ABSTRACT A new method for classification of Ig sequences is suggested. The defining characteristic of a class is presence of particular residues at several class-determining positions. Sequences within a class follow the same amino acid pattern, i.e., residues at identical positions are, in an overwhelming majority of sequences of that class, identical or chemically related. Thus, once the class of a sequence is determined, one can predict the residue(s) at almost any position in the sequence. In this paper, results of analysis of 1,172 human heavy chains are presented. It was shown that a sequence can be assigned to one of six classes depending on which residues are found at its positions 1, 3, 5, 6, 7, 9, 10, 12, and 13. It is important to note that it is possible to achieve same-six-class classification of the human heavy chains on the basis of a different set of positions found not at the beginning but near the end of the sequence (around position 80). For every class, an amino acid pattern of an entire sequence (complementarity determining regions excepting) has been determined. Our approach allowed us to reconstruct the incomplete human heavy chains in which residues at certain positions at the beginning or end of the chain are known. We developed a software tool for analysis, classification, and prediction of residues in sequences of the Ig family.

Antibodies comprise a large group of structurally similar proteins, which exhibit functional diversity (1–8). Antibody molecules consist of two heavy chains and two light chains. Each of the light and heavy chains has a variable region that folds into a variable domain (~115 amino acids), V\textsubscript{L} and V\textsubscript{H}, respectively. These domains play an essential role in immune response because through them contact with antigen molecules (“antigen recognition”) is achieved. Most of the sequence diversity in antibodies is due to variability in the V\textsubscript{H} domains. Approximately 1,200 human V\textsubscript{H} sequences presently are known (1). This paper focuses on their classification.

Traditional approaches to dividing sequence into classes are based on alignment of all amino acids or nucleotides sequences, followed by calculation of sequence homologies. Various methods of cluster analysis can be used further to find clusters of protein sequences (9–15). Application of this procedure to the human heavy chains in the Kabat database and nucleotide V\textsubscript{H} segments resulted in a classification in which chains belonging to the same class have at least 80% homology at the amino acid or nucleotide sequence level (1, 16–20). The drawback of this procedure is that it requires one to know (i) all, or almost all, residues or nucleotides in a sequence and (ii) how to compare (align) sequences with others in a cluster. Here, we suggest a way of classifying sequences that does not require either of these conditions.

In general, biological classification allows one to assign an object to its species, family, or class on the basis of only very few defining characteristics. In this work, we propose a classification of biological molecules, Ig human heavy chains, based on this principle. An assignment of a sequence into a class depends on what residues are found at several key positions of a sequence; these residues are the defining characteristics in the proposed classification.

Our approach to this problem follows along the lines of our previous investigations of the Ig family (7–8, 21–23). Process of classification of sequences of the variable domains is divided into four stages:

Stage I. Fragmentation of Sequences into “Words” and Determination of Positions of Residues in Words

The prediction of the secondary structures allows us to divide the Ig chains into 21 fragments (words). The fragments correspond to structural units of variable domains: strands, loops, or parts of them. Each residue in a sequence is assigned to a position in a word (21).

Stage II. Classification of Positions of Residues in Words

The essence of our approach is: For each position, a comprehensive dossier is compiled. The dossier includes (i) residue characteristic (21, 23); (ii) values of accessibility of residues to the solvent (23); (iii) data of the residue–residue contacts (23); (iv) the average coordinates of C\textalpha{} atoms in the invariant system of coordinates of Ig molecules (7) and other structural and functional properties (21). The role of residues at each position of a word was determined from the examination of >5,300 sequences in the Kabat database (1) and >100 three-dimensional Ig structures from a protein database (7–8, 21, 23).

Observation of residue frequencies in Ig sequences and analysis of the structural role of residues allowed us to divide 20 amino acids into 10 groups (Table 1). Amino acids are grouped together not only on the basis of chemical similarities but also on the structural role they play in Ig molecules. This Ig-specific division of amino acids into groups serves as the basis for our classification procedure.

Stage III. Classification and Patterns of Words

A simple set of rules is developed that allows us to divide the words into classes (22). Two words are in the same class if residues at a respective position are of the same amino group. For each class of words we found a main pattern of words (keyword).

Stage IV. Classification and Patterns of Sequences

In our approach, each sequence is divided into words, and then words are assigned to certain classes. It allows us to represent

Abbreviation: CDR, complementarity determining region.

To whom reprint requests should be addressed. e-mail: igelfand@math.rutgers.edu.
a chain in terms of keywords. Based on the results of the analysis of this representation, the chains are divided into classes. In effect, this classification furnishes us with several amino acid patterns, which describe the respective classes of the human heavy chains. Important to note, particular residues found at several specified class-determining positions characterize each class of sequences. These residues constitute the defining characteristics of a class.

1. There are simple explicit rules of attributing a sequence to a class. These rules involve determining which residues occupy several key positions (~5–10) in a sequence. Residues at these class-determining positions in the actual sequence are compared with residues at corresponding positions in the pattern sequences. It follows that information about residues at several crucial positions permits one to determine the class of the sequence and therefore to predict the residue (or more precisely, the amino acid group to which the residue belongs) at almost every positions of the sequence (complementarity determining regions excepting).

2. The proposed classification is unambiguous, i.e., a sequence belongs to only one class. In other words, there are no intersections (overlaps) between different classes.

3. The classification is universal, i.e., all, or almost all, sequences in a protein family can be classified in accordance with these rules.

Following these rules, we divided the human heavy chains into six classes. Our analysis of the chains uncovered two sets of class-determining residues: the first 13 residues in a chain and 7 residues near the end of the sequence. It is important to note that either of the two sets of positions can be used to arrive to the same six-class classification of the human heavy chains. A significant corollary of our approach is that once a class of a sequence is known, its secondary structure, values of surface accessibility of residues, most favorable residue–residue contacts in a structure, and coordinates of main chain atoms in the invariant system of coordinate for Ig family also are known (23).

In the first part of this paper, we describe the basic concepts of our approach (“position” and “word”) and the methods of its classification. In the second part, we present the classification of the human heavy chains. (However, to understand the results of classification, it is not necessary to master the concepts defined in the first part). In the second part, we illustrate, as well, the advantages of this method for prediction of residues in the incomplete human heavy chains in the Kabat database and for determination of their secondary structures.

### METHODS

#### Words and Positions of Residues Within Words.

Our approach to the analysis of the Ig molecules is based on two main concepts: a “word” and a “position of a residue” within a word. In our analysis, the Ig sequences were divided into fragments (words) based on our secondary structure analysis. The variable domain of Ig sequences were divided into 21 words: 0A, A, AA, A', A'B, B, BC, C, CC, C', C'C, C'C'D, D, DE, E, EF, F, FG, and G. Each of these words corresponds, approximately, to a strand, loop, or part of these structural units. Then, each residue in a sequence was assigned to a position within a word.

We illustrate the division of sequences into words and the assignment of positions to residues using the mAb 216 CL sequence as an example (Fig. 1). The 120 residues of that sequence, in addition to the Kabat numbering in Ig sequences, are referred to by two-part indices, such as 0A1. The letter(s) references the word in which a given residue is found.
and the number indicates the position within the word that the residue occupies (1st position). For example, the first residue in the mAb 216’CL sequence—O—is assigned to the position 0A1, and V (Kabat numbering: # 120)—the last residue in G strand—is assigned to the position G10 (Fig. 1b). This numbering is very convenient for comparing residues in identical positions in different sequences.

Much sequence and structural data were analyzed to define words and positions of residues in words. It was shown that residues occupying identical positions in different molecules revealed a large number of similarities; they had approximately same coordinates of Ca atoms in the invariant system of coordinates of the variable domains (7, 8), similar residue—residue contacts (23), similar H-bonds between main chain atoms (21), approximately same values of accessibility to a solvent (21, 23), and other common properties.

Classification of Amino Acids in Igs. Statistical analysis of residues with the same index revealed a number of positions that are occupied by a particular residue ("invariant residue positions") or by similar, chemically related residues ("similar residues positions") (23). On the basis of this observation, and taking into account the structural role of residues in strands and loops, we divided amino acids into 10 groups (Table 1). The number of different words strongly varies for various groups of the words. However, it is always much less than the number of all words in a group (22) (the only exception being the FG words—there were, approximately, as many different FG words as the number of chains) (22). This is the first important result of the statistical analysis of words; it provides an estimate of the number of different strands and loops found in the chains.

The words in each group of words are divided into classes. For example, all E words were divided into six classes. We require that a word belong to a given class if residues at any one position in a word and residues at a respective position of all other words in a class are from the same amino acid group. For example, all residues at the position E3 in E words of the first class are for or for Y (4 group, Table 1). This requirement gives us the possibility to construct main patterns or keywords for each class of the words. A keyword of a given class of word is a set of residues that are encountered most frequently at the positions of words of that class. For example, the keyword of the first class of the E words is: TLYLQMN; these residues occupy E1, E2, E3, E4, E5, E6, and E7, respectively.

We found a considerable number of words that cannot be assigned to a class because a residue at just one position in these words does not belong to the amino acid group of a residue in the respective position of keyword. Thus, we expanded our definition of a class to include the words with a single "mutation." For example, T at the D3 position in the D keyword is not in the same group as N, which occupies this position in the mAb 216’CL sequence. However, this word is incorporated in the same class, which is described by D keyword: RVTISV in D1, D2, D3, D4, D5, D6, and D7 positions, respectively (Fig. 1b). Thus, the definition of a class allows an exception in one position (see more details in ref. 22).

The classification of words was performed for words of 17 (excluded from this analysis: CB, C’C’, C”, and FG words, which correspond to CDR Fragment in the Ig sequences). An important result of the classification of words in the human heavy chains is that no more than six classes are found for each of the 17 groups of words. This means that a very limited number of keywords (patterns) describes all sequences in variable domains of the human heavy chains (22).

Representation of Sequences in Terms of Keywords. From the division of the VH sequences into words, it follows that the human heavy chains can be written as a sequence of 21 words. Each word in the sequences, except for four CDR words, can be assigned to a certain class. Because every class of words has a proper keyword, we can represent now any VH chain in a generalized way as a sequence of these keywords. For an illustration of this, see Fig. 1b, which contains the actual mAb 216’CL sequence (line 2) and its keyword representation (line 1).

A similar fragmentation into words and characterization of words by keywords is carried out for 1,172 human heavy chains. Thus, each human heavy chain is represented as a sequence of 17 keywords. This generalized representation gives us the possibility of dividing sequences into classes, with each class of sequences being subsumed under a particular pattern of keywords.

RESULTS

Classification of the Human Heavy Chains Based on the First 13 Residues of a Chain. Analysis of sequences in terms of keywords revealed that human VH sequences can be divided into six classes of which the I and III classes are subdivided further into two subclasses. As a rule, in all sequences of one class, identical positions are occupied by residue(s) from one amino acid group. Patterns for each of the six classes and their subclasses are presented in Table 2. Residues at positions of patterns are "representative" of their amino acid groups (Table 1). A representative residue at a position was selected over others from its group because, at that position, it was found more often than other residues of its group. Comparison of patterns shows that they differ from each other at 13–28 positions.

It was shown that residues at nine of the first 13 positions -numbered 1, 3, 5, 6, 7, 9, 10, 12 and 13 - are unique for every class and subclass of sequences. Thus, class affiliation of any sequence can be determined by comparing residues at the class-determining positions with residues at respective positions in the patterns of Table 2.

Consider, for example, the mAb 216’CL sequence (Fig. 1). The 1, 3, 5, 6, 7, 9, 10, 12, and 13 positions are occupied by the Q, Q, Q, Q, W, A, G, L, and K residues, respectively. Comparison of these residues with residues at respective positions in patterns (Table 2) shows that the sequence belongs to the III2 class. Note that the differing residues at position 12—residue L in the mAb 216’CL sequence and V in the pattern of III2 class—are from the same hydrophobic group (Table 1). An analogous procedure was applied to 1,172 human heavy chains in the Kabat database, and in this way class affiliation of >95% of the human VH chains was determined.

Determination of Lengths of the CDRs in the Human Heavy Chains. Left out of consideration in this work was classification of words composing CDR1, CDR2, and CDR3. This exclusion is because of the high degree of variability of residues in these regions. However, we were able to establish one-to-one correspondence between CDR1 and CDR2 length (numbers of residues) and class of the sequence (Table 2).

*Positions 2, 4, 8, and 11 were found to be conservative or "similar residue positions." For each of these four positions, in all human heavy chains, the residues belong to same amino acid group (21, 23). Thus, these residues can be used to test whether a given sequence is a human heavy chain.
The VH sequences’ human heavy chains vary in length from 111 to 125 amino acids. Such variability is caused by differences in lengths of the CB, C9C0, and FG words that describe CDR1, CDR2, and CDR3, respectively. In the present work, CB and C9C0 words are discussed; analysis of FG loops will be presented elsewhere. CB loops can be as short as 5 and as long as 7 residues, and C9C0 loops can be as short 4 and as long as 7 residues. However, we discovered that the lengths of CB and C9C0 words are constant for sequences within a class and subclass. For each class and subclass in Table 2 (with the exception of III1 subclass), we determined the lengths of these two variable loops. Thus, class y subclass of a chain uniquely defines CDR1 and CDR2 lengths.

Prediction of Residues in a Sequence Based on Information About the First 13 Residues. Once the class of a sequence is determined on the basis of its first 13 residues, one can predict residues in almost all remaining positions. This prediction is possible because the pattern for each class and subclass of sequences overlaps with the set of sequences from one of our six classes (here, “M” stands for “miscellaneous family” in the Kabat database).

Table 2. The patterns of the classes and subclasses in the human heavy chains

<table>
<thead>
<tr>
<th>#</th>
<th>A</th>
<th>AA</th>
<th>A'A</th>
<th>AA'B</th>
<th>A'B</th>
<th>B</th>
<th>C</th>
<th>C'B</th>
<th>C'C'</th>
<th>C'D</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>V</td>
<td>S</td>
<td>G</td>
<td>G</td>
<td>L</td>
<td>V</td>
<td>Q</td>
</tr>
<tr>
<td>2</td>
<td>E</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>V</td>
<td>S</td>
<td>G</td>
<td>G</td>
<td>L</td>
<td>V</td>
<td>Q</td>
</tr>
<tr>
<td>3</td>
<td>Q</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>Q</td>
<td>Q</td>
<td>W</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>L</td>
</tr>
<tr>
<td>4</td>
<td>Q</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>Q</td>
<td>Q</td>
<td>W</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>L</td>
</tr>
<tr>
<td>5</td>
<td>Q</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>Q</td>
<td>Q</td>
<td>W</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>L</td>
</tr>
<tr>
<td>6</td>
<td>Q</td>
<td>V</td>
<td>Q</td>
<td>L</td>
<td>Q</td>
<td>Q</td>
<td>W</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>L</td>
</tr>
</tbody>
</table>

First row, residues are numbered as in the Kabat database. Second row, division of the human heavy chains into words. Third row, positions in words are referenced by two-part indices. The Roman numerals in the first column stand for the class and subclass of the human heavy chains whose pattern is shown in the same row. The patterns are written in terms of keywords. The keywords for 17 out of 21 fragments are presented. Positions, which are occupied by the same residue or by residues from the same amino acid group in all keywords, are darkened. “X” marks the positions within a keyword, which are occupied by residues from more than two amino acid groups. Superscripts 1–7, the positions from which residues from two amino acid groups are found. Superscript 1 marks the positions with residues from 9th and 10th groups; 2, from groups 8 and 10; 3, from groups 7 and 9; 4, from groups 7 and 9; 5, from groups 6 and 7; 6, from groups 1 and 8; 7, from groups 2 and 7. For each class of sequences, lengths (numbers of residues of CB and C9C0 words are presented in columns CB and C9C0). Results of comparison with Kabat classification (1), and VH gene segments classification (18), are presented in K families from Kabat or germ-line classifications whose set(s) of sequences overlaps with the set of sequences from one of our six classes. Results of comparison with Kabat classification (1), and VH gene segments classification (18), are presented in K families from Kabat or germ-line classifications whose set(s) of sequences overlaps with the set of sequences from one of our six classes (here, “M” stands for “miscellaneous family” in the Kabat database).
known. Residues at positions of patterns are representative of their amino acid groups. Thus, our method allows one to first row, residues are numbered as in the Kabat database. Second row, division of the human heavy chains into words. Third row, positions in words are referenced by two-part indices. The Roman numerals in the first column (#) stand for the class and subclass of the human heavy chains whose pattern is shown in the same row. The patterns are written in terms of keywords. The keywords for 17 out of 21 fragments are presented. Positions, which are occupied by the same residue or by residues from the same amino acid group in all keywords, are darkened. ‘X’ marks the positions within a keyword, which are occupied by residues from more than two amino groups. Superscripts 1-7, the positions where residues from two amino acid groups are found. Superscript 1 marks the positions with residues from 9th and 10th groups; 2 from groups 8 and 10; 3 from groups 7 and 8; 4 from groups 7 and 9; 5 from groups 6 and 7; 6, from groups 1 and 8; 7, from groups 2 and 7. For each class of sequences, lengths (numbers of residues) of CB and C’C’ words are presented in columns |CB| and |C’C’|.

Results of comparison with Kabat classification (1), and VH this sequence belongs to the III2 subclass. The pattern sequence from the Kabat database how one can predict Xs in the patterns (Table 2). For example, the residues that make up the E words of mAb 216’CL sequence—Q, F, S, L, K, L, and S—are of the same amino acid groups as, and indeed coincide with, residues in the E keyword of the III2 subclass pattern. Thus, whichever set of residues is used, the sequence is classified into the III2 subclass. This simple alternative algorithm can be used to verify the class of a sequence.

Classification of Germ-Line Sequences. The human Ig VH locus, which contains 51 VH sequences, was classified into seven families on the basis of homology. Sequences were said to belong to the same family if they were at least 80% homologous at the nucleotide level (18–20). We applied a classification algorithm based on residues at class-determining positions to germ-line sequences. It was shown that the germ-line sequences naturally divide into same six basic classes as the sequences from the Kabat database (Table 2). We then compared the classes of sequences with the VH families of germ-line segments. There was a one-to-one correspondence between six classes obtained by using our procedure and six families of the human VH repertoire classification. A sequence from the 7th family, however, was classified into the IV class because they differed from the pattern of this class at only four positions. Finally, comparison of the classifications of amino acid sequences from the Kabat database and germ-line sequences confirms the role of residues at the class-determining positions.

DISCUSSION

The possibility of deducing regularities in Ig sequences was predicated on classification of amino acids into 10 groups. This 10-group classification takes into account the sequence features as well as the nature of residues and structural and functional properties of residues in Ig structures and is therefore specific to this particular protein family.

Another essential component of the analysis is breaking up of sequences into words and deduction of main patterns (keywords) for each group of words. Representation of sequences as strings of keywords (Table 2) amounts to re-writing amino acid sequences in terms of amino acid groups (10-letter code) rather than amino acids themselves (20-letter code). When VH heavy sequences are represented in such a way that they naturally divide into six classes. Each of these classes is described by a particular sequence pattern written in the 10-letter notation.

A further advantage of such a representation is that it allows one to find a small number of class-determining positions. In human heavy chains they are found in OA, A, AA’, A’, and E words. We can speculate that special importance of residues at these positions derives from the structural role they play in the chain. The specific nature of the Ig fold is such that residues at the beginning of the chain, where the first set of class-determining positions is located, can form H-bonds with residues at the end of the chain. Likewise, residues of E strands (the second set of class-determining positions) form H-bonds with the main chain atoms of B strand residues, which are ~60 amino acids away. (It can be suggested that E and B interact.
as anti-parallel β-strands because the sequence motif of both of them is characterized by interchange of hydrophobic residues at even positions with hydrophilic residues at odd ones.)

In this work, we propose a new approach to classification of Ig sequences and apply it to the V\textsubscript{H} human heavy sequences from the Kabat database and germ-line V\textsubscript{H} sequences. Resulting classification was compared with the classifications from Kabat database and of V\textsubscript{H} segments that were based on amino acid and nucleotide homologies. Comparison demonstrated that the classifications of the V\textsubscript{H} segments and developed in this work are coincident. However, in the Kabat database (1), the sequences were divided into 14 families, and \( \sim 30\% \) of all sequences were not assigned to any family (Table 2, column K), whereas, in the proposed classification, there are six classes and over 95\% of all currently known human heavy chains were assigned to one of these classes.

Suggested procedure for analysis of an Ig chain involves the following steps: (i) examination of residues at positions 2, 4, 8, and 11 of a sequence to check whether it is a human heavy chain (see Results); (ii) if the chain is a human heavy chain, residues at positions 1, 3, 5, 6, 7, 9, 10, 12, and 13 are examined to determine the class of the sequence; (iii) residues of the E strand can be used to verify the class of the sequence; (iv) if the sequence is incomplete, the sequence pattern of its class yields information about amino acid group of residues at almost every position in the entire sequence; and (v) sequence pattern is used to predict the secondary structure of the sequence.

Once the class, and, consequently the assignment residues to positions, are known, one can determine the coordinates and structural role of residues at most positions. It was shown recently that Ca atoms of almost all residues at identical positions in different molecules have approximately the same coordinates in the invariant system of coordinates (7). Exceptions were found only for coordinates of Ca atoms of CDR’ residues. However, the analysis of conformation of the main chain of CDR’ residues gives one the information about geometry of these regions. Chothia et al. (19) revealed three different canonical structures in CDR1 and five different canonical structures in CDR2. It is important that each canonical structures can be assigned to a proper class of sequence.

To evaluate the structural role of residues, we calculated conserved residue–residue contacts for residues (23) as well as accessible surface areas for most positions of the variable domains. On the basis of the latter criterion, positions were classified as either interior, exposed, or highly exposed, depending on accessibility of the residue in it to the solvent (23).

In summary: Following procedure outlined in this paper, one arrives at an essentially biological classification of protein molecules. Central to any biological classification is the notion of a defining characteristic. If a particular object is found to possess a particular defining characteristic, a large number of far-reaching consequences about its nature can be deduced. In our case, the defining characteristic of a V\textsubscript{H} human heavy chains is a set of residues found at specific class-determining positions. Once these residues are known, class of the sequence is immediately revealed, and one can deduce its amino acid sequence with a high degree of accuracy, as well as its secondary structure and three-dimensional characteristics.

We are grateful to Drs. C. Chothia, M. Hecht, C. Kulikowski, I. Muchnik, and O. Pitsyn for very helpful discussion. We thank I. Kister for critical review of the manuscript. We acknowledge with deep gratitude the support of the Gabriella and Paul Rosenbaum Foundation and also thank Mrs. M. Goldman for continuous encouragement. B.A.G. and A.E.K. were supported by the Gabriella and Paul Rosenbaum Foundation.