A CURRENT LOOK AT THE BIOLOGICAL BASIS OF ANTIBODY DIVERSITY AND SPECIFICITY

by

Paul Arlyn Barstad

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1975

(Submitted April 18, 1975)

This thesis is dedicated to my wife Glenna and my daughter Debbie who have tolerated my long hours in the lab and helped so much with the preparation of this thesis, and to the Tiger Toads.

ACKNOWLEDGMENTS

I would like to thank all the people in Lee Hood's lab group for their helpful discussions on scientific as well as non-scientific subjects. Special thanks must go to Barbara Eaton for her cheerfulness and superb technical assistance, to Lee Hood for his guidance, to Jim Black for the many favors, and to Ray Owen for his warm friendship and encouragement.

ABSTRACT

Using automated protein sequence analysis of BALB/c myeloma proteins, the genetic basis for antibody specificity and diversity is investigated. Studies on the N-terminal regions of the heavy chains from these immunoglobulins reveal that a large amount of diversity must exist in the V_H regions. Examination of the heavy and light chain sequences from myeloma proteins with hapten-binding activities indicates that the heavy chain variable region sequence correlates closely with all the specificities of intact molecules. The light chain appears to be less restricted in some specificities, however. The relevance of these data to the proposed mechanisms of antibody diversity is discussed.

TABLE OF CONTENTS

CHAPTER I

THE PROBLEM

THE PROBLEM

During evolution, about the time of the emergence of vertebrates, nature began to develop an active defense mechanism for higher creatures to protect themselves from bacteria and viruses as well as perhaps from abnormal or transformed cells generated within their own tissues. Such a system had three requirements: (1) a high degree of specificity to distinguish self from non-self or normal from abnormal cells; (2) extensive diversity to recognize a vast number of different foreign substances; and (3) memory, so that a second encounter with the same foreign material triggers a rapid and effective defensive response.

Although many of the functions of the immune system involve complex cellular interactions, one part called the humoral immune system utilizes protein molecules, called antibodies, which circulate and recognize the foreign materials, called antigens, and cause their elimination. These proteins must combine two functions--one part of the molecule must recognize and bind the antigen while another part must activate a mechanism which may lyse or opsonize the cell, if it is a bacterium, or destroy other antigens in other ways. Mammals possess different classes of antibodies whose effector and stability properties are specialized for their particular location in the

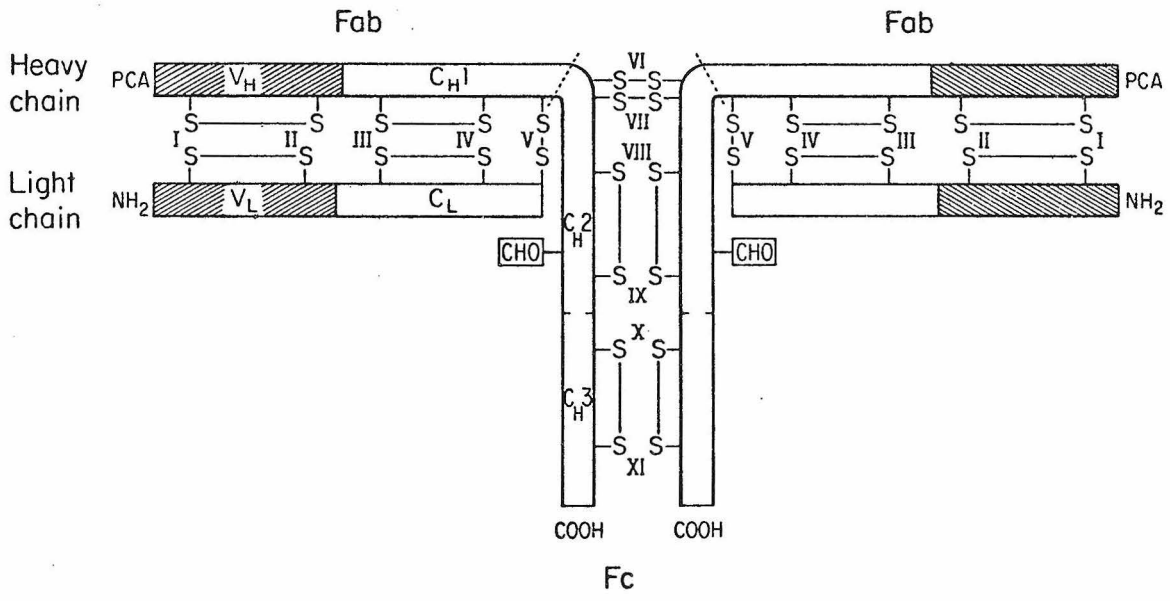
body. Thus, immunoglobulin G (IgG) is found as the dominant circulating antibody, IgM makes up the largest part of antibody produced in a primary response, IgA is specialized to be stable in the gut and secretory system of the respiratory tract, and IgE attaches to and triggers mast cells to release histamine.

The basic subunit of antibodies can be divided into functional fragments. The basic functional subunit of the antibody molecule is shown in Fig. 1-1. It has a molecular weight of about 150,000 daltons and has a sedimentation coefficient of 7S. Enzymatic digestion of this molecule has shown it to consist of three regions, each with a molecular weight of 50,000 daltons--two of which (called the Fab fragments) are identical and bind the antigens, and a third fragment named the Fc that performs at least some of the effector functions of the intact molecule.

Two types of polypeptide chains make up the basic functional antibody subunit. Two pairs of identical polypeptide chains make up the 7S subunit--two "light" chains with a molecular weight of 23,000 to 25,000 daltons and two heavy chains ranging in size from 50,000 to 70,000, depending on the class of the antibody. The amino terminal half of one heavy chain combines with a light chain to form each Fab fragment while the two carboxyl halves of the heavy chains form the Fc region.

Each polypeptide chain of an immunoglobulin has a variable portion and a constant portion. The normal antibody response to an antigen or antigens consists of a very heterogeneous

Fig. 1-1.--Structure of immunoglobulin 7S subunit. The IgG molecule shown here consists of a single 7S subunit. The locations of the various interchain and intrachain disulfide bonds are indicated. The dotted lines show the enzymatic cleavage points to form the Fab and Fc fragments. The shaded areas represent the variable regions of each chain. CHO indicates the location of carbohydrate on the heavy chain. After (1).


Fab
Fab
Heavy chain
Light chain
PCA
PCA
V_H
C_H1
V_L
C_L
NH_2
NH_2
CHO
CHO
C_H2
C_H3
COOH
COOH
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
Fc

collection of antibodies. The vast majority of sequence variation within the light and heavy chains occurs in the amino terminal 110 residues or so as shown by sequence studies on various homogeneous immunoglobulins from human and mouse as well as the polypeptide subunits from heterogeneous circulating IgG. These regions are called the variable regions or V_H and V_L of the heavy and light chains respectively. It is these regions that determine the antigen-binding specificity of the antibody and are the subject of this thesis.

The remainder of the heavy chain is highly similar or identical in sequence from one antibody to another within a subclass of antibody in a single creature. The carboxyl half of the light chain is also constant in sequence within a light chain family. These two regions are called the constant regions for the heavy (C_H) and light (C_L) chains. The V and C regions of the heavy or light chains are encoded by separate genes and thus the concept of "two genes--one polypeptide chain" (2).

Three families of immunoglobulin subunits exist and appear to be unlinked in the mammalian genome. Two families of light chains exist in every mammalian immune system--kappa and lambda chains. These two types of chains differ greatly in the sequence in both the constant and variable regions. A third family codes for the heavy chains. Genetic analysis of serologic markers on the constant regions of these chains indicates that they are unlinked in the mammalian genome (3).

The variable region can be divided into hypervariable and framework residues. The three-dimensional structure of an immunoglobulin Fab fragment has been shown by X-ray crystallography to consist of four nearly identical domains--two V and two C domains (4). The distal end of the V_H and V_L regions forms a cleft which binds the antigen. The light chain half of the cleft is formed by residues 26_L-32_L and 90_L-96_L, while the heavy chain half is formed by 31_H-37_H, 51_H-58_H, and 101_H-110_H. Within a series of immunoglobulins, these regions have been found to be significantly more variable in sequence and size than the rest of the V regions. One additional region in each chain, 48_L-55_L and 84_H-91_H, is also highly variable in sequence (5,6). Accordingly, these seven areas have been termed the hypervariable regions. The remainder of the variable region is called the framework region.

Each antibody-secreting cell synthesizes only one kind of antibody. The cell that secretes antibody molecules is called the plasma cell or B cell. Each plasma cell appears to make only one type of light and heavy chain. The variable regions of the light chains made by a single cell are all identical as are those of the heavy chain. The cell also produces either lambda or kappa chains, but not both. It is generally accepted now that multiple V genes exist in a creature and the mechanism for activation and expression of one of a number of variable region genes is of interest from a developmental standpoint as well as from an immunological one.

The number of different antibodies that can be made by a single creature must be very large. The more specific the interactions are between an antibody and antigen, the more diverse the antibody library must be in order for the creature to protect itself against a broad spectrum of potential pathogens. Thus, any factors concerning either diversity or specificity of antibodies yield information about the other component.

Since antibodies can distinguish self from non-self in many instances, they must be rather specific, analogous perhaps, to certain enzymes. The total antibody library must also recognize a large number of different antigens in order to keep the creature alive. The number of different antibodies that can be made by a single creature must then be very large and 10^5 different specificities may be necessary and sufficient to keep a creature healthy. This is merely an estimate and certainly the actual number may be an order of magnitude above or below this figure.

The amount of functional antibody diversity depends on four factors. In order to approach estimating the total functional diversity and specificity present within a creature, one can divide the problem into its two components: (1) how many different functional antibody molecules can an animal make? and (2) what are the contributions of the heavy and light chains in determining the antigen-binding specificity of the molecule? The number of functional antibody molecules can in turn be divided into three separate questions. (a) how

diverse are the light chains? (b) how diverse are the heavy chains? and (c) can any heavy chain pair with any light chain (and vice versa) to form functional antibody molecules? While a great deal of data have been obtained concerning the total light chain diversity, the other three questions have been relatively unanswered.

Four theories have been advanced to explain the diversity of antibody subunits. Since the three-dimensional structure of the antigen-binding site determines the specificity of the antibody, and since the primary structure of the variable region determines the tertiary structure, the question of the biological basis for information storage and expression for antibody subunits has been a subject of interest for several years. Two major theories have been advanced--the somatic theory and the germ line theory--to explain the diversity of primary sequences in antibodies.

The somatic theory (7) postulates that a limited number of genes for antibody variable regions are passed on in the germ line from parent to progeny. By this theory, the diversity of antibodies is generated by mutation in the soma during development. In this theory, antibody genes are coded by a small amount of germinal DNA. However, almost limitless diversity can be generated. It would also predict a large amount of random sequence flexibility within antibodies that bind similar antigens.

The germ line theory (8) postulates that all antibody genes are present in the germ line. This theory requires a large

number of variable region genes. Selection for specificity and diversity then occurs during the evolution of the species rather than during ontogeny. Subunit diversity is generated by expression of all the hundreds or thousands of variable region genes for each family.

The two other theories that have recently arisen involve modifications of germ line and somatic theories. Essentially, these two theories maintain that the hypervariable regions are under separate control for diversity from framework residues. Both of these hypotheses assume that a large (at least 200) number of genes code for the framework residues.

The antigen-driven somatic mutation theory (9) maintains that the basic antibody diversity is encoded in the germ line but that during expansion of the clone following antigenic stimulation some clones will accumulate random mutations in the hypervariable regions which give them a higher affinity for the original antigen. These mutant clones are then expanded to a greater extent than the corresponding clones with the original antibody. While mutations in these regions probably occur, most of the mutations must be either neutral or harmful and it is difficult to conceive how such a mechanism could have much advantage in vivo at the currently accepted rate of random mutation.

The second theory in this category is called the episomal theory and claims that the hypervariable regions are actually separate genes that are spliced into the framework genes (10).

Several different combinations of hypervariable genes could be spliced into the same framework gene generating a tremendous diversity from only a small amount of DNA. This theory requires "five genes--one polypeptide chain" and my data clearly dispute its existence.

This thesis investigates the heavy chain diversity and genetic basis of immunoglobulin specificities.

These four theories of subunit diversity were based primarily on data concerning the BALB/c light chains. In this thesis, investigations on the diversity present within the heavy chain subunit are reported. In addition, I have evaluated the relative importance of the heavy and light chain amino acid sequences in determining the function (i.e., hapten-binding activity) of an immunoglobulin. This work renders the episomal theory unlikely and strongly supports the concept of a germ line basis for antibody diversity.

CHAPTER II

THE SYSTEM

THE SYSTEM

Ideally, to approach the questions concerning antibody diversity one would like to study a series of antibodies made in a single individual in response to a single antigen. In practice, however, such antibodies can normally only be studied by micro methods such as isoelectric focusing since there is too little of any antibody species for sequence analysis. While such a study can give insight into the heterogeneity of an immune response, (11) it cannot give any data on the genetic relationship of the various molecules or their subunits.

Relatively restricted antibodies may be induced using special immunization techniques. Two examples of relatively large amounts of homogeneous antibodies exist. The first is the production of large quantities (up to 50 mg/ml of serum) of antibodies of restricted heterogeneity (apparently two or three homogeneous antibodies) to bacterial polysaccharides in rabbits by repeated immunization (12). Since rabbits are not inbred, however, the genetic relationships of these antibodies can only be meaningfully evaluated against the other few antibodies available in large amounts from that individual animal.

Recently it has been possible to raise large amounts of restricted antibodies in the BALB/c mouse to the hapten

p-azophenylarsonate by hyperimmunization with this antigen coupled to a carrier followed by induction of an ascites tumor (13). The ascites tumor is a free-growing tumor in the peritoneal cavity and results in the accumulation of 15 ml or more of fluid in that area which contains the immunoglobulins in concentrations similar to that found in the serum. This approach allows one to do a series of experiments within a single genome (BALB/c) but also has a couple of disadvantages.

First, since each mouse may respond differently to the antigen, the antibodies from many mice should not be pooled or one will only be detecting the major sequences of the response. Second, the antibodies made within a single mouse are usually heterogeneous as judged by isoelectric focusing. While such antibodies can be purified by preparative isoelectric focusing, many minor bands are present in too small amounts to analyze, and thus the total diversity within a specificity cannot easily be determined.

Multiple myeloma offers virtually unlimited amounts of homogeneous proteins in the BALB/c mouse. A type of cancer, known as multiple myeloma, results in the transformation and subsequent proliferation of a single plasma cell until its homogeneous immunoglobulins make up to 95% of that found in the circulation.

These plasmacytomas may be induced in the inbred strain of mouse, BALB/c (14,15). Such an induction program involves injection of 1 ml of mineral oil or pristane intraperitoneally

at 60-day intervals. Within six months, over 60% of the BALB/c mice develop myeloma tumors which are then transplantable to other syngeneic animals.

These proteins are called immunoglobulins rather than antibodies since they were not produced in response to stimulation by antigen. The mineral oil apparently provides a good environment for the transformation and subsequent proliferation of plasma cells in the gut. The transformation event seems to occur in any random antibody-secreting cell in the gut area since sequence analysis on these proteins has shown them to be very diverse.

The tumors are usually named by four letters followed by a number. The letters reflect the type of inducing agent, as in Mineral Oil Plasmacytoma (MOPC), or person or place of induction, e.g., Salk = S and Weigert = W. The number of the tumor is assigned by the originator.

The myeloma proteins are an excellent system for studying immunoglobulin structure. Myeloma immunoglobulins are the products of aberrant cells--tumors that are not responding to antigen but simply growing, dividing, and secreting a protein product. Early studies indicated the myeloma proteins in humans did not have any detectable antigen-binding activities and it was questioned how applicable studies on these immunoglobulins were to normal antibodies. Recent work, however, has shown the BALB/c myeloma proteins to share the same structures and binding characteristics as normal antibodies.

The human and BALB/c myeloma immunoglobulins are different in many respects. Much of the early work on immunoglobulin structure was done on myeloma proteins from human patients. In humans, multiple myeloma appears spontaneously in patients usually about 50 years of age and progresses for a few years resulting in the death of the patient. Large amounts of homogeneous protein may be obtained from the individual during the course of the disease. Human myeloma proteins exhibit the same kappa and lambda ratio as the circulating IgG in normal humans. These proteins also appear to possess all the heavy chain constant regions (which determine the class of the antibodies) and, indeed, were instrumental in the detection of some of the rare classes and subclasses of normal human antibodies. While these proteins share structure in the constant regions with normal immunoglobulins, very few of these proteins have been found to bind specifically to any hapten.

The human population is extremely polymorphic and variation in the structure of the human myeloma proteins from one patient to another may be due to either differences between several genes within the same individual or to a polymorphism between the individuals. It is not possible to evaluate which of these two factors is most important in the observed human immunoglobulin variability.

The BALB/c mouse is a highly inbred creature, on the other hand. Animals of this strain have been inbred to the

extent that all BALB/c mice must have the same genome except for a few spontaneous mutations which may have occurred recently. If 100 myeloma tumors are induced in 100 different BALB/c mice, then we are looking at 100 separate immunoglobulins which are products of the same genome. It is analagous to being able to isolate 100 different homogeneous antibodies from the same individual. Since these mice are highly inbred, a given tumor may be transplanted to many other BALB/c mice where it will continue to grow and produce virtually unlimited amounts of the original homogeneous immunoglobulin.

The BALB/c myeloma proteins, like their human counterparts, possess many of the same constant regions as the normal antibodies of the animal. Both kappa and lambda light chains have been found. The BALB/c myeloma proteins include IgM, IgA, IgG-1, IgG-2a, IgG-2b, and IgG-3. No IgD or IgE has been reported.

Unlike their human counterparts, however, the BALB/c proteins often show a relatively high specific binding to various haptens, particularly antigens present on gut flora and dinitrophenyl proteins. Binding constants for these haptens range from 10^5 to 10^7 M^{-1}. The BALB/c myeloma proteins, therefore, appear to represent the products of functional antibody-producing plasma cells which have been expanded by transformation.

The BALB/c myeloma proteins also differ from the human myeloma proteins in that while most of the human proteins are

IgGs, the vast majority of those in the BALB/c are IgAs. Since most of the human myelomas are found in the bone marrow (where mostly IgG is produced) and the BALB/c tumors arise in the peritoneal cavity (where the majority of immunoglobulins are IgAs) this is not surprising.

Many of the BALB/c myeloma proteins exhibit hapten-binding affinities analogous to antibodies. One criticism of the study of myeloma proteins as a model for antibodies was that they lacked a well-defined hapten-binding specificity. The discovery that MOPC 315 had a binding constant for dinitrophenyl of 10^7 M^{-1} demonstrated that myeloma proteins could have a high affinity for a hapten(16). Extensive screening programs were conducted and it was found that 5% to 10% of the hundreds of tumors examined bound to a few common antigens including dinitrophenyl, phosphorylcholine, and various carbohydrate moities. Since these tumors arise in the peritoneal cavity, it is reassuring that many of these haptens are found on gut flora indicating the gut associated lymphocytes have probably been clonally expanded by these antigens prior to transformation.

Several different carbohydrate specificities have been found. Six myeloma proteins specific for β(1→6)-D-galactan have been characterized by N-terminal sequence analysis. Both α(1→3) and α(1→6) dextran-binding BALB/c myeloma immunoglobulins have been studied in regard to the size of their binding site and serologic properties. Two levan specificities, β(2→1) and β(2→6), have been studied by sequence analysis also.

Eleven BALB/c myeloma proteins have been found which bind phosphorylcholine. This binding specificity was characterized by screening large numbers of myeloma proteins against gut bacteria antigens. These 11 proteins form a precipitin band against Pneumococcal c polysaccharide that can be inhibited by adding free phosphorylcholine to the system. The binding constants determined by equilibrium dialysis for phosphorylcholine range from 10^4 to 10^5 M^{-1} for the five tumors characterized.

Serological studies have shown many of these phosphorylcholine-binding tumors to be highly similar, if not identical, in their variable region structure to the normal phosphorylcholine-binding antibodies made in BALB/c. Antisera were made to several of the phosphorylcholine immunoglobulins by injecting them into A/He strain mice. After appropriate absorptions to remove non-specific antibodies, the antiserum made against each myeloma protein was assayed against over 100 other BALB/c immunoglobulins. The specific determinants recognized in this manner are called "idiotypic" determinants (from the Greek work "idios" meaning "one's own"). Since all these proteins are IgAs and have kappa chains, they all contain the same constant regions and thus the idiotypic determinants must be located on the V_H or V_L or on a combination of the two V regions. When the heavy and light chains are separated and checked for the idiotype of the intact molecule, neither subunit has much activity indicating that a large portion of the idiotypic

determinants require heavy and light chain interaction (17). The degree of specificity of the antiidiotypic antisera is often a problem. Some of the determinants recognized by the antisera may be shared by other heavy and light chain pairs. If, however, two proteins are able to compete equally well against an antiserum to one of them and vice versa, the two molecules must have all the idiotypic determinants in common, suggesting a great deal of sequence similarity, if not identity.

Seven of the phosphorylcholine-binding myeloma proteins have been found to be idiotypically identical (15,17). These proteins, T15, H8, 299, S63, S107, Y5170, and Y5236 are said to have the T15 idiotype. Sher and Cohn (18) have shown that the majority of the antibody to phosphorylcholine made in the BALB/c mouse also has the T15 or S107 idiotype. This suggests structural similarity or identity between the variable regions found in BALB/c myeloma proteins and antibodies.

The α(1→3) dextran-binding antibodies show similarity in idiotype and sequence for myeloma proteins. A second idiotype, one in which the myeloma proteins bind α(1→3) dextran, has been studied in the BALB/c mouse (19). Antiidiotypic antisera made to these myeloma proteins reacts with idiotypic identity with the normal α(1→3) dextran antibodies produced in the BALB/c. This α(1→3) dextran antibody can be produced in quantities sufficient for sequence analysis. The myeloma heavy chains from molecules which bind α(1→3) dextran have a very unique sequence,

vastly different from the normal pooled sequence(20) (Fig. 2-1). Although the protein was somewhat insoluble in various solvents (including TFA), I was able to do a preliminary sequence analysis of the pooled heavy chains from BALB/c antibodies raised to α(1→3) dextran on some material supplied by William Geckler at the Salk Institute. This sequence had the same unusual residues at all the detectable positions as J558, a myeloma protein which binds this polysaccharide, implying that at least some myeloma and antibody pools share the same variable region sequences (Fig. 2-1).

Myeloma proteins may be better for studying antibody diversity than bona fide antibodies. In addition, we have sequenced nearly the total variable regions from a series of heavy chains from molecules that bind phosphorylcholine. These chains exhibit limited heterogeneity. The nature of the variation is such that if we pooled all seven chains together, however, it would be very easy to mistake it for a single sequence with present day technology. Pooled sequences in general cannot be as absolute as claimed by some and indeed, it seems that the myeloma system is, in fact, better than normal antibodies for examining antibody diversity.

The myeloma immunoglobulins probably do not represent the total antibody (or immunoglobulin) diversity of the BALB/c mouse. This immunoglobulin population likely represents clones which have been preferentially expanded out by gut antigens prior to transformation. Undoubtedly, we are looking at only a subset of total diversity of antibodies. It appears to be a very

Fig. 2-1.--N-terminal sequences of BALB/c heavy chains from antibodies to α(1→3) dextran, pooled serum antibodies, and α(1→3) dextran-binding myeloma protein J558. Single letter abbreviations are: A, alanine; B, asparagine or aspartic acid (asx); C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine, R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; Z, glutamine or glutamic acid (glx). Boxes indicate residues shared between J558 and antibody to α(1→3) dextran that are not found in the normal pool. Data from this thesis.

Position	1				5					10					15				
J558	E	V	Q	L	Q	E	S	G	P	E	L	V	K	P	G	A	S	V	
α1,3 Dextran	Z	V	Z	L	Z	Z	S	G	P	Z	L	V	?	P	G	A	S	V	
Normal Pool V_H	E	V	K	L	L	E	S	G	G	G	L	V	Q	P	G	G	S	L	

diverse subset, but it is not clear how much more diverse the total immunoglobulin variability may be.

CHAPTER III

THE TOOLS

THE TOOLS

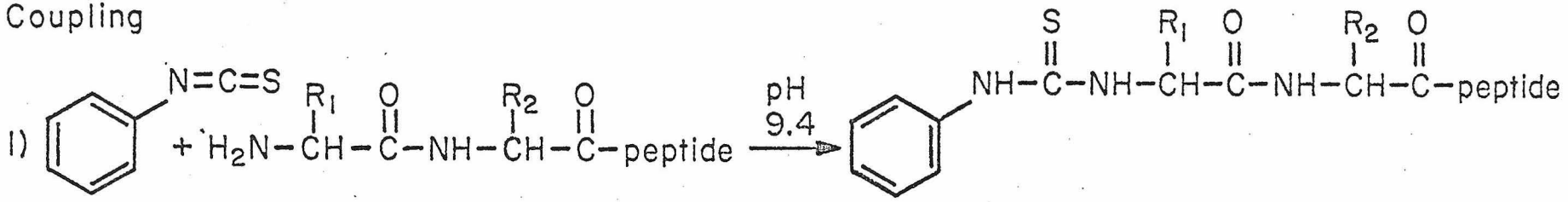
The vast majority of the data in this thesis were generated by four machines in our laboratory. Two of these machines are Beckman 890A sequencers which were updated two years ago to resemble the 890C that Beckman still markets. The third machine is a Hewlett-Packard gas chromatograph (GC) Model 7600 which is used to identify about half of the amino acids. Finally, we have a Durrum D-500 amino acid analyzer which analyzes 10 nanomoles or less of amino acids and quantitates the hydrophilic amino acids (which the GC does not) as well as the hydrophobic ones. Since the major portion of my time was spent in combat with these semi-reliable creations, I feel it only just to describe their salient features so that a reader unfamiliar with sequencing can more readily understand, interpret, and appreciate the data.

Edman reaction. The heart (or hearts) of the laboratory is the Beckman protein/peptide sequenator. This device employs essentially the same chemistry as that developed by Pehr Edman 20 years ago (21) appropriately termed the Edman reaction. (Fig. 3-1) The reaction degrades polypeptides from their N-terminus one step at a time. By analyzing which amino acid derivative is generated at each cycle of the degradation one can unequivocally establish the primary structure of the polypeptide.

Fig. 3-1.--Chemistry of the Edman degradation.

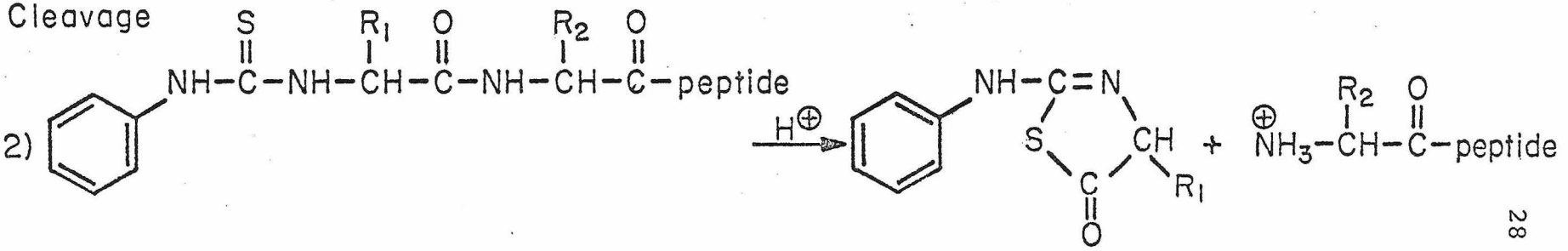
See text for details.

Coupling



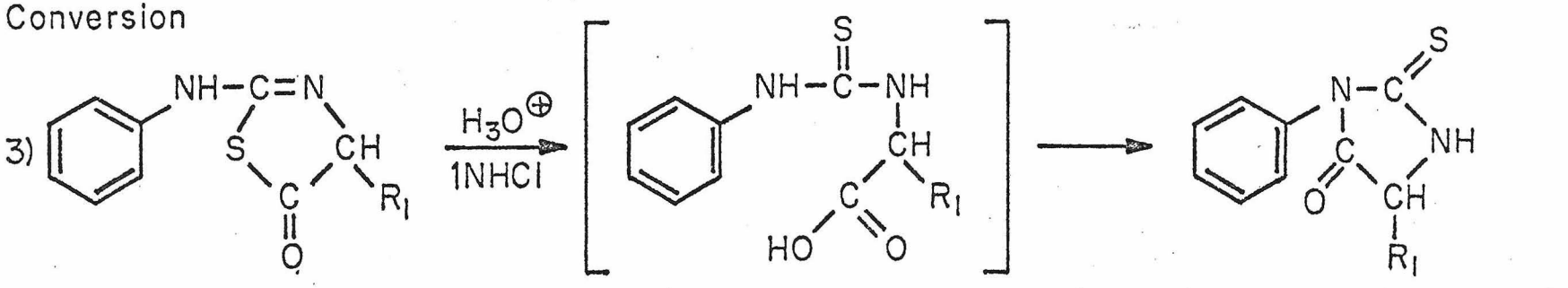
phenylisothiocyanate peptide phenythiocarbamyl peptide

Cleavage



amino acid-2-anilino-5-thiazolinone

Conversion



amino acid-phenylthiohydantoin

The Edman reaction begins by coupling phenylisothiocyanate (PITC) to the free N-terminus of the polypeptide (Step 1). This reaction is carried out at pH 9 or higher as it requires the unprotonated form of the amine group to form the phenylthiocarbamyl (PTC) derivative.

Cleavage of the PTC-peptide at low pH generates the 2-anilino-5-thiazolinone derivative of the first amino acid and a peptide one amino acid shorter than the original (Step 2). Anhydrous trifluoroacetic acid (TFA) and heptafluorobutyric acid (HFBA) are the most commonly used reagents for this cleavage reaction.

The final step is carried out manually and converts the 2-anilino-5-thiazolinone derivative to a more stable amino acid phenylthiohydantoin (PTH) (Step 3). This is normally accomplished at 80° in aqueous HCl. Analysis of PTH derivatives by physical-chemical means allows one to determine what the original amino acid was at the N-terminus.

Beckman sequenator. A sequenator does not give you a sequence, it merely takes the protein apart one step at a time from the N-terminus and delivers the substituted amino acids present at each step into a fraction collector tube for analysis by the other instruments and thin-layer chromatography. The instrument employs a precision ground cup which spins at 600 or 1200 rpm on whose inner walls the protein is spread from solution in a thin layer. The sequencing procedure consists of washing chemicals over the protein surface in such an order as to perform the Edman degradation.

The delivery of chemicals and application of vacuum and nitrogen to the reaction cup are controlled by a programmer which reads a paper tape that has been appropriately punched by the operator. A sequencer program consists of 35 or more steps during which different valves are activated for a period of time as designated by the tape.

The valves control the delivery of the reagents and solvents used by the 890 sequenator. Three reagents, PITC (R1), coupling buffer (R2), and HFBA (R3) and two solvents, benzene (S1) and butyl chloride (S3) are used for each cycle. Since any impurities in the chemicals interfere with the degradation, all solvents and reagents must be extremely free of contaminants; thus only sequencer grade chemicals are used.

The main parts of the reaction program are: (1) addition of 5% PITC in heptane and drying off heptane (R1); (2) addition of coupling buffer pH 9.5 (R2) followed by a 30-minute reaction time during which the protein dissolves in the R2 and the PITC couples to the N-terminus of the polypeptide; (3) after drying off the R2, benzene (S1) is washed over the sample to remove the reaction byproducts and non-volatile components of the buffer. Vacuum is applied to the reaction cell to remove the S1. The phenylthiocarbamyl amino acid is then cyclized in heptafluorbutyric acid (HFBA) (R3) for seven minutes at 57^{o}. HFBA is used instead of the original TFA to prevent evaporation and drying during cleavage. The PTC amino acid is then extracted in 2 ml of butyl chloride (S3) and delivered to

a fraction collector tube. The protein sample is dried down, the fraction collector advanced one step, and the whole process begins again.

All of my automated sequencing was done using one of two programs on the 890. Early work was done using a standard Quadrol D program. Quadrol is the Beckman trade name for N,N,N',N'-tetrakis (2-hydroxypropyl) ethylenediamine and is used in R2 as the buffer. The Quadrol D program includes two cleavages and two extractions (one delivers to the fraction collector and a second to waste) per cycle. This program suffered from a high degree of washout and acidolysis, making residues out 30 residues or more difficult to interpret.

The DMBA program was designed by Hermodson et al (22) and employs N,N-dimethylbenzyl amine (DMBA) with propanol and water as the coupling buffer (12:40:48) at pH 9.4. This program uses a single cleavage and extraction for each cycle and gives far better results than the Quadrol system.

If each reaction is only 98% efficient, a noticeable lag builds up in the sequence after 10 or more steps. This lag can be corrected for mathematically on a homogeneous protein.

Ideally, of course, one could make his reaction conditions such that 99.9% efficiency or better was obtained and thus hundreds of cycles could be performed from the N-terminus until the entire sequence was finished. Needless to say, this is not yet possible for the reasons I will discuss below.

Probably the greatest problem encountered in protein sequencing is the build up of background during the run. Exposure

of the protein to HFBA at 57° causes a small number of apparently random breaks due to acidolysis. Each break in the polypeptide generates a new N-terminus that begins sequencing along with the main region of interest. Random acidolysis usually results in the background building up so rapidly that by step 25, the residue present at each step is often not the major component in the analysis. One then has to resort to net change in the quantity of each amino acid from one step to the next to determine the sequence far out in the chain. Even this approach often yields questionable results out farther than 40 residues from the original N-terminus on a protein. Since acidolysis generates background by breakage in the polypeptide chain, obviously the shorter the chain, the less the background, and peptides should yield better data and indeed they do, except that they are beseiged by a problem of their own, namely "washout".

Washout is a problem generated by small peptides with PITC coupled to them. They are somewhat soluble in benzene and butyl chloride and thus when the peptide is washed with these solvents, a certain amount of peptide leaves the reaction chamber. The amount of peptide being lost at each step increases as the peptide gets shorter and by the time three or four residues are left, usually the peptide is all gone. The amount of washout depends upon the composition of the peptide, of course, and rarely certain long peptides can be sequenced to the penta peptide stage or better. As a rule of

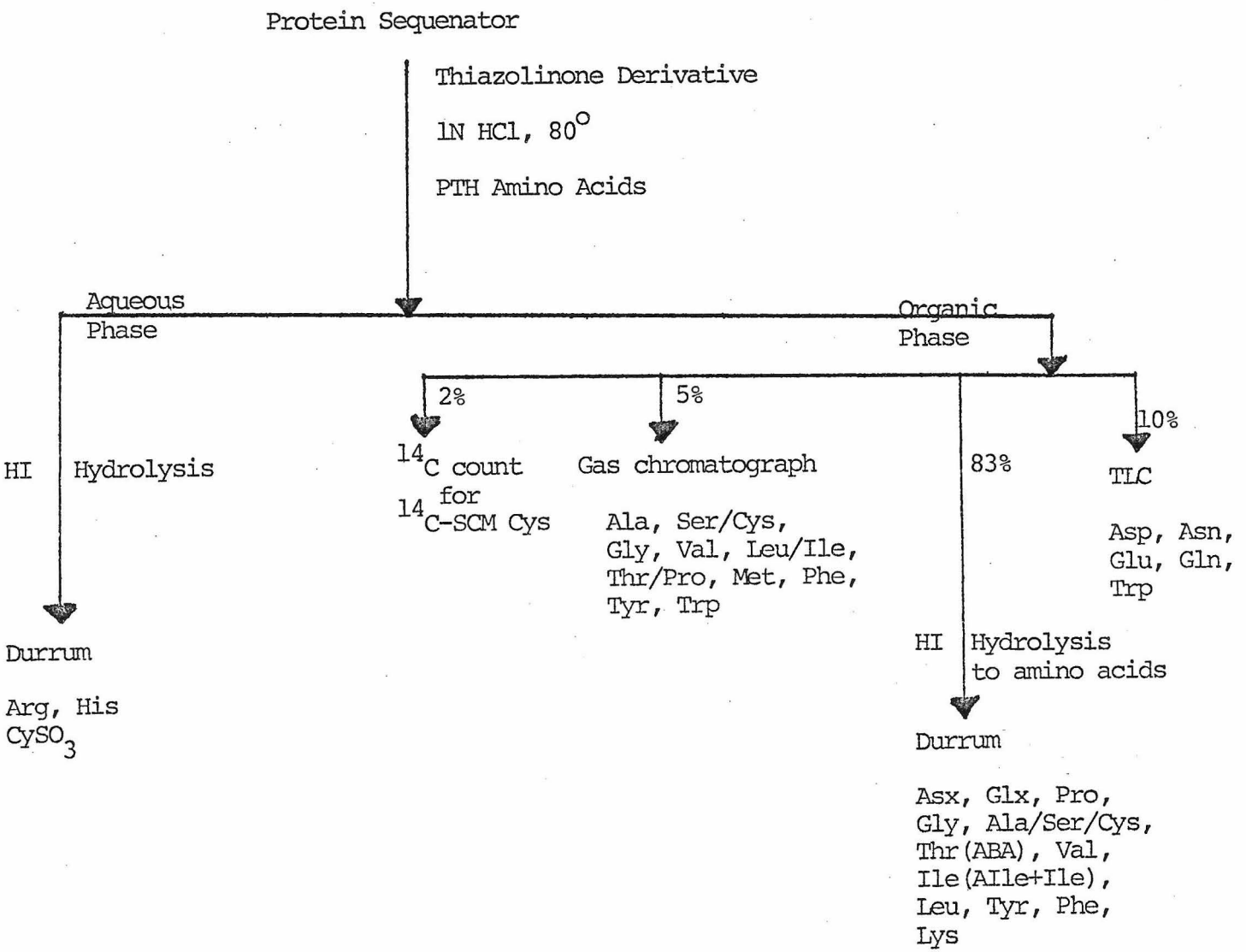
thumb, however, one should not expect to be able to do more than half of a random peptide with the present day spinning cup technology.

Other reactions including O→N acetyl migrations on acetylated hydroxy amino acids and α→β peptide shifts at aspartic acid and aparagine residues apparently cause some blockage of the N-terminus when these residues are encountered.

Other problems exist in the automated Edman degradation but by far these are the most limiting. Given a functional sequenator then, one can hope to turn out 40 or more steps for analysis at the current rate of 15 steps per day on 10 to 15 mg of protein.

Conversion procedure. Since the amino acid thiazolinone is relatively unstable, it must be converted to the more stable phenylthiohydantoin (PTH). This is done by adding 200λ 1N HCl to each tube after the butyl chloride has been dried off under nitrogen. Each sample is heated to 80°C for ten minutes in a temp-block, cooled slightly, and then extracted twice with 750λ of ethyl acetate. Vigorous vortexing is required to thoroughly mix the HCl and organic phase to insure complete removal of the PTHs. (PTH norleucine is added to each sequenator tube in the sequenator to normalize for any handling losses.) Since the ethyl acetate-aqueous partition is done at low pH, any charged PTHs, i.e., histidine and arginine, remain in the aqueous phase and must be analyzed for separately. The 1.5 ml of ethyl acetate is then dried down under nitrogen, redissolved in 50 µl

Fig. 3-2.--Flow diagram for analysis of various amino acids generated during automated sequence analysis. The percentage of the organic phase sample required for each analysis is indicated.


Protein Sequenator
Thiazolinone Derivative
1N HCl, 80°
PTH Amino Acids
Aqueous Phase
Organic Phase
2%
5%
10%
83%
HI
Hydrolysis
Durrum
Arg, His
CySO3
14C count for 14C-SCM Cys
Gas chromatograph
Ala, Ser/Cys, Gly, Val, Leu/Ile, Thr/Pro, Met, Phe, Tyr, Trp
TLC
Asp, Asn, Glu, Gln, Trp
HI
Hydrolysis to amino acids
Durrum
Asx, Glx, Pro, Gly, Ala/Ser/Cys, Thr(ABA), Val, Ile(AIle+Ile), Leu, Tyr, Phe, Lys

of ethyl acetate, and transferred to a gas chromatography vial for injection into the gas chromatograph (GC).

Gas chromatograph. The Hewlett-Packard 7600 gas chromatograph is a modular affair composed of a separate plug in chassis for each function required. Thus, the oven is separate, as is the recorder, temperature programmer, signal amplifiers (electrometer), auto injector and injection control modules. This offers one the chance to design a gas chromatograph to his needs.

The GC is used to monitor the sequencer run while it is in progress. Fortunately, each GC sample can run in 25 or 30 minutes and the auto injection device can hold 17 samples, so with a visit to the laboratory around midnight, one can process up to 48 steps per day.

In order to operate the gas chromatograph, one merely has to position the first sample under the syringe in the auto injection device, push the reset and run buttons and the GC does the rest. After vigorous purging of the syringe to remove any air, 2.5 µl of sample are injected into the sample column. The syringe is then washed with several volumes of ethyl acetate. A typical run includes six minutes of isothermal operation at 185^{o} followed by a 15^{o} per minute rise to 270^{o} and a ten-minute hold at that temperature. At the end of that period, the GC oven cools back down to 185^{o} and, after a short delay, injects the next sample. Two identical columns are used in this instrument--one for the sample and

one for reference--and the recorder display reflects the difference between the output of these two columns.

Only hydrophobic amino acids pass though our six-foot DC-560 columns (23) and are analyzed in the flame ionization detector. Thus, the amino acids alanine, serine, glycine, valine, threonine, proline, leucine, isoleucine, norleucine, methionine, phenylalanine, tyrosine, and tryptophan can be assayed. The results are semi-quantitative in the one nano-mole range, but this vastly improves as the sample load increases. Also, some amino acids are more stable than others as can be seen by the standard. (Fig. 3-3) Serine and tryptophan are particularly troublesome residues that refuse to show up on the chromatogram occasionally if the sample size is small. SCM-cysteine also appears as serine on the GC and its analysis requires radiolabelling with ^{14}C iodoacetamide prior to loading in the sequenator and counting the sequencer steps in a scintillation counter (Fig. 3-4).

Thin-layer chromatography. Before hydrolysis of the other 95% of the sample back to amino acids for analysis on the Durrum, it is necessary to distinguish glutamine from glutamic acid and asparagine from aspartic acid, since the hydrolysis causes deamidation. A thin-layer chromatography on silica gel plates coated with fluorescent indicator using a solvent composed of benzene: acetic acid (9:1) is used for this purpose (24). After removal of the solvent in an 110^{o} oven, the plate is sprayed with a 1% solution of ninhydrin in acetone which upon heating

Fig. 3-3.--Gas chromatogram of hydrophobic amino acid PTHs. Approximately one nanomole of each PTH was injected into the DC-560 column. The program was 185^{o}C for six minutes, then increased at 15^{o} per minute to 270^{o} and held at that temperature for 10 minutes.

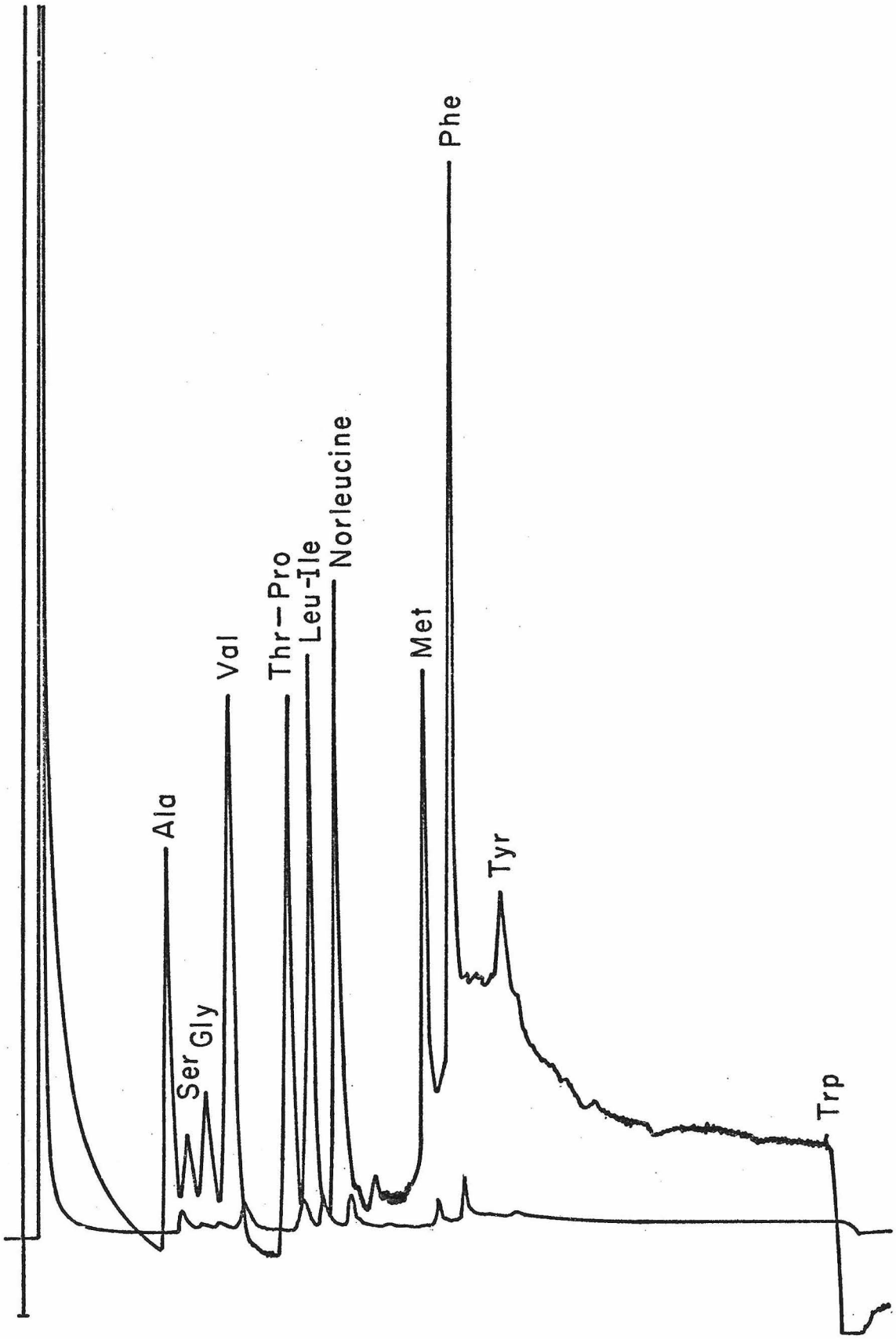
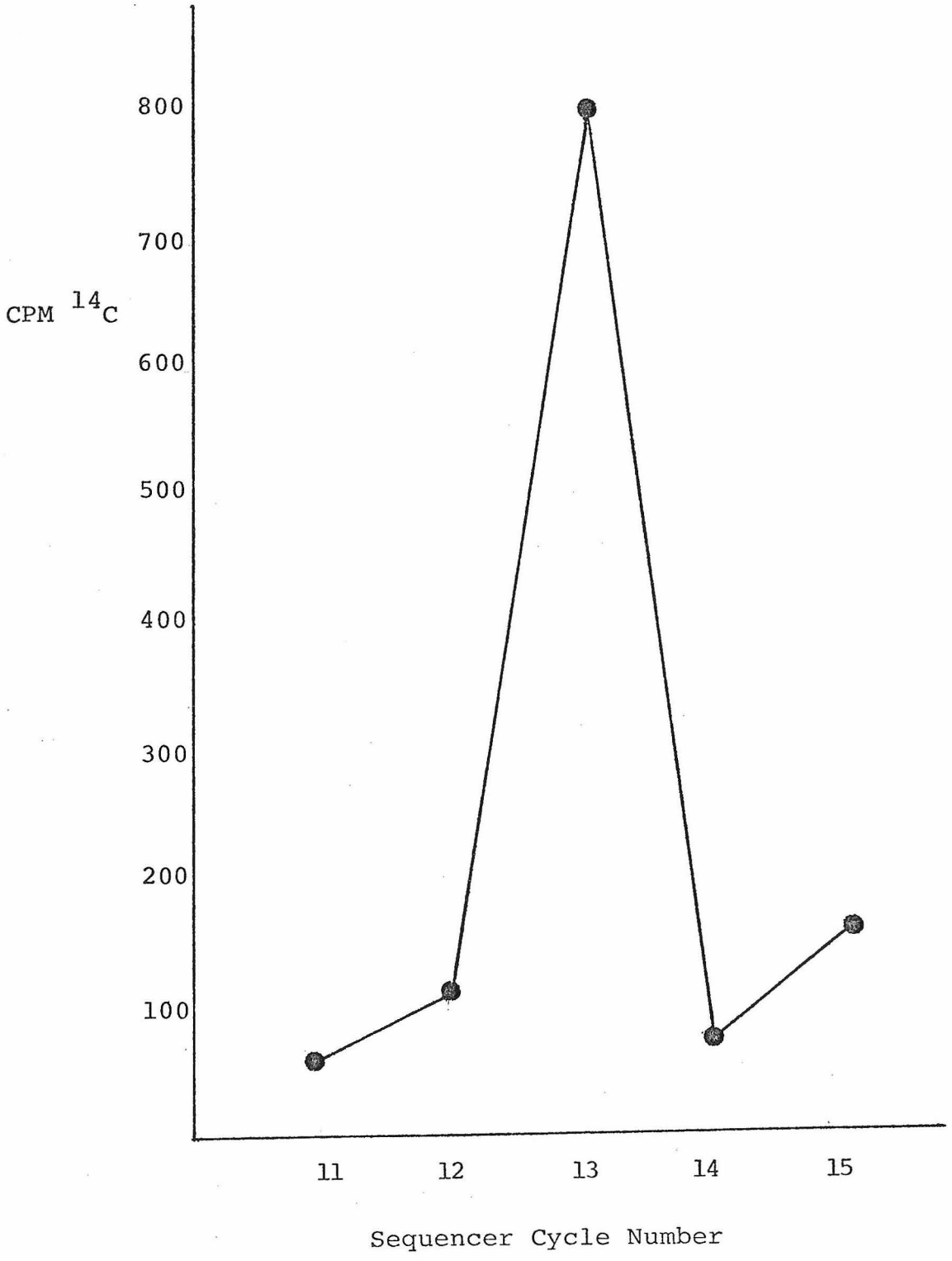

Ala
Ser
Gly
Val
Thr—Pro
Leu-Ile
Norleucine
Met
Phe
Tyr
Trp

Fig. 3-4.--Detection of Cysteine using ^{14}C S-carboxy-methylation. Two percent of each organic phase from sequencer steps 11 to 15 from H8CN1C was dissolved in 5 ml Aquasol and counted for ten minutes in a Beckman LS-100C scintillation counter. The Cysteine residue is at position 13.

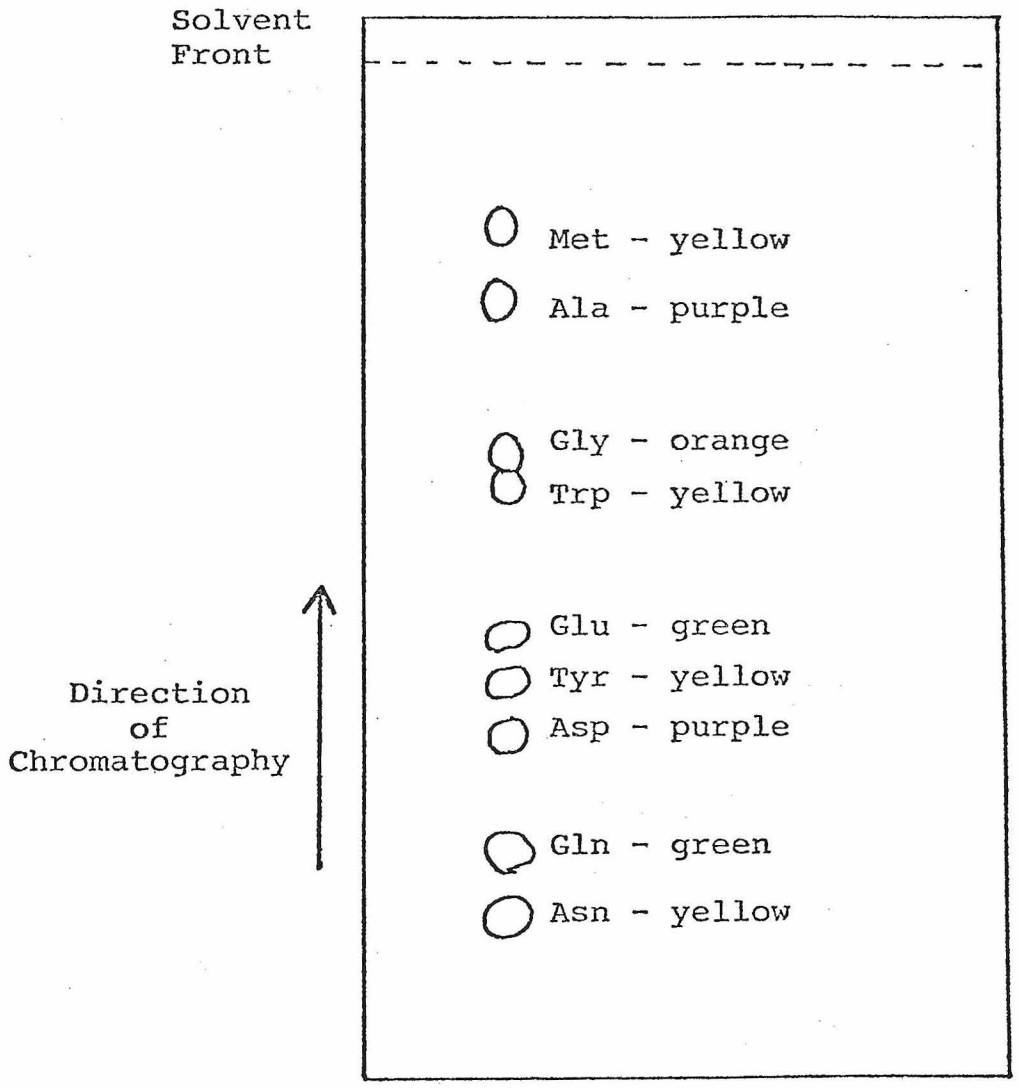

800
700
CPM ^{14}C
600
500
400
300
200
100
11
12
13
14
15
Sequencer Cycle Number

to 110°C for five minutes, causes glutamic acid and glutamine PTHs to show up bright green, while aspartic acid PTH is red or violet, and asparagine is yellow. Also giving colored products in this system are alanine (violet), glycine (orange), tyrosine (yellow), and tryptophan (dull yellow) (Fig. 3-5).

Hydrolysis of PTHs. Hydrolysis of the PTHs back to the amino acids is accomplished in a dessicator in 57° Hydroiodic acid flushed with N_2 and sealed under vacuum which is put in an autoclave for 18 hours at 124°(25). This procedure destroys cysteine and serine which give alanine in the process, converts threonine to alpha amino butyric acid (ABA), destroys methionine, converts tryptophan to gly and ala, and racimizes PTH-isoleucine to isoleucine and allo-isoleucine. Since histidine and arginine are left in the aqueous phase during conversion, only 11 amino acid peaks (counting ile and allo-ile as one) are of interest on the Durrum.

Durrum D-500 amino acid analyzer. The Durram D-500 amino acid analyzer was truly a breakthrough when it first entered the market three years ago. It cut analysis time from 3-4 hours to 1 or 2, increased sensitivity by an order of magnitude, and added its own data processing unit in the form of a PDP-8 computer. It employs a single column sulfonated polystyrene resin which separates all the amino acids normally encountered in acid hydrolysates. It also employs two photometers--one at 440 nm for proline, the other at 560 nm for the ninhydrin reaction products of all other amino acids.

Fig. 3-5.--Relative positions and colors of PTH amino acids on silica gel coated with fluorescent indicator. Relevant PTH AA are indicated.


Solvent
Front
Met - yellow
Ala - purple
Gly - orange
Trp - yellow
Glu - green
Tyr - yellow
Asp - purple
Gln - green
Asn - yellow
Direction
of
Chromatography

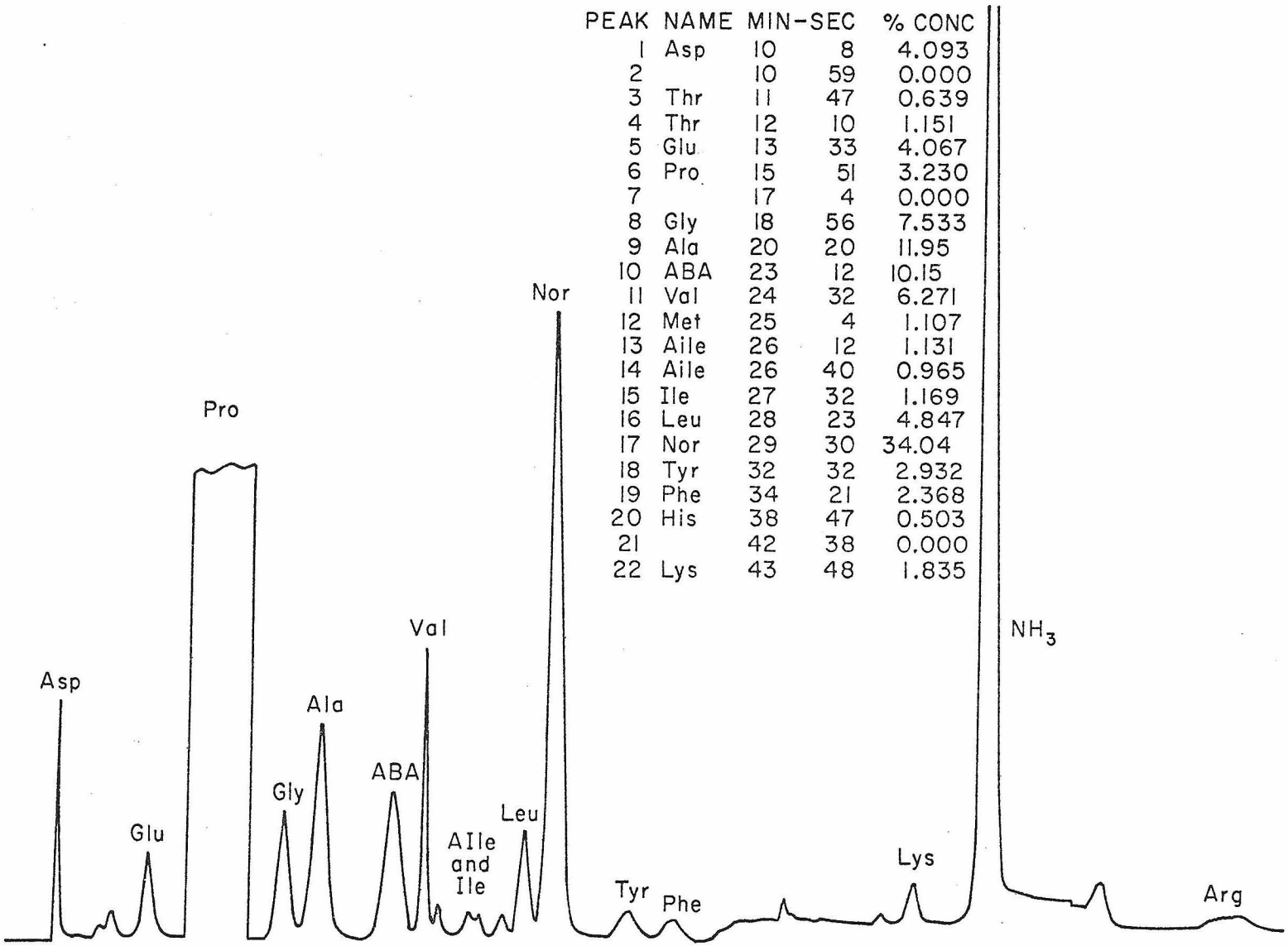
The sample is loaded in 40 µl or less into a stainless steel and teflon cartridge, called a shu, and rolled down the sample chute. When the run is initiated, the analyzer flushes the last shu, ejects it by pneumatic pressure, and aligns the next shu by the injector. The sample is injected in buffer A and this buffer is pumped through the column until alanine has been eluted (Fig. 3-5). Buffer B is pumped next and continues until tyrosine comes off and then buffer C is used for histidine, lysine, and arginine. After the analysis, the column is regenerated with NaOH and the procedure beings again.

Resolution of threonine and serine is the most difficult task for this system and if one requires good separation of these two amino acids a two-hour program is required. Since HI hydrolysis destroys these two amino acids, however, the buffer pumping rate can be increased and also since arginine is left in the aqueous phase, we can stop the run after lysine (and save 12 minutes per analysis) allowing us to run a sequenator sample in 60 minutes.

Figure 3-6 shows a typical Durrum chart from step 20 of a light chain, T15L. All significant peaks for a sequencer analysis are indicated by the three-letter abbreviations above each peak. This is a 72-minute analysis and includes arginine. The insert shows the mole percent data for each peak as computed by the PDP-8. Nor is norleucine and ABA is α-amino butyric acid.

Figure 3-7 illustrates the gradual decline in the signal to noise ratio as the sequencer run progresses. Two sets of Durrum analyses from T15 light chain run are illustrated.

Fig. 3-6.--Durrum chromatogram from a light chain, T15L. All significant peaks are indicated with the three-letter abbreviations. The mole percent data computed by the PDP-8 for each amino acid are shown in the insert. The order of elution of the peaks is from left to right and one inch on the chart paper equals five minutes.



PEAK	NAME	MIN	SEC	% CONC
1	Asp	10	8	4.093
2		10	59	0.000
3	Thr	11	47	0.639
4	Thr	12	10	1.151
5	Glu	13	33	4.067
6	Pro	15	51	3.230
7		17	4	0.000
8	Gly	18	56	7.533
9	Ala	20	20	11.95
10	ABA	23	12	10.15
11	Val	24	32	6.271
12	Met	25	4	1.107
13	Aile	26	12	1.131
14	Aile	26	40	0.965
15	Ile	27	32	1.169
16	Leu	28	23	4.847
17	Nor	29	30	34.04
18	Tyr	32	32	2.932
19	Phe	34	21	2.368
20	His	38	47	0.503
21		42	38	0.000
22	Lys	43	48	1.835

Fig. 3-7.--Representative data from cycles 20 and 30 on a T15 light chain. The relative amount of each amino acid as determined by the Durrum D-500 amino acid analyzer is shown by the mole percent figures. The prominent residue from the gas chromatography analysis of these samples is indicated on the line marked GC. The line marked Seq indicates the determined amino acid sequence in these two regions. As is apparent, the sequence around step 30 is slightly less obvious and more lag exists here than at step 20. These data were obtained on 10 mg of light chain using the DMBA program.

Step #	19	20	21	29	30	31
Asx	6.3	7.7	6.5	6.9	7.7	8.2
Glx	6.7	7.8	7.1	9.9	8.7	8.8
Pro	4.4	4.2	4.6	5.0	4.2	4.2
Gly	6.0	7.1	6.3	6.7	6.9	6.9
Ala	15.8	16.1	17.6	17.6	20.9	27.0
ABA	6.4	19.4	7.9	5.0	4.7	5.6
Val	20.7	8.6	4.3	4.0	3.7	4.1
AI + Ile	3.2	4.3	23.0	3.7	3.5	4.0
Leu	7.0	7.7	7.3	17.0	11.4	9.8
Nor	4.8	5.8	5.0	5.4	6.0	6.8
Tyr	4.0	4.9	4.6	5.5	12.6	8.8
Phe	2.8	2.9	2.8	3.4	3.4	3.5
Lys	10.5	3.8	4.0	3.6	4.8	4.2
GC	Val	Thr/Pro	Leu/Ile	Leu/Ile	Tyr	Ser
Seq	Val	Thr	Ile	Leu	Tyr	Ser

One is from 19 to 21 and the second from 29 to 31. As can be seen some amino acids are more obvious than others causing questionable results at some positions far out in the sequence.

The Durrum analyzer is a high pressure system and has the normal tendency to develop leaks, usually in the injection valve or around the ram. A leak is indicated by peaks running late on the chromatogram. Periodic inspection and tightening keeps this problem to a minimum. Sample quantity is important since the recorder can only be set for one of five linear scales and too little sample gives poor quantitation while too much drives the peaks off scale, also ruining the quantitation.

Even with these problems the Durrum can operate effectively with such low sample loads (2.5 nm full scale on recorder) that usually half of the sample or less is required per analysis and therefore, any step can be rerun if necessary.

This is how automated sequencing is done in our laboratory; start to finish requires about one week to sequence 40 residues, providing all the equipment is operational. After the first one hundred conversions, the work lacks excitement but the data obtained give insight into key biological questions, obtainable in no other manner. In the last few years, it has provided us with data relevant to the biological basis of antibody diversity and specificity, the subject of this thesis.

CHAPTER IV

The Heavy Chain Diversity

The publication in this chapter is from the Proceedings of the National Academy of Science, USA. It is reproduced here with their permission.

Reprinted from
Proc. Nat. Acad. Sci. USA
Vol. 71, No. 10, pp. 4096–4100, October 1974

52

Mouse Immunoglobulin Heavy Chains Are Coded by Multiple Germ Line Variable Region Genes

(amino-acid sequence/sequenator/BALB/c mouse/hapten binding/selection)

P. BARSTAD*, V. FARNSWORTH*, M. WEIGERT†, M. COHN‡, AND L. HOOD*

* Division of Biology, California Institute of Technology, Pasadena, Calif. 91109; † Institute for Cancer Research, Philadelphia, Pennsylvania 19111; and ‡ Salk Institute, San Diego, California 92112

Communicated by Ray D. Owen, June 20, 1974

ABSTRACT **The N-terminal 20 residues of 13 heavy immunoglobulin chains from myeloma proteins of the BALB/c mouse are compared with the same residues of 15 other heavy chains described in the literature. Sixteen of 28 sequences are different from one another. These proteins fall into four major sets, with 18 of the proteins in the largest set being further divisible into at least five subsets. This pattern of diversity suggests there are at least eight germ line genes coding for the variable regions of mouse heavy chain. Many of the immunoglobulins from which these heavy chains are derived exhibit binding activity for various haptens. The differing hapten specificities are closely correlated with distinct primary amino-acid sequences.**

An understanding of the genetic mechanism responsible for antibody diversity is one of the most intriguing problems in modern immunology, for its resolution may provide insights into the problems of information storage and expression in other complex systems (1, 2). Two alternative hypotheses for antibody diversity have been proposed—the *germ line* theory suggests that diversity arises in antibody genes during the evolution of the species, whereas the *somatic* theories propose that antibody diversity is created anew during the development of each individual by somatic mutation (1–3). Insights into the nature of this problem have come, in part, through the patterns observed from the detailed amino-acid sequence analysis of homogeneous myeloma immunoglobulins. The immunoglobulin molecule is comprised of two different polypeptides, light and heavy chains (4). All immunoglobulin chains can be divided into two segments, the variable (V) and the constant (C) regions, which code, respectively, for the antigen-binding and the various effector functions of the antibody molecule (5). There are three families of immunoglobulin genes, which are unlinked in the mammalian genome. Two code for light chains [lambda (λ) and kappa (κ)] and the third codes for heavy chains (1). The *V* and *C* genes of a given family appear to be coded by separate germ line genes (6). Myeloma proteins from the inbred BALB/c mouse have proven to be a fruitful system for the analysis of amino acid sequence patterns because tumors can be induced at will and are readily transplanted from one individual to a second (7). In addition, these inbred mice are presumably genetically homogeneous, in contradistinction to the genetic polymorphism present in most outbred populations such as humans and rabbits.

Abbreviations: V and C, variable and constant regions of immunoglobulin chains.

Two important patterns relevant to theories of antibody diversity have emerged from the sequence studies on myeloma proteins. First, the V regions from most immunoglobulin families can differ so extensively that all immunologists agree that they are coded by multiple germ line *V* genes. Thus, all theories of antibody diversity are multigenic, the question at issue is how multigenic. Second, when the V regions from large numbers of myeloma light or heavy chains are compared, three to four sections of extreme sequence variability are noted, which are designated "hypervariable regions" (8, 9). X-ray crystallographic studies have established that antibody polypeptides fold so that hypervariable regions constitute the walls of the antigen-binding crevice (10, 11, 32). The presence of hypervariable regions that make up the antigen-binding site has divided somatic theories into two groups. Both agree that the total antibody diversity is coded in part by germ line *V* genes and in part by *V* genes somatically derived from the germ line by mutation and stepwise selection for superior antigen binding sites. One class of somatic theories argues that the selection for somatic variants occurs predominantly, if not exclusively, in the hypervariable regions (2), whereas, a second group of somatic theories argues that somatic variation can occur throughout the entire V region (3).

Previous studies from our laboratories have assessed the diversity present in myeloma proteins from the two light chain families of the BALB/c mouse, λ and κ (12–16). In this paper we extend these studies to include an analysis of the N-terminal 20 amino-acid residues of 28 myeloma heavy chains derived from BALB/c mice. These sequences lie outside the hypervariable regions of the heavy chain and will, accordingly, place different constraints on the various theories of antibody diversity outlined above.

METHODS AND RESULTS

Mouse myeloma proteins were isolated from serum of tumor-bearing BALB/c mice by immunoabsorption for those examples with known specificities; J558 and MOPC 104E on dextran B-1355S–polyacrylamide gel (17), W3129 and W3434 on Sephadex G-75, and S23 on dinitrophenyl (Dnp) Sepharose (18). J606 and Y5606 were precipitated from serum by dialysis against 0.01 M NH_4HCO_3 (19). Y5476, Y2020, Y5444, W3082, and S176 were isolated by the methods described by Potter (20). The purified proteins were partially reduced and alkylated and heavy and light chains were separated as described by Bridges and Little (21). The heavy chains were desalted, lyophilized, and approximately 10 mg per run was loaded on the sequencer.

Tumor Number	Ig Class	Light Chain Type	1 5 10 15 20	Activity
Prototype V_{HI}			EVQLQESGPSLVKPSQ TLSL	
M460	IgA	κ	——————————*——	Dnp
S23	IgA	κ	——————————————	Dnp
M315[a]	IgA	λ	D————G————S——	Dnp
Prototype V_{HII}			EVQL QESGPELVKPGASVKM	
J558	IgA	λ	——*——*——————————	1,3D
M104E	IgM	λ	——————————————	1,3D
Prototype V_{HIII}			EVKL LESGGGLVQPGGSLKL	
Y5476	IgA	κ	——————————————	L
W3434	IgA	κ	——————————————	1,6D
W3129	IgA	κ	——VI*————*————	1,6D
M173[c]	IgA	κ	————————P——L————	U
H2020	IgA	λ	—Q—V————*————	U
Y5444	IgG2a	λ	—M—V————————*—	U
S10[d]	IgA	κ	——————————————	1,6G
X24[d]	IgA	κ	——————————————*	1,6G
X44[d]	IgA	κ	——————————————	1,6G
T191[d]	IgA	κ	——————————————	1,6G
J539[d]	IgA	κ	——————————————	1,6G
J1[d]	IgA	κ	———I————————	1,6G
H8[e]	IgA	κ	———V————————R—	PC
T15[e]	IgA	κ	———V————————R—	PC
S107[e]	IgA	κ	———V————————R—	PC
M603[e]	IgA	κ	———V————————R—	PC
M167[e]	IgA	κ	———VV———————R—	PC
Y5606	IgG3	λ	D—Q—V————*——Z——	TMA
M21A[b]	IgA	κ	D—Q—VQ———————M——	U
M406[b]	IgA	κ	D———Q——————————	AM
W3082	IgA	κ	———E————————M——	L
J606	IgG3	κ	———E————————M——	L
Prototype V_{HIV}			EVQLQZSGTVLARPGSSLKM	
S176 major	IgA	λ	S——————————————	5AU
S176 minor	IgA	λ	——————————————	5AU

FIG. 1. N-terminal sequences of BALB/c heavy chains. The one letter code of Dayhoff (28) is: A, Ala; B, Asx; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro, Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp, X, ?; Y, Tyr; Z, Glx; Ig indicates immunoglobulin; Dnp designates dinitrophenyl; 1,3D indicates α-1→3 dextran; 1,6D designates α-1→6 dextran; U indicates unknown specificity; L designates levan; AM designates *N*-acetyl-D-mannosamine; 1,6G indicates β-1→6 galactan; PC indicates phosphorylcholine; TMA designates trimethylamine; and 5AU is 5-acetyluracil. [a] From reference 29. [b] Quoted in ref. 7 and represents sequences. derived from a single sequentor run (J. M. Kehoe and J. D. Capra, personal communication). [c] From 30. [d] From 31. [e] From 24. * indicates uncertainty about amide or residue assignment. M104E indicates MOPC 104E; M406 indicates MOPC 406; the remainder of the complete names for the proteins is given in ref. 16.

Each chain was subjected to 20 steps of automated Edman degradation on an updated Beckman 890A protein sequenator employing a standard Quadrol or DMBA program. Samples were dried under N_2 and converted to the phenylthiohydantoin amino acids (PTH) by treatment with 1 M HCl–1 mM ethanethiol for 10 min at 80°. Following extraction in ethylacetate, 5% of each sample was run on a Hewlett Packard gas chromatograph equipped with dual glass 7.5% DC-560 columns. Glutamine–glutamic acid and asparagine–aspartic acid were resolved by thin-layer chromatography on silica gel (benzene:acetic acid, 9:1) or by silylation of the phenylthiohydantoin derivatives followed by gas chromatography. The amino acids were then regenerated by hydrolysis in hydriodic acid for 18 hr at 120° (22) and analyzed on a Durrum D-500 amino-acid analyzer. Throughout these experiments phenylthiohydantoin norleucine was added to the sequencer tubes prior to conversion, to normalize for any losses due to handling. The methodologies used have been described in detail in a previous publication (16). Yields of approximately 50% of the theoretical values were noted at step 1 and repetitive yields averaged 90% for all chains.

TABLE 1. *Percent homology between mouse and human prototype sequences for the N-terminal 20 residues*

	*MV_{HII}	MV_{HIII}	MV_{HIV}	*HV_{HI}	HV_{HII}	HV_{HIII}
MV_{HI}	65	45	55	40	65	45
MV_{HII}		60	75	70	50	60
MV_{HIII}			60	45	50	85
MV_{HIV}				55	40	60
HV_{HI}					40	50
HV_{HII}						50

* MV_{HI} and HV_{HI} respectively denote the prototypes from mouse and human V_H regions. The human prototype sequences are derived from the data given in ref. 26.

DISCUSSION

Selection of Proteins. The heavy chains used in this analysis were generally selected because the immunoglobulins from which they were derived exhibited specific binding for one or more haptens (Fig. 1).

Patterns of Diversity. The amino terminal 20 residues of 13 heavy chains are compared with 15 sequences taken from the literature (Fig. 1). Sixteen of the 28 sequences differ by one residue or more. By sequence homology, these sequences can be divided into four major sets, which contain, respectively, 22, 3, 2, and 1 proteins. Prototype sequences can be derived for each of these sets by selecting the major residue alternative expressed at each position (Fig. 1) and these are respectively designated prototypes V_{HIII}, V_{HI}, V_{HII}, and V_{HIV}.§ The four prototype sequences differ from one another by 25–55% of their residues (Table 1). Prototype set V_{HIII} can be broken down into at least four subsets that are defined by two or more linked residues. For example, the heavy chains from W3082 and J606 have a glutamic acid at position 5 and a methionine at position 18 that distinguish these proteins from the remaining proteins of this set (Fig. 1). In this fashion, the heavy chains from H8, T15, S107, M603, and M167 as well as Y5606 and M21A can also be divided into two additional subsets by the presence of linked amino acids (Fig. 1).

The variable region of the heavy chain (V_H region) is about 120 residues in length and hence the N-terminal 20 residues represent about 1/6 of the V_H region. Accordingly, a rough

§ The largest set of heavy chains is designated V_{HIII} because the chains show the greatest homology with the human heavy chains belonging to the V_{HIII} subgroup (see ref. 26). We have refrained from using the term V region subgroup for two reasons. First, with the accumulation of large amounts of sequence data it has become apparent that the definition of what constitutes a subgroup is uncertain. For example, within many subgroups, additional groupings are defined using smaller numbers of linked residues. Are these new groupings additional V region subgroups? Second, the term V region subgroup has had the genetic implication that it is encoded by at least one germ line gene. Yet each of the theories of antibody diversity incorporates very different ideas about what types of sequence patterns define additional germ line V genes. Thus we feel the term subgroup is undefinable, has confusing genetic implications, and should be replaced by a more neutral term such as "set."

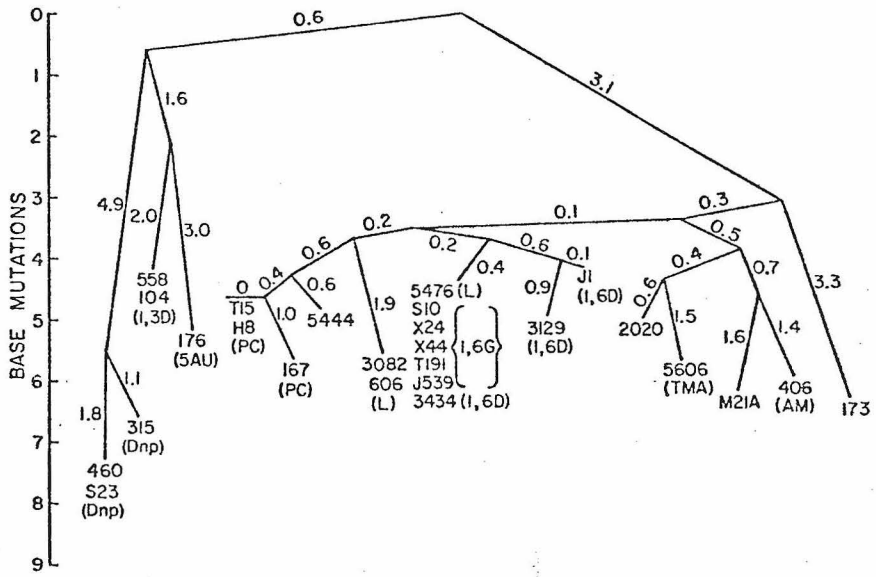


FIG. 2. A genealogic tree of heavy chains from BALB/c myeloma proteins. This tree was constructed using only the N-terminal 20 residues. Many of the immunoglobulins from which these V_H regions were derived bind to one or more haptens. These specificities are shown below the tumor number. Abbreviations: Dnp, dinitrophenyl; PC, phosphorylcholine; 1,6D, α-1→6 dextran; 1,3D, α-1→3 dextran; L, Levan; TMA, trimethylamine (7).

estimate of the number of amino-acid differences between two V_H regions can be determined by multiplying the difference in diversity shown over the N-terminal 20 residues by six (this is an underestimate because no hypervariable regions are included in this region). Thus, proteins in different sets will probably differ by 48 residues or more, whereas proteins within a set can, from this estimate, differ by as many as 24 or more residues. The only two nearly complete mouse V_H sequences from different sets, M315 and T173, are consistent with the above analysis in that they differ by 10 residues in the first 20 and by 62/104 residues and two sequence gaps.

Theories of V Gene Diversity. A more sophisticated examination of these proteins can be carried out by a genealogic analysis (1). The sequences within any set can be related to one another by an evolutionary tree which depicts the minimal number of genetic events (i.e., single base substitutions) required to generate this set of sequences from a single ancestral sequence (gene). This procedure permits a step by step analysis of the single base substitutions required to generate diversity in a given set of sequences. One can then ask what events are compatible only with germ line evolution and what events might be explained by somatic evolution. The genealogic analysis in Fig. 2 shows that the relationships of these sequences to one another are somewhat more complex at the nucleotide than at the protein level because certain single amino-acid differences are coded by two base substitutions. In spite of this, four distinct branches are noted on the genealogic tree, which correspond to the four sets of sequences described previously, V_{HI} through V_{HIV}. The V_{III} branch is divided into five or more sub-branches by the presence of linked nucleotide bases in certain of the sequences (e.g., T15-167-5444; 3082-606; 5476-3129-J1; 2020-5606-M21A-406; and 173—see Fig. 2). Obviously additional data will be required (more and longer sequences) to more sharply delineate this genealogic tree, particularly with regard to the sub-branch structure of V_{HIII}.

The somatic theory which argues for somatic variation only in hypervariable regions (2) and the germ line theory would agree on the conclusions to be drawn from this genealogic tree; namely, each of the V regions that differ from one another is coded by distinct germ line *V* genes that arose by gene duplication, sequence divergence (mutation, recombination, etc.) and selection during the evolution of the species. Thus both of these theories would agree that the diversity noted in these V_H regions must be coded by at least 16 germ line V_H genes. If, however, it is postulated that replacements in the N-terminal 20 residues result from somatic mutation and selection, two questions must be asked. First, how many single base substitutions are reasonable to postulate during the somatic development of a given lymphocyte line? Second, how many parallel (or identical) mutations may occur in lymphocyte lines in separate individuals? If somatic diversification occurs by ordinary random mutation followed by selection, then it would appear likely that relatively few (i.e., 1–5) mutations occur per lymphocyte line because of the need to select and expand out each mutant clone of lymphocytes (2). In addition, no selective forces adequate to explain significant parallel mutation have been postulated to occur somatically, thus each set of sequences with multiple linked residues is probably derived from a different germ line gene (2). Accordingly, from the viewpoint of random somatic mutation that can occur throughout the entire *V* gene, at least eight germ line V_H genes must be postulated for this group of V_H regions, one for each distinct branch and sub-branch of the genealogic tree. A discussion of other somatic theories is beyond the scope of this report.

The mouse heavy chain sequences examined here appear to exhibit less diversity than the mouse κ family and far more diversity than the mouse λ family. This generalization should be qualified, however, because most of these V_H sequences were derived from immunoglobulins selected because they exhibited binding activity for various haptens. Thus, these heavy chains certainly constitute a selected subset of BALB/c heavy chain myeloma pool. The degree of selection that occurred in examining primarily V_H regions derived from myeloma proteins with binding activity is entirely unknown.

It will be very important to analyze the diversity exhibited by unselected V_H regions to obtain a more reliable estimate of the diversity present in the BALB/c pool of myeloma heavy chains. In addition, mouse heavy chains blocked at their N-termini comprise about 78% of the normal serum immunoglobulin (23). Thus it will be particularly valuable to analyze the extent of diversity present in these blocked sequences.

Normal Pooled Mouse Heavy Chains Appear to be More Restricted in Sequence Than the Pool of Myeloma Heavy Chains. A recent study reports that normal pooled mouse heavy chains differ in two respects from the sequences reported here. First, as mentioned earlier, the normal mouse H chains have a blocked α-amino group in 78% of the proteins (23). Only one of 20 myeloma H chains we examined was blocked (HOPC-1). Second, the pool of unblocked mouse H chains obtained from the serum appeared to have a single residue at the 95% level at most of the N-terminal 28 positions. This is in striking contrast to the myeloma data we have reported here. For example, in the pool of myeloma heavy chains, aspartic acid is seen in 14% of the sequences at position 1; glutamine is seen in 32% of the sequences at position 3; the alternatives at position 5 are evenly spread among a number of residues. Thus, the myeloma pool of heavy chains reveals a heterogeneity not detectable by sequencing the normal pool, because the myeloma population, not surprisingly, expands individual sequences that are normally represented infrequently or not at all in the serum pool. Hence one must be cautious about drawing conclusions concerning the nature and extent of V region diversity from studies on normal pooled sequences.

It appears that the myeloma immunoglobulins are themselves highly selected in that 5% or more of them bind to a few determinants (7). This is unexpected in view of the enormous functional heterogeneity of the normal immune response. Accordingly, both the normal and myeloma pools could be interpreted not to represent the true V region diversity that is coded in the BALB/c genome.

One obvious difference between the myeloma pool and the normal serum pool is the fact that the artificial induction of myeloma proteins in BALB/c mice selects primarily IgA proteins, whereas the normal serum pool is mostly of the IgG type (7). Obviously, an analysis has yet to be carried out of the extent and nature of amino-acid sequence diversity in the V_H regions associated with BALB/c γ heavy chains. In any case, a critical question arises. Are the V_H regions associated with $C\gamma$ regions distinct from those associated with C_α regions? If the V_H regions from γ and α myeloma proteins are equally diverse, then the normal and myeloma heavy chain pools express the same family of germ-line V_H genes. If the V_H regions from γ myeloma proteins are less diverse than those from their α counterparts, then selection must occur in the association of certain germ-line V_H genes with given C_H genes.

Some Mouse V_H Regions Are Similar to Their Human Counterparts, Whereas Others Are Quite Distinct. A rough comparison of the similarity of mouse and human heavy chains at their N-termini can be made by comparing the prototype sequences of the mouse and human V_H sets (Table 1). The mouse and human prototypes are quite distinct from one another (30–60%) apart from the mouse and human V_{HIII} sequences (about 15% difference). The comparison of the single nearly complete human V_{HIII} sequence (Nie) with two mouse V_H regions (M173-39% and M315-61%) suggests that the group of mouse sequences to which M173 belongs is more closely related to the human V_{HIII} sequences than other mouse V_H regions are, although they are not as closely related as suggested by the N-terminal data. The fact that many mouse V_H prototypes are as closely related to one another as to their human counterparts suggests that multiple V_H genes existed in the mouse-human ancestor (Table 1). Thus some of the V_H genes in human and mouse may have been derived from a common V_H gene in the human–mouse ancestor, whereas others appear to have diverged from distinct genes in this ancestor. The heterogeneity of the myeloma heavy chains at the N-terminus compared with their normal counterparts also suggests that from the analysis of normal pooled sequences it is not possible to conclude that species- or phylogenetically associated residues are coded by a major portion of the germ-line V genes.¶

The Hapten Binding Properties of Certain Myeloma Proteins Appear to Correlate with Their Heavy Chain Sequences. Six groups of myeloma proteins with binding activity for different haptens are shown in Fig. 1. The binding activity for a specific hapten appears to correlate with the sequence of the N-terminal 20 residues (Figs. 1 and 2). Indeed, each of the major branches on the genealogic tree with just a few sequences correlates with a single hapten binding activity (e.g., dinitrophenyl–V_{HI}, α-1→3 dextran–V_{HII}, and 5-acetyluracil–V_{HIV}). The binding specificities for heavy chains in the V_{HIII} branch generally correspond to the linked sets of proteins or subbranches that occur within this branch (e.g., the phosphorylcholine and levan binding heavy chains). Other studies suggest that the correlation of specificity with distinct amino-acid sequences is much greater with the heavy than with the light chain sequences (24).

The Correlation of Particular Antigen Binding Specificities with Particular N-Terminal Sequences Suggests That Selection for Function Occurs Outside As Well As Within Hypervariable Segments of V Regions. This observation renders unlikely theories of antibody diversity that argue that special hypermutational mechanisms operate only in the hypervariable regions or that the hypervariable regions are coded by separate episome-like genes which are integrated into a framework V gene. Clearly, mutation and selection occur through the entire V gene. The relative contributions of germ line versus somatic diversification to this process remain uncertain.

Heavy Chains May Have a Precursor with Additional Residues at the N-Terminus. The amino-acid sequence analysis of S176 revealed a mixture of two sequences, probably identical to one another but for a single extra residue at the N-terminus (Fig. 1). A recent report has demonstrated that a myeloma light chain synthesized *in vitro* has an additional 10–15 residues on the N-terminus (25) which are, presumably, cleaved off by a post-translational event *in vivo*. Accordingly, we suggest that heavy chains also have a precursor form that is modified post-translationally. The presence of the normal heavy chain in this tumor product suggests that the extra N-terminal residue cannot be accounted for by a mutational event at the DNA level.

¶ Species- or phylogenetically associated residues are those found at a particular position in most of the serum immunoglobulin chains of one species and which are distinct from the homologous residues of a second species (23, 27).

Are There Any Differences in the Set of V_H Regions Associated with λ As Compared to k Light Chains? None of the heavy chains associated with λ chains has a lysine at position 3, a residue found in the heavy chains associated with 18 of 21 κ type molecules (Fig. 1). Eight of nine heavy chains derived from λ type molecules have glutamine at position 3. On the other hand, the V_H regions from M315 and M460 as well as Y5606 and M21A are quite similar, yet in each case one molecule is of the λ and the other of the κ type. Obviously a great deal more sequence data must be available before conclusions can be drawn regarding V_H associations with V_λ and V_κ regions.

This investigation was supported by NIH Grant AI 10781 and NSF Grant GB 27605 to L.H.; NIH A105875 and CA 05213 to M.C.; and NIH Grant GM 20964 to M.W. L.H. has an NIH Career Development Award. M.W. has an American Cancer Society Faculty Research Award no. PRA-59. We thank M. Potter for tumor MOPC 104E.

1. Hood, L. (1973) *Stadler Symp.* 5, 73–142.
2. Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R. & Weigert, M. (1974) in *The Immune System: Genes, Receptors, Signals*, eds. Sercarz, E. & Williamson, A. (Academic Press, New York & London), pp. 89–118.
3. Smith, G. P. (1973) *The Variation and Adaptive Expression of Antibodies* (Harvard University Press, Cambridge, Mass.).
4. Smith, G., Hood, L. & Fitch, W. (1971) *Annu. Rev. Biochem.* 40, 969–1012.
5. Edelman, G. (1971) *Ann. N.Y. Acad. Sci.* 190, 5–25.
6. Hood, L. (1972) *Fed. Proc.* 31, 177–187.
7. Potter, M. (1972) *Physiol. Rev.* 52, 631–719.
8. Wu, T. & Kabat, E. (1970) *J. Exp. Med.* 132, 211–250.
9. Capra, D. & Kehoe, M. (1974) *Proc. Nat. Acad. Sci. USA* 71, 845–848.
10. Poljak, R., Anzel, L., Avey, H., Chen, B., Phizackerley, R. & Saul, F. (1973) *Proc. Nat. Acad. Sci. USA* 70, 3305–3310.
11. Padlan, E., Segal, D., Cohen, G. & Davies, D. (1974) in *The Immune System: Genes, Receptors, Signals*, eds. Sercarz, E. & Williamson, A. (Academic Press, New York & London), pp. 7–14.
12. Weigert, M., Cesari, I., Yonkovich, S. & Cohn, M. (1970) *Nature* 228, 1045–1047.
13. Cesari, I. & Weigart, M. (1973) *Proc. Nat. Acad. Sci. USA* 70, 2112–2116.
14. McKean, D., Potter, M. & Hood, L. (1973) *Biochemistry* 12, 749–759.
15. McKean, D., Potter, M. & Hood, L. (1973) *Biochemistry* 12, 760–771.
16. Hood, L., McKean, D., Farnsworth, V. & Potter, M. (1973) *Biochemistry* 12, 741–749.
17. Carson, D. & Weigert, M. (1973) *Proc. Nat. Acad. Sci. USA* 70, 235–239.
18. Wofsy, L. & Burr, B. J. (1969) *J. Immunol.* 103, 380–382.
19. Grey, H. M., Hirst, J. W. & Cohn, M. (1971) *J. Exp. Med.* 133, 289–304.
20. Potter, M. (1967) in *Methods in Cancer Research*, ed. Busch, H. (Academic Press, Inc., New York), Vol. II, pp. 106–157.
21. Bridges, S. H. & Little, J. R. (1971) *Biochemistry* 10, 2525–2530.
22. Smithies, O., Gibson, D., Fanning, E., Goodfleish, R., Gilman, J. & Ballantyne, D. (1971) *Biochemistry* 10, 4912–4921.
23. Capra, J. D., Wasserman, R. & Kehoe, J. M. (1973) *J. Exp. Med.* 138, 410–427.
24. Barstad, P., Rudikoff, S., Potter, M., Cohn, M., Konigsberg, W. & Hood, L. (1974) *Science* 183, 962–964.
25. Milstein, C., Brownlee, G., Harrison, T. & Mathews, M. (1972) *Nature New Biol* 239, 117–120.
26. Gally, J. (1973) in *The Antigens*, ed. Sela, M. (Academic Press, New York & London), pp. 161–298.
27. Hood, L., Eichmann, K., Lackland, H., Kause, R. & Ohms, J. (1970) *Nature* 228, 1040–1044.
28. Dayhoff, M. (1972) *Atlas of Protein Sequence and Structure* (Nat. Biomed. Res. Found., Silver Spring, Md.).
29. Francis, S., Leslie, R., Hood, L. & Eisen, H. (1974) *Proc. Nat. Acad. Sci. USA* 71, 1123–1127.
30. Bourgois, A., Fougereau, M. & De Preval, C. (1972) *Eur. J. Biochem.* 24, 446–455.
31. Rudikoff, S., Mushinski, E., Potter, M., Glaudemans, C. & Jolly, M. (1974) *J. Exp. Med.*, in press.
32. Amzel, L. M., Poljak, R. J., Saul, F., Varga, J. M. & Richards, F. F. (1974) *Proc. Nat. Acad. Sci. USA* 71, 1427–1430.

Additional Data

We have determined the heavy chain variable region sequences from two molecules which bind phosphorylcholine, H8 and S107. In collaboration with Rudikoff, we have also determined the variable region of the T15 heavy chain. In addition, we have determined the N-terminal sequences from M511 and W3207. These studies reveal that the heavy chains from PC-binding myeloma proteins represent multiple variable region genes. Two identical heavy chain variable regions, S107 and T15, occur within molecules with the T15 idiotype. A third molecule with this idiotype H8 has a single amino acid difference from T15, however. This difference indicates that serologic identity may not imply variable region sequence identity. The high degree of sequence homology observed within these total heavy chain variable regions supports the concept of heavy chain sequence importance in determining the hapten-binding activity of the intact molecule.

Materials and Methods

H8 and T15 tumors were obtained from Dr. S. Rudikoff, NIH, Bethesda, Maryland. Tumors W3207, S107 and M511 were obtained from the Salk Institute, San Diego, California. All solid tumors were converted to ascites form by passage through an 80 mesh screen and injection intraperitoneally into (DBA/BALB) F_1 mice. Ascites fluid was collected after

three weeks by paracentesis with an 18 gauge needle. Each paracentesis yielded from 5 to 15 ml of fluid.

Purification of Immunoglobulins. Phosphorylcholine specific immunoglobulin from each tumor was purified by affinity chromatography according to Chesebro and Metzger (26).

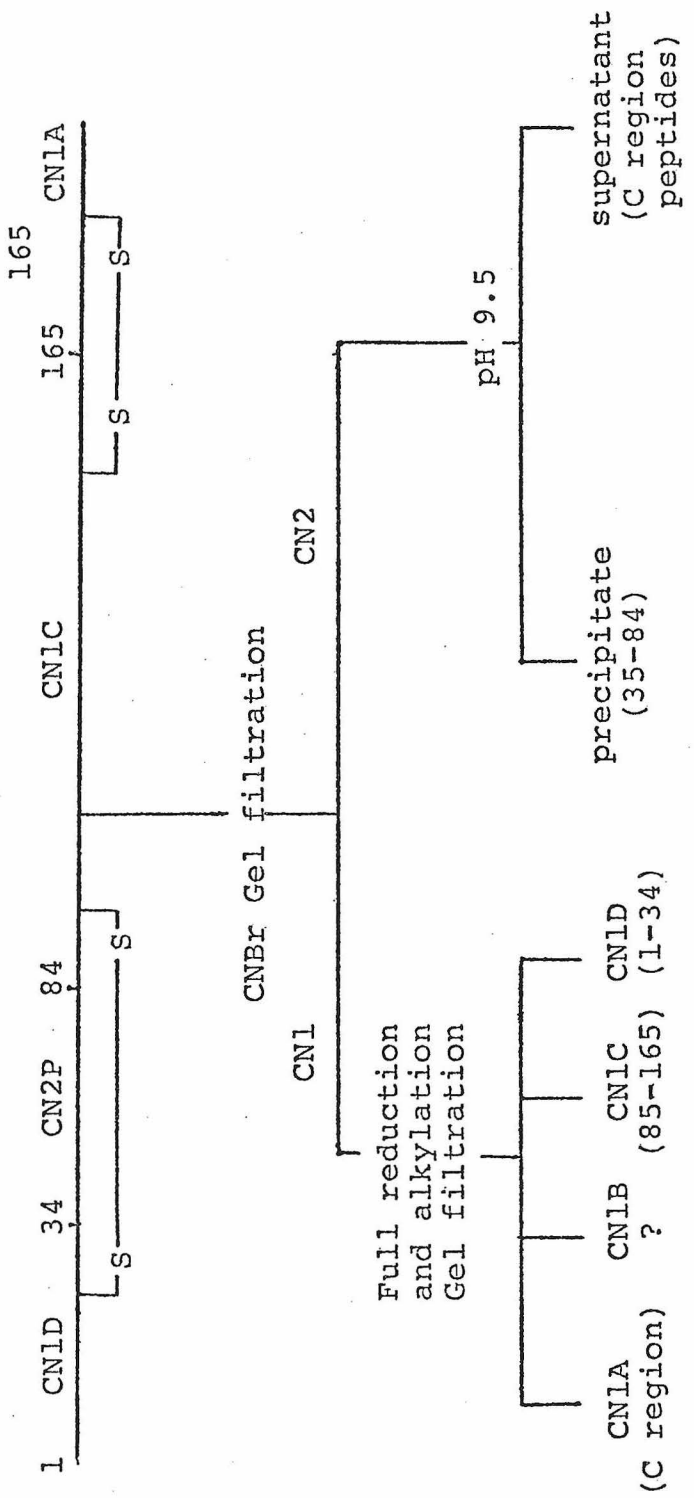
Chain Separation. Heavy and light chains were separated according to (27) on an upward flow 5 x 150 cm G-150 column in 5 M GuHCl buffer in .2 N NH_4HCO_3.

Preparation of Cyanogen Bromide Fragments. The generation and isolation of cyanogen bromide fragments as outlined in Fig. 3 was done according to Rudikoff and Potter (27) with the following changes: Fragment CN2 was generated from intact heavy chain. The CN2 was desalted on a Biogel P-2 column (6 x 50 cm) in 5% formic acid. The precipitation of CN2P (residues 85-165) was carried out by dissolving CN2 in .01 N acetic acid and changing the pH to 9.5 by the addition of 1 N NH_4OH. This procedure gave a better yield of CN2P in our hands than the technique previously published.

Sequence Analysis. Approximately one micromole of each peptide was subjected to 36 or 45 cycles of automated Edman degradation on an updated Beckman 890 A protein/peptide sequenator using the DMBA program of Hermodson et al (22). Each residue was analyzed as in (20).

Trypsin Digest of CN2P. Approximately 150 nm of CN2P was dissolved in 200 µl of .01 N acetic acid. Two nm trypsin (Calbiochem grade B) was added in 20 µl 10^{-3} N HCl. Two hundred

Fig. 3.--Protocol for isolation of variable region cyanogen bromide peptides of heavy chains from H8, T15, and S107. Numbers over line indicate methionine locations. For details, see text. Fragments of interest in this paper are CN1C (residues 85-165), CN1D (residues 1-34) and CN2P (residues 35-84). The supernatant fraction from CN2 contains two or more constant region peptides, CN1A is a large constant region peptide beginning at position 166 and CN1B appears to be a complex mixture of peptides generated in low yield.


1
CN1D
34
CN2P
84
CN1C
165
CN1A
165
S
S
S
S
CNBr
Gel filtration
CN2
CN1
Full reduction
and alkylation
Gel filtration
pH 9.5
precipitate
(35-84)
supernatant
(C region
peptides)
CN1A
(C region)
CN1B
?
CN1C
(85-165)
CN1D
(1-34)

µl of .2 N NH_4HCO_3 was added to start the digest and caused most of the CN2P to precipitate. The digest was incubated for one hour at 37°C with frequent vortexing. An additional 2 nm trypsin was then added and the vortexing and incubation were continued for one more hour after which the solution was frozen and lyophilized.

Peptide Maps of CN2P. The lyophilized digest was dissolved in .1 N acetic acid and 1/3 was placed on the analytical map and 2/3 on the preparatory map. Descending chromatography was performed in water saturated butanol:acetic acid (540:160) for 18 hours on 46 x 57 cm Whatman 3 mm chromatography paper. After drying at 80°C for 15 minutes, electrophoresis was run at pH 3.5 in pyridine:acetic acid:water (1:10:290) for one hour at 3000 V. After drying again the analytical map was dipped in a collidine-ninhydrin solution consisting of 1 g of ninhydrin in ethanol:acetic acid:collidine (600 ml:200 ml: 80 ml). Peptides were visible after 15 minutes at 80°C and were cut out, eluted in 1 N NH_4OH, dried under nitrogen, and hydrolyzed for 15 minutes at 160°C in 12 N HCl propionic acid (1:1) (28). The NH_4OH reacts with excess ninhydrin making it possible to obtain peptide compositional data from an analytical map (E. Loh, unpublished observation). Following hydrolysis, the composition of each peptide was determined by electrophoresis for two hours at pH 1.7 (29).

Preparative maps were sprayed with 1% ninhydrin in acetone (Pierce Chemical Co.) and the peptides were cut out and eluted

in 1 NH_4OH as soon as they became visible. Hydrolysis on the peptides of interest was done at 110°C in 6 N HCl for 24 hours. The amino acid compositions for these peptides were determined on a Durrum D-500 amino acid analyzer using a 72-minute program.

Results and Discussion

The amino acid sequences for the entire heavy chain variable regions of four myeloma proteins which bind PC are shown in Fig. 4 along with portions of two others. These variable regions are highly similar in sequence. This 90% or greater sequence homology within this set of heavy chains correlates with the concept that the heavy chain structure is important in determining the hapten-binding specificity of the molecule.

The N-terminal 20 residue sequences for the heavy chains W3207 and M511 are identical to T15. The T15 branch on the genealogical tree now represents six proteins, all of which were derived from tumors which bind PC.

Molecules which are serologically identical can differ in their amino acid sequence. Three of the heavy chains in Fig. 4, S107, T15, and H8, are from parent molecules which all have the T15 idiotype, (15) i.e., they are serologically indistinguishable. Two of these heavy chains, S107 and T15, are identical for their entire variable region amino acid sequences. As indicated in Fig. 5, however, sequence analysis shows that H8 differs from T15 and S107 at position 105. These data were obtained following purification and automated sequence analysis on peptides CN1C (residues 85-165) from all three

Fig. 4.--The total heavy chain variable region sequences from four BALB/c phosphorylcholine-binding myeloma proteins. W3207, M511, S107, and H8 are from this thesis. T15 was sequenced in collaboration with Rudikoff. M603 is from (27). Horizontal lines indicate homology with T15. Letters indicate the appropriate amino acid using the single letter amino acid code: A, alanine; B, aspartic acid or asparagine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; and Z, glutamine or glutamic acid. { } denotes a gap. ( ) indicates sequence by peptide composition homology only.

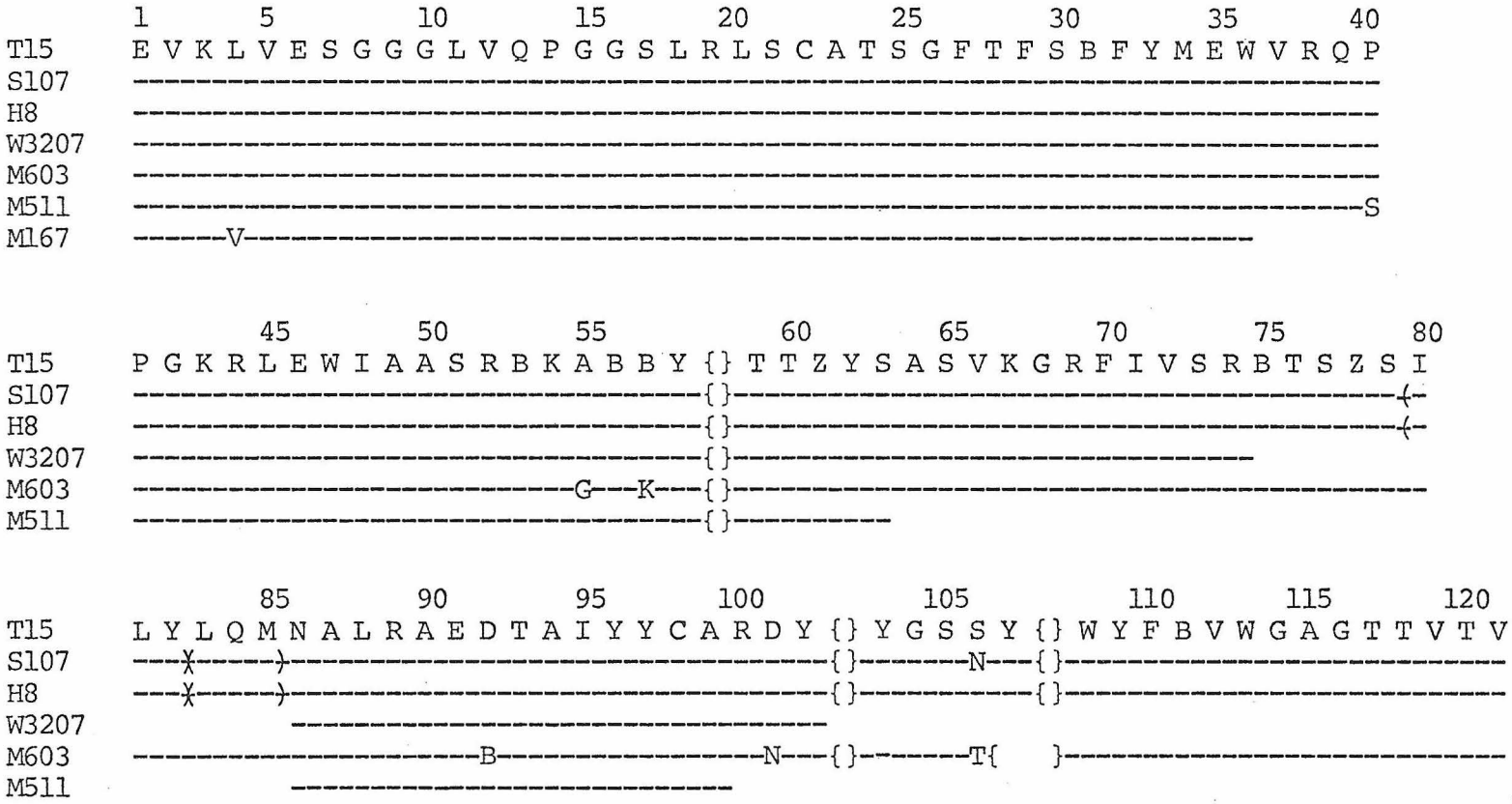

1 5 10 15 20 25 30 35 40
T15 E V K L V E S G G G L V Q P G G S L R L S C A T S G F T F S B F Y M E W V R Q P
S107
H8
W3207
M603
M511 S
M167 V
45 50 55 60 65 70 75 80
T15 P G K R L E W I A A S R B K A B B Y {} T T Z Y S A S V K G R F I V S R B T S Z S I
S107 {}
H8 {}
W3207 {}
M603 G K {}
M511 {}
85 90 95 100 105 110 115 120
T15 L Y L Q M N A L R A E D T A I Y Y C A R D Y {} Y G S S Y {} W Y F B V W G A G T T V T V
S107 {} N {}
H8 {} {}
W3207
M603 B N {} T{ }
M511

Fig. 5.--Mole percent data from the Durrum D-500 analyzer of steps 20, 21, and 22 from CN1C of T15, H8, and S107. The residue indicated by gas chromatography is shown on the line marked G.C. In HI hydrolysis PTH-serine is converted to alanine in approximately 50% yield (25). Prior to hydrolysis of the PTH's back to amino acids, the asx was resolved to be asn by thin-layer chromatography (24). The major residue at each position is underlined.

	T15			H8			S107		
	20	21	22	20	21	22	20	21	22
Asx	7.3	6.6	8.2	6.7	24.1	13.1	6.5	6.0	6.6
Glx	3.5	3.4	3.5	4.0	4.0	4.0	3.1	4.8	3.8
Pro	4.2	6.4	6.0	9.0	7.7	10.6	11.6	11.6	13.6
Gly	19.3	12.4	9.7	14.3	9.6	8.2	30.0	21.0	15.4
Ala	19.4	28.1	18.3	33.3	19.6	14.9	20.5	26.6	21.0
ABA	4.6	4.1	4.0	4.6	4.5	4.1	2.3	2.0	1.8
Val	5.0	3.9	3.7	3.7	3.5	3.1	1.6	1.8	2.2
AIle & Ile	5.6	4.6	4.1	4.4	4.3	4.2	3.5	3.2	3.1
Leu	5.0	5.3	4.6	8.4	6.7	5.7	5.8	4.8	4.6
Nor	13.9	10.2	11.3	4.5	13.2	13.4	2.8	3.6	3.6
Tyr	7.5	10.1	21.7	6.4	6.0	17.1	7.5	7.2	14.1
Phe	1.9	2.2	2.1	1.0	1.0	.5	3.0	2.8	4.8
Lys	.9	.7	.5	0	.9	0	0	0	0
G.C.	S	S	Y	S	?	Y	S	S	Y

heavy chains. Clearly position 21 in H8 CN1C (residue 105) is an asparagine in contrast to the serine found at this position in both T15 and S107. Each sample was sequenced twice to verify this finding.

The observation of a single amino acid substitution within molecules with the T15 idiotype indicates that the anti-idiotypic antisera is not absolutely specific. It did not detect the single heavy chain variable region difference and thus genetic implications drawn from serological studies must be made cautiously. In view of the fact that the light chains from T15 and H8 appear to be identical for the first 93 residues (this thesis) the intact immunoglobulins are extremely similar in structure and it is not surprising that serologically they are indistinguishable.

The genetic mechanism responsible for the H8 difference could be one of three alternatives. Since the T15 and S107 heavy chain variable region sequences are identical, they appear to be encoded in the BALB/c germ line DNA. Three alternative mechanisms could explain the single base substitution which would generate the H8 sequence: (1) H8 could represent a second germ line gene generated by a gene duplication of the T15 gene which has subsequently accepted a single base substitution, or (2) it could reflect a polymorphism of the T15 heavy chain gene which exists in a subset of the BALB/c mice, or (3) it could have been generated by mutation in the soma of a T15-producing cell which was subsequently transformed. H8

does not have a higher binding affinity (15) for PC than T15, thus this mutation does not appear to have a selective advantage, and therefore, it does not appear to be a case of antigen-driven somatic mutation (9). Further sequence analysis of other heavy chains from molecules with the T15 idiotype will determine if the H8 sequence is found again. If this sequence recurs, the third explanation would appear unlikely.

The PC-binding myeloma proteins must represent four germ line V_H genes. Since S107 and T15 have identical V_H regions and H8 differs by only a single base mutation, any somatic mutational mechanism responsible for additional diversity which might exist must operate at a very low level (if at all) in these cases. M167 (30), M511 (31), and M603 all differ from one another as well as from T15 by multiple amino acid substitutions and insertions or deletions (Fig. 6). If these sequences were all generated from the T15 germ line gene, the mechanism responsible for generating these differences must be very complicated. First of all it must be capable of operating at different rates in different cells since the T15 - S107 sequences are identical while M167 differs by ten substitutions and three size differences. Do different rates of mutation represent different enzymes or cofactors? Second, the mechanism also must operate in the framework as well as the hypervariable regions since the differences are located in both areas. Third, it must be able to break DNA at any of three different points in the variable region, insert codons

Fig. 6.--Comparison of the amino acid sequences and lengths of the variable regions from the heavy chains of phosphorylcholine-binding myeloma proteins. The line marked substitutions indicates the number of amino acid differences from T15. The line marked length compares the length of each V_H with that of T15, thus M603 is one residue shorter than T15, while M511 is one residue longer. M511 is from (31) and M167 from (30).

	S107	H8	M603	M511	M167
Substitutions	0	1	4	3	10
Length	0	0	-1	+1	+3

and splice the DNA back together in phase each time such that a functional molecule will be produced. Such a mechanism would be extremely complex and appears unlikely. It seems more reasonable to believe that T15, M167, M511, and M603 represent four different germ line V_H genes.

The N-terminal differences between V_H regions from normal antibodies and myeloma proteins must be due to the antigenic environments involved. A recent report (32) indicated that human IgA heavy chains are 67% unblocked compared to 25% for those from IgG. While this class difference in V_H expression is suggestive, it does not explain the large percentage of unblocked V_H regions found in our study of myeloma proteins. Our study contained three IgG molecules and all three are unblocked. Since each V_H from a mouse IgG molecule has only a .22 probability of being unblocked if the IgG myelomas and IgG normal antibodies express the same V_H regions, the probability of finding all three IgGs unblocked in our study would be $(.22)^3$ or about 1 in 100. Thus, the classes of the immunoglobulins involved does appear to adequately explain the difference in the fraction of unblocked heavy chains observed in the two studies.

A more reasonable explanation of this difference is that the antigens expressed in the gut and circulation differ and since the hapten-binding activity of the molecule correlates with the V_H N-terminal sequences, the V_H sequences observed should also differ. Heavy chains expressed in specificities

for gut flora may be predominantly unblocked while those specific for circulating antigens are mostly blocked.

A recent finding by Capra (13) supports this concept that the V_H sequence profile reflects the antigenic history of the organism. He reported that the normal IgG response to p-azophenylarsonate has unblocked V_H regions in the BALB/c mouse. Furthermore, these heavy chains are identical in sequence to our myeloma MV_{HII} subset for the first 20 residues. This sequence, which differs by 40% from the single sequence reported at the 95% level or greater, was undetected in the original study on unimmunized animals (33). Upon reexamination of the original data, this MV_{HII} sequence has been detected at the 5% level or less. This illustrates that sequence studies on pooled antibodies detect only a few major sequences at best and upper limits for total subunit diversity based upon these studies are of questionable value.

The N-terminal 20 residues appear to be the most highly conserved block in the entire V_H region. Four nearly total BALB/c myeloma V_H regions are compared in Table 2. The percent of sequence homology between each pair of V regions is indicated for the N-terminal 20 residues and the total variable region. For the purpose of this comparison, a deletion is counted as a single difference for each amino acid missing. In every comparison, the observed homology is greater for the N-terminal 20 residues than for the total variable regions. In examining all the total BALB/c V_H regions sequences to date, the N-terminal

Table 2.--Percent homology between four BALB/c heavy chain variable regions. The number above the line indicates the percent homology between the two V_H regions in the first 20 residues. The number below the line indicates total V_H homology. Total V_H homology is calculated from 116 residues for all chains except for T173 which employed only 104 residues. M21 is from (34), 315 is from (35), T173 is from (36), and T15 is from this thesis in collaboration with S. Rudikoff.

TABLE 2

	M315	T173	M21
T15	$\frac{60}{41}$	$\frac{80}{55}$	$\frac{80}{59}$
M315		$\frac{50}{40}$	$\frac{70}{41}$
T173			$\frac{70}{63}$

section is the most highly conserved 20 residue block in the entire variable region.

The heavy chain diversity must be analogous to that of the kappa chains. This N-terminal study examined the most highly conserved portion of a selected subset of heavy chain variable regions and yet considerable diversity in the V_H regions is evident. The sequence variation observed in the total BALB/c V_H regions appears comparable to that present in the BALB/c kappa chains. The BALB/c mouse thus appears to have two very diverse V region families available for the generation of antibody specificities. By combining various V_H and V_L regions to form different hapten-binding specificities, an organism could generate vast antibody diversity from a few thousand V region genes.

CHAPTER V

The N-Terminal Diversity of Myeloma Proteins Which Bind Phosphorylcholine

The publication in this chapter is from Science and is reproduced with the permission of the publisher.

Reprinted from
March 1974, Volume 183, pp. 962-964

SCIENCE

Immunoglobulin Structure: Amino Terminal Sequences of Mouse Myeloma Proteins That Bind Phosphorylcholine

P. Barstad, S. Rudikoff, M. Potter, M. Cohn, W. Konigsberg and L. Hood

Immunoglobulin Structure: Amino Terminal Sequences of Mouse Myeloma Proteins That Bind Phosphorylcholine

Abstract. *The amino terminal sequences of five light and heavy immunoglobulin chains from myeloma proteins of the BALB/c mouse with binding activity to phosphorylcholine are presented. Except for a single substitution in position 4, all five heavy chains have identical amino terminal sequences through the first hypervariable region. Proteins which share unique (idiotypic) antigenic determinants are identical through the first hypervariable region of their light and heavy chains. Proteins with differing idiotypic determinants have light chains of differing amino acid sequence. These observations suggest that the heavy chain plays a more important role than the light chain in determining the phosphorylcholine binding site.*

For the past several years, myeloma immunoglobulins produced by plasmacytomas in the BALB/c mouse have been screened against a limited series of antigens (*1*). Approximately 5 percent of these proteins have been shown to exhibit specific binding activity. In many instances it has been possible to assign this binding activity to chemically defined haptens. Indeed, groups of proteins have emerged which bind the same haptenic determinant, for

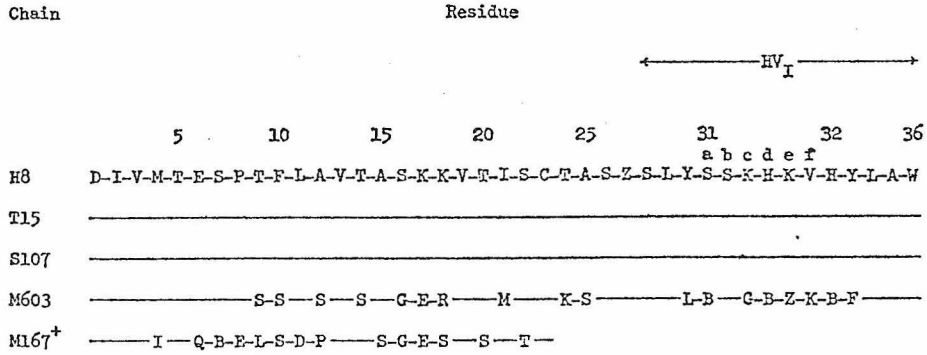


Fig. 1. The amino terminal sequences of κ chains from myeloma proteins with binding activity to phosphorylcholine. HV_I indicates the extent of the first hypervariable region; H8 indicates HOPC 8; T15 indicates TEPC 15; M603 indicates MOPC 603; and M167 indicates MOPC 167 [see (*14*)]. The one-letter amino acid code is: glycine, G; alanine, A; valine, V; leucine, L; isoleucine, I; serine, S; threonine, T; proline, P; cysteine, C; methionine, M; histidine, H; lysine, K; arginine, R; aspartic acid, D; glutamic acid, E; asparagine, N; glutamine, Q; aspartic acid or asparagine, B; glutamine or glutamic acid, Z; tyrosine, Y; phenylalanine, F; tryptophan, W. The numbering of residues for these chains is that taken from a homologous mouse κ chain, MOPC 41 (*19*).

example, dinitrophenol (*2*), phosphorylcholine (*3–5*), α-(1 → 3)dextran (*6*), β-(1 → 6)-D-galactan (*7*), α-(1 → 6)dextran (*8*), β-(2 → 1)fructosan (*8*), and β-(2→ 6) fructosan (*8*). Studies of the primary structure of proteins that bind the same hapten will be useful for delineating structure-function relationships as well as providing possible insights into the genetic mechanism of antibody diversity.

In this study we compare the partial amino acid sequences of light (L) and heavy (H) chains from five BALB/c myeloma proteins that bind phosphorylcholine. Previous studies have indicated that three of the five phosphorylcholine-binding proteins, H8, T15, and S107, bind identical groups of related antigens (*9*). The other two proteins, M603 and M167, bind groups of related antigens which distinguish them from each other and from those of the first group. In addition, immunologic studies have demonstrated that H8, T15, and S107 possess the same individual antigenic specificity (idiotype) suggesting a high degree of structural identity (*5*). [These proteins are members of the S63-T15 idiotypic group (*10*).] In contrast, the other two proteins, M603 and M167, have unique antigenic determinants, which differentiate them from each other and from the group with the shared idiotype (*5*). We were interested in examining the variable regions of the light (V_L) and heavy (V_H) chains from these proteins in order to determine how their amino acid sequences correlated with the antigen binding and idiotypic properties.

The amino acid sequence analyses of partially reduced and alkylated immunoglobulin chains were carried out on a Beckman model 890A or 890C sequencer with the use of standard buffers. The phenylthiohydantoin amino acid derivatives were analyzed by gas chromatography, by thin-layer chromatography, and by amino acid analysis after hydrolysis of the derivatives to free amino acids (*11*). At least two sequenator runs were carried out on each polypeptide chain.

The partial amino acid sequences for L chains from phosphorylcholine-binding immunoglobulins H8, T15, S107, M167, and M603 are given in Fig. 1. Each of these L chains is of the kappa (κ) type. Several important points can be derived from an analysis of these L chain data. It should be noted that L chains have three regions, residues 28 to 34, 50 to 56, and 90 to

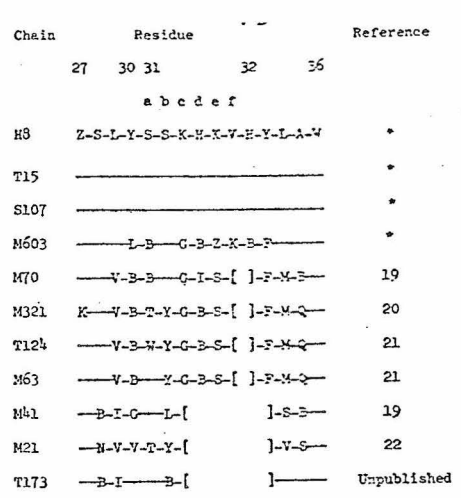


Fig. 2. The first hypervariable region from myeloma κ chains of the BALB/c mouse. The asterisk indicates the data presented in this report.

97, which are far more variable than the remainder of the V region (*12*). These are termed hypervariable regions. Indeed, these regions are so variable that two randomly chosen myeloma L chains, from man or mouse, have a very low probability of showing identity in any one of the hypervariable regions (*13*). Thus it is striking that the L chains from the three proteins demonstrating idiotypic identity, H8, T15, and S107, are identical for their first 41 residues which include the first hypervariable region (Figs. 1 and 2). The M603 L chain differs from the H8 group by 18 out of 41 residues, 8 of which are in the first hypervariable region. The M167 L chain differs from the H8 group at 14 out of the first 23 residues, none of which are in the first hypervariable region (*14*). Thus at least three L chain sequences are present among the five immunoglobulins with phosphorylcholine-binding activity. The first hypervariable region of the four L chains examined is two residues longer than that of any of the mouse κ chains previously reported (Fig. 2). This may reflect a common structural requirement for phosphorylcholine-binding L chains.

The partial amino acid sequences for H chains from phosphorylcholine-binding immunoglobulins H8, T15, S107, M603, and M167 are given in Fig. 3. The amino terminal sequences from these H chains are identical, apart from a single base substitution, leucine to valine, at position 4 in M167. Heavy chains also have hypervariable regions whose properties are similar to those described for L chains. The first hypervariable region of H chains appears to extend from positions 27 to 35 as indicated in Fig. 3. Perhaps the most striking finding in this study is the identity of all five H chains in the first hypervariable region, even though they were derived proteins with three different idiotypes. This observation assumes added importance in that the hypervariable regions appear to be in or near the antigen combining site as judged by affinity label and x-ray crystallographic studies (*15*). Thus these regions may play a particularly critical role in determining the three-dimensional configuration of the combining site. Indeed, the H chain sequence characteristic of the phosphorylcholine-binding myelomas has not been seen in any of the 23 other BALB/c V_H sequences that were derived from immunoglobulins lacking this activity analyzed to date (*16*).

This is, to our knowledge, the first report in which hypervariable regions from both the L and H chains of independently arising immunoglobulins with idiotypic identity have been shown, respectively, to be identical. Thus the results of this partial amino acid sequence analysis and some other studies on the remainder of the V regions are consistent with the supposition that immunoglobulins with identical idiotypic specificities are identical in the

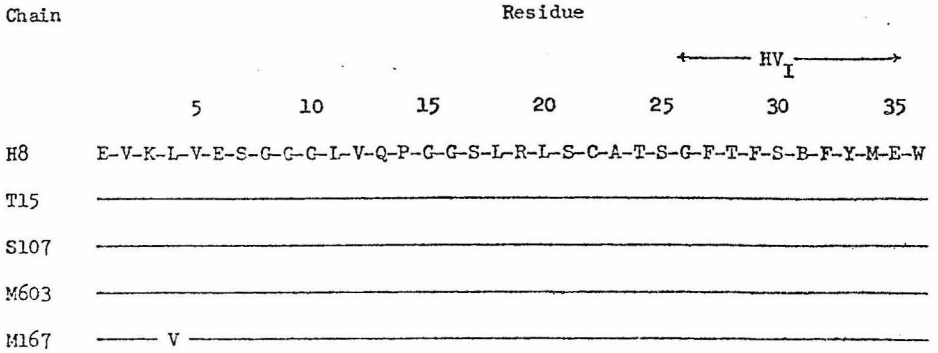


Fig. 3. The amino terminal sequences of H chains from myeloma proteins with binding activity to phosphorylcholine. HV_I indicates the span of the first hypervariable region.

V region sequences of their L and H chains.

If identical idiotypic specificities indicate structural identity, then there are two indications that immunoglobulins of the S63 idiotype are present in normal BALB/c mice. First, immunization of normal BALB/c mice against the phosphorylcholine determinant yields immunocytes producing specific antibody that can be inhibited in the Jerne plaque assay by idiotypic antiserum of the S63 type (*17*). Second, immunoglobulins of the S63 idiotype are found in the normal serum of BALB/c mice (*17a*). These results suggest that the myeloma proteins with phosphorylcholine-binding activity are indeed an excellent model system for studying the relation between structure and function in bona fide antibody molecules.

Since the H chains, even from those proteins with differing idiotypic specificity, are identical for 36 residues, except for a single amino acid substitution, it will be particularly interesting to determine whether the entire H chains from proteins of differing idiotypic specificity are identical (for example, T15 and M603). If so, perhaps this H chain sequence, or one very closely related to it, is a prerequisite for all immunoglobulins with phosphorylcholine-binding activity. It has been shown that all myeloma proteins with α-(1 → 3)dextran activity have an identical lambda L chain (*18*). Proteins with β-(1 → 6)-D-galactan binding activity have very similar sequences in both L and H chains (*7*). Thus in some instances either L or H chain may play a dominant role in hapten binding, while in others a specific L and H pair may be rigidly required.

If immunoglobulins of the S63 idiotype are identical, this suggests that the corresponding V_L and V_H regions are coded by genes in the germ line of the BALB/c mouse, for it is unlikely that a random somatic recombination or somatic mutation process would generate multiple identical antibodies in different individuals. In this regard, it will be interesting to determine whether the H chain from M167 is identical to the others apart from a single substitution outside the hypervariable region, or whether other differences will be found. If a single substitution is present, it could be explained by a single base substitution, either in the germ line or in the soma.

The complete analysis of these proteins as well as those from groups with binding activity to other haptens should continue to provide insights into the structure, genetics, regulation, and evolution of immunoglobulin molecules.

P. BARSTAD
Division of Biology, California Institute of Technology, Pasadena 91109

S. RUDIKOFF, M. POTTER
National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014

M. COHN
Salk Institute for Biological Studies, San Diego, California 92112

W. KONIGSBERG
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06510

L. HOOD
Division of Biology, California Institute of Technology

References and Notes

1. M. Potter, *Fed. Proc.* 29, 85 (1970).
2. H. N. Eisen, E. S. Simms, M. Potter, *Biochemistry* 7, 4126 (1968); B. M. Jaffe, H. N. Eisen, E. S. Simms, M. Potter, *J. Immunol.* 103, 872 (1969).
3. M. Cohn, *Cold Spring Harbor Symp. Quant. Biol.* 32, 211 (1967); M. Potter and M. A. Leon, *Science* 162, 369 (1968).
4. A. Sher, E. Lord, M. Cohn, *J. Immunol.* 107, 1226 (1971); M. A. Leon and N. M. Young, *Biochemistry* 19, 1424 (1971).
5. M. Potter and R. Lieberman, *J. Exp. Med.* 132, 737 (1970).
6. M. Weigert, M. Cesari, S. J. Yonkovich, M. Cohn, *Nature* 228, 1045 (1970); M. A. Leon, N. M. Young, K. R. McIntire, *Biochemistry* 9, 1023 (1970).
7. M. Potter, E. B. Mushinski, C. P. J. Glaudemans, *J. Immunol.* 108, 295 (1972); S. Rudikoff, E. B. Mushinski, M. Potter, C. P. J. Glaudemans, M. E. Jolley, *J. Exp. Med.*, in press.
8. J. O. Cisar, E. A. Kabat, M. G. Weigert, M. Cohn, M. Potter, *Fed. Proc.* 32, 957 (1973).
9. M. Potter, *Ann. N.Y., Acad. Sci.* 190, 306 (1970).
10. The S63-T15 group of phosphorylcholine-binding myeloma proteins is named after the S63 plasmacytoma induced at the Salk Institute and the TEPC 15 plasmacytoma induced at NIH, which have the same idiotype. Other proteins with the identical idiotype are MOPC 299 (NIH), HOPC 8 (NIH), S107 (Salk), Y5170 (Salk), and Y5236 (Salk) (*4*).
11. L. Hood, D. McKean, V. Farnsworth, M. Potter, *Biochemistry* 12, 741 (1973); J. Pisano and T. Bronzert, *J. Biol. Chem.* 244, 5597 (1969); O. Smithies, D. Gibson, E. M. Fanning, R. M. Goodfliesch, J. G. Gilman, D. M. Ballantyne, *Biochemistry* 10, 4912 (1971).
12. T. T. Wu and E. A. Kabat, *J. Exp. Med.* 132, 211 (1970).
13. G. Smith, L. Hood, W. Fitch, *Annu. Rev. Biochem.* 40, 696 (1971).
14. The sequence for the amino terminal 23 residues of the L chains from H8, T15, and M603 has been reported [L. Hood, D. McKean, M. Potter, *Science* 170, 1207 (1970)].
15. H. Eisen, in *Progress in Immunology*, B. Amos, Ed. (Academic Press, New York, 1971), p. 243; R. J. Poljak, L. M. Anzel, H. P. Avey, B. L. Chen, R. P. Phizackerley, F. Saul, *Proc. Natl. Acad. Sci. U.S.A.* 70, 3305 (1973).
16. P. Barstad, M. Weigert, M. Cohn, L. Hood, in preparation.
17. A. Sher and M. Cohn, *Eur. J. Immunol.* 2, 319 (1972); H. Cosenza and H. Kohler, *Proc. Natl. Acad. Sci. U.S.A.* 69, 2701 (1972).
17a. M. Potter and R. Lieberman, unpublished observations.
18. M. Weigert, M. Cesari, S. J. Yonkovich, M. Cohn, *Nature* 228, 1045 (1970).
19. W. R. Gray, W. J. Dreyer, L. Hood, *Science* 155, 465 (1967).
20. D. McKean, M. Potter, L. Hood, *Biochemistry* 12, 749 (1973).
21. ———, *ibid.*, p. 760.
22. C. Milstein and J. Svasti, in *Progress in Immunology*, B. Amos, Ed. (Academic Press, New York, 1971), p. 33.

16 November 1973

Additional Data

We have determined the amino acid sequences of the N-terminal regions of the light and heavy chains from two additional BALB/c phosphorylcholine (PC) binding myeloma proteins, W3207 and M511. In addition, we have extended the sequences of the H8 and T15 light chains to position 94. The N-terminal heavy chain sequences for both W3207 and M511 are identical to T15 through the first hypervariable region. W3207 light chain is highly similar to M603 light chain and M511 light chain closely resembles M167. The present data suggest that three subspecificities exist within the PC-binding myeloma proteins. These specificities have highly similar heavy chain sequences but vastly different light chains.

Methods and Materials

Sequence analysis was performed as in using the DMBA program of Hermodson et al. (22).

Preparation of T15L trypsin fragments. Fifty mg of T15 light chain was succinylated according to (37). The light chain was then fully reduced and alkylated according to (27). Trypsin digestion and gel filtration were similar to that of Svasti et al. (38) with the exception that a G-100 column (2.5 x 150 cm) in 1N NH_4OH was used. The first major peak contained two fragments in about equimolar concentration and

these were sequenced as a mixture. Since one of the fragments was from the constant region (positions 109 to 155) it was possible to deduce the second sequence by subtracting the known residue at each step.

Preparation of H8 light chain tryptic fragments. Approximately 100 mg of H8 light chain was fully reduced and alkylated (27). After this treatment, the light chain was marginally soluble in the buffer used for the trypsin digest. Digestion was carried out in 5 ml .2N NH_4HCO_3 using two 1% aliquots of trypsin (Calbiochem grade B) added at one-hour intervals. After two hours, the digestion was centrifuged and the supernatant removed. The pellet was resuspended in .1N acetic acid and centrifuged again. The acetic acid supernatant fraction was brought to pH 9.5 with .1N NH_4OH at which time the desired peptide precipitated from solution along with the N-terminal fragment. The precipitate was sequenced as a mixture of three components.

Results and Discussion

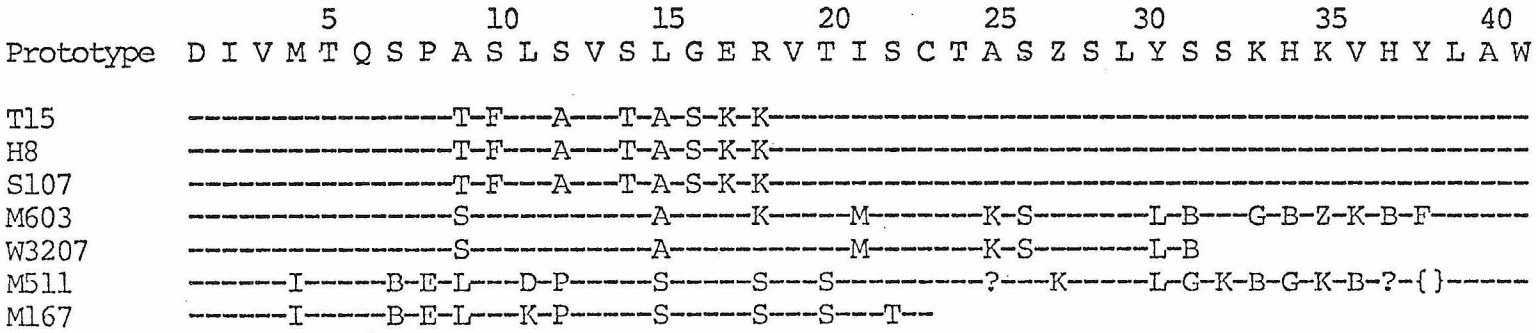
The heavy chain sequences for W3207 and M511 are identical to T15 for the first 35 residues. The extreme sequence correlation observed with the heavy chains from the first five PC-binding proteins applies to W3207 and M511. This is in agreement with the concept that the heavy chain sequence conservation is critical for the immunoglobulins which bind PC. Subsequent work reported in this thesis has shown that five of these heavy chains are unique in sequence, however. Thus, these heavy chains

represent a set of closely related molecules rather than the repeated expression of one variable region gene.

The light chains from the PC specificity now fall into one of three subsets. The partial amino acid sequences for the seven light chains from the PC-binding immunoglobulins are shown in Fig. 4. Three subsets can be delineated based upon their amino acid sequence similarity. The first subset contains molecules which have the T15 idiotype and these light chains are identical in sequence in the areas examined. A second subset contains light chains from two idiotypically distinct molecules, M511 and M167. These two light chains differ from each other by only two residues in the first 23, yet are vastly different from any of the other light chains in this study. A third subset contains M603 and W3207 light chains. These light chains differ at only a single position in the first 29 residues, but again differ greatly from the other of the seven light chains examined. Each of these subsets must represent at least one unique germ line gene, thus three germ line V_L genes must be expressed within the PC-binding BALB/c immunoglobulins.

Why do three light chain subsets exist? One explanation is that each subset represents immunoglobulins which bind to different natural antigens. Studies on the intact immunoglobulins indicate that they do indeed differ in their relative affinities for both synthetic (39) and bacterial (15) structures, although the binding to PC itself appears to be relatively

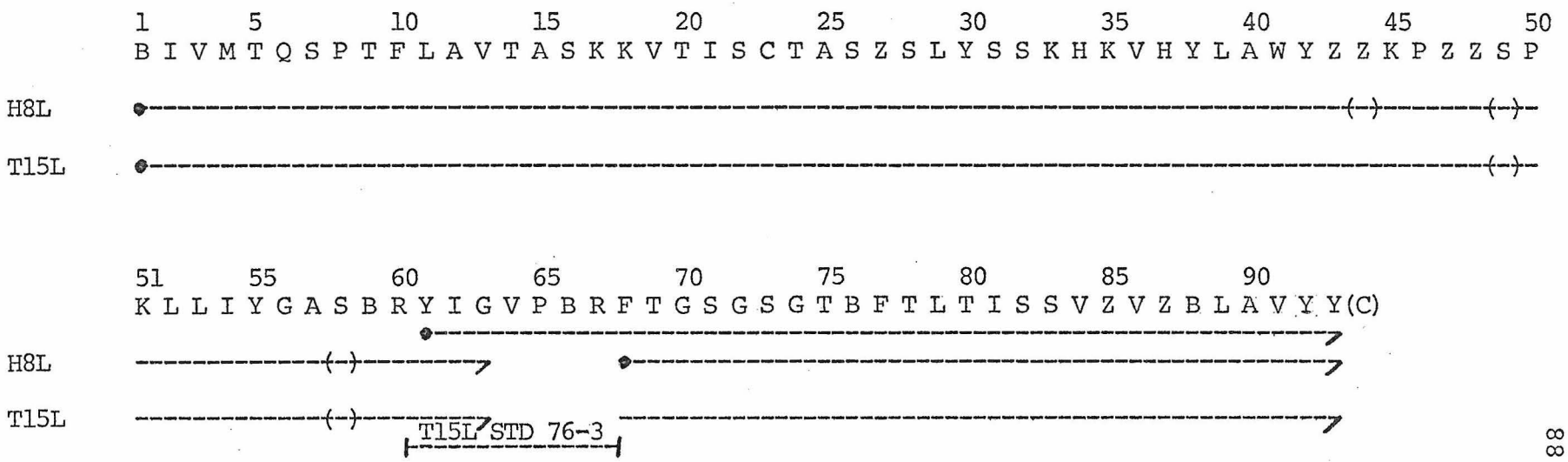
Fig. 4.--N-terminal sequences of seven light chains from molecules which bind phosphorylcholine. The prototype sequence was generated by taking the most common residue in all BALB/c kappa chain sequences available at each position. Lines indicate sequence identity with the prototype and letters indicate differences using the single letter amino acid code.


5 10 15 20 25 30 35 40
Prototype D I V M T Q S P A S L S V S L G E R V T I S C T A S Z S L Y S S K H K V H Y L A W
T15 ----------T-F---A---T-A-S-K-K----------
H8 ----------T-F---A---T-A-S-K-K----------
S107 ----------T-F---A---T-A-S-K-K----------
M603 ----------S----------A-----K-----M-------K-S-------L-B---G-B-Z-K-B-F------
W3207 ----------S----------A----------M-------K-S-------L-B
M511 ------I-----B-E-L---D-P-----S-----S---S---------?---K-----L-G-K-B-G-K-B-?-{}-----
M167 ------I-----B-E-L---K-P-----S-----S---S---T--

equal for most of the molecules (40). The PC-binding immunoglobulins may then represent molecules which bind to at least three different natural antigens, but which are all capable of reacting with phosphorylcholine.

Two light chains from molecules with the T15 idiotype are identical for the first 80% of their variable regions. Preliminary sequences for the first 93 residues of T15 and H8 light chains are shown in Fig. 5. The sequence for H8L could be determined since the mixture consisted of the N-terminal fragment plus peptides which began at 61 and 68. The sequence for the peptide beginning at position 68 could be confirmed by detection of the same residue seven steps later. Work is currently underway to separate these three fragments by gel filtration. In this case, light chains that are identical for the first 40 residues are identical for the majority of their variable regions. These two chains may represent repeated expression of the same V region gene. Since molecules with the T15 idiotype have been found seven times within the BALB/c myeloma proteins, a much higher frequency than any other known idiotype, the myeloma system appears to selectively express certain V gene products. This selective expression is probably a result of clonal expansion due to antigenic stimulation prior to transformation since molecules with the T15 idiotype can be found in the circulation of unimmunized BALB/c mice (41). In view of this unequal frequency of expression, the value of calculations concerning the amount of diversity present within the V gene libraries based on random expression of V genes is questionable.

Fig. 5.--The amino acid sequences of the first 93 residues from T15L and H8L. ⟷ indicates data from a sequenator run; ( ) indicate sequencer results of residues which are uncertain; ⟷ indicates the composition of a peptide isolated from a fully succinylated trypsin digest (STD) of T15L as determined on the Durrum analyzer. Letters in the sequence represent amino acids according to the single letter code.


1 5 10 15 20 25 30 35 40 45 50
B I V M T Q S P T F L A V T A S K K V T I S C T A S Z S L Y S S K H K V H Y L A W Y Z Z K P Z Z S P
H8L
T15L
51 55 60 65 70 75 80 85 90
K L L I Y G A S B R Y I G V P B R F T G S G S G T B F T L T I S S V Z V Z B L A V Y Y(C)
H8L
T15L
T15L/STD 76-3

CHAPTER VI

Immunoglobulins with Hapten-Binding Activity: Structure-Function Correlations

This chapter is being submitted to Science for publication.

Immunoglobulins with Hapten-Binding Activity: Structure-Function Correlations

P. Barstad[+], M. Weigert[†], B. Eaton[*], B. Black[*], and L. Hood[*]

Running title: Hapten-binding immunoglobulins

[*]Current Address: Division of Biology, California Institute of Technology, Pasadena, California 91125

[+]Current Address: L.C.C.P., Gerontology Research Center, Baltimore City Hospitals, Baltimore, Maryland 21224

[†]Current Address: Institute for Cancer Research, Philadelphia, Pennsylvania 19111

Abstract. The N-terminal amino acid sequences of light and heavy chains from immunoglobulins binding a variety of simple haptens are compared with available sequences from the literature. These comparisons suggest that 1) the heavy chain sequences correlate closely with hapten-binding specificity and 2) light chain sequences correlate closely with some specificities (i.e., carbohydrate moieties) but not with others (phosphorylcholine or dinitrophenyl {Dnp}). The following rationale is offered to explain these differing light chain results. Carbohydrate moieties ranging in size from trimers to hexamers must generally react with immunoglobulins through precise interactions involving hydrogen bonding of hydroxyl groups and van der Waals forces. Accordingly, the molecular complementarity must be precise, allowing little opportunity for light (or heavy) chain variation. In contrast, the Dnp and phosphorylcholine moieties are small and have many possibilities for ionic or charge transfer interactions, presumably with a variety of groups in the large active site crevice. Hence a variety of light chain sequences are compatible with active sites that bind these haptens. However, the light chains from myeloma proteins binding phosphorylcholine fall into three groups that correlate with various "subspecificities." Accordingly, we postulate that the light chain can play an important role in "fine tuning" the specificity of the immunoglobulin molecule.

The amino acid sequence analyses of homogeneous myeloma immunoglobulins from a variety of mammals have provided important insights into the structure, genetics, regulation and evolution of immunoglobulins[1]. The immunoglobulin molecule is composed of light and heavy chains each of which can be divided into a highly variable (V) region and a more conserved constant (C) region. The variable regions of both the heavy and light chains determine antigen-binding specificity of the molecule and exhibit segments of extreme diversity termed hypervariable regions[2,3]. Two of the three hypervariable regions in the light chain and three of four in the heavy chain form the walls of the antigen-binding crevice[4].

Myeloma tumors can be artificially induced in the inbred strain of mouse BALB/c by intraperontoneal injections of mineral oil or similar compounds[5]. These tumors represent many different classes of immunoglobulins and can be frozen for storage or transplanted to many additional BALB/c mice to produce virtually unlimited quantities of any given immunoglobulin. Banks with thousands of different tumors have been developed in several laboratories and are available for study[6], thus it is now possible to examine vast numbers of immunoglobulins produced by a single genome. Large numbers of these proteins have been screened for hapten-binding activity and approximately five percent bind specifically to one or more of a series of haptens including phosphorylcholine, 2→1 and 2→6 levans, 1→3 and 1→6 dextrans, 1→6 galactan, and Dnp. Two lines of evidence indicate that these myeloma proteins represent

the normal immunoglobulin population. First, BALB/c myeloma tumors arise in the gut cavity and presumably represent transformed lymphocytes from expanded clones of antibody-producing cells directed against gut antigens. Many of the haptens which bind to the myeloma proteins are found on the normal gut flora in BALB/c mice. Second, antibodies can be raised that are highly specific for the variable regions of myeloma proteins. The collection of unique antigenic determinants on a particular myeloma protein recognized by this antisera is termed its idiotype. The idiotype of certain myeloma proteins which bind phosphorylcholine and 1→3 dextran are indistinguishable from the idiotypes of the antibodies normally induced in the BALB/c mouse against these antigens[7,8]. This observation suggests that the structure of the myeloma protein is very similar, if not identical, to the structure of the normally induced antibody.

Our laboratories have recently initiated an analysis of the structure-function relationships of sets of myeloma proteins binding identical or related haptens. Initial studies of the light and heavy chains from five myeloma proteins binding phosphorylcholine demonstrated that the heavy chains were virtually identical through the first hypervariable region, whereas the light chains differed up to 50% of their N-terminal sequences[9]. A study of the N-terminal regions of 28 heavy chains from myeloma proteins binding various haptens has confirmed the striking conservation of heavy chain sequence among myeloma proteins binding a common hapten[10]. This paper reports additional

N-terminal V_H and V_L sequences of proteins which bind various haptens and compares the sequence conservation found in each subunit within a specificity. The variation of V_L regions is discussed with respect to the chemical nature of the hapten and the concept of "fine tuning" immunoglobulin specificity.

Myeloma tumors were grown in BALB/c and CDF1 mice. The immunoglobulins were purified by affinity chromatography[11] or as indicated in [10]. Light and heavy chains were separated [10,12] and analyzed on a Beckman 890A sequenator as previously described [10].

The sequence data can be summarized by use of a genealogical tree. This tree is a graphic representation of the minimal number of genetic events required to generate this set of sequences from a single ancestral gene. The number of base mutations required is depicted by distance along the lines joining two proteins.

A genealogical tree generated from the amino terminal 20 residues from all heavy chain sequences available as reported in [10] is shown in Fig. 1. The heavy chains from molecules of similar hapten-binding specificity are located on closely related branches of the tree, indicating a great deal of sequence homology.

The light chain sequences from these molecules is shown in Fig. 2, and summarized in a genealogic tree based on the amino terminal 23 residues in Fig. 3. Two degrees of sequence restriction are apparent. For the carbohydrate specificities, namely α(1→3) dextran, α(1→6) dextran, and (2→1) levan, the light

Fig. 1.--Genealogical tree generated from the N-terminal 20 residues from BALB/c myeloma heavy chains. The data for this tree are from (10). The minimum mutational distance between two sequences is represented by the distance along the lines (branches) which join them. The numbers along each branch denote the actual mutational distance. The abbreviations under each tumor number indicate the hapten-binding activity of the intact immunoglobulin and are: Dnp, dinitrophenyl; 5AU, 5-acetyluracil; 1,3D, α(1→3) dextran; 1,6D, α(1→6) dextran; 1,6G, β(1→6) galactan; 2,1L, β(2→1) levan; 2,6L, β(2→6) levan; and TMA, trimethylamine.

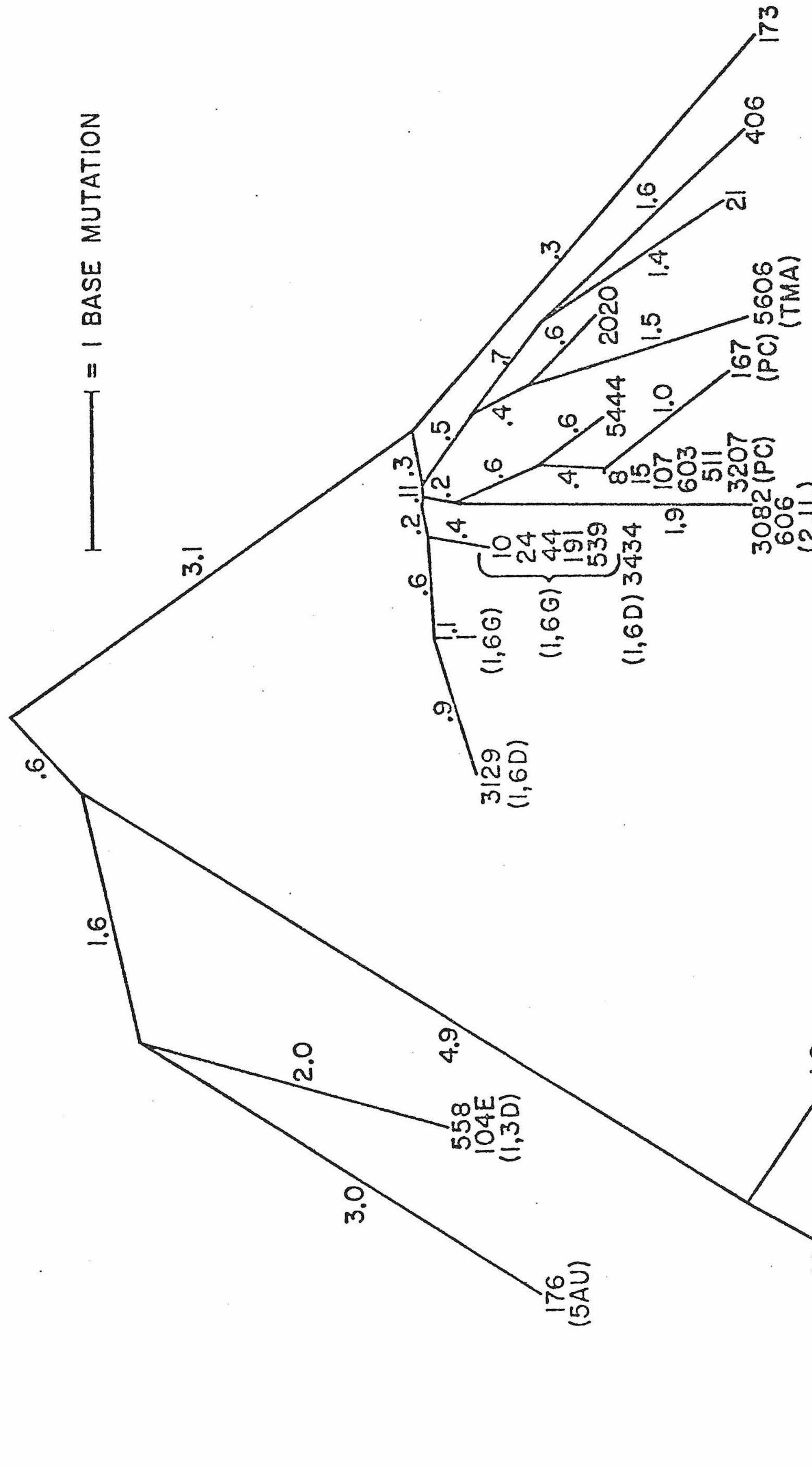

= 1 BASE MUTATION
173
406
21
2020
5606 (TMA)
167 (PC)
5444
8
15
107
603
511
3207 (PC)
3082
606 (2,1L)
10
24
44
191
539
(1,6G)
3434 (1,6D)
1 (1,6G)
3129 (1,6D)
558
104E (1,3D)
176 (5AU)
460
23 (Dnp)
315 (Dnp)
3.1
.3
1.6
1.4
.7
.6
1.5
.5
.4
.6
1.0
11.3
.2
.4
1.9
.2
.4
.6
1.1
.9
.6
1.6
2.0
4.9
3.0
1.8
1.1

Fig. 2.--N-terminal sequences from light chains which bind phosphorylcholine and levans. Spec. denotes the hapten-binding specificity of the intact molecule. Binding specificity abbreviations are the same as in Fig. 1. The prototype sequence represents the most common amino acid at each position when all available BALB/c kappa chain sequences are examined. Lines indicate sequence identity with the prototype. Letters indicate the amino acids using the single letter code: A, alanine; B, asparagine or aspartic acid; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; Z, glutamine or glutamic acid.

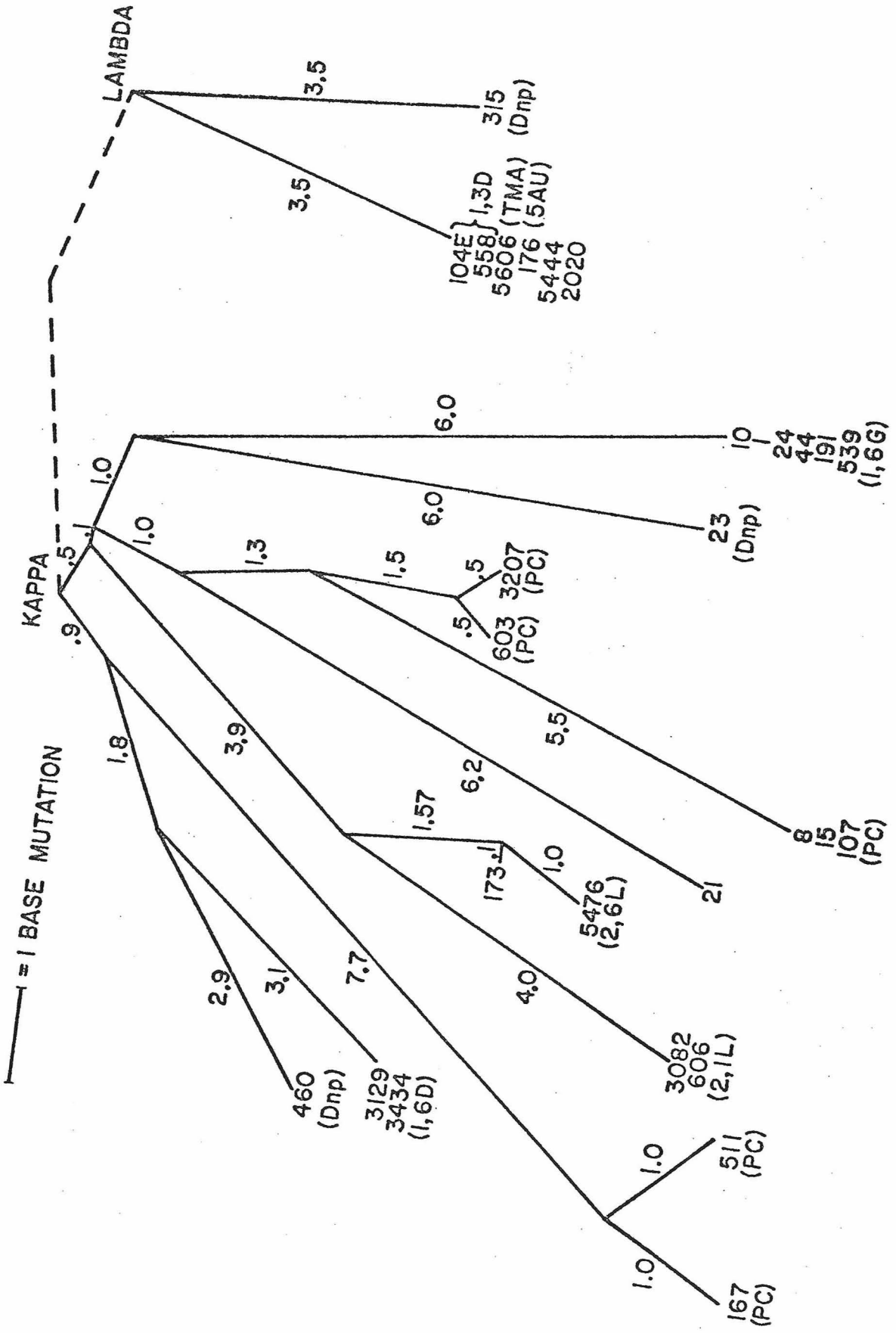
```
                  5         10        15        20        25        30        35        40     Spec.
Prototype  D I V M T Q S P A S L S V S L G E R V T I S C T A S Z S L Y S S K H K V H Y L A W

T15        ---------------T-F---A---T-A-S-K-K-------------------------------------------------  PC
H8         ---------------T-F---A---T-A-S-K-K-------------------------------------------------  PC
S107       ---------------T-F---A---T-A-S-K-K-------------------------------------------------  PC
M603       ---------------S----------A-----K-----M-------K-S-------L-B---G-B-Z-K-B-F------    PC
W3207      ---------------S----------A----------M-------K-S-------L-B                          PC
M511       ------I-----B-E-L---D-P-----S-----S---S----------?---K-----L-G-K-B-G-K-B-?-{}-----  PC
M167       ------I-----B-E-L---K-P-----S-----S---S---T--                                       PC


J606       --V-Q---I-------S-------A-------D-I-----M-T---Z-------G-T-?-I-B-L-B-W-F-Z          2,1L
W3082      --V-Q---I-------S-------A-------D-I-----M-T---Z-------G-T-?-I-B-?-B-W-F-Z          2,1L


5476       ----Z-------T-T-S-------A-------B----------R-S---G-?-I-S-B-Y-L-B-W-Y              2,6L
```

Fig. 3.--The genealogical tree from the N-terminal 23 residues of light chains from BALB/c myeloma proteins. Sequences are from (13, 14, 15, and 16) and Fig. 2. Abbreviations are the same as in Fig. 1.


LAMBDA
3.5
315 (Dnp)
3.5
104E } 1,3D
558 (TMA)
5606 (5AU)
176
5444
2020
KAPPA
.5
1.0
6.0
10
24
44
191
539
(1,6G)
6.0
23 (Dnp)
1.0
.9
1.3
1.5
.5
3207 (PC)
.5
603 (PC)
1.8
3.9
5.5
6.2
1.57
8
15
107
(PC)
173
1.0
5476
(2,6L)
21
2.9
3.1
7.7
4.0
460 (Dnp)
3129
3434
(1,6D)
3082
606
(2,1L)
1.0
511 (PC)
1.0
167 (PC)
= 1 BASE MUTATION

chains are all highly conserved in sequence much like the α(1→6) galactan specificity reported previously (12). For the noncarbohydrate specificities dinitrophenyl and phosphorylcholine, a great deal of flexibility of light chain sequence exists. These two categories of V_L sequence restriction can be explained by the nature of the hapten involved.

The carbohydrate binding myeloma proteins show maximum affinity for molecules ranging in size from trimers to hexamers[17]. These oligomers must bind in a highly complementary fashion to the antigen-binding site since their major interactions must involve hydrogen binding of the sugar's hydroxyl groups and van der Waals interactions. Since the carbohydrate moiety is rigid, it is unlikely that a large proportion of the hydroxyl groups line up in a linear fashion with appropriate groups on the protein necessary to give strong interactions. A great deal of molecular complementarity is also required for van der Waals interactions, analogous perhaps to the carbohydrate binding to lysozyme[18]. Accordingly, there is little room for variation in the active site and the V_L and V_H regions would be expected to be highly conserved.

In contrast, dinitrophenyl and phosphorylcholine are small haptens which have polar groups available for relatively strong ionic interactions. In addition, the dinitrophenyl group has pi electrons available for charge transfer interactions with aromatic side chains of the immunoglobulins. Thus, these moieties should be able to bind to the active site in a variety of different configurations employing a few strong noncovalent bonds.

The V_L region sequences from phosphorylcholine-binding myeloma proteins correlate with three distinct subspecificities [19]. 1) Proteins T15, H8, and S107 bind phosphorylcholine much more tightly than either choline or phosphonocholine. 2) M511 and M167 bind choline nearly as well as phosphorylcholine, but do not bind phosphonocholine tightly. 3) M603 binds phosphonocholine very well. The V_L sequences within each subspecificity are highly similar. (Fig. 2). V_L regions from T15, H8, and S107 are identical for their N-terminal 40 residues. V_L regions from M511 and M167 differ by only two residues in the first 23, yet they are very different from light chains of the other two subspecificities. Finally, the V_L region from M603 is quite distinct from those of the other subspecificities. In this regard, it is intriguing that W3207 possesses a light chain very similar to M603, but has not been well characterized with regard to its subspecificity. Accordingly, it appears that the light chain diversity in myeloma proteins binding phosphorylcholine correlates with the subspecificities of the respective molecules. Hence the light chains may play a role in "fine tuning" the more general specificity of the heavy chain. In view of this hypothesis, we expect that 3207 will possess a M603-like subspecificity.

Three of these proteins are idiotypically identical (T15, H8, and S107), implying a high degree of structural similarity. The idiotypes for M603, M167, and M511 are distinct from each other and from those of the T15 group[5]. W3207 has not been well characterized with regard to idiotype, although it differs

from T15[20]. As additional structural data are gathered it will be possible to determine which structural features of V_L and V_H regions correlate with the idiotype.

These results place two important constraints on theories of antibody diversity. First, the diversity of V_L regions of each subspecificity within the myeloma proteins that bind phosphorylcholine is so great that the corresponding sequences must be coded by separate V_L genes unless one concedes extensive parallel mutation[1]. Second, the correlation of structure outside the hypervariable region (i.e., framework residues) with antigen-binding specificity renders unlikely one recently proposed theory of antibody diversity[21]. A simple form of this theory proposed that the framework and hypervariable regions are coded by separate genes in the germ line and that the hypervariable genes are inserted into the framework gene to generate a complete V gene during differentiation. This view is based in part on the belief that only the hypervariable regions determine antigen-binding specificity since the framework regions are remarkably invariant in their tertiary structure[4]. If this theory is correct, a given set of hypervariable regions (i.e., those coding for Dnp binding) should be able to insert into any framework sequence to generate a Dnp antibody. Since the V_H and some V_L framework sequences correlate very well with their corresponding antigen-binding specificities, this theory of episomal insertion requires additional unattractive ad hoc assumptions.

The observation that many myeloma idiotypes (e.g., T15) can be found in normally induced antibodies has permitted the fine structure genetic mapping of these determinants[22]. These idiotypes are expressed on normal antibodies found in some but not all strains of inbred mice and are inherited as a dominant gene. Preliminary genetic studies indicate that the idiotypes are all linked to one another and to the C_H genes. Why these idiotypes are linked only to the C_H and presumably V_H regions is an intriguing question. Is the idiotype reflecting only V_H region determinants? Will the various idiotypes expressed in the molecules which bind phosphorylcholine map together or apart? Is there any correlation between the various branches on the genealogical tree and the V_H genetic map? Answers to these questions will have important implications for theories of antibody diversity.

REFERENCES

1. Hood, L. Stadler Symp. 5, 73 (1973)
2. Wu, T. T. and Kabat, E. A. J. Exp. Med. 132, 221 (1970)
3. Capra, J. D. and Kehoe, J. M. Proc. Nat. Acad. Sci. USA. 71, 845 (1974)
4. Segal, D. M.; Padlan, E. A.; Cohen, G. H.; Silverton, E. W.; Davies, D. R.; Rudikoff, S.; and Potter, M. In Progress in Immunology II, Vol. 1. Edited by L. Brent and J. Hoborrow. Oxford and Amsterdam: North-Holland Pub. Co. p93 (1974)
5. Potter, M. Physiol. Rev. 52, 631 (1972)
6. Tumors are available from Cell Distribution Center, Salk Institute for Biological Studies, San Diego, California.
7. Sher, A. and Cohn, M. Eur. J. Biochem. 2, 319 (1972)
8. Carson, D. and Weigert, M. Proc. Nat. Acad. Sci. USA. 70, 235 (1973)
9. Barstad, P.; Rudikoff, S.; Potter, M.; Cohn, M.; Konigsberg, W.; and Hood, L. Science 183, 962 (1974)
10. Barstad, P.; Farnsworth, V.; Weigert, M.; Cohn, M.; and Hood, L. Proc. Nat. Acad. Sci. USA. 71, 4096 (1974)
11. Chesebro, B. and Metzger, H. Biochem. 11, 766 (1972)
12. Rudikoff, S. and Potter, M. Biochem. 13, 4033 (1974)
13. Appella, E. Proc. Nat. Acad. Sci. USA. 68, 590 (1971)
14. Francis, S.; Leslie, R.; Hood, L.; and Eisen, H. Proc. Nat. Acad. Sci. USA. 71, 1123 (1974)

15. Hood, L.; McKean, D.; Farnsworth, V.; and Potter, M. Biochem. 12, 741 (1973)

16. Appella, E. and Inman, J. Contemporary Topics in Molecular Immunology. 2, 51 (1974)

17. Lunblad, S.; Stellar, R.; Kabat, E.; Hirst, J.; Weigert, M.; and Cohn, M. Immunochem. 9, 535 (1972)

18. Phillips, D. Scientific American. 215, #5, 78 (1966)

19. Leon, M. and Young, N. Biochem. 10, 1424 (1971)

20. Sher, A., Lord, E., and Cohn, M. J. Immunol. 107, 1226 (1971)

21. Capra, J. D. and Kindt, R. Immunogenetics. 1, 417 (1975)

22. Weigert, M., Potter, M., and Sachs, D. Immunogenetics. 1, 511 (1975)

SUMMARY

SUMMARY

The immune system is an excellent model with which to study one of the major evolutionary steps required to generate the complex organisms which exist today. To generate higher eukaryots, it was necessary to evolve a system for cell recognition and cooperation. Such a system required a high degree of specificity and diversity, probably manifested as different cell surface receptors. The genetic mechanisms responsible for antibody diversity and specificity most likely evolved from such a receptor system. By examining a series of gene products from immunoglobulins, we may learn more about the mechanisms operating in the immune system as well as in other complex mammalian systems.

The current data suggest fewer antibodies exist than was believed. Recent data from several sources suggest that the antibody population does not consist of a nearly infinite number of proteins. Work by Richards et al (42) has shown that the same immunoglobulin can bind to several similar haptens and even to a few which differ greatly in structure. This indicates that the antibody combining site is not monospecific. This concept is more consistent with observations on other macromolecular systems (e.g., the proteolytic enzymes) (43). A decrease in the specificity of the antibodies correspondingly reduces the diversity requirement for the system.

A second indication that the functional diversity observed in a response is more restricted than previously believed is the idiotypic studies on natural antibodies made to phosphoryl-choline (PC) and α(1→3) dextran (1,3D). The majority of antibodies made to these two antigens are each very restricted in idiotype. Our sequence studies have shown that the idiotypes represent molecules with a large degree of sequence similarity or identity in the variable regions. This suggests that there is a "best fit" molecule for many antigens and this molecule predominates after repeated exposures to the immunogen. Since PC and 1,3D represent common natural antigens encountered by the BALB/c, the concept of extremely diverse antibody responses to nearly any hapten is now questionable. The original response may be heterogeneous due to cross reactivity analogous to that noted by Richards, but hyperimmunization greatly decreases the heterogeneity observed.

One example of this phenomenon is the dinitrophenyl response in Guinea pigs (44). Dinitrophenyl has classically been used as an immunogen since it gives a very large heterogeneous antibody response. This hapten is capable of forming relatively strong non-specific interactions with proteins and thus the active site structural requirements to generate a binding affinity capable of stimulating antibody production are quite small. This allows a large number of different antibody-producing clones to be expanded out early in the response. Hyperimmunization with this hapten greatly reduces this heterogeneity

however, at least in the heavy chain variable region. Since the DNP molecule can bind to such a large number of different antibodies and myeloma proteins (15) through quite non-specific interactions, it is questionable how representative such a response is of that to natural antigens.

Our studies on the molecular subunits from myeloma proteins also indicate that immunoglobulins which bind to any determinant examined are all highly restricted in their heavy chain structure. In some specificities, the light chain sequence is similarly conserved. This again indicates that natural antibodies made in response to a hapten are quite restricted in structure.

Since the antibodies are not monospecific and since the response to many antigens seems to be restricted as to the number of different antibodies produced, it seems that 10^5 different specificities may be sufficient for maintaining the creature. Still, this is a large number of different molecules and new genetic mechanisms were required during evolution to provide this diversity.

Mixing a series of gene products to form different molecules was imperative to generate the antibody system. The major obstacle in evolving a very diverse system like that for antibodies was the amount of DNA a creature could afford to commit to the purpose. If an organism committed a single gene for each specificity, 10^5 different specificities would require 10^5 different genes, each of which contained 675 base pairs or more (1 V_H + 1 V_L = 225 AA = 675 base pairs). This would represent

an extremely large genetic load to the animal. In order to provide the 10^5 or more different antibody active sites, it was necessary to encode the site in two halves, one in the heavy chain and one in the light chain. By combining various heavy and light chains to form different active sites, the creature could drastically reduce the amount of germ line DNA required. To maximize the advantage derived from encoding the active site in two parts, the variable regions of the light and heavy chains must be equally diverse. One million different molecules can be generated from 2000 V genes if each subunit library contains 1000 genes. If the V_H library was only 1% as diverse as the V_L library, 10,100 genes would be required to generate the 10^6 V_HV_L pairs. Our data on the BALB/c myeloma heavy chains suggest that indeed the heavy and kappa chains are comparably diverse. This again is consistent with the concept of encoding the functional diversity in germ line genes.

Correlation of amino acid sequence and binding affinity reduces the somatic contribution to diversity. When we selected subsets of immunoglobulins based upon their function, i.e., binding characteristics, we found that the variable regions within hapten-binding specificities represent one or more unique germ line genes unless one accepts an extremely large rate of parallel mutation. Within phosphorylcholine-binding proteins, three subsets of light chains exist which must be encoded by at least three germ line V genes. The heavy chains appear to be encoded by at least four germ line genes which differ from

any other V_H genes found to date. If somatic mutation contributed a major role in generating antibody diversity and specificity, the active site structure would be expected to correlate much more closely in sequence to the hapten-binding activity than would the framework regions. Indeed, if somatic mutation operated at a high level, from a few genes a high degree of sequence flexibility should appear in the framework regions. In fact, we find just the opposite; the framework regions are highly conserved, thus a large amound of the structure must be encoded in the germ line.

One explanation: After four years of studying immunoglobulin structure, I believe that the model which best explains the current data is one in which there is nothing magic about the structure of the antibody molecule itself. The antibody population is a collection of 10^5 or more proteins which have active sites analogous to enzymes. Each molecule must cross react with a series of similar haptens. The information for the antibody subunits is encoded in the germ line of the animal and selection for function must occur during evolution. The primary structure of the heavy and light chain variable regions determines the specificity of the molecule and there is no reason to believe that somatic mutation contributes to the functional diversity. The hypervariable regions are probably under different evolutionary selective pressures and thus mutate at a higher rate--analogous to the outer surface residues of cytochrome C (45) . The number of different V genes in the

kappa and heavy chains is comparable and probably greater than 1000. If only 10% of the $V_H V_L$ combinations are functional, it could generate sufficient diversity to combat most potential pathogens encountered during evolution.

As with most fields of biology, however, each answer or partial answer to a question generates many additional questions. Now that it appears certain that the antibodies represent a large multigene system, many new questions relating to the control mechanisms which must exist to govern gene expression within V_H and V_L libraries have arisen. During ontogeny is the readout of variable regions programmed? Which is committed first, V_H or V_L? How does a creature generate 10^5 different antibody-producing clones in such a very short time? Is there a genetic mechanism which restricts the expression of various V_H and V_L pairs?

These and many other questions will continue to make the investigation of antibody diversity extremely active and intriguing. As a model for multigene systems, the antibodies should provide principles applicable to differentiation. As part of nature's prime method of protection from microorganisms for higher creatures, it will provide valuable therapeutic information if we can learn how to manipulate the system. Each new finding will again generate many more questions, keeping this field one of the most exciting areas in molecular biology for some time to come.

BIBLIOGRAPHY

1. Gally, J. A. and Edelman, G. M. Nature 227, 341 (1970)
2. Hood, L.; Gray, W. R.; Sanders, B. G.; and Dreyer, W. J. Cold Spring Harbor Symp. on Quant. Biol. 32, 133 (1967)
3. Hood, L. Stadler Symp. 5, 73 (1973)
4. Segal, D. M.; Padlan, E. A.; Cohen, G. H.; Silverton, E. W.; Davies, D. R.; Rudikoff, S.; and Potter, M. In Progress in Immunology II, Vol. 1. Edited by L. Brent and J. Hoborow. Amsterdam and Oxford: North-Holland Publishing Co., p. 93, (1974)
5. Capra, J. D. and Kehoe, J. M. Proc. Nat. Acad. Sci. USA. 71, 845 (1974)
6. Wu, T. T. and Kabat, E. A. J. Exp. Med. 132, 211 (1970)
7. Lederberg, J. Science 129, 1649 (1959)
8. Dreyer, W. J., Gray, W. R., and Hood, L. Cold Spring Harbor Symp. on Quant. Biol. 32, 353 (1967)
9. Cohn, M. In The Biochemistry of Gene Expression in Higher Organisms. Edited by J. K. Pollak and J. W. Lee. Holland: D. Reidel Publishing Co., p. 574 (1973)
10. Capra, J. D. and Kindt, T. J. Immunogenetics 1, 417 (1975)
11. Williamson, A. R. Eur. J. Biochem. 1, 390 (1971)

12. Eichman, K.; Lackland, A.; Hood, L., and Krause, R. M. J. Exp. Med. 131, 207 (1970)

13. Capra, J. D. and Nissonof, A. Personal Communication.

14. Potter, M. Methods Cancer Res. 2, 105 (1967)

15. Potter, M. Physiol. Rev. 52, 631 (1972)

16. Eisen, H. N., Simms, E. S., and Potter, M. Biochem. 7, 4126 (1968)

17. Sher, A., Lord, E., and Cohn, M. J. Immunol. 107, 1226 (1971)

18. Sher, A. and Cohn, M. Eur. J. Biochem. 2, 319 (1972)

19. Carson, D. and Weigert, M. Proc. Nat. Acad. Sci. USA 70, 235 (1973)

20. Barstad, P.; Farnsworth, V.; Weigert, M.; Cohn, M.; and Hood, L. Proc. Nat. Acad. Sci. USA 71, 4096 (1974)

21. Edman, P. Arch. Biochem. Biophys. 22, 475 (1949)

22. Hermodson, M.; Ericsson, L.; Titani, K.; Neurath, H.; and Walsh, K. Biochem. 11, 4493 (1972)

23. Pisano, J., Bronzert, T., and Brewer, H. Anal. Biochem. 45, 43 (1972)

24. Hood, L.; McKean, D.; Farnsworth, V.; and Potter, M. Biochem. 12, 741 (1973)

25. Smithers, O.; Gibson, D.; Fanning, E.; Goodfleish, R.; Gilman, J.; and Ballantyne, D. Biochem. 10, 4912 (1971)

26. Chesebro, B. and Metzger, H. Biochem. 11, 766 (1972)

27. Rudikoff, S. and Potter, M. Biochem. 13, 4033 (1974)

28. Westall, F. and Hesser, H. Anal. Biochem. 61, 610 (1974)

29. Dreyer, W. and Bynum, E. In Methods in Enzymology, Vol. XI. Edited by C. Hirs. New York and London: Acad. Press 32 (1967)

30. Rudikoff, S. Personal communication.

31. This thesis and Appella, E., personal communication.

32. Capra, J. D. and Kehoe, J. M. J. Immunol. 114, 678 (1975)

33. Capra, J. D., Wasserman, R., and Kehoe, J. M. J. Exp. Med. 138, 410 (1973)

34. Milstein, C.; Adetugbo, K.; Cowan, N.; and Secher, D. In Progress in Immunology II. Edited by L. Brent and J. Holborrow. Oxford: North-Holland Pub. Co. Vol. 1, 157 (1975)

35. Francis, S.; Leslie, R.; Hood, L.; and Eisen, H. Proc. Nat. Acad. Sci. USA. 71, 1123 (1974)

36. Bourgouis, A., Fougereau, M., and De Preval, C. Eur. J. Biochem. 24, 446 (1972)

37. Klotz, I. In Methods in Enzymology, Vol. XI. Edited by C. Hirs. New York and London: Acad. Press 576 (1967)

38. Svasti, J. and Milstein, C. Biochem. J. 128, 427 (1972)

39. Leon, M. and Young, N. Biochem. 10, 1424 (1971)

40. Metzger, H.; Chesebro, B.; Hadler, M.; Lee, J.; and Otchin, N. In Progress in Immunology I. New York and London: Acad. Press 244 (1971)

41. Lieberman, R.; Potter, M.; Mushinski, E.; Humphrey, W.; and Rudikoff, S. J. Exp. Med. 139, 983 (1974)

42. Richards, F. and Konigsberg, W. Immunochem. 10, 545 (1972)

43. Leonis, J., Li, C., and Chung, D. J. Am. Chem. Soc. 81, 419 (1959)

44. Cebra, J., Koo, P., and Ray, A. Science 186, 263 (1974)

45. Dickerson, R. J. Molec. Evol. 1, 26 (1971)