group. In fact, combined Fk and peptide treatment led to a slight increase (~ 2%) in cAMP levels compared to Fk alone. These results confirm that GtNPR-1 is agonized by GYIRFamide and further establish that this receptor is natively coupled to the Gαᵢ/o signaling pathway.
**In silico Target Selection**

The two ligands that most drastically stimulated and inhibited adenlyate cyclase activity in our primary ligand screen were 5-HT and NPF, respectively. These results were shown to extend to *S. mediterranea* (data not shown). To identify and rank 5-HT receptor candidates, a profile HMM was built with sequences procured from GPCRDB [107]. Training was focused on 62 full-length invertebrate 5-HT and 5-HT-like receptors. This was used to search against *S. mediterranea* GPCR sequence datasets [199], and the results were ranked by E-value. The top 10 receptor candidates were used as BLASTp [124] queries against the NCBI “nr” database. This was used to identify receptors displaying 5-HT receptor homology, and to filter against receptors that displayed a varied range of biogenic amine GPCR-related homology.

Receptors that survived the BLAST filter were compared to their nearest-related *S. mansoni* homologs, each returning either Smp_148210 and Smp_126730 (Table 3.2). While there is a great deal of bioinformatic evidence to suggest multiple receptor targets for 5-HT, we narrowed our list to two best-match receptors and selected mk4.001585.00.01 as the first target for RNAi-based deorphanization. This receptor is tentatively labeled Smed-SER85.

**RNAi-based Deorphanization of Planarian 5HT Receptor**

*Smed-SER85* transcript expression was confirmed via PCR, and knockdown was illicited following the protocol described for GtNPR-1. Similarly, a membrane preparation protocol was optimized for *S. mediterranea* and membranes were treated with $10^{-4}$ M Fk, $10^{-4}$ M 5-HT, $10^{-4}$ 8-hydroxy-2-(di-n-propylamino) tetralin (8-OH-DPAT), and $10^{-4}$ meta-Chlorophenylpipеразин (mCPP) (Figure 3.5). 8-OH-DPAT is an agonist of vertebrate 5-HT$_{1A}$ and 5-HT$_7$ receptors, while mCPP most potently stimulates 5-HT$_{2B}$ and 5-HT$_{2C}$. This initial profile establishes that these compounds can modulate cAMP levels in this phylum as well, presumably through the action of 5-HT receptors.
**Smed-SER85 Mediates Flatworm Motility**

The phenotypic effects of *Smed-SER85* knockdown were assayed using an automated video tracking system. Planarian motility was recorded over a 15 minute period after acclimation of individual worms to the recording well environment. Figure 3.6 shows \( \sim 52\% \) decrease in distance travelled over the recording duration in a comparison of control and *Smed-SER85* RNAi worms. This significant decrease in basal motility levels suggests that Smed-SER85 plays an important role in the maintenance of flatworm locomotory behavior. Previous studies have established the role of serotoninergic signaling in the regulation of flatworm locomotion [206,207], and this motility assay establishes Smed-SER85 as the only known receptor mediating such fundamental flatworm behaviors.

### 3.3 Conclusions

This study shows the utility of joining RNAi with biochemical endpoint assays to as a means of deorphanizing platyhelminth GPCRs in their native membrane environment. The approach was first validated using the only deorphanized flatworm neuropeptide GPCR (GtNPR-1), confirming agonism by GYIRFamide and providing new information about its endogenous G protein coupling profile. The orphan *S. mediterranea* GPCR Smed-SER85 was shown to respond to 5-HT, revealing its endogenous G protein pathway and illustrating the usefulness of applying an in silico strategy to candidate receptor selection. The translational suppression of Smed-SER85 led to a major decrease in overall flatworm motility. While 5-HT and 5-HT-related pharmacological agents had been shown to be involved in regulation of locomotory behavior, this represents the identification of the first receptor that mediates these observations. While this loss-of-function strategy side-steps some of the concerns and difficulties associated with heterologous GPCR expression, there is significant room for improving both the sensitivity and scalability of this assay.

The heavy tissue requirements of the membrane preparation protocols employed introduce a potential rate-limiting step. Further optimizations of membrane or whole cell preparation
protocols in this phylum could allow for more efficient and robust pharmacological analysis. This assay could also conceivably be adapted to higher-throughput platforms, and extended to include GPCRs that signal through the Goq pathway. Conveniently, establishing receptor-specific RNAi in planaria allows for the accumulation of loss-of-function phenotypic data in parallel to pharmacological data. In this regard, the study of planarians can inform flatworm parasite biology. Biasing the receptor and ligand pool to those best conserved between parasitic and free-living flatworms could lead to new targets for chemotherapeutic intervention.

3.4 Materials and Methods

Planarian maintenance protocol

*Dugesia tigrina* (Ward’s Natural Science, Rochester, NY) and *Schmidtea mediterranea* colonies were maintained in the laboratory with a regular feeding cycle. *Dugesia tigrina* colonies were fed three times a week, while *Schmidtea mediterranea* colonies were fed twice a week. Planarians were randomly selected and isolated in ~ 50-worm groupings for RNAi feeding cycles and cAMP assays.

RNA interference

Primers were designed with Primer3 [192] to selectively amplify 300-500 bp fragments of *GtNPR-1* and 5HT receptor candidate *Smed-SER85*. BLAT [163] was used to check for potential off-target silencing. A 465 bp fragment of *GtNPR-1* was amplified from a full length clone of *Gt-NPR1* housed in pcDNA3.1(+) with the primers 5’-TTGGATCTTTCCAGCGACTCT-3’ (forward) and 5’-ATGGTTCGTTCGACGTTTTC-3’ (reverse). A 388 bp fragment of *Smed-SER85* was amplified from *Schmidtea* cDNA isolated using RNAqueous (Ambion) and Retroscript (Ambion) with the primers 5’-CTCCGCTTTTAATTGGAGGA-3’ (forward) and 5’-CTGTTTCTTTTCCGGGGAT-3’ (reverse). An RNAi control sequence was amplified from *Aedis aegypti* cDNA with primers 5’-CTCCGCTTTTAATTGGAGGA-3’ (forward) and 5’-CTGTTTCTTTTCCGGGGAT-3’ (reverse), corresponding to a 413 bp fragment of a putative odorant receptor (VectorBase [208] id: AAEL013422). Second-round PCR was performed
for each target sequence using the original gene-specific primers flanked by Gateway Cloning system (Invitrogen) recombination sites: 5’-GGGG-attB1-3’ (forward) and 5’-GGGG-attB2-3’ (reverse). Entry sequences were subcloned into the pPR244 (pDONRdT7) [77] (Appendix D) destination vector with corresponding attP1 and attP2 recombination sites using the BP Clonase II (Invitrogen) enzyme mix. Clones were transformed into TOP10 Electrocompe
tent E. coli (Invitrogen) and sequence confirmed.

**Bacterial-mediated dsRNA feeding**

Propagation of RNAi vectors was carried out using competent RNase III-deficient HT115 (DE3) bacterial cells as previously described [77]. Colonies were scaled in liquid culture (500 ml) until an OD of 0.3-0.4 was reached, and T7 polymerase activity was IPTG-induced. After 2 hours, cells were pelleted by centrifugation and resuspended in 50 ml of media. This step was repeated with a resuspension volume of 10 ml. Glycerol stocks of dsRNA-containing bacteria (665 ul total volume; 20% glycerol) were mixed with 275 ul of blended organic beef liver and 100 ul red blood cells (RBCs). RBCs allowed for visual monitoring of planarian food intake. Planarians were dark-fed dsRNA-containing bacteria four times over a 10-day timeline with each feeding separated by a two-day period, and starved for at least four days prior to their use in experiments.

**Semi-quantitative RT-PCR**

Total RNA was extracted from individual planaria with RNAqueous (Ambion), followed by removal of DNA contaminants with TURBO DNase (Ambion). First strand cDNA synthesis was carried out with the RETROscript kit (Ambion), as part of a two-stage RT-PCR. PCR optimization was carried out with the QuantumRNA 18S Internal Standards kit (Ambion) per manufacturer instructions. 18S ribosomal RNA was used as an endogenous standard for normalizing measures of gene expression and reducing sample-to-sample variation. cDNA samples were used in parallel as templates for multiplex PCR with gene-specific and 18S rRNA primer pairs. PCR reaction products were visualized on 1.2% electrophoretic gel with the Kodak Gel
Logic 112 imaging system, and amplicon intensities were analyzed with standard software to derive relative transcript abundances.

**Membrane preparation and cAMP RIA**

Planaria were washed twice with cold cAMP buffer containing 50 mM sucrose, 50 mM glycyl-glycine, 10 mM creatine phosphate, 2 mM MgCl2, 0.5 mM isobutylmethylxanthine (IBMX), 1 mM dithiothreitol (DTT), 0.02 mM EGTA, 10 units/ml creatine kinase, and 0.01% bovine serum albumin. Worms were kept on ice for 5 min and then homogenized on ice for 2 min with a Teflon homogenizer. This preparation was centrifuged at 1,000 X g for 5 min, with the pellet that included cell debris discarded. This centrifugation step was then repeated. The supernatant was centrifuged at 40,000 X g for 30 min at 4°C. The supernatant was discarded, and the membrane-containing pellet was resuspended via sonication in cAMP buffer supplemented with 0.1 mM ATP and 0.1 mM GTP. Total suspension volume was set at 500 µl/sample, such that each sample would contain cell membranes from ~3 worms. 500 µl aliquots of this membrane preparation correspond to individual reactions in the cAMP assay.

Samples were incubated with various concentrations (and combinations) of forskolin and/or putative ligands (peptide or biogenic amine) at 37°C for 20 min to stimulate cAMP production. Forskolin and peptide ligands were dissolved in Me2SO, with final reaction mixtures containing <0.1% Me2SO. Me2SO has no measurable effect on cAMP in this range. Samples were centrifuged at 3,000 X g for 5 minutes after ligand incubation, and 400 µl of supernatant from each sample was transferred into a fresh tube. Three reactions was assayed for each treatment condition, and each reaction was sampled in triplicate. cAMP concentration was determined with a radioimmunoassay as described previously [209], with a lower detection bound of 1.6 fmol per tube.
Bioinformatics

A Profile HMM were built using HMMER-2.3.2 [126] using training sequence data from GPCRDB [107]. Available invertebrate 5-HT full-length receptor sequences were aligned and a profile HMM was constructed for each using hmmbuild. The model underwent calibration using hmm-calibrate, with the default parameters. *S. mediterranea* GPCR sequence datasets [199] were searched against this subfamily-specific profile HMM using hmmpfam. The resulting hits for were parsed with a Perl script and ranked according to E-value. The top 10 receptors candidates were used as BLASTp [124] queries against the NBCI “nr” database, and surviving receptors were similarly searched against the *S. mansoni* predicted proteome [24].

Phenotypic Assays

Planarian motility was tracked with the Ethovision 3.1 video tracking system (Noldus). Worms were individually tracked in wells, and allowed to settle for 10 minutes proceeded by a 15 minute recording period. Worms were recorded in groups of six in a six-well plate platform. Recorded movement tracks were examined to confirm that the tracking software properly distinguished the organism from background.

Statistical Analysis

Basal cAMP levels were set as a baseline for individual RIA experiments, and cAMP values were normalized with respect to the level of Fk-stimulated cAMP (set at 100%). This allowed us to join datasets from repeated experiments with differing basal cAMP levels due to variance in the quality and yield of individual membrane preparations. One-way analysis of variance (ANOVA) was used with Tukey’s post hoc test for multiple comparison analysis of cAMP levels associated with different treatments. A two-tailed T-test was used for motility comparison between control and RNAi groups. Significances are reported at $P < 0.05$, $P < 0.01$, and $P < 0.001$. 
3.5 Acknowledgements

We would like to thank Dr. Peter Reddien and Dr. Alejandro Sanchez Alvarado for providing the pPR244 (pDONRdT7) RNAi vector. This work was funded by a grant from the National Institutes of Health (NIH R01 AI49162) to TAD and AGM and a USDA Formula Fund Grant (IOWV-DAY-411-23-08) to TAD.

3.6 Figures and Tables

<table>
<thead>
<tr>
<th>Profile</th>
<th>Blank</th>
<th>Fk</th>
<th>L</th>
<th>Fk+L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>RNA-i (x)</td>
<td>A</td>
<td>B</td>
<td>C'</td>
<td>D'</td>
</tr>
</tbody>
</table>

I. If $C > A$, $D \geq B$: $L$ has aggregate stimulatory effect on [cAMP]

If $C' < C$, $D' \leq D$: GPCR "x" is L-responsive and G-s coupled

II. If $C \leq A$, $D < B$: $L$ has aggregate inhibitory effect on [cAMP]

If $C' \geq C$, $D' > D$: GPCR "x" is L-responsive and G-i coupled

III. If $C \approx A$, $D \approx B$: $L$ has no net effect on [cAMP]

Figure 3.1 RNAi-based deorphanization approach overview.

The general set of experimental outcomes for an RNAi-based deorphanization experiment focused on the $G_s$ and $G_i$ pathway are shown. Letters placed within wells represent assayed cAMP levels for membrane preparations in response to various treatments. Potential results are described with respect to the notion that a given ligand can act on multiple GPCRs that are not necessarily coupled to the same G protein ($G_s$ or $G_i$). Abbreviations: Fk, forskolin; L, ligand; RNAi (x), RNAi preparation with GPCR “x” knocked down.
Figure 3.2  cAMP ligand screen

Peptide and biogenic amine ligand screen performed against isolated D. tigrina membranes. RIA cAMP outputs are shown as mean ± SEM, and asterisks represent statistically significant differences compared with either control or treatment with Fk alone; * P < 0.05, *** P < 0.001, one-way ANOVA, Tukey post hoc test. Red bars are compared with Fk treatment: octopamine (OCT), GYRIFamide (GYIRF), and neuropeptide F (NPF) all inhibit Fk-stimulated cAMP at 100 uM. The green bar is compared with the control condition: serotonin (5-HT) stimulates basal cAMP. These changes in cAMP are likely GPCR-mediated, and should therefore be altered in a ligand-specific manner by subtraction of particular receptor targets from cell membranes via RNAi.
Figure 3.3  Semi-quantitative PCR reveals *GtNPR-1* knockdown.

Lane 1 is a 100 bp DNA ladder, lanes 2-5 represent individual *GtNPR-1* dsRNA-fed planarians, and lanes 6-9 represent control dsRNA-fed planarians. The bottom band (∼300 bp) is the 18S internal standard, and the top band (∼400 bp) shows *GtNPR-1* expression. The top band disappears in the experimental group, confirming near abolishment of receptor expression in these worms. Relative band intensities (*GtNPR-1*/18S rRNA) for *GtNPR-1* RNAi group: 0.44 ± 0.15. Relative band intensities for control group (band location manually selected): 0.08 ± 0.02. This corresponds to > 80% knockdown of *GtNPR-1* transcript.
Figure 3.4  RNAi-based GtNPR-1 deorphanization

Treatment groups are Control (control dsRNA) and GtNPR-1 RNAi (GtNPR-1 dsRNA). Treatments are C (control), Fk (10^{-4} M forskolin), and Fk + GYIRF (10^{-4} M forskolin and 10^{-4} M GYIRFamide). Each bar is the mean (± SEM) of 3 separate experiments. Basal cAMP levels were set as a baseline for each individual experiment, and cAMP values were normalized with respect to the level of Fk-stimulated cAMP (set at 100%). This allowed us to join datasets with differing basal cAMP levels, due to variance in the quality and yield of individual membrane preparations. Analysis of the raw cAMP values of individual experiments renders the same results. Asterisks indicate significance at P < 0.001 (**), and “ns” indicates no significant difference (one-way ANOVA, Tukey post hoc test).
Figure 3.5 Pharmacological response profile of 5-HT receptor agonists.

Treatments applied at $10^{-4}$ M: Fk, 5-HT, 8-OH-DPAT, and meta-Chlorophenylpiperazine (mCPP). This control profile shows that 5-HT and 8-OH-DPAT lead to an overall increase in cAMP, while mCPP has an overall inhibitory effect on cAMP levels. Treatment groups are compared to control group. Asterisks indicate significance at $P < 0.001$ (**) and $P < 0.01$ (one-way ANOVA, Tukey post hoc test).
Total locomotory distance calculated over a 15 minute recording period (n=20). Comparison of control and Smed-SER85 RNAi planarians reveals a > 50% decrease in overall motility (two-tailed t-test, P < 0.001), suggesting this receptor plays a fundamental role in maintenance of planarian locomotory behavior.
Table 3.1 RNAi-based *GtNPR*-*1* deorphanization cAMP raw values

<table>
<thead>
<tr>
<th>EXP</th>
<th>Treatment</th>
<th>Control</th>
<th><em>GtNPR</em>1 RNAi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>62.05 ± 2.46</td>
<td>60.83 ± 1.91</td>
</tr>
<tr>
<td></td>
<td>Fk</td>
<td>102.45 ± 4.06</td>
<td>101.47 ± 1.59</td>
</tr>
<tr>
<td></td>
<td>Fk + G</td>
<td>85.02 ± 1.59 ***</td>
<td>103.03 ± 4.27 ns</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>27.88 ± 0.97</td>
<td>33.54 ± 1.27</td>
</tr>
<tr>
<td></td>
<td>Fk</td>
<td>57.37 ± 2.68</td>
<td>58.78 ± 1.64</td>
</tr>
<tr>
<td></td>
<td>Fk + G</td>
<td>48.67 ± 1.23 **</td>
<td>57.89 ± 0.93 ns</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>81.49 ± 4.06</td>
<td>55.16 ± 1.60</td>
</tr>
<tr>
<td></td>
<td>Fk</td>
<td>215.96 ± 10.99</td>
<td>129.79 ± 3.61</td>
</tr>
<tr>
<td></td>
<td>Fk + G</td>
<td>195.63 ± 6.17 **</td>
<td>132.60 ± 4.62 **</td>
</tr>
</tbody>
</table>

RIA-determined cAMP values (pM) are provided for 3 separate experiments (mean ± SEM). Treatments: C (control), Fk (Forskolin), Fk + G (Forskolin + GYIRFamide). The amount of isolated membrane differs between experiments, as evidenced by basal cAMP levels. This is in part due to differences in the size, number, and feeding behavior of worm batches used for membrane isolation. Analysis (one-way ANOVA, Tukey) of these raw datasets establishes abolishment of cAMP inhibition brought on by GYIRFamide via *GtNPR*-1 suppression. For each experimental grouping, Fk is compared to Fk + G. Asterisks indicate significance at P < 0.001 (**), P < 0.01 (**), and “ns” means no significant difference.
<table>
<thead>
<tr>
<th><strong>S. mediterranea</strong></th>
<th>5-HT HMM</th>
<th>BLAST</th>
<th><strong>S. mansoni</strong></th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>mk4.013690.00.01</td>
<td>2.90E-108</td>
<td>*</td>
<td>Smp_148210</td>
<td>58%</td>
</tr>
<tr>
<td><strong>mk4.005939.01.01</strong></td>
<td>3.90E-85</td>
<td>*</td>
<td>Smp_126730</td>
<td>47%</td>
</tr>
<tr>
<td>mk4.011371.00.01</td>
<td>1.20E-71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mk4.001585.00.01</strong></td>
<td>3.00E-70</td>
<td>*</td>
<td>Smp_148210</td>
<td></td>
</tr>
<tr>
<td>mk4.007388.02.01</td>
<td>9.60E-69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mk4.029325.00.01</td>
<td>2.30E-65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mk4.000656.10.01</td>
<td>1.10E-61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mk4.004462.02.01</td>
<td>5.10E-51</td>
<td>*</td>
<td>Smp_148210</td>
<td></td>
</tr>
<tr>
<td>mk4.011006.00.01</td>
<td>3.30E-50</td>
<td>*</td>
<td>Smp_126730</td>
<td></td>
</tr>
<tr>
<td>mk4.003202.01.01</td>
<td>1.10E-49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2  5-HT candidate receptor selection

Top-ranked 5-HT profile HMM hits are shown with their associated E-values. Receptors that survived the BLAST filter are marked with an asterisk (*). The nearest related *S. mansoni* homolog is shown and % identity (PID) values are printed for the best match to each parasite receptor. Of these two receptors (shown in bold), mk4.001585.00.01 (Smed-SER85) was selected for RNAi-based deorphanization.
4 PROF1 Localization in Planaria

The PROF1 receptor family, as a flatworm-specific subset of the notoriously drugable GPCR superfamily, represents a rational target for anthelmintic drug discovery. As an initial step towards deciphering the biological functions of these receptors and assaying their potential exploitability as drug targets, whole mount in situ hybridization (WISH) was used to localize transcript expression for two planarian PROF1 receptors. The overarching motivation is that the different PROF1 receptors and receptor subtypes are likely to display unique mRNA expression patterns that will inform us about their cognate ligands and their potential signal transduction roles with respect to flatworm anatomy.

4.1 Planarian In situ hybridization protocol

PROF1 receptor fragments were PCR amplified with the minimal T7 polymerase promoter sequence appended to the 5’ anti-sense primer. Primers for SMDC2955.1: 5’-cgtgcctacctgattccatt-3’ (forward) and 5’-TAATACGACTCACTATAGGGTACTtttcctcgttgggagatttg-3’ (reverse). Digoxygenin (DIG)-labeled antisense riboprobes were synthesized using these PCR products (Roche). Whole-mount in situ hybridization (WISH) was performed at 55°C in hybridization solution (50% formamide, 5XSSC, 100 ug/ml yeast tRNA, 100 ug/ml heparin sodium salt, 0.1% Tween-20, 10 mM DTT, 10% dextran sulfate sodium salt). DIG-labeled riboprobe (40 ng/ml) was denatured at 72°C for 15 min immediately prior to hybridization. BCIP/NBT was used for chromogenic color development, followed by paraformaldehyde fixation and imaging.
4.2 PROF1 Localization Results

WISH staining revealed that the expression pattern for SMDC2955.1 (PROF1 group II) is localized to the planarian nervous system. From this we can conclude that PROF1 receptors are endoGPCRs expressed in the planarian nervous system, as opposed to chemosensory-type GPCRs. Further, their cognate ligands likely belong to either the neuropeptide or biogenic amine receptor families. The former is of greater likelihood, given the relative conservation of aminergic GPCRs in the metazoa in comparison to peptidergic GPCRs, and the highly-diverged nature of this receptor clade. More concrete evidence for the hypothesis that PROF1 receptors respond to neuropeptide ligands comes in the form of the transcript distributions of very recently identified flatworm-specific peptides [172]. It is also entirely possible that PROF1 ligand(s) have yet to be uncovered. This preliminary result is encouraging, given the likelihood of functional conservation with respect to schistosome biology, as the parasite nervous system and neuromusculature is widely recognized as fertile ground for drug targeting [?, 210].

![Figure 4.1 Localization of PROF1 transcripts in S. mediterranea](image)

(L) The localization of a single PROF-1 subtype II transcript (labeled SMDC2955.1) is shown along the longitudinal nerve chords (red) and the cerebral ganglion (green). (R) The planarian head region reveals PROF-1 expression in the cerebral ganglion (green).
CONCLUSIONS

This dissertation describes the successful application of a sophisticated bioinformatics protocol to comprehensively identify G protein-coupled receptors in two important flatworm species, the human parasite \textit{Schistosoma mansoni} and the model planarian \textit{Schmidtea mediterranea}. Transmembrane-focused hidden Markov models were used in combination with a set of filters to mine the genomic assemblies of these organisms for GPCRs. Subsequent rounds of manual gene curation and homology-based searches against the nucleotide assemblies further expanded the total receptor count and improved the quality of the underlying gene models. The final GPCR dataset houses 116 \textit{S. mansoni} and 333 \textit{S. mediterranea} GPCRs. Phylogenetic analysis confirmed the presence of the primary metazoan ‘GRAFS’ families, and revealed large numbers of lineage-specific receptors. Among these, the flatworm-specific PROF1 receptors and the planarian-specific PARF1 receptors represent the largest and most distinct groupings.

Transmembrane-focused SVMs were trained and used to sub-classify a subset of full-length \textit{Rhodopsin} and aminergic receptors. Together, these phylogenetic, homology, and machine learning-based outputs can guide future efforts to identify the cognate ligands of these GPCRs. To learn more about the potential roles of PROF1 receptors, \textit{in situ} hybridization was performed in \textit{S. mediterranea}. The transcript distribution of a representative of PROF1 subtype II was localized to the nervous system. This and other lines of evidence suggest that these receptors are endo-GPCRs that likely respond to flatworm-specific peptide ligands. Analysis of \textit{Glutamate} GPCRs revealed receptors with non-canonical ligand binding domains that are likely to exhibit atypical pharmacology or respond to other amino-acid derived ligands. To support the adoption of planarians as flatworm parasite models, parasite-planarian sequence
pairs were ranked by shared sequence identity.

In light of the unpredictable nature of heterologous approaches to GPCR deorphanization, significant steps were taken to validate an RNAi-mediated loss-of-function approach to characterizing GPCRs in their native membrane environment. A membrane preparation protocol was optimized for use with planaria, and RNAi was used in conjunction with a biochemical endpoint assay (cAMP RIA) to associate the planarian receptor GtNPR-1 with its neuropeptide ligand GYIRFamide. This process also coupled the receptor to its endogenous G protein signaling pathway (Gαi/o). A small ligand screen of biogenic amines and peptides yielded other leads for application of this method. Among these, we further bioinformatically ranked and pursued 5-HT receptors mediating the aggregate stimulatory effects of 5-HT on cAMP levels. Parallel phenotypic assays revealed that downregulation of Smed-5HT85 significantly decreases planarian motility.
APPENDIX A The genome of the blood fluke Schistosoma mansoni

A paper published in the journal Nature

Matthew Berriman et al. (Mostafa Zamanian, and Tim A Day)

This appendix includes the abstract and selected sections for which I shared primary responsibility in both the underlying analysis and write-up. Supplementary Tables 12, 13, 20, and Supplementary Figure 4 resulted from the work summarized in these sections.

Abstract

Schistosoma mansoni is responsible for the neglected tropical disease schistosomiasis that affects 210 million people in 76 countries. Here we present analysis of the 363 megabase nuclear genome of the blood fluke. It encodes at least 11,809 genes, with an unusual intron size distribution, and new families of micro-exon genes that undergo frequent alternative splicing. As the first sequenced flatworm, and a representative of the Lophotrochozoa, it offers insights into early events in the evolution of the animals, including the development of a body pattern with bilateral symmetry, and the development of tissues into organs. Our analysis has been informed by the need to find new drug targets. The deficits in lipid metabolism that make schistosomes dependent on the host are revealed, and the identification of membrane receptors, ion channels and more than 300 proteases provide new insights into the biology of the life cycle and new targets. Bioinformatics approaches have identified metabolic chokepoints, and a chemogenomic screen has pinpointed schistosome proteins for which existing drugs may be

1Reprinted with permission of Nature: Vol 460 | 16 July 2009 | doi:10.1038/nature08160
2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
3Department of Biomedical Sciences, Iowa State University, Ames, IA USA
4Interdepartmental Neuroscience Program, Iowa State University, Ames, IA USA
active. The information generated provides an invaluable resource for the research community to develop much needed new control tools for the treatment and eradication of this important and neglected disease.

**GPCRs, ligand-gated and voltage-gated ion channels**

GPCRs, ligand-gated and voltage-gated ion channels are targets for 50% of all current pharmaceuticals [211]. At least 92 putative GPCR-encoding genes are present (Supplementary Table 12), the bulk (82) of which are from the rhodopsin family. The largest groups are the α-subfamily (30), which includes amine receptors, and the β-subfamily (24), which contains neuropeptide and hormone receptors. The diversity of the former subfamily underlines the wide range of potential amine/neurotransmitter reactivities of schistosomes, but the tentative identities assigned need to be confirmed by functional studies, as has already been performed for a histamine receptor [139]. Schistosomes detect chemosensory cues, but a large, unique clade of the mediating receptors was not found. However, the 26 ‘orphan’ rhodopsin family GPCRs may include proteins with this role. Outside the large rhodopsin family, representatives from each of the smaller families of GPCRs, glutamate family (2), frizzled family (3), and the secretin/adhesion family (4) are present.

Each of the three major ligand-gated ion channel families— the Cys-loop family, glutamate-activated cation channels, and ATP- gated ion channels— are represented in the schistosome genome. Of the 13 Cys-loop family ligand-gated ion channels, nine encode nicotinic acetylcholine receptor subunits (Supplementary Fig. 4 and Supplementary Table 13). The remaining four anion channel subunits group among GABA (γ-aminobutyric acid), glycine and glutamate receptors, but it is not possible to assign precise identities. The seven schistosome glutamate-activated cation channels comprise at least two sequences from each of the three common sub-groupings. The presence of a functional P2X receptor for ATP- mediated signalling in schistosomes was already known [212], and the data here show at least four more.
Voltage-gated ion channels generate and control membrane potential in excitable cells, and are central to ionic homeostasis. There are examples of successful drugs targeting voltage-gated sodium, potassium and calcium channels [213]. Although voltage-gated sodium channels were not found, at least 41 members from each of the major six transmembrane (6TM) and four transmembrane (4TM) families of potassium channels (Supplementary Table 14) are present. The 6TM voltage-gated potassium channel family (20 members) is the largest, including the well-characterized Kv1.1 channel found in nerve and muscle of adult schistosomes [214]. Other classes of 6TM potassium channels include the KQT channels, large calcium-activated channels, small calcium-activated channels, and cyclic-nucleotide-gated groups. This last group, comprising eight members, is most often associated with signal transduction in primary olfactory and visual sensory cells (Caenorhabditis elegans has only five; [215]). S. mansoni possesses six 4TM inward-rectifying TWIK-related potassium channels (about 46 in C. elegans). There are four α and two β subunits of voltage-gated calcium channels in schistosomes, and a β subunit is implicated as a molecular target of the anti-schistosomal praziquantel [216].

Neuropeptides

Thirteen putative neuropeptides were identified (Supplementary Table 20), indicating that schistosomes may have much greater diversity than the two described previously. Apart from the neuropeptide Fs (NPFs), most are apparently restricted to the Platyhelminthes- their absence from humans making them a credible source of anthelmintic drug leads. The predicted product of npp-6 (the amidated heptapeptide AVRLMRLamide) resembles molluscan myomodulin, whereas the two NPP-13 peptides show 100% carboxy-terminal identity with vertebrate neuropeptide-FF-like peptides (peptides ending with a C-terminal sequence PQRFamide); neither of these has previously been reported in any non-vertebrate organism. The discovery of a second NPF (NPP-21b) as well as the known NPP-21a [201] is reminiscent of the vertebrate neuropeptide Y (NPY) superfamily, and strengthens the argument that NPFs and NPYs have a common ancestry.
<table>
<thead>
<tr>
<th>GeneID</th>
<th>Top BLAST</th>
<th>% ID</th>
<th>E-Value</th>
<th>Final Annotation</th>
<th>TreeFam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Glutamate Activated Cationic Channels</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NMDDA</td>
<td>TF314731</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smp_126350</td>
<td>RF:XP_321646.2</td>
<td>56</td>
<td>4.80E-180</td>
<td>Glutamate Receptor, NMDA</td>
</tr>
<tr>
<td></td>
<td>Smp_147390</td>
<td>RF:XP_971730.1</td>
<td>36</td>
<td>4.60E-121</td>
<td>Glutamate Receptor, NMDA</td>
</tr>
<tr>
<td></td>
<td>Smp_126350</td>
<td>RF:XP_321646.2</td>
<td>56</td>
<td>4.80E-180</td>
<td>Glutamate Receptor, NMDA</td>
</tr>
<tr>
<td></td>
<td>Smp_147390</td>
<td>RF:XP_971730.1</td>
<td>36</td>
<td>4.60E-121</td>
<td>Glutamate Receptor, NMDA</td>
</tr>
<tr>
<td></td>
<td>Kainate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smp_140920</td>
<td>GB:BAC06343.1</td>
<td>64</td>
<td>6.80E-120</td>
<td>Glutamate Receptor, Kainate</td>
</tr>
<tr>
<td></td>
<td>Smp_147430</td>
<td>RF:XP_966711.1</td>
<td>54</td>
<td>9.50E-136</td>
<td>Glutamate Receptor, Kainate</td>
</tr>
<tr>
<td></td>
<td>Smp_153780</td>
<td>RF:XP_611666.2</td>
<td>37</td>
<td>3.90E-124</td>
<td>Glutamate Receptor, Kainate</td>
</tr>
<tr>
<td></td>
<td>AMPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smp_133920</td>
<td>RF:XP_787239.1</td>
<td>30</td>
<td>2.70E-034</td>
<td>Glutamate Receptor, AMPA</td>
</tr>
<tr>
<td></td>
<td>Smp_023290</td>
<td>RF:XP_787239.1</td>
<td>29</td>
<td>3.00E-039</td>
<td>Glutamate Receptor, AMPA</td>
</tr>
<tr>
<td></td>
<td>Cys-Loop Family</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nicotinic Acetylcholine Receptor Subunits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smp_132070.1</td>
<td>GB:ABA60381.1</td>
<td>51</td>
<td>2.80E-086</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_132070.2</td>
<td>GB:ABA60381.1</td>
<td>51</td>
<td>2.80E-086</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_031680</td>
<td>GB:AR84361.1</td>
<td>99</td>
<td>0</td>
<td>nAChR subunit (ShAR1alpha)</td>
</tr>
<tr>
<td></td>
<td>Smp_037960</td>
<td>GB:ABA60381.1</td>
<td>50</td>
<td>2.80E-105</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_142690</td>
<td>GB:ABA60385.1</td>
<td>42</td>
<td>1.20E-081</td>
<td>nAChR subunit (ShAR1beta2)</td>
</tr>
<tr>
<td></td>
<td>Smp_142700</td>
<td>GB:ABA60381.1</td>
<td>49</td>
<td>3.10E-102</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_157790</td>
<td>GB:ABA60385.1</td>
<td>57</td>
<td>5.00E-098</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_176310</td>
<td>GB:ABA60381.1</td>
<td>48</td>
<td>3.80E-095</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_180570</td>
<td>GB:ABA60386.1</td>
<td>37</td>
<td>6.20E-062</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_130390</td>
<td>GB:ABA60386.1</td>
<td>76</td>
<td>4.40E-045</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_139330</td>
<td>GB:ABR84362.1</td>
<td>97</td>
<td>0</td>
<td>nAChR subunit (ShAR1beta)</td>
</tr>
<tr>
<td></td>
<td>Smp_101990</td>
<td>RF:NP_000070.1</td>
<td>30</td>
<td>6.30E-025</td>
<td>nAChR subunit</td>
</tr>
<tr>
<td></td>
<td>Non-nicotinic Cys-loop Receptor Subunits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smp_015630</td>
<td>GB:AAM23270.1</td>
<td>42</td>
<td>2.70E-065</td>
<td>Cys-loop LGIC Subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_096480</td>
<td>SP:O75311</td>
<td>37</td>
<td>2.30E-068</td>
<td>Cys-loop LGIC Subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_099500</td>
<td>RF:XP_974894.1</td>
<td>36</td>
<td>3.60E-031</td>
<td>Cys-loop LGIC Subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_176730</td>
<td>RF:NP_001024077.1</td>
<td>51</td>
<td>1.20E-028</td>
<td>Cys-loop LGIC Subunit</td>
</tr>
<tr>
<td></td>
<td>Smp_104890</td>
<td>SP:O57965</td>
<td>47</td>
<td>2.40E-062</td>
<td>Cys-loop LGIC Subunit</td>
</tr>
</tbody>
</table>

Table A.1 Summary of Ligand-Gated Ion Channels (LGIC)
<table>
<thead>
<tr>
<th>GeneID</th>
<th>Top BLAST</th>
<th>% ID</th>
<th>E-Value</th>
<th>Final Annotation</th>
<th>TreeFam</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP gated ion channels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smp_099640</td>
<td>GB:CAH04147.1</td>
<td>48</td>
<td>2.00E-097</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Smp_176860</td>
<td>GB:CAH04147.1</td>
<td>46</td>
<td>1.80E-073</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Smp_179310</td>
<td>GB:CAH04147.1</td>
<td>100</td>
<td>5.60E-164</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Smp_114060</td>
<td>GB:AAX11263.1</td>
<td>43</td>
<td>1.60E-031</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Smp_089780</td>
<td>RF:NP_990079.1</td>
<td>37</td>
<td>5.50E-024</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Smp_005030</td>
<td>GB:AAX11263.1</td>
<td>100</td>
<td>1.20E-035</td>
<td>P2x Receptor Subunit</td>
<td>TF328633</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclic-nucleotide-gated cation channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smp_056560</td>
<td>RF:XP_967432.1</td>
<td>50</td>
<td>1.10E-055</td>
<td></td>
<td>TF318250</td>
</tr>
<tr>
<td>Smp_152480</td>
<td>RF:XP_395071.2</td>
<td>54</td>
<td>5.80E-146</td>
<td></td>
<td>TF318250</td>
</tr>
<tr>
<td>Smp_152500</td>
<td>RF:XP_314248.2</td>
<td>66</td>
<td>7.50E-066</td>
<td></td>
<td>TF318250</td>
</tr>
<tr>
<td>Smp_155040</td>
<td>RF:XP_554755.1</td>
<td>43</td>
<td>8.90E-115</td>
<td></td>
<td>TF318250</td>
</tr>
<tr>
<td>Smp_194700</td>
<td>SP:Q90805</td>
<td>53</td>
<td>6.50E-166</td>
<td></td>
<td>TF318250</td>
</tr>
<tr>
<td>Amiloride-sensitive sodium channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smp_058270</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TF330663</td>
</tr>
<tr>
<td>Smp_083980</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TF317359</td>
</tr>
<tr>
<td>Smp_175020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TF317359</td>
</tr>
<tr>
<td>Smp_052630</td>
<td>GB:AAK20896.1</td>
<td>24</td>
<td>6.90E-014</td>
<td></td>
<td>TF317359</td>
</tr>
<tr>
<td>Smp_162680</td>
<td>GB:AAF80601.1</td>
<td>25</td>
<td>2.20E-017</td>
<td></td>
<td>TF317359</td>
</tr>
<tr>
<td>Smp_093210</td>
<td>GB:AAF80601.1</td>
<td>22</td>
<td>2.60E-026</td>
<td></td>
<td>TF317359</td>
</tr>
<tr>
<td>Smp_180260</td>
<td>GB:AAK20896.1</td>
<td>26</td>
<td>2.20E-012</td>
<td></td>
<td>TF317359</td>
</tr>
</tbody>
</table>

Table A.2  Summary of Ligand-Gated Ion Channels (LGIC)- Continued
<table>
<thead>
<tr>
<th>GeneID</th>
<th>Annotation</th>
<th>TreeFam</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMP_081250</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_129380</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_136440</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_035160</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_035180</td>
<td>voltage-gated potassium channel</td>
<td>TF315186</td>
</tr>
<tr>
<td>SMP_035870</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_146620</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_148670</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_151810</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_152350</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_063930</td>
<td>voltage-gated potassium channel SKv1.1</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_157490</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_069240</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_160780</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_161140</td>
<td>voltage-gated potassium channel</td>
<td>TF313130</td>
</tr>
<tr>
<td>SMP_163090.1</td>
<td>voltage-gated potassium channel</td>
<td>TF352511</td>
</tr>
<tr>
<td>SMP_163090.2</td>
<td>voltage-gated potassium channel</td>
<td>TF352511</td>
</tr>
<tr>
<td>SMP_121190</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_194710</td>
<td>voltage-gated potassium channel</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_144310</td>
<td>voltage-gated potassium channel, KQT</td>
<td>TF315186</td>
</tr>
<tr>
<td>SMP_144310</td>
<td>voltage-gated potassium channel, KQT</td>
<td>TF315186</td>
</tr>
<tr>
<td>SMP_008170</td>
<td>calcium-activated potassium channel, large conductance</td>
<td>TF313103</td>
</tr>
<tr>
<td>SMP_161450</td>
<td>calcium-activated potassium channel, large conductance</td>
<td>TF315015</td>
</tr>
<tr>
<td>SMP_166620</td>
<td>calcium-activated potassium channel, large conductance</td>
<td>TF313947</td>
</tr>
<tr>
<td>SMP_166910</td>
<td>calcium-activated potassium channel, large conductance</td>
<td>TF314283</td>
</tr>
<tr>
<td>SMP_156150</td>
<td>calcium-activated potassium channel, small conductance</td>
<td>TF315015</td>
</tr>
<tr>
<td>SMP_056560</td>
<td>cyclic-nucleotide-gated cation channel</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_152480</td>
<td>cyclic-nucleotide-gated cation channel</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_152500</td>
<td>cyclic-nucleotide-gated cation channel</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_155040</td>
<td>cyclic-nucleotide-gated cation channel</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_194700</td>
<td>cyclic-nucleotide-gated cation channel</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_153100</td>
<td>cyclic-nucleotide-gated cation channel, hyperpolarization-activated</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_168880</td>
<td>cyclic-nucleotide-gated cation channel, hyperpolarization-activated</td>
<td>TF318250</td>
</tr>
<tr>
<td>SMP_174860</td>
<td>cyclic-nucleotide-gated cation channel, hyperpolarization-activated</td>
<td>TF318250</td>
</tr>
</tbody>
</table>

Table A.3  Summary of Voltage-Gated and Other Ion Channels
<table>
<thead>
<tr>
<th>GeneID</th>
<th>Annotation</th>
<th>TreeFam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>4TM K+ Channels</strong></td>
<td></td>
</tr>
<tr>
<td>SMP_127170</td>
<td>twik family of potassium channels-related</td>
<td>TF316115</td>
</tr>
<tr>
<td>SMP_034850</td>
<td>twik family of potassium channels-related</td>
<td>TF313947</td>
</tr>
<tr>
<td>SMP_141570</td>
<td>twik family of potassium channels-related</td>
<td>TF316115</td>
</tr>
<tr>
<td>SMP_046640</td>
<td>twik family of potassium channels-related</td>
<td>TF316115</td>
</tr>
<tr>
<td>SMP_151120</td>
<td>twik family of potassium channels-related</td>
<td>TF313947</td>
</tr>
<tr>
<td>SMP_155970</td>
<td>twik family of potassium channels-related</td>
<td>TF313947</td>
</tr>
<tr>
<td></td>
<td><strong>Voltage-Gated Ca(^{2+}) Channels</strong></td>
<td></td>
</tr>
<tr>
<td>SMP_020270</td>
<td>high voltage-activated calcium channel Cav1</td>
<td>TF312805</td>
</tr>
<tr>
<td>SMP_159990</td>
<td>high voltage-activated calcium channel</td>
<td>TF312805</td>
</tr>
<tr>
<td>SMP_020170</td>
<td>high voltage-activated calcium channel Cav2A</td>
<td>TF312805</td>
</tr>
<tr>
<td>SMP_004730</td>
<td>high voltage-activated calcium channel Cav2B</td>
<td>TF312805</td>
</tr>
<tr>
<td>SMP_135140</td>
<td>high voltage-activated calcium channel beta subunit CavB1</td>
<td>TF316195</td>
</tr>
<tr>
<td>SMP_141660</td>
<td>high voltage-activated calcium channel beta subunit CavB2</td>
<td>TF316195</td>
</tr>
<tr>
<td>SMP_141780</td>
<td>Four domain-type voltage-gated ion channel alpha-1 subunit</td>
<td>TF312843</td>
</tr>
<tr>
<td></td>
<td><strong>Voltage-Gated Cl- Channels</strong></td>
<td></td>
</tr>
<tr>
<td>SMP_058360.1</td>
<td>chloride channel protein</td>
<td>TF313867</td>
</tr>
<tr>
<td>SMP_058360.2</td>
<td>chloride channel protein</td>
<td>TF313867</td>
</tr>
<tr>
<td>SMP_071970</td>
<td>chloride channel protein</td>
<td>TF313867</td>
</tr>
</tbody>
</table>

Table A.4  Summary of Voltage-Gated and Other Ion Channels- Continued
APPENDIX B  Yeast GPCR Assay

This appendix contains data related to progress towards the yeast-based deorphanization of a *Schistosoma mansoni* neuropeptide-like GPCR (Smp_011940). Figure B.1 provides a comparison of the endogenous (left) and modified (right) pathways that allow for these strains to be used for receptor deorphanization. The first candidate receptor chosen for this approach was originally identified in a regular expression-driven ORF screen of GPCR homologues in *S. mansoni*, and exhibits a great deal of sequence similarity with an allatostatin-responsive GPCR (Figure B.2).

![Figure B.1 Heterologous GPCR expression in yeast.](image)
Smp_011940 (0029329) is aligned with an allatostatin receptor from *Periplaneta americana* (AAK52473). Yellow and green signify sequence identity and sequence similarity, respectively.

### Full-length GPCR sequence characterization with 5',3'-RACE PCR.

5',3'-Rapid Amplification of cDNA Ends (RACE) PCR was used to elucidate full-length gene structure, and to establish with certainty the coding sequence of the GPCR to be heterologously expressed. Poly A+ RNA was extracted from adult *S. mansoni* tissue using TRI Reagent (Sigma) and Dynabead's mRNA Purification Kit (Dynal). This mRNA was used to synthesize cDNA with SMART RACE cDNA Amplification Kit (BD Biosciences). Gene specific primers were designed from ORF sequences and used for 5' and 3' RACE PCR in conjunction with the cDNA templates. Reactions were visualized on 1.2% agarose gel. Discrete amplicons were excised from gel, purified, and subcloned into pGEM-T Easy Vector (Promega). The resulting construct was transformed into JM109 competent *Escherichia coli* (Promega). Individual clones identified by a blue-white colony screen were filtered with colony PCR, and positive clones were cultured in LB broth overnight. Plasmic DNA was purified with the Wizard Plus SV Miniprep kit (Promega) and sequence confirmed. The final consensus sequence is shown in Figure B.3.
Sub-cloning receptor coding sequence into yeast expression vector

The Smp_011940 coding sequence was sub-cloned into the yeast expression vector Cp4258 (2µ ori AmpR LEU2 REP3 PGK-promoter-MFα1-(1-89)) [217, 218]. In the resulting construct, GPCR expression is driven by the PGK promoter, while the 89 amino acid Mfα1-leader sequence is fused to the receptor N terminus to promote GPCR export to the cell membrane [219]. Smp_019940 (in pcDNA3.1) was used as a template in a two-step PCR that added a 6-mer spacer and NcoI and XbaI restriction sites to the 5’ and 3’ ends of the coding sequence, respectively. The complete coding sequence was then ligated into Cp4258. The resulting construct was transformed into JM109 competent Escherichia coli (Promega). Individual clones were screened with a colony PCR, and positive clones were cultured in LB broth overnight. Plasmid DNA was purified with the Wizard Plus SV Miniprep kit (Promega) and sequenced. A perfect clone was identified, and used for yeast transformation.

Transformation of yeast with recombinant plasmid.

Cy14083 cells were grown in 10 mL YPD broth overnight (30°C, 250 RPM), achieving an optical density (OD$_{600}$) of about 1. The culture was scaled up by inoculation in 300 mL YPD in a 1 L flask until OD$_{600}$ measures fell in the 0.3-0.5 range. The sample was centrifuged at 4,000 x g for 5 min and resuspended in 10 mL H$_2$O. Another centrifugation step was performed at 5,000 x g for 5 min, the pellet was resuspended in 10 mL buffered lithium solution (1 mL 1M LiAc, 1 mL 10X TE and 8 mL H$_2$O) and incubated for 1 hr at 30°C. Carrier ssDNA (Salmon sperm) was boiled for 10 minutes and left to cool on ice for 45 min. 200 ug carrier ssDNA and 1 ug of recombinant plasmid (Smp_0119949∥Cp4258) was added to a sterile 1.5 mL tube. 200 uL of yeast suspension was added, followed by 1.2 mL of PEG 4000 solution (1 mL 10X TE, 1 mL 10X LiAc, 8 mL H$_2$O 50% PEG 4000). The mixture was quickly vortexed and incubated for 30 min at 30°C (250 RPM). Cells were heat shocked at 42°C for 15 min and placed in the microcentrifuge for 5 sec. 500 uL was diluted in 1 mL 1X TE, and 200 uL of the dilution were spread on selective (-LEU) plates. Plates were incubated for 2-4 days at 30°C until transformants appeared. Colonies were inoculated in 2 mL -LEU broth and incubated for
20-24 hrs at 30°C. Yeast genomic DNA was isolated and used in a PCR screen with primers specific for Smp_011940. Four positive colonies were propagated and stored for later use in functional assays.

Figure B.3  Full-length gene transcript of Smp_011940.

As revealed by RACE PCR, the Smp_011940 transcript is 2044 nucleotides in length and contains a 1248 nucleotide open reading frame encoding a 416 amino acid peptide.
APPENDIX C  PROF1 Primers and RT-PCR

S. mansoni and S. mediterranea PROF1 primers

> S. mansoni:

SMP084270

F: 5’ atgataagtatgaactcaagtgaattaatttttacctg 3’ (Tm=57)
R: 5’ tcagtaattgtggcctgatacaacgct 3’ (Tm=58)

Predicted length: 1,107 bp (full-length)

SMP041880

F-primer: 5’ atgttacataatactacaactatagattatagtcagttagt 3’ (Tm=59)
R-primer: 5’ tcaatcttcagttctttggggtctatgca 3’ (Tm=59)

Predicted length: 1,026 bp (full-length)

> S. mediterranea: (Tm: 59-60 C for all primers)

SMDC6472.1

F: 5’ tgcaacaaatgtgacgttt 3’ / R: 5’ ggcaagtgtagagatggcaaa 5’

Predicted length: 534

F: 5’ tgtgcacaataaactcctg 3’ / R: 5’ taaaaccgaagctgtgcat 5’

Predicted length: 1007

SMDC8510.2

F: 5’ cagccctttgggatattta 3’ / R: 5’ gcggccaaaatagcaaaaa 5’

Predicted length: 446

F: 5’ gcattgcctttccactctg 3’ / R: 5’ cttggaatgggtctttt 5’

Predicted length: 956

SMDC8510.3A

F: 5’ ccaaactccatgaccgaact 3’ / R: 5’ aggatccgcaaaggagaat 5’

Predicted length: 486

F: 5’ aacaaagggcatactctcg 3’ / R: 5’ cgcaatattttcgactcttc 5’

Predicted length: 1009
SMDC14380.2

F: 5' ccaaactccatgaccgaact 3' / R: 5' aggatccgcaaaaggagaat 5'
Predicted length: 493

F: 5' aacaacgggcaaatctctcg 3' / R: 5' gcgaatatgtgctgaccttc 5'
Predicted length: 1026

SMDC504.1

F: 5' tagagctcttgtgcaaaattgct 3' / R: 5' gcgaataaaccgcagttgacg 5'
Predicted length: 433

F: 5' tggacctgtgtaaatgacagg 3' / R: 5' gaggataatgtgctgtaacca 5'
Predicted length: 1042

SMDC1889.2

F: 5' gccaattgtaccatgtgtcg 3' / R: 5' aaattcgaatggtcgtatgc 5'
Predicted length: 425

F: 5' tccctgggatttatttgccttc 3' / R: 5' aaaaattcgaatggctgatgc 5'
Predicted length: 1039

SMDC15273.1

F: 5' aggcgaactacgcgttcata 3' / R: 5' ccccataagtccacgaagaa 5'
Predicted length: 411

F: 5' tcgacttgtcaaatgtcaaagg 3' / R: 5' gaggataatggtcgtatgc 5'
Predicted length: 1031

SMDC7587.4

F: 5' tgtgctgctgttgattcata 3' / R: 5' ttgcaaaaactaccaacgaca 5'
Predicted length: 460

F: 5' tgatctatatagccagatcagc 3' / R: 5' aaagattttgttaatcgtccaga 5'
Predicted length: 1001

SMDC2411.1

F: 5' tgtgcgctgttgattgaata 3' / R: 5' aattttcagctctttgttg 5'
Predicted length: 499

F: 5' tgtgctgctgttgattgaata 3' / R: 5' aacgcgggaaatgtgatttc 5'
Predicted length: 1059

SMDC2955.1

F: 5' acgataaggtttggttattg 3' / R: 5' ttccctcgttgagatttg 5'
Predicted length: 495

F: 5' cttgcctacttggattc 3' / R: 5' ttccctcgttgagatttg 5'
Predicted length: 949
SMDC5598.1A

F: 5' ttctctgcttggtcttcc 3' / R: 5' cgcaatttccctgtaat 5'
Predicted length: 500

F: 5' cggtaatggtgcttgg 3' / R: 5' cgaaagttccagtaacgtgaa 5'
Predicted length: 905

SMDC4570.1

F: 5' ttcatcgattgtggcctcg 3' / R: 5' tgatactgatgctggggtca 5'
Predicted length: 505

F: 5' caagtgatgtaatccccgaat 3' / R: 5' atggctggggtaccaataac 5'
Predicted length: 904

SMDC5598.1B

F: 5' ggtgagtagggccctgtga 3' / R: 5' tacgccaagatgttggtat 5'
Predicted length: 528

F: 5' ggtgagtagggccctgtga 3' / R: 5' gtagcatatgaacatataaggtgtg 5'
Predicted length: 942
RT-PCR Results

Performed using primer set two (400-600 bp products)

<table>
<thead>
<tr>
<th>Reaction component</th>
<th>V (uL)</th>
<th>Stage</th>
<th>Temp</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer 1 (10 uM)</td>
<td>3.0</td>
<td>Denature</td>
<td>94.0 °C</td>
<td>2:00</td>
</tr>
<tr>
<td>Primer 2 (10 uM)</td>
<td>3.0</td>
<td>Denature</td>
<td>94.0 °C</td>
<td>0:30</td>
</tr>
<tr>
<td>cDNA template</td>
<td>4.0</td>
<td>Anneal</td>
<td>55.0 °C</td>
<td>0:30</td>
</tr>
<tr>
<td>10X Buffer</td>
<td>5.0</td>
<td>Extented</td>
<td>72.0 °C</td>
<td>1:00</td>
</tr>
<tr>
<td>MgCl₂ (50 mM)</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dNTP (10 mM)</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₂O</td>
<td>31.5</td>
<td>Hold</td>
<td>4.0 °C</td>
<td>∞</td>
</tr>
<tr>
<td>Platinum Taq</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total volume</td>
<td>50.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table C.1 PROF1 RT-PCR conditions

PCR reaction (left), PCR thermocycler conditions (right). Template: *S. mediterranea* cDNA

Figure C.1 *S. mediterranea* PROF1 RT-PCR.

Lanes 1 and 15 are low DNA mass ladder. Lanes 2-14 are visualized PCR reactions with PROF1 primers corresponding to (in order): SMDC6472.1, SMDC8510.2, SMDC14380.2, SMDC8510.3A, SMDC, SMDC1889.2, SMDC15273.1, SMDC7587.4, SMDC2411.1, SMDC2955.1, SMDC5598.1A, SMDC4570.1, and SMDC5598.1B. Although spurious products can be seen in some lanes, correct-sized amplicons were present for all putative PROF1 receptors that were PCR amplified.
APPENDIX D  RNAi vector

Two T7 RNA polymerase promoters are flanked by two class I T7 transcriptional terminators to improve transcription efficiency. A gene fragment is subcloned between the \textit{attP1} and \textit{attP2} recombination sites, and the construct is transformed into RNase-III deficient bacteria with IPTG-inducible T7 polymerase activity to generate target-specific dsRNA.
APPENDIX E  cAMP Assay Optimization

This sample optimization experiment was performed using a *D. tigrina* membrane preparation, after alteration of a *S. mansoni* membrane preparation protocol. Establishes that DMSO is safe to use as a solvent at the concentrations used in experiments. Establishes that 10^{-4} M forskolin provides a more robust cAMP range, and allows for the measurement of NPF inhibition of cAMP with better resolution than forskolin used at 10^{-5} M. Confirms significant 5-HT stimulation of cAMP at 10^{-5} M, even as 10^{-4} M 5-HT membrane treatments have been previously published.
APPENDIX F  PERL Code

This appendix contains Perl source code used in the completion of this work. This includes scripts for extracting transmembrane domains, generating SVM feature vectors, calculating position-specific entropy in multiple sequence alignments, and parsing Blast output files.

**TM extraction**

```perl
#!/usr/bin/perl
use strict;
use warnings;

# Processes FASTA-formatted sequence file
open(FILE, "X.fasta") or die("Unable to open file");
my $a;
while ($_ = <FILE>) {
    if ($_ =~ />/) {
        chomp $_;$_="&";

    } else {
        chomp $_;$_="&";

    }
```
$_ =~s/\s//g;
$a.=$_;
}
$_ =~s/>/\&>/g;
$a = $a."&";
#Processes corresponding HMMTOP output file
open(FILE2, "X_HMMTOP.out") or die("Unable to open file");
my $b;
while ($_ = <FILE2>) {
    chomp $_;
    $_ = $_."\*";
    $b.=$_;
}
my $b2 = $b;
#Creates output files
open(OUTFILE, ">X_TMextract.fasta");
open(OUTFILE2, ">X-TMcount.csv");
my @TMcounter;
#Regular expression used to parse and identify headers-sequences pairs
while ($a =~ />(.*)\&([A-Z]*\{30,\})\&/g){
    my $geneid = $1;
    my $seq = $2;
    my $geneid2 = substr($geneid, 0, 99);
    #Regular expression used to identify the HMMTOP coordinate line
    #Extract and concatenate TM domains if TM-count is 3-15
    if ($b2 =~ /($geneid2)\s+[A-Z]{2,3}\s+(\[3-9]|1[0-5])\s+(\d+.*?)\*/g){
        my $TMtemp = $2;
        push @TMcounter, $TMtemp;
        my $COOR = $3;
        my $i = 1;
        my @TMaa = 'i';
        while ($COOR =~ /\d+\s+(\d+)/g){
            my $CR1 = $1;
            my $CR2 = $2;
            #+-5 aa
my $TMaatemp = substr ($seq,$CR1-6,$CR2-$CR1+11);
push @TMaa, $TMaatemp;
$i = $i+1;
}
while ($geneid =~ /(.*?)\s/) {
    $geneid = $1;
}

#Prints TM-only sequences to FASTA file
print OUTFILE ">$geneid,-TM:',$TMtemp,\n; 
print OUTFILE join "", @TMaa; 
print OUTFILE \n;
}

#Hash that maps TM numbers to their frequency
my %TM_counters;
for (@TMcounter){
    $TM_counters{$_}++;
}

#Prints hash in CSV table format
while (my ($key, $value) = each %TM_counters) {
    my $output = "$key ==> $value;"
    print $output, "\n";
    print OUTFILE2 "$key,$value; 
    print OUTFILE2 "\n";
}
close OUTFILE;close OUTFILE2;

TM extraction: joining of TM domains with character barrier

#Adds special character "-" between concatenated TM domains
#Required for SVM feature vector generation for model T7
#Swap first line for second in previous script:
my $TMaatemp = substr ($seq,$CR1-6,$CR2-$CR1+11);
my $TMaatemp = "-".substr ($seq,$CR1-3,$CR2-$CR1+5)."-";
SVM dipeptide frequency feature vector construction: full-length model (FL)

# This script streams one input file (FASTA):
# 1) Sequence file (with no special characters) with full-length seqs
# And generates one output file (txt):
# 1) 400-element frequency vectors taken over each full-length sequence

#!/usr/bin/perl
use strict;
use warnings;

# Declare amino acid alphabet in array form
my @aa1 = ( "A", "C", "D", "E", "F", "G" .. "Q", "R", "S", "T", "V", "W", "Y" );
my @aa2 = ( "A", "C", "D", "E", "F", "G" .. "Q", "R", "S", "T", "V", "W", "Y" );
my $z;
my @seqnames;
my @seqs;
# Read in sequence file
while ($_ = <> ) {
  if ($_ =~ />/) {
    $_ = $_ . "*";
  }
  chomp $_;
  $z .= $_;
}
$z =~ s/\s//g;
$z =~ s/>\>/\*>/g;
$z = $z . "*";
my $z2 = $z;
my $z3 = $z;
# Hash associating amino acids with numbers (1-20)
my %AA_numbering = ( 'A' => '1', 'C' => '2', 'D' => '3', 'E' => '4', 'F' => '5',
  'G' => '6', 'H' => '7', 'I' => '8', 'K' => '9', 'L' => '10',
  'M' => '11', 'N' => '12', 'P' => '13', 'Q' => '14', 'R' => '15',
  'S' => '16', 'T' => '17', 'V' => '18', 'W' => '19', 'Y' => '20',
  'B' => '21', 'Z' => '22', 'X' => '23', 'U' => '24', 'J' => '25',
  '0' => '26');
Generate duplet hash by first generating a duplet array

```perl
my @duplet_array;
my $i=0;
for ($i=0; $i<20; $i++) {
    my $one = $aa1[$i];
    my $j=0;
    for ($j=0; $j<20; $j++) {
        my $two = $aa2[$j];
        my $onetwo = $one.$two;
        push @duplet_array, $onetwo;
    }
}
```

Create a seed hash, mapping every aa pair to numbers 1-400

```perl
my %AA_duplet_numbering;
my $k=1;
for (@duplet_array) {
    $AA_duplet_numbering{$_}=$k;
    $k = $k +1;
}
```

Extract filename

```perl
my $ARGV2='';
while ($ARGV =~/(.*)\.fasta/g){
    $ARGV2=$1;
}
```

Declares feature vector output file

```perl
open(OUTFILE, ">$ARGV2_FV.txt");
```

Reads in and processes individual sequences

```perl
while ($z2 =~ />(.*)\*(\[ACDEFGHIKLMNPQRSTVWY-]+)\*/g) {
    my $seqnumber = $1; #sequence label
    my $sequence1 = $2; #raw sequence
    my $seq_length = length($sequence1); #sequence length
    #Reads in aa duplets in both frames and adds them to vector
    my @duplets;
```
while ($sequence1 =~ /([ACDEFGHIKLMNPQRSTVWY]{2})/g) {
    push @duplets, $1;
}

my $sequence2 = substr ($sequence1,1,);
while ($sequence2 =~ /([ACDEFGHIKLMNPQRSTVWY]{2})/g) {
    push @duplets, $1;
}

# Creates hash that maps duplets to their frequency
my %duplet_counter;
for (@duplets) {
    $duplet_counter{$_}++;
}

my @printarray2;
while (my ($key, $value) = each %duplet_counter) {
    my $output = "$key ===> $value";
    push @printarray2, $output;
}

# Sorts duplets alphabetically
my @sorted2 = sort {
    $a cmp $b
} @printarray2;

print join
    "\n",
    @sorted2;
# Prints progress

# Creates ordered frequency array
my @AA_duplet_ordered;
for (@duplets) {
    push @AA_duplet_ordered, $AA_duplet_numbering{$_};
}

# Creates ordered hash that maps duplet numbers to frequency
print
    "\n",
    "AA_duplet_numb => frequency:",
    "\n";
my %duplet_counterb;
for (@AA_duplet_ordered) {
    $duplet_counterb{$_}++;
}

print OUTFILE "x "; # Where x is an integer- typically 0
# Orders hash keys into new array
my @duplet_counterb_temp = sort {
    $a <=> $b
} (keys %duplet_counterb);
for (@duplet_counterb_temp) {
my $key = $_;  # dipeptide key
my $value = $duplet_counterb{$_};  # frequency value
my $valuef = $value/($seq_length-1);  # converts frequencies to proportions
print OUTFILE "$key:$valuef ";  # prints final fv values to output file
}
print OUTFILE "\n";
}
close (OUTFILE);

SVM dipeptide frequency feature vector construction: T1 model

# This script streams one input file (FASTA):
# 1) Sequence file (with no special characters) with full-length seqs
# And generates one output file (txt):
# 1) 400-element dipeptide frequency vector for each sequence
# FVs calculated over length of concatenated TM domains
# Same source code as previous script

SVM dipeptide frequency feature vector construction: T7 model

# This script streams one input file (FASTA):
# 1) TM-only sequence file with TM borders marked
# And generates one output file (txt):
# 1) 2800-element dipeptide frequency vector for each sequence
# Ordered concatenation of 400-element FVs calculated for each TM
# Replace everything past the line "#***" in FL script

# Reads in and processes individual sequences
while ($z2 =~ />(.*?)\*(\[ACDEFGHIKLMNPQRSTVWY-]+)\*/g){
    my $seqnumber = $1;
    my $sequence1 = $2;
    # Reads in individual TM partitions in order
    my @sequencepartitions;
    while ($sequence1 =~ /\[(ACDEFGHIKLMNPQRSTVWY)\]+\)/g){
        push @sequencepartitions, $1;
my $x=0;
for (@sequencepartitions) {
    print "\n","Duplets","\n";
    my @duplets = ();
    my $sequence1a = $_;
    my $seq_length1a = length($sequence1a);
    while ($sequence1a =~ /([ACDEFGHIKLMNPQRSTVWY]{2})/g) {
        push @duplets, $1;
    }
    my $sequence2a = substr ($sequence1a, 1,);
    while ($sequence2a =~ /([ACDEFGHIKLMNPQRSTVWY]{2})/g) {
        push @duplets, $1;
    }
    # Creates hash that maps duplets to their frequency
    my %duplet_counter = ();
    for (@duplets) {
        $duplet_counter{$_}++;
    }
    my @printarray2 = ();
    while (my ($key, $value) = each %duplet_counter) {
        my $output = "$key ===> $value";
        push @printarray2, $output;
    }
    my @sorted2 = ();
    my @sorted2 = sort { $a cmp $b } @printarray2; # sorts duplets alphabetically
    # Creates ordered hash that maps duplet numbers to frequency
    my @AA_duplet_ordered = ();
    for (@duplets) {
        push @AA_duplet_ordered, $AA_duplet_numbering($_);
    }
    print "\n","AA_duplet_numb => frequency:","\n";
    my %duplet_counterb = ();
    for (@AA_duplet_ordered) {

$duplet_counterb($_)++;  
}
my @printarray2f =();  
#Orders hash keys into new array  
my @duplet_counterb_temp =();  
my @duplet_counterb_temp = sort {a <=> b} (keys %duplet_counterb);  
for (@duplet_counterb_temp) {
    my $key = $_+(400*$x);  
    my $value = $duplet_counterb($_);  
    my $valuef = $value / ($seq_length1a -1);  
    my $output = "$key ===> $valuef";  
    push @printarray2f, $output;  
    print OUTFILE "$key:$valuef 
    }  
$x=$x+1;  
}
print OUTFILE "\n";
}
close (OUTFILE);

Multiple Sequence Alignment Shannon Entropy Calculator

#This script streams in one input file:  
#1) FASTA-formatted multiple sequence alignment  
#And generates one file:  
#1) CSV formatted-file containing MSA positions and corresponding entropy values

#!/usr/bin/perl
use strict;
use warnings;

#Spreadsheet header
open OUTFILE, (">entropy.csv
print OUTFILE "Position","","Entropy","\n"
#Read in MSA (fasta) and amass into single string
my $z;
while ($_ = <> ) {
    if ($_ =~ />/) {
        chomp $_;
        $_ = $_ . "&";
    }
    chomp $_;
    $z .= $_;
}
$z =~ s/\s//g;
my $z2 = $z;

# Logarithm subroutine
sub log_base {
    my ($base, $value) = @_; 
    return log($value)/log($base);
}

# Find sequence length
my $alength = ''; 
if ($z2 =~ />.*?\&([^ACDEFGHIKLMNPQRSTVWYX-]+)/) {
    my $atemp = $1;
    $alength = length ($atemp);
}

# Find number of sequences
my $k = 0;
while ($z2 =~ />/g) {
    $k = $k + 1;
}
my $seqnumber = $k;

# Calculate Shannon Entropy for each position in alignment
my $i = 0;
for ($i = 0; $i <$alength; $i++) {
    print "ARRAY POSITION NUMBER: ", $i + 1;
    print OUTFILE $i + 1, ",";
    my @singlets;
    while ($z =~ />.*?\&([^ACDEFGHIKLMNPQRSTVWYX-]+)/g) {
        # Continue with the rest of the code
    }
}
my $letters = substr ($1,$i,1);
push @singlets, $letters;
}
print join ' ', @singlets;
#Create a counter for this position
my %singlet_counter;
for (@singlets) {
    $singlet_counter{$_}++;
}
my @printarray;
my @printarray2;
#Only consider columns where seqs with gap < 1/4th of the total seq number
#Otherwise entropy set to default arbitrarily high value (5)
my $seqnumberb = $seqnumber/4;
my $entropy = 5;
if ($singlet_counter{'-'} < $seqnumberb){
    $entropy = 0;
    #Counts instances of each aa for given column, converts to probabilities
    #Calculates Shannon entropy with probabilities
    while (my ($key, $value) = each %singlet_counter) {
        my $prob = $value/$seqnumber;
        my $templog = log_base(2,$prob);
        $templog = (-1)*$prob*$templog;
        $entropy = $entropy+$templog;
        my $output2 = "$key ===> $value";
        push @printarray1, $output2;
        my $output = "$key ===> $templog";
        push @printarray2, $output;
    }
}
#Print frequencies, probabilities, and entropy to screen
print "AA => frequency: ", "\n";
print join "\n", @printarray1;
print "\n";
print "AA => probability: ", "\n";
print join "\n", @printarray2;
print "\n";
print "ENTROPY: ", $entropy,"\n";
#Print entropy to CSV file
print OUTFILE2 $entropy,"\n";
}
}
close (OUTFILE);

BLAST parser

#This script streams in one input file:
#1) Blast result output file (txt) generated with netblast-2.2.18
#And generates one output file:
#1) CSV formatted-file containing information about top Blast hits

#!/usr/bin/perl
use strict;
use warnings;

#Reads in Blast output file and stores in single string
my $a;
while ($_ = <>) {
    chomp $_;
    $a.=$_;
}

#Creates csv file, fills out first row with desired parameter list
open(OUTFILE, ">BLAST.csv");
print OUTFILE "query id,query length...","\n";
#Regular expression extract query and hit information
while ($a =~ /Query=\s(\[[^\s\]*?\])\s.*?\((\[0-9\]+)\sletters\)\((.*?)\)Lambda/g){
    my $queryid = $1; # query id
    my $querylength = $2; #query length
    my $region = $3; #region
    print OUTFILE $queryid,"",$querylength,"\n";
    while ($region =~ /\>(ref\.*\s\.)\s.*/g){
        my $regionid = \1; # region id
        my $regionlength = \2; #region length
        my $regionmatch = \3; #region match
        print OUTFILE $regionid,"",$regionlength,"\n";
    }
}
Expect\s=\s(.*)\s+Identities\s=\s(.*)\s+/g){
    my $hitid = $1; # hit id
    my $hitannotation = $2; # hit annotation
    my $hitspecies = $3; # hit species
    my $hitlength = $4; # hit length
    my $hiteval = $5; # hit E-value
    my $hitidentity = $6; # hit identity
    $hitannotation =~ s/\s{2,}/\s/g;
    $hitannotation =~ s/\s/,/\s/g;
    $hitspecies =~ s/\s{2,}/\s/g;
    print OUTFILE " ", ", ", ", ", ", ";
    # Choose and print parameters of interest to file
    print OUTFILE $hitid,...," \n";
}
}
close OUTFILE;
BIBLIOGRAPHY


