Comparisons of protein annotations in secondary databases

Rother K, Michalsky E, Leser U. How well are protein structures annotated in secondary databases?
Proteins. 2005 Sep 1;60(4):571-576. PMID: 16021624

SILS Biomedical Informatics Journal Club
http://ils.unc.edu/bioinfo/
2005-09-27

Background

- Structure information is most useful when associated with annotations, such as sequence, function, and active site.
- Manual curation of annotations is often required, but the volume of information is outstripping the ability for curators to keep up.
- The absence of links to secondary databases can be interpreted in several ways by users, but most frequently as an assumption that it is not available, which may be incorrect.
- The absence of inter-database links also inhibits the creation of aggregation & integration tools such as COLUMBA (the authors' tool).

Objectives

- Determine coverage of PDB entries by secondary resources (by % of total PDB, & by timing)
- Determine overlap of secondary sources by type (e.g., fold/family, sequence)
- Determine whether representative sets of structures show a better coverage by secondary annotation than the average PDB entry.
- Note: Some resources will never have 100% coverage due to the nature of their content (e.g., enzymes only).

Methods

- Examined out-links to PDB from 16 secondary protein resources
- Computed relationship between annotation presence and structure deposition date
- Overlap = $\sum_{i=1,j=1}^{16} \frac{D_i \cup D_j}{D_i \cap D_j}$ or $\frac{\bigcup_{i=1}^{16} D_i}{\bigcap_{i=1}^{16} D_i}$
- Overlap visualization constructed via a distance tree using UPGMA clustering (Unweighted Pair Group Method with Arithmetic Mean)

Results: Coverage

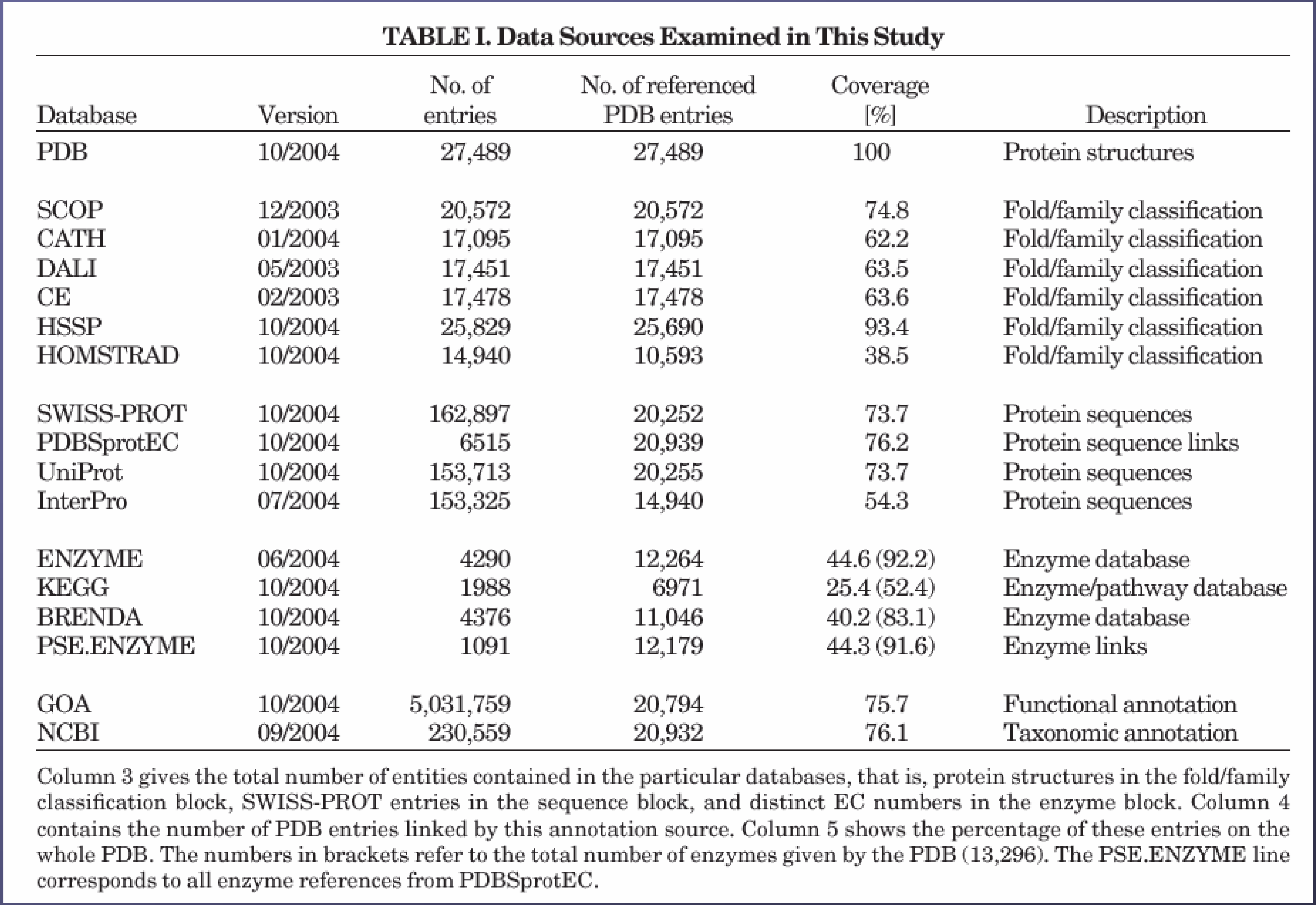
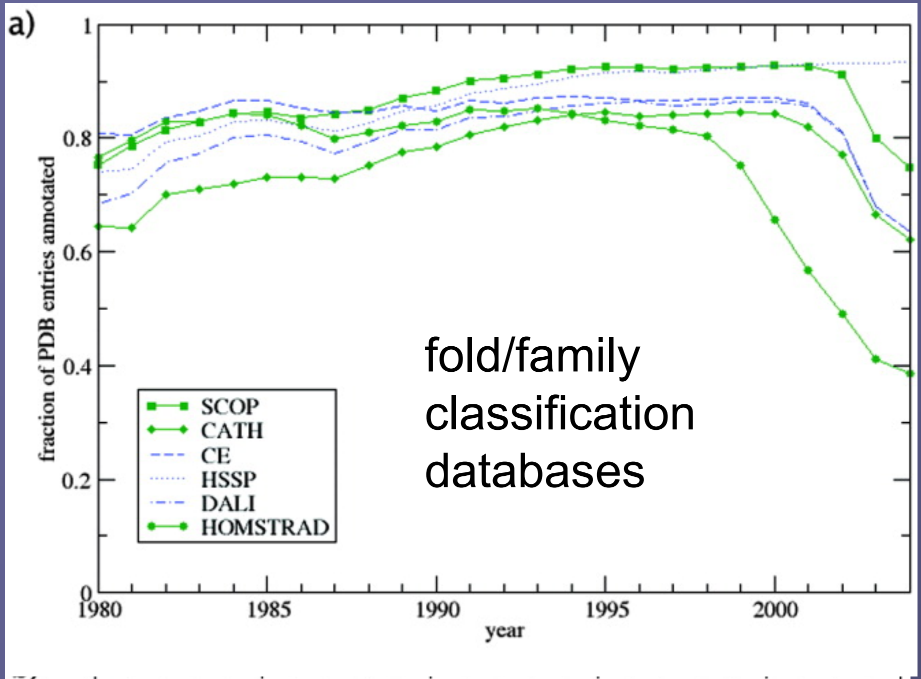


TABLE I. Data Sources Examined in This Study

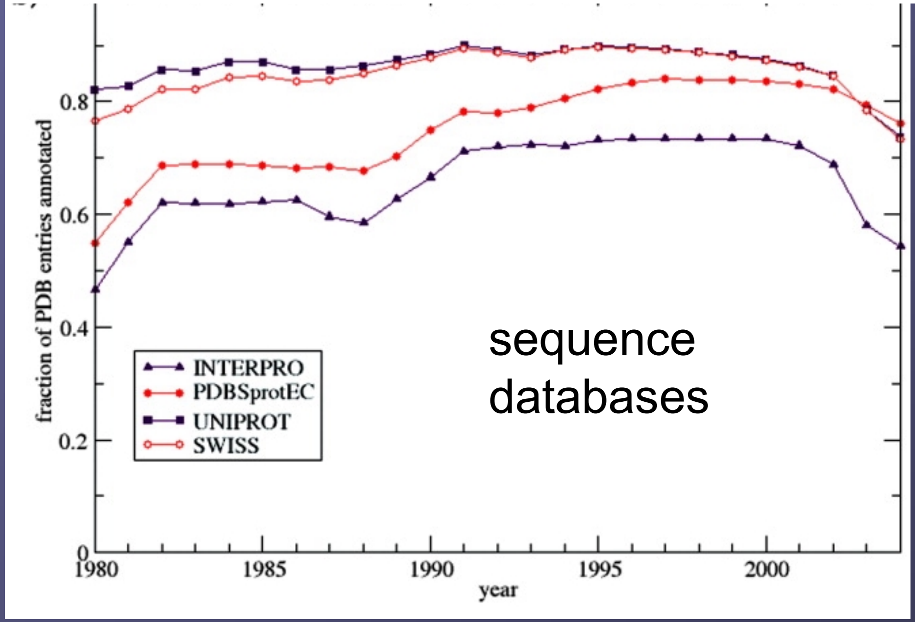
Database	Version	No. of entries	No. of referenced PDB entries	Coverage [%]	Description
PDB	10/2004	27,489	27,489	100	Protein structures
SCOP	12/2003	20,572	20,572	74.8	Fold/family classification
CATH	01/2004	17,095	17,095	62.2	Fold/family classification
DALI	05/2003	17,451	17,451	63.5	Fold/family classification
CE	02/2003	17,478	17,478	63.6	Fold/family classification
HSSP	10/2004	25,829	25,690	93.4	Fold/family classification
HOMSTRAD	10/2004	14,940	10,593	38.5	Fold/family classification
SWISS-PROT	10/2004	162,897	20,252	73.7	Protein sequences
PDBSprotEC	10/2004	6515	20,939	76.2	Protein sequence links
UniProt	10/2004	153,713	20,255	73.7	Protein sequences
InterPro	07/2004	153,325	14,940	54.3	Protein sequences
ENZYME	06/2004	4290	12,264	44.6 (92.2)	Enzyme database
KEGG	10/2004	1988	6971	25.4 (52.4)	Enzyme/pathway database
BRENDA	10/2004	4376	11,046	40.2 (83.1)	Enzyme database
PSE.ENZYME	10/2004	1091	12,179	44.3 (91.6)	Enzyme links
GOA	10/2004	5,031,759	20,794	75.7	Functional annotation
NCBI	09/2004	230,559	20,932	76.1	Taxonomic annotation

Column 3 gives the total number of entities contained in the particular databases, that is, protein structures in the fold/family classification block, SWISS-PROT entries in the sequence block, and distinct EC numbers in the enzyme block. Column 4 contains the number of PDB entries linked by this annotation source. Column 5 shows the percentage of these entries on the whole PDB. The numbers in brackets refer to the total number of enzymes given by the PDB (13,296). The PSE.ENZYME line corresponds to all enzyme references from PDBSprotEC.

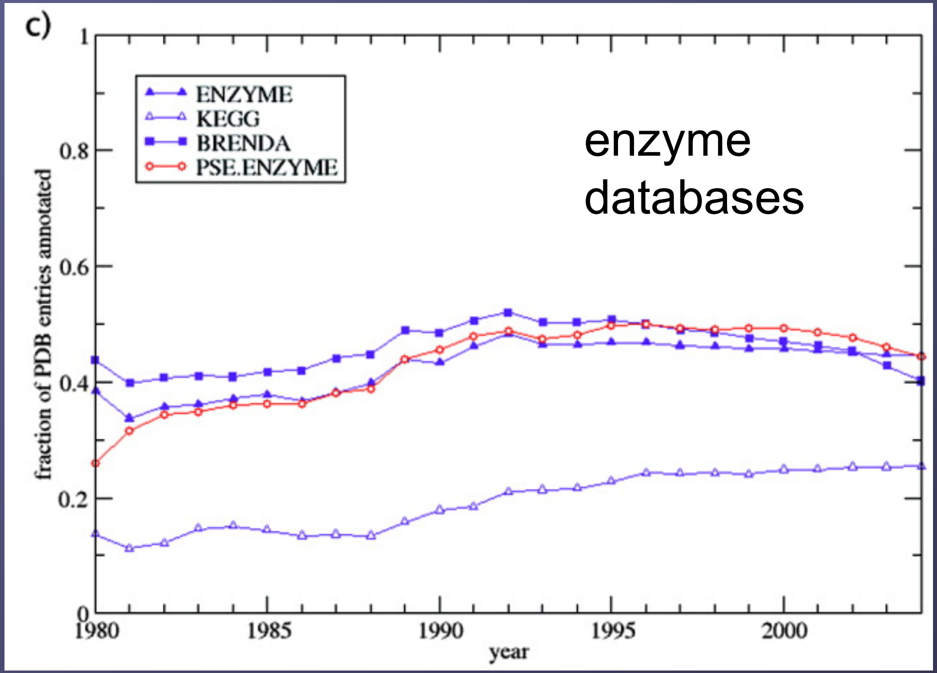
Results: Secondary annotation by year



fold/family classification databases



sequence databases



enzyme databases

Fig.1. Fraction of the PDB entries published up to a certain year for which links to a specific secondary-party database exists.The coverage over time is shown for (**a**) fold/family classification databases, (**b**) sequence databases, and (**c**) enzyme databases.During the 1970s, the PDB was very small and a large variations were observed. After 1997, many data sources for which the creation of cross-references requires manual interaction were not able to keep pace with the rapid growth of the PDB. The PSE.ENZYME curve corresponds to all enzyme references from PDBSprotEC.

Results: Data source overlap

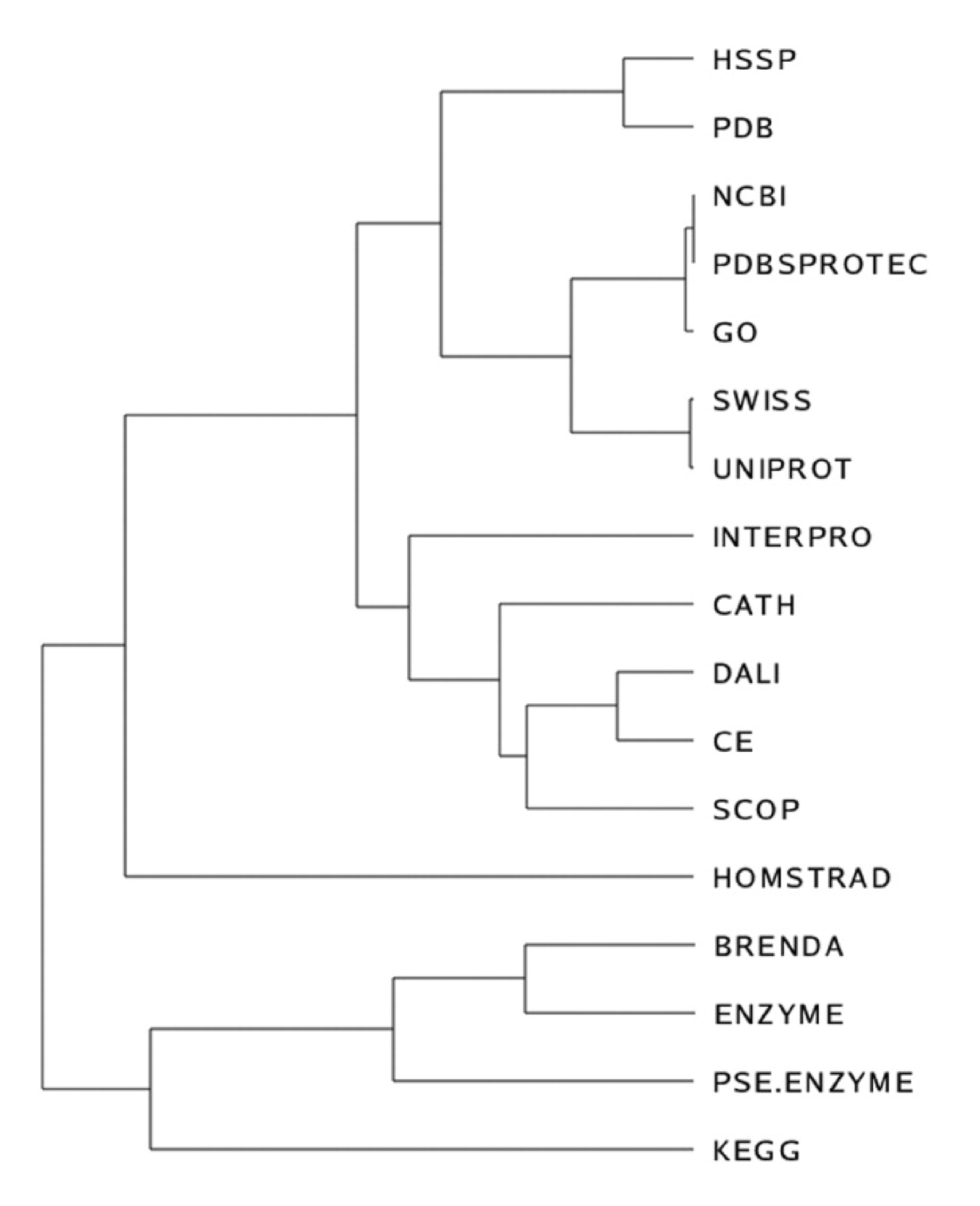


Fig.2. Degree of overlap between the sets of cross-referenced PDB entries for the 15 secondary-party databases we studied.The data sources are arranged in a tree computed using the UPGMA method.

Distances were calculated as the number of entries in the intersection divided by the union of 2 data sources.The best overlapping data sources were assigned a common node first. The branch lengths correspond to the calculated dissimilarity of 2 nodes.The PSE.ENZYME leaf corresponds to all enzyme references from PDBSprotEC.

Questions

- Need more details on:
  - How data was collected
  - How commonly-annotated structures were determined [574 col 2]
- How do we know that links exist from PDB to the extant secondary sources for the entries they annotate if the authors didn’t look at PDB out-links? (see 1) Analyzed Data: “Our reason for disregarding…” [572]; 2) Conclusions: “Roughly two-thirds…” [575])