Biochemistry

Volume 15, Number 3 February 10, 1976

Documentation of Results in the Determination of the Covalent Structure of Proteins

In 1974 a Committee was named by the Editors of *Biochemistry* and of *The Journal of Biological Chemistry* to conduct an inquiry into the possibility of developing a set of recommendations on criteria to be met in the publication of the results of research concerned with the determination of amino acid sequences in proteins. This Committee consisted of Ralph A. Bradshaw, R. David Cole (Co-chairman), C. H. W. Hirs (Co-chairman), Stanford Moore, and Hugh Niall.

This Committee prepared a statement summarizing the views of its members which was sent with a request for comment to some 80 laboratories around the world, selecting those laboratories in which sequence analysis was known to be undertaken. The responses from this survey were used to prepare a summary report. Following extended evaluation of the summary report, a series of recommendations were formulated with the concurrence of the Committee. These recommendations follow.

Recommendations to Authors

A. General Statement

The Committee found there is general agreement on a number of central issues: that sequences advanced with little or no supporting data are unacceptable; that those data minimally required to establish an amino acid sequence should be made available to the public in as complete detail as possible, although not necessarily in their entirety in the pages of a journal (appropriate repositories of scientific information and/or reduced format (miniprint) appendices may be used for much of the data); and that there should be guidelines as to how research in this field should be documented, but that there should be flexibility in these guidelines. Ultimately, the decision to accept or decline a manuscript will require the discretion of the referees and the editors in their interpretation of the following guidelines.

B. Organization of Manuscripts

The importance of careful forethought in how to organize a manuscript containing a large body of information, as is so often the situation in this field, cannot be overemphasized.

Although the referee system in present use is intended to ensure that inadequately documented sequences do not get published as having been fully established, i.e. without commas and parentheses in the abbreviation system advocated in the IUB/IUPAC Rules, the fact is that the recently conducted survey by the committee brought to light instances when sequences have been published as fully established when the evidence was actually incomplete. Although it is impossible to reconstruct after the event how such lapses can arise, it is clear that poor organization of data in a manuscript is the principal culprit, the chances being that poorly organized data misled not only the reviewers, who are usually inclined to give authors the benefit of the doubt, but the authors of the paper as well.

In general it is recommended that sequence data not be published piecemeal and that subdivision of a report on a major accomplishment in this field be kept to a minimum. Partial sequence data should not be published unless the partial sequence serves to elucidate some notable property of a protein, such as substrate binding in an enzyme, attachment of a prosthetic group, site of limited proteolysis, etc. Subdivision of reports on the determination of the primary structure of a protein should be allowed only for good scientific reasons and provided subdivision does not materially increase the overall length of the text in the journal.

The design of figures and of tabular material is of particular importance in making the text readily comprehensible not just to the specialist, but to the average reader who may wish to study the evidence in some depth. It is recommended that every manuscript describing a sequence contain a carefully designed summary figure which shows how the sequence was derived from the interrelation of data obtained with the various peptide fragments. Such figures should show (preferably with suitable symbols such as arrows) how residues were placed in each of the fragments studied. To make such figures more readily understood in relation to the text, it is recommended further that wherever possible a simple system of designation be employed to identify peptides and their subfragments.

Inasmuch as sequence analysis is not as yet a routine undertaking it is rare that a sequence is elucidated which did not present unusual or unanticipated difficulties, or in which the strength of the evidence is uniformly high throughout. It is recommended that manuscripts contain a section in which such questions are addressed.

C. Isolation of Peptides

Information presented on this topic is likely to be of more value to subsequent workers than any other aspect of the experimental detail in the manuscript. Therefore, it is strongly recommended that great care be taken in presenting descriptions as to how peptides and their subfragments were obtained. In particular, careful documentation is es-

sential for those peptides which proved to be difficult to isolate in good yield. While discretion should be used in selecting the number of figures to be included, authors are reminded that no limits on length are imposed on material which will be placed in data repositories and reduced format appendices. Information on the composition of peptides is essential and should be presented in appropriately designed tables which give compositions in terms of molar ratios of constituent residues at least to the nearest tenth of a residue. Such tables should also give the yield in which the peptides were isolated and, when appropriate, information on electrophoretic properties, staining with ninhydrin and other reagents, end groups, etc. It is particularly important that documentation be complete for those peptides minimally required for the establishment of a sequence. Less complete documentation is acceptable for those peptides which by virtue of composition or sequence are used to confirm structures derived from principal fragments.

D. Sequence Analysis

When an automated sequenator has been employed the following information should be provided: (*a*) the type of program used in a particular run; (*b*) the quantity of material subjected to degradation; (*c*) the number of times the experiment was repeated, and how many degradation cycles were made in each run; (*d*) the analytical procedures used at each step to identify and measure the quantity of amino acid derivative removed; and (*e*) the repetitive yield observed from quantitative measurement of those amino acid derivatives which can be determined accurately. It is recognized that quantitative data are not always obtained in assessing the results of a sequenator experiment and that some workers rely entirely on qualitative procedures. The intent of the present recommendations, however, should be obvious: it is incumbent on authors to provide information which will permit reviewers and ultimately the readers of the journal to form a clear picture as to how well the sequenator performed and at what point problems connected with overlap and low yield may objectively be deduced to have constituted a limitation on the conclusions drawn. The simplest way to present such information is to show for each step (or at representative points through the run) the yield in which the principal and subsidiary amino acid derivatives were obtained. If quantitative data are not available, as for example when most types of thin layer chromatography are used for identification, suitable indications should be provided as to the relative intensities observed for the derivatives at each step on the thin layer chromatogram.

In general, when a sequenator is used for the determinations of relatively long stretches of sequence, *i.e.* stretches of over 15 residues, the experiment reported should have been performed at least twice or confirmed by an independent method. Identification of residues corresponding to steps at which no derivative could be identified (i.e. "gaps" in the analytical record) should not be attempted unless corroborative evidence can be provided either from other peptide fragments or from the composition of the peptide being analyzed. For example, if a peptide with composition Glx_2, Pro, Gly, Ala, Leu, Tyr was sequenced and gave the following identifications:

Pro-Leu-Glu-Gly-Ala-X-Tyr

it would be appropriate to conclude X is Glx and that the final residue is in fact Tyr.

Especially clear-cut quantitative data are essential if mixtures of peptides are subjected to sequence analysis; the assumptions made in interpreting the data should be clearly stated.

The recommendations presented in connection with documentation of results obtained by application of sequenators are in general equally applicable to results obtained by manual methods. However, inasmuch as the results obtained by manual methods often involve different analytical procedures for assessing the progress of the degradation, the following additional recommendations are offered. In those experiments in which subtractive analysis has been employed it is essential to provide the composition of the peptide residue after each step of degradation. In presenting such results much space can be saved if suitably scaled tables are used in which each composition is presented in a single line of text and the lines show the residue composition to conform with the sequence deduced. When subtractive analysis is conducted with the aid of end group determinations, particularly by the dansyl method, the results should be presented to show both the principal and subsidiary derivatives identified at each step.

In dealing with the results obtained by analysis of reaction products formed on treatment of peptides with exopeptidases, authors should provide the following information: (*a*) the quantity and concentration of peptide taken for analysis; (*b*) the ratio of the concentration of substrate to enzyme; (*c*) the yield of each amino acid (as moles of amino acid released per mole of peptide) observed at each time; and (*d*) the presence, if any, of peptide intermediates that are detected by the analytical procedure used.

Reconstruction of large sequences by interrelation of the results obtained by sequence analysis of derived peptides should proceed on the basis of overlaps of at least 2 residues. A single residue overlap can be used only when the residue in question is (*a*) the only residue of that type involved in the alignment and (*b*) substantive evidence is provided that an additional fragment, beginning and ending with that residue, has not been lost. In such circumstances, a partially structured peptide, in which the NH_2- or COOH-terminal residue has been established to firmly position the peptide, may be used as proof.

E. Use of Reduced Format Appendices and Data Repository

Journals accepting these recommendations may make available directions for the preparation of sheets to be reproduced photographically in reduced format appendices or deposited in an information or data repository. Authors are reminded that these sheets cannot be corrected by the editors or the redactory and that, therefore, special care should be taken in typing and proofreading them. It is particularly important to make certain that abbreviations conform with those recommended by the IUB/IUPAC Commission on Biochemical Nomenclature.