Sequence similarity network reveals common ancestry of multidomain proteins

Nan Song∗ George B. Davis † Dannie Durand‡
Carnegie Mellon University, 4400 Fifth Avenue Pittsburgh, PA 15213
February 8, 2007

Abstract

Protein modularity allows the evolution of diverse function through combinatorial rearrangement of functional building blocks. This versatile evolutionary mechanism appears to have played a transformative role in key evolutionary transitions, including the emergence of multicellular animals and the vertebrate immune system. Identification of multidomain homologs is essential to studying the evolution of such families, as well as to many genomic applications that exploit evolutionary information. Traditional homology identification methods based on sequence similarity fall short when applied to multidomain families, which can have significant sequence similarity due to a shared domain despite having distinct evolutionary histories. Elimination of the resulting misassignments requires a method for distinguishing genes related through duplication and speciation from genes related only through domain insertion. We developed Neighborhood Correlation, a novel homology identification method based on the observation that gene duplication and domain insertion result in different topological structures in the sequence similarity network. In an empirical comparison with traditional approaches on manually curated families in mouse and human, we demonstrate that Neighborhood Correlation is more accurate and reliable. Neighborhood Correlation achieves high sensitivity and specificity in both single domain and complex multidomain families. In contrast, traditional methods that combine sequence similarity with additional criteria based on alignment length performed poorly, challenging the widely held view that homologous sequences have longer alignments than pairs that share an inserted domain. We predicted mouse and human homologs using Neighborhood Correlation. Those predictions, as well as our manually curated data set and a visualization tool for exploratory analysis of the network neighborhood structure are available at http://goby.compbio.cs.cmu.edu/NC. Protein family classification has been limited by the view that it is only possible to cluster individual domains or families with conserved domain architectures. By exploiting the structure of the sequence similarity network, our approach surmounts this limitation and opens the door to studies of families with complex domain architectures.

∗Biological Sciences
†Computer Science
‡Biological Sciences. Contact author: tel: 412-268-6036, fax: 412-268-7129, email: durand@cmu.edu
Introduction

Accurate identification of homologs, sequences that share common ancestry, is essential for accuracy in function prediction and comparative genomics. Homology identification is integral to the annotation of novel genes [1] and prediction of gene function by various methods, including phylogenetic clustering [2–4], gene fusion analysis [5, 6], phylogenomic inference [7–9] and genomic context [10, 11]. Homologous genes are used as markers to identify homologous chromosomal regions for comparative mapping [12, 13], analysis of whole genome duplication [14–16], phylogenetic footprinting [17–21] and operon prediction [22–25]. Pairwise homology detection is also an integral component of clustering approaches to protein family classification [1, 26, 27, and work cited therein].

In the era of high-throughput sequencing, automated methods are required that can be applied to large data sets with no human intervention. Many excellent methods have been proposed to address various aspects of homology detection: identification of very distantly related sequences [28–34], separation of paralogs from orthologs [35–42], and identification of homologs that share the same genomic context [43–50]. However, the problem of identifying multidomain homologs remains unsolved. Multidomain proteins represent 40% of the metazoan proteome and play important roles in signal transduction, cellular adhesion, tissue repair, and immune response [51]. Two aspects of multidomain proteins present particular challenges for homology identification: promiscuous domains that appear in many unrelated families and large families with varied domain architectures. For multidomain sequences, spurious similarity can arise from domain-based matches: two otherwise unrelated sequences with significant similarity arose through insertion of the same domain into both sequences. Box 1 shows a concrete example of this problem, illustrating the challenge of distinguishing multidomain homologs from unrelated pairs that share a domain.

Homology detection can be defined as a classification problem. The goal is to separate sequence pairs into two classes: homologs and unrelated pairs. Given a set of sequence pairs, scores are assigned to pairs in such a way that the score distributions for the two classes are disjoint, or nearly so. A threshold is then applied to the scores to predict homologs. Current approaches assign such scores based on properties of pairwise sequence comparisons. Typically, pairs are scored based on a measure of sequence similarity (e.g., E value or percent identity). In order to eliminate domain-based matches, pairs are also scored using alignment coverage, the length of the optimal local alignment, expressed as a fraction of the entire sequence length [52]. Pairs that meet both a sequence similarity threshold and an alignment coverage threshold are considered homologs. The assumption underlying these heuristics is that pairwise alignments of homologous sequences have distinctive properties that differ from those of domain-based matches and can therefore be used to eliminate false positives. Yet Box 1 shows an example where the alignments of a pair of homologous sequences and a domain-based match have similar properties. Both have weak sequence similarity, short alignments, and a combination of shared and unique domains, calling the above assumption into question. In fact, despite the prevalence of methods based on sequence similarity and alignment coverage [35, 44, 48, 52–59], to our knowledge their accuracy in classifying multidomain homologs has never been investigated.

There are two requirements for advances in multidomain homology detection: a formal definition of multidomain homology and curated benchmarks for validating multidomain homology detection methods. According to the traditional definition, homologs are sequences that descended from a common ancestor by vertical descent; that is, via speciation and gene duplication [60, 61]. This definition is not applicable to multidomain families that evolve via domain shuffling; that is, via insertion, deletion and rearrangement of domains. In this work, we propose an extended definition of homology that takes both
vertical descent and domain insertion into account. Our model provides a formal basis for curating a benchmark of homologous multidomain families. Although comprehensive data sets are available for testing methods for predicting homology of individual domains [62–64], we are unaware of a gold standard data set of known multidomain families with variable domain architectures. To meet this need, we hand-curated a data set of twenty well-studied families for which there is clear evidence of common ancestry and used it to evaluate the performance of multidomain homology detection.

Our empirical study using this data set shows that classification using alignment coverage results in both high false positive and high false negative rates for more than half the families studied. Our results challenge two unstated, but widely accepted hypotheses: (1) homologous sequences share similarity along the bulk of their length and (2) the local alignment between homologous sequences is usually a greater fraction of total length than those sharing only a domain. Based on the results of our study, we argue that pairwise sequence comparison alone does not contain enough information to discriminate between common ancestry and domain sharing. Additional information is required for reliable identification of multidomain homologs.

We propose a novel method, called Neighborhood Correlation, that exploits the structure of the weighted sequence similarity network to distinguish homologs from domain-based matches. The sequence similarity network is a network in which each sequence corresponds to a vertex and two vertices are connected by an edge if the associated sequences have significant similarity. Edge weights are proportional to the degree of similarity. The neighbors of a sequence in this network is the set of vertices adjacent to it; that is, the set of all sequences that match it above a predefined significance threshold. We predict that the neighborhood structure of genes that arose through domain shuffling will be characteristically different from that of genes that arose through duplication or speciation (illustrated in Box 1). Moreover, these differences in neighborhood organization will be detectable and can be exploited to distinguish homology from domain sharing. In an all-against-all comparison of coding sequences from one or more genomes, we assign to each pair of sequences a score that reflects the similarity of their neighborhoods. As we demonstrate empirically, the Neighborhood Correlation scores of homologs within a given gene family are consistently higher than the scores of unrelated pairs that share a domain, providing a reliable, quantitative basis for identifying homologs. Neighborhood Correlation is highly effective and dramatically outperforms traditional sequence comparison approaches. It is also easy to compute and hence suitable for high throughput genomic studies.

In the following sections, we first describe our multidomain homology model and review the current evidence from the molecular evolution literature on which it is based. In the following section, we define Neighborhood Correlation quantitatively. We discuss the competing influences of gene duplication and domain insertion on the structure of the sequence similarity network and describe how Neighborhood Correlation captures these structures. We next describe the curation of a novel benchmark data set for validating methods for separating homology due to gene duplication from matches due to domain shuffling. Our test set includes single domain families, as well as multidomain families with promiscuous domains that are at risk for domain matching. We offer this validation data set, which is based on published evidence by experts on each of the families, as a resource for future studies. Using this data set, we evaluate the performance of Neighborhood Correlation empirically and compare it to sequence similarity alone, and with an alignment coverage criterion. To illuminate specific advantages of Neighborhood Correlation, we discuss its performance on three complex families in depth. In addition to validating Neighborhood Correlation, we discuss its use in predicting homologs on novel gene sets. Predictions of mouse and human homologs using this method are provided in supplementary data. We conclude the
Results section by revisiting alignment coverage and consider why a method so widely accepted should yield such poor performance. In Discussion, we relate the processes of multidomain evolution to the structure of the sequence similarity network and consider the effectiveness of Neighborhood Correlation in light of known evolutionary processes.

Results

A model of multidomain homology

Modularity is a powerful mechanism for achieving functional variation and interaction specificity within a gene family that performs a core molecular function. Modular proteins arise through domain shuffling, mediated by mechanisms such as segmental duplication, non-allelic-homologous recombination, retrotransposition, and read-through [65–70, and work cited the rein]. Modularity also arises on a smaller scale, through acquisition of sequence motifs that provide functional specificity but do not encode structural domains. Examples include sequence motifs that confer substrate specificity in enzymes [71] and transactivation and transrepression motifs in transcription factors [72].

Evolution of modularity is characterized by extensive reuse of building blocks. Domains that determine the core molecular function of a given family are often specific to that family, while domains associated with interaction specificity frequently appear in many different families. Domains that occur in many different domain architectures are called promiscuous domains. The number of co-occurring domains is often used as a measure of promiscuity [51,73]. The Ank, Fn3 and Ig domains in the example in Box 1 are promiscuous domains.

Only a small fraction of domains are promiscuous, suggesting that these particular domains have properties that facilitate successful insertion into new genetic contexts. Two requirements of promiscuity are mobility and a tendency to be selectively favorable, once inserted into a new context. Size and phase are two indicators of domain mobility. Tordai et al. [51] observed a correlation between domain promiscuity and domain length and noted that typically promiscuous domains are at most 80 - 100 residues long. They concluded that small size may facilitate mobility. In addition, insertions that preserve the translational reading frame have a higher probability of survival [74]. These are called symmetric phase insertions because the flanking introns stop and start, respectively, in the same phase of the reading frame. Symmetric phase domains occur more frequently than expected by chance [69]. Moreover, promiscuous domains tend to have symmetric phase and, especially, 1-1 phase [69, 75, 76].

The number, size and complexity of multidomain families expanded dramatically at the metazoan boundary, with additional expansions at the base of the chordate and vertebrate lineages [58,66,77–83]. The correlation between increased domain shuffling and key taxonomic transitions suggests a link between the acquisition of novel architectures and the emergence of novel cellular processes, as well as physiological and morphological innovations [66,84,85]. For example, multidomain architectures promote cellular processes that require integration of several molecular functions, such as coupling of a recognition module and an adapter module to form a receptor. Such families enabled the evolution of signal transduction, extracellular matrix functions and cellular adhesion, prerequisites for multicellularity. Other multidomain families emerged during chordate and early vertebrate evolution. These include families associated with immune response, tissue formation and repair, blood-related functions and the vertebrate nervous system [66,77,79,80].

Domain shuffling has lead to large, complex multidomain families with variable domain architectures. The modularity of these families offers functional plasticity that is essential to the evolution of
metazoans and, in particular, vertebrates. That same modularity is an obstacle to abstract models and computational analysis. Multidomain homology is a case in point: homology has traditionally been defined in terms of families that evolve by vertical descent [60,61]; that is, by speciation and gene duplication. However, multidomain sequences evolve by vertical descent and horizontal transfer; that is, by speciation, gene duplication and insertion of domains from outside the family [86].

The traditional definition of homology does not apply in this case as previous authors have pointed out [61,87]. In the words of Walter Fitch [61], “We must recognize that not all parts of a gene have the same history and thus, in such cases, that the gene is not the unit to which the terms orthology, paralogy, etcetera, apply.” It has been proposed that smaller sequence fragments should be the units of interest and that the term partial homology be used when referring to sequences that have been the targets of insertions [87,88].

However, there are many applications, such as ortholog detection, comparative mapping, and phylogenetic footprinting, for which it is convenient and desirable to work with a definition of homology where the gene is the basic unit. Moreover, in order to study the evolution of multidomain gene families, it is necessary to focus on genes. A model is needed that applies to multidomain sequences, when considered as entire genes.

Here, we propose a model of multidomain homology based on vertical descent and one particular type of horizontal transfer, namely, insertion of a sequence fragment into an existing gene. In this case, it is possible to define common ancestry by vertical descent, even though horizontal transfers have occurred. In our model, two sequences are homologous if they share an ancestral locus. The rationale for this definition is illustrated in Fig. 2, which shows the evolution of genes through vertical descent and horizontal transfer of domains in the context of the chromosomes in which they reside. When genomic context is taken into account, it is clear that genes \( g_1 \) and \( g'_1 \) are homologous, despite the fact that \( g_1 \) contains a domain not present in \( g'_1 \) and vice versa. In contrast, genes \( g_1 \) and \( g_2 \) are not homologous, despite the fact that they share a homologous domain, since \( g_1 \) and \( g_2 \) are not located in chromosomal regions that share common ancestry. For comparative mapping applications, where homologous genes are used as markers for identifying chromosomal regions, this distinction is crucial. For example, phylogenetic footprinting [17,19–21] is a method for predicting transcription factor binding sites by identifying homologous genes and then searching their flanking chromosomal regions for conserved sequence motifs. In Fig. 2, the regions upstream of \( g_1 \) and \( g'_1 \) have a higher probability of sharing conserved motifs since they share common ancestry. However, there is no reason to expect an enrichment of shared motifs between the flanking regions of \( g_1 \) and \( g_2 \).

This situation can also be represented as a tree. Fig. 1 shows the evolution of a hypothetical gene family by duplication, and domain insertion, as well as further duplications not shown in Fig. 2. Sequences \( A \) and \( B \) have one unique domain each and one shared domain. Although on the sequence level they are not homologous along their entire length, they share a common ancestor by vertical descent and are homologous genes according to our definition. Sequences \( B \) and \( C \) also have one unique domain each and one shared domain. However, since they are not related by vertical descent, they are not homologs according to our definition, even though they share a homologous sequence fragment.

Our model is applicable to families that evolved through gene duplication followed by insertion of sequence fragments into existing genes. Our model does not capture the case where a new domain architecture is assembled de novo from several unrelated building blocks and subsequently acquires a

\*In our model, horizontal transfers originate outside the family. It is immaterial whether they originate in the same genome or different genome.
regulatory region. In this case, we consider the novel architecture to be the progenitor of a new family, since it is not clear that the ancestry of any one constituent is preferred.

We believe that this abstraction will prove to be a useful model for a large fraction of multidomain families. Several lines of evidence support this model. First, a substantial number of metazoan, chordate and vertebrate families have been identified that evolved through a pattern of duplication, insertion of domains and further duplication, a pattern consistent with this model [84, 85, 89]. A second line of evidence is the existence of promiscuous domains that lend themselves to insertion in new chromosomal environments. Third, domain insertion is more likely to be successful when a domain is inserted into an existing functional environment, e.g., into the intron of an existing gene. In this case, all of the regulatory and termination signals required for successful transcription are already present. Finally, evidence of the particular domain shuffling mechanisms that occurred is sometimes still visible in the flanking DNA of genes that arose very recently. A number of studies describe recently evolved metazoan genes that arose through duplication of an existing gene followed by insertion of one or more domains by unequal crossing over or by retrotransposition [70, 90–92]. Since the recipient gene is duplicated, one copy continues to perform the original function, leaving the newly generated, mosaic gene free to evolve new function.

Neighborhood Correlation

Our model makes it possible to define the problem more precisely. Our goal is to develop quick, simple and reliable methods for identifying pairs that are related by vertical descent, even if they have been the targets of horizontal insertions. For example, in Fig. 1, our classifier should accept sequence pair A and B, but reject sequences pair B and C. Both pairs have one unique domain and one shared domain. The challenge is to find a property that will allow us to distinguish between these cases, despite their apparent similarity. Current methods are not well suited to this task. Sequence similarity identifies candidate homologs but additional information is needed to eliminate domain-based matches. Probabilistic models of multiple sequence alignment, while enormously helpful in identifying distant homologs [34], do not help here because only the shared domains can be aligned. For similar reasons, phylogenetic methods are not appropriate for this problem either. The example in Box 1 suggests that alignment length cannot distinguish between homologs and domain-based matches either. Our empirical results, as shown below, confirm this. A completely new approach is required.

The structure of the sequence similarity network provides a basis for distinguishing pairs that share common ancestry by vertical descent from pairs that are related through horizontal insertions only. The history of gene duplications and domain insertions leaves its trace in the network. For a given protein pair, the sequences that match both members of the pair are shared neighbors, while the sequences that match only one member are unique neighbors. Gene duplication tends to increase the shared neighborhood while domain insertion increases unique neighborhoods. Thus, homologous pairs are likely to have a greater percentage of shared neighbors than domain-based matches. We express this intuition quantitatively by defining the Neighborhood Correlation score of two sequences to be the correlation coefficient of their respective neighborhoods:

\[
NC(x, y) = \frac{\sum_{i=1}^{N}(\varsigma(x, i) - \bar{\varsigma}(x))(\varsigma(y, i) - \bar{\varsigma}(y))}{N^2 \sqrt{\text{var}(x) \cdot \text{var}(y)}},
\]

where \(\varsigma(x, i)\) is a similarity score for query sequence \(x\) and data base sequence \(i\), and \(N\) is the number of sequences in the data base. \(\bar{\varsigma}(x)\) and \(\text{var}(x)\) are the mean and variance of \(\varsigma(x, i)\) over all sequences.
in the data base. When there is a significant match between \( x \) and \( i \), we define the similarity score to be \( \varsigma(x, i) = \log_{10} S(x, i) \), where \( S(x, i) \) is the normalized bit score [34] of the alignment of \( x \) and \( i \). Sequence pairs with no significant match were assigned a score of \( \log_{10} S_{\text{min}} \), where \( S_{\text{min}} \) is the bit score corresponding to the significance threshold (see Methods).

Neighborhood Correlation is similar in flavor to the overlap algorithm [93], developed independently by Medini et al., which calculates pairwise scores based on the relative size of the shared and unique neighborhoods in the sequence similarity network. These pairwise scores were then used to rewire the network. Clustering was applied to the resulting network to predict protein families in 120 bacterial genomes. Medini et al. used this approach to identify core sets of proteins associated with Type III and IV secretion systems.

Unlike the overlap algorithm, the Neighborhood Correlation score depends not only on the relative size of the shared and unique neighborhoods, but also on the weights of the edges. The inclusion of edge weights (i.e., sequence similarity) in the calculation of Neighborhood Correlation, is crucial to the effectiveness of our method. \( \text{NC}(x, y) \) increases with the size of the shared neighborhood and decreases with the size of the unique neighborhood. \( \text{NC}(x, y) \) also increases with the score \( S(x, i) \) (respectively, \( S(y, i) \)), if matching sequence \( i \) is in the shared neighborhood and decreases if \( i \) is in the unique neighborhood. Most important, since the Neighborhood Correlation score measures the correlation between scores, it is greatest when \( S(x, i) \) and \( S(y, i) \) are roughly of the same magnitude. These qualitative trends are derived quantitatively in Methods.

Consider the example in Box 1. The shared neighborhood of \( \text{Pdgfrb} \) and \( \text{Ankrd3} \) is large. Since \( \text{Ig} \) and \( \text{Ank} \) are promiscuous domains, the unique neighborhoods of \( \text{Pdgfrb} \) and \( \text{Ankrd3} \) are also quite large. However, like many promiscuous domains, the \( \text{Ig} \) and \( \text{Ank} \) superfamilies are quite diverged so the scores in the unique neighborhoods are weak, so that the Neighborhood Correlation score is not greatly decreased by the large unique neighborhoods. In our data set, the Neighborhood Correlation score of \( \text{Pdgfrb} \) and \( \text{Ankrd3} \) is 0.67. The unrelated pair, \( \text{Pdgfrb} \) and \( \text{Ncam1} \), have a fairly large shared neighborhood. However, the edge weights in this neighborhood do not correlate. Neighbors that match \( \text{Pdgfrb} \) with a strong score typically match \( \text{Ncam1} \) with a weak score and vice versa. The sizeable shared neighborhood does increase the Neighborhood Correlation score much. The unique neighborhood of \( \text{Ncam1} \) is of the same size as the shared neighborhood. The unique neighborhood of \( \text{Pdgfrb} \) is much larger, due to the size of the kinase family, and has higher scores, decreasing the Neighborhood Correlation score. \( \text{Pdgfrb} \) and \( \text{Ncam1} \) have a Neighborhood Correlation score of 0.25, which is much lower than 0.67. Neighborhood Correlation can distinguish between the pairs in Box 1, while sequence similarity, alignment coverage and domain count cannot. As we will see in the following sections, this is frequently the case.

A benchmark data set for multidomain homology

The lack of gold-standard validation data to test the effectiveness of new methods is a major obstacle to research on multidomain homology and classification. Structural data bases, such as SCOP [62] and CATH [63, 64], are used to test detection of individual homologous domains, but do not capture the ancestry of multidomain families. Functional ontologies, such as the Gene Ontology (GO) [94], can be used to test function prediction methods, but do not provide a suitable gold-standard for multidomain homology either. Although numerous protein family classifications exist [26, 88, 95, and work cited therein], these are either generated automatically and not curated or are designed for other goals and not suitable for our study. For example, COGs [40], a repository of orthologous groups, and Gene3D [96],
a compendium of protein families for structural genomics, both use a narrow definition of family, in which proteins with different domain architectures are placed in different families. Restricting families to essentially identical domain architectures is appropriate for the goals of these projects, but does not yield a suitable benchmark for testing methods to identify multidomain families with diverse architectures. The Treefam [97] data base is closer to our goals and may be appropriate for future studies once the curation has progressed further.

Although studies of many multidomain families have been described in the literature, this information has not been compiled in a central resource for use in automated validation of novel methods. To meet this need, we constructed a test set of twenty well studied families for the purpose of evaluating alignment coverage and multidomain homology methods. The curation of this test set included four steps: (1) selection of suitable families; (2) identification of sequences from each family from a large set of amino acid sequences; (3) construction of a set of pairs that meet a minimum similarity threshold through all-against-all comparison; (4) for each family, creation of a set of positive examples (pairs of similar sequences, where both sequences are members of the family) and negative examples (pairs where only one sequence is a member of the family). We constructed this data set using mouse and human sequences. We focus on vertebrate data because the multidomain families that challenge traditional homology identification methods tend to be larger and more complex in vertebrates [51, 58, 66, 78–80]. Many large, modular families found in vertebrates evolved through duplication of a simple ancestral gene, followed by domain accretion and rearrangement in the resulting paralogs, creating the progenitors of modern, mammalian subfamilies. The domain architectures of these progenitor sequences of subfamilies included a set of shared domains characteristic of the family as a whole, as well as auxiliary domains that distinguish each subfamily. These subfamilies expanded further through additional rounds of duplication [85, 98].

The primary criterion in selecting test families was availability of concrete evidence concerning their evolutionary history, from published articles and/or curation by a nomenclature committee. In the best cases, direct syntenic evidence of vertical descent can be found. In other cases, indirect evidence such as conserved intron/exon structure must be used. Phylogenetic evidence can confirm vertical descent, for example, if all domains in a family have consistent phylogenies. However, phylogenetic disagreement between core and auxiliary domains does not rule out homology according to our model. For each family, the evidence used is described in Methods.

After selecting a set of families, we identified sequences drawn from each family from a large set of amino acid sequences. To construct the data set, we extracted complete mouse and human protein sequences from Swiss-Prot [99], resulting in a test set of 18,198 sequences. Of these, 1,281 were assigned to one of our twenty families following a curation procedure described in Methods and the supporting text. To obtain pairs, we conducted an all-against-all BLAST [34] comparison of all sequences in our data set and calculated Neighborhood Correlation scores and alignment coverage ($\alpha$, defined in Methods) for each of the 2,008,066 significant sequence pairs. The significance threshold chosen was $\theta \leq 1$, where $\theta$ is the expected number of chance matches per sequence in our data set (see Methods). For each test family, we extracted all sequence pairs where the query sequence was a member of that family. This set was divided into two subsets: family (FF) pairs, where the matching sequence was also a member of the family, and non-family (FO) pairs, where the matching sequence was not a member of the family. The resulting sets of positive and negative examples (Table 2) were the basis for evaluating multidomain homology identification methods. A scatter plot Neighborhood Correlation and sequence similarity scores for all pairs (Fig. 3) shows that Neighborhood Correlation scores are not linearly related to
sequence similarity, indicating that Neighborhood Correlation reflects information that is not captured by sequence similarity.

The test families do not represent an unbiased sample, since we were limited to families that have been well characterized in the literature. Nevertheless, the families cover a broad range of functional categories, summarized in Table 1. A brief description of the functional and evolutionary properties of each family is given in Methods. The phylogenetic distribution of these families is diverse and biased toward metazoan and vertebrate families. Only ACSL and some Kinase subfamilies occur in bacteria. Six families appear in yeast, as well as in multicellular organisms. Two, the tumor necrosis factors and their receptors (TNF and TNFR), are specific to vertebrates and one (KIR) occurs almost exclusively in primates. The remainder are metazoan families.

With respect to domain architecture, the families represent a variety of challenges for homology identification. Four families (ACSL, FGF, TNF and WNT) encode single domain proteins and follow the classical model of gene family evolution, evolving via vertical descent. An additional three families (FOX, TBOX, and USP) are primarily single domain proteins that evolved via a combination of vertical descent and acquisition of sequence motifs associated with interaction specificity. For USP, these are subcellular localization sequences and functional motifs that confer substrate specificity. The other families are transcription factors. They have short, conserved DNA binding domains and various trans-activation or -repression motifs, sequences that determine conditions for DNA binding. These motifs can result in matches to unrelated sequences, while the short, conserved domains are a challenge for alignment coverage.

The remaining thirteen families encode multidomain proteins. Five of these (DVL, GATA, KIR, NOTCH, and TRAF) have highly conserved sequence and domain architectures, suggesting that a progenitor architecture formed by domain accretion, and was subsequently duplicated to produce several family members. Although the domain architectures are conserved in these families, they contain promiscuous domains or motifs, leading to matches with unrelated sequences. Laminin is similar to these small, multidomain families in the sense that the domain type is conserved. However, the copy number varies due to internal duplications. Seven families have complex and varied domain architectures, resulting from duplication, domain accretion and a second round of duplication as described above. These families are characterized by one or a small number of core domains that define the family and a broad variety of auxiliary domains that control specificity. The variation in domain architectures, as well as the presence of promiscuous domains, is a major challenge for homology identification.

**Accuracy of Homolog Identification**

We compared the ability of each method to discriminate between family and non-family pairs. Each of the ∼2M pairs was scored using Neighborhood Correlation, sequence similarity alone, and sequence similarity with alignment coverage criteria. We considered three alignment cutoffs, $\alpha > 0.3$, $\alpha > 0.6$, and $\alpha > 0.8$, values that span the range of length cutoffs used in the literature (e.g. [55, 57]). Performance was assessed using the Area Under the Curve (AUC) scores, which captures both false positives and false negatives (see Methods). An AUC score of 1.0 indicates perfect separation of the two classes. Random assignment of elements to two groups would result in an AUC score of ∼0.5.

We first evaluated the effectiveness of alignment coverage in separating family and non-family pairs. Remarkably, the addition of an alignment coverage criterion does not improve the performance of sequence similarity, as shown in Table 3. When applied to the combined set of all family pairs, sequence similarity alone achieves an AUC score of 0.8362, compared with a score of 0.7569 when $\alpha > 0.3$. Classi-
fication accuracy further decreases as the minimum required coverage increases: the performance barely differs from random guessing when $\alpha > 0.8$.

When families are considered individually, an alignment cutoff of 0.6 or 0.8 does not significantly improve accuracy in any family. For most families, the performance is worse. When a cutoff of 0.3 was imposed, classification performance improved in three families, remained unchanged in six, and decreased in the remaining eleven. Strikingly, the AUC score decreased by at least 10% in more than one-third of the families.

Are there any circumstances under which adding a minimum alignment coverage criterion improves the performance of sequence comparison alone? As a performance measure, AUC reflects overall discrimination power. However, in some cases specificity ($S_p$) is preferred over sensitivity ($S_n$) or vice versa. A modest length cutoff can be helpful if one is willing to sacrifice sensitivity for greater specificity. A cutoff of $\alpha > 0.3$ eliminates 77% of non-family pairs across all twenty families, but only 13% of family pairs. Plots of sensitivity and specificity (Fig. S4) illustrate this trade-off for individual families. For some families (FGF, GATA, PDE and TBOX), it is possible to obtain $S_p \approx 1$ and $S_n \gtrsim 0.9$ simultaneously given an appropriate significance threshold and an alignment coverage threshold of 0.3. For other families, however, high specificity comes at a high price. For example, if the threshold is selected to obtain a specificity close to one, the sensitivity drops to $S_n \leq 0.5$ for FOX, $S_n \leq 0.65$ for TNFR and $S_n \leq 0.4$ for USP (Figs. 6B, S4).

We next compared the performance of Neighborhood Correlation with sequence similarity. When applied to the combined set of sequence pairs from all families, Neighborhood Correlation dramatically outperforms sequence similarity with or without an alignment coverage criterion (Table 3). Neighborhood Correlation also yields the highest mean AUC score with the smallest standard deviation, showing that Neighborhood Correlation is not only accurate, but consistent. The larger standard deviations seen with sequence similarity and alignment coverage indicate that these methods achieve much better performance with some families than with others.

When applied to the twenty families individually, Neighborhood Correlation had equal or better discriminatory power in 19 out of 20 families, compared with sequence similarity. For ten families, there is no statistically significant difference between the methods; both achieve near perfect separation of the classes. These families all have simple or conserved domain architectures. For nine families, Neighborhood Correlation outperforms sequence similarity significantly ($p$-value $\leq 0.001$). Most of these families are complex, modular families where domain matching is a problem for sequence comparison. For the remaining family, Myosin, Neighborhood Correlation does not perform as well as sequence similarity. This may be because Myosins have very long coiled-coil regions ($>300$ residues, in some cases). Although several other test families also have coiled-coil regions, they are much shorter.

To demonstrate why Neighborhood Correlation is more effective for complex families, we discuss three complex families in detail: Kinase, ADAM (A Disintegrin And Metalloprotease) and USP (Ubiquitin Specific Protease). Similar analyses, including ROC curves and score distributions, are given for all families in Fig. S1-S5.

The Kinases, one of the largest known protein families, are involved in many aspects of cellular control, including cell cycle progression, signal transduction, metabolism and cell movement.

---

1 For KIR, the AUC score for sequence similarity with $\alpha > 0.6$ is higher than sequence similarity alone but this difference is not significant ($p = 0.11$).

2 Note that while the AUC score for TNF was lower for Neighborhood Correlation than for sequence similarity when $\alpha > 0.6$, the difference is not significant ($p = 0.02$).
Kinases achieve this variety of cellular activity through modular domain architectures. In addition to the kinase domain, which mediates phosphorylation, many Kinases have additional domains for recognition of binding partners. The repertoire of such domains is large and diverse and many appear in other protein families as well. In our data set, more than 100 non-kinase domains are found in combination with kinase domains according to the CDD data base [102].

USPs [103] are deubiquitinizing enzymes involved in coordinating many eukaryotic cellular processes that are mediated by modification of proteins by ubiquitin, such as cell cycle progression, DNA repair, transcriptional activation, vesicle trafficking and signal transduction. USPs are characterized by three highly conserved motifs responsible for catalytic activity. Outside of these conserved motifs, USP sequences vary greatly due to the broad range of substrate specificities exhibited by this family, as well as diverse cellular localization signals. USPs can match non-homologous sequences with which they share a limited region of sequence similarity due to these functional properties (e.g., substrate specificity).

ADAMs are zinc-dependent metalloproteases thought to carry out functions such as cell adhesion, fusion, signaling and proteolysis in the context of processes including fertilization, development and inflammation [104]. The typical extracellular ADAM sequence is characterized by a signaling peptide and metalloprotease, disintegrin, cysteine-rich and EGF-like domains. These last are promiscuous and cause ADAMs to match unrelated sequences such as Tenascin, Notch, Integrin and some Kinases. Members of a recently discovered subfamily, ADAMTS [105], also contain thrombospondin type-1 repeats, which are promiscuous as well.

All three families achieve a broad range of cellular function through modular organization, which also confers a tendency to match sequences with which they share only local homology. The specific modular organization of the three families differs. The Kinase family is defined by a single domain: all Kinases have a kinase domain. In contrast, there is no single domain that defines the ADAM family; rather, ADAM sequences are characterized by a specific combination of extracellular domains partnered with a variety of different C-terminal domains. In USP sequences, modularity is due to a variety of sequence motifs rather than structural domains.

Neighborhood Correlation effectively identifies homologous pairs for all three types of modularity. Fig. 4 shows the scores of all matches for three example query sequences, one from each family. Scores of family and non-family pairs are shown in blue and red, respectively. A scoring system is an effective classifier if the red and blue points form separate groups when the scores are sorted. For each of these examples, Fig. 4 shows that when the scores are ranked by significance ($-\log \theta$), there is no cutoff that correctly separates the true homologs from domain-sharing matches. In contrast, when ranked by Neighborhood Correlation score, the family and non-family matches form two distinct groups. Rank plots, such as those shown in Fig. 4, provide a helpful visualization tool, revealing both the structure of the neighborhood of the query sequence and organizational substructure within the family. For example, Fig. 4C shows the separation of the ADAM proteins into two known subfamilies, ADAM and ADAMTS, suggesting that, at least in some cases, Neighborhood Correlation can discover substructure within families, as well as separating family members from domain matches. Rank plots, giving a visual representation of the sequence neighborhood for each of the 18K sequences in our data set, are provided at http://goby.compbio.cs.cmu.edu/NC/. The id is “durand”, the password is “SomethingEasy”.

How typical are these examples? The score distributions in Fig. 5 show that the ability to separate family and non-family pairs extends to the entire family in each case. The $-\log \theta$ distributions of family and non-family pairs overlap completely, while the Neighborhood Correlation score distributions for family and non-family matches are largely distinct, with limited overlap in the tails of the distri-
butions. This demonstrates visually why Neighborhood Correlation achieves more accurate homology identification than sequence similarity.

Fig. 6 shows Receiver Operating Characteristic (ROC) curves (see Methods) for the three families. ROC curves for all families are given in S2. While AUC scores quantify all mis-assignments in a single measure, ROC curves show the relative importance of sensitivity and specificity. Fig. 6 shows that for all three families, Neighborhood Correlation delivers near-perfect sensitivity with very few false positives. In fact, Neighborhood Correlation delivers high sensitivity and specificity simultaneously ($S_n \geq 0.95$, $S_p \approx 0.9$) for most families (S2). In contrast, sequence similarity ($\alpha > 0.0$) achieves a good balance between sensitivity and specificity for ADAM, but is not able to obtain high sensitivity without a high false positive rate for the Kinase and USP families. Obtaining high sensitivity with sequence similarity requires a large sacrifice in specificity in a number of families, as discussed above.

Homolog Prediction

Our ultimate goal is a homolog prediction method that can be used in automated genome scale analyses. In this case, a universal Neighborhood Correlation threshold is needed to identify all homologous pairs. We obtain an AUC score of 0.9462 when all family pairs are treated as a single set. This suggests that a universal threshold can be found that performs well with most families. The distribution of Neighborhood Correlation scores for all sequence pairs in our data set (Fig. 7) has a trough around 0.4, suggesting that this is a good choice.

The threshold can be adjusted if either higher specificity or higher sensitivity is preferred. Since in most families, very few non-family pairs have Neighborhood Correlation scores above 0.7, a threshold of 0.7 improves specificity, although possibly at the expense of sensitivity. Similarly, in most families, few family pairs have Neighborhood Correlation scores below 0.2, suggesting a preferred threshold of 0.2 if sensitivity is the primary goal. Table S1 shows true positive and negative rates for all twenty families with thresholds of 0.2, 0.4, and 0.7. When a threshold of 0.4 is used on our test families, the true positive rate is above 0.95 in 17 families. The true negative rate exceeds 0.97 for 13 families and 0.75 for 18 families. We provide homolog predictions for our data set for all three thresholds as supplementary data.

For a given family, the degree of sequence conservation within the family modulates the Neighborhood Correlation threshold that yields the best classification. Low sequence conservation shifts the distribution of family pairs to the left. For diverged families such as FGF, TNF, TNFR, and USP, a lower threshold such as 0.2 is preferable to 0.4 since fewer family pairs are lost. Similarly, high sequence conservation increases Neighborhood Correlation scores for family pairs. In this case, a higher threshold can be used to eliminate more non-family pairs without increasing the false negative rate. This is particularly helpful in families with coiled coils or highly conserved promiscuous domains, since these increase the score of non-family pairs. For example, a threshold of 0.7 is a better choice for Myosins and Kinesins. In summary, in the absence of prior knowledge, a threshold of 0.4 is recommended. If something is known about the degree of sequence conservation in a particular family of interest, the threshold can be increased or decreased accordingly.

Reevaluation of Alignment Coverage

The poor performance of alignment coverage in classifying multidomain homologs is striking given its widespread use in software tools and genomic analyses [35, 44, 48, 52–59]. We therefore further
investigated the two hypotheses underlying the use of alignment coverage to eliminate domain-based matches:

1. The region of sequence similarity in homologous sequence pairs covers a significant fraction of their length.

2. The fraction of sequence length covered by the aligned region is typically larger in homologous pairs than in unrelated sequence pairs that share an inserted domain.

In order to test these hypotheses, we investigated how the length of the conserved region varies within each family compared with the lengths of alignments with sequences outside the family. For all family and non-family pairs in our data set, we calculated the length of the highest scoring local alignment and the fractional alignment coverage, \( \alpha \) (see Methods). Table 4 shows, for each family, the fraction of family pairs in which \( \alpha \) exceeds a specified threshold (True Positive Fraction), as well as the fraction of non-family pairs in which \( \alpha \) is below the same threshold (True Negative Fraction). Density plots of alignment coverage for individual families are shown for selected families in Fig. 8 and all families in Fig. S1.

Our results do not support the hypothesis that homologous sequences generally share a region of similarity that extends over the bulk of their length. A cutoff of 0.3 will eliminate 40% of true positives, a cutoff of 0.6 will eliminate 90% of true positives, and a cutoff of 0.8 will eliminate 97% true positives. To avoid the possibility that these numbers are biased by the Kinases, the largest family in our data set, we calculated the same statistics for the remaining 19 families, excluding the Kinases. This case is less extreme, but still surprising: a cutoff of 0.3 will eliminate 30% of true positives, a cutoff of 0.6 will eliminate 76% of true positives, and a cutoff of 0.8 will eliminate 82% true positives.

A better understanding of these trends can be obtained by considering individual families in detail. Only four families in our test set (ACSL, NOTCH, TRAF, and WNT) have long regions of similarity (e.g., Fig. 8A). These families are either single domain families (ACSL and WNT) or small multidomain families with conserved domain architecture (NOTCH and TRAF). Another six families have moderate regions of similarity.

Ten families (GATA, FOX, Kinase, Kinesin, Laminin, Myosin, PDE, TBOX, TNFR and USP) clearly violate the hypothesis that homologous sequences have long conserved regions. There are several reasons for this, related to specific properties of the families. First, some families are characterized by a short, conserved DNA binding domain and little conservation elsewhere (e.g., FOX, Fig. 8B). Motifs outside this domain determine the magnitude, duration and/or context of functional activity and are specific to individual family members. These families include the transcription factors in our test set and the enzymes PDE and USP. Second, alignment length is extremely variable in multidomain families that are characterized by a single domain that defines the family, partnered with a variety of auxiliary domains. The length of the alignment depends on the number of domains that a given pair has in common. Since different family members contain different auxiliary domains, this can vary considerably. For example, the density plot for the Kinesin family (Fig. 8C) shows two peaks: A small peak near \( \alpha = 1.0 \) corresponds to pairs with identical domain architectures, while a much larger peak between \( \alpha = 0.2 \) and \( \alpha = 0.6 \) represents Kinesin pairs with different auxiliary domains. Variation in \( \alpha \) can also occur due to differences in copy number. For example, while domain content and order in the Laminin family are quite conserved, differences in copy number result in great variation in alignment coverage (Fig. 8D). Finally, a broad \( \alpha \) distribution can occur if there is considerable variation of sequence
length within the family. In that case, even if the length of the conserved region is constant, the alignment coverage varies because it is expressed as a fraction of the total length.

With respect to the second hypothesis, the alignment coverage distributions show that alignment coverage can only distinguish family pairs from non-family pairs in a minority of our families: ACSL, DVL, NOTCH, SEMA and WNT (e.g., Fig. 8A). In the remaining families, the \( \alpha \) distributions for family and non-family pairs overlap at least partially. In five families, including the Kinases, Kinesins, Laminins, Myosins, and TNFR, promiscuous domains result in long matches with unrelated proteins (e.g., Figs. 8C and D). As a consequence, there is substantial overlap between the alignment length distributions for family and non-family pairs, suggesting that it is not generally true that alignment coverage is longer in homologous pairs than in unrelated pairs.

By considering the optimal alignment alone, we risk underestimating the extent of similarity between homologous sequences. Related sequences may share several distinct, statistically significant regions of similarity. If these appear as separate suboptimal alignments, the additional similarity will not be captured by \( \alpha \). Would our results be different had we used a definition of alignment coverage based on the combined the lengths of optimal and suboptimal alignments? To investigate this, we implemented a simple heuristic (see Methods) for selecting a consistent set of high-scoring local alignments; that is, alignments that appear in the same order in both sequences and do not overlap. We then recomputed the alignment coverage distributions for our twenty families using the combined lengths of this set of consistent alignments.

Surprisingly, including suboptimal alignments in the alignment coverage calculation has little impact on our results. The distributions of the combined alignment lengths, shown as dotted lines in Figs. 8 and S1, differ little from the distribution of optimal alignment lengths (solid lines). The true positive and true negative fractions for the combined set of all family and non-family pairs are essentially unchanged. When families are considered individually, we see slight improvements in two families (Laminin and USP, see Table S4). However, these improvements are not sufficient to change our original observations: the expectation that non-homologous pairs have shorter regions of similarity than homologous pairs holds only in a minority of families.

Our results in twenty families suggest that alignment coverage is not a reliable basis for homology identification. Alignment coverage also presents a number of theoretical concerns. First, the length of the optimal local alignment is not a robust quantity. Dynamic programming algorithms for local sequence alignment stop extending an alignment as soon as a score of zero is reached. A slight change to input parameters, such as the substitution matrix or gap penalty, can eliminate these stops, causing an optimal local alignment to be merged with an adjacent suboptimal alignment. A substantial increase in the value of \( \alpha \) can result. Thus, \( \alpha \) is extremely sensitive to the choice of parameter values, an undesirable property in a quantity used for classification. An additional concern arises for alignment coverage methods that combine the lengths of several high scoring alignments. Since high scoring local alignments in the same sequence pair are not independent events, identification of a maximal set of consistent and significant local alignments requires a multiple testing approach [106]. Most statistically motivated sequence comparison algorithms, including BLAST [34], evaluate the statistical significance of each local alignment independently. Although several methods for combining suboptimal alignments have been proposed [107–111], we are not aware of any studies that take the dependence of events into account in calculating statistics for incorporating suboptimal alignments in alignment coverage criteria.
Discussion

Two obstacles have impeded development of more effective homology identification methods for modular sequences. The first is the absence of formal models. Although models of gene family evolution based on vertical descent have been proposed and debated for more than three decades [60, 61], models of multidomain evolution are in their infancy [112–114]. A second obstacle is the lack of curated, benchmark data sets of multidomain homologs that can be used to evaluate and compare proposed methods. In the current paper, we offer preliminary solutions to both problems: we propose a model that forms a basis for defining multidomain homology and provide a curated test set of homologous mouse and human sequences from twenty well-studied families.

Current models focus on domains or families with identical domain architecture. Traditionally, gene homology is a yes/no question: genes either share common ancestry or they do not. Fitch et al. [61] argued that when different parts of sequences can have distinct evolutionary histories, it is not possible to determine gene homology. Similarly, Rost et al. [115] proposed that “dissecting proteins into domains” is the only reasonable way to study protein relationships. However, focusing exclusively on domains cannot help us to understand the evolutionary processes by which modular genes are formed. In addition, for some research, such as comparative mapping, identification of homologous loci is the prerequisite. Here we propose a model of multidomain homology for families that evolved by gene duplication and insertion of domains into existing genes. Recent evidence from studies of young genes [70,90–92,116], as well as indirect evidence of sequence shuffling, suggest that this model is consistent with a significant fraction of metazoan multidomain families. The utility of the model will be further tested in future studies involving other data sets and applications by ourselves and, we hope, other researchers. Our model does not describe families that evolved through other domain shuffling processes such as the fusion of adjacent genes resulting from read-through errors and de novo formation of novel architectures through insertions in intergenic regions. Extending the model to capture a broader range of domain shuffling scenarios is an important direction for future work.

The second obstacle to multidomain homology identification is the lack of a gold standard data set that can be used to evaluate the performance of new methods empirically. To address this problem, we curated such a data set: twenty, well-studied families for which there is convincing evidence in the molecular evolution literature that the sequences in the family are derived from a common ancestor. Using this data set, we evaluated the effectiveness of the two most commonly used heuristics for detecting multidomain homologs: sequence similarity and alignment coverage. Remarkably, given their widespread use [35,44,48,52–59], our results show that these methods do not produce reliable homology predictions. They are plagued by high false positive and false negative rates for a broad range of families, including families with short, conserved regions. Examples include transcription factors; families with variable sequence lengths; and families with diverse, multidomain architectures. These observations raise doubts about two assumptions underlying alignment coverage: (1) that homologs have long alignments and (2) that local alignments in homologous pairs are distinctly longer than local alignments arising from domain-based matches. Moreover, these results lead us to question whether pairwise comparisons alone contain enough information to distinguish homologous pairs from domain-based matches.

Here, we propose Neighborhood Correlation, a novel method that uses network neighborhood similarity as a basis for identifying homologous sequences. Surprisingly, despite the complexity of the network neighborhood structure, a single numerical value captures neighborhood similarity sufficiently well to classify homologs in our test set with great accuracy. Our method is based on the hypothesis that the evolutionary history of a gene family is reflected in the local structure of the sequence similarity network.
To better understand the effectiveness of Neighborhood Correlation in classifying sequence homology, we consider how the evolutionary processes that drive gene family evolution influence this structure. Recall that Neighborhood Correlation scores depend on the relative number and strength of matches in the shared and unique neighborhoods, as well as how well the edge weights in the shared neighborhood correlate. The families in our data set fall, roughly, into four different evolutionary patterns. For each of these, we consider how the interplay of gene duplication, domain insertion and sequence divergence modulates these properties.

In the traditional view \([60, 61, 117]\), gene families evolve via gene duplication and sequence divergence. In our test set, ACSL, FGF, TNF and WNT fit this model. Generally, proteins in these families match most other family members with strong sequence similarity scores, and a few non-family members with very weak sequence similarity scores. For family pairs, the shared neighborhood is made up primarily of other family members and is much larger than the unique neighborhood (Fig. 9A). Edge weights in the shared neighborhood are strong and well correlated, resulting in high Neighborhood Correlation scores. In contrast, non-family pairs have a very small shared neighborhood with weak edge weights, and large unique neighborhoods with strong edge weights (Fig. 9B). As a result, non-family pairs have much lower Neighborhood Correlation scores than family pairs.

A second evolutionary pattern is gene duplication combined with acquisition of sequence motifs that encode functions such as binding specificity or subcellular localization. Examples of such families in our data set include FOX, TBOX and USP. These functional motifs can also be found in other families, resulting in weak matches to unrelated sequences. This kind of modularity can increase size of the unique neighborhoods of family pairs and the shared neighborhoods of non-family pairs (Fig. 10). However, since the matches are weak, the impact on Neighborhood Correlation score is limited; most family pairs have higher Neighborhood Correlation scores than non-family pairs. Neighborhood Correlation performs well for this type of family.

Third, we observe small, multidomain families with highly conserved sequence and domain architecture. These families include DVL, NOTCH, and TRAF, all of which contain promiscuous domains. Unlike the previous cases, the shared neighborhoods of family pairs contain not only family members but also many unrelated sequences that match those promiscuous domains. Since the sequences in these families are conserved, the scores between both members of the pair and the shared neighbors correlate well, so that even weak edges to unrelated sequences contribute to a strong Neighborhood Correlation score (Fig. 11A). As a result, family pairs have strong Neighborhood Correlation scores. For non-family pairs, the primary contribution to the shared neighborhood comes from matches to a shared promiscuous domain (Fig. 11B). Since most promiscuous domains have weak sequence similarity, the contribution to the score is not great. In contrast, edge weights in the unique neighborhoods of non-family pairs are strong, resulting in low Neighborhood Correlation scores. Neighborhood Correlation performs exceptionally well for these families.

Two additional families represent variants of this pattern. KIR is a small, multidomain family with promiscuous Ig domains. Unlike the other small, multidomain families in our test set, KIR sequences underwent repeatedly domain exchanges within the family. However, since there were no domains insertions from outside the family and the Ig domains in the KIR family are highly conserved compared with other Ig domains, the sequence neighborhoods of the members of KIR family also follow the pattern shown in Fig. 11. Laminin is similar to the other small, multidomain families in that the domain type is conserved. However, the copy number varies due to internal duplications. Since the number of copies has minimal effect on the composition of the neighborhood or the strength of the matches, the
neighborhood organization of Laminin sequences also adheres to the pattern of families with conserved domain architectures.

Finally, we observe large multidomain protein families with diverse architectures, including ADAM, Kinase, Kinesin, Myosin, PDE, SEMA and TNFR. These families formed through duplication, followed by acquisition of auxiliary, and frequently promiscuous, domains, leading to the progenitors of subfamilies. The subfamilies subsequently expanded through additional gene duplications. These families present some of the greatest challenges for homology identification: cases where family pairs do not have identical domain composition (e.g., Box 1). In some family pairs, each member of the pair has one or more promiscuous domains that are not present in the other member. These promiscuous domains increase the unique neighborhoods of the family pair, depressing the Neighborhood Correlation score (Fig. 12A). Similarly, promiscuous domains increase the shared neighborhoods of non-family pairs, augmenting Neighborhood Correlation scores for non-family pairs (Fig. 12B). Nevertheless, our empirical results show that in all test families except Myosin, the Neighborhood Correlation scores of family pairs generally exceed those of non-family pairs. Why is this the case? First, the shared neighborhoods of family pairs are very large. This suggests that the growth of the shared neighborhood due to gene duplication outpaces the growth of the unique neighborhood due to domain insertion. Second, most promiscuous domain superfamilies exhibit a broad range of sequence divergence. This has consequences for both family and non-family pairs. For family pairs, the reduction in Neighborhood Correlation scores is minor, because although the unique neighborhoods are large, their edge weights are weak. For non-family pairs, the Neighborhood Correlation score is not greatly increased because the scores of matches in the shared neighborhood do not correlate well.

It is notable that Neighborhood Correlation can separate some major subfamilies for these sequence families. For example, Neighborhood Correlation separates ADAMTS from other ADAM proteins, Myosin II sequences from other Myosins, and Kinesins with coiled-coils from other Kinesins. This is because the auxiliary domain specific to the subfamily contributes to the shared neighborhood when we compare two sequences from the same subfamily, while the auxiliary domain contributes to the unique neighborhood when we compare sequences from different subfamilies. Hence, pairs from the same subfamily have higher Neighborhood Correlation scores than sequences from different subfamilies.

In summary, several factors are responsible for the fact that homologous pairs have higher Neighborhood Correlation scores than non-homologous pairs. The first is disparity in neighborhood size. Homologous pairs tend to have larger shared neighborhoods than non-homologous pairs, and smaller unique neighborhoods than non-homologous pairs. Second, the scores of shared neighbors for homologous pairs correlate better than the shared neighborhood scores of non-homologous pairs. These factors in turn depend on the particular interplay of the processes of gene duplication, domain shuffling and sequence divergence. Since the processes that govern gene family evolution are not uniform across all domains of life, Neighborhood Correlation may perform differently in other lineages. Many of our test families are specific to metazoan and perform functions that support multicellularity. Multicellularity has evolved several times [118, and work cited therein]. In each case, Nature has had to devise novel evolutionary solutions to the problems of coordinated cellular communication and control. Interaction with the environment also poses communication and control challenges for unicellular organisms. It is an intriguing question whether the same patterns of gene duplication and domain insertion that drive the evolution of metazoan signal transduction families also dominate in other lineages. Future work will determine whether we can exploit local organization of the sequence similarity network to obtain effective homology identification methods in other species.
Methods

Data

We extracted all complete mouse and human protein sequences from Swiss-Prot Version 44 [99], yielding 7,439 mouse protein sequences and 10,759 human protein sequences. This was done by downloading all sequences that were not annotated with the keywords “frag” or “fragment” and had NCBI_taxid equals to 9606 (human) or 10090 (mouse). We chose Swiss-Prot, a high quality, curated protein sequence database, as opposed GenBank, which would have resulted in a larger, but less reliable, data set.

We performed our analysis on the combined data set. To test whether Neighborhood Correlation performance depends on the choice of data set, we evaluated classification performance on the set of mouse sequences and the set of human sequences independently (Tables S2 and S3). These two data sets test classification of paralogs within a single mammalian species, as opposed to the combination of orthologs and paralogs seen in the combined data set. Since paralogous and orthologous sequences in these species exhibit different patterns of divergence, classifiers may perform differently on comparisons within and across genomes. For these species, we expect to see a large set of paralogs that arose at the base of the vertebrate lineage or earlier [55, 119], as well as some very recent paralogs. In contrast, the orthologs date from the primate-rodent divergence. While the details differ, the basic trends in the mouse-only and human-only data sets are the same as the combined data set for all experiments performed (Tables S2 and S3). This suggests that Neighborhood Correlation performance is not highly sensitive to the degree of sequence divergence, with the exception of the initial choice of significance threshold.

Family identification

For each family, we derived a list of designated gene symbols, PFAM [120] and/or InterPro [121] codes from publications by family experts, and reports from the Human Genome Nomenclature Committees (http://www.gene.ucl.ac.uk/nomenclature/genefamily.html). When the symbols of a family have been standardized, symbols are good criteria for family identification. For protein families that are characterized by specific domains, domain information can also be used to identify protein family members. These lists were used to generate a preliminary roster for each family, then confirmed by referring to recent analyses of gene family evolution in the literature. A detailed account of the curation procedure for each family with specific identification criteria and references is given in the supporting text, as well as the Swiss-Prot Accession ID’s for all sequences used. We briefly summary the families here:

**Acyl-CoA synthetase long-chain (ACSL):** Long chain acyl-CoA synthetases are members of an ancient superfamily of AMP-binding proteins involved in lipid metabolism [122]. Unlike other acyl-CoA synthetases, ACSLs primarily catalyze long-chain fatty acids. Eukaryotic ACSLs are uniquely characterized by a 30 to 70 residue linker sequence that is not present in bacterial ACSLs or in AMP-binding proteins that catalyze short-, medium- or very long-chain fatty acids [123]. The presence of this linker region, combined with a conserved single domain architecture, supports common ancestry for this family.

**A Disintegrin and Metalloprotease (ADAM):** ADAMs are zinc-dependent metalloproteases thought to carry out functions such as cell adhesion, fusion, signaling, and proteolysis in the context of processes including fertilization, development, and inflammation [104]. All ADAM sequences are characterized by a conserved extracellular domain architecture that includes metalloprotease, disinte-
grin, and cysteine-rich domains. Many ADAM sequences also contain EGF-like domains. Both the cysteine-rich and EGF-like domains are promiscuous and cause ADAMs to match unrelated sequences such as Tenascin, Notch, Integrin, and some Kinases. Members of a recently discovered subfamily, ADAMTS [105], also contain promiscuous thrombospondin type 1 repeats (TSP1). ADAMs have not been observed in yeast or bacteria, but do occur in fly and worm and, in much larger numbers, in human. The N-terminal domain architecture is conserved in all ADAM proteins, supporting common ancestry for the family.

**Dishevelled (DVL):** DVL is a small family of multidomain proteins that participate in the WNT pathway, a developmental pathway that mediates cell-cell interactions during embryogenesis [124]. The DVL domain architecture is largely conserved. In our data set, all members of the family have DEP, DIX, and PDZ domains. DVL proteins are adapters with many binding partners. This variety of interactions is achieved through its multidomain architecture, in which each domain is responsible for a specific set of interactions. Both the PDZ and the DEP domain occur in a number of unrelated signaling proteins. For this reason, DVL proteins match a large number of sequences outside the family.

DVL homologs have been found in fly, worm, mouse, and human, but not in yeast or bacteria, suggesting that this family arose in metazoans. DVL sequence and domain architecture are both highly conserved. This conservation, along with the small size of the family, suggest that the family arose through duplication of a single progenitor sequence.

**Fibroblast Growth Factor (FGF):** FGF is a single domain family of developmental proteins involved in regulating cell differentiation, proliferation and migration during embryogenesis as well as in tissue repair in adults [125]. In human, there are seven FGF subfamilies, each with similar biochemical and developmental properties. This family has a conserved core roughly 120 amino acids in length with 30% to 60% identity within the core [125]. Overall, FGF sequences have low conservation, although sequence similarity is greater within each subfamily. FGF sequences have been only been observed in metazoans [126]. Phylogenetic analysis, exon-intron structure and conserved synteny support the hypothesis that the vertebrate FGF gene family expanded by gene duplication in two phases, in early metazoan and early vertebrate evolution, respectively [126].

**Forkhead box (FOX):** FOX is a family of transcription factors involved in the regulation of developmental processes, including tissue determination, differentiation and homeostasis [127,128]. FOX genes have been observed in yeast and animals but not in plants. FOX proteins share a conserved winged helix DNA-binding domain approximately 110 residues in length. Family members differ in their trans-activation and -repression domains; regions outside the conserved domain are highly divergent. Most FOX proteins have single domain architectures, but a few also contain the promiscuous domain, FHA.

**GATA binding protein (GATA):** GATA is a family of transcription factors that play a role in determining cell fates during development. All six vertebrate GATA proteins have two tandem zinc finger domains. In addition, GATA 4/5/6 have an N-terminal transactivation domain that is unique to chordates [129, 130]. According to the phylogenetic evidence, these six genes were generated by duplication from a common ancestral gene [130].

**Kinases:** The Kinases [100,101], one of the largest known protein families, are involved in many aspects of cellular control, including cell cycle progression, signal transduction, metabolism and cell movement. Kinases achieve this variety of cellular activity through modular domain architectures. All Kinases share a kinase domain derived from a common ancestor [131,132]. In addition to the kinase domain, which mediates phosphorylation, many Kinases have additional domains for recognition of binding partners. The repertoire of such domains is large and diverse and many appear in other protein
families as well. In our data set, more than 100 non-kinase domains are found in combination with kinase domains according to the CDD data base [102].

Several lines of evidence support common ancestry for the Kinase family. All kinase domains share a common origin [131, 132]. In addition, there is evidence to suggest that kinase domains are not mobile; that is, multidomain kinases arose through insertion of other domains into existing Kinase genes. Tordai et al. have observed that mobile domains tend to be small and have 1-1 intron phase [51]. The pkinase domain is larger than typical promiscuous domains (< 100 residues) and is not 1-1. Moreover, mobile domains typically have auxiliary functions such as adaptor and adhesion domains [51]. In contrast, pkinase has enzymatic function. Finally, roughly 40% of Kinases are single domain proteins [133]. These occur primarily in more ancient Kinase subfamilies (AGC, CAMK, CK1, STE and CMGC). Single domain Kinases are relatively rare in the more recent TK and TKL subfamilies, which are specific to metazoans. This is consistent with the hypothesis that single domain Kinases represent the ancestral state and that multidomain Kinases arose through insertion of mobile domains into existing Kinases.

**Kinesins:** A family of microtubule-associated molecular motors that are transport regulators for organelles and vesicles. Kinesins are involved in cell division and play a role in pattern formation in embryogenesis. All Kinesins share a conserved motor domain ranging in length from 340 to 450 residues [134, 135]. A few family members contain other domains, including a number of coiled-coil (e.g., MAD, SF-assemblin, HOOK, Myosin-tail-1) and/or promiscuous domains (e.g., FHA, PH).

It has been proposed that Kinesins evolved through a first round of duplication of a simple Kinesin, followed by domain insertions which formed the progenitors of the 14 modern Kinesin subfamilies. Additional duplications resulted in proliferation within these subfamilies [134, 136, 137].

**Killer cell Immunoglobulin-like Receptors (KIR):** A family of type 1 transmembrane glycoproteins active in innate immune response through regulation of natural killer cells and some T cells. KIR proteins have an extracellular ligand-binding region containing two or three Ig domains and an intracellular portion responsible for signal inhibition or activation [138]. This family is largely specific to primates: Only two KIR-like genes have been identified in mouse and it is not known whether these are functional [139]. The analogous function in mouse is carried out by the lectin-resembling Ly49 receptors [140].

The pattern of gene family evolution observed in the KIR family is unique within our test set [141, 142]. The KIR family appears to have descended from an ancestral gene in the primate/rodent ancestor. Many KIR genes are specific to a small number of primate species, suggesting lineage-specific duplication and specialization. Human KIR genes are arranged in a tandem cluster on chromosome 19. Repetitive sequences are dispersed throughout this region, in both intergenic and intronic DNA, conferring an extraordinary degree of plasticity on the region. The number of KIR genes varies within the human population, consistent with frequent changes in copy number due to unequal crossing over. Moreover, there is evidence of repeated recombination at domain boundaries between individual KIR genes, resulting in continual swapping of Ig domains within the family [141]. The Ig domains in the KIR family are highly conserved (90% identity), when compared with other members of the highly promiscuous Ig domain superfamily. This evidence suggests that all human KIRs are descended from a single gene via duplication. While there has been domain swapping within the family, no insertions of domains from outside the family have occurred since its inception.

**Laminin:** Laminin is a key component of the extracellular matrix with a broad range of functions related to tissue morphogenesis. There are three laminin subfamilies, α, β and γ. In vertebrates, peptides from these subfamilies unite in different combinations to form trimeric subcomplexes with varying
functional specificity [143,144]. Laminins are composed of Laminin_N, EGF_Lam, Laminin_B and LamG domains repeated in tandem formation. The number of domain copies differs in each subfamily. The similarity in domain content is suggestive of common ancestry, followed by changes in copy number through unequal crossing over. Laminins contain a number of highly promiscuous and/or low complexity domains, including Laminin-II, EGF-Lam, Laminin-B, LamG, SF-assemblin, Laminin-N and Myosin-tail-1, that result in a large number of unrelated matches.

**Myosin:** A family of actin-associated molecular motors implicated in a broad range of functions, including cell polarity, cytokinesis, organelle transport, motility, muscle contraction, and, in some cases, signal transduction [145–147]. All Myosins exhibit the same general organization: an N-terminal motor domain (the Myosin head), a light-chain domain and a C-terminal tail that confers specific functional properties [148–150]. The domain content of the tail region is highly variable and can include a number of promiscuous (e.g., SH3, PDZ, B41) and coiled-coil regions (e.g., Myosin-tail-1). In mammals there are at least 11 Myosin subclasses. Each class is distinguished by a characteristic motor domain and a specific combination of C-terminal domains, implying a mode of evolution through early duplication and domain insertions, followed by additional duplications within the subclasses.

**Notch:** Notch is a small developmental gene family. Notch genes encode multi-functional transmembrane proteins that take on a broad range of signaling functions that determine cell fate [151–153]. Notch domain architecture consists of a series of tandem calcium-binding EGF domains in the extracellular region, several NOD domains and a number of intracellular ankyrin repeats. Although EGF-CA and Ank are promiscuous domains, the sequence of these domains is highly conserved within the family. This family has only been observed in metazoans. The degree of conservation at both the sequence and domain architecture level indicates common ancestry.

**Phosphodiesterases (PDE):** A family of enzymes that regulate intracellular concentrations of cAMP and cGMP [154]. Since these small molecules act as intracellular mediators, PDEs indirectly influence many biological processes. Mammalian PDEs share a C-terminal catalytic core of ~270 residues that is highly conserved. The N-terminal regions of these proteins contain sequence motifs that determine the magnitude, duration and location of their activity through regulation by small molecules specific to each PDE subfamily. The N-terminal portions of these proteins are highly variable due to their modular nature. However, since these regions do not contain promiscuous domains, PDEs match relatively few unrelated sequences. The PDE subfamilies arose from the same ancestral single-domain gene through repeated gene duplication and domain shuffling early in metazoan evolution [89].

**Semaphorin (SEMA):** A family of functionally diverse multidomain ligands. First studied in the context of neuronal cell migration during development, Semaphorins also play roles in a number of other processes related to tissue formation and restructuring during development [155,156]. Recent evidence suggests an immunological role for Semaphorins as well. Semaphorins encode both secreted and transmembrane proteins and are characterized by a conserved extracellular core roughly 500 residues long. Five classes of Semaphorins have been identified in vertebrates. These are distinguished by their C-terminal domain structure which confers the broad functional variation associated with this family. C-terminal domains found in Semaphorins include the promiscuous Ig and TSP1 domains. Semaphorins have been observed in metazoans and viruses, but not in yeast, protozoa or plants.

**T-box (Tbx):** T-box transcription factors are a single-domain family of developmental genes, characterized by a conserved segment of 180-190 amino acids (the t-box) as well as C-terminal sequences required for transcriptional regulation activity [157,158]. Both phylogenetic evidence and synteny support the hypothesis that members of the Tbx family are descended from a single ancestral gene through
several rounds of gene duplication [157, 159]. Tbx genes have been observed in worm, fly, mouse and human, suggesting an early metazoan origin.

**Tumor Necrosis Factors (TNF):** The Tumor Necrosis Factors, acting in concert with the TNF receptors, are primarily associated with signaling pathways in the adaptive immune system [160]. They control cell proliferation and differentiation, as well as apoptosis. TNF is a single domain family characterized by the ~150 amino acid THD domain. Although TNFs have a single domain, that domain that is not highly conserved within the family, having only 20-30% sequence identity. Tumor Necrosis Factors evolved through repeated duplication that coincided with the emergence of the adaptive immune system and MHC expansion [161–163]. Detailed studies, including comparison of phylogenies, and structural and functional relationships, suggest that all vertebrate TNF proteins share a common origin with the Drosophila TNF protein [163].

**Tumor Necrosis Factor Receptors (TNFR):** TNF receptors are transmembrane glycoproteins characterized by a cysteine-rich extracellular domain. Two of the three TNFR subfamilies are defined by cytoplasmic DEATH domains and TRAF interaction motifs, respectively. The third subfamily has no cytoplasmic signaling motifs. TNF receptors have low sequence conservation within the family. Phylogenetic evidence, as well as chromosomal location, supports a history of co-evolution with the TNF ligands [161–163]. This evidence is consistent with evolution of both families by vertical descent.

**Tumor necrosis factor Receptor Associated Factors (TRAF):** TRAFs mediate a broad range of functions associated with the innate and adaptive immune responses, stress response and embryogenesis [164]. They act through many signaling molecules including, but not restricted to, TNF receptors. Most TRAFs contain a C-terminal TRAF-specific RING finger domain, one or more zinc finger domains and an N-terminal MATH domain. The latter is also found in an otherwise unrelated family of metalloendopeptidases, the meprins. TRAF is an adaptable family of genes that has been recruited to varied functions over evolutionary history in response to changes in genomic environment. Small family size, highly conserved sequence motifs, and little variation in domain architecture support descent from a single common ancestor. Although TRAF proteins mediate TNFR function, as a family they are much older [165].

**Ubiquitin Specific Proteases (USP):** USPs are deubiquitinizing enzymes involved in coordinating many eukaryotic cellular processes that are mediated by modification of proteins by ubiquitin [166, 167]. These processes include cell cycle progression, DNA repair, transcriptional activation, vesicle trafficking and signal transduction. USPs are characterized by a catalytic domain containing two highly conserved motifs responsible for catalytic activity. Outside of these conserved motifs, USP sequences vary greatly due to the broad range of substrate specificities exhibited by this family as well as diverse cellular localization signals [103]. USPs can match non-homologous sequences with which they share a limited region of sequence similarity that confers shared functional properties (e.g., substrate specificity).

**Wingless-related MMTV integration site (WNT):** A highly conserved family of single-domain signaling proteins that regulate cell-cell interactions during development of the embryo [168]. There is also evidence that WNT proteins play a role in tissue homeostasis in adults. WNT homologs have been observed in fly, worm, and mammals but are absent from plants, yeast and bacteria. The WNT family is defined in terms of conserved sequence motifs, rather than functional properties. These motifs include a specific pattern of cysteines as well as other conserved residues [168]. The observation that WNT is a conserved single domain protein and that the WNT domain only exists in this family, supports the view that all WNT proteins arose from the same ancestor.
Sequence Comparison
We conducted an all-against-all BLAST search [34] for all sequences in our data set, using the BLOSUM 62 matrix and an affine gap penalty of \(-10 - k\) for a gap of length \(k\). The size of the search space was set to \(Y = n^2\) and the significance threshold to \(E = N\), where \(n\) is the size of the data base in residues and \(N\) is the number of sequences in the data set. The combined data set has \(N = 18,198\) sequences, composed of 7,439 mouse sequences and 10,759 human sequences. There are \(n = 9,466,573\) residues in the combined mouse and human data set.

The quantities \(Y\) and \(E\) are related by the expression \(E = Y \cdot 2^{-S}\), where \(E\) is the expected number of matches with bit score \(S\). \(Y\) and \(E\) must be chosen relative to one each other. Our choice of parameter values treats the all-against-all BLAST search as a single experiment. This approach is roughly equivalent to conducting \(N\) single query blast searches with \(E = 1\) and \(Y = m_x \cdot n\), where \(m_x\) is the length of query sequence \(x\). Treating the all-against-all BLAST comparison as a single experiment has the advantage that the resulting \(E\) values are symmetric (i.e., \(E(x,y) = E(y,x)\)). In contrast, when each query is treated as a separate experiment, the resulting \(E\) values depend on the length of the query sequence and, in general, will not be symmetric. We define \(\theta(x,y) = E(x,y)/N\) to be the expected number of chance hits per sequence in the data set with a score equivalent to, or better than, that of the alignment of query sequence \(x\) with matching sequence \(y\). The significance threshold of \(E = N\) corresponds to \(\theta = 1\) chance hits per sequence, in expectation. This is a reasonably lenient threshold that finds most biologically meaningful matches without accepting an excessive number of unrelated matches.

Asymmetries also occur due to low complexity filtering [169], which is applied only to the query sequence but not to data base sequences. In these cases, we forced symmetry by assigning the better of the two values to both matrix entries. The resulting data set had 2,008,066 significant sequence pairs, with 296,522 pairs within mouse alone and 802,476 pairs within human alone.

Neighborhood Correlation score calculation
We calculated the Neighborhood Correlation scores for all sequence pairs in our data set from Equation 1 using the similarity score

\[
\varsigma(x, i) = \log_{10} \begin{cases} 
S_{\min} & \text{if } \theta(x, i) > 1 \\
S(x, i) & \text{otherwise.}
\end{cases}
\]

\(S_{\min}\) is set to be 32, which is the normalized bit score to the nearest integer corresponding to \(\theta = 1\) for a data set of the size used in this study.

A key component in the Neighborhood Correlation score calculation is the similarity score, \(\varsigma(x, i)\), for significant matches. We experimented with several schemes for determining \(\varsigma(x, i)\). Since the correlation coefficient captures only linear associations, ideally, the scoring function should be linear. In other words, given related sequences, \(x\) and \(y\), with similar neighborhoods, the scatter plot of the points \(\{(\varsigma(x, i), \varsigma(y, i))\}\) should not deviate excessively from a straight line.

Initially, we tried a binary score, defined by setting \(\varsigma(x, i) = 1\) if there is a significant match between \(x\) and \(i\), and zero otherwise. This measure, which gives an unweighted comparison of neighborhood membership, performed extremely well on some families (e.g., Kinases) and poorly on others (e.g., Kinesins). This may be because binary score does not capture differences in the degree of similarity to sequences in the neighborhood.
To address this problem, we considered $\zeta(x, i) = S(x, i)$, where $S(x, i)$ is the bit score [34] of the alignment of $x$ and $i$. We assigned the score $S_{\text{min}}$ to sequence pairs that did not have significant sequence similarity, where $S_{\text{min}}$ is the bit score corresponding to the significance threshold to the nearest integer. Although this variant based on bit scores improved the performance of Neighborhood Correlation in classifying the Kinesins, it did not perform well overall. Scatter plots (not shown) indicates this scoring function does not have linear behavior, leading us to consider $\zeta(x, i) = \log_{10} S(x, i)$. The use of the logarithm compresses the range of $\zeta$, resulting in scatter plots that more closely approximate linearity. Neighborhood Correlation performed well overall when the scores were based on logarithm of the bit score. This is the variant used in the current report. We preferred bit score over $-\log \theta$. Bit score does not depend on the size of the search space, so that bit scores from different searches may be compared.

The choice of $S_{\text{min}}$, the score assigned to pairs without significant similarity, may influence Neighborhood Correlation performance in homology identification. We experimented with values of $S_{\text{min}}$ ranging from 32 to 26, corresponding to thresholds between $\theta = 1$ and $\theta = 100$ in our study, to the nearest integer. The results suggest that varying $S_{\text{min}}$ has little impact on the discriminatory ability of Neighborhood Correlation method as indicated by AUC scores in Table S5.

The factors that influence the Neighborhood Correlation score can be better understood under a transformation of the Neighborhood Correlation formula (Equation 1). Given two sequences $x$ and $y$, let $N_{xy}$ be the shared neighborhood of $x$ and $y$, and $N_x$ and $N_y$ be the unique neighborhoods of $x$ and $y$, respectively. Remember that $S(x, i) = S_{\text{min}}$ for all sequences outside of the neighborhood of $x$. Since $N \gg N_x$ and $N \gg N_{xy}$, this means that $\zeta(x) \approx S_{\text{min}}$. Replacing $\zeta(x)$ with $S_{\text{min}}$ in Equation 1 yields the following approximation for $\text{NC}(x, y)$:

$$\frac{\sum_{i \in N}(\zeta(x, i) - S_{\text{min}})(\zeta(y, i) - S_{\text{min}})}{N^2 \sqrt{\sum_{i \in N} (\zeta(x, i) - S_{\text{min}})^2 \sum_{i \in N} (\zeta(y, i) - S_{\text{min}})^2}},$$

which can further be transformed to

$$\frac{\sum_{i \in N_{xy}} (\zeta(x, i) - S_{\text{min}})(\zeta(y, i) - S_{\text{min}})}{N^2 \sqrt{\sum_{i \in N_{xy} \cup N_x} (\zeta(x, i) - S_{\text{min}})^2 \sum_{i \in N_{xy} \cup N_y} (\zeta(y, i) - S_{\text{min}})^2}}.$$

This equation shows that $\text{NC}(x, y)$ generally increases with $|N_{xy}|$ and decreases with $|N_x|$ and $|N_y|$. $\text{NC}(x, y)$ is also influenced by the magnitude of sequence similarity. $\text{NC}(x, y)$ increases with $S(x, i)$ and $S(y, i)$, when $i \in N_{xy}$ and decreases as $S(x, i)$ and $S(y, j)$ increase, when $i \in N_x$ and $j \in N_y$. Finally, the value of $\text{NC}(x, y)$ will be greatest when $x$ and $y$ match sequences in the $N_{xy}$ with comparable scores; i.e., when $\zeta(x, i) \approx \zeta(y, i)$, where $i \in N_{xy}$.

Alignment coverage

For every pair of sequences, $x$ and $y$, with significant similarity, we calculated the alignment coverage, defined as

$$\alpha(x, y) = 2 \frac{l_a}{l_x + l_y},$$

where $l_x$ and $l_y$ are the length of sequences $x$ and $y$, and $l_a$ is the length of the optimal local alignment. We define $l_a$ to be the number of columns needed to represent it; that is, it includes gapped positions. The length of the optimal alignment between query $x$ and match $y$ will not, in general, be the same as
the length of the optimal alignment between query \( y \) and match \( x \). We forced the alignment coverage to be symmetric by setting both \( \alpha(x, y) \) and \( \alpha(y, x) \) to the maximum of the two values.

In various studies, alignment coverage has also been defined as the ratio of length of the optimal local alignment to the shorter or longer sequence length. In a preliminary study, we compared the behavior of \( \alpha \) with measures based on these different definitions. While the specific values obtained using these measures differ, the trends are similar (data not shown).

By considering only the optimal alignment, we risk underestimating the extent of similarity between homologous sequences. To take suboptimal alignments into account, we used a simple heuristic method for selecting a set of high-scoring local alignments that do not conflict. Two alignments conflict if they overlap or do not appear in the same order in both sequences. More formally, let \( a^i(x, y) = \{(x^i_s, x^i_e), (y^i_s, y^i_e)\} \) be the \( i \)th local alignment, where \( x^i_s \) and \( x^i_e \) are the starting and ending positions of the alignment in sequence \( x \) and \( y^i_s \) and \( y^i_e \) are defined similarly. Two alignments \( a^i \) and \( a^j \) do not conflict if \( x^i_j > x^i_i \) or \( y^i_j > y^i_i \) or \( x^i_j < x^i_i \) and \( y^i_j < y^i_i \). The heuristic for finding a set of non-conflicting, local alignments between sequences \( x \) and \( y \) proceeds as follows:

Let \( S = \emptyset \)
Let \( C = \{a(x, y) | S(a(x, y)) > \text{threshold}\} \)

While (\( C \neq \emptyset \))

Let \( a^i(x, y) \) be an alignment in \( C \) with maximum score.

Remove \( a^i(x, y) \) from \( C \) and place it in \( S \).

Remove all alignments from \( C \) that conflict \( a^i(x, y) \).

The length of the region of similarity is then calculated by summing the lengths of the alignments in \( S \).

Validation

We evaluated classifier performance using the following measures:

**Sensitivity** \((S_n)\) and specificity \((S_p)\) are measures of how well a classifier can correctly identify true positives and eliminate false positives. Sensitivity and specificity are defined as follows:

\[
S_n = \frac{TP}{TP + FN},
\]

\[
S_p = \frac{TN}{TN + FP},
\]

where TP, FP, TN and FN refer to the number of True Positives, False Positives, True Negatives and False Negatives, respectively. In the context of our test, TP represents the number of sequence pairs that have common ancestry and have been correctly identified by the classifier. FP represents the number of pairs that are classified as homologs, but are not family pairs. TN and FN refer to the number of non-homologous pairs that are correctly ruled out and incorrectly included, respectively.

Note that sensitivity and specificity differ from the concepts of precision and recall used in Information Retrieval (IR) (see, for example, [170]). **Recall** denotes the fraction of relevant documents retrieved and is equivalent to sensitivity. **Precision**, however, differs from specificity in that it refers to the fraction of documents retrieved that are actually relevant \((TP/(TP + FP))\). In the IR context, specificity would describe the fraction of irrelevant documents that were not retrieved.
Receiver Operating Characteristic (ROC) curves express the trade-off between sensitivity and specificity as a function of the classifier threshold. Inspection of a ROC curve gives visual insight into classifier performance. A curve approaching the point (0,1) represents an effective classifier which can provide high sensitivity and specificity simultaneously. Note, again, that precision/recall curves, while similar to ROC curves, do not capture the same information. Precision/recall curves express the success of a particular retrieval method as a function of the number of items that were retrieved. ROC curves capture the success of a classification method applied to all elements in the set.

Area Under the ROC Curve (AUC score) provides a single measure of classification accuracy, corresponding to the fraction of correctly classified entities given the best possible choice of threshold [171]. Suppose we are given two classes \( \{a\} \) and \( \{b\} \) of size \( p \) and \( q \), respectively. Let \( a_i, i = 1, 2, ..., p \) and \( b_j, j = 1, 2, ..., q \), be the values assigned to the members of \( \{a\} \) and \( \{b\} \) respectively. The AUC score is calculated as follows:

\[
AUC = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} X(a_i, b_j)}{pq},
\]

where

\[
X(a_i, b_j) = \begin{cases} 
1 & \text{if } a_i > b_j \\
0.5 & \text{if } a_i = b_j \\
0 & \text{if } a_i < b_j 
\end{cases}
\]  

A value of 1.0 indicates perfect separation of the two classes, while random assignment of elements to the two classes would result in a score of roughly 0.5. If we consider sensitivity and specificity to be equally important, a larger AUC score indicates a better classifier.

AUC scores can be used for direct comparison of different methods applied to the same test data set. The statistical significance of the difference between the AUC scores for two classifiers was evaluated using the methods proposed in [172]. This method uses a nonparametric approach to test the null hypothesis that the difference in AUC scores is due to random fluctuations in the sampling process used to obtain the test data. Rejection of this null hypothesis indicates that the difference in AUC scores represents a true difference in the performance of the classifiers.

Acknowledgments

We thank S. F. Altschul, W. E. Brown, S. H. Bryant, L. Y. Geer, D. Hackney, R. A. Hoberman, A. J. Lopez, A. W. Moore, J. H. Nadeau, T. Przytycka, A. Rinaldo, J. Schneider, S. Tavazoie and N. R. Washburn for helpful discussions. This research was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.
Box 1: Disambiguating multidomain homology - a motivating example

Sequence comparison:

Beta platelet-derived growth factor receptor (Pdgfrb) and PKC-delta-interacting protein kinase (Ankrd3), shown at left, are enzymes involved in protein amino acid phosphorylation. Phylogenomic and structural evidence [100,131,132,173], as well as the extreme promiscuity of the Ank and Ig domains, supports the common ancestry of this pair (see Methods). They have a statistically significant alignment with an E value of $1.1e^{-08}$ that covers 29% of the average of their lengths. While they share a common domain (Pkinase), the Ig domains are unique to Pdgfrb and the Ank domain is unique to Ankrd3.

An example of domain-based matching is shown at right. Pdgfrb and Cell adhesion molecule 1 (Ncam1) share two Ig domains, resulting in a significant alignment ($E = 1.1e^{-08}$) with an alignment coverage of 48%. However, they are unrelated and perform different functions: Ncam1 is involved in cell-cell adhesion with no enzymatic functions.

In this example, sequence similarity alone is not sufficient to eliminate domain-based matches. About two thirds of the Kinases that match Pdgfrb have E values less significant than $1e^{-08}$. Setting a significance threshold to eliminate Ncam1 would also eliminate roughly 400 sequences that are related to Pdgfrb. Alignment coverage would not help distinguish these two cases: the homologous pair has a shorter alignment than the unrelated pair. Nor could we separate this case by comparing domain content, since Pdgfrb and Ankrd3 share one domain, while Pdgfrb and Ncam1 share two.
Neighborhood comparison:

Although the homologous pair, PDGFRB and ANKRD3, and the domain sharers, PDGFRB and NCAM1, have pairwise alignments with similar properties (E value, alignment length, number of shared domains), their neighborhoods with weighted sequence similarity network are very different. PDGFRB and ANKRD3 share 540 neighbors, mostly Kinases. These are strong matches due to a shared kinase domain. PDGFRB has 88 unique neighbors, mostly due to weak matches with Ig. ANKRD3 has 173 unique neighbors due to weak matches with Ank. In contrast, PDGFRB and NCAM1 have only 113 matches in common, while PDGFRB has 400 unique neighbors and NCAM1 has 115. Unlike sequence comparison, there is a clear difference in neighborhood structure that can be used to recognize multidomain homology.
<table>
<thead>
<tr>
<th>Functional category</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Process</td>
<td></td>
</tr>
<tr>
<td>Neural development</td>
<td>SEMA, Notch</td>
</tr>
<tr>
<td>Immune response</td>
<td>TNF, TNFR, KIR</td>
</tr>
<tr>
<td>Development and homeostatic regulators</td>
<td>ADAM, FGF, WNT</td>
</tr>
<tr>
<td>Cell-cell/cell-matrix interaction</td>
<td>ADAM, Laminin, Notch</td>
</tr>
<tr>
<td>Molecular Function</td>
<td></td>
</tr>
<tr>
<td>Transcription factors</td>
<td>FOX, GATA, Tbx</td>
</tr>
<tr>
<td>Intracellular signal transducer</td>
<td>Kinase, DVL, TRAF</td>
</tr>
<tr>
<td>Enzyme</td>
<td>ACSL, ADAM, Kinase, USP, PDE</td>
</tr>
<tr>
<td>Motor</td>
<td>Myosin, Kinesin</td>
</tr>
<tr>
<td>Structural molecule</td>
<td>Laminin</td>
</tr>
<tr>
<td>Ligand</td>
<td>FGF, SEMA, TNF, WNT</td>
</tr>
<tr>
<td>Receptor</td>
<td>TNFR, KIR, Notch</td>
</tr>
<tr>
<td>Cellular location</td>
<td></td>
</tr>
<tr>
<td>Extracellular</td>
<td>ADAM, FGF, Laminin, SEMA, WNT</td>
</tr>
<tr>
<td>Transmembrane</td>
<td>ADAM, SEMA, KIR, Kinase, Notch, TNF, TNFR</td>
</tr>
<tr>
<td>Intracellular</td>
<td>ACSL, DVL, FOX, GATA, Myosin, Kinesin, PDE,</td>
</tr>
<tr>
<td></td>
<td>Tbx, Kinase, TRAF, USP</td>
</tr>
</tbody>
</table>

Table 1: Functional properties of the twenty test families.
<table>
<thead>
<tr>
<th>Family</th>
<th>$k$</th>
<th>FF</th>
<th>FO</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1281</td>
<td>369047</td>
<td>77783</td>
</tr>
<tr>
<td>ACSL</td>
<td>10</td>
<td>100</td>
<td>41</td>
</tr>
<tr>
<td>ADAM</td>
<td>62</td>
<td>3674</td>
<td>4022</td>
</tr>
<tr>
<td>DVL</td>
<td>7</td>
<td>49</td>
<td>451</td>
</tr>
<tr>
<td>FGF</td>
<td>42</td>
<td>1724</td>
<td>14</td>
</tr>
<tr>
<td>FOX</td>
<td>69</td>
<td>4761</td>
<td>999</td>
</tr>
<tr>
<td>GATA</td>
<td>12</td>
<td>144</td>
<td>59</td>
</tr>
<tr>
<td>Kinase</td>
<td>671</td>
<td>345675</td>
<td>28709</td>
</tr>
<tr>
<td>Kinesin</td>
<td>44</td>
<td>1936</td>
<td>8225</td>
</tr>
<tr>
<td>KIR</td>
<td>14</td>
<td>196</td>
<td>322</td>
</tr>
<tr>
<td>Laminin</td>
<td>22</td>
<td>484</td>
<td>9343</td>
</tr>
<tr>
<td>Myosin</td>
<td>40</td>
<td>1600</td>
<td>17299</td>
</tr>
<tr>
<td>Notch</td>
<td>8</td>
<td>64</td>
<td>4314</td>
</tr>
<tr>
<td>PDE</td>
<td>38</td>
<td>1044</td>
<td>96</td>
</tr>
<tr>
<td>SEMA</td>
<td>34</td>
<td>1156</td>
<td>538</td>
</tr>
<tr>
<td>Tbxox</td>
<td>22</td>
<td>484</td>
<td>158</td>
</tr>
<tr>
<td>TNF</td>
<td>32</td>
<td>348</td>
<td>23</td>
</tr>
<tr>
<td>TNFR</td>
<td>53</td>
<td>1649</td>
<td>1322</td>
</tr>
<tr>
<td>TRAF</td>
<td>12</td>
<td>144</td>
<td>1223</td>
</tr>
<tr>
<td>USP</td>
<td>51</td>
<td>2371</td>
<td>597</td>
</tr>
<tr>
<td>WNT</td>
<td>38</td>
<td>1444</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 2: Test family statistics. $k$: the number of sequences. FF: homologous pairs. FO: non-homologous pairs.
<table>
<thead>
<tr>
<th></th>
<th>AUC scores for sequence similarity, alignment coverage, and Neighborhood Correlation.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC scores for different methods, combined data set. The largest value in each row is shown in bold. The significance of the difference in the AUC scores for Neighborhood Correlation compared with sequence similarity is expressed as a p-value for each value of $\alpha$. Dashes indicate that the AUCs are equal or not significantly different ($p &lt; 0.001$).</td>
</tr>
<tr>
<td></td>
<td>$\alpha \geq 0.3$</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td>TPF</td>
</tr>
<tr>
<td>All (Optimal)</td>
<td>0.6506</td>
</tr>
<tr>
<td>All (Combined)</td>
<td>0.5958</td>
</tr>
<tr>
<td>ACSL</td>
<td>1.0000</td>
</tr>
<tr>
<td>ADAM</td>
<td>0.9782</td>
</tr>
<tr>
<td>DVL</td>
<td>1.0000</td>
</tr>
<tr>
<td>FGF</td>
<td>0.9954</td>
</tr>
<tr>
<td>FOX</td>
<td>0.3900</td>
</tr>
<tr>
<td>GATA</td>
<td>0.8889</td>
</tr>
<tr>
<td>Kinase</td>
<td>0.6436</td>
</tr>
<tr>
<td>Kinesin</td>
<td>0.7789</td>
</tr>
<tr>
<td>KIR</td>
<td>1.0000</td>
</tr>
<tr>
<td>Laminin</td>
<td>0.7190</td>
</tr>
<tr>
<td>Myosin</td>
<td>0.8825</td>
</tr>
<tr>
<td>Notch</td>
<td>1.0000</td>
</tr>
<tr>
<td>PDE</td>
<td>0.9674</td>
</tr>
<tr>
<td>SEMA</td>
<td>1.0000</td>
</tr>
<tr>
<td>Tbx</td>
<td>0.9132</td>
</tr>
<tr>
<td>TNF</td>
<td>1.0000</td>
</tr>
<tr>
<td>TNFR</td>
<td>0.6992</td>
</tr>
<tr>
<td>TRAF</td>
<td>1.0000</td>
</tr>
<tr>
<td>USP</td>
<td>0.4129</td>
</tr>
<tr>
<td>WNT</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 4: True positive fraction (TPF) and true negative fraction (TNF) for all twenty families.

Top: all pairs. Alignment coverage calculated with optimal alignment (1st row) and combined alignments (2nd row). Bottom: Twenty families. Alignment coverage calculated with optimal alignment.
Figure 1: Evolutionary history of a hypothetical multidomain family showing both gene duplications and domain insertions.
Figure 2: Evolutionary history of multidomain sequences.
Figure 3: $-\log \theta$ vs NC for all 2,008,066 sequence pairs in data set
Figure 4: Rank plots for the query sequences PDGFRB, USP1 and ADAM10.
Figure 5: Distribution of significance and Neighborhood Correlation scores.
Figure 6: ROC curves for the Kinase, USP and ADAM families for Neighborhood Correlation, sequence similarity with and without alignment coverage.
Figure 7: The distribution of Neighborhood Correlation scores for all pairs in data set.
Figure 8: Alignment coverage density plots for WNT, Kinesin, FOX and Laminin.
Figure 9: Evolutionary pattern 1.
Figure 10: Evolutionary pattern 2.
Figure 11: Evolutionary pattern 3.
Figure 12: Evolutionary pattern 4.
Figure Legends:

1. The evolutionary history of a hypothetical multidomain family showing both gene duplications and domain insertions. Genes in the A and B subfamilies share a common ancestor but do not have identical domain composition. Gene C shares a homologous domain with the B genes, but there is no gene that is ancestral to both B and C.

2. Evolutionary history of multidomain sequences. (A) A hypothetical genome with two chromosomes (B) Both chromosomes are copied thorough duplication or speciation. Initially, the two resulting copies are identical. (C) Following sequence divergence, similarity is only retained in coding regions. (D) Domain insertions. (E) Although $g_1$ and $g'_1$ contain different domains, they are homologous. Evidence for this is provided by conserved genomic context.

3. Neighborhood Correlation vs $-\log \theta$ scatter plot.

4. Rank plots for the query sequences Pdgfrb(A), Usp1(B) and Adam10(C). Scores of sequences that matched the query sequence are ranked by significance (left) and Neighborhood Correlation score (right). Homologous (FF) and Non-homologous (FO) matches are shown in blue and red, respectively.

5. Distribution of significance and Neighborhood Correlation scores. Distributions of significance (left) and the Neighborhood Correlation (NC) scores (right) for all sequence pairs where the query sequence is a Kinase(A), USP(B) and ADAM(C) protein. Homologous (FF) and non-homologous (FO) matches are shown in blue and red, respectively. Areas where the distributions overlap are shown in yellow. For the convenience of comparison, the vertical axis in the histograms from the same family is shown on the same scale. As a result, the tops of some peaks are cut off.

6. ROC curves for all classification methods tested. ROC curves for the Kinase(A), USP(B) and ADAM(C) families for Neighborhood Correlation (NC) (blue) and significance with $\theta < 1$ and $\alpha > 0.0$ (green), $\alpha > 0.3$ (dark blue), $\alpha > 0.6$ (magenta), $\alpha > 0.8$ (cyan).

7. The distribution of Neighborhood Correlation scores for all 2,008,066 sequence pairs in our data set.

8. Figure showing density plots of alignment coverage calculated with the optimal alignment length only (solid lines) and with combined non-conflicting alignments (dashed lines) for (A) WNT (B) FOX (C) Kinesin and (D) Laminin. FF distributions are shown in red, FO distributions in blue.

9. Evolutionary Pattern 1: (A) The shared and unique neighborhoods of $F_a$ and $F_b$, two members of a hypothetical family that evolved by gene duplication. The shared neighborhood consists of strong matches to other members of the same family. The unique neighborhoods consist of a very few unrelated sequences with weak, chance similarity to $F_a$ (respectively, $F_b$). (B) The shared and unique neighborhoods of $F_a$ and an unrelated sequence, $O$. The shared neighborhood consists only of $F_a$ and $O$. The unique neighborhoods are made up of members of their respective families.

10. Evolutionary Pattern 2: (A) The shared and unique neighborhoods of $F_a$ and $F_b$, two members of a hypothetical family that evolved by gene duplication. Members of the family subsequently acquired sequence motifs, shown as colored tags, that modulate interactions specificity. The share
neighborhood consists of other members of the same family, some of which have different interaction motifs. Matches within the shared neighborhood are strong. The unique neighborhoods are made up of otherwise unrelated sequences with interaction motifs similar to those of $F_a$ and $F_b$. Matches in the unique neighborhood are weak. (B) The shared and unique neighborhoods of $F_a$ and an unrelated sequence, $O$. The shared neighborhood consists of sequences that contain a sequence motif. The unique neighborhood of $F_a$ consists of other members of the $F$ family. The unique matches to $O$ are sequences with green domains.

11. Evolutionary Pattern 3: (A) The shared and unique neighborhoods of $F_a$ and $F_b$, two members of a hypothetical modular family with conserved domain architecture. The shared neighborhoods consists of strong matches to other family members (shown in green, turquoise and yellow), as well weaker matches to other proteins that share one or more domains with $F_a$ and $F_b$. The unique neighborhoods consist of a very few unrelated sequences with weak, chance similarity to $F_a$ (respectively, $F_b$). (B) The shared and unique neighborhoods of $F_a$ and an unrelated sequence, $O$, that also has a turquoise domain. The shared neighborhood consists of sequences that contain a turquoise domain, including other members of the $F$ family. These family members have strong matches with $F_a$ and weak matches with $O$. The unique neighborhood of $F_a$ consists of sequences with yellow and/or green domains. The unique matches to $O$ are sequences with orange domains.

12. Evolutionary Pattern 4: (A) The shared and unique neighborhoods of $F_a$ and $F_b$, two members of a hypothetical modular family with variable domain architecture. The shared neighborhoods consists of strong matches to other family members, which carry a turquoise domain. The unique neighborhoods consist of sequences that match the domains that are unique to $F_a$ (respectively, $F_b$). (B) The shared and unique neighborhoods of $F_a$ and an unrelated sequence, $O$, that shares an orange domain with $F_a$. The shared neighborhood consists of sequences that contain a orange domain. The unique neighborhood of $F_a$ consists of other sequences in the $F$ family. The unique matches to $O$ are sequences with green domains.
Supplementary data

1. Alpha for all pairs
2. Homology predictions (0.2)
3. Homology predictions (0.4)
4. Homology predictions (0.7)
5. Sequences in gold standard data set

Supporting Information

• Figures:
  - Figure S1: Alignment coverage ($\alpha$) density plots for all twenty families.
  - Figure S2: ROC curves for all twenty families.
  - Figure S3: Distributions of significance and Neighborhood Correlation scores for all families. Each distribution shows all significant pairs where the query sequence was a member of the family. Homologous (FF) and non-homologous (FO) matches are shown in blue and red respectively.
  - Figure S4: Sensitivity and specificity for all twenty families. Sensitivity and specificity as a function of classification threshold for Neighborhood Correlation, sequence similarity alone and sequence similarity with a length criterion ($\alpha > 0.3$) for all families.
  - Figure S5: Scatter plots of significance ($-\log \theta$) versus alignment coverage ($\alpha$) for all twenty family.

• Tables:
  - Table S1: True Positives and Negatives for Neighborhood Correlation thresholds of 0.2, 0.4 and 0.7 for all families.
  - Table S2: AUC score for each family for all classification methods tested - mouse data only
  - Table S3: AUC score for each family for all classification methods tested - human data only
  - Table S4: True positive fraction (TPF) and true negative fraction (TNF) using alignment coverage calculated with combined alignments.
  - Table S5: Impact of $S_{\text{min}}$ on performance of Neighborhood Correlation performance.

• Supporting Text:
  - Curation of test families.
  - Test family accession IDs.
References


