Domain Architecture in Homolog Identification

N. Song, R.D. Sedgewick and D. Durand

Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Abstract. Homology identification is the first step for many genomic studies. Current methods, based on sequence comparison, can result in a substantial number of mis-assignments due to the alignment of homologous domains in otherwise unrelated sequences. Here we propose methods to detect homologs through explicit comparison of domain architecture. We developed several schemes for scoring the similarity of a pair of protein sequences by exploiting an analogy between comparing proteins using their domain content and comparing documents based on their word content. We evaluate the proposed methods using a benchmark of fifteen sequence families of known evolutionary history. The results of these studies demonstrate the effectiveness of comparing domain architectures using these similarity measures. We also demonstrate the importance of both weighting critical domains and of compensating for proteins with large numbers of domains.

Introduction

Homology identification arises in a broad spectrum of genomic analyses, including annotation of new whole genome sequences, construction of comparative maps, analysis of whole genome duplications and comparative approaches to identifying regulatory motifs. Currently, sequence comparison methods are widely used to identify homologous genes. These methods assume that sequences with significant similarity share common ancestry, *i.e.* are homologs. However, the existence of multi-domain proteins and complex evolutionary mechanisms pose difficulties for traditional, sequence based methods. A domain inserted into two unrelated protein sequences causes those sequences to have a region of similarity, resulting in mis-assignment of homologs.

To address this issue, researchers frequently also require that the alignment between a pair of potential homologs extend over a large fraction of their lengths [1]. However, the accuracy of this *alignment coverage* heuristic is unknown and there are many examples where alignment coverage makes an incorrect determination of protein homology. For example, human FOXB1 and FOXP3 are known homologs with a significant BLAST [2] e-value of $10^{-8.95}$, but the alignment covers only one fourth of the shorter protein sequence. A heuristic to determine homology based on alignment coverage would fail to recognize these proteins as homologs. Conversely, human KIF5C and mouse BICD2, which have a region of sequence similarity due to a shared HOOK domain, have a BLAST e-value of $10^{-7.26}$ and an alignment coverage of more than half of the length of the shorter sequence. They are not in the same family, but an alignment coverage heuristic would decide that they are. An accurate and automatic method for the identification of homologs remains an open problem.

In this work, we investigate explicit comparison of domain architecture in predicting homology. Complex, multidomain families, such as membrane-bound receptors and cellular matrix adhesion proteins, are characterized by varied domain architectures within a single family. For example, the Kinase domain partners with over 70 different domains in the eukaryotic Kinases. A given member of the Kinase family may contain from one to a dozen domains. Typically, a pair of Kinases has one or more shared domains, but each member of the pair also has domains that do not appear in the other. At the same time, these protein sequences also share domains with sequences not in the Kinase family. The challenge is to determine which aspects of domain architecture (e.g., total number of domains in the sequence, the set of distinct domains, copy number, domain promiscuity) are most informative for separating homologs from unrelated sequences that share a domain.

In order to develop measures of domain architecture similarity we exploit an analogy between domain architecture composition and a problem in information retrieval, namely, determining the similarity of two documents drawn from a corpus. In this metaphor, the word content of a document is analogous to the domain content of a protein sequence and the set of protein sequences under study is analogous to the set of documents in the corpus.

In this work, we adapt information retrieval techniques to domain architecture comparison as a method for identifying multi-domain homologs. We evaluate the effectiveness of several methods on fifteen different protein families by applying each method to test sets composed of positive examples (pairs of proteins both in the family) and negative examples (pairs of proteins with only one member in the protein family). We use this empirical approach to determine:

- whether domain content comparison is, in fact, an effective method for identifying homologs,
- what information about domain content is most informative for this purpose,
- what measure of domain architecture similarity is most effective in identifying homologs.

Model

A domain is a sequence fragment that will fold independent of context into a protein subunit with specific shape and function. Domains are natural evolutionary units. New domain architectures arise via complex mechanisms such as nonhomologous recombination, transposable elements and retrotransposition [3–10]. About two thirds of proteins in prokaryotes and 80% of proteins in eukaryotes are multi-domain proteins [11].

Domain databases [12–17] store probabilistic, sequence-based models of protein domains. These models, typically encoded as a Position Specific Scoring Matrix (PSSM) or a Hidden Markov Model (HMM), can be used to determine the location and type of domains in amino acids sequences. These models can therefore be applied to new sequences about which domain information is unknown.

Using domain databases, a protein can be represented as an unordered list of the domains it contains. For this paper, we compare proteins based solely on

this unordered list of domains, ignoring any differences in the linker sequences between domains. Any two instances of the same domain will have some variance at the sequence level. To simplify the comparison, we treat every instance of a domain as identical.

The use of domain architecture composition to define homology has not been previously investigated. A few papers in the literature have considered measures of domain content similarity in other contexts. CDART (Conserved Domain Architecture Retrieval Tool) [18], a web-based resource, presents a list of sequences with similar domain architecture to a given query sequence. Similarity is calculated by counting the number of domains common to the two proteins. A distance metric for domain architecture comparison has been considered in the framework of evolutionary events [19], where distance is defined to be the number of domain insertions and deletions needed to transform one protein to the other. The effectiveness of these methods in determining protein homology was not evaluated.

Now consider again the analogy with document similarity in information retrieval. A simple measure of the similarity of two documents is the number of shared words. Documents that share a significant number of words are more likely to be on the same subject than documents that share few words. Similarly, two proteins that share many domains are more likely to be homologs than sequences with few domains in common. This is the basis of the method used in CDART. Similarity scores based on common domains can take the number of instances of each domain (the copy number) into account (as done by CDART) or can ignore copy number and simply count the number of distinct types of shared domains.

Two long documents are also more likely to share a large number of words than two shorter documents. To correct for this effect, a length correction is used, where length is typically measured by word count. Correcting for the domain count is also useful in the context of comparing protein domain composition. Jaccard similarity, which is commonly used in information retrieval, is appropriate here. It is given by

$$J(p_1, p_2) = \frac{n_{12}}{n_1 + n_2 - n_{12}}, \tag{1}$$

where n_{12} is the number of domains common to both sequences p_1 and p_2, n_1 is the number of domains in p_1, and n_2 is the number of domains in p_2. Using a Jaccard similarity instead of an uncorrected score allows comparisons among a set of sequences with large differences in the number of domains.

These similarity measures treat all words (or domains) equally. However, words in a document are not equally important to determining its subject. The word "big" conveys less information about the subject of document than the word "parsimony," for example. Words should be weighted based on how informative they are. One measure of the power of a given word to distinguish between subjects is the *inverse document frequency* [20], a measure based on the observation that a word that occurs in very few documents is more likely to differentiate between subjects then a word that occurs in a large fraction of the documents in the corpus. Similarly, some domains are more informative than others. "Promiscuous" domains are domains that occur in many unrelated proteins [11, 21]. Intuitively, promiscuous domains and non-promiscuous domains

reveal different amounts of evolutionary information and should be treated differently. We consider several methods for weighting promiscuity.

Since promiscuous domains occur in many unrelated sequences, they are less useful for determining homology than relatively rare domains. We can adapt the inverse document frequency to obtain a weighting system to reduce the emphasis of promiscuous domains on the score. This gives an idf weight for a domain d of

$$w_{\text{idf}}(d) = \log_2 \frac{|\mathcal{P}|}{|\{p|d \in D(p), p \in \mathcal{P}\}|}, \tag{2}$$

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

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

$$w_{\text{tf}}(d, p) = \frac{N(d, p)}{|D(p)|}, \tag{3}$$

where $N(d, p)$ is the number of occurrences of the domain d in the protein p and $|D(p)|$ is the number of domains in the protein p.

Typically, words are weighted using the product of w_{tf} and w_{idf},

$$w_{\text{tf-idf}}(d, p) = w_{\text{tf}}(d, p) \times w_{\text{idf}}(d). \tag{4}$$

By comparing the results from idf with tf-idf weighting, we obtain a measure of the importance of copy number to multidomain homology detection. Note that since w_{tf} includes the total number of domains in protein sequence p in the denominator, the tf-idf weight includes a correction for the number of domains in a protein, while the idf weight does not.

We designed a third approach for weighting domains considering how multidomain protein sequences evolve. Both gene duplication and domain shuffling are involved in the evolution of multidomain protein families. Gene duplication results in a pair of related protein sequences sharing a domain, while domain shuffling results in unrelated protein sequences sharing a domain. A domain that appears in many unrelated sequences is not a strong indicator of homology. However, idf weighting simply measures the number of sequences in which a domain appears without considering whether or not those sequences are related. A domain that appears in many sequences, but always in the same context, could be very informative yet have a low idf weight. For example, if a domain appears in all sequences in a large gene family that arose through repeated gene duplication and does not appear in sequences from any other family, it should be a strong indicator of homology.

A measure of domain promiscuity that captures this idea is the number of distinct domain partners associated with it, where two domains are *partners* if

they co-occur in at least one protein. For this reason, we consider weighting a domain by the inverse of the number of distinct domain types with which the domain occurs in sequences. We refer to this weighting method as "distinct partner" weighting:

$$w_{\mathrm{dp}}(d) = \frac{1}{|\{d_i | d_i \in D(p), d \in D(p), p \in \mathcal{P}\}|} \tag{5}$$

Note that the denominator is the number of distinct domain partners of the domain d. This weighting was first developed for studying protein functions in [22], but, to our knowledge, the analogous weighting method has not been used in information retrieval.

Regardless of which weighting scheme is used (idf alone, tf-idf, or distinct partner weighting), we calculated the weighted domain comparison score as follows:

$$S(p_1, p_2) = \sum_i w(d_i, p_1) w(d_i, p_2), \tag{6}$$

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

Although the tf-idf and distinct partner weight can give a measure of the importance of different domains, they do not include any correction for the number of domains in a protein and they cannot be used with the Jaccard similarity, which does not apply in the weighted case. This is addressed in information retrieval using the cosine similarity, which is similar to the Jaccard similarity but can be used on either weighted or unweighted domains. For two proteins p_1 and p_2, the cosine similarity is given by

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

Treating the protein sequences as vectors in a vector space with number of dimensions equal to the number of domain types, cosine similarity is the cosine of the angle between the two weight-vectors. In the case of unweighted methods the cosine similarity reduces to $n_{12}/\sqrt{n_1 n_2}$, a variant of the Jaccard similarity.

Experiments

Data

We extracted all complete mouse and human protein sequences from SwissProt Version 44 [23], released in 09/2004 (`http://us.expasy.org/sprot/`), yielding 18,198 protein sequences. We focused on vertebrate data because the multi-domain families that challenge traditional homology identification methods tend to be larger and more complex in vertebrates.

We obtained protein domain architectures from CDART [18]. Among 18,198 sequences, 15,826 sequences have detectable domain architectures. An all-against-all comparison of all 15,826 domain architectures yielded 2,371,608 pairs which share at least one domain.

Family identification

To test the various methods we required a set of pairs of proteins with known homology. We constructed a list of sequences from fifteen known families using reports from nomenclature committees (`http://www.gene.ucl.ac.uk/nomenclature/`) and articles from the literature. Types of evidence presented in these articles include intron/exon structure, phylogenetics and synteny. As a result of this process, we have a list of 1,137 proteins each known to be from one of the following families: ADAM (a family of proteins containing A Disintegrin And Metalloproteinase domain) [24–26], DVL (Dishevelled protein family) [27, 28], FOX (Forkhead transcription factor family) [29, 30], GATA (GATA binding proteins) [31, 32], Kinase [33–36], Kinesin [37–39], KIR (Killer cell Immunoglobulin-like Receptors) [40, 41], Laminin [42, 43], Myosin [44–46], Notch [47–49], PDE (Phosphodiesterases) [50], SEMA (Semaphorin) [51, 52], TNFR (Tumor Necrosis Factor Receptors) [53, 54], TRAF (Tumor necrosis factor Receptor Associated Factors) [55], and USP (Ubiquitin Specific Proteases) [56, 57].

From this database of sequence families, we constructed a list of positive and negative examples of homologous sequences. For each family, all pairs of sequences with both members in the family are positive examples of homologous pairs, and all pairs sharing at least one domain but with only one member in the family are examples of non-homologous pairs. The number of homologous pairs for the fifteen families studied ranged from 75 pairs in the KIR family to 1,074,570 pairs in the Kinase family. The number of non-homologous pairs ranged from 78 in the FOX family to 184,107 in the Kinase family.

Performance Evaluation

Our goal is to assign a similarity score to each pair such that all pairs within the family have scores that are higher than scores between protein pairs with one member in the family and one non-family member. We attempted this using the number of common domains similarity score (from here on referred to as the unweighted score) and the weighted number of domains in common similarity score, trying distinct partner weighting as well as idf weighting and tf-idf weighting. We considered each scoring system both with and without domain-count correction. To evaluate how well these methods could perform in a situation where family identification is not known and therefore a universal cutoff must be used, we also applied these methods to the aggregate set, in which the homologous and non-homologous pairs from all the families are combined to a single set of positive and negative examples.

We evaluated classifier performance using Area Under receiver operating characteristic Curve (AUC) scores. AUC score provides a single measure of classification accuracy, corresponding to the fraction of correctly classified entities given the best possible choice of threshold [58]. The range of AUC scores is from 0.0 to 1. The higher the AUC score, the better the performance. When AUC is 1 the classifier has perfect performance and when AUC is 0.5 the classifier cannot distinguish homologous pairs from non-homologous pairs, *i.e.* the score distributions for positive and negative examples are not separable. When the AUC score is less than 0.5 the positive and negative examples are somewhat

separable, but the classifier is separating them in the wrong direction. For this range of AUC scores the classifier is under-performing randomly guessing if the pair is homologous or non-homologous.

Results and Discussion

	no weighting		distinct partner		idf		tf-idf	
		cosine		cosine		cosine		cosine
ADAM	0.94	0.96	0.98	0.98	0.97	0.98	0.84	0.88
DVL	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00
FOX	0.50	1.00	1.00	1.00	0.00	0.99	0.99	0.99
GATA	1.00	0.99	1.00	0.94	0.63	0.87	0.83	0.87
Kinase	0.44	0.71	0.16	0.49	0.09	0.53	0.47	0.46
Kinesin	0.52	0.90	0.99	0.96	0.85	0.93	0.94	0.93
KIR	0.50	0.86	0.50	0.86	0.50	0.77	0.77	0.77
Laminin	0.99	0.97	0.99	0.95	0.97	0.90	0.96	0.97
Myosin	0.64	0.56	0.94	0.65	0.98	0.74	0.76	0.70
Notch	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
PDE	0.54	0.84	0.84	0.87	0.84	0.82	0.82	0.62
SEMA	0.78	0.89	0.99	1.00	0.99	0.99	0.97	0.87
TNFR	0.57	0.96	0.83	0.95	0.30	0.94	0.96	0.95
TRAF	0.85	0.98	0.99	0.99	0.99	1.00	0.99	0.99
USP	0.55	0.84	0.82	0.77	0.57	0.82	0.87	0.85
all	0.46	0.73	0.19	0.55	0.13	0.60	0.51	0.52
all (no Kinase)	0.63	0.89	0.97	0.92	0.91	0.91	0.91	0.88

Table 1. AUC scores for identifying homologs from different families. Results are shown both with and without domain-count correction using cosine similarity. The last two rows contain the AUC scores for the aggregate system containing all the families, and the aggregate system containing all the families except for Kinase.

We tested the ability of domain content similarity to classify homologs using the unweighted score, distinct partner weighting, idf weighting, and tf-idf weighting. The results are shown in Table 1. For each method, results are reported both with and without the cosine similarity for domain-count correction. For the unweighted scoring method, Jaccard similarity was also tested, although the results are not reported here as they behave similarly to the cosine similarity results. To understand the effect of domain copy number, we also tested a similarity score based on types of domains in common that ignores domain copy number, but results are not shown as for most families it was similar to the unweighted method.

As the ability to separate positive and negative examples in the aggregate dataset best mirrors the way that these comparison methods will be used in practice, we use these results to judge the overall performance of each comparison method. Because of the large size of the Kinase family, the aggregate results

are dominated by the results on Kinase pairs. We therefore consider the aggregate set not including the Kinase family. We see that the unweighted method with no domain-count correction performs poorly with an AUC score of only 0.63. Methods that weighted domains to account for domain promiscuity significantly improve the performance. The aggregate AUC score increases to 0.91 or better by using any of the weighting methods discussed here. Among the three weighting methods considered, we obtain the best performance from distinct partner weighting. It best reflects the evolutionary information encoded by each domain. Cosine similarity provides a domain-count correction and boosts the performance of the unweighted similarity score to 0.89. While there is a significant benefit to using either weighting or a domain-count correction, there is little additional benefit to applying both. The best AUC score of 0.97 is obtained by distinct partner weighting with no domain-count correction. For the Kinase family, the methods based on domain weighting give similar or significantly worse AUC scores as compared to the unweighted methods. The best AUC score for the Kinase family is 0.71, obtained by the unweighted cosine similarity method.

When families are considered individually, the results show that for most families both weighting and domain-count correction are helpful, consistent with the results observed in the aggregate dataset. However, for some families, the performance under different comparison methods differs significantly from the trends of the aggregate dataset. To better understand what information is most informative for domain architecture comparison and the advantages and disadvantages of each method, we will analyze these families in details.

GATA: All methods perform well on the GATA family (AUC score greater then 0.80) except idf weighting, where an AUC score of only 0.63 is obtained. Unlike other scoring systems presented in Table 1, idf weighting does not reflect domain copy number in the similarity score. The poor performance of the idf weighting combined with the good performance of the other scoring systems implies that the domain copy number is important for GATA classification. Each member of the family has two GATA domains, while non-homologous proteins that share a domain have only one GATA domain. For similarity scores that incorporate the domain copy number, these duplicate domains give GATA family-family pairs improved similarity scores.

To better isolate the effect of domain copy number, we calculated the AUC score where the similarity between two protein sequences is given by the number of shared domain types (ignoring the duplicate copies). For most families, the AUC score using this method was not significantly changed from the unweighted similarity score (which included duplicate domains), but for GATA the AUC score drops from 1.0 to 0.61. The next most significant decline in AUC score occurred in Laminin, where the AUC score dropped from 0.99 to 0.91. For all other protein families the decline in AUC score was less than 0.05.

TNFR: The poor performance (AUC score of 0.57) of the unweighted similarity score in the TNFR family was improved with domain-count correction, distinct partner weighting, or tf-idf weighting. All TNFR family members in our dataset have at least one TNFR domain. The DEATH domain is often also present, and the TNFR_c6 domain is occasionally present, but no member of the TNFR family has more than three domains. TNFR sequences also match unrelated sequences with a large number of domains. Therefore any of the scoring methods that

incorporate a domain-count correction (either cosine similarity or tf-idf weighted score) will give a larger similarity score to the homologous TNFR pairs, which have fewer domains.

The distinct partner weighting method works relatively well, as the TNFR domain only has 7 distinct partners, while the DEATH domain has 14 distinct partners. Moreover, the majority of the non-homologous pairs share only a DEATH domain, while the homologous pairs all share a TNFR domain. Conversely, idf weighting fails since the TNFR domain occurs in 47 proteins in the database, while the DEATH domain occurs in only 36. Therefore, the DEATH domain has a larger idf weight than the TNFR domain, so that non-homologous pairs that share a DEATH domain have larger similarity scores than the homologous pairs that share only a TNFR domain. This results in a AUC score of 0.3 for the idf weighted score.

FOX: The FOX family stands out in our test set as the AUC score for idf method is zero. This means that all non-homologous FOX pairs have a higher score than the homologous pairs. Reason behind this is similar to the reason that idf weighting performed poorly with the TNFR family. In our test set, all FOX sequences except one are single domain proteins with the FH domain. The only multi-domain FOX sequence contains a promiscuous domain, the FHA domain. The FH domain is a strong indicator of homology and the FHA domain is not. Although the FH domain exists in a larger number of sequences than the FHA domain (69 vs 27), the FH domain has a smaller number of distinct partners than the FHA domain (FH occurs only with FHA, while FHA occurs with 11 other domains). Therefore, distinct partner gives a higher weight for the FH domain than the FHA domain, while idf gives a higher weight for the FHA domain. Since all homologous pairs share a FH domain and all non-homologous pairs share a FHA domain, the distinct partner method has a AUC score of 1.0 while the idf weighting method has a AUC score of 0.0.

Kinase: All weighting methods fail for the Kinase family. Surprisingly, the distinct partner weighting and idf weighting score distributions *are* separable, but the distributions of scores are inverted: non-homologous pairs are given larger scores than homologous pairs. This is because the Kinase family is a large, complex family. In our database, there are more than 600 Kinases. Since Kinase proteins, which are characterized by Pkinase domain, have varied domain architectures, the Pkinase domain has the largest number of distinct partners in our dataset. Therefore, although Pkinase domain characterizes the Kinase family, it was given a small score under all weighted scoring systems, and Pkinase is treated like a promiscuous domain. This explains why the score distributions are inverted: Kinase-Kinase pairs share a Pkinase domain which has a low weight, while the non-homologous pairs share a non-Pkinase domain with higher weight.

Myosin: Unlike all other families in our test set domain-count correction does not improve classification performance for the Myosin family. Domain-count correction gives higher similarity scores to domain matches in two proteins with a small number of domains than it does to domain matches between proteins with a larger number of domains. Therefore, when the family members have relatively many domains but share only a few, and the outsider proteins have relatively few domains, then domain-count correction will worsen the AUC score of a family.

This occurs in the Myosin family, causing the AUC scores for the comparison methods with domain count correction to be lower than the uncorrected scores.

Conclusion

For many families domain architecture comparison is an effective method for identifying homologs. However, these methods are challenged by large families defined by promiscuous domains, such as the Kinase family. For most families, two key aspects are useful for domain content comparisons: a scoring system that corrects for the bias of proteins with a large number of domains and a measure of the importance of the domain in determining family membership. Domain count correction significantly improves the performance for fourteen of the fifteen studied families. Weighting systems such as idf, tf-idf, and distinct partners are helpful for all families in the test set other than the Kinase family. The best performance is obtained by the distinct partner weighting system. For the Kinase family, the weighting systems under study did not work well as they gave a low weight to the Pkinase domain which is a major factor in determining Kinase family membership. For a few families, such as GATA and Laminin, domain copy number useful in determining sequence homology. Outstanding problems that we hope to address in future work include designing a weighting method that works well for the Kinase family as well as the other families and comparing these domain based methods to other homology identification methods based on sequence comparison and alignment coverage.

Acknowledgments

We thank S. H. Bryant, L. Y. Geer, and J. Joseph for helpful discussions. This research was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.

References

1. Huynen, M.A., Bork, P.: Measuring genome evolution. PNAS **95**(11) (1998) 5849–5856
2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**(17) (1997) 3389–3402
3. Gilbert, W.: The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol. **52** (1987) 901–905
4. Patthy, L.: Genome evolution and the evolution of exon-shuffling–a review. Gene **238**(1) (1999) 103–114
5. Eichler, E.E.: Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet **17**(11) (2001) 661–669
6. Emanuel, B.S., Shaikh, T.H.: Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet **2**(10) (2001) 791–800
7. Kaessmann, H., Zollner, S., Nekrutenko, A., Li, W.H.: Signatures of domain shuffling in the human genome. Genome Res. **12**(11) (2002) 1642–1650

8. Wang, W., Zhang, J., Alvarez, C., Llopart, A., Long, M.: The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in drosophila melanogaster. Mol Biol Evol **17**(9) (2000) 1294–1301
9. Long, M.: Evolution of novel genes. Curr. Opin. Genet. Dev. **11**(6) (2001) 673–680
10. Long, M., Thornton, K.: Gene duplication and evolution. Science **293**(5535) (2001) 1551
11. Apic, G., Gough, J., Teichmann, S.A.: Domain coalmbinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol **310**(2) (2001) 311–325
12. Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., Bork, P.: Recent improvements to the smart domain-based sequence annotation resource. Nucleic Acids Res **30**(1) (2002) 242–244
13. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L.L.: The pfam protein families database. Nucleic Acids Res **30**(1) (2002) 276–280
14. Corpet, F., Gouzy, J., Kahn, D.: The prodom database of protein domain families. Nucleic Acids Res **26**(1) (1998) 323–326
15. Gracy, J., Argos, P.: Domo: a new database of aligned protein domains. Trends Biochem Sci **23**(12) (1998) 495–497
16. Heger, A., Holm, L.: Exhaustive enumeration of protein domain families. J Mol Biol **328**(3) (2003) 749–767
17. Murzin, A., Brenner, S., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247**(4) (1995) 536–40
18. Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H.: Cdart: protein homology by domain architecture. Genome Res. **12**(10) (2002) 1619–1623
19. Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J., Elofsson, A.: Domain rearrangements in protein evolution. J Mol Biol **353**(4) (2005) 911–923
20. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523
21. Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor M., G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W.e.a.: Comparative genomics of the eukaryotes. Science **287**(5461) (2000) 2204–2215
22. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. Science **285**(5428) (1999) 751–753
23. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.: The universal protein resource (UniProt). Nucleic Acids Res. **33** (2005) D154–9
24. Nicholson, A.C., Malik, S.B., Logsdon, J.M.J., Van Meir, E.G.: Functional evolution of ADAMTS genes: evidence from analyses of phylogeny and gene organization. BMC Evol Biol **5**(1) (2005) 11
25. Stone, A.L., Kroeger, M., Sang, Q.X.: Structure-function analysis of the adam family of disintegrin-like and metalloproteinase-containing proteins (review). J Protein Chem **18**(4) (1999) 447–465
26. Wolfsberg, T.G., White, J.M.: Adams in fertilization and development. Dev Biol **180**(2) (1996) 389–401
27. Wharton, K.A.: Runnin' with the Dvl: proteins that associate with Dsh/Dvl and their significance to Wnt signal transduction. Dev Biol **253**(1) (2003) 1–17
28. Sheldahl, L.C., Slusarski, D.C., Pandur, P., Miller, J.R., Khl, M., Moon, R.T.: Dishevelled activates Ca2+ flux, PKC, and CamKII in vertebrate embryos. J Cell Biol **161**(4) (2003) 769–77

29. Mazet, F., Yu, J.K., Liberles, D.A., Holland, L.Z., Shimeld, S.M.: Phylogenetic relationships of the fox (forkhead) gene family in the bilateria. Gene **316**(Oct 16) (2003) 79–89
30. Kaestner, K.H., Knochel, W., Martinez, D.E.: Unified nomenclature for the winged helix/forkhead transcription factors. Genes Dev. **14**(2) (2000) 142–146
31. Lowry, J.A., Atchley, W.R.: Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. J Mol Evol **50**(2) (2000) 103–115
32. Patient, R.K., McGhee, J.D.: The GATA family (vertebrates and invertebrates). Curr Opin Genet Dev **12**(4) (2002) 416–22
33. Robinson, D.R., Wu, Y.M., Lin, S.F.: The protein tyrosine kinase family of the human genome. Oncogene **19**(49) (2000) 5548–5557
34. Hanks, S.K.: Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol **4**(5) (2003) 111
35. Cheek, S., Zhang, H., Grishin, N.V.: Sequence and structure classification of kinases. J Mol Biol **320**(4) (2002) 855–881
36. Shiu, S.H., Li, W.H.: Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. Mol Biol Evol **21**(5) (2004) 828–840
37. Iwabe, N., Miyata, T.: Kinesin-related genes from diplomonad, sponge, amphioxus, and cyclostomes: divergence pattern of kinesin family and evolution of giardial membrane-bounded organella. Mol Biol Evol **19**(9) (2002) 1524–1533
38. Lawrence, C.J., Dawe, R.K., Christie, K.R., Cleveland, D.W., Dawson, S.C., Endow, S.A., Goldstein, L.S., Goodson, H.V., Hirokawa, N., Howard, J., et. al.: A standardized kinesin nomenclature. J Cell Biol **67**(1) (2004) 19–22
39. Miki, H., Setou, M., Hirokawa, N.: Kinesin superfamily proteins (kifs) in the mouse transcriptome. Genome Res **13**(6B) (2003) 1455–1465
40. Welch, A.Y., Kasahara, M., Spain, L.M.: Identification of the mouse killer immunoglobulin-like receptor-like (Kirl) gene family mapping to chromosome X. Immunogenetics **54**(11) (2003) 782–790
41. Belkin, D., Torkar, M., Chang, C., Barten, R., Tolaini, M., Haude, A., Allen, R., Wilson, M.J., Kioussis, D., Trowsdale, J.: Killer cell Ig-like receptor and leukocyte Ig-like receptor transgenic mice exhibit tissue- and cell-specific transgene expression. J Immunol **171**(6) (2003) 3056–63
42. Engel, J.: Laminins and other strange proteins. Biochemistry **31**(44) (1992) 10643–10651
43. Hutter, H., Vogel, B.E., Plenefisch, J.D., Norris, C.R., Proenca, R.B., Spieth, J., Guo, C., Mastwal, S., Zhu, X., Scheel, J., Hedgecock, E.M.: Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. Science **287**(5455) (2000) 989–994
44. Richards, T.A., Cavalier-Smith, T.: Myosin domain evolution and the primary divergence of eukaryotes. Nature **436**(7054) (2005) 1113–1118
45. Goodson, H.V., Dawson, S.C.: Multiplying myosins. Proc Natl Acad Sci U S A **103**(10) (2006) 3498–3499 Comment.
46. Foth, B.J., Goedecke, M.C., Soldati, D.: New insights into myosin evolution and classification. Proc Natl Acad Sci U S A **103**(10) (2006) 3681–3686
47. Maine, E.M., Lissemore, J.L., Starmer, W.T.: A phylogenetic analysis of vertebrate and invertebrate notch-related genes. Mol Phylogenet Evol **4**(2) (1995) 139–149
48. Westin, J., Lardelli, M.: Three novel notch genes in zebrafish: implications for vertebrate notch gene evolution and function. Dev Genes Evol **207**(?) (1997) 51–63
49. Kortschak, R.D., Tamme, R., Lardelli, M.: Evolutionary analysis of vertebrate notch genes. Dev Genes Evol **211**(7) (2001) 350–354

50. Degerman, E., Belfrage, P., Manganiello, V.: Structure, localization, and regulation of cGMP-inhibited phosphodiesterase (PDE3). J Biol Chem **272**(11) (1997) 6823–6
51. Raper, J.: Semaphorins and their receptors in vertebrates and invertebrates. Curr Opin Neurobiol **10**(1) (2000) 88–94
52. Yazdani, U., Terman, J.R.: The semaphorins. Genome Biol **7**(3) (2006) 211
53. Locksley, R.M., Killeen, N., Lenardo, M.J.: The tnf and tnf receptor superfamilies: integrating mammalian biology. Cell **104**(4) (2001) 487–501
54. MacEwan, D.J.: TNF ligands and receptors–a matter of life and death. Br. J. Pharmacol. **135**(4) (2002) 855–875
55. Inoue, J., Ishida, T., Tsukamoto, N., Kobayashi, N., Naito, A., Azuma, S., Yamamoto, T.: Tumor necrosis factor receptor-associated factor (TRAF) family: adapter proteins that mediate cytokine signaling. Exp. Cell Res. **254**(1) (2000) 14–24
56. Wing, S.S.: Deubiquitinating enzymes–the importance of driving in reverse along the ubiquitin-proteasome pathway. Int J Biochem Cell Biol **35**(5) (2003) 590–605
57. Kim, J.H., Park, K.C., Chung, S.S., Bang, O., Chung, C.H.: Deubiquitinating enzymes as cellular regulators. J Biochem (Tokyo) **134**(1) (2003) 9–18
58. DeLong, E.R., DeLong, D.M.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44** (1988) 837–845