Journal of Chromatography, 193 (1980) 1–8
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHROM. 12,647

SIZE AND CHARGE DISTRIBUTION OF MACROMOLECULES IN LIVING SYSTEMS

ELISABETTA GIANAZZA and PIER GIORGIO RIGHETTI
Department of Biochemistry, University of Milano, Via Celoria 2, Milan 20133 (Italy)

(First received December 19th, 1979; revised manuscript received January 7th, 1980)

SUMMARY

We report the statistical distributions of size and charge in protein molecules in living systems. The data have been plotted from more than 500 entries. 57% of the native proteins have molecular weights between 25,000 and 150,000 daltons. In oligomeric proteins, 71% of the subunits have molecular weights in the range 10,000–60,000. As for the subunit stoichiometry in oligomeric proteins, 50% of all the possible structures are dimeric, 30% tetrameric and 8% hexameric. More than one third of the proteins (38%) have p*I* values within 1.5 pH unit (pH 4.5–6.0). 70% of the total proteins have p*I* values below pH 7 and 30% above it. On the basis of these data, guidelines are given for a purification strategy when using SDS-electrophoresis and isoelectric focusing.

INTRODUCTION

It is generally stated in the literature that "many proteins having molecular weights (MW) above 36,000 contain two or more polypeptide chains"[1] or that "the great majority of proteins with molecular weights above 50,000 are composed of subunits rather than a single chain"[2]. In regard to the charge of proteins, it is reported that "most globular proteins have isoelectric points (p*I*) between pH 4.5 and 6.5"[1]. However, we have been unable to find quantitative data on the distributions of size and charge of protein molecules in living systems. This might have been due to the paucity of such data in the biochemical literature of past years. However, with the advent of gel filtration[3], sodium dodecyl sulphate (SDS)-electrophoresis[4] and isoelectric focusing (IEF)[5,6], data on MWs, p*I* values and quaternary structure of proteins have been rapidly accumulating. A very comprehensive table on the subunit composition of proteins, listing more than 500 entries, has been compiled by Darnall and Klotz[7]. A similar table, listing protein p*I* values, as determined by IEF, has been compiled by us[8]. We have used these data to calculate the distribution frequencies of MWs and p*I* values of proteins.

RESULTS AND DISCUSSION

MW distribution of native proteins

Fig. 1 lists a total of 530 MWs[7]. Each bar in the graph spans 25,000 daltons. In this distribution we have distinguished three peaks, each comprising three bars. By far the largest peak, which includes 40% of the proteins listed, occurs in the MW range 50,000–125,000. If we include in this group also the two neighbouring bars, we find that 57% of the proteins have MWs between 25,000 and 150,000 daltons. The second peak, representing 18% of the total proteins, appears in the MW range 175,000–250,000. A third peak, comprising 8% of the total proteins, occurs in the MW range from 325,000 to 400,000 daltons. 83% of the total proteins fall into these three groups. We have noticed that, while the distribution is continuous throughout most of the MW range considered (the minimum in some bars being 0.2%) there are surprisingly, two major gaps: one in the region 600,000–650,000 and another in the region 700,000–775,000. We don't know whether there is any reason for this, or if it is only due to lack of data. We stress also that, while the general distribution pattern holds true, the first bar (5000–25,000) is definitely underestimated, since this graph include only proteins having a quaternary structure.

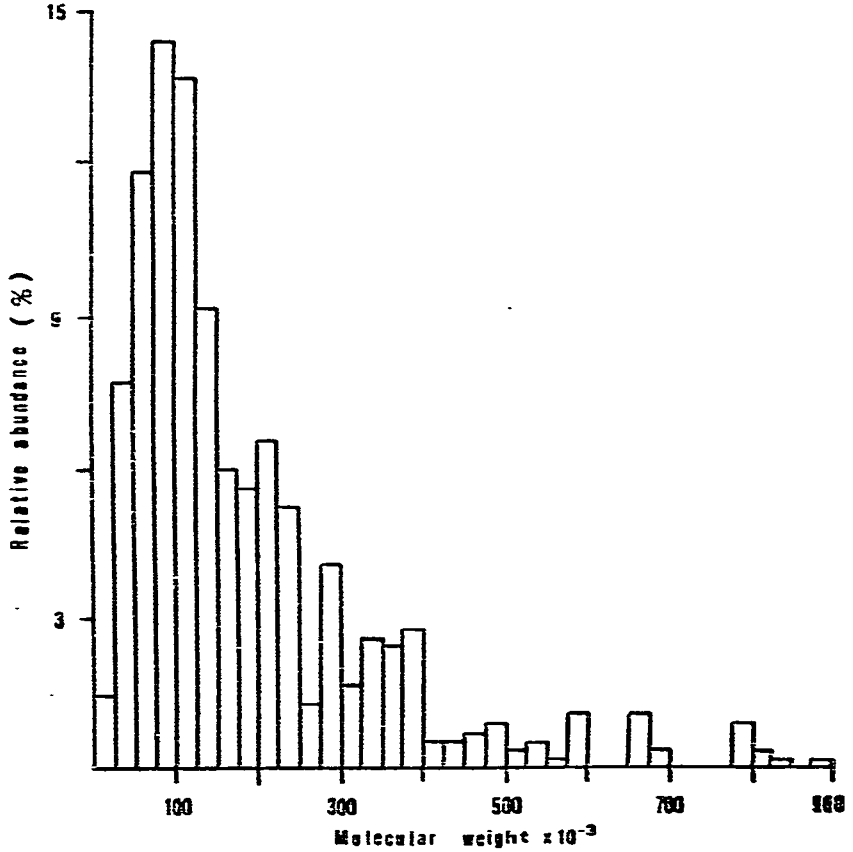

Fig. 1. Size distribution of native proteins. The relative percentages have been calculated from 530 entries. Each bar spans 25,000 daltons.

Quaternary structure

That proteins generally might be constituted of subunits seems to have first been suggested by Svedberg[9]. However, interest in the quaternary structure of proteins was mainly confined to the realm of physical biochemistry until it was realized[10–14] that broad aspects of cellular control mechanisms and the regulation of enzyme activity might operate at the molecular level through interactions between subunits of oligomeric macromolecules.

Fig. 2 gives the size distribution of subunits in oligomeric proteins. The total number of entries is still 530; each bar spans 10,000 daltons. It can be seen that 47% of the total polypeptide chains have MWs between 30,000 and 60,000 daltons, and that 71% of the subunits are in the range 10,000–60,000. Above 60,000 daltons there is a rapid decrease in frequency, until another small peak (7% of the total protein) is found centered around 100,000. We believe that this size distribution holds true not only for subunits of oligomeric proteins, but also, in general, for any polypeptide

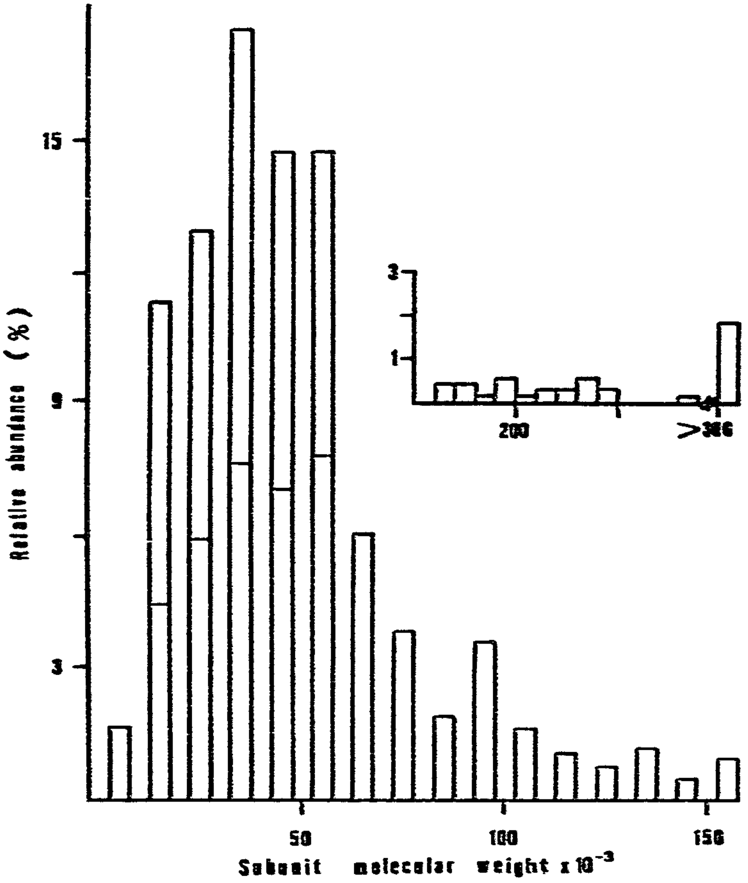

Fig. 2. Size distribution of polypeptide chains in oligomeric proteins. The relative abundances has been calculated from 530 entries. Each bar spans 10,000 daltons. The insert continues the distribution from 150,000 to 300,000 daltons. All the chains above this value have been grouped in a single bar. The five bars from 10,000 to 60,000 have been subdivided into groups of 5000 daltons.

chain made by living organisms. This same size distribution has been found, for instance, in chloroplast membrane proteins[15] as well as in rat liver mitochondrial[16,17], microsomal[17] and rough and smooth endoplasmic reticulum[18] membrane proteins. Perhaps one exception is the red cell membrane, where two exceptionally large chains (240,000 daltons) constitute 25% of the total membrane proteins[19].

Subunit stoichiometry in oligomeric proteins

The two most common models of subunit function and interaction, the concerted transition model (or symmetry, or all-or-none model) of Monod *et al.*[13] and the sequential (or induced fit) model of Koshland *et al.*[14], have been described for tetrameric proteins. The same applies to older models by Adair[20] and Pauling[21]. Even though these models could be modified to fit other subunit stoichiometries, it is of interest to know the frequencies of subunit combinations in oligomeric proteins. Klotz *et al.*[22], using a pool of 110 entries, had already calculated that 40% of the oligomeric proteins exist as dimers and 34% as tetramers. We have recalculated this distribution using their much larger set of data[7,8]. As shown in Table I, the conclusions of these authors generally still hold true. Dimeric proteins represent 50% and tetrameric structures 30% of the total. The two combined amount to 80% of all the total possible stoichiometries. Trimeric proteins comprise only 5% and pentameric 1% of the total. The next highest frequency is represented by hexamers (8%). There is only one known protein with seven and none with nine subunits. By far the largest body is represented by proteins with an even number of subunits (88%). If we exclude complex structures such as viruses, there are very few proteins with more than 12 subunits (2% of the total).

TABLE I

SUBUNIT STOICHIOMETRIES IN OLIGOMERIC PROTEINS

Subunit No.	*No. of proteins with given stoichiometry**	%*	*No. of proteins with given stoichiometry***	%**
2	269	49.4	44	40
3	27	5.0	6	5
4	159	29.2	37	34
5	6	1.1	2	1.8
6	40	7.3	8	7.3
7	1	0.2	—	—
8	14	2.6	5	4.5
9	—	—	—	—
10	4	0.7	4	3.6
12	13	2.4	4	3.6
14	2	0.4	—	—
16	2	0.4	—	—
24	3	0.5	—	—
48	2	0.4	—	—
60	1	0.2	—	—
162	1	0.2	—	—

* Calculated from a total of 530 entries.

** Calculated from a total of 110 entries (see ref. 22).

pI distribution

Fig. 3 lists a total of 500 entries from ref. 8. Each bar in the graph spans 0.5 pH units. It can be seen that more than a third of the proteins (38%) are grouped within 1.5 pH units (pH 4.5–6.0). If we now consider all the proteins having p*I* values below pH 7 and those having greater values, we see that 70% of the proteins appear in the first class (acidic proteins) and 30% in the second class (basic proteins). Thus it appears that all living systems have evolved in such a way as to have, at physiological pH, most of their proteins present and functioning in their anionic form.

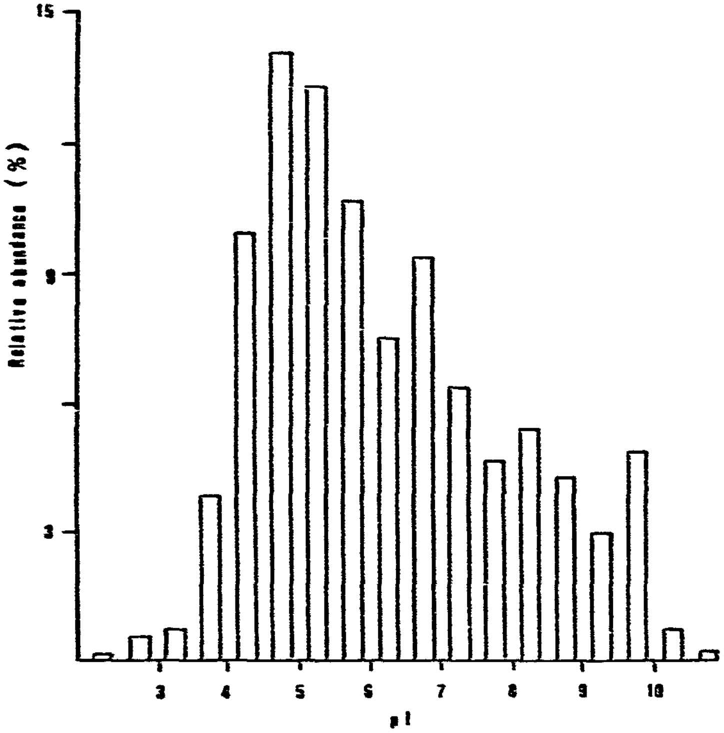

Fig. 3. p*I* distribution of proteins. The graph lists 500 entries. Each bar spans 0.5 pH units, from pH 2 to pH 11.

We wondered whether these data could be significantly altered by using a larger number of entries. We have thus redrawn Fig. 3 by adding also the p*I* values of minor isozyme components which had not originally been taken into consideration. Fig. 4 represents the p*I* distribution calculated from 800 entries. It can be seen that, while the height of some of the bands is altered, the overall picture is still the same. In this case, 36% of the proteins have p*I* values in the pH region 4.5–6.0, 71% of the proteins falling in the acidic group and 29% in the basic group.

CONCLUSIONS

The statistical data we have collected in the present work are not merely a

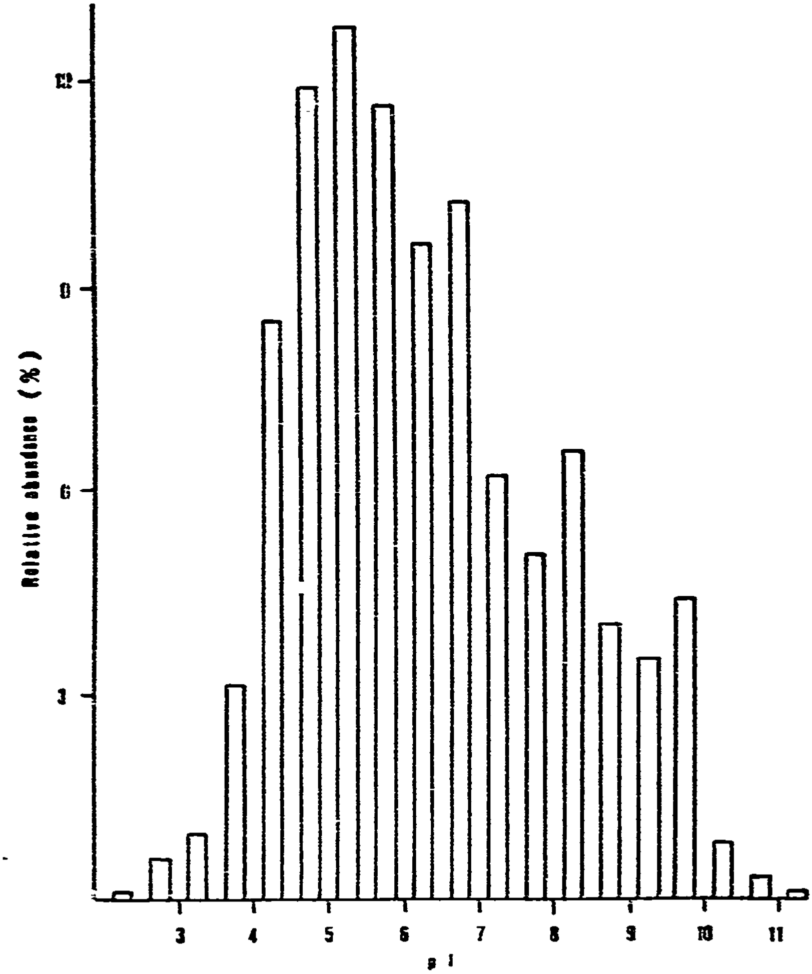

Fig. 4. p*I* distribution of proteins, as Fig. 3 but with 800 instead of 500 entries.

biochemical curiosity, but may represent the starting point for optimizing the purification of macromolecules. For instance, in the case of SDS-electrophoresis, since most polypeptide chains (71 %) have MWs in the low to medium range (10,000–60,000 daltons), the best separations should be achieved by running experiments in concave exponential acrylamide gel gradients, which optimize resolution in this MW range, rather than in gels of fixed acrylamide concentration or in linear gel gradients.

Even more dramatic is the case of separations performed by IEF, since *ca.* 40 % of all possible phenotypes are crowded into a narrow portion of the pH gradient, which spans only 1.5 pH units (pH 4.5–6.0). At present, eight narrow Ampholine pH intervals are commercially available from LKB (pH 2.5–4, 3.5–5, 4–6, 5–7, 6–8, 7–9, 8–10 and 9–11). While these narrow ranges are equally distributed along the pH scale, the proteins are not, since probably *ca.* 50 % of all possible macromolecules will focus in only one of the eight narrow intervals, *i.e.*, pH 4–6. Therefore, in order to maintain the same resolving power, the resolution of this narrow pH range should be increased by a factor of at least five, compared to other pH intervals, especially in the alkaline zone. This could be achieved by synthesizing the carrier ampholytes so as to have a preponderance of acidic species in the synthetic mixture. The pH 4–6 interval, thus enriched, could then be subfractionated into four narrower pH intervals, each spanning $\frac{1}{2}$ pH unit, which could then be made commercially available to the scientific community. Indeed, there are hints that things are moving in this direction. Thus Charlionet *et al.*[23] have recently described a method for the synthesis of highly diversified carrier ampholytes, which can be subfractionated in the pH 4–6 region in very narrow pH cuts, encompassing only 0.6 pH units (pH 4.3–4.9) or as little as 0.3 pH units (pH 4.4–4.7). In presence of these very shallow pH gradients the resolving power could be increased from the limit of 0.02 units (in p*I* differences between two proteins) as described by Vesterberg and Svensson[24], to as little as 0.001 pH units, an amazing twenty fold increase in resolution. Even though these concepts were applied to the very difficult separation of α_1-antitrypsin phenotypes, they hold generally true for all separations in the acidic pH range, as we have recently demonstrated also in the case of IEF of peptides[25,26].

Finally, it should be stated that the present-day practice of characterizing a new protein simply by MW and p*I* in a two-dimensional macromolecular map (for a review see ref. 27) may not be completely reliable. In fact, in the light of our data (see Figs. 2 and 3), it can be safely concluded that, in the case of p*I* determinations, a p*I* value can unequivocally be assigned to a given protein only if it is accurate to at least the third decimal place. Even with the most sensitive pH meters presently available, it is doubtful whether the second decimal place in a rending has any statistical significance. Moreover, the p*I* values given in the literature[8] are already uncertain in the first decimal place, due to persistent "malpractice" in the field (focusing at 4°, pH readings at 20–24°, data uncorrected for the presence of urea or other disaggregating agents, non-attainment of equilibrium conditions).

ACKNOWLEDGEMENTS

We thank Dr. M. Lederer for helpful hints and criticism. E. Gianazza is a fellow of the National Medical Research Council (CNR, Rome, Italy). Supported in part by grants No. 78.02249.04 and 78.01533.06 from Consiglio Nazionale delle Ricerche.

REFERENCES

1 A. L. Lehninger, *Biochemistry*, Worth Publishers, New York, 2nd ed., 1975, pp. 58 and 164.
2 H. R. Mahler and Y. Cordes, *Biological Chemistry*, Harper & Row Publishers, New York, 2nd ed., 1971, pp. 96 and 173.
3 L. Fischer, *An Introduction to Gel Chromatography, Vol. 1*, in T. S. Work and E. Work (Editors), *Laboratory Techniques in Biochemistry and Molecular Biology*, North-Holland, Amsterdam, 1969.
4 A. L. Shapiro, E. Vinuela and J. V. Maizel, *Biochem. Biophys. Res. Commun.*, 28 (1967) 815.
5 H. Svensson, *Acta Chem. Scand.*, 15 (1961) 325.
6 H. Svensson, *Acta Chem. Scand.*, 16 (1962) 456.
7 D. W. Darnall and I. M. Klotz, *Arch. Biochem. Biophys.*, 166 (1975) 651.
8 P. G. Righetti and T. Caravaggio, *J. Chromatogr.*, 127 (1976) 1.
9 T. Svedberg, *Nature (London)*, 123 (1929) 871.

10 J. C. Gerhart and A. B. Pardee, *J. Biol. Chem.*, 237 (1962) 891.
11 J. Monod, J. P. Changeaux and F. Jacob, *J. Mol. Biol.*, 6 (1963) 306.
12 J. C. Gerhart and H. K. Schachman, *Biochemistry*, 4 (1965) 1054.
13 J. Monod, J. Wyman and J. P. Changeaux, *J. Mol. Biol.*, 12 (1965) 88.
14 D. E. Koshland, Jr., G. Némethy and D. Filmer, *Biochemistry*, 5 (1966) 365.
15 F. Henriques, W. Vaughan and R. Park, *Plant Physiol.*, 55 (1975) 338.
16 J. H. J Tsai, H. Tsai and G. von Ehrenstein, *FEBS Lett.*, 10 (1970) 276.
17 C. A. Schnaitman, *Proc. Nat. Acad. Sci. U.S.*, 63 (1969) 412.
18 D. N. Hinman and A. H. Phillips, *Science*, 170 (1970) 1222.
19 G. Guidotti, *Annu. Rev. Biochem.*, 41 (1972) 731.
20 G. S. Adair, *J. Biol. Chem.*, 63 (1925) 529,
21 L. Pauling, *Proc. Nat. Acad. Sci. U.S.*, 21 (1935) 186.
22 I. M. Klotz, N. R. Langerman and D. W. Darnall, *Annu. Rev. Biochem.*, 39 (1970) 25.
23 R. Charlionet, J. P. Martin, R. Sesboüé, P. J. Madec and F. Lefebvre, *J. Chromatogr.*, 167 (1979) 89.
24 O. Vesterberg and H. Svensson, *Acta Chem. Scand.*, 20 (1966) 820.
25 P. G. Righetti and F. Chillemi, *J. Chromatogr.*, 157 (1978) 243.
26 E. Gianazza and P. G. Righetti, *Protides Biological Fluids, Proc. Colloq.*, (1980) 715.
27 E. Gianazza and P. G. Righetti, in P. G. Righetti, C. J. Van Oss and J. W. Vanderhoff (Editors), *Electrokinetic Separation Methods*, Elsevier/North-Holland, Amsterdam, New York, 1979, pp. 293–311.