Sequence Sizes of Eukaryotic Enzymes

Eugene Kolker and Edward N. Trifonov
Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

Abstract

We have shown in earlier studies that an appreciable fraction of proteins display sequence size periodicity with periods of ≈123 aa and ≈152 aa for eukaryotes and prokaryotes, respectively. For any firm conclusions to be made, the issue of possible bias due to an overabundance of some protein families should be addressed in more than one way. Here we present the size distributions for various sequence ensembles of eukaryotic enzymes that differ by level of data bank cleaning. The sequences were purged by applying several successive thresholds of relatedness irrespective of the sequence lengths. The previously observed preference to typical sizes is confirmed. Possible reasons for the observed excess of the typical size sequences are discussed.

Key words: eukaryotic enzymes, sequence length, typical size, cleaning, segments.

Introduction

In 1929 Svedberg [1] suggested that proteins could have standard molecular masses, multiples of a certain unit mass. Since then, this simple and attractive idea has been tested several times by different techniques on different data sets, along two main parallel lines: i) analysis of protein structures, and ii) analysis of primary sequences. When the structure of chymotrypsin was solved it was suggested that a sufficiently long polypeptide chain might be "piled on itself" or "folded around nuclei of highly stabilized local conformation" [2], thus making two distinct folding domains of that protein [3]. According to Matthews *et al.* [4], long polypeptide chains could be considered to be a combination of smaller independently folded substructures. On the basis of the limited data available at that time, they estimated that the typical masses of such structural units ranged between 10 to 16 kDa (≈ 90-150 aa). Concurrently, several attempts to find out possible regularities in the distributions of the protein sequence sizes were undertaken (e.g., [5-8]), but they all were based on nonrepresentative sequence sets. Our own recent reexamination of protein sequence length distributions [9, 10] revealed that the expected regularity in the protein organization indeed exists although in a rather weak form. We have shown that an appreciable fraction of proteins make a ladder with steps ≈123 aa and ≈152 aa for eukaryotes and prokaryotes, respectively. (We call these units "sequence segments" or "segments".)

The question of the possible reasons for such regularity is still open. One possibility is that this phenomenon is simply due to an overabundance of some protein families. Although we have already addressed this point earlier [9, 10], we are trying here to answer the question by presenting data at several levels of cleaning of an entire set of eukaryotic enzyme sequences (independent polypeptide chains). The cleaning procedure applied in this work involves only commonly used programs. It compares and purges similar sequences irrespective of their lengths. This not only eliminates any domination by one or another protein family, but also excludes any size preferences due to possible previously undetected gene duplications. There are several reasons why we choose eukaryotic enzymes. First, the sequence sample size of the set of eukaryotic enzymes is large enough to do such an analysis. Second, we have previously shown that for eukaryotic enzymes the size regularity is rather strong. The enzymes also show a less prominent peak at the typical unit size, unlike the total eukaryotic protein ensemble. The results of current studies clearly demonstrate that the observed segment-size regularity is of a universal nature, rather than the result of domination by some species or protein group.

Sequences and Methods

Eukaryotic enzyme sequences were obtained from a recent release of the Swiss-Prot protein sequence data bank [11]. The enzymes in our data set are those sequences which have an enzymatic code in their description. This simple definition of being an enzyme clearly overlaps largely with other more elaborate definitions, and thus involves a majority of the qualifying sequences. To construct sequence sets for the analysis of eukaryotic enzyme sequence lengths, we applied the following cleaning procedure. First, all irrelevant entries, such as fragments, unidentified reading frame sequences, viral, mitochondrial and chloroplast genome sequences, were removed. This reduced the data set from the original 7017 entries to 4391 eukaryotic enzymes with polypeptide chain lengths ≤600 aa. Then, we implemented the PURGE procedure provided by A. Neuwald, NCBI [12-14] to clean the sequence set. Based on the BLAST program [15], PURGE compares each sequence against all other sequences from the sequence set, calculates similarities between the sequences and, at a chosen threshold, constructs an output set of a maximal number of dissimilar sequences (for details see [12-14]). The PURGE threshold values are standard BLAST MSP (maximum segment pair) scores [15-17]. The PURGE mutational matrix is a standard Henikoff's BLOSUM 62 scoring matrix [18]. By applying progressively decreasing thresholds of similarity, the sequences were cleaned irrespective of their lengths.

For construction of unbiased sequence sets one could use one of generally known weighting schemes [19-21]. Our choice of weights 0 and 1 only makes the calculations simpler, although, perhaps, it is unfair to the signal. Such oversimplification (simple weights and purging of related sequences irrespective of their lengths) causes some reduction of the sample size, and, hence, of the signal-to-noise ratio as well. We used the PURGE [12-14] program at several different thresholds in order to check whether, and to what extent the observed size regularities would survive despite the excessive cleaning of the sequence set.

Smoothed distributions for protein lengths were obtained by calculating running averages with a 50-aa window. Fig. 1 represents normalized size distributions for four different sets of protein sequences, as specified in Table 1.

To evaluate typical periods in sequence lengths, we applied the cos-Fourier transform technique, based on the following equation:

$$A(j) = \frac{\sum_{i=1}^{n} H(i)\cos\left(\frac{2\pi i}{j}\right)}{\sum_{i=1}^{n} \cos^2\left(\frac{2\pi i}{j}\right)};$$

where H(i) is the oscillating part of the corresponding distribution; n - total number of points, in aa residues; i - the variable size in aa residues; j - a given period value; and A(j) - amplitude of cosine components for a given period. Oscillating components of the corresponding distributions were calculated by subtracting their smoothed versions from the actual distributions, with the averaging window equal to a given period [9, 10, 22].

Results

The length distribution of the initially cleaned set of eukaryotic enzymes (4391 sequences with lengths ≤600 aa) has four peaks (Fig. 1, curve A) at lengths approximately equal to a typical unit (segment) size and its multiples, as in our previous work [9, 10]. Assuming periodic distribution, we applied the cos-Fourier transform to the data in order to detect all possible periods (sizes). We confirmed that the most preferred unit size for eukaryotic enzymes is 125±5 aa (Table 1, set 1). In the amplitude spectrum (Fig. 2, curve A) along with the major maximum at ≈125 aa some smaller maxima (at ≈173 aa, ≈63 aa, ≈43 aa, ≈37 aa etc.) are also observed. This suggests that there could be some other types of protein sequence size regularity. The amplitudes and widths of smaller peaks, however, are insufficiently large to claim more than one preferred unit size, as was previously observed [9, 10].

We repeated the same procedure, using different thresholds of cleaning (Table 1, sets 2-7) with the same result each time: the most typical size of eukaryotic enzymes is ≈125 aa. We could not reduce the data set size below 30% of the original one simply because such a small number of sequences is no longer suitable for the above analysis. Notably, the main features of both the length distributions (Fig. 1, curves B-D) and the spectra (Fig. 2, curves B-D) remain unchanged irrespective of the degree of cleaning. That is, the sequence length distributions of eukaryotic enzymes have four peaks approximately at the typical unit size and multiples thereof (Fig. 1). Only the period ≈125 aa survives among all amplitude spectra as a main maximum (Fig. 2). All other peaks remain small compared to the main one. Most of the smaller maxima are not conserved (Fig. 2). (Along with the main maximum there are also three small peaks at ≈63 aa, ≈37 aa and ≈30 aa, which survived the cleaning, but to prove or disprove their validity much larger sequence ensembles would be required.) Altogether, these results demonstrate that the observed segment-size regularity of eukaryotic enzymes is *not* a result of bias at the primary sequence level, i.e., by domination of a few protein families and/or undetected gene duplications, but rather represents the general property of a protein sequence organization.

Discussion

The data presented here demonstrate that, largely independent of the degree of the protein sequence data bank purification, the distribution of the protein sequence sizes stays the same, with a period of $\approx$125 aa residues. When the data set is reduced to about 30% of its initial size, the periodicity becomes less obvious. The remaining sequences are maximally dissimilar and unrelated (see Table 1, set 7). The bulk of sequences (70%), those which clearly display periodicity, are interrelated, making many linked groups. According to our analysis [23], any large enough representative set of segment-size sequences can be separated into about 70 unrelated families covering 60-80% of the entire ensemble. Apparently, it is these different families which contribute most to the observed periodicity. This may indicate that proteins of the "right" lengths have had enough time to form families. Alternatively, there may be some modern selection pressure, perhaps of a structural nature, towards these specific protein sizes so that the "right" sizes have a better chance to develop families.

Apart from the possible domination by some protein families, which seems to be excluded by this and other [9, 10] studies, other plausible reasons for the observed underlying regularity in protein lengths include:

1) Potential size bias due to incomplete representation of genomes. At least in the near future this possibility cannot be excluded. We think, however, that this is unlikely, since the observed period does not depend on species, protein families or the degree of cleaning of the data bank;

2) Probable optimal protein folding pressure. We believe that the typical size for protein sequences could also be the typical size for protein structures. It is worth noting that the estimation of typical sizes of the most common structural folds comes very close to the segment size [24];

3) Possible involvement of DNA was discussed earlier [9, 10, 22, 25-27]. The

protein segment size might correspond to the optimal size for proposed ancient circular DNA elements; such DNA elements could have been fused (recombined), thus making longer genes, which resulted in longer proteins. The stronger regularity of the sequence sizes for conserved proteins (e.g. eukaryotic enzymes) perhaps implies that this phenomenon is ancient in origin. Interestingly enough, our recent analysis of amino acid distributions along the segment-size eukaryotic sequences showed that only methionine residues exhibit the typical size recurrence [22]. In other words, excess of methionines is observed at multiples of the segment size along the sequences. This could correspond to the hypothetical borders between the internal segments [22, 10]. Such excess methionines could be fossils of the early fusion (duplication or shuffling) of the segment-size genes, each starting with the initiation triplet (9, 10, 22, 25-27);

4) Exons could cause some size regularity, but typical exon sizes ($\approx$35-45 aa) do not correspond to the segment sizes (however, the spectra in Fig. 2 have a relatively small, but conserved peak at about the typical exon size ($\approx$37 aa), which, perhaps, hints at a possible exon-size regularity);

5) The observed segment size could be a result of modulation of other, smaller, sequence sizes (e.g., $\approx$30 aa and/or $\approx$63 aa). Unfortunately, not enough material is available to test this or the exon hypothesis. The size regularity in general could be due to one or another of the above pressures or could be a result of the combined action of different pressures both of divergent and convergent nature.

Acknowledgments

We are grateful to A. Girshovich, B. Henley, I. Ioshikhes, L. Lvovsky, S. Scharf, A. Schmitt, A. Weinberg, and E. Yakobson for a critical reading of the manuscript, to I. Ioshikhes for his technical assistance, to J. Prilusky for his help, and to all who made valuable comments to one of the authors (E.K.) during his recent visits in Europe and in the U.S.A. This work was partly supported by The Genome Committee of The Weizmann Institute.

References

1. Svedberg, T.: Nature 123, 871 (1929).
2. Matthews, B. W., Sigler, P. B., Henderson, R., and Blow, D. M.: Nature 214, 652-657 (1967).
3. Birktoft, J. J., and Blow, D. M.: J. Mol. Biol. 68, 107-112 (1972).
4. Matthews, B. W., Jansonius, J. N., Colman, P. M., Schoenborn, B. P., and Dupourque, D.: Nature 214, 37-41 (1970).
5. Svedberg, T.: Nature 139, 1051-1062 (1937).
6. Kieh, E. D., and Holland, J. T.: Nature 226, 544-545 (1970).
7. Sommer, S. S., and Cohen, J. E.: J. Mol. Evol. 15, 37-57 (1980).
8. Savageau, M. A.: Proc. Natl. Acad. Sci. USA 83, 1198-1202 (1986).
9. Berman, A. L., Kolker, E., and Trifonov, E. N.: Proc. Natl. Acad. Sci. USA 91, 4044-4047 (1994).
10. Kolker, E., and Trifonov, E. N.: Modelling of Biomolecular Structures and Mechanisms, 461-471, eds. Pullman, A., *et al.*, Kluwer Acad. Publish. (1995).
11. Bairoch, A., and Boeckmann, B.: Nucleic Acids Res. 21, 3093-3096 (1993).
12. Neuwald, A. F., and Green, P.: J. Mol. Biol. 239, 698-712 (1994).
13. ftp ncbi.nih.gov; cd pub/neuwald/asset.
14. Neuwald, A. F., Liu, J. S., and Lawrence, C. E.: Protein Sci. 4, 1618-1632 (1995).
15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.: J. Mol. Biol. 215, 403-410 (1990).
16. Altschul, S. F.: J. Mol. Biol. 219, 555-565 (1991).
17. Karlin, S., and Altschul, S.F.: Proc. Natl. Acad. Sci. USA 87, 2264-2268 (1990)
18. Henikoff, S., and Henikoff, J. G.: Proteins 17, 49-61 (1993).
19. Altschul, S. F., Carroll, R. J., and Lipman, D. J.: J. Mol. Biol. 207, 643-653 (1989).
20. Sibbald, P. R., and Argos, P.: J. Mol. Biol. 216, 813-818 (1990).

21. Vingron, M., and Sibbald, P. R.: Proc. Natl. Acad. Sci. USA 90, 8777-8781 (1993).
22. Kolker, E., and Trifonov, E. N.: Proc. Natl. Acad. Sci. USA 92, 557-560 (1995).
23. Kolker, E., Schmitt, A., and Trifonov, E. N., in preparation.
24. Kolker, E., and Trifonov, E. N., in preparation.
25. Trifonov, E. N.: J. Mol. Evol. 38, 543-546 (1994).
26. Trifonov, E. N. : Modelling of Biomolecular Structures and Mechanisms, 473-479, eds. Pullman, A., *et al.*, Kluwer Acad. Publish. (1995).
27. Trifonov, E. N.: J. Mol. Evol. 40, 337-342 (1995).

Table 1. Sequence unit sizes for eukaryotic enzymes.

No. of set	No. of sequences	Threshold	% compared to set 1	Calculated average unit length, aa*
1	4391	NO	100	125 ± 5
2	3854	2000	87.8	125 ± 5
3	3332	1500	75.9	124 ± 6
4	2714	1000	61.8	125 ± 6
5	2322	750	52.9	126 ± 7
6	1831	500	41.7	124 ± 7
7	1292	250	29.4	126 ± 8

* Highest maximum in the amplitude spectrum.

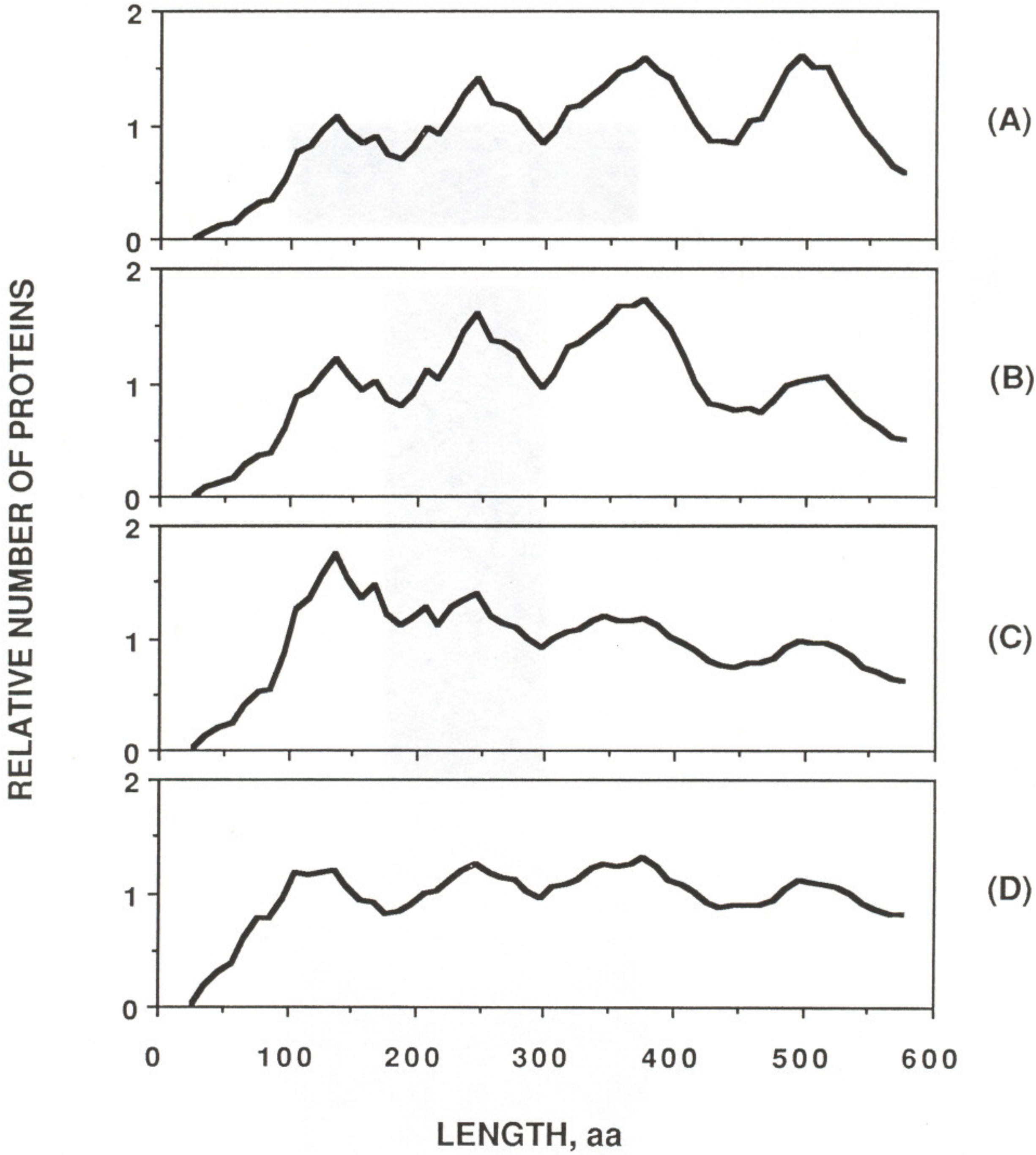

Fig. 1. Sequence length distribution of eukaryotic enzymes. The curves correspond to four cleaning levels (see Table 1): (A) the complete set, (B) 88%, (C) 62%, (D) 42% remaining. The smooth curves represent the running averages with a 50-aa window.

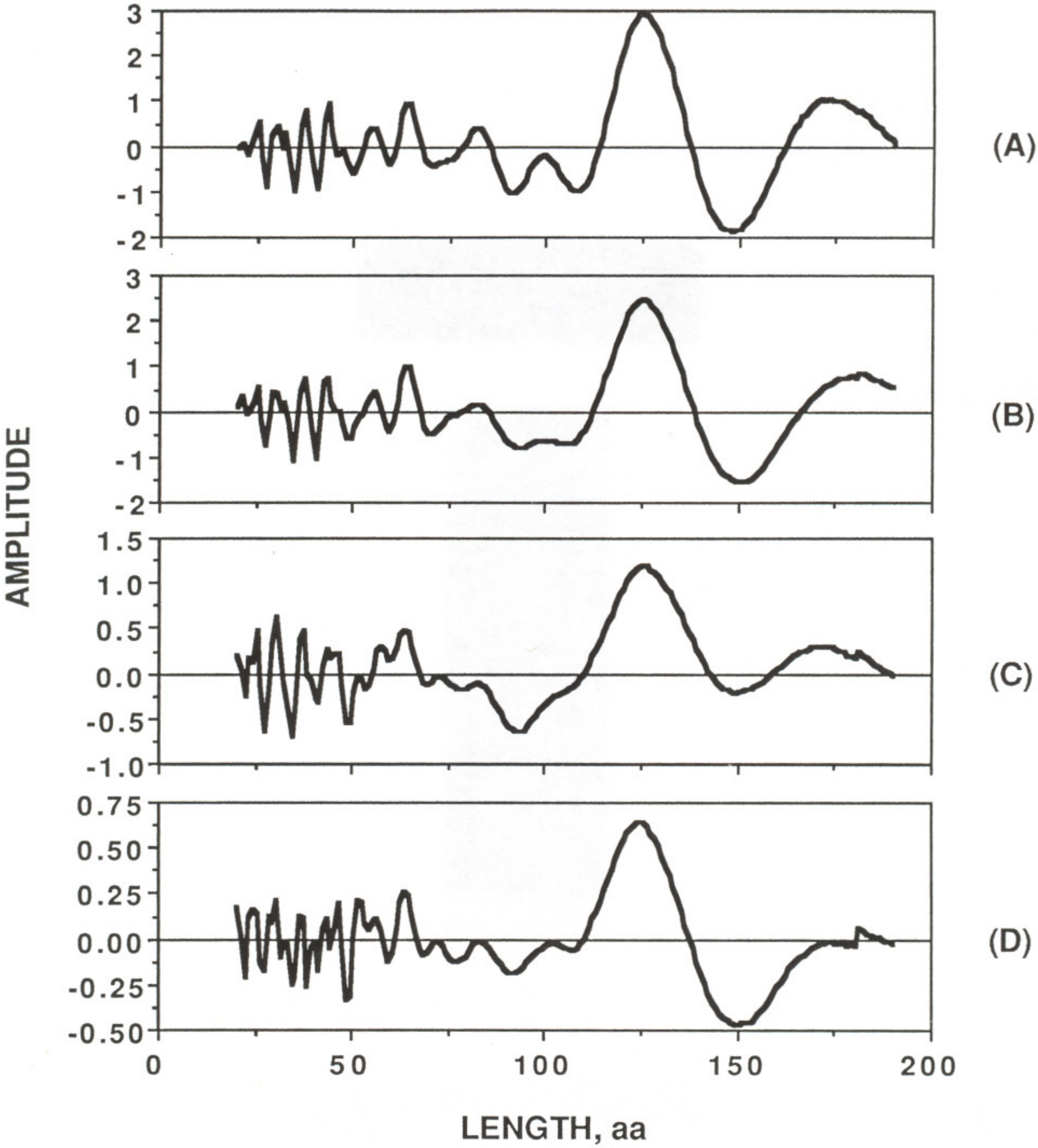

Fig. 2. Amplitude spectrum of the sequence length distributions of eukaryotic enzymes. The curves correspond to four cleaning levels: (A) the complete set, (B) 88%, (C) 62%, (D) 42% remaining.