AN ABSTRACT OF THE THESIS OF

George David Rose for the degree of Doctor of Philosophy
in Biochemistry and Biophysics presented on 3, Dec. 1975.

Title: Calculations Relating to the Structure of Biopolymers

Abstract approved: Redacted for Privacy
Kensal E. Van Holde

A testable, biphasic model for protein folding is formulated. In this model, linearly short and medium range interactions dominate early folding, causing the chain to assume independently nucleated modules of persisting structure termed LINCs. In a later stage of folding, the LINCs fold relative to each other, and it is only at this time that the protein assumes its characteristic interior and exterior and its overall globular structure.

In the perspective of the model, a computational approach is outlined, requiring first a systematic examination of steric and energetic constraints that can be calculated with some confidence by accepted means. To this end, calculations were conducted to determine the sterically allowed conformation for:

1) a post-helical residue situated at the carboxy-terminal end of a backbone-only helix,

2) various side-chains of an intra-helical residue, and

3) the constraints imposed on lysyl and arginyl side-chains if some accounting is made for hydration of the respective cationic side-chain moieties.

It is found that substantial steric constraints are engendered in all three cases.

In a second part of this thesis, the secondary structure of nucleic acids is examined. The secondary structure of ribonucleic acids and the genes from which they are transcribed is likely to be a parameter in any recognition and control processes involving these molecules. It is theoretically possible to enumerate the set of all messages, M, consistent with the observed amino acid sequence of a given protein. In practice, this set is computationally too large, being on the order of Avogadro's number for even a small protein. A method is developed to select two distinguished members of M without explicit enumeration. These members are:

$\overline{m}$ - the potential message with maximal secondary structure,

and $\underline{m}$ - the potential message with minimal secondary structure.

The distinguished members, $\overline{m}$ and $\underline{m}$, are extrema that bracket M. They are used to examine the properties of M relative to the degree of secondary structure forced upon the actual biological message and upon the structural gene from which it is transcribed. Although this study leads to some general conclusions about nucleic acid structure, the range between $\overline{m}$ and $\underline{m}$ is too large to permit specific predictions except in a few singular cases where further information is already available. With the exception of these cases, it appears likely that the quest for structural determination will be confined to thelaboratory until a larger library of nucleic acid sequence data can be accumulated.

Calculations Relating to the Structure of Biopolymers

by

George David Rose

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Commencement June 1976

APPROVED:

Redacted for Privacy

Professor of Biochemistry and Biophysics in charge of major

Redacted for Privacy

Head of Department of Biochemistry and Biophysics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented: 31 December 1975

Typed for G. Rose by G. Rose

Acknowledgement

Without the skilled midwifery of Ken Van Holde, this work would have been stillborn, while without the long-standing and patient support of Larry Hunter and the Oregon State University Computer Center, it would have starved to death at an early age. Nurtured on discussions with Robert R. Becker, as well as Ralph Quatrano, Ronald H. Winters, Ted Hopkins, and Rjay Murray, these ideas reached their maturity during exchanges with Don Wetlaufer.

Any work of mine will always bear the imprint of my teacher, Harry Goheen, who has long fostered the application of automata theory as a paradigm of cognition.

This thesis is dedicated to Harry and Molly Goheen.

TABLE OF CONTENTS

LIST OF TABLES AND ILLUSTRATIONS

CALCULATIONS RELATING TO THE STRUCTURE OF BIOPOLYMERS

1. Introduction

This thesis divides naturally into two distinct parts, each relating to the structure of informational biopolymers. The first part is concerned with the conformation problem for globular proteins, that is, with the relationship between a protein's amino acid sequence and its three dimensional configuration. The second part is concerned with the secondary structure characteristics of messenger ribonucleic acids and the structural genes from which they are transcribed.

In both parts, proteins and nucleic acids, a cannonical set of molecular structures is already inherent in the definition of the problem, and, in each case, the set in question is too large to be computationally useful. For example, it would be helpful to compute the free energy of all reasonable configurations of a small protein and display the result in some representation of protein conformation space. The astronomical number of members in the cannonical set precludes such an approach by exhaustion, not only for the example, but for any interesting computation one cares to make.

In both cases, proteins and nucleic acids, it is presumed that the large cannonical sets each have a single member (or perhaps a small equivalence class of members) that is the biologically active representative of that set. In the case of proteins, this member would be the native form of the protein molecule; in the case of messenger-RNA, it would be the actual biological message. In any case, it is this distinguished member that is being sought.

In lieu of a method that permits exhaustive inspection, alternative approaches must be invented. For the conformation problem, the alternative depends upon a new, testable model for protein folding which is presented at the beginning of the next chapter. Using the model, the conformation problem can be partitioned into the summation of several separable, smaller problems, each of which is computationally feasible. In effect, the model provides additional information that can be used to eliminate uninteresting subsets of the cannonical set.

No model springs readily to mind for the messenger-RNA problem, and a different approach was taken, requiring the selection of boundary elements from the cannonical set. These elements bracket the remaining members of the set, and they can be discovered without explicit enumeration. The selection and application of boundary elements, called characteristic extrema, is discussed in chapter three.

During the course of this work, many computer programs were defined and written. Some of these, such as the computer graphics routines, were of enough general interest to be included in the O.S.U. Computer Center library and will not be further discussed here. There remained, however, a large number of programs whose interest is particular to this work. This latter category can, in turn, be further divided into three libraries, known as:

a) PROTEINS - a library of programs used to manipulate protein sequence data and coordinates, and to compute and display molecular energies.

b) RNA - a library of programs used to manipulate nucleic acid sequence data, and to compute and display selected configurations satisfying the Tinoco stability criteria.

c) PLIB - a subroutine library containing support routines used in conjunction with other programs.

A synopsis of the major programs in all three libraries is given in the appendix.

II. Tertiary Structure of Globular Proteins

The transition of a denatured protein into its native structure is defined to be a global folding process, whereas any linearly piecewise folding that occurs in a nascent chain is a local folding process. Convincing instances of local folding have been demonstrated in various contexts (1,2). In general, the folded end product is expected to be process dependent because conformational states adopted by partial chains will be deprived of any information that accrues with additional chain growth. That is, a nascent chain cannot forsee its future.

Cases are known, however, in which both local and global folding processes yield the same final structure (3,4). One conception of how qualitatively differing initial states converge to the same final structure rests on the assumption that this structure is necessarily synonymous with a global free energy minimum for the molecule (3). Another conception will also rationalize the directed emergence of a unique conformation from differing initial states. In particular, a biphasic model for folding is proposed here. In this model, linearly short and medium range interactions dominate early folding from any state, and order the polypeptide chain into independently nucleated, persistent modular units of structure. Following this early assembly, linearly long-range interactions are then responsible for the further ordering of modular entities into the full three-dimensional configuration of the protein.

The general notion of a biphasic model is no longer novel inasmuch as elements thereof are to be found, either explicitly or implicitly, in several recent publications (5,6), and the concept of nucleation events proposed by Levinthal is, of course, well known (7). The attempt here, however, has been to provide a highly specific model that both takes into account the body of experimental evidence and includes sufficient detail to allow a quantitative examination of its consequences.

In detail, it is proposed that the polypeptide chain, dominated by linearly short and medium range interactions, folds initially into <u>L</u>ocal <u>I</u>ndependently <u>N</u>ucleated <u>C</u>ontinuous segments (LINCs). The ordering of the chain into LINCs is promoted during any local folding that takes place in a nascent chain, and LINC formation is also favored in a global folding process because the chain will fold into LINCs before it can fold into anything else.

LINCs are structurally persistent, separable, modular entities that are precursors to their counterparts in a folded protein. LINCs are usually, if not invariably, bounded by peptide chain turns (8,9) which are construed to be the conformationally permissive (10) hinges that allow an ensemble of LINCs to fold relative to each other.

In this model, a protein is comprised entirely of LINCs and interspersed hinges. Not until the occurrence of inter-LINCs folding does the protein take on its characteristic interior and exterior or its overall globular structure. It is at this latter stage in the folding pathway that linearly long-range forces come into play and the LINCs are disposed into their native conformation.

The LINCs and hinges model is consistent with the observation that both local and global processes can yield the same final configuration. The model is also consistent with the success of recent empirical efforts (11) to predict secondary structure based only upon correlations between local amino acid sequences. In the present model, alpha helices and anti-parallel beta pleated sheet are considered as particular instances of LINCs.

Viewed from a perspective prompted by this model, the problem of structure formation can be divided into two parts: prediction of LINCs conformation and prediction of inter-LINCs conformation. Some of the factors limiting inter-LINCs folding in the case of myoglobin suggest that packing constraints and hydrophobic interactions place major restrictions on any possible solution set (12,13).

Turning now to the question of LINC's conformation, a study by Gelin and Karplus (14) finds side-chain torsional angles in pancreatic trypsin inhibitor at or near their expected minima in the free amino acid. Such a result is consistent with the present model, for, within a LINC, short and medium range interactions direct the folding process for side-chains as well. Thus, when an independently nucleated oligopeptide 'jiggles' into a persisting conformational minimum, the side-chains are expected to populate their respective minima too, because the steric constraints at this stage in the folding process are not comparable to those imposed on a side-chain at the interior of a protein. It might be thought that when the LINCs subsequently fold relative to each other, displacement of the side-chain from a rotational minimum may find compensation in better inter-LINCs packing. In practice, this trade-off becomes less feasible because a side-chain displacement is no longer free to occur independently, but only in cooperation with other structural determinants in the LINC.

In the general case, the problem of predicting the conformation of only a single LINC by complete energy minimization (15) is still too complex to solve directly. In a recent attempt to reduce the computational complexity, each amino acid residue in the protein is represented by just two points (16). While this approximation is presented as being highly successful, it is difficult to believe that the information loss arising from a point representation of the side-chain can yet be consistent with predictive results.

The approach adopted here is to compile a catalog of constraints limiting the conformational freedom of a LINC. The catalog can then be used to winnow conformation space to a limited set of energetically favorable conformations for a given LINC. In this manner, the computational complexity will be suitably reduced without concomitant loss of information.

The remainder of this paper describes computations that reflect the stringent limitations inherent in LINCs packing, based primarily on steric restrictions.

2.1 Limitations affecting a post-helical residue

Upon termination of a right-handed alpha helix at its C-terminus, the first residue no longer in a helical orientation will be termed a post-helical residue. The subspace of conformation space that can be occupied by selected post-helical residues is now explored.

Figure 1 is a Ramachandran (phi,psi) plot with peptide coordinates taken from Marsh and Donohue (17). This (360 x 360) space was sampled every ten degrees and each 'x' marks a sample point where the dipeptide gly-ala is found to be sterically allowed. The contact distance criteria used to compute steric inhibition were taken from Ramachandran and Sasisekharan (18). Superimposed upon the 'hard-sphere' contact map in Figure 1 are energy contours of a 'soft-sphere' function (19). The good agreement between hard sphere and soft sphere functions is no longer surprising to us, as repulsive forces are known to play a dominant role in such functions. To facilitate discussion, dipeptide space is partitioned and named as shown in Figure 1.

Inspection of Figure 1 shows a narrow energy well in the map area corresponding to right-handed alpha helix. For helical residues populating this region of the map, narrowing of the well ought to be further enhanced by hydrogen bonding within the helix.

Allowed positions for the dipeptide gly-ala. Positions found to be sterically allowed are indicated by an X. Some favorable energy contours are outlined, and the regions are named.

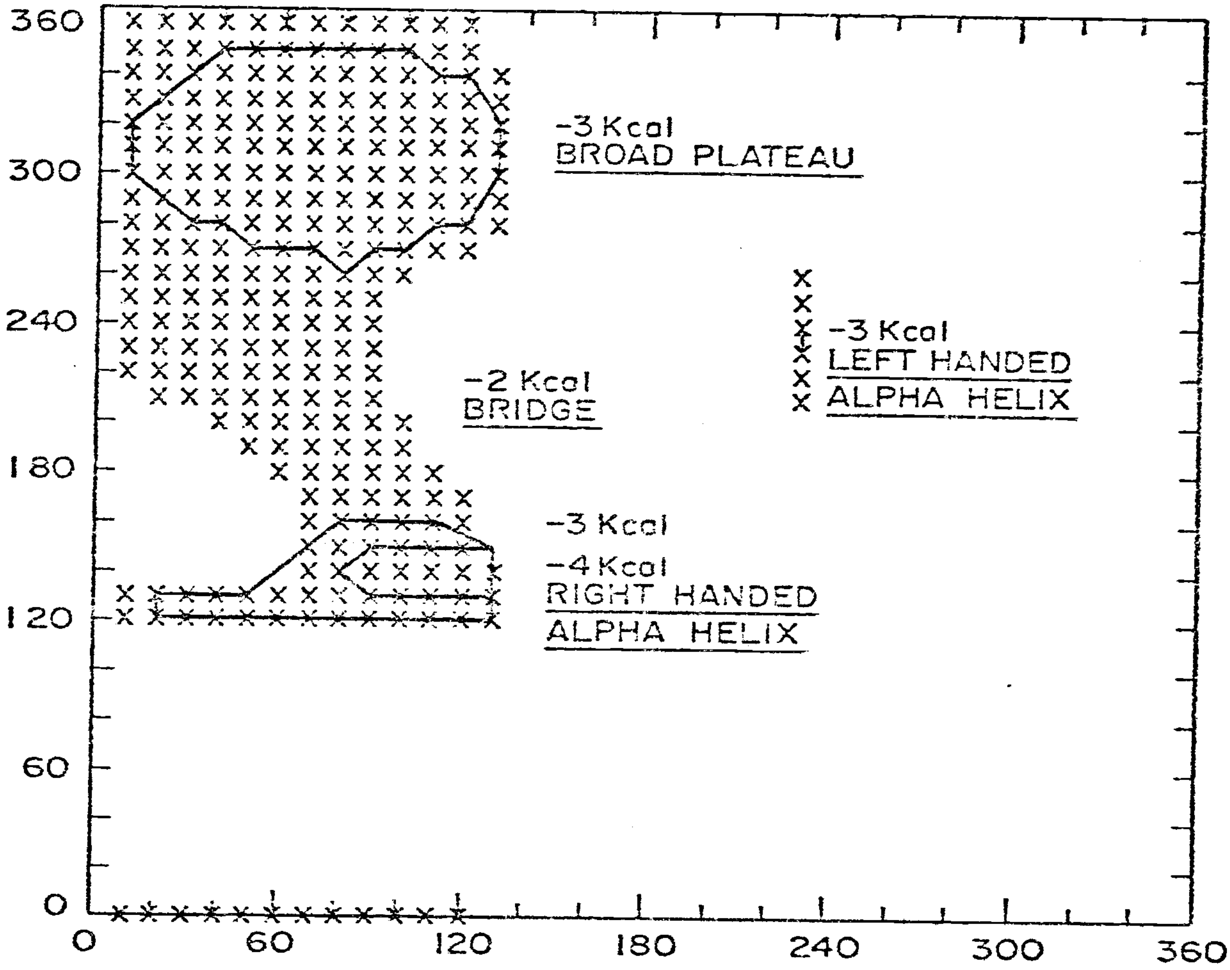


Figure 1

This expectation appears to be borne out for the refined x-ray structure of lysozyme (15,20) by the apparent clustering of (ϕ,ψ) values in the neighborhood of ϕ =120, ψ=130. This is the only high density cluster of points in the (ϕ, ψ) plot of lysozyme.

I first examine steric constraints resulting from backbone-only interactions between a post-helical residue at the carboxyl end of a right-handed alpha-helix and the four preceding residues; all five residues are backbone-only residues. A <u>backbone-only</u> residue is one without a side-chain; it can be viewed as a des-methyl L-alanyl residue. Steric constraints imposed on a backbone-only residue are the minimal constraints for any actual residue, regardless of the nature of the side-chain.

With one turn of backbone-only helix preceding a backbone-only post-helical residue,only the conformations shown in Figure 2(a) are allowed. This restriction of conformation space is due to steric interference between the backbone atoms in the post-helical residue and the adjacent carbonyl oxygen from the preceding turn of the helix. Since the restriction involves only backbone atoms, every post-helical residue is at least this restricted.

When the side-chain in a post-helical residue is also taken into consideration, further structural limitations are seen. While a post-helical backbone-only residue is not distinguishable in this analysis from a post-helical alanine, differences do begin to appear with further increases in side-chain size. Corresponding diagrams for the cases of histidine and tryptophan are shown in Figure 2(b)

Sterically allowed positions for the first post-helical residue adjoining the C-terminus of a backbone-only α-helix.

(a) Allowed positions for a backbone-only residue. Backbone-only residues are allowed only in the area shaded by diagonal lines.

(b) Allowed positions for Trp and His.

(c) Allowed positions for Trp only.

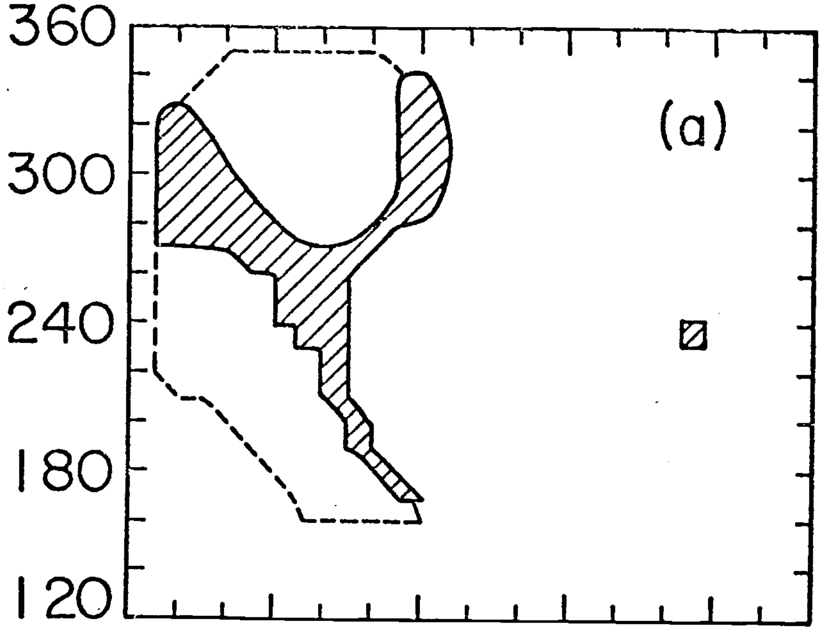
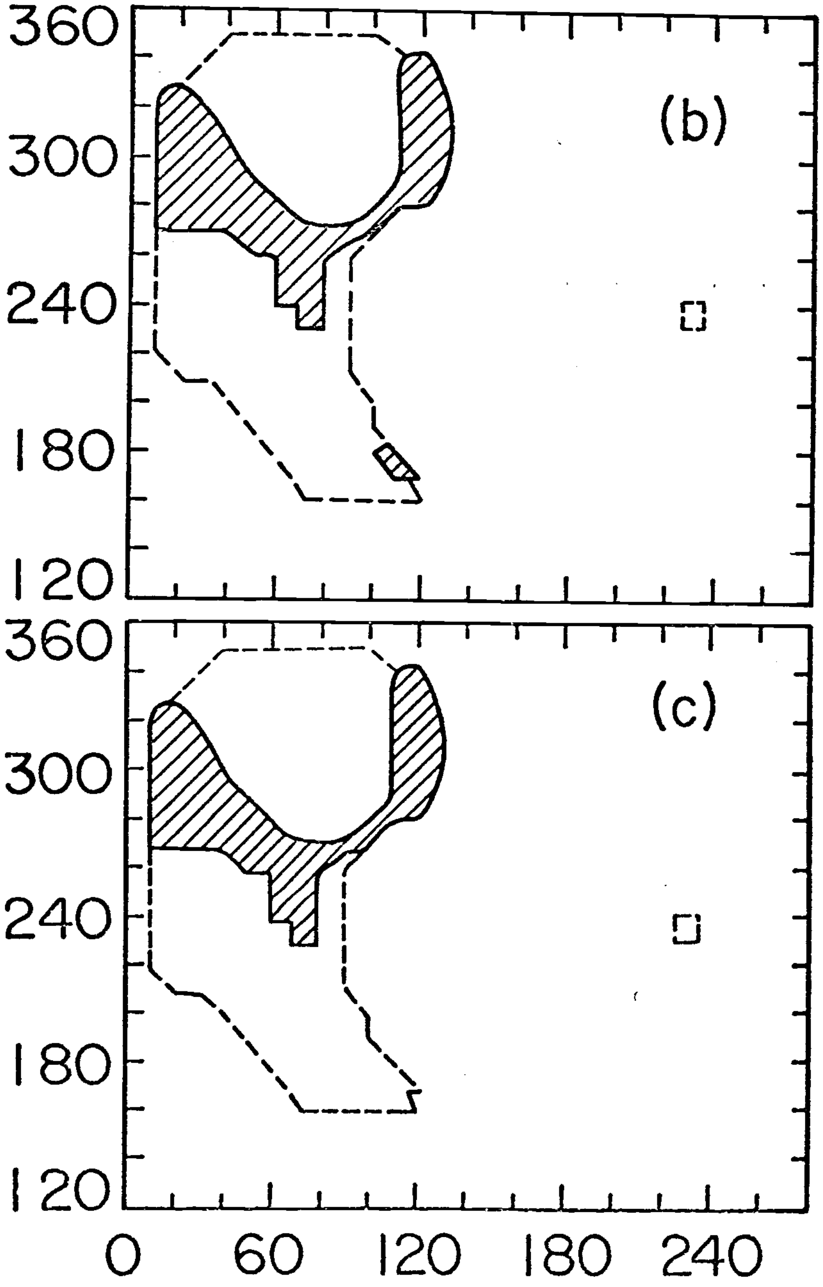


Figure 2

Figure 2 (Continued)



and Figure 2(c). In this computation, side-chain configurations arising from the domain $\chi^1 = 60^\circ, 180^\circ, 300^\circ$ ($\pm10^\circ$) and $\chi^2 = 0^\circ, 90^\circ, 180^\circ, 270^\circ$, were examined. It can be seen from the figure that the side-chains can impose significant additional constraints on the possible disposition of a post-helical residue.

The structural limitations shown for post-helical residues are based on the assumption of energetically well-formed helix (21). When the helix used for these computations is appropriately distorted at a constraining locus, there is an accompanying relaxation of the observed constraints.

In addition, deviation from the ideal peptide geometry used here may tend to reduce the limitations shown in Figure 2. However, an attempt has been made to compensate for this possibility by a conservative choice of contact distance criteria. Studies on steric hindrance show a sensitive dependence upon the choice of contact distance criteria (17), with the Ramachandran values being the most conservative set proposed.

2.2 Limitations affecting an intra-helical residue

A second example of stringent packing constraints is seen in the case of an intra-helical residue. The helix-breaking tendency of proline due to steric effects was observed some time ago (18,22). In this second example, attention is focused on the converse steric effect, limitation of side-chain freedom by the helical backbone.

Each of the amino acids listed in Table 1 was included as the middle residue between two turns of backbone helix (i.e. $(gly)_4$-X-$(gly)_4$ where X is the residue under inspection). The side-chains were then examined at configurations where side-chain groups are in one of the conventionally observed torsional minima. Aliphatic groups were varied over the domain 60^o, 180^o, and 300^o ($\pm 10^o$), while planar and aromatic groups were varied over the domain 0^o, 90^o, 180^o, and 270^o. Table 1 summarizes the positions found to be sterically allowed. Backbone helix is seen to strongly limit the accessible side-chain structures of several amino acid residues.

In the formation of a LINC, charged polar residues are probably hydrated. The attachment of a hydration shell to the terminal group of arginine or lysine, for example, will increase the packing constraints. To approximate hydration effects, x-ray data from salts of arginine and lysine (23, 24, 25) were examined and water molecules were attached to the terminal groups at loci where hydrogen bonding was observed in the crystal structures. The water was oriented so that its hydrogen atoms were symmetrically positioned above and below the plane of the side-chain group. The hydrated amino acid residues, $Lys\cdot(H_2O)_3$ and $Arg\cdot(H_2O)_5$ were then used in the intra-helical computation. In Table 1, it can be seen that the inclusion of hydration tends to force both arginyl and lysyl side-chains towards extended chain configurations.

Table 1

$(\text{Gly})_4\text{-X-}(\text{Gly})_4$ in Helix

Domain A	position	I = $60^\circ \pm 10^\circ$
		II = $180^\circ \pm 10^\circ$
		III= $-60^\circ \pm 10^\circ$

Domain B	position	I = 0°
		II = 90°
		III= 180°
		IV = -90°

The domains given for each residue are the domains of definition over which each side-chain group was varied, listed in sequential order of increasing distance from the C-alpha along the side-chain. For example, Tyr has two degrees of rotational freedom in its side-chain arising at the C^α-C^β bond and at the C^β-C^γ bond. With two degrees of freedom, it is necessary to specify two domains of definition. These are listed in the table below as A,B where domain A pertains to the C^α-C^β bond and domain B pertains to the C^β-C^γ bond.

Residue	Domain	Allowed Positions	Hydrated Form Allowed Positions
LYS	A, A, A, A	II, II, II, I-III II, I, II, I-III III, II, II, I-III III, III, II, I-III	II, II, II, II II, I, II, II III, II, II, II III, III, II, II
CYS	A	II III	
GLU	A, A, B	II, I, II or IV II, II, II or IV I, II, II or IV I, I, II or IV	
HIS	A, B	II, II or IV	
MET	A, A, A	II, I, I or II II, II, I-III III, II, I or II III, III, II or III	
ASP	A, B	II, II or IV	

Table 1 (continued)

Residue	Domain	Allowed Positions	Hydrated Form Allowed Positions
THR	A	III	
TYR	A, B	II, II or IV	
SER	A	II III	
VAL	A	III, III	
ILU	A, A	III, II or III	
LEU	A, A	II, II III, III	
PHE	A, B	II, II or IV	
TRP	A, B	II, II or IV	
ARG	A, A, A, B	II, II, II or III, I or II or IV	II, II, II or III, I or II
		II, II, I, I or IV	
		II, I, II, I or II or IV	II, I, II, I
		II, I, I, I or IV	II, I, I, I or IV
		III, II, II, I or II	III, II, II, I
		III, II, I, I or IV	III, II, I, I or IV
		III, III, II, I or or II or IV	III, III, II, I
		III, III, III, I or II	III, III, III, I or II

Table 2

(Gly)$_4$-LYS-ARG-(Gly)$_4$ in Helix

Domains are defined as in Table 1. Any of the allowed positions listed for lysine are sterically compatible with any of the allowed positions listed for arginine. All other pairwise positional arrangements are sterically incompatible.

Allowed Positions for Hydrated Lysine	Allowed Positions for Hydrated Arginine
II, II, II, II	II, II, II, I
II, I, II, II	II, II, III, I or II
III, III, II, II	II, I, II, I
	II, I, I, I or IV
	III, III, II, I
	III, III, III, I or II

2.3 Limitations affecting adjacent intra-helical lysyl and arginyl residues

As a final experiment, sequentially adjacent lysyl and arginyl residues, both intra-helical, were inspected to see whether such a juxtaposition imposes constraints in addition to those experienced by these residues taken individually. Additional constraints were observed, as summarized in Table 2.

2.4 Summary and conclusions

The values obtained from the preceding computations were not compared to values available from x-ray studies since a correspondence between individual torsion angles will depend in part on factors not included here. These initial computations have employed an idealized moiety called backbone-only helix, and with it, the assumption of a completely regular geometry for a helix. While helical fibers of poly-L-alanine appear to be compatible with these assumptions (26), it is not expected that a heterogeneous collection of helical residues will exhibit equivalent regularity. For these reasons, it is felt an appropriate test of the model must wait until predicted LINCs can be compared to their x-ray elucidated counterparts in solved structures.

In closing this section, it should be noted that the LINCs and hinges model is the simplest representative taken from a spectrum of related models. In the preceding paragraphs, emphasis has been placed on the similarity in structural identity of a LINC from the onset of structure formation through folding to incorporation in the final globular assembly. The model is simple in a computational sense, because, with these assumptions, the approximate structure of a given LINC can be calculated without regard for its neighbors and then treated as a single structural entity during subsequent computations. It is possible, however, that when the ensemble of LINCs is packed into a final globular assembly, a more extreme deformation of the original structures occurs. In the most extreme case, the original structure would be deformed beyond recognition, but for reasons given earlier, this extreme is thought to be unlikely. In the event that limited deformation takes place during inter-LINCs assembly, the initial conformation of the undeformed LINC would serve as a suitable starting structure.

In summary, a testable biphasic model for the folding of globular proteins has been proposed. In this model, linearly short and medium range interactions dominate early folding, causing the chain to assume independently nucleated, structurally persistent modular units of structure; these postulated entities are termed LINCs. In a later stage of folding, the LINCs fold relative to each other, forming a structure in which linearly long-range interactions also play a role. It is only at this time that the protein assumes its characteristic interior and exterior and its overall globular structure.

If these ideas about the folding process are valid, then demonstrable stabilizing forces must exist in oligopeptides of even moderate size. One strong source of structural stabilization is steric repulsion, and, to this end, some packing constraints for intra-helical and post-helical residues have been shown. Additional work will be necessary to further develop the catalog of structural determinants for a LINC. At the same time, an exploration of the interfaces between LINCs and hinges will be required. In the transition from a LINC to a hinge, steric constraints can no longer take such a key role, since by these working assumptions hinges are comparatively flexible. In order to predict the locations of these interfaces, it will be necessary to have some further accounting of hydrogen-bonding and hydrophobic interactions.

III. Secondary Structure of Ribonucleic Acids

The secondary structure of ribonucleic acids and the genes from which they are transcribed is likely to be a parameter in any recognition and control processes involving these molecules. For example, the half-life of mRNA varies widely in eukaryots, ranging from a few seconds to many hours (27). The notion that the secondary structure of RNA correlates with such half-life differences is consistent with the observation that rRNA and tRNA have both a high degree of secondary structure and a comparatively long half-life (28). The explanation for increased longevity as a function of increased secondary structure is probably related to the action of ribonuclease which will preferentially degrade single stranded RNA over double stranded RNA (29).

Evolutionary considerations are also of concern here. If there exists a structure:function correlation in ribonucleic acids, then particular configurations should enable or enhance biological function, and we may therefore expect selection pressures to play a role in the evolution of these molecules. Since there is no a priori reason to believe a mutation that confers benefit on the message will also benefit the protein, it is necessary to ask whether conflicts do occur, and, if so, how they are resolved.

Recent attention has been focused upon configurations of 'twofold rotational symmetry' or palindromes in the base sequences of genes. Such apparently diverse mechanisms as the Lac control region (30), the recognition sites for restriction enzyme action (31), and the DNA renaturation experiments of Wilson and Thomas (32) all implicate palindromic sequences in a seemingly central way. Palindromes fall naturally within the scope of interest of this study since an RNA sequence transcribed from a palindrome will energetically favor a hairpin secondary structure. Indeed, it seems likely that some of the helical regions in RNA are merely structural artifacts that are carried over from the DNA structure. Conversely, at least some palindromes must be forced in order to satisfy functional constraints imposed upon the RNA; the clover leaf structure of tRNA serves as a likely example.

3.1 Messenger RNA

To date a few mRNA's have been isolated, purified, and to some extent characterized. These include the message for silk fibroin, for wheat gluten, for the zymogen of cocoonase, and for a composite of the histones (33,34). While encouraging, this work has not yet been sufficient to prompt general conclusions.

One prospect for anticipating eventual experimental evidence is to make a statistical estimate of RNA secondary structure from a known amino acid sequence. This method may be coupled with deductions from available RNA sequence data. These techniques have been explored by White, Laux, and Dennis (35,36) and by Mark and

Petruska (37). Another approach lies in the possibility of enumerating and characterizing the entire set of messages that could code for a single protein. If fortuitously this set of potential mRNA's all have some interesting property in common, it follows that the actual mRNA would also share this property.

Unfortunately, the set of all messages, M, consistent with the amino acid sequence of even a small protein is computationally unwieldy, to say the least. In the case of ribonuclease, for example, M has 7.5 exp 22 members, almost on the order of Avogadro's number.

Without explicit enumeration, however, it is always possible to choose two distinguished members from the set M. These are:

1) $\overline{m}$ - the potential message that exhibits maximal secondary structure, and

2) $\underline{m}$ - the potential message that exhibits minimal secondary structure.

The algorithm for selecting $\overline{m}$ and $\underline{m}$ requires either a permissive or a restrictive choice whenever an unspecified base is encountered.

The distinguished elements $\overline{m}$ and $\underline{m}$ are characteristic extrema that bracket the set M. They are, in effect, a least upper bound and a greatest lower bound on the degree of secondary structure of any arbitrarily chosen member of M. As such, $\overline{m}$ and $\underline{m}$ may provide a way to characterize M without resorting to explicit enumeration. For example, if $\overline{m}$ exhibits only a small degree of secondary structure, then it is clear that the actual biological message also has only a small degree of secondary structure. On the other hand, if $\underline{m}$

exhibits a high degree of secondary structure, then the actual message also has a high degree of secondary structure.

The existence of characteristic extrema for any set M is interesting to the degree that it provides a tool to test interesting hypotheses about the secondary structure of mRNA's. One such hypothesis is that, in the general case, long-lived message confers a selective advantage on a cellular system, for in this case less metabolic energy is required to maintain the message pool in a steady state condition. Such a hypothesis is promoted by the observation of very long-lived message in a situation where the correlative protein is required in great abundance (27). The mechanism of action of ribonuclease further suggests that long-lived mRNA will have a high degree of nuclease resistant secondary structure (29).

Many realistic considerations have been excluded from this hypothesis such as the effect of ribosome attachment on nuclease action or the use of scarce tRNA's as protective masking devices. These simplifications are appropriate as the first object of this exercise is to see what kind of insight the application of characteristic extrema can provide into problems of this type.

A prediction of the hypothesis is that proteins with greater evolutionary lattitude will have messages with proportionately higher secondary structure since a mutation that confers a structural advantage on the message may well alter the amino acid sequence of the protein. Hence, arranging a set of proteins in order of increasing unit evolutionary period should arrange their respective messages in order of decreasing secondary structure.

To test this prediction, a set of sequenced proteins with differing unit evolutionary periods was chosen for examination. These included the histones f2b and f2al, human cytochrome c, and the alpha and beta chains of hemoglobin (38). In each case, the characteristic extrema were computed and inspected. The degree of secondary structure was measured using a method of Tinoco et al (39). The method consists of forming a matrix with diagonals that reflect all possible hydrogen bonded arrangements that can exist between bases. Thermodynamic criteria are then applied to assess the stability of each arrangement, and unstable loops are identified. While there is a minor disagreement about the thermodynamic criteria used to predict hairpin loops at the lower margin of stability in model systems, it is unlikely that computation of this threshold presents a problem in biological systems. In R17, MS II, and tRNA (38) the biologically significant configurations are more than stable; they are conspicuous.

The Tinoco matrix developed from a protein is going to be quite complex in appearance. If the protein has n amino acids, then the matrix has (6n-1) diagonals. In order to simplify this array, a program was developed to scan each diagonal in turn, remove any unstable structures, and extend stable folding trends to allow easier reading.

Examination of these data showed that the π for all five proteins could be entirely tied up in hairpin loops, while the $\underline{m}$ in each case was virtually devoid of loop structures. In retrospect, this result is hardly surprising since a loop alignment that avoids a 3:3 registration between the indeterminate third bases would stabilize loops for $\overline{m}$ and destabilize them for $\underline{m}$. Hence, M is too large, and the range between $\overline{m}$ and $\underline{m}$ is too broad to distinguish between the messages for these five proteins in this fashion. As a correlary to this conclusion, it appears that the set M is sufficiently rich that evolutionary changes in the protein need not occur at the expense of structural constraints on the message.

3.2 Large Palindromes

The set of messages, M, coding for a given protein has been shown to be very large. Nevertheless, it is always possible to single-out two boundary messages that bracket this set with respect to the degree of secondary structure. These distinguished members, $\overline{m}$ and $\underline{m}$, trap all remaining elements of M between them. In the general case, though, the range between $\overline{m}$ and $\underline{m}$ is too large to permit the existence of an effective forcing function on remaining members of M.

An article by Wilson and Thomas (32) reported the detection of very long palindromes in eukaryotic DNA. These palindromes are said to range from 300 to 6000 nucleotides in length, and experimental evidence indicates they are quite exact, with fewer than one percent base pairing differences.

If a very long, almost perfect palindrome is transcribed, its transcript should exhibit a hairpin loop of corresponding length. The loop, in turn, will appear as a long trace down a diagonal of the Tinoco matrix on the message. While imperfect pairing may cause a small gap in the trace, or even a jog over to another nearby diagonal where the trace is continued, such irregularities will not be sufficient to obscure the overall pattern in the matrix.

The existence of an extended trace in the Tinoco matrix imposes a severe structural constraint on the message, for in this case the high percentage of overall secondary structure must be packed into a single hairpin loop. For any protein, P, of known sequence, we can develop $\overline{m}$, the potential message in M that is most permissive of secondary structure, and this extremum can be examined for the existence of a long trace. If that trace is not apparent in $\overline{m}$, then we may conclude it does not appear in any message in the set M, and it follows that the gene for P does not contain a large palindrome. A representative collection of proteins was examined, and in each case the $\overline{m}$ for the protein was inspected for the existence of a long trace. A trace of suitable length was never found, and the tentative conclusion was reached that long palindromes do not reside in structural genes. Of course, the one configuration that cannot be excluded in an experiment of this sort is the possibility that a structural gene comprises half or less of an even larger palindrome.

The test set of proteins included the alpha and beta chains of hemoglobin, the histones f2al and f2b, and human cytochrome c. From each, an $\overline{m}$ was computed together with the Tinoco matrix on that $\overline{m}$. A computer program was written to pass a window of fixed size down every diagonal, advancing each frame one base pair at a time from upper right to lower left. Frames with complementary base pairing in excess of a specified threshold were marked. The window size and the percentage threshold were parameters to the program.

With a window of 35 base pairs and a pairing threshold in excess of 98%, no subtrace was found in the entire matrix of any protein in the test set. Relaxing these criteria to 85% pairing in a window of 25 base pairs, the trace patterns shown graphically in figure 3 were observed. These criteria are considered highly relaxed in view of the experimental evidence cited. A window size of at least an order of magnitude larger than the one used, as well as a pairing threshold in excess of 99% is indicated in the experimental studies.

The use of highly relaxed detection criteria coupled with the use of $\overline{m}$, which is really an upper bound on possible pairing, assure that no palindrome, as characterized by Wilson and Thomas, escaped notice by falling just below the margin of detection.

As a control for the above experiment, the MS2 coat protein cistron was subjected to the same treatment as the $\overline{m}$ for a protein

Figure 3

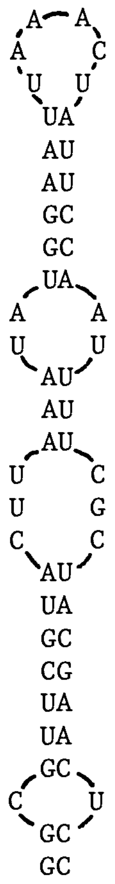
The graphs represent the Tinoco matrix of the MS2 coat protein cistron, an illustrative test region from the cistron, and proteins in the test set. Each diagonal was scanned by a program to find continuous regions with more than some specified threshold of pairing. The ratio shown in each figure is the pairing threshold over the window size in base pairs. Diagonal lines mark anti-parallel helical regions where these criteria were satisfied. The vertical scale is graduated in base pairs, while the horizontal scale is a nominal one; both scales divide the message into ten equal regions.

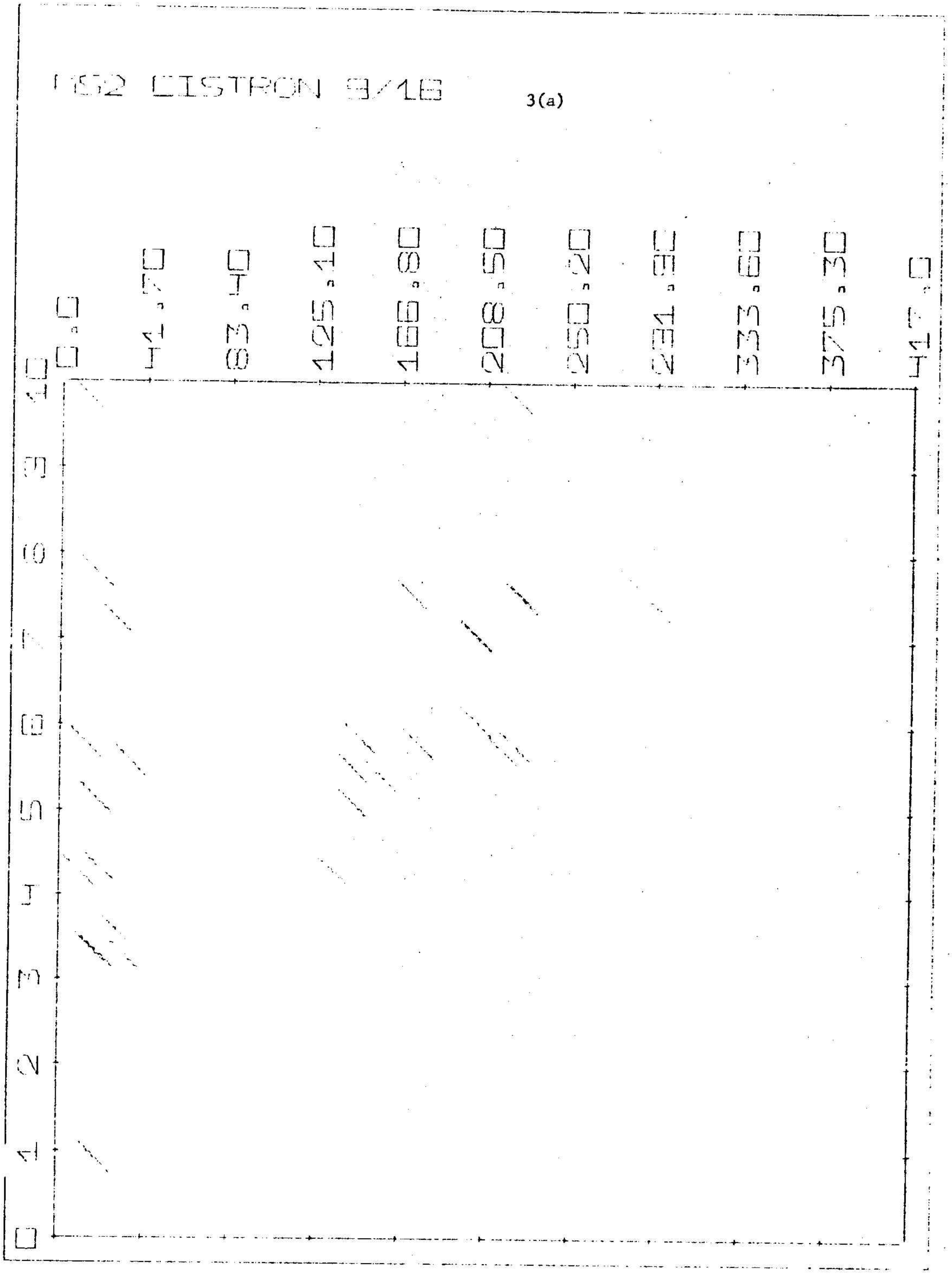
A diagonal line in one of the figures can be translated into the more typical hairpin diagram by using the axes to locate that diagonal within the matrix. The particulars of the pairing pattern can then be discovered by referring to a detailed printout of the Tinoco matrix. For example, the diagonal in the MS2 cistron test region that pairs bases 54,3,2,... with bases 2,3,4,... respectively is diagonal 55. In printout form, that diagonal appears as follows:

```
      Diagonal 55.

base             5555554444444444333333333322
number           5432109876543210987654321098

base             GCCUCAAGCAUCGCUUUUAACCUUAUCA
score             22 2112211   111  122111
base             UGGCGUUCGUACUUAAAUAUGGAAUUAA

base             1234567890123456789012345678
number                    1111111111222222222
```

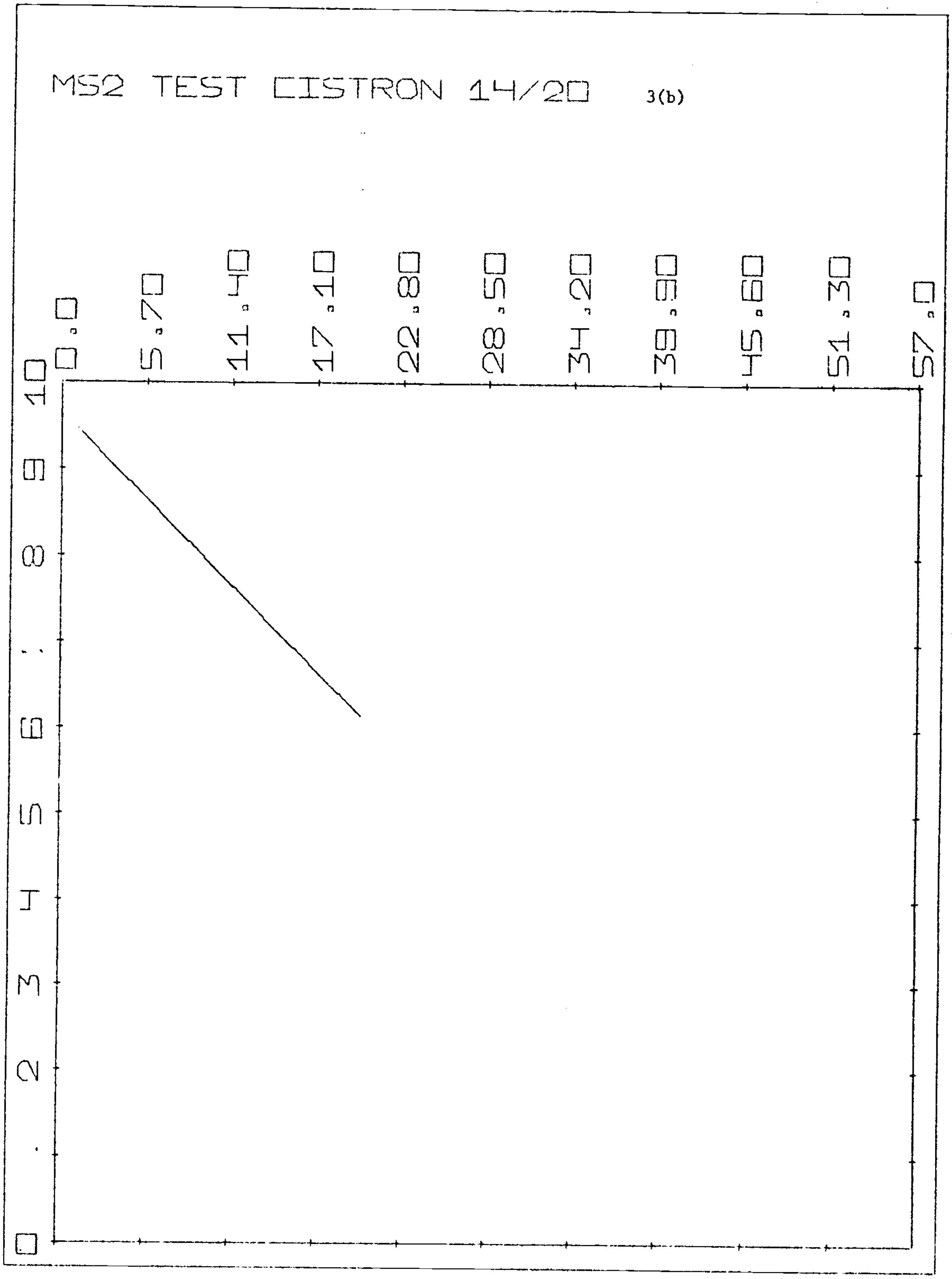
Figure 3(continued)

In the more familiar pictorial format, the helical region appears as:

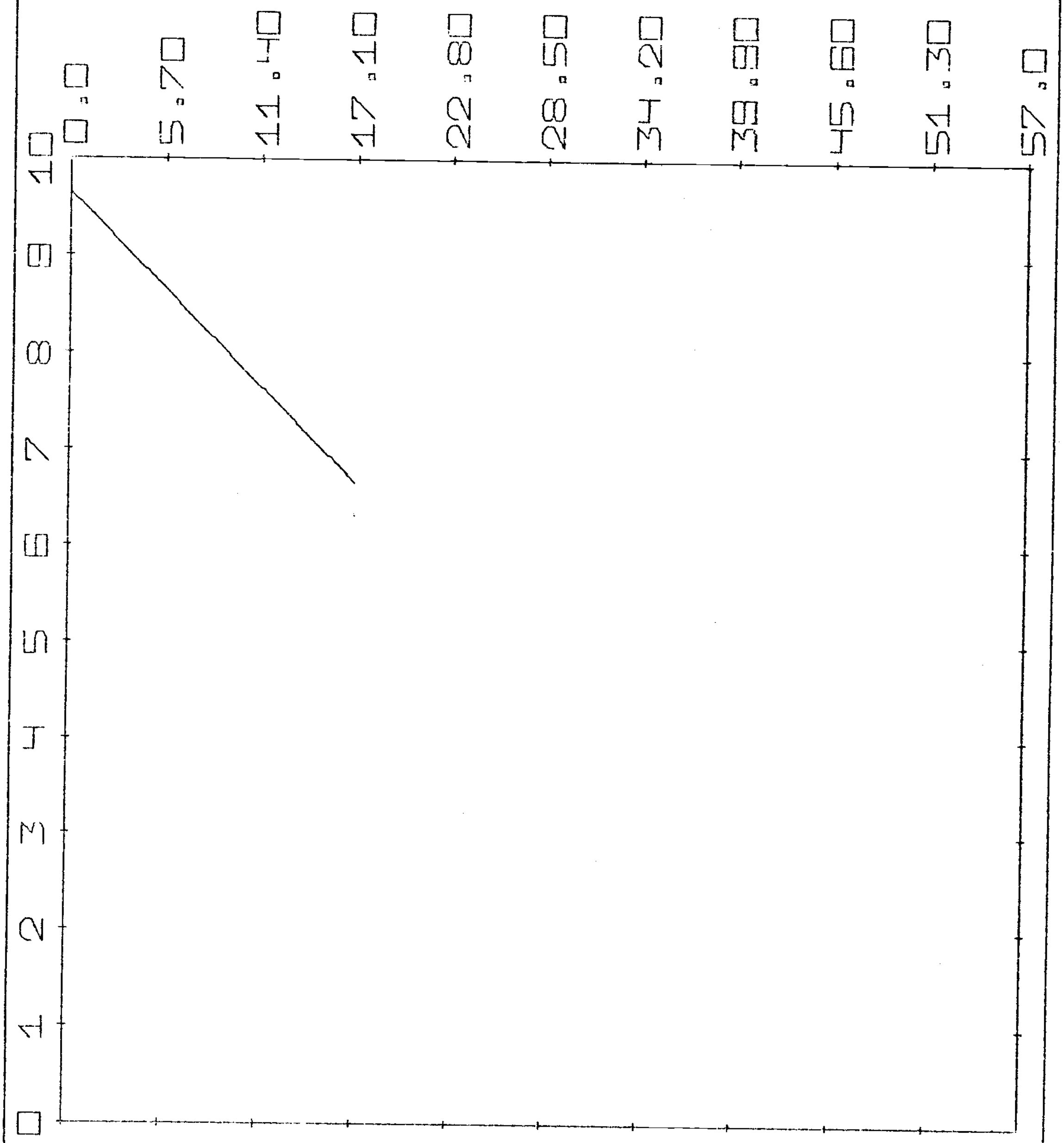


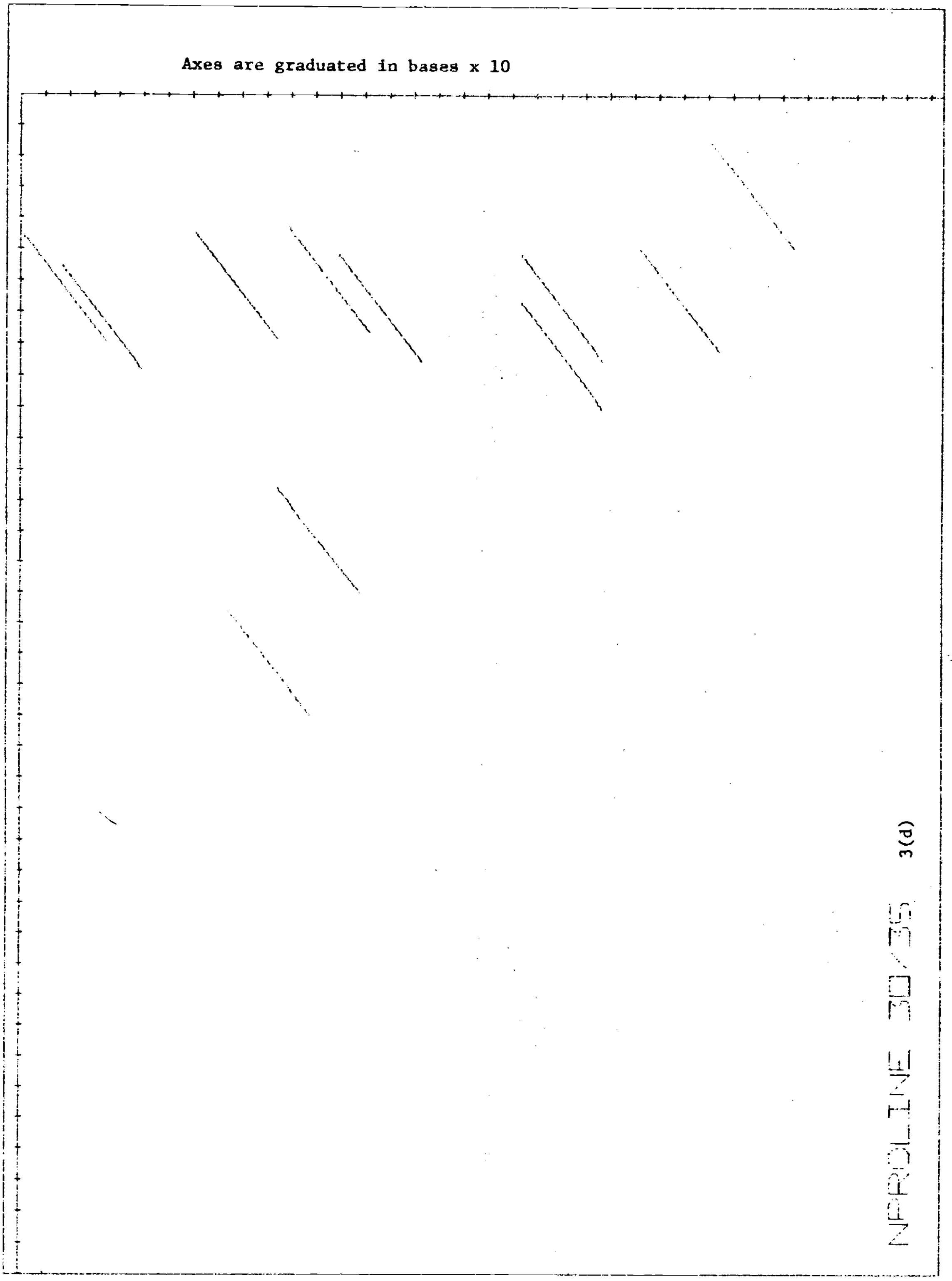

MS2 CISTRON 9/16
0.0
41.70
83.40
125.10
166.80
208.50
250.20
291.90
333.60
375.30
417.0
10
9
8
7
6
5
4
3
2
1
0

3(a)

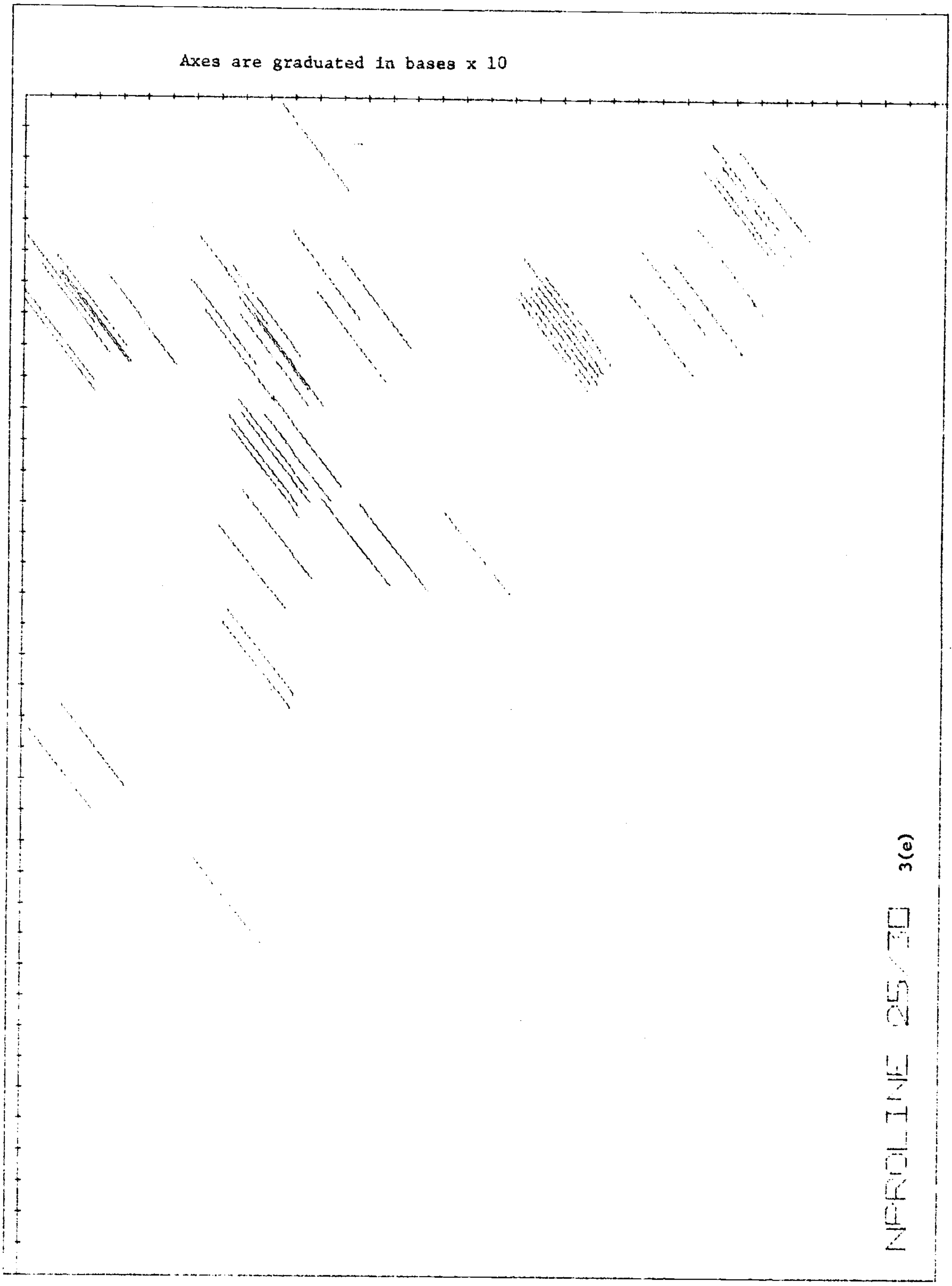

MS2 TEST CISTRON 14/20
3(b)
0.0
5.70
11.40
17.10
22.80
28.50
34.20
39.90
45.60
51.30
57.0
10
9
8
7
6
5
4
3
2
1
0

MS2 TEST CISTRON 12/18 3(c)



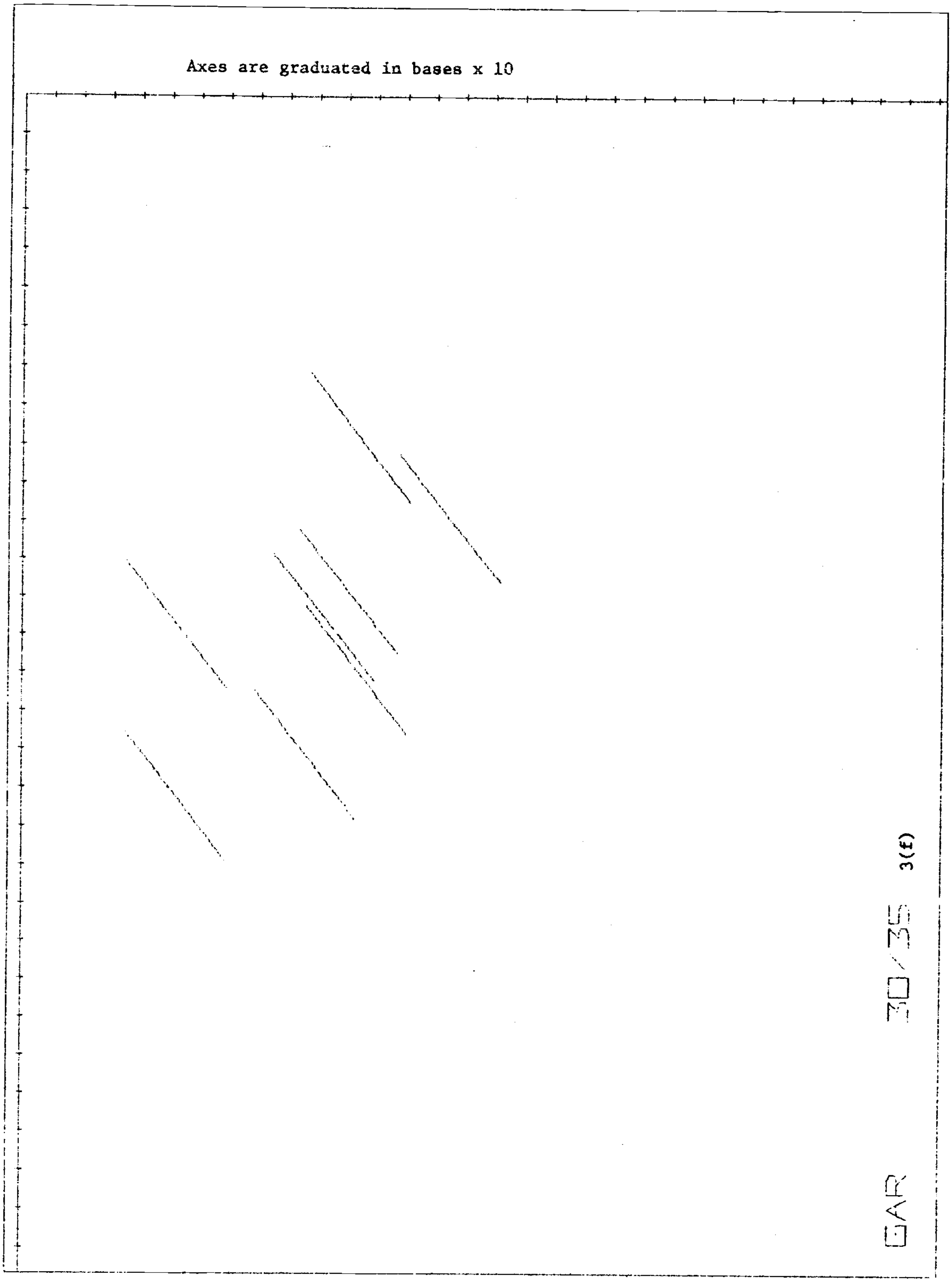

Axes are graduated in bases x 10

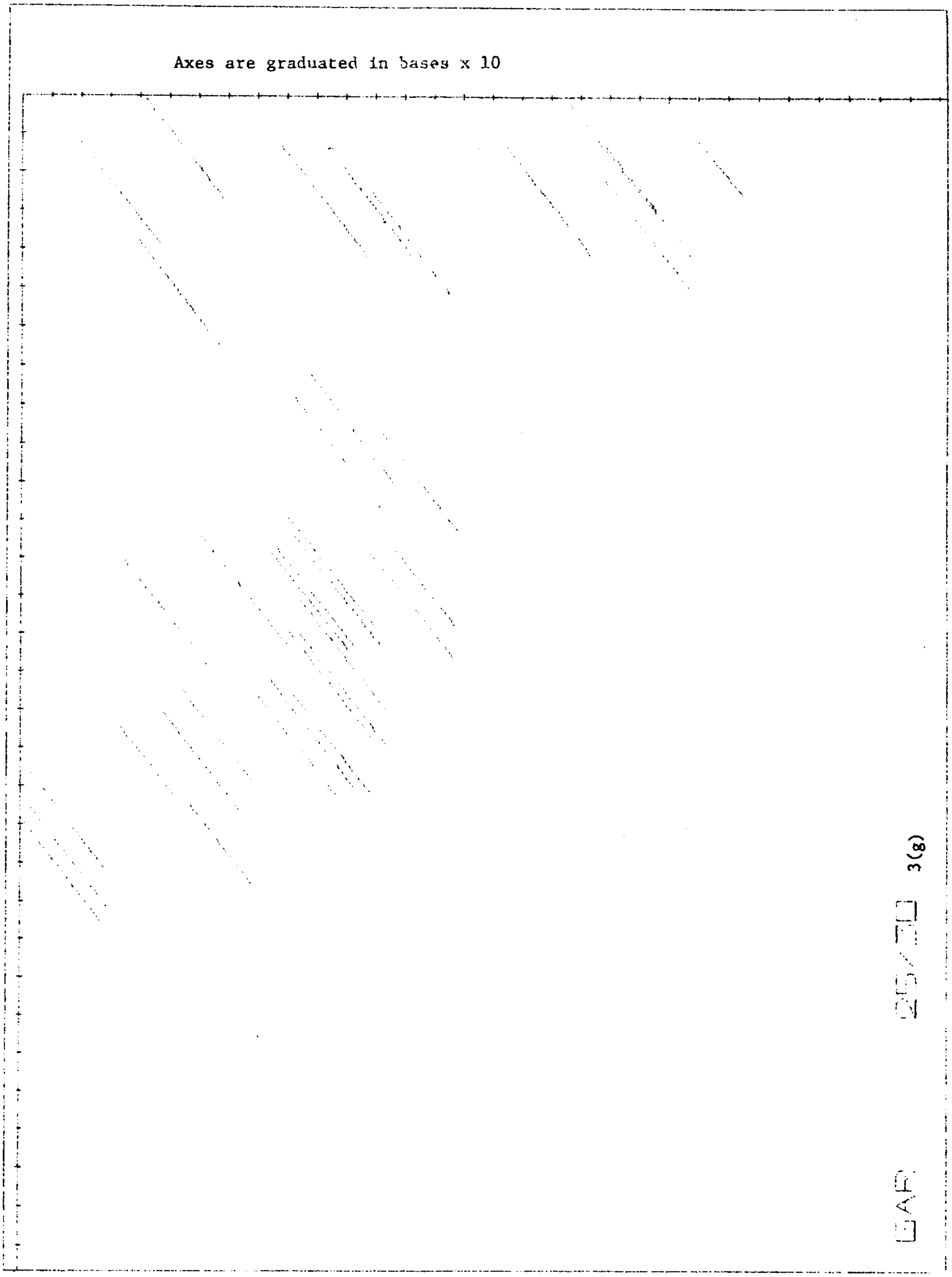
3(d)



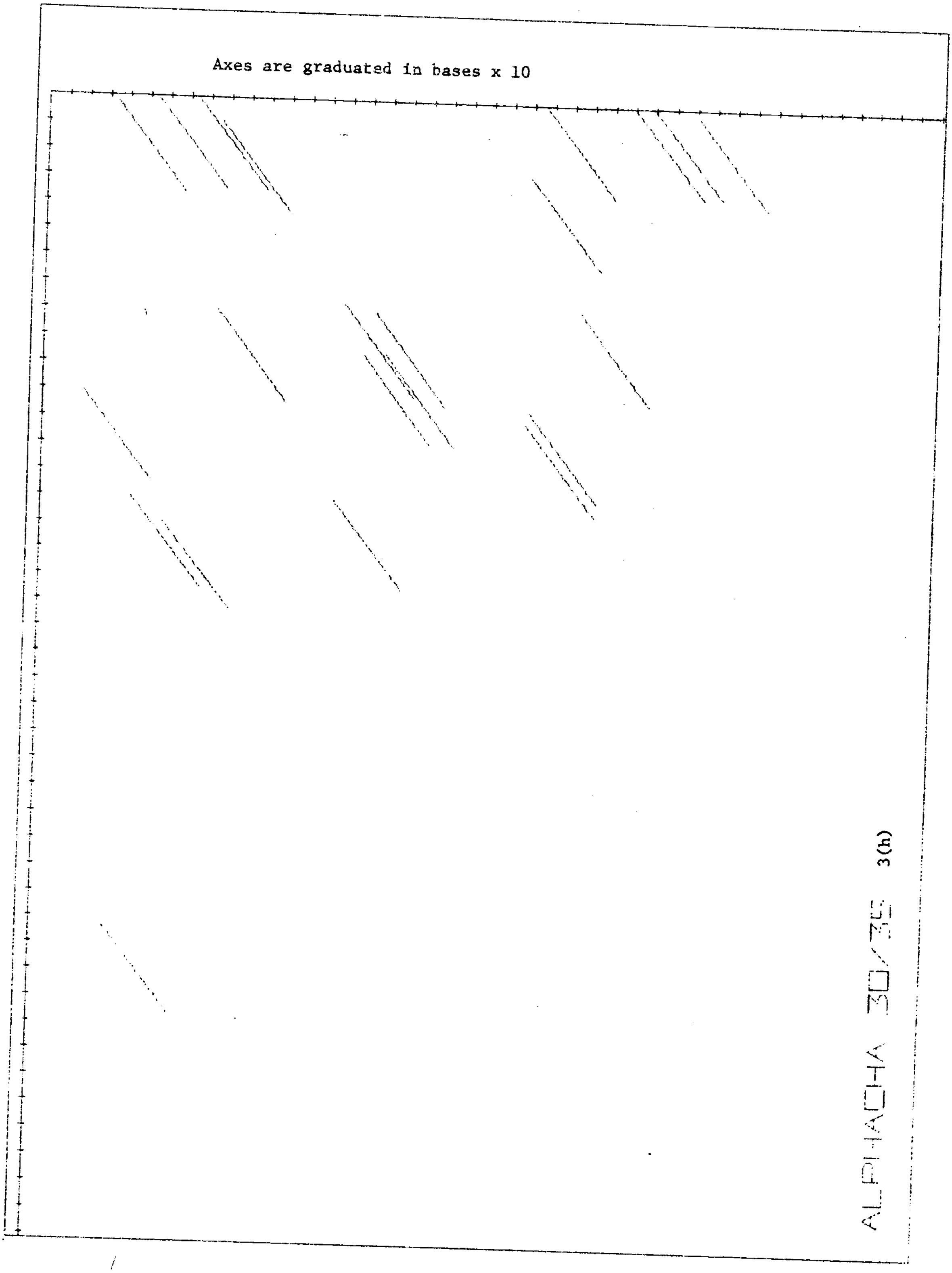
Axes are graduated in bases x 10

3(e)

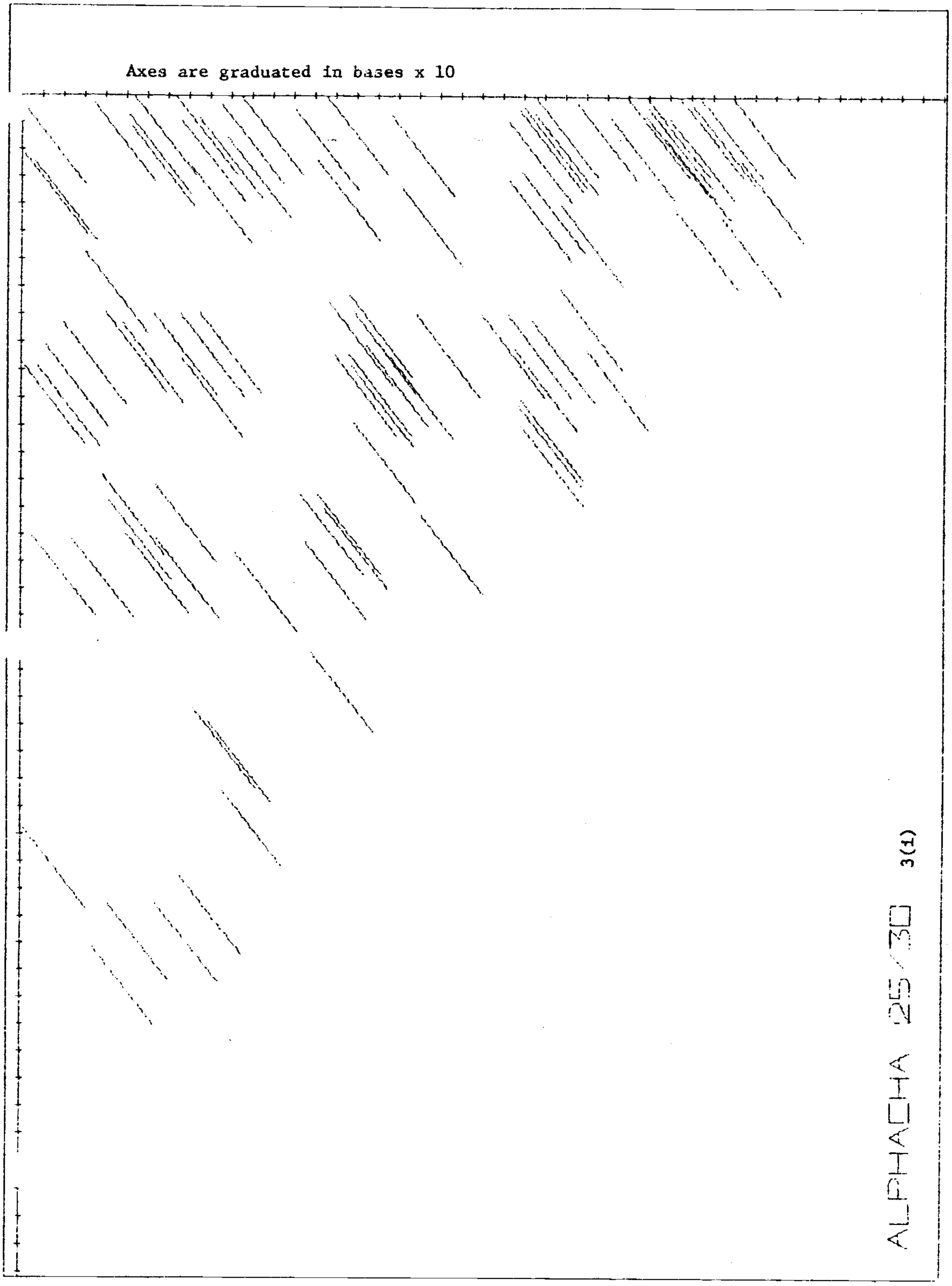

Axes are graduated in bases x 10
GAR 30/35
3(f)

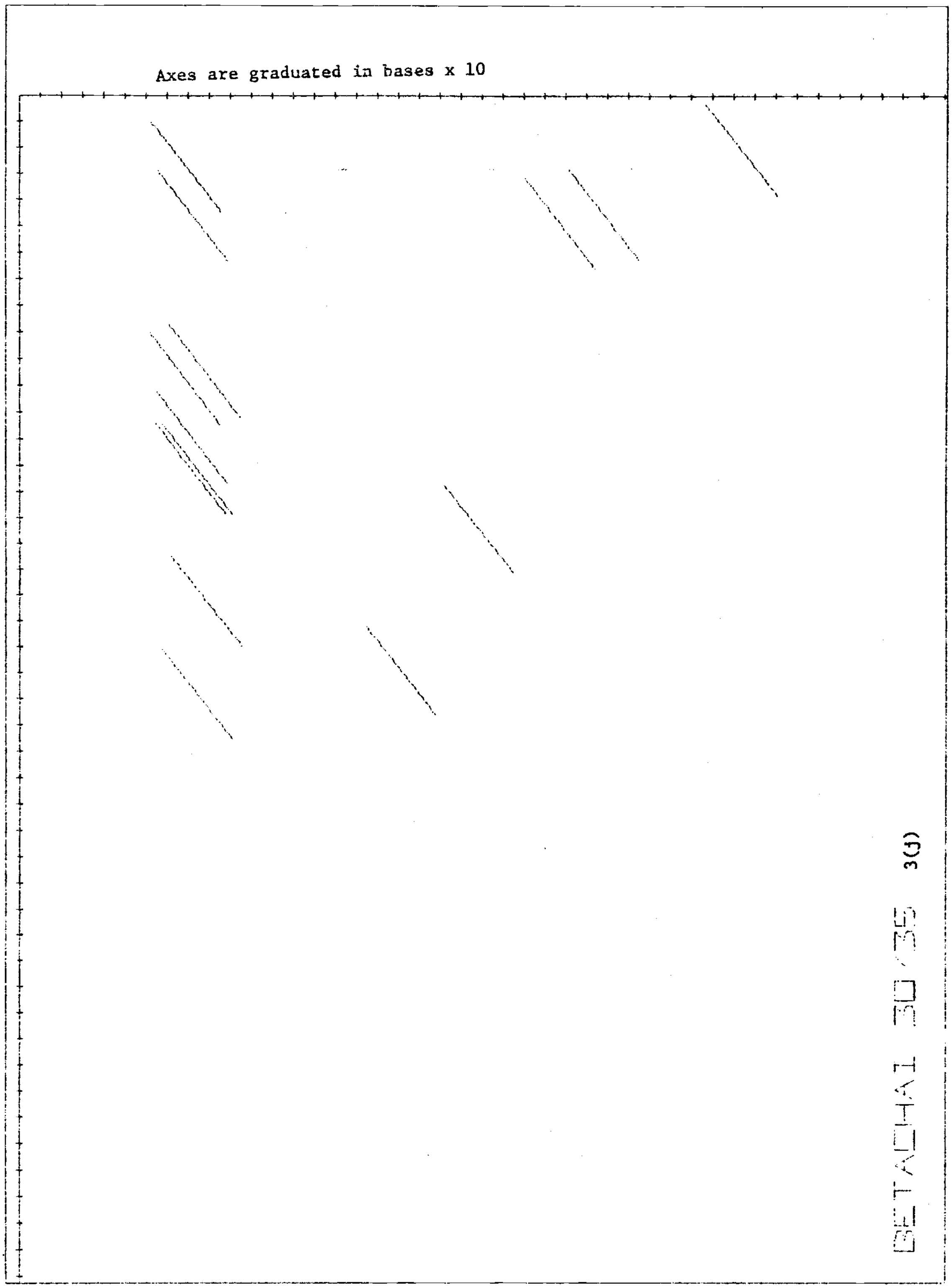


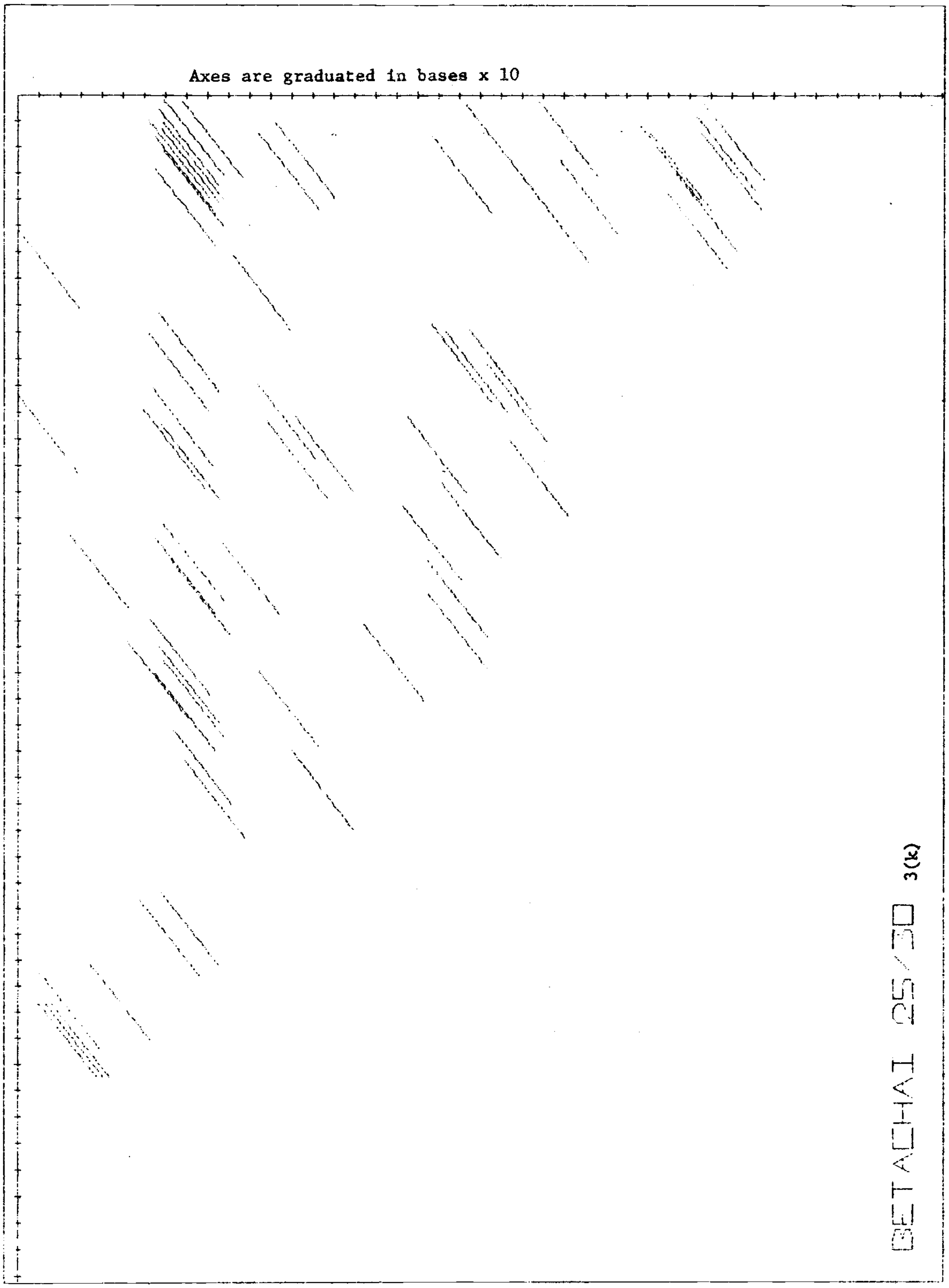
3(8)



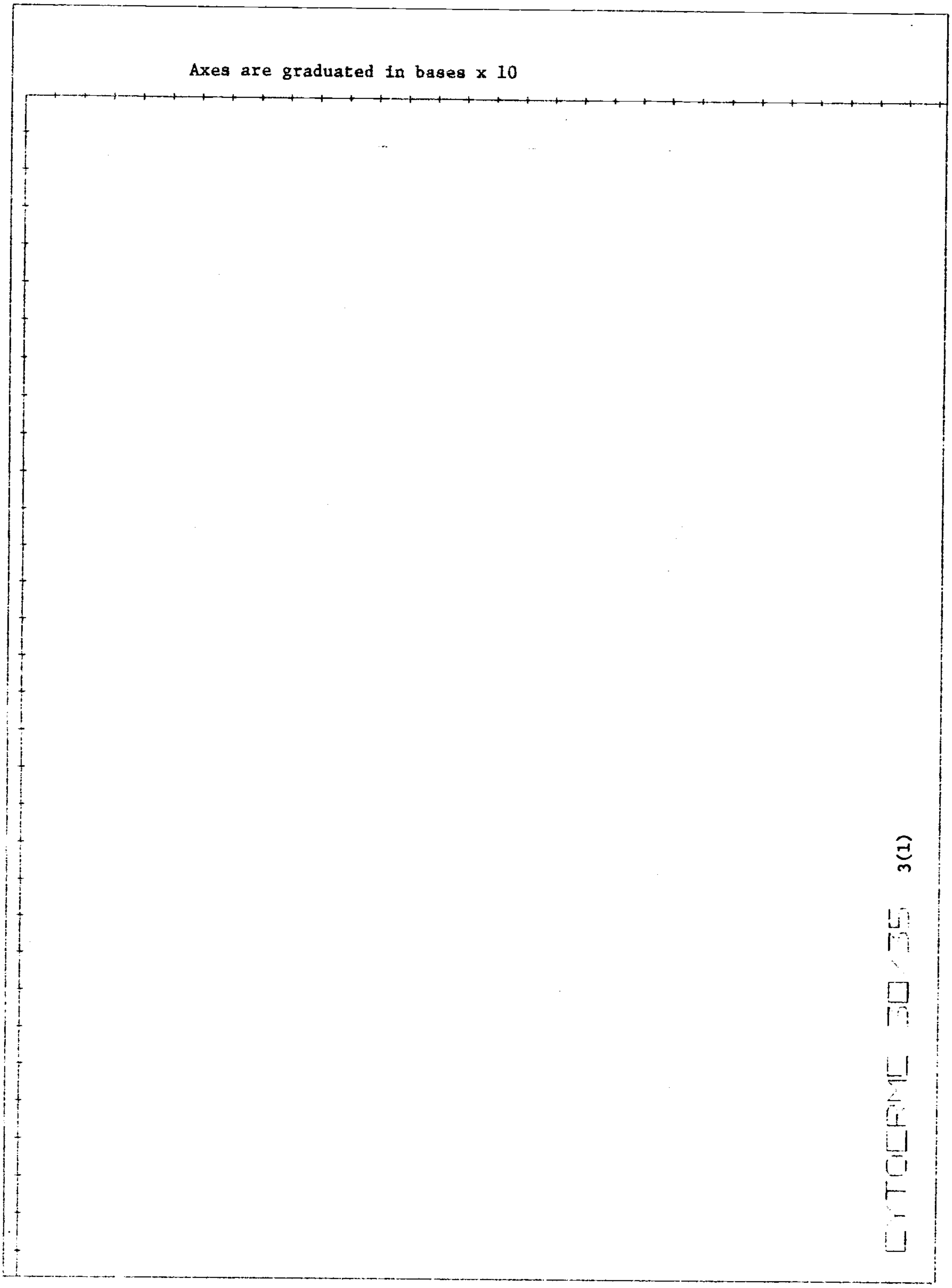
3(h)

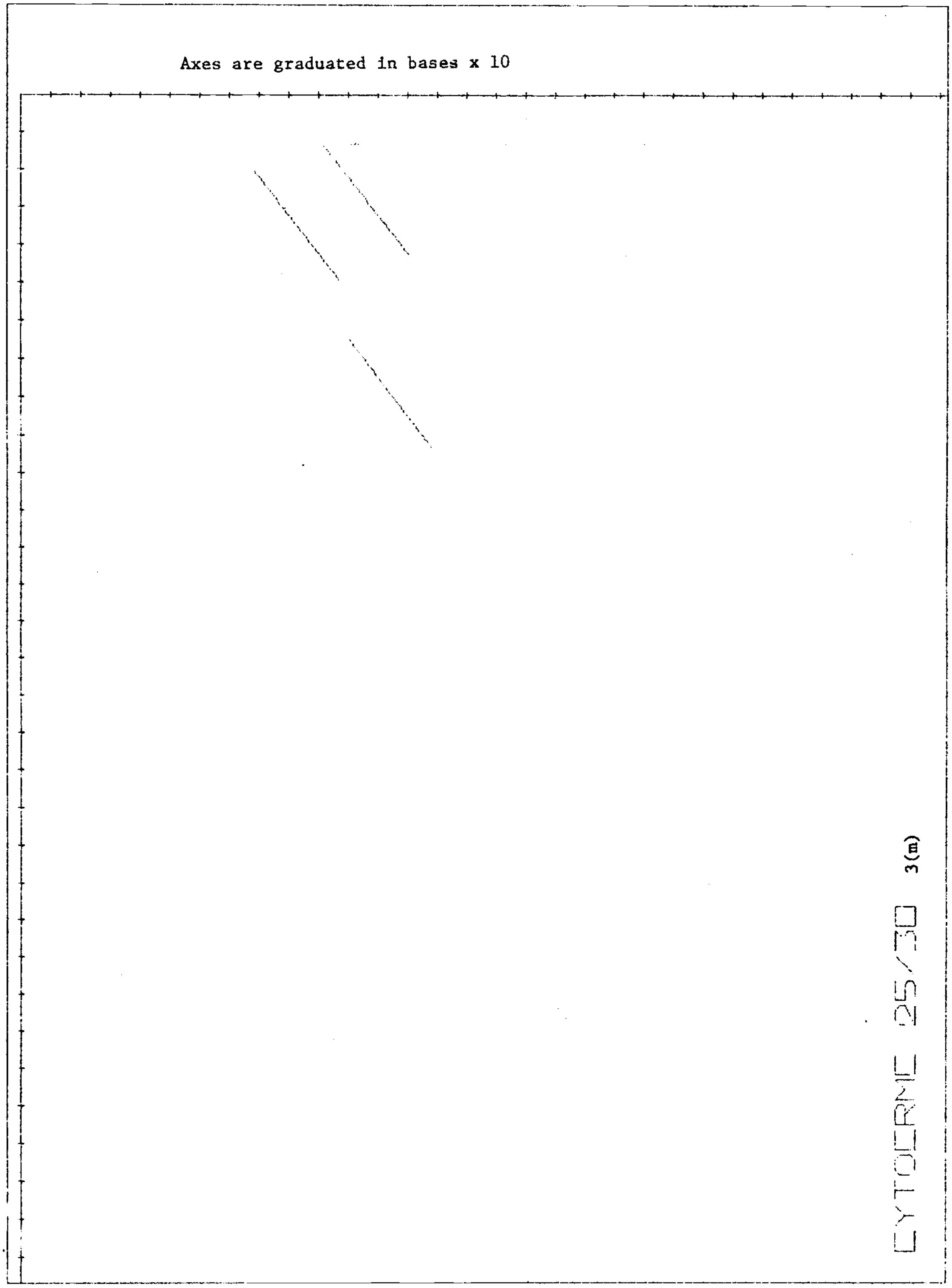

Axes are graduated in bases x 10
ALPHACHA 25/30
3(1)


Axes are graduated in bases x 10
BETACHAI 30/35 3(J)

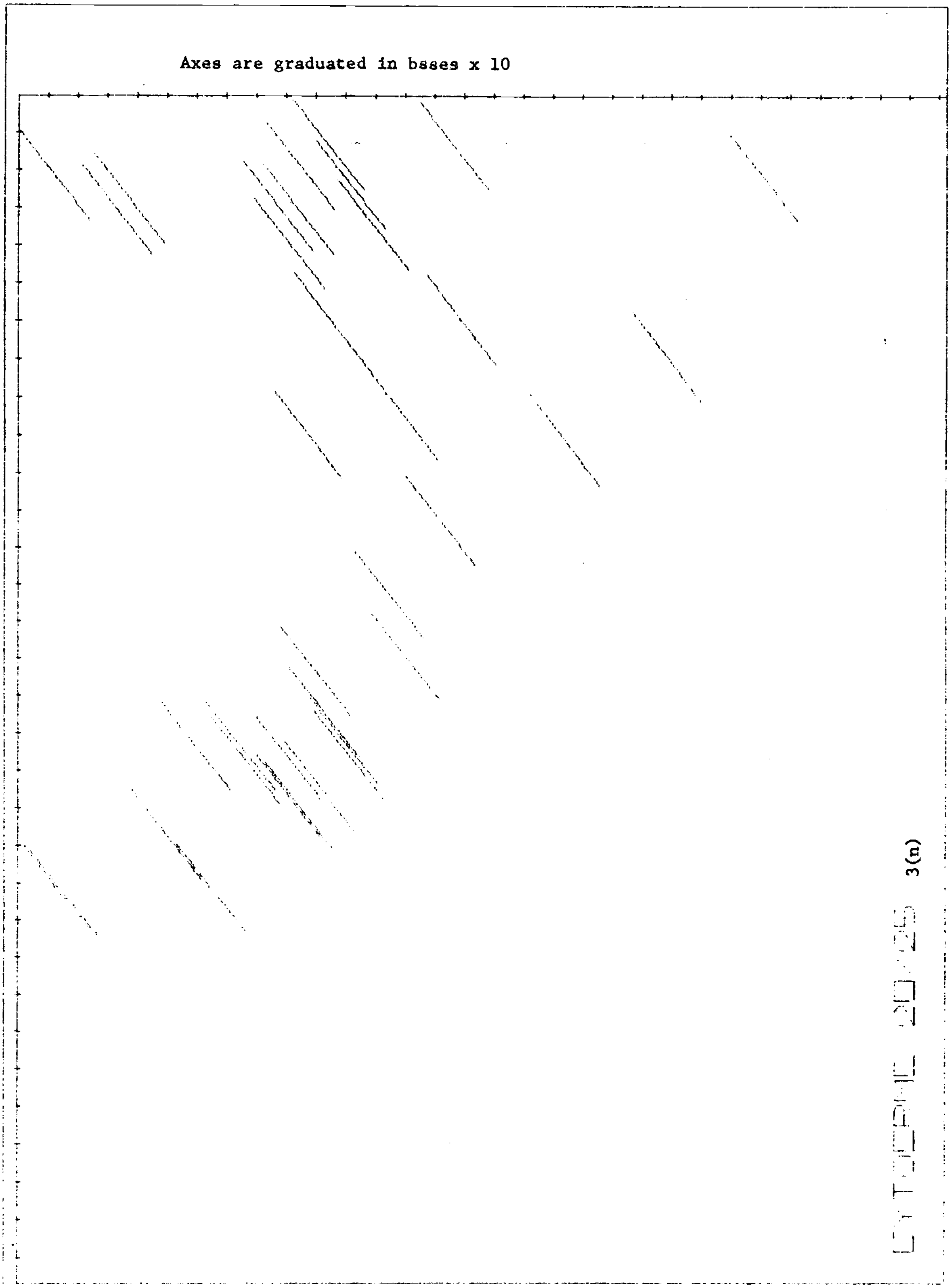


3(k)


Axes are graduated in bases x 10
CYTOCRME 30/35
3(1)



3(m)



3(n)

from the test set, with the computer held at constant conditions of temperature and pressure. Here the expected hairpin loop size is nine to twelve base pairs, and, in consequence, window sizes of sixteen to eighteen were chosen. This choice stands in sharp contrast to the previous computation in which the window size used was only one tenth of the expected loop size. At the thresholds shown, the trace patterns in figure 3 emerge; the base-paired 'petals'are readily apparent.

In passing, a technique was devised to winnow the set M to some proper subset M' by taking advantage of evolutionary data. In the case of cytochrome c, for example, sequence data for 34 species, from neurospora to human, are available (38). At a given amino acid position in the protein, there are in general only a small number of residues that occur; this number ranges between one and nine for the 34 species used. The assumption was made that amino acid substitutions in cytochrome c are the result of a single point mutation. Following this assumption, a program was written to examine all possible permutations of the amino acids at any position, and to discard all arrangements in which adjacent amino acids differed by more than one base in their respective codons.

After discarding all arrangements failing the assumption, three cases were found:

1) no arrangements remained - this could happen if an evolutionary precursor was not included in the 34 species.

2) multiple arrangements remained - in this case,

phylogenetic considerations were applied to choose a likely arrangement.

3) a unique arrangement remained - in this case, phylogenetic considerations were still needed to validate likelihood.

In cases two and three, there were instances where a unique evolutionary path was discovered that was both consistent with the assumption and seemed to make phylogenetic sense.

The codons for each amino acid along the discovered path were then examined, and it was often possible to eliminate codons that would have contradicted the assumption. Following this strategy, one can finally end up with a proper subset of codons for the amino acid used in human cytochrome c, and by applying the algorithm at every position, a winnowing of the whole set M is achieved. The process is shown schematically in Table 3.

Clearly, the reliability of this method depends upon a knowledge of the true evolutionary path taken. To this extent, the final result represents only an informed guess.

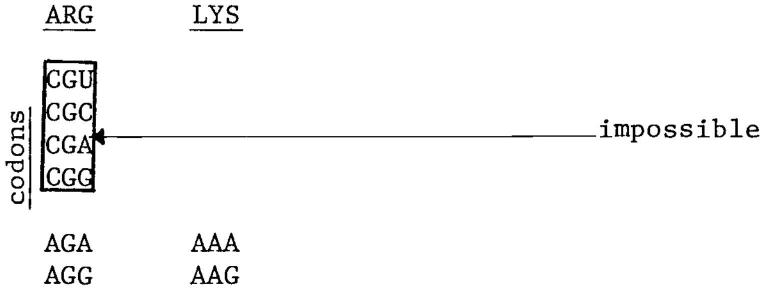
Substitution of the winnowed set, M', for the whole set M does modify the extrema $\overline{m}$ and $\underline{m}$. In practical terms, though, the use of modified extrema in the experiments previously described did not change or enhance their outcome. This is not to say, however, that other experiments will not be rendered possible by the use of this technique.

In summary, the method of characteristic extrema was used to examine the genes of a representative set of proteins for the

Table 3.

Winnowing of the Codon Set

At amino acid position 13 in cytochrome c, only two amino acids are used: lysine in mammals, other vertebrates, and in higher plants; and arginine in lower plants and insects. If LYS was substituted for ARG by a single point mutation, then the first four codons in the left hand column become logically impossible.



ARG	LYS
CGU	
CGC	
CGA	
CGG	
AGA	AAA
AGG	AAG

Multiple Paths

At amino acid position 39, three amino acids are used: lysine in mammals, other vertebrates, and insects; glutamine in higher plants; and histidine in two of the lower plants. Two arrangements are possible, each satisfying the assumption of a single point mutation.

Table 3 (continued)

1) $\underline{\text{LYS}} \longrightarrow \underline{\text{GLN}} \longrightarrow \underline{\text{HIS}}$

codons	AAA	CAA	CAU
	AAG	CAG	CAC

2) $\underline{\text{HIS}} \longrightarrow \underline{\text{GLN}} \longrightarrow \underline{\text{LYS}}$

codons	CAU	CAA	AAA
	CAC	CAG	AAG

Arrangement two is the preferred one based on phylogenetic criteria.

existence of very large palindromes. The non-existence of such palindromes in the set under test prompts a conclusion that long palindromic configurations do not occur in structural genes. The possibility that a structural gene participates as a fractional part of an even larger palindrome could not be excluded by the method used.

References

1. Brown, J. E. and Klee, W. A. (1971) Biochemistry 10, 470-6.

2. Villarejo, M.R. and Zabin, I. (1973) Nature New Biol. 242,50-2.

3. Anfinsen, C.B. (1973) Science 181, 223-30.

4. Saxena, V.P. and Wetlaufer, D.B. (1970) Biochemistry 9,5015-23.

5. Baldwin, R.L. (1975) Ann. Rev. of Biochemistry 44, 453-475.

6. Ptitsyn, O.B., Lim, V.I., and Finkelstein, A.V. (1972) Federation of European Biochemical Societies 25, 421-431.

7. Levinthal, C. (1968) J. Chim. Phys. 65, 44-45.

8. Kuntz, I.D. (1972) J. Am. Chem. Soc. 94, 4009-12.

9. Lewis, P.N., Momany, F.A., Scheraga, H.A. (1971) Proc. Nat. Acad. Sci., U.S.A. 65, 2293-97.

10. Wetlaufer, D.B. and Ristow, S. (1973) Ann. Rev. of Biochemistry 42, 135-158.

11. Schulz, G.E., Barry, C.D., Friedman, J., Chou, P.Y., Fasman, G.D., Finkelstein, A.V., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T.T., Levitt, M., Robson, B., and Nagano, K. (1974) Nature 250, 140-42.

12. Ptitsyn, O.B. (1975) Biophys. Chem. 3, 1.

13. Lim, V.I. (1974) J. Mol. Biol. 88, 857-894.

14. Gelin, B.R. and Karplus, M. (1975) Proc. Nat. Acad. Sci., U.S.A. 72, 2002-2006.

15. Warme, P.K. and Scheraga, H.A. (1974) Biochemistry 13, 757-82.

16. Levitt, M. and Warshel, A. (1975) Nature 253, 694-98.

17. Marsh, R.E. and Donohue, J. (1967) Adv. Prot. Chem. 22, 235-56.

18. Ramachandran, G.N. and Sasisekharan, V. (1968) Adv. Prot. Chem. 23, 326-438.

19. Brant, D.A. and Flory, P.J. (1965) J. Am. Chem. Soc. 87, 2791-2800.

20. Levitt, M. (1974) J. Mol. Biol. 82, 393-420.

21. Ramachandran, G.N. (1972) Conformation of Biological Molecules and Polymers; The Jerusalem Symposium on Quantum Chemistry and Biochemistry 5, 1.

22. Szent-Gyorgyi, A.G. and Cohen, C. (1957) Science 126, 697.

23. Karle, I.L. and Karle, J. (1964) Acta. Cryst. 17, 835-41.

24. Wright, D.A. and Marsh, R.E. (1962) Acta. Cryst. 15, 54-64.

25. Ramachandran, G.N., Mazumdan, S.K., Venkatesan, K., and Lakshminarayanan, A.V. (1966) J. Mol. Biol. 15, 232-42.

26. Arnott, S. and Dover, S.D. (1967) J. Mol. Biol. 30, 209-12.

27. Kafatos, F.C. (1972) Proceedings of the Vth Karolinska Symposium, Appendix C.

28. Brawerman, G. (1974) Ann. Rev. of Biochemistry 43, 621-42.

29. Hirs, C.H.W., Moore, S., Stein,W.H(1952) J. Biol. Chem. 200, 493-506.

30. Gilbert, W., Maizels, N., and Maxam, A. (1973) Cold Spring Harbor Symposium 38.

31. Nathans, D. and Smith, H.O. (1975) Ann. Rev. of Biochemistry 44, 273-293.

32. Wilson, D.A. and Thomas, C.A., Jr. (1974) J. Mol. Biol. 84, 115-144.

33. Robbins, E., Borun, T.W. (1967) Proc. Nat. Acad. Sci., USA 57, 409.

34. Gilmour, R.S., Dixon, G.H. (1972) J. Biol. Chem. 247,4621.

35. White, H.B., Laux, B.E. and Dennis, D. (1972) Science 175, 1264-1266.

36. Laux, B., Dennis, D., and White, H.B. (1973) Biochemical and Biophysical Research Communications 54, 894-898.

37. Mark, A.J., Petruska, J.A. (1972) J. Mol. Biol. 72, 609-17.

38. Dayhoff, M.O. (1972) Atlas of Protein Sequence and Structure Volume 5, National Biomedical Research Foundation, Wash. D.C.

39. Tinoco, I., Jr., Uhlenbeck, O.C. and Levine, M.D. (1971) Nature 230, 362-367.

40. Brezinski, D.P. (1975) Nature 253, 128-130.

APPENDIX

PROTEINS LIBRARY

```
      PROGRAM AALIST
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                              C
C  PROGRAM TO LIST THE AMINO ACID SEQUENCE                     C
C  FROM A PROTEIN                                              C
C  PARAMETERS ARE THE PROTEIN FILE NAME AND                    C
C  THE NUMBER OF THREE LETTER CODES PER LINE                   C
C                                                              C
C  OUTPUT DIRECTED TO LUN 12                                   C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                              C
```

```
      PROGRAM ADDHV2
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                              C
C  PROGRAM TO READ A CRYSTALLOGRAPHIC STRUCTURE AND            C
C  AUGMENT IT BY ADDING H ATOMS TO THE MAIN CHAIN AND TO       C
C  THE SIDE CHAINS OF THE FOLLOWING RESIDUES                   C
C      SER,THR,TYR,LYS,ARG                                     C
C                                                              C
C  OUTPUT TO LUN 12                                            C
C  OUTPUT AUGMENTED FILE TO LUN 13                             C
C                                                              C
C  WORKS WITH FILE OF COORDINATES*10.                          C
C                                                              C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
      PROGRAM BENDER
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                              C
C  PROGRAM TO COMPUTE SUCCESSIVE DIHEDRAL AND BEND ANGLES      C
C  FOR THE BYRON BENDER                                        C
C  INPUT IS THE ATOMIC COORDINATES FOR A PROTEIN               C
```

```
C  OUTPUT TO LUN 12, WHICH IS                                        C
C     EQUIPPED AS A FILE IF NOT ALREADY EQUIPPED                     C
C     REWOUND IF A FILE                                              C
C                                                                    C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM CONMAP
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                    C
C  PROGRAM TO PRINT OUT THE 10A CONTACT MAP                          C
C  FROM THE COORDINATES FOR A PROTEIN                                C
C                                                                    C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM DIHEDRAL
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                    C
C  PROGRAM TO PRINT OUT THE PHI,PSY,AND CHI DIHEDRAL ANGLES          C
C  FOR A PROTEIN WITH KNOWN COORDINATES                              C
C                                                                    C
C  OUTPUT DIRECTED TO LUN 12                                         C
C  OUTPUT CONSISTS OF COORDINATE DATA                                C
C  THEN A FILE MARK                                                  C
C  THEN DIHEDRAL ANGLES                                              C
C  FOLLOWED BY A SECOND FILE MARK                                    C
C                                                                    C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM EPATH
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C  PROGRAM TO FIND AN EVOLUTIONARY PATH THROUGH AN AMINO ACID        C
C  SEQUENCE.                                                         C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
      SUBROUTINE ALLPATHS(N,JAA,JNUM,JPATHS)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                    C
C  SUBROUTINE TO FIND ALL POSSIBLE EVOLUTIONARY PATHS WITH           C
C  DISTANCE 1 THROUGH N AMINO ACIDS                                  C
C  EACH POSSIBLE PATH IS CHARACTERIZED BY ITS                        C
C  PERMUTATION NUMBER.                                               C
C  OUTPUT IS RETURNED IN JPATHS. IT CONSISTS OF JNUM PERMUTATION     C
C  NUMBERS, ONE FOR EVERY LEGAL PATH.                                C
```

```
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM NEXTRES
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C  PROGRAM NEXT RESIDUE                                                C
C  PROGRAM TO START FROM THE N-TERMINAL END OF A PROTEIN               C
C  AND STOPPING AT CHOSEN SPOTS, TO ADD A RESIDUE                      C
C                                                                      C
C  INPUT COORDINATES SHOULD BE IN ANGSTROMS*10                         C
C                                                                      C
C  SINCE WORKATOM COORDINATES GET SUPERIMPOSED ON THE COORD ARRAY      C
C  IT IS NECESSARY TO WORK FROM THE C TERMINAL END BACK                C
C  TO THE N-TERMINAL END WHEN MAKING POTENTIAL MAPS                    C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM PDISPLAY
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C  PROGRAM TO DISPLAY A POLYPEPTIDE                                    C
C  OPTIONS INCLUDE                                                     C
C     DISPLAY WITH ALPHA CARBONS INDEXED                               C
C     WHOLE BACKBONE SELECTIVELY DISPLAYED                             C
C     SIDE GROUPS SELECTIVELY DISPLAYED                                C
C     INDIVIDUAL ACID TYPES MAY BE SELECTIVELY DISPLAYED               C
C     PARTICULAR RANGES MAY BE SELECTIVELY DISPLAYED                   C
C                                                                      C
C     H-HARD COPY                                                      C
C    SP-ROTATE THE PICTURE                                             C
C     R-RESET                                                          C
C   DEL-EXIT                                                           C
C     Z-ZOOM                                                           C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
      SUBROUTINE LABEL(JJ,JROT,KOPY)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C  SUBROUTINE TO LABEL THE ALPHA CARBONS.                              C
C  PARAMETERS ARE-                                                     C
C     JJ- ALPHA CARBON SEQUENCE NUMBER                                 C
C     JROT- 0 IF LEFT FIGURE, 4 IF RIGHT FIGURE                        C
C                                                                      C
C  LABEL DESTROYS THE CURRENT SCALING AND ROTATION                     C
```

```
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM PLANECK
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
C  PROGRAM TO CHECK THE BACKBONE PLANARITY OF A POLYPEPTIDE CHAIN. C
C  TAKES THE PLANES    CA-N-H    AND                               C
C                      CA-C O                                      C
C  AND ESTABLISHES A NORMAL TO EACH PLANE. THEN MEASURES THE ANGLE C
C  BETWEEN THE TWO NORMALS.                                        C
C                                                                  C
C  ALL OUTPUT IS DIRECTED TO LUN 12 WHICH IS EQUIPPED TO BE A FILE C
C  UNLESS OTHERWISE EQUIPPED                                       C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM POLYMER
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
C  PROGRAM TO PRODUCE A POLYMER WITH SPECIFIED                      C
C  PRIMARY STRUCTURE                                               C
C  ALL TORSION ANGLES IN BACKBONE ARE 180                          C
C  WILL SIDE CHAIN ANGLES ARE 0                                    C
C  INPUT FROM A NAMED FILE                                         C
C  OUTPUT TO LUN 12                                                C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM SHUFFLE
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
C  PROGRAM TO CANGE THE ORDER OF A FILE OF ATOMIC COORDINATES      C
C  FROM THE ORDER N,H,CA,HA,CB,...,C,O                             C
C  TO THE ORDER    CA,N,H,C,O,HA,CB,...                            C
C                                                                  C
C  INPUT FROM A NAMED FILE                                         C
C  OUTPUT TO LUN 12                                                C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
      PROGRAM SYNTHESIZE
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C   SYNTHESIZE A BACKBONE THAT IS NUMRES LONG.                         C
C   READS NUMRES (PHI,PSY) ANGLES FROM A FILE AND                      C
C   TWISTS THE BACKBONE ACCORDING TO THE INTOWELL TRANSFORM            C
C   OF THOSE SPECIFIED ANGLES                                          C
C   PRO AND GLY ARE NOT TRANSFORMED, HOWEVER.                          C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
      SUBROUTINE INTOWELL(A1,A2,A3,A4)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C   SUBROUTINE TO COMPUTE THE DIHEDRAL ANGLES OF AN ENERGY WELL        C
C   THAT IS CLOSEST TO THE RAMACHANDRAN POSITION SPECIFIED             C
C   BY A DIHEDRAL ANGLE PAIR (A1,A2)                                   C
C                                                                      C
C   A1,A2 ARE THE ACTUAL (PHI,PSY) ANGLES                              C
C   A3,A4 ARE DIHEDRAL ANGLES OF THE CLOSEST WELL                      C
C                                                                      C
C   ALL ANGLES ARE IN DEGREES                                          C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM TRUER
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C   PROGRAM TO COMPUTE A BASE PLANE AND MARK SELECTED                  C
C   RESIDUES FOR "TUNING-UP" A RENDER MODEL.                           C
C                                                                      C
C   INPUT CONSISTS OF                                                  C
C    1) ATOMIC COORDINATES FOR THE PROTEIN                             C
C    2) A FILE OF RESIDUES OF INTEREST. THE FIRST THREE OF             C
C       THESE RESIDUES ARE TAKEN TO DEFINE THE ZERO PLANE              C
C       FILE SHOULD CONTAIN ONE RESIDUE PER RECORD                     C
C                                                                      C
C   KEY RESIDUES 1 AND 2 ARE LINED UP ALONG THE + X-AXIS               C
C   IN THIS ORIENTATION, THE PROGRAM ASKS WHETHER KEY RESIDUE          C
C   3 HAS POSITIVE OR NEGATIVE Y COORDINATE (IN THE XY PLANE           C
C   THIS INFORMATION IS USED TO ROTATE THE COORDINATES                 C
C   INTO STANDARD ORIENTATION                                          C
C                                                                      C
C   OUTPUT CONSISTS OF COORDINATES OF THE BENCH MARK RESIDUES          C
C   TOGETHER WITH A SCALED PLOT OF THEIR (X,Y) POSITIONS               C
```

```
C                                                                    C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM UNBEND
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                    C
C  PROGRAM TO COMPUTE A SET OF CONSISTENT C ALPHA COORDINATES        C
C  FROM THE DIHEDRAL AND BEND ANGLES                                 C
C                                                                    C
C  INPUT IS A FILE CONSISTING OF RECORDS, EACH CONTAINING            C
C  DIHEDRAL ANGLE...BEND ANGLE...CA(I)-CA(I+1) LENGTH                C
C                                                                    C
C  OUTPUT TO LUN 12, WHICH IS                                        C
C     EQUIPPED AS A FILE IF NOT ALREADY EQUIPPED                     C
C     REWOUND IF A FILE                                              C
C                                                                    C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM OLIGOPEP
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                               C
C PROGRAM TO COMPUTE OLIGOPEPTIDE ENERGIES                       C
C                                                               C
C INPUT CONSISTS OF                                             C
C 1) SOME FILES OF (PHI,PSY) PAIRS WHICH COMPRISE THE DOMAINS  C
C    FOR THE BACKBONE ANGLES SETS                               C
C 2) SOME FILES OF SIDE CHAIN ANGLES THAT COMPRISE THE          C
C    DOMAINS OF SIDE CHAIN ANGLE SETS                           C
C 3) A FILE THAT MAPS EACH RESIDUE INTO ONE AND ONLY ONE        C
C    OF THE DOMAIN SETS, CONSISTS OF A DOMAIN NUMBER            C
C    PER RECORD AND ASSUMED TO BE IN RESIDUE ORDER              C
C 4) A FILE THAT MAPS EACH SIDECHAINS BONDS INTO ITS DOMAIN     C
C    OF DEFINITION, THE FILE HAS ONE RECORD FOR EACH RESIDUE    C
C    EVERY SUCH RECORD HAS K ITEMS, ONE FOR EACH ROTATABLE      C
C    BOND IN THE SIDE CHAIN OF THAT RESIDUE                     C
C 5) A FILE OF COORDINATES FOR THE OLIGOPEPTIDE AS IT           C
C   EXISTS IN TOTALLY UNFOLDED CONFORMATION,I.E.                C
C   WITH BACKBONE ANGLES=180 AND SIDECHAIN ANGLES=0             C
C                                                               C
C OUTPUT CONSISTS OF                                            C
C 1) A FILE NAMED WHY OF ALLOWED CONFIGURATIONS                 C
C 2) A FILE NAMED WHYNOT OF CONTACT INHIBITED CONFIGURATIONS    C
C                                                               C
```

```
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC



            RNA LIBRARY




      PROGRAM AATOAUG
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     PROGRAM TO READ A FILE OF AMINO ACIDS AND TRANSLATE              C
C     THEM TO AUGONS                                                   C
C     OUTPUT TO LUN 12                                                 C
C     INPUT HAS ONE AMINO ACID PER RECORD-3 LETTER CODES               C
C     OUTPUT HAS UP TO 2 AUGONS PER RECORD                             C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM EPATH
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     PROGRAM TO FIND THE EVOLUTIONARY PATHS THROUGH                   C
C     A SERIES OF AMINO ACIDS.                                         C
C     INPUT IS AN AMINO ACID SEQUENCE WITH EVOLUTIONARY                C
C     ALTERNATIVES. EACH RECORDS IS OF THE FORM                        C
C        ITH AA    AA1,AA2,...,AAJ                                     C
C     OUTPUT IS A SERIES OF AUGONS WHERE EACH RECORD IS                C
C     OF THE FORM                                                      C
C        ITH AA    AUG1,AUG2                                           C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM FOLDRNA
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     PROGRAM TO INSPECT AN AUGON CHAIN, IDENTIFY                       C
```

```
C     THE AUGONS (UNIONS OF CODONS) THAT CODE FOR EACH.                C
C     AND FORM A TINOCO MATRIX FOR THE SECONDARY STRUCTURE             C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE TINOCO(MINMAX)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     SUBROUTINE TO CONSTRUCT AN (NXN)/2 TRIANGULAR TINOCO             C
C     MATRIX ON A FILE OR RAF.                                         C
C     CONSTRUCTION TAKES PLACE 1 ROW AT A TIME                         C
C     THE I-TH ROW HAS   MAXBASE-I+1  COLUMNS                          C
C                                                                      C
C     THIS ROUTINE IS LIMITED TO AT MOST 2000 BASES                    C
C     PARAMETERS ARE MINMAX=                                           C
C        0=TAKE MINIMUM SECONDARY STRUCTURE                            C
C        1=TAKE MAXIMIM SECONDARY STRUCTURE                            C
C                                                                      C
C     OUTPUT IS TO LUN 48                                              C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE SIMPLIFY
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     SUBROUTINE TO SIMPLIFY THE TINOCO MATRIX ON LUN 49               C
C     OUTPUT NEW SIMPLIFIED MATRIX ON LUN 48                           C
C                                                                      C
C     SIMPLIFICATION OCCURRS BY                                        C
C        1) ELIMINATING ALL SINGLETONS                                 C
C        2) EXTENDING ANY RUNS                                         C
C                                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE SCREEN
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                      C
C     SUBROUTINE TO TRACE THROUGH ALL THE DIAGONALS OF A               C
C     SIMPLIFIED TINOCO MATRIX IN ORDER TO ELIMINATE THOSE             C
C     DIAGONALS THAT FALL BELOW SOME DESIGNATED THRESHOLD              C
C     VALUE                                                            C
C     THE MATRIX SHOULD BE A RAF ON LUN 48                             C
```

```
C                                                                  C
C     FOR EACH DIAGONAL, COMPUTES THE PERCENTAGE OF                C
C     BASE PAIRING, AND WRITES IT ONTO LUN 47                      C
C                                                                  C
C     THEN LOOKS ALONG EACH DIAGONAL FOR BLOCKS                    C
C     OF LENGTH > LB AND DISCARDS THE FIRST ELEMENT OF THE BLOCK  C
C     IF THAT BLOCK CONTAINS LESS THAN MIN BASE PAIRS             C
C     AFTER WHICH DOES A FRAME SHIFT OF ONE AND CONTINUES THROUGH C
C     THE DIAGONAL                                                 C
C     IF AT LEAST MIN BASE PAIRS ARE DISCOVERED WITHIN THE BLOCK  C
C     THEN ANY SPACES ARE 0 FILLED AND THE NEXT BLOCK IS TAKEN    C
C     FROM THE (LB+1)TH LOCATION                                   C
C     OUTPUTS THE DIAGONALS TO LUN 46                              C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE PLOTTM
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
C     SUBROUTINE TO PLOT THE DIAGONALS OF THE FINDCO MATRIX        C
C     STORED ON LUN 46                                             C
C     THE MATRIX IS TRIANGULAR AND IS STORED ONE DIAGONAL          C
C     PER RECORD                                                   C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM PALDROME
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
C     PROGRAM TO INSPECT A CODON CHAIN AND STEP DOWN               C
C     FROM 5' TO 3' WITH A WINDOW SIZE OF W, LOOKING               C
C     FOR PALINDROMES OF LENGTH L OR MORE                          C
C     THIS VERSION ALLOWS SELECTION FOR                            C
C        PERFECT OR IMPERFECT PALINDROMES                          C
C        MINIMAL OR MAXIMAL SECONDARY STRUCTURE                    C
C        ONE BASE CHOICE OR TWO BASE CHOICES                       C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      PROGRAM GCANAL
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                                  C
```

```
C      PROGRAM TO PERFORM A G+C ANALYSIS ON A NUCLEOTICE SEQUENCE  C
C      LOOKS AT EACH NUCLEOTIDE FROM 1ST+50 TO LAST-49             C
C      AND EXAMINES ITS NEAREST 50 NEIGHBORS BELOW AS WELL         C
C      AS ITS NEAREST 49 NEIGHBORS ABOVE                           C
C      THE G+C CONTENT OF THESE 100 NUCLEOTIDES IS RECORDED        C
C      AS WELL AS JUST THE G CONTENT                               C
C      THEN DOES A FRAMESHIFT OF 1 AND REPEATS                     C
C                                                                  C
C      INPUT IS FROM A FILE OF NUCLEOTIDES                         C
C      OUTPUT TO LUN 12 IN A FORMAT SUITABLE FOR INPUT             C
C      TO THE GRAPHIT PROGRAM                                      C
C                                                                  C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

PLIB LIBRARY

```
          PROGRAM     AMINOIDX
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE AMINOIDX(NAMEACID,INDXACID)
          SUBROUTINE TO TRANSLATE A 3 CHARACTER AMINO ACID NAME
          INTO ITS ALPHABETIC INDEX [1-20].
          PARAMETERS ARE-
             NAMEACID- A 3 CHARACTER BCD NAME
             INDXACID- THE INDEX NUMBER
          THE ROUTINE READS THE NAMEACID PARAMETER AND
          PLACES THE APPRNPRIATE INDEX NUMBER INTO
          INDXACID.
          THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL
          THREE CHARACTER CODE IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
          PROGRAM     ATLASIDX
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE ATLASIDX(NAMEACID,INDXACID)
          SUBROUTINE TO TRANSLATE A 1 CHARACTER AMINO ACID NAME
          INTO ITS ALPHABETIC INDEX (1-20).
          PARAMETERS ARE-
             NAMEACID- A 1 CHARACTER BCD NAME
             INDXACID- THE INDEX NUMBER
          THE ROUTINE READS THE NAMEACID PARAMETER AND
          PLACES THE APPROPRIATE INDEX NUMBER INTO
          INDXACID.
          THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL
          ATLAS CODE IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM     ATOMIDX
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE ATOMIDX(NAMEATOM,IDXATOM)
          SUBROUTINE TO TRANSLATE AN ATOM NAME INTO
          A UNIQUE INDEX NUMBER
          PARAMETERS ARE-
             NAMEATOM - A BCD ATOMIC NAME
              IDXATOM - A UNIQUE INDEX NUMBER
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM     CIRCLES
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE CIRCLES(N,NEXT)
          SUBROUTINE TO TAKE THE NEXT ROTATION OF AN N DIGIT NUMBER.
          USED WITH SUBROUTINE PERMUTE TO REDUCE THE PERMUTATIONS TO
          COMBINATIONS.
          PARAMETERS ARE-
             N      - THE NUMBER TO BE ROTATED. IF N=0, TAKE THE
                      NEXT ROTATION ON THE PREVIOUS NUMBER
             NEXT   - SET TO THE ROTATED VALUE OF N

          LIMITATION - WILL NOT WORK WITH N LARGER THAN 2)24-1
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM     CODONIDX
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
        SUBROUTINE CODONIDX(CNAME,INDXACID)
        SUBROUTINE TO TRANSLATE A 3 CHARACTER CODON NAME
        INTO THE INDEX NUMBER OF THE AMINO ACID IT CODES FOR.
        STOP IS CODED AS A 0 INDEX.
        PARAMETERS ARE-
           CNAME- A 3 CHARACTER BCD NAME
           INDXACID- THE INDEX NUMBER
        THE ROUTINE READS THE CNAME PARAMETER AND
        PLACES THE APPROPRIATE INDEX NUMBER INTO
        INDXACID.
        THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL
        CODON IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


        PROGRAM     CONTACT
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
        FUNCTION CONTACT
        DISTANCE=CONTACT(JATOM1,JATOM2)
        GIVES MINIMUM CONTACT DISTANCES IN ANGSTROMS FOR ANY
        TWO OF THE FOLLOWING SET
           H,O,N,C, AND CH=METHYL GROUP

        ATOM NAMES SHOULD BE LEFT JUSTIFIED
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


        PROGRAM     COVRADII
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
        SUBROUTINE COVRADII(IDXATOM,DIST)
        SUBROUTINE TO DETERMINE THE COVALENT RADIUS
        FOR AN ATOM
        PARAMETERS ARE-
           IDXATOM - THE ATOMIC SYMBOL INDEX NUMBER
              DIST - THE COVALENT RADIUS IN ANGSTROMS
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


        PROGRAM     DISTANCE
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
        SUBROUTINE DISTANCE(JCODON1,JCODON2,MEASURE)
        SUBROUTINE TO MEASURE THE EVOLUTIONARY DISTANCE
        BETWEEN 2 ARBITRARY CODONS. MEASURES THE NUMBER
        OF BASE CHANGES BETWEEN THE TWO. SO THE FUNCTION
        YIELDS A VALUE BETWEEN 0 AND 3.
```

```
         PARAMETERS ARE-
            JCODON1    - AN ARBITRARY CODON
            JCODON2    - ANOTHER ARBITRARY CODON
            MEASURE    - THE NUMBER OF BASE DIFFERENCES BETWEEN THEM

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM      IDXAMINO
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE IDXAMINO(INDXACID,NAMEACID)
         SUBROUTINE TO TRANSLATE AN INDEX NUMBER [1-20]
         TO A 3 CHARACTER AMINO ACID NAME.
         PARAMETERS ARE-
            INDXACID- THE INDEX NUMBER
            NAMEACID- A 3 CHARACTER BCD NAME
         THE ROUTINE READS THE INDXACID PARAMETER AND
         PLACES THE APPROPRIATE NAME INTO
         NAMEACID.
         THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL
         INDEX NUMBER IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM      IDXANAME
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE IDXANAME(INDEX,NUM,JARRAY)
         SUBROUTINE TO CONVERT AN ALPHABETIC AMINO ACID INDEX NUMBER
         TO THE LIST OF NON-HYDROGEN ATOM NAMES FOR THAT AMINO ACID.
         THE ORDER OF ATOMS WILL BE CA,N,C,O,CB,...
         ATOM NAME NUMENCLATURE IS TAKEN FROM THE SCHERAGA ARTICLE
            CALCULATIONS OF CONFORMATIONS OF POLYPEPTIDES
            ADV. IN PHY.ORG.CHEM. (1968)

         PARAMETERS ARE-
            INDEX - AN ALPHABETIC INDEX NUMBER
            NUM   - THE NUMBER OF NON-HYDROGEN ATOMS IN THE A.A.
            JARRAY- AN ARRAY IN WHICH NAMES WILL BE STORED

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM      IDXATLAS
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
         SUBROUTINE IDXATLAS(INDXACID,NAMEACID)
         SUBROUTINE TO TRANSLATE AN INDEX NUMBER [1-20]
         TO A 1 CHARACTER AMINO ACID NAME.
         PARAMETERS ARE-
            INDXACID- THE INDEX NUMBER
            NAMEACID- A 1 CHARACTER BCD NAME
         THE ROUTINE READS THE INDXACID PARAMETER AND
         PLACES THE APPROPRIATE NAME INTO
         NAMEACID.
         THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL
         INDEX NUMBER IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM     IDXATOM
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE IDXATOM(INDEX,NAMEATOM)
         SUBROUTINE TO TRANSLATE AN INDEX NUMBER TO
         AN ATOMIC SYMBOL.
         PARAMETERS ARE -
              INDEX - AN ATOM INDEX
           NAMEATOM - AN ATOMIC SYMBOL
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM     IDXAUGON
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE IDXAUGON(INDEX,NUM,JARRAY)
         SUBROUTINE TO TRANSLATE AN ALPHABETIC AMINO ACID INDEX
         NUMBER INTO A LIST OF COLLAPSED CODONS THAT CODE FOR
         THAT ACID.
         INDEX NUMBER 0 IS THE STOP CODE

         COLLAPSED CODONS INCLUDE A,C,G,U, AND
            M=MASTER=(U,A,G,C)
            Y=PYRIMIDINE=(U,C)
            R=PURINE=(A,G)

         PARAMETERS ARE-
            IDX - AN ALPHABETIC INDEX NUMBER
            NUM - THE NUMBER OF CODONS THAT CODE FOR THE A.A.
            JARRAY - AN ARRAY POINTER WHERE THE 3 CHARACTER CODES
                     OF THE CODONS WILL BE PLACED, 1 CODON/WORD,
                     LEFT JUSTIFIED
         THE ROUTINE READS THE INDEX AND PLACES THE APPROPRIATE
         NUMBER OF CODONS INTO NUM. THEN THE 3 CHARACTER CODON
         CODES ARE PLACED INTO EACH SUCCESSIVE ELEMENT OF JARRAY.

         THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL INDEX
         NUMBER IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM     IDXCODON
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE IDXCODON(INDEX,NUM,JARRAY)
         SUBROUTINE TO TRANSLATE AN ALPHABETIC AMINO ACID INDEX
         NUMBER [0-20] INTO A LIST OF CODONS THAT CODE FOR THAT ACID.
         INDEX NUMBER 0 IS THE STOP CODE.
         PARAMETERS ARE-
            IDX- AN ALPHABETIC INDEX NUMBER
            NUM- THE NUMBER OF CODONS THAT CODE FOR THE AMINO ACID
            JARRAY- AN ARRAY POINTER WHERE THE 3 CHARACTER CODES OF TH
                    CODONS WILL BE PLACED, 1 CODON/WORD, LEFT JUSTIFIE
         THE ROUTINE READS THE INDEX AND PLACES THE APPROPRIATE
         NUMBER OF CODONS INTO NUM. THEN THE 3 CHARACTER CODON CODES
         ARE PLACED INTO EACH SUCCESSIVE ELEMENT OF JARRAY.
         THE ROUTINE COMES TO AN ERROR HALT IF AN ILLEGAL INDEX NUMBER
         IS ENCOUNTERED.
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM     INVCODON
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE INVCODON(J1,J2)
         PARAMETERS ARE-
            J1 - A LEFT JUSTIFIED BCD 3 CHARACTER CODON
            J2 - ANTICODON FOR J1
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE JEFFREYS(A,B,POINT,THETA)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     OPERATOR TO ROTATE A POINT(X,Y,Z) BY THETA DEGREES
C     ABOUT A LINE FROM A TO B
C
C     A IS CONSTRUED TO BE THE ORIGIN
C     B IS A POINT ON THE AXIS LINE THROUGH THE ORIGIN
```

```
C      POINT IS COORDINATES OF THE POINT TO BE ROTATED
C      THETA IS THE NUMBER OF DEGREES OF ROTATION
C
C      IF A=B=0, THEN OLD VALUES ARE ASSUMED FOR THE ORIGIN AND
C      ROTATION MATRIX. OTHERWISE, THE ORIGIN AND ROTATION MATRIX
C      ARE RECOMPUTED.
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM      PAIRSMAX
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE PAIRSMAX(J1,J2,KSCORE)
          SUBROUTINE TO ASSIGN A SCORE TO A NUCLEOTIDE PAIR
          PARAMETERS ARE-
               J1 - NUCLEOTIDE 1 (RIGHT JUSTIFIED)
               J2 - NUCLEOTIDE 2 (RIGHT JUSTIFIED)
           KSCORE - THE SCORE        GC=2,AU=1,ANYTHING ELSE=0
          NUCLEOTIDES MUST BELONG TO THE SET
             G,C,U,A
             M=MASTER SET = (G,C,U,A)
             Y=PYRIMIDINE = (U,C)
             R=PURINE     = (A,G)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM      PAIRSMIN
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE PAIRSMIN(J1,J2,KSCORE)
          SUBROUTINE TO ASSIGN A SCORE TO A NUCLEOTIDE PAIR
          PARAMETERS ARE-
               J1 - NUCLEOTIDE 1 (RIGHT JUSTIFIED)
               J2 - NUCLEOTIDE 2 (RIGHT JUSTIFIED)
           KSCORE - THE SCORE        GC=2,AU=1,ANYTHING ELSE=0
          NUCLEOTIDES MUST BELONG TO THE SET
             G,C,U,A
             M=MASTER SET = (G,C,U,A)
             Y=PYRIMIDINE = (U,C)
             R=PURINE     = (A,G)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM      PERMUTE
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE PERMUTE(N,NEXTPERM)
          SUBROUTINE TO FIND ALL POSSIBLE PERMUTATIONS OF N OBJECTS
          PARAMETERS ARE
             N       - THE NUMBER OF OBJECTS. IF N=0, PERMUTE DELIVERS
                       THE NEXT PERMUTATION OVER THE LAST NON-ZERO N.
             NEXTPERM- THE NEXT PERMUTATION OF N OBJECTS, EXPRESSED AS
                       A BASE N NUMBER. IF ALL PERMUTATIONS HAVE BEEN
                       EXHAUSTED, NEXTPERMO=0

          LIMITATION- N CANNOT EXCEED 9 BECAUSE OF THE WAY NEXTPERM
                      IS EXPRESSED (I.E. AS A DIGIT). OF COURSE
                      IN PRACTICE, NEXTPERM CANNOT EXCEED 2¬24-1

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM     REVCODON
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          SUBROUTINE REVCODON(J1,J2)
          PARAMETERS ARE-
             J1 - A LEFT JUSTIFIED BCD 3 CHARACTER CODON
             J2 - REVERSE CODON FOR J1
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM     ASTRING,JSTRING,DSTRING
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
          FUNCTION ASTRING(BUFF,JSTART,LNGBUFF)
               OR
          FUNCTION DSTRING(BUFF,JSTART,LNGBUFF)

          FUNCTION TO SCAN OFF AN ALPHANUMERIC SYMBOL OR A NUMBER
          AND RETURN THE RESULTS IN AQ.

          PARAMETERS ARE-
             BUFF-THE FWA OF THE BUFFER IN WHICH THE STRING IS FOUND
             JSTART-A CHARACTER POSITION POINTER ON THAT BUFFER
             LNGBUFF-CHARACTER LENGTH OF THE BUFFER

          UPON EXIT, THE SCANNED ITEM WILL BE IN AQ, AND JSTART WILL BE
          UPDATED TO THE LAST CHARACTER SCANNED+1

          THIS ROUTINE WILL NOT SCAN BEYOND LNGBUFF. BLANKS AND SPECIAL
          CHARACTERS ARE IGNORED. AN ALPHA STRING IS TERMINATED
          AFTER 8 CHARACTERS ARE ACCRUED OR WHEN A NON-ALPHANUMERIC
          IS ENCOUNTERED.
```

```
         MEANT TO BE USED IN THIS FASHION-

            BUFFER IN(N,0) (BUFF(1),BUFF(LNGBUFF))
                OR
            READ(N,100) BUFF
     100    FORMAT(100A4)

            JSTART=1
            SYMBOL=ASTRING(BUFF,JSTART,LNGBUFF)
            IF(JSTART.GT.LNGBUFF) GOTO ERROR
         C  OTHERWISE USE THE SYMBOL AND JSTART WILL BE
         C  AUTOMATICALLY UPDATED

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


         PROGRAM      STAT
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE STAT(LUN,JTYPE,JSTAT)
         PROGRAM TO DETERMINE THE HARDWARE TYPE AND CURRENT
         STATUS OF A LOGICAL UNIT.

         PARAMETERS ARE-
            LUN - THE LOGICAL UNIT IN QUESTION
          JTYPE - THE HARDWARE TYPE OF LUN
          JSTAT - THE 9 BIT STATUS OF THE LUN

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      FUNCTION TORSION(A,B,C,D)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     FUNCTION TO COMPUTE THE TORSION ANGLE BETWEEN 4 ATOMS
C     A,B,C, AND D
C     COPLANAR CIS CONFIGURATION OF A AND D IS TAKEN AS ZERO
C     2ND HANDBOOK OF BIOCHEM CONVENTIONS ON TORSION ANGLES
C     ARE USED
C
C     THE TORSION ANGLE IS IN DEGREES
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
      SUBROUTINE ANGLE(X,Y,THETA,D1,D2,CTH)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO FIND THE ANGLE, THETA, BETWEEN VECTORS
C     X AND Y. ASSUME BOTH X AND Y HAVE TAILS AT THE ORIGIN
C
C     ALSO RETURNS D1,  THE LENGTH OF X  AND
C                  D2,  THE LENGTH OF Y
C     AS WELL AS  CTH,  THE COSINE OF THETA
C     ANGLE GIVES THE VECTOR DOT PRODUCT OF X AND Y
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE NORMAL(A,B,P)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO FIND THE NORMAL, P, BETWEEN TWO
C     VECTORS, A AND B. A AND B ARE ASSUMED TO HAVE
C     TAILS AT ORIGIN
C     P WILL BE A UNIT VECTOR ORIENTED SUCH THAT A RIGHT HANDED SCREW
C     DRIVEN IN THE DIRECTION OF P WILL CARRY A INTO B
C     NORMAL GIVES THE VECTOR CROSS PRODUCT, AXB
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE FINDROT1(VEC,THETA,PHI)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO FIND THE ANGLES (THETA,PHI)
C     NECESSARY TO ROTATE AN ARBITRARY VECTOR, VEC,
C     SO THAT IT IS PARALLEL TO THE X-AXIS  AND POINTING
C     IN THE POSITIVE DIRECTION
C        THETA - DISPLACEMENT FROM THE XZ PLANE
C                WHEN ROTATING AROUND Z AXIS
C        PHI   - DISPLACEMENT FROM THE X AXIS
C                WITHIN THE XZ PLANE
C                WHEN ROTATING AROUND Y AXIS
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE FINDROT2(VEC,THETA,PHI)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

```
C
C     SUBROUTINE TO FIND THE ANGLES (THETA,PHI)
C     NECESSARY TO ROTATE AN ARBITRARY VECTOR, VEC,
C     SO THAT IT IS PARALLEL TO THE X-AXIS  AND POINTING
C     IN THE POSITIVE DIRECTION
C        THETA  - DISPLACEMENT FROM THE YZ PLANE
C                 WHEN ROTATING AROUND X AXIS
C        PHI    - DISPLACEMENT FROM THE X AXIS
C                 WITHIN THE XZ PLANE
C                 WHEN ROTATING AROUND Y AXIS
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE ROTXY(THETA,V)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO PERFORM A ROTATION IN THE XY PLANE
C     ROTATION WILL BE IN THE CLOCKWISE DIRECTION VIEWED
C     FROM THE (+) Z-AXIS
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE ROTYZ(THETA,V)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO PERFORM A ROTATION IN THE YZ PLANE
C     ROTATION WILL BE IN THE CLOCKWISE DIRECTION
C     VIEWED FROM THE (+) X-AXIS
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


      SUBROUTINE ROTXZ(THETA,V)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     SUBROUTINE TO PERFORM A ROTATION IN THE XZ PLANE
C     ROTATION WILL BE IN THE CLOCKWISE DIRECTION
C     VIEWED FROM THE (+) Y-AXIS
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


          PROGRAM      VDWRADII
```

```
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
         SUBROUTINE VDWRADII(IDXATOM,DIST)
         SUBROUTINE TO DETERMINE THE MIMINUM VAN DER WAALS
         CONTACT DISTANCE FOR AN ATOM.
         PARAMETERS ARE -
            IDXATOM - THE ATOMIC SYMBOL INDEX NUMBER
               DIST - VAN DER WAALS RADIUS IN ANGSTROMS
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```