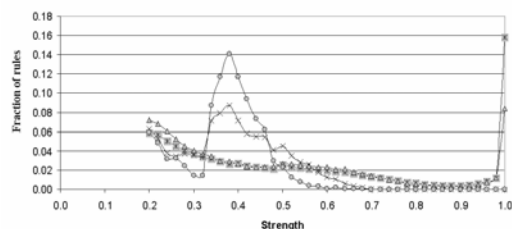


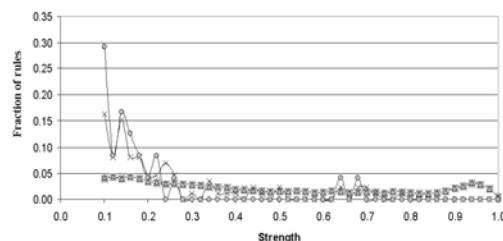
## Statistics of association rules in non-protein databases

Are rule strength distributions shown in Figures 1 and 2 of the *main text* characteristic reflect protein annotation databases, or is this a general property of all large databases? For comparison we investigated two non-protein databases obtained from the UCI Knowledge Discovery in Databases Archive (<http://kdd.ics.uci.edu/>). One of them, Forest CoverType database, contains 581012 entries and 54 attributes for which we defined 15958 items (in our analysis only binary variables indicating the presence or absence of every possible value of a given attribute), describing forest cover types for 30 x 30 meter cells obtained from US Forest Service Region 2 Resource Information System data. Using the *Apriori* algorithm we were able to extract from this database 375609 rules with strength values distributed as shown in Figure I. The shape of this curve is qualitatively similar to that obtained from Swiss-Prot, with an even more substantial fraction of strong rules. In particular, it appears that the database of forest cover types yields many perfect or nearly perfect rules, with just one or two exceptions, probably reflecting a more deterministic character of the underlying data. These data are the result of multiple consecutive measurements of the same type and are thus significantly less ambiguous than biological annotation. The random version of this database, generated using the same approach as for Swiss-Prot, also produced many medium-strength rules around 0.38 for the same reason as discussed above.



**Fig. I.** Distribution of the rule strength in a non-biological database (forest cover types) and the corresponding random database. The first two curves correspond to rules with minimal coverage 50 (—\*—) or 100 (—△—), the last two curves (—×—, —○—), to rules constructed from the random database (see *Methods* for details).

The second non-protein database we tested, that of the census bureau database (also known as “Adult” database) (Figure II), displayed a somewhat different behavior in that the peaks corresponding to strong and weak rules were much less pronounced and, in addition, much fewer rules with strength equal exactly to 1.0 were observed. This behavior may probably be explained by a much smaller database size (48842 instances with only 14 attributes and 135 items as opposed to 125642 entries with 9143 attributes in Swiss-Prot and 581012 entries and 15958 items in the database of forest cover types) and the resulting low (10906 for the real and merely 86 for random database) number of rules. The number of weak rules is low, presumably due to relatively small amount of combinatorial combinations possible in this small set of items. Notably, a randomly generated database of bureaucratic form demonstrates very unstable behavior of its statistics with respect to strength (see Figure II) due to the small number of rules: most peaks are formed by only two rules with the same item in their RHS.



**Fig. II.** Distribution of the rule strength in the database of filled bureaucratic forms and the corresponding random database. The first two curves correspond to rules with minimal coverage 50 (—\*—) or 100 (—△—), the last two curves (—×—, —○—), to rules constructed from the random database (see *Methods* for details).