UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

COLEGIO DE CIENCIAS Y HUMANIDADES UNIDAD ACADEMICA DE LOS CICLOS PROFESIONALES Y DE POSTGRADO

PROYECTO DE INVESTIGACION BIOMEDICA BASICA

INSTITUTO DE INVESTIGACIONES BIOMEDICAS

ESTUDIO DE LA RELACION ENTRE LA SECUENCIA, ESTRUCTURA Y LAS PROPIEDADES DE RECONOCIMIENTO MOLECULAR DE LOS ANTICUERPOS.

TESIS

QUE PARA PRESENTAR EL GRADO DE

DOCTOR EN CIENCIAS BIOMEDICAS

PRESENTA:

BIOL. MARIA DEL CARMEN RAMIREZ BENITEZ

MEXICO, D. F.

OCTUBRE 2001

A mis Padres
Alberto Ramírez y Felicitas Benítez
Por su amor

A mis Hermanos
Por su cariño incondicional

Agradecimientos

Al Dr. Juan Carlos Almagro Domínguez,
Por brindarme la oportunidad de realizar este proyecto bajo su asesoría

Al Comité Tutorial
Dr. Juan Carlos Almagro Domínguez
Dra. Adela Rodríguez Romero
Dr. Xavier Soberón Mainero
Por la enseñanza académica, que ayudó a mi formación dentro de este campo de la investigación.

Al Jurado de Tesis
Dr. Carlos Larralde Rangel
Dra. Adela Rodríguez Romero
Dr. Alejandro Alagón Cano
Dra. Goar Gevorkian Markosian
Dr. Edmundo Lamoyi Velázquez
Dr. Lorenzo Segovia Forcella
Dr. Mario Soberón Chávez
Por sus observaciones y consejos para la mejor presentación de este trabajo

Al Consejo Nacional de Ciencia y Tecnología
Por la beca que se me otorgó para la realización de este proyecto

A mis Amigos
Dra. Teresa Hernández Quiroz, M. en C. Gabriela Flores Ramírez, M. en C. Luis Ledezma Candanoza, M. en C. Deyanira Fuentes Silva.
Por su amistad y cariño

Al Dr. Gabriel Moreno Hagelsieb
Por haber compartido conmigo sus conocimientos y experiencias en el ámbito de la ciencia

A todos mis compañeros del Instituto de Biotecnología
Por su entusiasmo y ayuda durante mi estancia.

A la familia López
A quienes siempre agradeceré su confianza, cariño y apoyo

Resumen

El avance tecnológico ha generado un incremento exponencial de la información relacionada con las secuencias y las estructuras de anticuerpos. A pesar de ello, actualmente no es posible predecir la especificidad de un anticuerpo partiendo de su secuencia primaria y/o de su estructura. Estudios de varios grupos de investigación alrededor del mundo, incluyendo el nuestro, sugieren la existencia de reglas en el reconocimiento molecular mediado por anticuerpos. Esto pudiera ser la semilla de esquemas predictivos que relacionen la secuencia de amino ácidos de un anticuerpo con su capacidad para reconocer cierto antígeno.

Con la finalidad explorar si existen reglas o no del reconocimiento molecular mediado por anticuerpos, se analizó gran parte de las secuencias y estructuras de anticuerpos conocidas hasta la fecha. Para ello se construyó una herramienta de computo denominada VIR.II. VIR.II es una página de "World Wide Web" que permite un acceso rápido y sencillo a las secuencias de anticuerpos compiladas en la base de datos de secuencias de interés inmunológico, también conocida como la base de datos de Kabat. VIR.II permite además clasificar la especificidad de los anticuerpos en grupos tales como anti-proteínas, anti-péptidos, anti-haptenos, etc. Esto, junto con otra interfaz que permite compilar y analizar las estructuras tridimensionales de anticuerpos disponibles en el Protein Databank (PDB), nos hizo posible analizar y correlacionar patrones en la secuencia, la estructura y la función (especificidad gruesa) de los anticuerpos.

Comprobamos la existencia de reglas del reconocimiento molecular mediado por anticuerpos. Se describe como los anticuerpos reconocen a las proteínas, péptidos y haptenos. En el caso de los anticuerpos anti-proteína, la superficie del sitio de interacción con el antígeno es plana, lo que proporciona una superficie complementaria a los antígenos grandes. En el caso de los anticuerpos anti-péptido, debido a su tamaño menor con respecto a las proteínas, una ranura en el sitio de unión con el antígeno asegura una complementariedad adecuada con estos ligandos. Una observación interesante que emergió de este trabajo es el modo en que los anticuerpos reconocen haptenos. Debido al tamaño

pequeño de estos ligandos, no se requiere de cambios importantes en la topografía del sitio de unión al antígeno. Al parecer, sólo se requieren pequeños cambios en la topografía del sitio de interacción con los antígenos, mismos que pueden realizarse a través de cambios conformacionales o substituciones de las cadenas laterales durante el proceso de maduración de la respuesta inmunológica.

Finalmente analizamos el papel de la hipermutación somática en sustituir los residuos en contacto con el antígeno. Encontramos que los residuos en contacto con los antígenos son raramente modificados por el proceso de hipermutación somática. Una posible explicación para este patrón es que las mutaciones en los residuos en contacto con el antígeno son dañinas en un número importante de casos. Así, son eliminadas del "pool" de genes durante el proceso de maduración de la respuesta inmunológica. Esta explicación tiene, sin duda, implicaciones para las teorías que explican el origen y la evolución del repertorio de anticuerpos. También tiene consecuencia prácticas. Esto es, si se desea madurar la afinidad de un anticuerpo *in vitro*, lo que se debe hacer es identificar los residuos en contacto; pero en vez de mutar estos, lo que se debe mutar son lo vecinos.

Abstract

The technological advance has generated an exponential increase of the information related to the sequences and the structures of antibodies in the last few years. However, nowadays is not possible to predict the specificity of an antibody starting from its aminoacid sequence or even from its structure. Studies of several research groups around the world, including our own, have suggested the existence of rules in the molecular recognition by antibodies. Such rules could be the seed of predictive methods to relate the aminoacid sequence of a given antibody with its capability to recognize an antigen.

To explore whether or not rules in the molecular recognition by antibodies do exit, here we analyzed most of the antibody sequences and structures currently known. To this end, we constructed a World Wide Web interface denominated VIR.II. VIR.II allows a fast and simple access to the sequences of antibodies compiled in the database of sequences of immunological interest, also called The Kabat database. In addition, VIR.II allows the classification of the of the antibody specificity in gross specificities such as anti-proteins, anti-peptides, anti-haptenos, etc. This, together with another interface, which serves to compile and to analyze the three-dimensional structures of antibodies available in Protein Databank (PDB), made possible to correlate patterns in the sequence, the structure and the function (gross specificity) of the antibodies.

We verified existence of rules in the molecular recognition by antibodies. It is described how the antibodies recognize proteins, peptides and haptens. In the case of the anti-protein antibodies, the surface is flat, which provides a proper complementary to the large antigens. Anti-peptide antibodies, on the other hand, have a groove at the antigen-binding site, which assures the complementarity with peptides. An interesting observation that emerged from this work, it is the way in which the antibodies recognize haptenos. Due to the small size of the haptens, it is needed gross changes in the topography of antigen-binding site. Just small changes in the topography of the antigen-binding site are enough to bind the haptens. This changes could be achieved through conformational

changes or substitutions of the aminoacid side-chains during the process of the immune response maturation.

Finally, we analyzed the role of the somatic hypermutation in replacing the residues in contact with the antigen. We found that replacement of residues in contact with the antigen during the process of somatic hypermutation is rare event. A possible explanation for this finding is that mutation of residues in contact with the antigen is harmful. Therefore, the variant bearing those mutations will be eliminated from the "pool" of mutated genes during the process of maturation of the immune response. This explanation has implications for the theories that explain the origin and the evolution of the repertoire of antibodies. Also, it has practical consequences. That is, if the affinity of an antibody need to be improved *in vitro*, what we should do is identify the residues in contact in the antigen but, instead of select these for mutagenesis, the target of mutagenesis should be the nearby residues.

Indice

Introducción y Plan de la tesis

El objetivo central de nuestro trabajo ha sido estudiar la relación entre la secuencia, la estructura, la evolución y las propiedades de reconocimiento de los anticuerpos. Los anticuerpos tienen la capacidad de reconocer virtualmente a cualquier otra molécula con alta especificidad y afinidad. Así, esta familia de proteínas ha sido de gran utilidad en el diagnóstico clínico, la terapéutica y como herramienta de análisis en el laboratorio. Desde el punto de vista básico, son el paradigma de reconocimiento molecular por excelencia (Winter & Milstein, 1991).

Por esas razones, los anticuerpos han sido las moléculas más estudiadas. Lo anterior se ilustra por el número de secuencias y estructuras tridimensionales que actualmente se depositan en los bancos de datos (Figura 1 y 2). A la fecha, se han depositado 19,832 secuencias de anticuerpos en la base de datos de Kabat (Figura 1). Además, se han depositado 355 estructuras cristalográficas en el Protein Data Bank (Figura 2). Como se observa en las figuras, tanto el número de secuencias, como el de estructuras, se han incrementado exponencialmente de manera sostenida durante los últimos años.

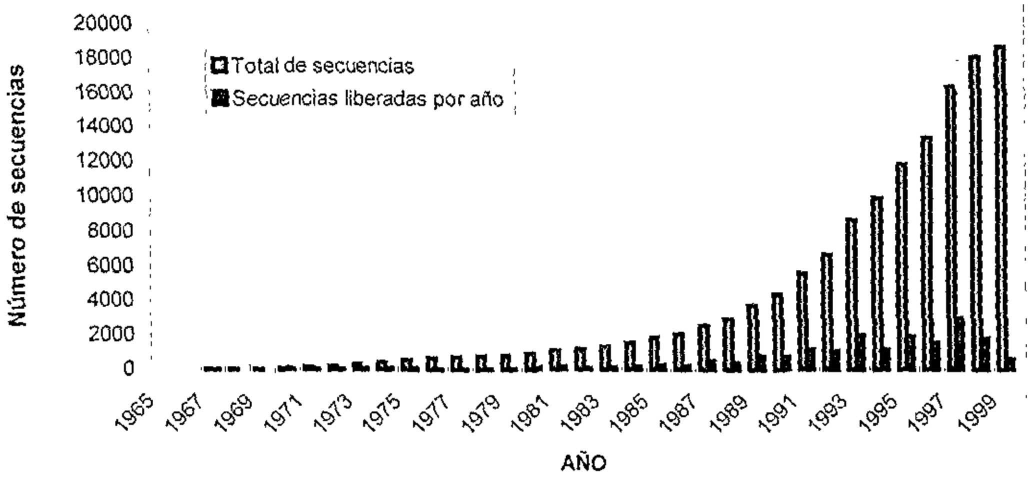


Figura 1. Crecimiento de la base de datos de Kabat de las secuencias de anticuerpos con especificidad reportada (Johnson & Wu, 2000)

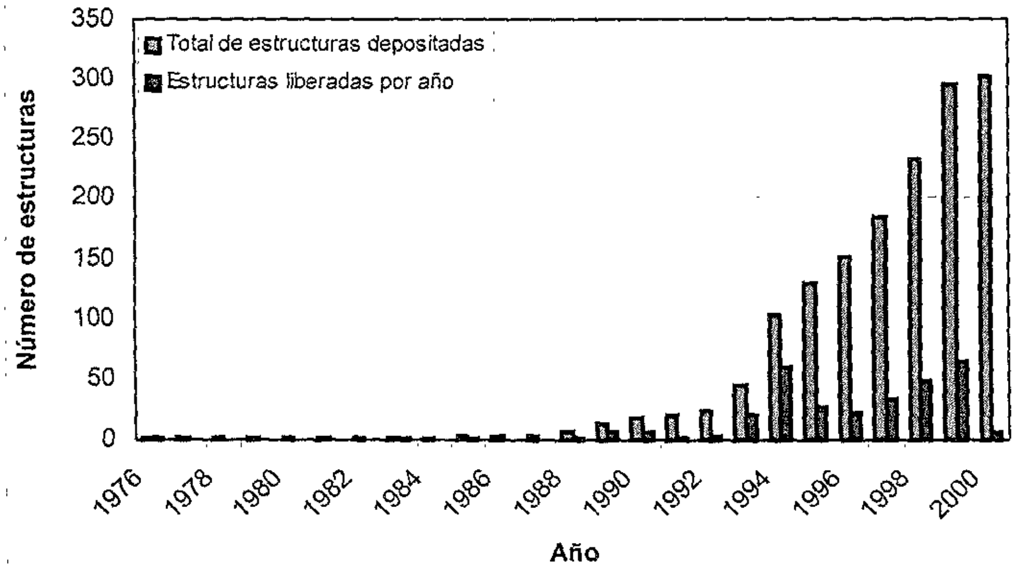


Figura 2. *Crecimiento exponencial de la base de datos de estructuras de anticuerpos depositadas en el Banco de Estructura tridimensional, PDB (Berman et al., 2000).*

A pesar de ello, actualmente no es posible predecir la especificidad o afinidad de un anticuerpo en particular, partiendo de su secuencia de aminoácidos, o incluso de su estructura. La falta de un modelo predictivo que relacione la estructura con las propiedades de reconocimiento de los anticuerpos, ha limitado la comprensión del reconocimiento molecular mediado por estos receptores del sistema inmune y ha impedido abordar el diseño racional de anticuerpos de especificidad predeterminada (Almagro *et al.*, 1995).

Los resultados que se exponen en esta tesis (Ramirez-Benitez *et al.*, 2001a, Ramirez-Benitez *et al.*, 2001b, Ramirez-Benitez *et al.*, 2001c), son parte de un conjunto de trabajos desarrollados dentro del proyecto de doctorado (Vargas-Madrazo *et al.*,1997, Almagro *et al.*, 1997, Almagro *et al.*, 1998, Ramirez-Benitez *et al.*, 2001a, Ramirez-Benitez *et al.*, 2001b, Ramirez-Benitez *et al.*, 2001c), que han tenido como objetivo común analizar la relación entre la estructura y las propiedades de reconocimiento de los anticuerpos. La intención en todos estos trabajos ha sido buscar reglas que permitan predecir las propiedades de reconocimiento de una secuencia o estructura de un anticuerpo dado.

Hipótesis de trabajo

Resultados previos obtenidos por varios grupos de investigación, incluyendo el nuestro (Capítulo I), sugieren que existen reglas del reconocimiento molecular mediado por anticuerpos. Dado el incremento reciente de la información, este trabajo busca aportar nueva información para probar o descartar esta hipótesis.

La tesis se organiza en los siguientes capítulos:

Capítulo I: *Estructura y Función de los anticuerpos.* En este capítulo se revisa el estado actual de conocimiento sobre la estructura, la genética y los modelos de reconocimiento mediado por anticuerpos. El objetivo de este capítulo es brindarle al lector los conceptos y elementos necesarios para comprender los capítulos que siguen, así como presentar los antecedentes de la tesis.

Capítulo II: *VIR.II: A new interface with the antibody sequences in the Kabat database.* Este es el primer artículo que se presenta de la tesis. Se describe una interfase, denominada VIRII, con la base de datos de Kabat (Jhonson & Wu, 2000). VIR.II tiene su antecedente en VIR (Almagro *et al.*, 1995) y fue diseñada para la búsqueda de secuencias de anticuerpos por tipo de cadena, especie y especificidad. Además, esta herramienta introduce una clasificación de las especificidades en términos de la naturaleza química y bioquímica de los antígenos. Esta clasificación ha permitido correlacionar patrones en el sitio de unión del antígeno y el tipo de antígeno reconocido. Art¡culo en presa, enviado a Biosystems.

Capítulo III: *Structure-function relationships in anti-protein, anti-peptide and anti-hapten antibodies*. Este capítulo incluye un segundo artículo. Se presentan los resultados del análisis de los determinantes estructurales de los anticuerpos anti-proteína, anti-péptido y anti-hapteno. Este estudio confirmó que la topografía del sitio de unión al antígeno determina el tipo de antígeno reconocido por el anticuerpo. Artículo publicado en PROTEINS: Structure, Function, and Genetics. 45:199-206 (2001)

Capítulo IV: *Analysis of antibodies of known structure suggest a lack of correspondence between the residues in contact with the antigen and those modified by somatic hypermutation.* Este tercer artículo es un análisis de la relación entre la hipermutación somática y los residuos en contacto con el antígeno. Se asignan los genes de línea germinal a la estructura tridimensional de los anticuerpos para obtener las posiciones mutadas, las cuales se comparan con los residuos que hacen contacto con el antígeno. Los resultados muestran que la hipermutación somática no correlaciona con los residuos en contacto. Esto implica que las mutaciones observadas pudieran ser reminiscencias del proceso de selección negativa y sugiere un esquema novedoso para madurar la afinidad de los anticuerpos. Artículo por enviarse.

Conclusiones. Se comentan los principales resultados de los capítulos anteriores, en el contexto de la tesis.

Anexos. Finalmente, en esta última sesión, se incluyen tres publicaciones que, junto con los resultados de los capítulos II, III y IV, resumen el trabajo realizado durante el doctorado.

Evolution of the structural repertoire of the human V(H) and Vkappa germline genes. Basados en la organización de la información de línea germinal en familias multigénicas, la organización de estas en clanes y en nuestras observaciones, donde mostramos que los patrones conservados en las secuencias de los anticuerpos brindan información de la existencia de un fuerza selectiva no solo en términos de restricciones funcionales, sino también en términos de las propiedades del reconocimiento, en este artículo, analizamos la evolución del repertorio estructural V_H y V_k en el contexto de las clases de estructuras canónicas, a fin de conocer cual es el papel de cada una de los componentes en la generación de la diversidad de reconocimiento mediado por anticuerpos. El análisis de la relación evolutiva entre los genes, las familias y los clanes indicaron que las clases no están distribuidas al azar. Nuestro análisis de clanes y clases de estructuras canónicas entre las diversas especies, nos permitió sugerir un repertorio estructural primordial. Así los clanes I, II y III están representados en las clases 1-2, 1-3 y 3-1, donde el principal determinante es H2. En V_k, los clanes I y II están representados por las clases 2-1 y 4-1, aquí el determinante es L1. Las clases V_H-V_k de estructuras canónicas resultantes solo se reducen a 1-2/3-2-1, 1-2/3-4-1, 3-1-2-1 y 3-1-4-1, las cuales representan en repertorio de topologías (planas, con cavidad y con ranuras) necesarias para el reconocimiento especifico de proteínas, péptidos y haptenos.

The differences between the structural repertoires of V_H germ-line gene segments of mice and humans: implication for the molecular mechanism of the immune response. Este trabajo, representa la primera compilación y caracterización del repertorio estructural en los segmentos de genes de línea germinal V_H de ratón. Nuestros resultados mostraron que el repertorio estructural está representado por la clase 1-3, ampliamente codificada por la familia V_H1. La comparación del repertorio de ratón y humano (clase 1-2 codificada por la familia V_H3), mostró que aun cuando la canónica 3 y 2 es igual en longitud, se presentan cambios conformacionales. Sin embargo, estos cambios no afectan la forma del sitio de unión al antígeno. Por otro lado, sugieren que estas diferencias podrían estar relacionadas con procesos de regulación específicos de la especie.

Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. Continuando con la caracterización del repertorio estructural en ratón, en este artículo, presentamos la compilación y el análisis del repertorio estructural de los genes de línea germinal de V_k, la comparación entre repertorios de ratón y humano y las implicaciones evolutivas de las diferencias observadas entre estas dos especies. Nuestros resultados mostraron que el repertorio estructural de V_k es más diverso en ratón, puesto que este codifica para 7 clases de estructuras canónicas, mientras que en humanos solo se restringe a 4 clases de estructuras canónicas. Una posible explicación a esta diferencia, es que el repertorio de genes de línea germinal V_k en el humano, está más distribuida, es decir 60% de sus genes son V_k y el 40% son V_λ, mientras que en ratón el

95% de sus genes son V_k. Las diferencias observadas pudieran ser debidas a una compensación en el contenido de genes que codifican para la cadena variable ligera lambda en el ratón.

Capítulo I

Estructura y función de los anticuerpos

Estructura Primaria de las Anticuerpos

Los anticuerpos son la fuente de especificidad de la respuesta inmune humoral. El rasgo distintivo de esta familia de proteínas es su capacidad para reconocer virtualmente un número ilimitado de antígenos (Tonegawa 1983); capacidad que se debe a la posibilidad para acomodar una gran diversidad tanto de formas como de aminoácidos en el sitio de unión con el antígeno.

La secuencia de aminoácidos de los anticuerpos se estableció hacia finales de los años 50's y principios de los 60's (Edelman & Gally 1962). Esta se forma por cuatro cadenas polipeptídicas: dos cadenas ligeras (L) idénticas y dos cadenas pesadas (H) idénticas. La cadena L tiene un peso de 25 000 Daltons y su nombre deriva de ser la mitad de la cadena pesada H, la cual posee un peso de 50 000 Daltons.

Hasta el presente se han descrito dos tipos de cadenas ligeras L: el tipo kappa (k) y el tipo lambda (λ). Estos tipos de cadena ligera, difieren en algunos detalles estructurales, estas diferencias no parecen tener un significado funcional. Para la cadena H se han descrito cinco clases o isotipos: α, β, δ, γ, y μ. A diferencia de la cadena L, los isotipos de la cadena H tienen propiedades funcionales diferentes. Los isotipos de cadena H pueden combinarse indistintamente con cualquiera de los tipos de cadena L para formar una molécula funcional. El isotipo más abundante en el suero es el γ, que da origen a la IgG de tipo κ o λ.

En una IgG las cuatro cadenas polipeptídicas se unen entre sí por interacciones covalentes (puentes disulfuro intercatenarios) y se estabilizan por interacciones no covalentes (interacciones de Van der Waals y de Coulomb) (Nissonoff *et al.*, 1959, Edelman & Gally 1962). La digestión de una IgG con papaína (Porter 1959), genera tres fragmentos, dos de ellos idénticos que se denominan fragmentos de unión al antígeno (Fab) y un tercer fragmento fácilmente cristalizable, denominado por tanto fragmento cristalizable (Fc).

Si se compara la secuencia primaria de varios anticuerpos, se observa que tanto las cadenas L como H, poseen porciones de aproximadamente 100 aminoácidos que son homologas entre sí, denominadas "dominios de homología" (Edelman *et al*, 1969). Esta observación sugiere que la secuencia primaria de los anticuerpos se originó por eventos de duplicación a partir de un segmento ancestral de aproximadamente 100 aminoácidos (Pascual & Capra, 1991). Los primeros dominios de homología de ambas cadenas, L y H, muestran gran variabilidad de aminoácidos (Wu & Kabat, 1970), por lo que reciben el nombre de dominios variables (V). El resto de los dominios son relativamente conservados y se les denomina dominios constantes (C).

Partiendo de la organización en dominios de una IgG (Figura 1), la cadena L está compuesta por dos dominios, uno Variable (V_L) en la porción amino terminal y uno constante (C_L) en la porción carboxilo terminal. La cadena H, está compuesta por cuatro dominios, uno V (V_H) en la porción amino terminal y tres constantes (C) hacia la porción carboxilo (CH_1, CH_2 y CH_3). De esta manera en el Fab se encuentran dos dominios V (V_L y V_H) y dos dominios C (C_L y CH_1), por lo que este fragmento puede describirse como un hetero-polimero conformado de V y C. El Fc, sin embargo, es un homo-polímero, puesto que solo lo constituyen dominios C (CH_2 y CH_3). La asociación de V_L y V_H hacia la porción amino terminal de una IgG, forma el fragmento de dominios variables o Fv. En esta porción del Fab reside la capacidad de los anticuerpos de interactuar de manera especifica con los antígenos (Kabat & Wu 1971).

Estructura tridimensional

La primera observación de la estructura que adquiere una IgG en el espacio se realizó en los años 60's por microscopía electrónica (Valentine & Green 1967). Las imágenes de baja resolución de estas moléculas revelaron que tienen una estructura en forma de Y como se representa en la Figura 1. Análisis más detallados de la estructura de una IgG por técnicas de difracción de rayos X a principios de los 70's, indicaron que los dominios de homología forman porciones compactas más o menos elipsoides (Figura 2), que siguen un patrón común de plegamiento (Schiffer *et al.*, 1973, Poljak *et al.*, 1973).

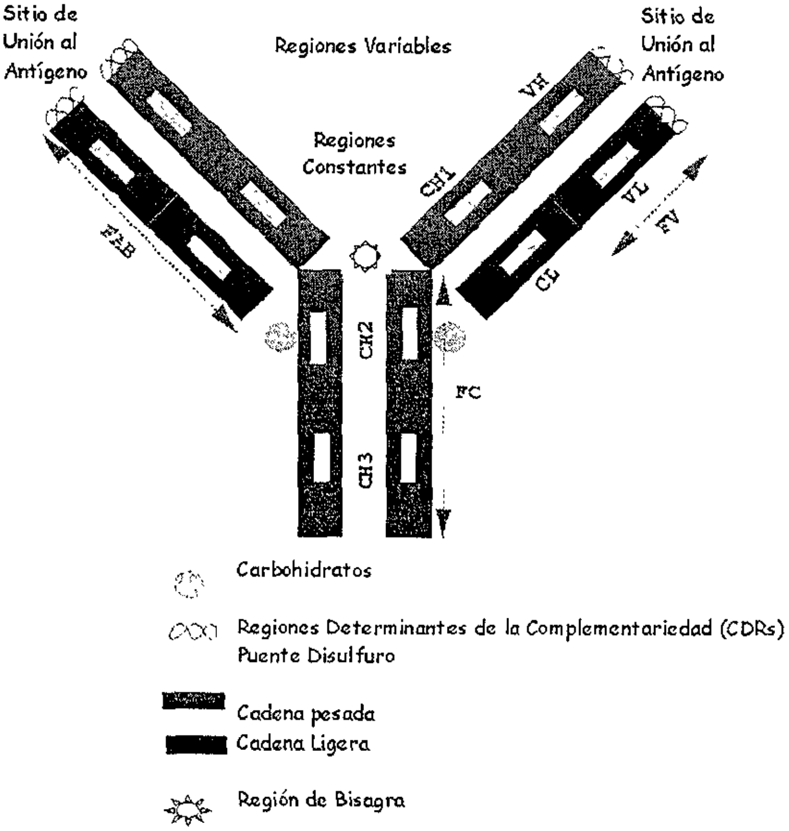


Figura 1. Representación esquemática de una IgG de humano. V y C indican las regiones variables y constantes de las cadenas pesadas (H) y Ligeras (L). Fragmento Fv, en el cual descansa el sitio de unión al antígeno. Así como los fragmentos generados por digestión proteolítica, Fab y Fc.

El plegamiento típico de los dominios V y C, está formado por dos láminas β antiparalelas estabilizadas por un puente disulfuro intracatenario e interacciones no covalentes entre los dominios. A su vez, cada lámina β está formada por varias hebras β, estabilizadas por puentes de hidrógeno entre los átomos de la cadena principal de las hebras β. Las láminas β difieren en el número de hebras β que las componen. La primera lámina, iniciando del termino amino, está formada por cuatro hebras β. La segunda lámina está formada por tres hebras β (Amzel & Poljak 1979). Las láminas β se denotan de la "A" a la "F" partiendo del termino amino del domino V.

La conexión entre las hebras β se realiza a través de asas, algunas de ellas conectando dos hebras β continuas dentro de la misma lámina y otras conectando hebras β de láminas diferentes. La topología resultante de estas conexiones se asemeja a un motivo griego. Este dominio, aunque es común a los dominios V y C, tienen una diferencia fundamental entre ellos. La diferencia consiste de una inserción de dos hebras β en el dominio V (C' y C'') con respecto al dominio C (Figura 2). Tal inserción tiene gran importancia puesto que allí se localiza una de las regiones que unen al antígeno (Davies *et al.*, 1975).

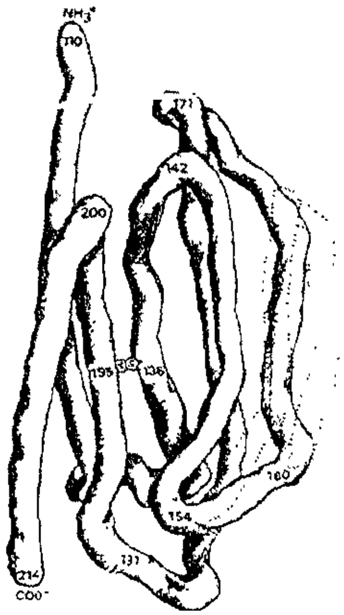


Figura 2. Representación esquemática del plegamiento básico de los anticuerpos. Las líneas marcadas representan el plegamiento de los dominios constantes CL y CH1. Las líneas punteadas indican le región extra de los dominios VL y VH. NH3 y COOH son los extremos amino y carboxilo. Tomado de Poljak et al., Proc. Nat. Acad. Sci. USA 703305 (1973)

Los dominios C se asocian formando aproximadamente un ángulo recto entre sí (ver por ejemplo el Fc en la Figura 3). Esta asociación se establece principalmente por interacciones de Van der Waals a través de la primera lámina de cada dominio. La asociación V-V es, sin embargo, muy diferente, puesto que la primera lámina β de cada dominio V queda hacia el exterior de la estructura. Esto hace que los dominios V queden aproximadamente paralelos uno con respecto al otro, garantizando así que las regiones que forman el sitio de unión con el antígeno converjan en una superficie continua en la punta del Fv (Figura 3).

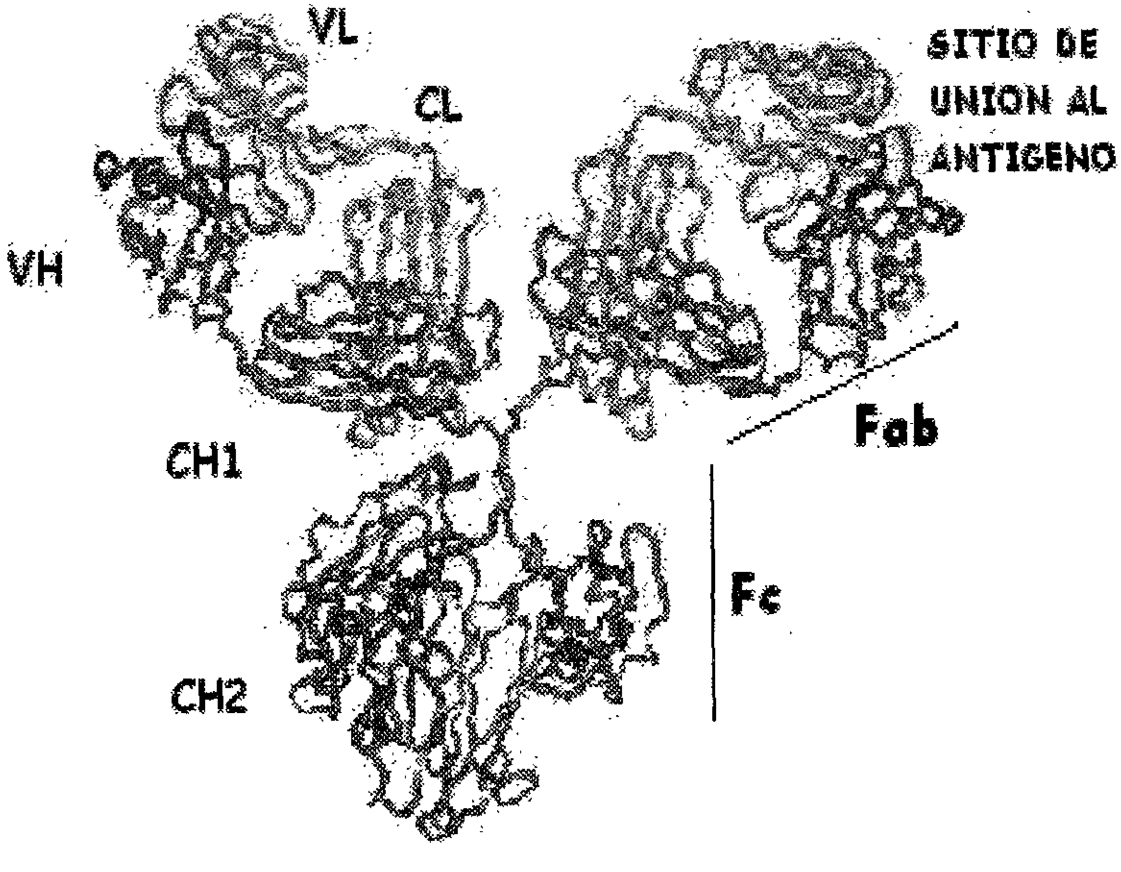


Figura 3. Representación en forma de listones de la estructura tridimensional de una IgG. En color morado los dominios de la cadena ligera (VL y CL). En azul los dominios de la cadena pesada (VH, CH1 y CH2).

Mecanismos Genéticos de Generación de la Diversidad

Un dominio V se genera a través de varios segmentos de genes. El dominio V_L se ensambla a partir de dos segmentos de genes: un segmento Variable ligero (*Igl-V*) y un segmento de unión (*Igl-J*; del inglés *Joining*), llamado así porque une al segmento V con el gen que codifica para el dominio C. Por otra parte, un dominio V_H se codifica a partir de tres segmentos de genes: un gen Variable pesado (*Igh-V*), un segmento de unión (*Igh-J*) y un tercer segmento de gen llamado de diversidad (*Ig-D*). Este último segmento de gen debe su nombre a que aporta diversidad adicional a los anticuerpos (Figura 4).

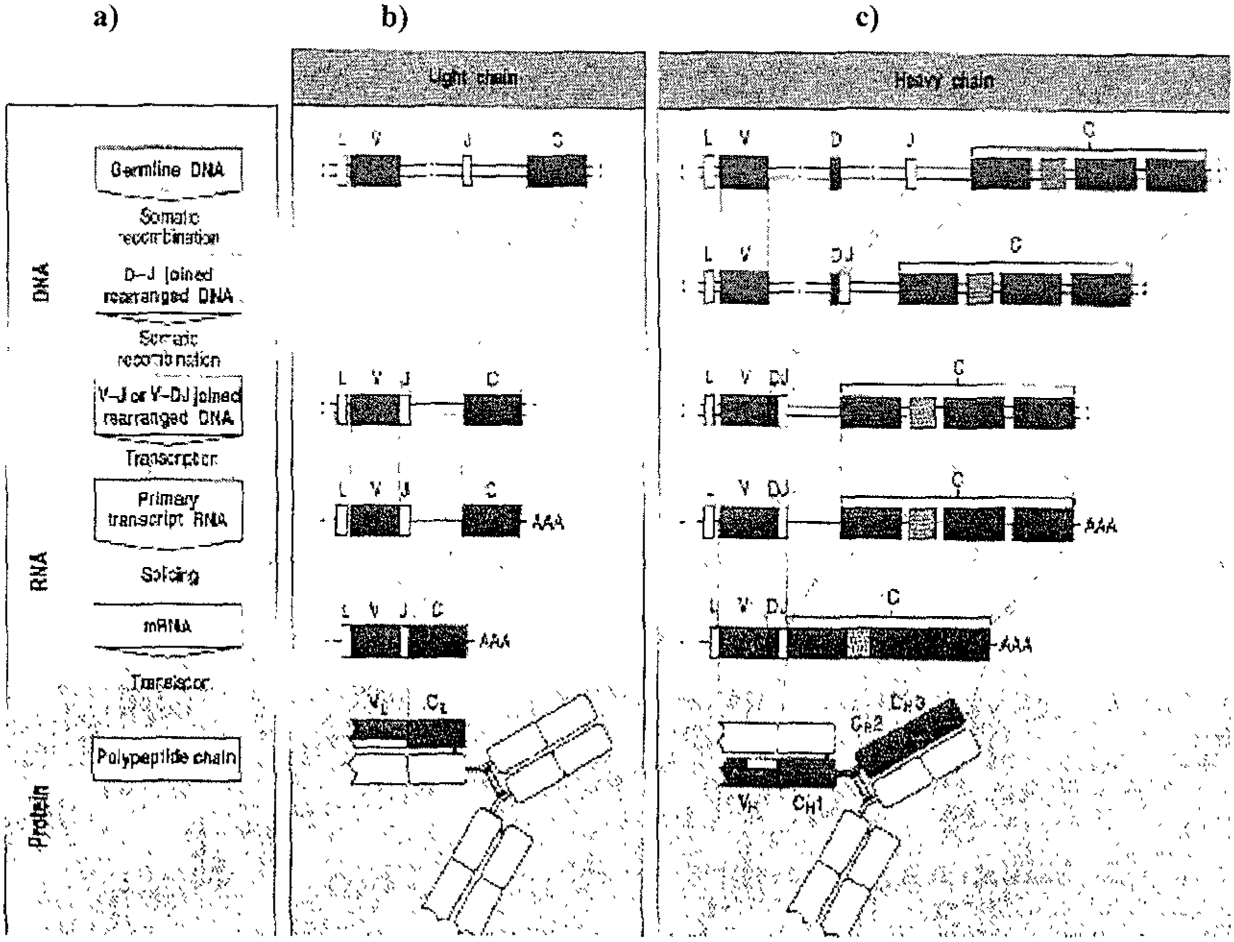


Figura 4. Representación del rearreglo de segmentos de genes de línea germinal . a) procesos que se llevan a cabo en el rearreglo desde la línea germinal hasta la formación de las cadenas polipeptídicas de un anticuerpo. b) rearreglo de segmentos de genes para generación de una cadena ligera. c)rearreglo de segmentos de genes para la generación de una cadena pesada. Tomado de CA. Janeway, P. Travers, M Walport and DJ Capra., Immunobiology. The immune system in health and Disease Fourth Edition 1999

Las especies mejor estudiadas hasta la fecha, en cuanto a los mecanismos que generan la diversidad de los anticuerpos, son el ratón y el humano. El ratón por ser el modelo experimental por excelencia en inmunología y el humano por razones obvias. En los últimos años se han completado los mapas físicos para cada uno de los loci que codifican para los anticuerpos humanos (Tomlinson 1995, Corbett *et al.*, 1997; Zachau 1993, Tomlinson 1994, Williams 1996).

El locus que codifica para V_H tiene una longitud aproximadamente de 1100 kilobases (Kb) y se encuentra en la extremidad telomérica del brazo largo del cromosoma 14 (Croce *et al.*, 1979). El número segmentos de genes V_H es 95 (Shin *et al.*, 1991, Matsuda *et al.*, 1993, Cook *et al.*, 1994, Cook *et al.*, 1995, Matsuda *et al.*, 1998), 27 segmentos D (Siebenlist *et al.*, 1981, Buluwela *et al.*, 1988, Ichihara *et al.*, 1988, Corbett., *et al.*, 1997), 6 segmentos J (Ravetch *et al.*, 1981) y 6 segmentos C (Rabbitts *et al.*, 1981). En la Figura 5 se muestra el mapa del locus V_H y la ubicación de cada uno de estos segmentos de genes. Como se observa, el locus está dividido en el conjunto de los segmentos V, seguido del conjunto de los segmentos D y finalmente hacia la región centromérica del cromosoma se encuentran los segmentos J y los genes que codifican para los dominios C.

La información completa del gran locus V_k se localiza en el brazo corto del cromosoma 2 (Malcolm *et* al., 1982, Mc Bride *et al.*, 1982), su longitud es de 1820 Kb. El mapa completo de este locus se presenta en la Figura 8, en este se localizan 82 segmentos de genes variables (Huber *et al.*, 1993, Schäble & Zachau 1993, Cox *et al.*, 1994, Schable *et al* 1994), de los cuales 76 corresponden a la región variable (32 funcionales con marco de lectura, 25 pseudogenes, 16 segmentos con menores defectos y 3 que son tanto defectuosos como funcionales), 5 segmentos J_k (Hieter *et al.*, 1982) y 1 segmento C (Hieter *et al.*, 1980).

La información completa del gran locus V_λ se localiza en el brazo largo del cromosoma 22 (Erikson *et al.*, 1981, Emanuel *et al.*, 1985, Dunham *et al.*, 1999), con una extensión de 1050 Kb. El mapa completo de este locus se presenta en la Figura 9, en este se localizan 82 segmentos de genes variables, de los cuales 69-70 corresponden a la región variable y se distribuyen en 30 funcionales con marco de lectura, 31 pseudogenes, 1 segmento con menores defectos y 3 que son tanto defectuosos como funcionales (Frippiat *et al.*, 1995, Kawasaki *et al.*, 1995, Williams *et al.*, 1996, Kawasaki *et al.*, 1997), 7 segmentos J_{λ} (Hieter *et al.*, 1981, Taub *et al.*, 1983, Dariavach *et al.*, 1987, Vasicek & Leder 1990) y 7 segmentos C (Taub *et al.*, 1983, Ghanem *et al.*, 1988, Kay *et al.*, 1992, Lefranc *et al.*, 1999).

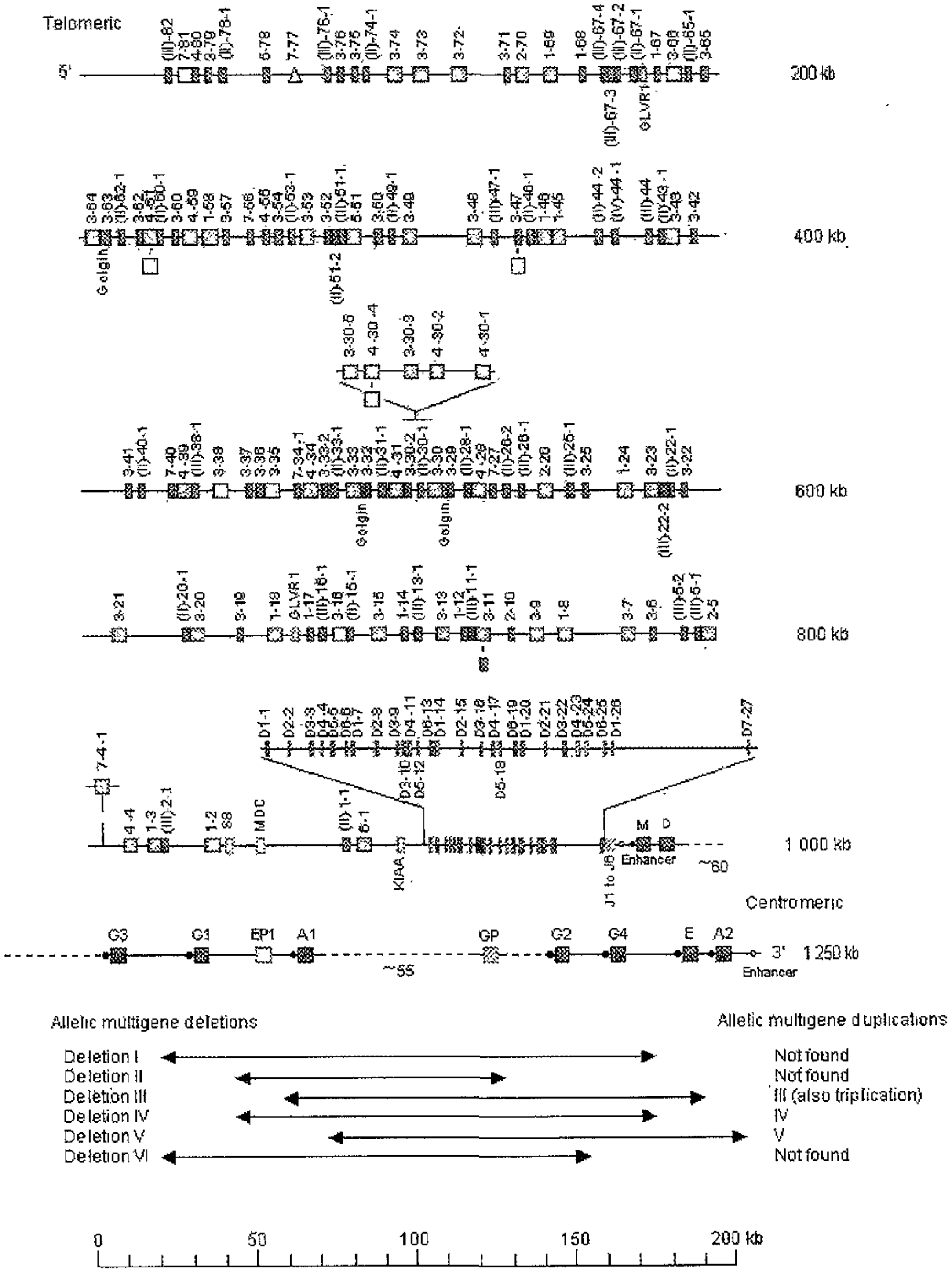


Figura 5. Locus V_H de humano. Localización de los segmentos de genes que codifican la cadena pesada de los anticuerpos de humano. Segmentos de genes Variables funcionales □ , *con marco de lectura* □ , *pseudogenes* ▨ , *segmentos D funcionales* ■ , *segmentos Joining* *y segmentos constantes* ▨ . *Tomado de IMGT, the international ImMunoGeneTics database http://imgt.cines.fr:8104 (Initiator and coordinator: Marie-Paule Lefranc, Montpellier, France).*

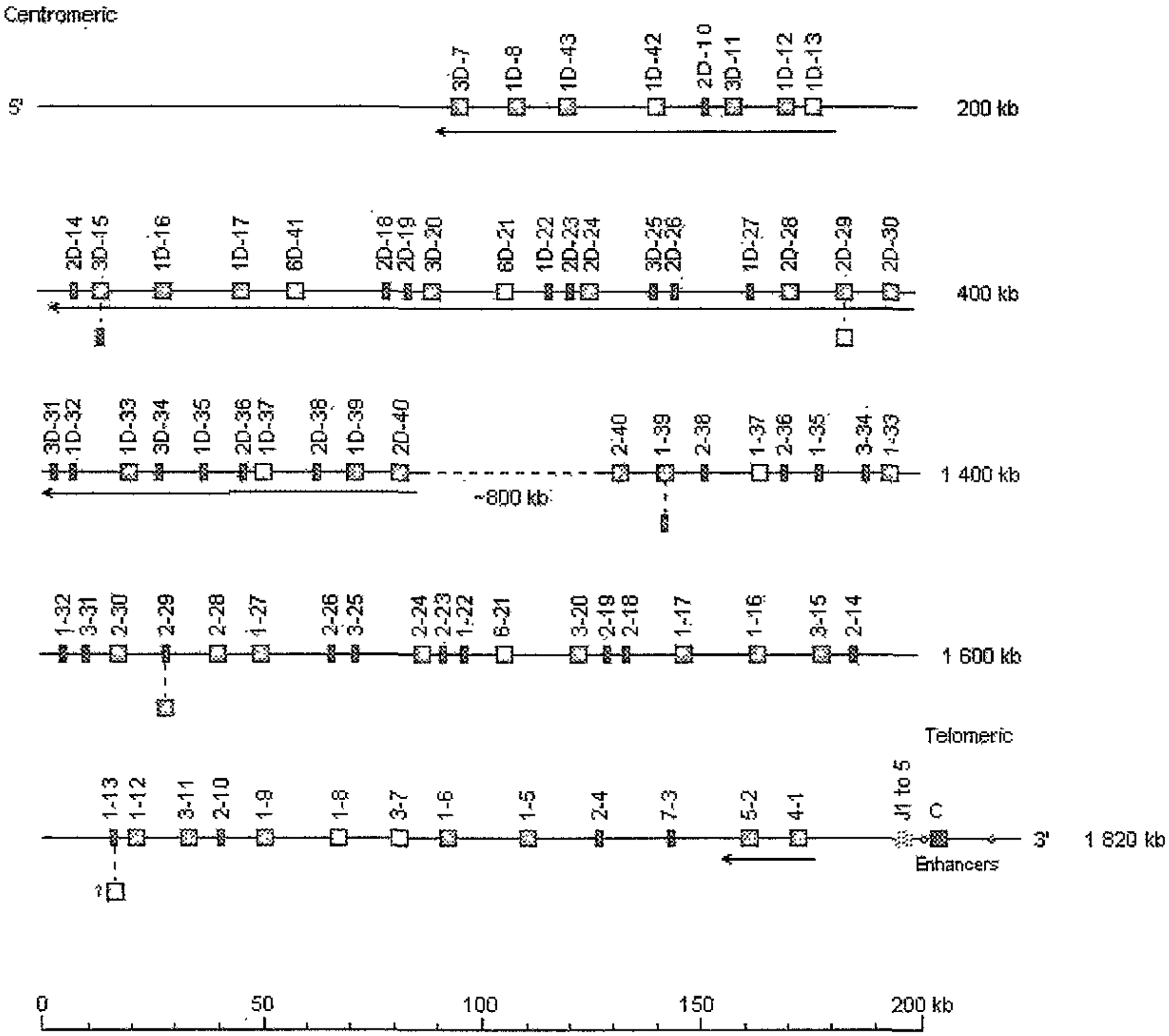


Figura 5. Locus V_k de humano. Localización de los segmentos de genes que codifican la cadena pesada de los anticuerpos de humano. Segmentos de genes Variables funcionales □ , con marco de lectura □ , pseudogenes ▨ , segmentos Joining ░ y segmentos constantes ▒ . IMGT, the international ImMunoGeneTics database http://imgt.cines.fr:8104 (Initiator and coordinator:Marie-Paule Lefranc, Montpellier, France).

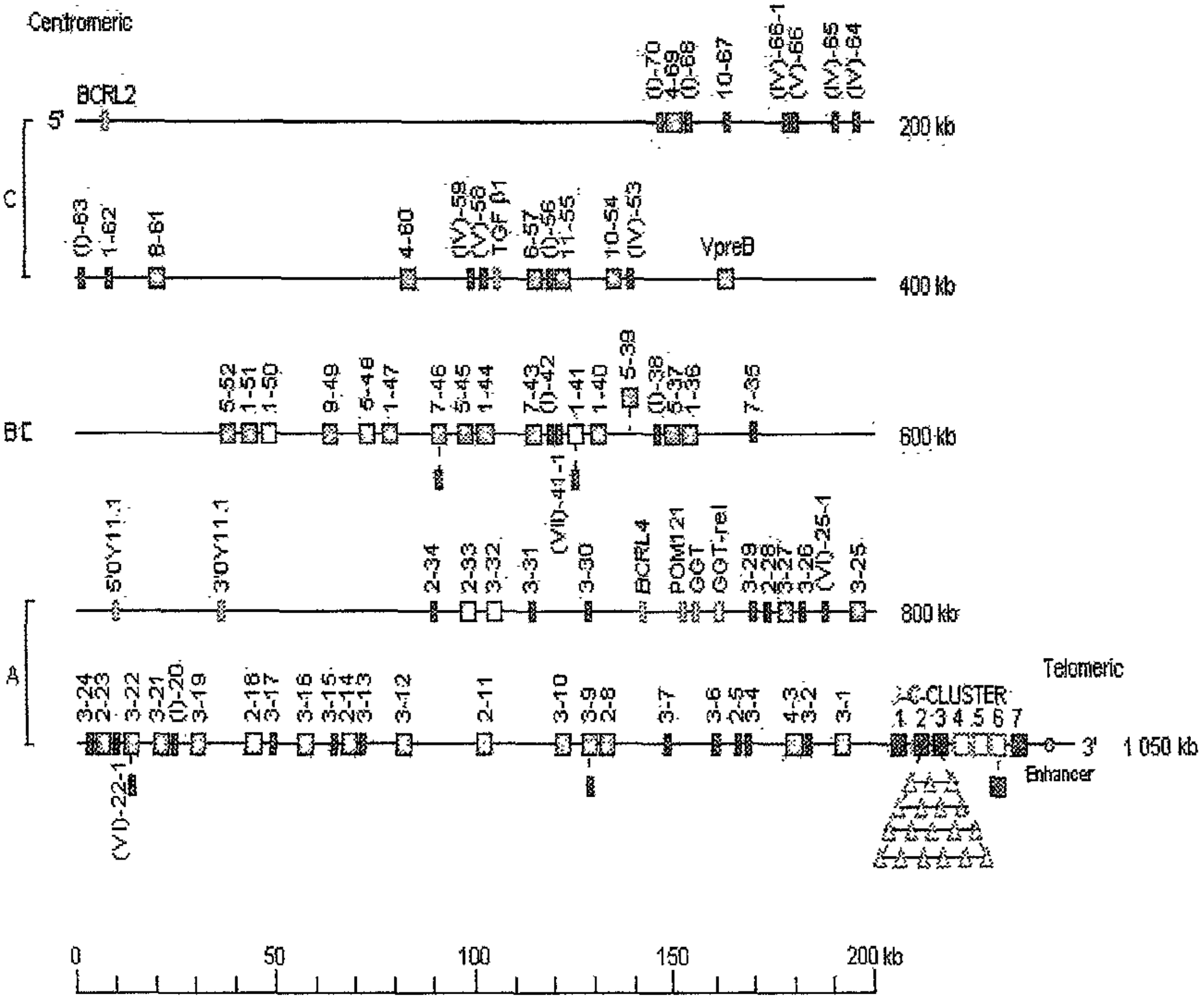


Figura 5. Locus V_{λ} de humano. Localización de los segmentos de genes que codifican la cadena pesada de los anticuerpos de humano. Segmentos de genes Variables funcionales ▧ , con marco de lectura □ , pseudogenes ■ , segmentos Joining ▨ y segmentos constantes ▩ IMGT, the international ImMunoGeneTics database http://imgt.cines.fr:8104 (Initiator and coordinator:Marie-Paule Lefranc, Montpellier, France

El sitio de Interacción con el Antígeno

En ausencia de estructuras tridimensionales de anticuerpos, Wu y Kabat (1970, Kabat & Wu 1971), compararon las secuencias primarias de los dominios V conocidos. Observaron que la variabilidad de los aminoácidos, típica de los dominios V, no se distribuye de manera homogénea a lo largo del dominio V. Así, se identificaron regiones que concentran la variabilidad de los aminoácidos, alternando con regiones relativamente conservadas. Debido a que la función esencial de los anticuerpos es interactuar con un diverso número de antígenos, se supuso que las regiones de máxima variabilidad o hipervariables determinaban la complementariedad con el antígeno. En consecuencia, se les llamó regiones determinantes de la complementariedad (CDRs, del ingles Complementarity Determining Regions). Por contraposición con los CDRs, se denominó al resto del dominio V, es decir, a aquellas regiones relativamente conservadas, armazón (FR, del ingles framework).

Al obtenerse la primera estructura tridimensional de un Fab (Schiffer *et al.*, 1973, Poljak *et al.*, 1973), se observó que la definición de CDR coincidía aproximadamente con las asas más externas del Fv; aquellas que forman una superficie continua en la punta del Fv. Al año siguiente, con la estructura del primer complejo antígeno-anticuerpo (Segal *et al.*, 1974), se mostró que, en efecto, la región definida por los CDRs contiene el sitio de interacción con el antígeno.

Dentro de V_L se identifican tres CDRs. El primero o CDR-1, se define como la región comprendida entre las posiciones 24 y 34, el segundo o CDR-2 se localiza de la posición 50 a la posición 56 y el tercero o CDR-3, se define desde la posición 89 hasta la posición 97. De manera similar, se identifican tres CDRs en V_H: el CDR-1 de la posición 31 a la posición 35, el CDR-2 de la posición 50 a la posición 65 y el CDR-3 de la posición 95 hasta la posición 102. Las FR, quedan definidas automáticamente, como aquellas regiones entre los CDRs. Así, tanto para V_L como para V_H, se definen cuatro regiones FR, el FR-1 comprendido entre el amino-terminal y el CDR-1, el FR-2, entre el CDR-1 y el CDR-2, el FR-3, entre el CDR-2 y el CDR-3, finalmente el FR-4, entre el CDR-3 y el carboxilo-terminal.

Aunque la definición de CDR coincide aproximadamente con las asas más externas del Fv, existen, sin embargo, algunas diferencias entre la definición de las asas y la región que ocupan los CDRs. Esto se debe a que la definición de CDR es una definición funcional (variación de los aminoácidos debido a la selección por el antígeno), mientras que la definición de asa responde a una definición meramente estructural. De hecho, en algunas aplicaciones, como la selección de posiciones para mutagénesis con objeto de optimizar la afinidad de los anticuerpos, se utiliza la definición de CDR (definición de Kabat). En otras aplicaciones, como la modelación por homología u otros tipos de análisis estructurales, se utiliza la definición estructural o de asa hipervariable (definición de Chothia; por su autor).

Las asas hipervariables se definieron por Chothia y Lesk en 1987 comparando los Fv de varias estructuras (Chothia & Lesk, 1987). Al superponer las estructuras, se observó que hay regiones dentro de los dominios V que difieren muy poco entre las estructuras analizadas. Estas regiones coinciden con la estructura secundaria (láminas β) y algunas de las asas que no forman parte del sitio de unión al antígeno. Por otra parte, se encontró que hay regiones con grandes diferencias estructurales, están localizadas precisamente en las asas que unen al antígeno.

Modelos de Reconocimiento Molecular

La definición de CDR como las regiones de máxima variabilidad o hipervariables resolvió, en parte, la pregunta de por donde los dominios V_L y V_H se unen con el antígeno. Sin embargo, dió origen a una pregunta que prevaleció durante los años 70s y gran parte de los 80s. Esta pregunta se refiere a si las asas que alojan al sitio de unión con el antígeno tienen una conformación particular en cada anticuerpo o su conformación se repite de un anticuerpo a otro.

Como se mencionó antes, al comparar los Fvs de estructura conocida, se observó que las regiones que constituyen el sitio de unión con el antígeno tienen grandes diferencias estructurales. No obstante, una comparación minuciosa entre las diferentes estructuras reveló que, aunque estas regiones varían significativamente, tanto en longitud como en aminoácidos, existen ciertas conformaciones que se repiten de un anticuerpo a otro. Estas conformaciones se llamaron conformaciones canónicas para hacer notar lo invariante de tales conformaciones (Chothia & Lesk, 1987). Una conformación canónica depende de la longitud del asa y de ciertos residuos clave. Estos residuos establecen interacciones de Van der Waals o puentes de hidrogeno con la propia asa y/o con las regiones estructuralmente conservadas del dominio V.

Estructuras canónicas en la cadena ligera.

Conformaciones de L1 kappa.

Dentro de V_L, se identifican tres asas hipervariables (ver Figura 11). La primera o L1 (nótese que se utiliza un término diferente al de CDR), se identifica como el asa que conecta las hebras β "B" y "C", comprendiendo desde la posición 26 hasta la 32 (Chothia & Lesk, 1987). El principal factor determinante de la conformación es un residuo hidrofóbico en la posición 29, el cual se empaca contra una cavidad hidrofóbica formada por los residuos 2, 32, 33 y 71 (Chothia & Lesk, 1987, Chothia *et al.*, 1989). Se han observado inserciones de residuos entre las posiciones 30 y 31, las cuales forman un "abultamiento". Estas inserciones no afectan significativamente la conformación del asa (Chothia *et al.*, 1989).

Se han identificado seis tipos se estructuras canónicas para esta región (Chothia *et al.*, 1989, Rini *et al.*, 1973, Haynes *et al.*, 1994, Tomlinson *et al.*, 1995). La principal diferencia entre estos seis tipos es el número de aminoácidos entre las posiciones 30 y 32. En la Tabla 1, se presenta un resumen de los tipos de estructuras canónicas observadas en L1 de V_k. Las Figuras 5, 6, 7 y 8 muestran los primeros cuatro tipos de conformaciones adoptadas por esta región.

	Tipos Observados		No de residuos entre la posición 30 y 32	Característica
Estructuras Canónicas para L1 V_k	Tipo 1		0	Carece de residuo en la posición 31
	Tipo 2	SubtipoA	1	Puente de Hidrogeno entre Y/71 y el residuo en 30
		Subtipo B	1	Posee una F en posición 71
	Tipo 3		7	
	Tipo 4		6	
	Tipo 5		5	
	Tipo 6		2	

Tabla 1. Tipos y subtipos de estructuras canónicas observadas en L1 de V k.

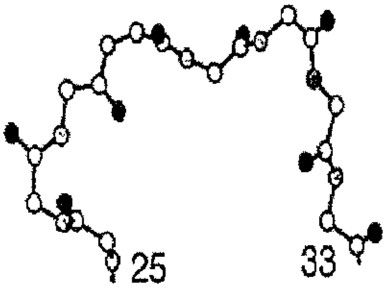


Figura 5. Estructura canónica tipo 1 de L1 Vk

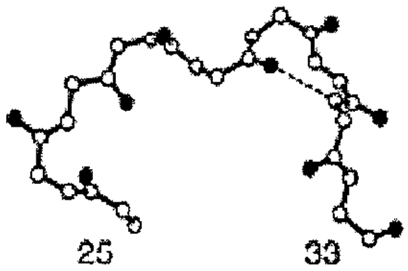


Figura 6. Estructura canónica tipo 2 de L1 Vk

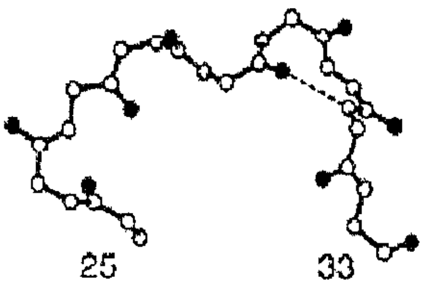


Figura 7. Estructura canónica tipo 3 de L1 Vk

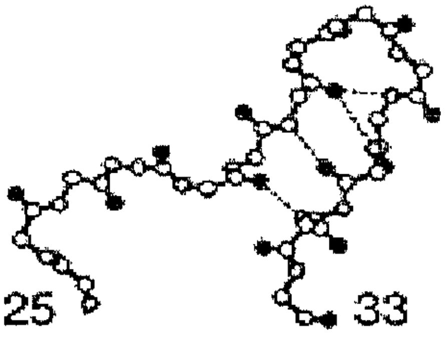


Figura 8. Estructura canónica tipo 4 de L1 Vk

Conformaciones de L1 lambda.

Esta asa se forma entre las posiciones 26 a 32 (Chothia & Lesk, 1987). Hacia 1993 Wu y Cigler determinaron que la posición 25 también forma parte de esta asa. Al igual que L1 de V_k, la conformación de L1 lambda, esta determinada por un residuo hidrofóbico que se empaca en una cavidad hidrofóbica. Se han identificado 4 conformaciones de esta región (Chothia & Lesk 1987, Wu & Cigler 1993, Martin & Thornton 1996, Al-Lazikani *et al.*, 1997). Las principales características de cada tipo de estructura canónica, se resume en la Tabla 2. Las representaciones de la conformación adoptada por esta región se observan en las Figuras 9, 10, 11 y 12.

	Tipo s Observados		**No de residuos entre la posición 25 y 32**	**Característica**
Estructuras	Tipo 1		10	
Canónicas	Tipo 2		11	Posee un residuo adicional (30c), que divide el asa en dos hélices irregulares.
Para L1 Vl	Tipo 3	Subtipo A Subtipo B	11	Orientación del péptido de la posición 28-29 es diferente en estos dos subtipos.
	Tipo 4		9	Forma una hélice distorsionada de la posición 26 a la 30

Tabla 2. Tipos de Estructuras canónicas observadas en L1 de Vλ. Algunas propiedades de cada región definen los tipos y subtipos de canónicas.

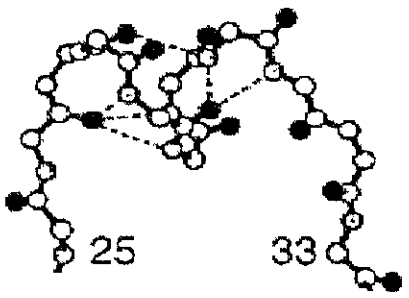


Figura 9. Estructura canónica tipo 1 de L1 lambda

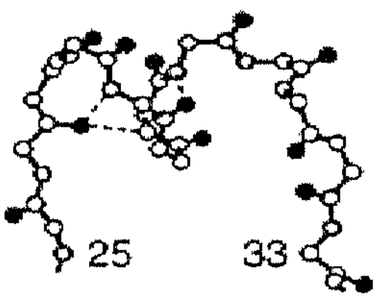


Figura 10. Estructura canónica tipo 2 de L1 lambda

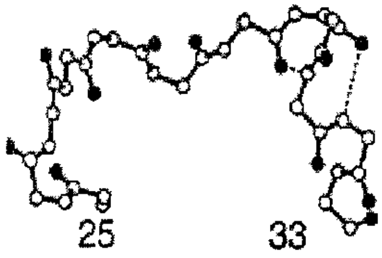


Figura 11. Estructura canónica tipo 3 de L1 lambda

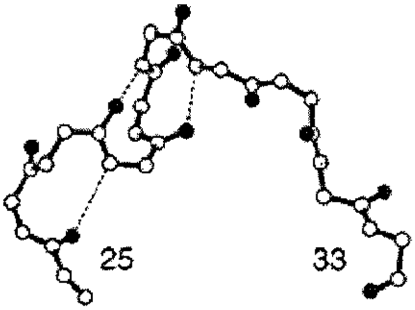


Figura 12. Estructura canónica tipo 4 de L1 lambda

Conformación de L2

La segunda asa hipervariable de V_L o L2, comprende el asa que une las hebras C' y C'', está localizada de la posición 50 a la posición 52, tanto en V_k como en V_λ (Chothia & Lesk, 1987). Estos tres residuos forman el asa y solo se ha identificado una estructura canónica (Figura 13).

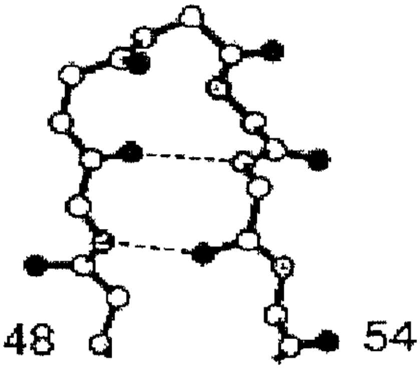


Figura 13.Unica estructura canónica de L2 kappa y lambda

Conformaciones de L3 kappa

La tercer asa hipervariable o L3, une las hebras β: G y F, comprendiendo las posiciones 91 a 96, tanto en V_k como en V_λ (Chothia & Lesk, 1987). Su conformación está determinada por la presencia de una Prolina en la posición 94 o 95 y Glutámico, Histidina o Asparagina en posición 91. Además por la presencia de dos puentes de Hidrógeno entre los residuos 92 y 95. Se han observado 6 estructuras en L3 de V_k (Chothia *et al.*, 1989, Brünger *et al.*, 1991, He *et al.*, 1992, Guarné *et al.*, 1996), el tipo 1 es el más común (Figura 14).

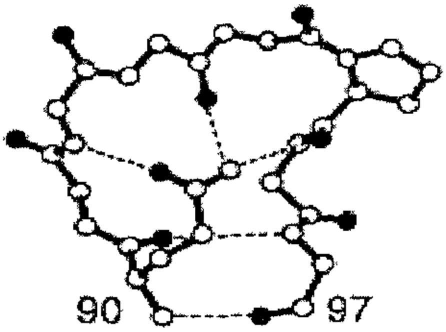


Figura 14. Estructura canónica tipo 1 de L3 kappa

Conformaciones de L3 lambda

Se han observado 2 conformaciones para esta asa, la primera contiene dos residuos en posición 93 y 94 los cuales forman el tope de la asa, diferencias en la orientación de este dipéptido da origen a tres subtipos denominados A, B y C (Figura 15, solo el tipo A). El segundo tipo posee 8 residuos, solo cuatro de ellos constituyen una horquilla de la punta de la asa (Al-Lazikani *et al.*, 1997).

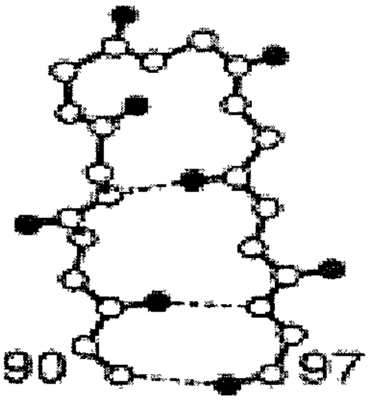


Figura 15. Estructura canónica tipo 1 de L3 lambda

Estructuras canónicas de V_H

Conformaciones de H1

En V_H, también se identifican tres asas hipervariables en posiciones similares a las de V_L (ver Figura 22). H1 comprende de la posición 26 a la posición 32. Al igual que L1, este posee un residuo hidrofóbico en posición 29, el cual se empaca contra los extremos de cadena de los residuos 34, 72 y 77. Los residuos en la posición 27 (Tirosina o Fenilalanina) son parcialmente introducidos en una cavidad. H1 presenta variación en tamaño e incluye inserciones entre la posición 31 y 32 (Chothia *et al.*, 1992). Tres estructuras canónicas fueron observadas para H1. El tipo 1 es el más común (ver Figura 16).

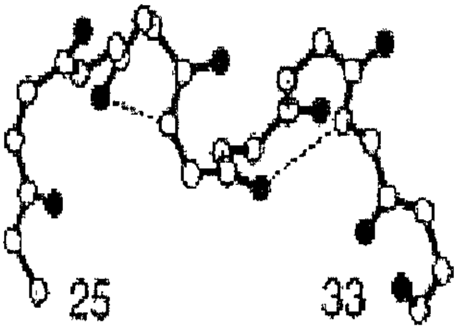


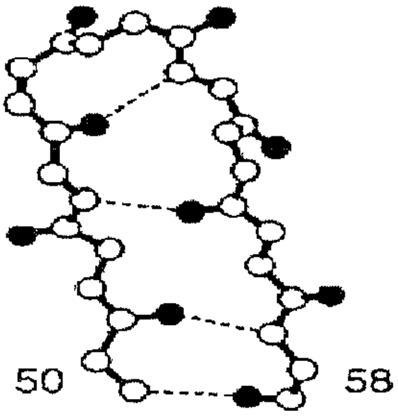
Figura 16. Estructura canónica tipo 1 de H1

Conformaciones de H2.

Esta asa se limita a la región comprendida entre la posición 52 a la 56. Se han observado cuatro tipos de conformaciones de la cadena principal para H2. Un resumen de las principales características de los tipos de canónicas se presenta en la Tabla 3. La representación de cada una de las conformaciones de H2 se observan en las Figuras 17, 18, 19 y 20.

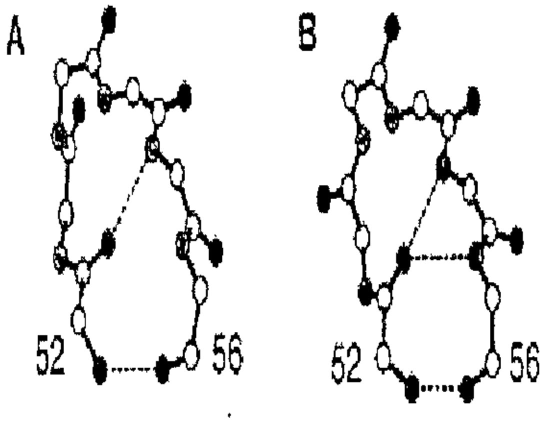
	Tipos Observados		**No residuos entre la posición 52a y 53**	**Característica**
Estructuras	Tipo 1		3	Conformación observada con mayor frecuencia, es la mas corta
Canónicas	Tipo 2	Subtipo A	4	Posee un residuo adicional (30c), que divide el asa en dos hélices irregulares.
de H2		Subtipo B	4	Orientación del péptido de la posición 52a-53 está rotado 160 con relación al tipo A
	Tipo 3	Subtipo A	4	La presencia de Glicinas podría ser el
		Subtipo B	4	Factor para diferenciar estos tres subtipos del
		Subtipo C	4	Del tipo 3 en H2
	Tipo 4		6	

Tabla 3. Tipos de Estructuras canónicas observadas en H2. Algunas propiedades de cada región definen los tipos y subtipos de canónicas.



Tipo 1 de H2

Figura 17. Estructura canónica tipo 1 de H2



Tipo 2 de H2

Figura 18. Estructura canónica tipo 2 de H2

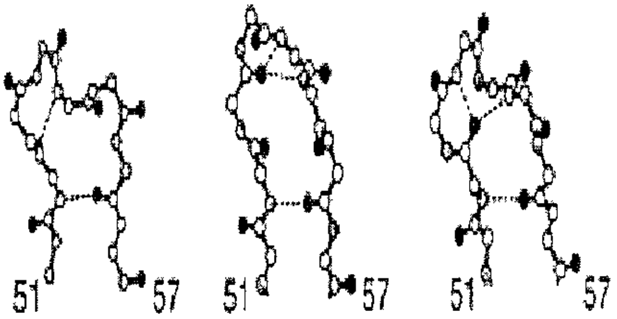


Figura 19.Subtipos A, B y C de la estructura canónica tipo 3 de H2

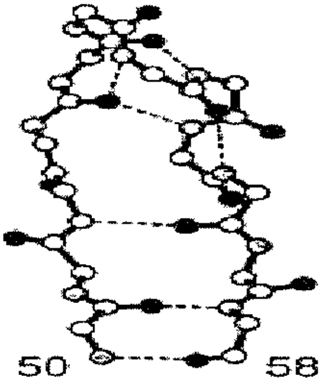


Figura 20. Estructura canónica tipo 4 de H2

Conformaciones de H3

Esta asa va de la posición 92 a 104 (Al-Lazikani *et al., 1997*). La gran variabilidad que presenta H3 tanto en secuencia como en longitud, no ha permitido hacer una correlación entre la secuencia y la longitud para esta asa y por lo tanto, predecir su conformación. Sin embargo, la existencia de residuos conservados en la base de esta región, permiten definir dos estructuras canónicas, denominadas "extendidas" o "torcidas" (Shirai *et al.*, 1996, Martín & Thornton., 1996, Morea *et al.*, 1998, Oliva *et al.*, 1998), las cuales dependen del patrón de puentes de hidrógeno que se formen en esta región de H3.

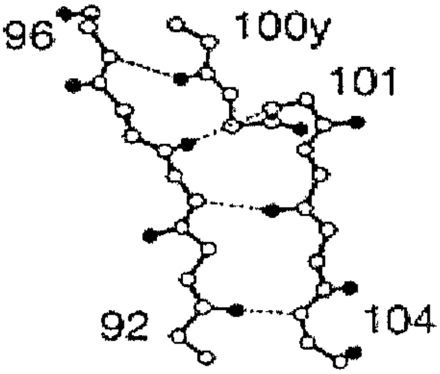


Figura 21. Estructura canónica extendida de H3

Partiendo de las conformaciones canónicas se puede entonces predecir la estructura tridimensional que tendrá una secuencia de aminoácidos de un Fv (Figura 22) en cinco de las seis asas que conforman el sitio de unión con el antígeno. Esto, sin duda, resultó un avance significativo en el análisis de los anticuerpos, puesto que permite, con base en la estructura tridimensional, interpretar resultados experimentales y tomar decisiones sobre que residuos a mutar para mejorar la afinidad de un anticuerpo en particular.

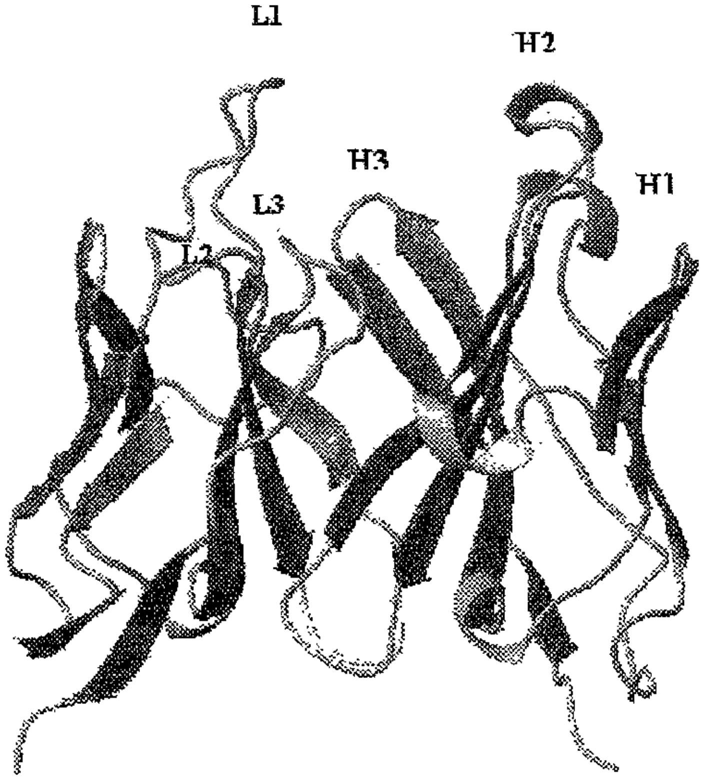


Figura 22. Representación del Fragmento Variable (Fv). El sitio de unión al antígeno en azul celeste , esta formado por seis asas denominadas L1, L2, L3 de la cadena Ligera y H1, H2, H3 de la cadena pesada.

El hallazgo de las estructuras canónicas permitió, además, establecer un esquema predictivo que correlaciona la secuencia de aminoácidos, la estructura tridimensional del sitio de unión con los antígenos y los tipos de antígenos reconocidos por los anticuerpos (Vargas-Madrazo *et al*., 1995a, Lara-Ochoa *et al*., 1996).

Estudios en las secuencias de aminoácidos de los genes de línea germinal (Chothia *et al.*, 1992, Tomlinson *et al.*, 1992, Cox *et al.*, 1994), pseudogenes (Vargas-Madrazo *et al.*, 1995b, Almagro *et al.*, 1995) y en secuencias de aminoácidos maduras de Igs, han confirmado la existencia de estructuras canónicas para la mayoría de estas asas, excepto para H3. Nuestros estudios, mostraron que de las 300 posibles combinaciones de estructuras canónicas, solo 10 combinaciones describían el 90% de todas las secuencias de anticuerpos con especificidad reportada (Vargas-Madrazo *et al.*, 1995a). Se mostró también una correlación entre la geometría del sitio de unión al antígeno (determinada por cierta combinación de estructuras canónicas) y el tipo de antígeno reconocido (especificidad). Por ejemplo anticuerpos que unen antígenos proteicos están caracterizados por sitios de unión con superficies planas, aquellos que unen péptidos o DNA poseen superficies con una pequeña ranura y los que unen haptenos poseen una cavidad (Vargas-Madrazo *et al.*, 1995a, Ramirez-Benitez *et a*l., 2001a, Ramirez-Benitez *et al.*, 2001b). Estas correlaciones han sido confirmadas por otros grupos de investigación utilizando diferentes metodologías (MacCallum *et al.*, 1996). Así, los resultados obtenidos a la fecha, permiten el avance en la comprensión del mecanismo de interacción antígeno-anticuerpo, lo que ayuda a mejorar los esquemas predictivos y el diseño racional de anticuerpos con especificidad predeterminada

Capítulo II

VIR.II: A new interface with the antibody sequences in the Kabat database

VIR.II:
A new interface with the antibody sequences in the Kabat database

Maria C. Ramirez-Benitez[1] , Gabriel Moreno-Hagelsieb[2] and
Juan Carlos Almagro[1] *

1. Instituto de Biotecnología, UNAM
2. Programa de Biología Computacional, Centro de Investigación sobre Fijación de Nitrógeno, UNAM.

***Corresponding author**

Instituto de Biotecnologia, UNAM
Av. Universidad # 2001 C.P. 62210
Cuernavaca, Morelos.
Mexico

Telephone (Lab): + 52 (7) 329-1605
Telephone (Office): +52 562-27845
FAX: +52 (7) 317-2388

Email: almagro@ibt.unam.mx

Keywords:

Immunoglobulin, Complementarity Determining Regions, CDRs, Antigen-binding site, amino acid sequences, nucleotide sequences.

Running head: Interfacing the Kabat database with VIR.II.

Total number of pages: 17, including 2 Figures and 1 Table.

Abstract

The Kabat database is the source of information par excellence on antibody sequences. In 1995 we developed an interface with the Kabat database, called VIR. VIR has been very useful in conducting studies aiming to find structure-function relationships in antibodies. Here we report a new version adapted to the World Wide Web, called VIR.II. VIR.II allows searches by type of chain (V_H or V_L), by species and by specificity. In contrast to other interfaces designed to deal with the Kabat database, the way to select a given species or specificity is by choosing from a list. This avoids mistakes and redundancies in the searches. Another feature, and probably the most important one, is that VIR.II introduces a classification of specificities in terms of the chemical and biochemical nature of the antigen, like anti-protein, anti-peptide, etc. This classification has been useful in discovering patterns in the antigen-binding site of antibodies that correlate with the type of antigen the antibody interacts with. To illustrate this, while showing the capabilities of VIR.II, we analyze all the murine anti-peptide and anti-protein antibody sequences currently compiled in the Kabat database (as of July, 2000).

Introduction

Antibodies recognize a seemingly unlimited number of antigens with exquisite specificity. The first antibody sequence was determined in the mid-1960s (Hilschmann and Craig, 1965). In the early 1970s, Tai Te Wu and Elvin A. Kabat compiled all the complete and partial sequences for antibodies published at that time; 77 in total (Wu and Kabat, 1970). That constituted the first version of the Kabat database (Johnson and Wu, 2000). The analysis of such information allowed the location, prior to the resolution of the first three-dimensional structure of an antibody (Poljak et al., 1972), of the so-called complementarity determining regions (CDRs); the regions where the capability of this family of proteins to recognize a diverse number of antigens rests (Wu and Kabat, 1970).

The number of antibody sequences in the Kabat database has increased exponentially (http://www.ibt.unam.mx/vir/VIR/stat_aim_vir.html), amounting today (July, 2000) to 19,382 sequences. These sequences have been isolated from more than 70 species and as many as 7,989 of them have the specificity annotated (4,547 from V_H and 3,442 from V_l [kappa + lambda]). Thus, the Kabat database is the source of information par excellence for studies aiming to find correlations among sequence patterns in the CDRs, the species and/or the specificity of the antibodies.

To access the ever-increasing and valuable information gathered in the Kabat database, two interfaces have been developed and are currently available in the World Wide Web (WWW): SeqhuntII (Johnson et al., 1995) and Kabatman (Martin A.C.R, 1996). SeqhuntII (http://immuno.bmc.nwu.edu/seqhunt.html) allows searches through the annotations and sequences (Johnson and Wu, 2000) and the output is a fixed-line length record of 80 characters per line. This output format constrains the user to see one sequence at once. Therefore, to find patterns in the sequences raised against a given specificity, the user needs to run many queries, and the information has to be put together by using home-written programs. There is an electronic mail server (seqhunt2@immuno.bme.nwu.edu), which is considerably more flexible, though it does not offer ease of use.

Kabatman allows searches using an SQL-like query language (http://www.bioinf.org.uk/abs/kabatman.html) and the output is a list of sequences aligned following the Kabat conventions (Kabat et al., 1991). Sequences having a given pattern in the CDRs, for instance: the CDR-L1 of 11 residues and a proline at position L29, can be obtained by using Kabatman (Martin A.C.R, 1996). That is not possible when using SeqhuntII. In addition, the URL http://www.bioinf.org.uk/abs/simkab.html provides access to a simple point-and-click interface, allowing user-friendly searches. However, the searches are limited to amino acid sequences.

Early in 1995 we developed an interface to manage the antibody sequences available in the Kabat database (Almagro et al., 1995). Our interface, developed to run on PCs, was called VIR (Variable domains of the **I**mmune system **R**eceptors). During the last six years, VIR has been very useful in conducting studies that have discovered structure-function relationship in antibodies (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996; Almagro et al., 1997; Almagro et al., 1998). This, together with the fact that VIR solved several of the SeqhuntII and Kabatman limitations, stimulates the implementation of a new version of VIR. The new version, designed to be accessible through the WWW, has been called VIR.II (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi).

Main features of VIR.II

VIR.II allows searches of amino acid and nucleotide antibody sequences. The information at the nucleotide level is indispensable for determining with precision the genetic mechanisms that originated a particular antibody, *i.e.*, the germline V, D, J gene combination, the addition and/or deletion of nucleotides at the V-J, V-D or D-J junctions and the putative somatic mutations (see for example IMGT/V-QUEST at http://imgt.cines.fr).

In addition, VIR.II introduces a classification of the specificities in terms of the chemical and biochemical nature of the antigen. This classification has allowed the finding of rules to correlate the primary structure of an antibody with its capability to recognize different kinds of antigens (Vargas-Madrazo et al., 1995). Also, in contrast to SeqhuntII and Kabatman, the way to select a given species is through a pulldown menu. This avoids mistakes and redundancies in the searches.

The output of VIR.II is a sequence alignment with a header showing the Kabat numbering schedule (Kabat et al., 1991). The sequences in the output are aligned following the conventions on the placement of deletions/insertions at CDRs proposed by Kabat and co-workers (Kabat et al., 1991). The Kabat ID identifies each sequence in the alignment and this has a hyperlink to the Kabat database. Thus, the original source can be consulted in any case.

Implementation

VIR.II searches through databases created in DBM (database management) format (Glover and Humprey, 1996). The DBM files are generated starting from the information contained in the dump files of the Kabat database (ftp://ncbi.nlm.nih.gov/repository/kabat/fixlen/). These files are downloaded after each Kabat database update.

The interface (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi) consists of three sections (Figure 1). In the first section, the user can select the kind of sequence (amino acid or nucleotide), the type of chain (VH, Vkappa or Vlambda) and the species.

The way to select a given species is by choosing in a pulldown menu. The pulldown menu is generated by searching for unique species in the raw information downloaded from the dump files of the Kabat database. This preprocessing of the information avoids mistakes and redundancies in subsequent searches. For example, some entries in the Kabat database have the species annotated in colloquial terms like "frog", whereas others use the scientific nomenclature: "*Xenopus laevis*". Moreover, it should be noted that the pulldown menu of species is in itself an inventory of the species, thus offering easy access to the number and the kind of species that have been used to isolate the antibody sequences compiled in the Kabat database.

The second section of VIR.II is designed to work with the specificities. As with species, a list of unique specificities is generated from the raw information downloaded from the dump files of the Kabat database. The list is available to the user. In this case, however, the specificity to seek should be written or pasted from the list in a text box. This allows more flexibility in the search. Also, we implemented the facilities "OR" and "AND". By using the list and the "OR" and "AND" facilities, the user can construct strings that could take into account all the variants of a given specificity; including misspellings and redundancies.

For example, we have found four different definitions for the specificity "ANTI-(4-HYDROXY-3-NITROPHENYL) ACETYL":

1. ANTI-(4-HYDROXY-3-NITRO-PHENYL) ACETYL ("-" after "NITRO"),

2. ANTI-(4-HYDROXY-3-NITROPHENY)ACETYL (no "L" after "PHENY"),
3. ANTI-(4-HYDROXY-3-NITROPHENYL) ACETYL (space before "ACETYL") and
4. ANTI-(4-HYDROXY-3-NITROPHENYL)ACETYL (no space before "PHENYL").

When the Kabat database (as of July, 2000) was queried with VIR.II and these definitions were used, sets of different sequences were obtained. Definition 1 gave 10 sequences, definition 2 gave 75 sequences, definition 3 gave 248 sequences and definition 4 gave 263 sequences. To obtain all of them, a search using all of the definitions and the facility "OR" was conducted. An alternative is to use the substring "ANTI-(4-HYDROXY-3-NITRO", which is common to all the definitions. The hapten "ANTI-(4-HYDROXY-3-NITROPHENYL) ACETYL" is not an exception, there are many more examples.

VIR.II also introduces a classification of antigens in seven groups (see Figure 1). These groups have been called gross specificities (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996). The classification is based on the chemical and biochemical nature of the antigen. By using a similar classification, different types of antigen-binding sites have been discovered, some of them with preference for the recognition of proteins, peptides, haptens, carbohydrates, nucleic acids, and so on, and others with multi-specific capabilities (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996).

There is no way to classify the specificities in gross specificities automatically. Once a number of antigens were classified, the process was achieved in a semi-automatic way. Today, in each update, the information downloaded from the Kabat database is compared with the specificities that have previously been classified. If a new sequence is specific, for instance, for lysozyme, then it is classified as anti-protein. Otherwise, a file with exceptions is generated and the specificity is classified *de novo*. Obviously, the process is more automatic as more antigens are classified. The list of antigens, as classified in gross specificities, can be consulted in the URL: http://www.ibt.unam.mx/vir/VIR/gross_specificity.html.

The third section of VIR.II contains two options that have been designed to filter the search. The first option has been denoted "completeness" to differentiate it from the definition of a complete antibody of Kabatman. Completeness allows selection of sequences with a certain percentage of undefined positions in a given segment. This kind of filtering is very useful when considering that a mere 20% of the sequences (3,850 out of 19,382) available in the Kabat database are 100% complete.

The other option, called "counterpart", is equivalent to the option "complete" in Kabatman. This allows searches of partners. If the user selects for instance V_H chains then, by choosing counterpart, all the V_L chains (kappa and lambda) that share the same name of the V_H chains, will be obtained. This filter is indispensable when patterns in the antigen-binding site as a whole (V_H+V_L) are wanted.

Example of application

Studies conducted with VIR have revealed features at the antigen-binding site of antibodies that correlate, in some cases, with the specificity (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996; Almagro et al., 1997; Almagro et al., 1998). For instance, we have found that antibodies with a short CDR-L1 preferentially recognize large antigens such as proteins (Vargas-Madrazo et al., 1995). In contrast, antibodies with a long CDR-L1 tend to bind smaller molecules such as peptides (Vargas-Madrazo et al., 1995). To illustrate that finding, while showing the potential use of VIR.II, in this section we analyze the lengths of the CDR-L1 in the murine anti-peptide and anti-protein antibody sequences available in the Kabat database (as of July, 2000).

We choose amino acid as kind of sequence, Vkappa as chain type and mouse as species. Then, we choose anti-peptide in gross specificity. Finally, we set 24 as the beginning position, and 34 as the ending one [according to the definition of CDR-L1 of Kabat et al. (1991)]. The completeness was defined as 100%. We found 103 anti-peptide sequences. In the case of anti-protein antibodies, we just changed to anti-protein in gross specificity. We found 624 sequences.

Figure 2 shows the distribution of the CDR-L1 lengths of the anti-protein and anti-peptide sequences. The lengths of the CDR-L1 follow a bimodal distribution. Interestingly, no sequence has a CDR-L1 of 13 residues. In anti-protein antibodies the proportion of short/long CDR-L1 (short < 13 residues; long >13 residues) is similar: 58% and 42%, respectively. The anti-peptide sequences are biased towards the use of a long CDR-L1 (84% of the sample).

To test whether this bias is related to the classification of the sequences in anti-protein and anti-peptide antibodies, a 2 x 2 contingency table (Table 1) was set up. The X^2_c gave 30.231 and $X^2_{0.001,1}$ is 10.828 (Zar, 1999), which means that the bias in the length of the CDR-L1 is related to the classification of the sequences in anti-peptide and anti-protein antibodies. A similar analysis could be conducted for the remaining CDRs, in particular for the CDR-H3, where biases in the length of this loop may be related to the specificity (Kabat and Wu, 1991). Also, a study of the combination of lengths in CDRs in antibodies that interact with different kind of gross specificities it could be conceived.

It is worth mentioning that this kind of analysis is difficult to be achieved by using Kabatman. By using that interface, many queries, one for each CDR length, should be run. Then, the sequences should be classified in terms of gross specificities. This latter task, as mentioned above, is not a trivial one since the classification of antigens within groups is not a straightforward process.

Conclusion

Antibodies constitute a paradigm of molecular recognition. They bind to a virtually infinite number of antigens with exquisite specificity. To decipher the molecular basis of such a feature several thousands of antibody sequences have been obtained in the last 40 years. As a consequence, several specialized immunogenetic databases have been created (Brusic et al., 2000; Dübel, 2000). Despite this accumulation of information, its classification and analysis, it is yet not possible to predict the specificity of a given antibody sequence; even if much is known about the antigen structure.

One of the features of VIR.II is that the amino acid and nucleotide sequences available at the Kabat database can be queried by using a classification of the antigens in terms of gross specificities. This classification may be the seed of a taxonomy of antigens. This would be useful in unraveling the common features of antibodies that recognize similar antigens. Actually, by using this classification, patterns at the antigen-binding site have emerged (Vargas-Madrazo et al., 1995; MacCallum et al., 1996). Importantly, such patterns correlate with the propensity of given antibody to recognize a particular kind of antigen (Vargas-Madrazo et al., 1995). Therefore, such patterns may be used to predict the specificity of a given antibody sequence (Vargas-Madrazo et al., 1995).

In the section "example of application", we showed how, by using VIR.II, a long CDR-L1 may be associated to antibodies that bind to peptides. Such a rule is still quite general and rudimentary. However, we envision that systematic analyses of the ever-increasing and valuable information available in the Kabat database, *via* VIR.II, would generate rules of increasing degree of complexity. Such rules probably allow, in the near future, better (fine-tuning) predictions of the recognition features of antibody sequences.

Acknowledgments. M.C. Ramirez-Benitez was supported by CONACyT. This work was partially supported by the grants of DGAPA-UNAM IN223399 and CONACyT 33352-N.

References

Almagro, J.C., Vargas-Madrazo, E., Zenteno-Cuevas, R., Hernandez-Mendiola,V. and Lara-Ochoa, F. (1995) VIR: A computational tool for analysis of immunoglobulin sequences. *BioSystems.*, **35**, 25-32.

Almagro, J.C., Hernandez, I., del Carmen Ramirez, M., Vargas-Madrazo, E. (1997) The differences between the structural repertoires of VH germ-line gene segments of mice and humans: implication for the molecular mechanism of the immune response. Mol Immunol., **34**, 1199-214.

Almagro, J.C., Hernandez, I., Ramirez, M.C., Vargas-Madrazo, E. (1998) Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. Immunogenetics., **47**, 355-63.

Brusic, V., Zeleznikow, J., Petrovsky, N. (2000) Molecular immunology databases and data repositories. J Immunol Methods., **238**, 17-28.

Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., **et al.** (1989) Conformations of immunoglobulin hypervariable regions. Nature., **342**, 877-83.

Dübel S. 2000. The antibody web. Immunol. Today. **21**: 355-357.

Glover, M., Humphey A. (1996) Chapter 10: DBM files. *Perl 5 how-to.* The Waite Group, Inc. pp. 247-275.

Hilschmann, N. and Craig, L,C. (1965) Amino acid sequence studies with Bence-Jones proteins. Proc. Natl. Acad. Sci. USA., **53**, 1403-1409.

Johnson, G., Wu, T.T., Kabat, E.A. (1995) SEQHUNT. A program to screen aligned nucleotide and amino acid sequences. Methods. Mol Biol., **51**, 1-15.

Johnson, G. and Wu, T.T. (2000) Kabat Database and its applications: 30 years after the first variability plot. Nucleic Acids Research., **28**, 214-218.

Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. and Foeller, C. (1991) *Sequences of proteins of immunological interest.* 5th Edn., Public Health Service. N.I.H. Washington. D.C.

Kabat, E.A., Wu, T.T., (1991) Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. J Immunol., **147**, 1709-19.

Lara-Ochoa, F., Almagro, J.C., Vargas-Madrazo, E. and Conrad, M. (1996) Antibody-antigen recognition: a canonical structure paradigm. J. Mol. Evol., **43**, 678-684.

MacCallum R.M., Martin A.C.R., Thornton J.M. (1996) Antibody-antigen interactions: contact analysis and binding site topography. J Mol Biol **262**:732-45

Martin, A.C.R. (1996) Accessing the Kabat Antibody Sequence Database by Computer PROTEINS: Structure, Function and Genetics., **25**, 130-133.

Poljak, R.J., Amzel, L.M., Avey, H.P., Becka, L.N. (1972) Structure of Fab' New at 6 A resolution. Nat. New. Biol., **235**, 137-40.

Vargas-Madrazo, E., Lara-Ochoa, F. and Almagro, J.C. (1995) Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. J. Mol. Biol., **254**, 497-504.

Wu, T.T. and 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 antibody complementarity. J. Exp. Med., **132**, 211-50.

Zar, J.H. (1999) Contingency tables. In: *Biostatistical analysis*. 4th ed. Prentice Hall, Inc. pp. 486-515.

Figure 1. Structure of VIR.II (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi)

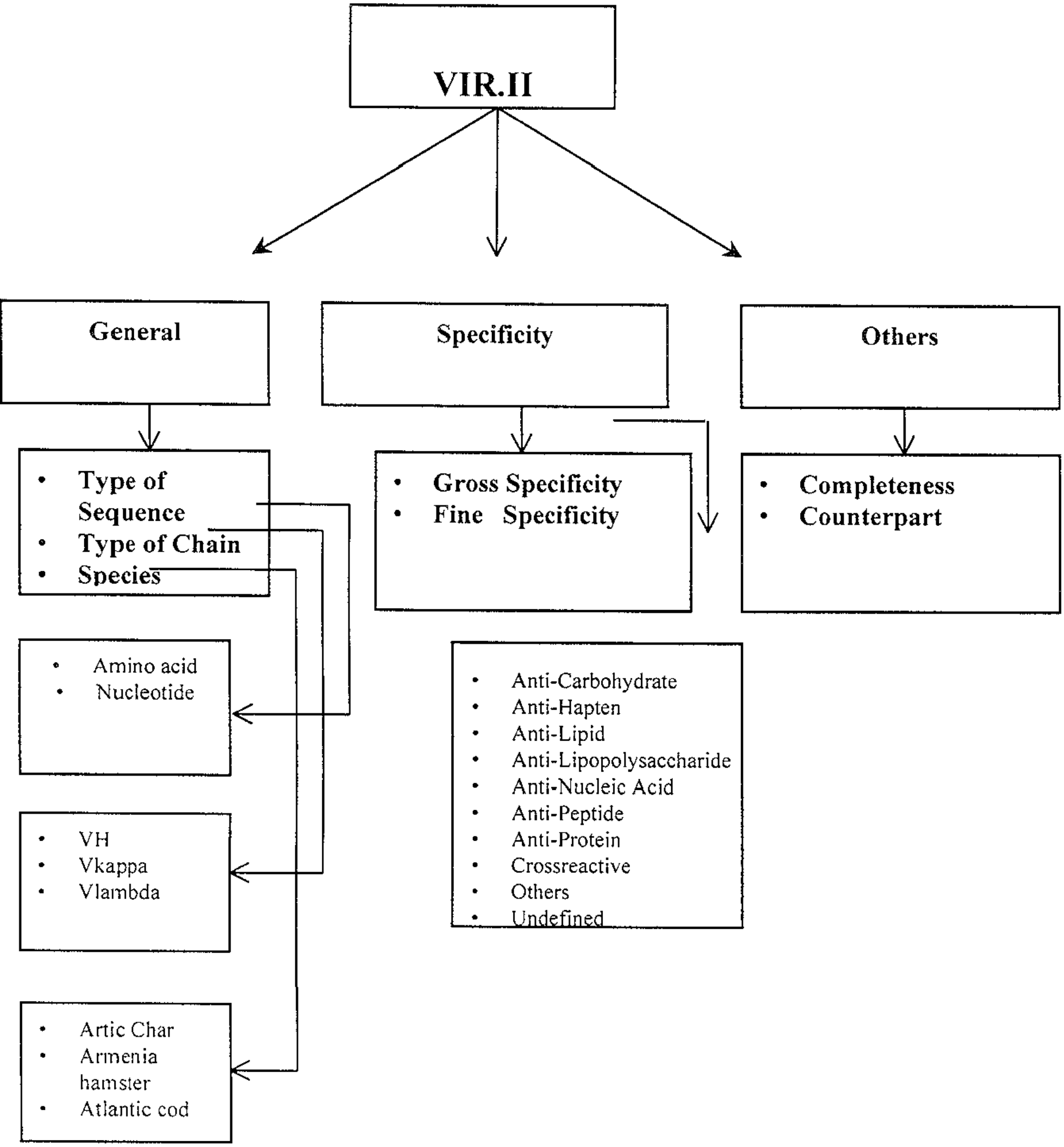


Figure 2. Distribution of the CDR-L1 lengths in the murine anti-protein and anti-peptide antibody sequences available in the Kabat database.

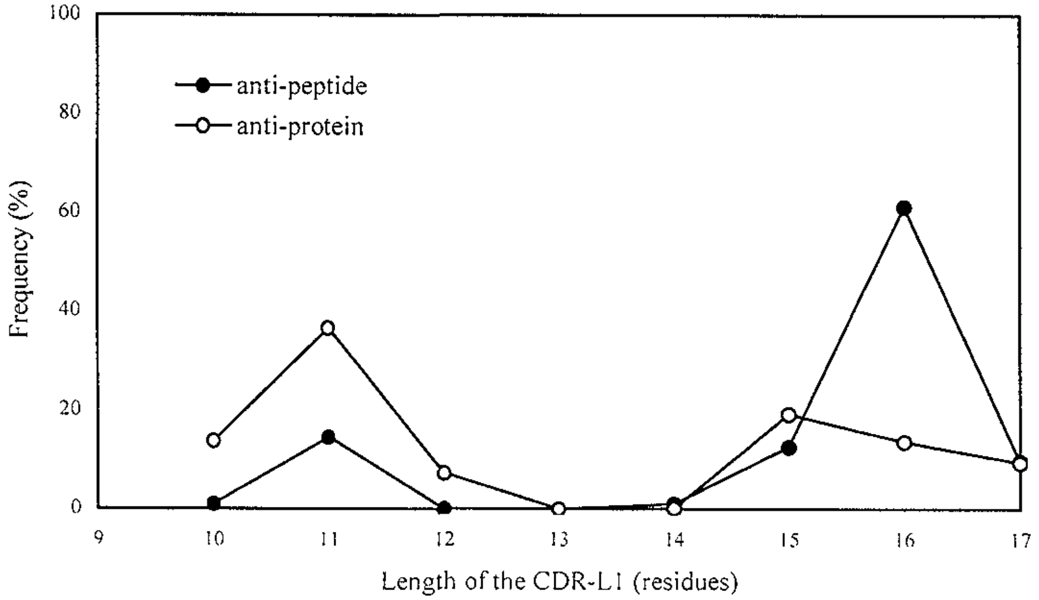


Table 1. Contingency table of the CDR-L1 length distribution, as grouped in short and long loops, and the sequences classified as anti-protein and anti-peptide antibodies.

	CDR-L1 Length		
Gross specificity	<13	>13	Total
Anti-protein	359	265	624
Anti-peptide	16	87	103
Total	375	352	727

Capítulo III

Structure-function relationships in anti-protein, anti-peptide and anti-hapten antibodies

Structure-function relationships in anti-protein, anti-peptide and anti-hapten antibodies

Maria C. Ramirez-Benitez, Hector A. Ceceña and Juan C. Almagro[1]

Instituto de Biotecnologia, Universidad Nacional Autonoma de Mexico.
Av. Universidad # 2001 C.P. 62210
Cuernavaca, Morelos. Mexico

[1]To whom correspondence should be addressed.

Current address:
Fred Hutchinson Cancer Research Center,
Human Biology Division
1100 Fairview Ave. N.
Mailstop C3-168
PO Box 19024
Seattle, WA 98109-1024

FAX: (206) 667-6524
Telephone: (206) 667-6837
Email: almagro@ibt.unam.mx

Keywords

Antigen-binding site, Topography, Hypervariable loops length, Immunoglobulins, Igs, vaccines.

Total number of pages: 40, including 1 Table and 7 Figures .

Word count: 7,000.

Abstract

Antibodies constitute the paradigm of molecular recognition per excellence. Here, to gain insight into the rules governing their mechanism of binding, we analyzed 94 unique antigen-antibody complexes of known three-dimensional structure: 49 anti-protein, 20 anti-peptide and 35 anti-hapten complexes.

On inspection of the hypervariable loop length, we found that 33 out of 49 (67%) anti-protein structures had a short loop at L1. In contrast, 17 out of 20 (85%) of anti-peptide antibodies had a long L1. Study of 1,017 and 121 anti-protein and anti-peptide sequences respectively, confirmed the finding in the structures. Analysis of the antigen-binding site topography indicates that insertion of residues at L1 switches the antigen-binding site topography from flat to grooved and flat antigen-binding sites grant a good protein-antibody complementarity, whereas the grooved ones accommodate peptides.

Anti-hapten and anti-peptide structures were similar regarding to the proportion of short/long L1 loops. However, analysis of 632 anti-hapten sequences yield that they are similar to anti-protein sequences. This lack of consistency suggested that the recognition of haptens might not be related to changes in the length of L1. The profile of residues in contact with the antigens indicates that hapten are recognized deep in the antigen-binding site. Haptens, being in the range of size of the side chains, do not need large modification at the antigen-binding site topography as those achieved by indels at L1.

Introduction

Antibodies recognize a virtually infinite number of molecules with exquisite specificity, thus constituting the molecular recognition paradigm per excellence. Although many aspects of such a feature have been clarified after decades of research, is still not very well understood what determine the capability of certain antibody to distinguish a particular antigen o group of antigens in the vast and complex antigenic universe. This has limited our ability to develop methods for predicting the specificity of antibody genes and designing *de novo* antibodies with desired specificity.

The antigen-binding domains, V_L and V_H, are composed of a very well-conserved two ß-sheets framework and six hypervariable loops, which defines the antigen-binding site, denoted L1, L2 and L3 for V_L and H1, H2 and H3 for V_H (Wu and Kabat, 1970; Amzel and Poljak, 1973). Although these loops are highly variable in length and sequence, five of them, exception is H3, are built of a very few structural solutions, called canonical structures (Chothia and Lesk, 1987; Chothia et al., 1989; Al-zakani et al., 1997). In addition, it have been shown that antibodies use a small amount of all possible combinations of canonical structures (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996), which suggest structural restrictions in the antigen-binding site at work in the process of antigen recognition.

We have found that antibodies combining short canonical structures in L1 and H2 have a flat antigen-binding site, in contrast to antibodies with long canonical structures in L1 and H2 that have a groove (Vargas-Madrazo et al., 1995; Lara-Ochoa te al., 1996). Flat antigen-binding sites have been found to correlate with the propensity of antibodies to bind large antigens such proteins, whereas grooved ones have been related to the recognition of small molecules, namely peptides and haptens (Wilson and Stanfield, 1993; Webster et al., 1994; Vargas-Madrazo et al., 1995; MacCallum et al., 1996). These correlation between then amino-acid sequence (understood as the hypervariable loop length) the antigen-binding site topography and the recognition properties of antibodies, though still rudimentary, may the seed of methods to predict the specificity of antibody genes (Vargas-Madrazo et al., 1995).

Over the last few years, the number and diversity of antibodies of known three-dimensional structure has increased largely (URL: http://www.ibt.unam.mx/vir/structure/structures.html). Here we have taken advantage of that fact to gain further insight into the structural determinants of antibodies that bind different kind of antigens. We have analyzed all anti-protein, anti-peptide and anti-hapten antibodies of known and public three-dimensional structure, 104 structures in total. Results confirmed previous findings, while add new elements our understanding of rules governing the recognition of different king of molecules by antibodies. We discuss the use of the knowledge to design repertories biased toward the recognition of different kind of molecules.

Results

Anti-protein, anti-peptide and anti-hapten antibodies of known structure

Until March of 2001, the PDB (Berman et al., 2000; URL: http://www.rcsb.org/pdb/) collected co-ordinates for 94 unique kappa type antibodies that recognize proteins, peptides or haptens: 45, 19, 30 structures respectively (Table 1). Unique means antibodies with different names. For antibodies having co-ordinates for the bound and free forms, we choose the complex. If the same antibody was determined at different resolutions, we used the structure of better resolution.

Antibodies from camelidaes (camel and lama) were not included in the analysis. On other hand, the majority of the structures, being from mice, were kappa type antibodies. We found only 10 lambda type antibodies: 4 anti-protein, 5 anti-hapten and 1 anti-peptide. Thus, we decide to focus our analysis in kappa type antibodies only.

Length of the hypervariable loops

The length of hypervariable loops is the main structural determinant of the antigen-binding site topography (Bolger and Sherman, 1991; Wilson et al., 1994; Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996), thus we analyzed it firstly.

L1 is the most diverse loop in terms of lengths, covering a range of 8 lengths, from 6 to 13 residues (column 8 of Table 1). The distribution of lengths was found to be bimodal. No loop has 9 residues, defining the boundary between a short (≤ 8 residues) and a long (≥ 10 residues) L1. Anti-protein structures showed 73% and 26% of short and long L1, respectively. In contrast, anti-peptide structures are biased towards the use of a long L1. All anti-peptide structures but 2 (10%) use a long L1. Anti-hapten structures are similar to anti-peptide antibodies. However, the distribution is less skewed: 63% anti-hapten structures have a long L1 and 37% have a short one.

L2, being highly conserved in structure (Al-zakari et al., 1997) was not further analyzed. L3, on other hand, has 6 lengths, from 3 to 8 residues. Most structures have 6 residues: 78%, 95% and 90% of anti-protein, anti-peptide and anti-hapten structures, respectively.

In V_H, H1 and H2 are relatively conserved in all structures. H1 has 3 lengths: 7, 8 and 9 residues. Structures with 7 residues in H1 predominate the three samples (90%, 84% and 90% of the anti-protein, anti-peptide and anti-hapten structures, respectively). H2 has 4 lengths, from 5 to 8 residues. The six-residues loop preponderate: 67%, 79% and 70% of the anti-protein, anti-peptide and anti-hapten structures, respectively. We found a slight bias toward the use of the shorter H2 loop in anti-protein structures (27%).

H3, as expected, is by far the most diverse loop of the antigen-binding site, with a range of lengths covering 17 lengths, from 4 to 20 residues. The length distribution in the three samples follows a Poisson distribution . The average length of H3 in anti-protein and anti-peptide structures: 11 and 12 residues, respectively, whereas anti-hapten structures have in average 10 residues. The difference of 1 or almost 2 residues in anti-haptens antibodies with respect to anti-protein and anti-peptide structures is due to the overuse of ten-residue loops in anti-hapten structures (see table 1).

Length distribution of L1 in sequences

Results from the above section point toward the length of L1 as the main determinant of the recognition of large antigens like proteins, from the recognition of small or medium antigens like peptide or haptens. To test the scope of that finding, we analyzed all the anti-protein, anti-peptide and anti-hapten kappa sequences compiled in the Kabat database (Johnson and Wu, 2000; Johnson and Wu, 2001; http://immuno.bme.nwu.edu).

By using VIR.II (Ramirez-Benitez et al., 2001; http://www.ibt.unam.mx/vir/VIR/vir_index.html) we found, up to April, 2001, 1,017 anti-protein kappa sequences with L1 completely determined. Among them, 624 sequences are murine, 372 come from humans, 14 are from rat, 6 from hamster and 1 is a chimpanzee sequence. The number of anti-peptide sequences is almost one order of magnitude lesser, amounting to a total of 121 sequences: 104 from mouse and the remaining 17 from humans. We found 632 anti-hapten sequences, almost all of them coming from mice (628 sequences), 3 were from rabbit and 1 from human.

As in the structures, the length of L1 followed a bimodal distribution with no loop of 9 residues (data not shown). Anti-protein sequences have 68% and 32% of short and long L1 loops, respectively. Anti-peptide sequences have 28% and 72%. Therefore, anti-peptide and anti-protein sequences show a proportion of short/long L1 loops similar to the structures. Similarity with respect to the structures indicates that the differences in L1 between antibodies that bind to proteins and peptides is a general trend. Moreover, since the sample of sequences is more heterogeneous in terms of species, the differences in L1 appear to be beyond the of species barrier.

Anti-hapten antibodies, on other hand, showed a frequency of short/long L1 loops similar to anti-protein sequences. Sixty-seven per cent and 33% of anti-hapten sequences had a short or long L1, respectively. This lack of consistency with respect to the structures suggests that the recognition of haptens may not be related with the length of L1.

Structural meaning of changes in the length of L1

To understand the meaning of differences in length of L1 and the relationship with protein or peptide recognition, we superposed the trace of all anti-protein structures with a short L1 as well as the trace of all anti-peptide structures with a long L1 (Figure 1 and 2). Counterexamples, namely anti-protein structures with a long L1 and anti-peptide structures with a short L1, will be analyzed in the next section.

Anti-protein structures with short L1 show a flattened antigen-binding site (Figure 1), with some variations in that basic topography due to differences in H3 (see below). Anti-peptide structures shared a groove (Figure 2a), capable of accommodating the peptides (Figure 2b). The groove is mainly a consequence of inserting residues at the tip of L1 (in red in the figure 2b). Therefore, insertion or deletion of residues at L1 switches the gross shape of the antigen-binding site from flat to grooved and thus the preference of antibodies to bind proteins or peptides.

Counterexamples that confirm the rule

Within the 69 anti-protein and anti-peptide structures we found 14 counterexamples: 12 anti-protein structures have a long L1 and then should have a grooved antigen-binding site, and 2 anti-peptide structures have a short L1 and then should have a flattened antigen-binding site (see Table 1). Within the anti-protein structures with a long L1, 5 of them have been co-crystallized with the antigen: 2jel, 1nsn, 1afv, 2vir and 1qfu (see Table 1).

1qfu is an exception to any rule, since the interaction with the antigen is established out of the antigen-binding site (Fleury et al., 1999). The complexes with the PDB code 2jel and 1nsn show a protein-antibody interaction that is not fully complementary; no contacts at the center of the grooved antigen-binding site are established (Figure 3a and 3b). Typical anti-protein structures guarantee an adequate protein-antibody complementarity establishing the proper interactions at the center of the antigen-binding site due to the flattened surface these kind of antibodies have (Figure 3c).

The third exceptional structure, 1afv, shows a good antigen-antibody complementarity (Figure 3d). The epitope involves the carboxy end of helix D and three residues from the helix E of the capside protein of human inmunodeficiency virus (Momany et al., 1996). Interestingly, the main portion of the epitope protrudes from the antigen (Figure 4), thus filling the grooved antigen-binding site. Therefore, this counterexample together with 2jel and 1nsn, suggests that antibodies with a long L1, but recognizing proteins, have poor antigen-antibody complementarity, except when the epitope protrude from the plane of antigen and fill the groove.

Counterexamples by the side of anti-peptide antibodies, i.e., anti-peptide antibodies with a short L1 (PBD codes: 1ikf and 1bog) have a grooved antigen-binding site as any other anti-peptide structure (Figure 5a and 5b). The short L1 of 1ikf is compensated by a long H3: 17 residues (see Table 1). Such a long H3 shapes the groove instead of the typical hairpin loop at L1 (Figure 5c). 1bog, on other hand, has a short, shorter than usual, H3: 4 residues (see Table 1). This short H3 "built" the groove instead a long L1or H3 (Figure 5d). Thus, alternative designs for anti-peptide antibodies based on inserting or deleting residues at H3 could be envisioned.

Worth mentioning is that 3 anti-protein antibodies have also a long H3 loop: 1opg has 17 at H3, 1dfb has 18 residues and 1gc1 has the longest H3 of all the structures: 20 residues (see Table1). Such a long H3 loops in anti-protein structures, however, has different conformations than in the anti-peptide antibody. In anti-protein structures, H3 is packed against the antigen binding site (Figure 6), whereas in the anti-peptide antibody, 1ikf, the long H3 has an open conformation (compare Figure 5a and 6). No anti-protein antibody has a H3 loop as short as 1bog, among the structures analyzed. The difference in conformation of H3 depending upon if the antibody bind a protein or a peptide suggest another clue for designing anti-protein or anti-peptide antibodies based on H3.

Anti-hapten antibodies

Changes in the length of L1 appear to be irrelevant for the recognition of haptens. The superposition of anti-hapten structures with a short L1 (Figure 7a) and those with a long L1 (Figure 7b) indicates as a distinguishing feature of the hapten recognition that the ligand is deeply buried in the antigen-binding site. Compare, for instance, the recognition of haptens and peptides (Figures 2b and 7b).

To confirm that observation we determined the by profile of positions in contact with the antigens for each sample. As can be see, the residues in contact with haptens are mainly located at the stems of the loops, in the framework. In contrast proteins and peptides establish contacts mainly at the middle of the loop.

Discussion

Differences in antigen-binding site topography correlate with the propensity of antibodies to rēcognize different kind of antigens (Bolger and Sherman, 1991; Wilson et al., 1991; Webster et al., 1994; Vargas-Madrazo et al., 1995; MacCallum et al., 1996; Lara-Ochoa et al., 1996). In previous studies we found that certain lengths in L1 and H2 imply flatted or grooved antigen-binding sites, thus finding a connection between pattern in the amino acid sequence (understood as the hypervariable loop length) and the propensity of antibodies to bind large molecules like proteins or small ones like peptides or haptens (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996). The number and diversity of antibodies of known three-dimensional structure has increased largely over the last years. This information allowed in this paper a refinement of the above findings. Also we have find new features of the antibody interaction that allows a better understanding of the mechanism of binding of different kind of molecules.

Anti-protein and anti-peptide antibodies and the structural basis for the protein/peptide cross-reactivity

Consistent with previous reports, we found that antibodies with a short L1 have a flattened antigen-binding site. This topography allows large antigens, like proteins, extend all along the antigen-binding site surface establishing a good antigen-antibody complementarity. No flat antibody recognizes a peptide, within the structures analyzed. Inability of flat antibodies to bind peptides may be related to the insufficiency of a flattened antigen-binding site to provide complementarity to peptides. Actually, all anti-peptide antibodies had a grooved antigen-binding site that ensured a proper peptide-antibody complementarity. This groove is built in 90% of the cases by inserting residues at the tip of L1.

The protrusion at L1 might also prevent interaction of the antibody molecule with a flat surface in the antigen. This is suggested by the structures that bind proteins but have a long L1. About one-third of the anti-protein antibodies (16 out of 49 structures) belong to this category. Four of these have been co-crystallized with their respective antigens and two evidenced a poor antigen-antibody complementarity, with no interaction at the center of the grooved antigen-binding site.

A third counterexample, indicated that if the epitope protrudes from the antigen, then it could fill the groove. From this, it would be proposed that if such a protrusion has a similar conformation in the context of the protein as well as when being bind as a peptide to the antibody molecule, then these antibodies might be able to bind both the free peptide as well as the peptide in the context of the cognate protein.

No direct experimental evidence supporting this view is available. However, indirect evidences have been obtained from the analysis of the antibody VP2(2156-2170)/8F5 (Tormo et al., 1994). This antibody was originally raised against the viral capside protein VP2, but was also found to cross-react with a synthetic peptide derived from the same protein (residues: 2156-2170). The conformation adopted by the peptide is very similar to and can be superimposed onto the corresponding region of the viral protein VP2 of human rhinovirus 1A (HRV1A) whose three-dimensional structure is known (Al-Lazikani et al., 1997).

The suggestion that anti-peptide-like antibodies, namely antibodies with a long L1, recognize protruding epitopes in the native structure of proteins, together with the fact that this kind of antibodies are the only ones that bind peptides, suggest them as the candidates for the target of deeper research in order to decipher the molecular basis of peptide/protein cross-reactivity.

Anti-hapten antibodies

An important finding of the present work is that the length of L1, or the length of any other loop, is not related to the recognition of haptens. The distinguishing feature of hapten recognition is the deep burial of the hapten in the antigen-binding site. Haptens, due to their size, could be fitted into the antigen-binding site with no large changes in the antigen-binding site topography, as those achieved by insertion or deletion of residues at the tip of L1.

Antibodies would create the complementarity with the hapten via conformational changes and/or mutation of side chain replacement during the affinity maturation process. For example, in the anti-steroid antibody DB3 the steroid-binding pocket have two conformations, open and closed (Arevalo et al., 1993). In the free structure, the bulky side chain of the Tryptophan at position 100 of V_H has an 'close' conformation, occupying the hapten-binding site as an surrogate ligand. In the steroid-DB3 complex, the side of the Trytophan has an 'open' conformation, creating the hapten-binding pocket (Arevalo et al., 1993).

Other example of conformational changes to accommodate a hapten is the catalytic antibody CNJ206 (Charbonier et al., 1995). The free form of CNJ206 has the phenil ring of Tyrosine 101 in V_H preventing the formation of a deep pocket in the hapten-combining site (Charbonier et al., 1995). This orientation of the phenil ring changes in the complex, enlarging the cavity where the hapten is bound (Charbonier et al., 1995).

Regarding changes during the affinity maturation, it have been showed that early variants of the anti-NP antibody 88C6/12 (Yuhasz et al., 1995), which are close to the germline gene configuration, have Tryptophan at position 33 in the VH chain. This residue allows a favorable interaction with the hapten via the indole group. However, the interaction appears to be suboptimal since the bulky nature of the Tryptophan prohibits a deep burial of the hapten within a cavity (Yuhasz et al., 1995). Such constrains is relieved by substituting a Tryptophan by Leucine while during the affinity maturation process (Yuhasz et al., 1995).

Toward the design of smart repertoires of antibodies

Recently, the vast majority of the V germline genes of man and mice have been sequenced (http://www.med.uni-muenchen.de/biochemie/zachau/kappa.htm), thus allowing to address the question about how many genes encode anti-peptide-like or anti-protein antibodies in those species. There are about 92 functional or potentially functional Vκ germline genes[9] and 3 functional Vλ germline genes in mice.[1] 63 genes (68%) encode short L1loops and 29 (32%) encode long ones (see http://www.med.uni-muenchen.de/biochemie/zachau/kappa.htm). This suggest, in turn, that the probability to obtain an anti-peptide-like antibody, assuming that the immune response is a random sample of the V genes, is about one-third, which is not too high. However, based in our findings repertoires biased toward the specific recognition of peptides can be built to increase the probability of obtaining anti-peptide-like antibodies.

References

Al-Lazikani B., Lesk A.M., and Chothia C. 1997. Standard conformations for the canonical structures of immunoglobulins. J. Mol. Biol. 273: 927-948.

Arevalo, J, H., Stura, E.A., Taussin, M. J. & Wilson, I.A. 1993. Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex. J.Mol.Biol. 231, 103-118.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne, PE. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235-242

Bolger, M.B., Sherman, M.A. 1991. Computer modeling of combining site structure of anti-hapten monoclonal antibodies. Methods Enzymol. 203: 21-45.

Charbonnier, J.B., Carpenter, E., Gigant, B., Golinelli-Pinpaneau, B., Eshhar, Z., Green, B.S., Knossow, M. 1995. Crystal structure of the complex of a catalytic antibody Fab fragment with a transition state analog: structural similarities in esterase-like catalytic antibodies. Proc. Natl. Acad. Sci. USA. 92: 11721

Chothia, C., Lesk, A. M. 1987. Canonical structures for the hipervariable regions of immunoglobulin. J.Mol. Biol. 196 :901-917

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. 1989. Conformations of immunoglobulins hypervariable regions. Nature (London). 342 : 877

Fleury, D., Barrere, B., Bizebard, T., Daniels, R.S., Skehel, J.J., Knossow, M. 1999. A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. Nat. Struct. Biol. 6:530.

Johnson, G. and Wu, T.T. 2000. Kabat Database and its applications: 30 years after the first variability plot. Nucleic Acids Research., 28, 214-218.

Johnson, G. and Wu, T.T. 2001. Kabat Database and its applications: future directions. Nucleic Acids Res. 29:205-6.

Lara-Ochoa, F., Almagro, J.C., Vargas-Madrazo, E. and Conrad, M. 1996. Antibody-antigen recognition: a canonical structure paradigm. *J. Mol. Evol.* 43: 678-684.

MacCallum, R.M., Martin, A.C. and Thornton, J.M. 1996. Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* 262: 732-745.

Momany, C.., Kovary, L.C., Prongay, A.J., Keller, W., Gittı, R.K., Lee, B.M., Gorbalenya, A.E., Tong, L., McClure, J., Ehrlich, L.S., Summers, M.F., Carter, C., Rossmann, M.G. 1996. Crystal structure of dimeric HIV-1 capsid protein. Nat. Struct. Biol. 3:763.

Tormo, J., Blaas, D., Parry, N. R., Rowlands, D., Stuart, D. and Fita, I. 1994. Crystal structure of a human rhinovirus neutralizing antibody complexed with a peptide derived from viral capsid protein VP2. *EMBO J.* 13: 2247-2256.

Vargas-Madrazo, E, Lara-Ochoa, F and Almagro, J.C. 1995. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J. Mol. Biol.* 254: 497-504.

Webster, D. M., Henry A. H. and Rees, A. R. 1994. Antibody-antigen interactions. *Curr. Op. Struct. Biol.* 4: 123-129.

Wilson, I. A., Rini, J. M., Fremont, D. H., Feiser, G. G. and Sture, E. A. 1991. X-ray crystalographic analysis of free and antigen-complexed Fab fragments to investigate structural basis of immune recognition. *Meth. Enzymol.* 203: 153-176.

Wilson, I.A., Ghiara, J.B., Stanfield., R.L. 1994. Structure of anti-peptide antibody complexes. Res. Immunol. 145:73-8

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 antibody complementary.

J.Exp.Med. 132:211

Yuhasz, S.C., Parry, C., Strand, M., Amzel, L.M. 1995. Structural analyisis of affinity maturation: the three-dimensional structures of complexes of an anti-nitrophenol antibody. Mol. Immunol. 32: 1143.

Figure Captions

Table 1. The hypervariable loop lengths (definition as in Chothia and Lesk, 1987).

Figure 1. C^{α} atoms superposition of the anti-protein antibodies having a short L1 (PDB codes: 1VFB, 1C08, 1BQL, 1JHL, 1FBI, 1MLC, 1DVF, 1NCA, 1NMB, 1IAI, 1IAI, 1OSP, 1AR1, 1WEJ, 1OAK, 1JRH, 1KB5, 1EO8, 1FJ1, 1C9R 1GC1, 1FOR, 1VGE, 1OPG, 12E8, 1A6T, 1BFO, 1CD5 and 1GHF).

Figure 2. C^{α} atoms superposition of the anti-peptide antibodies having a long L1 (PDB codes: 1A3R, 1IFH, 1GGI, 1TET, 1AI1, 1FRG, 1FPT, 2MPA, 2HRP, 2H1P, 1F58, 2IGF, 2AP2, 1QKZ, 1EJO, 1CU4, and 1MF2). **a**. Structures without peptides showing the groove. **b**. The structures with peptides (in orange) illustrating how the grove accommodates the peptides no matter their structure; L1 in red.

Figure 3. Connolly surface of the anti-peptide-like structures but binding proteins (a-c) as seen from the antigen perspective: **a.** 2JEL, **b.** 1NSN, **c.** 1AFV. The contacts of these structures with their corresponding antigens are shown in red and blue. **d.** Contact region of the typical anti-peptide (PDB code:1VFB) antibodies in the same view of figures a and b. The color red and blue represents contact (Hidrogens bonds and Van der Walls) in the complex. In all cases the residues in contact were calculated with the HB-plus[22] and visualized with Insight-II.

Figure 4. Solvent accessible surface of heavy atoms of the anti-protein antibody (PDB code: 1AFV). In red show the main portion of the epitope of immunodeficiency virus, which protrudes from the antigen filling the groove of the antigen binding site.

Figure 5. Ribbons representation of the anti-peptide antibodies having a short L1. **a.** 1IKF, **b.** 1BOG. In counterpart of the L1 short, the long H3 in ping form groove. In the second structure, the groove is the result of the a very short H3. **c** and **e.** Ribbons representation of the anti-peptide antibodies (PDB code: 1IKF and1BOG) having a short L1 superposed with a typical anti-peptide antibody (PDB code: 1IFH). In cyan antibodies having a short L1, in white the typical anti-peptide antibody, in red H3. **d.** A 90^0 rotation of the structures about the Y-axes with respect to a to show how the long H3 (blue) compensates the short L1 these structure have. Insertions at the L1 of the typical anti-peptide antibody are shown in red.

Figure 6. Ribbons representation of the anti-proteins antibodies showing the differents conformations of H3 (PDB code: 1DFB, 1GC1, 1OPG).

Figure 7. a Ribbons representation of the anti-haptens antibodies which have a large L1 in complex with haptens (1FLR, 1IGJ, 1DBB, 1YED, 1YEE, 1YEJ, 1HYX, 2PCP, 1A3L and 2MCP). **b**. Structures with haptens having a short L1 showing the groove (PDB code: 1BAF, 1EAP, 1KNO, 1NGP, 25C8, 1CF8, 1CT8, 1C12 and 1A0Q) . **b**. The structures with peptides (in orange) illustrating how the grove accommodates the peptides no matter their structure; L1 in red

Code	Name	Type	Species	Specificity	Status	Res (Å)	L1	L2	L3	H1	H2	H3	Reference
A. Anti-protein (49)													
1vfb	D1.3	κ	Mouse	Hen Egg White Lysozyme	Bound	1.80/0.185	7	3	6	7	5	9	Bhat, et al 1994
1c08	HyHel-10	κ	Mouse	Hen Egg White Lysozyme	Bound	2.30/0.235	7	3	6	7	5	6	Kondo., et al 1999
1bql	HyHel-5	κ	Mouse	Bobwhite Quail Lysozyme	Bound	2.60/0.191	6	3	5	7	6	8	Chacko, et al To be Published
1jhl	D11.15	κ	Mouse	Pheasant Egg Lysozyme	Bound	2.40/0.214	7	3	6	7	6	10	Chitarra, et al 1993
1fbi	F9.13.7	κ	Mouse	Guinea Fowl Lysozyme	Bound	3.00/0.190	7	3	6	7	6	14	Lescar, et al 1995
1mlc	D44.1	κ	Mouse	Hen Egg White lysozyme	Bound	2.10/0.184	7	3	6	7	6	8	Braden, et al 1994
1dvf	E5.2	κ	Mouse	D1.3 antibody	Bound	1.90/0.194	7	3	5	7	5	14	Braden, et al 1996
1nca	NC41	κ	Mouse	Neuraminidase (N9 Tern)	Bound	2.50/0.191	7	3	6	7	6	12	Tulip, et al 1992
1nmb	NC10	κ	Mouse	Neuraminidase N9	Bound	2.50/0.210	7	3	6	7	6	14	Malby, et al 1994
2jel	JEL42	κ	Mouse	Histidine-containing protein	Bound	2.50/0.210	12	3	6	7	6	10	Prasad, et al 1998
1nsn	N10	κ	Mouse	Staphylococcal nuclease	Bound	2.90/0.195	11	3	6	8	5	5	Bossart-Whitaker, et al 1995
1iai	730.1.4	κ	Mouse	409.5.3 antibody	Bound	2.90/0.210	7	3	6	7	6	13	Ban, et al 1994
1iai	409.5.3	κ	Mouse	730.1.4 antibody	Bound	2.90/0.210	8	3	6	7	8	11	Ban, et al 1994
1osp	184.1	κ	Mouse	Outer Surface Protein A	Bound	1.95/0.229	7	3	6	7	5	13	Li, et al 1997
1afv	25.3	κ	Mouse	HIV-1 Capsid Protein	Bound	3.70/0.217	11	3	6	7	6	12	Momany, et al 1996
1ar1	7E2	κ	Mouse	Cytochrome *c* oxidase	Bound	2.70/0.207	7	3	7	7	6	10	Ostermeier, et al 1997
1wej	E8	κ	Mouse	Horse cytochrome *c*	Bound	1.80/0.200	7	3	6	7	6	9	Mylvaganam, et al 1998
1oak	NMC-4	κ	Mouse	Von Willebrand Factor A1 domain	Bound	2.20/0.201	7	3	6	7	5	15	Celikel, et al 1998
1jrh	A6	κ	Mouse	Interferon-gamma receptor	Bound	2.80/0.246	7	3	5	9	5	13	Sogabe, et al

				alpha chain									1997
1kb5	Désiré-1	κ	Mouse	Kb5-C20 T-Cell Antigen Receptor	Bound	2.50/0.219	7	3	6	7	6	11	Housset, et al 1997
1bln	Mrk-16	κ	Mouse	P-Glycoprotein	Bound	2.80/0.209	12	3	3	7	6	11	Vasudevan, et al 1996
1eo8	Bh151	κ	Mouse	Hemagglutinin (HA1 Chain)	Bound	2.80/0.196	6	3	6	7	6	10	Fleury, et al 2000
1fj1	La2	κ	Mouse	Outer Surface Protein A	Bound	2.68/0.226	7	3	6	7	6	5	Ding, et al 2000
1e6j	L:1-210/H:1-219	κ	Mouse	HIV-1 Capsid Protein (P24)	Bound	3.00/0.210	6	3	5	7	6	12	Monaco-Malbet, et al 2000
1c9r	28	κ	Mouse	HIV-1 Reverse Transcriptase	Bound	3.50/0.262	7	3	6	9	5	14	Sarafianos, et al 1999
1qfu	HC45	κ	Mouse	Influenza Hemagglutinin (HA)	Bound	2.80/0.198	12	3	6	7	6	14	Fleury, et al 1999
1for	17-IA	κ	Mouse	Human Rhinovirus	Free	2.75/0.174	6	3	6	7	6	11	Liu, et al 1994
1opg	OPG2	κ	Mouse	Receptor (Platelet Glycoprotein)	Free	2.00/0.160	7	3	6	7	6	17	Kodandapani, et al 1995
1dqj	HyHel-63	κ	Mouse	Hen Egg White Lysozyme	Free	1.80/0.210	7	3	6	7	5	6	Li, et al 2000
1nld	1583	κ	Mouse	HIV-1, gp41	Free	2.90/0.198	12	3	6	7	5	6	Davies, et al 1997
1ghf	GH1002	κ	Mouse		Free	2.70/0.210	7	3	6	7	6	8	Ban, et al 1996
1ap2	C219	κ	Mouse		Free	2.40/0.204	13	3	6	7	6	13	Hoedemaeker, et al 1997
1a6t	1-IA	κ	Mouse	Human Rhinovirus 14	Free	2.70/0.169	6	3	6	7	6	9	Che, et al 1998
12e8	2E8	κ	Mouse	Low density lipoprotein	Free	1.90/0.221	7	3	6	7	6	12	Trakhanov, et al 1999
1a5f	7A9	κ	Mouse	E- Selectin	Free	2.80/0.195	13	3	6	7	6	12	Rodriguez-Romero, et al 1998
1ay1	TP7	κ	Mouse	Thermus aquaticus DNA polymerase	Free	2.20/0.196	6	3	6	8	5	11	Murali, et al 1998
1cl7	1696	κ	Mouse	HIV-1 protease	Free	3.00/0.183	12	3	6	7	6	12	Lescar, et al 1999
1cr9	3F4	κ	Mouse	Syrian hamster prion protein (SHaPrP)	Free	2.00/0.171	12	3	6	7	6	6	Kanyo, et al 1999
32c2	32C2	κ	Mouse	Cytochrome P450 Aromatase	Free	3.00/0.213	11	3	5	8	5	11	Sawicki, et al 1999
1bbd	8F5	κ	Murin	Human Rhinovirus Serotype	Free	2.80/0.190	13	3	6	7	6	10	Tormo, et al

			e	2								1992
1c5d	192	κ	Rat	Human Muscle Acetylcholine Receptor	Free	2.40/0.196	7	3	5	7	5	10 Kontou, et al 2000
1bfo	CAMPATH-1G	κ	Rat	CD52	Free	2.60/0.192	7	3	6	7	8	11 Cheetham, et al 1998
1gc1	17B	κ	Human	CD4	Bound	2.50/0.210	7	3	8	7	6	20 Kwong, et al 1998
1dfb	3D6	κ	Human	HIV-1, gp41	Free	2.70/0.177	7	3	4	7	6	18 He, et al 1992
1vge	TR1.9	κ	Human	Thyroid Peroxidase	Free	2.00/0.180	7	3	6	7	6	13 Chacko, et al 1996
1nfd	H57	λ	Mouse	N15 T-cell receptor	Bound	2.80/0.243	8	3	8	7	8	10 Wang, et al 1998
2vir	HC19	λ	Mouse	Influenza Virus Hemagglutinin	Bound	3.25/0.198	11	3	6	7	5	15 Fleury, et al 1998
1adq	Igg4 Rea	λ	Human	Rf-An Igm/lambda	Bound	3.15/0.225	8	3	8	7	6	15 Corper, et al 1997
1aqk	B7-15A2	λ	Human	Tetanus Toxoid	Free	1.84/0.185	11	3	7	7	6	15 Faber, et al 1998

B. Anti-peptide (20)

2igf	B13I2	κ	Mouse	Myohemerythrin; residues 69-87	Bound	2.80/0.220	12	3	6	7	6	11 Stanfield, et al 1990
1ggi	50.1	κ	Mouse	HIV-1 gp120; residues: 311-328	Bound	2.80/0.188	11	3	6	9	5	6 Rini, et al 1993
1tet	TE33	κ	Mouse	Cholera Toxin; residues: 50-64	Bound	2.30/0.148	12	3	6	7	6	8 Shoham, et al 1993
1frg	26/9	κ	Mouse	Influenza HA; residues: 101-108	Bound	2.80/0.190	13	3	6	7	6	12 Churchill, et al 1994
1ai1	59.1	κ	Mouse	HIV-1; V3 Loop	Bound	2.80/0.220	11	3	6	9	5	12 Ghiara, et al 1997
1ikf	R454511	κ	Mouse	Cyclosporin A	Bound	2.50/0.164	7	3	6	7	6	17 Altschuh, et al 1992
1fpt	C3	κ	Mouse	Poliovirus Type 1; residues: 86 - 103	Bound	3.00/0.230	12	3	6	7	6	12 Wien, et al 1995
2ap2	C21	κ	Mouse	α-Helical Epitope on P-Glycoprotein	Bound	2.40/0.221	13	3	6	7	6	13 Vandenelsen, et al 1999
2mpa	Mn12H2	κ	Mouse	*N. meningitidis*; residues: 180-187	Bound	2.60/0.202	12	3	6	7	6	13 Van Den Elsen, et al 1997
2hrp	F11.2.32	κ	Mouse	HIV-1 Protease; residues:	Bound	2.20/0.198	11	3	6	7	6	17 Lescar, et al

				36-46									1997
2h1p	2H1	κ	Mouse	*C. neoformans;* residues: 1-12	Bound	2.40/0.186	12	3	6	7	6	12	Young, et al 1997
1a3r	8F5	κ	Mouse	Rhinovirus Vp2; residues: 156-170	Bound	2.10/0.171	13	3	6	7	6	10	Tormo, et al 1994
1f58	58.2	κ	Mouse	HIV-1 gp120; residues: 308-333	Bound	2.00/0.196	11	3	6	6	7	18	Stanfield, et al 1999
1qkz	MN14C11.6	κ	Mouse	porin PorA from *N. meningitidis*	Bound	1.95/0.209	12	3	5	7	6	11	Derrick, et al 1999
1ero	4C4	κ	Mouse	Foot-and-Mouth Disease Virus	Bound	2.30/0.248	11	3	6	7	6	12	Ochoa, et al 2000
1cu4	3F4	κ	Mouse	Syrian hamster prion protein (SHaPrP)	Bound	2.90/0.159	12	3	6	7	6	6	Kanyo, et al 1999
1ifh	17/9	κ	Mouse	Influenza HA; residues: 75-110	Bound	2.80/0.170	13	3	6	7	6	12	Schulze-Gahmen, et al 1993
1bog	Cb 4-1	κ	Mouse	p24 (HIV-1)	Bound	2.60/0.246	7	3	6	7	6	4	Keitel, et al 1997
1sps	3A2	κ	Mouse	C-Terminal Peptide of Hcg	Free	2.00/0.180	13	3	6	7	8	13	Fotinou, et al 1998
1sm3	Sm3	λ	Mouse	Mucin; residues: 1-13	Bound	1.95/0.213	11	3	6	7	8	7	Dokurno, et al 1998

C. Anti-hapten (35)

1flr	4-4-20	κ	Mouse	Fluorescein	Bound	1.85/0.188	12	3	6	7	8	8	Whitlow, et al 1995
1igj	26-10	κ	Mouse	Digoxin	Bound	2.50/0.176	12	3	6	7	6	11	Jeffrey, et al 1993
1dbb	DB3	κ	Mouse	Progesterone	Bound	2.70/0.210	12	3	6	7	6	11	Arevalo, et al 1993
1baf	AN02	κ	Mouse	2,2,6,6-Tetramethyl-1-Piperidinyloxy-Dinitrophenyl	Bound	2.90/0.195	6	3	7	8	6	7	Brunger, et al 1991
2cgr	NC6.8	κ	Mouse	N-(P-Cyanophenyl)-N'-(Diphenylemethyl) Guanidineacetic Acid	Bound	2.20/0.214	12	3	6	7	6	8	Guddat, et al 1994
1fig	1F7	κ	Mouse	1,2,Dicarboxy-4-Hydroxy-7-Oxa-Bicyclon-2-Ene	Bound	3.00/0.220	8	3	6	7	6	11	Haynes, et al 1994
1eap	17E8	κ	Mouse	Phenyl [1-(1-N-Succinylamino)Pentyl]	Bound	2.50/0.186	7	3	5	7	6	11	Zhou, et al 1994

Figura 1.

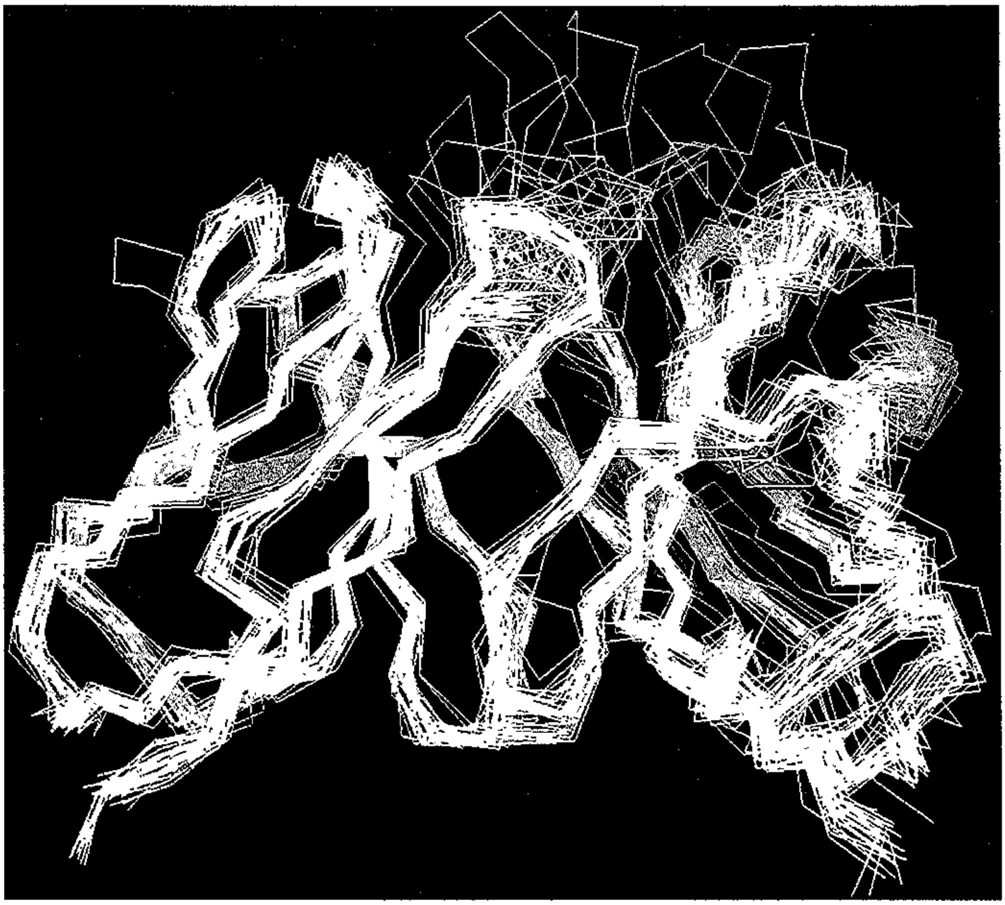


Figura 2.

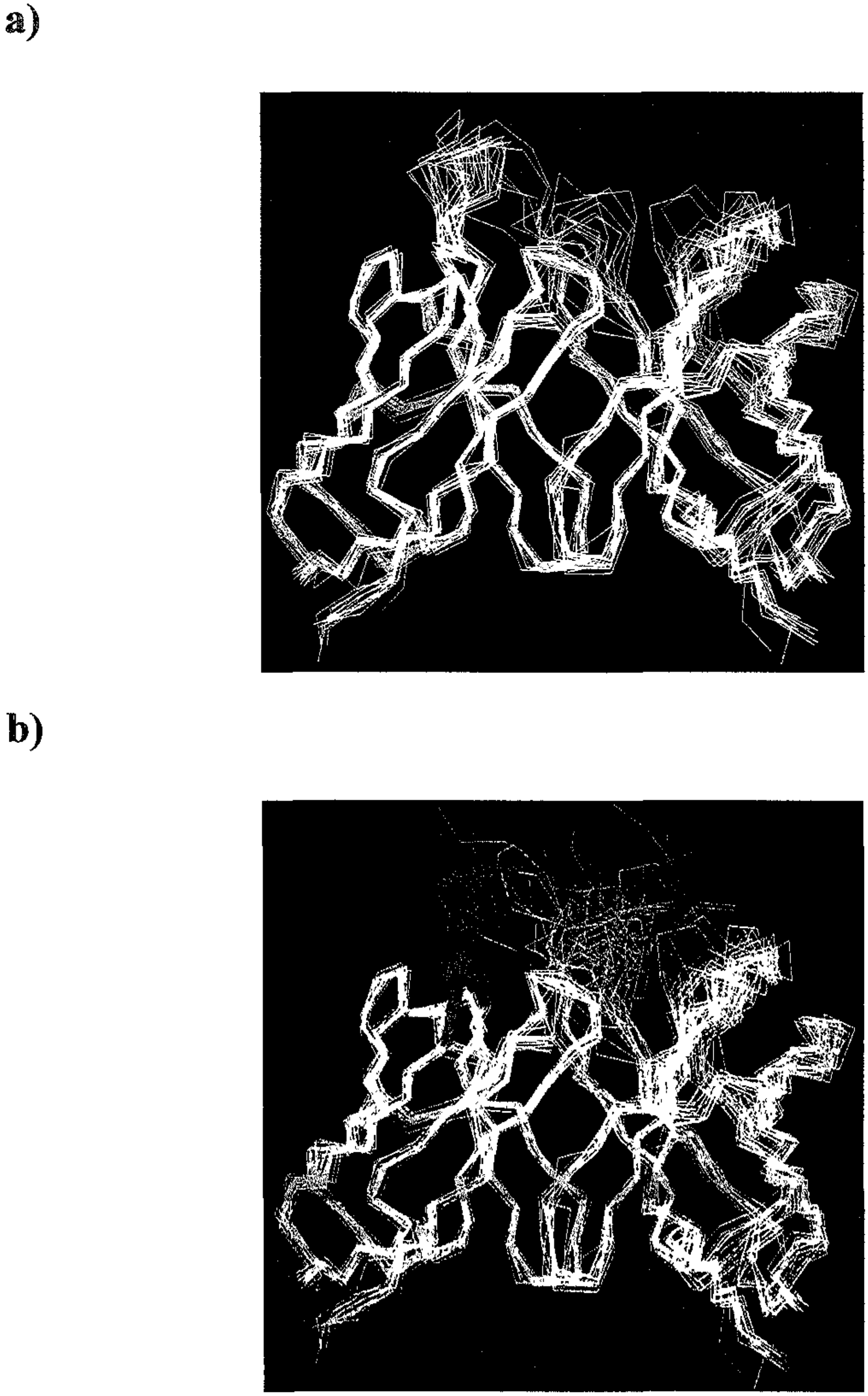


Figura 3.

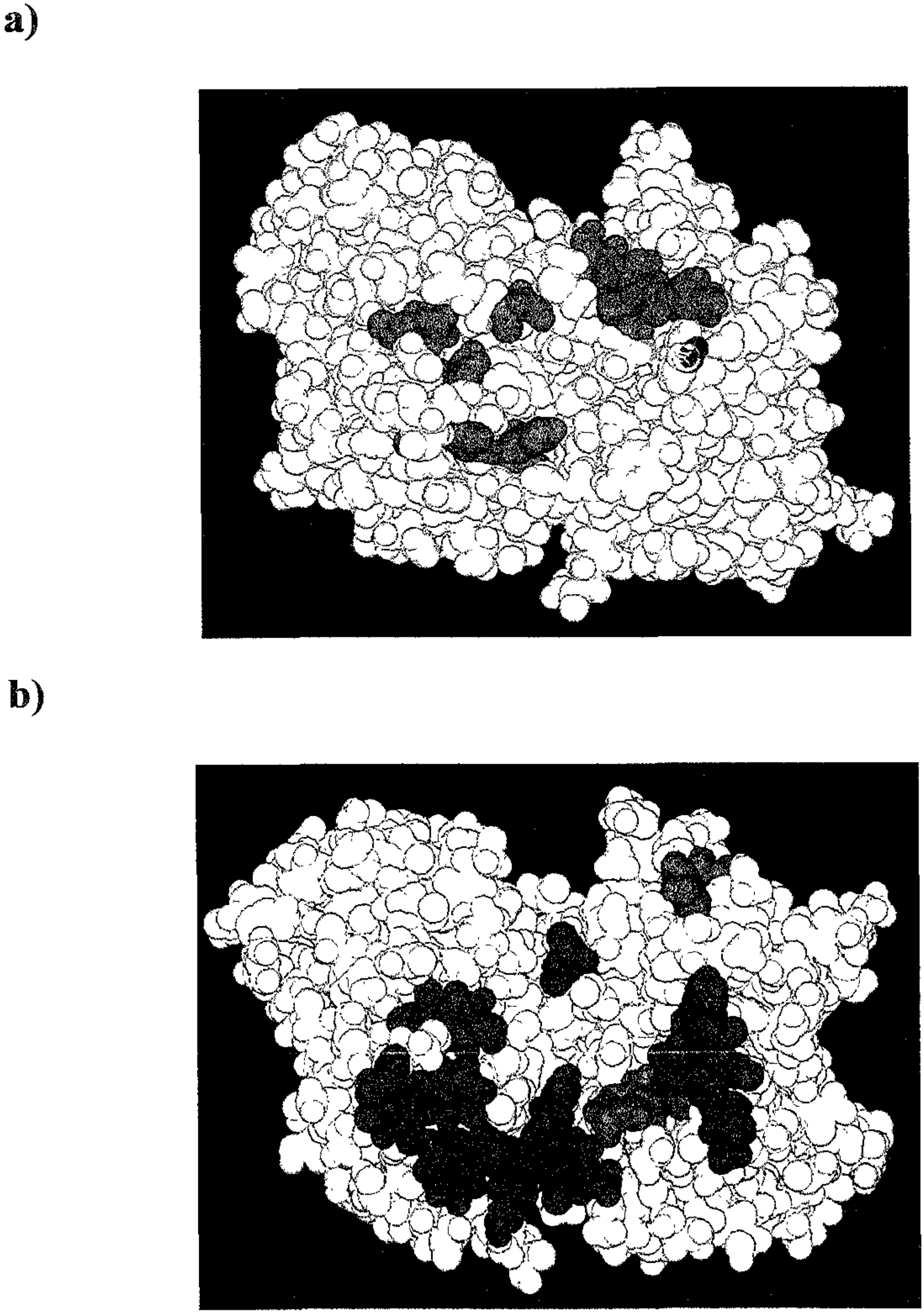


Figura 4.

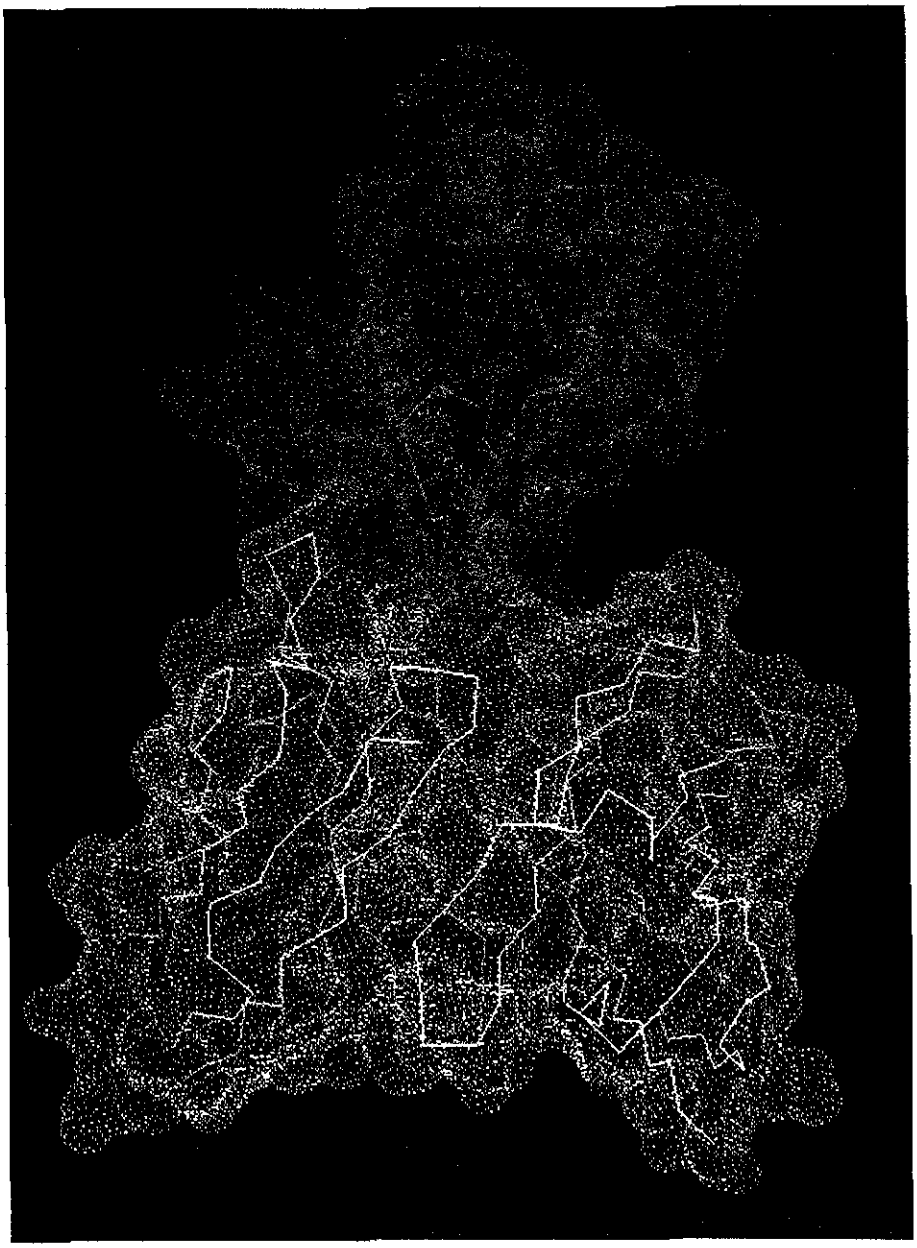
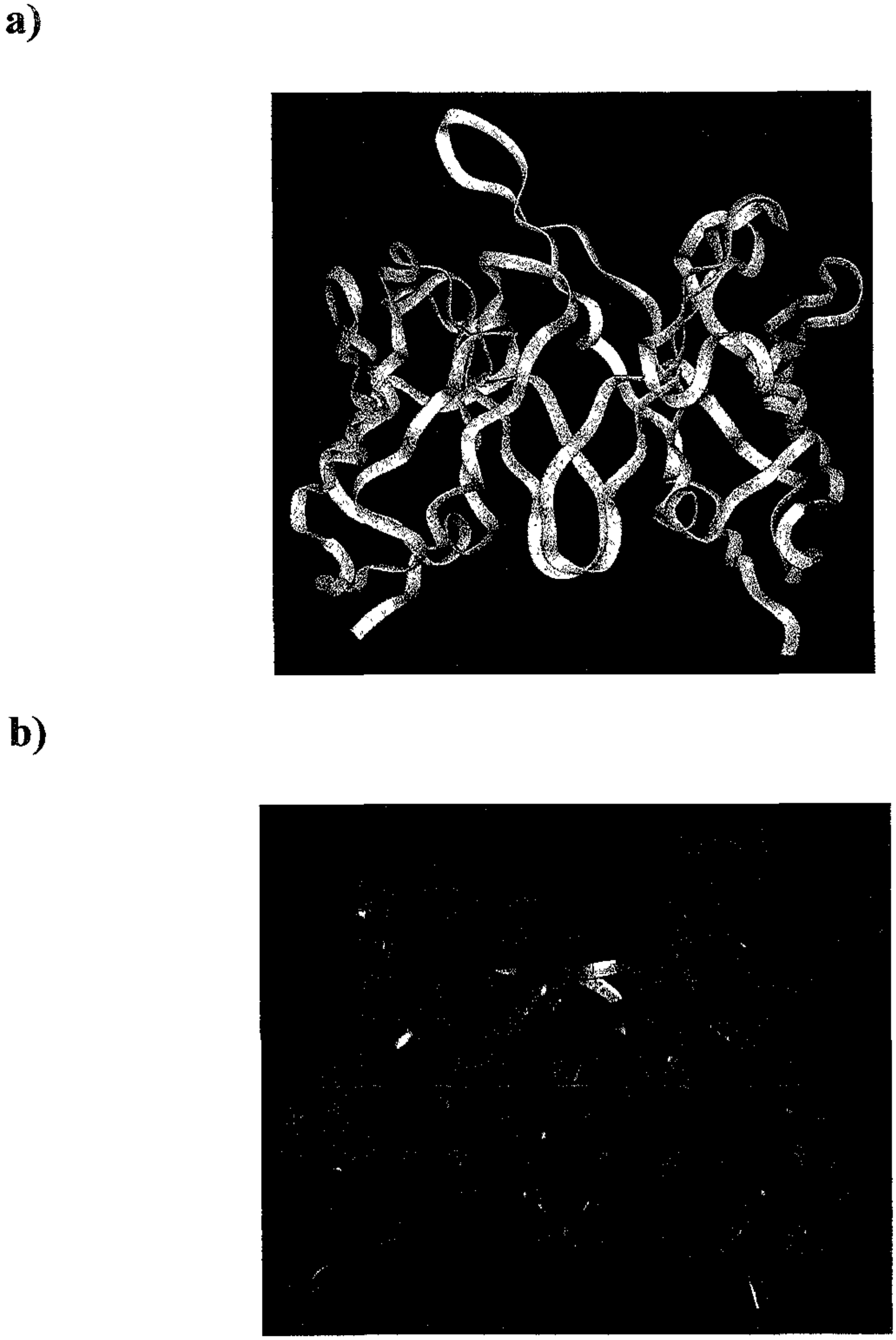


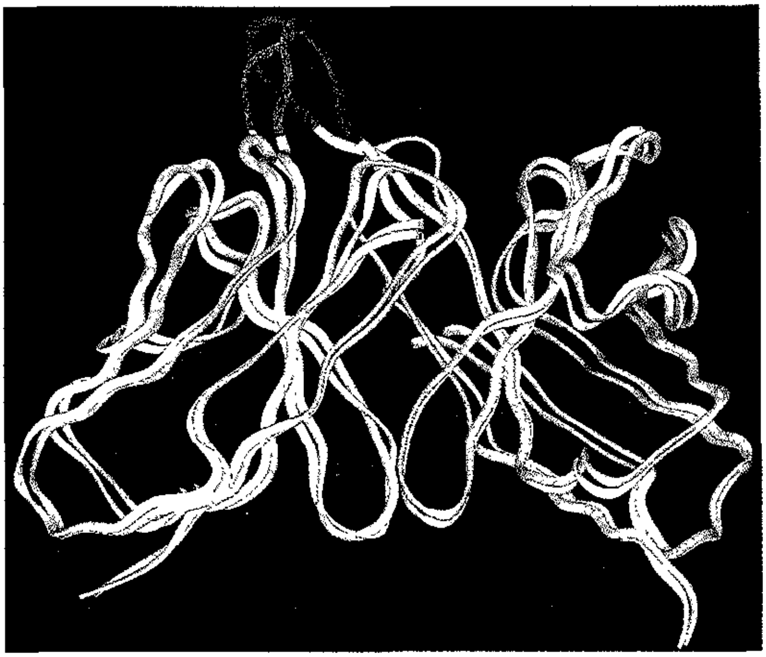
Figura 5.



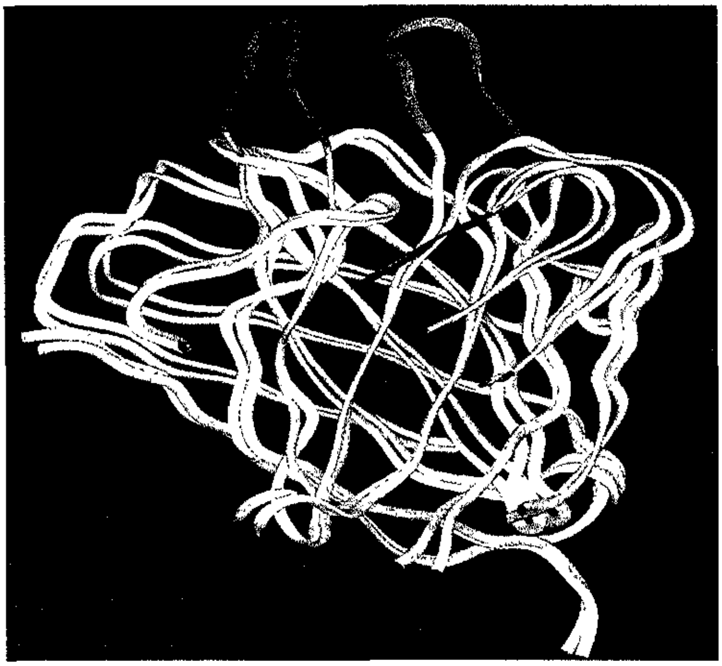
a)

b)

c)



d)



e)

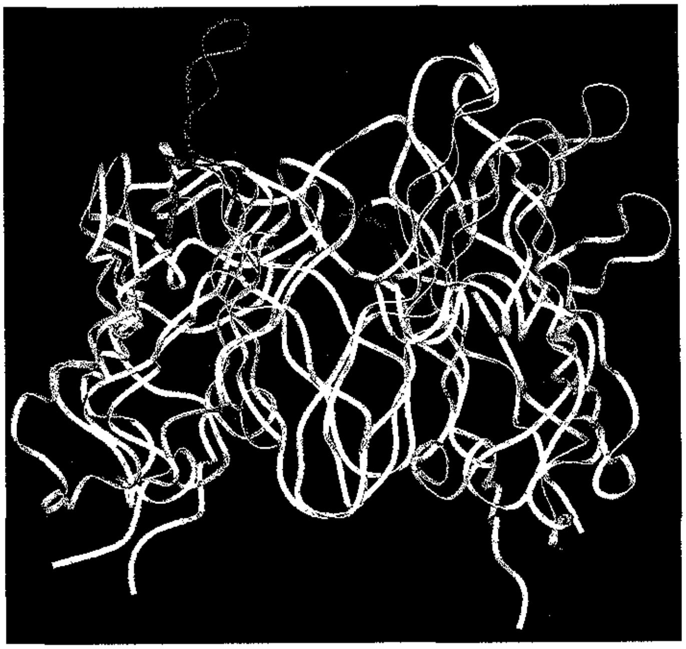


Figura 6.

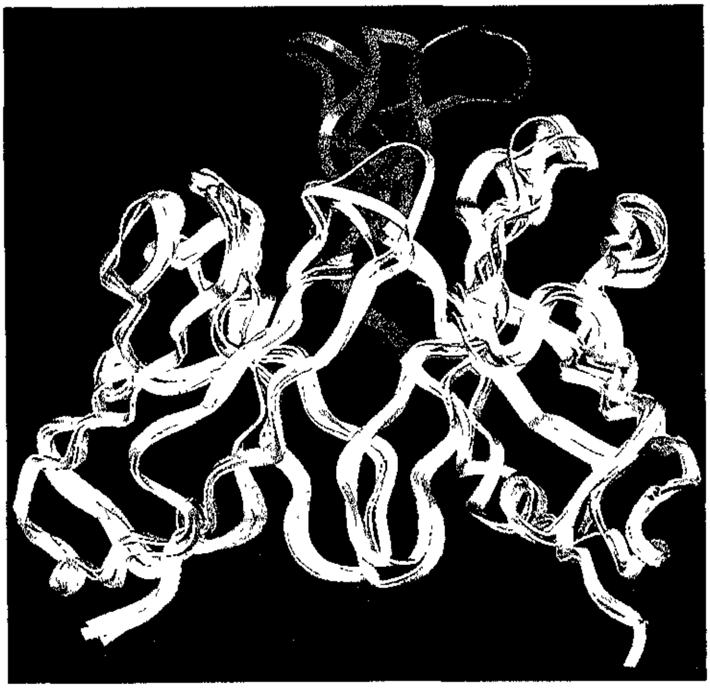
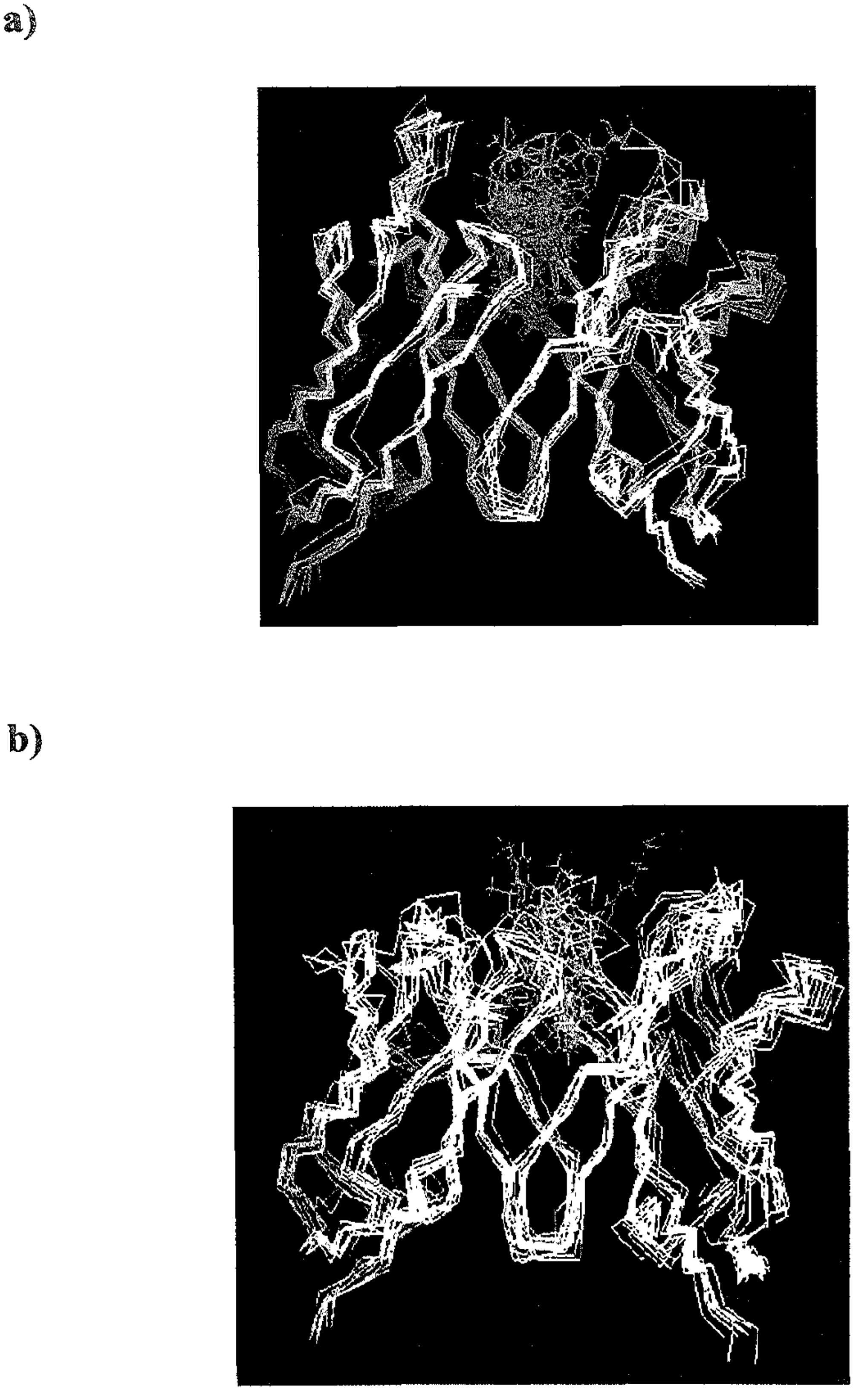


Figura 7.



Capítulo IV

Analysis of antibodies of known structure suggest a lack of correspondence between the residues in contac with the antigen and those modified by somatic hypermutation.

Analysis of Antibodies of Known Structure Suggests a Lack of Correspondence Between the Residues in Contact With the Antigen and Those Modified by Somatic Hypermutation

Maria del Carmen Ramirez-Benitez and Juan Carlos Almagro*
Instituto de Biotecnoligia, UNAM, Cuernavaca, Morelos, Mexico

ABSTRACT **Forty unique murine antibody–antigen complexes determined at 2.5 Å or less resolution are analyzed to determine whether the residues in direct contact with the antigen are modified by somatic hypermutation. This was done by taking advantage of the recent characterization of the pool of $V\kappa$ germline genes of the mouse. The average number of residues in contact with the antigen in the V_L gene, which contains the CDRL-1, CDRL-2, and all but one residue of CDRL-3, was six. The average number of somatic mutations was similar (around five). However, as many as 53% of the antibodies did not show somatic replacements of residues in contact with the antigen. Another 28% had only one. Overall, the frequency of antibodies with increasing number of somatic replacements in residues in contact with the antigen decreased exponentially. A possible explanation of this finding is that mutations in the contacting residues have an adverse effect on the antigen–antibody interaction. This implies that most of the observed mutations are those remaining after negative (purifying) selection. Therefore, efficient strategies of site-directed mutagenesis to improve the affinity of antibodies should be focused on residues other than those directly interacting with the antigen. Proteins 2001; 45:199–206.**

Key words: immunoglobulin; Ig; affinity maturation process; antigen–antibody interactions; site directed mutagenesis

INTRODUCTION

Antibodies constitute a paradigm of molecular recognition, recognizing, as they do, a seemingly unlimited number of antigens with exquisite specificity and high affinity. This ability is achieved in man and mouse through two steps. First, the combinatorial germline gene rearrangement produces a primary repertoire of antibodies.[1] Such a repertoire should be capable of recognizing any antigen with at least a low or medium affinity during the primary immune response.[2,3] Second, once the antigen is recognized, somatic hypermutation furnishes additional diversity to produce the raw material from where antibodies of improved affinity would be selected, as the secondary and tertiary immune responses proceed[2,3]

Comparisons of germline genes and antibody sequences product of immune responses indicate that, although somatic hypermutation occurs all along the V gene, it is more frequent at the antigen-binding site.[2,4,5] Mutational enrichment at the antigen-binding site has been explained in the presence of mutational hot and cold spots at the antigen-binding site and framework regions, respectively, as well as antigenic selection.[1–5]

The question that arises is whether somatic hypermutation modifies residues in direct contact with antigen, that is, those that were responsible for antigenic recognition during the primary immune response. Answers to this question could provide insight into the theories about the origin and evolution of antibody diversity.[3–7] In addition, it might be useful to define a conceptual framework to assist selecting positions to be altered by site-directed mutagenesis to improve the affinity, or even the specificity, of a given antibody. However, a systematic analysis of this subject has not been performed partially because human antibodies of known structure and crystallized in complex with the antigen are very scarce (see the URL: http://www.ibt.unam.mx/vir/structure/structures.html), and man is the only species in which all the germline V_L and V_H genes have been sequenced at present.[8–12]

The pool of murine germline V_L genes was recently characterized.[13] By taking advantage of this achievement, in this article we have analyzed the 44 different murine antibodies in complex with their respective antigens determined at 2.5 Å of resolution or less. That resolution allows the proper definition of the atomic interactions, particularly of the hydrogen bonds that require appropriate orientation of the atoms involved in the contact.[14] The contacting residues were then compared with the putative somatic mutations in the region coded by the V_L gene, which contain three of the six CDRs that conform the antigen-binding site: CDRL-1, CDRL-2, and all of CDRL-3 except one residue.[15–17] Somatic replacement of residues

Abbreviations: V_L, variable domain of the light chain; V_H, variable domain of the heavy chain; CDR, complementarity determining region, PDB, Brookhaven Protein Data Bank

Grant sponsor: DGAPA-UNAM, Grant number: IN223399, Grant sponsor: CONACyT, Grant number. N-33322

*Correspondence to: Juan Carlos Almagro, Instituto de Biotecnologia, UNAM, Av. Universidad # 2001, Col. Chamilpa C.P. 62210, Cuernavaca, Morelos, Mexico E-mail: almagro@ibt.unam.mx

Received 6 March 2001 Accepted 15 June 2001

TABLE I. Antibody-Antigen Complexes Used in This Study

PDBID[38]	Name	Type	Antigen	Resol. (Å)
1IND	CHA255	λ	4-[N′-(2-Hydroxyethyl)-Thioureido]-L-Benzyl-EDTA-In(3+) . . .	2.20
1MFA	SE155-4	λ	Trisaccharide	1.70
1SM3	Sm3	λ	Carcinoma-Associated Mucin; Residues 1–13	1.95
1NGP	N1G9	λ	(4-Hydroxy-3-Nitrophenyl) Acetate	2.40
1A0Q	29G11	κ	Phenyl[1-(1-N-Succinylamino)Pentyl] . . .	2.30
1A3L	13G5	κ	1-Carboxy-1′-[(Dimethylamino)-Carbonyl]Ferrocene	1.95
1A3R	8F5	κ	Human Rhinovirus Capsid Protein Vp2; Residues 156–170	2.10
1A4K	39-A11	κ	Bicyclo[2.2.2]Octene Derivative	2.40
1A6W	B1-8	κ	4-Hydroxy-5-Iodo-3-Nitrophenylacetyl-Epsilon-Aminocaproic . . .	1.80
1C1E	1E9	κ	Hexachloronorbornene Derivative 6 (Catalytic)	1.90
1CFV	FV4155	κ	Steroid	2.10
1CT8	7C8	κ	[4-(2,2,2-Trifluoro-Acetylamino)-Benzyl] . . .	2.20
1DQJ	HyHEL-63	κ	Hen Egg White Lysozyme	2.00
1DVF	E5.2	κ	D1.3 antibody	1.90
1EAP	17E9	κ	Phenyl [1-(1-N-Succinylamino)Pentyl] . . .	2.50
1EJO	4C4	κ	G-H Loop From Foot-And-Mouth Disease Virus	2.30
1F58	Fab 58.2	κ	HIV-1 Gp120; Residues 308–333	2.00
1FLR	4-4-20	κ	Fluorescein	1.85
1HYX	6D9	κ	[1-(3-Dimethylamino-Propyl)-3-Ethyl-Ureido] . . .	1.80
1IGJ	26-10	κ	Digoxin	2.50
1IKF	R454511	κ	Cyclosporin A	2.50
1JHL	D11.15	κ	Pheasant Egg Lysozyme	2.40
1KB5	Désiré-1	κ	Kb5-C20 T-Cell Antigen Receptor	2.50
1KEL	28B4	κ	1-[N-4′-Nitrobenzyl-N-4′-Carboxybutylamino] . . .	1.90
1MLC	D44.1	κ	Hen Egg White Lysozyme	2.10
1MRE	JEL103	κ	Guanosine-5′-Diphosphate	2.30
1NCA	Nc41	κ	N9 Neuraminidase	2.50
1NMB	Nc10	κ	N9 Neuraminidase	2.50
1OAK	NMC-4	κ	Domain 1 of Von Willebrand Factor	2.20
1OSP	184.1	κ	Outer Surface Protein A	1.95
1TET	TE33	κ	Cholera Toxin; Residues 50–64	2.30
1VFB	D1.3	κ	Hen Egg Write Lysozyme	1.80
1WEJ	E8	κ	Horse Cytochrome *c*	1.80
1YEE	D2.5	κ	4-Nitro-Benzylphosphonobutanoyl-Glycine	2.20
1YEJ	D2.3	κ	p-Nitrophenyl Phosphonate Glutaryl Caproate	1.85
25C8	5C8	κ	N-Methyl-N-(p-Glutaramidophenyl-Ethyl)-Piperidinium	2.00
2AP2	C21	κ	Alpha-Helical Epitope on p-Glycoprotein	2.40
2CGR	NC6.8	κ	N-(p-Cyanophenyl)-N′-(Diphenylemethyl) Guanidine acetic Acid	2.20
2FBJ	J539	κ	Galactan	1.95
2H1P	2H1	κ	Cryptococcus Neoformans; Residues 1–12	2.40
2HRP	F11.2.32	κ	HIV-1 Protease, Residues 36–46	2.20
2JEL	JEL42	κ	Histidine-Containing Protein	2.50
2PCP	6B5	κ	1-(1-Phenylcyclohexyl) Piperidine	2.20
43CA	43CA	κ	p-Nitrophenol	2.30

in direct contact with the antigen was found to be a rare event. Implications of this finding are discussed.

RESULTS

Antibody–Antigen Complexes and Identification of Residues in Contact

Table I lists the 44 antibody–antigen complexes currently available (as of September 1, 2000) determined at 2.5 Å of resolution or less (see the URL: http://www.ibt.unam.mx/vir/structure/structures.html). The structures, coming from mice, are mostly Vκ-type antibodies. Only 4 Vλ type antibodies were gathered.

The residues in the antibodies involved in direct contact with the antigen were identified via HB plus,[11] using the default parameters. In crystals with more than one molecule per asymmetric unit, the antigen–antibody contacts in all the molecules were taken into account in the calculations. Contacts established through water molecules were not considered. All the interactions were verified by visual inspection in Insight II (MSI, San Diego, CA).

The average number of residues in direct contact with the antigen in the V_L gene was found to be 6. The distribution of the contacts along the gene is shown in Figure 1(a). Contacts are more often established, let us say 0.4 or more in Figure 1(a), at positions 30a and 32 in the CDR-L1, 50 in the CDR-L2, and from positions 91 to 94 in the CDR-L3. From these, three positions are particularly

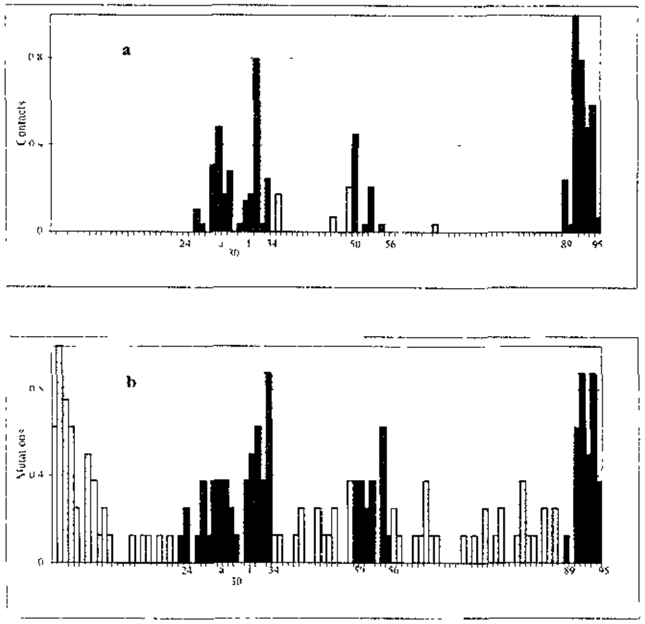


Fig 1 Frequency distribution of residues in contact with the antigen (a) and somatic hypermutation (b) along the Vκ gene. To make amenable the comparison between residues in contact and somatic hypermutations, the frequency is plotted in both cases by using a relative scale, calculated as 1-[(Pm-Pi)/Pm]; where Pm is the maximum value of contacts or mutations (30 and 8, respectively) and Pi is the number of contacts or mutations for the position "i." CDRs are represented by filled columns and are defined following the Kabat and Wu[15] conventions, that is, CDR-L1: positions 24–34, including the indels "a–f" at position 30; CDR-L2 positions 50–56 and CDR-L3: positions 89–95.

overused in establishing interactions (0.8 or more): positions 32, 91, and 92.

In 10 of the antigen–antibody complexes analyzed, the residues in contact with the antigen were previously identified.[18] A comparison to the previous identification indicates good agreement in both number and placement of the contacts. Discrepancies may be ascribed to slight differences in the definition of an atomic interaction, particularly a hydrogen bond.[14,18]

Assignment of Germline Genes

All the known functional and potentially functional genes of mice[13] were used to determine the putative somatic mutations in the V_L sequence of antibodies. About 90 functional or potentially functional Vκ germline genes[13] and 3 functional Vλ germline genes exits.[1] Probably 3–5 Vκ germline genes are still unknown.[13]

The formal translation product of the genes was compared with the antibody sequences, and the genes were sorted in decreasing order of identity for each antibody V_L sequence No established criterion defines whether the nearest germline V gene is the actual precursor of a given rearranged sequence. Klein and coworkers,[19] when working with rearranged human Vκ sequences at the nucleotide level, defined an assignment as unambiguous if the best match differed from the next one by at least 30%. Table II reports the PDB code of the antibodies, the names of the best and next matches, and the number of putative mutations for the antibody sequences and the difference between them, in percent.

The first four antibodies of Table II are the Vλ antibodies. The Vλ locus of mice, with low complexity, has been very well characterized.[1] Therefore, the Vλ antibodies provided a control to define whether the 30% criterion used by Klein and coworkers holds at the amino acid level and with murine sequences. Inspection of the Vλ antibodies indicated that this is the case. Thus, the 30% criterion was used to analyze the Vκ antibodies.

Thirty-six of the 40 Vκ antibodies fulfilled the 30% criterion. The four exceptions have 15 or more mutations for their nearest gene (in italics in Table II). This number of mutations exceeds the average number of mutations for the second choice when all antibodies are considered (see Table II). In addition, none of the Vλ antibodies have more than 9 putative mutations. Hence, it was assumed that the 4 Vκ antibodies with 15 or more mutations originated from unknown genes, and these were not further analyzed.

Putative Somatic Mutations

The average number of somatic mutations in the remaining 40 antibodies (4 Vλ and 36 Vκ) was 4.6 (Table III). This number is similar to the average number of somatic mutations in the human rearranged Vκ sequences.[4] The distribution of somatic mutations along the Vκ gene [Fig. 1(b)] is similar to the distribution of somatic mutations in the human counterpart,[4,5] mutations being concentrated in the N-terminal region and CDRs. On average, 52% of the somatic mutations are concentrated in CDRs. Mutational enrichment in the N-terminal region has been related to imprecise sequence determination,[4] whereas in CDRs it has been associated with antigenic selection.[2–5]

Some differences between germline genes and the antibody sequences may also be due to allelic variation. However, that artifact should be minimal because the predominant source of DNA used to derive the repertoire of the germline Vκ genes comes from the inbred strain C57BL/c[13], and many of the monoclonal antibodies listed in Table I were raised in BALB/c. Both strains belong to the haplotype c, where differences in the Vκ gene sequences of the inbred strains are small.[13]

Putative somatic mutations occurred with a frequency of 0.4 or more in Figure 1(b) at positions 31, 32, and 34 of the CDR-L1, 55 of the CDR-L2, and from positions 91–94 within CDR-L3. Very often mutated positions (≥0.8) are 34, 92, and 94. As can be seen, only residue 32 at CDR-L1 and residues within the CDR-L3 are frequently found in both contacts [Fig. 1(a)] and somatic mutations [Fig. 1(b)]. If positions very often used in contacts as well as mutated are the only ones considered, just residue 92 is found in both. This suggests a lack of correspondence between the residues in contact with the antigen and those modified by somatic hypermutation.

Residues in Contact With the Antigen and Somatic Replacements

Table III lists, for each antibody, the number of putative somatic mutations, the number of residues in contact with

TABLE II. Assignment of Germline V_L Genes to Antibodies of Known Structure, Based on Smallest Number of Putative Mutations

		Best match		Next best match		
PDB ID	Name	Gene	Mutations	Gene	Mutations	Diff. (%)
1IND	CHA255	11	2	12	9	78
1NGP	N1G9	11	9	12	17	47
1SM3	Sm3	11	0	12	8	100
1MFA	SE155-4	11	2	12	10	80
1A0Q	29G11	Gj38c	9	23-48	28	68
1A3L	13G5	He24	9	hf24	20	55
1A3R	8F5	8-19	9	8-21	13	31
1A4K	39-A11	Bb1	4	b11	9	56
1A6W	*B1-8*	*23-48*	*44*	*am4*	*55*	*20*
1C1E	1E9	Bb1	8	cr1	13	38
1CFV	*FV4155*	*bb1*	*15*	*cr1*	*19*	*21*
1CT8	*7C8*	*23-43*	*15*	*23-45*	*16*	*6*
1DQJ	HyHEL-63	23-43	0	23-45	2	100
1DVF	E5.2	Ce9	3	cp9	11	73
1EAP	17E9	Gj38c	10	23-48	28	64
1EJO	*4C4*	*21-5*	*20*	*21-10*	*22*	*9*
1F58	Fab 58.2	21-4	8	21-3	13	38
1FLR	4-4-20	Bb1	2	cr1	8	75
1HYX	6D9	Cr1	7	bb1	11	36
1IGJ	26-10	Bb1	4	cr1	9	56
1IKF	R454511	Ce9	6	cp9	11	45
1JHL	D11.15	Rf	7	23-48	29	76
1KB5	Désiré-1	12-44	2	12-41	9	78
1KEL	28B4	Cr1	2	bb1	9	78
1MLC	D44.1	23-43	3	23-45	5	40
1MRE	JEL103	Bb1	0	cr1	7	100
1NCA	NC41	19-25	6	19-17	12	50
1NMB	NC10	Ce9	5	cp9	10	50
1OAK	NMC-4	Cp9	8	ce9	13	38
1OSP	184.1	Gm33	6	gn33	13	54
1TET	TE33	Cr1	4	bb1	11	64
1VFB	D1.3	12-41	5	12-46	12	58
1WEJ	E8	12-41	0	12-44	8	100
1YEE	D2.5	Bj2	5	bd2	8	38
1YEJ	D2.3	Bj2	5	bd2	8	38
25C8	5C8	Ai4	2	ad4	12	83
2AP2	C21	8-19	0	8-28	5	100
2CGR	NC6 8	Bb1	7	cr1	12	42
2FBJ	J539	Kb4	5	kk4	18	72
2H1P	2H1	Bb1	3	cr1	10	70
2HRP	F11.2.32	21-2	5	21-1	13	62
2JEL	JEL42	Cr1	2	bb1	9	78
2PCP	6B5	Cr1	4	bb1	9	56
43CA	43CA	8-24	6	8-30	17	65
	Average		6.3		13.4	
	Std. Dev.		7.2		8.7	

the antigen, and the fraction of residues in contact that have been replaced by somatic mutation. As many as 21 of 40 antibodies (53%) do not show any somatic mutation in residues in contact with antigen, and another 11 (28%) have only 1. Overall, the frequency of antibodies with increasing number of replacements in the residues in contact with the antigen decreases following an exponential distribution (Fig. 2). Clearly, replacement of residues in contact with the antigen by somatic hypermutation constitutes a rare event

The number of residues in contact with the antigen that have been replaced by somatic mutation is not proportional to either the number of residues in contact with the antigen or the extent of somatic mutations. For instance, antibody 1IKF has 5 contacts with the antigen, of which 2 were replaced by a somatic mutation. In contrast, antibody 1YEJ has 11 residues in contact with the antigen of which 1 has been replaced. Again, antibody 1F58 has 8 mutations and 4 mutations are in residues in contact with the antigen, whereas antibody 1OAK has the same number of mutations but none in contacting residues.

Even more, no correlation between the number of somatic mutations in the CDRs and the number of replace-

TABLE III. Number of Residues in Contacts With the Antigen, Number of Putative Somatic Mutations, and Fraction of Residues in Contact With the Antigen That Were Replaced by Somatic Mutation (Fraction M-C)

PDB ID	Name	Contacts	Mutations	Fraction M-C
1IND	CHA255	3	2	0
1NGP	N1G9	3	0	0
1SM3	Sm3	5	2	0
1MFA	SE155-4	4	9	2
1A0Q	29G11	4	9	1
1A3L	13G5	4	9	0
1A3R	8F5	9	9	3
1A4K	39-A11	0	4	0
1C1E	1E9	5	8	1
1DQJ	HyHEL-63	9	0	0
1DVF	E5.2	5	3	0
1EAP	17E9	5	10	2
1F58	Fab 58.2	9	8	4
1FLR	4-4-20	4	2	1
1HYX	6D9	7	7	0
1IGJ	26-10	2	4	1
1IKF	R454511	5	6	2
1JHL	D11.15	5	7	0
1KB5	Désiré-1	8	2	1
1KEL	28B4	7	2	0
1MLC	D44.1	5	3	0
1MRE	JEL103	7	0	0
1NCA	NC41	7	6	1
1NMB	NC10	7	5	2
1OAK	NMC-4	5	8	0
1OSP	184.1	7	6	2
1TET	TE33	9	4	1
1VFB	D1.3	8	5	1
1WEJ	E8	8	0	0
1YEE	D2.5	7	5	0
1YEJ	D2 3	11	5	1
25C8	5C8	6	2	0
2AP2	C21	6	0	0
2CGR	NC6.8	5	7	1
2FBJ	J539	2	5	0
2H1P	2H1	8	3	0
2HRP	F11.2.32	9	5	2
2JEL	JEL42	7	2	1
2PCP	6B5	4	4	0
43CA	43CA	7	6	0
	Average	5.6	4.6	0.8
	Std. Dev	2 3	2.9	1.0

ments of residues in contact with the antigen was found. A more systematic analysis, in correlation coefficients, supports all these observations (data not shown), thus emphasizing the finding that residues in contact and somatic hypermutation do not correlate.

DISCUSSION

This report addresses the extent to which somatic hypermutation modifies the residues in direct contact with the antigen. To achieve this, we determined the residues in contact with the antigen in the region coded by the V_L gene of murine antibody–antigen complexes. The set of contacting residues was then compared with the putative somatic mutations. Results indicate that, although 52% of somatic hypermutations are concentrated in CDRs, somatic replacements occur mainly in residues that are not involved in the contacts with the antigen.

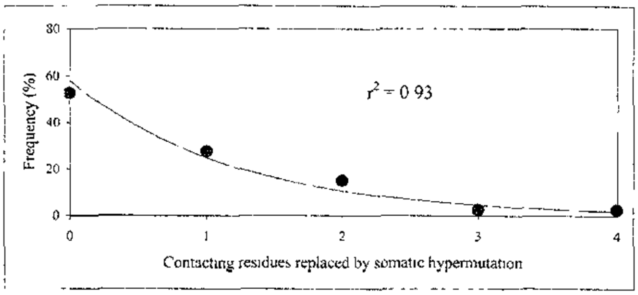


Fig. 2. Frequency of residues in contact with the antigen that have been modified by somatic hypermutation per antibody

Figure 3 depicts the patterns of putative somatic mutations and of residues in contact with the antigen on the surface of an antibody. Positions frequently involved in contacts, namely, 30a, 32, 50, and from 91 to 94, form a continuum patch at the center of the antigen-binding site [Fig. 3(a)]. Residues 32, 91, and 92, which are in contact with almost all the antigens, are just in the middle of that patch, conforming to the bottom of the antigen-binding site. Somatic mutations are spread all the way to the antigen-binding site periphery [Fig. 3(b)], with positions more frequently mutated, that is, 34, 55, and 94 flanking contacts with the antigen. The Cα of residue 34 is not at the surface as defined by running a probe of 10 Å radius over the trace, but we have indicated its relative position on the V_L surface. It is just down to residues 32, 91, and 92 in the interface with V_H. Residues 55 and 94, on the other hand, define mutational hot spots at the edge of the antigen-binding site [see Fig. 3(b)].

The spatial distribution of contacts and mutations observed in Figure 3 resembles the pattern found by Tomlinson and coworkers in human sequences.[4] These authors reported that germline gene diversity is located at the center of the antigen-binding site, in contrast to somatic mutations that take place mainly at the antigen-binding site periphery. Provided that most contacts are established at the center of the antigen-binding site [Fig. 3(a)], it can thus be proposed that the germline gene repertoire evolved to concentrate its diversity in that region, which guarantees the recognition of any antigen. The somatic hypermutation mechanism then evolved to spread mutations to the antigen-binding site periphery [Fig. 3(b)], which preserves the residues in contact. That strategy, as Tomlinson and coworkers[4] suggested, may have been favored by evolution as an efficient way for searching the sequence space.

However, the mechanism of somatic hypermutation predates the combinatorial gene rearrangement and diversification of the germline gene families.[20,21] In some species (e.g., sheep and rabbits), somatic hypermutation diversifies the primary repertory of antibodies.[3,21] Such a diversification generates antibodies capable of contending with the antigenic challenge. In man and mouse, where the somatic hypermutation is antigen dependent, this

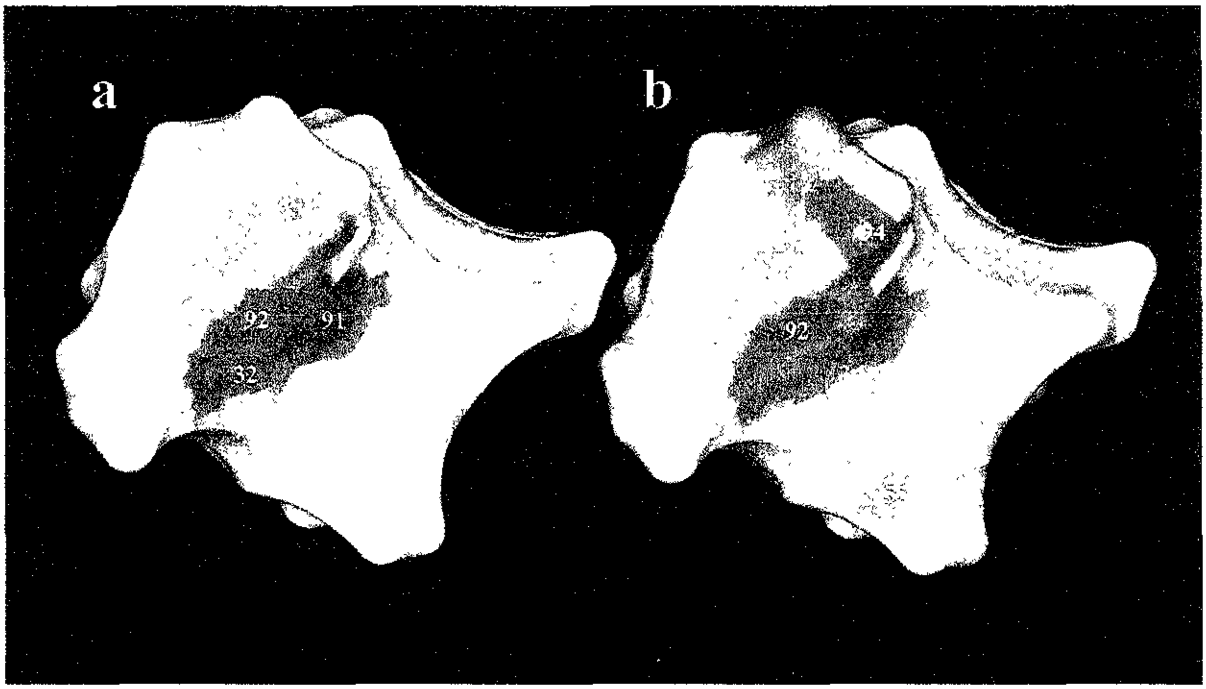


Fig. 3. Connolly surface[39] of the V_L domain of antibody MCPC603 (PDB code: 2IMM), shown from the antigen perspective. The color code corresponds to a gradient from red to white according the data of Figure 1. Red indicates a residue establishing contact (**a**) or mutated (**b**) in all complexes (equal to 1 in Fig. 1). White indicates a residue that never makes contacts or is never mutated (equal to 0 in Fig. 1). Positions very often found in contacts (a) or frequently mutated (b) are labeled. The surface was defined by running a probe of 10 Å radius over the trace to avoid variation due to changes in the amino acid sequence and to obtain a surface common to all antibodies.

mechanism produces the raw material from which antibodies of improved affinity would be selected.[2,3] But, during that process, it has been observed that harmful mutations exceed useful replacements.[6] The frequency and type of harmful mutations vary, depending on the region of the V gene where they occurred.[22,23] Quantitative estimates of the impact of punctual mutations in the CDR-H2 of anti-PC antibodies shows that >50% of the mutations are deleterious and none improved binding.[6] Therefore, the somatic hypermutation appears to be an inefficient strategy to search the sequence space in man and mouse.

Hence, an alternative explanation for our finding emerged. The replacement of the contacting residues had an adverse effect on the antigen–antibody interaction, and the variants bearing those mutations were removed from the pool of somatically diversified genes. In molecular terms, antigen–antibody interaction requires shape and physicochemical complementarity;[24] the better the complementarity, the higher the affinity. The probability that a mutation is harmful increases as the complementarity (affinity) increases. This is particularly true for residues at the center of the antigen–antibody interface, because atoms located in that region are, in most cases, very well packed and then would not be replaced without an energetic cost.[25] At the edge, atoms remain partly accessible to solvent and can be replaced by water molecules.[25,26] On the periphery, away from the atoms involved in the interface, replacements can be easily accommodated into the solvent.

Following this reasoning, the observed mutations occurred in residues more tolerant to replacements. Such replacements would have had an impact on the affinity via indirect optimization of the antigen–antibody interactions or simply be neutral. In fact, analysis of the effect of the frequency of mutations on single chain antibodies indicates that most mutations leading to higher affinity corresponds to residues distant from the antigen-binding site.[27] Mechanisms of indirect optimization of interactions with the antigen have been described in a number of instances.[28–31] For example, Wedemayer and coworkers[28] showed how the affinity can increase by more than four orders of magnitude through somatic mutations that are at least 10 Å away from the contacting residues. Chatellier and coworkers[29] established that modifications of some of the residues involved in the V_H:V_L interface are able to influence the antigen-binding properties of antibodies. Previously, Shillbach and coworkers[30] suggested how substitutions in the side chain of an amino acid, which is not in contact with the antigen but is adjacent to a residue in contact, would be accommodated without substantial rearrangement in the antibody but with impact on the affinity. Even before, Foote and Winter[31] demonstrated how substitutions in the framework region underlying the CDRs modulate the antibody affinity.

Affinity maturation can also be achieved through mechanisms other than somatic hypermutation. Analyses of immune responses indicate that in some cases the antibodies dominating the secondary or tertiary responses use different genes than those selected during the primary response.[32–35] In this phenomenon, called repertoire shift, the new gene, which is close to the germline gene configuration, grants a set of contacts or an antigen-binding geometry that lead to affinity improvements.[34] Worth mentioning is that mathematical models attempting to explain the repertoire shift within the framework of the affinity maturation have to recognize explicitly the destructive nature of somatic hypermutation.[36]

It is important that the complementarity of germline gene and somatic diversities observed by Tomlinson and coworkers[4] can also be explained in the deleterious effect of somatic hypermutation. That is, most somatic mutations that occurred at the center of the antigen-binding site, being in residues that are more frequently involved in direct contacts with the antigens and thus are less tolerant to replacements, have been removed from the pool of somatically diversified genes. The somatic diversity appears to have evolved to be complementary to the diversity of the primary repertoire but, the "actual" situation is that most of the observed mutations are those after purifying selection.

Finally, our results relate to the V_L gene only and may not hold for the V_H gene, which encodes the CDR-H1 and CDR-H2, nor for the V_LJ_L and V_HDJ_H junctions that define the last position of the CDR-L3 and the CDR-H3, respectively. Nonetheless, it should be noted that mechanisms of indirect optimization of residues in contact with the antigen, repertoire shift, and the pattern of germline gene and somatic diversities have been observed in both V_L and V_H. This finding suggests that our results for the V_L gene would be generalized to the V_H gene. As for V_LJ_L and V_HDJ_H junctions, we should be cautious in the making of predictions, because other mechanisms to produce diversity as recombination and addition and/or deletion of nucleotides are operative at those regions.[1]

CONCLUSIONS

We found a lack of correspondence between residues in contact with the antigen and those modified by somatic hypermutation. This finding may reflect the destructive nature of somatic hypermutation instead of an evolutionary strategy to search the sequence space. In species such as man and mouse, where the somatic hypermutation is antigen dependent, the primary repertoire of antibodies may have evolved to provide a set of residues capable of recognizing any antigen. Direct modification of those residues by somatic hypermutation, being in residues less tolerant to replacements, more often than not may be harmful. Mutations in the V_L and V_H interface, down to the antigen-binding site, at its edge or far away from the antigen-contacting residues would be more tolerated, while leading to affinity improvements. Our findings thus define a conceptual framework to propose a novel strategy to improve the affinity of antibodies. One should focus on identifying the residues in contact with the antigen but, instead of selecting these for mutagenesis, the target for mutagenesis should be nearby residues. Actually, strategies focused in residues other than those in contact with the antigen have already been exploited successfully for an anti-hapten antibody.[37]

ACKNOWLEDGMENTS

I thank J. Hermans, G. Odell, R. Stock, and T. Buss for the critical revision of the manuscript and very valuable comments, G. Moreno-Hagelsieb for the help in determining the residues in contacts with the antigens, and E. Horjales for discussions.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity Nature 1983;302:575–581.
2. Neuberger MS, Milstein C. Somatic hypermutation. Curr Opin Immunol 1995;7:248–254.
3. Weill J-C, Reynaud C-A. Rearrangement/hypermutation/gene conversion. when, where and why? Immunol Today 1996;17:92–97.
4. Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhammer EL, Winter G The imprint of somatic hypermutation on the repertoire of human germline V genes. J Mol Biol 1996;256:813–817
5. Foster SJ, Dorner T, Lipsky PE Targeting and subsequent selection of somatic hypermutation in the human $V\kappa$ repertoire. Eur J Immunol 1999;29:3122–3132.
6. Wiens GD, Roberts VA, Whitcomb EA, O'Hare T, Stenzel-Poore MP, Rittenberg MB Harmful somatic mutations. lessons from the dark side Immunol Rev 1998;162:197–209.
7 Blanden RV, Rothenfluh HS, Zylstra P, Weiller GF, Steele EJ. The signature of somatic hypermutation appears to be written into the germline IgV segment repertoire. Immunol Rev 1998;162:117–132.
8. Zachau HG. The immunoglobulin kappa locus-or-what has been learned from looking closely at one-tenth of a percent of the human genome. Gene 1993;135.167–173.
9. Cox JP, Tomlinson IM, Winter G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. Eur J Immunol 1994;24:827–836.
10. Williams SC, Frippiat JP, Tomlinson IM, Ignatovich O, Lefranc MP, Winter G. Sequence and evolution of the human germline V lambda repertoire J Mol Biol 1996;264:220–232
11. Tomlinson IM, Walter G, Marks JD, Llewelyn MB, Winter G. The repertoire of human germline VH segments reveals about fifty groups of VH segments with different hypervariable loops. J Mol Biol 1992,227·776–798.
12. Matsuda FK, Ishii P, Bourvagnet K-I, Kuma H, Hayashida T, Miyata T, Honjo T. The complete nucleotide sequence of the human immunoglobulin heavy chain variable locus. J Exp Med 1998,188·2151–2162.
13 Thiebe R, Schäble KF, Bensch A, Brensing-Küppers J, Heim V, Kirschbaum T, Lautner-Rieske A, Mitlöhner H, Ohnrich M, Pourrajabi S, Röschenthaler F, Schwendinger J, Wichelhaus DP, Zocher I, Zachau HG. The variable genes and gene families of the mouse immunoglobulin κ locus Eur J Immunol 1999;29:2072–2081.
14. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238.777–793.
15 Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 1970;132·211–250.
16. Amzel LM, Poljak RJ Three-dimensional structure of immunoglobulins Annu Rev Biochem 1979;48:961–997.
17 Al-lazikani B, Lesk A, Chothia C. Standard conformation for the canonical structures of immunoglobulins J Mol Biol 1997,273.927–948
18 Padlan EA, Abergel C, Tipper JP Identification of specificity-determining residues in antibodies FASEB J 1995,9·133–139
19 Klein R, Jaenichen R, Zachau HG. Expressed human immunoglobulin κ genes and their hypermutation Eur J Immunol 1993;23 3248–3271

20. Hinds-Frey KR, Nishikata H, Litman RT, Litman GW. Somatic variation precedes extensive diversification of germline sequences and combinatorial joining in evolution of immunoglobulin heavy chain diversity J Exp Med 1993;178:815–824.
21. Diaz M, Flajnik MF. Evolution of somatic hypermutation and gene conversion in adaptive immunity. Immunol Rev 1998;162:13–24
22. Hurle MR, Helms LR, Li L, Chan WN, Wetzel R. A role for destabilizing amino acid replacements in light-chain amyloidosis Proc Natl Acad Sci USA 1994,91.5446–5450.
23. Dul JL, Argon Y. A single amino acid substitution in the variable region of the light chain specifically blocks immunoglobulin secretion. Proc Natl Acad Sci USA 1990;87·8135–8139.
24. Pauling L. Molecular structure and intermolecular forces. In: Landsteiner K, editor. The specificity of serological reactions. New York Dover Publications; 1962. p 275–293.
25. Lo Conte, Chothia C, Janin J The anatomy structure of protein–protein recognition sites J Mol Biol 1999;285·2177–2198.
26. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280·1–9.
27. Daugherty PS, Chen G, Iverson BL, Georgiou G. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. Proc Natl Acad Sci USA 2000;97:2029–2034.
28 Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. Science 1997;276:1665–1669.
29 Chatellier J, Van Regenmortel MHV, Vernet T, Altschuh D. Functional mapping of conserved residues located at the VL and VH domain interface of a Fab. J Mol Biol 1996;264:1–6
30. Schillbach JF, Near RI, Bruccoleri RE, Haber E, Jeffrey PD, Novotny J, Sheriff S, Margolies MN. Modulation of antibody affinity by a non-contact residue. Protein Sci 1993,2.206–214
31. Foote J, Winter G Antibody framework residues affecting the conformation of the hypervariable loops J Mol Biol 1992;224:487–499.
32 Berek C, Milstein C. Mutation drift and repertoire shift in maturation of the immune response. Immunol Rev 1987;96:23–41.
33. Alzari PM, Spinelli S, Mariuzza RA, Boulot G, Poljak RJ, Jarvis JM, Milstein C. Three-dimensional structure determination of an anti-2-phenyloxazolone antibody. the role of somatic maturation of an immune response. EMBO J 1990;9:3807–3814
34. Foote J, Milstein C. Kinetic maturation of an immune response. Nature 1991;352:530–532.
35 Randen I, Potter KN, Li Y, Thompson KM, Pascual V, Forre O, Natvig JB, Capra DL. Complementarity-determining region 2 is implicated in the binding of staphylococcal protein A to human immunoglobulin VHIII variable regions. Eur J Immunol 1993;23·2682–2686.
36. Shannon M, Mehr R. Reconciling repertoire shift with affinity maturation. the role of deleterious mutations. J Immunol 1999;162: 3950–3959
37. Arkin MR, Wells JA. Probing the importance of second sphere residues in an esterolytic antibody by phage display. J Mol Biol 1998;284.1083–1094.
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000,8:235–242.
39 Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221·709–713.

Conclusiones

El objetivo principal de este proyecto es profundizar en el estudio de la relación entre la estructura y la función de los anticuerpos. Aunque actualmente se dispone de gran cantidad de información para los anticuerpos, no es posible predecir aun la especificidad o la afinidad de un anticuerpo en particular. Algunos trabajos de investigación relativamente recientes sugieren que es posible establecer reglas del reconocimiento molecular mediado por esta familia de proteínas (Capítulo I).

El primer paso del proyecto (Capítulo II) fue el desarrolló de una herramienta de computo denominada VIR.II. Esta herramienta tiene su antecedente en VIR (Almagro *et al.*, 1995). A diferencia de la versión anterior, VIR.II esta disponible en el WWW (URL: http://www.ibt.unam.mx/vir/VIR/vir_index.html). VIR.II permite obtener alineamientos múltiples de grupos de secuencias de anticuerpos por especie, cadena (VH o VL) y especificidad fina. Otra característica de VIR.II - tal vez la más importante – es que permite una clasificación de las especificidades finas de acuerdo a criterios de la naturaleza química del antígeno. Por ejemplo: proteínas, péptidos, haptenos, ácidos nucleícos, etc. Esta clasificación pudiera ser la semilla de una taxonomía de antígenos en el futuro.

Además de VIR.II, se desarrolló una interfaz con el Protein Data Bank (PDB) que permite tener acceso a toda la información sobre las estructuras tridimensionales de anticuerpos conocidas (http://www.ibt.unam.mx/vir/structure/structures.html). Esta página web ha sido unida por varias páginas del mundo entero y ha sido citada a la fecha en varias publicaciones (Dübel 2000). Con VIR.II y esta página Web, emprendimos el análisis de las secuencias y estructuras conocidas de anticuerpos.

El análisis de la relación entre la estructura y la función de las estructuras conocidas de anticuerpos (Capítulo III) mostró que de las 49 estructuras de anticuerpos anti-proteína, 33 (67%) tienen un asa corta en L1. En contraste, el 90% de la estructuras anti-péptido poseen una asa larga en L1. Estos resultados en las estructuras son consistentes con un análisis similar de las secuencias. De 1,017 secuencias depositadas en la base de datos de Kabat de anticuerpos anti-proteína 692 (68%) tienen un asa corta. En el caso de las secuencias de anticuerpos anti-péptido, de 121 secuencias, 87 (71%) tienen una asa larga.

Al estudiar la topografía del sitio de interacción con el antígeno, se observó que los anticuerpos anti-proteína, es decir, aquellos con un asa corta, tienen una superficie plana en su sitio de unión al antígeno. Los anticuerpos anti-péptido, es decir aquellos con L1 largo, tienen una ranura en el sitio de unión al antígeno. Desde el punto de vista de reconocimiento molecular, una superficie plana proporciona una superficie complementaria a los antígenos grandes, como proteínas. En el caso de los anticuerpos anti-péptido, debido a su tamaño menor con respecto a una proteína, una ranura en el sitio de unión con el antigeno asegura una complementariedad adecuada con estos ligandos. Estas observaciones, nos permiten proponer que un determinante estructural que dirige la diferencia de reconocimiento entre proteínas y péptidos es L1.

Una observación interesante que emergió de este trabajo fue que ningún anticuerpo con superficie plana reconoce péptidos. Una posible explicación es que una superficie plana difícilmente brinda una complementariedad adecuada a un péptido. En contraste, de las 49 estructuras de anticuerpos anti-proteína, 16 presentan un L1 largo y por lo tanto tienen la ranura típica de los anticuerpos que reconocen péptidos. De ellas, 5 fueron co-cristalizadas con el antígeno y en consecuencia pudieron ser analizadas. Estas estructuras muestran como tendencia poca complementariedad antígeno-anticuerpo, excepto cuando el epitope se proyecta del plano del antígeno y rellena la ranura en el anticuerpo. Esta observación sugiere el modo en que un anticuerpo obtenido con un péptido (anticuerpo anti-péptido por definición) pudiera ser reconocido por una proteína.

Por otra parte, del análisis de la topografía del sitio de unión al antígeno llamó la atención el modo en que los anticuerpos reconocen haptenos. Debido al pequeño tamaño de los haptenos, estos no requieren de cambios importantes en la topografía del sitio de unión al antígeno en el anticuerpo, como los que se realizan por la inserción o deleción de residuos en L1. Al parecer, solo se requieren pequeños cambios en la topografía del sitio de interacción con los antígenos, mismos que pueden realizarse a través de cambios conformacionales o substituciones de las cadenas laterales durante el proceso de maduración de la respuesta inmunológica.

Es importante mencionar que este resultado es totalmente novedoso, en trabajos previos se observó que el reconocimiento de haptenos y péptidos era similar (Vargas-Madrazo, *et al* 1995), suposición basada en que el tamaño de los haptenos es más cercano a los péptidos que a las proteínas,

En el capítulo IV analizamos cual es el papel de la hipermutación somática al sustituir los residuos en contacto con el antígeno. Para ello consideramos 40 estructuras no redundantes de ratón que presentan una resolución de 2.5 Å o menos y que han sido co-cristalizadas con el antígeno. Un primer paso en el análisis fue la asignación de las secuencias de línea germinal a las estructuras. Se observó que el gene V_k acumula en promedio 5 mutaciones durante el proceso de maduración de la respuesta inmunológica. El número promedio de contactos es similar al de mutaciones somáticas, alrededor de 6. Sin embargo, en el 60% de los anticuerpos no coinciden las posiciones mutadas con los residuos en contacto. En otro 27% de los anticuerpos sólo hay 1 mutación en los residuos en contacto con el antígeno. Esto indica que los residuos en contacto son modificados raramente por el proceso de hipermutación somática.

Si se identifican las posiciones mutadas y los residuos en contactos en la superficie de un anticuerpo (ver Figura 3, Capitulo IV), se observa que los contactos están en el centro del sitio de interacción, mientras que las mutaciones somáticas están en la periferia. Una posible explicación para este patrón, es que las mutaciones en los residuos en contacto son dañinas en un importante número de casos. Así, son eliminadas del "pool" de genes durante el proceso de maduración de la respuesta inmunológica. Esta expiación tiene, sin duda, implicaciones para las teorías que explican el origen y la evolución del repertorio de anticuerpos.

La observación de que no hay correlación entre los residuos en contacto y las mutaciones somáticas, tiene también conseecuencia prácticas. Esto es, si se desea madurar la afinidad de un anticuerpo in *vitro*, lo que se debe hacer es identificar los residuos en contacto, pero en vez de mutar estos, se deben mutar son los residuos circundantes a los residuos en contacto.

En resumen, el conjunto de trabajos realizados durante el doctorado confirman la existencia de reglas que gobiernan el reconocimiento molecular mediado por anticuerpos. Estas reglas permiten proponer esquemas predictivos para el diseño racional de anticuerpos, tanto a nivel de la topografía general del sitio de interacción con el antígeno (Capítulo III) como en los detalles de maduración de la afinidad (Capítulo IV).

Con respecto a nuestros primeros trabajos (ver anexos: Vargas-Madrazo *et al.*, 1997, Almagro *et al.*, 1997, Almagro *et al.*, 1998), estos estuvieron enfocados al análisis y comparación de los genes de línea germinal de las dos especies más estudiadas: humano y ratón. En el primer anexo presentamos la caracterización estructural del repertorio de genes de línea germinal de humano (Vargas-Madrazo *et al.*, 1997).

El análisis anterior permitió proponer un repertorio primordial de V_H, constituido por las clases de estructuras canónicas 1-2 , 1-3 (asas cortas en H1 y H2) y 3-1 (asa larga en H1 y corta en H2) codificadas por los clanes I, III y II respectivamente. En el caso de V_k, el repertorio estructural primordial para los genes de línea germinal está representado por las clases de estructuras canónicas 2-1 (asas cortas en L1 y L3) y 4-1 (asa larga en L1 y corta en L3) codificadas por los clanes I y II respectivamente. La combinación de estas clases de estructuras canónicas nos permitieron proponer que el repertorio V_H-V_k (H1-H2-L1-L3) de clases de estructuras canónicas se reduce solo a cuatro combinaciones: 1-2/3-2-1, 1-2/3-4-1, 3-1-2-1 y 3-1-4-1.

Desde el punto de vista estructural, estas clases V_H-V_k de estructuras canónicas representan el repertorio de topografías (planas, con cavidades y con ranuras) del sitio de unión al antígeno, necesarias para el reconocimiento molecular mediado por anticuerpos, específicamente el reconocimiento de antígenos proteicos, péptidos y haptenos. Las demás clases parecen ser variaciones sobre este esquema primordial.

En el anexo II analizamos el repertorio estructural V_H de ratón, el cual está codificado por la familia V_H1 y representa la clase 1-3, mientras que en humano, su repertorio estructural es codificado por la familia V_H3 (anexo I), la cual codifica la clase 1-2. Esta diferencia no resulta tan significativa, puesto que, la diferencia entre la estructura canónica 2 y 3 sólo representa cambios conformacionales.

El análisis del repertorio estructural de V_k mostró diferencias entre los repertorios estructurales de humano y ratón. El repertorio de ratón es más diverso, es decir, codifica siete clases de estructuras canónicas, mientras que humano solo codifica para tres clases. La clases de estructuras canónicas de ratón que no se observaron en el humano son: 5-1, 1-1 y 1-2. Otro dato sobresaliente es que el 60% de los genes V_k de humano codifican para la clase 2-1, mientras que en ratón, sólo el 30% de los genes codifican esta clase. Las diferencias encontradas en el repertorio estructural V_k, pueden ser resultado de la estrategia seguida por ratón para compensar las diferencias en el contenido de genes V_λ.

Estos resultados, tomados en su conjunto con la existencia de reglas que gobiernan el reconocimiento de diferentes tipos de moléculas, apuntan en la dirección de que el proceso de evolución de los anticuerpos no es del todo estocástico, como se ha propuesto; existen caminos preferenciales a través de los cuáles la evolución ha dado forma a los diferentes repertorios de anticuerpos en las especies conocidas.

Bibliografía

Almagro JC., Vargas-Madrazo E., Zenteno-Cuevas R., Hernández-Mendiola V., Lara-Ochoa, F. 1995. VIR: A computacional tool for analysis of immunoglobulin sequences. BioSystems. 35:25-32.

Almagro JC., Dominguez-Martinez V., Lara-Ochoa F., Vargas-Madrazo E. 1996. Structural repertoire in human VL pseudogenes of immunoglobulins: comparison with functional germline genes and amino acid sequences. Immunogenetics. 43:92-6.

Almagro JC., Hernandez I., del Carmen Ramirez M., Vargas-Madrazo E. 1997. The differences between the structural repertoires of VH germ-line gene segments of mice and humans: implication for the molecular mechanism of the immune response. Mol Immunol. 34:1199-214.

Almagro JC., Hernandez I., Ramirez MC., Vargas-Madrazo E. 1998. Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. Immunogenetics. 47:355-63.

Al-Lazikani B., Lesk AM., Chothia C. 1997. Standard conformations for the canonical structures of immunoglobulins. J. Mol. Biol. 273: 927-948.

Amzel LM., Poljak RJ. 1979. Three-dimentional structure of immunoglobulin. Ann. Rev. Biochem. 48: 916-997.

Berman JE., Mellis SJ., Pollock R., Smith CL., Suh H., Heinke B., Kowal C., Surti U., Cantor CR., Alt FW. 1988. Content and organization of the human Ig VH locus: Definition of three new VH families and linkage to the Ig CH locus. EMBO J. 7 : 727-738.

Berman HM., Westbrook J., .Feng Z., Gilliland G., Bhat TN., Weissig H., Shindyalov IN., Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Research. 28: 235-242

Buluwela L, Albertson DG., Sherrington P., Rabbitts PH., Spurr N., Rabbitts TH. 1988. The use of chromosomal translocations to study human immunoglobulin gene organization: mapping DH segments within 35 kb of the C mu gene and identification of a new DH locus. EMBO J. 7:2003

Chothia C., Lesk, AM., Tramontano A., Levitt M., Smith-Gill SJ., Air G., Sheriff S., Padlan EA., Davies D., Tulip WR., Colman PM., Spinelli S., Alzari, PM., Chothia, C., Lesk AM. 1987. Canonical structures for the hipervariable regions of immunoglobulin. J.Mol. Biol. 196 :901

Chothia C., Lesk AM., Tramontano, A., Levitt, M., Smith-Gill SJ., Air G., Sheriff S., Padlan EA., Davies D., Tulip WR., Colman PM., Spinelli S., Alzari PM., Poljak RJ. 1989. Conformations of immunoglobulins hypervariable regions. Nature (London). 342 : 877

Chothia C., Lesk AM., Gherardi E., Tomlinson IM., Walter G., Marks JD., Llewelyn MB., Winter G. 1992. Structural repertoire of the human VH segments. J. Mol. Biol. 227 : 799

Cook GP., Tomlinson IM., Walter G., Riethman H., Carter NP., Buluwela L., Winter G., Rabbitts TH. 1994. A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q. Nature Genetics. 7: 162

Cook GP., Tomlinson IM. 1995. The human immunoglobulin V_H repertoire.Immunology Today. 16 : 237

Corbett SJ., Tomlinson IM., Sonnhammer EL., Burk D, Winter G. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. J.Mol. Biol. 270:587.

Cox JP., Tomlinson IA., Winter G. 1994. A directory of human germ-line V-kappa segments reveals a trong bias in their usage. Eur. J. Immunol. 24: 827

Croce CM., Shander M, Martinis J., Cicurel L, D'Ancona GG., Dolby TW., Koprowski II. 1979. Chromosomal location of the genes for human immunoglobulin heavy chains. Proc Natl Acad Sci U S A. 76:3416

Davies RD., Padlan EA Segal DM. 1975. Three-dimensional structure of immunoglobulins. A. Rev. Biochem. 44: 639.

Dariavach P., Lefranc G., Lefranc MP. 1987. Human immunoglobulin C lambda 6 gene encodes the Kern+Oz-lambda chain and C lambda 4 and C lambda 5 are pseudogenes. Proc Natl Acad Sci U S A. 84:9074.

Dübel S. 2000. The antibody web. Immunology Today. 355

Dunham I., Shimizu N., Roe BA., Chissoe S., Hunt AR., Collins JE., Bruskiewich R., Beare DM., Clamp M., Smink LJ., Ainscough R., Almeida JP., Babbage A., Bagguley C., Bailey J., Barlow K., Bates KN., Beasley O., Bird CP., Blakey S., Bridgeman AM., Buck D., Burgess J., Burrill WD., O'Brien KP., et al. 1999. The DNA sequence of human chromosome 22. Nature. 402:489

Edelman GM., Cunningham BA., Gall WE., Gottkieb PD., Rutihauser U., Waxdal MJ. 1969. The covalent structure of the entire gammaG immunoglobulin molecule. Proc.Natl. Acad. Sci. U.S.A. 63 :78

Edelman GM., Gally JA. 1962. J. Exp.Med. 116 :207

Emanuel BS., Cannizzaro LA., Magrath I., Tsujimoto Y., Nowell PC., Croce CM. 1985. Chromosomal orientation of the lambda light chain locus: V lambda is proximal to C lambda in 22q11. Nucleic Acids Res. 13:381

Erikson J., Martinis J., Croce CM. 1981. Assignment of the genes for human lambda immunoglobulin chains to chromosome 22. Nature. 294:173-5.

Frippiat JP., Williams SC., Tomlinson IM., Cook GP., Cherif D., Le Paslier D., Collins JE., Dunham I., Winter G., Lefranc MP. 1995. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. Hum. Mol. Genet. 4: 983

Ghanem N., Dariavach P., Bensmana M., Chibani J., Lefranc G., Lefranc MP. 1988. Polymorphism of immunoglobulin lambda constant region genes in populations from France, Lebanon and Tunisia. Exp Clin Immunogenet. 5:186

Guarné A., Bravo J., Calvo J., Lozano F., Vives J., Fita I. 1996. Conformation of the hipervariable region L3 without the key proline residue. Prot. Sci. 5:167.

Haynes MR., Stura EA., Hilvert D., Wilson IA. 1994. Routes to catalysis structure of a catalytic antibody and comparasion with its natural counterpart. Science. 263:646

He XM., Ruker F., Casale E., Carter DC. 1992. Structure of a human monoclonal antibody Fab fragment against gp41 of human immunodeficiency virus type 1. Prot. Nat. Acad. Sci. USA. 89:7154

Hieter PA., Max EE., Seidman JG., Maizel JV Jr., Leder P. 1980. Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. Cell. 22:197

Hieter PA., Maizel JV Jr, Leder P. 1982. Evolution of human immunoglobulin kappa J region genes. J Biol Chem. 257:1516

Huber, C., Schäble, H. F., Huber, E., Klein, R., Meindl, A., Thiebe, R., Lamm, R. & Zachau, H. G. 1993. The V_k genes of the L regions and the repetoire of V_k gene sequences in the human germ line. Eur. J. Immunol. 23 : 2868

Ichihara Y., Matsuoka H., Kurosawa Y. 1988. Organization of human immunoglobulin heavy chain diversity gene loci. EMBO J. 7:4141-50.

Janeway CA., Travers P., Walport M., Capra DJ. (1999). Immunobiology. The immune system in health and Disease. Fourth Edition. Garland pp. 92

Johnson G., Wu TT., Kabat EA. 1995 SEQHUNT. A program to screen aligned nucleotide and amino acid sequences. Methods. Mol Biol. 51: 1-15.

Johnson G., Wu TT. 2000. Kabat Database and its applications: 30 years after the first variability plot. Nucleic Acids Research., 28, 214-218.

Kabat EA., Wu TT., Perry HM., Gottesman KS. Foeller C. 1991. Sequences of Proteins of Immunological Interest., 5th Edition, US Department of Health and Human Services, Public Health Service, National Institutes of Health (NIH Publication No. 91-3242). Washington. D.C.

Kabat EA., Wu TT. 1971. Attemps to locate complemetarity determining residues in the variable portions of light and heavy chains. Ann. NY Acad. Sci. USA 79:4118

Kay pH., Moriuchi J., Ma PJ., Saueracker E. 1992. An unusual allelic form of the immunoglobulin lambda constant region genes in the Japanese. Immunogenetics.35:341.

Kodaira M., Kinashi T., Umemura I., Matsuda F., Noma T., Ono Y., Honjo T. 1986. Organization and evolution of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. 190 : 529-541.

Kawasaki K., Minoshima S., Schooler K., Kudoh J., Asakawa S., de Jong PJ., Shimizu N. 1995. The organization of the human immunoglobulin lambda gene locus. Genome Res. 5:125.

Kawasaki K., Minoshima S., Nakato E., Shibuya K., Shintani A., Schmeits JL., Wang J., Shimizu N. 1997. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. Genome Res. 7:250.

Lara-Ochoa F., Almagro JC., Vargas-Madrazo E., Conrad M. 1996. Antibody-Antigen Recognition : A Canonical Structure Paradigm. J. Mol. Evol. 43:678

Lee KH., Matsuda F., Kinashi T., Kodaira M., Honjo T. 1987. A novel family of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. 195 : 761-768.

Lefranc MP., Pallares N., Frippiat JP. 1999 Allelic polymorphisms and RFLP in the human immunoglobulin lambda light chain locus. Hum Genet. 104:361

Mc Bride OW., Hieter PA., Hollis GF., Swam D., Otey MC., Leder P. 1982. Chromosomal location of human kappa and lambda immunoglobulin light chain constant regions genes J. Exp. Med. 155 :1480.

MacCallum RM., Martin ACR., Thornton J.M. 1996. Antibody-antigen interactions: contact analysis and binding site topography. J Mol. Biol 262:732-45

Malcolm S., Barton P., Murphy C., Ferguson-Smith MA., Bentley DL., Rabbitts., TH. 1982. Localization of human immunoglobulin k light chain variable region genes to the short arm of cromosome 2 by in situ hibridization. Proc. Natl. Acad.Sci. U.S.A 79 : 4957

Martin ACR. 1996. Accessing the Kabat Antibody Sequence Database by Computer PROTEINS: Structure, Function and Genetics. 25: 130

Martin ACR., Thornton JM. 1996. Structural families in loops of homologous proteins: automatic classification, modeling and application to antibodies. J. M. Biol. 26:815

Matsuda F., Lee KH., Nakai S., Sato T., Kodaira M., Zong SQ., Ohno H., Fukuhara S., Honjo T. 1988. Dispersed localization of D segments in the human immunoglobulin heavy-chain locus. EMBO J. 7 : 1047-1051.

Matsuda F., Shin EK., Nagaoka H., Matsumura R., Haino M., Fukita Y., Taka-ishi S., Imai T., Riley JH., Anand R., et al. 1993. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. Nat Genet. 3:88

Morea V., Tramontano A., Rustici M., Chothia C., Lesk AM. 1998. Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol. 275:269

Nisonoff A., Wisssler FC., Woernley., DL. 1959. Biochem Biophys. Res. Commun. 1 :318

Oliva B., Bates., PA., Querol E., Aviles FX., Sternberg MJE. 1998. Automatic classification of the complementarity-determining region 3 of the heaby chain (H3) loops into canonical forms and its applicationto protein structure prediction. J. Mol. Biol. 279:1193

Padlan EA., Segal DM., Spande TF., Davies DR., Rudikoff R., Pottter M. 1973. Structure at 4.5 Å resolution of a phosphorylcholine-binding Fab. Nature New Biol. 245:165-167.

Pascual V., Capra D. 1991. Human immunoglobulin heavy-chain variable region genes:organization, position and expression. Adv. Immunol. 49:1

Poljak RJ. 1989. Conformations of immunoglobulins hypervariable regions. Nature (London). 342 : 877

Poljak RJ., Amzel LM., Avery HP., Becka LN., Nisonof A. 1972. Structure of Fab New at 6 Å resolution. Nature New Biol. 233:137.

Poljak RJ., Amzel LM., Avery HP., Chen BL., Phizacherley RP., Saul F. 1973. Three - dimentional structure of the Fab' fragment of a human immunoglobulin at 2.8 A resolution. Proc. Natl. Acad. Sci. USA. 69 :3689

Porter RR. 1959. The hidrolysis of rabbit γ globulin and antibodies by crystalline papain. Biochemical J. 73:119.

Rabbitts TH., Forster A., Milstein CP. 1981. Human immunoglobulin heavy chain genes: evolutionary comparisons of C mu, C delta and C gamma genes and associated switch sequences. Nucleic Acids Res. 9:4509

Ramirez-Benitez CDM., Almagro JC. 2001a. Analisys of Antibodies of known structure suggests a lack of correspondence between the residues in contact with the antigen and those modified by somatic hipermutation. PROTEINS: Structure, Function, and Genetics. 45: 199-206

Ramirez-Benitez CDM., Moreno-Hagelsieb G., Almagro JC. 2001b. VIRII: A new interface with the antibody sequences in the Kabat database. Biosystems. En prensa

Ramirez-Benitez CDM., Ceceña HA., Almagro JC. Structure-function relationships in anti-protein, anti-peptide and anti-hapten antibodies. Por enviar.

Ravetch JV., Siebenlist U., Korsmeyer S., Waldmann T., Leder P. 1981.Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. Cell. 27:583

Rudikoff S., Potter M., Segal DM., Padlan EA., Davies DR. 1972. Crystal of phosphorilcholine-binding Fab-fragments from mouse myeloma proteins: preparation and X-ray analysis. Proc. Natl. Acad. Sci. USA. 69:3689.

Sarma VR., Silverton EW., Davies DR. Terry WP. 1971. Three-dimentional structure at 6 A resolution of human gamma G immunoglobulin molecule. J.Biol. Chem. 246 :3753

Schäble KF., Zachua HG. 1993. The variable genes of the human immunoglobulin k locus. J.Biol. Chem. 374:1001

Schäble HF., Thiebe R., Flügel A,. Meindl A., Zachua HG. 1994. The human immunoglobulin k locus: pseudogenes, unique and repetitive sequences. J. Biol. Chem. 375.189

Shen A., Humphries C., Tucker., P. Blattner F. 1987. Human heavy-chain variable region gene family nonrandomly rearranged in familial chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. USA. 84 : 8563

Shin EK., Matsuda F., Nagaoka H., Fukita Y., Imai T., Yokoyama K., Soeda E., Honjo T. 1991. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. EMBO J. 10:3641

Schiffer M., Girling RL., Ely KR., Edmunson AB. 1973. Biochemistry. 12 :4620

Shirai H., Kidera A,. Nakamura H. 1996. Structural classification of CDR-H3 in antibodies. FEBS Letters, 399: 1

Siebenlist U., Ravetch JV., Korsmeyer S., Waldmann T., Leder P. 1981. Human immunoglobulin D segments encoded in tandem multigenic families.Nature. 294:631

Taub RA., Hollis GF., Hieter PA., Korsmeyer S., Waldmann TA., Leder P. 1983. Variable amplification of immunoglobulin lambda light-chain genes in human populations. Nature 304:172-4.

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

Tomlinson IM., Walter G., Marks JD., Llewelyn MB. Winter G. 1992. The repertoire of human germline VH segments reveals about fifty groups of VH segments with different hypervariable loops. J. Mol. Biol. 227 : 776

Tomlinson IM., Cook GP., Carter NP., Elaswarapu R., Smith S., Walter G., Buluwel, L., Rabbitts TH., Winter G. 1994. Human immunoglobulin V_H and D segments on chromosomes 15q11.2 and 16p11.2. Human. Molec. Genetics. 3 : 853

Tomlinson IM., Cox JPL., Gherardi E., Lesk AM., Chothia C. 1995. The structural repertoire of the human VK domain. EMBO J. 14:4628

Tomlinson IM., Walter G., Jones PT., Dear PH., Sonnhammer EL., Winter G. 1996. The imprint of somatic hipermutation on the repertoire of human germline V genes. J..Mol. Biol. 256:813

Valentine RC., Green NM. 1967. J. Mol. Biol. 27: 615-17.

Vargas-Madrazo E., Lara-Ochoa F., Almagro JC. 1995a. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions asociated to the mechanism of immune recognition. J. Mol. Biol. 254:497

Vargas-Madrazo E., Almagro JC., Lara-Ochoa F. 1995b. Structural repertoire in VH pseudogenes of immunoglobulins: comparison with human germline genes and human amino acid sequences. J Mol Biol. 246:74

Vargas-Madrazo E., Lara-Ochoa F., Ramirez-Benites MC., Almagro JC. 1997. Evolution of the structural repertoire of the human V(H) and Vkappa germline genes. Int Immunol. 9:1801

Vasicek TJ., Leder P. 1990. Structure and expression of the human immunoglobulin lambda genes. J Exp Med. Aug 1;172(2):609-20

Walter MA., Surti U., Hofker MH., Cox DW. 1990. The physical organization of the human immunoglobulin heavy chain complex. EMBO J. 9 :3303

Weins GD., Roberts VA., Whitcomb EA., O'Hare T., Stenzel-Poore MP., Rittemberg MB. 1998. Harmful somatic mutations: lessons from the darkside. Immunol Rev. 162. 197

Wilson IA., Stanfield RL. 1994. Antibody-antigen interactions: new structures and new conformational changes. Curr. Opin. Sturct. Biol. 3: 113-118.

Williams SC., Frippiat J-P., Tomlinson IA., Ignatovich O., Lefranc M-P., Winter G. 1996. Sequence and evolution of the human germline V_L repertoire. J. Mol. Biol. 264: 220

Willems van Dijk K., Mortari F., Kirkham P.M., Schroeder HW., Milner ECB. 1993. The human immunoglobulin V_H7 gene family consist of a smal, polimorphic group of six to eight gene segments dispersed throughout the V_H locus. Eur. J. Immunol. 23:832

Winter G., Milstein C. 1991 Man-made antibodies. Nature 349: 293

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 antibody complementary. J. Exp. Med. 132:211

Wu S., Cygler M. 1993. Conformation of Complementarity Determining Region L1 Loop In Murine IgG 1 Light Chain Extends the Repertoire. J. Mol. Biol. 229:597

Zachau HG., 1989. The immunoglobulin Genes. Academic Press, London and San Diego pg.91.

Zachau HG. 1992. The human immunoglobulin k locus. Characterization of the duplicated A regions. Eur. J. Immunol. 22 : 1023

Zachau HG. 1993. The immunoglobulin kappa locus-or-what has been learned from looking closely an one-tent h of percent of the human genome. Gene 135: 167

Anexos

International Immunology Vol 9, No 12 pp 1801–1815

Evolution of the structural repertoire of the human V_H and V_κ germline genes

Enrique Vargas-Madrazo, Francisco Lara-Ochoa[1], Maria C. Ramirez-Benites and **Juan C. Almagro[1]**

Instituto de Investigaciones Biológicas, Universidad Veracruzana, Apartado Postal 495, Xalapa, Veracruz 91000, México
[1]Instituto de Química, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, CP 04510, México, DF

Keywords antigen-binding site, canonical structure, canonical structure class, clans, gene families, V_κ locus, V_H locus

Abstract

Variable genes of human Ig are classified in families and clans which reflect the early events of gene duplication in the evolution of the locus. This organization in multiple copies of variable genes plus the somatic processes of recombination and hypermutation allows the immune system to generate an antibody repertoire of great diversity. At present the role that somatic processes play in the generation of that diversity is understood with some detail. It is a matter of hard controversy, however, which selective pressures have shaped the evolution of the germline genes of Ig and, consequently, what the role of this germline component in the generation of the antibody diversity actually is. Previous studies of our group have showed that the structural repertoire of Ig—determined by the canonical structures—is an important factor to determine the recognition properties of the antibodies. Complete knowledge of the sequences of the human V_H and V_κ loci is available to analyze the evolution of the structural repertoire of these loci. Two phylogenetic gene trees were built from the functional germline genes and the evolution of the structural repertoire was studied. We report that for both loci the canonical structures are not randomly distributed within the tree. Conversely, it is shown that the evolution of the structural repertoire follows a gradual process of diversification. This indicates a correlation between the evolution of genes and the structural repertoire, although important differences are found in the patterns of evolution of the structural repertoire between V_H and V_κ. Based on those results we propose a primordial structural repertoire for V_H and V_κ. The general properties and an outline of the three-dimensional structure of this primordial repertoire are given.

Introduction

The variable locus of Ig is composed by multiple genes which have evolved through gene duplication in order to generate a diverse germline repertoire (1) Analysis of homology among the V gene segments has revealed that these can be grouped in discrete families (2,3) Additionally, it has been proposed that the V gene families can also be grouped in clans (4) which represent the early events in the evolution of the V genes (5,6)

The nature of the selective forces responsible for molding the evolution of the V genes is still a matter of controversy (for reviews, see 7,8). Analysis of nucleotide and amino acid substitutions at the coding region of the V genes has shown that the regions involved in the interaction with the antigen present high variability, in contrast to the relatively conserved remaining framework regions This results suggest different selective forces acting over these two regions (1 8–10) Additional analysis of others aspects like polymorphism (1,11,12), sequence variability (13,14) and phylogeny (4 5,15) has provided additional evidence of selective forces acting over V genes in order to shape their variability

The adopted model of immune recognition is the initial aspect of the controversy about the selective forces. A widely accepted model is based on the established correlation between the regions of random hypervariability and the

Correspondence to E Vargas Madrazo

Transmitting editor R Perlmutter

Received 22 July 1997 *accepted* 18 August 1997

specificity properties of the antibodies (16). In this model is assumed that a high rate of substitutions in the hypervariable regions can generate an antibody repertoire capable of contending successfully with the antigen challenge (8,10,13) That means (according to this model) it is possible to generate a sufficient diversity of antigen-binding sites only through the variation of side chains in hypervariable positions However, differences at the hypervariable loops in the germline genes were observed not only including amino acid substitutions but insertions and deletions that alter the loop length (1,8) Since the side chains and also the length of hypervariable loops are important in determining specificity (17,18), the presence of these insertion and deletion events can be understood in structural terms The former are due to an insufficient amount of possible surface topologies which will result from the mere variation of side chain types without altering (as well) the backbone of the hypervariable loops Thus, it can be proposed that not only amino acid substitutions but also length variations in hypervariable regions are important to determine the antibody repertoire and consequently have been subject to selection in the evolution of V genes

From the structural point of view the analysis of antibodies of known atomic structure has revealed a small number of main chain conformations or canonical structures for L1, L2, L3 as well as for H1 and H2 (19,20) This simple model has proven to be valid for almost all antibodies for which the three-dimensional structure has been reported at present (21) From the perspective of sequence analysis, a relevant feature of the canonical structure model is that the conformation of the loop is determined only by the loop length and the presence of certain residues in key positions in both the hypervariable loop and the framework regions (19), i e a loop that presents a defined length and adequate residues in a few key positions adopts a well-defined conformation According to this, it is possible through evolution to vary the canonical structure repertoire of the germline genes by altering the length and key residues of the hypervariable loops in order to generate a diverse antibody repertoire In fact, the analysis of functional germline genes (22–25), pseudogenes (26,27) and mature amino acid sequences (20,25–27) has revealed that almost all the sequences analyzed present canonical structures, indicating that the variations in length and sequence in key positions in the hypervariable loops have been constrained to maintain the canonical structures

In addition to those results our previous studies have shown that the combinations of canonical structures in five out of six hypervariable loops determine the antigen type which is recognized by the antibody (18) This result indicates that the conservation of the sequence patterns compatible with the canonical structure has got be an important selective force not only in terms of structural restrictions, but also in terms of the recognition properties of the antibody product of the V genes (18,28). According to this, it can be proposed that the diversity of the structural repertoire determined by the canonical structures has been subject to selective pressure in the evolution of V genes If this has occurred, it can be proposed that the evolution of the structural repertoire is correlated with the evolution of the V genes With the recent data available regarding the V_H and V_κ loci (21,29 31) it is possible to build complete phylogenetic trees and so to prove the above hypothesis.

In the present paper, in order to study the evolution of the structural repertoire of V genes of human Ig, we report the following analysis (i) An outline of the canonical structure model and its evolutionary paths are given (ii) The canonical structure repertoire is determined based on the germline gene sequences of the human V_H and V_κ available (iii) A gene tree is built for each V_H and V_κ locus, and the canonical structure class of each gene is assigned in the tree (iv) The phylogenetic relationships among the genes, families and clans are analyzed in terms of the structural repertoire (v) The primordial structural repertoires for V_H and V_κ loci are proposed

Canonical structures and evolution of V genes

In order to trace an evolutionary relationship among the canonical structures present at this time, it is necessary to analyze the common structural restrictions among the different conformations and the molecular evolutionary events required to convert one canonical structure type into another Here we briefly describe the main structural determinants for the hypervariable loops and outline the molecular evolutionary events (substitutions, insertions and deletions) to transform one canonical structure into another The sequence patterns and numbering scheme used here are as described by Chothia *et al* (20,22,25,32)

The first hypervariable loops of V_H and V_κ are both β-hairpin arms that link β-strands from different sheets (19) For all the canonical structures described for these loops, the presence of a hydrophobic residue in position 29 serves as an anchor of the loop that packs against a hydrophobic pocket This is formed by residues 24, 27 and 34 for H1 (19), and by residues 2, 32, 33 and 71 for L1 (19) The above-mentioned hydrophobic residue is the main stabilizing factor of the conformation Different lengths have been observed for these loops, but analysis of antibodies of known three-dimensional structure has shown that these extra residues form a bump which does not affect significantly the overall conformation of the loop (22) These insertions occur between residues 31 and 32 for H1 and 30 and 31 for L1 The molecular evolutionary events that have generated these length differences could have taken place among any other residues within the loop with the consequent disruption of the stabilizing pattern The placement of the insertion at the previously mentioned position maintains the key residues for all human germline genes (*cf* Table 1) This is a valid observation for mouse as well as for other vertebrate sequences (results not shown) This indicates that in most cases, the insertion events fixed during evolution were constrained to take place only at certain positions apparently to maintain the canonical structure pattern of these β-hairpin arms

According to this for instance, starting from type 1 of H1 the insertion of one or two residues between positions 31 and 32 and the conservation of the residue types in the determining positions 24, 27 and 34 are necessary in order to generate types 2 and 3 respectively

The second hypervariable loops of V_H and V_κ are both β-hairpin loops that connect β-strands from the same sheet

Table 1. Functional germline genes of the V_H locus

Locus[a]	H1-H2[b]	1 10 20 30ab 40 50 abc 60 70 80 abc 90
V_H1		....| ...|... |....| . ***..*|. . .*| ..| | ..| . * .**.. | .|. ..|* . |.. .| . . . |... |
1-2/VI-2 DP75	1-3	QVQLVQSGAEVKKPGASVKVSCKASGYTFTG--YYMHWVRQAPGQGLEWMGWINP--NSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
1-3/VI-3B/DP25	1-3	.... .S--.A .. ..R ......A--GN.N.K.S ... .T... .A........S ..E. .. ...
1-8/DP15	1-3	.... S--.DIN. ..T........ M..--. N.G.......... ..N..... . ...S.. E .. ..
1-18/DP14	1-2	.... S-- GIS .. .. SA--YN.N. . L .... T .T.... . RS ....
1-24/DP5	1-U	. ...V.. L E--LS. . ...K.....GFD.--ED.E.T............E...TD..... S ..E. .... T
1-45/DP4	1-3	.M.... . ..T.S ... ..Y--R L . .A .TT.--FN.N ......D.. T .R.M . ....S .E.. M ..
1-46/21-2/DP7	1-3	.. . . .S--...... . ...I...--SG.S.S........ ....T V.... S. .E . .
1-58/V71-5/DP2	1-3	.M..... P.... T...... . F.. S--SAVQ ..R..R. I...W--G N.. .. . E ..T.. M.T........S...E. .. .A
1-69/DP10	1-2	V. .... .S.. . . ..SS--.AIS.. ....G I.--IF TA...... . .T.A.E.T. ......S . E....
1-e/DP88	1-2	....S..... GT.SS-- AIS ....G.I.--IF.TA.. .. . .T.A.K.T.... .. S ..E.. ......
1-f/DP3	1-2	E ....T. I. .V .... D--......Q. ..K..... LVD.--ED.E.I..E. .....T.A.. TD...... S .E.... T
V_H2		
2-5/DP76	3-1	QITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGVGVGWIRQPPGKALEWLALIY---WDDDKRYSPSLKSRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR
2-26/DP26	3-1	.V S V... E . . V . .NARM. S...... ... . H.F---SN E.S..T.... . IS..T. S .T... ... RI
2-70/DP28	3-1	.V ...S..A . . .. MR S..... . ..R D---. ..F T ..T. .IS T . ....T .... .RI
V_H3		
3-7/DP54	1-3	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS--YWMSWVRQAPGKGLEWVANIKQ--DGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
3-9/DP31	1-3	....R. . .. . DD--.A.H. . ... SG.SW--NSGSIG.A... . .. . L... .KD
3-11/22-2B/DP35	1-3	Q ..... . K..... . . . . .D--.Y...I .. . ...SY.SS--S. TI A ....
3-13/13-2/DP48	1-1	.. ..--.D H . ..T. ..SA.G---TAGDT. PG ....E.. .... .G. . ..
3-15/9-1/DP38	1-U	....K.. .. .. .N--A. . ... .GR. .SKT. .GTTD.AAP. .. ...DS T.........KT. ..... TT
3-20/DP32	1-3	..... . .V.R... .... DD--.G... . ..... .SG.NW--N.GSTG A...... .. .. L.H..
3-21/WHG16/DP77	1-3	. . .. K...... . ....--.S N..... . ....SS.SS--SS.YI A . ...
3-23/VH26/DP47	1-3	....L .... --.A... . .... SA.SG--S GST. A ....S .T . . .. K
3-30/1 9III/DP49	1-3	Q .... .V ..R . .... .-- G.H . .. .V SY--...N A . .. . S..T . . .K
3-30.3/GL-SJ2/DP46	1-3	Q. . .. .V . R...... . . .--.A H ..... . .V.SY--. N ..A... . .. ...S .T...... . ...
3-30.5/DP49	1-3	Q.... . .V... R. . ...--.G.H. .... .V SY--. N A .. .S. T . . . . . . .K
3-33/3019B9/DP50	1-3	Q.... . .V.. R. . . .. . .--.G.H . . .. .. .V WY--...N A .... S..T.. . . . .
3-43/DP33	1-3	.... W... ....DD-- T.H........ .SL.SW--. GST. A . . S . . T. ..L. KD
3-48/DP51	1-3	.. . . -- S.N . ...SY SS--SS.TT..A ... .D. . ...
3-49	1-U	. .. . R.... .T . ..GD-- A. .F.... . ...GF.RSKAY GTTE.TA . .GS STA.. KT. . . .T
3-53/DP42	1-1	.... T ...I. . . . V --NY.... ...SV.Y---S.GST..A . .... . S..T.. ... .
3-64/DP61	1-3	. ...-- A.H....... .. .Y.SA.SS--N GST..A. . . .. . .S .T . . G. . .
3-66/Yac5/DP86	1-1	.... V --NY.. . ...SV.Y---S.GST..A .. .. S..T....... ....
3-72/12-2/DP29	1-4	. ...D--HY.D...... ......GRTRNKANSYTTE.AA . ..... DS . .... KT .
3-73/Yac9	1-4	.. ..K. ..G--SA.H S .. .. GR.RSKANSYATA.AA..... . .DS..TA. .....KT. . . T.
3-74/H11/DP53	1-3	. ... --...H.. .. ...V..SR.NS--...STS.A . ... ..T.. . .
3-D/COS12	1-6	.. ....R.V. .. .. V --NE .. . .. SS.S---- GST..A..R. . . ...S .T H .... KK
V_H4		
4-4/DP70	2-1	QVQLQESGPGLVKPSETLSLTCAVSGGSISSS-NWWSWVRQPPGKGLEWIGEIY---HSGSTNYNPSLKSRVTISVDKSKNQFSLKLSSVTAADTAVYYCAR
4-28/V12G-1/1.9II/VH4.13/DP68	2-1	.... D ... Y . .S- G I ..... Y. ---Y . Y..........M. .T.... . .V . .
4-30.1/DP65	3-1	....Q . . .T.. . GGYY. I H. . .. Y ---Y . Y ... T .
4-30.2/DP64	3-1	L . ..S Q . ...GGYS I ... .. Y..--- .. Y.... .. ...R . . ...
4-30.4/VH4.34/DP78	3-1	.... Q ... T . ...GDYY I . . . .. Y. ---Y Y . ..T.. .. . . .
4-31/DP65	3-1	... .Q T... GGYY...I H. ......Y. ---Y....Y ..... . T ... .
4-34/VH5/VH4-21/DP63	1-1	....QW.A. L .E .. . .Y ..F.G--YY .I.. . ....N--- ....T ... . .
4-39/VH4.18/DP79	3-1	.L .... .E . T .. .. SSYY G I. .. . .S..---Y. ..Y . . T. . .. .
4-59/VH4.11/DP71	1-1	. ...E ...T.. .. .--YY.. I .. . .. Y..---Y. . . .. T ... .
4-61/V71-2/DP66	3-1	....E ..T... V GSYY...I. .....Y ---Y.... . ... T ....
4-b/DP-67	2-1	. .. . .E . ...Y. .G-YY.G.I... .....S .--- ...Y. . . ..T. .. . .
V_H5		
5-51/VH251/DP73	1-2	EVQLVQSGAEVKKPGESLKISCKGSGYSFTS--YWIGWVRQMPGKGLEWMGIIYP--GDSDTRYSPSFQGQVTISADKSISTAYLQWSSLKASDTAMYYCAR
5-a/VH32	1-2	.. . . R . . . . .-- N S...... . . R D.--S Y N..... H . .. .
V_H6		
6-1/VH-VI/6-1G1/DP74	3-5	QVQLQQSGPGLVKPSQTLSLTCAISGDSVSSNSAAWNWIRQSPSRGLEWLGRTYYR-SKWYNDYAVSVKSRITINPDTSKNQFSLQLNSVTPEDTAVYYCAR
V_H7		
7-4 1/DP21	1-2	QVQLVQSGSELKKPGASVKVSCKASGYTFTS--YAMNWVRQAPGQGLEWMGWINT--NTGNPTYAQGFTGRFVFSLDTSVSTAYLQICSLKAEDTAVYYCAR

[a] Locus is presented clustered by families Within each family the sequences are ordered following the physical location of genes from D distal to proximal (31)

[b] Determined canonical structure class according to Table 1 (H1 and H2) Canonical structures were determined using the VIR package (33) 'U' denotes uncertain structure Key positions for canonical structures are labeled with '*' in the corresponding column The gene 1-24 has the length and sequence pattern of type 2 in H1, except Glu in position 71 which has not been observed in three-dimensional structures (22) Genes 3-15 and 3-49 have the length and sequence patterns of type 4 in H2, but Gly in position 55 instead of Tyr (19) For these three genes this makes the prediction of the conformation difficult Type 6 of H2 corresponds to a conformation for hairpin loops of length two as defined by Chothia *et al* (19)

(19) The variation both in length and residues is strongly restricted for L2 (19 25) One canonical structure has been defined for this loop (19) At present, almost all human and mouse sequences reported are compatible with this structure (25,27,28) In human V_κ germline genes all sequences present the canonical structure sequence pattern For H2, however significant variation has been observed (14,19). The main determinants of the conformation of this loop are the length and the presence of residues that can adopt special dihedral angles in positions 54 or 55 Structural (19) and sequence (34) analyses indicate that the insertions take place between positions 52 and 53 for all human (*cf* Table 1) mouse and

other species sequences (results not shown) Thus, for this loop as well as for H1, insertion events were fixed only when they occurred between residues 52 and 53 in almost all Ig

To transform one canonical structure into another, the insertion or deletion of one or more residues between positions 52 and 53, and substitutions in positions 54, 55 and 71 are required

The third hypervariable loop of V_κ is a β-hairpin loop that links β-stands from the same sheet (19) The analysis of this loop is more complex than the others because the recombination events between V_κ and J_κ might alter the length and sequence of the loop An analysis of rearranged human sequences shows that ~10–15% of the sequences result in a modified length of L3 (25,35) Considering this, we will analyze the canonical structures by assuming that the joining process will not alter the length of the loop. The main determinants for these conformations are the presence of Pro in position 94 or 95, which imposes special constraints to the loop structure, and the presence of Gln, His or Asn in position 91 forming a hydrogen bond network, which stabilizes the conformation of the loop (19). For almost all the germline gene sequences for both human and mouse, but not for the other vertebrates (results not shown), Pro is conserved at position 95 and Gln, His or Asn appears at position 91.

For example, starting from type 1, which is the most common canonical structure of human and mouse Ig (20), type 2 can be generated through substitutions that change the placement of Pro from position 95 to 94 (*cf* Table 2)

For all loops in general, it is concluded that key residues have been preserved and the placement of insertions/deletions has been strongly restricted in order to avoid disrupting the whole stabilizing sequence pattern As mentioned before, this has occurred for Ig genes in most vertebrates.

Results

Evolution of V_H genes and the structural repertoire

The physical map and sequences of the complete V_H locus in humans are now known (31) In Table 1 we present an alignment of all the functional germline genes of V_H. The canonical structure classes as defined by Chothia *et al.* (22) are determined for each gene and reported in the second column in Table 1 (see Table 1 footnote for details).

To analyze the evolutionary relationships between genes and gene families, we built a gene tree based on the amino acid translations of the 51 functional germline genes of V_H (Table 1). The evolutionary relationships among genes were analyzed considering the canonical structure class encoded by each gene. The phylogenetic tree was built using the distance matrix method (Protdist program) and applying Fitch's method for tree construction available in the Phylip package (37) An additional tree was built using the neighbor method giving identical topology to it (results not shown) In order to avoid the influence of length differences in hypervariable loops on the tree topology the regions with insertions/deletions (positions 31a and b for H1, and 52a, b and c for H2) were not considered for the analysis In Fig 1, a gene tree for the V_H locus is presented in the form of a phenogram with lengths of branches proportional to the number of differences among groups The canonical structure class encoded by the gene is reported adjacent to each one

V_H gene families and the structural repertoire In the tree introduced in Fig 1, seven well-established families for this locus (31) are observed and labeled (see the Fig 1 caption) The relevant information of Fig 1 is summarized in Table 3 In Table 3, the classes they encode, the clan they belong to and the number of functional genes are reported for each family Based on this information, the evolution of the structural diversity of the locus is analyzed.

As we can see in Fig 1 and Table 3, the V_H3 family is the principal contributor to the diversity of the structural repertoire supplying four different classes. This family encodes 1-X form classes, where X runs from type 6 (the shortest one) to type 4 (the longest one described) (22). This means that the evolutionary diversification has only taken place in H2 In Fig 2(a), the three-dimensional structures of V_H domains of antibodies with the superimposed classes 1-1, 1-3 and 1-4 are presented (38). At present, the three-dimensional structure of antibodies with class 1-6 has not been reported. As can be seen from Fig 2(a), a great diversity in the shape of the binding site can be generated by different types of H2.

V_H1 and V_H4 families encode for two and three classes respectively The family V_H1 encodes for structure types 2 and 3 for H2 These two types imply the same length, but differences are present at position 71 which alter the conformation of the loop (22). In Fig. 2(b), V_H domains of antibodies with classes 1-2 and 1-3 are shown. As can be seen, the shape of the binding site is not altered significantly by the change from type 2 to type 3 in H2 Consequently it is concluded that V_H1 has developed a very homogeneous structural repertoire through evolution. On the contrary, the V_H4 family encodes for three classes that present a type variation in H1 (X-1 general form) The V_H4 family is the only one that encodes for type 2 in H1 Loops with types 2 and 3 are one or two residues longer than type 1 respectively (20) In Fig. 2(c), V_H domains of antibodies with classes 1-1, 2-1 and 3-1 are superimposed. Types 2 and 3 in H1 modify considerably the general shape of the binding site when compared with class 1-1 According to this in terms of quantity and quality of the classes, V_H4 contributes to the antibody repertoire to a greater extent than V_H1

Within the small families, V_H2 and V_H6 are other families, in addition to V_H4, that encode for classes of different type to 1 in H1 (*cf.* Table 3). Both families encode for one class each The V_H2 family presents class 3-1 which is also encoded by the V_H4 family The V_H6 family encodes for class 3-5 which seems to be a variation of 3-1. These observations indicate a relationship from the structural point of view among the families V_H2, V_H4 and V_H6

It is worth mentioning that class 3-5 is the only one that combines long loops in both H1 and H2 (*cf.* Table 1) Classes 1-1 and 1-6 combine short loops in H1 and H2, whereas classes 1-2, 1-3, 1-4, 2-1 and 3-1 combine a short loop with a long one Unfortunately, no three-dimensional structure of an antibody for class 3-5 is available at present

The small families V_H5 and V_H7 both present the canonical structure class 1-2 which is also encoded by the V_H1 family The structural repertoire of these three families is very homo-

Table 2. Functional germline genes of the V_κ locus

Locus[a]	L1-L2-L3[b]	1 10 20 30bcdef 40 50 60 70 80 90
$V_\kappa1$		* | ...|.. | ..| ..*. *|. .. ...**|... |. ..|. * | .| . .|.. *|. |* ..|....| . |.. .*. ..*
L24/V13/Ve''/DPK10	2-1-1	VIWMTQSPSLLSASTGDRVTISCRMSQGIS------SYLAWYQQKPGKAPELLIYAASTLQSGVPSRFSGSGSGTDFTLTISCLQSEDFATYYCQQYYSFP
L23	2-1-1	A.R . FS ...V. . T.WA. . ------ .. ..A...K.F .Y S....... .. . .Y. S..P.... ..T.
L19b/Vb''/DPK6	2-1-1	D Q ... SV. .V..... T. A .,------ W. .... . . K.... S . .. . S .P .. .... AN .
L18/Va''	2-1-1	A QL. .. S . V . T .A ,------ A. . .. .K ...D. S E... .. . .S .P FNNY.
L15/HK101/HK146/HK189/DPK7	2-1-1	D Q .. S.. V . T .AR .------ W ..E.. KS. . ..S . ...S. P. . ...N.Y.
L14/DKP2	2-1-1	N Q . .AM.. V . T .AR .------N F .. V KH. . ..S. . . .. . . E ....S P. . L HN.Y.
L12/HK102/VI	2-1-6	D.Q . T .V... T A. S ------ W . K. D .S.E. . ...E . .S..PD. . . N.YS
L11/vf/DPK3	2-1-1	A.Q . . S. . V . T A . R------ND.G .....K. . .S...,. . . ... . .S. P .. .. L D.NY
L9/Ve	2-1-1	A.R. . SF .. .. T A. . .------.. .. .. ..K... . . . . Y
L8/Vd/DPK8	2-1-1	D.QL. . F. ..V..... T A.... ------..... ....K... .. .. .E......S..P .. .... LN.Y
L5/Vb/Vb'/Vab/DPK5	2-1-1	D.Q . SV...V. . T A . . ------.W.. .. ....K... S..... . ....S .P . .... AN.
L4/Va'/V4a	2-1-1	A.QL. . .S ...V. . .T .A.... ------.A.. . .. .K.. .D. S E.... . . ..S .P.... . . FNNY
L1/HK137	2-1-1	D.Q.. S V. T A... ------N . F . KS... . S . . . . .. .S .P ... N Y
O8/DPK1	2-1-1	D.Q.. . .S ...V. .. .T QA..D..------N..N.... .. .K ...D..N ET . ... F ..S..P. I . . DNL.
O2/DPK9	2-1-1	D.Q S . .V T A. S ------ .N ...K . S ... . . . S P... .. S T
O12/V3b/DPK9	2-1-1	D.Q. ...F.. .V..... T A. S..------...N.. . ....K... S . ...S..P . ....G T
O18/DPK1	2-1-1	D.Q.. .. S V . T QA D..------N..N . . K ...D N ET .. .F . S..P..I DNL.
A20/DPK4	2-1-U	D.Q . ...S....V. T..A . .------N .....V.K.. ....S..P..V .. K.N.A.
A30	2-1-1	D.Q .S .V. . T A . .R------ND G . . ... KR.. S . . .E . . S. P L HN Y
$V_\kappa2$		
A3/A19/DPK15	4-1-1	DIVMTQSPLSLPVTPGEPASISCRSSQSLLH-SNGYNYLDWYLQKPGQSPQLLIYLGSNRASGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQALQTP
A2/DPK12	4-1-1	. T . S .Q..... K .-.D KT .Y . .P . EV .F . . .. SI.L.
A1/DPK19	4-1-1	.V . ... ..L Q. .. VY-.D NT..N.FQ.R ....RR...KV..WD. . ... GTHW.
O1/DPK13/DPK13	3-1-1	T . .. . .. DSDD NT.. .. . .TL Y .. . . RIEF.
O11/V3a/DPK13	3-1-1	... .T . ... DSDD NT...C . . . TL Y . ..D . ... . .RIEF.
A17/DPK18	4-1-1	V L Q VY-.D NT .N FQ.R RR .KV D .... . . GTHW.
A18/DPK28	4-1-1	. . .T .S .. Q. K . . - D KT..Y . . . . ..EV S F .. GIHL
A19/DPK15	4-1-1	. . . - .. . . S . . . . .
A23/DPK16	4-1-1	. .T. S .L.Q. ...V.- D NT..S LQ.R. P.R ..KI. .FS . . .A . . T.F.
$V_\kappa3$		
L25/DPK23	6-1-1	EIVMTQSPATLSLSPGERATLSCRASQSVSS-----SYLSWYQQKPGQAPRLLIYGASTRATGIPARFSGSGSGTDFTLTISSLQPEDFAVYYCQQDYNLP
L20/Vg''/humkv3g''	2-1-6	L . . G ------ A . D N .. .P E RS WH
L16/humkv328/humkv328h2	2-1-1	. .. V . .. ------ N A. .. . . E. S .. . .YN W
A11/humkv305/DPK20	6-1-1	L G ----- A L D S D . R E. YGSS
A27/humkv325/VKRF/DPK22	6-1-1	.L G. .. .. ----- ..A.... .. . .S .D.. .. . . . R E. . . YGSS.
L2/humkv328h5/DPK21	2-1-1	V . . .. .. ------ N A . . . E. . . S YN W.
L6	2-1-1	L . . . . ------. A ... . D N . . . .E. ..RS W
$V_\kappa4$		
B3/vkIv/DPK24	3-1-1	DIVMTQSPDSLAVSLGERATINCKSSQSVLYSSNNKNYLAWYQQKPGQPPKLLIYWASTRESGVPDRFSGSGSGTDFTLTISSLQAEDVAVYYCQQYYSTP
$V_\kappa5$		
B2/EV15	2-1-1	ETTLTQSPAFMSATPGDKVNISCKASQDID--- --DDMNWYQQKPGEAAIFIIQEATTLVPGIPPRFSGSGYGTDFTLTINNIESEDAAYYFCLQHDNFP
$V_\kappa6$		
A14/DPK25	2-1-1	DVVMTQSPAFLSVTPGEKVTITCQASEGIG------NYLYWYQQKPDQAPKLLIKYASQSISGVPSRFSGSGSGTDFTFTISSLEAEDAATYYCQQGNKHP
A10/A26/DPK26	2-1-1	EI.L . D Q. K... .R QS --- --SS H .. S F . ... N . . . H SSSL
A26/DPK26	2-1-1	EI L . D Q K . R QS ------SS H .. S . ... . F . ...N. H.SSSL
$V_\kappa7$		
B1	5-1-1	DIVLTQSPASLAVSPGQRATITCRASESVSFLGI--NLIHWYQQKPGQPPKLLIYQASNKDTGVPARFSGSGSGTDFTLTINPVEANDTANYYCLQSKNFP

[a] Locus is presented clustered by subgroups or families Within each family the sequences are ordered following the physical location of genes from J_κ distal to proximal (35)

[b] Determined canonical structure class according to Chothia *et al* (19) Canonical structures were determined using the VIR package (33) 'U' denotes uncertain structure Key positions for canonical structures are labeled with '*' in the corresponding column L24 has all the features of type 2 in L1, except in key position 25, which presents Met instead of Ala or Ser The gene A20 presents Lys in position 90 instead Gln His or Asn The prediction of conformation for these loops is uncertain Type 6 of L3 corresponds to the conformation defined by Guarné *et al* (36)

geneous. This structural analysis indicates a relationship between this group of families

It is important to mention that the gene tree presented here was built without considering the insertion/deletion regions within the hypervariable loops, which avoids the fact that information regarding loop length variations affects the tree topology As mentioned above, loop length differences are the main determinant for canonical structure type Therefore, the fact that the same family genes encode for the same or similar classes indicates a correlation between the structural repertoire and the evolution of genes

In general, the evolutionary analysis of the V_H families' structural repertoire has shown that classes are not randomly distributed within the families. Thus, it is important to determine if these common structural features are also correlated with the established evolutionary relationship among gene families (5)

V_H family clans and the structural repertoire. Previous studies have shown that the seven families of the human V_H locus can be grouped in three ancestral clans (5,6,31) This clan organization is common for the mammalian V_H gene families (5,6,31,45) These three clans can be clearly identified as the three main branches labeled with Roman numbers in Fig 1 (cf Fig 1 caption) This information is summarized in Table 3 (third column)

In the previous section it has been noted that there are some classes encoded by only one family and others which are common to various families Members of the same clan encode for classes of similar structural characteristics (*cf* Fig 1 and Table 3) Clan I only encodes for the 1-X form classes (classes 1-2 and 1-3) As previously mentioned, the presence of these classes implies a very homogeneous structural repertoire (*cf* Fig 2b) Clan II encodes for X-1 form

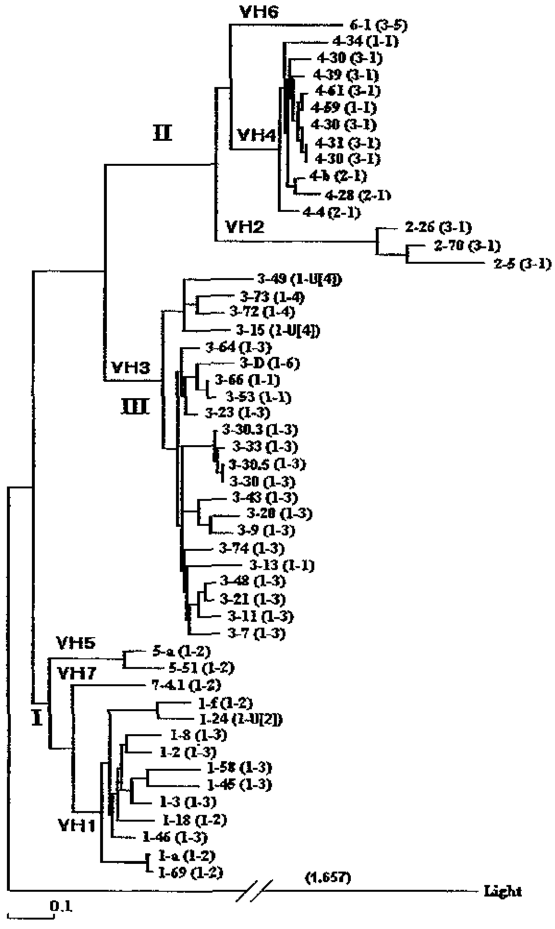


Fig. 1. Gene tree for V_H functional germline genes The three clans found for V_H are labeled with Roman numbers. Families defined for V_H are labeled Canonical structure classes encoded by the gene are indicated When it was not possible to assign a canonical structure type, the structure type which is related to the sequence of the loop is indicated in brackets.

Table 3. Structural repertoire of V_H locus

Family	Canonical structure class	Clan	No of functional genes
V_H1	1-2, 1-3	I	11
V_H5	1-2	I	2
V_H7	1-2	I	1
V_H2	3-1	II	3
V_H4	1-1, 2-1, 3-1	II	11
V_H6	3-5	II	1
V_H3	1-1, 1-3, 1-4, 1-6	III	22

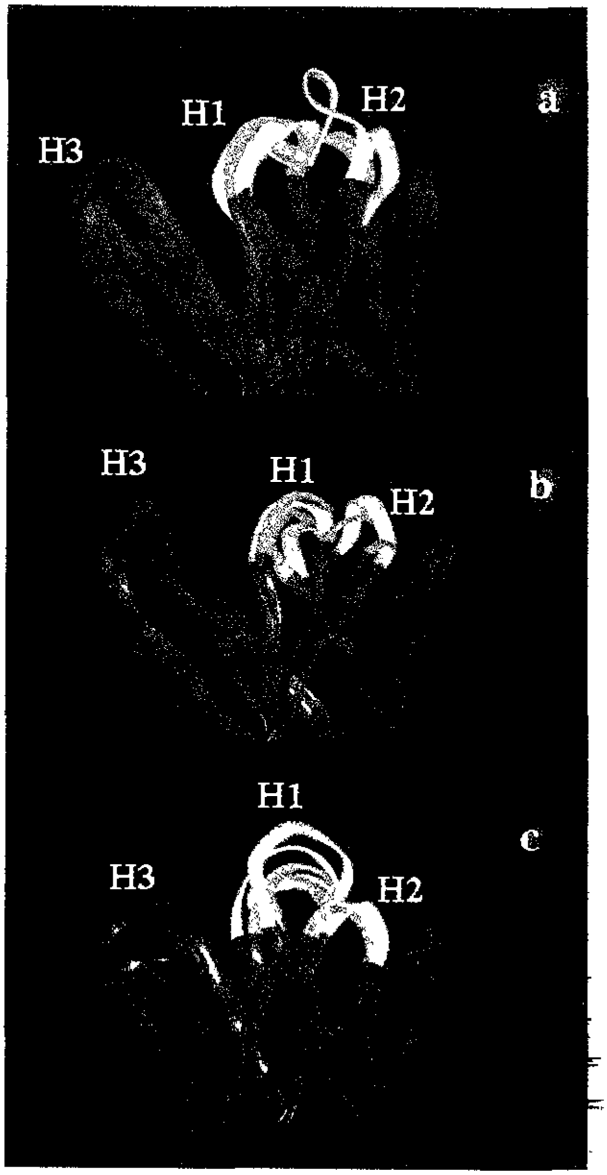


Fig. 2. Ribbon representation of antibody backbones with representative V_H canonical structure classes The V_H domains are superimposed by the framework according to the definitions of Chothia and Lesk (19). Framework and H3 are in red for all the antibodies, the hypervariable loops are colored to mark different canonical structure classes (a) Comparison among classes 1-1 [green, 1VFA (39)], 1-2 [yellow, 2HFL (40)] and 1-4 [white, 4FAB (41)]. (b) Comparison between classes 1-2 [yellow, 1BBD (42)] and 1-3 [green, 2FBJ (43)] (c) Comparison among classes 1-1 [green, 1VFA (39)], 2-1 [white, 1BAF (41)] and 3-1 [yellow, 1GGB (44)]

classes (classes 1-1, 2-1 and 3-1, except for gene 6-1 which encodes class 3-5) The presence of these four classes in clan II implies a great structural diversity (*cf.* Fig. 2c) It is clear that clan II has developed a structural repertoire significantly more diverse than clan I Clan III presents 1-X form classes (classes 1-1, 1-3, 1-4 and 1-6) This clan is exclusively formed by the large V_H3 family and generates a diverse structural repertoire with four different classes which present a significant variation in the length of H2 (*cf* Fig 2a)

During the analysis of common features among clans, it is observed that class 1-3 is encoded by clan I and III members, and that class 1-1 is common for clan II and III members This inter-clan identity in the structural repertoire is interesting if we consider the evolutionary distance among members of different clans, which suggests that recombination or convergent evolutionary events—among members of different clans—have taken place in the evolution of the locus.

Primordial structural repertoire of V_H Based on the observed correlation between the evolution of the germline genes and the structural repertoire it is possible to propose the most probable ancestral classes for each clan However, since it has been demonstrated that the clans identified in human are also present in mouse and other vertebrate species (6,9,15,45), the proposed ancestral repertoire must be supported by consistent results for these species. Here we present such comparative analysis. Figure 3 shows the resulting primordial structural repertoire from this analysis. The molecular evolutionary events necessary to inter-transform the classes are presented in Fig 3.

A comparative analysis of the corresponding members of clan I in human, mouse and other vertebrates gives the following results (i) Class 1-2 is the common class encoded by genes from human, mouse and other vertebrates (results not shown). (ii) Class 1-2 also is the most frequently encoded in these species (iii) Class 1-2 is present in both the large and the small families of human, and a similar analysis for mouse gives the same result. Based on these observations, it can be proposed that the clan I ancestral gene codified for class 1-2 The remaining genes from human V_H1 diverged from class 1-2 through a single evolutionary event (substitution in position 71) to generate class 1-3 (*cf.* Fig 3)

For clan II, the comparative analysis shows the following (i) Classes 1-1, 2-1 and 3-1 are present in the three groups of sequences (human, mouse and vertebrates). (ii) The most frequently encoded class is 3-1 for human, whereas for mouse and vertebrates the most frequent one is 1-1 Such results do not entitle us to certainly propose an ancestral class However, in the case of human, considering that class 3-1 is present in V_H2 and V_H4 families, and class 3-5 from the V_H6 family appears as a modification of the 3-1 one (*cf* Fig 1), we suggest class 3-1 as the ancestral class.

The following results were obtained for clan III (i) Classes 1-1, 1-3 and 1-4 are present in human, mouse and vertebrates, 1-6 is characteristic of human (ii) Class 1-3 is the most frequently encoded in the three groups of sequences These results shows a high consistency in the evolution of the structural repertoire for this clan for all vertebrates This allows us to propose with high confidence that the ancestral class for this clan was 1-3

The previous analysis suggests an outline for the early structural repertoire that encoded the three primordial V_H genes responsible for the generation of the three clans found This primordial repertoire is outlined in Fig 3 It is proposed that ancestral genes in clans I, II and III encoded for classes 1-2, 3-1 and 1-3 respectively (*cf* the three main branches in Fig 3) Due to the nearness in the sequence pattern between types 2 and 3 in H2 (*cf* above), we propose that the primordial genes in clans I and III have a common ancestry. The evolution from the primordial class to the present class is indicated with arrows in Fig 3 The legend to Fig 3 defines the molecular evolutionary events involved in each process. It is worth mentioning that in all cases, the present classes can be obtained by applying a reduced number of molecular evolutionary events to the ancestry class

Evolution V_κ genes and the structural repertoire

Here, in the same manner as for V_H, we analyze the V_κ structural repertoire evolution through building a gene tree with functional V genes (*cf.* Table 2) and assigning the gene canonical structures in the tree (*cf* Fig 4). In the analysis, the type that occurs in L2 is not taken into consideration since all genes present the type 1 canonical structure (*cf* second column of Table 4). Consequently, this loop does not contribute to the locus structural diversity The tree construction methodology is the same as the one used for V_H. The insertion segments in L1 of V_κ sequences (positions 30a–30f) were not considered in the construction of the tree. The relevant information of the V_κ gene tree is summarized in Table 4. The encoded classes, the clan to which they belong to and the size of each family is reported in second, third and fourth columns in Table 4 respectively

V_κ gene families and the structural repertoire. At present and to our knowledge, no systematic evolutionary studies of the complete human V_κ locus have been published in an extensive article (46). Only comparative studies between human and mouse V_κ germline genes have been reported (47) The seven gene families—usually termed subgroups—in which the V_κ germline genes can be grouped have been well established (25,30) The gene tree of functional V_κ germline genes is shown in Fig 4. The seven V_κ families in the built tree are clearly observed and labeled (*cf.* Fig 4 caption) Members of only six out of these seven families have been observed in rearranged sequences and are considered as functional germline genes (25) The B1 gene is the single member of $V_\kappa7$ family and has never been observed in rearranged gene sequences. Consequently, it is not considered functional (25) We have included this gene in the analysis because is representative of another family and, as shown later, it encodes for a unique canonical structure class.

For V_κ, the family that contributes the most to the structural repertoire diversity is $V_\kappa3$ with three canonical structure classes This family is the only one that encodes for classes with different canonical structures in both L1 and L3 (classes 2-1, 2-6 and 6–1). Also, this family is the only one that encodes for class 6-1. In L1, type 6 has one more residue than type 2 between positions 30 and 31 (20), and consequently that might alter the shape of the binding site Types 1 and 6, in L3, present the same V gene length, though type 6 lacks Pro in position 95 which is a determinant for the type 1 canonical structure (19) V_κ domains with classes 2-1 and 6-1 are shown in Fig 5(a) Ig structures with class 2-6 are not available at present in the Protein Data Bank The difference of types in L1 produces some modifications in the shape of the binding site However, this is not as major a difference as those

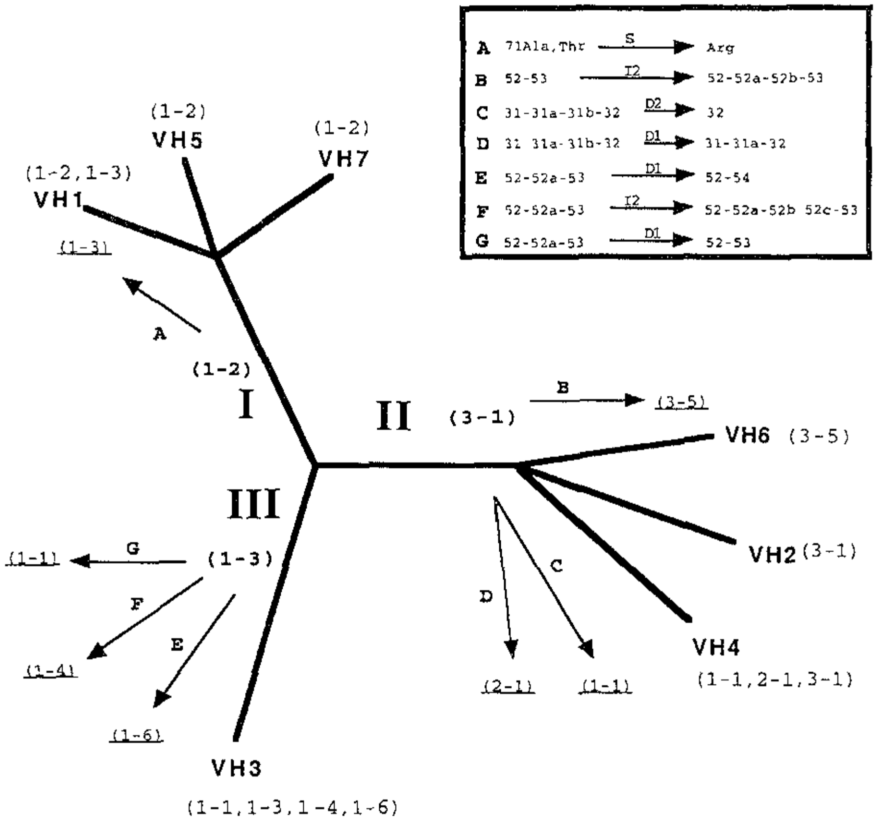


Fig. 3. Proposed evolutionary relationships among canonical structure classes of V_H families The three V_H clans are labeled with Roman numbers The proposed canonical structure class for the ancestral gene for each clan is in bold in parentheses, the classes that have appeared through the evolution of the structural repertoire are underlined in parentheses For each family the present structural repertoire is indicated The upper case bold letters behind the arrows indicate the transformations (molecular evolutionary events) to convert the ancestral class into the present class The legend presents the molecular evolutionary events. Three categories of events are proposed and are represented by letters, (i) S, substitution, the position in which the mutation must take place is indicated and examples of the amino acids are given (ii) I, insertion of residues in specific position, e.g in the transformation rule 'B' the event I2 means an insertion of two residues between positions 52 and 53. (iii) D, deletion of residues at specific positions, e g in the transformation rule 'C' the event D2 means a deletion of two residues between residues 31 and 32

observed when short loop genes are compared to long loop ones in L1 (see below)

The other large families ($V_\kappa1$ and $V_\kappa2$) encode for two canonical structure classes each Almost all genes in the largest $V_\kappa1$ family encode for class 2-1 except for class 2-6 as the only variation in this family (the gene L12 presents a substitution of Pro by Ser in position 95) This preserved pattern in the evolution of the largest V_κ family (19 functional genes) contrasts with the case of the largest V_H family (V_H3 with 22 functional genes) which presents a diverse structural repertoire (*cf* Tables 3 and 4) The encoded classes by $V_\kappa1$ are also present in families $V_\kappa3$, $V_\kappa5$ and $V_\kappa6$ This similarity in the structural repertoire suggests that these families have an evolutionary relationship The tree topology supports this fact (*cf* Fig 4). The structural repertoire diversity in the other large $V_\kappa2$ family is also restricted Classes 3-1 and 4-1 are structurally similar. The only difference between these classes is one insertion in L1 (20) V_κ domains of antibodies with classes 3-1 and 4-1 are presented in Fig. 5(b) The difference of one residue in length slightly alters the shape of the binding site However, the most relevant feature is the projection of L1 towards the solvent in both classes. Other families which present classes with long loops in L1 are $V_\kappa4$ and $V_\kappa7$ These three families ($V_\kappa2$, $V_\kappa4$ and $V_\kappa7$) with very similar structural repertoires do not appear close to each other in the tree topology, which suggests that gene conversion events have occurred among these families (*cf* Fig. 4).

The small V_κ families encode for classes which have already been encoded by large families (*cf* second and fourth columns in Table 4). The exception to the rule is the single member in family $V_\kappa7$ which as mentioned before encodes for a unique class (class 5-1) However, this is apparently a non-functional gene, i e. in terms of the structural repertoire the small functional V_κ families are redundant respect to the large families

An important feature of the evolution of the V_κ structural

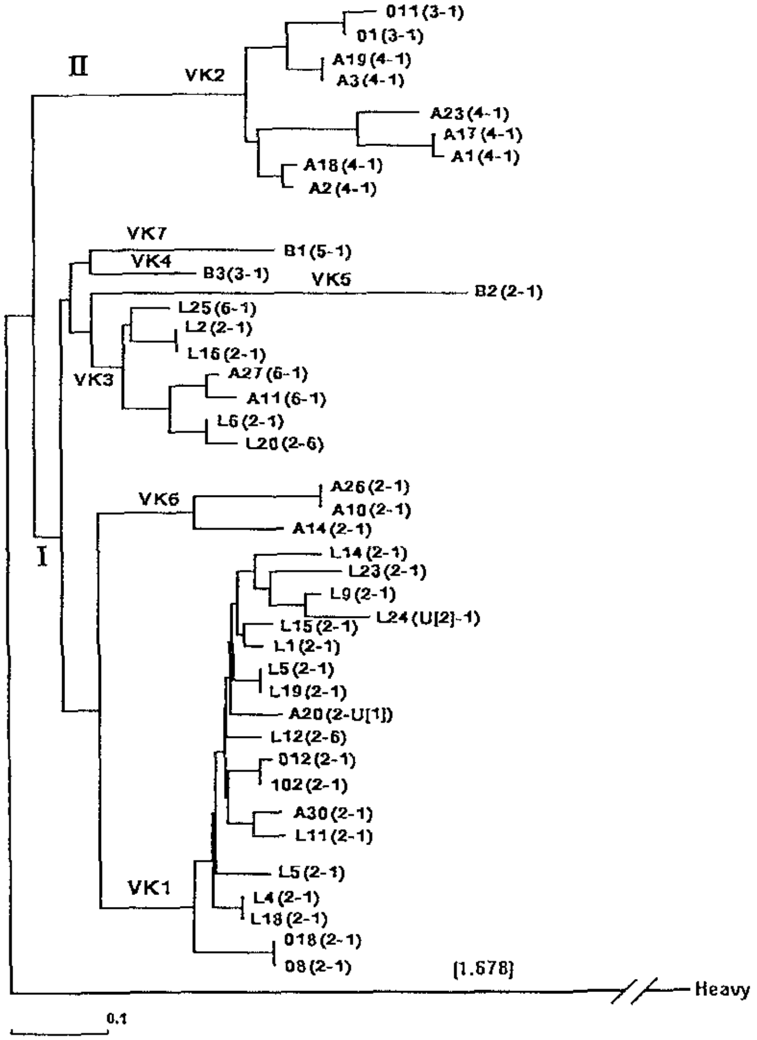


Fig. 4. Gene tree for V_κ functional germline genes. All the conventions as in Fig. 1

Table 4. Structural repertoire of V_κ locus

Family	Canonical structure class	Clan	No. of functional genes
$V_\kappa 1$	2-1, 2-6	I	19
$V_\kappa 3$	2-1, 2-6, 6-1	I	7
$V_\kappa 4$	3-1	I	1
$V_\kappa 5$	2-1	I	1
$V_\kappa 6$	2-1	I	3
$V_\kappa 7$	5-1	I	1
$V_\kappa 2$	3-1, 4-1	II	9

repertoire is that most of the V_κ structural diversity is contributed by L1. The distinctive feature which differentiates the canonical structure types in L1 is the length. Considering the types present in human V_κ, L1 types can be grouped as long (types 3, 4 and 5) and short (types 2 and 6). From this point of view $V_\kappa 2$, $V_\kappa 4$ and $V_\kappa 7$ are families that encode for long loops in L1, and $V_\kappa 1$, $V_\kappa 3$, $V_\kappa 5$ and $V_\kappa 6$ encode for short loops in L1 (*cf.* Table 4). In Fig. 5(c), the V_κ domain of antibodies with class 2-1 is presented as a representative for classes with a short loop in L1 and 3-1 as a representative for classes with a long loop in L1. The difference of the shapes generated by these two classes is huge. These two groups of loops in L1 can be combined with loops of V_H to form binding sites with flat, pocket or mixed forms (18,28,52,53).

Considering the results obtained here for V_κ families, it is worth mentioning that there is an important difference between V_H and V_κ families in the sense that V_H genes of the same family encode for classes with different shapes (*cf.* $V_H 3$ and $V_H 4$ families in Table 3), but for V_κ, classes within the families

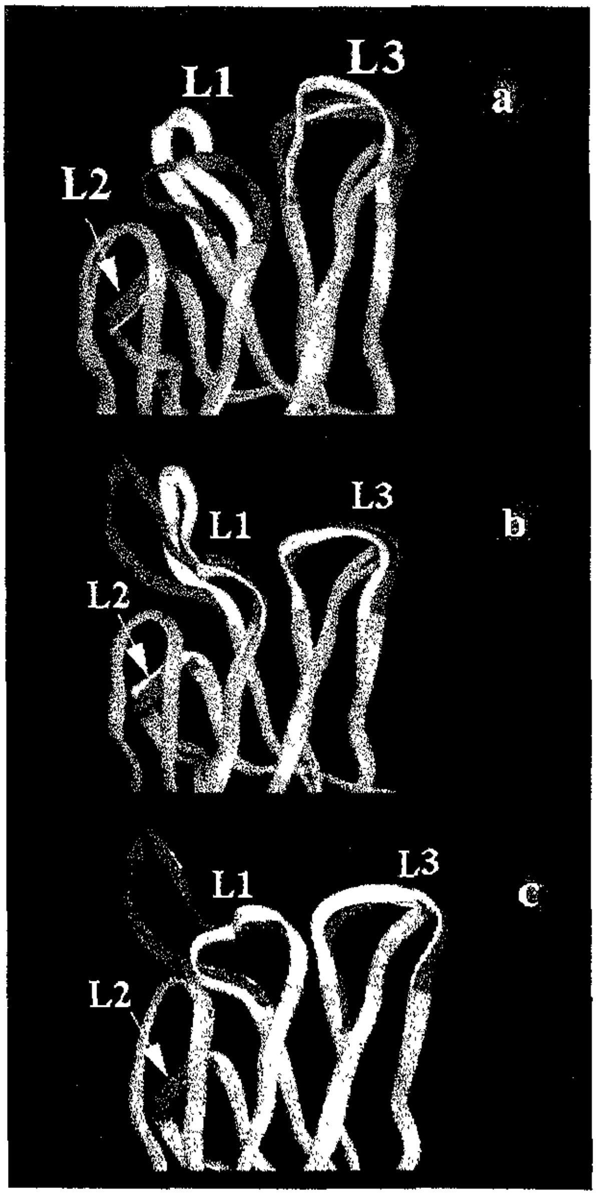


Fig. 5. Ribbon representation of antibody backbones with representative V_κ canonical structure classes. The V_κ domains are superimposed by the framework according to the definitions of Chothia and Lesk (19). The framework is in orange for all the antibodies, the hypervariable loops are colored to mark different canonical structure classes (a) Comparison among class 6-1 [green, 1FIG (48)] and class 2-1 [red, 1IGM (49)] (b) Comparison among class 3-1 [red, 2MCP (50)] and class 4-1 [green, 4FAB, (41)] (c) Comparison among class 3-1 [red, 1MCP (51)] and class 2-1 [green, 1BAF (41)]

are very similar The large $V_\kappa 2$ and $V_\kappa 3$ families are the only ones that have genes with loops different in length, but length variations are only of one residue and do not change significantly the binding site shape (*cf* Fig 5a and b) Conversely, e.g. in V_H, V_H3 presents genes with structure type 6 and others with structure type 4 in H2—this signifies a difference of four residues in length (*cf* Table 1).

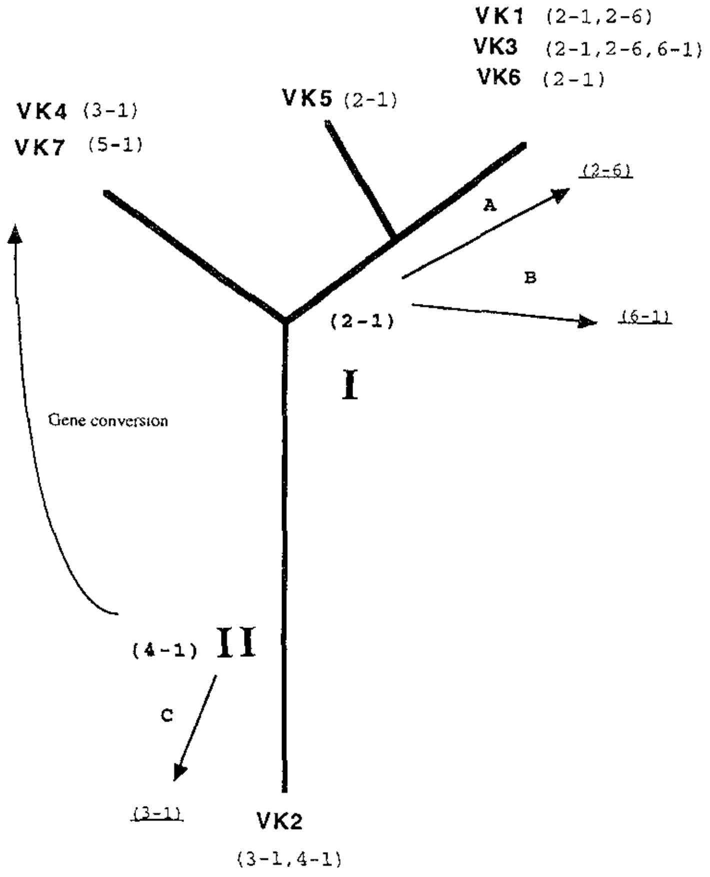
V_κ clans and the structural repertoire. The organization of three clans observed here for V_H is based on sequence conservation primarily from FR1 (4) (*cf.* Fig. 1) Unlike V_H, the FR1 sequence for V_κ varies significantly more, which does not allow a clear clan division of this locus (46,54). It has been proposed, however, that based on an FR3 analysis it is possible to identify two clans for V_κ: clan I comprising families $V_\kappa 1$, $V_\kappa 3$ and $V_\kappa 4$; and clan II only formed by family $V_\kappa 2$ (46). Both the tree presented here (*cf.* Fig. 4) and the one proposed by Kroemer *et al.* (47) agree with this clan clustering. Take into account that both trees are built considering the complete variable exon. In the tree presented here, in addition to the families mentioned above ($V_\kappa 1$, $V_\kappa 3$ and $V_\kappa 4$), families $V_\kappa 5$, $V_\kappa 6$ and $V_\kappa 7$ also appear clustered in clan I

An analysis of the clan clustering of V_κ in terms of the structural repertoire shows that clan I encodes mainly for classes with short loops at L1 (types 2 and 6), with the exception of the single member families $V_\kappa 4$ and $V_\kappa 7$, which encode for class 3-1. As previously mentioned, these families present a structural repertoire similar to the $V_\kappa 2$ family that forms another clan. It is interesting to observe that mouse families ($V_\kappa 8$, $V_\kappa 18/28$, $V_\kappa 21$ and $V_\kappa 22$) which are evolutionary correspondents to the human families $V_\kappa 4$ and $V_\kappa 7$ (47) also encode for class 3-1 (results not shown). This fact implies that gene conversion events among clan I and II members of V_κ occurred before the radiation between human and mouse.

V_κ primordial structural repertoire. In order to propose the ancestral structural repertoire for V_κ we performed a comparative analysis of the structural repertoire observed in human, mouse and other vertebrates

The following results were observed for clan I. (i) The most frequent class encoded in human families of this clan (class 2-1) is also the most common class for the corresponding families in mouse and other vertebrates (results not shown) (ii) Other canonical structure classes characteristic of this clan (6–1, 2-6) also appear in the corresponding families in mouse and other vertebrates. Based on these observations we propose the ancestry of clan I encoded for class 2-1 In Fig 6, the outline of the ancestral repertoire is presented It can be observed that classes 6-1 and 2-6 are only one molecular evolutionary event away from the ancestral class As mentioned above, classes 3-1 and 5-1 from families $V_\kappa 4$ and $V_\kappa 7$ respectively are outgroup with respect to the typical structural repertoire of this clan, and as a consequence do not follow the gradual evolutionary pattern observed in all the other cases.

For clan II the results are as follows. (i) The most frequent class for human (4-1) is also the most common one for the corresponding families in mouse and other vertebrates (ii) The other class (3-1) appears in human as well as in mouse and other vertebrates. Considering this, we conclude that the ancestry of clan II encoded for class 4-1. As can be seen in Fig 6, the evolutionary events between classes 4-1 and 3-1 are a minimum.



A 95Pro $\xrightarrow{S}$ His, Ser

B 30-31 $\xrightarrow{I1}$ 30-30a-31

C 30-30a-30b-30c-30d-30e-31 $\xrightarrow{D1}$ 30-30a-30b-30c-30d-30e-30f-31

Fig. 6. Proposed evolutionary relationships among canonical structure classes of V_κ families All the notations as in Fig 3

Discussion

We have analyzed in the preceding sections the evolution of the human V_H and V_κ structural repertoire. The distribution of the canonical structure classes in families and clans was proven to be compatible with a model of a minimum number of changes in order to transform one class into another The only exception to this rule was observed in clan I of V_κ, where apparently gene conversion events have created inter-clan diversity in the structural repertoire It is known that this kind of mechanism has played, in addition to vertical divergence, an important role in the evolution of the Ig genes (1,55) For both loci it is possible to propose a common ancestry class for each clan responsible for the generation of all the contemporary classes through a small number of molecular evolutionary events. This indicates that the evolution of the structural repertoire has followed the principle of maximum parsimony or minimum molecular evolution (56) in spite of the presence of mechanisms of horizontal evolution such as gene conversion The observed pattern in the distribution of classes within the families is apparently achieved through a strong selective pressure in order to avoid substitutions and/or insertions/deletions that might destroy/modify drastically the canonical structures These observations taken together support the hypothesis proposed here that states that the diversification process in the evolution of the structural repertoire was subject to evolutionary selective pressure

Analysis of the evolution of the V_H and V_κ structural repertoire has shown several common properties for these two loci (i) As mentioned before, both loci present a clan structure that correlates with the evolution of the structural repertoire. (ii) The large families encode for more than one class and there is one large family which contributes more than the others to the diversity of the structural repertoire. (iii) Most of the small families are redundant with respect to the large families in terms of the structural repertoire (iv) There is one small family for each locus (for V_H the one member family V_H6 and for V_κ the one member family $V_\kappa 7$) that encodes for a unique class

On the other hand, there are important differences in the evolution of the structural repertoire between V_H and V_κ The V_H families encode for classes that although evolutionary related, generate antigen-binding sites with different form (*cf* Fig 2a and c). For example, family V_H3 encodes for class 1-4 which presents the longest loop in H2 and simultaneously encodes for class 1-6 which presents the shortest type for this loop Conversely, the V_κ families encode for classes that generate a homogeneous structural repertoire For instance, family $V_\kappa 1$, the largest of V_κ, presents only classes with short loops in L1 (type 2)

It is worth mentioning another difference between V_H and V_κ. As previously mentioned, both loci presented one small family (V_H6 and $V_\kappa 7$) that encodes for a unique class (3-5 and 5-1 respectively) However, it appears that V_κ does not use this additional source of structural diversity. A recent study of usage of V_κ genes has revealed that gene B1 ($V_\kappa 7$), which apparently does not present genetic or structural defects, has not expressed counterparts in the known antibodies (25). Conversely, gene 6-1 (V_H6) is highly expressed (57) and is present in antibodies with a wide range of fine specificities (58–61). This gene is the most D proximal segment (31). Analysis of the sequences of higher primates revealed high conservation of this gene with a sequence variation of <2% between human and these species (62) Moreover, the canonical structure class characteristic of this gene (class 3-5) is present in Ig germline genes of a wide range of vertebrate species (*Xenopus*, rainbow trout and rabbit) (results not shown). The present work has shown that this gene is the only one that combines long loops in H1 and H2, and consequently is expected to generate antibodies with singular recognition properties All these results suggest an important function for this gene (1)

An important question that arises from this observation is, why are these important differences present in the evolution of the V_H and V_κ structural repertoire?

There is evidence from structural (18,63,64) and functional (65,66) analysis which indicates that the V_H domain plays a more important role than V_L in the recognition mechanism of the Ig The present work has proven that for V_H both H1 and H2 contribute to the diversity of the structural repertoire, but for V_κ only L1 varies. This result agrees with structural analysis which has shown that H1, H2 and L1 make numerous contacts with the antigen, whereas L2 does not since it is far from the binding site (64) Taking these observations together, it is reasonable to expect that the V_H locus has been subjected to different evolutionary selective pressures than V_κ as those observed here Antibodies of the V_λ isotype represent 30-40% of the human light chain repertoire (67) Consequently, it will be necessary to analyze the evolution and diversity of the structural repertoire of this locus to understand the differences between V_H and V_κ presented here, and to have the whole picture of the evolution of the human antibody structural repertoire However, this kind of analysis is difficult because the canonical structure repertoire of the V_λ isotype has been elusive to classify (19,23) and, at present, the nucleotide sequences of all the functional genes of this locus are not known (68)

There is evidence about the biased expression of gene families, and the correlation of this biased expression with normal and pathologic stages of the immune system (1,69–71) The proposed basis of this correlation is that members of the same family would have very similar recognition properties (69,72) In the present work, it was shown that genes with high identity belonging to the same family can generate very different structural repertoires Additionally, in a previous study we have shown that antibodies with different canonical structure classes possess different recognition properties (18) Consequently, the proposal that 'the expression of genes which belong to the same family will generate antibodies with similar recognition properties' must be managed cautiously

Primordial structural repertoire

In the present work based on the existence of a strong correlation between the evolution of the V_H and V_κ genes and the structural repertoire, we have proposed an outline of the primordial structural repertoire of V_H and V_κ loci For V_H it is proposed that this repertoire was composed of classes 1-2, 1-3 and 3-1 for clans I, III and II respectively For V_κ, the repertoire was 2-1 and 4-1 for clans I and II respectively This proposal is supported by the following observations: (i) all the existing classes in the clan can be generated in almost all cases by one insertion/deletion or substitution event from the ancestral class, consequently these ancestral classes are in the 'mid' distance by molecular evolutionary events between the present classes, (ii) the ancestral class is the most numerous of the clan, and (iii) the ancestral structural repertoire resulting from a comparative analysis in mouse and other vertebrates is the same as that found in human

A detailed analysis shows that the primordial V_H structural repertoire was formed by two classes combining short loops in H1 and H2 (1-2 and 1-3), and by a class that combines the largest loop of H1 with the shortest of H2 (3-1) (*cf.* Table 1 and Fig. 3). As previously shown, classes 1-2 and 1-3 represent only small variations of the same structural motif (*cf* Fig 2b) Because only one substitution is necessary in position 71 to convert type 2 into 3 in H2, it is probable that classes 1-2 and 1-3 arise from a common ancestral gene On the other hand, the primordial V_κ structural repertoire was composed by class 4-1 (clan II), which combines a large loop in L1 with a short loop in L3, and by class 2-1 (clan I), which combines short loops in L1 and L3 (*cf.* Table 2 and Fig 6) Thus, an outline considering the loop length of the primordial V_H and V_κ structural repertoire will involve two kinds of classes for each locus short–short 'S-S' (1-2/3 for V_H and 2-1 for V_κ) and long–short 'L-S' (3-1 for V_H and 4-1 for V_κ)

Recently, we have proposed the concept of potential structural repertoire of antigen-binding sites (28) This is defined as the $V_H V_L$ canonical structure classes that can be formed

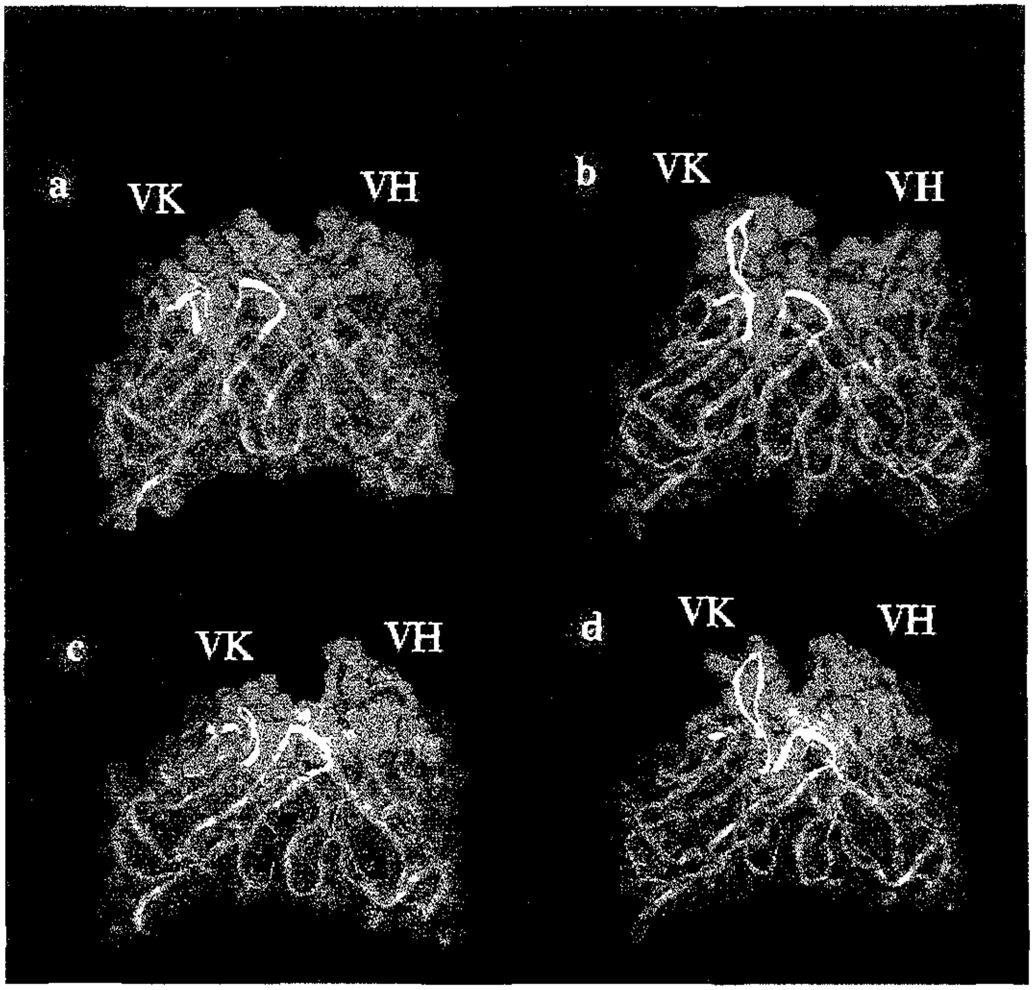


Fig. 7. Ribbon representation of antibody backbones with representative primordial V_H/V_κ canonical structure classes. Antibodies which present classes corresponding to the proposed ancestral human Ig structural repertoire are shown Framework and H3 are in blue, L1, L2 and L3 in yellow, and H1 and H2 in red (a) V_H/V_κ class 1-2-2-1, antibody 36-71 [6FAB (73)] (b) V_H/V_κ class 1-2-3-1, antibody 8F5 [1BBD (42)] Antibodies presenting the V_H/V_κ classes 3-1-2-1 and 3-1-2-1 have not been reported in the Protein Data Bank. To overcome this, we have built *ad hoc* V domains of the desired V_H/V_κ class by pairing the V_H domain from one antibody with the V_κ from another. For example, in (c), the V_H/V_κ class 3-1-2-1 was built by taking the V_H domain from antibody 50 1 [1GGB (44)] which presents class 3-1 (H1-H2) and the V_κ domain from antibody POT [1IGM (49)] which presents class 2-1 (L1-L3) The V_κ domain of POT was superimposed to the best fit (r.m.s. = 0.67 Å) to the framework of the V_κ domain of 50 1. Panel (d) was built with the V_H of 50.1 [1GGB (44)], class 3-1, and the V_κ from McPC603 [1MCP (51)], class 3-1 (r m s = 0 72 Å)

by the possible combinations between all the observed classes of V_H germline gene segments with those observed for V_L germline gene segments (being V_κ, V_λ or both) (28). Following this concept here we propose that the potential primordial structural repertoire of V_H and V_κ was composed by four different kinds of combinations of canonical structure classes for H1-H2-L1-L3: 1-2/3-2-1 (S-S-S-S), 1-2/3-4-1 (S-S-L-S), 3-1-2-1 (L-S-S-S) and 3-1-4-1 (L-S-L-S). These four V_H:V_κ canonical structure classes are presented in Fig 7(a-d) Within these four kinds of antigen-binding sites *there are antigen-binding sites with a flat shape (cf Fig. 7a),* a groove shape (*cf* Fig 7b and c) and a pocket shape (*cf.* Fig. 7d). It is worth mentioning that within these four V_H:V_κ classes, we only observe combinations of short types in all the loops or combinations of long loops in at most two hypervariable loops, but never long loops in all the cases

Several studies have shown that the main clusters or clans of V genes have been conserved through the evolution of vertebrate Ig (6,9,15,74,75). Analysis of Ig sequences revealed that canonical structures are present in a wide range of vertebrates species (26,32) Here we have shown that the structural repertoire is correlated with the evolution of families *and clans. Considering this, it can be expected that at least* some of the important features of the primordial human structural repertoire described here could be present in the primordial repertoire that was present in the early stages of evolution of the vertebrate immune system.

Acknowledgements

We thanks Dr I M Tomlinson for valuable comments, Dr J Felsenstein for providing the Phylip program and comments, R López-Hernández, E Paz-Garcia, E. Reyes-Rojano and E García-Hernández for technical assistance, and Haroldo González for revision of the English version of the manuscript E V. was supported by SNI and CONACyT grant 1833 P-N J C A was supported by the DGAPA grant IN-206093

Abbreviations

FR	framework region
H1, H2 and H1	first, second and third hypervariable loop of the heavy chain respectively
L1, L2 and L3	first, second and third hypervariable loop of the light chain respectively
V_H	variable heavy domain
V_L	variable light domain

References

1 Pascual, V and Capra, J D 1991 Human immunoglobulin heavy-chain variable region genes· organization, polymorphisms, and expression *Adv Immunol* 49·1

2 Kodaira, M , Kinashi, T , Umemura, I , Matsuda, F , Noma, T , Ono, Y and Honjo, T 1986. Organization and evolution of variable region genes of the human immunoglobulin heavy chain *J Mol Biol* 190·529

3 Lee, K. H , Matsuda, F , Kinashi, T , Kodaira, M and Honjo T 1987. A novel family of variable region genes of the human immunoglobulin heavy chain *J Mol Biol* 195 761

4 Schroeder, H W, Hillson, J L and Perlmutter, R M 1990 Structure and evolution of mammalian V_H families *Int Immunol* 2 41.

5 Kirkham, P M , Mortari, F , Newton, J A. and Schroeder, H W 1992 Immunoglobulin V_H clan and family identity predicts variable domain structure and may influence antigen binding *EMBO J* 11 603

6 Ota, T and Nei, M 1994 Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V_H gene family *Mol Biol Evol* 11·469

7 Möller, G , ed. 1990 *Immunol Rev* 115

8 Rothenfluh, H S., Blanden, R V and Steele, E J 1995 Evolution of V genes· DNA sequence structure of functional germline genes and pseudogenes *Immunogenetics* 42 159

9 Ghaffari, S H and Lobb, C J 1991 Heavy chain variable region gene families evolved early in philogeny *J Immunol* 146:1037.

10 Tanaka, T and Nei, M 1989 Positive Darwinian selection observed at the variable-region genes of immunoglobulins *Mol Biol Evol* 6:447

11 Perlmutter, R M., Kearney, J F , Chang, S P and Hood, L E 1985 Developmentally controlled expression of the immunoglobulin V_H genes *Science* 227 1597

12 Booker, J K and Haughton, G 1994 Mechanisms that limit the diversity of antibodies II Evolutionary conservation of variable region genes which encode naturally occurring autoantibodies *Int Immunol* 6 1427

13 Vargas-Madrazo E , Lara-Ochoa, F and Jiménez Montano, M A 1994 A skewed distribution of amino acid at the recognition sites of hypervariable region of immunoglobulins *J Mol Evol* 38 100

14 Tomlinson, I M , Walter, G , Jones, P T , Dear, P H Sonnhammer E L L and Winter, G 1996 The imprint of somatic hypermutation on the repertoire of human germline V genes *J Mol Biol* 256 813

15 Andersson, E and Matsunaga, T 1995 Evolution of immunoglobulin heavy chain region genes a V_H family can last for 150–200 million years or longer *Immunogenetics* 41·18.

16 Wu, T T. and 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 antibody complementarity *J Exp Med* 132 211.

17 Rees, A R and de la Paz, P 1986 Investigating antibody specificity using computer graphics and protein engineering *Trends Biochem Sci*. 11·144

18 Vargas-Madrazo, E , Lara-Ochoa, F and Almagro, J C 1995 Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J Mol Biol* 254 497

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

20 Chothia, C , Lesk, A M , Tramontano, A., Levitt, M , Smith-Gill, S J., Air, G., Sheriff, S , Padlan, E. A , Davies, D , Tulip, W. R , Colman, P M , Spinelli, S , Alzari, P M and Poljak, R J 1989 Conformations of immunoglobulins hypervariable regions *Nature* 342 877.

21 Wilson, I A. and Stanfield, R L 1994 Antibody–antigen interaction: new structures and conformational changes *Curr Opin. Struct. Biol* 4·857

22 Chothia, C , Lesk, A M., Gherardi, E , Tomlinson, J M , Walter, G , Marks, J. D , Llewelyn, M. B and Winter, G 1992. Structural repertoire of the human V_H segments *J Mol Biol* 227 799

23 Williams, S C and Winter, G 1993 Cloning and sequencing of human immunoglobulin Vλ gene segments. *Eur J Immunol* 23·1456

24 Cox, J P L , Tomlinson, I A and Winter, G 1994 A directory of human germ-line V-k segments reveals a strong bias in their usage *Eur J Immunol* 24 827

25 Tomlinson, I M , Cox, J P L., Gherardi, E , Lesk, A M and Chothia, C 1995. The structural repertoire of the human V_κ domain *EMBO J* 14 4628

26 Vargas-Madrazo, E , Almagro, J C and Lara-Ochoa, F 1995 Structural repertoire in V_H pseudogenes of immunoglobulins comparison with human germline genes and human amino acid sequences *J Mol Biol* 246 74

27 Almagro, J C , Lara-Ochoa, F , Dominguez-Martinez V and Vargas-Madrazo, E 1996 Structural repertoire in human Vl pseudogenes of immunoglobulins: comparison with functional germline genes and amino sequences *Immunogenetics* 43.92

28 Almagro, J C , Lara-Ochoa, F and Vargas-Madrazo, E 1997 A comparison between the potential and functional structural repertoire of human immunoglobulins reveals bias in the expression of V_H:V_κ canonical structure classes *Prot Eng*, submitted

29 Schable, K F. and Zachau, H G 1993. The variable genes of the human immunoglobulin κ locus *J Biol Chem*. 374 1001

30 Schable, H F , Thiebe, R , Flügel A , Meindl, A and Zachau, H G 1994 The human immunoglobulin κ locus pseudogenes, unique and repetitive sequences *J Biol Chem* 375.189

31 Cook, G P and Tomlinson, I M 1995 The human immunoglobulin V_H repertoire *Immunol Today* 16 237

32 Barré, S , Greenberg, A S , Flajnik, M F and Chothia, C 1994 Structural conservation of hypervariable regions in immunoglobulins evolution *Nature Struct Biol* 1·915

33 Almagro, J C , Vargas-Madrazo, E , Zenteno-Cuevas R , Hernandez-Mendiola, V and Lara Ochoa F 1995 VIR a computational tool for analysis of immunoglobulin sequences *Biosystems* 35 25

34 Kabat, E. A , Wu, T. T., Perry, H M , Gottesman, K S and Foeller, C 1993 *Sequences of Proteins of Immunological Interest*, 5th edn Public Health Service, NIH Washington, DC

35 Klein R , Jaenichen, R and Zachau, H G 1993 Expressed human immunoglobulin κ genes and their hypermutation *Eur J Immunol* 23 3248

36 Guarné, A , Bravo, J , Calvo, J Lozano, F , Vives, J and Fita, I 1996 Conformation of the hypervariable region L3 without the key proline residue *Prot Sci* 5 167

37 Felsenstein, J 1988 Phylogenies from molecular sequences inference and readability *Annu Rev Genet* 22 521

38 Bernstein, F C , Koetzle, T F , Williams, G J B , Meyer, E F Jr, Brice, M D , Rodgers, J R , Kennard, O , Shimandouchi, T and Tasumi, M 1977 The Protein Data Bank. A computer-based archival file for macromolecular structures *J Mol Biol* 112 535

39 Bhat, T N , Bentley, G A , Fischmann, T O , Boulot, G and Pojak, R J 1994. Bound molecules and conformational stabilization help mediate an antigen–antibody association *Proc Natl Acad Sci USA* 91 1089

40 Sheriff, S , Silverton, E W , Padian, E A , Cohen, G H , Smith-Gill, S J , Finzel, B C and Davies, D R 1987 Three dimensional structure of an antibody–antigen complex *Proc Natl Acad Sci USA* 84 8075

41 Herron, J N , He, X , Mason, M L , Voss, E W , Jr and Edmundson, A, B 1989 Three-dimensional structure of a fluorescein–Fab complex crystallized in 2-methyl-2,4-pentanediol *Proteins* 5 271

42 Tormo, J , Stadler, E Skern, T , Auer, H , Kanzler, O Betzel C , Blaas, D and Fita, I 1992 Three dimensional structure of the Fab fragment of a neutralizing antibody to human of a neutralizing antibody to human rhinovirus serotype 2 *Prot Sci* 1 1154

43 Suh, S W , Bhat, T N , Navia, M A , Cohen, G H , Rao, D N , Rudikoff, S and Davies D R 1986 The galactan-binding immunoglobulin Fab J539 An X-ray diffraction study at 2 6-angstroms resolution *Proteins* 1 74

44 Rini J M , Standfield, R L , Stura E A , Salinas, P A , Profy, A T and Wilson, I A 1993 Crystal structure of a human immunodeficiency virus type 1 neutralizing antibody, 50 1, in complex with its V3 loop peptide antigen *Proc Natl Acad Sci USA* 90 6325

45 Tutter, A and Riblet, R 1989 Conservation of an immunoglobulin variable-region gene family indicates a specific, noncoding function *Proc Natl Acad Sci USA* 86 7460

46 Kirkham, P M , Elgavish, R A and Schroeder, H W 1993 Structure and evolution of mammalian Vκ families *J Immunol* 150·150a

47 Kroemer, G Helmberg, A , Bernot, A , Auffray, C and Kofler, R 1991 Evolutionary relationship between human and mouse immunoglobulin κ light chain variable region genes *Immunogenetics* 32 42

48 Haynes, M R , Stura, E A , Hilvert, D and Wilson I A 1994 Routes to catalysis structure of a catalytic antibody and comparison with its natural counterpart *Science* 263 646

49 Fan, Z -C., Shan L , Guddat L W, He X M Gray W R Raison R L and Edmunson, A B 1992 Three dimensional structure of an Fv from a human IgM immunoglobulin *J Mol Biol*, 228 188

50 Padlan E A , Cohen, G H and Davies, D R 1985 On the specificity of antibody/antigen interactions, Phosphocholine binding to Mc/PC603 and the correlation of three-dimensional structure and sequence data *Ann Immunol (Paris) Sect C* 136 271

51 Satow, Y Cohen, G H Padlan, E A and Davies, D R 1986 Phosphocholine binding immunoglobulin Fab MC-PC603 An X-ray diffraction study at 2 7 angstoms *J Mol Biol* 190 593

52 Webster, D M , Henry A H and Rees A 1994 Antibody–antigen interactions *Curr Opin Struct Biol* 4 123

53 MacCallum, R M , Martin A C R and Thornton, J M 1996 Antibody-antigen interactions contact analysis and binding site topography *J Mol Biol* 262 732

54 Kirkham, P M and Schroeder, H W 1994 Antibody structure and the evolution of immunoglobulin V gene segments *Semin Immunol* 6·347

55 Wysocki, L J and Gefter, M L 1989 Gene conversion and the generation of antibody diversity *Annu Rev Biochem* 58 509

56 Li, W-H and Grau, D 1991 *Fundamentals of Molecular Evolution*, p 106 Sinauer Press, New York

57 Schroeder, H W., Hillson, J L and Perlmutter, R M 1987 Early restriction the human antibody repertoire. *Science* 238·791

58 Logtenberg, T , Young, F M , Van Es, J H , Gmelig-Meyling, F H and Alt, F W 1989 Autoantibodies encoded by the most J_H-proximal human immunoglobulin heavy chain variable region gene *J Exp Med* 170 1347

59 Singal, D P, Frame, B Joseph S , Blajchman, M A and Leber B F 1993 Nucleotide sequences of an antiidiotypic antibody from a transplant recipient *Immunogenetics* 38 242

60 Settmacher, U , Jahn, S , Siegel, P , von Baehr, R and Hansen A 1993 An anti-lipid antibody obtained from the human fetal repertoire is encoded by V_H6–$V_\lambda1$ genes *Mol Immunol* 30 953

61 Andris, J , Brodeur, B R and Capra, J D 1993 Molecular characterization of human antibodies to bacterial antigens utilization of the less frequently expressed VH2 and VH6 heavy chain variable region gene families *Mol, Immunol* 30 1601

62 Meek, K , Eversole, T and Capra, J D 1991 Conservation of the most J_H proximal Ig V_H gene segment (V_H/VL) through out primate evolution *J Immunol* 146 2434

63 Kabat, E A and Wu, T T 1991 Identical V region amino acid sequences and segments of sequences in antibodies of different specificities *J Immunol* 147 1709

64 Wilson I A and Stanfield, R L 1993 Antibody–antigen interactions *Curr Opin Struct Biol* 3 113

65 Zouali, M 1995 B-cell superantigens implications for selection of the human antibody repertoire *Immunol Today* 16 399

66 Ward, E S Gussow, D , Griffiths, A D , Jones, P T and Winter G 1989 Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli Nature* 341 544

67 Milstein, C 1965 Interchain disulfide bridge in Bence–Jones proteins and γ-globulin B chains *Nature* 205 1171

68 Frippiat, J P, Williams, S C, Tomlinson, I M Cook G P , Cherif, D , Le Peslier, D , Collins J E , Dunham I Winter, G and Lefranc, M -P 1995 Organization of the human immunoglobulin κ light-chain locus on chromosome 22q11 2 *Hum Mol Genet* 4 983

69 Marion, T N , Tillman, D M , Jou N T and Hill, R J 1992 Selection of immunoglobulin variable regions in autoimmunity to DNA *Immunol Rev* 128 122

70 Stewart, A K and Schwartz, R, S 1994 Immunoglobulin V region and the B cell *Blood* 83 1717

71 Schroeder, H W and Digheiro G 1994 The pathogenesis of chronic lymphocytic leukemia Analysis of the antibody repertoire *Immunol Today* 15 288

72 Stewart, A K , Huang C , Long, A A , Stollar, B D and Schwartz R S 1992 V_H-gene representation in autoantibodies reflects the normal human B-cell repertoire *Immunol Rev* 128 101

73 Strong, R K , Campbell, R Rose D R , Petsko, G A Sharon, J and Margolies M N 1991 Three-dimensional structure of murine anti-*p*-azophenylarsonate Fab 36-71 1 X ray crystallography site-directed mutagenesis, and modeling of the complex with hapten *Biochemistry* 30 3739

74 Hayzer, D J 1990 Immunoglobulin λ light chain evolution Igλ-like sequences form three major groups *Immunogenetics* 32 157

75 Rast J P, Anderson M K and Ota, T 1994 Immunoglobulin light chain class multiplicity and alternative organizational forms in early vertebrate phylogeny *Immunogenetics* 40,83

Pergamon
PII: S0161-5890(97)00118-1

Molecular Immunology, Vol. 34, No 16-17, pp 1199-1214, 1997
© 1997 Elsevier Science Ltd All rights reserved
Printed in Great Britain
0161-5890/97 $17 00 + 0 00

THE DIFFERENCES BETWEEN THE STRUCTURAL REPERTOIRES OF V_H GERM-LINE GENE SEGMENTS OF MICE AND HUMANS: IMPLICATION FOR THE MOLECULAR MECHANISM OF THE IMMUNE RESPONSE

JAUN CARLOS ALMAGRO,*‡ ISMAEL HERNANDEZ,*
MARIA DEL CARMEN RAMIREZ* AND ENRIQUE VARGAS-MADRAZO†

* Instituto de Biotecnologia, Universidad Nacional Autónoma México, Apdo Postal 04510-3, Cuernavaca, Morelos 62250, Mexico, † Instituto de Investigaciones Biológicas, Universidad Veracruzana, Araucarias 280 Col Animas, Xalapa, Ver. 91190, Mexico.

(*First received* 2 *June* 1997, *accepted in revised form* 4 *September* 1997)

Abstract—Although human and murine antibodies are similar when considering their diversification strategies, they differ in the proportion by which κ and λ type chains are present in their receptive V_L repertoires. It has been shown that this difference implies a divergence in the structural repertoire of the κ and λ genes of these species. Nonetheless, the differences in V_H have not been systematically studied. In this paper a systematic characterization of the V_H structural repertoire of mice is made, so that a comparison with the V_H structural repertoire of humans, described in detail elsewhere, could be properly accomplished. Our study shows the structural repertoire of mice to be dominated by canonical structure class 1–2 (~60%), while in humans the dominant one is class 1–3 (~40%) Analysis of the evolutionary relationships between human and mice suggest that this divergence may have a functional meaning The implications of such findings are discussed © 1997 Elsevier Science Ltd. All rights reserved

Key words: immunoglobulin, Ig, canonical structures, V_H repertoire, structural repertoire.

INTRODUCTION

The antigen-binding site of antibodies consists of six hypervariable loops; three from V_H and three from V_L denoted H1, H2, H3 and L1, L2, L3 respectively (Wu and Kabat, 1970; Kabat and Wu, 1971, Poljak *et al*, 1973). Although there is great sequence variability in these regions (Wu and Kabat, 1970; Kabat and Wu, 1971), it has been shown that excepting H3, the remaining five hypervariable loops have one of a small set of main-chain conformations or canonical structures (Chothia and Lesk, 1987; Chothia *et al*, 1989) Based on that fact, it has been found that from the total number of possible combinations of canonical structures only a few possibilities do exist in the known antibody sequences, named structural repertoire (Chothia *et al.*, 1992; Tomlinson *et al.*, 1995; Vargas-Madrazo *et al.*, 1995a Vargas-Madrazo *et al.*, 1995b; Almagro *et al.*, 1996) Furthermore, it has been suggested that the antigen-binding site shapes allowed by the structural repertoire correlate with the kind of antigen the antibody interacts with (Vargas-Madrazo *et al.*, 1995a; Lara-Ochoa *et al*, 1996) Taken together, these findings provide evidence concerning structural restrictions at work in the process of antigen recognition.

Genetically, the structural repertoires of human and murine antibodies are generated in a similar fashion (Weill and Reynaud, 1996): L1, L2 and most of L3 are encoded in the V_L gene segments (κ and λ type), while H1 and H2 are encoded in the V_H gene segments (Tonegawa, 1983). In spite of this similarity, it has been noticed that the corresponding repertories of humans and mice differ in the relative proportion by which κ and λ type chains are present in V_L. In humans, roughly 60% of the V_L repertoire is κ type [40 functional V_K germ-line genes versus 30 functional V_λ germ-line genes (Klein *et al.*, 1993, Tomlinson *et al.*, 1995; Williams *et al.*, 1996)] In mice, κ type preponderates, being as much as 95% (Hood *et al*, 1967) Such divergence implies differences in the structural repertoire of humans and murine V_κ and V_λ germ-line genes (Williams *et al.*, 1996, Almagro *et al.*, 1998) and, consequently, differences in the initial structural restrictions operating to recognize different types of antigens.

Although differences in V_H are less evident, recent studies we made in the rearranged V_H sequences of mice indicate that the combination of canonical structures most frequently used is the 1–2 class (combination of

‡ Author to whom all correspondence should be addressed. Tel : (52) (73) 291605; Fax: (52) (73) 172388, e-mail: almagro @ibt unam mx

Abbreviations V_H, Variable heavy domain, V_L, Variable light domain canonical structures in H1 and H2) (Vargas-Madrazo *et al* , 1995a; Lara-Ochoa *et al.*, 1996). In contrast, human V_H germ-line genes, which have been thoroughly characterized (Cook and Tomlinson, 1995), have shown to encode predominantly canonical structure class 1-3 (Chothia *et al.*, 1992; Vargas-Madrazo *et al.*, 1995b). We have also found that same difference at pseudogene level (Vargas-Madrazo *et al.*, 1995b). This suggests that V_H germ-line gene segments of mice and humans may encode different structural repertoires in V_H too.

Such difference, however, has not been properly characterized, partially because the structural repertoire of the mice V_H germ-line genes has not been systematically studied. A proper characterization of this subject could provide insight and additional ideas to the theories addressing the origin, organization, complexity and use of V_H genes. Furthermore, if such differences in V_H do exist, then taken together with the structural divergence in the repertoire of V_L germ-line genes, they could shed light on the different structural constrains at work when antigen recognition takes place in human and mouse (Vargas-Madrazo *et al.*, 1995a). In addition, such a characterization might prove useful as a criterion to choose human V_H genes for humanization of murine antibodies (Poul and Lefranc, 1995).

In this paper we compiled the information published on V_H gene germ-line segments of mice to characterize their structural repertoire. Comparison with its human counterpart corroborates the differences found in rearranged sequences and pseudogenes. Implications of such findings for the molecular mechanism of the immune response are discussed.

MATERIAL AND METHODS

The germ-line V_H gene segments of mice

We compiled all of the *Mus musculus* V_H gene segments reported as germ-line genes or pseudogene sequences at Genbank and LIGM, as well as in current literature up to April 1996 We found a total of 295 V_H gene segments and immediately discarded 42 of them because of being duplicates (different accession numbers but identical entries) or not comprising one or both hypervariable loops (see web site http://www ibt.unam mx/~almagro for a full description of the sequences)

Of the remaining 253 V_H gene segments, some were identical at nucleotide level, so we considered them to be the same V_H gene segment because current available information does not allow to distinguish if these sequences are recent copies of a particular V_H gene segment in the mice genome or if they have been sequenced more than once.

There were also present, pairs of sequences with one or two nucleotide differences (99 6% and 99.2% identities respectively). Those sequences having silent mutations (100% identical at amino acid level) were also considered to be the same gene segment. This is so because they might be alleles in different individuals or in different strains of mouse. Sequences in which the nucleotide difference resulted in replacements (different amino acid sequences) were considered as distinct V_H gene segments. Although this might seem very conservative, we relied on it because there is no established criterion to define alleles based only on the analysis of nucleotide identities. Thus, we preferred to include in the analysis all those sequences differing by at least one amino acid in order to avoid underestimating the available information.

A unique exception was made with those genes belonging to the S107 (V_H7) family which has been well characterized in two strains of mice. BALB/c (Crews *et al.*, 1981) and C57BL/10 (Perlmutter *et al.*, 1985). In this family we have taken into account only the alleles of BALB/c (the most represented strain within the compilation; see below), in spite of those from C57BL/10 which differ by more than one amino acid when compared with the BALB/c sequences. In this way, we managed to finally gather 185 sequences as representative of the mice V_H locus.

Classification of the known V_H gene segments in gene families

V_H gene segments in mice have been classified in 15 families based on Southern blot hybridization and sequencing (Brodeur and Riblet, 1984, Kofler *et al.*, 1992, Mainville *et al.*, 1996). Each family is represented by a prototype member defining the name of the family (Kofler *et al.*, 1992, Mainville *et al.*, 1996). V_H sequences within families share an identity of at least 80%, whereas among those belonging to different families the identity is at most 75% (Brodeur and Riblet, 1984). Following these criteria we clustered the 185 sequences finally gathered into the 15 established V_H families. In the case of the V_H14 family, in which some members are greater than 80% identical to sequences belonging to the V_H1 family, the assignment was made following the criteria established by Tutter and coworkers (1991)

The resultant sequence alignment, as organized by families, is given in Fig. 1 and can be retrieved from web site http://www.ibt.unam.mx/~almagro Within V_H families, the sequences are sorted according to the decreasing order or similarities they have with respect to prototype members.

Determination of functional V_H gene segments

From the 185 V_H gene segments depicted in Fig. 1, 47 were reported as pseudogenes in databases or in the literature (see status column of Fig. 1). This led us to assume they had serious genetic defects and, so, were not taken into account to determine the mice V_H gene segments functional repertoire The remaining 138 V_H gene segments reported as germ-line genes and potentially functional were examined to see what would their *in vivo* expression be. V_H gene segments not expressed *in vivo* might have defects within the coding region hindering the formation of a stable three-dimensional V_H domain Otherwise, they may have minor genetic defects outside the coding region, for example in splicing sites, regulatory

```
Ig-fold^a              B B    B  T  B B B B 111111111 B BIBI  T  I IBB  2222222        TT B        B B B  T   T B BIBI
Position^b             1        10        20        30          40        50          60        70        80          90
Family^l  Name^l       |...|. .|.. |...| ...|    | ab...|. ..|.. .|....|..abc..|.. .|.. .|....|   .|....|. abc..|....|. .|   Rearranged gene^e   Status^f
VH1       J558         QVQLQQPGAELVKPGASVKLSCKASGYTFTS--YWMHWVKQRPGRGLEWIGRIDP--NSGGTKYNEKFKSKATLTVDKPSSTAYMQLSSLTSEDSAVYYCAR
VH-186-2^a/V186-2^a/B21c^b/B10C^b   . ... ... . .. ......... ....--.. ..... ............--............................ ..... ...... ..   T099      (0)   F        1
C36e^c/B7c^b           P.. . .... ........ .   . .. -- . . . . . ...... ...-- .... ............... ...   .. .             E11-14    (15)  F        2
VH145^a/C1egc^c        .  ..   . .....      .S     ....--.    .. ... . ..... ..--.................................... ....... .....             NP^SD    3
VH186-1^a              ....  ....  ...... .  ..   .   --... .. . .. . ..... .--...  ............. TS... ...... . ....H....         NP       4
C14c^c                  ............... ...R .   . ..-- ...  .... Q.....  ...--........ .... .......   ...   .                    PS
B16c^b                 P ......... ....  ........     .--...    .     .   ..*....--.................. ...........               PS
C19c^c                 .....   . . .... ... S... .  ..-- ..........Q.......   --....    .  ............. . ... I              NP^SD    5
C22e^c                 E ... ........     .             ..--  .     ...Q.... .....--......S....................   .          NP       6
7C-07^a                ...   ... ....  ....  ...V   K.--....*.. ... ...........--............  .  ...  ... ..... .. .... ....  PS
C44gc^c                .. ... .........   ...    ...   -- .... ...  Q... ..M.H.--...S.N..  ...........S..... ...            NP       7
C31e^c                 E..   ...  ........ .. ........--.... ......Q......M.H.--...S.N.... ...    .. S.... ....   cyd-1     (17)  F        8
C20c^c                 ..  ....... .... ...   ....   .--. . ......Q .... M H.--  .S.N .... .........S..S.. ..I                  NP       9
C25c^c                  .............. . .... .....  .-- .. ......Q..... ....--G.SS.N.............TS........I               NP       10
VH28^b                 ....   ..... .... .... . ...   --...Q.. ....Q  ... N.N.--...S.N.....  .....  .............   .......TR  PS
C15c^c                 .. .. ..... ...... ..S    ..   --..........Q......N.N.--SN...N .................... ..             NP^SD    11
V23^a/C3egc^c/C35e^c/C45g^c/C44e^c   ......T........ . ..  .......--.  ...  ...Q  ....N.N.--SN...N......  .... ..S......... ...........   T077   (0)   F   12
B13c^b                  ...   ..... ... .... .....  .--..........Q.P......H.--SN...N.............TS......H..                 NP       13
VH3^b                  ................ .M.. ..... ..-- .IT*.... . ..   ...Y.--S....N .... ...... .TS   ...   ......... ...I   PS
C22c^c                 . ......T......    ... ..... ..--...... ....Q......N.N.--SN...N..............S..S.....I             NP       14
B3e^b                  . ......S... ......... . ...  .--S....A.....Q.... ......--.. N.N. . ..G.......TS ....VD             NP       15
B25c^b                 ................  ....  ... ..-- ..... ....Q......E.N.--SN.R.N..............S........S          163 100   (1)   F        16
VH6^a                  . ..... ...............  .. ..--.  ........E.....N.Y.--G.SS.N.. ... .......T...   ... ... ... .  ...   NP  17
C46g^c                 .......T.........  ..     .    -- .  ......AQ......N.N.--SN...N..............S... .....             NP       18
C11c^c                  ......T..................... --..  .......E......N Y.--G.SS.N.. ...........TS ..   ..        T210      (5)   F        19
B20c^b                 ...................M....   ..  --...T.... ..Q......D.Y.--G..S.N...  .........TS...... .          NP       20
C9gc^c                 .. .................R...... ...--... . .....E......N.Y.--G.SS.N..... ........TS ....  ..        PS
J558-122T^d            .   ....SV..R......     ..      --...N.......Q .... G.Y.--. .S.D... ..G.... ..TS. .T..D .....K........   NP   21
C40c^c                 ............. ...............--...Y.......Q ..R..E.N.--GN...N....    .... ..S..  .. ..          NP^SD    22
B12c^b                 .. . ........ ..P........     --.  N.... .....Q.. . ..--SDSE.H  Q.. ...... ...S.....I.        RF-4 PAN  (4)   F        23
C16c^c                 .. . ...........    M..    .... .-- .IT..... .Q......D.Y.--G..S.N... . ........TS.....  ..    H20-A15   (10)  F        24
C27c^c                 ..  ..........  ...M.. ....     --..IT ... . Q.   ...D Y --G..S.N...........TS....    I                NP       25
VH124^b/VH 124^b       .. ...............  ...   .     --...........Q......E...--SDSY.N..Q...G........S............  ... .....   L3 11D   (2)   F   26
C10g^c                 .. .. ............V........ ...--.  ..... ..Q........H.--SDSD.N..Q...G........S.....  .         NP       27
C33eg^c                ..  ........*...   ...      .. --.. .........Q......E...--SDSY.N..Q...G .......S... .....          PS
C23c^c                 . ...... .....  ...V... .......--. . . .....Q..   ....H --SDSD N..Q . G... ....S   ..   I         NP       28
C38e^c                 E. ..   .........  V....       --.. .... ...Q. .......H.--SDSD.N..Q...G. ..... S.........          NP       29
C8g^c                  ... ....... ...    .M..     . . .-- IT....S .Q  ....D.Y --G..S N .  .    .... .TS..                 NP       30
C2c^c                    ... ............M          --..IT.......Q... ..DTH.--G..S.N... .... .....TS. ... . .               NP       31
B9c^b                  .. ............T..  .. . ..N   --. IN...L  .Q.. ...D.Y.--G  S  ... .     ... .TS..  .               NP       32
C6e^c                     .     ......T  ....         --. IN...L...Q......D.Y.--G..S.N .    ........TS.......              NP       33
p2M5^e                 . ..... ......... ..       .    -- IN... ...Q. ....N.Y --G.SS N   .   ...    TS. .      ...D.  . . .   MRA7H   (0)   F   34
C11g^c                  . .   *... . ..V...            --          ..Q  ....  H --SDSD.N .Q...G....   .S.... ..          PS
B6c^b                   . ...  .....    M      .     -- .IN..     Q.... .D Y --GT I.N .       .. .L TS                      NP       35
```

Fig 1(a)

Ig-fold[a]	B B B T B B B B 11111111 B BIBI T I IBB 22222222 TT B B B B T T B BIBI				
Position[§]	1 10 20 30 40 50 60 70 80 90				
Family[I] Name[‡]	| . |....|....|. .| ..|....|.ab...| ...|... |. .| abc..|....|....|.. .| | .| abc..|....|.. |	Rearranged gene[t]		Status[*]	
VH1 J558 (continued)	QVQLQQPGAELVKPGASVKLSCKASGYTFTS--YWMHWVKQRPGRGLEWIGRIDP--NSGGTKYNEKFKSKATLTVDKPSSTAYMQLSSLTSEDSAVYYCAR				
B14e[b]/B2c[b]/B5e[b]	. . .. M ........--..IN.......Q... .D.Y.--GR.I.N............L.TS			NF	36
VH102[a]	E. .. ....V. .. ..-- ....Q . .H.--SDSD.N..Q .G. S I			NF	37
VH33[b]	... .K.... . .. .--......E*...Q.. ...E.N.--SN...N......R.. . .S..... . ... . . .TI			PS	
VH5(P1)	T. ...M .. ..--.T.. .....Q.... .Y N --S..Y.N..Q.. D .A S			PS	
H30[b]	......S......T. M .. ..--.T.........Q.. ...Y.N.--S..Y.N..Q...D.....A. S...	R	(3)	F	38
pcDFL.1[b]	.. ..V. RH. ..............--S....A...H.Q .. ..E.H.--...N.N......G.... . S VD..............			PS	
VH 104B[b]/VH104B[b]	.. .SV..R..T....... ... ..--.....A... .Q.. . .E.H.--.C.NIN......G.. . .TS.....VD.			NF	39
C26c[c]	....--.........................---PYSDI..S....N. . .....N. . .HI			NF^{SD}	40
B16e[b]	..............T.... . N ..--..IN.. L...Q..... D.Y.--G..S.N............. .TS... . .	A003×40/5G7	(2)	F	41
H8[b]	...S.P... .. .RI ........--.YI........Q......W.Y.--GNVN.... .G	13	(4)	F	42
H9[b]	...S.P..... .. M.. .....--.YI........Q......W Y.--GD.S........G.T...A	L77	(11)	F	43
B13e[b]	..E...S.PQ..... ..I ......S...--..........Q.....AM...--SDSE..*.Q............S			PS	
VH5[b]	*......... NT.....M..... .....-- T. . ..L.Q.... .Y.N.--S..Y.N..Q...D. ...A..S..			PS	
H13-1[b]	...S.P..... L. I...........--.DIN.......Q......W.Y.--GD.S...... G ....A..S.... .. . ...N..			NF^{SD}	44
VH105[b]	......S.P..........I .. ..--.YI.. .....Q......Y.Y.--RD.S.N......G.. . A.TS...... ... . . .F...	H72	(12)	F	45
H13-3[b]	......SA...AR.. .M...........--.T.........Q......Y.N.--S..Y.E..Q.. D.T...A..S..	A	(4)	F	46
J558-43y[d]	......S.P.........RI.... ......--.NI........Q......W.Y.--GD.N........G.T...A..S............ . ..F ..	111.68	(9)	F	47
VH3[a]	.... .R..S .. ..--...D.......Q..... N.Y.--SDSE.H..Q...D .. ....S......	MRA11H	(4)	F	48
B23c[b]/B18c[b]	. ...M......M ....D--...........Q......T..T--SDSY.S..Q...G .. .ES.	Gu0-2	(14)	F	49
VH-Id-11[d]/V(H)Id(CR)[d]	S ..M ..........--.GIN.......Q .....Y.N.--GN.Y........G.T ...S.........R..... . .F ..	ASWA1	(0)	F	50
J558-28[d]	......S.... R..T...V.......V..N--.LIE..... .Q.......V.N.--G....N......G.....A..S.. . ....D.....F...	CO17-1AC	(1)	F	51
VH31[b]	.A . .S.............M. ....S...--.YI........QE.. .*.FL--G..N.........G.....A.TS. ....HF. .			PS	
B1c[b]	... .R. . .. .--..IN.......Q.......N.Y.--LDSN.N..Q. D.. .. .S....	RP-2	(2)	F	52
J558-83[d]	......S............RI...T.......--.NI....E...Q... .W.Y.--GD.N........G.T...A..S.... . ...F .			NF	53
VH108a[b]	E... .S.P. .... I. ......D--.N......SH.KS.....Y.Y.--YN...G..Q... .. ..NS......E................			NF	54
pM11[e]	... ..S.....R..T...M....A.....N--..IG.......H.......D.Y.--GG.Y.N......G.. ..A.TS. ....I.....	H163-130H9	(0)	F	55
VH111[b]	.. ...S . ..R..T.. K..........AN--..IG.......H.......D.Y.--GD.V.N. ....A.....A..S... .E..R. ......*....			PS	
H16[b]	E.....S.P......... I..........D--.N......SH.KS.....Y.Y.--YN...G..Q............NS......DVR........			NF	56
J558-1.3[d]	E.. ..S.....R..S.......T.......--.GIN.......Q.......Y.YI--GN.N.E... .... .S.TS......E..........I.F. .			PS	
VGAM3-0[m]	E.....S.TV.AR.. ...M. .T. .....--...........Q.......A.Y.--GNSD.S..Q...G..K..AVTSA.. ..E ... N........T.			PS	
B26c[b]	... .R. .. ..S...--...N.......Q.. ...M.H.--SDSE.RL.Q...D...... .S........	L2 11C	(17)	F	57
B4c[b]	................S...........N...--..IN...L..YQ.I...*D.Y.--G..S.N..... .... TS.			PS	
MH	S.......T.......--.GIN.......Q.......Y.Y.--GN.Y.A...Q. G .....TS... ....R.....			NF	58
VH-Id-7[d]/VH-Id-14[d]	S. .. ..T. .....--.GIN......Q.......Y.Y.--GN.Y.A...Q..G.. S TS... .R.......	anti-(cyd-1)	(12)	F	59
BALB71	E... F . . .... I..........D--.N... ..SH.KS.....G.N.-- .N.A.S .Q...G. ......S......E.R	11F6	(16)	F	60
C57G5	E.....S.P........ .M........K..D--.Y......SH.KS.. .D.N.--.N...S..Q. .G.........S. . ...N	MOPC104E	(2)	F	61
VAR100	. ...S P. .R..L...........I.IT--...N.......Q.......Q.F --A..S.N...M.EG.......TS..........			NF	62
BALB17	E... .S.P.. ..I. T. . .E--.T......SH.KS.....G.N.--.N.. S..Q. .G... .. S. ...E R	129	(1)	F	63
BALB67	E.. ..P.P...........I...........D--.N.D....SH.KS.....D.N.--.N...I .Q...G.........S......E.R			NF	64
C57C27/C57G6	E.....S.P.. . ...I .. .. ..D--.Y.D....SH.KS.....D.N.--.N...I..Q ..G.. ... S ... .E R	H	(14)	F	65
37A11[b]	.. L S... M.... .I....T....S --..IE.. ....H... ..K.L.--G..S.N......G..KF.A IS.N...	19 1 2	(4)	F	66
BALB6	E.....S.P.......L .I .T......E--.T.. ...SH.KS.....G.N.--.N...S..Q...G..... . S.. E.R			NF	67
C57C18/C57G3/C57G14/C57C9	E. . S P . . . IP . . ....D--.N.D SH.KS. . D N -- N...I..Q...G. S . E R			NF	68

Fig 1(b)

Family	Name	Sequence	Rearranged gene		Status	
Ig-fold[a]		B B B T B B B B 111111111 B BIBI T I IBB 22222222 TT B B B B T T B BIBI			#	
Position[β]		1 10 20 30 40 50 60 70 80 90				
Family[ι]	Name[ι]	| .|....|...|....|..|...|.ab..|...|. .|..|..abc |. .|....|....|....|....|..abc..|...|. . .|	Rearranged gene[ι]		Status[ι]	
VH1	J558 (continued)	QVQLQQPGAELVKPGASVKLSCKASGYTFTS--YWMHWVKQRPGRGLEWIGRIDP--NSGGTKYNEKFKSKATLTVDKPSSTAYMQLSSLTSEDSAVYYCAR				
	VAR104[b]	....S P..R..T...I.........LT--...N...*..AQ......Q.F.--A..S.N ..M..G.......TS.... .. ..F..			PS	
	V104A[b]/VAR104A	....S.P ..R..T...I....... .LT--...N...*M..Q..... Q.F.--A..S.N...M..G... ...TS............. . .F..			PS	
	H24[b]	. ...S.....R..T...K...V.. .AN-- .IG.......H ... .D.Y.--GD.V.N... ..G.....A..S.... *. . ... S.			PS	
	VH36-65[c]	E.....S ... R..S...M...T...... --.GIN.......Q......Y.YI--GN.Y.G.. ..G.....S.TS............ ..			NF	69
	GLvh50[f]	E.....S.P.. ...D...M... ......D--.Y.D... SH.KS.....Y.Y.--.N. S..Q...G .......S......E.H..............	163.72	(0)	F	70
	BALB8	E.L...S.P..........IT.........D--.N.D....SH.KS.... D.N.--.N...I..Q...G........S......E R			NF	71
	BALB9	E.L .S.P..........IP... ... .D--.N.D....SH.KS.....D.N.--.N...I..Q ..G........S.. .. E R			NF	72
	J558-122B[d]	..... S.P...R..T...I.........IT--...N...*...Q.X... Q.F.--A .S.N...M..G... ...TS........ ... ... .HF..			PS	
	VHATAG-2[b]	....SDT............... .....D--HAI.......EQ. ....Y.S.--GN.DI ......G.....A..S.........N........ .F.K.			NF	73
	J558-186[d]	S.....S. ...R..S.......T.... .--.GIN.......Q......Y.YL--GN.Y.A......G.....S.TS....... .R.. ....VIKF. .			NF^{SD}	74
	C57C2	E ..S.P............... .S..G--.Y.N....S.EKS.....E.N.--.N...S .Q. .G........S.... .E.R	1410E.10e	(9)	F	75
	C57G9	E.. ..S.P....... ..I..........D--.Y.N...SH.KS.....D.N.--.N...S..Q.I.G........S......E.R	AC38 205.12	(1)	F	76
	C57G1	E ....S.P ..N.... .I..... .S..G--.Y.N... S. KS.....E.N.--ST...T..Q ..A........S........K			NF	77
	BALB58/BALB13	E.....S.P..........I.....*.. .D--.N......SH.KS.....G.N.--.N.A.S..Q...G... ....S .....E.R			PS	
	J558-43X[d]/21[d]/VHATAG-1[b]/H17[b]	.... .SD..........I..........D--HAI.....K.EQ......Y.S.--GN.DI.......G. ..A .S.........N...... ..F K.	3-1-3	(2)	F	78
	pHC103[b]	...S.P......P...I.......S..G--.Y......SH.KS.....E.N.--YN...S..Q...G.......TS... .E.H.......L.. ..			NF^{SD}	79
	C57C48	E ....S.P.... ....I....... ...D--.Y....SRAME.A.... .D.N.--.N...S..Q...G........S........N			PS	
	H26-1[b]	E.....S.P....L.P...I.......S..G--.Y......SH.KS.....E.N.--YN...S..Q. .G..... .TS......E.H.......L......			PS	
	Balb11/Balb19	E.. ..F.............I..........D--.N.D....SH.KS.....D.N.--.YDS.S..Q...G........S......E.R	mAb A41	(6)	F	80
	C57G26/C57C16	E.....S.P.... .....I.......S..G--.Y.N....S.EKS.....E.N.--ST...T..Q. .A.........S...... ..K	5G12-6	(1)	F	81
	C57C17	E. ...S.P..........I.......S..G--.Y.N....S.EKS.....E.N.--ST...T..Q...A.. .....S........K			PS	
	VH104A[b]	....S.P...R..T...I ........LT--...N...*M..Q.......A.F.--AG.S.N..QM..G.......TS..... . .. . .F...			PS	
	C57G30	E... .S.P..........I.... ..S..G--.Y.N....S.EKS ....E.N.--ST...T..Q...A.........S.....I*.K			PS	
	M34976(91A3)[d]	E ....S.. ..KT.S...M.............--SGIN.......Q......Y.H.--GK.YIH...R..G.T......S.........R..........F ..	91A3 CRI-	(0)	F	82
	J558-15[d]	E .. S....GR..S.......T........--.GIN.......QD. ...Y.Y.--GN.Y.A.....QGE....S.TS. ....R ..... .I.F .			NF	83
	H130/H18[b]	E.....S.P..... ....I.......S..G--.F.N..M.SH.KS... ...N.--YN.D.F..Q...G. ......S....H.E.R..A.	A12	(1)	F	84
	C57G15/C57G18	E ...S.P ........I..M....Q.SD--.Y .*...SH.KS.....Y.N.--.N.C S..Q...G.......TS. ...E.H			PS	
	C57C44	E ..S.P..........I..M....S.SD--.Y..*...SH.KS.....Y.N.--.N.C.S..Q...G.......TS......E H			PS	
	VH108B[b]	E.....S.P .. .. ..IT....D.S..G--.I.N....SH.KS.....E.N.--YN...S..Q...G ......TS......E.H.......L . ..			PS	
	H26-6[b]	..T.....I.......S..G--.Y......SH.KS.....Y.SC--YN.A.S..Q...G...F...TS.......FN......... ...			NF	85
	ASB9[b]	... ..S....M ......I....T..K.S.-- NIE......EQ ... .E.L.--G.DY.Y.I....G...F.A.TS.N. .. .G.			NF	86
	C57G22	E... .S.P..........I..M...LS.SD--.Y..*...SH.KS.....Y.N.--.N.C.S..Q...G.......TS......E.H			PS	
	pHC102[b]	E.H...SLPKV..A P...I.......S..G--.Y. ....SH.KI.QR.EYVN.--YN...G. ...D.....A..SF.... .F. .....L			PS	
	VAR34	E...K...TVV........I..Q... S..G--.Y......SHEKS.*.. L.I.--YN.N.SN.Q...G.........S....N.E.C			PS	
VH2	Q52	QVQLKESGPGLVAPSQSLSITCTVSGFSLTG--YGVNWVRQPPGKGLEWLGTIW---GNGSTDYNSTLKSRLTITKDNSKSQVFLKMNSLQTDDTARYYCASV				
	PJ14/V00767	. . ..--.......... .. .. M..---.D.......A ....S S................ ... . ..R	D1.3	(2)	F	87
	VHOx-1[b]/VOx-1[b]	. ...S-- ..H ... .V. ---AG...N...A.M...S.S.... .. ..M .. R	DB1-453.2	(0)	F	88
	M37808[b]	. ... .S--.. H......... ... .V. ---SD...N.I.A.....S.S... ... . M . R			PS	
	M26982[b]	. .T.... ....I...S--.. H . .. VV. ---SD.. N. A.. .S.S .. .. .M. ..R			PS	
	M26984[b]	.. .Q. ...*. ....S-- , H....S.... . ..V. ---SG.. ....AFI...S.S .. ... .F .. .A. M. ..K			PS	
	M26981[b]	Q.. .. Q. . ...S-- ..H ...S .. ..V .---SG. ..AAFI ..S.S... . F A. .I .R			PS	
	V(Ox2)[g]	.. Q......Q... . .. .S-- .H.. .S .. . V..---SG..... AAFI. S S ... .F . AN I			PS	
	Q5SH.100[h]	.. ..Q . .. .F. ...Y . .S-- EI. . . . ..V..---TG ..N ..A.I...S.S. .. L I VR			PS	

Fig 1(c)

```
Ig-fold^c                      B B     B  T  B B B B 11111111 B BIBI  T  I IBB  22222222        TT B        B B B  T   T B BIBI
Position^d                     1       10        20        30         40        50          60        70        80         90
Family^1  Name^4               |. |.. |    | . |    |....|.ab .|    .|....|....|..abc..|.   | . .|....|....|....|..abc..|....|....|  Rearranged gene^f  Status^g
VH2       Q52 (continued)      QVQLKESGPGLVAPSQSLSITCTVSGFSLTG--YGVNWVRQPPGKGLEWLGTIW---GNGSTDYNSTLKSRLTITKDNSKSQVFLKMNSLQTDDTARYYCASV
V(Ox2)^g                        .. Q      Q        ..    . . .S--   H   .S. ..  .  V..---SG......AAFI  .S.S. ...   ..F ... .AN...I... R                        PS
VH101^b                        ... .Q   .  Q . .  .... ... . S--...H..  S.........V..---SG.... .AAFI. .S.S. .    ...F......SN...I . R   D23          (1)   F        89
VH3       36-60                EVQLQESGPSLVKPSQTLSLTCSVTGDSITS--DYWNWIRKFPGNKLEYMGYIS---YSGSTYYNPSLKSRISITRDTSKNQYYLQLNSVTSEDTATYYCTSL
36-60                           ..   .        ...  ...... ....--. ...  .....  ........---........,......... ......... .. . ... ......... .AR  Pab 419      (4)   F        90
VH-36-60^d                          .    ..  ...  ... ... .  .  --    ...   .   .........---  .. ,......................  ..............P..  A               NP^SD    91
VH-36-60^b                     ..... . ...     ..         ... ...--G. .  ...   ..........---........ .......... .............T  .P....A                 NP^SD    92
SB32^b                         D .. ..X.X . .  S.....T.  Y  ..D-YA   . Q ......W.....---.... S  .. ..... ........FF...                     LB8          (2)   F        93
VH3A1^I                        D.... .. G  .....V....T...I...TGNYR S  .Q.... ..WI...Y---.. TIT.....T..TT ........FF.EM .L.A.....             NEO C72-3A1  (0)   F        94
VH4       X-24                 EVKVIESGGGLVQPGGSLKLSCAASGFDFSR--YWMSWVRQAPGKGLEWIGEINP--DSSTINYTPSLKDKFIISRDNAKNTLYLQMSKVRSEDTALYYCARL
V-H 441/V441^b                  .LL.....   .............. ...--....... ..... .........--..................................  .  ... ..     XRPC44       (0)   F        95
VH55^b                         . .LL ..         ..N...    ......--. ...A....  .Q........--G............. ......... .. ......... ..... .    XRPC24       (0)   F        96
VH5       7183                 DVQLVESGGGLVQPGGSRKLSCAASGFTFSS--FGMHWVRQAPEKGLEWVAYISS--GSSTLHYADTVKGRFTISRDNPKNTLFLQMTSLRSEDTAMYYCAR
61-1P^b                                     ..      .  .... ....-- . ........  .........-- .. IY..........  . ...... ............   ...  RF-3 PAN     (1)   F        97
98-3G^b                                                 .....   --YA.S....S...R..  ..E...--.G.YTY.P...T.........A....Y.E.S.............  H37-40       (0)   F        98
76-1EG^b/VH7183.9^b                       . K    L.. .  ...   .. --YA.S   .T...R.....T...--.G.YTY.P.S ...... .  A.  ..Y...S... . ........  H37-45       (3)   F        99
VH7183 13^b                                .K.. .L .  ..........--YT.S....T...R.........N--.GGSTY.P........... .A....Y...S..K..........  ASWA2        (5)   F       100
VH10-19^b                       K........ K... L.....  .   .....--YT S.. .T...R.....T...--.G.YTY.P.S.... ......A  ..Y...S..K... . ...T.  H35-C6       (0)   F       101
Vh7183(VH69.1)^b               E.K. ..    ..K    L  ............--YT.S....S...R.....T...--.G.YTY.P.S......... .A....Y...S..K....... .T.                     NF      102
VH7183.14^b                                .K... L.........A...--YD.S.   T...R.....T ..--.G.YTY.P.S.    ..... .AR...Y...S. ......L.... .  H37-60       (3)   F       103
VH283                          E.M   . .. K.   L. ........... --YT.S.  .T ..R.....T...--.GGNTY.P.S...........A..N.Y...S.......L.. ..    MRK16        (1)   F       104
VH37.1^b                       E.K.   .. ..K .. L ....T.... . --Y..S....T.. R.....T..G--.G.YTY.P.S....... ...A..N.Y...S.......L.....                      NF      105
VHB4-psi^b                     EL...  ....  .. .................--YA S....T...R.....A..T--DG.FIY*P.............A........S. .Y. ......L.                      PS
VH7183.11^b                                K..  .  ....  ......D--Y..A......G..P....P...--LAYSIY.....T.......E.A....Y.E.S ............                     NF      106
VH50.1^b                       E.K..        .. ...L     T  ..  D--YY.Y....T ..R..... ..N--.GGSTY.P. ......... A... Y...SR.K .. .......     BSPv         (0)   F       107
VH7183.10^b                                .K.. .L.. ...........--YY.S....T...R..L..A.N.--NGGSTY.P.... .........A....Y...S..K.....L... .   B13          (1)   F       108
57-1M^b/VH7183.12^b                        .K.  .L.... ........--YA.S  ..T.  .R.....S..---SGGSTY.P S ..........AR I.Y...S........... .    AN10         (3)   F       109
68-5N^b                                   .....  L.  ...  ........--Y..S....T.D R..L..T.N.--NGGSTY.P.S.......... A... Y...S..K......... ..  BI1279       (4)   F       110
VH81X^b                        E.....  ... ..RE.L.  ..ESNEYE.P.--HD S.  .KT...R..L..A.N.--DGGSTY.P .MER..I.....T.K..Y...S........L....                   NP^SD   111
VH6       J606                 EVKLEESGGGLVQPGGSMKLSCVASGFTFSN--YWMNWVRQSPEKGLEWVAEIRLKSNNYATHYAESVKGRFTISRDDSKSSVYLQMNNLRAEDTGIYYCTTG
VH22 1^b                       ....... ....  ........ ... .--...S....  ........Q....D.................. ...     .. ............ .G      68.2D8       (4)   F       112
VH7       S107                 EVKLVESGGGLVQPGGSLRLSCATSGFTFSD--FYMEWVRQPPGKRLEWIAASRNKANDYTTEYSASVKGRFIVSRDTSQSILYLQMNALRAEDTAIYYCARD
V1^b/pBV132^b                  ............. .    .. .........--................. ...... ......... .. .. ................ ...............  NQ10.3 8     (0)   F       113
V11^b/pBV19B4^b                .. .      . ....    .    .   ....T --Y..S...   ..A...LGFI.....G.............TI ..N......... T.. ..S T . ...  H220-7       (0)   F       114
V13^b                          ....M. .. .  ...A......EA.  . .T.--Y .S    L.R.SP..L.LI.....G.............TI...N..N.......T....AS.T....K                 NP^SD   115
V3^b                           ...    ...  .    A    .   .S.....T.--Y. N. HR....P...L.LI.... .G.I......M....TI...N..........T ST..S.T... ..                PS
VH8       3609 7               QVTLKESGPGILKPSQTLSLTCSFSGFSLSTSGMGVGWIRQPSGKGLEWLAHIW---WDDDKYYNPSLKSQLTISKDTSRNQVFLKITSVDTADTASYYCARV
CB17H-3^a                      ...... ... Q.... .......     . F. .I.. ......   .......---..........A...R........ N.   . .            T                     NF      116
CB17H-1^a                      .......... .QS        . ..............S...............Y---.... R.. ....R..................  .   .T.           ASWB1        (1)   F       117
CB17H-10^a                     .....       Q. . .   ....        N..I....    ...........---.N. ..........R........ N .. .. .T ..... T.        B6.2         (10)  F       118
CB17H-8^a                      ...    .. ..QS.. ......       N......S..... . .........Y---. ...R.......R...     . .     ........ ...T                  NF      119
CB17H-6^a                      .   .        QS             V   ...F   .S..          .....Y---  E .H K ....R..  .   N.     ....  ..   T                   NF      120
CB17H-9^a                       ....    . .Q...    .       V   N F .  S. ..    ..........Y---..E. H.K.....R...... N  . ....T.    ...T.                  NF      121
VH3609                                       . .P .. T          Y. M S MC   .   V ..L. ---CNN .G.. F.R ... .    .N             F.   T                     PS
```

Fig 1(d)

```
Ig-foldα                       B B     B  T  B B B B 11111111 B BIBI  T  I IBB  22222222         TT B         B B B  T    T B BIBI
Positionβ                      1       10        20         30         40         50          60         70         80        90
Familyχ  Nameδ                 | . |....|....|    |     |....| ab   |....|. .|    | .abc. |    |    |. ..|    |. .| abc. | .. |    |  Rearranged geneε  Statusφ
VH8      3609 7 (continued)    QVTLKESGPGILKPSQTLSLTCSFSGFSLSTSGMGVGWIRQPSGKGLEWLAHIW---WDDDKYYNPSLKSQLTISKDTSRNQVFLKITSVDTADTASYYCARV
V31b /VMU-3 1                  .. ..V.  ...Q   . G.A T . I........LS.L.K.Q.R-.. ..S. ----NN.N..      R .. .E..N.     L..     .STT ...WR                         PS
VH9      GAM3-8                QIQLVQSGPELKKPGETVKISCKASGYTFTN--YGLNWVKQAPGKGLKWMGWINT--YTGKSTYADDFKGRFAFSLETSAITAYLQINNLKNEDMATYFCTRS
VFM11b/VGK1Bj                   ...... .. .     .    . . ....--..M .........      . .--...EP... .. ..  .. ....S.       ......... .A.   L6'             (2)   F    122
VMS9b/VGK1Aj/251b              ......   .  .     .      . .. --. M.  .....  .   .....--.. EP           ..... ..S . ..........T.....A.  RFT2            (2)   F    123
161b                           .........   .     .   .........--..M  .... ..   .......--.. EP.. . .........   C S      .... Q.T .                          NF   124
VGK6j                           ..... .  .      ..............--..M........  .. ..... .--E .EP.........  . . ...S....... .... T... .   L69             (7)   F    125
VMS2b/VGK4j                    ..    .. ..   . ............. --. M........ ......... --N..EP...EE.... .........S..  .........T. ...A.  2E7             (3)   F    126
264b                             ...... .. .. .  .  .... .. T--..MS... .  ....     ...--.S.VP.....  ...........S.. ..........T.....A.  TE32            (1)   F    127
VFM1b/281b/VGK7j                 .... . ..  .   .......     .. D--.SMH.... .........  .--E..EP.... .  ...... ...S.. ..........T.... A   C55-7B3         (3)   F    128
VMS1b/141b/VGK3j               .. .........   .    .   .........--.AMH ..............KY --N..EP..G.      .....    S ..................A                       NF   129
VGK5j                            ......  .    ...     .......T--A.MQ..QKM.......I ....--HS.VPK. E............ ..S..........  . .                          NF   130
VGK2j                            ........  .....R........   T--A.MQ..QKM... ...I... .--HS.VPK..E ...... ......S......S......T.....     AN08            (5)   F    131
VH10     MRL-DNA4              EVQLVETGGGLVQPKGSLKLSCPASGFSFNT--NAMNWVRQAPGKGLEWVARIRSKSNNYATYYADSVKDRFTISRDDSQSMLYLQMNNLKTEDTAMYYC
MRL-RF24BGk /M21469            ......VWWRM.    ......A....T...--Y........  ..  .  ... ..S... .... .......... .... ... ....     .......                        PS
VH11     CP3                   EVQLLETGGGLVQPGGSRGLSCEGSGFTFSG--FWMSWVRQTPGKTLEWIGDINS--DGSAINYAPSIKDRFTIFRDNDKSTLYLQMSNVRSEDTATYFCMRY
VH12     CH27                              KPXQXLXXTCSITXFPITSG-YYWIWIRQSPGKPLEWMGYIT---HSGETFYNPSLQSPISITRETSKNQFFLQLNSVTTEDTAMYYCXGD
16-A                            GAVQESGPGLV.NS S.FLA  ..G .....*........ .............---.WENFLQPIPSRA.S........................  ..A..                       PS
VH13      vh3609N              QVQLVETGGGLVRPGNSLKLSCVTSGFTFSN--YRMHWLRQPPGKRLEWIAVITVKSDNYGANYAESVKGRFAISRDDSKSSVYLQMDRLREEDTATYYCSRG
VH14      vhsm7-13             EVQLQQSGAEVV-PGASVKLSCTASGFNIKD--DYMHWAKQRPDQGLEWIGRIDP--AIDDTDYAPKFQDKATMITDTSSNIAYLQSSSLTSEDTAVYYCPY
vhsm7-13l                       ...........*.. .. ...........--.............. ...--........ .................... ....... .  .P..                        NFSD  132
H2b-3b/VH2b-3b                 .. .  . ..L RS..........  ..   --Y....V....E. .....W...--ENG..E.   ..G. ..TA.. ..T....L.....   .....NA   MRL-Histone7 (7)   F    133
37A4b                             ..  .  .K..................--S....V....E...........--.NGN.K.D....G...ITA.. . T.H..L.R....  .                           NF   134
VH10b/H10b/M33391-7l           ...... .  L K.. ......... ..  .--T....V....E.......... --.NGN.K.D. ..G...ITA... .T.   L . ......    AR   87.92 6         (0)   F    135
17C1b                          ..........L K  .  . .. ........--T....V. ..E.......V...--.NGIPI.D.........ITA.....T..............  ..AR                        NF   136
14C3b                          ..........L.K...... ............--T....V....E.......V...--.NGFPN D  ..G. .ITA...  T.. ........ . ...AR                        NF   137
VH4a-3b/H4a-3b                  .  ... ..L.R...L...  K. ......--Y....V....E.......W...--ENGN.I.D....G. SITA.....T.. .L............AR   2G8            (15)   F    138
VH15      Vh15A                QVHLQQSGSELRSPGSSVKLSCKDFDSEVFPI-AYMSWVRQKPGHGFEWIGDILP--SIGRTIYGEKFEDKATLDADTVSNTAYLELNSLTSEDSAIYYCARD
```

Fig. 1(e)

Fig 1 Multiple amino acid sequences alignment of mice V_H germ-line gene segments. (α) Positions primarily responsible for the variable immunoglobulin fold (V-Ig-fold) conserved features (Chothia *et al*, 1988) and hypervariable loop definition (Chothia and Lesk, 1987). Within this, B stands for residues buried within the protein, T residues in turns, I Inter-domain residues, V. residues between B and C domains (Chothia *et al*, 1988), 1: H1 and 2: H2 definition (Chothia and Lesk, 1987) (β) Residue numbering as in Chothia and Lesk (1987). (χ) V_H family and prototype sequences. (δ) Name, clone or sequence access number in Genbank, or name of the sequence in the literature. Superscripts in the name of the sequence indicate the strain of the origin of each of the sequences as follows a. C57BL/6; b: BALB/c; c. C57BL/6J, d. A/J, e MRL/MpJ-LPR/LPR; f: MRL-LPR/LPR; g. BALB/cJ, h· NFS/N; i· BALB/b, j. BALB K, k MRL/MP-lpr/lpr, l· MRL/lpr, m C57BL/6 × BALB/c Only residues which diverge with respect to the prototype sequences of the family are represented (ε) Name (in the Kabat's Database) of the closest V_H rearranged gene and number of amino acid differences between this and the germ-line gene (φ) F stands for sequences with a rearranged counterpart (functional), NF Non-functional sequence due to not having a rearranged counterpart. Superscript "S.D " means structural defects, this underlined in the sequence. PS· Pseudogene Insertions or deletions that produce frame shift changes in the amino acid sequence were eliminated to obtain the most correct immunoglobulin-like sequences Aesthetics within the sequences means a stop codon Numbers at the right most part represent the code of each sequence in Fig 3 The multiple sequences alignment and all the calculations therein presented were made by using the VIR package (Almagro *et al*, 1995)

elements or recombination signals (Tomlinson *et al.*, 1992)

In vivo expression of the V_H gene segments was performed by assigning their acid sequences to their closest rearranged functional V_H sequence in a database of 627 V_H amino acid sequences compiled from the Kabat's Database on-line service (Kabat *et al.*, 1991; see web site http://immuno bme.nwu.edu). We chose the V_H rearranged sequences having a reported specificity, in order to avoid non-productive rearrangements, therefore guaranteeing assignment of functional V_H gene segments only The database with the 627 V_H amino acid sequences is available on request to the authors

It is worth mentioning that the criterion for choosing the V_H rearranged sequences, as those having a reported specificity may bias the assemble of rearranged sequences due to researchers' interests. However, inspection of the 627 sequences indicates 137 different specificities there included. Moreover, many of the sequences reported as possessing the same specificity probably correspond to antibodies elicited against different epitopes, particularly in the case of large antigens like proteins. This increases the actual amount of different specificities. Therefore, the database of V_H rearranged sequences would be sufficiently heterogeneous to detect most of the functional V_H gene segments of mice.

To determine structural defects, we analyzed those residues mainly responsible for the structural conserved features of antibodies V domains (Amzel and Poljak, 1979; Chothia and Lesk, 1987; Chothia *et al.*, 1988). Such residues were derived early from the analysis of the V_L and V_H domains of the seven antibodies of known three-dimensional structure (Chothia and Lesk, 1987; Chothia *et al.*, 1988). However, the pattern depends to some extent on the number of structures analyzed. Currently, there exist atomic structures of more than 50 antibodies with different amino acid sequences, thus allowing to update the pattern In addition we decided, to further improve updating, to add the 627 V_H amino acid sequences compiled from the Kabat's Database. This was done supposing that these sequences, having a reported specificity, are functional and should have no structural defects. The pattern is summarized in Table 1.

Table 1 Classification and repertoire of the mice V_H gene segments

V_H Gene	Prototype	Number of V_H gene segments	
family[a]	member[b]	Estimated[c]	Found[d]
V_H1	J558	60–1000	120
V_H2	Q52	15	10
V_H3	36-60	5–8	5
V_H4	X-24	2	2
V_H5	7183	12	16
V_H6	J606	10–12	1
V_H7	S107	2–4	4
V_H8	3609	7–8	8
V_H9	GAM3-8	5–7	10
V_H10	MRL-DNA4	2- 5	1
V_H11	CP3	1 -6	—
V_H12	CH27	1	1
V_H13	3609N	1	-
V_H14	SM7	3–4	7
V_H15	V_H15a	2–3	–
Total		123–1073	185

[a] V_H gene families defined for mice V_H1 to V_H14 (Kofler *et al*, 1992) and V_H15 (Mainville *et al*, 1996).

[b] Name of the prototype sequence of each family

[c] Number of sequences estimated by Southern blot hybridization and sequencing V_H families 1–14 (Kofler *et al*, 1992) and V_H15 (Mainville *et al.*, 1996)

[d] Number of V_H germ-line genes and V_H pseudogenes found in our compilation (see Fig 1)

Determination of the canonical structures in H1 and H2

In structural terms, H1 has been defined as the hypervariable loop beginning at position 26 and finishing at position 32 (see head of Fig 1) Three different sizes have been identified for this loop canonical structures type 1 (seven residues), type 2 (eight residues) and type 3 (nine residues) (Chothia and Lesk, 1987, Chothia *et al*, 1989; Chothia *et al.*, 1991)

On the other hand, H2 is defined from a structural point of view as the hypervariable loop running from position 52 to position 56 (Chothia and Lesk, 1987; Chothia *et al*, 1989) Currently, five different sizes have been found (Chothia *et al*, 1992; Tramontano *et al.*, 1990) Early works assigned canonical structural type 1 to the shortest loop (5 residues), the next length (6 residues) to types 2 and 3 (these types share the same length but differ in their conformation), and type 4 was identified with the longest loop (8 residues). Recently, two other sizes for H2 have been distinguished in the functional V_H gene segments of humans. one having 7 residues (between the size of types 2/3 and type 4) and named type 5 (Chothia *et al.*, 1992), and another one shorter than type 1 (4 residues) named type 6 [I M. Tomlinson, personal communication]

The patterns of residues determining the different canonical structures for H1 and H2 have been described in detail by Chothia *et al* (1992) Starting from this pattern, we defined a new one (Fig 2) This new pattern includes the recent analysis of Barré *et al.* (1994) in shark V_H sequences, as well as our own analysis of recently solved V_H X-ray structures (underlined amino acids in Fig. 2). For example, in H2, Valine (v) was added at position 71 in the pattern of type 2 because Fab 8F5 (Tormo *et al*, 1992) has this residue This residue was not previously considered in the patterns (Chothia and Lesk, 1987; Chothia *et al*, 1989; Chothia *et al*, 1992) and does not modify the H2 conformation [the rms of the 8F5 in H2 when compared with NC41, a prototype of H2 type 2 (Chothia *et al*, 1989), is 0 36 Å].

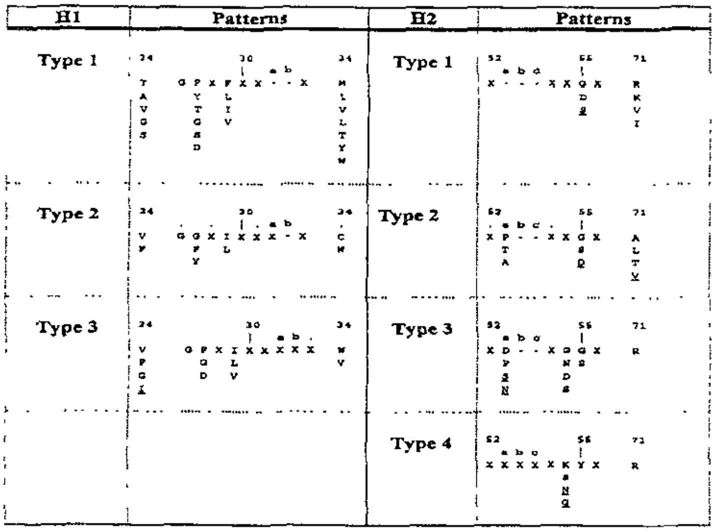


Fig. 2. Amino acid pattern for the canonical structure classes as defined as the simultaneous combination of canonical structures in a given sequence (Chothia *et al.*, 1992) The amino acid residues are shown in one letter code. X means any residue Underlined residues are those differing with respect to the original pattern (see Material and Methods for details)

RESULTS

The known V_H germ-line gene segments of mice

Although the exact number of gene segments in the entire mice V$_H$ germ-line gene repertoire is currently unknown, the complexity of most individual V$_H$ families has been established within a narrow range for several strains of the mouse (Kofler *et al.*, 1992). Only the size of the largest family (V$_H$1) is controversial, varying from 60 (Brodeur and Riblet, 1984) to ~1000 members (Livant *et al*, 1986). Several lines of evidence suggest, however, that the size of the V$_H$1 family is closer to 60 than to 1000 (Kofler *et al.*, 1992).

Based on the estimated complexity of the individual V$_H$ families of mice, we first established how representative our compilation of mice V$_H$ gene segments really was (Table 2). In most V$_H$ families the estimated number of genes and the amount we found are in good agreement. We compiled 120 V$_H$ gene segments in the V$_H$1 family (Fig. 1), supporting the proposition that the size of this family is indeed closer to 60 members than it is to 1000 (Kofler *et al.*, 1992). In other 9 V$_H$ families (V$_H$2, V$_H$3, V$_H$4, V$_H$5, V$_H$7, V$_H$8, V$_H$9, V$_H$10 and V$_H$12) the established quantities of V$_H$ gene segments are also similar to those we found (see Table 2), suggesting these 9 V$_H$ families to be well represented in our compilation.

Four V$_H$ families (V$_H$6, V$_H$11, V$_H$13 and V$_H$15) showed discrepancies when the estimated and found complexity were compared (see Table 2). In the V$_H$6 family, less segments than expected were assigned. For the V$_H$11, V$_H$13 and V$_H$15 families, no V$_H$ gene segments were found. Nonetheless, these families have one or only a few members (Table 2) and therefore their contribution to the whole mice V$_H$ germ-line gene repertoire should be minimal.

The functional V_H germ-line gene segments of mice

Analysis of the expression *in vivo* of the 138 V$_H$ gene segments reported as germ-line genes and potentially functional, suggests that only 72 of them are functional (Fig. 3) Of the 66 V$_H$ gene segments not expressed *in vivo*, and therefore defined as non-functional, 13 present structural defects when those residues responsible for the structural conserved features of V$_H$ domains are analyzed (see Table 1 and the status column of Fig. 1). For example, 3 sequences within the V$_H$1 family (VH145/C1egc, C19c and C15c; see Fig. 1) possess Serine (s) instead of Cysteine (c) at position 22. These sequences are unable to establish the disulfide bridge that stabilizes the standard fold of V$_H$ domains (Amzel and Poljak, 1979; Chothia and Lesk, 1987).

In the remaining 53 sequences not showing structural defects, it was difficult to define why they had not any counterpart in the V$_H$ rearranged sequences. Hence, we can only infer that they have minor genetic defects outside the coding region. This hypothesis however, could not be properly scrutinized because information outside the coding region is not reported in many sequences. In such way we cannot discard the possibility of some of these

Table 2. Pattern of residues determining the structural features of the V-Ig-fold[a]

	Intra-domain positions											
Position	Residues buried between the β-sheets											
4	L	M	*V*	*P*	*F*	*S*	*H*					
6	Q	E	*N*	*P*	*R*							
12	G	A	V	M	*T*	*L*	*I*	*E*	*K*	*C*		
18	L	V	*A*	*I*	*M*	*R*	*K*	*Q*				
20	L	I	M	*V*								
22	C											
24	G	A	S	T	V	*P*	*F*	*D*	*I*			
34	V	L	I	M	*A*	*F*	*W*					
36	W											
38	R	K	V	*I*	*M*	*N*	*Q*	*S*	*T*			
48	L	V	I	W	*R*	*M*	*F*	*S*				
49	G	A	*S*	*D*	*T*	*V*						
69	G	A	V	I	M	F	*L*	*S*				
78	A	L	F	Y	*V*	*T*	*I*	*G*	*S*			
80	L	M	*F*	*I*	*S*	*T*	*V*					
82	L	I	M	*V*	*F*	*S*						
88	G	A	S	*T*	*V*							
90	Y	*F*	*H*	*N*								
92	C											
	Residues in turns											
15	G	*S*	*K*	*E*	*N*							
42	G	*D*	*E*	*H*	*A*	*K*	*V*	*Q*	*R*	*S*	*T*	*W*
66	R	K	*A*	*E*	*H*	*Q*	*T*					
67	V	F	*T*	*S*	*L*	*I*	*A*	*G*				
82c	V	L	M	*A*	*P*	*I*	*T*					
86	D	*E*	*S*									
	Inter-domains positions Between variable domains											
37	V	*I*	*F*	*M*	*A*	*L*						
39	Q	*H*	*E*	*K*	*L*	*P*	*R*					
45	L	*F*	*R*	*P*	*Q*							
47	W	*Y*	*F*	*I*	*H*	*C*	*L*	*S*				
91	F	Y	*H*	*I*	*S*							
93	A	L	*V*	*T*	*D*	*K*	*S*	*G*	*H*	*M*	*N*	
	Between V_H and C_{H1} domains											
11	L	V	*I*	*S*	*F*	*P*	*T*					

[a] Residues differing with respect to the original pattern described by Chothia *et al.* (1988) are underlined In italic those residues identified in the 627 rearranged sequences.

V_H genes actually being functional even though no rearranged counterpart was found. This is so because the database of rearranged sequences was built chosen those sequences having a reported specificity to avoid non-productive rearrangements, in spite of the fact that this would introduce some bias due to the researcher's interests However, the sample of rearranged sequences would be sufficiently heterogeneous (see Material and Methods section) to lead to the conclusion that, if some of the V_H genes defined as non-functional are indeed functional, they should be exceptional.

The structural repertoire of functional V_H germ-line gene segments

In Fig. 4 the canonical structure classes implicit in the 72 defined as functional V_H germ-line genes of mice are shown. Seventy-one of them present patterns compatible with some canonical structure in H1. In H2, three sequences do not have a proper pattern to fit any of the canonical structure known to exist.

Analysis of the structural repertoire indicates that mice encode 6 canonical structures classes. Class 1 2 is the most frequent (64%), followed by class 1–3 (17%) and class 1–1 (7%). Classes 1–4, 3- 1 and 2–1 are very poorly represented in the sequences (3%, 3% and 1%, respectively).

Interestingly, the structural repertoire of mice is not randomly distributed among the V_H families. Almost all sequences within a family encode the same canonical structure class (Fig. 4). Therefore, their structural repertoire is family-specific, suggesting it to be preserved despite actual diversification of the V_H gene segments.

Comparison between the structural repertoire of mice and humans

To compare the structural repertoire of mice and humans, those canonical structure classes implicit in the 51 functional V_H germ-line genes of humans are depicted in Fig. 5. Differently from mice, humans encode 8 canonical structure classes (Fig. 6). Canonical structure classes 3–5 and 1–6 implicit in human sequences were not found in the functional V_H mice germ-line genes.

Canonical structure class 3–5 is encoded by germ-line 6-01/DP74; the only gene segment defining the human V_H6 family (see Fig. 5). In mice, neither the sequences nor the pseudogenes compiled in Fig. 1 possess the proper size to fit canonical structure 5 in H2. Inspection of the 627 functional rearranged V_H mice sequences indicates this size not to be present either. Therefore, it is unlikely that mice germ-line genes possess this class.

In the case of canonical structure class 1–6, one doubly sequenced pseudogene of mice (V31/VMU-3.1; Fig. 1) has 4 residues at the H2 loop which is the size corresponding to canonical structure type 6. Because this size is found in 5 functional rearranged V_H sequences [PY54, PY2 (Ruff-Jamison *et al.*, 1991); 8H3 (Mukherjee *et al.*, 1993) 246B.4g, 245F.6g (Limpanasithikul *et al.*, 1995), it would be responsible to expect that this pseudogene has its functional counterpart in some mouse or in certain strains of mouse. Alternatively, the pseudogene encoding this loop size might had given the segment comprising H2 to some functional gene segment by somatic gene conversion (Weill and Reynaud, 1996), so generating the rearranged sequences presenting this canonical structure class.

Differences between the structural repertoire of humans and mice are also found in the proportion by which these species encode classes 2–1 and 3 -1 (Fig 6) Class 2–1 is encoded only by one gene segment belonging to the mice V_H3 family whilst class 3–1 is encoded by two sequences: those belonging to the V_H8 family (see Fig 4)

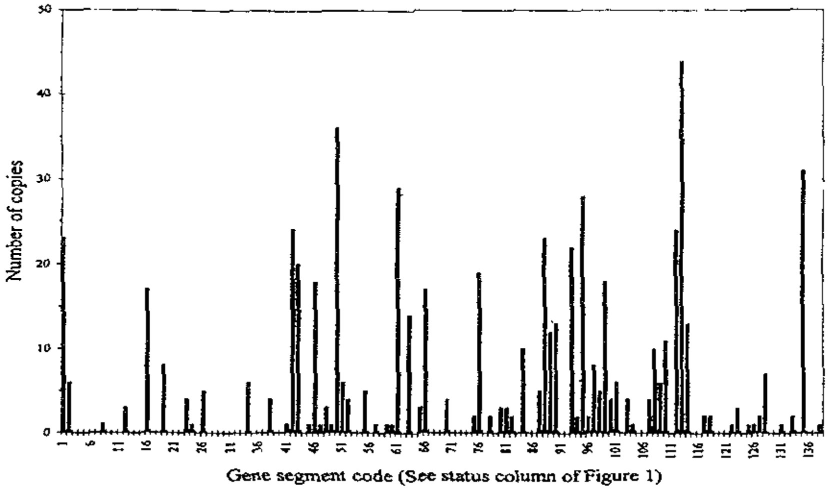


Fig. 3. Usage of V_H germ-line gene segments of mice.

			H1		H2		V_HCSC^a
		B	111111111	B	22222222	B	
osition		24	30	34	52 55	71	
amily	Name	.	....|.ab.	.	.abc..|.	.	
1	VH-186-2	A	GYTFTS--Y	M	DP--NSGG	V	1-2
1	C36a/B7c	A	GYTFTS--Y	M	DP--NSGG	V	1-2
1	C31e	A	GYTFTS--Y	M	HP--NSGS	V	1-2
1	V23	A	GYTFTS--Y	M	NP--SNGG	V	1-2
1	B25c	A	GYTFTS--Y	M	NP--SNGR	V	1-2
1	C11c	A	GYTFTS--Y	M	YP--GSSS	V	1-2
1	B12c	A	GYTFTS--Y	M	DP--SDSE	V	1-2
1	C16c	A	GYTFTS--Y	I	YP--GSGS	V	1-2
1	VH124	A	GYTFTS--Y	M	DP--SDSY	V	1-2
1	p2MS	A	GYTFTS--Y	I	YP--GSSS	V	1-2
1	H30	A	GYTFTS--Y	M	NP--SSGY	A	1-2
1	B16e	A	GYNFTS--Y	I	YP--GSGS	V	1-2
1	H8	A	GYTFTS--Y	I	YP--GNVN	-	1-?
1	H9	A	GYTFTS--Y	I	YP--GDGS	A	1-2
1	VH105	A	GYTFTS--Y	I	YP--RDGS	A	1-2
1	H13-3	A	GYTFTS--Y	M	NP--SSGY	A	1-2
1	J558-43y	A	GYTFTS--Y	I	YP--GDGN	A	1-2
1	VH3	A	GYTFTS--Y	M	YP--SDSE	V	1-2
1	B23c/B18c	A	GYTFTD--Y	M	DT--SDSY	V	1-2
1	VH-Id-11	A	GYTFTS--Y	I	NP--GNGY	V	1-2
1	J558-28	A	GYVFTN--Y	I	NP--GSGG	A	1-2
1	B1c	A	GYTFTS--Y	I	YP--LDSN	V	1-2
1	pM11	A	GYTFTN--Y	I	YP--GGGY	A	1-2
1	B26c	A	GYSFTS--Y	M	HP--SDSE	V	1-2
1	VH-Id-7	T	GYTFTS--Y	I	YP--GNGY	S	1-?
1	BALB71	A	GYTFTD--Y	M	NP--NNGA	V	1-2
1	C57G5	A	GYKFTD--Y	M	NP--NNGG	V	1-2
1	BALB17	T	GYTFTE--Y	M	NP--NNGG	V	1-2
1	C57C27	A	GYTFTD--Y	M	NP--NNGG	V	1-2
1	37A11	A	GYTFSS--Y	I	LP--GSGS	A	1-2
1	GLvh50	A	GYTFTD--Y	M	YP--NNGG	V	1-2
1	C57C2	A	GYSFTG--Y	M	NP--NNGG	V	1-2
1	C57G9	A	GYTFTD--Y	M	NP--NNGG	V	1-2
1	J558-43X	A	GYTFTD--H	I	SP--GNGD	A	1-2
1	BALB11	A	GYTFTD--Y	M	NP--NYDS	V	1-2
1	C57G26	A	GYSFTG--Y	M	NP--STGG	V	1-2
1	M34976(91A3)	A	GYTFTS--S	I	HP--GKGY	V	1-2
1	H130/H18	A	GYSFTG--Y	M	NP--YNGD	V	1-2
2	PJ14	V	GFSLTG--Y	V	W---GDGS	K	1-1
2	VHOx-1	V	GFSLTS--Y	V	W---AGGS	K	1-1
2	VH101	V	GFSLTS--Y	V	W---SGGS	K	1-1
3	VH-36-60	V	GDSITS--D	W	S---YSGS	R	1-1
3	SB32	V	GYSITSD-Y	W	S---YSGS	R	2-1
3	VH3A1	V	GISITTGNY	W	Y---YSGT	R	?-1
4	V-H 441	A	GFDFSR--Y	M	NP--DSST	R	1-3
4	V(H)55	A	GFDFSR--Y	M	NP--GSST	R	1-3
5	61-1P	A	GFTFSS--F	M	SS--GSST	R	1-3
5	98-3G	A	GFTFSS--Y	M	SS--GGSY	R	1-3
5	76-1BG	A	GFTFSS--Y	M	SS--GGSY	R	1-3
5	VH7183.13	A	GFTFSS--Y	M	SN--GGGS	R	1-3
5	VH10-19	A	GFTFSS--Y	M	SS--GGSY	R	1-3
5	VH7183.14	A	GFAFSS--Y	M	SS--GGSY	R	1-3
5	VH283	A	GFTFSS--Y	M	SS--GGGN	R	1-3
5	V(H)50.1	T	GFTFSD--Y	M	SN--GGGS	R	1-3
5	VH7183.10	A	GFTFSS--Y	M	NS--NGGS	R	1-3
5	57-1M	A	GFTFSS--Y	M	S---SGGS	R	1-1
5	68-SN	A	GFTFSS--Y	M	NS--NGGS	R	1-3
6	VH22.1	A	GFTFSN--Y	M	RLKSDNYA	R	1-4
7	V1/pBV132	T	GFTFSD--F	M	RNKANDYT	R	1-?
7	V11/pBV19B4	T	GFTFTD--Y	M	RNKANGYT	R	1-4
8	CB17H-1	F	GFSLSTSGM	V	Y---WDDD	K	3-1
8	CB17H-10	F	GFSLSTSNM	I	W---WNDD	K	3-1
9	VFM11/VGK1B	A	GYTFTN--Y	M	NT--YTGE	L	1-2
9	VMS9/VGK1A	A	GYTFTN--Y	M	NT--YTGE	L	1-2
9	VGK6	A	GYTFTN--Y	M	NT--ETGE	L	1-2
9	VMS2	A	GYTFTN--Y	M	NT--NTGE	L	1-2
9	264	A	GYTFTT--Y	M	NT--YSGV	L	1-2
9	VFN1	A	GYTFTD--Y	M	NT--ETGE	L	1-2
9	VGK2	A	GYTFTT--A	M	NT--HSGV	L	1-2
14	H2b-3	A	GFNIKD--Y	M	DP--ENGD	A	1-2
14	VH10	A	GFNIKD--T	M	DP--ANGN	A	1-2
14	VH4a-3	A	GFNIKD--Y	M	DP--ENGN	A	1-2

Fig. 4. Structural repertoire of the functional V_H germ-line gene segments of mice. (a) V_HCSC: Canonical structure of classes of V_H. ?: means that the loop does not fit the canonical structure pattern. Those residues responsible for the mismatch are underlined

			H1		H2		V_HCSC*
		B	111111111	B	22222222	B	
Position		24	30	34	52	71	
Family	Name	.	....|.ab.	.	.abc..|.	.	
1	1-02/DP75	A	GYTFTG--Y	M	NP--NSGG	R	1-3
1	1-03/DP25	A	GYTFTS--Y	M	NA--GNGN	R	1-3
1	1-08/DP15	A	GYTFTS--Y	I	NP--NSGN	R	1-3
1	1-18/DP14	A	GYTFTS--Y	I	SA--YNGN	T	1-2
1	1-24/DP5	V	GYTLTE--L	M	DP--EDGE	E	1-?
1	1-45/DP4	A	GYTFTY--R	L	TP--FNGN	R	1-3
1	1-46/DP7	A	GYTFTS--Y	M	NP--SGGS	R	1-3
1	1-58/DP2	A	GFTFTS--S	V	VV--GSGN	R	1-3
1	1-69/DP10	A	GGTFSS--Y	I	IP--IFGT	A	1-2
1	1-e/DP88	A	GGTFSS--Y	I	IP--IFGT	A	1-2
1	1-f/DP3	V	GYTFTD--Y	M	DP--EDGE	A	1-2
2	2-05/DP76	F	GFSLSTSGV	V	Y---WNDD	K	3-1
2	2-26/DP26	V	GFSLSNARM	V	F---SNDE	K	3-1
2	2-70/DP28	F	GFSLSTSGM	V	D---WDDD	K	3-1
3	3-07/DP54	A	GFTFSS--Y	M	KQ--DGSE	R	1-3
3	3-09/DP31	A	GFTFDD--Y	M	SW--NSGS	R	1-3
3	3-11/DP35	A	GFTFSD--Y	M	SS--SGST	R	1-3
3	3-13/DP48	A	GFTFSS--Y	M	G---TAGD	R	1-1
3	3-15/DP38	A	GFTFSN--A	M	KSKTDGGT	R	1-?
3	3-20/DP32	A	GFTFDD--Y	M	NW--NGGS	R	1-3
3	3-21/DP77	A	GFTFSS--Y	M	SS--SSSY	R	1-3
3	3-23/DP47	A	GFTFSS--Y	M	SG--SGGS	R	1-3
3	3-30/DP49	A	GFTFSS--Y	M	SY--DGSN	R	1-3
3	3-30.3/DP46	A	GFTFSS--Y	M	SY--DGSN	R	1-3
3	3-30.5/DP4	A	GFTFSS--Y	M	SY--DGSN	R	1-3
3	3-33/DP50	A	GFTFSS--Y	M	WY--DGSN	R	1-3
3	3-43/DP33	A	GFTFDD--Y	M	SW--DGGS	R	1-3
3	3-48/DP51	A	GFTFSS--Y	M	SS--SSST	R	1-3
3	3-49	A	GFTFGD--Y	M	RSKAYGGT	R	1-?
3	3-53/DP42	A	GFTVSS--N	M	Y---SGGS	R	1-1
3	3-64/DP61	A	GFTFSS--Y	M	SS--NGGS	R	1-3
3	3-66/DP86	A	GFTVSS--N	M	Y---SGGS	R	1-1
3	3-72/DP29	A	GFTFSD--H	M	RNKANSYT	R	1-4
3	3-73/YAC9	A	GFTFSG--S	M	RSKANSYA	R	1-4
3	3-74/DP53	A	GFTFSS--Y	M	NS--DGSS	R	1-3
3	3-d	A	GFTVSS--N	M	S----GGS	R	1-6
4	4-04/DP70	V	GGSISSS-N	W	Y---HSGS	V	2-1
4	4-28/DP68	V	GYSISSS-N	W	Y---YSGS	V	2-1
4	4-31/DP65	V	GGSISSGGY	W	Y---YSGS	V	3-1
4	4-30.2/DP64	V	GGSISSGGY	W	Y---HSGS	V	3-1
4	4-30.4/DP78	V	GGSISSGDY	W	Y---YSGS	V	3-1
4	4-30.1/DP65	V	GGSISSGGY	W	Y---YSGS	V	3-1
4	4-34/DP63	V	GGSFSG--Y	W	N---HSGS	V	1-1
4	4-39/DP79	V	GGSISSSSY	W	Y---YSGS	V	3-1
4	4-59/DP71	V	GGSISS--Y	W	Y---YSGS	V	1-1
4	4-61/DP66	V	GGSVSSGSY	W	Y---YSGS	V	3-1
4	4-b/DP67	V	GYSISSG-Y	W	Y---HSGS	V	2-1
5	5-51/DP73	G	GYSFTS--Y	I	YP--GDSD	A	1-2
5	5-a	G	GYSFTS--Y	I	DP--SDSY	A	1-2
6	6-01/DP74	I	GDSVSSNSA	W	YYR-SKWY	P	3-5
7	7-4.1/DP21	A	GYTFTS--Y	M	NT--NTGN	L	1-2

Fig. 5. Structural repertoire of the functional V_H germ-line gene segments of humans. V_HCSC: Canonical structure classes of V_H. ?: means that the loop does not fit the canonical structure pattern. Those residues responsible of the mismatch are underlined

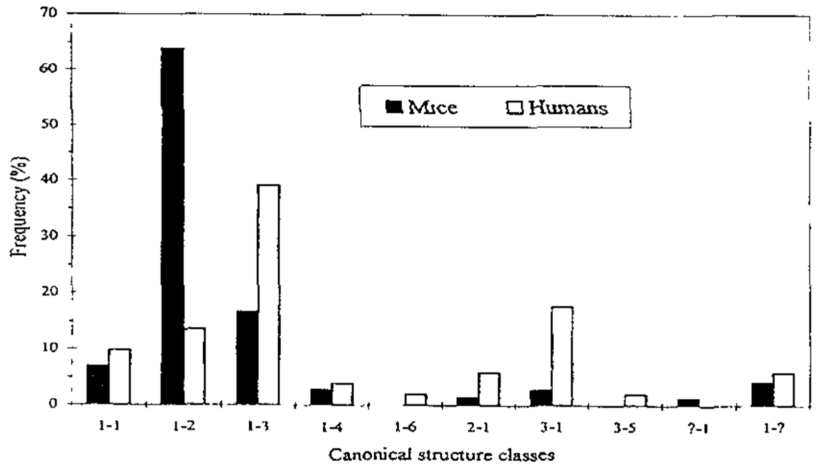


Fig 6 Comparison of the V_H structural repertoire in mice and humans.

In humans, canonical structure class 2-1 is encoded by three gene segments, while class 3-1 is implicit in 9 sequences: 6 from the V_H4 family (half the sequences belonging to this V_H family) and 3 from the V_H3 family (see Fig 5).

Besides the differences described above, classes 1-2 and 1-3 have inverted proportions in humans and mice (Fig. 6). The most common class in mice (1-2, ~64%) has a lower frequency in humans (~14%). Conversely, in humans the most frequent one is 1-3 (~39%), which has a relatively low frequency in mice (~17%). Among those found, this contrast is the most noticeable because it involves roughly half the structural repertoire of mice and humans. Since the human V_H locus has been completely determined (Cook and Tomlinson, 1995), the scope of this astounding difference depends on how complete and precise our compilation of the functional V_H gene segments of mice turns out to be. Nonetheless, several observations support the validity of the difference found.

First, as previously stated, the structural repertoire of mice is family-specific. So, due to the fact that the largest family (V_H1) encodes canonical structure class 1–2 (see Fig. 4), the structural repertoire of mice should remain dominated by this class, although we might have overestimated the number of functional gene segments in this family. Second, the amount of expected and found sequences in those V_H families encoding for class 1–3 are similar (see Table 2). Therefore, the estimation of the contribution of class 1–3 to the structural repertoire of mice should be correct. Third, in families where no gene segments were found (V_H11, V_H13 and V_H15) only the representative sequence of the V_H11 family encodes class 1–3 (see Fig. 1) and this family has from one to six members (see Table 2). Thus, the contribution of this family to the proportion of class 1–3 in the structural repertoire of mice should be marginal. Finally, within those other families in which no gene segments were found (V_H13 and V_H15 families), their representative members encode classes 1–4 and 2–2 (V_H13 and V_H15 families, respectively), therefore, they do not contribute to the total amount of classes 1–2 and 1–3. Altogether, these observations indicate that, when knowledge of the mice V_H repertoire is completed, the difference between humans and mice regarding classes 1–2 and 1–3 might change quantitatively but not qualitatively.

DISCUSSION

In the preceding sections we have shown that humans and mice encode inverted proportions of canonical structure classes 1–2 and 1 3 in their V_H germ-line genes. From a structural point of view, canonical structure classes 1–2 and 1-3 differ at the canonical structure of H2. The canonical structures 2 and 3 are the only two hypervariable loops that, having the same size (Fig. 2), display different conformations (Chothia and Lesk, 1987; Chothia *et al.*, 1989). However, this change does not contribute so much to the variations of the antigen-binding site shape (Vargas-Madrazo *et al.*, 1995a). Thus, the difference found may be fortuitous, i.e., irrelevant for the mechanism of the immune response or, alternatively, such structural divergence may have a functional meaning.

From an evolutionary perspective, V_H gene segments of mice and humans have been classified in three main groups or clans (Schroeder *et al.*, 1990; Tutter *et al.*, 1991, Kirkham *et al.*, 1992). These clans represent three progenitor elements whose descendants have coexisted in the vertebrate genome for 200 millions years (Anderson and Matsunaga, 1995) or more (Ota and Nei, 1994), before the divergence of humans and mice took place ~70 million years ago. Expansion and divergence from those three clans have generated the currently known 15 V_H mice families and the 7 V_H human families (Schroeder *et al.*, 1990; Kirkham *et al.*, 1992). Clans and families have preserved distinctive structural features, such as the framework 1 (FR1) and framework 3 (FR3) structures, throughout evolution. Structural preservation of these portions has been explained in terms of the essential roles they play in antibody function (Schroeder *et al.*, 1990; Kirkham *et al.*, 1992).

In contrast with the structurally conserved FR1 and FR3, it has been proposed that the hypervariable loop, being directly implied in the specific recognition of a wide variety of antigens, have been the target of strong environmental diversifying pressures in the course of evolution (Perlmutter *et al.*, 1985; Schroeder *et al.*, 1990; Kirkham *et al.*, 1992; Sims *et al.*, 1992, Litman *et al.*, 1993). However, as already mentioned, the structural repertoire of mice is family-specific (Fig. 4), which implies restrictions to the random diversification of the hypervariable loops conformations (canonical structures) and their combinations within the same V_H segment (canonical structure classes). Although less prominent, human repertoire follows this same family-specific feature (Fig. 5). Moreover, inspection of the structural repertoire of humans and mice, as classified by clans, shows that canonical structures are also clan-specific (Table 3). Therefore, preservation of the structural repertoire, even across species, strongly suggests restrictions operating to counteract the random diversification of the hypervariable loop structure.

A more detailed analysis of the evolutionary relationships of the V_H repertoire of mice and humans reveals that the largest family in mice (V_H1) belongs to clan I while the largest one in humans (V_H3) belongs to clan III (Schroeder *et al.*, 1990; Kirham *et al.*, 1992). The fact that the largest families in their respective species have developed from different ancestral elements suggests that the V_H gene segments of human and mice have followed different evolutionary pathways. Interestingly, this divergence correlates well with the difference found in the structural repertoire. That is, the V_H1 family of mice encodes canonical structure 1–2 (see Fig. 4), while the V_H3 family of humans mainly encodes class 1–3 (see Fig 5). Therefore, this correlation, jointly with the suggestion that some mechanism preserves the structural repertoire, supports the proposition that the found differences have a functional meaning.

A possibility to explain the different development of the structural repertoire of mice and humans relies on the indirect or direct interaction of classes 1–2 and 1–3 with bacterial or self-antigens named superantigens (Zouali, 1995). In humans, for example, the protein A of *Staphylococcus aureus* is highly specific to the V_H3 family (Sasso *et al.*, 1989; Sasso *et al.*, 1991). This specificity is probably due to a direct contact between the superantigen molecule and the V_H3 family-conserved FR3 region of antibodies (Sasso *et al.*, 1989, Sasso *et al.*, 1991) In structural terms, residue 71 within the FR3 segment is the major determining factor of the conformation of canonical structures 2 and 3 in H2 (Tramontano *et al.*, 1990). That implies a

Table 3. Comparison of the V_H structural repertoire between human and mice as classified by clans.

	Mice		Humans	
Clan[a]	V_κCSC	Frequency (%)	V_λCSC	Frequency (%)
I	**1–2**[b]	**94.1**	**1–2**	**50.0**
	1–?	3.9	1–?	7.1
	1–1	2.0	1–1	—
	1–3	—	1–3	42 9
II	1–1	**60.0**	1–1	13.3
	3-1	20.0	**3–1**	**60.0**
	2 1	10.0	2–1	20.0
	?–1	10 0	?–1	—
	3–5	—	3–5	6.7
III	**1–3**	**75.0**	**1–3**	**64.0**
	1–4	12 5	1–4	9.0
	1–1	6.3	1–1	14.0
	1–?	6.3	1–?	9.0
	1–6	--	1–6	4.0

[a] Clan I includes the human V_H1 and V_H5 families, and mice V_H1, V_H9 and V_H14 families. Clan II is defined by the human V_H2, V_H4 and V_H6 families, and the mice V_H2, V_H3, V_H8 and V_H12 families. Clan III consists of the human V_H3 family and the mice V_H4, V_H5, V_H7, V_H10 families (Schroeder *et al* , 1990). It should be noted that the V_H14 family of mice was described after the classification of Schroeder *et al.* (1990). However, their members are very similar to the V_H1 *family (>80%) and thus it is easy to assign them to the clan* 1

[b] The specific canonical structure classes for each clan are shown in bold

close relationship between the H2 conformation and the FR3 region which indirectly may account for differences in classes 1 2 and 1 3 A more direct interaction of classes 1 2 and 1 3 with superantigens might also be conceived. Since H2 is adjacent to FR3 in the three-dimensional structure, these regions jointly conform a continuous area exposed to solvent. Therefore, the shape of this area would change depending on the conformation of H2, which is in turn determined by position 71 in FR3. In that way, cononical structures 2 and 3 in H2 together with the FR3 structure might be recognized directly by different superantigens Since superantigens are family specific and might be important within the immune response (Zouali, 1995), they would account for the different conservation and development of the specific structural repertoires of mice and humans

A second explanation for the origin of the differences between the structural repertoire of mice and humans, its development and preservation once established, is that those genes having canonical structure classes 1-2 and 1–3 possess different regulatory roles in their respective species To support this, it is worth noting that the most frequently expressed sequence in the human repertoire is germ-line 3–23 (also called VH26, DP47, V_H30p1 and V_H182) (Stewart *et al.*, 1992; Schwartz and Stollar, 1994). The 3–23 V_H gene segment belongs to the V_H3 family and possesses canonical structure class 1–3 (Fig. 4). Several lines of evidence suggest that over-expression of this gene segment and its idiotype (Id 16/6) is associated with *important physiological roles* (Stewart *et al.*, 1992). In mice, frequent usage of the V_H gene segment H10 (VH10 in Fig. 1) has been reported (Schiff *et al.*, 1988), so making an equivalent example of the human germ-line 3–23. The H10 gene has cononical structure class 1 2 (Fig. 4) and, although being assigned to the V_H14 family, it shares more than 80% nucleotide identities with sequences belonging to the V_H1 family. This gene is used in response to different antigens (Schiff *et al.*, 1988) and, in its germ-line configuration, it is used in anti-GAT antibodies as well as in the GAT idiotypic cascade (Schiff *et al.*, 1988). That suggests a regulatory function for this gene segment within the immune response of mice, e.g., a role to play in the idiotypic network (Schiff *et al.*, 1986). Thus, the development of the V_H3 family in humans, particularly those members having canonical structure class 1–3, and the development of the V_H1 family as well as the closely related V_H14 family in mice (which encodes class 1–2) would be associated to regulatory roles these V_H families (and classes) have had in the immune response of their respective species.

Finally, a third argument is the one related to structural divergence of human and murine V_κ and V_λ germ-line genes on the one hand (Williams *et al.*, 1996; Almagro *et al.*, 1997), and the differences of human and murine repertoire of D gene segments on the other (Wu *et al.*, 1993). It has been suggested that different V_L impose restrictions to the use of some V_H gene segments or V_H families (Yurovky and Kelsoe, 1993). That indicates additional pressures acting on the divergence of V_H repertoire in humans and mice Furthermore, it has been shown that the length of H3 is significantly longer in human than in murine antibodies (Wu *et al.*, 1993), which has been related with the different lengths present in the repertoire of D gene segments (Wu *et al.*, 1993). Since a long H3 interact directly with H1 and H2 (Chothia *et al.*, 1987), this difference may also have given shape to the currently known repertoires of different human and murine V_H genes. Of course, these restrictions do not exclude any of the other two reasons, i.e., regulatory pressures and/or specific interaction with other molecules (like superantigens), which could perfectly happen to be complementary.

In summary, we have shown that the difference between the structural repertoire of V_H germ-line genes of mice and humans may have a functional meaning. Although such difference does not influence the antigen-binding site shape strongly and, thus, cannot be directly related with the initial structural restrictions operating to recognize different types of antigens, it may indeed be a reflection of species-specific regulatory and/or structural restrictions at work to balance the random diversification of the structural repertoire of V_H gene segments. Therefore, the difference here described could be very useful as a guide to choose the most human-compatible murine antibodies for human therapy

Acknowledgements—We thank I. Tomlinson for kindly providing the sequences of the 51 functional V_H gene segments of humans, to H. Ceceña and B. Levin for revision of the submitted manuscript. E. V. was supported by S-CONACyT grant no 1843. This work was partially supported by the grant from DGAPA-UNAM IN213796.

REFERENCES

Almagro, J C, Vargas-Madrazo, E., Zenteno-Cuevas, R., Hernandez-Mendiola, V. and Lara-Ochoa, F (1995) VIR A computational tool for analysis of immunoglobulin sequences *BioSystems* **35**, 25–32.

Almagro, J C., Dominguez-Martinez, V, Lara-Ochoa, F. and Vargas-Madrazo, E. (1996) Structural repertoire in human VL pseudogenes of immunoglobulins: Comparison with functional germline genes and amino acid sequences. *Immunogenetics* **43**, 92–96.

Almagro, J C, Hernandez, I., Ramirez, M. C. and Vargas-Madrazo, E (1998) Characterization of the differences between the structural repertoire of V_κ germ-line gene segments of mice and humans. *Immunogenetics* (in press)

Amzel, L. M and Poljak, R. J. (1979) Three dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* **48**, 961–997.

Anderson, A and Matsunaga, T. (1995) Evolution of immunoglobulin heavy chain variable region genes: a VH family can last for 150–200 million years or longer. *Immunogenetics* **41**, 18–28.

Barré, S., Greenberg, A. S., Flajnik, M. and Chothia, C. (1994) Structural conservation of hypervariable regions in immunoglobins evolution. *Nature Structural Biology* **1**, 915–920

Brodeur, P H and Riblet, R. (1984) The immunoglobulin heavy chain variable region (Igh-V) locus in the mouse I. One hundred Igh-V genes comprise seven families of homologous genes. *Eur. J Immunol.* **14**, 922–930.

Chothia, C and Lesk, A M (1987) Canonical structures for the hypervariable regions of imunoglobulins *J. Mol Biol* **196**, 901-917

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S, Padlan, E. A., Davies, D., Tulip, W R, Colman, P. M, Spinelli, S., Alzari, P. M. and Poljak, R J (1989) Conformations of immunoglobulin hypervariable regions *Nature* **324**, 877–883.

Chothia, C, Lesk, A M, Gherardi, E., Tomlinson, I. M., Walter, G, Marks, J. D., Llewelyn, M. B and Winter, G (1992) Structural repertoire of the human V_H segments *J Mol Biol* **227**, 799 817.

Chothia, C., Boswell, D R and Lesk, A (1988) The outline structure of the T-cell $\alpha\beta$ receptor. *EMBO J* **7**, 3745–3755

Cook, G P and Tomlinson, I. M (1995) The human immunoglobulin V_H repertoire *Immunol. Today*, **16**, 237–242

Crews, S, Griffin, J., Huang, H., Calame, K. and Hood, L (1981) A single VH gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody *Cell* **25**, 59–66

Hood, L, Gray, W R, Sanders, B G. and Dreyer, W. J (1967) *Cold Spring Harbor Symp. Quant Biol* **32**, 133

Kabat, E A and Wu, T T. (1971) Attempts to locate complementarity determining residues in the variable positions of light and heavy chains *Ann NY Acad Sci.* **190**, 382-383

Kabat, E A, Wu, T T, Perry, H M, Gottesmann, K S. and Foeller, C (1991) *Sequences of proteins of immunological interest* 5th Edn., Public Health Service N I H Washington, D C

Kirkham, P M, Mortari, F., Newton, J A and Schroeder, H. W Jr. (1992) Immunoglobulin VH clan and family identity predicts variable domain structure and may influences antigen binding. *EMBO J.* **11**, 603–609

Klein, R., Jaenichen, R and Zachau, H. G (1993) Expressed human immunoglobulin k genes and their hypermutation *Eur J. Immunol.* **23**, 3248–3271.

Kofler, R, Geley, S., Kofler, H. and Helmberg, A (1992) Mouse variable-region gene families: complexity, polymorphism and use in non-autoimmune responses *Immunol. Rev.* **128**, 5–21

Lara-Ochoa, F., Almagro, J C., Vargas-Madrazo, E. and Conrad, M. (1996) Antibody-antigen recognition a canonical structure paradigm. *J. Mol. Evol.* **43**, 678–684.

Limpanasithikul, W., Ray, S and Diamond, B (195) Cross-reactive antibodies have both protective and pathogenic potential. *J. Immunol.* **155**, 967-973

Litman, G. W., Rast, J. P., Shamblott, M. J, Haire, R. N, Hulst, M., Roess, W., Lipman, R. T., Hinds-Frey, K. R, Zilch, A. and Amemiya, C. T. (1993) Phylogenetic diversification of immunoglobulin genes and the antibody repertoire *Mol. Biol. Evol* **10**, 60–72.

Livant, D., Blatt, C. and Hood, L. (1986) One heavy chain variable region gene segment subfamily in the BALB/c mouse contains 500–1000 or more members. *Cell* **47**, 461–470.

Mainville, C. A., Sheehan, K. M., Klaman, L. D, Giorgetti, C A., Press, J. L. and Brodeur, P. H (1996) Deletional mapping of fifteen mouse VH gene families reveals a common organization for three Igh haplotypes. *J. Immunol.* **156**, 1038 1046.

Mukherjee, J., Casadevall, A. and Scharff, M. D. (1993) Molecular characterization of the humoral responses to Cryptococcus neoformans infection and glucuronoxylomannan-tetanus toxoid conjugate immunization. *J. Exp Med.* **17**, 1105–1116.

Ota, T. and Nei, M. (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol. Biol Evol.* **11**, 469–482.

Perlmutter, R. M., Berson, B, Griffin, J A. and Hood, L. (1985) Diversity in the germline antibody repertoire. Molecular evolution of the T15 VH gene family. *J. Exp. Med* **162**, 1998-2016

Poljak, R. J, Amzel, L. M., Avey, H. P., Chen, B L., Phizacherley, R. P. and Saul, F. (1973) Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8 Å resolution. *Proc Nat Acad. Sci U.S.A* **70**, 3305–3310

Poul, M-A and Lefranc, M-P. (1995) Structural correspondence between mouse and human immunoglobulin VH genes. Applications to the humanization of mouse monoclonal antibodies. *Ann. N Y. Acad. Sci.* **764**, 359–361.

Ruff-Jamison, S., Campos-Gonzalez, R. and Glenney, J. R., Jr (1991) Heavy and light chain variable region sequences and antibody properties of anti-phosphotyrosine antibodies reveal both common and distinct features. *J Biol Chem* **26**, 6607–6613.

Sasso, E.H., Silverman, G J, and Mannik, M (1989) Human IgM molecules that bind staphylococcal protein A contain VHIII H chains *J. Immunol.* **142**, 2778–2783

Sasso, E.H, Silverman, G. J., and Mannik, M. (1991) Human IgA and IgG F(ab')$_2$ that bind to staphylococcal protein A belong to the V_HIII subgroup. *J Immunol* **147**, 1877 1883

Schiff, C, Milili, M, Huc, I., Rudikoff, S and Fougereau, M (1986) Genetic basis for expression of the idiotypic network One unique Ig VH germline gene accounts for the major family of Ab1 and Ab3 (Ab1') antibodies of the GAT system *J Exp. Med* **163**, 573–587

Schiff, C., Corbet, S. and Fougereau, M (1988) The Ig germline gene repertoire: economy or wastage? *Immunol Today* **9,** 10–14.

Schroeder, H. W Jr., Hillson, J. L. and Perlmutter, R. M. (1990) Structure and evolution of mammalian VH families. *Int. Immunol.* **20,** 41–50

Schwartz, R. S. and Stollar, B. D. (1994) Heavy-chain directed B-cell maturation: continuous clonal selection beginning at the pre-B cell stage *Immunol. Today.* **15,** 27–32

Sims, M. J., Krawinkel, U. and Taussig, M. (1992) Characterization of germ-line genes of the VGAM3.8 VH family from BALB/c mice. *J. Immunol* **149,** 1642–1648.

Stewart, A. K., Huang, C., Long, A. A., Stollar, B. D. and Schwartz, R. S. (1992) VH-gene representation in autoantibodies reflects the normal human B-cell repertoire. *Immunol. Rev.* **128,** 101–122.

Tomlinson, I. A., Walter, G , Marks, J. D., Llewelyn, M. B. and Winter, G. (1992) The repertoire of human germline V_H sequences reveals about fifty groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* **227,** 776–798

Tomlinson, I. A , Cox, J. P., Gherardi, E., Lesk, A. M. and Chothia, C (1995) The structural repertoire of the human V kappa domain *EMBO. J.* **14,** 4628–4638.

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

Tormo, J., Stadler, E., Skern, T., Auer, H., Kanzler, O., Betzel, C., Blaas, D. and Fita, I. (1992) Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2. *Protein Sci* **1,** 1154–1161.

Tramontano, A., Chothia, C. and Lesk, A. M. (1990) Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J. Mol. Biol.* **215,** 175–182.

Tutter, A and Riblet, R. (1989) Conservation of an immunoglobulin variable-region gene family indicates a specific noncoding function. *Proc. Natl Acad. Sci. USA* **86,** 7460–7464.

Tutter, A., Brodeur, P , Shlomchik, M. and Riblet, R. (1991) Structure, map position, and evolution of two newly diverged mouse Ig VH gene families. *J. Immunol.* **147,** 3215–3223

Vargas-Madrazo, E , Lara-Ochoa, F and Almagro, J C (1995a) Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mech nism of immune recognition. *J Mol Biol* **254,** 487–504.

Vargas-Madrazo, E., Almagro, J C and Lara-Ochoa, F (1995b) Structural repertoire in V_H pseudogenes of immunoglobulins. comparison with human germline genes and human amino acid sequences. *J. Mol Biol.* **246,** 74 81

Weill, J-C. and Reynaud, C-A. (1996) Rearrangement/hypermutation/gene conversion: when, where and why? *Immunol Today* **17,** 92–97

Williams, S. C , Frippiat, J-P., Tomlinson, I A , Ignatovich, O., Lefranc, M-P and Winter, G. (1996) Sequence and evolution of the human germline V_λ repertoire. *J Mol Biol* **264,** 220–232.

Wu, T. T. and 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 antibody complementarity. *J. Exp. Med* **132,** 211–250.

Wu, T. T., Johnson, G. and Kabat, E. A (1993) Length distribution of CDRH3 in antibodies. *Proteins* **16,** 1- 7

Yurovky, V. and Kelsoe, G (1993) Pairing of V_H gene families with the λ1 light chain: evidence for a non-stochastic association *Eur. J. Immunol* **23,** 1975 1979.

Zouali, M. (1995) B-cell superantigens. implications for selection of the human antibody repertoire *Immunol Today* **16,** 399–405

Immunogenetics (1998) 47: 355–363

ORIGINAL PAPER

Juan Carlos Almagro · Ismael Hernández
Maria del Carmen Ramírez · Enrique Vargas-Madrazo

Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications

Received. 1 June 1997 / Revised: 6 October 1997

Abstract Although human and mouse antibodies are similar when one considers their diversification strategies, they differ in the extent to which kappa and lambda light chains are present in their respective variable light chain repertoires. While the *Igk-V* germline genes are preponderant in mice (95% or more), they comprise only 60% in humans. This may account for differences in the structural repertoire encoded in the *Igk-V* germline genes of these species. However, this subject has not been properly investigated, partially because a systematic structural characterization of the mouse *Igk-V* germline genes has not been undertaken. In the present study we compiled all available information on mouse *Igk-V* germline genes to characterize their structural repertoire. As expected, comparison with the structural repertoire of human *Igk-V* germline genes indicates differences. The most interesting is that the mouse *Igk-V* germline gene repertoire is more diverse in structural terms than its human counterpart: the mouse encodes seven canonical structure classes (combination of canonical structures in L1 and L3). In contrast, the human encodes only four. Analysis of the evolutionary relationships of human and mouse *Igk-V* germline genes led us to propose that the difference reflects a strategy of mice to compensate for the small lambda chain contribution to the repertoire of their variable light chains.

Key words Mice · Immunoglobulins · Canonical structures *Igk-V* · Evolution

Introduction

In contrast to other species such as shark, chicken, rabbit, or sheep, the human and mice species generate their antibody diversity in a similar fashion (Weill and Reynaud 1996). Prior to the antigenic challenge, these species produce a primary repertoire through the recombination of multiple germline genes (Berek and Milstein 1988; Neuberger and Milstein 1995, Tonegawa 1983). The variable kappa or lambda light chain is produced by the recombination of the *Igl-V* or *Igk-V* and *Igl-J* or *Igk-J* germline genes, respectively (Tonegawa 1983). The variable heavy chain is caused by a recombination of the *Igh-V* germline genes with two additional germline genes, *Igh-D* and *Igh-J* (Tonegawa 1983). Antibodies thus generated should be capable of interacting with any antigen at least with a low or medium affinity in the primary immune response (Berek and Milstein 1988; Neuberger and Milstein 1995). Upon selection by the antigen, the chosen human or mouse antibodies improve their affinity mainly by somatic hypermutation during a secondary or tertiary immune response (Berek and Milstein 1988; Neuberger and Milstein 1995; Weill and Reynaud 1996).

Despite the similarity that renders humans and mice "equivalent" in their diversification strategies for antibodies, these species possess different proportions of kappa and lambda light chains in their germline genes. In human, roughly 60% of the variable light chain repertoire is kappa [40 functional *Igk-V* germline genes (Klein et al. 1993; Tomlinson et al. 1995) vs 30 functional *Igl-V* germline genes (Williams et al. 1996)]. In mice, kappa preponderates, being as much as 95% or more [fewer than 160 functional *Igk-V* germline genes (Zocher et al. 1995) vs three functional *Igl-V* germline genes (Dildrop et al. 1987, Selsing et al. 1989)] This difference may account for a divergence in the primary repertoire of human and mouse antibodies. However, a systematic analysis has not been made.

A way to characterize that difference is through a comparison of the repertoire of antigen binding site structures implicit in the variable light chain germline genes of mice and humans Even though considerable sequence variability exists at the antigen binding site (Kabat and Wu 1971, Wu and Kabat 1970), it has been shown that the hypervariable loops do not adopt a large and unpredictable

J. C. Almagro (✉) · I. Hernández · M. del Carmen Ramírez
Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, APDO POSTAL 510-3, Cuernavaca, Morelos 62250, México

E. Vargas-Madrazo
Instituto de Investigaciones Biologicas, Universidad Veracruzana Xalapa, Mexico

Ig-fold[a]		B B B T B B B B111111111111B BIBI T I IBB 222 TT B B B B T T B BIBI 33333				
Position[b]		1 10 20 30 40 50 60 70 80 90				
Family[c]	Name[d]	. .' '. .'. ' . ' ..'abcdef .'. .'. ' ' ..'. ' ...' .. '....!....'....'....' . .'	Rearrenged gene[e]		Status[f]	
21	21B	NIVLTQSPASLAVSLGQRATISCRASESVD--SYGNSFMHWYQQKPGQPPKLLIYLASNLESGVPARFSGSGSRTDFTLTIDPVEADDAATYYCQQNNEDP	N19-8 scFv	(3)	F	1
21	21C/45 21.1	DIVLTQSPASLAVSLGQRATISCRASESVD--SYGNSFMHWYQQKPGQPPKLLIYRASNLESGIPARFSGSGSRTDFTLTINPVEADDVATYYCQQSNEDP	PC3741	(0)	F	2
21	18kb	DIVLTQSPASLAVSLGQRATIFCRASQSVD--YNGISYMHWFQQKPGQPPKLLIYAASNLESGIPARFSGSGSGTDFTLNIHPVEEEDAATYYCQQSIEDP	48 2 1	(7)	F	3
21	21E/21B	DIVLTQSPASLAVSLGQRATISCRASKSVS--TSGYSYMHWYQQKPGQPPKLLIYLASNLESGVPARFSGSGSGTDFTLNIHPVEEEDAATYYCQHSRELP	98QQ	(2)	F	4
21	1.6kb	DIVLTQSPASLAVSLGQRATISCRASQSVS--TSSYSYMHWYQQKPGQPPKLLIKYASNLESGVPARFSGSGSGTDFTLNIHPVEEEDTATYYCQHSWEIP	RF-4 PAN	(1)	F	5
21	21G	DIVLTQSPASLAVSLGQRATISCRASESVE--YYGTSLMQWYQQKPGQPPKLLIYAASNVESGVPARFSGSGSGTDFSLNIHPVEEDDIAMYFCQQSRKVP	S4 5A	(1)	F	6
21	21A	DIVLTQSPASLAVSLGQRATISCRASESVD--NYGISFMNWFQQKPGQPPKLLIYAASNQGSGVPARFSGSGSGTDFSLNIHPMEEDDTAMYFCQQSKEVP	S3 12D	(1)	F	7
21	P3-X63-Ag8/HNK20	DIVLTQSPASLAVSLGQRATISYRASKSVS--TSGYSYMHWNQQKPGQPPRLLIYLVSNLESGVPARFSGSGSGTDFTLNIHPVEEEDAATYYCQHIRE			PS	
21	CEA 66-E3	DIVLTQSPASLAVSLGQRATISYRASKSVS--TSGYSYMHWNQQKPGQPPRLLIYLVSNLESGVPARFSGSGSGTDFTLNIHPVEEEDAATYYCQHI*GAY			PS	
23	L7	DILLTQSPAILSVSPGERVSFSCRASQSIG------TSIHWYQQRTNGSPRLLIKYASESISGIPSRFSGSGSGTDFTLSINSVESEDIADYYCQQSNSWP	A26	(0)	F	8
23	A23A41	DVLLTQSPAILSVSPGERVSFSCRASQSIG------TSIHWYQQRTNGPPRLLIKYASESISGIPSRFSGSGSGTDFTLSISSVESEDIADYYCQQTNSWP			PS	
23	MMIG21	DILLTQSPAILSVSPGERVSFSCRASQSIG------TSIHWYQQRTNGSPRLLIKNASESISGIPSRFSGSGSGTDFTPSINSVESEDIAEYYCQQSYRWP			PS	
23	B1P8-7-2	DIVLTQSPATLSVTPGDRVSLSCRASQSIS------NYLHWYQQKSHESPRLLIKYASQSISGIPSRFSGSGSGTDFTLSINSVETEDFGMYFCQQSNSWX			PS	
23	MRL-RF33BL/MRL-n-RF33	DIVLTQSPATLSVTPGDSVSLSCRASQSIS------NNLHLYQ*KSHESPRLLIKYVFQSISEIPSKFSGSGSGTDFTLSINSVETEDFGMYFCQQSNSWP			PS	
4/5	H3/Ox1	QIVLTQSPAIMSASPGEKVTMTCSASSSV-------SYMHWYQQKSGTSPKRWIYDTSKLASGVPARFSGSGSGTSYSLTISSMEAEDAATYYCQQWSSNP	11G3	(0)	F	9
4/5	R9	QIVLTQSPAIMSASPGEKVTMTCSASSSI-------SYMHWYQQKPGTSPKRWIYDTSKLASGVPARFSGSGSGTSYSLTISSMEAEDAATYYCHQRSSYP	58.2C 10.3	(2)	F	10
4/5	H13	QIVLTQSPALMSASPGEKVTMTCSASSSV-------SYMYWYQQKPRSSPKPWIYLTSNLASGVPARFSGSGSGTSYSLTISSMEAEDAATYYCQQWSSNP	AN01	(1)	F	11
4/5	H9	QILLTQSPAIMSASPGEKVTMTCSASSSV-------SYMHWYQQKPGSSPKPWIYDTSNLASGFPARFSGSGSGTSYSLIISSMEAEDAATYYCHQRSSYP			NF	12
4/5	H4	QIVLTQSPAIMSASPGEKVTISCSASSSV-------SYMYWYQQKPGSSPKPWIYRTSNLASGVPARFSGSGSGTSYSLTISSMEAEDAATYYCQQYHSYP	163.42	(1)	F	13
4/5	H8	QIVLTQSPAILSASPGEKVTMTCSASSSV-------SYM*WFQQKPGSSPKLWIYSISNLASGVPARFSGSGSGTSYSLTISSVKAEDAATYYCQQKSSSP			PS	
4/5	H6/X24	EIVLTQSPAITAASLGQKVTITCSASSSV-------SYMHWYQQKSGTSPKPWIYEISKLASGVPARFSGSGSGTSYSLTISSMEAEDAAIYYCQQWNYPL	TEPC191	(0)	F	14
4/5	R13	ENVLTQSPAIMSASLGEKVTMSCRASSSV-------NYMYWYQQKSDASPKLWIYYTSNLAPGVPARFSGSGSGNSYSLTISSMEGEDAATYYCQQFTSSP	AN09	(3)	F	15
4/5	R2	EILLTQSPAIIAASPGEKVTITCSASSSV-------SYMNWYQQKPGSSPKIWIYGISNLASGVPARFSGSGSGTSFSFTINSMEAEDVATYYCQQRSSYP			NF	16
4/5	H2	GIVLTQSPTTMTAFPGENVTITCSASSSI-------NYIHWYQQKSGNTPKQ*IYKTSDLPSGVPTLPSGSGSGTSYSLTISSVEAEDAATYYCQQWSGYP			PS	
4/5	H1	QIVLTQSPAIMSASPGEKVTMTCSARSSVSS-----SYLYWYQQKPGSSPKLWIYSTSNLASGVPARFSGSGSGTSYSLTISSMEAEDAATFYCQQYSQYP	MRL-22	(4)	F	17
4/5	R1/s107b	ENVLTQSPAIMAASLGQKVTMTCSASSSVSS-----SYLHWYQQKSGASPKPLIPRTSNLASGVPARFSGSGSGTSYSLTISSVEAEDDATYYCQQWSGYP	T10-938	(15)	F	18
4/5	T3B	TQSPAIMAASLGEKVTMTCSASSSVSS-----SYLHWYQQKSGTSPKPWIYGTSNLASGVPVRFSGSGSGTSYSLTISSMEAEDAATYYCQQWSSYP			PS	
4/5	L8	ENVLTQSPAIMAASLGEKVTMTCSASSSVSS-----SYLHWYQQKSGTSPKLWIYGTSNLASGVPARFSGSGAGISYSLTISSMEAENDATYYCQQWSGYP			NF	19
4/5	R11	ENVLTQSPAIMAASPGEKVTMTCSASSSVSS-----SNLHWYQQKSGTSTKFWIYRTSNLASEVPAPFSGSGSGTSYSLTISSVEAEDAATYYCQQWSGYP	18-2-3	(12)	F	20
12/13	k2/MMIG27	DIQMTQSPASLSASVGETVTITCRASGNIH------NYLAWYQQKQGKSPQLLVYNAKTLADGVPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTP	SO2	(2)	F	21
12/13	k3	DIQMTQSPASLSVSVGETVTITCRASENIY------SNLAWLFSRNRENPPSLVYAATNLADGVPSRFSGSGSGTQYSLKINSQQPEDFGSYYCQHFWSAP			NF	22
11	V11	DVQMTQSPSSLSASLGERVSLTCQASQSIN------NFLKWFQQTLGKTARLLIYGANKLEDGVPSRFSGTGYGTDFTFTISSQEEEDVSTYFCLQHRYLP			PS	
9A	vk41/mpoc41	DIQMTQSPSSLSASLGERVSLTCRASQDIG------SSLNWLQQEPDGTIKRLIYATSSLDSGVPKRFSGSRSGSDYSLTISSLESEDFVDYYCLQYASSP	202.105	(10)	F	23
9A	m173b	DIQMTQSPSSLSASLGERVSLTCRASQDIH------GYLNLFQQKPGETIKHLIYETSNLDSGVPKRFSGSRSGSDYSLIIGSLESEDFADYYCLQYASSP	EXW-16	(24)	F	24
9B	L6	DIKMTQSPSSMYASLGERVTITCKASQDIN------SYLSWFQQKPGKSPKTLIYRANRLVDGVPSRFSGSGSGQDYSLTISSLEYEDMGIYYCLQYDEFP	E6	(3)	F	25
9B	9B 8	DIQTTQSPSSMSVSLGETVSITCRASQGIS------SYVSWLQQKPGKSPKTLISYATNLEDGITSWFSSSGSGADYSLTISSLESEDCKIYYCVQYVQLP			PS	
10	Ars/AJ1/Id(CR)/B1P8-7b	DIQMTQTTSSLSASLGDRVTISCRASQDIS------NYLNWYQQKPDGTVKLLIYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEQEDIATYFCQQGNTLP	Sulf-1	(0)	F	26
10	PERU1	DIQMTQTTSSLSASLGDRVTISCRASQDIS------NYLNWYQQKPDGTVKLLIYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEQEDIATYFCQQGSTLP	hVF65-107	(1)	F	27
10	AKR1	DIQMTQTTSSLSASLGDRVTISCRASQDIS------NYLNWYQQKPDGTVKLLIYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEQEDIATYFCQQDSKHP			NF	28
10	AKR2	DIQMTQTTSSLSASLGDRVTISCRASQDIS------NYLNWYQQKPDGTVKLLIYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEPEDIATYYCQQYSKLP	44 1	(4)	F	29
10	PERU2	DIQMTQTTSSLSASLGDRVTISCSASQGIS------NYLNWYQQKPDGTVKLLIYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEPEDIATYYCQQYSNLP			NF	30
10	AJ2/B1P8-7-3	DIQMTQTTSSLSASLGDRVTISCSASQGIS------NYLNWYQQKPDGTVKLLIYYTSSLHSGVPSRFSGSGSGTDYSLTISNLEPEDIATYYCQQYSKLP	L2-10CL	(0)	F	31
10	V10	DIQMTQTTSSLSASPGDRVTISCRTSQDIS------NFLYWFQQKSDGTVKLLIYYTSR*HSGVPSRFSGSGSGTDYSFTINNLS*EDVATYS*QQGILTL			PS	
24/25	167/24	DIVITQDELSNPVTSGESVSISCRSSKSLLYK-DGKTYLNWFLQRPGQSPQLLIYLMSTRASGVSDRFSGSGSGTDFTLEISRVKAEDVGVYYCQQLVEYP	C57BL 2857	(0)	F	32
24/25	24A	DIVMTQAAFSNPVTLGTSASISCRSSKSLLHS-SGNTYLYWFLQKPGQSPQLLIYYISNLASGVPDRFSGSGSGTDFTLRISRVEAEDVGVYYCMQGLEYP	H35-C6	(11)	F	33
24/25	24B	DIVMTQAAFSNPVTLGTSASISCRSSKSLLHS-NGITYLYWYLQKPGQSPQLLIYQMSNLASGVPDRFSSSGSGTDFTLRISRVEAEDVGVYYCAQNLELP	MS	(0)	F	34
1	V-1A/KS 1/KS 1	DVVMTQTPLSLPVSLGDQASISCRSSQSLVHS-NGNTYLHWYLQKPGQSPKLLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVP	K146-24B3	(0)	F	35
1	V-IB	DVVMTQTPLSLPVSLGDQASISCRSSQSLVHS-NGNTYLYWYLQKPGQSPKLLIYRVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCFQGTHVP	25.12	(1)	F	36
1	V-IC	DVVMTQTPLSLPVSLGDQASISCRSSQSIVHS-NGNTYLEWYLQKPGQSPKLLIYKVSNRLSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVP			NF	37
1	V-1C/V1A5/K1A5	DVLMTQTPLSLPVSLGDQASISCRSSQSIVHS-NGNTYLEWYLQKPGQSPKLLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVP	v16-19	(0)	F	38
1	K18 1	DAVMTQTPLSLPVSLGDQASISCRSSQSLENS-NGNTYLNWYLQKPGQSPQLLIYRVSNRFSGVLDRFSGSGSGTDFTLKISRVEAEDLGVYFCLQVTHVP	JV3	(0)	F	39
1	V1F	DVLLTQTPLFLPVSLGDQASISCSSSQSLVHS-NGNYYLEWHLQKSGQSLQLLIYEVSKRHSGVPDRFSGSGSGTDFTLKISRVEPEDLGVYYCFQGTHLP			PS	
1	1E	TSSKSLVHS-NGNSYLDWHLQKPGQSLQLLIYSVSKRNSGVPDRFSGSGSGTDFTLKISRVEPEDLGVYYCFQGTHLP			PS	
2	70/2	DVVMTQXLHSLSVTIGQPASISCKSSQSLLYS-NGKTYLNWLLQRPVQPPKRLIYLVSKLYSGVPDRFSGSGSGTDFTLKISRVXPEDLGVY*CMQDTHFP			PS	
2	70/3	DVVMTQTPLTLSVTIGQPASISCKSSQSLLDS-DGKTYLNWLLQRPGQSPKRLIYLVSKLDSGVPDRFTGSGSGTDFTLKISRVEAEDLGVYYCWQGTHFP	BALB/C1210 7	(0)	F	40
2	70/1	DVVMTQTPLTLSVTIGQPASISCKSSQSLLYS-NGKTYLNWLLQRPGQSPKRLIYLVSKLDSGVPDRFTGSGSGTDFTLKISRVEAEDLGVYYCVQGTHFP	4G11	(0)	F	41
8	ABPC48	DIVMTQSPSSLAMSVGQKVTMSCKSSQSLLSSSNQKNYLAWYQQKPGQSPKLLVYFASTRESGXPDRFIGSGSGTDFTLTISSVQAEDLALYYCQQHYSTP			PS	
8	GLvk50	DIVMTQSPSSLSVSAGDKVTMSCKSSQSLLNSRNQKNYLAWYQQKPWQPPKLLIYGASTRESGVPDRFTGSGSGTDFTLTISSVQAEDLAVYYCQNDYSYP	D20	(0)	F	42
19/28	V-Ser[b]	SIVMTQTPKFLLVSAGERVTITCKASQSVS------NDVAWYQQKPGQSPKLLIYYASNRYTGVPDRFTGSGYGTDFTFTISTVQAEDLAVYFCQQDYSSP	17F12	(7)	F	43
19/28	V-Ser[a]	SIVMTQTPKFLPVSAGDRVTMTCKASQSVG------NNVAWYQQKPGQSPKLLIYYASNRYTGVPDRFTGSGSGTDFTFTISSVQVEDLAVYFCQQHYSSP	A34	(22)	F	44
19/28	SK/CamRK	SIVMTQTPKFLPVTAEDRVTITCKASQSVS------NEVAWYQQKPGQSPKLLIYYASNRYTGVPDRFTGSGSGTDFTFTISSVQVEDLAVYFCQQHYSSP	AN12	(18)	F	45
19/28	PERA/Ei	SIVMTQSPKSLPVSAGDRVTMTCKASQSVS------NDVAWYQQKPGQSPKLLIYYASNRYTGVPERFTGSGSGTDFTFTISGVQAEDLAVYFCQQHYTTP	RF49B	(18)	F	46
32	MUSIGKA3G	DIQMNQSPSSLSASLGDTITITCHASQKIN------VWLSWYQQKKGNIPKLLIYRTSNLHTGVPSRFSGSGSGTGFTLTISSLQPEDIATYYCQQGQNYP			PS	
33/34	Vk34A	DIQMTQSSSSLSVSLGDRVTITCKASEHIN------SWLAWYQQKPGNAPRLLISGATSLETGVPSRFSGSASGKDYTLSITSLQTEDVAT	165 14	(24)	F	47
33/34	Vk34B	DIQMTQSSSYLSVSLGGRVTITCKASEHIN------SWLAWYQQKPGNAPRLLISGATSLETGVPSRFSGSGSGKDYTLSITSLQTEDVAT	T10-421	(6)	F	48
33/34	Vk34C	DIQMTQSSSSFSVSLGDRVTITCKASEDIY------NRLAWYQQKPGNAPRLLISGATSLETGVPSRFSGSGSGKDYTLSITSLQTEDVAT	C8.5	(43)	F	49
20	294A9	EITVTQSPASLSVATGEKVTIRRI--TDID------DEMH*YQQKPGEPPKLLISEGNTLHPGVPSQFSSSGYGTDFPFTIENTLSEDVADYYCLQSGNMP			PS	

Fig. 1 Multiple amino acid sequence alignment of mouse *Igk-V* germline genes. [a]Positions primarily responsible for the variable immunoglobulin fold (*V-Ig*-fold) conserved features (Chothia et al. 1988) and hypervariable loop definition (Chothia and Lesk 1987). *B* Residues buried in the protein; *T* residues in turns; *I* inter-domain residues. 1 L1; 2. L2, and 3· L3 [b]Residue numbering as in Chothia and Lesk (1987) [c]*Igk-V* gene family [d]Name, clone, or sequence access number in GenBank, or name of the sequence in the literature. [e]Name in the Kabat's Database of the closest *Igk-V* rearranged gene and number of amino acid differences between this and the germline gene. [f]Sequence status: *F* sequences with rearranged counterpart (functional); *NF* non-functional sequence because it has no rearranged counterpart or possesses structural defects (SD); *PS* pseudogene. *Numbers* on the *right* stand for the code of each sequence in Fig 2 The multiple sequences alignment and all the calculations presented therein were done using the VIR package (Almagro et al 1995)

number of conformations: they possess one of a small set of main-chain conformations or canonical structures (Chothia and Lesk 1987; Chothia et al. 1989; Martin and Thornton 1996). On the basis of this, it has recently been reported that of the total number of possible combinations of these canonical structures (denoted canonical structure classes) (Almagro et al. 1996; Chothia et al. 1992; Lara-Ochoa et al. 1996; Tomlinson et al. 1995; Vargas-Madrazo et al. 1995) only a few options effectively exist. The existence of canonical structures and canonical structure classes implies restrictions to a free diversification of hypervariable loops and their combination within the same gene. Therefore, if there are significant differences between the primary repertoires of human and mouse antibodies, they might

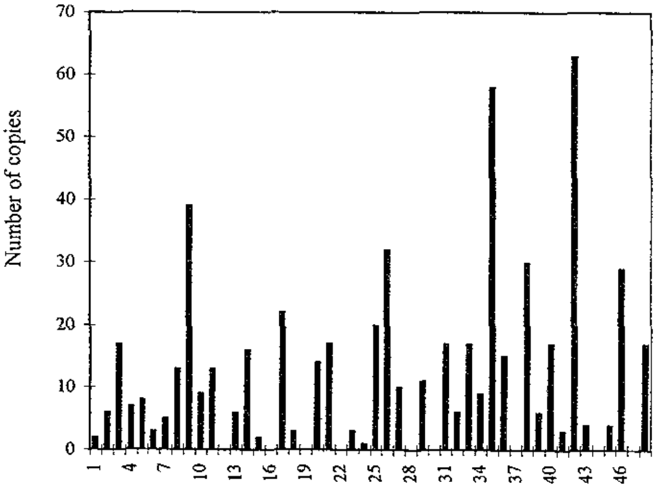


Fig. 2 Usage of *Igk-V* germline genes

become evident when their corresponding structural repertoires are analyzed.

Since the variable light chain repertoire in mice is kappa dominated, our analysis is mainly concerned with mouse *Igk-V* germline genes. The structural repertoire of human *Igk-V* germline genes has been described in detail by Tomlinson and co-workers (1995) but not the mouse. Thus in the first part of this paper we made a systematic characterization of the structural repertoire of mouse *Igk-V* germline genes so that in the second part we could make a comparison with the structural repertoire of human *Igk-V* germline genes. Finally, results are discussed in the light of the evolutionary relationships of human and mouse *Igk-V* genes.

Materials and methods

Mouse Igk-V germline genes

In order to estimate the *Igk-V* repertoire in the germlines of mice, we compiled all of the *Mus musculus Igk-V* genes reported as germline genes or pseudogene sequences in GenBank and LIGM, as well as in the literature published up to December 1996. We found a total of 97 *Igk-V* genes and immediately discarded eight of them because they were fragments of sequences (see web site http.//www.ibt.unam mx/nalmagro for a full description of the sequences).

Having collected 89 useful *Igk-V* genes, those fully identical at the nucleotide level in the coding region were considered to be the same. Similarly, sequences with only one or two nucleotide differences (99.6% and 99.2% identities, respectively) resulting in silent mutations (100% identical at amino acid level) were considered to be the same This was so decided because they might be alleles in different individuals or in different strains of the mouse.

Sequences in which the nucleotide difference resulted in replacements (different amino acid sequences) were considered different genes. In that way we finally gathered 67 sequences as representative of the mouse *Igk-V* locus

Table 1 *Igk-V* classification and germline gene repertoire

Igk-V family complexity		
Igk-V gene family[a]	Number of *Igk-V* germline genes (estimated)	Number of *Igk-V* germline genes (found)
21	6–13	9
23	2– 4	5
4/5	25–50	15
12–13	2– 8	2
11	4– 6	1
9A	4– 9	2
9B	2	2
10	2– 3	7
24–25	6	3
1	4– 6	7
2	1– 6	3
8	5–16	2
19–28	4– 6	4
38C	–	–
RF	0– 1	–
22	1– 2	–
20[b]	5– 7	1
32[c]	4– 8	1
33/34[d]	1– 3	3
Total	79–156	67

[a] *Igk-V* gene family nomenclature according to Strohal and co-workers (1989)

[b] Gene family described by Shefner and co-workers (1990)

[c] Gene family described by D'Hoostelaere and Klinman (1990)

[d] Two groups (D'Hoostelaere and Klinman 1990, Valiante and Caton 1990) have independently described this family and termed it *Igk-V33* and *Igk-V34*, respectively. To avoid confusion, it has been renamed *Igk-V33/34* (Kofler and Helmberg 1991)

Classification of the known mouse Igk-V germline genes in Igk-V families

On the basis of nucleotide comparisons, mouse *Igk-V* genes have been classified into 19 *Igk-V* families: 16 defined by Strohal and co-workers (1989) and three more defined by Valiante and Caton (1990), Shefner and co-workers (1990), and D'Hoostelaere and Klinman (1990). Therefore, in order to cluster the 67 *Igk-V* genes we had found into the 19 established *Igk-V* families, we followed the criteria established by the aforementioned authors. The resulting alignment of sequences, organized by families, is given in Fig. 1 and can be retrieved in a computer-ready format from web site: http:\\www.ibt unam mx/nalmagro.

The functional Igk-V germline genes of mice and their structural repertoire

Of the 67 *Igk-V* genes depicted in Fig. 1. 18 have been reported as pseudogenes in databases or in the literature (see status column of Fig. 1). This led us to assume they had serious genetic defects and were not taken into account to determine the functional repertoire of mouse *Igk-V* germline genes.

The remaining 49 *Igk-V* genes reported as germline and potentially functional were examined to observe their expression *in vivo* This was done by assigning the amino acid sequence of each of them to their closest rearranged functional *Igk-V* sequence in a database of 574 Igk-V amino acid sequences compiled from the Kabat's Database on-line service (Kabat et al 1991, web site. http://immuno bme nwu edu) We chose those *Igk-V* rearranged sequences having a reported specificity to avoid non-productive rearrangements, while guaranteeing the assignment of functional *Igk-V* genes only Since those segments not expressed *in vivo* might have minor genetic or structural defects hindering the formation of a stable three-dimensional V domain, they were considered to be non-functional (Klein et al 1993, Tomlinson et

Fig. 3 Structural repertoire of the functional *Igk-V* germline gene segments of mice. V_κCSC Canonical structures classes of light chain *Igk-V* genes. *?* A hypervarible loop which does not fit the canonical structure pattern. Residues responsible for mismatch are *underlined*. *U* Unknown

			L1			L3	
		,	B1111111111111B	B		333333	V_kCSC
Position[b]		2	25 30 33	71	90	95	
Family[c]	Name[d]	.	!....!abcdef...	.	!	....!	
21	21B	I	ASESVD--SYGNSFM	F	Q	NNEDP	5-1
21	21C/45.21.1	I	ASESVD--SYGNSFM	F	Q	SNEDP	5-1
21	18kb	I	ASQSVD--YNGISYM	F	Q	SIEDP	5-1
21	21E/21E5	I	ASKSVS--TSGYSYM	F	H	SRELP	5-1
21	1.6kb	I	ASQSVS--TSSYSYM	F	H	SWEIP	5-1
21	21G	I	ASESVE--YYGTSLM	F	Q	SRKVP	5-1
21	21A	I	ASESVD--NYGISFM	F	Q	SKEVP	5-1
23	L7	I	ASQSIG------TSI	F	Q	SNSWP	2-1
4/5	H3(Ox1)	I	ASSSV-------SYM	Y	Q	WSSNP	1-1
4/5	R9	I	ASSSI-------SYM	Y	Q	RSSYP	1-1
4/5	H13	I	ASSSV-------SYM	Y	Q	WSSNP	1-1
4/5	H4	I	ASSSV-------SYM	Y	Q	YHSYP	1-1
4/5	X24/H6	I	ASSSV-------SYM	Y	Q	WNYPL	1-2
4/5	R13	N	ASSSV-------NYM	Y	Q	FTSSP	?-1
4/5	H1	I	ARSSVSS-----SYL	Y	Q	YSQYP	?-1
4/5	R1/s107b	N	ASSSVSS-----SYL	Y	Q	WSGYP	6-1
4/5	R11	N	ASSSVSS-----SNL	Y	Q	WSGYP	6-1
12/13	k2/MMIG27	I	ASGNIH------NYL	Y	H	FWSTP	2-1
9A	vk41/mopc41	I	ASQDIG------SSL	Y	Q	YASSP	2-1
9B	L6	I	ASQDIN------SYL	Y	Q	YDEFP	2-1
10	Ars/AJ1/Id(CR)/B1P8-7b	I	ASQDIS------NYL	Y	Q	GNTLP	2-1
10	PERU1	I	ASQDIS------NYL	Y	Q	GSTLP	2-1
10	AKR2	I	ASQDIS------NYL	Y	Q	YSKLP	2-1
10	AJ2/B1P8-7-3	I	ASQGIS------NYL	Y	Q	YSKLP	2-1
24/25	167/24	I	SSKSLLYK-DGKTYL	F	Q	LVEYP	4-1
24/25	Vk24	I	SSKSLLHS-NGITYL	F	Q	MLERP	4-1
24/25	24B	I	SSKSLLHS-NGITYL	F	Q	NLELP	4-1
1	V-1A/K5.1/K5.1	V	SSQSLVHS-NGNTYL	F	Q	STHVP	4-1
1	V-IB	V	SSQSLVHS-NGNTYL	F	Q	GTHVP	4-1
1	V-1C/V1A5/K1A5	V	SSQSIVHS-NGNTYL	F	Q	GSHVP	4-1
1	18.1/K18,1	A	SSQSLENS-NGNTYL	F	Q	VTHVP	?-1
2	70/3	V	SSQSLLDS-DGKTYL	F	Q	GTHFP	4-1
2	70/1	V	SSQSLLYS-NGKTYL	F	Q	GTHFP	4-1
8	GLvk50	I	SSQSLLNSRNQKNYL	F	N	DYSYP	3-1
19/28	V-Ser[c]	I	ASQSVS------NDV	F	Q	DYSSP	2-1
19/28	SK/CamRK	I	ASQSVS------NEV	F	Q	HYSSP	2-1
19/28	PERA/Ei	I	ASQSVS------NDV	F	Q	HYTTP	2-1
33/34	Vk34B	I	ASEHIN------SWL	Y			2-U

al. 1995). The database with the 574 Igk-V amino acid sequences is available from the authors on request

Characterization of the canonical structures in L1 and L3

The patterns of residues determining the different canonical structures for L1 and L3 have been described in detail (Chothia and Lesk 1987, Chothia et al 1989; Tomlinson et al. 1995), and have recently been reviewed by Martin and Thornton (1996). This information is summarized at web site. Antibody Structure-function (http: //www.biochem - ucl.ac.uk/~martin/antibodies html:Chothia.dat.auto). Using these patterns, we analyzed the functional *Igk-V* germline genes of mice. It must be noted that L2 was not considered in the analysis because it has only one canonical conformation and does not influence the structural variability in antibodies.

Results

Known mouse Igk-V germline genes

Although the exact number of *Igk-V* germline genes in the mouse genome is currently unknown, several estimations exist (Cory et al. 1981; Kofler et al. 1992; Zeelon et al. 1981; Zocher et al. 1995). Early proposals ranged from 90–320 (Cory et al. 1981) up to 2000 (Zeelon et al. 1981). However, on the basis of restriction fragment length polymorphism (RFLP) criteria, as well as on current established knowledge about *Igk-V* germline gene sequences and expressed *Igk-V* sequences, it has been estimated that the entire *Igk-V* germline repertoire may not much exceed 160 (Kofler et al. 1992). Recently, cloning experiments have facilitated the proposal that the final number of gènes in the mouse *Igk-V* locus is notably smaller than 160 (Zocher et al. 1995). In agreement with this, we found 89 *Igk-V* germline genes, of which 67 turned out to be unique (Fig. 1).

In Table 1 a comparison between the established number of *Igk-V* genes within the individual *Igk-V* families and those we found is shown. Such a comparison indicates that nine *Igk-V* gene families (*21*, *23*, *12–13*, *9B*, *10*, *1*, *2*, *19–28*, and *33–34*) are well represented in our compilation. However, in seven *Igk-V* families (*4/5*, *11*, *9A*, *24–25*, *8*, *32*, *20*) we found fewer *Igk-V* genes than estimated (see Table 1). For the *38C*, *RF*, and *22* gene families, no *Igk-V*

Fig. 4 Structural repertoire of the functional human *Igk-V* germline gene segments. V_κCSC, Canonical structures classes of light chain *Igk-V* genes. ?, A hypervarible loop which does not fit the canonical structure pattern. Residues responsible for mismatch are *underlined*

			L1			L3	
			B1111111111111B	B		33333	V_kCSC
Position[b]		2	25 30 33	71	90	95	
Family[c]	Name[d]	.	'....!abcdef...	.	!	....'	
I	012/DPK9	I	ASQSISS------YL	F	Q	SYSTP	2-1
I	02/DPK9	I	ASQSISS------YL	F	Q	SYSTP	2-1
I	018/DPK1	I	ASQDISN------YL	F	Q	YDNLP	2-1
I	08/DPK1	I	ASQDISN------YL	F	Q	YDNLP	2-1
I	A20/DPK4	I	ASQGISN------YL	F	K	YNSAP	2-?
I	A30	I	ASQGIRN------DL	F	Q	HNSYP	2-1
I	L14/DKP2	I	ARQGISN------YL	F	Q	HNSYP	2-1
I	L1	I	ASQGISN------YL	F	Q	YNSYP	2-1
I	L15/DPK7	I	ARQGISS------WL	F	Q	YNSYP	2-1
I	L4	I	ASQGISS------AL	F	Q	FNSYP	2-1
I	L18	I	ASQGISS------AL	F	Q	FNSYP	2-1
I	L5/DPK5	I	ASQGISS------WL	F	Q	ANSFP	2-1
I	L19/DPK6	I	ASQGISS------WL	F	Q	ANSFP	2-1
I	L8/DPK8	I	ASQGISS------YL	F	Q	LNSYP	2-1
I	L23	I	ASQGISS------YL	Y	Q	YYSTP	2-1
I	L9	I	ASQGISS------YL	F	Q	YYSYP	2-1
I	L24/DPK10	I	MSQGISS------YL	F	Q	YYSFP	?-1
I	L11/DPK3	I	ASQGIRN------DL	F	Q	DYNYP	2-1
I	L12	I	ASQSISS------WL	F	Q	YNSYS	2-?
II	011/DPK13	I	SSQSLLDSDDGNTYL	F	Q	RIEFP	3-1
II	01/DPK13	I	SSQSLLDSDDGNTYL	F	Q	RIEFP	3-1
II	A17/DPK18	V	SSQSLVYS-DGNTYL	F	Q	GTHWP	4-1
II	A1/DPK19	V	SSQSLVYS-DGNTYL	F	Q	GTHWP	4-1
II	A18/DPK28	I	SSQSLLHS-DGVTYL	F	Q	GTHLP	4-1
II	A2/DPK12	I	SSQSLLHS-DGKTYL	F	Q	SIQLP	4-1
II	A19/DPK15	I	SSQSLLHS-NGYNYL	F	Q	ALQTP	4-1
II	A3/DPK15	I	SSQSLLHS-NGYNYL	F	Q	ALQTP	4-1
II	A23/DPK16	I	SSQSLVHS-DGNTYL	F	Q	ATQFP	4-1
III	A27/DPK22	I	ASQSVSSS-----YL	F	Q	YGSSP	6-1
III	A11/DPK20	I	ASQSVSSS-----YL	F	Q	YGSSP	6-1
III	L2/DPK21	I	ASQSVSS------NL	F	Q	YNNWP	2-1
III	L16	I	ASQSVSS------NL	F	Q	YNNWP	2-1
III	L6	I	ASQSVSS------YL	F	Q	RSNWP	2-1
III	L20	I	ASQGVSS------YL	F	Q	RSNWH	2-?
III	L25/DPK23	I	ASQSVSSS-----YL	F	Q	DYNLP	6-1
IV	B3/DPK24	I	SSQSVLYSSNNKNYL	F	Q	YYSTP	3-1
V	B2	T	ASQDIDD------DM	F	Q	HDNFP	2-1
VI	A26/DPK26	I	ASQSIGS------SL	F	Q	SSSLP	2-1
VI	A10/DPK26	I	ASQSIGS------SL	F	Q	SSSLP	2-1
VI	A14/DPK25	V	ASEGIGN------YL	F	Q	GNKHP	2-1

genes were detected and consequently these families remained empty. But they contain very few members and in some cases, as in the *38C* gene family, the sequences belonging to these *Igk-V* families might represent highly mutated genes from other families (Strohal et al. 1989). Therefore, we considered that they made only a marginal contribution to the diversity of the entire *Igk-V* repertoire.

The functional mouse Igk-V germline genes and their structural repertoire

Analysis of the expression in vivo of the 49 *Igk-V* genes reported as *Igk-V* germline (see status column of the Fig. 1) suggests that 42 of them are functional (Fig. 2). Within the seven *Igk-V* genes not expressed in vivo, and therefore defined as non-functional, four are seen to posses structural defects when compared with antibodies of known three-dimensional structure (data not shown). The remaining three *Igk-V* genes were considered to have minor genetic defects and were not analyzed further

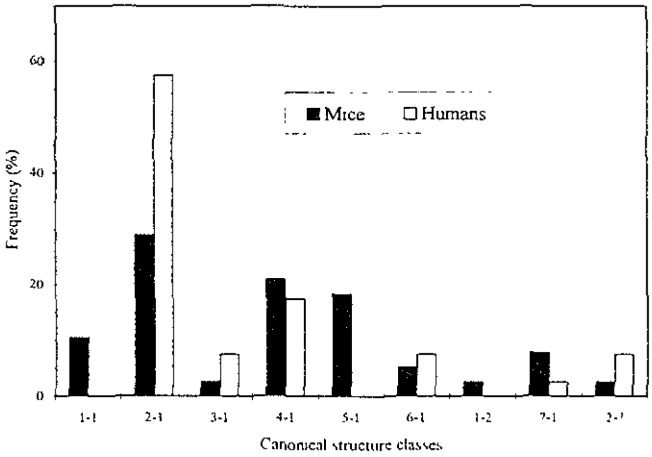
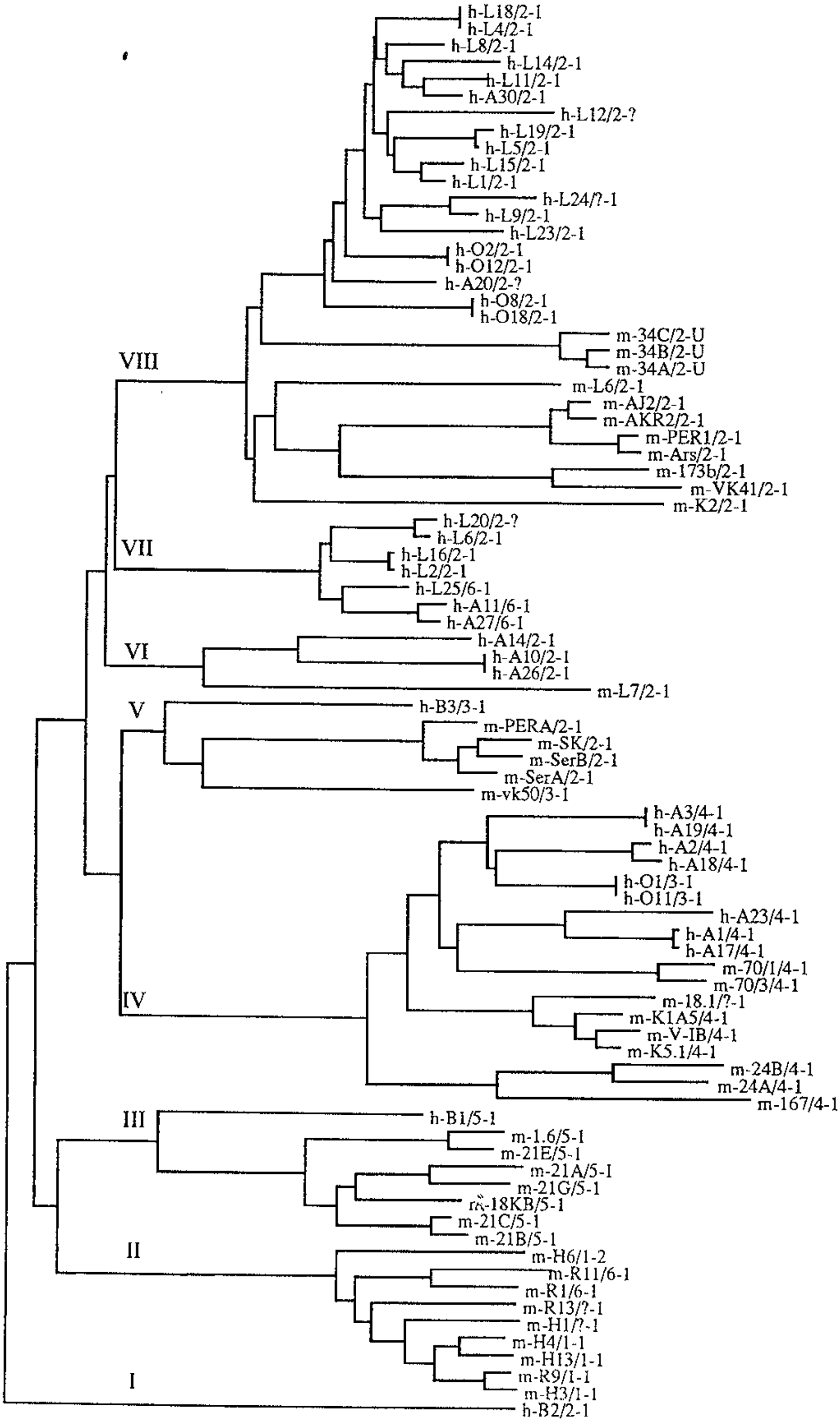


Fig. 5 Comparison of the use frequency of the canonical structural repertoire of mouse and human *Igk-V* germline genes

Fig. 6 Evolutionary relationship between human and mouse *Igk-V* germline genes. The evolutionary tree is based on the nucleotide similarities of mouse and human *Igk-V* germline genes shown in Figs. 3 and 4, respectively Mouse and human genes are distinguished by an "m" or "h" before the name of each gene. After the name, separated by a *slash*, is the canonical structure class encoded in each *Igk-V* germline gene. Being a human pseudogene, *B1* is not included in Fig. 4 and was added to the analysis for completeness. The *B1* sequence was obtained from V-base (web site: http://www.mrc-cpe.cam.ac.uk). The tree was obtained with the CLUSTALW program (Thompson et al. 1994)



Of the 42 functional mice *Igk-V* germline genes, 39 have patterns compatible with some canonical structure in L1 (Fig. 3). In L3, all sequences present canonical structures, with the exception of three *Igk-V* genes (*34A*, *34B*, and *34C*) for which this loop has not been sequenced (Valiante and Caton 1990) and could not be assigned any structure at all.

Inspection of the structural repertoire of the mouse *Igk-V* germline genes (Fig. 3) indicates that the mouse encodes seven canonical structure classes. Classes 2–1 and 4–1 are the most frequent (31% and 19%, respectively) followed by classes 5–1 and 1–1 (17% and 10%, respectively). Classes 6–1, 3–1, and 1–2 are poorly represented in the mouse *Igk-V* germline genes (5%, 2%, and 2%, respectively).

Comparison between the structural repertoires implicit in mouse and human Igk-V germline genes

To compare the mouse and human structural repertoires, the canonical structures classes implicit in the 40 functional human *Igk-V* germline genes are depicted in Fig. 4. In contrast to mice, the entire repertoire of human *Igk-V* germline genes encodes only four canonical structure classes. Canonical structure classes 5–1, 1–1, and 1–2 implicit in the functional *Igk-V* germline genes of mice were found not to have a human counterpart.

In addition, differences exist in the degree to which these species encode class 2–1 (Fig. 5). This class is encoded in mice by 13 of 42 genes (~30%), while in humans it is encoded by 23 of 40 genes (~60%). As can be seen, in humans just one class contributes ~60% of the structural repertoire, whereas in mice this class contributes ~30%. Classes 4–1 (19%), 5–1 (17%,), and 1–1 (10%) play an important role in the mouse, but the last two classes are not present in humans. Thus, the structural repertoire in mice is not only more diverse than in humans, it is also more heterogeneous in the sense of being less skewed or biased towards a particular class.

Discussion

Because the mice variable light chain repertoire is essentially, composed of kappa, while in humans the kappa chain is only ~60%, in the preceding section we compared the structural repertoires of mouse and human *Igk-V* germline genes. Results indicated that in structural terms mouse is more diverse than human. That suggests an evolutionary strategy of mice to compensate for the small lambda chain contribution in its repertoire of variable light chains. To expedite further examination of this suggestion, in Fig. 6 a phylogenetic tree relating the *Igk-V* germline genes of humans and mice is shown.

Tree topology is similar to that proposed by Kroemer and co-workers (1991), who classified human and mouse *Igk-V* genes into groups of *Igk-V* families or clans (Roman numerals in the figure). This classification represents the common ancestral human and mouse *Igk-V* germline genes or *Igk-V* gene families (Kroemer et al. 1991). As is evident from the tree, the structural repertoire implicit in functional *Igk-V* germline genes is far from being randomly distributed within *Igk-V* gene families or even within clans: namely, canonical structures classes are family- and clan-specific. This suggests that the structural repertoire was established 1) prior to the divergence of humans and mice, and 2) has been preserved despite the diversification of human and mouse *Igk-V* germline genes.

On the basis of these suggestions, we followed the evolutionary pathway of the structural repertoire and noticed that clans II and III encode almost exclusively the mouse-specific canonical structure classes 5–1, 1–1, and 1–2. Interestingly, clans II and III have no human counterpart except for sequence B1 (Lorenz et al. 1988) which possesses canonical structure 5–1. However, B1 is a pseudogene due to its having a modified start codon (Klein et al. 1993) and, being non-functional, does not contribute to actual human repertoire variability (we added this sequence to the tree for completeness). Therefore it seems reasonable to propose that *Igk-V* germline genes belonging to these clans were deleted or never developed in humans.

In contrast to the absence of the human functional *Igk-V* germline genes in clans II and III, we found that the mouse *Igk-V* gene families belonging to these clans are two of the three more complex families in that species. Clan II contains the *Igk-V* 4/5 family, the most complex mouse *Igk-V* gene family (see Table 1). Likewise, clan III consists of the mouse *Igk-V* 21 family, which is its third largest (see Table 1). The large number of *Igk-V* germline genes in these families suggested that there were influences to expand them, such as a demand to complement the "poor" diversity encoded by the remaining *Igk-V* gene families. Since expansion of the *Igk-V* 4/5 and *Igk-V* 21 families implied development of the canonical structure classes 1–1, 1–2, and 5–1, which are not present in the functional genes of humans, it can be assumed that these classes developed in mice to supplement the poor structural diversity inherited from the human and mouse ancestors: namely, the remaining four classes, 2–1, 6–1, 4–1, and 3–1.

In humans, the lambda chain might have furnished the additional structural diversity to set aside canonical structure classes 1–1, 1–2, and 5–1. In accordance with this assumption, in structural terms the repertoire of human *Igl-V* germline genes is more diverse than that of mice: human encodes nine canonical structure classes, while mouse only encodes two (Williams et al. 1996). This accounts for the converse relationship we found in the *Igk-V* germline genes and supports the suggestion that human *Igl-V* germline genes supply structural diversity to compensate for the lack of canonical structure classes 1–1, 1–2, and 5–1. This, together with the analysis of the *Igk-V* germline genes, indicates that the ultimate reason for the major diversification of the mouse *Igk-V* structural repertoire is related to the poor lambda chain contribution. In other words, the possible deletion of an important part of the *Igl-V* locus in mice might have forced its *Igk-V* locus to be structurally more diverse. The hypothesis that mice lost an important part of the *Igl-V* locus is supported by phylogenetic analysis showing that the human *Igl-V* genes were originated at early stages of vertebrate evolution, prior to the divergence of humans and mice (Haire et al. 1996).

These results have interesting implications for the evolution of the *Ig-V* genes. To explain their evolution it has been suggested that the *Igh-V, Igk-V,* and *Igl-V* loci evolved by stochastic processes, where no positive darwinian selection operated to retain the complexity of the *Ig-V* loci and the coherence within *Ig-V* gene families (Tutter and Riblet 1989). Alternatively, it has been proposed that some environmental selection pressures retain the complexity of the *Ig-V* loci (Kirkham et al. 1992; Schroeder et al. 1990). In the case of the *Igh-V* locus, this latter suggestion is supported by analysis of the mice and human sequences (Kirkham et al. 1992; Schroeder et al. 1990). It shows that although the number of *Igh-V* genes might vary between species, the genes can be clustered into *Igh-V* families and *Igh-V* clans, which share structural features like the framework regions 1 and 3 (Kirkham et al. 1992; Schroeder et al. 1990). This has been extended to species that diverged from human 200 million years ago (Anderson and Matsunaga 1995) or more (Ota and Nei 1994). The structural features preserved throughout evolution within the *Igh-V* gene families and *Igh-V* clans suggest a reflection of those environmental selection pressures (Kirkham et al. 1992; Schroeder et al. 1990).

Human and mouse *Igk-V* genes can also be clustered into families and clans. However, the nature of the environmental selection pressures operating to retain the complexity and coherence within *Igk-V* gene families remains speculative (Kroemer et al. 1991). Here we found that the evolutionary diversification of the *Igk-V* germline genes preserves the structural repertoire. Moreover, the divergence of the mouse *Igk-V* gene families from those of human could be explained in terms of the structural repertoire diversification strategies. Therefore, we suggest that the conservation of a "basic" structural repertoire, on the one hand, and its diversification to furnish a minimum of structural diversity, on the other, operate as opposite selective pressures to retain the complexity of the *Igk-V* locus and coherence within *Igk-V* gene families.

Similar observations have been made regarding the *Igl-V* locus. It is interesting to note that, for example, horses, in which the lambda chain predominates, have developed more canonical structure classes than have humans (Williams et al. 1996). Thus, further analysis of other species whose kappa and lambda light chain contributions differ would help test the consistency and generalization of an evolutionary model for the *Igl-V*and *Igk-V* loci based on the diversification and conservation of structural repertoires.

Acknowledgments We thank I A Tomlinson for kindly providing the sequences of the 40 functional human *Igk-V* gene segments, and H. Ceceña and B Levin for revision of the submitted manuscript. E.V. was supported by CONACyT grant 1833 PN This work was partially supported by grant DGAPA-UNAM IN213796.

References

Almagro, J.C., Vargas-Madrazo, E., Zenteno-Cuevas, R., Hernandez-Mendiola, V., and Lara-Ochoa, F. VIR: a computational tool for analysis of immunoglobulin sequences. *BioSystems 35:* 25–32, 1995

Almagro, J. C., Domínguez-Martinez, V., Lara-Ochoa, F., and Vargas-Madrazo, E. Structural repertoire in human *VL* pseudogenes of immunoglobulins: comparison with functional germline genes and amino acid sequences. *Immunogenetics 43:* 92–96, 1996

Anderson, A. and Matsunaga, T. Evolution of immunoglobulin heavy chain variable region genes: a *VH* family can last for 150–200 million years or longer. *Immunogenetics 41:* 18–28, 1995

Berek, C. and Milstein, C. The dynamic nature of the antibody repertoire. *Immunol Rev 105:* 5–26, 1988

Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol 196:* 901–917, 1987

Chothia, C., Boswell, D. R., and Lesk, A. The outline structure of the T-cell αβ receptor. *EMBO J 7:* 3745–3755, 1988

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., and Poljak, R. J. Conformations of immunoglobulin hypervariable regions. *Nature 342:* 877–883, 1989

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B., and Winter, G. Structural repertoire of the human V_H segments. *J Mol Biol 227:* 799–817, 1992

Cory, S., Tyler, B. M., and Adams, J. M. Sets of immunoglobulin V kappa genes homologous to ten cloned V kappa sequences: implications for the number of germline V kappa genes. *J Mol Appl Genet 1:* 103–116, 1981

D'Hoostelaere, L. A. and Klinman, D. Characterization of new mouse V kappa groups. *J Immunol 145:* 2706–2712, 1990

Dildrop, R., Gause, A., Muller, W., and Rajewsky, K. A new V gene expressed in lambda-2 light chains of the mouse. *Eur J Immunol 17:* 731–734, 1987

Haire, R. N., Ota, T., Rast, J. P., Litman, R. T., Chan, F. Y., Zon, L. I., and Litman, G. W. A third Ig light chain gene type in *Xenopus laevis* consists of six distinct VL families and is related to mammalian lambda genes. *J Immunol 157:* 1544–1550, 1996

Kabat, E. A. and Wu, T. T. Attempts to locate complementarity determining residues in the variable positions of light and heavy chains. *Ann N Y Acad Sci 190:* 382–383, 1971

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S., and Foeller, C. *Sequences of Proteins of Immunological Interest (5th edn),* Public Health Service. N.I.H. Washington. D.C.,1991

Kirkham, P. M., Mortari, F., Newton, J. A., and Schroeder, H. W. Jr. Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. *EMBO J 11:* 603–609, 1992

Klein, R., Jaenichen, R., and Zachau, H. G. Expressed human immunoglobulin *k* genes and their hypermutation. *Eur J Immunol 23:* 3248–3271, 1993

Kofler, R and Helmberg, A. Comment to the article "A new Igk-V gene family in the mouse". *Immunogenetics 34:* 139–140, 1991

Kofler, R., Geley, S., Kofler, H., and Helmberg, A. Mouse variable-region gene families: complexity, polymorphism and use in non-autoimmune responses. *Immunol Rev 128:* 5–21, 1992

Kroemer, G., Helmberg, A., Bernot, A., Auffray, C., and Kofler, R. Evolutionary relationship between human and mouse immunoglobulin kappa light chain variable region genes. *Immunogenetics 33:* 42–49, 1991

Lara-Ochoa, F., Almagro, J. C., Vargas-Madrazo, E., and Conrad, M. Antibody-antigen recognition: a canonical structure paradigm. *J Mol Evol 43:* 678–684, 1996

Lorenz, W., Schable, K. F., Thiebe, R , Stavnezer, J., and Zachau, H. G. The J kappa proximal region of the human K locus contains three uncommon V kappa genes which are arranged in opposite transcriptional polarities. *Mol Immunol 25·* 479–484, 1988

Martin, A. C. and Thornton, J. M. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol 263* 800–815, 1996

Neuberger, M. S. and Milstein, C. Somatic hypermutation. *Curr Opin Immunol 7:* 248–254, 1995

Ota, T. and Nei, M. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol Biol Evol 11:* 469–482, 1994

Schroeder, H. W. Jr., Hillson, J. L., and Perlmutter, R. M. Structure and evolution of mammalian VH families. *Int Immunol 20:* 41–50, 1990

Selsing, E., Durdik, J., Moore, M. W., and Persiani, D. ML. *In* T. Honjo, F.W. Alt, and T.H. Rabbits (eds.): *Immunoglobulin Genes (2nd edn),* p. 111, Academic Press, New York, 1989

Strohal, R., Helmberg, A., Kroemer, G., and Kofler, R. Mouse *Vk* gene classification by nucleic acid sequence similarity. *Immunogenetics 30:* 475–493, 1989

Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22:* 4673–4680, 1994

Tomlinson, I. A., Cox, J. P., Gherardi, E., Lesk, A. M., and Chothia, C. The structural repertoire of the human V kappa domain. *EMBO J 14:* 4628–4638, 1995

Tonegawa, S. Somatic generation of antibody diversity. *Nature 302:* 575–581, 1983

Tutter, A. and Riblet, R. Conservation of an immunoglobulin variable-region gene family indicates a specific noncoding function. *Proc Natl Acad Sci USA 86:* 7460–7464, 1989

Valiante, N. M. and Caton, A. J. A new *Igk-V* gene family in the mouse. *Immunogenetics 32:* 345–350, 1990

Vargas-Madrazo, E, Lara-Ochoa, F., and Almagro, J. C. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J Mol Biol 254:* 497–504, 1995

Weill, J.-C. and Reynaud, C.-A. Rearrangement/hypermutation/gene conversion: when, where and why? *Immunol Today 17:* 92–97, 1996

Williams, S. C., Frippiat, J.-P., Tomlinson, I. A., Ignatovich, O., Lefranc, M.-P., and Winter, G. Sequence and evolution of the human germline V_λ repertoire. *J Mol Biol 264:* 220–232, 1996

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

Zeelon, E. P., Bothwell, A. L. M., Kantor, F., and Schechte, I. An experimental approach to enumerate the genes coding for immunoglobulin variable-regions. *Nucleic Acids Res 9:* 3809–3820, 1981

Zocher, I., Roschenthaler, F., Kirschbaum, T., Schable, K. F., Horlein, R., Fleischmann B., Kofler, R., Geley, S., Hameister, H., and Zachau, H. G. Clustered and interspersed gene families in the mouse immunoglobulin kappa locus. *Eur J Immunol 25:* 3326–3331, 1995