Universidad Nacional
Autónoma de México

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

Proyecto de Investigación Biomédica Básica
Instituto de Investigaciones Biomédicas

"Reconocimiento Molecular: Patrones en la Geometría del Sitio de Reconocimiento en Inmunoglobulinas"

TESIS

QUE PARA OBTENER EL GRADO DE
DOCTOR EN INVESTIGACION BIOMEDICA BASICA
P R E S E N T A :

Bioq. Enrique Vargas Madrazo

MEXICO, D. F. ABRIL DE 1995

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

a mi esposa y compañera Gladis,
a mis hijos Paolo Seam
y Olaf Giuseppe,
sin los que nada de esto
tendría razón de ser.

a mi Madre, por su esencia y
apoyo.
a mi Familia.
a mi abuela Elvira por su vida.

a Carlos Larralde, quien con su amor
por la ciencia y su amistad siempre
me impulsó en este camino.
a Antonio Lazcano, quien guió
mis primeros pasos en la ciencia.

a lo "Sutil" que está mas allá de la
ciencia, pero que al mismo tiempo nos guía
a través de ella...

Agradecimientos

Por su apoyo en todo momento de este esfuerzo quisiera agradecer al Dr. Francisco Lara Ochoa.

A mis alumnos de la Universidad Veracruzana, con quien he compartido estos años de investigación.

A Augusto Hemández, Julia Trejo, Luis Zamora, Miguel A. Jiménez, Carmen Blazquez y en general a los miembros del Instituto de Investigaciones Biológicas de la Universidad Veracruzana por su apoyo.

A los miembros de mi comité tutorial:
Dr. Carlos Larralde,
Dr. Francisco Lara Ochoa,
y muy especialmente al Dr. Lino Díaz de León por todo su apoyo para hacer este doctorado con tantas facilidades para mi desarrollo.

A Jorge Pérez de la Mora, Luci y María Elena Ortega por su ayuda y apoyo en todo momento.

Por su apoyo académico a:
Dr. Eduardo Horjales,
Dr. Enrique Ortega,
Dr. Xavier Soberón,
Dr. Jaime Lagunez,
y muy especialmente al Dr. Edmundo Lamoyi y la Dra. Carmen Gómez por sus consejos y amistad desde un inicio.

Al Ing. Comadurán y su familia por su cariño y apoyo en el inicio de este camino.

Al Dr. Mario Amzel, por su amistad sencilla y sus consejos sobre el tema.

En especial con gran cariño a Juan Carlos Almagro por su amistad, y por todo lo que juntos hemos vivido y emprendido. También a Ligia y Ana Sofía por su cariño.

Este trabajo doctoral fue realizado en las instalaciones y con el apoyo de:

Instituto de Química, UNAM.

Instituto de Investigaciones Biomédicas, UNAM.

Instituto de Investigaciones Biológicas, Universidad Veracruzana.

bajo la dirección del Dr. Francisco Lara Ochoa.

Resumen

El objetivo principal de la presente tesis es el estudio de las características geométricas generales del sitio de reconocimiento de las Inmunoglobulinas (Igs). Se pretende a partir de este estudio encontrar ciertas reglas en el mecanismo de reconocimiento molecular mediado por Igs. Lo anterior se lleva a cabo mediante la utilización de técnicas de análisis de secuencias y de estructura de proteínas.

El reconocimiento específico en el sistema inmune es mediado por dos familias de proteínas: los receptores de células-T (Tcr's) y las Igs. El sistema inmune es capaz de generar receptores específicos contra un enorme número de moléculas. En la actualidad existe una explosión sin precedentes en la cantidad de información tanto en secuencias como en estructuras de Igs. La utilización de esta información mediante esquemas organizados de análisis permite el estudio de las propiedades generales y particulares del reconocimiento mediado por Igs. Asimismo, gran cantidad de datos funcionales como son; especificidad, afinidad, especies, etc. están disponibles. Como parte del trabajo investigación del doctorado se implementaron un conjunto de programas de cómputo que permiten compilar, manejar y analizar las bases datos de secuencias de Igs. Utilizando esta herramienta se pueden realizar estudios con una enorme cantidad de datos que involucra tanto información estructural como funcional. Se presenta el artículo correspondiente donde se reporta el paquete de programas (VIR) y ejemplos de la aplicación de el programa al estudio del reconocimiento inmune.

El sitio de unión al antígeno de las Igs se encuentra formado por seis lazos o "loops" hipervariables. Estos lazos conectan las hebras beta que forman el andamiaje conservado típico del plegamiento-Ig. El modelo de estructuras canónicas propone que las conformaciones posibles de estos lazos hipervariables estan limitadas a un conjunto pequeño de estructuras canónicas. Mediante el modelo es posible predecir la estructura que tendrá el lazo a partir sólo de la secuencia del dominio variable de la Ig. Para las aprox. 50 estructuras tri-dimensionales de Igs conocidas hasta el momento, este modelo ha resultado válido en su gran mayoría. Lo que implica que la variabilidad estructural de los lazos hipervariables se encuentra fuertemente limitada. En los dos artículos de investigación que se presentan en esta tesis, se utiliza el modelo de estructuras canónicas para caracterizar el repertorio estructural de distintas muestras de Igs.

En el primer artículo de investigación se determina la presencia de estructuras canónicas en las secuencias de pseudogenes del domino variable pesado de Igs. Ha sido reportado en la literatura que casi el 50% de los genes de Igs son pseudogenes, por lo que numerosas preguntas se han hecho respecto del posible papel funcional que estos pseudogenes puedan tener. En este trabajo se encuentra que más del 70% de los pseudogenes de Igs presentan estructuras canónicas. Dicho resultado junto con algunos otros que se presentan en el artículo, permiten evaluar la posible manera en que los pseudogenes pueden contribuir a la diversidad del repertorio estructural de las Igs.

En el segundo artículo de investigación se estudia la base de datos total de secuencias funcionales de Igs mediante el modelo de estructuras canónicas. Se propone el concepto de clases de estructuras canónicas como la combinación de estructuras canónicas que aparecen en los lazos hipervariables que forman el sitio de unión de cada dominio variable. Mediante este concepto, se caracteriza la diversidad del repertorio estructural presente en la base de datos total de secuencias de Igs. De las 300 posibles clases que pueden existir, se encuentra que solo 10 clases representan el 87% de las secuencias analizadas. Asimismo, se encuentra que 6 de estas clases tienen preferencia para reconocer a ciertos tipos de antígenos (haptenos, proteínas, etc.). Se analizan las estructuras tri-dimensionales de Igs con el objetivo de comparar los resultados obtenidos a partir de secuencias con los de las estructuras. Se observa una fuerte correspondencia entre las clases de estructuras canónicas y la geometría del sitio de unión en las estructuras de Igs. Estos resultados permiten proponer que el repertorio estructural de las Igs se encuentra fuertemente restringido a sólo unas cuantas formas. Por otra parte, nos permite proponer reglas entre la geometría del sitio de unión (que es determinada por las combianciones estructuras canónicas) y la función de reconocimiento de la Ig. Se propone un modelo de reconocimiento molecular mediado por Igs.

Abstract

The main purpose of the present thesis is the study of the general geometrical characteristics of the recognition site of the Immunoglobulins (Igs). Based on this, it is proposed to found some rules in the mechanism of molecular recognition mediated by Igs. The above is realized by using techniques of sequences analysis and protein structure.

The specific recognition that is characteristic of the immune system is mediated by two families of specific receptors: The receptors of T-cells (Tcr's) and the Igs. The immune system has the capacity to generate specific receptors against most of molecules. Actually, there is an explosion without precedents in the amount of information available both at sequence and structure level for Igs. The use by organized schemes of analysis of this information allows the study of the general and particular properties of the recognition mediated by Igs. In addition, a great amount of functional data like: specificity, affinity, specie, etc. are available. Due this, as a part of the research work of the doctorate, it was implemented a set of computational programs that allow to compile, manage and analyze the sequence databases of Igs. By the use this tool, it is possible to realize studies with an enormous amount of data that involve structural and functional information. In this thesis it is presented an article in which is reported the program (VIR) and examples of the application of the program.

The antigen-binding site of Igs is formed by six hypervariable loops. These loops link the beta-strands that form the conserved framework characteristic of the Ig-fold. The canonical structure model proposes that the possible conformation that could adopt the hypervariable loops is restricted to a small set of canonical structures. By this model it is possible to predict the structure of a loop based only on the sequence of the variable domain of the Ig. For the approx. the 50 three-dimensional structures of Igs reported now this model has been confirmed in most of cases. This implies that the structural variability of the hypervariable loops is strongly restricted. In the two research papers presented in this thesis, it is used the model of canonical structures to characterize the structural repertoire of different samples of Igs.

In the first research article it is determined the presence of canonical structures in the pseudogene sequences of the variable domain of the heavy chains of Igs. It has been reported in the literature that 50% of the Ig genes are pseudogenes. Therefore, several questions have been formulated respect the possible functional role that could play the pseudogenes of Igs. In this study it is found that more than 70% of the Ig pseudogenes have canonical structures. The above result, together with other results presented in the article, allows to evaluate the possible manner in which the pseudogenes could contribute to the diversity of the structural repertoire of the Igs.

In the second research article it is studied the total database of functional Igs with the model of canonical structures. It is proposed for this study the concept of canonical structure classes as the combination of canonical structures that appear in the hypervariable loops that form the antigen-binding site on each variable domain. By this concept, it is characterized the diversity of the structural repertoire present in the total database of Igs. Of the 300 possible classes that could exist, it is found that only 10 classes represent the 87% of all the sequences analyzed. Also, is found that six of the classes have preference to recognize certain types of antigens like hapten, protein, etc. Based on this, it is analyzed the three-dimensional structures of Igs with the objective to compare the results obtained from the sequences with those from the three-dimensional structures. It is observed a strong correspondence between the canonical structure classes and the geometry of the binding-site found in the three-dimensional structures. These results allow to propose that the structural repertoire of the Igs is strongly restricted to only a few geometrical forms. On the other hand, it allows to propose a set of rules to relate the geometry of the binding-site (determined by the combination of canonical structures) and the recognition function of the Ig. A model of the mechanism of inumune recognition mediated by Igs is proposed.

Indice

1. INTRODUCCION.

El fenómeno de la especificidad es una de las cualidades distintivas de los sistemas biológicos (Breckenridge 1991). Debido a que los seres vivos son esencialmente máquinas químicas los procesos de Reconocimiento Molecular (RM) constituyen una parte sustancial y la base mecanística de la mayoría de los procesos vitales relacionados con la especificidad. La comprensión de los mecanismos que determinan el RM ha sido una de las grandes incógnitas en química orgánica y bioquímica desde que Fisher propusiera el concepto de reconocimiento basado en la estereoquímica en 1894 y Ehrlich acomodara la imagen de Fisher en términos de la inmunología describiendo la interacción como "... el antígeno llave en la cerradura del anticuerpo..." en 1897 (Breckenridge 1991).

Actualmente el estudio del RM constituye una de las ramas de mayor crecimiento en química y bioquímica (Lehn 1990) y ocupa un papel central en la mayoría de los estudios y aplicaciones tanto en biología, biomedicina y biotecnología. Su rango de aplicaciones va desde la biocomputación (Conrad 1985), diseño de receptores, enzimas y demás macromoléculas, hasta el avance en la comprensión de la organización y funcionamiento de los sistemas inmune, endocrino y nervioso (Suckling 1991, Breckenridge 1991).

Una de las áreas de mayor importancia en el RM es la especificidad en el reconocimiento inmune, tanto por sus implicaciones prácticas, como por el conocimiento básico (Telford y Stimson 1991). El estudio de la interacción epítopo/parátopo constituye uno de los modelos de especificidad más característicos, debido a la gran versatilidad de este sistema y a la amplia disponibilidad de datos en estructura tridimensional, secuencia, función y genética.

Estudios en múltiples modelos de RM en años recientes, convergen a una visión en la cual el fenómeno de la especificidad entre un ligando y un receptor se encuentra determinado esencialmente por dos tipos de procesos: i) un ajuste entre superficies complementarias (Rebek Jr. 1991) que permitan una proximidad en el rango en el cual las interacciones débiles son significativas (Burley y Petzko 1988, Friedman y col. 1994), y ii)

una distribución complementaria de las interacciones débiles (puentes de Hidrógeno, puentes salinos, interacciones polares débiles) que determinen la especificidad y afinidad (Pauling 1945, Fersht y col. 1985, Rebek Jr. 1991).

Con la caracterización de los principales mecanismos genético-moleculares responsables de la generación de la diversidad de la respuesta inmune (Tonegawa 1983), la opinión que prevalecía en la literatura a principios de los 80's era que puestos en marcha estos procesos de generación de diversidad, sólo es necesario que el azar y el tiempo suficiente transcurra para lograr generar receptores que reconozcan con alta especificidad y afinidad a un epítopo dado. Sin embargo, en los últimos años con la disponibilidad de gran cantidad de secuencias y estructuras de inmunoglobulinas (Igs), se han encontrado numerosas evidencias que indican la existencia de sesgos y patrones a distintos niveles de organización de la respuesta inmune (Dildrop 1984, Ohno y col. 1985, Chothia y col. 1986, 1989, 1992, Padlan 1990, Mian y col. 1991, Vargas-Madrazo y col. 1993, 1994, Lara-Ochoa y col. 1995). Se considera que estos factores de información pre-codificada pueden complementar y eficientizar el funcionamiento de los procesos básicos que generan la diversidad de la respuesta inmune.

A nivel de secuencia en 1977 Kabat y col. reportan la existencia de sitios altamente conservados (similar al encontrado en las regiones de hebras beta) en las regiones hipervariables. Estudios posteriores (Ohno y col. 1985, Chothia y Lesk 1987, Padlan 1990, y Vargas-Madrazo y col. 1993, 1994) han confirmado dichas observaciones. Asimismo se ha detectado uso preferencial de ciertos aminoácidos tanto en el sitio de reconocimiento en su conjunto (Padlan 1990, Mian y col. 1991) como en cada una de las posiciones presumiblemente responsables de la especificidad (Kabat y col. 1977, Vargas-Madrazo y col. 1993, 1994, Lara-Ochoa y col. 1995). Desde el punto de vista estructural ha sido propuesto por Chothia y col. (Chothia y col. 1986, 1989, Chothia y Lesk 1987) que el número de conformaciones que adopta la cadena principal de los llamados lazos hipervariables se encuentra fuertemente restringido a sólo un pequeño conjunto. Estas "conformaciones canónicas" se determinan por la presencia de ciertos aminoácidos claves

en ciertas posiciones (Chothia y Lesk 1987). El modelo de estructuras canónicas ha resultado enormemente útil como base para predecir la estructura de las lazos hipervariables (Chothia y col. 1989) de un gran numero de inmunoglobulinas recientemente reportadas (Padlan 1994). Debido a esto y a la gran cantidad de secuencias disponibles actualmente es posible emprender estudios generales para tratar de caracterizar la propiedades globales tanto fisico-químicas (Padlan 1990, Mian y col. 1991, Vargas-Madrazo y col. 1994) como geométricas del sitio de unión al antígeno (Chothia y col. 1992, Cox y col. 1994, artículos presentados en la presente tesis).

Este conjunto de evidencias indican la existencia de un cierto grado de estructura tanto en el mecanismo mismo de reconocimiento molecular como en los factores genéticos, moleculares y celulares, que forman la base de la respuesta inmune.

En el presente trabajo, utilizando el análisis de secuencias y estructuras de Igs, bajo la premisa del modelo de estructuras canónicas se pretende proporcionar mayores evidencias que apoyen esta visión de la respuesta inmune. Asimismo se propone un modelo general del mecanismo de reconocimiento inmune.

2. ANTECEDENTES.

2.1. Enfoque Estructuralista en Biología e Inmunología.

Históricamente en ciencia se ha establecido una dicotomía que contrasta al funcionalismo con el estructuralismo (Judson 1979, Van Regenmortel 1989). En biología y en particular en biología molecular, esta dicotomía es claramente distinguible por el tipo de técnicas empleadas en el estudio de los procesos biológicos (Judson 1979). En general, se distingue a los estudios estructurales como aquellos que utilizan principios físicos para estudiar la **estructura** del material biológico. Algunos autores identifican la diferencia fundamental en el hecho de que el enfoque estructural no considera el componente tiempo, mientras que éste es parte fundamental para el enfoque funcional (Van Regenmortel 1989).

Específicamente en el problema del reconocimiento antígeno-anticuerpo, es un tema de gran discusión la determinación de los residuos en la interfase responsables del reconocimiento. Para varios complejos se encuentra contraposición entre los resultados obtenidos mediante estudios estructurales y funcionales (Getzoff y col. 1988). Los estudios estructurales identifican un numero elevado de residuos (entre 15-20) y generalmente ubicados en la superficie de las moléculas acomplejadas como los responsables del reconocimiento. Mientras que estudios funcionales identifican solo unas cuantos residuos (entre 3-7) y en ocasiones lejos de la superficie de interacción como los responsables del reconocimiento. Este hecho ejemplifica según Van Regenmortel (1989) la diferencia en el tipo de información y de las conclusiones a las que puede llegar cada enfoque. En general, desde el punto de vista del enfoque funcional, el enfoque estructural, **reduce** artificialmente la complejidad del sistema, al proponer solo una descripción de su estructura, considerando en el modelo solo algunos componentes del fenómeno estudiado (Van Regenmortel 1989). No obstante, por ejemplo en la actualidad en estructura de macromoléculas ha surgido una rama (la dinámica molecular) que permite estudiar los aspectos temporales de la estructura de las macromoléculas. Pudiéramos concluir que en la actualidad la diferencia entre estos dos enfoques no es tan radical y que en ocasiones se presentan enfoques intermedios (quizás aun más fructíferos, pero no muy comunes).

Desde un punto de vista más amplio en el contexto de la filosofía de la ciencia, el concepto de **estructuralismo** implica un enfoque más general de análisis y que más adelante definiremos. Este enfoque estructuralista en ocasiones se relaciona en la literatura con lo que anteriormente definimos como enfoque estructural en ciencias naturales.

El enfoque estructuralista debe considerarse como un enfoque **reduccionista, de naturaleza explicatoria** (Thom 1972). El origen formal de este enfoque científico y filosófico se remonta a los estudios lingüísticos de de Saussure y antropológicos de Lévi-Strauss de fines del siglo XIX y principios del XX. Este tipo de estudios se han extendido a otras disciplinas basándose en ciertos conceptos fundamentales.

A continuación se definen los conceptos esenciales del estructuralismo, que dan

una idea clara de lo que esta visión científica implica.

Noción de sistema. Se define a un sistema real, como un conjunto abierto conectado de componentes que interactúan en un espacio-substrato. El concepto de espacio-substrato implica las nociones de orden y jerarquía (Thom 1972).

Estructura. La estructura de un sistema implica la descomposición del sistema en un conjunto de componentes elementales (Thom 1972).

En la visión estructuralista, no se trata de explicar la morfología mediante su reducción a elementos obtenidos de otra teoría -supuestamente más elemental o fundamental- de la manera como se trata de explicar la biología mediante la física y/o la química, o la sociología mediante la biología y/o la psicología; solamente se trata de mejorar la descripción de la morfología empírica mediante la exhibición de sus regularidades, sus simetrías profundas, mostrando su unidad interna. En este aspecto el estructuralismo es una teoría modesta, ya que su único propósito es mejorar la descripción (Thom 1972).

Por lo tanto, creo que es importante el delimitar las diferencias que entre el enfoque estructural en ciencias naturales y el enfoque general filosófico del estructuralismo. El primero (en particular en biología molecular) se refiere principalmente al análisis de la estructura física y química de los sistemas y donde generalmente el componente tiempo no es incluido. El objetivo central de este enfoque es describir y entender los principales rasgos **morfológicos** del sistema, para así relacionarlos con la función del sistema. El enfoque estructuralista implica una posición de estudio en términos global ante el proceso natural. Como define Thom (1972), es un enfoque reduccionista y de naturaleza explicatoria, que trata de encontrar las **regularidades** internas (**lógicas**) del sistema. Es decir, trata a los procesos naturales más en cuanto a su estructura lógica desde un punto de vista muy cercano al matemático (Thom 1972) aunque no necesariamente lo implica. En conclusión, quizás la diferencia fundamental entre estos dos enfoques radica en cómo comprendamos los siguientes conceptos; estructura física (enfoque estructural) y estructura lógica (enfoque estructuralista)

En general el científico experimental trata de señirse más a sus propias

metodologías y técnicas antes de intentar tomar las de la filosofía. En caso contrario se corre el peligro de entrar quizás en contradicciones de gran complejidad y de poco valor científico (Horjales 1995). No obstante, una vez aclarados aquí los conceptos de estructuralismo y enfoque estructural, considero que puede resultar productivo el utilizar prudentemente (cuando menos como guía metodológica) los postulados principales del estructuralismo. Y por o tanto así lo realizo en el presente trabajo doctoral. Esto, ya que principalmente utilizaré aquí el enfoque estructural en el estudio del reconocimiento inmune. Como se verá más adelante, detrás de las concretas técnicas de biología molecular estructural utilizadas en mi investigación siempre está una búsqueda de la estructura lógica del sistema y de la consecuente existencia de reglas.

En resumen, considero que el enfoque utilizado en este trabajo incluye no solo los estudios de la **estructura física** del sistema (en este caso las inmunoglobulinas) que denominamos enfoque estructural, sino además aquellos estudios que forman parte de un esfuerzo por comprender los mecanismos fundamentales que rigen el funcionamiento del sistema (la interacción antígeno-anticuerpo) basándose en la noción de la **estructura del sistema.**

2.2. Antecedentes en el Estudio del Sitio de Reconocimiento en Igs.

2.2.1. Estudios a Nivel de Secuencia.

El estudio directo de las características del sitio de reconocimiento a nivel de secuencia se plantea formalmente por primera vez con los trabajos clásicos de Kabat y Wu en 1970 y 1971, en donde definen la localización precisa de las regiones hipervariables (Wu y Kabat 1970, Kabat y Wu 1971). Estos autores postulan que estas regiones determinan la formación de la superficie de interacción con el epítopo (Regiones que Determinan la Complementaridad o CDRs). Dicha propuesta se basa en la hipótesis de que si la propiedad distintiva de los anticuerpos consiste en su diversidad, esta propiedad debe expresarse en

una alta variabilidad de aminoácidos localizada en las posiciones que hacen contacto con el antígeno, y por lo tanto determinan la especificidad. En 1973 se dilucidan las primeras estructuras tridimensionales de inmunoglobulinas (New, McPC603, Meg y Rei) y la hipótesis de las regiones hipervariables se confirma desde el punto de vista de la estructura tridimensional. Se observa que las regiones hipervariables convergen en una porción del dominio variable formando una endidura, esta regiones hipervariables coinciden además con los lazos que unen las hebras beta, que son las porciones con mayor posibilidad de variación en conformación (Nisonoff 1975). Dado el éxito del enfoque propuesto por Kabat y Wu el problema de analizar el tipo de variabilidad que se presenta en el sitio de reconocimiento no es abordado en la bibliografía de los años posteriores. Sólo el mismo Kabat y col. y Padlan a finales de los 70's retoman este tipo de análisis. Kabat y col. (1977) determinan la existencia de residuos en los CDRs que se encuentran altamente conservados, proponiéndose por primera vez, en base a evidencias experimentales, que este hecho puede estar vinculado con el mantenimiento de ciertas características estructurales indispensables para el reconocimiento. Padlan (1977 y 1979) por otra parte demuestra que las sustituciones que se presentan en las regiones que establecen contacto con el epitopo son más radicales (en términos de una medida de disimilitud estructural[1]) que en las regiones de andamiaje. En 1985 Ohno y col. retoman el problema señalando que existen numerosas posiciones altamente conservadas en los CDRs y que aparentemente la variabilidad de los residuos hipervariables esta lejos de ser al azar (Ohno y col . 1985).

En 1990 Padlan realiza un análisis detallado del uso de aminoácidos en las posiciones de los CDRs, demostrando que los aminoácidos hidrófobos se encuentran poco usados y aquellos con múltiples capacidades para establecer interacciones con los átomos complementarios del epítopo se encuentran altamente representados (como es el caso de Tyr, Trp y Arg)(Padlan 1990). Mian y col. (1991) confirman los resultados de Padlan pero

---

[1] Medida de disimilitud entre los 20 aminoácidos en términos de las propiedades físico-químicas que tradicionalmente se han considerado como las más importantes para el plegamiento de las proteínas.

adicionan el criterio de solo considerar en los cálculos aquellas posiciones que se han identificado como en contacto en los seis complejos antígeno-anticuerpo que hasta el momento se conocían. Adicionalmente Mian y col. (1991) proponen una racionalización de estos resultados basada en un minucioso análisis de las propiedades físico-químicas de cada aminoácido. El trabajo realizado por nuestro grupo ha seguido esta línea de investigación cuyo objetivo es determinar el tipo de variabilidad que se presenta en cada una de las posiciones que pueden estar involucradas en la interacción del anticuerpo (Almagro y col. 1995, Cocho y col. 1993, Lara-Ochoa y col. 1994 y 1995, Vargas-Madrazo y col. 1993, 1994).

Más recientemente Kabat y col. han retomado el esfuerzo por comprender las propiedades generales del mecanismo de reconocimiento inmune. Han estudiado la frecuencia en que se presentan secuencias iguales entre anticuerpos con diferente especificidad para identificar cuales son las regiones que más contribuyen en la determinación de la especificidad (Kabat y col. 1991a). Asimismo estudian la distribución de longitudes en aminoácidos de la tercera región hipervariable de la cadena pesada para comprender con mayor detalle el papel que esta región juega en la especificidad y afinidad del anticuerpo (Wu y col. 1993).

2.2.2. Estudios de estructura tridimensional y descripción del sitio de unión al antígeno.

A lo largo de la historia de la inmunoquímica la noción de complementaridad ha sido el concepto central para comprender el reconocimiento molecular en el sistema inmune. Los estudios tridimensionales que se inician con la década de los 70's permiten evaluar toda la información recabada durante más de 80 años de investigación inmunoquímica. Al mismo tiempo dan la posibilidad de proponer nuevos modelos más ambiciosos, tratando de explicar más detalladamente la propiedades del reconocimiento inmune.

En 1972 aparecen dos trabajos (Epp y col. 1972, Poljak y col. 1972) que exponen por primera vez la estructura general de las inmunoglobulinas a baja resolución (6.0 Å). El

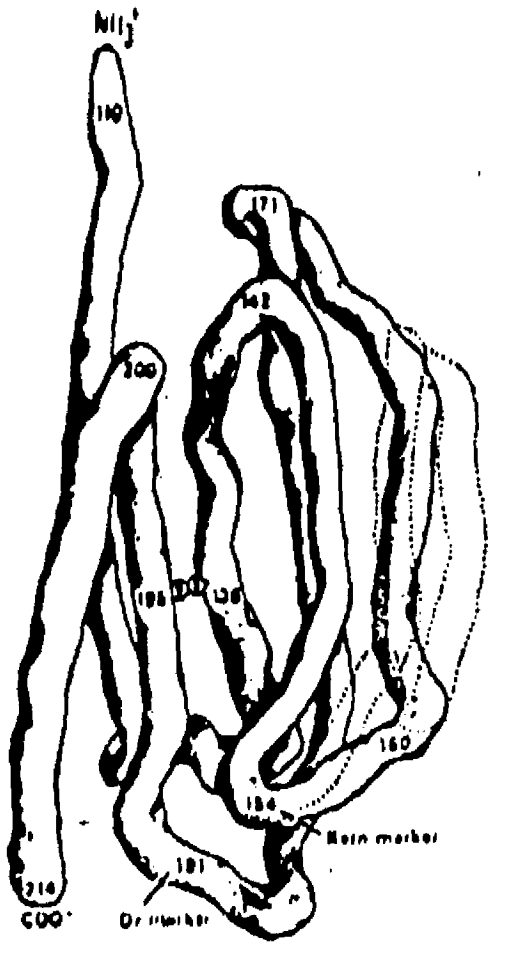
grupo de Huber trabaja con un dímero de cadena ligera (Rei), mientras que el de Poljak trabaja con el Fab de mieloma del paciente New. La característica más importante de la estructura de estas proteínas es el plegamiento de la cadena en dos láminas beta construidas por hebras antiparalelas, que generan el "plegamiento-Ig" (Poljak y col. 1973) (véase la Figura 1). Se identifica la región amino-terminal como la responsable del reconocimiento debido a que se aprecia un bolsillo en esta zona, además de que por difusión de haptenos en los cristales de Igs se determina que los haptenos difunden hasta este bolsillo (Poljak y col. 1973).

La determinación de la estructura de una proteína sin la ayuda (disponible actualmente) de programas de cómputo que procesan la información de los rayos difractados, constituía a principios de los 70's un trabajo enorme. Es por esto que todos los resultados y modelos proporcionados por los estudios en secuencias enumerados en la sección 2.2.1. fueron determinantes para el éxito de los estudios cristalográficos. De particular importancia para la interpretación de los resultados de la difracción de los cristales fue el concepto de dominio (Edelman y Gall, 1969) como unidad de repetición en las Igs.

Posteriormente, en 1973, estas dos moléculas son resueltas a media resolución 2.0-3.0 Å. También se reporta en este año la estructura del Fab McPC603 de ratón a media resolución (Padlan y col. 1973). Con esta resolución se puede identificar la localización aproximada de cada residuo (sobre todo si se conoce la secuencia), por lo que se reporta la espectacular comprobación de la hipótesis de los CDRs de Kabat y Wu (Poljak y col. 1973, Schiffer y col. 1973, Padlan y col. 1973). Estos segmentos que por su hipervariabilidad eran postulados como los responsables de la especificidad, se encuentran efectivamente convergiendo en un extremo de la molécula, formando las paredes del bolsillo de reconocimiento. Se determina las dimensiones aproximadas de este bolsillo como de 15 Å de diámetro por 10 Å de profundidad.

En estos trabajos y muchos otros posteriores, se describen innumerables características del "plegamiento-Ig", como son la interfase entre los dominios, el lazo extra

A



B

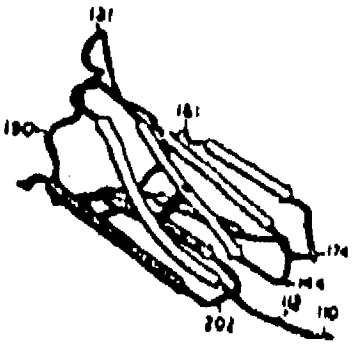


Figura 1. Plegamiento típico de las inmunoglobulinas. A) un modelo a baja resolución, que muestra las hebras beta como segmentos rectos de la cadena. Se observa el "extra-loop" característico de los dominios variables. B) Representación en listones del plegamiento-inmunoglobulina. Los listones con flecha representan las hebras beta. (tomados de Amzel y Poljak 1979).

que presentan los dominios variables respecto a los constantes (Figura 1), los ángulos entre los dominios variables, el ángulo entre el eje de simetría de los dominios variables con el eje de simetría de los constantes, etc., etc. El objetivo de esta revisión no es el de describir estos detalles, además de existir excelentes revisiones al respecto (Amzel y Poljak 1979, Marquart y Deisenhofer 1982, Davies y Metzger 1983, Mariuzza y col. 1987, Alzari y col. 1988, Davies y col. 1990, Janin y Chothia 1990, Padlan 1994), por lo que refiero al lector a estos artículos.

Otro paso muy importante fue la resolución de la molécula New en complejo con el derivado hidroxilo de la vitamina K (Amzel y col. 1974), así como del Fab de mieloma McPC603 de ratón en complejo con fosforil-colina (Padlan y col. 1973). De estos estudios se puede analizar en detalle la naturaleza de los enlaces que contribuyen a la energía de interacción. Se identifica como los más importantes los puentes de Hidrógeno, los puentes salinos y las interacciones de van der Walls. La estrecha complementaridad entre la geometría del hapteno y su anticuerpo fue un elemento de gran trascendencia que brindaron estos estudios (Day 1990).

Como ya se ha mencionado anteriormente el concepto de complemetaridad es esencial a lo largo de toda la historia del estudio de la interacción antígeno-anticuerpo. Históricamente este concepto fue redefinido rigurosamente en términos de la fisico-química por Pauling a principios de los años 40's. Pauling enfatiza la importancia de la complementaridad tanto geométrica como química entre las moléculas que interaccionan para poder compaginar la naturaleza inespecífica de las interacciones débiles con la especificidad del reconocimiento inmune (Pauling 1945). Por lo tanto, podemos considerar como el elemento más importante de esta primera etapa de estudios tridimensionales, la interpretación rigurosa de este principio expuesto por Pauling en términos de información tridimensional.

A mediados de los 80's aparece la primera estructura de un complejo de anticuerpo (D1.3) contra un antígeno protéico (lisozima de huevo) (Amit y col. 1986). Durante la segunda mitad de los 80's se resuelven otros tres complejos antígeno-anticuerpo más, otras

dos más contra distintos epitopes de lisozima, HyHEL-5 (Sheriff y col. 1987) y HyHEL-10 (Padlan y col. 1989), y otro del anticuerpo el NC41 contra la neuraminidasa del virus de la influenza (Colman y col. 1987). Un análisis detallado de estos estudios se encuentra en la revisión de Davies y col. (1990).

A partir de estos modelos se han podido detectar ciertas regularidades entre los distintos complejos, así como diferencias entre los complejos anti-hapteno y anti-macromolécula. Algunas de las características comunes en complejos anti-macromolécula son: 1) Superficie de interacción (entre 500 y 800 $Å^2$), 2) Gran número de puentes de Hidrógeno e interacciones de van der Walls, así como uno o dos puentes salinos, 3) Número de aminoácidos del anticuerpo que hacen contacto con el antígeno (entre 15 a 20), 4) Superficies planas, con numerosas protuberancias y depresiones, y 5) Una elevada **complementaridad** entre las superficies. Las diferencias más importantes entre complejos contra macromolécula y contra hapteno son: 1) En todos los casos para haptenos los lazos se prolongan hacia el solvente, generando un bolsillo que recibe al hapteno, mientras que para antígenos grandes en general el CDR-3 de la cadena pesada bloquea el bolsillo formando una superficie plana, 2) Los puentes salinos parecen tener mucho menor importancia para antígenos grandes que para haptenos. En ambos casos se presenta un elevado número de Tirosinas en la interfase. Por otro lado, la opinión generalizada de los investigadores hasta la aparición de estos modelos, era que como resultado de la interacción entre anticuerpo y ligando (ya sea hapteno o macromolécula) no se producen cambios conformacionales significativos en el anticuerpo (Davies y col. 1990).

Los últimos cuatro años pueden referirse como una etapa cualitativa y cuantitativamente distinta, ya que debido a la utilización de la técnica de reemplazamiento molecular (Brünger y col. 1991) que permite refinar con enorme facilidad estructuras nuevas a partir de estructuras similares conocidas, en combinación con las técnicas de ingeniería genética, ha habido una explosión en el número de estructuras tridimensionales y en el tipo de especificidad de los anticuerpos analizados. Actualmente se han reportado unos 50 complejos (no todos de alta resolución) incluyendo a anticuerpos en complejo con

haptenos, ácidos nucléicos, polisacáridos, péptidos y proteínas, es decir, casi toda la gama de posibles ligandos (Padlan 1994).

Considero que debido a la enorme cantidad de información surgida recientemente, así como por la complejidad del problema que requiere de nuevos enfoques, hasta el momento no se han publicado trabajos que integren la nueva información de manera global en la visión que se tiene del reconocimiento inmune. A mi entender, ésta es una fase muy similar a la que antecedió al artículo de Wu y Kabat en 1970, que marcó un cambio en los estudios posteriores.

No obstante, considero existen tres puntos importantes a resaltar como nuevos elementos y que, ligando los fenómenos de especificidad, reactividad cruzada y cambios conformacionales, han cuestionado la visión que se tenía a fines de los 80's. Poniendo en vigencia de nuevo las propuestas de Pauling de hace 50 años: 1) El componente de movilidad de las cadenas laterales de los residuos altamente expuestos (sobre todo de aminoácidos aromáticos) (Tainer y col. 1985, Geysen y col. 1987, Padlan 1990, Herron y col. 1991, Mian y col. 1991, Rini y col. 1992, Arévalo y col. 1993), que permite a un mismo anticuerpo reconocer distintos ligandos, 2) La presencia de cambios conformacionales, como efecto de la formación del complejo, y cómo estos cambios conformacionales pueden mejorar la interacción antígeno-anticuerpo (en términos geométricos y energéticos) (Edmundson y col. 1987, Herron y col. 1989, 1991, Fan y col. 1992, Wilson y Stanfield 1993), y 3) Como consecuencia de lo anterior (ya que estos cambios conformacionales ocurren principalmente al modificarse el ángulo entre el dominio variable ligero y el variable pesado), el que los residuos hipervariables que se encuentran en la interfase de los dominios variables juegan un papel en la especificidad, la diversidad y la reactividad cruzada (Edmundson y col. 1987, Herron y col. 1989, 1991, Fan y col. 1992). A este respecto, pudiera decirse que quizás nos encontramos en este momento (parafraseando a Kunh, 1986) en la etapa de crisis del paradigma de la llave-cerradura rígida, y que esta antecede a una revolución científica de sustitución del paradigma.

2.3. Estudios estructurales (siguiendo principios estructuralistas) realizados a partir de los años 70's sobre el problema del reconocimiento inmune.

Como mencioné en la sección 2.1. considero que en general existe confusión en cuanto a la definición de lo que implica el enfoque estructuralista en biología y en particular en inmunoquímica, por lo que es importante revisar en forma sistemática los trabajos que se han realizado en esta área. Esto tanto como un elemento de consulta posterior, como para servir de marco de referencia a la línea teórica que sustentan los estudios que aquí se presentan.

En esta sección se reportan los trabajos más importantes que han abordado con un enfoque estructural (pero siguiendo a mi entender algunos principios estructuralistas) el problema del reconocimiento inmune, ya sea utilizando datos de secuencias o de estructura tridimensional de inmunoglobulinas, así como de datos funcionales. Como aspectos principales de estos principios estructuralistas considero: 1) noción de sistema y 2) existencia de una estructura, que implica la existencia de reglas.

La literatura sobre este tema es escasa. Solo una treintena de artículos han aparecido durante el cuarto de siglo posterior a los trabajos de Kabat y Wu de 1970 y 1971. Los estudios realizados durante los 70's a nivel de secuencia se basan en la observación de dos características de los dominios variables como punto de partida para estudiar la existencia de reglas en el mecanismo de reconocimiento inmune; 1) Las secuencias del dominio variable se agrupan en familias de homología (Kabat 1968), es decir, existe cierta organización (familias de genes) entre las secuencias, la cual puede correlacionarse con una organización en el mecanismo de reconocimiento, y 2) En los dominios variables se localizan residuos altamente conservadas (Kabat 1967) aún en los segmentos hipervariables (Kabat y col. 1977), lo cual sugiere le existencia de restricciones estructurales en el mecanismo de reconocimiento.

Estudios estadísticos sobre el uso de aminoácidos en los segmentos hipervariables permiten identificar claramente la presencia de posiciones altamente conservadas en los CDRs. De acuerdo con estos estudios y datos de las estructuras conocidas se identifican los

residuos que son responsables de la especificidad en los anticuerpos (Kabat y col. 1976, 1977). Se propone que los residuos altamente conservados son necesarios para mantener cierta estructura mínima del sitio de unión del anticuerpo requerida para un reconocimiento adecuado (Kabat y col 1977). Es interesante, desde el punto de vista de una visión general del sitio de unión, que Edelman y Gall en 1969 ya habían postulado la existencia de tres tipos de residuos en el sitio de unión al antígeno: a) residuos de *contacto*, que están directamente involucrados en la unión con el antígeno, de manera similar a la unión de los substratos por las enzimas, b) residuos *moduladores*, que pueden variar para cambiar el plegamiento de las cadenas polipeptídicas que forman el sitio de unión y así permitir que los residuos en contacto se localicen en las posiciones adecuadas para unir al antígeno, y c) residuos *compensatorios* que permiten mantener el plegamiento general del sitio de unión y del dominio en general, para compensar la variación de los dos tipos anteriores de residuos (Edelman y Gall 1969).

Padlan (Padlan y Davies 1975, Padlan 1977, 1979), estudia en términos estructurales el grado de variabilidad de los segmentos hipervariables así como de la región de "andamiaje". Encuentra que los CDRs difieren entre las distintas Igs considerablemente más que la región de andamiaje. No obstante, en sus trabajos trata de buscar si ésta "hipervariación estructural" de las lazos, se encuentra acotada sobre cierto rango. Sin embargo, el reducido número de estructuras de que disponía no le permitió encontrar estos patrones a nivel estructural. Como veremos más adelante Chothia y col. (1986, 1989), obtienen resultados positivos, debido al crecimiento de la base de datos cristalográfica y a un esquema de análisis más completo.

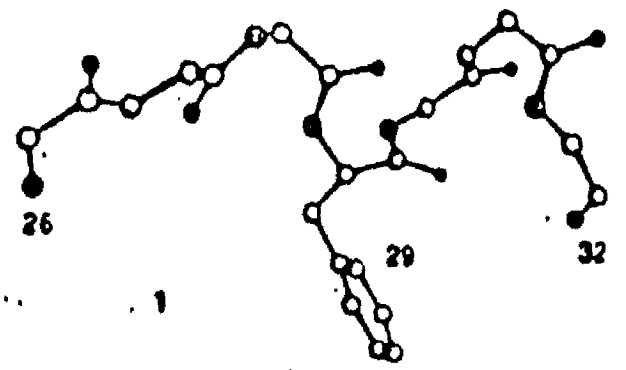
Aunque no se analizará en detalle el trabajo respecto a la hipótesis de minigenes propuesta por Kabat y col. a fines de los 70's (Kabat y col. 1978, 1980a, 1980b), es importante señalar que la búsqueda de patrones y un conjunto de reglas sencillas que explicaran a nivel genético la generación de la diversidad en la respuesta inmune, también fue abordada por Kabat con base al análisis de secuencias. Se refiere al lector a los artículos originales (Kabat y col. 1978, 1980a, 1980b).

Dildrop (1984), realiza estudios de homología entre las secuencias del dominio variable de la cadena pesada de Igs en ratón, y propone una nueva clasificación en 7 familias de secuencias basada en la homología entre las secuencias de todo el dominio variable. Estas familias de secuencias de aminoácidos corresponden con las siete familias de genes identificadas por métodos de genética molecular. Dildrop reporta que estas familias de genes de Igs generan anticuerpos con especificidades características. Es decir que las especificidades no se distribuyen aleatoriamente entre las familias de genes de la cadena pesada. Como es fácil deducir, este esquema propone una organización muy precisa de la respuesta inmune.

A pesar de que éste modelo no ha sido explorado con más detalle a nivel de secuencia, ni se han evaluado sus implicaciones en la estructura tridimensional, es un modelo muy ambicioso, basado en datos muy limitados, por lo que sus alcances también lo son. No obstante, propone una línea de trabajo muy interesante, sobre todo en el contexto de la información disponible actualmente.

Los trabajos iniciales de Padlan (Padlan y Davies 1975, Padlan 1977, 1979), para tratar de delimitar la variabilidad estructural de los lazos del sitio de unión al antígeno, fueron continuados por Chothia y col. (Chothia y col. 1986, 1989, Chothia y Lesk 1987), encontrando que efectivamente, las conformaciones que adoptan las lazos hipervariables, están limitadas a un número reducido de estructuras (que denomina estructuras canónicas) (ver la Figura 2). La presencia de cada una de estas estructuras canónicas, puede ser determinada por la longitud del lazo y por la presencia de residuos determinados en posiciones clave (Chothia y Lesk 1987). Este esquema de análisis fue corroborado en Igs que estaban por ser resueltas (Chothia y col. 1986, 1989), ya que las conformaciones propuestas resultaron sobre el margen de error respecto a las determinadas experimentalmente. Debido a que las variables que determinan las estructuras canónicas son a nivel de secuencia (longitud y residuos claves) este análisis puede ser aplicado a la base de datos de secuencias actualizadas de inmunoglobulinas de Kabat y col. (1991b), permitiendo un estudio detallado de la organización de las secuencias, su relación con las

A



B

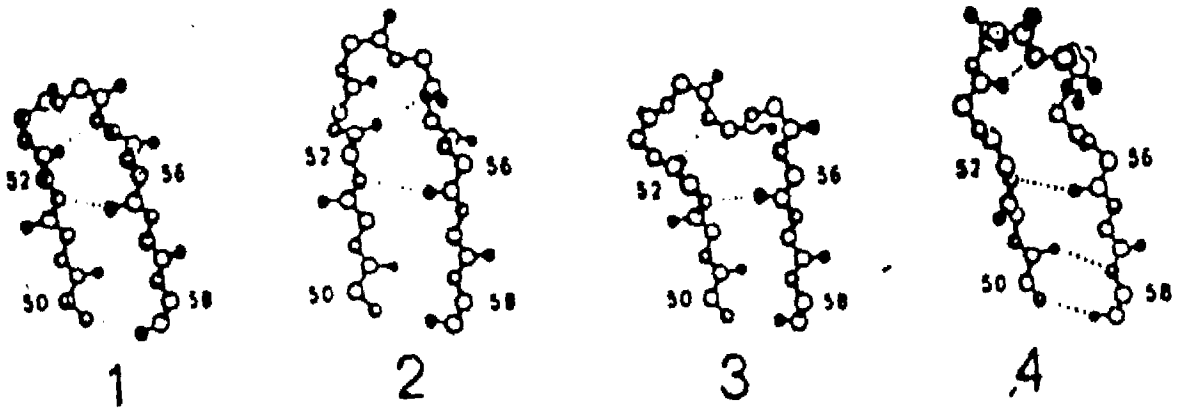


Figura 2. Estructuras canónicas para los "loops" hipervariables de la cadena pesada (según Chothia y col. 1992). A) Se muestra el tipo 1 y el tipo 2, este último tiene una inserción respecto al tipo 1. Se señalan los residuos que determinan la estructura. B) Se presentan las cuatro conformaciones canónicas para el segundo loop hipervariable de la cadena pesada.

estructuras canónicas y con los datos de especificidad. Es precisamente éste el tema que ocupa los dos trabajos de investigación reportados en esta tesis, en donde se caracteriza el repertorio estructural del sitio de unión al antígeno y se prueba que las estructuras canónicas se encuentran en un elevado porcentaje de las bases de datos de Igs, aun en pseudogenes (ver más abajo en resultados).

Recientemente, se ha retomado el enfoque propuesto por Kabat a finales de los 70's (Kabat y col. 1976 y 1977) en el sentido de analizar en detalle la variabilidad y el uso de aminoácidos en relación con la especificidad y las propiedades generales del sitio de unión (Padlan 1990, Mian y col. 1991, Vargas-Madrazo y col. 1992, 1993, 1994, Almagro y col. 1995, Lara-Ochoa 1994, 1995). Los resultados de estos trabajos indican una fuerte restricción en el uso de aminoácidos en el sitio de unión al antígeno. También se observa efectivamente que existen grupos de residuos que cumplen con distintas funciones para el mecanismo de reconocimiento. Estos resultados que de manera general indican la existencia de propiedades inherentes al reconocimiento general, están en concordancia con los estudios de los grupos de Edmundson (Herron y col. 1991, Bhat y col. 1990), Colman (Colman y col. 1987) y Wilson (Rini y col. 1992, Arévalo y col. 1993), que proponen la existencia de cambios conformacionales como elemento central del mecanismo de reconocimiento. El modelo de ajuste inducido propuesto por estos autores permite explicar el fenómeno de reactividad cruzada y la multi-especificidad de los anticuerpos y llega a conclusiones similares a las de estudios de uso de aminoácidos, ya que ambos indican la existencia de un sitio de unión con propiedades generales y cierto grado de versatilidad en el reconocimiento. Este modelo esta en contraposición (Wilson y Stanfield 1993) con la visión rígida del modelo tradicional de llave-cerradura (Amit y col. 1986). Todos estos resultados, indican la necesidad de revisar los conceptos de especificidad, reactividad cruzada y diversidad, para poder comprender la organización de la respuesta inmune.

El enfoque propuesto por Chothia y col. (Chothia y col. 1986, 1989), proporciona una extraordinaria herramienta para el análisis de secuencias. No obstante, es necesario sustentar estos análisis con información funcional y genética. Un excelente par de artículos

presentados por Chothia en el análisis estructural, y Winter y Walter en la secuenciación y análisis de genes de línea germinal de Igs en humano (Chothia y col. 1992, Tomlinson y col. 1992), presenta un análisis detallado en donde, en base a un directorio de secuencias de genes se analiza la existencia de estructuras canónicas para los CDRs 1 y 2. Ellos encuentran que el 100% para CDR-1 y el 96% para CDR-2 de los genes analizados presentan patrones de secuencia acorde con las estructuras canónicas. Adicionalmente se propone el concepto de clase de estructura canónica para la combinación de estructuras canónicas en CDR 1 y 2 en una misma secuencia. Se propone la siguiente nomenclatura: si en un gene variable se encuentra la estructura canónica tipo 2 en el CDR-1 (véase la Figura 2) y la estructura canónica tipo 3 para el CDR-2, ese gene tiene la clase de estructura canónica 2-3. Esta observaciones sirven de base para modelar los tipos de superficies que cada gene de línea germinal aportará al repertorio inmune (Chothia y col. 1992). Un análisis similar para la cadena ligera y para las combinaciones de cadenas ligeras y pesadas que se observan en las respuesta inmune permitiría predecir en su conjunto la geometría general de la superficie del sitio de unión al antígeno que se forma, únicamente conociendo las secuencias. A partir del análisis que realizan Chothia y col. (1992) se estima que son aproximadamente 50 las distintas superficies que genera el grupo de genes variables de la cadena pesada en humano (Tomlinson y col. 1992). Este tipo de estudios marca una nueva fase (a mi juicio similar al de los artículos de Kabat y Wu), dentro del estudio del reconocimiento inmune.

En esta misma dirección, Vargas-Madrazo y col. (1992, 1994, Lara-Ochoa 1995), han analizado para varios niveles la organización de la respuesta inmune (genes de línea germinal, pseudogenes, secuencias de aminoácidos de anticuerpos funcionales, para vertebrados superiores e inferiores, etc.) la validez de los patrones de uso de aminoácidos en el sitio de unión encontrados previamente en muestras totales. En estos trabajos también se ha introducido información funcional, por ejemplo, construyendo alineamientos con igual número de secuencias de las 253 especificidades reportadas, construyendo sub-muestras de grupos de especificidades bioquímicamente similares (anti-proteínas, anti-haptenos, anti-

polisacáridos, etc) A partir de estas muestras "controladas" se estudia en detalle el uso de aminoácidos, encontrándose que persisten los patrones detectados en las muestras globales no controladas (Lara-Ochoa y col. 1995).

La dos líneas de investigación mencionados anteriormente (el análisis de estructuras canónicas ligado a información funcional y el estudio de uso de aminoácidos), representa el tipo de análisis estructuralista en esta área que considerando información funcional a todos los niveles de organización, intenta proponer modelos generales de reconocimiento para entender la naturaleza de la diversidad de la respuesta inmune. Evidentemente estos modelos generales deben ir acompañados por descripciones de los casos particulares para los cuales no son validas algunas de las reglas, ya que como todo proceso biológico el reconocimiento inmune es una conjunción de regularidades con exquisitas particularidades.

3. METODOLOGIA.

3.1. Heurística General de Investigación.

Dentro de la investigación científica de los sistemas naturales, históricamente existen dos grandes enfoques que en ocasiones resulta difícil armonizar; el experimental y el teórico. El presente trabajo a pesar de no ser exclusivamente teórico si utiliza métodos característicos de este enfoque, como son; el empleo de herramientas matemáticas y cibernéticas, el análisis global de datos, la modelación por computadora de procesos naturales, etc. Considero que para que estudios que involucran metodología teórica aporten resultados relevantes y confiables acerca de fenómenos biológicos, se deben considerar ciertos aspectos que a continuación mencionaré; 1) La proposición de experimentos con base en la consideración de las conclusiones y modelos surgidos de los estudios experimentales, 2) La construcción de bases de datos considerando las restricciones físicas, químicas y biológicas del sistema, 3) La evaluación de los resultados en términos de los

datos experimentales conocidos, 4) En particular para este problema de biología molecular, la conjunción de resultados obtenidos del análisis de estructura tridimensional con los de secuencia, así como 5) el análisis detallado de casos particulares para su comparación con las tendencias generales.

4. DESARROLLO DE LA INVESTIGACION.

4.1. Secuencia de los resultados presentados.

Los artículos que se presentan aquí y los resultados que estos reportan se enfocan en el objetivo de caracterizar la variación que en términos estructurales se presenta en el sitio de unión al antígeno de las inmunoglobulinas. Estos estudios se realizan basándose en el modelo de estructuras canónicas para los lazos hipervariables de las inmunoglobulinas propuesto por Chothia y col. (Chothia y Lesk 1987). Partiendo de este modelo se estudia el grado de diversidad estructural presente en las distintas muestras estudiadas. Se caracteriza por primera vez de manera general la diversidad del repertorio estructural de las Igs y se correlacionan estos resultados con información referente a la especificidad de los anticuerpos. Los resultados obtenidos se comparan con un análisis de la geometría del sitio de unión en base a las estructuras tridimensionales reportadas. Considero que las conclusiones que se obtienen de estos trabajos permiten proponer un modelo del reconocimiento inmune más completo el cual incluye información tanto de secuencia, de estructura tridimensional así como funcional.

Se presenta en primer lugar el artículo donde se reporta la herramienta de trabajo -paquete de cómputo de análisis de secuencias de receptores de sistema inmune "VIR"- desarrollada por Juan C. Almagro y Enrique Vargas-Madrazo. Este paquete de cómputo y las bases de datos que se construyen utilizando el paquete constituyen el eje central con el cual se realizan todos los estudios.

Seguidamente se presenta el artículo donde se analiza la presencia de estructuras

canónicas en secuencias de pseudogenes del dominio variable de la cadena pesada de inmunoglobulinas

Finalmente se presenta el artículo más extenso, donde se caracteriza la diversidad estructural del repertorio de inmunoglobulinas. En este trabajo se describe en términos estructurales el repertorio de los sitios de unión al antígeno a partir de la base de datos total actualizada de inmunoglobulinas.

4.2. Métodos, Resultados y Discusión (Artículos de Investigación).

BIO1479
/RLB

BioSystems VOL (1994) PP-PP

VIR: A computational tool for analysis of immunoglobulin sequences

J.C. Almagro[a,*], E. Vargas-Madrazo[b], R. Zenteno-Cuevas[a], V. Hernández-Mendiola[b], F. Lara-Ochoa[a]

[a] *Instituto de Química. UNAM Circuito Exterior. Ciudad Universitaria. C.P. 04510. México, D.F., México*
[b] *Instituto de Investigaciones Biológicas. Universidad Veracruzana. Xalapa, Veracruz, México*

Received 27 May 1994, accepted 12 August 1994

Abstract

In this paper a microcomputer software named VIR (Variable domains of the Immune Receptors) is reported. This package can be used in sequence studies of immunoglobulin variable domains. The main features of the VIR software in the sequences management are: (1) ease of information recovery/extraction from amino acid sequences; and (2) its capability to obtain multiple sequence alignments with predefined characteristics (i.e. specie and/or specificity). As an analytical tool, the VIR package employs such multiple sequence alignments to compute: (1) tables showing amino acid frequencies, (2) three variability indexes, (3) identity matrices, (4) random samples, and (5) sequences with possible canonical structures. Thus the software reported here here is proposed as a useful tool to carry out detailed studies of immunoglobulin variable domains.

Keywords: Data bases; Variability analysis; Sequence analysis; Pattern recognition; Canonical structures

1. Introduction

One of the core problems in molecular biology concerns the constraints determining structure-function relationships in proteins. In principle, it can be safely assumed that such constraints can be found by analyzing amino acid patterns in key positions of a multiple sequence alignment (Zuckerkandl, 1976). In the particular case of the immune system, where the only source of specificity at molecular level seems to come from two proteins: T-cell receptors (TCR) and immunoglobulins (Ig), a long standing question prevails: is it possible to find amino acid patterns to reveal the structure-function relationships in the antigen binding sites of TCR and Ig?.

Prediction of the antigen binding site or complementarity determining regions (CDRs) in Ig by variability analysis, prior to the resolution of Ig structures (Wu and Kabat, 1970), made possible the beginning of understanding specificity (function) mediated by these molecules. Subsequently to the resolution of several Ig structures, analysis of its three-dimensional conformation have revealed canonical structures in five of its six CDRs (Chothia and Lesk, 1987; Chothia et al., 1989). Existence of these canonical structures in the CDRs of Ig implies that just a few main-chain conformations are present in a large set of Ig molecules with

* Corresponding author. Tel.: +525 616 2576, Fax: +525 616 2217; E-mail address: salazar@redvax1.dgsca.unam.mx.

SSDI 0303-2647(94)01479-Q

different loop amino acid sequences and specificities (Chothia and Lesk, 1987, Chothia et al., 1989). Recent studies have enhanced the knowledge of Ig structure-function relationships by finding definite amino acid bias when analyzing positions that interact with antigens in antigen-antibody complexes (Mian et al., 1991, Lara-Ochoa et al., 1994). This suggests a general mechanism underlying the molecular recognition mediated by Ig (Lara-Ochoa et al., 1994).

A next step is to correlate functional and structural information. In the opinion of the authors, the main limitation to achieve this, has been the absence of user-friendly computational tools to perform such analysis. To circumvent this difficulty in this paper a microcomputer software named VIR (Variable domains of the Immune Receptors), developed to easily retrieve and analyze Ig amino acid sequences, is reported. VIR interfaces with Kabat's data base, which is known to possess all currently updated information on Ig sequences (see for example Kabat et al., 1991). To do so, this program makes use of SEQHUNT internet service which, in turn, has the necessary tools to access this information. The data are translated to VIR data base format. This data base structure optimizes the information management and its recovery form Ig sequences, and allows use of the different analytical tools available with the VIR package.

On its analytical side, VIR uses two different approaches: statistical analysis and pattern recognition. Statistical analysis provides computation of: (1) tables of amino acid usage by position; and (2) three different variability indexes. Correspondingly, the pattern recognition approach provides: (1) patterns compatible with key residues responsible for determining the Ig-fold; and (2) canonical structures in the CDRs of Ig sequences.

In order to properly outline the work in the following sections, a complete — but not exhaustive — description of the VIR data bases format and the management and analytical tools are given in the first place. Secondly, examples of the two different approaches in Ig sequence analysis are shown and discussed. Finally, the significance of the VIR software is discussed in the light shed by current trends in data bases management and sequence analysis.

2. Description of the VIR package

All programs were written in the Turbo Pascal version 5.5 language, and were compiled to run on an IBM-compatible PC microcomputer. A user-friendly menu-driven graphical interface system was developed with on-line help available to explain each and every item appearing in the main menu or submenus.

In order to readily start using the package two data bases are available with it. One corresponds to the variable domains of light (V_L) chains while the other does so with the heavy (V_H) chains. Each data base contains a multiple sequence alignment as well as the information associated with each sequence. In the V_H case, the data base contains all the sequences compiled in the Kabat's data base up to April 1994. For V_L the data base contains the sequences from the same source of V_H sequences. Both of them are ASCII files. Therefore, data bases available with the program, updated ones, or those created by the user, can be displayed, edited or printed using a text processor capable of importing ASCII files (WordPerfect, Word, Word for Windows, and the like).

2.1. VIR data base structure

Functional, technical and bibliographic information in the VIR data bases is codified on a label associated with each sequence in the multiple alignment according to the following description:

Descriptor	Position on label
Technique	1
Year	2–3
Species	4–5
Journal	6–8
Volume	9–11
First page of the paper	12–16
Name given by Kabat et al. (1991)	17–30
Particular specificity	31–33
General specificity	36
Binding constant (if known)	37
Subclass (kappa or lambda)	38
Subgroups of Kabat et al. (1991)	39–40
Space	41–50
Sequence	51

Thus, each line of the VIR data base is composed of a label (left side) and the sequence codified in the one letter amino acid code (right side), for example

```
position   1         10        20        30        40        50
           |   |     |   |     |   |     |   |     |   |     |
           NW9RNPNAR6 2341 3.14.9          LEVANUNK5********DIQMT...AS
           |----------Label-----------------------------------|sequence|
```

The label is used by the management tools of the package, while the sequence is processed by the analytical tools (see below). In order to codify the describers of the label, a commitment was made between classifying the maximum information and avoidance redundancies or ambiguities. In the case of specificity, judged by the authors as the most important functional describer, it should be emphasized that two describers, particular and general, are used. General specificity refers to the groups of specificities (for example anti-protein Ig), and it can thus be used to directly obtain data bases with global specificities. Particular specificities refers to the specificities (for example anti-lysozyme Ig), allowing to stratify the search.

Concerning the multiple sequence alignments, two conventions have been proposed in the literature regarding insertions numbering and placement inside CDRs of Ig: (i) Kabat et al. (1991); and (ii) Chothia et al. (1987). We chose the convention proposed by Chothia et al. (1987) because it has been developed on the basis of comparisons among Ig of known three-dimensional structures.

2.2. Tools of management and analysis in the VIR package

The package consists of a master program that controls four main modules: Open, Data, Analysis and Up-to-date (Fig. 1). (Words in bold italics are those shown to be chosen in all menus and submenus below described).

1. The Open module is the one in charge of loading data bases, either those available with the package *Ig* or any *Other* crated by the user in VIR format.

2. The Data module is the data base manager of the package. It is divided in the two functions below indicated by the first indentation level:

(i) To provide a complete characterization of the data base currently being used, the Information submenu (see Fig. 1) is used. This submenu report.

Total number of sequences.

Number of sequences by *Technique* (amino acid sequences or nucleotide translation).

Number of sequences according to the *Year* of publication.

Number of sequences by *Specie*.

Number of sequences by *Specificity*.

Number of sequences with reported *Affinity*.

Number of kappa or lambda (*k* or *l*) Sequences (Only applicable to V_L.)

Number of sequences by Kabat's *Subgroup* (Kabat et al., 1991).

Matrix: this item allows self-comparison of the sequence data base (that is to say, with

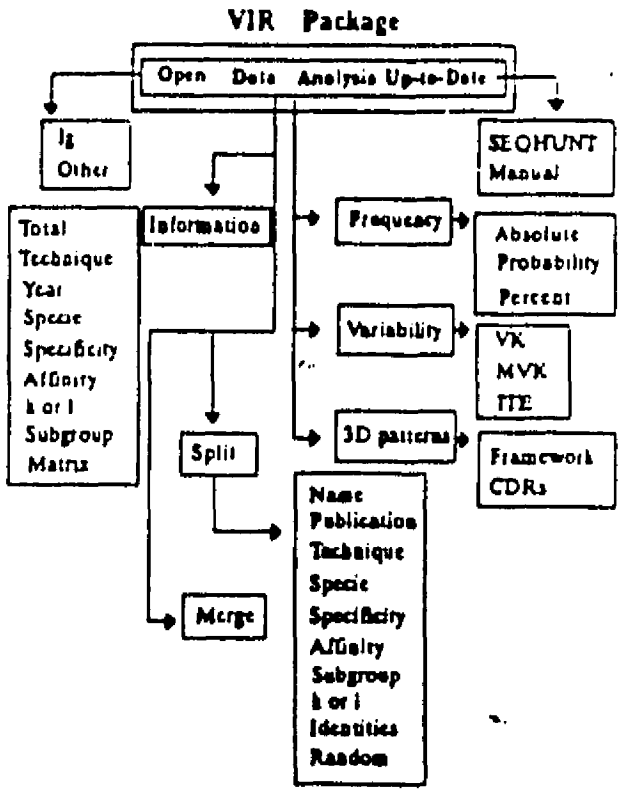


Fig. 1. VIR package menu system.

itself) making possible to identify redundancies in it (ii) To create new data bases with predefined features (sequences from a single specificity, for example). This is achieved using the Split submenu and the Merge option (see Fig. 1). The latter merges two data bases, while the former creates a new data base given one pre-established criterion Below, the different criteria concerning the Split submenu are explained. Generation of a data base with sequences given its *Name*(s).

Generation of a data base according to bibliographic information (*Publication*)

Generation of a data base according to the *Technique* used to obtain the amino acid sequences (amino acid sequences or nucleotide translation). Generation of a data base with sequences of a single *Specie*.

Generation of a data base with sequences of a single *Specificity* (General and particular).

Generation of a data base selecting those sequences with known *Affinity* constant.

Generation of a data base selecting those sequences within a Kabat *Subgroup* (Kabat et al., 1991).

Generation of a data base containing only kappa or lambda (*k* or *l*) sequences.

Generation of a data base obtained by analyzing *Identities* among amino acids with a query sequence within a range of maximum and minimum Identity threshold.

Generation of a *Random* data base with a given number of sequences to be taken from the data base in use.

For the Species, Specificity, Publication and Subgroup options, submenus listing the corresponding different classifications are displayed by the package.

Output files from the Split submenu are in VIR data base format (see VIR data bases structure above). Hence, combining the Split submenu and the Merge option, new data bases performing any combination of the described items can be obtained, for example, a not-before-hand included data base containing human anti-lysozyme with known Affinity constant.

3. The Analysis module is the third one of the package and was designed bearing two conceptually different techniques in mind to analyze retrieved data bases: statistical techniques and sequence pattern analysis. Concerning statistical analysis two methods are offered, the most simple one computes of amino acid frequencies via the Frequency submenu composed of three options: *Absolute*, *Probability* and *Percent* (see Fig. 1). Output coming from this submenu are tables in ASCII format listing the positions (rows) and the numerical values associated with each amino acid (columns). The second statistical technique is Variability analysis (Variability submenu, see Fig 1). Three variability indexes are provided: Kabat's variability index (*VI*) (Wu and Kabat, 1970), modified Kabat's variability index (*MVI*) (Jores et al., 1990), and informational-theoretical entropy index (*ITE*) (Shenkin et al., 1991). Output coming from this submenu consists of tables in ASCII format listing the positions in the amino acid sequence and the numerical values given to each position in the multiple sequence alignment.

The second analytical (non-statistical) technique before mentioned, namely pattern recognition (3D patterns submenu in Fig. 1), was developed to estimate the relationship among Ig of known three-dimensional structures and amino acid sequences. The *Framework* option compares amino acid sequences with a pattern of residues identified as mainly responsible for determining the Ig-fold (Chothia et al., 1988), while the *CDRs* option searches for canonical structure patterns in the data base currently being used.

4. The Up-to-date module is an interface with the SEQHUNT internet service to update data bases already created. As in previous examples, it is further subdivided in two items: *SEQHUNT* and *Manual*. SEQHUNT converts VIR data base format into Kabat's data base format and vice versa. This is the interface with the Kabat's data bases containing currently updated sequences of Ig variable domains. Manual allows the user to actualize the data bases or to analyze a sequence not reported by the SEQHUNT service simply by typing its sequence.

3. Applications

In order to show the potential use of the package, two applications of the VIR software are

described in this section. Application (a) shows an example of a statistical analysis, while application (b) performs an analysis of a typical immunologic problem

3.1. Example of predefined data bases and amino acid usage

It has recently been shown that several positions within CDRs of Ig — historically considered as randomly hypervariable — have preference for certain amino acids (Vargas-Madrazo et al., 1994a; Lara-Ochoa et al., 1994). The bias found in the amino acid usage of these positions could be due to: (i) a general feature of the Ig molecular recognition; or (ii) some artefact introduced by certain over-represented specificities resulting in predominance of some sequences in the data base (e.g. anti-hapten Ig). The second possibility could be discarded by building balanced samples with equal number of sequences of different specificities and then, comparing their amino acid usage with that of the total sample (total data base).

Thus, two samples were built as follows:

Sample 1. Using the Split submenu in the Data module a data base was generated by randomly choosing one sequence from each specificity (210 different specificities) in the total data base.

Sample 2. Following the already mentioned procedure another data base was built choosing five sequences from specificities having at least five sequences (52 different specificities) in the total data base.

Each data base was loaded with the Other option in the Open module, and using the Percent option in the Frequency submenu, the corresponding tables of amino-acids percent by positions were computed. The results for one position in the V_L are shown in Fig. 2. It was observed that the patterns of amino acid usage found in the three samples are rather similar. Statistical calculations confirm this result, and the same behavior was obtained for other positions in CDRs (Vargas-Madrazo et al., in preparation). Therefore, this analysis discards the existence of any bias as a consequence of a predominance of certain specificities in the total sample analyzed, suggesting that the preferential use of certain amino acids in CDRs is a general feature of the Ig molecular recognition.

3.2. Example of canonical structures

Genetically, the first two CDRs of V_H (H1 and H2) of Ig are produced by the V_H germline genes (Tonegawa, 1983). Additionally to V_H germline genes, most species retain a pool of V_H pseudogenes estimated to be 30% of the total number of V_H genes (Kodeira et al., 1986). Besides, polymorphism studies of human V_H pseudogenes report high sequence conservation among unrelated individuals (Pascual and Capra, 1991). These facts, suggest a possible functional role for pseudogenes in the human immune response diversity (Wysocki and Gefter, 1989, Pascual and Capra, 1991). In this application additional arguments in favor of this hypothesis are given by the combined use of Framework and CDRs items, within 3D patterns submenu (Vargas-Madrazo et al., 1994b).

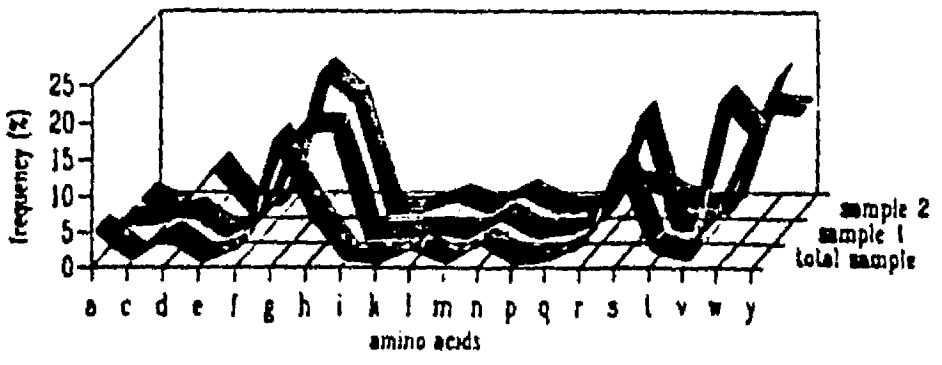


Fig. 2. Amino acid usage for position 91 of V_L. This position is selected as an example because it does not determine the canonical conformation (Chothia and Lesk, 1987) and contacts the antigen in 6 of the 8 antigen–antibody complexes (Padlan and Kabat, 1991). The number of sequences in each sample is: total sample: 1934 sequences; sample 1: 210 sequences; sample 2: 260 sequences.

Table 1
Structural divergence for V_H human germline genes and pseudogenes. We consider as defective or mutated those positions not having amino acids codified in the Framework item. Defective substitutions on each sample were averaged dividing the total number of defective positions by the number of sequences in each sample

Pseudogenes	Germline genes
6.3	0.5

It has been proposed that the number of mutations accumulated in pseudogenes with respect to germline genes is principally a function of either (i) the time elapsed since inactivation of the gene has occurred; or (ii) possible functional restrictions acting on the new state of the gene (Pascual and Capra, 1991). In the first case, random mutation distributions should be expected while in the latter a pattern — a canonical structure, that is to say — should emerge. Since human V_H germline genes show a large percentage of canonical structures within the first two CDRs (Chothia et al., 1992; Vargas-Madrazo et al., 1994b), then it might be possible that such patterns could also be present in these sequences.

In order to test the above mentioned hypothesis comparisons of mutations distributed in framework and CDR positions were carried out. So, the framework option was used on human V_H germline genes and pseudogenes to determine the degree of structural divergence (non-conservative amino acid substitutions) accumulated at framework positions (Table 1). As can be seen, mutations that may alter the framework stability of the human V_H pseudogenes are about 12 times as divergent as those of human V_H germline genes. A high divergence degree in pseudogenes, with respect to the germline genes, suggests that the production of a structurally stable variable domain is not the function of selection for the V_H pseudogenes. Therefore, if producing a structurally stable domain is not the function of selection, then pseudogenes could be a source of additional diversity for the CDRs or, otherwise, be irrelevant for the immune response.

Table 2
Sequences with canonical structures patterns in human V_H germline genes and pseudogenes

CDRs	Pseudogenes (%)	Germlines genes (%)
H1	74	100
H2	43	96

Contrasting with the fact that pseudogenes have an average of six destructive mutations per sequences in frame work positions (see Table 1), it was found using the CDRs item that 3/4 of the sequences have canonical structures for H1, while half of sequences do so for H2 (Table 2). The case of H2 having less sequences with canonical structures might be closely related to the fact that this CDR is, in general, less conserved (Chothia et al., 1992). Taken together, these results suggest that functional restrictions are directly responsible of canonical structure conservation (Vargas-Madrazo et al., 1994).

4. Discussion

In the last 13 years, since the first sequence compilation (Dayhoff and Eck, 1966), accumulation of quantitative information about nucleotide and amino acid sequences at an ever increasing rate makes it very difficult to assimilate and/or to analyze its possible meanings. Projects to map and sequence complex genomes including the human one are underway (Bell, 1990). Sequences analysis, in the molecular biology field can be divided into two main branches: (i) DNA or amino acid sequences collection in data bases; and (ii) tools development for sequence analysis research.

Regarding the first point, in the Ig case, the functional, technical, bibliographic data and sequences, have been collected from the Kabat's data base. In April 1994, the number of Ig sequences or fragments in this data base roughly amounted to 8000 (4700 for V_H and 3300 for V_L), classified under 460 different specificities. All this information made it possible for careful evolutionary and functional analysis in order to begin to understand the structural basis of the immune recognition mediated by Ig.

In parallel with this large amount of experimental data, in the last 20 years the Ig-fold has been the subject of many structural investigations, from which, antibodies are considered the best known structure of all proteins to date (Padlan, 1994). This structural knowledge permits starting from a more robust working hypothesis to study the relation between sequences and their three-dimensional structure. Thus, combining this kind of analysis with functional information (specificities and/or affinity constant), structure-function relationship studies can be improved in Ig to account for the second branch of the sequences analysis. The main limitation to achieve this, however, has been the absence of user-friendly 'immunologist-language-like' tools to manage and analyze Ig sequences. The software described here is a first attempt to carry out this demand.

In the Analysis module of the VIR package three main analytical tools are given: amino acids frequency; variability analysis; and pattern recognition analysis. Calculation of amino acid frequencies allows analysis of amino acid propensities in positions of interest in Ig sequences. Application (a) shows the combined use of functional information from Ig sequences and this kind of analysis. Such analysis allows to study the general and particular properties of CDRs throughout the construction of samples considering the functional information relevant to the process (i.e specificity).

In relation to the variability indexes, variability analysis by Kabat index, at the time of its development, played a major role in the structural analysis of Ig, successfully predicting the regions responsible for the antigen specificity, before X-ray data could be obtained (Wu and Kabat, 1970). Other two variability indexes have been developed; the modified Kabat index (Jores et al., 1990), and the informational-theoretical entropy index (Shenkin et al., 1991). Simultaneous use of the Kabat index and the modified Kabat index has been employed as an analytical tool to investigate the residues responsible for the different detailed specificities of various lectins (Young and Oomen, 1992). Recently, variability analysis was extended to simultaneously use the three currently known variability indexes to understand the interaction between the TCR and peptide-MHC complex (Almagro et al., 1994). This shows the potential uses of this analysis in other protein families or, in our case, in the detailed evolutionary or functional analysis of Ig.

Finally, the last implementation, pattern recognition, deals with the proposition of supplying an analytical tool to perform studies of the relation between sequences and their corresponding three-dimensional structures. Application (b) describes a structural analysis of Ig sequences with an unknown functional role. Thus, it is expected that this analysis will allow a more rational view of the molecular recognition process mediated by antibodies (Chothia et al., 1992).

In summary, the software described is an attempt to provide a package of tools with a unified format that brings together the most relevant methods to analyze the Ig variable domains. Of course, this proposition can be applied to other protein families. For the immune system it is currently being planned to extend it to TCR and MHC molecules.

4.1. Availability of the program

The executable programs and data bases can be obtained by sending an e-mail message to the following address: salazar@redvax 1.dgsca.unam.mx.

Acknowledgements

The authors wish to thank Hector Ceceña for very helpful comments and carefully reading the manuscript. This work was partially supported by DGAPA, grant No. IN-206093. EVM was supported by a grant of CONACyT.

References

Almagro, J.C., Zenteno-Cuevas, R., Vargas-Madrazo, E. and Lara-Ochoa, F., 1994, Variability analysis of the T-cell receptors using three variability indexes. (Submitted).

Bell, G.I, 1990, The human genome: An introduction, in: Computers and DNA, Bell, G.I. and Marr T.G. (eds.) Santa Fe Insutute studies in the Sciences of Complexity, Vol. 7 (Addison Wesley) pp. 3–11.

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

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

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 (Lond.) 342, 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 VH segments. J. Mol. Biol. 227, 799–817.

Dayhoff, M.O. and Eck, R.V., 1966. Atlas of Protein Sequence and Structure, Vol.2 (NBRF Press, Silver Spring).

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

Jores, R., Alzari, P.M. and Mro, T., 1990. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acids diversity. Proc. Natl. Acad. Sci. USA. 87, 9138–9142.

Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. and Foeller, C., 1991, Sequences of Proteins of Immunological Interest. 5th edn. (U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, NIH, Bethesda, MD) Publication No. 91–3242.

Lara-Ochoa, F., Vargas-Madrazo, E., Jiménez-Montaño, M.A. and Almagro, J.C., 1994, Patterns in the complementarity determining regions of immunoglobulin. BioSystem 32, 1–9.

Mian, I.S., Bradwell, A.R. and Olson, A.J., 1991. Structure, function and properties of antibody binding sites. J. Mol. Biol. 217, 133–151.

Padlan, E.A. and Kabat, E.A., 1991, Modeling of antibody combining sites. Methods Enzymol. 202-B, 3–21.

Padlan, E.A., 1994. The antibody molecule. Mol. Immunol. (in press).

Pascual, V. and Capra, D., 1991, Human immunoglobulin heavy-chain variable region genes: Organization, polymorphism, and expression. Adv. Immunol. 49, 1–74.

Shenkin, P.S., Erman, B. and Mastrandrea, L.D., 1991, Information-theoretical entropy as a measure of sequence variability. Proteins 11, 297–313.

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

Vargas-Madrazo, E., Ochoa-Lara, F. and Jiménez-Montaño, M., 1994a. A skew distribution of amino acids at recognition sites of the hypervariable regions of immunoglobulin. J. Mol. Evol. 38, 100–104.

Vargas-Madrazo, E., Almagro, J.C. and Lara-Ochoa, F., 1994b, Structural repertoire in VH pseudogenes of immunoglobulin: Comparison with human germline genes and human functional sequences (submitted).

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.

Wysocki, L.J. and Gefter, M.L., 1989. Gene conversion and the generation of antibody diversity. Annu. Rev. Biochem. 58, 509–531.

Young, M.N. and Oomen, R.P., 1992, Analysis of sequence variation among legume lectins: A ring of hypervariable residues forms the perimeter of the carbohydrate binding site. J. Mol. Biol. 228, 924–934.

Zuckerkandl, E., 1976, Evolutionary processes and evolutionary noise at molecular level. J. Mol. Evol. 7, 167–183.

JMB—MS 305 Cust Ref No CAM 230/94 [SGM.]

J. Mol. Biol. (1994) 245, 000–000

(1995) 246, 84-86

JMB

Structural Repertoire in V_H Pseudogenes of Immunoglobulins: Comparison with Human Germline Genes and Human Amino Acid Sequences

Enrique Vargas-Madrazo[1]*, Juan C. Almagro[2] and Francisco Lara-Ochoa[2]

[1]*Instituto de Investicaones Biológicas, Universidad Veracruzana, Xalapa Veracruz México.* Av. Dos S/N Km 7.5 Carr. Ver. 91000 e-mail evargas@
[2]*Instituto de Química Universidad Nacional Autónoma de México Circuito Exterior, Ciudad Universitaria, C.P. 04510 México, D.F.*

In the pool of human immunoglobulin V_H gene segments, pseudogenes amount to roughly 30% of the total number of genes. Some of them are highly conserved among unrelated individuals. These facts suggest a possible functional role for pseudogenes in the human immune response diversity. This paper intends to provide additional information about the structure of V_H pseudogene sequences to evaluate the possible role of pseudogenes in the immune response.

Mutations capable of altering framework stability in human V_H pseudogenes were analyzed. Results indicate that V_H pseudogenes are about 14 times as divergent as human V_H functional germline genes on the one hand, and four times as divergent in the case of human V_H amino acid sequences on the other. The high number of disruptive mutations in pseudogenes is an expected result because of the lack of functionality of these genes. In the second part of the work we analyze whether or not the same takes place in the positions that determine the existence of canonical structures in the hypervariable loops in V_H pseudogenes. An extension of such analysis is applied to all species with reported V_H pseudogenes. In contrast with results concerning framework positions, 69% of known human V_H pseudogenes have canonical structures in the first hypervariable loop, while 48% do so in the second one. Comparison of these results with those found in human V_H functional germline genes and human V_H amino acid sequences shows that in the former as many as 100% and in the latter 96% have canonical structures. In V_H amino acid sequences the result is similar to pseudogenes for H1. For H2, such value lies between the percentage of germline genes (96%) and the percentage of pseudogenes (48%). The possible significance of the existence of canonical structures in the hypervariable loops of V_H pseudogenes is discussed.

Keywords: canonical structures; canonical structure classes; gene conversion; antigen-binding site; V_H locus

*Corresponding author

Introduction

Antibody molecules are highly antigen-specific receptors of the immune system. Antigen–antibody interaction involves the former variable domains, each composed of a two β-sheet framework (Amzel & Poljak, 1979). The antigen binding site is composed of six hypervariable loops; three from the variable light domain (V_L) and three from the variable heavy domain (V_H) denoted L1, L2, L3 and H1, H2, H3, respectively (Wu & Kabat, 1970; Poljak *et al.*, 1973). Genetically, L1 and L2 are produced by the V_L gene, while L3 is produced by the recombination of an additional gene segment, J_L. In a similar way, H1 and H2 are produced by the V_H gene, and H3 is a result of the recombination of two additional gene segments, D and J_H (Tonegawa, 1983).

In addition to functional germline genes, most species retain a pool of V_H pseudogenes estimated to be approximately 30% of the total number of V_H genes (Kodeira *et al.*, 1986; Hsu *et al.*, 1989). In the human V_H3 family, half of the genes are pseudogenes (Pascual & Capra, 1991). Polymorphism studies of human V_H pseudogenes report high sequence conservation among unrelated individuals (Pascual & Capra, 1991; Tomlinson *et al.*,

† *Abbreviation used:* Ig, immunoglobulin.

0022-2836/95/000000-00 $08.00/0

Table 1

Percentage of loops in the samples with canonical structures

	Samples					
Loop	Human pseudogenes	Human functional germline genes	Human amino acid sequences	Mouse pseudogenes	Rabbit pseudogenes	*Xenopus* pseudogenes
H1	69	100	71	86	100	100
H2	48	96	70	86	80	25

acid sequences is equal to that in pseudogenes, while for H2 the value is between that of functional germline genes (96%) and pseudogenes (48%).

Considering that the structural divergence measurement in pseudogenes is significantly larger than in the amino acid sequences, and even larger for functional germline genes, the percentage of human V_H pseudogenes with sequence patterns compatible with canonical structures is striking. For mouse pseudogenes the frequency of canonical structures is larger than the one corresponding to human, and even greater for the amino acid sequences (see Table 1). The five sequences from rabbit have canonical structure patterns in H1 and four sequences for H2. The four sequences of *Xenopus* present canonical structure in H1 and one sequence has a canonical structure for H2. The sequence of elops (Pa8a) has a canonical structure in H1 but for H2 the sequence has a sequence pattern that does not match with the canonical structure.

Canonical structure types for H1

Considering only those sequences with canonical structures in the first V_H hypervariable loop, the percentage of each canonical structure type for H1 is shown in the first half of Table 2. Canonical structure type 1 represents 94% of human pseudogenes and 100% of mouse pseudogenes, about 75% for human functional germline genes and 82% for the amino acid sequences.

It is worth mentioning that while in most cases human pseudogenes present the type 1 for H1 (only two sequences with the type 3, see Table 2), functional germline genes possess a significative proportion of types 2 and 3 of H1 (13% and 14%, respectively). Type 2 have low frequency in the amino acid sequences (3.7%). The observation that human and mouse pseudogenes, in most cases, present the type 1 for H1 contrasts with rabbit and *Xenopus* pseudogenes which have the other two canonical structure types.

Canonical structure types for H2

The second half of Table 2 reports the frequency of the canonical structure types for the second V_H hypervariable region. The frequencies of canonical structures for this loop vary from one sample to another, although the canonical structure type 4 is the least represented in all the samples.

Frequencies of canonical structure types from human sequences are similar in the functional germline genes and in the amino acid sequences. Human pseudogenes do not present sequences with canonical structure type 4. In mouse, the canonical structure type 2 is the most frequent. The above indicates that unlike H1, where type 1 is the most

Table 2

Percentage of occurrence in the samples of different canonical structures

		Samples							
				Pseudogenes					
Loop	Canonical† structure type	Human functional germline genes	Human amino acid sequences	Human	Mouse	Rabbit	*Xenopus*	Elops	Shark
		(77)‡	(161)	(35)	(18)	(5)	(4)	(1)	—
	1	75	82	94	100	80	50	100	—
H1	2	12	4	0	0	20	0	0	—
	3	13	14	6	0	0	50	0	—
		(74)	(140)	(25)	(18)	(4)	(1)	—	—
	1	39	28	28	11	25	0	—	—
H2	2	18	18	16	78	25	0	—	—
	3	39	48	56	11	50	100	—	—
	4	4	[illegible]	0	0	0	0	—	—

† Canonical structures described by Chothia et al. (1989).
‡ Number of sequences analyzed.

representative type of canonical structure and the most frequent in human pseudogenes. H2 displays a heterogeneous distribution of canonical structure types.

Canonical structure classes

The frequency of canonical structure classes in the three sequence samples is shown in Table 3. As previously discussed, H1 type 1 occurs in a large number of sequences in the three sequence data bases (see Table 2). H2 types are rather variable among different samples. Thus, most of the combinations occur only between the canonical structure type 1 for H1 and any one of the other canonical structure types of H2. In pseudogenes this rule holds for sequences from all species, with the sole exception of one human sequence (5% of the 3-1 class in Table 3) and one rabbit sequence.

For human functional germline genes and the amino acid sequences, the 2-1 and 3-1 classes occur with a relatively high percentage (see Table 3). Correlation between these two samples is good for all classes.

Discussion

The finding of a vast amount of pseudogenes within the human V_H and V_K genes of other species studied have posed several questions about its possible function in the immune response (Wysocki & Gefter, 1989; Pascual & Capra, 1991). It has been proposed that the number of mutations accumulated in pseudogenes with respect to functional germline genes is principally a function of: (1) the time elapsed since inactivation of the gene has occurred (Blankenstein *et al.*, 1987); and (2) possible functional restrictions acting in the new state of the gene (Pascual & Capra, 1991; Tomlinson *et al.*, 1992; McCormack *et al.*, 1993). In pseudogenes of all the species analyzed (principally from human), a high divergence was observed in residues determining the Ig-fold when compared with the functional germline genes and amino acid sequences.

Contrasting with the fact that pseudogenes have an average of seven destructive mutations per sequence outside the antigen binding site, the finding that 69% of its sequences have canonical structures of H1 and 48% of its sequences have canonical structures for H2 is quite surprising. In the case of H2, having fewer sequences presenting canonical structures might be intimately related to the fact that this hypervariable loop is, in general, less conserved (Chothia *et al.*, 1992). In addition, the amino acid sequences have canonical structure frequencies similar to those in pseudogenes. The above combined with the results that show that pseudogenes from mouse and other species (as distant from human as *Xenopus* and chips) also present a large proportion of canonical structures reinforce the main result of this report, the presence of canonical structure sequence patterns in human pseudogenes.

Pseudogenes are not only more noisy than functional germline genes and amino acid sequences, but the number of mutations is also more variable among sequences as well. This observation seems to point towards the need of analyzing if the presence of a large number of destructive mutations in the framework imply the absence of sequence patterns of canonical structures. Therefore, discarding from the analysis all those sequences with more than ten mutations (12 sequences), the percentage of human pseudogene sequences with canonical structures was calculated. The percentages obtained were 68% for H1 (69% considering all sequences, see Table 1), and 52% for H2 (48% considering all sequences). For H1, the percentage of sequence patterns decreases, while for H2 it increases. These results seem to be an indication that a strong correlation is not observed between large framework mutations and the absence of canonical structure patterns in pseudogene hypervariable loops.

Sequence studies between functional germline genes and pseudogenes have shown the existence of a strong evolutionary link between the functional germline genes and the pseudogenes belonging to the same family (Pascual & Capra, 1991). Thus, the consistency of results reported in previous sections can be tested by analyzing the correlation among canonical structures and functional germline gene and pseudogene families. The distribution of canonical structures codified by gene families was analyzed. It was found in almost all cases that the canonical structures that are encoded by the functional germline genes of some families are encoded by pseudogenes that are members of the same family (results not shown). That is to say that a close correlation in the codification of canonical structures and the sequence families between functional germline genes and pseudogenes exists.

It has been proposed that the sequence conservation observed in pseudogenes is due to germline gene conversion (Haino *et al.*, 1994). As discussed previously, almost all the V_H functional germline

Table 3

Percentage of canonical structure classes

	Samples			
Canonical structure class	Human pseudogenes (19)†	Human functional germline genes (74)	Human amino acid sequences (116)	Mouse pseudogenes (15)
1-1	36.0	14.7	13.8	6.7
1-2	11.0	17.3	19.8	86.7
1-3	69.0	38.7	47.4	6.7
1-4	13.3	4.0	0.9	0.0
2-1	0.0	12.0	5.2	0.0
2-2	0.0	0.0	0.0	0.0
2-3	0.0	0.0	0.0	0.0
2-4	0.0	0.0	0.0	0.0
3-1	5.0	12.0	11.2	0.0
3-2	0.0	0.0	0.0	0.0
3-3	0.0	0.0	0.9	0.0
3-4	0.0	0.0	0.0	0.0

† Number of sequences analyzed

genes present canonical structures (Chothia *et al.*, 1992). Considering this, it could be proposed that the presence of canonical structures in V_H pseudogenes reported here is due to germline gene conversion events between functional germline genes and pseudogenes. At present no firm evidence has been presented about the presence of somatic gene conversion in human Ig genes (Berek, 1993). However, if this mechanism really contributes to the diversity of the functional repertoire in humans, then the presence of canonical structures at the hypervariable loops of the donor pseudogenes could be convenient for maintaining the restrictions that have been established that work within the structural repertoire generated by the functional germline genes (Chothia *et al.*, 1992).

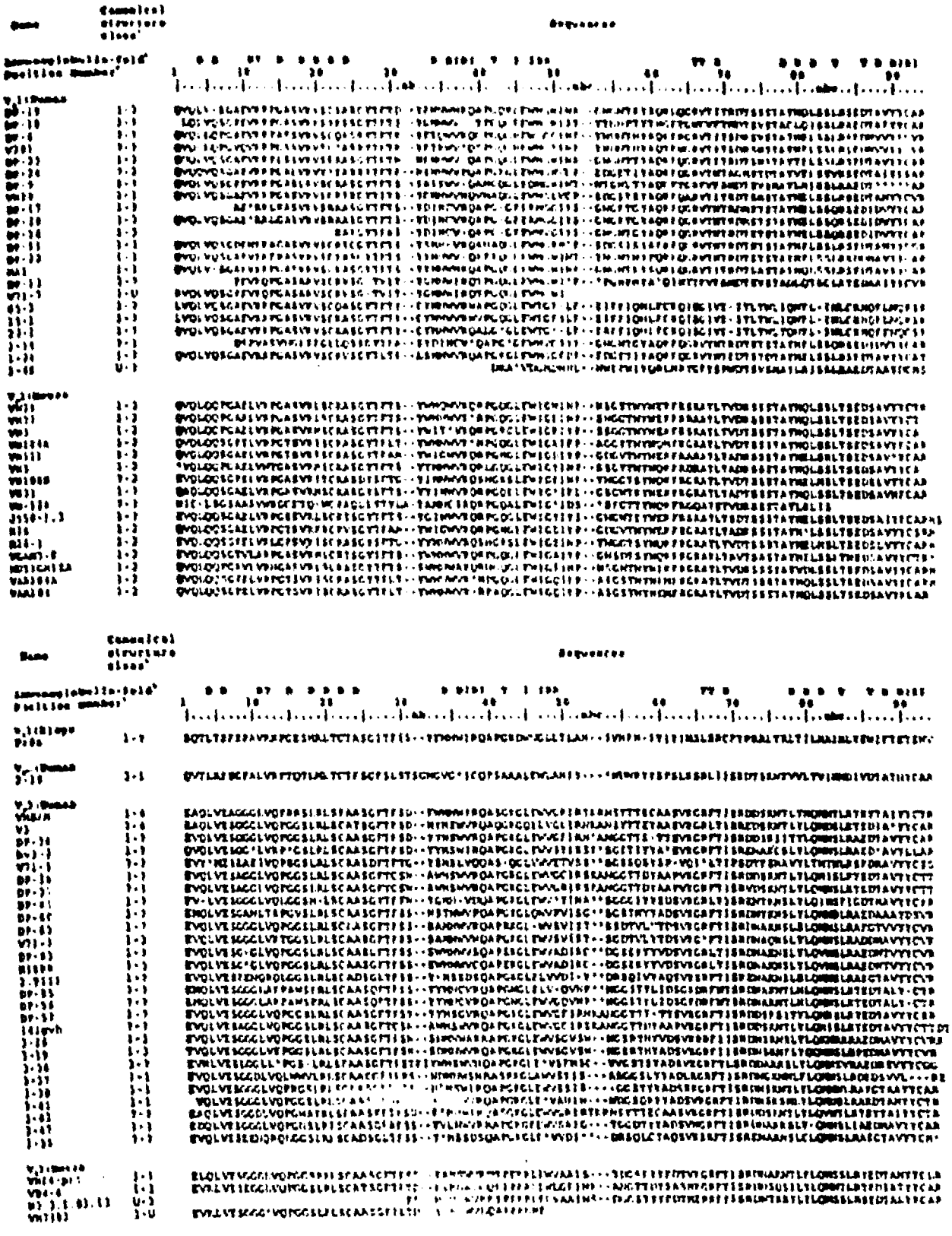


Figure 1. *(continued overleaf)*

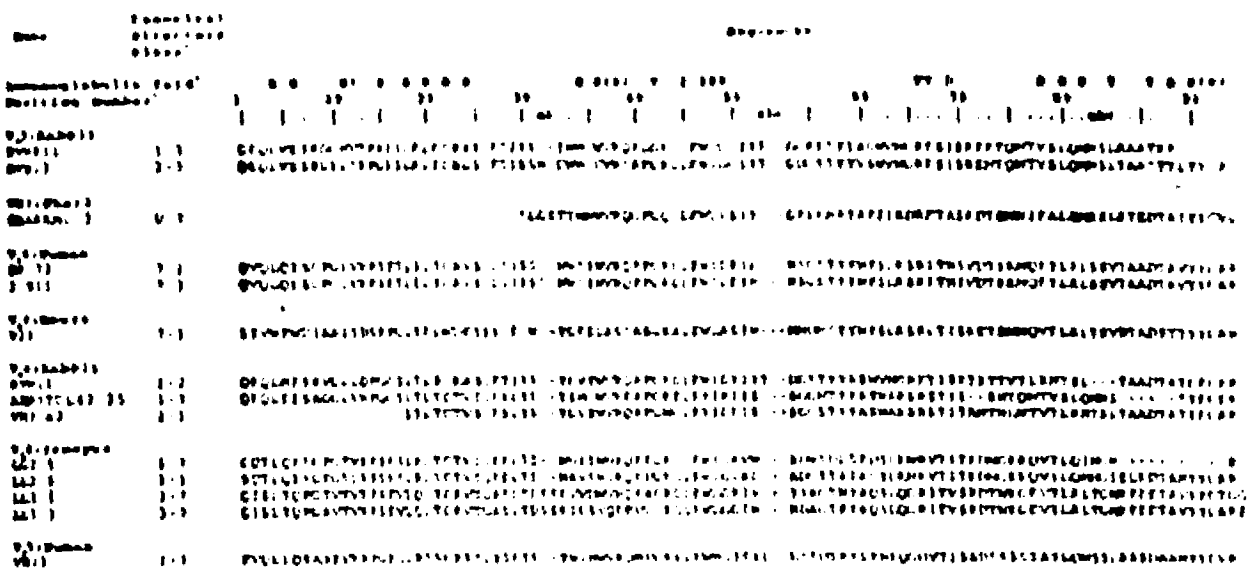


Figure 1. Pseudogenes multiple sequence alignment. Notation for canonical structure classes. The first number in the canonical structure class corresponds to the H1 canonical structure type while the second one corresponds to the H2 canonical structure type. ? Indicates that the loop has an uncertain conformation, U indicates an undetermined sequence for this loop. [b]Positions that are primary responsible for the immunoglobulin fold conserved features (B, buried in β-sheets, T, in turns, I, inter-domains (Chothia *et al.*, 1988)). [c]Numbering as in Chothia *et al.* (1992). [d] Indicates a stop codon. Insertions and deletions that produce frameshift changes in the amino acid sequence were eliminated to obtain the most correct Ig-like sequences. The sequence V71-7, from human pseudogenes has, in H1, the length and sequence pattern corresponding to canonical structure 1 but, lacking the second half of the sequence, residue 94, being important for conformation, it is not known. The identity of this sequence with respect to other pseudogene sequences available were calculated. A sequence having 82% identity with V71-7 was found. Such a sequence has an Arg at position 94. Thus, we conclude that V71-7 has the canonical structure type 1 for H1.

Materials and Methods

Ig V_H Sequences

In order to study the canonical structures in Ig V_H pseudogene sequences, a multiple alignment was formed for human pseudogene sequences and all pseudogene sequences reported for the other species in the Genbank data base release 80 (Figure 1). The main source of *Homo sapiens* (human) pseudogene sequences was the directory of pseudogene sequences reported by Tomlinson *et al.* (1992). Human sequences obtained by searching the nucleotide sequence data base were compared with the above-mentioned directory. Those sequences that have 95% or more identity with one or more sequences of the directory of Tomlinson *et al.* (1992) were discarded from the analysis. Incomplete sequences that do not encode at least one of the two hypervariable loops (H1 and H2) were discarded. We found 14 additional V_H sequences out of a total of 53 sequences. Concerning other species, the number of pseudogenes sequences found were: *Mus musculus* (mouse), 21; *Oryctolagus cuniculus* (rabbit), 5; *Xenopus laevis* (*Xenopus*), 4; *Heterodontus francisci* (shark), 1; and *Elops saurus* (elops), 1.

Two additional multiple alignments were formed to compare the results of human pseudogenes, one for human V_H functional germline genes, and the other one for human V_H amino acid sequences. The multiple alignment of human functional germline genes was obtained from Tomlinson *et al.* (1992). The multiple alignment of human V_H amino acid sequences was constructed considering all human V_H mature sequences reported by Kabat *et al.* (1991), thus making a total of 325 sequences.

Classification of pseudogenes sequences by residue identities

Sequences of V_H functional germline genes have been classified into six established families (Pascual & Capra, 1991). Tomlinson *et al.* (1992) have assigned human pseudogenes to the functional germline gene families. This classification was implemented according to the amino acid identity of each pseudogene sequence with members from different functional germline gene families. Applying this method, the 14 additional human pseudogene sequences and pseudogenes from the other species were assigned to the functional germline gene families established by Tomlinson *et al.* (1992) (Figure 1).

Structural divergence measurement

Analysis of Ig of known three-dimensional structure has shown that V_L and V_H domains contain a largely conserved β-sheet framework (Amzel & Poljak, 1979). Residues mainly responsible for the structural conserved features in Ig-variable domains have been proposed by Chothia *et al.* (1988) based on structural considerations (see top of Figure 1). For each framework position, several amino acid residues appear normally in the V_H domains of Ig (Chothia *et al.*, 1988). For the purposes of the present analysis those positions not having any of these amino acid residues were considered as defective or mutated. This procedure was carried out for every sequence in each sample. The average number of defective positions per sequence in each sample was calculated by the standard method of dividing the total number of defective positions by the total number of sequences in each sample. Such a description of sequence structural divergence might be considered as an

oversimplification of the substitution process since it is a well-established fact that proteins are capable, at least partially of accommodating the effects of mutations (Chothia *et al.*, 1985). In this analysis, however, the purpose was to estimate a divergence measurement of each sample with respect to some standard of reference.

Determination of canonical structures and canonical structure classes

Canonical structures for H1 and H2 in pseudogene sequences, functional germline genes, and amino acid sequences were determined on each multiple alignment by the sequence patterns as defined by Chothia *et al.* (Chothia & Lesk, 1987; Chothia *et al.*, 1989, 1992; Tramontano *et al.*, 1990). In order to search for these patterns in the three multiple sequence alignments, all conventions of numbering, placement of insertions, length and localization in sequences of hypervariable loops are utilized as reported by Chothia *et al.* (1992).

Canonical structure classes as defined by Chothia *et al.* (1992) are the different combinations of canonical structure types for H1 and H2. These classes are numbered in the form N–M, N being the number of the H1 canonical structure and M the corresponding one for the H2 structure.

Acknowledgements

We gratefully acknowledge Dr Mario L. Amzel and Dr Carlos Larralde for the critical revision of the manuscript. E. V. thanks V. Hernández-Mendiola, R. López-Hernández, M. Ramírez-Benítez and P. Reidy for technical assistance. We also thank Hector Ceceña for useful comments in the manuscript preparation. E. V. was supported by CONACyT, SNI and FOMES, J. C. A. was supported by DGAPA grant no. IN-206093.

References

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

Becker, R. S. & Knight, K. L. (1990). Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits. *Cell*, **63**, 987–997.

Berek, C. (1993). Somatic mutation and memory. *Curr. Opin. Immunol.* **5**, 218–222.

Blankenstein, T., Bonhomme, F. & Krawinkel, U. (1987). Evolution of pseudogenes in the immunoglobulin V_H gene family of the mouse. *Immunogenetics*, **26**, 237–248.

Brünger, A. T., Leahy, D. J., Hynes, T. R. & Fox, R. O. (1991). 2.9 Å resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment with bound hapten. *J. Mol. Biol.* **221**, 239–256.

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

Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). Domain association in immunoglobulin molecules. The packing of variable domains. *J. Mol. Biol.* **186**, 651–663.

Chothia, C., Boswell, R. & Lesk, A. M. (1988). The outline structure of the T cell receptor. *EMBO J.* **7**, 3745–3755.

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 immunoglobulin hypervariable regions. *Nature (London)*, **342**, 877–883.

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

Fan, Z. C., Shan, L., Guddat, L. W., He, X. M., Gray, W. R., Rasoo, R. L. & Edmundson, A. B. (1992). Three-dimensional structure of an Fv from an human IgM immunoglobulin. *J. Mol. Biol.* **228**, 188–207.

Fischmann, T. O., Bentley, G. A., Bhat, T. N., Boulot, G., Mariuzza, R. A., Phillips, S. E. V., Tello, D. & Poljak, R. J. (1991). Crystallographic refinement of the three-dimensional structure of the Fab D1.3 lysozyme complex at 2.5 Å resolution. *J. Biol. Chem.* **266**, 12915–12920.

Garcia, K. C., Desiderio, S. V., Ronco, P. M., Verroust, P. J. & Amzel, L. M. (1992). Recognition of angiotensin II antibodies at different levels of a idiotypic network are superimposable. *Science*, **257**, 528–531.

Hamo, M., Hayasida, H., Miyata, T., Shin, E. K., Matsuda, F., Nagaoka, H., Matsumura, R., Takaishi, S., Fukita, Y., Fujikura, J. & Honjo, T. (1994). Comparison and evolution of human immunoglobulin V_H segments located in the 3' 0.8-Megabase region. *J. Biol. Chem.* **269**, 2619–2626.

Hsu, E., Schwager, J. & Ali, F. W. (1989). Evolution of immunoglobulin genes: V_H families in the amphibian *Xenopus*. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8010–8014.

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

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

McCormack, W. T., Hurley, E. A. & Thompson, C. B. (1993). Germline maintenance of the pseudogene donor pool for somatic immunoglobulin gene conversion in chickens. *Mol. Cell. Biol.* **13**, 821–830.

Pascual, V. & Capra, D. (1991). Human immunoglobulin heavy-chain variable region genes: organization, polymorphism, and expression. *Adv. Immunol.* **49**, 1–74.

Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. & 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.

Reynaud, C.-A., Dahan, A., Anquez, V. & Weill, J.-C. (1989). Somatic hyperconversion diversifies the single V_H gene of the chicken with a high incidence in the D region. *Cell*, **59**, 171–183.

Rose, D. R., Przybylska, M., To, R. J., Kayden, C. S., Oomen, R. P., Vorberg, E., Young, N. M. & Bundle, D. R. (1993). Crystal structure to 2.45 Å resolution of a monoclonal Fab specific for the *Brucella* A cell wall polysaccharide antigen. *Protein Sci.* **2**, 1100–1113.

Strong, R. K., Campbell, R., Rose, D. R., Petsko, G. A., Sharon, J. & 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, 3739–3748.

Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline V_H segments reveals 50 groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* 227, 776–798.

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

Torno, J., Stadler, E., Stern, T., Auer, H., Kanzler, O., Betzel, C., Blaas, D. & 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. & Lesk, A. M. (1990). Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the V_H domains of immunoglobulins. *J. Mol. Biol.* 215, 175–182.

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 complementarity. *J. Exp. Med.* 132, 211–250.

Wysocki, L. J. & Gefter, M. L. (1989). Gene conversion and the generation of antibody diversity. *Annu. Rev. Biochem.* 58, 509–531.

Edited by J. Karn

(Received 20 May 1994, accepted in revised form 4 November 1994)

Canonical Structure Repertoire of the Antigen-Binding Site of Immunoglobulins Suggests Strong Geometrical Restrictions Associated to the Mechanism of Immune Recognition

Enrique Vargas-Madrazo[1*], Francisco Lara-Ochoa[2], Eduardo Horjales[2], Juan Carlos Almagro[2]

[1] Instituto de Investigaciones Biológicas,
Universidad Veracruzana,
Av. dos Vistas s/n, Carr. a Veracruz Km. 2.5,
Xalapa, Ver. 91000 México.
Telephone number: (28) 12-5757
Fax number: (28) 12-5757
E-mail: evargas@uv4.invest.uv.mx.

[2] Instituto de Química,
Universidad Nacional Autónoma de México,
Circuito Exterior, Ciudad Universitaria,
C.P. 04510, México, D.F.

*** Corresponding author:**

Total number of pages: 29 (including two Tables and three Figures).

Running title: Canonical Structure Repertoire of the Antigen-Binding Site.

Keywords: canonical structure, canonical structure classes, molecular recognition, H3, antigen-antibody complexes.

Abbreviations: Ig: immunoglobulin; V_L: Variable light domain; V_H: Variable heavy domain; PDB: Brookhaven Protein Data Bank.

Abstract

Is the structural repertoire of immunoglobulins free to adopt an almost infinite number of conformations or does it take advantage of only a few conformations to build the diversity of the immune response? In this paper we study this question by applying the canonical structure model to the characterization of the structural repertoire of the immunoglobulin. The results show that, from 300 possible different canonical structures classes (combinations of canonical structures in five of the six hypervariable loops that forms the antigen-binding site), merely 10 combinations comprise about the 87% of the roughly 400 sequences analyzed. Additionally, it was observed that H2 and L1 are the main determinants of the structural variability of the antigen-binding site. The possible functional significance of these results was correlated to a classification of the antibodies in terms of their gross specificity. Two sets of canonical structure classes were found, one of them with preference for some types of antigens like proteins, polysaccharides or haptens, and the other with multi-specific binding capabilities. Analysis of antibodies of known three-dimensional structure show that the specific classes each present a particular overall geometry of the antigen-binding site. In contrast, in at least one multi-specific class, it was observed that changes in the general geometry of the antigen-binding site are produced by different conformations of H3. The implications of these results for the molecular recognition process mediated by immunoglobulin are hereby discussed.

1. Introduction

Antibody molecules are highly antigen-specific receptors of the immune system. Antigen-antibody interaction involves the variable domains each composed of a two ß-sheet framework (Amzel & Poljak, 1979). The antigen-binding site is composed of six hypervariable loops; three from V_H and three from V_L denoted H1, H2, H3 and L1, L2, L3 respectively (Wu & Kabat, 1970; Poljak *et al.*, 1973). Genetically, L1 and L2 are encoded by the V_L gene, while L3 is produced by the recombination of an additional gene segment, J_L. In a similar way, H1 and H2 are encoded by the V_H gene, and H3 is a result of the recombination of two additional gene segments, D and J_H (Tonegawa, 1983).

Analysis of antibodies of known three-dimensional structure has revealed a small number of main-chain conformations or canonical structures for H1 and H2 as well as for L1, L2, L3 (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Tramontano *et al.*, 1990). Canonical structures in five out of six hypervariable loops imply that only a few main-chain

conformations are present in a large set of antibody molecules with different loop sequences. A canonical structure is determined by (1) the loop size and (2) by the presence of certain residues at key positions, in both the loop and the framework regions (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Tramontano *et al.*, 1990). On basis of this pattern that relates the sequence and the three-dimensional structure of the hypervariable loops, analysis of functional germline genes (Chothia *et al.*, 1992; Cox *et al.*, 1994; Williams & Winter, 1993), pseudogenes (Vargas-Madrazo *et al.*, 1995a) and mature amino acid sequences (Chothia *et al.*, 1989; Vargas-Madrazo *et al.*, 1995b) show canonical structures in almost all the Ig sequences.

From a functional point of view, it has been suggested that there are some geometrical features of the antigen-binding site that correlate with the type of antigen recognized (Davies *et al.*, 1990; Wilson *et al.*, 1991; Wilson & Stanfield, 1993). For example, antibodies specific for small molecules have concave antigen-binding sites, frequently found as a deep "pocket" or "groove", while antibodies that bind larger molecules, such as proteins, have flat antigen-binding sites (Rees & de la Paz 1986; Bolger & Sherman, 1991; Wilson *et al.*, 1991). Following this suggestion and the fact that canonical structures are present in almost all the Ig sequences, one can expect that certain combinations of canonical structures correlate with the type of recognized antigen. If some correlation between the general architecture of the antigen-binding site and antibody function is found, this will provide insight into the general mechanism of the molecular recognition mediated by Igs. In addition, heuristic schemes based on this correlation could be useful for more rational *de novo* design of antibodies of desired specificity.

In order to study the above proposition, the concept of canonical structure classes as the combination of canonical structures in H1 and H2 proposed by Chothia *et al.* (1992), is extended in this paper for the combinations of canonical structures in H1, H2, L1, L2 and L3. The concept of canonical structure classes allows us to characterize the structural repertoire of Igs. The possible correlation between different canonical structure classes and the recognized antigen was then studied by classifying the antigens in term of their gross chemical and biochemical nature, for example, protein, polysaccharide, hapten, etc.

Concerning to H3, it has been proposed that this hypervariable loop is the main contributor to the specificity of the antigen-binding site (Kabat *et al.*, 1991a; Wu *et al.*, 1993; Wilson & Stanfield, 1993; Wilson & Stanfield, 1994a; Padlan 1994). On the other hand, for H3 no canonical structure has been described at present (Chothia & Lesk, 1987;

Wilson & Stanfield, 1993). Thus, in order to characterize the role of this hypervariable loop in relation with the Ig specificity, we analyzed the length distribution of H3 in relation to the gross specificities above mentioned.

Finally, to find out the general geometric features of the antigen-binding site in relation to the canonical structure classes and the role played for H3 in the Ig specificity, the Igs of known three-dimensional structures were examined. Based on the results obtained in the analysis here presented, a general mechanism for the molecular recognition mediated by Igs is proposed.

2. Results

a) Distribution of canonical structure classes in the Ig sequences. For five of the six hypervariable loops that form the antigen-binding site, several canonical structure types have been described at present (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Chothia *et al.*, 1992). Three canonical structures types have been identified for H1, four types for H2, five types for L1, one type for L2, and five types for L3 (Chothia & Lesk 1987; Chothia *et al.*, 1989; Chothia *et al.*, 1992). From these canonical structure types, the total number of expected canonical structures classes for the Igs can be computed. Thus, the expected structural repertoire of Igs should comprise 300 canonical structure classes.

In order to study the canonical structure classes occurring in the functional domains of Igs, the V_H and V_L sequences compiled in the Kabat's Data Base (Kabat *et al.*, 1991b) were examined. From approximately 4000 sequences reported in the Kabat's Data Base for each V domain, not all the sequences have the complete information to be analyzed. That is, the sequences should fulfil three main criteria to be selected for the analysis. First, its should have V_H and V_L sequences for the same antibody. Second, the sequences should have the patterns corresponding to canonical structures for H1, H2, L1, L2 and L3. Third, the sequences should have known specificity. Application of these criteria to the total amount of sequences in the Kabat's Data Bases gives a set of 381 sequences (see Materials and Methods section for details).

Analysis of the 381 useful sequences shows that, from the 300 expected canonical structure classes, only 29 classes are found. If only are considered those classes which have over the 2% of the total number of the sequences analyzed the result is more surprising. Thus, ten classes (see first and second columns of Table 1) comprise the 86.9% of the sample. That is, a mere 3.3% of the expected canonical structure classes represents roughly

the 87% of the Ig sequences useful for the analysis. These results indicate that the observed structural repertoire of Igs is restricted to use preferentially a small number of combinations of canonical structures in five of the six hypervariable loops that form the antigen-binding site.

Another significative feature of the observed structural repertoire of Igs raising for the above results is that, H1, L2 and L3 always appear with the canonical structure type 1 (see first column of Table 2). This means that, H1, L2 and L3, do not contribute to the variation of the most frequent classes in the Igs, only H2 and L1 change from one class to another.

b) Canonical structure classes and gross specificities. Based on the previous results, we shall now analyze the specificities of the ten classes with highest frequency. With this aim, a classification in terms of gross chemical and biochemical nature of the recognized antigen was introduced (see Materials and Methods section). The results are reported in the columns three to eight of the Table 1. It should be noted that the specificities in these columns are arranged in decreasing order of antigen size, from protein to hapten. At a first approximation, this scale represents the convexity degree of the antigen surface.

The classification of the antigens based in chemical and biochemical criteria here proposed oversimplifies the complexity of the epitope structure. However, with this simple procedure two main groups of canonical structure classes can be found. In some classes one or two specificities appears with a high frequency, whereas in other classes, several specificities occur with similar frequencies. Therefore, based on these two types of distribution respect to the specificity, the canonical structure classes can be classified as specific 'S' or as multi-specific 'M' (see last column of Table 1 and table footnote). Following this classification, six classes are preferentially specific for one type of antigen, while the remaining four classes are multi-specific. The former represent about half of the sequences comprised in the 10 often occurring classes. Moreover these classes cover almost all the antigen types, from protein to hapten. The multi-specific classes on the other hand, appear capable of interact with various types of antigen and represent the other half of the sequences.

c) H3 length and its contribution to the diversity of the antigen-binding site. At present no canonical structure has been identified for H3. Recently, Wu *et al.* (1993) have analyzed the distribution of H3 lengths in the Ig sequences. They observed that the H3 lengths distribution from different species resembled an aleatory process. From such study

has been suggested that the role of H3 is intimately related to the fine specificity of the Igs (Wu *et al.*, 1993).

In Table 2 (column two) the frequency distribution of all the sequences reported in the Kabat's Data Base as a function of the H3 length is reported. The columns three to eight shows the distribution of H3 lengths in terms of the sequence specificity. For all sequences as well as for the specificities the distribution is similar. Those H3 with medium length (from 7 to 13 residues) have frequencies rather large. This range of lengths represent 72.5% of the sequences. The frequency of sequences with long loops (more than 14 residues) is also considerable large, comprising the 20.4% of the sequences. This is more evident for anti-protein and anti-surface sequences. In spite of that, the frequencies as function of H3 length for each specificity resemble a Poisson distribution, like in the analysis of Wu *et al.* (1993) for species. This observation suggest that the length of H3 is independent of the gross specificity.

d) Three-dimensional analysis of the Ig structures. The results presented in Table 1 show the existence of six specific classes and four multi-specific classes. In order to find the structural features of the canonical structure class associated to the differences in specificities above described, the three-dimensional structures of the Igs reported in the PDB (Bernstein *et al.*, 1977) were analyzed.

In the case of specific classes, those that preferentially bind the largest and the smallest antigen types were selected for the structural analysis. These are the class 1-1-2-1-1, preferentially specific for proteins and classes 1-4-3-1-1 and 1-4-4-1-1 preferentially specific for haptens. The most abundant class both in sequences and in structures, class 1-2-2-1-1, was selected to be analyzed among the multi-specific classes.

In Figure 1 the superposition of the Ig structures corresponding to the class specific for proteins is presented. The surface of the antigen-binding site is relatively flat, independently of the amino acid sequence of each particular molecule, that can alter this character only slightly. This observation is consistent with Wilson *et al.* (1991) suggesting that antibodies that bind larger molecules, such as proteins, tend to have flat antigen-binding sites. In the anti-hapten classes the antigen-binding site (Figure 2) presents a very deep cleft, approximately 10 Å deep and 15 Å long. Differences in geometry from an antibody to another do not change the general geometry of this cleft. This is also consistent with the suggestion of Wilson *et al.* (1991) for the general geometry of antigen-binding site of the anti-hapten Igs. In these anti-hapten classes, the two walls of the cleft are built by the

hypervariable loops H2 and L1. All the antigens found in these crystallographic structures are bound in some region close to the bottom of the cleft. This feature should allow the antibody to maximize the contact surface between this highly concave part of the antibody and the smaller molecules. It would be very difficult to tightly bind a globular protein to this surface because the area in contact at the tips of the cleft, would not be sufficient to provide interactions capable of producing a high binding constant. Thus, it can be concluded that both anti-hapten classes analyzed have a general geometry of the antigen-binding site formed by a deep cleft, in which all the antigens are bound. In addition this geometry exclude the possibility of bind large molecules. These structural features allow to differentiate the specific class for haptens from specific class for proteins.

Concerning to the multi-specific canonical structure class selected for the study (1-2-2-1-1), four structures belonging this class are displayed in Figure 3. Specificities as diverse as for neuraminidase and for arsonate are found among these molecules. The general geometry of the antigen-binding site results rather flat (see Figures 3???? and 3b). Even though, a central hole appears in some structures (see Figure 3b). The distinctive feature of the antibodies in this class is the large conformational differences in H3 (Figure 3a). This allow to classify the H3 conformations in open and twisted (see Figure 3c). When H3 is in an open conformation, the antigen-binding site presents a central hole as in the anti-phenylarsonate Ig. In contrast, when H3 is in a twisted conformation, the antigen-binding site results in a relatively flat surface. The twisted conformation is found in two antibodies: one binds a protein (neuraminidase), and the other binds a hapten (arsonate). In the anti-arsonate Ig the hole is not completely closed, and the hapten, being very small, is bound in the hole (**cita**). Thus, H3 works as a lid, covering the arsonate. This observations indicates that the surface of the antibodies in this multi-specific class, being flat can present a central hole depending on the conformation adopted by H3. Therefore, the differences in the geometry of the antigen-binding site allows the antibodies in this class bind both, large antigens and small molecules.

3. Discussion

In order to characterize the structural repertoire of Igs, in this paper we have studied the distribution of canonical structure classes for the antigen-binding site of a set of 381 V domain sequences. From 300 expected combinations of canonical structures that should comprise the possible structural repertoire of Igs, only 29 canonical structure classes

are observed in the Ig sequences. This result suggest that the observed structural repertoire of the Igs include a very small amount of the all possible combinations of canonical structures. Moreover, within the observed structural repertoire roughly the 87% of the sequences are comprised in only 10 canonical structure classes.

An interesting feature of the 10 often occurring classes is the conservation of the canonical structure type 1 in H1, L2 and L3. The canonical structure type 1 is the shortest conformation for these hypervariable loops (Chothia & Lesk, 1987). Topologically, these three shorter and conserved loops alternate with the three other loops of variable length to form the antigen-binding site (H2, H3 and L1; see Figure 2c to notice the relative location of each hypervariable loop in the antigen-binding site). Thus, variations in the length of H3, L1 and H2 can be related with the gross specificity of the antibody. Bolger & Sherman (1991) have suggested a similar proposition for H3 and L1.

In functional terms, the 10 canonical structure classes that mainly conform the observed structural repertoire of Igs can be classified in two main groups in relation with their ability to binds different types of antigens (see Table 1). One group has preferences for certain antigen type, such as proteins or haptens, while the other group has multi-specific capabilities. Within the specific classes the length of H2 and L1 correlate with the type of recognized antigen. That is, antibodies with short loops in H2 and L1 (canonical structures class 1-1-2-1-1) are preferentially specific for large antigens (proteins). In contrast, antibodies with long loops in H2 and L1 (canonical structure classes 1-4-3-1-1 and 1-4-4-1-1) are preferentially specific for small molecules (haptens). The three-dimensional structures analyzed for these classes have acknowledge of this finding.

Regarding to H3 (the other loop of variable length) our results from the sequences studies, indicate no correlation between the length and the gross specificity at this hypervariable loop (see Table 2). In the three-dimensional structure of the specific classes analyzed it can be noticed that this loop is not essential in determining the geometrical features of the antigen-binding site (see Figures 1 and 2). Thus, the contribution of H3 to the recognition capabilities in the specific classes seems to be more related to the fine specificity of the antibody. However, in the three-dimensional structures of the multi-specific class analyzed, it was concluded that the conformation of H3 is the main responsible of modulate the geometric characteristics of the antigen-binding site (see Figure 3).

The above results taken together suggest that, differences in the role of H2, H3 and L1 generate the main groups of canonical structural classes that form the observed structural

repertoire of Igs. In a first gross level of antigen-antibody interaction, antibodies with similar geometrical characteristics in the antigen-binding site recognizes type of antigens with similar overall geometrical features. This gross level of recognition in the specific classes, is determined by the length of H2 and L1. In the multi-specific class, the first level of antigen-antibody recognition is determined mainly by the H3 conformation. In both cases, on the other hand, the further fine adjustment for a high affinity of the antigen-antibody interaction should be contributed by: (1) side-chain interaction of residues mainly at most variable positions within the hypervariable loops (Alzari *et al.*, 1990; Mian *et al.*, 1991; Padlan *et al.*, 1995); (2) conformational rearrangements of the antibody in response to the ligand binding (Wilson & Stanfield, 1993, 1994a), and (3) water molecules at the antigen-antibody interface (Mariuzza & Poljak, 1993; Wilson & Stanfield, 1994b). Then, one can speculate that the process of molecular recognition mediated by Igs would comprise two main components: a gross geometrical complementarity mainly determined by the combination between H2 and L1 or by H3, and a fine complementarity, given by the three-above-mentioned components. According to this proposition the fine complementarity of the antigen-antibody interaction, first presuppose an adequate level of shape recognition as was early suggested by Pauling (1945). The scheme of variation by H1/L1 for specific classes that comprise about half of the observed structural repertoire of Igs, is useful as a framework to explain overall features of the mechanism of immune recognition mediated by Igs. In addition, this scheme makes possible to start from certain combination of canonical structures in order to design antibodies of desired specificity.

4. Materials and Methods

a) Ig sequences. The Ig sequences were obtained from Kabat's Data Base (Kabat *et al.*, 1991b) via the internet on-line service in August 1994. The total set of sequences comprises 4565 sequences for V_H and 3377 sequences for V_L. From this sample the number of sequences considered for the computation of the distribution of canonical structure classes was 381. This number results from several restrictions that must be satisfied by the sequences to be considered in the study. These include: 1) The sequence must be complete (3118 and 2415 sequences for V_H and V_L, respectively); (2) Both domains (V_H and V_L) of a given antibody must have reported (1192 sequences); (3) The sequences must simultaneously have sequence patterns compatible with some canonical structure type (Chothia & Lesk, 1987; Tramontano *et al.*, 1990) for H1, H2, L1, L2 and L3 (415

sequences), and (5) The sequences needs to have reported its specificity (381 sequences)

(b) Determination of canonical structures and canonical structure classes. Canonical structures for H1, H2, L1, L2 and L3 in the sequences were determined following all conventions of numbering, placement of insertions, length and localization of hypervariable loops proposed by Chothia *et al.* (Chothia & Lesk, 1987; Tramontano *et al.*, 1990) Sequence management, analysis and determination of canonical structures were made using the VIR package developed in our group (Almagro *et al.*, 1995). Canonical structure classes were defined as the different combinations of canonical structure types for H1, H2, L1, L2 and L3.

c) Determination of the length of H3. The H3 is a hairpin loop that begins at position 95 and ends at position 102 according to the conventional numeration of Kabat *et al.* (1991b). There is a maximum number of 19 residues for the length for H3 (positions 95 to 102 plus positions 100a to 100k). Nevertheless, sequences with lengths of up to 26 residues have been reported (Wu *et al.*, 1993). However, the number of sequences with very long lengths is small and represents extreme cases in the sequences. We consider only those sequences that fit into the conventional numeration of Kabat that allows a maximum of 19 residues for H3 for the present analysis.

d) Definition of general specificities. A classification in terms of the chemical and biochemical nature of the antigen was introduced. The following groups of antigens were defined: (1) protein, (2) surface antigen (3) polysaccharide, (4) nucleic acid, (5) peptide, and (6) hapten. The sample of 381 V_H and V_L sequences comprises 130 different specificities, distributed as follows: 64 specificities for protein, 10 specificities for surface antigen, 4 specificities for polysaccharide, 11 specificities for nucleic acid, 10 specificities for peptide and 27 specificities for hapten. There are several considerations about this classification. PROTEIN: If an antibody recognizes a soluble or membrane-imbibed protein. The recognized protein could be an immunoglobulin itself. Obviously there could be many cases in which the determinant recognized by the antibody is a glyco or glyco-peptide region of the protein. This information is not considered in this classification. SURFACE ANTIGEN: There are many cases in which the specificity of the antibody is ambiguous, as in "anti-RBC" or "anti-E. Coli". For these cases, an ambiguous group of "anti-surface antigen" was defined. It is expected that most of these specificities were anti-protein or anti-polysaccharide. POLYSACCHARIDE: All the carbohydrate polymers recognized were considered as members of this group. NUCLEIC ACID: DNA, RNA, single chain,

double chain and any compound formed by nucleotide are so classified PI:PTIDE. Defined as a segment of a protein or a small natural polypeptide like "angiotensin". HAPTEN. Any small molecule is considered as hapten. Antibodies that recognize lipids and catalytic antibodies were included within this group.

Acknowledgments. *We thanks to V. Hernandez-Mendiola, M. Ramirez-Benites and P. Reidy for technical assistance. E. V. was supported by SNI-Conacyt and FOMES-UV. J.C.A. was supported by DGAPA grant IN-206093.*

References

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

Alzari, P. M., Spinelli, S., Mariuzza, R. A., Boulot, G., Poljak, R. J., Jarvis, J. M. & Milstein, C. (1990). Three-dimensional structure determination of an anti-2-phenyloxazolone antibody: the role of somatic mutation and heavy/light chain pairing in the maturation of an immune response. *EMBO J.* **9**, 3807-3814.

Amzel, L. M. & Poljak, R. J. (1979). Three-dimensional structure of Immunoglobulins. *Annu. Rev. Biochem.* **48**, 961-997.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimandouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bolger, M. B. & Sherman, M. A. (1991). Computer modeling of combining site structure of anti-hapten monoclonal antibodies. *Meth. Enzimol.* **203**, 21- 45.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *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**, 877-883.

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

227, 799-817

Cox, J. P. L., Tomlinson, I. A. & Winter, G. (1994) A directory of human germ-line V-kappa segments reveals a strong bias in their usage *Eur. J. Immunol.* **24**, 827-836.

Davies, D. R., Padlan, E. A. & Sheriff, S. (1990). Antibody-Antigen Complexes *Annu. Rev. Biochem.* **59**, 439-473.

Kabat, E. A. & Wu, T. T. (1991a). Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. *J. Immunol.* **147**, 1709-1719.

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

Mariuzza, R. A. & Poljak, R. J. (1993). The basics of binding: mechanisms of antigen recognition and mimicry by antibodies. *Curr. Op. Immunol.* **5**, 50-55.

Mian, I. S., Bradwell, A. R. & Olson, A. J. (1991). Structure, function and properties of antibody binding sites. *J. Molec. Biol.* **217**, 133-151.

Padlan, E. A. (1994). The anatomy of the antibody molecule. *Mol. Immunol.* **31**, 169-217.

Padlan, E. A., Abergel, C. & Tipper, J. P. (1995). Identification of specificity-determining residues in antibodies. *FASEB J.* **9**, 133-139.

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

Pauling, L. (1945). Molecular structure and Intermolecular forces. In: *The specificity of serological reactions*. (Karl Landsteiner), pp. 276-294. M. D. Dover Publications, Inc. New York.

Rees, A. R. & de la Paz, P. (1986). Investigating antibody specificity using computer graphics and protein engineering. *TIBS.* **11**, 144-148.

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

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

Vargas-Madrazo, E., Almagro, J. C. & Lara-Ochoa, F. (1995a). Structural repertoire in V_H pseudogenes of immunoglobulins: Comparison with human germline genes and human

amino acid sequences *J Mol. Biol.* **246**, 74-81.
Vargas-Madrazo, E., Almagro, J. C. & Lara-Ochoa, F. (1995b) *manuscript in preparation.*
Williams, S. C. & Winter, G. (1993) Cloning ans sequencing of human immunoglobulin V-lambda segments *Eur. J. Immunol.* **23**, 1456-1461.
Wilson, I. A., Rini, J. M., Fremont, D. H., Feiser, G. G. & 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. & Stanfield, R. L. (1993). Antibody-antigen interactions. *Curr. Op. Struct. Biol.* **3**, 113-118
Wilson, I. A. & Stanfield, R. L. (1994a). Antibody-antigen interactions: new structures and new conformational changes. *Curr. Op. Struct. Biol.* **4**, 857-867.
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 complementarity. *J. Exp. Med.* **132**, 211-250.
Wu, T. T., Johnson, G. & Kabat, E. A. (1993). Length distribution of CDR3 in antibodies. *Proteins*, **16**, 1-7.

Tables Footnotes and Figure Captions.

Table 1.

a) The classes are numbered in the form M-N-O-P-Q; M being the type of canonical structure present in H1, N the corresponding one for the H2, O the structure for L1, P the structure for L2 and Q the structure for L3. b) The percentage for each canonical structure class in respect to the total number of sequences analyzed. Only classes with more than 2% are reported. c) Specificities as defined in Materials and Methods section. The percentages are normalized to consider the differences in the number of sequences analyzed for each gross specificity. The normalization implies that to calculate the percentage of each specificity within a class, the number of sequences is multiplied by a factor obtained from dividing the total number of sequences analyzed (381) by the number of sequences found for the specificity. This normalization make the percentages independent from the number of sequences analyzed for each specificity. d) Classification of canonical structure classes as specific or multi-specific. In the analysis of the distribution of percentages of classes by general specificity the classes are grouped accord to the form of the percentage

distribution. If in the class a specificity has more than 50% of the sequences, the class is categorized as specific and a 'S' is assigned. If a specificity with more than 50% does not appear in the class, the class is categorized as multi-specific and is labeled as 'M'.

Table 2.

a) Distribution H3 length. b) The percentage for each length in respect to the total number of sequences analyzed. c) For each specificity the percentage of each H3 length with respect to the total number of sequences of the specificity is reported.

Figure.1 Ig structures within the canonical structure class 1-1-2-1-1. This is a class specific for proteins. The superposition of the structures correspond to HYHEL-10 antibody, in blue (PDB entry: 1HFM) and D1.3 antibody, in yellow (PDB entry: 1FVB). Both Igs are anti-lysozyme antibodies but against different epitopes. The solvent accessible surface (1.7 Å rolling sphere) of the hypervariable loops [Kabat *et al.*, (1991b) definition] is presented to be able to appreciate differences derived from different side chains.

Figure.2 Ig structures in the canonical structure classes 1-4-4-1-1 and 1-4-3-1-1. Both classes are specific for hapten. Superimposed structures are: 4-4-20, anti-fluorescein, in magenta (PDB entry: 4FAB), BV04-01, anti-trinucleotide, in white (PDB entry: 1CVB), MCPC603, anti-phophorilcholine, in green (PDB entry: 2MCP). (a) Ca representation. (b) Ca representation and stick and boll of the bound antigens. (c) Ca representation, bound antigens and the solvent accessible surface of hypervariable loops (see figure 1 for definitions) from BV04-01 structure. (d) Solvent accessible surfaces for all the antibody molecules.

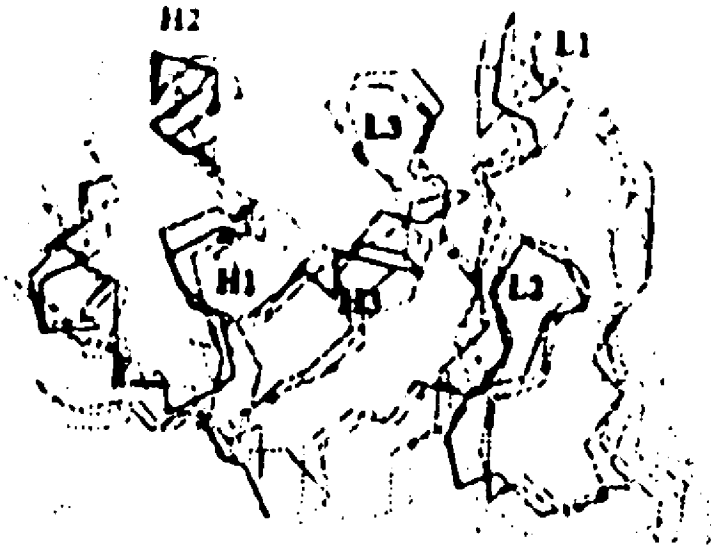
Figure.3 Ig structures in the canonical structure class 1-2-2-1-1. This is a multi-specific class. Superimposed structures are: 36-71, anti-phenylarsonate, in blue, (PDB entry: 6FAB); R19.9, anti-arsonate, in white (PDB entry: 2F19); 4D5, anti-P185^{HER2}, in magenta (PDB entry: 1FVC); NC41, anti-neuraminidase, in yellow (PDB entry: 1NCA). (a) Ca representation. (b) Molecules with twisted H3. It is possible to observe the hole in which the hapten binds in R19.9, which is partially covered by the twisted H3. (c) Comparison of two structures, one of them with H3 in twisted conformation (anti-arsonate) and the other with an open conformations (anti-phenylarsonate). Slice of the structures show the hole present in structures with H3 in open conformation.

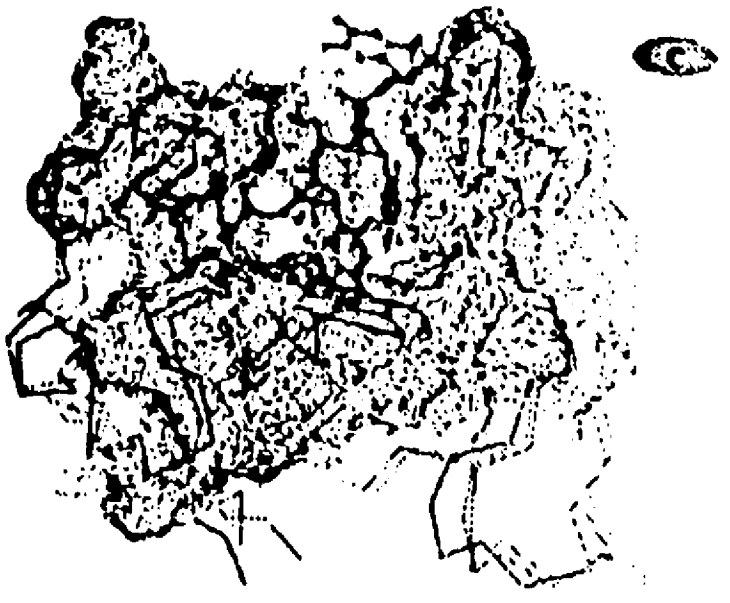
Table 1. Percentage distribution of canonical structure classes.

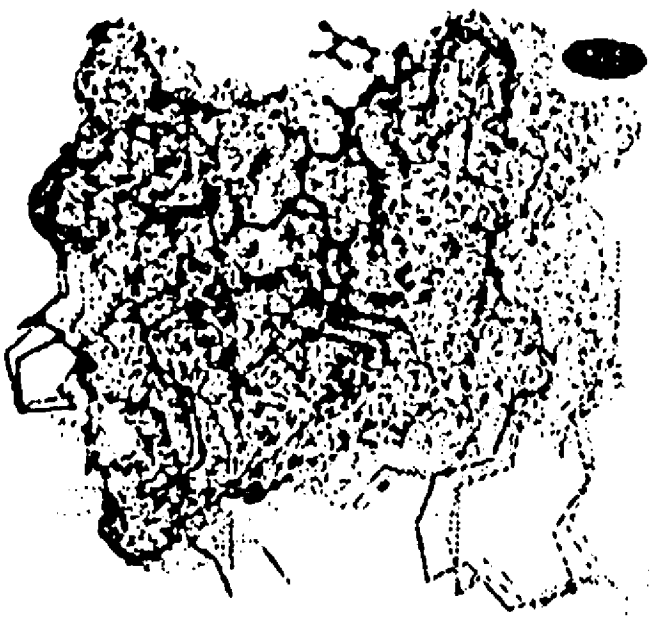
Canonical structure class[a]	Percentage of the class[b]	General specificity[c]						Type of general specificity[d]
		Protein	Surface antigen	Polysaccharide	Nucleic acid	Peptide	Hapten	
1-1-2-1-1	3.2	50[e]	0	0	25	0	19	S
1-1-4-1-1	3.7	10	0	33	13	0	44	M
1-2-1-1-1	7.1	5	5	80	5	0	4	S
1-2-2-1-1	24.5	16	44	0	18	0	23	M
1-2-3-1-1	2.9	57	43	0	0	0	0	S
1-2-4-1-1	14.2	11	4	5	24	52	4	S
1-3-2-1-1	7.9	15	26	0	31	20	8	M
1-3-4-1-1	10	43	0	14	11	25	6	M
1-4-3-1-1	6.8	11	0	0	0	0	89	S
1-4-4-1-1	4.6	4	0	0	41	0	55	S
Total of the categories	381	108	22	17	42	19	112	

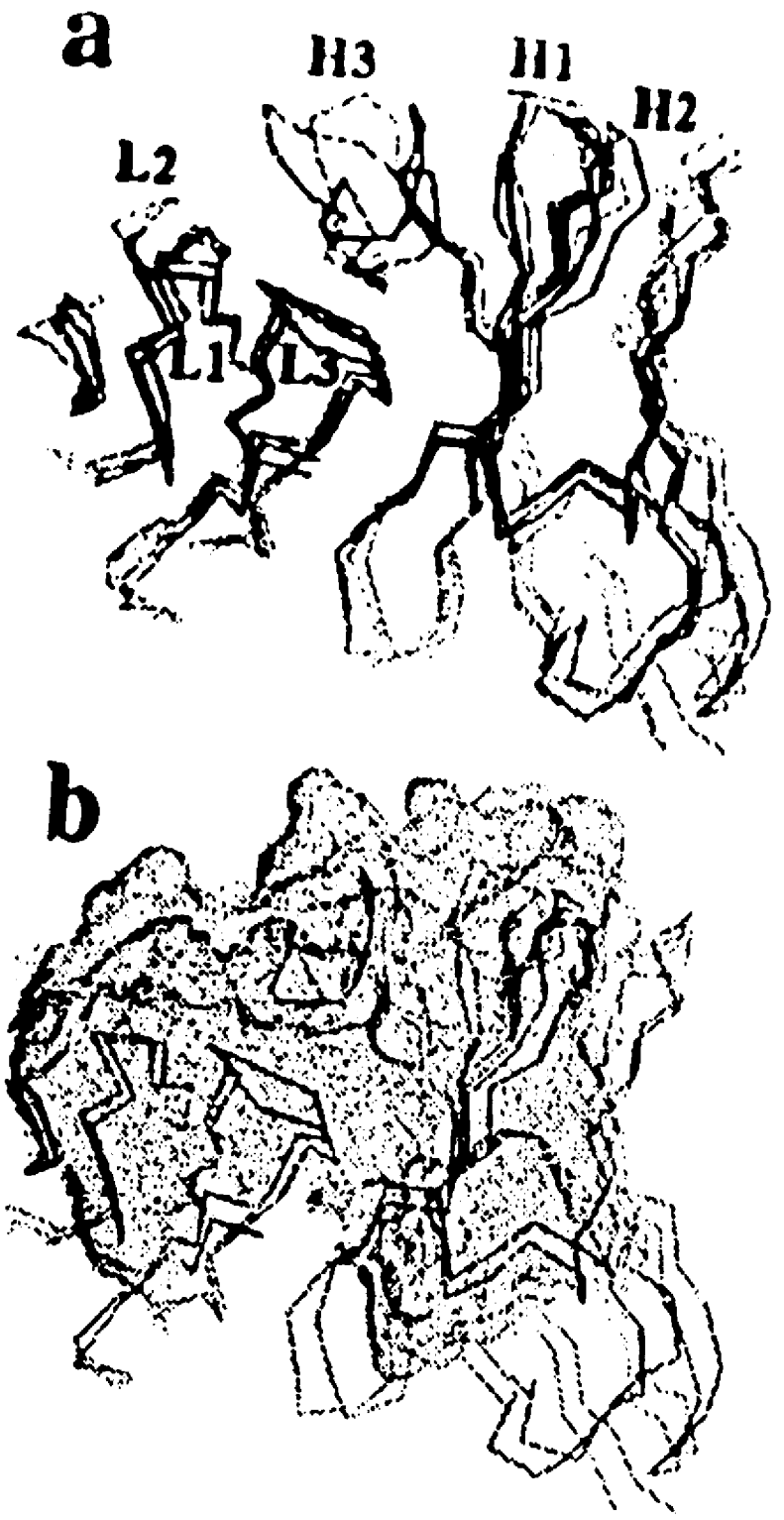
Table 2. Lenght distribution of the third hypervariable loop of the heavy chain.

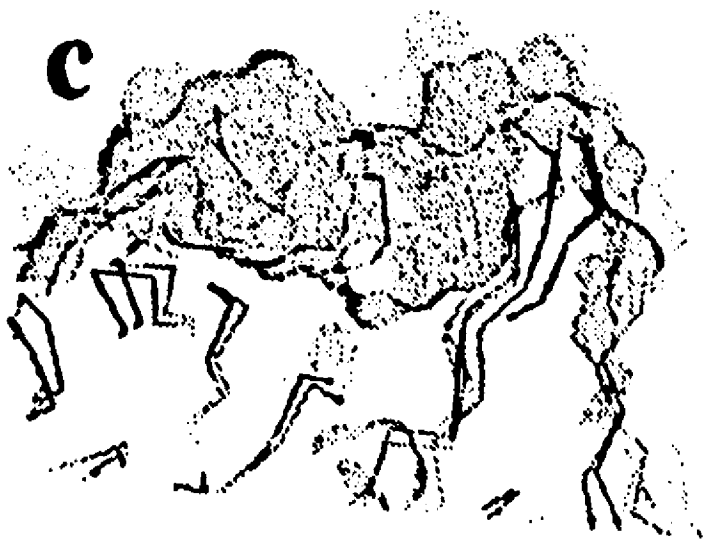
Lenght[a]	Percentage of the length[b]	General specificity[c]					
		Protein	Surface antigen	Polysac-charide	Nucleic acid	Peptide	Hapten
0	0.0	0	0	0	0	0	0
1	0.0	0	0	0	0	0	0
2	0.0	0	0	0	0	0	0
3	0.3	0	1	0	1	2	0
4	1.2	1	0	10	1	2	0
5	5.2	2	0	2	4	0	12
6	2.3	2	3	3	5	3	1
7	6.6	7	13	4	5	12	6
8	8.8	10	8	28	3	15	5
9	15.6	13	19	19	13	17	18
10	12.5	9	13	15	16	13	14
11	11.8	14	6	11	17	12	8
12	12.8	8	5	2	12	8	23
13	6.1	8	4	4	10	7	3
14	3.6	4	3	3	6	5	3
15	2.3	5	1	0	2	0	1
16	2.5	3	10	0	3	0	1
17	1.5	3	1	0	1	3	0
18	3.4	6	4	0	1	2	3
19	3.6	6	8	0	3	0	2
Size of the sample	1371	464	77	108	191	60	471











5. CONCLUSION.

5.1. Principales Resultados de la Investigación.

5.1.1. Paquete de cómputo.

La implementación del paquete de computo VIR constituye en si un resultado novedoso de la investigación, ya que este paquete permite la administración y análisis de secuencias de receptores del sistema inmune con posibilidades y criterios anteriormente no descritos en la literatura. Las principales cualidades del paquete son:

a) Ambiente gráfico amigable al usuario.

b) Actualización instantánea de la base de datos de secuencias de receptores vía Internet.

c) Brinda información detallada sobre la base de datos, incluyendo los atributos funcionales más importantes relacionados con las secuencias.

d) Construcción de sub-muestras en base a uno o varios criterios de selección.

e) Análisis de variabilidad mediante tablas de uso de aminoácidos e índices de variabilidad.

f) Análisis de patrones de secuencia de estructuras canónicas por separado y/o de acuerdo al concepto de clases propuesto por nuestro grupo.

5.1.2. Estudio del repertorio estructural en pseudogenes.

a) Utilización de criterios estructurales para determinar el estado funcional de pseudogenes del dominio variable de Igs.

b) Los pseudogenes de genes variables de la cadena pesada presentan estructuras canónicas en un porcentaje elevado muy similar al observado en secuencias de aminoácido funcionales de Igs.

c) La distribución de tipos y clases de estructuras canónicas es muy similar en genes funcionales y pseudogenes.

d) Los resultados obtenidos brindan evidencias experimentales adicionales a las descritas en la literatura respecto al posible papel que los pseudogenes de Igs pueden jugar en la generación de la diversidad de la respuesta inmune.

5.1.3. Repertorio estructural de Igs.

a) La presencia de un tipo de estructura canónica en una lazo aislada no es suficiente para determinar la especificidad general del anticuerpo

b) De las 300 posibles combinaciones de estructuras canónicas en el sitio de unión, solo 10 clases representan más del 85% de las 380 secuencias estudiadas. Tres clases representan el 50% de las secuencias.

c) Solo H2 y L1 contribuyen significativamente a la diversidad del repertorio estructural.

d) Realizando un estudio ampliado que incluye más de 600 secuencias se obtienen resultados muy similares a los de la muestra original de 380 secuencias. La base de datos de 380 secuencias incluye 110 especificidades. Estos dos resultados aumentan la confiabilidad de los resultados obtenidos.

e) Al analizar la distribución de clases respecto a los grupos de especificidad general propuestos, se encuentra que existen dos tipos de anticuerpos de acuerdo a la clases que presentan: 1) aquellos que tienen preferencia por una especificidad general, y 2) aquellos con capacidad para unirse a cualquier tipo de antigeno.

f) La determinación de la longitud de H3 por separado no presenta correlación con la especificidad general de la Ig.

g) El estudio de la combinación de H3 con las clases presentes en las cinco lazos restantes no permite observar correlación clara con la especificidad de las Igs. Esto indica que la participación de H3 en la especificidad involucra propiedades geométricas y fisico-químicas mucho más complejas que aquellas relacionadas con la longitud de H3.

h) El estudio de la estructura tridimensional de las Igs permite interpretar en términos geométricos de manera coherente los resultados obtenidos a partir de secuencias.

f) La interpretación de los resultados de estudios de secuencias y de estructura tridimensional permite proponer un modelo de reconocimiento inmune que incluye dos componentes, 1) una etapa de reconocimiento grueso donde la propiedades geométricas generales juegan el papel principal, y 2) el ajuste fino que implica la optimización de la complementaridad geométrica y química se da posteriormente a través de sustituciones de aminoácidos fundamentalmente en las posiciones hipervariables.

5.2. Conclusión.

Los resultados obtenidos en la investigación que se llevó a cabo durante el doctorado aportan numerosas evidencias acerca de la existencia de patrones en el mecanismo de Reconocimiento Molecular de las Igs. Los trabajos presentados aquí son la continuación de los estudios que sobre la distribución de aminoácidos por posición se realizaron durante la maestría (Vargas-Madrazo y col. 1992, 1993, 1994, Lara-Ochoa y col. 1994, 1995, Almagro y col. 1995). De estas investigaciones se concluye la existencia de fuertes restricciones en el uso de aminoácidos tanto en las posiciones relacionadas con la estructura del sitio de unión como en las posiciones hipervariables relacionadas con la especificidad del anticuerpo. Se encontró además que ciertas propiedades físico-químicas se mantienen entre anticuerpos de diferente especificidad y que esto es independiente de la muestra analizada. En los trabajos presentados aquí se encuentra que también a nivel de estructura tridimensional existen fuertes restricciones en la geometría del sitio de unión. Considero que no obstante los sesgos encontrados a varios niveles de organización y funcionamiento de la respuesta inmune es indispensable conjuntar este modelo con la propiedad de alta especificidad de la respuesta inmune. Esta capacidad de generar respuestas de exquisita especificidad constituye indudablemente una de las propiedades esenciales del sistema inmune que deben ser explicadas por el modelo de reconocimiento inmune. Los resultados y el modelo expuesto aquí y los expuestos por otros autores mencionados en los antecedentes intentan poner un poco de orden en el estudio de este sistema (Kabat y col. 1991a). A partir de la reducción en la complejidad del modelo que

describe al sistema es como podemos intentar comprender los aspectos particulares de respuestas inmune especificas e integrar toda esta información en un modelo global

6. BIBLIOGRAFIA.

Almagro, J.C., Zenteno, R., Vargas-Madrazo, E. & Lara-Ochoa, F. (1995) Variability analysis of the T-cell receptors using three variability indexes *Int. J. Peptide Protein Res.* in press

Alzari, P.M., Lascombe, M.B. & Poljak, R.J. (1988) Three-dimensional structure of antibodies *Annu. Rev. Immunol.* **6**:555-580.

Amit, A.G., Mariuzza, R.A., Phillips, S.E.V. & Poljak, R.J. (1986) Three-dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science.* **233**:747-753.

Amzel, L.M., Poljak, R.J., Saul, F., Varga, J.M. & Richards, F.F. (1974) The three dimensional structure of a combining region-ligand complex of immunoglobulin NEW at 3.5-Å resolution *Proc. Natl. Acad. Sci.* U.S.A. **71**:1427-1430.

Amzel, L. M. & Poljak, R. J. (1979). Three-dimensional structure of Immunoglobulins *Annu. Rev. Biochem* **48**:961-997.

Arévalo, J.H., Taussig, M.J. & Wilson, I.A. (1993) Molecular basis of crossreactivity and the limits of antibody-antigen complementarity. *Nature.* **365**:859-863.

Bhat, T.N., Bentley, G.A., Fischmann, T.O., Boulot, G. & Poljak, R.J. (1990) Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature.* **347**:483-485.

Breckenridge, R.J., (1991) "Molecular Recognition: Models for drug design", *Experientia.* **47**:1148-1161.

Brünger, A.T. (1991) Solution of a Fab (26-10)/digoxin complex by generalized molecular replacement. *Acta Crystallogr.* A47:195-204.

Burley, S.K. y Petsko, G.A., (1988) "Weakly Polar Interactions in Proteins", *Adv. Prot. Chem.* **39**:125-189.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, Phillips, S. E. V. & Poljak, R. J. (1986) The Predicted Structure of Immunoglobulin D1.3 and Its Comparision with the Crystal Structure. *Science.* **233**:755-758.

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

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**, 877-883.

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

Cocho, G., Lara-Ochoa, F., Vargas-Madrazo, E., Jimenez-Montaño, M.A. & Ruis, J.L. (1993) Structural Patterns in Macromolecules. In Thinking About Biology, Eds. W.D. Stein and F.J. Varela , SFI Studies in the Sciences of Complexity, Lect. Note Vol. III, Addison-Wesley. pp. 105-119.

Colman, P.M., Laver, W.G., Varghese, J.M., Baker, A.T., Tulloch, P.A., Air, G.M. & Webster, R.G. (1987) Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. *Nature.* **326**:358-363.

Conrad, M. (1985) On design Principles for a Molecular Computing *Comm. ACM* **28**:464.

Cox J. L. P., Tomlinson, I. M. & Winter, G. (1994) A distionary of human germ-line V-kappa segments reveals a strong bias in their usage. *Eur. J. Immmunol.* **24**:827-836.

Davies, D.R. & Metzger, H. (1983) Structural basis of antibody function. *Ann. Rev. Immunol.* **1**:87-117.

Davies, D.R., Padlan, E.A. & Sheriff, S. (1990) Antibody-antigen complexes. *Ann. Rev. Biochem.* **59**:439-473.

Day, E.D. (1990) Advanced Immunochemistry. Wiley-Liss. New York.

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

Edelman, G.M. & Gall, E. (1969) The Antibody Problem *Ann. Rev. Bioch.* **32**:699.

Edmundson, A.B., Ely, K.R., Herron, J.N. & Cheson, B.D. (1987) The binding of opioid peptides to the Meg light chain dimer: Flexible keys and adjustable locks. *Mol. Immunol.* **24**:915-935.

Epp, O., Palm, W., Felhammer, H., Ruhlmann, A., Steigemann, W., Schwager, P. & Huber, R. (1972) *J. Mol. Biol.* **69**:315.

Fan, Z. C., Shan, L., Guddat, L. W., He, X. M., Gray, W. R., Raison, R. L. & Edmunson, A. B. (1992). Three-

dimensional structure of an Fv from a human IgM immunoglobulin *J. Mol. Biol.* **228**, 188-207
Fersht, A.R. y col., (1985) Hydrogen bonding and biological specificity analysed by protein engineering *Nature* **314**:235
Friedman, A.R., Roberts, V.A. & Tainer, J.A. (1994) Predicting molecular interactions and inducible complementarity: Fragment docking of Fab-peptide complexes *Proteins.* **20**:15-24.
Getzoff, E.D., Tainer, J.A., Lerner, R.A. & Geysen, H.M. (1988) The chemistry and mechanism of antibody binding to protein antigens *Adv. Immunol.* **43**:1-98.
Geysen, H.M., Tainer, J.A., Rodda, S.J., Mason, T.J., Alexander, H., Getzoff, E.D. & Lerner, R.A. (1987) Chemistry of antibody binding to a protein *Science* **235**:1184-1190.
Herron, J.N., He X., Mason, M.L., Voss, E.W., Jr. & Edmundson, A.B. (1989) Three-dimensional structure of a fluorescein-Fab complex crystallized in 2-methyl-2,4-pentanediol *Proteins.* **5**:271-280.
Herron, J.N., He X.M., Ballard, T.W., Blier, P.R., Pace, P.E., Bothwell, A.L.M., Voss, E.W., Jr. & Edmundson, A.B. (1991) An autoantibody to single-stranded DNA: Comparison of the Three-dimensional structures of the unliganded Fab and a deoxinucleotide-Fab complex. *Proteins* **11**:159-175
Horjales, E. (1995). Comentario del revisor de tesis.
Janin, J. & Chothia, C. (1990) The structure of protein-protein recognition sites *J. Biol. Chem.* **265**:16027-16030.
Judson, H.F. (1979) The Eight Day of Creation. The Makers of Revolution in Biology. Simon & Schuster (New York, 686 pp.) p.61.
Kabat, E.A. (1967) A comparison of invariant residues in the variable and constant regions of human K, human L and mouse L Bence Jones proteins. *Proc. Natl. Acad. Sci.* U.S.A. **58**:229-233.
Kabat, E.A. (1968) Unique features of the variable regions of Bence Jones proteins and their possible relation to antibody complementarity. *Proc. Natl. Acad. Sci.* U.S.A. **59**:613-619.
Kabat, E.A. & Wu, T.T. (1971) Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N. Y. Acad. Sci.* **190**:382-393.
Kabat E. A., Wu T. T. & Bilofsky H. (1976) Attempts to locate residues in complementary-determining regions of antibody combining sites that make contact with antigen *Proc.Natl. Acad. Sci.* U.S.A. **73**:617-619.
Kabat, E. A., Wu T. T. & Bilofsky H. (1977) Unusual Distributions of Amino acids in complementary determining (Hypervariable) segments of Heavy and Light chains of Immunoglobulins and their Possible roles is specificity of Antibody- combining sites. *J. Biol. Chem.* **252**:6609-6616.
Kabat, E.A. Wu, T.T. & Bilofsky, H. (1978) Variable region genes for the immunoglobulin framework are assembled from small segments of DNA - A hypothesis. *Proc. Natl. Acad. Sci.* U.S.A. **75**:2429-2433.
Kabat, E.A., Wu, T.T. & Bilofsky, H. (1980a) Evidence supporting somatic assembly of the DNA segments (minigenes), coding for the framework, and complementarity-determining segments of immunoglobulin variable regions. *J. Exp. Med.* **149**:1299-1313.
Kabat, E.A., Wu, T.T. & Bilofsky, H. (1980b) Evidence indicating independent assortment of framework and complementarity-determining segments of the variable regions of rabbit light chains. Delineation of a posible J minigene. *J. Exp. Med.* **152**:72-84.
Kabat, E.A. & Wu, T.T. (1991a) Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. *J. Immunol.* **147**:1709-1719.
Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. & 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.
Kunh, A. en "La esructura de las Revoluciones cientificas". Breviarios del F.C.E., México D. F. (1986).
Lara-Ochoa, F., Vargas-Madrazo, E., Jimenez-Montaño, M.A. & Almagro, J.C. (1994) Patterns in the complementarity determining regions of immunoglobulins (CDR's). *BioSystems.* **32**:1-9.
Lara-Ochoa, F., Vargas-Madrazo, E. & Almagro, J.C. (1995) Distribution of the use frequencies of amino acids in the hypervariable regions of immunoglobulins. *J. Mol. Evol.* **39**: In press.
Lehn, J-M. (1990) Perspectives in Supramolecular Chemistry-From Processing and Self Organization. *Agnew.*

Chem. Int. Ed. Engl. **29**:1304-1319.
Marluzza, R.A., Phillips, S.E.V. & Poljak, R.J. (1987) The structural basis of antigen-antibody recognition. *Ann. Rev. Biophys. Biophys. Chem.* **16**:139-159.
Marquart, M. & Deisenhofer, J. (1982) The three-dimensional structure of antibodies. *Immunology Today* **3**:164-166.
Mian I. S., Bradwell A.R. & Olson A. J. (1991) Structure, Function and Properties of Antibody Binding Sites. *J. Mol. Biol.* **217**:133-151.
Nissonoff, A., In: The Antibody Molecule, Academic Press Inc. (London) 1975.
Ohno S., Mori N. & Matsunaga T. (1985) Antigen-binding Specificities of Antibodies are Primarily determined by Seven Residues of Vh." *Proc. Natl. Acad. Sci.* U.S.A. **82**:2945-2949.
Padlan, E. A. (1977) Structural Implications of Sequence Variability in Immunoglobulins. *Proc. Natl. Acad. Sci.* U.S.A. **74**:2551-2555.
Padlan E. A. (1979) Evaluation of the Structural Variation Among Light Chain Variable Domain. *Mol. Immunol.* **16**:287-296.
Padlan, E.A. (1990) On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins.* **7**:112-124.
Padlan, E.A. (1994) Anatomy of the antibody molecule. *Mol. Immunol.* **31**:169-193.
Padlan, E.A., Segal, D.M., Spande, T.F., Davies, D.R., Rudikoff, R. & Potter, M. (1973) Structure at 4.5 Å resolution of a phosphoril-choline Fab. *Nature New Biol.* **245**:165-167.
Padlan, E.A. & Davies, D.R. (1975) Variability of three-dimensional structure in immunoglobulins. *Proc. Natl. Acad. Sci.* U.S.A. **72**:819-823.
Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J. & Davies, D.R. (1989) Structure of an antibody-antigen complex: Crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci.* U.S.A. **86**:5938-5942.
Pauling, L. (1945) Molecular structure and Intermolecular forces. In The specificity of serological reactions. Karl Landsteiner, M.D. Dover Publications, Inc. New York.
Poljak, R.J., Amzel, L.M., Avery, H.P., Becka, L.N. & Nisonoff, A. (1972). *Nature* (London), New Biol. **235**:137.
Poljak, R.J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizacherley, R. P. & Saul, F. (1973). Three-dimensional structure of the Fab' fragment of a human Immunoglobulin at 2.8-Å resolution. *Proc. Natl. Acad. Sci.* U.S.A. **70**. 3305-3310.
Rebek Jr, J. (1991) Molecular recognition and the development of self-replicating systems. *Experientia.* **47**:1096-1104.
Rini, J.M., Schulze-Gahmen U. & Wilson, I.A. (1992) Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science.* **255**:959-965.
Schiffer, M., Girling, R.L., Ely, K.R. & Edmunson. A.B. (1973) Structure of a lambda-type Bence-Jones protein at 3.5-Å resolution. *Biochemistry.* **12**:4620-4631.
Sheriff, S., Silverton, E. U., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987) Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci.* U.S.A. **84**:8075-8079.

Suckling, C.J. (1991) Molecular Recognition: A universal molecularscience. *Experientia* **47**:1093.
Tainer J. A., Getzoff E. D., Paterson Y., Olson A. J. & Lerner R. A. (1985) The atomic mobility component of protein antigenicity. *Annu. Rev. Immunol.* **3**:501-535.
Telford, M.C. & Stimson, W.H. (1991) Molecular Recognition in antibodies and its application. *Experientia* **47**:1129.
Thom, R. (1972) in "Towards a Theoretical Biology". Ed. by C.H. Waddington, pp. 68-82.
Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992) The repertoire of human germline V_H segments reveals fifty groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* **227**:776-798.

Tonegawa, S. (1983) Somatic Generation of antibody diversity. *Nature.* **320** 575.
Van Regenmortel, M. H. V. (1989) Structural and functional approaches to the study of protein antigenicity. *Immunol. Today* **10** 266-271.
Vargas-Madrazo E., Almagro J. C., Lara-Ochoa F. & Jiménez-Montaño M. A. (1992) Proceedings of the Seventh Panamerican Biochemical Congress, Ixtapa, México, September 27 -October 2
Vargas-Madrazo, E., Almagro, J.C., Lara-Ochoa, F. & Jiménez- Montaño, M.A. (1993) Articulo in extenso en Memorias del Congreso ECAL '93. May 24-26th, Brussels, Belgium pp 1070-1089.
Vargas-Madrazo, E., Lara-Ochoa, F. & Jiménez-Montaño, M.A. (1994) A skewed distribution of amino acids at recognition sites of the hypervariable region of immunoglobulins. *J. Mol. Evol* **38**:100-104.
Wilson, I. A. & Stanfield, R. L. (1993) Antibody-antigen interactions. *Curr. Op. Struct. Biol.* **3**, 113-118.
Wu, T.T. & Kabat, E. A. (1970) An analysis of the sequences of the variable region of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med* **132**:211-249.
Wu, T.T., Johnson, G., & Kabat, E.A. (1993) Length distribution of CDR3 in antibodies. *Proteins.* **16**:1-7.