From the Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

ANALYSES OF PROTEIN EVOLUTION, FUNCTION, AND ARCHITECTURE

Anna Henricson

Karolinska Institutet

Stockholm 2010

Printed by Larserics Digital Print AB, Sundbyberg, Sweden

ISBN 978-91-7409-753-5

ABSTRACT

Proteins can evolve over time in many different ways. An ancestral protein sequence inherited in different species will gradually undergo changes in primary sequence and sometimes in domain architecture. Some of these changes will affect its function, and evolutionary analyses can be used to predict function shift. A common paradigm is that orthologs, *i.e.* genes in different species that derive from the same gene in the last common ancestor, are functional counterparts. Orthology is a special case of the more general concept homology, which means any form of shared ancestry. This thesis investigates the functional conservation of orthologs compared to non-orthologs, and further explores gene and protein domain architectural changes during evolution.

A set of 17 proteins were selected between human and the nematode *C. elegans* such that they were predicted to be orthologous, membrane-spanning, and did not have a known function. By experimental studies in the nematode, functional clues were obtained for 12 of them that thus have high relevance for the human orthologs. Several of the genes were expressed in the nervous system. One of them was a presenilin-like protein, which was subjected to further bioinformatic analysis, including prediction of its transmembrane topology. Mutations in presenilin are known to cause Alzheimer's disease, the main type of dementia in humans. Resolving the molecular structure of presenilin has not been possible yet because it is a transmembrane protein. Instead, many attempts to elucidate the transmembrane topology biochemically have been made, but the results were often contradictory. We therefore approached the problem by reconciling the output from several transmembrane topology predictors and previously published experimental studies. This allowed us to propose a novel nine-transmembrane topology with the C-terminus located in the extracytosolic space, which has subsequently been verified by several other researchers.

To study the evolution of protein domain architecture we developed a new algorithm based on the maximum parsimony criterion to infer ancestral architectures. We analyzed 96 species across all kingdoms to find cases where a domain architecture had been created multiple times independently. In contrast to previous studies we found that such events are relatively frequent, up to 12.4%. Among the architectures displaying reinvention we could find no strong functional bias, implying that it is a widespread phenomenon.

In this thesis, the focus is on evolutionary analysis and applying it when investigating various aspects of protein function and architecture. Incorporating new discriminating features is important to further enhance the accuracy of phylogenetic inference. To this end, we investigated conservation of intron positions among orthologs versus non-orthologs that are equally similar in sequence. We found that ortholog-ortholog gene pairs on average have a significantly higher degree of intron position conservation compared to ortholog-closest non-orthologs. This implies that shared intron positions could be used as an additional discriminating feature in evolutionary analysis.

LIST OF PUBLICATIONS

I. **Henricson A**, Sonnhammer EL, Baillie DL, and Gomes AV.

Functional characterization in Caenorhabditis elegans of transmembrane worm-human orthologs.
BMC Genomics. 2004, 5:85.

II. **Henricson A**, Käll L, and Sonnhammer EL.

A novel transmembrane topology of presenilin based on reconciling experimental and computational evidence.
FEBS Journal. 2005, 272:2727-2733.

III. Forslund K*, **Henricson A***, Hollich V, and Sonnhammer EL.

Domain tree-based analysis of protein architecture evolution.
Molecular Biology and Evolution. 2008, 25:254-264.

* These authors contributed equally to this work.

IV. **Henricson A**, and Sonnhammer EL.

Orthology confers intron position conservation.
Submitted.

CONTENTS

1 PREFACE

Evolution became a concept with the publication of Charles Darwin's "Origin of the Species" in 1859. His hypothesis of how species evolve revolutionized science. While the 19th and 20th century scientists studied evolution mostly through phenotypic observations, the 21st century has been characterized by genotypic analysis. With the accessibility of whole genomes from a multitude of organisms from all kingdoms and computing resources previously not available, evolutionary analysis and all of its aspects have stepped into the post genomic era.

2 HOMOLOGY, ORTHOLOGY, PARALOGY

Just as anyone can make an ancestral tree of their family going back in time, phylogenetic trees can be built for sequences of all kinds. Sequences can be DNA, RNA or protein. The concept of homology, orthology and paralogy was written down by Fitch in 1970 (Fitch 1970). Sequences are homologous if they have evolved from a common ancestor. To more specifically denote the type of evolutionary ancestry, orthology and paralogy were defined. Orthologs are sequences that arose from a last common ancestor due to a speciation split. In contrast to homology, orthology is not transitive; meaning that if gene A and B are orthologs, and gene B and C likewise, gene A and C are not necessarily orthologs. Paralogs are sequences that have been duplicated during evolution. Depending on when, in relation to a species split, the duplication occurred, paralogs can be divided into outparalogs and inparalogs (Sonnhammer and Koonin 2002). If the duplication predates the speciation split, the sequences are outparalogs and as such do not form orthologous relationships. However, if the duplication occurred after the speciation split, the sequences are inparalogs and form co-orthologous relationships with sequences in other species.

The definitions of orthologs, inparalogs and outparalogs are illustrated in figure 2.1. From the phylogenetic tree depicted, the following conclusions can be drawn:

- The yeast gene is ortholog to all human and worm genes, while the human and worm genes are co-orthologs to the yeast gene.
- The human gene HB and worm gene WB are orthologs.
- The human genes HA* and worm genes WA* are co-orthologs.
- The human genes HA* are inparalogs when comparing human to worm.
- The worm genes WA* are inparalogs when comparing human to worm.
- The human genes HA* and HB are outparalogs when comparing human to worm, however, they are inparalogs when comparing animal with yeast.
- The worm genes WA* and WB are outparalogs when comparing human to worm, however, they are inparalogs when comparing animal with yeast.

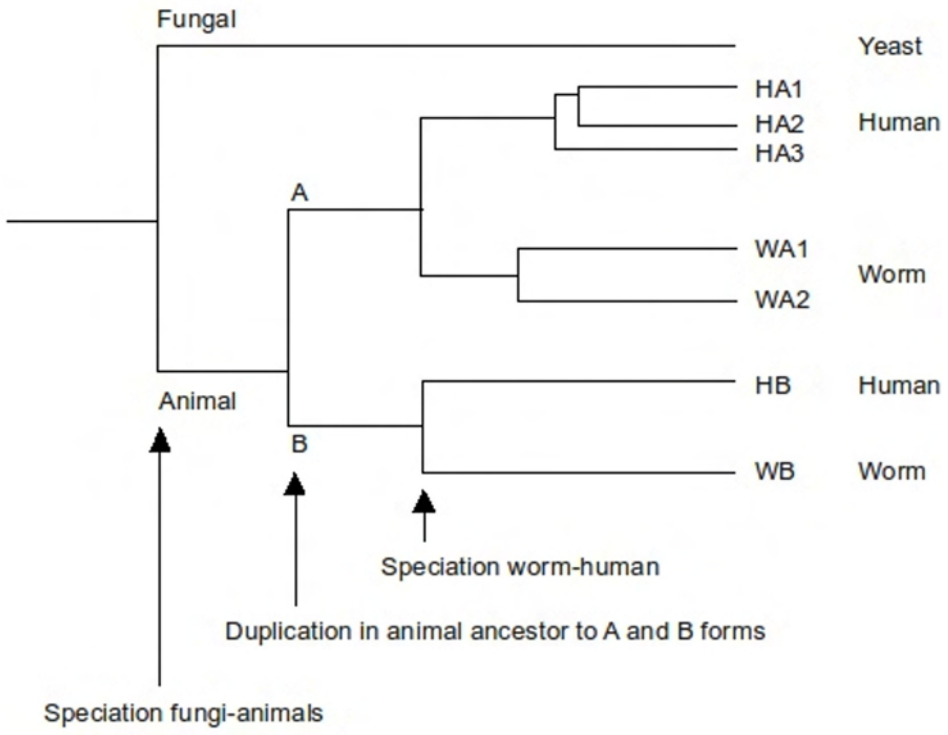

Figure 2.1: Phylogenetic tree illustrating the definition of orthologs, inparalogs and outparalogs.

Adapted from Sonnhammer and Koonin 2002.

Close orthologs in different species are likely to have retained the same biological role. Distant orthologs, on the other hand, are less likely to have the same function; however, they may still play the same role in corresponding pathways in the different species. In contrast, paralogs are likely to have diversified their function through neo- or sub-functionalization. The former implies that as genes duplicate, one of the copies is under positive selective pressure and therefore free to develop a new function, while the other retains the original function. In contrast, the latter means that both copies diverge in function, possibly dividing the original function between them. Analysis of inparalogs can be used to detect lineage-specific adaptations. Outparalogs, on the other hand, cannot be used to transfer functional assignments between species, since they do not form co-orthologous relationships.

2.1 INFERENCE OF PHYLOGENETIC RELATIONSHIPS

With the wealth of genomes that are now available it is impossible to experimentally deduce function for all proteins. Therefore, characterization in simple model organisms and subsequent functional transfer to orthologs in other species is becoming increasingly important. For this reason, correct and reasonably fast inference of evolutionary relationships between sequences is essential. There exists a multitude of approaches for assigning orthology (reviewed in Alexeyenko et al. 2006; Hulsen et al. 2006; Chen et al. 2007; Gabaldón 2008; Kuzniar et al. 2008; Altenhoff and Dessimoz 2009), however, they can be broadly classified into three different groups; the tree-based, the pairwise matching-based, and the hybrid methods.

2.1.1 Tree-based methods

The reasoning behind the tree-based methods is that since orthology is a phylogenetic relationship, phylogeny, as opposed to pairwise sequence comparisons, should be used to analyze it. To infer evolutionary relationships a gene tree is built for a group of sequences and subsequently reconciled with a species tree. However, there are drawbacks with this approach; the major ones being that both the gene and species tree have to be correct. Also, it is time consuming and computationally resource intensive and therefore not suitable for analysis of complete genomes.

2.1.2 Pairwise matching-based methods

The pairwise matching-based approaches are the ones most extensively used to infer orthology and they are applicable for genome-wide analysis. There are several methods available and they are all based on some pairwise sequence similarity calculated between all sequences. The most simplistic approach is the bidirectional best-hits (BBH) (Overbeek et al. 1999) that can assign only one-to-one orthologs across two species. InParanoid (Remm et al. 2001; Ostlund et al. 2009 *in press*) also assigns orthology across two species, however, in addition to one-to-one orthologs, it can identity one-to-many and many-to-many ortholog relationships. Furthermore, InParanoid separates inparalogs from outparalogs, something that is particularly important when analyzing phylogenies involving eukaryotes. Other methods assign orthology across several species, *e.g.* OrthoMCL (Li et al. 2003), COGs (Tatusov et al. 2003), and OMA (Roth et al. 2008). They use different clustering techniques to extend from two species to several. Unfortunately, in the resulting orthologous groups, orthologs and paralogs can be grouped together, and there is also no separation between in- and outparalogs. However, OMA tries to avoid classifying paralogs as orthologs by using an outgroup.

2.1.3 Hybrid methods

Hybrid methods use both pairwise matching-based methods and phylogenetic trees to infer orthology, *e.g.* Ensembl Compara (Hubbard et al. 2007), Homologene (Wheeler et al. 2007), and TreeFam (Li et al. 2006; Ruan et al. 2008). These methods are suitable for whole genome analysis since they are much more scalable than the strictly tree-based approaches.

2.2 PERFORMANCE OF ORTHOLOGY ASSIGNMENTS METHODS

A problem when assessing the performance of various orthology assignment approaches is that there is no "gold standard", *i.e.* a set of known orthologs and non-orthologs across several species that could be used to test the outcome of each approach. Despite this, there have been several studies comparing the performance of different automatic orthology detection methods (Hulsen et al. 2006; Chen et al. 2007; Altenhoff and Dessimoz 2009). Typically, tree-based methods exhibit high specificity and low sensitivity, whereas pairwise matching-based demonstrate high sensitivity and low specificity (Chen et al. 2007). Although these studies do not fully agree as to the ranking of ortholog detection methods, InParanoid (Remm et al. 2001; Ostlund et al. 2009 *in press*) was found to be one of the most accurate algorithms for pairwise orthology assignments, with both specificity and sensitivity being satisfactory high (>80%) (Chen et al. 2007). Also worth noting, is that the tree-based and hybrid methods generally performed worse than the pairwise matching-based.

2.3 INPARANOID

InParanoid finds non-overlapping clusters of orthologs and inparalogs across two species. The algorithm finds the bi-directionally best blast hits between the two genomes; these are the so-called seed orthologs (see figure 2.2). Around these seed orthologs, inparalogs from each species are clustered separately. Sequences in the same species that are more similar to the seed ortholog than to any sequence in the other species will be classified as an inparalog and added to the cluster. A confidence score is calculated for each inparalog, reflecting its similarity to the seed ortholog. There is also an InParanoid database, where the current version (InParanoid7) contains eukaryotic ortholog clusters from 100 organisms.

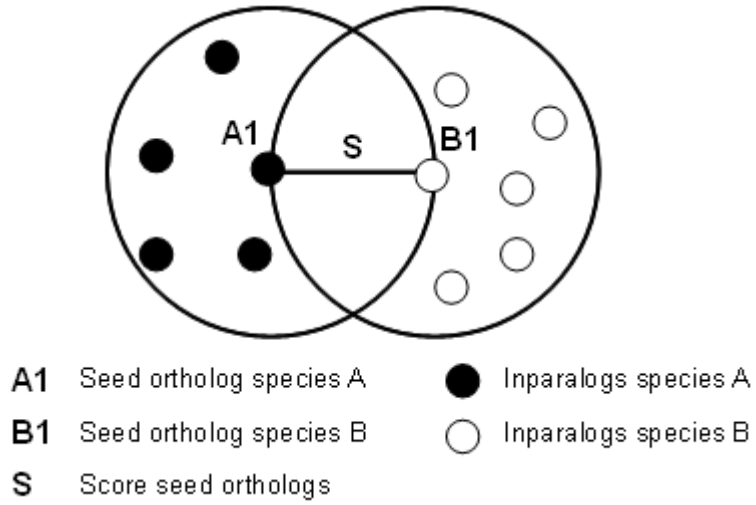

Figure 2.2: Graphical representation of an InParanoid ortholog cluster. The seed orthologs from the different species are denoted A1 and B1 and they are bi-directional best Blast hits. Their similarity score (S) is shown. Inparalogs with score S or higher to the seed ortholog are inside the circle with diameter S and hence, added to the cluster. Inparalogs are added to the cluster independently for each species.
Adapted from Remm et al. 2001.

3 MODEL ORGANISMS

Model organisms are organisms that have been extensively studied in science due to some tractable feature, such as ease of maintenance in the laboratory and rapid generation time. Common model organisms are yeast, roundworm, fruit fly, zebra fish, mouse and the plant *Arabidopsis thaliana*. Proteins can be experimentally characterized in these organisms and thereafter, through orthology, function can be transferred to other more complex organisms, such as humans.

3.1 CAENORHABDITIS ELEGANS

The roundworm *Caenorhabditis elegans* (see figure 3.1) has been used as a model organism since the 1960s when it was brought from the soil into the laboratory by Sidney Brenner. This 1.5 mm long, transparent animal that feeds on bacteria has proven to be very useful and has been extensively studied. The worm is easy and cheap to maintain in the laboratory. In addition, it can be frozen and subsequently thawed and still remain viable, allowing for long-term storage.

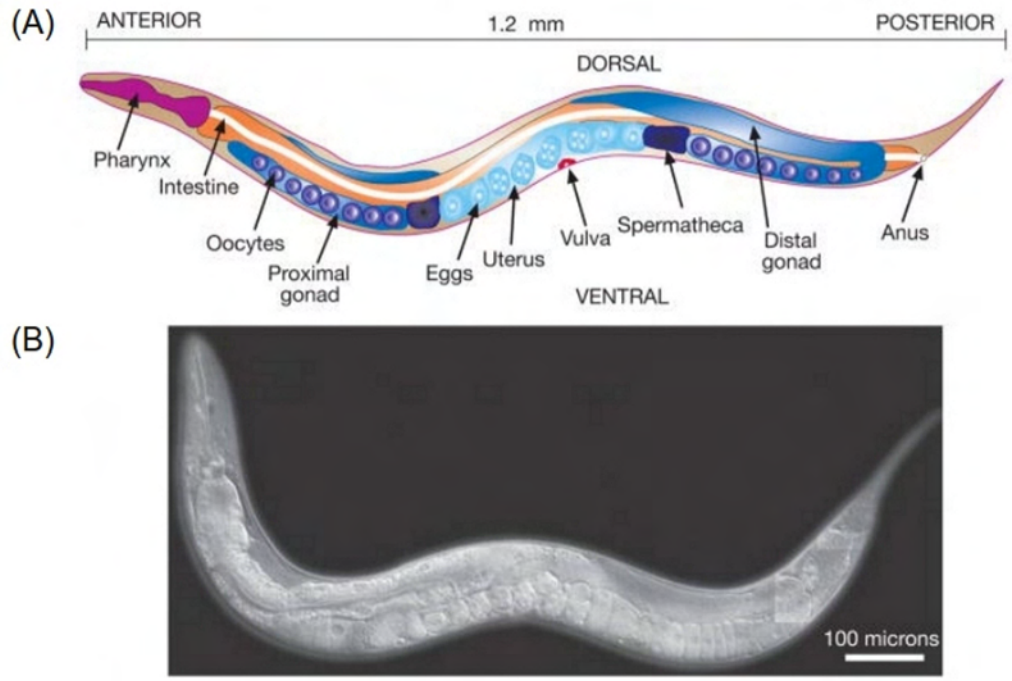

Figure 3.1: *Caenorhabditis elegans* adult hermaphrodite. (A) Schematic picture with labeled body parts. (B) Microscope picture.

Adapted from "Essentials of glycobiology", (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=glyco2).

C. elegans has two sexes; hermaphrodites and males (Riddle et al. 1997). The hermaphrodite is self-fertilizing and mainly produces hermaphrodite offspring. Males arise spontaneously due to X chromosome non-disjunction at meiosis with a frequency of 1/500 animals. Eggs are laid by the hermaphrodite and after hatching, they pass through four larval stages. The generation time is approximately three days. If conditions for growth and reproduction are unfavorable, *e.g.* nutrients are sparse, the larvae can enter an alternative third larval stage called dauer. This is an endurance

stage where the larvae have a unique morphology and are resistant to stress. They also have an altered energy metabolism, and are arrested in development and aging. The dauer larvae can live four to eight times longer compared to the normal three weeks lifespan of *C. elegans*.

The adult hermaphrodite has 959 somatic nuclei, which represent most major differentiated tissue types (Riddle et al. 1997). The complete lineage of these cells has been mapped out. Despite being such a simple organism, *C. elegans* has quite an elaborate nervous system that comprises in total 302 neurons. The complete wiring of all neurons has been mapped out. In 1998 the first genome from a multi-cellular organism was published, and it was that of *C. elegans* (C. elegans Sequencing Consortium 1998).

4 PROTEIN DOMAINS AND ARCHITECTURE

Proteins are the products of genes and the building blocks of nature. There are three major classes of proteins i) globular, ii) fibrous, and iii) membrane proteins. Globular proteins are soluble and many of them are enzymes. The fibrous proteins are the ones that make up structure, *e.g.* keratin and collagen. Membrane proteins are often channels or receptors that enable passing or transport of molecules across the membrane.

4.1 DOMAINS

A protein can contain one or several domains. A domain is defined as a functional element that can fold independently (Jaenicke 1987) and may combine with other domains to form a multi-domain protein (Rossmann et al. 1974). Domains can combine in different ways to form proteins with varying functions. Domains are sometimes viewed as Lego blocks that can be put together to build a protein. So-called supra-domains have also been identified (Vogel et al. 2004). These are two or three domain combinations that reappear in various proteins with different partner domains. A majority of protein domains have been shown to only recombine with one or two other domain families, whereas others are highly promiscuous combining with several other families (Apic et al. 2001; Park et al. 2001). This pattern of domain usage is that of a power law. Hence, the graph of domain combinations is a scale-free network (Wuchty 2001), with the promiscuous domains acting as hubs.

4.2 PROTEIN DOMAIN DATABASES

There are several databases where protein domains are collected using various classification schemes.

The SCOP (Structural Classification of Proteins) database describes the structural and evolutionary relationships between all proteins whose structure is known (Murzin et al. 1995; Andreeva et al. 2008). There are six main levels of hierarchical clustering; species, protein, family, superfamily, fold and class (top of the hierarchy). Another protein domain database that also only entails proteins with known structure is CATH (Orengo et al. 1997; Cuff et al. 2009). Here the proteins are classified according to four major levels; class, architecture, topology, and homologous superfamily (top of the hierarchy), hence, the abbreviated name CATH. Both SCOP and CATH classify proteins using both automatic methods and manual procedures.

Other protein domain databases are based more extensively on automatic methods, namely SUPERFAMILY (Gough et al. 2001) and Pfam (Finn et al. 2008). The former is based on a collection of hidden Markov models (HMMs) that are derived from protein domains on the SCOP superfamily level. These HMMs are subsequently used to search for sequences which match the models and hence belong to the same domain family.

The Pfam database is also based on HMMs; however, the starting point is manually curated multiple sequence alignments representing each domain family. From these

so-called seed alignments, an HMM is generated and subsequently used to find new members of the family, thus producing the full alignment of all the members of the domain family. This collection constitutes the Pfam-A part of the database. In addition, Pfam-A has a higher-level grouping of related families called clans. There is a fully automatic part of Pfam (Pfam-B) that is of lower quality. However, annotations in Pfam-B can be useful when no Pfam-A domains are found in a protein.

4.3 DOMAIN ARCHITECTURES

There are both single-domain and multi-domain proteins occurring in nature. While most domain families constitute single-domain proteins, a majority of domains are also found in multi-domain proteins. However, the number of existing domain combinations only comprise a small fraction of the possible number of combinations, implying that domain recombination is under strong selective pressure (Vogel et al. 2005). In eukaryotes, a majority of proteins have multiple domains, whereas prokaryotes have fewer multi-domain proteins (Apic et al. 2001, Ekman et al. 2005, Wang and Caetano-Anollés 2006). A schematic representation of a multi-domain architecture is given in figure 4.1.

Figure 4.1: Example of a multi-domain protein as depicted by the Pfam database. Shown is the human proto-oncogene vav (VAV_HUMAN, P15498). The yellow domain is C1_1, and the purple domain is SH3_1.

The order of domains in a protein can be referred to as the protein's domain architecture. Two domain families A and B can occur either in the sequential order AB or BA; however, the sequential order is always the same in different proteins in which they are found. Only for a minority of domain pairs (~2%) can both sequential orders be found (Apic et al. 2001; Bashton and Chothia 2002). During evolution, domains have combined in different ways, leading to loss of architectures or gain of new ones. Fusion and fission are two processes shaping the domain architecture repertoire. Fusion is the joining of domains creating multi-domain proteins. Fission, on the other hand, is the separation of domains possibly creating single-domain proteins. Studies have shown that fusion events are much more common than fission events in all kingdoms of life Kummerfeld and Teichmann 2005; Fong et al. 2007). Furthermore, domain losses and duplications have been shown to preferentially occur at either terminus (Björklund et al. 2005; Weiner et al. 2006). Proteins that have been circularly permuted, meaning that the domain order has been inverted, have also been identified (Weiner and Bornberg-Bauer 2006). Most architectures have been created once and through evolution it has spread across species (Dolittle 1995; Apic et al. 2001; Gough 2005; Kummerfeld and Teichmann 2005). However, there are also examples of where the same domain architecture have arisen several times independently, a phenomenon called convergent evolution (see paper III).

5 TRANSMEMBRANE TOPOLOGY

Membrane proteins are notoriously difficult to over express and crystallize making it extremely troublesome to determine their structure. However, elucidating the topology of transmembrane proteins is much more feasible. The transmembrane topology can be viewed as a description of which parts of the protein that lie within the membrane and which portions that are situated on the cytoplasmic or non-cytoplasmic side of the membrane, respectively (see figure 5.1).

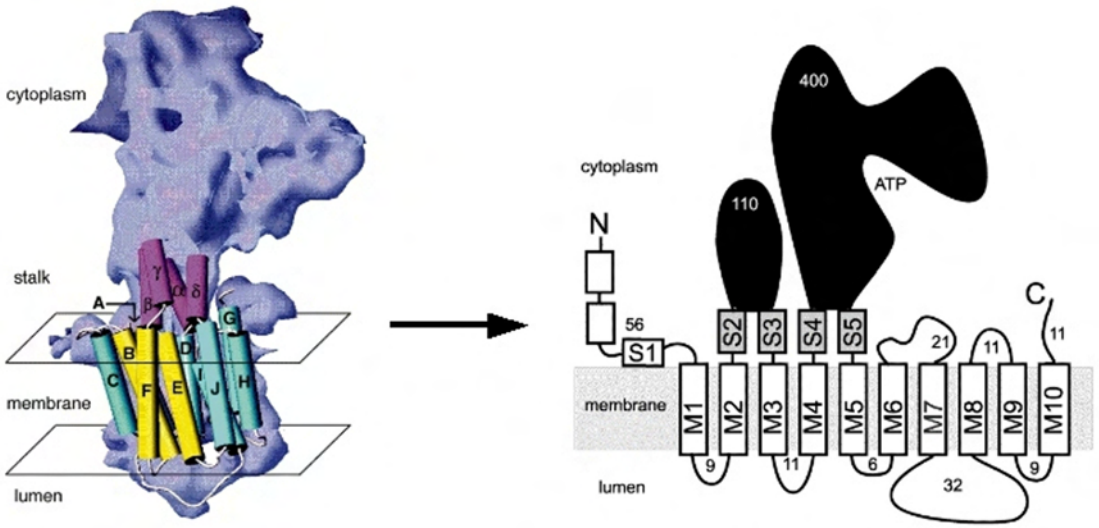

Figure 5.1: Transmembrane topology describes which parts of the protein that lie within the membrane and which portions that are situated on the cytoplasmic or non-cytoplasmic side of the membrane.

Adapted from Zhang et al. 1998.

There are two types of transmembrane proteins; α-helical and β-barrels. The former have segments of 18-35 amino acids long hydrophobic α-helices that traverse the membrane, while the latter have β-sheets. The first class is far more common, hence I will only refer to these when describing transmembrane proteins in this text. In eukaryotes, proteins with seven transmembrane regions are common and they have also been extensively studied mainly because a lot of them are drug targets. The numerous G-protein coupled receptors (GPCRs) belong to this group. Proteins with more than ten transmembrane regions usually form pores in the membrane. Depending on whether the protein has an even or odd number of transmembrane regions, it will either have the N- and C-terminal ends on the same side or on different sides of the membrane, respectively.

Proteins can also have signal peptides, usually at their N-terminal end. The signal peptide is used to guide the protein to its final location. When the protein is in place, the signal peptide is normally cleaved off, and if the protein is soluble it is released from the membrane. However, membrane proteins can also have signal peptides. Determining transmembrane topology can be done by using experimental approaches or prediction methods, or preferably, a combination of the two.

5.1 PREDICTION THROUGH EXPERIMENTAL METHODS

There are several strategies for experimental determination of transmembrane topology. The most widely used are biochemically based, such as reporter gene fusions, site tagging and antibodies.

5.1.1 Reporter gene fusions

A common experimental method for determining transmembrane topology is through reporter genes that are inactive on one side of the membrane but active on the other. Typically, several truncated forms of the membrane protein are made and each truncated form is fused with a reporter gene. The hybrid protein is expressed in bacteria or some other model organism such as *C. elegans*, and then activity of the reporter is screened for. Unfortunately, these types of studies do not always give clear-cut results, making it difficult to solely rely on them. This can be avoided to some extent by using two different reporter genes that are active on opposite sides of the membrane, freeing the researcher from relying on negative results. There is also the possibility that the reporter gene might alter the topology or that the truncated forms of the protein fold differently compared to the native protein.

5.1.2 Site-tagging

Instead of a reporter gene, site-tagging via N-glycosylation can be used to determine the topology of a protein. Glycosylation can only occur in the lumen of the endoplasmic reticulum (ER), which is equivalent to the extracellular space. If the protein is glycosylated it will be heavier compared to the native protein and, hence they can be separated on an SDS-PAGE gel. The N-glycosylation site consists of the amino acids N-X-S/T, where X can be any amino acid except proline (Hart et al. 1979). As a result of being much smaller compared to reporter genes, N-glycosylation sites are less likely to alter the native topology of the protein.

5.1.3 Antibodies

Antibodies directed towards various epitopes of the protein, usually the loop regions, can be also be utilized to determine the transmembrane topology of the protein. If the epitope is hidden within the cell, it cannot be reached by the antibodies; however, if location is opposite, the antibody can bind to the epitope. As an additional control, the cells can be made permeable to allow for the antibodies to enter. The pitfalls of these types of experiments are mainly the specificity of the antibodies used.

5.2 IN SILICO TOPOLOGY PREDICTION

The first attempts to model protein membrane topology arose once it was shown that loops on the cytosolic side of the membrane tend to include more positively charged amino acids, the so-called positive inside rule (von Heijne 1986). Today there are several different programs for predicting transmembrane topology or signal peptides. Predicting transmembrane topology involves several different aspects. The α-helices that traverse the membrane have an approximate length of 18-35 hydrophobic amino acids simply due to the thickness of the membrane layer. The loops can be of variable length; however, they contain hydrophilic amino acids and also follow the positive inside rule. A complicating factor is that sometimes there can be regions of a membrane protein that graze the membrane, thereby giving a somewhat weaker

hydrophobic signal than the traversing regions. Most transmembrane topology predictors cannot distinguish between a signal peptide and an N-terminal transmembrane region. Determining the presence of a signal peptide is very attractive since if present, the N-terminus of the protein must be located on the non-cytoplasmic side of the membrane. As a result, the orientation of the protein in the membrane can be easily determined. Accordingly, a combination of transmembrane topology and signal peptide predictor is preferable. Phobius (Käll et al. 2004) is such a predictor that combines the two. It is a hidden Markov model (HMM)-based prediction method. Other predictors based on HMMs include TMHMM (Krogh et al. 2001) and HMMTOP (Tusnády and Simon 2001). There are also predictors based on neural networks, *e.g.* PHDhtm (Rost et al. 1996), or dynamic programming, *e.g.* Memsat (Jones et al. 1994).

Another advantage of Phobius, besides the combined prediction of transmembrane regions and signal peptides, is that it can also perform constrained predictions, meaning that the N- or C-terminus or an internal loop can be constrained to be either cytosolic or non-cytosolic. Phobius then predicts the most likely topology with the selected loop(s) constrained. This is useful when the localization of a certain portion of the protein has been determined through experimental methods.

6 GENE STRUCTURE

It was long believed that a gene was an uninterrupted sequence of nucleotides coding for amino acids. This is indeed true for prokaryotes, although some self splicing introns have been found. However, in the 1970s it became clear that this is not the case for eukaryotic genomes (Gilbert 1978). A gene seemed to be split into several parts with some non protein coding sequence in between; introns had been discovered. There is a splicing machinery in eukaryotes (the spliceosome) that splices the pre mRNA so that the introns are removed and only the exons remain in the mature mRNA, for this reason these introns were called spliceosomal introns. From hereafter, I will only refer to these when describing introns in this text, unless explicitly stated. Alternative splicing was also discovered, *i.e.* that a pre mRNA transcript can be spliced in various fashions resulting in different forms of mature mRNA and consequently leading to alternate protein products. This discovery explained some of the mystery regarding how humans, although much more complex, only have roughly 5,000 more genes than the roundworm *C. elegans*. With extensive alternative splicing, the number of different proteins can greatly exceed the number of genes in the genome.

6.1 INTRON EVOLUTION

The evolution of the spliceosomal introns proved to be elusive. Two major opposing hypotheses soon became apparent, the introns-early and the introns-late theory (see figure 6.1). In the introns-early theory, it is believed that nearly all introns are ancient and inherited by eukaryotic genes from prokaryotic ancestors, and then subsequently they have been lost in the prokaryotes (Gilbert 1978; Doolittle 1978; Blake 1978; Gilbert 1987). This complete extinction of spliceosomal introns and also of the whole spliceosome complex in prokaryotes, have been explained by a need to maximize the replication rate to facilitate rapid growth (Gilbert 1987; Roy 2003). In line with the introns-early theory, the difference in gene structure among homologous eukaryotic genes that we see today is largely the result of differential intron loss. According to this theory, introns were the main driving force leading to the creation of early genes by shuffling of exons ("the exon theory of genes") (Blake 1979; Gilbert 1987; Holland and Blake 1987). In contrast, the introns-late hypothesis states that spliceosomal introns are an eukaryotic invention and that they have never existed in prokaryotes (Cavalier-Smith 1985; Stoltzfus et al. 1994; Logsdon 1998), and that new introns have been emerging continuously since then. With increased knowledge about intron evolution the more drastic versions of the two scenarios, which view nearly all introns as either ancient or new, have given way to more nuanced hypotheses. The current introns-early theory is that only a minority of moderns introns were present in prokaryotes (de Souza et al. 1998; Roy 2003), whereas the introns-late states that introns evolved from bacterial group-II-like self-splicing introns in relatively early eukaryotes (Cavalier-Smith 1991; Stoltzfus 1999).

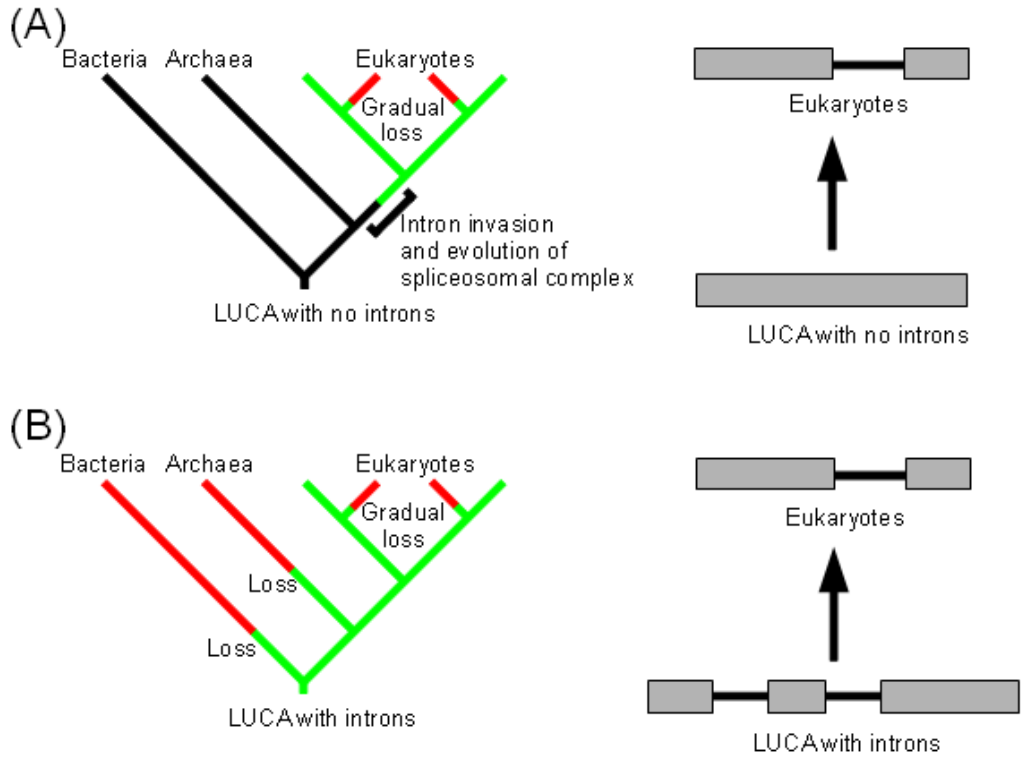

Figure 6.1: Intron evolution as explained by the (A) introns-late, and (B) introns-early hypothesis. In the trees to the left the green branches show lineages containing introns, the black branches indicate pre-intron stages and the red branches denote secondary loss of introns. LUCA is last universal common ancestor.

Adapted from Jeffares et al. 2006.

The scientific evidence points to an intron-rich eukaryotic ancestor; however, this is still up for debate. Parsimonious approaches and maximum likelihood models, or a mix thereof, have been used to study intron evolution dating back to the last common ancestor of animals and plants. The results show somewhat different conclusions as to just how intron-rich the ancestor genome was, although the methods agree that it was relatively intron-rich (Rogozin et al 2003; Qiu et al 2004; Nguyen et al. 2005; Roy and Gilbert 2005). A complicating factor when analyzing intron evolution is that the underlying eukaryotic phylogenetic tree has not yet been resolved. Also, the methods have different drawbacks; parsimony tends to underestimate the number of ancestral introns, since multiple losses of introns is penalized, whereas maximum likelihood models have a lot of parameters that are difficult to optimize, leading to, in particular, overestimation of ancestral introns (Koonin 2006). Another complicating factor when elucidating intron evolution, is that different lineages exhibit very divergent rates and patterns of intron loss or gain (Rogozin et al. 2003; Roy and Gilbert 2005; Carmel, Wolf et al. 2007). It seems that intron loss is generally more prevalent than gain (Robertson 1998; Robertson 2000; Mourier and Jeffares 2003; Roy et al. 2003; Roy and Penny 2007), although there are studies showing the opposite (Babenko et al. 2004).

6.2 CONSERVATION OF INTRON POSITIONS

It has been shown that introns are preserved across distant species (Fedorov et al. 2002; Rogozin et al. 2003). However, it is not the content or length of the introns that

is conserved, rather their positions within the gene. On the other hand, it was also shown that introns preferentially inserted into or are fixed at so-called protosplice sites (Dibb and Newman 1989; Dibb 1991; Sadusky et al. 2004; Sverdlov et al. 2004). As a consequence, another study proposed that the reason for finding introns in the same positions in different species was to a great extent due to parallel gain of introns into these protosplice sites (Qiu et al. 2004). These findings were later disputed and it was shown that protosplice sites are no more conserved during eukaryotic evolution than random sites (Sverdlov et al. 2005). In addition, simulation of intron insertion into protosplice sites with the observed protosplice sites frequencies and intron densities showed that parallel gain could account for only 5-10% of shared intron positions in distantly related species. Subsequently, this has been verified in another study, where on average ~8% of shared intron positions in distantly related species were found to be due to parallel gain (Carmel, Rogozin et al. 2007). However, across the eukaryotic lineages, the distribution of parallel gain was highly heterogeneous with evolutionary closer species showing virtually no shared introns due to parallel gain, whereas evolutionary more distant species, such as human and plants, exhibited up to 20% parallel gain.

7 PRESENT INVESTIGATION

7.1 PAPER I

A valid approach for assigning function in more complex organisms is through functional characterization of its ortholog in a model organism. Therefore, in this study we used *Caenorhabditis elegans* to functionally analyze worm-human orthologs of unknown function. These proteins had initially been identified in a study searching for worm-human orthologs of transmembrane proteins (Remm and Sonnhammer 2000). We analyzed the function of 17 of these orthologs through experimental studies in *C. elegans* as well as further bioinformatic analysis.

RNA interference to downregulate the genes of interest was performed in both wildtype worms and an RNAi sensitive worm strain. For 2 of the 17 genes (~12%) we could detect an RNAi phenotype, which is comparable to other RNAi studies (Kamath et al. 2003; Simmer et al. 2003). Transgenic worm strains carrying transcriptional *gfp* fusions were established for 14 of the genes. Gene expression was detected in various tissues in the transgenic worm strains, with the most predominant tissues being hypodermis, nervous system, pharyngeal muscle and intestine. However, 2 of the transgenic strains established showed no expression.

Bioinformatic analysis was also carried out to further investigate the evolutionary relationships and function of the proteins. A more in depth phylogenetic analysis revealed that the great majority of the genes were indeed true orthologs. Only in one case could we detect that the proposed ortholog was most likely an outparalog, hence, transfer of putative function between species was not valid. Protein domain assignments were made using Pfam; however, for 3 proteins, no Pfam-A domain could be detected. The transmembrane topology of the proteins was analyzed with nine different methods. In addition, the possible presence of signal peptides was investigated with two different approaches. Since all transmembrane topology and signal peptide predictors have some margin of error, a consensus result gives a better estimate of the true protein topology. The number of transmembrane regions varied between 6 and 10, with a majority of the proteins having 6 or 7 segments traversing the membrane.

To summarize, we could assign a putative function to 12 of the 17 genes studied. More specifically, we also proposed a novel transmembrane topology for a presenilin-like protein; a 9-transmembrane topology with the C-terminus located in the cytoplasm (see figure 7.1 B).

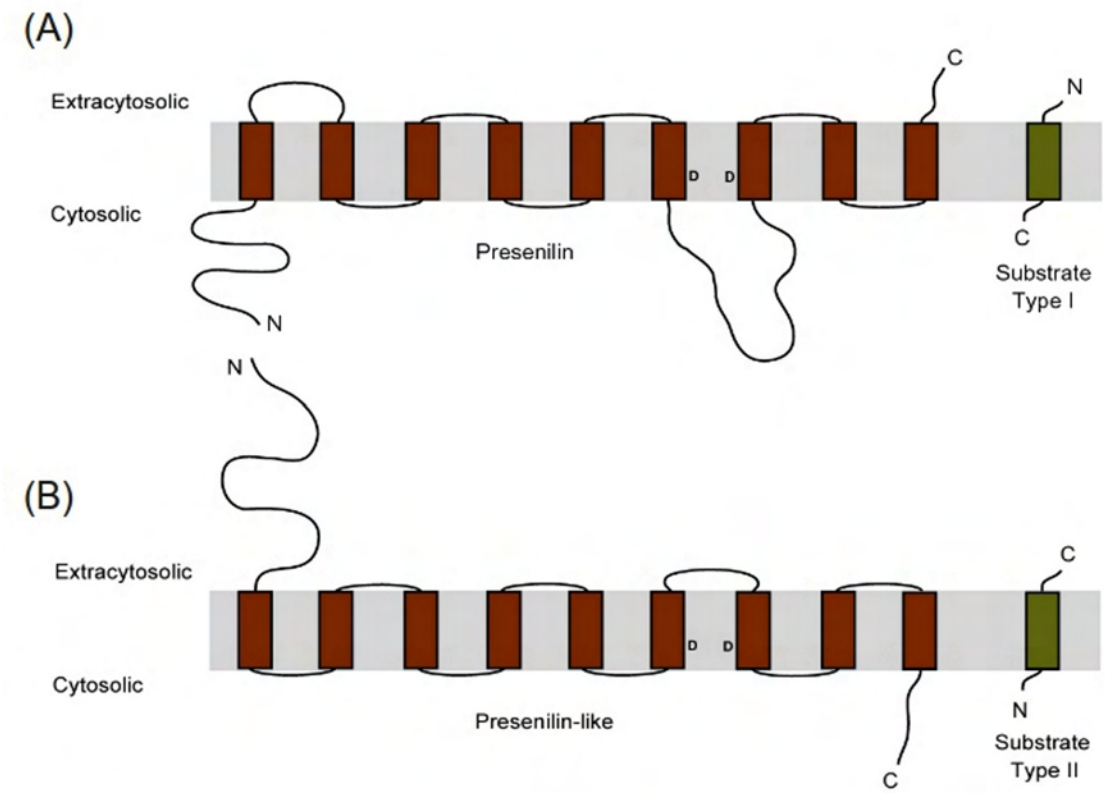

Figure 7.1: Transmembrane topology for the (A) presenilins, and (B) presenilin-like proteins. The active site aspartate residues are indicated with D. Due to their different topology orientation, presenilin and presenilin-like proteins cleave substrates of type I and type II, respectively.

7.2 PAPER II

Presenilins are transmembrane proteins that are part of γ-secretase, a multi subunit protease that also entails nicastrin, aph-1 and pen-2 (Kimberly and Wolfe 2003; Fraering et al. 2004). This complex is responsible for the intramembrane proteolysis of type I membrane proteins, such as amyloid-β precursor protein (APP) and Notch. The former protein is involved in Alzheimer's disease, which is the major cause of dementia in humans. The latter is a developmental protein responsible for critical signaling events. Elucidating the transmembrane topology of presenilin is important to fully understand its function, and several attempts had been made at experimentally determining the topology. Different approaches such as reporter gene fusions, antibodies, glycosylation studies, or a combination of these methods had been used (Doan et al. 1996; Li and Greenwald 1996; Dewji and Singer 1997; Lehmann et al. 1997; Li and Greenwald 1998; Nakai et al. 1999; Dewji et al. 2004). Unfortunately, no consensus could be reached.

In this study we investigated the presenilin topology by using several different prediction methods as well as incorporating published experimental results. We also used our previously published topology model for the homologous presenilin-like family of proteins (see paper I). Although the overall sequence similarity between the two protein families is fairly low, they share the conserved putative active site aspartate residues and the C-terminal “PAL” motif. We presented a novel 9-transmembrane topology with the C-terminus located in the extracytosolic space for the presenilin family (see figure 7.1 A). This model was heavily supported by

published data on γ-secretase function and presenilin topology. In fact, our model was supported by 70% of the experimentally determined loop locations.

7.3 PAPER III

It is believed that a great majority of protein domain architectures have arisen only once during evolution and the reason for finding them in different species today is due to speciation splits during the course of evolution (Doolittle 1995; Apic et al. 2001; Gough 2005; Kummerfeld and Teichmann 2005). However, sometimes multiple independent creation events can generate the same domain architecture due to functional necessity or random chance. A previous study had shown that although convergent evolution could be found it was extremely rare (Gough 2005). This analysis was heavily biased towards prokaryotes and it also relied on a species tree. We decided to use a completely different approach based on phylogenetic trees and in addition include more species, specifically increasing the fraction of eukaryotes.

Our novel ancestral architecture inference algorithm is based on maximum parsimony and runs in two passes. In the first pass, the phylogenetic tree of a domain family is traversed from the leaves to the root. The existing domain architectures at the leaves are used to initialize the tree. Subsequently, at each inner node, the ancestral architecture with the lowest cost according to the maximum parsimony criterion is assigned. If several architectures have equal costs, they are all enumerated. In the second pass, we traverse the tree in the opposite direction starting from the root. At each inner node, the ancestral architecture resulting in the lowest cost over the whole tree is selected. If there are several architectures giving the same overall cost, an architecture is randomly chosen. We used bootstrapping to ensure the quality of the phylogenetic trees. Only architectures where the phylogenetic tree and all of its pseudoreplicate trees were available for at least two domains were scored. A majority of the phylogenetic trees for the individual domains had to support the evolutionary classification as “multiple” or “single”, otherwise the evolution of the architecture was scored as “ambiguous”. There also had to be agreement regarding the architecture origin.

An inherent difficulty when analyzing domain architectures is that there are regions in the proteins that have no domains assigned to it. This is especially an issue when studying eukaryotes. To tackle this problem, we decided to investigate two different datasets, the *no-limit* and *max50*. In the *no-limit*, all proteins are included regardless of the length of unassigned regions. In the *max50* dataset a maximum of 50 amino acids was allowed between neighboring domains, in addition, this maximum length was applied to both the N- and C-terminus. For the architectures where we could determine either a single or multiple origin, we found that 12.4% had a multiple origin in the *no-limit* data set (see Table 7.1). The equivalent figure in the *max50* data set was 5.6%.

Table 7.1: Candidates for convergent protein architecture evolution found in the *no-limit* and *max50* data sets.

Adapted from paper III.

Data Set	N_{total}[1]	$N_{ambiguous}$[2]	N_{single}[3]	$N_{multiple}$[4]
no-limit	8367	1605	4613	650
max50	1798	301	1172	70

[1] Total number of architectures included in the analysis.

[2] Number of architectures where results from the individual domains were non-conclusive.

[3] Number of single-origin architectures.

[4] Number of multiple-origin architectures.

In a majority of cases, the architectures revealing convergent evolution have evolved through two independent creation events. We analyzed the kingdom distribution of domain architectures displaying multiple independent creation events further and found that a considerable amount was specific to eukaryotes. We also analyzed if there was some functional specificity among the reinvented architectures using GO terms. In the set of architectures classified as multiple there was a significant enrichment or underrepresentation of 22 GO terms compared to the architectures with a single origin. GO terms associated with signaling were found among the enriched set, but not represented in the depleted. For the architectures with single origin, GO terms associated with metabolic processes were overrepresented.

Our analysis indicated that convergent evolution of domain architectures could be more prevalent than previously thought. We found no strong functional bias among the domain architectures displaying multiple independent creation events, suggesting that convergent evolution is driven by chance rather than functional necessity.

7.4 PAPER IV

The evolution of introns is an elusive topic, and there are are many contradicting views on how and why they emerged and subsequently evolved. Nonetheless, intron positions have been shown to be conserved across long evolutionary timescales (Fedorov et al. 2002; Rogozin et al. 2003). Therefore, there is a possibility that this feature could be used when inferring evolutionary relationships between genes. Indeed, previous studies of individual gene families had used information regarding shared intron positions to elucidate phylogenetic relationships (Robertson 1998; Ferrier et al. 2000; Robertson 2000; Franck et al. 2004); however, no global analysis had been carried out. In this study, we wanted to analyze if orthologs shared more introns positions compared to non-orthologous sequences that were equally similar in sequence.

Clusters of orthologs between human and six other species were identified using the InParanoid algorithm (Remm et al. 2001; Ostlund et al. 2009 *in press*). The algorithm first finds the bi-directionally best Blast hits between the two genomes, the so-called seed orthologs. Around these seed orthologs, inparalogs from each species are clustered separately. Sequences in the same species that are more similar to the seed

ortholog than to any sequence in the other species will be classified as an inparalog and added to the cluster. The inparalogs are ranked by a confidence score that is calculated for each inparalog, reflecting its similarity to the seed ortholog. Subsequent to applying the InParanoid algorithm, the closest non-orthologs, meaning the sequences close in sequence space, yet falling just outside the ortholog clusters, were also added to the clusters (see figure 7.2).

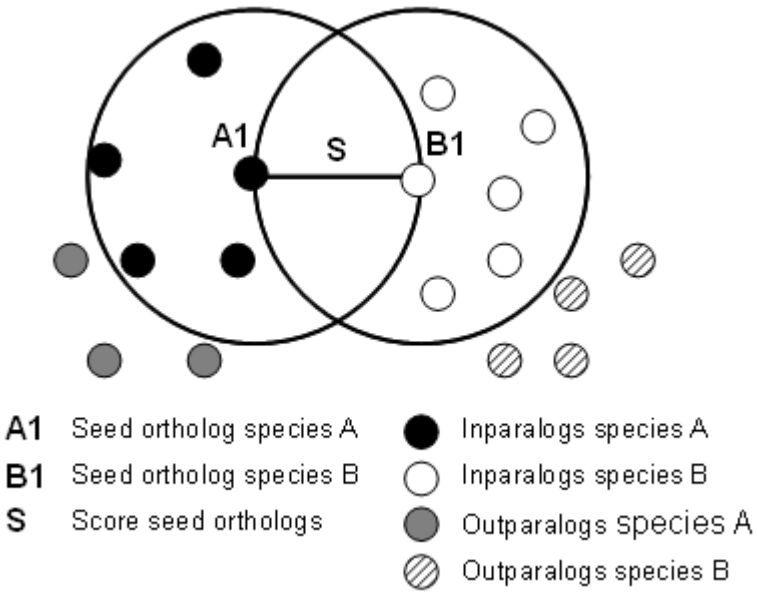

Figure 7.2: Graphical representation of an InParanoid ortholog cluster with the outparalogs outside the cluster indicated. The seed orthologs from the different species are denoted A1 and B1 and they are the bi-directional best Blast hits. Their similarity score (S) is shown. Inparalogs with score S or higher to the seed ortholog are inside the circle with diameter S and hence, belonging to the cluster. Inparalogs are added to the cluster independently for each species. The sequences with a lower score than S are outside the cluster and classified as outparalogs. For each inparalog in the cluster, the closest outparalog (non-ortholog or non-inparalog) from each species was added to the cluster.

Adapted from paper IV.

We developed a new score for intron position conservation (IPC) and applied it to the clusters. For all species comparisons, we found that ortholog-ortholog gene pairs on average have a significantly higher degree of IPC compared to ortholog-closest non-ortholog pairs (see figure 7.3). This was also found to be true for inparalog pairs versus inparalog-closest non-inparalog pairs. Furthermore, we verified that these differences could not simply be attributed to the generally higher sequence identity of the ortholog-ortholog and the inparalog-inparalog pairs.

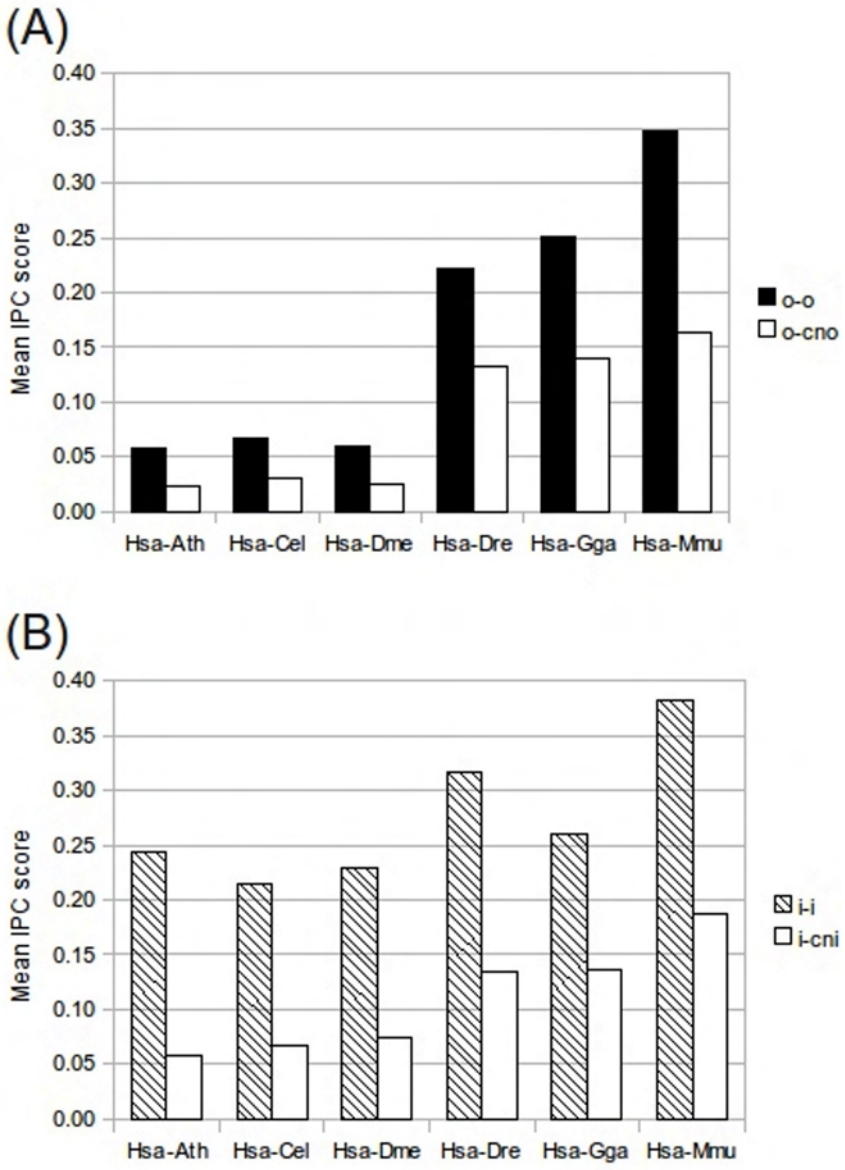

Figure 7.3: Mean intron position conservation score for the different pair types and species comparisons. (A) ortholog-ortholog (o-o) pairs versus ortholog-closest non-ortholog (o-cno) pairs, and (B) inparalog-inparalog (i-i) pairs versus inparalog-closest non-inparalog (i-cni) pairs.

Adapted from paper IV.

If IPC could be used as a discriminating factor when assigning orthology, it has to agree, at least to some extent, with reliable existing orthology detection methods. Therefore, we analyzed the agreement between InParanoid's ranking of inparalogs in an ortholog cluster and their IPC score to the seed ortholog in the other species. We found that for a great majority of multi clusters, *i.e.* clusters with several inparalogs in at least one species, the IPC score supports the seed ortholog assignments made by InParanoid. There is thus a correlation between IPC score and InParanoid seed ortholog assignments, meaning that a high IPC score generally implies a highly confident orthology relationship.

We concluded that orthologous genes tend to have more conserved intron positions compared to non-orthologous genes. As a consequence, our IPC score could be useful as an additional discriminating factor when assigning orthology.

8 REMARKS AND FUTURE PERSPECTIVES

The most exciting phrase to hear in science, the one that heralds new discoveries, is not Eureka! (I found it!) but rather, "hmm.... that's funny...."

Isaac Asimov

Mankind has always sought answers to how nature works. As technology has moved forwards at a rapid pace, the amount of biological data available has grown tremendously. How to interpret and understand this data is a challenging task. With this thesis, I have tried to make a small contribution by unraveling some of the aspects of protein evolution, function and architecture. Although, as is the nature of scientific studies; when answering one question, several new ones arise.

A general comment about studies involving analysis of data present in various databases is that only one instance of the database is captured. As more data become available, results will become more robust and conclusions drawn will more truthfully reflect the true nature of biology. This also reflects on bioinformatics tools available since they have been benchmarked on these databases. Hence, with accumulating scientific knowledge the databases will improve, leading to even better bioinformatics tools and more accurate biological analysis.

8.1 PAPER I

Unfortunately, since the publication of this paper there has not been a lot more learned about the function of the genes in the analysis. For a majority of the genes, their RNAi phenotypes and expression pattern have been confirmed by other studies (http://www.wormbase.org).

Gene *xbx-6* (F40F9.1) has been shown to be an Xbox-promoter element regulated gene (Efimenko et al. 2005). Such genes are involved in cilia formation; however, the precise function of this particular gene has not yet been elucidated.

In our analysis, we found that Y6B3B.10 (*lagr-1*) was associated with longevity and that the human ortholog was LASS1 (P27544). Another name for LASS1 is CerS (ceramide synthase); a protein that regulates ceramide synthesis. Ceramide is essential for apoptosis (Gulbins and Li 2006), however, how it exerts its effect has not yet been established. Indeed, mutant worm strains of *lagr-1* display apoptotic disturbances (Deng et al. 2008). More specifically, the mutant strains exhibited resistance to radiation-induced germ cell apoptosis. Conversely, if ceramide was injected into *lagr-1* mutants there was an increase in germ cell apoptosis. Further studies are needed to elucidate how ceramide synthase is involved in apoptosis, and *C. elegans* can certainly aid in this endeavor.

Our proposed 9-transmembrane topology with the C-terminus located in the cytosol for the presenilin-like proteins has been verified in subsequent studies (Friedmann et

al. 2004; Nyborg et al. 2004). This topology model also aided in our quest for elucidating a novel transmembrane topology of the presenilins (see paper II).

8.2 PAPER II

Our proposed 9-transmembrane topology with the C-terminus located in the extracytosolic space for presenilin has been verified by subsequent studies (Laudon et al. 2005; Oh and Turner 2005; Kornilova et al. 2006; Spasic et al. 2006). The elucidation of the presenilin topology has aided in understanding how the γ-secretase complex can perform intramembrane proteolysis of its substrates. As one of the substrates is amyloid-β precursor protein (APP), this could lead to a better understanding of the mechanisms behind Alzheimer's disease, and hopefully, ultimately, a prevention or cure for the disease.

8.3 PAPER III

Naturally, with even more data in the Pfam database, the evolutionary reconstruction of domain architectures will improve. Therefore, employing our analysis on a newer version of the database is likely to give somewhat different results. How the outcome would be different though, is difficult to foresee. Since our publication there has been no other study analyzing how frequent convergent evolution is.

There are some modifications to our method that could potentially improve the analysis. We used bifurcating trees in our study, which are by far the most commonly used in phylogeny. However, allowing for unresolved tree nodes in cases where the resolution is too low to determine the correct branching order could give a more accurate reflection of biology. Another potential improvement could be to not only consider the topology of the tree but also include branch lengths. If the branch length was considered in relation to the number of gains/losses of domains along the branch, an estimate of the probability of the assigned evolutionary events actually occurring could be calculated. This would give a confidence score to the phylogenetic events inferred. Furthermore, we used an equal cost model for gain or loss of a domain. This is probably not a correct reflection of the nature of these processes; however, we believed that with the current understanding of how domains recombine it was not feasible to design a suitable differential cost model. With increasing understanding of domain recombination events, it could be possible to design such a model.

8.4 PAPER IV

Predicting introns is not a trivial task and pinpointing their exact location is even more difficult. Therefore, with increased accuracy of genome annotations, the results of this analysis are most likely to improve and more accurately reflect the true nature of intron position conservation.

Our analysis shows that there is a fraction of ortholog-ortholog and inparalog-inparalog pairs that do not have any conserved intron positions. On the other hand, the sequences with highest levels of intron position conservation (IPC) are ortholog-ortholog or inparalog-inparalog pairs. This implies that there are two different groups of orthologs; one with a very low and one with a very high IPC. Whether this

grouping reflects ancientness, function or some other feature of the sequences, or is simply random, would be interesting to investigate.

Although some consensus has been reached in the intron evolution field, such as acknowledgement that some introns indeed have conserved positions and that the eukaryotic ancestor had a relatively intron-rich genome, there is still a lively debate ongoing still, more than 30 years after the initial discovery of the introns.

9 ACKNOWLEDGEMENTS

This work has been supported by grants from the Swedish Research Council, Pfizer Corporation, and the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Strategy and Development Office (SDO) at Karolinska Institutet.

I would like to thank my main supervisor, Professor Erik Sonnhammer, for guiding me through these years, teaching me about bioinformatics and how to become an independent researcher.

I would also like to thank my co-supervisor, Dr Ana Vaz Gomes, for introducing me to the worm and for being there even after pursuing a non-academic career.

Thank you to all my co-authors: Kristoffer Forslund for your contribution to the domain architecture project, input on this thesis, and for being an excellent office buddy. Volker Hollich for a good collaboration when launching the domain architecture project. Lukas Käll for a good collaboration on the presenilin project and your expertise in topology predictors.

Thanks to present members of the Sonnhammer group for both scientific and non scientific discussions. Dave Messina, thanks also for input on this thesis, Gabriel Östlund, Oliver Frings, Sanjit Roopra, and Thomas Schmitt.

Thanks to all former members of the Sonnhammer group. A special thanks to Timo Lassman and Isabella Pekkari for helping with some of the programming for the intron project.

Thanks to former members of the Vaz Gomes laboratory: Ivan Tamas, for your never ending efforts on germ line injections in *C. elegans*, I know that it is not easy. Josefin Friberg, our orphan lab member from Umeå, thanks for being a good friend both in and out of the office.

Thanks to past members of the former Center for Genomics and Bioinformatics for a lot of fun during the years.

Thanks to the administrative personnel at CMB for helping me sort out all the practical details.

Ett stort tack till svärmor Florence, svärfar Johnny, svåger Henke och svägerska Annicka för att ni alltid ställer upp.

Jag vill också tacka mina föräldrar Gunnel och Jan för deras stöd under alla dessa år. Min storasyster Lina och svåger Thomas, som alltid har en dörr öppen för oss i Barcelona, även om jag skulle önska att ni bodde närmare.

Till min älskade lilla familj Martin, Adrian, Wilmer och Syskon. Utan er vore jag inget. Nu flyttar vi till Kumla skola!

10 REFERENCES

Alexeyenko A, Lindberg J, Perez-Bercoff A, Sonnhammer ELL. 2006. Overview and comparison of ortholog databases. Drug Discov Today Tech. 3:137-143.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 5:e1000262.

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36:D419-425.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 310:311-325.

Babenko V, Rogozin I, Mekhedov S, Koonin E. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res. 32:3724-3733.

Bashton M, Chothia C. 2002. The geometry of domain combination in proteins. J Mol Biol. 315:927-939.

Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain rearrangements in protein evolution. J Mol Biol. 353:911-923.

Blake CCF. 1978. Do genes-in-pieces imply proteins-in-pieces? Nature 273:267.

Blake CC. 1979. Exons encode protein functional units. Nature 277:598.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Patterns of intron gain and conservation in eukaryotic genes. BMC Evol Biol. 7:192.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 17:1034-1044.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. Nature 315:283-284.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. Trends Genet. 7:145-148.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282:2012-2018.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2:e383.

Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA. 2009. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 37:D310-314.

Deng X, Yin X, Allan R, Lu DD, Maurer CW, Haimovitz-Friedman A, Fuks Z, Shaham S, Kolesnick R.
2008. Ceramide biogenesis is required for radiation-induced apoptosis in the germ line of C. elegans. Science 322:110-115.

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. 1998. Towards a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 95:5094-5099.

Dewji NN, Singer SJ. 1997. The seven-transmembrane spanning topography of the Alzheimer disease-related presenilin proteins in the plasma membranes of cultured cells. Proc Natl Acad Sci USA 94:14025-14030.

Dewji NN, Valdez D, Singer SJ. 2004. The presenilins turned inside out: implications for their structures and functions. Proc Natl Acad Sci USA 101:1057-1062.

Dibb NJ, Newman AJ. 1989. Evidence that introns arose at proto-splice sites. EMBO J. 8:2015-2021.

Dibb NJ. 1991. Proto-splice site model of intron origin. J Theor Biol. 151:405-416.

Doan A, Thinakaran G, Borchelt DR, Slunt HH, Ratovitsky T, Podlisny M, Selkoe DJ, Seeger M, Gandy SE, Price DL, Sisodia SS. 1996. Protein topology of presenilin 1. Neuron 17:1023-1030.

Doolittle RF. 1995. The multiplicity of domains in proteins. Annu Rev Biochem. 64:287-314.

Doolittle WF. 1978. Genes in pieces – were they ever together? Nature 272:581-582.

Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P. 2005. Analysis of xbx genes in C. elegans. Development 132:1923-1934.

Ekman D, Björklund AK, Frey-Skött J, Elofsson A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 348:231-243.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci USA 99:16128-16133.

Ferrier DE, Minguillon C, Holland PWH, Garcia-Fernandez J. 2000. The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. Evol Dev. 2:284-293.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. 2008. The Pfam protein families database. Nucleic Acids Res. 36:D281-288.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99-113.

Fong JH, Geer LY, Panchenko AR, Bryant SH. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 366:307-315.

Fraering PC, Ye W, Strub JM, Dolios G, LaVoie MJ, Ostaszewski BL, van Dorsselaer A, Wang R, Selkoe DJ, Wolfe MS. 2004. Purification and characterization of the human gamma-secretase complex. Biochemistry 43:9774-9789.

Franck E, Madsen O, van Rheede T, Ricard GN, Huynen MA, de Jong WW. 2004. Evolutionary diversity of vertebrate small heat shock proteins. J Mol Evol. 59:792-805.

Friedmann E, Lemberg MK, Weihofen A, Dev KK, Dengler U, Rovelli G, Martoglio B. 2004. Consensus analysis of signal peptide peptidase and homologous human aspartic proteases reveals opposite topology of catalytic domains compared with presenilins. J Biol Chem 279:50790-50798.

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9:235.

Gilbert W. 1978. Why genes in pieces? Nature 271:501.

Gilbert W. 1987. The exon theory of genes. Cold Spring Harb Symp Quant Biol. 52:901-905.

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

Gough J. 2005. Convergent evolution of domain architectures (is rare). Bioinformatics 21:1464-1471.

Gulbins E, Li PL. 2006. Physiological and pathophysiological aspects of ceramide. Am J Physiol Regul Integr Comp Physiol. 290:R11-R26.

Hart GW, Brew K, Grant GA, Bradshaw RA, Lennarz WJ. 1979. Primary structural requirements for the enzymatic formation of the N-glycosidic bond in glycoproteins. Studies with natural and synthetic peptides. J Biol Chem. 254:9747-9753.

Holland SK, Blake CC. 1987. Proteins, exons and molecular evolution. Biosystems 20:181-206.

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. 2007. Ensembl 2007. Nucleic Acids Res. 35:D610-D617.

Hulsen T, Huynen MA, de Vlieg J, Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7:R31.

Jaenicke R. 1987. Folding and association of proteins. Prog Biophys Mol Biol. 49:117-237.

Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. Trends Genet. 22:16-22.

Jones DT, Taylor WR, Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33:3038-3049.

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

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231-237.

Kimberly WT, Wolfe MS. 2003. Identity and function of gamma-secretase. J Neurosci Res. 74:353-360.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biology Direct 1:22.

Kornilova AY, Kim J, Laudon H, Wolfe MS. 2006. Deducing the transmembrane domain organization of presenilin-1 in gamma-secretase by cysteine disulfide cross-linking. Biochemistry 45:7598-7604.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567-580.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21:25-30.

Kuzniar A, van Ham RC, Pongor S, Leunissen JA. 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 24:539-551.

Laudon H, Hansson EM, Melén K, Bergman A, Farmery MR, Winblad B, Lendahl U, von Heijne G, Näslund J. 2005. A nine-transmembrane domain topology for presenilin 1. J Biol Chem. 280:35352-35360.

Lehmann S, Chiesa R, Harris DA. 1997. Evidence for a six-transmembrane domain structure of presenilin 1. J Biol Chem. 272:12047-12051.

Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 34:D572-580.

Li L, Stoeckert CJ Jr, Roos DS. 2003. Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178-2189.

Li X, Greenwald I. 1996. Membrane topology of the C. elegans SEL-12 presenilin. Neuron 17:1015-1021.

Li X, Greenwald I. 1998. Additional evidence for an eight-transmembrane-domain topology for Caenorhabditis elegans and human presenilins. Proc Natl Acad Sci USA 95:7109-7114.

Logsdon JM Jr. 1998. The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev. 8:637-648.

Mourier T, Jeffares DC. 2003. Eukaryotic intron loss. Science 300:1393.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247:536-540.

Nakai T, Yamasaki A, Sakaguchi M, Kosaka K, Mihara K, Amaya Y, Miura S. 1999. Membrane topology of Alzheimer's disease-related presenilin 1. Evidence for the existence of a molecular species with a seven membrane-spanning and one membrane-embedded structure. J Biol Chem. 274:23647-23658.

Nguyen HD, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol. 1:e79.

Nyborg AC, Jansen K, Ladd TB, Fauq A, Golde TE. 2004. A signal peptide peptidase (SPP) reporter activity assay based on the cleavage of type II membrane protein substrates provides further evidence for an inverted orientation of the SPP active site relative to presenilin. J Biol Chem. 279:43148-43156.

Oh YS, Turner RJ. 2005. Topology of the C-terminal fragment of human presenilin 1. Biochemistry 44:11821-11828.

Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. Structure 5:1093-1108.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. 2009. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. Nov 5 *(in press)*.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896-2901.

Park J, Lappe M, Teichmann SA. 2001. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol. 307:929-938.

Qiu WG, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol. 21:1252-1263.

Remm M, Sonnhammer E. 2000. Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs. Genome Res. 10:1679-1689.

Remm M, Storm CEV, Sonnhammer ELL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 314:1041-1052.

Riddle DL, Blumenthal T, Meyer BJ, Priess JR. 1997. C. elegans II: Chapter 1, Introduction to C. elegans. Cold Spring Harbor Laboratory Press. Editors: Riddle DL, Blumenthal T, Meyer BJ, Priess JR.

Robertson HM. 1998. Two large families of chemoreceptor genes in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. 8:449-463.

Robertson HM. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. 10:192-203.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13:1512-1517.

Rossmann MG, Moras D, Olsen KW. 1974. Chemical and biological evolution of nucleotide-binding protein. Nature 250:194-199.

Rost B, Fariselli P, Casadio R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci. 5:1704-1718.

Roth AC, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 9:518.

Roy SW. 2003. Recent evidence for the exon theory of genes. Genetica 118:251-266.

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci USA 100:7158-7162.

Roy SW, Gilbert W. 2005. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci USA 102:5773-5778.

Roy SW, Penny D. 2007. On the incidence of intron loss and gain in paralogous gene families. Mol Biol Evol. 24:1579-1581.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R. 2008. TreeFam: 2008 Update. Nucleic Acids Res. 36:D735-740.

Sadusky T, Newman AJ, Dibb NJ. 2004. Exon junction sequences as cryptic splice sites: Implications for intron origin. Curr Biol. 14:505-509.

Simmer F, Moorman C, Van Der Linden AM, Kuijk E, Van Den Berghe PV, Kamath R, Fraser AG, Ahringer J, Plasterk RH. 2003. Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions. PLoS Biol. 1:e12.

Sonnhammer EL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 18:619-620.

Spasic D, Tolia A, Dillen K, Baert V, de Strooper B, Vrijens S, Annaert W. 2006. Presenilin-1 maintains a nine-transmembrane topology throughout the secretory pathway. J Biol Chem. 281:26569-26577.

Stoltzfus A, Spencer DF, Zuker M, Logsdon JMJ, Doolittle WF. 1994. Testing the exon theory of genes: the evidence from protein structure. Science 265:202-207.

Stoltzfus, A. 1999. On the possibility of constructive neutral evolution. J Mol Evol. 49:169-181.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2004. Reconstruction of ancestral protosplice sites. Curr Biol. 14:1505-1508.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2005. Conservation versus parallel gains in intron evolution. Nucleic Acids Res. 33:1741-1748.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The cog database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tusnády GE, Simon I. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849-850.

Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA. 2004. Supra-domains: evolutionary units larger than single protein domains. Mol Biol. 336:809-823.

Vogel C, Teichmann SA, Pereira-Leal J. 2005. The relationship between domain duplication and recombination. J Mol Biol. 346:355-365.

von Heijne G. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. 5:3021-3027.

Wang M, Caetano-Anollés G. 2006. Global phylogeny determined by the combination of protein domains in proteomes. Mol Biol Evol. 23:2444-2454.

Weiner J 3rd, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. FEBS J. 273:2037-2047.

Weiner J 3rd, Bornberg-Bauer E. 2006. Evolution of circular permutations in multi-domain proteins. Mol Biol Evol. 23:734-743.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35:D5-D12.

Wuchty S. 2001. Scale-free behavior in protein domain networks. Mol Biol Evol. 18:1694-1702.

Zhang P, Toyoshima C, Yonekura K, Green NM, Stokes DL. 1998. Structure of the calcium pump from sarcoplasmic reticulum at 8-A resolution. Nature 392:835-839.