Annotation of protein residues based on a literature analysis: cross-validation against UniProtKB

Kevin Nagel*[1], Antonio Jimeno[1] , Tom Oldfield[1] and Dietrich Rebholz-Schuhmann[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Email: Kevin Nagel*- auyeung@ebi.ac.uk;

*Corresponding author

Abstract

Background: A protein annotation database, such as the Universal Protein Resource (UniProtKB), is a valuable resource for the validation and interpretation of predicted 3D structure patterns in proteins. Previously, results have been on point mutation extraction methods from biomedical literature which can be used to support the consuming work of manual database curation. However, these methods were limited on point mutation extraction and do not extract features for the annotation of proteins at the residue level.

Results: This work introduces a system that identifies protein residue sites in abstract texts and annotate them with features extracted from the context. The performances of all text mining modules were evaluated against a manually annotated corpus. The identified annotation features can be attributed to at least one of six targeted categories, e.g. enzymatic reaction. Extracted results were cross-validated against UniProtKB and for 13 annotations of residues that have not been confirmed in the UniProtKB a manual assessment was performed.

Conclusions: This work proposes a solution for the automatic extraction of protein residue annotation from biomedical articles. The presented approach is an extension to other existing systems in that a wider range of residue entities are considered and that features of residues are extracted as annotations.

Background

The understanding of the biological function of proteins remains to be a central challenge in biology. In protein science, sequence analysis of amino acids or studies of their spatial distribution have led to predictions and discoveries of a number of biological significant patterns and motifs, e.g. metal-binding sites, catalytic triads, and ligand binding sites [1–7]. Complementary to these mined data is the proliferation of protein annotations by extracting information from biomedical articles in the view of updating existing databases. Clearly, annotations can be used to verify data mined sequence/structure patterns and likewise predicted patterns can be used to search for association in the database. However, the major annotation effort at the current stage is the compilation of features at the protein level, while the actual target should be at the residue level, because biological function! s can be mapped to a defined group of residues in proteins (function sites). This is also reflected in the field of automatic information extraction from literature, where solutions

have been published for the extraction of interactions of proteins [8, 9], subcellular protein localisation [10], pathway discovery [11], and function annotation with Gene Ontology terminologies [12]. Few groups have investigated in point mutation extraction, but without feature extraction for residue annotation [13–17].

Works have been published that focused on the extraction of point mutations, which is one type of a residue entity [13–17]. The point mutation extraction systems called MEMA [16] and MuteXt [17] use a dictionary lookup approach to detect protein names and disambiguate multiple protein-residue pairs with a word distance measurement. MutationGraB [13], the successor of MuteXt, uses a graph bigram method to calculate the proximity by weighting the association of word-pairs. Another application called MutationMiner [15] focuses on the integration of extracted point mutations into a protein structure visualisation program.

These systems are all dedicated to the extraction of point mutations, but provide no extraction of residue annotation. In a recent publication [14], an ontological model was proposed that should hold information extracted from MutationMiner as well as point mutation annotations. However, the author did not provide any results of feature extraction nor was a strategy proposed. Residue annotation differs from functional annotation of proteins because the biological role of a residue is described rather in a biochemical context, which is then revealed in the function or property of the protein. At present, there is neither such an ontological model nor a terminological resource publicly available.

The goal of this research is the identification of biological function of mined structure patterns of proteins. For this purpose a novel approach that combines structure mining and text mining is proposed. The results of the combined mining study will be published elsewhere. This paper reports on the text mining part and introduces a strategy for the compilation of protein residue annotations that can be used for the interpretation of structure patterns. The result demonstrates that textual information can be captured and used to augment data in UniProtKB. Because the primary data resource is Medline, the extraction covers a broad range of biomedical fields, but is limited to abstract texts. The biological community benefits from the extracted annotations, for example, in that data mined structure patterns can be interpreted biologically or predicted function in proteins can be better characterised.

The contribution of this work is the automatic extraction of protein residue annotation from biomedical articles. Contextual information are exploited to identify features of residues that correspond to one of six chosen target categories (SCAT, Table 1). As a result, proteins can be selected with residues clustered by annotation types, which can lead to discovery of, for example, evolutionary relationships.

Results and Discussion

The following sections assess first the extraction system and then the extracted data.

Evaluation of the identification systems for mentions of organism, protein and residues and their associations.

In order to evaluate the performance of the NER and the AD systems used in this study, the results were compared against the results from manual curation of a set of 100 Medline articles, i.e. the gold standard corpus (GC) generated as part of this study.

Table 2 (top) shows the performance of each named entity recognition. With an F1 measure of 0.91 the performance of the residue tagger is within range of previous works where only the residue was identified as point mutation [13–17]. On the other hand, the performance of organism name recognition was lower with precision of 0.81 and recall of 0.72. The protein recognition has the lowest performance (precision = 0.65, recall = 0.60). The relatively low recall is due to permutation and lexical variants in text that are not covered by the dictionaries.

The evaluation of the organism-protein-residue AD module shows that the algorithm of [17] is suitable for association detection. The performance has a precision of 0.83 and a recall of 0.33 (Table 2, bottom). Two prominent reasons for the low recall is the correct organism-protein association but with a mismatch of protein sequence and residue, or the association of organism and protein was wrong in the first instance.

The implemented association detection system is able to extract associations in accordance to UniProtKB.

Cross-validation of organism-protein association with UniProtKB.
In this section the evaluation was performed automatically on a cross-validation test set (XC) derived from the UniProt corpus (UC). From the 136,566 citations listed in the UniProt a virtually complete set of 136,559 abstract texts were retrieved from Medline to build the UC. Subselection from UC to determine XC resulted in 5,253 abstract texts representing a range of diverse proteins (Table 3, top). Corresponding to this test corpus is the set of 70,401 triplet identifiers of UniProtID-TaxonomyID-PMID (UTP) for the protein-organism association evaluation and 68,008 triplet identifiers of UniProtID-ResidueID-PMID (URP) for the protein-residue association (Table 3, middle and bottom).

With a precision of 0.77 and recall of 0.08 (F1 = 0.14) the result for organism-protein association extraction indicates that although the system seems to extract correct relations with a reasonable number of TP the recall of the solution is too low to fully judge on the performance. The low recall is best explained by missing information in the scientific documents that would confirm the organism-protein association. The results shows that the stringent residue-sequence match resulted in a precision of 1.00 and recall of 0.14 (F = 0.25). The low recall can be explained by several factors: 1) differences between the protein sequence index between the author and the database; 2) changes in the sequence indexing rules by UniProtKB; 3) sequence variants which have not been reported in the database yet; 4) false protein-organism association with the consequence of retrieving the incorrect sequence.

Notice the evaluation of the extraction system was done on Medline abstracts for a range of diverse proteins indexed by UniProtKB as opposed to previous works with extraction from full texts for a few protein family examples. Therefore the results implicate that the extraction from only abstract texts is possible for a number of different UniProt proteins.

PDB citation enrichment.
For each PDB protein entry a link to a corresponding UniProt record is available. The AD system extracts only relations for proteins recorded in the UniProtKB. Therefore each Medline record with a found o-p-r association can be added to the citation set of the corresponding PDB entry. At the state of this analysis, the PDB contained 42,943 PDB protein structure with a sub-fraction of 42,653 having a unique corresponding UniProt protein identifier (11,912). For each of these proteins the whole Medline was scanned for abstracts with extracted organism-protein-residue associations. Figure 1 shows the comparison of the citation sets based on UniProtKB references and the whole Medline analysis.

For 2,535 out of 11,912 proteins the extraction system found a total of 18,748 corresponding PMIDs. Analysis with citation indices for this subset of proteins revealed that 680 out of 18,748 PMIDs were rediscoveries. The low number of rediscovery can be explained in that many annotations are done from sections only available in the full text. Although the analysis was based on Medline abstract texts, the extraction was already able to find for 21 percent of the target proteins a large number of citations. With a precision of 0.83 (determined by gold standard evaluation) the estimated number of TP from the novel discovered citations is 15,560. In context of the 16,560 references of the 2,535 proteins from UniProtKB, the extraction expands the citation set by 1.94 fold.

The extraction system can be used to expand the citation list of UniProtKB/PDB by using only Medline abstract texts. In this experiment the estimated number of overlooked citations for a subset of target proteins provide already a large set for feature extraction for the annotation of protein residues.

Evaluation of feature extraction.
The detection of domain specific features was done by a classification approach which required a labelled reference set and a defined set of categories. The precision, recall and F1-measure values were calculated for each category and summarised in Table 4. Two sets of categories were tested, each with different but corresponding semantic categories: (1) the six targeted categories (SCAT) and (2) the categories listed in the feautre table in UniProtKB (FCAT).

For SCAT, the classifiers for structure component, chemical modification, binding type yielded in F1 measures of 0.69, 0.61, and 0.67. For FCAT the top performing classifiers were: motif, variant, and binding with similar F1 scores (0.62, 0.61, 0.58). The remaining classifiers are still usable for feature detection, as they had precision scores comparable to the top F1 performing classifiers: enzymatic activity and cellular phenotype from SCAT, modified residue, ac-

tive site and site for FCAT. The figures indicate that the features used here are suitable for feature detection and their classification. The performance of feature detection was tested on the gold standard corpus (GC). Sentences with residue mentioning were examined and where applicable suitable features were annotated manually and compared with the extraction method. The number of validated and non-validated features was determined and performance measured.

The performance shows that the classification approach for feature detection had a reasonable coverage for SCAT and FCAT (recall of 0.61 and 0.59 for SCAT and FCAT) but is imprecise in capturing the correct annotation (precision of 0.21 for both, Table 5). This is not surprising, considering that features are expressed throughout the whole sentences, but have different attachments to named entities.

The association of residues and features was based on a syntactical analysis of their verbal and prepositional relations by using a shallow language parser. The approach was evaluated by the performance of detecting all manually annotated residue-feature pairs within the GC data set. With a precision of 0.54 and recall of 0.81 the performance of the shallow parser suggests it is highly usable for residue annotation extraction (Table 6). The low precision is explained by the current implementation of the parser which returns relations with nested prepositional phrases, thus the calculated precision tends to have a lower value. extraction performance decreases when additional extraction modules (NER, AD, FE) were used. This shows that the extraction of annotation is greatly sensitive to each extraction modules.

Despite the performance of each module can be improved, the result shows that the extraction system can deliver residue annotations.

Protein residue annotation extraction and comparison with UniProtKB.

The extraction system in this study delivered classified features of protein residues from Medline as annotations. This section provides examples of the validity of the drawn annotations by comparing extracted information from the gold standard corpus with entries in the UniProtKB.

Within this experiment, four UniProt proteins with a total of 19 annotations from seven sentences and five abstract texts were mined with the extraction system (Table 7). By comparing the mined annotations with correspondent entries in the UniProt six out of 19 annotations were equivalent to existing information in the database (rediscovery). Further, the semantic tags of the annotations, provided by the classification of extracted text features, are biologically meaningful. For example, "the putative catalytic triad" is correctly tagged as enzymatic, because it is a chemical reaction site and therefore a requirement for enzymatic function. In this example, the predicted semantic tag is equivalent to the category active site from the feature table in UniProt. In another example, "major phosphorylation sites" was evaluated as rediscovery of the database information "Phosphothreonine; by MAPK" and "Phosphoserine; by MAPK" while the predicted tag (structural com! ponent) and the assigned category in UniProt (modified residues) are not equivalent. This is still valid, because both pieces of information describe the function of the residues as modification site, while the predicted tag represented this as a substructure and UniProt emphasises on the modification of the residues.

For the remaining 13 extracted annotations there are no equivalent information represented in the UniProt. All are tagged with structural component which is biologically valid, for example, "highly conserved C-terminal region" is an important substructure of the protein and the extraction can aid in determining evolutionary important residues of protein families. However, the annotation "conserved phosphopantothenate binding" can arguably be discussed whether it should be tagged as structural component or binding.

In conclusion, the biological significance of the extracted annotations were studied by comparison with annotations from UniProt for the extracted proteins from the gold standard corpus. From the comparison, the rediscovery data shows that the used SCAT scheme and its feature sets are able to capture information correspondent to UniProt annotations. The predicted semantic tags are biologically valid and do not necessarily have to be equivalent to the categories found in the database. On the other hand, the novel discovery data indicates a potential contribution of the extraction for the automatic annotation of protein residues in UniProt.

Conclusions

The aim of this work was to compile protein residue features from Medline texts as annotation for UniProtKB proteins by combining a series of text mining methods. Although the performances of each module may not be at optimal level, the generated data output indicates that the strategy is able to deliver biological meaningful results. Cross-validation with UniProtKB analysis indicate that the extraction contains novel information that can complement and update the knowledge in UniProtKB and consequently provide annotations for PDB protein structures.

It is important to note that the extraction was done only on abstract texts from Medline. The advantage over full text is to exploit a publicly available broad range of scientific publications but on the cost on the information level of abstract texts. However, the results demonstrate that even with abstract texts a vast amount of annotation can be obtained.

As with high performing NER, AD, and FE systems become more available, this conceptual strategy in protein residue annotation extraction may yield optimal results for the biological community.

Methods

The extraction of protein residue annotation from text can be divided into three steps: 1) named entity recognition (NER) and extraction of residue mentions, 2) association detection (AD) of related named entities, 3) extraction of annotation features for associated entities.

NER for protein and species.

Named entity recognition for proteins was based on an approach that combined dictionary lookup with fuzzy matching and basic disambiguation [18–20]. All protein names were collected from UniProtKB/SwissProt. Names of species were extracted from the NCBI Taxonomy references from UniProtKB/SwissProt and then collecting scientific and common names of the referenced organisms. The dictionary was complemented with terminologies describing only the referenced genus and the collection of full organism name (genus + specie) augmented with abbreviated genus forms (first letter abbreviation of genus + specie). Web services for the identification of protein names and taxa names are available from the TM infrastructure at the EBI ([18]).

Identification of residue mentions from the text.

The extraction of residue mentions follows approaches of previous publications [16, 17]. Sets of regular expressions were constructed to identify three types of protein residue site mentions. The first basic type is the single protein sequence site reference which consists of a (wild-type) amino acid name, followed by the sequence position number (e.g. “Gly-12”, “arginine 4”, “Tyr74”, “Arg(53)”). A point mutation is the second type of residue site where the description details the change of an amino acid at given position. The common notation is the wild-type amino acid name, the sequence position followed by the substitution (e.g. “W77R”, “Cys560Arg”, “ser-52->ala”, “ala2-methionine”). Finally, the third type of residue site describes either a list of residues or an interaction pair (e.g. “Tyr 85 to Ser 85”, “Trp27–Cys29”). The common notation is an amino acid name, sequence position, a connection symbol or conn! ection word, amino acid name, and sequence position. In addition to the abbreviated notation residue sites can be expressed in grammatical form (e.g. “isoleucine at position 3”, “substitution of Ala at position 4 to Gly”, “Ser472 to glutamic acid”).

Identification of associations between mentions of species, proteins and residues.

The identification of a residue can only be validated, if it is part of the protein sequence as it is reported in a reference database (e.g., UniProtKB). This requires that the protein mention in the text is further supported by evidence for the species under scrutiny to select the appropriate protein sequence from the bioinformatics database; that excludes the risk of using orthologous protein sequences. The association of organisms with proteins and the proteins with residues was done based on the algorithm described by [17]. First, specie and protein mentions were associated by measuring the word distance between them. Associated proteins and their specie mention form a pair that correctly specifies the protein with a unique identifier in the reference database (UniProtKB). If no match was found, the association was relaxed to genus matching resulting in a list of protein identifiers. In case of multiple organisms matching, word proximity metric was used to pr!

efer the closest word-pair. The identifier was used to retrieve the protein sequence from the database in order to validate the residue mention. According to the algorithm proposed by [17], three cases can be distinguished: (1) the residue correctly matches the protein sequence, (2) several alternative sequences are matching from a list of protein mentions (identifiers), and (3) no match can be found for the residue in the available protein sequences. If several protein sequences were relevant candidates, then again the word distance metric was used to select the closest word pairs.

Feature extraction for the annotation of residues. The origin of a biological function of a protein is group of residues and their experimental characterisation are reported in scientific publications. In this study the feature extraction process was divided into two parts: in the first part the text was processed to extract NPs that served as candidate features, and in the second part the extracted candidate features were classified into categories of annotation features. Noun phrases are specified as nominal forms in combination with adjective and adverb mentions (NP = Det? (Adj—Adv—N)* N). Even though most NPs denote terms this is not always true [21].

In the first part, the abstract text was split into sentences and annotated with part-of-speech (pos) tags using the cistagger which has a similar performance as the treetagger but it has an integration of a large biomedical terminological resource. Then the shallow parser described in [22] was applied to extract verbal and prepositional dependencies. Since this parser does not deal with prepositional attachment ambiguity it has been extended with a prepositional phrase attachment disambiguation module explained in [23]. In the second part, the features were categorized using the endogenous classification approach described in [24]. Basically, the algorithm relies only on the mutual information of the lexical constituents of terms and their assigned categories. In contrast, the exogenous (corpus-based) approach requires large amounts of contextual cues which are difficult to obtain. The endogenous approach is therefore more reliable to produce results even under conditions of sparse data. During the training phase, lexical constituents of multi-word terms were extracted from a labelled reference set and represent features for a defined set of categories. The association between both, the features and the categories, were estimated based on their mutual information score and the association between the multi-word term and a category was computed as the sum of the associations of its constituents. The categorization of a multi-word term into one of the categories then amounts to t! he identification of the best fitting category for a term based on the term's components. The reference set for the relevant multi-word terms was generated using maximal length noun phrase (MLNP) analysis based on two different sets of NPs that were extracted from an whole Medline abstract texts analyses: the first set consists of NPs that co-occurred with residue mentions in the same sentence without nested residue terms (NP(not r)), and the second set represents NPs with nested residue terms (NP(r)); since the co-occurrence with a residue may indicate higher relevance. Once the set of MLNPs were extracted each NP was manually labelled using three different categorization schemes. The first scheme is binary labelling (BCAT) to separate domain relevant terms from non relevant ones. The second scheme uses six semantic categories identified from a study on the manual categorization of residue annotations based on scientific content from Medline (bottom-up approach). The identified categories and their definitions are shown in Table 1 (SCAT). The final set was defined through a top-down approach by reusing categories described in the feature table of the UniProtKB data resource for proteins (FCAT).

Generation of evaluation corpora.
For the evaluation of the extraction system, two test corpora were generated using the UniProt corpus (UC). The UC consists of those Medline abstract texts that are cited in the UniProt database for relevant protein-residue pairs. The complete corpus was automatically analysed for organism, protein and residue mentions and tagged appropriately. A gold standard corpus (GC) was created through manual curation since no corpora are available. A random sample of 100 Medline abstract texts was drawn from the UC where every abstract had to fulfil the condition that a mention of an organism, a protein and a residue was present (tri-co-occurrence). All mentions of an organism, a protein, the residue, the associations between the mentions, and the contained features of the residues (see above) were then annotated manually from two independent annotators with domain expertise. For the automatic evaluation

of extracted data a cross-validation corpus (XC) was derived from UC, because not all database information are necessarily expressed in abstract texts and vice versa. Documents in UC were scanned for tri-occurrences of organism-protein-residue mentions in text, and then analysed if the combinations of the four identifiers UniProtID-TaxonomyID-ResidueID-PMID can be found in the database. If at least a single match was found the document was selected. For the non-matching combinations the corresponding annotations were removed from text.

Authors contributions

Kevin Nagel carried out the experiments, developed and implemented the methods, assessed the annotations, and drafted the manuscript. Antonio Jimeno participated in the development of the methods and drafted the manuscript. Dietrich Rebholz-Schuhmann participated in design of the experiments, assessed the annotation and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Kim Henrick, Michael Ashburner and Rob Russell for their input in this project.

References

1. Barker J, Thornton J: **An algorithm for constraint based structural template matching: application to 3D templates**. *Bioinformatics* 2003.
2. Oldfield T: **Data Mining the Protien Data Bank: Residue Interactions**. *Proteins* 2002.
3. Nebel J, Herzyk P, Gilbert D: **Automatic generation of 3D motifs for classification of protein binding sites**. *BMC Bioinformatics* 2007.
4. Kristensen D, Ward M, Lisewski A, Erdin S, Chen B, Fofanov V, Kimmel M, Kavraki L, Lichtarge O: **Prediction of enzyme function based on 3D templates of evolutionarily important amino acids**. *BMC Bioinformatics* 2008.
5. Polacco B, Babbitt P: **Automated discovery of 3D motifs for protein function annotation**. *Bioinformatics* 2006.
6. Yoon S, Ebert J, Chung E, DeMicheli G, Altman R: **Clustering protein environments for function prediction: finding PROSITE motifs in 3D**. *BMC Bioinformatics* 2007.
7. Stark A, Sunyaev S, Russell R: **A model for statistical significance of local similarities in structure**. *J Mol Biol* 2003.
8. Marcotte E, Xenerios I, Eisenberg D: **Mining literature for protein-protein interactions**. *Bioinformatics* 2001.
9. Blaschke C, Andrade M, Ouzounis C, Valencia A: **Automatic extraction of biological information from scientific text: Protein-protein interactions**. *Proc Int Conf Intell syst Mol Biol* 1999.
10. Stapley B, Kelley L, Sternberg M: **Predicting the subcellular location of proteins from text using support vector machines**. *Pac Symp Biocomput* 2002.
11. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A: **GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles**. *Bioinformatics* 2001.
12. Blaschke C, Leon EA, Krallinger M, Valencia A: **Evaluation of BioCreAtIvE assessment of task 2.** *BMC Bioinformatics* 2005.
13. Lee L, Horn F, Cohen F: **Automatic extraction of protein point mutations using a graph bigram association**. *PLoS Computational Biology* 2007.
14. Witte R, Kappler T: **Enhanced semantic access to the protein engineering literature using ontologies populated by text mining**. *Int. J. Bioinformatics Research and Applications* 2007.
15. Baker C, Witte R: **Mutation Miner - Textual Annotation of Protein Structures**. *CERMM Symposium* 2005.
16. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H: **Automatic extraction of mutations from Medline and cross-validation with OMIM**. *Nucl. Acids Res.* 2004.
17. Horn F, Lau A, Cohen F: **Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors**. *Bioinformatics* 2004.
18. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A: **Text processing through Web services: Calling Whatizit**. *Bioinformatics* 2008.
19. Pezik P, Jimeno A, Lee V, Rebholz-Schuhmann D: **Static dictionary features for term polysemy identification**. *Building and evaluating resources for biomedical text mining, LREC Workshop* 2008.
20. Tsuruoka Y, Mcnaught J, Ananiadou S: **Normalizing biomedical terms by minimizing ambiguity and variability**. *BMC Bioinformatics* 2008, **9**.
21. Krauthammer M, Nenadic G: **Term identification in the biomedical literature.** *J Biomed Inform* 2004.
22. Leroy G, Chen H, Martinez J: **A shallow parser based on closed-class words to capture relations in biomedical text**. *J Biomed Inform* 2002.
23. Schuman J, Bergler S: **Postnominal Prepositional Phrase Attachment in Proteomics**. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, Association for Computational Linguistics 2006.
24. Cerbah F: **Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms**. *COLING* 2000.

25. Gaizauskas R, Demetriou G, Artymiuk P, Willett P: **Protein structures and information extraction from biological texts: the PASTA system**. *Bioinformatics* 2003.

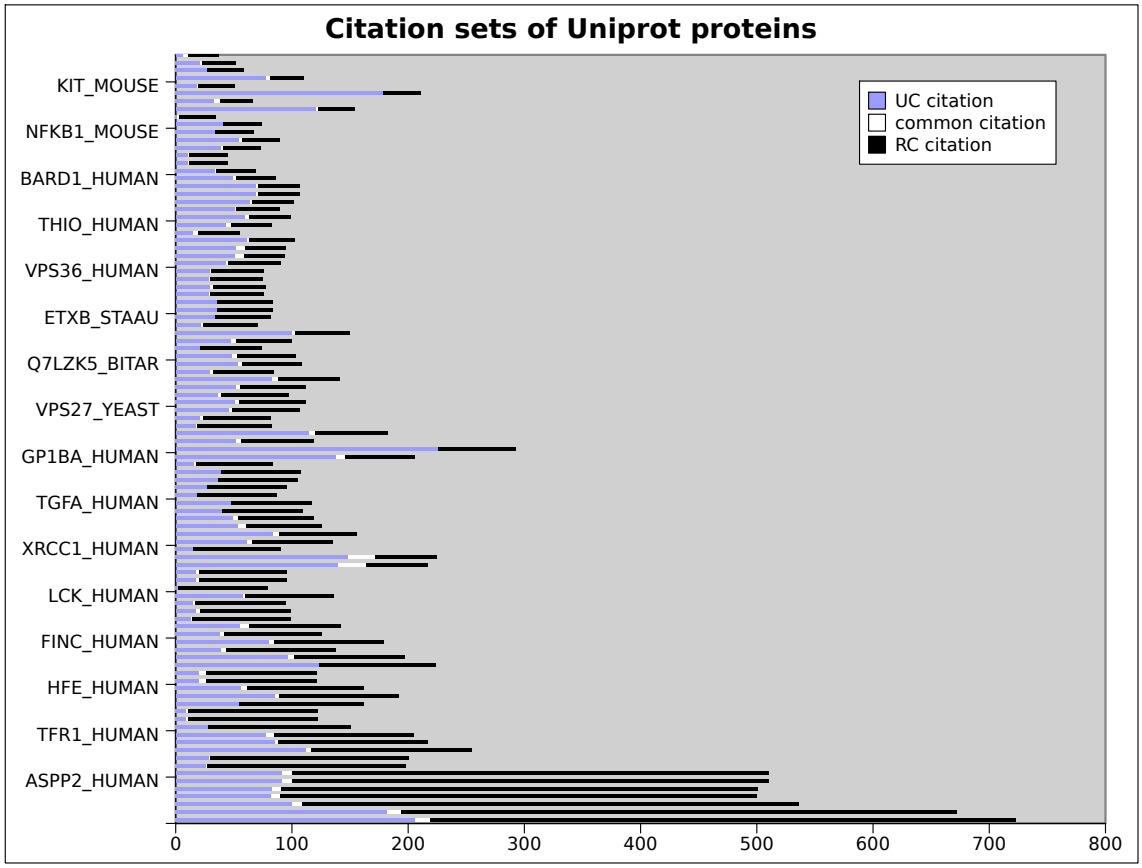
26. Ashburner M, Lewis S: **On ontologies for biologists: the Gene Ontology - uncoupling the web**. *Novartis Found Symp* 2002.

27. Bairoch A: **The ENZYME database in 2000**. *NAR* 2000.

Figures

Figure 1: Comparison of UniProt indexed citations and discovered citations from Medline.

The extraction system identified for a subset of all UniProt proteins the triple associations of organism-protein-residue in Medline abstract texts. The identified list of citations for these proteins were compared with the citations references from the correspondent UniProt entries.



Tables

Table 1: Six target categories of biological interest (SCAT).

The definition of each category of biological interest targeted in this study are listed together with their references to databases for extracting candidate terminologies. A mapping of these categories to equivalent/similar categories from UniProtKB (FCAT) is provided.

SCAT	FCAT	reference	definition
structure component	domain, motif, topo dom, chain, transmem, coil	PASTA [25]	Class denoting concepts that represent pieces and parts of the protein structure.
chemical modification	variant, mod res, peptide, var seq, lipid	n/a	Class denoting changes to the protein sequence and the chemical composition.
structural modification	region, site	n/a	Class denoting the changes to the protein structure without changes to the chemical composition.
binding type	binding, metal, disulfid, crosslnk, dna bind, np bind, zn fing, ca bind	GO [26]	Class denoting different physicochemical forces leading to a bond formation between a protein structure component and a chemical entity.
enzymatic activity	act site	EC [27], GO [26]	Types of enzymatic reactions as a subpart to protein functions.
cellular phenotype	n/a	n/a	Class denoting different cellular phenotypes that can be affected by structural or compositional changes of a protein.

Table 2: Named entity recognition and association detection performance evaluated on gold standard corpus.

Performance was measured in terms of precision, recall, and F1 measure. o = organism; p = protein; r = residue; o-p-r = association of o, p and r.

target	available	extracted	TP	precision	recall	F1
o	123	109	88	0.81	0.72	0.76
p	511	471	305	0.65	0.60	0.62
r	202	222	197	0.87	0.96	0.91
o-p-r	158	63	52	0.83	0.33	0.47

Table 3: Cross-validation of organism-protein-residue extraction with UniProtKB.

Automatic performance analysis of the extraction with UniProtKB as reference. Performance was measured in terms of precision, recall, and F1 measure. UC = UniProt corpus; XC = cross validation corpus; UTP = triplet identifiers of UniProtID-TaxonomyID-PMID; URP = triplet identifers of UniProtID-ResidueID-PMID; o = organism; p = protein; r = residue; o-p = association of o and p; p-r = association of p and r.

									ResID			
data	o	p	r	o-p	p-r	PMID	TaxID	UniProtID	conv	site	seq	range—pair
UC						136,559	11,348	175,695	28,950	33,750	4,021	2,281
UC	-	-	-			6,532	0	0	0			
UC	+					119,880	11,348	174,717	25,482	30,041	3,740	2,095
UC		+				129,792	11,328	175,695	28,932	33,723	3,991	2,278
UC			+			30,732	4,743	115,882	28,950	33,750	4,021	2,281
UC	+	+				119,653	11,328	174,717	25,470	30,014	3,713	2,092
UC	+	+	+			27,709	4,740	113,412	25,470	30,014	3,713	2,092
XC						5,253	1,536	45,869	9,519	7,342	227	421
XC	-	-	-			131,306	0	0				
XC	+					5,253	1,536	45,869	9,519	7,342	227	421
XC		+				5,253	1,536	45,869	9,519	7,342	227	421
XC			+			5,253	1,536	45,869	9,519	7,342	227	421
XC	+	+				5,253	1,536	45,869	9,519	7,342	227	421
XC	+	+	+			5,253	1,536	45,869	9,519	7,342	227	421
XC	+	+		+		5,253	1,536	45,869	9,519	7,342	227	421
XC	+	+	+	+		5,253	1,536	45,869	9,519	7,342	227	421
XC	+	+	+	+	+	4506	1301	3937	8804	5783	0	329

						UTP					
data	o	p	r	o-p	p-r	available	extracted	common	precision	recall	F1
XC	+	+		+		70,401	7,333	5,625	0.77	0.08	0.14

						URP					
data	o	p	r	o-p	p-r	available	extracted	common	precision	recall	F1
XC	+	+	+	+	+	68,008	9504	9504	1.00	0.14	0.25

Table 4: Feature classification performance.
The classification of contextual features of residues mentioned in text was used to identify annotations and to classify them into categories of biological interest. Cross validation was performed with training and test sets with 3600 and 400 features, respectively. Performance was measured in terms of precision, recall and F1 measure.

SCAT				FCAT			
category	recall	precision	F1	category	recall	precision	F1
structure component	0.8	0.6	0.69	motif	0.45	1	0.62
				domain	0.5	0.62	0.55
chemical modification	0.73	0.52	0.61	variant	0.77	0.5	0.61
				lipid	0.4	1	0.57
				modified res	0.47	0.59	0.52
				peptide	0.11	0.29	0.16
binding type	0.68	0.67	0.67	binding	0.63	0.54	0.58
				crosslink	0.25	0.67	0.36
				disulfid	0.17	0.62	0.26
				metal	0.12	0.25	0.16
structural modification	0.25	0.64	0.36	site	0.68	0.47	0.56
				region	0.59	0.46	0.52
enzymatic activity	0.42	0.49	0.46	active site	0.48	0.5	0.49
cellular phenotype	0.47	0.6	0.53	n/a			

Table 5: Feature detection evaluated on gold standard corpus.
The classification method was used to identify features of interest. The performance in detecting manually determined annotations was measured in terms of precision, recall and F1 measure. SCAT = feature detection using the six target categories; FCAT = feature dection using categories from the feature table in UniProtKB.

feature	available	extracted	common	precision	recall	F1
SCAT	164	474	100	0.21	0.61	0.31
FCAT	164	460	97	0.21	0.59	0.31

Table 6: Performance of residue-feature association detection evaluated on gold standard corpus.
The association of residue and annotation was done by shallow parsing and extracting verbal/prepositional relations. The performance was measured in precision, recall and F1 measure. GC = gold standard corpus; r = residue; f = feature; o = organism; p = protien; s = verbal/prepositional relation between r and f; o-p = association between o and p; p-r = association between p and r.

	extraction filter												
data	s	r	f	o	p	o-p	p-r	avail	extr	comm	prc	rec	f1
GC	+							88	132	68	0.52	0.77	0.62
GC	+	+	+					88	65	30	0.46	0.34	0.39
GC	+	+	+	+	+			82	62	27	0.44	0.33	0.38
GC	+	+	+	+	+	+	+	82	93	19	0.20	0.23	0.22

Table 7: Comparison of extracted protein residue annotations with UniProtKB.
The extraction system delivered protein annotation from Medline abstracts. Example of extraction were drawn from the gold standard corpus extraction.

UniProtID	ResidueID	PMID	SCAT	extracted feature	FCAT	UniProt annotation
P40380	THR13	12135491	str comp	major phosphorylation sites for MAPK	mod res	Phosphothreonine; by MAPK
“	SER19	”	”	”	”	Phosphoserine; by MAPK
“	SER19	12135491	chem mod	negative effect	mutagen	S → E:reduces activity as a cdc2 inhibitor; when associated with E-13
Q93K00	ASP123	12147465	enzymatic	the putative catalytic triad	act site	nucleophile (by similarity)
“	HIS279	”	”	”	”	proton acceptor (by similarity)
“	ASP250	”	”	”	”	proton donor (by similarity)
Q93K00	GLU55	12147465	str comp	putative oxyanion hole	n/a	n/a
“	TRP124	”	”	”	n/a	n/a
Q02809	W612	9617436	str comp	”	n/a	n/a
Q9HAB8	GLY43	12906824	str comp	conserved ATP binding residues	n/a	n/a
“	SER61	”	”	”	n/a	n/a
“	GLY63	”	”	”	n/a	n/a
“	GLY66	”	”	”	n/a	n/a
“	PHE230	”	”	”	n/a	n/a
“	ASN258	”	”	”	n/a	n/a
“	ASN59	”	”	conserved phosphopantothenate binding	n/a	n/a
“	ALA179	”	”	”	n/a	n/a
“	ALA180	”	”	”	n/a	n/a
“	ASP183	”	”	”	n/a	n/a