5-2011

Functional Classification of Divergent Protein Sequences and Molecular Evolution of Multi-Domain Proteins

Pooja K. Strope
University of Nebraska-Lincoln, poojastrope@gmail.com

FUNCTIONAL CLASSIFICATION OF DIVERGENT PROTEIN SEQUENCES AND MOLECULAR EVOLUTION OF MULTI-DOMAIN PROTEINS

by

Pooja K. Strope

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professor Etsuko Moriyama

Lincoln, Nebraska

May 2011

FUNCTIONAL CLASSIFICATION OF DIVERGENT PROTEIN SEQUENCES AND MOLECULAR EVOLUTION OF MULTI-DOMAIN PROTEINS

Pooja K. Strope, Ph.D.

University of Nebraska, 2011

Advisor: Etsuko Moriyama

Transmembrane proteins and multi-domain proteins together make up more than 80% of the total proteins in any eukaryotic proteome. Therefore accurately classifying such proteins into functional classes is an important task. Furthermore, understanding the molecular evolution of multi-domain proteins is important because it shows how various domains fuse to form more complex proteins, and acquire new functions possibly affecting the organismal level of evolution. In this thesis, I first investigated the performance of several protein classifiers using one of the most divergent transmembrane protein families, the G-protein-coupled receptor (GPCR) superfamily, as an example. Alignment-free classifiers based on support vector machines using simple amino acid compositions were effective in remote-similarity detection even from short fragmented sequences. While a support vector machine using local pairwise-alignment scores showed very well-balanced performance, profile hidden Markov models were generally highly specific and well suited for classifying well-established protein family members. We suggested that different types of protein classifiers should be applied to gain the optimal mining power. Including some of these methods, combinations of multiple protein classification methods were applied to identify especially divergent plant GPCRs (or seven-transmembrane receptors) from the *Arabidopsis thaliana* genome. We identified 394 proteins as the candidates and provided a prioritized list including 54 proteins for

further investigation. For multi-domain protein families, the distribution of urea amidolyase, urea carboxylase, and sterol-sensing domain (SSD) proteins across kingdoms was investigated. Molecular evolutionary analysis showed that the urea amidolyase genes currently found only in fungi among eukaryotes are the results of a horizontal gene transfer event from proteobacteria. Urea carboxylase genes currently found in fungi and other limited organisms were also likely derived from another ancestral gene in bacteria. Finally we showed the possibility of the bacterial origin of the eukaryotic SSD-containing proteins and that these ancestral sequences evolved into four different SSD-containing proteins acquiring specific functions. Two groups of SSD-containing proteins seemed to have been formed before the divergence of fungal and metazoan lineages by domain acquisition.

ACKNOWLEDGEMENTS

Dr. Etsuko Moriyama took me into her lab years ago when I had no experience with bioinformatics, but only an interest. She provided me with financial support and direction to start my career as a graduate student in bioinformatics. She gave me the needed time to figure out a PhD project on my own without imposing any of her own projects or interests. I thank you for your support and guidance during the entire process of my PhD studies.

Dr. Ken Nickerson came up to me with a problem and allowed me to work on it to ultimately make one project and a paper out of it. His positive remarks on my first draft of that paper was very encouraging. When much of success is judged only by the final results, such encouragement along the way is much needed for struggling students to know that what they are doing is infact leading them somewhere. Thank you for giving me the opportunity to work with you.

Dr. Audrey Atkin and Dr. Kathy Hanford both have acted as role models for me to know that family and career for a mother is possible with great organizational and time management skills. Dr. Stephen Scott always has been an easy person to talk to about work, family etc. I thank you all for the discussions and suggestions you provided during our meetings.

I would like to thank the past and present members of the Moriyama lab. It has been a pleasure to work along side all of you and discuss each other's work. I thank SBS and staffs for making the entire process smooth and providing me with financial support whenever I needed. The opportunity I had this year to be a teaching assistant has reminded me how I enjoy teaching.

I want to thank my family in Lincoln, Arnie, Jim, Carrie and Todd, for all their support. Arnie and Carrie allowed me to leave my daughter in their care whenever I needed to work during weekends in this final semester. I also want to thank the staffs at the UNL children's center where they took such good care of my daughter that gave me the peace of mind to focus on my work.

I want to thank my parents Radhaber and Barsha, who have worked hard to make sure their children got the opportunities that they never had, and for their support and patience during my graduate school years. I want to thank my grandmother, Radha Devi, for her love and blessings. My wonderful brothers and their families, Diwas, Rishika, Sameer, Michelle, Shuvalee, Jeman, and a newborn nephew who is yet to be named, have added so much happiness in my life and I hope to see them more often in the future.

My husband Cory is the one person who knows all the details of the rollercoaster that I (and he) went through in our graduate school. His positive attitude, his unwavering support and encouragement has been the guiding light for me to complete my dissertation. Thank you to my daughter Adi for being such a joy in our lives and I am looking forward to spending more time with you this summer.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

Chapter 1

Introduction

1.1 Objectives

The rapidly growing number of sequenced genomes warrants an efficient and dependable way of classifying the protein sequences into functional groups. The distribution of different types of protein sequences in different organisms allows us to study the evolution of protein sequences. This in turn allows us to understand how certain changes in the sequence affected the protein function and how these changes over time affected organismal evolution. To classify new protein sequences, we utilize the information that is already known. Thousands of protein sequences have already been characterized with structure and function. By comparing the features of known protein sequences to those of unknown ones, we can assess the degree of similarity, and by which we can assign potential functional classes to the new proteins. By performing phylogenetic analyses including these newfound proteins, for example, we can infer the evolutionary history of these proteins, when the proteins were formed, and how they have diverged and acquired various functions.

Two broad categories of protein families are used in this study. These are the transmembrane proteins and the multi-domain proteins. Divergent transmembrnae proteins such as the G-protein coupled receptor are difficult to identify, hence serve as excellent examples to study protein classifier performance. The molecular evolutionary study of multi-domain proteins are important because it can show how different domains could have come together to form a larger and more complex protein thereby changing the evolutionary path.

In this study I first analyzed and compared the accuracy of various protein classification methods to classify an extremely diverged family of proteins, the G-protein

coupled receptors (GPCRs). These methods were then utilized to identify putative GPCRs from a model plant *Arabidopsis thaliana*. I also studied the distribution of multi-domain proteins, urea carboxylase and urea amidolyase, in all kingdoms of life and studied its evolutionary history. I examined another set of proteins consisting of sterol-sensing domain in all kingdoms of life to understand its evolution and formation of proteins that possess this domain.

1.2 Transmembrane proteins

Transmembrane proteins make up 20-30% of the total proteins in a genome [1]. They function in detecting and conveying signals from outside into the cell thereby allowing cells to interact and respond to environmental signals [2]. These proteins are the targets for ~60% of the pharmaceuticals used today [3]. The transmembrane domains which embed these proteins into the membrane are predominantly alpha-helices, where each helix is made up of 20-25 hydrophobic amino acids. Analysis of transmembrane proteins in humans by Almen *et al.* [2] resulted in 1,352 receptors, 817 transporters and 533 enzymes. Two thirds of all the human transmembrane receptors were G-protein coupled receptors, a large superfamily of signal transducing proteins having seven transmembane domains. Detailed description of this superfamily is given in the later sections in this chapter. A survey of transmembrane proteins in eukaryotes, eubacteria and archaebacteria showed that these organisms have similar proportions of alpha-helical membrane proteins within their genomes [3]. Various methods have been developed to predict the transmembrane regions in a protein. These include HMMTOP [4], TMHMM [1] and Phobius [5].

1.3 Multi-domain proteins

Domains are functional units of protein sequences that are evolutionarily conserved. Two different families of proteins that serve different functions can share a common domain. Multi-domain proteins make up about 80% of eukaryotic proteins and about 65% of prokaryotic proteins [6]. One of the most important functions of a protein is its ability to interact with other proteins in order to carry out certain functions. These interactions are often carried out by domains, which are units of larger proteins [7]. Therefore any change in an interacting domain can affect the function of the protein, resulting in either loss of function or neofunctionalization. It has been proposed that organismal complexity especially in eukaryotes could be the result of complex domain organizations of proteins. Complex domain organizations allow for the increase in potential interactions between these domains and formation of signal transduction pathways [8]. The creation of new proteins by bringing in different domains is termed as domain shuffling. Kawashima *et al.* [9] identified 1,227 new domain pairs in the vertebrate lineage, among them 137 domain pairs were shared by all seven vertebrate species examined, pointing out that some of these pairs occur in vertebrate specific proteins, thereby linking domain shuffling with the evolution of vertebrates. Databases that store information of protein domains include: Pfam [10] that stores multiple alignments and profile hidden Markov models of families of protein domains, Prosite [11] that stores patterns and profiles that describe conserved protein domains, and SCOP [12] that stores domains based on protein structures. Recently, a domain-domain interaction database DOMINE was created from interactions inferred from the Protein Data Bank and other predicted interactions [13].

In my study, I have used multi-domain proteins such as the urea amidolyase, urea carboxylase, the sterol sensing domain proteins, and nuclear receptors, to study their distribution and molecular evolution. The SSD proteins fall under both categories, they are transmembrane proteins as well as multi-domain proteins. The next section describes these proteins in detail.

1.4 Protein families used in this thesis

1.4.1 G-protein coupled receptors

G-protein coupled receptors (GPCRs) are a superfamily of cell membrane proteins found in a wide range of eukaryotes. They act *e.g*., as light sensing molecules (rhodopsins), as odorant receptors, and as taste receptors [14]. They are characterized by seven hydrophobic transmembrane regions (Figure 1.1). Each GPCR has an extracellular amino terminal (N-terminal) followed by three sets of alternate intracellular and extracellular loops, which connect the seven transmembrane regions, and a final intracellular carboxyl terminal (C-terminal) region [15]. GPCRs are involved in signal transmission from the outside to the interior of the cell through interaction with heterotrimeric G-proteins, or proteins that bind to guanine (G) nucleotides. The receptor is activated when a ligand that carries an environmental signal binds to a part of its cell surface component. A wide range of molecules is used as the ligands including peptide hormones, neurotransmitters, pancrine mediators, ions, proteases, *etc*.

The heterotrimeric G-proteins have three subunits, namely, alpha, beta, and gamma. The G-protein activity is regulated by the alpha subunit, which binds guanine

(G) nucleotides. In an inactive state, the GDP (guanine diphosphate) bound alpha subunit is bound to the beta and gamma subunits. Ligand binding to the extracellular domain of the receptor induces a conformational change in the receptor, which causes the G-proteins to bind to the intracellular domain of the receptor. This stimulates the exchange of the GDP with a GTP (guanine triphosphate) in the binding site of the alpha subunit. The activated GTP-bound alpha subunit then dissociates from the beta and gamma subunits. The beta and gamma subunits remain bound to each other and function as the beta/gamma complex. The beta/gamma complex and the GTP-bound alpha subunit interact with their targets, for example, an enzyme or an ion channel, to transmit the signal. The bound GTP becomes a GDP due to hydrolysis after the transmission of the signal. The GDP-bound alpha subunit reassociates with the beta/gamma complex to form a heterotrimeric G-protein, which is ready for another cycle of transmission of a signal through a GPCR [16].

The GPCRDB, a database system for GPCRs [17], divides the GPCR superfamily into five major classes based on the ligand types, functions, and sequence similarities. The sequences of different GPCR classes are highly diverged from each other, except that they share one common structural feature, that is, they all have seven hydrophobic transmembrane regions. Identifying the function of GPCR sequences is important in biomedical and pharmaceutical research, because GPCRs play key roles in many biologically important functions and are related to many diseases (*e.g.,* neurological cardiovascular diseases, depression, obesity, pain, and viral infections). However, identifying and classifying this membrane protein family is a difficult task due to the high levels of divergence observed among the GPCR family members. GPCRs are used in this

study due to their scientific importance, and also as an example of highly diverged protein families.

1.4.2 Urea degradation enzymes

Urea is degraded into ammonia and carbon dioxide by two distinct enzymes urease and urea amidolyase. Urease breaks down urea in a one-step process while urea amidolyase carries out this reaction in a two-step process as shown below:

[Urea carboxylase] $\text{urea} + \text{ATP} + HCO_3^- \rightarrow \text{allophanate} + \text{ADP} + P_i$ (i)

[Allophanate hydrolase (amidase)] $\text{allophanate} \rightarrow 2NH_3 + 2CO_2$ (ii)

[Urease] $\text{urea} \rightarrow 2NH_3 + CO_2$ (iii)

where (i) and (ii) are carried out by two different domains of the urea amidolyase protein, namely urea carboxylase and amidase.

Urease is a nickel-binding enzyme that has been well-studied in plant, bacteria and fungi and it has been found to be a virulent factor in numerous bacteria and fungi [18]. The bacterial urease protein is a trimer of alpha, beta, and gamma subunits encoded by separate genes forming a gene cluster, whereas in eukaryotes a single gene encodes the urease protein (~800 amino acids), a fused protein representing the three bacterial subunits [19]. Plant and bacterial ureases have also shown anti-fungal properties [20]. This enzyme is of a historical importance as it was the first enzyme to be purified and crystallized [21], and the first enzyme that was shown to require nickel ions [22]. This enzyme is also used in the "rapid urease test" for testing for the presence of *Helicobacter pylori,* which is a bacteria that causes gastrointestinal disorders. A biopsy of the mucosa from the stomach is placed into a medium containing urea and the amount of the

ammonia is tested (raise in pH) to trace the presence of urease, which indicates the presence of the bacterium *H. pylori* [23].

Urea amidolyase is an energy dependent biotin-containing enzyme. It is encoded by the *DUR1,2* gene and was first characterized in the yeast *Candida utilis*, now known as *Pichia jadinii* [24]. The activity of this enzyme has been found in certain species of fungi and green algae, but the sequence itself is present only in fungi and one species of bacteria. This enzyme can be induced in fungal cells by addition of urea or other substances that degrade to urea, while it can be repressed by the lowering the amounts of urea in the medium [24]. Urea amidolyase is a ~1800 amino acid long protein. As shown in Figure 1.2, it consists of the amidase domain (~600aa) (also called allophanate hydrolase) and the urea carboxylase domain (~1200) making it a multi-domain protein. Both of these domains also exist as stand alone proteins. In many bacterial and green algal species, the urea carboxylase gene is in close proximity to the amidase gene, therefore implying that their transcription is regulated together. However, there are also species where these two genes are far apart, or one of them is missing and that leaves a question about its functions.

The urea carboxylase, which is a member of a biotin-dependent carboxylase family of enzymes, is further divided in to smaller domains: the carboxylation domain, the allophanate hydrolase subunit 1, allophanate hydrolase subunit 2, and the biotin lipoyl domains (Figure 1.2). The carboxylation domain and the biotin-lipoyl domain are also common in other biotin carboxylases such as pyruvate carboxylase, acetyl Co-A carboxylase, propionyl Co-A carboxylase, and methylcotonyol Co-A carboxylase. The absence of urea amidolyase and urea carboxylase in many eukaryotic lineages lead us to

study the molecular evolution of these enzymes. Kanamori *et al.* [18] showed that a bacteria *Oleomonas sagaranensis* consists of both pathways for urea utilization and we show in Chapter 4 that several fungal species also consist of both the enzymes.

1.4.3 Sterol-sensing domain proteins

Sterol-sensing domain (SSD) proteins are characterized by the presence of a 180 amino-acids long region called the sterol-sensing domain. This domain forms five hydrophobic membrane spanning helices interconnected with loop regions. The SSD region is believed to sense sterol levels in the cell through direct or indirect interaction with sterols, or other proteins, and is involved in cholesterol homeostasis in cells. This domain has been found to be present in members of six different protein families [25]:

1. 3-hydroxy-3-methylglutaryl coenzyme A-reductase (HMGCR)
2. the sterol regulatory element-binding protein (SREBP)-cleavage activating protein (SCAP)
3. Niemann-Pick disease type C1 (NPC1) protein
4. Patched (Ptc)
5. Dispatched (Disp)
6. Ptc-related (PTR)

Figure 1.3 illustrate these proteins.

HMGCR: The enzyme HMGCR is the rate-limiting enzyme for sterol synthesis and is regulated via negative feedback mechanism. It converts 3-hydroxy-3methylglutaryl-CoA (HMG-CoA) to mevalonic acid. In animals, HMGCR is rapidly degraded when sterol levels are high in the cell. The degradation is mediated by sterol-

induced binding of HMGCR's sterol-sensing domain to insigs, proteins in the endoplasmic reticulum (ER) [26]. Certain drugs such as statins are used to inhibit the function of HMGCR thereby lowering serum cholesterol to reduce the risk of cardiovascular diseases [27]. It is not clear whether HMGCR directly binds cholesterol, but it has been shown that four phenylalanine residues in the SSD is required for the regulated degradation [28]. In yeast, a sterol pathway derivative farnesol causes misfolding of Hmg2p (HMGCR isozyme), and this process requires an intact sterol-sensing domain in Hmg2p [29]. Opposite to animals, the yeast insig homologs, NSG1 and NSG2, inhibit degradation of Hmg2p by direct interaction with the SSD of Hmg2p [30].

SCAP: The SCAP protein acts as a chaperone to transport sterol regulatory element binding protein (SREBP) from ER to the Golgi for further processing. SREBP is a transcription factor for sterol synthesis genes. In mammals, higher cholesterol levels cause SCAP to bind to insigs, therefore causing it to not release SREBPs from the ER resulting in lower sterol synthesis. It has been shown that cholesterol binds to SCAP at an octahelical region, which contains the sterol-sensing domain [31], thereby changing its conformation and making it bound to insig, the ER retention protein. The SSD is required for the ER retention of SCAP, and the degradation of HMGCR in response to higher levels of sterols in the cell [32].

NPC1: NPC1 is a protein that is involved in vesicular trafficking of cholesterol and other lipids. It is one of the two proteins (NPC1 and NPC2) that when mutated can cause Niemann-Pick type C disease where there is accumulation of cholesterol and lipids in cells and neurons. The first evidence that a protein containing SSD region binds to a

cholesterol analog was shown by Ohgami *et al.* [33] where a NPC1 protein was shown to require an SSD region for cholesterol analog to bind. More recently, a binding site for cholesterol and oxysterols have been localized to the first luminal loop of the NPC1 [34]. The exact function of the SSD in NPC1 still remains unknown.

DISP/PTC/PTC-R: The proteins DISP and PTC are key players in the hedgehog (Hh) signaling pathway. The Hh pathway is conserved throughout metazoans and functions in development, patterning, and growth. Alterations in the signaling of this pathway can lead to developmental defects and tumorigenesis [35]. The signaling molecule Hh is covalently linked to cholesterol and is released from signaling cells by the protein DISP while PTC is its receptor in the receiving cells [36]. Once PTC receives the Hh signal, it turns on another protein Smo, thereby turning on a signaling cascade. The role of the SSD regions in DISP and PTC are not clear. Another group of proteins similar to PTC, are called PTC-R, but their functions are not known.

The conservation of the SSD in seven different families and the results shown by mutational studies [29, 33, 37] indicate the functional importance of this domain in cellular activities. In my study, I searched in eukaryotic and prokaryotic genomes to find proteins that contain SSD sequences.

1.4.4 Nuclear receptors

Nuclear Receptors (NRs), a multi-domain protein family of ligand activated transcription factors, play a key role in the process of development, metabolism and reproduction of the cell. In their inactive state, NRs reside in either the nucleus or the cytoplasm. Activation occurs when a ligand binds at the ligand-binding domain (LBD) of

the NR. This in turn causes the NR to bind to response elements (promoters) of their target genes via DNA-binding domain (DBD). Some NRs like the thyroid receptors are always bound to the DNA and are activated by ligand binding. The effect of this reaction is the regulation of the expression of the target genes.

NRs share a common organizational structure as shown in Figure 1.4: the N-terminal region (A/B domain) that is highly variable and consists of a transactivation region AF-1, the DBD (C domain) that is highly conserved and is also involved in the dimerization of NRs, the less conserved flexible hinge (D domain), the moderately conserved LBD (E domain), the extremely variable and sometimes absent F domain [38].

Depending upon the DBD and LBD, NRs are divided into six subfamilies as follows:

1. Thyroid hormone
2. Hepatocyte nuclear factor 4-gamma
3. Estrogen
4. Nerve growth factor 1B
5. Fushi tarazu-F1
6. Germ cell nuclear factor

In addition to these, there are two more subfamilies: 1) Knirps (NRs with no LBD) and 2) DAX (NRs with no DBD). Many of the annotated NRs do not have a known ligand and hence are called orphan nuclear receptors. It is likely that the ancestral protein of NRs was an orphan receptor and ligand binding was an acquired property of these proteins [39].

Natural activation of NRs typically occurs by the binding of lipophilic molecules (ligands), such as steroid hormones, bile acids, fatty acids, thyroid hormones, certain vitamins and prostaglandins [39]. Many orphan NRs have also been found to be activated by synthetic ligands. NRs are also responsible in diseases such as cancer, diabetes, and asthma [40]. Their potential to be regulated by exogenous compounds makes them an extremely important drug target in human disease [41].

NRs have been found in diverse metazoans but have been absent in plants and fungi [39]. Most likely, NRs in these kingdoms either are so diverged that current methods fail to find them, or these organisms may have a different kind of protein that do the same function. This hypothesis lead us to explore these genomes in search of proteins that are either NRs or a novel family of proteins that has some similarities with the LBD and DBD of known NRs.

1.5 Protein classification methods

Various types of classification methods exist for sequence classification. They can be grouped into three categories as below. Methods from each of the categories were used in the study.

1.5.1 Pairwise sequence comparison methods

One of the common sequence comparison methods, Basic Local Alignment Search Tool (BLAST) [42], has been used extensively in finding sequence similarity. It finds segments of the query sequence that match to segments of sequences in a database. It then extends these ‘seeds’ to find longer alignments. BLAST scores the alignments and

then ranks its results based on e-values, which is measure of the reliability of the score. The e-value of a database match is the number of times that one would find an alignment that has the equal or greater score than the given alignment by randomly matching any two sequences. It is dependent on the score of the alignment, the sequence database length and the query length. Similar to probabilities, e-values closer to 0 mean that such alignments cannot happen simply by chance. The results of BLAST must be carefully interpreted, however, as some results can be misleading especially when the entire sequences of multi-domain proteins are used for searches. For example, given a query protein X that has both domains A and B, when a BLAST search is done to identify proteins with a function defined by the domain A, proteins that do not have a domain A but another domain B often will show low e-values (high scores). This can introduce false positives in the search for proteins with domain A sequences.

Another local similarity method, SSEARCH [43] , uses the Smith-Waterman (SW) pairwise alignment, which uses the dynamic programming algorithm to find the optimum local alignment. This method is computationally expensive. Although SSEARCH is more sensitive than BLAST, it still produces only relatively close hits. Both BLAST and SSEARCH are often useful as the first step in a classification problem. BLAST has been used to search for G-protein-coupled receptors (GPCRs) from the genome of *Magnaporthe grisea* and a novel family of GPCR-like proteins was found [44]. These pairwise sequence comparison methods are very specific and are not sensitive enough when trying to find new proteins whose sequences have diverged significantly from the known sequence of a family but whose structure and function have retained similarity. For those sequences that have not diverged extremely, however, these methods

can efficiently identify them. BLAST is now part of many sequence databases such as its original site National Center for Biotechnology Information [45], Universal Protein Resource [46], Fungal Genome Initiative [47], and Joint Genome Institute [48].

1.5.2 Generative methods

A new era of protein sequence classification arose after the introduction of generative methods. These methods are based on multiple sequence alignments, and include methods such as PSI-BLAST [49] and profile hidden Markov models [50, 51]. A set of sequences from a family of interest is used in building the profile that represents the family. Profiles contain the position-specific amino acid information from the multiple alignment of a family of sequences. New sequences are aligned to this profile and the results are ranked based on the score calculated by the method. Higher scoring sequences can be thought of as being generated by this profile. These methods are more sensitive than pairwise alignment methods because the profile is made from a set of sequences, making it more general than methods using pairwise alignments based on a single sequence query. While pairwise alignment uses position-independent scoring parameters (*e.g.*, BLOSUM scoring matrices), profiles use position-specific parameters for amino acid substitutions (*e.g.*, position-specific scoring matrix or PSSM) and gap penalties. This property of profiles is important when certain regions of the protein are more conserved than other, and when certain regions can acquire more insertions or deletions than others. Generative methods have been shown to perform better than the pairwise sequence similarity methods in finding remote homology [52, 53].

Profile hidden Markov models (HMMs) [51] have been used widely in the

classification of protein sequences. In biological sequence analysis, profile HMMs are built based on a multiple alignment as shown in Figure 1.5. In general, the multiple alignments are generated from a training set consisting of positive examples of protein sequences that belong to a certain functional family sharing a level of sequence similarities. Given a multiple alignment of protein sequences, “match”, “insert”, and “delete” states are first identified. If a column of the multiple alignment has less than or equal to fifty percent gaps (*i.e.*, a half or more of the sequences emit an amino acid), then it is classified as a “match column” (columns 1-3 and 6-10 in Figure 1.5). A non-gap entry in a match column is a “match state” in the HMM, while a gap in a match column is a “delete state”. Delete states are presumed to be modifications that stem from an amino acid sequence losing one or more amino acids in an evolutionary event. The last type of state is the “insert” state. “Insert columns” (columns 4 and 5 in Figure 1.5) are similar to delete states, except that the evolutionary modification to the amino acid sequence is that of gaining amino acids. A non-gap in an insert column is an “insert state”, while a gap in an insert column is ignored since it does not represent an event of evolutionary significance. As shown in Figure 1.6, a profile HMM, which can be visualized as a finite state machine, has a start and an end state in addition to the previously identified match, insert, and delete states. Each of these states has position-specific transition probabilities for transitioning into each of these states from the previous state (represented by arrows in Figure 1.6). Match states have position-specific emission probabilities for each of the 20 amino acids. Insert states also have position-specific emission probabilities for inserting each of the 20 amino acids at that state. When no residue is associated with a node, it is a delete state, and no emission probability is associated with it.

To obtain the probability that a new sequence belongs to the family of the model, the new sequence is compared to the profile HMM by aligning it to the model. The most probable path taken to generate the sequence similar to the new sequence gives the similarity score. It is calculated by multiplying the emission and transition probabilities along the path. The most likely path through the model is computed with the *Viterbi* algorithm or the *forward* algorithm [54]. One could also generate the most probable sequence obtained from a particular HMM by summing over all possible paths and choosing the path with the maximum score. In both ways, the most probable path can be efficiently and optimally calculated. Two of the most common programs based on profile HMMs are SAM [50] and HMMER [51].

Profiles and profile HMMs can be created using either the entire protein sequences or only domains or motifs conserved between proteins belonging to the same family. Examples of databases of multiple alignments and profiles/profile HMMs from protein families and domains include PROSITE [11], Pfam [10], PANTHER [55], SMART [56] and Superfamily [52]. Certain domains belonging to the member proteins of a family are functionally constrained, causing these domains to be more conserved than other parts of the sequence. In this case, domain-specific profiles work better than entire sequence profiles in finding remote homology.

Wistrand *et al.* [57] developed a new GPCR detection method, GPCRHMM. It incorporates GPCR-specific TM features (*e.g.*, loop-region lengths, different amino acid composition among loop and TM regions) in a hidden Markov model architecture. With their method they were able to predict 120 novel GPCRs in various genomes including mouse and human.

One problem that arises from generative methods is that reliable multiple alignments cannot be created from protein sequences of a family whose members are highly diverged, such as the nuclear receptors and the G-protein coupled receptors. Another problem with these methods is that only positive sequences are used in building the models, since negative sequence information cannot be incorporated in building the alignments or profiles. Nonetheless, profile HMMs work well with not too extremely diverged proteins and have been used widely in protein classification.

The sequence similarity methods and the generative methods rank their scores based on e-values. In order to be able to compare the e-values from one database search to another using the same method, the "effective database length" needs to be kept constant. This is because the database lengths are used in calculating e-values. For example, one can use the database length of the NCBI nr database, which is currently more than 2.5×10^9 characters, as this parameter so that the e-values from the NCBI nr searches can be comparable to those from blast database searches using smaller databases whose lengths are significantly smaller in the range of only 2.5×10^6 characters (*e.g.*, against a single genome)

1.5.3 Discriminative methods

Discriminative methods are powerful in that they do not have to depend on sequence alignments. Added robustness comes as they are able to incorporate both positive and negative data. These methods are trained on positive sequence information as well as negative sequence information. Once trained, the methods can then

discriminate the test set into positive and negative sequences by using a threshold score. Discriminative methods have been shown to be very sensitive*, i.e.*, able to find distantly related sequences [58]. One popular discriminative method used today is the support vector machine (SVM).

A support vector machine (SVM) is a learning machine that makes a binary classification based on a separating hyperplane on a remapped instance space [59]. The goal of the classification is to remap the input vectors onto a multi-dimensional space so that the instances are linearly separable. SVMs learn from labeled examples from a training set including both positive and negative samples. Depending upon a set of attributes, SVMs find a hyperplane that classifies the positive and negative data in the training set (Figure 1.7). The hyperplane is optimized in such a way that the distance called the *margin*, between the hyperplane and the closest training example, is maximized. The data points nearest to the margin on both sides are called *support vectors*, marked with 'v' in Figure 1.7. We assume that there is a mapping or target function between the data and their labels the machine will learn [60]. A kernel function, which is a dot product that is used in remapping input feature vectors, is used to find the hyperplane. Once the hyperplane is found, unlabeled examples from the test set can be classified as shown in Figure 1.7. Classification can be done solely based upon the support vectors found. Some commonly used kernel function includes: linear, polynomial, radial basis, and sigmoid functions.

Many types of input can be used with the SVMs, *e.g.,* 20 amino acid composition, 400 dipeptide composition, and physico-chemical properties of the protein sequences. These properties represent the protein sequence where important regions have properties

that are conserved among functionally similar sequences. Matsuda *et al*. [61] have used localized amino acid compositions (N-terminal, middle, and C-terminal) and the local frequencies of distance between successive basic, hydrophobic, and other amino acids for cellular localization prediction, yielding 87 percent or higher accuracy. Park *et al*. [62] have used amino acid composition and dipeptide composition for classification of outer membrane proteins using SVMs, resulting in 94 percent accuracy. Bhasin and Raghava [63] have also used similar methods for classifying the subfamilies of NRs and achieved 97.5 percent accuracy by using the SVM with only dipeptide composition. Lin *et al*. [64] also used SVMs with amino acid compositions, physico-chemical properties (hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility) to classify lipid binding proteins into functional classes with high accuracies.

A combination of profile HMM and SVM was introduced by Jaakkola *et al*. [65], and an SVM using pairwise sequence similarity scores was developed by Liao and Noble [66]. Both these methods have performed well in their studies. Recently developed classification methods based on domain regions by Sadka and Linial [67] have used transmembrane (TM) domain regions of many TM proteins to build Gaussian profiles using 20 amino-acid composition, and then used SVMs to classify each family of TM proteins. Their method is based on the idea that the information encoded at the TM domains is enough to classify the protein into a functional family. Their method gave good results in classifying polytopic proteins with 80 percent sensitivity and 90 percent specificity. Another domain-based method was introduced by Chou and Cai [68] where a protein sequence was represented as a 2005-dimensional binary vector, representing 2005

functional domains from domain database SBASE-A [69], with 0s for absence and 1s for presence of the domain. Then SVM was applied to discriminate between the positive and negative sequences resulting in high success rates. This 'functional domain composition' method using SVMs, and additional nearest neighbor algorithm was used in the prediction of the functional class of yeast proteins [70].

1.6 Organization of the dissertation

This dissertation is divided into the following chapters. In Chapter 1, this chapter, I presented the objectives of this dissertation, a brief description on transmembrane and multi-domain proteins, and background on protein families and classification methods I have used.

Chapter 2 describes the comparative study of various classification methods. Alignment-based classifiers (*e.g.,* profile HMM, support vector machines with Fisher score and with pairwise alignment scores) are compared against alignment-free classifiers (*e.g.,* support vector machines and decision trees with amino acid composition) using extremely divergent G-proteins coupled receptors as an example. This chapter has been published in:

Strope, P. K. and Moriyama, E. N. (2007) Simple alignment-free methods for protein classification: a case study from G-protein coupled receptors. *Genomics* **89:** 602-612.

Chapter 3 involves the application of the methods I studied in mining the putative G-protein coupled receptors (also called seven transmembrane receptors) in the *Arabidopsis thaliana* genome. I was involved in training data preparation and prediction

of candidate GPCRs using profile hidden Markov models, support vector machines with amino acid composition and dipeptide composition. This chapter has been published in: Moriyama, E. N., Strope, P. K., Opiyo, S. O., Chen, Z. and Jones, A. M. (2006) Mining the *Arabidopsis thaliana* genome for highly-divergent seven transmembrane receptors. *Genome Biology* **7:** R96.

In Chapter 4, I studied the molecular evolution of related multi-domain protein families: urea amidolyase and urea carboxylase in both eukaryotes and prokaryoties. I presented the possible horizontal transfer scenario of urea amidolyase from bacteria to fungi. This study has been published in:
Strope, P. K., Nickerson, K. W., Harris, S. D. and Moriyama, E. N. (2011) Molecular evolution of urea amidolyase and urea carboxylase in fungi. *BMC Evolutionary Biology* **11**: 80.

Chapter 5 reports the study of sterol-sensing domain (SSD) proteins in eukaryotes. I thoroughly searched for SSD proteins in bacteria and eukaryotes, and performed phylogenetic analyses to understand their evolutionary history. The result from this study is in preparation for submission to the journal *Genome Biology and Evolution*.

Chapter 6 describes the conclusion of my study and future directions. In the Appendix, a study of Nuclear Receptors is also described with some preliminary results.

1.7 References:

1. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**(3):567-580.
2. Almen MS, Nordstrom KJ, Fredriksson R, Schioth HB: **Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin**. *BMC Biol* 2009, **7**:50.
3. Stevens TJ, Arkin IT: **Do more complex organisms have a greater proportion of membrane proteins in their genomes?** *Proteins* 2000, **39**(4):417-420.
4. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server**. *Bioinformatics* 2001, **17**(9):849-850.
5. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method**. *J Mol Biol* 2004, **338**(5):1027-1036.
6. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. *J Mol Biol* 2001, **310**(2):311-325.
7. Deng M, Mehta S, Sun F, Chen T: **Inferring domain-domain interactions from protein-protein interactions**. *Genome Res* 2002, **12**(10):1540-1548.
8. Basu MK, Carmel L, Rogozin IB, Koonin EV: **Evolution of protein domain promiscuity in eukaryotes**. *Genome Res* 2008, **18**(3):449-461.
9. Kawashima T, Kawashima S, Tanaka C, Murai M, Yoneda M, Putnam NH, Rokhsar DS, Kanehisa M, Satoh N, Wada H: **Domain shuffling and the evolution of vertebrates**. *Genome Res* 2009, **19**(8):1393-1403.

10. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R *et al*: **Pfam: clans, web tools and services**. *Nucleic Acids Res* 2006, **34**(Database issue):D247-251.

11. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database**. *Nucleic Acids Res* 2006, **34**(Database issue):D227-230.

12. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data**. *Nucleic Acids Res* 2004, **32**(Database issue):D226-229.

13. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R: **DOMINE: a comprehensive collection of known and predicted domain-domain interactions**. *Nucleic Acids Res*, **39**(Database issue):D730-735.

14. Foreman JC, Johansen T: **Textbook of receptor pharmacology**, 2nd edn: CRC Press; 2003.

15. Watson S, Arkinstall S: **The G-Protein linked receptor FactsBook**: Academic Press; 1994.

16. Cooper GM: **The Cell: A Molecular Approach**, 2nd edn: ASM Press; 2000.

17. Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G: **GPCRDB: an information system for G protein-coupled receptors**. *Nucleic Acids Res* 1998, **26**(1):275-279.

18. Kanamori T, Kanou N, Atomi H, Imanaka T: **Enzymatic characterization of a prokaryotic urea carboxylase**. *J Bacteriol* 2004, **186**(9):2532-2539.

19. Carter EL, Flugga N, Boer JL, Mulrooney SB, Hausinger RP: **Interplay of metal ions and urease**. *Metallomics* 2009, **1**(3):207-221.

20. Becker-Ritt AB, Martinelli AH, Mitidieri S, Feder V, Wassermann GE, Santi L, Vainstein MH, Oliveira JT, Fiuza LM, Pasquali G *et al*: **Antifungal activity of plant and bacterial ureases**. *Toxicon* 2007, **50**(7):971-983.

21. Sumner JB: **The isolation and crystallization of the enzyme urease**. *The journal of biological chemistry* 1926, **69**:435-441.

22. Dixon NE, Gazzola TC, blakeley RL, Zermer B: **Letter: Jack bean urease (EC 3.5.1.5). A metalloenzyme. A simple biological role for nickel?** *J Am Chem Soc* 1975, **97**(14):4131-4133.

23. Mobley HL, Hu LT, Foxal PA: **Helicobacter pylori urease: properties and role in pathogenesis**. *Scand J Gastroenterol Suppl* 1991, **187**:39-46.

24. Roon RJ, Levenberg B: **Urea amidolyase. I. Properties of the enzyme from Candida utilis**. *J Biol Chem* 1972, **247**(13):4107-4113.

25. Kuwabara PE, Labouesse M: **The sterol-sensing domain: multiple families, a unique role?** *Trends Genet* 2002, **18**(4):193-201.

26. Sever N, Yang T, Brown MS, Goldstein JL, DeBose-Boyd RA: **Accelerated degradation of HMG CoA reductase mediated by binding of insig-1 to its sterol-sensing domain**. *Mol Cell* 2003, **11**(1):25-33.

27. Istvan ES, Deisenhofer J: **Structural mechanism for statin inhibition of HMG-CoA reductase**. *Science* 2001, **292**(5519):1160-1164.

28. Chang TY, Chang CC, Ohgami N, Yamauchi Y: **Cholesterol sensing, trafficking, and esterification**. *Annu Rev Cell Dev Biol* 2006, **22**:129-157.

29. Shearer AG, Hampton RY: **Lipid-mediated, reversible misfolding of a sterol-sensing domain protein**. *EMBO J* 2005, **24**(1):149-159.

30. Flury I, Garza R, Shearer A, Rosen J, Cronin S, Hampton RY: **INSIG: a broadly conserved transmembrane chaperone for sterol-sensing domain proteins**. *EMBO J* 2005, **24**(22):3917-3926.

31. Radhakrishnan A, Sun LP, Kwon HJ, Brown MS, Goldstein JL: **Direct binding of cholesterol to the purified membrane region of SCAP: mechanism for a sterol-sensing domain**. *Mol Cell* 2004, **15**(2):259-268.

32. Song BL, Javitt NB, DeBose-Boyd RA: **Insig-mediated degradation of HMG CoA reductase stimulated by lanosterol, an intermediate in the synthesis of cholesterol**. *Cell Metab* 2005, **1**(3):179-189.

33. Ohgami N, Ko DC, Thomas M, Scott MP, Chang CC, Chang TY: **Binding between the Niemann-Pick C1 protein and a photoactivatable cholesterol analog requires a functional sterol-sensing domain**. *Proc Natl Acad Sci U S A* 2004, **101**(34):12473-12478.

34. Infante RE, Radhakrishnan A, Abi-Mosleh L, Kinch LN, Wang ML, Grishin NV, Goldstein JL, Brown MS: **Purified NPC1 protein: II. Localization of sterol binding to a 240-amino acid soluble luminal loop**. *J Biol Chem* 2008, **283**(2):1064-1075.

35. Eaton S: **Multiple roles for lipids in the Hedgehog signalling pathway**. *Nat Rev Mol Cell Biol* 2008, **9**(6):437-445.

36. Gallet A: **Hedgehog morphogen: from secretion to reception**. *Trends Cell Biol*.

37. Yabe D, Xia ZP, Adams CM, Rawson RB: **Three mutations in sterol-sensing domain of SCAP block interaction with insig and render SREBP cleavage insensitive to sterols**. *Proc Natl Acad Sci U S A* 2002, **99**(26):16672-16677.

38. Escriva Garcia H, Laudet V, Robinson-Rechavi M: **Nuclear receptors are markers of animal genome evolution**. *J Struct Funct Genomics* 2003, **3**(1-4):177-184.

39. **Essays in Biochemistry**, vol. 40: Portland Press; 2004.

40. Novac N, Heinzel T: **Nuclear receptors: overview and classification**. *Curr Drug Targets Inflamm Allergy* 2004, **3**(4):335-346.

41. Maglich JM, Watson J, McMillen PJ, Goodwin B, Willson TM, Moore JT: **The nuclear receptor CAR is a regulator of thyroid hormone metabolism during caloric restriction**. *J Biol Chem* 2004, **279**(19):19832-19838.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

43. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**(1):195-197.

44. Kulkarni RD, Thon MR, Pan H, Dean RA: **Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus Magnaporthe grisea**. *Genome Biol* 2005, **6**(3):R24.

45. [http://www.ncbi.nlm.nih.gov/]

46. [http://www.uniprot.org/]

47. [http://www.broadinstitute.org/science/projects/fungal-genome-initiative/fungal-genome-initiative]

48. [http://www.jgi.doe.gov]

49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

50. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method**. *Comput Appl Biosci* 1996, **12**(2):95-107.

51. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

52. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**(4):903-919.

53. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines**. *Bioinformatics* 2002, **18**(1):147-159.

54. Durbin R, Eddy SR, Krogh A, Mitchinson G: **Biological sequence analysis: Probabilistic models of proteins and nucleic acids**. Cambridge: Cambridge University Press; 1998.

55. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ *et al*: **The PANTHER database of protein families, subfamilies, functions and pathways**. *Nucleic Acids Res* 2005, **33**(Database issue):D284-288.

56. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32**(Database issue):D142-144.

57. Wistrand M, Kall L, Sonnhammer EL: **A general model of G protein-coupled receptor sequences and its application to detect remote homologs**. *Protein Sci* 2006, **15**(3):509-521.

58. Strope PK, Moriyama EN: **Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors**. *Genomics* 2007, **89**(5):602-612.

59. Christianini N, Shawe-Taylor J: **An Introduction to Support Vector Machines**, Ist edn: Cambridge University Press; 2000.

60. Karchin R: **Classifying G-protein Coupled Receptors with Support VectorMachines**. Santa Cruz: University of California; 2000.

61. Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T: **A novel representation of protein sequences for prediction of subcellular location using support vector machines**. *Protein Sci* 2005, **14**(11):2804-2813.

62. Park KJ, Gromiha MM, Horton P, Suwa M: **Discrimination of outer membrane proteins using support vector machines**. *Bioinformatics* 2005, **21**(23):4223-4229.

63. Bhasin M, Raghava GP: **Classification of nuclear receptors based on amino acid composition and dipeptide composition**. *J Biol Chem* 2004, **279**(22):23262-23266.

64. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ: **Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity**. *J Lipid Res* 2006, **47**(4):824-831.

65. Jaakkola T, Diekhans M, Haussler D: **A discriminative framework for detecting remote protein homologies**. *J Comput Biol* 2000, **7**(1-2):95-114.

66. Liao L, Noble WS: **Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships**. *J Comput Biol* 2003, **10**(6):857-868.

67. Sadka T, Linial M: **Families of membranous proteins can be characterized by the amino acid composition of their transmembrane domains**. *Bioinformatics* 2005, **21 Suppl 1**:i378-386.

68. Chou KC, Cai YD: **Using functional domain composition and support vector machines for prediction of protein subcellular location**. *J Biol Chem* 2002, **277**(48):45765-45769.

69. Vlahovicek K, Murvai J, Barta E, Pongor S: **The SBASE protein domain library, release 9.0: an online resource for protein domain identification**. *Nucleic Acids Res* 2002, **30**(1):273-275.

70. Cai YD, Doig AJ: **Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition**. *Bioinformatics* 2004, **20**(8):1292-1300.

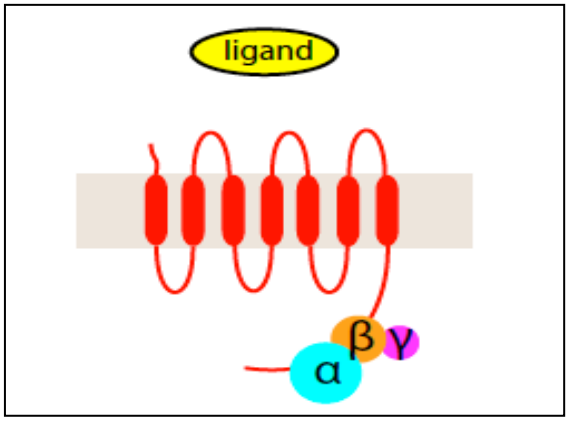

Figure 1.1. A model of G-protein coupled receptor protein. Seven transmembrane regions are shown. A ligand is present in the extracellular space and G-proteins (α, β, and γ) are present inside of the cell.

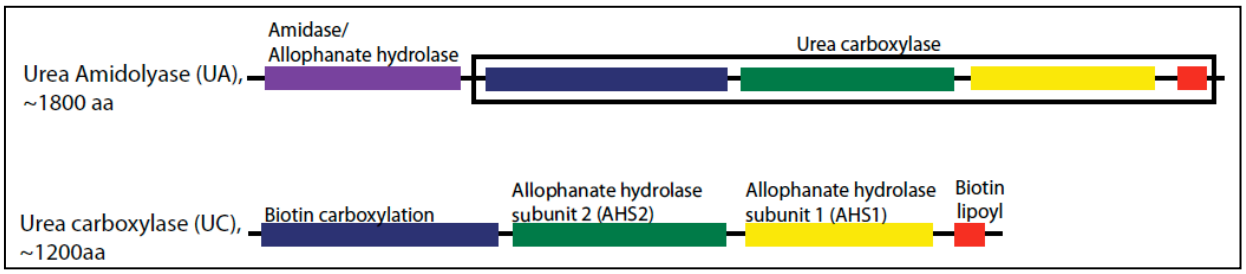

Figure 1.2. Domain structures of urea amidolyase and urea carboxylase. The abbreviations and approximate amino-acid lengths are given with the protein names. Amidase and urea carboxylase sequences exist as domains within the urea amidolyase protein or as single proteins by themselves.

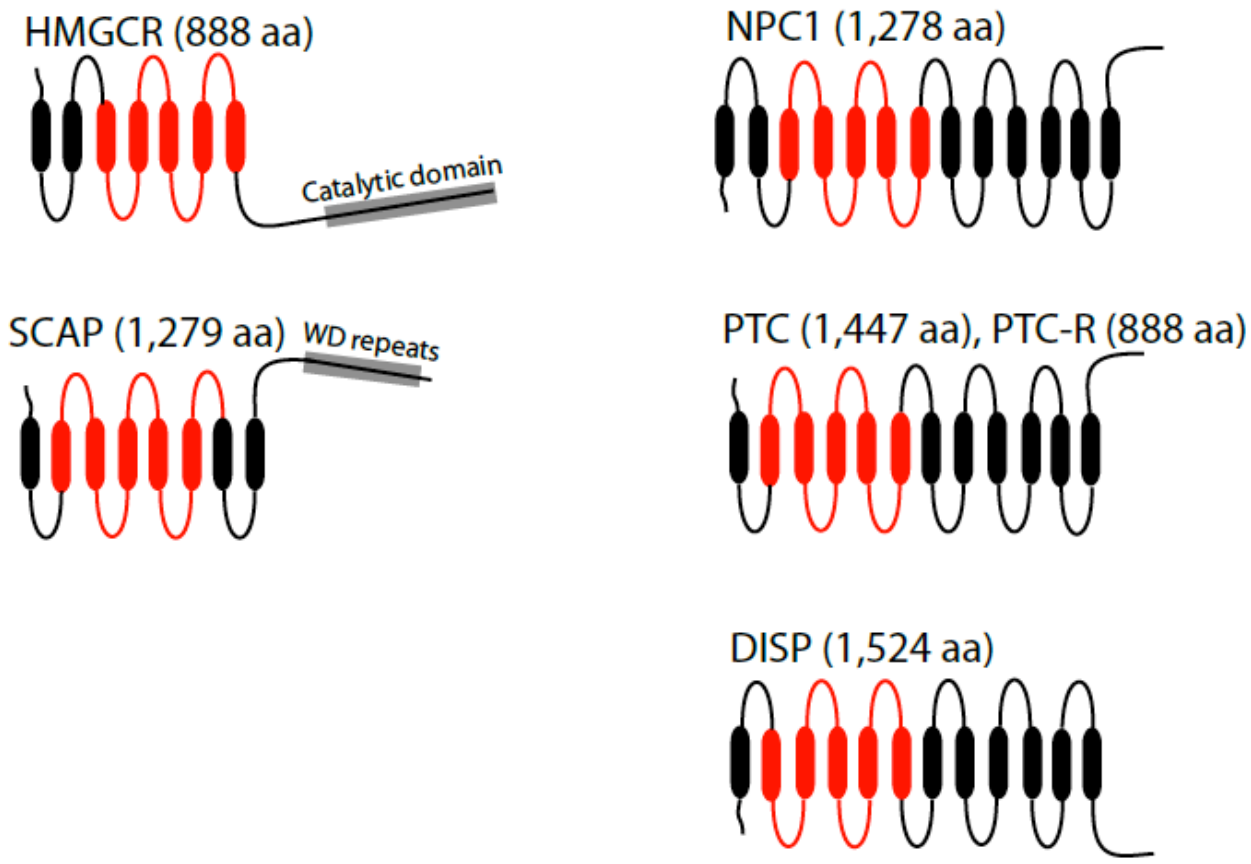

Figure 1.3. Topology of the SSD proteins. The lengths and topology of the proteins shown are based on the human SSD proteins. The cylindrical structures are the transmembrane regions. The SSD regions are indicated in red. The top side of each protein is cytoplasmic. Enzyme names are as follows. HMGCR: 3-hydroxy-3-methylglutaryl-coenzyme A reductase, SCAP: Sterol regulatory element binding protein clevage activating protein, NPC1: Niemann-Pick type C1 protein, PTC: Patched protein, PTC-R: Patched related protein, and DISP: Dispatched protein.

Figure 1.4. Organization of a typical nuclear receptor (Taken from Escriva Garcia *et al.* 2003).

	1	2	3	4	5	6	7	8	9	10	
. . .	V	G	A	-	-	H	A	G	E	Y	. . .
. . .	V	-	-	-	-	N	V	D	E	V	. . .
. . .	V	E	A	-	-	D	V	A	G	H	. . .
. . .	V	K	G	-	-	-	-	-	-	D	. . .
. . .	V	Y	S	-	-	T	Y	E	T	S	. . .
. . .	F	N	A	-	-	N	I	P	K	H	. . .
. . .	I	A	G	A	D	N	G	A	G	V	. . .

Figure 1.5: An example multiple alignment to create a profile hidden Markov model. A gap is represented by a '-'. Columns 1-3 and 6-10 are "match" columns, while the columns 4 and 5 are "insert" columns.

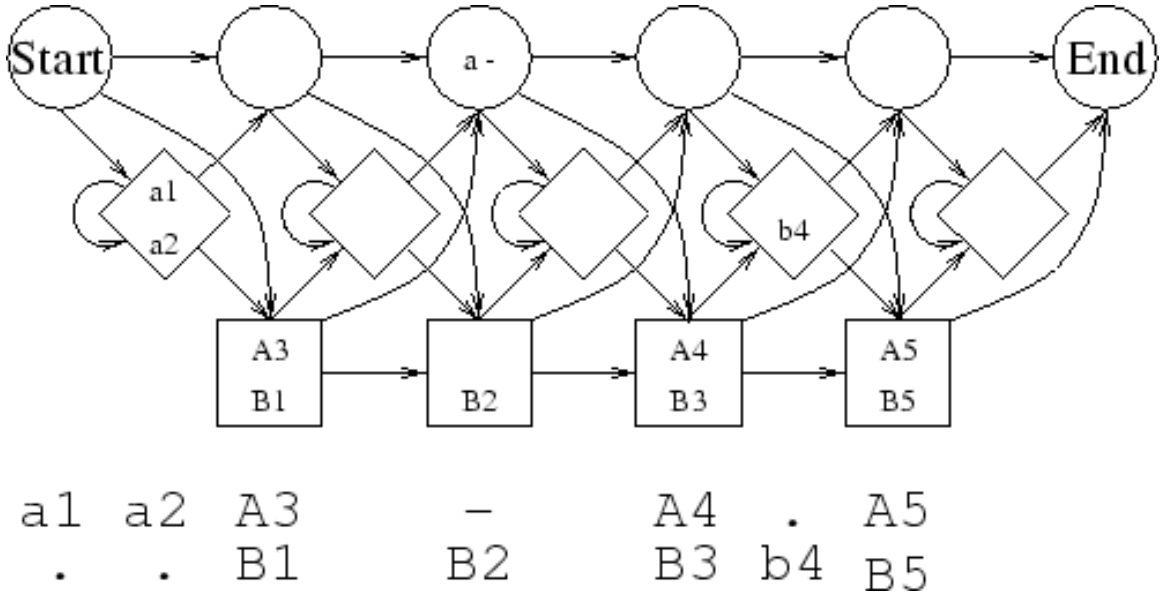

Figure 1.6: A profile hidden Markov model with delete (circle), insert (diamond), and match (square) states (taken from Hughey and Krogh, 1996). Transitions are allowed along each arrow. Delete and match states can only be visited once for each position along a path. Delete states do not emit any symbols. Insert states are allowed to insert multiple symbols. The alignment at the bottom is used to build the model in this example. The sequences begin in the start state. Amino acids a1 and a2 are inserted at the beginning of the sequence. A3 and B1 are the first matched symbols, followed by a deletion, where B2 is matched with a gap. A4 is then matched with B3, b4 is inserted, A5 is matched with B5, and finally the end state is reached.

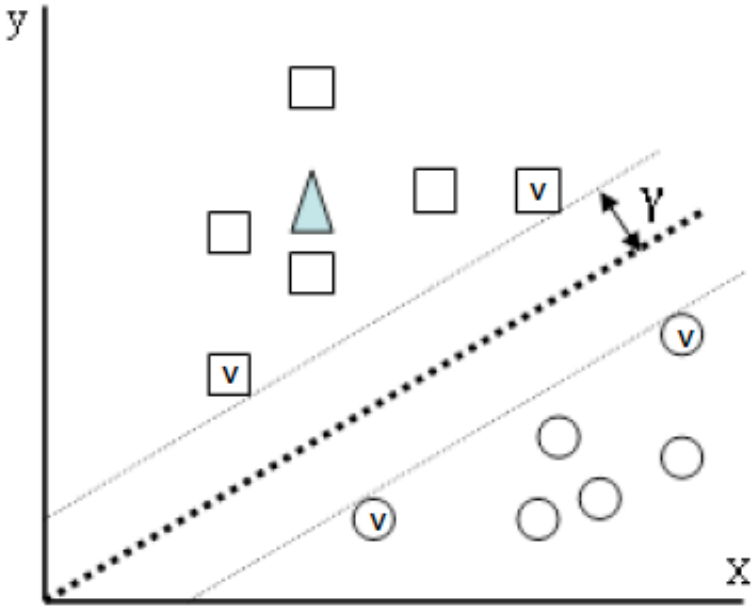

Figure 1.7: A hyperplane classifying two classes of data. A new sample of an unknown class can be classified based on the hyperplane. In this figure, the training data have two dimensions, represented by the x and y axes. Two classes of data are represented by squares and circles. The hyperplane that is calculated from these training examples is given by the bold dotted line, separated from the closest training vectors (support vectors marked with 'v') by the distance. The classification of an unknown sample (triangle) is done by determining which side of the hyperplane the new instance falls. In this example, the prediction for the unknown sample would be square.

Chapter 2

Simple Alignment-free Methods for Protein Classification: A Case Study from G-Protein Coupled Receptors

2.0 Preface for Chapter 2

Computational methods of predicting protein functions rely on detecting similarities among proteins. However, sufficient sequence information is not always available for some protein families. For example, proteins of interest may be new members of a divergent protein family. The performance of protein classification methods could vary in such challenging situations. This chapter describes the comparative study of various classification methods using an extrememly divergent superfamily of transmembrane proteins, G-proteins coupled receptors, as an example. Alignment-based classifiers (*e.g.,* profile HMM, support vector machines with Fisher scores and with pairwise alignment scores) are compared against alignment-free classifiers (*e.g.,* support vector machines and decision trees with amino acid composition). Alignment-free classifiers based on support vector machines using simple amino acid compositions were effective in remote-similarity detection even from short fragmented sequences. Although it is computationally expensive, a support vector machine classifier using local pairwise alignment scores showed very good balanced performance. More commonly used profile hidden Markov models were generally highly specific and well suited to classifying well-established protein family members. From these results, we suggested that different types of protein classifiers should be applied to gain the optimal mining power. This chapter has been published in:

Strope, P. K. and Moriyama, E. N. (2007) Simple alignment-free methods for protein classification: a case study from G-protein coupled receptors. *Genomics* **89:** 602-612.

2.1 Background

Predicting functions of new protein candidates is an essential part of post-genomic processing. Many effective protein classification methods have been developed for this purpose. Routinely applied methods include Pfam [1], SMART [2], Superfamily [3], PANTHER [4], PRINTS [5], and PROSITE [6]. InterPro [7] provides an integrated interface for various methods. These methods rely on multiple alignments to compare sequences and to build various forms of models. However, generating reliable multiple alignments becomes increasingly difficult when more divergent protein sequences are to be incorporated. Another disadvantage shared by these multiple alignment-based methods is that their models are built only from "positive samples" (protein sequences of interests), and information from "negative samples" (unrelated protein sequences) is not directly incorporated. Since subsequently found proteins are classified based on these models, possible initial sampling bias is kept and possibly reinforced.

Recent developments in protein classification methods addressed the above-mentioned problems. Kim *et al.* [8] and Moriyama and Kim [9] developed classification methods based on discriminant function analyses incorporating amino acid composition and physico-chemical properties in the descriptors. Their discriminant analysis methods were effective in discriminating G-protein coupled receptors (GPCRs) from non-GPCRs especially when only partial sequences were available. Support vector machines (SVMs) were used in other studies. Karchin *et al.* [10] used an SVM with a kernel function built on profile hidden Markov models (HMMs). Their results showed that their method, SVM_Fisher, could classify GPCR subfamilies within the superfamily better than a profile HMM method. SVM_pairwise developed by Liao and Noble [11] used pairwise similarity scores as input vectors. It performed better than other methods (*e.g.,* profile HMM and SVM_Fisher) for

discriminating SCOP protein families [12]. More recently, SVM classifiers were applied for GPCR family classification based on amino acid composition and dipeptide frequencies by Bhasin and Raghava [13] and Wang *et al.* [14]. Decision tree and naïve Bayes classifiers with n-gram (n-mer or n-residue string) frequencies were also used for GPCR subfamily classification by Cheng *et al.* [15; includes also extensive list of protein classifiers]. Another alignment-free descriptors, auto/crosscovariance vectors based on amino acid properties, were used with partial least squares regression [16; 17] and with self-organizing maps (SOMs, an artificial neural network) [18]. These methods (except for SOMs) are discriminative; they generate models based on both positive and negative samples. Remote similarity detection has also been studied in relation to protein structure prediction, since incorporation of structural information could improve the identification sensitivity [reviewed by e.g., 19; 20].

One example showing the power of *alignment-free* classifiers was in the discovery of odorant receptor (OR) genes, a divergent member group of GPCRs, from the *Drosophila melanogaster* genome. Although OR protein sequences were previously known in vertebrates, due to their extremely low similarities with vertebrate counterparts, *Drosophila* ORs could not be identified until Kim *et al*. [8] applied their *alignment-free* discriminant analysis method. Sixty-one Drosophila OR as well as gustatory receptors were then newly identified [21; 22]. We should also note that *alignment-free* methods do not require us to assume homologous relationship (common ancestry) among similar sequences. Descriptors are in general designed to extract sequence properties shared among functionally similar proteins regardless of their evolutionary relationships.

The main purpose of this study is to compare the performance among *alignment-based* and *alignment-free* protein classification methods and to identify their strengths and

weakness from the practical perspectives of the users. Using the GPCR superfamily and taking advantage of their extreme and various levels of divergence, we designed our comparative analyses simulating some practical situations: when a good number of samples is available for training classifiers, when only a limited amount of information is available for training classifiers, and when short partial sequences need to be identified. Identifying short partial sequences helps detecting candidate gene regions based on single-exon similarities even if gene prediction methods misidentify these genomic regions. It also provides an effective way of exploiting an underutilized short Expressed Sequence Tags (ESTs). Due to its economical advantage, not surprisingly EST data comprise currently the majority of available genomic information.

We examined the following classifiers: a profile HMM, SVM_Fisher, SVM_pairwise, and simple amino-acid-composition-based classifiers using SVMs and decision trees. Performance of the classifiers against short partial sequences was examined using both simulated datasets and *Drosophila melanogaster* EST sequences. The results we obtained will be useful to gain the optimal classification power using different protein classifiers for various identification problems we encounter in practice.

2.2 Results

We divided GPCR sequences into two groups: *Class A* datasets including GPCRs belonging to a single large class, and *non-Class A* datasets including GPCRs from other classes (see Table 2.1, and Materials and methods). While Class A GPCRs are relatively more conserved, non-Class A GPCRs are extremely heterogeneous. We trained classifiers on each group of datasets, and tested against the datasets derived from the same group (*within-class test*) or from another group (*between-class test*). Table 2.2 summarizes the

combinations of datasets used in each test. The *within-class tests* are to examine how well classifiers perform if they can be trained on samples sufficiently similar to those to be identified. The *between-class tests* simulate situations when we want to search protein sequences distantly related from currently available samples.

2.2.1 Within-class tests

Fig. 2.1 summarizes the performance of the eight classifiers. The accuracy and false positive (FP) rates are plotted with circles and X's, respectively. All classifiers had 92% or higher accuracy for identifying Class A GPCRs (Fig. 2.1a). Similarly high but slightly lower accuracy rates (85% or higher) were observed against non-Class A datasets (Fig. 2.1b). In order to examine sampling effects, we repeated the performance analysis after switching datasets used for training and testing. All classifiers showed very similar consistent results between the two repeating tests (data not shown). For non-Class A, leave-one-out cross-validation tests using a larger dataset including all 162 non-Class A sequences also showed the consistent results (data not shown).

All alignment-based classifiers, SAM (a profile HMM classifier), SVM_Fisher, and SVM_pairwise, showed almost perfect discrimination in these within-class tests regardless of the GPCR classes. Amino acid composition-based classifiers, SVM_AAs and DT, even though they do not rely on alignments to compare sequences, had also very high accuracy rates. Among SVM_AAs, SVM_AA(rbf) was the best performer with lower FP rates (higher specificity).

The median and maximum rates of false positives (MedRFPs and MaxRFPs) concisely summarize the performance behavior of each classifier (see Materials and methods). These FP rates are included in Table A2.1. For all classifiers MedRFPs were 0%

or very close to 0%, indicating that a half of GPCR samples were identified correctly before any negative samples being misidentified as false positives. SVM_pairwise showed very low MaxRFPs, and SVM_AAs had slightly higher MaxRFPs (9% or higher). Surprisingly, SAM and SVM_Fisher had very high MaxRFPs for within-non-Class-A tests (*e.g.,* 62% for SAM was the average between 49 and 75%). It indicates that some non-Class A GPCRs had very low scores, and could not be identified unless setting the threshold score very low and allowing many negative samples to become false positives. Consistent with this, almost all of the errors made by SAM and SVM_Fisher were false negatives (FNs). Higher divergence among non-Class A GPCR sequences must have contributed to these results.

2.2.2 Between-class tests

The results were quite different for between-class tests. As shown in Fig. 2.1 (plotted with squares and +'s), the accuracy rates of SAM and SVM_Fisher were only around 70-80%. Low Matthews correlation coefficients (MCC < 60%; Table A2.2) of both classifiers reflect very low sensitivity (high FN rates) even though specificity was not quite low. It implies that SAM and SVM_Fisher could not identify sequences only weakly similar to their trained models. MaxRFPs of these classifiers were 100% or close to 100%, indicating some non-Class A GPCRs scored lower than almost all of the non-GPCR test sequences. Since their MedRFPs (<24%) were lower, at least a half of positive samples were found before too many negative samples being misidentified.

Surprisingly, SVM_pairwise, even though it uses pairwise alignments to compare sequences, performed the best (higher than 90% accuracy), closely followed by alignment-free SVM_AA(rbf) or SVM_AA(pol). All of the amino acid composition based classifiers

(SVM_AAs and DT) performed better than SAM and SVM_Fisher. Accuracy levels of SVM_AAs were constantly close to 90% or higher. Although their MaxRFPs were sometimes higher than those of SVM_pairwise, their MedRFPs were still very close to 0%.

2.2.3 Subsequence test

Figs 2.2 and 2.3 summarize the performance (accuracy rates) of the eight classifiers against short subsequences. Overall patterns were consistent among different classifiers; performance increased when the subsequence lengths became longer. Fig. 2.2 shows that for the within-class tests, profile HMM-based SAM and SVM_Fisher had the advantage over the other classifiers. Even against 50 or 75-amino acid (aa) subsequences, these classifiers maintained the accuracy at 94% or higher (for Class A) or 88% or higher (for non-Class A). The performance of SVM_pairwise was slightly lower than these two classifiers. Among the amino-acid composition based classifiers, DT showed the lowest accuracy rates. The accuracy rates of SVM_AAs were close to but slightly lower than those of SVM_pairwise.

Consistent with the results obtained for the full sequence analysis, for the between-class tests, SAM and SVM_Fisher gave the worst performance regardless of the subsequence lengths (Fig. 2.3). Both SVM_pairwise and SVM_AAs performed similarly and constantly better than SAM, SVM_Fisher, and DT. Their discrimination performance was better when SVM_AAs were trained on non-Class A. On the contrary, SAM performed worse when trained on non-Class A. SVM_AAs maintained around 80% accuracy even against 50-aa subsequences.

2.2.4 *D. melanogaster* EST analysis

Since almost all EST sequences contain fragments of both non-translated exons as well as coding sequences, identifying their family memberships is more challenging than

subsequence identification. Table 2.3 compares the performance between SAM and SVM_AA(rbf). The majority of *D. melanogaster* ESTs that contained GPCR coding sequences were in fact derived from Class A GPCRs (1,937 out of 2,103). Against these Class A GPCR containing ESTs, SAM performed very well when trained on the same Class A (~90% accuracy). However, none of them was correctly identified when training was done using the non-Class A dataset. Similarly, when training was done with the Class A dataset, none of non-Class A containing ESTs was correctly identified. "Frizzled/smoothened" and "odorant/gustatory receptors" are another distant GPCR groups and these sequences were not included in our training data. Predictably, SAM failed to identify the majority of the ESTs containing these sequences. In the cases where SAM failed, SVM_AA(rbf) showed better identification performance. Furthermore, the majority of the Class A containing ESTs in fact coded highly conserved opsin proteins (1,807 of 1,937). Against the remaining 130 Class A ESTs, SAM showed only a slight advantage. In total, SVM_AA(rbf) identified more GPCR containing ESTs (145) than SAM did (95). Note that, although SVM_AA(poly) seemed to perform better than SVM_AA(rbf) for short subsequences (Figs 2.2 and 2.3), in this EST analysis, SVM_AA(poly) showed extremely high FP rates (50% or higher from 370,488 negative ESTs).

2.3 Discussion

Profile HMMs are currently the most used method in protein classification (*e.g,.* Pfam, SMART, Superfamily, PANTHER). Profile HMMs are built on multiple alignments generated from known protein families. Therefore, they cannot be optimized directly for discriminating positive samples from negative samples. SVM_Fisher developed by Jaakkola *et al.* [23] combines the power of generative model building of HMMs with the

discriminative power of SVMs. Our results showed only a small improvement of performance with SVM_Fisher over SAM when the classifiers were trained and tested to identify more diverged non-Class A GPCR sequences. Both profile HMM-based classifiers performed poorly in between-class tests and they misidentified many GPCRs as false negatives. While the higher specificity of profile HMMs contributed to very low errors when classification was against the same group of sequences they were trained on, such high specificity may have prevented profile HMMs to identify distantly related sequences not well-represented in their models. SVM_pairwise surpassed profile HMM-based classifiers, especially for between-class tests. It appears to combine the strength in profile-HMMs (high specificity) and flexibility in SVM_AAs. The simple use of amino acid frequencies with SVMs is completely free from alignments and was very effective for discriminating GPCRs from non-GPCRs regardless of how they were trained.

Based on the different results we obtained in this study, profile HMMs have an advantage when training and testing can be done using sufficiently similar sequences. SVM_AAs perform better when currently available sample proteins do not represent well the remotely similar new proteins that are needed to be identified. It is beneficial for the users to know how remote is too remote to select the best classifier for their interest. In order to examine further the relationships between the level of similarity and classifier performance, we performed the similar analyses using different families among Class A GPCRs as shown in Table 2.4 (see Materials and methods). Three major families (Amine/Rhodopsin, Peptide, and Olfactory) were chosen from Class A. One of these Class-A-family datasets was used for training, and the testing was done against the other two Class-A-family datasets. As shown in Table 2.5, SAM and SVM_pairwise performed better than SVM_AA(rbf). Such results were expected since the difference among these Class A families are not as great as between-class

tests. In fact, sensitivities of SVM_AA(rbf) were very close to those of SAM. Performance decrease observed in SVM_AA(rbf) was mainly caused by the misclassification of negative samples but not positives. Furthermore, the performance by SAM trained with the Olfactory family dataset (OL), the most conserved datasets, was the lowest, showing a possible overfitting effect. Compared to SAM and SVM_AA(rbf), SVM_pairwise showed again consistently almost perfect classification performance.

The disadvantage of using SVM_pairwise is its computational expense. It requires generating all combination of Smith-Waterman local pairwise alignments both in training and testing. It becomes computationally significantly expensive especially against larger datasets (*e.g.*, genomes). On the contrary, SVM_AA is quick and simple, requiring only amino acid composition from each protein. There are many public softwares that can be used to obtain amino acid composition from protein sequences. Using SVM_AA is easy and more practical especially for large-scale (*e.g.,* genome-scale) analyses.

We should note that the results shown so far were obtained at the minimum error point (MEP). It shows the best possible performance each classifier can produce, and such performance cannot be expected in the real life. In the reality, we have to rely on the classifiers optimized based on the training set used. When we used the results simply produced by each classifier as a default output (using e-value = 0.05 as the threshold for SAM), the results for within-class tests were close to those obtained at the MEP (see Supplementary Materials). However, the accuracy rates for between-class tests by SAM, SVM_Fisher, and SVM_pairwise were lower by as much as 20%. The difference was much smaller for SVM_AAs.

In Kim *et al*. [8] and Moriyama and Kim [9], they reported the performance of their alignment-free classifiers to be better than that of profile HMMs (Pfam) especially for short subsequence identification. The datasets they used to train and test their classifiers were

randomly sampled across the entire GPCR classes. For profile HMMs, however, multiple models were collected from the Pfam database, with each model corresponding to a different GPCR class (*e.g.,* 7tm_1 for the rhodopsin family). Therefore, their results for profile HMM/Pfam were equivalent to results combined from within and between-class tests in this study. In fact, this is generally what happens when we submit query sequences to profile HMM databases such as Pfam. For example, currently 22 GPCR proteins are known from *Arabidopsis thaliana* [24; 25; 26; 27]. Using multiple profile HMMs constructed from 14 GPCR groups, Fredriksson and Schioth [28] identified only six *Arabidopsis* GPCRs. In their recent study, Ono *et al.* [29] reported that combining profile HMMs with other methods including BLAST [30] and PROSITE [6], they could identify 21 of the *Arabidopsis* GPCRs. Compared to such a small number of GPCRs found in *Arabidopsis*, animal genomes encode much larger numbers of GPCRs (*e.g.,* >800 in human and ~1000 in *Caenorhabditis elegans*; [25]). It indicates either that the number of GPCRs exploded only in metazoan lineages after plants and metazoa parted their evolutionary histories, or that distant plant members have not been identified properly. Combining various alignment-free classifiers and transmembrane prediction methods, for example, our group recently identified about 400 GPCR candidates from the *A. thaliana* genome [31]. Although knowing how many of these candidates are actual GPCRs (true positives) needs experimental confirmation, relying only on highly specific results produced by profile HMMs does not allow us to explore such possibilities.

Recently a new alignment-free GPCR detection method, GPCRHMM, was developed by Wistrand *et al.* [32]. The authors analyzed TM topologies among GPCRs, and compared differences in loop lengths and amino acid composition between different GPCR regions. A hidden Markov model is built based on these regional features. Since their classifier was trained using positive samples collected across the entire GPCR families (except for plant Mlo and insect odorant receptor families), it is not possible to compare the results from our

within- and between-test analyses directly with those by GPCRHMM. Nevertheless such comparisons would be beneficial for the users when choosing classifiers. Therefore, we applied GPCRHMM against all of our datasets (Table A2.4). As expected, GPCRHMM discriminated Class A and non-Class A GPCRs from non-GPCRs with very high accuracies. All Class A sequences (AR, PE, and OL datasets in Table 2.4) were identified almost perfectly. On the other hand, of the two non-Class A GPCR datasets (N1 and N2 in Table 2.4) 70 sequences each were identified as negative (non-GPCR). This is, however, not surprising because the training samples used for GPCRHMM do not include those extremely diverged GPCRs such as plant Mlo's and insect odorant receptors. In each of the non-Class A GPCR datasets (N1 and N2), 68 sequences were obtained from these families and these sequences were missed by GPCRHMM. This result shows again that it is very important to understand how classifiers are trained and for what purpose we want to use each classifier.

2.4 Conclusions

SVM_pairwise is the most balanced classifier that is sensitive to remote similarity and can be also highly discriminative for classifying GPCR classes. However, use of SVM_pairwise for a large-scale analysis may not be practical for its computational cost. To identify member proteins from well-established protein families where a good number of representative samples are available, profile HMMs as well as GPCRHMM give highly accurate classifications. When protein sequences of our interests are distant members of divergent protein families and only a limited amount of information is available for training classifiers, SVM_AA(rbf) is the better alternative. Our recommendation is thus to use both SAM (or GPCRHMM) and SVM_AA(rbf) for the first stage analysis, and to follow up with SVM_pairwise to reduce false positives to achieve a thorough mining of divergent protein

family members.

2.5 Materials and methods

2.5.1 Data sources

GPCRs are seven-transmembrane proteins involved with G-protein mediated signal transduction. They form a large (the largest among eukaryotic transmembrane protein families) and highly diverged superfamily. GPCRDB (Information System for G Protein-Coupled Receptors) [25] divides the superfamily into five major classes (see Table 2.1). Class A is by far the most populated GPCR class with more than 4,300 entries in the database. Other families not listed in Table 2.1 are, for example, "Frizzled/Smoothened", "Insect odorant receptors", and "Plant Mlo receptors" (see http://www.gpcr.org/7tm for the complete listing of GPCR families). Other GPCR classification systems exist. For example, Fredriksson *et al*. [28] divide Class B into two major families: "Secretin" and "Adhesion". However, for the purpose of our current study, the difference is not significant. Each class is further divided into families, subfamilies, and so forth, based on their ligand-specificities as well as sequence similarities.

The GPCR sequences of different classes/families are highly diverged from each other. Their lengths are also varied especially in the 5' and 3'-terminal as well as loop regions. Such high variation makes reconstructing reliable multiple alignments across families or from the entire GPCR superfamily very difficult or practically impossible. This is, therefore, an ideal protein family for us to analyze classifier performance at various degrees of similarities. GPCRs have been also used in previous classifier developments [e.g., 8; 9; 10; 13; 14; 15; 16; 17; 18; 32].

As shown in Table 2.1, entries in GPCRDB are derived from Swiss-Prot Protein Knowledgebase [33], a curated protein database providing high quality annotations, as well as its computer-annotated supplement, TrEMBL. In order to use GPCR sequences less likely

to be misclassified, for our positive samples, we included only Swiss-Prot derived GPCR sequences.

2.5.2 Positive and negative samples

The lists of accession numbers for the sequences used in each dataset are available in Supplementary Materials. All sequences are available from: http://bioinfolab.unl.edu/emlab/gpcr/

Class A datasets

200 GPCR sequences were randomly sampled from Class A. Such random sampling may not represent all groups evenly since some groups are represented by only small numbers of entries in the database and other groups include many highly similar sequences. In order to examine the effect of training data sampling, we previously examined two other sampling methods: a phylogeny-based sampling using a certain cut-off similarity level, and family-wise sampling based on the Class A classification by GPCRDB. The phylogeny-based sampling avoids redundant representation by highly similar sequences, and the family-wise sampling avoids biased representation by large groups. While these sampling methods could cover the entire GPCR sequence space more evenly, no significant improvement was observed in classifier performance (for detailed descriptions, see Khati [34]). In this study we thus used only random sampling for preparing training datasets. Two independent datasets were prepared from Class A GPCRs.

Non-Class A datasets

Positive datasets were also generated by sampling from non-Class A (including Classes B, C, D, and E). As shown in Table 2.1, only 162 GPCR sequences were available

for non-Class A. One positive dataset was prepared including all of these sequences. Two other smaller but non-overlapping positive datasets were also generated by randomly dividing the 162 sequences into two (each including 81 non-Class A GPCRs).

Non-GPCR negative datasets

For negative samples, 200 non-GPCR sequences longer than 100 amino acids were randomly sampled from Swiss-Prot. We added also ten bacteriorhodopsin sequences. Bacteriorhodopsins are seven-transmembrane proteins. However, they do not couple with G proteins, nor function as GPCRs. Adding such somehow similar but unrelated negative samples may improve the discriminating power of classifiers, resulting in fewer false positives. Note, however, that Khati [34] reported that such performance increase was minimal. The total number of sequences in each negative dataset was thus 210. Two independent negative sets were prepared.

Datasets used for Class A family analysis

From Class A GPCRs, we chose four major subfamilies: Amine, Peptide, Opsin (rhodopsin), and Olfactory. Clustering patterns were examined by phylogenetic analysis using ClustalW multiple alignment [35], protein distance estimation based on the JTT model [36], and neighbor-joining phylogenetic reconstruction [37] implemented in Phylip (version 3.65) [38]. The consistent results were obtained by Fredriksson *et al*. [39] in their extensive analysis of human GPCRs. Amine and Opsin groups were closely clustered and Fredriksson *et al*. [39] included them in a single group α. Therefore, we combined these two groups and generated three Class A datasets AR, PE, and OL as shown in Table 2.4. Their average pairwise divergence (amino acid substitutions per site estimated by JTT protein distance) was the highest among the Peptide (PE) group and the lowest among the Olfactory (OL) group. Pairwise protein divergence of 0.3 was used to identify highly similar sequence clusters, and from each of such clusters single sequence was randomly chosen and others were excluded.

For non-Class A datasets, GPCR sequences were obtained from Classes B-E (Table 2.1) as well as "Frizzled/Smoothened", "Ocular albinism proteins", "Insect odorant receptors", "Plant Mlo receptors", "Nematode chemoreceptors", "Vomeronasal receptors", and "Taste receptors T2R". As before, highly similar sequences were removed by using pairwise protein divergence of 0.3 as the cut-off threshold. Two non-overlapped datasets (N1 and N2 in Table 2.4) were generated and one (N1) was used for training and the other (N2) for testing.

2.5.3 Training and test datasets preparation

Positive and negative datasets were combined to create Class A training and test sets, each including 410 sequences, and non-Class A training and test sets, one including 372 and two including 291 sequences. The two Class A datasets and the two smaller non-Class A datasets were mutually exclusive.

Subsequence test sets

Based on the average length of GPCRs (374 aa from Class A), six lengths were chosen: 50, 75, 100, 150, 200, and 300 aa. One subsequence with a given length was randomly taken from each sequence of the dataset. While all GPCR sequences were longer than 300 aa, some non-GPCR sequences were shorter than the required lengths and had to be replaced with new sequences obtained from Swiss-Prot. Six subsequence test sets were generated for one each dataset of Class A and non-Class A, each including 410 and 291 sequences, respectively.

Drosophila melanogaster EST datasets

374,229 *D. melanogaster* EST sequences (337,753 for 5' and 36,476 for 3' ESTs)

were collected from the EST division of Genbank (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) in October 2005. Using blastx similarity search program (http://www.ncbi.nlm.nih.gov/BLAST/) [40], we compared them against all 304 *D. melanogaster* GPCR protein sequences in GPCRDB. Using 90% for the amino acid identity and 5 aa (15 bp) length of HSPs (High-scoring Segment Pairs or regions aligned with GPCR coding sequences) as the threshold, we identified 2,103 ESTs (1,994 for 5' and 109 for 3' ESTs) that contain fragments of GPCR coding sequences. The average length of these ESTs was 557 bp (ranging from 151 bp to 871 bp). The average HSP length was 125 bp, and on average an HSP covered 20 – 25 % of each EST. These 2,103 ESTs were translated in three reading frames and used for testing classifier performance.

Class A analysis datasets

For the within-family tests, each of the three Class A datasets (AR, PE, and OL in Table 2.4) was randomly divided into two. One was combined with a non-Class A datatset N1 and used for training, and the other was combined with another non-Class A dataset N2 and used for testing. For the between-family tests, each of the three Class A datasets (AR, PE, and OL) was combined with the non-Class A dataset N1 for training. Two of the three Class A datasets not used for training was combined with another non-Class A dataset N2, and used for testing (*e.g.,* if AR+N1 dataset was used for training, PE+OL+N2 dataset was used for testing).

2.5.4 Classifiers used

Profile hidden Markov models (HMMs)

A profile HMM is a full probabilistic representation of a sequence profile [41].

Sample sequences need to be alignable, and thus only positive sample information is directly incorporated. We used the program package of Sequence Alignment and Modeling System (SAM version 3.5; http://www.cse.ucsc.edu/research/compbio/sam.html) [42] in this study, *buildmodel* was used to build profile HMMs with the nine-component Dirichlet mixture priors [43] and *hmmscore* was used to calculate scores and e-values. The 'calibration' option (for more accurate e-value calculation) and the fully local scoring option (-sw 2) were used. The *w0.5* script is to build profile HMMs especially for searching remotely similar sequences. We built profile HMMs with and without using the *w0.5* script. As shown in Appendices, especially for between-class test, *w0.5* did not consistently improve GPCR discrimination performance. Therefore, we discussed only results obtained without using *w0.5*.

Support vector machines (SVMs)

SVMs are learning machines that make binary classifications based on a hyperplane separating a remapped instance space [44]. Kernel functions are chosen so that the remapped instances on a multidimensional space are linearly separable. Both positive and negative samples are used in their training.

SVM_Fisher. This method introduced by Jaakkola *et al.* [23] combines generative models (trained only on positive samples as profile HMMs) with discriminative methods, SVMs. If an HMM, H_1, is built from a set of positive sequences, the probability model for a sequence X is denoted as $P(X|H_1,\theta)$, and a Fisher score vector (FSV) is given by $U_X = \Delta_q logP(X|H_1,\theta)$. The detailed derivation of the FSV is given by Karchin *et al.* [10].

Given a profile HMM, each sample sequence was compared against it using a SAM program, *get_fisher_scores*, and transformed into a 9n-component FSV based on the nine-

component Dirichlet mixture ('matchprior' option; *n* is the number of match states). This FSV was then used as an input vector for SVMs. A program *svm_learn* of the SVM^{light} package (version 5.0; http://svmlight.joachims.org/) [45] was used with a radial basis kernel, $\exp(-\gamma||x-y||^2)$, where γ was set based on the median of Euclidean distances between positive examples and the nearest negative example as described in Jaakkola *et al.* [23]. SVM classification was done by another SVM^{light} program, *svm_classify*.

SVM_pairwise. In this method developed by Liao and Noble [11], each sequence is compared to every sequence in the data set by the Smith-Waterman local pairwise alignment [46]. If *n* is the total number of proteins in the training set and f_{xi} is the e-value of the Smith-Waterman similarity score between a sequence *X* and the *i*-th training sequence ($i = 1, 2, \ldots, n$), the feature vector corresponding to a sequence *X* is in the form of $F_X = [f_{x1}, f_{x2}, \ldots, f_{xn}]$. *SSearch* (version 3.4) [47] was used as an implementation of the Smith-Waterman algorithm with the default options (open gap penalty = 12, gap extension penalty = 2, BLOSUM50 scoring matrix). SVM^{light} programs were used as above with the set of e-values as the input vector and with the radial basis kernel.

SVMs with amino acid composition. Simple nineteen amino acid frequencies of each protein sequence (the 20th amino acid frequency can be explained completely by the other 19) were used as an input vector for SVMs. The SVM^{light} package was used as before. Four kernel functions used are the linear kernel, $(x \bullet y + 1)$; the polynomial kernel, $(kx \bullet y + 1)^p$; the sigmoid kernel, $\tanh(kx \bullet y + c)$; and the radial basis kernel, $\exp(-\gamma ||x-y||^2)$. γ in the radial kernel function was set as described before (γ=122 for Class A and γ=126 for non-Class A were used. Also the regulatory parameter C was set as 0.5002 for Class A and 0.5003 for non-Class A data sets). The other parameter values were chosen for the most optimal discrimination. We call these SVM classifiers SVM_AA(lin), SVM_AA(pol), SVM_AA(sig), and SVM_AA(rbf), respectively.

Decision trees (DT)

The nineteen amino acid frequencies were also used as an input vector for decision trees. The program C4.5 (release 8; http://www.rulequest.com/Personal/) by Quinlan [48] was used. A decision trees classifier with boosting showed only a minimum performance gain [34]. Therefore, in this study, we used the decision trees without boosting.

GPCRHMM

Recently Wistrand *et al.* [32] developed a new GPCR detection method, GPCRHMM. It incorporates GPCR-specific TM features (*e.g.,* loop-region lengths, different amino acid composition among loop and TM regions) in a hidden Markov model architecture. GPCRHMM is available at http://gpcrhmm.cgb.ki.se/index.html.

2.5.5 Performance Analysis

Test statistics

Classification results are grouped as the following four categories:

• True positive (*TP*): the number of actual GPCRs predicted as GPCRs

• False positive (*FP*): the number of actual non-GPCRs predicted as GPCRs

• True negative (*TN*): the number of actual non-GPCRs predicted as non-GPCRs

• False negative (*FN*): the number of actual GPCRs predicted as non-GPCRs

Based on these numbers, following performance measures were calculated:

• Accuracy: (TP + TN) / (TP + TN + FP + FN) = 1 – error rate

• Sensitivity: TP / (TP + FN)

• Specificity: TN / (TN + FP) = 1 – FP rate

• Matthews correlation coefficient (MCC) =

$$(TP \times TN - FP \times FN) / \{(TP + FN)(TP + FP)(TN + FP)(TN + FN)\}^{1/2}$$

MCC provides more balanced evaluation of performance (reviewed in, *e.g.,* [49]).

Minimum error point (MEP)

The minimum error point (MEP) is the threshold score where the classifier produces the minimum number of errors (FP + FN) showing the best possible performance. MEP was used in Karchin *et al.* [10]. Unless specified, the performance statistics were obtained at the MEP for all classifiers except for DT.

Maximum and median rates of false positives (MaxRFP and MedRFP)

The maximum rate of false positives (MaxRFP) is the FP rate at a certain threshold score where all positive samples are correctly identified. Similarly, the median rate of false positives (MedRFP) is the FP rate at a certain threshold score where a half of the positive samples are correctly identified. These statistics (used in [23]) concisely summarize the behavior of each classifier performance. Therefore, we chose to show these statistics in Tables A2.1 and A2.2 instead of receiver operating characteristic (ROC) curves, which is the plot between TP rates (sensitivities) against FP rates (1-specificity) with a given range of threshold values.

Leave-one-out cross-validation test

Since non-Class A datasets were much smaller than Class A, and two independent datasets prepared from non-Class A included only 81 positive samples, in addition to independent test data analysis, we performed leave-one-out cross-validation analysis, too.

For non-Class A, the dataset including the entire positive samples (162 sequences) was used for this analysis.

2.6 References

[1] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J. Studholme, C. Yeats, and S.R. Eddy, The Pfam protein families database. Nucleic Acids Res 32 (2004) D138-141.

[2] I. Letunic, R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, and P. Bork, SMART 4.0: towards genomic data integration. Nucleic Acids Res 32 (2004) D142-144.

[3] J. Gough, K. Karplus, R. Hughey, and C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313 (2001) 903-19.

[4] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M.J. Campbell, H. Kitano, and P.D. Thomas, The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33 (2005) D284-8.

[5] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri, PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31 (2003) 400-402.

[6] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, and C.J. Sigrist, The PROSITE database. Nucleic Acids Res 34 (2006) D227-30.

[7] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R.R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S.E.

Orchard, M. Pagni, D. Peyruc, C.P. Ponting, J.D. Selengut, F. Servant, C.J.A. Sigrist, R. Vaughan, and E.M. Zdobnov, The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31 (2003) 315-318.

[8] J. Kim, E.N. Moriyama, C.G. Warr, P.J. Clyne, and J.R. Carlson, Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. Bioinformatics 16 (2000) 767-775.

[9] E.N. Moriyama, and J. Kim, Protein family classification with discriminant function analysis. in: J.P. Gustafson, R. Shoemaker, and J.W. Snape, (Eds.), Genome Exploitation: Data Mining the Genome, Springer, New York, 2005, pp. 121-132.

[10] R. Karchin, K. Karplus, and D. Haussler, Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18 (2002) 147-159.

[11] L. Liao, and W.S. Noble, Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol 10 (2003) 857-868.

[12] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A.G. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 (2004) D226-229.

[13] M. Bhasin, and G.P. Raghava, GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 32 (2004) W383-9.

[14] Y.F. Wang, H. Chen, and Y.H. Zhou, Prediction and classification of human G-protein coupled receptors based on support vector machines. Genomics Proteomics Bioinformatics 3 (2005) 242-6.

[15] B.Y. Cheng, J.G. Carbonell, and J. Klein-Seetharaman, Protein classification based on text document classification techniques. Proteins 58 (2005) 955-70.

[16] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, and J.E.S. Wikberg, Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci 11 (2002) 795-805.

[17] I. Gunnarsson, P. Andersson, J. Wikberg, and T. Lundstedt, Multivariate analysis of G protein-coupled receptors. J Chemometrics 17 (2003) 82-92.

[18] J.M. Otaki, A. Mori, Y. Itoh, T. Nakayama, and H. Yamamoto, Alignment-free classification of G-protein-coupled receptors using self-organizing maps. J Chem Inf Model 46 (2006) 1479-90.

[19] R.L. Dunbrack, Jr., Sequence comparison and protein structure prediction. Curr Opin Struct Biol 16 (2006) 374-84.

[20] X.F. Wan, and D. Xu, Computational methods for remote homolog identification. Curr Protein Pept Sci 6 (2005) 527-46.

[21] P.J. Clyne, C.G. Warr, M.R. Freeman, D. Lessing, J.H. Kim, and J.R. Carlson, A novel family of divergent seven-transmembrane proteins: Candidate odorant receptors in *Drosophila*. Neuron 22 (1999) 327-338.

[22] P.J. Clyne, C.G. Warr, and J.R. Carlson, Candidate Taste Receptors in *Drosophila*. Science 287 (2000) 1830-1833.

[23] T. Jaakkola, M. Diekhans, and D. Haussler, A discriminative framework for detecting remote protein homologies. J Comput Biol 7 (2000) 95-114.

[24] S. Pandey, and S.M. Assmann, The Arabidopsis putative G protein-coupled receptor GCR1 interacts with the G protein alpha subunit GPA1 and regulates abscisic acid signaling. Plant Cell 16 (2004) 1616-32.

[25] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F.E. Cohen, and G. Vriend, GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31 (2003) 294-297.

[26] M.-H. Hsieh, and H.M. Goodman, A novel gene family in Arabidopsis encoding putative heptahelical transmembrane proteins homologous to human adiponectin receptors and progestin receptors. J. Exp. Bot. 56 (2005) 3137-3147.

[27] N.-F. Chen, J.-Z. Yu, N.P. Skiba, H.E. Hamm, and M.M. Rasenick, A specific domain of $G_i\alpha$ required for the transactivation of $G_i\alpha$ by tubulin is implicated in the organization of cellular microtubules. J. Biol. Chem. 278 (2003) 15285-15290.

[28] R. Fredriksson, and H.B. Schioth, The repertoire of G-protein-coupled receptors in fully sequenced genomes. Mol Pharmacol 67 (2005) 1414-25.

[29] Y. Ono, W. Fujibuchi, and M. Suwa, Automatic gene collection system for genome-scale overview of G-protein coupled receptors in eukaryotes. Gene 364 (2005) 63-73.

[30] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (1997) 3389-3402.

[31] E.N. Moriyama, P.K. Strope, S.O. Opiyo, Z. Chen, and A.M. Jones, Mining the *Arabidopsis thaliana* genome for highly-divergent seven transmembrane receptors. Genome Biology 7 (2006) R96.

[32] M. Wistrand, L. Kall, and E.L. Sonnhammer, A general model of G protein-coupled receptor sequences and its application to detect remote homologs. Protein Sci 15 (2006) 509-21.

[33] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31 (2003) 365-370.

[34] P. Khati, Comparative Analysis of Protein Classification Methods, Master's Thesis, Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, 2004.

[35] J.D. Thompson, D.G. Higgins, and T.J. Gibson, Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22 (1994) 4673-4680.

[36] D.T. Jones, W.R. Taylor, and J.M. Thornton, The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8 (1992) 275-282.

[37] N. Saitou, and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4 (1987) 406-425.

[38] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author., Department of Genetics, University of Washington, Seattle, 2005.

[39] R. Fredriksson, M.C. Lagerstrom, L.G. Lundin, and H.B. Schioth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63 (2003) 1256-72.

[40] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, Basic local alignment search tool. J Mol Biol 215 (1990) 403-410.

[41] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.

[42] R. Hughey, and A. Krogh, Hidden Markov models for sequence analysis: Extension and analysis of the basic method. Comput Appl Biosci 12 (1996) 95-107.

[43] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler, Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. Comput Appl Biosci 12 (1996) 327-345.

[44] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1999.

[45] T. Joachims, Making large-Scale SVM Learning Practical. in: B. Schölkopf, C. Burges, and A. Smola, (Eds.), Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, 1999, pp. 169-184.

[46] T.F. Smith, and M.S. Waterman, Identification of common molecular subsequences. J Mol Biol 147 (1981) 195-197.

[47] W.R. Pearson, Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 11 (1991) 635-650.

[48] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.

[49] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16 (2000) 412-424.

Table 2.1: The five major classes of G-protein coupled receptors.

Classes	Examples	Numbers of entries[1]
A: Rhodopsin like	rhodopsin, adrenergic receptor	4,350 (1,593)
B: Secretin like	secretin receptor, calcitonin receptor	198 (107)
C: Metabotropic glutamate/pheromone	metabotropic receptor	135 (40)
D: Fungal pheromone	fungal pheremone receptor STE2-like	24 (11)
E: cAMP receptors (*Dictyostelium*)	cAMP receptor	5 (4)

[1]The numbers of entries are based on the GPCRDB July 2004 release. Numbers in parentheses are those including only Swiss-Prot derived entries.

Table 2.2: Datasets used in within- and between-class tests.[1]

Training datasets		Test datasets	
Positive	Negative	Positive	Negative
[Within-class test]			
Class A (200)	Non-GPCR (210)	Class A (200)	Non-GPCR (210)
Non-Class A (81)	Non-GPCR (210)	Non-Class A (81)	Non-GPCR (210)
[Between-class test]			
Class A (200)	Non-GPCR (210)	Non-Class A (162)	Non-GPCR (210)
Non-Class A (162)	Non-GPCR (210)	Class A (200)	Non-GPCR (210)

[1]The datasets used in training and test are independent to each other. The number of sequences included in each dataset is shown in the parentheses.

Table 2.3: Identification of *D. melanogaster* ESTs containing GPCR coding sequences.[1]

GPCR classes[2]	Numbers of ESTs identified by the classifiers (%)					
	SAM			SVM_AA(rbf)		
	Class A[3]	Non-Class A[3]	Combined[4]	Class A[3]	Non-Class A[3]	Combined[4]
A (1,937/130)	**1,672/55 (86.3/42.3)**	0/0 (0/0)	**1,703**/86 (**83.4**/36.6)	1,435/45 (74.1/34.6)	**251/64 (13.0/49.2)**	1,541/**105** (75.5/**44.7**)
Non-A (105)	0 (0)	**31 (29.5)**		**23 (21.9)**	22 (21.0)	
Fz (34)	0 (0)	0 (0)	0 (0)	**7 (20.6)**	**24 (70.6)**	**24 (70.6)**
OR (27)	9 (33.3)	0 (0)	9 (33.3)	**16 (59.3)**	**12 (44.4)**	**16 (59.3)**
[Total] (2,103/296)	**1,681/64 (79.9/21.6)**	31/31 (1.5/10.5)	**1,712**/95 (**81.4**/32.1)	1,481/51 (70.4/17.2)	**309/122 (14.7/41.2)**	1,581/**145** (75.2/**49.0**)

[1]The numbers (%) of ESTs after excluding possible opsin ESTs are given after '/'. The numbers (%) shown in boldface indicate where either of the classifiers has better performance compared to the other.
[2]A: Class A; Non-A: non-Class A (including B, C, D, and E); Fz: frizzled/smoothened; OR: odorant and gustatory receptors. The numbers of ESTs containing GPCR coding sequence fragments are shown in the parentheses.
[3]The dataset used to train each classifier.
[4]The numbers of ESTs identified by the classifier trained with either or both of Class A and non-Class A datasets.

Table 2.4: Datasets used for the Class A family analysis.

Dataset names (families)	Numbers of entries[1]	Average pairwise divergence ± SD
[Class A]		
AR (Amine/Rhodopsin)	126 (296)	2.14 ± 0.61
PE (Peptide)	139 (552)	2.44 ± 0.57
OL (Olfactory)	309 (476)	1.28 ± 0.33
[Non-Class A]		
N1	158	6.96 ± 3.80
N2	158	7.81 ± 5.30

[1]Protein sequences that have pairwise divergence (amino acid substitutions per site) lower than 0.3 were excluded. The numbers in parentheses are those before the exclusion. The total number of non-Class A entries before such exclusion was 597.

Table 2.5: Classifier performance for Class A between-family analysis.[1]

Methods	Family[2]	Errors (FP/FN)	Accu-racy	Sensi-tivity	Speci-ficity	MCC	MaxRFP	MedRFP
SAM	AR	6 (1/5)	0.99	0.99	0.99	0.97	0.52	0
SAM	PE	4 (4/0)	0.99	1	0.97	0.98	0.03	0
SAM	OL	89 (38/51)	0.79	0.81	0.76	0.56	0.95	0
SVM_pairwise	AR	4 (0/4)	0.99	0.99	1	0.98	0.22	0
SVM_pairwise	PE	4 (3/1)	0.99	1.0	0.98	0.98	0.03	0
SVM_pairwise	OL	8 (2/6)	0.98	0.98	0.99	0.96	0.27	0
SVM_AA(rbf)	AR	124 (114/10)	0.80	0.98	0.28	0.39	0.94	0.22
SVM_AA(rbf)	PE	64 (41/23)	0.89	0.95	0.74	0.72	0.65	0
SVM_AA(rbf)	OL	114 (66/48)	0.73	0.82	0.58	0.41	0.97	0

[1]The results from Class A within-family tests are shown in Table A2.3.
[2]The Class A family dataset used to train each classifier. The between-family tests were performed using the two families that were not used for the training. See Table 2.4 for these datasets.

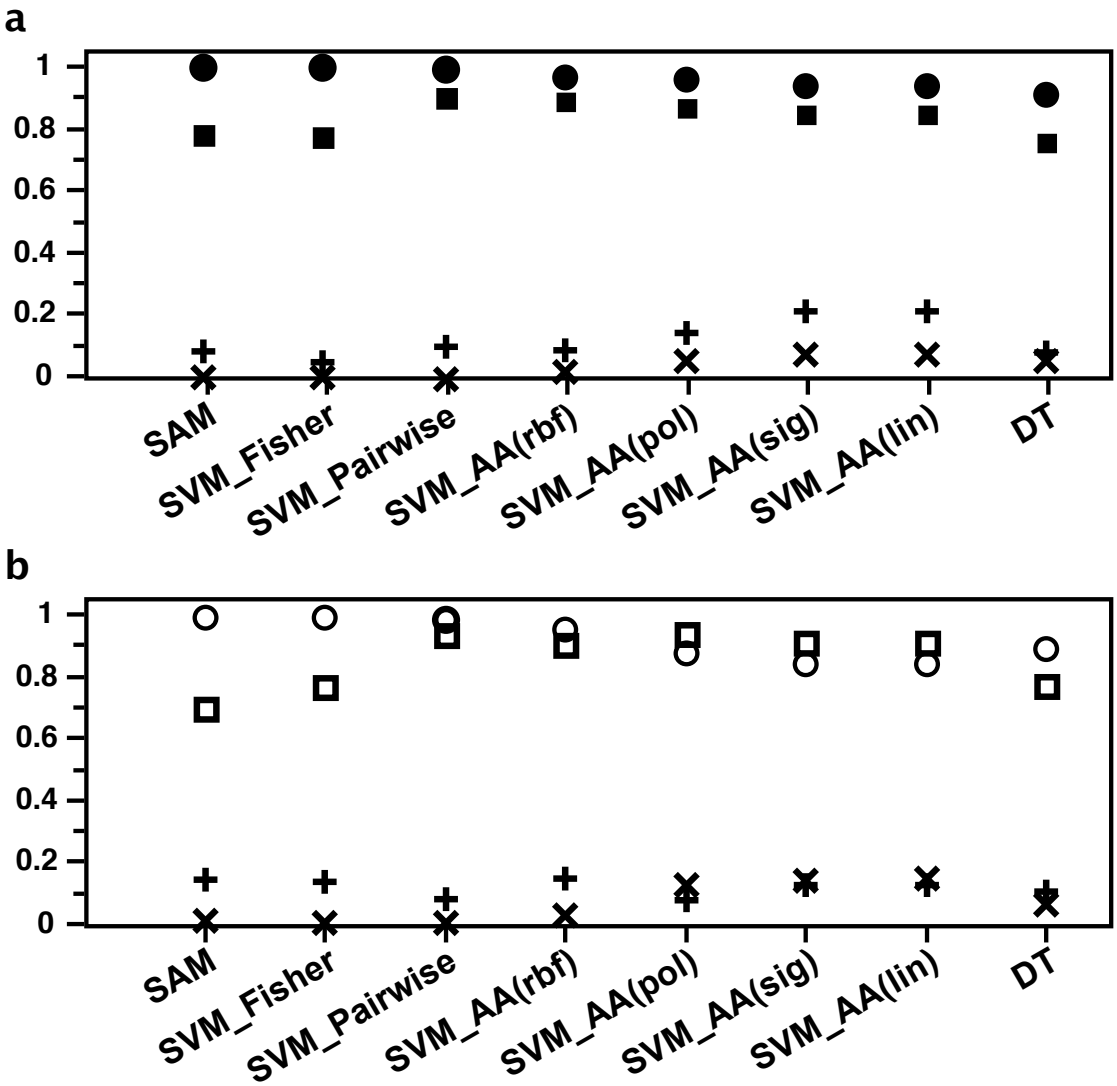

Fig 2.1. Performance comparison among eight classifiers. Classifiers were trained on the Class A dataset (a) or trained on the non-Class A dataset (b). Circles and squares plot the accuracy rates for the within-class and for the between-class tests, respectively. 'X' and '+' show the FP rates for the within-class and for the between-class tests, respectively. The detailed statistics are listed in Tables A2.1 and A2.2.

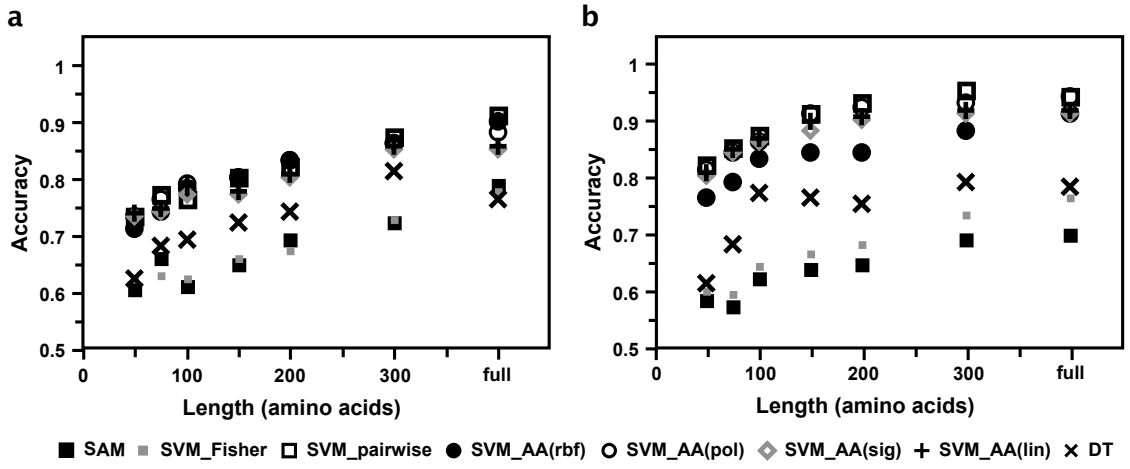

Fig 2.2. Performance comparison among eight classifiers for within-class subsequence tests. Classifiers were trained and tested on the Class A datasets (a) or trained and tested on the non-Class A datasets (b). The accuracy rates when classifiers were tested on the full test sequences are plotted above 'full'.

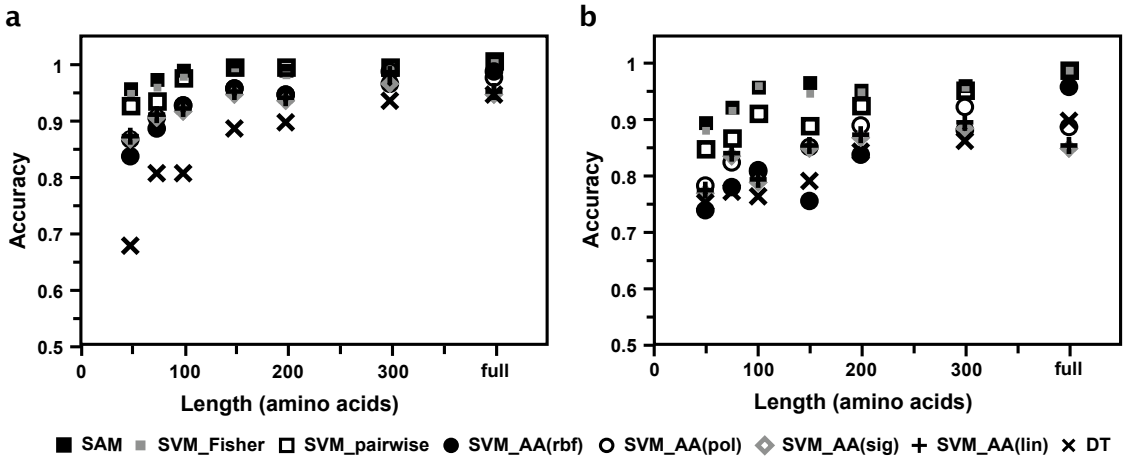

Fig 2.3. Performance comparison among eight classifiers for between-class subsequence tests. Classifiers were trained on the Class A dataset and tested on the non-Class A dataset (a) or *vice versa* (b). The accuracy rates when classifiers were tested on the full test sequences are plotted above 'full'.

Additional files

Table A2.1: Classifier performance for within-class tests. [1]

Methods	Class[2]	Errors (FP/FN)	Accu-racy	Sensi-tivity	Speci-ficity	MCC	MaxRFP	MedRFP
SAM	A	1 (0/1)	1.00	1.00	1	0.99	0	0
SAM(w0.5)[3]	A	0.5 (0/0.5)	1.00	1.00	1	1.00	0.00	0
SVM_Fisher	A	1 (0/1)	1.00	1.00	1	1.00	0.18	0
SVM_Fisher (w0.5)[3]	A	1.5 (1/0.5)	1.00	1.00	1.00	0.99	0.01	0
SVM_pairwise	A	1 (0/1)	1.00	1.00	1	1.00	0.01	0
SVM_AA(rbf)	A	10.5 (4/6.5)	0.97	0.97	0.98	0.95	0.09	0
SVM_AA(pol)	A	14.5 (11.5/3)	0.96	0.99	0.95	0.93	0.25	0.00
SVM_AA(sig)	A	24 (16/8)	0.94	0.96	0.92	0.88	0.18	0.01
SVM_AA(lin)	A	23.5 (16/7.5)	0.94	0.96	0.92	0.89	0.18	0.01
DT	A	33.5 (11.5/22)	0.92	0.89	0.95	0.84	-	-
SAM	N	4.5 (1/3.5)	0.98	0.96	1.00	0.96	0.62	0
SAM(w0.5)[3]	N	2 (0/2)	0.99	0.98	1	0.98	0.43	0
SVM_Fisher	N	3.5 (0.5/3)	0.99	0.96	1.00	0.97	0.52	0
SVM_Fisher (w0.5)[3]	N	2 (0.5/1.5)	0.99	0.98	1.00	0.98	0.05	0
SVM_pairwise	N	4 (1/3)	0.99	0.96	1.00	0.97	0.04	0
SVM_AA(rbf)	N	12.5 (7.5/5)	0.96	0.94	0.96	0.89	0.26	0.00
SVM_AA(pol)	N	33.5 (27.5/6)	0.88	0.93	0.87	0.75	0.22	0.03
SVM_AA(sig)	N	45 (31.5/13.5)	0.85	0.83	0.85	0.65	0.26	0.06
SVM_AA(lin)	N	44.5 (32.5/12)	0.85	0.85	0.85	0.66	0.26	0.06
DT	N	29.5 (14/15.5)	0.90	0.81	0.93	0.75	-	-

[1]Values shown are the average from two independent tests. Class A and non-Class A datasets included 410 and 291 sequences, respectively.
[2]The dataset used to train each classifier. A; Class A, N: non-Class A.
[3]Results obtained using w0.5 of the SAM package.

Table A2.2: Classifier performance for between-class tests.

Methods	Class[1]	Errors (FP/FN)	Accu-racy	Sensi-tivity	Speci-ficity	MCC	MaxRFP	MedRFP
SAM	AN	80 (14/66)	0.78	0.59	0.93	0.57	1.00	0.03
SAM(w0.5)[2]	AN	79 (18/61)	0.79	0.62	0.91	0.57	0.96	0.03
SVM_Fisher	AN	84 (8/76)	0.77	0.53	0.96	0.56	0.99	0.02
SVM_Fisher (w0.5)[2]	AN	69 (20/49)	0.81	0.70	0.90	0.62	0.90	0.03
SVM_pairwise	AN	34 (19/15)	0.91	0.91	0.91	0.81	0.52	0.01
SVM_AA(rbf)	AN	38 (16/22)	0.90	0.86	0.92	0.79	0.66	0.02
SVM_AA(pol)	AN	46 (28/18)	0.88	0.89	0.87	0.75	0.40	0.06
SVM_AA(sig)	AN	54 (42/12)	0.85	0.93	0.80	0.72	0.39	0.08
SVM_AA(lin)	AN	54 (42/12)	0.85	0.93	0.80	0.72	0.40	0.08
DT	AN	89 (14/75)	0.76	0.54	0.93	0.52	-	-
SAM	NA	125 (26/99)	0.70	0.51	0.88	0.41	1	0.12
SAM(w0.5)[2]	NA	145 (36/109)	0.65	0.46	0.83	0.31	1	0.24
SVM_Fisher	NA	98 (25/73)	0.76	0.64	0.88	0.53	1	0.04
SVM_Fisher (w0.5)[2]	NA	91 (46/45)	0.78	0.78	0.78	0.56	0.88	0.06
SVM_pairwise	NA	25 (14/11)	0.94	0.95	0.93	0.88	0.14	0.00
SVM_AA(rbf)	NA	36 (29/7)	0.91	0.97	0.86	0.83	0.31	0.00
SVM_AA(pol)	NA	23 (15/8)	0.94	0.96	0.93	0.89	0.20	0.01
SVM_AA(sig)	NA	35 (25/10)	0.91	0.95	0.88	0.83	0.30	0.01
SVM_AA(lin)	NA	35 (25/10)	0.91	0.95	0.88	0.83	0.30	0.01
DT	NA	89 (20/69)	0.78	0.66	0.90	0.58	-	-

[1]AN: trained on a Class A dataset and tested on a non-Class A dataset; NA: trained on a non-Class A dataset and tested on a Class A dataset. Class A and non-Class A datasets included 410 and 372 sequences, respectively.

[2]Results obtained using w0.5 of the SAM package.

Table A2.3: Classifier performance for Class A within-family tests.[1]

Methods	Family[2]	Errors (FP/FN)	Accu-racy	Sensi-tivity	Speci-ficity	MCC	MaxRFP	MedRFP
SAM	AR	0 (0/0)	1	1	1	1	0	0
SVM_pairwise	AR	0 (0/0)	1	1	1	1	0	0
SVM_AA(rbf)	AR	19 (6.5/12.5)	0.91	0.80	0.96	0.78	0.49	0
SAM	PE	0 (0/0)	1	1	1	1	0	0
SVM_pairwise	PE	1.5 (1/0.5)	0.99	0.99	0.99	0.98	0.01	0
SVM_AA(rbf)	PE	25.5 (12/13.5)	0.89	0.81	0.92	0.73	0.54	0.01
SAM	OL	0 (0/0)	1	1	1	1	0	0
SVM_pairwise	OL	0 (0/0)	1	1	1	1	0	0
SVM_AA(rbf)	OL	6.5 (1/5.5)	0.98	0.96	0.99	0.96	0.06	0

[1]Values shown are the average from two independent tests.
[2]The Class A family dataset used to train and test each classifier.

Table A2.4: Classification performance of GPCRHMM against various datasets. [1]

Datasets (no. samples) [2]	Errors (FP/FN)	Accu-racy	Sensi-tivity	Speci-ficity	MCC	MaxRFP	MedRFP
Class A Training (410)	0 (0/0)	1	1	1	1	0	0
Class A Test (410)	2 (0/2)	1.00	0.99	1	0.99	0.04	0
Non-Class A (372)	4 (0/4)	0.99	0.98	1	0.98	0.03	0
AR (126)	1 (-/1)	1.00	-	-	-	-	-
PE (139)	1 (-/1)	1.00	-	-	-	-	-
OL (309)	0 (-/0)	1	-	-	-	-	-
N1 (158)	70 (-/70)	0.44	-	-	-	-	-
N2 (158)	70 (-/70)	0.44	-	-	-	-	-

[1]All statistics were obtained at MEP.
[2]Class A and non-Class A datasets include both positive (GPCR) and negative (non-GPCR) samples (see Table 2.2). AR, PE, and OL datasets include only Class A GPCR samples, and N1 and N2 datasets include only non-Class A GPCR samples (see Table 2.4).

Chapter 3

Mining the *Arabidopsis thaliana* Genome for Highly-divergent Seven Transmembrane Receptors

3.0 Preface for Chapter 3

In this chapter multiple protein classification methods, including both alignment-based and alignment-free classifiers, were combined to identify divergent seven-transmembrane receptor (7TMR) candidates from the *Arabidopsis thaliana* genome. Inclusion of both types of classifiers resolved problems in optimally training individual classifiers using limited and divergent samples, and increased stringency for candidate proteins. The methods included the ones I studied in the previous chapter as well as some new ones. I was involved in the training data preparation and prediction of candidate 7TMRs using profile hidden Markov models, and support vector machines with amino acid composition and dipeptide composition. We identified 394 proteins as 7TMR candidates and highlighted 54 with corresponding expression patterns for further investigation. This chapter has been published in:

Moriyama, E. N., Strope, P. K., Opiyo, S. O., Chen, Z. and Jones, A. M. (2006) Mining the *Arabidopsis thaliana* genome for highly-divergent seven transmembrane receptors. *Genome Biology* **7:** R96.

3.1 Background

Seven-transmembrane (7TM)-region containing proteins constitute the largest receptor superfamily in vertebrates and other metazoans. These cell-surface receptors are activated by a diverse array of ligands, and are involved in various signaling processes such as cell proliferation, neurotransmission, metabolism, smell, taste, and vision. They are the central players in eukaryotic signal transduction. They are commonly referred to as G protein-coupled receptors (GPCRs) because most transduce extracellular signals into cellular physiological responses through the activation of heterotrimeric guanine nucleotide binding proteins (G proteins) [1]. However, an increasing number of alternative "G protein-independent" signaling mechanisms have been associated with groups of these 7TM proteins [2-5]. Thus, for precision and clarity, we refer to these proteins simply as 7TM receptors (7TMRs), and candidate proteins in organisms greatly divergent to humans are designated here as 7TM putative receptors (7TMpRs).

The human genome encodes approximately 800 or more 7TMR, both with known cognate ligands and without or so-called orphan GPCRs, thus, constituting >1% of the gene complement [6]. More than 1,000 genes or 5% of the *Caenorhabditis elegans* genome are predicted to encode 7TMRs; the majority of them appear to be chemoreceptors [7]. Approximately 300 7TMR-encoding genes (about 1-2% of the genome) have been recognized in the *Drosophila melanogaster* genome [6]. Compared to such large numbers of 7TMRs found in animal genomes, very few 7TMpRs have been reported in plants and fungi. Only 22 Arabidopsis 7TMpRs have been described so far. Fifteen of them constitute the "mildew resistance O" (MLO) family, whose direct interaction with G-protein a subunit (Ga) has not been shown [8, 9]. While another 7TMpR, GCR1 [10], directly interacts with the plant Ga subunit GPA1 [11], it has been shown that GCR1 can act independently of the

heterotrimeric G-protein complex as well [2]. Hsieh and Goodman [12] recently reported five expressed proteins predicted to have 7 TM regions (heptahelical transmembrane proteins 1-5 or HHP1-5) but these like the other 16 do not have candidate ligands. Finally, an unusual regulator of G signaling protein (AtRGS1) has been predicted to have 7 TM regions [13]. RGS proteins function as a GTPase activating protein (GAP) to de-sensitize signaling by de-activating the Gα subunits of the heterotrimeric complex. Because Arabidopsis seedlings lacking AtRGS1 have reduced sensitivity to D-glucose [2, 13, 14], the possibility exists that AtRGS1 is a novel D-glucose receptor having an agonist-regulated GAP function. Although we designate them 7TMpRs here, it should be noted that neither a ligand nor a full signaling cascade has been demonstrated yet for any of these plant proteins and only for a barley MLO, the 7TM topology was experimentally confirmed [8].

None of the reported Arabidopsis 7TMpR proteins share substantial sequence similarity to known metazoan GPCRs constituting six different subfamilies. It appears that plant 7TMpRs dramatically diverged from known metazoan GPCRs over the 1.6 billion years since the plant and metazoan lineages bifurcated. It should be noted that Arabidopsis GCR1 shares weak but significant similarity to the cyclic AMP receptor, CAR1, found in the slime mold [2, 10, 15]. There is also very weak similarity to the Class B Secretin family GPCRs. However, other than GCR1, currently used search methods have not robustly identified plant 7TMpR proteins as candidate GPCRs. This great sequence divergence highlights the need for new approaches to identify divergent 7TMR candidates in non-metazoan genomes.

The human genome contains 16 Gα, 5 Gβ, and 12 Gγ genes. In stark contrast, both fungi and plants have much simpler G-protein coupled signaling systems. For example, the Arabidopsis genome contains one canonical Gα, one Gβ, and two Gγ genes [16]. Similarly a

small number of G-proteins are found in fungi; there are two Gα, one Gβ, and one Gγ in *Saccharomyces cerevisiae* [17-19] while *Neurospora crassa* and some fungi have more of each subunit genes [20-22]. Therefore, it may be reasonable to assume that plants and fungi have fewer GPCRs than human, and while ~200 Arabidopsis proteins were predicted to have 7 TM regions, sequence divergence precludes unequivocal assignment of any as an orphan GPCR [23]. However, at least 61 7TMpRs have been recently predicted from the plant pathogenic fungus *Magnaporthe grisea* genome [24], raising the possibility that more divergent groups of 7TMpR proteins likely remain undiscovered in non-metazoan taxa.

In this report, we describe our comprehensive computational strategy for identifying 7TMpR candidates from the entire protein sequence set predicted from the *A. thaliana* genome, and compile their tissue-specific expression and co-expression patterns with G-proteins. In order to take advantage of different approaches, we combined multiple protein classification methods including more specific (conservative) alignment-based classifiers and more sensitive alignment-free classifiers to predict candidate 7TMpRs in divergent genomes more effectively.

3.2 Results and Discussion

3.2.1 Identifying 7TMpR candidates using various protein classification methods

Among many protein classification methods commonly used, the current state-of-the-art and most used is the profile hidden Markov models (profile HMMs) [25]. It is used to construct protein family databases such as Pfam [26], SMART [27], and Superfamily [28].

However, profile HMMs and other currently used classification methods such as PROSITE [29] and PRINTS [30] share an important weakness. These methods rely on multiple alignments for generating their models (patterns, profile HMMs, *etc.*). Generating robust multiple alignments is difficult or impossible when extremely diverged sequences are included in the analysis. 7TMRs are one such protein family whose sequence similarities between subgroups can be lower than 25%. Furthermore, alignments are generated only from known related proteins (positive samples), and therefore no information from negative samples (unrelated protein sequences) is directly incorporated in the model building process. Identifiable "hits" are therefore constrained by initial sampling bias, which becomes reinforced when models are iteratively rebuilt from accumulated sequences. Consequently the predictive power, especially the sensitivity, of these classifiers decreases when they are applied against extremely diverged protein families.

In order to overcome this disadvantage and to increase sensitivities against such non-alignable similarities, several *alignment-free* methods have been proposed recently. These methods quantify various properties of amino acid sequences and convert them into a descriptor array. Once multiple sequences with different lengths are transformed into a uniform matrix, various multivariate analysis methods can be applied. Kim et al. [31] and Moriyama and Kim [32] used parametric and non-parametric discriminant function analysis methods. Karchin *et al.* [33] incorporated profile HMMs with support vector machines (SVMs) using the Fisher kernel (SVM-Fisher) so that negative sample information can be taken into account when training the classifier. SVMs can be applied with completely *alignment-free* sequence descriptors, *e.g.,* amino acid and dipeptide compositions. Such alignment-free classifiers are shown to outperform profile HMMs as well as Karchin *et al.*'s SVM-Fisher [34; Strope and Moriyama submitted]. Another multivariate method, partial

least squares (PLS) regression, was used by Lapinsh *et al.* [35] with physico-chemical properties of amino acids. We recently re-evaluated the descriptors used with PLS and optimized them to discriminate 7TMRs from other proteins [Opiyo and Moriyama submitted].

We applied these methods against the entire predicted protein sequence set derived from the *Arabidopsis thaliana* genome. As shown in Table 3.1, among the 28,952 protein sequences, SAM, a profile HMM method, predicted only 16 (excluding one alternatively spliced gene sequence) as 7TMpR candidates. Fifteen of them are identified as MLO or similar to MLO and one as GCR1 in The Arabidopsis Information Resource (TAIR) [36]. It clearly shows that SAM is highly specific (discriminating) with no false positive assuming that current annotations are correct. SAM failed to identify only one known MLO (MLO4: At1g11000). This protein as well as AtRGS1 and five recently-predicted 7TM proteins (HHP1-5) were of the 16 previously-predicted Arabidopsis 7TMpRs not included in the randomly sampled 500 7TMR training sequences (see Materials and methods). Thus, we concluded that the predictive power of SAM alone is insufficient to identify highly diverged and potentially novel 7TMpR sequences.

The results obtained by SAM were compared with those by alignment-free methods. As shown in Table 3.1, alignment-free methods (LDA, QDA, LOG, KNN, SVM-AA, SVM-di, and PLS-ACC) predicted 2,000 – 3,400 proteins as 7TMpR candidates, which is about 10% of the entire predicted Arabidopsis proteome and about 30-50% of the all possible transmembrane proteins (6,475 proteins) [23]. These alignment-free methods clearly call many false positives, and need further optimization to improve their discrimination power.

One advantage of alignment-free methods to be noted is their sensitivity against short or partial sequences [31, 32]. Many of the 28,952 protein sequences used in this study are based only on *ab-initio* gene prediction results, and hence are likely to contain various types of errors. If only a part of a 7TMR protein is predicted correctly, alignment-free methods could have a better chance to identify it.

Table 3.1 lists Arabidopsis proteins that were predicted to have 5-10 transmembrane regions and bins them by the number of transmembrane regions. Two hundred and one proteins were predicted by HMMTOP 2.0 [37] to have 7 TM regions. This number is close to a previous prediction (184 proteins) [23]. We should note, however, that no single method predicts exactly 7 TMs from all known 7TMRs (see Materials and methods). As mentioned above, it is also possible that some deduced Arabidopsis proteins we analyzed do not contain the entire coding region correctly. 952 Arabidopsis proteins were predicted to have five to nine TM regions. Based on the distribution of predicted TM numbers obtained from the entire GPCRDB entries, this range (5-9 TMs) could cover almost all of 7TMR candidates (99.1%; see Figure 3.1 and Materials and methods). The 22 previously-predicted Arabidopsis 7TMpRs were predicted to have seven to ten TM regions (Figure 3.1). If we extend the range to 5-10 TMs, the number of Arabidopsis 7TMpR candidates becomes 1,179 proteins.

3.2.2 Choosing 7TMpR candidates by combining prediction results

Among the ten alignment-free classifiers, LOG misclassified seven previously-predicted Arabidopsis 7TMpRs. KNN with *K* set at 5, 10, and 15 missed one, while KNN with *K* set at 20 classified them all correctly (See Materials and methods on KNN). In order to reduce the number of false positives (non-7TMRs predicted as 7TMRs) as well as false negatives (7TMRs predicted as non-7TMRs) and to obtain a set of 7TMpR candidates with

higher confidence, we examined combinations of the prediction results by the remaining six alignment-free methods (LDA, QDA, KNN with K=20, SVM-AA, SVM-di, and PLS-ACC). 652 proteins were predicted as 7TMpR candidates by all six methods (by choosing the strict intersection). Using the number of predicted TM regions to be 5-10, 394 (342 after removing duplicated entries due to alternative splicing) proteins were identified as 7TMR candidates. These Arabidopsis proteins are listed in Additional data file 1 (http://genomebiology.com/2006/7/10/R96/additional). Twenty of the 22 previously-predicted 7TMpRs were found in this list. Although HHP4 and HHP5 were not included in this list, both were identified by two of the alignment-free methods: KNN and SVM-AA. Note that RGS1 and five HHP (as well as nine MLO and GCR1) sequences were excluded from the training set, and these six were not identified as candidate 7TMpRs by SAM.

A further restriction to protein topology of exactly 7 TM regions and an N-terminus located extracellularly reduced the candidate number to 64 (54 excluding duplications due to alternative splicing). This set included nine of the 22 previously-predicted 7TMpRs. These 54 7TMpR candidates are the first targets for our further analysis and are summarized in Table 3.2 (also listed in Additional data file 2 http://genomebiology.com/2006/7/10/R96/additional). Eighteen are described as simply "expressed proteins" in the TAIR database (except for AT3G26090, which encodes RGS1). Interestingly, one of them (AT5G27210) is known to have weak similarity to a mouse orphan 7TMR. While others are known to belong to certain protein families (*e.g.,* nodlin MtN3 family), in many cases, their molecular functions have not identified, and further investigation on these 7TMpR candidates is warranted.

The 54 proteins were grouped into families based on similarities to known protein sequences. Eight of the 54 7TMpR candidates, including GCR1 and RGS1, are encoded by

single copy genes. In addition to the 7 MLO proteins identified, there are 8 nodulin MtN3 family members, 2 proteins of an unnamed family consisting of 6 expressed proteins, as well as multiple (2-3) members from smaller gene families (<=5). All members of the TOM3 family and the Perl1-like family, as well as the majority of the GNS/SUR4 family and an unnamed family consisting of 5 expressed proteins (expressed protein family 2) were included in the list. The identification of multiple members from these gene families using our alignment-free methods supported the consistency of this approach. However, for most of these families, not all members were found. Additionally, 8 single representatives of small protein families consisting of 2-5 members and 4 single representatives of large protein families were found in the list. Some of these proteins, especially those from large protein families, may represent false positives as 7TMpR candidates. This 7TMR mining method can be refined, for example, by re-training models as well as using more flexible hierarchical classification.

The five predicted heptahelical proteins (HHP1-5) reported by Hsieh and Goodman [12] were identified by sequence similarity to human adiponectin receptors (AdipoRs) and membrane progestin receptors (mPRs) that share little sequence similarity to known GPCRs. HHP1- 3 were identified in our initial list of 394 but were culled from the final list of 54 Arabidopsis 7TMpR candidates. This is because HMMTOP predicted HHP1, 2, 4, and 5 to have 7 TMs with the intracellular N-termini in contrast to known GPCRs. This unusual structural topology was also found in AdipoRs [12, 38]. HHP3 had 8 predicted TM regions. Eight of the 15 MLO proteins were also predicted to have 8-10 TM regions by HMMTOP (Figure 3.1). Recently, Benton *et al.* [39] experimentally showed that Drosophila odorant receptors, another extremely diverged 7TMR family, have intracellular N-termini. Among our 394 candidate list, 23 proteins were predicted to have 7 TM regions with intracellular N-

termini (Additional data file 1 http://genomebiology.com/2006/7/10/R96/additional). Therefore, we consider these 54 as a minimum working set of 7TMpR candidates, and many of the other proteins included in the list of 394 should be examined in the second stage.

3.2.3 Expression patterns of genes encoding the 7TMpR candidates and G-protein subunits

We utilized the Meta-Analyzer server of Genevestigator web site to study spatial expression patterns of Arabidopsis genes encoding the 7TMpR candidates and G-protein subunits. Note that the expression of MLO genes were not included in this analysis since we reported them recently [40]. As is shown in Figure 3.2, expression patterns of analyzed 7TMpR candidates can be divided into two major groups; about half of them show distinct tissue specificity, whereas the other half either exhibit less distinct expression patterns or display ubiquitous expression. All genes encoding G-protein subunits fall into the latter major group. Ubiquitous expression of genes encoding G-protein subunits allows overlap with genes in both groups, and makes, in principle, co-functioning of G-proteins with these 7TMpR candidates spatially and temporally possible. All 8 genes encoding the MtN3 family proteins appear to have distinct tissue specific expression. Among them, At3g48740 and At4g25010 have the highest sequence similarities to At5g23660 and At5g50800, respectively. Both pairs of genes share similar or overlapping expression patterns, suggesting relatedness/similarity of their functions. Confirming the actual functions of the 7TMpR candidates as GPCRs requires further extensive testing. A possible involvement of these candidate proteins in "G protein-independent" signaling mechanisms also needs to be explored.

3.3 Conclusions

We showed that the profile HMM protein classification method, currently one of the most used, is overly specific (conservative) when applied to extremely diverged 7TMpR proteins. Our premise is that there are more 7TMpRs yet to be identified in the *A. thaliana* and other genomes divergent to humans. The limitations were that the lack of available samples limits the effectiveness of profile HMM methods, and while alignment-free methods are more sensitive, they have high rates for false positives. The candidate 7TMpR proteins provided in this study, for example, can be included to expand the training set and re-iteration using refined training sets can be done in order to reduce false positive rates. However, this is possible only after these new candidates are confirmed as true positives experimentally.

The strategy we described here overcomes the "chicken-or-egg"; predictions by multiple protein classification methods and the number of predicted transmembrane regions were used to identify more likely and reduced number of 7TMR candidates. By setting up various methods as hierarchical multiple filters, one can prioritize target protein sets for further experimental confirmation of their functions.

3.4 Materials and methods

3.4.1 Arabidopsis protein data

28,952 protein sequences were downloaded from The Institute for Genomic Research (TIGR; *Arabidopsis thaliana* Database Release 5, dated on June 10, 2004) [41]. Among the 28,952 proteins, 2,760 are derived from alternative splicing.

3.4.2 Training data preparation for protein classification

Positive training samples (known 7TMR sequences) were obtained from GPCRDB (Information System for G Protein-Coupled Receptors, Release 9.0, last updated on June 28, 2005) [6]. In the GPCRDB, 2,030 7TMRs (originally collected from the Swiss-Prot protein database) were grouped into six major classes (Classes A - E plus the Frizzled/Smoothened family) and six putative families (ocular albinism proteins, insect odorant receptors, plant MLO receptors, nematode chemoreceptors, vomeronasal receptors, and taste receptors). Five hundred 7TMR sequences were randomly sampled and used as the positive samples. Note that "putative/unclassified" (orphan) 7TMRs and bacteriorhodopsins were not included in this dataset. These 500 7TMRs included six of the15 known Arabidopsis MLO proteins. Among the 22 currently known Arabidopsis 7TMpRs, in addition to the nine MLO proteins, GCR1 as well as six recently identified Arabidopsis 7TMpRs (AtRGS1 and HHP1-5; GPCRDB does not list these proteins) were not included in the random 500 7TMR samples. Note that the 15 Arabidopsis 7TMpRs not included in the training set can be used to assess the classifier performance as test cases.

For negative samples, 500 non-7TMR sequences longer than 100 amino acids were randomly sampled from the Swiss-Prot section of the UniProt Knowledgebase [42]. The average length of the 500 non-TMR sequences was 401 amino acids (with the maximum length of 2,512 amino acids). Positive and negative samples were combined to create a training dataset. Note that only positive samples were used to train the profile HMM classifier, SAM (see below).

3.4.3 Protein classification methods used

One alignment-based method (profile HMM) and four types of alignment-free multivariate methods were included in our analysis.

Profile hidden Markov models (profile HMMs). Profile HMMs are full probabilistic representation of sequence profiles [25]. Sample sequences need to be alignable, and thus only positive samples can be used for training. Two programs in Sequence Alignment and Modeling System (SAM version 3.4) [43] were used: *buildmodel* to build profile HMMs with the nine-component Dirichlet mixture priors [44], and *hmmscore* to calculate scores and e-values. The 'calibration' option (for more accurate e-value calculation) and the fully local scoring option (-sw 2) were used. The e-value threshold was set at 0.01 for choosing 7TMR candidates.

Discriminant function analysis. In Moriyama and Kim [32], we described three parametric (linear, quadratic, logistic) and nonparametric K-nearest neighbor methods that performed better than the profile HMM method. Therefore, we included these four alignment-free methods (LDA, QDA, LOG, and KNN) in our analysis. For KNN, K was set at 5, 10, 15, or 20, where K is the number of neighbors. The four variables used (amino acid index and three periodicity statistics) were described in Kim *et al*. [31]. S-PLUS statistical package (Insightful Corporation, version 6.1.2 for Linux) with the MASS module [45] was used for the classifier development.

Support vector machines with amino acid composition (SVM-AA). Support vector machines (SVMs) are learning machines that make binary classifications based on a

hyperplane separating a remapped instance space [46]. A kernel function can be chosen so that the remapped instances on a multidimensional feature space are linearly separable. The radial basis kernel, $\exp(-g||x-y||^2)$, was used in this study. The parameter g was set to 102 based on the median of Euclidean distances between positive examples and the nearest negative example as described in Jaakkola *et al.* [47]. Simple 19 amino acid frequencies (the 20th amino acid frequency can be explained completely by the other 19) of each protein sequence were used as an input vector for SVMs. Programs *svm_learn* and *svm_classify* of the SVM^{light} package version 5.0 [48] were used for training and classification by SVM, respectively. The default value of the regulatory parameter C (0.5006) was used with *svm-learn*. Our comparative analysis showed that SVM-AA performs better than profile HMMs when they are applied to remote similarity identification, the same problem we deal with in this study (Strope and Moriyama submitted).

Support vector machines with dipeptide composition (SVM-di). We also included an SVM classifier with dipeptide composition following Bhasin and Raghava [34]. The SVM^{light} package version 5.0 [48] were used for training and classification as before. The regulatory parameter C=1 and the radial basis kernel function parameter g=90 were chosen by the grid analysis using 5-fold cross-validation.

Partial least squares with amino acid properties (PLS-ACC). Partial least squares (PLS) regression is a projection method that takes into account correlations between independent and dependent variables [49]. We used the pls.pcr package, an R implementation developed by Ron Wehrens [50], with the SIMPLS method, four latent variables, and cross-validation options. Each amino acid in the protein sequences was first converted to a set of five

principal component scores developed from twelve physico-chemical properties. The auto/cross covariance (ACC) method developed by Wold *et al.* [51] was then applied to each of the converted sequences. ACC describes the average correlations between two residues a certain lag (amino acids) apart. The lag size of 30 was chosen for optimal classification performance. We found that the performance of PLS-ACC is robust even when only a small number of positive samples (5 or 10) are available for training. In contrast, the performance of profile HMMs suffered extremely when positive sample size was small. The twelve physico-chemical properties used and more details on the use of PLS in protein classification are described elsewhere (Opiyo and Moriyama submitted). The cutoff value of 0.4999 was used for choosing 7TMR candidates in this study, which was determined as the average of the minimum error points [33] obtained from 500 replications of 10-fold cross-validation analysis using the training dataset.

3.4.4 Transmembrane region prediction

HMMTOP 2.0 [37] and TMHMM [originally 52, implemented as S-TMHMM by 53] were used for predicting transmembrane regions. Figure 3.1 shows the numbers of TM regions predicted by the two methods for the 500 7TMR sequences used for classifier training. HMMTOP predicted 7 TMs from 433 7TMRs (86.6%), while only 165 7TMRs (33%) were predicted to have 7 TMs by TMHMM. HMMTOP predicted 97% or more of 7TMRs to have 6-8 TMs, and with 5-9 TMs more than 99% of 7TMRs were included. Using TMHMM, in order to include 97% of 7TMRs, the range of predicted TM numbers needs to be between 4 and 10. Therefore, we decided to use HMMTOP in our further analysis. With HMMTOP using the range of 5-9 TMs, we should be able to cover almost all possible 7TM proteins.

3.4.5 Grouping of the candidate proteins

The candidate proteins were grouped based on the e-values obtained by BLASTP protein similarity search [54] against the Arabidopsis protein database using the default parameter set (*e.g.,* BLOSUM62) at the Arabidopsis Information Resource (TAIR) web site [55]. The e-value threshold of 10^{-20} was used to identify protein families similar to the candidate proteins.

3.4.6 Expression patterns of genes encoding 7TMR candidates and G-protein subunits

Expression patterns of genes encoding 7TMpR candidates and G-protein subunits among tissues was studied by using the Meta-Analyzer server of the Genevestigator web site (last updated in Nov. 2005) [56]. All data were generated using the 22K Affymetrix ATH1 Arabidopsis Genome array. Gene expression profiles based on microarray data were clustered according to similarity in expression patterns. Hierarchical clustering results were generated by default settings using pairwise Euclidean distances and the average linkage method.

3.5 References

1. Pierce KL, Premont RT, Lefkowitz RJ: **Seven-transmembrane receptors**. *Nat Rev Mol Cell Biol* 2002, **3**:639-650

2. Chen JG, Pandey S, Huang J, Alonso JM, Ecker JR, Assmann SM, Jones AM: **GCR1 can act independently of heterotrimeric G-protein in response to brassinosteroids and gibberellins in Arabidopsis seed germination**. *Plant Physiol* 2004, **135**:907-915

3. Kimmel AR, Parent CA: **The signal to move: *D. discoideum* go orienteering**. *Science* 2003, **300**:1525-1527

4. Lefkowitz RJ, Shenoy SK: **Transduction of receptor signals by beta-arrestins**. *Science* 2005, **308**:512-517

5. Kristiansen K: **Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function**. *Pharmacol Ther* 2004, **103**:21-80

6. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G: **GPCRDB information system for G protein-coupled receptors**. *Nucleic Acids Res* 2003, **31**:294-297. [http://www.gpcr.org/7tm/]

7. Bargmann CI: **Neurobiology of the *Caenorhabditis elegans* Genome**. *Science* 1998, **282**:2028-2033

8. Devoto A, Piffanelli P, Nilsson I, Wallin E, Panstruga R, von Heijne G, Schulze-Lefert P: **Topology, subcellular localization, and sequence diversity of the Mlo family in plants**. *J Biol Chem* 1999, **274**:34993-35004

9. Devoto A, Hartmann HA, Piffanelli P, Elliott C, Simmons C, Taramino G, Goh CS, Cohen FE, Emerson BC, Schulze-Lefert P *et al*: **Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family**. *J Mol Evol* 2003, **56**:77-88

10. Josefsson LG, Rask L: **Cloning of a putative G-protein-coupled receptor from *Arabidopsis thaliana***. *Eur J Biochem* 1997, **249**:415-420

11. Pandey S, Assmann SM: **The Arabidopsis putative G protein-coupled receptor GCR1 interacts with the G protein alpha subunit GPA1 and regulates abscisic acid signaling**. *Plant Cell* 2004, **16**:1616-1632

12. Hsieh M-H, Goodman HM: **A novel gene family in Arabidopsis encoding putative heptahelical transmembrane proteins homologous to human adiponectin receptors and progestin receptors**. *J Exp Bot* 2005, **56**:3137-3147

13. Chen J-G, Willard FS, Huang J, Liang J, Chasse SA, Jones AM, Siderovski DP: **A seven-transmembrane RGS protein that modulates plant cell proliferation.** *Science* 2003, **301**:1728-1731

14. Ullah H, Chen JG, Wang S, Jones AM: **Role of a heterotrimeric G protein in regulation of Arabidopsis seed germination**. *Plant Physiol* 2002, **129**:897-907

15. Josefsson LG: **Evidence for kinship between diverse G-protein coupled receptors**. *Gene* 1999, **239**:333-340

16. Jones AM, Assmann SM: **Plants: the latest model system for G-protein research**. *Embo Rep* 2004, **5**:572-578

17. Nakafuku M, Itoh H, Nakamura S, Kaziro Y: **Occurrence in Saccharomyces cerevisiae of a gene homologous to the cDNA coding for the alpha subunit of mammalian G proteins**. *Proc Natl Acad Sci U S A* 1987, **84**:2140-2144

18. Nakafuku M, Obara T, Kaibuchi K, Miyajima I, Miyajima A, Itoh H, Nakamura S, Arai K, Matsumoto K, Kaziro Y: **Isolation of a second yeast Saccharomyces cerevisiae gene (GPA2) coding for guanine nucleotide-binding regulatory protein: studies on its structure and possible functions**. *Proc Natl Acad Sci U S A* 1988, **85**:1374-1378

19. Whiteway M, Hougan L, Dignard D, Thomas DY, Bell L, Saari GC, Grant FJ, O'Hara P, MacKay VL: **The STE4 and STE18 genes of yeast encode potential beta and gamma subunits of the mating factor receptor-coupled G protein**. *Cell* 1989, **56**:467-477

20. Baasiri RA, Lu X, Rowley PS, Turner GE, Borkovich KA: **Overlapping functions for two G protein alpha subunits in Neurospora crassa**. *Genetics* 1997, **147**:137-145

21. Turner GE, Borkovich KA: **Identification of a G protein alpha subunit from Neurospora crassa that is a member of the Gi family**. *J Biol Chem* 1993, **268**:14805-14811

22. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*: **The genome sequence of the filamentous fungus *Neurospora crassa***. *Nature* 2003, **422**:859-868

23. Schwacke R, Schneider A, van der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugge UI, Kunze R: **ARAMEMNON, a novel database for Arabidopsis integral membrane proteins**. *Plant Physiol* 2003, **131**:16-26. [http://aramemnon.botanik.uni-koeln.de]

24. Kulkarni R, Thon M, Pan H, Dean R: **Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus *Magnaporthe grisea***. *Genome Biol* 2005, **6**:R24

25. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**. Cambridge: Cambridge University Press; 1998.

26. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-141. [http://pfam.wustl.edu/]

27. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32**:D142-144. [http://smart.embl.de/]

28. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919

29. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database**. *Nucleic Acids Res* 2006, **34**:D227-230. [http://www.expasy.org/prosite/]

30. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P *et al*: **PRINTS and its automatic supplement, prePRINTS**. *Nucleic Acids Res* 2003, **31**:400-402. [http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/]

31. Kim J, Moriyama EN, Warr CG, Clyne PJ, Carlson JR: **Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties**. *Bioinformatics* 2000, **16**:767-775

32. Moriyama EN, Kim J: **Protein family classification with discriminant function analysis**. In: *Genome Exploitation: Data Mining the Genome.* Edited by Gustafson JP, Shoemaker R, Snape JW. New York: Springer; 2005.

33. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines**. *Bioinformatics* 2002, **18**:147-159

34. Bhasin M, Raghava GP: **GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors**. *Nucleic Acids Res* 2004, **32**:W383-389. [http://www.imtech.res.in/raghava/gpcrpred/]

35. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES: **Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences**. *Protein Sci* 2002, **11**:795-805

36. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community**. *Nucleic Acids Res* 2003, **31**:224-228. [http://www.arabidopsis.org]

37. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server**. *Bioinformatics* 2001, **17**:849-850. [http://www.enzim.hu/hmmtop]

38. Yamauchi T, Kamon J, Ito Y, Tsuchida A, Yokomizo T, Kita S, Sugiyama T, Miyagishi M, Hara K, Tsunoda M *et al*: **Cloning of adiponectin receptors that mediate antidiabetic metabolic effects**. *Nature* 2003, **423**:762-769

39. Benton R, Sachse S, Michnick SW, Vosshall LB: **Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo**. *PLoS Biol* 2006, **4**:e20

40. Chen Z, Hartmann HA, Wu MJ, Friedman EJ, Chen JG, Pulley M, Schulze-Lefert P, Panstruga R, Jones AM: **Expression analysis of the AtMLO Gene Family Encoding Plant-Specific Seven-Transmembrane Domain Proteins**. *Plant Mol Biol* 2006, **60**:583-597

41. **The Institute for Genomic Research (TIGR) Arabidopsis thaliana Database ftp site** [ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep.gz]

42. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **The Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2005, **33**:D154-159. [http://www.uniprot.org]

43. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: Extension and analysis of the basic method**. *Comput Appl Biosci* 1996, **12**:95-107. [http://www.cse.ucsc.edu/research/compbio/sam.html]

44. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology**. *Comput Appl Biosci* 1996, **12**:327-345

45. **S-plus MASS module** [http://www.stats.ox.ac.uk/pub/MASS4/]

46. Vapnik VN: **The Nature of Statistical Learning Theory**, 2nd edn. New York: Springer-Verlag; 1999.

47. Jaakkola T, Diekhans M, Haussler D: **A discriminative framework for detecting remote protein homologies**. *J Comput Biol* 2000, **7**:95-114

48. Joachims T: **Making large-Scale SVM Learning Practical**. In: *Advances in Kernel Methods - Support Vector Learning.* Edited by Schölkopf B, Burges C, Smola A. Cambridge: MIT Press; 1999: 169-184.

49. Geladi P, Kowalski BR: **Partial least squares regression: A tutorial**. *Anal Chim Acta* 1986, **185**:1-17

50. **The Comprehensive R Archive Network** [http://cran.r-project.org/]

51. Wold S, Jonsson J, Sjostrom M, Sandberg M, Rannar S: **DNA and Peptide Sequences and Chemical Processes Multivariately Modeled by Principal Component Analysis and Partial Least-Squares Projections to Latent Structures**. *Anal Chim Acta* 1993, **277**:239-253

52. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182

53. Viklund H, Elofsson A: **Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information**. *Protein Sci* 2004, **13**:1908-1917

54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410. [http://www.ncbi.nlm.nih.gov/BLAST/]

55. **The Arabidopsis Information Resource (TAIR)** [http://www.arabidopsis.org]

56. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W: **GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox**. *Plant Physiol* 2004, **136**:2621-2632. [https://www.genevestigator.ethz.ch]

57. ***Arabidopsis thaliana* 7TMR Mining** [http://bioinfolab.unl.edu/emlab/at7tmr/index.html]

Table 3.1. Numbers of 7TMpR candidates identified by various methods from the *A. thaliana* genome

Methods	Numbers of 7TMpR candidates [1]
HMMTOP (7TMs) [2]	236 (201)
(6-8 TMs) [2]	633 (545)
(5-9 TMs) [2]	1,091 (957)
(5-10 TMs) [2]	1,343 (1,179)
SAM	16 (15)
LDA	3,211 (2,935)
QDA	2,006 (1,820)
LOG	2,626 (2,394)
KNN (*K*=5)	3,125 (2,839)
KNN (*K*=10)	3,202 (2,906)
KNN (*K*=15)	3,298 (3,004)
KNN (*K*=20)	3,347 (3,043)
SVM-AA	2,263 (2,043)
SVM-di	2,004 (1,807)
PLS-ACC	2,671 (2,466)

[1]The numbers in parentheses show 7TMpR candidates after removing proteins derived from alternative splicing.
[2]The numbers of TM regions predicted by HMMTOP.

Table 3.2. Summary of the 54 7TMpR candidates identified in this study[1]

Groups[2]	TAIR locus IDs
[Multiple members from gene families]	
Nodulin MtN3 family proteins (8/17)	At1g21460, At3g16690, At3g28007, At3g48740, At4g25010, At5g13170, At5g23660, At5g50800
MLO proteins (7/15)	At1g11000 (MLO4), At1g26700 (MLO14), At1g42560 (MLO9), At2g33670 (MLO5), At2g44110 (MLO15), At4g24250 (MLO13), At5g53760 (MLO11)
Expressed protein family 1 (2/6)	At1g77220, At4g21570
GNS1/SUR4 membrane family proteins (3/4)	At1g75000, At3g06470, At4g36830
Perl1-like family protein (2/2)	At1g16560, At5g62130
TOM3 family proteins (3/3)	At1g14530, At2g02180, At4g21790
Expressed protein family 2 (3/5)	At1g10660, At2g47115, At5g62960
Expressed protein family 3 (2/4)	At3g09570, At5g42090
Expressed protein family 4 (2/5)	At1g49470, At5g19870
Expressed protein family 5 (2/5)	At3g63310, At4g02690
Single copy genes (8)	At1g48270 (GCR1), At1g57680, At2g41610, At2g31440, At3g04970, At3g26090 (RGS1), At3g59090, At4g20310
Single member from small gene families (8)	At2g01070, At3g19260, At2g35710, At2g16970, At1g15620, At1g63110, At4g36850, At5g27210
Single member from big gene families (4)	At1g71960, At3g01550, At5g23990, At5g37310

[1]See Additional data file 2 (http://genomebiology.com/2006/7/10/R96/additional) for more detailed information.
[2]The number of candidates identified in this study belonging to each group is shown in parentheses (the number of all proteins in each group is given after '/').

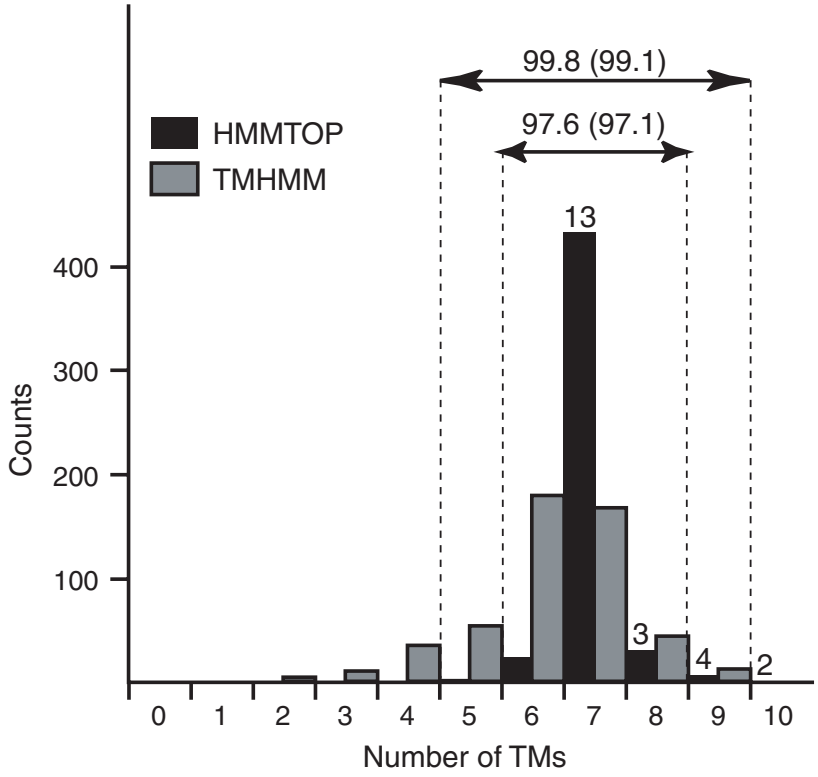

Figure 3.1. Distribution of transmembrane numbers predicted by HMMTOP (black bars) and TMHMM (gray bars) from the 500 7TMR sample sequences. Proportions (%) of the proteins predicted to have 6-8 and 5-9 TMs by HMMTOP are shown at the top. The percentages shown in parentheses were obtained from the entire 7,674 7TMR dataset in GPCRDB. The numbers shown on the top of black bars are the number of the previously-predicted 22 *Arabidopsis* 7TMpR proteins.

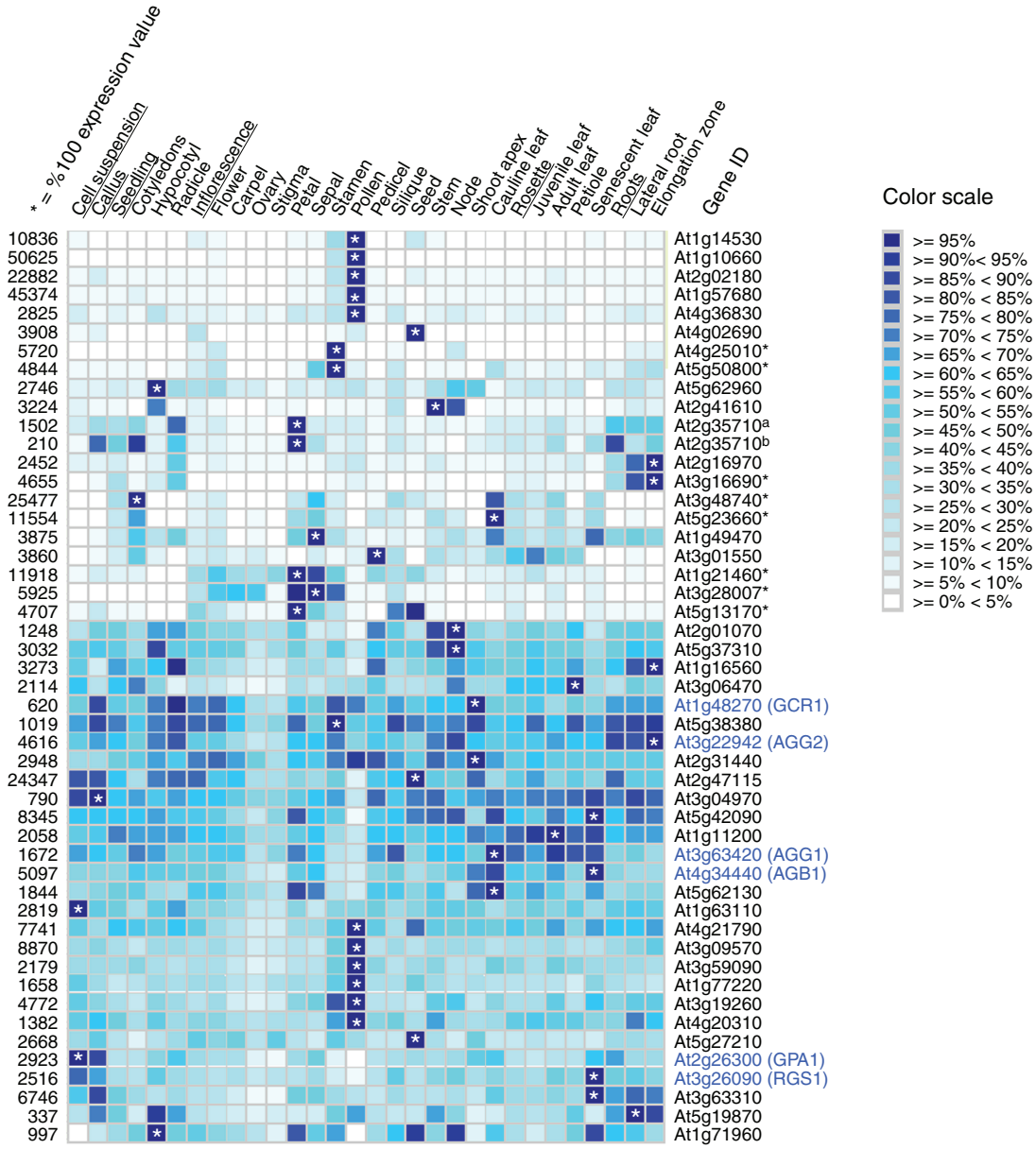
* = %100 expression value
Cell suspension
Callus
Seedling
Cotyledons
Hypocotyl
Radicle
Inflorescence
Flower
Carpel
Ovary
Stigma
Petal
Sepal
Stamen
Pollen
Pedicel
Silique
Seed
Stem
Node
Shoot apex
Cauline leaf
Rosette
Juvenile leaf
Adult leaf
Petiole
Senescent leaf
Roots
Lateral root
Elongation zone
Gene ID
10836 At1g14530
50625 At1g10660
22882 At2g02180
45374 At1g57680
2825 At4g36830
3908 At4g02690
5720 At4g25010*
4844 At5g50800*
2746 At5g62960
3224 At2g41610
1502 At2g35710a
210 At2g35710b
2452 At2g16970
4655 At3g16690*
25477 At3g48740*
11554 At5g23660*
3875 At1g49470
3860 At3g01550
11918 At1g21460*
5925 At3g28007*
4707 At5g13170*
1248 At2g01070
3032 At5g37310
3273 At1g16560
2114 At3g06470
620 At1g48270 (GCR1)
1019 At5g38380
4616 At3g22942 (AGG2)
2948 At2g31440
24347 At2g47115
790 At3g04970
8345 At5g42090
2058 At1g11200
1672 At3g63420 (AGG1)
5097 At4g34440 (AGB1)
1844 At5g62130
2819 At1g63110
7741 At4g21790
8870 At3g09570
2179 At3g59090
1658 At1g77220
4772 At3g19260
1382 At4g20310
2668 At5g27210
2923 At2g26300 (GPA1)
2516 At3g26090 (RGS1)
6746 At3g63310
337 At5g19870
997 At1g71960
Color scale
>= 95%
>= 90%< 95%
>= 85% < 90%
>= 80% < 85%
>= 75% < 80%
>= 70% < 75%
>= 65% < 70%
>= 60% < 65%
>= 55% < 60%
>= 50% < 55%
>= 45% < 50%
>= 40% < 45%
>= 35% < 40%
>= 30% < 35%
>= 25% < 30%
>= 20% < 25%
>= 15% < 20%
>= 10% < 15%
>= 5% < 10%
>= 0% < 5%

Figure 3.2. Expression patterns of *Arabidopsis* genes encoding 7TMpR candidates and G-protein subunits among tissues. The figure was modified from an output of the Meta-Analyzer of Genevestigator (last updated in Nov. 2005), which illustrates expression levels of each gene in different organs. Relative expression levels of a gene in different organs/tissues are given as heat maps in blue-scale coding that reflects absolute signal values, where darker colors represent stronger expression. All gene-level profiles are normalized for coloring such that for each gene the highest signal intensity obtains value 100% (shown in the darkest blue and marked with *) and absence of signal obtains value 0 % (shown in white). Probe-sets of five 7TMpR candidates (At1g15620. At1g75000, At4g21570, At4g36850, and At5g23990) were not present in the 22K chip, and therefore their tissue-specific expression could not be assessed. For At2g35710, two probe-sets (265797_at[a] and 265841_at[b]) were designed on the chip. Gene names for those belonging to the MtN3 family are shown in boldface and marked with *. Genes encoding G-protein subunits (*AGB1*, *GPA1*, *AGG1*, and *AGG2*) as well as two reported 7TMpRs (*RGS1* and *GCR1*) are labeled accordingly in boldface.

Chapter 4

Molecular evolution of urea amidolyase and urea carboxylase in fungi

4.0 Preface to Chapter 4

In this chapter, I studied the molecular evolution of related multi-domain protein families: urea amidolyase and urea carboxylase in both eukaryotes and prokaryotes. Urea amidolyase contains two major domains: the amidase and urea carboxylase domains. A shorter form of urea amidolyase is known as urea carboxylase and has no amidase domain. In order to elucidate the evolutionary origin of urea amidolyase and urea carboxylase, we studied the distribution of these enzymes across kingdoms. Phylogenetic analysis showed that these two enzymes appeared to have gone through independent evolution since their bacterial origin. The amidase domain and the urea carboxylase domain sequences from fungal urea amidolyases clustered strongly together with the amidase and urea carboxylase sequences, respectively, from a small number of beta- and gammaproteobacteria. On the other hand, fungal urea carboxylase proteins clustered together with another copy of urea carboxylases distributed broadly among bacteria. We concluded that the urea amidolyase genes currently found only in fungi are the results of a horizontal gene transfer event from beta-, gamma-, or related species of proteobacteria. Urea carboxylase genes currently found in fungi and other limited organisms were also likely derived from another ancestral gene in bacteria. Our study presented another important example showing plastic and opportunistic genome evolution in bacteria and fungi and their evolutionary interplay. This study has been published in:

Strope, P. K., Nickerson, K. W., Harris, S. D. and Moriyama, E. N. (2011) Molecular evolution of urea amidolyase and urea carboxylase in fungi. *BMC Evolutionary Biology* **11**: 80.

4.1 Background

Fungi exhibit great metabolic flexibility in the diversity of carbon and nitrogen sources they can use. We have been especially interested in their nitrogen sources, most recently urea [1, 2]. In a previous study [1], a dichotomy was observed with regard to urea utilization in fungi. Hemiascomycetes (yeasts and yeast-like fungi; the majority belongs to the class Saccharomycetes of the phylum Ascomycota) possess the urea amidolyase (*DUR1,2*; Degradation of URea) genes whereas all other fungi examined possess the nickel-containing urease sequences. Urea amidolyase is an energy dependent biotin-containing enzyme. It is encoded by the *DUR1,2* gene and was first characterized in the yeast *Candida utilis*, now known as *Pichia jadinii* [3]. The activity of this enzyme was also detected in green algae such as *Asterococcus superbus* and *Chlamydomonas reinhardii*. Urease and urea amidolyase activities were not observed together in the same green algal species; it was either one or the other [4, 5]. This cytoplasmic, biotin-dependent enzyme [6] consists of a single polypeptide chain with regions for urea carboxylase (EC 6.3.4.6) and allophanate hydrolase (also known as amidase; EC 3.5.1.54) activity. Two adjacent genes (*DUR1* and *DUR2*) were originally considered to encode the two enzymes; but later they were renamed as a single gene, *DUR1,2* [7].

Urea amidolyase breaks down urea into ammonia and carbon dioxide in a two-step process, while urease (EC 3.5.1.5) does this in a one-step process [1] as shown in the following equations:

[Urea carboxylase] $\quad$ urea + ATP + HCO_3^- → allophanate + ADP + P_i $\quad$ (i)

[Allophanate hydrolase (amidase)] $\quad$ allophanate → $2NH_3 + 2CO_2$ $\quad$ (ii)

[Urease] $\quad$ urea → $2NH_3 + CO_2$ $\quad$ (iii)

There are two forms of urea amidolyase proteins. Figure 4.1 shows the domain structure of urea amidolyase and related proteins. A shorter form of urea amidolyase is known as urea carboxylase, and has no amidase domain attached to it. This protein is found in several fungal species [1], green algae [8], and has been also characterized in bacteria [9].

The urea carboxylase protein (as well as the domain) is further divided into sub-domains: the biotin-carboxylation domain, allophanate hydrolase subunit 1 (AHS1) domain, allophanate hydrolase subunit 2 (AHS2) domain, and the biotin-lipoyl domain (Figure 4.1). The function of the AHS1 and AHS2 domains is still unknown. The biotin-carboxylation domain and the biotin-lipoyl domain of urea carboxylase are commonly found in various other carboxylases including pyruvate carboxylase (Pyc), methylcrotonoyl-CoA carboxylase (MccA), acetyl-CoA carboxylase (Acc1), and propionyl-CoA carboxylase (PccA) [10].

In Navarathna *et al.* [1], we suggested that urea amidolyase likely arose before the divergence of the hemiascomycetes and the euascomycetes (filamentous fungi; the subphylum Pezizomycotina of the phylum Ascomycota), *c.* 350 - 400 million years ago, by insertion of a gene encoding allophanate hydrolase into a methylcrotonyl CoA carboxylase (*mccA*) gene, thus creating *DUR1,2* and inactivating *mccA*. This suggestion was made because of the corresponding dichotomies: the hemiascomycetes have *DUR1,2* but do not have *mccA* whereas the rest of the fungi have both urease and *mccA* [1]. The present paper investigates the evolutionary origin of *DUR1,2,* the urea amidolyase gene, more thoroughly. We studied the distribution of urea amidolyase, urea carboxylase, and urease proteins in various species across all kingdoms, and biotin-carboxylation domain containing proteins, *i.e.*, Acc1, Pyc, PccA, and MccA, in various fungal species. Contrary to our previous speculation, an ancestral urea amidolyase gene likely arose in bacteria and then appeared in the fungal lineage before the divergence of the subphyla Pezizomycotina and

Saccharomycotina by prokaryote-to-eukaryote horizontal gene transfer. There have been studies indicating such bacteria-to-fungi horizontal transfers [*e.g.,* 11, 12-15]. Our study adds yet another important example showing evolutionary interplays between bacteria and fungi and how plastic and opportunistic the fungal genome evolution can be.

4.2 Results and Discussion

4.2.1 Urea amidolyase is unique to the kingdom fungi among eukaryotes

We have previously shown that long and short forms of urea amidolyase are present in fungi [1]. The urea amidolyase protein of the yeast *Saccharomyces cerevisiae* (phylum Ascomycota; subphylum Saccharomycotina) is 1,835 amino acids (aa) long. As shown in Figure 4.1, the first 632-aa region in the N-terminus of the protein consists of the amidase domain. The remainder of the sequence is the urea carboxylase domain, which consists of four smaller sub-domains. As mentioned before, the shorter form of urea amidolyase lacks the amidase domain and the urea carboxylase domain exists as a whole protein. This urea carboxylase sequence (1,241 aa) has been identified from a filamentous fungus *Aspergillus nidulans* (phylum Ascomycota; subphylum Pezizomycotina). Using these protein and domain sequences, we first examined if these two forms of urea amidolyase exist in eukaryotes outside of the fungal kingdom.

As shown in Table 4.1 (see also Table A4.1), urea amidolyase is absent in non-fungal eukaryotic genomes we examined. Blastp similarity search against the NCBI non-redundant (nr) database also showed no sequence similar to urea amidolyase from any other eukaryotic species. However, urea carboxylase and amidase genes are present in all four green algae we examined. In three of the four green algae, the amidase genes are located near the urea

carboxylase genes but not adjacent to them. The distance between these two genes ranged from 588 to 6,236 bp in these green algae (see Table A4.2). The absence of urea amidolyase gene but the presence of urea carboxylase and amidase genes in *C. reinhardtii* suggests that the activity of urea amidolyase seen previously in this species [3-5] is not due to the urea amidolyase protein but the combined activity of urea carboxylase and amidase proteins. Although we did not find sequences similar to urea carboxylase from any of the metazoan genomes we examined, similarity search against NCBI nr database turned up two sequences from Hydra (*Hydra magnipapillata*). One of them, however, was found actually to be a sequence of a putative bacterial symbiont. These Hydra sequences are discussed further later. No amidase sequence was found from Hydra or any other eukaryotes other than fungi and green algae.

Urease was found in both plant genomes we examined: *Arabidopsis thaliana* (a dicot) and *Oryza sativa* (a monocot). Similarity search against NCBI nr database also showed a wide distribution of urease in higher plants. While none of the green algal genomes we examined had urease (Table 4.1), it was identified in distantly related and more ancestral types of green algae (*Ostreococcus* and *Micromonas*) by searching against NCBI nr database. On the other hand, in metazoa, urease was found only in a limited number of genomes. In addition to *Nematostella vectensis* (a sea anemone, Table 4.1), only three metazoan urease sequences were found in the NCBI nr database (from *Strongylocentrotus purpuratus, Branchiostoma floridae,* and *Ixodes scapularis*). These observations are not consistent with what we observed earlier in fungi, where all fungi that lack urea amidolyase seemed to possess urease ([1]; also described next).

4.2.2 Distribution of urea amidolyase and other related proteins among fungi

We searched 64 fungal genomes for urea amidolyase, urea carboxylase, and amidase. For selected 27 fungal genomes, we further searched urease as well as proteins that share the biotin-carboxylation and the biotin-lipoyl domains (Acc1, Pyc, MccA, and PccA proteins; see Figure 4.1). These searches were conducted to examine the earlier hypothesis of Navarathna *et al.* [1] that the fungal urea amidolyase may have been formed by the extension of a biotin carboxylation gene that was already present in fungi.

Our search results are summarized in Tables 4.2 and A4.3 (see also Tables A4.4 and A4.5). The results are also mapped on the current consensus of the fungal phylogeny [16, 17] in Figure 4.2. Among the fungi we examined, only the class Sordariomycetes (subphylum Pezizomycotina; except for *Neurospora crassa* and its close relative in the order Sordariales) and the class Saccharomycetes (subphylum Saccharomycotina) had the urea amidolyase sequences. In one species, *Yarrowia lipolytica*, there were two copies of urea amidolyase. Urea carboxylase was found in many but not all of the species in the Pezizomycotina while being completely absent from the Saccharomycotina. Interestingly, except for *Fusarium graminearum* (known also as *Gibberella zeae*), the species belonging to the order Hypocreales (*Nectria, Fusarium,* and *Trichoderma*) had both the urea carboxylase and the urea amidolyase sequences. Many of these species are found in soils and associated with plants [18-20]. Dothideomycetes species did not have urea amidolyase, but many contained amidase as well as urea carboxylase sequences. However, the location of these two genes (amidase and urea carboxylase) was not near each other in their genomes. They were located in different scaffolds or supercontigs (see Table A4.2).

Consistent with the earlier observation [1], the urease protein was present in all the fungal species examined except for those of the Saccharomycotina. Two species (*F. graminearum* and *F. oxysporum*) had two copies of this protein. Previously only two Sordariomycetes species (*Magnaporthe oryzae*, previously known as *M. grisea*, and *F. graminearum*) were observed to possess both of urease and urea amidolyase. We now confirmed that all Sordariomycetes species except for *N. crassa* and closely related species have both of these enzymes.

Why do the Saccharomycetes species use the energy-dependent, biotin-containing urea amidolyase system and abandon the urease that accomplishes the same overall reaction in a simpler process? This question becomes even more germane when we consider that all strains of *C. albicans* are biotin auxotrophs [21], and it has long been known that 2 to 4 times as much biotin is required for maximum growth of *S. cerevisiae* on urea, allantoic acid, or allantoin as sole nitrogen sources [22]. However, the dichotomy in distribution of urease and urea amidolyase among some fungal lineages coincides precisely with that for the Ni/Co transporter (Nic1p), which is present in those fungi that use urease and absent in those that do not [23]. In Navarathna *et al.* [1], we suggested that the selective advantage of using urea amidolyase over using urease is that it allowed the Saccharomycetes species to jettison all Ni^{2+} and Co^{2+} dependent metabolisms and thus to have two fewer transition metals whose concentrations need to be regulated. However, while reasonable for the Saccharomycetes, such selective advantages may not be great enough to abandon the use of urease particularly in the Sordariomycetes species. Further investigation is needed to elucidate whether retaining two types of urea degradation enzymes in the Sordariomycetes species is in fact selectively advantageous rather than redundant.

We also examined the distribution of biotin-carboxylation domain containing enzymes. Acc1 and Pyc were present in all the fungal species we examined. MccA was absent almost completely from the Saccharomycetes and *Schizosaccharomyces pombe* (phylum Ascomycota; subphylum Taphrinomycotina), but was present in the rest of the fungi we examined. PccA was present in fewer species than MccA was, and was completely absent from the classes Saccharomycetes and Sordariomycetes. MccA was present along with urea amidolyase and urea carboxylase in three species (*Fusarium verticilloides, F. oxysporum,* and *Nectria haematococca*), and along with only urea amidolyase in three other species (*F. graminearum, M. oryzae*, and *Y. lipolytica*). A phylogenetic analysis using the biotin-carboxylation domains of Pyc, Acc1, MccA, PccA, urea amidolyase, and urea carboxylase from fungi showed that these domain sequences were highly diverged. Bootstrap analysis did not show any significantly supported clustering of urea amidolyase and urea carboxylase with any of the other four enzymes (see Figure A4.1). Urea amidolyase and urea carboxylase appear to have no clear direct origin among the other biotin-carboxylation domain containing proteins. Or such divergence may have happened such a long time ago that we can no longer identify the origin.

4.2.3 Distribution of urea amidolyase and other related proteins among eubacteria

In order to elucidate further the origin of long and short forms of urea amidolyase found in fungi: whether they share a common evolutionary origin or arose independently, we performed extensive similarity searches using these protein and domain sequences among 56 bacterial genomes. As shown in Table 4.3 (see also Table A4.6), the longer form of urea amidolyase (~1,800 aa) was found only in one bacterium, *Pantoea ananatis* (class

Gammaproteobacteria). This bacterium, which previously belonged to the genus *Erwinia* but was recently reclassified into the genus *Pantoea*, is a well-known plant pathogen with a reported case of it also being a human-pathogen [24, 25]. This bacterium and its related species are usually isolated from soil, fruits, and vegetables [24]. Urea carboxylase (~1,200 aa), the shorter form of urea amidolyase, was found in bacterial species scattered among a wide range of groups. Almost all bacteria with urea carboxylase also had amidase. These two enzymes are encoded in two different genes in bacteria, but are located next to each other in most of the bacterial genomes we examined (see Table A4.7). In two species (*Wolinella succinogenes*, class Epsilonproteobacteria; and *Gloeobacter violaceus*, phylum Cyanobacteria), the two genes were not adjacent to each other but only 943 bp and 1,701 bp apart, respectively, while in another Cyanobacteria species (*Cyanothece* sp.), the two genes were located far apart (979,743 bp). *Sorangium cellulosum* (class Deltaproteobacteria) and *Nitrosomonas europaea* (class Alphaproteobacteria) had urea carboxylase but lacked amidase. Three Gammaproteobacteria species have two urea carboxylase genes, only one of which lies next to the amidase gene. *P. ananatis*, a gammaproteobacteria, which has urea amidolyase (the long form), also has urea carboxylase (the short form). Furthermore, *P. ananatis* has no independent amidase gene. The only amidase sequence present in this bacterium is the domain of the urea amidolyase gene. It seems reasonably likely that fusion of the amidase and urea carboxylase genes occurred in *P. ananatis* to generate the long form of the urea amidolyase gene similar to those found in fungi.

The urease protein in bacteria occurs as a trimer of alpha, beta, and gamma subunits encoded by separate genes forming a gene cluster, whereas in eukaryotes a single gene encodes the urease protein, a fused protein representing the three bacterial subunits [26]. In some bacteria, beta and gamma subunits are fused and encoded by one gene (denoted with

ß/γ in Table 4.3) while in others either beta- or gamma-subunit gene was missing. As shown in Table 4.3, existence of these urease-subunit genes was scattered throughout the bacterial groups. Of 56 bacterial genomes we examined, 31 had either or both of urease and amidase/urea carboxylase (or urea amidolyase). Only seven of 31 bacterial species had all three genes. Consistent with what we observed in fungi, there appears to be a certain degree of dichotomy in possession of urease genes or amidase/urea carboxylase (or urea amidolyase) genes among bacterial genomes.

4.2.4 Phylogenetic analysis of amidase domain sequences

In order to elucidate the evolutionary origin of eukaryotic urea amidolyase proteins, we performed phylogenetic analysis among amidase, urea amidolyase, and urea carboxylase identified across kingdoms. Phylogenies were reconstructed using amidase and urea carboxylase sequences separately.

Figure 4.3 is the maximum-likelihood phylogenetic tree reconstructed from the amidase domain sequences from urea amidolyase and the amidase protein sequences from fungi, green algae, and bacteria (the minimum-evolution tree is shown in Figure A4.2). It shows that the fungal amidase domain from urea amidolyase (shown in blue and denoted by UA in Figure 4.3), and the stand-alone fungal amidase protein that exists on its own (shown in blue and denoted by A in Figure 4.3) cluster separately, implying that they have evolved independently. The amidase sequences from green algae (shown in green in Figure 4.3) cluster with the stand-alone amidase protein from fungi, however, with not very strong bootstrap support (76%).

Bacterial amidase sequences also cluster into two groups (shown in red in Figure 4.3). Amidases from four gammaproteobacteria species (*P. ananatis, Pantoea sp. At-9b,*

Pectobacterium carotovorum, and *Cellvibrio japonicus*) and one betaproteobacteria species (*Achromobacter piechaudii*) form a cluster (denoted by A1 in Figure 4.3). Notably, the amidase sequence of *P. ananatis* is part of the urea amidolyase, and the amidase genes of the other three gammaproteobacteria species lie immediately adjacent to their urea carboxylase genes (see Table A4.7). These bacterial amidases cluster with fungal amidases from urea amidolyase with a strong bootstrap support (100%). Compared to the fungal stand-alone amidases (Fungi A), the fungal amidase-domain sequences from urea amidolyase (Fungi UA) are clearly more closely related to the bacterial amidases, especially to those from *P. ananatis* and a small number of gamma- and betaproteobacteria species (Bacteria A1).

4.2.5 Phylogenetic analysis of urea carboxylase domain sequences

Figure 4.4 shows the result of maximum-likelihood phylogenetic analysis using the urea carboxylase protein and urea carboxylase domain sequences from urea amidolyase (the minimum-evolution tree is shown in Figure A4.3). The urea carboxylase sequence (~1,200 aa) is twice longer than the amidase sequence (~600 aa), which resulted in a better resolution in the reconstructed phylogeny. Bacterial urea carboxylase sequences were clearly divided into two clusters (denoted by UC1 and UC2 in Figure4. 4) where both were supported by 100% bootstrap values. The UC1 group, which consists of the five species of gamma- and betaproteobacteria (*P. ananatis, Pantoea At-9b, P. carotovorum, C. japonicus*, and *A piechaudii*), clustered closely with the fungal urea amidolyase (UA) with a high bootstrap value (97%). These five bacterial species are the same five species found in Figure 4.3 (A1) whose amidases clustered with the amidase-domain sequences of the fungal urea amidolyase. Four of these five bacterial species have a second urea carboxylase gene. Thus, the

duplication event that created these two sets of urea carboxylase genes must have happened before the divergence of the five proteobacteria. Based on the deep divergence between the paralogous groups (UC1 and UC2) and the somewhat slower evolution observed in UC1 (the urea carboxylase genes found only in five gamma/betaproteobacteria species), we speculate that the close functional association with amidase likely arose in the UC1 group to create a fused single gene, urea amidolyase, in *P. ananatis*, and thus changed the evolutionary rate and pattern in this copy of urea carboxylase.

We also see two separate and strongly supported clusters of urea carboxylase sequences in fungi. One cluster is of the urea carboxylase domain from urea amidolyase (UA, 100% bootstrap support) whereas the other cluster is of the urea carboxylase protein sequence (UC, 97% bootstrap support). It shows that the urea carboxylase sequences in the two groups have independently evolved over a long period of time. Since urea carboxylase was found in the phylum Basidiomycota (represented by *Cryptococcus neoformans* in Figure 4.4) and it clustered with other urea carboxylase proteins, the divergence between urea carboxylase and urea amidolyase in fungi must have preceded the Ascomycota-Basidiomycota divergence. As we discuss in the next section, the formation of urea amidolyase with acquisition of the amidase domain seems to have happened most likely in a bacterial lineage. Note that the urea carboxylases from green algae clustered with the fungal urea carboxylases (with 100% bootstrap support) rather than with the fungal urea amidolyases. This clustering pattern is consistent with what we observed in the amidase phylogeny (Figure 4.3) where green algal genes clustered with the stand-alone version of the fungal amidase genes rather than with the amidase-domain sequence of urea amidolyase. Although in some green algae, amidase and urea carboxylase genes are located relatively

closely (within 588 to 6,236 bp; Table A4.2), their evolution is completely independent from urea amidolyase genes found in fungi.

As mentioned before, two Hydra urea carboxylase sequences were found from the NCBI nr database search. One of them was actually found to be a sequence of a putative bacterial symbiont, *Curvibacter* (betaproteobacteria) (described in NCBI gi|260221606 entry). Phylogenetic analysis clearly showed that this sequence belongs to the bacterial urea carboxylase (UC2) group (see Figure A4.4). The other Hydra sequence clustered with urea carboxylase sequences from green algae and fungi (93% bootstrap support).

4.2.6 Bacterial origins of the fungal urea amidolyase and urea carboxylase

Our phylogenetic analysis did not support the previous hypothesis that the fungal urea amidolyase and urea carboxylase sequences are formed from fungal biotin-carboxylation domain containing proteins such as MccA or PccA. Instead, our conclusion is that the urea amidolyase and urea carboxylase genes currently found in fungi and green algae, as well as in Hydra, are the results of horizontal gene transfer events from bacteria. This is based on observations such as the abundant distribution of the shorter form of urea amidolyase, *i.e.,* urea carboxylase, as well as the single occurrence so far of urea amidolyase (the long form) in bacteria, coupled with the rarity of both forms of urea amidolyase in eukaryotes except in the fungal kingdom, in some green algae, and in Hydra.

Phylogenetic analysis of amidase and urea carboxylase sequences across kingdoms showed that the urea carboxylase domain in urea amidolyase and the urea carboxylase protein itself have undergone extensive independent evolution. Fungal urea amidolyase proteins are more closely related to one of the two groups of bacterial urea carboxylase. Furthermore, one of these bacteria (*P. ananatis*) has a unique urea amidolyase gene, a

product of amidase/urea carboxylase gene fusion. The direction of the horizontal gene transfer seems to be from a bacterial lineage to a fungal lineage, since in bacteria other than *P. ananatis,* urea carboxylase and amidase exist as two independent genes although they are located next to each other. Inspection of introns in fungal urea amidolyase genes corroborates this hypothesis further. Fungal urea amidolyases are either single or double-exon genes (see Table A4.2). All Saccharomycetes species except for *Y. lipolytica* have single-exon urea amidolyase genes. While in the three Sordariomyetes species (*M. oryzae, N. haematococca,* and *F. graminearum*) the single intron was inserted towards the end of the urea carboxylase domain, in the duplicated *Y. lipolytica* genes the single intron was inserted at the beginning of the amidase domain. These observations indicate that the introns in these fungal urea amidolyase genes must have been acquired independently during their evolution as fungal genes. Therefore, fusion of the two genes appears to have happened in the ancestral bacterial species close to *P. ananatis*, and this fused gene must have been transferred to a fungal lineage.

Since so far we found the urea amidolyase protein only in one bacterial species, it is probable that the fusion of urea carboxylase and amidase to form bacterial urea amidolyase is a recent event specific to this bacterial lineage. If this is the case, the fusion event in *P. ananatis* could be also independent from those that produced fungal urea amidolyases. However, we did not find any unfused fungal urea carboxylase sequences clustered with urea amidolyase in our phylogenetic analysis (Figure 4.4), nor did we find any unfused fungal amidase sequences clustered with urea amidolyase (Figure 4.3). Therefore, if the fusion happened in fungal lineage, it must have happened soon after the two bacterial genes (amidase and urea carboxylase) were acquired by an ancestral fungal species. Regardless of the timing of the fusion event, association between the amidase and urea carboxylase

sequences for the urea amidolyase function and subsequent divergence of these sequences from the other paralogous set must have started in bacterial lineage.

Compared to urea amidolyase, urea carboxylase genes in fungi have a wider range in the number of exons, 1-16 exons, implying again their independent evolution as well as a greater number of accumulated changes. Note that the single introns found in the urea carboxylase genes of *N. haematococca* and *F. oxysporum* are both at the beginning of the genes and of similar lengths (55-56bp; see Table A4.2). It indicates that the common ancestor of these species acquired a single intron in the urea carboxylase gene and it happened independently from the intron acquisition in *N. haematococca* urea amidolyase. Interestingly, the number of introns in urea carboxylase and amidase genes in green algae is much higher than the number of introns in the fungal orthologues. This is in agreement with the observation that the *Chlamydomonas reinhardtii* genome has much higher percentage of genes with introns and a much greater number of exons per gene (88% and 7.4) as compared to *S. cerevisiae* (5% and 1) and *S. pombe* (43% and 2) [27].

There have been studies presenting cases of bacteria-to-fungi horizontal gene transfers. For example, Hall *et al.* [11] found ten potential such cases in *S. cerevisiae* and one in *Ashbya gossypii*. Fitzpatrick *et al.* [12] reported two *Candida parapsilosis* genes as bacterial origin. Garcia-Vallvé *et al.* [13] showed that many glycosyl hydrolase genes in the rumen fungus *Orpinomyces joyonii* were acquired from bacteria. Schmitt and Lumbsch [14] showed that the polyketide synthase in lichen-forming fungi were results of ancient horizontal gene transfer from Actinobacteria. A recent study, the largest of its kind, by Marcet-Houben and Gabaldón [15] detected 713 transferred genes in 60 fungal genomes. Therefore, horizontal gene transfers from bacteria to fungi do not appear to be rare events. We identified yet another such example.

4.2.7 Proposed model for the urea carboxylase and urea amidolyase evolution

Figure 4.5 illustrates our proposed model for the evolution of urea carboxylase and urea amidolyase genes in fungi. As presented in Figure 4.5A, an ancestral urea carboxylase sequence in bacteria duplicated in the beta/gammaproteobacteria lineage and evolved into two genes (UC1 and UC2). Since in many bacterial genomes, urea carboxylase and amidase genes are located adjacent to each other (see Table A4.7), it is plausible that before the duplication, the ancestral urea carboxylase gene already had an associated function with the amidase gene. However, the creation of duplicated redundant copies of the urea carboxylase gene in beta/gammaproteobacteria species appears to have reinforced the association between the two genes and changed their evolutionary pattern and rate in these bacteria. This amidase-associated copy of bacterial urea carboxylase gene (UC1) was subsequently fused with the amidase gene to form a single urea amidolyase gene. The fused gene was later transferred to an ancestral ascomycete lineage before the divergence of the Pezizomycotina and Saccharomycotina. Alternatively, the gene fusion could have happened in an ancestral fungal species soon after the region containing amidase and urea carboxylase genes was transferred from bacteria.

The other bacterial urea carboxylase gene (UC2) may have also been acquired by fungi, green algae, as well as Hydra. Since our phylogenetic analysis did not show independent origins for these urea carboxylase genes, the acquisition of this enzyme into fungi, green algae, and Hydra must have happened around the time of divergence among these groups of organisms. It may have been by a single event, likely before the divergence of these organisms. Then we cannot eliminate the possibility that what we observed in the

urea carboxylase genes is the result of simple vertical evolution from bacteria to eukaryotes. Either way, however, many eukaryotes including the entire metazoa and land plants must have lost these genes. As we mentioned before (and shown also in Figure 4.5B), even within fungi, the urea carboxylase gene is not retained in many species. Considering that either scenario requires such a high number of loss events, there would be other possible scenarios. One group of organisms (either green algae, Hydra, or fungi) may have acquired a urea carboxylase gene from bacteria first. Later this gene may have been transferred to other organisms. Although this scenario requires fewer loss events, the main question is how such horizontal gene transfers can happen between green algae, Hydra, and fungi, or among any of their ancestral organisms.

In fungi, the introduction of the urea carboxylase gene happened earlier than that of the urea amidolyase gene as shown in Figure 4.5B. The urea carboxylase gene (red circle) was acquired in fungi before the divergence of the phyla Ascomycota and Basidiomycota. The acquisition could have been after the divergence of the phylum Zygomycota or alternatively the gene was lost from the Zygomycota lineage. Some Basidiomycota species subsequently lost the gene (the lost events are indicated with grey symbols in Figure 4.5B). In the phylum Ascomycota, this gene was again lost in the subphyla Taphrinomycotina (it includes *S. pombe*) and Saccharomycotina. Further losses of this gene happened in some species of the subphylum Pezizomycotina. The urea carboxylase gene appears to become easily dispensable in many species, which may be related to the genomic and metabolic environment of the organisms. The same seems to be the case with MccA and PccA. The introduction of the urea amidolyase gene (black square) in fungi took place before the divergence of the subphyla Pezizomycotina and Saccharomycotina but probably after the divergence of the subphylum Taphrinomycotina (at least after the phylum Ascomycota

diverged from the ancestral lineage). Within the subphylum Pezizomycotina, the urea amidolyase gene was lost in many groups but retained in almost all species in the class Sordariomycetes (absent in the order Sordariales species). The urea amidolyase gene was retained in all Saccharomycotina species, and even recently duplicated in *Y. lipolytica*.

4.3 Conclusions

We have presented a possible scenario of horizontal gene transfer of the urea amidolyase and urea carboxylase genes from bacteria to fungi. Plastic and opportunistic genome evolution in bacteria and fungi and their evolutionary interplay must have allowed the Saccharomycetes fungi to abandon the use of nickel-containing urease. It contributed to optimizing these organisms toward Ni^{2+} (and Co^{2+})-independent cellular metabolisms. Further detailed studies of a wider range of gene families would reveal the importance of acquisition of bacterial genes in fungal evolution.

4.4 Methods

4.4.1 Similarity searches

Similarity searches for protein sequences were performed using blastp (version 2.2.17 [28]). For urea amidolyase search, the *S. cerevisiae* sequence (P32528) was used as a query. Search was performed using both the full sequence as well as only the amidase domain of this sequence. To search for urea carboxylase sequences, *A. nidulans* sequence (P38095) was used as a query. To search for other urea carboxylase domain containing proteins, the *S. cerevisiae* Acc1 (Q00955) and Pyc (P11154), *A. nidulans* MccA (Q6T5L7), and *Aspergillus*

related *Neosartorya fischeri* PccA (A1DF70) were used as query sequences. The urease sequence from *A. fumigatus* (Q6A3P9) was used as a query sequence to search for urease.

We performed these searches against 56 bacterial genomes, 64 fungal genomes, and 10 non-fungal eukaryotic genomes (including 4 green algae, 2 land plants, 1 amoebozoa, and 3 animals). The species names, taxonomical groups, and the sources of the sequences are listed in Tables A4.1, A4.4, A4.5 and A4.6. The species were chosen such that all major bacterial, fungal, and other eukaryotic groups were represented from a tree of life [*e.g.*, 29]. For fungi, preliminary search for urea amidolyase, urea carboxylase, and amidase was done in 64 genomes and further analysis was done using 27 selected fungal genomes (noted with * in Table A4.3). The non-redundant (nr) database at National Center for Biotechnology Information (NCBI) was also searched for urea amidolyase, urea carboxylase, and urease protein sequences using blastp.

All protein sequences were highly conserved, and similar sequences were clearly identifiable in the results obtained by blastp similarity search. The E-value threshold for each protein hit was as follows: 1×10^{-49} for amidase, 0 for urea amidolyase and urea carboxylase, 1×10^{-12} for urease, 1×10^{-111} for MccA, 1×10^{-115} for PccA, and 0 for Pyc and Acc1. The default parameters were used with blastp program (version 2.2.17), which include BLOSUM62 scoring matrix, low-complexity filtering, gap-open and gap-extend penalties of 11 and 1, respectively. In order to obtain the E-values comparable among different genome sizes, the "effective length of database" was set to 500,000,000 (using -z option). This also makes the E-values obtained from each genome search equivalent to those obtained against NCBI nr database.

4.4.2 Multiple alignment and phylogenetic analysis

Multiple alignments of protein sequences were generated using MAFFT (version 6.240 [30]) with default parameters (FFT-NS-2, a progressive FFT alignment with two tree-building cycles). The maximum-likelihood phylogeny [31] was reconstructed as implemented in raxmlHPC-MPI (version 7.0.4 [32]) using the following options: '-m PROTMIXWAG' to use WAG amino-acid substitution model [33] with a fixed number approximation followed by a refined gamma-model of rate heterogeneity, '-f a' for a rapid bootstrap analysis, '-x 1234' to set the random seed, and '-# 1000' for 1000 pseudoreplicates for bootstrap analysis. To gather the bootstrap values, the 'consense' program of the Phylip package (v. 3.68 [34]) was used. The minimum-evolution phylogeny [35] was reconstructed as implemented in MEGA4 [36] using the JTT amino-acid substitution model [37] with 1000 pseudoreplicates for bootstrap analysis.

4.5 References

1. Navarathna DH, Harris SD, Roberts DD, Nickerson KW: **Evolutionary aspects of urea utilization by fungi**. *FEMS Yeast Res* 2010, **10**(2):209-213.
2. Ghosh S, Navarathna DH, Roberts DD, Cooper JT, Atkin AL, Petro TM, Nickerson KW: **Arginine-induced germ tube formation in Candida albicans is essential for escape from murine macrophage line RAW 264.7**. *Infect Immun* 2009, **77**(4):1596-1605.
3. Roon RJ, Levenberg B: **Urea amidolyase. I. Properties of the enzyme from Candida utilis**. *J Biol Chem* 1972, **247**(13):4107-4113.
4. Leftley JW, Syrett PJ: **Urease and ATP:urea amidolyase activity in unicellular algae**. *J Gen Microbio* 1973, **77**.
5. Al-Houty FAA, Syrett PJ: **The occurrence of urease/urea amidolyase and glycollate oxidase/dehydrogenase in Klebsormidium spp. and members of the Ulotrichales**. *European J of Phycology* 1984, **19**:1-10.
6. Roon RJ, Hampshire J, Levenberg B: **Urea amidolyase. The involvement of biotin in urea cleavage**. *J Biol Chem* 1972, **247**(23):7539-7545.
7. Cooper TG, Lam C, Turoscy V: **Structural analysis of the dur loci in S. cerevisiae: two domains of a single multifunctional gene**. *Genetics* 1980, **94**(3):555-580.
8. Hodson RC, Williams SK, 2nd, Davidson WR, Jr.: **Metabolic control of urea catabolism in Chlamydomonas reinhardi and Chlorella pyrenoidosa**. *J Bacteriol* 1975, **121**(3):1022-1035.
9. Kanamori T, Kanou N, Atomi H, Imanaka T: **Enzymatic characterization of a prokaryotic urea carboxylase**. *J Bacteriol* 2004, **186**(9):2532-2539.

10. Jitrapakdee S, Wallace JC: **The biotin enzyme family: conserved structural motifs and domain rearrangements**. *Curr Protein Pept Sci* 2003, **4**(3):217-229.

11. Hall C, Brachat S, Dietrich FS: **Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae**. *Eukaryot Cell* 2005, **4**(6):1102-1115.

12. Fitzpatrick DA, Logue ME, Butler G: **Evidence of recent interkingdom horizontal gene transfer between bacteria and Candida parapsilosis**. *BMC Evol Biol* 2008, **8**:181.

13. Garcia-Vallve S, Romeu A, Palau J: **Horizontal gene transfer of glycosyl hydrolases of the rumen fungi**. *Mol Biol Evol* 2000, **17**(3):352-361.

14. Schmitt I, Lumbsch HT: **Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi**. *PLoS One* 2009, **4**(2):e4437.

15. Marcet-Houben M, Gabaldon T: **Acquisition of prokaryotic genes by fungal genomes**. *Trends Genet* 2010, **26**(1):5-8.

16. Marcet-Houben M, Gabaldon T: **The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome**. *PLoS One* 2009, **4**(2):e4357.

17. Wang H, Xu Z, Gao L, Hao B: **A fungal phylogeny based on 82 complete genomes using the composition vector method**. *BMC Evol Biol* 2009, **9**:195.

18. Oren L, Ezrati S, Cohen D, Sharon A: **Early events in the Fusarium verticillioides-maize interaction characterized by using a green fluorescent protein-expressing transgenic isolate**. *Appl Environ Microbiol* 2003, **69**(3):1695-1701.

19. Enya J, Togawa M, Takeuchi T, Yoshida S, Tsushima S, Arie T, Sakai T: **Biological and phylogenetic characterization of Fusarium oxysporum complex, which causes yellows on Brassica spp., and proposal of F. oxysporum f. sp. rapae, a**

novel forma specialis pathogenic on B. rapa in Japan. *Phytopathology* 2008, **98**(4):475-483.

20. Gunawardena U, Rodriguez M, Straney D, Romeo JT, VanEtten HD, Hawes MC: **Tissue-specific localization of pea root infection by Nectria haematococca. Mechanisms and consequences**. *Plant Physiol* 2005, **137**(4):1363-1374.

21. Odds FC: **Candida and candidiasis**. Bailliere: Tindall; 1988.

22. DiCarlo FJ, Schultz AS, Kent AM: **The mechanism of allantoin catabolism by yeast**. *Arch Biochem Biophys* 1953, **44**:468-474.

23. Zhang Y, Rodionov DA, Gelfand MS, Gladyshev VN: **Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization**. *BMC Genomics* 2009, **10**:78.

24. De Baere T, Verhelst R, Labit C, Verschraegen G, Wauters G, Claeys G, Vaneechoutte M: **Bacteremic infection with Pantoea ananatis**. *J Clin Microbiol* 2004, **42**(9):4393-4395.

25. De Maayer P, Chan WY, Venter SN, Toth IK, Birch PR, Joubert F, Coutinho TA: **Genome sequence of Pantoea ananatis LMG20103, the causative agent of Eucalyptus blight and dieback**. *J Bacteriol* 2010, **192**(11):2936-2937.

26. Carter EL, Flugga N, Boer JL, Mulrooney SB, Hausinger RP: **Interplay of metal ions and urease**. *Metallomics* 2009, **1**(3):207-221.

27. Labadorf A, Link A, Rogers MF, Thomas J, Reddy AS, Ben-Hur A: **Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii**. *BMC Genomics* 2010, **11**:114.

28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

29. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283-1287.

30. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30**(14):3059-3066.

31. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach**. *J Mol Evol* 1981, **17**(6):368-376.

32. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.

33. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18**(5):691-699.

34. Felsenstein J: **Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5**:164-166.

35. Rzhetsky A, Nei M: **Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference**. *J Mol Evol* 1992, **35**(4):367-375.

36. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

37. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

38. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.

39. Lucking R, Huhndorf S, Pfister DH, Plata ER, Lumbsch HT: **Fungi evolved right on track**. *Mycologia* 2009, **101**(6):810-822.

Table 4.1. Distribution of urea amidolyase and related proteins in eukaryotic species other than fungi.[a]

		Enzymes[b]			
Kingdom	**Species**	**UA**	**UC**	**A[c]**	**Urease**
Plantae (green algae)					
	Chlamydomonas reinhardtii	-	1	1[+]	-
	Volvox carteri	-	1	1[+]	-
	Chlorella sp. NC64A	-	1	1[+]	-
	Coccomyxa sp. C-169	-	1	1	-
Plantae (land plants)					
	Arabidopsis thaliana	-	-	-	1
	Oryza sativa	-	-	-	(1)[d]
Amoebozoa					
	Dictyostelium discoideum	-	-	-	-
Animalia					
	Nematostella vectensis	-	-	-	1
	Drosophila melanogaster	-	-	-	-
	Homo sapiens	-	-	-	-

[a]See Table A4.1 for the sequence sources.
[b]See Figure 4.1 for the enzyme name abbreviations. The number of sequences found from each genome is shown. '-' indicates that no similar sequence was found.
[c]The amidase gene located close to the urea carboxylase gene (less than 6,250 bp) is indicated with [+]. See Table A4.2 for the distance between the genes.
[d]Blastp similarity search against the downloaded rice genome showed no sequence similar to urease. However, similarity search against NCBI nr database showed urease from *Oryza sativa*.

Table 4.2. Distribution of urea amidolyase and related proteins in fungal species.[a]

Taxonomical group[b]	Species	Enzymes[c]					
		UA	UC	A[d]	Urease	MccA	PccA
[Zygomycota]	*Rhizopus oryzae*	-	-	-	1	1	1
[Basidiomycota]	*Ustilago maydis*	-	-	-	1	1	-
	Cryptococcus neoformans	-	1	-	1	1	-
	Coprinus cinereus	-	-	-	1	1	1
[Ascomycota/Taphrinomycotina]							
Schizosaccharomycetes	*Schizosaccharomyces pombe*	-	-	-	1	-	-
[Ascomycota/Pezizomycotina]							
Eurotiomycetes	*Coccidioides immitis*	-	-	-	1	1	1
	Aspergillus nidulans	-	1	-	1	1	1
	Aspergillus fumigatus	-	1	-	1	1	1
	Aspergillus terreus	-	1	-	1	1	1
	Aspergillus oryzae	-	-	-	1	1	-
Dothideomycetes	*Mycosphaerella graminicola*	-	-	1	1	1	1
	Stagonospora nodorum	-	1	1	1	1	1
	Cochliobolus heterostrophus	-	1	1	1	1	1
Leotiomycetes	*Botritis cinerea*	-	-	-	1	1	1
Sordariomycetes	*Neurospora crassa*	-	-	-	1	1	-
	Magnaporthe oryzae	1	-	(1)	1	1	-
	Nectria haematococca	1	1	(1)	1	1	-
	Fusarium graminearum	1	-	(1)	2	1	-
	Fusarium oxysporum	1	1	(1)	2	1	-
	Fusarium verticillioides	1	1	(1)	1	1	-
[Ascomycota/Saccharomycotina]							
Saccharomycetes	*Yarrowia lipolytica*	2	-	(2)	-	1	-
	Candida albicans	1	-	(1)	-	-	-
	Candida lusitaniae	1	-	(1)	-	-	-
	Debaryomyces hansenii	1	-	(1)	-	-	-
	Ashbya gossypii	1	-	(1)	-	-	-
	Candida glabrata	1	-	(1)	-	-	-
	Saccharomyces cerevisiae	1	-	(1)	-	-	-

[a]See Tables A4.4 and A4.5 for the sequence sources.
[b]The phylum/subphylum (in square brackets) and class are given.
[c]See Figure 4.1 for the enzyme name abbreviations. The number of sequences found from each genome is shown. '-' indicates that no similar sequence was found.
[d]The amidase sequences that are a part of the urea amidolyase sequences are shown in parentheses.

Table 4.3. Distribution of urea amidolyase and related proteins in eubacterial species.[a]

		Enzymes[b]			
Phylum or Class	**Species**	**UA**	**UC**	**A[c]**	**Urease[d]**
Alphaproteobacteria	*Caulobacter crescentus NA1000*	-	1	1*	-
	Asticcacaulis excentricus CB 48	-	1	1*	-
	Sinorhizobium medicae WSM419	-	-	-	α,β,γ
Betaproteobacteria	*Achromobacter piechaudii ATCC 43553*	-	1	1*	-
	Bordetella pertussis Tohama I	-	-	-	α,β,γ
	Nitrosomonas europaea ATCC 19718	-	1	-	-
	Neisseria meningitidis FAM18	-	-	-	-
	Burkholderia sp. CCGE1001	-	1	1*	α,β,γ
Gammaproteobacteria	*Escherichia coli O111:H- str. 11128*	-	-	-	α,β,γ
	Yersinia pestis Angola	-	-	-	α,β
	Haemophilus influenzae 86-028NP	-	-	-	α,β,γ
	Pantoea ananatis LMG 20103	1	1	(1)	-
	Pantoea sp. At-9b	-	2	1*	-
	Shewanella oneidensis MR-1	-	-	-	-
	Pseudomonas aeruginosa LESB58	-	-	-	α,β,γ
	Coxiella burnetii Dugway 5J108-111	-	-	-	-
	Pectobacterium carotovorum subsp. carotovorum PC1	-	2	1*	-
	Xanthomonas campestris pv. campestris str. B100	-	-	-	-
	Cellvibrio japonicus Ueda107	-	2	1*	-
	Teredinibacter turnerae T7901	-	1	1*	α,β,γ
	Marinomonas sp. MED121	-	1	1*	α,γ
	Klebsiella pneumoniae 342	-	1	1*	α,β,γ
	Pseudomonas fluorescens SBW25	-	-	-	α,β,γ
Deltaproteobacteria	*Geobacter sp. M21*	-	-	-	-
	Sorangium cellulosum 'So ce 56'	-	1	-	α,β/γ
Epsilonproteobacteria	*Helicobacter pylori B38*	-	-	-	α,β/γ
	Wolinella succinogenes DSM 1740	-	1	1[+]	-
Acidobacteria	*Acidobacterium capsulatum ATCC 51196*	-	-	-	-
	Solibacter usitatus Ellin6076	-	1	1*	-
Cyanobacteria	*Synechococcus sp. PCC 7002*	-	-	-	α,β,γ
	Gloeobacter violaceus PCC 7421	-	1	1[+]	-
	Cyanothece sp. PCC 7425	-	1	1	α,β,γ
Deinococcus-Thermus	*Thermus thermophilus HB8*	-	-	-	-
	Deinococcus deserti VCD115	-	-	-	-
Chloroflexi	*Dehalococcoides ethenogenes 195*	-	-	-	-
Aquificae	*Aquifex aeolicus VF5*	-	-	-	-
Thermotogae	*Thermotoga maritima MSB8*	-	-	-	-
Fusobacteria	*Fusobacterium nucleatum subsp. Nucleatum ATCC 25586*	-	-	-	-
Verrucomicrobia	*Verrucomicrobium spinosum DSM 4136*	-	1	1*	α,β,γ
Chlamydiae	*Chlamydophila pneumoniae CWL029*	-	-	-	-
	Chlamydia trachomatis B/TZ1A828/OT	-	-	-	-
Bacterioidetes	*Porphyromonas gingivalis W83*	-	-	-	-
Chlorobi	*Chlorobium limicola DSM 245*	-	-	-	-
Fibrobacteres	*Fibrobacter succinogenes subsp. succinogenes*	-	-	-	-

	S85				
Actinobacteria	*Mycobacterium tuberculosis F11*	-	-	-	α,β,γ
	Corynebacterium aurimucosum ATCC 700975	-	-	-	-
	Streptomyces avermitilis MA-4680	-	1	1*	α,β,γ; α,β/γ
	Bifidobacterium longum subsp. infantis ATCC 15697	-	-	-	α,β/γ
Spirochaetes	*Borrelia burgdorferi ZS7*	-	-	-	-
	Treponema denticola ATCC 35405	-	-	-	-
Planctomycetes	*Rhodopirellula baltica SH 1*	-	-	-	-
Firmicutes	*Clostridium botulinum A2 str. Kyoto*	-	-	-	-
	Mycoplasma hyopneumoniae 7448	-	-	-	-
	Streptococcus pneumoniae 70585	-	-	-	-
	Bacillus anthracis str. CDC 684	-	-	-	-
	Roseburia intestinalis L1-82	-	1	1*	-

[a]See Table A4.6 for the sequence sources.
[b]See Figure 4.1 for the enzyme name abbreviations. The number of sequences found from each genome is shown. '-' indicates that no similar sequence was found.
[c]The amidase gene located next to (within 200 bp) the urea carboxylase gene is indicated with *. The amidase gene located close to (within 6,500 bp) but not next to the urea carboxylase gene is indicated with $^{+}$. See Table A4.7 for the distance between the genes. The amidase sequences that are a part of the urea amidolyase sequences are shown in parentheses.
[d]For urease, the search results for three subunits (α, β, or γ) are shown. β/γ indicates that the β and γ subunits are fused into one gene.

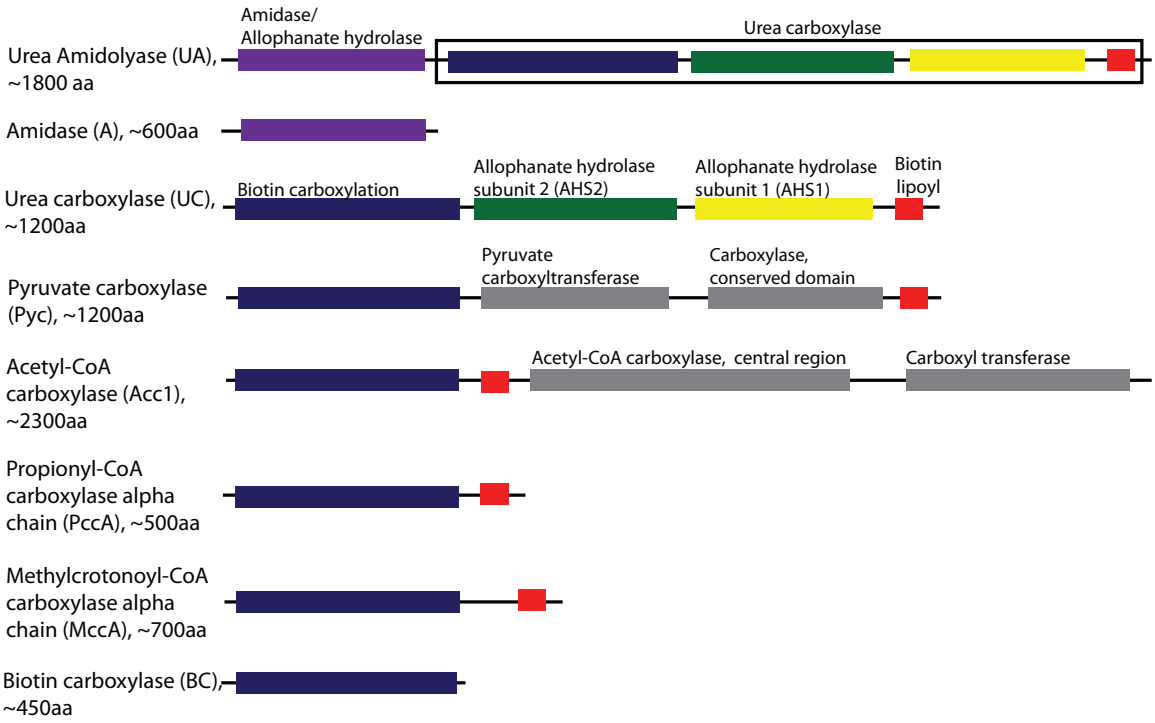

Figure 4.1. Domain structures of urea amidolyase and related proteins. Proteins that share the amidase (allophanate hydrolase) or the biotin-carboxylation domain are listed. The domains colored in grey are those that are not shared among these proteins. The domain structures are based on the InterPro protein domain database [38]. The abbreviations and approximate amino-acid lengths are given with the protein names. Amidase and urea carboxylase sequences exist as domains within the urea amidolyase protein or as single proteins by themselves. Similarly, the biotin-carboxylation sequence exists as a domain in various proteins as well as by itself as in the biotin-carboxylase protein.

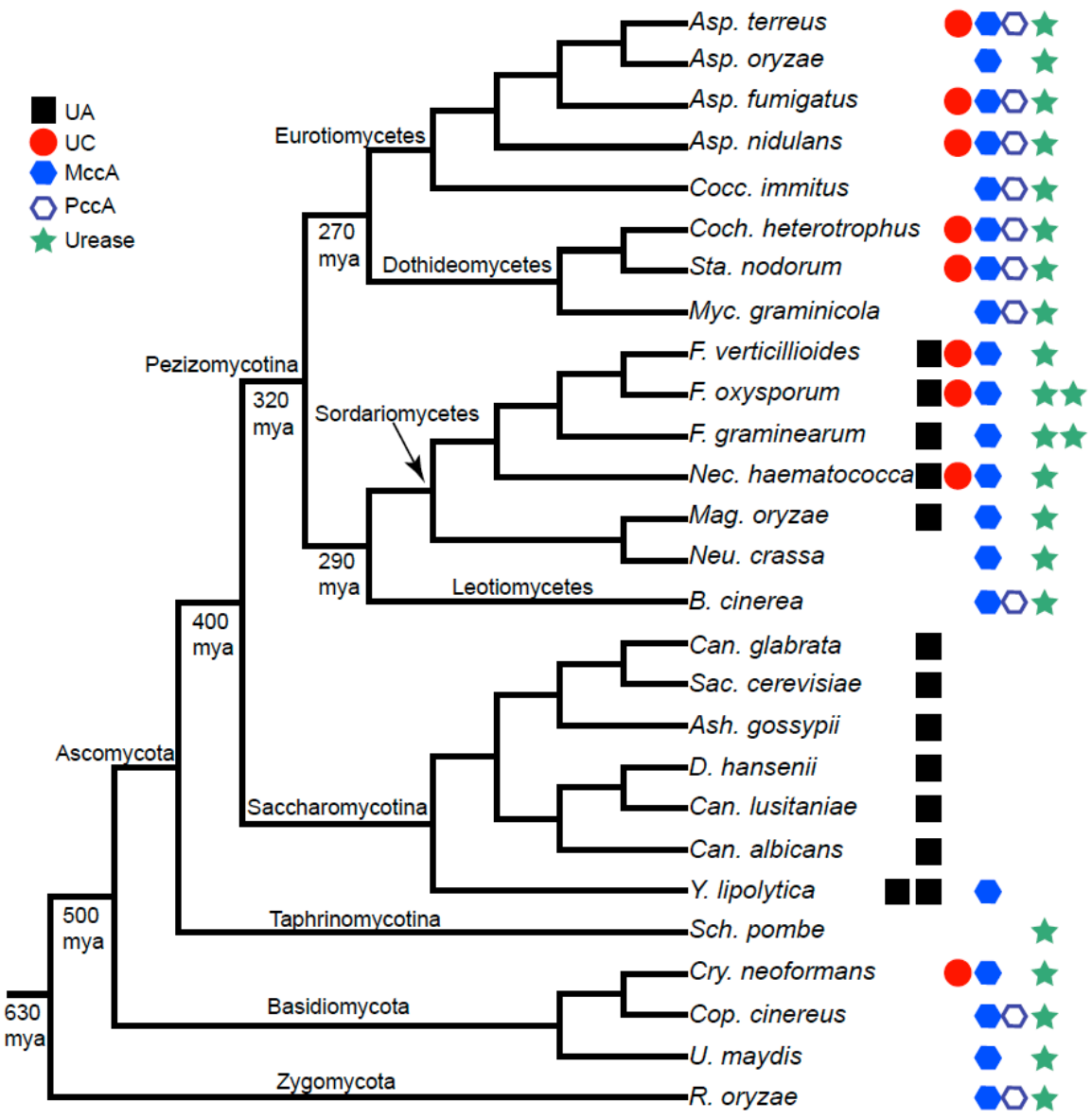

Figure 4.2. Distribution of urea amidolyase and related proteins in fungi. Existence of urea amidolyase and four other proteins are mapped along the current consensus of the fungal phylogeny (summarized from [16, 17]). The estimated divergence times (million years ago or mya) are taken from [39]. Refer to Figure 4.1 for protein name abbreviations. See Tables 4.2 and A4.3 for the complete search results.

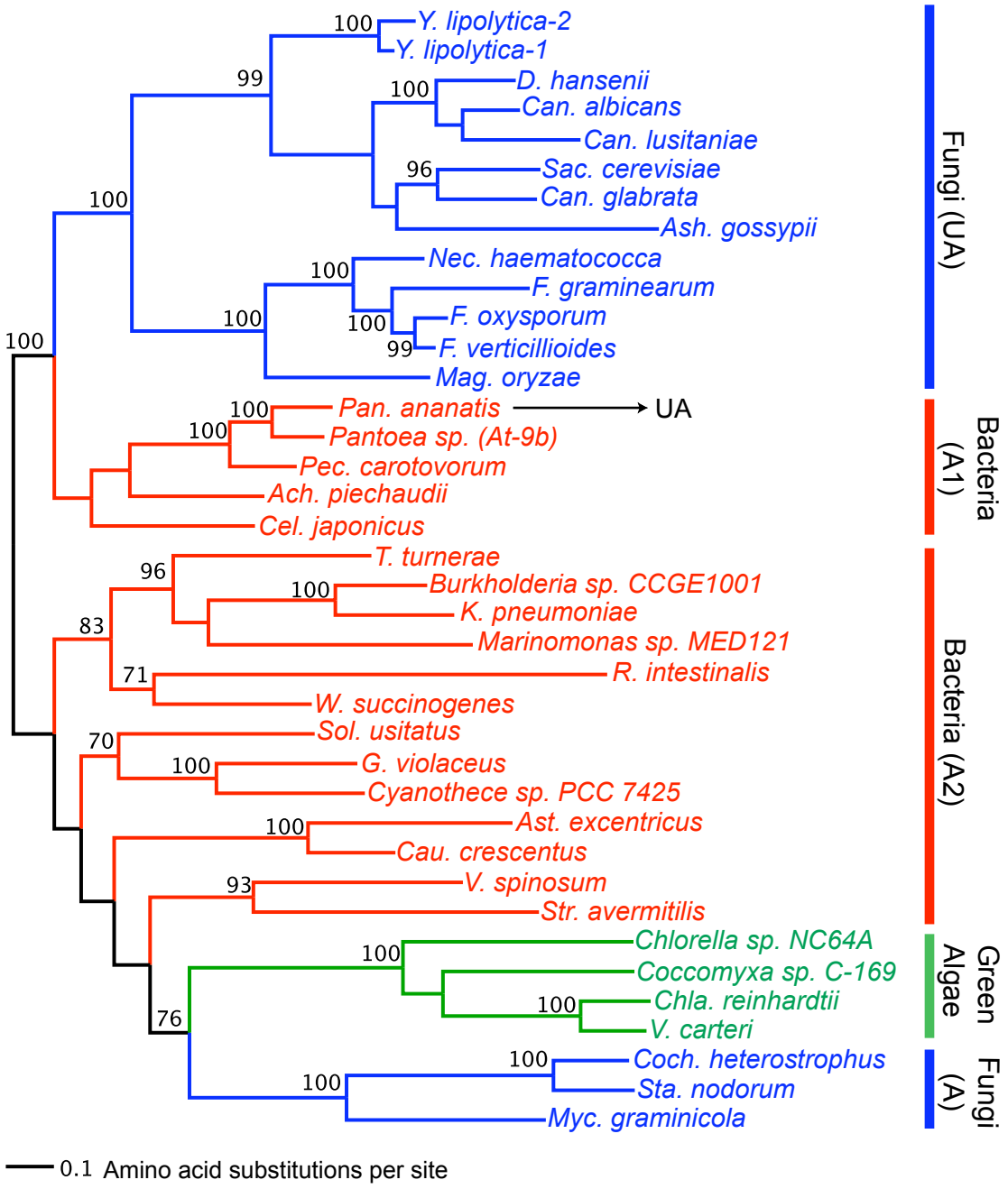

Figure 4.3. Maximum-likelihood phylogeny of amidase protein sequences. The maximum-likelihood phylogeny was reconstructed using the protein sequences from the amidase domains of the urea amidolyase proteins and the amidase proteins. The numbers above or below the internal branches show bootstrap values (%). Only bootstrap values equal to or higher than 70% are shown. Branches are colored as follows: blue for fungi, green for green algae, and red for bacteria. The bacterial urea carboxylase forms two separate groups denoted by A1 and A2. See Tables A4.1, A4.4, and A4.6 for the sequence sources.

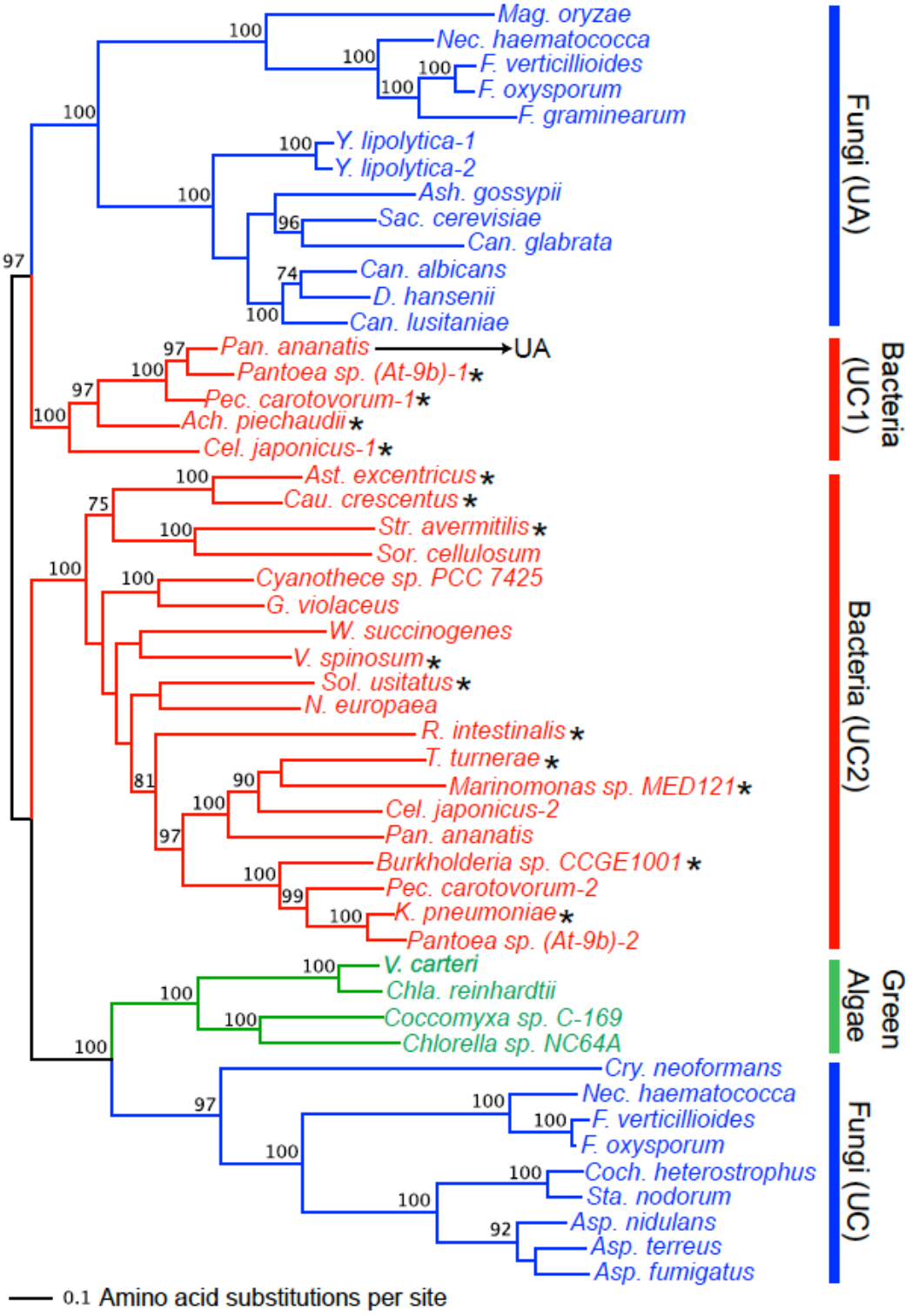
Mag. oryzae
Nec. haematococca
F. verticillioides
F. oxysporum
F. graminearum
Y. lipolytica-1
Y. lipolytica-2
Ash. gossypii
Sac. cerevisiae
Can. glabrata
Can. albicans
D. hansenii
Can. lusitaniae
Pan. ananatis
UA
Pantoea sp. (At-9b)-1*
Pec. carotovorum-1*
Ach. piechaudii*
Cel. japonicus-1*
Ast. excentricus*
Cau. crescentus*
Str. avermitilis*
Sor. cellulosum
Cyanothece sp. PCC 7425
G. violaceus
W. succinogenes
V. spinosum*
Sol. usitatus*
N. europaea
R. intestinalis*
T. turnerae*
Marinomonas sp. MED121*
Cel. japonicus-2
Pan. ananatis
Burkholderia sp. CCGE1001*
Pec. carotovorum-2
K. pneumoniae*
Pantoea sp. (At-9b)-2
V. carteri
Chla. reinhardtii
Coccomyxa sp. C-169
Chlorella sp. NC64A
Cry. neoformans
Nec. haematococca
F. verticillioides
F. oxysporum
Coch. heterostrophus
Sta. nodorum
Asp. nidulans
Asp. terreus
Asp. fumigatus
Fungi (UA)
Bacteria (UC1)
Bacteria (UC2)
Green Algae
Fungi (UC)
0.1 Amino acid substitutions per site

Figure 4.4. Maximum-likelihood phylogeny of urea carboxylase protein sequences. The maximum-likelihood phylogeny was reconstructed using the protein sequences from the urea carboxylase domains of the urea amidolyase proteins and the urea carboxylase proteins. The numbers above or below the internal branches show bootstrap values (%). Only bootstrap values equal to or higher than 70% are shown. Branches are colored as follows: blue for fungi, green for green algae, and red for bacteria. The bacterial urea carboxylase forms two separate groups denoted by UC1 and UC2. The asterisks beside the bacterial names indicate that their urea carboxylase genes are adjacent to the amidase genes in their genomes. See Table A4.7 for the distance between these genes. See Tables A4.1, A4.4, and A4.6 for the sequence sources.

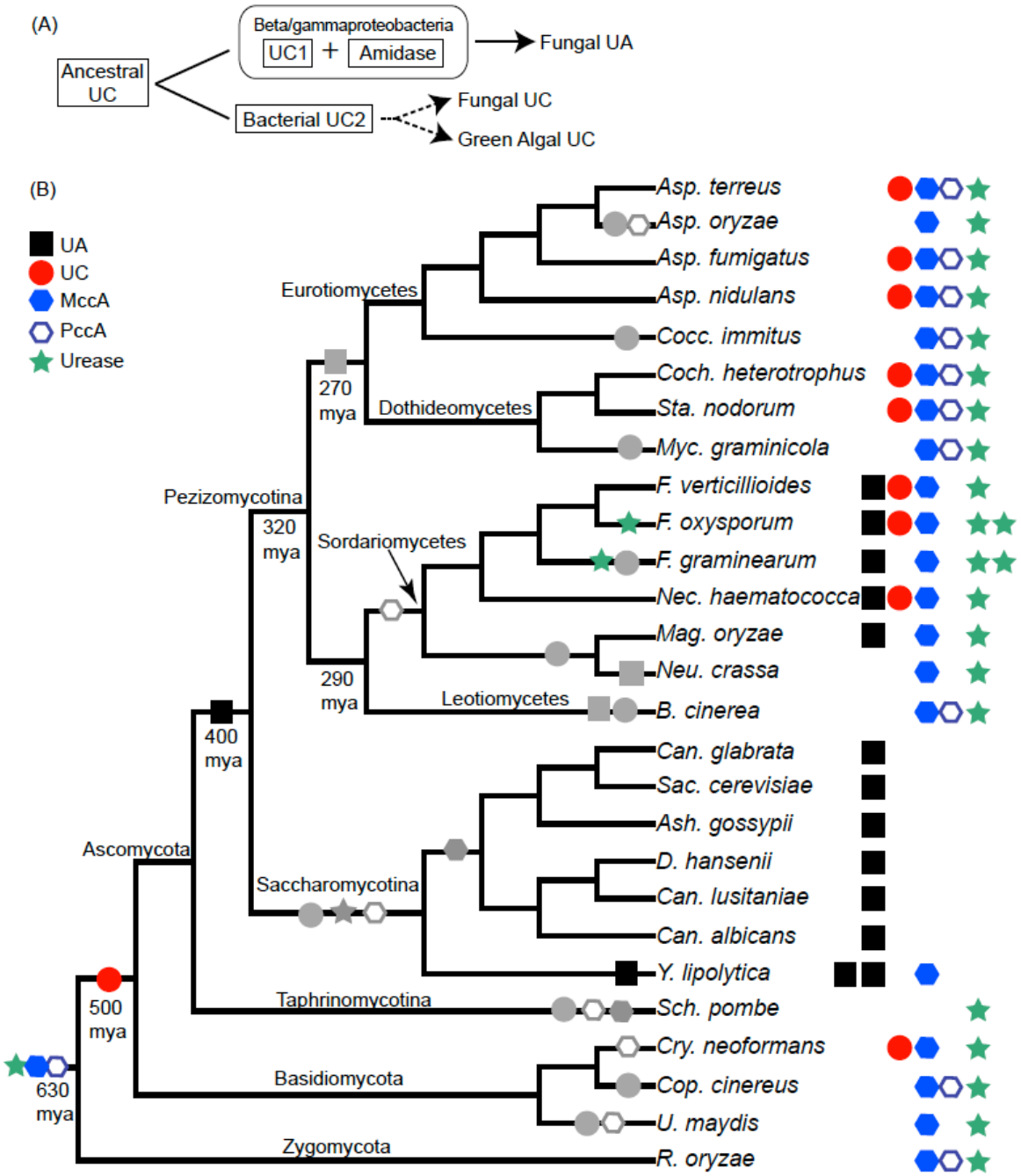

Figure 4.5. Evolutionary model of urea carboxylase and urea amidolyase in fungi. (A) The evolution of the two types of bacterial urea carboxylases, UC1 and UC2, and the subsequent transfer of those genes to fungi and green algae. The arrows represent possible horizontal gene-transfer events. Dashed arrows indicate that both horizontal transfer and vertical transmission are possible. (B) Acquisition and loss events of the urea amidolyase and related proteins inferred within fungal evolution. The fungal consensus phylogeny and the presence/absence table for five proteins are the same as Figure 4.2. Within the tree, the colored symbols indicate gene-acquisition events while the grey symbols indicate the deletion of that gene.

Additional files

Table A4.1. Sequence sources for the non-fungal eukaryotic sequences used in this study.

Kingdom		Enzymes[b]			
Species [genome ver. or Acc #]	**Source[a]**	**UA**	**UC**	**A**	**Urease**
Plantae (green algae)					
Chlamydomonas reinhardtii [v3.1]	JGI	-	133000	196482	-
Volvox carteri f. nagariensis [v1.0]	JGI	-	98356	98357	-
Chlorella sp. NC64A [v1.0]	JGI	-	133810	57824	-
Coccomyxa sp. C-169 [v2.0]	JGI	-	19857	30676	-
Plantae (land plants)					
Arabidopsis thaliana [NC_003070, NC_003071, NC_003074, NC_003075, NC_003076]	NCBI	-	-	-	15220459
Oryza sativa v6.1	Rice Genome	-	-	-	-
Amoebozoa					
Dictyostelium discoideum	DictyBase	-	-	-	-
Animalia					
Nematostella vectensis [v1.0]	JGI	-	-	-	98292
Drosophila melanogaster [rel 5.12]	Flybase	-	-	-	-
Homo sapiens	UniProtKB	-	-	-	-

[a]JGI: Joint Genome Institute (http://www.jgi.doe.gov), Rice Genome: Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/), NCBI: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), DictyBase: *Dictyostelium discoideum* database (http://dictybase.org/), Flybase: A Database of *Drosophila* Genes & Genomes (http://flybase.org/), and UniProtKB: The UniProt Knowledgebase (http://www.uniprot.org/).

[b]See Figure 4.1 for the enzyme name abbreviations. The IDs of sequences found from each genome are shown. '-' indicates that no similar sequence was found.

Table A4.2. Number of exons in urea amidolyase and related genes and their distance in eukaryotic genomes.

Kingdom / Species	No. of exons			No. of bp between UC and A[b]
	UA[a]	UC[a]	A[a]	
Plantae (green algae)				
Chlamydomonas reinhardtii v3.1		23	9	1,567 (0)
Volvox carteri f. nagariensis v1.0		25	6	588 (0)
Chlorella sp. NC64A v1.0		32	14	6,236 (3)
Coccomyxa sp. C-169 v2.0		25	9	Scaffolds 15 and 20
Fungi				
Cryptococcus neoformans H99		16		
Aspergillus nidulans FGSC A4		8		
Aspergillus fumigatus Af293		8		
Aspergillus terreus NIH2624		7		
Mycosphaerella graminicola v2.0			1	
Stagonospora nodorum SN15		3	2	Supercontigs 3 and 12
Cochliobolus heterostrophus C5		4	1	Scaffolds 1 and 8
Magnaporthe oryzae ATCC 64411	2[c]			
Nectria haematococca v2.0	2[c]	2[e]		
Fusarium graminearum PH-1 (NRRL 31084)	2[c]			
Fusarium oxysporumi 4286	1	2[e]		
Fusarium verticillioides 7600	1	1		
Yarrowia lipolytica CLIB122	2[d], 2[d]			
Candida albicans SC5314	1			
Candida lusitaniae ATCC 42720	1			
Debaryomyces hansenii CBS767	1			
Ashbya gossypii ATCC 10895	1			
Candida glabrata CBS138	1			
Saccharomyces cerevisiae S288C	1			

[a]See Figure 4.1 for the enzyme name abbreviations.
[b]The number of genes present between UC and A genes are given in parentheses. When exact distance between the two genes is not known, the locations of the two genes are given.
[c]In these genes, the intron is located towards the end of the urea-carboxylase domain.
[d]In these genes, the intron is located at the beginning of the amidase domain.
[e]In these genes, the intron is located at the beginning of the urea carboxylase sequence. The lengths of the intron and second exon in *N. haematococca* (56 bp and 3,502 bp) is similar to those in *F. oxysporum* (55 bp and 3,499 bp).

Table A4.3. Distribution of urea amidolyase, urea carboxylase, and amidase proteins in 64 fungal species.[a]

Taxonomical group[b]	Species	Enzymes[c]		
		UA	UC	A[d]
[Zygomycota]				
Zygomycetes/Mucorales	*Rhizopus oryzae* RA 99-880*	-	-	-
	Phycomyces blakesleeanus NRRL1555 v2.0	-	-	-
	Mucor circinelloides CBS277.49, v2.0	-	-	-
[Chytridiomycota]				
Chitridiomycetes/Chytridiales	*Batrachochytrium dendrobatidis* JEL423	-	-	-
[Basidiomycota/Agaricomycotina]				
Tremellomycetes /Tremellales	*Cryptococcus neoformans* H99*	-	1	-
Homobasidiomycetes/Agaricales	*Coprinus cinereus* okayama7#130*	-	-	-
	Laccaria bicolor S238N-H82	-	-	-
Homobasidiomycetes/Boletales	*Serpula lacrymans* S7.3 v2.0	-	1	-
[Basidiomycota/Ustilaginomycotina]				
Ustilaginomycetes/Ustilaginales	*Ustilago maydis* 521*	-	-	-
[Basidiomycota/Pucciniomycotina]				
Microbotryomycetes/Sporidiobolales	*Sporobolomyces roseus* v1.0	-	1	-
[Ascomycota/Taphrinomycotina]				
Schizosaccharomycetes/Schizosaccharomycetales	*Schizosaccharomyces pombe* 972h-*	-	-	-
[Ascomycota/Pezizomycotina]				
Eurotiomycetes/Onygenales	*Microsporum gypseum* CBS118893	-	-	-
	Microsporum canis CBS113480	-	-	-
	Trichophyton equinum CBS127.97	-	-	-
	Coccidioides immitis RS*	-	-	-
	Coccidioides immitis RMSCC 2394	-	-	-
	Coccidioides immitis RMSCC 3703	-	-	-
	Coccidioides immitis H538.4	-	-	-
	Coccidioides posadasii RMSCC 3488	-	-	-
	Coccidioides posadasii str. Silveira	-	-	-
	Histoplasma capsulatum G186AR	-	-	-
	Histoplasma capsulatum H143	-	-	-
	Histoplasma capsulatum H88	-	-	-
	Histoplasma capsulatum NAm1	-	-	-
	Blastomyces dermatitidis SLH14081	-	-	-
	Blastomyces dermatitidis ER-3	-	-	-
	Paracoccidioides brasiliensis Pb01	-	1	-
	Paracoccidioides brasiliensis Pb03	-	1	-
	Paracoccidioides brasiliensis Pb18	-	1	-
Eurotiomycetes/Eurotiales	*Aspergillus nidulans* FGSC A4*	-	1	-
	Aspergillus fumigatus Af293*	-	1	-
	Neosartorya fischeri NRRL 181	-	1	-
	Aspergillus terreus NIH2624*	-	1	-

	Aspergillus oryzae RIB40 / ATCC 42149*	-	-	-
	Aspergillus carbonarius ITEM 5010 v3	-	1	-
	Aspergillus clavatus NRRL 1	-	1	-
	Aspergillus flavus NRRL 3357	-	1	-
	Aspergillus niger ATCC 1015	-	1	1
Dothideomycetes/Capnodiales	*Mycosphaerella graminicola* v2.0*	-	-	1
	Mycosphaerella fijiensis v2.0	-	1	1
Dothideomycetes/Pleosporales	*Alternaria brassicicola* ATCC 96866	-	1	-
	Stagonospora nodorum SN15*	-	1	1
	Cochliobolus heterostrophus C5*	-	1	1
	Pyrenophora tritici-repentis Pt-1C-BFP	-	1	1
Leotiomycetes/Helotiales	*Botrytis cinerea* B05.10*	-	-	-
	Sclerotinia sclerotiorum 1980	-	-	1
Sordariomycetes/Sordariales	*Neurospora crassa* OR74A*	-	-	-
	Chaetomium globosum CBS 148.51	-	-	-
Sordariomycetes/Magnaporthales	*Magnaporthe oryzae* ATCC 64411*	1	-	(1)
Sordariomycetes/Hypocreales	*Nectria haematococca* v2.0*	1	1	(1)
	Fusarium graminearum PH-1 (NRRL 31084)*	1	-	(1)
	Fusarium oxysporum 4286*	1	1	(1)
	Fusarium verticillioides 7600*	1	1	(1)
	Trichoderma virens Gv29-8 v2.0	1	1	(1), 1
[Ascomycota/Saccharomycotina]				
Saccharomycetes/Saccharomycetales	*Yarrowia lipolytica* CLIB122*	2	-	(2)
	Candida albicans SC5314*	1	-	(1)
	Candida albicans WO1	1	-	(1)
	Candida parapsilosis isolate 317 from CDC	1	-	(1)
	Candida lusitaniae ATCC 42720*	1	-	(1)
	Debaryomyces hansenii CBS767*	1	-	(1)
	Ashbya gossypii ATCC 10895*	1	-	(1)
	Candida glabrata CBS138*	1	-	(1)
	Saccharomyces cerevisiae S288C*	1	-	(1)
	Saccharomyces cerevisiae RM11-1a	1	-	(1)

[a]See Table A4.4 for sequence sources.

[b]The phylum/subphylum (in square brackets) and class are given.

[c]See Figure 4.1 for the enzyme name abbreviations. The number of sequences found from each genome is shown. '-' indicates that no similar sequence was found.

[d]The amidase sequences that are a part of the urea amidolyase sequences are shown in parentheses.

*These fungal species are used in our further analysis

Table A4.4. Sequence sources of the urea amidolyase, urea carboxylase, and amidase from 64 fungal species.

Taxonomical group[a]	Species	Source[b]	Enzymes[c] UA	UC	A[d]
[Zygomycota]					
Zygomycetes/ Mucorales	*Rhizopus oryzae* RA 99-880*	FGI	-	-	-
	Phycomyces blakesleeanus NRRL1555 v2.0	JGI	-	-	-
	Mucor circinelloides CBS277.49, v2.0	JGI	-	-	-
[Chytridiomycota]					
Chitridiomycetes/ Chytridiales	*Batrachochytrium dendrobatidis* JEL423	FGI	-	-	-
[Basidiomycota/ Agaricomycotina]					
Tremellomycetes / Tremellales	*Cryptococcus neoformans* H99*	JGI	-	CNAG_07944	-
Homobasidiomycetes/ Agaricales	*Coprinus cinereus* okayama7#130*	JGI	-	-	-
	Laccaria bicolor S238N-H82	JGI	-	-	-
Homobasidiomycetes/ Boletales	*Serpula lacrymans* S7.3 v2.0	JGI	-	169686	-
[Basidiomycota/ Ustilaginomycotina]					
Ustilaginomycetes/ Ustilaginales	*Ustilago maydis* 521*	FGI	-	-	-
[Basidiomycota/ Pucciniomycotina]					
Microbotryomycetes/ Sporidiobolales	*Sporobolomyces roseus* v1.0	JGI	-	21475	-
[Ascomycota/ Taphrinomycotina]					
Schizosaccharomycetes/ Schizosaccharomycetales	*Schizosaccharomyces pombe* 972h-*	Sanger	-	-	-
[Ascomycota/ Pezizomycotina]					
Eurotiomycetes/ Onygenales	*Microsporum gypseum* CBS118893	FGI	-	-	-
	Microsporum canis CBS113480	FGI	-	-	-
	Trichophyton equinum CBS127.97	FGI	-	-	-
	Coccidioides immitis RS*	FGI	-	-	-
	Coccidioides immitis RMSCC 2394	FGI	-	-	-
	Coccidioides immitis RMSCC 3703	FGI	-	-	-
	Coccidioides immitis H538.4	FGI	-	-	-
	Coccidioides posadasii RMSCC 3488	FGI	-	-	-
	Coccidioides posadasii str. Silveira	FGI	-	-	-
	Histoplasma capsulatum G186AR	FGI	-	-	-
	Histoplasma capsulatum H143	FGI	-	-	-
	Histoplasma capsulatum H88	FGI	-	-	-
	Histoplasma capsulatum NAm1	FGI	-	-	-
	Blastomyces dermatitidis SLH14081	FGI	-	-	-
	Blastomyces dermatitidis ER-3	FGI	-	-	-

	Paracoccidioides brasiliensis Pb01	FGI	-	PAAG_02163	-
	Paracoccidioides brasiliensis Pb03	FGI	-	PABG_02398	-
	Paracoccidioides brasiliensis Pb18	FGI	-	PADG_00734	-
Eurotiomycetes/ Eurotiales	*Aspergillus nidulans* FGSC A4*	FGI	-	ANID_00887T0	-
	Aspergillus fumigatus Af293*	FGI	-	Afu1g15520	-
	Neosartorya fischeri NRRL 181	FGI	-	NFIA_009890	-
	Aspergillus terreus NIH2624*	FGI	-	ATET_05246	-
	Aspergillus oryzae RIB40 / ATCC 42149*	FGI	-	-	-
	Aspergillus carbonarius ITEM 5010 v3	JGI	-	10485	-
	Aspergillus clavatus NRRL 1	FGI	-	ACLA_019830	-
	Aspergillus flavus NRRL 3357	FGI	-	AFL2T_01101	-
	Aspergillus niger ATCC 1015	FGI	-	e_gw1_1.1117	fge1_pg_C_12000388
Dothideomycetes/ Capnodiales	*Mycosphaerella graminicola* v2.0*	JGI	-	-	75341
	Mycosphaerella fijiensis v2.0	JGI	-	41182	82172
Dothideomycetes/ Pleosporales	*Alternaria brassicicola* ATCC 96866	JGI	-	AB06360.1	-
	Stagonospora nodorum SN15*	FGI	-	SNOT_02186	SNOT_08324
	Cochliobolus heterostrophus C5*	JGI	-	57707	29777
	Pyrenophora tritici-repentis Pt-1C-BFP	JGI	-	PTRG_09405	PTRG_11638
Leotiomycetes/ Helotiales	*Botrytis cinerea* B05.10*	FGI	-	-	-
	Sclerotinia sclerotiorum 1980	FGI	-	-	SS1T_04628
Sordariomycetes/ Sordariales	*Neurospora crassa* OR74A*	FGI	-	-	-
	Chaetomium globosum CBS 148.51	FGI	-	-	-
Sordariomycetes/ Magnaporthales	*Magnaporthe oryzae* ATCC 64411*	FGI	MGG_04386	-	
Sordariomycetes/ Hypocreales	*Nectria haematococca* v2.0*	JGI	79968	44732	
	Fusarium graminearum PH-1 (NRRL 31084)*	FGI	FGSG_10913	-	
	Fusarium oxysporum 4286*	FGI	FOXG_12848	FOXG_07646	
	Fusarium verticillioides 7600*	FGI	FVEG_11593T0	FVEG_04571T0	
	Trichoderma virens Gv29-8 v2.0	JGI	53233	67729	42211
[Ascomycota/ Saccharomycotina]					
Saccharomycetes/ Saccharomycetales	*Yarrowia lipolytica* CLIB122*	Géno	YALI0E07271g YALI0E35156g	-	
	Candida albicans SC5314*	CGD	orf19_780	-	
	Candida albicans WO1	FGI	CAWT_00928	-	
	Candida parapsilosis isolate 317 from CDC	FGI	CPAG_03627	-	
	Candida lusitaniae ATCC 42720*	FGI	CLUT_00442	-	
	Debaryomyces hansenii CBS767*	Géno	DEHA2D07040g	-	
	Ashbya gossypii ATCC 10895*	NCBI	45187924	-	
	Candida glabrata CBS138*	Géno	CAGL0M05533g	-	
	Saccharomyces cerevisiae S288C*	SGD	YBR208C	-	
	Saccharomyces cerevisiae RM11-1a	FGI	SCRT_02761	-	

[a]The phylum/subphylum (in square brackets) and class/order are given.

[b]FGI: Fungal Genome Initiative (http://www.broadinstitute.org/science/projects/fungal-genome-initiative/fungal-genome-initiative), JGI: Joint Genome Institute (http://www.jgi.doe.gov), Sanger: The *S. pombe* Genome Project (http://www.sanger.ac.uk/Projects/Fungi/), Géno: Génolevures Genomic Exploration of the Hemiascomycete Yeasts (http://www.genolevures.org]), CGD: Candida Genome Database (http://www.candidagenome.org/), NCBI: National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), and SGD: Saccharomyces Genome Database (http://www.yeastgenome.org/).

[b]See Figure 4.1 for the enzyme name abbreviations. '-' indicates that no similar sequence was found.

*****These fungal species are used in our further analysis.

Table A4.5. Sequence sources of the urease, methylcrotonoyl-CoA carboxylase, and propionyl-CoA carboxylase from the selected 27 fungal species.[a]

	Enzymes[b]		
Species	**Urease**	**MccA**	**PccA**
Rhizopus oryzae RA 99-880	RO3G_06489	RO3G_06576	RO3G_06560
Ustilago maydis 521	UM06045	UM04382	-
Cryptococcus neoformans H99	CNAG_05540	CNAG_01680	-
Coprinus cinereus okayama7#130	CC1G_10059	CC1G_13741	CC1G_05511
Schizosaccharomyces pombe 972h-	SPAC1952_11c	-	-
Coccidioides immitis RS	CIMT_05193	CIMT_07030	CIMT_05331
Aspergillus nidulans FGSC A4	AN10079	AN4690	AN7764
Aspergillus fumigatus Af293	Afu1g04560	Afu5g08910	Afu5g07580
Aspergillus terreus NIH2624	ATET_03748	ATET_06576	ATET_08368
Aspergillus oryzae RIB40 / ATCC 42149	AO090003000879	AO090020000495	-
Mycosphaerella graminicola v2.0	85598	70525	109805
Stagonospora nodorum SN15	SNOT_11285	SNOT_09555	SNOT_12342
Cochliobolus heterostrophus C5	95543	78650	105664
Botritis cinerea B05.10	BC1T_13063	BC1T_08870	BC1T_02620
Neurospora crassa OR74A	NCU03127	NCU00591	-
Magnaporthe oryzae ATCC 64411	MGG_01324	MGG_10320	-
Nectria haematococca v2.0	65875	92030	-
Fusarium graminearum PH-1 (NRRL 31084)	FGSG_00740 FGSG_10627	FGSG_08688	-
Fusarium oxysporum 4286	FOXG_01071 FOXG_17146	FOXG_03110	-
Fusarium verticillioides 7600	FVEG_00443	FVEG_01973	-
Yarrowia lipolytica CLIB122	-	YALI0B14619g	-
Candida albicans SC5314	-	-	-
Candida lusitaniae ATCC 42720	-	-	-
Debaryomyces hansenii CBS767	-	-	-
Ashbya gossypii ATCC 10895	-	-	-
Candida glabrata CBS138	-	-	-
Saccharomyces cerevisiae S288C	-	-	-

[a]See Table A4.4 for the data source for each genome.

[b]See Figure 4.1 for the enzyme name abbreviations. The sequences IDs found from each genome is shown. '-' indicates that no similar sequence was found.

Table A4.6. Sequence sources of urea amidolyase, urea carboxylase, and amidase in eubacterial genomes.

			Enzymes[b]			
Phylum or Class	**Species**	**ACC#[a]**	**UA**	**UC**	**A**	**Urease**
Alphaproteobacteria	*Caulobacter crescentus NA1000*	NC_011916	-	221234842	221234843	-
	Asticcacaulis excentricus CB 48	NZ_ACQR00000000	-	241771960	241771961	-
	Sinorhizobium medicae WSM419	NC_009636	-	-	-	150397583 150397586 150397588
Betaproteobacteria	*Achromobacter piechaudii ATCC 43553*	NZ_ADMS00000000	-	293607215	293607216	-
	Bordetella pertussis Tohama I	NC_002929	-	-	-	33594086 33594087 33594089
	Nitrosomonas europaea ATCC 19718	NC_004757	-	30250344 30250340*	-	-
	Neisseria meningitidis FAM18	NC_008767	-	-	-	-
	Burkholderia sp. CCGE1001	NZ_ADDJ00000000	-	282888296	282888297	282888448 282888449 282888450
Gammaproteobacteria	*Escherichia coli O111:H- str. 11128*	NC_013364	-	-	-	260867324 260867323 260867322
	Yersinia pestis Angola	NC_010159	-	-	-	162421917 162421306
	Haemophilus influenzae 86-028NP	NC_007146	-	-	-	68249136 68249137 68249138
	Pantoea ananatis LMG 20103	NC_013956	291616199	291619625	291616199	-
	Pantoea sp. At-9b	NZ_ACYJ00000000	-	258639802 258639881	258639803	-
	Shewanella oneidensis MR-1	NC_004347	-	-	-	-
	Pseudomonas aeruginosa LESB58	NC_011770	-	-	-	218893963 218893960

						218893962
	Coxiella burnetii Dugway 5J108-111	NC_009727	-	-	-	-
	Pectobacterium carotovorum subsp. carotovorum PC1	NC_012917	-	253688548 253688770	253688549	-
	Xanthomonas campestris pv. campestris str. B100	NC_010688	-	-	-	-
	Cellvibrio japonicus Ueda107	NC_010995	-	192360305 192360281	192360851	-
	Teredinibacter turnerae T7901	NC_012997	-	254787389	254787390	254788040 254788041 254788042
	Marinomonas sp. MED121	NZ_AANE00000000	-	87119094	87119095	87120670 87120669
	Klebsiella pneumoniae 342	NC_011283	-	206578981	206581101	206580665 206579658 206580264
	Pseudomonas fluorescens SBW25	NC_012660	-	-	-	229588129 229588133 229588130
Deltaproteobacteria	*Geobacter sp. M21*	NC_012918	-	-	-	-
	Sorangium cellulosum 'So ce 56'	NC_010162	-	162453191	-	162454831 162454830
Epsilonproteobacteria	*Helicobacter pylori B38*	NC_012973	-	-	-	254778798 254778799
	Wolinella succinogenes DSM 1740	NC_005090	-	34557492	34557494	-
Acidobacteria	*Acidobacterium capsulatum ATCC 51196*	NC_012483	-	-	-	-
	Solibacter usitatus Ellin6076	NC_008536	-	116619994	116619993	-
Cyanobacteria	*Synechococcus sp. PCC 7002*	NC_010475	-	-	-	170076824 170079032 170077934
	Gloeobacter violaceus PCC 7421	NC_005125	-	37520527	37520530	-
	Cyanothece sp. PCC 7425	NC_011884	-	220907713	220908629	220907679 220907680 220907681

Deinococcus-Thermus	*Thermus thermophilus HB8*	NC_006461	-	-	-	-
	Deinococcus deserti VCD115	NC_012526	-	-	-	-
Chloroflexi	*Dehalococcoides ethenogenes 195*	NC_002936	-	-	-	-
Aquificae	*Aquifex aeolicus VF5*	NC_000918	-	-	-	-
Thermotogae	*Thermotoga maritima MSB8*	NC_000853	-	-	-	-
Fusobacteria	*Fusobacterium nucleatum subsp. nucleatum ATCC 25586*	NC_003454	-	-	-	-
Verrucomicrobia	*Verrucomicrobium spinosum DSM 4136*	NZ_ABIZ00000000	-	171912641	171912640	171911815 171911816 171911817
Chlamydiae	*Chlamydophila pneumoniae CWL029*	NC_000922	-	-	-	-
	Chlamydia trachomatis B/TZ1A828/OT	NC_012687	-	-	-	-
Bacterioidetes	*Porphyromonas gingivalis W83*	NC_002950	-	-	-	-
Chlorobi	*Chlorobium limicola DSM 245*	NC_010803	-	-	-	-
Fibrobacteres	*Fibrobacter succinogenes subsp. succinogenes S85*	NC_013410	-	-	-	-
Actinobacteria	*Mycobacterium tuberculosis F11*	NC_009565	-	-	-	148823061 148823060 148823059
	Corynebacterium aurimucosum ATCC 700975	NC_012590	-	-	-	-
	Streptomyces avermitilis MA-4680	NC_003155	-	29833240	29833239	29833648 29829257 29829258 29833647 29833646
	Bifidobacterium longum subsp. infantis ATCC 15697	NC_011593	-	-	-	213691032 213691031
Spirochaetes	*Borrelia burgdorferi ZS7*	NC_011728	-	-	-	-
	Treponema denticola ATCC 35405	NC_002967	-	-	-	-

Planctomycetes	*Rhodopirellula baltica SH 1*	NC_005027	-	-	-	-
Firmicutes	*Clostridium botulinum A2 str. Kyoto*	NC_012563	-	-	-	-
	Mycoplasma hyopneumoniae 7448	NC_007332	-	-	-	-
	Streptococcus pneumoniae 70585	NC_012468	-	-	-	-
	Bacillus anthracis str. CDC 684	NC_012581	-	-	-	-
	Roseburia intestinalis L1-82	NZ_ABYJ00000000	-	240144639	240144640	-

[a]All the bacterial sequences were downloaded from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

[b]See Figure 4.1 for the enzyme name abbreviations. '-' indicates that no similar sequence was found.

*This UC sequence was only 780 amino acids long, consisting of incomplete urea carboylase domain. Hence it was not used in phylogenetic analysis.

Table A4.7. Distance between amidase and urea carboxylase genes in eubacterial genomes.

Species	No. of bp between UC and A[a]
Caulobacter crescentus NA1000	-2 (0)
Asticcacaulis excentricus CB 48	0 (0)
Achromobacter piechaudii ATCC 43553[b]	18 (0)
Burkholderia sp. CCGE1001	61 (0)
Pantoea sp. At-9b[b]	2 (0)
Pectobacterium carotovorum subsp. carotovorum PC1[b]	-6 (0)
Cellvibrio japonicus Ueda107	118 (0)
Teredinibacter turnerae T7901	-6 (0)
Marinomonas sp. MED121	25 (0)
Klebsiella pneumoniae 342	-2 (0)
Wolinella succinogenes DSM 1740	943 (1)
Solibacter usitatus Ellin6076	2 (0)
Gloeobacter violaceus PCC 7421	1,701 (2)
Cyanothece sp. PCC 7425	979,743 (916)
Verrucomicrobium spinosum DSM 4136	118 (0)
Streptomyces avermitilis MA-4680	16 (0)
Roseburia intestinalis L1-82	16 (0)

[a]See Figure 4.1 for the enzyme name abbreviations. The number of genes present between UC and A genes are given in parentheses. Negative distances indicate that these two genes are overlapped.

[b]These species have two copies of the urea carboxylase (UC) gene. The UC gene in this table is the one that is closest to the A gene in the respective genome.

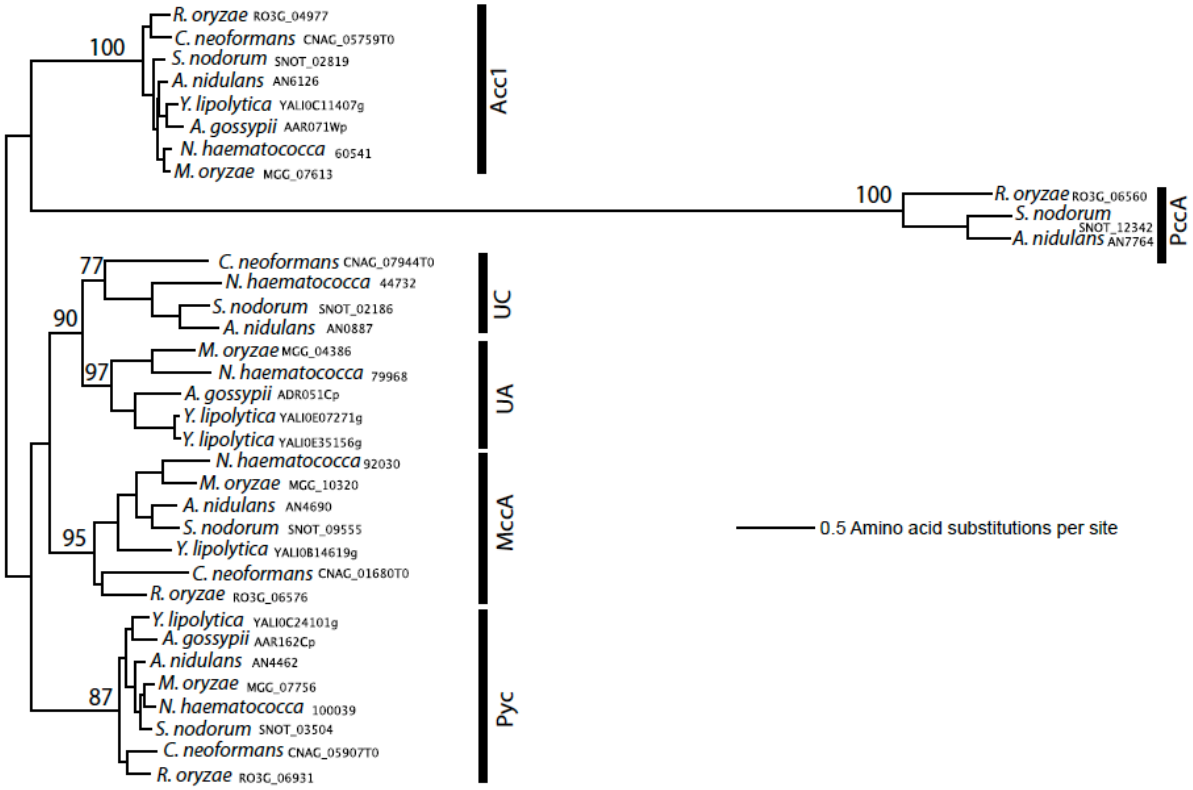

Figure A4.1. Maximum-likelihood phylogeny of carboxylation domain sequences. The maximum-likelihood phylogeny was reconstructed using the carboxylation domain sequences of the following fungal proteins : urea amidolyase (UA), urea carboxylase (UC), acetyl-CoA carboxylase (Acc1), propionyl-CoA carboxylase (PccA), methylcrotonoyl-CoA carboxylase (MccA) and pyruvate carboxylase (PyC). The numbers above the internal branches show bootstrap values (%). Only values ≥ 70 are shown.

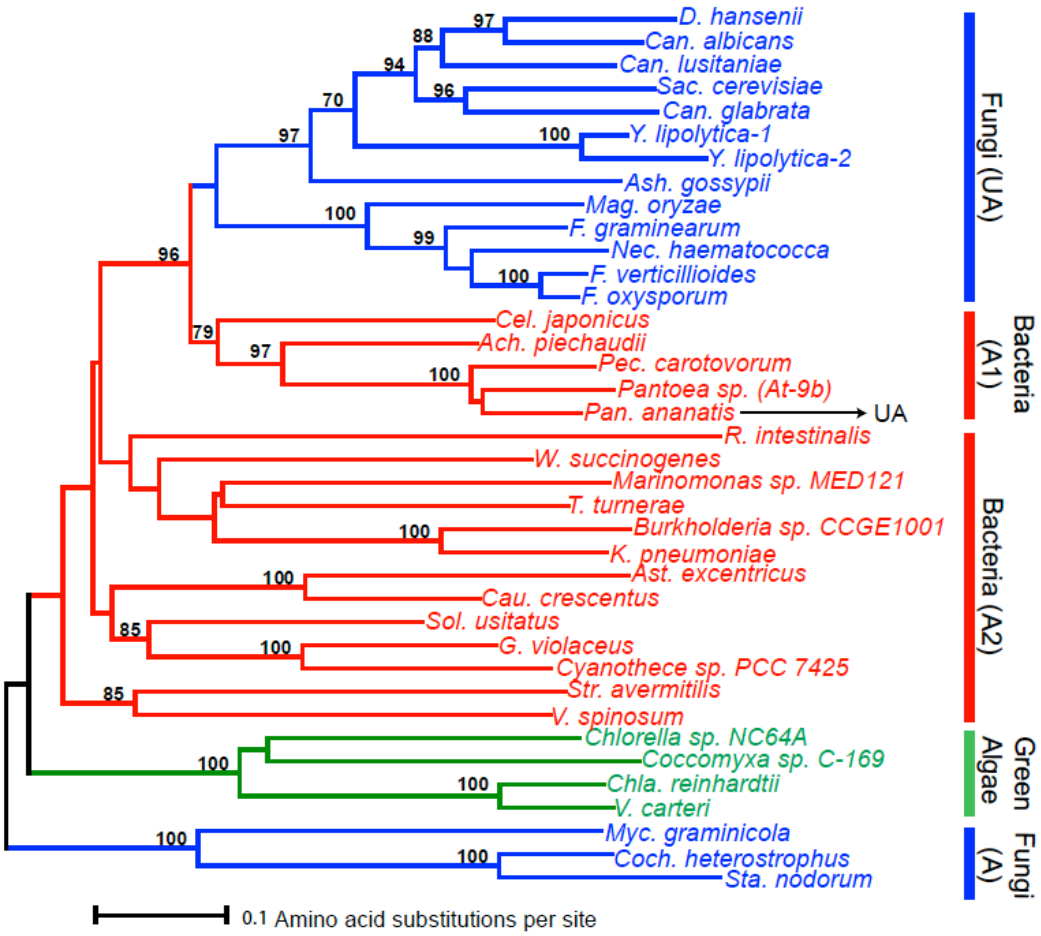

Figure A4.2. Minimum-evolution phylogeny of amidase protein sequences. The minimum-evolution phylogeny was reconstructed using the protein sequences from the amidase domains of the urea amidolyase proteins and the amidase proteins. The numbers above the internal branches show bootstrap values (%). Only values ≥ 70 are shown. Branches are colored as follows: blue for fungi, green for green algae, and red for bacteria. The bacterial urea carboxylase groups denoted by A1 and A2 correspond with the same groups in Figure 3.

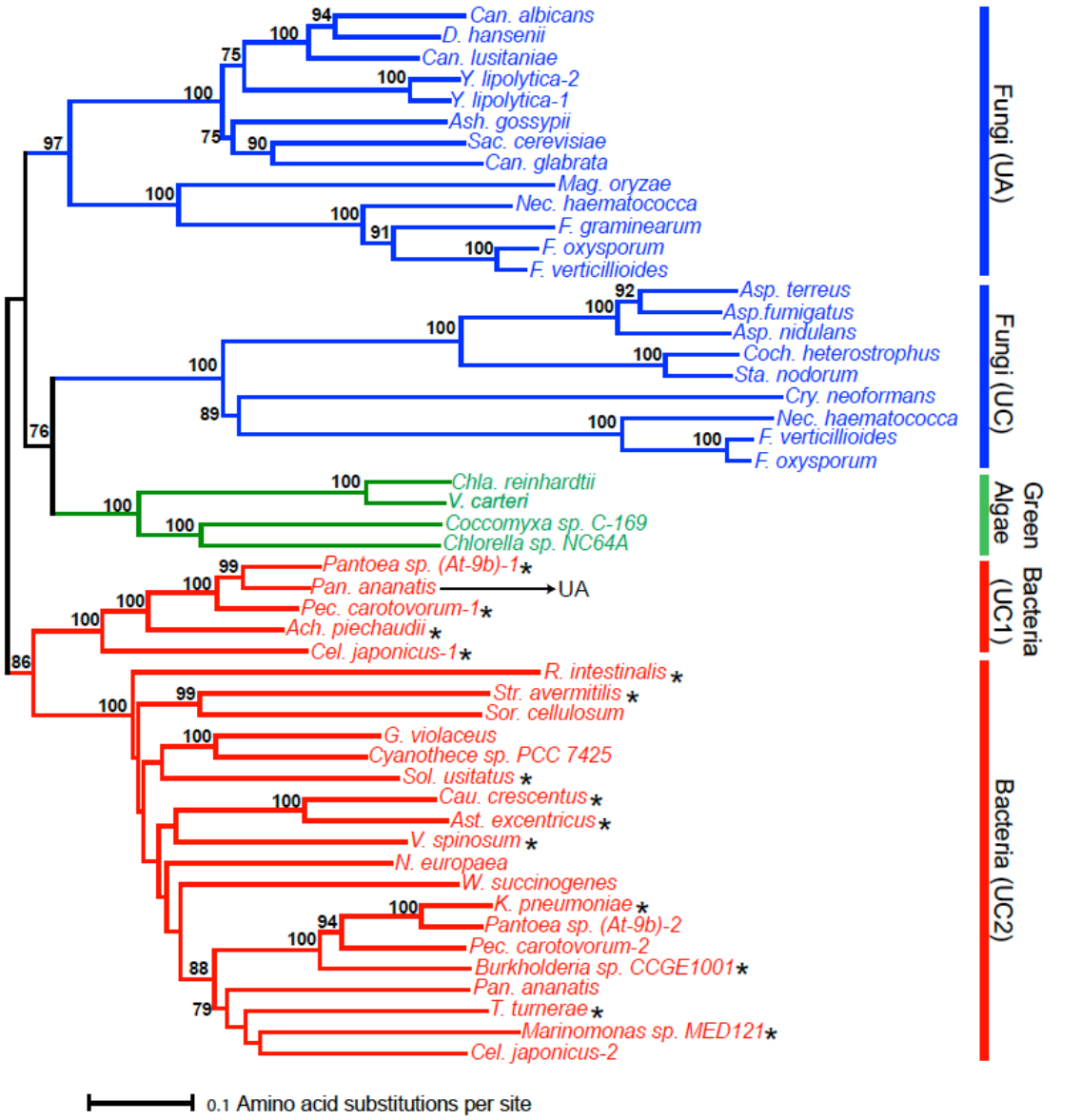

Figure A4.3. Minimum-evolution phylogeny of urea carboxylase protein sequences. The minimum-evolution phylogeny was reconstructed using the protein sequences from the urea-carboxylase domains of the urea amidolyase proteins and the urea carboxylase proteins. The numbers above the internal branches show bootstrap values (%). Only values ≥ 70 are shown. Branches are colored as follows: blue for fungi, green for green algae, and red for bacteria. The bacterial urea carboxylase groups denoted by UC1 and UC2 correspond with the same groups in Figure 4. The asterisks beside the bacterial names indicate that their urea carboxylase genes are next to the amidase genes in their genomes.

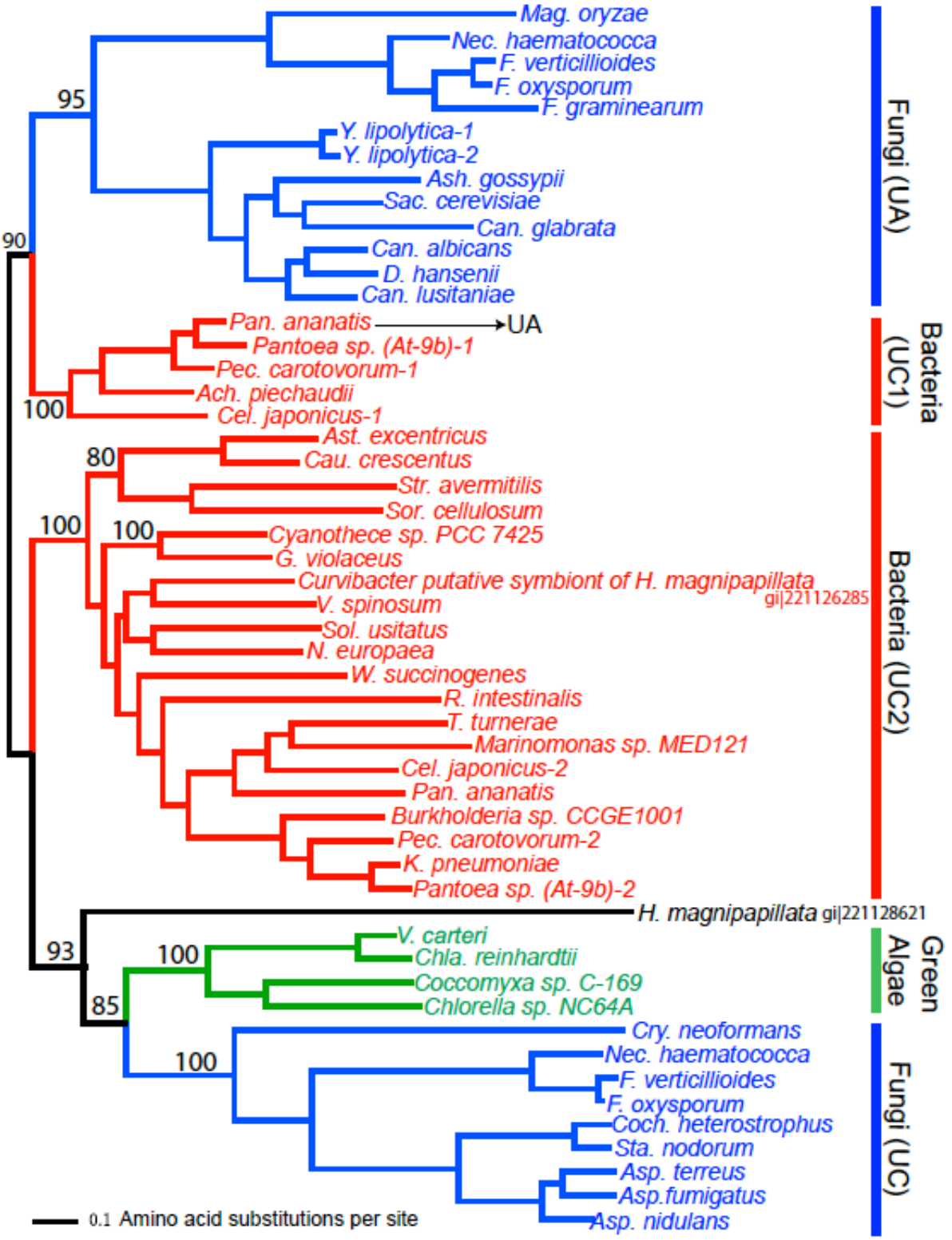

Figure A4.4. Maximum-likelihood phylogeny of urea carboxylase protein sequences including the two sequences found in *Hydra magnipapillata*. The maximum-likelihood phylogeny was reconstructed using the protein sequences from the urea carboxylase domains of the urea amidolyase proteins and the urea carboxylase proteins. The numbers above or below the internal branches show bootstrap values (%). Only bootsrap values equal to or higher than 70% are shown. Branches are colored as follows: black for metazoan, blue for fungi, green for green algae, and red for bacteria.

Chapter 5

Molecular Evolution of Sterol-Sensing Domain Proteins in Eukaryotes

5.0 Preface for Chapter 5

In this chapter, I studied the molecular evolution of sterol-sensing domain (SSD) proteins in representative species of all kingdoms of life. Sterol-sensing domain is known to "sense" sterol-levels in the cell and thereby regulates the synthesis and transport of sterols by various different pathways. These proteins include the hydroxymethylglutaryl-CoA reductase (HMGCR), SREBP (sterol regulatory element binding protein) cleavage activating protein (SCAP), and Niemann-Pick C-1 type protein (NPC1). The SSD is also a part of signaling proteins Patched (PTC) and Dispatched (DISP), as well as Patched-related (PTC-R) proteins. Both PTC and DISP are involved in hedgehog signaling pathway for cell differentiation, where a cholesterol-bound ligand molecule is involved. The distribution of the SSD domain showed that this is present in all eukaryotes and remotely similar sequences are also present in bacteria. Phylogenetic analyses showed that these ancestral proteins evolved into DISP, PTC, PTC-R, and NPC1 acquiring their specific functions and in some cases getting lost or replaced in various lineages. We also showed that HMGCR with SSD, and SCAP may have been formed as results of domain acquisition.

5.1 Introduction

Sterols are important components of cell membranes and are precursors to hormones. The major sterol of vertebrates and fungi is cholesterol and ergosterol, respectively. Plants possess varying compositions of stigmasterol and sitosterol as the major sterol [1]. In animals, cholesterol helps to regulate membrane fluidity and create a semipermeable barrier between cellular compartments. It modulates functions of membrane proteins and plays roles in membrane trafficking and transmembrane-signaling processes. It also plays significant roles in diabetes, cancers, and other diseases related to the heart and brain [2]. For the proper functioning of the cell, the amount of sterol present at any time needs to be carefully regulated. Sterol homeostasis is maintained by several feedback controls that include transcriptional and post-transcriptional mechanisms [3].

The sterol-sensing (or sterol-regulatory) domain is present in multiple proteins that have a common property of sterol homeostasis with varying functions. The sterol-sensing domain (SSD) is ~180 amino-acids long and conserved in at least six families of proteins (see Figure 5.1): hydroxymethylglutaryl-CoA reductase (HMGCR), SREBP (sterol regulatory element binding protein) cleavage activating protein (SCAP), Niemann-Pick C-1 type protein (NPC1), Patched (PTC), Patched-related (PTC-R), and Dispatched (DISP) [4]. All these classes of proteins have functions related to sterols. The SSD encompasses five transmembrane helices and is involved in sterol-level sensing in the cell.

HMGCR is a sterol biosynthetic enzyme that is degraded when sterol levels are high. Along with the SSD, it consists of a catalytic domain about 400 amino acids long. SCAP is responsible for regulating SREBP, a transcription factor of cholesterol biosynthetic genes. Apart from SSD, the only other known domain this protein has is the WD-40 repeat (InterPro: IPR001680). For example, in the human SCAP, there are seven

WD repeats spanning 450 amino-acid long region. In both of these proteins, higher levels of sterols cause the SSD to bind to Insigs (proteins coded by the insulin induced gene) in the endoplasmic reticulum (ER). While Insig-bound SCAP is retained in the ER, Insig-bound HMGCR is ubiquitinated and degraded [5]. Yabe *et al*. [6] showed that three different point mutations in the SSD of SCAP prevent sterol-induced binding of SCAP to Insig.

NPC1 is responsible for intracellular transport of sterol. Niemann-Pick type C disease, caused by mutations in the NPC1 gene, is a fatal lipid storage disorder, which is characterized by lysosomal cholesterol accumulation. Millard *et al*. [7] showed that mutations in the SSD regions of NPC1 disrupt normal transportation of cholesterol in the cells. PTC plays a role in cell differentiation during development and morphogenesis. It is a receptor of the hedgehog protein, a ligand that is bound to cholesterol [8]. DISP is involved in releasing the cholesterol-bound hedgehog from the cell [9]. Both PTC and DISP are key proteins of the hedgehog-signaling pathway that is essential for development and differentiation.

The actual binding of cholesterol or cholesterol analog has only been shown in NPC1 and SCAP. A functional SSD is required in NPC1 for binding of a cholesterol analog [10]. Binding of cholesterol has been demonstrated in SCAP at an octahelical region that includes the SSD [11]. Mutations in SSD can be lethal to cells and cause various diseases due to the abruption of cholesterol homeostasis in cells and signaling pathways. The role of SSD in sterol homeostasis in cells and cell differentiation makes it an important target for bio-medical research in understanding and curing cholesterol-related diseases.

Among the six different families of SSD proteins in eukaryotes, only HMGCR proteins have high sequence similarities to prokaryotic proteins. The prokaryotic HMGCR, however, lacks the SSD region. Similarities in membrane topology have been found between NPC1 and members of the resistance-nodulation-division (RND) family

of prokaryotic permeases [12]. The RND protein superfamily is a ubiquitous group of permeases originally identified in bacteria; now known to have representation in all major kingdoms [13]. This superfamily has been defined into seven subfamilies by Tseng *et al*. [13]. One of them consists of the eukaryotic sterol homeostasis (ESH) family of proteins, which includes HMGCR, SCAP, PTC and NPC1, while the rest of the subfamilies is made up of bacterial and archaeal permeases. Expression of human NPC1 in *Escherichia coli* showed that this protein was able to transport acriflavine and fatty acids and act as a bacterial permease [14]. There are also RND transporters in bacteria that are predicted to have functions related to hopanoid biosynthesis (InterPro: IPR017841). Hopanoids are sterol analogs in bacteria [1]. These results, and the similarities between SSD proteins and the other members of the RND superfamily show that the bacterial permeases could be the ancestral proteins to the eukaryotic SSD proteins [15].

Despite its importance, SSD proteins have not been thoroughly studied as a protein family distributed among all kingdoms of life. In order to elucidate the molecular evolution of SSD and related protein families, we examined SSD sequences in various eukaryotic species. While metazoans consisted of all the six types of SSD-containing proteins, fungi lacked DISP, PTC, and PTC-R. Land plants consisted of only the NPC1 while some green algae consisted only of PTC, PTC-R and DISP. Basal eukaryotes possessed NPC1, DISP, PTC and PTC-R. We also identified HMGCR proteins with and without SSDs. Based on the evolutionary relationships among these HMGCR proteins, we discussed how their functions and domains have been acquired during the evolution of this protein family.

5.2 Materials and Methods

SSD sequences used

Seventy-two annotated SSD-containing proteins (Prosite profile PS10156 [16])

were gathered from the UniProt database [17]. The SSD regions predicted from these proteins (only those longer than 100 amino acids) were extracted. The resulting 67 SSD sequences were used to build a multiple alignment using MAFFT (version 6.240; [18]) with default parameters (FFT-NS-2, a progressive FFT alignment with two tree-building cycles). The maximum likelihood phylogeny [19] was reconstructed by RAXML (version 7.0.4 [20]) using '-m PROTMIXWAG' to use WAG amino-acid substitution model [21] with a fixed number approximation followed by a refined gamma-model of rate heterogeneity and '-x 1234' to set the random seed. Based on the phylogeny identical or highly similar SSD sequences were removed. Four bacterial sequences were annotated as SSD-containing by Prosite. However, these bacterial sequences are extremely diverged compared to the eukaryotic SSD sequences. As they do not align well with the eukaryotic SSD sequences, we did not include these bacterial sequences in our training dataset. After this, we had a total of 35 SSD sequences: 6 from DISP, 5 from PTC, 5 from PTC-R, 5 from NPC1, 12 from HMGCR, and 2 from SCAP.

Building the profile hidden Markov model for SSD sequences

A multiple alignment of these 35 sequences were built using MAFFT (version 6.240; [18]) with default parameters (FFT-NS-2, a progressive FFT alignment with two tree-building cycles). The alignment was used to build a profile hidden Markov model (HMM) using the HMMER software (version 3.0 [22]) with its program *hmmbuild* using default parameters: '-fast >=symfrac' where symfrac = 0.5 (for defining consensus columns as those that have at least 50% residues as opposed to gaps); '-wpb' (for using Henikoff position-based sequence weighing scheme [23] so that uneven phylogenetic representation in the training set will not bias the model); '-seed 42'; and a Dirichlet mixture priors.

Organisms searched

For both SSD and HMGCR, 98 complete eukaryotic genomes were searched. It included 54 fungi, 21 plants (including 9 green algae), 9 basal eukaryotes, and 14 metazoa. For SSD we also searched in 56 bacterial genomes from 14 phyla as representatives for the prokaryotes. These seuqnces were downloaded from the National Center for Biotechnology Information [24]. The sequence names and ACC # are given in Table 5.S1.

Profile HMM searches for SSD sequences

The entire predicted protein set from each genome was searched using the profile HMM. The program *hmmsearch* from HMMER(version 3.0 [22]) was used with default parameters such as -seed 42, reporting threshold e-value of 10 (-E) and inclusion threshold of evalue 0.01 (-incE). The total number of sequences in the database was set to 50,000 (-Z option) in order to obtain the e-values comparable among different genome sizes. The e-value cut-off used for eukaryotic SSD proteins was 1×10^{-11}. For bacterial SSD hits, it was set at 1×10^{-4} because bacterial proteins are very diverged from eukaryotic proteins unless they are results of recent horizontal transfer.

Phylogenetic analysis

Only the predicted SSD region was used to reconstruct phylogenetic trees. The multiple alignments were reconstructed using MAFFT (version 6.240; [18]) with default parameters (FFT-NS-2, a progressive FFT alignment with two tree-building cycles). The maximum-likelihood phylogeny [19] was reconstructed as implemented in raxmlHPC-MPI (version 7.0.4; [20]) using the following options: '-m PROTMIXWAG' to use WAG amino-acid substitution model [21] with a fixed number approximation followed by a refined gamma-model of rate heterogeneity, '-f a' for a rapid bootstrap analysis, '-x 1234'

to set the random seed, and '-# 1000' for 1000 pseudoreplicates for bootstrap analysis. To gather the bootstrap values, the 'consense' program of the Phylip package (v. 3.68, [25]) was used. Due to their extreme divergence, the sequences from *Branchiostoma floridae* and *Caenorhabditis elegans* were removed from the tree reconstruction.

Classification of SSD-containing proteins

SSD-containing proteins identified were classified into one of the six classes based on the phylogenetic clustering and reciprocal BLASTP (version 2.2.17 [26]) results as follows. All the search results were first classified according to the clustering pattern of the SSD phylogeny. These sequences were used as a query to search against the human proteome using BLASTP. When each sequence search resulted in any one of the six types of SSD proteins from the human proteome as the top hit(s), that query sequence was considered to be an SSD protein of that type. The default parameters were used with BLASTP program (version 2.2.17), which include BLOSUM62 scoring matrix, low-complexity filtering, gap-open and gap-extend penalty of 11 and 1, respectively. In order to obtain the E-values comparable across different size of the genomes, the "effective length of database" was set to 500,000,000 (using -z option).

HMGCR search

The human HMGCR sequence HMDH_HUMAN (P04035) of length 888 amino acids was used to search for HMGCR sequences that had the SSD as well as those that did not using BLASTP. This human HMGCR has the SSD region and the catalytic domain (see Figure 5.1). The SSD region was not found in all HMGCR proteins while the catalytic domain was present in all HMGCRs. Even though the SSD profile HMM search was able to find those HMGCRs with SSD, this BLASTP (version 2.2.17) search

was required to find those HMGCRs that lacked the SSD. For HMGCR hits, e-value cut-off was set at $1x10^{-16}$. The parameters used for BLASTP is the same as above. For the phylogenetic analysis of HMGCR in fungi, the complete sequences (~880 amino acid) as well as only the catalytic domain (~440 amino acid) were used. Phylogenies were reconstructed using RAXML with parameters described above.

5.3 Results and Discussion

5.3.1 Distribution of SSD proteins among eukaryotic genomes

The SSD was searched in genomes of 14 metazoan species, 54 fungal species, 21 plant species (including 9 green algae), and 9 basal eukaryotic species. SSD sequences were found in all 98 eukaryotic genomes we searched. Table 5.1 shows the distribution of SSD proteins found in metazoa. Besides the absence of SCAP in *Hydra magnipapillata,* all the metazoans had 1-2 copies each of HMGCR, SCAP, and NPC1. The HMGCR of *Nematostella vectensis* was missing the catalytic domain but had the SSD region, while the *Caenorhabditis elegans* HMGCR lacked the SSD region and had only the catalytic domain. The number of PTC, DISP, and PTC-R varied, with *C. elegans* and *Branchiostoma floridae* having the most number of PTC-R: 29 and 40, respectively. This may be due to similar gene expansions observed for a nuclear receptor gene in *C. elegans* [27] and G-protein coupled receptor genes in *B. floridae* [28].

In fungi, only three SSD-containing proteins, HMGCR, SCAP, and NPC1, were found (Table 5.2). DISP, PTC, and PTC-R were completely missing from all the 54 fungal genomes we searched. Furthermore, the SCAP protein is completely absent from Eurotiomycetes (phylum Ascomycota; subphylum Pezizomycotina) and Saccharomycetes (phylum Ascomycota; subphylum Sachcaromycotina) except for *Yarrowia lipolytica*. Although we did not find the SCAP protein from Chitridiomycetes either, we only have one representative (*Batrachochytrium dendrobatidis*) from this phylum. These results

indicate that there have been at least two independent gene-loss events during fungal evolution. The NPC1 is present in all fungal species except *B. dendrobatidis* (phylum Chitridiomycotina), *Schizosaccharomyces pombe* (phylum Ascomycota; subphylum Taphrinomycotina), and *Aspergillus niger* (phylum Ascomycota; subphylum Pezizomycotina). The loss of the NPC1 gene from *A. niger* is recent because all its closely related species from the Aspergillus group have this gene. However, we cannot determine if the loss of NPC1 from *B. dendrobatidis* and *S. pombe* is species- or lineage-specific since we have only one representative species each from the phylum Chitridiomycetes and the subphylum Taphrinomycotina. All fungi had a copy of SSD-containing HMGCR except for *Chaetomium globosum* (phylum Ascomycota; subphylum Pezizomycotina), which only had the HMGCR that lacked SSD. Several species consisted of both types of HMGCRs, with and without SSD. These occurrences were dispersed throughout the fungal kingdom, but seemed to be more prominent in the Aspergillus group of species. We will discuss the HMGCR in fungi later.

In plants, there was a complete absence of SSD-containing HMCGR and SCAP (Table 5.3). All higher plants possessed multiple copies of HMGCR without the SSD, while all the green algae lacked this enzyme. Green algae are shown to have an alternative sterol synthesis pathway called the deoxyxylulose 5-phosphate (DXP) pathway, which takes place in their plastids [29]. While fungi, animals and some bacteria appear to use the mevalonate pathway, many bacteria (including many human pathogens) and green algae appear to rely exclusively on the DXP pathway, and some algae, mosses and liverworts, marine diatoms, and higher plants appear to use both pathways [30]. So perhaps there is a relation between having the SSD in HMGCR or having an alternate sterol synthesis pathway in eukaryotes.

Most of the plants had NPC1 except for prasinophyte green algae (*Micromonas* and *Ostreococcas* species), which is considered to retain features of the ancestral green

lineage [31]. While PTC, DISP, and PTC-R were absent from higher plants (with an exception of a moss *Physcomitrella patens*), these genes were present in the prasinophyte green algae.

In basal eukaryotes (Table 5.4), the SSD-containing HMGCR was absent. The NPC1 was found in *Naegleria gruberi* (amoeboflagellate) and in two *Dictyostelium* species (Amoebozoa). PTC was present only in *Monosiga brevicolis* (Choanozoa). DISP and PTC-R were found in the species of Haptophyta, Stramenopiles, and Choanozoa. The presence of these genes in these basal eukaryotic organisms indicates that the ancestral SSD-containing proteins in eukaryotes may have been similar to NPC1, DISP, PTC, and PTC-R. This hypothesis is further discussed in the following sections.

5.3.2 Distribution of SSD proteins among prokaryotes

We also searched 54 eubacterial genomes for SSD sequences. We found 46 sequences that were similar to SSD (Table 5.S1). These were distant relatives as shown by their higher e-values ($e < 1 \times 10^{-4}$). These proteins were permeases/transporters, and they were in various bacterial classes with species having one or more similar sequences to the SSD (Table 5.S1). These sequences have high number of transmembrane regions (9-14). Some of these sequences are shorter (~350 amino acids) and are subunits of a larger transporter unit. Among the transporters are two sequences (gi|282886364 and gi|253699522) annotated as "hopanoid biosynthesis associated RND transporter like". Hopanoids are bacterial pentacyclic compounds whose primary function is in maintaining plasma membrane fluidity [1]. This function is similar to what cholesterol does in higher eukaryotes. Therefore, these proteins would be very likely candidates for bacterial versions of SSD proteins. As was found in our study as well as in a previous study [15], these bacterial RND genes were closest to the eukaryotic PTC, DISP, and NPC1 proteins in sequence similarity.

5.3.3 Phylogenetic analysis of the entire SSD-containing proteins

Figure 5.2 is the phylogenetic tree based on SSD sequences from all SSD-containing proteins found both in eukaryotes and in bacteria. The prokaryotic sequences are shown in red. The bootstrap support to separate the cluster of all bacterial proteins from eukaryotic proteins is high (93%). The outermost eukaryotic protein group is the DISP sequences (94% bootstrap support). The large number of transmembrane regions in DISP proteins (12) is also similar to the ones found in bacterial transporters. Metazoans, prasinophyte green algae, and basal eukaryotes have DISP proteins but none of the fungal species has it. The last common ancestral species of fungi must have lost this gene. The PTC-R and PTC proteins seem to have diverged next, although their phylogenetic relationships are not well supported except that the PTC proteins cluster together with 74% bootstrap support. Neither PTC nor PTC-R proteins were found in fungal species; however, metazoans, green algae and basal eukaryotes were well represented in these protein groups, as in the case for DISP. It is thus likely that fungi have lost PTC and PTC-R genes as well. The inner cluster encompassing NPC1, SCAP, and HMGCR protein groups is well supported (82% bootstrap value). The SSD regions of these three proteins seem to be more closely related to each other than with DISP, PTC, and PTC-R. Fungal and metazoan sequences are represented in all of the NPC1, SCAP, and HMGCR groups while basal eukaryotes and plants are represented only in the NPC1 protein group. As described later, plants do have HMGCR proteins. However, they lack SSD regions and this is why the plant HMCGRs are not included in this SSD-sequence based phylogeny. The phylogenetic analysis showed that SSD sequences of HMCGR and SCAP proteins are most closely related and their sister relationship is highly supported (95% bootstrap value). These proteins also have similar numbers of transmembrane regions (7-8, see Figure 5.1).

5.3.4 Phylogenetic analysis of HMGCR and SCAP proteins

We performed more detailed phylogenetic analysis using SSD sequences of representative HMGCR and SCAP proteins, the most closely related SSD-containing protein groups (Figure 5.3). The bootstrap analysis showed that HMGCR and SCAP proteins form strongly supported clusters with the 95% supporting value. These two genes were clearly present in the last common ancestor of fungi (colored in blue) and metazoans (colored in black). We did not find any duplication events during the SCAP gene evolution except for *B. floridae*. On the other hand, we found several cases of duplications in the HMGCR gene: in the Aspergillus group, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Takifugu rubripes*, and again in *B. floridae*.

5.3.5 Phylogenetic analysis of DISP, PTC, PTC-R, and NPC1 proteins

Figure 5.4 shows the phylogenetic analysis of SSD sequences extracted from the other four SSD-containing proteins (DIPS, PTC, PTC-R, and NPC1). All DISP proteins cluster together with 100% bootstap support. Within the DISP protein group, most of the metazoan DISP sequences form a well-supported cluster (93% bootstrap value). Two of the human, *T. rubripes, Xenopus tropicalis, N. vectensis*, and one of *Lottia gigantea* DISP proteins fall in this cluster whereas another DISP copies from these organisms fall outside in one of the two different metazoan clades. Among plants, only the prasinophyte green algae (*Ostreococcus* and *Micromonas*) were found in the DISP group. Interestingly, these ancestral types of green algae have all DISP, PTC, and PTC-R. More derived types of green algae (*Chlamydomonas* and *Volvox*) have only PTC-R. Furthermore, DISP, PTC, and PTC-R are also absent in higher plants except for two PTC sequences found in the moss *P. patens*. The higher land plants seem to have lost the DISP, PTC, and PTC-R genes after the divergence from green algal lineages. The NPC1 proteins are clearly divided into groups specific to higher plants, non-prasinophyte green algae, basal eukaryotes, metazoans, and fungi (96%, 98%, 94%, 62%, and 98% bootstrap values,

respectively). This shows that NPC1 was present in the last common ancestor of all these organisms and they have been maintained in all these organismal groups. NPC1, like PTC, PTC-R, and DISP, has high number (13) of transmembrane regions.

5.3.6 Evolution of fungal HMGCR proteins

As mentioned before, we found many duplicated copies of HMGCR proteins with and without SSD in fungi (see Table 5.2). To understand the evolutionary process of such duplications, we reconstructed a phylogeny using the entire HMGCR sequences, which includes the catalytic domain and the SSD, where available, from representative fungal and metazoan species (Figure 5.5). Trees were also constructed using only the catalytic domain (~440 amino acids) of the HMGCR. Probably due to the short length and high conservation of this region, the phylogeny did not result in a good resolution (data not shown). The phylogeny based on the entire HMGCR protein sequences shows two distinct clusters for fungal and metazoan HMGCRs supported by 98% bootstrap support. HMGCR sequences shown in red lack SSD while those in black have SSD (Figure 5.5). One cluster with high bootstrap support (91%) includes only fungal HMGCR proteins that have no SSD (cluster 'a'). The fungal species included in this cluster also have at least one other copy of HMGCR that has SSD. Even within this cluster of HMGCRs with no SSD there are various numbers of duplicated HMGCRs identified within the same species. In this figure we also see that the Zygomycetes *Rhyzopus oryzae* and *Mucor circilloides* have undergone duplications most likely before speciation of these two species. One of the *R. oryzae* copy then lost the SSD (*R. oryzae*-1 in Figure 5.5). Species-specific duplications are also seen in *S. cerevisiae* and *Laccaria bicolor*. While *S. cerevisiae* kept SSD in both its HMGCR, *L. bicolor* has lost SSD in one of its HMGCR. Another loss of SSD can be seen in the only copy of HMGCR in *Chaetomium globosum*. There are multiple duplication events as well as loss of SSD during the evolution of fungal species as shown in Figure 5.5. One sub-cluster of cluster 'a' has a long branch

length indicating the changes in evolutionary rates in these sequences after the duplication and loss of SSD. This could be a possiblility of long branch attraction and therefore not likely a representation of the true phylogeny reflecting the evolution of those SSD sequences. Nonetheless, what we see is that the HMGCR is prone to duplication and it also tends to be lost. Also from Figure 5.3, we see that HMGCR is prone to duplication not only in fungi but also in metazoans (*D. melanogaster, T. rubripes, X. tropicalis, B. floridae*).

5.3.7 Evolution of SSD and SSD-containing proteins

The distribution of the SSD proteins in various eukaryotic lineages is summarized in Figure 5.6. Based on a eukaryotic tree of life (Parfrey *et al*. [32]), we hypothesize the evolutionary history of SSD proteins as follows. The presence of SSD sequences in all the eukaryotic organisms we examined indicates that SSD existed before the eukaryotic divergence. A bacterial permease, member of the RND superfamily, could have been a bacterial-sterol transporter or functionally related to hopanoid transporter [15]. This protein may have evolved into the ancestral protein of current SSD-containing proteins gaining new functions and evolving to contain a domain for sterol-sensing. The transfer of this bacterial sequence to eukaryotes could have been either by vertical descent or lateral transfer at around the origin of eukaryotes. Among all the SSD-containing proteins, the DISP, PTC and PTC-R seems to be the ancestral types, while NPC1 seems to be closely related to the more recently formed SSD proteins, HMGCR and SCAP. From NPC1 proteins, the SSD sequence appeared to be transferred to HMGCR proteins, and also merged with WD-40 repeat sequences to form SCAP in the lineage before the metazoan/fungal divergence.

The HMGCR without the SSD was already present in a wide range of eukaryotes and bacteria. The SSD in HMCGR and SCAP have regulatory functions [4]. These regions add another level of control in the cell for sterol homeostasis, and this may have

played part in the evolution of the organisms in these lineages. The NPC1 protein, whose function is in transporting of sterols in the cell, was found in most of the lineages except choanozoa, prasinophyte green algae, haptophyta, and stramenophiles. DISP, PTC, and PRC-R, all three occur in metazoa and choanozoa. All three are absent from fungi, amoebozoa, heterobolosea, and plants. The PTC and DISP are known to function in body patterning. Thus their presence in metazoa and choanozoa are understandable. Only PTC-R was also found in non-prasinophyte green algae while both PTC-R and DISP was found in haptophyta and stramenophiles. Figure 5.6 also shows the dichotomy between NPC1 and PTC/PTC-R/DISP proteins except in metazoa where both groups of proteins are present. It is possible that PTC/PTC-R/DISP proteins are acting as sterol transporters wherever NPC1 is absent.

5.4 Conclusions

We examined the distribution of the SSD proteins in various organisms. Previous studies have indicated their remote relationship with the bacterial permeases [4, 13, 33]. Our evolutionary analyses confirmed the possible bacterial origin of eukaryotic SSD sequences. We showed that these ancestral proteins evolved into DISP, PTC, PTC-R, and NPC1 acquiring their specific functions and in some cases getting lost or replaced in various lineages. We also showed that HMGCR with SSD, and SCAP may have been formed as results of domain acquisition. In general, the fungi and animals that use the mevalonate pathway have SCAP and HMGCR with SSD. The green algae that use only the DXP pathway have neither the SCAP, nor the HMGCR. The plants that use both pathways, do not have SCAP but have HMGCR without SSD. Therefore it seems that SSD in HMGCR and the SCAP protein is related to having only the mevalonate pathway for sterol synthesis where they both provide regulatory functions.

5.5 References

1. Dufourc EJ: **Sterols and membrane dynamics**. *J Chem Biol* 2008, **1**(1-4):63-77.
2. Ikonen E: **Cellular cholesterol trafficking and compartmentalization**. *Nat Rev Mol Cell Biol* 2008, **9**(2):125-138.
3. Espenshade PJ, Hughes AL: **Regulation of sterol synthesis in eukaryotes**. *Annu Rev Genet* 2007, **41**:401-427.
4. Kuwabara PE, Labouesse M: **The sterol-sensing domain: multiple families, a unique role?** *Trends Genet* 2002, **18**(4):193-201.
5. Goldstein JL, DeBose-Boyd RA, Brown MS: **Protein sensors for membrane sterols**. *Cell* 2006, **124**(1):35-46.
6. Yabe D, Xia ZP, Adams CM, Rawson RB: **Three mutations in sterol-sensing domain of SCAP block interaction with insig and render SREBP cleavage insensitive to sterols**. *Proc Natl Acad Sci U S A* 2002, **99**(26):16672-16677.
7. Millard EE, Gale SE, Dudley N, Zhang J, Schaffer JE, Ory DS: **The sterol-sensing domain of the Niemann-Pick C1 (NPC1) protein regulates trafficking of low density lipoprotein cholesterol**. *J Biol Chem* 2005, **280**(31):28581-28590.
8. Callejo A, Culi J, Guerrero I: **Patched, the receptor of Hedgehog, is a lipoprotein receptor**. *Proc Natl Acad Sci U S A* 2008, **105**(3):912-917.
9. Burke R, Nellen D, Bellotto M, Hafen E, Senti KA, Dickson BJ, Basler K: **Dispatched, a novel sterol-sensing domain protein dedicated to the release of cholesterol-modified hedgehog from signaling cells**. *Cell* 1999, **99**(7):803-815.

10. Ohgami N, Ko DC, Thomas M, Scott MP, Chang CC, Chang TY: **Binding between the Niemann-Pick C1 protein and a photoactivatable cholesterol analog requires a functional sterol-sensing domain**. *Proc Natl Acad Sci U S A* 2004, **101**(34):12473-12478.

11. Radhakrishnan A, Sun LP, Kwon HJ, Brown MS, Goldstein JL: **Direct binding of cholesterol to the purified membrane region of SCAP: mechanism for a sterol-sensing domain**. *Mol Cell* 2004, **15**(2):259-268.

12. Ioannou YA: **Multidrug permeases and subcellular cholesterol transport**. *Nat Rev Mol Cell Biol* 2001, **2**(9):657-668.

13. Tseng TT, Gratwick KS, Kollman J, Park D, Nies DH, Goffeau A, Saier MH, Jr.: **The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins**. *J Mol Microbiol Biotechnol* 1999, **1**(1):107-125.

14. Davies JP, Ioannou YA: **Topological analysis of Niemann-Pick C1 protein reveals that the membrane orientation of the putative sterol-sensing domain is identical to those of 3-hydroxy-3-methylglutaryl-CoA reductase and sterol regulatory element binding protein cleavage-activating protein**. *J Biol Chem* 2000, **275**(32):24367-24374.

15. Hausmann G, von Mering C, Basler K: **The hedgehog signaling pathway: where did it come from?** *PLoS Biol* 2009, **7**(6):e1000146.

16. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ: **The 20 years of PROSITE**. *Nucleic Acids Res* 2008, **36**(Database issue):D245-249.

17. [http://www.uniprot.org/]

18. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30**(14):3059-3066.

19. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach**. *J Mol Evol* 1981, **17**(6):368-376.

20. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.

21. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18**(5):691-699.

22. [http://hmmer.org/]

23. Henikoff S, Henikoff JG: **Position-based sequence weights**. *J Mol Biol* 1994, **243**(4):574-578.

24. [http://www.ncbi.nlm.nih.gov/]

25. Felsenstein J: **Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5**:164-166.

26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

27. Robinson-Rechavi M, Maina CV, Gissendanner CR, Laudet V, Sluder A: **Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes**. *J Mol Evol* 2005, **60**(5):577-586.

28. Nordstrom KJ, Fredriksson R, Schioth HB: **The amphioxus (Branchiostoma floridae) genome contains a highly diversified set of G protein-coupled receptors**. *BMC Evol Biol* 2008, **8**:9.

29. Lange BM, Rujan T, Martin W, Croteau R: **Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes**. *Proc Natl Acad Sci U S A* 2000, **97**(24):13172-13177.

30. Eisenreich W, Rohdich F, Bacher A: **Deoxyxylulose phosphate pathway to terpenoids**. *Trends Plant Sci* 2001, **6**(2):78-84.

31. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV *et al*: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas**. *Science* 2009, **324**(5924):268-272.

32. Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ, Katz LA: **Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life**. *Syst Biol* 2010, **59**(5):518-533.

33. Davies JP, Chen FW, Ioannou YA: **Transmembrane molecular pump activity of Niemann-Pick C1 protein**. *Science* 2000, **290**(5500):2295-2298.

Table 5.1. Distribution of SSD proteins in metazoa.

Phylum or subphylum	Species	Total SSD	SSD proteins[a]					
			HMGCR	NPC1	SCAP	PTC	DISP	PTC-R
Placozoa	*Trichoplax adharaens*	5	1	1	1	-	1	1
Cnidaria	*Nematostella vectensis* (sea anemone)	16	1*	1	1	1	6	6
	Hydra magnipapillata	7	1	2	-	-	2	2
Mollusca	*Lottia gigantean* (sea snail)	12	1	1	1	1	2	6
Annelida	*Hellobdella robusta* (leech)	6	1	1	1	1	1	1
	Capitella teleta (segmented worm)	16	1	1	1	1	8	4
Nematoda	*Caenorhabditis elegans* (roundworm)	39	(1)	2	1	5	2	29
Arthropoda	*Drosophila melanogaster*	8	2	2	1	1	1	1
	Daphnia pulex	8	1	2	1	1	1	2
Chordata								
Cephalochordata	*Branchiostoma floridae* (lancelet)	51	2	1	2	2	4	40
Urochordata	*Ciona intestinalis* (sea squirt)	6	1	2	1	1	1	-
Vertebrata	*Takifugu rubripes* (pufferfish)	13	2	2	1	2	3	3
Amphibia	*Xenopus tropicalis* (Western clawed frog)	12	1	2	1	2	3	3
Mammalia	*Homo sapiens*	12	1	2	1	2	3	3

[a] See Figure 5.1 for protein name abbreviations. '-' indicates absence of similar protein sequence. Numbers in parentheses indicate that HMGCR sequence had no SSD region.

* This sequence was found to have SSD and clustered along with other HMGCR sequences in the phylogeny. However, no catalytic domain was found in this sequence.

Table 5.2. Distribution of SSD proteins in fungi.

Taxonomical group	Species	Total SSD	SSD proteins[a]		
			HMGCR	NPC1	SCAP
[Zygomycota]					
Zygomycetes/Mucorales	*Rhizopus oryzae*	3	1 (1)	1	1
	Phycomyces blakesleeanus	4	1	1	2
	Mucor circinelloides	6	3	1	2
[Chytridiomycota]					
Chitridiomycetes/Chytridiales	*Batrachochytrium dendrobatidis*	1	1	-	-
[Basidiomycota/Agaricomycotina]					
Tremellomycetes /Tremellales	*Cryptococcus neoformans*	3	1	1	1
Homobasidiomycetes/Agaricales	*Coprinus cinereus*	3	1	1	1
	Laccaria bicolor	3	1 (1)	1	1
	Pleurotus ostreatus	3	1	1	1
Homobasidiomycetes/Russulales	*Heterobasidion annosum*	3	1	1	1
Homobasidiomycetes/Boletales	*Serpula lacrymans*	3	1	1	1
[Basidiomycota/Ustilaginomycotina]					
Ustilaginomycetes/Ustilaginales	*Ustilago maydis*	3	1	1	1
[Basidiomycota/Pucciniomycotina]					
Microbotryomycetes/Sporidiobolales	*Sporobolomyces roseus*	3	1	1	1
[Ascomycota/Taphrinomycotina]					
Schizosaccharomycetes/Schizosaccharomycetales	*Schizosaccharomyces pombe*	2	1	-	1
[Ascomycota/Pezizomycotina]					
Eurotiomycetes/Onygenales	*Microsporum gypseum*	2	1	1	-
	Microsporum canis	2	1 (1)	1	-
	Trichophyton equinum	2	1	1	-
	Coccidioides immitis RS	2	1	1	-
	Coccidioides posadasii str. Silveira	2	1	1	-

	Histoplasma capsulatum H143	2	1	1	-
	Blastomyces dermatitidis ER-3	2	1	1	-
	Paracoccidioides brasiliensis Pb01	2	1	1	-
Eurotiomycetes/Eurotiales	*Aspergillus nidulans*	2	1 (1)	1	-
	Aspergillus fumigatus	3	2	1	-
	Neosartorya fischeri	3	2 (1)	1	-
	Aspergillus terreus	3	2 (2)	1	-
	Aspergillus oryzae	2	1 (4)	1	-
	Aspergillus carbonarius	4	3 (1)	1	-
	Aspergillus clavatus	2	1	1	-
	Aspergillus flavus	3	2 (3)	1	-
	Aspergillus niger	2	2 (2)	-	-
Dothideomycetes/Capnodiales	*Mycosphaerella graminicola*	3	1	1	1
	Mycosphaerella fijiensis	3	1	1	1
Dothideomycetes/Pleosporales	*Alternaria brassicicola*	3	1	1	1
	Stagonospora nodorum	3	1	1	1
	Cochliobolus heterostrophus	3	1	1	1
	Pyrenophora tritici-repentis	3	1	1	1
Leotiomycetes/Helotiales	*Botrytis cinerea*	3	1	1	1
	Sclerotinia sclerotiorum	3	1	1	1
Sordariomycetes/Sordariales	*Neurospora crassa*	3	1	1	1
	Chaetomium globosum	2	(1)	1	1
Sordariomycetes/Magnaporthales	*Magnaporthe oryzae*	3	1	1	1
Sordariomycetes/Hypocreales	*Nectria haematococca*	3	1	1	1
	Fusarium graminearum	3	1	1	1
	Fusarium oxysporum	3	1 (4)	1	1
	Fusarium verticillioides	3	1	1	1
	Trichoderma virens	3	1	1	1
[Ascomycota/Saccharomycotina]					

Saccharomycetes/Saccharomycetales	*Yarrowia lipolytica*	3	1	1	1
	Candida albicans SC5314	2	1	1	-
	Candida parapsilosis	2	1	1	-
	Candida lusitaniae	2	1	1	-
	Debaryomyces hansenii	2	1	1	-
	Ashbya gossypii	2	1	1	-
	Candida glabrata	2	1	1	-
	Saccharomyces cerevisiae	3	2	1	-

[a] See Figure 5.1 for protein name abbreviations. '-' indicates absence of similar protein sequence. Numbers in parentheses indicate that these HMGCR sequences had no SSD region.

Table 5.3. Distribution of SSD proteins in plants.

Plant Group	Species	Total SSD	SSD proteins[a]					
			HMGCR	NPC1	SCAP	PTC	DISP	PTC-R
Green algae (prasinophytes)	*Micromonas pusilla*	4	-	-	-	1	3	-
	Micromonas RCC299	8	-	-	-	-	6	2
	Ostreococcus lucimarinus	5	-	-	-	1	3	1
	Ostreococcus tauri	6	-	-	-	1	4	1
	Ostreococcus RCC809	6	-	-	-	1	4	1
Green algae	*Chlorella sp.* NC64A	1	-	1	-	-	-	-
	Coccomyxa sp. C-169	1	-	1	-	-	-	-
	Chlamydomonas reinhardtii	3	-	1	-	-	-	2
	Volvox carterii	2	-	1	-	-	-	1
Spikemoss	*Selaginella moellendorffii*	2	(2)	2	-	-	-	-
Moss	*Physcomitrella patens*	4	(3)	2	-	2	-	-
Grass	*Sorghum bicolor*	1	(3)	1	-	-	-	-
	Oryza sativa	1	(3)	1	-	-	-	-
	Brachypodium distachyon	1	(3)	1	-	-	-	-
Flowering plants	*Arabidopsis thaliana*	2	(2)	2	-	-	-	-
	Arabidopsis lyrata	2	(2)	2	-	-	-	-
	Cucumis sativus	3	(3)	3	-	-	-	-
	Mimulus guttatus	1	(6)	1	-	-	-	-
	Ricinus communis (castor)	2	(3)	2	-	-	-	-
	Manihot esculenta (cassava)	3	(6)	3	-	-	-	-
Tree	*Populus trichocarpa*	3	(6)	3	-	-	-	-

[a] See Figure 5.1 for protein name abbreviations. '-' indicates absence of similar protein sequence. Numbers in parentheses indicate that these HMGCR sequence had no SSD region.

Table 5.4. Distribution of SSD proteins in basal eukaryotes.

Plant Group	Species	Total SSD	SSD proteins[a]					
			HMGCR	NPC1	SCAP	PTC	DISP	PTC-R
Haptophyta	*Emiliania huxleyi* CCMP1516	10	(3)	-	-	-	8	2
Stramenopiles (Heterokonta)								
microalga	*Aureococcus anophagefferens*	8	-	-	-	-	6	2
diatoms	*Fragilariopsis cylindrus*	3	(1)	-	-	-	2	1
	Phaeodactylum tricornutum	3	(1)	-	-	-	-	3
	Thalassiosira pseudonana	3	(1)	-	-	-	2	1
Heterobolosea (amoebaflagellate)	*Naegleria gruberi*	1	(2)	1	-	-	-	-
Amoebozoa	*Dictyostelium discoideum*	2	(2)	2	-	-	-	-
	Dictyostelium purpureum	2	(2)	2	-	-	-	-
Choanozoa (choanoflagellate)	*Monosiga brevicollis*	4	(1)	-	-	2	1	1

[a] See Figure 5.1 for protein name abbreviations. '-' indicates absence of similar protein sequence. Numbers in parentheses indicate that those HMGCR sequence had no SSD region.

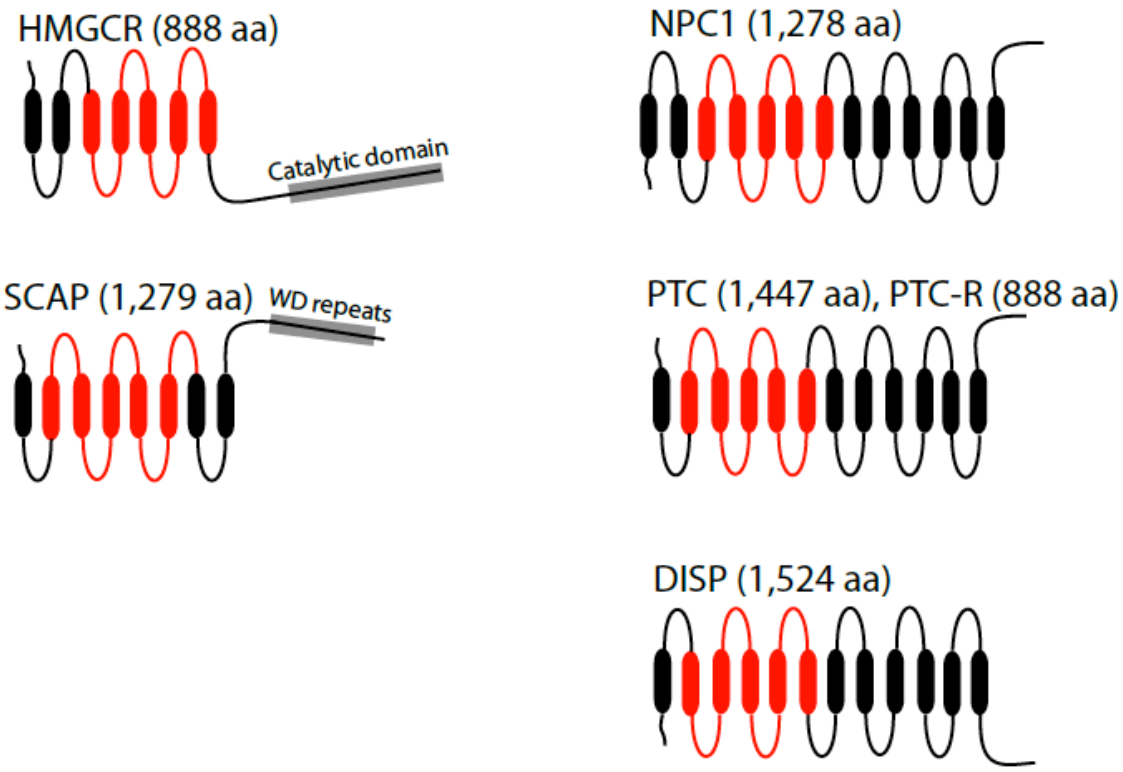

Figure 5.1. Topology of the SSD proteins. The lengths and topology of the proteins shown are based on the human SSD proteins. The cylindrical structures are the transmembrane regions. The SSD regions are indicated in red. The top side of each protein is cytoplasmic. Enzyme names are as follows. HMGCR: 3-hydroxy-3-methylglutaryl-coenzyme A reductase, SCAP: Sterol regulatory element binding protein clevage activating protein, NPC1: Niemann-Pick type C1 protein, PTC: Patched protein, PTC-R: Patched related protein, and DISP: Dispatched protein.

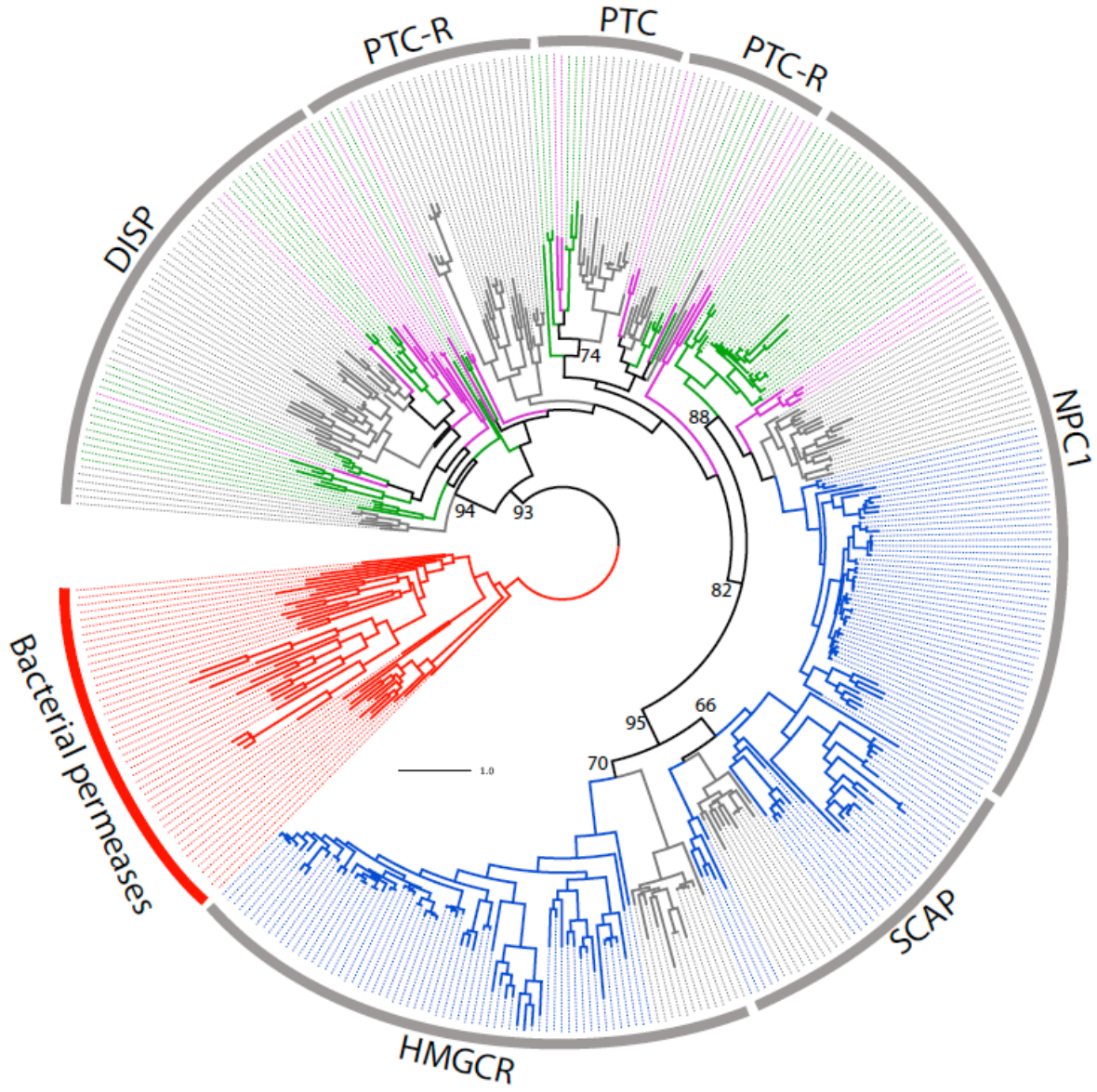

Figure 5.2. Maximum-likelihood phylogeny of the SSD protein family. Sequences marked in red are bacterial, blue is fungi, gray is metazoans, green is plants, and magenta is basal eukaryotes. Numbers at the nodes are the bootstrap support values (%). Only values higher than 65% are shown for major nodes. See Figure 5.1 for protein name abbreviations.

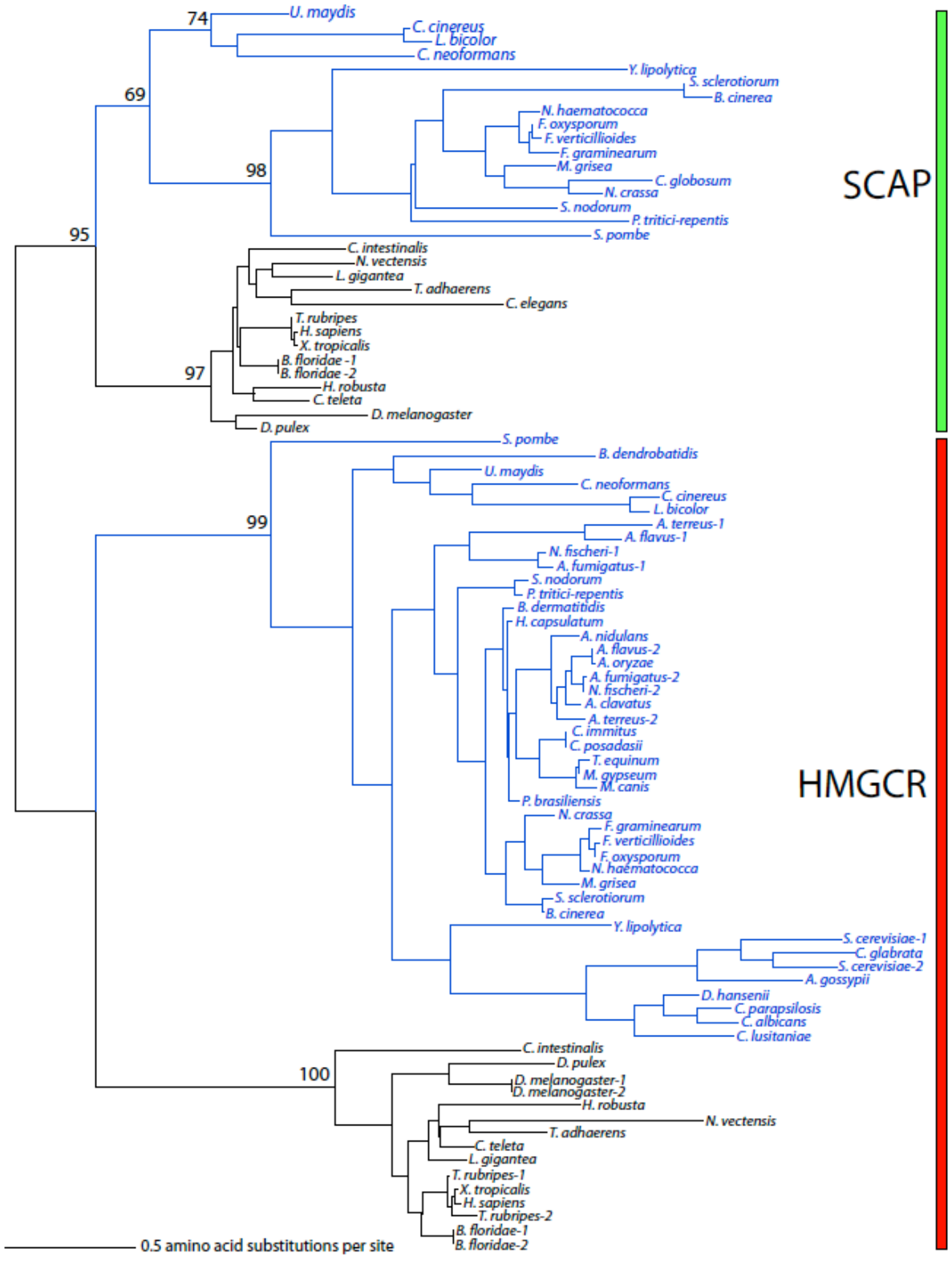

Figure 5.3. Maximum-likelihood phylogeny of the SSD regions of SCAP and HMGCR. Sequences marked in blue are fungi and those in black are metazoans. Numbers at the nodes are the bootstrap support values (%). Only values higher than 65% are shown for major nodes. See Figure 5.1 for protein name abbreviations.

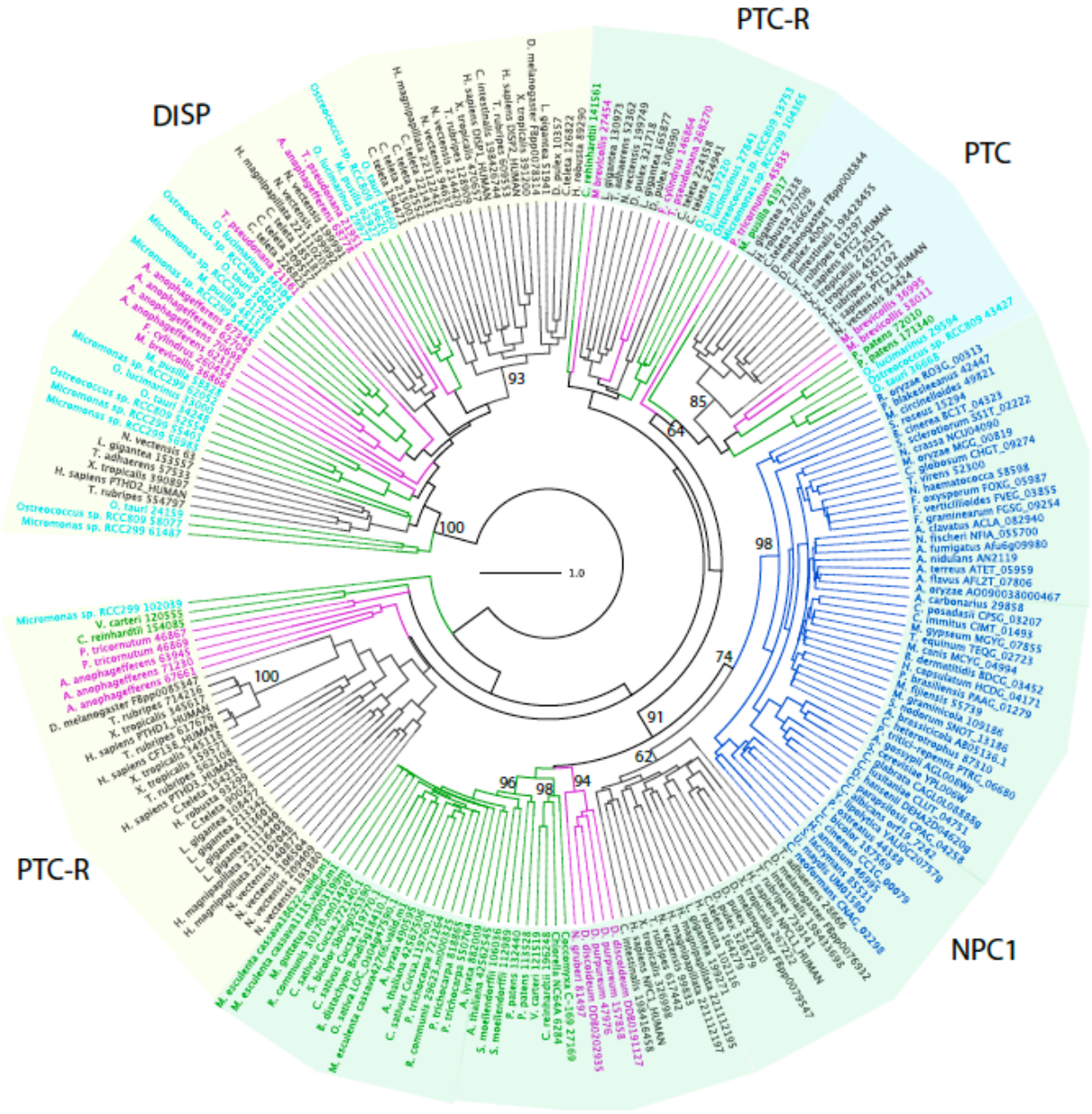

Figure 5.4. Maximum-likelihood phylogeny of the SSD regions of DISP, PTC, PTC-R, and NPC1. Sequences marked in blue are fungi, black are metazoans, cyan are prasinophyte green algae, green are plants and other non-prasinophyte green algae, and magenta are basal eukaryotes. Numbers at the nodes are the bootstrap support values (%). Only values higher than 60% are shown for major nodes. See Figure 5.1 for protein name abbreviations.

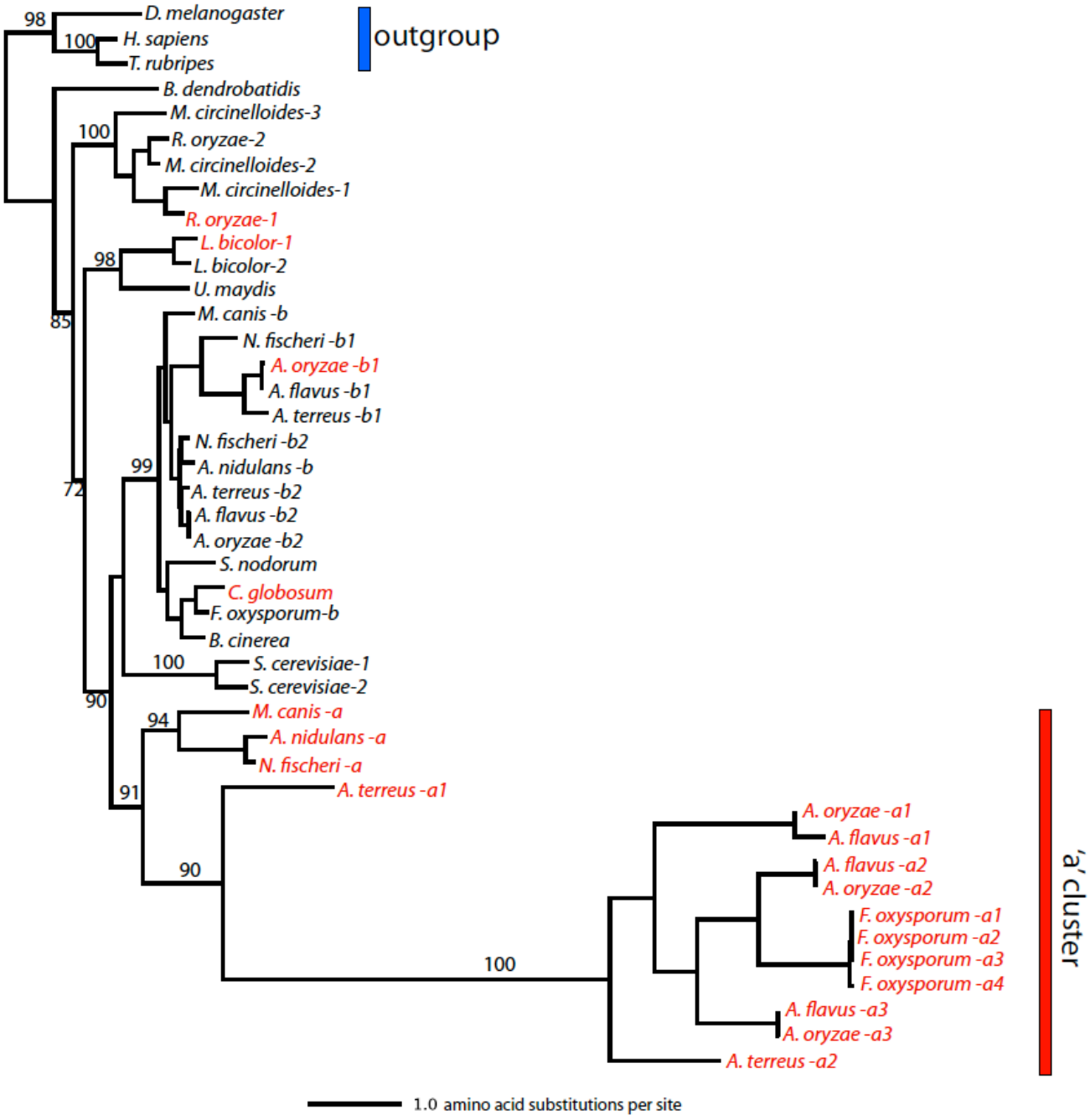

Figure 5.5. Maximum-likelihood phylogeny of the fungal HMGCR protein sequences. Sequences in red are HMGCRs that lack the SSD region while the sequences in black have the SSD region. Numbers at the nodes are the bootstrap support values (%). Only values higher than 70% are shown for major nodes.

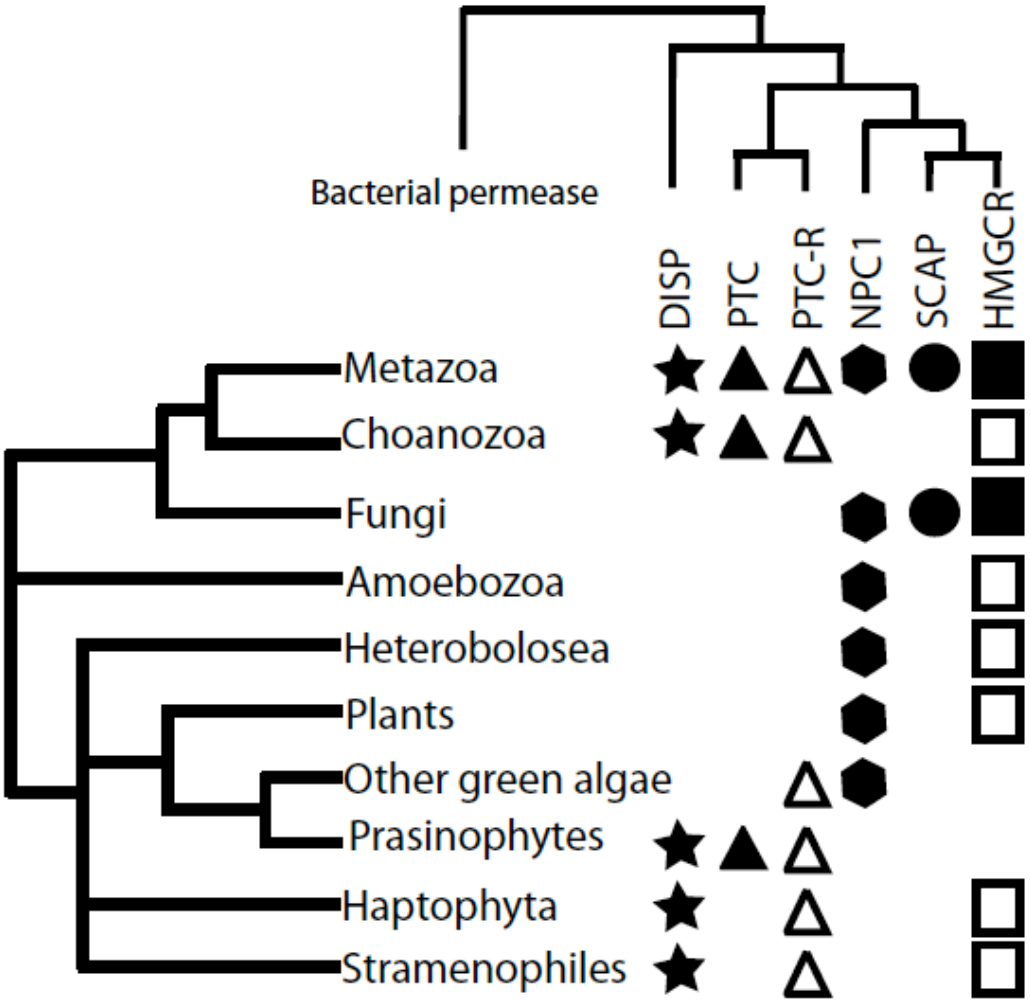

Figure 5.6. Distribution of the SSD-containing proteins among eukaryotes. The eukaryotic tree is based on Parfrey *et al*. [32]. The polytomies are due to the uncertainties of the placement of those groups in the tree of life. Green algae-1 include *Chlorella vulgaris, Chlamydomonas reihnardtii, Volvox carteri, and Coccomyxa sp*. Green algae-2 include the prasinophytes *Micromonas* and *Ostreococcus* species. The solid box represents the HMGCR with SSD, and the open box represents the HMGCR without SSD.

Table 5.S1. Bacterial species used in the study and the presence of SSD-like sequences.

Phylum or Class	Species[a]	ACC#	SSD-like
Alphaproteobacteria	*Caulobacter crescentus NA1000*	NC_011916	-
	Asticcacaulis excentricus CB 48	NZ_ACQR00000000	-
	Sinorhizobium medicae WSM419	NC_009636	-
Betaproteobacteria	*Achromobacter piechaudii ATCC 43553*	NZ_ADMS00000000	-
	Bordetella pertussis Tohama I	NC_002929	-
	Nitrosomonas europaea ATCC 19718	NC_004757	1
	Neisseria meningitidis FAM18	NC_008767	-
	Burkholderia sp. CCGE1001	NZ_ADDJ00000000	1
Gammaproteobacteria	*Escherichia coli O111:H- str. 11128*	NC_013364	-
	Yersinia pestis Angola	NC_010159	-
	Haemophilus influenzae 86-028NP	NC_007146	-
	Pantoea ananatis LMG 20103	NC_013956	1
	Pantoea sp. At-9b	NZ_ACYJ00000000	-
	Shewanella oneidensis MR-1	NC_004347	1
	Pseudomonas aeruginosa LESB58	NC_011770	1
	Coxiella burnetii Dugway 5J108-111	NC_009727	-
	Pectobacterium carotovorum subsp. carotovorum PC1	NC_012917	-
	Xanthomonas campestris pv. campestris str. B100	NC_010688	-
	Cellvibrio japonicus Ueda107	NC_010995	2
	Teredinibacter turnerae T7901	NC_012997	3
	Marinomonas sp. MED121	NZ_AANE00000000	3
	Klebsiella pneumoniae 342	NC_011283	-
	Pseudomonas fluorescens SBW25	NC_012660	-
Deltaproteobacteria	*Geobacter sp. M21*	NC_012918	1
	Sorangium cellulosum 'So ce 56'	NC_010162	2
Epsilonproteobacteria	*Helicobacter pylori B38*	NC_012973	-
	Wolinella succinogenes DSM 1740	NC_005090	1
Acidobacteria	*Acidobacterium capsulatum ATCC 51196*	NC_012483	-
	Solibacter usitatus Ellin6076	NC_008536	1
Cyanobacteria	*Synechococcus sp. PCC 7002*	NC_010475	-
	Gloeobacter violaceus PCC 7421	NC_005125	-
	Cyanothece sp. PCC 7425	NC_011884	-
Deinococcus-Thermus	*Thermus thermophilus HB8*	NC_006461	-
	Deinococcus deserti VCD115	NC_012526	2
Chloroflexi	*Dehalococcoides ethenogenes 195*	NC_002936	1
Aquificae	*Aquifex aeolicus VF5*	NC_000918	-
Thermotogae	*Thermotoga maritima MSB8*	NC_000853	1
Fusobacteria	*Fusobacterium nucleatum subsp. nucleatum ATCC 25586*	NC_003454	-
Verrucomicrobia	*Verrucomicrobium spinosum DSM 4136*	NZ_ABIZ00000000	-
Chlamydiae	*Chlamydophila pneumoniae CWL029*	NC_000922	-
	Chlamydia trachomatis B/TZ1A828/OT	NC_012687	-
Bacterioidetes	*Porphyromonas gingivalis W83*	NC_002950	1
Chlorobi	*Chlorobium limicola DSM 245*	NC_010803	-
Fibrobacteres	*Fibrobacter succinogenes subsp. succinogenes S85*	NC_013410	3
Actinobacteria	*Mycobacterium tuberculosis F11*	NC_009565	5
	Corynebacterium aurimucosum ATCC 700975	NC_012590	-

	Streptomyces avermitilis MA-4680	NC_003155	4
	Bifidobacterium longum subsp. infantis ATCC 15697	NC_011593	-
Spirochaetes	*Borrelia burgdorferi ZS7*	NC_011728	1
	Treponema denticola ATCC 35405	NC_002967	3
Planctomycetes	*Rhodopirellula baltica SH 1*	NC_005027	3
Firmicutes	*Clostridium botulinum A2 str. Kyoto*	NC_012563	1
	Mycoplasma hyopneumoniae 7448	NC_007332	-
	Streptococcus pneumoniae 70585	NC_012468	-
	Bacillus anthracis str. CDC 684	NC_012581	2
	Roseburia intestinalis L1-82	NZ_ABYJ00000000	1

Chapter 6

Conclusions

In this dissertation, alignment-based and alignment-free protein classification methods were compared for their accuracy in classifying highly divergent transmembrane proteins, and study of molecular evolution of multi-domain proteins were performed.

In Chapter 2, I examined various protein classification methods and carried out comparative performance analyses of these methods in classifying a group of transmembrane protein families, G-protein coupled receptors (GPCRs). The methods included profile hidden Markov model (HMM), GPCRHMM, decision trees, and support vector machines using the following input vectors: Fisher scores (derived from profile HMMs), amino acid compositions, and pairwise alignment scores. We tested the classifiers' performance to identify GPCRs when the classifiers were trained using highly similar or only remotely similar GPCRs, to identify short subsequences of the GPCRs, and also to identify GPCRs from the actual Drosophila EST sequences (including mostly short partial sequences). Our results showed that using simple amino acid compositions with support vector machines was effective in classifying GPCRs from non-GPCRs even when only small fragments of protein sequence was available. The computationally expensive method using sequence pairwise alignment scores with support vector machines (SVM_pairwise) is the most balanced classifier that is sensitive to remote similarity and can be also highly discriminative for classifying GPCR classes. However, use of SVM_pairwise for a large-scale analysis may not be practical for its computational cost. To identify member proteins from well-established protein families where a good number of representative samples are available, profile HMMs as well as GPCRHMM give highly accurate classifications. We suggested that a combination of protein sequence

classifiers be used in order to achieve a thorough mining of divergent protein family members.

Chapter 3 described an application of some of these methods in predicting putative GPCRs (also known as seven transmembrane receptors) from the *Arabidopsis thaliana* genome. The following six classifiers were utilized: linear and quadratic discriminant analysis, K-nearest neighbor, two different support vector machines with amino acid composition and dipeptide composition, and partial least squares with amino acid properties. Candidate proteins included in the intersection of the positives identified by these classifiers were then filtered according to the number of predicted transmembrane regions resulting in 54 proteins expanding the number of GPCR candidates in *Arabidopsis* from current 22 proteins. We showed that the strategy of combining different classifiers effectively provides prioritized lists of GPCR candidates for further experimental analyses to analyze their functions.

In the second part of the dissertation I examined the distribution of multi-domain proteins such as urea amidolyase, urea carboxylase, and sterol-sensing domain proteins in various species across kingdoms to elucidate their evolutionary history. In Chapter 4, I examined the distribution of urea amidolyase and urea carboxylase in eukaryotes as well as in prokaryotes. Phylogenetic analyses using amidase and urea carboxylase domain sequences revealed an interesting evolutionary pathway that eventually formed these proteins through gene fusion and also bacteria-to-fungus horizontal gene transfer. Urea amidolyase probably entered the fungal kingdom by horizontal gene transfer from a proteobacterial species either as a single gene, or as two separate genes (urea carboxylase and amidase) that later fused in the fungal lineage. Urea carboxylase could have either

evolved into fungi, green algae, and hydra through vertical descent followed by a large number of losses in various eukaryotic lineages, or is another case of bacterial gene transfer to these organisms. Acquisition of genes in this way may have important implications on the fungal evolution adaptability to newer environments.

The analyses of sterol-sensing domain (SSD) containing proteins in Chapter 5 showed that this domain is present in all the eukaryotes and has remote similarity with bacterial permeases. Our phylogenetic analyses showed that it is likely that the bacterial permease evolved into four different types of SSD-containing proteins, namely Dispatched, Patched, Patched-related and Niemann-Pick type C-1 acquiring specific functions in eukaryotes. I showed that some of these proteins have been lost on various lineages. The dispatched, patched, and patched-related proteins are completely absent from fungi but are present in some green algae and basal eukaryotes. Two types of SSD proteins, hydroxymethylglutaryl-CoA reductase and sterol regulatory element binding protein cleavage activating protein, seem to have formed by domain acquisition just before the divergence of fungi and metazoans.

The methods used in these studies can be applied to many other protein families to study their distribution and evolutionary history. For future works, such evolutionary analyses can be carried out for the nuclear receptor proteins. I have preformed a preliminary data mining of these proteins in eukaryotes and the results are presented in the Appendix. Both methods applied (profile HMMs and support vector machines using pairwise similarity scores) performed well in identifying the nuclear receptors from Drosophila species, but profile HMMs were able to give more remotely similar sequences to nuclear receptors in plants and fungi. However, more analysis is required to see if

these are distantly related to nuclear receptors or false positives. These proteins are also multi-domain and it would be interesting to see how the different domains have arrived to form this group of proteins. Additionally, differences in certain domains in nuclear receptors can be studied to understand the variation in function of these types of proteins. Another future work is the evolutionary study of biotin ligase proteins in fungi. These proteins modify the carboxylase proteins such as urea amidolyase and urea carboxylase by attaching biotin to it. The distribution and evolutionary information of these ligases can possibly tie with the results of urea amidolyase and urea carboxyase proteins that I have already worked with. It will be interesting to see if fungi has more than one biotin protein ligase, and if the distribution pattern of these biotin ligases match with the distribution pattern we have seen with urea amidolyse and urea carboxylase.

Appendix

Identification of candidate nuclear receptor proteins in the eukaryotic species using multi-domain information

Introduction

Nuclear Receptors (NRs), a multi-domain protein family of ligand activated transcription factors, play a key role in the process of development, metabolism and reproduction of the cell. In their inactive state, NRs reside in either the nucleus or the cytoplasm. Activation occurs when a ligand binds at the ligand-binding-domain (LBD) of the NR. This in turn causes the NR to bind to response elements (promoters) of their target genes via DNA binding domain (DBD). Some NRs like the thyroid receptors are always bound to the DNA and are activated by ligand binding. The effect of this reaction is the regulation of the expression of the target genes.

NRs share a common organizational structure as shown in Figure A1: the N-terminal region (A/B domain) that is highly variable and consists of a transactivation region AF-1; the DBD (C domain) that is highly conserved and is also involved in the dimerization of NRs; the less conserved flexible hinge (D domain); the moderately conserved LBD (E domain); the extremely variable, and sometimes absent, F domain [1].

Depending upon the DBD and LBD, NRs are divided into six subfamilies as follows:

1. Thyroid hormone
2. Hepatocyte nuclear factor 4-gamma
3. Estrogen
4. Nerve growth factor 1B
5. Fushi tarazu-F1
6. Germ cell nuclear factor

In addition to these, there are two more subfamilies: 1) Knirps (NRs with no LBD) and 2) DAX (NRs with no DBD). Many of the annotated NRs do not have a known ligand and hence are called orphan nuclear receptors. It is likely that the ancestor protein of NRs was an orphan receptor and ligand binding was an acquired property of these proteins [2].

Natural activation of NRs typically occurs by the binding of lipophilic molecules (ligands), such as steroid hormones, bile acids, fatty acids, thyroid hormones, certain vitamins and prostaglandins [2]. Many orphan NRs have also been found to be activated by synthetic ligands. NRs are also responsible in diseases such as cancer, diabetes, and asthma [3]. Their potential to be regulated by exogenous compounds makes them an extremely important drug target in human disease [4].

NRs have been found in diverse metazoans but have been absent in plants and fungi [2]. Most likely, NRs in these kingdoms either are so diverged that current methods fail to find them, or these organisms may have a different kind of protein that do the same function. This hypothesis lead us to explore these genomes in search of proteins that are either NRs or a novel family of proteins that has some similarities with the LBD and DBD of known NRs. One such possibility can be found in *Candida albicans*. A quorum-sensing molecule, farnesol, has been found to be produced by this organism [5], although no farnesol-binding protein has been found. This is interesting since farnesol and its metabolites are generated in the cell and is required during the synthesis of cholesterol, bile acids, steroids, retinoids, and farnesylated proteins [6], and are ligands for some mammalian NRs. Based on this, we expect to find farnesol-binding proteins in the fungal genomes, specifically in *Candida albicans*, as putative NRs. Recently, proteins with LBD

that is similar to animal NRs was discovered in yeast, and their function also resembles the function of PPAR-alpha/RXR which belongs to the NR superfamily [7]. This shows that there is a possibility that NRs are present in yeast and other fungal genomes.

At the time of this research, it was found that humans had 48 nuclear receptors [4], sea squirt had 17 [8], pufferfish had 70 [9], Drosophila had 21 [10], and *C. elegans* had more than 250 [4]. We wanted to see if by using a protein classification method that is different from the commonly used sequence similarity method Blastp, we can trace some NR-like protein sequences in fungi and plants.

Materials and Methods

Training data

Training sequences were gathered from the swissport database. 370 NR sequences were downloaded. Not all of these sequences were labeled to have both DBD and LBD. The numbers are given in Table 1. Negative data, 250 protein sequences that are not NRs, were also gathered from swissprot.

Classification methods

Two different methods were chosen for protein classification: profile hidden Markov models (HMMs) and support vector machines (SVMs). These methods have been shown to identify related proteins very accurately. For both of these methods, we chose to only use the LBD and the DBD regions because they are the most invariable regions among these proteins, and functionally very important.

For the profile HMM method, a multiple alignment is first required. Two multiple alignments were first created, each from all the 345 DBD sequences, and from all the 277 LBD sequences (see Table A.1). Similar multiple alignments (for LBD and DBD) were created using sequences from the 7 NR subfamilies. Since one subfamily had no LBD sequences annotated, we ended up with a total of 13 subfamily-level multiple alignments. Clustalw (version 2.0) was used to build these multiple alignments. These alignments were then used to build a profile HMM using HMMER (version 2.3.2). The profile HMM was calibrated using "hmmcalibrate" command, which takes the HMM and empirically determines parameters that are used to make searches more sensitive, by calculating more accurate e-values. A database size of 50,000 was used so that all the e-values using different databases could be comparable to each other. The subfamily-level profile HMM was used separately to classify sequences, and the results were all combined together. The combination of all the subfamily-level profile HMMs are called multi-HMM in the Result section.

Two models, each for LBD and DBD was created using SVM. The pairwise alignment e-value between two sequences were used as input to the SVM. This method, SVM-pairwise (SVM-pw), was explained in Chapter 2. SVM requires negative data for training. Since we are only using the LBD or DBD regions, random subsequences of similar lengths were gathered from the negative datasets. We used SSearch to make pairwise alignments of all the training sequences (both positive and negative sequences). Their e-values were used in creating a matrix of input data for the SVM. The radial basis kernel function was used as the kernel function. We did not make subfamily-level SVM models because from our previous work in Chapter 2, we found that SVM-pairwise was

able to classify well sequences from sister subfamilies even when it wasn't trained on them.

Genomic data

We searched the genomes including: 10 fungal species, 6 *Drosophila* species, and 7 plants species (including one green alga). The list of these species, the number of proteins and their sources are listed in Tables A.2, A.3 and A.4.

Results

Among the *Drosophila* genomes we examined, the SVM-pw classifiers and the family and subfamily-level profile HMMs (multi-HMM), found hits for all the six species (Table A.7). The numbers for NRs identified from the *D. melanogaster* genome included those that have been already known (21). This shows that our methods are working to find NRs. As we see, not all NRs in *Drosophila* have both the DBD and LBD in one protein. Among the 21 *Drosophila* NRs, only 17 had LBDs based on the profile HMM and SVM-pairwise results. Whether the rest of them have LBD or not is a question that can be answered with further study. It could be that they either lack an LBD, or that these sites are very different from what we know, so that our models could not trace them. From the combined results of profile HMM and SVM-pairwise methods, we found that three Drosophila species (*D. melanogaster, D. ananassae, and D. willistoni*) had 21 NRs with DBD region while the other three species (*D. pseudoobscura, D. virilis and D. grimshawi*) had 22 NRs with DBD region. Again, three of the *Drosophila* species (*D. melanogaster, D. willistoni, and D. virilis*) had only 17 NR sequences with LBD while

the rest of the three species (*D. ananassae, D. pseudoobscura and D. grimshawi*) had 18 NRs with LBD. Most of the time, the profile HMM methods classified more sequences as LBD and DBD, than the SVM-pairwise method. The sequences that were classified as having DBD or LBD by any one of the methods but not all, are the ones that need to be looked at again to see if they are remote homologs of NRs in these species.

For fungal genomes, the SVM-pw classifiers found no hits for either the DBD or the LBD. The LBD profile HMM (family-level) did not find any hits from the fungal genomes using the e-value threshold of 1, but at threshold of 10, five fungal species (*Rhyzopus oryzae, Schizosaccharomyces pombe, Aspergillus nidulans, Neurospora crassa* and *Ashbya gossypii*) had hits (Table A.5). Similarly, there were three fungal species with hits for the DBD profile HMM (family-level) at the e-value threshold of 1, while seven fungal species (*Rhyzopus oryzae, Ustilago maydis, Schizosaccharomyces pombe, Aspergillus nidulans, Magnaporthe grisea, Neurospora crassa* and *Saccharomyces cerevisiae*) had hits at the e-value threshold of 10. Using multi-class profile HMMs, one fungal species (*Rhyzopus oryzae*) had both DBD and LBD hits at the e-value threshold of 1, while another species (*Schizosaccharomyces pombe*) only had the DBD hits. Raising the e-value threshold to 10 gave 8 fungal species to have both LBD and DBD hits, and two species to have only DBD hits. Even though very strong hits were not found in the fungal genomes for both DBD and LBD, these hits could be remotely similar to the metazoan NRs and further analysis could determine the confidence in these sequences.

For plant genomes, as shown in Table A.6, the SVM-pw classifiers showed two DBD hits only for rice. No other hit was found neither for DBD nor LBD from any other plant genomes. Using the profile HMMs, at the e-value threshold of 1, there was one hit

each for DBD in the *Chlamydomonas* and poplar genomes. Raising the e-value threshold to 10, we found DBD and LBD hits from all except the maize genome, which only had hits of DBD. Using the multi-class HMMs, at the e-value threshold of 1, the *P. patens* had one hit each for LBD and DBD, while *A. thaliana* and poplar had four and one hit(s) for DBD only, respectively. Raising the e-value threshold to 10 resulted in hits for DBD and LBD for all plant species. The number of DBD hits was 171 for the maize genome, which seems too high compared to what we have seen on other plants. This could be the result of errors in sequencing and annotations of this genome.

In this analysis, we found that profile HMMs give more probable remote homologs than SVM-pairwise although these could be false positives. SVM-pairwise seems to be more specific.

Even though an NR-like activity was traced in fungi by a previous study, we found no such hits with high confidence. These organisms may have a very different type of NR-like proteins that perform similar functions. More experimental work would be needed to characterize such a protein, which would then help bioinformatics study in finding similar proteins from other fungal species.

References

1. Escriva Garcia H, Laudet V, Robinson-Rechavi M: **Nuclear receptors are markers of animal genome evolution**. *J Struct Funct Genomics* 2003, **3**(1-4):177-184.
2. **Essays in Biochemistry**, vol. 40: Portland Press; 2004.
3. Novac N, Heinzel T: **Nuclear receptors: overview and classification**. *Curr Drug Targets Inflamm Allergy* 2004, **3**(4):335-346.
4. Maglich JM, Watson J, McMillen PJ, Goodwin B, Willson TM, Moore JT: **The nuclear receptor CAR is a regulator of thyroid hormone metabolism during caloric restriction**. *J Biol Chem* 2004, **279**(19):19832-19838.
5. Shchepin R, Hornby JM, Burger E, Niessen T, Dussault P, Nickerson KW: **Quorum sensing in Candida albicans: probing farnesol's mode of action with 40 natural and synthetic farnesol analogs**. *Chem Biol* 2003, **10**(8):743-750.
6. Forman BM, Goode E, Chen J, Oro AE, Bradley DJ, Perlmann T, Noonan DJ, Burka LT, McMorris T, Lamph WW *et al*: **Identification of a nuclear receptor that is activated by farnesol metabolites**. *Cell* 1995, **81**(5):687-693.
7. Phelps C, Gburcik V, Suslova E, Dudek P, Forafonov F, Bot N, MacLean M, Fagan RJ, Picard D: **Fungi and animals may share a common ancestor to nuclear receptors**. *Proc Natl Acad Sci U S A* 2006, **103**(18):7077-7081.
8. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM *et al*: **The draft genome of Ciona**

intestinalis: insights into chordate and vertebrate origins. *Science* 2002, **298**(5601):2157-2167.

9. Bertrand S, Brunet FG, Escriva H, Parmentier G, Laudet V, Robinson-Rechavi M: **Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems**. *Mol Biol Evol* 2004, **21**(10):1923-1937.

10. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.

Figure A.1: Organization of a typical nuclear receptor (taken from Escriva Garcia *et al.* 2003).

Table A.1. Number of nuclear receptor sequences and the subfamilies used in the study

NR subfamily	Total	With DBD	With LBD
Knirps and DAX	16	5	11
Thyroid hormone	143	140	119
Hepatocyte nuclear factor 4-gamma	67	67	32
Estrogen	107	96	88
Nerve growth factor 1B	16	16	13
Fushi tarazu-F1	14	14	14
Germ cell nuclear factor	7	7	0
Total	370	345	277

Table A.2. List of fungal species used in the study

Fungal species	Phylum	No. of Proteins	Sequencing group
Rhyzopus oryzae	Zygomycota	17,467	FGI
Ustilago maydis	Basidiomycota	6,522	FGI
Schizosaccharomyces pombe	Ascomycota	5,025	Sanger
Aspergillus nidulans	Ascomycota	10,665	FGI
Magnaporthe grisea	Ascomycota	11,054	FGI
Neurospora crassa	Ascomycota	9,845	FGI
Fusarium graminearum	Ascomycota	13,321	Stanford BRI-NRC of Canada
Candida albicans	Ascomycota	5,993	yeastgenome.org
Saccharomyces cerevisiae	Ascomycota	6,717	NCBI
Ashbya gossypii	Ascomycota	4,714	FGI

Table A.3. List of Drosophila species used. Data from flybase.org

Drosophila species	No. of Proteins
D. melanogaster	21,243
D. ananassae	15,070
D. pseudoobscura	16,071
D. willistoni	15,513
D. virilis	14,491
D. grimshawi	14,986

Table A.4. List of plant species used in the study

Plant species	Common name	No. of Proteins	Sequencing group
Chlamydomonas reinhardtii	Green algae	14,598	JGI
Physcomitrella patens ssp patens	Moss	35,938	JGI
Selaginella moellendorffii	Spikemoss	34,697	JGI
Oryza sativa	Rice	66,710	rice.plantbiology.msu.edu
Zea Mays	Maize	78,966	ftp.maize.sequence.org
Arabidopsis thaliana	Mouse-ear cress	32,615	NCBI
Populus trichocarpa	Poplar	45,555	JGI

Table A.5. LBD and DBD identified by HMM and SVM in fungi.

Fungal species	HMM (LBD/DBD)		Multi-HMM (LBD/DBD)		SVM-pw (LBD/DBD)
	(E <1)	(E <10)	(E <1)	(E <10)	
Rhyzopus oryzae	0/1	1/3	1/1	3/3	0/0
Ustilago maydis	0/0	0/1	0/0	1/2	0/0
Schizosaccharomyces pombe	0/1	3/2	0/1	3/2	0/0
Aspergillus nidulans	0/0	1/1	0/0	0/2	0/0
Magnaporthe grisea	0/1	0/1	0/0	1/6	0/0
Neurospora crassa	0/0	1/2	0/0	1/2	0/0
Fusarium graminearum	0/0	0/0	0/0	5/1	0/0
Candida albicans	0/0	0/0	0/0	3/1	0/0
Saccharomyces cerevisiae	0/0	0/2	0/0	0/2	0/0
Ashbya gossypii	0/0	1/0	0/0	3/3	0/0

Table A.6. LBD and DBD identified by HMM and SVM in plants

Species	HMM (LBD/DBD)		Multi-HMM (LBD/DBD)		SVM-pw (LBD/DBD)
	(E <1)	(E <10)	(E <1)	(E <10)	
Chlamydomonas reinhardtii	0/1	2/1	0/0	2/7	0/0
Physcomitrella patens ssp patens	0/0	1/2	1/1	13/11	0/0
Selaginella moellendorffii	0/0	4/5	0/0	4/12	0/0
Oryza sativa	0/0	2/8	0/0	8/22	0/2
Zea Mays	0/0	0/16	0/0	26/171	0/0
Arabidopsis thaliana	0/0	2/9	0/4	8/18	0/0
Populus trichocarpa	0/1	2/10	0/1	6/14	0/0

Table A.7. LBD and DBD identified by HMM and SVM in Drosophila

Drosophila Species	HMM (LBD/DBD)		Multi-HMM (LBD/DBD)		SVM-pw LBD/DBD	Common[a] (LBD/DBD)
	(E <1)	(E <10)	(E <1)	(E <10)		
D. melanogaster	17/22	18/31	17/22	25/29	17/21	17/21
D. ananassae	18/22	20/33	18/21	22/28	19/22	18/21
D. pseudoobscura	18/24	18/34	18/23	23/34	18/23	18/22
D. willistoni	17/22	18/25	17/21	27/31	17/21	17/21
D. virilis	17/22	18/28	17/22	23/32	17/22	17/22
D. grimshawi	18/23	22/28	18/23	23/33	18/22	18/22

[a] This number is the common sequences (LBD abd DBD) found by all classifiers.