AMINO ACIDS PATTERNS IN THE RECOGNITION SITES OF IMMUNOGLOBULINS

E. Vargas-Madrazo[1,2], J.C. Almagro[2], F. Lara-Ochoa[2], M.A. Jiménez-Montaño[3]

[1] Instituto de Investigaciones Biologicas, Universidad Veracruzana, Xalapa, Ver. Mexico
[2] Instituto de Quimica, UNAM, Ciudad Universitaria, Coyoacan, Mexico, D:F: 04510, faral@unamvm1.bitnet
[3] Universidad de las Americas, Puebla, Pue, Mexico

Abstract.

An analysis of the frequency of use of amino acids on the CDR-1 and CDR-2 of 1500 immunoglobulins showed that the frequencies of amino acids in different positions could be fitted by two types of distributions. For some positions the frequencies were fitted by an inverse power law, and for other positions by an exponential distribution. In order to see whether the more frequently used amino acids for specific positions had physicochemical properties or attributes in common, they were clustered using an algorithm normally applied to artificial intelligence problems. It was found that the amino acids in those positions fitted by the inverse power law have similar hydrophobicity and volume, which are commonly attributes of amino acids in structural positions. Thus, if these positions are critical to maintaining the structural features of the CDR domains, the rest of the positions should be either properly involved in the recognition process or irrelevant. The frequencies of amino acids in these recognition positions, were fitted by the exponential law, and it was found by the clustering analysis that these amino acids share properties of a more general type, such as capability to form hydrogen bonds, polarity, etc. This suggests that at least part of the recognition mechanism requires general properties, rather than specific amino acids. Amino acids sharing the required attributes for each one of these positions, are then used with a random frequency.

INTRODUCTION

Immunoglobulins (Igs) are key proteins that mediate in the immune response, i.e. Igs bind foreign epitopes blocking their ability to bind to receptors on target cells or mark the invading microorganisms for destruction. A main challenge in immunology has been to understand how the immune system recognizes the presence of foreign epitopes.

It has been proposed that the number of main chain conformations of at least five of the six hypervariable loops of the Igs is limited to a very few canonical conformations (Chothia and Lesk, 1987). The adoption of specific backbone conformations is believed to reflect the existence of a few key conserved residues in the loop of the antibody molecule (Chothia and Lesk, 1987). This suggests that some of the amino acids on the Complementary Determining Regions (CDRs) may have a structural role, while the rest are either involved in the recognition function or are irrelevant. Thus, Kabat *et al.* (1977) proposed that there must be evolutionarily conserved residues within CDRs, and that a few hypervariable positions in the immediate neighborhood of such conserved positions must determine antigen specificity. Furthermore, Ohno *et al.* (1985) proposed that if some sites on the CDR-l and CDR-2 are conserved for the shaping of the primordial antigen-binding cavity, the remaining seven sites can generate 20^7 or 1.28×10^9 varieties of different amino acid sequence combinations. However, a recent analysis (Mian *et al.*, 1991) suggested that there are some constraints on the wide range of possible and meaningful amino acid sequence combinations in the recognition pocket, implying amino acids would have non-random frequency distributions.

In order to explore the existence of these constraints, an analysis of the frequency of amino acids in 36 sites in the CDRs (its great majority) of the variable light domain (V_L), and of 42 sites of the variable heavy domain (V_H), was performed in an alignment of approximately 1500 sequences of Igs made by Kabat *et al.* (1991).

MATERIALS AND METHODS

Data Base. The variability of the amino acids in each site of the CDRs was examined in the amino acid distribution table reported by Kabat *et al.* (1991).

Determination of the Degree of variability of each site. A compilation of structural and variability data for each site was performed in the following manner. Each one of the sites analyzed that are identified as responsible of the canonical conformations (structural set) for some of the hypervariable loops are marked in the second column of Tables 1 and 2 (Chothia and Lesk, 1987). The third column indicates the number of antigen-antibodies complexes in which each site has been found in contact with the epitope (recognition set), according to crystallographic data (Padlan and Kabat, 1991). The following two columns mark the sites proposed as conserved by Kabat *et al.* (1977) (Kabat's set) and the non-conserved set according with the present work, respectively. Finally in the sixth column, the percent frequencies of the most used amino acids for each site are calculated (composition column). Since there is not established objective criterion to decide if an amino acid is or not conserved in a given position, in the present work a site was classified as conserved if it was occupied by the same amino acid in more than 42% of the examined sequences. This value is obtained of the determination of the frequency distribution of the most used amino acids in the sites that are buried in the core of a protein family (Lim and Sauer, 1989;

Go and Miyasawa, 1980). As can be seen in Tables 1 and 2, this threshold identifies the structural site set correctly in 16 of 18 cases.

Determination of the non-random distribution of the amino acids in the hypervariable sites. To analyze the nature of the frequency distribution of the hypervariable site set, the existence of amino acids over-represented in the sample of sequences was tested considering that: i) there are various sources of error in the data base (Shenkin *et al.*, 1991), ii) it is assumed that mutations at the codon level are equiprobable, and consequently the frequency expected for each amino acid is equal to the proportion of the number of codons that codify for this amino acid to the total number of codifying codons (not considering the termination codons). For example, Ala is codified by 4 codons, and the total number of codifying codons are 61; thus, the expected frequency for Alanine is P(ALA)= 4/61= 0.065. According to this calculation, the distribution of observed frequencies for each site was contrasted to that expected by means of a Student's t-test for proportions (Spiegel, 1975), with a significant level of 1%. The over-represented amino acids are indicated in the composition column. This over-representation of amino acids in some of the sites should skew the use distribution frequency from a random. It is interesting then to try to find the particular type of distribution of amino acid frequencies for each site on the CDRs, and to interpret their behavior or function in terms of their distribution.

A type of skewed distribution, which has been found in different biological phenomena (West, 1985; Nicolis, 1987) (as well as non-biological; for instance, in relation with linguistics (Nicolis, 1986) ,social sciences (Montroll and Badger, 1974), and several processes of physics (Montroll and Shlesinger, 1983; Meyer *et al.*, 1981)), is the inverse power law. This distribution was proposed by Pareto (1897) for the first time, in relation with the annual income distribution of the wealthy. An improved version of this relation proposed by Mandelbrot (1977) has the form

$$P(r) = K(r + \rho)^{-\beta} \qquad (1)$$

where K, ρ and ß are parameters (ß > 0) intrinsic to each system. In our system, P(r) is the probability of appearance of each amino acid with rank order r.

The linearized form of relation (1) was plotted for 21 sites on the heavy chain and 18 sites on the light chain. These plots of log P(r) versus log (r + ρ) were statistically tested for their fit to a straight line.

A characteristic of a skewed distribution, similar to equation (1), is that a small group of objects (amino acids) with higher frequency may share one or more properties, or have abilities in common, which explain their belonging to the group (Montroll and Shlesinger, 1982). An algorithm usually applied in artificial intelligence (Quinlan, 1983) was used to explore whether the more frequent amino acids may be clustered in terms of common physicochemical properties or attributes. This algorithm takes objects of a known class, described in terms of a fixed collection of properties or attributes, and produces a decision tree over these attributes that correctly classifies all the given objects. Each attribute has its own set of discrete values. For he present application, the objects are amino acids, the class is the most frequent amino acid in a site, and the attributes were defined in terms of the physicochemical properties of the amino acids (Sneath, 1966).

RESULTS

Tables 1 and 2 allow to visualize the likely rationality of the variability observed for each site, based on the complementation of structural and variability data. As it has been previously noted the 42% threshold criteria for conserved sites allowed the identification of 16 of the 18 structural sites responsible of the canonical conformations. Likewise, this criterion identified the majority of the sites proposed by Kabat *et al.* (1977). The third column of Tables 1 and 2, point to the existence of regularities in the sites responsible for interaction in the antigen-antibody complexes. In this recognition set, there are a group of sites that make contact in only one or two Of the eight complexes considered. Nevertheless, there is another group of sites that are common to the majority of the antigen-antibodies complexes. As can be seen in the same tables, these common contact sites are distributed in the three CDRs of both chains. This observation has been reported previously by Davies and Padlan (1990). Within this recognition set, 14 of the 20 sites in V_L that are in contact with the epitope, and correspond to the non conserved set (fifth column of Table 1); while, for V_H the frequency is 16 of 22. If the sites that are in contact in only one of the eight complexes considered are not taken into account, the data are 9 of 14 for V_L and 13 of 17 for V_H. Increasing the restriction to consider only the sites with more than 2 of the 8 complexes in contact, it is obtained 7 of 8 for V_L , and 11 of 12 for V_H are obtained.

If the sets of structural function (Structural and Kabat columns are joined and are compared with the total number of sites that have one amino acid with frequency greater than 42%, a total of 3 sites in V_L and 14 in V_H remain as unclassified. Within this group there are sites like V_H-32, with composition of Tyr 65%, Phe 11%; V_H-52b with Lys 87% or V_H-100j with Phe 72%, Met 22%. That is, there are 17 sites in both chains that have high conservation that has not been reported previously. Some of these sites are in contact with the epitopes.

In the composition column of both chains (Tables 1 and 2) it is observed a strange repetitive pattern in CDR-3. In the light chain Tyr is highly used in 4 of 9 sites and the combination Ser/Tyr within the same site is present in 3 sites. In the heavy chain, Tyr is highly used in 9 of 11 sites, and the combination Tyr/Gly is present in 6 sites.

The results of the Students t-test reported in the first and last columns of Tables 1 and 2, shows that all the sites that have been identified previously as non-conserved present at least one amino acid over-represented in the sample. As it can be seen in the last column, some amino acids that have apparently high percentage frequencies are not identified as over-used, because they have a great number of codons assigned in the genetic code. That is the case of Ser, Leu and Gly.

The results of plotting the logarithm of the probability of use of each amino acid versus the $\ln(r + \rho)$, i.e. the linearized equation (1), are illustrated in Figure 1 for some of the positions analyzed. The Figure shows a good fit with a straight-line.

On the other hand, certain positions fit an exponential distribution of the form

$$P(r) = \lambda \exp.(-\lambda r) \qquad (2)$$

with correlation coefficients greater than 0.95. In relation (2) λ represents a parameter; and the other variables have the same meaning as in equation (1). The results of these plots are shown in Figure 2.

By application of the Quinlan (1983) algorithm, to cluster the more frequently used amino acids in terms of their common attributes for each position, it is found that

there are two distinct groups: positions in which the more frequently used amino acids have similar physicochemical properties of hydrophobicity and volume, and positions in which the attributes of the most abundant amino acids are of a more general type, such as polarity, volume, or hydrogen bonding capacity (see Table 3). Generally, the positions of the former group fit equation (1), and those of the latter adjusted with equation (2). There were three exceptions(site 31 in V_L and sites 35 and 54 in V_H), which showed an almost equal fit to both the linearized inverse power plot and the exponential distribution. The correlation coefficients of these plots are shown in Table 4. These sites however were not considered as belonging to the inverse power-law group, because their more abundant amino acids did not have similar hydrophobicity and volume, and their frequencies of use were not as high as those of the amino acids of the positions belonging to the inverse power law distribution.

For the exponential distribution group the clustering in terms of attributes of a general type was possible for all the positions on V_H, with the exception of three cases (see Table 3). For V_L the clustering was possible only for three cases (Table 3).

DISCUSSION

The results describe some regularities in the recognition pocket, both at structural and sequence level. This type of study has been proposed by Kabat *et al.* (1977), Dildrop (1984), Ohno *et al.* (1985), Tainer *et al.* (1985), Chothia and Lesk (1987), Padlan (1990), Davies and Padlan (1990) and Mian *et al.* (1991).

Analysis of the data obtained from the crystallography of antigen-antibody complexes reaffirm the existence of common sites in the different specificities (Tables 1 and 2). But the vision is completed only by the observation that these common sites agree with hypervariable sites in almost all the cases. It is important that the contact sites are distributed in the six CDRs, as it does compete to the discussion about the relative role of each one of the CDR's in the genetic organization of immune recognition (Ohno *et al.*, 1985).

The results of plotting the logarithm of the probability of use of each amino acid reveal that apparently there are two different types of positions, at least on the CDR-1 and CDR-2, of the Igs (this plot was not attempted in the present work for the CDR-3, this study is in development and will be published elsewhere): those which better fit the inverse power law, and those which obey an exponential distribution. The more frequently used amino acids in the positions of the first group have similar properties of hydrophobicity and volume (Table 5), and those of the second group share different attributes depending on the position (Table 3).

The relative frequencies of use of the predominant amino acids for each position fitted by the inverse power law are shown in Table 5. From this data it can be observed that some amino acids are highly conserved, strengthening the suggestion that these sites should play fundamentally a structural role. For some positions, the frequencies of some amino acids is as low as 43 % (see site 63 of V_H in Table 5), nonetheless in these cases an amino acid second or third in order of abundance (secondary amino acids), with similar properties of hydrophobicity and volume, can be found in the same site. That is, these secondary amino acids have the same physicochemical capabilities for maintaining the general structure of the CDRs, which gives additional support to the suggestion of a structural role for these positions. Moreover, a comparison of our results with crystallographic data studies (Chotia, *et al.*, 1977), which propose the sites to be responsible for the canonical conformations for some of the hypervariable

loops shows 16 of the 18 structural sites to be the same (Table 5). Furthermore, the majority of sites proposed as structurals by Kabat *et al.* (1977) are also coincident with those listed in Table 5.

It has been shown that there exists a parallel between an inverse power law and a deterministic chaotic dynamics, with multiple coexisting attractors (Nicolis, 1991). This chaoticity evolves in time, and consequently for a long time a regular motion in the state space, suddenly interrupted by randomly distributed bursts of strong chaoticity, can be observed. Such dynamics results from an external noise which is instrumental in shifting an initial condition from a basin, where it normally belongs, to some other basin. In our case, the initial state is characterized by the physicochemical properties of a prototype amino acid, which due to an external noise (maybe an antigen) is drawn intermittently from attractor to attractor until an energy minimum is reached, or the external perturbation disappears, being then trapped by some other attractor. These attractors are separated by fractal basin boundaries, and the transition probabilities P_{ji} of jumping from attractor j to attractor i are proportional to the area of the basin attractor i, and the shape of the separatrices. These properties of the phase space generate two distinct types of memory (Nicolis, 1991): a global memory, resulting from a slow diffusion from one part of the attractor to another, and a memory resulting from the connectivity and communication among independent global memories (coexisting attractors). In the latter case, the magnitude of the noise should be enough to induce the transition between attractors (memories). In our particular case of antigen-antibody interaction, the perturbation possibly throws the initial state, characterized by the properties of a prototype amino acid, to a new state of the phase space, in which a new amino acid is that which minimizes the energy, and the system is trapped by a new attractor. This means that there exists a restricted trajectory which constitutes a memory and assures a strict function for these positions. This conservativity assures in our case the maintaining of structural features in the CDRs which are critical for the antigen-antibody interaction.

The above type of dynamics has been shown (Nicolis, 1991) to lead to a 1/f noise spectrum These fluctuations are time-scale analogous to the shapes of fractals, which have the property of self-similarity. Very recently, these types of patterns were also found in the intron sequences of DNA (Yam, 1992). A characteristic of these distributions is that the elements maintain a long range interaction, a feature expected among elements coherently involved in a structural function; given that, to maintain a specific spatial structure a long range coherence among the key building blocks of the structure should be necessary .

That the aforementioned sites mainly play a structural role is a striking finding, since it has usually been considered that sites 26-30 and 53-55 of V_H , and 26-32 and 50-56 of V_L were part of the recognition positions, being located in the loop region. With these findings, it can be proposed that the concept of a recognition site should be based not only on whether it is on a loop, but on the type of distribution followed by the amino acids in the site.

What arises from the above results is that if the discussed positions are critical in maintaining the canonical structures of the family, then the rest of the positions of the analyzed CDRs are those properly responsible for the recognition process. That is, those positions which better fitted an exponential distribution are those which should really be involved in the recognition processes. This experimental frequency distribution confirms the traditional view of random variation (Ohno *et al.*, 1985). Apparently, however there are some general restrictions which limit the range of amino acids allowed in the recognition sites of the hypervariable regions on the CDRs. The importance of these

restrictions resides not in the frequency of use of the allowed amino acids, but in the fact that they reduce the possible recognition solutions to those with similar attributes, as indicated in Table 5.

The above findings suggests that the recognition process, at least for those sites listed in Table 5 requires only general properties or attributes, rather than specific complementary amino acids. The general attributes, or physicochemical properties required for each site are satisfied by a reduced set of amino acids with a random selection.

The importance of having these general properties in specific positions may be illustrated by the case where large amino acid residues are the more frequently used. These may cover large spaces, facilitating their participation in a wide variety of van der Waals and electrostatic interactions to bind a range of epitopes (Mian, *et al.*, 1991). Another case is when the common attribute is a great flexibility which facilitates secondary interactions of macromolecules brought about by specific epitope binding or facilitating the molding of the Igs to different but closely related pathogens, and favoring the immune response. In this way amino acids with flexible side chains may generate structurally plastic regions able to mold themselves around the antigens, improving the interacting surfaces complementary.

The above finding seems to conciliate the proposals that the recognition positions may generate a wide number of solutions, given by a random selection of amino acids in these positions (Ohno *et al.*, 1985), with proposals of the existence of general interactions which constrain the amino acid selection to those sharing similar physicochemical properties (Mian *et al.*, 1991).

Finally, there are also those sites for which it was not possible to characterize the amino acids in terms of general attributes. In these cases, either the specificity of the amino acids in these sites play a fundamental role in the recognition process or there was a lack of identification of the appropriate attributes shared by the elements of the cluster.

A paper in the literature (Shenkin, 1991) reports that the relative use frequencies of amino acids on the CDRs of immunoglobulins fit reasonably with a log-linear frequency-rank distribution. Nonetheless, the main objective of that paper was to propose the use of information thermodynamical concepts to measure variability of the CDRs rather than to analyze the origin of the distribution and its deviations.

The results of this study seems to indicate that the codification for the preferential use of certain amino acids with analogous physicochemical properties in the recognition sites is indeed a feature of the Igs, suggesting the existence of a general mechanism for the recognition process, based in general properties instead of specific interactions Of course additional mechanisms involved in the complicated process of antigen-antibody recognition should be considered, such as the somatic hypermutation, which is very possibly the origin of specific or more fine interactions. The undoubtable existence of these and other mechanisms may explain the difficulties to classify clearly some of the sites within one of the two identified groups.

Acknowledgment

We gratefully acknowledge the valuable technical help of E. Salazar in writing the manuscript.

REFERENCES

Bashford, D., Chothia, C., Lesk, A.M. 1987. Determination of a protein Fold: unique feactures of the globin amino acids sequences. *J. Mol. Biol.* **196**, 199-216.

Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S., Air, G., Sheriff,S., Padlan, E., Davies, D., Tulip, W., Colman, P.M., Spinelli, S., Alzari, P.M., Poljak, R.J., 1985, Conformations of Immunoglobulin Hypervariable Regions, *Nature*, **342**, 877-883.

Chotia, C. and Lesk, A.M. 1987. Canonical structures for the hypervariable regions of immunoglobulins., *J. Mol. Biol.* **196**, 901-918.

Chothia, C., Novotny, J., Bruccoleri, R., Karplus, M. 1985. Domain association in immunoglobulin molecules: the packing of Variable domains. *J. Mol. Biol.* **186**, 651-663.

Davies, D.R., and E.A. Padlan, 1990, Antibody-Antigen Complexes, *Ann. Rev. Biochem.*, **59**, 439-473.

Dildrop, R. 1984. A new classification of mouse Vh sequences. *Immunol. Today* **5**, 85-86.

Go, M., Miyazawa, S. 1980. Relationships between mutability, polarity and exteriority of amino acids residues in Protein evolution. *Int. J. Peptide Protein Rec.* **15**, 211-224.

Grantham, R., 1984, Amino acid difference formula to help explain protein evolution, *Science*, **185**, 862-864.

Hunkapiler, T., Hood, L. 1989. Diversity of the Immunoglobulin gene superfamily. *Adv. Immunol.* **44**, 1-63.

Kabat, E.A., Wu, T.,T., and Bilofsky, H., 1977, Unusual Distributions of Amino acids in complementary-determining (Hypervariable) segments of Heavy and Light chains of Immunoglobulins and their Possible roles in specificity of Antibody-combining sites., *J. Biol. Chem.*, **252**, 6609-6616.

Kabat, E.A., Wu, T., Perry, H.M., Gottesman, K.S., Foeller, C., 1991, In: Sequences of Proteins of Immunological Interest. 5th. Ed. National Institutes of Health, Bethesda.

Lim, W. A., Sauer. R. T. 1989 . Alternative packing arrangement in the hydrophobic core of proteins. *Nature* **339**, 31-36.

Meijer, P.H.E., Mountain, R.D. and Souler, Jr, R.J.. eds., 1981, Sixth International Conference on Noise in Physical Systems, (National Bureau Standards, Washington, D.C., Special Publication No. 614..

Mian, I.S., Bradwell, A.R., Olson, A.J., 1991, Structure, Function and Properties of Antibody Binding Sites, *J. Mol. Biol.*, **217**, 133-151.

Montroll, E. and W.W. Badger, 1974, Introduction to Quantitative Aspects of Social Phenomena, Gordon and Breach Science Pub.

Montroll, E. W. and Shlesinger, M. F. 1982, On 1/f noise and other distributions with long tails, *Proc. Nattl Acad. Sci. USA*, **79**, 3380-3383.

Montroll, E. and M.F.Shlesinger, 1983, Maximum Entropy Formalism, Fractals, Scaling Phenomena, and 1/f Noise: A tale of tails, *J. Statist. Phys.* **32**, 209-230.

Nicolis, J.S., 1986, Chaotic Dynamics as applied to information Processing, *Rep. Prog. Phys.* **49**, 1109-1187.

Nicolis, J.S., 1987, Chaotic Dynamics of logical paradoxes in: Dynamical Systems and Environmental Models, eds. Bothe, Ebeling, Kurzhanski and Peschel (Academic-Verlag), pp. 105-113.

Nicolis, J. S. 1991, Chaos and Information Processing. A Heuristic Outline. World Scientific.

Novotny, J., Haber, E. 1985. Structural invariants of antigen binding: comparison of immunoglobulins V_L-V_H and V_L-V_L domains dimers. *Proc. Natl. Acad. Sci. USA.* **82**, 4592-4596

Ohno, S., Mori, N., Matsunaga, T., 1985, Antigen-binding Specificities of Antibodies are Primarily determined by Seven Residues of Vh. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 2945- 2949

Pareto, V., 1897, Course d'Economie Politique (Lausanne).

Padlan, E A 1979. Evaluation of the Structural variation among Light chain Variable domain *Mol Immunol.* **16** 287-296.

Padlan, E A 1990 On the Nature of Antibody Combining Sites: Unusual Structural Features That May Confer on These Sites an Enhanced Capacity for Binding Ligands. *Proteins* 7, 112-124.

Padlan, E A , Kabat, E.A. 1991. Modeling of the Antibody Combining Site. *Methods in Enzymol.* **203**, 3-21.

Quinlan, J.R., 1983, In Machine Learning: An Artificial Intelligence Approach, Vol. 1. Michalski, R. S., Carbonell, J.G., Mitchell, T.M. (Eds.), Morgan Kauffman Publishers Inc. pp. 463-482.

Sneath, P.H.A., 1966, Relation between chemical structure and biological activity in peptides. *J. Theor. Biol.*, **12**, 157-195.

Shenkin, P.S., Erman, B. and Mastrandrea, 1991, Information-Theoretical Entropy as a measure of Sequence Variability, *Proteins*, **11**, 297- 313.

Swanson, R., Trus, B. L., Mandel, N., Mandel, G., Kallai, O.B., Dickerson, R. E. 1985. Tuna cytochrome C at 2.0 A resolution. *J.Biol. Chem.* **252**, 759-775.

Tainer, J.A., Getzoff, E. D. Paterson,Y., Olson, A.J., Lerner.R.A. 1985. The atomic mobility component of protein antigenicity. *Ann. Rev. Immunol.* **3**, 501-535.

Vargas-Madrazo, E., Almagro, J.C., Lara-Ochoa, F., Jimenez Montano, M.A. 1992. Proceedings of the Seventh Panamerican Biochemical Congress, Ixtapa, Mexico.

West, B.J., 1985, An Essay on the importance of being nonlinear, Lectures Notes in Biomathematics, 62, Springer Verlag.

Wu, T.T., Kabat, E.A. 1970. An Analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for the antibody complementarily. *J. Exp. Med.* **132**, 211-250.

Yam, P. , 1992, Noisy Nucleotides, *Sci. Amer.*, Sept., 13-14.

Table 1. Variable domain light chain

	Site(A)	Structural(B)	Recognition(C)	Kabat(D)	Non-conserved(E)	Composition(F)
CDR-1	24			+		R48, S21, K16
	25	+		+		A55, S30, G10
	26			+		S87
	27			+		Q51, S21
	27a	+				S81, G11
	27b					L47, V19, I12, A11
	27c					V42, L32, D8
	27d*		1		+	H39*, N18*, Y15*, S12
	28*		1		+	N25*, D21*, S13, T8, V8
	29*		1	+	+	G30*, I23*, S21
	30*		4		+	S27*, N24*, V13, K12*
	31*		2		+	S31*, N27*, T23*
	32		5	+		Y72
	33	+		+		L58, M17, V10
	34*		1		+	A24*, H23*, D21*, S9
CDR-2	50*		4		+	G16, D15*, K13*, Y13*
	51*				+	A34*, T30*, V15
	52		2	+		S79, N8
	53*		2		+	K29*, T24*, S16, R13
	54			+		R48, L48
	55*		1	+	+	A37*, E13*, P12, F12*
	56*		1	+		S72, P8
CDR-3	89					Q53, A10, L8
	90	+	1	+		Q70
	91*		6		+	W22*, Y17*, G17*, S15
	92*		4		+	S20, Y17*, D16*, N15*
	93*		4		+	S40*, H14*, E12*
	94*	+	4		+	S18, Y15*, N14*, L14,V14
	95	+	2	+		P71
	96*		5		+	L21, Y20*, W19*, R12
	97			+		T78, V13
FR	2	+				I66, V15
	48	+				Y84
	49		2			I94
	64	+				G97
	71	+				F52, Y28, A14

1981

Table 2. Variable domain heavy chain

	Site(A)	Structural(B)	Recognition(C)	Kabat(D)	Non-conserved(E)	Composition(F)
CDR-1	26	+				G98
	27	+				F48, Y42
	28					T62, S22
	29	+				F76, L11
	30		2			T50, S41
	31		4			S48, D25
	32		3			Y65, F11
	33*		5		+	Y32*, W22*, G21*
	34	+		+		M61, I14, V8
	35*		1		+	H26*, N24*, S20, E9
CDR-2	50*		2	+	+	Y20*, R12, V10, A10, E9
	51		2	+		I88
	52*		6		+	N28*, S14, Y13*, R12, D11*, W8*
	52a	+				P56, N15, S10
	52b		1			K87
	53*		3		+	G26*, N21*, Y13*, D9, A8
	54*		2	+	+	N27*, G27*, S23*, D14*
	55	+	1			G61, S14, Y13
	56*		4		+	S23, Y17*, T16, G13, N9, D8
	57		1			T76, I10
	58*		3		+	N22*, Y19*, K16*, E10
	59			+		Y94
	60					N48, A21, S12
	61*				+	E27*, P18*, D15*, N14*, A13
	62*				+	K40*, S37*, A9
	63			+		F46, V28, L19
	64					K77, N9
	65					G61, S23, D9
CDR-3	95*		4		+	D14*, G14, S13, Y10
	96*		3		+	Y20*, G17*, R8
	97*		4		+	Y33*, G19*
	98*		3		+	Y30*, G21*
	99*		3		+	G25*, Y18*, S13
	100*		1		+	S20, G19*, Y13*
	100a*				+	S25*, G15, Y12*
	100j					Y49, A26, W14
	100k					F72, M22
	101			+		D73, A17
	102					Y72, V17
FR	47		2			W92
	71*	+			+	R39*, V30*, A12, K10
	94	+				R81

Table 1. Variable domain light chain

	Site(A)	Structural(B)	Recognition(C)	Kabat(D)	Non-conserved(E)	Composition(F)
CDR-1	24			+		R48, S21, K16
	25	+		+		A55, S30, G10
	26			+		S87
	27			+		Q51, S21
	27a	+				S81, G11
	27b					L47, V19, I12, A11
	27c					V42, L32, D8
	27d*		1		+	H39*, N18*, Y15*, S12
	28*		1		+	N25*, D21*, S13, T8, V8
	29*		1	+	+	G30*, I23*, S21
	30*		4		+	S27*, N24*, V13, K12*
	31*		2		+	S31*, N27*, T23*
	32		5	+		Y72
	33	+		+		L58, M17, V10
	34*		1		+	A24*, H23*, D21*, S9
CDR-2	50*		4		+	G16, D15*, K13*, Y13*
	51*				+	A34*, T30*, V15
	52		2	+		S79, N8
	53*		2		+	K29*, T24*, S16, R13
	54			+		R48, L48
	55*		1	+	+	A37*, E13*, P12, F12*
	56*		1	+		S72, P8
CDR-3	89					Q53, A10, L8
	90	+	1	+		Q70
	91*		6		+	W22*, Y17*, G17*, S15
	92*		4		+	S20, Y17*, D16*, N15*
	93*		4		+	S40*, H14*, E12*
	94*	+	4		+	S18, Y15*, N14*, L14,V14
	95	+	2	+		P71
	96*		5		+	L21, Y20*, W19*, R12
	97			+		T78, V13
FR	2	+				I66, V15
	48	+				Y84
	49		2			I94
	64	+				G97
	71	+				F52, Y28, A14

Table 2. Variable domain heavy chain

	Site(A)	Structural(B)	Recognition(C)	Kabat(D)	Non-conserved(E)	Composition(F)
CDR-1	26	+				G98
	27	+				F48, Y42
	28					T62, S22
	29	+				F76, L11
	30		2			T50, S41
	31		4			S48, D25
	32		3			Y65, F11
	33*		5		+	Y32*, W22*, G21*
	34	+		+		M61, I14, V8
	35*		1		+	H26*, N24*, S20, E9
CDR-2	50*		2	+	+	Y20*, R12, V10, A10, E9
	51		2	+		I88
	52*		6		+	N28*, S14, Y13*, R12, D11*, W8*
	52a	+				P56, N15, S10
	52b		1			K87
	53*		3		+	G26*, N21*, Y13*, D9, A8
	54*		2	+	+	N27*, G27*, S23*, D14*
	55	+	1			G61, S14, Y13
	56*		4		+	S23, Y17*, T16, G13, N9, D8
	57		1			T76, I10
	58*		3		+	N22*, Y19*, K16*, E10
	59			+		Y94
	60					N48, A21, S12
	61*				+	E27*, P18*, D15*, N14*, A13
	62*				+	K40*, S37*, A9
	63			+		F46, V28, L19
	64					K77, N9
	65					G61, S23, D9
CDR-3	95*		4		+	D14*, G14, S13, Y10
	96*		3		+	Y20*, G17*, R8
	97*		4		+	Y33*, G19*
	98*		3		+	Y30*, G21*
	99*		3		+	G25*, Y18*, S13
	100*		1		+	S20, G19*, Y13*
	100a*				+	S25*, G15, Y12*
	100j					Y49, A26, W14
	100k					F72, M22
	101			+		D73, A17
	102					Y72, V17
FR	47		2			W92
	71*	+			+	R39*, V30*, A12, K10
	94	+				R81

Table 3. Attributes which identify the cluster of the more frequently used amino acids for those positions which fitted the exponential distribution. The sites were clustered using the algorithm proposed by Quinlan (1983). For the classification, nine physicochemical properties or attributes, were considered: composition, polarity, volume, bulk, molecular weight, charge, hydrogen bonding capacity, aromatic ring. The amino acids were grouped in a discrete way for each one of these attributes. For sites 33, 53, 58 on V_H , and for site 28 on V_L , only one attribute was necessary to characterize the cluster. For the other sites two attributes were necessary.

Site*	Attributes
V_H	
33	VOLUME
35	VOLUME, COMPOSITION
53	BULK**
54	VOLUME, COMPOSITION
56	POLARITY, COMPOSITION
58	BULK
61	BULK, H-BONDING
V_L	
28	COMPOSITION
31	VOLUME, H-BONDING**
53	VOLUME, POLARITY**

*Site numeration according to Kabat *et al* (1977).
**In these sites, an amino acid was included in the final clusters which was not one of the more frequently used.

Table 4. Correlation coefficients which fit both functions (1) and (2).

Site	r_{ipl}*	r_e*
31	0.97	0.95
35	0.98	0.91
54	0.96	0.94

r_{ipl}* stand for correlation coefficient of the inverse power law, and r_e* for exponential distribution.

Table 5. Relative Frequencies of amino acids in percent with the highest constancy on the CDR-1 and CDR-2 of the heavy and light domains. The frequencies of amino acids for all positions listed fit an inverse power distribution.

Site*	Predominant amino acid	Relative frequency**	Secondary amino acids	Sum of frequencies
V_H				
26	G	98%		
27	F	48%	Y(42%)	90%
28	T	62%	S(22%)	84%
29	F	76%	L(11%)	87%
30	T	50%	S(41%)	91%
31	S	48%	T(5%)	53%
32	Y	65%	F(11%)	76%
34	M	61%	I(14%), V(8%)	83%
51	I	88%		
52a	P	56%	S(10%)	66%
52b	K	87%		
55	G	61%	S(14%)	75%
57	T	76%		
59	Y	94%	F(2%)	96%
60	N	48%	A(21%), S(12%)	81%***
63	F	46%	V(28%), L(19%)	93%
64	K	77%		
65	G	61%	S(23%)	83%
V_L				
24	R	48%	K(16%)	64%
25	A	55%	S(30%), G(10%)	95%
26	S	87%	T(2%)	89%
27	Q	51%	N(1%)	52%
27a	S	81%	G(11%)	92%
27b	L	47%	V(19%),I(12%),A(11%)	89%
27c	V	42%	L(32%)	74%
32	Y	72%	F(5%)	77%
33	L	58%	M(17%), V(10%)	85%
52	S	79%	T(7%)	86%
54	L	48%	R(48%)	96%***
56	S	72%	T(11%)	83%

*Sites 26-30 are considered for some authors (Chotia, *et al*, 1985) as part of CDR-1.
**Each relative frequency is defined with respect to the total number of Igs analyzed (1500).
***For these cases, both amino acids have different physicochemical properties; however, the sum of both frequencies shows a high constancy.

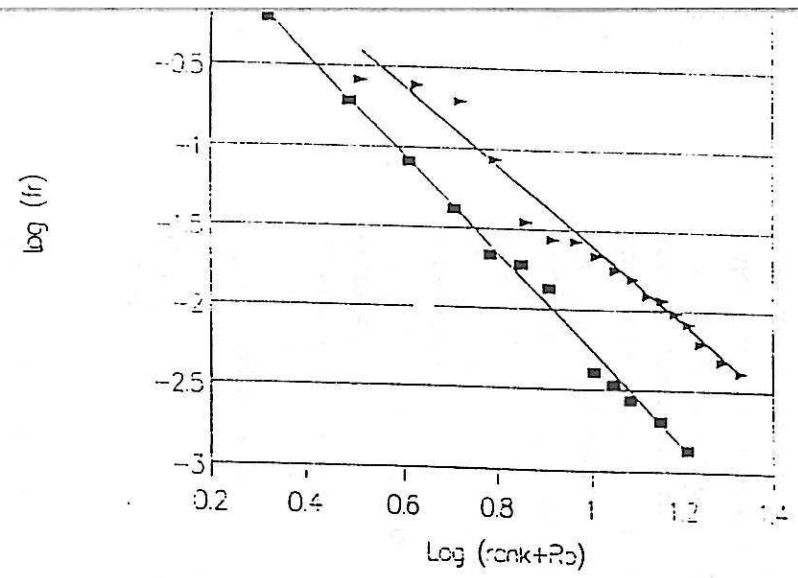
Figures and Tables footnotes.

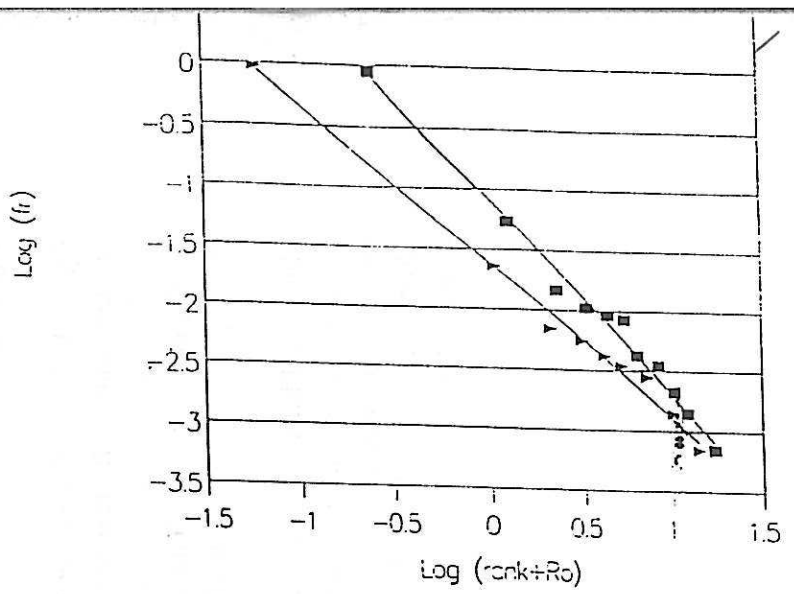
Table 1.
(A) Site numeration according to Kabat *et al.* 1991. Only are analyzed the sites of the Kabat's table that have at least 1/3 of the number of sequences considered (aprox. 1500). The hypervariable sites that have amino acids over-represented are marked with an asterisk. The framework sites analyzed are considered only if determine some canonical conformation or if are in contact with the epitope in some antigen-antibodly complex
(B) Sites that determine the conformation of some loops, according to Chothia and Lesk 1987.
(C) Sites that are in contact with the antigen according to crystallographic data (Padlan and Kabat, 1991). There are considered 8 complexes, it is indicated the number of complexes in which the site is reported as in contact.
(D) Sites proposed as conserved by Kabat *et al.* 1977.
(E) Sites identified as non-conserved because they not have amino acids with frequencies greater than 42% (see methods).
(F) The percentage composition of the most used amino acid for each site was calculated. The amino acids identified as over represented by the Student's t-test are marked with an asterisk. The single letter amino acid code are used.

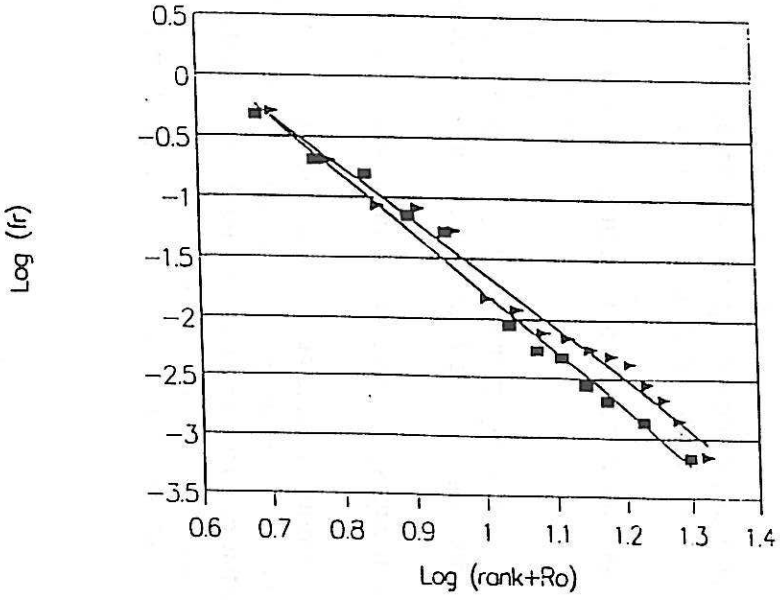
Table 2.
(A). In the CDR-1 of V_H the sites 26-30 are added to the sites considered by Kabat . because in this segment reside the hypervariable loop according to Chothia and Lesk (1987), in all the other CDR's the hypervariable loops are located inside the CDR's segments. All the other notation are the same as in the Table 1.

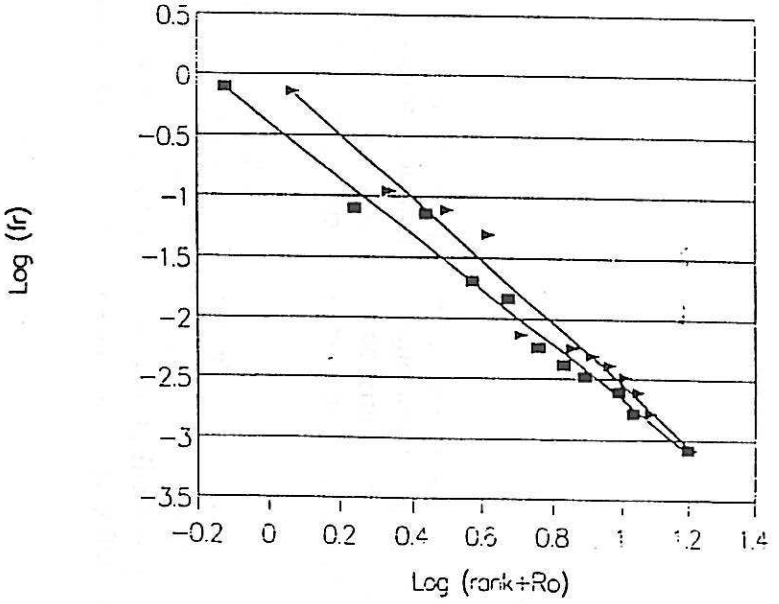
Figure 1. Plots of the logarithm of the rank of use of each amino acid versus the logarithm of their relative frequencies; for those sites identified as structurals. For all the positions, the fit with a straight line was very good, obtaining for all the cases correlation coefficients $r > 0.98$. The illustrated plots are: for the heavy chain a) site 33, b) site 53, site 56; for the light chain c) site 29, site 30, d) site 50, site 53.

Figure 2. Plot of the rank of the used amino acids versus the logarithm of their relative frequencies, for those sites identified as of recognition. For all the positions, the fit with a straight line was good, obtaining for all the cases correlation coefficients $r > 0.91$. The illustrated plots are: for the heavy chain a) site 34, site 35, b) site 51, site 59; for the light chain c) site 24, site 27, d) site 52, site 56.


Log (fr)
Log (rank+Ro)


Log (fr)
Log (rank+Ro)


Log (fr)
Log (rank+Ro)


Log (fr)
Log (rank+Ro)

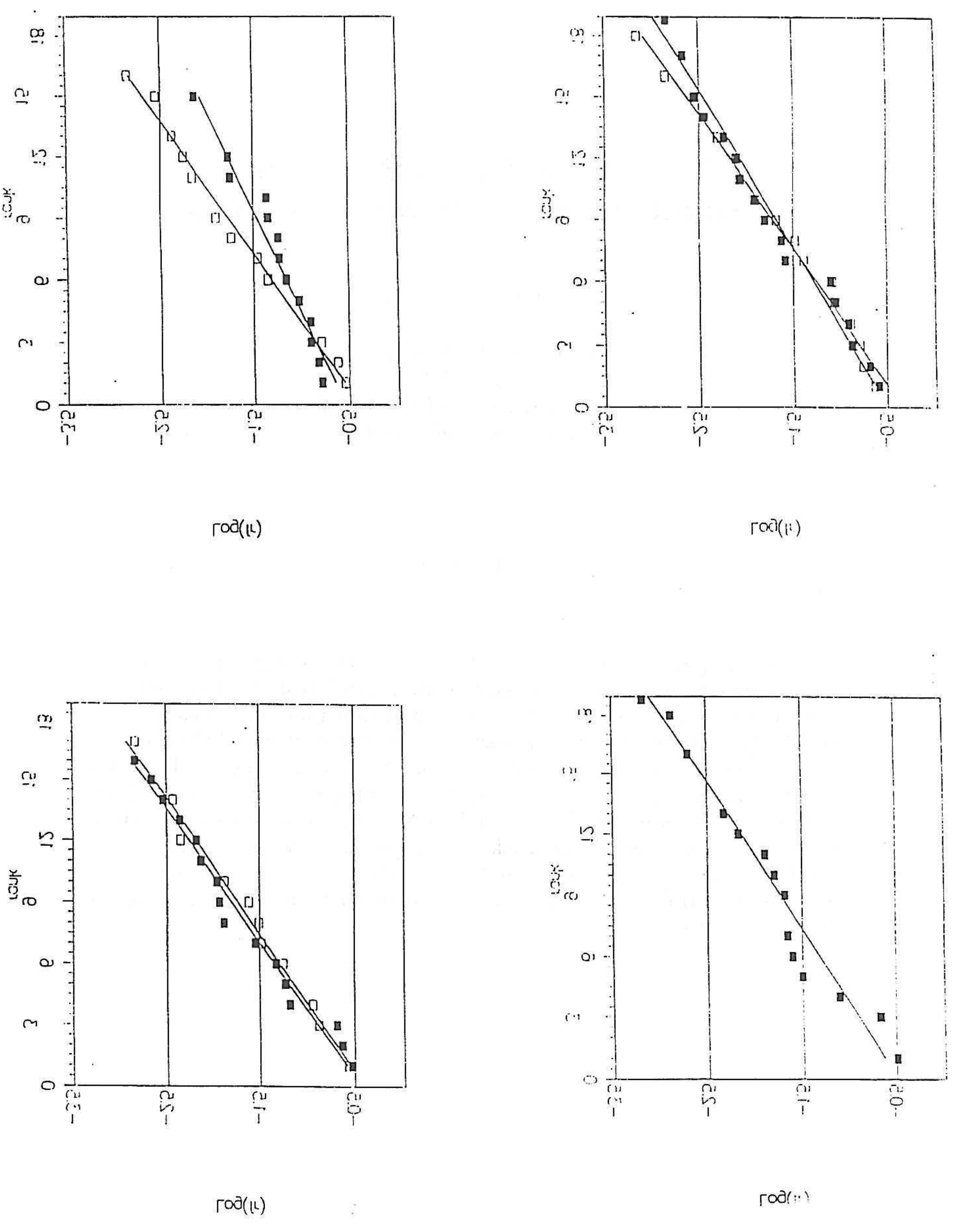


FIG 2