Sequence analysis

Automated Improvement of Domain ANnotations using context analysis of domain arrangements (AIDAN)

Francois Beaussart, January Weiner 3rd and Erich Bornberg-Bauer*

Evolutionary Bioinformatics, Institute for Evolution and Biodiversity, Westfälische Wilhelms University, Schlossplatz 4, D48149 Münster, Germany

Received on December 8, 2006; revised on April 18, 2007; accepted on April 27, 2007

Advance Access publication May 5, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Since protein domains are the units of evolution, databases of domain signatures such as ProDom or Pfam enable both a sensitive and selective sequence analysis. However, manually curated databases have a low coverage and automatically generated ones often miss relationships which have not yet been discovered between domains or cannot display similarities between domains which have drifted apart.

Methods: We present a tool which makes use of the fact that overall domain arrangements are often conserved. AIDAN (Automated Improvement of Domain ANnotations) identifies potential annotation artifacts and domains which have drifted apart. The underlying database supplements ProDom and is interfaced by a graphical tool allowing the localization of single domain deletions or annotations which have been falsely made by the automated procedure.

Availability: http://www.uni-muenster.de/Evolution/ebb/Services/AIDAN

Contact: ebb@uni-muenster.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Proteins are composed of domains and these are units of evolution as their integrity at the sequence level remains conserved by selection for functionality. Databases of domain signatures take advantage of this fact and allow a sensitive and selective sequence analysis based on a precompiled set of either HMMs/profiles (Bateman et al., 2002; Bru et al., 2005; Letunic et al., 2004) or regular expression (Attwood et al., 2003; Sigrist et al., 2002) (see also Mulder et al., 2005). However, protein evolution is characterised by a high degree of promiscuity as domains appear in many different combinations (Apic et al., 2003; Bornberg-Bauer et al., 2006; Vogel et al., 2005). Consequently, simple sequence comparisons can fail. For example when two proteins are circularly permutated their domain arrangements could be ABCD and CDAB (where every letter stands for a domain ID) and thus both proteins are represented as a string of four such characters (Weiner et al., 2005). In a series of recent papers it was suggested that protein domain arrangements arise not so much via recombination and exon-shuffling but primarily via gene fusion and loss of domains, predominantly at the ends (Bjorklund et al., 2005; Pasek et al., 2005; Weiner et al., 2006). The process of domain loss is characterized by accumulation of mutations which become more frequent (i.e. are less selected against) once a mutation renders a domain non-functional without decreasing the fitness of the organism. Consequently, and in particular for automatically annotated domain sets such as ProDom, several IDs may exist for domains which have just one ID in another, manually curated database (e.g. PFAM). The reason may be that they are structurally almost identical and have the same function but have drifted apart into smaller sub-clusters in sequence space (Finn et al., 2006). This means that many database annotations may be improved since they miss some relationships between domains or unannotated sequences. While some tools exist to look for domain combinations (Geer et al., 2002; Krishnadev et al., 2005), none of these makes use of the fact that domain combinations tend to be evolutionarily maintained and thus similarities which are overlooked can be tracked using context analysis (Coin et al., 2003; Huynen et al., 2000; Pellegrini et al., 1999). For example, if domains X and Y almost always occur in the context ABC and AXC, respectively, and have a high sequence similarity, it is reasonable to assume that they are in fact functionally identical. Indeed, the distribution of E-values obtained from an all-against-all SSEARCH (Pearson et al., 1988) of consensus sequences of ProDom families shows (i) a significant proportion of similarities below $10^{-4}$ and (ii) a strong over-representation of similarities below this threshold for all pairs of domains which match the same PFAM domain (Fig. 1). We exploit this redundancy and, in combination with context analysis, identify new annotations to improve the coverage of an automatically obtained database as explained in the following.

2 METHODS AND RESULTS

We first generated protein clusters as follows: all proteins annotated in the ProDom file version 2005.1 are used and proteins with domain

*To whom correspondence should be addressed.
arrangements containing more than six domains are selected as query domain arrangements for the cluster creation. For each of those queries all the proteins with domain arrangements containing at least 80% of the domain IDs in common with the query are grouped in the same cluster. Multiple occurrences of one domain are considered as separate domains. A multiple alignment of the domain arrangements is then performed for every newly created protein cluster (Feng, 1987) by (i) aligning two domain arrangements, (ii) creating a consensus of these two, (iii) aligning the consensus to a third domain string, etc. The first alignment is done using an adaptation of the Needleman–Wunsch-like global alignment algorithm we used for identification of circular permutations (Weiner et al., 2006). Matches between characters of domain IDs are given a positive score of 5, a mismatch is penalized with $-5$, a gap with $-3$ and gap extension with $-1$. Only clusters with at least 10 proteins were kept providing this way a data set of 3499 clusters containing 141 932 different proteins and 79 529 different ProDom domains. The cluster creation thresholds (at least six domains for the domain arrangement queries and at least 10 proteins per cluster) were chosen in order to provide a first data set to investigate annotation irregularities. A second, larger data set has been created using at least three domains for the domain arrangement queries and at least five proteins per cluster. Almost all (97%) of the domain events included in the first, smaller data set are also included in the second, larger set. The characteristics of those data sets as well as the results of the investigations performed on them are given in Supplementary Material (Table 1).

Clusters are automatically inspected to detect irregularities as follows: a camouflage is found when two differently annotated domains occur at the same position (column) in two proteins of a cluster and if their sequences present a high similarity corresponding to an E-value of $10^{-4}$ and less. Almost all the domain pairs defined as being camouflage domains have both domains belonging to the same Pfam clan (Supplementary Material Table 2). An erosion is found if a domain and an unannotated region are found at the same position in two proteins of a cluster and if their sequences are similar (according to the E-value threshold $10^{-4}$). We chose this rather conservative value because the idea is to suggest where domains may be identical or may have been overlooked without introducing new errors. We estimate that around 64.76% of all domains having a high similarity with other domains are detected (Fig. 1, yellow histogram). On the other hand, 99.95% of domains that correspond to different PfamA domains show an $E$-value above this threshold (Fig. 1, dark blue histogram). Therefore, we estimate that with this threshold, in $<0.05\%$ of all cases new errors are introduced.

Thus, based on amino acid similarity of protein fragments, we distinguish two groups of events. If the amino acid sequences are similar ($E < 10^{-4}$), we annotate these events as annotation artifacts (erosions or camouflage, respectively). If no similarity is detected, we annotate the events as real evolutionary events (physical domain deletions and substitutions). The remaining fragments of proteins that are not annotated after this procedure and are not similar to other domains within the cluster are called shadow domains. The above procedure was followed for both (small and large) cluster sets.

All domain investigations are annotated and stored in a database and can be downloaded in flat file format from the AIDAN website. This web site allows browsing clusters (Fig. 2), a domain-centric view by querying domains with a potentially flawed original annotation, a sequence-centric perspective by entering a sequence and looking for a possible corresponding cluster as well as browsing a list (IDs with annotations) of each category of annotation error and real evolutionary event. Inferred 'correct' annotations are suggested in case of 'eroded' or 'camouflage' domains. It is possible to download all data from AIDAN in the same format (XDOM) as ProDom but with additional information suggesting corrected IDs or corresponding similarities.

In conclusion, we have devised a tool which suggests corrected domain IDs using context analysis, when the sequence identity is too weak to correctly identify homologies in an automated procedure such as ProDom. The tool presented here is primarily meant to make suggestions about domains which are potentially identical or have not been annotated automatically. It will not allow a full functional analysis of sequences by itself but will be a valuable add-on to existing tools for sequence analysis. It may also be useful in speeding up the process of transferring information from ProDom to PFAM.

Conflict of Interest: none declared.
Fig. 2. Screenshots of the AIDAN database: cluster made from the human potassium voltage-gated channel subfamily A member 1 protein (SwissProt access number Q09470). Left: List of proteins belonging to a cluster. Top right: alignment of the protein domain arrangements. Colored boxes represent domains (all domains are depicted at equal length), colored bars represent erosion events and black bars shadow domains (unannotated sequences), empty spaces represent physical deletions. Bottom (‘Legend’): legend of the domain arrangement alignment. Bottom right: dotplots of the sequences involved in domain events. The left dot plot refers to the sequence of the domain J and the sequence of the upper colored bar in the first column of the alignment (erosion). The right dotplot refers to the sequences of the two domains J and K in the first column of the alignment (camouflage). As sequence similarity is displayed by a straight line, the ‘blobs’ on the dotplots indicate repetitions of very short identical sequences. Here, these sequence fragments are stretches of short glycine repeats. Colour version of this figure is available as Supplementary material online.