Nishizuka, Y., and Lipmann, F. (1966b), *Arch. Biochem. Biophys. 116*, 344.
Okura, A., Kinoshita, T., and Tanaka, N. (1970), *Biochem. Biophys. Res. Commun. 41*, 1545.
Parmeggiani, A., and Gottschalk, E. M. (1969), *Biochem. Biophys. Res. Commun. 35*, 861.
Raeburn, S., Goor, R. S., Schneider, J. A., and Maxwell, E. S. (1968), *Proc. Nat. Acad. Sci. U. S. 61*, 1428.
Randerath, K., and Randerath, E. (1964), *J. Chromatogr. Sci. 16*, 111.
Siler, J., and Moldave, K. (1969), *Biochem. Biophys. Acta 195*, 138.
Skogerson, L., and Moldave, K. (1968a), *Arch. Biochem. Biophys. 125*, 497.
Skogerson, L., and Moldave, K. (1968b), *J. Biol. Chem. 243*, 5354.
Smulson, M. E., and Rideau, C. (1970), *J. Biol. Chem. 245*, 5350.
Sutter, R. P., and Moldave, K. (1966), *J. Biol. Chem. 241*, 1698.
Tanaka, N., Kinoshita, T., and Masukawa, H. (1968), *Biochem. Biophys. Res. Commun. 30*, 278.

Quantitative Procedures for Use with the Edman–Begg Sequenator. Partial Sequences of Two Unusual Immunoglobulin Light Chains, Rzf and Sac*

O. Smithies,† D. Gibson,‡ E. M. Fanning, R. M. Goodfliesh, J. G. Gilman, and D. L. Ballantyne

ABSTRACT: Two methods of hydrolysis, with HI and with NaOH + $Na_2S_2O_4$, are described which permit amino acids to be regenerated from the thiazolinones produced by the Edman–Begg sequenator. Neither requires any additional extractions or prior conversion of the thiazolinones to other derivatives. Use of both methods enables all residues normally encountered in the automatic degradation of proteins to be unambiguously identified and quantitated with an amino acid analyzer, except that cysteine is not distinguishable from serine without some additional manipulation. Equations are developed to correct the resulting data for the systematic errors of automatic sequencing and to reduce them to an easily inspectable form in which all the quantitative information relevant to a sequence determination can be displayed in a single graph. An outline of computer programs to process the data is given. The procedures are illustrated by partial sequence determinations for 67 and 50 positions, respectively, of two smaller than usual immunoglobulin light chains. The two chains are Rzf [Deutsch, H. F. (1965), *Immunochemistry 2*, 207] and Sac [Lewis, A. F., Bergsagel, D. E., Bruce-Robertson, A., Schachter, R. K., and Connell, G. E. (1968), *Blood 32*, 189]. Rzf shows no evidence of a deletion in the region sequenced, but the light chain Sac has a major deletion of about 68 residues in its variable region.

The sequenator of Edman and Begg (1967) has already proved of considerable value in sequencing purified proteins (see, for example, Niall and Edman, 1967; Niall *et al.*, 1970). When mixtures of proteins are unavoidable, as with the light chains of antibodies from normal individuals, the usefulness of the sequenator can be considerably enhanced by making the procedure quantitative. In this paper we describe methods permitting the recognition and quantitation by regular amino acid analysis of all amino acids normally encountered in the sequential degradation of proteins. Equations are developed to correct the resulting data for systematic errors of automatic sequencing and to reduce them to an easily inspectable form. In an Appendix programs are described for the reduction of these data by a computer, but it can be done manually. The application of the quantitative procedures is illustrated by the partial sequencing of two unusual immunoglobulin light chains, Rzf (Deutsch, 1965) and Sac (Lewis *et al.*, 1968; Smithies *et al.*, 1971; Parr *et al.*, 1971). Quantitative data obtained by application of the same procedures to a study of mixtures of immunoglobulin light chains from normal humans and rabbits will be described elsewhere.

Materials and Methods

Proteins. Much of the work in developing the methods was carried out with a typical κ_{II} Bence-Jones protein, Dil. In preparing this protein, the patient's urine was deionized with 0.05–0.1 its volume of 20–50 mesh mixed bed resin, filtered, and then treated with approximately 6 g of water-washed DEAE-cellulose in its acetate form per liter of urine. These procedures were repeated, if necessary, until essentially all protein was removed from the urine. The batches of DEAE-cellulose on which the protein was absorbed were combined and packed in a column on top of approximately twice the

* Paper No. 1482 from the Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706. *Received July 30, 1971.* Supported by the National Institutes of Health (GM 15422 and FR 07098), the National Science Foundation (GB 4362), the Wisconsin Alumni Research Foundation, and a training grant (GM 00398) from the National Institute of General Medical Sciences.

† Author to whom correspondence should be addressed.

‡ Recipient of a postdoctoral fellowship from the National Research Council of Canada. Present address: Department of Biochemistry, University of Sherbrooke, Sherbrooke, Quebec, Canada.

quantity of DEAE-cellulose also in the acetate form and previously washed with distilled water. A linear gradient of unbuffered ammonium acetate (0–0.2 M) was used to elute the protein from the DEAE-cellulose. In some experiments a bulk elution with 0.1 M ammonium acetate was used. The purity of the product was checked by electrophoresis in alkaline and acidic starch gels.

Rzf is a crystalline fragment of a Bence-Jones protein and was a gift from Dr. Harold Deutsch, University of Wisconsin; it had been isolated from the urine of a myeloma patient together with normal-sized Bence-Jones protein which was also crystallized (Deutsch, 1965). The exact nature of the Rzf fragment has not yet been determined, although the sequence data presented here together with other currently available data suggest that it may be the variable region of the molecule (compare Baglioni *et al.*, 1967). Further work is in progress to determine whether this is correct.

The protein Sac is the light chain of a myeloma-like immunoglobulin and was prepared by Parr *et al.* (1971). This immunoglobulin was originally characterized by Lewis *et al.* (1968) who had provisional evidence that both the light and heavy chains had major deletions.

Edman–Begg Sequenator, a prototype from the Illitron Division of the Illinois Tool Works, Chicago, Ill., was designed by Robert G. Kobetsky following the Edman–Begg specifications closely. The most significant design modifications were the introduction of a cooled copper cylinder axial to the spindle permitting the base of the glass cup to be kept cooler than the neighboring vacuum-seal areas, the installation of a continuously flowing oil supply between the vacuum seals, and the use of a liquid-nitrogen trap between the instrument and the mechanical vacuum pump.

Sequenator Procedure. The sequenator machine program adopted as standard differs from that of Edman and Begg only in minor ways. Deviations in the chemical procedure were the inclusion of 0.001 M dithiothreitol in the three solvents, the inclusion of 0.001 M ammonium formate in the Quadrol[1] buffer, and the loading of the sample in the presence of dithiothreitol. The addition of dithiothreitol to the solvents was aimed at reducing the consequences of any oxygen leaks into the system. It also permits the protein to be applied with intact disulfide bonds since disulfide cleavage can be carried out in the instrument without the need for subsequent protection of the protein sulfhydryl groups formed. This avoids an extra operation on the protein before sequencing and, more importantly, takes advantage of the fact that proteins are frequently more soluble when native than after their disulfide bonds are cleaved. The loading procedure was designed to enable the disulfide cleavage to be performed during the first cycle. Typically, the sample (up to 10 mg) was applied to the sequenator dissolved in 0.2 ml of 0.1 M formic acid containing 0.001 M ammonium formate and 1 mg of dithiothreitol. After drying down the sample, Quadrol buffer was added (the dithiothreitol trapped in the protein seems to increase its ease of solution) followed by 5 min of delay during which time the dithiothreitol can cleave the disulfide bonds. Phenyl isothiocyanate (PITC) reagent was then added and the standard Edman–Begg program picked up at stage 3. In subsequent cycles the reagents were added in the normal order. We found that, in addition to the advantages derived from the application procedure, dithiothreitol in the solvents markedly reduces the formation of intractable protein gels during long runs. Hermodson *et al.* (1970) have found that the inclusion of a thiol, in their case dithioerythritol, in the butyl chloride improves the stability of the sequenator products. Our addition of 0.001 M ammonium formate to the Quadrol buffer was precautionary, being aimed at reducing the effects of any aldehydes which might be present in the Quadrol, but we do not have quantitative data to support its effectiveness.

Many parameters were investigated prior to adopting the standard procedure of Edman and Begg with the small chemical modifications indicated above. Some unnecessary duplication of future efforts by other investigators may be avoided by listing the variables that were examined.

Mechanical variables investigated included the following: one *vs.* two vacuum traps; traps in various places other than as described above; Dry Ice *vs.* liquid nitrogen as trap coolant; a diffusion pump in addition to the mechanical vacuum pump; a higher speed of cup rotation. None of these had any marked effects.

Procedural variables investigated included the following: 0, 10^{-3}, 10^{-4} M dithiothreitol in some or all of solvents; 0.5, 1, and 1.5 times the Quadrol coupling reaction time; 1 or 2 coupling stages; inclusion of 0.001 M NH_3 in the Quadrol; 0.33, 0.5, 1, and 2 times HFBA reaction time; 2 or 3 acid-cleavage steps; first cycle without any PITC; application of sample under N_2; sample as a suspension in water or in 50% aqueous acetone; sample dissolved in 4 M, 2 M, 0.1 M formic acid, dilute aqueous triethylamine, pH 9 triethylamine–formic acid buffer, or 0.1 M formic acid together with 0.1 M cysteamine; protein sample oxidized with performic acid or Br_2–water; protein reduced with dithiothreitol and alkylated with iodoacetamide; protein as mixed disulfide with (*S*)-aminoethane; protein sample with 1 or 4 mg of dithiothreitol. None of these variations gave consistently better results than the standard procedure described above.

Reagents. In most of the early work we used Pierce Chemical Co., Rockford, Ill., Sequanal grade reagents and solvents, but in the later work the benzene and ethyl acetate were Burdick and Jackson Laboratories, Muskegon, Mich., glass-distilled solvents. No change in the quality of the results was observed when the less expensive solvents were used. No further purifications were carried out with any of the commercial reagents or solvents and only normal care was taken in handling them. Dithiothreitol and ammonia were used as additives, as described above.

Sequenator Sample Processing. Two methods of hydrolysis were employed to regenerate the parent amino acids from their thiazolinones; neither requires any extractions or prior conversion of thiazolinones to phenylthiohydantoins (PTH's) or other derivatives. The methods are essentially those developed by Gibson (in preparation) with modifications to permit large numbers of samples to be routinely processed with few accidents and minimal handling.

The primary method was an acid hydrolysis with HI, which enables the following residues to be recognized and quantitated: methionine sulfone, aspartic acid, asparagine (as aspartic acid + ammonia), glutamic acid, glutamine (as glutamic acid + ammonia), proline, glycine, alanine or serine or carboxymethylcysteinamide or cysteine (as alanine), threonine (as α-aminobutyric acid, Aab), valine, isoleucine (as alloisoleucine + isoleucine), leucine, tyrosine, phenylalanine, histidine, lysine, tryptophan (as glycine + alanine), arginine. Small amounts of breakdown products are occasionally found

[1] Abbreviations used are: Aab, α-aminobutyric acid; HFBA, heptafluorobutyric acid; PITC, phenyl isothiocyanate; PTH, phenylthiohydantoin; Quadrol is a registered trademark of the Wyandotte Chemical Corp., Wyandotte, Mich., for the substance *N,N,N',N'*-tetrakis(2-hydroxypropyl)ethylenediamine; TDG, thiodiglycol.

with proline and leucine in addition to the parent amino acids; the product from proline elutes during amino acid analysis at the same time as histidine in our analyzer; the product from leucine is slightly later. Methionine as such is destroyed but gives a small amount of degradation product eluting between histidine and lysine.

The second method of hydrolysis was with NaOH–dithionite. This method enables methionine and tryptophan to be recovered unaltered, methionine in close to theoretical yield and tryptophan in about 50% yield. Alanine is recovered, but serine or cysteine or carboxymethylcysteinamide is almost completely destroyed. This permits alanine to be distinguished from serine or cysteine, but an extra procedure is needed to distinguish serine from cysteine. Asparagine and glutamine give less than half the yield with NaOH that is obtained from aspartic and glutamic acids. Threonine is almost completely destroyed but gives a little α-aminobutyric acid. Proline sometimes gives a small amount of breakdown product in addition to proline. Glycine, valine, isoleucine (as alloisoleucine + isoleucine), leucine, tyrosine, phenylalanine, histidine, and lysine are also recovered with this hydrolysis method. Arginine is usually almost completely destroyed but sometimes gives a breakdown product which is probably ornithine. By using both methods of hydrolysis all amino acids can be recognized and measured except that cysteine is not distinguishable from serine without some additional manipulation such as prior oxidation of the protein to convert all cysteines to cysteic acid, which is no longer recovered as alanine after HI hydrolysis, or the use of radioactive label (Leroy Hood, personal communication) to show the presence of carboxymethylcysteine. Table I below lists typical overall recoveries to be expected with both hydrolytic methods (overall recoveries include the efficiency of butyl chloride extraction as well as the efficiency of hydrolysis).

The butyl chloride extracts from both cleavage steps for each cycle of the sequenator were collected as a single fraction in 18 × 150 mm Pyrex disposable culture tubes. To each tube was added 20 μl of PTH norleucine solution (2.5 mM dissolved in butyl chloride containing 0.001 M dithiothreitol) to act as a semi-internal standard and the tubes were dried at room temperature on a Buchler Evapo-Mix. The tubes were then transferred to a 170-mm Pyrex crystallizing dish in a 250-mm Pyrex desiccator and further dried on a high-vacuum pump for at least 15 min. The samples can be stored at this stage for several weeks under vacuum at freezer temperatures. We have also kept samples in the freezer for shorter periods while still in solution in the butyl chloride without noticing any differences in the final results. For the most reproducible results the actual hydrolysis of *all* the samples for a given run should be done in a single batch and yields calculated relative to the yield of the semi-internal standard, norleucine.

HI (57%; Fisher Chemical A-135) was used for the acid hydrolysis. This material contains hypophosphorous acid as a preservative which in some experiments was oxidized before using the acid for hydrolysis. To do this, solid I_2 was added to the HI 1 hr before use, the amount being adjusted to yield a slight excess of I_2 as judged by a residual pale orange color. Extensive comparative tests of HI *vs.* HI–I_2 were not made, and we have recently begun to use the preserved HI without added I_2 as our standard procedure; 50 μl of HI was placed in each tube and approximately 50 ml in the bottom of the desiccator. The tubes in the 170-mm crystallizing dish were covered with a 190-mm crystallizing dish to prevent dried-out silicone grease from falling into them at the end of the hydrolysis. The desiccator was sealed with Dow Corning 11 silicone compound, evacuated, flushed three times with N_2, and finally evacuated once again. A clamping ring was placed around the lid of the evacuated desiccator and hydrolysis was carried out in an autoclave at 21 lb pressure (127°) for 20 hr. After the hydrolysis, the desiccator was reevacuated while still hot to dry down the samples. Each was dissolved in 0.2 ml of 0.1 M HCl containing 1% v/v thiodiglycol (TDG) and applied to Technicon sample cartridges in the H^+ form for amino acid analysis.

The following procedure was used for the alkaline hydrolysis. After the samples had been dried down on the Buchler Evapo-Mix, the last traces of HFBA were removed by placing them in a vacuum oven at room temperature, increasing the temperature to 60° for 1 hr, and allowing the samples to cool under vacuum. A 0.1 M solution of $Na_2S_2O_4$ (sodium dithionite) in 0.2 M NaOH was made by degassing 0.2 M NaOH, bubbling N_2 through it, adding solid $Na_2S_2O_4$, and bubbling again with N_2; 0.2 ml of this solution was added directly to the vacuum oven-dried samples in the 18 × 150 mm tubes. The tubes were placed in a desiccator with sufficient extra NaOH–dithionite solution in the bottom to ensure contact of each tube with liquid; if a crystallizing dish was used, solution was also added between it and the outside of the desiccator. These procedures help prevent any of the samples from drying out or being diluted during the heating-up and cooling period. The desiccator was sealed with Dow Corning 11 compound, evacuated, and clamped. Hydrolysis was carried out for 3.5 hr in an autoclave at 21 lb pressure (127°). The desiccator was then cooled in water and 20 μl of 3 N HCl containing 10% TDG was added to each tube prior to sample loading.

Amino acid analyses were carried out with a Technicon TSM-1 amino acid analyzer modified for the "1-hr system." In place of the two columns normally used with this system, the long column at 60° was used with three buffers all containing 1% v/v TDG and approximately 1.5% v/v of 30% w/v Brij 35. An important feature of this analytical system is that the time per run (90 min) is short enough to permit 16 analyses per 24 hr, which is adequate to keep up with the sequenator (14 cycles per 24 hr) without sample splitting. The buffers pumped at 0.5 ml/min were: 10 min of 0.1 M citric acid–0.2 M NaOH with HCl to pH 3.25; 30 min of the same solution at pH 4.20; 24 min of 0.3 M NaCl–0.07 M NaOH plus solid boric acid to pH 9.35; 2 min of 0.2 M NaOH; 20 min of pH 3.25 buffer; sample change. This system gives essentially base line resolution of all the amino acids that occur in HI and NaOH hydrolysates of the amino acid thiazolinones. Samples were loaded onto cartridges that were in the acid form. The place normally used in the TSM-1 analyzer for the basic column was used to prepare these acid cartridges with the following solutions pumped at 0.25 ml/min: 3 min of 0.1 M HCl containing 1% v/v TDG; 17 min of 0.2 M NaOH; 66 min of 0.1 M HCl with TDG; sample change. To keep the analytical column free from basic ninhydrin-positive contaminants in the acidic buffers that elute with the pH 9.5 buffer, the 3.25 and 4.20 buffers were run through a 6 mm × 20 cm column of Fisher Chemical Company Rexyn 101 sulfonic acid resin 40–100 mesh in the Na^+ form. This column was placed before the inlet of the high-pressure pump. During the pH 9.35 stage of each analysis it was automatically stripped of NH_3, etc., with 0.2 M NaOH and reequilibrated with the pH 3.25 buffer. The data for each analysis were recorded graphically in the usual way and also on punched paper tape as times and areas determined by an Infotronics CRS110A

integrator. Computation of the results was carried out by the programs described in the Appendix.

Results and Discussion

General Characteristics of Quantitative Sequencing. The general characteristics of a quantitative sequencing experiment can be illustrated by plotting the yield in nanomoles of a given amino acid recovered after hydrolysis of the thiazolinones against the number of degradations performed. Figure 1A shows such a plot for the yield of leucine (after HI–I_2 hydrolysis) obtained during 40 degradations with 240 nmoles of the Bence-Jones protein Dil. Several general features are apparent: the "background" rises, the yield of the "in-step" residues fall off, and the sequencing gets "out-of-step." These features rarely lead to any ambiguous sequences when sequencing a pure protein until the yields fall below the analytical noise levels, but they can be serious when attempting to sequence mixtures of proteins. It is, for example, easy to confuse the out-of-step contribution of the major component in a mixture with the in-step contribution of a minor component if the two happen to have the same amino acid one residue apart. These various systematic errors can be corrected by computation.

An adequate base line can be drawn in by hand on a graph such as that shown in Figure 1A to permit subtraction of the "background" of each residue at each position, or the data can be processed by the computer using the base line equation developed below. The approach adopted in making corrections for the "out-of-stepness" involves the following considerations. The desired outcome of a cycle in a sequenator experiment is that the PITC coupling and HFBA cleavage should be successful. We define the probability of this occurring as α_n for the nth cycle. If the possibility of cleaving more than one residue in a given cycle can be neglected, then the only way that the nth residue will be observed at the nth cycle is for the reaction to succeed all n times. In over 300 sequencing experiments, many with proteins of known homogeneity, we have no evidence that double cleavage occurs at a detectable level. Edman (1970) makes this same comment and also notes that histidine presents no special problem. We can confirm that degradation continues past histidine in the usual way from several experiments with hemoglobin. Consequently the yield of the nth residue at the nth cycle can be taken as being proportional to $\alpha_1\alpha_2\alpha_3 \ldots \alpha_n$ or α^n where α is the geometric mean of the n individual values of α_n. (Other factors by which this expression must be multiplied to give the actual yield are considered below.)

If either the HFBA cleavage or the PITC coupling is unsuccessful in any cycle, out-of-step degradations will result. We define β_n as the probability that during the nth cycle either an unsuccessful coupling and/or an unsuccessful cleavage occurred without irreversible chemical or mechanical loss of the relevant molecule. (The probability of irreversible loss is $1 - \alpha_n - \beta_n$.) The nth residue will be recovered at the $(n + 1)$th cycle if there was a failure of coupling or cleavage at any one of the preceding n cycles and if the $(n + 1)$th cycle was successful. The recovery of the nth residue at the $(n+1)$th cycle is consequently proportional to $(\beta_1\alpha_2\alpha_3 \ldots \alpha_{n+1}) + (\alpha_1\beta_2\alpha_3 \ldots \alpha_{n+1}) + \ldots + (\alpha_1\alpha_2\alpha_3 \ldots \beta_n\alpha_{n+1})$ or approximately $n\beta\alpha^n$ where β is the arithmetic mean of $\beta_1, \beta_2, \ldots \beta_n$. Two or more failures can lead to the recovery of the nth residue at the later cycles, $(n + 2)$th and $(n + 3)$th etc., approximately in proportion to $n(n + 1)/2\beta^2\alpha^n$ and to $n(n + 1)(n + 2)/(3 \times 2)\beta^3\alpha^n$ etc. The calculation of α and β is straightforward. For example,

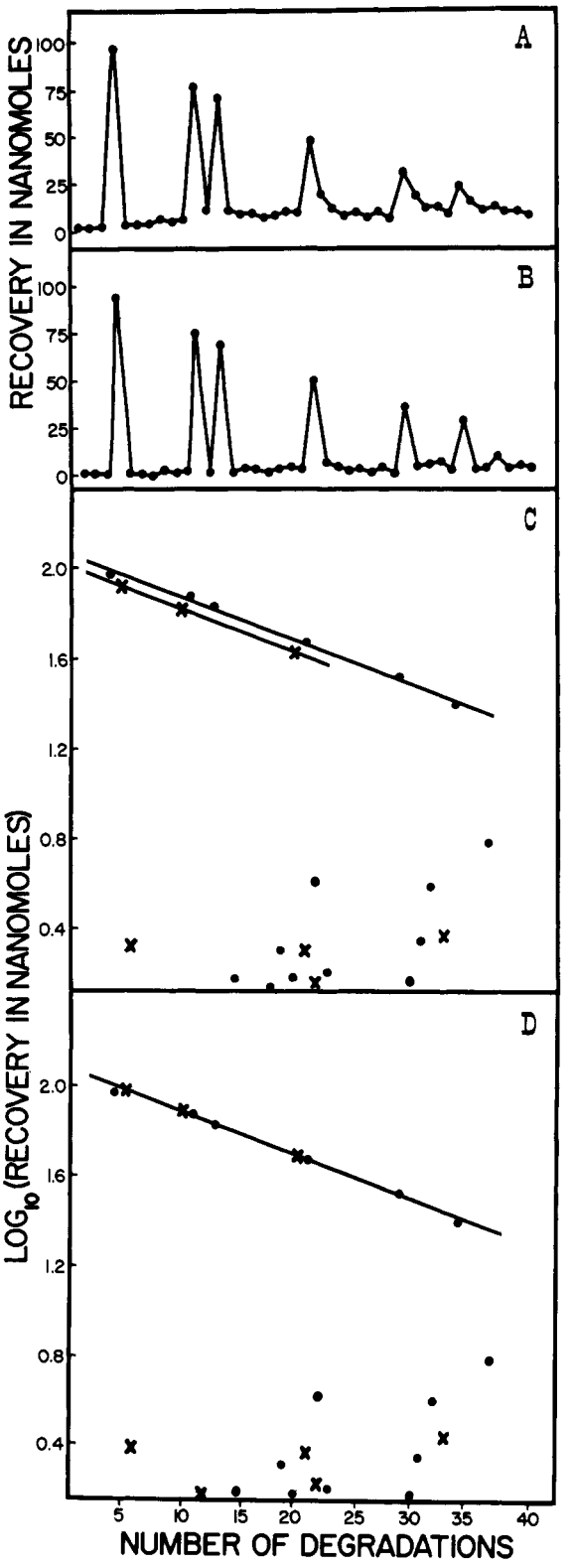

FIGURE 1: Stages in the processing of the data. (A) The yield in nanomoles of leucine, ●, recovered after HI hydrolysis of the thiazolinones plotted against the number of degradations performed. (B) Same data after corrections for "background" and "out-of-step" degradations with out-of-step recoveries added back onto in-step recoveries. (C) Semilog plot of the corrected leucine data with threonine, ×, data also added. (D) Semilog plot of data corrected for threonine recovery; 240 nmoles of the Bence-Jones protein Dil were used.

TABLE I: Typical Ranges of Individual Amino Acid Recoveries (Relative to an Average Recovery for Leucine of 1.00) after Acid and Alkaline Hydrolyses of the Sequenator Products.[a]

Residue in Sequence	HI Hydrolysis		NaOH Hydrolysis	
	Observed	R_i	Observed	R_i
Asp	Asp	0.8–1.2	Asp	0.8–1.2
Asn	Asp (+ NH_3)	0.8–1.2	Asp	0.5–0.7
Thr	Aab	0.9–1.1	Aab	0.01–0.2
Ser	Ala	0.5–1.0	Ala	0.05–0.2
Glu	Glu	0.8–0.9	Glu	0.5–0.7
Gln	Glu (+ NH_3)	0.8–1.0	Glu	0.2–0.4
Pro	Pro	0.7–1.0	Pro	0.7–0.9
Gly	Gly	0.8–1.0	Gly	0.8–1.0
Ala	Ala	0.9–1.2	Ala	0.8–1.3
Cys	Ala	0.6–1.0	Ala	0.05–0.1
Val	Val	0.7–0.9	Val	1.3–1.5
Met	Destroyed	0	Met	0.9–1.2
Ile	alloIle + Ile	0.7–1.0	alloIle + Ile	1.0–1.4
Leu	Leu	0.9–1.1	Leu	0.9–1.1
Tyr	Tyr	0.6–0.8	Tyr	1.0–1.2
Phe	Phe	0.8–1.2	Phe	1.0–1.4
His	His	0.2–0.8	His	0.2–0.5
Trp	Gly + Ala	0.6–1.0	Trp	0.4–0.7
Lys	Lys	0.7–0.9	Lys	0.3–0.5
Arg	Arg	0.3–0.6	Arg + Orn	0.05–0.3

[a] The recovery factors include the efficiency of butyl chloride extraction as well as the efficiency of the hydrolysis. The absolute recoveries of alanine and leucine with HI and of valine with NaOH are close to theoretical (see text).

$\alpha^{11}/\alpha^4 = \alpha^7$ = (yield of leucine above background at 11th cycle)/(yield above background at 4th cycle), and 11β = (yield of leucine above background at 12th cycle)/(yield above background at 11th cycle).

Using the equations given above and the parameter β the out-of-step contributions can then be calculated and substracted from the relevant steps. In making the correction with the computer for this experiment a smooth function was used for β which varied linearly from 0 before step 1 to 0.018 after 40 steps. The actual values of β measured after the 4th, 11th, 13th, 21st, 29th, and 34th cycles were 0, 0.005, 0.004, 0.014, 0.016, and 0.014. As can be seen, in this experiment β increased during the run. This could be due to at least two factors. Some peptide bonds may be more difficult to cleave than others and so might give rise to a high β_n for a single step which then leaves the (average) value of β higher for many subsequent steps. We have not observed that any particular residues (all have been encountered) are consistently more difficult to cleave than any others, but we have observed that in some proteins there may be a peptide bond that is reproducibly difficult to cleave. This is uncommon and we have insufficient data to say whether it is due to particular pairs of adjacent residues. Another possible cause of a changing β is that the levels of the Quadrol buffer and HFBA in the cup are not correctly adjusted (see Edman and Begg, 1967), so allowing protein to rise up in the cup to a position that in some cycles is inaccessible to the coupling reagents. We have some evidence that this second effect can be important, since at one time we had a series of experiments in which β was particularly high (around 0.035) but on increasing the amount of Quadrol relative to the HFBA the value of β decreased markedly to about 0.020. Increasing the reaction temperature can also cause a decrease in β, but the undesirable side effects of higher temperatures (see below) prevent unlimited use of this means of reducing out-of-stepness.

The parameters α and β are interdependent. A high value of β at any step must always be accompanied by a correspondingly lower value of α, since $(\alpha + \beta)$ cannot exceed 1. On the other hand, a low β is not always accompanied by a high α if, for example, mechanical or chemical losses are high. In optimizing overall recoveries it is helpful to keep the temperature as low as is consistent with an acceptable value of $\beta < 0.02$. (Note, however, that when β is equal to 0.02, which corresponds to 2% out-of-step degradations per cycle, the yield of the 50th residue at the 50th and 51st cycle are equal.) The temperature of the nitrogen in the reaction chamber that we selected in this way is 52°.

Since residues that are cleaved at out-of-step positions are eventually recovered, albeit late, the data were usually further corrected by adding back the out-of-step recoveries to the in-step recoveries. The expected out-of-step recovery is first calculated from β and the in-step value. The expected recovery is then compared with the observed out-of-step recovery. The smaller of these two values is then added back. In this way the add-back values are never allowed to exceed the values actually observed at the out-of-step position (although they can be less). Figure 1B shows the data from 1A after correcting for background and for out-of-step degradations using the add-back feature.

The corrected data shown in Figure 1B can be represented in a logarithmic form, as shown by the solid points and upper line in Figure 1C. The analytical data for threonine (recovered as α-aminobutyric acid in the HI–I_2 hydrolysis) are shown by crosses and the lower line in the same figure. The two lines are parallel, and displaced relatively by 0.065 log unit, equivalent to the constant factor 0.86. This factor permits us to define and determine parameters related to the recoveries of each amino acid. These are R_i, R_j, etc., which are the overall recoveries for the *i*th, *j*th, etc., amino acids relative to the standard amino acid leucine to which the recovery factor unity is arbitrarily assigned. The value 0.86 for R_{Thr} implies that the product of the extraction efficiency and of the efficiency of the HI–I_2 hydrolysis for threonine in this experiment was such that its overall yield is 0.86 times the yield for leucine.

Since the values of R_i, R_j, etc., are subject to the specific conditions of cup temperature, length of vacuum stages, extraction schedules, etc., and to the exact hydrolytic methods in use at any one time, they need to be determined by each investigator for each instrument under a given set of conditions. However, as a guide to the orders of magnitude to be expected, the range of typical individual values for both hydrolytic methods is presented in Table I. The fact that none of the R_i values with the HI method is significantly greater than one (with the possible exception of the value for alanine) indicates that leucine is recovered in equal or better amounts than other residues when using HI hydrolysis. On the other hand the recoveries of valine, and possibly also of several other amino acids, are significantly better than of leucine with the NaOH hydrolytic method. The recoveries of histidine and arginine appear to be particularly sensitive to the amount of HFBA removed after the cleavage stage; when the cup temperature is low, and less of the HFBA is removed, their yields

decrease. The hydrolysis of arginine with the NaOH method has been erratic with respect to yield and in the extent of its conversion to ornithine.

Figure 1D shows the leucine and threonine data plotted logarithmically after making a correction for R_{Thr}. The slope of the line corresponds to a value of 0.956 for $(\alpha + \beta)$, *i.e.*, to a relative repetitive yield of 95.6% per cycle. If the line plotted in Figure 1D is extrapolated back to zero degradations, the absolute initial yield of amino terminus, including losses from all parts of the procedure, can be determined accurately. In this experiment it was 122 nmoles, corresponding to a recovery of 51% of the applied protein. (This absolute initial yield is not to be confused with the 95.6% relative repetitive yield of subsequent degradations.) In other experiments with the same protein Dil the absolute initial yield has been as high as 60% but we have been unable to find conditions that reproducibly give a significantly better initial recovery with Dil. The less than theoretical yield is probably due to some property of this particular protein preparation, since in the experiment described below with the protein Rzf the absolute initial recovery was approximately 90% of the protein applied (a slight uncertainty in the molecular weight of Rzf prevents the calculation of a more precise yield). Our combined data indicate that the most favorable amino acids, alanine and leucine for HI and valine for NaOH, can be recovered in close to theoretical yields in good experiments, and that even the least good residues (see Table I) can be obtained in greater than 30% yield by one or other of the two hydrolytic methods. In practice it is often unnecessary to make any corrections for the recovery factors of the different amino acids.

The chemical basis for the increased background of amino acids observed during the course of a sequencing experiment was discussed briefly by Edman and Begg (1967). They suggested that the background was probably proportional to the amino acid composition of the protein and Edman (1970) noted that it was more pronounced with larger proteins. The suggestion was made that it might be due to nonspecific cleavage along the polypeptide chain. To put the background on a quantitative basis we define a hydrolytic coefficient, h, as the average probability that any given peptide bond will be nonspecifically cleaved during a given cycle. The actual probability of cleaving a given peptide bond is dependent on the nature of the residues on either side of the bond, but an average value of this probability is still a useful practical concept. The average number of "nicks" generated per cycle in a single molecule of a protein of length N residues can be calculated using h by summing the weighted probabilities of one, two, three, etc. nicks; this average number is approximately Nh. The mole fraction N_i/N of the ith amino acid will determine what proportion of the new amino termini will be of the ith type, and the recovery factor R_i will govern its recovery as an amino acid following extraction and hydrolysis. Additional nicks are generated on each successive cycle, although in progressively decreasing numbers since the length of the protein is decreasing. These nicks all contribute to the background. But the newly generated nonspecific amino termini decrease progressively as does the true amino terminus by the factor $(\alpha + \beta)$ per cycle. The equation combining these relationships is [background of ith type of amino acid after n cycles, in moles per mole of protein] $= R_i(N_i/N)h[N(\alpha + \beta)^n + (N - 1)(\alpha + \beta)^{n-1} + (N - 2)(\alpha + \beta)^{n-2} \ldots (N - n + 1)(\alpha + \beta)]$. This equation is used in fitting base lines to the data during the computer handling of quantitative experiments; it shows how the background is related to the composition and length of the protein, and to the hydrolytic coefficient, h. The data in Figure 1B were derived from those in 1A by the application of this equation.

The hydrolytic coefficient (and so the background) would be expected to be temperature dependent. In the extensive experiment described below, the accidental turning-off of a circulating cooling pump for part of one cycle caused a sudden rise in the background which persisted for many cycles; this is readily explained in terms of a transient increase in h. An additional factor contributes to the background besides nonspecific cleavage and renders the calculation of h of only limited value with the HI hydrolysis method; it is that any protein or peptide extracted by the butyl chloride will be hydrolyzed by the HI and will also give rise to background amino acids. This effect is sufficiently severe that the background (and the calculated value of h) may be nearly twice as high when the HI hydrolysis method is used as with NaOH. The NaOH procedure hydrolyzes peptides to only a limited extent and so can give a more accurate value of h. For this reason alkaline hydrolysis of samples suspected of gross protein or peptide contamination may help in determining what part of a result is due to contamination. Another useful index of protein contamination in samples is the ratio of alloisoleucine to isoleucine in the analyses. Proteins or peptides give isoleucine only on hydrolysis; the thiazolinone of isoleucine gives alloisoleucine in addition to and at 1.3 times the level of isoleucine.

One other consequence of the nonspecific hydrolytic "nicking" of proteins during sequencing is probably important in addition to the background effect—a nick terminates further sequencing of the relevant molecule distal to the lesion. For a protein molecule to be successfully sequenced for n residues, the first n peptide bonds must survive through the first cycle, $n - 1$ through the second, $n - 2$ through the third, and so on. The expression for this survival factor is $(1 - h)^n(1 - h)^{n-1} \ldots (1 - h) = (1 - h)^{n(n+1)/2}$ or approximately $1 - hn^2/2$. A typical value of h in our system is around 0.0002 as calculated from the background equation with NaOH hydrolysis. This means that with a protein of 200 residues about 4% of all molecules will be nicked somewhere in their length at each cycle. After 50 cycles, this amount of nicking will cause the yield of the amino terminus to be reduced to around 77% of that without any nicking; after 60 cycles the yield will be 69% and after 70 cycles 61%. This fall-off due to nicking is one of the factors which limits the extent to which the reaction temperature can be increased in attempts to decrease β. It may account in part for the fall-off from linearity of the logarithmic recovery of the major residue as the number of degradations increases.

Partial Sequence of Bence-Jones Protein Fragment Rzf. The application of the quantitative sequencing procedures can be illustrated with the Bence-Jones protein fragment Rzf. As indicated in materials and methods, this protein is approximately half the size of a usual Bence-Jones protein; it is crystalline and of excellent homogeneity (Deutsch, 1965). The results with it are among the best we have obtained with an average repetitive yield $(\alpha + \beta)$ for the first 40 degradations of 95% and an extrapolated absolute initial yield of approximately 90% of the applied protein.

Figure 2 shows the final log plot of *all* the analytical data after HI–I_2 hydrolysis of the thiazolinones from 67 degradations of the Bence-Jones protein fragment, Rzf. The plot was obtained as for Figure 1D, that is with background corrections and with add-back of out-of-step recoveries, but with recovery corrections in this particular experiment being made for only those amino acids which had an R_i consistently less than 0.8 (serine, tyrosine, lysine, and arginine which had the values

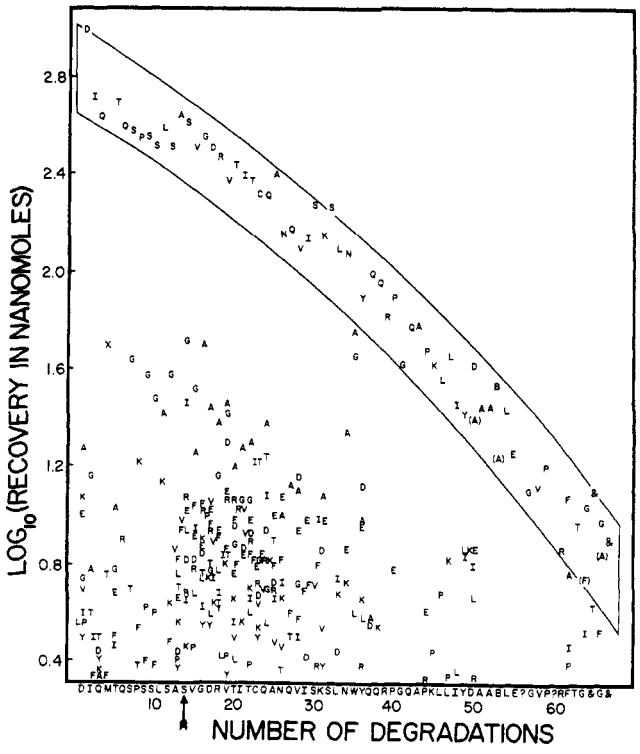

FIGURE 2: Corrected semilog plot of the amino acid recoveries in nanomoles after HI hydrolysis of the thiazolinones obtained in the degradation of the Bence-Jones fragment Rzf. The outlined area is the "sequence box" referred to in the text. The vertical arrow indicates the point at which the reaction temperature was accidentally and briefly increased (see text). The sequence deduced from this experiment and from a second experiment with NaOH–dithionite hydrolysis is shown below the graph in the one-letter code for amino acids: A, alanine; B, aspartic acid or amide; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, theonine; V, valine; W, tryptophan; Y, tyrosine; Z, glutamic acid or amide. An additional symbol, &, is used for serine, alanine, or cysteine. The number of nanomoles of protein used in the experiment was 750.

0.70, 0.60, 0.70, and 0.55). A constant value for β of 0.0125 was used in correcting for out-of-step recoveries. Since the data were obtained with the HI–I_2 hydrolysis method serine and cysteine cannot be distinguished from alanine. We have, however, plotted the relevant points as serine, alanine, or cysteine on the basis of a later sequencing experiment with NaOH hydrolysis (which destroys serine and cysteine and so identifies the alanine residues which are not destroyed) together with the assumption that cysteine and not serine is at position 23 (the cysteine at this position is essentially invariant in Bence-Jones proteins and amino acid analysis of Rzf shows the expected number of cysteines per molecule).

The figure shows that at 59 of the 67 positions plotted the residue in the sequence is unambiguously identifiable; at each of these positions only one analytical value falls within the outlined sequence box. No evidence for double cleavages is found since at no position is there any appreciable yield of the residue that will be found at the next position. No residue is in the sequence box at position 4, but there is a small amount of an unidentified amino acid, labelled X, at about 10% of the expected level; this is the HI breakdown product of methionine, which the experiment with NaOH hydrolysis showed was at position 4. Position 35 shows both glycine and alanine, but again they are below the sequence box. This result with HI hydrolysis suggests the presence of tryptophan at 35 which was confirmed in the NaOH hydrolysis experiment. The residue at position 56 could not be identified because of a considerable contamination of the butyl chloride fraction with protein. The contamination was obvious, since almost all amino acids were found in the fraction in significant amounts above background; presumably a protein flake was washed out of the cup. The computer program will make out-of-step corrections to positions succeeding such contamination in the same way that it does after a true peak; this causes difficulties, but they can be avoided by inserting zeroes for all the analytical values at a badly contaminated position. This was done for position 56, which is equivalent to discarding the position. At position 50, alanine, shown in parentheses in the figure, occurs in the sequence box in addition to aspartic acid; this position was also slightly contaminated but not sufficiently to obscure the true residue, aspartic acid, which was confirmed in the second experiment. The alanine in addition to aspartic acid at 53 could be due to a slight undercorrection for second order $(n + 2)$ out-of-step degradations (There are other similar indications in the data that molecules which are one residue out-of-step are more likely to get further out-of-step than are in-step molecules, possibly because of their physical state or position in the reaction cup.) We failed to find a residue at position 60. At positions 64 and 66 the second (lower) amino acid in the sequence box may also be due to insufficient correction for out-of-step degradations. The partial sequence deduced from the two experiments with Rzf after considering these minor complications is given as the bottom line of Figure 2 in the one letter code. It is D I Q M T Q S P S S L S A S V G D R V T I T C Q A N Q V I S K S L N W Y Q Q R P G Q A P K L L I Y D A A B L E ? G V P ? R F T G & G &. (The key to the one letter code is in the legend of Figure 2.) This sequence is comparable to published sequences of κ_I immunoglobulin light chains (reviewed recently by Smith *et al.*, 1971) and shows no evidence of a deletion in the first 67 residues.

Partial Sequence of Light Chain Sac. A considerably more complicated sequencing experiment is illustrated in Figure 3. In this experiment (192) 39 cycles were carried out with the light chain of the myeloma-like immunoglobulin Sac; the hydrolysis was with HI–I_2. The analytical data were processed in the same way as those shown in Figure 2, but no corrections were made for R_i, R_j, etc. The value of β used in correcting the data was increased linearly from 0.001 before cycle 1 to 0.017 at cycle 18 and was held constant thereafter. The experiment was technically less good than most with an average repetitive yield, $(\alpha + \beta)$, of only 92%. A second longer sequencing experiment (194) was also carried out with the same protein for 55 cycles, and in this experiment half of each of the first 20 fractions was hydrolyzed with NaOH–dithionite; HI–I_2 was used for the other half and for all of the remaining fractions. As indicated in the Materials section, the Sac protein was originally characterized by Lewis *et al.* (1968) and the present light chain was prepared by Parr *et al.* (1971) who had already determined a provisional sequence of the first 16 residues using conventional methods. The results of all of these experiments have been taken into account in identifying alanine *vs.* serine or cysteine and in assigning amides. The assignments of serine *vs.* cysteine were based on the following reasoning. The Sac light chain has four residues of cysteine per mole (Parr *et al.*, 1971). If three of these are assigned to the constant region the most likely position for the remaining cysteine is at position 20 as judged by homology with pre-

viously published sequences, but we do not regard this assignment as being secure. Within these limitations, the sequence of 34 out of the 39 positions plotted can be clearly determined from the figure, since only one analytical value falls in the outlined sequence box at each of the 34 places. No residue was detected at position 4 in the HI experiment, but methionine was recovered in excellent yield in the NaOH hydrolysis experiment. Positions 18, 35, and 39 gave lysine in a near normal yield, together with an amino acid identified by the computer program as aspartic acid (marked "D" in the figure). Careful inspection of the chromatograms showed that the amino acid had a retention time slightly different from aspartic acid. We have since noticed it to varying extents after HI hydrolysis whenever lysine is encountered in the degradation of a protein that has been exposed to urea during preparation (urea was used at one stage of the preparation of the present sample of Sac light chain). The amino acid is probably the ϵ-carbamic acid of lysine.

Three positions in the experiment 192 illustrated in Figure 3 show features which require special comment. At position 5, threonine was recovered in approximately 5% yield and at an even lower yield in the second sequencing experiment 194; yet Parr *et al.* (1971) obtained threonine in a good yield as the amino terminal acid at position 5 during a conventional Edman degradation using dansylation and hydrolysis to identify the amino terminal residues, and also when cyanogen bromide cleavage of the methionine at 4 was followed by dansylation and hydrolysis. At position 13 alanine in less than the expected amount was recovered together with threonine, and again the proportions were not the same in the two sequencing experiments. At position 28 both alanine and isoleucine were recovered but in lower amounts than expected; in experiment 194 only isoleucine was recovered at position 28 but again in a low amount. Yet isoleucine at position 30 was recovered in the expected yield in experiment 192 (the corresponding analysis in 194 was lost). In many previous and subsequent sequenator experiments with Bence-Jones proteins, no difficulty has been encountered in obtaining the expected yields of threonine at position 5, or of alanine at position 13, nor have we found isoleucine to give different relative recoveries in two different places in one sequence. We therefore suggest that these three peculiarities are related in the case of the light chain Sac. As indicated below, this protein has a large deletion relative to the other Bence-Jones proteins that have been studied. Conceivably this deletion causes the protein to take up an unusual configuration allowing the side chains of residues 5, 13, and 28 to interact in such a way that the extraction of their thiazolinones is hindered even after the residues are no longer linked by peptide bonds. The implication is that the threonine cleaved at cycle 5 was not significantly extracted until the alanine at position 13 was cleaved, at which time both thiazolinones were then partially extracted. We also suggest that the correct residue at position 28 is isoleucine (see below for additional related evidence to this effect) and that the alanine observed in experiment 192 at position 28 is actually from position 13. The broken line joining the circled points in the figure indicates the relationships. We realize that they are far from being established, but various alternative explanations in terms of biological peculiarities of the Sac light chain seem even more unlikely. Perhaps future applications of the quantitative sequencing technique will reveal other comparable cases. The sequence deduced after considering all these factors is shown as the bottom line of Figure 3. It is D I Q M T Q S P S S L S A S V G D K S C Z Z Z B & & T I P I G G G T K V B V K. The following additional sequence

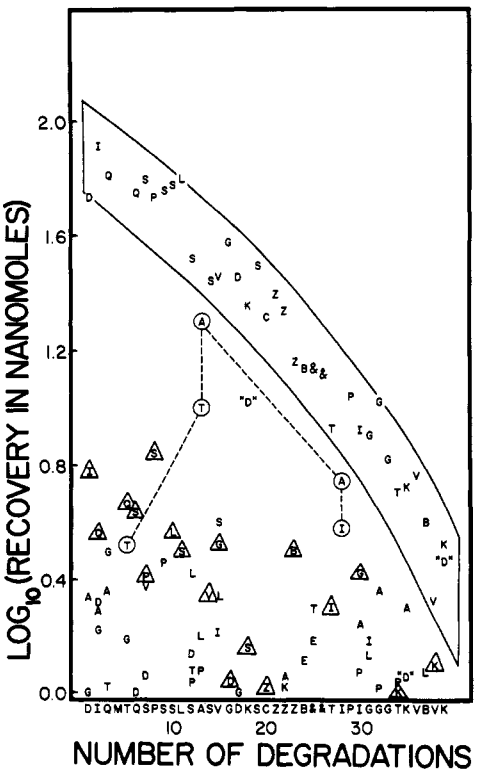

FIGURE 3: A plot equivalent to that shown in Figure 2 but with the light chain of the myeloma-like immunoglobulin Sac. The sequence deduced from this experiment with HI hydrolysis (192), from a second experiment (194) with NaOH–dithionite hydrolysis and from the data of Parr *et al.* (1971), is shown below the graph. See text for explanation of dotted line and triangles. Approximately 4 mg of dried protein ($\leq$270 nmoles) was used in the experiment.

was determined in the longer sequenator experiment: (40) R T V & & P ? ? F I F (50).

The main sequence so far considered clearly confirms the provisional data of Lewis *et al.* (1968) in that the light chain Sac contains a major deletion (of about 68 residues) in its variable region. We have discussed the biological implications of this finding already (Smithies *et al.*, 1971) and will not consider them further here. An additional interesting finding can however be deduced from the data in Figure 3. In the case of the Bence-Jones fragment Rzf considered above there were no signs of residues being recovered at the positions before their true positions. This has also been the case with all other purified proteins we have studied. But it is not so for the Sac light chain. At more than one-third of all the positions up to position 39, the main sequence amino acids are signalled by their detection at the immediately preceding degradation at a level higher than any amino acid other than the "true" residue. For example, isoleucine is recovered in the first cycle as well as aspartic acid. The relevant analytical values are marked with triangles in Figure 3. The simplest conclusion from this finding, which was unambiguously found in both sequencing experiments, is that the Sac light chain contains a mixture of two proteins. The major component of the mixture starts with a typical κ light chain sequence Asp-Ile-Gln-Met etc., while the minor component, present at about 7% of the major component, is one residue shorter and starts with the sequence Ile-Gln-Met etc. (We may note that the isoleucine tentatively assigned to the main sequence at position 28 is anticipated at position 27 in the minor component.) Two possibilities

are that the shorter chain is the result of degradation during the isolation of the protein or that it is a consequence of an *in vivo* process. Some evidence favors the first interpretation, since M. E. Percy (personal communication) found amino terminal aspartic acid essentially alone in another sample of freshly isolated Sac light chain, but isoleucine was detected in addition to aspartic acid after extended dialysis of the preparation against an acidic buffer—a step similar to one used in preparing the present sample. The large internal deletion of about 68 residues in the Sac light chain together with an unusual configuration might make it more sensitive to degradation than a normal light chain. The possibility that another unusual feature of the Sac immunoglobulin (the deletion of about 102 amino acids from the beginning of its heavy chain) might be the consequence of degradation of the heavy chain facilitated by an unusual light chain structure has been mentioned by Parr *et al.* (1971).

In conclusion, the present examples indicate that the methods we have described are adequate to measure with reasonable precision the quantitative recovery of all amino acids normally encountered in sequencing proteins. Log plots of *all* of the anaytical data of the type illustrated in Figures 2 and 3 are very helpful for understanding the outcome of any particular experiment, especially if the protein under study is in any way unusual.

Appendix

In processing the large amount of quantitative data from the amino acid analysis of sequenator samples, a combination of three computer programs was used: the "amino acid" program for the identification of amino acids and quantitation of their yields, a "graph" program used at several stages to plot the data, and a "statistics" program to determine the necessary parameters so that the graph program can correct for rising background, out-of-step degradations, and recovery factors.

Amino Acid Program. For each sequenator sample the amino acid analyzer generates a graphically recorded chromatogram, a printed teletype output, and a punched paper tape of the times of the chromatogram peaks together with the integrated areas under the peaks. The chromatograms are first inspected for any faulty integrations, due, for example, to incorrect base line selections, and appropriate corrections are made. If there are many corrections an edited punched paper tape is prepared; a small number can be corrected directly by supplying the program with amended values for specified integrals.

The amino acid program is supplied before operation with a list of amino acids, their respective color factors, and the nominal intervals, NI_{ij}, NI_{jk}, etc., between neighboring amino acids i and j, and j and k, etc. The nominal intervals are chosen as the longest intervals observed in a small sampling of normal chromatograms. The method of identifying amino acids depends on the relative constancy of these intervals rather than on the absolute elution time of an amino acid since this varies somewhat in different analyses.

The program is also supplied with a reference time and a reference area so that it can identify a selected reference peak in each chromatogram. When the program starts analyzing the data it first locates this reference peak, usually norleucine or NH_3, by choosing the first peak before the specified reference time with an area greater than the specified reference area. Having located the reference peak, "i," the program then computes two intervals in which the next amino acid, "j," may be expected to occur

$$\text{search interval 1} = NI_{ij} + 0.2(NI_{ij})$$

$$\text{search interval 2} = NI_{ij} + 0.4(NI_{jk})$$

The smaller of these intervals is searched for the occurrence of any peaks and the largest peak in the interval is then identified as "j"; optionally the program can be set to identify as "j" the first peak above a minimal cutoff area, or the largest peak in the interval if none are above the cutoff. If search interval 1 was used and no amino acid was found a second search is then made using search interval 2. (Note that search interval 2 extends the nominal interval ij by a proportion of the following interval jk while search interval 1 uses a proportion of the interval ij itself.) If search interval 2 was used originally a second search is not made. This procedure is necessary so that correct identifications are made when a large interval is followed by a small one, or *vice versa*.

Once "j" is found the search for "k" begins but this time with the peak time of "j" being the zero time for the next interval jk. If "j" does not occur in the analysis, the next nominal interval used is still NI_{jk} but the reference time zero is the time of "i" plus NI_{ij}. In this fashion all amino acids can be successfully identified. Their yields are then determined by dividing the respective areas by the appropriate color factors and, optionally, by the yield of norleucine which is added to the sequenator samples as a semi-internal standard to correct for mechanical losses.

The yields, as nanomoles of amino acids, are stored on disk and are printed on a line printer.

Graph Program. The graph program is used to create graphs on the line printer or teletype without the need for a plotter. To do this the program reads the yields from disk and places the one letter code for each amino acid in a two dimensional array. The array row is determined by the amino acid yield in nanomoles or in per cent yield of amount of protein applied. The array columns correspond to positions in the sequence; by using string handling statements four positions can be contained in each column. The array is printed one array row at a time. To avoid excessive overlap of the data for different amino acids, suitable base line displacements are added to the ordinates for each amino acid. The graph produced is equivalent to that shown in Figure 1A for Leu only.

Statistics Program. After inspecting the first graph (the "plain plot") the operator selects and specifies several points to be used as typical background values for each amino acid. The statistics program then derives the empirical parameter $(\alpha + \beta)$ from the following equation for the background which was described in the earlier section: [background of ith type of amino acid after n cycles] $= R_i(N_i/N)h[N(\alpha + \beta)^n + (N-1)(\alpha+\beta)^{n-1} + \ldots (N-n+1)(\alpha+\beta)]$.

The selected method of solving this equation is to compare the observed ratio of the background values at two points with the ratio calculated using a suitable starting value for $(\alpha + \beta)$. The value of $(\alpha + \beta)$ is then altered programatically until the agreement of the observed and calculated ratios is as close as needed. A mean of all the values determined for $(\alpha + \beta)$ for all the background points for all the given amino acids is then used as a single parameter for that sequenator experiment, after excluding from the final mean any points deviating by more than two standard deviations from a provisional mean.

Using the final mean value of $(\alpha + \beta)$ the program then returns to the background equation and solves for the "slope,"

$R_i(N_i/N)h$, for each selected background point for each amino acid. The means of the values of the "slope" for each amino acid and the final mean value of $(\alpha + \beta)$ describe the backgrounds. This information is used by the graph program to correct the yields for background and on demand to replot the data. (A graph is not usually plotted at this stage.)

To solve for out-of-step errors the statistics program requires the operator to specify positions where peaks occur that are suitable for use in calculating β. Peaks of a repeated residue may not be used; for example, Ala in the sequence Ala-Ala is not suitable but Ala in the sequence Ala-Leu is suitable. The statistics program then calculates β by solving the equation described above: $n\beta$ = (yield above background of nth at $n + 1$)/(yield above background of nth at n). The values are printed graphically and numerically on the teletype.

Graph Program Rerun. The operator inspects the values of β and determines whether β varies in any systematic manner. The graph program is arranged to accept suitable simple equations for specifying the variation of β with n. These equations are used in correcting the graphs for out-of-stepness. A third graph, now corrected for background and out-of-step degradations, can then be obtained, although again this graph is not in practice often plotted.

As discussed previously, the graph program is usually set to employ the "add-back" feature which causes either the observed or the calculated value for out-of-step recovery, whichever is smaller, to be added back to each amino acid peak. The calculated value is then subtracted from the out-of-step positions, a device which generates a negative value if the calculated correction exceeds the observed out-of-step recovery and serves to emphasize any over-corrections. Since the out-of-step "tails" at the end of a sequencing experiment are truncated by the absence of further analytical data, the values added back for the last few positions are the calculated ones rather than the (un)observed ones. The out-of-step correction program cannot distinguish true amino acid peaks from false peaks due to any contaminants, but it can be set not to correct small base line fluctuations (*e.g.*, less than 1 nmole). The corrected graph with add-back is then plotted (as in Figure 1B) and in its logarithmic form (as shown in Figure 1C). From this type of log plot the recovery factors, R_i, R_j, etc., can be evaluated (or they can be supplied beforehand) so permitting the final fully corrected log plot of the type shown in Figure 1D, 2, and 3 to be obtained. Once R_i, R_j, etc., are known, h is calculated from each of the slopes, $R_i(N_i/N)h$, $R_j(N_j/N)h$, etc., and a mean value for h can then be obtained.[2]

Acknowledgment

We thank Mr. Robert Kobetsky, Illitron Division, Illinois Tool Works, Chicago, Ill , and Mr. Joe Swoboda, Laboratory of Genetics, University of Wisconsin, for their invaluable engineering assistance in the course of setting-up the sequenator, and Dave J. Lapin and Dennis White, University of Wisconsin Computing Center, for helping us with our programs.

References

Baglioni, C., Cioli, D., Gorini, G., Ruffilli, A. and Alescio-Zonta, L. (1967), *Cold Spring Harbor Symp. Quant. Biol. 17*, 147.

Deutsch, H. F. (1965), *Immunochemistry 2*, 207.

Edman, P. (1970), *in* Protein Sequence Determination, Needleman, S. B., Ed., New York, Heidelberg, Berlin, Springer-Verlag, p 211.

Edman, P., and Begg, G. (1967), *Eur. J. Biochem. 1*, 80.

Hermodson, M. S., Ericsson, L. H., and Walsh, K. A. (1970), *Fed. Proc., Fed. Amer. Soc. Exp. Biol. 29*, 728.

Lewis, A. F., Bergsagel, D. E., Bruce-Robertson, A., Schachter, R. K., and Connell, G. E. (1968), *Blood 32*, 189.

Niall, H. D., and Edman, P. (1967), *Nature (London) 216*, 262.

Niall, H. D., Sauer, R., and Allen, D. W. (1970), *Proc. Nat. Acad. Sci. U. S. 67*, 1804.

Parr, D. M., Percy, M. E., and Connell, G. E. (1971), *Immunochemistry* (in press).

Smith, G. P., Hood, L., and Fitch, W. M. (1971), *Annu. Rev. Biochem. 40*, 969.

Smithies, O., Gibson, D. M., Fanning, E. M., Percy, M. E., Parr, D. M., and Connell, G. E. (1971), *Science 172*, 574.

[2] Copies of all the programs outlined here will appear following these pages in the microfilm edition of this volume of the journal. Single copies may be obtained from the Reprint Department, ACS Publications, 1155 Sixteenth St., N.W., Washington, D. C. 20036, by referring to author, title of article, volume, and page number. Remit check or money order for $5.00 for photocopy or $2.00 for microfiche.