

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

September 22nd, 2008



Kevin Nagel, Antonio Jimeno, Tom Oldfield, Dietrich Rebholz Schuhmann

Rebholz Group

EBI, WT Genome Campus

Hinxton, Cambridge, U.K.

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

EMBL-EBI

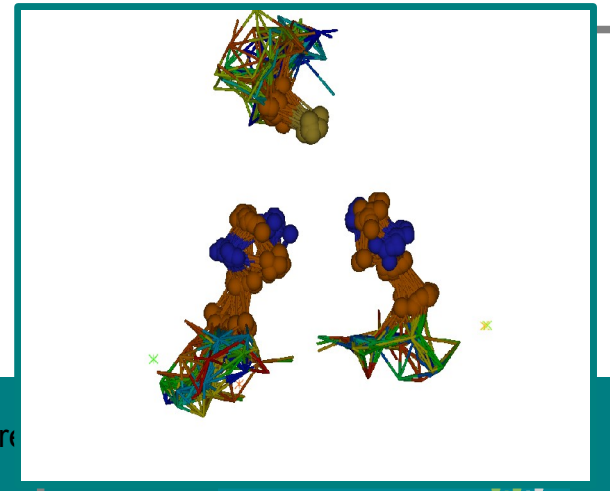
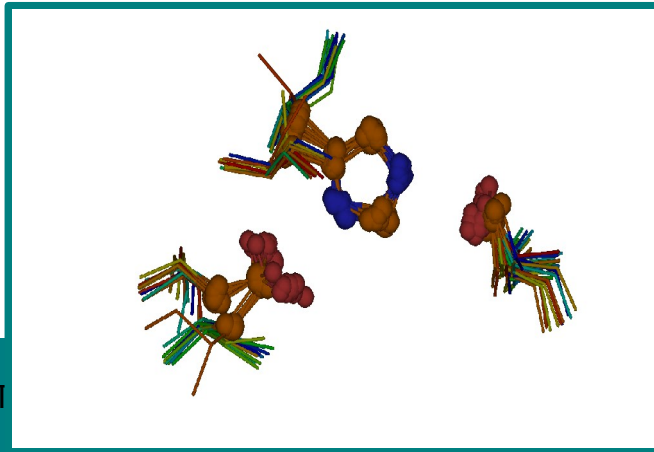
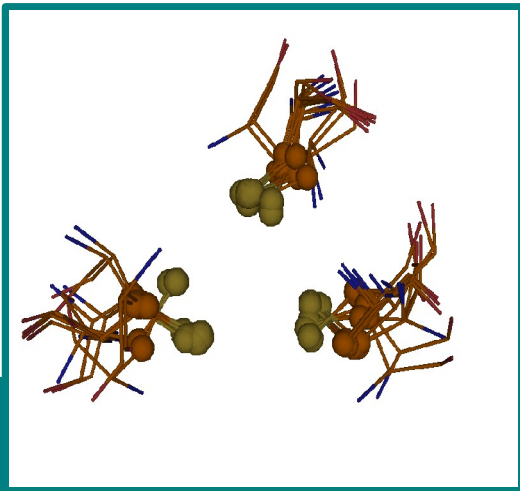


Overview

- We can mine Protein Data Bank (PDB) for potentially novel structure components.
- These novel components lack any functional annotation.
- We can explore and automatically mine the scientific literature to identify potential annotations of the structure components
- Protein residues is a very important part of the structural components from PDB that we want to annotate with function.
- This presentation gives an overview on the results of the literature mining and comparisons to UniProtKb

Protein residues require functional annotations

- PDB provide structural data of proteins
 - > structural data are useless if we don't know the function
- Conserved structural features can reflect biological relevance
 - > not all represent a functional site
- Structural data mining finds conserved 3D patterns of residues
 - > biological significance of data needs to be validated



Question

What evidence for functional annotation do we find in the literature, if we focus on protein structure data (e.g., residues or triads from residues)?

Description of biological function of protein residues are available in literature

[enzyme related]

Using the semi-empirical method of conformational analysis low energetic conformations were found for **trypsin's catalytic triad** Asp102, His57, Ser195 in the field of the active centre that consisted of about 800 atoms. (*PMID:6438492*)

[structure component]

We find that near neutral pH the **binding energetics** are influenced by a shift in the pKa of an **ionizable group**, most likely histidine 57 in the protease active site. (*PMID:9159490*)

[binding event]

Covalent bonds are formed between Val-P1 of the inhibitor and His-57 NE2 and Ser-195 OG of the enzyme. (*PMID:3391280*)

Objective: IE for the annotation of protein residues

- Information extraction for the annotation of protein residues
 - > develop methodology for automatic extraction
- Challenges:
 - > identification of organism, protein, residue and their association
 - > association of contextual features with residue mentions
 - > evaluation of domain relevance of contextual features w/o elaborate ontology
- Proposal: We expect that the annotation of the residues fits into the following categories (according to our domain knowledge).

Six semantic categories of biological interest

| Category | Reference | Definition |
|-------------------------|-----------------------|---|
| structure component | PASTA, PO, CATH, SCOP | pieces and parts of protein structure. |
| chemical modification | PSI-MOD | changes to protein sequence/chemical composition. |
| structural modification | n/a | changes to spatial arrangement of structural component. |
| binding type | GO | physicochemical forces leading to bond formation. |
| enzymatic activity | EC, GO | types of catalytic reaction as subpart of protein function. |
| cellular phenotype | n/a | cellular phenotypes affected by changes in protein |

Approach: combination of several text mining techniques

- dictionary of terminologies from Uniprot, NCBI Taxonomy / Regexp for residue terminology

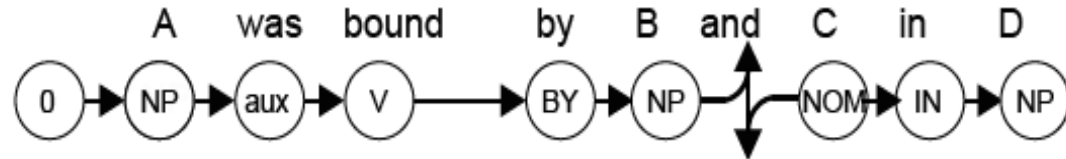
Whatizit pipeline from Rebholz Group (EBI)

- organism-protein-residue association / [Horn, 2004]

$$A(O, P) := P \rightarrow \text{taxid} = O \rightarrow \text{taxid}$$

$$A(P, R) := R \rightarrow \text{seqid} \in P \rightarrow \text{seq}$$

- syntactical relation analysis / VP and PP attachments / [Leroy, 2002; Schuman, 2006]



- assign semantic categories to NP/term / [Cerbah, 2000]

$$\text{term} := T \text{ with } \{w_i\}_{i=1}^n$$

$$T \rightarrow C$$

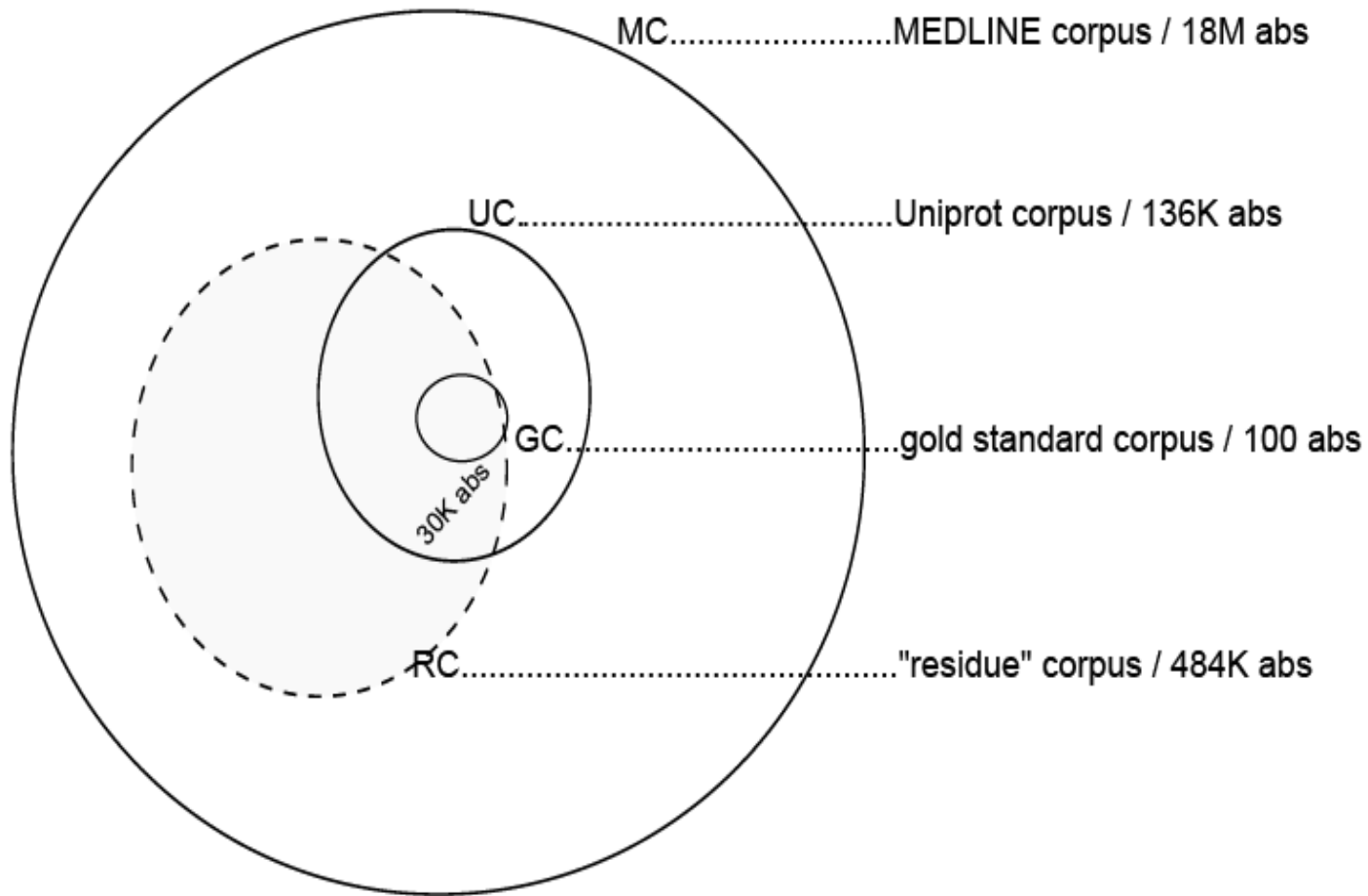
$$\text{dependency} := l(w, C)$$

$$\text{association} := A_t(T, C) = \alpha \sum_{i=1}^n l(w_i, C)$$

$$\text{classification} := C^* = \arg \max_C A_t(T, C)$$



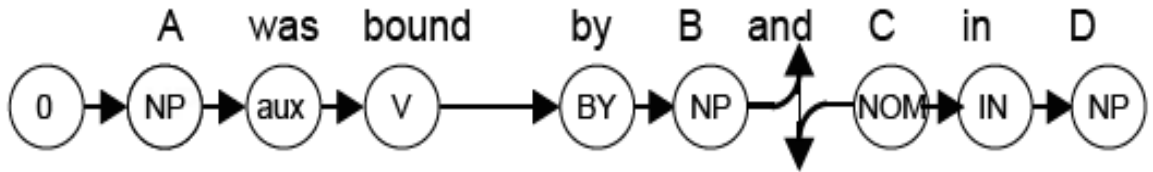
Medline is the primary resource for information extraction



Performance of NE recognition and NE association

| Dataset | org | prot | resi | o-p-r | precision | recall | F1 |
|---------|-----|------|------|-------|-----------|--------|------|
| GC | + | | | | 0.81 | 0.72 | 0.76 |
| " | | + | | | 0.65 | 0.60 | 0.62 |
| " | | | + | | 0.87 | 0.96 | 0.91 |
| " | + | + | + | + | 0.83 | 0.33 | 0.47 |

Performance of contextual feature classification

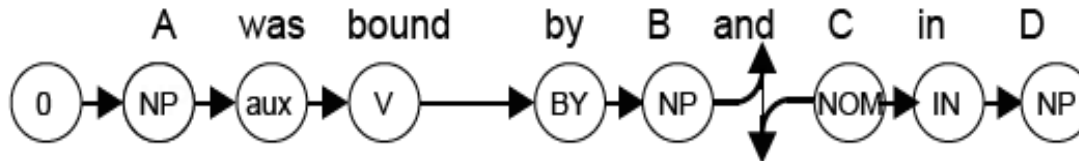


- Manual categorisation of text features in semantic categories
- Automatic categorisation of extracted features into the same semantic categories
- Categories are not disjoint

10 fold cross validation

| Category | prc | rec | F1 |
|-------------------------|------|------|------|
| structure component | 0.60 | 0.80 | 0.69 |
| binding type | 0.67 | 0.68 | 0.67 |
| chemical modification | 0.52 | 0.73 | 0.61 |
| cellular phenotype | 0.60 | 0.47 | 0.53 |
| enzymatic activity | 0.49 | 0.42 | 0.46 |
| structural modification | 0.64 | 0.25 | 0.36 |

Performance of contextual feature detection and association with residue mention



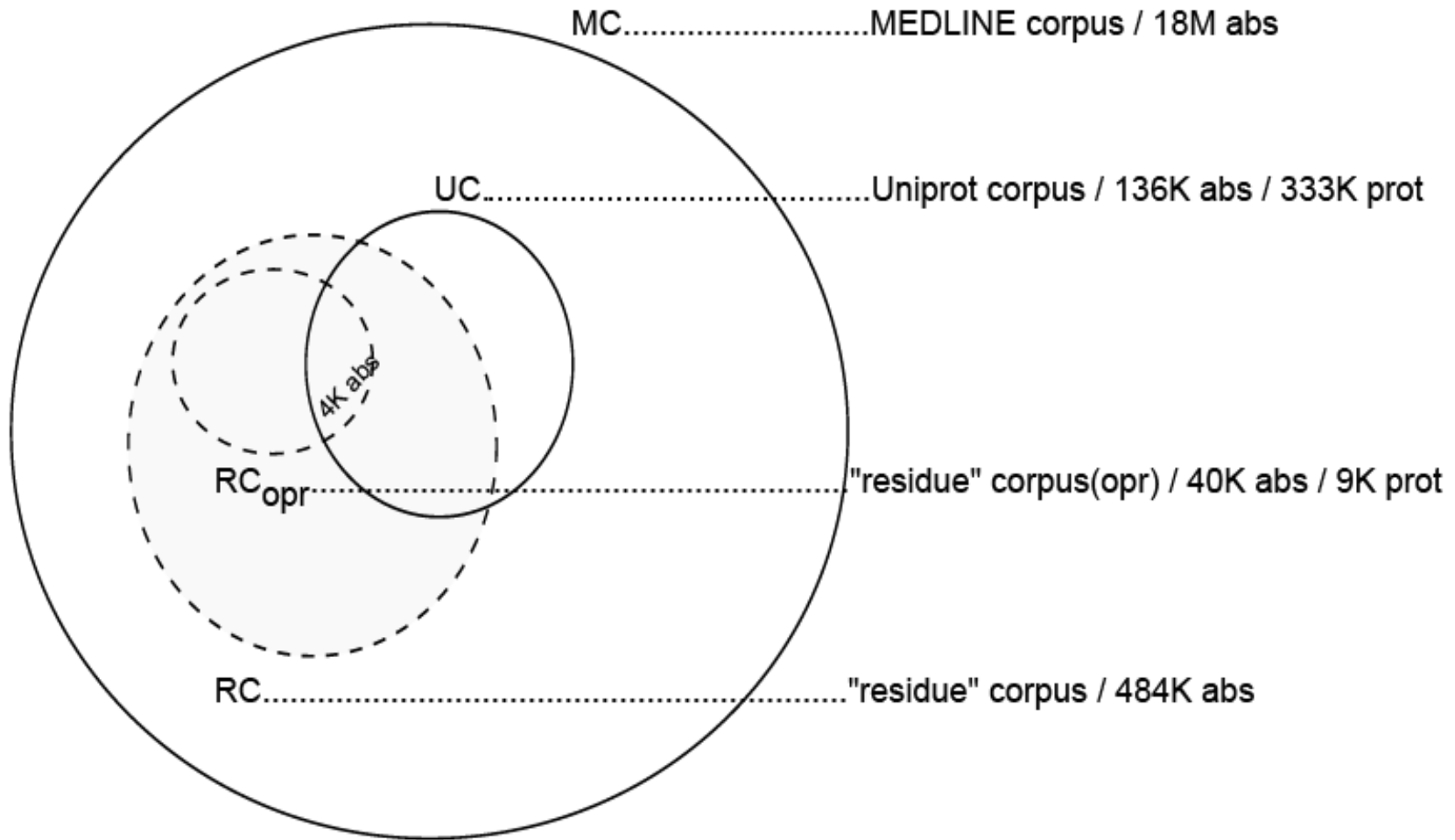
| Dataset | s | f | r | o | p | o-p-r | prc | rec | F1 |
|---------|---|---|---|---|---|-------|------|------|------|
| GC | | + | + | | | | 0.21 | 0.61 | 0.31 |
| " | | + | + | + | | | 0.46 | 0.34 | 0.39 |
| " | | + | + | + | + | | 0.44 | 0.33 | 0.38 |
| " | | + | + | + | + | + | 0.31 | 0.23 | 0.22 |

s = VP / PP structures
 f = contextual features of residues
 r = residue
 o = organism
 p = protein
 o-p = association of o and p
 p-r = association of p and r

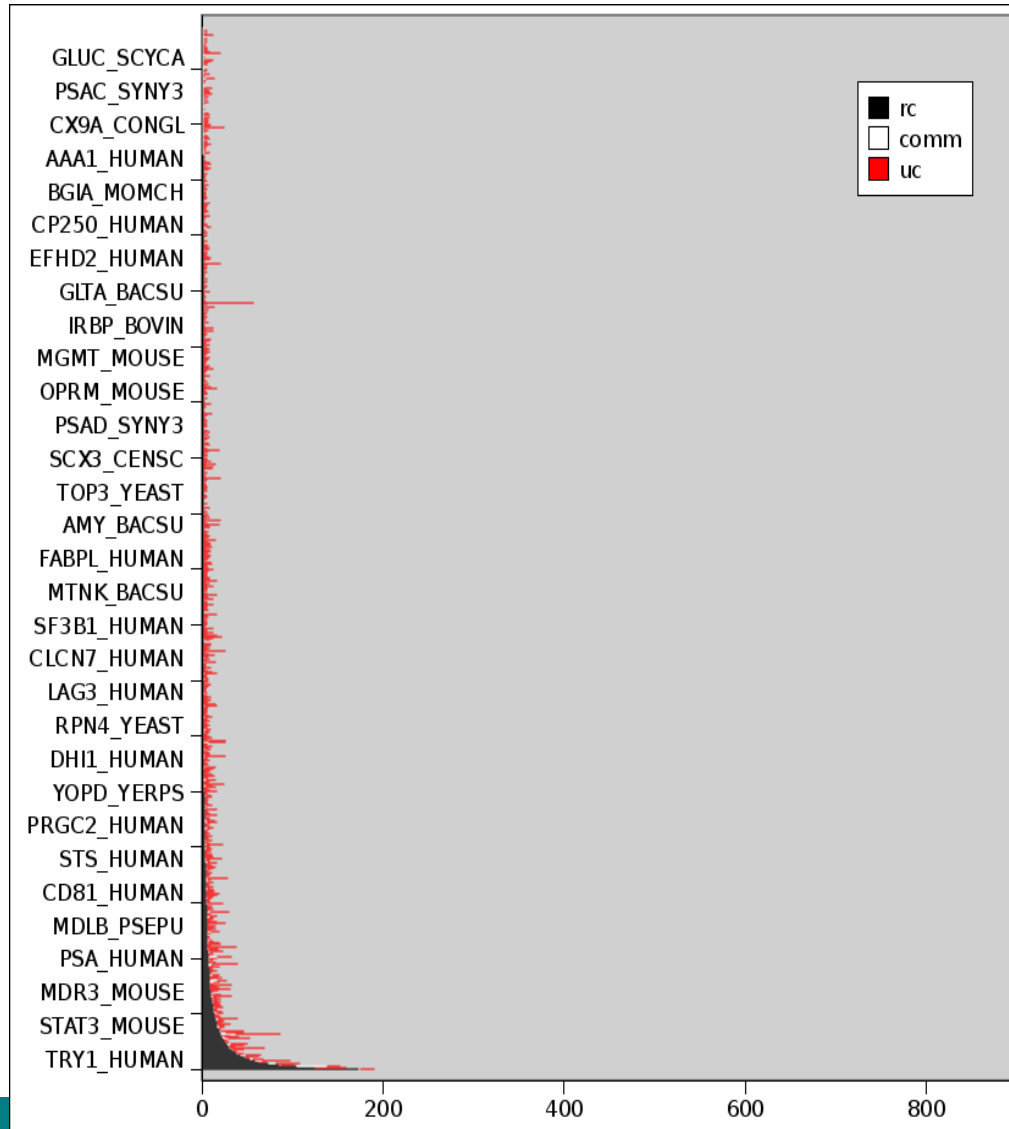
Residue annotation extraction compared to UniProtKB

| UniProtID | ResID | PMID | Category | Annotation | UniProtKB/FT |
|------------|--------|----------|-----------|--------------------------------------|---|
| RUM1_SCHPO | Thr13 | 12135491 | str comp | major phosphorylation sites for MAPK | Phosphothreonine; by MAPK. |
| | Ser19 | | | | |
| | Ser19 | | chem mod | negative effect | Phosphoserine; by MAPK. S->E:reduces activity as a cdc2 inhibiton. |
| DHMA_MYCAV | Asp123 | 12147465 | enzymatic | the putative catalytic triad | nucleophile (by similarity). proton acceptor (by similarity). proton donor (by similarity). |
| | His279 | | | | |
| | Asp250 | | | | |
| PPCS_HUMAN | Gly43 | 12906824 | str comp | conserved ATP binding residues | n/a |
| | Ser61 | | | | |
| | Gly63 | | | | |
| | Gly66 | | | | |
| | Phe230 | | | | |

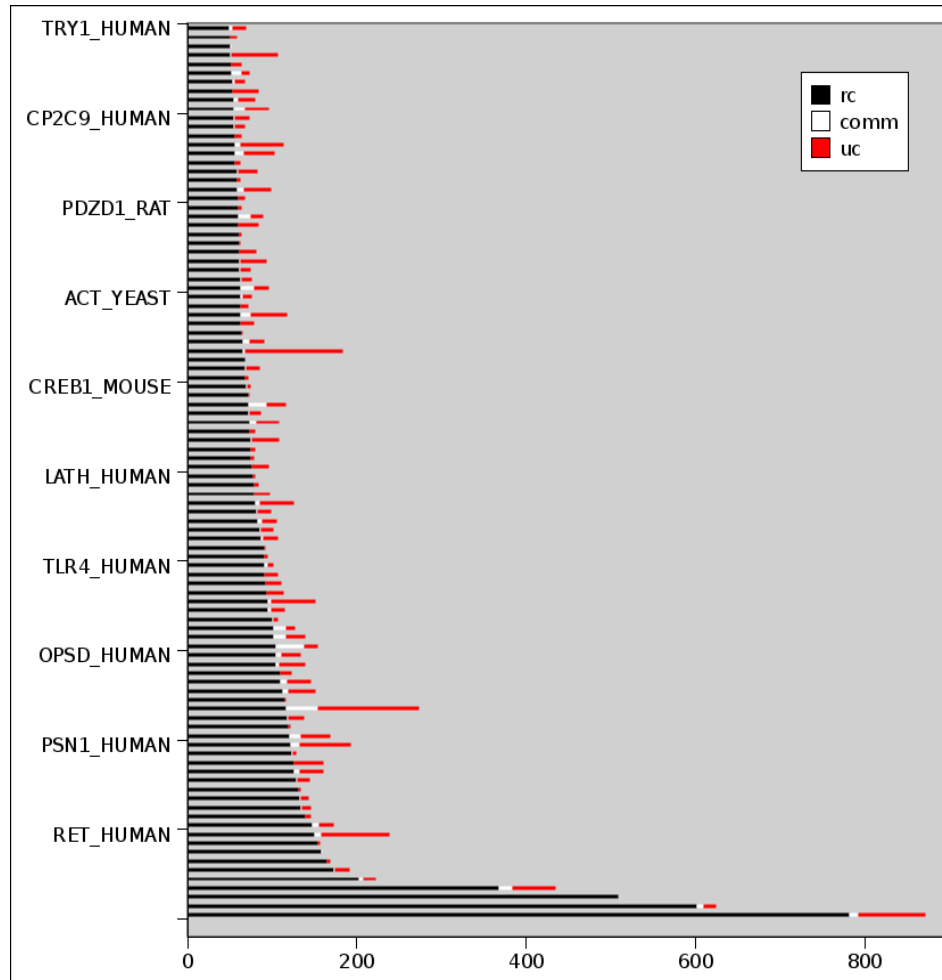
Citations found for Uniprot proteins



Comparison of identified citation sets with Uniprot citations



Comparison of identified citation sets with Uniprot citations



Conclusions

- Introduced IE methodology for the annotation of residues
- Contextual features of residues in text used
- Performances of each module may not be at optimal level
 - > Results justify the approach
- Extraction with Medline data
 - > rediscovery of known knowledge (x-validation with Uniprot)
 - > found other complementing information (update Uniprot)

Acknowledgements



Dietrich Rebholz-Schuhmann

Antonio Yepes



Kim Henrick

Tom Oldfield



Michael Ashburner



Rob Russell