Trends in Antibody Sequence Changes during the Somatic Hypermutation Process

Louis A. Clark, Skanth Ganesan, Sarah Papp and Herman W. T. van Vlijmen

_J Immunol_ 2006; 177:333-340; doi: 10.4049/jimmunol.177.1.333

http://www.jimmunol.org/content/177/1/333

References

This article cites 32 articles, 12 of which you can access for free at:
http://www.jimmunol.org/content/177/1/333.full#ref-list-1

Subscription

Information about subscribing to _The Journal of Immunology_ is online at:
http://jimmunol.org/subscription

Permissions

Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts

Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Trends in Antibody Sequence Changes during the Somatic Hypermutation Process

Louis A. Clark,1 Skanth Ganesan, Sarah Papp, and Herman W. T. van Vlijmen1

Probable germline gene sequences from thousands of aligned mature Ab sequences are inferred using simple computational matching to known V(D)J genes. Comparison of the germline to mature sequences in a structural region-dependent fashion allows insights into the methods that nature uses to mature Abs during the somatic hypermutation process. Four factors determine the residue type mutation patterns: biases in the germline, accessibility from single base permutations, location of mutation hotspots, and functional pressures during selection. Germline repertoires at positions that commonly contact the Ag are biased with tyrosine, serine, and tryptophan. These residue types have a high tendency to be present in mutation hotspot motifs, and their abundance is decreased during maturation by a net conversion to other types. The heavy use of tyrosines on mature Ab interfaces is thus a reflection of the germline composition rather than being due to selection during maturation. Potentially stabilizing changes such as increased proline usage and a small number of double cysteine mutations capable of forming disulfide bonds are ascribed to somatic hypermutation. Histidine is the only residue type for which usage increases in each of the interface, core, and surface regions. The net overall effect is a conversion from residue types that could provide nonspecific initial binding into a diversity of types that improve affinity and stability. Average mutation probabilities are ~4% for core residues, ~5% for surface residues, and ~12% for residues in common Ag-contacting positions, excepting the those coded by the D gene. The Journal of Immunology, 2006, 177: 333–340.

The versatility of the mammalian adaptive immune system rests on its ability to generate high-affinity Abs to foreign Ags. The set of naive B cells expresses a repertoire of surface-bound Ab precursors called B cell receptors or BCRs, a subset of which can bind most new Ags with low affinity. To make this possible, the V(D)J recombination process has evolved to mix V, D, and J genes during cell development to provide a large sequence diversity in the germline L and H chains. Considering only the number of human V, D, and J genes and neglecting a gain sequence diversity in the germline L and H chains. This article must therefore be hereby marked charges. The work presented here seeks to better define the results of the somatic hypermutation process and to investigate sequence-related aspects of protein-protein recognition. Given a mature sequence, a probable germline sequence resulting from the recombination of species-specific V, D, and J genes (15, 16) can be derived using sequence matching and consideration of the known mechanistic rules of V(D)J recombination (see Materials and Methods). Once the germline chain sequence is known, simple comparison with the aligned mature sequence yields the position and type of mutation that occurred during the affinity maturation process.
Tomlinson et al. (17) have previously used a similar type of analysis to study the diversity of amino acids at specific positions in the germline and mature Ab sequences. They found that the frequency of somatic hypermutation and the diversity of the germline sequences is highest in the CDRs. Rather than focus on the mutation frequencies, we examine the type of mutation and its functional implications deduced from the location in the structure. The results indicate that residue type changes during the somatic hypermutation process are significant and have underlying functional rationales.

Materials and Methods

Germline V(D)J fitting procedure

The V(D)J germline fitting algorithm is intended to provide reasonable solutions for the germline Ab sequence based on the mature protein sequence for the large majority of cases. It is recognized that no such fitting procedure can be completely free of ambiguities. In some cases, the length of the D region or the lack of residues in many publicly available sequences gives insufficient information for a certain fit. The flexible joining mechanisms of the natural process also limit the reliability of the method. For example, the natural N nucleotide addition process inserts random bases into the sequence, making the concept of a precursor gene at those positions inapplicable. In all situations, the default algorithm behavior is to assume that there was no mutation and that the germline sequence at ambiguous positions is identical with the mature sequence.

The gene fitting algorithms for the H and L chains share some basic features. In each case, the mature protein sequence is converted to a most probable DNA sequence using mouse or human codon probability tables. Comparisons with the germline V(D)J gene sequences must be done with the DNA sequence, because the germline genes may be read in multiple reading frames (only D and J genes) depending on their position in the final sequence and because some sequences, especially D genes, are too short to reliably match using the reduced information in the protein sequences. Readers interested in the most exhaustive germline fitting procedure for the D gene region should consult the work of Monod et al. (18). Once the most probable mature DNA sequence is determined, V, J, and, if appropriate, D genes from the known library of genes are sequentially fit to the sequence. V(D)J genes are taken from the ImMunoGeneTics (IMGT) database (15, 16). The fitting is done exhaustively for individual gene possibilities by sliding the library gene along the mature DNA sequence and recording the number of matching bases at each offset. Although simple and easy to code, this approach is more reliable than a BLAST search because of the short sequences involved. It is affordable because of the small number of possible germline V(D)J genes; analysis of a L/H chain pair takes a couple minutes. Fitting for the whole data set can be done in 1–2 days on a computer cluster and could be done considerably faster if the method and code were to be optimized. It should be noted that a useful, fast, and accessible algorithm for blasting against the V genes (IgBLAST) is available on the National Center for Biotechnology Information website (www.ncbi.nlm.nih.gov/igblast).

The L and H chain matching is handled using a similar exhaustive procedure. In each case the V genes are matched first near the N terminus of the sequence, and the best match is retained. The unmatched portion of the sequence is then extracted, and the best J gene match is found using a similar procedure. For the L chains, this process is repeated for each possible V gene to account for the possibility that an overall long V gene was chosen because it simply adds extra base matches near the C-terminal end rather than improving the final fit. There are cases where a V gene with a lower match score leads to a better overall fit because it allows room for a better-fitting J gene. The H chain is handled similarly; for each possible V gene match the entire set of possible J genes is placed at the end of the sequence. For each V-J combination, the remaining unmatched sequence is fit to the D gene library. Most but not all H chains have a D gene segment inserted between the V gene and J gene portions.

The necessity in the flexible joining mechanism is accounted for by allowing two amino acids worth of flexibility during fitting. For example, after the V gene segment is fit, the unmatched portion to which the J gene will be fit is extended two residues (six bases) of mature DNA sequence into the V gene region. This is done to allow for the possibility that part of the germline V gene was deleted during splicing. For a similar reason the unique regions of the D gene region is extended six bases into the bordering V and J gene regions. Even with this flexibility, many D genes are often excluded because the gap between the V and J genes is too small. Conversely, if there is a gap between segments, then bases unmatched by the best gene are always assumed to be identical with those in the mature sequence, and no mutations are recorded.

Algorithm testing was performed in cases where the V(D)J fitting procedure was performed manually. Additionally, D gene fitting was tested on cases where the DNA sequence was known. In the small number of cases examined, translating from the protein sequence to the most probable DNA sequence was sufficient to distinguish between available D gene sequences. Typically, one or two related D genes had match scores clearly separated from all other lower-scoring D gene matches.

In some cases the algorithm clearly returns nonsensical matches, as indicated by long stretches of poor base matches. The failures can be traced back to humanized Abs, to Abs from species outside the human and mouse libraries used, or to insertions/deletions. Germline sequence determinations requiring >20-amino acid mutations relative to the mature sequence are discarded on the assumption that the problem is inappropriate to the sequence or that the algorithm has failed. Sequences with more than five consecutive mutated residues are discarded for similar reasons. Some sequences have so little information in the D and J gene regions that fitting is impractical. For this reason, solutions having mutations at >40% of the D and J gene region positions are also discarded.

Calculation of mutation probabilities

A simple estimate for the probability of making a transition from residue type i to residue type j (Pij) with a single base change can be calculated using the codon definitions and their known usage frequencies. A given number of codons, M (N) code for each residue type (i). For a given i to j transition, the probability is the sum of individual possible codon m to n transitions (Equation 1).

\[ P_{ij} = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{1}{12} f_m S_{mn} \]  

(1)

The factor (1/12) is the probability of picking one of the three bases in the codon (1/3) times the probability of picking the correct change (1/4). The frequency \( f_m \) is the fraction of residue type i that is coded using codon m in human genes. If codons m and n differ by only a single base, then such a transition, which could occur, and the function \( S_{mn} \) is unity. This simple transition probability \( P_{ij} \) can be renormalized using the residue type usage in the germline genes to form the background of Fig. 1A.

An estimate for the probability of random mutations leading to a cysteine pair capable of forming a new disulfide bond (Pcys) can be calculated from the probability of making a single cysteine mutation (Pcys = 1/19), the probability of a residue pair being close enough to form a disulfide (Ppair) and the average number of mutation pairs per chain (Nmutpair) (Equation 2).

\[ P_{\text{p}} = N_{\text{mutpair}} P_{\text{cys}} P_{\text{pair}} \]  

(2)

An average of eight mutations per chain yields 28 unique pairs (Nmutpair). Approximately 4% (Ppair ~ 0.04) of residue pairs have \( C_{\text{Cys}} \) separation distances of <6 Å. This leads to an estimate of 68 disulfide-producing events (Ppair ~ 0.003) in the ~22,320 sequences examined.

Regional assignments from sequence position

All Ab sequences were first aligned to the set (AAAAA, Aho’s Amazing Atlas of Antibody Anatomy) provided by Honegger et al. (19). We use this alignment and numbering system because it leaves gaps for almost all loop regions, gives a structural correspondence between light and heavy positions, and provides numerical values for all positions. The Kabat system uses alphanumeric numbering for some loop positions and can be cumbersome in some cases.

Position definitions were derived from Protein Data Bank structures of Abs in contact with peptide or protein Ags. Ab-Ag interface positions are occupied by residues that face the Ag and are within 12 Å of the nearest Ag atom for at least one Protein Data Bank structure. Because the positions were assigned based on a small set of structures, residues occupying Ab-Ag positions do not always contact the Ag for every sequence and may be solvent exposed. Residue positions on the VL-VH interface were defined similarly using an 8-Å cutoff but were required to meet this requirement in at least half of the structures. Core positions are those that point inward to the core of the variable domain and typically expose <10 Å. The surface positions are taken to be the remainder of the occupied positions. Some positions may have more than one assignment, particularly those that may both contact the Ag and be present at the VL-VH interface.

Downloaded from http://www.jimmunol.org/ by guest on April 14, 2017
Results
A large number of mature Ab chain sequences are available from public databases (20–22), including Ig sequences from IgBLAST (subset of the “nr” database with similarity of at least 50% over one-third of germline V gene length). After removal of duplicates and sequences not labeled as mouse or human, 23,116 heavy and 11,095 A or κ variable domain sequences were fit to known germline V, D, and J genes using the methodology described in the Materials and Methods section. Seventy-two percent of the light and 62% of the heavy sequences yielded plausible (see Materials and Methods section for definition) germline solutions and contributed to the results presented in this section. Approximately half of the sequences that contribute to the final dataset are human. All results are for the combined mouse and human datasets unless noted. Due to the higher uncertainty of fitting the short D gene segments to the mature sequence, all mutation information from the D gene region is omitted unless specifically noted.

Full dataset: mutation type and location
A comparison of the residue mutation type frequency with that observed in regular evolutionary relationships shows some basic similarity. Fig. 1A shows the prevalence of residue type change mutations in the full dataset (circles) using a color scale. For comparison, the expected mutation frequency has been calculated assuming a single DNA base change per codon and human codon usage. These expected frequencies are visible as the background color in each box. There is a clear qualitative relationship between expected and observed frequencies. An additional comparison with expected mutation frequencies comes from sequence alignment scoring matrices. Mutation frequencies can be derived from a multiple sequence alignment by observing residue type variability at a given position. If this information is compiled in a residue type-specific fashion and averaged over all positions, then a global view of how easily a given residue type can substitute for another can be compiled. The BLOSUM62 data provides such mutation frequencies derived from protein sequences with >62% sequence homology (23).

There is a similarly loose relationship between the observed frequency and the mutation frequency expected from the BLOSUM62 data as shown in Fig. 1B. From these tests we conclude that the residue type changes during somatic hypermutation are similar overall to those seen during evolution but may show deviation in the finer details.

In this overall dataset, there is some conservation of hydrophobicity (clustering around the diagonal in Fig. 1A), a relatively small number of mutations involving proline, cysteine, and tryptophan amino acids, and a number of scattered higher-probability mutation types. Comparison of observed Ab (Fig. 1A, overlaid circles) with the expected (Fig. 1A, background squares) mutations indicates that at least some of the higher-probability mutation types, such as T→S and A→V, have similar frequencies. Others, such as F→L and Y→G transitions, are less frequently and more frequently seen, respectively, in Abs. Further trends will become more pronounced in region-specific subsets of the data.

There is often a clear imbalance in the frequency of mutations from a given residue type to another (X→Y) compared with the reverse (Y→X). This imbalance is apparent in the lack of symmetry in the matrix and is presented in Fig. 2, which shows the same data processed to show deviations from the average of X→Y and Y→X entries. Part of the imbalance can be accounted for by differences in codon usage. For example, K→R mutations are 3.3 times more common than R→K mutations, and this difference could be explained by the 6:2 ratio of R to K codons and slightly lower relative usage of arginine in the germline. Other imbalances, such as A→V, have no codon number or germline usage bias. For comparison, the BLOSUM matrices and other residue substitution scoring matrices are symmetric by construction.

Fig. 3 shows the probability of finding a mutation at a given position in the variable domain of the L chain or the H chain. Each residue position has been assigned to an Ab structural region using the sequence alignment and crystal structures (see Materials and Methods). There is a clear relationship between the position assignment and the mutation probability. Ab-Ag interface and solvent-exposed surface positions are more likely to be targets of productive (nonsilent) mutations. The distributions agree qualitatively with those from Tomlinson et al. (17), but it is now possible to see that the previously unanalyzed H3 loop has more mutations than the other CDR loops even when mutations in the D gene region are omitted. In agreement with these results, we also see a sizable mutation probability (>10%) for the first three positions on the N terminus of the L chain. Despite being outside of the traditional CDR loops, these residues are often in contact with the Ag
Average mutation frequencies broken down by region are 4, 7, 12, and 5% for the core, VL-VH interface, Ab-Ag interface, and surface, respectively.

The germline sequences were analyzed for four-base RGYW and WRCY (see Ref. 3 for definitions) intrinsic mutation hotspots. There is a proximity relationship between hotspot and mutation position. Fig. 4 shows the distance expressed in the number of bases between the mutated codon center and the nearest base of a hotspot. For example, a distance of two bases means that the codon precedes the four-base hotspot and that they are adjacent with no bases separating them. At a distance of −1, the codon is after the hotspot with its first base forming the last base of the hotspot. The mean distances to the RGYW and WRCY hotspots are 12 and 10 bases, respectively, and mutations can be seen to be much more likely near hotspots. There is a clear periodicity of 3 in Fig. 4. This means that the first three bases (for example, RGY) of the hotspots tend to be in-frame. Specific analysis shows that 62% of RGYW hotspots and 70% of WRCY hotspots are in-frame.

Despite the tendency for mutations to be near hotspots within specific Ab sequences, there is no apparent relationship between the overall frequency of mutation at a given position and the probability of a hotspot occurring at that position. The CDRs do not have a higher density of intrinsic mutation hotspots (data not shown). It is not rare for the probability of finding a hotspot to approach unity, meaning that some residue positions are nearly always part of hotspots.

Region-specific mutation type and location
The region of the variable domain where a mutation occurs strongly determines its effect, if any, on the function of the Ab. To a first approximation, the residues on the Ab-Ag interface determine specificity and affinity, whereas those in the core partially determine stability. Mutations at the interface between the two variable domains will also contribute to stability (24) and binding affinity (25) and may contribute to the relative orientation of the domains. Surface mutations outside the interfaces may affect solvation properties and aggregation propensity but are expected to have little effect as long as polar surface area is maintained.

Fig. 5 gives the distribution of number of mutations within each region. The residue positions were distributed into the Ab-Ag interface, VL-VH interface, core, and surface regions using analysis of Protein Data Bank Ab-Ag complex structure as described in the Materials and Methods section. It can be seen that the mutations...
on the Ab-Ag interface span the broadest range, with a mean of 3.1 mutations per chain. Surface and VL-VH mutations are less frequent at 2.3 and 1.5 mutations per chain on average. Core mutations are least frequent at 1.3 mutations per chain. The distributions for the L and H chains are qualitatively similar (not shown). For the H chain, region-specific averages are ~30% higher than those in the L chain. On average, L and H chains have 6.8 and 8.6 mutations per chain, respectively.

As a reference point for describing composition changes, Fig. 6 shows the region-specific amino acid compositions in the mature sequence. The germline compositions are qualitatively similar. As should be expected, polar residues are favored on the surface and hydrophobic residues are favored in the core regions. One of the largest contributions to Ab-Ag interface composition comes from tyrosine at 15%. Glycine and serine are also strong contributors at 11 and 18%, respectively. As a comparison for only the interface with the Ag, the data of Lo Conte et al. (26) from an analysis of Ab-Ag crystal structures are also shown. Their numbers are the percentages of the interface area contributed by each amino acid type and, thus, are lower in value for smaller residues such as glycine and serine.

The net creation and removal of amino acid types from the germline to the mature sequence varies considerably by residue type. Figs. 7 and 8 give the percentage increase of each amino acid type in the overall, surface, core, and Ab-Ag interface regions for mouse and human sequences, respectively. Some of the most striking net changes occur at the Ab-Ag interface. Tyrosine residues undergo the greatest change in absolute numbers, partially because they are so prevalent in the germline structure. They are reduced by 13% (10% in mouse and 18% in human) from their germline frequency. The loss is distributed throughout the variable domain but is more pronounced in the CDRs and, in particular, at Ag-contacting positions of the H3 loop. Tyrosine is over-represented in the germline JH genes for both human and mouse. This has been noted by Zemlin et al. (Fig. 3 of Ref. 27). One of the six human JH genes (JH6) and its alleles can code for strings of 4–6 tyrosines. The JH6 gene is used in 22% of the germline sequence solutions found in our analysis. The JH4 gene can code for two tyrosines separated by two residues and is used in 37% of human H chain germline gene fitting solutions. This overall tyrosine change is large enough to skew the average side chain size change at the H
chain to a net loss of 3.4 heavy atoms over the whole Ab-Ag interface. For comparison, there is a net gain of 1.1 heavy atoms in the corresponding region of the L chain. Glutamine, serine, and tryptophan residues are also depleted on the interface. Several residue types are used more frequently in the mature Ab on the interface with the Ag. Phenylalanine is the most enhanced residue at the interface, increasing by \(\frac{80}{80}\%\) (31% in mouse and 104% in human) from the germline composition. Proline usage also increases during maturation by 42% (32% in mouse and 50% in human), as does histidine (36% in mouse and 82% in human). The residues showing these large increases in mature Ab use are almost absent in the germline but occur significantly in mature sequences (see Fig. 6).

Proline usage changes in L and H chains are seen mainly in turn regions and are qualitatively very similar between chains. Residue positions 15, 48, and 98 (Kabat light, 14, 40, and 80; Kabat heavy, 14, 41, and 84) are in the \(\beta\)-hairpin turns nearest the constant domain and show pronounced mutation activity. The turn in the L3 and H3 loops tends to gain prolines at position 135 (Kabat light, 95e; Kabat heavy, 100h). Similarly, prolines tend to appear at position 52 (Kabat light, 44; Kabat heavy, 45), which is a kinked position in the strand that forms part of the \(V_L-V_H\) interface.

Other interesting composition trends occur during maturation. The most striking is an increase in the relative number of cysteine residues on the surface. The low abundance of cysteines on the surface means that 181 net mutation events in the human dataset translate to a 178% increase (see Fig. 8B). One expects that additional cysteines could lead to creation of nonnative disulfide bonds and inhibit expression. Presumably, the new cysteines observed in this work do not have a destabilizing effect large enough to be selected against. In a small number (<50) of cases, two cysteines are apparently created, and in 12 cases they have sequence positions that could lead to the formation of an additional stabilizing disulfide bond. A probability estimate of this event occurring randomly indicates that six times more potential disulfide bonds should have been observed (see Materials and Methods). The discrepancy could be partially attributed to the observed order of magnitude lower probability of mutating to a cysteine than the random mutation probability used in the calculation. Histidine residues are gained consistently in each of regions for both species except on the surface of the mouse Abs. There is an \(\sim 80\%\) increase in histidine usage on both the surface and Ab-Ag of human Abs during maturation. For the mouse, the increase on the Ab-Ag interface is less pronounced at 36%. This increase in the number of histidines is the largest of the core residue position changes (Figs. 7C and 8C), where comparatively little composition change is seen.

It is reasonable to ask whether the residue types in which the hotspots occur (RGYW or WRCY) bias the residue composition changes. According to Fig. 4, the highest probability location for a mutation is inside a hotspot. Consequently, one might expect a tendency for residue types that are consistently mutated to be more frequently present in hotspots. Fig. 9 shows the fraction of each germline residue type in each of the RGYW or WRCY motifs. More than 50% of the tyrosine residues in the germline are located in WRCY motifs. Generally, the residue types that are lost on the Ab-Ag interface (glutamine, serine, tyrosine, and tryptophan) all have some of the highest propensities to be coded in hotspots.

FIGURE 8. Amino acid composition changes during the human somatic hypermutation process expressed as percentage changes from the germline. For reference, the expected direction (+ or −) of net change from the simple model (see Materials and Methods) is shown above each bar. Error bars are at one SD.

FIGURE 9. Tendencies to find specific residue types in the mutation hotspot motifs. The bar heights are the fraction of a specific germline residue type that is coded within the four-base (two incomplete residues) motif.
The most important region for Ab affinity maturation is the interface with the Ag. As seen in Figs. 7A and 8A, trends in the residue usage are apparent, but no information is offered about the specific mutations that occur from given germline residue types. Fig. 10 shows the intrinsic mutation probabilities for residues on the interface between the Ab and the Ag. On the Ab-Ag interface, some pronounced residue type changes are E→D and K→R, which are charge conserving. Others, such as N→D and Q→E, conserve size but change charge or change predominantly the hydrophobicity, such as S→T and F→L.

A germline tyrosine residue at a potentially Ag-contacting position has a 13% chance to mutate to another residue type (see Figs. 7A and 8A for species-specific numbers). This tendency is also apparent in Fig. 10, which shows that tyrosines in the germline tend to have a substantial intrinsic tendency to mutate. In the unnormalized Ab-Ag interface data, the horizontal tyrosine stripe is the most pronounced feature. Conversions from tyrosine and serine to other residue types are the dominant mutation types occurring on the Ab-Ag interface. When germline tyrosine is mutated, it has a tendency to convert to aspartate, histidine, glycine, asparagine, serine, and phenylalanine. Of these residue types, each can be reached with a single DNA base change, except glycine. When a germline serine is mutated, larger residues such as asparagine and threonine, both still capable of hydrogen bonding, are common replacements. Tryptophans are often converted to leucines and tyrosines, but also may become small residues such as serine and glycine.

Discussion

With the results presented here, it is possible to speculate on the underlying reasons for the mutation propensities and what they tell us about Ab maturation. Why is tyrosine, serine, and tryptophan usage on the Ab-Ag interface decreased during the maturation process? Similarly, why is there a substantial relative increase in the usage of histidine, proline, and phenylalanine?

One of the most striking results shown is the overall tendency for germline tyrosine, serine, and tryptophan residues to mutate to other residue types. Germline genes are biased to place these versatile residue types on the Ab-Ag interface. We speculate that the tyrosines are there to promote low-affinity binding of the naive germline Abs to new Ags. Tyrosine can provide hydrogen bonding and substantial hydrophobic interactions. Tryptophan is the most hydrophobic of residues and arguably the best for nonspecific interactions. Serine can hydrogen bond while minimizing potential steric conflicts. Both tyrosine and serine have a polar character and are likely to help maintain solution stability. In support of these ideas, Fellouse et al. (28, 29) have used phage display techniques to show that tyrosine and a combination of small flexible residues are sufficient to recognize a number of Ags, including human and mouse vascular endothelial growth factor molecules.

Not only does nature bias the interface with promiscuous binding residue types, but it provides a maturation strategy capable of making efficient and productive changes to them. Tyrosine, serine, and glycine residues are well represented in germline mutation hotspots (Fig. 9), increasing their tendency to mutate. The presence of serines in hotspots has been previously noted and suggested to be a strategy for improving the targeting of CDRs (30). If one examines into which residue types the tyrosines mutate, one sees that they most often become the small residues glycine and serine. Aspartate and asparagine, two short-chain charged and uncharged residues, are also favored, perhaps to provide ion-pair and H-bond interactions. Phenylalanine may be favored because of its similarity to tyrosine and greater hydrophobicity. Tryptophan often becomes arginine, glycine, serine, and leucine. It may be converted to eliminate steric conflicts and minimize nonspecific interactions that could lead to aggregation.

The low usage of histidine and methionine at the Ab-Ag interface (Fig. 6) is puzzling. Next to proline and cysteine, which often play specific structural and functional roles, histidine and methionine are least used on the interface. The work of Lo Conte et al. (26) shows that these residues are used less on Ab-Ag interfaces than on other protein-protein interfaces. One possible explanation for the low use of histidine and methionine is the hypothesis that they have only recently been introduced into the amino acid repertoire on an evolutionary time scale (31). These amino acids may be underused relative to their potential. If under-utilization in the germline can explain their low usage, then one would expect their usage to increase in the mature Abs. Figs. 7 and 8 shows that this...
is consistently the case for histidine in most regions, but only at the Ab-Ag interface for methionine. A more structurally oriented hypothesis can also be used to explain low methionine usage on the interface. Methionine has a long side chain and is likely to lose comparatively more entropy on binding relative to its gain in enthalpy, resulting in net lower binding affinity. Relative to methionine, histidine usage may be favored at interfaces, because it can provide both hydrogen bonding and hydrophobic interactions.

One may also speculate on the usage of certain amino acids to modify loop characteristics with benefits for Ab stability and affinity. Proline residues are frequently created on the Ab-Ag interface (Figs. 7A and 8A), particularly in the H3 loop where they could stabilize beneficial loop conformations. Loop reorganization should reduce the entropy cost of binding and increase affinity (see Ref. 32 for a possible example). Similarly, Ab maturation sometimes leads to the introduction of disulfide bonds, which may stabilize the Ig fold or result in subtle conformational changes that lead to higher affinity. An example may be the anti-gp120 peptide complex structure (1ACY), where an additional disulfide is formed between CDR H1 and H2, which actually contacts the Ag (33). It is also notable that many of the excess tyrosines on the Ab-Ag interface of the germline chains are converted to glycines (see Fig. 10). In this case, the creation of glycines may allow for the backfolding flexibility necessary for the remaining tyrosines or other nearby Ag-directed side chains to better contact the Ag.

Ab-Ag interaction systems are ideal for examining factors and strategies for improving binding affinity at protein-protein interfaces. The trends uncovered and discussed in this work illuminate the strategies that nature uses to bias immature Ab properties and subsequently refine them during the affinity maturation process. Residue type changes are clearly biased (Fig. 1A), by which codons are accessible given a single base change. Serine, tyrosine, and tryptophan are over-represented in the germline at common Ab-Ag positions and are some of the most frequently coded in the known mutation hotspot motifs. Remaining influences on the residue type changes are presumed to result from functional selection during the affinity maturation process. These functional selection aspects make the somatic hypermutation process a good model for evolution on an accelerated time scale.

Acknowledgments

The advice and supporting environment provided to L.A.C. as postdoctoral fellow at Biogen Idec are greatly appreciated. Helpful comments from The Immunoglobulin Facts Book. Aca- demica Press, London.


