J Mol Evol (1994) 38:100–104

Journal of Molecular Evolution



# A Skewed Distribution of Amino Acids at Recognition Sites of the Hypervariable Region of Immunoglobulins

**E. Vargas-Madrazo,[1,2] F. Lara-Ochoa,[2] M. Jiménez-Montaño[3]**

[1] Instituto de Investigaciones Biológicas, Universidad Veracruzana, Carr. Xal-Ver Km. 2.5, Av. 2 Vistas s/n, Xalapa, Veracruz, México
[2] Instituto de Química, UNAM, Circuito exterior, Cd. Universitaria, Coyoacan 04510, México D.F.
[3] Universidad de las Américas, Cholula, Pue, México

Received: 16 November 1992/Revised and accepted: 3 May 1993

**Abstract.** Antibody binding site are formed by six hypervariable regions or complementarity determining regions (CDRs). The CDRs, three from the heavy chain and three from the light chain, are known as hypervariable segments and provide a surface complementary to that of the epitope. In recent work it was found that the amino acids in these positions fulfill different functions: Some play a structural role and others are involved in the specificity-determining function. It is reported here that the frequency of amino acids at hypervariable sites is skewed. By means of an informational algorithm, key physicochemical attributes of the dominant residues were identified for some of those sites. The results for about 1,500 antibodies suggest that approximately 35% of sites involved in the recognition process require only general properties such as composition, volume, and bulk or hydrogen bonding which are satisfied by a small set of amino acids instead of any one particular complementary amino acid.

**Key words:** Immunoglobulins — Hypervariable region — Pattern recognition

It has been proposed that the number of main-chain conformations of at least five of the six loops of the complementarity determining regions (CDRs) seem to be limited (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. This suggests that some of the amino acids on the CDRs may have a structural role, while the rest are involved in the recognition function or are irrelevant. Moreover, it has been proposed by Ohno et al. that there are only seven specificity-determining sites on the CDR-1 and CDR-2 of the heavy chain and that these follow a variability distribution far from random (Ohno et al. 1985). A question then is, do the amino acids responsible for recognition have general properties that enable them to bind a diversity of antigens, thus reducing the broad repertoire of possibilities to a few solutions?

To study this possibility, an analysis of the frequency of amino acids in 39 sites of the CDRs of the variable heavy domain and 31 of the variable light domain was performed in an alignment of approximately 1,500 immunoglobulins (Igs) (Kabat et al. 1991). The results of the analysis showed that some of the sites are occupied by highly conserved amino acids (column of conserved sites in Tables 1 and 2), which indicates that these sites participate in maintaining some structural characteristics of the recognition region of Igs (Kabat et al. 1977; Ohno et al. 1985; Padlan 1990; Mian et al. 1991). In some sites, the frequencies, in percent, of some of the amino acids are apparently not as high as in other cases of Tables 1 and 2 (for example, sites 27, 30, and 34 of

*Correspondence to:* E. Vargas-Madrazo

**Table 1.** Composition of the CDRs of the variable heavy chain

| Site[a] | Percentage composition[b] | |
|---|---|---|
| | Conserved sites | Hypervariable sites |
| CDR-1 | | |
| 26 | G98% | — |
| 27 | F48%, Y42% | — |
| 28 | T62%, S22% | — |
| 29 | F76%, L11% | — |
| 30 | T50%, S41% | — |
| 31 | S48%, D25% | — |
| 32 | Y65%, F11% | — |
| 33 | — | Y32%(20.60)*, W22%(17.09)*, G21%(11.48)* |
| 34 | M61%, I14%, V8% | |
| 35 | — | H26%(17.33)*, N24%(16.54)*, S20%(7.53)*, E9%(6.07)* |
| CDR-2 | | |
| 50 | — | Y20%(13.83)*, R12%(2.15), V10%(3.36)*, A10%(3.11)*, E9%(6.20)*, W6%(6.06)* |
| 51 | I88% | — |
| 52 | — | N28%(18.29)*, S14%(3.55)*, Y13%(9.82)*, R12%(2.05), D11%(8.12)*, W8%(7.96)* |
| 52a | P56%, N15%, S10% | — |
| 52b | K87% | — |
| 53 | — | G26%(13.72)*, N21%(14.02)*, Y13%(9.47)*, D9%(6.18)*, A8%(1.28) |
| 54 | — | N27%(17.54)*, G27%(14.37)*, S23%(9.37)*, D14%(10.20)* |
| 55 | G61%, S14%, Y13% | — |
| 56 | — | S23%(9.57)*, Y17%(12.31)*, T16%(8.09)*, G13%(5.64)*, N9%(5.99)*, D8%(5.54)* |
| 57 | T76%,I10% | — |
| 58 | — | N22%(15.26)*, Y19%(13.49)*, K16%(11.79)*, E10%(6.98)* |
| 59 | Y94% | — |
| 60 | N48%, A21%, S12% | — |
| 61 | — | E27%(17.67)*, P18%(9.18)*, D15%(10.83)*, Q14%(10.08)*, A13%(6.14)* |
| 62 | — | K40%(23.77)*, S37%(17.37)*, A9%(2.17) |
| 63 | F46%, V28%, L19% | — |
| 64 | K77%, N9% | — |
| 65 | G61%, S23%, D9% | — |
| CDR-3 | | |
| 95 | — | D14%(13.85)*, G14%(7.07)*, S13%(2.61)*, Y10%(7.28)* |
| 96 | — | Y20%(13.60)*, G17%(8.61)*, R8%(−1.40), D6%(3.59)* |
| 97 | — | Y33%(19.82)*, G19%(9.66)*, D6%(2.91)* |
| 98 | — | Y30%(17.77)*, G21%(10.60)*, D6%(3.67)* |
| 99 | — | G25%(12.05)*, Y18%(10.96)*, S13%(2.70)* |
| 100 | — | S20%(6.28)*, G19%(7.99)*, Y13%(7.72)*, D6%(2.71)*, W4%(2.60)* |
| 100a | | S25%(7.20)*, G15%(4.74)*, Y12%(5.77)* |
| 100j | Y49%, A26%, W14% | — |
| 100k | F72%, M22% | — |
| 101 | D73%, A17% | — |
| 102 | Y72%, V17% | — |

[a] Site number according to Kabat et al. 1991. Only those positions of Kabat's table, where at least 60% of the total number of sequences (aprox. 1,500) were gapless, were analyzed. In the CDR-1 of $V_h$, sites 26–30 are added to the sites considered by Kabat et al. because in this segment resides the hypervariable loop, according to Chothia and Lesk (1987). In all the other CDRs the hypervariable loops are located inside the CDR segments.

[b] The percentage composition of the most frequently used amino acids for each site was calculated. The values of the t-statistic are reported in parentheses. The amino acids in the hypervariable sites identified as overrepresented by the Student's t-test are marked with an asterisk. The considered significance level was of 1% (t = 2.58). The single-letter amino acid code is used

heavy chain and 24, 27b, and 27c of light chain). Nevertheless, in such cases a second or third amino acid in abundance was found at the site, having the same physicochemical properties (Lim and Sauer 1989) as those of the more frequent amino acid. Thus, presumably, in such cases the more abundant amino acids share the same ability to maintain the general structure of the CDRs. In contrast with this finding, it has theretofore usually been considered that the amino acids at sites 26–30 and 53–55 of the heavy-chain domain and 26–32 and 50–56 of the light-chain domain, being in the loop region, were part of the specificity-determining positions.

Our observations also suggest that if the positions discussed are critical for maintaining the canonical conformations or impose other restrictions

**Table 2.** Composition of the CDRs of the variable light chain

| | Percentage composition | |
|---|---|---|
| Site | Conserved site | Hypervariable site |
| 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 | — | H39%(13.40)*, N18%(7.32)*, Y15%(6.24)*, S12%(1.11) |
| 28 | — | N25%(15.27)*, D21%(13.20)* S13%(2.39), T8%(1.72), V8%(1.04), Y6%(3.46)* |
| 29 | — | G30%(15.73)*, I23%(18.86)*, S21%(;7.89)* |
| 30 | — | S27%(11.67)*, N24%(15.52)*, V13%(5.87)*, K12%(8.20)*, Y6%(2.95)* |
| 31 | — | S31%(13.76)*, N27%(17.26)*, T23%(12.35)* |
| 32 | Y72% | — |
| 33 | L58%, M17%, V10% | — |
| 34 | — | A24%(12.76)*, H23%(15.04)*, N21%(14.01)*, S9%(−0.95), Y6%(3.50)* |
| CDR-2 | | |
| 50 | — | G16%(7.63)*, D15%(10.12)*, K13%(8.50)*, Y13%(8.56)*, W4%(3.99)* |
| 51 | — | A34%(17.05)*, T30%(15.18)*, V15%(6.98)*, M4%(3.41)* |
| 52 | S79%, N8% | — |
| 53 | — | K29%(13.09)*, T24%(9.23)*, S16%(3.57)*, R13%(1.63) |
| 54 | R48%, L48% | — |
| 55 | — | A37%(18.08)*, E13%(8.74)*, P12%(4.29)*, F12%(8.36)*, H7%(3.74)* |
| 56 | S72%, P8% | — |
| CDR-3 | | |
| 89 | Q53%, A10%, L8% | — |
| 90 | Q70% | — |
| 91 | — | W24%(15.81)*, Y19%(11.58)* G19%(8.01)*, S15%(2.98), H7%(4.46)* |
| 92 | — | S20%(6.94)*, Y17%(11.04)*, D16%(10.51)*, N15%(10.17)*, T10%(2.84)* |
| 93 | — | S40%(17.40)*, H14%(9.24)*, E12%(8.38)*, T10%(2.91)* |
| 94 | — | S18%(5.96)*, Y15%(9.80)*, N14%(9.34)*, L14%(2.91)*, V14%(5.10)* |
| 95 | P71% | — |
| 96 | — | L21%(8.44)*, Y20%(12.50)*, W19%(13.44)*, R12%(1.81) |
| 97 | T78%, V13% | — |

Same as for Table 1, for the light chain

on the recognition pocket, the rest of the positions on the CDRs could be responsible for the recognition process or are irrelevant.

To analyze the nature of the frequency distribution of the hypervariable site set, the existence of amino acids overrepresented in the sample of sequences was tested, keeping in mind that (1) there are various sources of error in the database (Shenkin et al. 1991) and that (2) 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 code for this amino acid relative to the total number of codons (not considering the termination codons). For example, Ala is coded by four codons and the total number of codons is 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 with that expected by means of a Student's t-test for proportions (Spiegel 1975), with a significance level of 1% (t = 2.58). The overrepresented amino acids are indicated with an asterisk in the column of composition for the hypervariable sites in Tables 1 and 2. The results of the Student's t-test are reported in parentheses. The column of hypervariable sites of Tables 1 and 2 shows that all the sites that have been identified as hypervariable present at least one amino acid overrepresented in the sample. As can be seen in this column, some amino acids that have an apparently high frequency are not identified as overused because of the number of codons assigned to them in the genetic code. That is the case of Ser, Leu, and Arg. One may then question the traditional view that the CDRs are random hypervariable regions, since most of the recognition solutions for each site are given by a selection from a reduced number of amino acids of the skewed part of the distribution. Admittedly, however, the existence of mutational biases has not been excluded.

In order to test the possibility that the most fre-

**Table 3.** Sites with amino acids overrepresented with common physicochemical properties[a]

| Site | Attributes |
|---|---|
| Vh-33 | Size |
| Vh-35 | Size and composition* |
| Vh-53 | Bulkiness |
| Vh-54 | Bulkiness* |
| Vh-58 | H-bond |
| Vh-60 | Size |
| Vh-61 | Bulkiness |
| Vl-28 | Composition |
| Vl-31 | Size and H-bond |
| Vl-53 | Polarity and size* |

[a] Attributes that identify the cluster of the more frequently used amino acids. The sites were clustered in terms of one or two common attributes, which are shown in column 2. Sites marked with an asterisk have an amino acid that constitutes an exception in the cluster

quently used amino acids in the hypervariable sites share a common attribute or physicochemical property, an algorithm normally applied in artificial intelligence (Quinlan 1983) was used for the analysis. This algorithm allows the clustering of elements in a complex system in terms of their common attributes. By applying it to the 17 sites in the heavy chain and 15 sites in the light chain, and using as attributes those physicochemical properties commonly employed in protein structure studies (Sneath 1966; Zimmerman et al. 1968; Grantham 1974; Go and Miyazawa 1980), it was found that in some cases only a single attribute was necessary to characterize the cluster (Table 3). In 10 of the 17 sites analyzed in the heavy chain and in 12 of 15 sites from the light chain (most of them from the CDR-3), it was not possible to find a reduced set of attributes to characterize the cluster, although this could be due to various reasons: (1) Some of the sites may be irrelevant to the recognition process; (2) the imprecise rearrangement mechanism of the germ-line genes that generates the CDR-3 (Tonegawa 1983) may produce much more variability than found in the CDR-1 and CDR-2; and (3) the selected attributes do not represent all the physicochemical restrictions relevant for the recognition process (Quinlan 1983). The above results suggest that approximately 35% of the sites apparently involved in the recognition process require only amino acids with a general property (Mian et al. 1991)—in particular, composition, volume, bulk, or hydrogen-bonding capacity—instead of a specific complementarity amino acid.

One question that arises from the above results touches on the origin of the non-random distribution of some amino acids found at the hypervariable sites. Is this, indeed, a feature of the molecular recognition of Igs, or is it only a result of some bias existing in the sample analyzed (only 1,500 Igs)? In order to elucidate this question two subpopulations of sequences were tested: (1) all the reported sequences for the Igs germ-line genes were translated and aligned (Genbank, release 69), and (2) all the antimacromolecule antibodies were aligned (Kabat et al. 1991). The amino acid frequencies for each site were then analyzed for the type of distribution. The analyzed frequencies obey the same skewed distribution, similar to that found for the total protein sample (Vargas-Madrazo et al. 1992). This could indicate that the codification for the preferential use of certain amino acids with similar physicochemical properties in some recognition sites is indeed a feature of the Igs, suggesting the existence of a general mechanism for the recognition process. A more extensive study of the found skewed distribution reported here was published elsewhere (Lara-Ochoa et al. 1993).

*Acknowledgments:* We thank Dr. A. Lazcano-Araujo and Dr. C. Larralde for critical reading of the manuscript and for helpful discussions. EVM was supported by a grant from CONACYT.

## References

Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 196:901–918

Go M, Miyazawa S (1980) Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. Int J Peptide Protein Res 15:211–224

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Kabat EA, Wu TT, 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 EA, Wu TT, Perry HM, Gottesman KS, Foeller C (1991) Sequences of proteins of immunological interest, 5th ed. National Institutes of Health, Bethesda, MD

Lara-Ochoa F, Vargas-Madrazo E, Jiménez-Montaño MA, Almagro JC (1993) BioSystems (in press)

Lim WA, Sauer RT (1989) Alternative packing arrangement in the hydrophobic core of proteins. Nature 339:31–36

Mian IS, Bradwell AR, Olson AJ (1991) Structure, function and properties of antibody binding sites. J Mol Biol 217:133–151

Ohno S, Mori N, Matsunaga T (1985) Antigen-binding specificities of antibodies are primarily determined by seven residues of Vh. Proc Natl Acad Sci USA 82:2945–2949

Padlan EA (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

Quinlan JR (1983) Learning efficient classification procedures and their application to chess and games. In: Michalski RS, Carbonell JG, Mitchell TM (eds). Machine learning: an artificial intelligence approach, vol I. Morgan Kaufmann, San Mateo, CA, pp 463–482

Shenkin PS, Erman B, Mastrandrea LD (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 9:154–178

Sneath PHA (1966) Relations between structure and biological activity in peptides. J Theor Biol 12:157–195

Spiegel MR (1975) Probability and statistics. McGraw-Hill Book, New York

Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581

Vargas-Madrazo E, Almagro JC, Lara-Ochoa F, Jiménez-Montaño MA (1992) Proceedings of the seventh Panamerican biochemical congress. Ixtapa, México, September 27–October 2

Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for the antibody complementarity. J Exp Med 132:211–250

Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21:170–201