Digital Computer Analysis of Electrophoresis and Ultracentrifugation Patterns by a Sum of Gaussian-like Distributions

E. Tränkle

Institut für Theoretische Physik, Freie Universtät Berlin, Berlin

(Z. Naturforsch. **30 c**, 311–317 [1975]; received January 9/February 17, 1975)

Computer Analysis, Electrophoresis, Ultracentrifugation

A Gaussian-like distribution function with three additional parameters is introduced which describes well the electrophoresis patterns of albumin, prealbumin and transferrin. Electrophoresis and ultracentrifugation patterns with 10–15 overlapping peaks are analyzed by means of the FORTRAN-program DIANA. One obtains the (relative) areas, positions and widths of the peaks. The analysis of a series of patterns proceeds in an automated way after the number of molecular components as well as starting values of the positions and the widths have been chosen in a test period.

1. Introduction

It has been argued that about twelve major serum components constitute more than 95% of the serum protein mass [1]. The rest of the mass consists of some twenty plasma proteins, occurring in such low concentration, that they do not significantly influence the electrophoretic plasma pattern. In filter-paper electrophoresis only five or six peaks are observed, whereas in the case of less adsorptive media, such as agar gel or cellulose acetate as many as eight or nine peaks show up. Moreover, in the regions of the dominant fractions the scanning curve has a gaussian-like shape. Some of the peaks have, however, a very large width and a non gaussian-like shape, which suggests that they could actually be a superposition of two or three strong overlapping serum components.

Clinical applications of electrophoresis are based on the fact that a disease causes a characteristic change of the concentration of several or most of its major serum components with respect to its normal values. For filter-paper electrophoresis normal concentration values as well as values for special diseases are known. These values must certainly be corrected and completed in the case of agar gel electrophoresis. As far as we know a systematical approach to determine the concentration values of all twelve major serum components, which probably is also important for the improvement of diagnosis, is still missing.

Requests for reprints should be sent to Dr. E. Tränkle, Institut für Theoretische Physik, Freie Universität Berlin, *D-1000 Berlin 33*, Arnimallee 3.

The concentration values of the serum components are obtained from the scanned electrophoresis pattern by evaluating the areas of all molecular fractions. To this end the scanning curve is usually cut off at the minima or Gaussians are hand-drawn to the pronounced maxima of the curve. Then the areas are measured by counting squares or by means of an automatic integrator. In this way reliable results are obtained only, if the peaks are well separated by minima. A higher accuracy of pattern evaluation can certainly be reached with an analogue computer: First Gaussians are generated by the computer. Their parameters are successively adjusted, until the superposition of the Gaussians fits the scanning curve. Then the areas are computed from the values of the adjusted parameters. By this method one could also determine the areas of adjacent fractions which are not separated by a minimum. However, we doubt that the reproducibility of the results is always guaranteed. This problem may become serious, if results of different authors are compared.

Using a digital computer the values of the parameters are adjusted by minimizing the sum χ of squares of differences. The main advantages of this computer method are the greater precision of the evaluation and the reproducibility of the parameter adjustment. Moreover one obtains printed information about the goodness of fit: χ is a measure of the overall accuracy, whereas the local accuracy can be judged from the remaining difference between the observed curve and the adjusted distribution.

The great precision of the evaluation is of little use if the mathematical function to describe the distribution is not carefully chosen. Since the conditions in agar gel electrophoresis are only approximately those of ideal one-dimensional diffusion, deviations from the Gaussians should be taken into account by such a function.

In section 2 we list the most important physical effects which may cause a deviation from the Gaussian. Then we introduce a Gaussian-like function with three additional parameters C_i, which have a simple interpretation: C_0 is the scale of the deviation, C_1 and C_2 are the strengths of the asymmetrical and symmetrical deviation, respectively. By analyzing some patterns of albumin, prealbumin and transferrin, we examine the dependence of the deviation parameters C_i on the concentration and the kind of serum components in section 3. In section 4 we point out how to choose the starting values of the parameters in the analysis of agar gel electrophoresis and ultracentrifugation patterns of $10-15$ overlapping fractions. The degree of automation in pattern analysis by means of the Fortran-program DIANA is discussed in section 5.

2. The Gaussian-like Distribution

In the analysis of electrophoresis patterns it is usually assumed that the distribution of a molecular fraction has the shape of a Gaussian. This assumption is supported not only by experience in pattern analysis but also by diffusion theory. To explain the latter aspect clearly, we first have a look at ideal one-dimensional diffusion [2]. For $c(x,t)$, which is the number of molecules in x-space and time, the diffusion equation

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - V\frac{\partial c}{\partial x} \qquad (1)$$

holds, where D is the diffusion constant, and V is the constant transport velocity due to an external force in x-direction. We now take the initial distribution to be the Dirac δ-function, so that the solution of Eqn (1)

$$c(x,t) = \frac{F}{2\sqrt{\pi D t}} \exp\left(-\frac{(x-Vt)^2}{4Dt}\right) \qquad (2)$$

has the shape of a Gaussian for each point of time t. What one observes in pattern analysis is the distribution of the molecules after switching off the voltage and fixation of the molecules. Let us choose at this time T a still simpler parameterization

$$c(x,T) = \frac{F}{\sqrt{\pi} B} \exp\left(-\frac{(x-S)^2}{B^2}\right) \qquad (3)$$

where S is the position (first moment), B is the width and F is the area (concentration) of the distribution. Comparing Eqn (3) and Eqn (2) we obtain S and B in terms of D, V and T: $S=VT$, $B=2\sqrt{DT}$.

In agar gel electrophoresis [3] the conditions are only approximately those of one-dimensional diffusion with a Dirac δ-function initial distribution. Let us now list the most important physical effects, which can cause a deviation from the Gaussian: The width of the split in the gel is not zero as in the δ-function, although it is still small compared with the width of the observed peaks. The existence of the boundaries of the gel in y and z-directions slightly disturbs the one-dimensionality of diffusion. Because of the heterogeneity of a molecular fraction D and V are not simple constant quantities but are rather average values of two or more molecular subconfigurations. In addition V has a small subtractive component due to electro-osmosis, which can be coordinate-dependent. Taking into account the protein protein interaction, another nonlinear term appears in the diffusion equation, which changes the mathematical expression of the solution. The shape of this solution, however, is similar to the Gaussian, as long as the serum protein mass of the sample is small.

The deviation from the Gaussian due to these effects can probably be kept small by improving the agar gel electrophoresis techniques. This, however, may not be true for the protein agar gel interaction, since this interaction is the basis for the agar gel electrophoresis itself. First of all the agar binds water and thereby increases the viscosity of the medium. Then there is a direct protein agar gel interaction due to electrical or chemical forces, which may be described as a temporary binding of the protein molecules to the agar by a two-state model [4]. This shows, that Eqn (1) is to be understood as an approximation with D and V being the reduced values of the diffusion constant and transport velocity.

In order to take into account in pattern analysis the deviation from the Gaussian due to one or more of the above listed effects, we modify the

Gaussian in the following way

$$f(x) = \frac{F}{\sqrt{\pi}\,B} \exp\left(-\left(\frac{x-S}{B}\right)^2 \cdot\left(1 + C_1 \tanh \frac{x-S}{C_0 B}\, C_2 \tanh^2 \frac{x-S}{C_0 B}\right)\right). \quad (4)$$

The first additional term describes the asymmetry of the distribution, whereas the second term represents the symmetrical part of the deviation (excess of the distribution). This function is Gaussian-like as long as the exponent is negative, which is the case if the condition

$$-|C_1| + C_2 \rangle - 1 \quad (5)$$

holds. We call C_1 the strength of the asymmetrical deviation and C_2 the strength of the symmetrical one. Since $C_0 B$ is the characteristic length of the deviation, we call C_0 the scale of deviation.

If two different effects cause the symmetrical and asymmetrical deviations respectively, then the scale of deviation in general is different in the two terms. Instead of introducing another parameter, we prefer to replace $\tanh \hat{x}$ by $\tanh^3 \hat{x}$, or $\tanh^2 \hat{x}$ by $\tanh^4 \hat{x}$ ($\hat{x}$ is the abbreviation of $(x-S)/(C_0 B)$).

Taking into account additional higher powers of $\tanh \hat{x}$ in Eqn (4) certainly would increase the accuracy of evaluation for each pattern. However, this gain of accuracy seems to be insignificant, since the statistical error in agar gel electrophoresis has the same order of magnitude as the remaining difference between the observed pattern and the fit.

The choice of a function like Eqn (4) is not unique, even if we restrict ourselves to that class of functions, which have three deviation parameters and the property to reduce to the Gaussian in a particular parameterization. From our point of view the main advantage of function (4) is the simple interpretation of the parameters C_i. Obviously in Eqn (4) we can replace $\tanh \hat{x}$ by another similar function, such as $\tan^{-1} \hat{x}$ or the error function, without changing the meaning of C_i. A test run has shown, that the accuracy of pattern evaluation usually is only slightly changed by this replacement, as long as the sum (5) does not approach -1. If this should happen in the analysis for several patterns of the same type, then we propose to replace $\tanh \hat{x}$.

For some applications of pattern analysis one may not be interested in the explicit deviation function but only in values which characterize the strength of the deviation. Then it is more convenient to measure the strength of the deviation by computing the higher moments of the distribution. The third central moment or the skewness is the strength of the asymmetrical deviation, whereas the fourth central moment or the kurtosis is a measure for the symmetrical deviation [5]. Since this method does not work for patterns with more than one molecular fraction, it is not an alternative approach to pattern analysis.

Different molecular components of the serum protein have in general different diffusion constants and transport velocities. As usual we sum up the distributions of the single components to obtain the distribution function of the serum protein, the pattern itself, *i.e.*

$$f_p(x) = \sum_{k=1}^{KZ} f_k(F_k, B_k, S_k, C_i^k), \quad (6)$$

where f_k is the Gaussian-like function (4) of the k-th molecular component and KZ is the number of molecular components.

3. The Distribution of a Single Molecular Fraction

We now apply function (4) in the analysis of several patterns of albumin with different molecular concentration as well as of prealbumin and transferrin. The values of the parameters F, S, B, C_0, C_1, C_2 are determined by minimizing χ,

$$\chi(F, S, B, C_0, C_1, C_2) = \sum_{i=1}^{N} \cdot (Y_i - f(X_i, F, S, B, C_0, C_1, C_2))^2, \quad (7)$$

where X_i, Y_i are the N digitized coordinates of the scanning curve. In fact we replace $\tanh^2 \hat{x}$ by $\tanh^4 \hat{x}$ in Eqn (4), since a test run has shown, that χ is then smaller for most of the patterns.

Table I. The parameter values from the analysis of a typical agar gel electrophoresis pattern of albumin by a Gaussian and Gaussian-like (4) distribution. C_0 is the scale of deviation, C_1 and C_2 are the strengths of the asymmetrical and symmetrical deviation, respectively (for curves *cf.* Fig. 1).

Distribution	B	C_0	C_1	C_2	χ
Gaussian	3.91	—	0.0	0.0	.729
Gaussian-like	3.66	1.71	−.22	−.43	.066

In Fig. 1 and Table I we compare the results of the analysis of a typical pattern with Eqn (4) to those with a truly Gaussian distribution. This comparison confirms that the precision of the digital computer analysis is considerably improved by

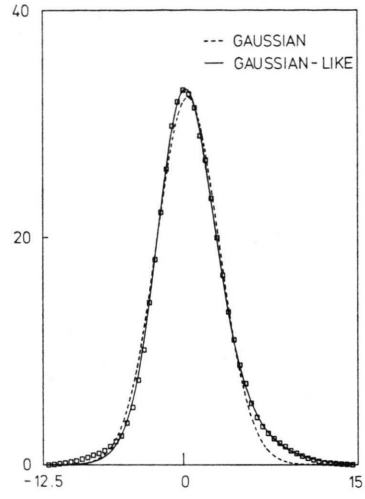

Fig. 1. A Gaussian (– – –) and a Gaussian-like (———) distribution adjusted to the digitized scanning curve (□) for a typical agar gel electrophoresis pattern of the albumin in arbitrary units (for parameters *cf.* Table I).

taking into acount these modifications of the Gaussian. The sum (5) is far from the bound -1. Since the transport velocity is in positive x-direction, $C_1<0$ means that the ascending boundary of the distribution is flattened out. The negative excess ($C_2<0$), the enlarged scale ($C_0>1$) and the replacement of $\tanh^2 \hat{x}$ by $\tanh^4 \hat{x}$ suggest, that the symmetrical part of the deviation shows up mainly as an extension of the tails of the distribution.

The dependence of the parameters C_i on the serum protein mass (concentration) of a molecular fraction is studied by analyzing patterns of albumin. Three patterns are evaluated for each concentration, which permits us to get a rough value of the statistical error of the separation and measuring procedure. A check has confirmed that the area of the distribution indeed rises linearly with the concentration. The analysis shows an independence, or only a small systematic dependence of the deviation on the concentration (Fig. 2). Assuming independence, we calculate the mean values of C_i to be $C_0 = 1.75 \pm .13$, $C_1 = -0.24 \pm .06$, $C_2 = -0.45 \pm .06$.

Since the protein protein interaction would produce a symmetrical deviation increasing with

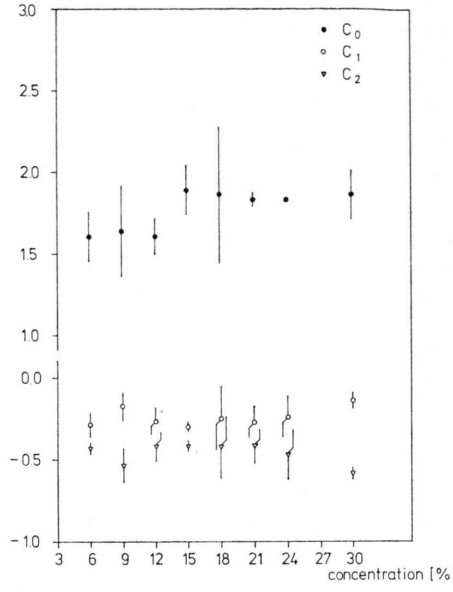

Fig. 2. The dependence of the deviation parameters C_i on the serum protein mass (concentration) from a series of agar gel electrophoresis patterns of albumin.

serum protein mass, whereas the small width of the split would cause a symmetrical deviation with positive C_2, none of these effects can explain the observed deviation. As for the heterogeneity of the molecular fractions it has been argued that it might be relevant, since molecules of different shape and different charge configuration differ in mobility. On the other hand diffusion is a stochastic process with a tendency to average. Therefore we expect a large effect of heterogeneity only, if two or more relatively stable subconfigurations exist. Since there is no reason to assume that different molecular fractions have similar sets of subconfigurations one would expect that the corresponding deviations are quite different for different molecular fractions and similar only by accident. Now the analysis of a pattern of prealbumin ($C_0 = 1.32$, $C_1 = -0.23$, $C_2 = -0.38$) and of transferrin ($C_0 = 1.35$, $C_1 = -0.12$, $C_2 = -0.59$) shows that the deviation is in fact similar for albumin, prealbumin and transferrin. This result therefore suggests, that the heterogeneity of a molecular fraction is probably not the main contribution to the observed deviation.

The influence of the protein agar gel interaction cannot be easily estimated without regard to a model. In a two state model [4], which describes the interaction as a temporary binding of the protein molecules to the agar, we obtain asymmetry of the

distribution due to the relaxation of the process. If the number of protein molecules in the bound state is larger than that in the free state, the ascending boundary of the distribution is flattened out ($C_1 < 0$). This model shows that the protein agar gel interaction can indeed explain the asymmetrical and possibly also the symmetrical part of the observed deviation.

4. Choosing the Starting Values of the Parameters

The analysis of a pattern with overlapping peaks is the more difficult the more the fractions overlap. To be more precise, we introduce a simple measure for the overlap of two Gaussian-like distributions j and k

$$U_{jk} = \exp\left(-\frac{|S_j - S_k|}{B_j + B_k}\right) \qquad (8)$$

and we call the overlap large for $1 > U_{jk} > 0.5$ and small for $0.5 > U_{jk} > 0$. For small overlaps the U_{jk} can be estimated directly by a hand-drawn evaluation of the scanning curve whereas for large overlaps this is not possible. Of course one can always compute the U_{jk} from the adjusted parameters S_k and B_k after the analysis. The knowledge of the U_{jk} can be very useful for the judgement of the goodness of fit.

The analysis of a pattern involves the adjustment of the parameters of Eqn (6) by a minimizing procedure, such that the "best" approximation of the observed curve by the function (6) is obtained. In the χ-squares method the best approximation is equivalent to the "exact" (absolute) minimum of χ. It is well known that the number of additional minima (good approximations) rapidly increases with increasing number of parameters and growing overlap of the peaks. Then the exact minimum is often only reached, if the starting values of the parameters are already close to their values at the exact minimum [6]. This clearly shows the importance of choosing the starting values carefully.

As an example we shall describe how to fix the number of molecular fractions and the starting values for some agar gel electrophoresis patterns of the cerebrospinal fluid [7] and 60 ultracentrifugation patterns of polysomes [8]. These patterns contain both regions with small and large overlap of the molecular fractions.

In the regions of small overlap the number of components KZ and good starting values for S_k and B_k can directly be read off from the scanning curve, whereas this is not possible for most of the deviation parameters C_0, C_1, C_2. Fortunately we know from the analysis of albumin, prealbumin and transferrin patterns that the C_i do not appreciably depend on F and k (the kind of serum component). The further dependence of C_i^k on the parameters S_k and B_k may be obtained either from a model or from an analysis of a series of patterns of a single fraction with varying transport velocity due to different values of the external voltage. At present we neglect this dependence. Then the deviation is the same for all fractions and the number of parameters is reduced from $6KZ$ to $3KZ+3$ for KZ serum components. Also the interpretation of the deviation parameters changes slightly in the sense, that now C_0, C_1, C_2 are the mean values of the scale and the strengths of deviation. As starting values of C_i we take the values from the analysis of albumin patterns.

In the regions of large overlap the direct determination of KZ and the starting values of S_k and B_k from the scanning curve is unreliable. Therefore we need additional information from another method of molecular separation, such as immunoelectrophoresis, or predictions of KZ, S_k and B_k by a model. Having thoroughly analyzed $10-20$ patterns of one type, the mean values of the adjusted S_k and B_k of these samples probably are good starting values for the analysis of further patterns of this type, since S_k and B_k mainly depend on the mobility of the serum components and not on their concentration.

In the case of agar gel electrophoresis for two of the samples, which we have analyzed, two-dimensional Laurell immunolectrophoresis patterns are available. There in the regions of large overlap (α_1, α_2 and γ fraction) the components of the fractions clearly show up. The number of components KZ, the positions S_k and the widths B_k of the components can be estimated from these two-dimensional patterns. Taking these values as starting values for the computer analysis of the corresponding agar gel electrophoresis patterns, we obtain a good fit. In the future we shall analyze about 20 samples of the cerebrospinal fluid by this method. We expect to obtain mean values of S_k and B_k with small variance, which then are fixed as starting values for the

analysis of further patterns of the cerebrospinal fluid. It can happen though, that another molecular component shows up which has not yet been recognized in the two analyzed patterns.

In ultracentrifugation with a suitable sucrose gradient [9] the distribution of a polysome is Gaussian-like and the transport velocity is approximately proportional to the mass of the polysome. Commonly it is assumed that the polysomes are aggregates of 1, 2, 3 ... molecules of the same mass. This picture suggests that the masses of the polysomes and consequently also the positions S_k of the polysome distributions form a monotonously increasing sequence of values. We therefore assume the starting values of S_k and also those of B_k to be monotonously increasing sequences $(k=1-11)$. Moreover three subcomponents of the monosome and one complex of it with one of its subcomponents are added as fractions $k=12-15$. Preliminary starting values for most of the fractions are estimated directly from the scanning curve. The value for the heaviest polysomes are obtained by extrapolating the sequences. With these values we analyze 20 patterns and compute the mean values of the adjusted parameters S_k and B_k. We take these mean values as our final starting values of S_k and B_k for the analysis of all 60 patterns, including the first twenty patterns. It turns out that the mean deviation from the Gaussian (5) is near its bound $(-|C_1|+C_2=-0.95)$. For all patterns we obtain a good approximation which, however, might not be the best approximation for a few of the patterns. A typical pattern as well as the remaining difference between the observed curve and the adjusted distribution is shown in Fig. 3.

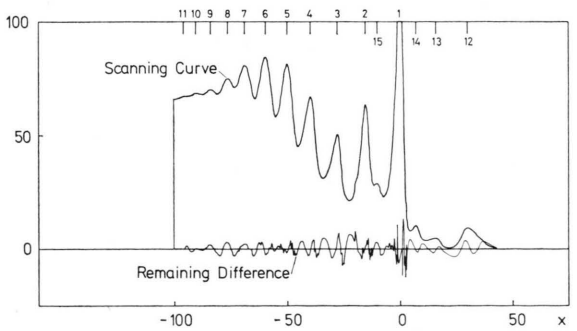

Fig. 3. The scanning curve and the remaining difference (enhanced by a factor 2) between the observed curve and the adjusted distribution (6) for a ultracentrifugation pattern of polysomes in arbitrary units. # 1–11 are polysomes, # 12–14 are subcomponents of the monosome, # 15 is a complex of the monosome with one of its subcomponents.

5. The Automation of Pattern Analysis

In this paper we have pointed out a way how to analyze automatically agar gel electrophoresis patterns with a digital computer. The minimizing algorithme, used in the Fortran-program DIANA [10], needs as input: The number $K\,Z$ of molecular fractions, starting values of the positions S_k, the widths B_k and the mean deviation parameters C_i. Starting values of the concentration parameters F_k are not required by DIANA. The computer running time for the pattern evaluation is fairly small and rises at most with the square of $K\,Z$. The running time for a typical pattern of 15 fractions is 3 min on a CDC 7200 computer.

We admit that this concept of pattern analysis may not be satisfactory from the point of view of a biochemist, who has no or little experience of computer programming. For, at present, the analysis is not fully automated as for each series of patterns the number of molecular fractions and a set of starting values of the parameters must be chosen in a test period. Only after these values have finally been fixed the analysis of further patterns of this series proceeds in a fully automatic way.

In principle, the starting values can be determined by a grid or Monte Carlo searching routine [11]. However, these methods tend to waste much computer time, especially in the case of a large number of parameters, and can hence be applied at most in the test period. In the production period, which may comprise the analysis of several hundreds of patterns for clinical purpose, one cannot afford to use these methods.

The proposed concept is as usual not completely save in attaining the best approximation. As already discussed, the best approximation of the pattern can only be reached, if the starting values are carefully chosen. In fact we do not determine the set of starting values for each pattern separately but only once for a series of patterns of the same type. Usually all patterns of one kind of fluid (blood, urine, cerebrospinal fluid, etc.), obtained under the same experimental conditions, belong to the same type. However, it may happen that one or a few patterns are quite different from the rest, so that we may hit an additional minimum rather than the exact one. Since the evaluation of these patterns results in a larger value of χ and probably also in a rather bad approximation in a certain region of the

pattern, these exceptional cases can in general be recognized in the computer print. They may be eliminated from the analysis or reevaluated with corrected starting values.

Starting from the same parameter values, it can happen that a minimizing procedure leads to the exact minimum whereas another one does not. *E. g.* the simplex method is known to be particularly save in that it avoids shallow additional minima [11]. In DIANA we use a problem oriented strategy, whose main advantage is a fast parmaeter adjustment but whose safety properties have not been checked yet.

I like to thank Prof. M. Siegert and Dr. I. Dornacher for stimulating discussions and for allowing me to use their experimental results from electrophoresis and ultracentrifugation.

[1] C. B. Laurell, Clin. Chem. **19**, 1, 99 [1973].
[2] P. Frank and R. v. Mises, Die Differential- und Integralgleichungen der Mechanik und Physik, Teil II, Vieweg, Braunschweig 1927. — J. Crank, The Mathematics of Diffusion, Oxford University Press, London 1956.
[3] R. J. Wieme, Agar Gel Electrophoresis, Elsevier Publishing Company, Amsterdam 1965.
[4] E. Tränkle, in preparation.
[5] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, Statistical Methods in Experimental Physics, North Holland Publishing Company, Amsterdam 1971.
[6] F. James, Proceedings of the 1972 CERN Computing and Data Processing School at Pertisan, CERN 72-21.
[7] M. Siegert and H. Siemes, Methodische Untersuchungen zur Agarosegel-Mikroelektrophorese des Liquor Cerebrospinalis von Kindern unter Anwendung eines Analogrechners zur Auswertung der Ergebnisse, F. U. Berlin preprint 1974.
[8] I. Dornacher, Untersuchungen zur Auftrennung und Charakterisierung von Polysomen aus Warmblüter-Embryonen und über den Einfluß einiger embryotoxischer Pharmaka auf die Polysomenprofile. Inaugural-Dissertation des Fachbereichs Chemie der Freien Universität Berlin.
[9] H. Noll, Nature **215**, 360 [1967].
[10] E. Tränkle, A Computer Analysis of Electrophoresis and Ultracentrifugation Patterns, F. U. Berlin preprint.
[11] CERN, computer program library, D 506, D 516.