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.

1 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.

1.1 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 non-homologous recombination, transposable elements and retrotransposition [3,4,5,6,7,8,9,10]. About two thirds of proteins in prokaryotes and 80% of proteins in eukaryotes are multi-domain proteins [11].

Domain databases [12,13,14,15,16,17] 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