Figure 1 A comparison of protein length distributions for *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *E. coli* and the archaebacterium *Methanococcus jannaschi*. Protein number, based on genome analysis, is plotted against protein length (number of amino acids), where length windows correspond to 50 amino acids. The total number of proteins analysed per genome (*n*), the average polypeptide length and the fraction of proteins exceeding a length of 500 amino acids (aa) are given. The last 1% of proteins (>2,000 aa) are not included. Data are derived from the following sources and represent essentially complete genomes, except for *C. elegans*, for which about 95% of the genome is sampled here: *C. elegans*, Batch Entrez NCBI protein database, http://www3.ncbi.nlm.nih.gov:80/cgi-bin/EntrezBatch/nph-batch/result; *S. cerevisiae*, FTP://darcy.bu.edu/pub/genomes/; *E. coli* and *M. jannaschi*, TIGR Microbial Database (MDB) of the Institute for Genomic Research, http://www.tigr.org:80/tdb/mdb/mdb.htlm. The following individuals are acknowledged for their help in performing this analysis: M. Adams and O. White (Institute for Genomic Research, Rockville, MD), T. Smith and J. Freeman (The BioMolecular Engineering Center, Boston, MA), and L. Hillier (Washington University, St Louis, MO).