Domain Architecture Comparison for Multidomain Homology Identification

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412) 268-7129

R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412) 268-7129

D. Durand*,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

March 9, 2007

*corresponding author

Abstract

Homology identification is the first step for many genomic studies. Current methods, based on sequence comparison, can result in a substantial number of mis-assignments due to the similarity of homologous domains in otherwise unrelated sequences. Here we propose methods to detect homologs through explicit comparison of protein domain content. We developed several schemes for scoring the homology of a pair of protein sequences based on methods used in the field of information retrieval. We evaluate the proposed methods and methods used in the literature using a benchmark of fifteen sequence families of known evolutionary history. The results of these studies demonstrate the effectiveness of comparing domain architectures using these similarity measures. We also demonstrate the importance of both weighting promiscuous domains and of compensating for the statistical effect of having a large number of domains in a protein. Using logistic regression we demonstrate the benefit of combining similarity measures based on domain content with sequence similarity measures.

1 Introduction

Figure 1

In this work, we investigate explicit comparison of domain architecture in predicting homology. A domain is a sequence fragment that will fold independent of context into a protein subunit with specific shape and function. Multi-domain families evolve through replication of genes (speciation and gene duplication) and domain shuffling (insertion, deletion, rearrangement, and internal duplication of domains). As a result, multi-domain families have properties that challenge current homology identification methods. Sequences in the same family can have different domain composition (e.g., A, B, C and D in Fig. 1), while sequences that are not descended from a common ancestral gene may share a homologous domain (e.g., D and E in Fig. 1). Current methods, based on sequence comparison result in false positives: unrelated sequences that have significant sequence similarity due to a shared domain will be incorrectly identified as homologs. This problem could be solved by making identical domain content or full-length alignments a requirement for homology identification. However, that would eliminate true homologs with different domain architectures (e.g., C and D), resulting in false negatives.

The need for accurate multi-domain homology identification is urgent. Multi-domain proteins represent a substantial fraction of the proteome: about 27% of proteins in bacteria and 39% of proteins in metazoa are multi-domain proteins (Tordai et al., 2005). Moreover, these are proteins of particular functional importance. In metazoa, complex multi-domain families are involved in cell-cell signaling,

cellular adhesion, and cellular migration, functions crucial to the evolution of multicellularity (Ben-Shlomo et al., 2003; Miyata and Suga, 2001). Multi-domain proteins are also implicated in many areas of human health including immune response and tissue repair (Patthy, 2003; Aravind et al., 2001; International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Since cancer typically arises from failures in signaling or apoptosis, most oncogenes are multi-domain.

Accurate homology detection is fundamental to a wide range of genomic applications, including the investigation of whole genome duplications (Durand and Hoberman, 2006; Van de Peer, 2004; Simillion et al., 2004), comparative mapping (Nadeau and Sankoff, 1998; O'Brien et al., 1997), phylogenetic footprinting (Tagle et al., 1988; Blanchette and Tompa, 2002; Duret and Bucher, 1997; Zhang and Gerstein, 2003; Dickmeis and Muller, 2005) and functional annotation (Wu et al., 2003). Pairwise homology identification is also the first step in graphical protein classification methods (Redfern et al., 2005; Orengo and Thornton, 2005; Wu et al., 2003, and work cited therein).

In this work, we propose and evaluate homology identification methods for this important and challenging class of sequences. We investigate the use of explicit comparison of domain architecture for this purpose. To develop measures of domain architecture comparison, we exploit an analogy between domain architecture composition and a problem in information retrieval, namely, determining the similarity of two documents drawn from a textual corpus. In this metaphor, the word content of a document is analogous to the domain content of a protein sequence and the set of protein sequences under study is analogous to the set of documents in the corpus. Based on this metaphor, we adapt information retrieval techniques to domain architecture comparison as a method for identifying multi-domain homologs. We evaluate the effectiveness of the proposed methods empirically, applying each method to test sets composed of positive examples (pairs of sequences drawn from the same family) and negative examples (pairs drawn from different families). We use this empirical approach to determine:

- whether domain architecture comparison is, in fact, an effective method for identifying homologs,
- what properties of domain architectures are most informative for this purpose.

For testing, we manually curated a novel benchmark dataset of multi-domain families. Using this testing procedure, we also compare our comparison methods with other methods previously proposed in the literature.

The goal of the current paper is to determine the effectiveness of domain architecture comparison for multi-domain homology identification. For some applications, such as the Conserved Domain Architecture Retrieval Tool (CDART) (Geer et al., 2002), this is an end in itself. For other applications,

a method that works well for both single- and multi-domain families is desired. Methods based on domain architecture comparison are not applicable to single domain families, which can be effectively classified using existing sequence comparison methods. Combining sequence similarity with domain architecture comparison is a promising direction for future studies. To investigate the potential of this approach, we also evaluated one method for combining domain architecture comparison and sequence comparison.

The rest of the paper proceeds as follows: In the next section, we introduce a formal definition of homology for multi-domain families and discuss the curation of our dataset according to this definition. After reviewing related work by other authors in Section 3, we introduce a set of similarity and distance measures based on domain architecture comparison in Section 4. These take various domain architecture properties into account, including the number of domains in a protein, repeated domains, and the promiscuity of individual domains. We describe our experimental methodology and the results of our experiments in Sections 5 and 6, respectively. We conclude that the best of the domain architecture comparison methods studied identify homologs with few prediction errors and that these methods can be further improved by combining them with sequence similarity. We summarize the domain architecture properties that proved to be most informative for homology identification and outline promising directions for future work.

2 A model and benchmark for multi-domain homology

Multi-domain sequences evolve by gene replication, domain shuffling and sequence mutation. Domain shuffling is mediated by non-allelic homologous recombination, unequal crossing over, retrotransposition and read-through errors (Gilbert, 1987; Patthy, 1999; Eichler, 2001; Emanuel and Shaikh, 2001; Kaessmann et al., 2002; Wang et al., 2000; Long, 2001; Long and Thornton, 2001). These events can cause duplication and insertion of a domain into a novel genomic context, tandem duplication of domains, domain deletion, and gene fusion and fission. Certain domains appear in many, unrelated domain architectures. These *promiscuous* domains (Marcotte et al., 1999) typically carry out auxiliary functions, such as conferring interaction specificity, acting as adapters, or mediating adhesion. Promiscuous domains have signature properties associated with the requirement that promiscuous domains be mobile (Patthy, 1999). In order to be retained following insertion, promiscuous domains must be able to fold autonomously. Promiscuous domains are typically short (100 residues or less), consistent with this requirement (Tordai et al., 2005). In addition, promiscuous domains must be encoded by

symmetric phase exons; that is, exons that are a multiple of three nucleotides in length (Patthy, 1985). A non-symmetric fragment that is inserted into an intron will disrupt the reading frame of downstream exons and most likely be deleterious. It has been shown that promiscuous domains have 1-1 symmetric phase, i.e. are flanked by phase 1 introns, more often than expected by chance (Patthy, 1999).

2.1 Model

In the context of molecular evolution, sequences are defined to be homologous if they are descended from a common ancestral gene via speciation (orthology) or gene duplication (paralogy) (Fitch, 1970, 2000). While this definition was proposed before domain shuffling was well understood, it can be directly extended to multi-domain families that evolve through domain insertion *into existing genes* and domain deletion, in addition to speciation and gene duplication. This is because for these domain shuffling events, it is still possible to trace ancestry by vertical descent. Fig. 1 illustrates this definition of multi-domain homology. Genes y and D originated by duplication of gene x. Although both genes subsequently sustained a domain insertion, they are clearly descended from the same ancestral gene. Therefore, A, B, C, and D form a homologous family. In contrast, although genes D and E share a homologous domain, they do not share a common ancestral gene and are not homologous.

This model does not apply to novel architectures that arise through independent insertions of multiple domains into a non-coding region, followed by *de novo* acquisition of regulatory machinery necessary for transcription. Since in this case it is not possible to trace the ancestry back to a single ancestral gene, we consider the new gene to be the originator of a novel gene family. Current evidence suggests that this latter scenario is relatively rare (Fong et al., 2007) and that new domain architectures are more likely to arise through acquisition of a new domain by an existing gene. This scenario is consistent with the histories of several recently discovered genes that arose so recently that evidence of specific domain shuffling events is still discernible (Sayah et al., 2004; Long et al., 2003; Vinckenbosch et al., 2006; Finnegan, 1989; Begun, 1997; Jones et al., 2005). In addition, it has been shown that many metazoan families evolved through a pattern of duplication, insertion of domains, and further duplication (Koyanagi et al., 1998; Miyata and Suga, 2001; Ben-Shlomo et al., 2003). These families are also consistent with the model.

Our model exhibits the correct behavior for comparative genomics applications, such as comparative map construction, identification of paralogous chromosomal regions and phylogenetic footprinting, that use homologous genes as markers. This is illustrated in Fig. 2. According to our model, $g2$ and

$g2'$ are homologous because they are descended from the same ancestral region in Chromosome A. In contrast, $g2$ and $g3'$ descended from different ancestral regions and are not homologous. For the purposes of comparative mapping, it is appropriate to associate $g2$ with $g2'$ but not $g3'$. Figure 2

2.2 Benchmark

Empirical evaluation of multi-domain homology detection methods requires a benchmark of multi-domain families that share common ancestry according to definition of multi-domain homology proposed above. Proteins are encoded by genes that are embedded in a specific chromosomal context, as shown in Fig. 2. This fact provides an operational definition for identifying sets of homologous multi-domain sequences. Even when domain shuffling has occurred, evidence of common ancestry can be obtained by considering gene structure and genomic context. Examples of such evidence include conserved synteny (i.e., genes that have homologous neighbors are likely to be homologous), conserved intron/exon structure, and conserved domain architecture. Congruent domain phylogenies also supports common ancestry, as does coevolution with a single-domain binding partner (e.g., coevolution of ligand and receptor families.)

Using this operational definition, we curated a benchmark dataset for empirical evaluation of the methods proposed in our study. Our test set is based on a synthesis of over 70 publications. Table 1 Our primary criterion in selecting families was strong evidence of common ancestry. In particular, we required evidence that all members of the family were descended from the same ancestral gene in the manner illustrated in Figs. 1 and 2. No restriction was placed on the age of this common ancestor. Some families, e.g. , ACSL, originated in bacteria, while others, e.g. , TNFR, originated in the chordate lineage. Families were selected to represent a range of functional and architectural properties. The benchmark includes intracellular, extracellular, and transmembrane proteins that carry out a broad range of molecular functions in the context of varied cellular processes (see Table 2). Table 2

The formation of multi-domain families is influenced by the structural, functional, and selective constraints of domain shuffling. For this reason, multi-domain families tend to fall into a small number of organizational patterns. These include families with conserved domain architectures, families with varied domain architectures, and families characterized by arrays of tandem repeats, often with variations in copy number. Families with variable domain architectures can further be partitioned into families characterized by a single defining, or *key*, domain, and families characterized by two or more key domains. Families with variable architectures often have subfamilies.

All of these patterns are represented in our test set, allowing us to study how the various or-

ganizational patterns influence homology identification methods. Families with conserved domain architectures, represented by DVL, FOX and NOTCH, expanded by gene duplication following formation of a progenitor domain architecture by domain shuffling. Kinase, Kinesin, Myosin and TNFR are examples of multi-domain families that are characterized by a single key domain that carries out the primary function of the family. This key domain is partnered with different auxiliary domains that perform "helper" functions and determine the specificity of individual family members. These families arose through duplication of an ancestral gene, followed by insertion of different domains into the resulting paralogs to create a family of diverse domain architectures (Ben-Shlomo et al., 2003; Miyata and Suga, 2001). These architectures were the progenitors of modern subfamilies, which further expanded through a second wave of duplication, resulting in multiple copies of each architecture. Fig. 3 shows seven members of the Kinesin family. All have at least one kinesin domain and most are partnered with other domains, many of which are promiscuous. Figure 3

Other families are characterized by two or more key domains that define the family. For example, all ADAM proteins are characterized by a pre-prodomain (Pep_M12B) that is processed after activation, and a catalytic domain (Astacin) that carries out ADAM's enzymatic function. These two domains Figure 4 are partnered with a variety of C-terminal domains that determine functional specificity. In particular, the C termini of members of the ADAMTS subfamily contain arrays of promiscuous thrombospondin (TSP1) repeats. Fig. 3 shows two members of the ADAM family, one representative from each of the ADAM subfamilies, ADAM and ADAMTS. In our dataset, the TRAF and GATA families are also characterized by two or more key domains.

Architectures characterized by tandem repeats of promiscuous domains are represented by KIR, NOTCH and Laminin. These families share properties with both conserved and variable families, Figure 5 since the number of different domain types is small, but the variations in organization and copy number can be complex. Fig. 5 shows examples of the three Laminin subfamilies, all of which are characterized by long, complex architectures made up of a small number of domain types. The promiscuous domains and high copy number repeats found in these families are a particular challenge for homology identification.

3 Previous Work

Several studies have proposed similarity or distance measures for domain architectures in the context of other applications. Since these methods were developed for other purposes, the effectiveness of these

methods in determining multi-domain homology has not been systematically evaluated on a manually curated dataset. CDART (Geer et al., 2002), a web-based resource, presents a list of sequences with similar domain architecture to a given query sequence. Similarity is calculated by counting the number of domains shared between the query and matching sequences. This tool is designed to retrieve similar architectures but does not consider whether or not similar architectures are homologous.

Lin et al. (2006) designed a domain architecture similarity measure to reconstruct eukaryotic orthologous groups in the KOG database (Tatusov et al., 2003). Their similarity score is a weighted average of three terms, reflecting the number of shared domain types, domain order conservation, and similarity in domain copy number, respectively. The coefficients were determined using an optimization procedure on a training set derived from the KOG database. The authors did not carry out an empirical evaluation of the ability of the similarity measure to distinguish homologous pairs from non-homologous pairs on a separate test set. The goal of that work differs from ours in that it it focuses on orthologs, which are typically closely related: 65% of KOGs contain a single domain architecture and 85% contain no more than two distinct architectures (Lin et al., 2006). In contrast, our goal is to design methods that can classify all multi-domain homologs, including families with diverse architectures.

Two recent papers have proposed distance measures for domain architecture comparison in the context of studies to infer ancestral domain architectures and the rates of various domain shuffling events. Bjorklund et al. (2005) define the domain architecture distance to be the number of domain insertions, deletions, substitutions, and internal duplications needed to transform one architecture into the other. Fong et al. (2007) embed domain architectures in the leaves of a species tree. Based on this embedding, they use a parsimony condition to infer the domain architectures in the ancestral species and the number of events that occurred on each branch of the species tree. The use of a species tree is a promising approach to more accurate distance measures, but can only be applied to domain architectures in species that have been fully sequenced.

4 Domain Architecture Comparison Measures

We here introduce new domain architecture comparison methods, based on an analogy with document similarity in information retrieval. In each case, a score is assigned to each pair of sequences based on domain content. The scores are then used as a basis for classifying sequence pairs as homologous or unrelated. In practice this means choosing a cutoff so that all pairs with scores above that cutoff are

predicted to be homologous.

The current work focuses on proteins represented as domain architectures. Domain databases (Letunic et al., 2002; Bateman et al., 2002; Corpet et al., 1998; Gracy and Argos, 1998; Heger and Holm, 2003) store probabilistic, sequence-based models of protein domains that can be used to determine the location and type of domains in amino acids sequences. By comparison with a set of such models, each protein sequence is converted into linear sequence of the domains it contains, ignoring any differences in the linker sequences between domains. Any two instances of the same domain will have some variance at the sequence level. To simplify the comparison, we treat all instances of a domain as though they were identical.

A simple measure of the similarity of two documents is the number of shared words. Documents that share a significant number of words are more likely to be on the same subject than documents that share few words. Similarly, two proteins that share many domains are more likely to be homologs than sequences with few domains in common. We consider two similarity measures based on shared domain content: the number of distinct *domain types* found in both architectures and the total number of *domain copies* that are shared. For example, the domain architectures ABBBC and ABBD share three domain copies (one A and two B's), but only two domain types (A and B). In counting domain copies, we do not distinguish between distant and tandem copies. In our dataset, 91% of all duplicated domains are found in tandem arrays, suggesting that this distinction is unlikely to be informative.

Under this naïve similarity measure all words (or domains) are equally important. However, words in a document are not equally important to determining its subject. The word "cheap" conveys less information about the subject of document than the word "parsimony." One measure of the power of a given word to distinguish between subjects is the *inverse document frequency* (Salton and Buckley, 1988), a measure based on the observation that a word that occurs in very few documents is more likely to differentiate between subjects than a word that occurs frequently.

Similarly, some domains are more informative than others. Since promiscuous domains occur in many unrelated sequences, they are less useful for determining homology than relatively rare domains. To deemphasize promiscuous domains, we adapt the inverse document frequency to obtain an *idf weight* for a domain d of

$$w_{\text{idf}}(d) = \log_2 \frac{|\mathcal{P}|}{|\{p|d \in D(p), p \in \mathcal{P}\}|}, \tag{1}$$

where $\mathcal{P}$ is the set of proteins in the dataset, and $D(p)$ is the multi-set of domains in a protein sequence, p. The denominator is the number of proteins in the dataset that contain the domain d.

The frequency of a word in a given document is also an indication of its importance to that

document. The word "parsimony" is more likely to be relevant to a document in which it appears five times than to a document in which it occurs only once. This aspect can be expressed by the *term frequency*, the number of times a word appears in a document amortized by the total number of words in that document. In multi-domain sequences, term frequency is analogous to domain copy number. We define the term frequency,

$$w_{\text{tf}}(d,p) = \frac{N(d,p)}{|D(p)|}, \tag{2}$$

where $N(d,p)$ is the number of occurrences of the domain d in the protein p and $|D(p)|$ is the number of domains in the protein p. In document retrieval, words are typically weighted using a *tf-idf weight*, the product of w_{tf} and w_{idf},

$$w_{\text{tf-idf}}(d,p) = w_{\text{tf}}(d,p) \times w_{\text{idf}}(d). \tag{3}$$

Note that comparing the results from idf weighting with *tf-idf weighting* yields a measure of the importance of copy number to multi-domain homology detection.

We also consider a measure of domain promiscuity we call *distinct partner* weighting, based on the observation that promiscuous domains co-occur with many different domains types. Each domain is weighted by the inverse of the number of distinct domain types with which it co-occurs:

$$w_{\text{dp}}(d) = \frac{1}{|\{d_i | d_i \in D(p), d \in D(p), p \in \mathcal{P}\}|}. \tag{4}$$

We refer to this weighting method as *distinct partner* weighting. To our knowledge, the analogous weighting method has not been used in information retrieval.

For each of these three weighting schemes (idf alone, tf-idf, or distinct partner weighting), the weighted domain comparison score is given by

$$S(p_1,p_2) = \sum_i w(d_i,p_1) w(d_i,p_2), \tag{5}$$

where $w(d_i,p_1)$ and $w(d_i,p_2)$ are the weights for domain d_i in sequence p_1 and p_2. The sum runs over all domains in the dataset, where $w(d_i,p) = 0$ if d_i does not occur in the protein p.

A final consideration is size. Long documents are more likely to share a large number of words than shorter documents. To correct for this effect, a length correction is used, where length is typically measured by word count. In the domain architecture comparison, correcting for *domain count* allows comparisons between sequences with different numbers of domains. A measure of unweighted similarity with a domain-count correction is given by the *Jaccard similarity*:

$$J(p_1,p_2) = \frac{n_{12}}{n_1 + n_2 - n_{12}}, \tag{6}$$

where n_{12} is the number of domains common[1] to both sequences p_1 and p_2, n_1 is the number of domains in p_1, and n_2 is the number of domains in p_2. *Cosine similarity* can be used with either weighted or unweighted domains. For two proteins p_1 and p_2, the cosine similarity is given by

$$C(p_1, p_2) = \frac{\sum_i w(d_i, p_1) w(d_i, p_2)}{\sqrt{\sum_i w(d_i, p_1)^2 \sum_i w(d_i, p_2)^2}}. \quad (7)$$

If the protein sequences are treated as vectors in a vector space with number of dimensions equal to the number of domain types, then cosine similarity is the cosine of the angle between the two weight-vectors. Note that since w_{tf} includes the total number of domains in protein sequence p in the denominator, the tf-idf weight implicitly corrects for the number of domains in a protein even when cosine similarity is not used.

In this study, we evaluate the classification performance of unweighted similarity based on domain-type and domain-copy number, as well as all three weighting schemes (idf, tf-idf, and distinct partner). Each scoring scheme is tested with and without a domain-count correction. The set of similarity scoring schemes that we proposed and tested is summarized in lines 1 - 12 of Table 3.

For comparison, we also implemented and tested several domain architecture comparison measures proposed by other authors. The domain architecture similarity used in CDART (Geer et al., 2002) is identical to the number of shared domain copies without domain-count correction (S1 in Table 3). The similarity score proposed by Lin et al. (2006) is a weighted average of three terms. The first is the Jaccard similarity using common domain types given in Equation 6. The second term, the Goodman-Kruskal γ index, counts the number of domain pairs that occur in the same order in both sequences. The third term is a measure of domain-copy number in the two architectures. We have implemented this weighted similarity score (listed as S13 in Table 3) using the optimal weights given by Lin et al. (2006) (0.36, 0.01, 0.63).

In addition to similarity scoring, we also investigated the edit distance between two domain architectures proposed by Bjorklund et al. (2005), using a dynamic programming algorithm similar to the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Note that edit distance depends on the order of the domains: the distance between proteins that have the same domain content, but with the domains in a different order, is non-zero. To determine the importance of order for multi-domain homology identification, we also implemented a distance metric that simply counts the number of domains that differ between the two proteins and does not depends on the order of the domains.

Although this was not proposed by Bjorklund et al. (2005), we also considered the impact of

[1]Note that Jaccard similarity can be calculated for shared domain copies or shared domain types by selecting the appropriate values of n_1, n_2 and n_{12}.

penalizing promiscuous domains by using a distinct partner weight in the determination of the edit distance. Additionally, we consider domain-count correction on the distance methods, both with order and without order. For distance measures, a domain-count correction was obtained by dividing the uncorrected distance by

$$\sum_{i \in D(p_1)} w(d_i, p_1) \; + \sum_{i \in D(p_2)} w(d_i, p_2) \; - \sum_{i \in D(p_1) \cap D(p_2)} \sqrt{w(d_i, p_1)\, w(d_i, p_2)}, \tag{8}$$

which, in the unweighted case, reduces to the size of the union of the sets of domains in both proteins. The distance methods tested are listed in lines 14 - 21 of Table 3. The method proposed in Fong et al. (2007) could not be tested as our dataset is not appropriate for its application. Table 3

5 Materials and Methods

In this section, we discuss evaluation of the methods in Table 3 and introduce our testing procedure. For empirical evaluation, we use the benchmark dataset discussed in Section 2. We developed this benchmark as no suitable, manually curated public datasets exist, to our knowledge. Structural data bases, such as SCOP (Murzin et al., 1995) and CATH (Orengo et al., 2003; Pearl et al., 2003), are excellent for validating homology predictions for individual domains but are not suitable for multi-domain families with diverse architectures. Similarly, COGs (Tatusov et al., 2003) and Gene3D (Yeats et al., 2006) focus on conservatively defined families in which the domain architectures are identical or nearly so. Gene Ontology (GO) (Ashburner et al., 2000) is suitable for validating predictions of shared function but not shared ancestry.

5.1 Sequence Data

To construct our dataset, we extracted all complete mouse and human protein sequences from Swiss-Prot Version 44 (Bairoch et al., 2005), released in 09/2004, yielding 18,198 protein sequences. We focused on vertebrate data because the multi-domain families that challenge traditional homology identification methods tend to be larger and more complex in vertebrates (Aravind et al., 2001; Chothia et al., 2003; Patthy, 2003; Li et al., 2001; International Human Genome Sequencing Consortium, 2001; Venter et al., 2001; Wuchty, 2001; Ye and Godzik, 2004; Wuchty and Almaas, 2005).

From this initial set of sequences, we constructed a list of sequences from fifteen known families using reports from Human Genome Nomenclature Committees[2], gene symbols, and high throughput

[2] http://www.gene.ucl.ac.uk/nomenclature/

annotations such as PFAM (Bateman et al., 2004) and Interpro (Mulder et al., 2005) codes. We then consulted articles by experts on specific families to refine the initial roster. This process resulted in a list of 1,137 proteins each known to be from the fifteen test families. A sample of articles consulted is given in Table 1.

5.2 Positive and Negative Examples

From this database of sequence families, we constructed a list of positive and negative examples of homologous pairs. We obtained protein domain architectures from CDART (Geer et al., 2002). Among 18,198 sequences, 15,826 sequences have one or more detectable domains. An all-against-all comparison of all 15,826 domain architectures yielded 2,371,608 pairs that share at least one domain. For each family, all pairs of family members that share at least one domain are positive examples of homologous pairs. All pairs sharing at least one domain, but where only one architecture in the pair is a member of the family, are negative examples. The number of homologous pairs for the fifteen families studied ranged from 75 pairs in the KIR family to 1,074,570 pairs in the Kinase family. The number of non-homologous pairs ranged from 78 in the FOX family to 184,107 in the Kinase family. The ability of all domain architecture comparison methods studied to correctly classify this set of positive and negative examples was the basis of our evaluation.

5.3 Combining E-value and Domain Architecture Information

In addition to evaluating various domain architecture comparison measures, we also considered the benefit of combining domain architecture comparison with sequence similarity. The difficulty with combining information from two sources is deciding how much each type of information should contribute to the final score. In this study, we address this problem using logistic regression to classify the data. Each pair of sequences in the dataset was assigned a score $a+bx+cy$, where x is the uncorrected domain architecture similarity with distinct partner weighting (S7) and y is $-\log$(e-value). E-values were calculated using all-against-all BLAST comparison (Altschul et al., 1997) using the BLOSUM 62 matrix and an affine gap penalty of $-10-k$ for a gap of length k.

Given these inputs, logistic regression classifies the data by maximizing the likelihood of the data given the scores, and implicitly learns the parameters a, b, and c. We used the logistic regression routine in the R statistical package (R Development Core Team, 2006), with ten-fold cross-validation to estimate the classification accuracy. For comparison purposes, we also classified the data using the same logistic regression methodology applied to sequence similarity alone and to the domain

architecture comparison scores alone.

5.4 Performance Evaluation

We evaluated classifier performance using Area Under the receiver operating characteristic Curve (AUC) scores. AUC score provides a single measure of classification accuracy, corresponding to the fraction of correctly classified entities given the best possible choice of threshold (DeLong and DeLong, 1988). The range of AUC scores is from zero to one. The higher the AUC score, the better the performance. When the AUC is 1 the classifier has perfect performance. When the AUC is 0.5, the classifier cannot distinguish homologous pairs from non-homologous pairs, i.e. the score distributions for positive and negative examples are not separable. When the AUC score is less than 0.5 the positive and negative examples are somewhat separable, but the classifier is assigning higher scores to negative examples and lower scores to positive examples.

We also applied these methods to the combined set of all protein sequences in our dataset, in which the homologous and non-homologous pairs from all the families are combined to a single set of positive and negative examples, respectively. Because of the large size of the Kinase family, the combined results are dominated by the results on Kinase pairs. We therefore also considered the combined dataset of all families except the Kinase family.

Logistical regression combining sequence similarity results with domain architecture comparison was evaluated using three quantities: the True Positive Fraction (TPF), the True Negative Fraction (TNF), and the True Fraction (TF). The TPF is given by the number of homologous pairs predicted correctly, divided by the total number of homologous pairs in the dataset. Analogously, the TNF is the ratio of the number of correctly predicted, non-homologous pairs to the total number of non-homologous pairs. The TF is the ratio of the number of correct predictions to the total number of homologous and non-homologous pairs. The TF gives an indication of the overall performance of the classifier. For a test dataset, we use the combined dataset of all families except for Kinase, so that the results of the logistical regression are not dominated by the performance on the Kinase family.

6 Results

We used our manually curated benchmark to evaluate the performance of the domain architecture comparison methods summarized in Table 3. We first consider performance of unweighted similarity, the impact of correcting for domain-count, and the relative merits of domain-type similarity and

domain-copy similarity. We next evaluate the weighted similarity measures, followed by a comparison of similarity and distance approaches to domain architecture comparison. Finally, we consider the benefit of combining domain architecture comparison with sequence similarity.

6.1 Unweighted similarity

Table 4

Table 4 compares the performance of unweighted similarity scores based on shared domain-count and shared domain-type, with and without domain-count correction (S1 - S6). Unweighted similarity without a domain-count correction is an effective classifier for the ADAM, DVL, GATA, Notch, and Laminin families. Each of these families is characterized by two or more key domains, which occur in most family members. Since most non-homologous pairs share only one domain, homologous pairs have higher similarity scores. Unweighted similarity scoring is less effective in the other ten families, which generally have only one domain that is conserved throughout the family. In these families, different auxiliary domains occur in different family members and exist in many proteins outside of the family. Since homologous pairs typically share only one domain, they cannot be distinguished from non-homologous pairs that also share one domain.

A domain-count correction, using either Jaccard or cosine similarity, improves performance for most families. The exception is Myosin, where performance degrades with a domain-count correction. When family members have many domains but share only a few, a domain-count correction reduces the similarity score for homologous pairs. This occurs in the Myosin family.

How important is copy number in classifying multi-domain homologs? For both the average AUC score over all families and the combined dataset excluding Kinase, shared copy similarity performs slightly better then shared type similarity. When individual families are considered, a substantial performance improvement is obtained from shared copy similarity for the GATA and Laminin families. For the GATA family, domain-copy number is a defining characteristic of the family. All family members have two GATA domains, while sequences outside the family that have a GATA domain have only one. Therefore, shared copy similarity can classify the GATA family, while shared type similarity cannot. Laminins have long tandem arrays of promiscuous domains (see Fig. 5). Although the promiscuous domains result in matches with many unrelated sequences, homologous pairs have greater shared copy similarity because these repeats boost the number of shared copies. This advantage is not obtained with shared type similarity.

The Lin similarity score (S13) is a linear combination of shared type similarity with a Jaccard correction, a measure of domain order, and the number of shared copies. The overall performance is

similar to Jaccard-corrected shared copy similarity (S2). Lin performs well on GATA and Laminin, the families that particularly benefited from counting copy number. The impact of the domain order term is probably negligible because it is given little weight (a coefficient of 0.01). In summary, shared copy similarity with domain-count correction (either Jaccard or cosine, S2 or S3) gives similar performance to the Lin score for all families but is much simpler to calculate. Table 5

6.2 Weighted similarity

The performance of the weighted similarity scoring methods is compared in Table 5. Since unweighted shared type and shared copy similarity exhibit similar performance, we only consider shared copy similarity in these analyses. For each method, results are reported both with and without domain-count correction. For comparison, AUC scores for unweighted similarity scoring (S1 and S3) are reproduced in this table.

All weighting schemes to penalize promiscuous domains improved the performance of uncorrected, unweighted similarity scoring. The best results are obtained with distinct partner weighting. This can be understood by noting that idf penalizes domains that appear in many proteins. A domain may appear in many proteins either through duplication of a gene that carries the domain, or through duplication and insertion of the domain into a new gene. Thus, idf penalizes promiscuous domains, but also domains that define a family that has sustained repeated gene duplication. In contrast, distinct partner weighting, which assigns lower scores to domains with many partners, only penalizes domains that proliferated through domain duplication followed by insertion.

For example, the domain represented by the black square in Fig. 1 was inserted once into an ancestral protein. It appears in three different proteins due to gene duplication ($w_{idf} = 0.33$) but only partners with one other domain, the grey oblong ($w_{dp} = 1.0$). In contrast, the black triangle appears in two proteins due to domain insertions ($w_{idf} = 0.5$) and partners with two other domains ($w_{dp} = 0.5$). The black triangle, which experienced two insertions, is more promiscuous than the black square, which experienced only one. The distinct partner weight correctly weights the black triangle lower then the black square. The idf weight is lower for the black square, incorrectly implying it is more promiscuous then the black triangle. Figure 6

The TNFR and DEATH domains shown in Fig. 6 exhibit this type of behavior. The DEATH domain, which is truly promiscuous, also appears in ANK2, resulting in a non-homologous match. Distinct partner weighting correctly penalizes the DEATH domain, lowering the scores for non-homologous pairs. In contrast, idf penalizes the TNFR domain more than the DEATH domain and

does not perform well on this family.

All weighting methods fail for the Kinase family. Surprisingly, the distinct partner and idf score distributions *are* separable, but the distributions of scores are inverted: non-homologous pairs are given higher similarity scores than homologous pairs. The difficulty is that distinct partner weighting cannot distinguish between promiscuous domains that are frequently inserted into new contexts and non-promiscuous domains that are frequent targets of insertions. For example, there are four copies of the grey oblong in Fig. 1, all due to gene duplication. This domain has never been inserted into another gene. However, genes in its lineage have been the target of insertions by two other domains. Because of this, it has a distinct partner weight of 0.5 and appears to be just as promiscuous as the black triangle.

The evidence suggests that pkinase behaves similarly to the grey oblong. The pkinase domain has the largest number of distinct partners in our dataset and is penalized by all scoring schemes tested. However, the evidence in the literature does not support promiscuity. It is substantially longer than typical promiscuous domains and does not have 1-1 phase (Tordai et al., 2005). In addition, phylogenomic analysis indicates that the Kinase family evolved by duplication of an ancestral, single-domain Kinase, followed by insertion of different domains into the resulting paralogs, and a second round of duplication (Miyata and Suga, 2001; Ben-Shlomo et al., 2003). This suggests that the pkinase domain proliferated primarily by duplication. Identifying a scoring scheme that weights pkinase correctly remains an open problem.

What are the relative merits of domain-count correction and penalizing promiscuous domains? While there is a significant benefit to using either distinct partner weighting or a domain-count correction, there is little additional benefit to applying both. Domain-count correction can improve idf weighting, however. This is because promiscuous domains frequently occur in proteins with many domains. For example, the non-homologous match between TNFR1 and ANK2 in Fig. 6 receives a lower score because of the large number of domains in ANK2. In general, although idf alone performs poorly on the TNFR family, idf combined with cosine similarity performs well. Uncorrected tf-idf also performs well for TNFR because tf-idf implicitly adjusts for domain-count.

6.3 Unweighted and weighted distance

Table 6

Table 6 shows the classification performance of all distance scoring methods tested in this study. The results obtained with and without domain order were nearly identical: the AUC scores differ only in the third significant digit. For this reason, we show only one set of numbers; each column represents

classification accuracy both with and without order. Similarly, all discussions in this section apply equally well to the distance methods with and without order.

That domain order plays such a small role in determining multi-domain homology is consistent with the observation made in prior studies that co-occurring domains appear predominately in one order (Apic et al., 2001; Bashton and Chothia, 2002). A study by Apic et al. (2001) showed that pairs of domains from different superfamilies appeared in both orders in only 2% of cases studied. This observation is also consistent with the very small coefficient (0.01) assigned to the domain order term in the classifier proposed by Lin et al. (2006).

When uncorrected, unweighted measures are used, distance gives better overall performance than similarity on our dataset. However, distance scoring performs poorly for some families, such as Myosin and SEMA, due to the existence of subfamilies with differing domain architecture. These differences increase the distance between homologous pairs.

A domain-count correction improves the performance of unweighted distance scoring. The unweighted distance method with domain-count correction gives the same results as the unweighted similarity method with Jaccard domain-count correction (S2) (this is approximately true for the distance method with order correction, and exactly true for the distance method without order). This occurs because, when domain-count correction is applied, the distance and the similarity between two architectures sum to one. As a result, unweighted similarity with a cutoff of c identifies the same set of homologs as unweighted distance with a cutoff of $1 - c$, yielding identical AUC scores.

Penalizing promiscuous domains improved the classification performance for the ADAM, Laminin, Notch, and SEMA families. In ADAM and SEMA, the distance between homologous pairs is due to promiscuous domains present in one member of the pair but not the other. By giving promiscuous domains a lower weight, the distance between homologous pairs is reduced. For example, the two promiscuous TSP1 domains that are present in ADAMTS20 but not ADAM7 will contribute to the distance between them (Fig. 4). NOTCH and Laminin have long tandem arrays of promiscuous domains (e.g. , Fig. 5). Homologous pairs will share roughly the same number of copies, yielding a low distance score. There will be many matches to unrelated sequences containing the same promiscuous domains, but the copy numbers will be very different, resulting in larger distances.

For the majority of families, however, unweighted distance scoring outperforms the weighted model. Because, in many cases, matches with unrelated proteins are due to promiscuous domains, penalizing promiscuous domains reduces the distance between non-homologous pairs. Combining domain-count correction with domain weights compensates for this problem, yielding high performance overall. The

distance method with distinct partner weighting and domain-count correction has similar performance to the best similarity method observed, distinct partner weighting without domain-count correction.

6.4 Combining Sequence Based Methods with Domain Architecture Based Methods

Table 7

The results of the previous sections show that considering the domain organization of multi-domain sequences can be helpful in determining homology. Here we report the results of an initial experiment to determine whether combining domain architecture and sequence comparison provides additional benefits. Logistic regression was used to classify the data and to determine the relative weight of the contributions from domain architecture and sequence comparison. Four domain architecture comparison methods were considered: unweighted similarity, alone and with Jaccard correction;and weighted similarity with distinct partner weighting, alone and with cosine correction. The results are shown in Table 7. For comparison, the results of logistic regression applied to sequence similarity alone and domain architecture comparison alone are also shown.

When logistic regression is performed with domain architecture comparison alone, the TNF is close to one, indicating that the vast majority of non-homologous pairs were correctly eliminated. However, for all methods, the TPF was low. The best TPF was obtained with cosine-corrected distinct partner weighting, which also outperformed sequence similarity alone. For sequence similarity alone, TNF performance was good, but TPF performance is lackluster: only 57% of homologs were correctly identified. When sequence similarity information and domain architecture comparison methods are combined, the TNF performance remains excellent and the TPF performance significantly improves. Again, the best performance is obtained using distinct partner weighting with cosine length correction.

Our results show that combining sequence similarity with domain architecture comparison is a promising approach. The combined scores gave better accuracy than either measure alone. The addition of sequence information improves performance because it provides additional information about the linker sequences between domains and differences between two instances of the same domain type. Domain architecture comparison contributes information about high-level protein organization that is not captured by sequence comparison. How these two measures are best combined remains an open problem. Other machine learning approaches should be considered, with particular attention to the vagaries of specific datasets and methodological problems, such as over-fitting. The maximum likelihood approach associated with logistic regression optimizes overall classification accuracy. However, for many studies, an emphasis on minimizing false positives (or false negatives) may be preferred.

7 Discussion

Homology identification is a challenging problem for multi-domain sequences. Homologous families may have varied domain architectures, while otherwise unrelated sequences may share a homologous domain. Current methods, based on sequence comparison, are not suited to distinguishing between these two cases. Here, we proposed and evaluated the use of domain architecture comparison for classifying multi-domain homology.

Our results show that domain architecture can be highly effective for this purpose. Three key properties of domain architecture are informative for domain architecture comparisons: domain copy number, domain architecture length, and domain promiscuity. First, although for many families shared count similarity and shared type similarity work equally well, for families characterized by internal tandem duplications, domain count number is critical in determining sequence homology. Second, normalizing domain architecture comparison scores using the number of domains in each architecture significantly improved the performance for fourteen of the fifteen families studied. This normalization allows comparison of architectures with very different lengths. It also discounts the importance of the presence of any one domain in an architecture that has many. Finally, weighting each domain based on an estimate of its promiscuity greatly increased the accuracy of the method for all families except the Kinases. Of the weighting systems we tried, distinct partner weighting was most helpful. This is because distinct partner weighting can distinguish between domains that are prevalent due to repeated gene duplication and truly promiscuous domains that are prevalent because of frequent insertion into new contexts.

None of the weighting schemes, including distinct partner weighting, were able to distinguish between frequently inserted domains and domains that are frequently targets of insertions. It is for this reason that none of the proposed methods were effective in classifying the Kinase family. The pkinase domain co-occurs with many different partners, not because it is easily inserted, but because it is frequently in genes that are targets of insertions of other domains. Outstanding problems that we plan to address in future work include designing a weighting method that accurately captures true promiscuity in difficult cases like the pkinase domain.

Sequence comparison is an accurate classifier for single-domain homologs, while the methods presented here are effective in classifying multi-domain homologs. Combining sequence similarity with domain architecture comparison to obtain a universal method is a natural direction to pursue. Our preliminary results, based on logistic regression, show that this is a promising avenue for future re-

search. Relaxing the assumption that all instances of a domain are equivalent is another approach to incorporating sequence similarity in domain architecture comparison.

Acknowledgments

We thank S. H. Bryant and L. Y. Geer for providing the domain architecture models used in this work; J. Joseph for assistance with the manuscript; and A. Goldman for assistance with the figures. This research was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.

References

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

Apic, G., Gough, J., and Teichmann, S. A., 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310, 311–325.

Aravind, L., Dixit, V., and Koonin, E., 2001. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. Science 291, 1279–84.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S., 2005. The universal protein resource (UniProt). Nucleic Acids Res. 33, D154–9.

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

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. L., 2002. The Pfam protein families database. Nucleic Acids Res 30, 276–280.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R., 2004. The Pfam protein families database. Nucleic Acids Res 32, D138–41.

Begun, D. J., 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in drosophila. Genetics 145, 375–382.

Belkin, D., Torkar, M., Chang, C., Barten, R., Tolaini, M., Haude, A., Allen, R., Wilson, M. J., Kioussis, D., and Trowsdale, J., 2003. Killer cell Ig-like receptor and leukocyte Ig-like receptor transgenic mice exhibit tissue- and cell-specific transgene expression. J Immunol 171, 3056–63.

Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, W., Haili, and Hsueh, A. J. W., 2003. Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. Sci STKE 2003, RE9.

Bjorklund, A. K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A., 2005. Domain rearrangements in protein evolution. J Mol Biol 353, 911–923.

Blanchette, M. and Tompa, M., 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 12, 739–748.

Cheek, S., Zhang, H., and Grishin, N. V., 2002. Sequence and structure classification of kinases. J Mol Biol 320, 855–881.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A., 2003. Evolution of the protein repertoire. Science 300, 1701–1703.

Corpet, F., Gouzy, J., and Kahn, D., 1998. The ProDom database of protein domain families. Nucleic Acids Res 26, 323–326.

Degerman, E., Belfrage, P., and Manganiello, V. C., 1997. Structure, localization, and regulation of cGMP-inhibited phosphodiesterase (PDE3). J Biol Chem 272, 6823–6826.

DeLong, E. R. and DeLong, D. M., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Dickmeis, T. and Muller, F., 2005. The identification and functional characterisation of conserved regulatory elements in developmental genes. Brief Funct Genomic Proteomic 3, 332–350.

Durand, D. and Hoberman, R. A., 2006. Diagnosing duplications: can it be done? Trends Genet 22, 156–64.

Duret, L. and Bucher, P., 1997. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7, 399–406.

Eichler, E. E., 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 17, 661–669.

Emanuel, B. S. and Shaikh, T. H., 2001. Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet 2, 791–800.

Engel, J., 1992. Laminins and other strange proteins. Biochemistry 31, 10643–10651.

Finnegan, D. J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet 5, 103–107.

Fitch, W. M., 1970. Distinguishing homologous from analogous proteins. Syst Zool 19, 99–113.

Fitch, W. M., 2000. Homology: a personal view on some of the problems. Trends Genet 16, 227–231.

Fong, H., Jessica, Geer, Y., Lewis, Panchenko, R., Anna, and Bryant, H., Stephen, 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol 366, 307–315.

Foth, B., Goedecke, M., and Soldati, D., 2006. New insights into myosin evolution and classification. Proc Natl Acad Sci U S A 103, 3681–3686.

Geer, L. Y., Domrachev, M., Lipman, D. J., and Bryant, S. H., 2002. CDART: protein homology by domain architecture. Genome Res. 12, 1619–1623.

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

Goodson, H. V. and Dawson, S. C., 2006. Multiplying myosins. Proc Natl Acad Sci U S A 103, 3498–3499. Comment.

Gracy, J. and Argos, P., 1998. DOMO: a new database of aligned protein domains. Trends Biochem Sci 23, 495–497.

Hanks, S. K., 2003. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol 4, 111.

Heger, A. and Holm, L., 2003. Exhaustive enumeration of protein domain families. J Mol Biol 328, 749–767.

Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C., Mastwal, S., Zhu, X., Scheel, J., and Hedgecock, E. M., 2000. Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. Science 287, 989–994.

Inoue, J., Ishida, T., Tsukamoto, N., Kobayashi, N., Naito, A., Azuma, S., and Yamamoto, T., 2000. Tumor necrosis factor receptor-associated factor (TRAF) family: adapter proteins that mediate cytokine signaling. Exp. Cell Res. 254, 14–24.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Iwabe, N. and Miyata, T., 2002. Kinesin-related genes from diplomonad, sponge, amphioxus, and cyclostomes: divergence pattern of kinesin family and evolution of giardial membrane-bounded organella. Mol Biol Evol 19, 1524–1533.

Jones, D., Corbin, Custer, W., Andrew, and Begun, J., David, 2005. Origin and evolution of a chimeric fusion gene in Drosophila subobscura, D. madeirensis and D. guanche. Genetics 170, 207–219.

Kaessmann, H., Zollner, S., Nekrutenko, A., and Li, W. H., 2002. Signatures of domain shuffling in the human genome. Genome Res. 12, 1642–1650.

Kaestner, K. H., Knochel, W., and Martinez, D. E., 2000. Unified nomenclature for the winged helix/forkhead transcription factors. Genes Dev. 14, 142–146.

Kim, J. H., Park, K. C., Chung, S. S., Bang, O., and Chung, C. H., 2003. Deubiquitinating enzymes as cellular regulators. J Biochem (Tokyo) 134, 9–18.

Kortschak, R. D., Tamme, R., and Lardelli, M., 2001. Evolutionary analysis of vertebrate Notch genes. Dev Genes Evol 211, 350–354.

Koyanagi, M., Suga, H., Hoshiyama, D., Ono, K., Iwabe, N., Kuma, K., and Miyata, T., 1998. Ancient gene duplication and domain shuffling in the animal cyclic nucleotide phosphodiesterase family. FEBS Lett 436, 323–8.

Lawrence, C. J. et al., 2004. A standardized kinesin nomenclature. J Cell Biol 67, 19–22.

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P., and Bork, P., 2002. Recent improvements to the smart domain-based sequence annotation resource. Nucleic Acids Res 30, 242–244.

Li, W. H., Gu, Z., Wang, H., and Nekrutenko, A., 2001. Evolutionary analyses of the human genome. Nature 409, 847.

Lin, K., Zhu, L., and Zhang, D.-Y., 2006. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics 22, 2081–2086.

Locksley, R. M., Killeen, N., and Lenardo, M. J., 2001. The TNF and TNF receptor superfamilies: integrating mammalian biology. Cell 104, 487–501.

Long, M., 2001. Evolution of novel genes. Curr. Opin. Genet. Dev. 11, 673–680.

Long, M., Betran, E., Thornton, K., and Wang, W., 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4, 865–75.

Long, M. and Thornton, K., 2001. Gene duplication and evolution. Science 293, 1551.

Lowry, J. A. and Atchley, W. R., 2000. Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. J Mol Evol 50, 103–115.

MacEwan, D. J., 2002. TNF ligands and receptors—a matter of life and death. Br. J. Pharmacol. 135, 855–875.

Maine, E. M., Lissemore, J. L., and Starmer, W. T., 1995. A phylogenetic analysis of vertebrate and invertebrate Notch-related genes. Mol Phylogenet Evol 4, 139–149.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D., 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751–753.

Mazet, F., Yu, J.-K., Liberles, D. A., Holland, L. Z., and Shimeld, S. M., 2003. Phylogenetic relationships of the fox (Forkhead) gene family in the bilateria. Gene 316, 79–89.

Miki, H., Setou, M., and Hirokawa, N., 2003. Kinesin superfamily proteins (KIFs) in the mouse transcriptome. Genome Res 13, 1455–1465.

Miyata, T. and Suga, H., 2001. Divergence pattern of animal gene families and relationship with the Cambrian explosion. Bioessays 23, 1018–27.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al., 2005. InterPro, progress and status in 2005. Nucleic Acids Res 33, D201–5.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247, 536–40.

Nadeau, J. H. and Sankoff, D., 1998. Counting on comparative maps. Trends Genet 14, 495–501.

Needleman, S. and Wunsch, C., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443–53.

Nicholson, A. C., Malik, S.-B., Logsdon Jr., J. M., and Van Meir, E. G., 2005. Functional evolution of ADAMTS genes: evidence from analyses of phylogeny and gene organization. BMC Evol Biol 5, 11.

O'Brien, S. J., Wienberg, J., and Lyons, L. A., 1997. Comparative genomics: lessons from cats. Trends Genet 10, 393–399.

Orengo, C. A., Pearl, F. M. G., and Thornton, J. M., 2003. The CATH domain structure database. Methods Biochem Anal 44, 249–271. Historical Article.

Orengo, C. A. and Thornton, J. M., 2005. Protein families and their evolution—a structural perspective. Annu Rev Biochem 74, 867–900.

Patient, R. K. and McGhee, J. D., 2002. The GATA family (vertebrates and invertebrates). Curr Opin Genet Dev 12, 416–22.

Patthy, L., 1985. Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell 41, 657–663.

Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling–a review. Gene 238, 103–114.

Patthy, L., 2003. Modular assembly of genes and the evolution of new functions. Genetica 118, 217–231.

Pearl, F. M. G., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C. A., 2003. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res 31, 452–455.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raper, J. A., 2000. Semaphorins and their receptors in vertebrates and invertebrates. Curr Opin Neurobiol 10, 88–94.

Redfern, O., Grant, A., Maibaum, M., and Orengo, C., 2005. Survey of current protein family databases and their application in comparative, structural and functional genomics. J Chromatogr B Analyt Technol Biomed Life Sci 815, 97–107.

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

Robinson, D. R., Wu, Y. M., and Lin, S. F., 2000. The protein tyrosine kinase family of the human genome. Oncogene 19, 5548–5557.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523.

Sayah, D. M., Sokolskaja, E., Berthoux, L., and Luban, J., 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 430, 569–573.

Sheldahl, L. C., Slusarski, D. C., Pandur, P., Miller, J. R., Khl, M., and Moon, R. T., 2003. Dishevelled activates Ca2+ flux, PKC, and CamKII in vertebrate embryos. J Cell Biol 161, 769–77.

Shiu, S. H. and Li, W. H., 2004. Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. Mol Biol Evol 21, 828–840.

Simillion, C., Vandepoele, K., and Van de Peer, Y., 2004. Recent developments in computational approaches for uncovering genomic homology. Bioessays 26, 1225–35.

Stone, A. L., Kroeger, M., and Sang, Q. X., 1999. Structure-function analysis of the ADAM family of disintegrin-like and metalloproteinase-containing proteins (review). J Protein Chem 18, 447–465.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., and Jones, R. T., 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203, 439–455.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tordai, H., Nagy, A., Farkas, K., Banyai, L., and Patthy, L., 2005. Modules, multidomain proteins and organismic complexity. FEBS J 272, 5064–5078.

Van de Peer, Y., 2004. Computational approaches to unveiling ancient genome duplications. Nat Rev Genet 5, 752–763.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S.,

Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X., 2001. The sequence of the human genome. Science 291, 1304–1351.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H., 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103, 3220–3225.

Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M., 2000. The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in drosophila melanogaster. Mol Biol Evol 17, 1294–1301.

Welch, A. Y., Kasahara, M., and Spain, L. M., 2003. Identification of the mouse killer immunoglobulin-like receptor-like (Kirl) gene family mapping to chromosome X. Immunogenetics 54, 782–790.

Westin, J. and Lardelli, M., 1997. Three novel Notch genes in zebrafish: implications for vertebrate Notch gene evolution and function. Dev Genes Evol 207, 51–63.

Wharton, K. A., 2003. Runninẃith the Dvl: proteins that associate with Dsh/Dvl and their significance to Wnt signal transduction. Dev Biol 253, 1–17.

Wing, S. S., 2003. Deubiquitinating enzymes—the importance of driving in reverse along the ubiquitin-proteasome pathway. Int J Biochem Cell Biol 35, 590–605.

Wolfsberg, T. G. and White, J. M., 1996. ADAMs in fertilization and development. Dev Biol 180, 389–401.

Wu, H., Cathy, Huang, H., Yeh, L., Lai-Su, and Barker, C., Winona, 2003. Protein family classification and functional annotation. Comput Biol Chem 27, 37–47.

Wuchty, S., 2001. Scale free behavior in protein domain networks. Mol. Biol. Evol 18, 1694–1702. Powerlaw, domain network.

Wuchty, S. and Almaas, E., 2005. Evolutionary cores of domain co-occurrence networks. BMC Evol Biol 5, 24.

Yazdani, U. and Terman, J. R., 2006. The semaphorins. Genome Biol 7, 211.

Ye, Y. and Godzik, A., 2004. Comparative analysis of protein domain organization. Genome Res 14, 343–353.

Yeats, C., Maibaum, M., Marsden, R., Dibley, M., Lee, D., Addou, S., and Orengo, A., Christine, 2006. Gene3D: modelling protein structure, function and evolution. Nucleic Acids Res 34, 281–284.

Zhang, Z. and Gerstein, M., 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol 2.

List of Figures

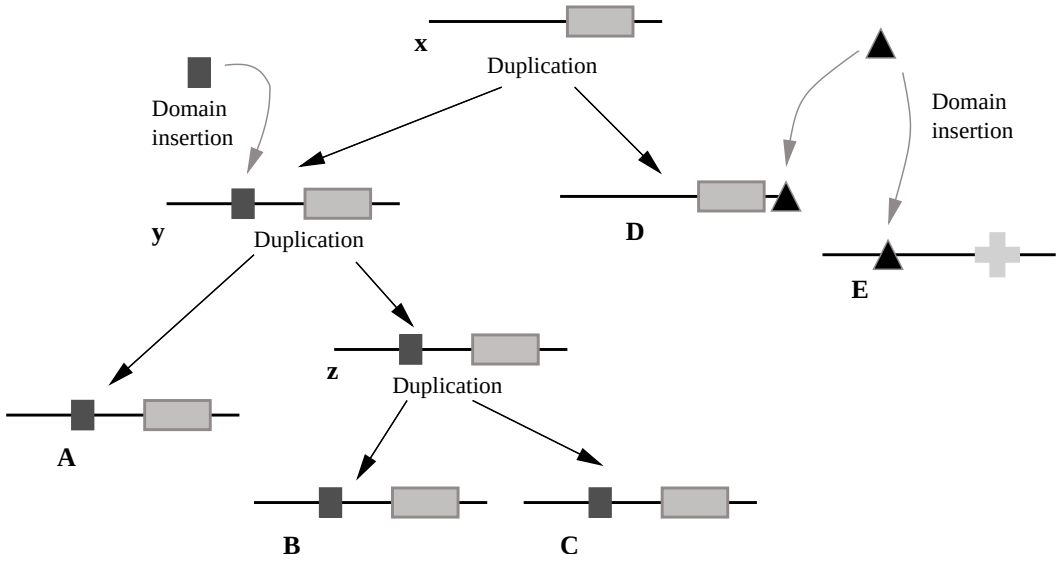

Figure 1: Evolutionary tree of a hypothetical multi-domain protein family evolving through duplication and domain insertion.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 1 (of 6)

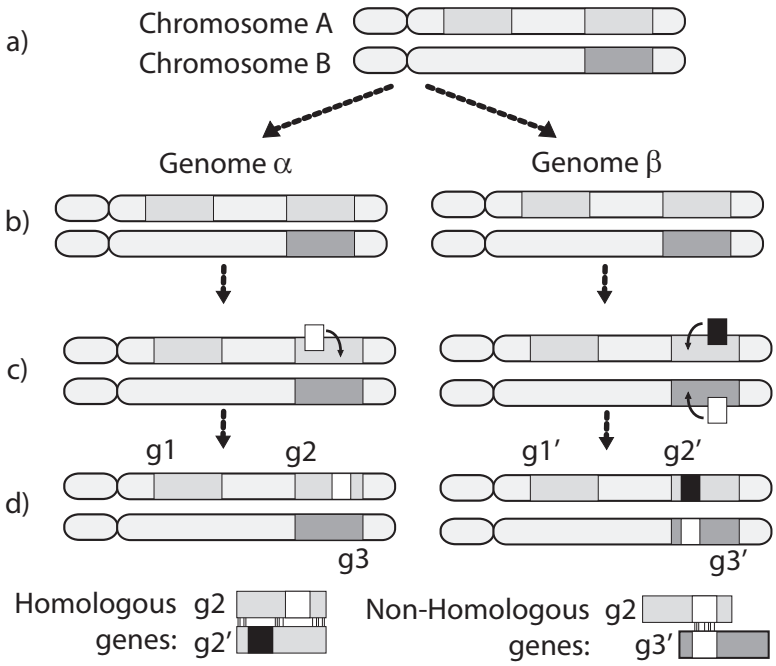

Figure 2: Genome evolution with domain insertions. (a) A hypothetical genome (b) Replication by speciation or whole genome duplication results in two offspring genomes. (c) Domain insertions. (d) g2 and g2' are descended from the same ancestral region. g2 and g3' share a homologous domain but are descended from different regions in the ancestral genome.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 2 (of 6)

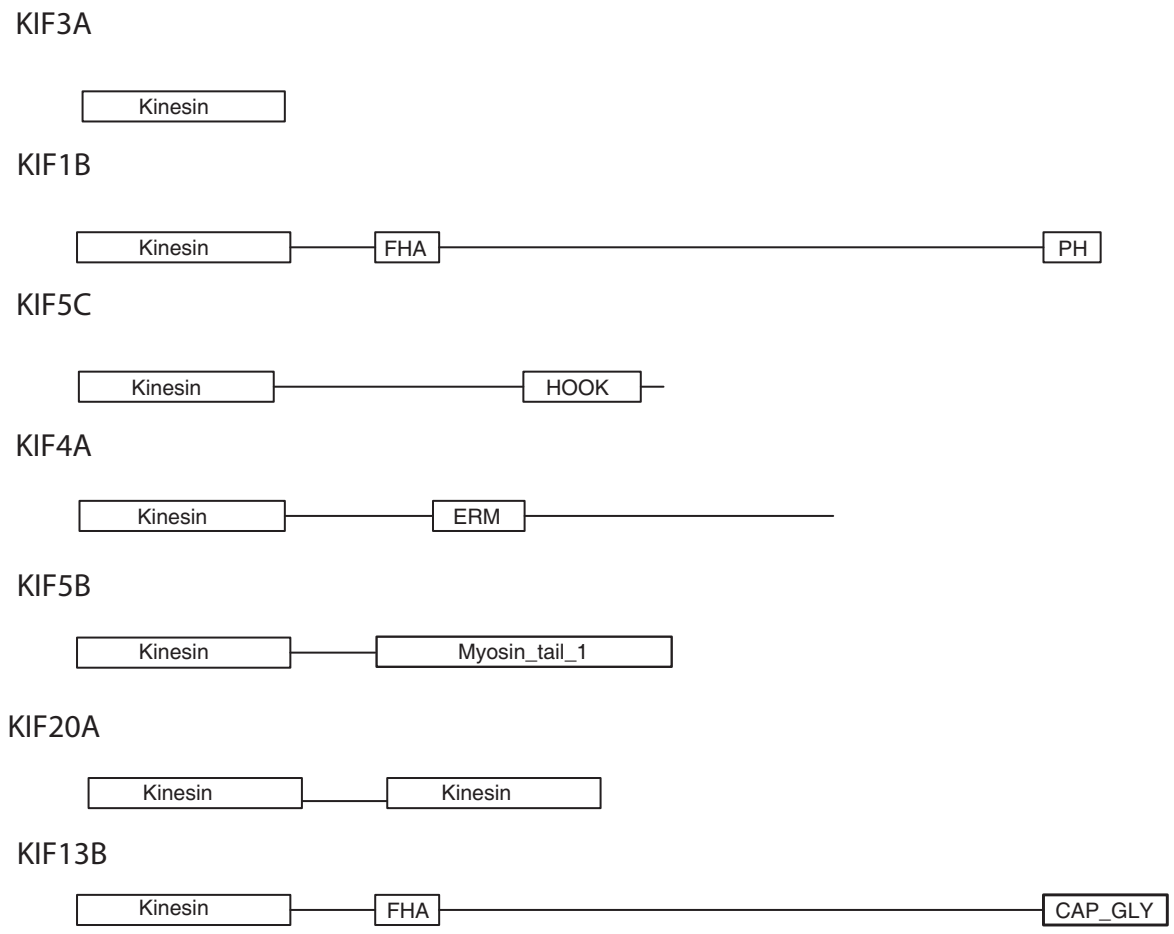

Figure 3: Domain architectures for representative members of the Kinesin family.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 3 (of 6)

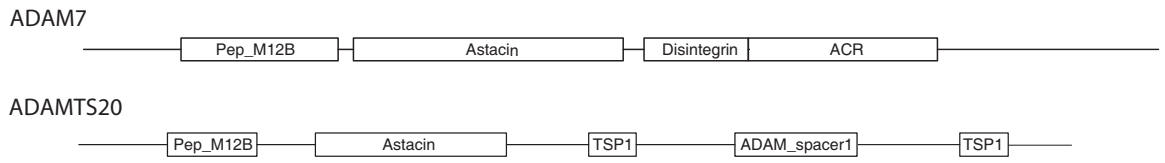

Figure 4: One representative domain architecture from each of the two ADAM subfamilies, ADAM and ADAMTS.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 4 (of 6)

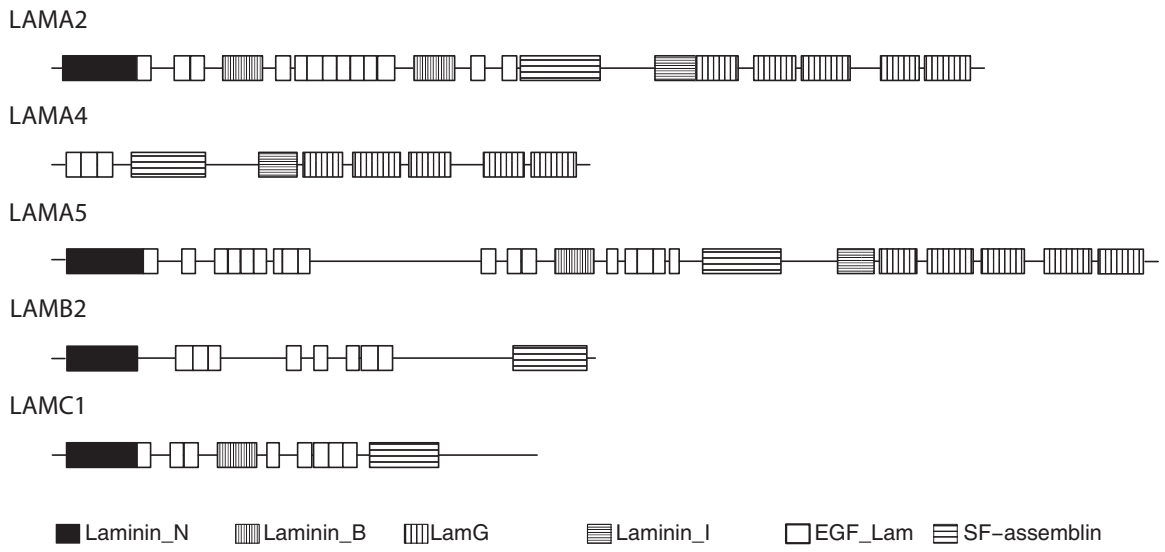

Figure 5: Representative architectures from the three Laminin subfamilies, α, β and γ.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 5 (of 6)

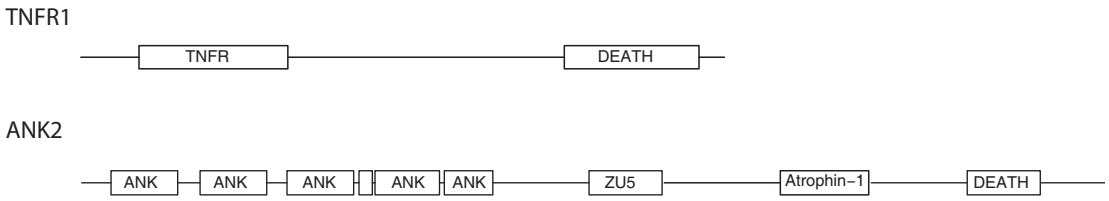

Figure 6: Domain architectures for TNFR1, a member of the TNFR family and ANK2, a protein that is not a member of the TNFR family. Both have a DEATH domain.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Figure 6 (of 6)

List of Tables

ADAM (A Disintegrin And Metalloproteinase containing family):

Nicholson et al. (2005); Stone et al. (1999); Wolfsberg and White (1996)

DVL (Dishevelled protein family):

Wharton (2003); Sheldahl et al. (2003)

FOX (Forkhead transcription factor family):

Mazet et al. (2003); Kaestner et al. (2000)

GATA (GATA binding proteins):

Lowry and Atchley (2000); Patient and McGhee (2002)

Kinase:

Robinson et al. (2000); Hanks (2003); Cheek et al. (2002); Shiu and Li (2004)

Kinesin:

Iwabe and Miyata (2002); Lawrence et al. (2004); Miki et al. (2003)

KIR (Killer cell Immunoglobulin-like Receptors):

Welch et al. (2003); Belkin et al. (2003)

Laminin:

Engel (1992); Hutter et al. (2000)

Myosin:

Richards and Cavalier-Smith (2005); Goodson and Dawson (2006); Foth et al. (2006)

Notch:

Maine et al. (1995); Westin and Lardelli (1997); Kortschak et al. (2001)

PDE (Phosphodiesterases):

Degerman et al. (1997)

SEMA (Semaphorin):

Raper (2000); Yazdani and Terman (2006)

TNFR (Tumor Necrosis Factor Receptors):

Locksley et al. (2001); MacEwan (2002)

TRAF (Tumor necrosis factor Receptor Associated Factors):

Inoue et al. (2000)

USP (Ubiquitin Specific Proteases):

Wing (2003); Kim et al. (2003)

Table 1: Protein families and selected sources used for curation

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 1 (of 7)

Functional category	Family
Biological Process	
Cell-cell/cell-matrix interaction	ADAM, Laminin, Notch
Cell motility and polarity	Myosin
Cellular transport	Kinesin, Myosin
Development and homeostatic regulation	DVL, FOX, ADAM, GATA
Immune response	ADAM, TNFR, KIR
Neural development	FOX, SEMA, Notch
Tissue morphogenesis	FOX, Laminin, SEMA
Molecular Function	
Transcription factor	FOX, GATA
Intracellular signal transducer	Kinase, DVL, TRAF
Enzyme	ADAM, Kinase, USP, PDE
Motor	Myosin, Kinesin
Structural molecule	Laminin
Ligand	SEMA
Receptor	TNFR, KIR, Kinase, Notch
Cellular location	
Extracellular	ADAM, Laminin, SEMA
Transmembrane	ADAM, KIR, Kinase, Notch, PDE, SEMA, TNFR
Intracellular	DVL, FOX, GATA, Kinase, Kinesin, Myosin, PDE, TRAF, USP

Table 2: Functional categories of families in the test dataset.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 2 (of 7)

Similarity			
	weight	**domain-count**	**comparison**
S1	unweighted	uncorrected	domain copies
S2	"	Jaccard	"
S3	"	cosine	"
S4	"	uncorrected	domain types
S5	"	Jaccard	"
S6	"	cosine	"
S7	distinct partner	uncorrected	domain copies
S8	"	cosine	"
S9	idf	uncorrected	"
S10	"	cosine	"
S11	tf-idf	uncorrected	"
S12	"	cosine	"
S13	unweighted	Lin	Lin

Distance			
	weight	**domain-count**	**comparison**
D1	unweighted	uncorrected	unordered distance
D2	"	Eq. 8	"
D3	distinct partner	uncorrected	"
D4	"	Eq. 8	"
D5	unweighted	uncorrected	ordered distance
D6	"	Eq. 8	"
D7	distinct partner	uncorrected	"
D8	"	Eq. 8	"

Table 3: Summary of scoring systems evaluated.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 3 (of 7)

Similarity							
	domain copies			domain types			
	(S1)	Jaccard (S2)	cosine (S3)	(S4)	Jaccard (S5)	cosine (S6)	Lin (S13)
ADAM	0.94	0.96	0.96	0.98	0.99	0.99	0.98
DVL	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FOX	0.50	1.00	1.00	0.50	1.00	1.00	1.00
GATA	1.00	0.99	0.99	0.63	0.87	0.87	0.94
Kinase	0.44	0.69	0.71	0.50	0.67	0.69	0.68
Kinesin	0.52	0.89	0.90	0.52	0.89	0.90	0.90
KIR	0.50	0.86	0.86	0.50	0.77	0.77	0.86
Laminin	0.99	0.99	0.97	0.91	0.90	0.86	0.96
Myosin	0.64	0.56	0.56	0.64	0.54	0.53	0.56
Notch	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PDE	0.54	0.82	0.84	0.54	0.89	0.89	0.84
SEMA	0.78	0.87	0.89	0.79	0.93	0.93	0.93
TNFR	0.57	0.96	0.96	0.56	0.97	0.98	0.97
TRAF	0.85	0.97	0.98	0.85	0.97	0.98	0.98
USP	0.55	0.86	0.84	0.53	0.87	0.86	0.88
Mean	0.72	0.89	0.90	0.70	0.88	0.88	0.90
St Dev	0.21	0.12	0.12	0.19	0.13	0.13	0.12
All	0.46	0.71	0.73	0.51	0.70	0.71	0.70
All(no kinase)	0.63	0.89	0.89	0.65	0.88	0.87	0.90

Table 4: AUC scores for homology identification using unweighted similarity. Bottom two rows are for the entire dataset combined and for the combined set, excluding the Kinases.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 4 (of 7)

Similarity								
	no weighting		distinct partner		idf		tf-idf	
	(S1)	cosine (S3)	(S7)	cosine (S8)	(S9)	cosine (S10)	(S11)	cosine (S12)
ADAM	0.94	0.96	0.98	0.98	0.97	0.98	0.84	0.88
DVL	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00
FOX	0.50	1.00	1.00	1.00	0.00	0.99	0.99	0.99
GATA	1.00	0.99	1.00	0.94	0.63	0.87	0.83	0.87
Kinase	0.44	0.71	0.16	0.49	0.09	0.53	0.47	0.46
Kinesin	0.52	0.90	0.99	0.96	0.85	0.93	0.94	0.93
KIR	0.50	0.86	0.50	0.86	0.50	0.77	0.77	0.77
Laminin	0.99	0.97	0.99	0.95	0.97	0.90	0.96	0.97
Myosin	0.64	0.56	0.94	0.65	0.98	0.74	0.76	0.70
Notch	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
PDE	0.54	0.84	0.84	0.87	0.84	0.82	0.82	0.62
SEMA	0.78	0.89	0.99	1.00	0.99	0.99	0.97	0.87
TNFR	0.57	0.96	0.83	0.95	0.30	0.94	0.96	0.95
TRAF	0.85	0.98	0.99	0.99	0.99	1.00	0.99	0.99
USP	0.55	0.84	0.82	0.77	0.57	0.82	0.87	0.85
Mean	0.72	0.90	0.87	0.89	0.71	0.89	0.88	0.86
St Dev	0.21	0.12	0.23	0.14	0.34	0.13	0.13	0.15
All	0.46	0.73	0.19	0.55	0.13	0.60	0.51	0.52
All (no Kinase)	0.63	0.89	0.97	0.92	0.91	0.91	0.91	0.88

Table 5: AUC scores for homology identification using weighted similarity. AUC scores for unweighted similarity are given for comparison. The last two rows give AUC scores for the entire dataset combined and for the combined set, excluding the Kinases.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 5 (of 7)

	Similarity		Distance			
	no weighting		no weighting		distinct partner	
	(S1)	cosine (S3)	(D1,D5)	Eq. 8 (D2,D6)	(D3,D7)	Eq. 8 (D4,D8)
ADAM	0.94	0.96	0.84	0.96	0.92	0.99
DVL	1.00	1.00	1.00	1.00	0.97	0.99
FOX	0.50	1.00	1.00	1.00	1.00	1.00
GATA	1.00	0.99	0.95	0.99	0.89	0.94
Kinase	0.44	0.71	0.72	0.69	0.62	0.47
Kinesin	0.52	0.90	0.89	0.89	0.88	0.96
KIR	0.50	0.86	0.86	0.86	0.86	0.86
Laminin	0.99	0.97	0.75	0.99	0.78	0.99
Myosin	0.64	0.56	0.53	0.56	0.41	0.69
Notch	1.00	1.00	0.99	1.00	1.00	1.00
PDE	0.54	0.84	0.82	0.82	0.62	0.85
SEMA	0.78	0.89	0.80	0.87	0.96	1.00
TNFR	0.57	0.96	0.95	0.96	0.91	0.95
TRAF	0.85	0.98	0.95	0.97	0.88	0.99
USP	0.55	0.84	0.85	0.86	0.75	0.80
Mean	0.72	0.90	0.86	0.89	0.83	0.90
St Dev	0.21	0.12	0.12	0.12	0.16	0.14
All	0.46	0.73	0.73	0.71	0.67	0.53
All(no Kinase)	0.63	0.89	0.84	0.89	0.81	0.94

Table 6: AUC scores for homology identification using distance scoring. AUC scores for unweighted similarity are shown for comparison. Domain-count corrections were calculated using the length correction formula given in Eq. 8. The last two rows give AUC scores for the entire dataset combined and for the combined set, excluding the Kinases. Note that the unweighted similarity score with Jaccard correction column is identical to the unweighted distance score with Jaccard correction for the reasons described in the text.

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 6 (of 7)

	No sequence similarity			**Sequence similarity**		
Domain similarity	TPF	TNF	TF	TPF	TNF	TF
none				0.57	0.97	0.88
unweighted (S1)	0.31	0.98	0.83	0.68	0.97	0.90
unweighted, Jaccard (S2)	0.67	0.95	0.89	0.79	0.97	0.93
distinct partner (S7)	0.58	0.99	0.90	0.79	0.98	0.94
distinct partner, cosine (S8)	0.74	0.92	0.85	0.84	0.98	0.95

Table 7: Logistic regression results for domain architecture comparison without sequence similarity scores (first column) and with (second column). The test set used is the collected set of all families except for the Kinase family

⇑

N. Song,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: nsong@andrew.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, R.D. Sedgewick,

Department of Biological Sciences,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: bobsedge@cs.cmu.edu, Phone: (412) 268-3199, Fax: (412)

268-7129, D. Durand,

Departments of Biological Sciences and Computer Science,

Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Email: durand@cmu.edu, Phone: (412) 268-6036, Fax: (412) 268-7129

Table 7 (of 7)