METHODS & REVIEWS

Determination of Molecular Masses of Proteins in Solution: Implementation of an HPLC Size Exclusion Chromatography and Laser Light Scattering Service in a Core Laboratory

Ewa Folta-Stogniew and Kenneth R. Williams

HHMI Biopolymer Laboratory and W. M. Keck Foundation Biotechnology Resource Laboratory, Yale University, New Haven, CT

Size exclusion chromatography (SEC) coupled with "on-line" laser light scattering (LS), refractive index (RI), and ultraviolet (UV) detection provides an elegant approach to determining the molecular weights of proteins and their complexes in solution. SEC serves solely as a fractionation step to minimize the ambiguity that otherwise can result from the fact that light scattering provides the weight-average molecular weight (MW) of all species in solution. Our goal is to establish realistic expectations for MW determination using LS coupled with SEC, define sample requirements, and identify possible limitations of SEC/LS analysis. Analyses of 14 protein standards that range from 12 to 475 kd suggest that the molecular weights of native proteins may be determined in a single SEC/LS experiment with an accuracy of ±5%. The MW determination depends only on the downstream LS and RI detectors, and it is independent of elution position. Unusual elution because of nonglobular shape or interaction with the SEC support has no impact on the MW determination by SEC/LS. With the instrument configuration that was used, the optimal amount of protein needed for SEC/LS is about 50 μg for proteins with molecular weight greater than 40 kd. However, analyses of ovalbumin and transferrin demonstrate that even 10 μg is sufficient to determine the MW with an error of less than ±6%. Although SEC/LS has some limitations, such as proteins that contain chromophores whose absorption spectrum overlaps that of the emission spectrum of the laser, it represents a fast and robust approach to determining MW and to monitoring protein oligomerization in solution. (J Biomol Tech 1999;10:51–63)

KEY WORDS: laser light scattering, native molecular weight, protein association, size exclusion chromatography.

ADDRESS CORRESPONDENCE AND REPRINT REQUESTS TO Ewa Folta-Stogniew, Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520-8024 (email: Ewa.Folta-Stogniew@yale.edu).

The remarkable success of genome-level DNA sequencing has put us on a threshold of knowledge that was unimaginable when the textbooks that introduced us to biochemistry were written.[1] To cross this threshold will require that we not only know, for instance, the sequences of the approximately 80,000 human proteins but also that we uncover and understand the myriad array of interactions between each of these proteins and their nucleic acid and other ligand targets and partners. For it is through these interactions that functions that are encrypted within the primary sequences are expressed. The paradigm shift being wrought by "the genome project" is well evidenced by the evolution in the questions that are being asked. Increasingly, the goal is to uncover and to understand the function of proteins rather than to simply sequence them and clone their respective genes. Often, one of the first approaches taken to identify or more closely define function is to identify interacting proteins and ligands and to then quantify the relative binding affinities.

Several excellent technologies exist for detecting and quantifying macromolecular interactions under equilibrium or near equilibrium conditions in solution. These include isothermal and differential scanning microcalorimetry, size exclusion chromatography (SEC) coupled with in-line laser light scattering (LS) detectors and a variety of fluorescence techniques. However, few biopolymer core laboratories offer these or other biophysical services. The goal of this study was to evaluate the technical capabilities of high performance liquid chromatography (HPLC) SEC/LS, particularly with respect to its suitability for inclusion in a core laboratory.

In theory, HPLC SEC/LS provides the capability of determining the "absolute" molecular weights of proteins and their protein:protein complexes. Molecular weights determined by SEC/LS depend only on the readings obtained from the downstream LS and refractive index (RI) detectors.[2–5] The resulting weight-average molecular weight (MW) is independent of the HPLC SEC elution position. Nonglobular shape and interaction with the SEC support, which otherwise pose severe limitations with regard to the use of SEC for estimating MW, have no impact on MWs determined by LS. Other advantages of HPLC SEC/LS are that it is a "noninvasive" technique that does not require the incorporation of a radioactive or fluorescent tag. SEC/LS is nondestructive, and the samples may be recovered for use in subsequent studies. Compared with techniques such as analytical centrifugation, SEC/LS is rapid, and samples may be analyzed easily at various pH values, ionic strengths, and temperatures and in the presence or absence of ligands. The second virial coefficient, which provides a measure of solute-solvent interactions, can be determined from a batch mode experiment. Although a variety of mass spectrometric (MS) approaches can be used to determine MWs more accurately than by SEC/LS, the MS determination is carried out in the gas phase, under nonequilibrium and generally nonnative conditions. MS therefore has limited utility in terms of detecting and quantifying noncovalent protein:protein interactions under native, equilibrium conditions in solution.

SEC/LS is a particularly useful tool for studying homoassociations and heteroassociations of proteins and other biologic macromolecules.[2–6] Fascinating examples are studies of the dimerization of several receptor proteins. For instance, a ligand-induced dimerization was detected for the erythropoietin receptor[7] and the tyrosine kinase KIT[8] while measuring the molar masses of the receptor proteins in the free- or ligand-bound state. Studies on the extracellular domain of the epidermal growth factor receptor provide a nice example in which the dissociation constant for ligand-induced dimerization has been determined by evaluating the MW as a function of receptor concentration.[9] Another example is a study of the assembly of the giant hemoglobin, which attained a molar mass in the range of 3 to 4 Md.[10]

These observations suggest strongly that SEC/LS offers a powerful tool for characterization of the biophysical properties of proteins and their biologically relevant complexes. Because incorporation of SEC/LS services into multi-user biotechnology core laboratories would provide much wider access to this technology, we have carried out a preliminary study directed at identifying parameters that contribute to the success of SEC/LS in a core laboratory setting and particularly at helping to establish realistic expectations for such a service.

MATERIALS AND METHODS

HPLC Size Exclusion Chromatography and Laser Light Scattering Service

SEC was carried out at RT at a flow rate of 0.4 mL/min using one of the following two columns and buffers:

1. Superdex 200 HR10/30 column (Pharmacia, Piscataway, NJ), which separates globular proteins with MWs that range from approximately 10 to 600 kd. The total column volume is approximately 24 mL, and the standard buffer is 20 mM N-2-hydroxyethyl-piperazine-N-2-ethane-sulfonic acid (HEPES), 100 mM KCl, 1 mM EDTA, with or without 1 mM dithiothreitol (DTT), at pH 8.0.
2. TSK-GEL G3000$_{SWXL}$ (TosoHaas, Montgomeryville, PA), which separates globular proteins with MWs that range from approximately 10 to 500 kd. The total column volume is 14.3 mL and the standard buffer is 20 mM phosphate, at pH 7.5, with 150 mM NaCl.

The HPLC size exclusion column was connected in-line with the following flow detectors:

1. UV detector (Model 773 variable wavelength, KRATOS, Applied Biosystems, Foster City, CA)
2. LS detector (DAWN DSP, Wyatt Technology, Santa Barbara, CA)
3. RI detector (OPTILAB DSP, Wyatt Technology, Santa Barbara, CA)

The solvent delivery system was a Waters' 510 pump (Waters Corp., Milford, MA) that was equipped with two transducers to dampen pump pulsation and a Rheodyne 7125 valve which served as an injector.

In the case of the two-detector approximation, the system was calibrated using five proteins (cytochrome *c,* carbonic anhydrase, ovalbumin, bovine serum albumin, and alcohol dehydrogenase) that range from 12.3 to 146 kd.

Proteins

The following proteins were used, and the biologic source and Swiss-Prot accession number are indicated within parentheses. alcohol dehydrogenase (yeast, P00331), serum albumin (BSA, bovine, P02769), carbonic anhydrase (bovine, P00921), cytochrome *c* (horse, P00004), apo-ferritin (horse, P02791), α-lactalbumin (bovine, P00711), aldolase (rabbit, P00883), β-lactglobumin (bovine, P02754), enolase (rabbit, P25704), enolase (yeast, P00924), myoglobin (horse, P02188), transferrin (human, P02787), trypsin inhibitor (soy bean, P01071), ovalbumin (chicken, P01012) were from Sigma (St. Louis, MO), and glutamate dehydrogenase (bovine, P003666) was from Boehringer Mannheim (Indianapolis, IN). The UP1 protein, which is a proteolytic fragment that contains residues 1–195 in A1 hnRNP, was made by means of limited tryptic cleavage, as described by Kumar et al.[11] The sample was obtained by limited trypsin proteolysis of recombinant hnRNP A1 (*Escherichia coli* expressed) at 0°C, with a 1:120 w/w enzyme to substrate ratio.[11] After 30 minutes of digestion, the sample was diluted with 7 volumes of column buffer (50 mM TRIS-HCl, at pH 7.5, 1 mM EDTA, 10 mM $Na_2S_2O_5$, 1 mg/mL pepstatin A, 1 mM phenylmethylsulfonyl fluoride [PMSF]) and applied to a XK 16/20 column packed with 10 mL AGPOLY(U) resin (Pharmacia Biotech Inc., Piscataway, NJ). The protein was eluted with a linear 30-minute gradient of 0 to 2 M NaCl in the running buffer. A fraction containing UP1 (in 0.6 M NaCl) was used as the LS sample.

Molecular Weight Determination

MWs were determined using the "two-detector" approximation[2,3] or ASTRA calculations (see the Theory section). The "two-detector" approximation uses data acquired at the maximum of the peak, and ASTRA uses data acquired across the entire peak. In the case of the ASTRA calculations the MW was determined from a Debye plot, as illustrated in Figure 1 for ovalbumin. A dn/dc value of 0.185 mL/g was used for all proteins when analyzed in the HEPES buffer, and a dn/dc value of 0.186 mL/g was used for all analyses performed in the phosphate buffer. These values are comparable to a dn/dc value of 0.186 mL/g used in previous studies.[5]

THEORY

Background

The amount of light scattered is directly proportional to the product of the weight-average molar mass and the solute concentration[1]: LS $\sim$ MW·c. This relationship is based on Zimm's formalism of the Rayleigh-Debye-Gans light scattering model for dilute polymer solutions.[2–5,12,13] The relation between the excess scattered light and MW is given by Equation 1:

$$\frac{K^*c}{R(\theta)} = \frac{1}{MW \cdot P(\theta)} + 2A_2c \qquad [1]$$

where

R(θ)   is the excess intensity of scattered light at DAWN angle θ

c   is the sample concentration (g/mL)

MW   is the weight-average molecular weight (molar mass)

A_2 is the second virial coefficient (mL·mol/g²)

K*   is an optical parameter equal to $4\pi^2n^2(dn/dc)^2/(\lambda_0^4N_A)$

n   is the solvent refractive index and dn/dc is the refractive index increment

N_A is Avogadro's number

λ_0 is the wavelength of the scattered light in vacuum (cm)

There are several approximations that can be used while solving Equation 1. The second virial coefficient term ($2A_2c$) can be neglected when $2A_2cMW \ll 1$. This condition is met at the relatively low concentrations (usually below 0.1 mg/mL) used with HPLC SEC. A typical value of A_2 is 10^{-5} mL·mol/g² for a globular protein[2]; the $2A_2cMW$ term is equal to only 0.002 for a 100-kd protein at a concentration of 0.1 mg/mL. The second virial coefficient term therefore can be neglected while analyzing proteins at concentrations below 0.1 mg/mL.

The function P(θ) describes the angular dependence of scattered light. The expansion of 1/P(θ) to first order gives

$$1/P(\theta) = 1 + (16\pi^2/3\lambda^2)\langle r_g^2\rangle\sin^2(\theta/2) + \ldots$$

When using the DAWN detector (incident light of 633 nm), the $[(16\pi^2/3\lambda^2)\langle r_g^2\rangle\sin^2(\theta/2)]$ term has a maximum value of only 0.03 for all proteins or complexes with root mean square radii ($\langle r_g^2\rangle^{1/2}$) smaller than 15 nm. This $[(16\pi^2/3\lambda^2)\langle r_g^2\rangle\sin^2(\theta/2)]$ term is negligible for globular proteins with root mean square radii ($\langle r_g^2\rangle^{1/2}$) of less than 15 nm, which includes
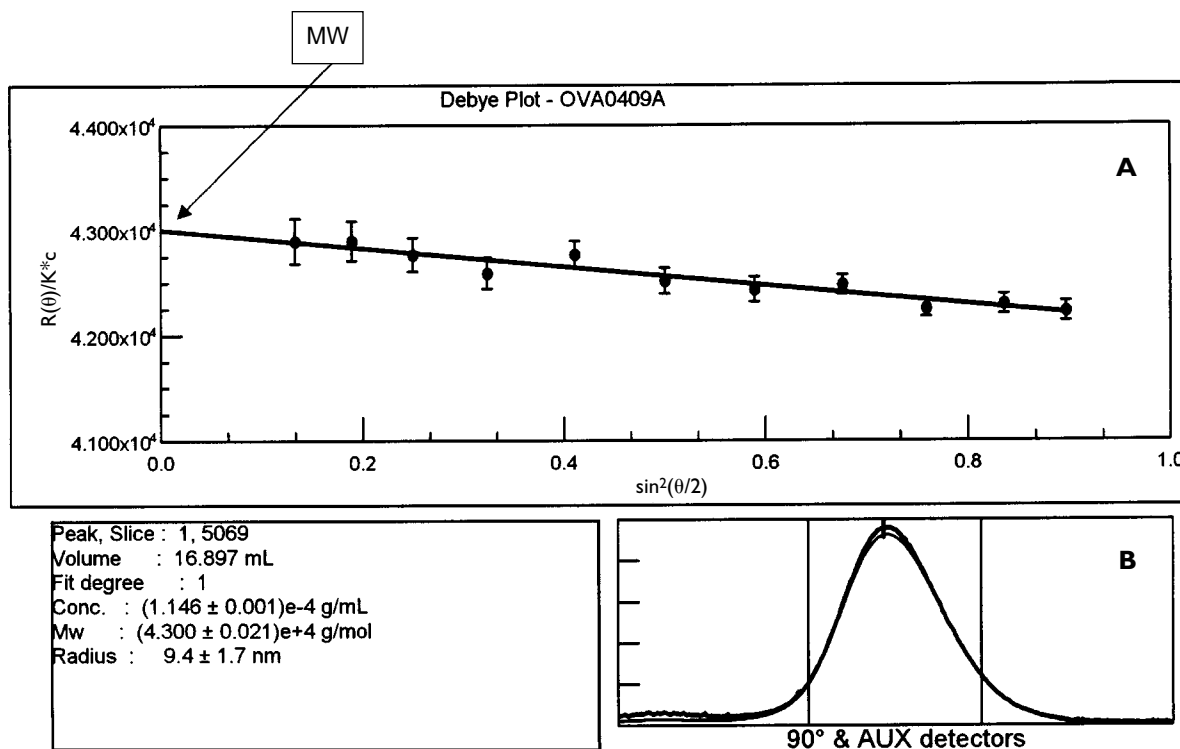
FIGURE 1

Debye plot for ovalbumin. **(A)** Debye plot: a plot of $R(\theta)/K^*c$ versus $\sin^2(\theta/2)$ and fit of a polynomial to the data to obtain the molecular weight from the intercept and the root mean square radius ($\langle r_g^2 \rangle^{1/2}$) from the slope at zero angle (see http://info.med.yale.edu/wmkeck/6_16_98/Astra2a.htm for more details). **(B)** Superposition of the refractive index (*thin line*) and light scattering (*thick line*) traces with the peak boundaries marked by vertical lines; the short vertical mark on the traces indicates the slice for which the Debye plot is shown.

proteins and their complexes with MW smaller than approximately 5000 kd.[2,4]

At low angles, the angular dependence of light scattering depends only on the root mean square radius ($\langle r_g^2 \rangle^{1/2}$) and is independent of molecular conformation or branching.[3] A plot of $K^*c/R(\theta)$ against $\sin^2(\theta/2)$ (Zimm plot)[12,13] yields a curve whose intercept gives $(MW)^{-1}$ and whose slope at low angles gives the root mean square radius ($\langle r_g^2 \rangle^{1/2}$), which is often (but apparently inappropriately[3]) referred to as the *radius of gyration*.

Two-Detector Approximation for Calculating Molecular Weight

Based on the previous discussion, the HPLC SEC/LS conditions used in this study permit Equation 1 to be simplified to

$$(K^*c)/R(\theta) = 1/MW \qquad [2]$$

Substituting the expression, $4\pi^2n^2(dn/dc)^2/(\lambda_0^4 N_A)$, for K^* and introducing K_{LS} (an instrument calibration

constant that includes the expression $4\pi^2n^2/[\lambda_0^4 N_A]$), the *measured* intensity of scattered light at a given -angle (θ) is

$$(LS) = K_{LS}(c)MW(dn/dc)^2 \qquad [3]$$

Similarly, the refractive index signal (RI) can be expressed as

$$(RI) = K_{RI}(c)(dn/dc) \qquad [4]$$

where K_{RI} is an instrument calibration constant.

For proteins and their protein:protein complexes that contain only polypeptides (no carbohydrates), the dn/dc value is nearly constant ($\sim$0.19 mL/g)[2–5] and independent of amino acid composition. The MW can be determined from the ratio of the signals from the two detectors (LS and RI):

$$MW = K'(LS)/(RI) \qquad [5]$$

where $K' = K_{RI}/[K_{LS}(dn/dc)]$.

This is the so-called two-detector method.[5] This method is widely used, but it is valid only when the

dn/dc value is known.[4,5] With this approach, the MW is *not* determined from the absolute light scattering measurements but rather is calculated from Equation 5. The instrument calibration constant (K') is determined by analyzing protein standards.

"ASTRA" Approach for Calculating Molecular Weight

Alternatively, the MW can be determined directly from the absolute light scattering measurements by solving Equation 1, providing that the concentration of eluting protein is determined independently. There are several ways in which this equation can be solved for the MW and $\langle r_g^2 \rangle^{1/2}$. The ASTRA software (Wyatt Technology Corporation, Santa Barbara, CA) package provides the following fitting methods to solve Equation 1.

Zimm Fitting Method

The Zimm fitting method[12,13] relies on constructing a plot of $K^*c/R(\theta)$ against $\sin^2(\theta/2)$ and fitting a polynomial in $\sin^2(\theta/2)$ to the data, thereby obtaining the MW and $\langle r_g^2 \rangle^{1/2}$ from the intercept and slope at zero angle. Zimm plots have been reported to work well for mid-sized molecules (rms radius ~20–50 nm).[14]

Debye Fitting Method

The Debye fitting method[15] uses a plot of $R(\theta)/K^*c$ against $\sin^2(\theta/2)$ and fits a polynomial in $\sin^2(\theta/2)$ to the data, thereby obtaining MW and $\langle r_g^2 \rangle^{1/2}$ from the intercept and slope at zero angle. An example of such a plot is shown in Figure 1 for ovalbumin. This approach is applicable over a wider range of MWs than the Zimm formalism.[14]

Berry Fitting Method

The Berry fitting method[16] constructs a plot of the SQRT[$K^*c/R(\theta)$] against $\sin^2(\theta/2)$ and then fits a polynomial in $\sin^2(\theta/2)$ to the data. It has been reported to be particularly useful for large molecules.

Random Coil Method

The random coil method[17] approach solves Equation 1 after inserting the theoretical form factor $P(\theta)$ for random coils, which is given by

$$P(\theta) = \frac{2}{u^2}(e^{-u} - 1 + u)$$

where $u = (4\pi/\lambda)^2 \langle r^2 \rangle \sin^2(\theta/2)$.

$P(\theta)$ is a nonlinear function of $\langle r^2 \rangle$; an iterative nonlinear least square fit is used during fitting. This fitting method may be advantageous for large random coil molecules.[17]

Solving Equation 1 is done automatically by ASTRA software for selected parts of chromatograms after the data are collected and the baselines are assigned.

Our experience is that all the above methods can give similar results for small molecules (rms radius < 10 nm). Any of these fitting formalisms can be used interchangeably for proteins with MWs of less than 5×10^5 daltons. Regardless of which approach is taken, the ASTRA software automates the curve fitting to Equation 1.

RESULTS AND DISCUSSION

Precision and Accuracy of Protein Molecular Weights Determined by SEC/LS

We assessed the reliability of SEC/LS by carrying out 46 analyses on 14 standard proteins (Table 1). The MWs were calculated using two methods. In the first approach, the MW was calculated for each "slice" of the chromatogram (ie, for a volume interval of ~3.3 μL) using ASTRA software with the Debye fitting method. The reported values represent the weight-average MW calculated for the majority of the eluting peak. In the second approach, the MW was calculated for the single slice at the maximum of the peak using the two-detector approximation as described earlier. Both analyses provided similar results, with the ASTRA approach being faster and more informative because of its ability to determine automatically the MW distribution across an eluting peak (Figs. 2 through 6).

As summarized in Table 1, the overall median standard deviation was 0.9 kd for all replicate determinations carried out within a period of 2 months. The largest standard deviation observed was 3.9 kd, which corresponds to 2.9% of the predicted value of 133 kd for the BSA dimer. The average error between the observed and predicted MW ranged from 0.8% to 16.3%, with an overall median average error of 2.3% for the 14 proteins examined.

In a few instances, the difference between the observed and predicted MW exceeded 3%. Further examination of these examples points to plausible and interesting explanations for these above-average errors. As shown in Table 1, the MW determined for both enolases studied were lower than expected. Figure 2 illustrates that, in the standard running buffer that contains EDTA, the elution peak for yeast enolase

T A B L E   1

Molecular Weights Determined From ASTRA Analysis

Protein	Oligomeric state	No. of runs[a]	Pred. MW[b] (kd)	Average MW[c] (kd)	Standard deviation[c] (kd)	Average error[d] (%)
Cytochrome *c*	Monomer	5	12.3	12.0 ± 0.2	0.57	2.4[e]
α-Lactalbumin	Monomer	2	14.2	14.320 ± 0.007	0.010	0.84
Myoglobin	Monomer	3	17.0	14.2 ± 0.5	0.91	16[e]
β-Lactoglobulin	Monomer	2	18.3	20.1 ± 0.2	0.33	9.7
Trypsin inhibitor	Monomer	1	20.0	20.5		2.3
Carbonic anhydrase	Monomer	4	29.0	29.2 ± 0.1	0.20	0.76
Ovalbumin	Monomer	10	42.8	42.52 ± 0.07	0.68	1.4
Bovine serum albumin (monomer)	Monomer	5	66.4	66.41 ± 0.04	1.0	1.2
Transferrin	Monomer	2	75.2	76.9 ± 0.7	0.98	2.3
Enolase (yeast)	Dimer	3	93.3	80.7 ± 0.7	1.2	13[f]
Enolase (rabbit)	Dimer	4	93.7	86.4 ± 0.8	1.9	7.8[f]
Bovine serum albumin (dimer)	Dimer	5	132.9	137 ± 2	3.9	3.2
Alcohol dehydrogenase	Tetramer	4	147.4	144.0 ± 0.4	0.86	2.4
Aldolase (rabbit)	Tetramer	2	156.8	154 ± 1	1.9	1.1
Apo-ferritin	24 × Monomer	2	475.9	470 ± 2	2.6	1.2
				Median	0.90	2.3

[a]Column: Superdex 200 (Pharmacia); buffer: 20 mM HEPES, 100 mM KCl, 1 mM EDTA, pH of 8.0.

[b]Predicted (pred.) molecular weight was calculated from the protein sequence as retrieved from the ExPASy molecular biology WWW server of the Swiss Institute of Bioinformatics (SIB): http://expasy.hcuge.ch/

[c]Average molecular weight (MW), error in the mean value (E), and standard deviation (SD) were calculated as[18]

$$MW = [sum(MW_i)/n]; \quad E = SD/[SQRT(n)]; \quad \text{and} \quad SD = SQRT\{[sum(MW_i - M)^2]/(n - 1)\},$$

where MW is an arithmetic mean calculated as $MW = sum(MW_i)/n$ (given in column "average MW"); MW_i is a result of the ith measurement, ie, experimental MW determined in the ith run; and n is the number of runs.

[d]The percent error is calculated as an average from the absolute values of $\{100 \times [(\text{experimental MW} - \text{pred. MW})/\text{pred. MW}]\}$ calculated for each run.

[e]Colored proteins absorb at the wavelength of the laser beam (633 nm), and the instrument is not capable of correcting for the absorbed light. The amount of scattered light is smaller and leads to underestimated molecular weight.

[f]These dimers require Mg^{2+} for stability, and they are unstable under the chromatographic condition used (ie, buffer with 1 mM EDTA).
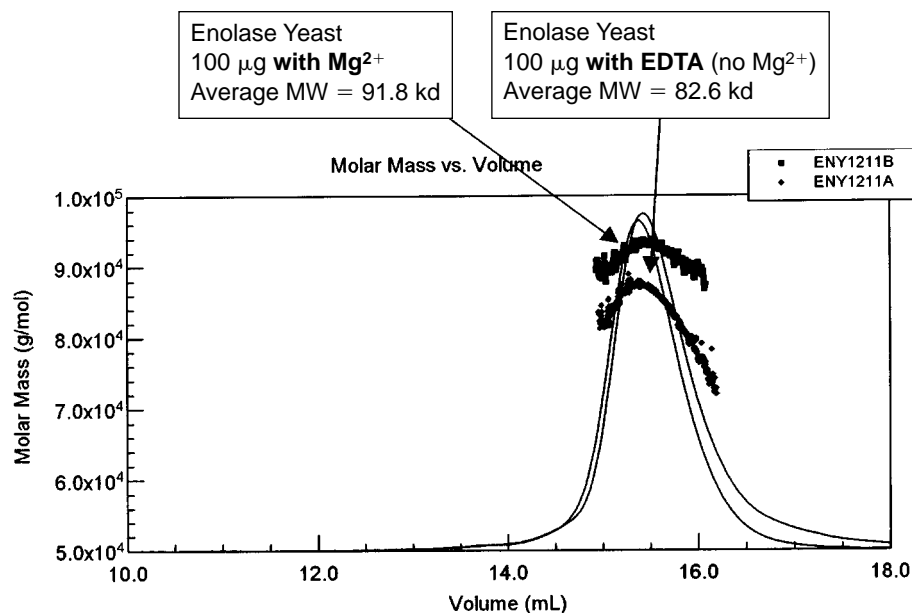
is not monodisperse (ie, the peak is not homogenous with respect to MW). The MW varies from ~90 kd at the maximum of the peak to ~55 kd at the trailing edge. Consequently, the estimated MW reported by the ASTRA algorithm (82.6 kd) is a weight-average value for all species present in the eluting peak, as predicted by light scattering theory. Because yeast enolase has a predicted monomeric MW of 46.9 kd, it appears the heterogeneity across the eluting peak results from partial dissociation of the dimer during HPLC SEC/LS. Because the dimer is stabilized by Mg^{2+} buffer [http://www.expasy.ch/cgi-bin/get-sprot-entry?P00924], it is not surprising the dimeric species would be destabilized by the 1 mM EDTA that was

included routinely in the running buffer. In the presence of Mg^{2+}, the enolase peak was observed with an average MW of 91.8 kd, in good agreement with the predicted value of 93.3 kd (Fig. 2). Similarly, the initial HPLC SEC/LS study (Table 1) on rabbit enolase was carried out in the presence of EDTA, which undoubtedly led to destabilization of the dimer and to a lower than predicted MW of 86.0 kd.

The highest discrepancy between the expected and experimentally determined MW was observed for myoglobin. Because myoglobin is a monomeric protein, an alternative explanation other than oligomerization must be invoked to understand why its experimental MW was 16.3% less than predicted (Table 1).

FIGURE 2

Determination of the molecular weight of enolase in the presence and absence of Mg^{2+}: molar mass distribution plot. The solid lines indicate the trace from the refractive index detector, and dots are the weight-average molecular weights for each slice (ie, measured every 0.5 s). The yeast enolase was analyzed on a Superdex 200 column in a buffer (20 mM HEPES, 100 mM KCl, pH 8.0) that contained 1 mM EDTA or 0.5 mM Mg^{2+}.

Enolase Yeast 100 μg **with Mg^{2+}** Average MW = 91.8 kd

Enolase Yeast 100 μg **with EDTA** (no Mg^{2+}) Average MW = 82.6 kd

Molar Mass vs. Volume

ENY1211B
ENY1211A

In this instance, the most likely origin of the discrepancy is the presence of prosthetic heme group. Apparently, this group absorbs sufficient light at the wavelength of the incident laser light (633 nm) that the amount of scattered light is less than expected, which leads to an underestimation of MW (Michelle Chen, personal communication, April 1998). Although the error was smaller, a similar effect was observed for cytochrome *c* (ie, average experimental MW was 0.3 kd [2.4%] less than the predicted value of 12.3 kd). We believe the results for cytochrome *c* were closer to the expected value for several reasons. First, the concentration of the cytochrome *c* sample was twofold less than that for myoglobin, leading to a lower absorbance at 633 nm. Second, myoglobin absorbs more strongly at 633 nm than cytochrome *c,* but the extinction coefficients at 280 nm are comparable for both proteins. The 0.1% extinction coefficients at 280 nm are 0.83 mL/(mg·cm) for myoglobin and 0.98 mL/(mg·cm) for cytochrome *c,* and the ratios of A_{633}/A_{280} are 0.13 and 0.05, respectively.

Our HPLC SEC/LS study demonstrates also that limitations of using HPLC SEC *without* on-line light scattering detection are overcome when the LS detector is present. In particular, HPLC SEC is an empirical approach that relies on external calibration based on the elution position, which is readily altered by nonglobular shape and interaction with the SEC matrix. An example of such behavior is shown in Figure 3*B*. In the absence of EDTA, yeast enolase eluted from SEC in a single, monodisperse peak with an elution volume of 9.75 mL, which is between that for a monomer of BSA (MW = 66.3 kd; elution volume 9.25 mL) and ovalbumin (MW = 42.8 kd; elution vol-

ume 10.15 mL) (Fig. 3*B*). Without the LS detector, SEC analysis of yeast enolase would have provided an estimated MW of ~50 kd (which corresponds most closely to the MW predicted for the monomer, 46.7 kd), whereas light scattering demonstrates that the eluting peak actually has a MW of 92.3 kd, which corresponds to a dimer (Fig. 3*B*).

In addition to detecting the presence of monomer:oligomer equilibria in solution, the ability of the ASTRA algorithm to provide a nearly continuous plot of MW across the eluting peak provides a powerful means for detecting co-eluting proteins. The ASTRA software allows determination of a weight-average molar mass for each "slice" of an eluting peak (ie, every 3.3 μL or 0.5 s in this study). The solid lines in Figures 2, 3, and 4 indicate the trace from an RI detector, and the dots are the weight-averaged MWs for each slice. From such a plot, one can see whether the eluting peak is monodisperse (ie, homogenous with respect to the molar mass, such as the peak at 15.6 mL in Fig. 3*A* or the peaks at 9.3, 9.75, and 10.1 mL in Fig. 3*B*) or the peak is polydisperse and contains a mixture of molecules with different molar masses (eg, the peak at ~13 mL in Fig. 3*A* or the peak at 9.0 mL in Fig. 3*C*). Although the polydispersity of the BSA peak at ~13 mL in Figure 3*A* results from incomplete SEC separation of the BSA trimer and tetramer species, the polydispersity of the peak shown in Fig. 3*C* results from co-elution of two different proteins. In this instance, the major SEC peak contains a mixture of the BSA monomer (66.3 kd) and the yeast enolase dimer (93.3 kd).

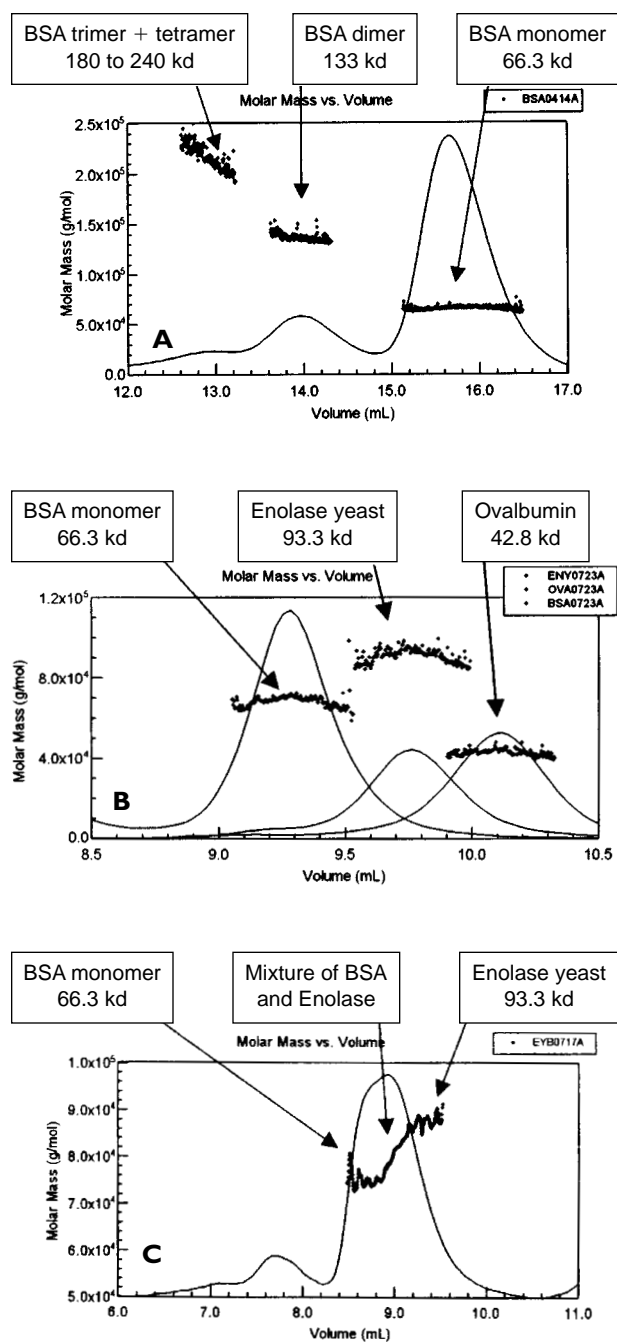The MW determined from a light scattering experiment is a weighted average of the MWs for the

BSA trimer + tetramer 180 to 240 kd

BSA dimer 133 kd

BSA monomer 66.3 kd

Molar Mass vs. Volume

A

BSA monomer 66.3 kd

Enolase yeast 93.3 kd

Ovalbumin 42.8 kd

Molar Mass vs. Volume

B

BSA monomer 66.3 kd

Mixture of BSA and Enolase

Enolase yeast 93.3 kd

Molar Mass vs. Volume

C

FIGURE 3

Molar mass distribution plot. The solid lines indicate the trace from the refractive index detector, and the dots are the weight-average molecular weights for each slice (ie, measured every 0.5 s). **(A)** Bovine serum albumin (column: Superdex 200 [Pharmacia]; buffer: 20 mM HEPES, 100 mM KCl, 1 mM EDTA, pH 8.0; for additional information see http://info.med.yale.edu/wmkeck/6_16_98/Bsarepa.htm). **(B)** Enolase from yeast, bovine serum albumin, and ovalbumin (column: TSK-GEL G3000$_{SWXL}$ [TosoHaas]; buffer: 20 mM phosphate, 150 mM NaCl, pH 7.5). **(C)** Mixture of bovine serum albumin and enolase from yeast (column: TSK-GEL G3000$_{SWXL}$ [TosoHaas]; buffer: 20 mM phosphate, 150 mM NaCl, pH 7.5).

individual components. In the example shown in Figure 3C, the MW changes from ~70 kd at the leading edge of the major peak at ~8.5 mL to ~93 kd at the trailing edge at ~9.5 mL. These results indicate that (as expected from the results shown in Fig. 3B) the BSA monomer (66.3 kd) elutes *before* the yeast enolase dimer (93.3 kd) and that the leading edge of the peak contains mostly BSA while the trailing edge contains mostly enolase.

Had such a result been obtained on an unknown sample, several approaches could have been used to determine whether the polydispersity resulted from co-elution of different proteins or partial dissociation of an aggregate of a single protein. First, partial dissociation of an oligomer produces molar mass distribution that is similar to the one presented in Figure 2; the MW increases along with the increase in the protein concentration. In the example shown in Figure 3C, the component with lower MW elutes ahead of the component with higher MW. Second, different proteins usually have different RI/UV absorbance ratios. The polydispersity evident in Figure 3C is also apparent in the corresponding RI/UV ratio plot (data not shown). Third, because HPLC SEC/LS is nondestructive, the eluent from the SEC/LS run can be collected and fractions across any observed peaks can be subjected to sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), two-dimensional PAGE, mass spectrometry, or a combination of these methods to positively identify the proteins eluting in each portion of the peak.

Sample Requirements for HPLC SEC/LS

We next addressed questions related to the sample requirements, particularly the minimum amount of protein needed for a reliable MW estimation. The results summarized in Table 1 were acquired on 50- to 300-μg amounts of protein that were injected in various sample volumes up to 3% of the total column volume. Because these amounts of protein may not always be available, various amounts of two standard proteins, ovalbumin (42.8 kd) and transferrin (75.2 kd), were analyzed to estimate the impact of sample amount on the accuracy of the resulting MW.

As shown in Table 2, the precision decreased, as evidenced by a fivefold increase in the standard deviation for replicate runs on ovalbumin. The range of errors observed in individual determinations increased as the amount of protein was decreased from the range of 100 to 150 μg to the range of 6 to 10 μg. However, even the 6- to 10-μg amounts provided a MW that always had an error of less than 6% for an

TABLE 2

Correlation Between the Amount of Protein Analyzed and the Accuracy of Molecular Weight Determination

Protein	Amount loaded[a] (μg)	No. of runs	Pred. MW[b] (kd)	Average MW[c] (kd)	Standard deviation[d] (kd)	Average error[e] (%)	Range of error[e] (%)
Ovalbumin	150	4	42.8	42.4	0.3	0.93	0.2–1.6
	100	7	42.8	42.3	0.8	1.2	0.2–2.4
	45–50	4	42.8	41.6	1	2.8	0.5–5.8
	6–10	5	42.8	42.9	2	0.23	1.4–4.5
Transferrin	100	3	75.2	76.5	1	1.7	0.7–3.2
	8	5	75.2	76.3	2	1.5	0.3–5.2

[a]Column: TSK GEL G3000$_{SWXL}$ (TosoHaas); buffer: 20 mM phosphate, 150 mM NaCl, pH of 7.5.

[b]Predicted molecular weight was calculated based on the protein sequence as retrieved from the ExPASy molecular biology WWW server of the Swiss Institute of Bioinformatics (SIB): http://expasy.hcuge.ch/

[c]Average molecular weight (MW) is a mean value calculated from weight averaged MW determined in the individual experiments.

[d]Error is calculated as $\sigma/(N)^{1/2}$, where σ is the standard deviation and N is a number of measurements in each set.[18]

[e]The percent error is calculated as the absolute value of $\{100 \times [\text{average MW} - \text{pred. MW})/\text{pred. MW}]\}$ calculated for the average value.

[f]Experimental accuracy (%) is calculated as the absolute value of $\{100 \times [(\text{average MW} - \text{pred. MW})/\text{pred. MW}]\}$.

individual analysis and that provided a precision of less than 2 kd, which corresponds to ~3% of the predicted MW for replicate determinations.

As shown in Figure 4, HPLC SEC/LS analyses on 7.7 μg amounts of transferrin and ovalbumin provided good quality data (on a TSK-GEL G3000$_{SWXL}$ column); the limiting factor was the LS signal, because the voltages recorded by the UV and RI detectors can be scaled up 10-, 100-, or even 1000-fold. During these analyses, the LS signal to noise (S/N) ratio was ~10 for transferrin and ~5 for ovalbumin (at the peak maxima). Nevertheless, the average MWs calculated by ASTRA (from a Debye analysis carried out on the entire peak) were 78 kd for transferrin and 41 kd for ovalbumin, which are both within 5% of the expected values of 75.2 and 42.8 kd, respectively.

FIGURE 4

Determination of molar masses for less than 10-μg amounts of protein. The solid lines indicate the trace from the refractive index detector, and the dots are the weight-average molecular weights for each 0.5-s "slice." A total of 7.7 μg of each protein was subjected to analysis (column: TSK-GEL G3000$_{SWXL}$ [TosoHaas]; buffer: 20 mM phosphate, 150 mM NaCl, pH 7.5).
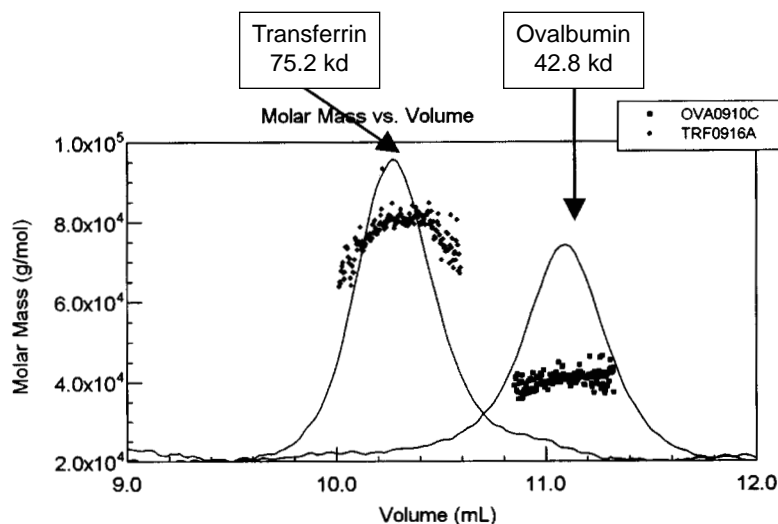
T A B L E   3

Optimal Amounts of Protein Required for HPLC Size Exclusion Chromatography With Laser Light Scattering

Column	Optimal amount of protein for expected MW >40 kd (μg)	Optimal amount of protein for expected MW 10 to 40 kd (μg)	Optimal amount of protein for expected MW <10 kd (μg)	Total volume of the eluting peak (mL)
Superdex 200 (Pharmacia)	100	200–300	Not applicable	~2
Superdex 75 (Pharmacia)	50	100–200	400	~1
TSK GEL G3000$_{SWXL}$ (TosoHaas)	50	200	Not applicable	~1

Because the LS signal is proportional to the product of the MW and the concentration, the smaller the MW, the higher the protein concentration (mg/mL) that is needed for analysis. Because an oligomeric form gives a proportionally larger LS response, the predicted MW for the monomer should be used as a guide for estimating the minimal amount of protein needed for the SEC/LS analyses.

Although it is possible to obtain reasonable SEC/LS data from 6- to 10-μg amounts of >40-kd proteins, when larger amounts of protein are available, studies should be carried out with amounts that are closer to the estimated, optimal sample amounts given in Table 3. Based on the data acquired, the optimal sample amounts given in Table 3 usually should provide an LS signal with a S/N >100, assuming an LS baseline noise that is below 2 mV and that >80% of the injected protein is recovered.

Another potential determinant of the LS signal (and therefore the amount of required protein) is the volume in which the protein elutes from the SEC. The total volume in which the sample elutes from the SEC affects the concentration of the eluting peak. As the column resolution is increased, the peaks sharpen, the eluting concentration is increased, and the amount of sample that is required may be decreased. Similarly, increasing the injected sample volume may decrease resolution and increase the amount of sample needed to provide the same degree of mass accuracy. A sevenfold increase in the sample volume caused an increase in the MW error from less than 1.9% to 3.3% (Table 4). Figure 5 demonstrates that increasing the injection volume causes small changes in elution position; nevertheless, these changes do not affect the MW determined from the SEC/LS analysis. The results summarized in Table 4 indicate that changing the sample volume has only a minor impact on the accuracy of SEC/LS analysis providing that the injection volume is kept below about 3% of a total column volume.

T A B L E   4

Correlation Between the Volume of the Injected Sample and the Accuracy of the Molecular Weight Determination

Column[a]	Injected volume (μL)	Sample volume as % of total column volume	Eluted volume (mL)	Injected amount (μg)	Eluted amount (μg)	Average MW (kd)	Error[b] (%)
Superdex 200 (Pharmacia)	100	0.4	2.1	150	109	43.6	1.9
	500	2.1	2.3	135	109	41.7	2.6
	750	3.1	2.3	160	160	44.2	3.3
TSK GEL G3000$_{SWXL}$ (TosoHaas)	30	0.2	1.0	45	36	42.2	1.4
	200	1.4	1.5	45	39	41.4	3.3

[a]Columns: Superdex 200 (Pharmacia); buffer: 20 mM HEPES, 100 mM KCl, 1 mM EDTA, pH of 8.0; or TSK GEL G3000$_{SWXL}$ (Toso-Haas); buffer: 20 mM phosphate, 150 mM NaCl, pH of 7.5.

[b]The percent error is calculated as the absolute value of $\{100 \times [(MW - \text{pred. MW})/\text{pred. MW}]\}$.

FIGURE 5

Determination of sample volume requirements. Molar mass distribution plot for ovalbumin injected in different sample volumes. The solid lines indicate the trace from the refractive index detector, and the dots are the weight-average molecular weights for each slice (ie, measured every 0.5 s). The TSK-GEL G3000$_{SWXL}$ (TosoHaas) column was eluted with 20 mM phosphate, 150 mM NaCl, pH 7.5 buffer.

Ovalbumin 45 μg in 200 μL

Ovalbumin 45 μg in 30 μL

Molar Mass vs. Volume

OVP0716A
OVP0716B
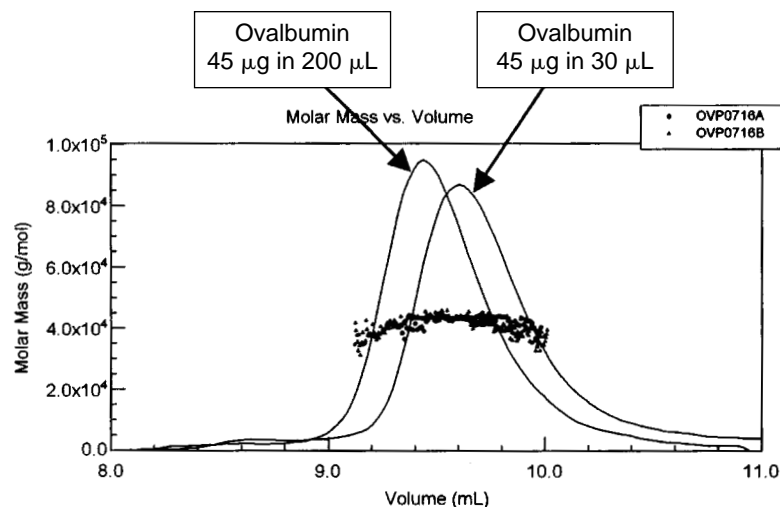
HPLC-SEC/LS Analysis of the UP1 Fragment of the A1 hnRNP Protein

A previous study[19] suggested that intact A1 hnRNP protein (residues 1–319) undergoes strong aggregation in solution and that removal of the C-terminal, glycine-rich region (residues 196–319) to yield the UP1 fragment prevents this aggregation, as evidenced by increased absorption at 320 nm and visible turbidity. Because this previous report did not determine the oligomerization state of UP1, it was of interest to do so by means of HPLC SEC/LS. As shown in Figure 6, the protein produced only one peak during SEC (the increase in the RI readings beyond 19 mL results from elution of salts in the injected sample). The MW for the major peak was estimated as 21.7 kd, in near perfect agreement with the expected value of 22.1 kd. The scattering of MW at the edges of the eluting peak

results from a low S/N ratio in the RI and LS signals, which propagates into low precision in the MW determination (only 30 μg of protein with a MW of about 20 kd was used for this analysis). The peak was monodisperse, as demonstrated by the homogeneous molar mass distribution across the peak (Fig. 6), indicating that the UP1 fragment is monomeric in solution.

An SEC/LS Service in the Core Laboratory Setting

Analysis of nonstandard proteins supplied by users presents special challenges that are not unique to HPLC SEC/LS. These include overestimation of the amount of protein submitted, contamination by nucleic acids, precipitation and loss of the protein

UP1 30 μg
MW$_{measured}$ = 21.7 kd

Molar Mass vs. Volume

UP10715A
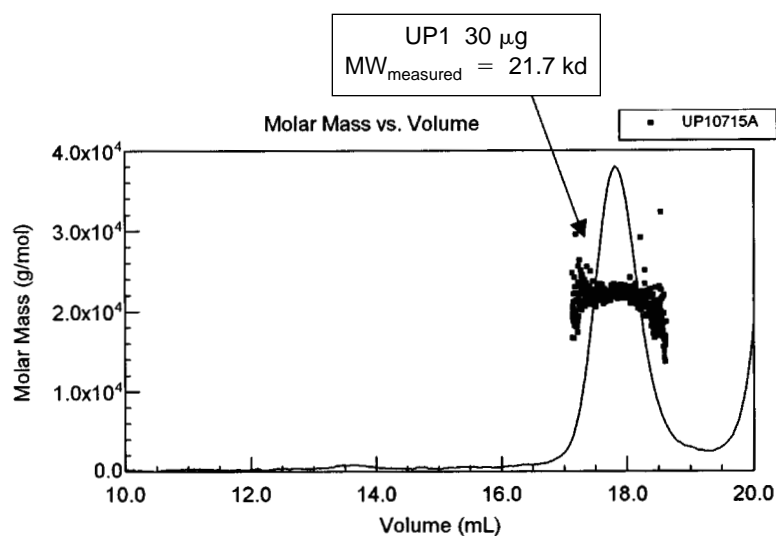
FIGURE 6

Determination of molar masses for UP1 protein; molar mass distribution plot. The solid lines indicate the trace from the refractive index detector, and the dots are weight-average molecular weights determined for each slice (ie, measured every 0.5 s). The amount of protein used for the analysis is indicated (column: Superdex 200 [Pharmacia]; buffer: 20 mM HEPES, 100 mM KCl, 1 mM EDTA, 1 mM DTT, pH 8.0). The scattering of molecular weights at the edges of the peaks results from a low signal to noise ratio in the refractive index and light scattering signals, which propagates into low precision for the molecular weight determination.

during HPLC SEC, and submission of nonhomogeneous samples. Careful description of the service, of its requirements, and of the "standard operating procedures" (eg, http://info.med.yale.edu/wmkeck/6_16_98/Lsmemoa.htm), as well as the sample submission sheet (eg, http://info.med.yale.edu/wmkeck/ls-sampl.htm), can greatly lessen these problems. For instance, precipitation and quantitation problems generally can be avoided by equilibrating the sample for at least 1 hour in the HPLC SEC running buffer, filtering it, and then measuring the absorbance at 280 and 260 nm. Coupled with an extinction coefficient at 280 nm calculated from the protein's content of tryptophan, tyrosine, and cystine,[20] the former reading can provide a reasonably accurate estimate of the amount of protein that is about to be injected onto the HPLC SEC system.

To ensure that the results of the HPLC SEC/LS analysis derive from the major components of the sample, the overall recovery of protein (usually calculated from the RI trace) must be reasonable. The average recovery of most proteins appears to range from about 60% to 100%. The 280/260 ratio provides a sensitive indicator of nucleic acid contamination, with a ratio below about 1.0 signifying more than 5% nucleic acid by weight.[21]

The appearance of multiple HPLC SEC/LS peaks poses the question of whether these arise from aggregation of single or multiple proteins or merely reflect the presence of contaminating proteins. Although requiring a photograph of an SDS-PAGE gel of the sample provides reasonable assurance of homogeneity, a single SDS-PAGE gel band may still contain multiple proteins, and some proteins may not stain with Coomassie blue or silver. Low MW proteins in particular may be recovered in poor yield after Coomassie blue staining.[22] The RI/UV absorbance ratio of the eluting peaks can be helpful in quickly determining which peaks are composed of only a single protein. The individual HPLC SEC/LS peaks may be collected and subjected to "off-line" two-dimensional gel analysis, MS analysis, and other analyses to positively confirm their identity.

The advantages of HPLC SEC/LS for determining MW and monitoring protein:protein interactions are manifold and include the fast turn around (~1 hour/sample), relative ease of automating the analyses and of data interpretation, ability to carry out studies in solution under "native" conditions, and the possibility of recovering the sample. The amounts of protein required for a single analysis are reasonable, assuming that most proteins being subjected to biophysical studies have been cloned and expressed in bacterial or other systems. HPLC SEC/LS is equally applicable to analyzing nucleic acids (data not shown) and other polymers, and we believe it holds particular promise for monitoring protein–nucleic acid interactions.

CONCLUSIONS

Based on the analysis of 14 standard proteins, HPLC SEC/LS represents a valuable and robust approach to determining MWs and monitoring protein oligomerization in solution under native conditions. When sufficient sample is available to permit multiple analyses, it is possible to achieve a mass accuracy of ±3% and a precision of ±1 kd for proteins that range in size from 12 to 475 kd. When sample amounts are limiting, we demonstrated that even 10 μg is sufficient to determine the MW with an error of less than ±6% (with proteins that are well recovered from HPLC SEC).

Unusual SEC elution position, caused by nonglobular shape or interactions with the SEC matrix or both, has absolutely no effect on the resulting MW determination. One minor limitation found was analyses that involved colored proteins, for which the MWs were underestimated because of absorption of the incident laser light. Other examples of minor SEC/LS limitations were manifested during analyses of unstable oligomers. These proteins were dissociating during the SEC fractionation. Although these proteins eluted in a single peak, the peak was not homogeneous with respect to MW.

Our study suggests that HPLC SEC/LS is technically suitable for inclusion in core laboratories, and we believe that the completion of numerous genome projects will provide a wealth of proteins and other biomolecules in the future that can benefit from this kind of analysis.

REFERENCES

1. Greene EA, Pietrokovski S, Henikoff S, et al. Building gene families. *Science* 1997;278:615–630.
2. Harding SE, Jumel K. Light scattering. *Curr Protocols Protein Sci* 1998;7.8.1–7.8.14.

3. Wyatt PJ. Light scattering and the absolute characterization of macromolecules. *Anal Chim Acta* 1993;272: 1–40.

4. Takagi T. Application of low-angle laser light scattering detection in the filed of biochemistry. *J Chromatogr* 1990;506:409–416.

5. Wen J, Arakawa T, Philo JS. Size-exclusion chromatography with on-line light scattering, absorbance, and refractive index detectors for studying proteins and their interactions. *Anal Biochem* 1996;240:155–166.

6. Wen J, Arakawa T, Talvenheimo J, et al. A light scattering/size exclusion chromatography method for studying the stoichiometry of a protein-protein complex. *Tech Protein Chem* 1996;7:23–31.

7. Philo J, Aoki K, Arakawa T, Narhi LO, Wen J. Dimerization of the extracellular domain of the erythropoietin (EPO) receptor by EPO: one high-affinity and one low-affinity interaction. *Biochemistry* 1996;35:1681–1691.

8. Philo J, Wen J , Wypych J, Schwartz MG, Mendiaz EA, Langley KE. Human stem cell factor dimer forms a complex with two molecules of the extracellular domain of its receptor, kit. *J Biol Chem* 1996;271:6895–6902.

9. Odaka M, Kohda D, Lax I, Schlessinger J, Inagaki F. Ligand-binding enhances the affinity of dimerization of the extracellular domain of the epidermal growth factor receptor. *J Biochem* 1996;122:116–121.

10. Zhu H, Ownby DW, Riggs CK, Nolasco NJ, Stoops JK, Riggs AF. Assembly of the gigantic hemoglobin of the earthworm *Lumbricus terrestris* roles of subunit equilibria, non-globin linker chains, and valence of the heme iron globin. *J Biol Chem* 1996;271:30007–30021.

11. Kumar A, Williams KR, Szer W. Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-stranded DNA-binding proteins. *J Biol Chem* 1986;261:11266–11273.

12. Doty PM, Zimm BM, Mark H. Some light scattering experiments with high polymer solutions. *J Chem Physics* 1944;12:144–145.

13. Zimm BH. The scattering of light and the radial distribution function of high polymer solutions. *J Chem Physics* 1948;16:1093–1116.

14. Jeng L, Blake ST. Evaluation of light scattering detectors for size exclusion chromatography. II. Light scattering equation selection. *J Appl Polymer Sci* 1993;49:1135–1374.

15. Jackson C, Nilsson L, Wyatt PJ. Online absolute measurements of radius and molecular weight of polysterene using size exclusion chromatography and laser light scattering. *J Appl Polymer Sci* 1990;45:191–202.

16. Berry GC. Thermodynamic and conformational properties of polysterene. I. Light-scattering studies of dilute solutions of linear polystyrenes. *J Chem Physics* 1966; 44:4550–4564.

17. Debye P. molecular-weight determination by light scattering. *J Physics Coll Chem* 1947;51:18–32.

18. Bevington PR, Robinson DR. *Data Reduction and Error Analysis for the Physical Sciences,* 2nd ed. New York: WCB/McGraw-Hill, 1992:53–74.

19. Casas-Finet JR, Smith JD Jr, Kumar A, Kim JG, Wilson SH, Karpel RL. Mammalian heterogeneous ribonucleoprotein A1 and its constituent domains: nucleic acid interaction, structural stability and self-association. *J Mol Biol* 1993;229:873–989.

20. Pace CN, Vajdos V, Fee L, Grimsley G, Gray T. How to measure and predict the molar absorption coefficient of protein. *Protein Sci* 1995;4:2411–2423.

21. Glasel JA. Validity of nucleic acid purities monitored by 260 nm/280 nm absorbance ratios. *Biotechniques* 1995; 18:62–63.

22. Williams KR, LoPresti M, Stone K. Internal protein sequencing of SDS-PAGE-separated proteins: optimization of an in gel digest protocol. In Marshak D (ed): *Techniques in Protein Chemistry VIII.* San Diego: Academic Press, 1997:79–90.