BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 21 2005, pages 4060–4066
doi:10.1093/bioinformatics/bti614

Databases and ontologies

ProRule: a new database containing functional and structural information on PROSITE profiles

Christian J. A. Sigrist*, Edouard De Castro, Petra S. Langendijk-Genevaux, Virginie Le Saux, Amos Bairoch[1] and Nicolas Hulo

Swiss Institute of Bioinformatics (SIB) 1 rue Michel Servet CH-1211 Geneva 4 Switzerland and [1]Structural Biology and Bioinformatics Department, University of Geneva, Geneva, Switzerland

ABSTRACT

Motivation: Increase the discriminatory power of PROSITE profiles to facilitate function determination and provide biologically relevant information about domains detected by profiles for the annotation of proteins.

Summary: We have created a new database, ProRule, which contains additional information about PROSITE profiles. ProRule contains notably the position of structurally and/or functionally critical amino acids, as well as the condition they must fulfill to play their biological role. These supplementary data should help function determination and annotation of the UniProt Swiss-Prot knowledgebase. ProRule also contains information about the domain detected by the profile in the Swiss-Prot line format. Hence, ProRule can be used to make Swiss-Prot annotation more homogeneous and consistent. The format of ProRule can be extended to provide information about combination of domains.

Availability: ProRule can be accessed through ScanProsite at http://www.expasy.org/tools/scanprosite. A file containing the rules will be made available under the PROSITE copyright conditions on our ftp site (ftp://www.expasy.org/databases/prosite/) by the next PROSITE release.

Contact: christian.sigrist@isb-sib.ch

INTRODUCTION

Motif descriptor databases contain information derived from alignments of multiple homologous sequences, giving them the notable advantage of identifying distant relationships between sequences that would have passed unnoticed, based solely on pairwise sequence alignment. Due to an exponential increase in the number of sequences coming from genome sequencing projects, such databases play a critical role in the analysis of the resulting protein sequences. In the absence of any experimental characterization of most of these proteins, motif descriptor databases will often be the unique source of information. This makes them essential for functional prediction derived from *in silico* characterization, which needs to be as sensitive and accurate as possible.

PROSITE is one of these motif descriptor databases. It is an annotated collection of biologically meaningful motif descriptors dedicated to the identification of protein families and domains (Hulo *et al*., 2004). The PROSITE database uses two kinds of motif descriptors, each having its own strengths and weaknesses defining its area of optimum application (Sigrist *et al*., 2002).

The first motif descriptors used by PROSITE are patterns or regular expressions in which all but the most significant residue information is discarded. Patterns are qualitative descriptors: either they match or they do not. If there is a mismatch at one of the positions the pattern will not match, even if the mismatch is a conservative, biologically feasible substitution. Therefore good patterns are usually located in short well-conserved regions, such as enzyme catalytic sites, prosthetic group attachment sites (haem, pyridoxal phosphate, biotin, etc.), metal ion binding amino acids, cysteines involved in disulfide bonds and regions involved in binding a molecule. Even though the ideal scope of a regular expression is limited to these particular biological regions, patterns are still very popular as they are easy to formulate and to use.

The second motif descriptors used by PROSITE, the generalized profiles (or weight matrices), are quantitative motif descriptors that consider the overall similarity on the entire length of domains or proteins and not just the most conserved parts of them. They provide numerical weights for each possible match or mismatch between a sequence residue and a profile position. A mismatch at a highly conserved position can thus be accepted provided that the rest of the sequence displays a sufficiently high level of similarity. This gives profiles an enhanced sensitivity when compared with patterns, and makes them able to detect highly divergent domains or families with only few very well-conserved sequence positions.

Despite their obvious advantages, profiles are not always superior to patterns. In fact the two types of descriptors have complementary qualities. While profiles are well suited for the detection of remote similarity spreading over entire domains or proteins, patterns often perform well in the identification of meaningful amino acid residues with a functional or structural role. The combined use of profiles and patterns render functional prediction much more accurate and can avoid some misinterpretation. A mismatch with a pattern directed against residues playing a biologically important role could possibly indicate that the function of the domain or protein is no longer conserved. Hence, while a profile (PS50240) is well designed to detect the structural relationship of the non-enzymatic haptoglobin with the trypsin family of serine proteases (Fig. 1), patterns directed against the proteolytic active site residues of the proteases (PS00134 and PS00135) allow the distinction between the

*To whom correspondence should be addressed.

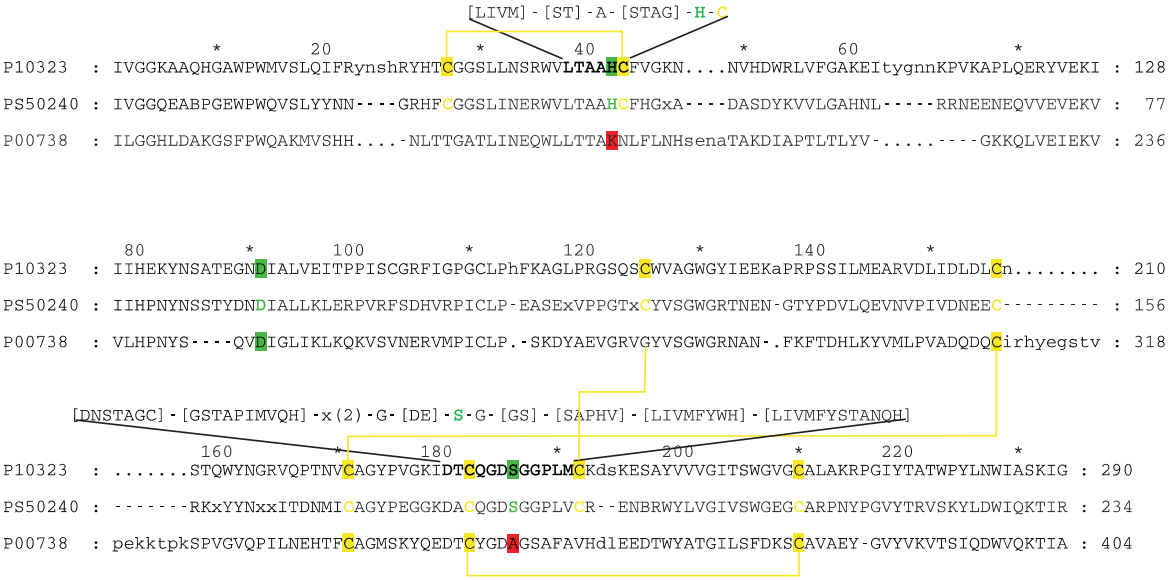

Fig. 1. Sequences/profile alignment between the trypsin-type serine protease domain (PS50240) and the matched region of the catalytically active human acrosin (P10323) and of the non-enzymatic human haptoglobin (P00738). The numbering corresponds to the position in the profile. The patterns TRYPSIN_HIS (PS00134) and TRYPSIN_SER (PS00135) are also shown. The region matched by the patterns in the catalytically active human acrosin is shown in bold. In the profile sequence and in the pattern, residues corresponding to the catalytic triad and cysteines involved in disulfide bonds are shown in bold and colored in green and yellow, respectively. The corresponding residues are shown in colored boxes in the protein sequences. Residues from the catalytic triad are shown in green boxes if they are compatible with the catalytic activity and in red boxes if not. The accession numbers (AC) correspond to records in the PROSITE and Swiss-Prot databases.

two subfamilies. A relevant *in silico* characterization should take both types of information into account. The profile result indicates that the protein of interest contains a structural domain of the serine protease type and the patterns provide additional information about the possible function or lack of function of this domain.

METHODS

The ProRule database

In order to combine the respective advantages of profiles (high sensitivity and full-length coverage of the domain or protein) and patterns (position-specific information), we have created a new database to extract more information from PROSITE profiles to help function prediction and annotation of the Swiss-Prot section (Boeckmann *et al*., 2003) of the UniProt Knowledgebase (Bairoch *et al*., 2005). The ProRule database is a set of manually created rules containing additional biologically meaningful information about domains detected by PROSITE profiles. It is aimed at providing domain-specific information in the Swiss-Prot format using standardized nomenclature and controlled vocabularies. ProRule uses the UniRule format (http://www.expasy.org/sprot/hamap/unirule_manual_short.html), which contains (complete or incomplete) annotation blocks in the form existing in the Swiss-Prot database, to provide template-based feature propagation. The UniRule format is identical for all types of rules created to annotate Swiss-Prot, like the HAMAP (High-quality Automated and Manual Annotation of microbial proteins) family rules (Gattiker *et al*., 2003). While HAMAP rules are family rules designed to annotate bacterial, archaeal as well as plastidial proteins in their entirety, ProRule has been developed mainly for the specific annotation relevant to protein domains in all kingdoms.

Each ProRule entry consists of the following parts:

- The **header section** contains the accession number, the data class, the identifier of the motif or metamotif that triggers the rule, as well as some basic information on the rule and its motif.
- The **annotation section**, where all Swiss-Prot annotations relevant for a ProRule is indicated. The minimal information supplied by ProRule is the name and the predicted boundaries of the domain, but some rules also provide data for description (DE) lines (protein name, EC numbers), keywords (KW), various comments (CC) and feature (FT) lines for active sites, post-translational modification (PTM) sites, binding sites, disulfide bonds and transmembrane regions. The kind of annotation propagated can be general, i.e. always associated with the domain or protein family, or conditional, i.e. depending on the presence of some features. The condition required can be the taxonomic range, the presence of an additional domain or more generally the presence of particular residue(s) in functionally or structurally critical positions. Some residues with a key role can be indispensable for the function or structure of a domain or protein. In the absence of the physico-chemical properties associated with them, the proper reaction cannot take place and the function or structure will be altered. The fact that profiles generally cover domains or proteins over their entire length makes the identification of such biologically relevant residues less obvious than with patterns. However, as domains and proteins might have acquired different functions during evolution as a consequence of the substitution of critical residues, information about the status of these residues is essential to make accurate function predictions. If the properties of a particular amino acid are needed for biological activity, its substitution by any other residue devoid of these features will mean that the function is probably not conserved and the domain or protein is inactive or plays another role. In order to increase the accuracy of biological predictions inferred from hits to PROSITE profiles, the rule contains information about the position of biologically meaningful residues in the profile and the condition required for their role. The corresponding residues can be identified in the matched domain from the alignment between the profile and the protein sequence (Fig. 1). The retrieved residues are then tested according to the condition they must fulfill to play their functional and/or structural role, and

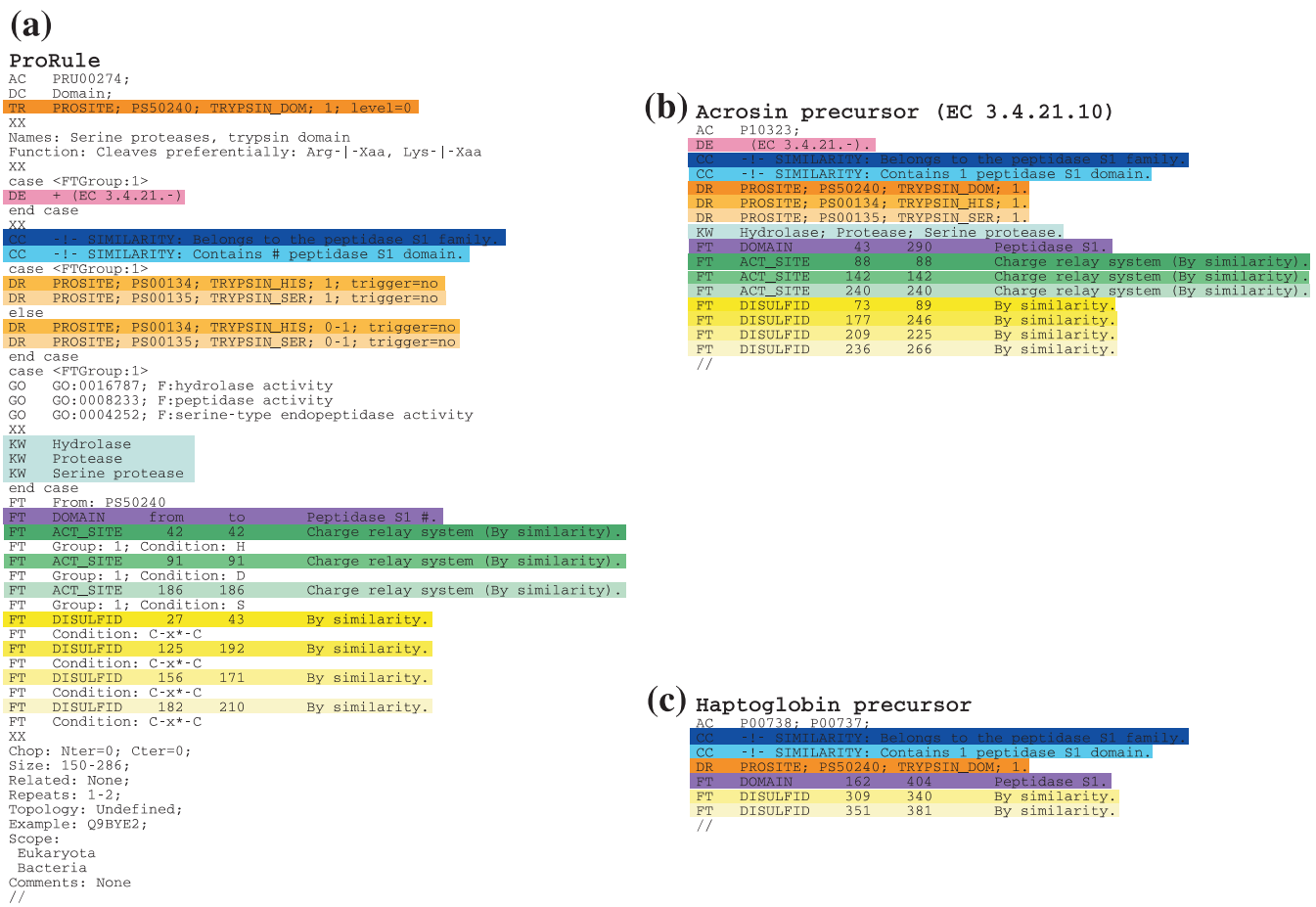

Fig. 2. Annotation of Swiss-Prot entries with a ProRule triggered by a profile. The amount of information to be transferred depends on the conditions contained in the rule. (**a**) The ProRule triggered by the trypsin domain of serine proteases. The information that is transferred from this rule to (**b**) the catalytically active human acrosin (P10323) and (**c**) the non-enzymatic human haptoglobin (P00738).

different biological functions can be proposed in agreement with their status.

- The **computing section** covers general information on a domain, like the observed range of domain size and copies or the taxonomic classes in which it is found, to evaluate the pertinence of the suggested annotation. If the observed values exceed the ranges indicated in the rule a warning will be generated. The computing section also contains one or more examples, whose domain specific annotation has been derived from the rule. This allows to detect changes in the Swiss-Prot annotation format and, as a consequence, the relevant modification of rules to keep them up-to-date.

RESULTS AND DISCUSSION

Annotation in Swiss-Prot format with ProRule

The kind of information that can be provided by ProRule is illustrated using Figure 2, which corresponds to the rule for the trypsin domain of serine protease (PRU00274). This rule belongs to the domain data class (DC), meaning that it is used only to propagate blocks of annotation associated with the trypsin domain and not to annotate completely a protein. It is triggered by at least one hit with the TRYPSIN_DOM profile (PS50240) above the trusted cut-off level (Fig. 2a). The trigger (TR) line indicates which PROSITE profile triggers the rule and how many hits above the trusted cut-off level are required for the rule to be applied. The TR line is also used to automatically generate a database cross-reference (DR) line with the number of hits for the profile. The header section ends with the name(s) of the domain and its function.

The data stored in the annotation section of the rule will be used to produce annotation in the Swiss-Prot format for all the relevant information associated with the presence of a trypsin domain. The following general information is provided and it applies to every sequence matched by the trigger profile:

(1) a comment (CC) line indicating the protein family membership, which in this case is the peptidase S1 family

(2) a feature (FT) line providing the N- and C-termini of the domain. As PROSITE profiles generally cover entire domains, the size is in most cases derived from the beginning (from) and end (to) of the alignment between the sequence and the profile (Figs 1 and 2). In some rare instances the domain is bigger or smaller than the profile requiring that the beginning and end of the domain are indicated by numbers in relation to the profile.

All the other information provided by the rule is conditional, depending on the presence of mandatory residues at some precise positions within the domain. A feature line for a predicted disulfide bond will only be created if cysteine residues are found at the two positions connected by the bond. Hence, the condition expected for the first disulfide is the presence of two cysteines at positions 27 and 43 of the profile. From the alignment of the profile with the protein sequences (Fig. 1), the residues found at these positions are shown to be, respectively, two cysteines in acrosin, for which a disulfide feature line with the positions (73 and 89) in the sequence is created, and threonine and asparagine which do not fulfill the condition in haptoglobin, for which no disulfide feature line is generated (Fig. 2b and c).

As the active site is formed by a catalytic triad, the annotation of the catalytic residues and other information resulting from the presence of the active site requires that the matched sequence contain a histidine, an aspartate and a serine corresponding to positions 42, 91 and 186 of the profile (Fig. 1). These three conditions have to be fulfilled simultaneously (i.e. be grouped) for a group to be validated (Fig. 2a). If only one or two of the three expected residues are found the condition of the group will not be considered as fulfilled and no annotation linked to the catalytic activity will be transferred. In the case of acrosin, the residues at positions 88, 142 and 240 fulfill their respective condition as they correspond to the active site residues. As the three are found together, the group they form will be validated and they will be used to annotate the active site residues (Fig. 2b). The validation of the group formed by the presence of the active site residues will also generate a description (DE) line with the minimal EC number corresponding to the catalytic activity common to all active trypsin domains and the additions of keywords (KW) linked to this activity. GO terms (Harris *et al*., 2004) are also listed in the rule but are currently not used for the annotation. As the active site residues are present, matches with the PROSITE patterns for the histidine and the serine active sites are mandatory and the corresponding database cross-reference lines are created (with the false negative status if they are not detected).

On the other hand, the trypsin domain of haptoglobin only contains one conserved residue of the catalytic triad and hence none of the residues corresponding to the positions of the active site is annotated. As the features of the active site are not conserved, no annotation depending on the presence of the group is transferred. In addition, no database cross-reference lines for the active site patterns are generated as they are not expected to match if the catalytic center is not functional (Fig. 2c).

The computing section contains additional data to ensure that the rule is applied in an appropriate way. The size, repeats and scope lines indicate, respectively, which size range, number of repeats and taxonomic distribution is currently known for the domain. The topology line specifies the subcellular location in which a domain may occur. The chop line indicates the range by which the bounds of a domain may be trimmed in order to annotate successive domains in an exactly consecutive manner. If two rules apply to the same region, the related line indicates which one should be disregarded. The example line contains the accession number of a Swiss-Prot entry targeted by the rule.

ProRule triggered by metamotifs

Some ProRules do not use a single PROSITE profile as a trigger, but rather a specific combination of PROSITE motif descriptors called metamotifs, which are defined by the syntax of mmsearch (Junier *et al*., 2001). Metamotifs allow the definition of arrangements of domains separated by spacers of variable size, as well as the anchoring to the N- and/or C-termini and the exclusion of a feature.

The specific information linked to metamotif-triggered rules results from the simultaneous presence of domains in a precise arrangement and cannot be reduced to the sum of the information contained in the rules triggered by the domain-specific profiles. This approach is useful in providing information for protein families defined by a specific and discriminatory arrangement of domain profiles and for subregions with properties that cannot be uniquely deduced from the individual properties of the domains that constitute them. For example, the actin binding domain (ABD) has probably arisen from duplication of the calponin homology (CH) domain, which can also be found in a single copy in a number of proteins like calponin or the vav proto-oncogene (Stradal *et al*., 1998) (Fig. 3). The identification of a single CH domain should be treated with caution and special care should be taken in attributing actin-binding properties to proteins simply due to the presence of this domain. Therefore, the rule triggered by the CH domain profile will only transfer minimal information about the domain name and boundaries, whereas the rule triggered by the metamotif made of two CH domains separated (=) by a spacer between 7 and 91 residues will allow the annotation of the actin-binding region, the keyword actin-binding, and the DR lines for the linked PROSITE patterns for the actinin-type actin-binding domain (Fig. 3). The condition line with the prefix c? is used to restrict the search for the metamotif to cases where the required features are present in order to save time (Fig. 3b).

ProRule in the annotation pipeline of Swiss-Prot

ProRule is currently used to assist the annotation of the Swiss-Prot knowledgebase (Bairoch *et al*., 2004). As illustrated in the examples above, a great variety of line types from Swiss-Prot entries can benefit from the annotation suggested by ProRule. However, it should be noted that the ProRule database is not used to generate automated annotation of the Swiss-Prot knowledgebase. It rather informs Swiss-Prot annotators that the presence of a domain has been detected and that this occurrence should imply the presence of a predefined type of annotation in order to increase homogeneity and consistency. During the annotation process ProRule is extensively used and the suggested annotation is validated or, if needed, modified. For example, the EC number for human acrosin in the Swiss-Prot knowledgebase has been completed as more information about the catalytic activity is known and the boundary of the serine protease trypsin domain of human haptoglobin has been manually extended by two residues in the Swiss-Prot knowledgebase to reach the C-terminus of the protein (Fig. 2b and c).

ProRule increases the discriminatory power of ScanProsite

The PROSITE profile information supplied by ProRule is not only useful to those annotating Swiss-Prot entries, but it is also of general interest to the PROSITE user in making function and/or structure predictions for a protein of interest. As a domain might have acquired different functions through substitutions of critical residues during evolution, it is necessary to check whether such residues fulfill the condition expected for the functional and/or structural properties associated with the domain. If not, it means that the expected biological property is no longer associated with the domain or that

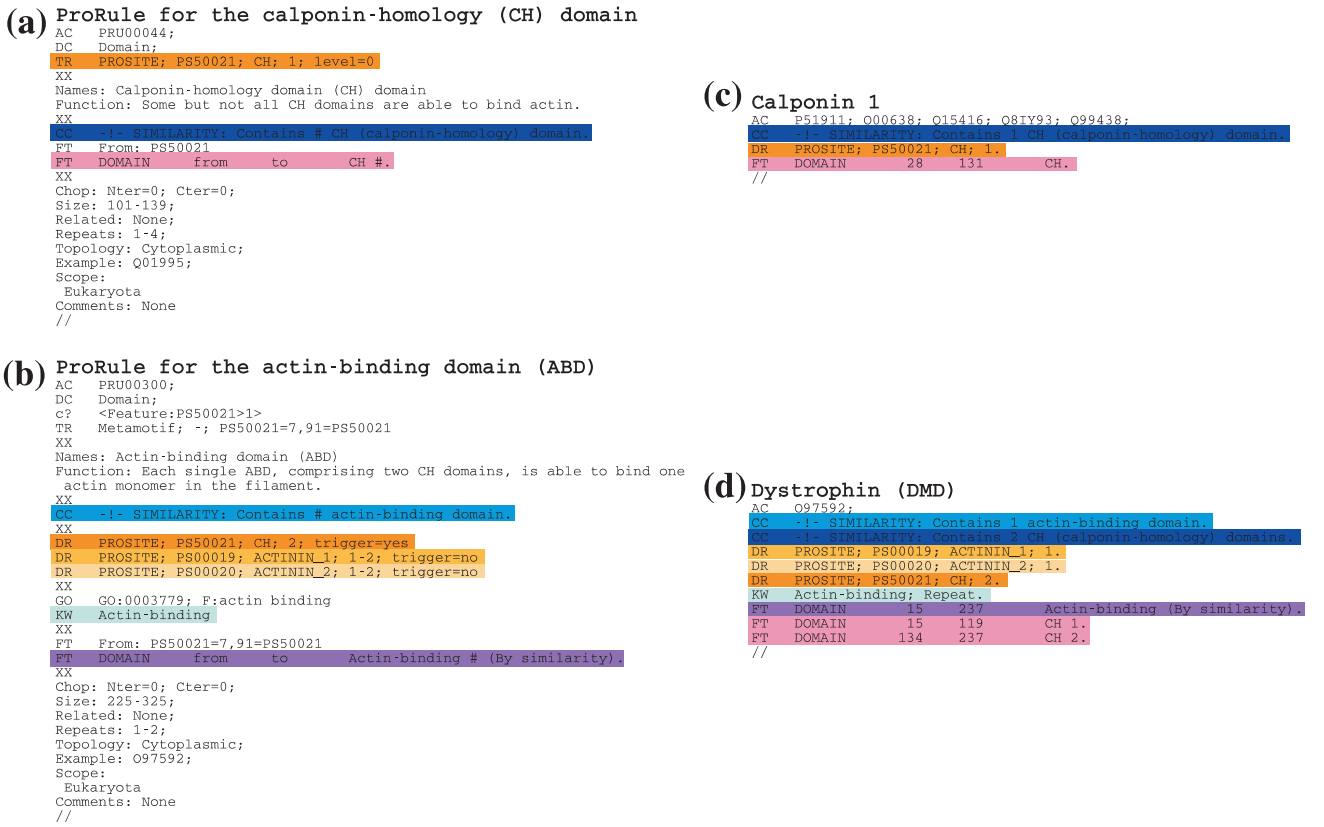

Fig. 3. Metamotifs can be used to provide information specific for particular domain combinations. (**a**) The Prorule triggered by the calponin-homology (CH) domain contains general information about the CH domain that can be used to annotate (**c**) human calponin 1 (P51911) and (**d**) dog dystrophin (O97592). (**b**) The ProRule triggered by the metamotif formed by the combination of two CH domains can be used to provide more specific information to annotate (d) dystrophin but not (c) calponin.

residues at other positions in the sequence compensate for it. Additional information about residues potentially involved in disulfide bonds, catalytic activity, phosphorylation, prosthetic group attachment (heme, pyridoxal-phosphate, biotin, etc.), metal, protein or molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) binding, or forming a transmembrane region could be of great help in predicting whether biological properties associated with a domain are in a given query sequence.

The positional features contained in ProRule have been implemented in the ScanProsite interface (http://www.expasy.org/tools/scanprosite/) to provide the user not only with domain boundaries but also with the position and potential role of critical residues or regions that might be found within the domain (Fig. 4). This additional service should help protein analysis by making functional prediction more accurate through the combination of the sensitivity of overall similarity with the specificity of site specific information.

Current status of ProRule

The current release of PROSITE 19.0, of 26 April 2005, is linked to 433 ProRules among which 427 are based on a PROSITE profile and 6 on a metamotif. The sum of the 427 profile matchlists is 32 860, corresponding to 26 992 different Swiss-Prot entries (as a given Swiss-Prot entry may be detected by more than one profile). Of the 427 profile-based rules 114 (representing ∼25% of the PROSITE profiles) contain information about functionally and/or structurally critical amino acids. This information is also made available to the user via the ScanProsite web page (http://www.expasy.org/tools/scanprosite/). Of the 427 profile-based rules 21 contain a DE line, 115 contain an associated PROSITE pattern and 203 contain keywords. 175 rules only contain information about the domain name and its boundaries.

Quality control of ProRule

There are two means to evaluate the quality and validity of the rules after their creation. The first way is the usage of ProRule by Swiss-Prot annotators. During the validation process annotators evaluate the accuracy of the rule and suggest modifications if needed. In addition, each rule contains the accession (AC) number of a reference protein containing the information provided by the rule. The rule is regularly applied against the reference protein to detect discrepancies resulting from modification to the Swiss-Prot format or the availability of new information that could be added to the rule.

Limitation of ProRule

The efficiency and reliability of ProRule depends on a number of constraints imposed not only by technical aspects of profiles, but also by the biological complexity, which cannot always be reduced to a simple rule. ProRule primarily depends on the sensitivity and specificity of the profile, as the rule should obviously only be applied

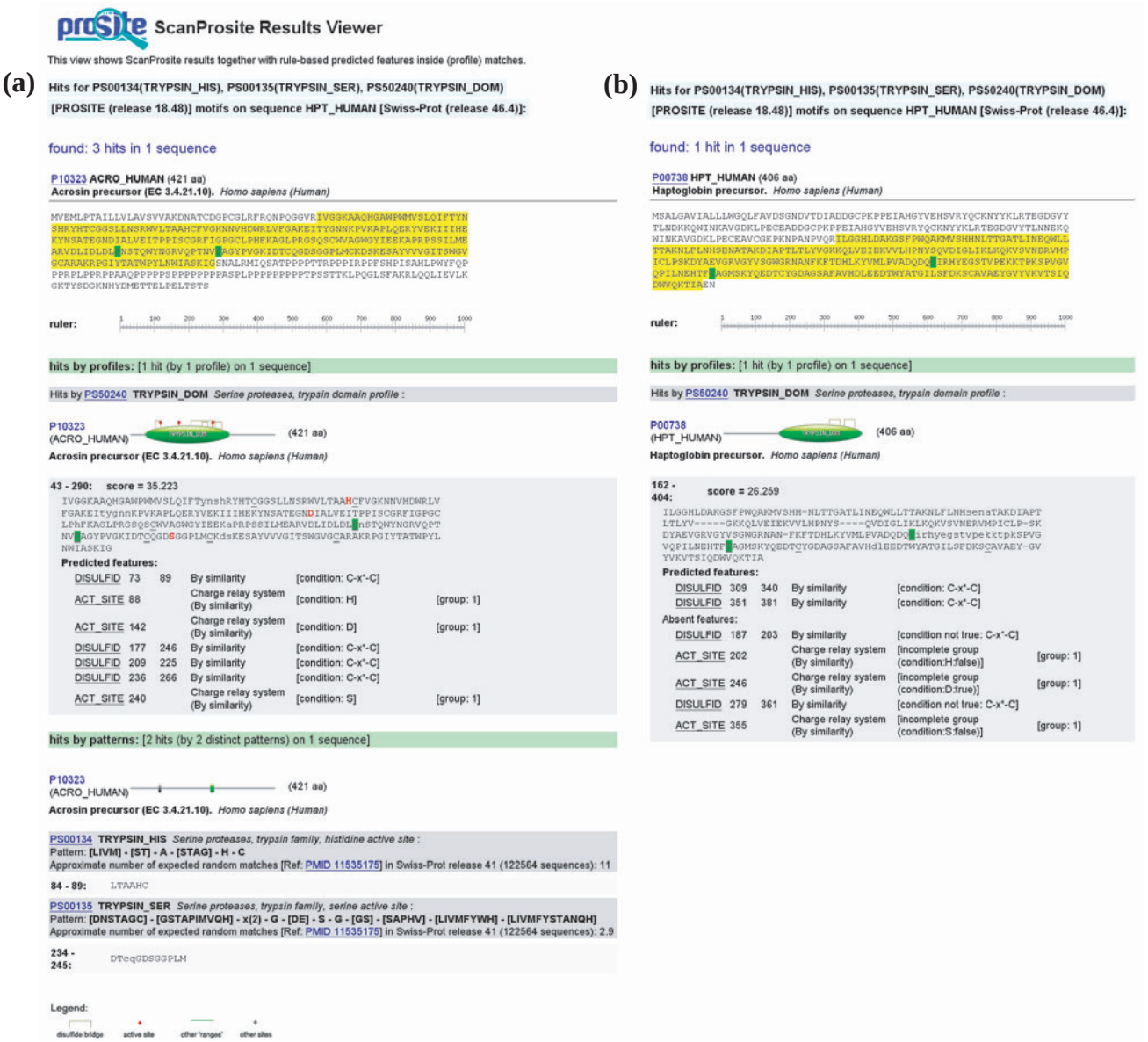

Fig. 4. ScanProsite results for human (**a**) acrosin (P10323) and (**b**) haptoglobin (P00738) scanned using the trypsin-type serine protease domain profile (PS50240) and the histidine active site (PS00134) and serine active site (PS00135) signatures. The predicted features come from the ProRule triggered by the profile (Fig. 2).

to true positive hits. The sensitivity and specificity of the profiles can be evaluated by the matchlist against the Swiss-Prot knowledgebase, which indicates the number of true positives, false positives and false negatives (Sigrist *et al.*, 2002). On the other hand, the positional features contained in a rule can also be used to evaluate the validity of a hit with a given profile. If all the features corresponding to a given profile are missing, it could be that the domain is degenerate or that the hit corresponds to a false positive. It is also possible that the alignment between the profile and the sequence produces an amino acid with the expected properties at a given position by chance, but this probability is low and can even be further reduced by grouping conditions (the higher the number of conditions, the lower the probability). The quantity of annotation to be transferred from a rule also depends on the quality of the alignment between the profile and the sequence. Any local misalignment in the region of an interesting site can affect the detection of biologically meaningful residues.

Finally, the additional information linked to a given profile must depend on features that can be included in the form of a simple rule. If the situation is too complex, the conditions cannot be formulated in the simple syntax of a rule. Conditions that predict which cysteines are linked together by a disulfide bridge are very complex as it is known that disulfide bridges are not always invariant among homologous domains: conserved cysteine residues can exhibit different

disulfide bond patterns from one protein type to the other (Calvete *et al*., 2000; Chong *et al*., 2002). In some instances, the situation is so complex that it is not possible to reduce the biological diversity to a set of conditions predicting unambiguously which bond will be made and no rules can be made. In addition, it is also possible that residues important for domain function are located in a region handled as an insert and thus do not correspond to a position in the profile. These limitations reflect biological variation for which the flexibility of ProRule may be further improved.

Nevertheless, despite these limitations, ProRule increases the discriminatory power of profiles for function determination and facilitates the annotation process, concomitantly making the content of Swiss-Prot more homogeneous and consistent.

ACKNOWLEDGEMENTS

The authors would like to thank Tania Lima for the correction of the manuscript. PROSITE is supported by grant no. 3152A0-103922/1 from the Swiss National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bairoch,A. *et al*. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.*, **5**, 39–55.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Calvete,J.J. *et al*. (2000) The disulfide bond pattern of catrocollastatin C, a disintegrin-like/cysteine-rich protein isolated from Crotalus atrox venom. *Protein Sci.*, **9**, 1365–1373.

Chong,J.M. *et al*. (2002) Disulfide bond assignments of secreted Frizzled-related protein-1 provide insights about Frizzled homology and netrin modules. *J. Biol. Chem.*, **277**, 5134–5144.

Gattiker,A. *et al*. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Hulo,N. *et al*. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.

Junier,T. *et al*. (2001) mmsearch: a motif arrangement language and search program. *Bioinformatics.*, **17**, 1234–1235.

Sigrist,C.J.A. *et al*. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.

Stradal,T. *et al*. (1998) CH domains revisited. *FEBS Lett.*, **431**, 134–137.