COMPARATIVE GENOMICS OF EARLY ANIMAL EVOLUTION

A thesis submitted to the National University of Ireland for the Degree of Doctor of Philosophy

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Presented by:
Roberto Feuda
Department of Biology,
NUI Maynooth,
Maynooth,
Co. Kildare, Ireland.

November 2012

Supervisor: Dr. Davide Pisani B.Sc., Ph.D. (Bristol)

Head of Department: Professor Paul Moynagh, Dip. Biology BA (mod.), PhD (Dublin)

Table of Contents

Index of figure

Index of table

Acknowledgments

One of the joy of completion is to look back to the past journey and remember all the friends who have helped and supported me during this long but fulfilling road.

I would like to express my heartfelt gratitude to Dr. Davide Pisani, who is not only a mentor but also a dear friend. I could not have asked for a better model, at the same time inspirational, supportive, and patient. I could not be more proud of my academic roots and I hope that I can pass on the research values and the dreams given to me.

I would also like to thank my examiners Prof. Emma Teelling and Dr. David Fitzpatrick, for spending their time reading this thesis. It is not an easy task, reviewing a thesis, and I am grateful for their thoughtful and detailed comments.

Additionally I would like to thank Prof. James McInerney and Dr. Omar Rota-Stabelli for their teaching and for having always inspired me to improve myself.

All the members (Lahern, Therese, AJ, Karen, David A.P., Leanne, Sinead, Aoife, Carla, Marco, Rob, Brian) of the Bioinformatics lab in Maynooth, who made it a convivial place to work. In particular, I would like to thank Dr. Carla Cummins, Dr. Therese Holton and Mr. Washiu Akanni for helping me with the English.

My research would not have been possible without the founding of IRCSET, the computational resources provided by the Irish Center for High End Computing (ICHEC) and The High Performance Computing (HPC) Facility at NUI Maynooth.

And finally I would like to thank Alessandra, I could not have done it without your support, your humor and the strength you have given me.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed: _____
Roberto Feuda

Abstract

The explosion of genomics permits investigations into the origin and early evolution of the Metazoa at the molecular level. In this thesis, I am particularly interested in investigating the molecular foundation of the animal senses (i.e. how animals perceive their world).

To understand the directionality of evolutionary innovation a well-developed phylogenetic framework is necessary. On one hand, the combination of molecular and morphological data sets has revolutionized our views of metazoan relationships over the past decades, but on the other hand, a number of nodes on the metazoan tree remain uncertain. Uncertainty is particularly high with reference to the taxa generally named "early branching metazoans". Unfortunately, understanding the relationships among these taxa is key to understanding the evolution of sensory perception (Nielsen 2008). In this thesis I will investigate both animal phylogenetics (to attempt to resolve the phylogeny among the early branching Metazoa) and the evolution of the metazoan sensory receptors.

The G-protein coupled receptor superfamily (GPCR) superfamily is the main family of metazoan surface receptors. In this thesis, after an initial introduction (Chapter 1), I address and substantially clarify the relationship among the early branching animals (Chapter 2) using novel genomic data and publicly available expressed sequence tags (ESTs). I then move forward (Chapter 3) to use network-based methods to study the early evolution of the GPCR superfamily in Eukaryotes and animals. Finally (Chapter 4), I focus on the study of a specific subset of GPCRs (the a-group, Rhodopsin-like receptors). This GPCR group is particularly interesting as it includes the best studied and, arguably, one of the most interesting among the GPCR families: the Opsin family. Opsins are key proteins used in the process of light detection, and the origin and early evolution of this family are still substantially unknown. Chapter 4 addresses both these problems. The thesis is then concluded by a general discussion (Chapter 5) and a future directions (Chapter 6) section.

Overall, this thesis provides new insights into the origin and early evolution of the Metazoa and their senses.

Chapter 1

Introduction

1.1.1 The Animal kingdom: the Metazoa

The kingdom *Animalia* was introduced by Linnaeus in the first edition of the *Systema Naturae* (1735). Linnaeus defined animals as natural objects which grow, live and sense in contrast to plants, which grow and live but do not sense, and minerals, which grow, but neither live nor sense. This definition was retained almost unchanged in the 10th edition of *Systema Naturae*, which forms the baseline zoological nomenclature.

Ernst Haeckel was the first to propose a classification of living organisms consistent with Darwin's principles of 'descent with modification', a principle implicitly stating that a classification needs to be strictly genealogical. Haeckel, a great admirer of Darwin, was responsible for drawing the first "animal tree of life" and he gave a remarkably modern definition of the kingdom *Animalia*. Based on the presence of tissues and organs he divided the Animal kingdom from the Protista. This definition excluded the sponges from the animals, however, these organisms were successively included in a group he called Metazoa. In modern zoology Animalia and Metazoa are used as synonyms and the sponges are considered animals. In this thesis, I will be studying metazoan evolution as well as the evolution of sensory reception. In a sense, therefore, this thesis is about animals as intended by Linnaeus: animals defined in the most traditional way.

In this first, introductory chapter, I will delineate current understandings of animal relationships and pinpoint open questions. I will then move forward to provide a general introduction to the methods used in this thesis.

1.1.2 Metazoa as Eukaryotes

There are three generally recognized domains of life (Woese and Fox 1977): Eukaryota, Archaebacteria and Eubacteria. Eukaryotes are set apart from the other two by distinct features that are indicative of a more complex form and structure. In detail, eukaryotes are characterized by membrane-delimited compartments supported by a cytoskeleton (Parfrey *et al.* 2006). They possess cellular subunits (organelles) and a membrane bound nucleus. One of the eukaryotic organelles, the mitochondrion (or its derivatives; see Embley and Martin 2006; Hjort *et al.* 2010), is present in the majority of extant eukaryotes and was a feature of the last common eukaryotic ancestor. The origin of the eukaryotes is an important unresolved enigma (contrast (Embley and Martin 2006; Gribaldo *et al.* 2010), representing a major challenge for evolutionary biology (Koonin 2010), even though the monophyly and chimerical origin of the eukaryotes now seems unquestionable (Pisani *et al.* 2007; Cotton and McInerney 2010; Koonin 2010).

Eukaryotes are currently classified in five supergroups (Excavata, Plantae, Chromalveolata, Rhizaria and Unikonta-see Table 1.1), but relationships among these supergroups are still highly debated (Koonin 2010).

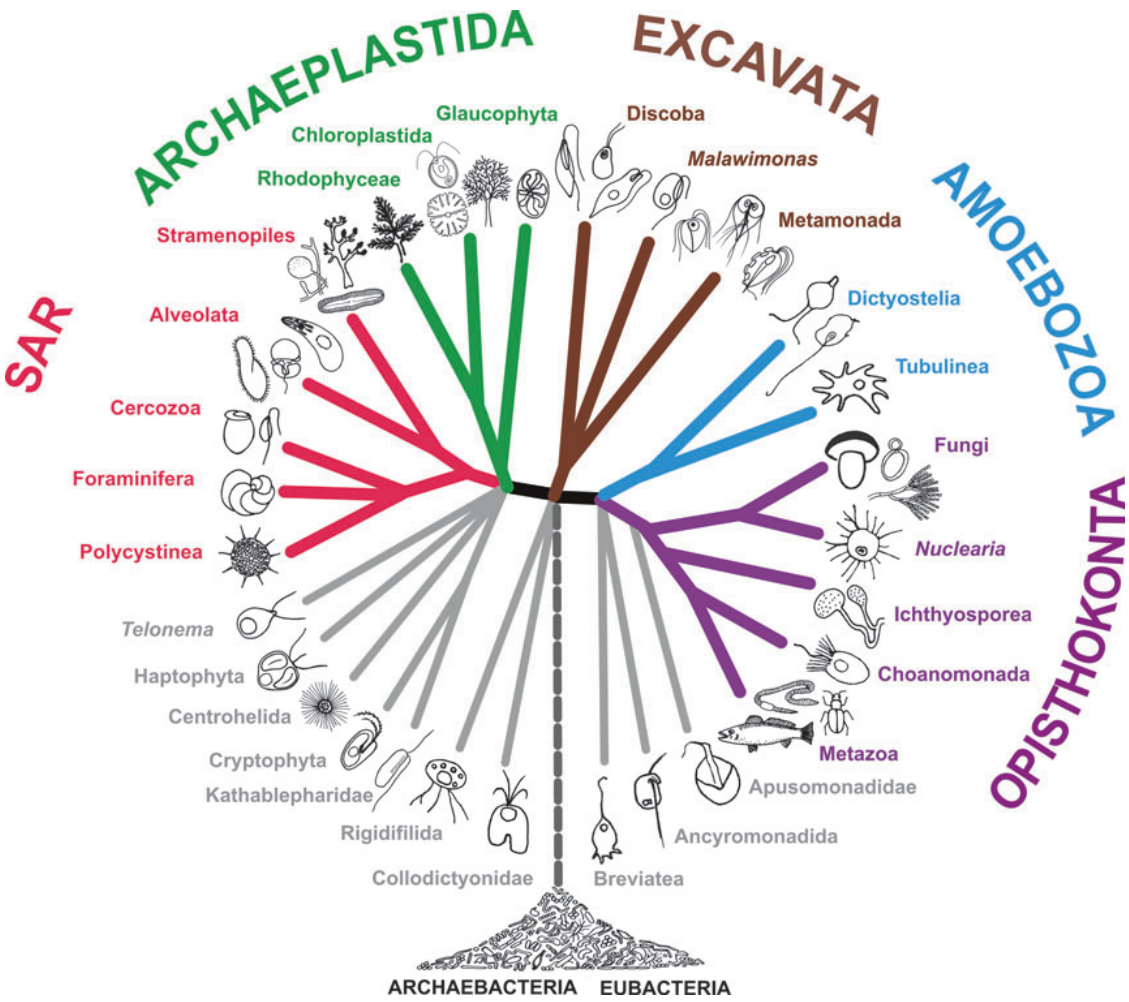


Figure 1.1: Evolutionary relationship of the eukaryotes (from Adl *et al.* 2012).

Despite current uncertainty in eukaryotic origins and early evolution (including the relationships of the supergroups-see figure 1.1), the metazoans and their closest outgroup the Choanoflagellata (see below) are known to belong to the Unikonta (Koonin 2010; Derelle and Lang 2012), a group characterised as having motile cells (like animal sperm) with a single cilium. Motile cells in all the other eukaryotic groups have two cilia and are sometimes called bikonts because of this (see Koonin 2010). The monophyly of the unikonts seems unquestionable, and is supported by rare genomic changes such as the fusion of two (dihydrogolate reductase and thymdylate synthase) enzyme genes (Richards and Cavalier-Smith 2005), the domain structure of myosins (Richards and Cavalier-Smith 2005) and from phylogenomic data (Derelle and Lang 2012). On the other hand, whether the bikonts are monophyletic or not is still a matter of debate, with many studies suggesting that the root of the eukaryotes should be found within the bikonts on the branch

separating the plants from all the other eukaryotes (Adl *et al.* 2012). If the Bikonta are not monophyletic then motile cells with two cilia represent a plesiomorphic condition within eukaryotes, with motile monociliated cells representing a unikont apomorphy.

		Super group	Example
Euckaryota	Unikonta	Amoebozoa	Dictyostelia
		Oposthokonta	Metazoan, Choanoflagellates, Fungi
	Bikonta	Chromalveolata	*Trypanosoma brucei*, *Thalassiosira pseudonana*
		Excavata	*Trichonoma vagianalis*, *Naegleria gruberi*
		Rhizaria	*Bigelowiella natans*
		Plants	*Vitis vinifera, Glycine max*

Table 1.1: Taxonomic definition for the Eukaryotes used in this thesis. The taxonomy follows Koonin (2010).

1.1.3 The Choanoflagellata: our unicellular cousins

Within Unikonta, the closest outgroup of the Metazoa is indubitably represented by the Choanoflagellata, and the Choanoflagellata-Metazoa group is generally referred to as the Holozoa. This sister group relationship is supported by both molecular and morphological data (King 2004; Philippe *et al.* 2005; King *et al.* 2008). Choanoflagellata is a small group currently containing only 200 species (Nielsen 2012). All choanoflagellates are free-living and they show both unicellular and colonial behaviour. As the name suggests, the choanoflagellates (collared flagellates) have a distinctive cell morphology characterized by an ovoid or spherical cell with a single apical flagellum surrounded by a collar of 30-40 microvilli. The flagellum is used to create a current that can propel free-swimming choanoflagellates through the water column, and trap bacteria and detritus against the collar of microvilli where they are then engulfed. The monophyly of the group seems unquestionable (Carr *et al.* 2008).

The origin of the Metazoa is associated to two fundamental evolutionary novelties: multicellularity and sexual reproduction. Multicellularity is considered probably the most import

apomorphy of Metazoa. Indeed, despite some choanoflagellate species exhibiting a colonial behaviour (coloniality) reminiscent of multicellularity, this similarity is only superficial (Nielsen 2012). Multicellularity is characterised by the division of labour, cell specialisation, and the presence of cell-to-cell connection (junctions – Nielsen 2012) allowing for the exchange of nutrients between different cells. In choanoflagellate colonies, on the contrary, cells might have different functions or shape but they are not joined and cannot exchange nutrients (Nielsen 2012).

The second important hallmark of the Metazoa is sexual reproduction (Nielsen 2012). Indeed, Choanoflagellata, despite the presence of conserved meiotic genes (Carr *et al.* 2008), reproduce by binary fusion only (Nielsen 2012). It is, however, important to point out that both multicellularity and sexual reproduction are known in other eukaryotic groups, including the Fungi, which are closely related to the Holozoa (Rokas 2008). The most primitive Fungi are unicellular, suggesting that the advent of multicellularity in Fungi and in Metazoa represent two independent events. However, it is unclear whether sexual reproduction should be considered an apomorphy of the Fungi-Holozoa clade that was lost in Choanoflagellata, or whether, as in the case of multicellularity, both Fungi and Metazoa independently acquired sexual reproduction.

1.1.4 Origin of Metazoa

The origin of the metazoans has received considerable attention for more than a century and to some extent still remains an open question (for a deeper discussion see Mikhailov *et al.* 2009). Recently, molecular clock analyses clarified that Metazoa separated from a choanoflagellate-like ancestor ~900 million years ago (Mya) (Erwin *et al.* 2011). The current controversy centres on what these first animals were like, what environments they inhabited, and how the change from unicellularity to multicellularity took place.

Historically, the two hypotheses that have enjoyed most support are usually referred to as the colonial theory and the cellularisation theory (Nielsen 2012). Haeckel's colonial theory was

the first widely accepted model for the origin of animals. According to this author the transition from unicellularity to multicellularity proceeded through two consecutive stages named Blastaea and Gastrea. The Blastaea consists of unicellular flagellates aggregated to form a hollow ball-shaped floating colony of identical cells. Ball-shaped colonies of flagellated cells are also observed outside the Metazoa, for example in the green algae *Volvox*.

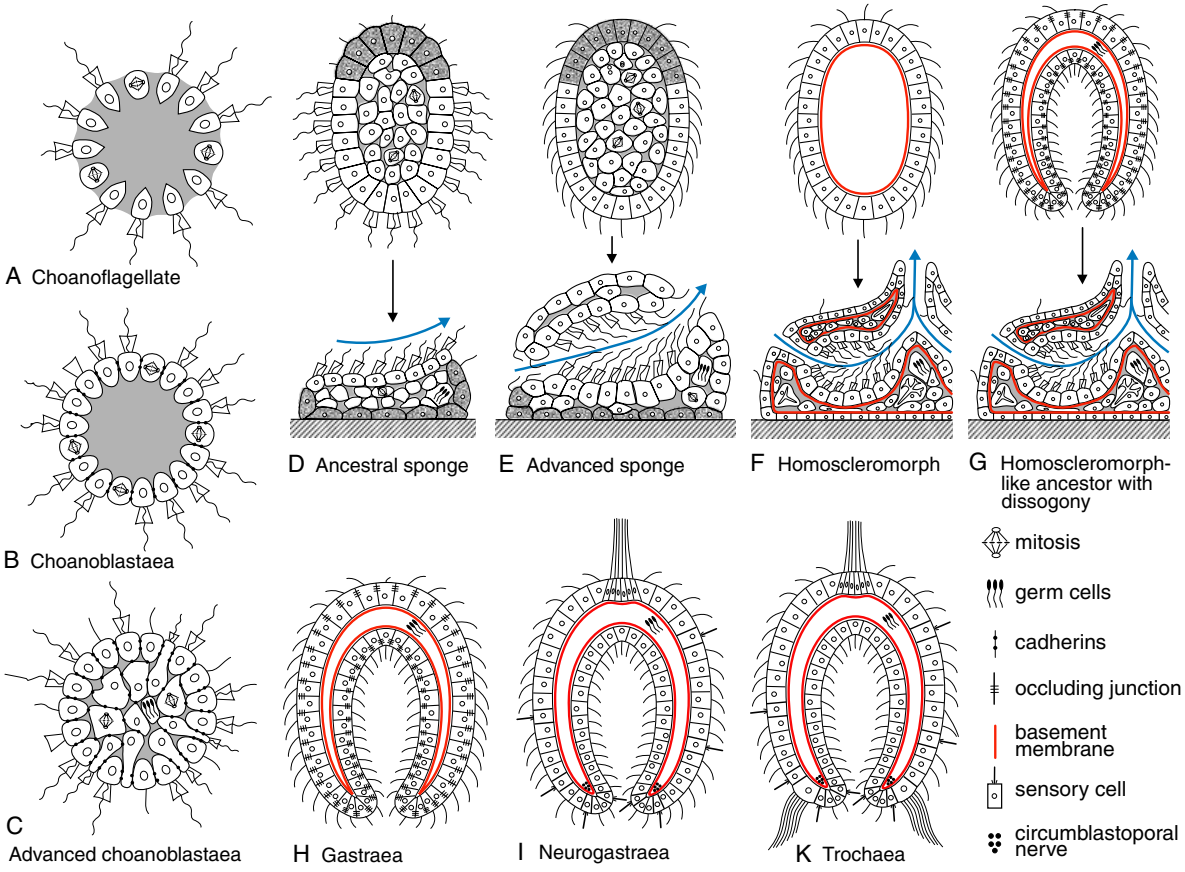


Figure 1.2: Diagrammatic representations of various stages in the evolution of the bilaterians from a choanoflagellate ancestor to the major bilaterian groups as proposed here. Extracellular matrix is represented in grey. The characters related to cell contacts are only indicated in the early stages after appearance. The blue arrows indicate the major water currents of the sponges; the currents around the individual choanocytes are not shown (from Nielsen 2008).

The ball-shaped aggregate of undifferentiated cells that compose the Blastea, according to Haeckel invaginated to reach the second (Gastrea) stage of his proposed process of multicellularisation. Invagination of the Blastea allowed for the origin of a second cell layer and of the precursor of the primary gut. After this important event this cellular aggregate acquired

primary cell differentiation (spatial cell differentiation). Haeckel who advocated that ontogeny recapitulates phylogeny, named these hypothetical ancestral forms Blastea and Gastraea to indicate their similarity with the blastula and gastrula stages in animal embryogenesis. In Haeckel's view, existing cnidarians and sponges are the first direct descendants of the ancestral Gastraea, because their body plans derive from two embryonic layers that can make a gastrula-like structure by bending its back.

The presence of a uniform blastula and gastrula with differentiated ectoderm and endoderm were postulated by Haeckel to be the characteristic of all Metazoa. However, recently Mikhailov and co-workers (Mikhailov *et al.* 2009) suggested that different cell types might have already been present in the common ancestor of the Metazoa. Mikhailov and co-workers ideas are inspired by the Synzoospore hypothesis of Zakhvatkin (Zakhvatkin 1949). This hypothesis suggests that the blastula might represent the pelagic, dispersing, and primarily non-feeding larva of a benthic sedentary metazoan ancestor. According to this hypothesis, multicellularity was not a trigger to the emergence of cell differentiation; rather multicellularity emerged as a result of the integration of different cell types.

There is one last set of alternative 'cellularisation' theories, which derive a turbelliform-metazoan ancestor through compartmentalization of a ciliate, or ciliate-like ancestor. However, theories belonging to this family of ideas are now only of historical interest (Nielsen 2012) and will not be discussed here.

1.1.5 Introduction to basal metazoan

Metazoans are currently categorised into 38 taxa (Nielsen 2012) that are generally regarded as phyla. The taxonomic status of "phylum" for some of these taxa is hotly debated (see next paragraph) and it is not commonly accepted. Within metazoan there is a general consensus on the recognition of some monophyletic supergroups i.e. Bilateria, Lophotrochozoa, Ecdysozoa and Deuterostomia.

With the concept of "basal metazoans" biologists usually refer to an assembly of four (or six, if the sponges are not assumed to represent a monophyletic group) phyla: sponges (or Porifera – if monophyletic), Cnidaria, Ctenophora and Placozoa. It is important to point out that despite these phyla are often referred to as a collective, they show substantial differences in biological organization and complexity (Valentine *et al.* 1994), and whether they form a monophyletic group is still unknown (but highly unlikely). The phylogeny of these taxa is one of the arguments that will be covered in this thesis (Chapter 2).

A good example of the different biological organisation of these taxa can be seen when comparing the sponges with the Ctenophora. The sponges, despite being multicellular, function largely like organisms with a unicellular grade of complexity, whilst the ctenophores are triploblastic animals with a complex nervous system, eyes, and digestive systems. The aim of the following section is to introduce the general characteristics of the basal metazoan lineages.

Sponges are formally named Porifera (Latin *porus,* "pore"; *ferre,* "to bear"). Poriferans are restricted to benthic marine environments, and can be described as sessile, suspension-feeding, multicellular animals that utilize choanocytes (flagellated cells) to circulate water through a unique system of water canals.

A simple level of organization characterizes the sponge *bauplan*; in fact, they lack true tissue (except possibly sponges in the class Homoscheromorpha – Nielsen 2012), a nervous system, eyes and gut. They possess an aquiferous system and some morphologically distinguished cells (Brusca and Brusca 2003). This aquiferous system changes substantially among sponges, in both size and shape, and it is used to channel water through the sponge and close to cells responsible for food gathering and gas exchange (the choanocytes). At the same time, excretory and digestive wastes and reproductive products are expelled by way of the water current flowing through the aquiferous system.

Sponges possess generally mineralised skeletal components (the spicules). Classical phylogenetic analyses were based on the anatomy of the spicules and have proven inadequate for

developing stable phylogenetic hypotheses and classifications (Brusca and Brusca 2003). There is a general agreement on the identification of four sponge classes (see figure 1.3) named Calcarea, Hexactinellida, Demospongiae, and Homoscleromorpha (Sperling *et al.* 2007; Philippe *et al.* 2009; Sperling *et al.* 2009; Pick *et al.* 2010; Sperling *et al.* 2010; Philippe *et al.* 2011; Nielsen 2012) but the phylogenetic arrangement of these classes with reference to the other animals is still debated (see chapter 2).

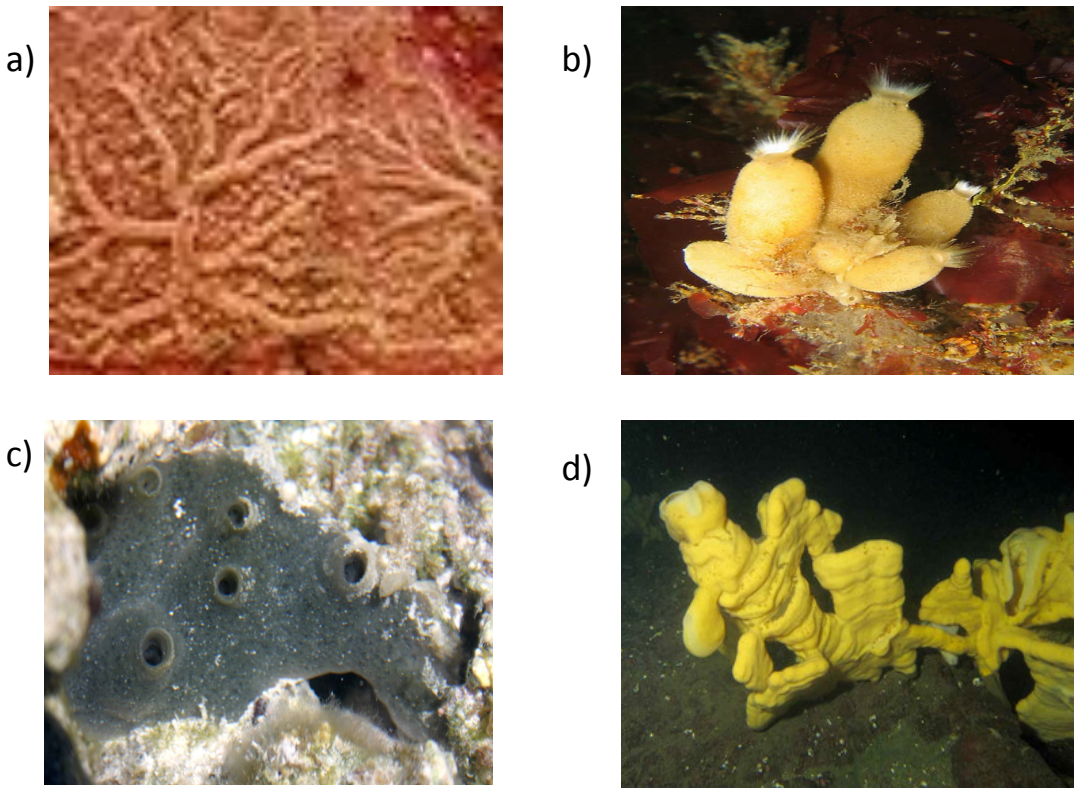


Figure 1.3: The four families of sponges. a) The Homoscheromorpha *Oscarella carmela*. b) The calcarean sponge *Sycon sp.* c) The demospongia *Amphimedon queenslandica* d) The exactinellida *Aphrocallistes vastus.*

The phylum Placoazoa includes only one described species *Trichoplax adhaerens* (see figure 1.4). However molecular studies have shown that many species, genera or families are probably included within this single species (Nielsen 2012). These cryptic taxa are generally referred to as strains. Morphologically, the placozoan are extremely simple animals constituted only of two cells layers. Placozoa are asymmetric and have no clear anterior–posterior axis. The cells of the upper and lower layers differ in shape, and there is a consistent dorsal–ventral orientation of the body relative to the substratum. Some authors (see Nielsen 2012) consider *Trichoplax* to be a true diploblastic metazoan and suggest that the upper and lower epithelia are homologous to ectoderm and entoderm, respectively. Most evidence suggests that *Trichoplax* feeds by phagocytosis on organic detritus. Although there is no evidence for extracellular digestion, *Trichoplax* may secrete digestive enzymes onto its food within a ventral digestive pocket, which is created by means of body invagination. *Trichoplax* reproduces asexually by fission of the entire body into two new individuals and by a budding process that yields numerous multicellular flagellated "swarmers," each of which forms a new individual. Sexual reproduction is also known, followed by a developmental period of holoblastic cell division and growth. Fertilised eggs have been observed within the mesenchyme, but their origin is unknown (Nielsen 2012).

Figure 1.4: The Placozoa *Trichoplax adherens.*

The phylum Cnidaria is a highly diverse assemblage of diploblastic metazoans that includes jellyfish, sea anemones and corals (see figure 1.5). Their current biodiversity accounts for approximately 11,000 extant species. The cnidarians lifestyle is characterized by a marked tendency to form colonies by asexual reproduction. Many cnidarian species exhibit a dimorphic life cycle that includes two entirely different adult morphologies: a polypoid form and a medusoid form. The dimorphic life cycle has major evolutionary implications touching on nearly every aspect of the cnidarian biology. Cnidaria possess primary radial symmetry, tentacles, and stinging or adhesive structures called *cnidae***.** An incomplete gastrovascular cavity is their only "body cavity", and a middle layer (mesenchima) derived primarily from the ectoderm, separate the two main cell layers and give consistency to their body. Cnidarians lack cephalisation, a centralised nervous system, and discrete respiratory, circulatory, and excretory organs. This basic *bauplan* is retained in both the polypoid and medusoid forms. Cnidarians are mostly marine animals, but a few groups have successfully invaded fresh waters. Most are sessile (polyps) or planktonic (medusae) carnivores, although some employ suspension feeding and many species harbour symbiotic intracellular algae from which they may derive energy (e.g. corals).

The nature of cnidarians was long debated. Until the nineteenth-century naturalists considered them plants, and it was not until the eighteenth century that the animal nature of the cnidarians was widely recognized (Brusca and Brusca 2003). Although some workers have been inclined to retain the cnidarians and ctenophores together in the Coelenterata (Philippe *et al.* 2009), these two groups are sometime recognized as paraphyletic, a view upheld by the recent molecular analysis of Pick *et al.* (2010). The older term "Coelenterata" is still preferred by some specialists, who regard it as a synonym of Cnidaria, even though it should probably only be employed to identify a superphylum including Cnidaria and Ctenophora if they were ultimately shown to be monophyletic (contrast Pick *et al.* 2010 and Philippe *et al.* 2009, and Chapter 2).

Cnidarians possess only two embryonic germ layers (the ectoderm and the endoderm) that become the adult epidermis and gastrodermis, respectively. In fact, the terms "ectoderm"

and “endoderm” were originally coined to refer to the outer and inner tissues of cnidarians, and many specialists still use them in that way. Their radial symmetry demands certain anatomical arrangements, particularly of those parts that interact directly with the environment, such as feeding structures and sensory receptors. Thus, we typically find a ring of tentacles that can collect food from any direction, and a diffuse, non-centralized nerve net with radially distributed sense organs.

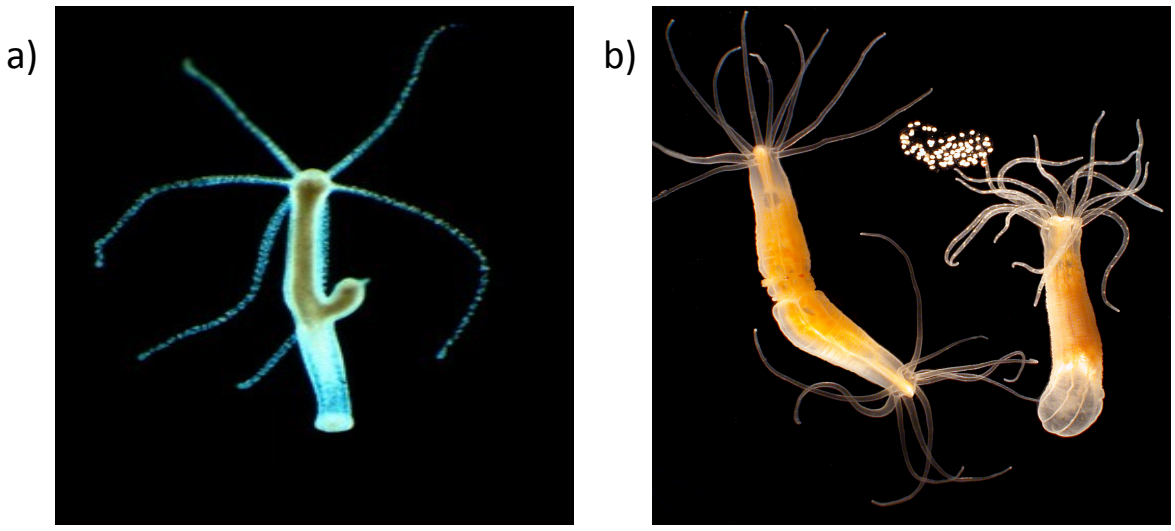


Figure 1.5: Two species of cnidarians: a) The medusozoa *Hydra magnipapillata* b) The anthozoa *Nematostella vectensis*.

Ctenophores (Greek *cten,* "comb"; *phero,* "to bear") — commonly called comb jellies, sea gooseberries, or sea walnuts—are transparent, gelatinous triploblastic animals (see figure 1.6). Most of them are planktonic, living from surface waters to depths of at least 3,000 meters; a few species are epibenthic. However, they are now known to form a major portion of the planktonic biomass in many areas of the world, and they may periodically be the predominant zooplankton in some areas. About 150 species have been described. Ctenophores are biradially symmetric, triploblastic animals. They are significantly different from cnidarians in their more extensively organized digestive system, their wholly mesodermal musculature and other minor features. Ctenophores also differ fundamentally from cnidarians in that they are monomorphic throughout their life history, they are never colonial, and do not have forms with a benthic creeping existence. They occur in all the world's seas and at all latitudes. Ctenophores do possess true tissues. Between the epidermis and the gastrodermis is a well developed middle layer, which is always a cellular mesenchyme (Brusca and Brusca 2003). Within this mesenchyme true muscle cells develop, a condition that also characterizes the triploblastic Metazoa.

The nervous system of the ctenophores is in the form of a simple, non-centralized nerve net. These organisms have locomotor structures that are arranged radially about the body. Other features that characterize the Ctenophora include: retractile tentacles and often tentacle sheaths; anal pores; adhesive prey-capturing structures called colloblasts; locomotor structures called ctenes or comb plates, arranged in comb rows; and an apical sense organ containing a statolith that regulates the activity of the comb rows. The sheathed tentacles, colloblasts, comb plates, and nature of the apical sense organ are unique features of ctenophores.

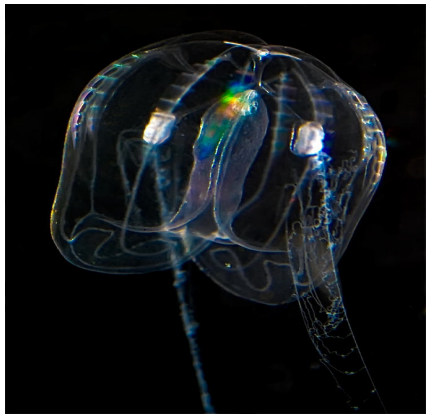
a)  b) 

Figure 1.6: Two species of ctenophores a) *Pleurobrachia pileus* b) *Mnemiopsis leidyi.*

1.1.6 Uncertainty in early animals relationships.

In the last few years multi-genes analyses (i.e. phylogenomics) clarified the relationship between the major animals clades (Philippe *et al.* 2005; Holton and Pisani 2010) confirming the existence of the Ecdysozoa, the Lophotrochozoa, and the Deuterostomia. In addition, phylogenomics found support for the monophyletic origin of the animals (Sperling *et al.* 2007; Philippe *et al.* 2009; Pick *et al.* 2010; Erwin *et al.* 2011). However, the relationships among the basal metazoans are still debated (see Figure 1.7), and resolving the branching order among the early metazoans is proving difficult (Philippe *et al.* 2011). However, from a biological point of view, alternative trees represent different evolutionary histories, and solving the animal tree is necessary to fully understand animal evolution.

In 2008, Dunn and co-workers (Dunn *et al.* 2008) analysed a data set of 150 genes and recovered a tree suggesting the monophyly of the Bilateria (Lophotrochozoa, Ecdysozoa and Deuterostomia). With reference to the "basal metazoans" they recovered a monophyletic sponge + Cnidaria clade, whilst the Ctenophora appeared as the sister-group of all the other metazoans. This position of the Ctenophora, supported also by the analysis of a 1450 gene data set later

performed by Hejnol and co-workers (Hejnol *et al.* 2009), implies an unparsimonious scenario for the origin of both the nervous system and the gut (Philippe *et al.* 2009; Renard *et al.* 2009). However, a more recent reanalysis of the data set of Dunn et al. (2008), performed by Pick *et al.* (2010), seems to suggest the sister group relation between Ctenophora and all the other metazoans is most likely a tree reconstruction artifact (Pick *et al.* 2010).

In 2009 Schierwater and co-workers (Schierwater *et al.* 2009), based on the analysis of 49 genes, suggested that Cnidaria, Ctenophora, sponges and Placoazoa form a monophyletic group. They named this hypothesis Diploblastica (see table 1.2 - even though Ctenophora have three germ layers and some cnidarians – Anthozoa seems to have muscle fibers that might be of mesodermal origin). Philippe and collaborators (Philippe *et al.* 2011) showed Schierwater and co-workers' topology to be the result of species misidentification, hidden paralogy and the use of a poorly fitting model of evolution. Accordingly, this hypothesis should be dismissed.

In 2009 Philippe and co-workers (Philippe *et al.* 2009) recovered a more classical view of the relationships between the early branches of the animal tree. Their analysis of 128 genes supported the monophyly of Cnidaria + Ctenophora (i.e. the Coelenterata hypothesis), and the Placozoa as the sister-group of the Neuralia (see table 1.2 - Cnidaria + Ctenophora + Bilateria). In their phylogeny the sponges appear as the monophyletic sister group of all the other animals (i.e. as the Phylum Porifera).

Finally, a series of studies by Sperling and co-workers (Sperling *et al.* 2007; Sperling *et al.* 2010) suggested that Sponges are paraphyletic, with the Homoscleromorpha and Calcarea being more closely related to the Eumetazoa than they are to Demospongiae + Hexactinellida. This hypothesis (named Epithelizoa see table 1.2) has found support in some morphological analyses (Nielsen 2012).

As mentioned above some of these hypotheses can be dismissed as the result of phylogenetic artefacts (i.e. Diploblastica and Ctenophora as the sister-group of all the remaining animals –Philippe et al 2011 and Pick et al. 2010). This implies that there are only two

alternative hypotheses among those that have been suggested that are still available to explain the relationships among the basal metazoans. The first sees the sponges as the monophyletic sister group of the remaining animals (Figure 1.7c and 1.7e), and the second sees the sponges as a paraphyletic assemblage of lineages with the Homoscheromorpha and in some cases the Calcarea more closely related to the remaining animals than they are to the Demospongiae. If we are to understand the early evolution of the Metazoa, we must first try to understand whether, within the context of a monophyletic Metazoa, the sponges are monophyletic or paraphyletic.

Taxonomic group		Refererence
Homoscleromorpha+Eumetazoan	Epitelizoa	Nielsen 2012, Sperling *et al.* 2007 and Sperling *et al.* 2009
Placozoan+Cnidarias+Cthenophore+Bilateria	Eumetazoan	Nielsen 2012
Cnidarians+Cthenohpore+Bilateria	Neuralia	Nielsen 2012
Cnidarians+Cthenohpore	Coelenterata	Philippe *et al.* 2009
Sponges+Placozoan+Cnidarias+Cthenophore	Dipoblastica	Schierwater *et al.* 2009

Table 1.2: Taxonomic definitions for animal relationships used in this thesis.

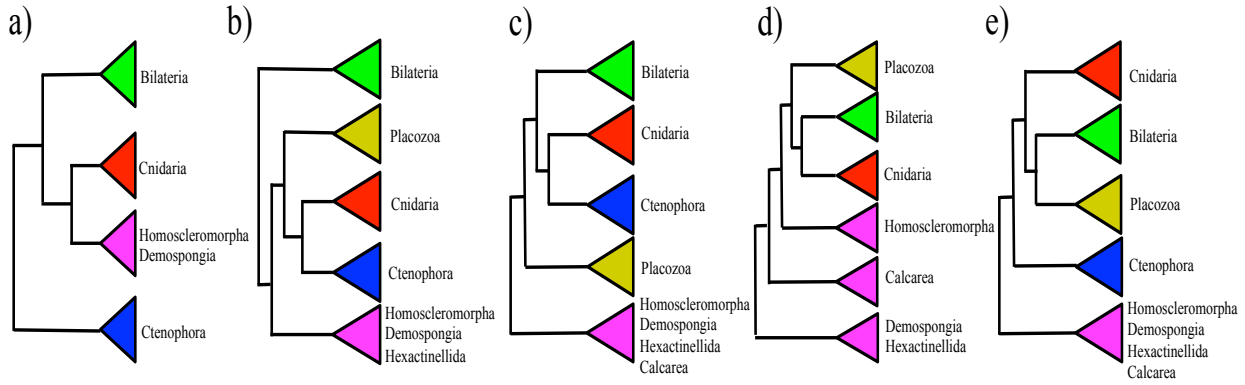


Figure 1.7: Competing hypothesis on the relationships between early animal branches. a) the phylogeny of Dunn *et al.* 2008 from the analysis of 150 genes. In this case Ctenophore are the sister group of other metazoan b) the phylogeny of proposed by Schierwater et al. 2008 from the analysis of 49 genes. In this case Cnidarians, Ctenophores and Sponges are monophyletic c) topology proposed by Philippe *et al.* 2009 from the analysis of 128 genes. In the hypothesis the sponges are monophyletic as well as Cnidaria and Ctenophora d) phylogeny from Sperling *et al.* 2009 and Nielsen 2012, the sponges are paraphyletic with Calcarea and Homoscheromorpha more close related to the eumetazoan e) phylogeny form Pick et al 2010. In this case sponges are monophyletic.

1.2 Metazoan Complexity

Complexity is a difficult concept to define, and it has been used to describe so many objects and phenomena that it has lost any generally recognized precision or meaning (Carroll 2001). A simple and widely accepted view to estimate the complexity of living organisms is by assuming that complexity relates with the number of cell-types (Valentine *et al.* 1994). From this point of view, animals with more cell-types are more complex than animals with fewer cell-types. The rationale underlying this perspective being that increasing cell-types increases the potential physiological and anatomical complexity of the organisms allowing for a finer division of labour and the formation of specialized tissues and organs (Arendt 2008).

Placozoans have only few cell types, while the Porifera (sponges) and Cnidaria (including jellyfish and sea anemones) possess 10–12 cell-types (Valentine *et al.* 1994). Cnidarians have only two distinct germ layers (that is, they are ‘diploblastic’), whereas bilaterians possess a third, mesodermal germ layer and considerably more cell types. The

evolution of the mesoderm and its derivatives had profound consequences for the evolution of animal body cavities, locomotion and overall size (Carroll 2001).

The number of cell-type can be view as a proxy for complexity but it leaves open the question of how complexity evolved. In recent years the advent of genomics allowed us to start comprehending the genetic bases of metazoan complexity. A variety of authors have tackled this issue suggesting a link between complexity and: (1) presence of transcription factors (Degnan *et al.* 2009), (2) the number coding genes (Carroll 2001) and (3) the number of microRNAs (Peterson *et al.* 2009). Other authors (Davidson and Erwin 2006) have suggested that the morphological difference between phyla is a system level problem that can only be explained by differences in the architecture of gene regulatory networks.

In this thesis I will address the problem of animal complexity focusing on the study of the G-protein coupled receptors, GPCRs superfamily. GPCRs are located on cellular membranes, making GPCRs keys elements in cellular signal transduction, which underpins biological complexity.

The level of complexity in the early diverging branches of the animal tree is very variable (see above). Sponges, for example, despite their multicellular level of organisation, function like choanoflagellate colonies. On the other hand the Ctenophora are triploblastic animals with a nervous system, gut and specialized organs. More generally the organization of complex systems (e.g. nervous, digestive) show a high level of variability in basal metazoans, and their physiological and anatomical complexity is extremely diverse. The participation of GPCRs in signal transduction and physiological systems, combined with the observation that key physiological systems evolved in the non-bilaterian Metazoa, suggest that the study of GPCRs in basal metazoan might be worthwhile to understand potential links between these proteins and animal complexity. An important aspect of this thesis will thus be testing whether GPCRs played a role in the evolution of animal complexity by modulating cell-cell communications, and

mediating interactions between the animals and their environments.

1.2.1 Gene Duplication and evolutionary novelties

The gene content of living objects changes over time by gene duplications (Ohno 1970) and horizontal transfer (Keeling and Palmer 2008). Whilst the horizontal gene transfer is ubiquitous in prokaryotes (McInerney *et al.* 2011) its influence in the evolutionary history of Metazoans seems to be reduced. Consequently, variations in gene content observed between animals are mostly a consequence of gene or genome duplications and gene deletions. Biologists have underlined the importance of gene duplication as a source of raw material for evolution since the origin of genetics (Taylor and Raes 2004). However, the milestone work on the subject is Ohno's *Evolution by Gene Duplication* (Ohno 1970), which made the case for the importance of gene duplication and considered the various types of duplications and their potential for yielding novel functions. Thirty-five years later, we are aware of mechanisms explaining the origin of genes through gene duplications (Innan and Kondrashov 2010), and biologists are aware that there are biological processes influencing gene duplicability (Conant and Wolfe 2008; Doherty *et al.* 2012)

Gene duplications are an important source of evolutionary innovation (Olson 2006). The presence of the two copies of the same gene allows them to evolve independently. One of the two copies is often free to accumulate mutations whilst the other could maintain the original function (Wagner 2011). Gene duplicates might thus be subject to opposing evolutionary forces. A duplicate can thus mutate and acquire new functions (driven by positive selection) while the second paralog would maintain the ancestral function.

1.2.2 GPCRs and animal complexity

Multicellularity has been one of the "major steps" in the evolution of animals (*sensu*

Nielsen 2008) and it involved a series of changes in the organism architecture. From few multifunctional cells, animals specialized their cellular repertoire forming tissues and organs. At the genomic level, this process happened through a series of gene duplications, acquisition of miRNAs, and transcription factors (Arendt 2008). This gene expansion allowed for an increased, cellular specialization.

GPCRs are located on cellular membranes, they probably played a key role in animal evolution, and they might have played a key role in the origin of multicellularity. Animal GPCRs can be broadly classified in two groups: the non-chemosensory GPCRs, and the chemosensory GPCRs (see below). Chemosensory GPCRs are involved in the detection of sensory signals of external origin as vision, odours, pheromones, or tastes (Vassilatis *et al.* 2003). Non-chemosensory GPCRs respond to endogenous signals, such as peptides, lipids, neurotransmitters, or nucleotides (Vassilatis *et al.* 2003) and they are involved in a multitude of physiological processes.

These two components of the GPCRs metazoan diversity contribute in two different ways to animal complexity. The non-chemosensory GPCRs responding to the endogenous stimuli are involved in communications between different cells and in the maintenance of homeostasis. The specialization of cells in tissues and organs allow the specialization of cell-cell communications. On the other hand the chemosensory GPCRs, responding directly to the external stimuli, allow animals to explore new ecological niches.

The large range of external stimuli detected by GPCRs are transduced downstream using an ancient, modular, intracellular signalling cascade present in unikont and Chromalveolata (see table 1.1) (the G-Protein signalling network) (Nordstrom *et al.* 2011; Krishnan *et al.* 2012). The G-protein signalling network consists of a receptor (the GPCR), a heterotrimeric G protein and an effector (Wettschureck and Offermanns 2005). In addition, each component, the receptor, the G protein, and the effector can be regulated independently by additional proteins, soluble

mediators, or at the transcriptional level (Wettschureck and Offermanns 2005). The relatively complex organization of the G signalling system provides the basis for a huge variety of trans-membrane signalling pathways that are tailored to serve particular functions in distinct cell types.

1.2.3 Origin and classifications of GPCRs

All GPCRs share the same structure based on a 7 trans-membrane region (7TM; see Figure 3.1). Proteins with a 7TM region are also present in prokaryotes. These include light-sensitive proteo-, bacterio- and halorhodopsins that are involved in non-photosynthetic energy harvesting in Archaeabacteria and Eubacteria (Sharma *et al.* 2006). Although structurally similar to the sensory rhodopsins found in eukaryotes (Sineshchekov *et al.* 2002; Waschuk *et al.* 2005), their phylogenetic relation to the eukaryotic GPCRs remains unclear (Soppa 1994). Sequence similarity between eukaryotic and prokaryotic 7TM proteins is low and a common origin cannot be inferred from sequence data using traditional phylogenetic methods (see Chapter 3). Indeed, it is not even clear how the eukaryotic 7TM core evolved. Notably, the highest sequence similarity between the bacteriorhodopsins and the mammalian GPCR is found in non-homologous helices. Some authors have explained this finding suggesting an evolutionary mechanism that involves exon shuffling (Pardo *et al.* 1992). An alternative hypothesis proposes gene duplication of an ancestral three trans-membrane module that gave rise to both helices 1 through 3 and 5 through 7 (Taylor and Agarwal 1993). With reference to the animal GPCRs, a variety of classification systems have been proposed: (Kolakowski 1994) grouped GPCRs in six subfamilies (from A to F); (Bockaert and Pin 1999) proposed a classification with five subfamilies. The first phylogenetic-based GPCRs classification was proposed by (Fredriksson *et al.* 2003). This author identified five clades: Glutamate (G), Rhodopsin (R), Adhesion (A), Frizzled/Taste2 (F) and Secretin (S) (Fredriksson *et al.* 2003), thus naming his classification system as GRAFS.

The Rhodopsin family in the GRAFS system, which corresponds to class A or 1 of other classification systems, is the largest family with about 672 members in the human genome including about 388 olfactory receptors (Fredriksson *et al.* 2003). The Glutamate family, which corresponds to class C or 3, is characterized by the presence of the so-called "Venus Flytrap" mechanism, which is found in the N-termini and is crucial for ligand binding. The Frizzled receptors, which correspond to class F or 5, play a role in cell polarity. Both the Secretin and the Adhesion families correspond to class B or 2. Secretins have a hormone-binding domain at their N-terminal end that interacts with peptide hormones (Schioth *et al.* 2007). Members of the Adhesion family are characterized by very long serine and threonine rich N-terminal end that displays multiple domains often found in other types of proteins such as tyrosine kinases (Bjarnadottir *et al.* 2007).

In this work I will use, as necessary, GPCRs functional and evolutionary classifications. GPCRs will thus be referred either as non-chemosensory / chemosensory, or with reference to the GRAFS system.

1.2.3.1 Non Chemosensory GPCRs

Non-chemosensory GPCRs represent a multitude of GPCRs that are not involved in vision, olfaction and taste. A physiological discussion of the GPCRs functions is beyond the scope of this work. Here, I shall only emphasize the key role played by GPCRs in the physiological processes that are characteristic of the animals (Wettschureck and Offermanns 2005). GPCRs are fundamental to process information in a wide range of animal systems. The most remarkable are the cardiovascular system, endocrine system, immune system, nervous system, development, cell growth and transformations (Wettschureck and Offermanns 2005). Furthermore, GPCRs are involved in all the physiological processes among which electrolyte and water homeostasis, metabolism, growth and reproduction, are controlled by a complex

system of cell-cell communications (i.e. the endocrine system) that produce, store, and secrete hormones directly into the circulatory system. Finally GPCRs are involved in embryogenesis, which is the highly coordinated assembly of distinct cellular communities, orchestrating the formation of a defined body plan. Numerous cell surface receptors have been implicated in the establishment of tissue polarity, including the evolutionarily conserved adhesion-GPCRs, Flamingo proteins (Yona *et al.* 2008).

1.2.3.2 Chemosensory GPCRs

Chemosensory GPCRs allow animals to perceive the external environment. In Neuralia, contrary to all the other living organisms, this interaction is meditated by a complex sensory system. The sensory system is part of the nervous system responsible for processing sensory information.

The olfactory system is based on the expression of a huge variety of GPCRs specifically in the olfactory epithelium. The vomeronasal system responds to pheromones that mediate effects on individuals of the same species and modulate social, aggressive, reproductive, and sexual behaviours (Smith 2000). The gustatory system perceives sweet, bitter, and amino acid (umami) signal through GPCRs. This system is known only in vertebrates. During recent years, two families of candidate mammalian taste receptors, T1 receptors and T2 receptors, have been implicated in sweet, umami, and bitter detection (Smith 2000). Finally, light detection in *Neuralia* (table 1.2) is mediated by the opsins (see Chapter 4). Opsins are GPCRs expressed in specific cell types called photoreceptors. Opsins perform their function by binding a light sensitive chromophore that reacts with visible light, leading to a conformation change in the opsin, switching on the physiological signal cascade (see chapter 4) (Terakita 2005).

The presence of complex sensory structures seems to be well established in Arthropods and Deuterostomia. However, the distribution of complex sensory structure seems to be less clear in the basal metazoan. Anatomically, complex structures such as eyes or the nervous system are present in Cnidaria, Ctenophora and Bilateria. Intriguingly, recent genomic data

suggest that genes involved in sensory functions pre-dates the presence of the related phenotype (Liebeskind *et al.* 2011).

Interactions between environment and animals had a crucial relevance during early animal evolution, and it is still is of crucial relevance for extant animals. From a paleontological prospective the distribution of sensory structures had changed early in animal history. Peterson and co-workers (Peterson *et al.* 2008) pointed out that Ediacaran organisms were fundamentally confined to an essentially two-dimensional world, conscribed by biomass. In contrast, the early Cambrian world was recognizably three-dimensional, with both an emergence in fauna and the first known pelagic eumetazoans (Vannier *et al.* 2009). The shift from the Ediacaran two-dimensional world to the Cambrian three-dimensional one was accompanied by, and inextricably linked with, the evolution of macroscopic sense organs (Plotnick *et al.* 2010). The change in repertoire of chemosensory GPCRs had indissolubly influenced the history of the animals, permitting the exploration of new niches that could have affected the evolution of complex sensorial structures (see Chapter 4).

1.2.4 The GPCRs repertories in basal metazoan

GPCRs have ancient origins (Krishnan *et al.* 2012) and most animal GPCRs are not animal-specific. Indeed, Krishnan and co-workers (2012) suggested an origin of the Adhesion and the Frizzled family in the unikont stem lineage and of the Glutamate and the cAMP receptor families in the common ancestor of the Chromalveolata and Unikonta (see table 1.1). The results of Krishnan and co-workers demonstrate that evolutionary divergent eukaryotes, like the unicellular chromalveolatas and the complex multicellular metazoans, share a basal signal transduction system that was present already in early eukaryotic evolution (Krishnan *et al.* 2012). Yet they did not explain how this signal transduction system evolved and what was it used for.

An unusual feature of the evolution of the eukaryotic GPCR repertoire is that it is highly

dynamic (Fredriksson *et al.* 2003). It increases from representing the 0.05% of the proteome in many unicellular eukaryotes (e.g. unicellular yeast) to more than 3% of the proteome in many metazoan lineages (Semyonov *et al.* 2008).

Although almost every GPCR family (e.g. Secretin, Glutamate, Rhodopsin) is found in the branching metazoans, the number of GPCRs varies greatly from lineage to lineage. For example, the Rhodopsin family underwent a strong expansion in the vertebrates (Fredriksson *et al.* 2005; Nordstrom *et al.* 2011). In addition, in vertebrates, non-chemosensory GPCRs were retained with a higher probability than expected after whole genome duplications (Semyonov *et al.* 2008). This may indicate that GPCR signalling is generally positively selected for, a finding that might explain why GPCR expansions are seen in some lineages of complex animals. GPCR families that were subject to independent, rapid, lineage-specific (sometime species-specific) expansions, include the olfactory (Kratz *et al.* 2002; Krautwurst 2008) chemokine (Zlotnik *et al.* 2006), aminergic (Le Crom *et al.* 2003), trace amine-associated (Hashiguchi and Nishida 2007), vomeronasal (Grus *et al.* 2007) and nucleotide receptor-like receptors (Schoneberg *et al.* 2007).

Nordström and co-workers (Nordstrom *et al.* 2011) attempted a study of the origin of the metazoan GPCRs using 13 complete animal genomes and the general pattern they observed suggests a continuous GPCR expansion from unicellular eukaryotes to *H. sapiens.* However, they also found that the number of Class A GPCRs found in *Nematostella vectensis* (Nordstrom *et al.* 2011) is bigger than that found in humans, which is counterintuitive. Unfortunately, there is not enough information available to interpret the strangely high numbers of GPCRs found in *Nematostella*.

Only a few early branching metazoan genomes are currently available. These are the genomes of the demosponge *Amphimedon queenslandica* (Srivastava *et al.* 2010), the placozoan *Trichoplax adhaerens* (Srivastava *et al.* 2008), and the, Cnidarians, *Nematostella vectensis* (Putnam *et al.* 2007) and *Hydra magnipapillata* (Chapman *et al.* 2010). *Amphimedon queenslandica* has been suggested to possess over 200 GPCRs (but see chapter 3), which

probably includes a large lineage-specific expansion of rhodopsin-related GPCRs. This result is surprising because rhodopsin-related GPCRs are mostly involved in the nervous system and sponges do not have anatomical structures that could be identified to represent a nervous system or some sort of predecessor of such a system. The number of GPCRs in *Trichoplax adhaerens* is debated. The *Trichoplax* genome project (Srivastava *et al.* 2008) had revealed the presence of several GPCRs that could be candidate sensory transducers, but the accuracy of this result is questionable. This is because these authors, for example, identify the presence of a "true" functional opsin in *Trichoplax,* which cannot be found in the deposited genomic data (personal observation). This genome seems to include also eighty-five members of the class 3 GPCR family, including putative metabotropic glutamate receptors. Members of the class 3 GPCR do not have any sequence similarity to other GPCRs. However, these numbers are significantly smaller than those reported by Nordström and co-workers (2011), as the latter authors identified ~530 GPCRs in *Trichoplax adhaerens*. Whatever the correct number of GPCRs in Placozoa, its relatively large GPCR repertoire is still surprising if one consider the morphological simplicity of *Trichoplax*. Similarly, the cnidarian *Nematostella vectensis* has around ~900 GPCRs (of which 826 are Class A – a number that exceeds those found in human) despite his relatively simple morphology.

It is clear that different metazoan lineages have different numbers of GPCRs. The questions that arise are thus (1) what is the evolutionary significance of the observed differences? (2) What are the advantages associated with an expanded GPCR repertoire? Most GPCRs do not play a primary vital role in organisms. Experiments performed on mice shown that more than 50% of the individuals with one knocked out GPCR display only an associated moderate phenotype or no phenotype at all. Only when the knockout mice are challenged with extreme conditions a defective phenotype become evident (Strotmann *et al.* 2011). A larger GPCR repertoire probably provides the organisms with more sensory information and improved homeostatic regulation. Hence, expanding the GPCR repertoire might be important to fine tune

regulatory and sensory processes.

Because of their different functionalities, the "two functional types" of GPCRs (chemosensory and non-chemosensory) could have evolved in response to distinct selective forces. The chemosensory ones, involved in processing external information, might have been strongly affected by extrinsic factors (e.g. colonization of new niches). Non-chemosensory GPCRs, that are implicated in cell-cell communications, differently, might have evolved to respond to the origin of new organs and systems. Yet, it is clear that also the evolution of non-chemosensory GPCRs might have been affected by extrinsic factors as the colonization of new habitats might involve the necessity of substantially alter homeostatic responses that might be regulated by GPCRs (e.g. when animals colonized the land).

1.3 Phylogenetics & data mining

Bioinformatics, Phylogenetic and data mining methods are the *leitmotiv* of this thesis and they will be widely used to address the study of both organisms and their proteins. In this section, I will provide a general introduction to the methods used in this thesis and to phylogenetics more broadly.

1.3.1 Homology, BLAST and Hidden Markov Models

Homology, from the Greek *Homologia*, meaning agreement, is a concept that was originally introduced by Richard Owen in 1843 and has proven key in modern biology. Owen defined homologous as: "[the] part or organ in one animal which has the same function as another part or organ in a different animal". Homology represents the foundation of any comparative analysis and the comparative approach is the key tool used across biology. Organs, systems, or genes are routinely compared to identify similarity and differences and understand what specific functionalities these similarities and differences underpin. Homology is normally contrasted with analogy (similarity of function) and/or with homoplasy (similarity arising through independent descent). The Darwinian idea of descent with modification can be best understood when comparing homologous organs of related species where differences (modifications) appear within the context of a common, underlying structure inherited from a shared ancestor (descent). This is why the similarity observed between homologous structures was referred to as "special similarity" (Fitch 2000).

The identification of homology is a central theme in bioinformatics and molecular evolution. In bioinformatics homology is statistically detected. The Basic Local Alignment Search Tool (BLAST -Altschul *et al.* 1990) and Hidden Markov Model (HMM -Eddy 2004a) respectively use sequence similarity and statistical properties of sequence alignments to identify sequences that are more similar than expected by chance. These sequences are putative

homologues. The rationale behind this idea is that homologous sequences will tend to be more similar because they arose from a common ancestor, but they will not be identical because they will have accumulated mutations since their last common ancestor. BLAST, a method for homologous gene detection uses a database representing a set of potentially homologous sequences, and a seed sequence for which we want to identify homologs. Homology between the seed and each sequence in the database is estimated by testing how likely it is for the considered seed to return a match of the observed level of significance when the seed is compared against each of the sequences in the considered database of possible homologues.

Significance of similarity for the compared sequences is measured using E-values (Expected Values). E-values are not probabilities, and are used as proxy for homology, whereby only sequences with a specified maximum E-value are considered potential homologs. The E-value generally used as the minimum requirement for two sequences to be considered homologs is 10e-8. However, the smaller the E-value, the higher the likelihood that the compared proteins are homologs. Generally, proteins with E-values < 10E-50 are considered close homologs whilst sequences with an E-value 10E-20 < E-value < 10E-8 are quite distant (but still quite certainly) homologs. Proteins with E-values > 10E-8 are very distant homologs and might be false positives (the observed similarity might be due to compositional or functional constraints). Sequences with E-values > 10E-5 are unlikely to be homologs (most likely they represent false positives).

In the case of highly divergent protein families (like the GPCR superfamily) BLAST might be unable to detect significant similarity for true, but very distant homologs. One way to overcome the limits of BLAST when dealing with distantly related sequences is using profile–sequence comparison methods such as PSI-BLAST (Altschul *et al.* 1997) and Hidden Markov Models (HMM). These approaches compare an alignment of homologs of the seed sequence against a database of sequences. They use positional specific information (e.g. the presence of

conserved sites) as landmarks to improve the detection of distantly related homologues in which very few key sites might have retained the same residue observed in the seed. PSI-BLAST works better than BLAST essentially because a multiple sequence alignment of homologous sequences contains more information about the sequences in a family than a single sequence does. The profile allows one to distinguish between conserved positions that are important for defining members of the family and non-conserved positions that are variable among the members of the family. More than that, it describes exactly what variation in amino acids is possible at each position by recording the probability for the occurrence of each amino acid along the multiple alignment (Soding 2005). The development of PSI-BLAST led to a great improvement in sensitivity in database searching and the possibility to identify much more distantly related homologues. A problem with PSI-BLAST is that it is sensitive to alignment errors when searching the database. If the alignment is corrupted (i.e. it includes proteins that are not related to the seed sequence) the results obtained using PSI-BLAST are likely to be misleading. Accordingly, this approach needs to be used with care. Another commonly used approach for the detection of distant homologues is the use of a HMM. This approach (which is related to PSI-BLAST) assumes that homologs share the same statistical properties. Statistics are inferred from a specific set of sequences (known homologs – i.e. the learning dataset). These statistics describe how sequences belonging to the protein family represented in the alignment should look like (i.e. what is the probability of observing, at each site every possible amino acid or a gap). These statistics are then used to score a set of sequences and identify which one of these would fit the alignment used to seed the search (i.e. the learning set). Sequences that fit the seed alignment well are retained as putative homologs. It is clear that HMMs are similar to sequence profiles (as in PSI-BLAST), but in addition to the amino acid frequencies in the columns of a multiple sequence alignment, they also include position-specific probabilities for insertions and deletions along the alignment, i.e. gaps (Soding 2005).

Molecular homology can be of three types: paralogy, orthology and xenology. Genes are

defined as paralogs when the homology is due to a gene duplication; orthologs when homology is the result of a speciation event, and xenologs when homology arises due to a lateral transfer of genetic material (Fitch 2000). BLAST, PSI-BLAST or HMMer do not distinguish between these types of homology. Accordingly, data mining steps are generally followed by downstream analyses performed to detect whether a homolog is an ortholog, a paralog or a xenolog.

1.3.2 Alignment and positional homology

Alignment is the procedure by which the hypothesis of homology, defined at the level of the whole sequence, is refined to identify homologous sites by placing gaps at sites where insertions or deletions have occurred since the last common ancestor (Boussau and Daubin 2010). Despite the alignment being crucial and strongly algorithm dependent (Wong *et al.* 2008), it is usually performed prior to a phylogenetic analysis and never questioned afterwards. Indeed, it is well known that misleading identifications of positional homologies can affect downstream analyses. However, homologous sites can only be defined based upon a description of the phylogenetic relationships among the considered sequences, and because such a description is not available *a priori*, most alignment algorithms start from a 'quick-and-dirty' low-quality phylogenetic tree, the guide tree (Boussau and Daubin 2010), that is then used to perform the alignment. As part of the alignment process, the software determines whether it is necessary to insert what is commonly known as a gap character (represented in the sequence by a '-') at a given site, to uphold the parallel confirmation of sites downstream. This is done to account for deletions and insertions that can have happened in some (but not all) the sequences in the dataset. Additionally, point mutations are also accounted for. This is done by means of an inbuilt weighting scheme (i.e. the use of a substitution matrix like one of the BLOSUM matrices – (Eddy 2004b), which can be defined by the user and tailored specifically to the demands of each study.

The explosion of bioinformatics resulted in a plethora of alignment software

implementing alternative algorithms. Available alignment software implementations include, for example, Clustal, Muscle and PRANK (Thompson *et al.* 1994; Chenna *et al.* 2003; Edgar 2004a; Loytynoja and Goldman 2008). For a recent review on multiple sequence alignment see (Kemena and Notredame 2009). Multiple sequence alignment software, and the algorithms they are based upon, have different strengths and weaknesses. Ultimately, they all produce an alignment that is the best estimate of the true (but unknown) alignment, given the considered algorithm, and the parameters (e.g. penalty score for gap insertion and expected frequency of amino acid substitutions) used. It is important to stress that one cannot be certain that the recovered alignment is a perfect representation of the true, unknown, alignment.

The most frequently used software for multiple sequence alignment is ClustalW (Thompson *et al.* 1994; Thompson *et al.* 2002). This is because of its long established reputation and low computational cost. However, the accuracy of this method when analysing long sequences (Edgar and Batzoglou 2006), and its handling of indels (Loytynoja and Goldman 2008) have been criticised.

Contrary to all the other multiple sequence alignment implementations, Loytynoja and Goldman's (2008) PRANK algorithm attempts to produce an alignment that more accurately reflects the evolutionary history of the considered sequences. To do this, it treats insertions and deletions as discrete events, and uses phylogenetic information to determine which of these events is responsible for every observed gap. Alignment software accounts for positional homology, providing a configuration that best explains the biological likeness of the nucleotides or amino acids of each sequence, at each site.

Once the alignment is complete (no matter what software is used to generate it), curation of the resulting sequence alignment is often necessary to eliminate regions that, for whatever reason (e.g. they might be highly variable), could have been misaligned. Manual curation is routine. However, for genomic scale studies where hundreds or families might need to be aligned

and curated, automated approaches like Gblocks (Talavera and Castresana 2007), TrimAL (Capella-Gutierrez *et al.* 2009) or BMGE (Criscuolo and Gribaldo 2010) are used. Because these programs are based on different statistical procedures they can reach different results. This simple observation implies that despite being a key step in any genomic analysis, alignments still need to be threated with caution.

1.3.3 Maximum likelihood and Bayesian estimation: A brief overview

One of the most important intellectual inheritances of the early population geneticists is the application of statistical methods to the study of evolutionary biology. Indeed, the best-known statistical framework for evolutionary inference is maximum likelihood (ML), which was initially introduced by R.A. Fisher (Fisher 1912; Fisher 1922).

For any two hypotheses H_1 and H_2 and an actualized result (the data – D), the likelihood ratio for the two hypotheses (H_1 and H_2) can be used to rank the considered hypotheses.

(1) $$L\left(\frac{H1}{H2}\right) = \frac{\text{Prob}\,(D|H1)}{\text{Prob}\,(D|H2)};$$

Given the data, if many hypotheses exist (H_1, H_2, H_3, … H_{ref}), a global ranking of the considered hypotheses can be obtained by comparing each hypothesis against a reference one. In order to simplify calculations, it is customary to compare each considered hypothesis against a hypothetical reference hypothesis for which:

(2) $$\text{Prob}\,(D|Href) = 1;$$

Accordingly

(3) $$\mathrm{L}\left(\frac{\mathrm{Htested}}{\mathrm{Href}}\right) = \frac{\mathrm{Prob\ (D|Htested)}}{1}.$$

That is, to rank hypotheses according to their likelihood (when multiple hypotheses are tested), the compared hypotheses are ranked with reference to the hypothetical (but unknown) hypothesis under which the probability of observing the data is equal to 1. This allows ranking hypotheses, under maximum likelihood, by simply calculating (for each hypothesis) the probability of the data. In the case of a set of phylogenetic trees, the likelihood of the data are calculated for each topology, given a fixed substitution model.

ML is now a well-established, hugely popular, method of phylogenetic inference, with many software implementations, including the relatively recent PhyML (Guindon *et al.* 2010) and RAxML (Stamatakis 2006), with the latter being generally considered the better performing of all currently available ML software.

A second, important statistical framework used in bioinformatics, and computational biology more broadly, is the Bayesian one. In Bayesian analysis, one tries to estimate the posterior probability of a hypothesis given the data and a prior distribution over all possible hypotheses.

Statistically, Bayesian methods are closely related to likelihood methods. The important difference between these probabilistic methods is that the Bayesian approach uses an informative prior distribution over the considered hypotheses (Felsenstein 2004).

Bayesian phylogenetics as well as Bayesian statistics is centered on the Bayes Theorem:

(4) $$\mathrm{PP(H)} = \frac{\mathrm{Prob\ (H|D)} * Prob(H)}{\sum_n \mathrm{Prob\ (H|D)}}.$$

Equation N.4 states that, given a prior distribution over the considered hypotheses and the data, one can estimate the posterior probability of the considered hypothesis by multiplying the likelihood of the hypothesis (given the data) by the prior probability of the hypothesis and dividing this value by the sum of the likelihoods of all considered hypotheses. Application of the Bayes theorem can be tricky when a prior distribution for the considered set of hypotheses is difficult to define. This is typically the case when there are an infinite number of hypotheses that have to be considered. In phylogenetics, where the number of hypotheses is always finite (i.e. the number of trees on n. taxa) one can always use an uninformative prior assigning a probability that is equal to 1/Bn (where Bn = number of binary trees on n.taxa) to every possible tree in order to estimate the posterior probability of a given tree topology (i.e. a uniform prior distribution can always be used).

A seemingly insurmountable problem with Bayesian phylogenetics has long been computational complexity. Indeed, calculating the denominator of equation n.4 is impossible for all cases where there are more than ~ 10 taxa to be considered (Yang and Rannala 1997). However, the implementation of Markov chain Monte Carlo (MCMC) algorithms, and the introduction of the Metropolis-Hasting algorithm (a mathematical trick allowing to avoid computing the denominator of equation N.4 – (Metropolis *et al.* 1953; Hastings 1970) has greatly helped popularize Bayesian MCMC methods (Yang and Rannala 1997) so that the application of the Bayesian principles to genetics has been defined a "revolution" (Beaumont and Rannala 2004).

In Bayesian phylogenetics, support for each node is represented by its posterior probability. Unlike other methods of estimating support, this has the advantage of being a measure of the probability that a particular node could be true (given the data and the model). Some authors however have contended that posterior probabilities overestimate the true support of a node (Rannala and Yang 1996). An additional benefit of Bayesian phylogenetics is that it

allows for the use of models of high dimensionality (Lartillot and Philippe 2004). This allows the integration of more realistic aspects of the substitution process into the considered evolutionary model. Bayesian inference continues to see a steady uptake in phylogenetic studies and currently boasts several software implementations, including MrBayes (Ronquist and Huelsenbeck 2003) and PhyloBayes (Lartillot *et al.* 2009).

As I previously pointed out, phylogenetics is a key aspect of this thesis. In the following section I will introduce the main procedures I used for the inference of phylogenetic trees. The focus of this section is only on the methods that I have used during my PhD. For an historical prospective of the evolution of phylogenetics see Felsenstein (2004).

1.3.4 Modelling the evolutionary process

In this thesis I only performed analyses of protein coding genes. Following Stabelli *et al.* (2012), sequences were analysed at the amino acid level. All the methods that I used are parametric and explicitly rely on the use of a model of protein evolution. Many such models exists, and they all attempt to represent the relative rates of the amino acid replacement process at homologous sites using weighting matrices derived from the analyses of real data sets.

Historically, the first method used to estimate substitution matrices was maximum parsimony (Dayhoff *et al.* 1978). Dayhoff and collaborators used parsimony and matrix multiplication to generate a class of substitution matrices named PAM (point accepted mutations) matrices. In the PAM matrices, relative rates of amino acid replacements were estimated by counting, for each amino acid, the inferred numbers of amino acid substitutions that occurred along a tree. Only closely related species and well conserved sequences were considered. The PAM 1 matrix, representing frequencies of substitutions expected to happen in a million years. Further matrices (e.g. PAM60 or PAM120) were inferred by matrix

multiplication. These matrices were supposed to model the evolutionary process between more distantly related sequences, e.g. sequences that separated 60 or 120 millions of years ago. More recently, Jones and others (Jones *et al.* 1992) used a faster (parsimony based) automated procedure to estimate a replacement matrix from a larger database of protein families. In so doing they generated a general replacement matrix (known as the JTT matrix), which is still used for phylogenetic reconstruction, multiple sequence alignment and other types of evolutionary analyses. More recently, the development of faster algorithms for maximum-likelihood (ML) allowed the development of ML-derived substitution matrices. The first such matrix to be developed was the WAG matrix (Whelan and Goldman 2001). This matrix should be seen as an update of the JTT matrix, where ML is used instead of parsimony to infer relative substitution rates. Indeed, using ML allows estimating substitution rates with greater precision. The WAG matrix, exactly as the JTT matrix, is still widely used. Further refinements of the WAG matrix have been performed with the latest one being incarnated in the recently released LG matrix (Le and Gascuel 2008).

Models like those implemented in the WAG matrix are generally referred as empirical general time reversible models. This is because (1) they are time reversible, i.e. the rate of substitution from amino acid X to amino acid Y (X->Y) is equal to the rate of substitution of (Y->X). In addition (2) their parameters are empirically derived from a set of pre-existing alignments, rather than from the data that are currently being analysed. If a dataset specific substitution model is being derived instead (as it is customary when analysing nucleotide data sets), the inferred model is generally referred to as Mechanistic General Time Reversible model. Mechanistic models tend to fit the data better, but are computationally more costly, particularly when dealing with amino acids. This is because to define an amino acid mechanistic General Time Reversible (GTR) model one need to estimate 211 parameters from the given alignment. However, the inference of mechanistic, amino acid, general time reversible models (generally referred to as GTR models) has become possible in a Bayesian framework (e.g. using MrBayes

3.0; Ronquist and Huelsenbeck 2003). The emergence of Bayesian phylogenetics also allowed more complex heterogeneous models, such as the CAT model (Lartillot and Philippe 2004; Quang *et al.* 2008) and the CAT-based models (like CAT-GTR) to be developed. The CAT model allows for a number *K* of classes, each of which is characterized by its own set of equilibrium frequencies, and lets each site "choose" the class under which its substitution history is better described. The model can be constrained, with the number of classes fixed to one as in the standard one-matrix model, or such that each site is described by its own class. Because of the amount of parameters to be inferred, CAT models can generally be used only with large data sets usually more 1000 sites long. Quang and co-workers (Quang *et al.* 2008) recently generated a series of CAT-based models in which the parameter *K* is fixed; these models, being pre-computed, are generally referred to as empirical CAT models and are suitable for single gene analyses.

Substitution models simply describe the frequency with which amino acids interchange among each other. However, it is well known that the rate at which different sites in an alignment can accept mutations vary substantially (the frequency at which alternative amino acids interchange remaining constant – i.e. as in the GTR matrix). The biological explanation for this phenomenon is that different sites are differently constrained because of functional and structural reasons. A common way to model this rate heterogeneity is to use a gamma distribution (Γ). Essentially, the rate at which sites accept mutations is modelled sampling the acceptance rates from a G distribution, which defining parameter (α) is estimated from the data (Yang 1994).

1.3.5 Model selection

Model selection can be performed using ether ML or Bayesian analysis. Models of evolution are a set of assumptions about the process of nucleotide and amino acid substitution. Whilst, in

maximum parsimony the model is implicitly built in the method, in the maximum likelihood and Bayesian analysis the model is explicit. This implies that its parameters need to be estimated (Posada 2009). If the model used in ML and Bayesian analyses is correct, then these methods are robust to phylogenetic reconstruction artifacts. However, it is important to underline that models are always approximations of a "true but unknown model" and if the model is misspecified. Furthermore ML and Bayesian analysis are sensitive to phylogenetic artifacts (Sperling *et al.* 2009; Philippe *et al.* 2011). For example, when the model assumed is wrong, branch length and divergence times may be underestimated, while the strength of rate variation among sites may be overestimated. In other words, the model makes assumptions in order to make complex computational problems tractable, and if these assumptions are incorrect the results of the analysis will be incorrect.

The evaluation of the statistical fit of a model can be performed using a series of approaches: hierarchical likelihood ratio test (hLRT), the Akaike information Criterion, the Bayes Factor, the Bayesian information Criterion (BIC) and the Bayesian Cross-validation.

In hLRT the log likelihoods of two competing model are contrasted using the following formula

(5) $$LRT= 2(l_1-l_0)$$

Where l_1 is the maximum log likelihood under the more parameter-rich model and l_0 is the log likelihood under less parameter-rich model (the null hypothesis). When the models compared are nested (i.e. the null hypothesis is a special case of the alternative hypotheses) this statistic is asymptotical distributed as a χ^2. If LRT is sufficiently large for the χ^2 to be significant, the parameter rich model should be selected. However, the hLRT is essentially an out-dated approach (Posada 2009) and better model selection strategies like the Akaike information criterion or Bayesian cross-validation (see below) are now more frequently used.

The Akaike information criterion (Akaike 1973) is used to simultaneously compare all

competing models.

(6) $$AIC = -2l + 2K$$

Where l is the log likelihood and K the number of free parameters in the model. The reasons why I preferred the AIC to the hLRT are (1) in AIC there is a penalty to be paid to accept a parameter rich model and (2) The AIC can be used to compare also non-nested models.

Model selection can be implemented in a Bayesian framework using the Bayes Factor, the Bayesian information criterion (BIC), and the Bayesian Cross-validation. The Bayes factor (BF) is similar to the LTRs in that they compare evidence (e.g. model likelihoods of competing topologies; see Pisani *et al.* 2012) for two competing models. Indeed, the BF can be considered the probability of the data given the null hypothesis, over the probability of the data given the alternative hypothesis (Goodman 1999). In this sense, essentially the BF is a measure of evidence for one hypothesis as opposed to another (Kass and Raftery 1995). The difference between the BF and the likelihood ratio test is that BF values are calculated using likelihood values marginalised across all tree topologies, rather than on a fixed optimal topology. In this way, the BF can take into consideration statistical uncertainty when comparing two hypotheses. The BF returned when two hypotheses are compared is generally interpreted according to the table of (Kass and Raftery 1995).

The Bayesian information criterion (BIC) (Schwarz 1978) provides an approximate solution of the natural log of the Bayes Factor. The smaller the BIC, the better the fit of the model to the data. Given an equal prior for all competing models, choosing the model with the smallest BIC is equivalent to selecting the model with the maximum posterior probability (Posada 2009).

Cross-validation (Browne 2000) is a very general and reliable method for comparing models. The rationale is as follows: the dataset is randomly split into two (unequal) parts, the learning set

and the test set. The parameters of the model are estimated on the learning set (i.e. the model is 'trained' on the learning set), and these parameter values are then used to compute the likelihood of the test set (which measures how well the test set is 'predicted' by the model). The overall procedure has to be repeated (and the resulting log likelihood scores averaged) over several random splits (Browne 2000). The Bayesian cross-validation has been used in chapter 4 because it allows the comparison between site homogenous model (e.g. GTR, WAG) and site heterogeneous (e.g. CAT, CAT-GTR and empirical CAT) models.

1.3.6 Assessment of support

Evaluating the reliability of a phylogenetic hypothesis is important. Two related approaches that can be used to estimate the level of support for a phylogeny are the bootstrap and the jackknife. Bootstrap is a statistical technique that was first applied in phylogeny by Felsenstein (Felsenstein). In the bootstrap analysis the original alignment is used to generate multiple (pseudoreplicate) alignments of the same dimensions. This process is replicated a certain number of times (e.g. 100 times), and each resultant alignment is individually used to build a phylogeny using the phylogenetic method of choice. A majority rule consensus method (Margush and McMorris 1981) is then used to merge the resulting trees into a single consensus solution with support values for each node. Values at the nodes represent the proportion of times a given clade is found by the analysis of the pseudoreplicated data sets.

The jackknife, which is an older statistical method, was also first used in a phylogenetic context by (Felsenstein 1985). Jackknife randomly purges a proportion of the sites from the original alignment so that the jackknifed alignment will be shorter than the original one. This resampling procedure typically will be repeated many times to generate numerous new samples. Each new sample will be subjected to regular phylogenetic reconstruction (Van de Peer 2009).

Bootstrapping and jackknifing only reflect the phylogenetic signal (or noise) in the dataset as detected by the phylogenetic method and model. Accordingly, if the inference is performed using a model that does not fit the data, the resulting support values will be misleading.

1.3.7 Phylogenetic reliability

There are two types of error that can occur in phylogenetics: systematic errors and stochastic errors. Stochastic errors affect all tree reconstruction methods equally, however, the use of genomic scale data sets largely reduce these errors (Delsuc *et al.* 2005). Accordingly, the emergence of genomic scale data sets allowed for the emergence of a form of "phylogenetic positivism" leading many biologists to suggest that the end of phylogenetic incongruence was near (Gee 2003).

However it has then been shown that systematic errors (that are positively misleading) strongly affect phylogenomic dataset, and this led Jeffroy and collaborators (Jeffroy *et al.* 2006) to correctly state in my opinion that phylogenomics was the beginning of incongruence. Systematic errors occur when a reconstruction method arrives upon an incorrect solution with stronger support as the amount of data considered increases. This situation occurs when certain characteristics of the data cause the method to be misled (Pick *et al.* 2010; Philippe *et al.* 2011). There are a variety of sources of systematic error e.g. compositional bias and long-branch attraction. The last part of this chapter will cover the main sources of systematic errors, the methods used to recognise them, and the strategies used to eliminate or reduce them.

1.3.7.1 Compositional bias

Compositional biases cause sequences to be erroneously grouped together based upon their analogous nucleotide or amino acid composition. This source of systematic error can affect both nucleotides and amino acids. Detection of compositional problems in a dataset can be performed using principal component analysis (PCA Stabelli *et al.* 2012) or a Bayesian posterior predictive

analysis for composition homogeneity (Foster 2004). Once compositional heterogeneity is confirmed one needs to test if the topology recovered from the analysis is driven by the compositional bias or by real phylogenetic signal. Compositional heterogeneity induced biases can be ameliorated or avoided using direct or indirect methods. The Dayhoff recoding strategy (i.e. recoding amino acids in their functional classes- see figure 1.8) has been shown to significantly reduce compositional biases (Hrdy *et al.* 2004). Phylogenetic analyses are then performed on the recoded data set. A problem with this approach is that recoding a dataset can cause a reduction (erosion) of "good" phylogenetic signal. Another approach is to account directly for compositional problems using heterogeneous models. When heterogeneous models are used compositional biases are directly accounted for whilst performing the phylogenetic analyses,

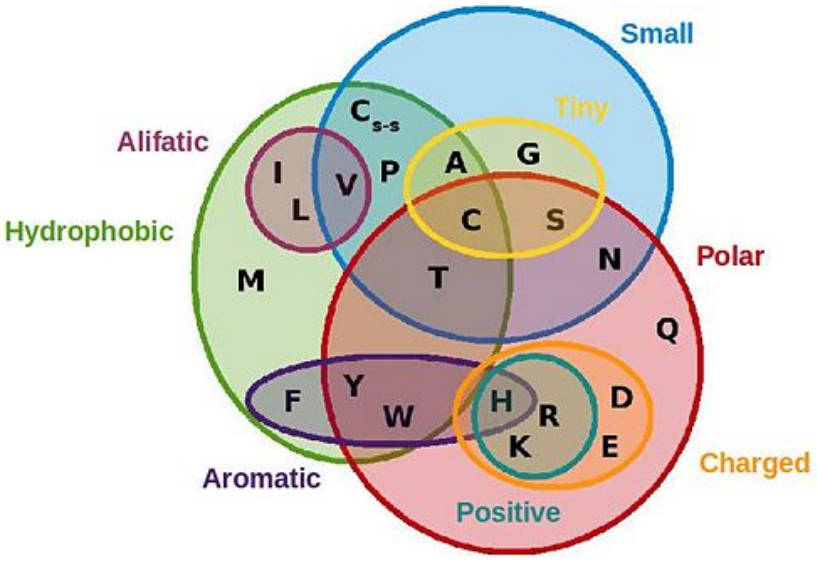


Figure 1.8: Venn diagram for the 20 most common amino acids. This diagram graphically represents the classes into which amino acids are recoded when using the Dayhoff strategy.

as the model used implements multiple compositional vectors across the tree (e.g. Foster 2004). The problem with direct approaches is that they are computationally expensive and, as such, of limited utility (Rodriguez-Ezpeleta *et al.* 2007). A third way to reduce or overcome compositional problems is site-stripping (Rodriguez-Ezpeleta *et al.* 2007). This technique

involves splitting a dataset into slow evolving partitions and fast evolving ones, based on the sites' evolutionary rates. As shown in many studies partitions containing fast evolving sites also contains the compositionally most heterogeneous sites (Rodriguez-Ezpeleta *et al.* 2007). Accordingly, removing the fast partitions reduces compositional problems (Feuda and Smith in prep – but see Cummins and McInerney 2011). Finally, compositionally heterogeneous sites can be directly removed from an alignment, after having performed a χ^2 test, as it has been proposed by Criscuolo and Gribaldo (2010).

1.3.7.2 Long branch attraction

Long branch attraction (LBA) is the most infamous and well-documented systematic error affecting phylogenetic reconstruction, and it was initially identified by Felsenstein (Felsenstein 1978). It occurs when species in a data set have heterogeneous rates of evolution. If a poorly fitting model is applied to such a data set it is often the case that slow and fast evolving species partition according to their rate: i.e. fast evolving species attract each other and the slowly evolving species are thus equally clustered in an artificial group of slowly evolving ones. Phylogenetic methods are differently affected by LBA, and non-parametric approaches (like Maximum Parsimony) are particularly strongly affected. However, even parametric approaches (like ML and Bayesian analysis) are not immune from this artifact if the data are analysed using a misspecified model.

LBA can be addressed in a variety of different ways. One of the most widely used approaches is to increase taxonomic sampling. When taxonomic sampling is increased, the introduction of new species serves to break up long branches (Rota-Stabelli *et al.* 2011; Campbell *et al.* 2011). Another way to minimize LBA is the use of optimal outgroups (Wheeler 1990; Lyons-Weiler *et al.* 1998; Rota-Stabelli and Telford 2008). When outgroups that are too divergent (i.e. long-branched) are selected, fast evolving ingroup taxa may be artifactually

attracted to the (long-branched) outgroups (Philippe and Laurent 1998). Indeed, the use of an extremely inappropriate outgroup becomes equivalent to using a random, highly saturated, sequence to root your tree (Wheeler 1990).

Various strategies can be employed to ensure the selection of an appropriate outgroup (Sanderson and Shafer 2002). In addition, LBA can also be circumvented by the adoption of a selective sampling strategy. In this approach, the evolutionary rate of large clades is assessed, with taxa exhibiting a particularly rapid rate being removed. Lastly, LBA can be alleviated or eliminated by the removal of fast evolving sites (Brinkmann and Philippe 1999; Hirt *et al.* 1999; Ruiz-Trillo *et al.* 1999; Pisani 2004). This is the same approach discussed above for the elimination of compositionally biased sites.

1.3.8 Phylogenomics

Phylogenomics is a discipline laying at the intersection of evolution and genomics. This term comprises several areas of research at the interplay between molecular biology and evolution. Characterising aspects are: (1) using molecular data to infer species relationships (see chapter 2), and (2) using information on species' evolutionary history to gain insights into the mechanisms of molecular evolution (see chapter 4). These two main applications of phylogenomics rely on different methods (see chapter 2 and 4).

When used to infer species relationship, phylogenomic analysis relies on two classes of methods: the supertrees and the supermatrix. Von Haeseler (2012) discussed the main differences between these two approaches. Supertrees methods combine source-trees, or trees obtained from the literature, with overlapping species sets into one tree. On the other hand, supermatrix methods (used in chapter 2 of this thesis) use a concatenation of multiple genes alignments. Because it is claimed that supermatrix approaches use the phylogenetic information encoded in the characters more fully than supertree methods (von Haeseler 2012), supermatrix approaches seem to be

superior (de Queiroz and Gatesy 2007). However, the supermatrix methods have potential pitfalls. Almost all phylogenetic tools treat the characters in the supermatrix as independent. This is not true for most sequences and therefore it may lead to systematic errors. Another potential pitfall is that although tree reconstruction methods include very complex models of sequence evolution, they cannot yet account for the complexity in super alignments. Finally, the assumption that gene trees are identical to speciation trees is not necessarily true and this introduces another potential bias (von Haeseler 2012).

When both supermatrix and supertree approaches deal with molecular data we have to ensure that genes sequences included in the alignment are orthologous. If the orthology assumption does not hold, then both approaches will produce misleading trees (see chapter 2).

1.3.9 Phylogenomic network

The complexity of the evolutionary process sometimes is difficult to describe with a phylogenetic tree. In a recent paper Chan and Ragan (2013) suggested the limitations of the traditional phylogenetic methods when they deal with complicated evolutionary history (e.g. gene fusion, gene deletion copy-number variation and recombination). Some of the processes mentioned above – recombination, duplication, gain and loss – play out *within* genes as well, yielding regions that can be aligned only ambiguously, or not at all. Given the heuristic nature of key steps in standard phylogenomic workflows, the relevance of alignment scores to homology can be difficult to assess statistically. As alternative to the traditional phylogenetic methods, Chang and Ragan (2012) invoke the development of a next generation methods for the phylogenetic inference. Among these next generation methods for the phylogenetic inference, phylogenetic networks are extremely powerful to describe complicated evolutionary history, because they make fewer assumptions than traditional phylogenetic methods. Mostly important they do not assume full-length sequence contiguity. These features allow the application of this class of methods to analyse the relationships between proteins that share a low level of sequence

similarity (e.g. 7TM proteins, see chapter 3).

Exactly like trees, networks are mathematical structures composed of vertices (nodes – entities) linked by edges (branches – relationships) representing the interactions between these entities. However, differently from trees, networks can contain cycles (i.e. closed circuits). Similar to phylogenetic trees, phylogenetic networks can be reconstructed from various data types including molecular sequences, evolutionary distances, presence/absence data and trees (Dagan 2011).

1.3.10 Ancestral state reconstruction and protein evolution

Reconstructions of ancestral character states make it possible in principle to describe what the past was like and to discover how traits evolved (Pagel 1999). Statistically, the evolution of a trait is modelled using the Markov process that adopts only a finite number of states. Ancestral state reconstruction has been widely used in the evolutionary biology including protein evolution, studies of sexual selection, and diet preferences (Pagel 1999).

Maximum parsimony, maximum likelihood and Bayesian methods have all been used to infer ancestral states. Whereas ML and MP assumes a tree and model parameters when inferring ancestral states, Bayesian approaches incorporate uncertainty by summing likelihoods over a distribution of possible trees or parameter values, all weighted by their posterior probabilities.

Williams *et al.* (2006) showed that alternative approaches to character state reconstruction have different properties and differently affect the thermodynamic stability of the reconstructed proteins. Notably, they found that maximum parsimony and maximum likelihood methods that reconstruct the "best guess" amino acid at each position tend to overestimate thermodynamic stability of the inferred proteins. Differently, Bayesian methods that sometimes choose less-probable residues from the posterior probability distribution, result in the smallest and most unbiased errors in stability. Accordingly, Bayesian methods should probably be

preferred when performing ancestral character state reconstruction.

1.3.11 The approximately unbiased test.

The likelihood function described in section 1.3.3 can be used also to test tree topologies (Goldman *et al.* 2000). A very common used method to compare among different topologies is the Approximately unbiased test (AU test) (Shimodaira 2002). This method produces a number ranging from zero to one for each tree. This number is the probability value or P-value, which represents the possibility that the tree is the true tree. The greater the P-value is, the greater the probability that the tree is the true tree. Relative certainty, or uncertainty, in tree selection can also be represented as the confidence set—the set of trees that are not rejected by the test. It is expected that the true tree will be included in the confidence set.

1.4 Aims of this thesis

The general aim of this thesis is to conduct a comprehensive investigation of several aspects of the early animal evolutionary history.

The increased availability of genomic-scale data, together with major advances in computational power, makes it possible to investigate the origin and early evolution of the Metazoan at the molecular level.

In particular, in chapter two, using the phylogenomic approach, I will reconstruct the phylogenetic relationship among basal metazoan. To do this, I will carefully assemble a new phylogenomic dataset considering the several sources of systematic error (i.e. outgroup and compositional bias). The aim of this chapter will be to generate a working hypothesis to be used in chapters three and four.

Therefore in chapter three, I will study the phylogenetic relationship among GPCRs. This protein superfamily shares a common structure of 7 transmembrane domains, it is involved in several physiological processes (see section 1.2.2) and most likely it has played a role in the diversification of animals. Interestingly proteins with 7TMD are also present in Archaebacteria and Eubacteria. I will try to better understand the origin and the diversifications of the 7TMD. In order to do this, I will build a broad dataset of genomes including Archaebacteria, Eubacteria and representative genomes from all the five supergroups of Eukaryotes (see Table 1.1). To reconstruct the phylogenetic relationship among proteins with 7TMD architecture I will use the phylogenetic network, which offers several advantages for studying the relationship among highly divergent protein families (see 1.3.9).

Finally, in chapter 4 I will try to infer the origin and the duplication pattern of the opsin (a sub-family of GPCRs) that plays a fundamental role in the visual process in metazoan. To do that, I will assemble a large dataset of metazoan opsins including all the possible putative outgroups. Other sources of systematic error (i.e. model of evolution and alignment) will be also considered.

Chapter 2

Phylogenomics of the basal metazoan and the evolutionary relationships of the sponges

Abstract

Early animal relationships are still hotly debated, and three main hypotheses have been proposed in the last few years (see figure 1.7). The first suggests that the sponges represent the monophyletic sister group of all the other Metazoa. The second hypothesis suggests that sponges plus the Coelenterata (see table 1.2) and perhaps the Placozoa represent the sister group of all the other Metazoa (an hypothesis named Diploblastica), and the third suggests that sponges are paraphyletic, with the Demospongiae representing the sister group of all the other Metazoa, and a Homoscleromorpha representing the sister group of the Eumetazoa (see table 1.2 and section 1.1.6).

Recent evidences suggested that Diploblastica could be dismissed as the result of paralogy, alignment errors and tree reconstruction artifacts. However, it is still unclear whether sponges represent the monophyletic or the paraphyletic sister group of Eumetazoa. In this chapter I have assembled a new phylogenomics data set of 146 nuclear genes (146-NG), illustrating how the outgroup choice and the compositional hetetogeneity are understimated issueses in te basal metazoan phylogeny.

My results confirm that Diploblastica is a phylogenetic artifact. In addition to that they also provide evidence suggesting that sponge monophyly might also represent a tree reconstruction artifact as previously postulated by Sperling *et al.* (2009). In any case, these results indicate that current evidence to resolve the phylogeny of the sponges is scant and that the problem of posed by the sponge phylogeny cannot be considered resolved yet.

2.1 Introduction

It is now generally agreed that the Animalia or Metazoa is a monophyletic group with the Choanoflagellata as their sister group (Carr *et al.* 2008; Dunn *et al.* 2008; King *et al.* 2008; Hejnol *et al.* 2009; Philippe *et al.* 2009; Mallatt *et al.* 2010; Pick *et al.* 2010; Philippe *et al.* 2011). The monophyly of the Bilateria is also strongly supported by both morphological and molecular data (Nielsen 2012). However, rooting the metazoan tree has proven to be difficult and the interrelationships of the non-bilaterian appeared unresolved. A number of alternative phylogenies, suggesting different arrangements of the non-bilaterian metazoans have been proposed. In particular, recent large scale phylogenomic analyses have proposed Diploblastica (Porifera, Placozoa, Cnidaria, and Ctenophora) as sister group of Bilateria (Schierwater *et al.* 2009); monophyletic Porifera as sister group of Eumetazoa (table 1.2; Philippe *et al.* 2009; Pick *et al.* 2010); Ctenophora (Dunn *et al.* 2008) as the sister group of all the other Metazoa with a monophyletic Porifera plus Cnidaria as the sister group of the Bilateria. Schierwater et al.'s (2009) work on the Diploblastica hypothesis has been recently shown by Philippe *et al.* (2011) to be artifactual: the result of paralogy, incorrect gene assignments and tree reconstruction artefacts. Pick *et al.* (2010) performed a series of reanalyses of a modification of the Dunn *et al.* (2008) super-alignment, showing that when some problematic genes were excluded and the taxonomic sampling of Dunn *et al.* (2008) was improved, significant topological changes could be observed. Ctenophora was henceforth not recovered as the sister group of all the other Metazoa, but as the sister group of a Cnidaria plus Placozoa and Bilateria group (see Pick *et al.* 2010). In addition, the monophyletic Porifera plus Cnidaria group found by (Dunn *et al.* 2008) disappeared. Instead, a monophyletic Porifera was found as the sister group of all the other Metazoa (as in Philippe *et al.* 2009; Philippe *et al.* 2011). Results from Pick and co-workers (Pick *et al.* 2010) are more in line with traditional (morphology-based) views of animal evolution, than those from Dunn and co-workers (Dunn *et al.* 2008).

Another alternative hypothesis of metazoan relationships is the Epitheliozoa hypothesis (see table 1.1; Sperling *et al.* 2007; Sperling *et al.* 2009; Sperling *et al.* 2010). This hypothesis suggests that the sponges are paraphyletic with (the Homoscleromorpha, the Calcarea and the Demospongiae plus Hexactinellida) being sequential (increasingly more distant) sister groups of the Eumetazoa. The Epitheliozoa hypothesis has been repeatedly supported by the alignment of seven selected nuclear housekeeping genes (Sperling *et al.* 2007; Sperling *et al.* 2009; Sperling *et al.* 2010). Recently, using a large EST alignment Hejnol and co-workers (Hejnol *et al.* 2009) recovered a tree showing the sponges as a paraphyletic assemblage with the Homoscleromorpha which are more closely related to the Cnidaria and the Bilateria, than it is to the Demospongiae. However, (Hejnol *et al.* 2009) had a very poor sampling of sponges and their tree found the Ctenophora as the sister group of all the other Metazoa (exactly as Dunn *et al.* 2008). Similar to the data set from Dunn *et al.* (2008) of which it represents an updated version, one could thus speculate that the data set of Hejnol *et al.* (2009) might also be problematic (see Pick *et al.* 2010). In addition, a recent study of Roure *et al.* (2012) suggested that when missing data were added to the outgroup taxa in the data set of Pick *et al.* (2010) the Epitheliozoa hypothesis was recovered, suggesting that this topology might also be artifactual. If the results of (Roure *et al.* 2012) were generalizable, then the results of (Philippe *et al.* 2009), suggesting a monophyletic sponges to be the sister group of a monophyletic Eumetazoa (Placozoa plus Neuralia), with the Neuralia composed of a monophyletic Coelenterata representing the sister group of Bilateria, would remain as the only viable hypothesis to describe the relationships among the basal Metazoa. However, the results of Roure *et al.* (2012) are difficult to generalise. In part this is because the Pick *et al.* (2010) data set is directly derived from the data set of Dunn and co-workers (Dunn *et al.* 2008), which was shown by Philippe and others (Philippe *et al.* 2009) and Rota-Stabelli and others (Rota-Stabelli *et al.* 2011) to be quite saturated, and not adequate to study the high-level relationships among the animals. It is certainly true that Pick *et al.* (2010) improved the quality of Dunn and co-workers Dunn *et al.* (2008) data set, for example by

removing paralogs and improving the sampling of sponges. Yet the relatively high level of saturation of Dunn *et al.* (2008) cannot be ameliorated since they represent a feature of the genes in the Dunn et al.'s data set. In addition, this result is difficult to extend to the datasets of (Sperling *et al.* 2007 and Erwin *et al*. 2011; Sperling *et al.* 2009) because the 7-housekeeping genes used in these studies includes few missing data.

Current uncertainty on non-bilaterian metazoan relationships revolve around whether the sponges are the monophyletic sister group of all the Metazoa (Philippe *et al.* 2009; Pick *et al.* 2010; Philippe *et al.* 2011), or a paraphyletic assemblage in which the Demospongiae and Hexactinellida (Sperling *et al.* 2009; Sperling *et al.* 2010) sister group of all the other Metazoan, with the other two sponge classes (Calcarea and Homoscleromorpha) more closely related to the Eumetazoa than they are to Demospongiae and Hexactinellida (Sperling *et al.* 2007; Sperling *et al.* 2009; Sperling *et al.* 2010). Notably, morphology is ambiguous with reference to this problem, and morphological analyses (depending on the interpretation of some key characters) support either sponge monophyly or the Epitheliozoa hypothesis (compare Nielsen 2008; Philippe *et al.* 2009; but see Nielsen 2012).

It is worth mentioning that uncertainty in the phylogeny of the non-bilaterian metazoans should be seen as a rooting problem. Indeed, the differences between Philippe *et al.* (2009) EST-based phylogeny and the trees obtained by Sperling *et al.* (2007), Sperling *et al.* (2009) and Sperling *et al.* (2010) disappears when the root is suppressed and the outgroups are not considered. This is because these studies found identical unrooted trees. Given that the differences between Pick *et al.* (2010) topology and Sperling *et al.* (2007), Sperling *et al.* (2009) topologies (if one were to exclude the way in which the relationships among the sponges were resolved) are inconsequential, and given that neither Sperling *et al.* (2007), Sperling *et al.* (2009), Sperling *et al.* (2010) nor Pick *et al.* (2010) or Philippe *et al.* (2009), Philippe *et al.* (2011) used explicitly described objective criteria (e.g. Rota-Stabelli and Telford 2008) to select the outgroups they used. In this chapter I will present an analysis of the effect of outgroup

selection on our understanding of early metazoan evolution. Accordingly, I assembled and analysed a new EST data set (see methods) based on the scarcely saturated data set of Philippe *et al.* (2009). In contrast to previous works, the key improvements of this study are (1) new data for three key lineages the homoscleromorph sponge *Oscarella carmela* and the choanoflagellate outgroups (*Monosiga ovata* and *Proterospongia sp.*) were added to the data set of Philippe *et al.* (2009)– reducing the amount of missing data. (2) A more thorough (manual and tree-based) ortholog-gene selection strategy was implemented. (3) The potential misleading effect of several sources of phylogenetic inaccuracy (particularly compositional heterogeneity) were thoroughly considered. (4) Objective outgroup selection Rota-Stabelli and Telford (2008), a key aspect of phylogenetic reconstruction that has not been considered in previous analyses of the basal metazoan relationships (e.g. Dunn *et al.* 2008; Philippe *et al.* 2009; Sperling *et al.* 2009; Pick *et al.* 2010) was implemented.

The results here presented, suggest that the sponges are paraphyletic and provide support for the Epitheliozoa hypothesis (see table 1.2). Additionally, my results confirm that outgroup selection can have a powerful influence on the results of phylogenetic analyses, and hence on our understanding of early metazoan evolution. In particular, I show that assuming that the phylogenetically closest taxon (the Choanoflagellata in the case of Metazoa) must be, by definition, the best outgroup to be used in a phylogenetic analysis, is erroneous (see also Lyons-Weiler *et al.* 1998; Rota-Stabelli and Telford 2008). Also, I show that compositional heterogeneity and the presence of missing data can have non-trivial effects on results of deep-time phylogenetic analyses. With reference to early metazoan evolution, my results illustrate that, despite a large body of evidence accumulated in recent years favouring sponge monophyly, (Philippe *et al.* 2009; Pick *et al.* 2010; Philippe *et al.* 2011), it is still uncertain whether sponges are monophyletic or paraphyletic. Nonetheless, it is clear that a greater availability of multiple sponge genomes, as well as multiple outgroup genomes, and a denser gene sampling (i.e. less

sparse matrices) will be necessary before the difficult problem of correctly rooting the animal tree of life could be finally resolved.

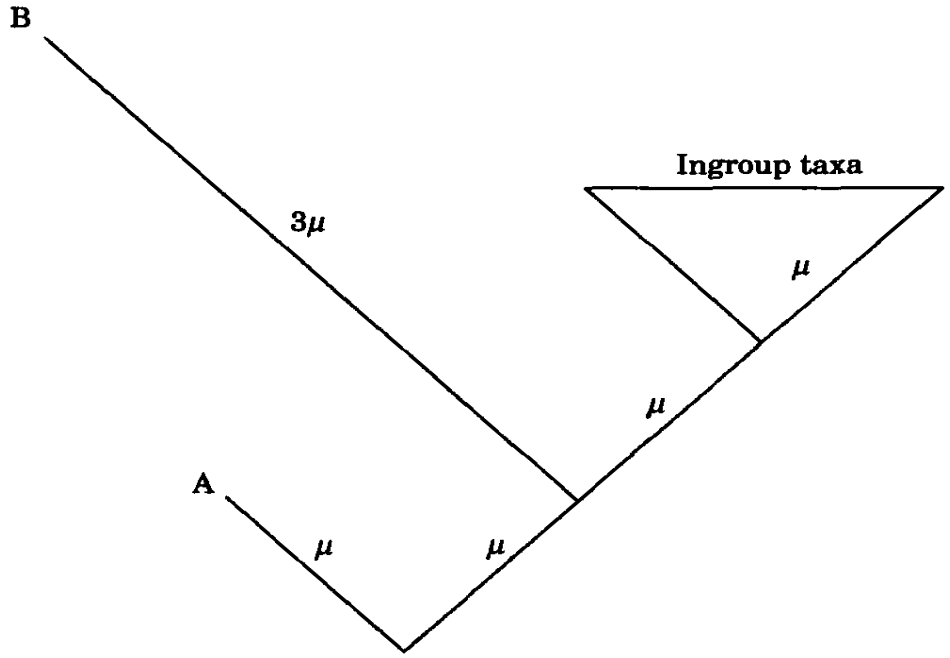


Figure 2.1: The complexity of outgroup choice. The validity of the assumption that the sister taxon (B) of the ingroup is ideal depends on specific aspects of the outgroup itself (e.g. its rate of evolution). In the case reported in this figure, Taxon A is likely to be of greater utility in the study of the ingroup. The problem with the phylogenetically-closest outgroup (taxon B) is that it is fast evolving and thus long branched. As a consequence it does not minimize the tip-to-tip (pairwise) distances between the outgroup and the ingroups (from Lyons-Weiler *et al.* 1998).

2.2 Methods

To generate the data set used in this study I modified the data set kindly provided by Professor Hervé Philippe. With reference to published results the data that Prof. Philippe sent me corresponds to the alignment used for the saturation plot in Philippe *et al.* (2009). This data set has been shown to be less saturated than that of Dunn *et al.* (2008) and Pick *et al.* (2010); it

includes 148 orthologs for 85 species, 8 of which are outgroups (see appendix A). From the raw data a super-matrix was assembled. To improve over the other EST-data sets (Dunn *et al.* 2008; Philippe *et al.* 2009; Pick *et al.* 2010), the amount of missing data in this data set was reduced. To do so, new data was added for two key species *Oscarella carmela* (Nichols *et al.* 2012) and *Proterospongia sp.* As suggested by Roure *et al.* (2012), missing data can have three negative effects on phylogenetic inference: (i) cause parameter misestimations, (ii) decrease resolving power, and (iii) reduce the detection of multiple substitutions. Furthermore, ortholog gene selection was performed using a rigorous (manual) phylogenetic approach (see above), and before gene concatenation, orthologs genes were aligned using Prank (Loytynoja and Goldman 2008) to generate an alignment of the highest possible quality.

Because the *Oscarella carmela* genome has been sequenced using Next Generation Sequencing, and has not been annotated (http://compagen.zoologie.uni-kiel.de/datasets.html), I performed gene prediction using Augustus (Stanke *et al.* 2008) trained on the closest species (i.e *Amphimedon queenslandica*) in order to estimate the parameters for the gene prediction.

New genes were added to reduce missing data with reference to the data set of (Philippe *et al.* 2009). Putative ortholog genes were identified using BLAST-P (Altschul *et al.* 1990). A representative of each orthologs in the considered set of 148 genes was searched against the complete proteomes of *Oscarella carmela, Amphimedon queenslandica* and *Proterospongia sp.* Sequences with e-values lower than 1^{-10} were retained as potentially homologues genes. New data are available also for *Monosiga ovata* from NCBI Trace archive. However, genomic data for *Monosiga ovata* are restricted only to ESTs. Accordingly, a different data mining procedure was adopted for this taxon. Each of the 148 genes in Philippe *et al*. (2009) alignment was searched using tBLASTn against the complete ESTs sequences available for *Monosiga ovata.* Sequences with an e-value below 1^{-10} were retained as putative homologues and then translated into protein using TranslatorX (Abascal *et al.* 2010).

For each of the 148 genes, the putative orthologs identified for *Oscarella carmela*, *Amphimedon queenslandica, Proterospongia sp.* and *Monosiga ovata* were then aligned to the original data set using a profile method as implemented in Muscle (Edgar 2004b).

Each one of the 148-ortholog data sets were analysed using Maximum Likelihood to generate a gene phylogeny. The best-fitting model for each of 148 alignments was chosen using the Akaike information criterion as implemented in Modelgenerator (Keane *et al.* 2006). ML analyses were then performed under the best-fitting model, using RAxML (Stamatakis 2006). Support for the nodes on these trees was inferred using the bootstrap 108 replicates. 108 replicates were result of the parallelization of the analysis on 12 processors (i.e. the number of processors available per core in Stokes). Despite I chose 100 bootstrap replicates the calculation of the 100 replicates cannot be divided into 12 tasks (100/12= 8.33), and so each processor is doing 9 replicates, 9*12=108.

Each of the 148 trees was manually inspected to select orthologs genes for each of the newly added taxa (*Oscarella, Amphimedon, Monosiga ovate and Proterospongia sp.*). Chimerical sequences were then generated following the scheme used by Pick *et al.* (2010)

Fast evolving sequences (associated with long branches in the gene trees) were identified and removed to avoid increasing the level of saturation of the genes in the super-matrix using a manual procedure. To reduce the amount of missing data, two genes with low species coverage were removed from the analysis (these genes are rplA and rplB). The final dataset consisted of 146 genes (hereafter I refer to this data set as the 146-NGs data set). Each of these genes was de-aligned and realigned using Prank (Loytynoja and Goldman 2008). Single gene alignments were then trimmed using Gblocks (Talavera and Castresana 2007) with the same parameters of Pick *et al.* (2010). Gene concatenation was performed using FASconCAT (Kuck and Meusemann 2010). The new, complete alignment score 85 species and 32432 amino acid positions (see appendix A). From this original alignment, fast-evolving bilaterian species (*Ciona intestinalis, Spinochordodes tellinii, Schimdea mediterranea, Paraplanocera sp., Dugesia japonica,*

Echinococcus granulosus, Macrostatum lignano, Xenoturmbella bocki, Richetersius conifere and Hypsilisbus dujardini) were excluded. Constant sites were also removed to reduce computational complexity. The final alignment scored 75 species and 23328 positions.

2.2.1 Phylogenetic Analyses

The 146-NGs data set was analysed using the CAT model. This is a site-heterogeneous model that is well known for his robustness to tree reconstruction artifacts like LBA. In the CAT models sites are partitioned in categories that are biochemically defined, and category-specific substitution matrices are applied to the data. This is in stark contrast to models like WAG where one single GTR matrix is applied to every site in the alignment, irrespective of the amino-acid equilibrium frequencies specific of each site in the alignment. Other CAT-based models such as CAT-GTR exist. These models might fit the data better than CAT but they are extremely costly from a computational point of view and were not applicable to my data set. Similarly to Pick *et al.* (2010) and Philippe *et al.* (2009), I did not perform analyses to evaluate whether CAT fits the data better than other models. This was because the fit of the CAT models can only be tested using Bayesian cross-validation, but this method is too computationally intense for a data set as large as the one used here. In any case, there is ample evidence that for large data sets the CAT-based models (including CAT) always fit the data better than any of the homogeneous time reversible models (like WAG, LG and GTR), making model testing somewhat redundant. The only model that was likely to fit the data better than CAT is CAT-GTR (see Phylobayes manual) but this model was computationally too expensive to be applied to my data set.

Phylogenetic analyses were performed using Phylobayes 3.3e (Lartillot *et al.* 2009). For all Phylobayes analyses 2 runs were performed and convergence was investigated using the bpcomp software (which is part of the Phylobayes package – see also Sperling *et al.* 2007; Sperling *et al.* 2009; Sperling *et al.* 2010; Campbell *et al.* 2011; Rota-Stabelli *et al.* 2011). For all analyses, among site rate variation was taken into consideration and modelled using a discrete Gamma distribution (4 rate categories). The Gamma distribution was preferred to a Dirichlet process to model among site rate variation, because convergence problems might arise, under Dirichlet in Phylobayes (see Phylobayes manual).

2.2.2 Dealing with Compositional Heterogeneity

Compositional heterogeneity can cause attraction artifacts that can sway phylogenetic analyses. Posterior Predictive Analysis (PPA; see Phylobayes manual) was used to evaluate whether the 146-NGs data set contained compositionally heterogeneous taxa. PPA identified several compositionally heterogeneous lineages (see appendix B). To ameliorate compositional problems and attempt alleviating potential compositional attractions the Dayhoff recoding was used. The 146-NGs data set was thus reanalysed, under CAT (same specifications reported above), after the data were recoded in the six Dayhoff categories (see also Stabelli *et al.* 2012).

Dayhoff recoding is well known to ease compositional problems, but can result in some signal erosion. To monitor whether signal-erosion had a substantial impact on the obtained results I monitored changes in support values by contrasting Bayesian Posterior Probabilities (PP) for corresponding nodes between the CAT and the Dayhoff-CAT tree.

2.2.3 Objective outgroup analysis versus "common sense" outgroup selection and outgroup ranking

The original data set of Philippe *et al.* (2009) included a total of eight outgroups: *Monosiga brevicollis*, *Monosiga ovata*, *Proterospongia sp.*, *Amoebidum parasiticum*, *Sphaeroforma artica*, *Capsaspora owczarzaki, Saccaromices cerevisiae and Cryptococcus neoformans.* In the new data set assembled here, the number of outgroup taxa was left unchanged. What was changed was the gene-coverage for two key taxa that were under-sampled in the data set of Philippe *et al.* (2009): *Monosiga ovata* and *Proterospongia sp.*

To investigate the effect of outgroup selection on phylogenetic results, analyses were performed to rank these outgroups. It is often considered "common sense" to use the phylogenetically closest outgroup to root a tree. However, phylogenetic proximity does not necessarily correspond to phylogenetic optimality. In many cases using members of the closest

outgroup might be a sensible idea (or even the only viable option). However, there are conditions in which such a choice can be counterproductive. As pointed out by Lyons-Weiler *et al.* (1998) (see also Holton and Pisani 2010), the closest outgroup is not necessarily the most adequate choice when its rate of evolution is greater than that of other available outgroups (Figure 2.1). Similarly, as Rota-Stabelli and Telford (2008) pointed out in a very compelling way, compositional heterogeneity and skews in amino acid usage patterns should also be considered when selecting outgroups for phylogenetic analyses.

Following Rota-Stabelli and Telford (2008) I explicitly analysed the quality of the considered outgroups and their potential biasing strength. The eight potential outgroups were ranked according to (1) the Z-score value from the PPA (this will allow selecting taxa with optimal composition), (2) their average pairwise genetic distance from the ingroup taxa (to select slowly evolving taxa), (3) their average pairwise compositional distance from the ingroup taxa (to further identify taxa that could cause compositional attractions) and (4) their amount of missing data (to take into consideration potential missing-data-induced LBA artifacts). Compositional and genetic distances were calculated using MEGA 5 (Tamura *et al.* 2011).

Outgroup ranking (see Table 2.1) was used to inform a series of taxon subsampling experiments. Accordingly, a series of independent analyses were performed using only the following outgroups: (1) *Monosiga ovata* and *Proterospongia sp.* (the two best Choanoflagellate outgroups). (2) *Amoebidium*, *Proterospongia sp.* and *Monosiga ovata* (the three best outgroups). (3) *Monosiga ovata* and *Monosiga brevicollis* (the two worst choanoflagellate outgroups), (4) *Monosiga brevicollis* and *Proterospongia sp.* (the worst and the best among chonaflagellates outgroups).

Analyses were also performed using “common sense” selected outgroups. This was done to compare results obtained using the “common sense” approach with results obtained using the rigorous outgroup selection approach. As examples of the “common sense” outgroup selection strategy two data sets were generated and analysed. The first excluded the Fungi (i.e. the

phylogenetically more distant outgroups), and the second excluded the Fungi, *Sphaeroforma*, *Amoebidum* and *Capsaspora* (i.e. the only considered phylogenetically closest outgroup: the Choanoflagellata). For all considered sets of outgroups phylogenetic analyses were performed under CAT and CAT with Dayhoff recoding (same specifications used above).

2.3 Results

2.3.1 Standard phylogenetic analysis & "common sense" outgroup selection

Figure 2.2 summarizes the result of the analysis performed using all outgroups. This analysis returns a tree where, *contra* Philippe *et al.* (2009), the Porifera and the Coelenterata are not monophyletic. More precisely, the calcarean sponges are recovered as the sister group of all the other metazoans (PP=0.6), whilst the Ctenophora are recovered as the sister group of all the other metazoans but the calcarean sponges (PP=0.94). Silicea (Demospongiae plus Hexactinellida) is found to be monophyletic (PP =1) and the Placozoa are recovered as the sister group of these two taxa (an unexpected result), but with low posterior probability PP=0.45. Figure 2.3 shows results of analyses where the Fungi are excluded (the first "common sense" data set). The tree recovered from this analysis is identical to that of Figure 2.2 (i.e. the one recovered using all the outgroups). However, inclusion of the distantly related fungi in the analysis of Figure 2.2 seems to have an impact on the support levels for the relationships of the Ctenophora, which drops from PP=0.94 to PP=0.56. This suggests that the inclusion of the distantly related Fungi might have participated in causing an attraction of the Ctenophora toward the base of the tree. However, as the position of the Ctenophora is unchanged when the Fungi are excluded they seem to be only a minor player in the definition of the topology in Figure 2.2.

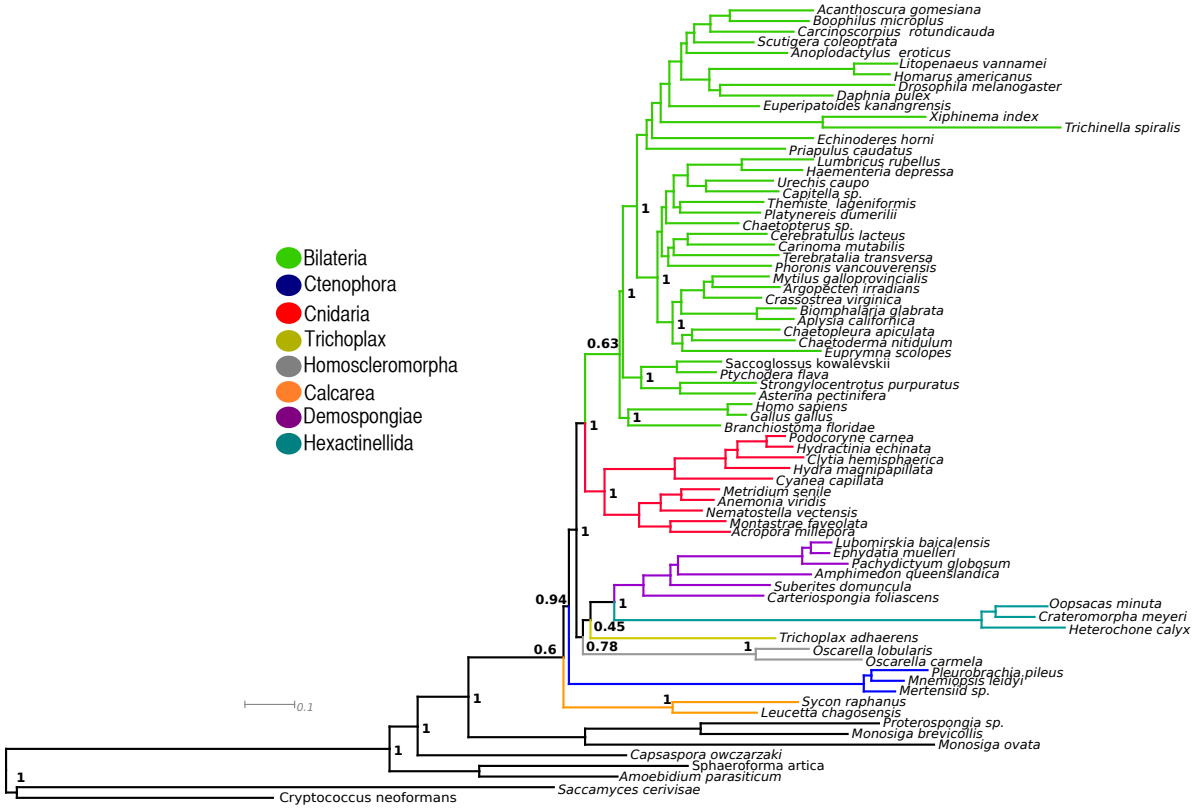


Figure 2.2: Bayesian analysis of 146-NGs data set with all the outgroups under CAT+Γ model.

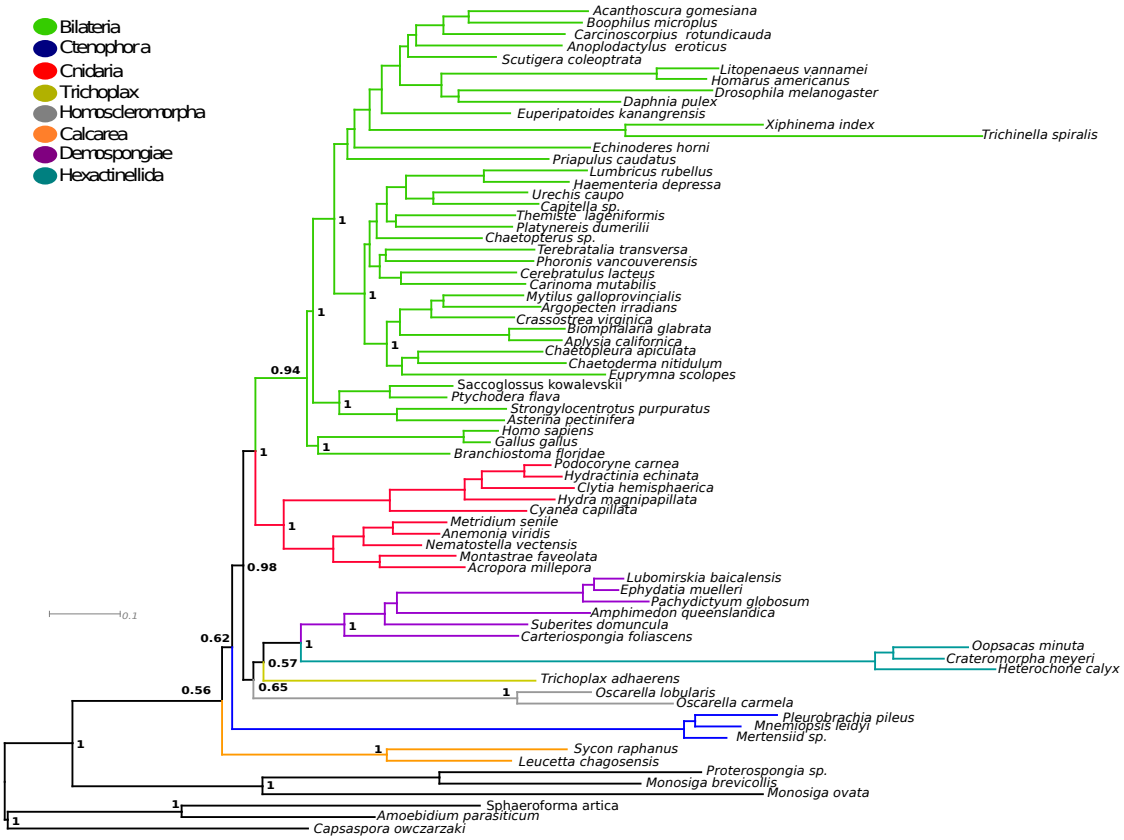


Figure 2.3: Bayesian analysis of 146-NGs data set with all the outgroups under CAT+Γ model, excluding Fungi.

When the Choanoflagellata (the phylogenetically closest outgroup, and the outgroup that was considered to be best by Philippe *et al.* (2009) is used as the sole outgroup for phylogenetic analyses the tree in Figure 2.4 is recovered. The topology of Figure 2.4 is consistent with the sponge monophyly hypothesis (even though it suggests that the Placozoa are also members of the "Porifera"). This shows that, exclusion of all the non-choanoflagellate outgroups had a significant effect on the position of the Calcarea, which is now no longer at the root of the tree. However, support for the Porifera + Placozoa group is insignificant (PP = 0.18). The Ctenophora are still placed toward the root of the tree and the crown-ward movement of the Calcarea as left them as the sister group of all the remaining metazoan (PP = 0.57). The support for a root-ward position of the Ctenophora does not change between Figure 2.3 (PP = 0.57) and Figure 2.4 (PP = 0.57). This suggests that the removal of *Capsaspora, Amoebidum* and *Sphaeroforma*, despite having a strong and seemingly beneficial effect on the position of Calcarea, which is now recovered as the sister group of the Homoscleromorpha with PP = 0.65 (a result previously reported by Philippe *et al.* 2009; Pick *et al.* 2010; Erwin *et al.* 2011), was invariant to the position of Ctenophora. With reference to the Placozoa, it can be noted that no significant change in support is observed when removing the Fungi, *Capsaspora, Amoebidum* and *Sphaeroforma*. This suggests that none of these taxa seem to be responsible for the placements of Ctenophora in Figures 2.2 to 2.4.

From this initial series of experiments it is obvious that one must conclude that serial removal of outgroups under a "common sense scheme" does not allow resolving the Metazoan relationships, even though removal of Fungi plus *Capsaspora, Amoebidum* and *Sphaeroforma* seems to alleviate attraction artifacts affecting the calcareans.

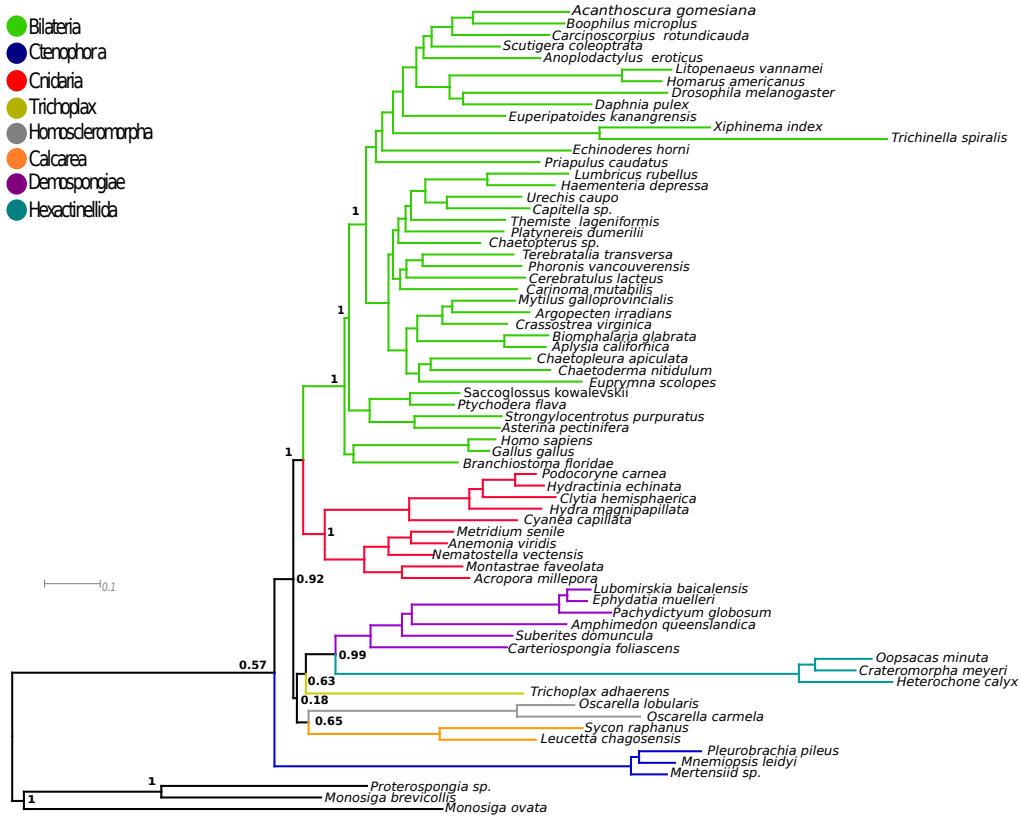


Figure 2.4: Bayesian analysis of 146-NGs data set using the closest outgroups under CAT+Γ model.

2.3.2 Compositional heterogeneity and its effect on "common sense" phylogenies

The PPA suggests that compositional heterogeneity affects the outgroups (see table 2.1). It is important to note that a certain amount of composition heterogeneity is present in every data set and that it can potentially affect the topology by causing the groupings of unrelated taxa. To assess the effects of compositional heterogeneity, the "common sense" data sets of Figs. 2.2 to 2.4 were re-analysed using Dayhoff recoding (see methods). The results suggest that the phylogenetic position of the calcarean sponges in Figure 2.2 (i.e. as the sister group of all the remaining metazoans) does not seem to be the result of a compositional attraction.

On the other hand low support values (Figure 2.5a, b and c) suggest that either Dayhoff recoding is causing signal erosion or that some support for the topology in Figs. 2.2 to 2.4 represent a compositional bias. In any case, substantial topological changes can be observed with reference to the Ctenophora that in figure 2.5a and 2.5c are found as the sister group of Cnidaria and thus as member of the Coelenterata. In particular in figure 2.4c this result is associated with a relatively high support PP = 0.74 (ruling out a signal-erosion effect at the least for the position of this taxon). With reference to the sponges, Dayhoff analyses found variable topologies all of which are poorly supported (suggesting that signal erosion might be a problem with reference to these taxa). The topology of figure 2.5c, in addition to finding relatively high support for Coelenterata is also consistent with the Epitheliozoa hypothesis as it shows the Homoscleromorpha to be closer to the Eumetazoa than the other sponges are, and the Silicea as the sister group of all the other Metazoa (see Sperling *et al.* 2009). In any case, it is clear that also in the case of the Dayhoff recoding analyses; the three "common sense data sets" cannot resolve the metazoan relationships with confidence.

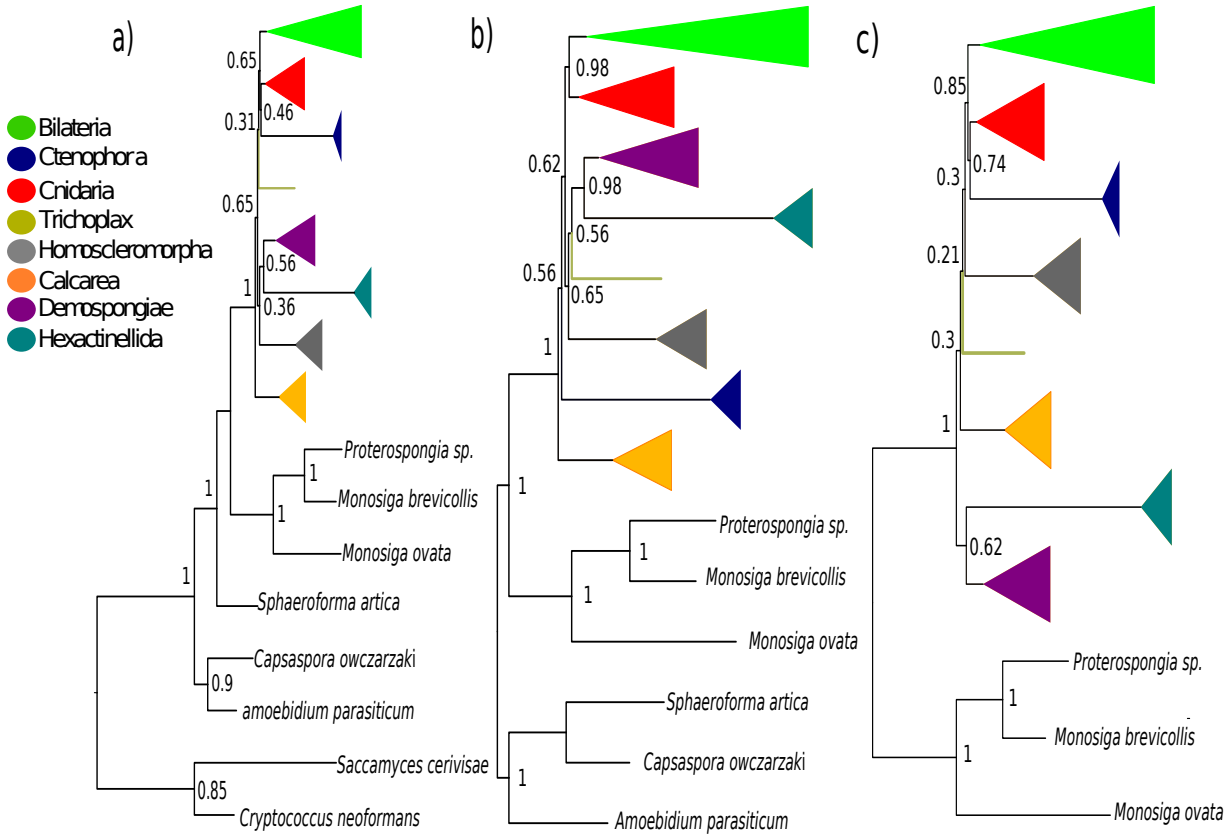


Figure 2.5: Bayesian Dayhoff recoding analysis under CAT+Γ (a) Dayhoff recoding all outgroup data set (b) Dayhoff recoding all out-group but fungi (c) Dayhoff recoding choanoflagellates data set.

2.3.3 Objective outgroup analysis & outgroup ranking.

Table 2.1 summarizes the results of the outgroup analyses, and of the outgroup ranking. It is evident that the 8 potential outgroups do not have the same compositional profile, and show different average pairwise genetic distances to the ingroup. It is also important to note that they do not have the same amount of missing data and this to some extent can affect the analyses performed here by masking potential, compositional problems. None of the considered outgroups are compositionally homogeneous with reference to the ingroups. Interestingly two choanoflagellates (the two *Monosiga* species) that *a priori* should be excellent outgroups (phylogenetically closest) are highly heterogeneous and have quite high genetic distances from the ingroup (see table 2.1). From a compositional point of view the optimal out-group is *Amoebidum*, which is relatively distantly related to the Metazoa. However, *Amoebidum* is the taxon with the highest degree of missing data. From a compositional point of view, the best choanoflagellate outgroup is *Proterospongia sp*. (Table. 2.1), and despite *Sphaeroforma* (another non-Choanoflagellata) is more heterogeneous than *Proterospongia sp*., it still is more homogeneous than the *Monosiga* species. Pinpointing the potentially scarce value of the two *Monosiga* species as outgroups for this data set. From an inspection of Table 2.1 it is also evident that despite *Monosiga brevicollis* and *ovata* having very similar PPA Z-scores and genetic distances, the average *Monosiga brevicollis* compositional distance from the ingroups is far higher than that of *Monosiga ovata* making the latter a better outgroup (despite his higher amount of missing data). Overall, given these results, I moved forward to carry out analyses in which outgroup taxa were subsampled with reference to their objective qualities.

Species	Taxonomy	Compostional heterogenity (z-max from ppred)	Genetic distance	Compositional distance	% Missing data
Proterospongia sp.	*Choanoflagellate*	6.152	29.142	137.598	19%
Monosiga brevicollis	*Choanoflagellate*	10.953	29.357	189.94	6%
Monosiga ovata	*Choanoflagellate*	10.123	29.508	123.675	62%
Sphaeroforma artica	*Ichthyosporea*	7.052	29.617	94.837	51%
Amoebidium parasiticum	*Ichthyosporea*	3.505	26.762	67.312	72%
Capsaspora owczarzaki	*Filasterea*	10.369	27.341	193.599	7%
Cryptococcus neoformans	*Fungi*	10.02	33.493	228.564	6%
Saccharomyces cerevisae	*Fungi*	4.646	35.384	209.789	7%

Table 2.1: This table illustrates the statistics used to rank the outgroups. Yellow: Best outgroup. Green: Second best outgroup (but has high level of missing data). Purple: Third best outgroup (less missing data but high genetic distance). Blue: Fourth best outgroup. Orange: Fifth best outgroup (worst of the Choanoflagellata). Grey: Poor fungal outgroups.

Analyses performed based on the subsampling of outgroups on the grounds of their properties in Table 2.1 are intriguing. Despite the fact that the two *Monosiga* species show the same Z-score values, they are characterized by different compositional distances, with *Monosiga ovata* having a shorter branch than *Monosiga brevicollis*. When *Proterospongia sp*. (the less compositional heterogeneous outgroup) is used in combination with *Monosiga brevicollis* (i.e. the best and worst choanoflagellates are used in combination) the results support, albeit with a low confidence (PP = 0.24), the monophyly of sponges (see Figure 2.6a). In this analysis the ctenophores are still in the same position in which (Dunn *et al.* 2008) found them. That is, they are found as the sister group of all the other animals (PP = 0.54). Interestingly, by improving the compositional profile of the outgroups (i.e. using *Proterospongia sp.* and *Monosiga ovata*) the sponges become paraphyletic (PP = 0.98; Figure 2.6b). Additionally, improving the outgroups also causes the ctenophores to shift their position, and in Figure 2.6b they appear as the sister group of the Cnidarians in a monophyletic Coelenterata (PP = 0.81). Analyses performed using *Monosiga ovata* and *Monosiga brevicollis* (Figure 2.6c) find the Ctenophores as sister group of Cnidaria+Bilateria (PP = 0.81 – as in Pick *et al.* 2010).

When the analysis is performed using the three outgroups with the best compositional profile (i.e. *Amoebidum*, *Proterospongia sp*. and *Monosiga ovata*) the calcarean sponges move at the root of the tree. It has recently been shown that gap-rich taxa can increase long-branch attraction artefact (Roure *et al.* 2012) and this result can be explained as a LBA artefact caused by the inclusion of the gap-rich (72% of missing data) *Amoebidum*.

An important aspect of the results of Figure 2.6 is that the placozoans appear to be unstable. Both sources of systematic error considered here seems to affect the position of the placozoan. However under the best phylogenetic conditions, when the compositional skew among the outgroups is minimized (i.e. *Monosiga ovata* and *Proterospongia sp.* are used) the placozoa is the sister group of the Neuralia plus Calcarea plus Homoscleromorpha group.

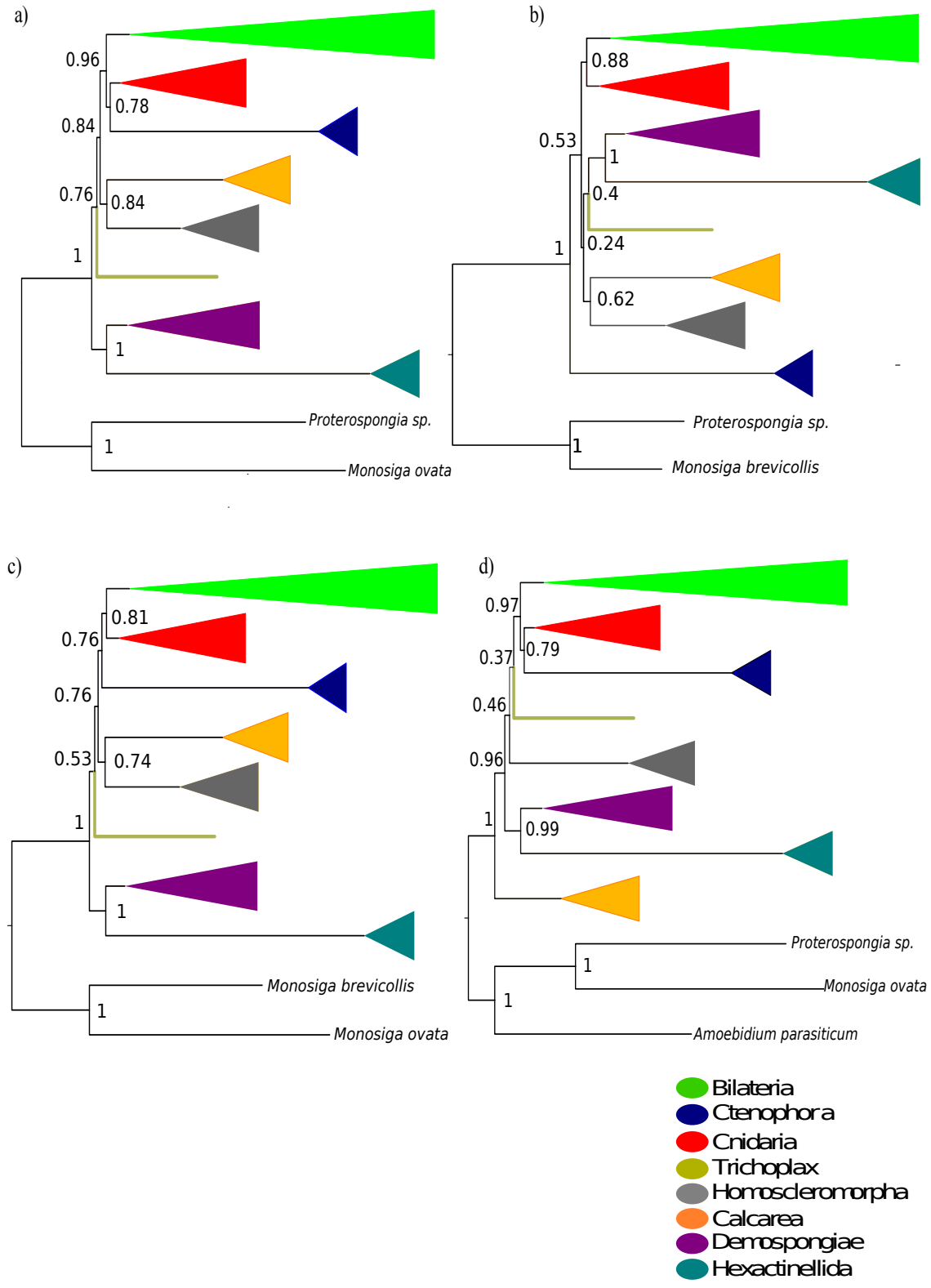


Figure 2.6: Phylogenetic relationship performed on with a sub-sample of the out-groups using CAT+Γ model. (a) *Monosiga ovata* and *Proterospongia sp.* (the bests among the Choanoflagellates) (b) *Monosiga ovata* and *Monosiga brevicollis* (two worst two choanoflagellates), (c) Only *Monosiga brevicollis* and *Proterospongia sp*. (the worst and the best among choanoflagellates). (d) *Amoebidium*, *Proterospongia sp.* and *Monosiga ovata* (three best out-groups)

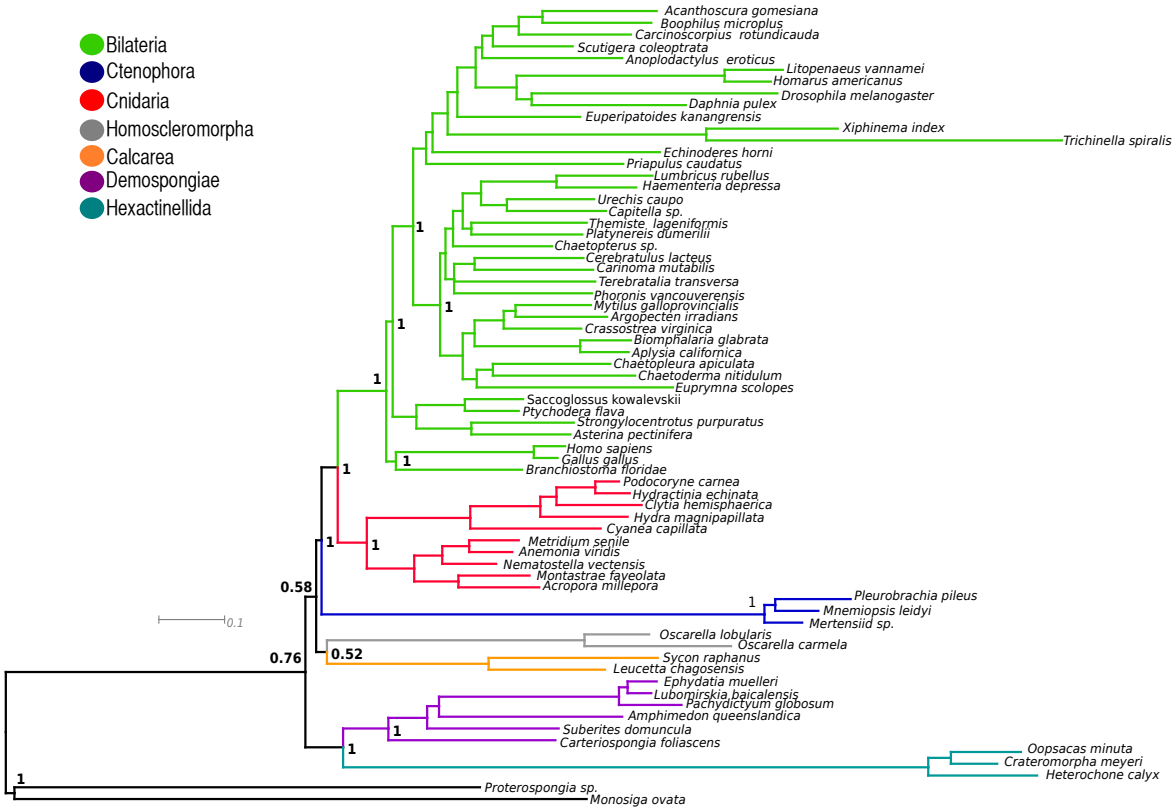


Figure 2.7: Phylogenetic analysis performed under CAT+Γ model using the two best outgroups where the Placozoa were excluded from the analyses.

When the three best outgroup are used (see figure 2.6d) Placozoa are found as the sister group of *Neuralia (*see table 1.2 as in Philippe *et al.* 2009; Sperling *et al.* 2009). As here the focus is not the phylogenetic position of the Placozoa, analyses were performed in which this taxon was considered a "nuisance factor" and was thus excluded – "marginalised". These analyses were performed using the two best choanoflagellate outgroups only *Proterospongia sp.* and *Monosiga ovata*. Results are reported in figure 2.7 do not topological changes when compared with figure 2.6b. However, support for Epitheliozoa decrease from PP = 0.84 to PP = 0.72. In figure 2.7 Coelenterata are not supported, instead, Ctenophora are the sister group of Cnidaria + Bilateria.

This analysis confirms that the relationships among the sponge taxa are invariant to the inclusion/exclusion of Placozoa from the analyses.

2.4 Discussion

The first important result of this chapter is that the data set considered here represent an improvement over that of (Philippe *et al.* 2009), particularly as it includes more data for key outgroups and for a key ingroup taxon (the sponge *Oscarella carmela*), and it does not find any support for the monophyly of the sponges. Furthermore, the results presented here suggest that compositional heterogeneity and outgroup selection are substantially underestimated issues in the study of metazoan evolution.

Excluding the work of Hejnol *et al.* (2009), which seems to be problematic in terms of missing data (Roure *et al.* 2012), this is the first study of a large-scale (EST) data set that supports the paraphyly of the sponges. Two methodological improvements implemented in this work could explain the differences between the results presented here and those of previous EST-analyses (Dunn *et al.* 2008; Philippe *et al.* 2009; Pick *et al.* 2010). First, the selection of the orthologous genes in this work has been performed using a rigorous procedure involving gene-

tree reconstruction and manual inspection of all the gene trees. All previous studies relied on automated approaches that did not properly identifying potential problems at the gene-phylogeny level. Second, the software used for my final gene alignments (Loytynoja and Goldman 2008) provides a better description of the evolutionary process when compared with software used in previous studies. It is clear that a rigorous selection of orthologous genes, and a more reliable alignment can substantially change the results obtained from a phylogenetic analysis.

A more general consideration can be drawn from the point of view of circumventing systematic errors in phylogenies. Indeed, as argued by (Rota-Stabelli and Telford 2008) the results presented here suggest that choosing a set of outgroups for phylogenetic analyses based only on prior phylogenetic knowledge can be problematic. Firstly, prior knowledge can be misleading (if phylogenetic relationships among the outgroups are uncertain it might be impossible to identify the closest one). Secondly (and most importantly) the phylogenetically closest outgroup is not necessarily the ideal outgroup to be used with reference to the ingroup because of lineage specific factors. As I show here for the choanoflagellates, it is possible that at the least some of the phylogenetically closest outgroups might not have the same compositional profile of the ingroup taxa, and this can potentially affect both the ingroup topology and the support level observed.

The results presented here suggest that *Monosiga brevicollis* (because of its composition) is unlikely to be a good outgroup to study metazoan evolution. A similar conclusion can be reached for *Monosiga ovata*. Among the choanoflagellates considered in this analysis the most adequate outgroup to study metazoan evolution (with reference to its composition) is *Proterospongia sp*. Another less closely related outgroup with a good compositional profile is *Amoebidum*. However, despite the good compositional profile, this taxon has the highest amount of missing data among the considered outgroups and this can affect phylogenetic results negatively (Roure *et al.* 2012). The results presented here also suggest that "common sense", *a priori*, outgroup choice is potentially misleading and rigorous outgroup analyses should be

routinely performed in phylogenetics. In the specific case of this data set, "common sense" based outgroup choice was shown to be a particularly inefficient way to try to analyse the data and to reach a coherent and acceptable conclusion (i.e. recovering a tree supporting one of the proposed, alternative hypotheses – sponge monophyly or paraphyly).

From a more applied perspective, the topologies recovered in the analyses presented here seem to suggest that sponges are most likely a paraphyletic assemblage of taxa and that the Ctenophora are indeed the sister group of the Cnidaria in a monophyletic Coelenterata. Placozoa proved quite unstable but where never found to be more closely related to the Bilatera than the Coelenterata are (*contra* Pick *et al.* 2010), and whilst inclusion of the Placozoa in the analysis has an effect on the phylogenetic position of the Ctenophora, the presence of the Placozoa in the data set does not affect the resolution of the sponges, which from this point of view are thus robust.

2.5 Conclusions

A general conclusion that can be drawn from this chapter, in line with Philippe *et al.* (2011), is that phylogenomic-scale data sets might not be sufficient to solve the relationship among the non-bilaterian Metazoa. It might be necessary to use other sources of data (like microRNAs), as well as a thorough investigation of all possible biases that could affect the considered data. Indeed it is clear from the trees presented here that different sources of phylogenetic bias differently affect the phylogeny of the basal Metazoa, and rejecting one of the currently available alternatives might prove more difficult than previously thought. Indeed, even though the results here presented take us a long way forward toward gaining a better understanding of metazoan evolution, many problems still persists. We can state with confidence that Ctenophora are clearly not the sister group of all the other Metazoa and that this result, as presented in Dunn *et al.* (2008) and Hejnol *et al.* (2009) was thus caused by a tree

reconstruction artefact. Outgroup analysis suggests that sponge paraphyly is more likely to be correct than sponge monophyly (in agreement with Sperling *et al.* 2007; Sperling *et al.* 2009; Pick *et al.* 2010; Sperling *et al.* 2010; Erwin *et al.* 2011 contra Philippe *et al.* 2009), but further investigations and more data will be necessary to further validate the relationships of the sponge classes. Indeed, the problem of understanding the relationships among the non-bilaterian animals is far from resolved, and it will be so until one of the two alternative hypotheses (sponge monophyly and sponge paraphyly) will be strongly rejected by the data.

With reference to the work I will perform in other chapters of this thesis (study of the evolution of the GPCR protein superfamily in Metazoa), I shall assume Epitheliozoa (see table 1.1) as my working hypothesis as it is favoured by the analyses presented in this Chapter. However, it is clear that I am fully aware that my current results do not allow for a robust distinction of the two competing hypotheses (see above), as support values for key nodes are low in figure 2.6b and figure 2.7.

Chapter 3

Phylogenomics of 7TMD/GPCR receptors and the origin of the metazoan GPCRs

Abstract

Proteins with 7TMD are present in Archaeabacteria and Eubacteria and Eukaryotes and they are key elements in the relationship between intracellular environment and extracellular environment. The lack of genomes for key taxa (i.e. unicellular Eukaryotes) and a high level of divergence have hampered the reconstruction of the phylogenetic history of 7TMD receptor.

In this chapter I have analysed the distribution and phylogenetic relationship among proteins with 7TMD in 1214 genomes including Archaebacteria, Eubacteria and representative genomes from all the five supergroups of Eukaryotes. This broad genomic sample and newly methods for the phylogenetic reconstruct (i.e. phylogenomic network) clarify the early history of 7TMD.

The results presented in this chapter suggest 1) an expansion of the 7TMD and GPCRs in *Neuralia* lineage (see Table 1.2); 2) a multiplied independent evolution of the 7TMD architecture and 3) the possible existence of the GPCRs in the last eukaryotic common ancestor.

3.1 Introduction

The ability to respond to stimuli is a necessity for every cell, allowing them to grow, explore the surrounding environment, and communicate with other cells. This allows inner-module communication (between different cell-types, tissues and organs) in multicellular organisms. 7-trans-membrane domains receptors (7TMDs) constitute a large protein super-family, and mediate responses to stimuli in eukaryotes. These proteins are characterized by the presence of seven alpha helices, crossing the cell membrane seven times.

7TMDs are also present in Archaeabacteria and Eubacteria, where they are named proteorhodopsins and are functionally classified in two main categories: transporters and receptors (Sharma *et al.* 2006). Additionally, from an ecological prospective, the proteorhodopsins are key elements in the marine ecosystem, capturing and transforming solar energy (Fuhrman *et al.* 2008).

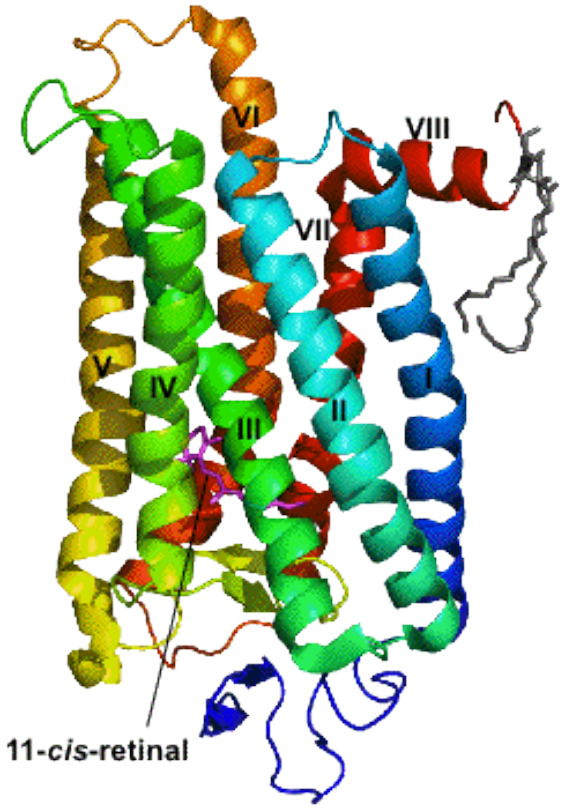


Figure 3.1: Three-dimensional structure of the bovine rhodopsin. This is the first 7TMD protein for which a christal structure was derived (Terakita 2005).

A particular type of 7TMD (named G-protein coupled receptors, see below) receptors are activated by a diverse array of ligands, and are involved in various signalling processes such as cell proliferation, neurotransmission, metabolism, smell, taste, and vision (Smith 2000). The presence of GPCRs in Metazoa, and more generally Unikonta and Cromoalveolata, is well established (Krishnan *et al.* 2012), while the presence of these proteins in plants is still debated (Devoto *et al.* 1999; Moriyama *et al.* 2006)

A common mechanism that characterizes both proteorhodopsin and GPCRs is that they undergo a conformational change in response to activation by an external agent. This process results in a cascade of chemical reactions, which affects the physiological condition and the transcriptional landscape of the cell (see figure 3.2 and Marinissen, Gutkind 2001).

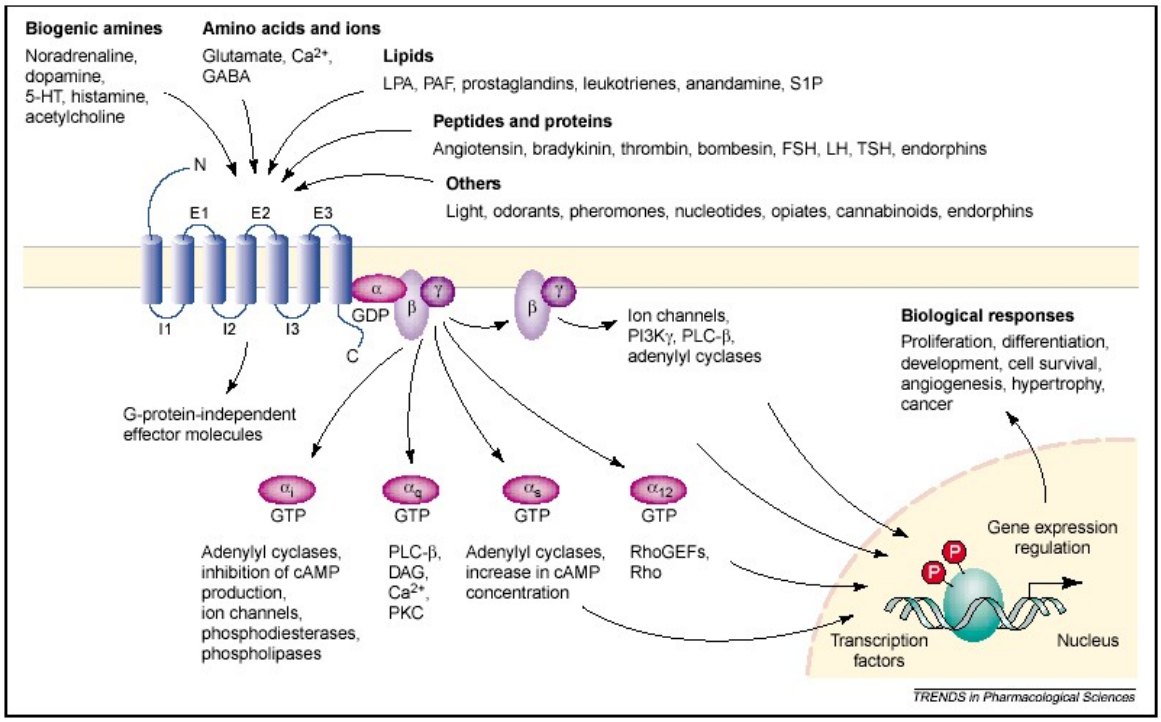


Figure 3.2: Schematic view of the GPCRs pathway (from Marinissen and Gutkind 2001).

Despite the presence of a common architecture, 7TMD receptors in prokaryotes and eukaryotes works differently. Proteorhodopsin mediates phototaxis by regulating cell motility using a two-component signalling cascade (Klare *et al.* 2004). Unlike proteorhodopsins, the

majority of the 7TMDs receptors in eukaryotes use a G-protein system for the signal transduction (because of this, they are defined G-protein coupled receptors). This signalling system has a modular design consisting of a receptor, a heterotrimeric G protein, and an effector (Wettschureck and Offermanns 2005). The relatively complex organization of the GPCRs signaling system provides the basis for a huge variety of transmembrane signalling pathways that are tailored to serve particular functions in distinct cell types. Although the majority of 7TMDs receptors in eukaryotes are G-coupled, there are notable examples of proteins with the 7TMD architecture that do not rely on the G-protein signalling pathway (e.g the insects Olfactory Receptor - ORs (Kaupp 2010).

The relationship between different 7TMDs receptors is only structural (i.e. they share a common architecture but no sequence similarity) and it is thus unclear whether these proteins are phylogenetically related (Soppa 1994). To explain the origin of the GPCRs from bacterial rhodopsins two different hypotheses have been proposed. Given that the highest sequence similarity between GPCRs and proteorhodopsins is in non-homologous helices, some authors have suggested that they are related via an evolutionary mechanism that involves exon shuffling (Pardo *et al.* 1992). An alternative hypothesis proposes that gene duplication of an ancestral three-transmembrane module gave rise to helices 1 through 3 and 5 through 7 (Taylor and Agarwal 1993). However, Larusso *et al.* (2008) showed that the animal opsins (and hence GPCRs more generally) do not appear to have originated through an internal domain duplication event. The work of Larusso *et al.* (2008) provides further evidence that the animal opsins are non-homologous, indicating a convergent evolutionary origin, in which both groups of opsins evolved a seven-TM structure and light sensitivity independently.

A striking feature of GPCR evolution is their highly dynamic repertoire in eukaryotic organisms (Nordstrom *et al.* 2011; Krishnan *et al.* 2012). Current data suggest the presence of GPCRs in Chromoalveolata, Unikonta (Nordstrom *et al.* 2011; Krishnan *et al.* 2012) and probably in plants. These results are difficult to explain in light of the most likely, among the

alternative rooting positions proposed for the eukaryotes (between Unikont-Bikont; see table 1.1 and Derelle and Lang 2012; Baldauf 2003; Stechmann and Cavalier-Smith 2002; Richards and Cavalier-Smith 2005). This is mostly because current evidence on the distribution of GPCRs is patchy, and a comprehensive analysis of the distribution of the GPCRs in eukaryotes is still lacking. In other words, as suggested by (Strotmann *et al.* 2011): the answer to how the 7TM core of the eukaryotic GPCRs has evolved still needs elucidation.

The GPCR content of the basal metazoans (the main subject of this thesis) has been systematically investigated in various genomic papers (see introduction). However, a comprehensive analysis of GPCR evolution, with reference to the origin and early evolution of the animals, is still lacking. In addition, it is clear that as more data are being accumulated, and better methods devised, further re-analysis of the available evidence might improve our understanding of both GPCR evolution in animals and more broadly in Eukaryota.

In this chapter, I present an analysis of 7TMD evolution across the three domains of life. The aim of this study is to better understand the relationship among the several 7TMD receptors, clarify GPCRs evolution within eukaryotes, and elucidate the evolution of this protein family with reference to the origin of animals. To investigate GPCR evolution in basal Metazoa, I obtained genomic data for two new sponge lineages, the Calcarea and the Homoscleromorpha (see table 1.2). These data were used to supplement publicly available databases. This allowed me to have a genomic-scale representation of the GPCR repertoire in all the basal metazoan lineages except the Ctenophora. Dr. Scott Nichols and Prof. Nicole King kindly provided sequence data for the homoscleromorph *Oscarella carmela* (Nichols *et al.* 2012), while the unpublished genome of the calcarean sponge *Sycon sp*. was provided by Dr. Maja Adamska.

In addition to using novel, genomic-scale data sets, I have used new, network-based approach to study GPCR evolution. Major problems that hampered previous investigations of the evolution of this protein family include low levels of sequence similarity, and problems of positional homology. It is important to underline (as it has already been done in the

introduction) that alignment errors result in incorrect phylogenetic trees. Limited sequence homology between distant paralogs can introduce substantial errors in alignments and phylogenies, which might result in incorrect evolutionary reconstructions. In order to avoid problems caused by the unreliability of alignments generated for sets of distantly related sequences, the relationship between the GPCRs were here reconstructed using phylogenomic networks (Atkinson *et al.* 2009; Dagan 2011).

Phylogenomic networks are useful to overcome problems related to the complexity of molecular evolution. Indeed, they allow for the identification of non-tree like processes (i.e. protein fusion, horizontal gene transfers, and domain shufflings). Networks do not rely on a global alignment to infer potential sets of relationships, thus substantially reducing errors caused by alignment misspecification (Wong *et al.* 2008).

In a similarity network, the nodes of the network represent sequences, and relationships are represented by the edges joining these nodes. An edge will be drawn to join two nodes if a pairwise alignment of significant level (i.e. a significant BLAST hit) was obtained between the two considered sequences. Similarity networks are typically composed of multiple connected components (CCs), each of which comprises a number of nodes that share similarity relationships with elements within the CC, but not with genes outside the CC. These CCs represent groups of directly or indirectly related sequences, without the requirement that all sequences exhibit a detectable similarity to each other. Accordingly, these CCs represent an extension of the classic gene families (Bapteste *et al.* 2012). For example, within a network framework, we can think of a three-gene CC with the topology "A-B-C". In such a CC, A exhibits detectable similarity to B, and B exhibits detectable similarity to C, but no significant similarity can be detected between A and C, e.g. as a result of a high degree of divergence.

In this thesis the integration of a denser taxon sampling, and new types of phylogenetic methods, has allowed for a clarification of crucial aspects of GPCR evolution. First, the results presented here suggest that the proteorhodopsin are not related to the eukaryotic GPCRs. Rather,

the 7TMD architecture has been explored several times independently in the tree domain of life. Furthermore, my results extend the previous finding of (Krishnan *et al.* 2012) suggesting the presence of animal-like GPCRs in Rhizaria, Excavata and probably in plants (see table 1.1). This result, in the light of alternative rooting positions suggested for the Eukaryotes, implies that these receptors were a component of the genetic tool kit of the last-eukaryotic common ancestor.

3.2 Material and Methods

With the aim of clarifying the origin and deep time history of the GPCR superfamily and the relationship among proteins with a 7TMD architecture, I sampled 7TMD/GPCR receptors from genomes belonging to the three domains of life. Contrary to every previous study (Nordstrom *et al.* 2011; Krishnan *et al.* 2012), the genome sampling used here includes representatives from the entire set of eukaryotic super-groups, and a large sample of prokaryotic 7TMD from both Archaebacteria and Eubacteria (see below), as well as all the receptors in PFAM with a 7TMD architecture. Furthermore, this work presents the first phylogenetic analysis of the GPCR/7TMD repertoire of three sponge genomes: that of the demosponge *Amphimedon queenslandica* (Srivastava *et al.* 2010), the homoscherlomorph *Oscarella carmela* (Nichols *et al.* 2012) and the unpublished genome the calcarean sponge *Sycon sp*.

Protein coding sequences for the three sponge genomes were predicted using the software Augustus (Stanke *et al.* 2008), with parameters trained on *Amphimedon queenslandica.* The number of putative protein coding genes identified was 28,831 for *Sycon sp*., 33,045 for *Amphimedon queenslandica* and 14,679 *Oscarella carmela*.

In total, I analysed 20 plant genomes, 22 unikont genomes (including 10 Metazoan), 5 excavate genomes, 10 chromalveolate and the only rhizarian genome currently available, that of *Bigelowiella natans.* The total number of sequences in database was thus 1,351,617 (see appendix B). In addition, I included in the analysis 1,074 eubacterial and 82 archaebacterial

sequences (3,792,506 sequences in total) that were provided by David Alvarez-Ponce (Alvarez-Ponce and McInerney 2011 and appendix c and table 1 electronic appendix).

3.2.1 Data mining

7TMD and GPCR receptors are characterized by a low level of sequence similarity and BLASTP would fail to identify distantly related homologs. Therefore, I used PSI-BLAST (see Chapter 1) as the primary data-mining tool. This method is more sensitive, and better suited for identifying distant homologues because it uses a sequence profile, which is built from a multiple alignment of homologous sequences, and contains more information about the sequence family being considered than a single sequence does. The profile allows one to distinguish between conserved positions that are important for defining members of the family, and non-conserved positions that are variable among the members of the family. Moreover, it describes exactly what variation in amino acids is possible at each position by recording the probability for the occurrence of each amino acid along the multiple sequence alignment (Soding 2005).

To identify putative GPCR homologues, a series of PSI-BLAST searches (Altschul *et al.* 1997) were performed. PFAM alignments of protein families with a 7-transmembrane domain (7TMD; CL0192-GPCRS_A; CL0176-Chemosensory 7tm receptor; MLO-receptor-PF03094; ABA-GPCRS-PF12430 receptor) were downloaded and used to seed searches performed against the considered 60 complete genomes (see appendix B). Sequences with e-values below 10^{-6} were retained as putative 7TMD homologues, and merged in a single database from which redundancy was eliminated using Cd-hit (Fu *et al.* 2012). This program was used to identify subsets of sequences with 100% identity, and eliminate all but one of them. For the retained sequences, secondary structure prediction was carried out using Phobius (Kall *et al.* 2004), and proteins with 7 trans-membrane domains were retained as likely 7TMD homologues.

3.2.2 Phylogenetic networks

2,589 proteins featured in my final dataset; 2,408 of these were of eukaryotic origin, 30

of archeabacterial origin and 151 of eubacterial origin (see figure 3.3 for details). These sequences were merged together and an all-versus-all BLAST search was performed. Two thresholds (10^{-5}, 10^{-10}) were used to construct similarity networks from the results of the BLAST analysis. Because results using the 10^{-5} threshold level could generate many false positives (i.e. too many connections between phylogenetically unrelated groups and proteins), here only results obtained with a 10^{-10} threshold will be presented. To make sense of the complexity of the generated networks, a variety of colouring schemes were applied. First, a general colouring scheme was used, where only the eukaryotic supergroups (Excavata, Plantae, Unikonta, Rhizaria and Chromalveolata), and the Metazoa within the Unikonta, were identified. After that, a second scheme was applied which allowed a specific focus on the Unikonta. This scheme represented all the non-unikonts in one single colour (black), but distinguished all the key groups within Unikonta (e.g. Fungi, Choanoflagellata Amoebozoa etc.). In addition, each basal-metazoan species considered (*Nematostella, Hydra, Sycon, Amphimedon, Oscarella, Trichoplax*) was colour coded and thus identified. Finally, a third colouring scheme was applied where, as in the second scheme, only the unikonts where identified. However, within Metazoa all the Cnidaria were represented using one single colour. In addition, in this analysis, sequences of archaebacterial and eubacterial origin where also highlighted to identify possible inter-domain lateral gene transfers within Unikonta.

To investigate whether the 7TMD domain is evolutionary related to other transmembrane domains (with 2, 3, 4, 5, and 6 transmembrane helices), a network was built that also included proteins with less than seven alpha helices. This network was built imposing Blast e-value of 10^{-10}. Also in this case, colouring schemes were applied to visualize the distribution of 2-6 TMD proteins with reference to the 7TMD proteins. Because proteins with less than 7TMD might represent incompletely sequenced 7MD, two visualisations were performed. Initially, all the proteins with 2 to 6 TMD were visualized. Subsequently, a second visualisation was carried out in which proteins with 5 and 6 TMD were assigned a different colour.

In addition to these analyses, that used all the sequences having a significant level of similarity (i.e. e-value < 10^{-10}), we performed a series of more stringent analyses to evaluate the robustness of the inferred results. Accordingly, networks were generated where only connections between proteins with at the least 30%, 40% and 50% sequence identity (and a minimum 10^{-10} blast hit) were retained. This series of analyses were performed to evaluate whether GPCRs in taxa belonging to eukaryotic supergroups where these proteins are rarely found (e.g. Excavata – more below) represent ancestral eukaryotic GPCRs, or more recently (Lateral Gene Transfer - LGT) acquired ones. The rationale underlying this analysis is that if these sequences were of ancestral origin, one would expect the branch connecting them to the including CC to disappear when stringency increase. Alternatively, if these proteins were acquired via recent LGTs, one would expect the branches connecting them to their included CC to be retained when stringency was increased.

The networks were visualized using Cytoscape (Smoot *et al.* 2011), using the organic layout. This layout uses only node connectivity to illustrate groups and inter-group relationships (Atkinson *et al.* 2009), and is therefore suitable for visualizing threshold sequence similarity networks where the high-dimensional graph is defined by all the pairwise sequence alignments that are better than a chosen cut-off.

3.4 Results and Discussion

To my knowledge, the results here presented represent the first attempt to reconstruct the evolutionary origin of the 7TMDs/GPCRs across the three domains of life using non-tree based methods. The distribution of 7TMDs/GPCRs (see Figure 3.3) suggests the existence of GPCRs in Rhizaria, plants and (albeit in low numbers see below) in Excavata. This finding increases the resolution of the previous results of Krishnan *et al.* (2012) and Nordstrom *et al.* (2011). However for some of the protein analysed in this chapter, the association with a G-protein pathway (the condition for a 7TMDs receptor to be define a GPCRs) is unclear (see Figure 3.4b).

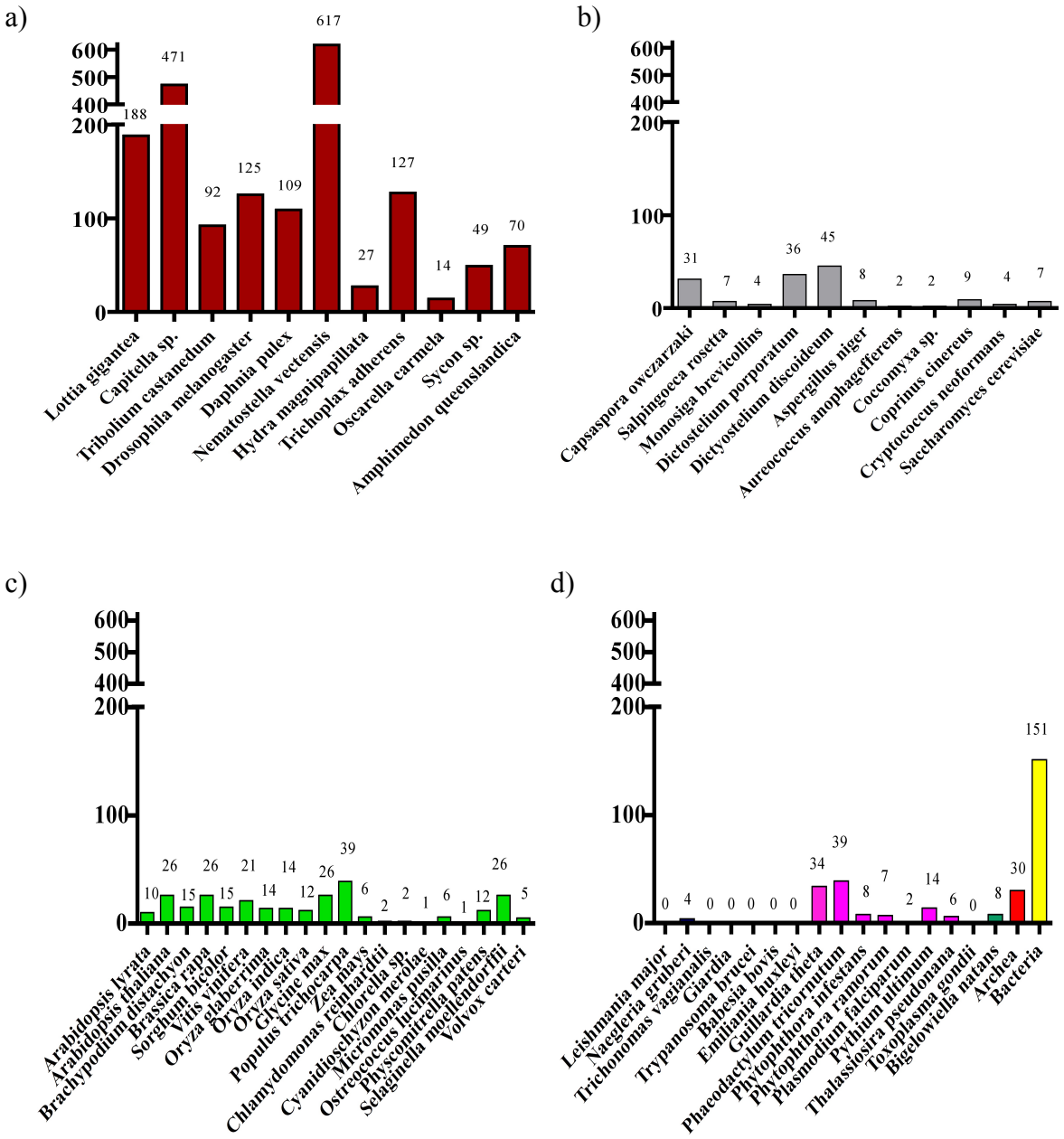


Figure 3.3: Distribution of 7TMD/GPCRs in the tree domains of life. (a) Animals (b) other unikonts (c) Plantae (d) Excavata, Chromolavealata, Rhizaria, Archaebacteria and Eubacteria

As expected the 7TMD/GPCR receptors vary largely between, and within, the eukaryotic supergroups. In animals, and particularly within *Neuralia* (*sensu* Nielsen 2012), a massive expansion of the GPCR repertoire is visible (Figure 3.3). On the other hand, the Fungi, and more generally the other unikonts (with the notable exception of the two considered amoebozoan species), seem to be characterized by small 7TMDs/GPCRs repertoires (Figure 3.3b).

Plants are also characterized by a 7TMD expansion (see figure 3.3c), with some of the plant-specific 7TMDs representing a lineage specific expansion (i.e. MLO-receptor, see Figure 3.4b). That is, they probably represent an independent evolution of the 7TDM. Instead, other plant 7TMD receptors seem to share homology with a pool of GPCRs that is common to all eukaryotes (see below and figure 3.4a). The analyses presented here also identified the presence of 7TMD proteins in Rhizaira. These include five glutamate receptors and several proteins that cluster with the Rhodopsins/Secretin/Frizzled/GPCR-1/cAMP group (Figure 3.4a). In Metazoa, glutamate receptors are expressed in the nervous system, the origin of which they substantially predate. Chromoalveolata have already been suggested by Krishnan *et al.* (2012) to possess eukaryotic 7TMD (see Fig 3.4a and 3b). In addition, *Guillardia theta* has proteorhodopsins that have been laterally transferred from the prokaryotes (Figure 3.4b). Not all chromoalveolates have the same number of GPCRs, with Toxoplasma and Plasmodium, which are endo-parasites, possessing only a few. The same conclusion seems to hold true for the Excavata. Indeed, the only excavate in which we could identify putative GPCR homologues was *Naegleria gruberi*: the only free living species among the considered ones. Archaebacteria and Eubacteria show that proteorhodopsins are characterized by high level of LGTs (see figure 3.4b). Additionally, Bacteria possess a lineage specific 7TMD receptor family (the bacterial ribonuclease) that does not have homologs outside this domain (Figure 3.4b).

Another interesting aspect of figure 3.4a and b, is that most of the considered proteins cluster in the same large CC. This is the Rhodopsin, Frizzled, Secretin, GPCR1/cAMP component (CC 1 in Figure 3.4a).

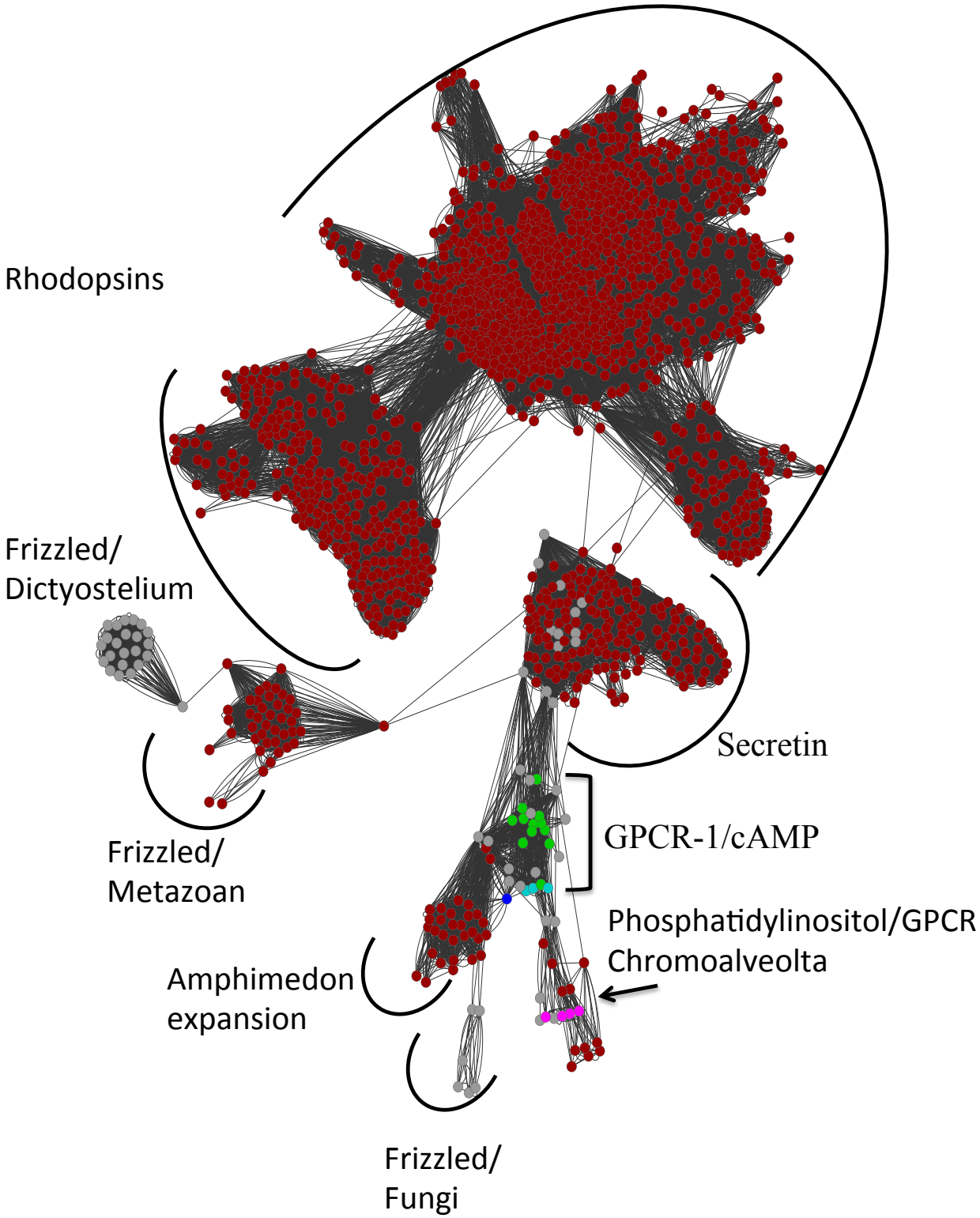
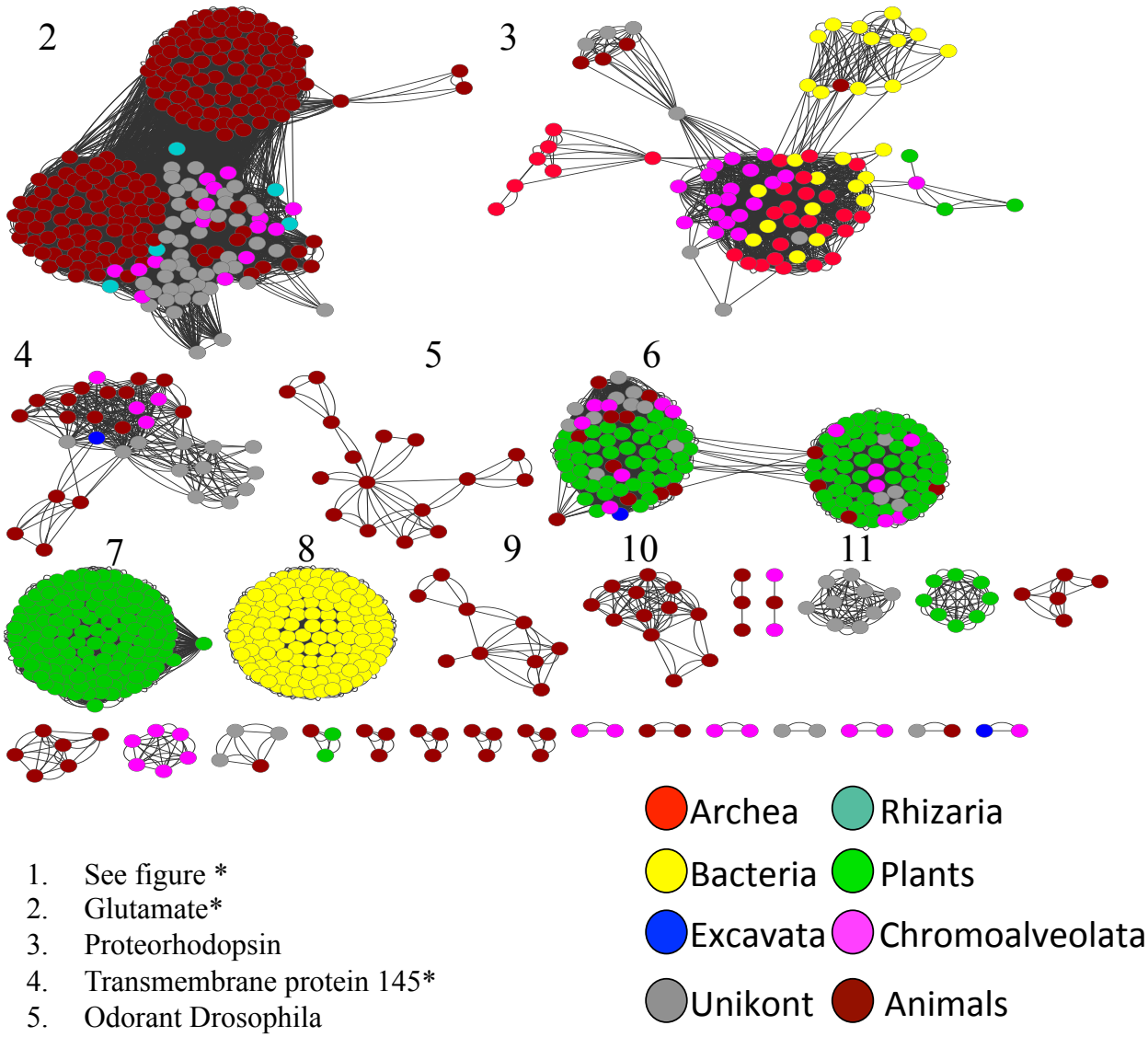


Figure 3.4a: Phylogenetic network of CC 1. Colour scheme and associated pathway is showed in figure 3.4b.



1. See figure *
2. Glutamate*
3. Proteorhodopsin
4. Transmembrane protein 145*
5. Odorant Drosophila
6. Transmembrane 187/gpr 107*
7. MLO
8. Ribonuclease
9. Gustatory receptor Lophotrocozoa
10. Ocular albinism*
11. Fungal Pheromone

Figure 3.4b: Phylogenetic network of all the other CCs. * indicates whether these sequences are known to use a G-protein signalling pathway.

This component is mostly composed of animal Rhodopsin, and if one were to exclude these proteins, this very large CC would immediately decrease in size and become comparable with most other components in figure 3.4b (e.g. CC 2 – Glutamate receptors or CC – 7 MLO). If one excludes the Rhodopsins, that are limited to the animals, one can identify other interesting aspects of CC 1. The first is that the Fungi Frizzled sequences are separated from those of the animals and from those of the Amoebozoa, which on the other hand are connected. This suggests that the Frizzled family is polyphyletic. Frizzled sequences have different evolutionary origins, and simply converged on the same function. I suggest the Metazoa plus Amoebozoa Frizzled sequences should be considered to be the original Frizzled group, whilst the fungal sequences should probably be best referred to as "Frizzled-like".

A further intriguing aspect of the history of the Frizzled group that figure 3.4a suggests, is that these proteins, that are of key relevance in cell-cell communications, and underlie the origin of multicellularity, might have been acquired by the Amoebozoa via LGT from an early animal. Figure 3.5a and 3.6a illustrate that this early animal might have been related to the Placozoa. One could thus conjecture that within Unikonta, there might have been two independent origins of multicellularity, in animal and fungi, whilst the tendency of Amoebozoa toward a simple form of multicellularity might have a common origin with that observed in animals.

Figure 3.4a and b can give us an idea on the origin of the GPCRs in general and of how many times they evolved. It is clear that every one of the 27 clusters in this Figure might represent an independent origin of the 7TMD domain. Some of the components in figure 3.4a and b include members of most eukaryotic supergroups. In particular, CC 1 includes Chromaleveolata, Rhizaria, Unikonta and plants. The taxonomic diversity of CC 1 is concentrated in the Secretin/GPCR-1/cAMP part of the network and one can conjecture that this GPCR block evolved in the last common eukaryotic ancestor. After that, a variety of expansions took place. In particular, an expansion of the Secretins and Frizzled, in Metazoa. Subsequently,

the Rhodopsins seem to emerge as an expansion of the Secretins. Amphimedon has a lineage specific GPCR1/cAMP expansion which is, however, recent (possibly demosponge-specific).

Figure 3.4b also highlights the Glutamate receptors (CC 2) as an ancient protein family, which similarly to CC 1 underwent a massive expansion in animals. Finally, a third, very interesting group, is represented by CC 7 (the MLO receptors) that seems almost certainly to be plant specific (Figure 3.4b).

Figure 3.5a and 3.5b are similar to figure 3.4a and b, but here the non-unikonts have been coloured in black to allow for a better definition of the history of the GPCRs in Unikonta. Within Unikonta a variety of groups have been marked out in colour. This Figure essentially illustrates the expansion of the Rhodopsins in animals, in the Neuralia first and in the Bilateria after that. Given the various hypotheses that have been proposed for the origin of the unikont GPCRs (Pardo *et al.* 1992; Taylor and Agarwal 1993) in figure 3.6a and 3.6b I highlighted what GPCR group might be of prokaryotic origin. I find that bacterial Rhodopsins may be ruled out entirely as representing the source of the unikont GPCRs.

In addition, I looked at whether some GPCR groups were in some way related to protein families with less transmembrane domains (Figure 3.7a and b and figure 3.8a and b). To do this, I first included in the network all the proteins with 2 to 6 transmembrane domains (Figure 3.7a and b) that had a BLAST hit of at least 10^{-10} with at least one of the proteins in figure 3.4a and b. Further to this, I performed a second analysis (Figure 3.8a and b), in which only proteins with 2 to 4 domains were retained (i.e. I assumed that proteins with 5 and 6 domains where partial 7TMDs sequences). These analyses showed that most of the proteins with less than 7TMD are randomly scattered amongst the various CCs. The only significant exception seems to be represented by the MLO (where the 7TMD proteins are sandwiched between two sets of proteins with 2,4 and 5 domains). This suggests that MLO have similarity with two sets of proteins with less than 5 domains, and that these two independent sets do not have similarity with each other.

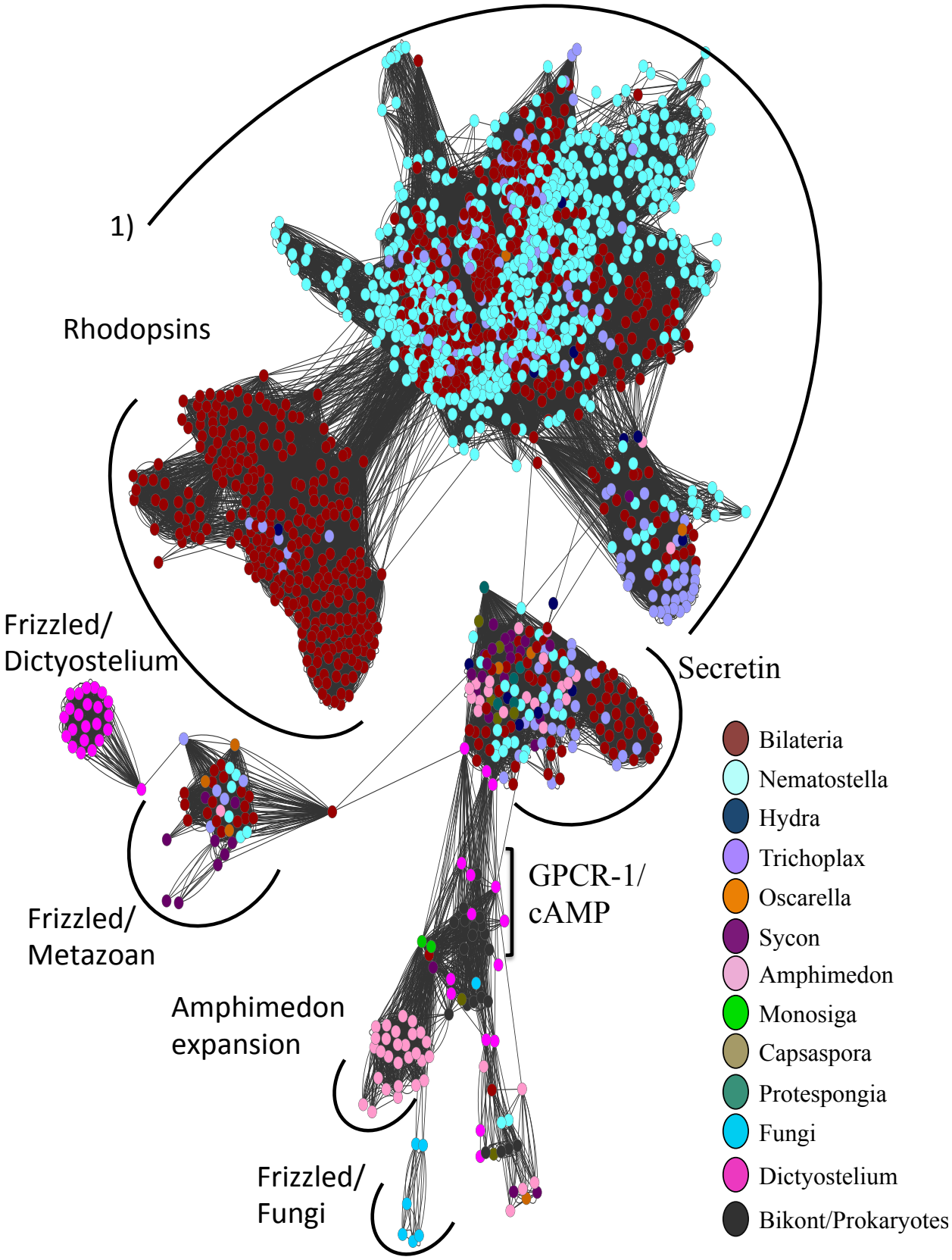
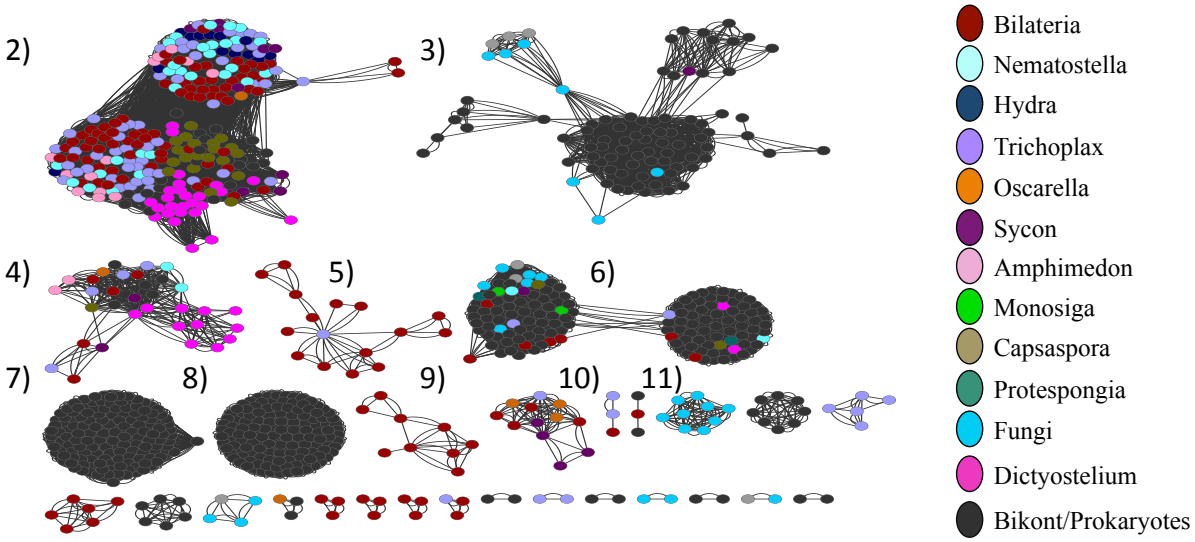


Figure 3.5a: Phylogenetic network of CC 1 with emphasis on the unikonts. Function and associated pathway are indicated in figure 3.5b.



1. See figure *
2. Glutamate*
3. Proteorhodopsin
4. Transmembrane protein 145*
5. Odorant Drosophila
6. Transmembrane 187/gpr 107*
7. MLO
8. Ribonuclease
9. Gustatory receptor Lophotrocozoa
10. Ocular albinism*
11. Fungal Pheromone

Figure 3.5b: Phylogenetic network of all the other CCs with emphasis on unikonts. * indicates whether these sequences are known to use a G-protein signalling pathway.

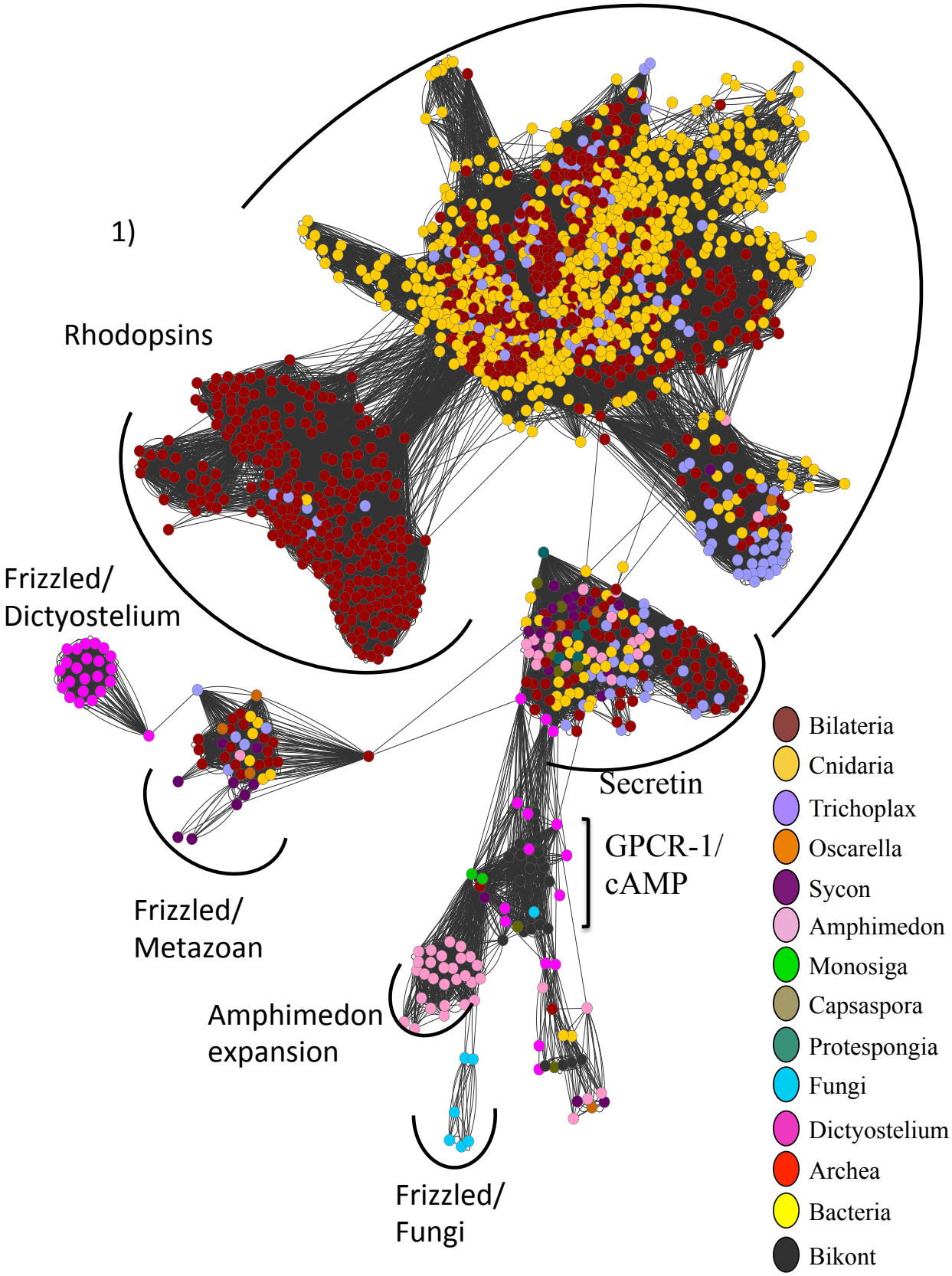
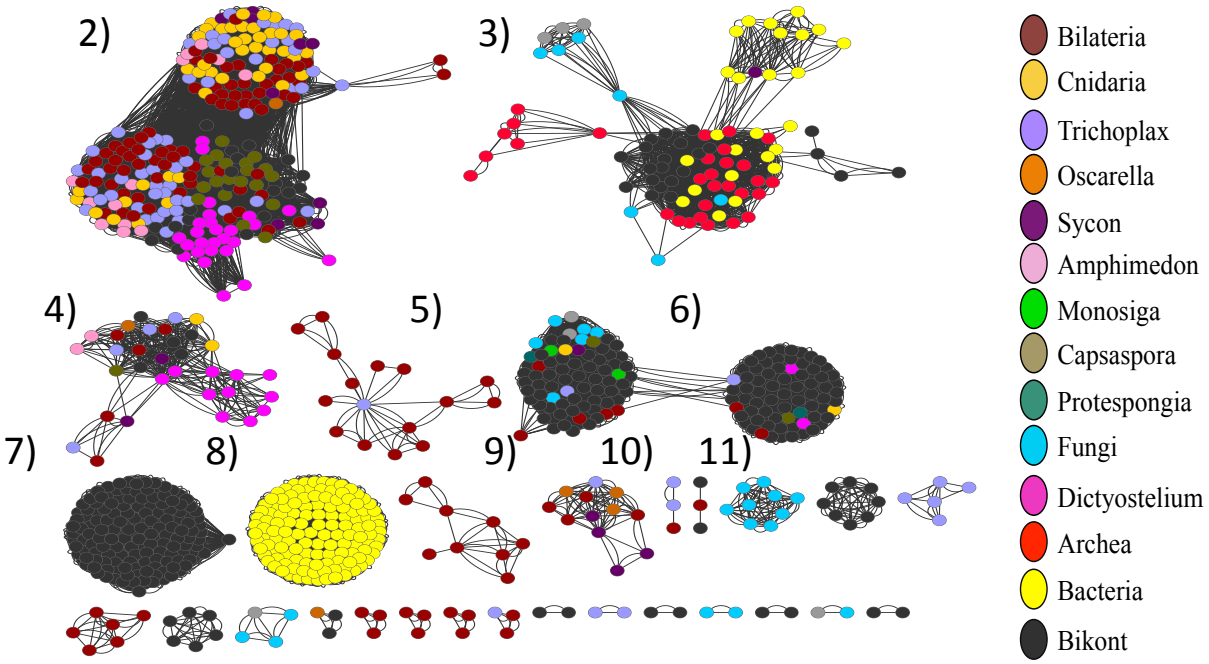


Figure 3.6a: Phylogenetic network of CC 1 with emphasis on the unikonts and the Prokaryotes. Function and associated pathway are indicated in figure 3.6b.



1. See figure *
2. Glutamate*
3. Proteorhodopsin
4. Transmembrane protein 145*
5. Odorant Drosophila
6. Transmembrane 187/gpr 107*
7. MLO
8. Ribonuclease
9. Gustatory receptor Lophotrocozoa
10. Ocular albinism*
11. Fungal Pheromone

Figure 3.6b: Phylogenetic network of all the others CC with emphasis on the unikonts and the Prokaryotes. * indicates whether these sequences are known to use a G-protein signalling pathway.

That is, the MLO probably evolved through the gene fusion of unrelated proteins with less than 7TMD.

Finally, I tested whether the clusters in figure 3.4a and b were robust by suppressing all the nodes in the network between proteins with less than 30% sequence identity. This was done to limit potential false positives (i.e. random hits). The results obtained are reported in figure 3.9. The effect of this test is visible in the key groups, particularly in CC 1. In this group, exclusion of proteins with low similarity causes the animal Frizzled and Amoebozoa Frizzled to separate, suggesting that these proteins might not be related to the Secretin/Rhodopsin group after all. If this is correct, we will have to assume three independent origin of the Frizzled group. On the other hand, the relation between the Rhodopsin and Secretin families is now apparent. The presence of unikont sequences in the Secretin sub-CC suggests a possible polarization of this network and indicates that the Rhodopsin-like proteins most likely evolved from Secretins. In Figure 3.9 the plant GPCR-1 form a cluster nested between a unikont cAMP receptor and a unikont Secretin, suggesting that these proteins might have evolved in plant through the fusion of two independently transferred sequences. In figure 3.9, the Fungi Frizzled sequences are still strongly associated with the Secretins (as in figure 3.4a) confirming that these are not related to the other Frizzled groups, but are modified Secretins instead. Overall, if one were to look at all the connected clusters in figure 3.9, it is clear that the only CC that has members from across three out of four eukaryotic supergroups is CC 2. Therefore, this is the only one that could have originated in the stem eukaryotic lineage. Further analyses performed removing sequences with a level of identity less than 40% (Figure 3.10) and 50% (Fig 3.11) suggest that it is quite unlikely that these sequences in CC2 have been horizontally transferred.

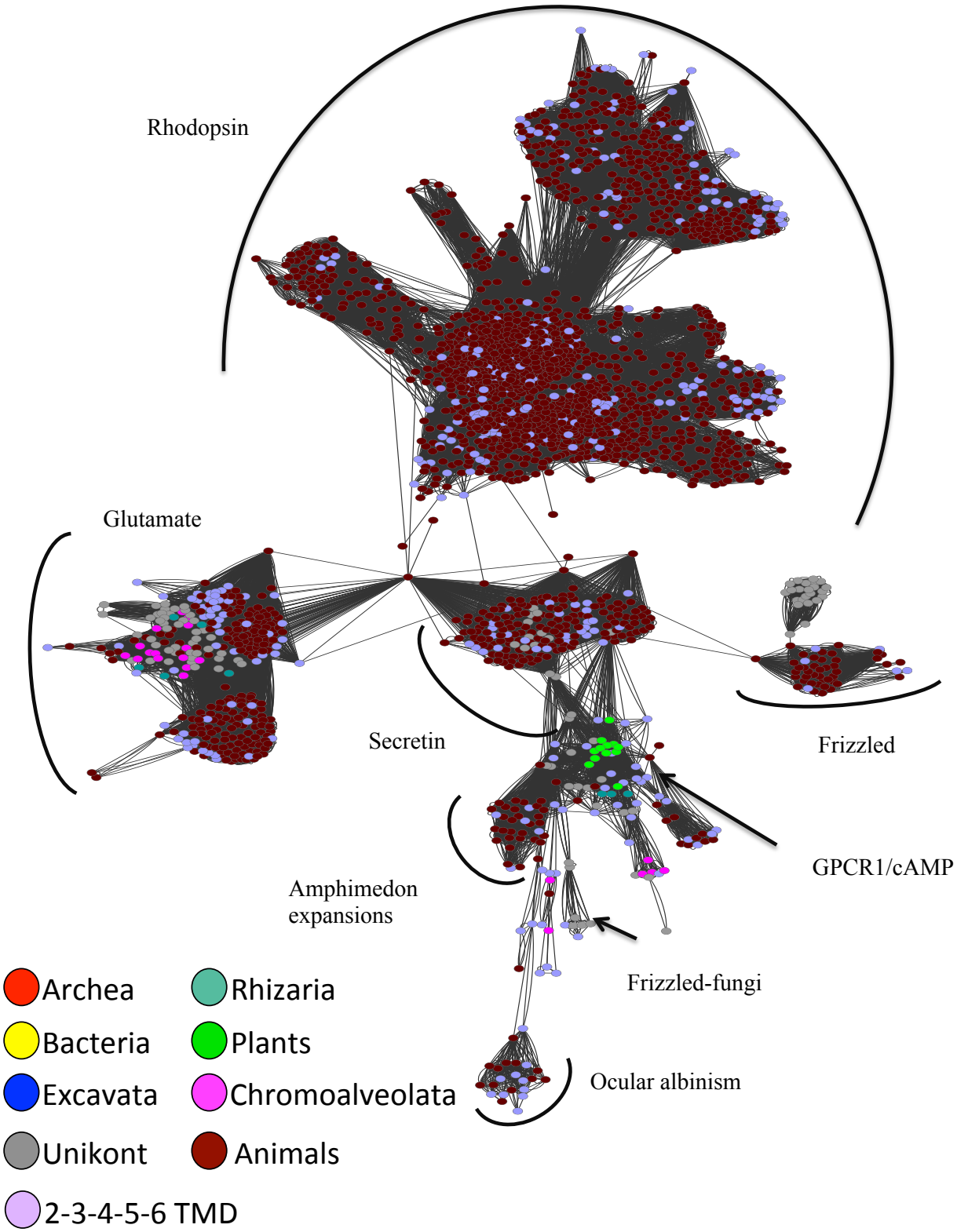


Figure 3.7a: Phylogenetic network of CC1 including also proteins with less then 7TMD. Function is indicated in figure 3.7b.

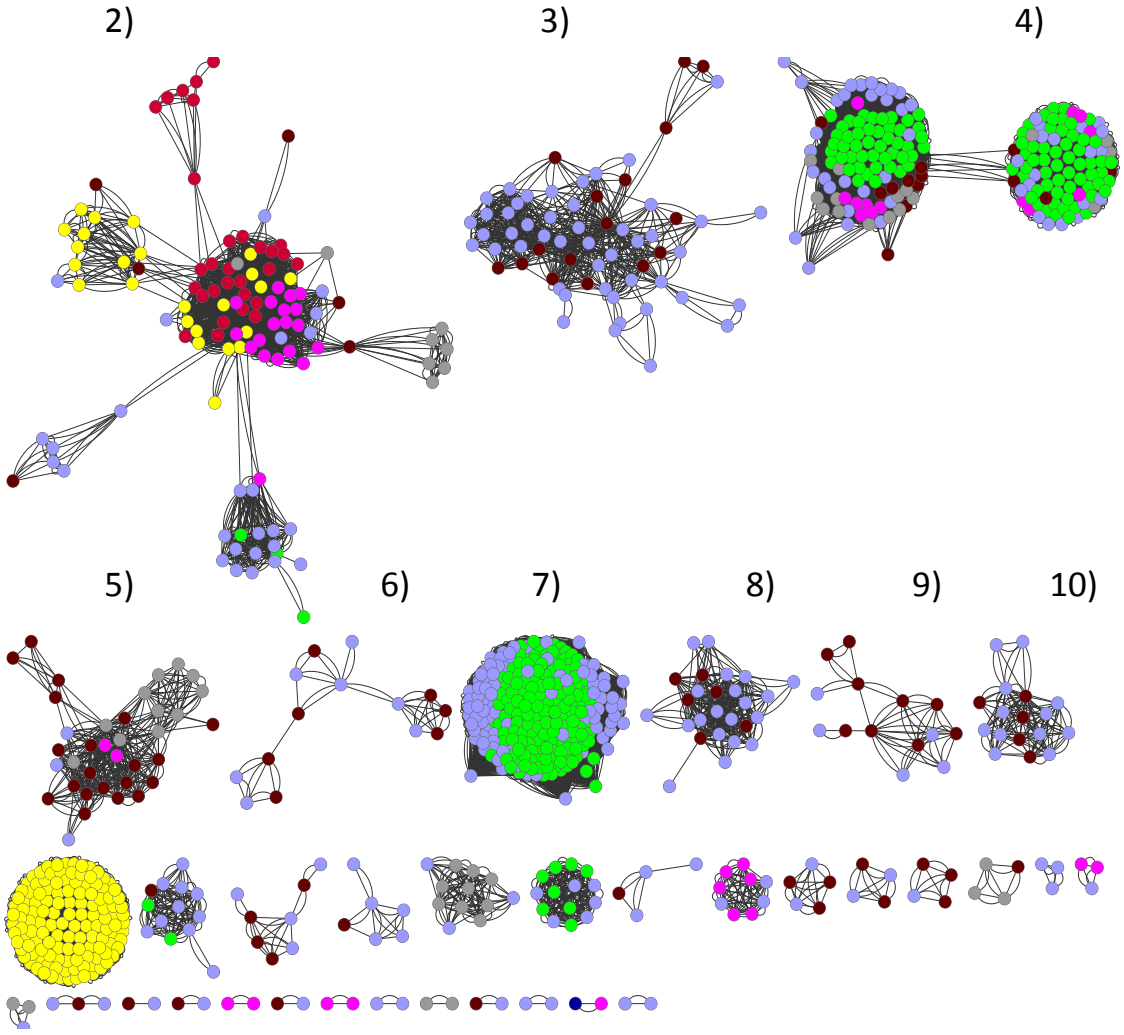


Figure 3.7b: Phylogenetic network of all the other CCs including also proteins with less then 7TMD.

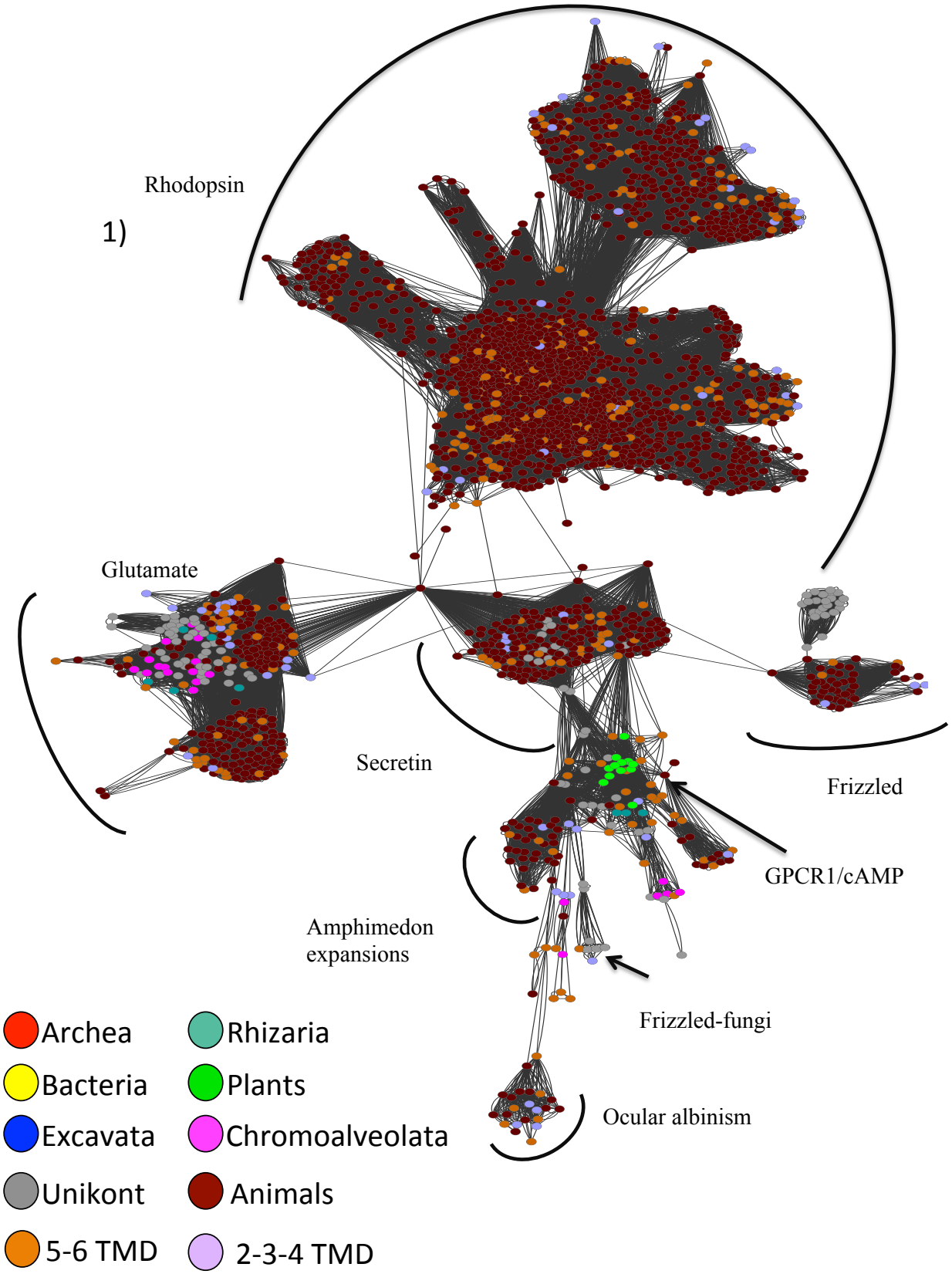


Figure 3.8a: Phylogenetic network CC1 and including also proteins with less then 7TMD (but showing proteins with 5 and 6 domains in a different colour). Function is indicated in figure 3.8b.

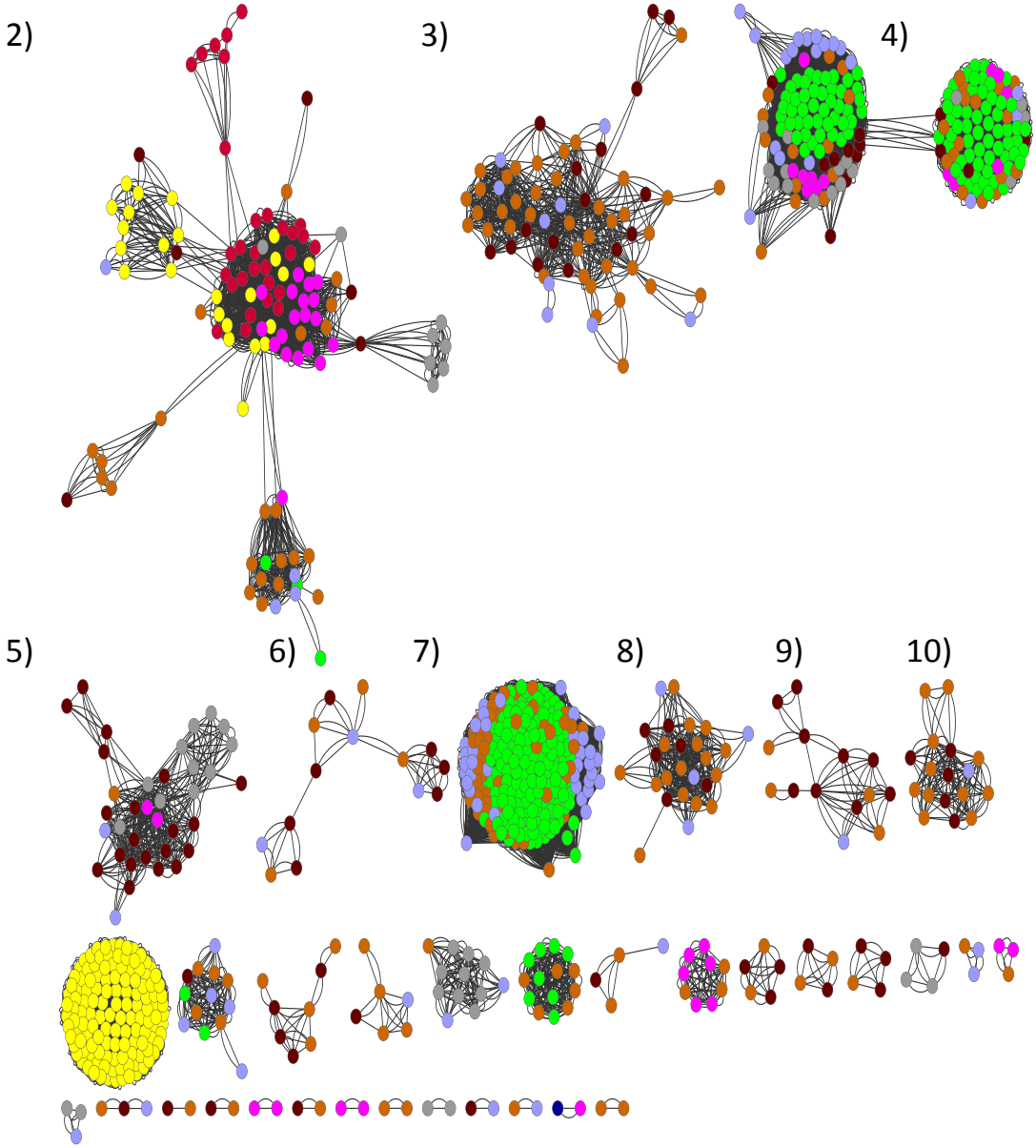
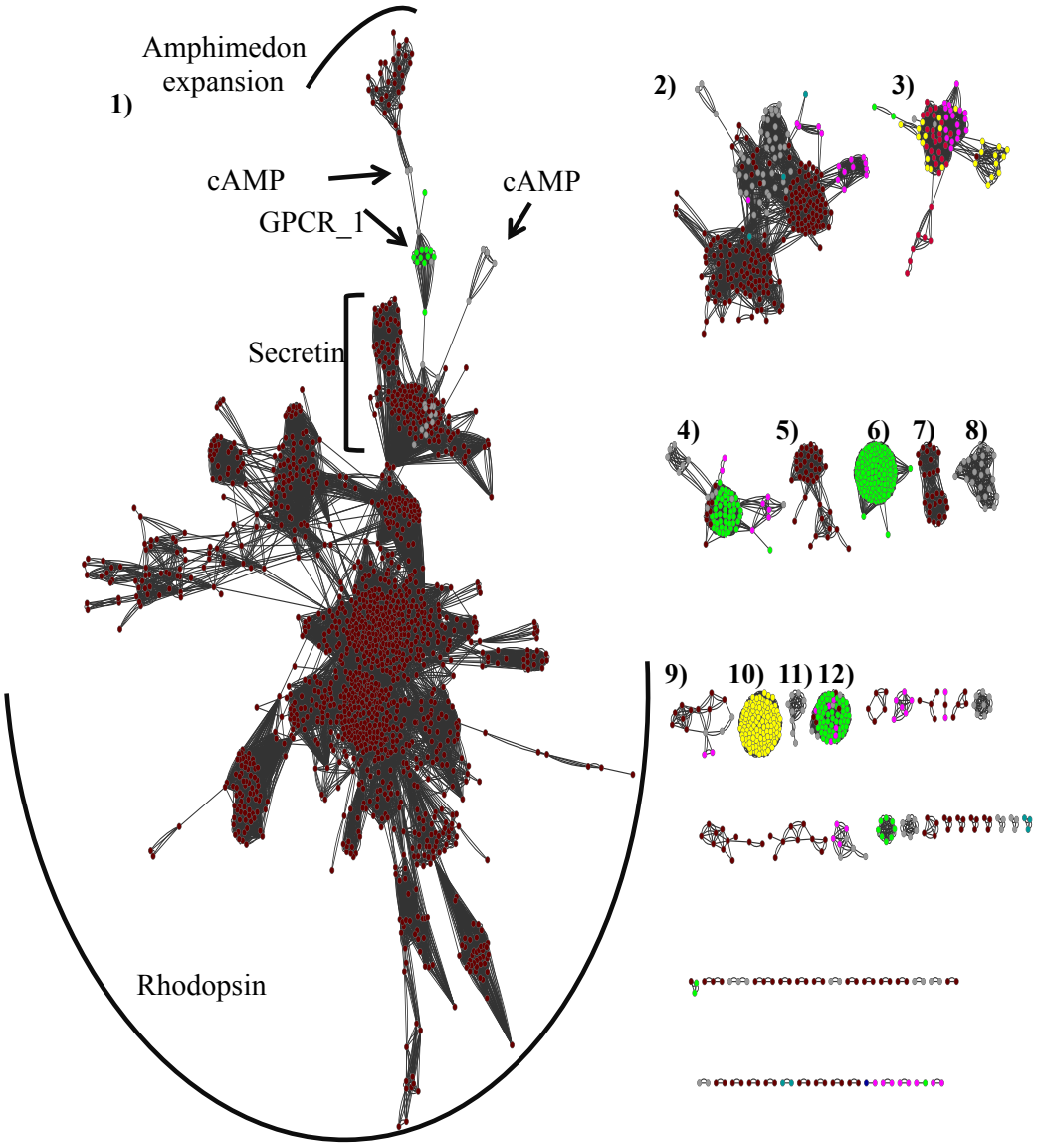


Figure 3.8b: Phylogenetic network of all the others CCs and including also proteins with less then 7TMD (but showing proteins with 5 and 6 domains in a different colour).



Archea
Bacteria
Excavata
Rhizaria
Plants
Chromoalveolata
Animals
Unikont

1. See Figure
2. Glutamate
3. Protorhodpsin
4. Transmembrane 87
5. Frizzled Metazoan
6. MLO
7. Frmf-Rhodopsin
8. Frizzled Dictyostelium
9. Transmembrane 145
10. Ribonuclease
11. Pheromone Fungi
12. GPR-107

Figure 3.9: Phylogenetic network where nodes with less then 30% similarity network are suppressed. This Figure is the same of figure 3.4a and b. However, in this case all the nodes in the network between proteins with less than 30% sequence identity are suppressed. Singletons are not shown

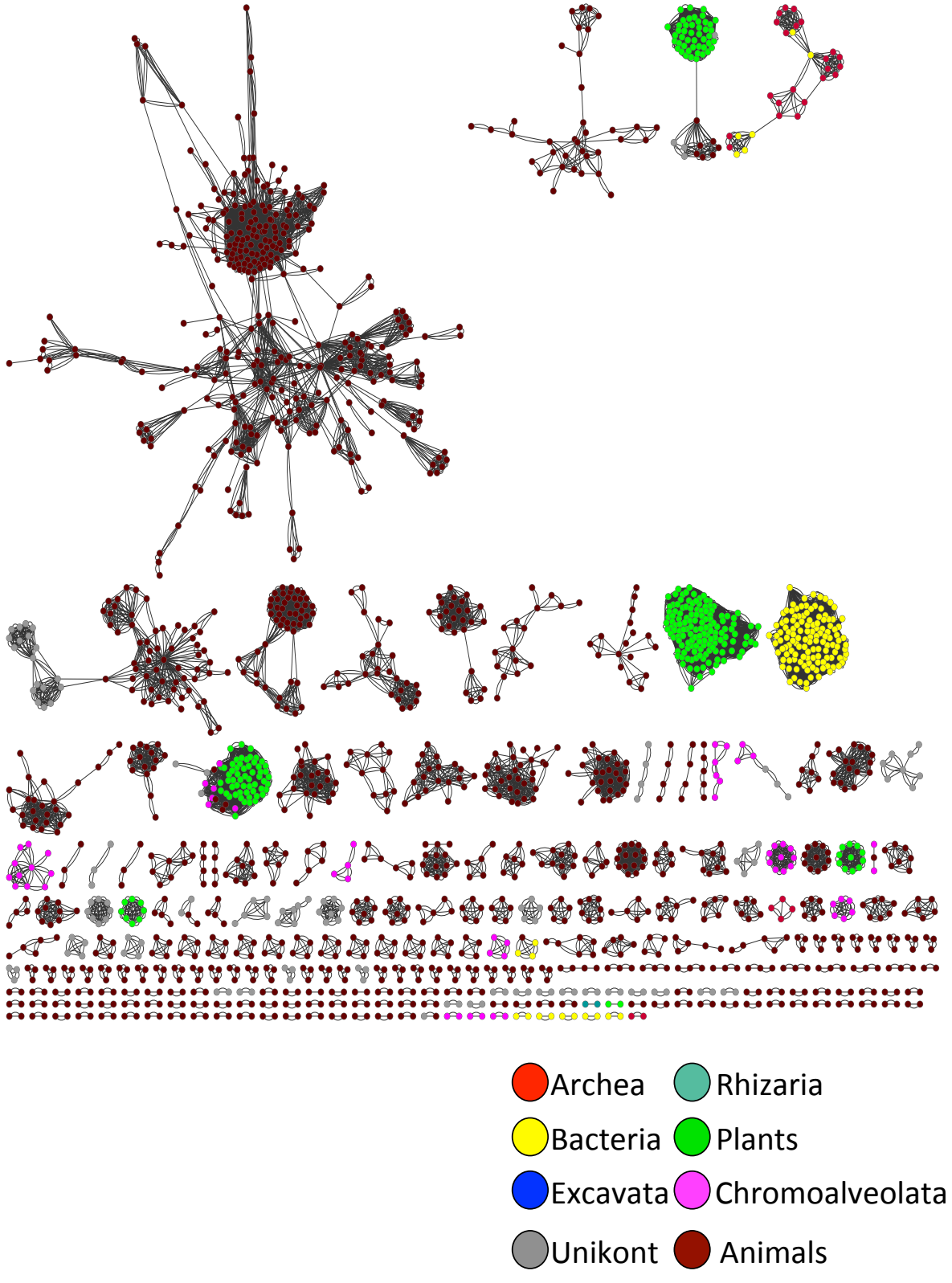


Figure 3.10: Phylogenetic network where nodes with less then 40% similarity network are suppressed**.** This Figure is the same of Figure 3.4a and b but in this case all the nodes in the network between proteins with less than 40% sequence identity are suppressed. CC1 and CC2 are now separated in several small CC. Singletons are not shown

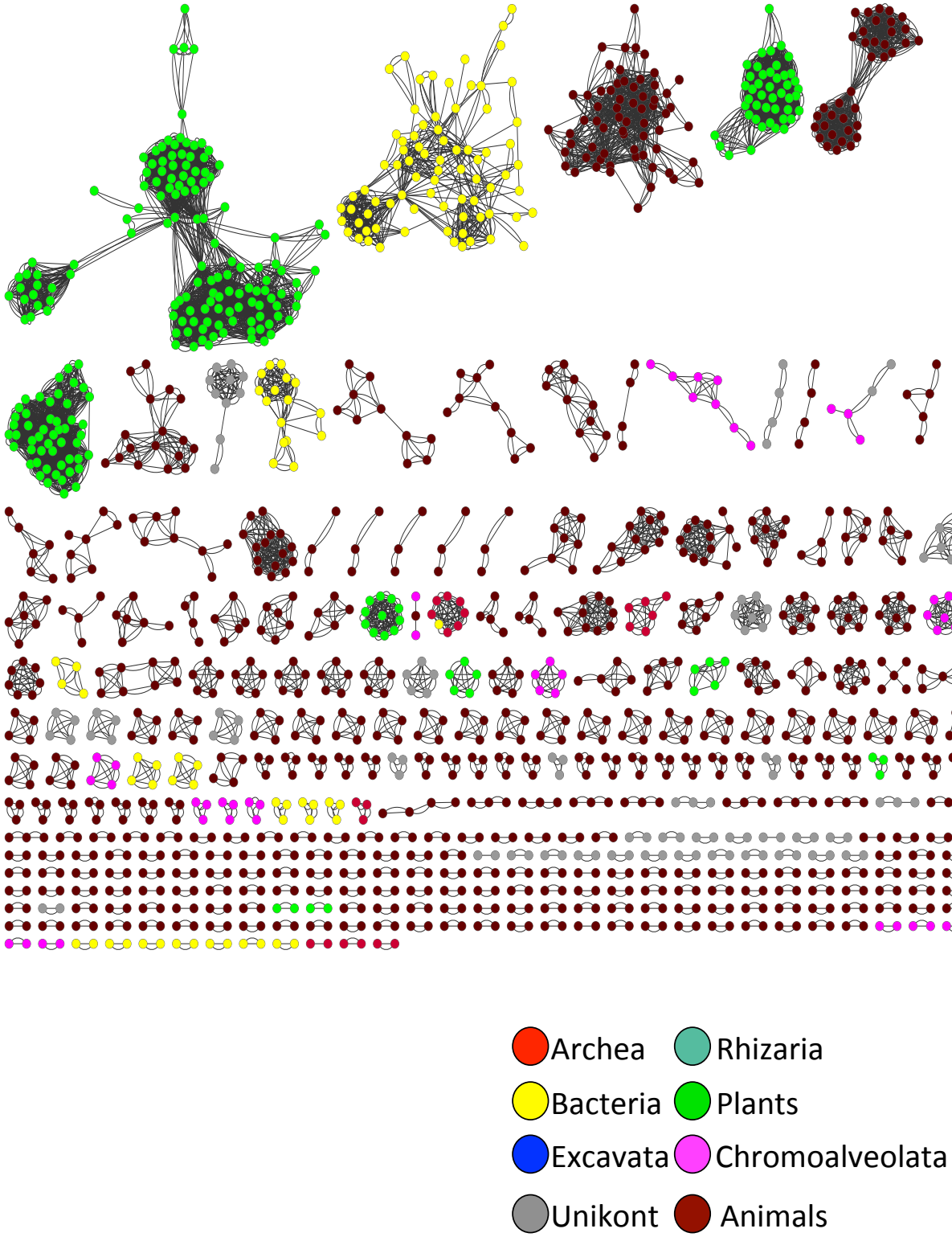


Figure 3.11: Phylogenetic network where nodes with less then 50% similarity network are suppressed. This Figure is the same of Figure 3.4a and b but in this case all the nodes in the network between proteins with less than 50% sequence identity are suppressed. CC1 and CC2 are now separated in many small CC. Singletons are not shown.

3.4.1 Is the 7TMDs architecture an example of convergent evolution?

The networks in figure 3.4a and b to figure 3.9 are quite intriguing. A total of 27 CCs were defined in figure 3.4a and b and these increased to 60 CCs in figure 3.9 when branches joining proteins with less than 30% identity were suppressed. Every CC in these figures could potentially be viewed as an independent evolution of the 7TMDs architecture, and if this were proven to be correct it would have profound implications for the origin of G-protein signalling, that should be considered to have been recruited multiple times in 7TMD-based signalling. Alternatively, although less likely, the entire pathway could have appeared multiple times.

With reference to the early evolution of the GPCRs it can be stated with confidence that the proteorhodopsins are not involved in the two most important unikont groups (CC1 and 2 in figure 3.4a and 3.9). That is, relationships among the proteorhodopsins and the eukaryotic rhodopsins are only structural, and most likely represent convergence (see also Soppa 1994). Convergent evolution to the same enzymatic function is widespread in nature (for a review, see (Zakon 2002). However, it is also possible, as it has been proposed by (Strotmann *et al.* 2011), that prokaryotic and eukaryotic 7TMDs have diverged so much that no residual sequence similarity remains between these proteins. Yet, because the fold-sequence relationship is degenerate (i.e. multiple evolutionary independent proteins with no sequence similarity are known to fold to the same three dimensional structure), arguments of homology based on structure alone are fundamentally unreliable, and should be considered with caution. This is particularly true in cases such as that of the eukaryotic GPCRs and of the prokaryotic 7TMD, where the receptors act in totally different ways (exploiting the G-signalling pathway in the case of the eukaryotic GPCRs, and opening/closing a ion pump in the case of the proteorhodopsins).

The results presented here (Figure 3.9) suggest that at the least one of the eukaryotic GPCR groups (the Glutamate receptors – CC2 in Figure 3.4b and 3.9) might be very ancient, as proteins belonging to this group are found in all eukaryotic supergroups except the excavates and plants. According to figure 3.4a also the Rhodopsin-like/Secretin group might be equally

ancient, as proteins of the GPCR1, Secretin and c-AMP receptors are distributed across most eukaryotic supergroups. Figure 3.9 shows that when proteins with less than 30% sequence identity are removed, the chromalveolate and rhizarian (see table 1.1) members of this group get disconnected suggesting that they might be ancient. However, plant members of this group (GPCR1) seem to have been acquired through LGT and as a consequence, the ancestry of this group is less certain. Consequently, the glutamate receptors remain as the only, potentially ancient, GPCRs CC. One that was likely to have been part of the genetic tool-kit of LECA (see below).

However, it is also possible that the GPCRs are separated in several CCs in my analyses because they evolved for hundreds of million of years under different selective pressures, despite having a single origin, although I accounted for this by using a PSI-BLAST in my database-searches.

The results presented here also suggest that for at least two 7TMDs (the gustatory receptor of *Tribolium* or the MLO-receptors in plants), the separation into independent isolated CC (Figure 3.9) is likely to reflect independent origins of the 7TMD. This is because, particularly in the case of the insects, the origin of these lineage specific receptors would not be particularly ancient. In addition, independent evidence exists to support the possibility that these CC represent a new invention of the 7TMD. For example, the insect receptors do not use G-protein signalling, differently from "true" GPCRs. In the case of the MLO, Figure3.6b illustrates that these receptors might have evolved from the fusion of receptors with 2/4 TMD, and that this is unique to the MLO receptors.

A remarkable feature of the GPCRs is the absence of sequence homology between CCs. From a structural point of view, the GPCRs and more generally the 7TMD architectures, seem to have extreme sequence plasticity. That is, it is able to accept mutations without loosing its thermodynamic stability. In other words, this fold seems to be characterized by a high level of designability (Shakhnovich *et al.* 2005).

3.4.2 The Rooting position of the Eukaryotes and GPCRs repertoire in LECA

My analyses have shown that some types of GPCRs are present in most eukaryotic super-groups (see figure 3.4 a and b and 3.9). My results expand the findings of (Nordstrom *et al.* 2011; Krishnan *et al.* 2012) and suggest that at least some type of GPCRs may have been part of the ancestral genetic tool kit of the LECA. If one excludes groups where multiple lines of evidence suggesting independent 7TMD evolution, and considers that proteins with a 7TMD that exploit the G-signalling pathway (i.e. the GPCRs) might be homologs, one should conclude that: (1) GPCRs were part of the toolkit of LECA, and (2) the Glutamate receptors are the most likely candidate for the ancestral GPCR.

3.4.3 The expansion of the GPCRs in Metazoa

The high dynamism characterizing the GPCRs finds its best example in the Metazoa. As expected, the amount of GPCRs increases in Metazoa. However, the results presented here suggest that the GPCRs expansion coincides with the origin of *Neuralia* (*sensu* Nielsen 2012- see table 1.2), rather than with the origin of Metazoa. In light of the evolution of complex structures, and as already widely described in the introduction, this result is coherent. Sponges are animals that largely work as unicellular organisms and lack tissues (possibly with the exception of *Oscarella carmela*). Differently, cnidarians have relatively complex organs and systems (i.e., a nervous net and a digestive system).

The GPCR increase in the neuralian lineage (see Figure 3.3a) suggests that these proteins played a central role in the evolution of complex structures, and in increasing physiological potential. However, some of the results here presented are quite surprising, for example they suggest an unexpectedly high number of GPCRs (617) in the Cnidarian *Nematostella vectensis*. One hypothesis that can be made with reference to the GPCR expansion, particularly the Rhodopsins (that are mainly involved in the processes of the nervous system), is that their expansion

coincides with the origin of the nervous system. Even though this hypothesis is fascinating, it remains untestable.

3.5 Conclusion

In this chapter, I provided evidence for the ancient origin of at least one type of 7TMD (the Glutamate receptors), which was present in LECA. This study also suggests that the 7TMD originated several times independently in the eukaryotes (at least three times). A high level of thermodynamic stability characterizes the 7TMD architecture and it is thus imaginable that it evolved several times.

Another, interesting and associated suggestion from this chapter is the potential multiple co-option of the G-protein pathway. However, it seems more likely that only 7TMDs, that do not use G-signalling, might represent independent acquisitions of the 7TMD in eukaryotes. This would imply that G-protein signalling was acquired only once by the 7TMD.

Finally, with reference to the animals, the number of GPCRs observed indicates that the GPCRs underwent an incredible expansion in Neuralia and this is consistent with the role they played in increasing the physiological potential of the neuralians (see Chapter 1).

In the next part of the thesis I will analyse the phylogenetic relationship among the opsins, which, with reference to the result presented in this chapter, represent a monophyletic group of opsins belonging to the Rhodopsins CC 1 (Figure 3.4a).

Chapter 4

Opsin evolution and the origin of vision

In Chapter 3 I investigated the origin and evolution of the GPCR in animals and pinpointed the existence of a massive expansion of Rhodopsin-like GPCRs in animals (Figure 3.9). In this chapter I shall focus on the Rhodopsin-like superfamily, and within this superfamily, I shall investigate the origin and early evolution of the animal visual opsins.

Abstract

All known visual pigments in Neuralia (Cnidaria, Ctenophora, and Bilateria) are composed of an opsin (a seven-transmembrane G protein-coupled receptor), and a light-sensitive chromophore, generally retinal. Accordingly, opsins play a key role in vision. There is no agreement on the relationships of the neuralian opsin subfamilies, and clarifying their phylogeny is key to elucidating the origin of this protein family and of vision. We used improved methods and data to resolve the opsin phylogeny and explain the evolution of animal vision. We found that the Placozoa have opsins, and that the opsins share a common ancestor with the melatonin receptors. Further to this, we found that all known neuralian opsins can be classified into the same three subfamilies into which the bilaterian opsins are classified: the ciliary (C), rhabdomeric (R), and go-coupled plus retinochrome, retinal G protein-coupled receptor (Go/RGR) opsins. Our results entail a simple scenario of opsin evolution. The first opsin originated from the duplication of the common ancestor of the melatonin and opsin genes in a eumetazoan (Placozoa plus Neuralia) ancestor, and an inference of its amino acid sequence suggests that this protein might not have been light-sensitive. Two more gene duplications in the ancestral neuralian lineage resulted in the origin of the R, C, and Go/RGR opsins. Accordingly, the first animal with at least a C, an R, and a Go/RGR opsin was a neuralian progenitor.

4.1 Introduction

Understanding the origin and early evolution of vision at the molecular level has proven difficult (Plachetzki *et al.* 2007;Plachetzki et al. 2010; Suga *et al.* 2008; Porter *et al.* 2011). Both Protostomia (e.g. Mollusca and Arthropoda) and Deuterostomia (e.g. Vertebrata) have eyes and it is plausible that the last common ancestor of the Bilateria (i.e. the Urbilateria), possessed simple eyespots and some limited ability to detect light (Land and Nilsson 2002). In addition, eyes are known in jellyfishes (e.g. Nilsson *et al.* 2005; Kozmik *et al.* 2008), and the common use of a Pax-6 regulated kernel (*sensu* Davidson and Erwin 2006) to control eye development in Cnidaria and Bilateria suggests a single origin of the neuralian eye (Gehring 2011). Furthermore, all neuralians for which data are available detect light using visual pigments composed of an opsin and a chromophore, generally retinal (Porter *et al.* 2011); and their opsins link the chromophore through a Schiff-base involving the Lysine found at position 296 (K296) of the reference bovine rhodopsin (Nathans and Hogness 1983) K296 is the key residue in the neuralian Retinal Binding Domain (RBD).

Opsins are seven-transmembrane proteins belonging to the GPCR superfamily (Terakita 2005), and according to the GRAFS (Fredriksson *et al.* 2003) classification system, they are members of the a-group of the Rhodopsin-like receptors (Figure 3.3). The opsin family includes several well-characterised subfamilies (Terakita 2005), and given the universal distribution of opsins in Neuralia (Plachetzki *et al.* 2007; Koyanagi *et al.* 2008; Kozmik *et al.* 2008; Suga *et al.* 2008; Plachetzki *et al.* 2010) it is clear that to understand the molecular foundations of vision we must focus on the non-bilaterian animals: the Cnidaria, the Ctenophora, the Placozoa and the sponges. Unfortunately, the phylogenetic relationships of the neuralian opsin subfamilies are still debated (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) and consequently, the early history of gene duplications and deletions within this family is still unknown (see Figure 4.1). Should we wish to understand the origin of vision, the pattern of opsin duplications and deletions must be clarified first, and the only way to accomplish this goal

is resolving the opsin phylogeny. A further consequence of uncertainty in opsin relationships is that the evolutionary timescale of visual evolution is still unknown. Divergence times among the animal phyla (Erwin *et al.* 2011) lets us very loosely bracket the early evolution of vision in the 105 Million Years (Ma) interval delimited by the divergence between the Demospongiae and the other metazoans (~773 Ma), and that between the Protostomia and the Deuterostomia (~668 Ma). However, this is a maximal time estimate, and the crucial steps in opsin evolution most likely unfolded in a significantly shorter time. Only by resolving the opsin phylogeny can we also clarify the evolutionary *tempo* of vision.

The current gap in our understanding of the evolution of vision is, at least in part, the consequence of an absence of genomic information for key, early branching metazoans. Data are still missing for two non-bilaterian lineages: the Ctenophora and the calcarean sponges. However, the genomes of four key taxa, the placoazoan *Trichoplax adherens* (Srivastava *et al.* 2008), the cnidarians *Hydra magnipillata* (Chapman *et al.* 2010; Srivastava *et al.* 2010) and *Nematostella vectensis* (Putnam *et al.* 2007), and the demosponge *Amphimedon queenslandica* (Srivastava *et al.* 2010) have recently been released, improving data availability. Further to this, the genome of *Oscarella carmela*, a representative of a second sponge class (the Homoscleromorpha), has now been sequenced (Nichols *et al.* 2012)

The relationships among the sponge classes are still debated (Hejnol *et al.* 2009; Philippe *et al.* 2009; Sperling *et al.* 2009; Sperling *et al.* 2010; Erwin *et al.* 2011), and two competing hypotheses exist. The first suggests that the sponges are monophyletic (Philippe *et al.* 2009; Pick *et al.* 2010), whilst the second (Hejnol *et al.* 2009; Sperling *et al.* 2009; Erwin *et al.* 2011; Nielsen 2012) suggests that they are paraphyletic. However, (see Chapter 2) the phylogenomic analyses presented here favour the sponge paraphyly hypothesis over the sponge monophyly hypothesis. According to the sponge monophyly hypothesis, Porifera is the sister group of Eumetazoa and hence both the Demospongiae and the Homoscleromorpha are valid outgroups to study GPCR (and opsin) evolution in Eumetazoa. In contrast, according to the paraphyly

hypothesis (that the results in Chapter 2 favour) only the Homoscleromorpha are a valid outgroup to study the eumetazoan GPCRs. It follows that the inclusion of the *Oscarella* genome in this study is key to ensure that the closest, putative sister group of the Eumetazoa is included in the analyses. Here genomic information from all the above-mentioned taxa has been used, together with a large sample of well-characterised eumetazoan opsins (see Table 2 in Electronic Appendix), to investigate the origin and evolution of the opsin family and the origin of animal vision.

Animal opsins have been classified in three major subfamilies (Terakita 2005): Rhabdomeric opsins (R– opsins), Ciliary opsins (C–opsins), and Go–coupled (Go) plus Retinal G–protein coupled Receptor (RGR) opsins (Go/RGR–opsins). Usually there is an association between light receptors (i.e. the cells expressing these proteins) and specific opsin subfamilies, with the ciliary receptors expressing C– and Go/RGR–opsins, and the rhabdomeric receptors expressing R–opsins (Fain *et al.* 2010; Porter *et al.* 2011). A fourth opsin subfamily was suggested by Plachetzki and colleagues (Plachetzki *et al.* 2007). These authors (Figure 4.1a) identified a large clan (*sensu* (Wilkinson *et al.* 2007) of cnidarian-specific opsins that they named "Cnidopsins". In addition, they found that one cnidarian opsin in their data set clustered with the bilaterian C–opsins (Figure 4.1a) a result that is consistent with the observation that cnidarians have ciliary receptors (Fain *et al.* 2010).

Four studies (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) have previously addressed the relationships among the main opsin groups with a view of clarifying the gene duplication and deletion history within this family, but these studies reached contradictory results (see Figure 4.1).

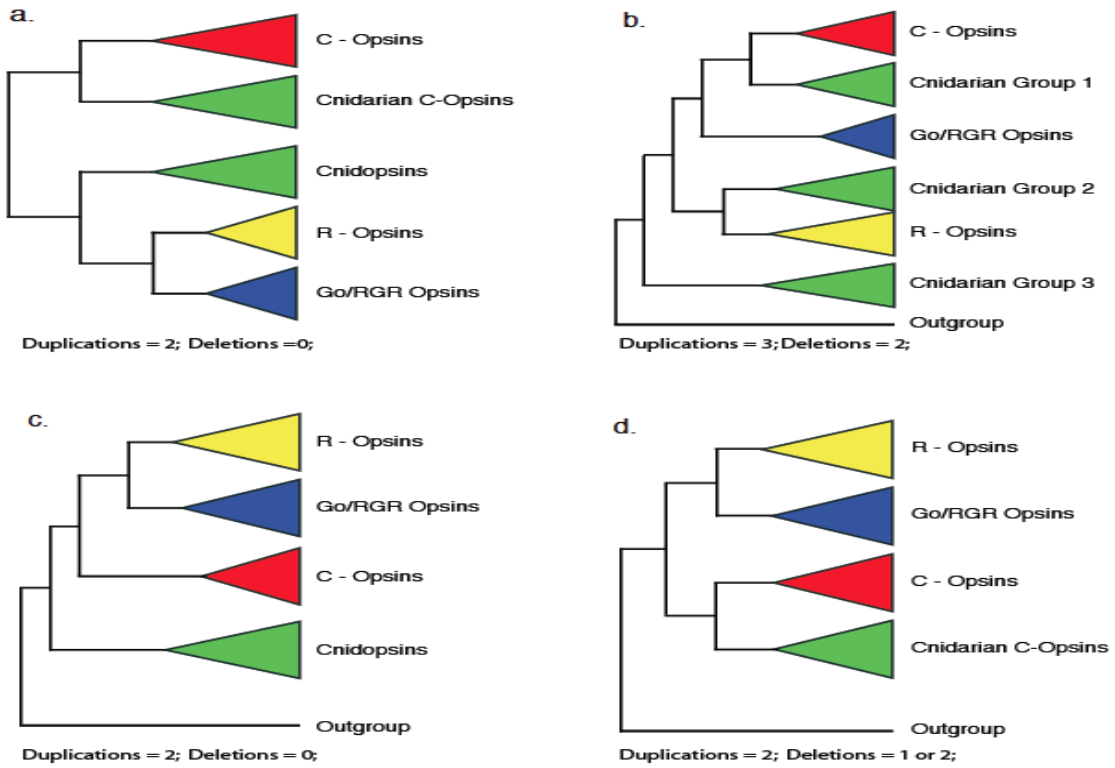


Figure 4.1: Alternative hypotheses of opsin relationships. (A) The phylogeny of (Plachetzki *et al.* 2007). In this tree the cnidarian–specific opsins form two groups. One represent the sister group of the C–opsins and includes only one sequence. The second represent the sister group of the R– plus Go/RGR–opsins. This phylogeny can be explained with two duplications only. The first duplication happened in the stem Neuralia lineage and separated the C–opsin lineage from the Cnidopsin, plus R, plus Go/RGR lineage. The second duplication happened in the stem bilaterian lineage and separated the R– from the Go/RGR–opsins. (B) The phylogeny of (Suga *et al.* 2008). In this hypothesis the cnidarian specific opsins are split into three groups. These represent the sister group of the C–opsins, of the R–opsins and of all the other opsins. In (Suga *et al.* 2008) these cnidarian–specific opsins were referred to as: Group 1, Group 2 and Group 3, respectively (see Figure 4.1b). To explain the opsin distribution in Figure 4.1b three duplications and two deletions are necessary. The first duplication separates the Group 3 opsins from all the other opsins. The other two duplications separate the C– plus Go/RGR–opsins from the C–opsins, and the C– from the Go/RGR–opsins, respectively. The first of the two deletions affected the Bilateria that loose the Group 3 opsins. The second deletion affected Cnidaria that loose their Go/RGR opsin paralog. (C) The phylogeny of (Plachetzki *et al.* 2010). This phylogeny implies that the Cnidopsins are the Cnidarian ortholog of the bilaterian opsins, and can be explained with two duplications in the stem bilaterian lineage only. The first of these duplications separated the C–opsins from the R plus Go/RGR lineage. The second separated the R–opsins from the Go/RGR– opsins. (D) The phylogeny of (Porter *et al.* 2011). This phylogeny can be explained with two duplications and one or two deletions. The first duplication separated the C–opsins from the R plus Go/RGR lineage and happened in the stem eumetazoan lineage. The second duplication separated the R–opsins from the Go/RGR–opsins. The two deletions happened in the Cnidarian lineage and caused the loss of the R and Go/RGR paralogs. If the duplication separating the R–opsins from the Go/RGR opsins happened in the stem bilaterian lineage then only one deletion (of the R–opsin plus Go/RGR–opsin ortholog) happened in the cnidarian lineage.

A major source of uncertainty in these studies is that (Plachetzki *et al.* 2007; Plachetzki *et al.* 2010; Porter *et al.* 2011) failed to include a representative sample of Cnidarian opsins (Figure 4a, c, and d) and did not have the power to test every possible hypothesis of opsin evolution. In addition the studies of (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) used precomputed, empirical time reversible matrices to model amino acid substitutions. These matrices – WAG (Plachetzki *et al.* 2007; Plachetzki *et al.* 2010), MtRev (Porter *et al.* 2011), and JTT (Suga *et al.* 2008) – are unlikely to fit an opsin dataset well because they were not derived from an opsin alignment. Consequently, the opsin phylogenies in Figure 4.1 might be affected by tree reconstruction artifacts (Philippe *et al.* 2009; Sperling *et al.* 2009; Holton and Pisani 2010). Further to this (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) used uncritically selected outgroups. Plachetzcki et al. (Plachetzki *et al.* 2007) recognised that the use of inadequate outgroups might have affected their results, but their solution to the outgroup selection problem was invalid. This is because these authors did not include outgroups in their analyses as they "destabilize[d] the ingroup topology". Instead, they used the AU test to select the branch where their unrooted (and outgroup-less) phylogeny should have been rooted. However, the time reversible model (WAG + Γ + I) that they used to estimate site-wise likelihood values for the AU test does not discriminate between the rooted resolutions of an unrooted tree. Accordingly, the differences between alternative rooting positions that they observed for a given unrooted topology must represent sampling and stochastic errors. Indeed, from a careful inspection of Plachetzki's *et al.* (2007) Table 3, it is clear that their AU tests (as expected) only let them discriminate between the three unrooted topologies in Plachetzki's *et al.* (2007) Figure 3, but not between the 15 rooted topologies reported in the same figure. This invalidates the most important criterion used in (Plachetzki *et al.* 2007) to select among alternative opsin phylogenies.

Here I performed new, detailed analyses, to better understand opsin evolution. Unlike previous studies I used the software PRANK (Loytynoja and Goldman 2008), a modern, well-

performing multiple sequence alignment software that can better discern insertions from deletions. I implemented better fitting evolutionary models, and considered all available genomic information for the deeply branching metazoans, including the newly sequenced genome of the homoscleromorph sponge *Oscarella carmela*. Finally, I thoroughly tested a large sample of putative outgroups and performed my analyses using only the closest, and less divergent, opsin outgroups. Most importantly, I used a comprehensive set of cnidarian opsins, including all the sequences specific to the studies of Plachetzki *et al.* (2007) and Suga *et al.* (2008). With the use of additional data, a well-performing multiple sequence alignment algorithm, better-fitting models, and a range of more adequate outgroups, I can test every possible hypothesis of opsin evolution and I expect to be able to achieve a greater precision in pinpointing duplications and losses in the opsin family.

4.2 Methods

4.2.1 Data mining, data set assembly, and alignment.

Taxonomic nomenclature in this study follows Nielsen (Nielsen 2012). I assembled a large sample of well-characterised opsins from across Eumetazoa (see Table 2 in Electronic Appendix), including key sequences like the putative cnidarian C–opsin of Plachetzki *et al.* (2007) and the putative cnidarian R–opsins of (Suga *et al.* 2008). In order to identify the closest outgroup(s) of the eumetazoan visual opsins, representatives of each monophyletic α–group of Rhodopsin–like receptors, and a set of sequences from the β–, γ–, and δ–groups (for a total of 139 sequences) were downloaded from GPCRdb (www.gpcrdb.org) and added to my data set (see Table 2 in Electronic Appendix). Sequences in GPCRdb are of vertebrate origin. To enrich my data set of putative GPCRs from non-bilaterian metazoans I mined the genomes of *Hydra magnipapillata*, *Nematostella vectensis*, *Trichoplax adherens, Amphimedon queenslandica* and

Oscarella carmela. These searches were seeded using the sequences I obtained from GPCRdb. To further enrich my data set of putative opsin homologues from non-bilaterian metazoans, I used my set of opsins to seed a series of BLAST-P searches against the genome of the placozoan *Trichoplax adherens*, and against a large set of predicted GPCRs from the two available sponge genomes (that of the demosponge *Amphimedon queenslandica*, and that of the homoscleromorph *Oscarella carmela*). This data-mining step was performed according to the following protocol: (I) each sequence in GPCRdb (a total of 42110 sequences) was used to seed a TblastN search of every scaffold of both sponge genomes. (II) Gene predictions were performed for all positive hits using Augustus (Stanke *et al.* 2008), trained against the *Amphimedon queenslandica* genome. (III) Predicted genes from both sponges (a total of 13059 Oscarella sequences and 23858 Amphimedon sequences) were merged into a database that also included the entire proteome of the placozoans *Trichoplax adherens*. (IV) A series of BLAST-P searches seeded using my set of 449 well-characterised opsins (see above) was performed against this database. All sequences with an e-value 1^{-20} were retained as representing putative opsin homologues. This procedure identified several putative opsin homologues from *Trichoplax*, one putative opsin homologue from *Oscarella*, but no putative opsin homologues from *Amphimedon.* Accordingly (V) a final BlastP analysis of the *Amphimedon* sequences was performed using, as a seed, the putative opsin homologue I identified in *Oscarella*. The two best hits from this final BLAST-P search (E-values = $1e^{-08}$ and $1e^{-07}$) were added to my data set.

My final data set included 625 GPCRs (499 opsins and 176 putative opsin outgroups). From this data set, I generated two master alignments (Loytynoja and Goldman 2008). The first alignment I generated, the All-Opsins Master alignment (AOM), included only the 499 neuralian opsins. The second alignment, the GPCR&Opsins Master alignment (G&OM), included all putative opsin outgroups (176 GPCRs in total) and a sample of 80 selected opsins (see below or details). The AOM and the G&OM alignments were, respectively, 317 and 366 positions long. A third alignment was generated *a posteriori* after having inspected the results of the analyses of

the G&OM data set (see below, Fig 3b and Fig 3 in Electronic Appendix) to identify the closest sister group of the animal opsins. This third alignment, the Opsins&Outgroup (O&O) alignment, included the 80 opsins in G&OM plus the closest sister group of the animal opsins (i.e. the MLT receptors – Figure 3b and Figure 3 in Electronic Appendix). O&O included 104 sequences and was 366 positions long. To build my two-master multiple sequence alignment (AOM and G&OM) I used Prank (Loytynoja and Goldman 2008) with the +F option. The two master alignments were visualized and manually edited using Jalview (Waterhouse *et al.* 2009) to eliminate gap-rich regions, as well as regions of dubious alignment quality.

In contrast to classical multiple sequence alignment software, Prank can distinguish insertions from deletions and has been suggested to have the potential to generate more realistic alignments. Indeed, previous investigations (Holton and Pisani 2010) shown that using Prank's alignments in phylogenetic studies based on single gene alignments, results in the recovery of more accurate phylogenetic trees (Holton and Pisani 2010). This suggests that Prank's alignments efficiently capture the phylogenetic signal single gene alignments.

4.2.2 Phylogenetic analyses

In this section I will focus on the logic of the analytical pipeline scheme used. The AOM alignment was analysed to recover an unrooted phylogeny including only well-characterised opsins from the three known bilaterian subfamilies (C, R, and Go/RGR), and an inclusive sample of cnidarian opsins. This analysis allowed an evaluation of the relative relationships among the cnidarian opsins in my data set and the opsins of Plachetzki *et al.* (2007), Suga *et al.* (2008) and Plachetzki *et al.* (2010). Results of the AOM analyses were used to select a subset of 80 opsins (20 C–, 20 R–, 20 Go/RGR–, and 20 cnidarian opsins) to be included in the G&OM and O&O data sets. Opsin subsampling was necessary to (I) reduce computational complexity and (II) minimise the likelihood of tree reconstruction artifacts (see below). Accordingly, fast evolving,

extremely short, and compositional heterogeneous sequences were not included in the G&OM and O&O alignments. However, a representative sample of sequences from every opsin clan identified in the AOM analysis was retained.

The G&OM alignment was analysed to identify the closest outgroup of the opsin family. This alignment included the complete set of 176 putative opsin outgroups I identified. Because the closest opsin outgroup must belong to the a-group of Rhodopsin-like receptors, the G&OM phylogeny was rooted using two g-group receptors: two Galanin-like receptors (Fredriksson *et al.* 2003).

To clarify the duplication and deletion history within the opsin family I analysed the O&O alignment, which I rooted using the closest opsin outgroup (identified from the results of the G&OM analyses) only. Accordingly, O&O is simply a modification of G&OM from which distantly related opsin outgroups were excluded to minimise systematic artifacts (Philippe *et al.* 2009; Sperling *et al.* 2009; Holton and Pisani 2010).

RAxML 7.2.6 (Stamatakis 2006) was used to estimate dataset specific GTR matrices for my data sets. The AIC test was then used to rank the fit to my data sets of the available empirical GTR matrices (like WAG, JTT and MtRev) and of my dataset specific GTR matrices. The difference between my GTR matrices and alternative, pre-computed, empirical GTR matrices (WAG and MtRev) was further evaluated by comparing their absolute substitution rates, and graphically displaying, for each amino acid, the difference (**Δ**-abs) between the GTR absolute substitution rate and the WAG or MtRev absolute substitution rate. Finally, 12-fold Bayesian Cross-validation, as implemented in Phylobayes 3.2 (Lartillot *et al.* 2009), was used to evaluate whether any of the precomputed CAT based models (Quang *et al.* 2008) would fit my data sets significantly better than a dataset specific GTR matrix. I thus compared the site-heterogeneous C20 + Γ, C30 + Γ, C40 + Γ, C50 + Γ, C60 + Γ, UL3+ Γ, WLSR5+ Γ (Quang *et al.* 2008) and sites homogeneous JTT+ Γ, WAG+ Γ, LG+ Γ against GTR + Γ. Because of computational limitations the 12-fold Bayesian cross validation analysis was only performed for the O&GM

and O&O data sets. Results of the cross validation analyses showed that none of the precomputed CAT-based models fit my data better than a data set specific GTR matrix.

All the analyses were performed under dataset specific GTR + Γ models in Phylobayes 3.2. For all analyses, two independent runs were performed and convergence was monitored using the maxdiff statistics calculated using the bpcomp program (see the Phylobayes manual). Analyses were considered to have converged when maxdiff dropped below 0.3 (see the Phylobayes manual). Results of the analyses of the O&O data sets where further confirmed by performing Maximum Likelihood (ML) analyses (under LG + Γ) in RAxML (Foster 2004; Stamatakis 2006) Support for the nodes in the ML phylogeny were estimated using the bootstrap (108 replicates-see page 62). ML analyses were performed under LG + Γ, rather then GTR + Γ, in order to test also the sensitivity of my results to the use of less fitting models.

Posterior Predictive analysis (PPA; implemented in Phylobayes3.2) was used to evaluate whether my data sets contained compositionally heterogeneous sequences and to evaluate whether compositional heterogeneity could have affected my results.

The Approximately Unbiased (AU) test (Shimodaira 2002) implemented using RAxML under GTR + Γ, was used to evaluate whether my data set (O&O data set – see main text) allowed to statistically discriminate between my results and those of previously published studies (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011).

I performed Bayesian and ML–based ancestral character state reconstruction for the O&O data set and recovered the ancestral retinal-binding domain for two key, internal nodes. These nodes are the one identifying the last common ancestor of all the opsins (LOCA), and the one identifying the last common ancestor of all the eumetazoan opsins (LOCNA). Bayesian Ancestral character state reconstruction was performed using MrBayes3.2 (Ronquist and Huelsenbeck 2003) under the dataset specific GTR substitution matrix I derived from the O&O data set in RAxML.

For the MrBayes analyses 2 runs of four chains were run until convergence and a burnin of 25% of the points in the chains was used. ML–based character state reconstruction was performed using PAML (Yang 2007) under GTR + Γ.

4.3 Results

Common problems with previous studies of opsin evolution (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) were the use of under sampled data sets and substitution models that might not have fit the data well (precomputed empirical GTR matrices). In addition, problems relating to outgroup selection and the adverse effect of inadequate outgroup selection on the opsin phylogeny have been pinpointed (Plachetzki *et al.* 2007), but had never been properly tackled (see above). To avoid such problems, I assembled three GPCR and opsin alignments scoring hundreds of sequences and for each of these alignments I estimated a dataset-specific GTR matrix. These matrices are substantially different from available, precomputed GTR matrices (see Figure 4.2 and Table 3 Electronic Appendix).

Using the Akaike Information Criterion (AIC) I was able to demonstrate (as expected) that my GTR matrices fit the data set from which they were inferred significantly better then any precomputed empirical GTR matrix, with LG + G as the second best fitting model (see Table 4.1). I also tested the use of site-heterogeneous empirical mixture models, but none of these models could be shown to fit my alignments significantly better than a data-set specific GTR matrix (see Table 4.2). Accordingly results obtained using these models were not considered.
Figure 3a represents the phylogeny derived from my All Opsin Master alignment (AOM; see Methods). AOM includes only neuralian opsins (no outgroups) and Figure 4.3a is thus an unrooted phylogeny of my opsin data set (see Table 2 in Electronic Appendix). Figure 4.3a (Figure 1 Electronic Appendix) is consistent with the monophyly of the traditionally recognised bilaterian opsin subfamilies (C, R and Go/RGR).

Data set	Model	Log-likelihood	AIC
GPCR & Opsin Master alignment	GTR+Γ	-78625.11026	157672.2205
	LGF+Γ	-79573.94395	159149.8879
	WAGF+Γ	-79929.86067	159861.7231
Opsins & Outgroups Alignment	GTR+Γ	-33759.69164	67941.38327
	LGF+Γ	-34197.46653	68393.93306
	WAGF+Γ	-34414.51953	68831.03907
All Opsin Master Alignment	GTR+Γ	-78875.35627	158172.7125
	LGF+Γ	-79821.76843	159625.5369
	WAGF+Γ	-80417.84551	160837.691

Table 4.1: Model selection: This analysis illustrate that data set specific GTR+Γ models fits each of my data set better than precomputed GTR model.

Dataset	Compared Models	Mean score	Stdev(+/-)
GPCR & Opsin Master alignment	C20 versus GTR	-458.095	1412.83
	C30 versus GTR	-263.715	1271.08
	C40 versus GTR	-868.692	1100.33
	C50 versus GTR	-394.801	1335.45
	C60 versus GTR	-615.158	1212.5
	JTT versus GTR	-680.076	1185.88
	LG versus GTR	-241.062	1674.67
	UL3 versus GTR	-344.796	1110.3
	WAG versus GTR	-12.2742	1463.64
Opsins & Outgroups alignment	C20 versus GTR	-440.047	543.972
	C30 versus GTR	-149.996	769.156
	C40 versus GTR	-390.47	373.572
	C50 versus GTR	107.343	608.496
	C60 versus GTR	-135.062	470.151
	JTT versus GTR	-25.1208	545.841
	LG versus GTR	-66.6875	744.275
	UL3 versus GTR	-151.482	587.579
	WAG versus GTR	-138.591	331.638

Table 4.2: Bayesian cross validation. This analysis was performed to compare the GTR+Γ models against precomputed site heterogeneous (CAT) models. The All Opsin Master alignment was not tested because of computational limitations. Note: In the cross validation a negative value implies that the reference model (GTR+Γ) is better than the tested model. Only in the case of O&O one of the heterogeneous models (C50) performs marginally better than GTR. However, for all considered models (including C50) the standard deviation around the cross validation scores is too large to claim that one of the two models fits the data better. As none of the precomputed empirical CAT models was found to fit the data significantly better than GTR+Γ, these models were not used to analyse the data.

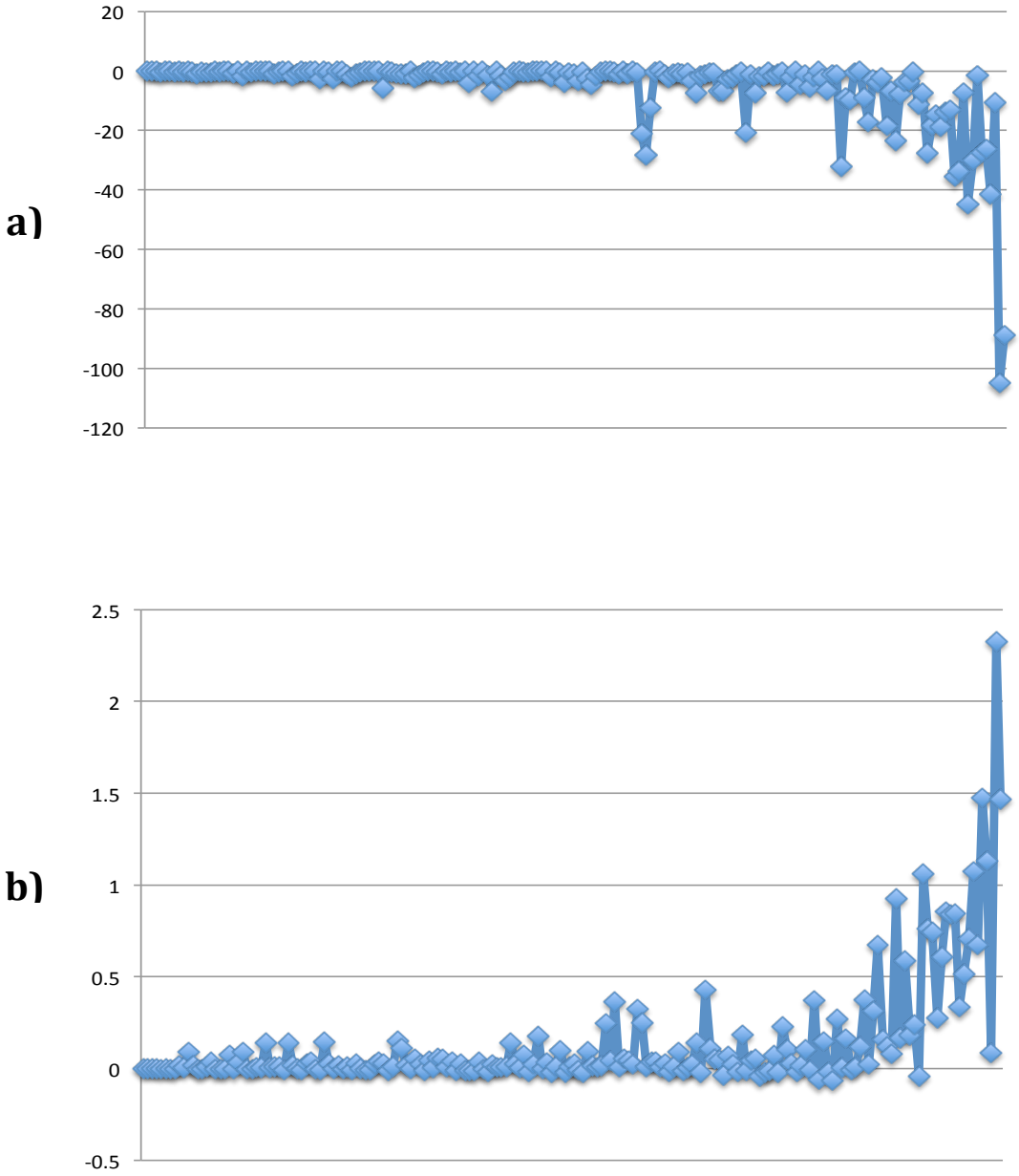


Figure 4.2: (A) A plot of the difference (Δ-abs), for each substitution in Table 3 Electronic Appendix, between the GTR-O&O and WAG global exchange rates. A value of zero means that the rate is the same in both matrices. A positive value means that the GTR- O&O global exchange rate is higher then the WAG global exchange rate. A negative value means that the GTR-O&O global exchange rate is lower than the WAG global exchange rate. (B) A plot of the difference (Δ-abs), for each substitution in Table 3 Electronic Appendix, between the GTR-O&O and mtRev global exchange rates. A value of zero means that the rate is the same in both matrices. A positive value means that the GTR-O&O global exchange rate is higher then the mtRev global exchange rate. A negative value means that the GTR-O&O global exchange rate is lower than the mtRev global exchange rate. On the X-axis: amino acid substitutions (ordered with reference to their Δ-abs – from smaller to big). On the Y-axis Δ-abs values.

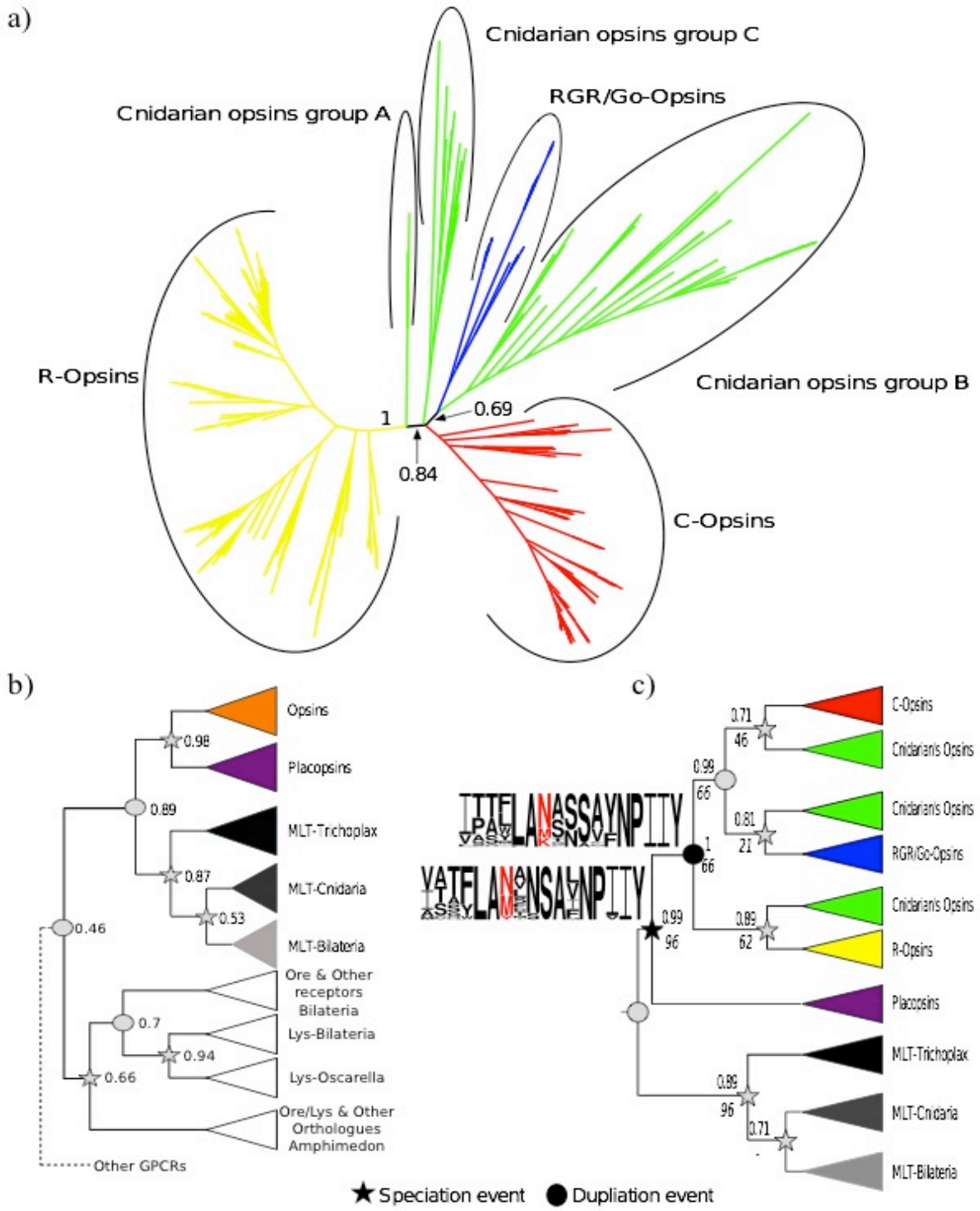


Figure 4.3: The phylogeny of the opsin family. (A) Unrooted phylogeny of the neuralian opsins. (B) Rooted phylogeny of the neuralian opsins and of other GPCRs showing that the Placopsins are members of the opsin family. Ore = Orexin; Lys = Lysosphingolipid. (C) Opsin phylogeny rooted using only the MLT receptors, and showing that cnidarians have orthologs of each bilaterian opsin subfamily: the C, R, and Go/RGR subfamilies. Support values (Bayesian posterior probabilities) are reported only for key nodes. (C) Bootstrap support values are showed in italic. The ancestral RBD of the LOCA and of the LOCNA are reported and are identified, respectively, by a black star and a black circle. The red position in the logos identify site 296.

In contrast, the Cnidarian opsins are split into three clans (which I named Group–A, –B, and –C). This is in agreement with the results of Suga and collaborators Suga *et al.* (2008) but in disagreement with Plachetzki *et al.* (2007), Plachetzki *et al.* (2010) and Porter *et al.* (2011). Group–A only includes two sequences and sits on the branch separating the R–opsins from all the other sequences in my data set (Posterior Probability – PP = 0.84). The sequences in Group–A are from the study of Suga and collaborators (Suga *et al.* 2008) where they were named Group–3. These sequences were not included in Plachetzki *et al.* (2007), Plachetzki *et al.* (2010) and Porter *et al.* 2011). Group–B form a relatively poorly supported clan with the Go/RGR opsins (PP = 0.69), while Group–C is found in a polytomy with the C–opsins and the Go/RGR plus Group–B clans (see Figure 4.3a). Group–C includes both the sequences that in the study of Suga and collaborators (Suga *et al.* 2008) emerged as the sister group of the R–opsins (their Group–2 opsins) and the single sequence that Plachetzki and colleagues (Plachetzki *et al.* 2007) classified as a C–opsin. The phylogeny shown in Figure 4.3a rejects the possibility that Suga and collaborators (Suga *et al.* 2008) Group–2 opsins could be the sister group of the bilaterian R–opsins. However, the tree in Figure 4.3a could neither confirm nor reject the C–opsin nature of Plachetzki and colleagues (Plachetzki *et al.* 2007) putative C–opsin. This is because Figure 4.3a shows that all above mentioned sequences belong to Group–C: a group that in this analysis could not be placed with confidence with reference to the C–, and the Go/RGR plus Group–B opsins, but that certainly is not the sister group of the R–opsins.

Posterior predictive analysis (Table 4 in Electronic Appendix) showed that some of the sequences in AOM were compositionally heterogeneous. Because of their skewed amino acid composition these sequences can mislead phylogenetic analyses (Foster 2004). These sequences were included in AOM for the purposes of testing to which major opsin clan they belong. However, most of these sequences have been excluded from further analyses (see below) to avoid their potentially biasing effect. Other sequences, for example short ESTs, like the putative cnidarian C–opsin of Plachetzki and colleagues (Plachetzki *et al.* 2007), were also excluded from

further analyses. This was done because in Figure 4.3a these sequences were unequivocally identified as members of one of the Cnidarian opsins clans (Group–A, –B or –C), and more complete representatives of each of these clans were retained for further analyses.

I analysed the GPCR & Opsin Master alignment (G&OM; see methods) to test what GPCR family represents the closest sister group of the opsin family (see Figure 4.3b and figure 2 Electronic Appendix). This is important to select the best possible outgroup for my opsin data set, and to elucidate the origin of the opsin family. Analyses of G&OM show that the neuralian opsins form a monophyletic group. Relationships among the major neuralian opsin clades are consistent with those of Figure 4.3a. That is, the tree in figure 4.3b is a rooted version of that in figure 4.3a. Figure 4.3b shows that the sister group of the neuralian opsins is composed of a set of placozoan "opsin–like" sequences (PP = 0.98), and that the sister group of the neuralian opsins plus the placozoan "opsin–like" sequences is represented by the MLT receptors (PP = 0.89). Figure 4.3b also shows that both the placozoans and the cnidarians have MLT receptors, and most importantly, that the placozoan "opsin–like" receptors are orthologs of the neuralian opsins. This implies that from an evolutionary point of view, the placozoans "opsin-like" receptors are members of the opsin family even though they lack a RBD with a K296 residue. Accordingly, Opsins are universally distributed within Eumetazoa (Placozoa plus Neuralia). Opsin and/or MLT receptors could not be identified in the eumetazoan outgroups (*Oscarella* and *Amphimedon*). That is, both the opsins and the MLT receptors are eumetazoan specific families, and the duplication from which they emerged happened after the split between *Oscarella* and the Eumetazoa (no matter whether the sponges are monophyletic or paraphyletic). The MLT + Opsin clade is then the sister group of the Lysosphingolipid and Orexin receptors (albeit with very low support: PP = 0.46, Figure 4.3b and Figure 2 Electronic Appendix). Both *Oscarella* and *Amphimedon* have sequences belonging to this group (see figure 4.3b; PP = 0.94). These results confirm the eumetazoan nature of the opsin family, and are in agreement with recent results showing that light sensitivity in *Amphimedon* is mediated by a cryptochrome, rather than an

opsin (Rivera *et al.* 2012).

I tested whether distant outgroups in the G&OM data set (results in figure 4.3b) could have caused tree-reconstruction artifacts with reference to the opsin ingroup topology. To do so I analysed a data set, the Opsins & Outgroups alignment (O&O – see Methods), in which the MLT receptors were used as the sole outgroups of an opsin data set that included also the placozoan "opsin–like" receptors. The Bayesian O&O phylogeny is reported in figure 4.3c (see also figure 3 in Electronic Appendix). The O&O data set was also analysed using ML (see below, and figure 4 in Electronic Appendix). Analyses of O&O confirmed the results obtained from the analysis of G&OM (compare figure 4.3b and 4.3c). To summarise, both O&O and G&OM show that the Cnidarian opsins can be classified in three groups (A, B, and C). These groups represent, respectively, the cnidarian orthologs of the bilaterian R–opsins (Group–A; PP-GTR = 0.89 and ML Bootstrap support under LG + G – BP-LG = 62%), the cnidarian orthologs of the bilaterian Go/RGR opsins (Group–B; PP = 0.81 and BP-LG < 50), and the cnidarian orthologs of the bilaterian C–opsins (Group–C; PP = 0.71 and BP-LG < 50). ML bootstrap support values for the internal opsin relationships are low. Therefore, I used the Approximately Unbiased (Shimodaira 2002) test to evaluate whether the data, under the best fitting GTR + G model, can discriminate between alternative opsin phylogenies. Results of these analyses (Table 4.3) confirm that the data can indeed discriminate between alternative opsin phylogenies, and that under my O&O–specific GTR + G model the trees in (Plachetzki *et al.* 2007; Suga *et al.* 2008; Plachetzki *et al.* 2010; Porter *et al.* 2011) fit my O&O data set significantly worse than the topology of figure 4.3c.

In order to provide further insights into opsin evolution, I carried out Bayesian and ML ancestral character state reconstruction of the RBD at key internal nodes. Results of the Bayesian analyses are reported as logos in figure 4.3c and in figure 5 in Electronic Appendix, and indicate that the Last Common Opsin Ancestor (LOCA) most likely did not have the key K296 residue (PP-K296 = 0.0034). Instead, with reasonable confidence I can say that position

296 was either occupied by an asparagine (PP for N296 = 0.51) or by a methionine (PP for M296 = 0.37). Absence of K296 in LOCA is confirmed by ML, which suggests with reasonable confidence that asparagine was the most likely amino acid in position 296 (P-N296 = 0.81 and P-K296 = 0.054). K296 is necessary to link the chromophore, and my results suggest that K296-mediated chromophore binding was not a feature of LOCA: it evolved within the opsin family. Indeed, even in the case of the last opsin common neuralian ancestor (LOCNA), the Bayesian reconstruction suggests that the RBD might not have had a K296 residue (PP for K296 = 0.15; figure 4C). However, ML contradicts this result, as it finds a P-K296 value of 0.99. This incongruence leaves the question of occupancy of position 296 in LOCNA unresolved. No matter what the amino acid in LOCNA was, my results strongly suggest that a K296-based RBD was not a feature of LOCA. If that were the case then K296-mediated retinal binding would be the result of a functional parallelism in the C– plus Go/RGR–opsins and in the R–opsin. However, ML-based character state reconstruction suggests the RBD of the LOCNA had a K296 residue (P =0.99), leaving the question of occupancy of position 296 in the LOCNA substantially unresolved. No matter whether K296 originated once or twice independently, my results suggest that a RBD with a K296 residue was not a feature of the LOCA.

Hypotheses	Probability
Fig 4.3c	0.7
Plachetzki et al. (2007)	0.04
Porter et al. 2011	0.03
Plachetzki et al. (2010)	0.008
Suga et al. (2008)	5e-18

Table 4.3: Results of the AU tests

4.5 Discussion

My results are markedly different from those of previous investigations. These differences reflect data and methodological dissimilarities between my study and previous ones. I used a combination of recently developed multiple sequence alignment software that can better differentiate between insertions and deletions, extensive model selection analyses resulting in the use of significantly better fitting substitution models and I was careful to include the closest outgroups of the eumetazoan opsins (including sequences from the Placozoa). Finally, and probably most importantly, I used a very inclusive set of cnidarian opsins allowing for the simultaneous test of the hypotheses of Plachetzki *et al.* (2007), Suga *et al.* (2008), Plachetzki *et al.* (2010) and Porter *et al.* (2011). Because previous studies, with the exception of Suga *et al.* (2008), did not include all these key opsins, they did not have the power to discriminate among all possible scenarios of opsin evolution.

My results (summarised in figure 4.3) allow for a substantial clarification of the *tempo and mode* of opsin evolution. They confirm the results of Fredriksson *et al.* (2003) that the sister group of the opsin family is represented by the MLT receptors, and they show that the opsin family originated from the duplication of the MLT plus opsin ancestral gene in the stem eumetazoan lineage. An important result of my study is that I could show that the placozoan genome contains sequences that are in an orthologs relationship with the eumetazoan opsins and therefore, from an evolutionary point of view, they are members of the opsin family (Figure 4.3b), irrespective of whether they have the ability to detect light or not. I thus propose to refer to these "opsin-like" receptors as "Placopsins". In addition, I show for the first time that cnidarians most likely have R–, Go/RGR– and C–opsin orthologs. Accordingly, these opsin subfamilies evolved in the stem neuralian lineage, rather then in the stem bilaterian lineage: that is, earlier than is currently accepted. My results are largely phylogeny-independent. Nonetheless, uncertainty in the placement of the Placozoa still persists and deserves discussion. Consistently with my results, some of the most thorough analyses to date (Philippe *et al.* 2009; Sperling *et al.*

2009) agree that the Placozoa are the sister group of Neuralia, even though Schierwater *et al.* (2009) and Pick *et al.* (2010) found different results. However, Philippe and collaborators (Philippe *et al.* 2011) have shown the results of Schierwater *et al.* (2009) to be invalid. Differently, despite the study of Pick and collaborators (Pick *et al.* 2010) is sound; its conclusion that Placozoa is a member of Neuralia is questionable. This is because the data set of Pick *et al.* (2010) is based on that of Dunn *et al.* (2008), which has been shown to be unreliable (Philippe *et al.* 2009; Pick *et al.* 2010; Philippe *et al.* 2011). Importantly, even if Bilateria and Placozoa were confirmed to be sister groups Pick *et al.* (2010), my results would still be valid, but my scenario would become less parsimonious as it would imply independent losses of the placopsin in Bilateria and Cnidaria, and of the C–, R– and Go/RGR opsins in Placozoa.

Ancestral character state reconstruction suggests the LOCA (i.e. the ancestor of the placopsins and of the neuralian opsins) did not have a RBD containing a K296 residue. Accordingly, K296–mediated light detection might have evolved in the stem Eumetazoan lineage through autogeneous evolution and neofunctionalisation of a protein that was not light sensitive. Neither of the two sponge taxa considered in this study has MLT or opsin receptors. This implies that the opsins evolved after the split between the Eumetazoa and both the demosponges and the homoscleromorph sponges. However, figure 4.3b shows that both considered sponges have receptors belonging to the clade representing the sister group of the MLT plus Opsin group. Overall these results confirm that the first opsin originated in the stem neuralian lineage. In addition, they imply that my conclusions are robust irrespective of whether sponges are monophyletic (Philippe *et al.* 2009) or paraphyletic (Sperling *et al.* 2009).

Identification of the duplication of the ancestral MLT plus Opsin gene in the stem eumetazoan lineage lets us better constrain the timing of this event as this lineage was dated to have existed between 755 and 711 Ma (Erwin *et al.* 2011). In addition, the neuralian stem lineage was dated to have existed between 711 and 700 Ma (Erwin *et al.* 2011). This relatively short time (11 million years) was a crucial period in opsin evolution, because it was during this

time that the K296-based RBD most likely evolved, and the duplications separating the C– plus Go/RGR–opsin ancestor from the R–opsins, and the C– from the Go/RGR–opsins were fixed.

My results suggest that the Go/RGR–opsins represent the sister group of the C–opsins. This is in disagreement with (Plachetzki *et al.* 2007; Plachetzki *et al.* 2010; Porter *et al.* 2011), but is in agreement with (Terakita 2005; Suga *et al.* 2008) among others. An additional line of evidence that seems to support my conclusion is that the Go/RGR–opsins, in the same manner as the C–opsins, are expressed in ciliary receptors (Fain *et al.* 2010; Porter *et al.* 2011). My results also predict that Rhabdomeric receptors should exist in Cnidaria. This has not yet been proven but cells with a strong resemblance to the bilaterian rhabdomeric receptors, and that could be cnidarian rhabdomeric receptors, have been observed in cnidarian larvae (Nordstrom *et al.* 2003; Fain *et al.* 2010; Gehring 2011).

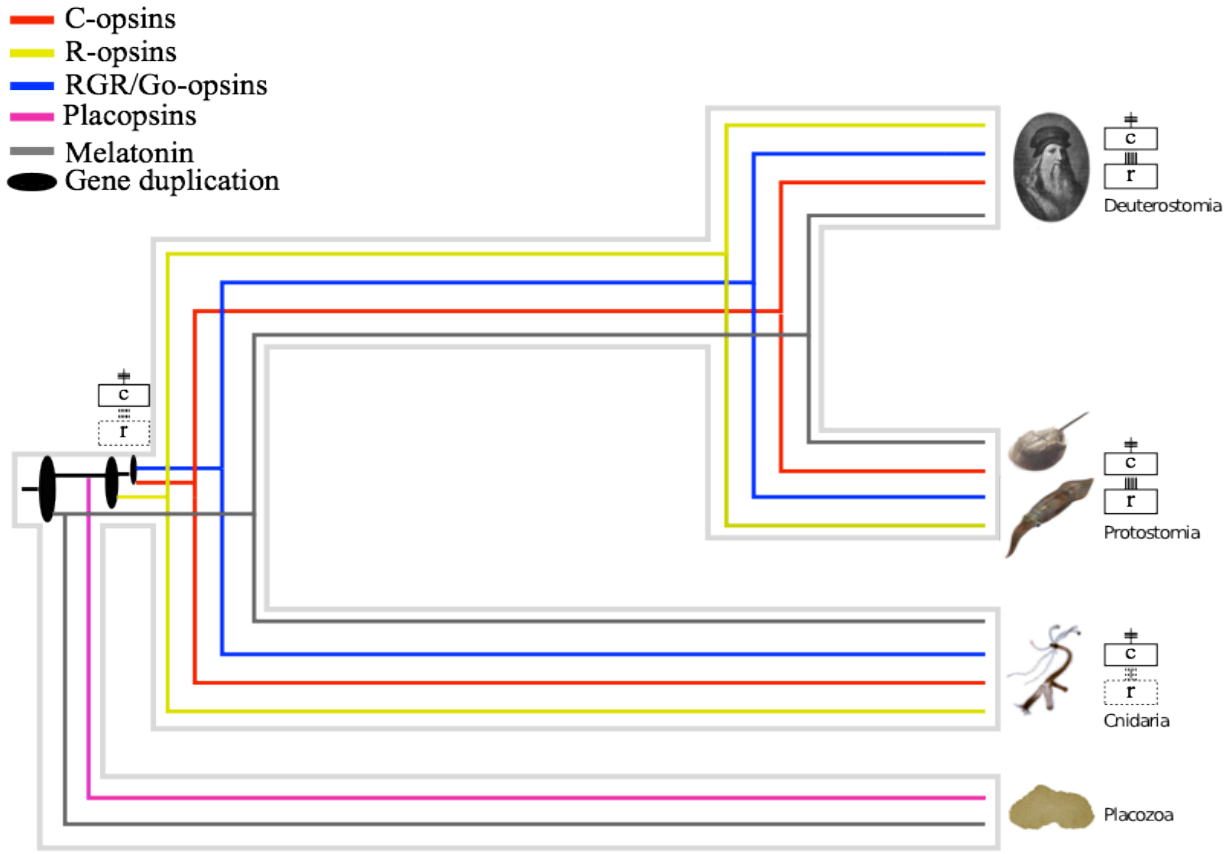


Figure 4.4: A synopsis of the opsin evolutionary history**.** This figure represent a gene tree embedded within a species tree illustrating the evolutionary history of the opsins and MLT receptors within Metazoa. It shows that only 3 duplications and no deletions are necessary to explain the origin and evolution of the opsin family.

4.6 Conclusion

I suggest a novel, early, and very parsimonious explanation for the diversification of the opsin family (summarised in figure 4.4), and show that the LOCA most likely did not have a RBD with a K296 residue. Scarcity of signal for the deepest event in the history of the opsin family implies that some level of uncertainty in opsin evolution still remains, and might be unavoidable. However, results of the AU tests show that the topology uncovered in this study fits the data (under a GTR + G model) significantly better than any previously proposed opsin phylogeny. My results also indicate that a short 11 million year period (711-700 Ma) was key in opsin evolution. During this time, two duplications in the stem neuralian lineage resulted in the evolution of the extant opsin paralogs. During this same time, a K296-based RBD most likely evolved, probably through a process of neofunctionalisation. From a point of view eye evolution my results suggest a monophyletic origin of this complex structure with the common ancestor of neuralia that posses both the photoreceptor and a single multifunctional cell (*sensu* Arendt 2008). In agreement with previous studies (e.g. Erwin 2009), my results are compatible with the view that the last common eumetazoan ancestor might have been more complex than it is generally thought. More precisely, it has been suggested by Erwin (2009) that current evidence (including the existence of muscular fibres in anthozoans), suggests that extant Cnidaria are simplified organisms. Indeed, the existence of a cnidarian eye lead to ward the same conclusion. My results can be extend to address the root of the animal tree and the origin of this complexity. As suggested by Erwin (2009) my results are compatible with a view were the existing Cnidarian species represent the remnant of a previously very successful animal phylum with a variety of body forms and complex morphologies. This thought provoking idea has never been tested and is essentially untestable. Yet, this hypothesis is appealing as it would provide a sensible framework to allow the interpretation of complex ediacaran fossils (like the various frondose and triradiate ones) that have been impossible to classify up to now. I suggest that the results of Erwin are indeed likely to be correct that and the Precambrian might have been dominated by a

variety of cnidarians of complex and different morphologies (including bilaterally symmetrical ones), with complex sensory systems.

Chapter 5

General discussion

Undoubtedly genome scale analyses have radically changed the current prospective in evolutionary problems. On one hand, they have allowed clarification of long-standing problems (e.g Holton and Pisani 2010), while on the other, they posed many new challenging questions (e.g. McInerney *et al.* 2011).

The study of animal evolution has not been immune to this radical change of prospective. ESTs sequences and complete genomes provide raw material for a substantial clarification of key aspects of early animal evolution. For example, the analysis of complete animal genomes (particularly those of the cnidarians *Hydra* and *Nematostella*) illustrated that the neuralian eye development is homologous (Gehring 2011) and that as a consequence there has been one single origin of the eye and vision. At the same time, it is clear that comparative genomics is no *panacea*. Simply increasing the dimension of the dataset analysed is not enough to solve difficult questions. Paraphrasing Philippe *et al.* (2011), if looking for a phylogenetic tree is like looking for a needle in a haystack, comparative genomics has simply made the haystack bigger. The message that Philippe and co-workers tried to convey is clear, as the amount of data increases, the problem to be solved becomes more complex, and more sophisticated analytical tools are needed to address it. This is particularly true in the case of phylogenomic analyses (see Chapter 2), as the most pervasive pitfalls of molecular phylogenetics (long branch attractions and compositional attractions) are positively misleading and will increase in strength as the amount of analysed data increases (Jeffroy *et al.* 2006).

Indeed, the realisation that comparative genomics was not going to be a *panacea* was a rude awakening, and since the publication of Jeffroy *et al.* (2006) the initial hope that scientists had, that genomes might have solved all remaining problems in evolutionary biology (Gee

2003), has been abandoned. Obviously the mistrust is not with the genomic data which arguably contain the information to solve most evolutionary problems (Zuckerkandl and Pauling 1965), but with the results of actual analyses, which have been shown to be quite frequently erroneous due to positively misleading tree reconstruction artifacts (e.g. Campbell *et al.* 2011; Philippe *et al.* 2011; Rota-Stabelli *et al.* 2011 and Chapter 2 of this thesis). Indeed, methods of analysis have re-emerged as tools of paramount importance in the genomic era, and the three chapters of results presented in this work provide additional evidence for the strong impact that alternative ways to analyse the data can have on the results of genomic-scale analyses. In particular, I have shown here that the use of more sophisticated methods and models can improve the resolution of the relationships among the animal lineages (Chapter 2), among proteins that share the same architecture (Chapter 3), and finally between opsins paralogs (Chapter 4). Each of the three result chapters has specific implications that I have discussed within the individual chapters. The general discussion I am presenting here has the sole aim of identifying what these results imply more broadly, with reference to early animal evolution.

However, before discussing the implications of the results presented in this thesis to my understanding of animal evolution, I would like to discuss some methodological considerations that permeate the thesis.

5.1 Better methods and more sequences

Since its inception, phylogenomics (Eisen 1998) has been successfully applied to clarify long-standing problems in evolutionary biology (Rokas *et al.* 2003). However, analysing genomic scale data sets soon turned out to be a methodological nightmare, and the inadequacy of increasing gene sampling alone, was soon noted (Jeffroy *et al.* 2006; Philippe *et al.* 2011). It is now almost universally accepted that increasing gene sampling does not automatically result in more reliable trees. The use of multi-gene datasets has reduced the impact of the stochastic

errors, but it has exacerbated the impact of systematic ones (Jeffroy *et al.* 2006).

Systematic errors unfortunately affect gene phylogenies also. A multitude of sources of such errors have been identified during the last decade (Jeffroy *et al.* 2006; Philippe *et al.* 2011; Roure *et al.* 2012). Within this long list, my results in Chapter 2 and 4 suggest that a key role is played by misalignments, inadequate outgroup selection, incorrect ortholog selection and compositional heterogeneity, all of which are frequently underestimated issues in tree reconstruction. My results confirm previous observations (Rota-Stabelli *et al.* 2011) that the use of phylogenomic data sets should always be complemented with a regular exploration of the sources of non-phylogenetic signal. Another powerful approach, especially when the question at hand has been shown to be particularly hard to resolve, is to investigate the congruence of alternative lines of evidence e.g. microRNAs and gene synteny.

Alignment problems (see above) probably require further discussions since the "quick and fast" (Boussau and Daubin 2010) approach of aligning sequences seems to be widespread. The influence of the alignment software on phylogenetic inference is well documented (Wong *et al.* 2008). As explained in the introduction, the explosion of bioinformatics has seen the multiplications of alignment software, each with individual strengths and weaknesses. At the same time, it is clear that under certain conditions (i.e. alignability) there are software, such as Prank (Loytynoja and Goldman 2008), that provide a more realistic alignment, and this coincides with the ability to capture the *bona fide* phylogenetic signal. In Chapter 2 and 4, the alignment improvement is one of the factors that could explain the differences obtained between my hypotheses and previous ones.

Additionally, the results of Chapter 3 suggest that the use of traditional phylogenetic methods, which rely on global alignments, are not applicable to the study of highly divergent protein families. It is clear that the ability to reconstruct reliable alignments is one of the limiting factors for these problems. There are a variety of tools that can be used as alternatives to global alignment methods. Among these tools, clustering methods, such as MCL (Enright *et al.*

2002), are the most used. In this approach, a similarity matrix (scoring BLAST results) is transformed through a weight matrix and subsequently a random walk conducted to separate gene sub-families that are then independently analysed. However, this process of compartmentalization in MCL removes evidence of non-tree like processes such as domain shuffling and protein fusions. For these reasons, and to overcome this problem, in Chapter 3 I used similarity networks derived directly from all-versus-all BLAST analyses to reconstruct phylogenetic relationship among proteins with 7TMDs. This approach relies only on local (pairwise) alignments and consequently eliminates biases related to inappropriate homology definitions, whilst retaining information about horizontal evolutionary processes. In my opinion, this feature makes this approach extremely attractive for studying highly divergent protein superfamilies (such as the 7TMD one). My results demonstrate that visualizing sequence similarity networks allows both vertical and horizontal relationships to be represented and that these relationships correlate well with known functional relationships (Chapter 3).

5.2 The evolution of the early animals

The results in Chapter 2 suggest, as expected, the monophyly of animals. In addition, they confirm the Ctenophora as a close relatives of the Bilateria and of the Cnidaria, whilst rejecting a possible sister group relationship between the Ctenophora and all the other animals (Dunn *et al.* 2008; Hejnol *et al.* 2009). In addition, in Chapter 2 I found, for the first time, support for the Epitheliozoa hypothesis (Sperling *et al.* 2007; Sperling *et al.* 2009; Sperling *et al.* 2010) using an EST data set. Because this topology, that rejects the monophyly of the traditionally defined Porifera, is the result of a series of methodological improvements and enhancements in data quality, I conclude, in agreement with (Sperling *et al.* 2009), that sponge monophyly (Philippe *et al.* 2009; Pick *et al.* 2010) is most likely a phylogenetic artifact. From an ecological point of view the topology suggested by the analyses performed in Chapter 2 imply that the last common

ancestor of all living animals was a benthic, sessile microsuspension-feeding organism (Sperling *et al.* 2007). Finding the Ctenophora on a crownward position within the tree, rather then as the sister group of all the other animals, is also important, as it allows for a more parsimonious interpretation of the evolution of animal morphology, and for a more derived origin of predation (Ctenophora being predators). Indeed, it is difficult to imagine that the first split within the animal tree (as suggested by Dunn *et al.* 2008) separated a carnivorous lineage (Ctenophora) from a clade composed of all the other animals (and with a common ancestor that most likely was a filter feeder).

The analysis performed in Chapter 2 using the optimal outgroups (fig 2.7), which is the analysis that is most likely to have returned a correct result, suggests that the most likely position of the Ctenophora is as the sister group of the Cnidaria within the Coelenterata. Because Cnidaria are mostly diploblastic, whilst the Ctenophora and the Bilateria are triploblastic, the Coelenterata hypothesis implies that the mesoderm evolved independently two times (in Bilateria and Ctenophora). Alternatively, it might have evolved in the neuralian ancestor (with the Cnidaria being secondarily simplified). I suggest that the second hypothesis is more likely as, histologically identical striated muscles exist in the entocodon of the hydromedusae and in Bilateria. In addition, Cnidaria express 'mesodermal' genes, and coelom-like structures exist in the hydromedusan subumbrellar structure (Erwin 2009).

If my conclusions are correct, then the last common ancestor of the Neuralia (Cnidaria, Bilateria and Ctenophora) possessed, as pointed out by Erwin 2009), the toolkit for bilaterality, triploblasty, and at least some elements of mesodermal muscle development. In addition (Chapter 4), it possessed all known bilaterian opsins and thus a fairly complex visual system. This implies that extant Coelenterata (particularly Cnidaria), with their simple morphologies and radial symmetry, are highly simplified organisms that as suggested by Erwin 2009), might represent the remnant of a once, much more successful independent animal radiation. More generally, it is clear that the picture emerging is that the last common neuralian ancestor was far

more complex than currently imagined.

The sister group of the Neuralia has previously been suggested to be represented by the Placozoa. However, in Chapter 2 I was not able to cluster with certainty this phylum, as its only representative (*Trichoplax*) was unstable. This result seems to be in line with those of other studies, where the placozoans have been found to be the sister group of a multitude of alternative lineages (Philippe *et al.* 2009; Sperling *et al.* 2009; Pick *et al.* 2010; Philippe *et al.* 2011). The lack of a nervous system, digestive system, symmetry and their extreme "simplicity" suggest that the most likely position for the Placozoa is not inside the Neuralia, as proposed by Pick *et al* (2010), but as the sister group of this lineage (see Chapter 2 and Philippe *et al.* 2009; Sperling *et al.* 2009).

The distribution of 7TMD receptors has been used in this work to make sense of the diversity observed in the organization of the early metazoans. In Chapter 3, it has been shown that proteins with a 7TMD architecture (including the GPCRs) have multiple origins. Whether 7TMD were present in the last eukaryotic ancestor is questionable, but possible (see Chapter 3). In any case, the 7TMD underwent a protein super-family expansion in the stem neuralian lineage, probably in association with the origin of the nervous system. However, rather than acting during development (body plan formation), the big 7TMD expansion has increased the physiological potential of these animals, allowing for cross signal integration between highly specialized cells and the origin of information processing. This 7TMD expansion in Neuralia correlates well with the level of complexity observed in these animals, as sponges are much more simple and do not have many 7TMD. Notably, 7TMD variability is not only quantitative but also qualitative, as it mostly involves an expansion of the Rhodopsin-like superfamily in Neuralia (which are mostly expressed in the nervous system).

In Chapter 4, I show that the opsins arose from a group of opsins-like GPCRs around 700 millions of years ago. Since then, animal ecology has changed dramatically. Being able to detect light has deeply changed the evolutionary history of the animals. Furthermore, in Chapter 4, I

suggest that the common ancestor of the Neuralia, had a complex visual system and expressed R, C and Go opsins. This result together with the unusual distribution of GPCRs in *Nematostella*, and the diversity of transcription factors and signalling pathway genes in Cnidaria, suggest the complex nature of at the neuralian common ancestor (Erwin 2009).

Chapter 6

Future prospective

Understanding how animal diversity and complexity arose is one of the key challenges ahead of evolutionary biology. Genomics has provided a substantial clarification of the relationships between living organisms and at the same time allowed investigating at the molecular level the differences between such organisms. It is clear that these two levels of investigation are consequential since a robust phylogeny of species is the first mandatory step to polarize character evolution. In other words, as suggested by Nielsen (2012), the phylogenetic trees are *naked.* It is the morphological, and genomic differences that should be explained. In this context, phylogenetic trees are powerful tools but at the same time, explanation of all the aspects of the living organisms are necessary to make evolutionary reconstruction non trivial. Some aspects of morphological variation have been shown in this work to correlate with an increase in the number of 7TMD. This protein family acts by integrating signals between the inner-modules of the organisms. Furthermore, the evolution of opsins ~711 Mya, and the ability to detect light, had a strong impact in the evolutionary history of the animals.

However, it is clear that the genomic program of the organism is encoded at a different level in the genome. This program is a complex network composed by the interactions of cis-regulatory elements and transcription factors (Peter and Davidson 2011).

Davidson and Erwin (2006) have defined a hierarchical structure for gene regulatory networks. Some elements, termed kernels, are composed of associations of genes with recursive expression patterns dedicated to basic functions. Other elements include, for example, plug-ins, which contain sub circuits that are dedicated to producing functional units or modules, others, such as, signalling cassettes are commonly found to serve in multiple pathways and finally

largely non-regulatory batteries (i.e. gene batteries), composed of structural genes, are found at the periphery of the networks. The expression of the genes in the batteries acts to differentiate cells, tissues and organs. Alteration of the architecture of the kernel, plug-ins, and gene batteries explain differences among different levels of the Linnaean hierarchy (Davidson and Erwin 2006).

Erwin and collaborators (Erwin *et al.* 2011) have proposed several factors that could explain the increasing morphological complexity and developmental stability of bilaterian lineages: (1) an increase in the diversity and number of GRN subcircuits, (2) the continued and hierarchical incorporation of miRNAs into these networks in a lineage-specific manner (3) other forms of RNA regulation, such as alternative splicing of transcripts, and combinatorial control of enhancers.

Identifying the components and then resolving the architecture of the developmental network, will explain the observed differences between extant animals. Integrating several fields of evolutionary biology will make it possible to understand how these changes took place. Furthermore, resolving the space-time structure of the network will allow to make predictions using in *silico* methods, as proposed by Peter *et al.* (2012). This change of prospective will require an increment in the number of genomic samples available for the basal metazoans, and as proposed by Jenner and Wills (2007) an increase in the phylogenetic coverage of animals (i.e. data for non-model systems will be necessary). Finally and probably mostly importantly, this shift will require a change in prospective, as it will be necessary to start looking at organisms as integrated protein-protein interaction networks rather than as sum of genes.

With this thesis, I hope I have been able to increase our understanding of animal evolution, but there is still much that needs to be done and further work that needs to be completed. Understanding the position of the Placozoans is certainly one topic for this further work. A better understanding evolutionary dynamics of the 7TMD is a second one. Finally, I suggest that a further focus on Opsin evolution to understand specific differences between R and

C opsins would be necessary. For example, an important aspect to investigate is what were the original functions of GPCRs expressed in LECA. All these are interesting projects that still lay ahead of us (and me).

Chapter 7

Bibliography

References

Abascal, F, R Zardoya, MJ Telford. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38:W7-13.

Adl, SM, AG Simpson, CE Lane, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol 59:429-514.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: BN IN PETROV, F CSAKI, editors. Proceedings 2nd International Symposium on Information Theory. Budapest: Akademia Kiado.

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

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

Alvarez-Ponce, D, JO McInerney. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. Genome Biol Evol 3:782-790.

Arendt, D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet 9:868-882.

Atkinson, HJ, JH Morris, TE Ferrin, PC Babbitt. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. Plos One 4:e4345.

Baldauf, SL. 2003. The deep roots of eukaryotes. Science 300:1703-1706.

Beaumont, MA, B Rannala. 2004. The Bayesian revolution in genetics. Nat Rev Genet 5:251-261.

Bjarnadottir, TK, R Fredriksson, HB Schioth. 2007. The adhesion GPCRs: a unique family of G protein-coupled receptors with important roles in both central and peripheral tissues. Cell Mol Life Sci 64:2104-2119.

Bockaert, J, JP Pin. 1999. Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J 18:1723-1729.

Boussau, B, V Daubin. 2010. Genomes as documents of evolutionary history. Trends Ecol Evol 25:224-232.

Brinkmann, H, H Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-825.

Browne, MW. 2000. Cross-Validation Methods. J Math Psychol 44:108-132.

Brusca, RC, GJ Brusca. 2003. Invertebrates. Sunderland, MA: Sinauer.

Campbell, LI, O Rota-Stabelli, GD Edgecombe, T Marchioro, SJ Longhorn, MJ Telford, H Philippe, L Rebecchi, KJ Peterson, D Pisani. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc Natl Acad Sci U S A 108:15920-15924.

Capella-Gutierrez, S, JM Silla-Martinez, T Gabaldon. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Carr, M, BS Leadbeater, R Hassan, M Nelson, SL Baldauf. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. Proc Natl Acad Sci U S A 105:16641-16646.

Carroll, SB. 2001. Chance and necessity: the evolution of morphological complexity and diversity. Nature 409:1102-1109.

Chan, CX, MA Ragan. 2013. Next-generation phylogenomics. Biol Direct 8:3.

Chapman, JA, EF Kirkness, O Simakov, et al. 2010. The dynamic genome of Hydra. Nature 464:592-596.

Chenna, R, H Sugawara, T Koike, R Lopez, TJ Gibson, DG Higgins, JD Thompson. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497-3500.

Conant, GC, KH Wolfe. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938-950.

Cotton, JA, JO McInerney. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci U S A 107:17252-17255.

Criscuolo, A, S Gribaldo. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210.

Cummins, CA, JO McInerney. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Systematic Biology 60:833-844.

Dagan, T. 2011. Phylogenomic networks. Trends Microbiol 19:483-491.

Davidson, EH, DH Erwin. 2006. Gene regulatory networks and the evolution of animal body plans. Science 311:796-800.

Dayhoff, M, R Schwartz, B Orcutt. 1978. A model of evolutionary change in protein. Atlas of Protein Sequences and Structure 5:345-352.

de Queiroz, A, J Gatesy. 2007. The supermatrix approach to systematics. Trends Ecol Evol 22:34-41.

Degnan, BM, M Vervoort, C Larroux, GS Richards. 2009. Early evolution of metazoan transcription factors. Curr Opin Genet Dev 19:591-599.

Delsuc, F, H Brinkmann, H Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-375.

Derelle, R, BF Lang. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol 29:1277-1289.

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

Doherty, A, D Alvarez-Ponce, JO McInerney. 2012. Increased Genome Sampling Reveals a Dynamic Relationship between Gene Duplicability and the Structure of the Primate Protein-Protein Interaction Network. Mol Biol Evol.

Dunn, CW, A Hejnol, DQ Matus, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.

Eddy, SR. 2004a. What is a hidden Markov model? Nat Biotechnol 22:1315-1316.

Eddy, SR. 2004b. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22:1035-1036.

Edgar, RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar, RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Edgar, RC, S Batzoglou. 2006. Multiple sequence alignment. Curr Opin Struct Biol 16:368-373.

Eisen, JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163-167.

Embley, TM, W Martin. 2006. Eukaryotic evolution, changes and challenges. Nature 440:623-630.

Enright, AJ, S Van Dongen, CA Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575-1584.

Erwin, DH. 2009. Early origin of the bilaterian developmental toolkit. Philosophical Transactions of the Royal Society B-Biological Sciences 364:2253-2261.

Erwin, DH, M Laflamme, SM Tweedt, EA Sperling, D Pisani, KJ Peterson. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science 334:1091-1097.

Fain, GL, R Hardie, SB Laughlin. 2010. Phototransduction and the evolution of photoreceptors. Current Biology 20:R114-124.

Felsenstein, J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology 27:401-410.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

Felsenstein, J. 2004. Inferring phylogenies. Sunderland: Sinauer Associates.

Fisher, R. 1912. On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society London, Series A 222:309-368.

Fitch, WM. 2000. Homology a personal view on some of the problems. Trends Genet 16:227-231.

Foster, PG. 2004. Modeling compositional heterogeneity. Systematic Biology 53:485-495.

Fredriksson, R, MC Lagerstrom, LG Lundin, HB Schioth. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63:1256-1272.

Fredriksson, R, MC Lagerstrom, HB Schioth. 2005. Expansion of the superfamily of G-protein-coupled receptors in chordates. Ann N Y Acad Sci 1040:89-94.

Fu, L, B Niu, Z Zhu, S Wu, W Li. 2012. CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics.

Fuhrman, JA, MS Schwalbach, U Stingl. 2008. Proteorhodopsins: an array of physiological roles? Nat Rev Microbiol 6:488-494.

Gee, H. 2003. Evolution: ending incongruence. Nature 425:782.

Gehring, WJ. 2011. Chance and necessity in eye evolution. Genome Biol Evol 3:1053-1066.

Goldman, N, JP Anderson, AG Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Systematic Biology 49:652-670.

Goodman, SN. 1999. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 130:1005-1013.

Gribaldo, S, AM Poole, V Daubin, P Forterre, C Brochier-Armanet. 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? Nat Rev Microbiol 8:743-752.

Grus, WE, P Shi, J Zhang. 2007. Largest vertebrate vomeronasal type 1 receptor gene repertoire in the semiaquatic platypus. Mol Biol Evol 24:2153-2157.

Guindon, S, JF Dufayard, V Lefort, M Anisimova, W Hordijk, O Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59:307-321.

Hashiguchi, Y, M Nishida. 2007. Evolution of trace amine associated receptor (TAAR) gene family in vertebrates: lineage-specific expansions and degradations of a second class of vertebrate chemosensory receptors expressed in the olfactory epithelium. Mol Biol Evol 24:2099-2107.

Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97.

Hejnol, A, M Obst, A Stamatakis, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci 276:4261-4270.

Hirt, RP, JM Logsdon, Jr., B Healy, MW Dorey, WF Doolittle, TM Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580-585.

Hjort, K, AV Goldberg, AD Tsaousis, RP Hirt, TM Embley. 2010. Diversity and reductive evolution of mitochondria among microbial eukaryotes. Philos Trans R Soc Lond B Biol Sci 365:713-727.

Holton, TA, D Pisani. 2010. Deep Genomic-Scale Analyses of the Metazoa Reject Coelomata: Evidence from Single- and Multigene Families Analyzed Under a Supertree and Supermatrix Paradigm. Genome Biology and Evolution 2:310-324.

Hrdy, I, RP Hirt, P Dolezal, L Bardonova, PG Foster, J Tachezy, TM Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618-622.

Innan, H, F Kondrashov. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97-108.

Jeffroy, O, H Brinkmann, F Delsuc, H Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225-231.

Jenner, RA, MA Wills. 2007. The choice of model organisms in evo-devo. Nat Rev Genet 8:311-319.

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

Kall, L, A Krogh, EL Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027-1036.

Kass, R, A Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773-795.

Kaupp, UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci 11:188-200.

Keane, TM, CJ Creevey, MM Pentony, TJ Naughton, JO McLnerney. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol 6:29.

Keeling, PJ, JD Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605-618.

Kemena, C, C Notredame. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25:2455-2465.

King, N. 2004. The unicellular ancestry of animal development. Dev Cell 7:313-325.

King, N, MJ Westbrook, SL Young, et al. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451:783-788.

Klare, JP, VI Gordeliy, J Labahn, G Buldt, HJ Steinhoff, M Engelhard. 2004. The archaeal sensory rhodopsin II/transducer complex: a model for transmembrane signal transfer. FEBS Lett 564:219-224.

Kolakowski, LF, Jr. 1994. GCRDb: a G-protein-coupled receptor database. Receptors Channels 2:1-7.

Koonin, EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol 11:209.

Koyanagi, M, K Takano, H Tsukamoto, K Ohtsu, F Tokunaga, A Terakita. 2008. Jellyfish vision starts with cAMP signaling mediated by opsin-G(s) cascade. Proc Natl Acad Sci U S A 105:15576-15580.

Kozmik, Z, J Ruzickova, K Jonasova, et al. 2008. Assembly of the cnidarian camera-type eye from vertebrate-like components. Proc Natl Acad Sci U S A 105:8989-8993.

Kratz, E, JC Dugas, J Ngai. 2002. Odorant receptor gene regulation: implications from genomic organization. Trends Genet 18:29-34.

Krautwurst, D. 2008. Human olfactory receptor families and their odorants. Chem Biodivers 5:842-852.

Krishnan, A, MS Almen, R Fredriksson, HB Schioth. 2012. The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. Plos One 7:e29817.

Kuck, P, K Meusemann. 2010. FASconCAT: Convenient handling of data matrices. Mol Phylogenet Evol 56:1115-1118.

Land, M, DE Nilsson. 2002. Animal eyes: Oxford University Press.

Lartillot, N, T Lepage, S Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288.

Lartillot, N, H Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Le Crom, S, M Kapsimali, PO Barome, P Vernier. 2003. Dopamine receptors for every species: gene duplications and functional diversification in Craniates. J Struct Funct Genomics 3:161-176.

Le, SQ, O Gascuel. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-1320.

Liebeskind, BJ, DM Hillis, HH Zakon. 2011. Evolution of sodium channels predates the origin of nervous systems in animals. Proc Natl Acad Sci U S A 108:9154-9159.

Loytynoja, A, N Goldman. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632-1635.

Lyons-Weiler, J, GA Hoelzer, RJ Tausch. 1998. Optimal outgroup analysis. Biological Journal of the Linnean Society 64.

Mallatt, J, CW Craig, MJ Yoder. 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. Mol Phylogenet Evol 55:1-17.

Margush, T, F McMorris. 1981. Consensus n-trees. Bulletin of Mathematical Biology 43:239-244.

Marinissen, MJ, JS Gutkind. 2001. G-protein-coupled receptors and signaling networks: emerging paradigms. Trends Pharmacol Sci 22:368-376.

McInerney, JO, D Pisani, E Bapteste, MJ O'Connell. 2011. The Public Goods Hypothesis for the evolution of life on Earth. Biol Direct 6:41.

Metropolis, N, A Rosenbluth, M Rosenbluth, A Teller, E Teller. 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21:1087.

Mikhailov, KV, AV Konstantinova, MA Nikitin, et al. 2009. The origin of Metazoa: a transition from temporal to spatial cell differentiation. Bioessays 31:758-768.

Moriyama, EN, PK Strope, SO Opiyo, Z Chen, AM Jones. 2006. Mining the Arabidopsis thaliana genome for highly-divergent seven transmembrane receptors. Genome Biol 7:R96.

Nathans, J, DS Hogness. 1983. Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. Cell 34:807-814.

Nichols, SA, BW Roberts, DJ Richter, SR Fairclough, N King. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/beta-catenin complex. Proc Natl Acad Sci U S A 109:13046-13051.

Nielsen, C. 2008. Six major steps in animal evolution: are we derived sponge larvae? Evol Dev 10:241-257.

Nielsen, C. 2012. Animal Evolution: Interelationship of the living phyla. United States: Oxford.

Nilsson, DE, L Gislen, MM Coates, C Skogh, A Garm. 2005. Advanced optics in a jellyfish eye. Nature 435:201-205.

Nordstrom, K, R Wallen, J Seymour, D Nilsson. 2003. A simple visual system without neurons in jellyfish larvae. Proceedings of the Royal Society of London Series B-Biological Sciences 270:2349-2354.

Nordstrom, KJ, M Sallman Almen, MM Edstam, R Fredriksson, HB Schioth. 2011. Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. Mol Biol Evol 28:2471-2480.

Ohno, S. 1970. Evolution by gene duplication. New York: Springer.

Olson, EN. 2006. Gene regulatory networks in the evolution and development of the heart. Science 313:1922-1927.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Pardo, L, JA Ballesteros, R Osman, H Weinstein. 1992. On the use of the transmembrane domain of bacteriorhodopsin as a template for modeling the three-dimensional structure of guanine nucleotide-binding regulatory protein-coupled receptors. Proc. Natl. Acad. Sci. USA 89: 4009–4012.

Parfrey, LW, E Barbero, E Lasser, M Dunthorn, D Bhattacharya, DJ Patterson, LA Katz. 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS Genet 2:e220.

Peter, IS, EH Davidson. 2011. Evolution of gene regulatory networks controlling body plan development. Cell 144:970-985.

Peter, IS, E Faure, EH Davidson. 2012. Feature Article: Predictive computation of genomic logic processing functions in embryonic development. Proc Natl Acad Sci U S A 109:16434-16442.

Peterson, KJ, JA Cotton, JG Gehling, D Pisani. 2008. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond B Biol Sci 363:1435-1443.

Peterson, KJ, MR Dietrich, MA McPeek. 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. Bioessays 31:736-747.

Philippe, H, H Brinkmann, DV Lavrov, DT Littlewood, M Manuel, G Worheide, D Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. Plos Biology 9:e1000602.

Philippe, H, R Derelle, P Lopez, et al. 2009. Phylogenomics Revives Traditional Views on Deep Animal Relationships. Current Biology 19:706-712.

Philippe, H, N Lartillot, H Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246-1253.

Philippe, H, J Laurent. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-623.

Pick, KS, H Philippe, F Schreiber, et al. 2010. Improved Phylogenomic Taxon Sampling Noticeably Affects Nonbilaterian Relationships. Mol Biol Evol 27:1983-1987.

Pisani, D. 2004. Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: An Example
from the Arthropoda. Systematic Biology 53:978-989.

Pisani, D, JA Cotton, JO McInerney. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol 24:1752-1760.

Pisani, D, R Feuda, KJ Peterson, AB Smith. 2012. Resolving phylogenetic signal from noise when divergence is rapid: a new look at the old problem of echinoderm class relationships. Mol Phylogenet Evol 62:27-34.

Plachetzki, DC, BM Degnan, TH Oakley. 2007. The Origins of Novel Protein Interactions during Animal Opsin Evolution. Plos One 2.

Plachetzki, DC, CR Fong, TH Oakley. 2010. The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. Proceedings of the Royal Society B-Biological Sciences 277:1963-1969.

Plotnick, R, S Dornbosb, J Chen. 2010. Information landscapes and sensory ecology of the Cambrian Radiation. Paleobiology 36:303-317.

Porter, ML, JR Blasic, MJ Bok, EG Cameron, T Pringle, TW Cronin, PR Robinson. 2011. Shedding new light on opsin evolution. Proc Biol Sci.

Posada, D. 2009. Selecting models of evolution. In: P Lemey, M Salemi, A Vandamme, editors. The phylogentic handbook. Cambrige: University Press Cambridge.

Putnam, NH, M Srivastava, U Hellsten, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86-94.

Quang, LS, O Gascuel, N Lartillot. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317-2323.

Rannala, B, Z Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304-311.

Renard, E, J Vacelet, E Gazave, P Lapebie, C Borchiellini, AV Ereskovsky. 2009. Origin of the neuro-sensory system: new and expected insights from sponges. Integr Zool 4:294-308.

Richards, TA, T Cavalier-Smith. 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436:1113-1118.

Rivera, AS, N Ozturk, B Fahey, DC Plachetzki, BM Degnan, A Sancar, TH Oakley. 2012. Blue-light-receptive cryptochrome is expressed in a sponge eye lacking neurons and opsin. J Exp Biol 215:1278-1286.

Rodriguez-Ezpeleta, N, H Brinkmann, B Roure, N Lartillot, BF Lang, H Philippe. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology 56:389-399.

Rokas, A. 2008. The molecular origins of multicellular transitions. Curr Opin Genet Dev 18:472-478.

Rokas, A, BL Williams, N King, SB Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Ronquist, F, JP Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rota-Stabelli, O, L Campbell, H Brinkmann, GD Edgecombe, SJ Longhorn, KJ Peterson, D Pisani, H Philippe, MJ Telford. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci 278:298-306.

Rota-Stabelli, O, MJ Telford. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol 48:103-111.

Roure, B, D Baurain, H Philippe. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol Biol Evol.

Ruiz-Trillo, I, M Riutort, DT Littlewood, EA Herniou, J Baguna. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919-1923.

Sanderson, M, HB Shafer. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33:49-72.

Schierwater, B, M Eitel, W Jakob, HJ Osigus, H Hadrys, SL Dellaporta, SO Kolokotronis, R Desalle. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol 7:e20.

Schioth, HB, KJ Nordstrom, R Fredriksson. 2007. Mining the gene repertoire and ESTs for G protein-coupled receptors with evolutionary perspective. Acta Physiol (Oxf) 190:21-31.

Schoneberg, T, T Hermsdorf, E Engemaier, K Engel, I Liebscher, D Thor, K Zierau, H Rompler, A Schulz. 2007. Structural and functional evolution of the P2Y(12)-like receptor group. Purinergic Signal 3:255-268.

Schwarz, G. 1978. Estimating the dimension of a model. The annals of statistics 6:461-464.

Semyonov, J, JI Park, CL Chang, SY Hsu. 2008. GPCR genes are preferentially retained after whole genome duplication. Plos One 3:e1903.

Shakhnovich, BE, E Deeds, C Delisi, E Shakhnovich. 2005. Protein structure and evolutionary history determine sequence space topology. Genome Res 15:385-392.

Sharma, AK, JL Spudich, WF Doolittle. 2006. Microbial rhodopsins: functional versatility and genetic mobility. Trends Microbiol 14:463-469.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51:492-508.

Sineshchekov, OA, KH Jung, JL Spudich. 2002. Two rhodopsins mediate phototaxis to low- and high-intensity light in Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A 99:8689-8694.

Smith, C. 2000. Biology of Sensory Systems. New York: John Wiley & Sons.

Smoot, ME, K Ono, J Ruscheinski, PL Wang, T Ideker. 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431-432.

Soding, J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960.

Soppa, J. 1994. Two hypotheses - one answer: Sequence comparison does not support an evolutionary link between halobacterial retinal proteins including bacteriorhodopsin and eukaryotic G-protein-coupled receptors. FEBS Letter 342:7-11.

Sperling, EA, KJ Peterson, D Pisani. 2009. Phylogenetic-Signal Dissection of Nuclear Housekeeping Genes Supports the Paraphyly of Sponges and the Monophyly of Eumetazoa. Mol Biol Evol 26:2261-2274.

Sperling, EA, D Pisani, KJ Peterson. 2007. Poriferan paraphyly and its implications for Precambrian palaeobiology. Rise and Fall of the Ediacaran Biota 286:355-368.

Sperling, EA, JM Robinson, D Pisani, KJ Peterson. 2010. Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. Geobiology 8:24-36.

Srivastava, M, E Begovic, J Chapman, et al. 2008. The Trichoplax genome and the nature of placozoans. Nature 454:955-960.

Srivastava, M, O Simakov, J Chapman, et al. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466:720-U723.

Stabelli, OR, N Lartillot, H Philippe, D Pisani. 2012. Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. Systematic Biology.

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

Stanke, M, M Diekhans, R Baertsch, D Haussler. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637-644.

Stechmann, A, T Cavalier-Smith. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89-91.

Strotmann, R, K Schrock, I Boselt, C Staubert, A Russ, T Schoneberg. 2011. Evolution of GPCR: change and continuity. Mol Cell Endocrinol 331:170-178.

Suga, H, V Schmid, WJ Gehring. 2008. Evolution and functional diversity of jellyfish opsins. Current Biology 18:51-55.

Talavera, G, J Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56:564-577.

Tamura, K, D Peterson, N Peterson, G Stecher, M Nei, S Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Taylor, E, A Agarwal. 1993. Sequence homology between bacteriorhodopsin and G-protein coupled receptors: exon shuffling or evolution by duplication? FEBS Letter 325:161-166.

Taylor, JS, J Raes. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615-643.

Terakita, A. 2005. The opsins. Genome Biol 6:213.

Thompson, JD, TJ Gibson, DG Higgins. 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2:Unit 2 3.

Thompson, JD, DG Higgins, TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Valentine, J, A Collins, C Porter Meyer. 1994. Morphological Complexity Increase in Metazoans. Paleobiology 20:131-142.

Van de Peer, Y. 2009. Phylogenetic inference based on distance methods. In: P Lemey, M Salemi, A Vandamme, editors. The phylogentic handbook. Cambrige: University Press Cambridge.

Vannier, J, DC Garcia-Bellido, SX Hu, AL Chen. 2009. Arthropod visual predators in the early pelagic ecosystem: evidence from the Burgess Shale and Chengjiang biotas. Proc Biol Sci 276:2567-2574.

Vassilatis, DK, JG Hohmann, H Zeng, et al. 2003. The G protein-coupled receptor repertoires of human and mouse. Proc Natl Acad Sci U S A 100:4903-4908.

von Haeseler, A. 2012. Do we still need supertrees? BMC Biol 10:13.

Wagner, A. 2011. The origins of evolutionaty innovations. New Yourk: Oxford University Press.

Waschuk, SA, AG Bezerra, Jr., L Shi, LS Brown. 2005. Leptosphaeria rhodopsin: bacteriorhodopsin-like proton pump from a eukaryote. Proc Natl Acad Sci U S A 102:6879-6883.

Waterhouse, AM, JB Procter, DM Martin, M Clamp, GJ Barton. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.

Wettschureck, N, S Offermanns. 2005. Mammalian G proteins and their cell type specific functions. Physiological Review 85:1159-1204.

Wheeler, WC. 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics 6:363-367.

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

Wilkinson, M, JO McInerney, RP Hirt, PG Foster, TM Embley. 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol Evol 22:114-115.

Williams, PD, DD Pollock, BP Blackburne, RA Goldstein. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2:e69.

Woese, CR, GE Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088-5090.

Wong, KM, MA Suchard, JP Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. Science 319:473-476.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang, Z, B Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14.

Yona, S, HH Lin, WO Siu, S Gordon, M Stacey. 2008. Adhesion-GPCRs: emerging roles for novel receptors. Trends Biochem Sci 33:491-500.

Zakhvatkin, AA. 1949. The comparative embryology of the low invertebrates. Sources and method of the origin of metazoan development. Moscov Soviet Science.

Zakon, HH. 2002. Convergent evolution on the molecular level. Brain Behav Evol 59:250-261.

Zlotnik, A, O Yoshie, H Nomiyama. 2006. The chemokine and chemokine receptor superfamilies and their molecular evolution. Genome Biol 7:243.

Zuckerkandl, E, L Pauling. 1965. Molecules as documents of evolutionary history. J Theor Biol 8:357-366.

Appendix

Appendix A

Taxon	Phylum	#Genes	% missing data
Acanthoscurria gomesiana	*Arthropoda*	88	50.8
Acropora millepora	*Cnidaria*	116	34.9
Amphimedon queenslandica	*Porifera*	138	5.8
Amoebidium parasiticum	*Choanoza/Ichthyophonida*	71	67.7
Anemonia viridis	*Cnidaria*	103	39.9
Anoplodactylus eroticus	*Arthropoda*	53	66.6
Aplysia californica	*Mollusca*	143	5.8
Argopecten irradians	*Mollusca*	78	57.1
Asterina pectinifera	*Echinodermata*	78	53.6
Biomphalaria glabrata	*Mollusca*	122	25.9
Boophilus microplus	*Arthropoda*	111	22.8
Branchiostoma floridae	*Chordata*	145	0.6
***Capitella* sp.**	*Annelida*	144	1.3
Capsaspora	*Filasterea*	144	6
Carcinoscorpius rotundicauda	*Arthropoda*	15	92.6
Carinoma mutabilis	*Nemertea*	48	69.4
Carteriospongia foliascens	*Porifera*	49	78.9
Cerebratulus lacteus	*Nemertea*	15	92.8
Chaetoderma nitidulum	*Mollusca*	31	82.7
Chaetopleura apiculata	*Mollusca*	20	90.3
***Chaetopterus* sp.**	*Annelida*	43	75.4
Ciona intestinalis	*Chordata*	142	1.6
Clytia hemisphaerica	*Cnidaria*	89	36.3
Crassostrea virginica	*Mollusca*	130	23.5
Crateromorpha meyeri	*Porifera*	35	82.9
Cryptococcus neoformans	*Fungi/Basidiomycota*	135	5.3
Cyanea capillata	*Cnidaria*	35	85.3
Daphnia pulex	*Arthropoda*	142	1.6
Drosophila melanogaster	*Arthropoda*	137	4.7
Dugesia japonica	*Platyhelminthes*	104	30.2
Echinococcus granulosus	*Platyhelminthes*	105	40.7
Echinoderes horni	*Cephalorhyncha*	43	74.8
Ephydatia muelleri	*Porifera*	62	62.7
Euperipatoides kanangrensis	*Onychophora*	49	72.1
Euprymna scolopes	*Mollusca*	112	37.2
Gallus gallus	*Chordata*	134	10.4
Haementeria depressa	*Annelida*	34	82.4
Heterochone calyx	*Porifera*	54	71.7
Homarus americanus	*Arthropoda*	123	26.7
Homo sapiens	*Chordata*	144	2
Hydra magnipapillata	*Cnidaria*	145	2
Hydractinia echinata	*Cnidaria*	81	54.4
Hypsibius dujardini	*Tardigrada*	72	64.6

Leucetta chagosensis	*Porifera*	68	63
Litopenaeus vannamei	*Arthropoda*	125	18.5
Lubomirskia baicalensis	*Porifera*	45	77.5
Lumbricus rubellus	*Annelida*	87	51.4
Macrostomum lignano	*Platyhelminthes*	66	57.2
***Mertensiid* sp.**	*Ctenophora*	46	71.4
Metridium senile	*Cnidaria*	97	40.4
Mnemiopsis leidyi	*Ctenophora*	96	40.7
Monosiga brevicollis	*Choanoflagellatea*	140	4
Monosiga ovata	*Choanoflagellatea*	68	58.6
Montastraea faveolata	*Cnidaria*	81	64.4
Mytilus galloprovincialis	*Mollusca*	122	23.1
Nematostella vectensis	*Cnidaria*	146	0.1
Oopsacas minuta	*Porifera*	23	88.5
Oscarella carmela	*Porifera*	100	40.1
Oscarella lobularis	*Porifera*	19	89.4
Pachydictyum globosum	*Porifera*	40	84
***Paraplanocera* sp.**	*Platyhelminthes*	41	73.4
Phoronis vancouverensis	*Phoronida*	31	81.4
Platynereis dumerilii	*Annelida*	67	56.9
Pleurobrachia pileus	*Ctenophora*	114	23
Podocoryne carnea	*Cnidaria*	69	59.3
Priapulus caudatus	*Priapulida*	58	68
***Proterospongia* sp.**	*Choanoflagellatea*	115	15.6
Ptychodera flava	*Hemichordata*	48	70.3
Richtersius coronifer	*Tardigrada*	151	2.6
Saccamyces cerivisae	*Fungi/Ascomycota*	133	5.2
Saccoglossus kowalevskii	*Hemichordata*	144	3.2
Schmidtea mediterranea	*Platyhelminthes*	138	5.1
Scutigera coleoptrata	*Arthropoda*	42	74.6
Sphaeroforma artica	*Choanoza/Ichthyophonida*	92	44.3
Spinochordodes tellinii	*Nematomorpha*	9	96.8
Strongylocentrotus purpuratus	*Echinodermata*	145	1.2
Suberites domuncula	*Porifera*	41	74.3
Sycon raphanus	*Porifera*	59	65.2
Terebratalia transversa	*Mollusca*	57	68.5
Themiste lageniformis	*Sipuncula*	40	77.2
Trichinella spiralis	*Nematoda*	135	8
Trichoplax adhaerens	*Placozoa*	143	2.6
Urechis caupo	*Echiura*	50	68.3
Xenoturbella bocki	*Xenacoelomorpha*	73	54.2
Xiphinema index	*Nematoda*	94	51

Appendix A. Number of genes and amount of missing data for the 146-NGs supermatrix.

Appendix B

taxon	p-value	z-score
Acanthoscu_1_1	0.43	-0.163
Acroporami_1_1	0.15	0.932
*** Amoebidium_1_8**	0	3.505
Anemoniavi_1_1	0.505	-0.203
*** Anoplodact_1_9**	0	3.269
Aplysiacal_1_2	0.215	0.701
Argopecten_1_1	0.075	1.346
Asterinape_1_1	0.15	1.202
Biomphalar_1_2	0.225	0.757
*** Branchiost_1_3**	0	3.033
*** Capitellas_1_3**	0.01	2.381
*** Capsaspora_1_3**	0	10.369
Carcinosco_1_2	0.086	1.533
Carinomamu_1_8	0.086	1.392
Carteriosp_1_5	0.204	0.796
Cerebratul_1_2	0.107	1.33
*** Chaetoderm_1_4**	0.021	2.146
Chaetopleu_1_2	0.43	0.019
Chaetopter_1_6	0.376	0.301
Clytiahemi_1_1	0.086	1.28
Crassostrea_1	0.064	1.57
Crateromor_1_5	0.924	-1.378
*** Cryptococc_1_3**	0	10.02
Cyaneacapi_1_4	0.086	1.215
*** Daphniapul_1_3**	0.01	2.778
Drosophila_1_3	0.053	1.699
*** Echinodere_1_7**	0.01	2.638
Ephydatiam_1_1	0.129	1.163
*** Euperipato_1_7**	0.01	3.683
Euprymnasc_1_1	0.268	0.597
*** Gallusgall_1_2**	0.032	2.089

Haementeri_1_4	0.053	1.646
Heterochon_1_8	0.075	1.512
Homarusame_1_2	0.408	0.11
*** Homosapien_1_3**	0	2.604
Hydractini_1_1	0.29	0.518
*** Hydramagni_1_3**	0	4.016
*** Leucettach_1_1**	0	3.608
*** Litopenaeu_1_2**	0.01	3.124
Lubomirski_1_6	0.172	0.687
Lumbricusr_1_1	0.322	0.266
*** metridium**	0.043	1.678
Mnemiopsis_1_1	0.053	1.984
*** Monosigabr_1_3**	0	10.953
*** Monosiga_ovata**	0	10.123
Montastrae_1_9	0.086	1.256
Mytilusgal_1_2	0.064	1.687
*** Nematostel_1_3**	0.021	1.882
*** Oopsacasmi_1_3**	0	3.822
*** Oscarellal_1_2**	0.021	2.359
*** Oscarella_nost**	0.01	2.675
*** Pachydicty_1_4**	0	2.678
Phoronisva_1_5	0.161	0.943
Platynerei_1_1	0.516	-0.06
Pleurobrac_1_2	0.193	0.893
Podocoryne_1_1	0.387	0.006
Priapulusc_1_8	0.365	0.084
*** Ptychodera_1_8**	0.01	2.584
Renierasp._1_2	0.064	1.826
*** Bhoophilus**	0	5.149
*** Saccharomy_1_3**	0	4.646
Saccogloss_1_3	0.537	-0.014
Scutigerac_1_7	0.086	1.261
*** Sphaerofor_1_1**	0	7.052
Strongyloc_1_3	0.204	0.838
Suberitesd_1_7	0.129	1.02
Syconrapha_1_9	0.204	0.866

*** Terebratal_1_9**	0.021	2.698
Themistela_1_6	0.053	1.646
*** Trichinell_1_2**	0	2.804
*** Trichopla__1_3**	0	4.665
Urechiscau_1_8	0.376	0.228
Xiphinemai_1_1	0.376	0.336
mertensiid_1_7	0.698	-0.53
*** proterospT_1_2**	0	6.152
global test:		
succeeded		
observed : 0.00143927		
mean pred : 0.00147748		
p-value : 0.548387		
z-score : -0.168684		

Appendix B. PPA for compositional homogeneity results of 146-NGs

Appendix C

Eukaryote	**unikont/metazoa**	**Lottia gigantea**
Eukaryote	unikont/metazoa	*Capitella sp.*
Eukaryote	unikont/metazoa	*Tribolium castanedum*
Eukaryote	unikont/metazoa	*Drosophila melanogaster*
Eukaryote	unikont/metazoa	*Daphnia pulex*
Eukaryote	unikont/metazoa	*Nematostella vectensis*
Eukaryote	unikont/metazoa	*Hydra magnipapillata*
Eukaryote	unikont/metazoa	*Trichoplax adherens*
Eukaryote	unikont/metazoa	*Oscarella carmela*
Eukaryote	unikont/metazoa	*Sycon sp.*
Eukaryote	unikont/metazoa	*Amphimedon queenslandica*
Eukaryote	unikont	*Capsaspora owczarzaki*
Eukaryote	unikont	*Salpingoeca rosetta*
Eukaryote	unikont	*Monosiga brevicollins*
Eukaryote	unikont	*Dictostelium porporatum*
Eukaryote	unikont	*Dictyostelium discoideum*
Eukaryote	unikont	*Aspergillus niger*
Eukaryote	unikont	*Aureococcus anophagefferens*
Eukaryote	unikont	*Coccomyxa sp.*
Eukaryote	unikont	*Coprinus cinereus*
Eukaryote	unikont	*Cryptococcus neoformans*
Eukaryote	unikont	*Saccharomyces cerevisiae*
Eukaryote	Plants	*Arabidopsis lyrata*
Eukaryote	Plants	*Arabidopsis thaliana*
Eukaryote	Plants	*Brachypodium distachyon*
Eukaryote	Plants	*Brassica rapa*
Eukaryote	Plants	*Sorghum bicolor*
Eukaryote	Plants	*Vitis vinifera*
Eukaryote	Plants	*Oryza glaberrima*
Eukaryote	Plants	*Oryza indica*
Eukaryote	Plants	*Oryza sativa*
Eukaryote	Plants	*Glycine max*
Eukaryote	Plants	*Populus trichocarpa*
Eukaryote	Plants	*Zea mays*
Eukaryote	Plants	*Chlamydomonas reinhardtii*
Eukaryote	Plants	*Chlorella sp.*
Eukaryote	Plants	*Cyanidioschyzon merolae*
Eukaryote	Plants	*Micromonas pusilla*
Eukaryote	Plants	*Ostreococcus lucimarinus*
Eukaryote	Plants	*Physcomitrella patens*
Eukaryote	Plants	*Selaginella moellendorffii*
Eukaryote	Plants	*Volvox carteri*
Eukaryote	Excavata	*Leishmania major*
Eukaryote	Excavata	*Naegleria gruberi*
Eukaryote	Excavata	*Trichonoma vagianalis*
Eukaryote	Excavata	*Giardia*

Eukaryote	Chromalveolata	*Trypanosoma brucei*
Eukaryote	Chromalveolata	*Babesia bovis*
Eukaryote	Chromalveolata	*Emiliania huxleyi*
Eukaryote	Chromalveolata	*Guillardia theta*
Eukaryote	Chromalveolata	*Phaeodactylum tricornutum*
Eukaryote	Chromalveolata	*Phytophthora infestans*
Eukaryote	Chromalveolata	*Phytophthora ramorum*
Eukaryote	Chromalveolata	*Plasmodium falciparum*
Eukaryote	Chromalveolata	*Pythium ultimum*
Eukaryote	Chromalveolata	*Thalassiosira pseudonana*
Eukaryote	Chromalveolata	*Toxoplasma gondii*
Eukaryote	Rhizaria	*Bigelowiella natans*

Appendix C. Eukaryotic species analysed in chapter 3

Publications

Metazoan opsin evolution reveals a simple route to animal vision

Roberto Feuda[a], Sinead C. Hamilton[a], James O. McInerney[a], and Davide Pisani[a,b,1]

[a]Department of Biology, National University of Ireland Maynooth, Kildare, Ireland; and [b]School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol BS8 1UG, United Kingdom

Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved September 13, 2012 (received for review March 21, 2012)

All known visual pigments in Neuralia (Cnidaria, Ctenophora, and Bilateria) are composed of an opsin (a seven-transmembrane G protein-coupled receptor), and a light-sensitive chromophore, generally retinal. Accordingly, opsins play a key role in vision. There is no agreement on the relationships of the neuralian opsin subfamilies, and clarifying their phylogeny is key to elucidating the origin of this protein family and of vision. We used improved methods and data to resolve the opsin phylogeny and explain the evolution of animal vision. We found that the Placozoa have opsins, and that the opsins share a common ancestor with the melatonin receptors. Further to this, we found that all known neuralian opsins can be classified into the same three subfamilies into which the bilaterian opsins are classified: the ciliary (C), rhabdomeric (R), and go-coupled plus retinochrome, retinal G protein-coupled receptor (Go/RGR) opsins. Our results entail a simple scenario of opsin evolution. The first opsin originated from the duplication of the common ancestor of the melatonin and opsin genes in a eumetazoan (Placozoa plus Neuralia) ancestor, and an inference of its amino acid sequence suggests that this protein might not have been light-sensitive. Two more gene duplications in the ancestral neuralian lineage resulted in the origin of the R, C, and Go/RGR opsins. Accordingly, the first animal with at least a C, an R, and a Go/RGR opsin was a neuralian progenitor.

ancestral character state reconstruction | Metazoa | protein evolution

Understanding the origin and early evolution of vision at the molecular level has proven difficult (1–4). Both Protostomia (e.g., Mollusca and Arthropoda) and Deuterostomia (e.g., Vertebrata) have eyes, and it is plausible that the last common ancestor of the Bilateria possessed simple eyespots and some limited ability to detect light (5). In addition, eyes are known in jellyfishes (e.g., refs. 6, 7), and the common use of a Pax-6 regulated kernel [*sensu* Davidson and Erwin (8)] to control eye development in Cnidaria and Bilateria suggests a single origin of the neuralian eye (9). Furthermore, all neuralians for which data are available detect light by using visual pigments composed of an opsin and a chromophore, generally retinal (3), and their opsins link the chromophore through a Schiff base involving a lysine found at position 296 (K296) of the reference bovine rhodopsin sequence (10).

Opsins are seven-transmembrane proteins belonging to the G protein-coupled receptor (GPCR) superfamily (11). According to the glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin (GRAFS) (12) classification system, opsins are members of the α-group of the rhodopsin-like receptors, and they are further classified in several subfamilies (11). Given that the opsins seem to be universally distributed within Neuralia (1, 2, 4, 7, 13), it is clear that, to understand the molecular foundations of vision, we must focus on the early branching metazoans: the Cnidaria, the Ctenophora, the Placozoa, and the sponges. Unfortunately, the phylogenetic relationships of the neuralian opsins are still debated (1–4), and, as a consequence, the early history of gene duplications and deletions within this family is still unknown (*SI Appendix*, Fig. S1). Should we wish to understand the origin of vision (in both its tempo and mode), the pattern of opsin duplications and deletions must be clarified first, and this can only be done by resolving the opsin phylogeny.

The current gap in our understanding of the evolution of vision is, at least in part, the consequence of an absence of genomic information for key, early branching metazoans. Data are still missing for two nonbilaterian lineages: the Ctenophora and the calcarean sponges. However, the genomes of four key taxa, the placoazoa *Trichoplax adhaerens* (14), the cnidarians *Hydra magnipapillata*(15) and *Nematostella vectensis* (16), and the demosponge *Amphimedon queenslandica* (17), have recently been released, improving data availability. Further to this, the genome of *Oscarella carmela*, a representative of a second sponge lineage (the Homoscleromorpha), has now been sequenced (18) and deposited in Compagen (http://compagen.zoologie.uni-kiel.de/).

The relationships among the sponges are still debated (19–23), and two competing hypotheses exist. The first suggests that the sponges are monophyletic (21, 22), whereas the second (19, 20, 23) suggests that they are paraphyletic. According to the sponge monophyly hypothesis, Porifera is the sister group of Eumetazoa, and both the Demospongiae and the Homoscleromorpha are valid outgroups to study the eumetazoan GPCRs (opsins included). According to the paraphyly hypothesis, the Homoscleromorpha is the sister group of the Eumetazoa, and proteins that are most closely related to the eumetazoan GPCRs should be found in this group only. Inclusion of the *Oscarella* genome is thus key to ensure that the closest sister group of the Eumetazoa is being considered when studying GPCR evolution, irrespective of what the relationships among the sponge classes are. Here, genomic information from all aforementioned taxa (*Oscarella* included) was used, together with a large sample of well-characterized neuralian opsins (*SI Appendix*, Table S1), to investigate the origin and evolution of the opsin family and of vision.

Bilaterian opsins have been classified in three major subfamilies (11): rhabdomeric (R) opsins, ciliary (C) opsins, and go-coupled plus retinochrome, retinal G protein-coupled receptor (Go/RGR) opsins. Usually there is an association between light receptors (i.e., the cells expressing these proteins) and specific opsin subfamilies, with the ciliary receptors expressing C and Go/RGR opsins, and the rhabdomeric receptors expressing R opsins (3, 24). A fourth opsin subfamily was suggested by Plachetzki et al. (1). These authors (*SI Appendix*, Fig. S1*A*) identified a large clan (*sensu* ref. 25) of cnidarian-specific opsins that they named Cnidopsins. In addition, they found that one cnidarian opsin in their data set clustered with the bilaterian C opsins, a result that is consistent with the observation that cnidarians have ciliary receptors (24).

Four studies (1–4) have addressed the relationships among the main opsin groups with a view of clarifying the gene duplication and deletion history within this family, but they reached contradictory

Author contributions: J.O.M. and D.P. designed research; R.F. performed research; S.C.H. contributed new reagents/analytic tools; R.F. and D.P. analyzed data; and R.F., S.C.H., J.O.M., and D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: davide.pisani@nuim.ie.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1204609109/-/DCSupplemental.

results (*SI Appendix*, Fig. S1). A major source of uncertainty in these studies is that three of them (1–3) failed to include a representative sample of cnidarian opsins (*SI Appendix*, Fig. S1 *A*, *C*, and *D*). Accordingly, these studies did not have the power to test every possible hypothesis of opsin evolution. In addition, all four (1–4) used precomputed, empirical time reversible matrices to model amino acid substitutions. These matrices—WAG (1, 2), MtRev (3), and JTT (4)—are unlikely to fit an opsin dataset well because they were not derived from an opsin alignment. Further to this, all the aforementioned studies used uncritically selected outgroups. Plachetzki et al. (1) recognized that the use of problematic outgroups might negatively affect the opsin phylogeny, but failed to find a valid solution to this problem (*SI Appendix*). Consequently, all phylogenies in *SI Appendix*, Fig. S1, are questionable.

Here we performed detailed analyses to better understand opsin evolution. Unlike previous studies, we used modern, well-performing multiple sequence alignment software (26). We implemented better fitting evolutionary models, and considered all available genomic information for the deeply branching metazoans, including the newly sequenced genome of the homoscleromorph sponge *O. carmela*. We thoroughly tested a large sample of putative opsin outgroups and performed analyses by using only the less divergent ones. Most importantly, we used a comprehensive set of cnidarian opsins, including all sequences specific to two previous studies (1, 4). Accordingly, our data set has the power to test every proposed hypothesis of opsin relationships, and its analysis should allow the achievement of greater precision in pinpointing duplications and losses within the opsin family.

Results

Common problems with previous studies (1–4) were the use of under-sampled data sets, substitution models that did not fit the data [precomputed empirical time reversible (GTR) matrices], and inadequate outgroup selection (as detailed earlier). To avoid such problems, we assembled three GPCR and opsin alignments scoring hundreds of sequences (*Methods*), and estimated alignment-specific GTR matrices. Our matrices differ from available, precomputed GTR matrices (*SI Appendix*, Table S2 and Fig. S2), with the Akaike information criterion and Bayesian cross validation showing that they fit the data significantly better than any precomputed GTR matrix, and at least as well as any precomputed site-heterogeneous model (*SI Appendix*, Tables S3 and S4).

Fig. 1*A* represents the phylogeny derived from our all opsin master (AOM) alignment (*Methods*). AOM includes only neuralian opsins (no outgroups), and Fig. 1*A* is thus an unrooted phylogeny of our opsin data set (*SI Appendix*, Table S1). Fig. 1*A* (see also *SI Appendix*, Fig. S3) is consistent with the monophyly of the traditionally recognized bilaterian opsin subfamilies (C, R, and Go/RGR). In contrast, the Cnidarian opsins are split into three clans (hereafter referred to as groups A, B, and C). This is in agreement with the results of Suga et al. (4), but in disagreement with others (1–3). Group A includes only two sequences and sits on the branch separating the R opsins from all the other sequences in our dataset [posterior probability (PP) of 0.84]. The sequences in group A are from the study of Suga et al. (4), in which they were named group 3. These sequences were not included in the other three studies (1–3). Group B forms a relatively poorly supported clan with the Go/RGR opsins (PP = 0.69), whereas group C is found in a polytomy with the C opsins and the Go/RGR plus group B clans (Fig. 1*A*). Group C includes both the sequences that, in the study of Suga et al. (4), emerged as the sister group of the R opsins (their group 2 opsins) and the single sequence that Plachetzki et al. (1) classified as a C opsin. The phylogeny shown in Fig. 1*A* rejects the possibility that Suga et al.'s (4) group 2 opsins could be related to the R opsins. However, it could neither confirm nor reject the C opsin nature of Plachetzki et al.'s (1) putative C opsin. This is because Fig. 1*A* shows that all the aforementioned sequences belong to group C: a group that could not be placed with confidence with reference to the C and the Go/RGR plus group B opsins.

Posterior predictive analysis (*SI Appendix*, Table S5) showed that some of the sequences in AOM were compositionally heterogeneous. Because of their skewed amino acid composition, these sequences can mislead phylogenetic analyses (27). Heterogeneous sequences were included in AOM for the purposes of testing to which major opsin clan they belong. However, most of these sequences were excluded from further analyses (*Methods* and *SI Appendix*) to avoid their potentially biasing effect. Other sequences, such as short expressed sequence tags (ESTs) that, in Fig. 1*A*, were unequivocally identified as members of one of the opsin clans, were also excluded from further analyses.

We analyzed the GPCR and opsin master alignment (G&OM; *Methods*) to test what GPCR family is most closely related to the opsin family. These analyses (Fig. 1*B* and *SI Appendix*, Fig. S4) shown that the neuralian opsins form a monophyletic group. Importantly, the relationships among the neuralian opsins in Fig. 1*B* are consistent with those of Fig. 1*A*. That is, the tree in Fig. 1*B* is a rooted resolution of Fig. 1*A* in which the polytomy from which the C opsins, the Go/RGR plus group B opsins, and the group C opsins stem is resolved according to one of its possible resolutions. Fig. 1*B* also shows that the neuralian opsins are most closely related to a set of placozoan "opsin-like" sequences (PP = 0.98). By turn, the neuralian opsins and the placozoan opsin-like sequences are most closely related to the melatonin (MLT) receptors (PP = 0.89). Fig. 1*B* shows that both the placozoans and the cnidarians have MLT receptors, and, most importantly, that the placozoan opsin-like receptors are orthologues of the neuralian opsins. This implies that from an evolutionary point of view, the placozoan opsin-like receptors are members of the opsin family, even though they lack a retinal binding domain (RBD) with a K296 residue and might thus be unable to detect light. Neither an opsin nor an MLT receptor could be identified in *Oscarella* and *Amphimedon*, and we can thus conclude that both these protein families are eumetazoan specific. This confirms recent results showing that light sensitivity in *Amphimedon* is mediated by a cryptochrome, rather than an opsin (28). Fig. 1*B* shows that the MLT-plus-opsin clade is most closely related to a group including the lysosphingolipid and the orexin receptors (albeit with very low support; PP = 0.46; Fig. 1*B* and *SI Appendix*, Fig. S4). *Oscarella* and *Amphimedon* have sequences belonging to the latter (PP = 0.94; Fig. 1*B*), further confirming the eumetazoan nature of the opsin family.

We tested whether distant outgroups in the G&OM data set could have caused tree-reconstruction artifacts with reference to the opsin phylogeny. To do so, we analyzed the opsins and outgroups (O&O) alignment (*Methods*). The MLT receptors are the sole outgroups of O&O, which also include the placozoan opsin-like receptors. The Bayesian O&O phylogeny is reported in Fig. 1*C* (*SI Appendix*, Fig. S5), and the O&O maximum likelihood (ML) phylogeny is reported in *SI Appendix*, Fig. S6. Analyses of O&O confirmed the results obtained using G&OM (compare Fig. 1*B* vs. Fig. 1*C*). Both data sets show that the Cnidarian opsins can be classified in three groups (A, B, and C). These groups represent, respectively, the cnidarian orthologue of the bilaterian R opsins [group A; GTR PP = 0.89 and ML bootstrap proportion (BP) under an LG plus Γ model = 62%], the cnidarian orthologue of the bilaterian Go/RGR opsins (group B; PP = 0.81 and LG BP < 50), and the cnidarian orthologue of the bilaterian C opsins (group C; PP = 0.71 and LG BP < 50). ML bootstrap support values for the opsin internal relationships are low. Therefore, we used the approximately unbiased (AU) test (29) to evaluate whether the data, under the best-fitting GTR plus Γ model, can discriminate between alternative opsin phylogenies. The results of the AU test (Table 1) confirm that the data are informative and that the trees in Fig. 1*C* fit the O&O data set significantly better than the trees of the aforementioned previous publications (1–4).

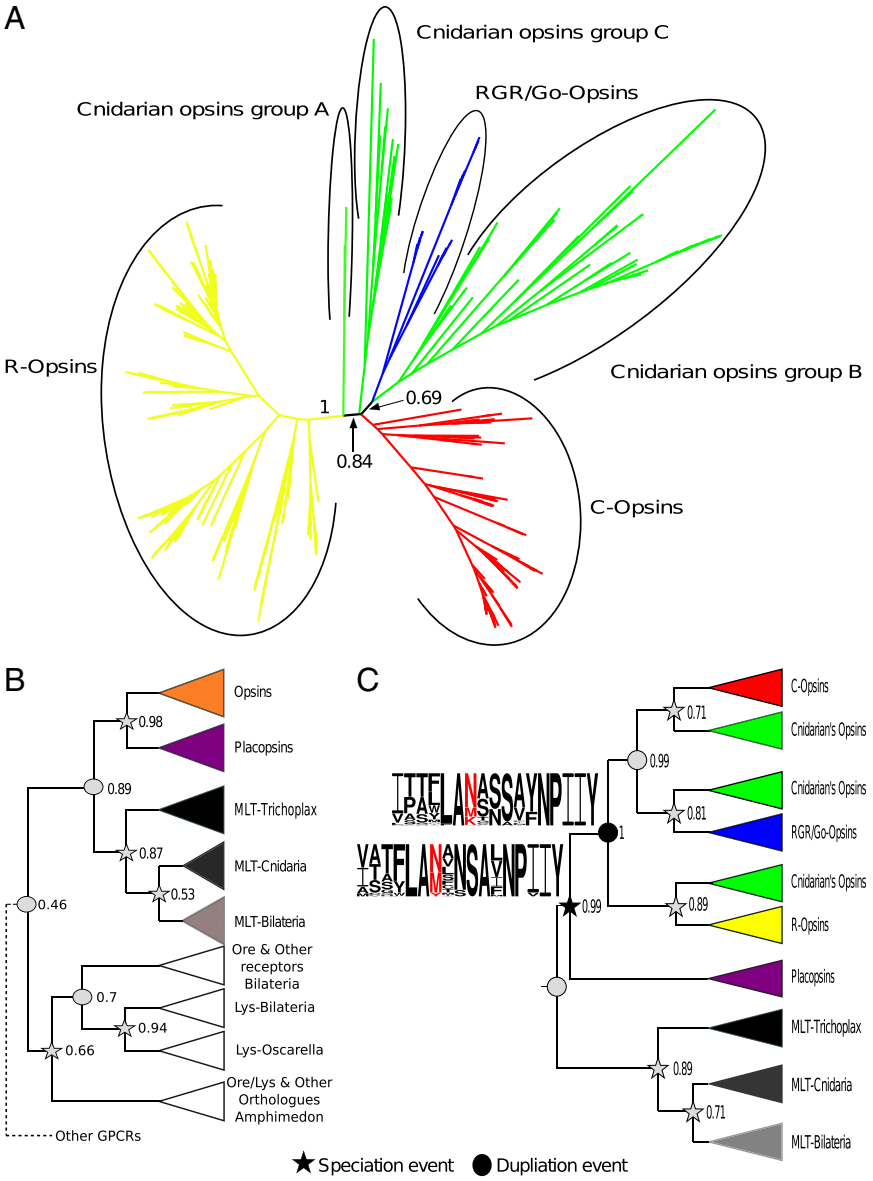


Fig. 1. Phylogeny of the opsin family. (*A*) Unrooted phylogeny of the neuralian opsins. (*B*) Rooted phylogeny of the neuralian opsins and of other GPCRs showing that the Placopsins are members of the opsin family (Ore, orexin; Lys, lysosphingolipid). (*C*) Opsin phylogeny rooted by using only the MLT receptors, and showing that cnidarians have orthologues of each bilaterian opsin subfamily: the C, R, and Go/RGR subfamilies. Support values (Bayesian PPs) are reported only for key nodes (*SI Appendix* shows all support values). The ancestral RBD of the LOCA and of the LOCNA are reported and are identified, respectively, by a black star and a black circle (*SI Appendix*, Fig. S7). The red position in the logos identifies position 296.

To provide further insights into opsin evolution, we carried out Bayesian and ML ancestral character state reconstruction of the RBD at key internal nodes. Results of the Bayesian reconstruction are reported as logos in Fig. 1*C* (*SI Appendix*, Fig. S7), and indicate that the last opsin common ancestor (LOCA) most likely did not have the key K296 residue (PP for K296 = 0.0034). Instead, position 296 was either occupied by an asparagine (PP for N296 = 0.51) or by a methionine (PP for M296 = 0.37). Absence of K296 in LOCA is confirmed by ML, which suggests with reasonable probability (P) that asparagine was the most likely amino acid in position 296 (P-N296 = 0.81 and P-K296 = 0.054). K296 is necessary to link the chromophore, and our results suggest that K296-mediated chromophore binding was not a feature of LOCA: it evolved within the opsin family. Indeed, even in the case of the last opsin common neuralian ancestor (LOCNA), the Bayesian reconstruction suggests that the RBD might not have had a K296 residue (PP for K296 = 0.15; Fig. 1*C*). However, ML contradict this result, as it finds a P-K296 value of 0.99. This incongruence leaves the question of occupancy of position 296 in LOCNA unresolved. No matter what the amino acid in LOCNA was, our results strongly suggest that a K296-based RBD was not a feature of LOCA.

Table 1. Results of AU tests

Hypothesis	Probability
Fig. 1C and *SI Appendix*, Fig. S5	0.7
Plachetzki et al. (1)	0.04
Porter et al. (3)	0.03
Plachetzki et al. (2)	0.008
Suga et al. (4)	5×10^{-18}

Discussion

Our results are markedly different from those of previous investigations. These differences reflect data completeness and methodological dissimilarities. We used a combination of recently developed multiple sequence alignment software that can better differentiate between insertions and deletions and performed extensive model selection analyses, resulting in the use of significantly better-fitting substitution models. We were careful to include the closest outgroups of the neuralian opsins (including sequences from the Placozoa and the Homoscleromorpha), and we used a very inclusive set of cnidarian opsins allowing for the simultaneous test of previous hypotheses (1–4).

Our results (summarized in Fig. 2) allow for a substantial clarification of the tempo and mode of opsin evolution. They confirm the results of Fredriksson et al. (12) that the sister group of the opsin family is represented by the MLT receptors, and they show that the opsin family originated from the duplication of the MLT and opsin ancestral gene in the stem eumetazoan lineage. Importantly, we were able to show that the placozoan genome contains sequences that are in an orthologous relationship with the neuralian opsins. From an evolutionary point of view, these sequences are members of the opsin family (Fig. 1*B* and Fig. 2) irrespective of whether they have the ability to detect light, and we propose to refer to these opsin-like receptors as placopsins. In addition, we show that cnidarians have R, Go/RGR, and C opsin orthologues. Accordingly, these opsin subfamilies must have evolved in the stem neuralian lineage, rather then in the stem bilaterian lineage, i.e., earlier than currently accepted.

Our results are largely phylogeny-independent. Nonetheless, uncertainty in the placement of the Placozoa still persists and deserves discussion. Consistently with our results, some of the most thorough analyses to date (20, 21) agree that the Placozoa are the sister group of Neuralia, even though some investigators (22, 30) found different results. However, Philippe et al. (31) have shown the results of Schierwater et al. (30) to be invalid. Differently, even though the study of Pick et al. (22) is sound, its conclusion that Placozoa is a member of Neuralia is questionable. This is because their dataset (22) is based on that of Dunn et al. (32), which has been shown to be unreliable (21, 22, 31). Importantly, even if Bilateria and Placozoa were confirmed to be sister groups (22), our results would still be valid, but our scenario would become less parsimonious as it would imply independent losses of the placopsin in Bilateria and Cnidaria, and of the C, R, and Go/RGR opsins in Placozoa.

Ancestral character state reconstruction suggests LOCA did not have a RBD containing a K296 residue. Accordingly, K296-mediated light detection most likely evolved in the stem Eumetazoan lineage, perhaps through autogenous evolution and neofunctionalization of a protein that was not light-sensitive. Neither of the two sponge taxa code for MLT or opsin receptors, yet their genomes include sequences clustering in the Lysosphingolipid plus orexin group (i.e., they code for proteins belonging to the sister group of the MLT-plus-opsins clade; Fig. 1*B*). These results confirm that the first opsin originated in the stem eumetazoan lineage, and imply that our conclusions are robust irrespective of whether sponges are monophyletic (21) or paraphyletic (20).

Identification of the duplication of the ancestral MLT plus opsin gene in the stem eumetazoan lineage lets us better constrain the timing of this event, as this lineage was dated to have existed between 755 and 711 Ma (19). In addition, the neuralian stem lineage was dated to have existed between 711 and 700 Ma (19). This relatively short time (11 million y) was a crucial period in opsin evolution. It was during this time that the K296-based RBD most likely evolved and the duplications separating the C plus Go/RGR opsin ancestor from the R opsins, and the C from the Go/RGR opsins, were fixed.

Our results suggest that the Go/RGR opsins represent the sister group of the C opsins. This is in disagreement with some previous findings (1–3), but is in agreement with others (4, 11). An additional line of evidence that seems to support our conclusion is that the Go/RGR opsins, exactly as the C opsins, are expressed in ciliary receptors (3, 24). Our results also predict that rhabdomeric receptors should exist in Cnidaria. This has not yet been proven, but cells with a strong resemblance to the bilaterian rhabdomeric receptors, which could be cnidarian rhabdomeric receptors, have been observed in cnidarian larvae (9, 24, 33).

Conclusions

We suggest an early and parsimonious explanation for the diversification of the opsin family (summarized in Fig. 2), and show that LOCA most likely did not have a K296-based RBD. Scarcity of signal for the deepest event in the history of the opsin family implies that some level of uncertainty in opsin evolution still remains, and might be unavoidable. However, results of the AU tests show that the topology uncovered in this study fits the data (under a GTR-plus-Γ model) significantly better than any previously proposed opsin phylogeny. Our results also indicate that a short 11-million-y period (711–700 Ma) was key in opsin evolution. During this time, two duplications in the stem neuralian lineage resulted in the evolution of the extant opsin paralogues. During this same time, the K296-based RBD most likely evolved, probably through a process of neofunctionalization. Our results are compatible with the view that the last common neuralian ancestor might have been more complex than generally assumed (34).

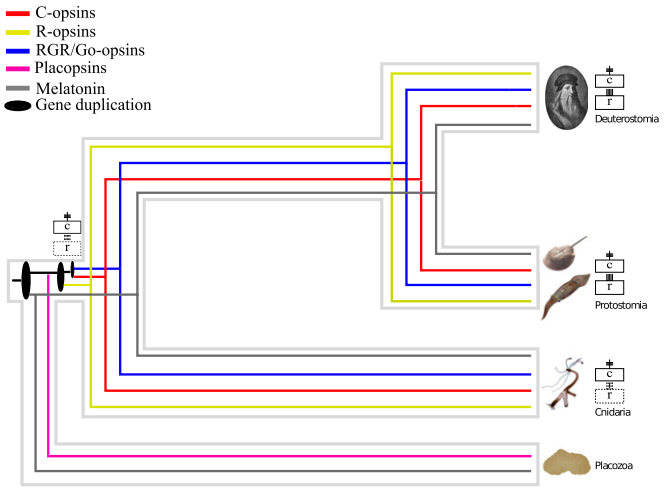


Fig. 2. Synopsis of opsin evolution. This figure represents a gene tree embedded within a species tree illustrating the evolutionary history of the opsins and MLT receptors in Metazoa. It shows that only three duplications and no deletions are necessary to explain opsin evolution.

Methods

Data Mining, Dataset Assembly, and Alignment. Taxonomic nomenclature follows the work of Nielsen (23). We assembled a large sample of well-characterized opsins from across Neuralia (*SI Appendix*, Table S1), including key sequences like the putative cnidarian C opsin of Plachetzki et al. (1) and the putative cnidarian R opsins of Suga et al. (4). To identify the closest outgroup(s) of the neuralian opsins, representatives of each monophyletic α-group of Rhodopsin-like receptors, and a set of sequences from the β-, γ-, and δ-groups (for a total of 139 sequences) were downloaded from GPCRDB (www.gpcrdb.org) and added to our dataset (*SI Appendix*, Table S1). Sequences in GPCRDB are of vertebrate origin. To enrich our data set of putative GPCRs from early branching metazoans, we mined the genomes of *H. magnipapillata*, *N. vectensis*, *T. adhaerens*, *A. queenslandica*, and *O. carmela* (*SI Appendix*). Our final dataset included 625 GPCRs (499 opsins and 176 putative opsin outgroups). From this data set, we generated two master alignments (26) (*SI Appendix*). The first alignment, the AOM alignment, included only the 499 neuralian

opsins. The second alignment, the G&OM, included all putative opsin outgroups (176 GPCRs in total) and a sample of 80 selected opsins (as detailed later; *SI Appendix*). The AOM and G&OM alignments were, respectively, 317 and 366 positions long. A third alignment was generated a posteriori after having inspected the results of the analyses of G&OM (as detailed later; Fig. 1*B*) to identify the closest sister group of the animal opsins. This third alignment, O&O, included the 80 opsins in G&OM plus the closest sister group of the animal opsins only (i.e., the MLT receptors; Fig. 1*B*). O&O included 104 sequences and was 366 positions long. All alignments are available upon request.

Phylogenetic Analyses and Ancestral Character State Reconstructions. In this section, we will focus on the logic of our analytical scheme. Technical details of the analyses performed are reported in *SI Appendix*. The AOM alignment was analyzed to recover an unrooted phylogeny including only well-characterized opsins from the three known bilaterian subfamilies (C, R, and Go/RGR) and an inclusive sample of cnidarian opsins. This analysis allowed the evaluation of the relative relationships among the cnidarian opsins in our data set, including those of Plachetzki et al. (1) and Suga et al (4). Results of the AOM analyses were used to select a subset of 80 opsins (20 C opsins, 20 R opsins, 20 Go/RGR opsins, and 20 cnidarian opsins) to be included in the G&OM and O&O data sets. Opsin subsampling was necessary to (*i*) reduce computational complexity and (*ii*) minimize the likelihood of tree reconstruction artifacts. Accordingly, fast-evolving, extremely short, and compositional heterogeneous sequences were not included in the G&OM and O&O alignments. However, a representative sample of sequences from every opsin clan identified in AOM was retained.

The G&OM alignment was analyzed to identify the closest outgroup of the opsin family. This alignment included the complete set of 176 putative opsin outgroups we identified. Because the closest opsin outgroup must belong to the α-group of Rhodopsin-like receptors, the G&OM phylogeny was rooted by using two γ-group receptors: two Galanin-like receptors (12).

To clarify the duplication and deletion history within the opsin family, we analyzed O&O, which we rooted by using the closest opsin outgroup (identified from the results of the G&OM analyses) only. Accordingly, O&O is simply a modification of G&OM from which distantly related opsin outgroups were excluded to minimize systematic artifacts (20–22, 31, 35).

The three alignments (AOM, G&OM, and O&O) were analyzed by using Bayesian tree reconstruction methods. O&O was also analyzed by using ML. The AU test was used to compare our O&O phylogeny against those from previous studies (1–4). Bayesian and ML-based ancestral character state reconstruction were performed to infer the sequence of the RBD at key internal nodes (LOCA and LOCNA).

ACKNOWLEDGMENTS. We thank Scott Nichols and Nicole King for providing access to the genome of *Oscarella carmela*, Stuart Longhorn for help in assembling our opsin data set, and Omar Rota Stabelli for discussion and suggestions. This work was supported by Irish Research Council for Science, Engineering, and Technology PhD fellowships (to R.F. and S.C.H.); Science Foundation Ireland Research Frontier Programme Grants 11/RFP/EOB/3106 and 09/RFP/EOB2510 (to D.P. and J.O.M.). All analyses were performed using the infrastructures provided by the Irish Centre for High End Computing and the National University of Ireland Maynooth supercomputing facility.

1. Plachetzki DC, Degnan BM, Oakley TH (2007) The origins of novel protein interactions during animal opsin evolution. *PLoS ONE* 2(10):e1054.
2. Plachetzki DC, Fong CR, Oakley TH (2010) The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. *Proc Biol Sci* 277(1690):1963–1969.
3. Porter ML, et al. (2011) Shedding new light on opsin evolution. *Proc Biol Sci* 279:3–14.
4. Suga H, Schmid V, Gehring WJ (2008) Evolution and functional diversity of jellyfish opsins. *Curr Biol* 18(1):51–55.
5. Land M, Nilsson DE (2002) *Animal Eyes* (Oxford Univ Press, Oxford, UK).
6. Nilsson DE, Gislén L, Coates MM, Skogh C, Garm A (2005) Advanced optics in a jellyfish eye. *Nature* 435(7039):201–205.
7. Kozmik Z, et al. (2008) Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proc Natl Acad Sci USA* 105(26):8989–8993.
8. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311(5762):796–800.
9. Gehring WJ (2011) Chance and necessity in eye evolution. *Genome Biol Evol* 3: 1053–1066.
10. Nathans J, Hogness DS (1983) Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. *Cell* 34(3):807–814.
11. Terakita A (2005) The opsins. *Genome Biol* 6(3):213.
12. Fredriksson R, Lagerström MC, Lundin LG, Schiöth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* 63(6):1256–1272.
13. Koyanagi M, et al. (2008) Jellyfish vision starts with cAMP signaling mediated by opsin-G(s) cascade. *Proc Natl Acad Sci USA* 105(40):15576–15580.
14. Srivastava M, et al. (2008) The Trichoplax genome and the nature of placozoans. *Nature* 454(7207):955–960.
15. Chapman JA, et al. (2010) The dynamic genome of Hydra. *Nature* 464(7288):592–596.
16. Putnam NH, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317(5834):86–94.
17. Srivastava M, et al. (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466(7307):720–726.
18. Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N (2012) Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proc Natl Acad Sci USA* 109(32):13046–13051.
19. Erwin DH, et al. (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* 334(6059):1091–1097.
20. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol Biol Evol* 26(10):2261–2274.
21. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19(8):706–712.
22. Pick KS, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27(9):1983–1987.
23. Nielsen C (2012) *Animal Evolution* (Oxford Univ Press, London).
24. Fain GL, Hardie R, Laughlin SB (2010) Phototransduction and the evolution of photoreceptors. *Curr Biol* 20(3):R114–R124.
25. Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM (2007) Of clades and clans: Terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol* 22(3): 114–115.
26. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320(5883):1632–1635.
27. Foster PG (2004) Modeling compositional heterogeneity. *Syst Biol* 53(3):485–495.
28. Rivera AS, et al. (2012) Blue-light-receptive cryptochrome is expressed in a sponge eye lacking neurons and opsin. *J Exp Biol* 215(Pt 8):1278–1286.
29. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492–508.
30. Schierwater B, et al. (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol* 7(1):e20.
31. Philippe H, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* 9(3):e1000602.
32. Dunn CW, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.
33. Nordström K, Wallén R, Seymour J, Nilsson D (2003) A simple visual system without neurons in jellyfish larvae. *Proc Biol Sci* 270(1531):2349–2354.
34. Erwin DH (2009) Early origin of the bilaterian developmental toolkit. *Philos Trans R Soc Lond B Biol Sci* 364(1527):2253–2261.
35. Holton TA, Pisani D (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol* 2:310–324.

Molecular Phylogenetics and Evolution 62 (2012) 27–34

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships

Davide Pisani [a], Roberto Feuda [a], Kevin J. Peterson [b], Andrew B. Smith [c,*]

[a] The National University of Ireland, Maynooth, Co. Kildare, Ireland
[b] Department of Biological Sciences, Dartmouth College, N. College Street, Hanover, NH 03755, USA
[c] Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW75BD, UK

ARTICLE INFO

Article history:
Received 1 March 2011
Revised 8 July 2011
Accepted 31 August 2011
Available online 14 September 2011

Keywords:
Molecular evolution
Long-branch attraction
Echinoderms
Phylogeny
MicroRNA
Molecular clocks

ABSTRACT

Resolving evolutionary relationships in groups that underwent fast radiation in deep time is a problem for molecular phylogeny, as the scant phylogenetic signal that characterises short internal branches is generally swamped by more recent substitutions. We implement an approach, that maps how the support for rival phylogenies changes when analysing subsets of sites with either faster and more heterogeneous rates or slower and more homogeneous rates, to address a long-standing problem in deuterostome phylogeny – the interrelationships of the eleutherozoan echinoderm classes. We show that miRNA genes are phylogenetically uninformative as to the relationships of asteroids, echinoids and ophiuroids, consistent with a rapid radiation of these groups as suggested by their fossil record. Using three nuclear rRNAs and seven nuclear housekeeping genes, we map the support for the three possible phylogenetic arrangements of asteroids, ophiuroids and echinoids when moving between subsets of the data with very similar or very different rates of evolution. Only one of the three possible topologies (asteroids (ophiuroids + echinoids)) strengthens when the most rate-homogeneous subset of data are analysed. The other two possible pairings become stronger in a less reliable data subset, which includes the fastest and thus homoplasy-rich data in our alignment. Thus, while superficial analysis of our concatenated alignment identifies asteroids and ophiuroids as sister taxa, more thorough analyses suggest that ophiuroids may be more closely related to echinoids. Divergence of these echinoderm groups, using a relaxed molecular clock, is estimated to have occurred within ~5 million years. Our results illustrate that the analytic approach of phylogenetic signal dissection can be a powerful tool to investigate rapid radiations in deep geologic time.

1. Introduction

Both morphological and molecular approaches to phylogenetic reconstruction work well when divergences between taxa are separated by relatively long time intervals, as the accumulation of substantial numbers of derived characters in the stem lineages creates a strong phylogenetic signal. However, when divergence occurred rapidly in deep time and stem lineages are of short duration, accurate phylogenetic reconstruction is difficult. This is because continuing evolution results in convergence and reversals that ultimately overwhelm the weak signal in short internal branches. In such situations, unequal rates of evolution can lead some branches to accumulate a significantly larger number of substitutions leading to the well-known problem of long-branch attraction (LBA: Felsenstein, 1978). While LBA has long been recognised as a problem, how best to identify trees affected by LBA and tease out historical signal from systematic biases remains a major challenge (Brinkmann and Philippe, 1999; Ruitz-Trillo et al., 1999; Pisani, 2004; Lartillot and Philippe, 2008; Jeffroy et al., 2006; Sperling et al., 2009; Rota-Stabellli et al., 2010). Indeed, while the signature of rapid divergence is a phylogenetic tree where branching order cannot be resolved with confidence, LBA can confuse the picture causing the recovery of artefactual groups with very high support (Jeffroy et al., 2006).

One problematic area of the metazoan tree concerns how the five echinoderm classes are related (Smith et al., 2004; Janies et al., 2011). Both morphology and molecular data place crinoids as sister group to the other classes (echinoids, asteroids, ophiuroids, holothurians), and pair echinoids and holothurians together. Yet the interrelationships of asteroids, ophiuroids and the echinoid–holothurian clade remain disputed. Morphological data favours either an asteroid–ophiuroid pairing (Mooi and David, 2000) or an ophiuroid plus echinoid–holothurian pairing (Littlewood et al., 1997), whereas different molecular analyses have found support for all three possible groupings (Field et al., 1988; Littlewood et al., 1997; Janies, 2001; Mallatt and Winchell, 2007; Pereske et al., 2010; Janies et al., 2011; Letsch and Kjer, 2011).

* Corresponding author.
E-mail address: a.smith@nhm.ac.uk (A.B. Smith).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.08.028

These echinoderm clades pose a particularly acute problem for molecular phylogenetic analyses because they underwent crown group diversification long after they had split from one another and all three have long stem groups that cannot be broken up by selective sampling of the modern fauna, making them particularly susceptible to LBA.

2. Materials and methods

2.1. Molecular data assembled

Total RNA was collected from the ophiuroid *Ophiopholis* and a small RNA library constructed and sequenced following Wheeler et al. (2009), resulting in 3804 parsed non-redundant reads. These were then compared with previously published small RNA libraries drawn from an asteroid (*Henricia sanguinolenta*), echinoid (*Strongylocentrotus purpuratus*), hemichordate (*Saccoglossus kowalevskii*) and other metazoans published previously and analysed by miRMiner (Wheeler et al., 2009) for known and potentially novel miRNAs (Table S1).

Six nuclear housekeeping genes (aldolase, methionine adenosyltransferase, ATP synthase beta chain, elongation factor 1 alpha, triosephosphate isomerase and phosphofructokinase) were sequenced from the ophiuroid *Ophiopholis* sp. following the protocol described in Sperling et al. (2009). These sequences have been deposited in Genbank under accession numbers (JN716365–JN716370). Sequences for *Aplysia californica*, *Alvinella pompejana* and *Tubifex tubifex*, as well as three genes for *Carinoma mutabilis*, were downloaded from the NCBI trace archives. Unpublished sequences from *Chaetopleura apiculata* and *Leptochiton asellus* were kindly provided by J. Vinther (Yale University). Sequences for other lophotrochozoan taxa were taken from previously published reports (Peterson et al., 2004), and new sequences were manually added to the pre-existing alignment used, for example in Sperling et al. (2011). Data for ribosomal 5.8S, 18S and 28S ribosomal genes for 22 deuterostome, 35 lophotrochozoan, and 15 ecdysozoan taxa were assembled, either taken directly from Mallatt et al. (2010) or downloaded from the NCBI Genbank website and manually aligned to the Mallatt et al. (2010) sequences. Chimaeras at the generic level were permitted when data for the same species were not available. After the removal of minor indels, the amino acid matrix was 88% complete and the ribosomal matrix was 76% complete. The seven nuclear housekeeping genes (2049 amino acids in total) and three ribosomal genes (4682 nucleotides in total) were concatenated for analysis.

2.2. Sequence analysis

2.2.1. Conventional phylogenetic analysis

The protein and rRNA partitions were first independently analysed to investigate the nature of the principal signal (Pisani and Wilkinson, 2002) in these data sets. Protein analyses were performed using the heterogeneous CAT-GTR model, and rDNA analyses were performed using the GTR + G model, which proved to be the best fitting model (selected using MrModeltest) for our nucleotide data. CAT-GTR analyses were performed in Phylobayes V. 3 (Lartillot and Philippe, 2004). We used posterior predictive analysis as implemented in Phylobayes (see also Sperling et al., 2009) to discover whether the taxa of interest (i.e. the echinoderms) were compositionally homogeneous or heterogeneous.

The rRNA and protein partitions were concatenated and analysed under mixed models using Bayesian and Maximum Likelihood (ML) analyses. Maximum Parsimony (MP) and Neighbour Joining (NJ) (with uncorrected P distances and no gamma correction) were also performed. Bayesian analyses were performed using MrBayes 3.1 (Huelsen-beck and Ronquist, 2001), ML analyses were performed using RAxML (Stamatakis, 2006), while MP and NJ analyses were performed using PAUP4b10 (Swofford, 2002). Support for nodes found in the MP, NJ and ML analyses was estimated using the bootstrap, with 500 replicates for MP and NJ (but see Supplementary information) and 5000 replicates for ML.

For all Bayesian mixed models analyses both the rRNA and protein partitions were modelled using GTR + G. Sperling et al. (2009) showed that for this protein data set, GTR + G is the best fitting amongst the homogeneous substitution models implemented in MrBayes, whilst we showed here that GTR + G is the best fitting model for our nucleotides partition. CAT-GTR analyses could not be performed for the concatenated data set because of software limitation (Lartillot, pers. comm.). For the ML analyses the protein partition was modelled using LG + G. The nucleotide partition was modelled using GTR + G.

2.2.2. Phylogenetic signal dissection

Both the rRNA and the Protein data sets were partitioned into sets of "homogeneously evolving" and "heterogeneously evolving" sites using a modification of Brinkmann and Philippe's (1999) slow–fast approach (see Sperling et al., 2011 for justifications). This method assigns rates to characters semi-independent of tree topology. The characters in the rRNA and protein data sets were independently ranked according to their evolutionary rate (estimated as slow–fast parsimony scores) and partitioned into four quartiles. For each data set (proteins and rRNAs) characters were split into two groups: the first containing all the sites in the fourth quartile plus invariant sites, the second contained all the variant sites in the first, second and third quartiles. The characters in the first data partition represent a combination of sites with highly heterogeneous rates (i.e. very fast and constant sites only). This partition included 1247 AA and 3206 NN positions, of which 748 AA and 2332 NN positions were constant and 499 AA and 874 NN where deemed to be fast evolving. Because of the extreme rate variation (including constant and fast evolving sites only), and the high substitution rates and homoplasy levels of the variable characters it includes, this data partition presents a hard phylogenetic problem, and is prone to generate phylogenetic artefacts (e.g. LBA) even when analysed using well-fitting, parameter-rich models. The second data partition is composed of phylogenetically more reliable, rate-homogeneous, characters of slow to intermediate evolutionary rate. This partition includes 811 AA and 1476 NN (all of which are parsimony informative) and is more likely to support relationships that represent historical signal (see Sperling et al., 2009, 2011; Rota-Stabellli et al., 2010).

We then evaluated the strength of the signals supporting the three possible arrangements of asteroids, echinoids and ophiuroids residing in the three data sets (i.e. all sites, rate-heterogeneous sites and rate-homogeneous sites), under three, differently performing, methods – Parsimony, Neighbour Joining and Bayesian analysis. The fit to data of the three topologies (see Fig. 1) into which asteroids, echinoids and ophiuroids can be arranged were compared using Bayes Factors (BF; e.g., Sperling et al., 2010; Holton and Pisani, 2010) as follows. For each data set (homogeneous, heterogeneous and all sites), and each sister-group hypothesis (E + O, A + E and E + A), a constrained tree search (of 2 runs and four chains per run) was performed in MrBayes (Huelsenbeck and Ronquist, 2001). Each constrained tree search was run for 5,000,000 generations and a burn-in of 2,500,000 generations was used. This burn-in period was sufficiently long to allow each analysis to converge, and generated an identical number of data points (per data set and hypothesis) to calculate the BF. For each data set, the MrBayes ".p" file corresponding to the chain of maximal marginal likelihood across all trees (estimated using the harmonic mean) was selected, and used to estimate the BF for each pair of considered hypotheses in Tracer v1.5.1 (Rambaut and

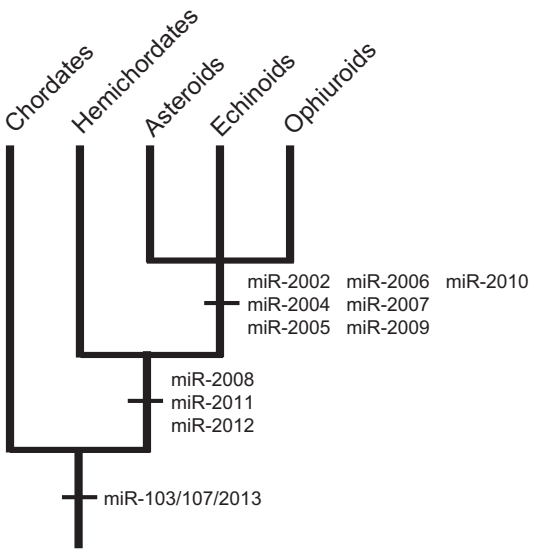


Fig. 1. MicroRNA phylogeny of deuterostomes highlighting the lack of phylogenetically informative characters for resolving echinoderm class relationships.

Drummond, 2007). Because the variance around the BF harmonic means can be extremely large (Lartillot and Philippe, 2005), we followed the suggestion of Marc Suschard (unpublished but see http://groups.google.com/group/beast-users/browse_thread/thread/3e9d7da1eeb9d6c8/9e3aa8eb29c76978?pli=1), that BF be calculated multiple times from the same data to estimate how much the results vary. Here, we have calculated the BF twice, starting from two independent MrBayes runs (time and computational limitations prevented us from performing more independent tests). In addition, all our BF results are presented in association with Standard Errors around the calculated harmonic means. For all BF analyses the protein and the nucleotide partitions were modelled using two unlinked GTR + G models.

To estimate the strengths of alternative signals in the three data sets we performed bootstrap analyses under MP, NJ and ML and compared the support for the three alternative topologies. Similar analyses could not explicitly be performed in a Bayesian framework, because, for each of the tree data sets, only one of the alternative hypotheses was supported. For each data set, some of the alternative hypotheses received low to extremely low levels of support, indicating that the strength for that signal in that data set was minimal. To evaluate whether these low support values represented an artefact of the bootstrap resampling, or a real feature of the data, for each data set we performed multiple bootstrap analyses (only under MP and NJ). In these analyses the number of replicates was incrementally augmented. Bootstrap analyses were performed (in Paup4b10) using 100, 500, 1000, 5000 and 10,000 replicates by which time the signal had stabilised around a given value.

2.3. Molecular clock analyses

All relaxed molecular clock analyses were performed using the software Phylobayes version 3 (Lartillot and Philippe, 2004) following the protocol of Sperling et al. (2010) and using the CIR model, an autocorrelated model that fits this data set better than uncorrelated models (Sperling et al., 2010). An additional 29 sponge, seven cnidarian and two non-metazoan outgroups (the choanoflagellate *Monosiga brevicollis* and the yeast *Saccaromices cerevisiae*) were added to the dataset to maximise the number of calibration points. Clock analyses used a fixed topology based on the results of the homogeneous data set only (i.e. with an ophiuroids plus echinoids grouping), combined with the results of Sperling et al. (2010). Branch lengths for this fixed topology (Table S9) were re-estimated under the CAT-GTR model using only the protein alignment. A total of 24 calibration points, spread phylogenetically throughout Metazoa and spaced temporally from the Miocene to Cryogenian, were used (Supplementary data, Table S2). Using Phylobayes two chains were initially run using soft bounds and allowing 5% of the prior probability density to lie outside of the minimum–maximum interval defined for each calibration point. Further analyses were performed to test the effect of different levels of relaxation on the recovered ages. We calculated divergence times allowing 10%, 25% and 50% of the prior probability density of each calibration point to lie outside the min–max interval defined by the provided calibration points. Analyses were also run with no-data to test the effect of our calibrations on the unconstrained nodes; this was done to test whether "composite calibration points" (i.e. the effect of multiple surrounding calibration points on intervening nodes) could have biased our results. The root node in our molecular clock analyses represents the split between Fungi and the Holozoa, and all the above-mentioned analyses were run using a prior root age of 1000 Ma and a standard deviation of 100 Ma. Analyses performed using the 5% relaxation level were also performed using a significantly deeper prior root (1600 Ma) and a SD of 700 Ma to test the effect of the root-prior on our divergence times.

3. Results

3.1. MicroRNA markers in echinoderms

Virtually all expected miRNAs were discovered in our ophiuroid small RNA library, including the deuterostome-specific miR-103/107/2013, the ambulacrarian-specific miRNAs miR-2008, -2011 and -2012, and seven echinoderm-specific miRNAs (Fig. 1). No miRNA shared between any two of the three echinoderms to the exclusion of the third was found (Supplementary data, Table S1). Only six potential miRNA sequences were shared among at least two of these three taxa, but none of these was a novel miRNA (three were transfer RNA sequences, and the other three were edits to known miRNAs).

3.2. Phylogenetic analyses of standard sequence data

Posterior predictive analysis showed that amino acid sequences in all species relevant to this study (and the great majority of the species in this data set) are compositionally homogeneous (Supplementary data, Table S3). The nucleotide data do, however, show compositional heterogeneity, although the ophiuroid and most echinoid sequences are compositionally homogeneous (Supplementary data, Table S4). This is not considered a problem for our analyses, as we never observe compositionally heterogeneous taxa grouping together.

As the relationships between outgroup organisms are essentially static across analyses, only the three taxa under consideration – echinoids (E), asteroids (A) and ophiuroids (O) – are discussed. Analyses of the rDNA (Supplementary Fig. S1) and of the protein data (Supplementary Fig. S2) partitions found low support for the grouping A + O (Posterior Probabilities (PP) = 0.56 (rDNA) and 0.49 (protein)). This group is then sister to the echinoids in both the rDNA (PP = 0.97) and protein data (PP = 1). Analyses of the concatenated (proteins and rRNA) data set performed under mixed models (Fig. 2A, left; Supplementary Fig. S3) also support a sister group of A + O, but with higher support (PP = 0.97; ML-BP = 0.71). This group is then the sister of the echinoids (PP = 1). Although data concatenation increased the

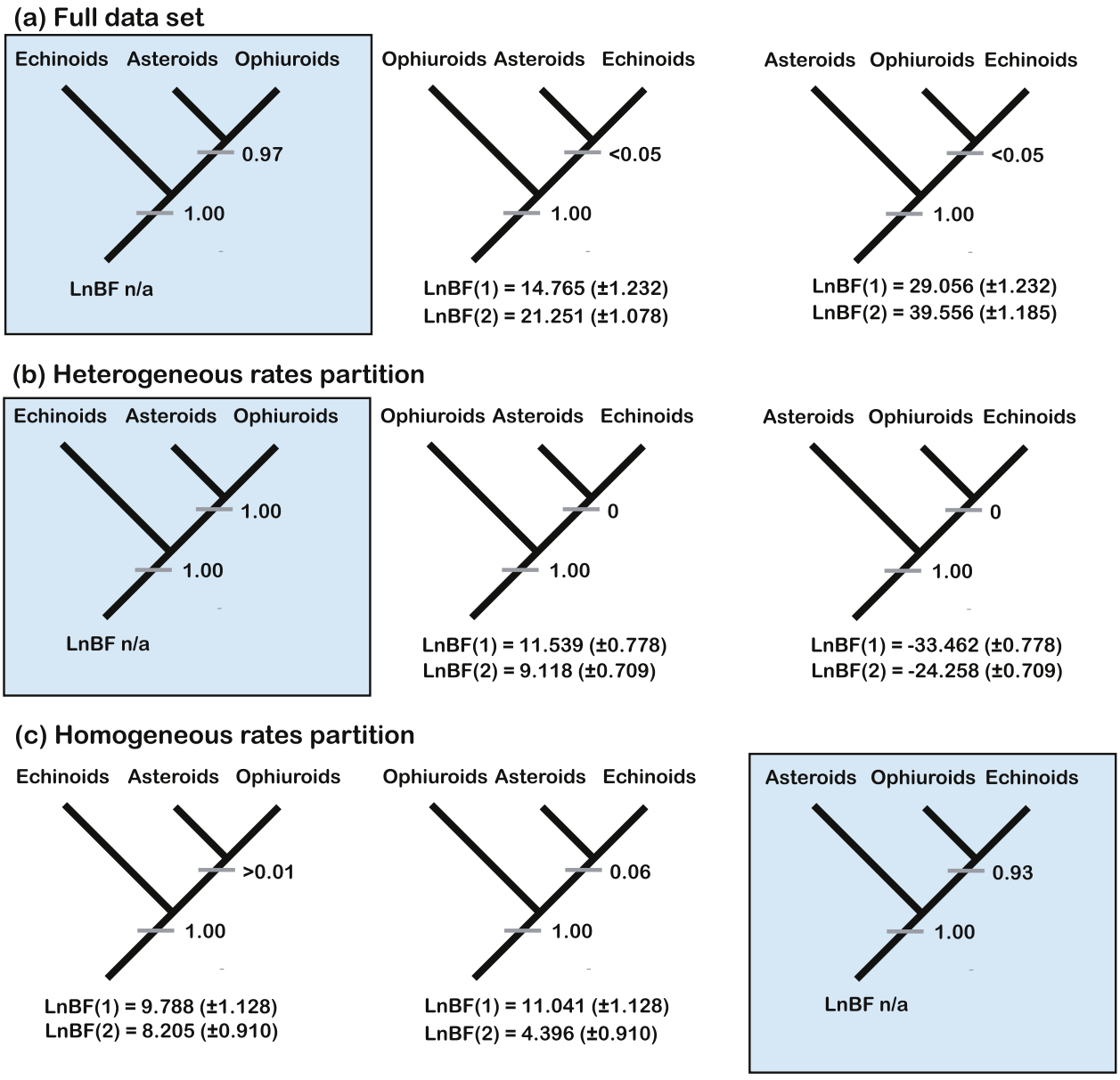


Fig. 2. Cladograms summarising levels of Bayesian Factors support for alternative potential topologies under different data partitions of the combined rDNA and protein sequences A, full sequence; B, partition of the quartile of fastest evolving sites plus invariant sites (heterogenous partition); C, partition of the slow and intermediate evolving sites only (homogeneous partition). Shaded box indicates best-supported topology in each case.

support for A + O (Fig. 2A), a common feature of these trees (Supplementary Figs. S1-3) is that the asteroid and ophiuroid terminal branches are long whilst the internal branch uniting them is very much shorter, raising the possibility that the pairing of asteroids and ophiuroids could be the result of LBA.

3.3. Phylogenetic signal dissection of rate-homogeneous and rate-heterogeneous data partitions

The fit of our data sets to the three alternative phylogenetic hypotheses was tested using Bayes Factors (Supplementary data, Tables S5-7). Results obtained in the two independent BF analyses are in full agreement, and comparisons of the estimated harmonic means (with their bootstrapped confidence intervals) show that uncertainty around the estimated harmonic means should not be a problem for our analyses.

When all sites are used BF clearly supports the A + O pairing and finds least support for the E + O pairing (Fig. 2A). Analysis of the heterogeneous data partition also identified strong support for the pairing A + O (PP = 1; ML-BP = 79 Fig. 2B). The homogeneous data partition, however, supports a different set of relationships, pairing O + E (PP = 0.93 Fig. 2C). Support for an A + O, or E + A pairing is minimal in this subset of data (less than 0.1 for both hypotheses; Fig. 2C). Exactly the same pattern is recovered when the analyses are repeated under ML, with higher support for an E + O pairing appearing in the homogeneous data partition, although support values are much lower and non-significant. Thus, the signal that groups A + O is strongest in the heterogeneous partition, whereas that for E + O is strongest in the homogeneous partition. Indeed, E + O is the only grouping that is better supported in the homogeneous partition (Fig. 2C), than in either the heterogeneous data partition or the full alignment. Note that the precise level of support for this group is method dependent, being higher under Bayesian analysis than maximum likelihood. This in part is to be expected as the bootstrap is known to be over-conservative whilst posterior probabilities might be too optimistic (e.g. Douady et al.,

2003). However, in this case the lower support obtained under ML (given also the substantial difference in support observed), most likely reflects the poor ability of MCMC methods to deal with complex models and mixed data sets (e.g. Lartillot and Philippe, 2004). Accordingly, we suggest the results of the Bayesian analysis, in this specific case, may better describe the strength of the signal in the compared data sets.

When the full data set is analysed using NJ and observed distances, a method and a distance measure that perform poorly and are easily swayed by LBA, support is again found for an A + O pairing (BP = 55%) (Fig. 3). The support for this group reaches a maximum of 65% in the NJ analysis of the heterogeneous data, and drops to 33% in the NJ analysis of the homogeneous data. Parsimony analysis (which is also easily swayed by LBA) of the complete data set finds virtually no support for E + O pairing (BP = 3%) and minimal support for the A + O group (BP = 15%), favouring instead a pairing of E + A (BP = 81%) (Fig. 3). As found with the NJ analyses, when the heterogeneous data are analysed with MP, the support for A + O rises to 36%, whereas support for E + A decreases to BP = 63% and support for E + O drops to zero. However, when the homogeneous data are analysed with MP, support for E + O increased to 55%, whilst support for E + A and A + O decreased to 26% and zero, respectively. That these differences are not stochastic variations associated with the heuristic nature of the bootstrap is demonstrated by the consistency of the differences observed (Fig. 3).

3.4. Molecular divergence estimates

Using a relaxed molecular clock methodology we find that the divergence amongst the sampled eleutherozoan echinoderms is estimated to be Early Ordovician ~480 Ma (95%CI = 505–446) (Fig. 4). Consistent with the fossil record (Dean-Shackleton, 2005; Smith and Savill, 2001), we estimate that ophiuroids and echinoids diverged very soon afterwards, roughly 475 Ma (95% CI = 501–440) (Fig. 3). Thus, this is indeed a very rapid divergence, spanning approximately 5 million years. Sensitivity analyses indicate that our dates are robust and unlikely to have been caused by the use of inappropriate fossil calibrations. Running the analyses under the priors shows that our set of calibrations seem appropriate to address the problem at hand (not shown). Relaxing the soft bounds to allow up to 10%, 25%, or 50% of the prior probability density to lie outside of the minimum–maximum interval of each considered calibration point caused negligible changes to estimated echinoderm ages, and in two cases (10% and 25%) cases still recovered divergence times that lay within the 95% confidence interval of the analysis run under the default 5% relaxation level (Supplementary data, Table S8). Finally, changing the root prior age did not significantly affect our recovered divergence times (Supplementary data, Table S8).

4. Discussion

MicroRNAs are a diverse family of small, non-coding regulatory genes present throughout Bilateria. Because they are continually added to over time, rarely change in primary sequence and are only rarely secondarily lost in most taxa, they are considered reliable phylogenetic markers (Sperling et al., 2010; Sperling and Peterson, 2009; Heimberg et al., 2010; Rota-Stabellli et al., 2010). Yet unexpectedly we found no unique microRNAs to resolve the asteroid–echinoid–ophiuroid trichotomy (Fig. 1). Polytomies that cannot be resolved using microRNAs must be considered as potentially having undergone rapid divergence, a possibility also suggested

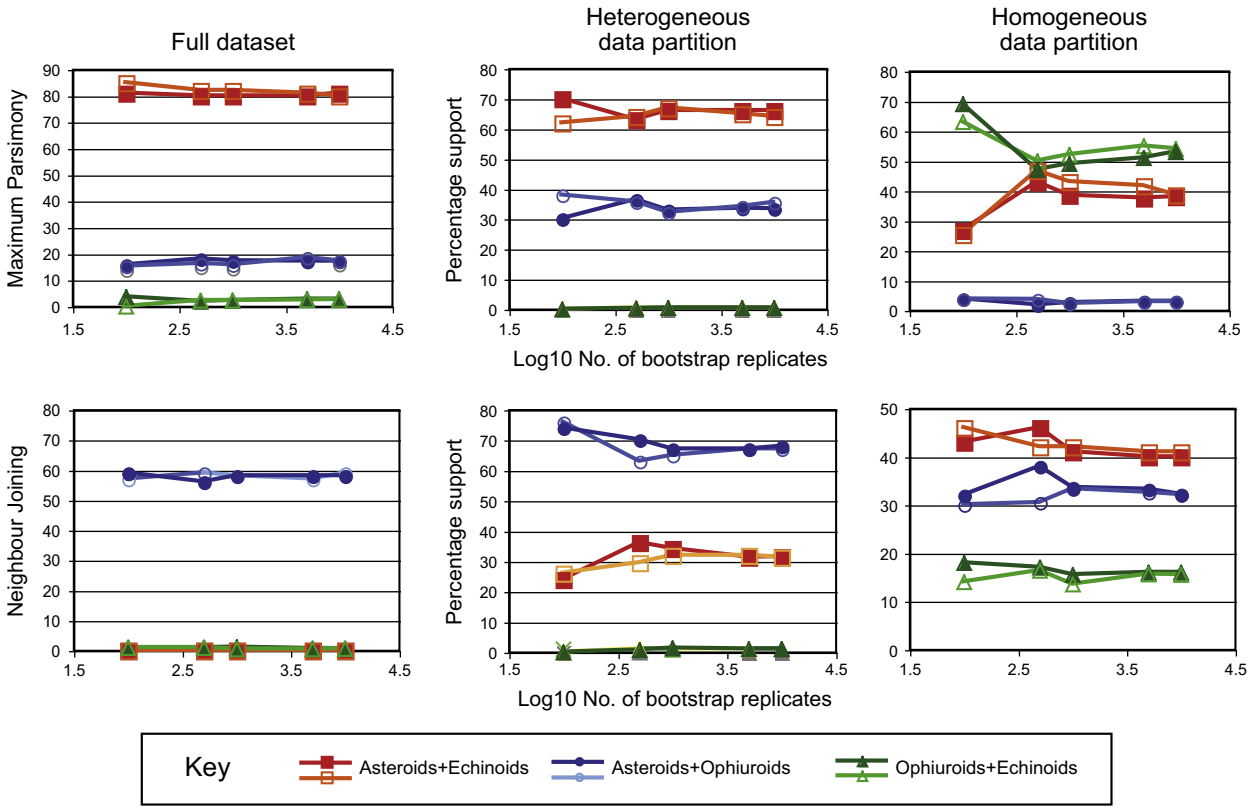


Fig. 3. Plots of bootstrap support for the asteroid + ophiuroid, asteroid + echinoid and ophiuroid + echinoid pairings in the rDNA and protein coding gene sequences as described in this paper. Bootstrap analyses were carried out using 100, 500, 1000, 5000 and 10,000 replicates under both maximum parsimony and neighbour joining for the full dataset, the heterogeneous data partition and the homogeneous data partition.

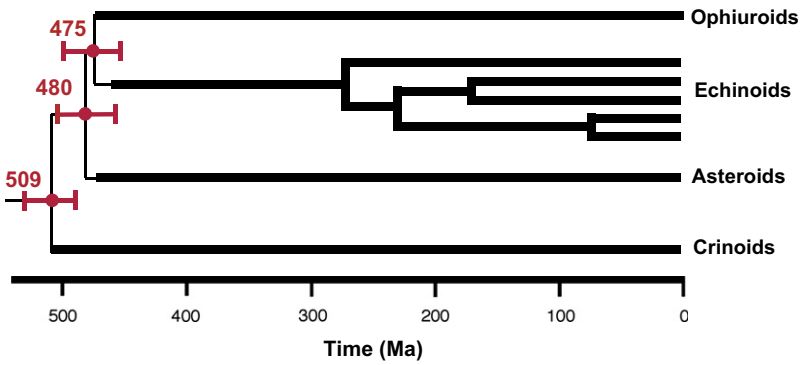


Fig. 4. Time of divergence estimates based on a relaxed molecular clock approach (see text). Thick lines = known fossil record.

by the fossil record (Smith, 1988), the volatile phylogenetic signal that emerges from gene sequence data (Fig. 2), and our molecular divergence estimates (Fig. 4).

Reconstructing relationships of clades that have radiated rapidly deep in geological time has proved to be particularly difficult, and the general approach adopted to address such problems has been to try and tease out a weak signal using larger and larger data sets (e.g. Holton and Pisani, 2010; Rota-Stabellli et al., 2010; Dunn et al., 2008; Hejnol et al., 2009). However, while increasing the dimension of the data set can eliminate stochastic errors, it will also exacerbate systematic errors like LBA (Sperling et al., 2009; Pick et al., 2010). Indeed, it is a misconception that simply adding more data will eventually lead to the recovery of the correct phylogeny; if the data are affected by LBA then the opposite will happen (Jeffroy et al., 2006).

The comparison of phylogenies obtained using differently fitting substitution models (and differently performing phylogenetic methods) has previously been used to identify LBA artefacts. This is because well-fitting substitution models (e.g. the CAT model of Lartillot and Philippe, 2008), and optimal outgroup selection (Rota-Stabellli and Telford, 2008) can help reduce LBA (Sperling et al., 2009; Rota-Stabellli et al., 2010; Rodriguez-Ezpeleta et al., 2007). An alternative, but less frequently used, strategy to circumvent LBA is to exclude sites with high evolutionary rate (i.e. site stripping) from the analyses (e.g. Brinkmann and Philippe, 1999; Rodriguez-Ezpeleta et al., 2007; Pisani, 2004; Sperling et al., 2009; Rota-Stabellli et al., 2010).

Wägele (1999) identified three classes of LBA artefact: symplesyomorphy trap (Type I LBA), erosion of phylogenetic signal (Type II LBA), and misleading and invisible attraction due to non-homologous similarities (Type III LBA). Each affects tree topology in a different way, producing artefactual topologies with different characteristics. However, all stem from the same phenomenon: the existence of substantially different lineage-specific substitution rates. Site-stripping approaches that exclude sites that accumulate substitutions at high rate (thus contributing to LBA), certainly help circumvent Type II and III artefacts, but it is unclear how much site stripping can help circumventing Type I artefacts. However, the application of site stripping should not exacerbate Type I LBA artefact, so long as only fast evolving sites are excluded. This is because rapidly evolving sites tend to be saturated and hence rich in homoplasy (including reversals) and poor in true apomorphies. True apomorphies are concentrated rather in sites of either slow or (most likely) intermediate rate, which we retain. Accordingly, exclusion of fast sites should not increase the true plesiomorphy to true apomorphy ratio in the data set, which is ultimately responsible for Type I LBA artefacts (see Wägele, 1999). In any case, it is clear that, even for data affected by Type I LBA, if noisy (fast) sites are excluded true but weak phylogenetic signal is more likely to emerge. Hence, we would expect that exclusion of sites of high rate (where multiple substitution are more likely to accumulate) should have a generally positive (or in the worst case neutral) effect independent of the LBA type affecting a data set.

It is important to bear in mind that site-stripping based methods are not the only possible approach to attempt circumventing systematic artefacts (see Jeffroy et al., 2006 and references therein), and they should not be considered a generalised panacea. Their utility is limited to deep time studies where anciently acquired substitutions at fast evolving sites are likely to have been erased by subsequent substitutional events.

In contrast to standard site stripping approaches, where only sets of slowly evolving sites are analysed, our approach (see also Sperling et al., 2009, 2011) compares the strength of phylogenetic signals in both the slow and fast evolving data, thus effectively associating the various signals to subsets of data. Signals associated with fast evolving sites most likely characterise artifactual groups, while those associated with slowly evolving sites are more likely to support real clades. Our results suggest that there is a partitioning of the signals within this data set, with the signal supporting the pairing of asteroids and ophiuroids concentrated in the fast (i.e., heterogeneous) positions and that supporting the pairing of echinoids and ophiuroids concentrated in the slow (i.e., homogeneous) positions. Support for the pairing of echinoids and asteroids is more widely distributed with some support for this group present in the homogeneous partition but with the majority residing in the heterogeneous partition.

These results clearly illustrate a serious, often underestimated potential pitfall of supermatrix analyses, that a clade with strong support might not necessarily be real. Our conventional analysis of aligned gene sequence finds strong and unambiguous support for an asteroid–ophiuroid pairing. However, data partitioning suggests this is most likely an LBA artefact since the data set scoring the most unreliable sites in our alignment strongly support an asteroid–ophiuroid pairing and provide a better fit than either the slow-evolving sites or the complete data to trees displaying this clade. In contrast, the analyses of the homogeneous data set tend to support an ophiuroid + echinoid grouping. This indicates that the signal for this clade is concentrated in the slowly evolving (more reliable) sites (Fig. 2), and is thus likely to represent phylogenetic signal. While this signal is not very strong, this is to be expected given that the signal for this grouping is swamped by other signals in the complete data set. Our signal dissection approach therefore provides a simple means of distinguishing those groupings more likely to be driven by LBA (e.g., asteroid + ophiuroid) from those more likely to represent genuine phylogenetic signal (e.g., echinoid + ophiuroid).

These results help explain why previous molecular analyses have come to different conclusions concerning the interrelationships among eleutherozoan echinoderms. Littlewood et al. (1997)

excluded all sites that could not be unambiguously aligned (thus avoiding the most rapidly-evolving sites), and discovered a weak signal for an echinoid–ophiuroid pairing. In contrast, Janies (2001) and Janies et al. (2011) analysed RNA data using POY (Wheeler et al., 1996; Varón et al., 2010), a technique that carries out tree building and sequence alignment simultaneously on the complete sequence and thus includes regions of highly ambiguous alignment. Initially Janies (2001) found strong support for an asteroid–ophiuroid pairing, and no signal for the echinoid–ophiuroid pairing. Later Janies et al. (2011) showed that class relationships could not be resolved because the outcome was very sensitive to tree search parameters being used. By including poorly aligned (i.e., faster evolving) regions, the direct optimisation approach implemented in POY is much more likely to find support for artifactual clades. Similarly we suspect that Pereske et al.'s (2010) analysis of mitochodrial genome architecture and amino acid sequences, which consistently recovered only one clade comprising asteroids, echinoids and holothurians, is most likely an artefact. Ophiuroids proved to be both very long-branched and highly divergent in genome architecture while crinoids had markedly different nucleotide compositions of protein coding genes.

Confirmation that echinozoans and ophiuroids form a clade will require analysis of many more slowly evolving genes. However, our preliminary results point to this being the only grouping subtended by an historical signal in our data. If correct this has important implications for the morphological evolution of echinoderms. First it confirms that that the morphologically similar pluteus larval stages of echinoids and ophiuroids are indeed homologous rather than convergent, as first suggested by Hyman (1955). It also supports the view that neurulation of the radial nerve in echinoids, holothurians and ophiuroids evolved just once, as argued by Heinzeller and Welsch (2001). Finally, the stellate body plan of asteroids and ophiuroids must be plesiomorphic with the globular echinozoan body plan derived from it. The outstanding problem now for palaeontology is to identify whether any of the early fossil asterozoans are potential stem group echinoids + holothurians.

Acknowledgements

DP is supported by a Science Foundation Ireland Research Frontier Programme grant (08/RFP/EOB1595). All molecular analyses were performed using the computing infrastructures provided by the Irish Center for High End Computing (ICHEC), and the National University of Ireland Maynooth High Performance Computing facility. RF is supported by an IRCSET (Irish Research Council for Science Engineering and Technology) PhD Studentship. KJP is supported by the National Science Foundation and a NASA National Astrobiology Institute grant. We should like to thank two anonymous referees for their helpful suggestions, and J. Robertson, V. Moy and E. Sperling for technical assistance.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.08.028.

References

Brinkmann, H., Philippe, H., 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16, 817–825.
Dean-Shackleton, J., 2005. Skeletal homologies, phylogeny and classification of the earliest Asterozoa. J. Syst. Palaeontol. 3, 29–114.
Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20, 248–254.
Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sørensen, M.V., Haddock, S.H.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.
Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., Raff, R.A., 1988. Molecular phylogeny of the animal kingdom. Science 239, 748–753.
Heimberg, A.M., Cowper-Sallari, R., Sémon, M., Donoghue, P.C.J., Peterson, K.J., 2010. MicroRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. Proc. Natl. Acad. Sci. USA 107 (45), 19379–19783. doi:10.1073/pnas.1010350107.
Heinzeller, T., Welsch, U., 2001. The echinoderm nervous system and its phylogenetic interpretation. In: Roth, G., Wullimann, M.F. (Eds.), Brain Evolution and Cognition. John Wiley and sons, New York, pp. 41–75.
Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Bagu, J., Bailly, X., Jondelius, U., Wiens, M., Mueller, W.E.G., Seaver, E., Wheeler, W.C., Martindale, M.Q., Giribet, G., Dunn, C.W., 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Roy. Soc. Lond. B 276, 4261–4270.
Holton, T.A., Pisani, D., 2010. Deep genomic-scale analyses of the Metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biol. Evol. 2, 310–324.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: bayesian inference of phylogeny. Bioinformatics 17, 754–755.
Hyman, L., 1955. The Invertebrates: Echinodermata, vol. 4. McGraw Hill, New York.
Janies, D., 2001. Phylogenetic relationships of extant echinoderm classes. Can. J. Zool. 79, 1232–1250.
Janies, D., Voight, J.R., Daly, M., 2011. Echinoderm phylogeny including *Xyloplax*, a progenetic asteroid. Syst. Biol. doi:10.1093/sysbio/syr044.
Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.
Lartillot, N., Philippe, H., 2004. Bayesian phylogenetic software based on mixture models. Mol. Biol. Evol. 21, 1095–1109.
Lartillot, N., Philippe, H., 2005. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.
Lartillot, N., Philippe, H., 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos. Trans. Roy. Soc. B 363, 1463–1472.
Letsch, H.O., Kjer, K.M., 2011. Potential pitfalls of modelling ribosomal RNA data in phylogenetic tree reconstruction: evidence from case studies in the Metazoa. BMC Evol. Biol. 11, 146.
Littlewood, D.T.J., Smith, A.B., Clough, K.A., Emson, R.H., 1997. The interrelationships of the echinoderm classes: morphological and molecular evidence. Biol. J. Linn. Soc. 61, 409–438.
Mallatt, J., Winchell, C.J., 2007. Ribosomal RNA genes and deuterostome phylogeny revisited. More cyclostomes, elasmobranchs, reptiles, and a brittle star. Mol. Phylogenet. Evol. 43, 1005–1022.
Mallatt, J., Craig, C.W., Yoder, M.J., 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. Mol. Phylogenet. Evol. 55, 1–17.
Mooi, R., David, B., 2000. What a new model of skeletal homologies tells us about asteroid evolution. Am. Zool. 40, 326–339.
Pereske, M., Bernhard, D., Fritzsch, G., Brümmer, F., Stadler, P.F., Schlegel, M., 2010. Mitochondrial genome evolution in Ophiuroidea, Echinoidea, and Holothuroidea: insights in phylogenetic relationships of Echinodermata. Mol. Phylogenet. Evol. 56, 201–211.
Peterson, K.J., Lyons, J.B., Nowak, K.S., Takacs, C.M., Wargo, M.J., McPeek, M.A., 2004. Estimating metazoan divergence times with a molecular clock. Proc. Natl. Acad. Sci. USA 101, 6536–6541.
Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alie, A., Morgenstern, B., Manuel, M., Wörheide, G., 2010. Improved phylogenomic taxon sampling noticably affects nonbilaterian relationships. Mol. Biol. Evol. 27, 1983–1987.
Pisani, D., 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53, 978–989.
Pisani, D., Wilkinson, M., 2002. MRP, taxonomic congruence and total evidence. Syst. Biol. 51, 151–155.
Rambaut, A., Drummond, A.J., 2007. Tracer v1.5.1 [Internet]. <http://beast.bio.ed.ac.uk/Tracer>.
Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.
Rota-Stabellli, O., Telford, M., 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48, 103–111.
Rota-Stabellli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M., 2010. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. Roy. Soc. Lond. B Biol. Sci. doi:10.1098/rspb.2010.0590.
Ruitz-Trillo, I., Riutort, M., Littlewood, D.T.J., Herniou, E.A., Bagu, J., 1999. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. Science 283, 1919–1923.
Smith, A.B., 1988. Fossil evidence for the relationships of extant echinoderm classes and their times of divergence. In: Paul, C.R.C., Smith, A.B. (Eds.), Echinoderm Phylogeny and Evolutionary Biology. Clarendon Press, Oxford, pp. 85–97.

Smith, A.B., Savill, J.J., 2001. *Bromidechinus*, a new Middle Ordovician Echinozoa (Echinodermata), and its bearing on the early history of echinoids. Trans. Roy. Soc. Edinburgh 92, 137–147.

Smith, A.B., Peterson, K.J., Wray, G., Littlewood, D.T.J., 2004. From bilateral symmetry to pentaradiality: the phylogeny of hemichordates and echinoderms. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling the Tree of Life. Oxford University Press, New York, pp. 365–383.

Sperling, E.A., Peterson, K.J., 2009. MicroRNAs and metazoan phylogeny: big trees from little genes. In: Telford, M.J., Littlewood, D.T.J. (Eds.), Animal Evolution – Genomes, Trees and Fossils. Oxford University Press, Oxford, pp. 157–170.

Sperling, E.A., Peterson, K.J., Pisani, D., 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol. Biol. Evol. 26, 2261–2274.

Sperling, E.A., Robinson, J.M., Pisani, D., Peterson, K.J., 2010. Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. Geobiology 8, 24–36.

Sperling, E.A., Pisani, D., Peterson, K.J., 2011. Molecular paleobiological insights into the origin of the Brachiopoda. Evol. Develop. 13, 290–303.

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

Swofford, D.L., 2002. Paup*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.

Wägele, J.W., 1999. Major sources of errors in phylogenetic systematics. Zool. Anz. 238, 329–337.

Wheeler, W.C., Gladstein, D.S., De Laet, J., 1996–2003. POY. Version 3.0 (current version 3.0.11). Documentation by Daniel Janies and Ward Wheeler.

Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., Peterson, K.J., 2009. The deep evolution of metazoan microRNAs. Evol. Develop. 11, 50–68.