Identification of function-associated loop motifs and application to protein function prediction

Jordi Espadaler1,2, Enrique Querol2, Francesc X. Aviles2 and Baldo Oliva1,*

1Grup de Bioinformàtica Estructural (GRIB-IMIM), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain and 2Institut de Biotecnologia i Biomedicina and Departament de Bioquímica, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Catalonia, Spain

Received on March 30, 2006; revised on July 4, 2006; accepted on July 6, 2006
Advance Access publication July 26, 2006
Associate Editor: Anna Tramontano

1 INTRODUCTION

The ultimate goal of functional genomics is to determine the function of genes and proteins as a means to better understand life, health and illness. Currently, most approaches to protein function prediction rely on searching sequence databases for homologous sequences with prior annotation. However, the function for one protein cannot be inferred from another when similarity is <40% sequence identity (Todd et al., 2001). Moreover, studies on enzyme proteins have shown that the precise function diverges below identities of 60% (Tian and Skolnick, 2003).

On the other hand, recent improvements in structural biology have greatly increased the number of protein three-dimensional (3D) structures (Deshpande et al., 2005). Structural Genomics projects (Burley, 2000) aim to solve the structures for representatives of all protein folds as a means to understanding their function (Shapiro and Harris, 2000; Todd et al., 2005). However, the number of structures in the Protein Data Bank with unassigned function is increasing exponentially (Pazos and Sternberg, 2004). Therefore, methods to annotate function through structure are now of growing importance. Proteins of known structure but of unknown function are typically compared with databases of other structures to discover functional relationships. However, it has been shown that 10% of remote homologues in a SCOP superfamily have quite different functions (Russell et al., 1998). An alternative strategy is to obtain functional clues by detecting local structural patterns associated with a particular function, which can be common to proteins with different folds. A number of approaches use 3D patterns known to be associated to particular functions to attempt to assign a function to newly determined structures (Ausiello et al., 2005; Di Genaro et al., 2001; Pazos and Sternberg, 2004; Stark and Russell, 2003). However, most of these methods can only be applied to 3D patterns whose function has been already described in the scientific literature.

Protein loops play important roles in protein function, stability and folding (Fetrow, 1995). Functional differences between the members of the same protein family are usually a consequence of structural differences on the protein surface. In a given fold, structural variability is a result of substitutions, insertions and deletions of residues between members of the family. Such changes frequently correspond to loop regions that connect elements of secondary structure in the protein fold, and therefore, loops often determine the functional specificity of a given protein framework (Fiser et al., 2000). In a recent work on clustering of octapeptides by geometric invariants, functional clusters were found to be typically made up of peptides in the loop regions (Tendulkar et al., 2004).

There are many examples in the scientific literature that relate loops to protein function: (1) recognition sites, such as CDRs, (Kim et al., 1999); (2) protein–protein interactions, such as signalling cascades (Bernstein et al., 2004; Zomot and Kanner, 2003), dimerization (Feng et al., 2003; Fritz-Wolf et al., 1996) and protease inhibitors (Jackson and Russel, 2000); (3) ligand binding, such as the P-loop (Saraste et al., 1990), EF-hand (Kawahashi and Krebsinger, 1995), NAD(P)-binding loops (Wierenga et al., 1986) and glycine-rich-loop (Schenk and Snaar-Jagalska, 1999); (4) DNA-binding (Tainer et al., 1995); (5) forming enzyme active sites, such as Ser-Thr kinases (Johnson et al., 1998) and serine proteases (Wlodawer et al., 1989); (6) ‘triggering’ loops whose conformational change is required for the catalytic process of enzymes such as β1,4-Galactosyltransferase (Gunsanakan and Nussinov, 2004), class II FBPA (Zhiby et al., 2002), triose-phosphate isomerases (Joseph et al., 1990) and protein kinases (Adams 2003; Johnson et al., 1997).
et al., 1996); (7) driving membrane insertion of pore-forming bacterial proteins, such as the anthrax protective antigen (Benson et al., 1998) and aerolysin (Iacovache et al., 2006).

Owing to their flexibility and non-periodic nature, loops long escaped structural classifications. As more structures were solved, structurally conserved patterns were found, and there have been many attempts to classify loops according to various conserved features (Burke et al., 2000; Efimov, 1991; Espadaler et al., 2004; Kwasigroch et al., 1996; Li et al., 1999; Oliva et al., 1997; Rufino et al., 1996). The program ArchType (Oliva, et al., 1997) defines a loop motif as one loop plus its bracing secondary structures. Loop motifs are clustered according to loop length, the $\Phi/\Psi$ conformation of the loop residues and the type and geometry of the bracing secondary structures. Clusters of loop motifs are further grouped into subclasses and classes, as reported in the ArchDB database (Espadaler et al., 2004), where each subclass corresponds to a particular structural pattern.

The complexity of protein function makes the establishment of any functional classification problematic (Shrager, 2003). Today, an extensively used functional classification is derived from the Gene Ontology (GO) project, which provides a controlled vocabulary to describe the attributes of a protein in any organism (Ashburner et al., 2000). Moreover, GO terms allow establishing a functional description that progresses from general functions to more specific ones.

Sequence signatures such as those found in the PROSITE, PRINTS and BLOCKS databases (Mulder et al., 2005) are a well-characterized feature of proteins. PROSITE signatures are described as pattern or profiles. Patterns are defined as sequence-based qualitative descriptors, which adopt the form of a regular expression, while profiles correspond to quantitative descriptors adopting the form of a position-specific scoring matrix. PROSITE patterns are considered to include those residues that play important roles in the function of the protein or the formation of the core of its structure, and are helpful at identifying proteins of a given family (Hulo et al., 2006). Function prediction based on sequence similarity has been shown to improve significantly when specific knowledge of residues involved in protein function is used (George et al., 2005).

In this paper, first we apply the ArchDB loop classification scheme to the SCOP95 set of protein structures (Andreeva et al., 2004). Second, we use the PROSITE database of sequence signatures as a gold standard to evaluate three methods to identify putative function-related loop motifs. Third, we assess the ability of sequence-patterns derived from the function-related motifs to discriminate between similar proteins performing and not performing the same function. Finally, the implications of our results are discussed in Section 4.

2 METHODS

Loop motif subclasses were obtained for all X-ray structures from the ASTRAL 95 set of SCOP v.1.67 with resolution better than 3 Å, as in Espadaler et al. (2004). This resulted in 4063 subclasses. Loops with missing residues and/or main chain atoms (including C$_\beta$, except for Gly) were not included.

GO terms were collected for all protein chains in the ASTRAL 95. In this study we focused on terms describing molecular function. The mapping between GO and PDB was taken from the GO Annotation project (Camon et al., 2003). Each loop motif was assigned all the GO terms referring to the protein chain from which it was extracted, as well as their parent terms. However, GO terms belonging to the first two levels of the GO hierarchy or occurring in >10% of the classified loop motifs were excluded from this study to avoid broad functional descriptors.

2.1 Measuring the association between GO terms and loop subclasses

We have compared three different methods to score the degree of association between a GO term $i$ and a subclass of $n$ loops containing $k$ loops annotated with term $i$. The most straightforward approach is to use the raw frequency of term $i$ within the subclass:

$$F = \frac{k}{n}. \quad (1)$$

The rationale behind this approach is that the more often proteins displaying a particular structural pattern share an annotation term, the more likely the structural pattern will be related to the annotation term.

Another method is based on the logarithm of the odds ratio between the observed and the expected frequency of term $i$ in a subclass of $n$ loops, given that $K$ loops display term $i$ in the classification containing $N$ loops, and is calculated as follows:

$$\text{Log-odd} = \log \left( \frac{\frac{K}{N}}{\frac{n}{N}} \right) = \log \left( \frac{k}{n} \right) \quad (2)$$

Finally, a more elaborate method is derived from information theory. Intuitively, the mutual information score measures the information of term $i$ that is shared by a given subclass of $n$ loops:

$$\text{MI} = \left( \frac{k}{N} \right) \log \left( \frac{\frac{K}{N}}{\frac{n}{N}} \right) = \frac{k}{N} \log \left( \frac{k}{n} \right) \quad (3)$$

The above methods were applied to the loop classification using a $k_{min} = 2$ (this is the minimal number of loops displaying term $i$ in a subclass).

2.2 Estimation of statistical significance of association

The significance of the scoring for the aforementioned methods was calculated by comparison with the distribution of 500 random classifications. Briefly, loops were randomly shuffled among subclasses of the same loop type, as defined by bracing secondary structure, and the frequency of occurrence of an association score for all three methods was recorded. Association scores were calculated for all pairs of subclasses and GO terms for all three methods, and the frequency of occurrence of each degree of association was recorded for each score. This process was repeated 500 times, and the $p$-value of obtaining an association score equal or better than a certain value was set as the average of the frequencies observed in the 500 random classifications.

2.3 Obtaining motif-derived sequence patterns

Structure-based alignments were obtained from the loop classification for subclasses associated to a GO term $i$ with a $p$-value < 0.01. If an alignment was associated to more than one GO term of the same branch the most accurate term describing the function was used for the association. Alignments were filtered by removing all loops not annotated with term $i$. The resulting seed alignments were further expanded by including the aligning sequence regions from all those Swiss-Prot homologues (Bairoch et al., 2005) in HSSP (Dodge et al., 1998) also annotated with GO term $i$. Sequences from HSSP displaying gaps in the aligning regions were removed, and alignments containing <10 sequences were not considered for further study. This resulted in a set of 375 multiple sequence alignments, each one associated to a GO term. Finally, sequence patterns with the form of regular expressions were calculated for each of these alignments.
2.4 Identifying functional homologues

To validate the ability of the patterns to correctly discriminate between homologous proteins performing the same function we used proteins from the well-annotated Swiss-Prot database (Bairoch et al., 2005). Proteins in Swiss-Prot were annotated using GO term assignments from the GO Annotation project (Canon et al., 2003), as well as all their parent terms.

We used each alignment associated to a GO term and containing >10 sequences to find homologues in Swiss-Prot. The procedure was performed by choosing one sequence of the alignment at random. These sequences were used as queries to retrieve putative homologues from Swiss-Prot using BLAST (Altschul et al., 1990). We used a total of 282 sequences (out of 375 alignments), since each alignment is associated to one GO term but sequences are often annotated with multiple terms and may appear in more than one alignment. Sequences in Swiss-Prot were used to build the patterns that were retrieved by BLAST were not considered. For putative homologues found by BLAST that also matched at least one subclass-derived sequence pattern we assigned the GO terms associated to such motif, and these GO terms were compared with the annotation in the Swiss-Prot database. This procedure is referred hereby as ArchFun. To compare sequence patterns with sequence similarity alone, all putative homologues found by BLAST were assigned the GO terms associated to the alignments where the query sequence was found, regardless of whether they also matched the corresponding patterns or not. Accuracy was calculated for both methods as the number of correct pairs (correct hits) found at a given identity threshold divided by the total number of pairs found at the same identity threshold.

3 RESULTS

A detailed benchmark is difficult to achieve since there is no large reliable resource of function-associated and non-function associated 3D-motifs. Therefore, we used a set of subclasses containing loops matching known protein signatures from PROSITE as a gold standard. A loop motif was considered to match a PROSITE signature whenever the signature and the loop plus its bracing secondary structures overlapped by >50% of residues. Moreover, we forced at least 50% of the loops of a subclass to match the same PROSITE signature in order to reduce false positive matches. A total of 67 PROSITE signatures were found using these criteria. PROSITE signatures matching more than one subclass can be considered as structural variants of the same functional motif. Therefore, our gold standard set of 3D-motifs consists of the 73 subclasses matching one of the aforementioned 67 PROSITE signatures.

We have evaluated three different methods to score the degree of association between a loop subclass and the GO function annotations of the proteins whose loops belong to that subclass. Distributions of scores from random classifications were obtained for each scoring method. These distributions were compared with the ones corresponding to the scores of the 3D-motifs from our gold standard set (Fig. 1). The mutual information score provides the best discrimination between motifs from the gold standard set and random ones (Fig. 1a). It is worth noting that at least one GO term is shared by 100% of the loops in the subclass (frequency = 1) in 30 out of the 73 subclasses from the gold standard set (Fig. 1c).

Random score distributions allow the calculation of a p-value associated to each score for each scoring method. All three scoring methods were further compared on the basis of their ability to identify the largest number of subclasses from the gold standard set at a given p-value threshold (Fig. 1d). As expected, the mutual information method yields the best results. At a p-value threshold of 0.05, all three methods yield similar results (~65 out of 73 subclasses are identified). At lower p-values, mutual information performs better than the other two methods, while the results for larger p-values are meaningless. For instance, at a p-value of 0.01 the frequency method finds 49 subclasses and the log-odds method finds 50, all of which are also found among the 62 subclasses identified by the mutual information method. Therefore, we can conclude that the method based on the mutual information score clearly outperforms the other two.

3.1 Function-related motifs found in ArchDB

We used the mutual information score to identify subclasses in our classification that corresponds to putative function-related 3D-motifs. At a p-value threshold of 0.01 we found 682 loop subclasses associated to 852 GO terms. Clearly, some subclasses were associated to more than one GO term, owing to the particular structure of the GO annotation vocabulary. Occasionally, multiple GO terms may appear associated to the same proteins within a subclass (e.g. ‘metalloendopeptidase activity’ and ‘zinc ion binding’). Also, GO terms describing the same function at different levels of precision may display a significant association score to the same subclass. Out of the 682 subclasses, 75 contained at least one loop matching PROSITE signatures, of which 62 were also found in the gold standard set. As above, a loop motif was considered to match a PROSITE signature whenever the PROSITE signature and the loop plus its bracing secondary structures overlapped by >50% of residues. A list of examples of loop-PROSITE associations that have been manually checked for their being related to proteins performing the same function can be found in the Supplementary Materials. Similarly, the number of subclasses matching PROSITES increases to 111 when the required residue overlap is reduced to 10%. In conclusion, between 80 and 90% of the putative function-related motifs identified by our method are new, when compared with PROSITE.

3.2 An example: laccase activity-related motifs

Laccases are enzymes found in fungi and plants, which oxidize many different types of phenols and diamines. For instance, laccases are involved in lignin degradation and detoxification of lignin-derived products. Laccases belong to the multicopper oxidase family of enzymes. All multicopper oxidases contain three spectroscopically different copper centres. In addition to laccases, multicopper oxidases also include l-ascorbate oxidases (EC 1.10.3.3) and ceruloplasmin (EC 1.16.3.1), a protein found in the serum of mammals and birds that oxidizes a great variety of inorganic and organic substances (Messerschmidt and Huber, 1990).

Two PROSITE patterns have been described that match multicopper oxidases: PS00079 and PS00080. These two patterns match 23 out of the 25 laccases found in Swiss-Prot. However, these patterns cannot distinguish between laccases and other multicopper oxidases. When applied to laccases, our method finds two β–β-link motifs that are associated to GO term ‘laccase activity’, none of which is found in the PROSITE, PRINTS or BLOCKS databases. Mapping the two motifs onto PDB structure 1KYA (a laccase from Trametes versicolor) shows that both contain residues located at <4 Å from two bound ligands: N-acetyl-D-glucosamine and 2,5-dimethylaniline (Fig. 2).

PROSITE-like sequence patterns were derived from these two motifs as described in Section 2. These patterns match 22 and 23 out of the 25 laccases found in Swiss-Prot, respectively.
PROSITE patterns PS00079 and PS00080 match most laccases, in Swiss-Prot, as well as most L-ascorbate oxidases and ceruloplasmins. However, our patterns are highly specific for laccases, as they match none of the L-ascorbate oxidases and the ceruloplasmins found in Swiss-Prot and there were not any other matches in Swiss-Prot.

3.3 Function prediction using motif-derived patterns

For the purpose of this work, we have assumed that all hits from the BLAST search would be assigned the same function as the query, in line with the way a researcher would typically transfer functional annotation between proteins. On the other hand, using sequence patterns derived from putative function-related subclasses to filter BLAST results (ArchFun) categorizes BLAST hits into two separated groups (i.e. matching or not the pattern). Moreover, of the 375 multiple sequence alignments (MSAs), only two of them contain exactly the same proteins, yielding identical patterns. The reason is that these MSAs are significantly associated to two terms that belong to two different branches in the GO ontology: ‘dioxygenase’ and ‘iron ion binding’. Therefore, this produces two patterns that are reciprocally matching.

3.3.1 Accuracy of the predictions

The accuracy of a BLAST search decreases as the GO terms used to describe the function become more precise. For instance, at the 60% sequence identity threshold, accuracy for GO terms of level 3 is 97%. At the same identity threshold, accuracy for GO terms of level 5 decreases to 85%. However, when combining sequence similarity and subclass-derived patterns (ArchFun), accuracy is nearly 100% for both levels 3 and 5 (Fig. 3). At low sequence identity thresholds, BLAST reports a much larger number of correct hits than our method,
Loop associated to protein function

but many false too. For instance, at the 40% sequence identity
threshold, accuracy of BLAST drops to 58% for GO terms of
level 3, and 34% for level 5. On the other hand, our method
finds a much lower number of hits but with accuracy >97%
(both for levels 3 and 5).

3.3.2 Coverage of the predictions The benefit of using ArchFun
in addition to BLAST can be easily seen as follows: At 99% accu-
curacy for GO terms of level 3, BLAST finds 179 hits, while
our method finds 220 hits (Fig. 3). BLAST reports 21 hits not found
by our method, while our method reports 62 hits that were not found
by BLAST. Thus, a 30% increase is achieved in the number of
highly reliable function predictions compared with BLAST. The
benefit of using subclass-derived patterns is even larger in the case
of GO terms of level 5. At 98% accuracy, BLAST reports 203 hits
while our method finds 386, of which 207 were not found by
BLAST. Therefore, applying our method yields a 100% increase
compared to using BLAST alone.

4 DISCUSSION
We have evaluated three different measures to quantify the degree
of association between a functional annotation and a 3D-motif
consisting of one loop plus its bracing secondary structures.
The method is fully automated and based on the widely used
GO classification. An important consequence of our work is the
identification of previously unreported function-associated 3D
motifs and their corresponding sequence patterns. Moreover, we
show that patterns derived from these loops can help to improve
the accuracy of similarity-based function assignment.

To benchmark our approach, we first required a set of 3D-motifs
known to be involved in protein function. Since a database of
known function-related and non-function related 3D-motifs does
not exist, we relied on the PROSITE database of sequence signa-
tures. PROSITE signatures include residues known to participate in
protein functions, and have been previously used to evaluate the
ability of an automated method to identify functional motifs
(Lu et al., 2004).

We have evaluated three scoring schemes of increasing complex-
ity to identify putative function-related motifs. The most elabor-
ated score, mutual information, yields the best results: 62 out of 73
PROSITE-matching loop subclasses were identified with a p-value
better than 0.01. Most putative function-related motifs identified
by this method in our loop classification correspond to new
motifs, when compared with a database of known motifs such as
PROSITE. Besides, previous studies identified functional motifs on
the basis of the frequency of a given annotation within a cluster
(Fernandez-Fuentes et al., 2004; Tendulkar et al., 2004). Our results
suggest that the use of mutual information could significantly
improve their results.

Functional motifs identified with this method do not necessarily
 correspond to fold signatures. For example, our method identifies
a hairpin motif significantly associated to the serine-protease
function, which is found in proteins adopting either the 7-bladed
β-propeller or the 8-bladed β-propeller folds. Also, a helix–loop–
helix motif significantly associated to the molybdenum binding
annotation is found in proteins either adopting the FAD-binding
domain fold or the formate dehydrogenase fold.

Our results show that sequence patterns derived from function-
related subclasses can substantially improve the accuracy of function
assignment, especially when putative homologues are distant. If a
standard database search method (i.e. BLAST) finds a distant
homologue annotated with a set of GO terms and the sequences of
both match a loop subclass-derived pattern associated to some of
these GO terms, then chances are ~95% that the two proteins
performs the same functions. This results in an increase of the
number of proteins that can be reliably annotated when compared
with using BLAST alone or motif patterns alone (see Fig. 3 in the
text and S1 in Supplementary Material). This improvement is
more significant when more precise functional descriptors are
evaluated (e.g. for GO terms of level 5 an increase of 100%
is obtained).

GO terms of levels 2 and 3 correspond to broad functional
descriptors. However, most function-related loop motifs found in
this work are related to GO terms of levels 4 and 5, which corre-
spond to more precise functional descriptors. Motifs associated to
GO terms of levels 4 and 5 account ~25 and ~50% of the func-
tional associations identified in our database, respectively. There-
fore, more precise function assignments can be achieved using
patterns derived from function-related loop subclasses as a filter
to correctly discriminate protein performing the same function
among homologues. This finding is not unexpected, since loops
are known to be the most variable part of protein structure, and
therefore are more likely to vary with the function of the protein.
Overall, our results further support the idea that loops play a key

Fig. 3. Comparison of ArchFun and BLAST in function prediction. Accuracy
(circles) and number of correct hits (triangles) for GO terms of level 3 (a) and
5 (b). Solid circles and triangles correspond to ArchFun; open circles and
triangles correspond to BLAST.
role at determining the specific function of a given protein (Fiser et al., 2000).

A key advantage of our subclass-derived patterns is that they can be applied to identify proteins performing the same function detected by any database search tool. This feature is particularly important for high-throughput genome sequencing projects, where no hand annotation can be provided in time before release, as well as for large databases such as GenBank (Pruitt et al., 2005), where annotation depends on each depositor, or UniProt (Bairoch et al., 2005). Errors in database annotation are a well-known problem of the post-genome era. Moreover, since databases are interconnect-

ded and function annotation methods rely on data extracted from previously annotated proteins, errors tend to propagate (Karp, 1998). Therefore, methods identifying proteins performing the same function with high accuracy can be useful to build core sets of proteins whose function annotations are highly reliable, which can latter be used to annotate other proteins.

Structural biology is an increasingly popular tool for obtaining functional clues for a protein. However, there is a large gap between the number of known sequences and the number of proteins with solved structures. There is also a large gap between the number of solved structures and the number of structures with annotated 3D motifs. Here we show that function-related 3D motifs can be automatically identified in known structures. Moreover, this structural information can be used to obtain sequence patterns that can be further applied to accurately predict the function of sequences without known structure. A further advantage of our method is that it is able to find a function-related pattern regardless of the fold of the proteins performing such function.

Future developments of the method include considering ligand information as well as obtaining specific p-value distributions for each subclass size to improve the identification of function-related loops. Also, we plan to include gap information from the HSSP alignments into pattern calculation. BLAST searches in the Swiss-Prot subset of UniProt could be further used to identify more homologues of the proteins with function-related loops. In this way, the Prot subset of UniProt could be further used to identify more homologues per-

matically identified in known structures. A further advantage of our method is that it is possible to find a function-related pattern regardless of the fold of the proteins performing such function.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for the useful comments. B.O. acknowledges grant from Spanish Ministerio de Educacion y Ciencia (MEC, BIO2005-00533). J.E. was supported by predoctoral fellowship from the Generalitat de Catalunya. E.Q acknowledges grant from the Spanish Ministerio de Educacion y Ciencia (BFU2004-06377). F.X.A acknowledges support from grants BIO2004-05879, GEN2003-20642 (Ministerio de Educacion y Ciencia, MEC, Spain) and by Centre de Referencia en Biotecnologia (Generalitat de Catalunya, Spain).

Conflict of Interest: none declared.

REFERENCES


Bernstein, L.S. et al. (2004) RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling. J. Biol. Chem.


