University of Reading

School of Biological Sciences

PhD thesis

Evolution of Protein Interdependence
From Pairs to Networks

Warren J. Read

Supervisor

Professor Mark Pagel

20 February 2015

Declaration

I confirm that this is my own work and that the use of all material from other sources has been properly and fully acknowledged.

Warren J. Read
20 February 2015

Abstract

I present a method for inferring a protein interdependency network, based on correlated evolution between proteins on a large phylogeny of 72 diverse eukaryotic species. My original contribution is in the span of the phylogenetic tree used to generate the network: similar studies have concentrated on more localised regions in the tree of life and have been undertaken with more limited intent. I show that the whole-eukaryotic correlated evolution network is a real network and has interesting features of its own. The method can be broken down into three major, sequential parts: binary trait derivation, phylogenetic inference and likelihood analysis. I describe the implementation of a reciprocal BLAST protocol for the inference of a binary trait matrix corresponding to the presence or absence of orthologues in each species in the analysis, based on a reference human proteome. Rows in the matrix correspond to the reference proteins, columns to species: entries are 1, denoting presence, or 0, denoting absence. The matrix that I derive is mapped onto a phylogeny of the same set of species, to facilitate the detection of correlated evolution between proteins, or orthologous sets thereof, based on the pattern of gains and losses. The phylogeny is inferred from a set of genes, which are selected according to the criterion that they be present in all 72 species. 15 genes meeting this criterion are identified from the trait matrix; 14 of them are aligned and used for phylogenetic inference. The inference itself is performed using the program BayesPhylogenies, which implements a phylogenetic mixture model using a Markov Chain Monte Carlo (MCMC) method. A consensus phylogeny (tree) is calculated after the chain has been run for many millions of iterations; trees based on the genes individually were also inferred, for purposes of comparison. I use the program BayesTraits to perform a likelihood analysis on pairs of proteins from the trait matrix. This method detects correlated evolution by means of a likelihood ratio statistic, relating the likelihood of the two proteins having evolved independently, to the likelihood of their having evolved in an interdependent, or correlated, fashion. If the likelihood ratio statistic exceeds a certain threshold, this is interpreted as the signature of correlated evolution. Using presumptively interacting protein pairs from the Human Protein Reference Database, and a control (or null) set of pairs where no interaction is expected, I present evidence for the efficacy of the method in detecting correlated evolution. I proceed to infer a network based on correlated evolution, wherein each link represents an instance of pairwise correlation, and demonstrate that a power law gives a good fit to the distribution of nodal degree within the network, which is also the case for a network of presumptive protein interactions with no filter for correlated evolution. Finally, I infer a new equation to characterise the evolutionary rules which fashioned the network. I propose a method for testing the equation, and discuss future directions.

Acknowledgements

As everyone who has undertaken a thesis or any long-term project knows, the work and energy which go into such a task rest on the shoulders of many others who contribute in crucial ways to the overall end result.

Above all, I am incredibly thankful to Professor Mark Pagel for suggesting the topic of the thesis, for his advice and guidance throughout the duration of the project, and for securing funding from the BBSRC to support this research. His patience and mentorship have sustained me in my goal.

My sincere thanks are due to Dr Andrew Meade and Dr Chris Venditti for their help with the methodology used, the computational methods, the writing process, and for being stable pillars of encouragement in both a personal and academic sense. In particular, Dr Venditti's patient, detailed and knowledgeable explanations of unfamiliar concepts, coupled with his unfailing sense of humour, helped me through some difficult phases, while without Dr Meade's expertise and input on the computational side this work would never have reached the stage it has.

I am very grateful also to Dr Ali Johnston of the British Trust for Ornithology for her assistance with matters mathematical, and to Daniel Scott, who assisted me in the initial phases of the project.

Further thanks are due to two people whom I have not met in person, but who nonetheless helped me at critical phases in the project when I was unsure where else to turn. So my thanks to Dr Daniel Barker for his advice and meticulous attention to detail in the preparation of input data for the workflow, and to Professor Sandra Baldauf for her help in relation to understanding issues around rooting the eukaryotic tree.

The whole analysis would have never been possible without the use of the ThamesBlue IBM supercomputer which the ACET team afforded me.

Finally, my eternal gratitude goes to my parents, who have sustained me morally and financially over the course of this project, to an extent far exceeding what it was reasonable of me to ask.

I apologise in advance if I have missed anyone's name, and am very grateful to you all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The World, the Cell and the Network

Consider our world. Wherever we look, most of what we see, most of the time, is quite familiar. Even if we find ourselves in a strange town, the very fact that it is a town lends it familiarity, because it tends to operate by rules with which we are intimately, and intricately, acquainted. The town probably has a bus station; if we are lucky, perhaps a railway station too. Moreover, to find these familiar places, with familiar purposes, we follow familiar protocols for discovering their locations.

On a smaller scale, one pebble on the beach looks much like another pebble, one oak leaf much like another oak leaf. Our lifelong familiarity with these objects lends us the capacity to classify and make sense of our world, of course. We know how to behave on the road, or inside a shop; we see oak leaves emerge, grow large, turn orange and fall. And we know how pebbles behave when thrown, or skimmed across the water. Our exposure to several, or many, such objects gives us a framework of rules which we can extend to predict the behaviour of similar objects. We are seldom called upon to deal with complete novelty; more rarely still do we relish it.

How does all the world's teeming diversity, and even the inorganic things which are not obviously related to each other, come to be divisible into categories, and sets of categories, which are comprehensible to its denizens? And comprehensible not only to its human denizens, but to all creatures possessed of neural cells[1], able to sense and react to their various environments by means

[1] Of course, the neuron is a signature cell type specifically of metazoa. However, recent work discusses structural and molecular motifs common to both neurons and certain types of plant cell [Baluska, 2010].

of the physical expression of sets of rules, however simple ("stay upright", "avoid sunlight", "seek out sunlight", "flee from predators" and so on)? Even a tiny animal with a rudimentary nervous system, and negligible capacity for developing new rules based on its individual lived experience, expresses rules which are hard-wired into its genome, a kind of collective unconscious encoded in its genetic inheritance[2]. The nematode *Caenorhabditis elegans* provides the most comprehensively studied example of such an organism, having a fixed number of 302 neurons, each of which has relatively few synaptic connections. Interestingly, despite their invariance between individuals, the level of morphological diversity among neurons in a single individual is surprisingly high: *C. elegans* neurons may belong to one of fully 118 distinct classes [White et al., 1986]. I shall return to this point shortly.

1.1.1 Evolution and information

If knowledge is viewed as a set of rules dictating the appropriate responses to given stimuli (or environmental information)[3], it is clear that living organisms of every sort must possess this information-processing capacity, this knowledge, at some level, in order simply to survive and reproduce. Seen in this light, evolutionary adaptation itself is in a sense a form of learning, wherein the acquisition of knowledge is confined and nurtured not within the particular neural structures of the individual, but distributed among an aggregate of individuals—a species or population. This form of "learning" allows for redundancy; individuals are expendable, both because they have no necessary ability to reinvent their own rule systems based on information acquired during their own lifespan, and in that this very knowledge is replicated more or less faithfully in each genome, i.e. in each individual. This is not to say that organisms are necessarily born with a full set of rules ready to run "out of the box". All but a few multicellular organisms have differentiated cell types and typically exhibit marked localisation of function within different parts of the body [Carroll, 2001]. This necessitates a developmental process characterised by progressive cellular specialisation following rounds of mitosis, and determined by factors extraneous to the genome itself, which is (for the most part) typically identical between different somatic cell types [Medvedev, 1995]; this is Waddington's "epigenetic landscape" [Waddington, 1957] in action. As such specialisation of function becomes established, the organism's full functional potential is realised when

[2] If one subscribes to the minimalist interpretation [Gooch, 1972] of Jung's notion of the collective unconscious, this might literally be true.

[3] Epistemologicially, knowledge can be equated with truth, plus belief in that truth on the part of the holder of knowledge, sometimes with an additional stipulation that the belief be justified rather than simply coincident [Ichikawa and Steup, 2014]. Parallel characterisations of knowledge, for example arising from research into Artifical Intelligence (AI), emphasise knowledge representation and have been described in terms of stimulus, perception, internal model and response [Newell, 1994]; more modern perspectives from AI have questioned the centrality even of an internal model [Brooks, 1990].

it becomes capable of marshalling all the now diversified biological resources at its disposal. This is the richest expression of the set of behavioural rules encoded in its genes, but throughout the developmental process, the rules which the organism is able to follow must facilitate its survival during the respective phases of maturation. This requirement might be thought significantly to constrain the possibility space of what comes later, analogously to the unplanned growth of a modern European city around its mediaeval marketplace—quite unlike a new town, with its designated schools, shopping malls and leisure facilities, conceived and realised as a thing fully formed, abstracted from history.

And yet there are animals that undergo spectacular metamorphosis between life stages adapted to inhabit distinctly different ecological niches; moreover, these niches commonly allow adult and juvenile forms to live in the same geographical location, i.e. sympatrically[4], without competing with each other for resources [Herder and Freyhof, 2006]. In the familiar case of butterflies and moths, metamorphosis entails the loss of pseudopods and mouth parts, along with the gain of feeding tubes and wings, among other remarkable transformative events [Judy and Gilbert, 1969; Mercer, 1900]. How does one genome encode these radically different, but both highly sophisticated and independent-living phenotypes? Are battalions of genes retired to make way for fresh conscripts during the change, or does the old guard have to learn to campaign under new leadership? In less allusory terms, we might ask whether changes in genetic expression are widespread or confined to a few controlling genes? In fact, recent work indicates that expression levels of a high proportion of genes undergo substantial up- or down-regulation during fruit fly metamorphosis, but that a much smaller proportion of genes acts to trigger the cascade of metamorphic events at the biomolecular level [O'Keefe et al., 2012].

The process of organismal maturation is not peculiar to animals: plants too must undergo differentiation and specialisation, although their body plans, and their genomes, are radically different from those of animals. For instance, a mature plant often has the ability to regenerate from the smallest vegetative part [Grime, 2002], while hybridisation and polyploidy are commonplace spurs both to phenoptypic diversity and evolution in plants, through the founding of reproductively isolated communities—potentially, the early stages of speciation [Soltis and Soltis, 2009]. Plants, along with most other major multicellular groups, lack nervous systems, but not the capacity for environmental response. They certainly respond to sunlight, to gravity, to moisture, as well as to stimuli from other living organisms. Pollinating insects and mycorrhyzal fungi [Barker, Tagu and

[4] More or less roughly speaking, that is: a pond inhabited by dragonfly nymphs is clearly spatially distinct from the airborne realm of the adult organism, yet the two habitats bound each other directly; strictly, this would be an instance of parapatry, rather than sympatry, as between conspecific juvenile and adult forms. Other geographical distributions involving terrestial larvae and winged adults could be regarded as more properly sympatric, but even here, trophic partitioning, perhaps between different parts (e.g. flowers and leaves) of the same food plant, will tend to be reflected in the respective stages favouring one microhabitat over another.

Delp, 1998] are just two examples. Such organisms' own "rules"—the particular ways in which they respond to their environment, e.g. in the case of some pollinating insects, by following an airborne chemical gradient "upstream" [Raguso, 2008]—can co-evolve with those of plants. Looser symbioses, or their logical antitheses—those evolutionary arms races between predator and prey, or parasite and host—also demonstrate example systems comprising agents each of which lives by rules which co-evolve in concert with the rules followed by its reciprocal partner, or partners, in this great interplay [Dawkins and Krebs, 1979].

The rules governing development and maturation, of a multicelluar organism's parts, and of the changing rules of behaviour which these parts acting in concert shall unfurl and deploy at each developmental stage, we may regard as "rules for rules" or meta-rules, existing alongside and co-ordinating the genes encoding the rules of behaviour themslves. These meta-rules are themselves encoded within the genome; the developmental genetic cascade—comprising, for example, HOX genes [Lemons and McGinnis, 2006], effector genes [Artero et al., 2003] and micro-RNAs [Pasquinelli and Ruvkun, 2002]—forms a kind of directed network (or directed graph)[5], wherein the active nodes at each stage of development define organismal phenotype. Behavioural rules are an aspect of phenotype, the manifestation of a larger network of molecular interactors [Hartwell et al., 1999]. This larger network changes with the organism, essentially as different genes are brought into play as the organism matures [Xue et al., 2013]. The "phylotypic stage"[6] of metazoan development has been correlated with signature genetic expression profiles, and has prompted the conjecture that the divergence of animal body plans over evolutionary time is related to differential synchronisation of genetic expression corresponding with key developmental milestones [Levin et al., 2012].

1.1.2 Ontogenic rules and the cell

In the most parsimonious formulation, we might anticipate that the genes encoding behavioural rules are recycled—that is, that they are employed in ways that may themsleves develop as the

[5] Generally, we may regard any aggregation of things which interact, and the interactions or relationships between these things, as a network (the equivalent term "graph", while technicially correct, is a formalism which we might choose to disfavour if only because of its propensity to confuse laypersons). Conventionally, each thing in the network constitutes a "node" or "vertex"; each relationship between things is a "link" or "edge". I give preference but not exclusivity to the terms "network", "node" and "link" over their respective counterparts for most of this study. For a slightly fuller discussion, see Section 1.1.3. Note that while it may appear that the network is an abstraction which distills reality down to a type of cartoon, networks nonetheless provide a paradigmatic framework which is extraordinary both in the breadth of its applicability, and in the growing power of the analytical toolkit which it can bring to bear on some very diverse fields of scientific inquiry [Watts, 2004].

[6] The embryonic stage following neurulation, during which diverse organisms within a single phylum are supposed most closely to resemble one another morphologically [Slack, Holland and Graham, 1993].

organism develops. As in fact may the genes encoding meta-rules, because they are often connected to other genes in the cascade by both inputs and outputs (forming internal nodes within the graph representing the developmental cascade) [Ben-Tabou de Leon and Davidson, 2007; Schütte, Moignard and Göttgens, 2012]. Or it could be that at certain developmental stages, certain genes are switched on, and having a single and very specific function, are switched off again, forever to remain quiescent, when that stage comes to a close [Ezhova, 2007]. Interestingly, this conception leads us to consider the possibility of the existence of yet another type of network, one wherein a protein or gene might have several connections to other proteins or genes, but where a specific subset only—of this aggregate of all such pairwise connections—is active at any given time[7]. So the network may change not only over evolutionary time, but within one lifetime. Recent work suggests that the transition from pluripotency to differentiated cellular phenotypes can be mediated by the binding of various histones to the other components of chromatin, including other histones; as they transition from being quite promiscuous to very specific in their affinities during the course of commitment, this in turn sets the general pattern for future gene expression, and by extension phenotype, in the differentiated cell [Meshorer et al., 2006]. These two conceptions of deterministic mutability in networks of gene (and by extension protein) interactions—the evolutionary and the developmental—surely do not exist in isolation one from the other, and indeed, the field of evolutionary developmental biology (or “evo-devo”) is concerned with just such connections between evolution and development [Carroll, 2008].

In this connection, we might also ponder the rules of behaviour encoded within the genome of a unicellular organism. Compared with the corresponding rules within multicellular organisms, the rules for a unicell appear to be more limited in scope, constrained as they seemingly must be to encompass only its immediate phenotype and associated behaviours, which pass down directly from the parent cell upon mitosis[8]; we may then expect these rule-based behaviours to become manifest in the progeny almost immediately (or at least within a cell cycle), so there is on the face of it no requirement to encode an auxiliary set of rules for growth and development of the organism. This is

[7] Although the immediate discussion in the context of developmental biology pertains chiefly to genetic regulatory networks, exactly this kind of temporal plasticity, relating to the pattern of connections (i.e. adjacency relationships) involving particular nodes, has been recorded within physical protein-protein interaction networks. It is not so much that such a network taken as a whole has a distinctive temporal signature, but rather that individual, highly-connected proteins within it can often be assigned to one of several different categories depending on their particular patterns of temporal variability, as these relate to observed (or inferred) physical connections to other proteins. Han et al. [2004] divide highly-connected proteins into “party hubs”, which are capable of forming many concurrent connections, and “date hubs”, whose affinites change over time. The terms “party” and “date” allude, respectively, to the parallel and serial character of the human interactions typical of the analogous social situations.

[8] Unicellular reproduction, even in eukaryotes, is predominantly asexual [Tibayrenc et al., 1991].

until one considers those unicellular organisms exhibiting several distinct life stages[9], such as, for example, the *Plasmodium* parasite, which causes malaria in humans[10]. Its different stages reflect adaptive resposes to the different environments which the organism finds itself inhabiting during the course of its life. The transitions from one form to another are discrete in character; they do not typically stop at some half-way point and we may therefore allow that the forms themselves are truly categorical in nature[11]. This is equally true of the cell types found at the same time, and often in close proximity, within multicellular organisms, from early development (e.g. gastrulation in metazoa [Smith, 1989]) through to the mature phenotype. The distinct cell types within the human being are sufficiently defined in character that we may meaningfully enumerate them; a recent investigation reckoned their number at 411 [Vickaryous and Hall, 2006]. Interestingly, of this number, 145 were reckoned to be types of neuron, 55 of these in turn belonging to the central nervous system. On the face of it, this makes for a considerable contrast with a past estimate of human neuronal diversity, which identified only 5 distinct classes of cerebellar neuron [Eccles, Itō and Szentägothai, 1967]. However, because the cell types specific to the cerebellum itself were not explicitly enumerated in the later study, this comparison is drawn only tentatively. Nonetheless, even if any reckoning of the ratio of neuronal diversity in *C. elegans* to that found in Homo sapiens is liable to overstatement[12], there is a very strong case for believing that the source of the greater complexity of the human organism as compared with the nematode lies not in the diversity of its cell types but in the way in which vast numbers of cells of similar type can in concert generate complex emergent behaviours. This is arguably lent support by the finding that the appearance of a single novel neuronal cell type in the lineage of higher primates may be associated (causally or otherwise) with other architectural changes in the brain, and most particularly with increases in both brain volume and cognitive facility in these species—which, of course, include *Homo sapiens* [Nimchinsky et al., 1999]. Such complex behaviours may be quite inaccessible to an organism possessing a fixed complement of 302 neurons *in total*, which are nonetheless sufficient to facilitate its survival and reproduction. The hypothesis that human complexity is an emergent quality of myriad yet largely homogenous basic components, if true, has important implications [Carroll, 2001].

So even within an organism, the categorical footprint of nature itself is crisp and unmistakable. But what is the reason for the discretisation of cell type? The absence of a continuum of cellular

[9] Interestingly, temporal polymorphism in eukaryotic unicells has been linked at the genomic level to cell type polymorphism in metazoa, which if true has profound implications for the evolutionary emergence of multicellular forms [Mikhailov et al., 2009].

[10] Concession should be made that *Plasmodium* is in some senses arguably a poor example here, being a rather atypical unicell not least in its predominantly sexual mode of reproduction [Tibayrenc et al., 1991; Kooij and Matuschewski, 2007].

[11] The discretisation of cellular phenotype within distinct *Plasmodium* life stages is well supported by biomolecular evidence [Gunderson et al., 1987; Florens et al., 2002].

[12] This is perhaps understandable given the very precise and thorough morphological studies of *C. elegans* published at a time when data of equivalent quality were not available for *any* other organism.

characteristics is the more surprising if one considers that most somatic cells of the metazoan body have at their nucleus precisely the same, full, genetic complement. This may in a sense be a manifestation of the unfolding of yet another set of rules, which are typically run to completion, or to some sort of provisional end state representing an attractor in possibility space—a point corresponding to a certain relatively stable configuration of gene expression and metabolism which can be sustained over time[13]. Much work in recent years has contributed to our understanding of the developmental cascade in terms of epigenetics, the framework beyond the genome, within whose intricate but not infinite interconnect the genome is drawn preferentially to certain nexi, each embodying a particular stable state of differential expression associated with a characteristic cell type [Goldberg et al., 2007]. These states may be real enough, but the thoeretical underpinnings of such stable states have been considered outside the context of the genome, even in the context of the origin of life itself, in the emergence of autocatalytic chemical networks [Kauffman, 1995].

More broadly, rules dictate category at every level. The pebbles on the beach look similar because similar sets of forces (rules) have operated on them over geological time to make them that way. They are very different from the stones found inland, although stones found in one inland area may be different from stones found in another, depending, for example, on whether what is now rolling countryside was in the deep past an ancient seafloor, or an eroded Hadean expression of volcanic forces.

These objects, as well as being similar, have relationships to each other. It is these, the relationships between like things, which ramify into the networks that give us a way of quantifying association. Smaller "things" may act in concert to form greater things with properties seemingly dissociated—or perhaps emergent—from those of their parts. Tiny grains of silica, themselves the product of forces on the Earth's surface acting to break apart the extended covalent structures forged deep below, now aggregate to form sand, which has particular and notable properties all its own—properties which differ depending on whether it is wet or dry. When wet, such properties are to do with the forces acting between the individual grains and water, and in this sense are deterministic and predictable. Even so, the emergent thing itself, the sand, has surprising empirical behaviours [Hornbaker et al., 1997]. Wolves [Marino et al., 2012], some cetaceans [Mann, 2000], and primates including humans [Marino, 2002], may act in concert within social groups to advance one other's interests, or the interest of the collective, at least when faced with challenges or threats from outside. Some insects[14] form highly complex groups of closely related individuals which act in ways the significance of which the tiny-brained actors cannot themselves be conscious of, but

[13] This is encapsulated in the notion of "homeostasis", which may be something of a misnomer in that what it is really describing is an ordered *dynamic* system [Lloyd, Aon and Cortassa, 2001].

[14] Chief among these are the Hymenoptera, comprising wasps, bees and ants. All three groups provide many instances of species which form "social", or super-organismal structures [Wilson, 1990*a*]—although solitary wasps and bees are also familiar [Linsley, 1958].

whose combined effect may be to generate both social and architectural structures of impressive complexity [Wilson, 1990*b*; Theraulaz et al., 1998].

Rules build upon rules. The tendency of a thing to follow one set of rules may cause such changes over time as to predispose it to observe new sets of rules: sand does not behave like so many pebbles; sandstone does not behave like sand. And a moth does not behave like a caterpillar. This layering of rules is observable at different spatial scales as well as over chronological time. The rule which attracts a bee to a flower is laid down in the bee's neuronal system, but has specific dependencies on the observance of finer-grained rules at the molecular scale, e.g. the formation of specific protein complexes at the synapses [Burne et al., 2011]. If one of the proteins involved in such a complex is mutated, it is possible that this dependency, this link, may ramify upwards in scale to impair the insect's ability to find its food source (and thereby to disperse the plant's pollen). It is also possible that the protein complex may form in a slightly modified way, or that different sensory cues are activated or made primary within the bee[15]. In this way—the conscription, activation or substitution of one link for another, if the latter becomes broken or weakened—robustness (arising from redundancy) at one scale in a network may help to preserve network integrity at another [Lehár et al., 2008].

These networks are not, of course, ordained by any designer, any authority, or in the case of biological networks, including the various types of biomolecular network, even necessarily any physical law[16]. At the organismal level, including not only all the molecular interaction and interconversion networks within the cell, but networks encompassing the nervous system, the endocrine system and other systems of more proximate intercellular communication[17], they are evolved phenotypes. At the level of communities or societies of organisms, this may still be true: a termite colony, including its intricate networks of cooperations between workers, and its labyrinthine passageways connecting living quarters, nurseries, etc., is arguably a phenotypic expression in the sense that the rules that built it are hard-wired into the termites at the level of the genome [Dawkins, 1999]. Yet in what manner does the structure of such networks embody their evolutionary or adaptive "purpose"? And what might be the precise role of selection in effecting changes to the network over evolutionary time?

[15] If our objective is actually to observe this rule-based cascade, note that although mutation can plausibly account for many behavioural modifications, arguably a more informative mechanism when we observe differential behaviour in bees of the same species, and especially in bees from the same colony, is the way in which environmental stimuli may influence underlying patterns of gene expression to generate behaviours which are adaptive in different habitats, e.g. where different feeding strategies might be optimal [Withers, Fahrbach and Robinson, 1993].

[16] All biomolecular networks are, of course, necessarily *constrained* by physical law, however.

[17] E.g. the paracrine, juxtacrine and autocrine signalling systems [Singh and Harris, 2005].

1.1.3 The network is more than a metaphor

In the sense that processes involving relationships between similar agents at all scales of both size and complexity (sand grains, bacterial cells, human beings, etc.) are ubiquitous, and that each agent usually has multiple such relationships with its fellows (other agents of the same general type), networks surely pervade our world. Yet, perhaps because it is easy to regard a network as a rather esoteric distillation of the bonds which exist within a real community[18], and one moreover which appears to discard much essential data, the true analytical potential of the network has begun to be understood and wielded in earnest within the general scientific community perhaps only in the last decade and a half.

A network, in formal mathematics known as a graph, can be defined as a set of nodes (vertices) plus the set of links (edges) which connect them. A node is an abstraction that can represent any one of a huge range of real-world agents; links may represent specific types of relationship among such (typically similar[19]) agents. An illustration of a simple network is provided in Fig. 1.1.

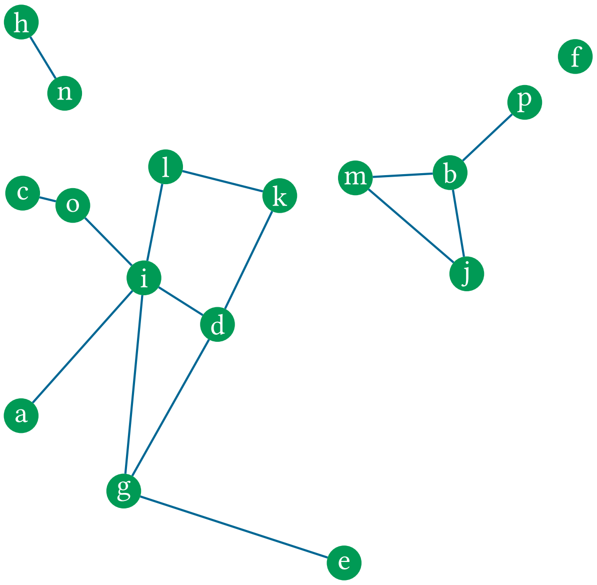


Figure 1.1: A simple network with a few nodes and links

Note certain features of the figure. There is one node (f) which has no links at all, yet formally it is still part of the network and we say that its "degree" is zero. There are visible "islands" of connected nodes which are completely disconnected from other such islands[20,21]. Because of the

[18] I use term "community" very loosely here, to refer not only to communities of organisms, but to aggregations of inanimate or even virtual objects (such as web pages).

[19] Although see Section 1.1.5.

[20] This is the very property that defines them as islands, of course.

[21] The more technical word for "island", in network terms, is "component"; I shall henceforth adopt this convention.

existence of separate components, with no way to reach one from another (e.g. in this network, there is no path from **a** to **b**), the network is said to be "disconnected". It would be "connected" only if there were at least one path between any two nodes in the network[22]. A series of "hops" over adjacent links[23] forms a "path" through the network. Any pair of nodes that are connected by a path are part of the same component. There are closed paths, or loops, within the network, such as the path connecting nodes **b**, **m** and **j**. There is one very well connected node (**i**) with links to five others: this is a "hub" (in this case a small hub of degree 5). Another point of note is that the network is *planar*: it is capable of being represented in two dimensions without any of its links crosssing over each other. While for a subset of real-world networks this might be true[24], in general it is not, and Fig. 1.1 is drawn as planar for clarity's sake only, in that the introduction of points of crossing might obsure its explanatory purpose. However, if the network were fully connected (retaining the same complement of nodes) it should be fairly clear that there is no way that it could then be drawn without links crossing one another. For the purposes of this study, we assume that molecular networks, including those derived by the application of novel inferential methods, are non-planar.

Further significant features of Fig. 1.1 include the observation that there are no nodes which are linked back to themselves (i.e. no "self-edges"); nor are there are any multiple links ("multi-edges") between any one pair of nodes. Furthermore, the links, like the nodes, have no visible weighting or other qualitative distinction between themselves and any other link—apart, that is, from their lengths. With regard to the lengths of links, note that we are forced to represent a set of relationships on a plane, and that the same set of relationships can be laid out in different ways, dependent upon our intentions, our aesthetic sensibilities, or both. We should therefore be careful in ascribing too much significance to the apparent length of a link in a network, possibly even if urged to do so. With this caveat in mind, certain features of a given network topology should remain apparent, regardless of its layout. Finally, there is no directionality in this network, as there are no arrows on the links; relationships denoted by the links run equally in both directions, which makes this an undirected network. In fact, this is an instance of a subclass of network of a very simple type[25] and I have here sacrificed complexity for the sake of clarity.

[22] In a *fully* connected graph, every node has a link to every other and every shortest path between any two nodes is of length 1.

[23] One can think of links as being joined together by nodes much as one thinks more conventionally of nodes being joined together by links.

[24] E.g. the road network is mostly planar, although bridges and tunnels provide occasional points at which links cross; such points are the exception rather the rule however.

[25] In fact, by virtue of its lack both of self-edges and multi-edges this is formally a simple graph, specifically a simple undirected graph.

1.1.4 Networks in the real world

Many different types of real-world network have come under scrutiny by scientists from a surprising range of disciplines. Friendship and other social networks have long been studied by sociologists, particularly since the publication of pioneering studies in the 1960s [Milgram, 1967] and 1970s [Granovetter, 1973]. Here the nodes and links are of the most familiar kind: people and their relationships to one another, respectively. Normally, people's relationships (friendships, acquaintanceships, etc.) are difficult to quantify on a very large scale. Yet statistics for certain kinds of relationship between people are more readily available: arguably, the two most famous examples are scientific citation networks [Leicht et al., 2007] and the network of film actors who have appeared in the same film [Collins and Chow, 1998]. In the former, the links are citations (strictly the nodes are scientific papers rather than people), while in the latter, the links are films[26].

Computerised networks have also been widely favoured as a subject of study, partly because of the ready availability of data. These networks may be actual computer networks, representing cabled links between routers and DHCP[27] clients, as well as directly between routers themselves. Or they may be less tangible, as with the network of links to and from pages on the World Wide Web [Dorogovtsev et al., 2002].

Biomolecular networks too have formed primary objects of scrutiny in recent years, with protein-protein interaction networks [Pellegrini et al., 2004], genetic regulatory networks [Guelzim et al., 2002] and signal transduction networks [McCormick, 2003] all providing a focus for examination as data from sequencing projects, genetic expression arrays and other large-scale assays[28] has become more readily available. These various sorts of network, while sharing some of the same participants, are each slightly different in kind. Protein-protein interactions are straightforwardly represented with proteins forming the nodes, and physical interactions between them the links. Contrastingly, in a metabolic network of the kind familiar from biochemistry textbooks, metabolites are typically represented as nodes while enzymatically-catalysed interconversions among them (if not the enzymes themselves) are usually shown as links [Jeong et al., 2000].

Some networks are more tangible, with both nodes and links having a (potentially observable) physical incarnation. The neural network within *C. elegans*, and its vastly more complex equivalent within us, are examples. Conventionally, the neurons (of however many kinds) form the nodes, while links are provided by the dendrites (which carry the electrical inputs from other neurons) and the axon, which conveys the action potential from the neuron when it fires, and branches into various terminals, all of which enable the signal to propagate to the synapses, on the other side of

[26] But again, see the discussion on bipartite networks in Section 1.1.5.

[27] Dynamic Host Configuration Protocol [Droms, 1997, RFC2131]

[28] E.g. the yeast two-hybrid assay, designed to detect protein-protein interactions [Fields and Song, 1989].

which lie the terminals of other neurons' dendrites [Braitenberg and Schüz, 1998]. Note that the simple (or simplistic) representation of a neural network as a set of interconnected neurons with undifferentiated links rather disregards the specific roles of dendrites, axons and synapses, which are subsumed into the concept of the link, but none of which in real neural networks[29] are any less a physical reality than the neurons themselves.

Another tangible network is our circulatory system, the system of veins, arteries and capillaries which distributes oxygen through our bodies before cycling deoxygenated blood back to the lungs [Tortora and Derrickson, 2011]. Analogously, the roots and branches of a tree, or the mycelial system of a fungus, may form physically realised network structures, wherein the nodes are simply the bifurcations where tubelike somatic tissue forks[30].

1.1.5 Unifying themes and categorical divisions

Actually, the examples of real-world networks given above are fairly diverse not only in their essential natures but also in their topologies; drawing contrasts between them may help to illuminate a number of the nuances that are sometimes alluded to when talking about and analysing networks. These nuances transcend the simple, but archetypal, representation of a network illustrated above, which is in some regards a special case, but is nonetheless often invoked to illustrate more general principles.

Consider for example the network of film actors mentioned above. Plainly, two actors might well have co-starred in multiple films. Do we then draw a link for each film in which they have appeared together? This is one possbility. The network of actors thus formed, wherein the same pair of actors might have an arbitrary (but integral) number of links joining them directly together (i.e. without any intervening nodes) is an example of a "multigraph". What about the strengths of the links? Maybe we could apply a weighting to each link, denoting how much on-screen time the two had shared together? Now we have a weighted graph: the strength of each link could be represented by its thickness. There is of course nothing to stop us drawing a weighted multigraph, the thickness of each link representing time shared on screen in a particular film. The two quantities (weightedness and multi-ness) differ in that the former is continuous, the latter discrete; they may be distinctly portrayed within the same network as multiple links of varying thickness between certain pairs of

[29] Because the "neural network" is now a paradigm in computer science, the term has become semi-detached from real brain tissue, in that often, it refers to an implementation in software or (more rarely, electronic hardware) of principles abstracted from biology. Perhaps it would be helpful to apply a separate term to real biological networks of neurons, e.g. "neuronal network", but here I simply allow that the intended meaning of each use of the accepted term is given clarity by context.

[30] This makes for a particular kind of directed network wherein each node has one incoming and two outgoing links.

nodes.

Continuing with the film analogy, what if an actor plays more than one part in a film, and perhaps even features as two different characters within the same scene, such as Eddie Murphy in *"Coming to America"*? We might draw a link in the network, from the node representing Eddie Murphy, back to himself—a self-link.

Considering again the notion of the multigraph among film actors, there is another, arguably more expressive, and certainly more information-rich, way of representing the same information: as a bipartite network, i.e. a network comprising two different types of node, and just one type of link (which can exist only between unlike nodes). In the case of the co-star network, one type of node represents actors, the other films; all links occur only between actors and films, never between two films, or between two actors. So a co-star relationship is represented by two separate links terminating at the same film. In this case there is no need for (nor any meaning in) drawing more than one link between any two nodes, since each link fully expresses the involvement of one actor in one film. Moreoever, we know which film. This is more informative than the equivalent multigraph representaton with a single type of node: consider three actors who once starred in the same film together, but whose paths have never crossed apart from that. The multigraph represents this as three links: one link between each pair of actors. Yet we cannot tell whether the triangular closed loop thus formed represents one film, or three[31]. The bipartite graph neatly encapsulates this information, and expands on it, by drawing a link between each actor and the same film: the relationship is unambiguous.

Of what relevance are these considerations to the present study? Actually, they are tenably of some importance. The co-star network shares with protein-protein interaction networks the feature that it is undirected, i.e. relationships are reciprocal and cast each partner in an equal role. This cannot be said of the network of links on the World Wide Web, or of genetic regulatory networks, wherein the relationships are decidedly non-commutative. Perhaps more interesting still is the idea that we might attempt to represent protein-protein interactions in the form of a bipartite network. Traditionally, this has not been done. And yet the mathematical and software toolkits at our disposal for the purpose of anlaysing network structures is richer now than ever before. With a protein-protein interaction network, it is processes involving physical contact between macromolecules that are being characterised. Proteins come together to form complexes with a certain effect. Participation in the same protein complex is typically depicted as a link between two proteins, even though such a complex may have multiple participants, rather than just two. That being the case, it seems plain that to represent the protein-protein interaction network as a set of pariwise interactions

[31] Even though we *do* know that it cannot be two, because at least one of the actors would necessarily have had a part in both films, implying that the remaining two actors had never starred in a film together at all—thus breaking the third link.

is actually to throw away important information[32]. If we could characterise the interaction clusters themselves and view them too as nodes, it would be possible to derive our bipartite form of the protein-protein interaction network. In combination with the proposed phylogenetic protocol for the inference of protein-protein relationships (admittedly not necesssarily confined to physical protein-protein *interactions*), this holds out a tantalising prospect of brand new science.

1.1.6 Towards formal network knowledge representation

Relationships between things are often represented within ontologies, as subject-verb-object triples; in our case protein A (subject) interacts physically (verb) with protein B (object). In fact the converse, commutative relation also applies. What we are concerned with here, however, is the fact that both protein A and protein B participate physically in protein cluster X, which may include yet more participants—proteins C, D and E, for instance. Clearly proteins A, B, C, D and E are related to each other physically by virtue of their participation in the cluster, but to represent this as a set of binary relationships in the form suggested above for proteins A and B would require $\frac{5!}{2!(5-2)!} = 10$ pairings, even if we enumerate only one of the two possible directionalities (orderings) of each pair[33]. In fact, in ontological terminology, the relationship between these five proteins, at its simplest, is *n*-ary rather than constituting a set of binary relations. Here we have $n = 5$, which we should be able to represent as a set of five statements (or relations), as long as we relate our five existing entities (proteins) not directly to each other, but to the entity (protein cluster X) of which they are all part. So we can add a new class of thing (protein cluster) to our ontology, and equivalently to our network—making it bipartite, but still undirected. In doing so, we save ourselves the inconvenience, consfusion and associated information *loss* of recording $10 - 5 = 5$ redundant relations in our ontology, or drawing the same number of extra lines on our network diagram. The number of redundant binary links between like entities increases roughly in proportion with the square of cluster size, as compared with using the *n*-ary relation, or bipartite graph. In fact, the number *r* of redundant, or "wasted" links in the former over the latter, in a protein cluster of size *s* (i.e. containing *s* proteins) is given by

[32] Admittedly, this assumes that we have the information to throw away in the first place: because of the pairwise nature of the screens often used to infer protein-protein interactions, this is not necessarily the case.

[33] In fact this is correct: because the protein-protein interaction network is undirected, we would not wish to enumerate ordered pairs (2-permutations) from our total of five elements (i.e. proteins in cluster X), but rather, the number of possible 2-combinations, or *un*ordered pairs.

$$\begin{aligned} r &= \frac{s!}{2!\,(s-2)!} - s \\ &= \frac{s(s-1)}{2} - s \\ &= \frac{s^2 - 3s}{2} \end{aligned} \tag{1.1}$$

Researchers are now well placed to start using, as well as enriching, ontological data concerning protein-protein interactions, as well as other intermolecular relationships within the cell, thanks in no small part to standards-based approaches to biological data curation and exchange like the HUPO PSI[34] Molecular Interaction (PSI-MI) format [Hermjakob et al., 2004]; several public repositories of protein-protein interactions now supply data in PSI-MI format, which includes a facility to represent and curate protein clusters distinctly from their constituent proteins. A broader list of ontologies, built upon and incorporating "controlled vocabularies" for the definition of allowable entities and relations between them, relating specifically to the represention of biological data, is available at the OBO[35] Foundry web site: OBO maintains a subset of these ontologies directly, in its own OBO format, but others are available in OWL[36] format. OBO has a primary goal of facilitating the exchange and interoperability of biological data, including specifically biomolecular data.

With the spread, recognition and adoption of these standards-based ways of encoding biomolecular knowledge, the opportunities for novel discovery through the integration and analysis of data from disparate sources—mRNA array data, DNA sequence data, analytical proteomics from mass spectrometry, metabolic networks, two-hybrid assays for detecting protein interactions, inferred genetic networks (increasingly accounting for the role of non-coding DNA control sequences), and, not least in the context of this study, relationships among genes and proteins inferred from various computational studies including evolutionary analyses—are as never before. Controlled vocabularies allow the matching, comparison and extension of raw and pre-filtered data acquired from a spectrum of disciplines within the biological sciences. Taking advantage of ontology-based knowledge representation, derived from the ever more precise characterisation and formalisation of biomolecular relationships, may ultimately allow us to couple networks of knowledge from these different realms in new ways, and open wholly new and unimagined vistas of discovery, which can be conducted from the desktop.

[34] Human Proteome Organizaton Proteomics Standards Initiative

[35] The Open Biological and Biomedical Ontologies [Smith et al., 2007]

[36] Web Ontology Language

1.2 Statistical features of undirected networks

Some types of biomolecular network are directed. Metabolic networks are a case in point, meaning that the enzymatic conversions of substrates to products are usually one-way only, which is why they are often drawn as arrows. This is one instance where it is meaningful to ascribe a direction to relationships within a network, even though many networks, including not only protein-protein interaction networks but other molecular networks within the cell, are better characterised as undirected. As far as our preferred objects of study are concerned, if protein A interacts with protein B, then protein B interacts with protein A; the relationship is straightforwardly commutative. A node may be characterised in part by its degree (or "valency"), which is simply the number of links attached to it at one end[37]. Degree is typically denoted k. Because an undirected link has two ends, each representing a connection possessed by a node, the sum of all values of k for every node will come to exactly twice the number of links in an undirected network.

In considering various ways in which we might characterise the features of a network as a whole, one of the primary statistics of interest is the node degree distribution, which can equivalently be thought of as the way in which the network's links are distributed among its nodes. We speak of a network's "degree distribution" rather than its "connectedness distribution" (much less its "connectivity distribution", which actually has a distinct formal meaning[38]). We might ask ourselves whether an observed degree distribution provides evidence for a network's links having been laid down in some random process, or whether the acquisition of links appears to be weighted or directed in some manner which merits further investigation. A note of caution should be sounded in connection with the word "random", however. In employing it, we mean specifically that links have been laid down between nodes in a stochastic or non-deterministic fashion. We may stipulate an additional condition, that the probability of an end attaching to a node is equal for all nodes each time a new link (pair of ends) is laid down, but this condition is not a necessary corollary of randomness. This is an important distinction to which we shall return later. For the time being, we note that in a random network which has been shaped by a probabilistically uniform attachment process (one which does not discriminate among nodes based on their existing properties, either innate or emergent), values of k follow a Poisson distribution, wherein the probability of a node being connected to k other nodes (in principle possibly including itself) is described by

$$p\left(k\right) = \frac{\bar{k}^2 \mathrm{e}^{-\bar{k}}}{k!} \tag{1.2}$$

[37] This is precisely equivalent to the number of nodes to which the node in question is "adjacent".

[38] Conventionally, by connectivity, we mean not the number of links terminating—or originating, which is effectively the same thing for an undirected network—at a node, but the number of independent paths between two given nodes. So connectivity differs fundamentally from degree, not least in being a function of two nodes rather than a single node.

where $\bar{k}$ is simply the mean value of k. Note that $k = 2l/n$ where l is the number of links and n the number of nodes. The distribution is peaked where $(k{-}1) \leq \bar{k} \leq k$. If k is integral, there are joint maxima at $k{-}1$ and k. In its tail, where $k \gg \bar{k}$, the Poisson distribution approximates an exponential distribution, i.e. one described by the equation

$$p\left(k\right) = \mathrm{e}^{-k} \tag{1.3}$$

Until the late 1990s, the often unspoken but prevalent assumption was that most naturally occurring networks, including the molecular networks which interest us, were fairly well characterised by Poisson-distributed degree. This assumption was shared even by investigators whose principal focus consisted in the theoretical aspects of molecular networks within the cell, an area where network topology, reflected in degree distribution, might be thought to have profound implications for any conclusions reached. Kauffman [1995] proposed that the origins of life itself could be traced to the formation of a "giant component" of reactions within an increasingly complex chemical soup. The giant component, an autocatalytic chemical set, emerges as a distinct phase transition within the soup as more reactions (links) are added to the substrates (nodes) in a random process wherein there are no probabilistically preferred nodes.

It is interesting therefore that in the space of two short years at the very end of the twentieth century, the predominance of the random graph as the expected form taken by real networks, natural and man-made alike, was first subverted and then effectively toppled. Although several clues had already been left lying on the trail for decades, because these clues came from the social sciences rather than from physics or biology, the key insights into real-world network topology did not follow until 1999. It was in this year, following a revival in interest in the "small world" phenomenon [Watts and Strogatz, 1998], which describes a social network wherein each individual is separated from any other by no more than six links (hence "six degrees of separation"), that Barabási and Albert [1999] published work showing that the true topology of several real world networks was better characterised as being "scale-free". Networks as varied and unrelated as the network of pages on the World Wide Web—the "nodes" being the pages themselves and the "links" being directed hyperlinks—and the metabolic networks within micro-organisms, shared a trait in common: instead of k following the assumed Poisson distribution, the observed degree distributions roughly followed a power law, where $p\left(k\right) = ak^{-\lambda}$, λ being dubbed the "degree exponent". The "fat tail"[39] of a power-law distribution allows for the existence of numerous nodes with very high k as compared with the mean ($\bar{k}$); modal degree, by contrast, is less than the mean. The

[39] An exponential distribution or a Poisson distribution will see the value of $f\left(k\right)$ (frequency, or, in the context of a network, number of nodes) approach zero asymptotically as k increases; in a power-law distribution, the decline is very much shallower. So $f\left(k\right)$ represents the probability of k (degree) taking a certain value.

term “scale-free” is used to indicate that one can expect to find nodes within the network of every level of connectedness (degree), no matter how large, given a big enough sample of nodes; this is characteristic of power-law distributions.

The findings of Barabási and Albert are counter-intuitive in the sense that the kinds of distribution with which people are typically better acquainted, such as the frequency of male adult individuals falling within certain height ranges, separated by (say) an inch, are approximately normal, and have the familiar bell-curve shape on normal axes. There are no men who are ten feet tall, let alone 100 feet tall, and yet this is what we would see if human height followed a power-law distribution. It should be noted that in a random graph, given large $\bar{k}$ and larger n, the degree distribution approaches a normal distribution; by contrast, a power-law distribution is skewed to the right, i.e. it has a fat tail to the right. In many real networks, such as metabolic networks, $\bar{k}$ can be quite small (<3), but plotting frequency against k on log-normal axes (y being logarithmic, x normal) should reveal that the distribution in the tail approaches a straight line where $k \gg \bar{k}$, if the degree distribution is random. For metabolic networks, for social networks of friends, and for networks of web links, the degree distribution does not follow this pattern. Rather, it is approximately linear when plotted on log-log axes, i.e. when both x and y axes are logarithmic. This is the key signature of a power-law distribution; the slope of the line fitted to such a plot gives the degree exponent λ, according to

$$\ln\left(p\left(k\right)\right) = \ln\left(a\right) - \lambda \ln\left(k\right) \tag{1.4}$$

where ln is the natural logarithm, or $\log_e$. This is a straightforward reformulation of $y = \mathrm{m}x + \mathrm{c}$.

Real-world networks with power-law degree distributions typically have several other interesting features. In common with random graphs, they tend to have very short mean path lengths [Barabási and Oltvai, 2004], path length being the minimum number of intermediate nodes lying on any path between a pair of nodes, out of all possible paths. However, unlike random graphs, many real world networks exhibit a high degree of clustering [Ravasz et al., 2002], meaning that if nodes A and B are directly connected and nodes B and C are directly connected, it is highly probable that nodes A and C are directly connected too. The archetypal networks in which this is observed are social networks of friends, but clustering has been found to exist in less obvious networks too, including molecular networks. One way of quantifying the extent of clustering within a network is to use the clustering coefficient C proposed by Watts and Strogatz [Watts and Strogatz, 1998], which is defined for each node; for a node with k links, it is given by

$$C(k) = \frac{2N(k)}{k(k-1)} \tag{1.5}$$

where $N(k)$ is the number of direct links between the k direct neighbours of the node. Clustering has been suggested, for example, as the mechanism by which cellular functions are modularised, or isolated from one other, enabling the cell to perform many tasks in parallel without too much mutual interference. Thus the network can be viewed as being constructed from building blocks which are themselves mini-networks built from still smaller network modules. It has even been argued that clusters within networks can exhibit self-similarity at different scales, in a manner analogous to fractals [Dorogovtsev et al., 2002]. In an evolutionary context, this kind of modularisation of roles may allow selection, both purifying[40] and positive[41], to operate on discrete traits which are associated with such modules. This conjecture is interesting in connection with our current inquiry in the sense that the kind of correlated evolutionary linkage between proteins which we seek to elucidate may provide supporting evidence for modularity in real networks of interacting proteins, or of proteins which are interdependent in other ways. We will return to this later.

It is worth pausing to note that one consequence of clustering or modularity that we might expect to observe is a more extended mean path length for the network as a whole; indeed this is the case for highly clustered networks whose degree distributions do not follow a power law. The reason why scale-free networks can simultaneously exhibit both a short mean path length and a relatively high level of clustering is the presence of highly-connected nodes, or "hubs", which, like the larger stations in a rail transport network, together form a high-speed backbone which is easily accessible to nearby, less well connected nodes, analogously to small stations on a branch line. This observation neatly accounts for the "small world" phenomenon mentioned earlier, and does so in a way which requires no special trick or clever wiring in the network to ensure that the journey between any two nodes contains no more than a few hops.

1.3 Network emergence as evolutionary process

We have noted that many real-world networks, including such protein-protein interaction networks as have been investigated, are characterised by degree distributions corresponding roughly to a power law. This being the case, they could not have arisen from a process of random at-

[40] Acting to eliminate miscopies resulting in deleterious mutations. Most mutations are deleterious, so genes coming under purifying selection tend to be conserved.

[41] Being pulled by selection along a certain evolutionary trajectory, many mutations at the locus concerned having a positive or neutral effect on organismal fitness.

tachment in which the probability of attachment is invariant between nodes. Various alternative mechanisms have been proposed to explain these networks' scale-free topologies. Of these, the most frequently cited is "preferential attachment" [Newman, 2001], whereby nodes possessing more links are more likely to acquire new links than are nodes possessing relatively few. More precisely, in this scheme, the probability of a fresh attachment being made to a particular node is proportional to its current degree, plus some constant term. Because of the dependence of current probabilities (the instantaneous probability of each node forming a new attachment) on prior events, this is a dynamic process in which accretions of new connections form around already well connected nodes in a manner loosely analogous to the way in which localised concentrations of matter in the early solar system are thought to have seeded the planets [Lin and Papaloizou, 1985].

Preferential attachment provides a plausible and intuitively pleasing explanation for observed power-law scaling in social networks (we refer here to traditional social networks, and specifically exclude online social networks such as Facebook, for no reason other than that they have not been considered in this study) because gregarious people have more friends of friends to make friends with themselves [Jin et al., 2001]. It has similar power in explaining the growth of some artificial networks such as the network of web hyperlinks and the scientific citation network [Jeong et al., 2003], where highly referenced pages or papers will attract further viewers in proportion to the number of existing references; the potential for explosive growth in the exposure of such highly linked or cited content is not of course itself necessarily correlated with the intrinsic merit thereof.

Despite its obvious attraction in certain realms, however, preferential attachment remains controversial as a candidate causal agent for the emergence of scale-free topology within the cell's molecular networks. Consider the protein interaction network. By what process may the acquisition of a link (an interaction with another protein) increase the probability that the protein will acquire yet more such links in the future? Although such a mechanism has many supporters, it is not clear that it exists. It has additionally been suggested that selection may operate to promote and preserve scale-free topology in the network as a whole, because scale-free networks are generally very robust to perturbation, i.e. there is a very low probability that random loss of any node (protein) within such a network will prove highly deleterious, let alone lethal to the organism. This is because modal connectivity, which by inference we presume to correlate with phenotypic effect, is so low (when compared with the mean). This may tie in neatly with Fisher's notion that most mutations will tend to have low phenotypic effect, because the kind of large instantaneous jump in fitness which would be associated with a change to a gene of high phenotypic effect is a poor mechanism for allowing an organism to evolve towards a peak on a unimodal fitness landscape; in this view tiny, incremental changes are required for efficient selection.

Plainly, losses of proteins will occur at a certain rate through evolutionary time; the same is true for gains of proteins. Taken together, protein gains and losses yield a net rate of protein turnover which has been estimated in the context of larger investigations [Wagner, 2003] into the reasons for power-law scaling in protein-protein interaction networks. If preferential attachment does operate to sculpt the protein interaction network, we may go on to ask whether selection favours the conservation of highly-connected nodes, whose loss could be potentially a lot more deleterious than losses of nodes with low or modal connectivity. This is the conclusion of one early study [Albert et al., 2000]. In contrast, Wagner suggests that the evidence for such selectional bias is poor; rather, he argues that preferential attachment plus node addition[42] constitute a set of "local rules" which are sufficient in themselves to maintain scale-free degree distribution within the protein-protein interaction network, without the necessity to invoke selection.

Wagner also examines the role of gene duplication in fashioning the observed network topology. Immediately subsequent to any duplication event, the reciprocal partner proteins in each interaction with the original protein will acquire an extra interaction with the new paralogue[43]. So such events always increase the average connectivity within the network. The question here is whether the interactions with both the original protein and the paralogue are likely to be retained, or whether redundancy favours the retention of one interaction only. Wagner argues that redundant links tend to be dispensed with relatively quickly; this is intuitively pleasing, as purifying selection seems unlikely to act to retain redundant function. Simulations of gene duplication and subsequent erosion of connectivity (which we might reasonably expect to be much higher for recent duplicates than the mean rate for the network as a whole) indicate that gene duplication by itself has little effect on connectivity distribution over time, although the historical role of large-scale gene duplications in landmark bifurcations in the eukaryotic tree is now beyond reasonable doubt [Christoffels et al., 2004].

Another possible cause of power-law scaling could be at work. The absence of any mechanism which is both intuitive and widely accepted for the operation of preferential attachment in both intra- and extracellular molecular networks, and protein-protein interaction networks in particular, may give us cause to consider the nature of such an alternative. Pagel, Meade and Scott [2007] have proposed that the rate of attachment (i.e. the rate at which new links are formed) in the protein-protein interaction network does not vary in proportion to the number of existing links, which would suggest an historically accelerating rate of attachment in those proteins which are now highly connected, but rather that each different protein possesses its own instantaneous rate of attachment, which is invariant with time. This rate, different for each protein, is dubbed "stickiness", and is a function of the protein's intrinsic properties, which do not change simply by virtue

[42] Note that "addition" here refers specifically to the acquisition of its first network link by a protein, regardless of the antiquity or otherwise of the protein itself. So by "addition" we mean "addition to the network".

[43] A homologous (duplicated) gene within the same genome.

of gaining an interaction with another protein. If a gamma-distributed coefficient of stickiness λ is defined such that

$$p\left(k\right) \propto \mathrm{e}^{-\lambda k} \tag{1.6}$$

it is possible by integration to find the two parameters to the gamma distribution, α and β, which provide the optimal fit to the observed connectivity distribution within the protein-protein interaction network (or other molecular network). The goodness of fit can then be compared with that of a simple power law. One aim of the current study will be to refine this investigation in the light of new data and theoretical considerations.

1.4 A phylogenetic approach to network inference

My interest here is specific. I concern myself neither with stellar accretion discs, nor crystal lattices, nor the social whirl nor yet any of the artefactual patterns of linkage spawned by our technology—all of which could in some sense be characterised as networks—but rather with the interior of the biological cells which constitute the building blocks of Earthly life. Each such cell can be characterised on one level by patterns of molecular interactions whose collective flux can usefully be understood in terms of networks. These networks are of several conceptually distinct but interrelated types, including: the familiar and relatively well-characterised metabolic networks of enzyme-catalysed interconversions between small molecules; the genetic regulatory network [Guelzim et al., 2002], comprising molecules both simple and complex which variously enhance, inhibit and stabilise the manufacture of proteins within the cell; the network of interactions among those proteins once formed; and the network of interactions triggered by molecules transmitting signals between cells, most particularly the hormonal and signal transduction systems of multicellular organisms. These, and more subtle networks still, are being elucidated in ever more detail by studies of patterns of correlated gene expression using RNA microarrays, among several other advanced techniques falling within what we might call the paradigm of direct molecular analysis. I point this out not by way of offering a methodological critique, but to draw a contrast between this paradigm and an alternative approach which characterises other elucidative methods, specifically those falling within the realm of computational inferential statistics including phylogenetic methods; I argue, with supporting evidence, that these techniques too offer a valid means of investigating molecular networks within the cell. Such a perspective is facilitated in large measure by the exponential growth in computing power over the past four decades and should be seen as being complementary, rather than antagonistic, to "wet-lab" inquiry.

In this study, I employ such computational techniques, combining traditional statistical methods with more modern Bayesian approaches to phylogenetic inference, in order that we might investigate pairwise linkages between proteins and the characteristics of the network which we discern when these linkages are aggregated. For this I use a phylogeny of eukaryotes and a matrix of protein presence/absence data to derive a protein network based on pairwise correlated evolution between proteins, on the assumption that the latter indicates some correlation in function; a direct physical interaction between two proteins would be the simplest instance of this. I lay to one side any attempt at direct investigation into metabolic networks, which are comparatively well characterised, but make few assumptions about the exact nature of the inter-protein linkages which do emerge. Although part of my concern is to compare our network with others that are already partially characterised, for which very reason we focus much attention on protein-protein interaction networks (i.e. those wherein the "linkage" between two proteins is physical rather than metaphorical), it is nonetheless possible that a sizeable component of the network that we derive using our computational method represents not these physical interactions but an interdependency based on shared participation within the same molecular pathway, be it related to signal transduction, genetic regulation, or indeed metabolism.

1.5 Synopsis

The body of research which this project records is presented in four discrete parts, each a smaller project in its own right (plus some more speculative, unfinished work presented in Appendix A). Each part is detailed in a separate chapter, and each was dependent upon all of the preceding parts having been completed first, with the exception of Chapter 5, which was informed by certain mistakes in Chapter 2, but can otherwise stand independently. Still, in a sense all the following five chapters follow an order which is both logical and chronological, in accurately reflecting the time at which the various stages were completed.

Chapter 2 describes the inference of a binary trait matrix of protein presence/absence across eukaryotic species. By "protein" we specifically mean an orthologue[44] of a source protein found in a reference human proteome; such orthologues will generally have undergone amino acid substitutions since they diverged from their common ancestor, in addition to synonymous substitutions in their corresponding genes, so in reality we seek to identify orthologous sets of proteins across species, rather than sequence identities.

By treating protein presence/absence as a discrete trait, we can trace the evolution of that trait on

[44] A homologue found in another species; an orthologue can have paralogy in its past and may not necessarily therefore be found at a similar locus within its own genome as can its human equivalent.

a phylogeny. Each protein will either have been present at the root of the phylogeny, or gained at some point. It may have been lost never, once, twice or several times, on different branches of the tree. Ultimately, our matrix will allow us to extend our investigation to infer instances of correlated evolution between pairs of proteins. Although the methods we use for such inference do not require us directly to infer the branch on which a protein may have been lost, we expect that detection of correlated evolution will be reflected typically in two or more instances of correlated loss (i.e. both proteins in a pair being eliminated from the proteome on the same branch). Phylogenetic inference will be dealt with in Chapter 3.

My penultimate aim is to facilitate the inference of a network based on correlated evolution, and to describe the important features of this network and whether it successfully recapitulates actual functional relationships between proteins, both in terms of specific links and overall topology. Network inference is dealt with in Chapter 4.

Chapter 5 records some recent work undertaken by me to refine the protocols laid out in more detail in Chapter 2.

Chapter 6 is reserved for concluding remarks and recommendations for future investigation. Appendices and a bibliography are located after the body of the thesis.

As what ultimately turned into an aside to our investigation, in Appendix A, I outline some thought experiments aimed at elucidating the rules underlying the evolution of the correlated evolution network. From this I would like to have been able to go on to divine the evolutionary rules which fashioned the network, by reconstructing ancestral states [Omland, 1999].

Chapter 2

A binary trait matrix for proteins

2.1 Aims

The overarching goals for this project are: firstly, to demonstrate the utility of a set of computational statistical methods in predicting functional linkages between genes; secondly, to investigate and characterise the structural motifs in the network that emerges from the aggregate of the presumptive links thus inferred. We use raw sequence data from publicly accessible repositories as the primary (but not the exclusive) input to our experimental protocols. We break these protocols down into discrete parts, corresponding to the chapters of this thesis, apart from the Introduction and the final chapter—the Conclusion, wherein we discuss several of the many potential ramifications of our findings.

We set out by following the logic of Barker and Pagel [2005] in employing a computational procedure to detect the existence of particular genes (or proteins) within several dozen well-characterised organisms. Specifically, we seek to derive a trait matrix T, representing the presence or absence of proteins within a set of eukaryotic species each having a putatively complete published proteome. One dimension of T corresponds to a reference set of proteins, the other to our list of species. We can visualise species as the row index and protein as the column index, as in Fig. 2.1. Each element of T is a trait of the simplest possible kind: a binary entry t_{ij} which, if 1, denotes that the j^{th} protein is present in the i^{th} species, or absent if $t_{ij} = 0$.

In determining whether a protein is present in a given species, we are asking not whether a given sequence corresponding to that protein is perfectly replicated, but rather whether there is a close orthologue of some reference protein. Reference proteins come from a reference proteome—in

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & \dots \\ 1 & 1 & 0 & 1 & 1 & 0 & \dots \\ 1 & 0 & 0 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Figure 2.1: A binary trait matrix

this case that of *Homo sapiens*. I use the human proteome because it is at once extensive, well-characterised and of immediate scientific and medical interest [Legrain et al., 2011]. Our ideal, for each source sequence in the human proteome, is to derive a set of true orthologues whose pairwise relationships one to another are generally as close as the relationship between each non-human orthologue and the human reference sequence itself[1].

The trait matrix will serve not only as one of two principal inputs (the other being the phylogeny) for the correlated evolution analysis to follow, but has interesting properties in its own right. Most critically for this project as a whole, it will allow me to select a suitable set of genes from which to infer the phylogeny, described in Chapter 3, "A phylogeny of eukaryotes". To do so, in the best case we require proteins—and, by extension, genes—which are observed to be present in all species in the study.

It will be of interest also to observe how particular proteins are distributed across species [White, Hightower and Schultz, 1994]. I shall later make use of this knowledge to help check my control group in the test for correlated evolution [Pagel, 1994][2].

[1] By this I do not mean that all proteins within each interspecies orthologous set should be equidistant one from another in sequence space, but (rather less rigidly) that in the ideal case, a protein's evolutionary history—its gene tree—should track as closely as possible the tree of species themselves. See Section 2.4 for an example of the sorts of inferential problem that can arise if there is a mismatch between gene and species trees (i.e. if the gene tree has extra "hidden" branches within a single species' lineage).

[2] Of course, species are not independent, and traits will tend to be shared preferentially between closely-related species. To avoid a naive across-species comparison wherein closely-related species are wrongly treated as independent points, comparative methods can be used to apply a phylogenetic correction to trait data before any inferences are derived [Harvey and Pagel, 1991; Stone et al., 2011]. Although the comparative method lies at the heart of the current project taken in the round, the control group (of protein pairs) described later was not derived using a phylogentic correction, because it was felt that even the use of a "rough and ready" control group would be sufficient, and a considerable innovation of itself as compared with previous research. This is discussed more fully in Section 4.2.

2.2 Methods: Multiple whole-proteome reciprocal BLAST

I start with the human proteome, for the reasons outlined above. Specifically, I wish to identify, for every protein within the human proteome, whether a direct orthologue is present in each of the other species considered in this study. In order that it should be possible to determine this reliably, we require not only a complete set of sequences for the human proteome, but similarly complete sets for all these other species too. If we search for a human orthologue within an incomplete set of proteins from another species, and fail to find it, we can have little confidence that our negative "result" is not the consequence of an omission in the source data rather than a true absence. Thus we expect false negatives (type II errors) if the protein data for any of our species are incomplete.

Because it was my every intention to include as many complete eukaryotic proteomes as possible in the study, I had envisaged importing proteomes from several different curated repositories, which specialise in different sets of eukaryotic species. However, in the event and for various reasons—including considerations relating to curational consistency, fragile and ad hoc custom data retrieval protocols (developed as part of this study), inconsistency of species coverage among the various sources, as well as time constraints—I settled on NCBI's RefSeq [Pruitt et al., 2005], or Reference Sequence project, as my exclusive source of proteomic data. RefSeq's curators aim to make available a unique consensus sequence for each gene or protein recorded for many species within the NCBI's taxonomy database[3]. This is helpful because various ecotypes and allelic variants of any such gene may exist concurrently in GenBank [Benson et al., 2005]. When searching for orthologues of a given human gene in other species, and still more when trying to infer a deep phylogeny of eukaryotes using a sample of such orthologues, we are unconcerned with the small sequence differences at a particular locus that may exist between populations of a single species; these will generally be insignificant compared with inter-species differences but one needs to decide on a single sequence per gene, per species, in order to perform the analysis.

My methodology for inferring presence/absence is conceptually straightforward. For each protein in the human proteome, I use the program NCBI BLAST [Altschul et al., 1990] to determine which proteins in each other species' proteome are orthologous. Taking the top hit in the target organism, i.e. the sequence with the highest BLAST score, I again use the BLAST algorithm (in effect reciprocally) to find its orthologues in the human proteome. If the original source protein in the human proteome comes out as the top hit, I deem that the two are exclusively orthologous and that the protein in question is, in fact, present in the target species. This reciprocal BLAST proto-

[3] Most species (several thousand) which feature within RefSeq have curated reference sequences for a small subset only of their total complement of proteins: particularly in metazoa, this is often confined to mitochondrial orthologues of the 13 mitochondrial proteins identified within human beings—although a fourteenth such protein (humanin or HN) was identified just after the turn of the century [Hashimoto et al., 2001].

col is completely without novelty and has been widely used by others in the past [Hutter, 2000; Fuchsman and Rocap, 2006].

The proteomic data within RefSeq are stored in FASTA [Pearson and Lipman, 1988] text files, a popular format which is readily convertible into BLAST databases using the formatdb[4] command, part of the NCBI BLAST package itself. However, the data are not stored using a single file for a single species; rather, RefSeq's FASTA proteomes are stored in seven different directories, corresponding to different branches of the tree of life. One directory is reserved for prokaryotes, which do not form part of this study. The remaining six each accommodate a set of sequentially numbered FASTA files, all of which contain multiple protein entries, potentially from multiple species. My chosen methodology required me to separate the entries into species-specific files.

Because neither taxon IDs, nor any other means of identifying the species directly, are stored as part of the FASTA protein sequence record, the only way of matching species to protein entry is to take the RefSeq ID and cross-reference it with the RefSeq catalogue, which is also available for download[5]. Although the catalogue does contain the species name directly, to avoid any inconsistencies in identifying and naming the proteomes, I instead chose to take the taxon ID and then match it against the NCBI Taxonomy database, which I also downloaded separately. The most up-to-date version of NCBI Taxonomy is available as a gzip archive; the file `names.dmp` matches taxon ID against scientific name; it is a tab-separated-variable file, which I parsed to exclude all but the two fields of interest before importing into a MySQL [Kofler, 2001] database table. I then downloaded the catalogue (also a tab-separated-variable file) and imported it into a separate table.

Next, I downloaded and decompressed (using gunzip) the raw FASTA protein files. To split these up into files each containing one species' proteome only, I created a Perl program which uses the BioPerl [Stajich et al., 2002] module Bio::SeqIO to read and write individual protein sequence records. This it achieves by stripping out the RefSeq accession number, cross-referencing it with the taxon ID in the catalogue, which is in turn cross-referenced with the species name in the NCBI taxonomy table; the species name is used for the output FASTA proteome file, which is also suffixed with the taxon ID. The catalogue and all FASTA files downloaded as part of this study were from RefSeq release 36.

In contrast to sources like Ensembl [Hubbard, 2002] or the JGI [Grigoriev et al., 2012] for example, RefSeq makes no explicit claims about the completeness of the proteomes which it contains. In fact, the vast majority of species present are represented by a small subset only of proteins—typically mitochondrial. Other proteomes, including that of *Homo sapiens*, appear to be relatively enriched (at over 30 000) as compared with the most widely accepted estimates of gene number—even if this

[4] In BLAST+, a.k.a. BLAST2 (especially on some Linux systems), formatdb has been superseded by makeblastdb.

[5] Available (current version only) by ftp from `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/`

remains itself to some extent an open question [Pertea and Salzberg, 2010]. Despite this uncertainty, the unexpectedly high number of human proteins in RefSeq seems likely to be a result of the inclusion (in well-annotated RefSeq proteomes) of splice variants, which will be elevated in comparison to gene count [Sorek, Shamir and Ast, 2004]. Even with such relatively substantial proteomes, I sought assurance that coverage was comprehensive for each species in the study: otherwise, I could not have had confidence that a given human orthologue was truly absent when I came to perform the reciprocal BLAST—carrying with it the aforementioned risk of generating type II errors. Complete proteomes exist in varying states of annotation, and hence of reliability. Many proteomes, even those described as complete, rely mainly on predictions from homology of known genes in other species. In this case the initial prerequisite is that the genome itself be fully sequenced. Although an authoritative list of all complete eukaryotic proteomes is elusive, in that the several repositories hosting complete proteomes do not claim to provide proteomes for every completed species, nevertheless equivalent lists of completely sequenced *genomes* do exist. Note however that annotated proteomes follow in the wake of genome sequencing: the intevening steps require further computation and curation. Moreover, putatively complete proteomes may be available from multiple sources, in various states of reliability and completeness. Because RefSeq seeks to provide reference sequences with high curational integrity, “complete” proteomes for most sequenced organisms will usually become available from other sources before they appear in RefSeq. Nonetheless, RefSeq continues to grow. From the “Fungi” tab on the “statistics” page of the RefSeq web site, on 7 July 2008, there were 100 fungal proteomes available; by 7 July 2014, this number had grown to 2 859. It is not clear from RefSeq’s statistics web page whether all the fungal proteomes at the point at which the research for this study was conducted (i.e. relating closely to the 2008 figure) were deemed to be complete, but this appears not to have been the case, particularly in light of the plethora of minimal mitochondrial proteomes only (for many eukaryotes), which are currently available in RefSeq. In any event, this study incorporates only 23 fungal species, based on the filtering procedure for complete proteomes described below.

Probably the oldest list of completely sequenced genomes is GOLD [Liolios et al., 2006], the Genomes OnLine Database. My method was to ascertain which eukaryotic genomes had been fully sequenced, then to check, for each species, whether its genomic sequence was available in RefSeq, and furthermore to verify that a compete proteome seemed to be available as well. For the first part of this process, GOLD seemed an appropriate resource, and one that lends itself, albeit a little clumsily, to formal proceduralisation, thanks to its downloadable tab-delimited table listing all published fully sequenced genomes, including their source repositories. I used this table to generate a list of all published eukaryotic genomes, which I then checked against the species-specific proteome file names already generated in parsing the proteomes from RefSeq. Instances where a match was found on taxon ID were checked again to ensure that the size of the proteome from RefSeq was consistent with a complete set of proteins. My first step was to filter for a proteome of $\geq$ 500 kB, a lower limit just a little smaller than that of the smallest proteome in the study, belonging to *Bombyx mori*, the silkworm, whose proteome I had presumed to be reduced thanks to millennia of artificial selection[6]; crucially however, see Chapter 5, "A trait matrix remade: improved protocols". Where the proteome met this size criterion, I selected the species for inclusion in the reciprocal BLAST analysis, with some checks for consistency of proteome size with other curated sources: again, see Section 2.4. A few of the species' taxon IDs seemed to be misspecified in GOLD, referring to different strains or variants of species which were clearly complete in RefSeq. Some manual selection was therefore necessary, but this was the exception rather than the rule. To the final list thus derived, I added two species: *Strongylocentrotus purpuratus*, the purple sea urchin, and *Trichomonas vaginalis*, the causal agent of trichomoniasis. Both organisms, despite having been absent from the GOLD listing, have complete proteomes in RefSeq 36.

Before creating the BLAST databases, one more step was necessary. Each of the complete proteome files contains several thousand entries, a certain proportion of which are duplicated sequences. Had I left the duplicates in, there would have been the potential for the reciprocal BLAST process to generate false negatives, because if the source protein in the first step has one or more exact duplicates, there is a chance that it will not be returned as the top hit in the second step, even though it will score exactly the same as the duplicate which *is* returned. I therefore wrote a Perl program, which again invokes BioPerl's Bio::SeqIO module, to flush out all duplicated sequences from each proteome by extracting, comparing and eliminating them. It matters not which specific sequence(s) among a set of duplicates is/are eliminated, just that each one that remains be unique. See Section 2.3 for details of the gross number of sequences, plus the number of duplicates, for each organism.

I used a Bash shell script to output a BLAST database from each of the duplicate-free FASTA proteomes. Because to perform a reciprocal BLAST from the human proteome to all the other complete proteomes in the study would have taken an inordinate amount of of time (up to a few weeks) on a standard desktop PC, I uploaded the databases to a 100-processor cluster housed at the Universiy of Reading, and wrote two Perl programs: one to perform the BLAST processing itself, the other to split the BLAST jobs among the cluster's 62 nodes (38 of which housed two processors). I used NCBI BLAST 2.2.19, accepting most of that program's defaults. However, I used a modified expectation (E) value, which denotes the number of matches that would be expected purely by chance (i.e. if there were no actual homology between the source sequence and any of the sequences in the target database) from a BLAST query. By default this is set to 10, which seems unduly lax. To constrain my result set to those target sequences which are overwhelmingly likely to possess and exhibit true homology to the query sequence, I set the expectation value to one in a million,

[6] In fact, it appears that while artificial selection might indeed have left its imprint on *B. mori*, that imprint emphatically does not manifest itself as a reduction in genome size [Xia et al., 2009].

i.e. $E = 10^{-6}$. Occasionally this resulted in the BLAST query returning no results at all, in which case I simply recorded the protein as absent in the target.

Output from the reciprocal BLAST was produced in the form of a series of text files containing one line per human protein, with the GI accession number of the protein itself in the first column, the rest of the columns being a tab-separated set of ones and zeroes denoting presence or absence in all species in the study; obviously, this will always take a value of 1 for *Homo sapiens*, but the column was included nonetheless[7]. These text files were concatenated into a single file encapsulating a complete presence/absence matrix with one row for each unique sequence in the human proteome (over 30 000 in total). The single file was stripped of its tabs before being loaded into a MySQL table, which stores the two fields of interest (GI number and the presence/absence fixed-width string of ones and zeroes) plus another field computed by the loading script, this being the overall number of species out of my sample in which the protein was reckoned to be present. This number is used later in the likelihood analysis to select proteins with a spread of presence and absence, since in the extreme cases of ubiquity on the one hand, and exclusivity to the human organism on the other, a protein cannot meaningfully be tested for correlated evolution with any other, regardless of the latter's pattern of presence/absence. I note that a simple sum of the number of organisms in which a protein is deemed to be present is an inherently biased statistic, in that it takes no account of phylogeny—which is is my precise objective in the context of the overall study. It is nonetheless presented in the results as a distribution of protein frequency among my particular, phylogenetically biased set of organisms.

In tandem with the BLAST databases, I needed a way of mapping unique human source sequences (from which duplicates are deliberately excluded) back to the actual, potentially multiple sequences in the original human proteome that they represent. This would later enable me to represent accurately the frequencies of the original proteins in the likelihood analyses, for both curated interactors from HPRD[8] and presumptive non-interactors. I therefore wrote one program to populate a MySQL table with all the human sequences, and another to map each genuinely unique RefSeq ID in the full proteome to a GI number representing a sequence from the unique set which might be common to multiple RefSeq IDs. (The GI number is real; for each unique sequence, it represents the GI from the first record found with that particular sequence). This is a many-to-one mapping, even though a majority of sequences do not have exact duplicates. I would later rebuild the trait matrix, reassigning its entries to interactor IDs from HPRD: see Chapter 4, "A correlated evolution network".

[7] I did not run a reciprocal BLAST from the human proteome back to itself, because this would have been redundant.

[8] The Human Protein Reference Database [Peri et al., 2004], henceforth referred to in the text only as HPRD.

2.3 Results

According to my stated criteria, I determined that 85 complete eukaryotic proteomes were available in RefSeq release 36. For reasons to be explained in Chapter 3, I would go on to eliminate 13 species from the study, leaving 72. Total number of sequences varies widely among the eukaryotes. The totals found in this study, together with a summary of duplicates for each of the 85 species, are shown in Table 2.1. Those species which did not feature in the likelihood analysis described in Chapter 4 are greyed in comparison to the other entries.

Table 2.1: Total and duplicate sequence counts[9] for complete eukaryotic proteomes from RefSeq 36

	total	**unique**							
species	**sequences**	**sequences**	**1**	**2**	**3**	**4**	**5–9**	**10–99**	**100+**
Aedes aegypti	16 798	16 059	15 478	483	71	17	8	2	
Anopheles gambiae str. PEST	13 123	12 889	12 707	145	29	5	3		
Apis mellifera	9 268	9 199	9 136	57	6				
Arabidopsis thaliana	33 402	31 220	29 508	1 379	249	54	28	2	
Ashbya gossypii ATCC 10895	4 725	4 720	4 716	3	1				
Aspergillus fumigatus Af293	9 630	9 630	9 630						
Aspergillus nidulans FGSC A4	9 541	9 528	9 515	13					
Aspergillus niger CBS 513.88	14 086	14 077	14 074	2			1		
Aspergillus oryzae RIB40	12 074	12 051	12 037	9	3	1	1		
Babesia bovis T2Bo	3 706	3 690	3 675	14	1				
Bombyx mori[10]	1 680	1 679	1 678	1					
Bos taurus	22 879	22 322	21 866	406	35	4	8	3	
Branchiostoma floridae	50 755	49 779	48 872	878	19	5	3	2	
Brugia malayi	11 472	11 349	11 250	85	10	2	2		
Caenorhabditis elegans	23 906	23 726	23 611	102	5	3	1	4	
Candida albicans SC5314	14 633	9 283	4 122	4 986	164	9	2		

[9] The second column, "total sequences", refers to the total count of protein sequences from RefSeq for each species, including all duplicate and multiplicate instances of identical sequences originating from different genomic loci. The third column, "unique sequences", gives the count, by species, of sequences with additional copies of any identical sequences removed: in fact it is the straightforward sum of all 7 columns to its right, which simply sort the unique sequences into buckets according to how many times they occur within the proteome. So counts under the column heading "2" refer to the non-redundant number of sequences which occur twice: to get the total for all the duplicates, multiply this number by 2. Similarly, to get the number for all the triplicates, multiply the numbers in the column headed "3" by 3—and so on. In general, most sequences have no exact intraspecific matches: species counts for these appear in the column headed "1". Note that the total sequence count in the second column can be obtained by summing the products, i.e. of all the uniquely defined sequence counts, and the number of times each sequence occurs (2 for duplicates, 3 for triplicates, and so on). However, it is not generally possible to obtain such products definitively from columns covering a range of multiplicates (e.g. 5–9).

[10] The count of 1 680 proteomic sequences for *B. mori* is an anomaly: in fact it is artefactually low, for reasons explained in Section 2.2, especially Footnote 6 on page 30.

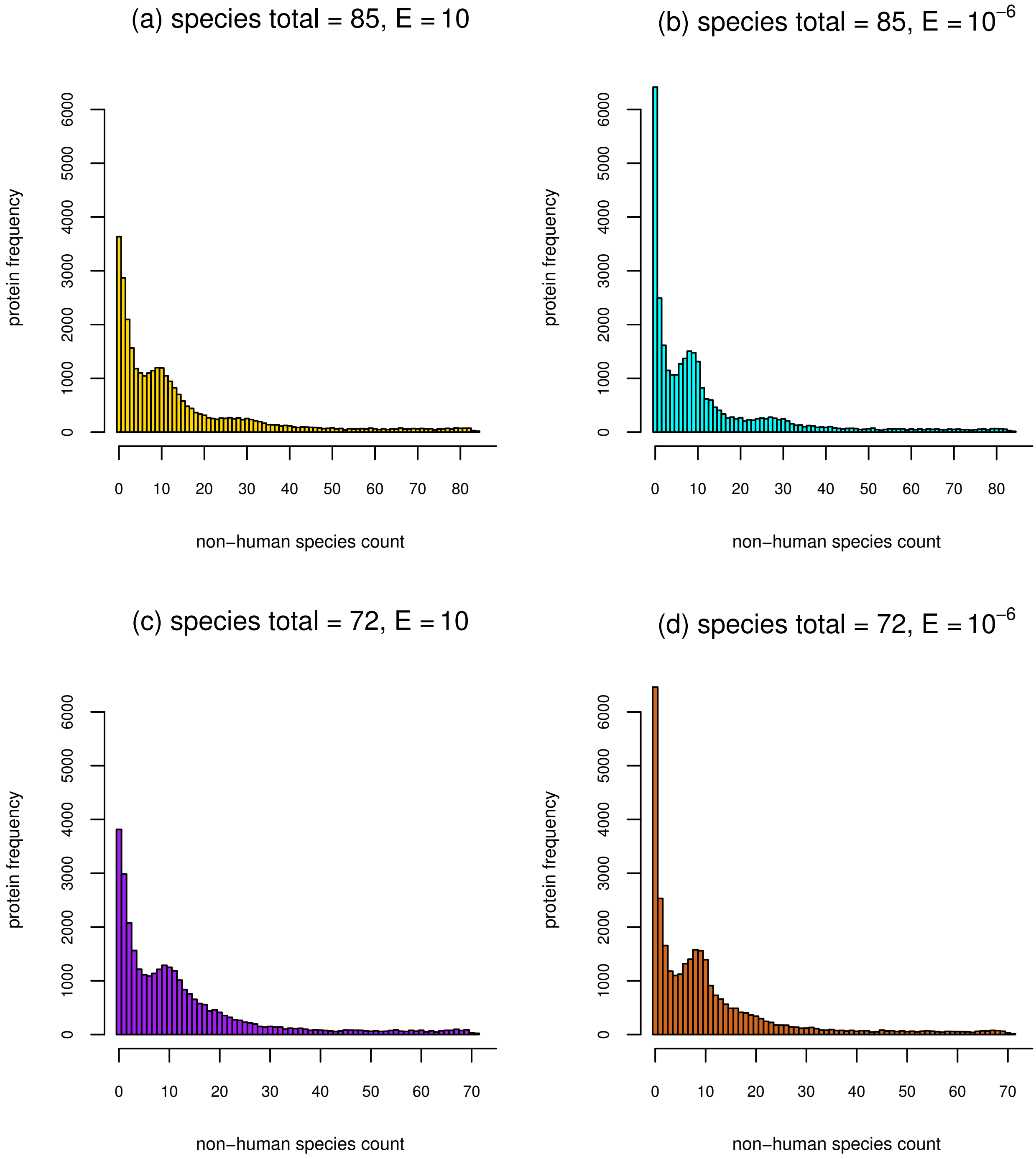
species	total sequences	unique sequences	1	2	3	4	5–9	10–99	100+
Candida glabrata CBS 138	5 192	5 180	5 170	8	2				
Canis lupus familiaris	33 502	31 980	30 942	775	158	55	46	4	
Chlamydomonas reinhardtii	14 489	14 338	14 280	48	2	1	4	3	
Ciona intestinalis	13 945	13 802	13 677	110	12	3			
Cryptococcus neoformans JEC21	6 594	6 443	6 309	120	11	3			
Cryptosporidium hominis TU502	3 885	3 885	3 885						
Cryptosporidium parvum Iowa II	3 805	3 805	3 805						
Danio rerio	27 850	27 068	26 500	514	31	6	10	7	
Debaryomyces hansenii CBS767	6 317	6 310	6 303	7					
Dictyostelium discoideum AX4	13 408	13 032	12 698	324	5	1	2	2	
Drosophila ananassae	15 070	14 954	14 913	28	6	1	1	5	
Drosophila erecta	15 048	14 868	14 827	28	6	1	1	5	
Drosophila grimshawi	14 986	14 741	14 620	113	4			4	
Drosophila melanogaster	21 099	18 122	16 213	1 309	379	122	94	5	
Drosophila persimilis	16 878	16 741	16 645	80	9	2	5		
Drosophila pseudoobscura pseudoobscura	16 071	15 995	15 930	58	5	1	1		
Drosophila sechellia	16 484	16 138	16 012	105	11	1	5	4	
Drosophila virilis	14 491	14 382	14 347	27	4			4	
Drosophila willistoni	15 513	15 434	15 375	49	6	1	3		
Drosophila yakuba	16 095	15 953	15 812	140	1				
Encephalitozoon cuniculi GB-M1	1 996	1 908	1 853	38	10	1	6		
Entamoeba histolytica HM-1:IMSS	8 163	7 958	7 803	121	24	6	4		
Gallus gallus	18 734	18 509	18 354	119	19	8	9		
Giardia lamblia ATCC 50803	6 502	6 329	6 176	137	13	2	1		
Gibberella zeae PH-1	11 640	11 638	11 636	2					
Homo sapiens	37 946	31 120	26 781	2 393	1 658	186	94	8	
Kluyveromyces lactis NRRL Y-1140	5 327	5 308	5 292	14	1	1			
Laccaria bicolor S238N-H82	18 215	17 895	17 607	266	17	2	3		
Leishmania infantum JPCM5	7 992	7 872	7 760	107	4		1		
Leishmania major strain Friedlin	8 265	8 001	7 865	101	16		15	4	
Macaca mulatta	37 851	32 836	29 982	1 748	573	291	229		
Magnaporthe grisea 70-15	14 010	13 459	12 972	468	8	4	4	3	
Malassezia globosa CBS 7966	4 286	4 274	4 266	4	4				
Monodelphis domestica	20 206	20 120	20 060	49	7	2	1	1	
Monosiga brevicollis MX1	9 171	9 155	9 145	5	4	1			
Mus musculus	35 651	30 713	26 752	3 476	309	101	60	15	
Nematostella vectensis	24 780	24 430	24 239	161	10	6	7	7	
Neurospora crassa OR74A	9 841	9 813	9 787	24	2				
Ornithorhynchus anatinus	16 680	16 590	16 560	23			5	2	
Oryza sativa Japonica Group	26 940	26 858	26 790	62	3	1	2		
Ostreococcus lucimarinus CCE9901	7 603	7 403	7 206	195	1	1			

species	total sequences	unique sequences	1	2	3	4	5–9	10–99	100+
Pan troglodytes[11]	51 405	41 822	37 215	2 594	905	457	570	81	
Paramecium tetraurelia strain d4-2	40 043	39 399	38 828	541	12	4	12		
Phaeodactylum tricornutum CCAP 1055/1	10 408	10 336	10 268	65	2	1			
Physcomitrella patens subsp. *patens*	35 894	35 552	35 294	213	33	3	7	2	
Pichia stipitis CBS 6054	5 816	5 797	5 782	13		2			
Plasmodium falciparum 3D7	5 262	5 244	5 231	10	2		1		
Plasmodium yoelii yoelii str. 17XNL	7 353	7 302	7 270	23	6	1	2		
Podospora anserina DSM 980	10 219	10 219	10 219						
Populus trichocarpa	42 344	41 784	41 312	419	36	11	5	1	
Postia placenta Mad-698-R	9 083	8 983	8 889	88	6				
Rattus norvegicus	30 920	26 709	22 728	3 841	89	41	8	2	
Saccharomyces cerevisiae	5 884	5 821	5 769	45	3	4			
Schizosaccharomyces pombe	4 994	4 948	4 911	31	3	3			
Sorghum bicolor	32 520	32 353	32 227	104	18	1	2	1	
Strongylocentrotus purpuratus	42 324	24 420	6 924	17 294	51	140	9	2	
Tetrahymena thermophila	24 770	24 743	24 717	25	1				
Thalassiosira pseudonana CCMP1335	11 673	11 612	11 562	46	1	1	2		
Theileria annulata strain Ankara	3 792	3 787	3 782	5					
Theileria parva strain Muguga	4 061	4 050	4 039	11					
Tribolium castaneum	9 883	9 853	9 825	27		1			
Trichomonas vaginalis G3	59 679	50 187	48 037	1 228	343	154	236	178	11
Trichoplax adhaerens	11 537	11 520	11 505	13	2				
Trypanosoma brucei TREU927	8 712	8 514	8 383	108	11	4	5	3	
Trypanosoma cruzi strain CL Brener	19 607	19 245	18 951	254	26	10	3	1	
Ustilago maydis 521	6 522	6 518	6 514	4					
Vanderwaltozyma polyspora DSM 70294	5 367	5 345	5 326	16	3				
Vitis vinifera	23 493	23 081	22 733	302	39	2	5		
Yarrowia lipolytica CLIB122	6 448	6 425	6 417	3	5				

After running the reciprocal BLAST itself, based on the unique sequences only, I observed the patterns of frequency distribution among the found orthologues displayed in Fig. 2.2 on the facing page. I include both my favoured reciprocal BLAST results from the later run with $\mathrm{E} = 10^{-6}$ and the results from my initial run with the BLAST default $\mathrm{E} = 10$ for comparison. The left-hand column represents the older default expectation parameter, the right-hand my one-in-a-million criterion. The top row gives the results for all 85 species, the bottom for the 72 species in the

[11] Total proteomic sequence tally for *P. troglodytes* is substantially overstated in comparison to the *H. sapiens* figure. This is presumed to be due to the relatively immature status of the chimpanzee genome assembly within RefSeq 36. Good evidence for this is that of 51 405 chimpanzee protein sequences, fully 50 080 had an "XP" prefix, denoting a predicted protein, with only 725 proteins based on the accepted assembly (denoted by the prefix "NP"). Compare this to *H. sapiens*, with 9 942 predicted proteins, and 27 991 already in the assembly; there are also 13 human "AP" proteins, presumed to be mitochondrial. See Table 5.1 on page 78 for a list of RefSeq sequence prefixes relating to proteins.

likelihood analysis only. Therefore Fig. 2.2(d) gives the results which relate specifically to the inputs to the likelihood analysis. The quoted species totals include *Homo sapiens*; human proteins with no orthologues appear in the "zeroth" bar in each histogram.


(a) species total = 85, $E = 10$

(b) species total = 85, $E = 10^{-6}$

(c) species total = 72, $E = 10$

(d) species total = 72, $E = 10^{-6}$

Figure 2.2: Human protein frequency by count of species exhibiting orthology: for different E and species totals

Finally for this section, in Fig. 2.3 on page 43 I present totals of human orthologues in each individual non-human species, leaving aside the trivial case of *Homo sapiens* itself, with 31 120 unique sequences. Species are presented in order of the number of orthologues found, left-to-right. Totals appear beside each bar, to the right.

2.4 Discussion

There are pitfalls in the reciprocal BLAST methodology. As a tool for detecting orthology specifically, and homology more generally, it could be judged to be effective but crude. I draw attention to two problems in particular: firstly, the weakness of heuristic local alignment in identifying best matches; secondly, the potential for questionable categorisation of orthologous sets, chiefly owing to gene duplication and loss.

NCBI BLAST [McGinnis and Madden, 2004], and WU-BLAST [Gish, 1996] for that matter, employ a heuristic search algorithm which finds short aligned sections of sequence (protein–protein in this case) and tries to produce an alignment by extending the aligned sections at either end using an amino acid substitution matrix—usually BLOSUM62 [Henikoff and Henikoff, 1992]—which also enables the overall alignment to be given a score. Low-scoring alignments can be rejected, while high-scoring alignments are retained in the BLAST report. While this approach is relatively quick, and is typically the first resort of the researcher searching for homologues of a given sequence, we should note that because it is based on short exact matches in the "seed" sequences, it is not necessarily optimal for discovering the best possible pairwise alignment, particularly in more distantly-related sequences. For this we might look toward a dynamic programming algorithm [Eddy, 2004] which can perform a global alignment by thoroughly weighing the possible state transformations between the sequences, albeit at the cost of very significantly increased processing time, typically around two orders of magnitude as compared with BLAST (and variants thereof). This in fact forms part of the hybrid approach adopted by Barker, Meade and Pagel [2007], who used blastp to isolate up to 20 of the highest-scoring matches from whole proteomes before using EMBOSS [Rice et al., 2000] needle, an implementation of the Needleman-Wunsch global alignment algorithm [Needleman and Wunsch, 1970], to find the best match out of this reduced set of candidates. This two-stage process is still acceptably quick and is very likely to find the "true" best match; using needle to query the entire proteome would be very much slower.

In spite of these reservations, arguably our reciprocal BLAST process does a fairly good job on its own of identifying the closest orthologue (to a given human protein) present in another species, given constraints of processing time. However, assuming that for each non-human species we can successfully infer the closest one-to-one orthologue mapping between human and non-human proteomes, are we justified in going on from here to characterise the collection of proteins found in this manner in the various non-human species as a true orthologous set? This has implications both for the derivation of presence/absence, and by extension (and perhaps even more importantly) for the phylogenetic inference, which depends critically on identifying such exclusive orthologies with accuracy. I will try to infer the tree of species based on a tree of genes, which constitute my data; because for this purpose I wish to use those genes which are both present in all species, and whose evolutionary bifurcations correspond with speciation events, a misidentified orthologue has the potential to give rise to a misleading species tree. Yet how could this happen under the already stated (if not completely realistic) assumption that given the sequence for any human protein, we can unfailingly pick out its closest match from another species' proteome database?

The chief source of difficulty is the fact that genes' evolutionary histories can be more convoluted than those of the organisms in which they are found. A gene can undergo a duplication within the same genome, giving rise to a paralogue. This can happen an arbitrary number of times on a single (organismal) evolutionary branch, i.e. during a period when no speciation takes place. Several rounds of such duplications can give rise to multiple paralogy. Not only can paralogues accumulate in this way; they can be eliminated as well. In fact, this may be typical and to be expected given that in the most probable case, paralogy will confer not evolutionary advantage but simply redundancy [Wagner, 2003]. On a gene tree inferred from the sequences of a set of paralogues, the historical evolutionary relationships among the paralogues may be apparent, but what will be missing are these losses—gene extinction events, effectively. Only when orthologues from other species are included in the gene tree may some such losses be exposed, because a speciation event following a proliferation of paralogues in the common ancestor gives us two repository species in which all the paralogues extant at the time of speciation are carried forward—to be eliminated, retained or duplicated further. Note that such a gene tree will contain bifurcations corresponding both to intragenomic duplications and speciation events [Tatusov, 1997]. To illustrate the potential complications which this might present for our simple reciprocal BLAST process, we may (hypothetically) use phylogenetic inference to construct a very simple gene tree for a hypothetical gene which may have been present in an ancestral primate, whose hypothetical evolution we trace into the human, chimpanzee and macaque descendent lineages. Let's say the gene underwent a prehistoric duplication before the speciation event which separated the ancestral macaque from the common ancestor of humans and chimps. Finally, the human and chimp lineages themselves split apart. Because the gene is present in each lineage at each bifurcation, effectively it became present in six separate versions. However, because it has been lost on two separate occasions following the most recent bifurcations, only four versions of the gene have survived until the present: both duplicates in the chimpanzee and one each in humans and macaques. This is illustrated in Fig. 2.4 on page 44, where α and β represent the initial duplicates, while the subscript m represents the macaque lineage, h the human, c the chimpanzee, and hc that of the most recent common human/chimp ancestor.

Starting with the human proteome, if we BLAST α_h against the chimp proteome, we are likely to score a reciprocal top hit on α_c, whereas if we BLAST it against the macaque proteome, we are likely to score a reciprocal top hit on β_m. So, if we didn't know any better, we would infer a mini-set of direct orthologues across the three species. However, consider starting from the macaque proteome, with its one remaining copy of the ancestral protein, β_m. We have already stated that this will probably undergo a successful reciprocal BLAST with α_c in the human proteome. However, targeting the chimp proteome, in the most likely event β_m will BLAST reciprocally with β_c. So if we start from the human proteome, we come up with the inference that $\{\alpha_h, \alpha_c, \beta_m\}$ constitutes an orthologous set, with β_c being unlikely to get included in any other orthologous sets and therefore being left out altogether. If on the other hand we start from the macaque proteome, we infer $\{\alpha_h, \beta_c, \beta_m\}$ to be an orthologous set, this time with α_c excluded altogether. Of course, if we were to start from the chimp proteome, in which both descendants of the initial intragenomic duplication survive, we would get two orthologous sets of proteins, $\{\alpha_h, \alpha_c\}$ and $\{\beta_c, \beta_m\}$; the former orthologue would be marked as absent in the macaque, the latter as absent in the human, since in each case the reciprocal BLAST would return not to the source protein but to its paralogue in the chimp proteome. So not just the extent, but the character of our binary trait matrix too, is in some measure dependent on the proteome we choose to start from, and the particular evolutionary relationships it shares with the other proteomes in our study, at a genomically localised level which has the potential to elude the rather coarser-grained methodology which we employ.

The hypothetical sequence of events outlined above is problematic not only for the derivation of the trait matrix itself, but perhaps even more importantly, in its implications for phylogenetic inference. In fact, it is possible that something close to this very scenario has already manifested itself during the course of the preparatory work for this project [Read, 2007]. In my 85-species reciprocal BLAST, the very last bar on the right of Fig. 2.2(b) actually represents 15 proteins found to be common to all species; admittedly this is difficult to see given the scale of the y axis. As will be explained in more depth in Chapter 3, it was these 15 genes which I elected to use for phylogenetic inference. An earlier run of the reciprocal BLAST, using only 48 species, identified a slightly different set of proteins which had an orthologue in all species. In like manner to this study, these genes[12] were then used to infer a phylogeny, or rather multiple phylogenies, since as well as using the concatenated sequences of all such genes to infer the main phylogeny, I had been interested to see how the topology would be affected by generating phylogenies from each gene sequence individually. In fact, the most striking feature of all the trees based on individual genes was the presence of one very long branch (i.e. at least three times longer than the second longest branch); more interestingly, the long branch belonged to a different species in each tree. In the tree generated using sequences from the orthologous set of the human gene isoleucyl-tRNA synthetase

[12] For reasons which will also be explained in the Chapter 3 (strictly in the context of the present study), the nucleic acid sequences from the gene orthologues, rather than the protein sequences, were used to infer the phylogeny.

(IARS), the long branch was that of *Macaca mulatta*; in fact, this was the longest branch observed in *any* of the single-gene trees.

A reciprocal BLAST from the macaque IARS orthologue against all other species' proteomes might have thrown further light on this apparent anomaly, but unfortunately this was not performed. Therefore any inference drawn at this point is strictly speculative, but it seems unlikely that the macaque protein's extraordinary apparent rate of evolution was real. More likely (and assuming the reciprocal BLAST did not simply miss the true orthologue for artefactual reasons to do with the less-than-perfect BLAST-only algorithm) is that the macaque protein found is not a true orthologue of the human source protein at all, but that the real orthologue has been lost from the macaque. We may extend the speculation to wonder if the last common ancestor protein of the human and macaque proteins predates by many millions of years the last common ancestral organism. We may imagine a gene duplication, a subsequent divergence of function between the two variant genes, followed by loss of one variant in the human lineage and of the other in the macaque lineage, which could account for the "success" of the reciprocal BLAST.

All of which prompts the question, what happens after gene duplication? Gene duplication has been proposed as one of the main driving forces behind evolution itself [Ohno, 1970]; plausibly, almost all novel genes which confer evolutionary advantage by means of functional innovation arise through duplication events. If both child genes arising from the duplication of a functional parent gene are themselves fully functional, an immediate redundancy in function arises. Until one of the duplicates undergoes a mutation which destroys or compromises its function in some way, purifying selection is likely to become more relaxed [Zhang, 2003]. Once one of the duplicates does start to diverge, however, the overwhelmingly likely consequence is that its function will be impaired [Wagner, 1998], while the other is likely to come back under more stringent purifying selection, and to remain strongly constrained. One can imagine selection operating rapidly to separate the functions of the sibling proteins once one of them starts mutating, purifying selection acting as negative feedback on the unmutated sibling and maintaining stasis, while genetic drift slowly pulls the other away, through sequence space, as it undergoes further mutation without check.

If the divergent duplicate remains free of purifying selection, it may meet one of several fates. It may be lost; the evidence is that this is the most likely outcome [Lynch and Conery, 2000] and would be expected eventually if its retention conferred no selective advantage. It may exploit its freedom from the shackles of purifying selection to wander over a fitness landscape before climbing towards a local optimum, bringing it under positive selection for some novel trait. Alternatively, the (otherwise potentially lethal) loss of its sibling duplicate may bring it suddenly back under purifying selection; in this scenario we would expect it to have retained aspects of its original function. In the case where it is free to wander over a fitness landscape, it might scale some peak which enables it to perform its original function better than the gene which had remained constrained, at which point it could come back under purifying selection itself, while liberating its sibling to wander the local fitness landscape in a kind of evolutionary ping-pong. Dehal and Boore [2005] have made a powerful case that major gene duplication events involving multiple chromosomes are associated with major historical evolutionary innovations in the eukaryotic tree. Smaller, more localised gene duplications might account for less punctuated evolutionary innovation, whose steady accumulation nonetheless contributes significantly to the diversity of organismal form and function in existence today [Shubin and Marshall, 2009].

Phylogenetic reconstruction of gene trees is one way of establishing the history of gene lineages with more certainty than a BLAST protocol alone can offer, and is interesting in its own right, but at the time of writing does not scale well given the existing limits to computational power. It might be reserved for problematic genes, however.

If we wish in future to address some of the issues around orthology discussed above, the various strategies are expensive in terms of computing time, whether we seek to employ phylogenetic methods (see Chapter 3) or simply to expand our BLAST search to give us the top hit(s) from every species' database found by matching every other protein in every other species' database against it. In the latter case, this applies whether we employ BLAST alone (in any of its variants), or whether we use it as a filter for a more limited, but definitive, search using a dynamic programming algorithm like Needleman-Wunsch. Either way, run time in the all-species-against-all-species scenario will increase by a factor roughly equal to the number of species in the analysis, compared with using the human proteome as the sole source proteome as per our reported method; also, run time would vary in rough proportion to the square of the number of species in the analysis, rather than linearly, as at present.

We might therefore profit from taking an approach suggested by Korf [2003], whereby BLAST searches are run serially, using different input parameters on each iteration, two iterations being typical. In this scheme, the first run uses very insensitive parameters, while subsequent runs become increasingly stringent: using high stringency for BLASTing against unrelated sequences is very costly in terms of processor time, but we *can* use stringent criteria in a fast BLAST protocol if we confine ourselves to a more limited set of targets which are already presumed to show some level of similarity to our source sequence.

2.5 Addendum: Protocols for incorporating additional proteomes

Although this study has been limited to proteomes extracted from the repositories in RefSeq, the original intention was to incorporate complete proteomes from as many different sources as possible, so as to maximise the coverage of eukaryotic species. A number of institutions maintain curated collections of such whole-proteome sequence data; work on bringing them into this project made some progress before it was abandoned owing both to considerations of time and data consistency. I briefly describe these efforts below.

Apart from RefSeq, projects hosting a range of complete eukaryotic proteomes include Ensembl [Hubbard, 2002], Integr8 [Pruess et al., 2005], the International Protein Index (IPI) [Kersey et al., 2004] and the Joint Genome Institute (JGI) [Grigoriev et al., 2012]. Like RefSeq, at the time this study was originlly conducted (2009), all of these repositories were updated every few weeks or months. Because none of them provides a simple downloadable archive containing all the proteomes in each release, I wrote Perl programs of various levels of complexity to retrieve all the files automatically, anticipating that the process might benefit from being re-run occasionally to accommodate updates to the data. For the purpose of combining the data after download, I also developed various ways of incorporating taxon IDs into the file names on my local machine.

I wrote a Perl program to download FASTA proteomes from Ensembl by calling ftp. It retrieved the respective taxonomy IDs from the Ensembl online databases using Perl's MySQL DBI interface.

Intergr8 from the EBI provides a SOAP[13] module which I called from within another Perl program to search for the required files by superkingdom (or "superregnum")—in our case, eukaryota. These FASTA proteomes are based on the UniProt Knowledge Base; taxon IDs are returned along with the remote file names in the SOAP search, so were easy to incorporate in the local file names after download.

Some of the proteomes rely heavily on computational inference from nucleotide sequence data, so can vary dependent on the inferential algorithms used. Therefore, because there is some overlap in the species for which these various sources (including RefSeq) hold data, I also developed a way of selecting unique sequences from the combined sources, so as to make the coverage of the proteomes themselves as comprehensive as possible. This was done in preference to simply choosing a favoured data source for each species. In making this choice I ran the risk of incorporating multiple non-consensus variants of the some of the same sequences in each file, and significantly overstating each species' real number of unique proteins. However, this was a conscious

[13] Simple Object Access Protocol: one of the standards defined for the provision of web services [Scribner and Stiver, 2000].

calculation. Crucially, if I chose a single data source, namely RefSeq, for the human proteome, my reciprocal BLAST process (which starts from the human proteome) should still have returned a plausible number of orthologues, although the individual orthologue sequences themselves (i.e. the top BLAST hits) might have been different. If there was any systematic variation in the way the various institutes curated their proteomic data, my strategy was going to be to try and minimise its effect by incorporating all of them. Essentially this process was the same as the one already referred to in Section 2.2, which eliminates any duplicate sequences. The only difference was the use of multiple input files, one per source proteome for each species; species were identified by placing the taxon ID in the file name as described above (in the current section).

This avenue of research was laid to one side soon after the stage where the proteome files could be downloaded and combined automatically. It might be profitable to revisit it in the future, as it would certainly facilitate the incorporation of many more species in the study. The total number of species for which I generated combined proteome files was 189. At the time when this much larger study was shelved, there remained unresolved problems in generating BLAST databases from a handful of the combined proteomes. Were the study to be conducted again in full at the time of writing (2014), it is unlikely that all the original data sources would be accessible using the protocols employed during this early phase of the study (2008–9), given the difficulties experienced at the time. Based on observed species counts within RefSeq however, it is altogether possible that a study incorporating 200+ complete proteomes would be feasible.

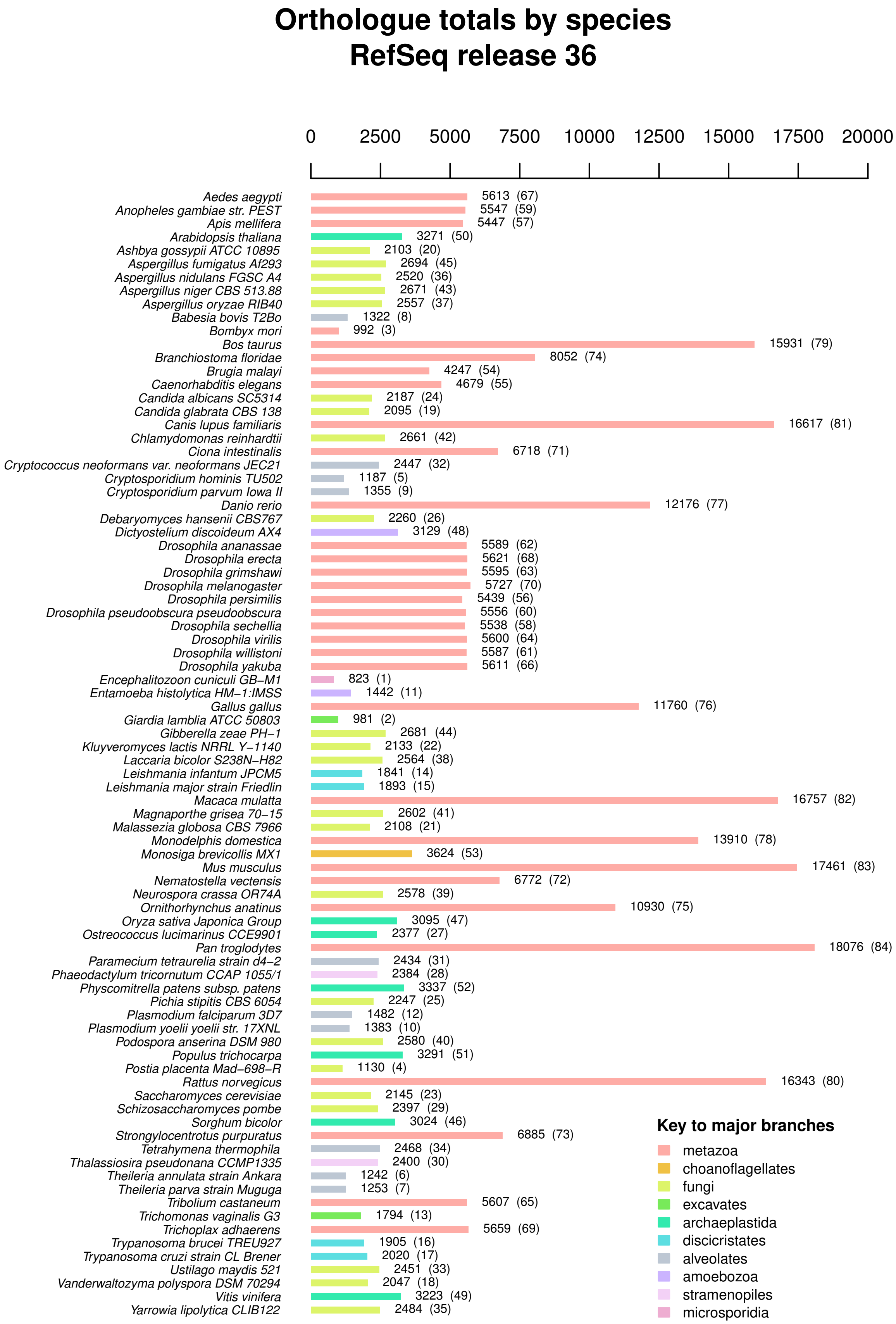


Figure 2.3: Nos. of human orthologues in other species, coloured by major phylogenetic grouping. Counts are next to bars (ranks in parentheses). Rank is assigned by descending orthologue count.

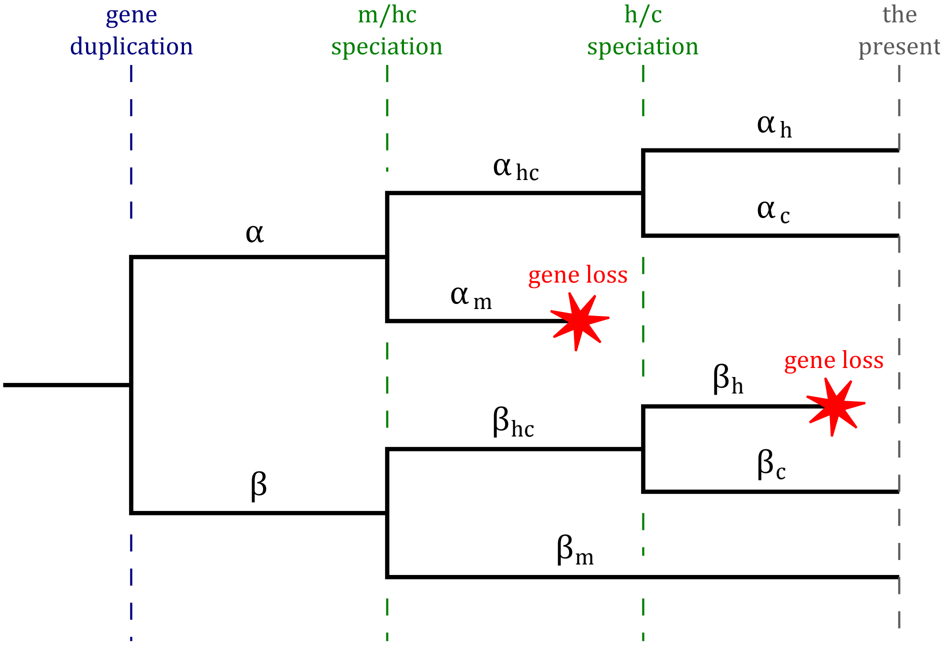


Figure 2.4: Evolutionary history of a hypothetical gene in primates

Chapter 3

A phylogeny of eukaryotes

3.1 Aims: Building a new phylogeny of fully sequenced eukaryotes

We sought to infer a phylogeny of all the organisms in our BLAST study. By building such a tree of relatedness between these disparate eukaryotic organisms, we aimed to facilitate the subsequent analysis of each protein (or, more correctly, each set of orthologous proteins[1]) for signs of correlated evolution. Using the patterns of protein presence/absence whose derivation was described in Chapter 2, it is possible to apply statistical techniques to correct these patterns for phylogeny [Freckleton et al., 2002], and thereby to go on to infer how likely it is that a given pairing of orthologous sets (proteins) shows evidence of a codependency. When it comes to ancestral state reconstruction [Omland, 1999], coincident points of protein gain and loss testify to such relationships [Barker and Pagel, 2005]. Multiple coincident points of loss (by which we mean two or more instances of two given proteins being lost on the same branch of the phylogeny) are generally an indicator of correlation. This will be discussed at greater length in Chapter 4; suffice for now to say that phylogenetic inference is a necessary precursor to inference of correlated evolution.

The process of phylogenetic inference, performed from scratch using a modern implementation of Bayesian statistical methods within a Markov Chain Monte Carlo (henceforth referred to as MCMC) framework [Pagel and Meade, 2004], is itself revealing, potentially offering insights into larger questions in evolutionary biology such as the correct placement of the root of the eukaryotic tree, which is still a matter of contention [Lasek-Nesselquist and Gogarten, 2013]. The correct

[1] An orthologus set is a multi-species set of homologues (i.e. orthologues) to a reference human protein sequence, where each sequence in the set is uniquely defined for its particular species. Few human protein sequences have orthologues in all the other species studied here, so these sets typically come from a subset of the species in the study; only rarely is the set complete, in the sense of having an entry for every species studied.

placement of major monophyletic branches within the eukaryotes is also a matter of ongoing debate, probably thanks to rapid adaptive radiation close to the basal eukaryotic divergence [Baldauf, 2008].

I will present not just a single tree, but a series, including a consensus tree based on multiple concatenated sequences. Trees based on individual genes were generated but are not included here as they were not used as inputs to the subsequent trait analysis.

3.2 Methods: Gene choice, MCMC and the consensus tree

The first step towards deriving a tree was to produce a set of sequence alignments that could serve as the inputs to a phylogenetic inference program; such programs are generally fairly similar in their data requirements, if not necessarily in their algorithmic implementation. In principle, sequences either of peptides or of nucleic acids could be used, although certain programs specifically require one type or the other. In either case, the first requirement is a set of genes or proteins which are common to all the species in the investigation; as discussed in Chapter 2, we want the gene tree to match the species tree if the true set of relationships among the species is to be computed reliably.

Because this study encompasses such a diversity of organisms, the quest for a reasonably sizeable set of genes which are common to all of them might appear ambitious. However, based on the full set of species which were originally intended for inclusion in the phylogeny, I isolated 15 proteins from the reciprocal BLAST which appeared to be common to all 85 species, without exception (that is, their orthologous sets spanned all 85 species). These were judged to constitute a sufficient quantity of data that we might have some confidence in the phylogeny that would be inferred from them. The human protein descriptions (from NCBI Entrez) for these proteins are displayed below.

Proteins found to be common to all 85 species, based on reciprocal BLAST results

chaperonin containing TCP1, subunit 4 (delta)
eukaryotic translation initiation factor 5
proteasome 26S subunit, non-ATPase 14
proteasome beta 3 subunit
ribosomal protein L9
ribosomal protein L13
ribosomal protein L15
ribosomal protein L17
ribosomal protein L18a
ribosomal protein L21
ribosomal protein L23
ribosomal protein S2
ribosomal protein S3a
ribosomal protein S6
ribosomal protein SA

I planned to use the program BayesPhylogenies [Pagel and Meade, 2006] to infer the phylogeny based on all of the above proteins; this would require first performing a separate multiple alignment on each of the fifteen orthologous protein sets, and then aligning the corresponding genes based on the protein alignment. Before these steps were undertaken, however, it was decided to remove 10 species from the analysis. Because 10 species from the genus *Drosophila* (fruit flies) had complete proteomes in RefSeq, I had included all of them in the reciprocal BLAST analysis, where the clustered character of their results extends beyond the closeness of the gross orthologue totals reported in Fig. 2.3, to a discernible similarity in their specific patterns of presence and absence. For example, the average number of human protein orthologues found in the 10 Drosophila species was 5 586.3, with a sample standard deviation of 72.1. *D. melanogaster* had the most orthologues with 5 727 and *D. persimilis* the fewest, with 5 439. A full 3 893 of these orthologues were found in all 10 species. Now consider the effect of including just two more species forming a sister clade to the drosophilae, namely the mosquitoes *Aedes aegypti* and *Anopheles gambiae*, with 5 613 and 5 547 orthologues respectively (i.e. both well within one standard deviation of the *Drosophila* mean): the number of orthologues common to all 12 species falls by 616 to 3 277. By contrast, cutting out any two of the drosophilae, to leave 8 from the original 10, increases the number of orthologues common to all eight by a maximum of 286 to 4 179, if *D. persimilis* and *D. sechellia* are taken out; the figure is lower is any other pair of species is removed[2]. This testifies to the comparative closeness of the relationships among the drosophilae. Because it was my intention to build a broad phylogeny of eukaryotes which would bring out the relationships between distantly related organisms, I was concerned not to bias the inference unduly by weighting the input data heavily towards a large clade of organisms with a comparatively recent common ancestor, which in the case of the drosophilae is reckoned to have existed around 60 million years ago[3]. Another consideration is the nature of the genes used for phylogenetic inference; because they are present in every organism in the study and therefore by definition highly conserved, it is likely that their range of sensitivity is better attuned towards accurate elucidation of deeper phylogenies. These considerations led me to eliminate 9 out of the 10 *Drosophila* species from the study before the alignment stage. *D. melanogaster* was retained as the best-studied and best-annotated representative of the genus.

Unfortunately, and ironically given that much previous work in this area has focussed on its proteome, it was also necessary to eliminate *Saccharomyces cerevisiae* at this stage, because the enhanced detail in its genomic annotation in RefSeq made it difficult to align, and would have required extra programming. With the drosophilae, this accounts for the 10 species excluded from the study

[2] The overall average for Drosophila of 5 586 orthologues is relatively high because they are concentrated at the ends of the distribution. By this is meant that while a majority of the 7 492 orthologues found in these fruit flies are common to all 10 species, the largest minority of orthologues (867) are exclusive to one species only, i.e. to any single species within the clade of drosophilae.

[3] The phylogeny does include closer relationships than this, particularly among the three primates. The closest relationship in the study, between humans and chimpanzees, reflects very recent common ancestry of about 5 million years ago; this was felt to be potentially illuminating given that the human proteome is our starting point. The primate clade, including the macaque, is unlikely to bias the result unduly because of the very few species involved. Common ancestry among the mammals as a whole, even excluding marsupials, stretches back considerably further than 60 million years.

at this stage. This left 75 species, which formed the subject of the phylogenetic inference.

My chosen program for performing the phylogenetic inference was BayesPhylogenies, which requires a multiple alignment of nucleotides. I therefore wrote a Perl program which re-ran the reciprocal BLAST for the fifteen ubiquitous proteins only[4], before retrieving the required 15 sets of 75 protein (amino acid) sequences from RefSeq, together with their coding mRNA (equivalent to sense DNA) sequences. This was facilitated by BioPerl's interface to RefSeq, Bio::DB::RefSeq, which makes it possible to code such operations quite ergonomically; coding sequence accessions were retrieved from protein sequence annotations before both the protein and the coding sequences themselves were automatically downloaded in the required FASTA format. The program stripped any starting methionines from the protein sequences at this point; methionine is employed as the eukaryotic start codon. The protein sequences must be matched to their coding (nucleotide) sequences prior to producing a nucleotide alignment. Because there is no start codon in the curated nucleotide sequences, to have left it in the protein sequences would have caused the matches between protein and nucleotide sequences to fail.

Separate multiple alignments were performed on all 15 sets of protein sequences, using the program MUSCLE [Edgar, 2004]; this was preferred to the older and more widely used ClustalW [Thomopson et al., 1994] on the grounds of the former's claimed superior performance and accuracy [Edgar and Batzoglou, 2006]. This produced fifteen files corresponding to the alignments.

I used the program AAtoDNA, developed by Andrew Meade of the University of Reading, to match the aligned protein sequences, i.e. including gaps, to the nucleotides, outputting a set of 15 multiple nucleotide alignments. By default, this program uses a standard eukaryotic genetic coding, which takes no account of coding modifications which have occurred along some branches of the eukaryotic tree. In this study, two organisms use one such alternative coding and another uses a third distinct coding. *Paramecium tetraurelia* and *Tetrahymena thermophila* transcribe the RNA codons UGA and UGG to glutamine; both are stop codons in the standard genetic code. *Pichia stipitis* transcribes CUG to serine, whereas in the standard code it makes leucine. I therefore had to modify AAtoDNA's source code and recompile to generate different implementations accounting for these alternative codings.

It was then necessary to split each of the 15 protein alignment files into three, one for each coding; because AAtoDNA requires a protein alignment plus a set of RNA sequences, the same split was applied to the raw RNA FASTA files. So with three codings, 6 files were needed per gene; I wrote a short program to do this automatically, plus a shell script to call the various incarnations of

[4] This did not take very long in terms of processing time; otherwise it would have been appropriate to store all the orthologues computed during the complete reciprocal BLAST performed earlier, rather than retaining just the binary presence/absence score. However, see the Discussion section at the end of this chapter.

AAtoDNA with appropriate inputs. This produced three separate nucleotide alignments per gene, which were then appended to give one alignment per gene. At this stage, one of the genes, ribosomal protein S3a, was excluded from the study, because it was not possible to match it to its RNA sequence for every species. Subsequent manual investigation of the raw nucleotide file using BioEdit [Hall, 2005] revealed a probable annotational insertion of a single base in the coding for one species only, throwing the matching algorithm out. BioEdit could have been used to correct this had it been noticed at an earlier stage; by this point, however, the phylogeny had already been inferred.

Not all the positions in such alignments are of equal value to us. In general, the third base in each codon carries little phylogenetic signal; because of the degeneracy of the code, most changes at the third codon position do not produce a change in the amino acid residue. Consider: there are $4^3 = 64$ possible codons, any one of which might constitute a "ground state" prior to a substitution[5] involving one of its three bases. So there are 64 possible pre-substitution ground states. Any one of the bases within the ground state codon can undergo a substitution, and can therefore assume one of three possible states after the substitution (assuming we are excluding trivial substitutions such as $\mathrm{A} \rightarrow \mathrm{A}$ or $\mathrm{C} \rightarrow \mathrm{C}$). So multiplying the number of possible ground states by the number of post-substitution states, at any given base within the codon, gives $64 \times 3 = 192$ possible base changes in any one of the three positions[6], some of which are substitutions to or from stop codons. If we exclude substitutions involving stop codons, we are left with 174 possible changes at the first base position within a codon, of which all but 8 are non-synonymous, i.e. they effect a change to the amino acid residue. At the second base, of 176 possible substitutions not involving stop codons, none is synonymous. By contrast, only 50 of the possible 176 substitutions at the third base are non-synonymous; it is therefore likely that purifying selection will have a greatly reduced effect in constraining rates of evolution at this position. We should therefore expect it to evolve much more rapidly than the first and second positions. This simply means that it will undergo more state changes over time than the first two positions, which is why its phylogenetic signal becomes swamped in noise. Therefore, because not only would the retention in the alignments of the third base be of marginal if any value, but would certainly impose an extra processing burden, I wrote a Perl program to strip it from the RNA alignments.

The final stage in preparing the alignments is to remove any stretches of sequence which are obviously of no value, and which could potentially confound the phylogenetic inference, possibly by

[5] In this context, I have preferred the term "substitution" over "transition" because the former refers unambiguously to a *general* state transition, which encompasses any change in a base, and includes purine-pyrimidine "transversions" as well as purine-purine and pyrimidine-pyrimidine "transitions".

[6] Of course, taking all possible ground states and all possible substitutions at *any* of the 3 bases within the codon gives us $64 \times 3 \times 3 = 576$ possible state changes which could potentially arise at a codon (which can assume any ground state) as the result of a single substitution event.

virtue of their very length. Many genes have long insertions in relatively few species, typically but not exclusively at the beginnings and ends of sequences, which means that the sequence data at that position will just be long gaps for most species. Ideally, I wanted to infer the phylogeny based on alignments with as few gaps as possible. To this end I wrote another short program, employing the simplest possible alignment cleaning algorithm: it snips from the start of the alignment to the start of the second longest trailing[7] start and from the second longest trailing end to the end of the alignment. Thus trailing starts and ends, where there are data for one species only, are excised. Note that using this algorithm, if at least two sequences were to run right from the start of the alignment, it would not be shortened at the beginning at all; similarly, if at least two sequences run to the end of the alignment, it will not be truncated. Internal sections of sequence are not affected at all; a long (or short) insertion in a single sequence would not have been removed by the program described.

While more sophisticated means of preparing the alignments could have been employed, by convention these would be manual, using a program like BioEdit to remove unwanted sections of sequence by hand. Because part of my aim was to develop a process which was fully replicable, I felt that it was important to minimise such manual interventions, without of course overlooking any glaring incongruence in the data. No such incongruence was apparent to me in viewing these alignments; they were in fact checked in BioEdit, but not altered manually.

Another option would have been to rationalise the alignments using the program Gblocks [Castresana, 2000]. This has the advantage of at least being fully replicable and was in fact attempted, but the program was found to have such conservative defaults (for example, deleting all locations at which a single gap was found in any species) as to reduce the length of the alignment by over 70%. Although the parameters could have been modified, it was nonetheless felt, given the highly conserved nature of the sequences chosen, that the alignments were of sufficient quality to serve as the inputs to the phylogenetic inference without such intricate tinkering. Also, quite apart from its stringency, Gblocks seemed unnecessarily complicated and sufficiently obscure in its implementation that true replication of any method employing it would be possible only with the program itself to hand: effectively, Gblocks is a “black box”.

The final 14-gene alignments, concatenated together, were 19 021 nucleotides long. Thus prepared, they were converted from FASTA to NEXUS format, ready for phylogenetic inference. The inference program BayesPhylogenies was used to run a Markov chain of 100 000 iterations of trees.

In case the reader is interested in implementing the protocols above for him/herself, I have included

[7] The second longest trailing start (or end), or the very longest for that matter, might not be “trailing” at all, in the sense of extending beyond the bulk of the other sequences in the alignment; all, most or at least more than a few of them might start or end in exactly the same place.

some technical notes in Appendix B, relating to configuring and running the alignments.

3.3 Results: Trees

The overall analysis using all 14 genes for all 75 species yielded some interesting results. The tree did not settle on a single topology; after rooting, the two main alternative topologies became very evident visually and are illustrated below. Sample trees were taken every 10^4 iterations of the chain, which ran for just over 9.964×10^7 iterations in total; the last 5 000 samples only were used as inputs to the consensus trees. This was more than sufficient time for burn-in, so the phylogeny was at convergence. The first tree topology, shown in Fig. 3.1 on the next page, was observed in 4 082 of the 5 000 samples.

The other relatively common topology, observed in 936 out of the 5 000 samples, is illustrated by way of contrast in Fig. 3.2 on page 53.

The consensus trees' branch lengths are derived by averaging the branch lengths from the individual trees in the sample of 5 000, for those trees in which that branch exists. Typically, nodal support values (not pictured in the trees above) in a consensus tree refer not to the likelihood but to the percentage of trees in the original sample in which that node is found; the nodes that make their way into the consensus tree are simply those which are observed most frequently in the sample. Each consensus is relevant for one root only, i.e. all trees contributing to the consensus must possess the same root node. Therefore, strictly speaking, nodal support values refer to the percentage of trees with the given node, out of those in the sub-sample having the same root. If, as here, the phylogenetic inference has not succeeded in resolving the root node[8] unequivocally, it may be that certain branches are given to skipping around the tree with high likelihood of being retained for many iterations of the chain.

Observe the effect of removing three species from the inference, demonstrated in Fig. 3.3 on page 54. In fact, this figure is a consensus tree of all 5 000 samples with *Dictyostelium discoideum* AX4, *Encephalitozoon cuniculi* GB-M1 and *Entamoeba histolytica* HM-1:IMSS removed. Without these species, one of which (*E. cuniculi*) has a noticeably long branch in both trees shown above, the root node stabilises. When only those 72 species which will feature in the likelihood analysis remain in the phylogeny, all 5 000 trees in the sample share the same root node; therefore the consensus tree for the 72 species is derived from all 5 000 trees. In one of the above trees, the two other species

[8] The tree was rooted manually; by saying that its root was "unresolved", I mean that the node chosen for rooting is not topologically valid in the all the trees in the sample. The presence of *E. cuniculi* on different sides of the "root" in different samples illustrates this.

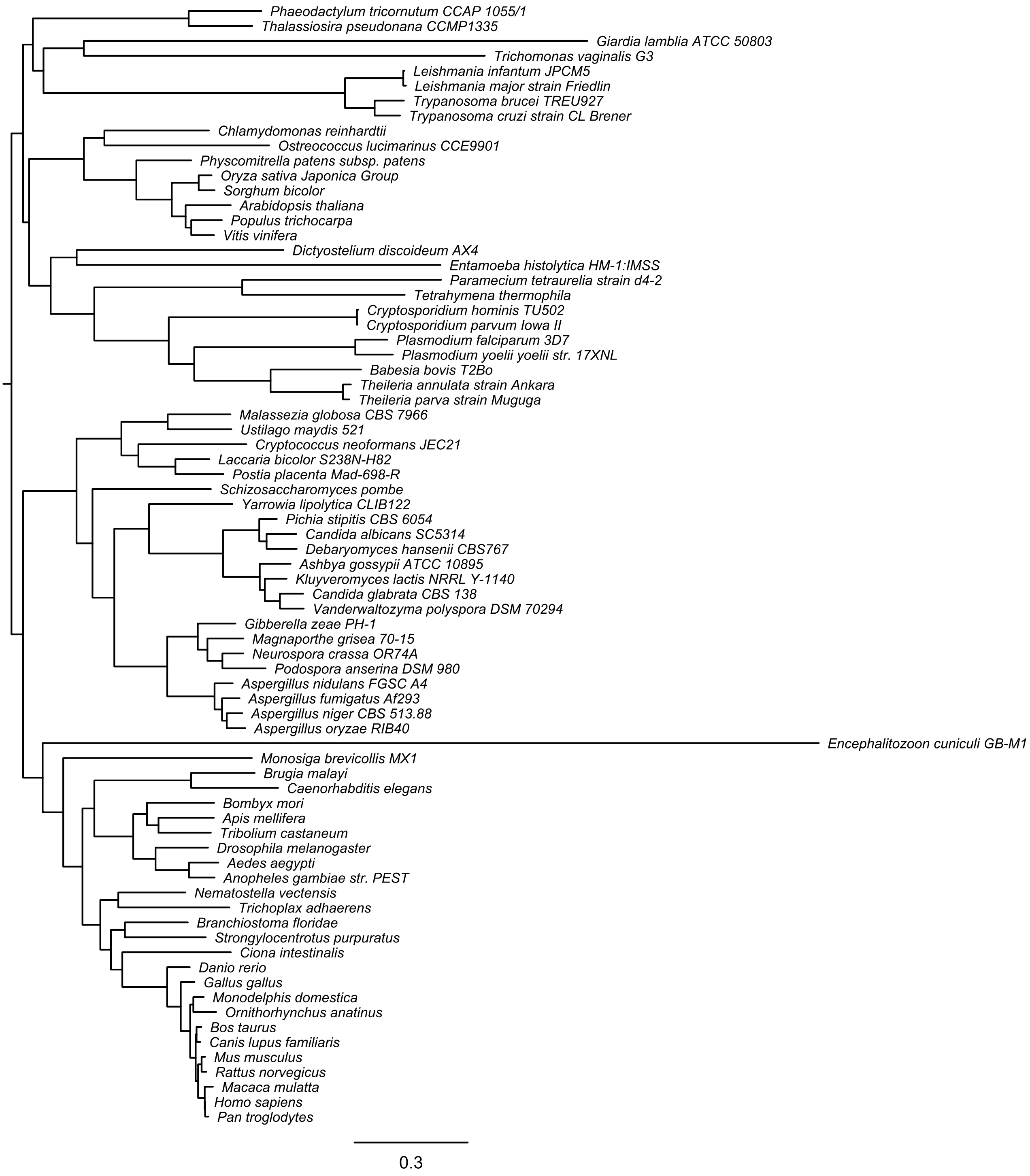

Phaeodactylum tricornutum CCAP 1055/1
Thalassiosira pseudonana CCMP1335
Giardia lamblia ATCC 50803
Trichomonas vaginalis G3
Leishmania infantum JPCM5
Leishmania major strain Friedlin
Trypanosoma brucei TREU927
Trypanosoma cruzi strain CL Brener
Chlamydomonas reinhardtii
Ostreococcus lucimarinus CCE9901
Physcomitrella patens subsp. patens
Oryza sativa Japonica Group
Sorghum bicolor
Arabidopsis thaliana
Populus trichocarpa
Vitis vinifera
Dictyostelium discoideum AX4
Entamoeba histolytica HM-1:IMSS
Paramecium tetraurelia strain d4-2
Tetrahymena thermophila
Cryptosporidium hominis TU502
Cryptosporidium parvum Iowa II
Plasmodium falciparum 3D7
Plasmodium yoelii yoelii str. 17XNL
Babesia bovis T2Bo
Theileria annulata strain Ankara
Theileria parva strain Muguga
Malassezia globosa CBS 7966
Ustilago maydis 521
Cryptococcus neoformans JEC21
Laccaria bicolor S238N-H82
Postia placenta Mad-698-R
Schizosaccharomyces pombe
Yarrowia lipolytica CLIB122
Pichia stipitis CBS 6054
Candida albicans SC5314
Debaryomyces hansenii CBS767
Ashbya gossypii ATCC 10895
Kluyveromyces lactis NRRL Y-1140
Candida glabrata CBS 138
Vanderwaltozyma polyspora DSM 70294
Gibberella zeae PH-1
Magnaporthe grisea 70-15
Neurospora crassa OR74A
Podospora anserina DSM 980
Aspergillus nidulans FGSC A4
Aspergillus fumigatus Af293
Aspergillus niger CBS 513.88
Aspergillus oryzae RIB40
Encephalitozoon cuniculi GB-M1
Monosiga brevicollis MX1
Brugia malayi
Caenorhabditis elegans
Bombyx mori
Apis mellifera
Tribolium castaneum
Drosophila melanogaster
Aedes aegypti
Anopheles gambiae str. PEST
Nematostella vectensis
Trichoplax adhaerens
Branchiostoma floridae
Strongylocentrotus purpuratus
Ciona intestinalis
Danio rerio
Gallus gallus
Monodelphis domestica
Ornithorhynchus anatinus
Bos taurus
Canis lupus familiaris
Mus musculus
Rattus norvegicus
Macaca mulatta
Homo sapiens
Pan troglodytes
0.3

Figure 3.1: Consensus tree for the most common root in a phylogeny of 75 eukaryotes

which were removed cluster closely around the long branch. More on this later; see this chapter's Discussion section. Support values for the 72-species consensus tree were 100% (i.e. of trees in the sample) for most nodes, with exceptions highlighted in dark blue in Fig. 3.3 on page 54. In all cases, nodal support was 60% or higher. Note that even where nodal support is significantly below 100%, these values are not likelihoods; the individual likelihood-based support values in those trees in

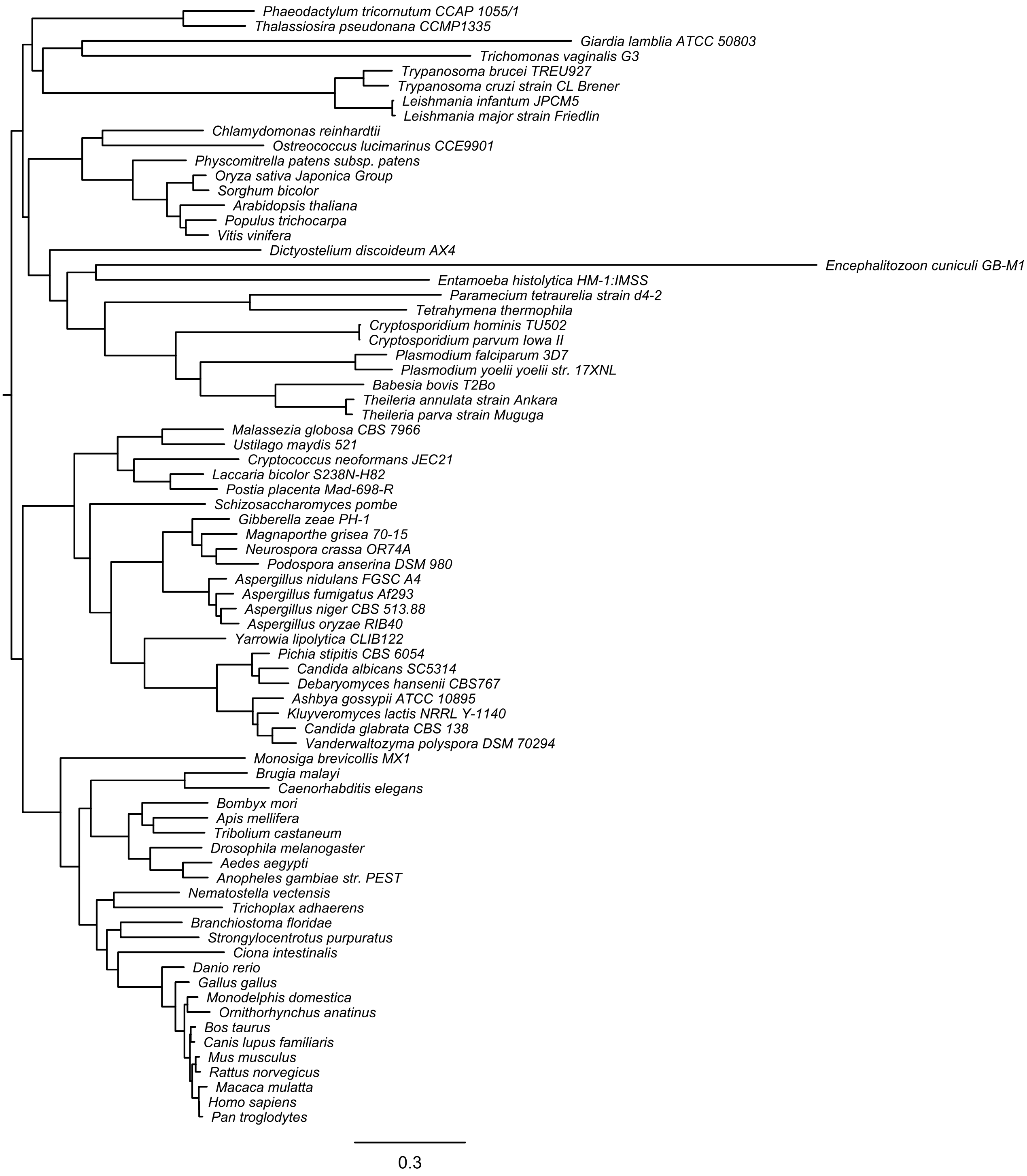

Phaeodactylum tricornutum CCAP 1055/1
Thalassiosira pseudonana CCMP1335
Giardia lamblia ATCC 50803
Trichomonas vaginalis G3
Trypanosoma brucei TREU927
Trypanosoma cruzi strain CL Brener
Leishmania infantum JPCM5
Leishmania major strain Friedlin
Chlamydomonas reinhardtii
Ostreococcus lucimarinus CCE9901
Physcomitrella patens subsp. patens
Oryza sativa Japonica Group
Sorghum bicolor
Arabidopsis thaliana
Populus trichocarpa
Vitis vinifera
Dictyostelium discoideum AX4
Encephalitozoon cuniculi GB-M1
Entamoeba histolytica HM-1:IMSS
Paramecium tetraurelia strain d4-2
Tetrahymena thermophila
Cryptosporidium hominis TU502
Cryptosporidium parvum Iowa II
Plasmodium falciparum 3D7
Plasmodium yoelii yoelii str. 17XNL
Babesia bovis T2Bo
Theileria annulata strain Ankara
Theileria parva strain Muguga
Malassezia globosa CBS 7966
Ustilago maydis 521
Cryptococcus neoformans JEC21
Laccaria bicolor S238N-H82
Postia placenta Mad-698-R
Schizosaccharomyces pombe
Gibberella zeae PH-1
Magnaporthe grisea 70-15
Neurospora crassa OR74A
Podospora anserina DSM 980
Aspergillus nidulans FGSC A4
Aspergillus fumigatus Af293
Aspergillus niger CBS 513.88
Aspergillus oryzae RIB40
Yarrowia lipolytica CLIB122
Pichia stipitis CBS 6054
Candida albicans SC5314
Debaryomyces hansenii CBS767
Ashbya gossypii ATCC 10895
Kluyveromyces lactis NRRL Y-1140
Candida glabrata CBS 138
Vanderwaltozyma polyspora DSM 70294
Monosiga brevicollis MX1
Brugia malayi
Caenorhabditis elegans
Bombyx mori
Apis mellifera
Tribolium castaneum
Drosophila melanogaster
Aedes aegypti
Anopheles gambiae str. PEST
Nematostella vectensis
Trichoplax adhaerens
Branchiostoma floridae
Strongylocentrotus purpuratus
Ciona intestinalis
Danio rerio
Gallus gallus
Monodelphis domestica
Ornithorhynchus anatinus
Bos taurus
Canis lupus familiaris
Mus musculus
Rattus norvegicus
Macaca mulatta
Homo sapiens
Pan troglodytes
0.3

Figure 3.2: Consensus tree for the second most common root in a phylogeny of 75 eukaryotes

the sample which do not contain a node (i.e. one which is present in the majority of trees) tend to be lower for alternative nodes which they contain than those for the majority of nodes. So from a likelihood perspective, it could be argued that consensus-based nodal support values are actually significantly understated. Whatever the true interpretation, the tree in Fig. 3.3 on the next page was the one I went on to use in the likelihood analysis of protein pairs, described in Chapter 4.

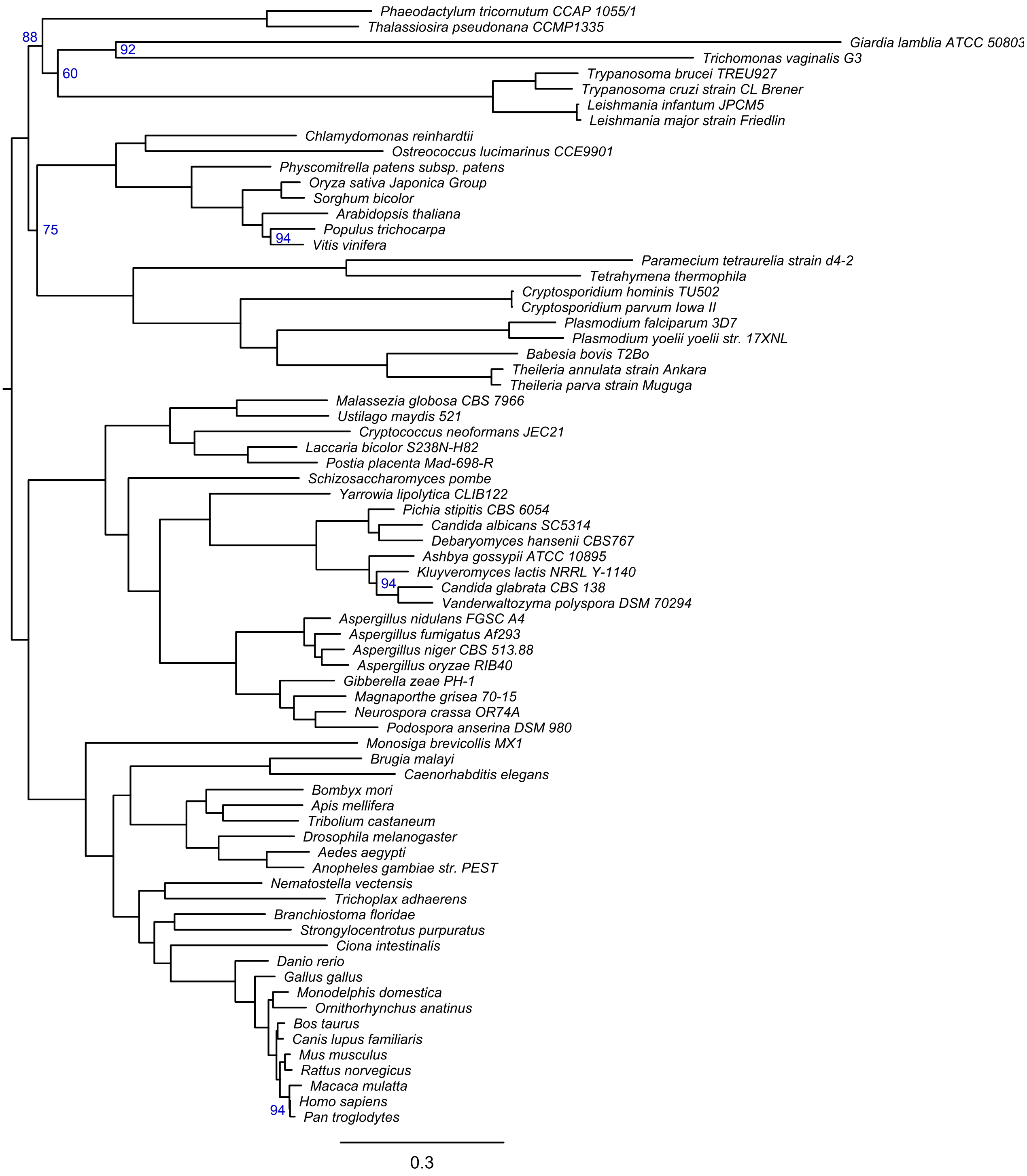

Phaeodactylum tricornutum CCAP 1055/1
Thalassiosira pseudonana CCMP1335
Giardia lamblia ATCC 50803
Trichomonas vaginalis G3
Trypanosoma brucei TREU927
Trypanosoma cruzi strain CL Brener
Leishmania infantum JPCM5
Leishmania major strain Friedlin
Chlamydomonas reinhardtii
Ostreococcus lucimarinus CCE9901
Physcomitrella patens subsp. patens
Oryza sativa Japonica Group
Sorghum bicolor
Arabidopsis thaliana
Populus trichocarpa
Vitis vinifera
Paramecium tetraurelia strain d4-2
Tetrahymena thermophila
Cryptosporidium hominis TU502
Cryptosporidium parvum Iowa II
Plasmodium falciparum 3D7
Plasmodium yoelii yoelii str. 17XNL
Babesia bovis T2Bo
Theileria annulata strain Ankara
Theileria parva strain Muguga
Malassezia globosa CBS 7966
Ustilago maydis 521
Cryptococcus neoformans JEC21
Laccaria bicolor S238N-H82
Postia placenta Mad-698-R
Schizosaccharomyces pombe
Yarrowia lipolytica CLIB122
Pichia stipitis CBS 6054
Candida albicans SC5314
Debaryomyces hansenii CBS767
Ashbya gossypii ATCC 10895
Kluyveromyces lactis NRRL Y-1140
Candida glabrata CBS 138
Vanderwaltozyma polyspora DSM 70294
Aspergillus nidulans FGSC A4
Aspergillus fumigatus Af293
Aspergillus niger CBS 513.88
Aspergillus oryzae RIB40
Gibberella zeae PH-1
Magnaporthe grisea 70-15
Neurospora crassa OR74A
Podospora anserina DSM 980
Monosiga brevicollis MX1
Brugia malayi
Caenorhabditis elegans
Bombyx mori
Apis mellifera
Tribolium castaneum
Drosophila melanogaster
Aedes aegypti
Anopheles gambiae str. PEST
Nematostella vectensis
Trichoplax adhaerens
Branchiostoma floridae
Strongylocentrotus purpuratus
Ciona intestinalis
Danio rerio
Gallus gallus
Monodelphis domestica
Ornithorhynchus anatinus
Bos taurus
Canis lupus familiaris
Mus musculus
Rattus norvegicus
Macaca mulatta
Homo sapiens
Pan troglodytes
88
92
60
75
94
94
94
0.3

Figure 3.3: Consensus tree for a phylogeny of 72 eukaryotes

3.4 Discussion

A word should be said about rooting. A phylogenetic tree provides a topological representation of species' divergence one from another, but the root of the tree is not defined: it has to be assigned based on prior knowledge. Internal nodes within the tree (as opposed to the tips, which represent extant species) describe points where a speciation event occurred, i.e. where one lineage diverged into two. An internal node is a last common ancestor of the two branches which emerge from it; one branch only goes into it. So in network terms, an internal node has one in-link and two out-links. In seeking the root however, we are also seeking another last common ancestor—in this case the ancestor of every extant species (tip) and of all the defined internal nodes (last common ancestors) as well. The root is undefined if we consider a phylogenetic inference as a whole, because we cannot assume rates of evolution to be constant. Consider the simple tree relating three extant species. Toplogically, there will be one internal node representing a point where the three branches leading to the tips join together. But if our species are A, B and C, how do we know whether the A+B lineage split from the C lineage before A and B themselves split, or whether A+C split from B, or B+C split from A? If A and B split recently, but B has experienced relatively rapid molecular evolution since the split, it might look like A and C split more recently if we assumed that time was closely correlated with the amount of molecular evolution. It may be, but not necessarily. How, under circumstances like this, can we assign a position for the root and thereby infer the correct set of relationships between A, B and C? A standard procedure is to use an outgroup. An outgroup is a related species (or the relevant set of genes or proteins from such a species) which is known to have a last common ancestor (i.e. with the group of species in the study) which predates any of the last common ancestors shared between the species in the study themsleves. Thus one might include a set of bacterial orthologues to the eukaryotic genes used to infer a phylogeny in a particular study. We know that the bacterium is more distantly related to all the eukaryotic species (which are our actual focus of interest) than the eukaryotes are to each other, so having derived an unrooted tree using our preferred inferential method, we look to see where the node that joins the bacterium to the eukaryotes attaches, i.e. to which two (probably but not necesssarily internal) nodes in the eukaryotic part of the tree is it adjacent? If we can pinpoint that with confidence, we have successfully rooted our tree. The two adjacent nodes represent the immediate descendents of the first speciation event, before the two derived lineages themselves underwent speciation, if indeed they both have.

Attempts to root a tree in the way described above can be confounded in distinct ways. Firstly, because the outgroup will by defintion be distantly related to the group of species in which we are really interested, does it share with those species an orthologue of the gene(s) or protein(s) used to infer the phylogeny, at all? The orthologue might not be conserved across such a large phylogenetic divide, which might mean that we have to restrict the sequences used for the phylogenetic inference, just to accommodate the need to include an outgroup. Another consideration of particular relevance to the root of the eukaryotic tree is the very short branch lengths close to the root, during what is assumed to be a time of rapid speciation and adaptive radiation, when the major phyla still present today quickly became resolved. With such short branches, the confidence associated with the correct topological placement of a root inferred using an outgroup will tend to be quite low, and there has indeed been much debate about the correct position of the true root of the eukaryotes. I took professor Sandra Baldauf's advice in assigning a root to my trees.

Note that for the root of the tree to be "resolved", the nodes adjacent to the root must be common ancestors each to a consistent set of descendents (terminal nodes); if different runs of the phylogenetic inference generate trees with different adjacent nodes, then the branch which, in the absence of the root, joins those two nodes together, cannot be resolved if both its endpoints are not themselves resolved. And by defintion, if one node adjacent to the root is unresolved, so is the other, because it means that at least one terminal node must have "swapped" from one side to the other of the first bifurcation in the tree[9].

Because the set of genes which were found to be common among all species in the study was relatively poor, both in terms of breadth and overall quantity, an alternative approach would have been to look for genes which are not of necessity common to every species in the study, but merely to all but perhaps one or two, and to derive a complementary set wherein species missing from any one gene's compass are nonetheless included within the ambit of most or all of the other genes in the set. Thus every species could be adequately covered, but there could be bits missing. Because the sort of multiple alignment of multiple non-contiguous sequences (i.e. of a set of intragenomically distant sequences) which can be used as input for computational phylogenetic inference is typically no more than the set of multiple alignments of the individual genes or proteins, concatenated end to end, it will contain gaps. The gaps represent genetic insertions and deletions. If an entire sequence, gene or protein, is missing, it can therefore just be treated as a large gap if it is part of a much longer alignment.

By treating missing genes as gaps in this way it would undoubtedly have been possible to expand the number of genes used in the alignment. This would certainly have been desirable from the perspective that the value of the small set of 14 genes that were used would have been enhanced had they come from a more functionally diverse set of backgrounds. Because a full 10 of our proteins are ribosomal, they hardly constitute a particularly representative sample; from a phylogenetic perspective, they are not "independent" points. In fact, the prevalence of ribosomal genes here is not a surprise, as ribosomal function is highly constrained; this is true not only of the structurally significant stretches of rRNA, but also of the polypeptide constituents of the ribosome which are our focus here. As stated in Section 3.2, had it been my original intention to infer the phylogeny from only the 72 species in the final likelihood analysis, and had time been available to correct one of the alignments manually, not just 14 but all 19 genes found to be present in every species could have been used; 6 out of these 19 genes were non-ribosomal. To go further, of the 33 genes which

[9] Possibly, the rearrangement might involve too many species for a notion like "swapping" to have much meaning; the point is that the topology has changed, at the root.

were found to be present in 71 out of the 72 species, 8 were absent from *Postia placenta* Mad-698-R, 2 each from *Aspergillus oryzae* RIB40, *Nematostella vectensis* and *Ornithorhynchus anatinus*, and 6 from *Bombyx mori* (which is almost certainly artefactual in that its proteome was incomplete and therefore included in the study in error, as already stated). Regardless of errors introduced by the last species, this still leaves 13 proteins which were absent from species in which they were the sole absentee orthologue. So adding these 13 proteins to the 19 which were ubiquitous within the species in the study would have introduced only 13 gaps, none of which would have been aligned with each other. Of these 13 genes, 5 were non-ribosomal, making 11 non-ribosomal genes in total. To broaden the scope of the data, by including the total of 7 new non-ribosomal genes (i.e. 2 extra genes from corrected alignments, plus the 5 genes which were absent in one species only), would be very desirable if inferring the phylogeny again.

Although this study would have benefited from increasing the diversity among the genes used, it is more open to question whether simply increasing the quantity of genes (and therefore the length of the alignment) would of itself have improved the reliability of the phylogenetic reconstruction. Current models of sequence evolution arguably deal well with sequences of up to several thousand base pairs; the complexity of these models reflects the complexity inherent in that order of quantity of data. Such complexity in the models is a function not only of the complexity of the data but also of the computing power available to developers of the models. In recent years, despite ongoing technological innovation in computing, the trend has been for the growth in the quantity of publicly available sequence data to outstrip the growth in average processing power. It is possible therefore that the available models, including the mixture model used here, are underspecified in relation to the complexity of the data. It is in turn possible that this explains why one artefactual consequence of using increased quantities of data is that the level of confidence in a possibly incorrect result becomes elevated. In phylogenetic terms, this can mean that nodal support is artificially high; this has been demonstrated in simulation studies performed by the Evolutionary Biology Group at the University of Reading. At this stage I offer the cautious conclusion that it would have been better to use a few more genes from a much broader range of functional categories, using alignments with whole gene gaps where necessary. Functional categorisation could have been derived from the Gene Ontology.

Chapter 4

A correlated evolution network

4.1 Aims: Network inference based on correlated protein evolution

In this study I aimed to undertake a large-scale comparative analysis that treats protein presence/absence as a binary trait. This allows us to select any pair of proteins (orthologous sets) based on the reference human proteome and perform statistical tests to establish whether the two proteins' evolutionary histories are correlated. Critically, these tests employ a phylogenetic correction, so that our species are not treated as independent data points as per the simplifying, and simplistic, across-species method. Species which are closely related are likely to have traits in common merely by virtue of evolutionary proximity. Because the signifiers of correlated evolution are shared patterns of protein gain and loss on the (in our case) eukaryotic tree, and because these patterns are historical and can only be inferred from existing data rather than observed directly, correlations in patterns of presence/absence between widely diverged species should be afforded more significance, or weight, than those seen in close relatives; this is precisely what the incorporation of a phylogeny allows us to do. Therefore, the binary trait matrix described in Chapter 2, plus the phylogeny described in Chapter 3, form the joint inputs to the likelihood analysis described here.

Correlated evolution between proteins is interesting on the micro level in terms of the clues that it may provide to interdependent function. This alone has many potential implications for diverse fields of research, including medicine. On the macro level, correlated evolution offers pointers to grand questions about the evolution of genetic and organismal modularity, based on the topology of the network described by aggregating the inferred pairwise correlations. In this, the logic of Barker and Pagel [2005], as described in Section 2.1, is again followed closely, but with certain modifications and a greatly expanded set of input data.

4.2 Methods: Likelihood-based correlated evolution on a phylogeny

If correlated evolution is taken to imply a true functional relationship between two proteins, then we should expect to see its signature in curated sets of data in which known functional relationships are recorded. Such known interactions are generally those inferred from direct laboratory analysis of the proteins in question, rather than gross computational investigations such as this one, which, if they are of use in this context at all, typically serve more to give pointers which may then be followed up in the wet laboratory. The kinds of interaction which might show up in correlated evolution include participation in the same metabolic or genetic regulatory pathway, shared roles in some subcellular structure or organelle, and direct physical interaction to perform some function which depends on the participation of both proteins, or possibly of more than two[1].

So I sought a reliably curated database of known protein-protein relationships. Beyond reliability, I was concerned that any such data source should be sufficiently extensive and mature to provide a broad coverage specifically of the human proteome, as that was my starting point in inferring the orthologous sets. Of the types of relationship that could have been chosen, those which fitted my criteria most closely were direct protein-protein interactions. There are several curated online databases of human protein-protein interactions, including BIND, MINT, IntAct and the BIOGRID. Each of these, and particularly the last (because it was a relatively recent initiative and very active in terms of updates), was considered for use as a reference data set against which to measure the effectiveness of the test for correlated evolution. However, in the end, HPRD was chosen as being the best-attested and offering the widest coverage specifically of the human proteome, with some 38 037 curated interactions (not all of them merely pairwise). Although releases of HPRD are relatively infrequent in that it can go more than a year without any new data being added, this was felt to be an advantage in that it seems to reflect the thoroughness of the curation; HPRD entries are manually curated from published research papers. I used HPRD Release 8, which is still current and was released in July 2009.

[1] Interactions involving more than two proteins are not strictly pairwise, of course. There is no reason in principle why a physical interaction between proteins should not require the participation of three or more proteins, the loss of any one of which would destroy the function; multiple interactions of this kind are recorded in HPRD, in fact. However, in this study, we treat all interactions as pairwise. So for simplicity's sake, we would expect a three-way interdependence of the type described to manifest itself as three separate correlations between the three possible pairs (combinations) which can be chosen from the three proteins. The same combinatorial logic applies for larger fully interdependent clusters (physical or otherwise) of proteins. See also this chapter's Discussion section.

The HPRD interaction data were downloaded in the form of a single XML file[2], obtainable from the HPRD website. I wrote small Perl programs employing the CPAN module XML::Twig to parse the XML data into tab-separated-variable format text files, which were then loaded from the command line into a MySQL database, incorporating separate tables for interactors and interactions, plus a cross-referencing table relating one to the other. A separate archive containing a series of flat files[3] was also downloaded from HPRD; one of these flat files, `HPRD_ID_MAPPINGS.txt`, was used to populate another MySQL table, which can map HPRD interactor IDs to protein accessions from RefSeq, among other repositories.

A major part of my concern in the conduct of this likelihood analysis was to go beyond simply performing a raw statistical test on pairs of proteins which we already know (or have very good reason to presume) to be interactors. I wanted to derive a reliable control group of pairs of presumptive non-interactors[4]. More than this, with very high confidence, members of pairs featuring in the control group should have no interdependencies of *any* kind one upon the other, including participation in the same pathways, so that we should expect to see no evidence whatsoever of correlated evolution between the members of any pair in such a set. This is a fairly problematic requirement. While one might reasonably argue that true interdependencies between proteins constitute a small subset of all possible pairwise relationships (of which, given the 31120 human proteins recorded in RefSeq, there are

$$\frac{31120^2 - 31120}{2} = 4.84 \times 10^8 \tag{4.1}$$

or nearly half a billion), this is a long way from being able to say that the absence of a given relationship from any repository of interactions is a reliable indicator that the relationship does not, in fact, exist. A recent study of publicly available repositories of protein interaction data found surprisingly little intersection between the data sets [Turinsky et al., 2010]. Because the set of possible unrecorded interactions is so very much larger than the set of interactions, I have taken this as being likely to mean that the set of as yet unknown interactions is still comparatively poorly explored. For this reason I suggest that we are on much safer ground in trusting that a curated interaction in any publicly available database is real, than we are in assuming that an interaction does not take place simply because it has never been observed (i.e. observed and recorded).

[2] `HPRD_SINGLE_PSIMI_070609.xml`, which contains, among other things, records specific to the interactors, the interactions, and the experiments in which the interactions were recorded.

[3] `HPRD_FLAT_FILES_070609.tar.gz`, which is an archived directory in which are placed thirteen text files containing data of various kinds, mostly in tab-separated-variable format, plus a separate README file.

[4] The same statistical test should be performed on the control group as on the presumptive interactors. If the control group is properly specified, this should yield our preferred null distribution of likelihoods, against which to measure statistical significance correctly in the interactors: more on this later.

How then *should* we attempt to identify pairs of proteins between which no interaction or other form of interdependency exists? The approach I adopted was to use the Gene Ontology [Harris et al., 2004]. The Gene Ontology, or GO, seeks to provide researchers with a way of annotating genes/proteins using a hierarchy of identifiers, all of which fall within one of the three main branches of the ontology: cellular component (effectively localisation), molecular function and biological process. Each main branch ramifies into a series of sub-branches, which become increasingly specialised towards the tips of the ontological tree. Each node in this structure corresponds to a GO ID and can have many child nodes; nodes can also have more than one parent, as long as the relationship to each parent is of a distinct type. Five different relationships are in current use in GO[5]: "is a", "part of", "negatively regulates", "positively regulates" and just "regulates". Take the GO term "apical part of cell" (accession 45 177), for example: it bears the relationship "is a" to GO term "cell part" (accession 44 464), which in turn bears the relationships "is a" to "cellular component" (accession 5 575 – one of the three root nodes) and "part of" to "cell" (accession 5 623), which itself "is a" "cellular component".

It was fortunate for the purposes of this study that one of the flat files provided by HPRD, namely `GENE_ONTOLOGY.txt`, maps HPRD IDs directly to GO accessions, albeit not in a format which lent itself to simple parsing into a database table because, in general, one protein (interactor) maps to multiple GO accessions, with multiple GO accessions being present on each line, wrapped in superfluous text which had to be removed. Compounding the difficulty, individual HPRD IDs can repeat over multiple lines. A custom Perl program was therefore written to parse the mapping into yet another MySQL database table.

Having thus enabled the mapping of IDs to their respective GO terms, it was necessary to download and install the GO database itself. Because the final stage of filtering for our control set of presumptively non-interacting protein pairs requires protein IDs bearing similar but non-identical GO identifiers to be excluded, I needed to be able to traverse the GO hierarchy to identify terms having identical parents or grandparents, for example. Exactly how this was accomplished is explained in more detail later on. The GO database itself is available in the form of an archive[6] which when unpacked gives a set of SQL table definitions, together with tab-separated-variable files ready to be loaded into each separate table. Two short shell scripts were written to automate this process; I defined an empty MySQL database, within which each GO table was first created and then populated in turn.

Having used the flat file to enable the necessary mapping between RefSeq ID and HPRD ID (as described earlier in this section), I was able to rebuild the binary trait matrix against HPRD IDs (including ones without any curated interactions—i.e. many of these proteins do not appear in the

[5] Although more are listed as available within the GO database, only five are actually employed.

[6] Available at `http://archive.geneontology.org/latest-termdb/go_daily-termdb-tables.tar.gz`

interactors table at all). To accomplish this I wrote a Perl program which stored the rebuilt trait matrix in another MySQL table. This was complicated somewhat by the consideration that the already stored matrix had been flushed of duplicate sequences—correctly so, but this meant that some of the original RefSeq IDs had been excised, which would have meant that HPRD IDs mapping to them would have been overlooked. I therefore imported the complete human proteome into a MySQL table and joined it to the original trait matrix on GI and then back to itself on sequence string, in order to pick up the missing RefSeq IDs so that the HPRD trait matrix (itself restricted to unique sequence strings) would not be missing anything. The number of HPRD IDs, at 19 390 as of Release 8, is considerably fewer than the number of RefSeq IDs (or GIs) in the human proteome, possibly because only those proteins of well established function, or properly analysed in the literature, have been curated. Without a fairly comprehensive one-to-one mapping between the two data sets, of which we are clearly short, it seemed very likely that some trait patterns would be missed without matching HPRD against the original, full set of RefSeq proteins.

Having reached this stage, it was possible, using the binary trait matrix and the phylogeny as twin inputs, to run the likehood analyses for all the protein pairs identified by the protocols outlined above. The pairwise analyses were submitted each as separate cluster jobs under the batch scheduler. The likelihood outputs from the linked and unlinked analysis allowed a likelihood ratio to be calculated, i.e. to tell us whether there was any evidence that gene loss events involving the deletion of both genes simultaneously were correlated across the phylogeny (i.e. after applying a phylogenetic correction). I used both 95% and 99% confidence intervals for subsequent stages of this investigation.

4.3 Results: Likelihood ratio distributions

From the 28 337 protein pairs identified as linked by HPRD, and in which the reciprocal BLAST analysis identified both members of the pair as being present in no fewer than 8 species (out of the 72), but no more than 65, giving us a good set of candidates for the likelihood analysis for correlated evolution, I identified 26 207 unique combinations of presence/absence vectors across the species[7]. Taking both partners in the pair, there would have been some redundancy in running the analysis had I not first excised the duplicate combinations, which although constituting less than 10% of the whole, would nonetheless have been costly in terms of computer time. Similarly for the presumptive non-interactors, in my sample of 60 000 pairs, I identified only 48 879 unique combinations

[7] By presence/absence vector, I mean the row in the binary trait matrix, consisting of a string of ones and zeros (72 in total) for a particular protein. Generally, I would expect the proportion of unique combinations of 2 such vectors to be much higher than the proportion of unique vectors among the proteins taken individually. However, see the next comments in the main text.

of presence/absence vectors. The surprisingly many non-unique vector combinations in the latter data are probably attributable to the stringency of my selection criteria for non-interacting pairs. Before charting the likelihood ratio distributions below, I expanded the data set out again, weighting the results to reflect the original numbers of protein pairs represented in each result (i.e. mostly 1 pair per result, sometimes 2 and occasionally more). This was accomplished by joining data tables in MySQL. I then excluded those pairs where, for whatever reason, the likelihood ratios had tended towards infinity before being aborted within BayesTraits [Pagel and Meade, 2007]; my practical criterion was to exclude ratios whose absolute values exceeded 1 000. This left 27 485 curated interacting pairs from HPRD, and 58 280 presumptively non-interacting, non-codependent pairs from my sample. It is possible that excluding these pairs could have been a source of systematic error, if the likelihood calculations "went wrong" in a systematic way. The best way of testing this would have been to run the likelihood analysis again on the excluded set; this was not done, owing mainly to time constraints.

A word of explanantion is due regarding likelihood ratios. Each likelihood ratio is formally twice the difference in the log likelihoods for each observed presence/absence vector existing under the two different models—dependent and independent evolution. In the independent model, where 1 and 0 represent protein presence and absence, respectively, there are four different instantaneous probabilities associated with the trait state transitions $0 \rightarrow 1$ and $1 \rightarrow 0$, which are different for the two proteins, which is how we get 4 rates in total. Each protein in a pair {A,B}, which are not functionally related, will undergo transitions according to the independent model, as the state of A at time $t+1$ is dependent only upon its own state at time t; it is unaffected by the state of B at t. Contrast this with the dependent model of trait evolution, wherein, for a pair of proteins, there are now 8 instantaneous probabilities[8], each associated with one of the 8 possible state transitions: $\{0,0\} \rightarrow \{0,1\}$, $\{0,0\} \rightarrow \{1,0\}$, $\{0,1\} \rightarrow \{0,0\}$, $\{0,1\} \rightarrow \{1,1\}$, $\{1,0\} \rightarrow \{0,0\}$, $\{1,0\} \rightarrow \{1,1\}$, $\{1,1\} \rightarrow \{0,1\}$ and $\{1,1\} \rightarrow \{1,0\}$. The dependent model thus has more variables and more explanatory power. If there is no dependence of A on B or vice versa, however, the dependent model collapses to the independent model, which is actually a special case of the dependent model where $p(\{0,0\} \rightarrow \{0,1\}) = p(\{1,0\} \rightarrow \{1,1\})$, $p(\{1,0\} \rightarrow \{0,0\}) = p(\{1,1\} \rightarrow \{0,1\})$, etc. In general, the dependent model will yield a higher likelihood than the independent model[9]. The question is whether the difference between the two is sufficiently large that we would be led to infer a *real* dependency between the two proteins, i.e. whether, based on observed patterns of coupled evolutionary gain and loss on the tree, we can make the inference that the evolution of the two proteins is *correlated.*

[8] We exclude trivial transitions like $\{0,0\} \rightarrow \{0,0\}$ as well as infinitessimally probable simultaneous transitions like $\{0,1\} \rightarrow \{1,0\}$.

[9] Both likelihoods are very small. They are an estimate of the likelihood of the observed data given the model. Because the model can generate many results, particularly in the space of two presence/absence vectors, any one result will almost certainly be very unlikely, but may fulfil a maximum likelihood criterion as compared with all the others.

All results are from the likelihood runs using 144 maximum likelihood tries. Likelihood ratios are plotted as frequency distributions in buckets of width 0.5. The x axes extend from -1 to 45; there are a few outliers which do not fall within this range but they are very few in number and would barely be noticeable given the scale of the y axes. Percentages quoted and represented graphically include these outliers, but not the excluded absolute likelihood ratios of greater than 1 000.

The distribution of likelihood ratios for the presumptive interactors from HPRD is illustrated in Fig. 4.1. The differently coloured tail of the distribution represents pairs whose likelihood ratios are higher than the 95% confidence interval for the χ^2 test with 4 degrees of freedom. 4 degrees of freedom corresponds to the difference in the number of transition rates in the models of dependent and independent evolution: the former has 8 rates, the latter 4. Fig. 4.2 on the next page highlights pairs with likelihood ratios higher than the 99% confidence interval for the same test.

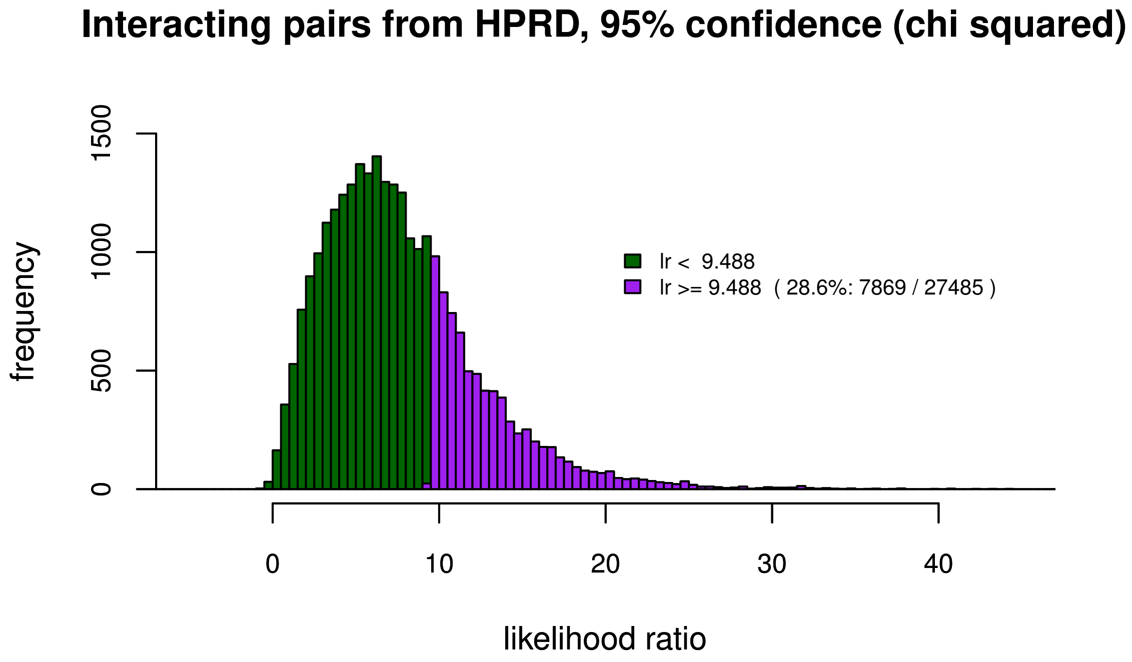


Figure 4.1: Likelihood ratio distribution for HPRD curated interacting pairs, showing χ^2 95% confidence interval

These charts seem to show a very positive result, with close to 30% of pairs showing up in the tail where we would expect to find around 5% by chance, i.e. if there were no correlated evolution. However, this is just the raw χ^2 result; what happens when the distribution of ratios in our control set of presumptively non-interacting pairs is plotted? Figures 4.3 and 4.4 on page 67 and on page 68 show the same distribution of likelihood ratios for the presumptive non-interactors, but highlighting the 95% and 99% confidence intervals, respectively.

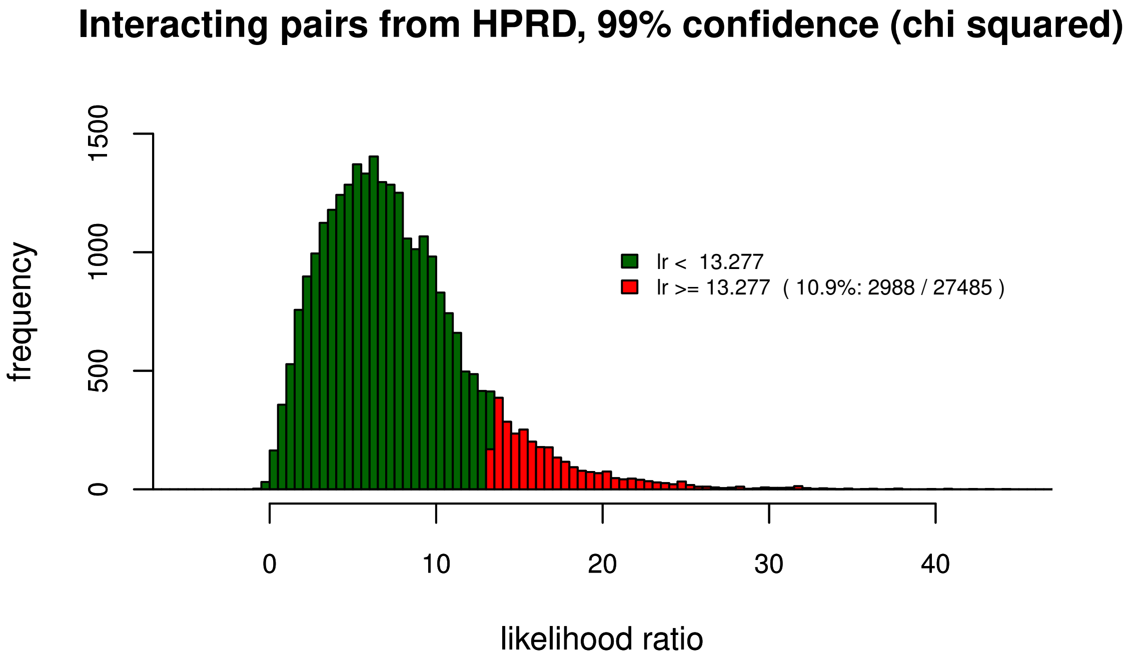


Figure 4.2: Likelihood ratio distribution for HPRD curated interacting pairs, showing χ^2 99% confidence interval

The picture now looks a little different. Although the results are weaker than for the HPRD interactors, they still appear strongly positive, even after all the care taken to ensure that no close relationship could exist between the two members of any of these pairs. What if these plots do however represent the true null set? Maybe we should dispense with the χ^2 criterion altogether. It will be interesting to see how the distribution for the known interactors looks if the confidence intervals are adjusted to reflect the distribution seen in the null set rather than the χ^2, which I have supposed from first principles rather than empirical observation to be correct. Of the 58 280 protein pairs in the null set, 2 914 lie above the 95th percentile and 583 (rounded) above the 99th. Looking at the empirical likelihood thresholds at these percentiles, I established that these are approximately 13.74 and 18.84, for the 95th and 99th respectively. Figures 4.5 and 4.6 on page 69 and on page 70 show what happens when we apply these to the interacting pairs from HPRD.

Now the predictive power of the method based on correlated evolution, while undeniably significant, seems far from overwhelming. Above the 95% confidence interval based on the null set from the randomised and filtered sample of pairs, we have barely twice the number of pairs that we would expect by chance alone. At the 99% confidence interval, things are a little more promising, with closer to three times as many pairs as we would expect by chance falling into this bracket; the greater preponderance at the higher percentile also argues in favour of the method's significance and reliability.

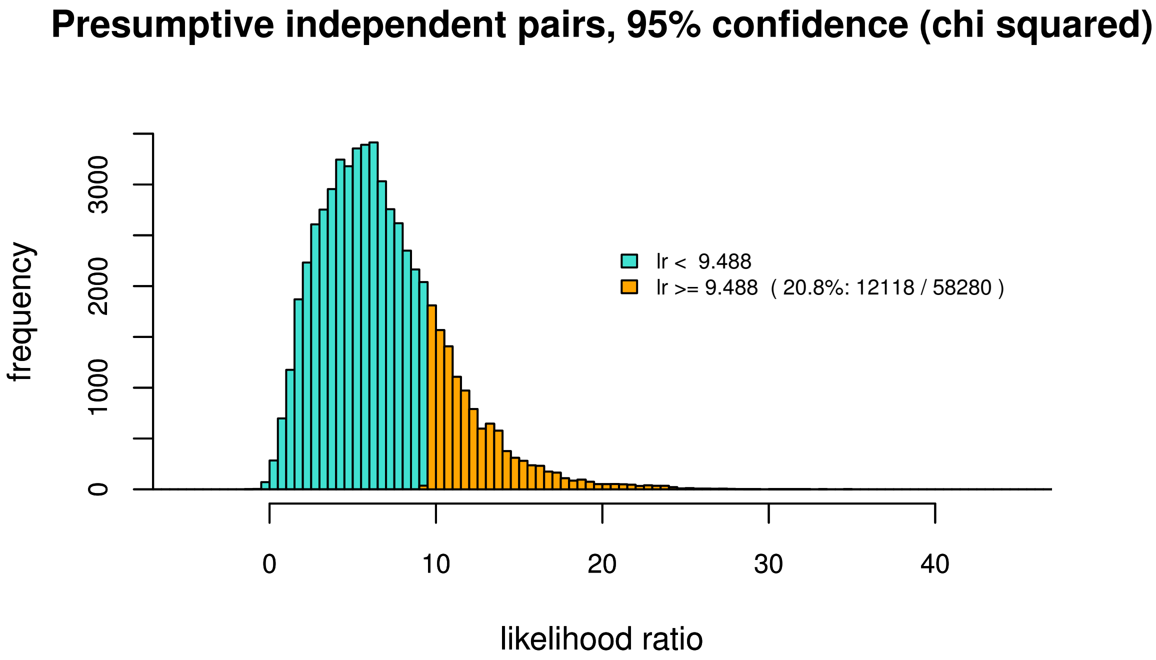


Figure 4.3: Likelihood ratio distribution for presumptively independent pairs, showing χ^2 95% confidence interval

We can use the correlated evolution method to infer a network and divine its features. There are various repositories of curated protein interactions and other network-related data available online, any of which we may examine both on its own, and then using a filter based on correlated evolution. It will be interesting to see how or indeed whether the imposition of such a filter materially affects the character of the network, starting with nodal degree distribution, k.

We do indeed see the characteristic power-law signature of the scale-free distribution in nodal degree, looking at either our HPRD data as a whole (i.e. we presume it to be a real network) or looking at just that part of it which exceeds the 95% confidence limit, based on either the χ^2 criterion and containing 7 869 links, or on the control set and containing just 2 657 links (links being instances of correlated evolution). See the three degree distribution plots below (Figs. 4.7 to 4.9 on pages 71–73).

The last network (Fig. 4.9 on page 73) is quite small but retains its scale-free topology. I felt that it might be instructive to render it visually, using Cytoscape [Shannon et al., 2003]. Because it is an interesting extension to our results, but is not central to such inferences as we might draw from the quantitative method detailed in this chapter, the network diagram is presented separately, in Fig. E.1 on page 106; the round "giant component" sits atop isolated islands of interaction, usually involving just two participants. Hubs—highly connected nodes—seem to be discernible within the

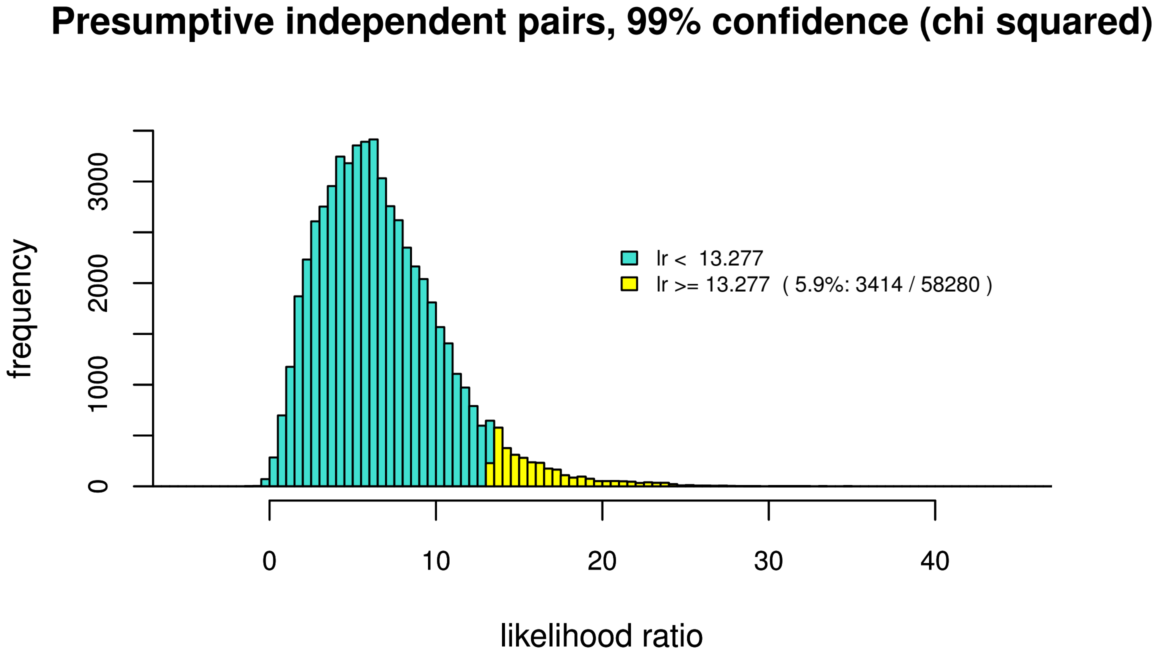


Figure 4.4: Likelihood distribution for presumptively independent pairs, showing χ^2 99% confidence interval

giant component. This might argue in favour of the idea that modularity is a key architectural feature of this and larger protein-protein interaction networks. Whether the sculpting of modules is as important as this might lead us to suppose, could be greatly informed by a more quantitative treatment of this network, and it would have been instructive and possibly revealing to have undertaken such a study. However, the network derived retains its pairwise character. Trait evolution has not been given a deeper treatment, e.g. involving a larger number—3, 4 and upwards—of binary presence/absence "traits" in the transition matrix, whereby we might have probed for a larger number of instantaneous transition probabilities. We could in this way potentially have derived a truly bipartite network of (physical) protein clusters, represented by one type of node, and proteins, by the other. As discussed in Chapter 1, this approach exposes more of the network's intrinsic information content. Arguably this more complex treatment might also have a role in allowing us to mine the network's secrets algorithmically.

4.3.1 Probing for particular functional relationships

I present below a distillation of the functional pairs with the highest credible (i.e. <1 000) likelihood ratios (of dependent to independent model), from the unpaired set. It is these that we should expect to be linked in some way if we are to bring some solid evidence to the table in support of our

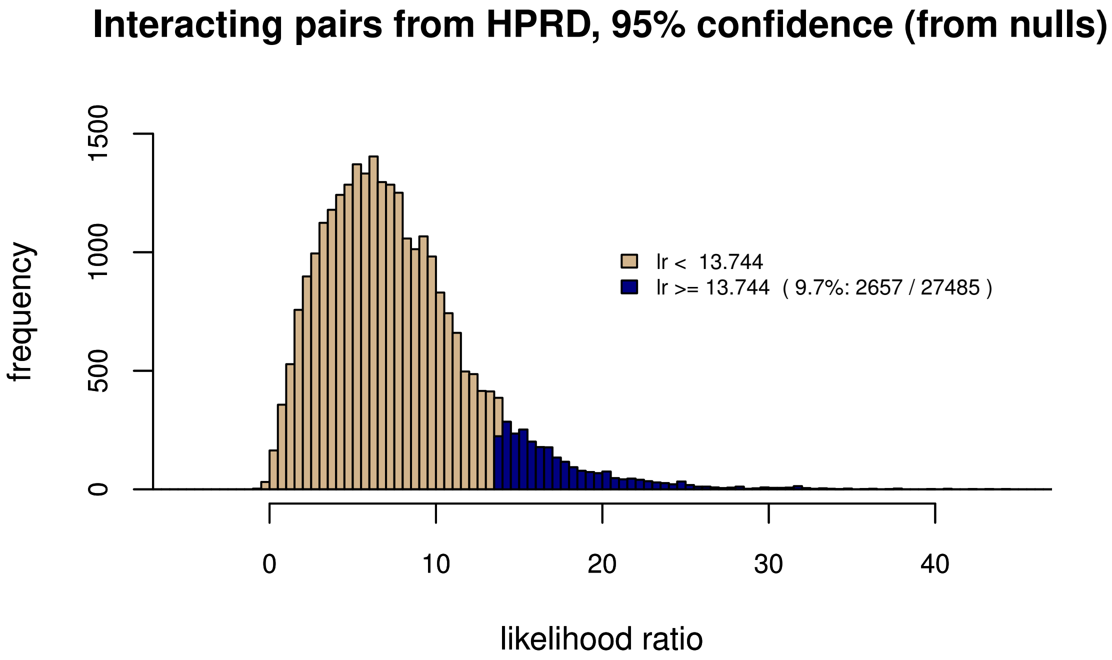


Figure 4.5: Likelihood ratio distribution for HPRD interacting pairs: χ^2 95% confidence interval based on control set

method. Accession numbers are from RefSeq.

Table 4.1: Putative functional protein-protein relationships in RefSeq 36

aceession 1	accession 2	lr
NP_036575	NP_061840	96.76
NP_005949	NP_000886	68.27
NP_004357	NP_055819	34.66
NP_003729	NP_003551	33.21
NP_005756	NP_361014	32.45
NP_722576	NP_000194	31.60
NP_065723	NP_361014	31.13
NP_115916	NP_068775	30.56
NP_009158	NP_000421	30.23
NP_065828	NP_361014	30.00

The first pair of proteins, with the highest of all the non-artefactual likelihood ratio scores, are NP_036575 (sperm-associated antigen 6 isoform 1) and NP_061840 (gamma-1-syntrophin isoform 1). A literature search revealed no obvious association betweeen them. However, the second pair of proteins, NP_005949 (melatonin receptor type 1A) and NP_000886 (leukotriene A-4 hydrolase isoform 1) showed more promise: a literature search on "melatonin" and "leukotriene" revealed a pos-

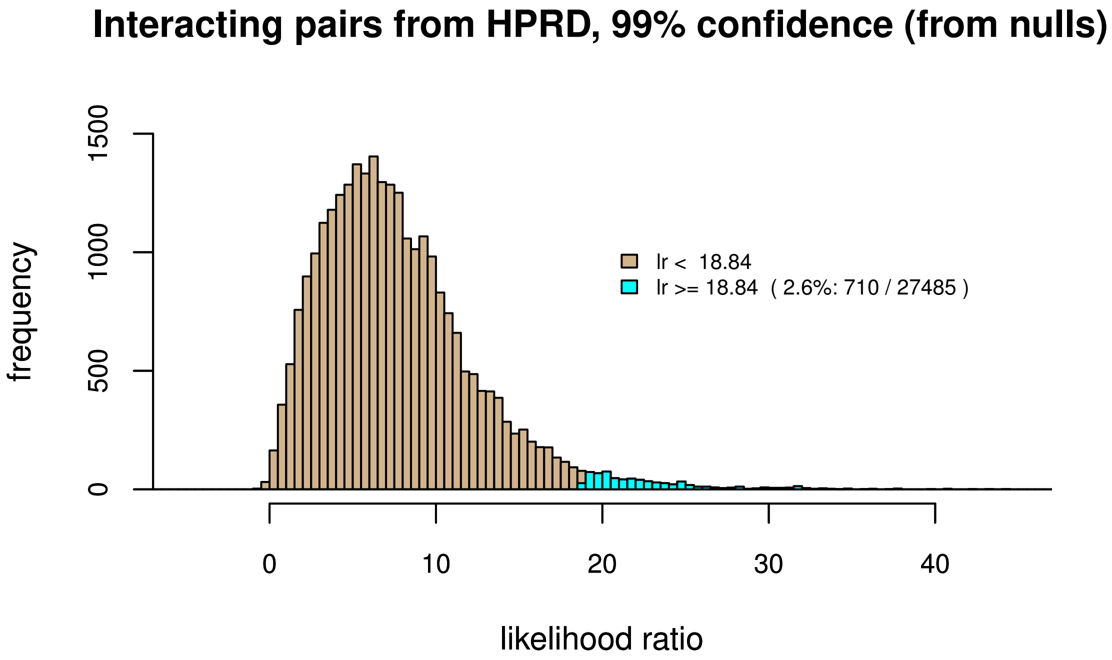


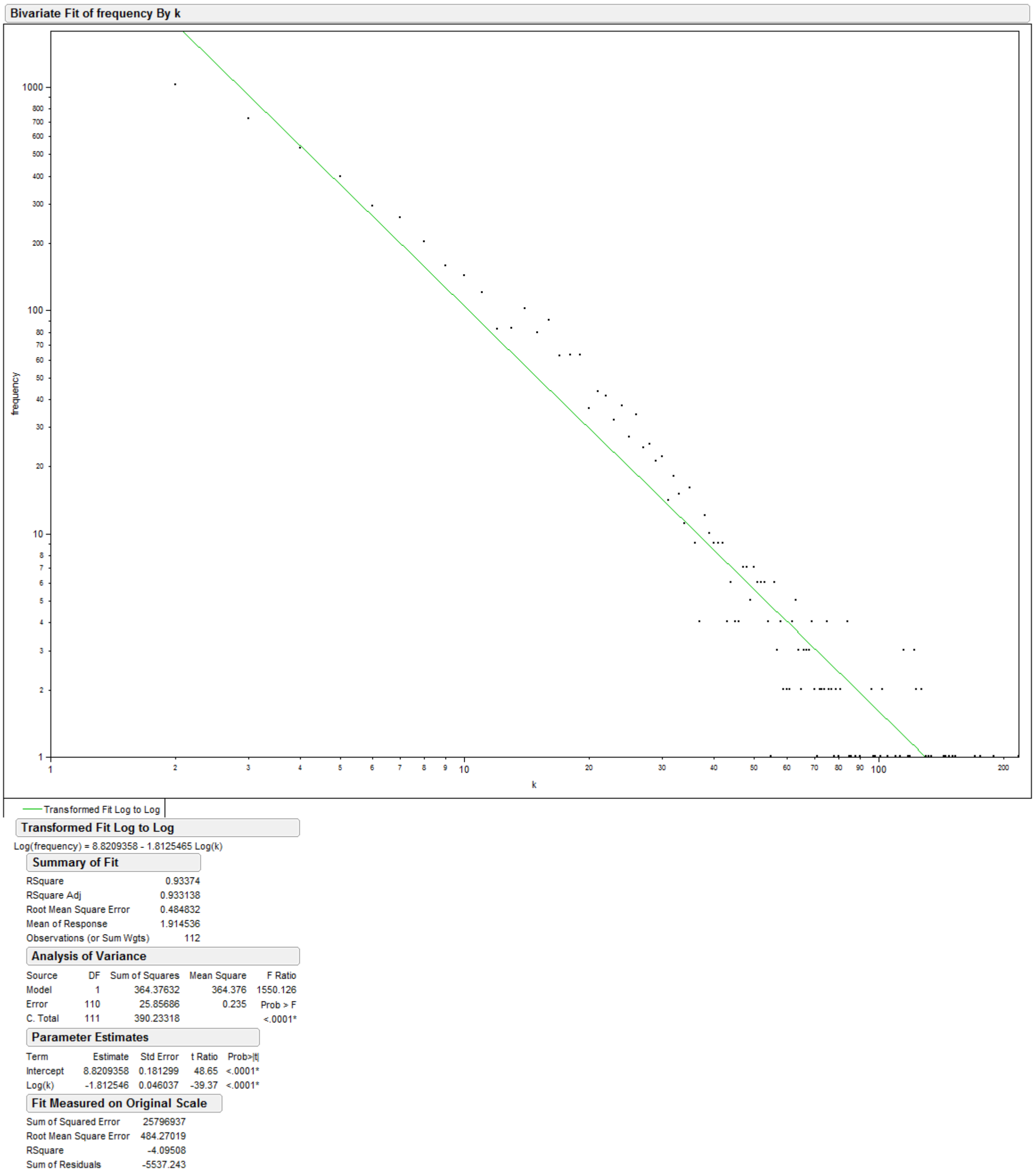
Figure 4.6: Likelihood ratio distribution for HPRD interacting pairs: χ^2 99% confidence interval based on control set

sible link between melatonins and leukotrienes: leukotriene A-4 in particular is mentioned [Chen et al., 2006]. The remaining pairs of proteins listed in Table 4.1 constitute further candidates for investigation.

4.4 Discussion

One feature of note in the degree distribution plots, both for the raw HPRD network and the correlated evolution network, is that each approximately follows the expected power law, i.e. the plots appear roughly linear on log-log axes. This is unsurprising, although perhaps reassuring for the predictive power of our inferential method. However, in each case, there appears to be a systematic deviation from pure power-law linearity; the plot is slightly convex, and bulges in the middle toward the top right of the chart. *In silico* experiments [Read, 2007] revealed that preferential attachment too could produce this effect in a degree distribution plot, although this is far from a claim that I have detected its fingerprint; it seems likely that other dynamic processes might have a similar manifestation. However, this is certainly one area which merits further invesigation.

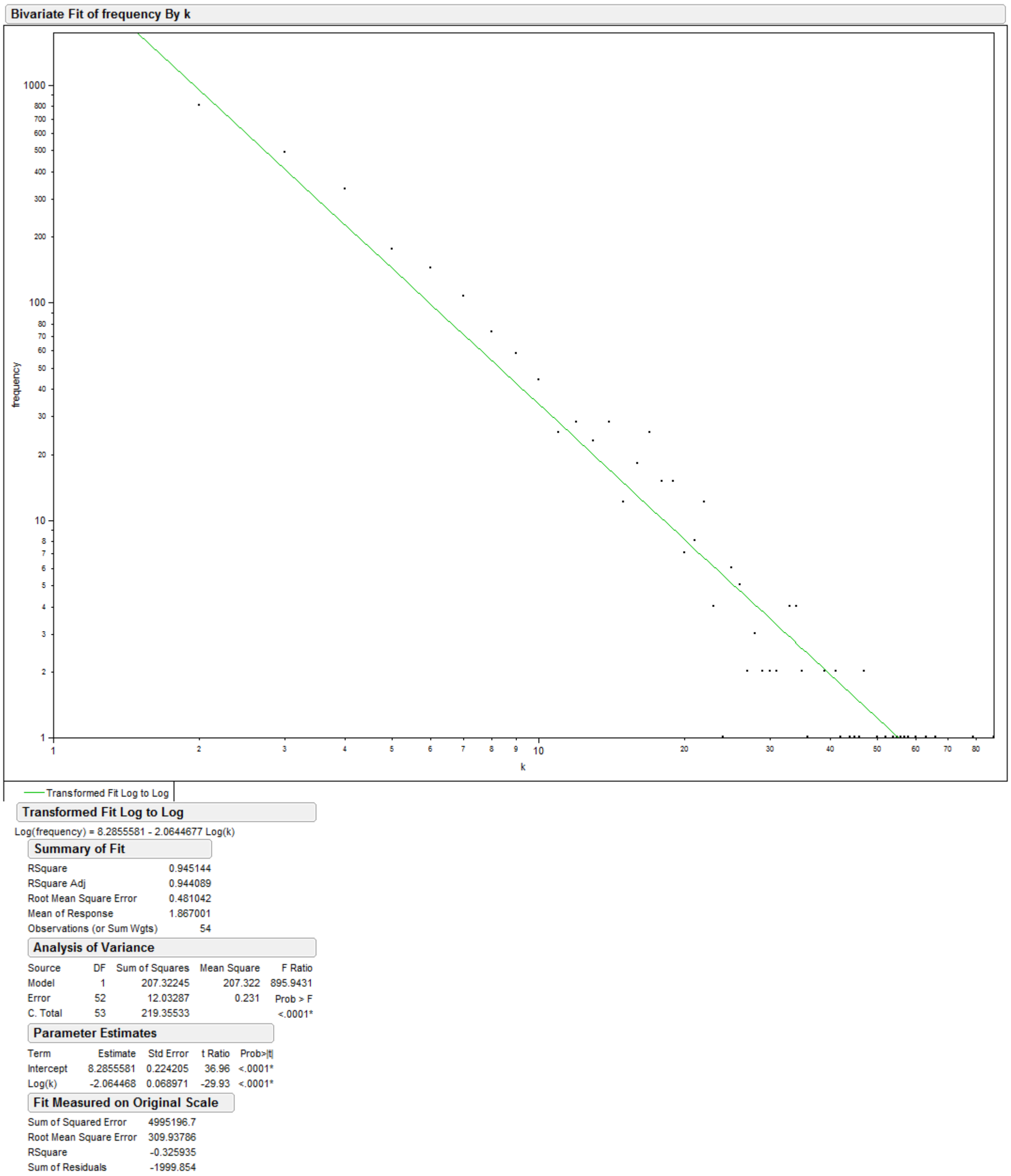
As well as the appearance of hubs in the correlated evolution network diagram produced by



```
Transformed Fit Log to Log
Log(frequency) = 8.8209358 - 1.8125465 Log(k)
  Summary of Fit
  RSquare                        0.93374
  RSquare Adj                   0.933138
  Root Mean Square Error        0.484832
  Mean of Response              1.914536
  Observations (or Sum Wgts)         112
  Analysis of Variance
  Source     DF  Sum of Squares  Mean Square   F Ratio
  Model       1       364.37632      364.376  1550.126
  Error     110        25.85686        0.235  Prob > F
  C. Total  111       390.23318                <.0001*
  Parameter Estimates
  Term        Estimate  Std Error  t Ratio  Prob>|t|
  Intercept  8.8209358   0.181299    48.65   <.0001*
  Log(k)     -1.812546   0.046037   -39.37   <.0001*
  Fit Measured on Original Scale
  Sum of Squared Error      25796937
  Root Mean Square Error   484.27019
  RSquare                   -4.09508
  Sum of Residuals         -5537.243
```

Figure 4.7: Degree distribution plot for 27 485 HPRD protein pairs
(excludes proteins present in many or few species)

Cytoscape, the most notable feature is the dominant presence of Kauffman's "giant component", surrounded by smaller islands of more limited connectness. Localised clusters of nodes too seem to be evident, and in this sense we may infer that, rewardingly from the point of view of the investigator, a relatively high level of clustering, or modularity, is also present. It will be interesting to take the analysis forward and augment these visual impressions with some hard statistics, but Chapters 2 and 5 give some reasons why we may have less than complete faith in the underlying data, based on the erroneous inclusion of incomplete proteomes—not to say that our observations

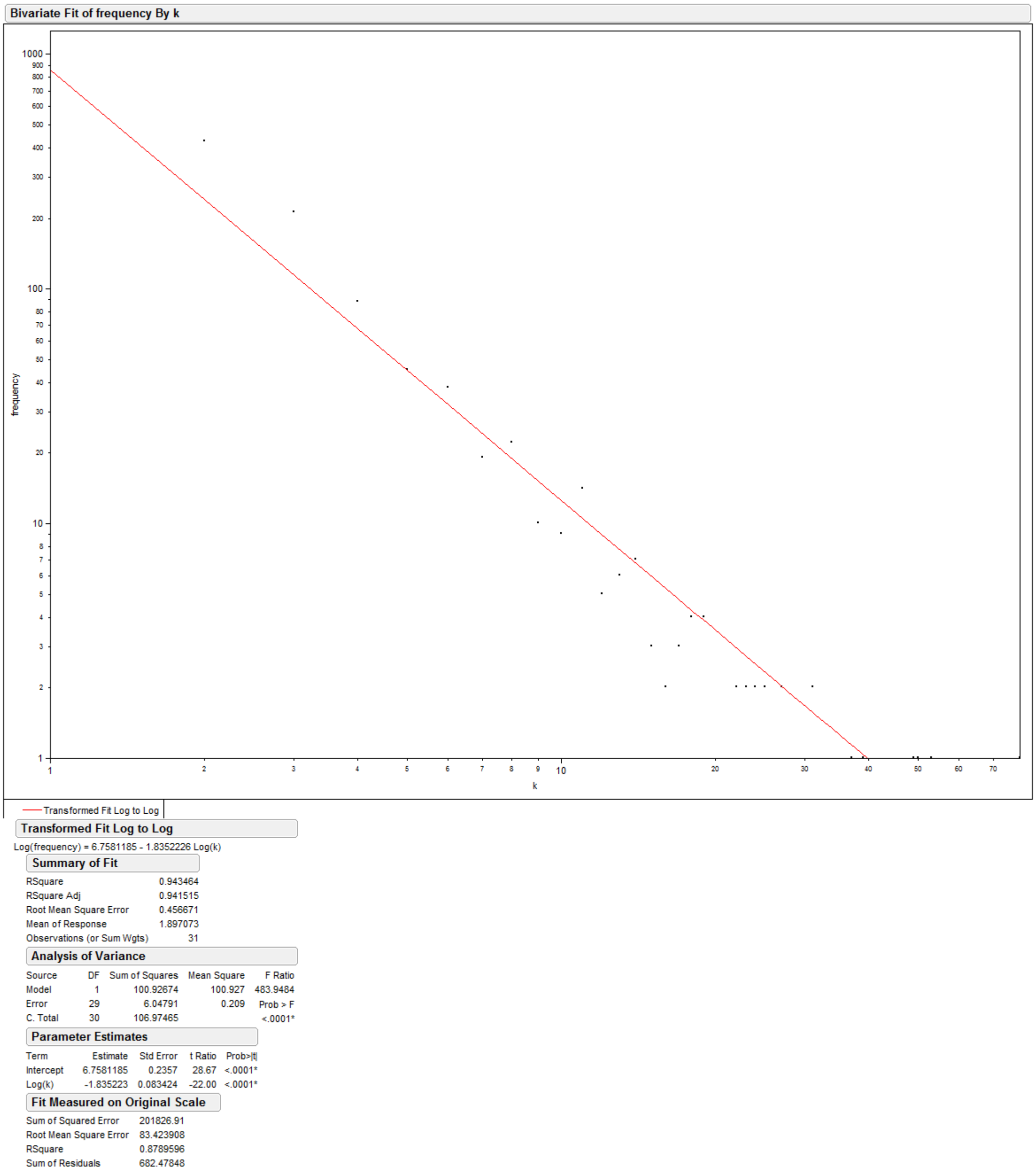


```
Transformed Fit Log to Log
Log(frequency) = 8.2855581 - 2.0644677 Log(k)
 Summary of Fit
 RSquare                       0.945144
 RSquare Adj                   0.944089
 Root Mean Square Error        0.481042
 Mean of Response              1.867001
 Observations (or Sum Wgts)          54
 Analysis of Variance
 Source      DF  Sum of Squares  Mean Square   F Ratio
 Model        1       207.32245      207.322  895.9431
 Error       52        12.03287        0.231  Prob > F
 C. Total    53       219.35533                <.0001*
 Parameter Estimates
 Term         Estimate  Std Error  t Ratio  Prob>|t|
 Intercept   8.2855581   0.224205    36.96   <.0001*
 Log(k)      -2.064468   0.068971   -29.93   <.0001*
 Fit Measured on Original Scale
 Sum of Squared Error     4995196.7
 Root Mean Square Error   309.93786
 RSquare                  -0.325935
 Sum of Residuals         -1999.854
```

Figure 4.8: Degree distribution plot for correlated evolution network 1
(7 869 HPRD protein pairs filtered on χ^2)

and inferences here are entirely undermined, but to hold out the prospect that we could do rather better with a few refinements to the protocols.

Mention deserves to be made again of the GO process used for filtering out a control set of presumptive non-interactors. It appears to have been successful in its stated aim, but I am not aware of its having been applied for a similar purpose anywhere else. Its recursive aspect in particular was not straightforward to implement. While I may on reflection have been over-conservative in the application of the filter, requiring that GO "ancestor" accessions very close to the root of GO



```
Transformed Fit Log to Log
Log(frequency) = 6.7581185 - 1.8352226 Log(k)
  Summary of Fit
  RSquare                        0.943464
  RSquare Adj                    0.941515
  Root Mean Square Error         0.456671
  Mean of Response               1.897073
  Observations (or Sum Wgts)           31
  Analysis of Variance
  Source     DF  Sum of Squares  Mean Square   F Ratio
  Model       1       100.92674      100.927  483.9484
  Error      29         6.04791        0.209  Prob > F
  C. Total   30       106.97465               <.0001*
  Parameter Estimates
  Term        Estimate  Std Error  t Ratio  Prob>|t|
  Intercept  6.7581185     0.2357    28.67  <.0001*
  Log(k)     -1.835223   0.083424   -22.00  <.0001*
  Fit Measured on Original Scale
  Sum of Squared Error     201826.91
  Root Mean Square Error   83.423908
  RSquare                  0.8789596
  Sum of Residuals         682.47848
```

Figure 4.9: Degree distribution plot for correlated evolution network 2
(2 657 HPRD protein pairs filtered on control set)

be separate (for the non-interacting pairs), I nonetheless hold out the principle to the community as one way of using GO curation to isolate not the relationships between genes, but the absence of any such relationship. This may be a useful method of enforcing experimental controls at a time when computational biology has become ubiquitous, but when it is not always clear how to subject computational experiments to the same sort of stringency that would be insisted on in the "wet" laboratory.

Finally, it will be most intersting to try and elucidate some of the more subtle features of the correlated evolution network. The notion that we may use our likelihood technique to identify relationships beyond the straightforwardly pairwise holds out the prospect, for one, that we could recast the whole network in bipartite terms, adding some information back into it. This would enable us to derive an altogether novel structure, prior to applying some of the many new statistical techniques now available for the analysis of the structure itself.

Chapter 5

A trait matrix remade: improved protocols

5.1 Background

There are numerous aspects of the protocols outlined in Chapter 2 in particular which may have been slightly flawed, which may have become obsolete, or which may be less than fully generalisable, in terms of requiring specialised computing resources to which the typical academic researcher may not have ready access. Therefore, several potential methodologies for improving this workflow suggested themselves. Most pertinently, the quality and completeness of some of the proteomes included in the analysis from Chapter 2 are questionable, owing in part to the method used to identify complete proteomes. The silkworm, *Bombyx mori*, was previously mentioned as having been misidentified as having a complete proteome in RefSeq 36. Far more probable (than its proteome being unusually small as a result of thousands of generations of artificial selection) is that the presence of a putatively complete genome sequence in GOLD, together with a RefSeq proteome that exceeds a small absolute mimimum size threshold, constitute an insufficiently stringent set of criteria reliably to identify and thereby exclude proteomes (such as that of *B. mori*) which are incomplete in significant proportion. In fact, the filter applied based on RefSeq proteome size took no account of the size of the genome reported in GOLD; a test based on the *relative* sizes of the Refseq proteome and GOLD genome might arguably have made for a more effective indicator of proteomic completeness. In any event, it seems likely that the inclusion of several such incomplete proteomes might have compromised the study described in Chapter 2, *Bombyx mori* being only the most glaring example[1]. The method lacks both rigour and robustness, because it takes direct account neither of annotation, nor of actual numbers of protein sequences. Consider again the protein count for *B. mori*: if the total recorded proteins are less than one sixth the true figure, this might affect not only the likelihood inference, but the preceding phylogenetic inference too, in that, potentially, I was artefactually excluding some proteins which might have been more diverse in their function and ancestral derivation than the set which in the end I was constrained to employ. In short, the Chapter 2 study had employed slightly impoverished data, and a better method would have been desirable; such a method was therefore sought, developed, tested and implemented.

Additionally, while the improvements to both the quality and quantity of proteomic data in RefSeq, as between releases 36 and 58 (i.e. during the interval between the original investigation of Chapter 2 and its later counterpart reported here), are in a sense incidental and insufficient in themselves to justify a robotic recapitulation of the old protocol, they may nonetheless offer enhanced power to an updated investigation of this general type. The prospect of achieving improved resolution in correctly identifying functional relationships between proteins, and the topology of the network emerging from this set of relationships, seemed tantalising enough to warrant the renewed effort. Because the correct identification of orthologues underpins and ramifies through each subsequent step of this overall investigation, the imperative to get things right at this stage appeared to be all the stronger. Therefore, some new work along these lines was undertaken from 2012 to 2013. A full refashioning and up-to-date reimplementation of the protocol might have been possible, had sufficent time been available to follow it to fruition. As it is, my implementation of a revised workflow for constructing the trait matrix is presented below.

Note additionally that not only is the quantity of data held by RefSeq increasing daily (with major release versions announced every two months), but that annotational quality within RefSeq is subject to continuous revision and improvement. GOLD too appears to have seen an improvement in scope and quality since the Chapter 2 investigation was undertaken [Pagani et al., 2012].

[1] *B. mori* is glaring indeed: in RefSeq 58, 2 224 proteins only are recorded. Contrast this with an ORF count from GOLD of 14 623 and the extent of the mistake appears clear. Although it is just one species, *B. mori* was evidently a major source of false negatives in the investigation reported in Chapter 2, and should not therefore have been included. The weakness in the protocol, which this points up, seems likely to have led to other (albeit less glaringly) anomalous species being included in the Chapter 2 study, too.

5.2 Revised reciprocal BLAST method

Methods presented here differ in the detail rather than the generality from those of Chapter 2; the principle is largely unaffected. The reference proteome used is again that of *Homo sapiens*; RefSeq remains the exclusive source for proteomic data and an unrefined full-proteome reciprocal BLAST process is used, as before. I contend that there is a marked improvement in the potential of the (incomplete) study reported here over its predecessor reported in Chapter 2, based not so much on the reciprocal BLAST protocol used, but on the choice of species, and the prior methodology used to derive them, for reasons explained in Section 5.1. The BLAST protocol outlined herein has the advantage of being fully reproducible by other investigators because it no longer depends upon the use of in-house hardware resources.

5.2.1 Obtaining the proteomes from RefSeq

I used the same programs that I used with RefSeq 36 to download proteomes from RefSeq 58, the latter having been released by the NCBI on 11 March 2013. A few modifications were necessary to accommodate slight changes in the native filenames available for download from RefSeq, but this was straightforward. The catalogue was again downloaded and stored locally in a database table. This time round I switched from MySQL to PostgreSQL [Momjian, 2001] as my database engine of choice, mainly because of familiarity, having worked with PostgreSQL more recently. More generally, PostgreSQL is an attractive option because of its more sophisticated feature set, including the availabilty of analytic or "windowing" funtions, which enable group data to be associated with each record that belongs within the group—so that, for example, a quantity associated with a species (such as the number of proteins in its proteome) may straightforwardly be represented as a percentage of the average number of proteins for the phylum to which it belongs[2]. PostgreSQL also allows custom functions to be written in a variety of external languages, including Java, Perl, R and C. Such a feature has the potential not only to enable sophisticated user-written functions to be embedded directly within SQL queries, but to facilitate the integration of biological databases (even small ones like my own) with various analytical bioinformatics and other toolkits for which public APIs are availiable in one or more of these languages[3]. However, within the parameters of this investigation, there was neither occasion nor time to take advantage of such rich extensibility.

All records obtained from the downloaded sequence files were summed for each species and compared against the species totals derived from the catalogue, as a check. The catalogue includes

[2] This particular percentage may of course be less than, equal to or greater than 100.

[3] The availabiltiy for several programming languages of extensive libraries of bioinformatics-related code modules, e.g. BioPerl [Stajich et al., 2002] and BioJava [Holland et al., 2008], greatly facilitates the development of such APIs.

microbial records; in fact, because I did not exclude microbial files from my sequence downloads, it was possible to make these comparisons for all species. RefSeq includes not only protein sequences, but nucleic acid sequences as well, comprising several different categories of DNA and RNA. Protein sequences themselves are allocated in RefSeq to one of five different categories, each category being associated with a standard two-letter prefix at the beginning of the RefSeq protein accession, wherein each such prefix has a "P" as its second character—this is definitive for proteins. These five prefixes and my own brief distillation of their associated categorical types (based on information on the RefSeq web site) are shown in Table 5.1.

Table 5.1: Meanings of RefSeq protein prefixes

AP	Alternate Protein product, derived from an alternate assembly.
NP	Protein product, based on accepted assembly, usually but not always full-length protein precursors.
XP	Protein product, predicted using a genome annotation process.
YP	Protein product lacking a matching transcript, typically reserved for microbes, viruses, and most importantly for us, mitochondria.
ZP	Protein product, predicted (typically computationally) by a genome annotation process on shotgun sequence data.

So, a typical protein product for *Homo sapiens* would be "NP_115634.2" (TBC1 domain family member 3F), while an example predicted protein product is "XP_003960205.1" (PREDICTED: E3 SUMO-protein ligase PIAS3-like isoform 4). The human proteome from RefSeq 58 contains 35 922 sequences in total, comprising 33 981 "NP" sequences, 1 928 "XP" sequences and a further 13 mitochondrial sequences which are not part of the human chromosomal complement itself, but which are nonetheless included in the study, as the mitochondrion's reduced genome is likely to be highly conserved, being under stringent purifying selection [Stewart et al., 2008]. These mitochondrial sequences start with "YP" and are all concerned with the electron transport chain [Hatefi, 1985], e.g. "YP_003024035.1" (NADH dehydrogenase subunit 4). The RefSeq catalogue makes it straightforward to assign the "YP" sequences definitively to a particular type of intracellular entity, because the field "refseq_dir" contains a list of directories on the RefSeq web site where the accession can be found (directly beneath `ftp://ftp.ncbi.nlm.nih.gov/refseq/release/`). As well as "complete"[4] and "vertebrate_mammalian", all 13 human "YP" sequences have the additional entry

[4] Here, "complete" does not signify completeness of proteome. It simply refers to a directory on the RefSeq ftp server from which all available data can be downloaded, no taxonomic filter having been applied. Taxonomic filtering at source (i.e. using a remote directory other than "complete") may be important simply in view of the quantity of data available for download from RefSeq, which, as at release 58 contained over 30 million protein accessions alone and amounts to many gigabytes.

"mitochondrion"[5]. The RefSeq catalogue additionally provides a status field denoting the level of confidence associated with the sequence. There are six status categories which apply to the human proteome. In order from lowest to highest confidence, these are "MODEL", "INFERRED", "PREDICTED", "PROVISIONAL", "VALIDATED" and "REVIEWED". All 1 928 human "XP" sequences are classed as MODEL, while the 13 mitochondrial "YP" sequences are classed as PROVISIONAL. None of the 33 981 "NP" sequences was classed as MODEL; the breakdown of human NP sequences by status is shown in Table 5.2[6], again from lowest to highest confidence.

Table 5.2: Breakdown by status of human "NP" proteins

status	sequence count
INFERRED	77
PREDICTED	158
PROVISIONAL	1 656
VALIDATED	12 116
REVIEWED	19 974

A simple numeric check consisted in summing the counts of all five protein accession categories from Table 5.2, for each species, and to compare the resulting total with the relevant downloaded FASTA sequence count. These two numbers were found to tally for every eukaryotic organism, but not for all microbes. I discovered that if the "ZP" accession counts were removed, the numbers for microbes too now tallied. These "ZP" accessions are protein products inferred from shotgun genome sequencing, and are not recorded for eukaryotes. The fact that the numbers from the catalogue and the sequence files tallied gave me added confidence that I was correctly interpreting the source data, and correctly allocating the various eukaryotic species from the aggregated source files to their respective single-species files, prior to conversion into BLAST+[7] databases.

Having downloaded RefSeq 58 in full and run a revised collation script (`collate_organisms_pg.pl`—see Appendix F) to create the single files for all available species, my next concern was to identify with accuracy those species whose proteomes were effectively complete, as it would be these alone which I sought to include in the reciprocal BLAST. To make this identification more robust, I felt

[5] This use of accession prefixes in RefSeq appears not to be entirely consistent. For example, the 13 mouse (*Mus musculus*) proteins listed under "mitochondrion" are prefixed "NP". This is true too of the 117 proteins from the model brassica *Arabidopsis thaliana* listed under "mitochondrion", and also of its 85 proteins listed under "plastid" (i.e. proteins manufactered in the chloroplast).

[6] This table provides a simple, pared-down distillation of RefSeq annotational status for one species (*H. sapiens*) only; in general, for each species, we could represent annotational state as a sparse matrix, with accession prefix along one axis, and status along the other. It is conceivable that some sort of algorithmic treatment of such matrices might assist us in deriving some useful metric of per-species annotational maturity. The incorporation of such a protocol is offered here simply as a suggestion to be considered in the design of future studies of this kind.

[7] The updated suite of programs from the NCBI is known collectively as BLAST+ [Camacho et al., 2009]; the older eqivalent programs used in the Chapter 2 study are referred to simply as BLAST.

that it was imperative to adopt an approach somewhat different from the one employed in the Chapter 2 investigation.

5.2.2 A new proteome filter based on ORF count

Although RefSeq, as previously noted, provides no information regarding the state of completion of the proteomes it offers for download, multiple species being mixed up together in the same files, others too have encountered this issue and developed their own protocols for reckoning the degree of completeness. I followed the protocol adopted by the developers of the ProRepeat repository [Luo et al., 2012] in comparing the ORF counts from genome assemblies listed as completed in GOLD, with the sequence counts from my downloaded proteomes. These authors, like me, used RefSeq as their source for proteomic data; they then applied a filter specifying that the counts from RefSeq should differ by <5% from the ORF count reported by GOLD, for each corresponding species. Conformance to this stipulation they took to indicate a putatively complete proteome.

Because of the general variability in the state of proteomic annotation for all but a select subset of model organisms, I could find no better way (than following the ProRepeat protocol) of systematically determining whether a given proteome was complete, although further literature searches on individual species might have been conducted, had time allowed. Still I encountered several difficulties in adapting the method to my particular data. A lot of these difficulties arose as a result of transient issues in the software chosen, most of which is under continuous development and is therefore subject to change. For the consequent reason that many such issues, at least in their specifics, are unlikely to concern future investigators, I have detailed some of these rather esoteric software problems (often specific to particular software *versions*), and my solutions to them, not in this methods section but in Appendix C.

5.2.3 Applying the ORF count filter

In comparing the RefSeq protein counts with GOLD ORFs, I first discounted from the former any proteins originating in mitochondria or chloroplasts. This number is usually fairly negligible anyway, typically 13 mitochondrial proteins for metazoa such as *Homo sapiens* and *Drosophila melanogaster*—although the number can be higher for other kingdoms, e.g. 43 (provisional) for the marine green alga *Ostreococcus tauri*[8] and 50 (provisional) for the fungus *Giberella zeae*, a major pathogen of wheat. While it is actually not entirely clear that mitochondrial and chloroplast genes

[8] 43 *mitochondrial* proteins: *O. tauri* has an addtional 61 choroplast proteins (60 provisional), which are of course entirely absent from metazoa, but which I also excluded when making the comparison with ORF counts.

are absent from the ORF counts, it seems probable. That being the case, and because their impact is so marginal anyway, I felt that their removal (at least for the purposes of comparing counts) was appropriate.

In looking at RefSeq's proteomic content, it seems that for the vast majority of non-microbial species for which any proteomic data is held, it is, critically, only the mitochondrial[9] proteome which has any entries. In fact, RefSeq typically holds no more nor fewer than 13 mitochondrial proteins, and no chromosomal proteins, for most metazoan species for which it holds any data at all. It seems likely that these 13 proteins are homologous, but I have not performed the BLAST analysis (outside of the putatively complete species) necessary to confirm that this is so.

Applying a threshold value, of 90% RefSeq proteins to GOLD ORFs, yields 85 species[10] for RefSeq release 58, a number which we may infer to be proteomically complete or nearly so. Too stringent a filter, such as setting the filter at 95%, gives undesirable results in the sense that the value for *Homo sapiens* is lower than we might expect, at 93.0%; clearly the human proteome is required as the basis for the study and I adjusted my criterion accordingly. There will be some artefactuality asociated with the respective figures anyway, possibly arising from inferences about the level of post-translational modification at a whole-proteome level; still, there is little doubt that *Homo sapiens* has a complete proteome, by any current standard.

As previously mentioned[4], RefSeq holds (and duplicates) proteomic data in various directories on its ftp server. For eukaryotes only, the breakdown of all RefSeq proteomic entries by classification—counts of those species having chromosomal entries, as against all species held in RefSeq (most of which lack chromosomal entries)—together with a count, for each category, of the few species reaching our completeness threshold of ≥90% protein count (from RefSeq) over ORF count (from GOLD) is given in Table 5.3. Note that beneath the current RefSeq release's ftp directory, eukaryotic proteomes are placed in one of six subdirectories: `vertebrate_mammalian`, `vertebrate_other`, `invertebrate`, `plant`, `fungi` and `protozoa`. These subdirectory names do not correspond exactly to kingdoms, or to phyla, but reflect the historical bias in the sequencing and annotation of genomes to date. Mammals, crop plants, certain pathogens, and model organisms like fruit flies have been preferentially studied.

It is very apparent that while RefSeq's coverage of (presumptively highly conserved) mitochondrial proteins is fairly broad, its coverage of chromosomal proteins is limited to a few key species, of which I have inferred a significant proportion ($85/235 = 36\%$) to be complete. Interestingly, of

[9] For plants, and some "protozoa", plastid proteins too are often populated in RefSeq where there are no entries for the organism's own chromosomal proteins. Protozoa, plants and fungi sometimes have entries for plasmid proteins as well.

[10] See the bottom rightmost figure in Table 5.3.

Table 5.3: Depth of coverage of eukaryotic proteomes in RefSeq 58

taxonomic directory	all	chromosomal	complete
`vertebrate_mammalian`	417	29	5
`vertebrate_other`	1 627	17	3
`invertebrate`	1 049	41	18
`plant`	318	21	11
`fungi`	210	85	32
`protozoa`	159	42	16
totals	3 780	235	85

the 85 complete proteomes, only 5 are mammalian, including *Homo sapiens*. The mouse proteome failed the filter, as did that of the chimpanzee. This appears to be because GOLD is reporting very high ORF counts for both: 60 745 and 48 910 respectively, as against 38 612 human ORFs. Given that the chimpanzee's total protein count from RefSeq 58, at 34 718 including post-translational modifications, etc., is reassuringly close to the human protein count of 35 909[11], we might expect the ORF count too to be similar. It seems reasonable to conclude that the reasons for the mismatch are artefactual, and possibly to do with relatively poor or provisional annotation; yet the opposite could be the case for the mouse ORFs, which one might expect to be better studied and more richly annotated than their human equivalents. Whatever the reason, neither *Mus musculus* nor *Pan troglodytes* made it into the investigation reported in Chapter 5 on page 75.

5.2.4 An updated reciprocal BLAST implementation

Having identified a set of proteomes which were roughly complete, I was again faced with building the BLAST+ databases and running the necessary reciprocal BLAST jobs. To recap the method briefly, I aimed to run a reciprocal BLAST of each unique human protein sequence against each of the other 84 proteomes which passed the filter. A reciprocal top hit would be taken as indicating the presence of a human orthologue in the target species' BLAST+ database, which would therefore be counted as "present" in the trait matrix. Owing partly to considerations of access to computing equipment, and partly to the desire to enable other investigators ro replicate and extend this work, I chose to use a cloud computing service for this updated protocol. I omit the details here, because in most regards it remains the same as the protocol employed in Chapter 2. However, for those to whom any modifications might be of interest, the updated procedure is detailed in Appendix D.

[11] Contrast this with the counts from RefSeq 36, reported in Chapter 2: 51 405 chimpanzee proteins, as against 37 946 human (without removing duplicates).

5.3 Results

As with the data in the Chapter 2 investigation, I present my results graphically, for the most part. Figure 5.1 on page 85 shows how many orthologues to human proteins were found (using the reciprocal BLAST protocol) in each of the 84 non-human eukaryotes in the study. Each species is also ranked out of 84, from highest to lowest orthologue count. Ordering is alphabetical, by genus, then specific name. Major branches (kingdoms) within the eukaryote tree are colour coded[12].

Of at least as much importance as the number of orthologues themselves is the number of proteins common to all, which together or separately will make good candidate sequences for phylogenetic inference (using their RNA template sequences). This study revealed 18 such proteins, shown below.

NP_006422.1 : T-complex protein 1 subunit beta isoform 1
NP_002786.2 : proteasome subunit beta type-3
NP_000962.2 : 60S ribosomal protein L7
NP_000969.1 : 60S ribosomal protein L23
NP_001012321.1 : 40S ribosomal protein SA
NP_060561.3 : elongator complex protein 3
NP_009057.1 : transitional endoplasmic reticulum ATPase
NP_000973.2 : 60S ribosomal protein L21
NP_057388.1 : probable ribosome biogenesis protein RLP24
NP_004757.1 : coatomer subunit beta'
NP_001240308.1 : 60S ribosomal protein L15 isoform 1
NP_001003.1 : 40S ribosomal protein S8
NP_000959.2 : 60S ribosomal protein L4
NP_001257411.1 : 26S proteasome non-ATPase regulatory subunit 11
NP_057439.2 : probable ATP-dependent RNA helicase DDX47 isoform 1
NP_001001.2 : 40S ribosomal protein S6
NP_001020092.1 : 60S ribosomal protein L9
NP_038203.2 : isoleucine—tRNA ligase, cytoplasmic

[12] The use of 10 colours for the 10 branches is admittedly possibly excessive, as they are not so easy to distinguish.

5.4 Discussion

In terms of the findings of the new investigation, in overall terms these results are preliminary only, a precursor to running updated implementations of the phylogenetic and likelihood analyses reported in Chapters 3 and 4. The aim of such a full investigation, thus brought up to date, would have been the more accurate identification of functionally interacting proteins, with higher likelihood, based on a better control set of presumptive non-interactors. I contend that the more diverse set of proteins found to be common to all species in the study to this point would have provided a credible springboard for a renewed, full investigation of similar type.

Another consideration relating to the reimplemented protocol reported above lies in its demonstration that the availability of modern cloud computing services such as Amazon's EC2, among an increasing number of others, brings Bayesian and likelihood methods within the reach of the typical phylogenetic researcher, even for the inference of large phylogenies (potentially involving hundreds of species). Not only that, but such services greatly aid the sharing of HPC[13] code, since it becomes relatively trivial for researchers to configure similar instances of an operating environment, without relying on in-house systems managers at multiple academic institutions.

[13] High Performance Computing

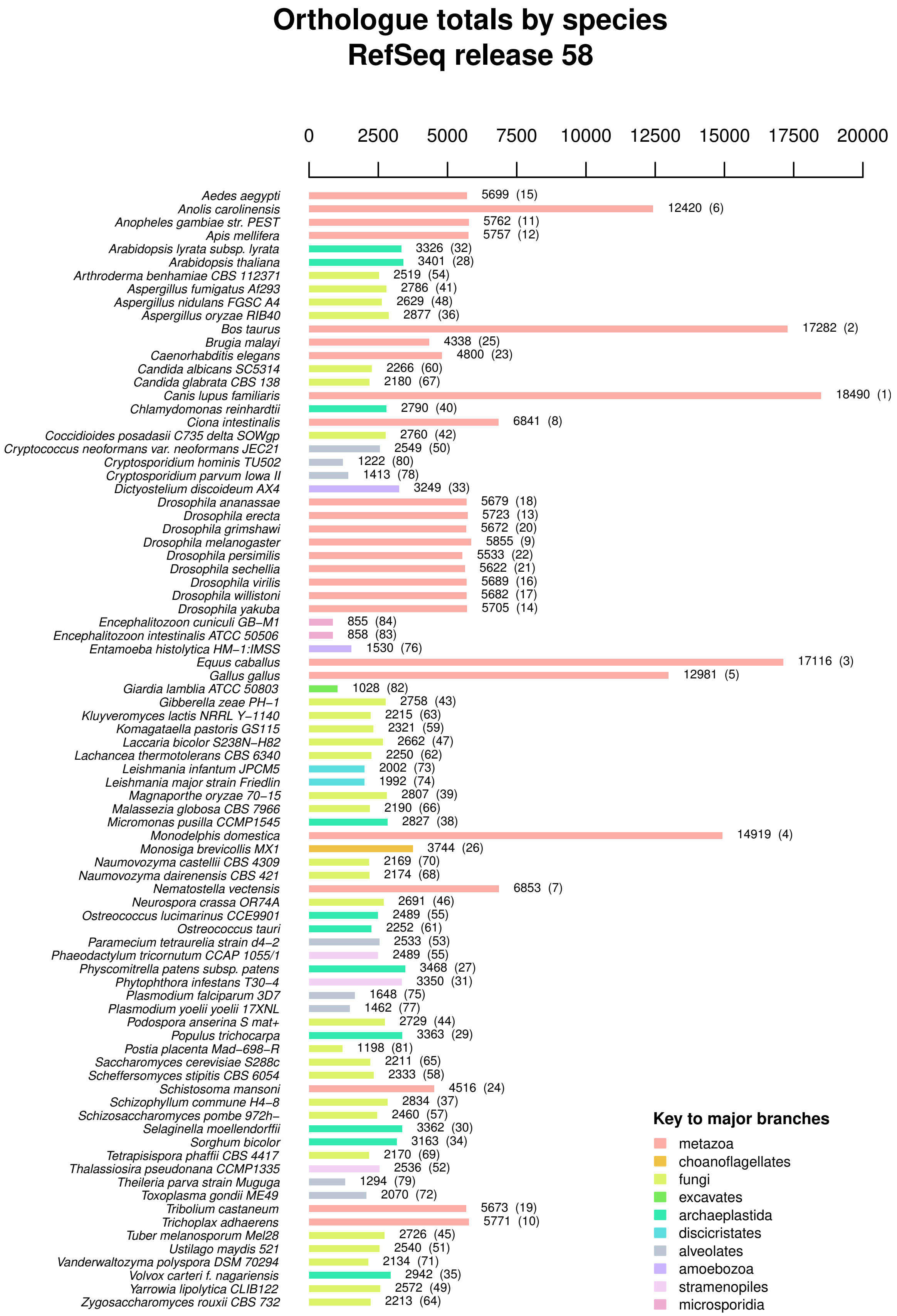


Figure 5.1: Nos. of human orthologues found in other species in 2013, colour keyed by major phylogenetic grouping. Counts appear next to the bars (ranks in parentheses).

Chapter 6

Conclusion

6.1 Network evolution

In attempting to divine the implications of this study, we may reflect that it remains incomplete, and that this adds a heavy ladleful of provisionality to any inferences we may draw. It does seem to demonstrate that there is a level of predictive power in the method; certainly some of our more quotidian predictions—such as the signature of a power law in the degree distribution of the inferred correlated evolution network—appear to be borne out. This is not to say that such a conclusion is mundane: far from it, actually, because an approximate power law is exactly what we would expect if the network were an analogue of the actual protein interaction networks that have already been teased out from multiple curated wet-lab investigations. So in that sense we have a method that yields up a network with similar topological characteristics to those biomolecular networks which have been inferred by more traditional means. This is a result of some note for a study whose raw material is just that—a set of raw protein and nucleic acid sequences downloaded from the NCBI's servers, without any recourse to annotation of the associated proteins' functions themselves. So potentially we have a method for the *de novo* prediction of individual functional linkages between proteins based on the digital alphabet of a set of amino acid sequences alone. We have not had to augment the method with structural prediction algorithms, or other sources of information exogeneous to the sequences themselves. This alone speaks to the applicability and the accessibility of the method, which within the formalism of a strict workflow is in principle available to a very broad academic community.

6.2 Data quantity and resolving power

Of course, here the method has been used to elucidate a network topology, which has interesting features in its own right, but which is only a particular outcome of a protocol with more general applicability. With refinement, might we not be able to use this protocol to pinpoint individual functional relationships of medical or even biotechnological significance? Perhaps the results are too tentative, and we canot have sufficient confidence that in the case of a particular protein pair, we are not just dealing with so much noise. Yet arguably, the salient feature of the weakly positive results is not their weakness, but that they are positive at all. Something is there. And how rich really are our input data? In the just over three years between the late 2009 and early 2013 reciprocal BLAST runs, the number of species whose proteomes were inferred to be complete in RefSeq remained unchanged, at 85. While the later protocol for determining "completeness" was undoubtedly more accurate, more to the point is that the RefSeq project is focussed on curational accuracy, with its primary emphasis on quality over quantity, in offering investigators a reliable consensus sequence for key model organisms. In reality, both the quantity and annotational quality of protein sequences in GenBank has increased considerably in this period; as attested by the 2013 build of the trait matrix, the quantity of protein records in RefSeq does not reflect this. Arguably the conservative instinct in RefSeq's curators is sound (the power to curate may not scale linearly even with the quality of the data, let alone the quantity), yet it allows the possibility that RefSeq is no longer the best source, because it leaves too much out.

What of the now many other eukaryotic genomes said to be complete[1]? Not all proteomes will be fully annotated of course, but of those that appear to be complete in RefSeq based on the ORF count comparison method, most are in a much more provisional state than that of *Homo sapiens*, as reflected in the counts falling into each sub-category. It may have been such a curational imprecision which caused the apparent discrepancy between the number of proteins recorded respectively for human and chimpanzee in the 2009 study (using RefSeq release 36), in which the chimpanzee number appeared grossly overstated at over 50 000. These were "proteins" whose state of annotation was overwhelmingly provisonal. The particular issue affecting the chimpanzee proteome appears now to be more or less resolved.

Consider now the effect of adding new species to the study, assuming that the proteomic data for each such species are truly reliable. The very limited subset of eukaryotes with whose proteomes I have been working affords some opportunities to observe concurrent losses of proteins in unrelated species, but these are at best very rare occurrences in the natural course of events, even over evolutionary time. We will also see retentions, which we are unlikely to be able to identitify, of pro-

[1] 186 according to GOLD at the time of writing (June 2013)

teins which are functionally redundant owing to the loss of a partner protein but which selection has not yet eliminated. We would like to afford ourselves the opportunity of observing independent, concurrent losses in as many species as possible, to increase the chance that we will observe anything at all. If we double the number of species, we may not precisely double our chances of seeing events of correlated loss (this will depend also on the place of any additional species within the phylogeny), but we do increase the probability: the acuity of our method depends crucially on its having sufficient example species to correlate these unusual correspondences. Even in cases where we see correlated loss, more species means a better chance of seeing even more: a threefold independent loss of protein pairs is a surer reckoner of functional interdependence than a twofold loss. The method stands to gain much from the input of more and improved data, to the point where it may become of genuine predictive utility in fields far outside the inference of network topology.

6.3 Unrealised promise in the 2013 method?

In the light of the above, it is instructive to reflect on what was discovered in the later 2013 trait study. *Bombyx mori* has already been identified as a "rogue" species which should never have made it into the 2009 study. Yet the effect of admitting it may have been rather more profound than simply to reduce the diversity of the proteins available for running the phylogeny. The mere fact that it was there at all, and correctly positioned on the phylogeny among the other insects, means that in all probability, several thousand falsely identified correlated loss events[2] had crept into the study, most of which were purely artefactual—an artefactual consequence of poor data screening on my part.

The positive aspect of this analysis of past mistakes, of course, is to wonder whether the 2013 study, if it were carried further (as with the 2009 study), might not show a stronger statistical effect in correctly identifying funtional linkages between proteins, and in inferring the network based on this identification. It seems not unreasonable to propose that continued investigation along this line might yet bear fruit. Or if not, that at least we would have a stronger foundation for doubting the usefulness of the method itself.

[2] Even had *B. mori*'s apparently reduced proteome not been illusory, its inclusion in the study would still have been inappropriate. The correlated loss events would still have been evident, but in the case where an entire genome undergoes a significant *generalised* reduction, correlated loss of any particular protein pair (of which there will of course have been many) is quite unlikely to denote true functional linkage, even when the same pair is lost independently on another branch of the tree, in which case the anomalous inclusion of *B. mori* would have served only to mislead.

6.4 Potential methodological modifications

As new commercial DNA sequencing technologies, such as IonTorrent [Merriman et al., 2012], come on stream, the rate at which new genomic data become available is likely to undergo a still sharper rate of increase than it has already experienced to date. Yet for our method, the extra step of inferring the proteome, even if it involves relying to an extent on approximate computational techniques, is still necessary. A length of genomic sequence which is no longer transcribed, and has passed out of purifying selection altogether, may in theory yet be a top BLAST hit for our source organism's fully transcribed DNA sequence, and the reciprocal step might also succeed—potentially leading to many false positives. So despite the suffiency of raw sequence data, the method is constrained by a curational bottleneck in that we must be as sure as is possible that each proteome is roughly complete but not overstated. In fact, the reciprocal BLAST protocol, in producing a binary result, is somewhat self-limiting.

What if we were to specify a threshold for the level of homology? Perhaps if we could weight the results in the trait matrix—the weights varying inversely with genetic distance—we might expect genomic sequences no longer under purifying selection (because they are no longer transcribed) to have been subject to sufficient drift over evolutionary time that they would score very weakly. And yet, by the same token, proteins in organisms at a greater distance from us would tend to score lower than equivalent proteins in organisms with whom we as human beings share a more recent last common ancestor. One approach might be to use a nucleic acid reciprocal BLAST as our primary orthology filter and set a homology threshold as the second stage, which would ultimately therefore still yield a binary result—our presence/absence "trait". Could we not in that way use BLAST at the nucleic acid level (i.e. blastn) to derive our putatively orthologous protein sets? This would free us from the tyranny of curation, which cannot hope to keep pace with the availability of raw genome sequence data. The quality of the method depends on both the quality *and* the quantity of input sequence data.

6.5 New networks, new insights

If we can produce stronger evidence that we are able reliably to make inferences about functional linkages between proteins[3], we then have the option of extending the existing study in several tempting directions, e.g. by exploring modularity, and bipartite representations, including data

[3] I.e. by rerunning the phylogeny with our more diverse set of proteins and then rerunning the likelihood analysis without the anomalous species—of which we know there to have been at least one in the investigation described in Chapter 2.

describing multiple relationships.

What exact form these networks might take is still cryptic, and all the more interesting for it. Are we potentially looking at a sort of power-law-distributed epistasis, modulated by the distribution of numbers of participants in functional modules of various sorts, including physical protein clusters? It will be easier to begin an investigation of this type on the foothills, owing to the limitations of computing power which apply even in 2015. By “the foothills”, I mean that we might seek to begin with a comprehensive binary inference of pairwise correlated evolution, in much the manner whose methodology, as well as suggested enhancements from which it could benefit, are outlined in this report. We might go on to apply more complex transition matrices using combinations of multiple binary traits (i.e. protein presence/absence), to discern the outline of multiple interactions. Other subtleties might be more difficult to capture: what is the nature of a multiple interaction? Is it more like a phylogenetic tree?—essentially comprising not, as in the case of the phylogeny, two bifurcations giving rise first to a common ancestor and a terminal node and then two further terminal nodes from the common ancestor, but, analogously, a kind of hierarchy wherein protein A interacts with a protein complex containing proteins B and C, and arising from the latters’ prior interaction exclusively with each other? In other words, is the biomolecular network in general constrained to follow certain paths only to functional interdependence? How shall we discern whether this is the case, in general, using phylogenetic methods? How shall we represent and utilise the insights that might emerge? The line of investigation which I have reported on here is but one of many avenues, yet, with the growing understanding of phylogenetic methods among biologists and ecologists, the accessibility of powerful in-house and cloud computing resources, and a spreading appreciation in the wider scientific community for the aesthetic appeal and explanatory power of the network paradigm, it is an approach that might just be coming of age.

Appendix A

Network assembly: new rules?

A.1 Aims: Observing the rules

The question of what determines the topology of molecular networks within the cell remains open. An original aim of this project was to use the inferred history of the protein-protein correlated evolution network to test the various mechanisms proposed—prominent among which is the notion of preferential attachment, with or without selection acting on the topology itself. For example, it has been proposed that selection acts to preserve the scale-free character of protein interaction networks because such topologies are robust against perturbations, i.e. random nodal loss [Barabási and Albert, 1999]. Others argue that preferential attachment operates blindly, simply forging and maintaining scale-free molecular networks as an emergent consequence of node throughput (i.e. nodal migration into and out of the network) [Wagner, 2003].

Pagel, Meade and Scott [2007] have proposed a different model, which they characterise as "variable rates of attachment". Its distinguishing feature is that rates of acquisition of attachments within the network vary among proteins but not over time; in fact each protein has its own particular fixed rate of attachment in the network, which does not change in proportion to its existing degree. This contrasts with the "rich get richer" model of preferential attachment. These authors used ancestral state reconstruction to model changes over evolutionary time in a protein-protein interaction network based on correlated evolution. They derived a formula based on their model and adjusted its parameters so as to give a best fit to their protein degree data: they discovered that the model gave a better fit than a simple power law.

My intention was to extend the variable rates of attachment model by using my own correlated evolution network together with a revised formula describing network topology. Although this aspect of the project has not yet been fully implemented, I nonetheless feel that it is worthwhile taking the trouble to describe my proposed methodology. I argue that my proposed refinements might be used to challenge the ascendancy of preferential attachment in the academic discourse if, as I anticipate, the intuition that it is problematic in the biomolecular realm is well-founded.

A.2 Proposed methods: Breaking the rules

I propose a development of the equation of Pagel, Meade and Scott, to describe the emergence of network topology under the variable rates of attachment model. My approach is based on the incorporation of the equation for the Poisson distribution, which correctly describes the emergence of connectivity distribution in the (unrealistic) case of random, unbiased attachment. In terms of the usual variable used to denote nodal degree k, under random attachment the probability of any given node having degree k would be given by

$$p\left(k\right)=\frac{\bar{k}^{k}\mathrm{e}^{-\bar{k}}}{k!} \tag{A.1}$$

My desire is to answer as comprehensively and accurately as possible the question, "What is is the probability that a node picked out at random will possess a given integer connectivity (k) in an arbitrarily large randomly connected network, where the ratio of links to nodes, i.e. half the average nodal degree, or $\bar{k}/2$, is known?"

What the Poisson formula specifically provides is a set of probabilities for each value of k—which always sum to 1, whatever the value of $\bar{k}$; the formula is correct for the ideal of random attachment with uniform stickiness. The modified integral incorporating gamma stickiness which I present towards the end of this section also sums to 1 for all values of k.

The Poisson formula above is an accurate predictor of degree distribution under random attachment; although the random attachment model itself is in no way a description of biomolecular reality, the formula would give the degree distribution under purely hypothetical conditions of perfectly random attachment. To this extent it is the correct null model; compared with the purely exponential distribution $p\left(k\right) = \mathrm{e}^{-\lambda k}$, which is generally used to describe the probability that a node has a given degree under random attachment, its decline is steeper, but tends towards exponential as k increases well beyond $\bar{k}$[1]. Although the exponential distribution is widely discussed in the context of random networks, it is only an approximation for degree distribution under random

[1] Interestingly, in the special case that $\bar{k}$ takes an integer value, $p\left(k\right)$ has joint maxima at k and k−1

attachment, and at that, one which works well only in the tail.

I would therefore argue that if we seek to use the variable rates of attachment model to account for the (approximately) power-law distribution of protein degree in an organism, then our starting point, or null model, should be the correct model for degree distribution under random attachment, or Eq. (A.1) above, rather than the simplistic $p\left(k\right)=\mathrm{e}^{-\lambda k}$.

If protein stickiness (relative fixed rate of attachment, if you like) has a probability density conforming to a gamma distribution, and we denote the coefficient of stickiness as ζ (zeta), we get

$$f\left(\zeta\right)=\frac{\zeta^{\alpha-1}\mathrm{e}^{-\zeta/\beta}}{\beta^{\alpha}\Gamma\left(\alpha\right)} \tag{A.2}$$

If we also impose the restriction that the mean value of the stickiness coefficient itself ($\bar{\zeta}$) be unity, its effect can then be logically distinguished from that of average connectivity, $\bar{k}$. This gives us

$$\frac{\int f\left(\zeta\right)\cdot\zeta\cdot\mathrm{d}\zeta}{\int f\left(\zeta\right)\cdot\mathrm{d}\zeta}=1 \tag{A.3}$$

In fact, because we know that $\int f(\zeta)\cdot\mathrm{d}\zeta=1$, we can simply say

$$\int f\left(\zeta\right)\cdot\zeta\cdot\mathrm{d}\zeta=1 \tag{A.4}$$

In practice, I have found that Eq. (A.4) is satisfied when $\alpha\beta=1$. It also appears to be the case that, more generally, $\alpha\beta=\bar{\zeta}$. Under gamma stickiness plus random attachment, we now get a new function for degree distribution defined by the integral

$$p(k)=\int f\left(\zeta\right)\left(\frac{\left(\zeta\bar{k}\right)^{k}\mathrm{e}^{-\zeta\bar{k}}}{k!}\right)\cdot\mathrm{d}\zeta \tag{A.5}$$

or, to state it fully,

$$p\left(k\right)=\int\left(\frac{\zeta^{\alpha-1}\mathrm{e}^{-\zeta/\beta}}{\beta^{\alpha}\Gamma\left(\alpha\right)}\right)\left(\frac{\left(\zeta\bar{k}\right)^{k}\mathrm{e}^{-\zeta\bar{k}}}{k!}\right)\cdot\mathrm{d}\zeta \tag{A.6}$$

where $\alpha\beta=1$. More simply,

$$\begin{aligned} p\left(k\right) &= \int \left(\frac{\zeta^{\alpha-1}\mathrm{e}^{-\zeta\alpha}}{\left(1/\alpha\right)^{\alpha}\Gamma\left(\alpha\right)}\right)\left(\frac{\left(\zeta\bar{k}\right)^{k}\mathrm{e}^{-\zeta\bar{k}}}{k!}\right)\cdot\mathrm{d}\zeta \\ &= \int \left(\frac{\zeta^{\alpha-1}\mathrm{e}^{-\zeta\alpha}\alpha^{\alpha}}{\Gamma\left(\alpha\right)}\right)\left(\frac{\left(\zeta\bar{k}\right)^{k}\mathrm{e}^{-\zeta\bar{k}}}{k!}\right)\cdot\mathrm{d}\zeta \\ &= \int \left(\frac{\zeta^{\alpha-1}\mathrm{e}^{-\zeta\left(\alpha+\bar{k}\right)}\alpha^{\alpha}\left(\zeta\bar{k}\right)^{k}}{\Gamma\left(\alpha\right)k!}\right)\cdot\mathrm{d}\zeta \end{aligned} \tag{A.7}$$

It may be possible to solve this integral analytically, and I present a preliminary result below.

$$p\left(k\right) = \frac{\left(1+\alpha\right)^{-\beta-k}\Gamma\left(\beta+k\right)}{\alpha^{\beta}\Gamma\left(\beta\right)k!} \tag{A.8}$$

As far as I was able to progress empirical work on this, which owing to time constraints was not very far, I succeeded in obtaining a degree distribution curve which looks very like a power-law distribution. However, it is not *exactly* a power-law distribution, in that seemingly, its deviation from the linear signature of a power law on log-normal axes is systematic, having the form of a slightly concave function; in this notable regard it recapitulates the general shape of the plots in Figs. 4.7 to 4.9 on pages 71–73. I believe that if we are confident in, and wish to pursue, random attachment with variable rates, then it is important that we start from the right set of axioms. Arguably, my suggested variant of the formula for degree distribution under random attachment with uniform stickiness, which is explicitly derived from that of Pagel, Meade and Scott, is a tenable candidate for one such axiom.

My proposal was therefore to reformulate the variable rates of attachment model to incorporate the Poisson process, which has more explanatory power than a simple exponential model. Curve fitting to test the model could be performed numerically. Preferential attachment has itself been taken by some researchers as axiomatic, but the approach suggested here, in combination with the historical perspective offered by ancestral state reconstruction in the light of ever-improving data, has the power to lend sophistication and weight to the already attractive and highly plausible variable rates model.

Appendix B

Multiple alignment procedure: technical notes

MUSCLE can handle several genes, so we want ideally to include as many genes in the input as are shared among all species. This we can determine from the presence database, which records the output from our reciprocal BLAST. Look at the field `present_qty` as it applies to all species in the analysis; ideally, any species that are in the database but not required in the inputs for the phylogeny (e.g. multiple *Drosophila* species) are not required to have the gene, so one would need to look at the appropriate substrings of the field `presence_string`. It is quite likely that this would make no difference, but it would be better to check, as there is a possibility that more genes could be used.

`retrieve_and_align_multiple_genes.pl` was the original program which not only retrieved the proteins and RNAs, but went on to align the proteins by calling ClustalW (using BioPerl), before in turn calling AAtoDNA to spit out RNA alignments as well. So effectively this one program took the relevant genes and species and produced inputs ready for generating the phylogeny, thus incorporating several logical stages. Partly because we prefer to use MUSCLE over ClustalW, and partly because a process with discrete stages is easier to understand and debug, the preferred program is now `generate_rna_alignment_inputs.pl`, which generates only the protein and RNA FASTA files—one per gene, containing as many sequence records as there are species. We then use MUSCLE to produce protein alignments in FASTA format. Once the derived RNA alignment is generated, one has the option to chop out the third base in each codon (one would also chop out every third dash for gaps). This was implemented in `retrieve_and_align_multiple_genes.pl`, but is implemented separately for the MUSCLE alignments (see below); although it is not strictly necessary in either case, the third base is under relatively little constraint and carries correspondingly little phylogenetic signal. Leaving it in also increases run times.

Note that both `retrieve_and_align_multiple_genes.pl` and `generate_rna_alignment_inputs.pl` need editing for the desired genes. They also both require BioPerl 1.6 or higher; for some reason they cannot iterate through the species under BioPerl 1.5.2, although they might well work under BioPerl 1.5.1 as this was the environment within which the original code was developed. We can use MUSCLE (assuming it is installed—check this!) to generate the protein alignments. Run `muscle_align.sh` to generate files (one per gene) with the compound extension `.protein.aln.fasta`.

Note that to trim the alignments down after stripping the third base, we could have used the program Gblocks, but it is arcane and possibly too stringent even with its most relaxed settings, which is why a simple "topping and tailing" was preferred. Also, strictly speaking, NEXUS [Maddison et al., 1997] format input is required to infer the phylogeny, but this can be achieved by converting the output FASTA files. Alternatively, we could output the alignments in NEXUS format as easily as we currently do in FASTA, since we are using BioPerl.

Appendix C

Extracting ORF counts from GOLD

In order to extract ORF counts for the investigation reported in Chapter 5, I first needed access to the GOLD data, ideally in a way that would be reproducible by future investigators. In the Chapter 2 investigation, it had been possible to download the bulk of the then-current GOLD data in the form of a comma-separated-variable (CSV) file, from the GOLD web site. In April 2013 however (by which time the GOLD site had undergone a redesign as part of GOLD version 4.0 [Pagani et al., 2012]), although the documentation spoke of such a download, it was nowhere to be found[1]. An attempt to contact the GOLD maintainers by email, requesting a download link, was unsuccessful; ORF count and associated data by species appeared to be available on-screen only. I could simply have copied the data, which would have been time-consuming in itself, but arguably a pragmatic option if the experiment were not going to be designed to be repeated. However, favouring future reproducibility in as painless a way as possible, I instead decided to write a "screen scraper"[2] program in Perl, to automate the process of retrieving the ORF counts.

C.1 Issues with GOLD client-side code

Here I encountered some difficulties. I could specify that the GOLD results be limited to complete genomes, and in turn to eukaryotes only, with just a couple of interactive mouse clicks on the GOLD search page. However, the correct display of the page proved to be JavaScript-dependent [Flanagan, 2002]. In other words, because JavaScript runs within the browser rather than on the web server, client-side code was being executed to generate the requisite table of

[1] This appears to have been addressed: in June 2013 a CSV download link became available on the GOLD home page.

[2] A screen scraper is a program written to parse data embedded within HTML output designed for the screen rather than for data processing directly.

species names, ORF counts, sequencing centres, etc. Screen scraping typically relies on sending an HTTP [Leach et al., 1999] request to a web server and parsing the HTML markup wrapped up in the response; Perl offers several add-on libraries for processing this markup, transforming the various elements of the World Wide Web's tree-like "Document Object Model" (or DOM) into objects which are easy to handle programatically. The library I used was HTML::Tree. The problem in this case with sending a request to the web server (from within the Perl program) was that the markup returned was incomplete; JavaScript code is embedded in the markup, or at least called from tags embedded within the markup, and does not run until the response from the server has loaded, after which the DOM itself is transformed into the required state after rendering the table of species data. The difficulty is that Perl lacks a JavaScript interpreter (a feature common to any modern web browser) within which to run the downloaded code. Thus the markup seen by Perl is the intermediate version sent by the server and not the final, JavaScript-generated version which contains the species table. How then to generate the end-state DOM prior to parsing it with Perl?

C.2 Automating Firefox to populate code-dependent HTML

The solution was to use the browser itself to generate the finalised markup. Firefox in particular has a vast array of add-on programs available thanks to its popularity, longevity and open API[3], which makes it possible for a huge community of developers to make available their own independent enhancements to Firefox's out-of-the-box functionality. Traditionally, such code contributions were made in the form of plug-ins, which tend to be written in C, have their own process space and take over over from Firefox in the rendering of certain content—such content often being served in a proprietary format, exclusive to large software companies. Firefox still offers an API for plug-ins, but latterly, in an attempt to open the field to smaller-scale, incremental and distributed innovation, Firefox offers an alternative API for extensions, which run in Firefox's own interpreter, are written in JavaScript and XUL[4] and natively extend the browser's capabilities. One of the more unusual such extensions is moz-repl[5], which offers a command-line interface for issuing Javascript commands to the browser. This interface is exposed over a UNIX port to which one can attach using a simple Telnet [Postel and Reynolds, 1983] client. Essentially, one connects to port 4242 with the `telnet` command; after installing moz-repl 1.1 as an add-on to Firefox 19.0[6], Linux-derived

[3] Application Programming Interface

[4] "XML User interface Language" [Feldt, 2007], a dialect of XML markup developed by the Mozilla Foundation, which developed Firefox itself out of the old Mozilla browser code base.

[5] This is an elision of "Mozilla" and "REPL", itself an acronym for "Read-Eval-Print-Loop", which refers to a programming environment wherein a command is read and evaluated immediately before a response is returned to the calling interface, without an intervening compilation step—effectively, a command-line interpreter.

[6] moz-repl 1.1 does not work with Firefox 20.0 and upwards, which require moz-repl 1.1.2.

Telnet clients worked well in general[7] but the DOS command line version seemed prone to undiagnosable problems and errors. However, because the objective was still to use Perl for the screen scrape, I sought also to enable Telnet programatically. This was achieved with the use of another Perl module, Net::Telnet, which uses the same Telnet implementation on whichever platform it is run. In any case, it worked well with moz-repl. The process is not fully automatic, because it is necessary to open Firefox interactively to enable the moz-repl port. With Firefox open, the Perl program (`gold_digger.pl`) can be run, and the scrape performed after JavaScript has rendered the table. At one point the program pauses, because the user still has to click an option to retrieve all completed eukaryotic genomes; this step could probably have been automated but the use of a third-party JavaScript library[8] on the GOLD pages made the necessary JavaScript syntax somewhat cryptic. Nevertheless, using moz-repl to retrieve the resultant markup, it was possible to parse it as normal with HTML::Tree. This allowed me to extract values enclosed in the HTML tags and generate a CSV file containing ORF counts by taxon ID. The Perl code listing which implements this process, `gold_digger.pl`, is provided in Appendix G.

I then loaded the CSV file into a database table, which enabled the ORF figures to be compared with the protein counts from RefSeq by means of a simple SQL query.

[7] This includes clients originally written for Linux but ported to Windows—so moz-repl can be run successfully on a Windows system with a non-Microsoft Telnet client.

[8] YUI, the Yahoo User Interface library [Wellman, 2008].

Appendix D

Amazon EC2 BLAST with MIT StarCluster

From the total 35 922 human proteins (including the 13 human mitochondrial proteins), I identified 31 452 unique sequences using a simple SQL `select distinct` statement. The average size of target proteomes[1], based on a total of 1 079 891 non-human sequences (each one unique within the bounds of its own genome) and dividing by 84 gives 12 856, to the nearest whole number. After undertaking some performance testing on a home machine with a single core Intel Pentium M processor and 1 GB RAM, running Ubuntu Linux 13.04 with BLAST 2.2.27+ and BioPerl 1.6.9, it appeared that each reciprocal BLAST round, generating a single row of the trait matrix from a unique human source sequence, was taking about 10 to 12 minutes on average. Assuming an optimistic 10 minutes' run time per human protein, this would require just over 218 days, i.e. more than 7 months, to run on this one machine. Plainly, as before, I required the capacity to distribute the BLAST jobs across multiple nodes: I required a computing cluster.

Wishing to have both access to and control over such a cluster, I investigated a couple of options, including building my own ephemeral cluster using a dedicated master with slave nodes booting from CDs[2], but this did not seem easy, practical or a particularly good solution given the magnitude of the computation needed and the limited hardware resources available to me. My preferred avenue was therefore to investigate buying some cluster computing time as an online service. The recent trend towards "cloud computing" [Weinhardt et al., 2009] has seen a plethora of computing services become available to home, business and academic users on both free and paid-for bases. One of several such services provided by Amazon.com is its EC2 "Elastic Compute Cloud" service, which in simple terms offers arbitrarily large amounts of computing capacity in virtual Linux or

[1] As with the human (source) proteome, before creating the target proteomes, duplicate sequences were removed to save processing time. Unique instances can be mapped to potentially multiple RefSeq IDs by means of a database query.

[2] I.e. any machines used could in principle have continued to be used for their normal purposes when booted from their own hard disks (and not slaved to the cluster)—hence "ephemeral", rather than "virtual".

Windows environments running on Amazon's own servers "in the cloud". The compute instances offered come in a variety of default configurations, each of which may be further configured after purchase. Amazon also has a "free channel", which enables an individual or organisation free use for a year of a small compute instance of very limited memory and processing power. Nonetheless it is fully configurable as a web server, for example. So it provides a sandbox within which to experiment prior to purchasing time on a larger compute instance or group of instances. I secured a free instance with a view to exploring the possibilities.

I subsequently discovered StarCluster [Ragan-Kelley et al., 2013] from MIT, which provides an interface for neatly securing Amazon's cluster resources, and also enables bidding on nodes rather than paying Amazon's fixed prices. If one is prepared occasionally to be outbid, and for one's job to have to wait until the price falls, instances of very considerable computing power (e.g. 8 nodes and 32 GB of RAM) can be secured for a relatively low price. The type, number of nodes and bid can be specified on the StarCluster command line; StarCluster itself can be run locally or on a free channel Amazon instance, which was my preferred option. Configuration was difficult, e.g. setting up customised images to be propagated to each node in the cluster, but was accomplished with some care and attention to detail. Bidding proved to be quite a game—e.g. by default, StarCluster sets up instances of a single type, with one master and however many slave nodes one requires, in the Amazon cloud. One problem is that large instances are absolutely expensive, and the master node comes at a fixed (high) price rather than a bid price. Yet larger instances are often relatively cheap in terms of "bang for the buck". So one viable strategem is initially to buy a slaveless cluster with a small (cheap) master node at Amazon's fixed price, and add the other (large) nodes at a bid price later. This I did, before running my BLAST analysis in Amazon's cloud, using the two programs in Appendices H and I (the first of which serially batches the second to the Oracle Grid Engine[3] queue manager).

[3] Formerly the Sun Grid Engine [Gentzsch, 2001].

Appendix E

Visualising the network

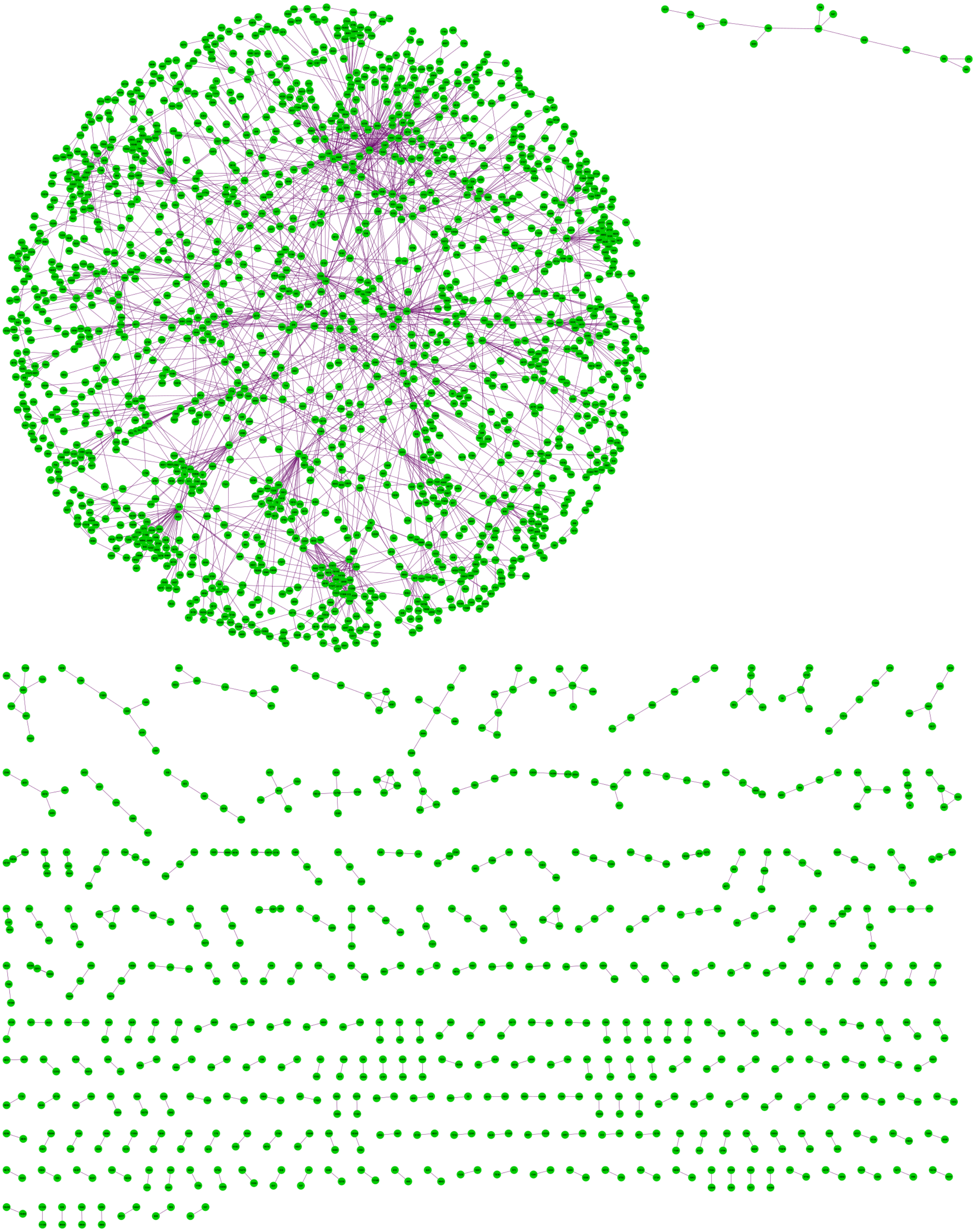


Figure E.1: Visualisation of correlated evolution network 2
(2 657 HPRD protein pairs filtered on control set)

Appendix F

collate_organisms_pg.pl

```
#!/usr/bin/perl

#  Changes made 17th August 2007:
#
#   1. The loop construct used to process both protein and rna files has been removed,
#      as we are interested only in proteins.
#   2. ”my” declarations sorted so they apply universally and exclusively to variables
#      used only within loops.
#   3. Braces stripped out of species’ filenames.
#
#  Changes made 20th August 2007:
#
#   4. Arrays of sequences now built and written out on a per−input−file, rather than
#      per−directory basis.
#   5. User now needs to supply MySQL username and passwords as first and second
#      command−line arguments, respectively.
#
#  Changes made 21st August 2007:
#
#   6. Embedded slashes in species names converted to underscores to allow files to
#      be written out successfully.
#
#  Changes made 4th September 2007:
#
#   6. Embedded colons in species names converted to underscores to allow files to
#      be written out successfully.
#
#  Changes made 9th October 2007:
#
#   7. Full stops in species names converted to underscores to allow files to
#      be written out successfully; multiple underscores converted to singles.
#
#  Changes made 12th May 2008:
#
#   8. $seqtype taken out of loop, as effectively it’s constant.
#
#  Changes made 17th December 2008:
#
#   9. $species (file stem) now has taxon id added in after a period, following
#      various searches and replaces on the name part of the stem.
#
#  Changes made 16th May 2009:
#
#   10. A count is no longer used to increment file names until the file is not
#       found, since some numbers are missing, and the process can’t skip over
#       the gaps, hopping to the next directory instead!
#
#  Changes made October 2012:
#
#   11. Adapted for PostgreSQL, and new RefSeq filename format.

use Cwd;
use Bio::SeqIO;
use DBI;

$dbh = DBI−>connect( ’DBI:Pg:dbname=evolution’, $ARGV[0], $ARGV[1] ) or die
    ’DB connection error: check your username/password combo!’;
$sth = $dbh−>prepare( ”select species_name, taxon_id from refseq.catalog where refseq_id = ?” );

$cwd = getcwd();
print ”$cwd\n\n”;

$seqtype = ’protein.faa’;
@phyla = qw( fungi invertebrate plant protozoa vertebrate_mammalian vertebrate_other microbial );

foreach my $phylum ( @phyla ) {

    print ”Processing $phylum species ...\n”;

    opendir( PHYLUMDIR, $phylum );
    my @seqfiles = grep { m/^$phylum\.\d+\.$seqtype$/ } readdir( PHYLUMDIR );
```

```perl
    foreach my $seqfile ( @seqfiles ) {

        # print "Parsing $seqfile ...\n";

        my %file_seqs = ();
        my $infile = Bio::SeqIO->new( -format => 'Fasta',
                                      -file   => "$phylum/$seqfile" );

        while ( my $sequence = $infile->next_seq ) {
            my $refseq_id = $sequence->display_id();
            $refseq_id =~ s/^.*ref\|//;
            $refseq_id =~ s/\|.*$//;

            $sth->execute( $refseq_id );
            my $species_ref = $sth->fetchrow_arrayref();
            my $species = $species_ref->[0];
            $species =~ s/(\s|\/|:|\.|\*)/_/g;
            $species =~ s/_+/_/g;
            $species =~ s/(\(|\)|\[|\]|')//g;
            $species .= '.' . $species_ref->[1];

            if ( exists $file_seqs{$species} ) {
                push ( @{$file_seqs{$species}}, $sequence );
            } else {
                $file_seqs{$species} = [ $sequence ];
            }
        }

        while ( ( my $key, my $value ) = each( %file_seqs ) ) {
            my $open_pref = '>';
            if ( -e "$phylum/$key\.$seqtype" ) {
                $open_pref = '>>';
            }
            my $species_file = Bio::SeqIO->new( -file   => "$open_pref$phylum/$key\.$seqtype",
                                                -format => 'Fasta' );

            foreach my $outseq ( @$value ) {
                $species_file->write_seq( $outseq );
            }
        }
    }
    closedir( PHYLUMDIR );
}
$sth->finish();
$dbh->disconnect();
print "Done.\n";
```

Appendix G

gold_digger.pl

```perl
#!/usr/bin/perl

#  gold_digger.pl
#
#  This script requires Firefox with moz-repl to be running on the local host.
#  Won't work if there are any other moz-repl sessions open, either!

use HTML::Tree;
use Net::Telnet;

my $goldmine = 'http://www.genomesonline.org/cgi-bin/GOLD/index.cgi' .
               '?page_requested=Complete+Genome+Projects&subset_requested=EUKARYAL';
my $mozza = new Net::Telnet( Host => 'localhost',
                             Port => 4242,
                             Timeout => 30,
                             Prompt => '/repl> /',
                             # Dump_log => 'repl_dump.log',
                             # Option_log => 'repl_opt.log',
                             # Input_log => 'repl_in.log',
                             # Output_log => 'repl_out.log',
                             Telnetmode => 0,
                             Cmd_remove_mode => 0 ) or die "Oh dear; no connection!";

my $ok = $mozza->cmd( "content.location.href = \"" . $goldmine . "\"" ) or die
    "Oh dear: command 1 failed!";

print "\nYou must now go to Firefox and select ALL eukaryotes before proceeding!";
print "\nHit <CR> when you're done ...";
my $stall = <STDIN>;
print "\n";

$ok = $mozza->cmd( "content.document.body.innerHTML" ) or die
    "Oh dear: HTML retrieval failed!";

#  It's weird, but you have to issue a second command to get the first's response.

my @pyrite = $mozza->cmd() or die "Oh dear: dummy command failed!";

=pod

These bits don't seem to work; onchange doesn't even work interactively, so some
weird event attachment model may be in use by the yui library.

$ok = $mozza->cmd( "content.document.getElementById('yui-pg0-0-rpp').selectedIndex = 5" );
$ok = $mozza->cmd( "content.document.getElementById('yui-pg0-0-rpp').onchange()" );
$ok = $mozza->cmd( "content.document.getElementById('yui-pg0-0-rpp').value = '1000'" );
=cut

$ok = $mozza->close() or die "Oh dear: could not close moz-repl session!";

my $linecnt = 0;
my $nuggetty = "";
foreach my $nugget ( @pyrite ) {
    chomp $nugget;
    $nuggetty .= $nugget;
    $linecnt++;
}
# print @pyrite;
# print $nuggetty;
print "\nNo of HTML lines in output: $linecnt\n";
print "No of array elements: " . scalar( @pyrite ) . "\n";
print "Length of HTML in characters: " . length($nuggetty) . "\n";
```

```perl

my $goldtree = HTML::Tree->new();
$goldtree->parse( $nuggetty );
my @rows = $goldtree->look_down( '_tag', 'tr' );
print "No of table rows: " . scalar( @rows ) . "\n";

my @todata = localtime(time);
$todata[5] += 1900;
$todata[4]++;
my $today = sprintf( "%04d%02d%02d", ( $todata[5], $todata[4], $todata[3] ) );
open( GOLDDATA, '>', "gold_${today}.data" );

my $cnt = 0;
foreach ( @rows ) {
    my @columns = $_->look_down( '_tag', 'td' );
    if ( scalar( @columns ) == 13 ) {

        my $species_element = $columns[1]->look_down( 'class', 'yui-dt-liner' );
        my $species_name = $species_element->as_text();

        my $taxon_element = $columns[3]->look_down( 'class', 'yui-dt-liner' );
        $taxon_element = $taxon_element->look_down( '_tag', 'a' );
        my $taxon_id = $taxon_element->attr( 'href' );
        $taxon_id =~ s/^.*=//;

        my $orf_element = $columns[4]->look_down( 'class', 'yui-dt-liner' );
        my $orf_qty = $orf_element->as_text();
        $orf_qty =~ s/^.*Kb//;
        $orf_qty =~ s/\s.*$//;

        #  Output only records with ORF counts - otherwise you'll get species duplicates
        $orf_qty ne "" && print GOLDDATA "$taxon_id\t$species_name\t$orf_qty\n";

        $cnt++;
    }
}

close( GOLDDATA );
print "Species count: $cnt\n";
```

Appendix H

cluster_presence_absence.pl

```
#!/usr/bin/perl −w

#  cluster_presence_absence.pl
#
#  This script takes an input list of RefSeq IDs, and for each one submits a
#  BioPerl reciprocal BLAST job by making a system call to the Oracle Grid
#  Engine. Jobs remain queued until a compute node becomes free; the more
#  available nodes, the faster the totality of jobs will be processed.

use Bio::Seq;
use Bio::SeqIO;

$root_species_ti = $ARGV[0]; #  Taxon ID only

@root_container = glob ”../combiner/preblast/*.$root_species_ti.unique.faa”;
# @root_container = glob ”/mnt/sgeadmin/preblast/*.$root_species_ti.unique.faa”;
$root_species_file = $root_container[0] or die ”Seq file not found for $root_species_ti”;

$root_sp_seqs_fh = Bio::SeqIO−>newFh( −format => 'fasta',
                                      −file   => $root_species_file );

$seq_count = 0;
while (<$root_sp_seqs_fh>) {
    $seq_count++;

    #  Set the following counters to process accessions within a range

    $seq_count >= 1 && $seq_count < 100000 &&
        system(
                ”qsub”, ”−V”, ”−b”, ”y”, ”−cwd”,
                ”perl”,
                ”onegene_presence_absence.pl”,
                $root_species_ti,
                ”\”” . $_−>display_id() . ”\””,
                $_−>seq(),
                $seq_count
              );
}
```

Appendix I

onegene_presence_absence.pl

```perl
#!/usr/bin/perl -w

#  onegene_presence_absence.pl
#
#  Runs reciprocal BLAST for one protein (gene product) at a time: should be
#  invoked iteratively within a batch script.

use Bio::Seq;
use Bio::Tools::Run::StandAloneBlastPlus;
use Sys::Hostname;

$hostname = hostname;
$hostname =~ s/\..*//;

$blast_db_dir = '../blast';
# $blast_db_dir = '/mnt/sgeadmin/blast';

#  For root species, supply taxon ID
( $root_species_ti, $this_id, $this_seq, $ordinality ) = @ARGV;

@root_db_container = glob "$blast_db_dir/*_$root_species_ti.phr";
$root_species_db = $root_db_container[0] or die "BLAST db not found for $root_species_ti";
$root_species_db =~ s/\.phr$//;

opendir( BLASTDIR, $blast_db_dir );

$open_pref = '>';
$prots_out = "${hostname}_proteins.${ordinality}.presence";
if ( -e $prots_out ) {
    $open_pref .= '>';
}
open( PROTSOUT, "${open_pref}$prots_out" );

$root_sp_seq_obj = Bio::Seq->new( -seq => $this_seq,
                                  -id  => $this_id );

$species_cnt = 0;
my @trait_list = ();

print "BLASTING $this_id ...\n";
SPECLOOP: foreach my $species_db ( glob "$blast_db_dir/*.phr" ) {
    $species_db =~ s/\.phr$//;
    ### print "BLASTing $species_db ...\n";

    if ( $species_db eq $root_species_db ) {
        $trait_list[$species_cnt] = 1;
    } else {
        $trait_list[$species_cnt] = &reciprocate( $root_sp_seq_obj, $root_species_db, $species_db );
    }

    $species_cnt++;
}

print PROTSOUT &get_accession( $root_sp_seq_obj, 3 );
print PROTSOUT "\t$_" foreach ( @trait_list );
print PROTSOUT "\n";

close PROTSOUT;
closedir BLASTDIR;

# print "Accessions processed: $acc_cnt\n";
print "Done.\n";


sub reciprocate {

    my ( $root_sp_seq_obj, $source_db, $target_db ) = @_;

    my $root_sp_accession = &get_accession( $root_sp_seq_obj, 3 );

    my $target_hit_accession = &blast_it( $root_sp_seq_obj, $target_db );

    my $reciprocal_hit_accession = '';

    my $target_hit_seq_obj;

    if ( $target_hit_accession ne '' ) {

        #  Ideally, you should use blastdbcmd from BLAST+ 2.2.28 or above.
        #  Install and use it locally if that's not your default release!

        open FASTA_IN, "blastdbcmd -dbtype prot -db $target_db -entry $target_hit_accession |";
        my $target_hit_seq_io = Bio::SeqIO->new( -format => 'fasta',
                                                 -fh     => \*FASTA_IN );
        $target_hit_seq_obj = $target_hit_seq_io->next_seq();
        close FASTA_IN;

        $reciprocal_hit_accession = &blast_it( $target_hit_seq_obj, $source_db );
    }

    my $present = ( $root_sp_accession eq $reciprocal_hit_accession || 0 );

    # print $species_cnt . ' ' . $present . ' ' . $source_db . ' ' . $target_db . ' ' .
    #       $root_sp_accession . ' ' . $target_hit_accession . ' ' .
    #       $reciprocal_hit_accession . "\n";

    return $present;
}


sub blast_it {

    my ( $source_seq_obj, $target_db ) = @_;

    # print $source_seq_obj->id() . " " . $target_db . "\n";

    my $hit_accession = '';
    my @params = (
        -db_name => $target_db
    );

    # map { print "$_\n" } @params;

    my $blast_obj = Bio::Tools::Run::StandAloneBlastPlus->new( @params );
    my $blast_result = $blast_obj->blastp(
        -query       => $source_seq_obj,
        -method_args => [
            '-evalue' => 0.000001
        ]
    );

    my $num_hits = $blast_result->num_hits();

    ### print $source_seq_obj->id() . " " . $target_db . " No. hits = " . $num_hits . "\n";

    if ( $num_hits > 0 ) {
        my $hit_obj = $blast_result->next_hit();
        $hit_accession = $hit_obj->name();
        my $hit_rank = $hit_obj->rank();
        my @hit_accession = split( /\|/, $hit_accession );

        #  This time we use 1 rather than 3 because only the RefSeq key-value
        #  pair was returned above.

        $hit_accession = $hit_accession[1];
        ### print "First hit accession is $hit_accession, ranked $hit_rank\n";
    }
    $blast_obj->cleanup();

    return $hit_accession;
}

```

```

sub get_accession {

    # Generally, acc_type 1 is a GI and acc_type 3 is a RefSeq ID. You probably
    # want the latter, for more straightforward compatibility with HPRD. And in
    # fact, it seems that either works with blastdbcmd.

    my ( $seq_obj, $acc_type ) = @_;

    my $accession = $seq_obj->display_id();
    my @accession = split( /\|/, $accession );
    $accession = $accession[$acc_type];

    return $accession;
}
```

Bibliography

Albert, R., Jeong, H. and Barabasi, A. [2000], 'Error and attack tolerance of complex networks', *Nature* **406**(6794), 378–82.
URL: *http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. [1990], 'Basic local alignment search tool.', *Journal of molecular biology* **215**(3), 403–10.
URL: *http://dx.doi.org/10.1016/S0022-2836(05)80360-2*

Artero, R., Furlong, E. E., Beckett, K., Scott, M. P. and Baylies, M. [2003], 'Notch and Ras signaling pathway effector genes expressed in fusion competent and founder cells during Drosophila myogenesis.', *Development (Cambridge, England)* **130**(25), 6257–72.
URL: *http://dev.biologists.org/content/130/25/6257.short*

Baldauf, S. L. [2008], 'An overview of the phylogeny and diversity of eukaryotes'.
URL: *http://uu.diva-portal.org/smash/record.jsf?pid=diva2:405839*

Baluska, F. [2010], 'Recent surprising similarities between plant cells and neurons.', *Plant signaling & behavior* **5**, 87–89.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2884105/*

Barabási, A. and Albert, R. [1999], 'Emergence of Scaling in Random Networks', *Science* **286**(5439), 509–512.
URL: *http://www.sciencemag.org/content/286/5439/509.abstract*

Barabási, A.-L. and Oltvai, Z. N. [2004], 'Network biology: understanding the cell's functional organization.', *Nature reviews. Genetics* **5**(2), 101–13.
URL: *http://dx.doi.org/10.1038/nrg1272*

Barker, D., Meade, A. and Pagel, M. [2007], 'Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes.', *Bioinformatics (Oxford, England)* **23**(1), 14–20.
URL: *http://bioinformatics.oxfordjournals.org/content/23/1/14.short*

Barker, D. and Pagel, M. [2005], 'Predicting functional gene links from phylogenetic-statistical analyses of whole genomes.', *PLoS computational biology* **1**(1), e3.
URL: *http://dx.plos.org/10.1371/journal.pcbi.0010003*

Barker, S. J., Tagu, D. and Delp, G. [1998], 'Regulation of Root and Fungal Morphogenesis in Mycorrhizal Symbioses', *PLANT PHYSIOLOGY* **116**(4), 1201–1207.
URL: *http://www.plantphysiol.org/content/116/4/1201.full*

Ben-Tabou de Leon, S. and Davidson, E. H. [2007], 'Gene regulation: gene control network in development.', *Annual review of biophysics and biomolecular structure* **36**, 191.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/17291181*

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. [2005], 'GenBank.', *Nucleic acids research* **33**(Database issue), D34–8.
URL: *http://nar.oxfordjournals.org/content/33/suppl_1/D34.short*

Braitenberg, V. and Schüz, A. [1998], *Cortex: Statistics and Geometry of Neuronal Connectivity*, Springer Berlin Heidelberg, Berlin, Heidelberg.
URL: *http://www.springerlink.com/index/10.1007/978-3-662-03733-1*

Brooks, R. A. [1990], 'Elephants don't play chess', *Robotics and Autonomous Systems* **6**(1-2), 3–15.
URL: *http://www.sciencedirect.com/science/article/pii/S0921889005800259*

Burne, T., Scott, E., van Swinderen, B., Hilliard, M., Reinhard, J., Claudianos, C., Eyles, D. and McGrath, J. [2011], 'Big ideas for small brains: what can psychiatry learn from worms, flies, bees and fish?', *Molecular psychiatry* **16**(1), 7–16.
URL: *http://dx.doi.org/10.1038/mp.2010.35*

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. [2009], 'BLAST+: architecture and applications.', *BMC bioinformatics* **10**(1), 421.
URL: *http://www.biomedcentral.com/1471-2105/10/421*

Carroll, S. B. [2001], 'Chance and necessity: the evolution of morphological complexity and diversity.', *Nature* **409**(6823), 1102–9.
URL: *http://dx.doi.org/10.1038/35059227*

Carroll, S. B. [2008], 'Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.', *Cell* **134**(1), 25–36.
URL: *http://www.sciencedirect.com/science/article/pii/S0092867408008179*

Castresana, J. [2000], 'Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis', *Molecular Biology and Evolution* **17**(4), 540–552.
URL: *http://mbe.oxfordjournals.org/content/17/4/540.short*

Chen, H.-M., Chen, J.-C., Ng, C.-J., Chiu, D.-F. and Chen, M.-F. [2006], 'Melatonin reduces pancreatic prostaglandins production and protects against caerulein-induced pancreatitis in rats.', *Journal of pineal research* **40**(1), 34–9.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/16313496*

Christoffels, A., Koh, E. G. L., Chia, J.-M., Brenner, S., Aparicio, S. and Venkatesh, B. [2004], 'Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes.', *Molecular biology and evolution* **21**(6), 1146–51.
URL: *http://mbe.oxfordjournals.org/content/21/6/1146.short*

Collins, J. J. and Chow, C. C. [1998], 'It's a small world.', *Nature* **393**(6684), 409–10.
URL: *http://europepmc.org/abstract/MED/9623993*

Dawkins, R. [1999], *The Extended Phenotype: The Long Reach of the Gene (Popular Science)*, Oxford Paperbacks.
URL: *http://www.amazon.co.uk/The-Extended-Phenotype-Popular-Science/dp/0192880519*

Dawkins, R. and Krebs, J. R. [1979], 'Arms Races between and within Species', *Proceedings of the Royal Society B: Biological Sciences* **205**(1161), 489–511.
URL: *http://rspb.royalsocietypublishing.org/content/205/1161/489.short*

Dehal, P. and Boore, J. L. [2005], 'Two rounds of whole genome duplication in the ancestral vertebrate.', *PLoS biology* **3**(10), e314.
URL: *http://dx.plos.org/10.1371/journal.pbio.0030314*

Dorogovtsev, S., Goltsev, A. and Mendes, J. [2002], 'Pseudofractal scale-free web', *Physical Review E* **65**(6), 066122.
URL: *http://link.aps.org/doi/10.1103/PhysRevE.65.066122*

Droms, R. [1997], 'Dynamic Host Configuration Protocol', *RFC2131* **March 1997**.
URL: *http://tools.ietf.org/html/rfc2131*

Eccles, J., Itō, M. and Szentägothai, J. [1967], *The cerebellum as a neuronal machine*, Springer-Verlag.
URL: *http://link.springer.com/book/10.1007%2F978-3-662-13147-3*

Eddy, S. R. [2004], 'What is dynamic programming?', *Nature biotechnology* **22**(7), 909–910.
URL: *http://www.nature.com/nbt/journal/v22/n7/full/nbt0704-909.html*

Edgar, R. C. [2004], 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic acids research* **32**(5), 1792–7.
URL: *http://nar.oxfordjournals.org/content/32/5/1792.short*

Edgar, R. C. and Batzoglou, S. [2006], 'Multiple sequence alignment.', *Current opinion in structural biology* **16**(3), 368–73.
URL: *http://www.sciencedirect.com/science/article/pii/S0959440X06000704*

Ezhova, T. A. [2007], 'Genetic control of early stages of leaf development', *Russian Journal of Developmental Biology* **38**(6), 363–373.
URL: *http://link.springer.com/10.1134/S1062360407060045*

Feldt, K. C. [2007], *Programming Firefox: Building Rich Internet Applications with XUL*, "O'Reilly Media, Inc.".
URL: *http://books.google.com/books?hl=en&lr=&id=ryEKOKnHFa8C*

Fields, S. and Song, O. [1989], 'A novel genetic system to detect protein-protein interactions.', *Nature* **340**(6230), 245–6.
URL: *http://europepmc.org/abstract/MED/2547163*

Flanagan, D. [2002], *JavaScript: The Definitive Guide*, "O'Reilly Media, Inc.".
URL: *http://books.google.com/books?hl=en&lr=&id=vJGlu9t9LNYC*

Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R. and Carucci, D. J. [2002], 'A proteomic view of the Plasmodium falciparum life cycle.', *Nature* **419**(6906), 520–6.
URL: *http://dx.doi.org/10.1038/nature01107*

Freckleton, R. P., Harvey, P. H. and Pagel, M. [2002], 'Phylogenetic analysis and comparative data: a test and review of evidence.', *The American naturalist* **160**(6), 712–26.
URL: *http://www.jstor.org/stable/10.1086/343873*

Fuchsman, C. A. and Rocap, G. [2006], 'Whole-genome reciprocal BLAST analysis reveals that planctomycetes do not share an unusually large number of genes with Eukarya and Archaea.', *Applied and environmental microbiology* **72**(10), 6841–4.
URL: *http://aem.asm.org/content/72/10/6841.short*

Gentzsch, W. [2001], Sun Grid Engine: towards creating a compute power grid, *in* 'Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid', IEEE Comput. Soc, pp. 35–36.
URL: *http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=923173*

Gish, W. [1996], 'Wu-blast'.

Goldberg, A. D., Allis, C. D. and Bernstein, E. [2007], 'Epigenetics: a landscape takes shape.', *Cell* **128**(4), 635–8.
URL: *http://www.sciencedirect.com/science/article/pii/S0092867407001869*

Gooch, S. [1972], *Total man: notes towards an evolutionary theory of personality*, London: Allen Lane, Penguin Press.
URL: *https://www.amazon.co.uk/Total-Man-Towards-Evolutionary-Personality/dp/071390237X*

Granovetter, M. S. [1973], 'The Strength of Weak Ties', *American Journal of Sociology* **78**(6), 1360.
URL: *http://ci.nii.ac.jp/naid/30017740362/en/*

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R. A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D. and Dubchak, I. [2012], 'The genome portal of the Department of Energy Joint Genome Institute.', *Nucleic acids research* **40**(Database issue), D26–32.
URL: *http://nar.oxfordjournals.org/content/40/D1/D26.short*

Grime, J. P. [2002], *Plant Strategies,Vegetation Processes 2e*, Wiley-Blackwell.
URL: *http://www.amazon.co.uk/Plant-Strategies-Vegetation-Processes-2e/dp/047085040X*

Guelzim, N., Bottani, S., Bourgine, P. and Képès, F. [2002], 'Topological and causal structure of the yeast transcriptional regulatory network.', *Nature genetics* **31**(1), 60–3.
URL: *http://dx.doi.org/10.1038/ng873*

Gunderson, J., Sogin, M., Wollett, G., Hollingdale, M., de la Cruz, V., Waters, A. and McCutchan, T. [1987], 'Structurally distinct, stage-specific ribosomes occur in Plasmodium', *Science* **238**(4829), 933–937.
URL: *http://www.sciencemag.org/content/238/4829/933.short*

Hall, T. [2005], 'BioEdit: biological sequence alignment editor for Win95'.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. and Vidal, M. [2004], 'Evidence for dynamically organized modularity in the yeast protein-protein interaction network.', *Nature* **430**(6995), 88–93.
URL: *http://dx.doi.org/10.1038/nature02555*

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. [2004], 'The Gene Ontology (GO) database and informatics resource.', *Nucleic acids research* **32**(Database issue), D258–61.
URL: *http://nar.oxfordjournals.org/content/32/suppl_1/D258.short*

Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. [1999], 'From molecular to modular cell biology.', *Nature* **402**(6761 Suppl), C47–52.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/10591225*

Harvey, P. and Pagel, M. [1991], *The comparative method in evolutionary biology*, Oxford University Press.
URL: *http://dannyreviews.com/h/The_Comparative_Method_in_Evolutionary_Biology.html*

Hashimoto, Y., Ito, Y., Niikura, T., Shao, Z., Hata, M., Oyama, F. and Nishimoto, I. [2001], 'Mechanisms of neuroprotection by a novel rescue factor humanin from Swedish mutant amyloid precursor protein.', *Biochemical and biophysical research communications* **283**(2), 460–8.
URL: *http://www.sciencedirect.com/science/article/pii/S0006291X01947655*

Hatefi, Y. [1985], 'The mitochondrial electron transport and oxidative phosphorylation system.', *Annual review of biochemistry* **54**, 1015–69.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/2862839*

Henikoff, S. and Henikoff, J. G. [1992], 'Amino acid substitution matrices from protein blocks.', *Proceedings of the National Academy of Sciences* **89**(22), 10915–10919.
URL: *http://www.pnas.org/content/89/22/10915.short*

Herder, F. and Freyhof, J. [2006], 'Resource partitioning in a tropical stream fish assemblage', *Journal of Fish Biology* **69**(2), 571–589.
URL: *http://doi.wiley.com/10.1111/j.1095-8649.2006.01126.x*

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. and Apweiler, R. [2004], 'The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data.', *Nature biotechnology* **22**(2), 177–83.
URL: *http://dx.doi.org/10.1038/nbt926*

Holland, R. C. G., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., Foisy, S., Dräger, A., Yates, A., Heuer, M. and Schreiber, M. J. [2008], 'BioJava: an open-source framework for bioinformatics.', *Bioinformatics (Oxford, England)* **24**(18), 2096–7.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2530884/*

Hornbaker, D., Albert, R., Albert, I., Barabasi, A. and Schiffer, P. [1997], 'What keeps sandcastles standing?', *Nature* **387**.
URL: *http://www.nature.com/nature/journal/v387/n6635/full/387765a0.html*

Hubbard, T. [2002], 'The Ensembl genome database project', *Nucleic Acids Research* **30**(1), 38–41.
URL: *http://nar.oxfordjournals.org/content/30/1/38.short*

Hutter, H. [2000], 'Conservation and Novelty in the Evolution of Cell Adhesion and Extracellular Matrix Genes', *Science* **287**(5455), 989–994.
URL: *http://www.sciencemag.org/content/287/5455/989.full*

Ichikawa, J. J. and Steup, M. [2014], The Analysis of Knowledge, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spr. 2014 edn, Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA 94305.
URL: *http://plato.stanford.edu/archives/spr2014/entries/knowledge-analysis/*

Jeong, H., Néda, Z. and Barabási, A. L. [2003], 'Measuring preferential attachment in evolving networks', *Europhysics Letters (EPL)* **61**(4), 567–572.
URL: *http://dx.doi.org/10.1209/epl/i2003-00166-9*

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A. L. [2000], 'The large-scale organization of metabolic networks.', *Nature* **407**(6804), 651–4.
URL: *http://dx.doi.org/10.1038/35036627*

Jin, E., Girvan, M. and Newman, M. [2001], 'Structure of growing social networks', *Physical Review E* **64**(4), 046132.
URL: *http://link.aps.org/doi/10.1103/PhysRevE.64.046132*

Judy, K. J. and Gilbert, L. I. [1969], 'Morphology of the alimentary canal during the metamorphosis of Hyalophora cecropia (Lepidoptera: Saturniidae).', *Annals of the Entomological Society of America* **62**(6), 1438–46.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/5374180*

Kauffman, S. [1995], 'Random chemistry', *Perspectives in Drug Discovery and Design* **2**(2), 319–326.
URL: *http://link.springer.com/10.1007/BF02172070*

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. and Apweiler, R. [2004], 'The International Protein Index: an integrated database for proteomics experiments.', *Proteomics* **4**(7), 1985–8.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/15221759*

Kofler, M. [2001], *MySQL*, Apress, Berkeley, CA.
URL: *http://link.springer.com/10.1007/978-1-4302-0853-2*

Kooij, T. W. and Matuschewski, K. [2007], 'Triggers and tricks of Plasmodium sexual development.', *Current opinion in microbiology* **10**(6), 547–53.
URL: *http://www.sciencedirect.com/science/article/pii/S1369527407001403*

Korf, I. [2003], 'Serial BLAST searching', *Bioinformatics* **19**(12), 1492–1496.
URL: *http://bioinformatics.oxfordjournals.org/content/19/12/1492.abstract*

Lasek-Nesselquist, E. and Gogarten, J. P. [2013], 'The effects of model choice and mitigating bias on the ribosomal tree of life.', *Molecular phylogenetics and evolution* **69**(1), 17–38.
URL: *http://www.sciencedirect.com/science/article/pii/S1055790313002066*

Leach, P. J., Berners-Lee, T., Mogul, J. C., Masinter, L., Fielding, R. T. and Gettys, J. [1999], 'Hypertext Transfer Protocol – HTTP/1.1', *RFC2616* **June 1999**.
URL: *https://tools.ietf.org/html/rfc2616*

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., Srivastava, S., Uhlén, M., Wu, C. H., Yamamoto, T., Paik, Y.-K. and Omenn, G. S. [2011], 'The human proteome project: current state and future direction.', *Molecular & cellular proteomics : MCP* **10**(7), M111.009993.
URL: *http://www.mcponline.org/content/10/7/M111.009993.short*

Lehár, J., Krueger, A., Zimmermann, G. and Borisy, A. [2008], 'High-order combination effects and biological robustness.', *Molecular systems biology* **4**, 215.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2538911/*

Leicht, E. A., Clarkson, G., Shedden, K. and Newman, M. E. [2007], 'Large-scale structure of time evolving citation networks', *The European Physical Journal B* **59**(1), 75–83.
URL: *http://www.springerlink.com/index/10.1140/epjb/e2007-00271-7*

Lemons, D. and McGinnis, W. [2006], 'Genomic evolution of Hox gene clusters.', *Science (New York, N.Y.)* **313**(5795), 1918–22.
URL: *http://science.sciencemag.org/content/313/5795/1918*

Levin, M., Hashimshony, T., Wagner, F. and Yanai, I. [2012], 'Developmental milestones punctuate gene expression in the Caenorhabditis embryo.', *Developmental cell* **22**(5), 1101–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22560298*

Lin, D. N. C. and Papaloizou, J. [1985], 'On the dynamical origin of the solar system', *Protostars and Planets II* **-1**, 981–1072.
URL: *http://adsabs.harvard.edu/abs/1985prpl.conf..981L*

Linsley, E. [1958], 'The ecology of solitary bees', *Hilgardia* **27**, 543 – 599.

Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyrpides, N. C. [2006], 'The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide.', *Nucleic acids research* **34**(Database issue), D332–4.
URL: *http://nar.oxfordjournals.org/content/34/suppl_1/D332.short*

Lloyd, D., Aon, M. A. and Cortassa, S. [2001], 'Why homeodynamics, not homeostasis?', *TheScientificWorldJournal* **1**, 133–45.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/12805697*

Luo, H., Lin, K., David, A., Nijveen, H. and Leunissen, J. A. M. [2012], 'ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins.', *Nucleic Acids Research* **40**(Database issue), D394–9.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245022/*

Lynch, M. and Conery, J. S. [2000], 'The evolutionary fate and consequences of duplicate genes.', *Science (New York, N.Y.)* **290**(5494), 1151–5.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/11073452*

Maddison, D. R., Swofford, D. L. and Maddison, W. P. [1997], 'Nexus: An Extensible File Format for Systematic Information', *Systematic Biology* **46**(4), 590–621.
URL: *http://sysbio.oxfordjournals.org/content/46/4/590.short*

Mann, J., ed. [2000], *Cetacean Societies: Field Studies of Dolphins and Whales*, University of Chicago Press.
URL: *http://books.google.com/books?hl=en&lr=&id=W-UQNoxMONwC*

Marino, J., Sillero-Zubiri, C., Johnson, P. J. and Macdonald, D. W. [2012], 'Ecological bases of philopatry and cooperation in Ethiopian wolves', *Behavioral Ecology and Sociobiology* **66**(7), 1005–1015.
URL: *http://link.springer.com/10.1007/s00265-012-1348-x*

Marino, L. [2002], 'Convergence of complex cognitive abilities in cetaceans and primates.', *Brain, behavior and evolution* **59**(1-2), 21–32.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/12097858*

McCormick, F. [2003], Signal Transduction Networks, *in* 'Oncogene-Directed Therapies', Springer, pp. 35–46.
URL: *http://link.springer.com/chapter/10.1007/978-1-59259-313-2_3*

McGinnis, S. and Madden, T. L. [2004], 'BLAST: at the core of a powerful and diverse set of sequence analysis tools.', *Nucleic acids research* **32**(Web Server issue), W20–5.
URL: *http://nar.oxfordjournals.org/content/32/suppl_2/W20.full*

Medvedev, Z. A. [1995], *Molecular-Genetic Mechanisms of Development*, Springer US, Boston, MA.
URL: *http://www.springerlink.com/index/10.1007/978-1-4684-1764-7*

Mercer, W. [1900], 'The development of the wings in the Lepidoptera', *Journal of the New York Entomological Society* .
URL: *http://www.jstor.org/stable/25002881*

Merriman, B., R&D Team, I. T. and Rothberg, J. M. [2012], 'Progress in Ion Torrent semiconductor chip based sequencing', *ELECTROPHORESIS* **33**(23), 3397–3417.
URL: *http://doi.wiley.com/10.1002/elps.201200424*

Meshorer, E., Yellajoshula, D., George, E., Scambler, P. J., Brown, D. T. and Misteli, T. [2006], 'Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells.', *Developmental cell* **10**(1), 105–16.
URL: *http://www.sciencedirect.com/science/article/pii/S1534580705004211*

Mikhailov, K. V., Konstantinova, A. V., Nikitin, M. A., Troshin, P. V., Rusin, L. Y., Lyubetsky, V. A., Panchin, Y. V., Mylnikov, A. P., Moroz, L. L., Kumar, S. and Aleoshin, V. V. [2009], 'The origin of Metazoa: a transition from temporal to spatial cell differentiation.', *BioEssays : news and reviews in molecular, cellular and developmental biology* **31**(7), 758–68.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/19472368*

Milgram, S. [1967], 'The small world problem', *Psychology Today* .
URL: *http://ci.nii.ac.jp/naid/10018230069/en/*

Momjian, B. [2001], *PostgreSQL: introduction and concepts*, Addison-Wesley Longman Publishing Co., Inc.
URL: *http://dl.acm.org/citation.cfm?id=359414*

Needleman, S. B. and Wunsch, C. D. [1970], 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology* **48**(3), 443–453.
URL: *http://www.sciencedirect.com/science/article/pii/0022283670900574*

Newell, A. [1994], *Unified Theories of Cognition*, Harvard University Press.
URL: *http://books.google.com/books?hl=en&lr=&id=1lbY14DmV2cC*

Newman, M. [2001], 'Clustering and preferential attachment in growing networks', *Physical Review E* **64**(2), 025102.
URL: *http://link.aps.org/doi/10.1103/PhysRevE.64.025102*

Nimchinsky, E. A., Gilissen, E., Allman, J. M., Perl, D. P., Erwin, J. M. and Hof, P. R. [1999], 'A neuronal morphologic type unique to humans and great apes', *Proceedings of the National Academy of Sciences* **96**(9), 5268–5273.
URL: *http://www.pnas.org/content/96/9/5268.full.pdf&a=bi&pagenumber=1&w=100*

Ohno, S. [1970], *Evolution by gene duplication.*, London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.
URL: *http://link.springer.com/book/10.1007%2F978-3-642-86659-3*

O'Keefe, D. D., Thomas, S. R., Bolin, K., Griggs, E., Edgar, B. A. and Buttitta, L. A. [2012], 'Combinatorial control of temporal gene expression in the Drosophila wing by enhancers and core promoters', *BMC Genomics* **13**(1), 498.
URL: *http://www.biomedcentral.com/1471-2164/13/498*

Omland, K. E. [1999], 'The Assumptions and Challenges of Ancestral State Reconstructions', *Systematic biology* **48**(3), 604–611.
URL: *http://www.jstor.org/stable/2585327*

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M. and Kyrpides, N. C. [2012], 'The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.', *Nucleic acids research* **40**(Database issue), D571–9.
URL: *http://nar.oxfordjournals.org/content/40/D1/D571.short*

Pagel, M. [1994], 'Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters', *Proceedings of the Royal Society B Biological Sciences* **255**(1342), 37–45.
URL: *http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.1994.0006*

Pagel, M. and Meade, A. [2004], 'A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data.', *Systematic biology* **53**(4), 571–81.
URL: *http://sysbio.oxfordjournals.org/content/53/4/571.short*

Pagel, M. and Meade, A. [2006], 'BayesPhylogenies 1.0', *Software distributed by the authors* .
URL: *http://www.evolution.rdg.ac.uk*

Pagel, M. and Meade, A. [2007], 'BayesTraits, version 1.0', *Software distributed by the authors* .
URL: *http://www.evolution.rdg.ac.uk*

Pagel, M., Meade, A. and Scott, D. [2007], 'Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes.', *BMC evolutionary biology* **7 Suppl 1**(Suppl 1), S16.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1796610/*

Pasquinelli, A. E. and Ruvkun, G. [2002], 'Control of developmental timing by micrornas and their targets.', *Annual review of cell and developmental biology* **18**, 495–513.
URL: *http://www.annualreviews.org/doi/abs/10.1146/annurev.cellbio.18.012502.105832*

Pearson, W. R. and Lipman, D. J. [1988], 'Improved tools for biological sequence comparison.', *Proceedings of the National Academy of Sciences* **85**(8), 2444–2448.
URL: *http://www.pnas.org/content/85/8/2444.short*

Pellegrini, M., Haynor, D. and Johnson, J. M. [2004], 'Protein interaction networks.', *Expert review of proteomics* **1**(2), 239–49.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/15966818*

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K. B., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S. and Pandey, A. [2004], 'Human protein reference database as a discovery resource for proteomics.', *Nucleic acids research* **32**(Database issue), D497–501.
URL: *http://nar.oxfordjournals.org/content/32/suppl_1/D497.short*

Pertea, M. and Salzberg, S. L. [2010], 'Between a chicken and a grape: estimating the number of human genes.', *Genome biology* **11**(5), 206.
URL: *http://genomebiology.com/2010/11/5/206*

Postel, J. and Reynolds, J. [1983], 'Telnet Protocol Specification', *RFC854* **May 1983**.
URL: *https://tools.ietf.org/html/rfc854*

Pruess, M., Kersey, P. and Apweiler, R. [2005], 'The Integr8 project–a resource for genomic and proteomic data.', *In silico biology* **5**(2), 179–85.
URL: *http://europepmc.org/abstract/MED/15972013*

Pruitt, K. D., Tatusova, T. and Maglott, D. R. [2005], 'NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.', *Nucleic acids research* **33**(Database issue), D501–4.
URL: *http://nar.oxfordjournals.org/content/33/suppl_1/D501.short*

Ragan-Kelley, B., Walters, W. A., McDonald, D., Riley, J., Granger, B. E., Gonzalez, A., Knight, R., Perez, F. and Caporaso, J. G. [2013], 'Collaborative cloud-enabled tools allow rapid, reproducible biological insights.', *The ISME journal* **7**(3), 461–4.
URL: *http://dx.doi.org/10.1038/ismej.2012.123*

Raguso, R. A. [2008], 'Start making scents: the challenge of integrating chemistry into pollination ecology', *Entomologia Experimentalis et Applicata* **128**(1), 196–207.
URL: *http://doi.wiley.com/10.1111/j.1570-7458.2008.00683.x*

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A. L. [2002], 'Hierarchical organization of modularity in metabolic networks.', *Science (New York, N.Y.)* **297**(5586), 1551–5.
URL: *http://www.sciencemag.org/content/297/5586/1551.abstract*

Read, W. J. [2007], Special Report, Technical report, University of Reading.

Rice, P., Longden, I. and Bleasby, A. [2000], 'EMBOSS: the European Molecular Biology Open Software Suite.', *Trends in genetics : TIG* **16**(6), 276–7.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/10827456*

Schütte, J., Moignard, V. and Göttgens, B. [2012], 'Establishing the stem cell state: insights from regulatory network analysis of blood stem cell development.', *Wiley interdisciplinary reviews. Systems biology and medicine* **4**(3), 285–95.
URL: *http://europepmc.org/abstract/MED/22334489*

Scribner, K. and Stiver, M. C. [2000], *Understanding Soap: Simple Object Access Protocol*, Sams.
URL: *http://dl.acm.org/citation.cfm?id=556856*

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. [2003], 'Cytoscape: a software environment for integrated models of biomolecular interaction networks.', *Genome research* **13**(11), 2498–504.
URL: *http://genome.cshlp.org/content/13/11/2498.short*

Shubin, N. H. and Marshall, C. R. [2009], 'Fossils, genes, and the origin of novelty'.
URL: *http://www.jstor.org/stable/1571664*

Singh, A. B. and Harris, R. C. [2005], 'Autocrine, paracrine and juxtacrine signaling by EGFR ligands.', *Cellular signalling* **17**(10), 1183–93.
URL: *http://www.sciencedirect.com/science/article/pii/S0898656805000768*

Slack, J. M., Holland, P. W. and Graham, C. F. [1993], 'The zootype and the phylotypic stage.', *Nature* **361**(6412), 490–2.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/8094230*

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L. and Lewis, S. [2007], 'The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.', *Nature biotechnology* **25**(11), 1251–5.
URL: *http://dx.doi.org/10.1038/nbt1346*

Smith, J. C. [1989], 'Mesoderm induction and mesoderm-inducing factors in early amphibian development.', *Development (Cambridge, England)* **105**(4), 665–77.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/2689132*

Soltis, P. S. and Soltis, D. E. [2009], 'The role of hybridization in plant speciation.', *Annual review of plant biology* **60**, 561–88.
URL: *http://www.annualreviews.org/doi/abs/10.1146/annurev.arplant.043008.092039*

Sorek, R., Shamir, R. and Ast, G. [2004], 'How prevalent is functional alternative splicing in the human genome?', *Trends in genetics : TIG* **20**(2), 68–71.
URL: *http://www.sciencedirect.com/science/article/pii/S0168952503003433*

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. [2002], 'The Bioperl toolkit: Perl modules for the life sciences.', *Genome research* **12**(10), 1611–8.
URL: *http://genome.cshlp.org/content/12/10/1611.short*

Stewart, J. B., Freyer, C., Elson, J. L., Wredenberg, A., Cansu, Z., Trifunovic, A. and Larsson, N.-G. [2008], 'Strong purifying selection in transmission of mammalian mitochondrial DNA.', *PLoS biology* **6**(1), e10.
URL: *http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060010*

Stone, G. N., Nee, S. and Felsenstein, J. [2011], 'Controlling for non-independence in comparative analysis of patterns across populations within species.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **366**(1569), 1410–24.
URL: *http://rstb.royalsocietypublishing.org/content/366/1569/1410.short*

Tatusov, R. L. [1997], 'A Genomic Perspective on Protein Families', *Science* **278**(5338), 631–637.
URL: *http://www.sciencemag.org/content/278/5338/631.short*

Theraulaz, G., Bonabeau, E. and Deneubourg, J.-L. [1998], 'The origin of nest complexity in social insects', *Complexity* **3**(6), 15–25.
URL: *http://www.santafe.edu/research/working-papers/abstract/77afdfc30583b2eaacab31a58477425b/*

Thomopson, J. D., Higgins, D. G. and Gibson, T. J. [1994], 'ClustalW', *Nucleic Acids Res* **22**, 4673–4680.

Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Breniere, S. F., Darde, M. L. and Ayala, F. J. [1991], 'Are eukaryotic microorganisms clonal or sexual? A population genetics vantage.', *Proceedings of the National Academy of Sciences* **88**(12), 5129–5133.
URL: *http://www.pnas.org/content/88/12/5129.short*

Tortora, G. J. and Derrickson, B. H. [2011], *Principles of Anatomy and Physiology International Student Version (2 Volume Set) (Isv 13th Edition)*, John Wiley & Sons.
URL: *http://www.amazon.co.uk/Principles-Anatomy-Physiology-International-Student/dp/0470929189*

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M. and Wodak, S. J. [2010], 'Literature curation of protein interactions: measuring agreement across major public databases.', *Database : the journal of biological databases and curation* **2010**, baq026.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21183497*

Vickaryous, M. K. and Hall, B. K. [2006], 'Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest.', *Biological reviews of the Cambridge Philosophical Society* **81**(3), 425–55.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/16790079*

Waddington, C. H. [1957], *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, London: George Allen & Unwin, Ltd.
URL: *https://books.google.co.uk/books/about/The_strategy_of_the_genes.html?id=PdU9AAAAIAAJ*

Wagner, A. [1998], 'The fate of duplicated genes: loss or new function?', *BioEssays : news and reviews in molecular, cellular and developmental biology* **20**(10), 785–8.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/10200118*

Wagner, A. [2003], 'How the global structure of protein interaction networks evolves.', *Proceedings. Biological sciences / The Royal Society* **270**(1514), 457–66.
URL: *http://rspb.royalsocietypublishing.org/content/270/1514/457.short*

Watts, D. [2004], 'The New science of networks', *Annual review of sociology* **30**, 243 – 270.
URL: *http://www.jstor.org/stable/29737693*

Watts, D. J. and Strogatz, S. H. [1998], 'Collective dynamics of 'small-world' networks.', *Nature* **393**(6684), 440–2.
URL: *http://dx.doi.org/10.1038/30918*

Weinhardt, C., Anandasivam, A., Blau, B., Borissov, N., Meinl, T., Michalk, W. and Stößer, J. [2009], 'Cloud Computing – A Classification, Business Models, and Research Directions', *Business & Information Systems Engineering* **1**(5), 391–399.
URL: *http://link.springer.com/10.1007/s12599-009-0071-2*

Wellman, D. [2008], *Learning the Yahoo! User Interface Library: Get Started and Get to Grips with the YUI JavaScript Development Library!*, Packt Publishing Ltd.
URL: *http://books.google.com/books?hl=en&lr=&id=omJDJJUa7rUC*

White, C., Hightower, L. and Schultz, R. [1994], 'Variation in heat-shock proteins among species of desert fishes (Poeciliidae, Poeciliopsis)', *Mol. Biol. Evol.* **11**(1), 106–119.
URL: *http://mbe.oxfordjournals.org/content/11/1/106.short*

White, J. G., Southgate, E., Thomson, J. N. and Brenner, S. [1986], 'The Structure of the Nervous System of the Nematode Caenorhabditis elegans', *Philosophical Transactions of the Royal Society B: Biological Sciences* **314**(1165), 1–340.
URL: *http://rstb.royalsocietypublishing.org/content/314/1165/1.abstract*

Wilson, E. O. [1990*a*], *Success and dominance in ecosystems: the case of the social insects.*, Ecology Institute.
URL: *http://onlinelibrary.wiley.com/doi/10.1002/iroh.19920770320/abstract*

Wilson, E. O. [1990*b*], *The Insect Societies (Harvard paperbacks)*, Harvard University Press.
URL: *http://www.amazon.co.uk/The-Insect-Societies-Harvard-paperbacks/dp/0674454952*

Withers, G. S., Fahrbach, S. E. and Robinson, G. E. [1993], 'Selective neuroanatomical plasticity and division of labour in the honeybee.', *Nature* **364**(6434), 238–40.
URL: *http://dx.doi.org/10.1038/364238a0*

Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R., Cheng, T., Jiang, T., Becquet, C., Xu, X., Liu, C., Zha, X., Fan, W., Lin, Y., Shen, Y., Jiang, L., Jensen, J., Hellmann, I., Tang, S., Zhao, P., Xu, H., Yu, C., Zhang, G., Li, J., Cao, J., Liu, S., He, N., Zhou, Y., Liu, H., Zhao, J., Ye, C., Du, Z., Pan, G., Zhao, A., Shao, H., Zeng, W., Wu, P., Li, C., Pan, M., Li, J., Yin, X., Li, D., Wang, J., Zheng, H., Wang, W., Zhang, X., Li, S., Yang, H., Lu, C., Nielsen, R., Zhou, Z., Wang, J., Xiang, Z. and Wang, J. [2009], 'Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx).', *Science (New York, N.Y.)* **326**(5951), 433–6.
URL: *http://www.sciencemag.org/content/326/5951/433.short*

Xue, L., Cai, J.-Y., Ma, J., Huang, Z., Guo, M.-X., Fu, L.-Z., Shi, Y.-B. and Li, W.-X. [2013], 'Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis.', *BMC genomics* **14**(1), 568.
URL: *http://www.biomedcentral.com/1471-2164/14/568*

Zhang, J. [2003], 'Evolution by gene duplication: an update', *Trends in Ecology & Evolution* **18**(6), 292–298.
URL: *http://www.sciencedirect.com/science/article/pii/S0169534703000338*