ESTIMACIÓN DE LA RELACIÓN, ENTRE LA SECUENCIA DE LOS GENES FUNCIONALES hORs *(RECEPTORES DEL OLFATO HUMANO*) CON SU LOCALIZACIÓN CROMOSOMICA, MEDIANTE HERRAMIENTAS BIOINFORMÁTICAS

HENRY FABIÁN TOBAR TOSSE

UNIVERSIDAD DEL CAUCA
FACULTAD DE CIENCIAS NATURALES, EXACTAS Y DE LA EDUCACIÓN
DEPARTAMENTO DE BIOLOGÍA
GRUPO DE BIOLOGÍA MOLECULAR, AMBIENTAL Y CÁNCER
POPAYÁN
2006

ESTIMACIÓN DE LA RELACIÓN, ENTRE LA SECUENCIA DE LOS GENES FUNCIONALES hORs *(RECEPTORES DEL OLFATO HUMANO*) CON SU LOCALIZACIÓN CROMOSOMICA, MEDIANTE HERRAMIENTAS BIOINFORMÁTICAS

HENRY FABIÁN TOBAR TOSSE

Trabajo de Grado presentado como requisito
Para optar el titulo de Biólogo

Directora
PATRICIA EUGENIA VELEZ VARELA
M.Sc.

Codirector
PEDRO ANTONIO MORENO TOVAR
Ph.D.

UNIVERSIDAD DEL CAUCA
FACULTAD DE CIENCIAS NATURALES, EXACTAS Y DE LA EDUCACIÓN
DEPARTAMENTO DE BIOLOGÍA
GRUPO DE BIOLOGÍA MOLECULAR, AMBIENTAL Y CÁNCER
POPAYÁN
2006

Nota de aceptación:

_____
_____
_____

Director _____
M.Sc. PATRICIA E. VELEZ VARELA

Jurado _____
M.Sc. SULMA MUÑOS BENITEZ

Jurado _____
Ing. EMBER UBEIMAR MARTINEZ

Fecha de sustentación:

Popayán, 4 de Febrero de 2007

Hoy doy gracias a Dios por permitirme culminar exitosamente esta etapa de mi vida, por permitirme compartir mi trabajo y vida con personas maravillosas, por regalarme verdaderos amigos, y por permitirme estar con toda mi familia compartiendo este triunfo.

Fabián Tobar.

AGRADECIMIENTOS

El autor expresan sus agradecimientos a:

M.Sc. Patricia Eugenia Vélez Varela. Directora. Por su amistad, orientación colaboración y dirección de mi trabajo de grado.

Ph.D. Pedro Antonio Moreno Tovar. Codirector. Por brindarme la oportunidad de trabajar a su lado, por su orientación, enseñanza y asesoría de mi trabajo de grado.

Ing. Ember Martinez y M.Sc. Sulma Muños Benites. Jurados. Por las sugerencias ofrecidas a mi trabajo.

Ing. Luís Garreta. Ingeniero de la Universidad del Cauca. Por su colaboración, enseñanza y apoyo en el desarrollo de mi trabajo de grado

Grupo de Biología Molecular, Ambiental y Cáncer (BIMAC), por permitirme realizar mi trabajo de grado en el Laboratorio de Bioinformática de la Universidad del Cauca.

Weimar Pérez, Anderson Muñoz, Héctor Ramírez, Adalberto Trujillo y demás integrantes del grupo GAIA, por su amistad y apoyo a lo largo de mi carrera.

A Dios y su Genoma Humano. Por ser la herramienta e inspiración para el presente trabajo.

Universidad del Cauca. Por la formación brindada.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE GRAFICAS

LISTA DE ANEXOS

GLOSARIO

El siguiente glosario fue creado bajo (http://fbio.uh.cu/bioinfo/glosario.html):

Algoritmo: Secuencia de pasos que conforman una tarea específica, usualmente en el ámbito computacional

Alineamiento de secuencias: Arreglo mutuo de dos o más secuencias, que muestra donde estas son similares y donde difieren. Un alineamiento óptimo es aquel que muestra la mayor cantidad de correspondencias y la menor cantidad de diferencias.

Base de datos: Conjunto de datos consistente y usualmente persistente, organizado en un modo específico que permita acceder a la información de forma fácil y rápida. En el mundo de la informática existen varios sistemas que permiten gestionar bases de datos, muchos de ellos basados en el leguaje SQL, sin embargo, en la práctica, cualquier fichero que contenga información organizada según un formato específico puede considerarse como una base de datos, por ejemplo: un fichero escrito en formato FASTA o en un lenguaje de marcado como XML.

Bioinformática: De manera simplificada, la aplicación de las tecnologías de la computación y la información al manejo y análisis de información de origen biológico.

BLAST: (Basic Local Alignment Search Tool) Herramienta para la comparación de secuencias biológicas a través del alineamiento. Utilizado para la búsqueda en bases de datos

Bootstrap: test aleatorio para la verificación de los árboles filogenéticos correctos

CDS: Secuencia Codificante. Región codificante de un gen

Clados: Grupo de datos con igual ancestro (nodo)

Clustal W: Algoritmo para el alineamiento múltiple de secuencias.

Dendograma: (Ideograma) Árbol creado tras la estimación de distancias a partir de otras características como las matrices de puntuación.

Dominio: Segmento de cadena polipeptídica que se pliega de forma independiente del resto de la cadena y que usualmente tiene una función específica. También usado para designar al segmento de gen que codifica para éste.

Exón: Región codificante dentro de un gen discontinuo.

Familia: Grupos de proteínas (y sus genes codificantes) que comparten características funcionales semejantes y una obvia relación entre sus secuencias.

FASTA: 1)**.** (algoritmo, herramienta) Una herramienta desarrollada por Pearson y Lipman para el alineamiento de secuencias de ácidos nucleicos y proteínas. 2) (formato) Uno de los formatos más simples utilizados para almacenar secuencias nucleotídicas o aminoacídicas. Una entrada con formato FASTA tiene dos bloques fundamentales, el primero está formado por una sola línea que comienza con '>' y no es más que una descripción de la secuencia y el segundo está formado por la secuencia en sí e implica tantas líneas como sea necesario

Filogenética: La rama de la biología que se ocupa de descubrir las líneas de origen o **filogenias** de los organismos, a fin de construir las relaciones antepasado-descendientes (**relaciones evolutivas**) entre los grupos de organismos vivos y extinguidos. En el contexto molecular estas relaciones se infieren a partir de las similitudes y diferencias en secuencias aminoacídicas o nucleotídicas.

Formato: En el mundo de la informática, el conjunto de reglas o especificaciones mediante las cuales se pueden organizar datos de diversa naturaleza, para poder acceder posteriormente a estos a través de los intérpretes adecuados. v. lenguaje de marcado, FASTA.

GAPS: Espacios o huecos dentro de una secuencia necesarios para el alineamiento.

hOR: Receptor Olfativo Humano

INDELS: Inserciones y Deleciones de una secuencia biológica

Itrón: Región no codificante dentro de un gen discontinuo.

Librería de cDNAs: Una colección de moléculas bicatenarias de ADN (cDNAs) obtenidas a partir de las moléculas de los ARN mensajeros correspondientes. Puesto que los cDNAs se obtienen a partir de moléculas de ARN mensajero, las librerías de cDNAs permiten obtener información acerca de los genes estructurales que se están expresando en la célula en un momento dado. Las librerias de cDNAs también se usan experimentalmente para conocer la secuencia codificante de los genes discontinuos típicos de eucariontes, después de que los intrones han sido escindidos.

Linux: Sistema operativo derivado de UNIX que, ha mantenido casi todas las ventajas que este último ofrece, como el ser multitarea y basado en bibliotecas dinámicas. Fue desarrollado originalmente por el estudiante finlandés de informática Linus Torvalds, que publicó su código fuente en 1990, en la forma de código abierto, esto es, accesible para toda la comunidad, sin restricciones para modificarlo y ampliarlo. Este hecho, unido a la estructura modular del sistema operativo (basado en la integración de componentes de software independientes) generó una nueva visión de desarrollo informático y ha permitido que Linux se haya expandido notablemente, gracias al trabajo, muchas veces voluntario y sin ánimo de lucro, de miles de programadores a todo lo largo del mundo. Actualmente están disponibles varias distribuciones de Linux, ofertadas por diversos proveedores, como RedHat, SuSE o Mandrake Inc.

Matrices de Sustitución: Matrices que evalúan los cambios de aminoácidos o nucleótidos presentes en secuencias alineadas.

Motivo: Regiones conservadas relativamente cortas (de 10 a 20 residuos) en un alineamiento múltiple de varias secuencias de proteínas que pertenecen a la misma familia. Los motivos, también conocidos como bloques**,** representan usualmente elementos importantes desde el punto de vista estructural o funcional y pueden utilizarse como descriptores de la familia en cuestión.

Mutación: Proceso mediante el cual el material genético sufre un cambio detectable y heredable, generalmente en la forma de una variación en la secuencia del ADN. Estas variaciones pueden ser puntuales (cambio de un nucleótido por otro) o deberse a la inserción (ganancia) o deleción (pérdida) de un segmento de nucleótidos. Las mutaciones son la base de la existencia de los alelos.

Phylographer: herramienta que representa los resultados del BLAST por medio de grafos.

Secuencia: La disposición ordinal de los monómeros que forman parte de las biomoléculas de tipo "polimérico" como los ácidos nucleicos y las proteínas. La mayor parte de la información en el contexto de los sistemas biológicos está representada en la forma de secuencias nucleotídicas (en los ácidos nucleicos) y aminoacídicas (en las proteínas). Las variaciones en la secuencia nucleotídica de los genes que codifican para proteínas, debidas a mutaciones, pueden mediar variaciones en la secuencia aminoacídica de estas proteínas. En tanto que, tales variaciones en la secuencia aminoacídica de las proteínas, usualmente acarrean variaciones en su estructura tridimensional y por ende, en su función.

Similitud: Una medida de la semejanza entre dos secuencias, que no necesariamente implica una relación de parentesco entre estas. En la práctica, la similitud entre dos secuencias se expresa en base al por ciento que representa el número de residuos idénticos en el alineamiento, con respecto al número total de residuos en la secuencia más corta, valor que se conoce como por ciento de similitud o identidad**.**

Software: En computación, procedimientos y reglas lógicas escritas en la forma de programas y aplicaciones, que definen el modo de operación de la computadora. Tienen carácter virtual (en contraposición con el hardware) y están almacenadas en los diferentes tipos de memoria de lectura/escritura.

TSS: Del Ingles transcription Start Site o sitio de comienzo de la transcripción

Query: Secuencia de entrada (u otro tipo de termino de búsqueda) que serán comparada en la base de datos.

ESTIMACIÓN DE LA RELACIÓN, ENTRE LA SECUENCIA DE LOS GENES FUNCIONALES hORs *(RECEPTORES DEL OLFATO HUMANO*) CON SU LOCALIZACIÓN CROMOSOMICA, MEDIANTE HERRAMIENTAS BIOINFORMÁTICAS.

Tesis de Pregrado en Biología.

Por

Henry Fabián Tobar Tosse

RESUMEN

La familia de genes del receptor olfativo humano (hOR) esta constituida por cerca de 800 genes y pseudogenes. Un estudio previo demostró que existe una correlación entre los agrupamientos cromosómicos de los genes y los agrupamientos topológicos de los árboles filogenéticos, sugiriendo una expansión de la familia a partir de cierto grupo ancestral de genes (Niimura & Nei, 2003). A fin de investigar si dicha relación compromete a las regiones promotoras de los genes, fueron analizados 385 Genes hOR obtenidos de la base de datos HORDE, Mediante diferentes algoritmos bioinformáticos de alineamiento, filogenia y minería de datos. El descubrimiento y definición de dos modelos estructurales de promotor a partir de motivos altamente conservados dentro de las regiones promotoras, ha permitido afirmar que los mecanismos de regulación transcripciónales son comunes a grupos de genes filogenéticamente distantes que no se encuentran asociados a los agrupamientos cromosómicos y filogenéticos establecidos anteriormente y en el presente trabajo.

Palabras claves: Bioinformática, Familias de genes, receptores olfativos (OR), motivos, agrupaciones, filogenia y topología de genes.

Directora: M.Sc. Patricia Eugenia Vélez Varela
Codirector: Ph.D. Pedro Antonio Moreno Tovar

ESTIMATING THE RELATIONSHIP BETWEEN THE SEQUENCE OF THE FUNCIONAL hOR GENES (HUMAN OLFACTOY RECEPTORS) WITH ITS CROMOSOMAL LOCATION, BY MEAN OF BIOINFORMATICS TOOLS.

Biology Grade Tesis.

By

Henry Fabián Tobar Tosse

ABSTRACT

The human olfactory receptor gen family (hOR) is consist of about 800 genes and pseudogenes. One previous study showed that there exist a relationship between the chromosomic clustering and phylogenetic topological clustering suggesting how the hOR gene family has expanded for the entire genome from an ancestral gene group (Niimura & Nei, 2003). In this work we study 385 genes hOR of the HORDE database (by mean of different bioinformatics tools, such as, aligment, phylogenetic trees and datamining), in order to find out whether the promoter sequences are involved on this relationship. We found and characterized two structural models of promoter, starting from high conserved motifs in the promoter regions, and we stablish that the transcriptional regulation mechanims are the same to group of genes distantly in the phylogenetic tree, different to the chromosomic clustering and phylogenetic topological clustering stablish in this staudy and others works.

Keys: Bioinformatics, gene families, olfactory receptors (OR), motifs, clustering, phylogenetic and topological clustering.

Director: M.Sc. Patricia Eugenia Vélez Varela
Codirector: Ph.D. Pedro Antonio Moreno Tovar

INTRODUCCIÓN

El conocimiento de las secuencias del genoma humano, así como de la jerarquía estructural y funcional de los genes que lo conforman, pone de manifiesto la gran y estrecha relación que existe entre ellos para la formación de un ser vivo. Es por tal motivo, que las investigaciones actuales buscan entender la forma en la que estos sistemas biológicos se regulan, expresan y mantienen durante el desarrollo de un organismo.

Conocer y entender estos mecanismos es un trabajo muy complejo debido a la gran cantidad de información con la que se cuenta tras la secuenciación del genoma humano y las pocas herramientas bioinformáticas óptimas para su estudio y análisis. Ahora bien, como un aporte al conocimiento y entendimiento a fondo de los genomas, nosotros realizamos un trabajo exploratorio en una familia de genes del genoma humano comenzando a estudiar patrones o regularidades dentro de la familia, debido principalmente a que las familias de genes pueden ofrecer indicios de la estructura y función de los genes y proteínas, además de cómo la información biológica se desarrolla y especifica desde los organismos primitivos hasta los superiores mediante la duplicación, mutación y replicación de sus genes (Abascal, 2003).

Esta estrategia de estudio surge tras los descubrimientos hechos dentro de la familia de genes MSP en *C.elegans*, en donde se establece una relación directa entre los agrupamientos de los árboles filogenéticos de los genes completos y la ubicación topológica de los genes dentro de los cromosomas (Moreno & Fox, 2004). Esto permitió establecer, que los genes estrechamente relacionados filogenéticamente y que se encuentran dentro de una misma región cromosómica, comparten así mismo un mecanismo regulatorio y una función asociada.

Es importante mencionar que las regularidades más notables y que jerarquizan a los genes en Superfamilias, Familias y Subfamilias, son los agrupamientos filogenéticos que ofrecen los árboles filogenéticos, esto debido a que pueden mostrar la forma como los genes se expanden a través del genoma y los efectos estructurales y funcionales que esto implica.

La familia con la cual se siguió esta estrategia de búsqueda y análisis, es la familia de genes codificantes de receptores olfativos humanos (hOR), debido principalmente a que en la actualidad sus CDS (secuencias codificantes) y Proteínas se encuentran caracterizados molecular (Buck & Axel, 1991) y filogenéticamente (Zozulya et al, 2001; Glusman et al, 2001 & Niimura y Nei, 2003). Permitiendo a su vez no solo buscar nuevas regularidades, sino también involucrar otras regiones poco o no estudiadas como es el caso de las regiones

promotoras. Este conjunto de regularidades fueron buscadas y caracterizadas mediante diferentes métodos y herramientas bioinformáticas, las cuales son: el alineamiento y análisis filogenético de secuencias, Blast y análisis de grafos, caracterización topológica y estructural de genes y relaciones bibliográficas asociadas al trabajo.

El presente trabajo es un aporte al entendimiento de los mecanismos de expansión y regulación de la familia de receptores olfativos, ya que a partir de nuestro estudio se ha podido conocer que los procesos de expansión de la familia dentro del genoma humano se han dado por procesos variables de expansión y en mayor grado sobre los genes el cromosoma 11 y que sobre las regiones promotoras de los genes hOR se encuentran grandes regiones conservadas permitiendo postular dos modelos estructurales de promotor, posiblemente de regulación negativa de la familia hOR y que son independientes a la homología que existe entre los genes.

Finalmente es importante mencionar que una de las metas principales del presente trabajo y del Grupo de Biología Molecular, Ambiental y Cáncer (BIMAC) de la Universidad del Cauca, es incentivar a los biólogos, ingenieros y a todos los interesados en la investigación, a relacionarse mutuamente en el desarrollo de este tipo de proyectos, en donde convergen diferentes áreas del conocimiento hacia la solución de un problema científico.

1. PLANTEAMIENTO DEL PROBLEMA

La gran cantidad de secuencias de ADN y Proteínas almacenadas, son el producto de la secuenciación automática de muchos genomas; estos datos biológicos se encuentran disponibles ante todo el mundo por la necesidad de relacionar y entender la forma como se organiza la información biológica para la formación de un ser vivo completo. Este trabajo de análisis y entendimiento de la información que contienen nuestros genes, no puede ser realizado por una sola persona o por un grupo, esto depende principalmente del esfuerzo global para poder abarcar toda esta información biológica que día tras día crece y se almacena esperando a ser analizada.

El análisis de datos biológicos almacenados debe ser realizado con la ayuda de un computadores o informática, debido ha que la gran cantidad de datos utilizados para el estudio de los genomas, puede llevar fácilmente al investigador a cometer errores y a perder grandes cantidades de tiempo en ellos. Es así como las herramientas bioinformática son muy útiles en cuanto a análisis de secuencias se refiere y pueden ofrecer muchos y variados resultados para el entendimiento de la organización y relación de los genes.

El genoma humano y en general el de todos los eucariotas, poseen una estructura y organización muy compleja, en donde las características de cada gen son únicas, aún encontrándose en una misma familia o grupo; es por tal motivo que el estudio de la información biológica debe comenzar con modelos o pilotos, y con grupos de genes no muy complejos como lo son los genes olfativos. Las diferentes características de estos genes, los convierten en un modelo ideal para la búsqueda de relaciones estructurales, funcionales y de regulación, que no han sido establecidas con claridad.

2. HIPÓTESIS

La relación entre la secuencia de los genes funcionales hOR con su localización cromosómica, involucra a sus regiones promotoras.

3. OBJETIVOS

3. OBJETIVO GENERAL

Estimar la relación entre la secuencia de los genes funcionales de la familia de receptores del sistema del olfato Humano (hORs) con su localización cromosómica, mediante la utilización de Herramientas Bioinformáticas.

3.1 OBJETIVOS ESPECÍFICOS

- Establecer la estructura del gen completo y funcional hOR

-Asociar las características funcionales del receptor olfativo, a sus mecanismos de regulación.

4. JUSTIFICACION

La necesidad de obtener información de las secuencias de ADN y Proteínas que se encuentran almacenadas en las bases de datos de todo el mundo, es lo que ha llevado a los biólogos a aplicar metodologías informáticas para su análisis; el cual representa principalmente el significado biológico de diferentes alineamientos de secuencias, filogenética molecular de genomas, estructura de proteínas, entre otras.

Por tanto, es necesario establecer diferentes y variadas relaciones entre las secuencias de ADN o Proteínas que se encuentran almacenadas en las bases de datos y que puedan ofrecer significado biológico y estructural de los mismos; asociar diferentes características de estas moléculas como la secuencia establecida de los genes y su localización cromosómica, para evaluarlas de forma rigurosa mediante herramientas bioinformáticas, nos permitirán plantear hipótesis o soluciones al dilema de la organización de los genomas.

Recientemente, se descubrió en la familia de genes MSP (mayor sperm protein) en *C. elegans* una relación entre los agrupamientos cromosómicos y las secuencias de los genes (Moreno & Fox, 2004). Con el fin de demostrar si esta propiedad puede ser extensible a otras familias de genes, se ha propuesto el presente estudio para encontrar una relación similar entre la familia de genes funcionales de los receptores olfativos humanos hORs, dada las características moleculares especiales de esta familia para el análisis de genomas.

Los receptores OR del olfato humano pertenecen a una gran familia de genes y proteínas que en la actualidad sirven como modelo de estudio en el genoma humano para el entendimiento de la organización y expresión de los genes, permitiendo un mejor y exacto análisis de los mecanismos moleculares que rigen sus genes, de ahí la importancia de tomar esta familia como base biológica del presente estudio.

La identificación, clonación y análisis de todos los genes OR funcionales es un importante paso en el entendimiento de la especificidad receptor-ligando y la codificación combinatorial del estimulo oloroso en la olfación humana (Zozulya et al, 2001).

5. MARCO TEORICO

5.1 ESTUDIO DE GENOMAS Y FAMILIAS DE GENES

El estudio de los genomas surge tras la necesidad de conocer y entender la forma como se organiza los genes y proteínas en la formación de un organismos complejo como lo es el hombre. Muchas técnicas fueron y son utilizadas para este fin como los son la reacción cruzada con anticuerpos, la hibridación y los marcadores moleculares (Kumar & Filipski, 2001). Estos estudios brindaban importante información acerca de la variabilidad de los genes y el polimorfismo de ellos dentro de las especies, pero muy poca acerca de la estructura y organización de los genes dentro de los genomas.

Es así como el conocimiento estructural de los genes y genomas, se convierte en una necesidad científica solo resuelta tras la aparición de nuevas técnicas moleculares o la combinación de estas como la secuenciación directa de ADN mediante mapeo genético y de ligamiento (Venter et al., 2001). Estas técnicas algunas ya automatizadas en la actualidad, permitieron secuenciar genomas completos y por ende establecer parámetros y relaciones que permitieran estudiar la estructura y función de cada uno de los genes que forman un organismo.

Las investigaciones actuales sobre genomas y en especial el genoma humano, toman como base las características generales encontradas en los primeros borradores del genoma humano, en donde las regularidades entre genes o proteínas fueron y son materia de estudio debido a la particularidad de asociarlas a procesos complejos de programación, expresión y regulación de genes. Una de estas regularidades y que hace parte de un análisis extenso dentro de el proyecto genoma humano, es la duplicación de los genes y su agrupamiento dentro de los cromosomas, debido a la incertidumbre acerca de las implicaciones y mecanismos moleculares por los cuales se rige este fenómeno (Venter et al., 2001).

La duplicación y otros mecanismos evolutivos de los genomas, se ponen de manifiesto tras los primeros secuenciamientos de genes y el estudios en ellos, en donde el resultado de la clonación y amplificación de estos, permitía no solo encontrar el gen estudiado sino una gran variedad de genes asociados o con pocas diferencias en su secuencia y cuya función se asociaba a procesos o estructuras semejantes (Otha, 2003). Es así como en los siguientes años estos genes se agruparon en una jerarquía de superfamilias, familias y subfamilias mediada por el grado de homología que existía entre las secuencias encontradas (Henrissat & Romeu, 1995; Murzin *et al.*, 1995; Yona *et al.*, 1999).

La secuenciación del genoma humano no solo ha permitido aumentar el numeró de genes por familia, sino inferir funciones específicas para sus miembros, tratando de entender la forma en que los genes están organizados y controlan sus funciones (Nei & Kumar, 2000). De esta forma las familias de genes y proteínas nos pueden ofrecer información acerca de los mecanismos por los cuales se regulan y expresan los miembros de cada una de ellas, debido a que los genes pertenecientes a una familia multigénica están usualmente bajo un control regulatório y funcional común (Otha, 2003). Esta característica es realmente importante debido a que nos podría aclarar los procesos por los cuales son adquiridos los mecanismos de regulación y expresión, al igual que la comprensión de los fenómenos estructurales conocidos tras la secuenciación del genoma humano.

De acuerdo a Davidson (2001), el estudio en las familias de genes de los mecanismos regulatorios podría ser el tema más prominente en los próximos años, dada la necesidad de asociar estos mecanismos al control y prevención de enfermedades génicas.

5.2 FAMILIA DE GENES CODIFICANTES DE RECEPTORES OLFATIVOS HUMANOS (hOR)

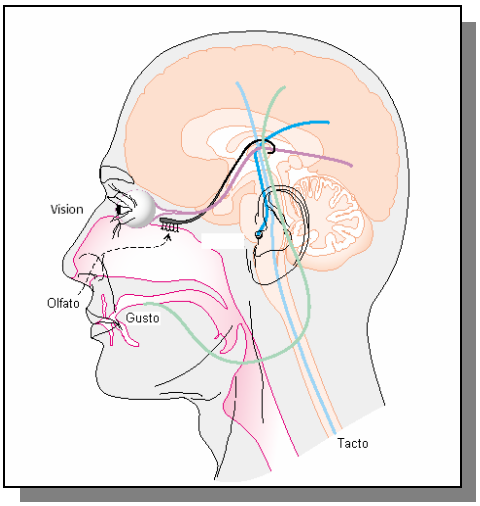
Teniendo claro las implicaciones moleculares que abarca el análisis de y dentro de las familias de genes, es claro que para el correcto estudio de estas se deben plantear modelos asociados a caracteres fisiológicos complejos, es decir estudiar familias de genes asociadas a grandes complejos celulares que cumplan una función igualmente importante dentro de las células, con el fin de que las implicaciones moleculares estudiadas tengan correlación o asocio con las demás moléculas y estructuras implicadas en el proceso fisiológico.

La familia de genes codificantes de Receptores Olfativos OR cumple muchas de las características buscadas para este tipo de estudios, debido a que esta hace parte de procesos complejos de comunicación celular en donde las proteínas son la maquinaria fundamental para la percepción de moléculas del exterior.

La familia de genes OR hace parte de una gran superfamilia de siete dominios extramembranales o 7TM acopladas a proteínas G, siendo esta, de gran importancia para la célula debido a que la mayor parte de las señales que vienen del exterior son demasiado grandes o polares para atravesar la membrana celular. Por tanto estos son los responsables de trasmitir la información iniciada por señales tan diversas como fotones, olores, sabores, hormonas y neurotransmisores a través de la membrana (Shenoy S. K. et al, 2005) siendo así, los responsables en la percepción o conocimiento del ambiente, e incluso de agentes patógenos, ayudando al sistema inmune en su función.

La figura 1 muestra los diferentes procesos a los cuales son asociados los receptores 7TM, permitiendo comprender la importancia del estudio en estas moléculas.

Figura 1. Sistemas asociados a los receptores 7TM

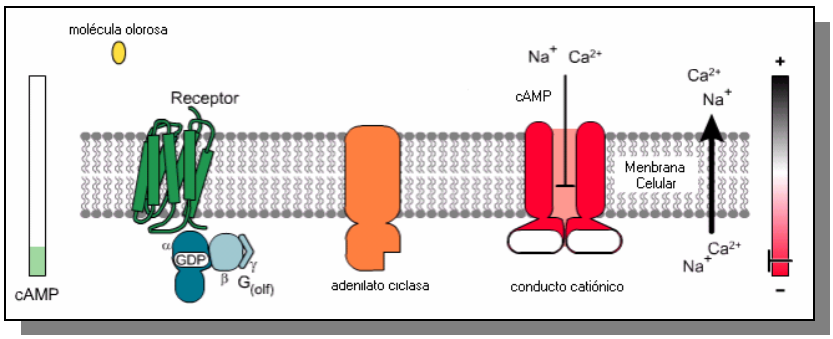


Los receptores olfativos pertenecen a una superfamilia de 7 dominios extramenbranales, responsables de trasmitir información iniciada por señales tan diversas como fotones, olores, sabores, hormonas y neurotransmisores a través de la membrana. Tomado de (Stryer, L. Biochemistry, 5ed. 2002)

Estas moléculas receptoras se encuentran acopladas a una proteína G especificas o $G_{(olf)}$, la cual permite activar toda la maquinaria intracelular para la acción iónica de las membranas, característica del sistema del olfato que fue descubierta gracias a las observaciones en epitelio olfativo de rata, donde se mostraba el aumento exponencial de moléculas implicadas en la recepción como GTP y cAMP (Buck & Axel, 1991).

Este estudio sugería que los receptores olfativos eran 7TM y que se encontraban acoplados a proteínas G, pero fue hasta su identificación mediante la hibridación con cDNA, que se pudo afirmar estas hipótesis y establecer que el complejo “Receptor 7TM-Proteína G“ permite percibir las moléculas olorosas que vienen del exterior con considerable sensibilidad y especificidad (Buck & Axel, 1991). La figura 2, representa las moléculas implicadas en la percepción de una molécula olorosa incluyendo el complejo Receptor 7TM-Proteína G especifica.

Figura 2. Moléculas implicadas en la recepción del olor



El complejo "Receptor 7TM-Proteína G "permite percibir las moléculas olorosas que vienen del exterior con considerable sensibilidad y especificidad. Tomado de (Stryer, L. Biochemistry, 5ed. 2002).

La familia de genes OR cumple su función de percepción olfativa gracias a los siete dominios extramenbranales que caracterizan sus proteínas y en general a todos los miembros de la superfamilia 7TM, la diferencia más marcada y que define a la familia OR dentro de la superfamilia 7TM es el N-terminal que en los receptores OR es extramembranal y el C-terminal que en los receptores OR es citoplasmático (Buck & Axel, 1991).

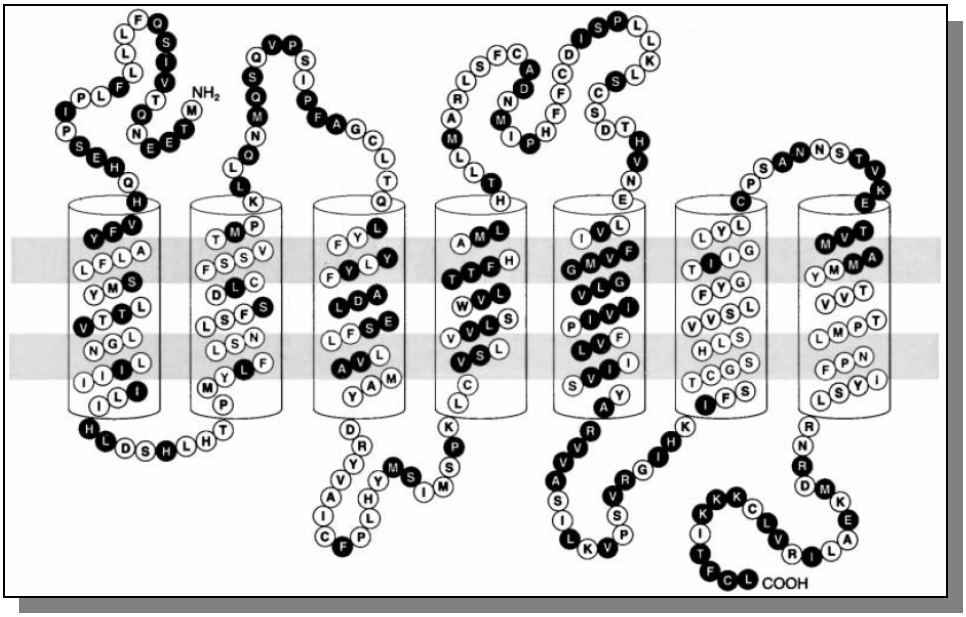
Los receptores que son codificados por los miembros de la familia OR se expresan principalmente sobre las neuronas sensoriales del epitelio olfativo en las cavidades nasales de la mayoría de las especies, siendo importante hacer hincapié en el hecho de que cada neurona olfativa expresa sólo un gen OR entre todos los existentes (Chess A. et al., 1994 & Malnic B. et al., 1999).

Algunos genes OR en mamíferos son expresados en las células espermatogenéticas, y recientes estudios indican que ellos tienen una función en la quimiotaxia espermática, es decir ayudan al espermatozoide en la localización del óvulo (Spehr et al., 2003).

La familia de receptores olfativos OR es una gran familia dentro de los genomas de mamíferos. En el ratón y la rata se han encontrado más de 1000 genes, mientras que en el genoma humano entre 500 y 800 hORs (Zozulya et al, 2001& Niimura & Nei, 2005b). Esta característica de la familia contrarresta con el hecho de que más de la mitad de estos son pseudogenes, es decir contienen mutaciones que impiden la generación de un OR completo o funcional (Rouquier et al, 1998). Por el contrario, todos o la gran mayoría de los genes OR de roedores son funcionales. Estas características ha permitido considerar que la pérdida de agudeza en el sentido del olfato de los mamíferos superiores, va ligada probablemente al hecho de que son menos dependientes de este sentido para su supervivencia (Zozulya et al, 2001).

Los receptores OR son, normalmente en un 30 y un 60% idénticos entre si en varias regiones de su secuencia, dado que muchas características específicas y necesarias para su función, están presentes en la mayoría o en todos los miembros de la familia OR (Zozulya *et al*, 2001). Un ejemplo de ello es la cisteina conservada dentro del dominio C-terminal que ancla el receptor a la membrana celular (Buck & Axel, 1991). La figura 3 ilustra las características peptídicas que se conservan en la proteína OR, mostrando una región central particularmente variable compuesta por las hélices transmembranales 4 y 5. Se conoce actualmente que este es el punto de unión de la molécula olorosa (Malnic et al, 1999).

Figura 3. Regiones conservadas y variantes en los receptores olfativos humanos.



Los receptores olfativos son miembros de la superfamilia de receptores 7TM. Los cilindros representan las posibles siete hélices transmembranales, los círculos claros muestran las regiones altamente conservadas y los círculos oscuros son las regiones variantes. Tomado de (Buck & Axel, 1991).

Una de las características más relevantes de los receptores OR es que pueden ser o no ser específicos para una molécula olorosa. Es decir la percepción del olor puede involucrar a un gran número de distintos receptores, cada uno capaz de asociarse con uno o con un pequeño número de olores, en este caso el cerebro puede distinguir cual receptor o cual neurona debe ser activada para permitir la discriminación entre diferentes estímulos olorosos (Buck & Axel 1991). Así pues el sistema olfativo emplea un código combinatorial para la detección de moléculas olorosas u odorantes (Malnic et al, 1999).

La comprensión de este mecanismo de percepción olfatoria implica la identificación de los receptores específicos por molécula, el análisis de la gran

diversidad de receptores, la especificidad del receptor y tal vez lo mas importante, entender los patrones o mecanismos de regulación y expresión en el epitelio olfativo (Buck & Axel 1991).

5.3 GENES Y PROTEÍNAS OR

Los genes OR están clasificados en dos clases, la clase I se especializa en el reconocimiento de olores solubles en agua y los de clase II en el reconocimiento de olores en el aire (Todd D. Taylor. et al, 2006). Los humanos poseen estos dos tipos de genes OR, pero el significado funcional de los genes OR clase I se desconoce (Niimura & Nei, 2003). Todos los OR genes tienen aproximadamente 310 codones de longitud, además no poseen intrones dentro de la región codificante (Buck & Axel, 1991 y Nef P. et al., 1992), característica que junto a los dominios extramembranales facilitan su identificación en el genoma (Zozulya et al, 2001).

En la actualidad se han encontrado aproximadamente 338 genes funcionales y 414 pseudogenes (Nimura & Nei, 2005a). Datos que aumentan o disminuyen de acuerdo a la investigación.

Los genes OR en mamíferos están típicamente organizados en agrupamientos de decenas o más miembros, y localizados en muchos cromosomas o en la gran mayoría de ellos (Ben et al., 1994 y Reed R. 1992). En el ser humano se conoce actualmente que en el cromosoma 11 se encuentran agrupados la gran mayoría de los genes OR y que tal vez a partir de este agrupamiento cromosómico divergieron los demás genes OR hacia los otros cromosomas (Glusman et al, 2001).

Finalmente, podemos decir que recientemente ha comenzado una carrera por descifrar los mecanismos de regulación de los genes para la percepción de olores, dada la gran coplejidad de este sentido y las implicaciones estructurales que el conocimiento de estos mecanismos conlleva (Malnic et al, 1999). Esto a permitido solventar algunas características del sistema del olfato, ya que algunas hipótesis han favorecido el entendimiento de los mecanismos regulatorios que presentan los miembros de esta familia OR.

Una de las hipótesis mas certeras, es la arrojada tras la obtención experimental de regiones promotoras dentro de diferentes áreas del epitelio olfativo de rata, pues esto ha permitido afirmar que los genes OR que se encuentran en áreas iguales del epitelio olfativo, comparten igualmente motivos o regiones conservadas dentro de sus regiones promotoras. Sugiriendo así, que estos elementos regulatorios pueden contribuir para determinar que gen se expresa por neurona y en que área del epitelio olfativo (R. Hoppe, 2006). Esto además ha permitido encontrar que en

ratas, los mecanismos de regulación no son iguales para genes relacionados filogenéticamente y por tanto los receptores olfativos se regulan mediante patrones globales de control.

Las características anteriormente expuestas de la familia y genes OR forman parte de la caracterización Molecular, Fisiológica y Evolutiva de esta familia durante los últimos años. Estas características y regularidades dentro de la secuencia de los genes OR permiten la búsqueda de nuevos patrones y en otras regiones del gen como es el caso de las regiones promotoras, debido a que esto implica un menor esfuerzo en cuanto a la correlación entre estudios del gen, la familia OR y el trabajo computacional.

A continuación se expone las estrategias de estudio de la familia de genes hOR dentro del presente trabajo. La cual ha permitido correlacionar todas o la gran mayoría de las características expuestas anteriormente.

5.4 BIOINFORMÁTICA, BUSQUEDA Y ANÁLISIS DE SECUENCIAS

El presente trabajo pretende establecer diferentes y variadas relaciones entre la secuencia establecida de los genes funcionales hOR y su localización cromosómica, lo cual implica establecer o caracterizar a cada uno de los genes a nivel filogenético, topológico, conocer la estructura en cada una de sus regiones y estudiar las características estructurales de los genes dentro del genoma humano. Esta labor no es posible realizarla mediante métodos estadísticos comunes por el gran volumen de información con la que se cuenta, es por ello que necesitamos acudir a diferentes herramientas y métodos en donde el poder de procesamiento de los computadores debe ser utilizado a niveles diferentes al de escritura o entretenimiento.

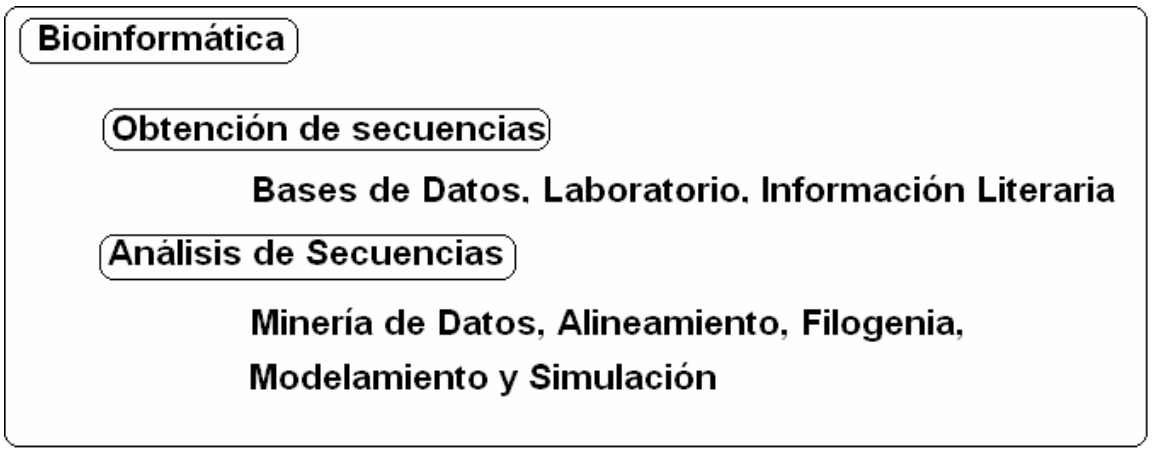
Esta relación entre herramientas informáticas y biología molecular se conoce como bioinformática y se manifiesta principalmente en la aplicación de diferentes algoritmos (procedimientos estructurados) computacionales, en la búsqueda, predicción, organización y visualización de genes, proteínas y genomas (Pevsner, 2003); Con el propósito de establecer relaciones estructurales, funcionales y de filogenia dentro de secuencias biológicas, que permitan a su vez descubrir nuevos genes, proteínas, funciones, descifrar la complejidad de los genomas y plantear fortalecidas hipótesis acerca de los mecanismos de regulación y expresión que rigen la formación y funcionamiento de un ser vivo.

Actualmente, la gran mayoría de investigaciones en donde el análisis de un gran número de secuencias biológicas es el eje principal, las herramientas bioinformáticas son el apoyo ideal para desarrollarlas, debido a que estas han sido creadas bajo un fuerte fundamento biológico y matemático con el fin de plantear

hipótesis biológicamente aceptables.

El análisis de secuencias biológicas, implica buscar, conocer y entender las posibles relaciones estructurales, funcionales o de filogenia que se presentan dentro de los genomas, ya que a partir de este conocimiento es posible establecer y entender los posibles mecanismos moleculares que se encuentra dentro de las secuencias de genes o proteínas. Este análisis se desarrolla bajo diferentes estrategias bioinformáticas, en donde la búsqueda, predicción e incluso visualización esta mediada por diferentes programas, algoritmos, scripts, etc. El presente trabajo no es ajeno a ello, por tal razón a continuación presentamos de forma clara las estrategias y herramientas utilizadas en el presente trabajo y bajo la estructura metódica en bioinformática (figura 4), la cual es particular para la gran mayoría de investigaciones en bioinformáticas y pretende ser una guía para el lector.

Figura 4. Estructura metódica en bioinformática



La bioinformática implica la búsqueda y obtención de secuencias biológicas con el propósito de analízalas mediante la aplicación de diferentes herramientas computacionales (Pevsner, 2003).

5.4.1 Bases de Datos y Obtención de Secuencias

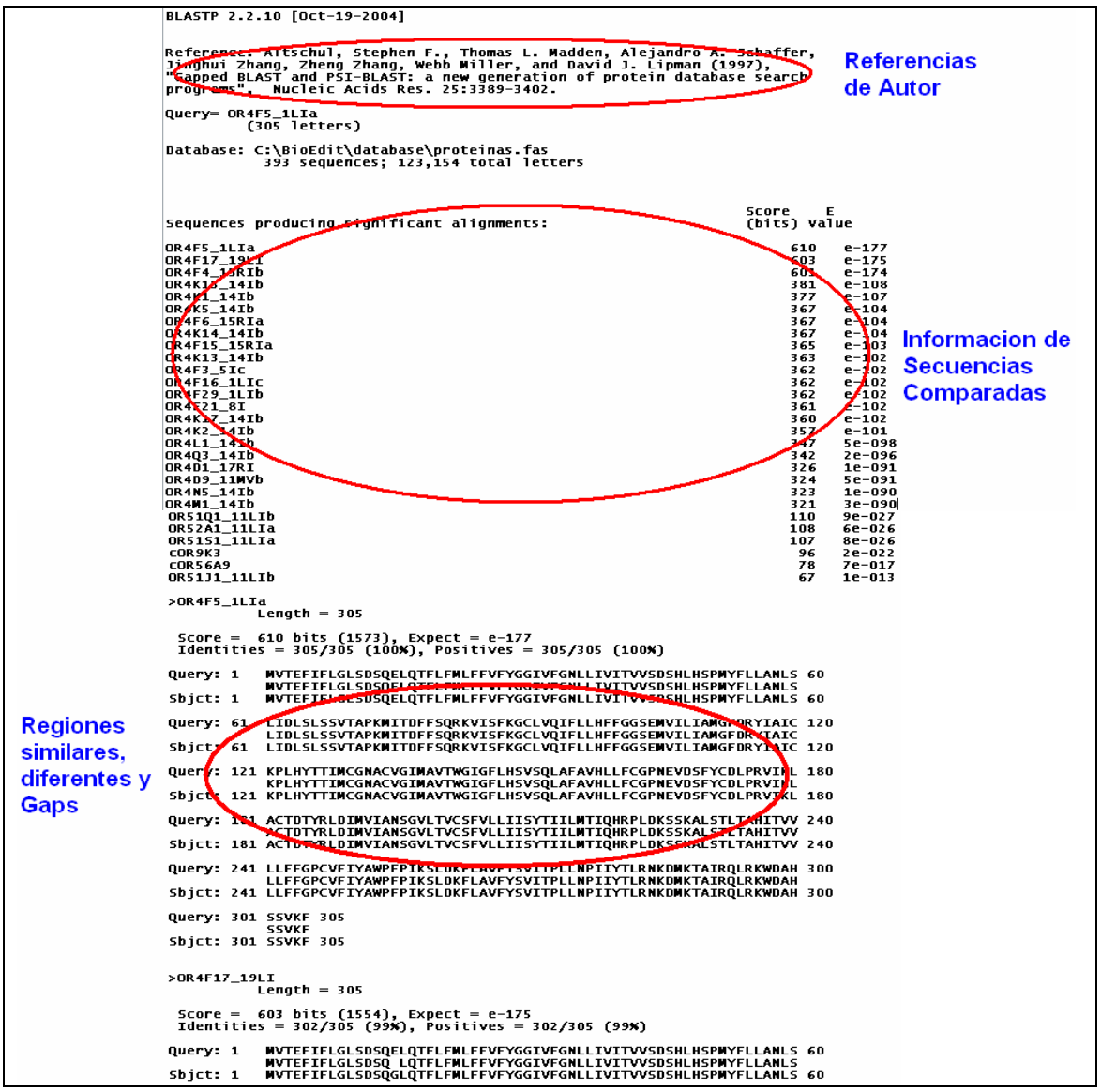
Las bases de datos son la base y el producto de todas las herramientas bioinformáticas actuales (Dopazo & Valencia, 2001), las cuales almacenan secuencias biológicas o análisis diferentes de ellas dentro de computadores distribuidos en todo el mundo y su acceso se realiza mediante el Internet, siendo así el primer recurso al cual se acude en la búsqueda de secuencias o información biológica. Es difícil enumerar las diferentes bases de datos que existen, pero hay tres que son de gran importancia por ser los más grandes almacenadores de datos biológicos públicos: el Genbank perteneciente al Centro Nacional de Información Biotecnológica (NCBI por sus siglas en ingles), el EMBL perteneciente

al Instituto Europeo de Bioinformática (EBI por sus siglas en ingles) y la base de datos de DNA de Japón (DDBJ por sus siglas en ingles). La información biológica almacenada en estas y en otras bases de datos esta organizada de tal forma que puedan ser localizados con facilidad mediante un identificador global (ID), el cual conduce a la secuencia en si determinada y a toda información asociada a ella. Es importante mencionar que esta información biológica puede ser usada de forma libre, directamente en las páginas Web de cada sitio o mediante la extracción de los datos para ser utilizados en los ordenadores o computadores propios.

Todo estudio a nivel bioinformático comienza con la obtención de secuencias de interés dentro de bases de datos, lo cual se realiza mediante la obtención directa de todas las secuencias de una base de datos, o mediante la búsqueda especifica mediante un Query (secuencia con información desconocida o de interés) dentro de ella. La búsqueda mediante un Query, se refiere al hecho de buscar dentro de las bases de datos información asociada a una secuencia cuyas características funcionales o estructurales deseamos conocer, a partir de la similitud de ésta con las secuencias que se encuentran dentro de las bases de datos. Esta búsqueda se realiza generalmente mediante la herramienta de alineamiento local BLAST (Altschul et al, 1997) y permite conocer de manera general las secuencias que mas se asemejan al Query.

El resultado visual del BLAST es una plantilla de texto, en donde se muestran las referencias de autor, seguido de información relevante a los genes encontrados dentro de la base de datos y las regiones conservadas frente a cada gen tal como lo muestra la figura 5.

Figura 5. Resultado del BLAST




```
BLASTP 2.2.10 [Oct-19-2004]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= OR4F5_1LIa
       (305 letters)

Database: C:\BioEdit\database\proteinas.fas
          393 sequences; 123,154 total letters


                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

OR4F5_1LIa                                                          610    e-177
OR4F17_19LI                                                         603    e-175
OR4F4_15RIb                                                         601    e-174
OR4K15_14Ib                                                         381    e-108
OR4K1_14Ib                                                          377    e-107
OR4K5_14Ib                                                          367    e-104
OR4F6_15RIa                                                         367    e-104
OR4K14_14Ib                                                         367    e-104
OR4F15_15RIa                                                        365    e-103
OR4K13_14Ib                                                         363    e-102
OR4F3_5Ic                                                           362    e-102
OR4F16_1LIc                                                         362    e-102
OR4F29_1LIb                                                         362    e-102
OR4F21_8I                                                           361    e-102
OR4K17_14Ib                                                         360    e-102
OR4K2_14Ib                                                          357    e-101
OR4L1_14Ib                                                          347    5e-098
OR4Q3_14Ib                                                          342    2e-096
OR4D1_17RI                                                          326    1e-091
OR4D9_11MVb                                                         324    5e-091
OR4N5_14Ib                                                          323    1e-090
OR4M1_14Ib                                                          321    3e-090
OR51Q1_11LIb                                                        110    9e-027
OR52A1_11LIa                                                        108    6e-026
OR51S1_11LIa                                                        107    8e-026
cOR9K3                                                               96    2e-022
cOR56A9                                                              78    7e-017
OR51J1_11LIb                                                         67    1e-013

>OR4F5_1LIa
          Length = 305

 Score =  610 bits (1573), Expect = e-177
 Identities = 305/305 (100%), Positives = 305/305 (100%)

Query: 1   MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS 60
           MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS
Sbjct: 1   MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS 60

Query: 61  LIDLSLSSVTAPKMITDFFSQRKVISFKGCLVQIFLLHFFGGSEMVILIAMGFDRYIAIC 120
           LIDLSLSSVTAPKMITDFFSQRKVISFKGCLVQIFLLHFFGGSEMVILIAMGFDRYIAIC
Sbjct: 61  LIDLSLSSVTAPKMITDFFSQRKVISFKGCLVQIFLLHFFGGSEMVILIAMGFDRYIAIC 120

Query: 121 KPLHYTTIMCGNACVGIMAVTWGIGFLHSVSQLAFAVHLLFCGPNEVDSFYCDLPRVIKL 180
           KPLHYTTIMCGNACVGIMAVTWGIGFLHSVSQLAFAVHLLFCGPNEVDSFYCDLPRVIKL
Sbjct: 121 KPLHYTTIMCGNACVGIMAVTWGIGFLHSVSQLAFAVHLLFCGPNEVDSFYCDLPRVIKL 180

Query: 181 ACTDTYRLDIMVIANSGVLTVCSFVLLIISYTIILMTIQHRPLDKSSKALSTLTAHITVV 240
           ACTDTYRLDIMVIANSGVLTVCSFVLLIISYTIILMTIQHRPLDKSSKALSTLTAHITVV
Sbjct: 181 ACTDTYRLDIMVIANSGVLTVCSFVLLIISYTIILMTIQHRPLDKSSKALSTLTAHITVV 240

Query: 241 LLFFGPCVFIYAWPFPIKSLDKFLAVFYSVITPLLNPIIYTLRNKDMKTAIRQLRKWDAH 300
           LLFFGPCVFIYAWPFPIKSLDKFLAVFYSVITPLLNPIIYTLRNKDMKTAIRQLRKWDAH
Sbjct: 241 LLFFGPCVFIYAWPFPIKSLDKFLAVFYSVITPLLNPIIYTLRNKDMKTAIRQLRKWDAH 300

Query: 301 SSVKF 305
           SSVKF
Sbjct: 301 SSVKF 305


>OR4F17_19LI
          Length = 305

 Score =  603 bits (1554), Expect = e-175
 Identities = 302/305 (99%), Positives = 302/305 (99%)

Query: 1   MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS 60
           MVTEFIFLGLSDSQ LQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS
Sbjct: 1   MVTEFIFLGLSDSQGLQTFLFMLFFVFYGGIVFGNLLIVITVVSDSHLHSPMYFLLANLS 60
```

La búsqueda de secuencias biológicas se hace dentro de bases de datos mediante programas de búsqueda como el Blast local. El resultado visual del Blast es una plantilla de texto versátil que muestra la comparación hecha entre el query y las secuencias de la bases de datos.

Hasta este punto el resultado del BLAST es igual para cada una de las bases de datos que lo aplican en la búsqueda de secuencias, pero la versatilidad de la plantilla ofrecida por BLAST es lo que diferencia las bases de datos, es decir, los diseñadores o programadores de bases de datos pueden a partir de esta plantilla crear vínculos, eliminar información redundante o tal vez lo que marca más a las

bases de datos, ofrecer a cada gen encontrado una puerta (link) que conlleve a la caracterización del gen como tal, como es el caso del NCBI que ofrece información detallada de cada gen encontrado en su base de datos (GenBank) tras la realización de la búsqueda, con solo hacer un *click*. Del mismo modo, cada base de datos ofrece diferentes opciones de búsqueda y resultado con ayuda del programa Blast, ya que este reúne diferentes algoritmos de alineamiento como el Blastp y Blastn para proteínas y nucleótidos respectivamente y por tanto podemos aplicar diferentes parámetros para la realización de un mejor análisis. Así pues, es importante conocer los principios básicos del alineamiento para comprender la forma cómo ésta es una de las estrategias de estudio más relevantes dentro de la bioinformática.

5.4.2 Alineamiento de Secuencias

Cuando se conoce una secuencia de un gen o una proteína, siempre surgirá una pregunta fundamental: ¿Que otros genes o proteínas están relacionados y en qué proporción se relacionan?

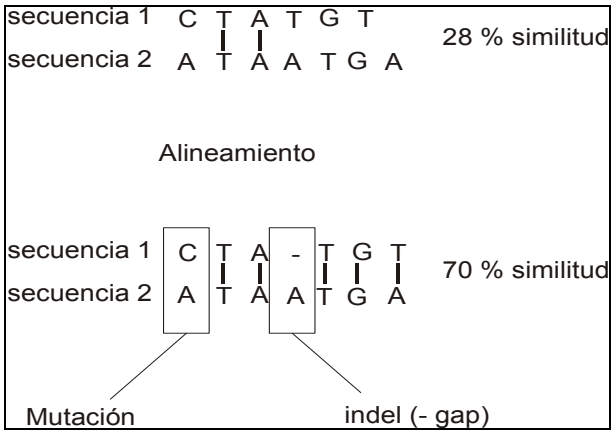
Esta pregunta es muy importante porque a partir de su respuesta podemos descifrar funciones, establecer estructuras, definir dominios y motivos que agrupen nuestros genes y proteínas en conjuntos ya definidos dentro de cualquier genoma. El primer paso para ello, se realiza generalmente a través de la comparación entre secuencias mediante el alineamiento, el cual es un proceso que compara dos o más secuencias para lograr niveles máximos de identidad (y conservación, en el caso de secuencias de aminoácidos), evaluando el grado de similitud y la posibilidad de homología entre secuencias. (Pevsner, 2003).

El alineamiento, indica el número de nucleótidos o aminoácidos que se conservan y no se conservan entre las secuencias de estudio (Nei & Kumar, 2000). Los sitios conservados son un parámetro importante dentro de la biología, ya que podemos definir dominios con implicaciones funcionales para la célula o estructurales en el genoma; al igual que los sitios que no se conservan, debido a que estos sitios representan los posibles eventos evolutivos por los cuales han divergido otras secuencias, divergencia que se puede dar mediante procesos de sustitución (mutaciones) o procesos de inserción o eliminación (indels) (Pevsner, 2003).

Ya que las mutaciones como la substitución no pueden ser predichas mediante algoritmos bioinformáticos, la inserción y eliminación se convierte en los parámetros biológicos que el algoritmo informático utiliza para el alineamiento, esto quiere decir que las inserciones o deleciones de nucleótidos o aminoácidos dentro de una secuencia pueden ser establecidas y cuantificadas por medio de algoritmos informáticos y por tanto ofrecer un significado biológico. Así pues, el algoritmo de alineamiento realiza este proceso mediante la introducción de

espacios (gaps) dentro de las secuencias, de tal forma que la introducción de gaps dentro de las regiones de baja similitud aumenta la similitud en otras regiones (Harold P. 2004). La figura 6, representa el alineamiento de un par de secuencias con los gaps y los principios básicos del mismo.

Figura 6. Principios básicos del alineamiento



El alineamiento implica la introducción de gaps dentro de secuencias para aumentar la similitud entre ellas.

El alineamiento es una estrategia que se desarrolla de diferentes formas o mediante diferentes algoritmos, siendo la forma de clasificación más sencilla la que los clasifica de acuerdo al número de secuencias que alinea, así pues este se puede clasificar en alineamiento simple (Pairwise) y Alineamiento múltiple, los cuales son utilizados en el presente trabajo y descritos a continuación.

5.4.2.1 Alineamiento Simple.

El alineamiento simple es la comparación de dos secuencias de aminoácidos o nucleotidos, en donde cada cambio o alteración es valorado bajo parámetros biológicamente aceptables (Pevsner, 2003). Este es un método efectivo para la búsqueda de secuencias en bases de datos, ya que a partir de la similitud entre secuencias podemos establecer posibles homologías y por tanto diferentes características estructurales y funcionales. El alineamiento simple se puede realizar mediante diferentes métodos y diferentes algoritmos computacionales, que valoran la similitud de las secuencias mediante una serie de matrices de sustitución que tratan de establecer el mejor emparejamiento mediante los valores de cada cambio o igualdad de las secuencias.

Los algoritmos mas conocidos son el de Needleman & Wunsch (1970), el cual alinea globalmente las dos secuencias penalizando (evaluando) cada cambio de forma dinámica, es decir cada cambio altera la matriz creada por el algoritmo. El

algoritmo de Smith y Waterman (1981) es un algoritmo utilizado en alineamiento simple, pero es utilizado cuando se busca el segmento mejor alineado dentro de un par de secuencias; es por tal motivo que se conoce como algoritmo para alineamiento local.

Ambos algoritmos crean matrices de sustitución que buscan capturar el significado biológico de las semejanzas en las secuencias considerando los cambios conservativos y diferencias en las secuencias conservadas de los mismos. Por lo tanto la puntuación que se le ofrece a cada cambio depende de la frecuencia evolutiva del mismo. Las matrices mas utilizadas son las PAM (Daykoff et al, 1978) y las BLOSUM (Steven Henikoff y Jorja G. Henikoff, 1992), las cuales difieren en la estrategia que se utiliza para la valoración de los cambios, ya sea por comparaciones globales de cada cambio en el caso de las PAM, o por comparación y agrupación de bloques conservado en el caso de las BLOSUM.

El alineamiento de un par de secuencias, puede realizarse mediante la herramienta de búsqueda de alineamiento local BLAST (Altschul et al, 1997), la cual establece la similitud de las secuencias a partir del algoritmo de Smith y Waterman (1981) y amplía la búsqueda a fragmentos poco iguales o con emparejamientos que estadísticamente tengan una baja probabilidad de ocurrir, permitiendo exceptuar los emparejamiento productos del azar. Como se mencionó esta herramienta es ideal para la búsqueda de secuencias en las bases de datos.

5.4.2.2 Alineamiento Múltiple.

En la comparación mediante alineamiento de secuencias de ADN y Proteínas, se observa que estas presentan regiones semejantes y diferentes, de acuerdo a la posición establecida durante el alineamiento. Estás regiones permiten afirmar que se parecen porque tienen un origen común y difieren porque a lo largo del tiempo los genes han divergido mediante la acumulación de cambios o mutaciones (Abascal, 2003). Estas comparaciones han permitido establecer homologías (origen común) y analogías (origen diferente) de las secuencias de los genes y proteínas, siendo la homología la que permite clasificar a los genes y proteínas en una jerarquía de superfamilias, familias y subfamilias (Henrissat & Romeu, 1995; Murzin et al., 1995; Yona et al., 1999). Así pues, el alineamiento múltiple es utilizado en la búsqueda de homologías y en la reconstrucción filogenética, ya que por lo general las secuencias con iguales características estructurales poseen una misma función (Pevsner, 2003).

El alineamiento múltiple podemos definirlo como la comparación de tres o más secuencias mediante la inserción de gaps dentro de ellas, con el fin de emparejar los residuos (regiones homologas) en posiciones estructurales comunes y/o los residuos hereditarios en la misma columna. Cuando un grupo de secuencias es

alineado, es posible predecir la función y estructura de los miembros del grupo además de asociarlos a complejos celulares (Pevsner, 2003).

El alineamiento múltiple es un problema computacional complejo dado el volumen de información que se pretende analizar y la incertidumbre en la penalización de gaps. Para solventar esto se han desarrollado diferentes métodos que buscan de forma eficiente alinear un gran número de secuencias con una inserción de gaps correcta o a aproximada a los cambios que han sufrido las secuencias. El método mas utilizado en la actualidad es el desarrollado por Feng y Doolittle (1987), conocido también como "*alineamiento progresivo de secuencias*". Progresivo, debido a que se rige por una serie de pasos que involucran vincular secuencia a secuencia al alineamiento. Este método comienzan con el alineamiento local (Pairwise) entre todas las secuencias, luego selecciona las dos secuencias de mayor similitud e involucra una tercera secuencia, la cual será la que mejor alineé con el primer par, y de esa manera se continúa hasta que todas las secuencias hayan sido incluidas (Pevsner, 2003). Se puede pensar que este método es la continuidad del alineamiento simple, pero la verdad es algo más complejo, dado que involucra estimar distancias evolutivas entre secuencias, a partir de las matrices de puntuación que crea el alineamiento simple. Estas distancias son necesarias para la construcción de un árbol guía, el cual es el que permite el orden en el cual las secuencias se involucran en el alineamiento, dando paso a la penalidad de gaps.

El algoritmo mas utilizado en el alineamiento múltiple bajo el método de alineamiento progresivo de secuencias, es ClustalW (Thompson, J.D., 1994) el cual aplica algoritmos filogenéticos en la construcción del árbol guía y permite penalizar la inserción y extensión de gaps. Parámetros que convierten a este algoritmo en uno de los más efectivos.

El alineamiento múltiple es uno de los primeros pasos en la búsqueda de regularidades dentro de grupos de genes, dado que sobre el resultado de este alineamiento son aplicados gran cantidad de algoritmos bioinformáticos que permiten reconstruir la historia evolutiva de genes, definir motivos y dominios funcionales y descifrar la estructura de genes y proteínas entre otros (Feng & Doolittle, 1987).

5.4.3 Análisis Filogenético

La inferencia filogenética o análisis filogenético de secuencias, busca establecer la maquinaria o proceso evolutivo de los Genes o Proteínas, mediante diferentes métodos estadísticos y computacionales (Nei & Kumar, 2000), a partir de comparaciones estructurales previas como el alineamiento.

En si, el algoritmo filogenético se basa en una estructura estadística, en donde se valora el grado evolutivo de cada cambio o alteración en las secuencias de estudio; cada algoritmo posee como base un componente (ecuación) estadístico que es interpretado por los computadores como una instrucción de selección y exclusión entre miles de datos biológicos, ofreciendo como resultado una serie de valoraciones e interpretaciones que se ven reflejadas en los árboles filogenéticos que se crean (Nei & Kumar, 2000). Estos árboles, se caracteriza porque la longitud de las ramas son proporcionales a la distancia evolutiva entre las secuencias o al tiempo de divergencia entre ellas. Los nodos corresponden a la secuencia ancestral a partir de la cual derivaron las secuencias incluidas en ese nodo. Los grupos que cumplen con esto, se denominan monofiléticos y cada uno de ellos se denomina taxa o clado (Pevsner, 2003).

El concepto de "árbol filogenético" es una visión algo limitada: suponer un grupo monofilético, es equivalente a suponer que las secuencias que conforman un clado descendieron directamente del ancestro común, y se diferencian de él solo por procesos mutacionales que acumularon substituciones. Es decir, los procesos biológicos como recombinación, duplicación y evolución paralela (paralogía), transferencia horizontal (xenología), o selección natural, no se reflejan claramente en este tipo de representaciones (Harold P. 2004)[1].

Para reconstruir un árbol filogenético a partir de un alineamiento, primero es necesario determinar un dendrograma, el cual es creado automáticamente tras el alineamiento múltiple, luego se comienza calculando las relaciones entre las secuencias, hasta tener una estructura de clados que agrupe pares de grupos (si algún nodo se divide en más de dos ramas, se dice que el nodo no está resuelto). Luego de tener esta estructura, se calculan las distancias moleculares entre las secuencias, y se ajusta por algún algoritmo a la longitud de las ramas que mejor ajuste a las distancias de todos los pares de secuencias. Los principales métodos y algoritmos utilizados en el análisis filogenético de secuencias, son los de parsimonia, de alta verosimilitud y los de distancias de secuencias; los cuales son aplicados de acuerdo a la similitud que estas poseen (Nei & Kumar, 2000).

5.4.4 Herramientas Bioinformáticas para el Estudio de la Familia hOR

Parte del presente estudio implica aplicar sobre las secuencias hOR, algunos programas con el propósito de visualizar y analizar la familia de una mejor forma. Estos programas reúnen los resultados del alineamiento, de árboles filogenéticos y de la minería de datos, permitiendo asociar y plantear hipótesis.

Es importante comprender que dentro de la bioinformática, pueden ser utilizados

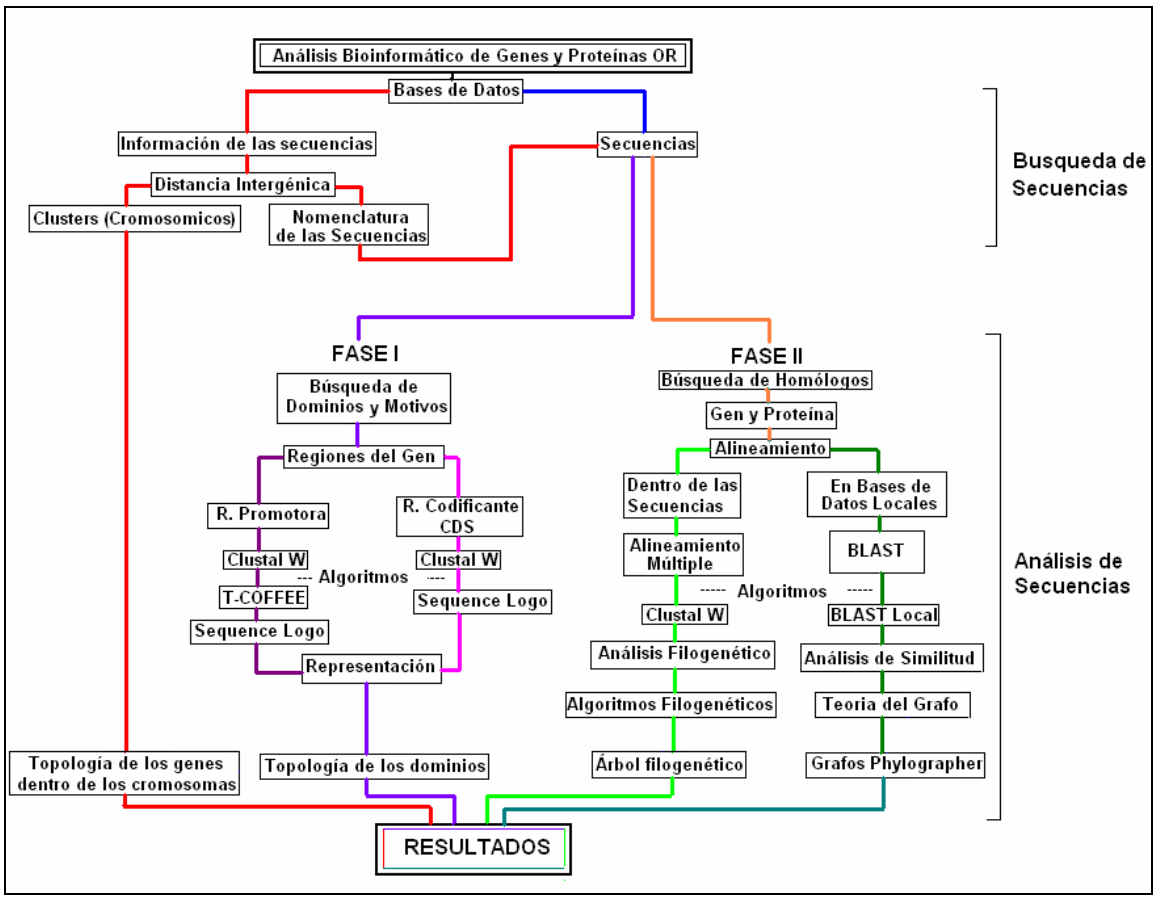
[1] Harold P. de Vladar. (2004). Introducción a la Bioinformática/ http://www.idea.org.ve/cristalograf/hpvladar/BioinformaticaIDEA.pdf

una gran cantidad de herramientas y algoritmos sin límites. Es por tal motivo que a medida que los resultados del estudio se vayan presentando, también aparecerán herramientas bioinformática para su análisis y la descripción de cada una.

6. MATERIALES Y MÉTODOS

La estructura metodológica de la presente investigación se rige bajo el siguiente esquema (figura 7), el cual es propuesto como estrategia de estudio de la familia de receptores olfativos humanos.

Figura 7. Estructura metodológica de la Investigación



6.1 BÚSQUEDA Y OBTENCIÓN DE GENES OR HUMANOS

El primer paso para el estudio de la familia OR, es la obtención de todos los genes funcionales, con sus regiones promotoras y con toda la información asociada a ellos. Esto se hace a partir de la búsqueda de y en bases de datos, tal como se presenta a continuación.

6.1.1 Búsqueda y Selección de Bases de Datos.

Para la búsqueda de la base de datos se realizó un sondeo intensivo mediante la Internet de toda información relacionada a la familia hOR, al igual que bases de datos creadas tras investigaciones en esta y que pudieran albergar las secuencias de los genes que codifican receptores olfativos humanos (hOR) e información detallada de cada uno de ellos.

La selección de la base de datos se realizó bajo los siguientes parámetros:

a. Facilidad de búsqueda dentro de la base de datos
b. Calidad y tipo de información, en donde definimos calidad como la relevancia de la información para nuestro estudio.
c. Las referencias bibliográficas que involucran y dan validez a la misma.

La anterior metodología permitió seleccionar a la base de datos The Human Olfactory Receptor Data Exploratorium (HORDE)[2] para la obtención de las secuencias de estudio. Selección que es explicada en el análisis de resultados.

6.1.2 Búsqueda Dentro de la Base de Datos HORDE

La búsqueda dentro de la base de datos implicó, obtener toda la información necesaria para el trabajo, la cual se representa en la secuencia en sí de los genes y la información más relevante tras su secuenciamiento.

Los siguientes pasos metodológicos se realizaron en la fecha **13/09/2005** de la siguiente forma:

a. Se selecciono de las secuencias presentes en la base de datos de HORDE, un gen prototipo que cumplirá con las características propias de un receptor olfativo, principalmente con longitud de 310 codones de longitud, la presencia del codón ATG en la ubicación definida bibliográficamente y que no contara con regiones desconocidas dentro de las regiones codificantes del gen o CDS. Lo anterior ofreció como resultado un grupo de genes óptimos para el estudio, de los cuales fue seleccionado aleatoriamente el del gen OR2C3, presente en el cromosoma 1 como **Query** para la búsqueda dentro de la base de datos.

[2] Olender T, Feldmesser E, Atarot T, Eisenstein M, Lancet D., The olfactory receptor universe--from whole genome analysis to structure and evolution. Genet Mol Res. 2004 Dec 30;3(4):545-53

b. El query es llevado a la base de datos HORDE en su sitio web www.weizmann.ac.il/HORDE , para la búsqueda mediante BLAST del resto de secuencias pertenecientes a la familia OR.

c. Tras realizar la búsqueda del query en la base de datos mediante BLAST, el resultado es una plantilla con gran cantidad de información referente al Query y las secuencias semejantes dentro de la base de datos. Para lo cual seleccionamos de esta información al gen más semejante al Query.

d. La selección del gen más semejante al Query nos lleva a la caracterización topológica y molecular del gen dentro de otro vínculo, el cual además de esta información, contiene un enlace directo con el centro de notación de genomas "*UCSC Genome Bioinformatics*". Nos dirigimos hacia este vínculo e ingresamos hacia la opción de búsquedas *Table Browser*.

e. Dentro del vínculo *Table Browser* buscamos dos tipos de archivo, el primero, las secuencias en formato (fasta)[3] de los genes hOR y el segundo con las características de la familia en formato (xls). Esto se realiza mediante las herramientas de búsqueda propias de este vinculo, en este caso las tablas "*sequence*" para la secuencia de los genes y "*all fields from selected table*" para la información de ellos.

f. Para la obtención de las secuencias con las regiones no codificantes, se flanquearon (delimitaron) todos los genes, 500 nucleótidos por encima del comienzo de la región codificante, es decir sobre el extremo 5' del sitio de comienzo de la trascripción o TSS. Mediante las opciones de búsqueda de la bases de datos.

g. Se creó una copia de los genes obtenidos, para su posterior trascripción a proteínas y poder continuar con los siguientes pasos de la metodología.

6.1.3 Obtención de los hOR Funcionales.

Es de aclarar que los resultados de la búsqueda que ofrece la base de datos HORDE, abarca todos los genes secuenciados o predecidos hasta la fecha de obtención, incluyendo genes funcionales, pseudogenes, y genes con secuencias incompletas o desconocidas. Es por tal motivo que se realizó un tratamiento u optimización en los datos, en donde se eliminaron los genes incompletos y los pseudogenes, con la ayuda de *algoritmos* y *revisión manual*.

La obtención de los genes funcionales OR se hizo a partir de la identificación y

[3] Tipo de archivo que almacena secuencias. Se compone de un signo mayor que (>), seguido por el nombre de la secuencia y los aminoácidos o nucleótidos que componen la secuencia.

eliminación de los pseudogenes, para lo cual se siguió los siguientes pasos de la metodología de Niimura & Nei, 2003:

a. Se tradujeron las secuencias de ADN a Proteína, mediante algoritmos computacionales, debido a que el número de caracteres a revisar visualmente (aminoácidos), es menor que en secuencias de ADN (nucleótidos).

b. Se eliminaron las secuencias con longitudes iguales o inferiores a 250 aminoácidos. Esto se debe a que este número no alcanza para que la proteína (hOR) atraviese la membrana celular 7 veces.

c. Se realizó un pre-alineamiento de las secuencias en busca de las regiones conservadas de los receptores OR, principalmente las regiones comunes necesarias para conservar la función. Estas regiones fueron determinadas por Buck & Axel (1991) y Zozulya y colaboradores (2001).

d. Se eliminan las secuencias que contengan deleciones o inserciones de más de 3 aminoácidos en los dominios más conservados.

e. Se eliminan las secuencias que no comiencen con el aminoácido metionina.

f. De acuerdo al identificador global de las secuencias, tomamos de una de las copias los genes OR (punto 5.1.1.g), las secuencias de ADN correspondientes para cada proteína que quedan después del anterior proceso.

Las secuencias que se eliminan después de los pasos anteriores, se consideran Pseudogenes y las secuencias que quedan se consideran Genes funcionales (Niimura & Nei , 2003).

Los anteriores pasos metodológicos se resumen en la obtención de 2 tipos de archivos, el primero las secuencias de genes y proteínas OR funcionales y el segundo el que contiene toda la información de cada una de estas secuencias. Los siguientes pasos de la metodología, muestran las estrategias que hemos seguido para la búsqueda de nuevas regularidades dentro de la familia de genes y proteínas OR a partir de estos dos tipos de archivos.

6.2 ANÁLISIS DE LA FAMILIA OR HUMANA

Tras la obtención de los genes OR e información de ellos, se puedo dar paso al estudio como tal de la familia, lo cual implico definir una nomenclatura propia, analizar la información ofrecida por la base de datos, comparar todos los genes y proteínas, definición de regiones conservadas, dominios y motivos; y finalmente alinear y reconstruir la filogenia de los genes. Los siguientes pasos representan las estrategias seguidas en el estudio de la familia OR.

6.2.1 Definición de Agrupamientos Cromosómicos y Nomenclatura de Secuencias.

Los agrupamientos cromosómicos es una regularidad que se observó tras el secuenciamiento del genoma humano, en donde los genes homólogos o muy semejantes se encuentran dentro de una misma región cromosómica. Los genes hOR no son una excepción a esta característica por lo cual los siguientes pasos indican la forma como se definieron estos agrupamientos dentro de cada uno de los cromosomas:

a. Mediante la información obtenida en la base de datos HORDE, principalmente la tabla "*all fields from selected table*", se realizó una nueva tabla o **Tabla de datos hOR** con los genes seleccionados anteriormente, clasificando los genes OR por cromosoma y ubicación dentro del genoma humano.

b. Se calculó la distancia entre los genes de acuerdo a su posición, en cada uno de los cromosomas en donde se encuentran, mediante la siguiente formula:

Distancia intergénica (pb) = (Comienzo del gen B (pb) - fin del gen A (pb)) - 1

En donde (pb) es pares de bases, B es el gen que sigue en forma sucesiva al gen A, y A es el primer gen que se encuentra en el cromosoma de acuerdo a su posición.

c. Se agruparon los genes, de acuerdo a la distancia intergénica, tomando en cuenta intervalos de distancia en Kpb (kilo pares de bases) y Mpb (Mega pares de bases) por agrupamiento.

d. Se definió la nomenclatura de los genes a partir de los agrupamientos encontrados y se le asigno a las secuencias de ADN y Proteína.

6.2.2 Análisis de las Secuencias Fase I: Búsqueda de Regiones, Dominios y Motivos

En el presente trabajo se caracterizó la estructura de los genes olfativos funcionales y completos es decir el gen con todas sus regiones, tanto codificantes (CDS) como promotoras. Esta caracterización implico definir regiones conservadas dentro de las secuencias, dominio, motivos o toda característica asociada a las secuencias y que pudieron ser definidas mediante la siguiente estrategia.

Como un primer paso para ello, se crearon dos archivos a partir del de secuencias de genes, uno que contenía las regiones promotoras y otro que contenía las regiones codificantes CDS. Esto es necesario para realizar una búsqueda más puntual de regularidades.

6.2.2.1 Caracterización de regiones codificantes CDS

Las regiones codificantes de los genes hOR, poseen grandes motivos que pueden ser caracterizados mediante el alineamiento tal como se indica en los siguientes pasos:

a. Se realizó un alineamiento múltiple de las secuencias mediante el algoritmo ClustalW y luego mediante el algoritmo T-COFFEE bajo los siguientes parámetros:

```
htobar@bioinfo:~$ t_coffee -in CDS.fas -special_mode=dna -outfile=outaln
```

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*) haga una alineamiento mediante (*t_coffee*) con las secuencias (*-in*) CDS.fas que son (*-special_mode*=) de dna y genere un archivo (*-outfile*=) alineado (*outaln*).

b. El alineamiento hecho se visualizó en el programa *BioEdit* como imagen pixelada y dentro del programa *Indonesia* como representación “Sequence Logo”. se identificaron los motivos y regiones característicos de los miembros de la familia hOR.

6.2.2.2 Caracterización de la Región Promotora

La variabilidad de las secuencias no intrónicas o no codificantes como lo son las promotoras, dificulta la búsqueda y definición de dominios mediante algoritmos comunes, es por tal motivo que para caracterizar estos fue necesario realizar gran

cantidad de alineamientos seguidos de una inspección visual constante.

a. Se realizó un pre-alineamiento de todas las secuencias mediante el algoritmo Clustal W.

b. Se realizó una inspección visual y se seleccionan grupos de genes que presentaron regiones en común.

c. Se realizó un nuevo alineamiento de cada uno de los grupos obtenidos, con el resto de genes y se realiza una nueva selección de grupos.

d. Se repitió el anterior proceso, hasta que quedara un solo grupo o se consolidarán fuertemente los demás.

e. Se realizó un alineamiento de cada grupo y juntos, mediante ClustalW para la búsqueda de regiones en común.

f. Dado que persistieron grupos independientes de secuencias, se realizó el alineamiento anterior, pero esta vez con el algoritmo T-COFFEE.

g. Dado que los grupos no mostraban asocio o regiones conservadas en común, los dominios y regiones de cada grupo fueron graficados y definidas sus características, es decir por separado.

6.2.3 Análisis de las Secuencias Fase II: Análisis de Similitud y Búsqueda de Homólogos

El análisis de similitud y la búsqueda de homólogos implica asociar a cada uno de los genes y proteínas hOR con todos los miembros de la familia a partir de sus características filogenéticas y de similaridad. Para tal fin se siguieron dos estrategias de búsqueda.

6.2.3.1 BLAST y Análisis de Grafos

El BLAST realizado ahora difiere al utilizado anteriormente en dos aspectos: 1) la base de datos en la cual se realizó la búsqueda, esta conformada por los 385 miembros de la familia hOR, 2) El Query, son las mismas 385 secuencias; en otras palabras, realizamos una búsqueda de los genes y proteínas hOR dentro de nuestra propia base de datos de hOR. Esta estrategia se conoce como BLAST local y se realizó mediante los siguientes pasos:

a. Se creó una base de datos local mediante los siguientes parámetros

htobar@bioinfo:~$ formatdb -p T -o T -i secuencia.fas -n db

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*), cree una base de datos (f*ormatdb*) de proteínas (*-p*) verdadero (*T*) o falso (*F)*, creando datos extras (*-o*) verdadero (*T*) a partir (*-i*) del archivo (*secuencias.fas*) y nombrar esta base de datos (*-n*) como (*db*)

b. Se comprueba la creación correcta de la base de datos si dentro de los archivos creados se encuentra el archivo *formatdb.log* y la búsqueda de cualquier secuencia dentro de ella. Esto se realizó mediante los siguientes parámetros :

htobar@bioinfo:~$ fastacmd -d db -s secuenciaX

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*) haga una búsqueda (*fastacmd*) en la base de datos (*-d*) creada (*db*) la secuencia (*secuenciaX*)

c. Luego se realizó el BLAST local de las secuencias mediante los siguientes parámetros:

htobar@bioinfo:~$ blastall -p blastp -d db -i secuencias.fas -m 0 -o listo

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*) realice una búsqueda (*blastall*) mediante un blast tipo (*-p*) proteínas (*blastp*), dentro de la base de datos (*-d*) db, de las secuencias (*-i*) secuencias.fas y ofrezca como resultado una matriz (*-m*) típica de blast (*0*) llamada (*-o*) listo.

d. Se eliminó la redundancia de información que posee el archivo creado mediante:

-La aplicación del algoritmo Blast2Phylopix bajo los siguientes parámetros:

htobar@bioinfo:~$ tclsh Blast2Phylopix_024.tcl

Enter the SOURCE file name: listo
Enter the DESTINATION file name: filtrado
extract DESCRIPTION line (Y/N): N
type of BLAST search was blast(n) blast(x) blast(p): n
Extract strand direction info (y)es/(n)o: n
normalize EXP value (Y/N): n

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*)

invocar (*tclsh*) el algoritmo Blast2Phylopix y ejecutarlo bajo los parámetros deseados.

-La aplicación del script phylopix_redundancy, de la siguiente forma:

htobar@bioinfo:~$ perl phylopix_redundancy.pl filtrado sin_re 0.2

Esta instrucción quiere decir dentro del servidor bioinfo (*htobar@bioinfo*) invocar (*perl*) el script phylopix_redundancy, eliminar la redundancia del archivo (*filtrado*) y crear una archivo (*sin _re*).

e. Los resultados del BLAST local, fueron llevados al programa Phylographer, para visualizar las asociaciones entre secuencias mediante grafos.

6.2.3.2 Alineamiento Múltiple y Análisis Filogenético

El alineamiento múltiple al igual que el análisis filogenético implican la aplicación de diferentes algoritmos y herramientas computacionales sobre las secuencias, a continuación se describe la forma como estos fueron aplicados y los parámetros impuestos a cada uno:

a. Se realizaron alineamientos múltiples de las secuencias mediante el algoritmo Clustal W y bajo los siguientes parámetros:

   - Se realizo un Test de inserción de Gaps para la realización de un correcto alineamiento de las secuencias. para ello en los parámetros *Gap Open* y *Gap Extend* del algoritmo, se ingresaron los valores de penalización de gaps, asociados al grado de homología de las secuencias de acuerdo a Wiley-LISS (2003).

   - Tras cada alineamiento se definió dentro del dominio mas conservado de las secuencias, el porcentaje de similitud.

   - Este porcentaje junto con los valores ingresados, fueron graficados, analizados y seleccionados los mejores parámetros para el alineamiento.

b. El alineamiento hecho fue llevado al programa *MEGA 3.1* siguiendo las instrucciones propias de este, frente a las secuencias alineadas, es decir se seleccionó el tipo de dato, tipo de análisis, tipo de visualización, entre otros.

c. Establecidas las secuencias alineadas en el programa, se dio paso a la

construcción de los árboles filogenéticos mediante los métodos de distancia y parsimonia, como son: UPGMA, Evolución Mínima (ME), Neighbor Joining (NJ) y Máxima Parsimonia (MP).

- La distancia genética se calculó mediante el método de *Kimura 2-parameter*, para las secuencias de ADN y para proteínas el método de *Poisson correction*; esto exceptuando a MP.

- Se realizo un Test de exactitud en los árboles, mediante el método Bootstrap, bajo 100 replicaciones o muestreos, y repeticiones al azar por defecto del método.

- Se seleccionó el árbol filogenético correcto bajo parámetros discutidos mas adelante, se señalaron los diferentes clados del árbol filogenético (agrupamientos filogenéticos) y se marcan para el posterior análisis.

7. RESULTADOS Y DISCUSIÓN

El desarrollo de la anterior metodología ha permitido establecer una gran cantidad de características estructurales, asociaciones e hipótesis acerca de la estructura de los genes, su filogenia, dominios y mecanismos de regulación asociados a motivos. Esto, junto con todo el análisis que se presenta a continuación representa algunos de los alcances que puede llegar a tener una investigación bioinformática.

A continuación se presentan los resultados ofrecidos por la anterior metodología, y se discuten sus alcances.

7.1 BÚSQUEDA Y OBTENCIÓN DE LOS GENES OR HUMANOS

Durante los últimos años la familia de receptores olfativos ha sido muy estudiada por las características fisiológicas y moleculares que la gobierna. Estos estudios han arrojado igualmente una gran cantidad de información almacenada en bases de datos, a la cual se puede acceder de forma libre. Esto ha llevado a que dentro del presente trabajo no solo se haya realizado la búsqueda de los genes OR, si no de la base de datos correcta para su obtención; siendo ésta, la mejor de entre las creadas tras cada estudio sobre la familia OR.

Ahora bien, dos de las bases de datos encontradas más trascendentales son la de Niimura y Nei (2003) y Zozulya y colaboradores (2001), por almacenar el mayor número de genes OR humanos y funcionales. Como era de esperarse, cada una de estas pequeñas bases de datos, organizaban los genes o los clasificaba de forma diferente, por lo tanto fue necesario buscar información adicional que permitiera relacionar o asociar estas bases de datos con el fin de trabajar sobre un grupo de secuencias estándar e igual para toda la comunidad científica. Esta búsqueda nos condujo a dos nuevas bases de datos que se caracterizan principalmente por su actualización constante, es decir en ellas se encuentran el total de genes OR que hasta la fecha están secuenciados. *Human Olfactory Receptor Data Exploratorium* (HORDE)[4] y *Olfactory Receptor DataBase* (ORDB)[5], en ellas las secuencias biológicas son almacenadas bajo una nomenclatura internacional que fue definida bajo la base de 4 publicaciones que buscaron

[4] Olender T, Feldmesser E, Atarot T, Eisenstein M, Lancet D., The olfactory receptor universe--from whole genome analysis to structure and evolution. Genet Mol Res. 2004 Dec 30;3(4):545-53.

[5] Crasto C..,Marenco L., Miller P.L., and Shepherd G.S. (2002) Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Research1:354-360

organizar esta familia de genes incluyendo la publicación del genoma humano. (Zozulya et al, 2001; Glusman et al, 2001; Nimura & Nei 2003; Venter et al, 2001). Teniendo en cuenta las características definidas para la selección de la base de datos, se seleccionó de entre estas dos, la base de datos HORDE, debido principalmente a que esta pertenece a un centro de notación genómica, en donde se organiza la información de diferentes bases de datos en una serie de tablas de búsqueda, el "*UCSC Genome Bioinformatics*" perteneciente al "*Center for Biomolecular Science & Engineering*". La notación que ofrece este centro implica conocer entre otros, el tamaño de los genes, la hebra en donde se encuentran (5' o 3'), el cromosoma, desde que nucleótido comienza y termina y por su puesto obtener la secuencia de los genes deseados.

Ahora bien, tras la selección de la base de datos HORDE como la ideal, para nuestro estudio se procedió a la obtención de las secuencias e información de ellas, tal como se mencionó en la metodología. Esto ofreció como resultado la obtención de 852 genes de receptores olfativos hOR, los cuales incluyen todo tipo de gen secuenciado, predecido u asociado a la familia hOR. Se debe mencionar que gran parte de estos genes son bioinformáticos, donde su definición parte de estrategias computacionales de búsqueda o predicción. Estos genes incrementan el total de genes de la base de datos y aunque no necesariamente todos son pseudogenes, estos pueden producir errores en la predicción y selección de los verdaderos genes funcionales determinados experimentalmente.

Para evitar estos errores, fue necesario realizar una "purificación o limpieza" a los 852 genes, lo cual indica eliminar todos o la gran mayoría de los pseudogenes y genes con secuencias incompletas o desconocidas. Para ello se buscó dentro de la secuencias de los 852 genes, codones de parada que cortaran la síntesis de la proteína, secuencias incompletas, genes sin codon de iniciación y otros parámetros mencionados en la metodología. La figura 7 muestra la estructura de las secuencias buscadas y definidas como pseudogenes.

Figura 7. Definición de pseudogenes

Gen funcional:	
Secuencia estándar	**ATGCCTACTGTAAACCACAG final**
Pseudogenes:	
Con secuencias incompletas:	**ATGCCTACTGTAAAC final**
Con secuencias desconocidas:	**ATGCCTACTGTNNNNCACAG final**
Sin codon de iniciación:	**CAGCCTACTGTAAACCACAG final**

La obtención de los genes funcionales OR se hizo a partir de la identificación y eliminación de los pseudogenes. Las letras azules representan la normalidad dentro de las secuencias y las letras rojas representan los errores que definen un pseudogen.

El anterior proceso condujo a la eliminación de 467 pseudogenes, alrededor del 55 %, y por ende la selección de 385 genes funcionales OR o el 45 %. Se destaca que no existen genes funcionales dentro de los cromosomas 4, 13, 18, 20, 21 y Y, debido probablemente al hecho de fueron eliminados tras el tratamiento de los datos o que son genes funcionales, pero están fuera de los parámetro que hemos tenido en cuenta para seleccionarlos. Sin embargo, Zozulya y colaboradores (2001) mediante un trabajo experimental, estimaron que no existía en los cromosomas 2, 4, 18, 20, 21 y Y. Lo cual representa que en dicho trabajo no encontraron genes dentro del cromosoma 2, como sí lo hemos evidenciado nosotros, de igual forma no se puede decir que sobe el cromosoma 13 se encuentran genes que nosotros no hemos encontrado. Por otra parte Niimura y Nei (2003) mediante una investigación bioinformática estimaron que no existía genes funcionales dentro de los cromosomas 20 y Y, lo cual da certeza de lo expuesto. La tabla 1 muestra estas relaciones, la cual permite decir que nuestro resultado esta más relacionado con el trabajo experimental desarrollados por Zozulya y colaboradores (2001), pero ambos se ven soportados por lo encontrado por Niimura y Nei (2003).

Tabla 1. Cromosomas sin genes funcionales

Investigación	**Cromosomas sin genes funcionales**						
Experimental: Zozulya y colaboradores (2001)	2	4		18	20	21	Y
Bioinformática: Niimura y Nei (2003)					20		Y
Bioinformática: Nosotros (2006)		4	13	18	20	21	Y

La relación entre los datos ofrecidos por diferentes publicaciones, sustentan el correcto desarrollo del presente trabajo, además de permitir generar discusión a partir de ellos.

Ahora bien, en cuanto al total de genes funcionales nuestros resultados van de la mano con todas las publicaciones actuales, pues el promedio definido de genes funcionales hOR es de 350 (Crasto et al 2002). Cabe destacar que Niimura y Nei (2005b), bioinformáticamente determinaron un total de 388 genes y 414 pseudogenes, entonces podemos tener certeza que la cantidad de genes funcionales seleccionados corresponde a un valor promedio de genes funcionales. El incremento del total de pseudogenes con respecto a Niimura y Nei (2005b) se debe probablemente a la predicción de genes bioinformáticos o al secuenciamiento total de algunos cromosomas humanos como el 11 y 17.

7.2 ANÁLISIS DE FAMILIA OR HUMANA

El análisis de la familia OR humana parte del estudio de la información que sobre la secuencia ofrece la base de datos, para luego tener claridad en el análisis de la secuencia en sí, de cada gen o proteína. Es por tal motivo que fue organizada toda la información de la secuencia dentro de una **Tabla de Datos** hOR (anexo A),

con el propósito de conocer las características de cada gen. Esta tabla permitió establecer entre otros la hebra en donde se encuentran, el tamaño, la distancia intergénica etc. (Tabla 2). Estas características hacen parte de la caracterización estructural de los genes, las cuales son expuestas a continuación.

Tabla2. Parte de la matriz creada mediante información ofrecida por la base de datos

Cr	Nombre-Agrupamiento	Distancia Intergénica	Hebra	Comienzo	Fin	ID	Tamaño Gen	DI Mb
1	OR4F5_1Lia	58953	+	58953	59868	OR4F5	915	**0,059**
1	OR4F29_1Lib	347653	+	407521	408457	OR4F29	936	0,348
1	OR4F16_1Lic	252504	-	660961	661897	OR4F16	936	0,253
1	OR10T2_1Mia	154519490	-	155181387	155182329	OR10T2	942	154,519
1	OR10K2_1Mia	20464	-	155202793	155203729	OR10K2	936	0,020
1	OR10K1_1Mia	44695	+	155248424	155249363	OR10K1	939	0,045
1	OR10R2_1Mia	13410	+	155262773	155263745	OR10R2	972	0,013
1	OR6Y1_1Mia	66248	-	155329993	155330968	OR6Y1	975	0,066
1	OR6P1_1Mia	14548	-	155345516	155346467	OR6P1	951	0,015
1	OR10X1_1Mia	15317	-	155361784	155362711	OR10X1	927	0,015

La presente tabla hace parte de la elaborada para la familia hOR y se encuentra dentro de los anexos. La información ofrecida por la base de datos permitió calcular la distancia intergénica, una nomenclatura propia, el tamaño de los genes y los promedios entre ellos.

7.2.1 Definición de Agrupamientos Cromosómicos y Nomenclatura de Secuencias

Para establecer agrupaciones cromosómicas de la familia OR, se partió de la distancia que existe entre los genes de cada cromosoma. Esta distancia se conoce como Distancia Intergénica (DI), la cual se calcula mediante la información que ofrece la base de datos en cuanto a la ubicación de los genes y la ecuación expuesta en la metodología (punto 5.2.1.b). La DI permite conocer la topología de los genes dentro de los cromosomas, identificar posibles agrupamientos dentro de ellos y definir una nomenclatura basada en la distancia que existe entre los genes.

Nosotros hemos calculado la DI para cada gen funcional de la familia hOR, la cual permitió definir una nomenclatura y una gran cantidad de agrupamientos dentro de los cromosomas. Esta nomenclatura se basó en la distancia de los genes dentro de los cromosomas, para lo cual se tuvo en cuenta que: L(left), M(Medium) y R(right), definen regiones en donde la distancia entre genes es de mas de 1000

Kb, los números romanos (I, II, III, etc) definen a los genes que se encuentran a distancias de 100 kb y por ultimo las letras (a,b,c, etc) definen a los genes que se encuentran a distancias de 10kb. Así pues se pudo establecer la nomenclatura de los genes de la siguiente forma:

OR_1LIa				
OR	1	L	I	a
Gen o Proteína OR	Cromosoma	D*≥1000kb	D≥100kb	D≥10kb

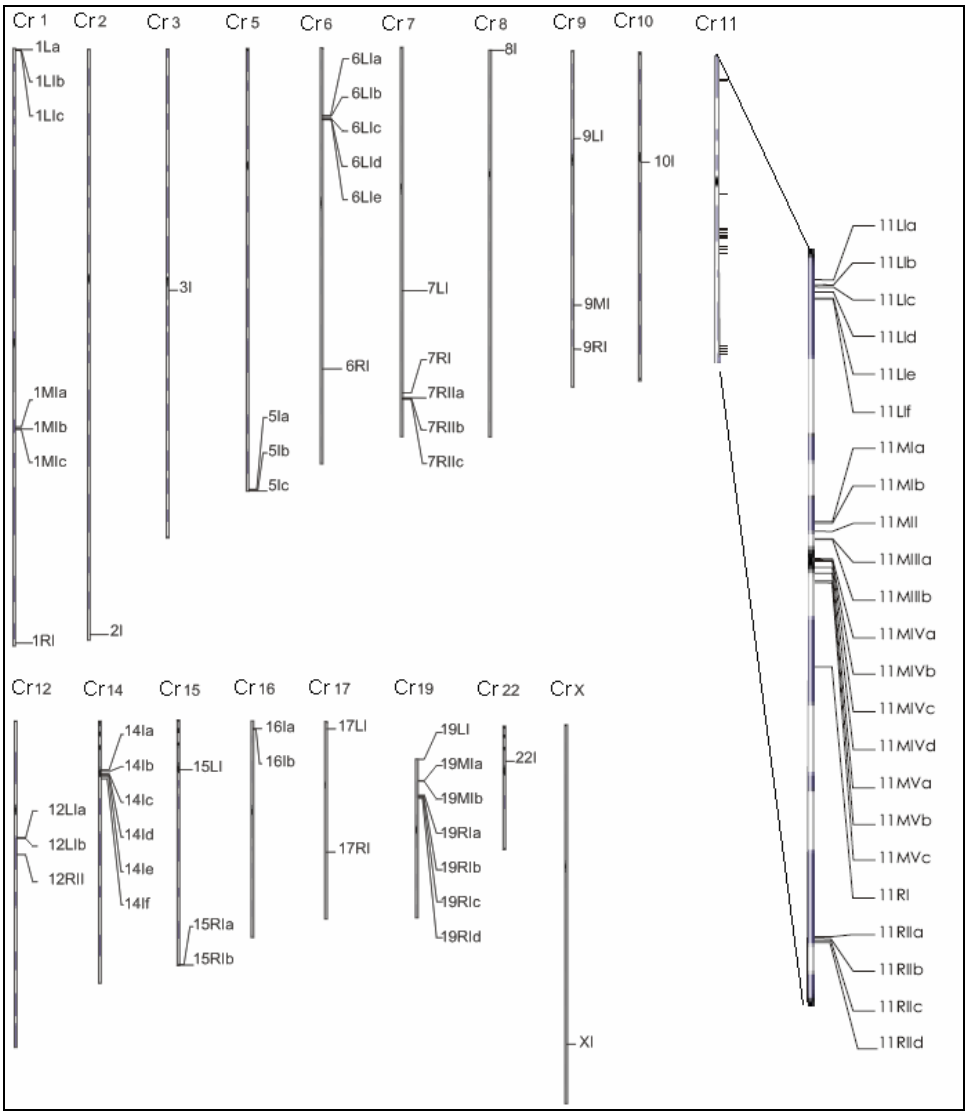
*Distancia

Los agrupamientos encontrados, los que llamaremos ahora **Agrupamientos Cromosómicos (AC)** se determinan igualmente por la distancia que existe entre los genes de un cromosoma de tal forma que la nomenclatura establecida va de acuerdo a los agrupamientos encontrados dentro de la familia de receptores olfativos hOR así:

OR_1LIa				
OR	1	L	I	a
Gen o Proteína OR	Cromosoma	Región	Agrupamiento	Sub-Agrupamierno

Estos Agrupamientos hacen parte de la topología de los genes dentro de los cromosomas la cual es necesaria para conocer los mecanismos por los cuales los genes de la familia hOR se han distribuido sobre el genoma. La figura 8 muestra la disposición de los genes hOR dentro de los cromosomas, reconstruida a partir de la distancia intergenica. La figura permite observar las tres regiones definidas por (R, M, L), los Agrupamientos definidos por (I, II, III, etc) y los sub-agrupamientos definidos por (a, b, c, etc).

Figura 8 Distribución de los genes sobre los cromosomas “Agrupamientos Cromosómicos”

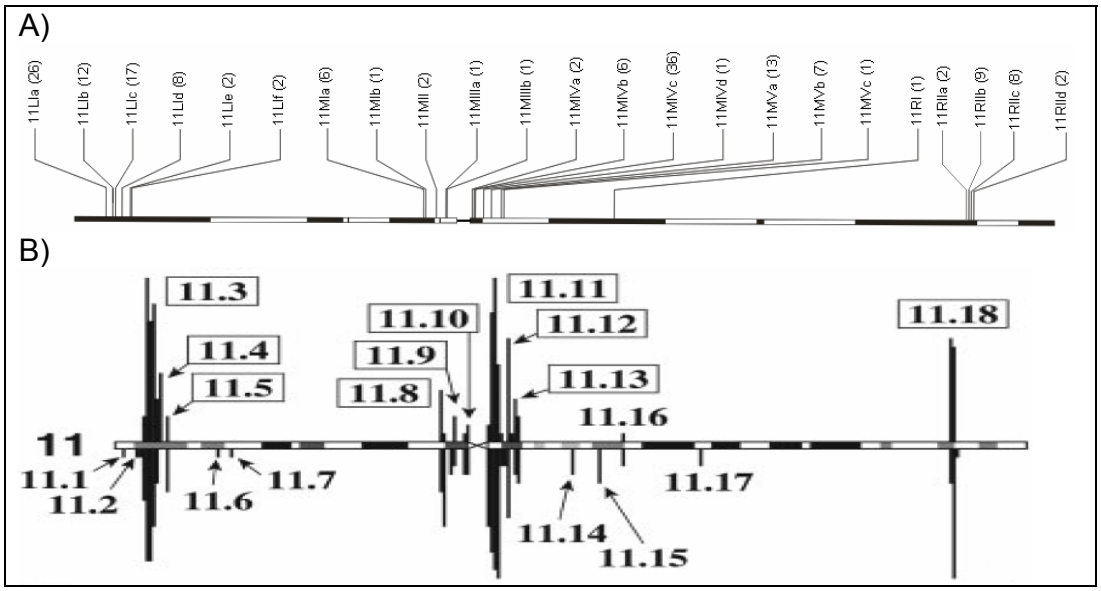


La distribución de los genes dentro de los cromosomas es necesaria en el conocimiento de la historia evolutiva de la familia hOR. La figura representa la topología de genes hOR a manera de agrupaciones cromosómicas. Las líneas indican la ubicación, L(left), M(Medium) y R(right), definen regiones en donde la distancia entre genes es de más de 1000 Kb, los números romanos (I, II, III, etc) definen genes que se encuentran a distancias de 100 kb y por ultimo las letras (a,b,c, etc) definen genes que se encuentran a distancias de 10kb. Se resalta el cromosoma 11 por contener el mayor número de los genes ROH.

La figura anterior presenta solo los cromosomas en donde los genes hOR se encuentran, siendo importante destacar que sobre el cromosoma 11 se encuentra el mayor número de genes hOR y es utilizado para compararlo con los trabajos

realizados por los investigadores Niimura y Nei (2003), tal como se muestra en la figura 9.

Figura 9. Agrupamientos cromosómicos del cromosoma 11



La disposición de los genes dentro de los cromosomas en forma de racimos se conoce como agrupamientos cromosómicos o genómicos. A) Agrupamientos presentes en el cromosoma 11 creados en el presente trabajo, las líneas indican la ubicación, los números y letras son el nombre del agrupamiento y el número entre paréntesis es el número de genes por agrupamiento. B) Cromosoma 11 creado por Niimura y Nei (2003). Las líneas indican la ubicación, la longitud indica el número de genes y el número indica la identificación del agrupamiento.

La figura 9, nos permite apreciar que la topología de los genes hOR dentro de los cromosomas es igual para ambos trabajos, lo cual da certeza de que la estrategia usada para conocer la disposición de los agrupamientos dentro de los cromosomas es correcta y por tanto puede ser utilizada para los análisis posteriores.

La distribución de los genes dentro de todos los cromosomas puede resumirse mediante la Tabla 3 "***Tabla Resumen***", en donde se destaca de una forma mas detallada el número de regiones, agrupaciones, sub-agrupaciones y número de genes para cada cromosoma en donde se encuentran genes funcionales hOR.

Tabla 3. ***Tabla Resumen***. Agrupaciones cromosómicas

Cr	Regiones	Agrupaciones por región	Sub-Agrupaciones por agrupamiento	N° DE GENES	Total genes
1	L	I	a	1	63
			b	1	
			c	1	
	M	I	a	13	
			b	1	
			c	2	
	R	I		44	
2		I		2	2
3		I		10	10
5		I	a	1	4
			b	2	
			c	1	
6	L	I	a	2	16
			b	1	
			c	4	
			d	7	
			e	1	
	R	I		1	
7	L	I		1	15
	R	I		1	
		II	a	2	
			b	8	
			c	3	
8		I		1	1
9	L	I		2	25
	M	I		9	
	R	I		14	
10		I		1	1
11	L	I	a	26	166
			b	12	
			c	17	
			d	8	
			e	2	
			f	2	
	M	I	a	6	
			b	1	
		II		2	
		III	a	1	
			b	1	
		IV	a	2	
			b	6	
			c	36	
			d	1	
		V	a	13	
			b	7	
			c	1	
	R	I		1	
		II	a	2	
			b	9	
			c	8	
			d	2	
12	L	I	a	1	17
			b	1	
	R	II		15	
14		I	a	1	23
			b	16	
			c	1	
			d	1	
			e	3	
			f	1	
15	L	I		2	5
	R	I	a	2	
			b	1	
16		I	a	1	2
			b	1	
17	L	I		11	13
	R	I		2	
19	L	I		1	20
	M	I	a	1	
			b	7	
	R	I	a	5	
			b	1	
			c	4	
			d	1	
22		I		1	1
X		I		1	1

La tabla 3, permite afirmar que la mayor parte de los genes hOR se encuentran dentro del cromosoma 11 es decir, 166 genes los cuales representan el 56 % del total de genes funcionales. Esta característica también fue encontrada por (Todd D. et al 2006) tras la secuenciación total del cromosoma 11.

Se destaca igualmente la ausencia de regiones (L, M, R) dentro de los cromosomas 2, 3, 5, 8 y 22, esto debido a que todos los genes se encuentran dentro de una misma región y por lo tanto no hay genes a más de 1000kb de distancia. Esta característica se refleja igualmente en los sub-agrupamientos, dado que la ausencia de ellos representa que no hay genes a una distancia de 10kb. Además es importante resaltar el número de genes, dado que esto permitirá al reconstruir del árbol filogenético, definir los sitios dentro de los cromosomas donde se han dado los procesos de expansión de la familia. Por lo pronto, podemos decir que en promedio el número de genes por agrupamiento es 10 y para sub-agrupamientos 4, valores que nosotros hemos establecidos y son necesarios para confrontar la posible historia evolutiva de los genes.

El cromosoma 14 es un cromosoma destacable respecto al número de genes que alberga y la forma como estos se distribuyen dentro de él, ya que todos se encuentran sobre una región y en varios sub-agrupamientos abarcando sus 23 miembros, lo que constituye un número grande de genes en relación a los de más cromosomas. Este tipo de característica permite pensar en los posibles procesos de duplicación que la familia hOR ha sufrido o la divergencia de la familia dentro del genoma. Todas estas características topológicas de los genes, serán discutidas en el análisis final de resultados, dada la necesidad de reconstruir el árbol filogenético y de asociar características estructurales con las hipótesis planteadas.

7.2.2 Análisis de Secuencias FASE I: Búsqueda de Regiones, Dominios y Motivos

Los dominios y motivos son regiones altamente conservadas dentro de las secuencias de proteínas y su significado difiere con respecto a la estructura y función que tienen dentro de la célula, pues los dominios son regiones con funciones específicas y los motivos son regiones que no siempre son funcionales, pero permiten agrupar filogenéticamente a las secuencias que los contienen (Pevsner, 2003). Para determinar si las regiones conservadas son motivos o dominios se acude a las bases de datos especializadas que ofrecen un posible significado a las regiones encontradas o a programas bioinformáticos específicos como eMOTIF (Huang et al, 2001).

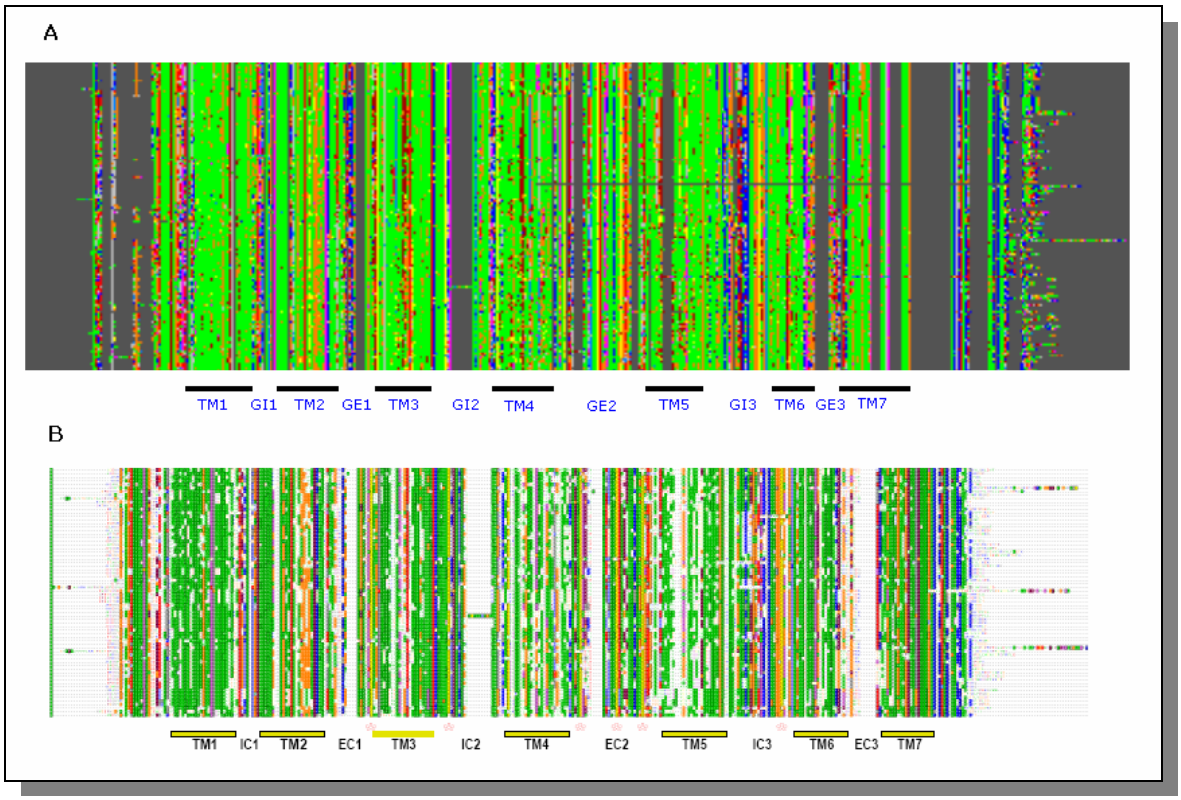
Dentro del presente trabajo se ha establecido la estructura del gen completo hOR, lo cual implica conocer las diferentes regiones conservadas que lo conforman, a partir de los dominios y motivos que se encuentran en la proteína y la búsqueda

de consensos en la región promotora del gen.

Como se mencionó anteriormente para poder llevar a cabo lo enunciado, fue necesario crear a partir de la secuencias de los genes olfativos hOR, dos tipos de archivos nuevos, uno con las regiones codificantes CDS y otro con las regiones promotoras del gen hOR; esto con el propósito de facilitar la búsqueda en cada una de las partes del gen.

Ahora bien, la estructura del receptor (proteína) hOR ya ha sido establecida en diferentes publicaciones, desde el descubrimiento del mismo (Buck & Axel, 1991), hasta la ultima caracterización estructural hecha (Zozulya et al, 2001). Esto nos permitió establecer la estructura de la región codificante (CDS) de forma más precisa, al confrontar los dominios establecidos bibliográficamente para proteínas, con los dominios y motivos establecidos en el presente trabajo (figura 10), para luego asociar estos dominios a las regiones conservadas de la secuencia de los genes.

Figura 10. Imagen pixelada de alineamiento múltiple de proteínas.



La estructura de la región codificante del gen hOR se ha establecido tras la búsqueda de dominios y motivos dentro de las secuencias de proteínas como los dominios transmembranales (TM) y los giros intra-extra celulares (GI o IC). A) Alineamiento Múltiple de Proteínas hecho en el presente trabajo B) Alineamiento Múltiple de Proteínas hecho por Zozulya y colaboradores (2001), cada punto de las imágenes representa un aminoácido.

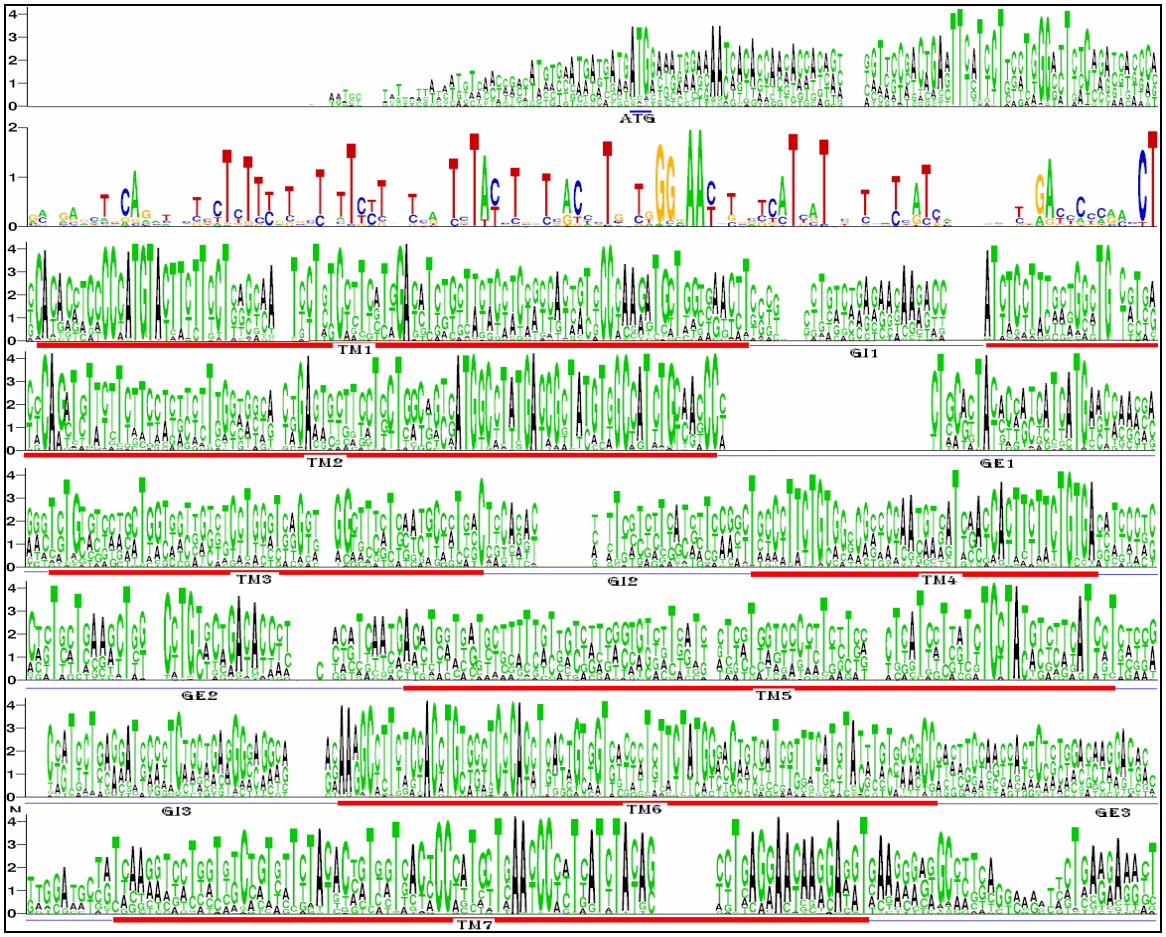
Este tipo de imagen pixelada permiten definir las características estructurales de las secuencias, ya que cada punto coloreado dentro de la gráfica representa un aminoácido o nucleótido y los espacios en gris son lugares sin secuencia o representan los gaps insertados; como se observa las dos imágenes son prácticamente iguales, lo cual indica que la búsqueda de dominios y motivos en las proteínas hOR puede resultar muy sencilla si se aplica el método y algoritmo correcto.

Se debe destacar que en los trabajos de Buck & Axel, (1991) y Zozulya et al, (2001) la estructura como tal del gen no se ha establecido o simplemente no se ha ilustrado, lo cual da trascendencia a la estructura del gen que aquí se presenta. Esta estructura partió de alineamientos hechos con los CDS de los agrupamientos cromosómicos 11LIa, 11LIb y 11LIc, alrededor de 55 secuencias; esto debido a que dentro de estos agrupamientos se encuentran los genes ancestrales de la familia y por ende el gen original cuya estructura al ser modificada dio origen al resto de miembros de la familia hOR (Glusman et al, 2001; Niimura & Nei, 2003).

Como se mencionó las regiones conservadas encontradas dentro de la región codificante, fueron confrontadas con los dominios establecidos para proteínas, lo cual da certeza de que los dominios aquí propuestos corresponden a verdaderas regiones de ADN con función conocida en la célula.

La figura 11, muestra la estructura de la región codificante del gen hOR, mediante una ilustración "*Sequence Logo*", establecida a partir del alineamiento múltiple de secuencias y la búsqueda de regiones conservadas mediante el algoritmo eMOTIF.

Figura 11. Estructura de la región codificante del gen hOR



Los receptores del olfato poseen en su estructura 7 dominios extramenbranales que los caracterizan. La figura es una representación “sequence logo” del alineamiento hecho. Cuanto mas grande las letras mayor es el porcentaje de su presencia.

El siguiente paso para la reconstrucción de la estructura del gen hOR, fue caracterizar la estructura del promotor dentro de la región 5’ del comienzo de la trascripción (TSS, por sus siglas en inglés), lo cual es el primer paso en la comprensión de los mecanismos moleculares que rigen la función del receptor olfativo, y que solo en este trabajo es propuesto.

La estructura de la región promotora se realizó mediante una gran variedad de alineamientos selectivos con ayuda de los algoritmos ClustalW (Thompson, J.D., 1994) y T-Coffee (Notre dame, et al 2000) tal como se mencionó en la metodología.

Esta estrategia de búsqueda intensiva, permitió observar que la región promotora

es muy variable en cuanto a la secuencia de los genes OR, lo cual impide generar árboles filogenéticos de toda la familia hOR a partir de estas regiones. En contraste a esto, se encontró una gran variedad de motivos estructurales que permitieron agrupar a las secuencias por regiones conservadas, además de algunos genes con copias idénticas de la región promotora.

Los motivos encontrados fueron clasificados en dos tipos: 1) Motivos Altamente Conservados (MAC) y 2) Motivos Ligeramente Conservados (MLC), los cuales se definieron de acuerdo al número de secuencias que cada uno agrupó. A si pues se consideró motivo altamente conservado a aquel que agrupó a 10 o más secuencias y motivo ligeramente conservado a los demás. Esta clasificación obedece al hecho de que una región conservada puede ser o no ser un motivo, dependiendo exclusivamente del número y tipo de secuencias que agrupa, si el número de secuencias es muy poco, estas regiones conservadas pueden ser el producto de la duplicación de los genes y no de su maquinaria regulatoria como si lo son los motivos altamente conservados dentro de las regiones promotoras.

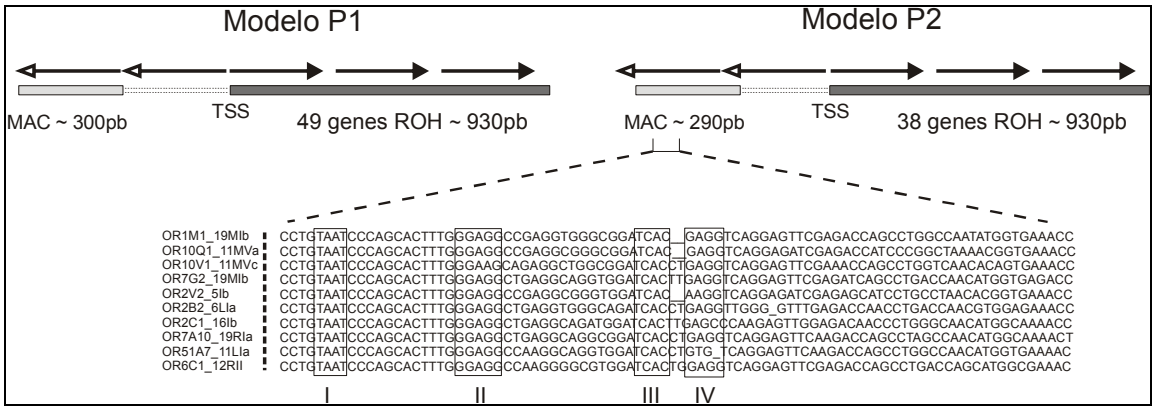
Esta estrategia dio como resultado la determinación de 2 Motivos Altamente Conservados y 25 Motivos Ligeramente Conservados, lo cual permite proponer 2 modelos estructurales de promotor (P1 y P2) y 25 rastros de duplicación de la familia o de modelos de regulación (Figura. 12). Para poder definir los MAC como modelos, se tuvieron en cuenta tres parámetros de evaluación basados en la longitud y ubicación de los MAC que componen cada modelo figura 12, pues en P1 esta longitud es de 300pb y se encuentra a -300pb en promedio; y el modelo P2 de 290pb ubicado a -250pb del sitio de comienzo de la trascripción (TSS, por sus siglas en ingles).

El primer parámetro involucró realizar una búsqueda de exones dentro de los genes olfativos humanos a través de minería de datos sobre la notación del genoma humano ofrecida por el GenBank, esto con el fin de demostrar la presencia o no, de más de un exón. Tras realizar esto se encontró que solo 5 genes poseen dos o más exones, los cuales no corresponden a los que conforman los modelos, favoreciendo los MAC como modelos.

La ubicación de los MAC fue un parámetro importante para aceptar el modelo, ya que los estudios realizados por Cooper et al, (2006), demuestran que los factores de regulación encontrados a esta distancia dentro del Genoma Humano, corresponden a elementos de control negativo del gen, característica que nosotros la hemos asociado al hecho de que solo un receptor se expresa por cada neurona del epitelio olfativo.

Por ultimo hemos encontrado que sobre el modelo P2 se encuentran factores de transcripción similares a los establecidos experimentalmente en regiones promotoras del gen olfativo de ratón (Hoppe, 2006), valorando la presencia dentro de estos motivos de factores de transcripción.

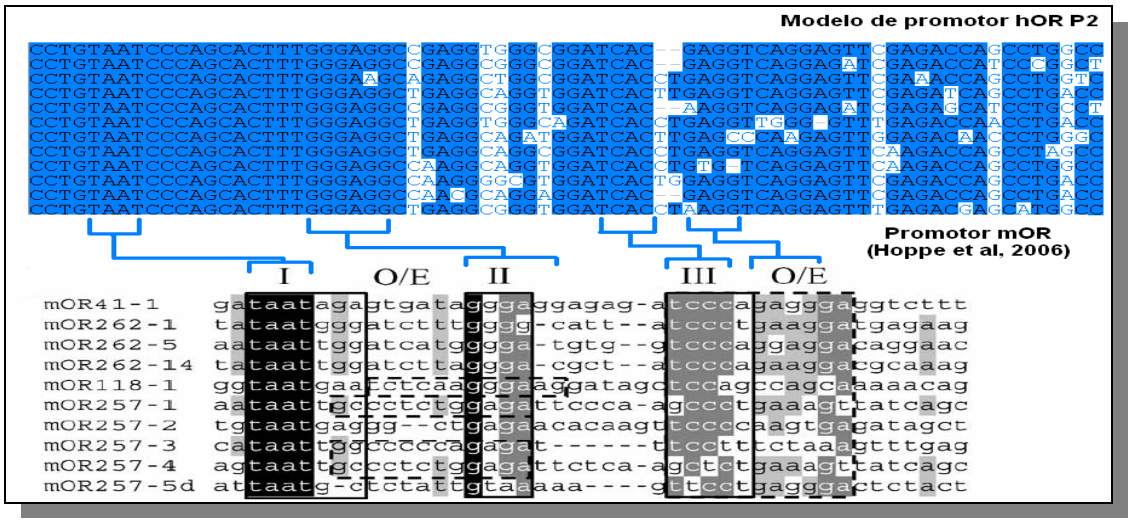
Figura 12. Esquema del gen ROH con los modelos estructurales de regulación encontrados P1 y P2



La distancia del modelo al punto TSS, indica una posible función de regulación negativa. Una Fracción del modelo P2 es ampliado, en donde es indicado los factores de transcripción I, II, III y IV, conocidos experimentalmente en genes hOR de ratón.

La figura 13 pretende comparar los resultados obtenidos por Hoppe (2006) y nuestro trabajo, ya que este es un punto a favor de la historia que pretendemos contar acerca de la familia OR humana.

Figura 13. Regiones conservadas del promotor.

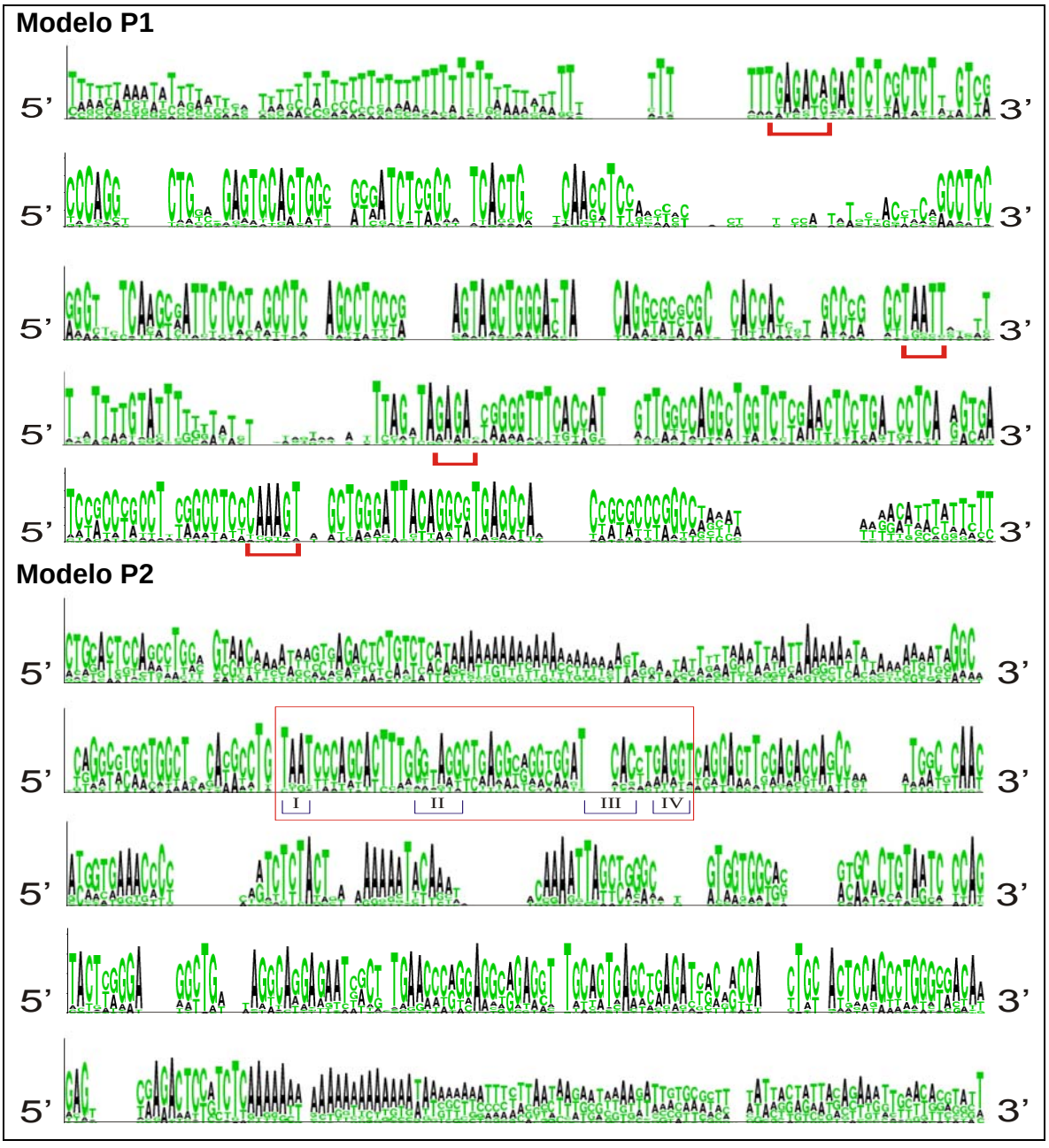


En el presente trabajo se han establecido 2 modelos estructurales de promotor. La figura compara el modelo de promotor P2 del presente trabajo con el establecido por Hoppe y colaboradores (2006). Las líneas indican la ubicación de los factores de trascripción y motivos dentro del promotor mOR (ratón), con el promotor P2 hOR (humano).

Estos modelos de promotor son un aporte grande en la comprensión de los mecanismos de regulación de la familia hOR y de los genomas. la figura 14

muestra los 2 modelos de promotor que se han establecido; representados mediante “Sequence Logo”, a partir de los alineamientos hechos y la búsqueda de motivos o regiones conservadas.

Figura 14. Estructura de los modelos de promotor encontrados



Dentro del presente trabajo se han descubierto 2 posibles modelos de promotores, a partir de los dominios conservados encontrados dentro de las secuencias. La figura es una representación ”sequence logo” de los modelos de promotores. Las letras inferiores representan los posibles factores de trascripción del promotor asociados a hoppe, 2006.

La anterior figura, permite apreciar la alta similitud entre las secuencias, encontrando pocas regiones con similitudes bajas; además de la trascendencia de esta parte del presente estudio, debido a que puede apreciarse como los factores de trascripción definidos para ratón, son solo una pequeña fracción del modelo P2 que hemos encontrado para el hombre. Esto implica que las características de cada receptor olfativo como la inhibición de otros receptores sobre la membrana celular y su función combinatoria con otros receptores, debe estar regido por grandes complejos de regulación que involucren desde inductores, silenciadores e incluso aisladores tal como lo planteó Cooper y colaboradores (2006).

Por ultimo se relacionaron los modelos de promotor con los agrupamientos cromosómicos establecidos anteriormente, ya que esto podría ser una característica fundamental en la reconstrucción filogenética de la familia hOR. La tabla 4, muestra la relación entre los modelos de promotor P1 y P2 frente a los agrupamientos cromosómicos, como puede verse los modelos se distribuyen a lo largo de la familia, lo cual indica que no son producto del ensamble de ***contigs*** tras la secuenciación del Genoma Humano.

Tabla 4. ***Tabla Resumen***. Relación entre agrupamientos cromosómicos y modelos de promotor.

Cr	Regiones	agrupamientos por región	Sub-agrupamientos por agrupamientos	Promotor-(# Seq)	N° DE GENES	Total genes Cr
1	L	I	a		1	63
			b		1	
			c		1	
	M	I	a	P1(1)	13	
			b		1	
			c		2	
	R	I		P1(3)- P2(1)	44	
2		I			2	2
3		I			10	10
5		I	a		1	4
			b	P1(1)- P2(1)	2	
			c		1	
6	L	I	a	P2(1)	2	16
			b		1	
			c		4	
			d	P1(2)	7	
			e		1	
	R	I		P1(1)	1	
7	L	I			1	15
	R	I		P1(1)	1	
		II	a	P1(1)	2	
			b		8	
			c	P1(1)	3	
8		I			1	1
9	L	I			2	25
	M	I		P1(1), P2(1)	9	
	R	I		P1(1)- P2(4)	14	
10		I		P1(1)	1	1
			a	P1(4)- P2(1)	26	
			b	P1(2)	12	
			c	P1(1)- P2(1)	17	

11	L	I	d		8	166
			e		2	
			f		2	
	M	I	a	P2(2)	6	
			b		1	
		II			2	
		III	a		1	
			b	P1(1)	1	
		IV	a		2	
			b		6	
			c	P1(6)- P2(1)	36	
			d		1	
		V	a	P(1)-P2(1)	13	
			b	P1(1)- P2(2)	7	
			c	P2(1)	1	
	R	I			1	
		II	a		2	
			b	P1(1)	9	
			c	P2(1)	8	
			d		2	
12	L	I	a		1	17
			b	P1(1)	1	
	R	II		P1(3)- P2(1)	15	
14		I	a		1	23
			b	P1(3)	16	
			c		1	
			d		1	
			e		3	
			f		1	
15	L	I			2	5
	R	I	a		2	
			b		1	
16		I	a		1	2
			b	P2(1)	1	
17	L	I		P1(2)- P2(1)	11	13
	R	I			2	
19	L	I		P1(1)	1	20
	M	I	a	P1(1)	1	
			b	P1(3)-P2(3)	7	
	R	I	a	P1(1)- P2(2)	5	
			b	P1(1)	1	
			c	P1(1)- P2(1)	4	
			d		1	
22		I			1	1
X		I			1	1

La anterior tabla “tabla resumen”, nos permite apreciar como ciertos cromosomas y agrupamientos cromosómicos, albergan lo que tal vez sea el punto de origen para cada modelo de promotor, factor que será solo establecido tras la reconstrucción del árbol filogenético de la familia hOR. Por lo pronto, inquieta un poco el conocer que hay agrupamientos cromosómicos que albergan los dos modelos de promotores, mostrando la existencia de un gran complejo de regulación, tal como se evidencia en el cromosoma 19. *Estas son hipótesis que solo podrán ser demostradas mediante trabajo experimental y el conocimiento detallado de los otros 25 motivos encontrados.*

7.2.3 Análisis de Secuencias FASE II: Análisis de Similitud y Búsqueda de Homólogos

Debemos considerar que sobre la familia hOR se realizaron dos investigaciones trascendentales para definir la organización filogenética de sus miembros ofreciendo como resultado la nomenclatura internacional con la cual cuentan. Estas investigaciones definen a todos los miembros de la familia hOR como homólogos sin características análogas (Zozulya et al, 2001; Niimura & Nei 2003), es decir, los miembros de la familia hOR no han surgido por procesos de divergencia filogenética sino bajo un ancestro común. Esto nos permite afirmar que todos los genes con los que hemos trabajamos son homólogos y que la "búsqueda de homólogos" aquí enunciada, indica la caracterización de todos los genes y proteínas homólogos hOR a partir de sus características filogenéticas y de similitud con respecto a los demás miembros de la familia hOR.

Esto es importante mencionarlo debido a que en el presente trabajo, nosotros asociamos la similitud calculada mediante algoritmos BLAST de cada gen, con la homología estimada mediante algoritmos filogenéticos; estrategia descrita en la metodología y discutida a continuación.

7.2.3.1 BLAST y Análisis de Grafos

El algoritmo BLAST es utilizado generalmente en bases de datos, para la búsqueda de secuencias asociadas a un Query o secuencia pregunta y así poder asociar características estructurales o funcionales a las secuencias buscadas (Pevsner, 2003).

Dentro del presente trabajo este algoritmo fue aplicado con el propósito de estimar la similitud de cada una de las secuencias frente a todos los miembros de la familia hOR y poder asociar este resultado al árbol filogenético creado. Esta estrategia se conoce como "BLAST local" e involucró la creación de dos bases de datos propias a partir de los genes y proteínas hOR, tal como se mencionó en la metodología.

Construida la base de datos se pudo dar paso a la búsqueda de las secuencias, tanto de proteínas como de genes dentro de las bases de datos creadas ofreciendo como resultado, una matriz BLAST con el porcentaje de similitud de cada gen o proteína frente a cada miembro de la familia hOR.

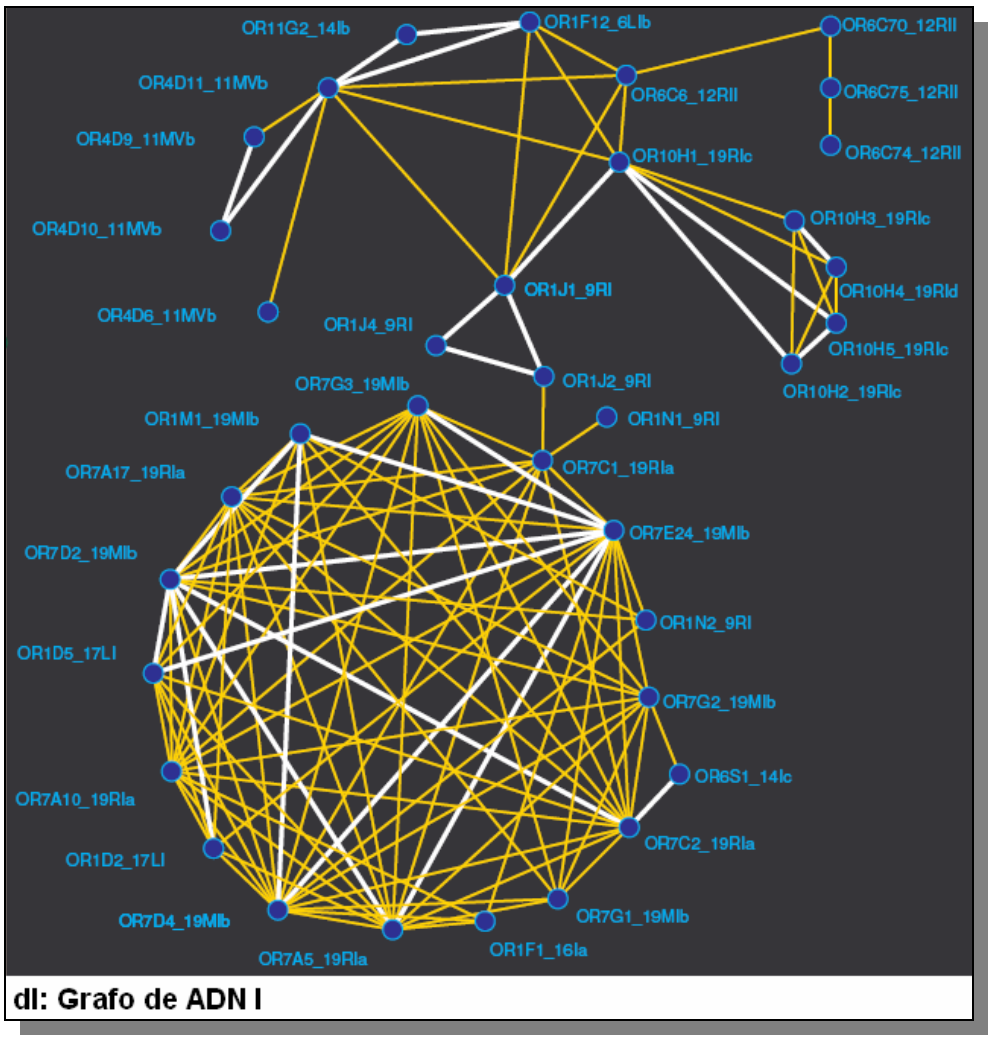
Dado que el análisis de estas matrices es muy tedioso debido al volumen de datos con el que se cuenta, se procedió a la visualización de estos resultados de tal

forma que se facilitara el trabajo. Esto se realizó gracias la aplicación de la *teoría del grafo*[6] sobre el resultado del BLAST, herramienta utilizada en el estudio del genoma de *C. elegans* con excelentes resultados. La teoría del grafo, permite visualizar diferentes y variadas relaciones entre las variables que el investigador desee, siendo en biología molecular, la similitud entre genes y proteínas. Este trabajo se realizó mediante el programa Phylographer, el cual según palabras de sus creadores, "*es un programa para visualizar y estudiar las relaciones evolutivas entre familias homólogas de genes o proteínas, mediante el dibujo de grafos asociados correspondiente a la similaridad establecida entre ellos*[7]".

De los grafos obtenidos, fueron seleccionados los que bajo una semejanza del 60 %, se agruparon en un número de secuencias mayor o igual a 9 miembros, dando como resultado, la selección de 9 grafos para proteínas y 2 para DNA. A estos grafos se les designó una nomenclatura de acuerdo al numero de secuencias por grafo y tipo de secuencias; por ejemplo el grafo **dI**, indica que es el grafo mas grande de ADN o el **plX**, que indica, que es el grafo mas pequeño de proteínas. La figura 15 muestra el grafo dI como ejemplo de los creados.

[6] "Graph Teory" es la rama de la matemática que implica el estudio de graficas
[7] http://www.atgc.org/PhyloGrapher/

Figura 15. Representación mediante grafos del BLAST hecho a los genes hOR



El análisis de matrices resultado del BLAST es un poco complicado cuando los volúmenes de información son demasiado grandes, por ello La representación mediante grafos facilita el proceso de análisis. La figura representa la similaridad existente entre los genes hOR determinadas mediante BLAST. Los círculos azules indican cada uno de los genes, las líneas atan los genes similares y el color de la línea el grado de similitud así: (1.0-0.90) blanco, (0.90-0.80) amarillo, (0.80-0.75) naranja, (0.75-0.70) gris oscuro, (0.70-0.65) gris y (0.65-0.60) gris claro.

Como se puede apreciar, los grafos asocian a todos los genes de las base de datos con las secuencias incluidas en la búsqueda. Esto ha permitido definir puntos de anclaje para grupos de genes o proteínas distantes, y cuya similitud es difícil de observar en las plantillas que ofrece el BLAST. Ahora bien, lo importante de este punto es poder asociar estos grafos con los agrupamientos cromosómicos anteriormente establecidos, dada la necesidad relacionar la similitud de los genes y proteínas a la topología en los cromosomas de ellos. La tabla 5 reúne esta información y permite establecer relaciones.

Tabla 5. ***Tabla Resumen***. Relación entre agrupamientos cromosómicos y grafos

Cr	Regiones	agrupamiento por región	Sub-agrupamiento por agrupamiento	Promotor-(# Seq)	Grafo (# Seq)	N° DE GENES	Total genes Cr
1	L	I	a		pI (1)	1	63
			b		pI (1)	1	
			c		pI (1)	1	
	M	I	a	P1(1)		13	
			b			1	
			c			2	
	R	I		P1(3)- P2(1)	pII(4), pIII(13)	44	
2		I				2	2
3		I				10	10
5		I	a		pII(1)	1	4
			b	P1(1)- P2(1)		2	
			c		pI (1)	1	
6	L	I	a	P2(1)	pII(2)	2	16
			b		dI (1)	1	
			c		pII(4)	4	
			d	P1(2)	pII(1)	7	
			e		pII(1)	1	
	R	I		P1(1)	dII (1), pIX(1)	1	
7	L	I				1	15
	R	I		P1(1)		1	
		II	a	P1(1)		2	
			b		dII (5), pIX(5)	8	
			c	P1(1)	dII (3), pIX(3)	3	
8		I			pI (1)	1	1
9	L	I			pVIII(1)	2	25
	M	I		P1(1), P2(1)	pVIII(9)	9	
	R	I		P1(1)- P2(4)	dI (5)	14	
10		I		P1(1)		1	1
11	L	I	a	P1(4)- P2(1)		26	166
			b	P1(2)		12	
			c	P1(1)- P2(1)		17	
			d			8	
			e			2	
			f			2	
	M	I	a	P2(2)	pIV (2)	6	
			b		pIV (1)	1	
		II			pIV (2)	2	
		III	a		pIV (1)	1	
			b	P1(1)	pIV (1)	1	
		IV	a		pIV (2)	2	
			b		pIV (2)	6	
			c	P1(6)- P2(1)		36	
			d			1	
		V	a	P(1)-P2(1)		13	
			b	P1(1)- P2(2)	dI (4)	7	
			c	P2(1)		1	
	R	I				1	
		II	a			2	
			b	P1(1)	pVI(1)	9	
			c	P2(1)	pVI(8)	8	
			d		pVI(2)	2	
12	L	I	a			1	17
			b	P1(1)		1	
	R	II		P1(3)- P2(1)	dI (4), pV(12)	15	
			a			1	
			b	P1(3)	dI (1), pI (8)	16	

14		I	c		dI (1)	1	23
			d			1	
			e			3	
			f			1	
15	L	I				2	5
	R	I	a		pI (2)	2	
			b		pI (1)	1	
16		I	a		dI (1)	1	2
			b	P2(1)	pII(1)	1	
17	L	I		P1(2)- P2(1)	dI (2)	11	13
	R	I				2	
19	L	I		P1(1)	pI (1)	1	20
	M	I	a	P1(1)		1	
			b	P1(3)-P2(3)	dI (7), pVII (6)	7	
	R	I	a	P1(1)- P2(2)	dI (5), pVII (5)	5	
			b	P1(1)		1	
			c	P1(1)- P2(1)	dI (4)	4	
			d		dI (1)	1	
22		I				1	1
X		I				1	1

Los grafos obtenidos al igual que la anterior tabla, permiten fortalecer las hipótesis que a través de la filogenia se han planteado, por lo cual solo podemos en este momento comentar las características de agrupamiento que los grafos nos ofrecen al igual que ingentes hipótesis que solo se valorarán mas adelante, con la filogenia. Teniendo claro este punto, se puede entonces definir las características que los grafos muestran tal como es el agrupamiento por similaridad de genes y proteínas.

Tras el estudio detallado de los grafos podemos darnos cuenta que los grafos de DNA pueden agrupar mas secuencias que los grafos de proteínas, aunque el número seleccionados sea menor. Un ejemplo claro de ello es el caso del grafo de DNA (dI) que abarca dentro de el, lo que en proteínas es un solo grafo, el (pVII). Por otro lado, si las secuencias son muy similares entre si, esto se debe ver en los grafos, pues la similaridad de las secuencias de proteínas y de DNA no debería variar demasiado. Los grafos (dII) y (pIX) reflejan esta característica, pues son iguales en cuanto a las secuencias que agrupan.

Otras características de los grafos son el tipo de secuencias que agrupan, pues son pistas de los posibles eventos moleculares que han sucedido durante la divergencia de la familia. Tal es el caso del grafo (pI), ya que agrupa a 9 secuencias de diferente Agrupamiento Cromosómico **AC** compuestos por un solo gen, además de 8 secuencias de un mismos AC el 14Ib; Esto podría implicar que las 9 secuencias ubicadas en diferentes lugares del genoma, divergieron de un mismo punto dentro del cromosoma 14 mas exactamente del agrupamiento cromosómico 14Ib. Otro dato para el posterior análisis filogenético, es el grafo (pVIII) ya que agrupa la totalidad de secuencias del AC 9MI, lo que implicaría que son el producto de la duplicación de un gran agrupamiento filogenético.

Por último se destaca el grafo (pIII), dado que esta compuesto por 13 secuencias

de un solo AC el 1RI, característica que se manifiesta de igual forma en el grafo (pV) compuesto por 12 secuencias de un solo ACel 12RII.

7.2.3.2 Alineamiento Múltiple de Genes hOR

Como se mencionó anteriormente, el primer paso para la construcción de árboles filogenéticos a nivel bioinformático es realizar un alineamiento múltiple de las secuencias. Este alineamiento implica estimar los posibles eventos evolutivos sucedidos a cada una de las secuencias, mediante la inserción de gaps (huecos) dentro de las regiones de poca similitud para aumentarla en otras (Pevsner, 2003). Esta inserción de gaps nosotros la hemos manipulado de tal forma que la falta o el exceso de estos no afecten la estructura del alineamiento drásticamente y por tanto refleje los posibles eventos filogenéticos de la familia hOR.

Para tal fin, fueron evaluados los porcentajes de similitud de los dominios mas conservados de los genes tras una serie de alineamientos múltiples de las secuencias y bajo el cambio de los parámetros de inserción o penalización de gaps "*Gap Open* y *Gap Extend*" dentro del algoritmo *ClustalW*, tal como se mencionó en la metodología.

Es importante tener en cuenta que dada la necesidad de asociar las características filogenéticas de los genes y proteínas hOR, los parámetros para el alineamiento de sus secuencias debe ser constante y por tanto los valores de penalización de gaps iguales para genes y proteínas; Esto permite exceptuar la evaluación de penalidad de gaps en alguno de los dos archivos, en este caso las secuencias de proteínas.

La tabla 6 presenta los valores seleccionados para la evaluación, los cuales son los predefinidos para secuencias estrechamente relacionadas como lo son las de la familia hOR (Pevsner, 2003).

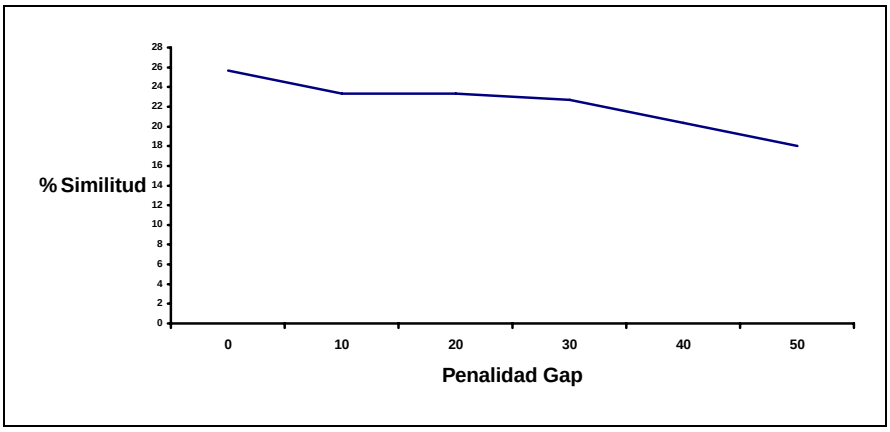
Tabla 6. Valores de penalidad de gaps evaluados

Gap extension	Gap opening
0	0
0,2	10
0,4	20
0,6	30
0,8	40
1	50

La realización de un buen alineamiento implica penalizar correctamente la inserción de gaps dentro de las secuencias. Los valores de *Gap opening* y Gap extension, son los predefinidos para secuencias estrechamente relacionadas (Pevsner, 2003).

El resultado de evaluación de la similitud de las secuencias tras cada alineamiento hecho bajo los parámetros anteriores, se representa en la gráfica 1, donde se ilustra la forma como la inserción de gaps altera esta similitud y por tanto permite seleccionar el conjunto óptimo de parámetros para el alineamiento final.

Gráfica 1. Inserción de gaps



La inserción correcta gaps permite realizar un buen alineamiento y por tanto conocer los eventos filogenéticos de la familia hOR. Aquí se gráfica la relación entre la penalidad del gap y el porcentaje de similitud del dominio mas conservado.

La curva que ofrece la anterior gráfica muestra tendencia al aumento de la similitud cuando las penalizaciones son bajas, esto quiere decir de acuerdo a Pevsner, (2003), que las secuencias necesitan la inserción de un porcentaje medianamente bajo de gaps para alinear correctamente las secuencias.

Esta gráfica además permitió seleccionar los valores de (10) para *Gap opening* y (0.2) para *Gap extension*, como los óptimos para el alineamiento múltiple de la familia hOR mediante el algoritmo *ClustalW*, pues son el valor promedio en donde la similitud alta de las secuencias no es afectada o se mantiene constante tras la penalización de los gaps. Estos valores coinciden con los predefinidos por el algoritmo de alineamiento ClustalW, los cuales se han establecido como valores estándar para secuencias con un grado medio de similitud (Thompson, J.D., 1994).

El anexo B, muestra los alineamientos hechos para las secuencias de ADN y Proteínas en una imagen píxelada (*cada punto de la imagen representa un aminoácido o un nucleótido*), esto debido a la cantidad y longitud de las secuencias que cada alineamiento contiene. Dentro de estas imágenes podemos observar el resultado del alineamiento múltiple tras la penalización de los gaps.

7.2.3.3 Análisis Filogenético de la Familia hOR

Realizado el alineamiento múltiple de las secuencias, podemos decir que los sitios en donde las secuencias varían o hay indels son los posibles eventos mutaciónales que han sufrido los miembros de la familia hOR en el transcurso de su historia evolutiva, historia que se representa mediante la construcción de árboles filogenéticos y que en bioinformática se realizan mediante la aplicación de algoritmos filogenéticos sobre las secuencias alineadas (Nei & Kumar 2000).

Nosotros realizamos todo este trabajo con la ayuda del programa *MEGA 3.1*, lo cual permitió definir el árbol correcto, los agrupamientos filogenéticos o agrupamientos y por ende contar la historia filogenética de la familia hOR.

El primer paso para esta inferencia fue la reconstrucción de los árboles filogenéticos bajo los parámetros establecidos en la metodología, ofreciendo como resultado 4 diferentes árboles para ADN y Proteínas, predichos bajo los métodos de distancia: UPGMA, Evolución Mínima (EM), Neighbor Joining (NJ) y el método de parsimonia: Máxima Parsimonia (MP).

La aplicación de estos métodos sobre las secuencias, buscó conocer si todos los árboles contaban la misma historia filogenética al distribuir los genes y proteínas de una forma particular e invariable en todos los árboles, esto debido a que la corta longitud de las secuencias (310 codones) implica que cada método puede reconstruir el árbol de forma correcta (Kumar & Filipski, 2001) y por tanto no se puede dudar de la fiabilidad de ningún método, en otras palabras “todos los árboles pueden ser correctos para esta familia, solo hay que seleccionar el mejor”.

El resultado de ello permitió observar que dentro de la familia si existen agrupaciones filogenéticas asociadas a cromosómicas y constantes en todos los árboles, pero la distribución de ellos a lo largo de cada árbol varió. Este resultado no fue el ideal o el que se buscaba, pues la historia filogenética no se puede contar sino se tiene certeza de los ancestros y clados del árbol, por lo cual se vio la necesidad de seleccionar el método y árbol ideal pero esta vez bajo dos conceptos importantes: 1) los parámetros predefinidos bibliográficamente para la familia hOR y 2) la inserción de 4 nuevas secuencias o grupo externo (OutGroup) para definir la raíz del árbol y la distribución de las agrupaciones de genes y proteínas a partir de él. (Kumar & Filipski, 2001).

Ahora bien, estos dos parámetros tenidos en cuenta en la elección del árbol correcto, implicaron comparar los árboles creados en el presente trabajo, con los reconstruidos en anteriores trabajos para la familia hOR. Además, de la búsqueda de nuevas secuencias que permitieran orientar el árbol o que sirvieran como grupo externo “OutGroup”.

Teniendo en cuenta lo anterior, se recurrió a los trabajos realizados por Niimura y Nei (2003), Zozulya y colaboradores (2001) y Glusman y colaboradores (2001); los cuales construyeron árboles filogenéticos de la familia hOR bajo sus propios parámetros. A manera de resumen corto, se puede mencionar que Niimura y Nei (2003) construyeron el árbol filogenético de la familia hOR mediante el método (NJ), sobre 388 secuencias de proteínas obtenidas en bases de datos y con un bootstrap de 100, siendo este el árbol que mejor representa la historia filogenética de la familia; Zozulya y colaboradores (2001) construyeron un árbol sin raíz para observar la distribución de los receptores del cromosoma 11 y 1, además de un árbol filogenético estándar o generado por defecto, al aplicar el algoritmo de alineamiento múltiple CLUSTALX sobre 347 secuencias de proteínas hOR; y por ultimo en el trabajo realizado por Glusman y colaboradores (2001), fue creado un árbol consenso mediante el método (NJ) con 322 secuencias de proteínas obtenidas en bases de datos. Se destaca que todos ellos trabajaron sobre secuencias de aminoácidos, esto debido a que los cambios o mutaciones en las secuencias son más fácilmente predecibles en proteínas que en DNA cuando hay un gran numero de secuencias (Nei & Kumar, 2000).

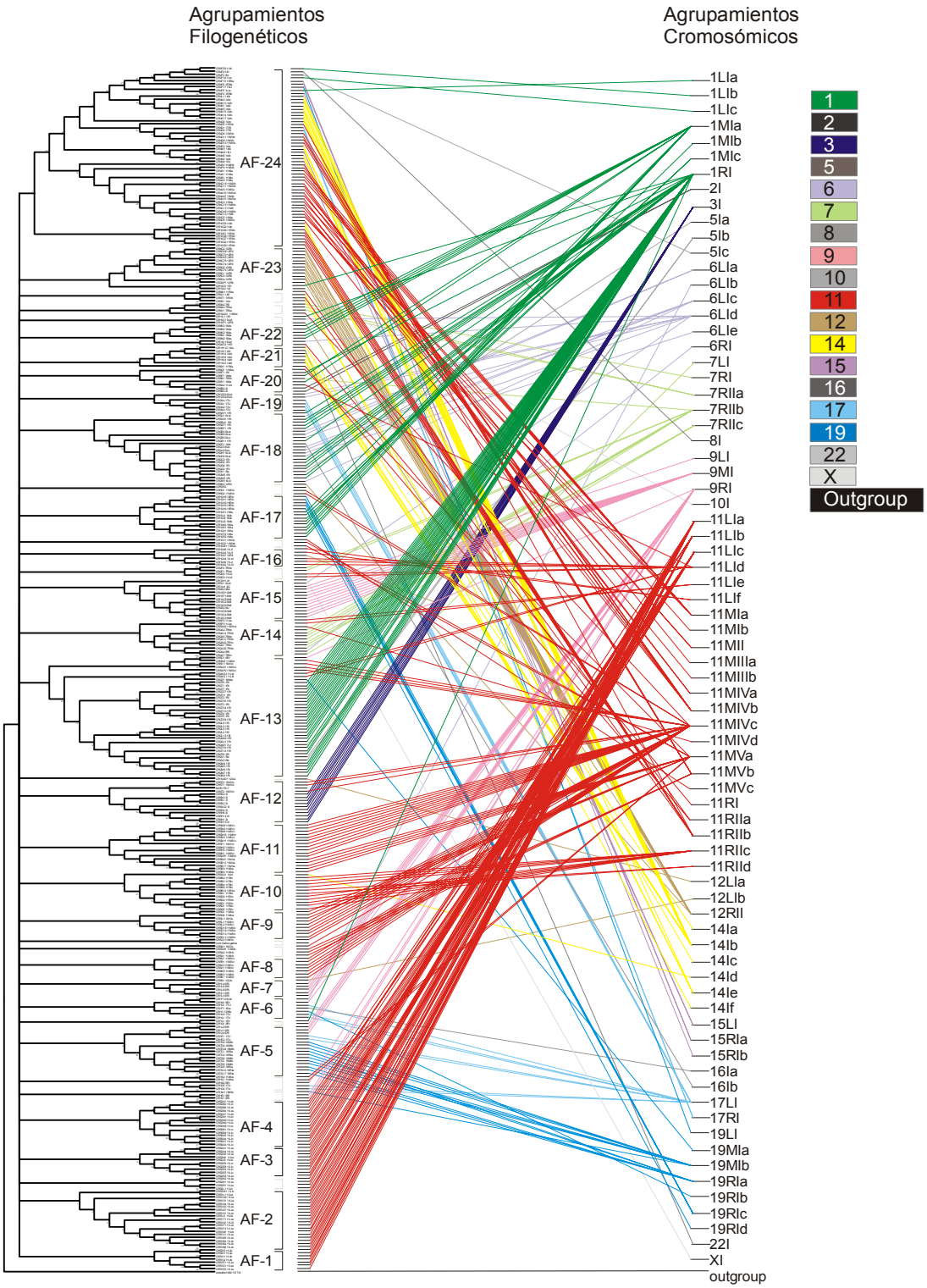
Estas investigaciones permitieron orientar la tarea de selección del árbol filogenético ideal, hacia la comparación de los árboles encontrados con los creados en el presente trabajo, permitiendo predefinir de todos los árboles propios el creado mediante el método (NJ) como árbol ideal, debido principalmente a que la distribución de las secuencias concuerda mejor con lo expuesto bibliográficamente, destacándose la definición del cromosoma 11 como el ancestro para los demás miembros de la familia y la organización de algunos genes a manera de agrupaciones sobre el árbol, como es el caso de los miembros del cromosoma 3.

Ahora bien, no se puede aun sentar una discusión profunda sobre la selección de este árbol, debido a que aún hace falta la inserción del grupo externo sobre las secuencias hOR. Este parámetro fue ejecutado en primera medida buscando y seleccionando un conjunto pequeño de genes dentro de las bases de datos, dando como resultado 4 nuevas secuencias que corresponden a receptores olfativos de perro (ID AY355591), ratón (ID NM_146365), gallo (ID NM_001031543) y pez (ID NM_131741) respectivamente, los cuales al ser involucrados en la reconstrucción del árbol filogenético, permitieron establecer la raíz del árbol, la dirección o la polaridad del proceso evolutivo de la familia hOR en todos los árboles anteriormente mencionados, tanto de distancia como de parsimonia.

Tras la aplicación de estos dos parámetros para la selección del árbol ideal, parámetro teórico (bibliográfico-comparativo) y uno práctico (grupo externo-selectivo) sobre los árboles elaborados; se pudo tener certeza en la elección del árbol ideal, permitiendo generar hipótesis y discusión acerca de la historia filogenética de la familia hOR.

Como primera medida se puede decir que el árbol seleccionado como el ideal para la familia hOR a partir de los parámetros anteriormente expuestos es el creado por el método MAXIMA PARSIMONIA (figura 16), debido principalmente a que la distribución de las proteínas hOR corresponde a lo predefinido por Niimura y Nei (2003); y el grupo externo (outgroup) se encuentra en la base del árbol filogenético.

Figura 16. Árbol filogenético de la familia hOR



Tras la selección del árbol ideal se pudo continuar con la búsqueda de agrupaciones o clados dentro de él, para poder así asociar las características filogenéticas de la familia hOR, con todas las características estructurales establecidas anteriormente. Esta búsqueda se refiere al hecho de definir las agrupaciones que más se destacan dentro del árbol, ya sea por el porcentaje de bootstrap (evaluación aleatoria) de los nodos o por el número de genes por agrupación. Ambas características son importantes para la definición de agrupaciones filogenéticas, pero dado que los valores de bootstrap del árbol (MP) son muy bajos con relación al (NJ) de Niimura y Nei (2003), 40% contra 90% por nodo, este factor de selección no fue adoptado. En contraste a esta falencia se debe tener claro que el método máxima parsimonia aquí seleccionado, establece el árbol correcto a partir de la inferencia o análisis de todos los posibles árboles o un gran numero de ellos (dependiente del computador), seleccionando al más simple y que más se aproxime a la verdadera historia filogenética de los genes (Nei & Kumar, 2000), siendo esta otra razón por la cual este método fue elegido.

Dado que el árbol parsimonioso es más exacto y simple en cuanto al número de nodos o eventos evolutivos, las ramas del árbol pueden verse simplificadas en comparación con otros árboles tal como se ilustra en nuestro árbol. Esta característica del método implicó definir las agrupaciones del árbol a partir del número de genes por grupo, considerando dentro de esta definición, a todas las secuencias del árbol aunque solo se le dió relevancia a los agrupamientos con un número mayor o igual a 5 miembros por agrupamiento.

Estas características permitieron definir 24 **Agrupamientos Filogenéticos (AF)**, los cuales se ven claramente sobre el árbol y fueron asociados a los agrupamientos cromosómicos con el propósito de plantear hipótesis o reforzar las existentes acerca de la filogenia de la familia hOR.

Ahora bien, La distribución de los agrupamientos filogenéticos es muy particular en los árboles parsimoniosos, debido a que los eventos evolutivos son menos y por tanto el árbol se define por la semejanza de cada uno de los taxas al ancestro y no por procesos iterativos de selección y exclusión como es el caso de los métodos de distancia. Considerando esto, podemos decir que la distribución de los agrupamientos filogenéticos dentro del árbol corresponde a la escala que filogenéticamente se ha calculado y por tanto los AF vecinos dentro del árbol son muy homólogos, pero no tanto como con el ancestro (Nei & Kumar, 2000). Esto es importante considerarlo en este momento, dado que la forma como se analiza la historia filogenética de la familia hOR de aquí en adelante corresponde directamente a la forma cómo los agrupamientos filogenéticos están distribuidos sobre el árbol.

Teniendo claro todos los conceptos y características que llevaron a la reconstrucción del árbol filogenético de la familia hOR, se puede entonces adentrar en la historia filogenética de la familia. Como primer paso para ello, se

establecieron las características que sobre el árbol se ven reflejadas a gran escala; y como segundo paso, el posible significado de estas características a nivel biológico.

Adentrándose al primer paso, se puede decir que el grupo de genes ancestrales de la familia hOR son los que se encuentran sobre el brazo Q del cromosoma 11, mas exactamente los genes pertenecientes al agrupamiento cromosómico 11LIa, 11LIb y 11LIc con 55 miembros o 14,3 % del total de genes hOR, este grupo ancestral se distribuye dentro de los 4 primeros AF del árbol, mostrando que esta región cromosómica es probablemente el eje de la duplicación de la familia de receptores olfativos humanos.

Todas las investigaciones acerca de la reconstrucción de la historia evolutiva de la familia hOR proponen la misma hipótesis, destacando además que la gran mayoría de los genes que se encuentran dentro de esta región cromosómica son de clase I, es decir, los encargados de percibir olores dentro del Agua. (Glusman et al, 2001 y Niimura & Nei 2003). Esta hipótesis se ve muy fortalecida por el árbol que se ha creado en este trabajo, ya que el outgroup mas ancestral es el pez y todos los genes de los AC 11LIa, 11LIb y 11LIc se encuentran consecutivamente y sin interrupción sobre los AF 1, 2 y 3 que son los mas homólogos al Outgroup Pez. Esta es una característica importante en el estudio de la familia hOR, debido que indica que es una familia de tipo ancestral y por tanto la estructura que el gen presenta, es un modelo básico de gen eucariota, del cual probablemente partieron nuevos modelos de genes con una estructura y función mucho mas compleja, como lo son las moléculas del sistema inmunológico, o de acuerdo a lo postulado por (Wedekind & Penn, 2000) posiblemente las moléculas del complejo de histocompatibilidad mayor (MHC).

Ahora bien, el análisis de las características del árbol muestra que el agrupamiento filogenético AF-12, es uno de los mas destacados debido a que abarca la totalidad de los genes pertenecientes al AC3I (color púrpura). Los agrupamientos filogenéticos 1, 2, 3, 4, 8, 9, 10 y 11 se destacan también, dado que abarcan casi el total de genes del cromosoma 11 y son los genes más cercanos al ancestro junto a los AF 5, 6, 7 (color rojo). Por otro lado, es destacable la distribución de los genes del cromosoma 1 sobre el árbol (color verde), dado que forman uno de los agrupamientos filogenéticos más grandes (AF-13), característica que se fortalece mucho más ya que los genes que lo conforman son del cromosoma 1 pero exclusivamente del agrupamiento cromosómico **1RI**. Esta particularidad se observa de igual forma sobre el AF-15, pues la gran mayoría de genes que lo conforman son del cromosoma 9 o del AC9MI.

Esta serie de agrupamientos filogenéticos, ponen de manifiesto la expansión que la familia hOR a sufrido dentro del genoma humano, pero dado el gran numero de genes por agrupamiento, es difícil proponer un modelo de expansión adecuado, apoyando los planteamientos de Niimura y Nei (2003), en donde se niega la

existencia de un único modelo de expansión como la duplicación. Por lo pronto podemos afirmar que: “Si existe una correlación directa entre los agrupamientos cromosómicos y los agrupamientos filogenéticos de los genes hOR, lo cual sugiere que la expansión de la familia hOR, depende de la función que los genes semejantes cumplen; es decir, la función de los receptores olfativos depende de la ubicación que cada uno tiene sobre los cromosomas, como consecuencia de la ubicación que estos tienen dentro del epitelio olfativo y el gasto energético necesario para su ubicación.”

7.3 ANÁLISIS GLOBAL DE LOS RESULTADOS.

Tras la definición de todas las características que definen a la familia hOR tanto estructurales como filogenéticas, se procedió a asociar cada una de estas, con el propósito de estimar y exponer las relaciones moleculares, genéticas y evolutivas que caracterizan a la familia hOR, esperando ser un aporte más al estudio de los genomas.

Ahora bien, todas estas características entre la secuencia, topología y estructura se pueden observar más fácilmente sobre la tabla 7, la cual ha venido creciendo a lo largo de este trabajo y es la herramienta de la cual partiremos este análisis final.

Tabla 7. ***Tabla Resumen***. Secuencia, topología y estructura de la familia hOR

Cr	Regiones	agrupamiento	Sub-agrupamiento	Promotor (# Seq)	Grafo (# Seq)	AF (# Seq)	Nº DE GENES	Total genes Cr
1	L	I	a		pI (1)	24(1)	1	63
			b		pI (1)	24(1)	1	
			c		pI (1)	24(1)	1	
	M	I	a	P1(1)		22(5)-20(2)-17(6)	13	
			b			17(1)	1	
			c			17(2)	2	
	R	I		P1(3)- P2(1)	pII(4), pIII(13)	23(1)-20(1)-18(12)-13(28)	44	
2		I				20(2)	2	2
3		I				**12(10**)	10	10
5		I	a		pII(1)	18(1)	1	4
			b	P1(1)- P2(1)		13(2)	2	
			c		pI (1)	24(1)	1	
6	L	I	a	P2(1)	pII(2)	18(2)	2	16
			b		dI (1)	6(1)	1	
			c		pII(4)	18(4)	4	
			d	P1(2)	pII(1)	22(1)-19(2)-18(2)-15(1)	7	
			e		pII(1)	18(1)	1	
	R	I		P1(1)	dII (1), pIX(1)	14(1)	1	
7	L	I				13(1)	1	15
	R	I		P1(1)			1	
		II	a	P1(1)			2	
			b		dII (5), pIX(5)	20(1)-14(5)	8	
			c	P1(1)	dII (3), pIX(3)	14(2)-13(1)	3	
8		I			pI (1)	24(1)	1	1
	L	I			pVIII(1)	15(2)	2	

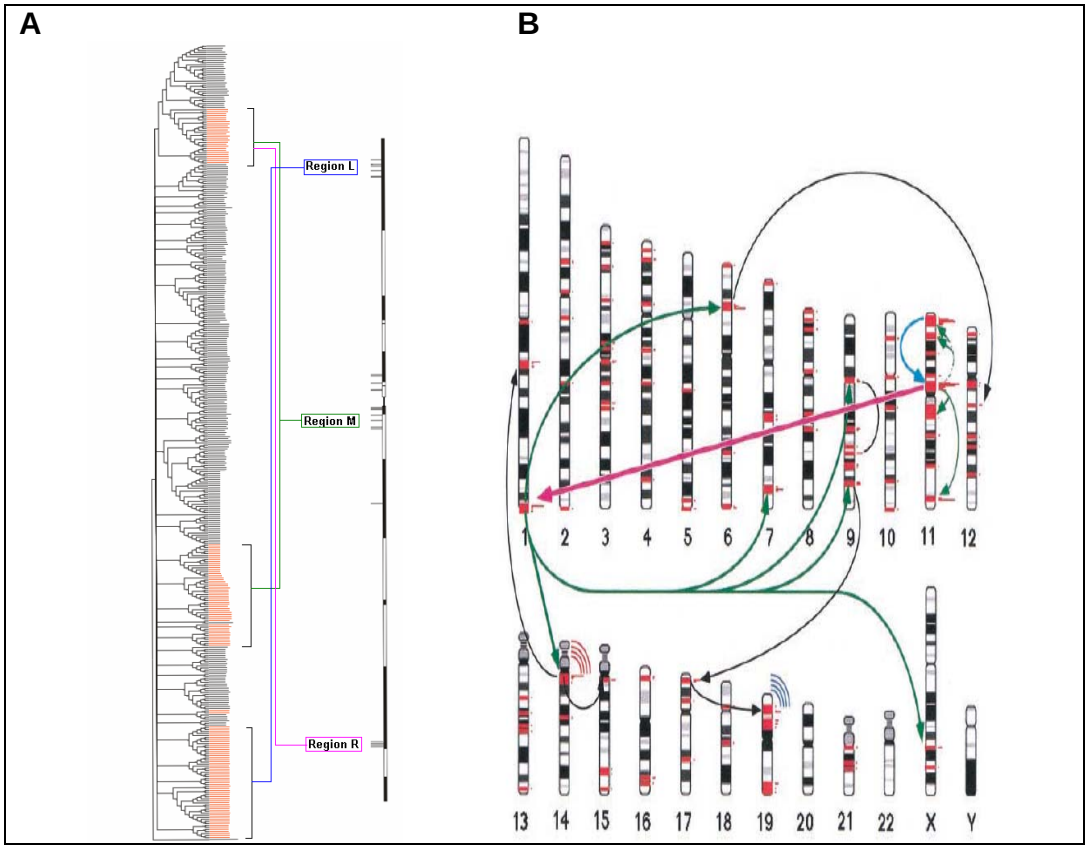
9	M	I		P1(1), P2(1)	pVIII(9)	15(9)	9	25
	R	I		P1(1)- P2(4)	dI (5)	5(3)- 6(1)-7(5)-13(1)	14	
10		I		P1(1)		23(1)	1	1
11	L	I	a	P1(4)- P2(1)		1(5)-2(11)-3(6)-4(1)	26	166
			b	P1(2)		1(2)-2(7)-4(3)	12	
			c	P1(1)- P2(1)		4(11)-3(4)-2(1)	17	
			d			13(2)-16(3)20(1)	8	
			e			14(2)	2	
			f			16(2)	2	
	M	I	a	P2(2)	pIV (2)	24(6)	6	
			b		pIV (1)	24(1)	1	
		II			pIV (2)	24(2)	2	
		III	a		pIV (1)	24(1)	1	
			b	P1(1)	pIV (1)	24(1)	1	
		IV	a		pIV (2)	24(2)	2	
			b		pIV (2)	24(6)	6	
			c	P1(6)- P2(1)		8(7)-9(6)-11(11)-12(3)-13(4)	36	
			d			14(1)	1	
		V	a	P(1)-P2(1)		9(3)- 11(5)-20(1)	13	
			b	P1(1)- P2(2)	dI (4)	24(4)	7	
			c	P2(1)			1	
	R	I				22(1)	1	
		II	a			21(1)	2	
			b	P1(1)	pVI(1)	10(1)- 24(7)	9	
			c	P2(1)	pVI(8)	10(8)	8	
			d		pVI(2)	10(2)	2	
12	L	I	a				1	17
			b	P1(1)		7(1)	1	
	R	II		P1(3)- P2(1)	dI (4), pV(12)	16(1)-23(12)	15	
14		I	a			21(1)	1	23
			b	P1(3)	dI (1), pI (8)	21(4)-24(12)	16	
			c		dI (1)		1	
			d			10(1)	1	
			e			24(3)	3	
			f				1	
15	L	I				24(2)	2	5
	R	I	a		pI (2)	24(2)	2	
			b		pI (1)	24(1)	1	
16		I	a		dI (1)	5(1)	1	2
			b	P2(1)	pII(1)	18(1)	1	
17	L	I		P1(2)- P2(1)	dI (2)	5(2)-6(3)-19(4)	11	13
	R	I				24(2)	2	
19	L	I		P1(1)	pI (1)	1(24)	1	20
	M	I	a	P1(1)		1(13)	1	
			b	P1(3)-P2(3)	dI (7), pVII (6)	5(6)	7	
	R	I	a	P1(1)- P2(2)	dI (5), pVII (5)	5(5)-	5	
			b	P1(1)		6(1)	1	
			c	P1(1)- P2(1)	dI (4)	17(4)	4	
			d		dI (1)	17(1)	1	
22		I				21(1)	1	1
X		I				15(1)	1	1

Un primer paso para el análisis es detallar la distribución que presentan los genes del cromosoma 11 sobre el árbol filogenético, ya que este es probablemente el cromosoma con mayor número de genes y sobre el que se encuentra los genes mas antiguos de la familia.

Esta distribución se puede resumir en tres grandes regiones de agrupamiento; una

cercana, una media y una lejana en relación con el ancestro (Figura 17). La primera región (cercana) y ultima (lejana) están constituidas por genes pertenecientes a regiones cromosómicas diferentes, lo cual ha permitido proponer la migración de los genes ancestrales del cromosoma 11 sobre los demás cromosomas (Glusman et al, 2001). Este mecanismo de migración expuesto por Glusman y colaboradores (2001), es un modelo cuestionado duramente por Niimura y Nei (2003), debido a que los mecanismos por los cuales esta familia se expandió en el genoma es mas complejo que la duplicación, y por ende esta debe ser producto de muchos otros procesos de selección, inversión e incluso de translocasión de genes.

Figura 17. Distribución de los genes del cromosoma 11 sobre el árbol filogenético



Sobre el árbol filogenético se observan 3 grandes regiones de agrupamiento de los genes del cromosoma 11. A) Árbol Maxima Parsimonia vs. Cromosoma 11. Permite observar la distribución de las regiones cromosómicas del cromosoma 11 L, M y R sobre el árbol filogenético. B) Hipótesis de migración del ancestro mediante procesos de duplicación, a partir de los agrupamientos filogenéticos encontrados (Glusman et al 2001).

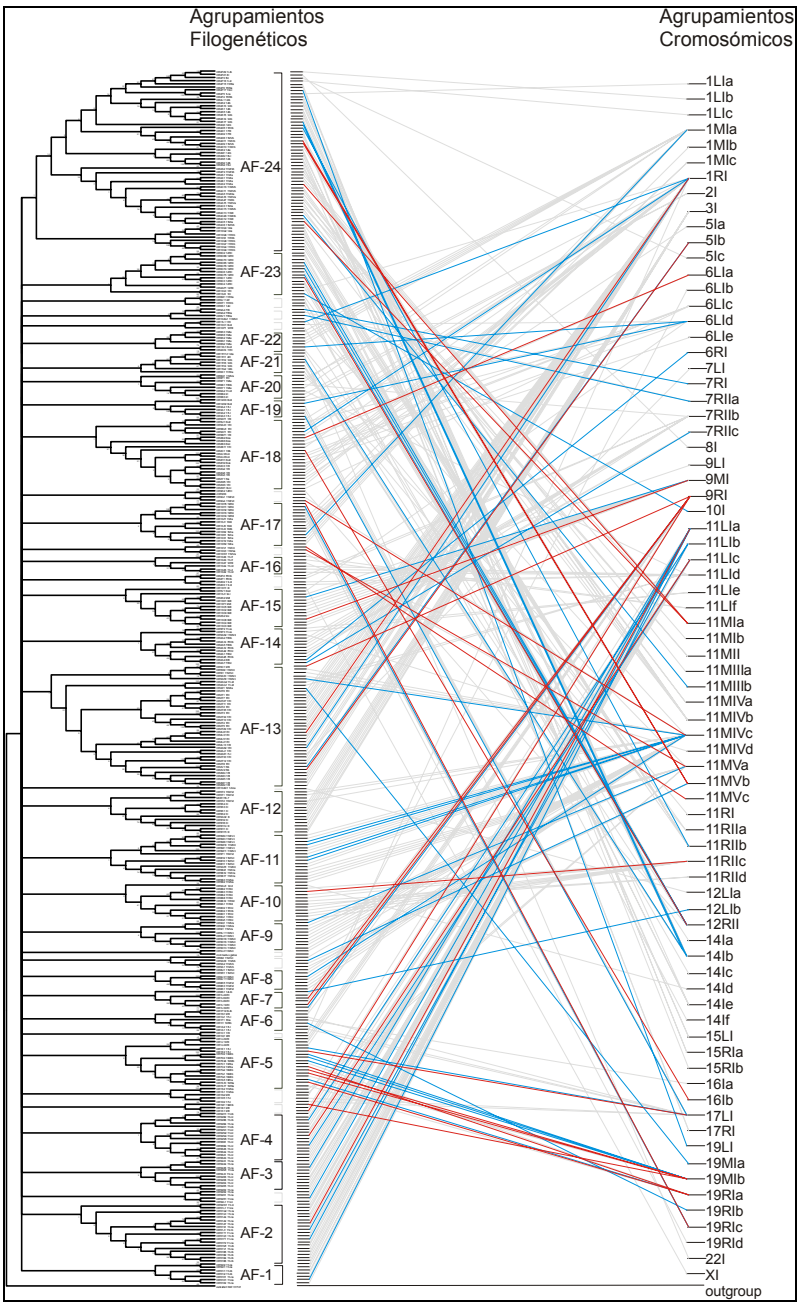
¡Claro! sobre la familia de receptores olfativos humanos si han sucedido mecanismos de expansión de la familia, he incluso el árbol aquí creado lo muestra (AF 12, 18, 19, 20, 21, 22, 23), pero como se ha podido analizar, estos procesos

han sido pasos lentos de duplicación en todos los cromosomas pero no con todos los genes. Sin embargo es de considerar, que el punto en donde diverge la familia hOR dentro del genoma humano (hot points), parece ser igualmente el cromosoma 11, pues como muestra la figura 17.A, hay un gran numero de secuencias (Región R) que no hacen parte de los ancestros, sino de los últimos genes en expandirse dentro del genoma humano.

Esta teoría puede estar reforzada por los últimos descubrimientos sobre el total de genes del cromosoma 11, ya que el mayor número de genes dentro de este cromosoma son defectuosos desde el punto de vista de desórdenes genéticos que causan enfermedades como el cáncer, diabetes, miopía entre otras (Todd D. et al 2006). Esto puede significar que dada la vulnerabilidad o inestabilidad genética encontrada dentro del cromosoma 11, los genes hOR de este cromosoma, han sido susceptibles a mutaciones que han generado los últimos genes hOR, es decir los propiamente llamados “receptores olfativos humanos”.

Ahora bien, nosotros hemos asociado las características filogenéticas y topológicas determinadas para la familia hOR con los modelos de promotor encontrados dentro del gen con el fin de favorecer la hipótesis (figura 18). Esta relación puede establecer puntos en donde posiblemente los mecanismos de regulación han surgido o se han mantenido en el genoma, al igual que las regiones codificantes de los genes. Pero el resultado de estas asociaciones muestra una distribución “Aleatoria” de los modelos sobre el árbol filogenético y sobre la topología de los genes dentro de los cromosomas, lo cual indica que los modelos estructurales de regulación aquí propuestos no están asociados directamente a la filogenia de los genes, y por tanto la regulación del gen se encuentra ligada a distintos procesos de organización y no hacen parte estrictamente de la expansión de la familia hOR. Los trabajos realizados por R. Hoppe, (2006), ratifican esta hipótesis, ya que el manifiesta que los mecanismos de regulación de los receptores olfativos no están asociados a la filogenia de los genes, pero si a la ubicación que este tiene sobre el epitelio olfativo.

Figura 18. Relación entre promotores encontrados y filogenia de sus genes



Los mecanismos de regulación de los genes hOR no siempre están asociados a la filogenia de los genes que los contienen (Hoppe, 2006). La figura representa la distribución que presentan los genes con modelo de promotor P1 (azul) y P2 (rojo) encontrados en el presente trabajo, sobre el árbol.

La existencia de relaciones entre la función del gen y su localización dentro de los cromosomas, son huellas de los mecanismos de expansión de la familia hOR en el genoma humano (Niimura y Nei, 2003), lo cual ha permitido estudiar patrones evolutivos o el desarrollo evolutivo de familias de genes entre diferentes especies como los OR humanos y de ratón (Niimura y Nei 2005). Ahora bien, si estas características se manifiestan igualmente en los mecanismos de regulación del gen, estaríamos hablando sin ninguna duda de un posible barajamiento de regiones cromosómicas que dieron origen a la expansión de la familia. Pero la verdad nosotros hemos demostrado que la regulación del gen no hace parte de este mecanismo de expansión, dado que la alta homología de las secuencias de cada modelo encontrado ~ 80 % y la distribución de ellos sobre el árbol filogenético de CDS, permite afirmar que la función que estos cumplen para el gen ya sea como regiones de control negativo de acuerdo a Cooper et al (2006) o de factores de transcripción de acuerdo a Hoppe (2006), es altamente efectivo y por tanto se mantiene desde los genes ancestrales hasta los últimos en la escala evolutiva, sin asociaciones filogenéticas aparentes entre modelos; es decir el mecanismo de regulación del gen hOR no están asociados a la función que el gen cumple dentro de la célula, si no a otro tipo de características estructurales o funcionales, que en el caso de raton es asociada a la topología de los genes sobre el epitelio olfativo (R. Hoppe, 2006).

Finalmente, la distribución de los modelos sobre el árbol filogenético no permite apreciar nodos ancestrales puntuales para alguno de los dos modelos, pero la topología que estos presentan dentro de los cromosomas ofrece algunos parámetros topológicos que precisan una distribución organizada, fuera del sorteo aparente que muestra la figura 18. Es por tal motivo que es necesario consolidar los 25 motivos ligeramente conservados (MLC) que hemos encontrado como verdaderos motivo altamente conservados (MAC), mediante la creación de estrategias de búsqueda más complejas y puntuales, que permitan establecer verdaderos nodos de divergencia para los modelos de promotor y la familia de receptores olfativos humanos.

8. CONCLUSIONES

La distribución de los genes dentro de los cromosomas y sobre el árbol filogenético creado, corresponde a lo establecido previamente para la familia hOR, lo cual implica que la estrategia seguida es correcta y pueden ser utilizados en el estudio comparativo de las regiones promotoras.

Se ha establecido la estructura del gen completo y funcional hOR, desecándose el descubrimiento y definición de 2 modelos estructurales de promotor, a partir de motivos altamente conservados encontrados sobre la región promotora, siendo un gran aporte al conocimiento del gen, la familia y el sistema olfativo humano.

Dentro de la región promotora de los genes hOR se han encontrado 25 diferentes motivos ligeramente conservados, los cuales son indicios de la expansión de la familia sobre el genoma o posibles modelos estructurales de promotor.

Bajo un análisis BLAST de la familia hOR se puede afirmar que las secuencias muy similares se encuentran dentro de iguales regiones cromosómica; al igual que la similitud entre pares de genes son lasos entre grupos distintos o no tan similares, lo cual filogenéticamente hablando pueden representar el nodo ancestral de dos o mas grupos de genes.

Existe una correlación directa entre los agrupamientos cromosómicos y los agrupamientos filogenéticos de los genes hOR, lo cual sugiere que la expansión de la familia hOR, depende de la función que los genes semejantes cumplen, es decir la función de los receptores olfativos, depende de la ubicación que cada uno tiene sobre los cromosomas.

La alta mutabilidad del cromosoma 11, puede ser la razón de que sobre el se encuentre el 56% de los genes hOR y que sus genes sean el grupo ancestral de la familia hOR.

El resultado de la relación entre los modelos de promotor encontrados, la topología de los genes y su filogenia, muestra una distribución “Aleatoria” de los modelos sobre el árbol filogenético y sobre la topología de los genes dentro de los cromosomas, lo cual indica que los modelos estructurales de regulación aquí propuestos no están asociados directamente a la filogenia de los genes, y por tanto la regulación del gen se encuentra ligada a distintos procesos de organización y no hacen parte estrictamente de la expansión de la familia hOR.

La distribución de los modelos de promotor sobre el árbol filogenético de CDS, permite afirmar que la función que estos cumplen en el gen, ya sea como regiones de control negativo o de factores de trascripción, es altamente efectivo y por tanto se mantiene desde los genes ancestrales hasta los últimos en la escala evolutiva, sin asociaciones filogenética aparentes entre modelos; es decir el mecanismo de regulación del gen hOR no están asociados a la función que el gen cumple dentro de la célula, si no a otro tipo de característica estructural o funcional, que en el caso de ratones es asociada a la topología de los genes sobre el epitelio olfativo.

Es necesario consolidar los 25 motivos ligeramente conservados (MLC) que hemos encontrado como verdaderos motivo altamente conservados (MAC), mediante la creación de estrategias de búsqueda más complejas y puntuales, que permitan establecer verdaderos nodos de divergencia para los modelos de promotor y la familia de receptores olfativos humanos.

9. RECOMENDACIONES

Los resultados de este trabajo, demuestran la necesidad de crear mejores herramientas bioinformáticas, para la búsqueda de regularidades dentro de las regiones no codificantes de los genes.

Es necesario realizar un trabajo y análisis experimental de los modelos de promotor encontrados, con el propósito de definir la función que cumplen ante el gen hOR.

El árbol filogenético de la familia hOR, demuestra la necesidad de profundizar en los mecanismos moleculares que rigen la expansión de los genes olfativos humanos, al igual que realizar comparaciones entre genes de diferentes especies.

BIBLIOGRAFÍA

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Abascal, F. (2003). Análisis de genomas, Métodos para la predicción y anotación de la función de las proteínas. Tesis doctoral, Universidad Autónoma de Madrid, Departamento de Biología Molecular, Centro Nacional de Biotecnología.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* 65: 175-187

Ben-Arie N, Lancet D, Taylor C, Khen M, Walker N, Ledbetter DH, Carrozzo R, Patel K, Sheer D, Lehrach H, North MA. (1994) Olfactory receptor gene cluster on human chromosome 17: Possible duplication of an ancestral receptor repertoire. Hum Mol Genet 1994, 3:229-235.

Chees A., Simon I., Cedar H., Axel R. (1994). Allelic inactivation regulates olfactory receptor gene expresion . cell 78:823-834.

Claus Wedekind and Dustin Penn. (2000). MHC genes, body odours, and odour preferences.Oxford Journal 15: 1269-1271.

Crasto C..,Marenco L., Miller P.L., and Shepherd G.S. (2002). Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Research1:354-360.

Dopazo, J., Valencia, A. (2001). Bioinformática y Genómica. Genómica y Mejora Vegetal. Mundi-Prensa Libros SA y Junta de Andalucía Editores Nuez F, Carrillo JM, Lozano R., 149-198.

Eric H. Davidson. (2001). Genomic Regulatory Systems: Development and Evolution (Hardcover). Academic Press Boston, MA, 2001.

Feng, D.F., and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisire to correct phylogenetic tree J. Mol. Evol. 25, 351 - 360.

Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001). Genome Res 11: 685–702.

Henrissat, B., Romeu, A. (1995). Families, superfamilies and subfamilies of

glycosyl hydrolases. Biochem J. 311: 350-351.

Huang J.Y., Brutlag D.L. (2001). The EMOTIF database. Nucleic Acids Res. 2001;29:202–204

Kumar, S., Filipski, A. (2001). Molecular Phylogeny Reconstruction. Encyclopedia of Life Sciences. Macmillan Publisher Ltd, Nature publishing Group / www.els.net

Malnic, B., Hirono, J., Sato, T., and Buck, L. (1999). combinatorial receptor codes for odors. *Cell* 96: 713-723.

Margaret O. Dayhoff. (1978). Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, April 1979.

Moreno, P.A, Fox, G.E. (2004). A relationship between the MSP gen family in C.elegans and chromosome clustering. *Molecular by Phylogenetic Analysis J.* (sometido a evaluación).

Murzin, A. G., Brenner, S.E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Molecular biology J 247: 536-540.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.

Nef P, Hermans-Borgmeyer I, Artieres-Pin H, Beasley L, Dionne VE, Heinemann SF. (1992) Spatial pattern of receptor expression in the olfactory epithelium. Proc Natl Acad Sci USA 1992, 89:8948-8952.

Nei, M. and Kumar, S. (2000). Molecular evolution and phylogenetics. Oxford Press. New York.

Niimura, Y. and Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. Proc. Natl. Acad. Sci. U.S.A. 100: 12235– 12240.

Niimura, Y. and Nei, M., (2005a). Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene* 346: 13-21.

Niimura, Y. and Nei, M. (2005b). Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* 346: 23–28.

Notre dame, C. , Higgins, D. G. & Heringa, J. (2000) J. Mol. Biol. 302, 205-217

Olender T, Feldmesser E, Atarot T, Eisenstein M, Lancet D., The olfactory receptor

universe--from whole genome analysis to structure and evolution. Genet Mol Res. (2004). Dec 30;3(4):545-53.

Otha, T. (2003). Gene Families: Multigene Families and Superfamiles. Encyclopedia of the human genome. Macmillan Publisher Ltd, Nature Publishing Group / www.ehgonline.net

Pevsner J. (2003). Bioinformatics and functional genomics. Wiley-LISS, Baltimore, Maryland.

Reed RR (1992): Mechanisms of sensitivity and specificity in olfaction. Cold Spring Harbor Symp Quant Biol 1992, 57:501-504.

Rouquier, S., Taviaux, S., Trask, B.J., Brand-Arpon, V., Van Den Engh G., Demaille, J., Giorgi, D. (1998). Distribution of olfactory receptor genes in the human genome. *Nature Genet* 18: 243-250.

Sara J. Cooper, Nathan D. Trinklein,1 Elizabeth D. Anton, Loan Nguyen, and Richard M. Myers. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome. Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

S. K. Shenoy, R. J. Lefkowitz. (2005). Seven-Transmembrane Receptor Signaling Through ß-Arrestin. Sci. STKE.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J. Mol. Biol. 147, 195-197.

Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J. A., Wetzel, C. H., Zimmer, R. K. and Hatt, H. (2003). *Science* 299: 2054–2058.

Steven Henikoff and Jorja G. Henikoff (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89: 10915-10919.

Stryer, L., Beig, J., Tymoezko, J. (2002). Biochemistry, 5ed. Reverte, SA.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research, submitted, June 1994.

Todd D. Taylor, Hideki Noguchi, Yasushi Totoki, Atsushi Toyoda, Yoko Kuroki, Ken Dewar,Christine Lloyd, Takehiko Itoh, Tadayuki Takeda, Dae-Won Kim, Xinwei She, Karen F. Barlow, Toby Bloom,Elspeth Bruford, Jean L. Chang, Christina A. Cuomo, Evan Eichler, Michael G. FitzGerald, David B. Jaffe,Kurt LaButti, Robert

Nicol, Hong-Seog Park, Christopher Seaman, Carrie Sougnez, Xiaoping Yang, Andrew R. Zimmer, Michael C. Zody, Bruce W. Birren, Chad Nusbaum, Asao Fujiyama, Masahira HattoriJane Rogers, Eric S. Lander & Yoshiyuki Sakaki. (2006) Human chromosome 11 DNA sequence and analysis including novel gene identification. Nature 440: 497 – 500.

Venter, J. C. et al., (2001). The Sequence of the Human Genome. Science. 291:1304-1351.

Yona, G., Linial, N., Linial, M. (1999). ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 37: 360-378.

Zozulya, S., Echeverri, F. & Nguyen, T. (2001). The human olfactory receptor repertoire. Genome Biol. 2, research 0018.1–0018.12.

URL Bibliográficos

Human Olfactory Receptor Data Exploratorium (HORDE). http://bioinformatics.weizmann.ac.il/HORDE/

National Center for Biotechnology Information (NCBI) www.ncbi.nlm.nih.gov/

www.mbio.ncsu.edu/BioEdit/bioedit

www.megasoftware.net

www.atgc.org/PhyloGrapher

ANEXOS

Anexo A. Tabla de Datos hOR. Creada par la Familia hOR

Cr	Nombre	Clusters Mb	Nombre Final	Distancia Intergenica	Strand	comienzo	PositionEnd	Gene size	DI Mb	score
1	OR4F5	C1-Ia-L	OR4F5_1LIa		+	58953	59868	915		0
1	OR4F29	C1-Ib-L	OR4F29_1LIb	347653	+	407521	408457	936	0,348	0
1	OR4F16	C1-Ic-L	OR4F16_1LIc	252504	-	660961	661897	936	0,253	0
1	OR10T2	C1-IIa-M	OR10T2_11MIa	154519490	-	155181387	155182329	942	154,519	0
1	OR10K2	C1-IIa-M	OR10K2_11MIa	20464	-	155202793	155203729	936	0,020	0
1	OR10K1	C1-IIa-M	OR10K1_11MIa	44695	+	155248424	155249363	939	0,045	0
1	OR10R2	C1-IIa-M	OR10R2_11MIa	13410	+	155262773	155263745	972	0,013	0
1	OR6Y1	C1-IIa-M	OR6Y1_11MIa	66248	-	155329993	155330968	975	0,066	0
1	OR6P1	C1-IIa-M	OR6P1_11MIa	14548	-	155345516	155346467	951	0,015	0
1	OR10X1	C1-IIa-M	OR10X1_11MIa	15317	-	155361784	155362711	927	0,015	0
1	OR10Z1	C1-IIa-M	OR10Z1_11MIa	26590	+	155389301	155390240	939	0,027	0
1	OR6K2	C1-IIa-M	OR6K2_11MIa	92303	-	155482543	155483515	972	0,092	0
1	OR6K3	C1-IIa-M	OR6K3_11MIa	16518	-	155500033	155500978	945	0,017	0
1	OR6K6	C1-IIa-M	OR6K6_11MIa	36805	+	155537783	155538707	924	0,037	0
1	OR6N1	C1-IIa-M	OR6N1_11MIa	9902	-	155548609	155549545	936	0,010	0
1	OR6N2	C1-IIa-M	OR6N2_11MIa	10002	-	155559547	155560498	951	0,010	0
1	OR10J3	C1-IIb-M	OR10J3_11MIb	536037	-	156096535	156097522	987	0,536	0
1	OR10J1	C1-IIc-M	OR10J1_11MIc	125132	+	156222654	156223581	927	0,125	0
1	OR10J5	C1-IIc-M	OR10J5_11MIc	94362	-	156317943	156318870	927	0,094	0
1	OR2B11	C1-IIIa-R	OR2B11_1RI	87621504	-	243940374	243941325	951	87,622	0
1	OR2W5	C1-IIIa-R	OR2W5_1RI	39145	+	243980470	243981430	960	0,039	0
1	OR2C3	C1-IIIa-R	OR2C3_1RI	39464	-	244020894	244021854	960	0,039	0
1	OR2G2	C1-IIIa-R	OR2G2_1RI	55848	+	244077702	244078653	951	0,056	0
1	OR2G3	C1-IIIa-R	OR2G3_1RI	16275	+	244094928	244095855	927	0,016	0
1	OR13G1	C1-IIIa-R	OR13G1_1RI	65605	-	244161460	244162384	924	0,066	0
1	OR6F1	C1-IIIa-R	OR6F1_1RI	38790	-	244201174	244202098	924	0,039	0
1	OR5AX1	C1-IIIa-R	OR5AX1_1RI	10346	-	244212444	244213386	942	0,010	0
1	OR5AY1	C1-IIIa-R	OR5AY1_1RI	14571	+	244227957	244228899	942	0,015	0
1	OR1C1	C1-IIIa-R	OR1C1_1RI	17908	-	244246807	244247749	942	0,018	0
1	OR5AT1	C1-IIIa-R	OR5AT1_1RI	56396	-	244304145	244305072	927	0,056	0
1	OR11L1	C1-IIIa-R	OR11L1_1RI	25201	-	244330273	244331239	966	0,025	0
1	OR2W3	C1-IIIa-R	OR2W3_1RI	53690	+	244384929	244385871	942	0,054	0
1	OR2T8	C1-IIIa-R	OR2T8_1RI	24489	+	244410360	244411296	936	0,024	0
1	OR2AJ1	C1-IIIa-R	OR2AJ1_1RI	11812	+	244423108	244424095	987	0,012	0
1	OR2L8	C1-IIIa-R	OR2L8_1RI	14105	+	244438200	244439136	936	0,014	0
1	OR2AK2	C1-IIIa-R	OR2AK2_1RI	15583	+	244454719	244455679	960	0,016	0
1	OR2L5	C1-IIIa-R	OR2L5_1RI	55611	+	244511290	244512226	936	0,056	0
1	OR2L2	C1-IIIa-R	OR2L2_1RI	15384	+	244527610	244528546	936	0,015	0
1	OR2L3	C1-IIIa-R	OR2L3_1RI	21478	+	244550024	244550960	936	0,021	0

1	OR2L13	C1-IIIa-R	OR2L13_1RI	37758	+	244588718	244589654	936	0,038	0
1	OR2M5	C1-IIIa-R	OR2M5_1RI	44836	+	244634490	244635426	936	0,045	0
1	OR2M2	C1-IIIa-R	OR2M2_1RI	33902	+	244669328	244670369	1041	0,034	0
1	OR2M3	C1-IIIa-R	OR2M3_1RI	22041	+	244692410	244693346	936	0,022	0
1	OR2M4	C1-IIIa-R	OR2M4_1RI	34925	+	244728271	244729204	933	0,035	0
1	OR2T33	C1-IIIa-R	OR2T33_1RI	32993	-	244762197	244763157	960	0,033	0
1	OR2T12	C1-IIIa-R	OR2T12_1RI	20804	-	244783961	244784921	960	0,021	0
1	OR2M7	C1-IIIa-R	OR2M7_1RI	28054	-	244812975	244813911	936	0,028	0
1	OR5BF1	C1-IIIa-R	OR5BF1_1RI	24206	+	244838117	244839053	936	0,024	0
1	OR2T4	C1-IIIa-R	OR2T4_1RI	11924	+	244850977	244851967	990	0,012	0
1	OR2T6	C1-IIIa-R	OR2T6_1RI	24983	+	244876950	244877874	924	0,025	0
1	OR2T1	C1-IIIa-R	OR2T1_1RI	17615	+	244895489	244896443	954	0,018	0
1	OR2T7	C1-IIIa-R	OR2T7_1RI	34105	+	244930548	244931472	924	0,034	0
1	OR2T2	C1-IIIa-R	OR2T2_1RI	10667	+	244942139	244943111	972	0,011	0
1	OR2T3	C1-IIIa-R	OR2T3_1RI	19581	+	244962692	244963646	954	0,020	0
1	OR2T5	C1-IIIa-R	OR2T5_1RI	14302	+	244977948	244978875	927	0,014	0
1	OR2G6	C1-IIIa-R	OR2G6_1RI	32113	+	245010988	245011936	948	0,032	0
1	OR2T29	C1-IIIa-R	OR2T29_1RI	35952	-	245047888	245048815	927	0,036	0
1	OR2T34	C1-IIIa-R	OR2T34_1RI	14330	-	245063145	245064099	954	0,014	0
1	OR2T10	C1-IIIa-R	OR2T10_1RI	18075	-	245082174	245083110	936	0,018	0
1	OR2T11	C1-IIIa-R	OR2T11_1RI	32412	-	245115522	245116470	948	0,032	0
1	OR2T35	C1-IIIa-R	OR2T35_1RI	11161	-	245127631	245128600	969	0,011	0
1	OR2T27	C1-IIIa-R	OR2T27_1RI	10675	-	245139275	245140226	951	0,011	0
1	OR5BU1	C1-IIIa-R	OR5BU1_1RI	30487	-	245170713	245171646	933	0,030	0
2	OR6B2	C2-Ia-L	OR6B2_2I		-	240688900	240689836	936		0
2	OR6B3	C2-Ia-L	OR6B3_2I	14650	-	240704486	240705479	993	0,015	0
3	OR5AC2	C3-Ia-L	OR5AC2_3I		+	99288706	99289633	927		0
3	OR5H1	C3-Ia-L	OR5H1_3I	44598	+	99334231	99335170	939	0,045	0
3	OR5H14	C3-Ia-L	OR5H14_3I	15749	+	99350919	99351849	930	0,016	0
3	OR5H15	C3-Ia-L	OR5H15_3I	18384	+	99370233	99371172	939	0,018	0
3	OR5H6	C3-Ia-L	OR5H6_3I	94694	+	99465866	99466793	927	0,095	0
3	OR5H2	C3-Ia-L	OR5H2_3I	17643	+	99484436	99485363	927	0,018	0
3	OR5K4	C3-Ia-L	OR5K4_3I	70024	+	99555387	99556350	963	0,070	0
3	OR5K3	C3-Ia-L	OR5K3_3I	35849	+	99592199	99593162	963	0,036	0
3	OR5K1	C3-Ia-L	OR5K1_3I	77948	+	99671110	99672034	924	0,078	0
3	OR5K2	C3-Ia-L	OR5K2_3I	27180	+	99699214	99700162	948	0,027	0
5	OR2Y1	C5-Ia-L	OR2Y1_5Ia		-	180098731	180099664	933		0
5	OR2V1	C5-Ib-L	OR2V1_5Ib	384301	-	180483965	180484910	945	0,384	0
5	OR2V2	C5-Ib-L	OR2V2_5Ib	29638	+	180514548	180515493	945	0,030	0
5	OR4F3	C5-Ic-L	OR4F3_5Ic	211400	+	180726893	180727829	936	0,211	0
6	OR2B2	C6-Ia-L	OR2B2_6LIa		-	27987005	27988076	1071		0
6	OR2B6	C6-Ia-L	OR2B6_6LIa	44921	+	28032997	28033936	939	0,045	0
6	OR1F12	C6-Ib-L	OR1F12_6LIb	115136	+	28149072	28149993	921	0,115	0
6	OR2W1	C6-Ic-L	OR2W1_6LIc	969978	-	29119971	29120931	960	0,970	0
6	OR2B3	C6-Ic-L	OR2B3_6LIc	41134	-	29162065	29163004	939	0,041	0
6	OR2J3	C6-Ic-L	OR2J3_6LIc	24642	+	29187646	29188579	933	0,025	0

6	OR2J2	C6-Ic-L	OR2J2_6LIc	60812	+	29249391	29250327	936	0,061	0
6	OR5U1	C6-Id-L	OR5U1_6LId	132118	+	29382445	29383408	963	0,132	0
6	OR5V1	C6-Id-L	OR5V1_6LId	47580	-	29430988	29431951	963	0,048	0
6	OR12D3	C6-Id-L	OR12D3_6LId	18144	-	29450095	29451043	948	0,018	0
6	OR12D2	C6-Id-L	OR12D2_6LId	21412	+	29472455	29473376	921	0,021	0
6	OR11A1	C6-Id-L	OR11A1_6LId	29076	-	29502452	29503397	945	0,029	0
6	OR10C1	C6-Id-L	OR10C1_6LId	12374	+	29515771	29516707	936	0,012	0
6	OR2H1	C6-Id-L	OR2H1_6LId	20818	+	29537525	29538473	948	0,021	0
6	OR2H2	C6-Ie-L	OR2H2_6LIe	125227	+	29663700	29664636	936	0,125	0
6	OR2A4	C6-IIa-R	OR2A4_6RI	102398668	-	132063304	132064234	930	102,399	0
7	OR2AE1	C7-Ia-L	OR2AE1_7LIa		-	99118338	99119307	969		0
7	OR9A4	C7-IIa-R	OR9A4_7RIa	41952552	+	141071859	141072801	942	41,953	0
7	OR9A2	C7-IIIa-R	OR9A2_7RIIa	1167325	-	142240126	142241056	930	1,167	0
7	OR6V1	C7-IIIa-R	OR6V1_7RIIa	25218	+	142266274	142267213	939	0,025	0
7	OR2F2	C7-IIIb-R	OR2F2_7RIIb	802760	+	143069973	143070924	951	0,803	0
7	OR2F1	C7-IIIb-R	OR2F1_7RIIb	23787	+	143094711	143095662	951	0,024	0
7	OR6B1	C7-IIIb-R	OR6B1_7RIIb	43075	+	143138737	143139670	933	0,043	0
7	OR2A5	C7-IIIb-R	OR2A5_7RIIb	45472	+	143185142	143186075	933	0,045	0
7	OR2A25	C7-IIIb-R	OR2A25_7RIIb	22885	+	143208960	143209890	930	0,023	0
7	OR2A12	C7-IIIb-R	OR2A12_7RIIb	19958	+	143229848	143230778	930	0,020	0
7	OR2A2	C7-IIIb-R	OR2A2_7RIIb	13545	+	143244323	143245277	954	0,014	0
7	OR2A14	C7-IIIb-R	OR2A14_7RIIb	18576	+	143263853	143264783	930	0,019	0
7	OR2A42	C7-IIIc-R	OR2A42_7RIIc	101871	-	143366654	143367584	930	0,102	0
7	OR2A7	C7-IIIc-R	OR2A7_7RIIc	25855	-	143393439	143394369	930	0,026	0
7	OR2A1	C7-IIIc-R	OR2A1_7RIIc	58496	+	143452865	143453795	930	0,058	0
8	OR4F21	C8-Ia-L	OR4F21_8I		-	106088	107024	936		0
9	OR13J1	C9-Ia-L	OR13J1_9LI		-	35859462	35860398	936		0
9	OR2S2	C9-Ia-L	OR2S2_9LI	86740	-	35947138	35948095	957	0,087	0
9	OR13F1	C9-IIa-M	OR13F1_9MI	68398003	+	104346098	104347055	957	68,398	0
9	OR13C4	C9-IIa-M	OR13C4_9MI	21036	-	104368091	104369045	954	0,021	0
9	OR13C3	C9-IIa-M	OR13C3_9MI	8563	-	104377608	104378559	951	0,009	0
9	OR13C8	C9-IIa-M	OR13C8_9MI	32444	+	104411003	104411963	960	0,032	0
9	OR13C5	C9-IIa-M	OR13C5_9MI	28332	-	104440295	104441249	954	0,028	0
9	OR13C2	C9-IIa-M	OR13C2_9MI	5260	-	104446509	104447463	954	0,005	0
9	OR13C9	C9-IIa-M	OR13C9_9MI	11623	-	104459086	104460040	954	0,012	0
9	OR13D1	C9-IIa-M	OR13D1_9MI	76313	+	104536353	104537295	942	0,076	0
9	OR2K2	C9-IIIa-M	OR2K2_9MI	6632025	-	111169320	111170268	948	6,632	0
9	OR1J1	C9-IVa-R	OR1J1_9RI	11148573	-	122318841	122319759	918	11,149	0
9	OR1J2	C9-IVa-R	OR1J2_9RI	32875	+	122352634	122353573	939	0,033	0
9	OR1J4	C9-IVa-R	OR1J4_9RI	7400	+	122360973	122361912	939	0,007	0
9	OR1N1	C9-IVa-R	OR1N1_9RI	6281	-	122368193	122369126	933	0,006	0
9	OR1N2	C9-IVa-R	OR1N2_9RI	25918	+	122395044	122395992	948	0,026	0
9	OR1L8	C9-IVa-R	OR1L8_9RI	13391	-	122409383	122410310	927	0,013	0
9	OR1Q1	C9-IVa-R	OR1Q1_9RI	46260	+	122456570	122457512	942	0,046	0
9	OR1B1	C9-IVa-R	OR1B1_9RI	12902	-	122470414	122471368	954	0,013	0
9	OR1L1	C9-IVa-R	OR1L1_9RI	32180	+	122503548	122504478	930	0,032	0

9	OR1L3	C9-IVa-R	OR1L3_9RI	12484	+	122516962	122517934	972	0,012	0
9	OR1L4	C9-IVa-R	OR1L4_9RI	47888	+	122565822	122566755	933	0,048	0
9	OR1L6	C9-IVa-R	OR1L6_9RI	24925	+	122591680	122592613	933	0,025	0
9	OR5C1	C9-IVa-R	OR5C1_9RI	38152	+	122630765	122631725	960	0,038	0
9	OR1K1	C9-IVa-R	OR1K1_9RI	10230	+	122641955	122642903	948	0,010	0
10	OR13A1	C10-Ia-L	OR13A1_10I		-	45118892	45119819	927		0
11	OR52B4	C11-Ia-L	OR52B4_11LIa		-	4345159	4346101	942		0
11	OR52K2	C11-Ia-L	OR52K2_11LIa	81044	+	4427145	4428087	942	0,081	0
11	OR52K1	C11-Ia-L	OR52K1_11LIa	38619	+	4466706	4467648	942	0,039	0
11	OR52M1	C11-Ia-L	OR52M1_11LIa	55348	+	4522996	4523947	951	0,055	0
11	OR52I2	C11-Ia-L	OR52I2_11LIa	40776	+	4564723	4565668	945	0,041	0
11	OR52I1	C11-Ia-L	OR52I1_11LIa	6203	+	4571871	4572816	945	0,006	0
11	OR51D1	C11-Ia-L	OR51D1_11LIa	44780	+	4617596	4618568	972	0,045	0
11	OR51E1	C11-Ia-L	OR51E1_11LIa	11764	+	4630332	4631286	954	0,012	0
11	OR51E2	C11-Ia-L	OR51E2_11LIa	28271	-	4659557	4660517	960	0,028	0
11	OR51F1	C11-Ia-L	OR51F1_11LIa	86270	-	4746787	4747723	936	0,086	0
11	OR52R1	C11-Ia-L	OR52R1_11LIa	33518	-	4781241	4782186	945	0,034	0
11	OR51F2	C11-Ia-L	OR51F2_11LIa	17041	+	4799227	4800217	990	0,017	0
11	OR51S1	C11-Ia-L	OR51S1_11LIa	25828	-	4826045	4826969	924	0,026	0
11	OR51T1	C11-Ia-L	OR51T1_11LIa	32736	+	4859705	4860686	981	0,033	0
11	OR51A7	C11-Ia-L	OR51A7_11LIa	24489	+	4885175	4886111	936	0,024	0
11	OR51G2	C11-Ia-L	OR51G2_11LIa	6416	-	4892527	4893469	942	0,006	0
11	OR51G1	C11-Ia-L	OR51G1_11LIa	7713	-	4901182	4902145	963	0,008	0
11	OR51A4	C11-Ia-L	OR51A4_11LIa	21822	-	4923967	4924906	939	0,022	0
11	OR51A2	C11-Ia-L	OR51A2_11LIa	7674	-	4932580	4933519	939	0,008	0
11	OR51L1	C11-Ia-L	OR51L1_11LIa	43269	+	4976788	4977733	945	0,043	0
11	OR52J3	C11-Ia-L	OR52J3_11LIa	46598	+	5024331	5025264	933	0,047	0
11	OR52E2	C11-Ia-L	OR52E2_11LIa	11194	-	5036458	5037433	975	0,011	0
11	OR52A4	C11-Ia-L	OR52A4_11LIa	61039	-	5098472	5099384	912	0,061	0
11	OR52A5	C11-Ia-L	OR52A5_11LIa	10116	-	5109500	5110448	948	0,010	0
11	OR52A1	C11-Ia-L	OR52A1_11LIa	18791	-	5129239	5130175	936	0,019	0
11	OR51V1	C11-Ia-L	OR51V1_11LIa	47368	-	5177543	5178488	945	0,047	0
11	OR51B4	C11-Ib-L	OR51B4_11LIb	100334	-	5278822	5279752	930	0,100	0
11	OR51B2	C11-Ib-L	OR51B2_11LIb	21415	-	5301167	5302103	936	0,021	0
11	OR51B5	C11-Ib-L	OR51B5_11LIb	18291	-	5320394	5321330	936	0,018	0
11	OR51B6	C11-Ib-L	OR51B6_11LIb	7983	+	5329313	5330249	936	0,008	0
11	OR51M1	C11-Ib-L	OR51M1_11LIb	36988	+	5367237	5368182	945	0,037	0
11	OR51J1	C11-Ib-L	OR51J1_11LIb	12220	+	5380402	5381350	948	0,012	0
11	OR51Q1	C11-Ib-L	OR51Q1_11LIb	18656	+	5400006	5400957	951	0,019	0
11	OR51I1	C11-Ib-L	OR51I1_11LIb	17421	-	5418378	5419320	942	0,017	0
11	OR51I2	C11-Ib-L	OR51I2_11LIb	11974	+	5431294	5432230	936	0,012	0
11	OR52D1	C11-Ib-L	OR52D1_11LIb	34282	+	5466512	5467466	954	0,034	0
11	OR52H1	C11-Ib-L	OR52H1_11LIb	54903	-	5522369	5523311	942	0,055	0
11	OR52B6	C11-Ib-L	OR52B6_11LIb	35434	+	5558745	5559687	942	0,035	0
11	OR56B1	C11-Ic-L	OR56B1_11LIc	154644	+	5714331	5715294	963	0,155	0
11	OR52N4	C11-Ic-L	OR52N4_11LIc	17252	+	5732546	5733509	963	0,017	0

11	OR52N5	C11-Ic-L	OR52N5_11LIc	21959	-	5755468	5756440	972	0,022	0
11	OR52N1	C11-Ic-L	OR52N1_11LIc	9222	-	5765662	5766622	960	0,009	0
11	OR52N2	C11-Ic-L	OR52N2_11LIc	31519	+	5798141	5799104	963	0,032	0
11	OR52E6	C11-Ic-L	OR52E6_11LIc	19660	-	5818764	5819703	939	0,020	0
11	OR52E8	C11-Ic-L	OR52E8_11LIc	14854	-	5834557	5835496	939	0,015	0
11	OR52E4	C11-Ic-L	OR52E4_11LIc	26602	+	5862098	5863034	936	0,027	0
11	OR52E5	C11-Ic-L	OR52E5_11LIc	15548	+	5878582	5879563	981	0,016	0
11	OR56A3	C11-Ic-L	OR56A3_11LIc	45589	+	5925152	5926097	945	0,046	0
11	OR56A5	C11-Ic-L	OR56A5_11LIc	19264	-	5945361	5946300	939	0,019	0
11	OR52L1	C11-Ic-L	OR52L1_11LIc	17449	-	5963749	5964691	942	0,017	0
11	OR56A4	C11-Ic-L	OR56A4_11LIc	15168	-	5979859	5980798	939	0,015	0
11	OR56A1	C11-Ic-L	OR56A1_11LIc	23758	-	6004556	6005498	942	0,024	0
11	OR56B4	C11-Ic-L	OR56B4_11LIc	80086	+	6085584	6086541	957	0,080	0
11	OR52B2	C11-Ic-L	OR52B2_11LIc	60622	-	6147163	6148132	969	0,061	0
11	OR52W1	C11-Ic-L	OR52W1_11LIc	28897	+	6177029	6177989	960	0,029	0
11	OR2AG2	C11-Id-L	OR2AG2_11LId	567827	-	6745816	6746764	948	0,568	0
11	OR2AG1	C11-Id-L	OR2AG1_11LId	16080	+	6762844	6763792	948	0,016	0
11	OR6A2	C11-Id-L	OR6A2_11LId	8742	-	6772534	6773515	981	0,009	0
11	OR10A5	C11-Id-L	OR10A5_11Ld	49974	+	6823489	6824440	951	0,050	0
11	OR10A2	C11-Id-L	OR10A2_11LId	23121	+	6847561	6848470	909	0,023	0
11	OR10A4	C11-Id-L	OR10A4_11LId	5957	+	6854427	6855399	972	0,006	0
11	OR2D2	C11-Id-L	OR2D2_11LId	13984	-	6869383	6870307	924	0,014	0
11	OR2D3	C11-Id-L	OR2D3_11LId	28549	+	6898856	6899798	942	0,029	0
11	OR5P2	C11-Ie-L	OR5P2_11LIe	874301	-	7774099	7775065	966	0,874	0
11	OR5P3	C11-Ie-L	OR5P3_11LIe	28097	-	7803162	7804095	933	0,028	0
11	OR10A6	C11-If-L	OR10A6_11LIf	101748	-	7905843	7906785	942	0,102	0
11	OR10A3	C11-If-L	OR10A3_11LIf	9916	-	7916701	7917643	942	0,010	0
11	OR4B1	C11-IIa-M	OR4B1_11MIa	40277294	+	48194937	48195864	927	40,277	0
11	OR4X2	C11-IIa-M	OR4X2_11MIa	27367	+	48223231	48224140	909	0,027	0
11	OR4X1	C11-IIa-M	OR4X1_11MIa	17848	+	48241988	48242903	915	0,018	0
11	OR4S1	C11-IIa-M	OR4S1_11MIa	41447	+	48284350	48285277	927	0,041	0
11	OR4C3	C11-IIa-M	OR4C3_11MIa	17872	+	48303149	48304055	906	0,018	0
11	OR4C5	C11-IIa-M	OR4C5_11MIa	39560	-	48343615	48344593	978	0,040	0
11	OR4A47	C11-IIb-M	OR4A47_11MIb	122327	+	48466920	48467847	927	0,122	0
11	OR4C13	C11-IIIa-M	OR4C13_11MII	1462703	+	49930550	49931477	927	1,463	0
11	OR4C12	C11-IIIa-M	OR4C12_11MII	28209	-	49959686	49960613	927	0,028	0
11	OR4C45	C11-IIIa-M	OR4C45_11MII	58122	+	50018735	50019653	918	0,058	0
11	OR4A5	C11-IVa-M	OR4A5_11MIIIa	1248373	-	51268026	51268971	945	1,248	0
11	OR4C46	C11-IVb-M	OR4C46_11MIIIb	102886	+	51371857	51372784	927	0,103	0
11	OR4A16	C11-Va-M	OR4A16_11MIVa	3494468	+	54867252	54868236	984	3,494	0
11	OR4A15	C11-Va-M	OR4A15_11MIVa	23789	+	54892025	54892967	942	0,024	0
11	OR4C15	C11-Vb-M	OR4C15_11MIVb	185553	+	55078520	55079468	948	0,186	0
11	OR4C16	C11-Vb-M	OR4C16_11MIVb	16711	+	55096179	55097109	930	0,017	0
11	OR4C11	C11-Vb-M	OR4C11_11MIVb	30386	-	55127495	55128425	930	0,030	0
11	OR4P4	C11-Vb-M	OR4P4_11MIVb	33963	+	55162388	55163345	957	0,034	0
11	OR4S2	C11-Vb-M	OR4S2_11MIVb	11610	+	55174955	55175888	933	0,012	0

11	OR4C6	C11-Vb-M	OR4C6_11MIVb	13330	+	55189218	55190145	927	0,013	0
11	OR5D13	C11-Vc-M	OR5D13_11MIVc	107344	+	55297489	55298431	942	0,107	0
11	OR5D14	C11-Vc-M	OR5D14_11MIVc	21176	+	55319607	55320549	942	0,021	0
11	OR5L1	C11-Vc-M	OR5L1_11MIVc	14969	+	55335518	55336451	933	0,015	0
11	OR5D18	C11-Vc-M	OR5D18_11MIVc	7230	+	55343681	55344620	939	0,007	0
11	OR5L2	C11-Vc-M	OR5L2_11MIVc	6650	+	55351270	55352203	933	0,007	0
11	OR5D16	C11-Vc-M	OR5D16_11MIVc	10600	+	55362803	55363787	984	0,011	0
11	OR5W2	C11-Vc-M	OR5W2_11MIVc	73917	-	55437704	55438634	930	0,074	0
11	OR5I1	C11-Vc-M	OR5I1_11MIVc	20876	-	55459510	55460452	942	0,021	0
11	OR10AG1	C11-Vc-M	OR10AG1_11MIVc	31160	-	55491612	55492515	903	0,031	0
11	OR5F1	C11-Vc-M	OR5F1_11MIVc	25220	-	55517735	55518677	942	0,025	0
11	OR5AS1	C11-Vc-M	OR5AS1_11MIVc	35793	+	55554470	55555442	972	0,036	0
11	OR8I2	C11-Vc-M	OR8I2_11MIVc	61917	+	55617359	55618289	930	0,062	0
11	OR8H2	C11-Vc-M	OR8H2_11MIVc	10805	+	55629094	55630030	936	0,011	0
11	OR8H3	C11-Vc-M	OR8H3_11MIVc	16394	+	55646424	55647360	936	0,016	0
11	OR8J3	C11-Vc-M	OR8J3_11MIVc	13465	-	55660825	55661770	945	0,013	0
11	OR8K5	C11-Vc-M	OR8K5_11MIVc	21678	-	55683448	55684369	921	0,022	0
11	OR5J2	C11-Vc-M	OR5J2_11MIVc	16300	+	55700669	55701605	936	0,016	0
11	OR5T2	C11-Vc-M	OR5T2_11MIVc	54555	-	55756160	55757114	954	0,055	0
11	OR5T3	C11-Vc-M	OR5T3_11MIVc	19137	+	55776251	55777271	1020	0,019	0
11	OR5T1	C11-Vc-M	OR5T1_11MIVc	22485	+	55799756	55800668	912	0,022	0
11	OR8H1	C11-Vc-M	OR8H1_11MIVc	13513	-	55814181	55815114	933	0,014	0
11	OR8K3	C11-Vc-M	OR8K3_11MIVc	27244	+	55842358	55843294	936	0,027	0
11	OR8K1	C11-Vc-M	OR8K1_11MIVc	26796	+	55870090	55871047	957	0,027	0
11	OR8J1	C11-Vc-M	OR8J1_11MIVc	13251	+	55884298	55885246	948	0,013	0
11	OR8U1	C11-Vc-M	OR8U1_11MIVc	14414	+	55899660	55900602	942	0,014	0
11	OR5R1	C11-Vc-M	OR5R1_11MIVc	40710	-	55941312	55942284	972	0,041	0
11	OR5M9	C11-Vc-M	OR5M9_11MIVc	44239	-	55986523	55987453	930	0,044	0
11	OR5M3	C11-Vc-M	OR5M3_11MIVc	6175	-	55993628	55994549	921	0,006	0
11	OR5M8	C11-Vc-M	OR5M8_11MIVc	19940	-	56014489	56015422	933	0,020	0
11	OR5M11	C11-Vc-M	OR5M11_11MIVc	50972	-	56066394	56067309	915	0,051	0
11	OR5M10	C11-Vc-M	OR5M10_11MIVc	33519	-	56100828	56101773	945	0,034	0
11	OR5M1	C11-Vc-M	OR5M1_11MIVc	34836	-	56136609	56137554	945	0,035	0
11	OR5AP2	C11-Vc-M	OR5AP2_11MIVc	27989	-	56165543	56166494	951	0,028	0
11	OR5AR1	C11-Vc-M	OR5AR1_11MIVc	21243	+	56187737	56188667	930	0,021	0
11	OR9G1	C11-Vc-M	OR9G1_11MIVc	35772	+	56224439	56225354	915	0,036	0
11	OR9G4	C11-Vc-M	OR9G4_11MIVc	41528	-	56266882	56267818	936	0,042	0
11	OR5AK2	C11-Vd-M	OR5AK2_11MIVd	245146	+	56512964	56513891	927	0,245	0
11	OR6Q1	C11-VIa-M	OR6Q1_11MVa	1041109	+	57555000	57555951	951	1,041	0
11	OR9I1	C11-VIa-M	OR9I1_11MVa	86599	-	57642550	57643492	942	0,087	0
11	OR9Q1	C11-VIa-M	OR9Q1_11MVa	60000	+	57703492	57704422	930	0,060	0
11	OR9Q2	C11-VIa-M	OR9Q2_11MVa	10116	+	57714538	57715480	942	0,010	0
11	OR1S2	C11-VIa-M	OR1S2_11MVa	11774	-	57727254	57728190	936	0,012	0
11	OR1S1	C11-VIa-M	OR1S1_11MVa	10641	+	57738831	57739767	936	0,011	0
11	OR10Q1	C11-VIa-M	OR10Q1_11MVa	12199	-	57751966	57752923	957	0,012	0
11	OR10W1	C11-VIa-M	OR10W1_11MVa	38068	-	57790991	57791906	915	0,038	0

11	OR5B17	C11-VIa-M	OR5B17_11MVa	90270	-	57882176	57883118	942	0,090	0
11	OR5B3	C11-VIa-M	OR5B3_11MVa	43398	-	57926516	57927458	942	0,043	0
11	OR5B2	C11-VIa-M	OR5B2_11MVa	18925	-	57946383	57947310	927	0,019	0
11	OR5B12	C11-VIa-M	OR5B12_11MVa	15948	-	57963258	57964200	942	0,016	0
11	OR5B21	C11-VIa-M	OR5B21_11MVa	67027	-	58031227	58032154	927	0,067	0
11	OR5AN1	C11-VIb-M	OR5AN1_11MVb	856353	+	58888507	58889440	933	0,856	0
11	OR5A2	C11-VIb-M	OR5A2_11MVb	56590	-	58946030	58947002	972	0,057	0
11	OR5A1	C11-VIb-M	OR5A1_11MVb	20215	+	58967217	58968162	945	0,020	0
11	OR4D6	C11-VIb-M	OR4D6_11MVb	12847	+	58981009	58981951	942	0,013	0
11	OR4D10	C11-VIb-M	OR4D10_11MVb	19527	+	59001478	59002411	933	0,020	0
11	OR4D11	C11-VIb-M	OR4D11_11MVb	25213	+	59027624	59028557	933	0,025	0
11	OR4D9	C11-VIb-M	OR4D9_11MVb	10404	+	59038961	59039903	942	0,010	0
11	OR10V1	C11-VIc-M	OR10V1_11MVc	197064	-	59236967	59237894	927	0,197	0
11	OR2AT4	C11-VIIa-R	OR2AT4_11RI	15239552	-	74477446	74478406	960	15,240	0
11	OR6X1	C11-VIIIa-R	OR6X1_11RIIa	48651094	-	123129500	123130436	936	48,651	0
11	OR6M1	C11-VIIIa-R	OR6M1_11RIIa	50892	-	123181328	123182267	939	0,051	0
11	OR8D4	C11-VIIIb-R	OR8D4_11RIIb	100081	+	123282348	123283290	942	0,100	0
11	OR4D5	C11-VIIIb-R	OR4D5_11RIIb	32243	+	123315533	123316487	954	0,032	0
11	OR6T1	C11-VIIIb-R	OR6T1_11RIIb	2299	-	123318786	123319755	969	0,002	0
11	OR10S1	C11-VIIIb-R	OR10S1_11RIIb	32860	-	123352615	123353575	960	0,033	0
11	OR10G6	C11-VIIIb-R	OR10G6_11RIIb	16507	-	123370082	123371078	996	0,017	0
11	OR10G4	C11-VIIIb-R	OR10G4_11RIIb	20413	+	123391491	123392424	933	0,020	0
11	OR10G9	C11-VIIIb-R	OR10G9_11RIIb	6505	+	123398929	123399862	933	0,007	0
11	OR10G8	C11-VIIIb-R	OR10G8_11RIIb	5677	+	123405539	123406472	933	0,006	0
11	OR10G7	C11-VIIIb-R	OR10G7_11RIIb	7513	-	123413985	123414918	933	0,008	0
11	OR8G1	C11-VIIIc-R	OR8G1_RVIIIc	210714	+	123625632	123626544	912	0,211	0
11	OR8G5	C11-VIIIc-R	OR8G5_11RIIc	13493	+	123640037	123640970	933	0,013	0
11	OR8D1	C11-VIIIc-R	OR8D1_11RIIc	43978	-	123684948	123685872	924	0,044	0
11	OR8D2	C11-VIIIc-R	OR8D2_11RIIc	8498	-	123694370	123695303	933	0,008	0
11	OR8B2	C11-VIIIc-R	OR8B2_11RIIc	62207	-	123757510	123758449	939	0,062	0
11	OR8B3	C11-VIIIc-R	OR8B3_11RIIc	13069	-	123771518	123772457	939	0,013	0
11	OR8B4	C11-VIIIc-R	OR8B4_11RIIc	26593	-	123799050	123799977	927	0,027	0
11	OR8B8	C11-VIIIc-R	OR8B8_11RIIc	15281	-	123815258	123816191	933	0,015	0
11	OR8B12	C11-VIIId-R	OR8B12_11RIId	101639	-	123917830	123918760	930	0,102	0
11	OR8A1	C11-VIIId-R	OR8A1_11RIId	26465	+	123945225	123946152	927	0,026	0
12	OR10AD1	C12-Ia-L	OR10AD1_12LIa		-	46882391	46883342	951		0
12	OR8S1	C12-Ib-L	OR8S1_12LIb	322339	+	47205681	47206617	936	0,322	0
12	OR9K2	C12-IIa-R	OR9K2_12RII	6603268	+	53809885	53810824	939	6,603	0
12	OR10A7	C12-IIa-R	OR10A7_12RII	90251	+	53901075	53902023	948	0,090	0
12	OR6C74	C12-IIa-R	OR6C74_12RII	25315	+	53927338	53928274	936	0,025	0
12	OR6C6	C12-IIa-R	OR6C6_12RII	46067	-	53974341	53975283	942	0,046	0
12	OR6C1	C12-IIa-R	OR6C1_12RII	25367	+	54000650	54001586	936	0,025	0
12	OR6C3	C12-IIa-R	OR6C3_12RII	10165	+	54011751	54012684	933	0,010	0
12	OR6C75	C12-IIa-R	OR6C75_12RII	32477	+	54045161	54046097	936	0,032	0
12	OR6C65	C12-IIa-R	OR6C65_12RII	34482	+	54080579	54081515	936	0,034	0
12	OR6C76	C12-IIa-R	OR6C76_12RII	24789	+	54106304	54107240	936	0,025	0

12	OR6C2	C12-IIa-R	OR6C2_12RII	25024	+	54132264	54133200	936	0,025	0
12	OR6C70	C12-IIa-R	OR6C70_12RII	16053	-	54149253	54150189	936	0,016	0
12	OR6C68	C12-IIa-R	OR6C68_12RII	22224	+	54172413	54173364	951	0,022	0
12	OR6C4	C12-IIa-R	OR6C4_12RII	57913	+	54231277	54232204	927	0,058	0
12	OR2AP1	C12-IIa-R	OR2AP1_12RII	22261	+	54254465	54255392	927	0,022	0
12	OR10P1	C12-IIa-R	OR10P1_12RII	61550	+	54316942	54317881	939	0,062	0
14	OR11H12	C14-Ia-L	OR11H12_14Ia		+	18447593	18448571	978		0
14	OR11H2	C14-Ib-L	OR11H2_14Ib	802366	-	19250937	19251882	945	0,802	0
14	OR4Q3	C14-Ib-L	OR4Q3_14Ib	33544	+	19285426	19286365	939	0,034	0
14	OR4M1	C14-Ib-L	OR4M1_14Ib	31956	+	19318321	19319260	939	0,032	0
14	OR4N2	C14-Ib-L	OR4N2_14Ib	46187	+	19365447	19366368	921	0,046	0
14	OR4K2	C14-Ib-L	OR4K2_14Ib	47898	+	19414266	19415208	942	0,048	0
14	OR4K5	C14-Ib-L	OR4K5_14Ib	43397	+	19458605	19459574	969	0,043	0
14	OR4K1	C14-Ib-L	OR4K1_14Ib	14091	+	19473665	19474598	933	0,014	0
14	OR4K15	C14-Ib-L	OR4K15_14Ib	38991	+	19513589	19514561	972	0,039	0
14	OR4K14	C14-Ib-L	OR4K14_14Ib	37701	-	19552262	19553192	930	0,038	0
14	OR4K13	C14-Ib-L	OR4K13_14Ib	18653	-	19571845	19572757	912	0,019	0
14	OR4L1	C14-Ib-L	OR4L1_14Ib	25286	+	19598043	19598979	936	0,025	0
14	OR4K17	C14-Ib-L	OR4K17_14Ib	56519	+	19655498	19656434	936	0,057	0
14	OR4N5	C14-Ib-L	OR4N5_14Ib	25300	+	19681734	19682658	924	0,025	0
14	OR11G2	C14-Ib-L	OR11G2_14Ib	52778	+	19735436	19736369	933	0,053	0
14	OR11H6	C14-Ib-L	OR11H6_14Ib	25408	+	19761777	19762698	921	0,025	0
14	OR11H4	C14-Ib-L	OR11H4_14Ib	18122	+	19780820	19781762	942	0,018	0
14	OR6S1	C14-Ic-L	OR6S1_14Ic	396935	-	20178697	20179690	993	0,397	0
14	OR5AU1	C14-Id-L	OR5AU1_14Id	513248	-	20692938	20693871	933	0,513	0
14	OR10G3	C14-Ie-L	OR10G3_14Ie	413905	-	21107776	21108715	939	0,414	0
14	OR10G2	C14-Ie-L	OR10G2_14Ie	63193	-	21171908	21172838	930	0,063	0
14	OR4E2	C14-Ie-L	OR4E2_14Ie	30298	+	21203136	21204075	939	0,030	0
14	OR6J1	C14-If-L	OR6J1_14If	968440	-	22172515	22173556	1041	0,968	0
15	OR4M2	C15-Ia-L	OR4M2_15LI		+	19869939	19870878	939		0
15	OR4N4	C15-Ia-L	OR4N4_15LI	12958	+	19883836	19884784	948	0,013	0
15	OR4F6	C15-IIa-R	OR4F6_15RIa	80278661	+	100163445	100164381	936	80,279	0
15	OR4F15	C15-IIa-R	OR4F15_15RIa	11531	+	100175912	100176848	936	0,012	0
15	OR4F4	C15-IIb-R	OR4F4_15RIb	103022	-	100279870	100280785	915	0,103	0
16	OR1F1	C16-Ia-L	OR1F1_16Ia		+	3194247	3195183	936		0
16	OR2C1	C16-Ib-L	OR2C1_16Ib	150758	+	3345941	3346877	936	0,151	0
17	OR1D5	C17-Ia-L	OR1D5_17LI		-	2912715	2913651	936		0
17	OR1D2	C17-Ia-L	OR1D2_17LI	28453	-	2942104	2943040	936	0,028	0
17	OR1G1	C17-Ia-L	OR1G1_17LI	33616	-	2976656	2977595	939	0,034	0
17	OR1A2	C17-Ia-L	OR1A2_17LI	69967	+	3047562	3048489	927	0,070	0
17	OR1A1	C17-Ia-L	OR1A1_17LI	17175	+	3065664	3066591	927	0,017	0
17	OR1D4	C17-Ia-L	OR1D4_17LI	24128	+	3090719	3091655	936	0,024	0
17	OR3A2	C17-Ia-L	OR3A2_17LI	36361	-	3128016	3128961	945	0,036	0
17	OR3A1	C17-Ia-L	OR3A1_17LI	12720	-	3141681	3142626	945	0,013	0
17	OR3A4	C17-Ia-L	OR3A4_17LI	17728	+	3160354	3161398	1044	0,018	0
17	OR1E1	C17-Ia-L	OR1E1_17LI	86114	-	3247512	3248454	942	0,086	0

17	OR3A3	C17-Ia-L	OR3A3_17LI	22175	+	3270629	3271574	945	0,022	0
17	OR1E2	C17-Ia-L	OR1E2_17LI	11342	-	3282916	3283885	969	0,011	0
17	OR4D1	C17-IIa-R	OR4D1_17RII	50303628	+	53587513	53588443	930	50,304	0
17	OR4D2	C17-IIa-R	OR4D2_17RII	13572	+	53602015	53602936	921	0,014	0
19	OR4F17	C19-Ia-L	OR4F17_19LI		+	61678	62593	915		0
19	OR2Z1	C19-IIa-M	OR2Z1_19MIa	8639797	+	8702390	8703332	942	8,640	0
19	OR1M1	C19-IIb-M	OR1M1_19MIb	361588	+	9064920	9065859	939	0,362	0
19	OR7G2	C19-IIb-M	OR7G2_19MIb	8088	-	9073947	9074919	972	0,008	0
19	OR7G1	C19-IIb-M	OR7G1_19MIb	11587	-	9086506	9087439	933	0,012	0
19	OR7G3	C19-IIb-M	OR7G3_19MIb	10251	-	9097690	9098626	936	0,010	0
19	OR7D2	C19-IIb-M	OR7D2_19MIb	58831	+	9157457	9158393	936	0,059	0
19	OR7D4	C19-IIb-M	OR7D4_19MIb	27184	-	9185577	9186513	936	0,027	0
19	OR7E24	C19-IIb-M	OR7E24_19MIb	36206	+	9222719	9223736	1017	0,036	0
19	OR7C1	C19-IIIa-R	OR7C1_19RIa	5547252	-	14770988	14771948	960	5,547	0
19	OR7A5	C19-IIIa-R	OR7A5_19RIa	27148	-	14799096	14800053	957	0,027	0
19	OR7A10	C19-IIIa-R	OR7A10_19RIa	12709	-	14812762	14813689	927	0,013	0
19	OR7A17	C19-IIIa-R	OR7A17_19RIa	38551	-	14852240	14853167	927	0,039	0
19	OR7C2	C19-IIIa-R	OR7C2_19RIa	60133	+	14913300	14914257	957	0,060	0
19	OR1I1	C19-IIIb-R	OR1I1_19RIb	144619	+	15058876	15059822	946	0,145	0
19	OR10H2	C19-IIIc-R	OR10H2_19RIc	640031	+	15699853	15700798	945	0,640	0
19	OR10H3	C19-IIIc-R	OR10H3_19RIc	12404	+	15713202	15714150	948	0,012	0
19	OR10H5	C19-IIIc-R	OR10H5_19RIc	51708	+	15765858	15766803	945	0,052	0
19	OR10H1	C19-IIIc-R	OR10H1_19RIc	12090	-	15778893	15779847	954	0,012	0
19	OR10H4	C19-IIId-R	OR10H4_19RId	140970	+	15920817	15921765	948	0,141	0
22	OR11H1	C22-Ia-L	OR11H1_22I		-	14823380	14824325	945		0
X	OR13H1	CX-Ia-L	OR13H1_XI		+	130403582	130404506	924		0

Anexo B. Imagen pixelada de alineamiento de Proteínas y ADN

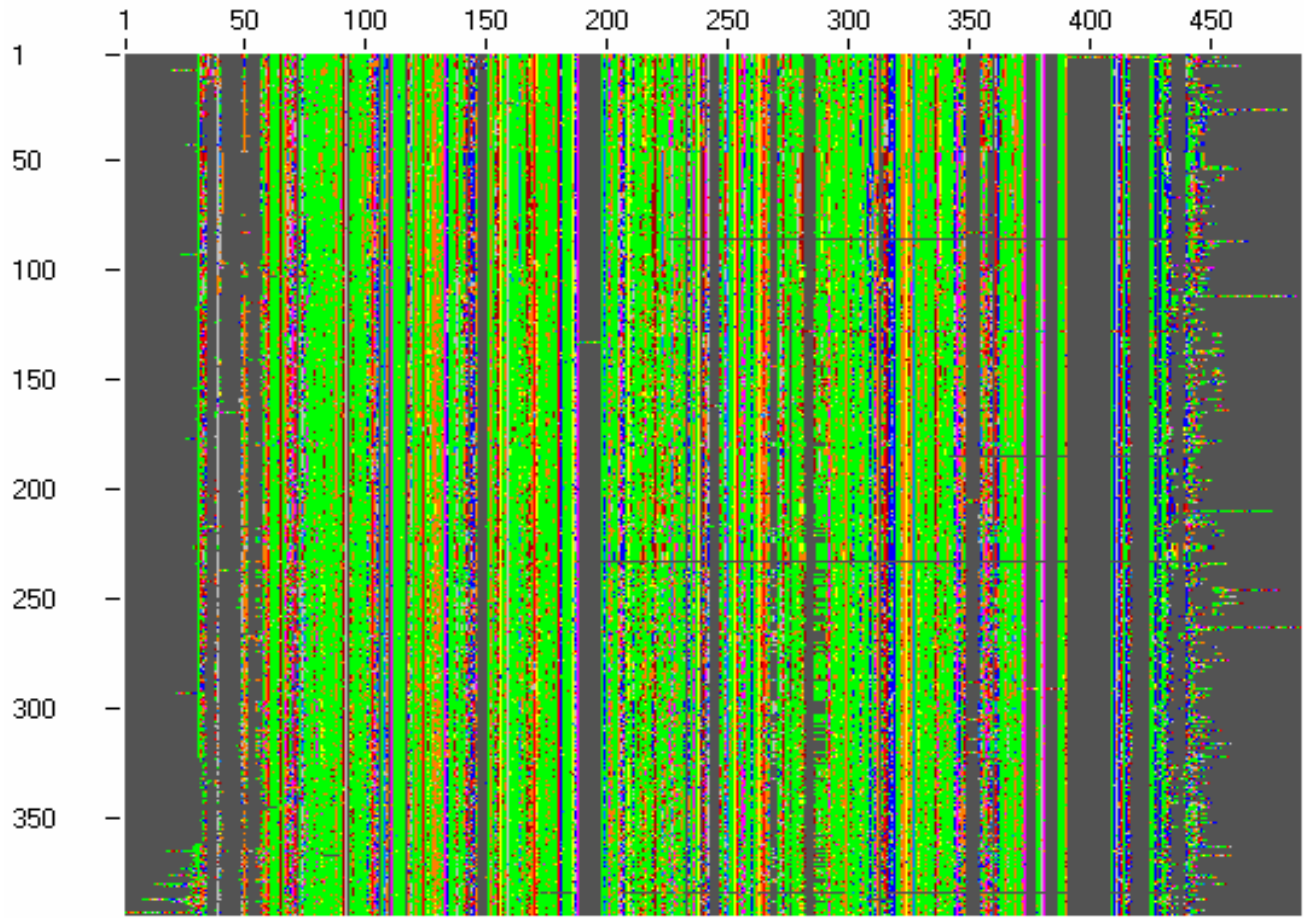
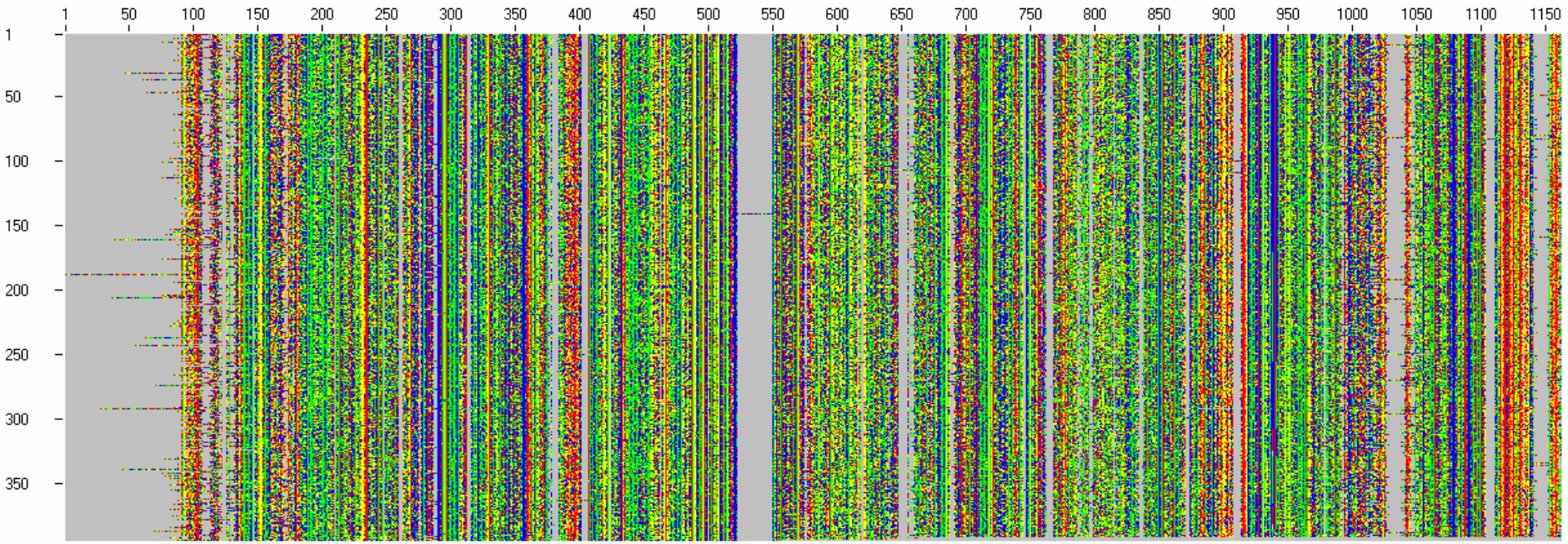
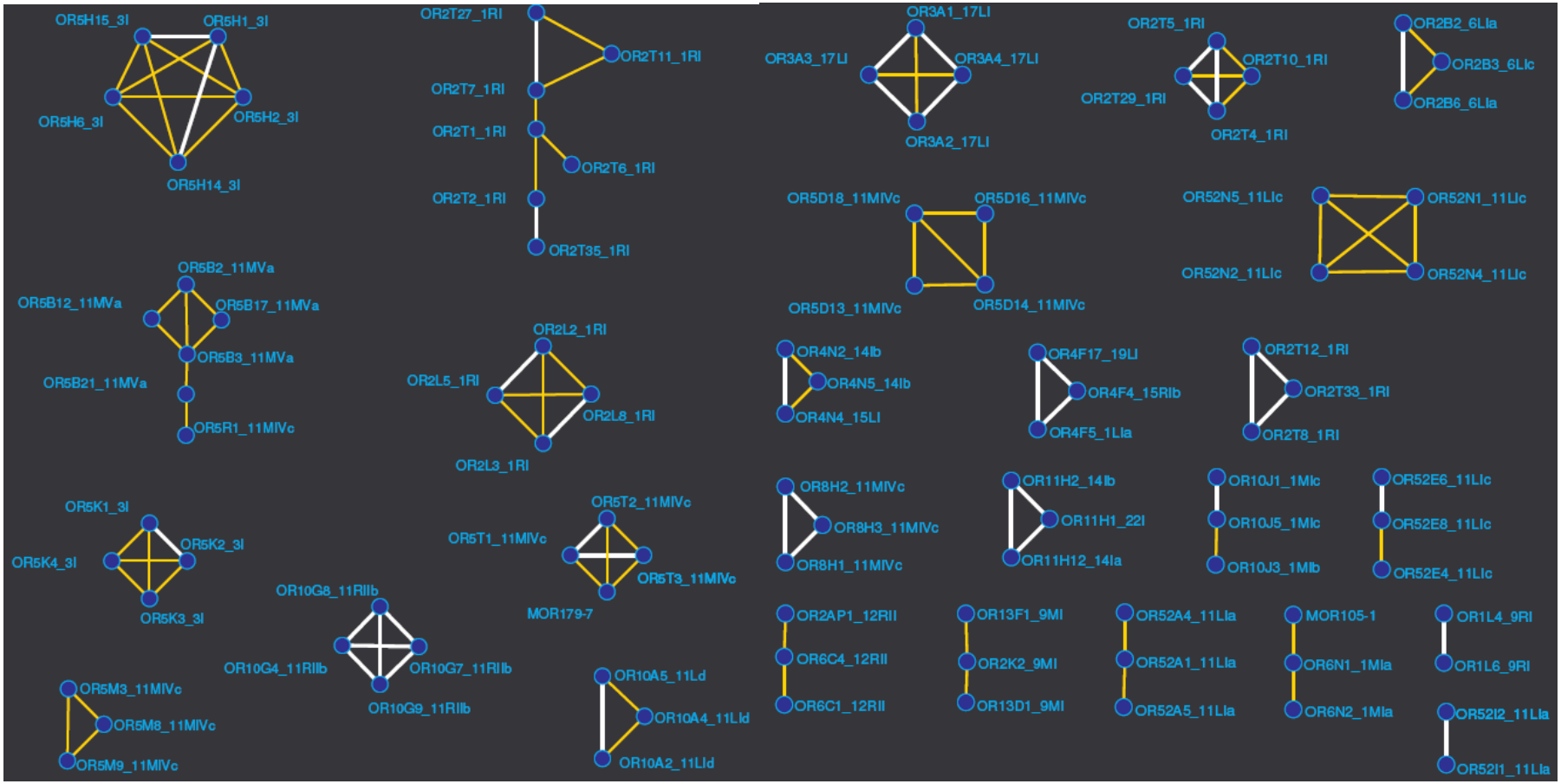
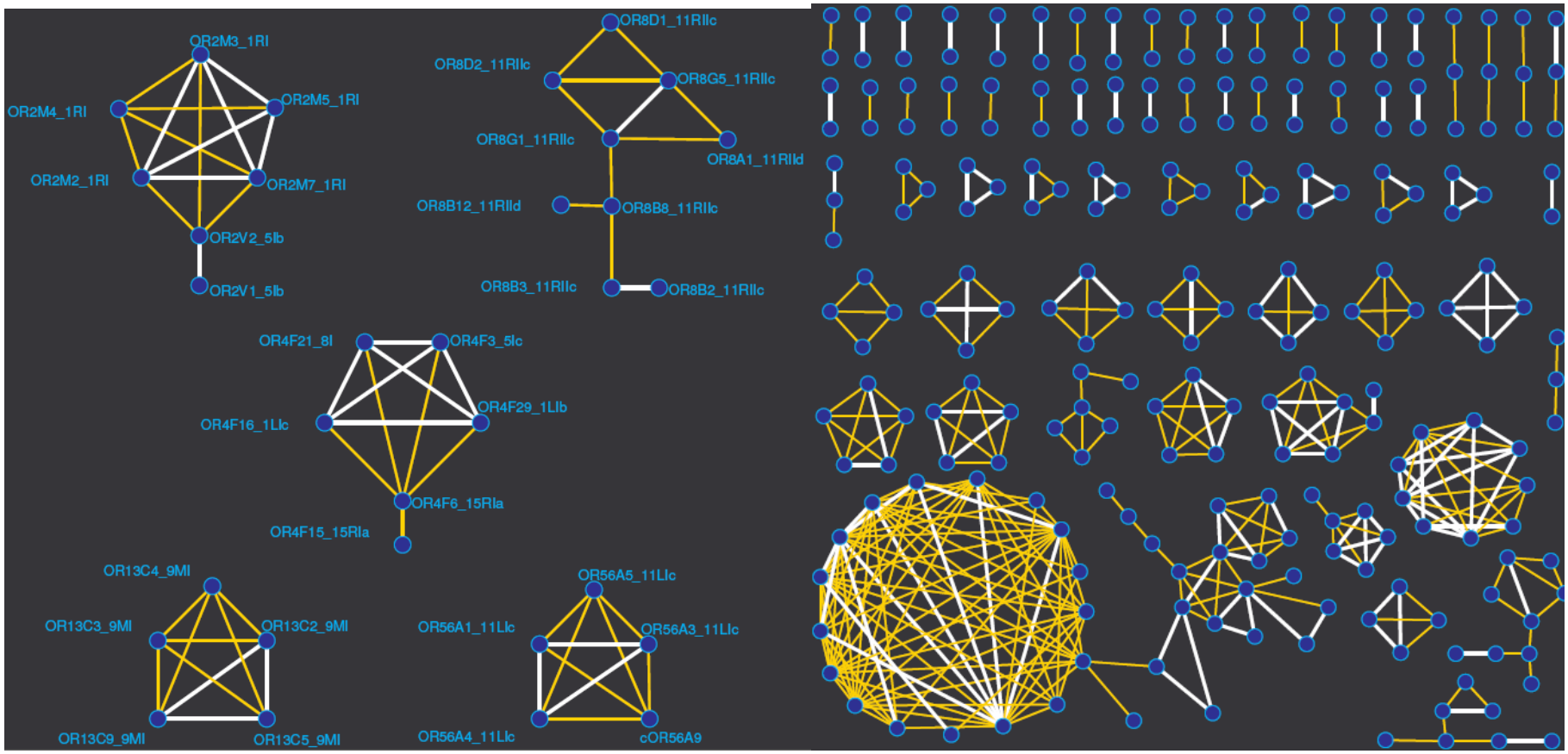


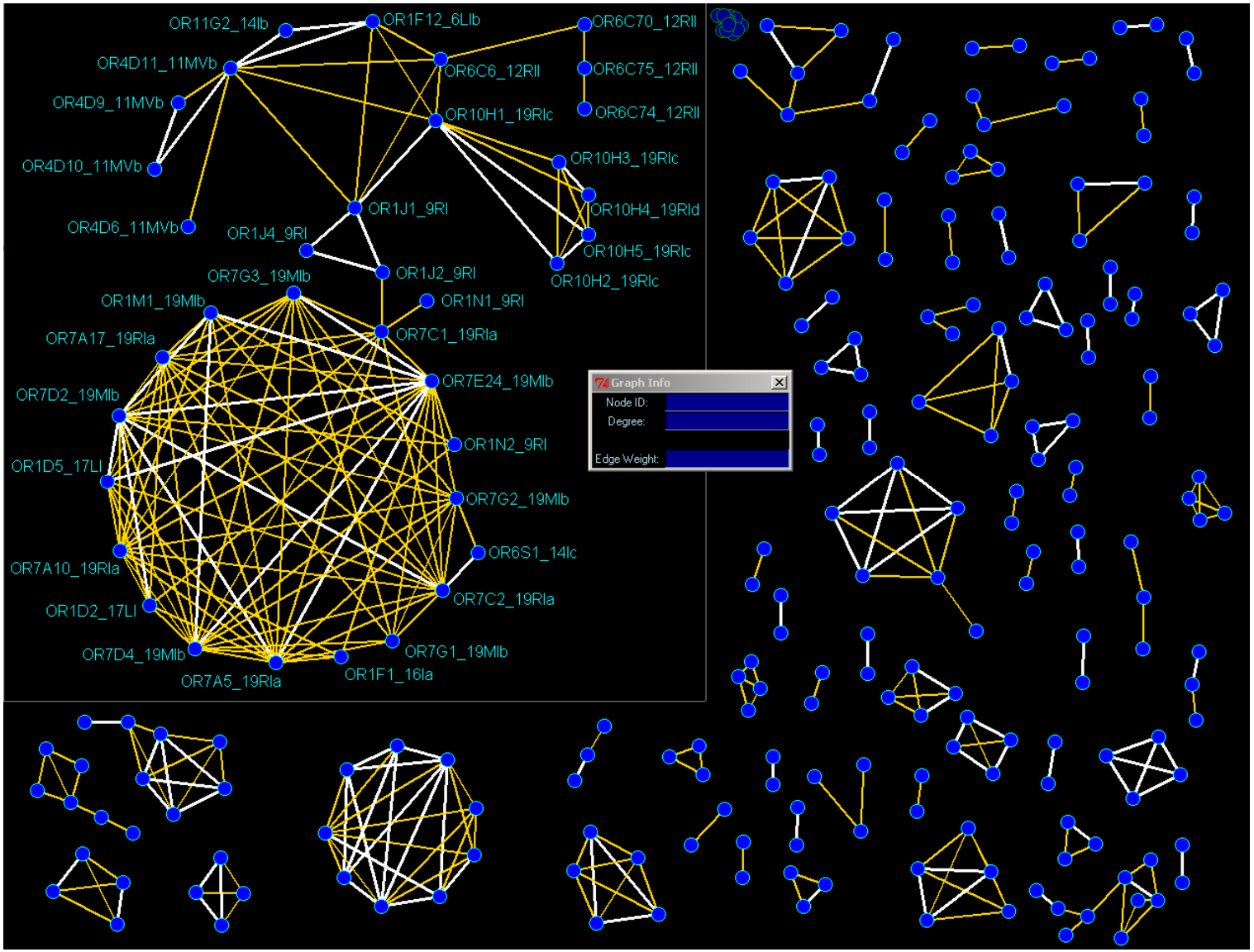
Imagen pixelada de Alineamiento de ADN



Anexo C. Grafos de phylographer




OR2M3_1RI
OR2M4_1RI
OR2M5_1RI
OR2M2_1RI
OR2M7_1RI
OR2V2_5Ib
OR2V1_5Ib
OR8D1_11RIIc
OR8D2_11RIIc
OR8G5_11RIIc
OR8G1_11RIIc
OR8A1_11RIId
OR8B12_11RIId
OR8B8_11RIIc
OR8B3_11RIIc
OR8B2_11RIIc
OR4F21_8I
OR4F3_5Ic
OR4F29_1LIb
OR4F16_1LIc
OR4F6_15RIa
OR4F15_15RIa
OR13C4_9MI
OR13C3_9MI
OR13C2_9MI
OR13C9_9MI
OR13C5_9MI
OR56A5_11LIc
OR56A1_11LIc
OR56A3_11LIc
OR56A4_11LIc
cOR56A9


OR11G2_14Ib
OR1F12_6LIb
OR6C70_12RII
OR4D11_11MVb
OR6C6_12RII
OR6C75_12RII
OR4D9_11MVb
OR10H1_19RIc
OR6C74_12RII
OR4D10_11MVb
OR10H3_19RIc
OR1J1_9RI
OR10H4_19RId
OR4D6_11MVb
OR1J4_9RI
OR10H5_19RIc
OR1J2_9RI
OR10H2_19RIc
OR7G3_19MIb
OR1M1_19MIb
OR1N1_9RI
OR7A17_19RIa
OR7C1_19RIa
OR7E24_19MIb
OR7D2_19MIb
OR1N2_9RI
OR1D5_17LI
OR7G2_19MIb
OR6S1_14Ic
OR7A10_19RIa
OR7C2_19RIa
OR1D2_17LI
OR7D4_19MIb
OR7G1_19MIb
OR7A5_19RIa
OR1F1_16Ia
Graph Info
Node ID:
Degree:
Edge Weight: