Literature and its referents: Analyzing PubMed citations across PFAM

Richard K. Belew [§]
Robert Finn[‡]
Alex Bateman[‡]

[§] Computer Science & Engr. Dept., Univ. California – San Diego
[‡] Wellcome Trust Sanger Institute

ISMB 2002 - Poster

It is now common to make an explicit reference to a PubMed entry as part of various gene and protein annotations. This creates an explicit linkage between a structured data element and (its author's) verbal characterization of why this piece of data is scientifically important. In this poster we consider an analysis of reference pattern rather than of the text of individual articles. This analysis is motivated by early bibliometric analyses of citation patterns across the scientific literature, and more recent linkage analyses of WWW pages [1]. Figure 1 shows a common first analysis. Considering a corpus of approximately 600,000 TREMBL/SwissPROT protein entries, the number of references made to particular articles follows the ubiquitous Zipfian distribution. A few central references accrue exponentially many references.

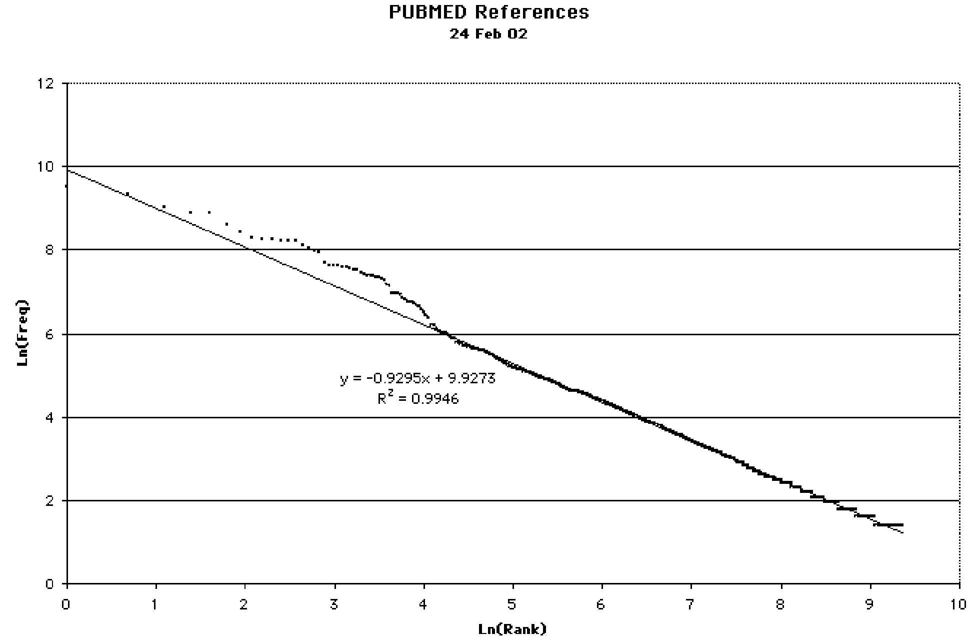

Figure 1: Zipfian distribution of PubMed references in TREMBL/SwissProt

The next step of the analysis is to consider the reference pattern of all that are part of the PFAM protein classification system. PFAM is a widely used effort to classify proteins, based on similar sequence structure, into one of approx. 3800 classes or "families" [2]. Often sequence similarity gives clues as to shared structural and functional features of class's of proteins as well. Our hypothesis here is that proteins sharing literature references might also share biological properties discussed in these papers and provide independent evidence concerning sequence-based classifications.

Further, it can be anticipated that proteins' references to PubMed will be due to a mixture of two quite different causes. Many, high-frequency, "generic" references are associated with proteins simply they are part of major, exhaustive sequencing efforts. Others will be to lower-frequency publications focusing on particular biological functions of proteins. By considering these patterns of reference with respect to the PFAM classification, we can expect the former to show nearly uniform reference across PFAM categories, while the latter can be expected to (at least in many cases) remain within a single category. A sample of the five most frequently-cited references is shown in Table 1.

Rank	Nref	PubMedID	Title	Cite (SO)
1	13751	10731132	The genome sequence of Drosophila melanogaster.	Science 2000 Mar 24;287(5461):2185-95
2	11458	7906398	2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegan	Nature 1994 Mar 3;368(6466):32-8
3	8465	11217851	Functional annotation of a full-length mouse cDNA collection	Nature 2001 Feb 8;409(6821):685-90
4	7281	8843436	A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb Streptomyces coelicolor A3(2) chromosome.	Mol Microbiol 1996 Jul;21(1):77-96.
5	7259	11214968	Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti	DNA Res 2000 Dec 31;7(6):331-8

Table 1: Most frequent PubMed references across TREMBL/SwissPROT

Repeating the basic rank/frequency analysis for the approx. 145000 PFAM-classified protein references produces a Zipfian distribution (not show), very similar to the one for all proteins shown above. Figure 2 shows a correlational analysis of the frequency-rank of documents' references, as a function of (the frequency rank of) how many different PFAM families contain a protein mentioning this reference. As expected, the approximately 80 most frequent, "generic" publications are indeed scattered across the most PFAM families. Beyond this threshold however, correlation diminishes considerably. The threshold is sharp enough that it may provide the basis for an operational definition of "generic" citation in this context, for example as part of a literature filter for human annotators searching the literature.

Questions also arise concerning exceptions this rule: What are those references which receive many citations but that are focused within a small number of PFAM? Conversely, what are those publications that are broadly distributed across many PFAM but which garner relatively few citations? The first category turns out to correspond to mutated proteins. These are of sufficient importance to receive great attention (e.g., HIV or hepatitis C viral coat proteins). Each is posted as a separate protein, but since its sequence is very similar to wild-type and other mutants, all are classified within the same PFAM family.

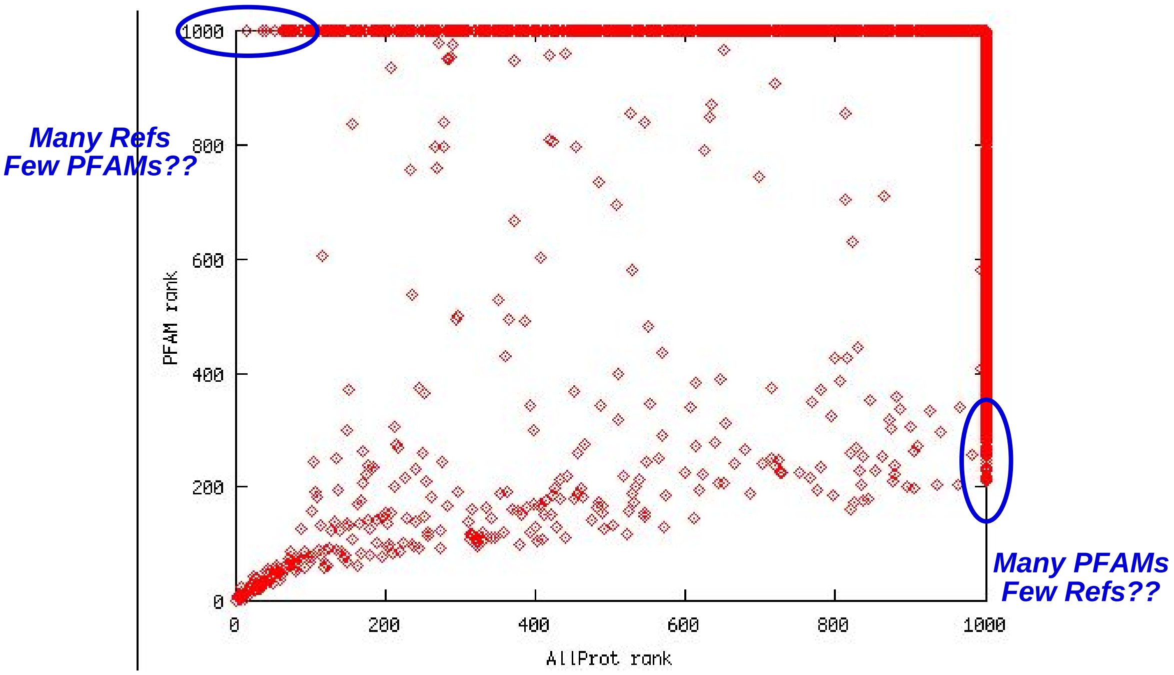

Figure 2: PubMed frequency rank vs. Number of PFAM families

The second type of exceptions (infrequent reference, broad PFAM coverage) seems due to two factors. First, the single protein may have a number of sub-unit enzymes (domains) that have individually been classified into different PFAM families (such as a ribosome), with individual papers discussing the functionality of every subunit. Alternatively, a series of proteins may have all been cloned from the same region (e.g., bacillus subtilils) and because they are all part of an important pathway, generate many publications; these various clones will often be in unrelated PFAM families.

The central point is that a relatively straight-forward analysis of purely statistical features of scientific publication patterns can be used to explore and perhaps infer semantic characteristics of the biological phenomena being investigated. Of course these analyses share the same basic statistics-to-semantics goals with, and can be performed as a compliment to, the analysis of individual publications' word features that are most common within the text retrieval community.

[1] **Finding Out About: A cognitive perspective on search engine technology and the WWW**. R. K. Belew. Cambridge Univ. Press, 2000.

[2] "The Pfam Protein Families Database" Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) **Nucleic Acids Research 30(1):276-280**