PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Facultad de Ciencias Biológicas
Programa de Doctorado en Ciencias Biológicas

ESTUDIO DE LA RELACIÓN SECUENCIA/ESTRUCTURA/FUNCIÓN EN DOMINIOS CATALÍTICOS DE POLIMERASAS DE ADN

Tesis presentada a la Pontificia Universidad Católica de Chile como parte de los requisitos para optar al grado de Doctor en Ciencias Biológicas mención Genética Molecular y Microbiología.

Por

ALEX WILLIAM SLATER MORALES

Director de Tesis: Dr. Francisco Melo Ledermann

Comisión de Tesis: Dra. Marta Bunster
Dr. Rodrigo Gutiérrez
Dr. Rafael Vicuña

Junio, 2013

Índice General

Índice de Figuras

Índice de Tablas

Abreviaturas

RMSD	*Root Mean Square Deviation*
SEQIDE	*Sequence Identity*
RSIM	*Relative Similarity*
SO	*Structural Overlap*
STOVCA	*Structural Overlap Calculator*
UmuC	*UV Mutator C*
DinB	*Damaged induced B*
PDB	*Protein Data Bank*
5'-3' Exo	*5'-3' exonuclease*
8 kDa	*Eight kiloDaltons domain*
PAD	*Polimerase Associated Domain*
PHP	*Polymerase and Histidinol Phosphatase*
3'-5' Exo	*3'-5' exonuclease*
RB69	*Enterobacteria phage RB69*
PIR	*Protein International Resource*
PAM	*Point Accepted Mutation*
BLOSUM	*Blocks Substitution Matrix*
DALI	*Distante Alignment*
CE	*Combinatorial Extension*
MAMMOTH	*Matching molecular models obtained from theory*
SALIGN	*Strutural Alignment*
MUSTANG	*Multiple Structural Alignment*
CATH	*Class Architecture Topology Homology*
SCOP	*Structural Classification of Proteins*

Resumen

La clasificación de proteínas en familias comienza con el trabajo propuesto por Dayhoff en la década de los 70. Desde entonces la clasificación basada en secuencia ha sido el método principal para el agrupamiento de proteínas. En general las familias proteicas manifiestan una diversidad contenida en sus secuencias, no obstante existen casos conocidos donde la divergencia puede llegar a los límites de la detección por parte de los métodos basados en secuencia. El hallazgo de Chothia y Lesk en la década de los 80 (la estructura es más conservada que la secuencia), permite la introducción de la información estructural en los sistemas de clasificación de proteínas. Sin embargo su uso se vio restringido por la gran diferencia entre estructuras y secuencias disponibles. Pese a que esta diferencia aún existe, en la actualidad es posible encontrar familias de proteínas para las cuales existe una gran cantidad de estructuras resueltas y que además poseen gran divergencia a nivel de secuencia (*e.g* globinas, polimerasas de ADN). Estas familias constituyen buenos modelos para mejorar nuestro entendimiento de la relación secuencia/estructura/función en proteínas. De particular interés son las polimerasas de ADN, debido a que participan en una función esencial en los sistemas biológicos, constituyen un gen muy antiguo en la historia de la evolución, cuentan con una gran cantidad de estructuras disponibles, y además concentran un interés biotecnológico asociado.

Inicialmente, las polimerasas de ADN fueron clasificadas en tres familias (A, B y C) de acuerdo a la similitud a nivel de secuencia con tres genes de polimerasas de *E. coli* (polA, polB y polC). A principios de los 90, se efectúa la primera revisión de su clasificación, resultando en una extensión a las familias D y X (polimerasas pertenecientes

a euriarqueotas y similares a polimerasa beta de humanos, respectivamente). A fines de esta década una nueva extensión de la clasificación se hace necesaria con el descubrimiento de polimerasas involucradas en la reparación de daño por radiación UV (familia Y).

En síntesis, actualmente se acepta una clasificación en 6 familias de polimerasas (A, B, C, D, X e Y), mediante un criterio basado en la similitud de sus secuencias únicamente. Miembros de distintas familias comparten muy baja similitud entre ellos, por debajo de los límites de detección confiables. También se ha observado que dentro de una misma familia es posible detectar gran diversidad de secuencias.

En la literatura se ha demostrado cuantitativamente que alineamientos de secuencias muy disímiles (porcentaje de identidad de secuencia menor a 30%) contienen un número considerable de aminoácidos mal alineados. En consecuencia, surge el cuestionamiento de la validez de la clasificación actual de las polimerasas de ADN, dado que ésta ha sido construida en base a comparaciones entre proteínas remotamente relacionadas que presentan una baja similitud de secuencia. En contraposición a esta diversidad a nivel de secuencia, el depósito de estructuras cristalizadas de polimerasas revela similitudes importantes a nivel estructural entre enzimas que pertenecen a distintas familias. Por tanto, es posible que existan relaciones entre familias de polimerasas de ADN que aún no han sido descritas debido a una limitación tecnológica en el estudio de tales relaciones.

En esta tesis doctoral, se realizó un estudio sistemático de las relaciones estructurales en los dominios catalíticos de polimerasas de ADN utilizando herramientas avanzadas de comparación estructural (*i.e.* alineamientos estructurales). Considerando que la estructura de las proteínas es más conservada que su secuencia, este tipo de técnica permite indagar en relaciones de similitud estructural y evolutivas de proteínas con alta

divergencia a nivel de sus secuencias, como es el caso de las polimerasas de ADN. Además esta información puede ser utilizada para mejorar la relación señal-ruido en alineamientos de secuencia incrementando la sensibilidad y especificidad en la detección de relaciones remotas.

Los resultados obtenidos muestran que a nivel de cadenas completas y dominios catalíticos funcionales, la clasificación propuesta en familias sigue siendo válida, aunque la variación estructural observada es mayor que la esperada inicialmente, mostrando que existe diversidad dentro de familias tanto a nivel de secuencias como a nivel estructural. Por otra parte, a nivel de dominios estructurales, las palmas corresponden al dominio más conservado, detectándose relaciones entre las palmas de las familias A,B e Y, mientras que las palmas de las familias C y X forman grupos estructurales independientes. La comparación del dominio de pulgares y dedos muestran que éstos corresponden a los dominios con mayor grado de variación estructural y de secuencia. Adicionalmente, la clasificación de polimerasas de ADN no se respeta cuando se realizan agrupamientos con esta clase de dominios estructurales. Por otra parte se detectaron relaciones cruzadas entre dedos y pulgares de diferentes familias. En el caso de los dedos de la familia A y pulgares de la familia B, los datos obtenidos sugieren que estos segmentos no son análogos estructurales. Por otra parte los dedos de la familia X y pulgares de la familia Y presentan una similitud estructural que además pudo ser comprobada mediante análisis de alineamientos de múltiples secuencias derivados de la estructura. No obstante, esta relación no pudo ser recuperada cuando se emplearon alineamientos de secuencia derivados por métodos tradicionales, mostrando que la estrategia de alineamientos estructurales mejora la relación señal-ruido en la detección de relaciones remotas. Finalmente, los resultados de

esta comparación cruzada de dedos de familia X y pulgares de familia Y, sugieren que estos segmentos son análogos estructurales.

Por último la información recopilada de las comparaciones estructurales fue resumida en un mapa de relaciones estructurales de polimerasas de ADN donde se muestran no sólo las relaciones a nivel de dominios estructurales, sino también complementada con los dominios accesorios presentes en polimerasas de ADN. La evidencia recopilada, sugiere que las cinco familias de polimerasas para las que se tiene estructura se habrían originado en tres eventos independientes, lo que se fundamenta en la existente de tres grupos estructurales principales: el grupo formado por polimerasas de las familias A, B e Y, las polimerasas de la familia C y las polimerasas de la familia X.

Abstract

Protein classification in families started with the work proposed by Dayhoff in the seventies. Since then, protein classification systems have been based mainly in sequence information as principal source of information. Proteins within a family exhibit a contained degree of divergence in their sequences, however there are cases where divergence can reach the limit of detection by sequence analysis techniques. The contribution made by Chothia and Lesk in the eighties (structure is more conserved than sequence), allowed the use of structural information in protein classification systems. However, the use of this kind of information was restricted by the great gap between the amount protein structures crystalized compared with the number of sequences available. The gap still remains, but some families have an increased number of structures solved (some of these families also have great amount of sequence divergente). These families are good models to improve our understanding of the sequence/structure/function relationship in proteins. Of particular interest are the DNA polymerases familiy, because they perform an essential function in biological systems, they are a very ancient gene in the history of evolution, there are a vast number of structures available, and finally, biotechnological applications are associated with this type of enzimes.

DNA polymerases were classified in three families (A, B and C) according to the similarity of their sequences with three polymerases genes present in *E. coli* (polA, polB and polC). In the beginning of the nineties, the first revision to their classification was made, giving the extension to the families D and X (polymerases of euriarcheotas and those similar to polymerase beta of human). At the end of this decade, a new extension to the

classification was made with the creation of family Y, that comprise all polymerases involved in repairing UV-damaged DNA.

Today, a classification of DNA polymerases in six families is currently accepted. This classification is based only in the similiarity of their sequence. Members of distinct families share very low sequence identity, which is under the limit of detection by classic sequence analysis methods. Also, members within the same family share very low sequence identity.

Aligments of highly dissimilar sequences (with a sequence identity below 30%) contain a great number of misaligned positions. As a consequence, one can ask about the reliability of current classifications, specially in families with high sequence divergence, like DNA polymerases. For that reason, it is posible that some relatioships among DNA polymerases of the same family, or between polymerases of different families have been not detected by current sequence analysis methods.

In this doctoral thesis, a systematic study of the structural relationships of the catalytic domain of DNA polymerases was performed, using advanced structural comparison tolos. Because protein structure is more conserved than sequence, this approach allows the exploration of structural and evolutionary relationships of proteins with high sequence divergence. Also this information can be used to improve sequence alingments improving the signal-noise ratio and subsequently improving the sensibility and specificity in the detection remote relationships.

The results obtained show that at chain and catalytic domain level, the current classification of DNA polymerases is still valid, but the degree of structural variation is higher than expected. At structural domain level, the palm domain is the most conserved

domain. Palms of families A, B and Y are closely related to each other, while palms of families C and X form independent structural groups. Fingers and thumbs domains are the most variable domains at structural level.

Relationships between fingers and thumbs from different families were detected. In the case of family A fingers and family B thumbs, the obtained results suggests that these segments are not structural analogues. Fingers of family X and thumbs of family Y share a structural similarity that can also be recovered using structural alingments, but cannot be recovered using classic sequence analysis techniques, showing that structural alignments improve the signal-noise ration in the detection of remote relationships. Finally, the results of these comparison suggest that family X fingers and family Y thumbs are structural analogues.

Finally, the information collected was synthezised in a map of structural relationships of DNA polymerases at domain level, but also complemented with accesory domains present in different polymerases. The relationships depicted, suggest that the five families studied in this work should have arisen in at least three independent events in the history of evolution, which is supported by the existance of three structural groups: the first group formed by polymerases of families A, B and Y, the second group formed by polymerases of family C and the third groups formed by polymerases of family X.

Agradecimientos

Quisiera expresar mi más profundo y especial agradecimiento a mis padres, quienes me han dado la lección de vida más importante: el amor incondicional, el valor del esfuerzo y el trabajo. Gracias a ustedes por todo lo que han dedicado de sus vidas a su familia, espero que este trabajo los haga sentir felices pues son parte de el.

También expreso mi gratitud a mi tutor de tesis, Dr. Francisco Melo quien siempre ha estado dispuesto a enseñarme y transmitirme su experiencia de manera sincera. Te agradezco profundamente toda tu dedicación y ganas de empujarme en este arduo camino de la ciencia. Gracias por enseñarme el rigor en el trabajo, creo que sin duda es un valor tremendamente importante en el mundo actual donde se quiere hacer mucho pero profundizar poco. Gracias por darme la oportunidad de trabajar en el MeloLab, del cual me llevo una gran cantidad de aprendizajes y hermosos descubrimientos.

Quisiera agradecer también de manera especial a un profesor, colega y amigo, junto al cual, en el último tramo de esta tesis, aprendí cosas muy importantes de la vida, que creo estuvieron pendientes de aprender por mucho tiempo. Gracias querido profesor Santiago Vasconcello por haber sido nuevamente mi profe y también por tu sincera amistad.

A mis compañeros de trabajo en el MeloLab, les agradezco por el diario compartir, por la amistad y el cariño que día a día compartieron con este aprendiz de científico ambulante que paseaba de ciudad a ciudad diariamente trabajando por sus sueños.

Finalmente agradezco a los miembros del comité de esta tesis doctoral, por sus comentarios que contribuyeron de manera muy importante a la corrección final de esta tesis.

Concluyo feliz de haber descubierto y comprendido una fracción mínima de la naturaleza y de haber descubierto nuevamente a Dios en ella.

Financiamiento

1. Beca para estudios de Doctorado, CONICYT, Gobierno de Chile.
2. FONDECYT 1110400.
3. Fundación Ciencia Translacional, Instituto Milenio de Inmunología e Inmunoterapia.

1. Introducción

1.1. La familia de las polimerasas de ADN.

Las polimerasas de ADN corresponden a una familia de proteínas esenciales para la mantención de la vida celular y la preservación de otros sistemas biológicos tal como los conocemos en la actualidad. En términos generales podemos decir que se trata de una familia muy diversa en términos de secuencia, estructura y función. Sin embargo existen aspectos claves que son altamente conservados como su capacidad de incorporar nuevos nucleótidos a polímeros de ADN, así como los mecanismos de reacción empleados para ello (T. A. Steitz, 1999).

Comprender cómo se originó la diversidad y conservación actual de esta familia es una tarea compleja, considerando que se trata de un gen muy antiguo en la historia de la evolución, el cual ha acumulado un gran número de cambios. De esta manera, la información que se pueda obtener al estudiar y comparar secuencias tendrá con certeza una baja relación señal/ruido; en cambio, el estudio a nivel estructural puede mejorar esta relación y entregar información relevante para establecer relaciones.

Hoy existen unas pocas familias de proteínas que poseen muchas estructuras disponibles y que además cuentan con una importante diversidad de secuencia. Entre éstas se encuentran las globinas y las polimerasas de ADN. Estas familias constituyen muy buenos modelos de estudio de la relación secuencia/estructura/función en proteínas.

Las polimerasas de ADN también han sido objeto de atención en relación a sus aplicaciones en biotecnología. Con respecto a esto, es posible encontrar numerosos estudios de mutantes a gran escala para modificar su función en pos de una aplicación particular.

En los siguientes puntos se describirá la clasificación actual de polimerasas de ADN, así como también algunas características estructurales y funcionales conocidas.

1.2. Clasificación de las polimerasas de ADN.

El comienzo del estudio de las polimerasas de ADN se remonta hace poco más de cincuenta años. En aquellos tiempos algunos científicos especulaban que dada la exquisita complejidad de la replicación del ADN, sería casi imposible reproducir dicho proceso fuera de la célula. Sin embargo, en 1956 el grupo de Arthur Kornberg, anuncia la síntesis enzimática de ADN en extractos de *Escherichia coli*. Esta enzima fue denominada polimerasa I (Bessman, Kornberg et al. 1956). Al año siguiente se describe la polimerasa α en eucariontes (en extractos de hígado de rata), la cual tiene baja similitud de secuencia con la polimerasa I de *E. coli*. A fines de la década del 60 y principios del 70, Thomas Kornberg aísla dos nuevas polimerasas en *E. coli*: polimerasas II y III. La primera de ellas es similar en secuencia a la polimerasa α de eucariontes, mientras que la segunda no comparte similitud con ninguna de las polimerasas descritas hasta esa fecha (Knippers 1970; Kornberg and Gefter 1970; Moses and Richardson 1970; Kornberg and Gefter 1971). En estos experimentos también se estableció que la replicación de ADN en esta bacteria se lleva a cabo por la polimerasa III. Paralelamente, Baltimore y Temin describen la transcripción reversa (Baltimore 1970; Temin and Mizutani 1970), lo que marca un hito

importante al demostrar que era posible la polimerización de ADN utilizando ARN como molde. Estos son algunos ejemplos de la prolífica actividad científica asociada a este tema, la que se prolongó durante las décadas siguientes (Friedberg 2005).

En un principio, la clasificación de las enzimas contemplaba dos familias, las que resultaban de agrupar las nuevas secuencias de acuerdo a su similitud con polimerasa I de *E. coli* (familia A), o bien con polimerasa α, la cual es similar a la polimerasa II de *E. coli* (familia B) (Jung, Leavitt et al. 1987). La gran acumulación de datos de secuencias motiva a Ito y Braithwite el año 1991 a revisar la clasificación de las polimerasas de ADN; para ello realizaron alineamientos múltiples de las secuencias disponibles. Sus hallazgos les permitieron proponer una clasificación de tres familias: A, B y C, basados en criterios de similitud de secuencia con las polimerasas I, II y III de *E. coli,* respectivamente (Ito and Braithwaite 1991). En virtud del constante depósito de nuevas secuencias, algunas de las cuales comparten una baja similitud con miembros de las familias antes definidas, se hace necesaria una revisión de la clasificación dos años más tarde. Como consecuencia, la clasificación se extiende estableciendo las nuevas familias D y X (que comprenden polimerasas de euriarqueaotas y homólogos a polimerasa β de humanos, respectivamente) (Braithwaite and Ito 1993).

Hacia fines de la década de los 90, ocurren numerosos descubrimientos de polimerasas que participan en la reparación de regiones lesionadas en el ADN (*e.g.* lesiones ocasionadas por radiación ultravioleta). Estas polimerasas comparten poca similitud con las familias ya descritas. Esto lleva a una nueva extensión de la clasificación, creándose la

familia Y. Estas polimerasas pertenecen a la superfamilia de proteínas UmuC/DinB/Rev1/Rad30, descritas en *E. coli, S. cerevisiae* y *H. sapiens* (Ohmori, Friedberg et al. 2001) (Tabla I).

Explorando más en detalle la clasificación anterior es posible notar que ninguno de los tres dominios de la vida posee polimerasas de las seis familias definidas. En efecto, existen familias que están restringidas a dominios específicos, como son los casos de las familias C y D. Sorprende que un proceso esencial a nivel biológico, y que se supone debiese ser más bien conservado a nivel de secuencia, presente distribuciones tan heterogéneas y de alta diversidad a nivel de secuencia (Koonin 2006). La clasificación presentada en los párrafos anteriores corresponde al estado del arte en esta materia. Pese a tener seis familias claramente definidas, las relaciones entre ellas no han sido suficientemente exploradas. Un estudio de este tipo permitirá establecer posibles conexiones estructurales y funcionales entre estas proteínas, lo que ayudará a comprender uno de los aspectos relacionados con la evolución de la replicación de ADN en sistemas biológicos.

Tabla I Clasificación de polimerasas de ADN aceptada actualmente [1]

Familia	Ejemplos	Función asociada	Dominios
A	Pol γ (mitocondrial). Pol I *E.coli*, *T.aquaticus* Pol. Pol T3, T5 y T7 (fagos)	Replicación del genoma mitocondrial. Procesamiento de fragmentos de Okazaki. Reparación asociada la vía NER. Replicación de genomas de bacteriófagos.	Bacteria, Eurakya, Virus.
B	Pol II *E.coli*. Pol α,β,ε *H.sapiens*. RB69 Pol.	Polimerasas replicativas en eucariontes, arqueas y bacteriófagos.	Bacteria, Eukarya, Archaea, Virus.
C	Pol III *E.coli*	Replicación de genomas bacterianos.	Bacteria.
D	*P. furiosus* Pol	Se sugiere rol en la replicación de algunos genomas de arqueas.	Archaea.
X	Pol β,μ,λ,σ *H.sapiens*. Pol X viral, Pol IV *S.cerevisiae*	Reparación de lesiones en sitios abásicos.	Archaea, Bacteria, Eukarya, Virus.
Y	Pol IV y V *E.coli*	Reparación de lesiones ocasionadas por radiación UV	Archaea, Bacteria, Eukarya.

[1] **(Ohmori et al., 2001; Filee et al., 2002; Rothwell and Waksman, 2005).**

No obstante, explorar las posibles relaciones entre familias a nivel de secuencia es muy complejo dada la divergencia acumulada. Por ejemplo, es posible encontrar algunos casos límite donde se observan porcentajes de identidad de secuencia menores al 20% en polimerasas de una misma familia. En la literatura se ha demostrado en forma cuantitativa que alineamientos pareados de secuencias bajo el 30% de identidad tienen aproximadamente un 20% de residuos incorrectamente alineados (Marti-Renom, Madhusudhan et al. 2004), y que secuencias en la llamada "*Twilight zone*"[2] no consiguen más de un 30% de residuos correctamente alineados (Armougom, Moretti et al. 2006). Por lo mismo, es conveniente preguntarse acerca de la exactitud y validez de la clasificación actual. En este sentido, es razonable pensar que en la propuesta actualmente aceptada, exista una relación ruido/señal elevada, lo que penaliza al momento de conseguir clasificaciones confiables e impide explorar posibles relaciones que conecten diferentes familias.

1.3. Características estructurales de polimerasas de ADN.

Un concepto importante al momento de estudiar características estructurales de proteínas es el de dominio. Se reconocen dos principales definiciones según la disciplina que lo estudie. En biología estructural, los dominios fueron definidos inicialmente como segmentos de una cadena polipeptídica que tienen la capacidad de plegarse en unidades globulares, las que pueden tener funciones especializadas (dominio estructural). En contraste, los genetistas y bioquímicos definieron dominio como el fragmento mínimo de un gen (usualmente identificado mediante experimentos de deleciones parciales en

---

[2] Se denomina *Twilight zone* a aquella región donde pares de secuencias tienen un porcentaje de identidad menor al 20%. En esta región puede existir similitud estructural, pero no está garantizada.

segmentos de ADN), que es capaz de realizar una función determinada (dominio funcional).

El primer estudio de dominios estructurales fue llevado a cabo en la década de los 70. En aquel entonces se utilizó la inspección visual de estructuras resueltas por cristalografía de rayos X. Para identificarlos se buscaron unidades globulares compactas conectadas de manera holgada entre ellas. Al poco tiempo se utilizó una definición más estricta basada en el cálculo de distancias entre carbonos alfa de la proteína. Las distancias medidas, al ser representadas en un gráfico bidimensional, permiten buscar grupos de carbonos-alfa compactos con distancias cortas, los que representan dominios estructurales. Actualmente, el concepto detrás de los programas que identifican dominios estructurales en proteínas es que las interacciones atómicas dentro de un dominio son mayores que entre dominios. Luego, la metodología para identificar dominios en una estructura consistiría en ubicar grupos de residuos con un número máximo de contactos entre ellos y con un número mínimo de contactos entre residuos de distintos grupos. Por lo demás, esta aproximación también fue utilizada para predecir regiones de una proteína estables por sí mismas y capaces de plegarse de manera independiente.

La definición revisada parece bastante razonable cuando los dominios identificados son continuos en la estructura primaria de la proteína. Sin embargo, es frecuente encontrar casos donde lo anterior no se cumple, puesto que las unidades compactas desde el punto de vista estructural se encuentran repartidas a lo largo de la estructura primaria, con múltiples interrupciones producto de la inserción de otros segmentos. A los dominios estructurales

que ocurren de manera continua en la estructura primaria los llamaremos dominios monosegmento, mientras que a los que ocurren de manera interrumpida en ésta los denominaremos dominios multisegmento.

En las polimerasas de ADN el dominio que históricamente ha recibido mayor atención es el dominio catalítico. Este dominio es el responsable de la actividad de polimerización y, de acuerdo a la definiciones expuestas en los párrafos anteriores, corresponde a un dominio de tipo funcional. Este dominio catalítico tiene una estructura tridimensional característica y conservada. La arquitectura fue comparada a la de una mano derecha (Joyce and Steitz 1994), y es posible reconocer segmentos de la estructura que son análogos a las distintas partes de la mano: dedos, palma y pulgar (Figura 1). Estos segmentos corresponden a dominios de tipo estructural y pueden ser del tipo mono o multisegmento. El dominio de la palma es el central puesto que en él se encuentran los aminoácidos responsables de la catálisis de la reacción de polimerización del ADN a través de un mecanismo conservado que involucra a dos iones bivalente (T. A. Steitz & Steitz, 1993), mientras que los dedos y el pulgar tienen un rol en acomodar la doble hélice ADN en una orientación adecuada para que ocurra la polimerización. Tanto el pulgar como los dedos interactúan a través del surco mayor del ADN mediante interacciones inespecíficas con el esqueleto de fosfatos de la molécula de ADN (T. A. Steitz, 1999).

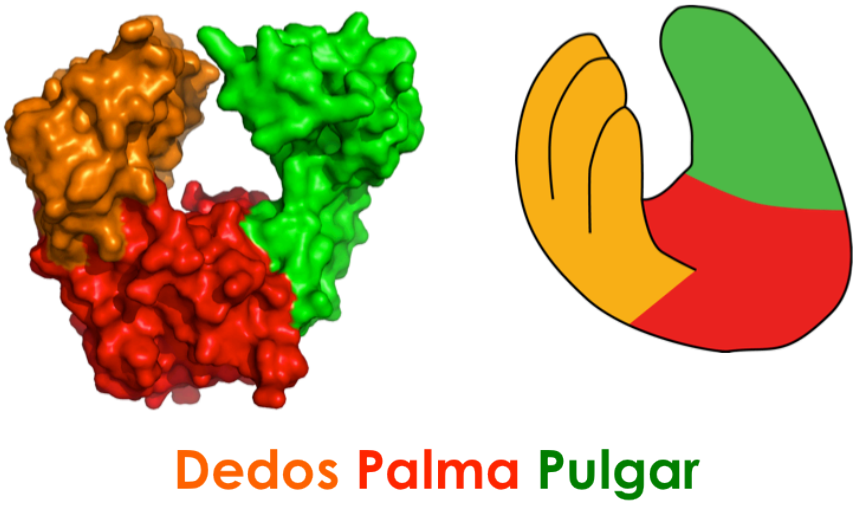


Figura 1. Analogía de la estructura del dominio catalítico de polimerasas de ADN. Analogía estructural del dominio catalítico de polimerasas de ADN con una mano derecha. A la izquierda se presenta el dominio catalítico de la polimerasa I de *E. coli* (código PDB 1taq) con representación de la superficie de la proteína. En ella se han destacado con los colores presentes en la base de la figura los diferentes dominios estructurales presentes. En color naranja se destaca el dominio estructural de los dedos, en rojo la palma, y en verde el pulgar.vA la derecha, una representación esquemática de la analogía de los diferentes dominios estructurales a las partes de una mano derecha según lo propuesto por Steitz.

Además de los dominios palma, pulgar y dedos, es posible reconocer una serie de dominios accesorios del tipo estructural. La presencia de estos últimos es variable según la familia de polimerasas de ADN que se esté estudiando. En este tipo de dominios es posible encontrar funciones tales como: 5'-3' exonucleasa, 3'-5' exonucleasa y fosfoesterasa. Los dominios accesorios junto con el dominio catalítico forman una cadena proteica continua, llamada en adelante cadena completa. El estudio realizado en esta tesis se concentró en tres niveles principales de la estructura terciaria: cadenas completas, dominio catalítico y dominios estructurales (Figura 2).

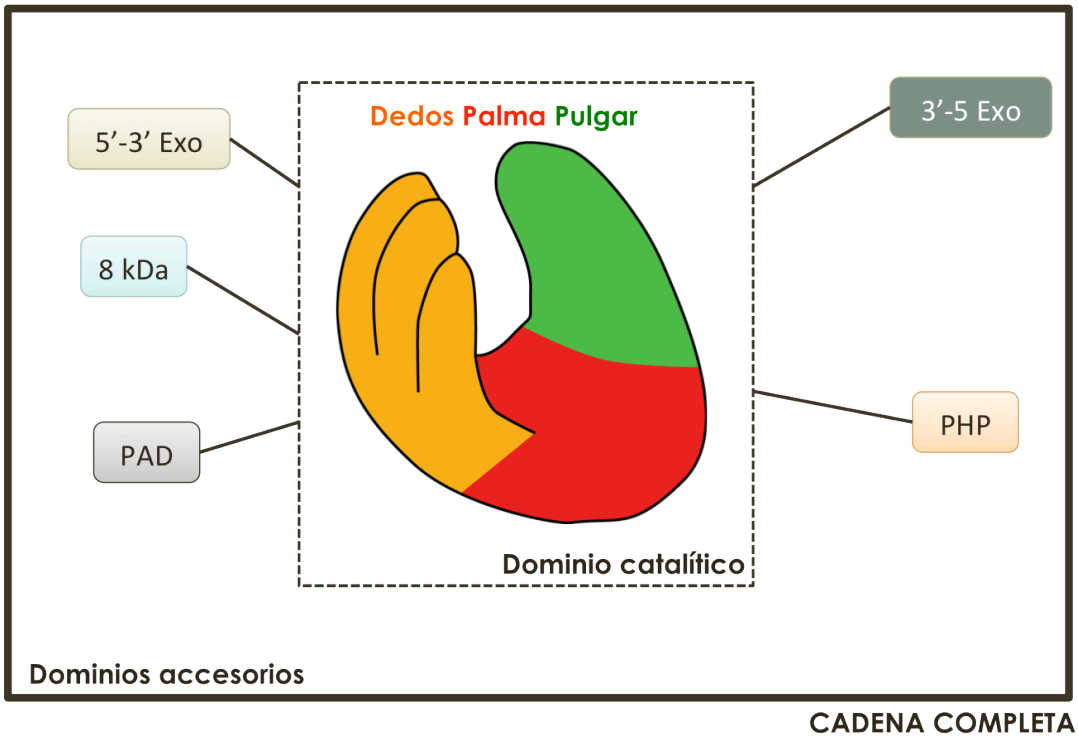


Figura 2. Dominios funcionales y estructurales presentes en polimerasas de ADN. El dominio catalítico de polimerasas de ADN se encuentra representado mediante la forma de una mano derecha y encerrado en un cuadro de línea discontinua. En la mano se han destacado con colores los tres dominios estructurales que conforman este dominio funcional (fingers, palm y thumb). Fuera del cuadro de línea discontinua se encuentran una serie de cajas que representan a los dominios accesorios que pueden ser encontrados en las distintas familias de polimerasas de ADN, los que se conectan al dominio catalítico. El cuadro más externo de línea continua representa a la unión del dominio catalítico y sus dominios accesorios, los que conforman la cadena completa de la polimerasa de ADN. En color gris PAD: Polymerase associated domain, en color celeste 8kDa: 8 kiloDaltons domain, en color damasco PHP: Phosphosterase domain, en color café claro 5'-3' Exo: 5'-3' Exonuclease domain, y en color verde grisáceo 3'-5': 3'-5' Exonuclease domain.

Conforme se fueron cristalizando ejemplares de otras familias se observó que la arquitectura del dominio catalítico era conservada. Sin embargo, hasta la fecha no se ha realizado ningún análisis a gran escala de las relaciones de similitud entre los diferentes ejemplares de una misma familia, así como tampoco entre ejemplares de diferentes familias. Para llevar a cabo un análisis de esta clase, se requiere del uso de herramientas avanzadas, tales como los alineamientos estructurales. Esta clase de alineamientos permite identificar relaciones en proteínas que poseen alta divergencia en sus secuencias, como es el caso de las polimerasas de ADN, debido a que la estructura de las proteínas es más conservada que su secuencia (Chothia & Lesk, 1986). Sin embargo, la comparación de estructuras de proteínas tiene algunas desventajas en términos del tiempo de cálculo requerido, comparada con el alineamiento de secuencias. Pese a ello, algunos algoritmos se especializaron en resolver estas dificultades generando soluciones muy rápidas que compiten directamente con los algoritmos basados en secuencias únicamente (Ortiz, Strauss, & Olmea, 2002). Un ejemplo de este tipo de programas es MAMMOTH, que permite manejar de manera simple una lista de estructuras (archivos en formato PDB) para ser comparados a gran escala y, posteriormente, generar clasificaciones basadas en información estructural. De acuerdo a lo anterior, pareciera ser que los avances logrados ofrecen una solución prácticamente completa a esta tarea, sin embargo existen excepciones.

En la literatura científica actual, es posible encontrar con alta frecuencia, publicaciones basadas en análisis a gran escala. Poco se ha indagado en la confiabilidad y reproducibilidad de estos resultados, pues es factible cometer errores asociados a la ejecución o bien por casos extremos para los que un algoritmo no está preparado. Más

importante aún, es la posibilidad real de detectar dichos errores, de manera tal que los resultados y conclusiones derivadas sean correctas. Un ejemplo de esta situación fue observado por nosotros en un análisis preliminar de los dominios catalíticos de polimerasas de ADN. En una comparación a gran escala de dominios palma de polimerasas de ADN se identificó un caso donde proteínas clasificadas como miembros de una misma familia (Polimerasa RB69 y Polimerasa Phi29, ambas familia B) reportaron valores de similitud estructural muy bajos cuando fueron alineadas con el programa MAMMOTH. Al inspeccionar la superposición, se identificó como región de similitud estructural sólo una alfa-hélice que corresponde a la base de la palma. Ésta similitud estructural no corresponde a la superposición reportada en la literatura para este par de proteínas (Kamtekar et al., 2004). Al reproducir los resultados del artículo utilizando el programa DALI (Holm & Park, 2000), se logró obtener el alineamiento reportado, donde prácticamente la totalidad de los aminoácidos se encuentran alineados estructuralmente (Figura 3). Si se pretende generar una clasificación estructural de proteínas, problemas como el mencionado anteriormente pueden tener un impacto importante en ella, sobre todo si se piensa en realizar un número importante de comparaciones que no pueden ser verificadas manualmente. De lo anterior aparece la necesidad de efectuar dichas comparaciones con más de un programa de alineamiento estructural, de manera que se pueda comparar los resultados producidos por ellos y escoger aquel que identifique de mejor manera la similitud entre las dos estructuras. Sin embargo, esta tarea no puede ser realizada directamente, pues los resultados entregados por los programas de alineamiento estructuras no son directamente comparables. En el desarrollo de esta tesis se abordó este problema con mayor profundidad.

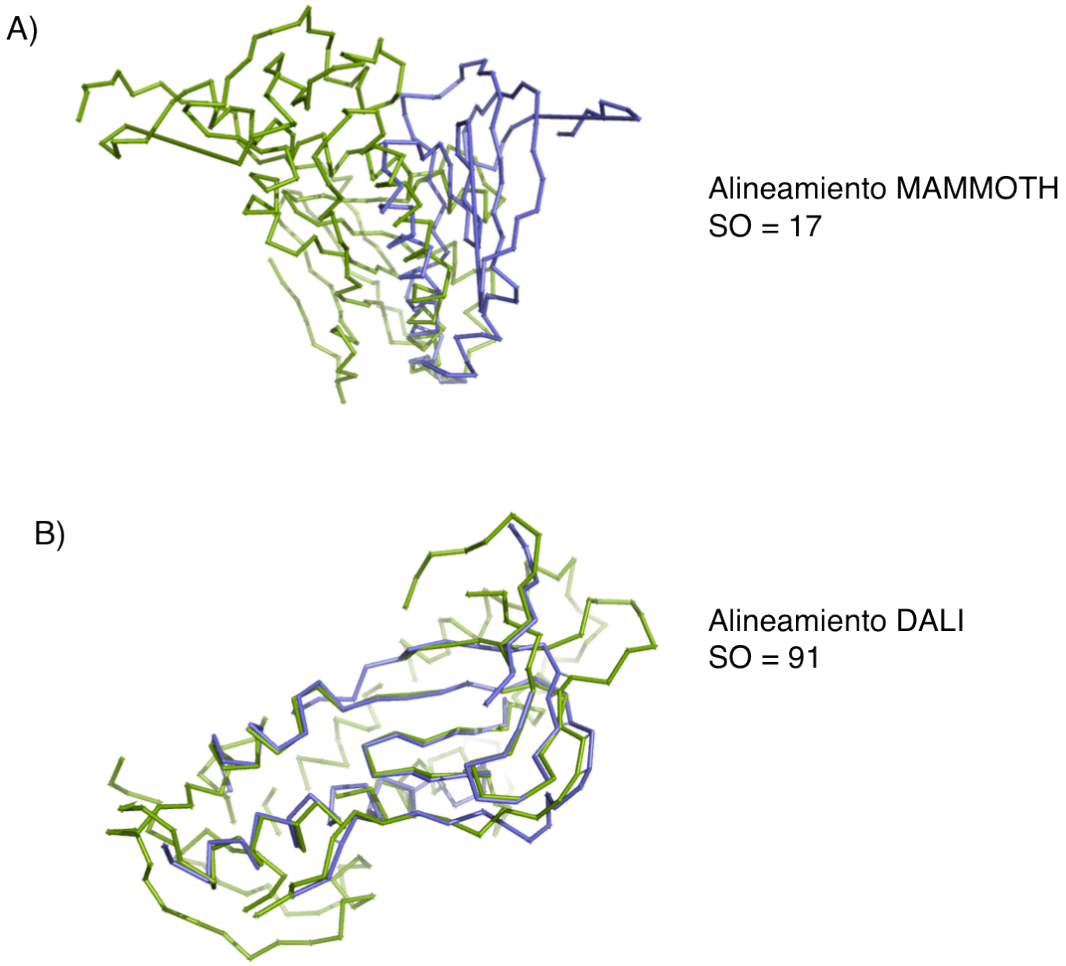


Figura 3. Caso extremo de diferencia en la comparación de dominios palmas de polimerasas de ADN de una misma familia. Se muestra un caso ejemplo de la diferencia extrema en la comparación de dos palmas de polimerasas de ADN de la familia B, utilizando dos programas distintos. En ambos casos se encuentra representada la cadena principal de ambas proteínas. La estructura de la polimerasa Phi29 se colorea en azul, mientras que la del fago RB69 se muestra en verde. A) Superposición óptima entregada por MAMMOTH, el *Structural Overlap* calculado para esta superposición es de 17%. B) Superposición óptima reportada por DALI, cuyo *Structural Overlap* es de 91%.

El aspecto funcional de las polimerasas de ADN también es diverso. Hoy en día se reconoce que las polimerasas de ADN no sólo participan en la replicación de los genomas, sino también en los procesos de reparación. Estas funciones se encuentran también repartidas entre las distintas familias de polimerasas. No obstante, dentro de las características más interesantes de esta clase de enzimas es su selectividad para la incorporación de nucleótidos. Esta propiedad, conocida como fidelidad de copiado, se expresa como la razón entre las eficiencias de incorporación del nucleótido correcto versus el incorrecto con respecto a la hebra templado sobre la cual se realiza la polimerización. Al observar el proceso replicativo en su totalidad podemos distinguir al menos tres elementos responsables de la fidelidad de la replicación: dominio catalítico de polimerasas de ADN, actividad correctora exonucleasa y factores accesorios. Estos tres elementos en conjunto dan cuenta de la tasa de error existente durante la replicación. En una revisión reciente, se resume la importancia relativa de cada uno de estos elementos (Joyce & Benkovic, 2004). El dominio catalítico por sí solo permite obtener tasas de errores de hasta 10^{-5}. El resto de los componentes mencionados permite completar el 10^{-9} observado (Kunkel & Bebenek, 2000).

Conocer los determinantes estructurales de la fidelidad de copiado en el dominio catalítico de las polimerasas de ADN, además de tener un interés científico, también posee uno de tipo tecnológico. En este sentido en la literatura es posible encontrar numerosos trabajos donde se han realizado *screenings* de mutantes que puedan tener un incremento o un decremento en la fidelidad de copia. Sin embargo, muy pocos trabajos han explotado una aproximación racional al problema tratando de identificar y resumir los determinantes

de la especificidad, permitiendo el diseño racional de mutantes y su posterior evaluación.

En síntesis de lo revisado en los puntos 1.2 y 1.3, podemos decir que las polimerasas de ADN se encuentran clasificadas en 6 familias distintas a partir de la información de sus secuencias y que, dada su gran diversidad, es difícil poder explorar relaciones entre familias e incluso dentro de una misma familia. Lo anterior deja abierta la posibilidad de cuestionar la clasificación actual. A nivel estructural se reconocen algunas características conservadas, sin embargo no existe un estudio sistemático de las relaciones estructurales intra e interfamilias. Este estudio debe considerar la posibilidad de utilizar varios programas de alineamiento estructural, ya que algunos de ellos pueden entregar resultados artificiales o erróneos. En los puntos siguientes de la introducción, se revisarán en mayor profundidad aspectos teóricos relacionados al problema de la clasificación de proteínas y de la comparación estructural de éstas.

1.4. El problema de la clasificación en biología.

El ser humano en su proceso de comprensión de la naturaleza requiere realizar procesos de clasificación de entidades. La clasificación requiere comparar dichas entidades en función de un parámetro que permita distinguir clases de entidades. Una vez establecida la clasificación, es posible establecer relaciones de orden superior entre dichas entidades.

En la perspectiva del conocimiento biológico, el proceso descrito anteriormente se aplica de la misma manera. De hecho, los biólogos desde los tiempos de Linneo, han procurado sistematizar el proceso de clasificar diferentes clases de entidades biológicas. Dependiendo del tipo de ésta última, se han diseñado diferentes sistemas y métodos de

clasificación. En la actualidad, este problema sigue siendo esencialmente el mismo, sin embargo es más común ver que el tipo de entidades biológicas sobre el cual se trabaja son moléculas tanto a nivel de su estructura como los monómeros que componen su secuencia.

Dado lo anterior, las condiciones bajo las cuales se trabaja este problema ciertamente han cambiado. El proceso ya no parte de la medición de rasgos discretos, cualitativos y humanamente definidos, por el contrario, es frecuente encontrar definiciones de rasgos más concretos y precisos que pueden ser medidos incluso mediante una función matemática continua de carácter objetivo. De acuerdo a esto último, es posible entregar una respuesta precisa del grado de similitud entre dos entidades biológicas.

Según lo expuesto en los párrafos anteriores, pareciera ser entonces que esta tarea no representa una dificultad, sino que se trata de una tarea fácilmente sistematizable y abordable, al momento que se cuenta con una definición precisa de lo que se espera comparar para clasificar.

Sin embargo, esto último es lo que precisamente provoca numerosos problemas, pues las características de los sistemas biológicos a sus distintos niveles de organización, distan de ser fácilmente definibles y reducibles a una cantidad discreta. En los siguientes puntos será descrito un tipo particular de este problema en biología a nivel molecular, el cual tiene que ver con la clasificación de estructuras de proteínas. Para adentrarse en esta problemática, se revisarán diferentes métodos de agrupamiento y técnicas de comparación de estructuras moleculares.

1.5. Agrupamientos o *clustering*.

Definiremos un agrupamiento como la clasificación no supervisada de patrones (observaciones, ítemes de datos numéricos o vectores de características) en grupos (en adelante *clusters*). De manera intuitiva, los patrones que ocurren dentro de un *cluster,* son más similares entre sí que con aquellos pertenecientes a un *cluster* distinto. Por ejemplo en la Figura 4(A), se muestra un patrón de entrada para el agrupamiento de ciertos elementos. Estos elementos no se encuentran etiquetados y se espera organizarlos de acuerdo a la similitud de sus patrones. En la Figura 4(B), se enseña el resultado de este agrupamiento con la asignación de los distintos elementos a un *cluster*.

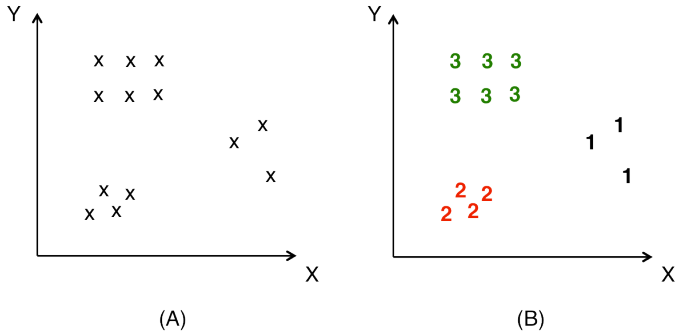


Figura 4 Agrupamiento de datos. En la figura se muestra un ejemplo del carácter intuitivo del agrupamiento de datos. En el gráfico (A) se muestran 13 elementos distribuidos en un plano X,Y sobre los cuales no se conoce su pertenencia a un grupo determinado. En el gráfico (B) los mismos elementos fueron asignados a diferentes grupos, lo que se encuentra indicado mediante un número y color.

Antes de continuar es importante mencionar la diferencia existente entre el agrupamiento (clasificación no supervisada) y el análisis discriminante (clasificación supervisada). En la primera, la tarea es clasificar un conjunto de patrones no etiquetados en grupos significativos. Esta clasificación se consigue únicamente a partir de los datos de los elementos a clasificar. Por otra parte la clasificación supervisada consiste en incorporar un nuevo elemento a un conjunto de datos previamente etiquetados en grupos, es decir, establecer la similitud de uno o más elementos con un patrón previamente establecido.

En lo que respecta al agrupamiento propiamente tal, se reconocen dos tipos principales: **particional** y **jerárquico**. El agrupamiento particional, como su nombre lo indica, realiza una partición sobre todos los datos y de una sola vez (Figura 5A); mientras que el agrupamiento jerárquico produce grupos anidados (uno dentro de otro) con distintos niveles de similitud y se representa mediante **dendrogramas** (Figura 5B).

El agrupamiento de tipo particional, en algunos casos requiere definir *a priori* el número de grupos a los que se dará origen. Esta elección consiste en obtener un número óptimo llevando a cabo una búsqueda combinatorial, que generalmente utiliza una gran cantidad de recursos computacionales, lo que hace al método particional muchas veces inaplicable. Además en biología generalmente esto no es conveniente, dado que lo que se pretende es describir y explorar las características del espacio de similitud de un conjunto de entidades biológicas. Por otra parte, agrupamiento jerárquico presenta ventajas tales como: posibilidad de ser aplicado a grandes cantidades de datos e insensibilidad al orden de entrada de los datos.

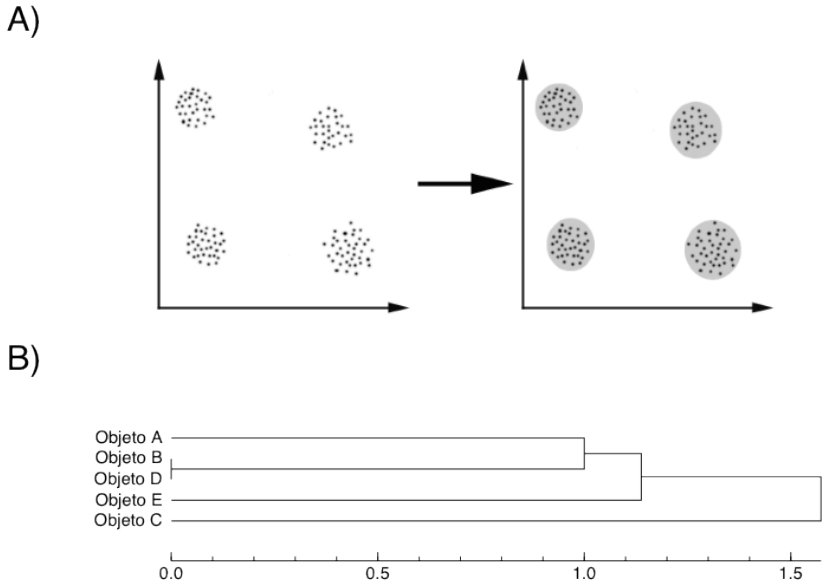


Figura 5. Diferencias entre un agrupamiento particional y uno jerárquico. A) Agrupamiento de tipo particional. El gráfico de la izquierda muestra un conjunto de puntos sobre un espacio bidimensional sobre el cual se desea determinar su agrupamiento. A la derecha, la partición del espacio en cuatro grupos discretos considerando la disposición de los puntos genera el agrupamiento particional. B) Agrupamiento jerárquico. Se muestra un dendrograma, que corresponde a la representación de un agrupamiento jerárquico. En él, cinco objetos se conectan mediante nodos a través de ramas cuya longitud está en relación a la distancia que tienen los elementos.

Adicionalmente no es necesario definir de manera previa el número de grupos que se generan en el agrupamiento. Nos centraremos en describir más en detalle el agrupamiento jerárquico que será unos de los principales componentes de esta tesis.

La realización de un agrupamiento jerárquico involucra una serie de tareas que han sido descritas previamente en la literatura (Jain, Murty, & Flynn, 1999):

1. Definición de la proximidad o similitud de los patrones a través de una métrica consistente con el dominio de los datos.
2. Agrupamiento.
3. Representación gráfica del agrupamiento (si se requiere)
4. Evaluación del agrupamiento (si se requiere)

La cuantificación de la distancia entre las entidades es fundamental para la definición de un agrupamiento. Típicamente se define una función matemática que tenga las propiedades de una distancia (llamada en la jerga disimilitud). La distancia es una función $D_{a,b}$ que se aplica sobre dos objetos cualesquiera *a* y *b*. Esta función entrega como resultado un número real. La función de distancia debe cumplir los siguientes axiomas que se presentan en la Ecuación 1:

$$\begin{aligned} D_{a,a} &= 0 \\ D_{a,b} &\geq 0 \\ D_{a,b} &= D_{b,a} \\ D_{b,c} &\leq D_{a,b} + D_{a,c} \end{aligned} \quad \text{[Ec. 1]}$$

Es importante recalcar que normalmente lo que se mide directamente entre dos entidades es su similitud, la que posteriormente puede ser transformada en una distancia. Es importante notar que la similitud debe cumplir los axiomas presentados anteriormente, luego la transformación de la similitud a una distancia requiere adecuaciones matemáticas simples. A modo de ejemplo, cuando se comparan secuencias de proteínas es común utilizar la identidad de secuencia para cuantificar la similitud entre pares de secuencias alineadas. La identidad de secuencia (S*eqIde*) se define como el porcentaje de las posiciones alineadas que son idénticas (Ecuación 2). El rango de esta similitud varía entre 0 y 100, donde 0 significa que de las posiciones alineadas ninguna comparte el mismo carácter, mientras que si es 100, la totalidad de las posiciones alineadas comparten el mismo carácter. Para transformar esta similitud a una distancia bastará con restar a 100 al valor de la identidad de secuencia medida anteriormente (Ecuación 3).

$$SeqIde(A,B) = \frac{Identidades}{N^{o}\ posiciones \quad alineadas} \quad \textbf{[Ec. 2]}$$

$$Dist(A,B) = 100 - SeqIde(A,B) \quad \textbf{[Ec. 3]}$$

El proceso de construcción de agrupamientos jerárquicos, contempla la obtención de distancias para todos los pares posibles, los que se guardan en una matriz de dimensión *N x N*. Esta matriz es la entrada requerida por los algoritmos de agrupamiento jerárquico. En la literatura, existe una gran variedad de algoritmos descritos para la realización de esta clase de agrupamientos (Jain et al., 1999).

1.6. Clasificación de proteínas.

Las diferentes propuestas de clasificación de proteínas utilizan agrupamientos de tipo jerárquico para describir las relaciones de similitud que existen entre ellas. En la actualidad, la principal fuente de información para la construcción de las clasificaciones proviene del análisis de la similitud de sus secuencias. El primer criterio de clasificación aparece en el año 1976, cuando Margaret Dayhoff publica un artículo proponiendo dos niveles jerárquicos para el agrupamiento de secuencias: familias y superfamilias (Dayhoff 1976). Este esquema de clasificación dio origen a la base de datos PIR (Barker et al., 1999), la que sigue siendo mantenida hasta el día de hoy (http://pir.georgetown.edu/pirsf/). Posteriormente, los nuevos desarrollos en el área preservaron la premisa inicial de clasificar a las proteínas de acuerdo a la similitud de sus secuencias, sin embargo incorporaron variaciones tales como restringir los segmentos donde se efectúa la comparación (*i.e.* comparación a nivel de dominios), la derivación de matrices de sustitución para cuantificar la similitud (*e.g.* series PAM y BLOSUM), la búsqueda de patrones y la inclusión de modelos probabilísticos para la detección de homologías remotas (Henikoff and Henikoff 1992; Attwood, Beck et al. 1994; Sonnhammer, Eddy et al. 1997). Dentro de esta última categoría tenemos a los perfiles de secuencia, que al reunir la información procedente de varias secuencias de proteínas de una misma familia pueden tomar en cuenta dicha variación para establecer relaciones más profundas que las que podrían detectarse mediante alineamientos pareados de secuencias. Su efectividad ha sido ampliamente demostrada en la literatura (Gribskov, McLachlan, & Eisenberg, 1987).

En general, la tendencia observada es que la diversidad dentro de una familia de proteínas sea baja, lo cual simplifica el problema de su clasificación. No obstante, existen casos donde las similitudes descienden hasta el límite de lo detectable a nivel de secuencia. En tales casos, la relación ruido-señal se eleva en los alineamientos de secuencia incrementando la dificultad del problema de la clasificación.

Una mirada alternativa al planteamiento anterior proviene de estudios realizados en la década de los ochenta y principios de los noventa, donde se demostró que la estructura de las proteínas es más conservada que su secuencia (Chothia and Lesk 1986; Holm and Sander 1996). Para ello se compararon a nivel estructural y de secuencia un conjunto de proteínas pertenecientes a diferentes familias definidas según los criterios mencionados anteriormente. Este hallazgo impactó a la metodología tradicional de clasificación de proteínas, posicionando a los alineamientos estructurales como una poderosa herramienta para la comparación de proteínas, pues tienen la capacidad de tomar en cuenta aspectos no-locales que reducen la relación ruido-señal. Esto último les permite identificar relaciones entre proteínas que no son posibles de detectar por métodos basados únicamente en la secuencia. Sin embargo estos métodos no fueron adoptados inmediatamente debido al bajo número de estructuras cristalográficas disponibles, comparado con el número de secuencias disponibles.

Una forma de mejorar la calidad del alineamiento de secuencias es realizar una superposición óptima de estructuras, y a partir de éstas obtener un alineamiento estructural. La razón por la que estos alineamientos son más exactos es un efecto de la conservación de

aminoácidos importantes para la estabilidad termodinámica, plegamiento y función de la proteína. Las estructuras varían en forma más lenta que las secuencias, por lo que es posible establecer relaciones de homología remota, aún cuando el grado de similitud de las secuencias que originan dichas estructuras ha caído por debajo de los límites de reconocimiento (Chothia and Lesk 1986; Sippl 2008; Hasegawa and Holm 2009).

Teniendo en cuenta las consideraciones expuestas en párrafos anteriores, es conveniente plantear un esquema de clasificación utilizando información estructural (*i.e.* alineamientos estructurales). La ventaja de utilizar esta aproximación es que no se tienen los errores de los alineamientos a nivel de secuencias, cuando la similitud a este nivel es muy baja (Pei, 2008).

Resumiendo, las polimerasas de ADN poseen una gran diversidad a nivel de sus secuencias. Por lo anterior, la tarea de encontrar nuevos genes de polimerasas de ADN *in silico*, y clasificarlos por metodologías basadas en secuencia puede ser una tarea compleja y no exenta de errores. Si bien en la actualidad los métodos de anotación basados en la búsqueda de patrones locales o firmas de secuencia han mostrado un buen desempeño en la detección de homologías remotas (Elofsson 2002), existen casos reportados donde los métodos tienen un comportamiento diferencial dependiendo de cómo se emplean. Un ejemplo concreto proviene nuevamente de las polimerasas de ADN. En un estudio enfocado a conocer las relaciones evolutivas entre polimerasas de una misma familia, se comenta explícitamente esta dificultad. Los autores señalan que dependiendo de la secuencia que se utilice como sonda de partida en PSI-BLAST (Altschul, 1997; Altschul &

Koonin, 1998) se recuperan secuencias distintas (Fileé, Forterre et al. 2002). Según lo anterior, estos métodos dependen del alineamiento de secuencias y de su calidad para construir un perfil que represente en forma adecuada el patrón a buscar (esto es maximizar la señal por sobre el ruido en el perfil de secuencias). Una forma de construir perfiles que se aproximen de buena manera al requisito anterior, es hacerlos a partir de alineamientos de múltiples estructuras de proteínas. Por los argumentos mencionados anteriormente, estos alineamientos múltiples no contienen los errores que se encuentran en aquellos derivados únicamente de la secuencia.

Otro punto que ha sido poco explorado en la búsqueda de nuevos genes es la posibilidad de emplear parámetros estructurales para la nueva predicción. Los métodos basados en secuencia asignan un valor de significancia estadística a cada coincidencia local de un patrón en una base de datos (*e.g. e-value* en PSI-BLAST). Asimismo, los métodos basados en cadenas de Markov dependen estrictamente del orden del patrón en una dimensión y están fuertemente restringidos a señales locales (Eddy 2004). Tomando las estructuras tridimensionales de proteínas, podemos observar que aminoácidos que se encuentran separados en la secuencia convergen en el espacio, en la conformación nativa de la proteína. En algunos casos estos aminoácidos pueden ser claves en la función. Los métodos basados en secuencia pueden fallar frecuentemente en la detección de este tipo de señales.

Sin embargo, dado que la cantidad de estructuras conocidas hoy en día es varios órdenes de magnitud menor que la cantidad de secuencias de proteínas depositadas en bases

de datos, es necesario recurrir a técnicas de predicción de la estructura tridimensional para explorar los aspectos relacionados a las señales de interacción no local. Hoy en día existen técnicas de modelado de proteínas que reconocen el pliegue de una secuencia de estructura desconocida, alineando dicha secuencia contra el espacio de estructuras conocido en la actualidad, y posteriormente evaluando mediante una función de energía cuán bien se ajusta dicha secuencia a la estructura que se le propone. Esta técnica, que se conoce como *threading* (Miller, Jones et al. 1996), se emplea para modelar secuencias para las cuales no es posible encontrar una estructura con la que comparta más de un 30% de identidad de secuencia (sobre este porcentaje de identidad las estructuras son modeladas mediante modelado comparativo (Marti-Renom, Stuart et al. 2000). Una forma alternativa de emplear el reconocimiento de pliegues sería fijar el espacio estructural de búsqueda, esto es, elegir un pliegue de interés, y luego tomar un conjunto de secuencias provenientes de un genoma y alinear cada una de ellas contra la estructura de interés. De esta forma lo que se intenta es conocer cuales de esas secuencias problemas se ajustarían de mejor forma a dicho pliegue de interés, luego de la evaluación del modelo. Con esta aproximación es posible capturar patrones de interacción no local entre aminoácidos que convergen en tres dimensiones y que pueden ser claves en la estabilidad termodinámica de la estructura, su plegamiento o la función (Panjkovich, Melo, & Marti-Renom, 2008).

1.7. Alineamientos estructurales

Conforme se fueron depositando más estructuras en la base de datos del PDB, fue necesario diseñar métodos que permitieran comparar geométricamente estas moléculas, pues pronto se hizo evidente que la organización tridimensional contenía información relevante que podía ser interpretada a nivel bioquímico en relación a la función que puede cumplir una proteína, así como en términos de sus relaciones evolutivas.

1.7.1. Definición de un alineamiento estructural.

Un alineamiento estructural viene a resolver el problema de comparar a nivel estructural dos proteínas cuya organización tridimensional es conocida de manera previa por métodos experimentales (cristalografía de rayos X o resonancia nuclear magnética). El propósito de todo alineamiento estructural es identificar los residuos de una proteína que tienen un rol estructural equivalente en ambas estructuras, que se enuncia como aminoácidos estructuralmente equivalentes (Hendrickson, 1979).

Los alineamientos estructurales son especialmente útiles cuando se quiere explorar proteínas que se encuentran distantemente relacionadas en término de sus secuencias. Una forma de mejorar la calidad del alineamiento de secuencias es emplear un alineamiento estructural, obtenido a partir de la previa superposición óptima de estructuras. La razón por la que estos alineamientos son más exactos es un efecto de la conservación de aminoácidos importantes para la estabilidad termodinámica, plegamiento y función de la proteína.

Los alineamientos de secuencias principalmente consideran la optimización de identidades de aminoácidos sin considerar la información estructural. Esto puede generar resultados que cuando se examinan en la perspectiva estructural, carecen de todo sentido (Figura 6). Si se trata entonces de determinar qué aminoácidos cumplen el mismo rol en ambas estructuras, desde luego hay que considerar su ubicación espacial. En este sentido el ejemplo mostrado en la Figura 6 es categórico. Aquí, se comparan los resultados de una alineamiento de secuencias y uno de estructuras. Ambos alineamientos son graficados en la forma de una superposición estructural y de un alineamiento de secuencias. En el caso del alineamiento de secuencia, que optimiza una función que es dependiente del tipo de aminoácido, se obtienen pobres relaciones estructurales. En efecto, en este ejemplo los segmentos alineados tampoco son similares en su composición de estructura secundaria. Por otra parte, incorporar la información estructural (*i.e.* alineamiento estructural), permite identificar no sólo la real relación entre las estructuras comparadas, sino que además se puede incrementar la relación ruido-señal en alineamientos de secuencia.

Este tipo de ejemplo es típico de proteínas que tienen bajos porcentajes de identidad de secuencia entre ellas, donde se sabe que los métodos basados en el uso exclusivo de esta información generan alineamientos defectuosos (Pei, 2008). Sin embargo, es posible utilizar la información estructural para generar alineamientos de mejor calidad en proteínas distantemente relacionadas (*i.e.* alineamientos estructurales). Es interesante considerar que el proceso de construcción de alineamientos estructurales en algunos algoritmos no considera el tipo de aminoácido que se está alineando, sino que se basan de manera exclusiva en el uso de la información de coordenadas atómicas (Ortiz et al., 2002).

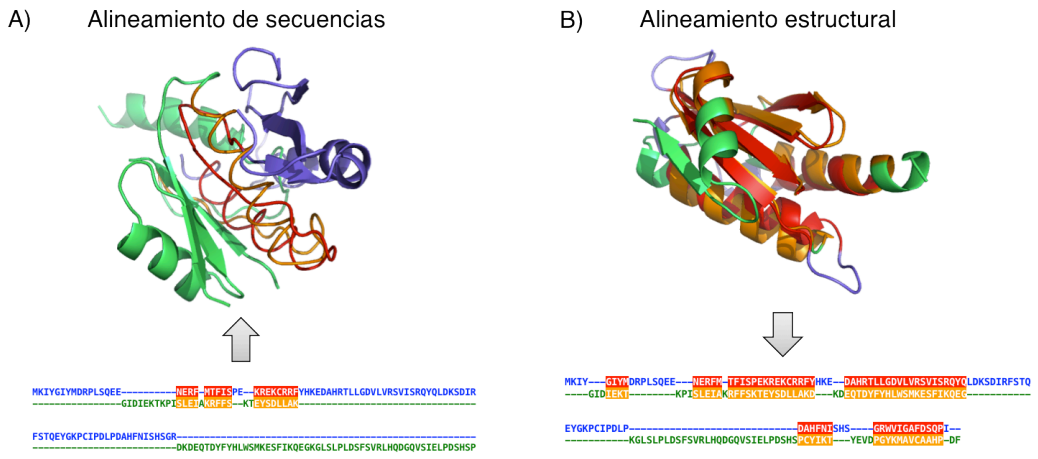


Figura 6. **Comparación entre un alineamiento de secuencias y un alineamiento estructural y los efectos observados cuando se realizan sobre proteínas distantemente relacionadas.** En la figura se compara el alineamiento de secuencias con el alineamiento estructural. Se estudia un par de proteínas distantemente relacionadas a nivel de sus secuencias. Las proteínas se encuentran coloreadas en azul (polimerasa Dpo4, código PDB 2iwm) y verde (polimerasa iota, código PDB 2wtf) tanto en la estructura como en sus secuencias. En el panel A, se muestra la construcción de alineamiento de secuencias (base de la figura) y cómo queda representado éste desde el punto de vista estructural. En el alineamiento de secuencias se destacaron en rojo para la secuencia azul y en naranjo para la estructura verde, los aminoácidos alineados según un algoritmo de alineamiento de secuencias. Esos mismos aminoácidos fueron utilizados para generar una superposición óptima de ambas estructuras que se encuentran representadas en la modalidad de *cartoons* (el mismo código de colores utilizado en la secuencia se aplica acá). En el panel B, se muestra una un alineamiento estructural del mismo par de proteínas. La superposición óptima las estructuras se encuentran en color azul y verde, y los aminoácidos identificados como estructuralmente equivalentes según el algoritmo de alineamiento estructural se encuentran en rojo para la estructura azul y naranjo para la estructura verde. Posteriormente dicho alineamiento estructural se representó en la forma de un alineamiento de secuencias. Los códigos de colores que se aplican son los mismos que los descritos anteriormente.

Desde luego, el problema tiene una importante complejidad, por lo que en sus inicios se intentó tener una representación simple de la proteína que diera cuenta de su estructura terciaria. Para estos efectos es posible aproximarla tomando únicamente los carbonos alfa de la cadena principal. La comparación de dos conjuntos de estos carbonos pertenecientes a dos proteínas diferentes es el proceso del alineamiento estructural.

A continuación se describirá de manera breve en qué consiste un alineamiento estructural. Sin embargo es conveniente advertir que existen numerosos métodos, lo que se comentarán de forma breve más adelante en el texto.

Los primeros métodos y programas de superposición estructural datan de fines de la década de los 70. La comparación de estructuras de proteínas es un problema NP-completo, que se resuelve de manera heurística por todos los métodos. Si bien diferentes heurísticas pueden reconocer pliegues similares, no necesariamente entregan el mismo alineamiento estructural. Los métodos de alineamiento estructural tampoco garantizan que se entregue una solución que sea biológicamente correcta.

A lo largo de la historia se han hecho algunas revisiones del estado del arte de estos métodos, quizás las más relevantes son las publicadas por Orengo y Gibrat (Gibrat, Madej, & Bryant, 1996; W. R. Taylor & Orengo, 1989). En términos generales, los métodos de alineamiento estructural consideran tres a cuatro pasos en el procedimiento:

1) Representar a las proteínas A y B (*e.g.* cadenas polipeptídicas, dominios u otros fragmentos aminoacídicos) en un espacio de coordenadas independiente, para poder compararlas.
2) Comparar estructuralmente A y B.
3) Optimizar el alineamiento entre A y B.
4) Medir la similitud estructural mediante una métrica específica, o evaluar la significancia estadística del alineamiento comparándola con un conjunto de comparaciones aleatorias.

Estos pasos aplican para la comparación pareada de estructuras de proteínas. A continuación se describirán de manera breve los detalles más importantes de los algoritmos de interés en el desarrollo de esta tesis.

En la actualidad existen cerca de cien programas diferentes[3], sin embargo se pueden clasificar en 3 grandes grupos de acuerdo a la forma en que están construidos sus algoritmos. El primer grupo corresponde a aquellos que utilizan la información geométrica contenida en la disposición espacial de sus carbonos alfa (Csaba, Birzele, & Zimmer, 2008; Guerler & Knapp, 2008; L. Martínez, Andreani, & Martínez, 2007), el segundo corresponde a aquellos algoritmos que utilizan la información de la composición de la estructura secundaria para realizar el alineamiento (Birzele, Gewehr, Csaba, & Zimmer, 2007; Leslin, Abyzov, & Ilyin, 2007; Sacan, Toroslu, & Ferhatosmanoglu, 2008), y finalmente el tercer grupo que utiliza una combinación de mapas de contacto e información de la secuencia de aminoácidos para producir el alineamiento estructural (Friedberg et al.,

[3] http://en.wikipedia.org/wiki/Structural_alignment_software

2007; S. Wang & Zheng, 2008). De los algoritmos descritos en la literatura, sólo cuarenta tienen disponible algún programa para ser descargado y ser ejecutado de manera local en un computador. De esta lista, veintidós pueden recibir como entrada pares de archivos PDB para realizar alineamientos estructurales que pueden ser recuperados en la forma de superposiciones óptimas de proteínas. De esta lista de veintidós programas, se escogieron siete programas que representan algoritmos diferentes para la construcción de superposiciones óptimas de proteínas. A continuación, se describen de manera breve cada uno de ellos:

DALI: En el primer paso, se habla acerca de la representación de estructuras de proteínas. En este aspecto es posible encontrar una gran variedad de alternativas. Por ejemplo el programa DALI utiliza matrices de distancia para representar cada estructura que será comparada, específicamente como una matriz bidimensional que contiene las distancias entre carbonos alfa de una proteína, conocidos como mapas de contacto. Las similitudes en las diagonales representan conformaciones similares del esqueleto principal de las proteínas, mientras que fuera de la diagonal corresponden a similitudes de la estructura terciara en general. Luego los mapas de contacto se pueden acomodar de diferente manera uno sobre otro, permitiendo la inserción de *gaps* y finalmente identificando los pares estructuralmente equivalentes entre ambas estructuras. Por ejemplo, DALI para lograr acomodar las matrices de distancia utiliza un algoritmo de ramificación y poda (*branch and bound*). Para evaluar la significancia de alineamientos estructurales, DALI emplea un z-score derivado de una distribución de alineamientos (Holm & Park, 2000)

CE: El caso de CE también emplea matrices de distancia, pero a diferencia de DALI, éstas son de fragmentos octaméricos de la proteína. CE utiliza tres criterios de corte en el proceso de construcción de los alineamientos. El primer criterio de corte sirve para detectar los FAPs (fragmentos alineados de proteínas), el segundo para evaluar la posibilidad de nuevos FAPs candidatos en relación al alineamiento existente y, finalmente, el tercer criterio que evalúa todos los alineamientos para escoger aquellos que son óptimos. Para la evaluación de la significancia de los alineamientos, también se emplea un *z-score* (Shindyalov & Bourne, 1998).

MAMMOTH: por su parte utiliza una representación de los carbonos alfa de la proteína basado en vectores unitarios, concentrándose únicamente en información estructural y descartando información acerca de la secuencia. MAMMOTH utiliza una aproximación heurística que parte con la identificación de un alineamiento de tipo local de los esqueletos de las proteínas a comparar. A partir de ahí continua para encontrar el subconjunto más amplio de residuos que se encuentren bajo un umbral de distancia predefinido en el espacio tridimensional. En caso de MAMMOTH se emplea un *P*-valor que corresponde a la probabilidad de obtener una proporción determinada de residuos alineados. La estimación de los *P*-valores se basa una función ajustada a partir de alineamientos estructurales aleatorios descritos en la literatura (Ortiz et al., 2002).

TM-ALIGN: utiliza la información del esqueleto de la proteína y la información de la estructura secundaria. Por su parte, TM-align comienza su ejecución con tres tipos de alineamientos iniciales. El primero de ellos utiliza los elementos de estructura secundaria y

un algoritmo de programación dinámica. El segundo se basa en la comparación estructural sin *gaps*. Esto se efectúa mediante un deslizamiento de la estructura más pequeña sobre la más grande empleando como métrica el TM-score descrito previamente (Zhang & Skolnick, 2004). El tercer de los alineamientos es una combinación de los dos primeros, pero utilizando una matriz de programación dinámica y penalidades de *gaps*. La evaluación de los alineamientos se hace utilizando el TM-score (Zhang, 2005).

SALIGN: Este programa utiliza programación dinámica para generar alineamientos de pares de estructuras. La matriz de programación dinámica empleada es una combinación lineal de seis matrices de distancias independientes. La penalidad de gap es lineal y su valor es dependiente de los parámetros de apertura y extensión. Las matrices de distancia se basan en diferentes características que se miden dentro de las estructuras moleculares. Estas son: i) el tipo de residuo de acuerdo a lo descrito en la matriz BLOSUM62, ii) distancias ente carbonos alfa de los residuos de una proteína, iii) accesibilidad a solvente de las cadenas laterales categorizadas en tres clases discretas (enterrado, semi-expuesto y expuesto), iv) estructura secundaria asociada a un residuo particular, v) conformación local del residuo en la forma de un dRMSD en torno a una vecindad de 5 residuos y vi) matriz de disimilitudes definida por el usuario que puede ser opcional (Madhusudhan, Webb, Marti-Renom, Eswar, & Sali, 2009).

TOPMATCH: Para la construcción de alineamientos se parte con una serie de bloques que corresponden a alineamientos sin gaps entre segmentos de dos estructuras a comparar. Posteriormente, el alineamiento es construido a partir de la suma de los distintos bloques.

Se utiliza la información de los carbonos alfa para la construcción de los alineamientos ignorando la información de la secuencia presente en ellos. La evaluación de los alineamientos se realiza utilizando varias métricas. La primera de ellas es la longitud del alineamiento que corresponde al total de residuos alineados. La segunda corresponde a una función Gausiana que relaciona el RMSD[4] con la longitud del alineamiento. La similitud de dos estructuras es mayor a medida que se incrementa el número de posiciones alineadas, sin embargo esto tiene como costo que el error medido en el RMSD se incrementa (Sippl & Wiederstein, 2012).

MUSTANG: Este programa utiliza la información contenida en los carbonos-alfa e ignora completamente la información referente a la identidad del residuo. Utiliza representaciones en la forma de matrices de contacto que sirven para buscar segmentos contiguos similares entre dos estructuras que son evaluados a través del RMSD. Una vez obtenida la lista de segmentos contiguos se genera el alineamiento pareado de ambas estructuras. Finalmente se calcula un puntaje de similitud basado en las correspondencias residuo-residuo (Konagurthu, Whisstock, Stuckey, & Lesk, 2006).

1.7.2. Criterios de optimización en alineamientos estructurales.

Una pregunta que aún no tiene una respuesta plenamente satisfactoria es cómo se puede obtener un alineamiento más adecuado u óptimo. La razón por la cual no se tiene una respuesta es porque en este problema se deben satisfacer dos objetivos que se confrontan.

[4] RMSD: *root mean square deviation.*

Uno de esos objetivos corresponde a maximizar el número de residuos equivalentes (*i.e.* la longitud del alineamiento). El segundo corresponde a minimizar el RMSD de esos residuos equivalentes. Evidentemente es posible minimizar el RMSD a expensas de acortar un alineamiento, o bien maximizar la longitud del alineamiento, pero incrementando el RMSD.

Hoy existe una gran cantidad de programas de alineamiento estructural destinados a resolver la misma tarea, que es la superposición pareada de dos estructuras de proteínas (Hasegawa & Holm, 2009; Pei, 2008). Sin embargo, todo este esfuerzo ha sido realizado con definiciones poco precisas y arbitrarias de lo que constituye un alineamiento óptimo. Cada programa entrega sus propios criterios numéricos y puntajes para describir las propiedades de los alineamientos obtenidos. Desde el punto de vista de un usuario, esta situación es confusa, sobre todo si se pretende comparar resultados que provienen de diferentes programas. El estudio y comparación de estructuras de proteínas requiere herramientas de alineamiento estructural, sin embargo no está claro cuales programas producen los resultados más precisos en una situación dada, además de cómo esos resultados pueden ser interpretados (Slater, Castellanos, Sippl, & Melo, 2013).

1.8 Estudio de la relación secuencia-estructura-función en dominios catalíicos de polimerasas de ADN.

En resumen, la información presentada en los párrafos anteriores describe a las polimerasas de ADN como un caso de estudio interesante, pues se trata de familias que poseen una alta divergencia a nivel de secuencias y con una larga historia evolutiva. Finalmente, corresponden a una familia que posee una cantidad importante de estructuras diferentes disponibles en el PDB, lo que permite estudiarlas mediante comparaciones estructurales.

La clasificación actualmente propuesta para esta clase de enzimas, las divide en 6 familias basada en la similitud de sus secuencias. Dado que las identidades de secuencias que comparten polimerasas de una misma familia y de familias distintas son muy bajas, es posible que existan relaciones entre sus dominios catalíticos que no puedan ser detectadas por este tipo de métodos. A la fecha no existe un estudio sistemático que indique cómo se relacionan a nivel estructural, y que explore posibles relaciones evolutivas profundas entre los dominios catalíticos de polimerasas de ADN de diferentes familias. Los alineamientos estructurales de proteínas son una herramienta ideal para explorar esta clase relaciones en secuencias altamente divergentes. Adicionalmente, la información que provenga de los alineamientos de estructuras también puede ser utilizada para mejorar la relación señal-ruido en alineamientos de secuencias para mejorar la sensibilidad y especificidad de detección de relaciones remotas en proteínas.

En esta tesis doctoral se abordó el problema de estudiar de manera sistemática las relaciones estructurales de los dominios catalíticos de polimerasas de ADN, a través de superposiciones de los dominios catalíticos de este tipo de enzimas y alineamientos de secuencias derivados de comparaciones estructurales. Posteriormente, la información recopilada fue interpretada para establecer relaciones estructurales y evolutivas entre los dominios catalíticos de diferentes familias de polimerasas de ADN.

2. Objetivos e Hipótesis

2.1. Hipótesis

La utilización combinada de información estructural y de secuencia permitirá comprender de mejor manera la relación secuencia/estructura/función de los dominios catalíticos de polimerasas de ADN.

2.2. Objetivo General

El objetivo general de esta tesis consiste en estudiar la relación secuencia-estructura-función de los dominios catalíticos de polimerasas de ADN basado en la comparación de sus secuencias y estructuras.

2.3. Objetivos Específicos

1. Creación de una metodología para la comparación estandarizada de alineamientos estructurales provenientes de diferentes programas. Esta etapa consiste en el desarrollo de algoritmo que permita comparar superposiciones óptimas de estructuras de proteínas provenientes de diferentes programas, a través de la construcción de alineamientos estructurales estandarizados. El propósito es poder extraer el mejor alineamiento estructural de un conjunto de alineamientos. De esta manera se pretende maximizar la calidad de la información de similitud estructural que se pueda obtener de una serie de alineamientos estructurales, considerando que no necesariamente un programa pueda

comportarse de manera óptima en forma consistente.

2. Clasificación de los dominios catalíticos de polimerasas de ADN basado en información de estructura y secuencia. Esta etapa consiste en extraer desde la base de datos PDB (*Protein Data Bank*) todas las estructuras resueltas de polimerasas de ADN. Los datos serán sometidos a un proceso de anotación de dominios catalíticos a partir de información contenida en bases de datos de clasificación estructural, bases de datos de secuencias e información disponible en la literatura. Posteriormente, se generarán agrupamientos de los dominios catalíticos basados en secuencia y estructura. El resultado anterior será contrastado con la clasificación basada en secuencia propuesta actualmente.

3. Búsqueda de relaciones estructurales y de secuencia a nivel de cadenas completas y dominios estructurales en polimerasas de ADN. Los grupos generados en el objetivo anterior servirán para la construcción de perfiles de secuencia a partir de alineamientos de múltiples estructuras. Estos perfiles permitirán capturar señales de composición de secuencias a nivel local para dominios catalíticos putativos polimerasas de ADN. Los perfiles, al igual que la información estructural serán utilizadas de manera combinada para la búsqueda de relaciones entre polimerasas de ADN. Adicionalmente se realizarán búsquedas estructurales en todo el espacio de estructuras disponibles para detección de posibles relaciones remotas.

3. Materiales

3.1. Equipos

El procesamiento de los datos de la presente investigación fue realizado en un *cluster* de 5 computadores MacPro, cada uno con 2 procesadores Quad-Core de 2.28 Ghz y a lo menos 6 Gb de RAM, que corren el sistema operativo Mac OS X 10.5 Leopard. También se efectuaron cálculos en un *cluster* de 16 computadores Linux con procesador Pentium IV de 3.4 Ghz y 512 Mb de RAM, que corren el sistema CentOS 5.1. Los análisis posteriores fueron realizados en una estación de trabajo iMac con procesador Intel Core 2 Duo de 2.4 Ghz, que corre el sistema operativo Mac OS X 10.5 Leopard.

3.2. Programas de alineamiento estructural

Para efectuar los alineamientos estructurales de proteínas se utilizaron 6 programas distintos que se detallan en la Tabla II. Esencialmente, todos los programas de alineamiento estructural tienen la capacidad de recibir como entrada dos cadenas de proteínas.

Un segundo recurso para la comparación estructural, disponible como servidor web, se utilizó para buscar todas las estructuras similares a una estructura de consulta en el PDB, calculado en tiempo real. Esta aplicación recibe el nombre de TopSearch. Este servidor no es un programa de alineamiento estructural propiamente tal, y se basa en el algoritmo de TopMatch para la identificación de similitudes estructurales. Disponible en http://topsearch.services.came.sbg.ac.at

Tabla II. Programas de alineamiento estructural empleados

Nombre	Autor	Descarga	Sistema Operativo	Código Fuente
TopMatch	Sippl, MJ	No aplica	Windows, Linux	No
Mustang	Kognarthuturu, A.	http://www.cs.mu.oz.au/~arun/mustang	Windows, Linux, Mac OS	Sí
TM-Align	Zhang, Y.	http://zhanglab.ccmb.med.umich.edu/TM-align	Linux, Mac OS	Sí
Salign	Madhusudhan, MS.	http://salilab.org/salign/	Windows, Linux, Mac OS	No
Mammoth	Ortiz, A.	http://predictioncenter.org/local/mammoth.txt	Linux	Sí
DaliLite	Holm, L.	http://ekhidna.biocenter.helsinki.fi/dali_server/start	Linux	Sí

3.3. Programas de alineamiento de secuencia

BLAST (*Basic local alignment tool*): Se utilizó el programa BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) para búsqueda de secuencias proteicas en bases de datos. También se empleó la variante PSI-BLAST para la búsqueda iterativa de secuencias con homólogas de baja identidad de secuencia. Disponible en: http://blast.ncbi.nlm.nih.gov/Blast.cgi

MAFFT: programa de alineamiento de múltiples secuencias basado en transformadas de Fourier (Katoh, 2002). Código fuente disponible en http://align.bmr.kyushu-u.ac.jp/mafft/online/server/.

ClustalW: programa de alineamiento de múltiples secuencias basado en programación dinámica (Thompson, Gibson, & Higgins, 2002). Disponible en http://www.ebi.ac.uk/Tools/msa/clustalw2/

Para la construcción de perfiles derivados de las secuencias y búsquedas basadas en secuencia se empleó el programa HMMer versión 3 (Finn, Clements, & Eddy, 2011). Disponible en http://hmmer.janelia.org/software.

3.4. Programas escritos de manera especial para el desarrollo de esta tesis

Los programas escritos de manera especial para esta tesis fueron trabajados en los lenguajes de programación Python y C++. También se emplearon las librerías BioPython (Cock et al., 2009) y BOOST, disponibles en http://biopython.org/ y http://www.boost.org/.

CompileAln.py: programa que toma como entrada un conjunto de alineamientos pareados y genera un alineamiento de múltiples secuencias.

StructuralComparer.py: programa en Python que toma como entrada una lista de estructuras en formato PDB y calcula las comparaciones pareadas de todas las estructuras contra todas utilizando un programa de alineamiento estructural definido por el usuario. Entrega como salida una tabla recopilando la información de similitud estructural, superposiciones óptimas y alineamientos estructurales. Puede utilizar la información que sale directamente de los programas de alineamiento estructural o acoplado al programa STOVCA.

STOVCA (Structural Overlap Calculator): programa que toma como entrada un par de estructuras en formato PDB, previamente superpuestas con algún algoritmo de comparación de estructuras proteicas. El programa infiere el alineamiento estructural a partir de dicha superposición sin alterar la matriz de rotación/traslación entregada. Como salida se obtiene el alineamiento estructural y una serie de medidas de la similitud estructural, calculadas a partir del alineamiento óptimo calculado.

3.5. Programas para la visualización de datos

iTOL (*Interactive Tree of Life*): software de visualización y manipulación de árboles filogenéticos (Letunic & Bork, 2006). Disponible en http://itol.embl.de/.

JalView: editor de alineamientos de secuencia multiplataforma, escrito en el lenguaje de programación Java (Clamp, Cuff, Searle, & Barton, 2004). Disponible en http://www.jalview.org/

3.6. Datos

Los datos estructurales fueron extraídos desde la base de datos PDB (Protein Data Bank), versión 3.2 (Rose et al., 2010).

Los datos de secuencias complementarias fueron obtenidos desde la base de datos Swissprot, Uniprot y NR (Boeckmann, 2003), (Bairoch, 2004) y (Pruitt, Tatusova, & Maglott, 2007). Disponibles en ftp://ftp.ncbi.nih.gov/blast/db/

El conjunto de datos de proteínas homólogas distantemente relacionadas fue obtenida desde la base de datos HOMSTRAD (http://tardis.nibio.go.jp/homstrad/). Se obtuvieron luego de aplicar los siguientes filtros a los 9,538 pares disponibles: (i) cada una de las cadenas debe estar en el rango de 100-150 residuos, (ii) el porcentaje de identidad de secuencia debe ser < 25% y (iii) el solapamiento estructural debe estar en el rango 30-75%.

Estos valores fueron determinados sobre los alineamientos entregados por HOMSTRAD (Mizuguchi, Deane, Blundell, & Overington, 1998).

Datos de clasificación estructural de proteínas fueron obtenidos desde la base de datos CATH (Pearl et al., 2003), disponible en http://www.cathdb.info/) y SCOP (Murzin, Brenner, Hubbard, & Chothia, 1995), disponible en http:// http://scop.mrc-lmb.cam.ac.uk/scop/). Los datos fueron consultados a partir de los códigos PDB disponibles de cada estructura. La descomposición automática de dominios fue obtenida desde la base de datos de COPS-*TopDomain* (https://topdomain.services.came.sbg.ac.at) (Sippl, Suhrer, Gruber, & Wiederstein, 2008)

4. Métodos

4.1 Conjunto de estructuras y base de datos

Un conjunto de estructuras de polimerasas de ADN se obtuvo desde la base de datos del PDB (*Protein Data Bank*), versión 3.2 (Marzo 2010). La extracción de datos se llevó a cabo mediante el módulo de búsqueda avanzada, donde se utilizó como palabra clave: *DNA polymerase*. A los resultados obtenidos se les aplicó filtro manual para descartar aquellas estructuras que fuesen factores accesorios de polimerasas de ADN.

Las estructuras filtradas fueron sometidas a un segundo filtro de redundancia por secuencia, descartando todas aquellas proteínas que tuviesen un 75% ó más de identidad de secuencia a nivel de cadena completa.

Sobre este conjunto no redundante a nivel de secuencia se realizó una descomposición de la cadena proteica en dominios estructurales, la que empleó tres fuentes de información. Por una parte se tomó la información proveniente de la asignación realizada en las publicaciones de la resolución de las estructuras. En caso que dicha información no estuviese presente en la publicación, se utilizó la información presente en bases de datos CATH y SCOP. Si la proteína en cuestión tampoco se encontró depositada y asignada en estas bases de datos, se utilizó el algoritmo de descomposición automática *TopDomain* (Figura 7).

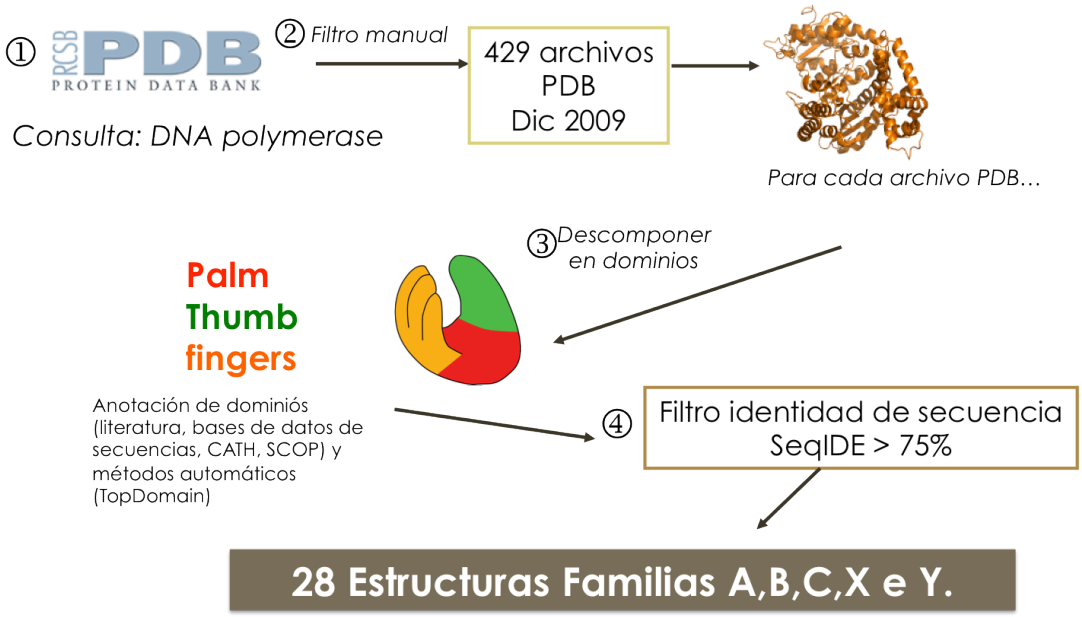


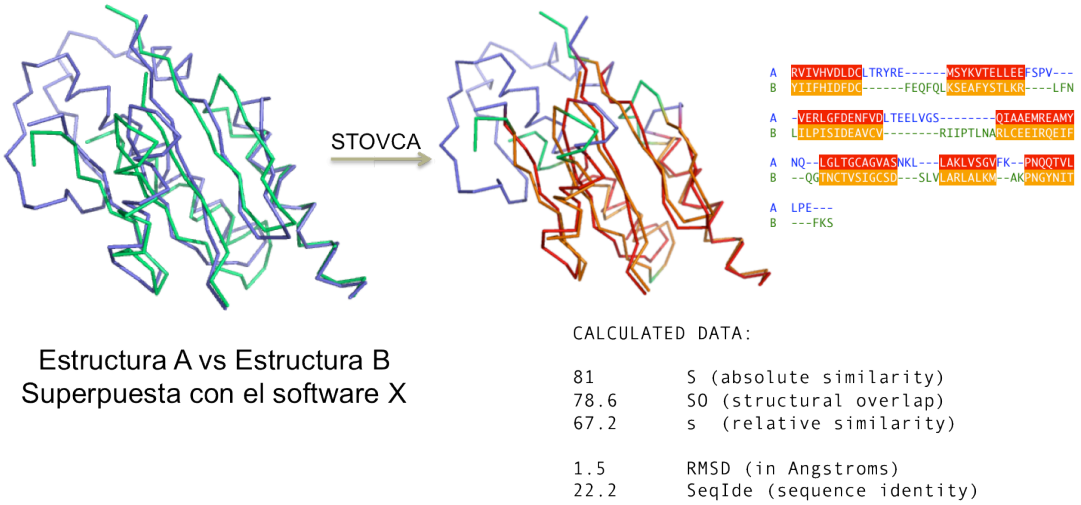
Figura 7. Construcción de la base de datos de polimerasas de ADN. 1) Consulta en la base de datos del PDB utilizando una palabra clave genérica (*DNA polymerase*). 2) Filtros manuales para descartar todas aquellas estructuras etiquetadas como polimerasas de ADN pero que no contienen el dominio catalítico. 3) Cada estructura fue dividida en sus dominios estructurales de palma, pulgar y dedos a partir de información combinada contenida en la literatura, bases de datos y métodos de descomposición automática en dominios. 4) Filtro por identidad de secuencia descartando aquellas secuencias que tengan una identidad superior al 75%.

4.2 Obtención de alineamientos estructurales mediante STOVCA (*Structural overlap calculator*).

Conforme a lo expuesto en la sección 1.7 de esta tesis, se infiere que se requiere un esfuerzo para generar un proceso de estandarización de alineamientos estructurales, considerando la variedad de puntos de vista y definiciones existentes. A continuación se describe una aproximación original para comparar alineamientos estructurales obtenidos de cualquier programa que reporte transformaciones geométricas requeridas para superponer dos estructuras. De hecho, las transformaciones geométricas (que consisten de un vector de traslación y una matriz de rotación), son el resultado esencial de cualquier técnica de alineamiento estructural. Usando la transformación, las estructuras pueden ser superpuestas, y, de las estructuras superpuestas, el alineamiento estructural de sus secuencias puede ser derivado. La transformación (superposición) puede ser ejecutada de manera independiente de la construcción del alineamiento. Por lo tanto, es posible utilizar las transformaciones reportadas por programas individuales para calcular alineamientos estructurales estandarizados y algunas medidas que se desprenden de este alineamiento. Finalmente, demostramos la posibilidad de calcular un alineamiento estandarizado, que es óptimo, empleando un procedimiento basado en programación dinámica.

Para estos efectos se creó un programa computacional llamado STOVCA (*Structural Overlap Calculator*), que calcula el alineamiento estructural óptimo para una superposición de un par de estructuras previamente dada, considerando parámetros definidos. El criterio de optimización utilizado en este algoritmo se basa en la longitud del alineamiento, que

corresponde al número de pares de aminoácidos estructuralmente equivalentes entre dos estructuras superpuestas a un umbral de corte definido por el RMSD (*i.e.* umbral de distancias para definir pares de residuos equivalentes). Es muy importante mencionar que STOVCA no altera la superposición previa entregada por otro programa de superposición estructural. Como resultado el programa reporta el alineamiento estructural correspondiente junto con algunas medidas de la similitud estructural de un par de proteínas. El software utiliza una implementación similar al algoritmo de programación dinámica de Smith-Waterman con un modelo de penalidad de gap afín (Figura 8).



Estructura A vs Estructura B
Superpuesta con el software X

```
CALCULATED DATA:

81        S (absolute similarity)
78.6      SO (structural overlap)
67.2      s  (relative similarity)

1.5       RMSD (in Angstroms)
22.2      SeqIde (sequence identity)
```

Figura 8. Esquema del funcionamiento general del programa STOVCA (*Structural Overlap Calculator*). A la izquierda un par de proteínas superpuestas con un programa X. Las proteínas se encuentran representadas con su cadena principal. Esto sirve de entrada al programa STOVCA que calcula el alineamiento estructural óptimo sin alterar la superposición recibida. A la derecha, se identifican los pares estructuralmente equivalentes (que se muestran en rojo para la estructura azul y naranja para la estructura verde), los que son representados sobre la superposición recibida como entrada y en la forma de un alineamiento de secuencias derivado de las estructuras (alineamiento estructural). A derecha y abajo, se muestran algunas medidas de similitud estructural y de secuencia que son calculadas a partir del alineamiento estructural y reportadas por el programa.

En la descripción de los alineamientos estructurales se utilizará la siguiente nomenclatura, llamamos a la primera estructura *query* (q) y a la segunda *target* (t). Los alineamientos estructurales son caracterizados por varios parámetros, de los cuales el más importante es la longitud del alineamiento, el cual será llamado en adelante *absolute similarity S(q,t)*. Este valor depende de parámetros que el usuario puede modificar en tiempo de ejecución: o (*gap-opening penalty*), e (*gap-extension penalty*), m *(matched pair score)*, u (*unmatched pair score)*, t (*distance threshold)* y el conjunto {*A*} (lista de tipos atómicos). Una posición equivalente entre dos residuos es asignada cuando todos los átomos correspondientes (*i.e.* los que tienen los mismos nombres de átomos) del conjunto de nombres de átomos definidos en {*A*} se encuentran a una distancia igual o menor que t en Ångstroms. Además de *S(q,t),* STOVCA tiene la capacidad de calcular otras medidas de la similitud estructural tales como:

Similitud relativa *(relative similarity)*: La similitud relativa definida por *s(q,t)*, es una medida de la similitud a nivel global de dos estructuras de proteínas, y se calcula mediante:

$$s(q,t) = 100 \times \frac{2 \times S(q,t)}{L_q + L_t}$$

donde L_q and L_t corresponden a longitud en aminoácidos de la proteína *query* y *target* respectivamente.

Superposición estructural (*structural overlap)*: El solapamiento estructural definido por *SO(q,t)* da cuenta de la similitud a nivel local entre dos estructuras de proteínas, y se calcula por:

$$SO(q,t) = 100 \times \frac{S(q,t)}{\min(L_q + L_t)}$$

Además es importante mencionar que *SO(q,t)* es un valor aproximado de la cobertura sobre la estructura *query cq = 100 x S(q,t)/Lq*, y sobre la estructura *target ct = 100 x S(q,t)/Lt* cuando las dos proteínas son similares en longitud. Sin embargo, cuando las estructuras difieren mucho en su tamaño, SO(q,t) es redundante con *cq* o con *ct*.

Uno de los parámetros más importante es el *gap-opening* y corresponde al inverso aditivo del tamaño del más pequeño de los fragmentos permitidos en el alineamiento. Este parámetro puede ser modificado a gusto por el usuario en el tiempo de ejecución. Por defecto los parámetros de STOVCA corresponden a: o = -3, e = 0, t = 3.5 Å, m = 1, $\{A\}$ = 'Cα' y u = el entero negativo más grande disponible para la plataforma computacional donde se ejecuta el programa. Estos mismos parámetros fueron empleados en las ejecuciones de STOVCA para esta tesis.

La longitud del alineamiento y los valores de similitud estructural descritos arriba dependen del alineamiento estructural óptimo que es calculado, el cual en último término depende de los parámetros del algoritmo de programación dinámica.

4.3 Obtención de alineamientos estandarizados y selección del mejor alineamiento estructural disponible.

En el proceso de construcción de alineamientos estructurales para la construcción de una jerarquía estructural de los dominios catalíticos de polimerasas de ADN, quedó en evidencia que algunos programas no entregan soluciones completamente convincentes cuando se inspeccionan sus superposiciones de manera visual. En algunos casos es posible encontrar ejemplos donde definitivamente el programa empleado genera superposiciones absolutamente sin sentido desde la perspectiva estructural. Lo anterior puso una advertencia, que es especialmente relevante si se piensa en realizar ejercicios de comparación a nivel estructural en gran escala.

Por otra parte, cada programa tiene sus ventajas y se comporta mejor en algunas situaciones, por lo tanto la idea de poder aprovechar cada fuente de información disponible es importante.

Dado que esencialmente cada programa maneja su propia definición de pares equivalentes para construir un alineamiento de secuencias a partir de un alineamiento estructural, se hace imposible utilizar de manera directa los resultados entregados por cada programa de alineamiento, dado que éstos no son comparables entre sí.

Luego, la creación de un programa que pueda estandarizar dichos resultados es necesaria.

A continuación se formula el problema de manera genérica (Figura 9):

1. Sea *P(a,b)* un par de estructuras de proteínas a comparar.
2. Sea *S* un conjunto de superposiciones óptimas del par *P(a,b)* generadas con distintos programas.
3. Calcular los alineamientos estructurales óptimos de manera estandarizada para cada miembro S_i del conjunto *S* sin alterar su matriz original de rotación/traslación.
4. Calcular una métrica estándar de la similitud estructural para comparar varios alineamientos estructurales.
5. Ordenar la superposiciones en forma decreciente de acuerdo al valor de la métrica estándar.
6. Escoger la superposición con el mejor valor de similitud estructural.

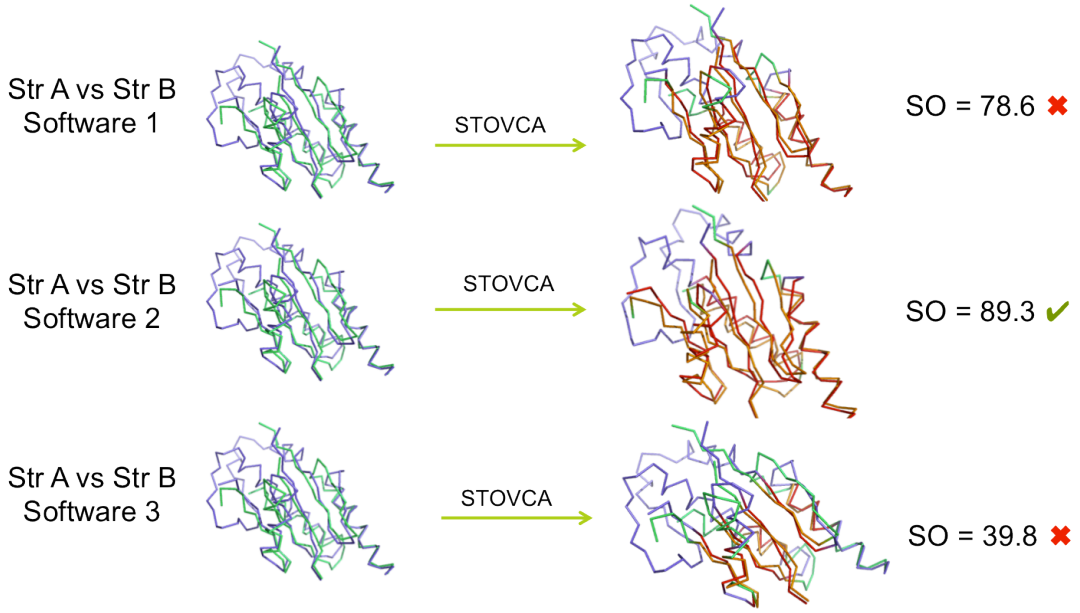


Figura 9. Ejemplo del funcionamiento del algoritmo de STOVCA. En la figura se muestra un ejemplo de cómo opera el algoritmo de selección del mejor alineamiento a partir de un conjunto de alineamientos estandarizados. A la izquierda aparecen tres proteínas óptimamente superpuestas con diferentes programas. Cada una de ellas analizada con STOVCA para obtener el correspondiente alineamiento estructural óptimo estandarizado y las medidas de similitud estructural que puede calcular el programa. Posteriormente se examinan los valores de similitud estructural y se selecciona el mayor obtenido. Esa superposición será conservada para posteriores análisis.

4.4 Alineamiento de múltiples estructuras (MStAs)

Uno de los aspectos importantes de esta tesis consiste en la creación de alineamientos de múltiples estructuras o MStAs (*Mutiple Structural Alignments*). Estos alineamientos consideran en primer lugar la similitud de las estructuras para encontrar aquellos residuos equivalentes y, por tanto, alineables a nivel de secuencia. Para clarificar la nomenclatura que se utilizará en adelante en esta tesis entenderemos por un MStA a un alineamiento de múltiples secuencias obtenido a partir de la superposición de múltiples estructuras de proteínas. Esto es para distinguirlo de un alineamiento de múltiples secuencias (MSA), que corresponde al alineamiento de varias secuencias mediante un algoritmo que utiliza únicamente la información de la secuencia aminoacídica para construir dicho alineamiento.

La construcción de los MSAs típicamente sigue un procedimiento llamado alineamiento de tipo progresivo. Este procedimiento consta en generar en primer lugar todos los alineamiento pareados posibles para un conjunto de *n* secuencias a analizar. Esto permite agrupar las secuencias a través de un árbol guía que sirve para determinar en qué orden se irán compilando las secuencias para crear el alineamiento de múltiples secuencias. Este procedimiento de alineamiento progresivo fue creado en respuesta a que la realización de una optimización mediante programación dinámica en forma simultánea para más de 3 secuencias, es impracticable desde el punto de vista del tiempo de cálculo requerido para ello. Para la construcción de los MStAs se emplea el mismo procedimiento, sólo que el árbol guía utiliza el grado de similitud estructural para indicar el orden en que las estructuras se irán incorporando progresivamente al alineamiento. Al mismo tiempo se

compilan las secuencias con las respectivas posiciones alineadas de acuerdo al criterio estructural.

4.5 Agrupamiento jerárquico estructural

El agrupamiento jerárquico requiere como entrada una tabla de diferencias que debe ser construida con observaciones de distancias pareadas entre los elementos a comparar. Para el caso de las comparaciones estructurales se emplearon dos métricas diferentes que cuantifican el grado de similitud estructural entre dos proteínas. Las métricas corresponden al SO (S*tructural Overlap*) y RSIM (R*elative Similarity*). Ambas tienen en común que poseen las propiedades matemáticas esperadas para una función que describe una distancia entre dos entidades, y que dependen de una cantidad denominada longitud del alineamiento. Ésta última corresponde al número de pares estructuralmente equivalentes luego de una superposición óptima de dos estructuras de proteínas. Ha sido demostrado en la literatura que la longitud del alineamiento también tiene las propiedades de una función de distancia, sin embargo no permite hacer comparaciones directas entre varios pares de superposiciones óptimas, pues cada proteína tiene longitudes variables, por lo que es necesario normar esta distancia para hacerla comparable. Para ello se definió el SO, el cual se obtiene al dividir la longitud del alineamiento estructural por la longitud mínima en aminoácidos del par de estructuras comparadas (Ecuación 4). Por otra parte el RSIM, se obtiene al dividir el doble de la longitud del alineamiento por la suma de la longitud en aminoácidos de las dos estructuras comparadas (Ecuación 5).

$$SO = \frac{Ceq}{\min(La,Lb)}, \quad \text{[Ec.4]}$$

$$RSIM = \frac{2 \times Ceq}{La+Lb}, \text{[Ec. 5]}$$

De esta manera, el SO cuantifica el grado de similitud local que puede existir entre dos proteínas, mientras que el RSIM cuantifica el grado de similitud global entre éstas.

4.6 Búsqueda de estructuras relacionadas mediante *TopSearch*

TopSearch es una herramienta que permite realizar búsquedas estructurales a gran escala en el PDB (*Protein Data Bank*). De manera simple, se parte con una estructura *query*, a partir de la cual se buscan todas aquellas estructuras que tengan algún grado de similitud con ella. Para esto se utiliza la propiedad de desigualdad triangular que se cumple para las métricas de similitud estructural. Por lo tanto en un primer paso *TopSearch*, infiere la cota inferior de similitud que tiene la estructura *query* con el resto de las estructuras presentes en el PDB. *TopSearch* se encuentra acoplado a *TopMatch*. Esto permite que una vez construida la lista de proteínas similares a la proteína *query*, se puede realizar el cálculo del alineamiento estructural correspondiente entregando el valor de similitud estructural real.

TopSearch tiene algunas características que hacen muy interesante su uso:

1. Trabaja con una versión del PDB actualizada semanalmente.
2. Muy eficiente en términos del tiempo de cálculo (los resultados son prácticamente instantáneos).
3. Está acoplado a otras herramientas de exploración de la similitud estructural como *TopMatch*.
4. Puede trabajar con unidades biológicas, cadenas o dominios de proteínas.

En esta tesis se realizaron búsquedas con *TopSearch* utilizando como *query*: a) las estructuras de cadenas completas de polimerasas de ADN seleccionada para la base de datos, b) dominios catalíticos y c) dominios estructurales que conforman el dominio catalítico.

En cada caso se recuperaron las listas no redundantes de los hits obtenidos y fueron analizadas manualmente en busca de resultados relevantes, los que fueron evaluados en detalle utilizando TopMatch.

4.7 Construcción de perfiles basados en alineamientos de múltiples estructuras

Los datos provenientes de estructuras de proteínas aún son limitados en la actualidad. De hecho, aún persiste la brecha entre el número de estructuras y secuencias disponibles. Para complementar y obtener diversidad en los alineamientos múltiples, se realizó una

estrategia de complementar los MStAs con secuencias provenientes de bases de datos curadas manualmente, como es el caso de *Swissprot* (Boeckmann, 2003). La curación manual garantiza que los datos que se están incorporando a este MStA sean efectivamente polimerasas de ADN. Además se consigue incrementar la diversidad presente en alineamiento.

El procedimiento consiste en tomar como base el MStA generado según lo descrito en los puntos anteriores, y mediante un programa de alineamiento de secuencias incorporar los nuevos miembros. Cabe destacar que al alinear nuevas secuencias contra un alineamiento pre-existente, las posiciones determinadas como homólogas posicionales se preservan, dicho de otra manera, se respeta el alineamiento original, y se busca que las nuevas secuencias se acomoden a dicho alineamiento. De esta manera, se preserva la información adquirida en el alineamiento estructural. En todos los casos, las nuevas secuencias que se incorporan al alineamiento tienen un 40% o más de identidad de secuencia y cobertura superior al 75%, con alguna de las proteínas ya existentes en el MStA, lo que permite asegurar que se puede encontrar un alineamiento razonable entre esta nueva secuencia y el MStA pre-existente (*e.g.* no se introducen errores al alineamiento original, aumentando así las relación señal-ruido en éste).

Para realizar la búsqueda de nuevas secuencias se utilizó el programa BLAST (Altschul et al., 1990) en su variante *blastp*. Cada una de las secuencias de las estructuras presentes en la base de datos de polimerasas de ADN fue utilizada como *query* para realizar buscar mediante BLAST nuevas secuencias en la base de datos SwissProt (Boeckmann,

2003). Dado que pueden existir resultados coincidentes al realizar búsquedas con secuencias provenientes de una misma familia, los resultados fueron ordenados y filtrados para eliminar elementos redundantes. Posteriormente se filtraron todos aquellos resultados cuya identidad de secuencia fuese menor al 40% y mayor a 99% con la secuencia *query*. Con esto se evita tener secuencias muy divergentes, donde un alineamiento de secuencia podría fallar al alinearlas; también se eliminaron las secuencias altamente similares pues sólo aportan redundancia al alineamiento y no información. Adicionalmente, se descartaron todos aquellos resultados donde la cobertura de la secuencia fuese menor al 75% (Figura 10).

Finalmente, para completar la construcción del perfil, el MStA complementado es utilizado como entrada para un programa que convierte un alineamiento en un perfil. Para estos efectos se empleó el programa HMMer con sus parámetros por defecto (Finn et al., 2011).

4.8 Alineamientos de múltiples secuencias (MSAs)

Se construyeron alineamientos de múltiples secuencias para diferentes clases de secuencias de proteínas de polimerasas de ADN. Para esto se empleó el programa ClustalW con sus parámetros por defecto (Thompson et al., 2002). Los alineamientos fueron guardados en formato FASTA para su posterior tratamiento.

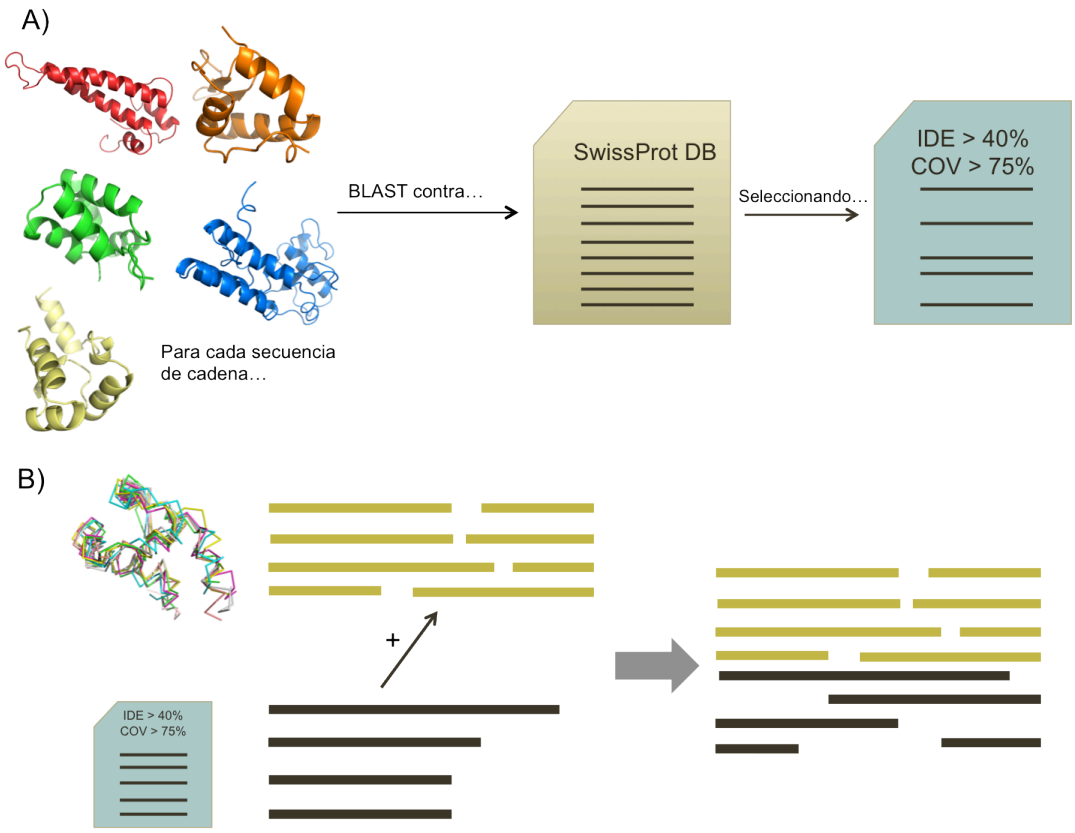


Figura 10. Complemento de MStAs con secuencias provenientes de bases de datos curadas manualmente. A) Procedimiento realizado para buscar secuencias para complementar el alineamiento. La secuencia de cada estructura se utilizó para buscar nuevas secuencias mediante BLAST contra la base de datos SwissProt y se descartaron secuencias según el filtro que se indica en el último cuadro. B) Procedimiento para complementar los MStAs. Arriba y a la izquierda, las barras verdes representan un MStA; a un costado izquierdo se encuentra la superposición de múltiples estructuras que origina este MStA. Abajo y a la izquierda, las barras negras representan el conjunto de secuencias recuperadas desde las bases de datos de secuencia empleando los criterios descritos en A. La flecha inclinada indica que dichas secuencias serán incorporadas al MStA. A la derecha se muestra el resultado de complementar el MStA. Las nuevas secuencias fueron alineadas respetando el esquema impuesto por el MStA. De esta manera se genera un MStA complementado con nuevas secuencias.

5. Resultados

5.1. Conjunto de datos para la validación de STOVCA (*Structural Overlap Calculator)*.

Antes de utilizar este programa de forma sistemática en la comparación de dominios catalíticos de polimerasas de ADN, se efectuó una validación del programa. Para estos efectos se construyó un conjunto de datos destinado exclusivamente a este propósito. Los datos fueron obtenidos desde la base de datos HOMSTRAD.

La base de datos obtenida consta de 215 pares de proteínas homólogas, cuya similitud estructural se encuentra en un rango de entre 30% y 75%. Sobre el 75% de similitud estructural, es posible considerar que la solución del alineamiento estructural es trivial, pues en efecto todos los programas de alineamiento estructural presentan un desempeño similar. Este conjunto de datos representa a un total de 42 familias de proteínas con diferentes estructuras, representando una amplia variedad de arquitecturas, topologías y pliegues.

5.2. Ningún programa de alineamiento estructural entrega de manera consistente el mejor alineamiento estructural.

La validación contempló la utilización inicial de 7 programas de alineamiento estructural que representan algunos de los algoritmos más frecuentemente utilizados. Cada uno de los 215 pares de proteínas fue alineado utilizando DALI, CE, MUSTANG, SALIGN, TMALIGN, MAMMOTH y TOPMATCH. Las superposiciones resultantes fueron analizadas con STOVCA. En la Tabla III se presentan algunas estadísticas derivadas de estos datos.

Tabla III. Estadísticas de alineamientos estructurales

Método o software	Número de Casos		Diferencia en longitud de alineamiento		Promedio SO (*Structural Overlap*)
	Mejor	Peor	Acumulativo	Promedio	
CE	15/32	35	1,554	7	58.0
DALI	15/46	14	820	4	60.8
MAMMOTH	8/16	61	1,738	8	57.1
MUSTANG	10/18	77	3,013	14	52.2
TMALIGN	13/41	14	733	3	61.2
TOPMATCH	47/84	5	529	2	62.0
SALIGN	35/72	8	603	3	61.7
HOMSTRAD	7/19	37	1,821	8	57.0
Mejor Valor	215	0	0	0	64.1

Las columnas con los números de mejores y peores casos registran el número de veces que un programa en particular genera la mejor (los números a continuación de la barra inclinada representan el número de veces que el mejor resultado es compartido con otras herramientas) y peores soluciones. La diferencia en longitud del alineamiento acumulativa y promedio representa la suma y promedio sobre los 215 superposiciones utilizadas en la evaluación. La última columna contiene el promedio de SO para las 215 superposiciones utilizadas. La fila Mejor Valor, representa los valores máximos (SO) o mínimos (diferencia en longitud del alineamiento) obtenidos con cualquiera de los programas utilizados y HOMSTRAD para cada caso particular en la evaluación.

De los resultados obtenidos es importante destacar que todos los programas empleados obtuvieron al menos en una ocasión el mejor o peor resultado. Esto es relevante, ya que demuestra que el uso de una única herramienta no garantiza que se obtenga siempre el mejor resultado. De forma similar, una herramienta dada con un desempeño global bueno, también puede obtener el peor resultado en alguna ocasión. Por lo tanto, se infiere que el uso de varios programas de alineamiento estructural junto con un método de estandarización para calcular el alineamiento estructural óptimo de proteínas, como el descrito con la herramienta STOVCA, puede ser útil para seleccionar la solución más apropiada de entre un conjunto de alineamientos estructurales producidos por varios programas.

Los datos obtenidos en esta evaluación revelan una tendencia de buen desempeño global para los programas DALI, TMALIGN, TOPMATCH y SALIGN. Estas cuatro herramientas producen un bajo número de peores soluciones, a la vez que logran un alto número de mejores soluciones. También es posible notar que en general se distancian poco de la mejor solución cuando producen la peor. Por último tienen los promedios más altos entre los programas evaluados. Siguiendo en lo referente al desempeño, encontramos un segundo grupo de software compuesto por CE, MAMMOTH y HOMSTRAD. Al final, en la línea base se encuentra MUSTANG, que generalmente produce los alineamientos con el promedio de *Structural Overlap* más bajo, las mayores diferencias con la mejor solución y el mayor número de peores soluciones. Esta tendencia se observa de manera clara cuando se grafican distribuciones acumulativas de los valores de *Structural Overlap* (Figura 11).

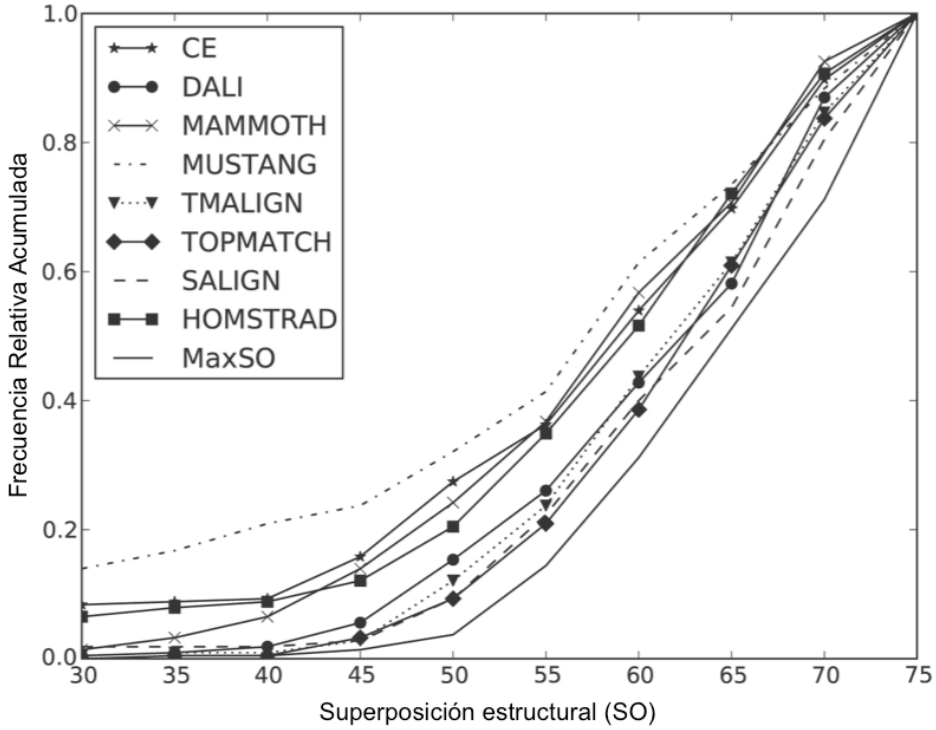


Figura 11. **Comparación del desempeño de programas de alineamiento estructural**. Distribución acumulativa de los valores de SO obtenidos con siete programas de alineamiento estructural y la base de datos HOMSTRAD para los 215 pares de estructuras de proteínas en la evaluación. Adicionalmente se muestra la distribución acumulativa para los valores máximos de SO (MaxSO) obtenida de los programas evaluados. Para esta distribución se tomaron todos los valores entre 30 y 75 de SO según lo reportado por STOVCA.

Posteriormente se determinó si las diferencias que se observan en este gráfico son estadísticamente significativas. Para estos efectos de utilizó la prueba no paramétrica de Kolmogorov-Smirnov . De esta manera se corroboraron las diferencias expresadas en el gráfico de distribuciones acumulativas. En efecto, el desempeño observado para TOPMATCH, DALI, TMALIGN y SALIGN con los otros programas utilizados en la evaluación es estadísticamente significativa a un nivel de confianza del 95%. Un análisis posterior de correlación, reveló que las correlaciones entre varios programas puede ser baja, de hecho HOMSTRAD es quien obtiene las correlaciones más bajas, siendo estas menores a 0.61 (Tabla IV).

Tabla IV. Análisis estadístico de la diferencia en el desempeño.

	CE	DALI	MAMMOTH	MUSTANG	TMALIGN	TOPMATCH	SALIGN	HOMSTRAD	MaxSO
CE		0.75	0.67	0.62	0.78	0.76	0.73	0.53	0.8
DALI	0.021		0.77	0.73	0.89	0.89	0.82	0.59	0.92
MAMMOTH	**0.653**	0.015		0.62	0.81	0.76	0.74	0.44	0.79
MUSTANG	0.027	$<5x10^{-4}$	0.015		0.68	0.66	0.63	0.61	0.7
TMALIGN	0.001	**0.935**	0.003	$<5x10^{-4}$		0.88	0.89	0.58	0.93
TOPMATCH	$<5x10^{-4}$	**0.575**	$<5x10^{-4}$	$<5x10^{-4}$	**0.575**		0.85	0.56	0.93
SALIGN	$<5x10^{-4}$	**0.422**	$<5x10^{-4}$	$<5x10^{-4}$	**0.422**	**0.738**		0.55	0.88
HOMSTRAD	**0.356**	0.027	**0.815**	0.048	0.048	0.003	0.001		0.59
MaxSO	$<5x10^{-4}$	0.001	$<5x10^{-4}$	$<5x10^{-4}$	0.006	0.008	**0.102**	$<5x10^{-4}$	

Triángulo superior derecho: coeficientes de correlación de Pearson para los valores de SO de todos los pares superpuestos en la evaluación (n = 215). Triángulo inferior izquierdo: *P*-valores de la prueba estadística de Kolmogorov-Smirnov (KS) para los valores de SO de las 215 estructuras de proteínas superpuestas en la evaluación. La fuente en negrita indica aquellas diferencias que no son estadísticamente significativas ($\alpha = 0.05$)

Para ilustrar las potenciales diferencias obtenidas cuando se utilizan dos herramientas de alineamiento estructural diferentes, fueron seleccionados algunos casos ejemplo de aquellos pares de proteínas que se utilizaron en la evaluación del programa STOVCA. Los ejemplos incluyen la herramienta con mejor y peor desempeño para el caso seleccionado, ilustrándose de manera gráfica que se generan alineamientos completamente distintos, en los cuales se puede identificar claramente uno que es el mejor. Todos los alineamientos efectuados con su análisis detallado a nivel de superposición, alineamiento de secuencia y medidas de similitud se encuentran disponibles en el sitio web http://melolab.org/stovca.

En algunos casos es posible encontrar diferencias extremas entre dos programas de alineamiento estructural. Por ejemplo el caso de las proteínas de códigos PDB 1d2z (Xiao, Towb, Wasserman, & Sprang, 1999) y 1e41 (Berglund et al., 2000), que corresponden a dominios DEATH. Cuando este par de proteínas se alinean empleando el software MAMMOTH se obtiene una transformación que identifica mínimas similitudes estructurales, comparado con el caso cuando se utiliza SALIGN que identifica una similitud del 55%. Lo mismo ocurre con otro par de proteínas y programas distintos. Para el caso de las proteínas de códigos PDB 1efu (Kawashima, Berthet-Colominas, Wulff, Cusack, & Leberman, 1996) y 1tfe (Jiang, Nock, Nesper, Sprinzl, & Sigler, 1996), que corresponden a la familia de factores de elongación Tu, el programa CE entrega una transformación totalmente errada. Por otra parte MUSTANG, identifica en este caso particular la máxima similitud con un 58% de solapamiento estructural. (Figura 12)

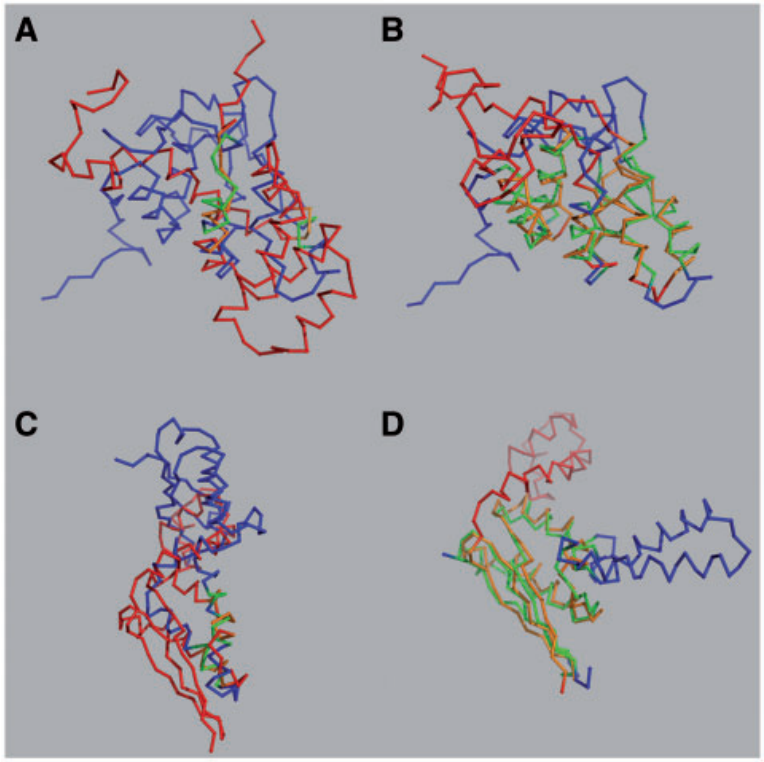


Figura 12. Ejemplos de superposiciones de proteínas con desempeño diferente. En la figura se muestran ejemplos de superposiciones que presentan diferencias extremas cuando son obtenidas con programas diferentes. Se encuentra representado el esqueleto de la proteína. La proteína *query* se encuentra en color azul, mientras que la *target* en color rojo. Los pares de aminoácidos alineados se muestren en color verde para la estructura azul y en naranjo para la estructura roja. La superposición de las proteínas 1d2zb y 1e41a con MAMMOTH (A) y SALIGN (B) entrega valores de SO de 9.45% y 55.91% respectivamente. La superposición de 1efud1 y 1tfe con CE (C) y MUSTANG (D) da valores de 10.79% y 58.99% respectivamente.

Un ejemplo donde ocurren diferencias más finas, pero relevantes, se puede encontrar en las polimerasas de ADN, específicamente de la comparación de la parte de los dominios catalíticos correspondiente a la palma de dos polimerasas de ADN: la polimerasa I de *Bacillus strearothermophilus* (Código PDB 1xwl, (Kiefer et al., 1997)) y la palma de la polimerasa de ADN Dpo4 de *Sulfolobus solfataricus* (Código PDB 2imw, (Fiala et al., 2007)). El alineamiento estructural óptimo resultante con STOVCA a partir de las superposiciones con TOPMATCH y MUSTANG muestran valores de SO de 74% y 66% respectivamente. El porcentaje de identidad de secuencia de estos alineamientos estructurales es 14% para TOPMATCH y 13% para MUSTANG. Cuando se analizan más en detalle los alineamientos estructurales entregados por STOVCA, se observa que en el caso de TOPMATCH, los ácidos aspárticos involucrados en la catálisis (Asp-829 y Asp-652 de la ADN Pol I; Asp-104 y Asp-7 de la ADN Pol Dpo4) están correctamente alineados. Esto no ocurre cuando se revisa el alineamiento estructural proporcionado por MUSTANG, donde Asp-829 de la ADN Pol I no es equivalente con el Asp-104 de la ADN Pol Dpo4. Este es un ejemplo muy específico de alineamientos incorrectos que pueden tener un impacto profundo en análisis posteriores (Figura 13).

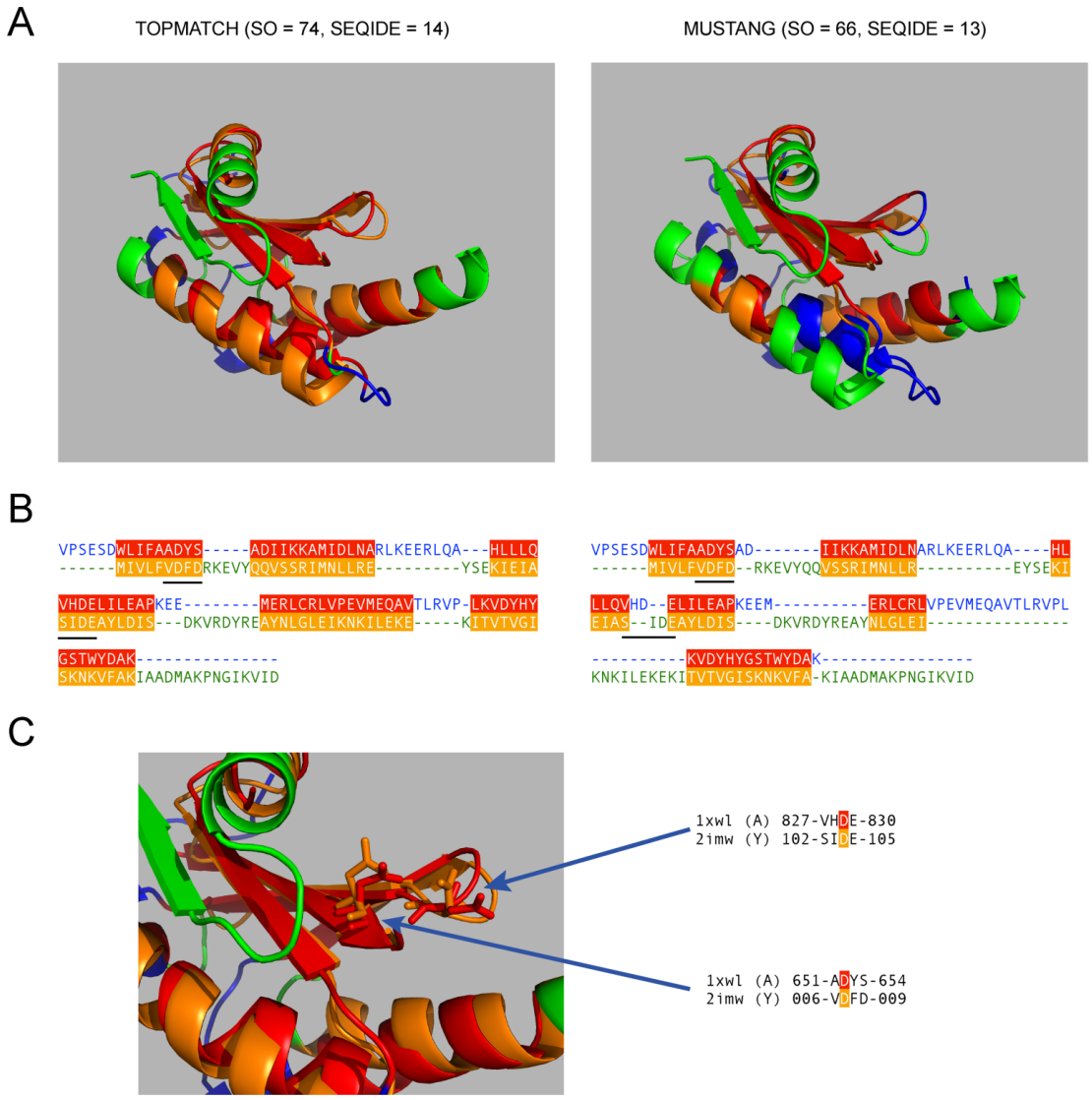


Figura 13. Ejemplo de superposiciones de dominios palma de polimerasas de ADN. (A) Superposiciones de las estructuras de ADN polimerasa I de *B. Stearothermophilus* (Código PDB 1xwl) y polimerasa de ADN Dpo4 de *S. Solfataricus* (Código PDB 2imw) generadas con TOPMATCH (izquierda) y MUSTANG (derecha). (B) Alineamientos estructurales óptimos generados con STOVCA para la superposición reportada con TOPMATCH (izquierda) y MUSTANG (derecha). Las dos regiones que contienen al aparato catalítico se encuentran subrayadas en el alineamiento. (C) Vista en detalle de los residuos de ácido aspártico en la superposición generada con TOPMATCH y los alineamientos estructurales óptimos generados con STOVCA.

Finalmente, se determinó que para posteriores análisis estructurales sobre los dominios catalíticos de polimerasas de ADN, se utilizará una estrategia basada en múltiples fuentes de información de alineamientos estructurales, que posteriormente son evaluados de manera estándar con la herramienta STOVCA, para seleccionar el alineamiento que maximice la similitud estructural.

5.3. Base de datos de estructuras de polimerasas de ADN.

De acuerdo al procedimiento descrito en la sección de métodos, la base de datos no redundante quedó conformada por un total de 28 estructuras de polimerasas de ADN. Dentro de los criterios empleados para filtrar las estructuras se privilegió en primer lugar que la estructura fuese lo más completa posible y, en segundo lugar, que tuviese la mayor resolución. Dado que los estudios realizados en esta tesis corresponden a comparaciones de tipo estructural que se basan en la información disponible para los carbonos alfa de las proteínas, no se requiere ser tan exigente en términos de la resolución con la que fue obtenida la estructura. Por último, cabe destacar que sólo se incluyeron en este conjunto de datos aquellas proteínas clasificadas mediante el número EC 2.7.7.7, que corresponden a polimerasas de ADN dependientes de ADN. No todas las familias cuentan con igual representación; la más numerosa es la familia B con ocho estructuras, mientras que la menos numerosa es la familia C, que posee un total de tres estructuras. Todos los dominios de la vida y virus se encuentran representados en esta base de datos. Tal como se mencionó en la revisión bibliográfica, en el año 1998 se describió una nueva polimerasa de ADN, que en términos de secuencia no comparte ninguna similitud con las descritas previamente. Por

ello se le asignó a una familia nueva, llamada familia D. Durante el desarrollo de esta tesis se depositó en el PDB la primera estructura de una polimerasa de esta familia (código PDB 3o59). Sin embargo, se trata de un fragmento de 300 aminoácidos en el segmento N-terminal de la proteína, que según lo descrito en la literatura no tiene relación con el dominio catalítico de esta polimerasa, por lo tanto se excluyó del análisis. Los detalles de la base de datos no redundante de estructuras de polimerasas de ADN, se resumen en la Tabla V.

Tabla V. Base de datos no redundante de estructuras de polimerasas de ADN.

Código PDB	Nombre Común	Familia	Organismo	Dominio	Referencia
1taq	Taq Polymerase	A	*Thermus aquaticus*	Bacteria	(Y. Kim et al., 1995)
1lv5	Bacillus Fragment DNA polymerase	A	*Geobacillus stearothermophilus*	Bacteria	(S. J. Johnson, Taylor, & Beese, 2003)
2ajq	T7 Phage DNA Polymerase	A	Phage T7	Virus	(Ellenberger, Doublié, Tabor, Long, & Richardson, 1998)
2kfn	Klenow fragment DNA polymerase II	A	*Escherichia coli*	Bacteria	(Brautigam, Sun, Piccirilli, & Steitz, 1999)
3ikm	Polymerase Gamma	A	*Homo sapiens*	Eukarya	(Y.-S. Lee, Kennedy, & Yin, 2009)
2p5o	Phage RB69 DNA polymerase	B	Phage RB69	Virus	(Hogg, Wallace, & Doublié, 2004)
3maq	DNA Polymerase II	B	*Escherichia coli*	Bacteria	(F. Wang & Yang, 2009)
1tgo	Polymerase TGO	B	*Thermococcus gorgonarius*	Archaea	(Hopfner et al., 1999)
2gv9	HSV-1 DNA polymerase	B	Herpes simplex virus	Virus	(Liu et al., 2006)
2jgu	Pfu DNA polymerase	B	*Pyrococcus furiosus*	Archaea	(S. W. Kim, Kim, Kim, Kang, & Cho, 2008)
1s5j	DNA polymerase B1	B	*Sulfolobus solfataricus*	Archaea	(Savino et al., 2004)
1xhx	Phage Phi29 DNA polymerase	B	Phage Phi29	Virus	(Kamtekar et al., 2004)
3iay	DNA polymerase delta	B	*Homo sapiens*	Eukarya	(Swan, Johnson, Prakash, Prakash, & Aggarwal, 2009a)
2hnh	DNA polymerase III	C	*Escherichia coli*	Bacteria	(Lamers, Georgescu, Lee, O'Donnell, & Kuriyan, 2006)
3f2b	DNA polymerase III	C	*Geobacillus kaustophilus*	Bacteria	(Evans et al., 2008)
2hpi	DNA polymerase III	C	*Thermus aquaticus*	Bacteria	(Bailey, Wing, & Steitz, 2006)
1huo	DNA polymerase beta	X	*Rattus norvegicus*	Eukarya	(Arndt et al., 2001)
1xsl	DNA polymerasa lambda	X	*Homo sapiens*	Eukarya	(Garcia-Diaz, Bebenek, Krahn, Kunkel, & Pedersen, 2005)
2ihm	DNA polymerase mu	X	*Mus musculus*	Eukarya	(Moon et al., 2006)
1jaj	ASFV polymerase X	X	African swine virus	Virus	(Maciejewski et al., 2001)
2w9m	DNA polymerase X	X	*Deinococcus radiodurans*	Bacteria	(Leulliot et al., 2009)
2asd	Dpo4 DNA polymerase	Y	*Sulfolobus solfataricus*	Archaea	(Rechkoblit et al., 2006)
2dpi	DNA polymerase iota	Y	*Homo sapiens*	Eukarya	(Nair, Johnson, Prakash, Prakash, & Aggarwal, 2006)
2wtf	DNA polymerase eta	Y	*Saccharomyces cerevisiae*	Eukarya	(Enoiu, Jiricny, & Schärer, 2012)
2oh2	DNA polymerase kappa	Y	*Homo sapiens*	Eukarya	(Uljon et al., 2004)
1k1s	DinB DNA polymerase	Y	*Sulfolobus solfataricus*	Archaea	(Silvian, Toth, Pham, Goodman, & Ellenberger, 2001)
2aq4	yRev1 DNA polymerase	Y	*Saccharomyces cerevisiae*	Eukarya	(Nair, Johnson, Prakash, Prakash, & Aggarwal, 2005)
3gqc	hRev1 DNA polymerase	Y	*Homo sapiens*	Eukarya	(Swan, Johnson, Prakash, Prakash, & Aggarwal, 2009b)

5.4. Alineamientos estructurales de cadenas completas de polimerasas de ADN.

Cuando se compara a nivel de cadena, podemos ver que la clasificación de familias actualmente propuesta para polimerasas de ADN se mantiene. Todos los miembros de una misma familia descienden de un nodo común, en ausencia de un umbral de corte definido. Sin embargo, es posible notar que existe variación estructural importante entre familias, así como dentro de familias de polimerasas de ADN. Algunos casos notables ocurren en las familias A y B. Cuando se establecen grados discretos de similitud estructural en la forma de criterios de corte en el agrupamiento jerárquico por cadenas, se observa de mejor manera esta diversidad. Por ejemplo, la Familia A posee cuatro miembros que se pueden considerar altamente similares, ya que comparten sobre un 70% de similitud estructural. Un quinto miembro (código PDB 3ikm) se distancia de gran manera del resto de los miembros de esta familia. Esta proteína corresponde a la polimerasa Gamma de *H.sapiens*. En el caso de la familia B también es posible notar un grado interesante de variación estructural. Primero, se observa un grupo formado por 6 estructuras, y luego una séptima estructura que se separa del resto de las polimerasas de la familia B al umbral *Relacionado*[5] (código PDB 1xhx). Esta última corresponde a una polimerasa del fago Phi29. En el caso de la familia C los tres miembros agrupan bajo el umbral *Relacionado*; los elementos más cercanos comparten un máximo de 68% de similitud estructural. En la familia X se distinguen tres miembros que comparten una alta similitud. Por otra parte, un cuarto miembro (código PDB 1jaj) agrupa bajo el umbral *Relacionado*. Finalmente, de acuerdo al dendrograma, la familia con menor grado de variación estructural es la Familia Y, pues todos sus miembros se encuentran

[5] Los umbrales de corte se designan según la nomenclatura de Sippl et al., 2008 *Distant* (distante) corresponde a proteínas que comparten 30% o más de similitud estructural, *Remote* (remoto) para aquellas que comparten 40% o más de similitud estructural, y finalmente *Related* (relacionado) para aquellas que comparten un 60% o más de la similitud de sus estructuras.

agrupados bajo el umbral *Relacionado*. (Figura 14).

El agrupamiento jerárquico construido, selecciona las mejores superposiciones pareadas para este conjunto de estructuras. luego el dendrograma que se muestra acá ha optimizado la búsqueda de similitudes estructurales, por lo que cuando se observan diferencias grandes, podemos esperar que sean efectivas y no se trate necesariamente de artefactos en la construcción de las superposiciones.

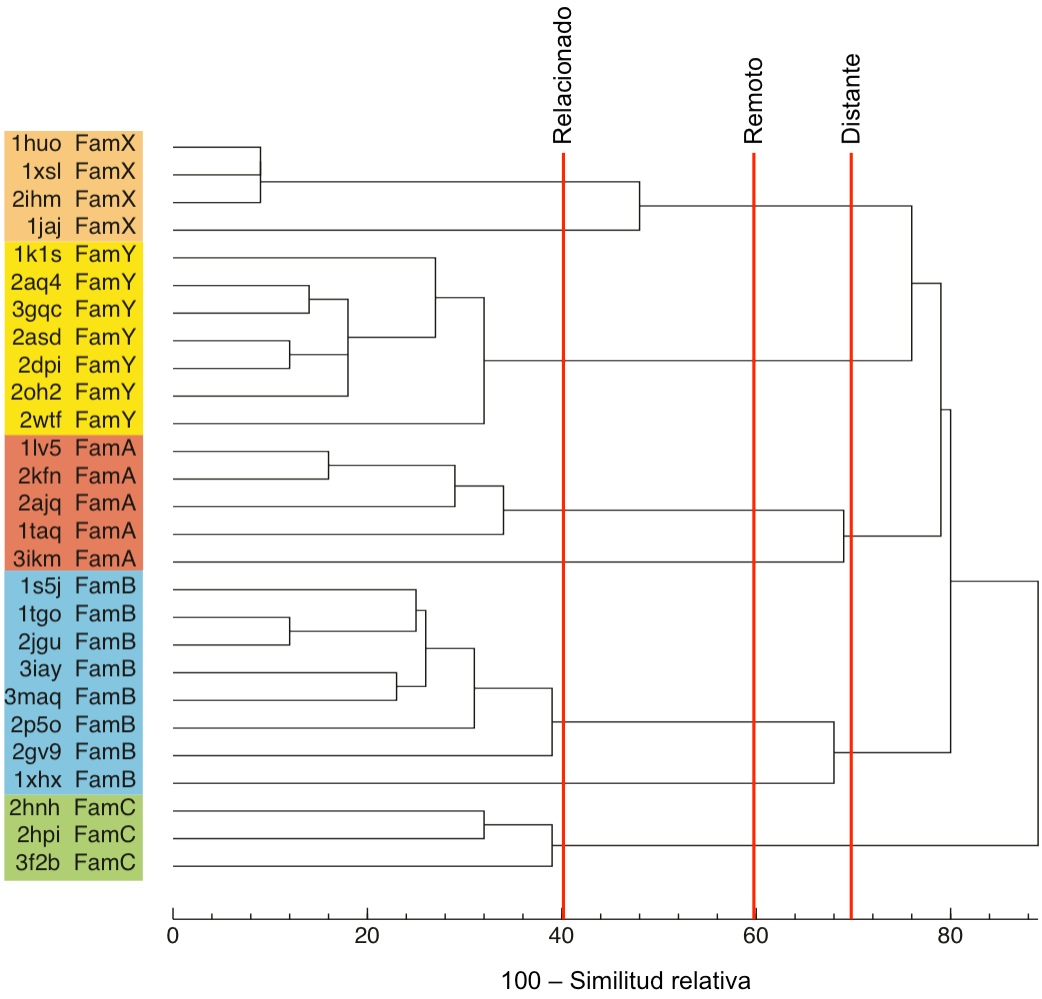


Figura 14. Comparación a nivel estructural de cadenas completas de polimerasas de ADN. En la figura se muestra un dendrograma que da cuenta de las relaciones jerárquicas a nivel de cadenas de polimerasas de ADN. Los alineamientos estructurales y el cálculo de la similitud estructural fueron generados de acuerdo a la metodología descrita en el punto 4.4. En la base del árbol se encuentran las etiquetas de los códigos PDB de cada estructura y la familia a la que pertenece. A cada familia le fue asignado un código de color (Familia A: rojo, Familia B: celeste, Familia C: verde, Familia X: naranjo, Familia Y: amarillo). La medida de similitud estructural empleada es la *Similitud Relativa*, que en la figura se encuentra expresada en la forma de una distancia. Tres líneas rojas perpendiculares al eje del dendrograma fueron dibujadas para establecer criterios de corte de similitud en este agrupamiento jerárquico. Los criterios de corte fueron extraídos de la base de datos COPS **(Suhrer, Wiederstein, Gruber, & Sippl, 2009)**

5.5. Alineamientos estructurales a nivel del dominio catalítico.

Como se mencionó en la introducción de esta tesis, el dominio catalítico es un dominio funcional formado por tres dominios estructurales (palma, pulgar y dedos). Cuando se analizan las relaciones estructurales de los dominios catalíticos de polimerasas de ADN, es posible observar que la clasificación actual se mantiene para algunos de los criterios de corte posible según la similitud estructural. Si se toma como umbral de corte *Distante*, encontramos dos grandes grupos de dominios catalíticos, el primero formado por las familias A, B, X e Y, el segundo formado sólo por la familia C. Cuando el umbral utilizado es *Remoto*, aparecen las cinco familias actualmente definidas. Finalmente si se escoge el umbral *Relacionado*, se distinguen cinco grupos y dos *singletons*. Uno de ellos corresponde a la polimerasa Gamma (código PDB 3ikm, Familia A) y el otro a la polimerasa del fago Phi29 (código PDB 1xhx, Familia B). Con respecto a la variabilidad por familia, se puede constatar que las familias A y B son las que tienen el mayor grado de diversidad a nivel estructural. En el caso de la familia A es importante destacar que esta diversidad es ocasionada por uno de los miembros (código 3ikm), el que se distancia de manera considerable de los otros 5 que pertenecen a la familia ésta familia; este último grupo tiene un grado de conservación muy alto (similitudes relativas superiores al 80%). Por otra parte, la familia B es la que en promedio tiene mayor diversidad con un promedio de 40% de similitud estructural. En el caso de la familia C sus miembros comparten una similitud estructural de 75% en promedio, a nivel del dominio catalítico. Por su parte, la familia X, presenta tres grupos estructurales marcados. El primero se encuentra constituido por las polimerasas beta, lambda y mu (códigos PDB 1huo, 1xsl y 2ihm respectivamente), el

segundo está representado por la polimerasa ASFV (código PDB 1jaj), y el tercero por la polimerasa X (código PDB 2w9m). Finalmente, la familia Y presenta un alto grado de conservación estructural. Las similitudes relativas observadas son superiores al 75% (Figura 15).

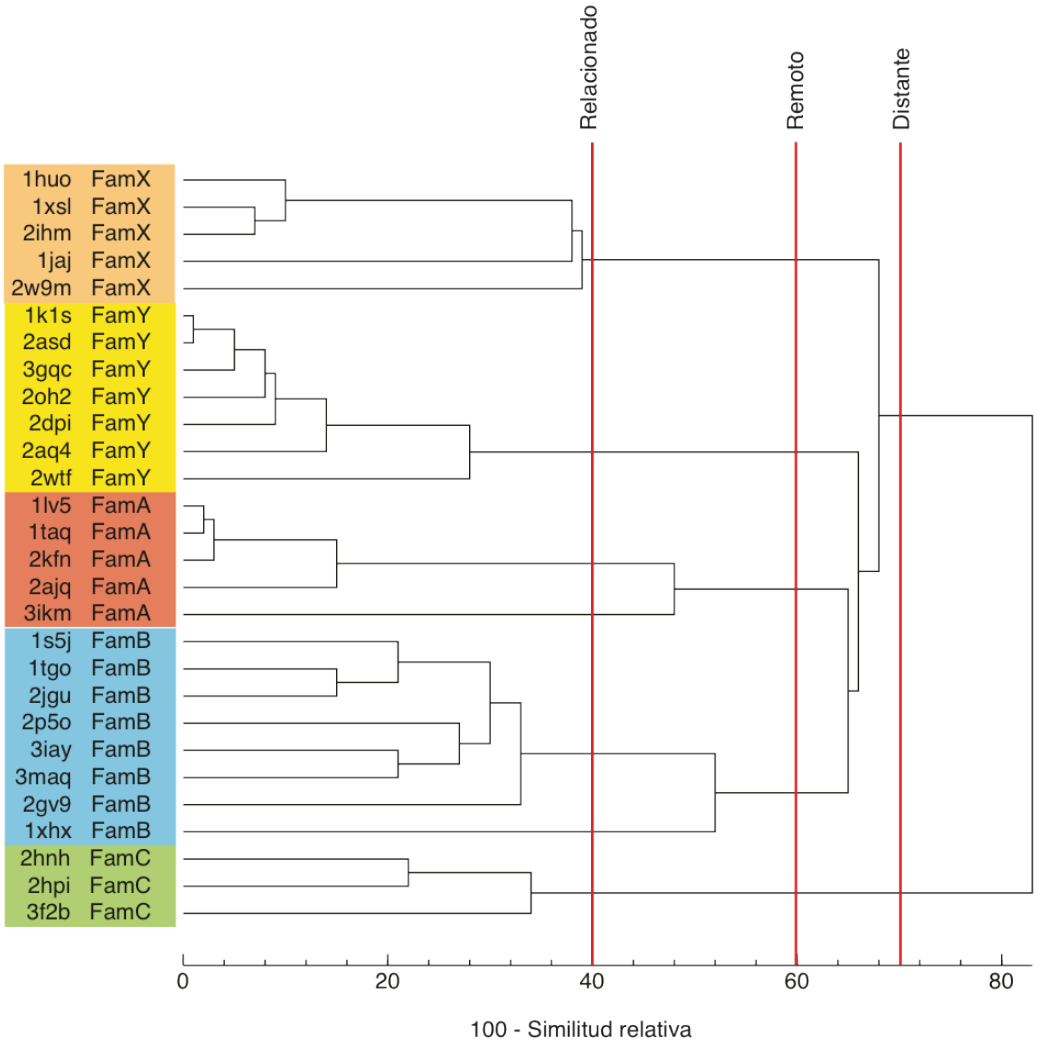


Figura 15. Relaciones de similitud estructural en el dominio catalítico de polimerasas de ADN. El dendrograma muestra las relaciones de similitud estructural del dominio catalítico de polimerasas de ADN. Los alineamientos estructurales y el cálculo de la similitud estructural fueron hechos de acuerdo a la metodología descrita en el punto 4.4. En la base del árbol se encuentran las etiquetas de los códigos PDB de cada estructura y la familia a la que pertenece. A cada familia le fue asignado un código de color (Familia A: rojo, Familia B: celeste, Familia C: verde, Familia X: naranjo, Familia Y: amarillo). La medida de similitud estructural empleada es la *Relative Similarity*, que en en la figura se encuentra expresada en la forma de una distancia. Tres líneas rojas perpendiculares al eje del dendrograma fueron dibujadas para establecer criterios de corte de similitud en este agrupamiento jerárquico. Los criterios de corte fueron extraídos de la base de datos COPS **(Suhrer et al., 2009)**.

5.6. Alineamientos estructurales a nivel de dominios palma, dedos y pulgares.

Tal como se comentó en la sección 1.4, el dominio catalítico de polimerasas de ADN ha sido tradicionalmente dividido en tres dominios estructurales de acuerdo a la analogía propuesta por Steitz en el año 1994 (Joyce & Steitz, 1994). En las comparaciones realizadas a nivel de dominios se empleó la medida de *Similitud Relativa* (RSIM) para cuantificar la similitud estructural entre éstos. Se escogió esta medida, puesto que permite cuantificar el grado de similitud global entre dos dominios de polimerasas de ADN, pues al ser segmentos con un menor número de aminoácidos, interesa conocer que tan similares son en su totalidad más que por segmentos locales.

5.6.1. Similitud estructural en el dominio de la palma.

En primer lugar se describirán las relaciones de similitud estructural observadas en el dominio palma. Estas relaciones se exploraron en tres niveles de acuerdo a lo descrito en la literatura (Suhrer et al., 2009). De manera general, es posible observar que la clasificación en familias propuesta actualmente para esta clase de enzimas se conserva, pues es posible encontrar un nodo en este agrupamiento jerárquico, a partir del cual se agrupan todos los miembros de una familia. A nivel de agrupamiento de familias, vemos que al umbral de corte *Distante* se observan dos grandes grupos de palmas; uno formado las familias A, B, X e Y, y un segundo grupo conformado por todas las palmas de las polimerasas de la familia C. Cuando se hace más estricto el criterio de corte y se escoge el umbral *Remoto*, se observan tres grandes grupos: el primero formado por las palmas de las familias A, B e Y,

el segundo constituido por las palmas de la familia X y el tercer grupo formado por las palmas de la familia C. Finalmente con el umbral de corte de similitud estructural más estricto (*Relacionado)*, todas las familias quedan como grupos independientes, pero además, algunos miembros de familias forman *singletons*. Esto ocurre en las familias A y B, donde los mismos miembros que se distanciaron a nivel de cadena completa, lo hacen a nivel del palmas (códigos PDB 3ikm y 1xhx respectivamente). En el caso de la familia A, el elemento que deja de pertenecer al umbral estricto *Relacionado*, corresponde a la palma de la polimerasa gamma de *Homo sapiens* (código PDB 3ikm), mientras que en el caso de la familia B corresponde a la polimerasa del fago Phi29 (código PDB 1xhx). Las familias que tienen menor variabilidad corresponden a las familias C, X e Y que en promedio poseen un 75% de similitud relativa. En la situación opuesta quedan las familias A y B que presentan el mayor grado de variabilidad estructural en este segmento específico (con porcentajes promedio de 68% y 66% respectivamente).

Finalmente, de acuerdo a los criterios de corte establecidos en este dendrograma es posible establecer que las palmas de tres familias comparten una importante similitud estructural. Estas corresponden a las palmas de las familias A, B e Y (Figura 16). En efecto, cuando se estudia en mayor detalle este tipo de relaciones se puede resumir que ellas presentan una arquitectura y topología similar constituida por cuatro hebras betas y dos hélices alfa, lo que es consistente con las relaciones de similitud encontradas entre los dominios palma de estas familias. Por otra parte las polimerasas de la familia C y X se caracterizan por presentar una arquitectura de la palma conformada por cinco hebras beta y dos hélices alfa.

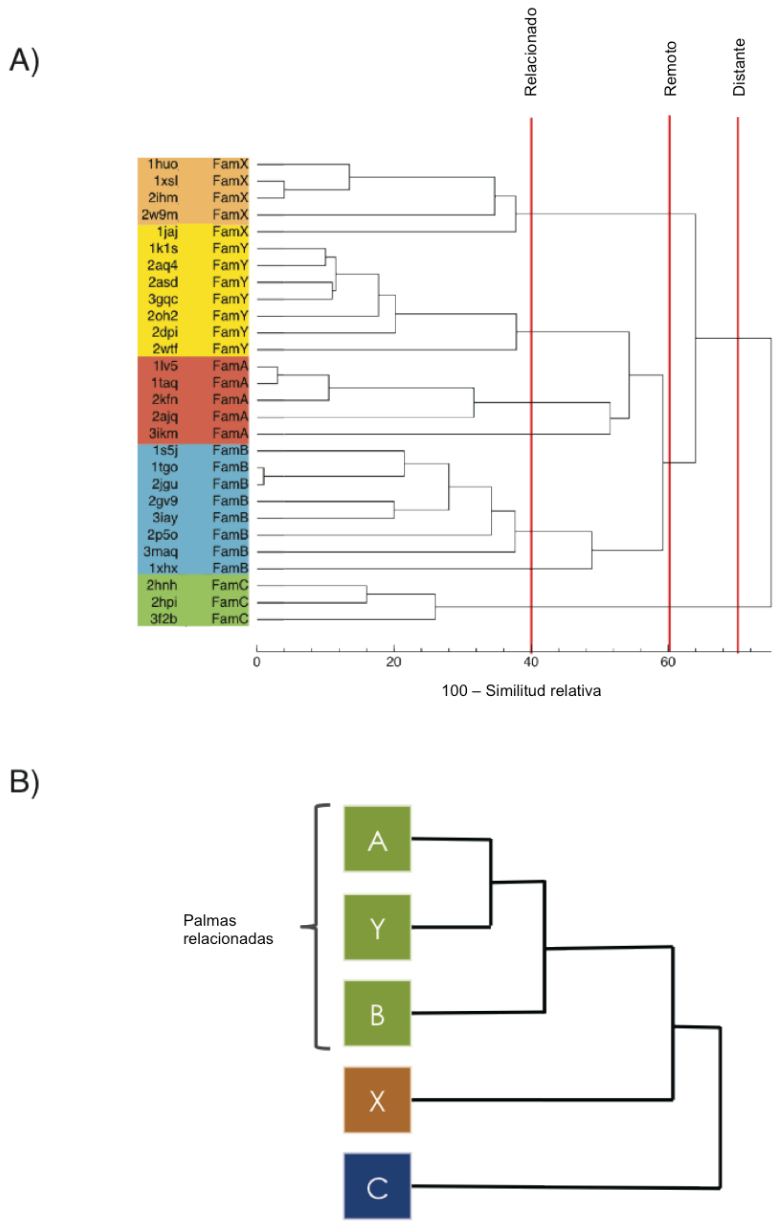


Figura 16 Agrupamiento jerárquico de dominios palma de polimerasas de ADN. A) Dendrograma de la similitud estructural de dominios palma de polimerasas de ADN. Las similitudes fueron expresadas mediante *Similitud Relativa* (RSIM) y convertidas posteriormente a distancia. En las hojas del árbol, cada palma fue etiquetada con el código PDB de donde proviene y las familias a las que pertenecen con códigos de colores (Familia A: rojo, Familia B: celeste, Familia C: verde, Familia X: naranjo, Familia Y: amarillo). B) En este árbol se resumen las relaciones de similitud de palmas de polimerasas de ADN. Las palmas de las familias A, B e Y descienden de un nodo bajo el umbral de corte "*Remoto*" en A). Por otra parte las palmas de las familias X y C forman grupos independientes.

5.6.2. Similitud estructural del dominio de los dedos.

Desde el punto de vista funcional, el dominio de los dedos, está involucrado en la interacción con la hebra templado y en acomodar al sustrato deoxinucleótidotrifosfato en una posición apropiada para su interacción con los residuos del aparato catalítico (Arndt et al., 2001). Cuando se revisa el agrupamiento generado para los dedos de polimerasas de ADN se observa que la clasificación en familias no se cumple completamente, pues no es posible encontrar un nodo bajo el cual desciendan todos los miembros de una familia (Figura 17). A un umbral de corte *Distante*, todos los miembros aparecen agrupados bajo un mismo nodo. Cuando se incrementa la exigencia al umbral de corte *Remoto*, se observan tres grandes grupos: el primero formado por los dedos de la familia C, el segundo por los dedos de la familia Y, y el tercero por dedos de las familias A, B y X. Finalmente al utilizar el umbral de corte más estricto (*Relacionado*), aparecen ocho grupos diferentes. De estos ocho grupos, tres corresponden a elementos únicos: dedos de la polimerasa del fago T7 (código PDB 2ajq), dedos de la polimerasa Gamma (código PDB 3ikm) y dedos de la polimerasa II (código PDB 3maq). A nivel más específico es posible notar que hay algunos grupos que tienen una variabilidad estructural muy baja. Tal es el caso de los dedos de la familia X, donde se ve las tres estructuras presentes acá tienen en promedio una similitud relativa del orden del 90%. La misma situación se observa para algunos miembros de la familia A (códigos PDB 1lv5, 1taq, 2kfn) y B (códigos PDB 1s5j, 2gv9, 2jgu, 3iay, 2p5o, 1tgo, 1xhx). Finalmente, la familia X presenta un caso especial donde una de las estructuras de esta familia no está presente en este agrupamiento, dado que no posee el dominio dedos (código PDB 1jaj).

En un análisis más fino del agrupamiento jerárquico obtenido, es posible notar que los dedos de la familia C, en relación a los dedos de otras se encuentran en el último nivel de la jerarquía establecida, inmediatamente por debajo del umbral *Distante*. Los dominios dedos de esta familia se caracterizan por tener la mayor longitud en la cadena proteica (mayor a 200 aminoácidos), mientras que en las otras familias tienen largos inferiores a 100 aminoácidos.

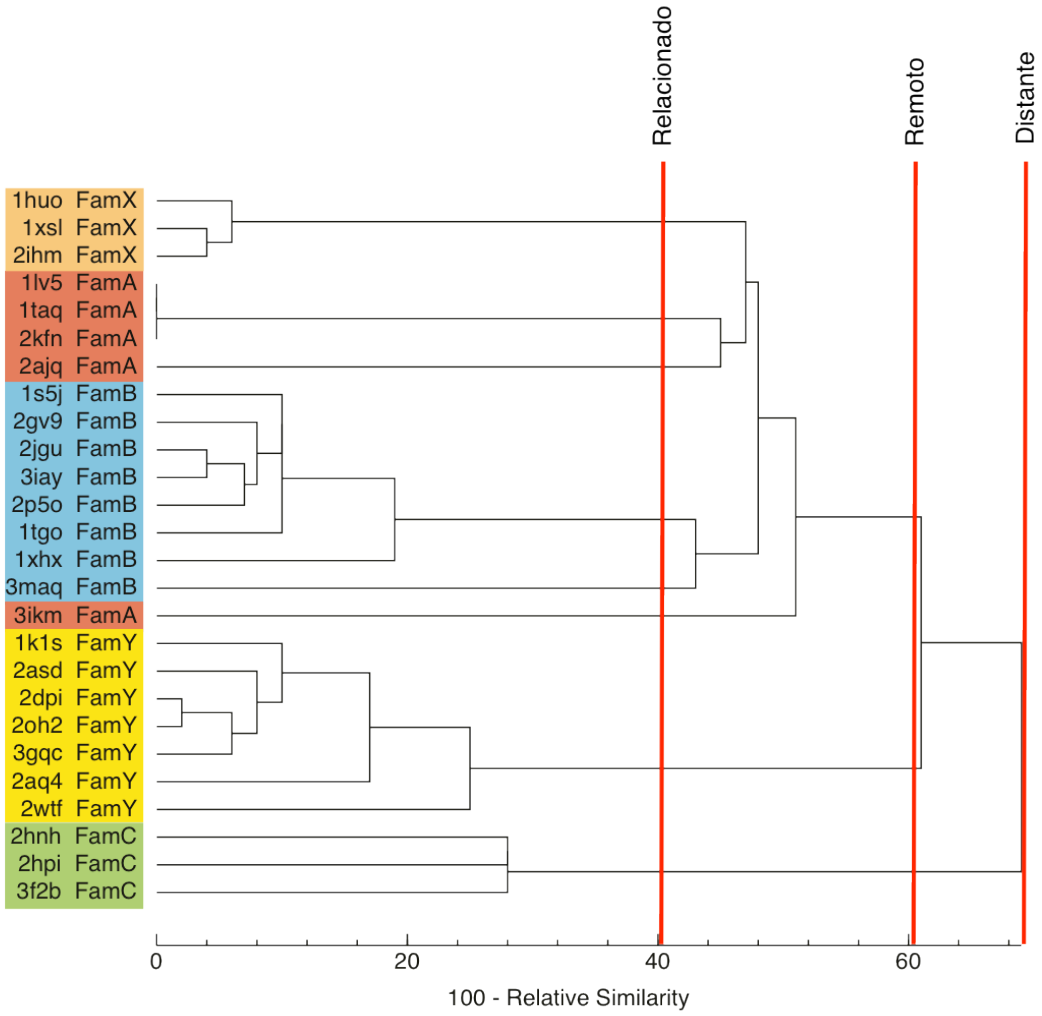


Figura 17. Dendrograma de las relaciones estructurales del dominio dedos de polimerasas de ADN. En las hojas del árbol se encuentra escrito el código PDB de la estructura y la familia a la cual pertenece esa polimerasa; adicionalmente se destaca con color la familia a la que pertenece cada dominio agrupado. Tres criterios de corte de acuerdo a la distancia estructural se presentan en la forma de líneas rojas perpendiculares al eje del árbol (*Relacionado*, *Remoto* y *Distante*). El árbol fue construido empleando el algoritmo de *Group Average*.

5.6.3. Similitud estructural del dominio de los pulgares.

En último lugar analizaron las relaciones de similitud estructural del dominio de los pulgares. Este dominio está involucrado fundamentalmente en interacciones con hebras dobles de ADN. Las interacciones ocurren fundamentalmente a través del surco menor del ADN por contactos entre el esqueleto de la proteína con los fosfatos del esqueleto del ADN. Cuando se analizan los resultados del agrupamiento jerárquico, se observa que no es posible encontrar un nodo común desde el cual desciendan todos los miembros de una misma familia. Con el criterio de corte *Distante*, se obtienen 6 grupos, uno de ellos formado por el pulgar de la polimerasa del fago Phi29 (código PDB 1xhx). Del agrupamiento jerárquico, es posible notar dos variaciones importantes, una de ellas ocurre en la familia A, donde la polimerasa Gamma (código PDB 3ikm) presenta un pulgar diferente a los observados en el resto esa familia. Según el agrupamiento su estructura es más similar a los pulgares de la familia B. Si bien este pulgar de la polimerasa Gamma conserva las dos hélices alfa antiparalelas, características de este dominio, en el extremo superior carece de un segmento de 30 aminoácidos, que posee dos hélices alfa antiparalelas que se orientan en un ángulo de 45º con respecto a las hélices de mayor longitud. En su lugar es posible apreciar la inserción de un segmento de 310 aminoácidos. Se utilizó la herramienta *TopSearch* para buscar estructuras similares, sin embargo todas las coincidencias encontradas tienen alineamientos con longitudes de alineamiento menores a 30. Lo anterior sugiere que este es un motivo estructural nuevo en la base de datos del PDB. En la familia B también se aprecian variaciones importantes, tal como ocurre con la estructura del pulgar de la polimerasa del fago Phi29 (código PDB 1xhx). Al emplear el criterio de corte *Remoto*, se obtiene un total de 7 grupos estructurales, siendo la familia B la

más diversa en este aspecto, pues aporta con 3 grupos. Finalmente, cuando se utiliza el criterio de corte *Relacionado*, se obtiene un total de 11 grupos estructurales diferentes de dedos, de los cuales 6 corresponden a elementos simples (Figura 18).

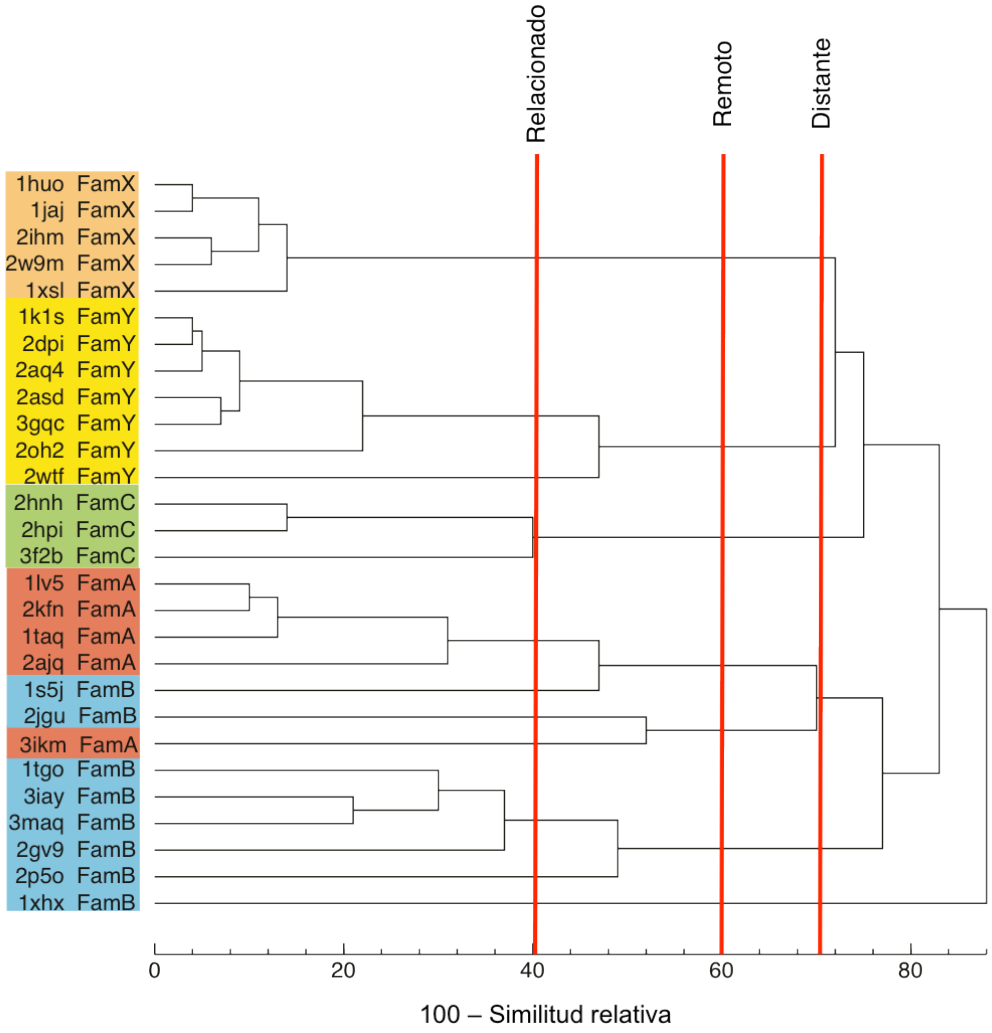


Figura 18. Dendrograma de las relaciones estructurales del dominio pulgares de polimerasas de ADN. En las hojas del árbol se encuentra escrito el código PDB de la estructura y la familia a la cual pertenece esa polimerasa; adicionalmente se destaca con color la familia. Tres criterios de corte de acuerdo a la distancia estructural se presentan en la forma de líneas perpendiculares al eje del árbol (*Relacionado*, *Remoto* y *Distante*). El árbol fue construido empleando el algoritmo de *Group Average*.

5.7. Las polimerasas de ADN constituyen una familia diversa a nivel estructural y de secuencia.

5.7.1. Diversidad en la estructura tridimensional y la secuencia.

De acuerdo a lo observado en la comparación a nivel de cadenas completas, en la familia A se aprecian dos grupos claramente definidos. El primero de ellos, al que se denominará canónico, está representado por la polimerasa I de *E. coli*. El segundo grupo se encuentra representado por la polimerasa Gamma de *H. sapiens*. Algunas polimerasas del grupo canónico están involucradas en la síntesis de los fragmentos de Okazaki durante el proceso de replicación del ADN; otras están involucradas en la replicación de genomas virales, como es el caso de la polimerasa del fago T7. Por su parte, la polimerasa Gamma es la enzima encargada de replicar el genoma mitocondrial. Las polimerasas presentes en el grupo canónico comparten una identidad de secuencia promedio de un 47%. En lo que se refiere a la similitud estructural, este grupo cuenta con una similitud promedio del 69%. La polimerasa Gamma es altamente divergente en relación a los demás miembros de la familia A; con su vecino más cercano comparte un 20% de identidad de secuencia (polimerasa TAQ de *T. aquaticus*, código PDB 1taq). Desde el punto de vista estructural, la similitud alcanza un máximo de 38% con el resto de los miembros. Cuando se estudia en detalle este grupo a partir de los alineamientos de múltiples estructuras complementados según el procedimiento descrito en esta tesis, se observa que la identidad de secuencia para las polimerasas del grupo Gamma es de un 77% en promedio. Cuando se comparan estas secuencias con aquellas que tradicionalmente se han utilizado para la identificación de

nuevas polimerasas de la familia A, se observan diferencias importantes como las que ocurren en el dominio de la palma. Este dominio, que usualmente es muy conservado dentro de una familia, presenta diferencias importantes entre el grupo canónico y el grupo Gamma. A modo de ejemplo, entre las hebras 2 y 3 solamente se observa la conservación de tres aminoácidos, dos de los cuales son responsables de la catálisis (ácido aspártico y glutámico). El resto del segmento está poco conservado (Figura 19). Los datos estructurales y de secuencia, sugieren que la polimerasa Gamma representaría una clase particular dentro de la familia A.

Desde el punto de vista estructural, la polimerasa Gamma presenta algunas variaciones destacables. En primer lugar, a diferencia de lo que se observa en el grupo canónico, el pulgar de esta polimerasa presenta una inserción en la forma de un dominio espaciador, el cual tiene una longitud de 310 aminoácidos. Con esto, el pulgar queda convertido en un dominio multisegmento; el resto de los dominios pulgares presentes en el grupo canónico es del tipo monosegmento. La región correspondiente a la inserción presenta dos segmentos claramente definidos desde el punto de vista estructural. Una de ellas es altamente estructurada y se encuentra próxima al dominio exonucleasa, la segunda presenta una baja estructuración secundaria y se encuentra hacia el exterior de la proteína. La búsqueda de estructuras similares mediante la herramienta *TopSearch* no arroja resultados relevantes, dado que los alineamientos estructurales entregados tienen longitudes del alineamiento muy bajas (menores a 30 pares alineados). La excepción a lo anterior lo constituye la recuperación de otras estructuras de polimerasas Gamma, con las que evidentemente comparte la máxima similitud posible

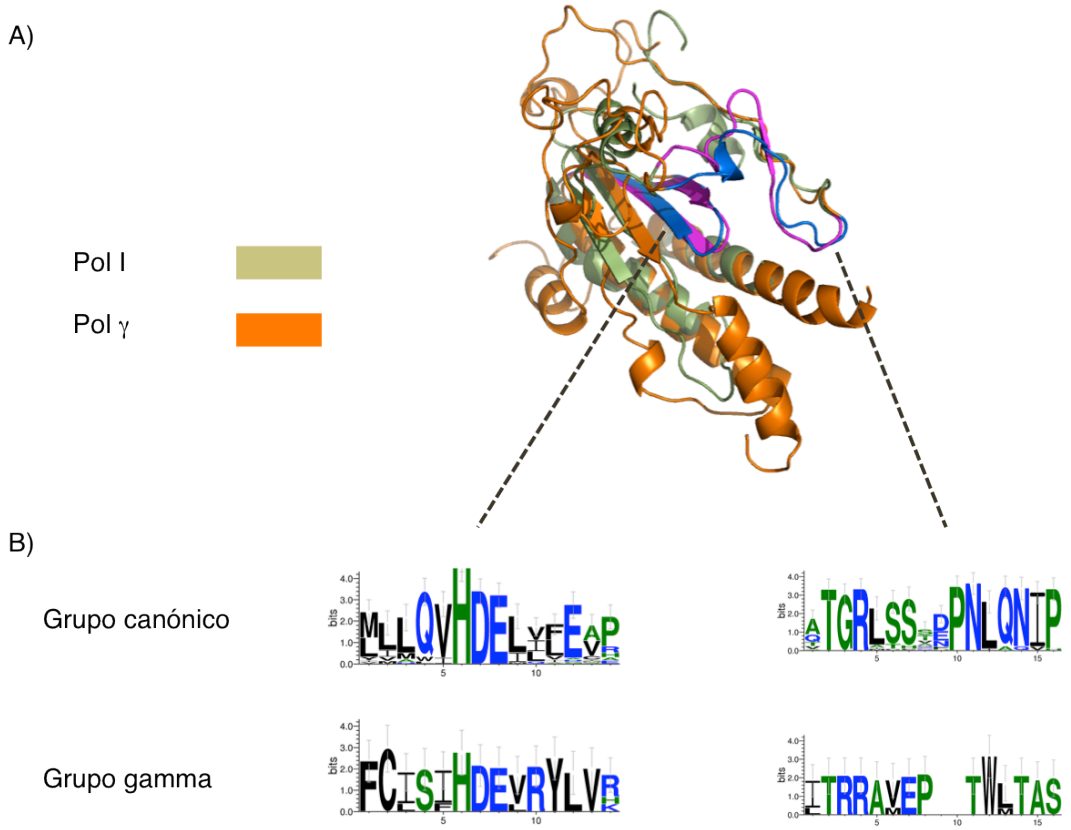


Figura 19. Diversidad de secuencia en polimerasas de ADN de la familia A. Ejemplo de la diversidad de secuencia observada en polimerasas de la familia A. A) Superposición óptima de la polimerasa I (código PDB 2kfn, en color verde) y la polimerasa Gamma (código PDB 3ikm en color naranjo). En las estructuras se destacaron regiones que usualmente han sido utilizadas como firmas de secuencia de la familia A, en color magenta para la polimerasa I y en azul para la polimerasa Gamma. B) Logos de secuencia de las regiones estructurales destacadas en A. Se muestran logos de secuencia para el grupo canónico, representado por la polimerasa I y para el grupo gamma representado por la polimerasa Gamma.

. Lo anterior sugiere que este segmento de 310 aminoácidos es una característica particular de esta polimerasa y que además podría constituir un pliegue nuevo, dada la baja similitud con otras estructuras presentes en el PDB. Búsquedas basadas exclusivamente en la información de secuencia mediante la herramienta BLAST, utilizando la secuencia del segmento espaciador, recuperan únicamente polimerasas tipo Gamma. Lo mismo ocurre empleando búsquedas con PSI-BLAST. El dominio de la palma también tiene algunas diferencias en relación a lo observado en las otras polimerasas canónicas de la familia A. Se observa en el grupo de las polimerasas Gamma una inserción de 45 aminoácidos en el extremo C-terminal. Esta inserción se encuentra a continuación de la hebra 4 que conforma el núcleo del dominio de la palma. Un tercer aspecto en el cual se aprecia diversidad es el de la organización en la estrutura primaria. El grupo de la polimerasa Gamma está constituido por dos segmentos, mientras que el grupo canónico es un dominio de tres segmentos (Figura 20).

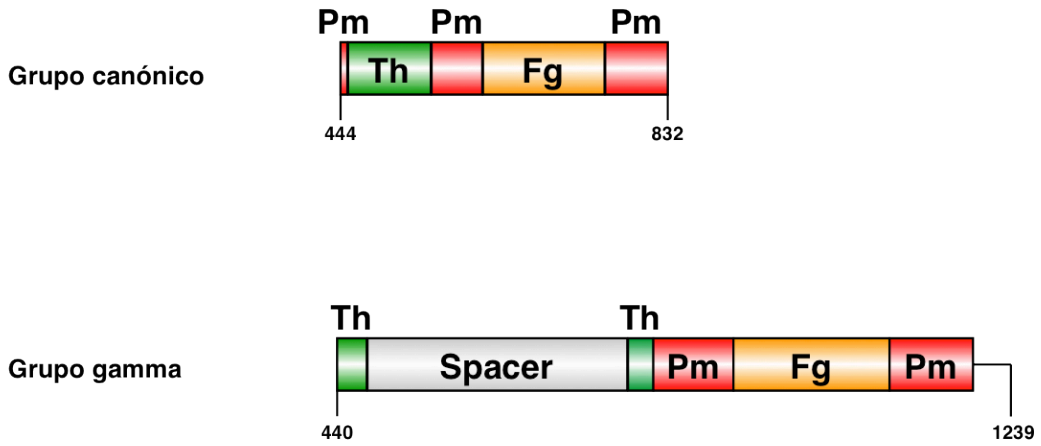


Figura 20. Organización del dominio catalítico de polimerasas de la familia A. Representación de la organización de los dominios catalíticos de polimerasas de la familia A en la estructura primaria. Los dominios palma (Pm) se encuentran destacados en color rojo, los pulgares (Th) en color verde, y los dedos (Fg) en color naranja. Las polimerasas del grupo canónico se caracteriza por tener dominios palmas compuestos de tres segmentos, mientras que los pulgares y dedos son monosegmento. El grupo gamma, se caracteriza por tener pulgares interrumpidos por un dominio espaciador (*Spacer*, color gris). Por otra parte, las palmas son de dos segmentos en este grupo.

En la familia B también es posible encontrar otro ejemplo de variación, tanto a nivel estructural como de secuencia. En los alineamientos estructurales se observa que la estructura de la polimerasa del fago Phi29 (código PDB 1xhx) tiene un 20% de identidad de secuencia como máximo con miembros de su misma familia. Desde el punto de vista estructural comparten a lo menos un 30% de similitud relativa, lo que según los criterios de corte empleados establece una relación estructural entre las proteínas, pero distante.

Al estudiar las superposiciones, se observa de manera clara que la diferencia fundamental entre la polimerasa del fago Phi29 y el resto de las polimerasas de la familia B se encuentra en la estructuración del dominio pulgar. Búsquedas de similitud estructural de este dominio mediante la herramienta *TopSearch* no recuperan similitud con otros dominios de polimerasas de ADN, o de alguna otra proteína asociada a la replicación. Lo anterior sugiere que este dominio es una característica exclusiva de este tipo de polimerasas. Desde el punto de vista funcional se ha sugerido que este segmento tiene un rol equivalente a las abrazaderas de ADN (*DNA clamps*) que están presentes como factores accesorios en varias polimerasas de ADN, los cuales tienen relación el incremento de la procesividad de los complejos de la replicación.

Un análisis más detallado de la diversidad y conservación de secuencia presente en la familia B fue efectuado a partir de los alineamientos de múltiples estructuras que pueden ser derivados de las superposiciones de los miembros de esta familia (Figura 21). Si se toma de manera específica el dominio de la palma, es posible encontrar dos firmas características de la familia B que aparecen en todos los miembros, tanto aquellos del grupo

canónico, como aquellos del grupo Phi29. El primer elemento de secuencia altamente conservado es la firma `DxxSYLPsii` (en minúscula se indican aquellos aminoácidos que tienen se conservan en un 75% o menos del los casos). A nivel estructural, este motivo corresponde a un segmento con estructura secundaria del tipo hélice alfa que está ubicada en la base de la palma. El segundo motivo de secuencia está en directa relación con el aparato catalítico de esta polimerasa (`YxDTD`), que contiene dos ácidos aspárticos responsables de la catálisis.

La conservación descrita anteriormente es estrictamente local y no da cuenta de la real diversidad que existe en esta familia. Por ejemplo, al realizar búsquedas mediante el algoritmo de PSI-BLAST sembradas con la secuencia de la polimerasa del fago Phi29, sólo se recuperan polimerasas virales altamente similares a la del bacteriófago. Por otra parte, cuando se siembra una búsqueda en PSI-BLAST con cualquiera de las otras secuencias presentes en la base de datos de estructura, tampoco es posible recuperar polimerasas similares a la del bacteriófago Phi29 (código PDB 1xhx). Esto sugiere que a nivel de secuencia existen dos grupos diferentes de polimerasas dentro de esta familia.

Una interpretación alternativa del resultado anterior es que la separación en dos grupos a nivel de secuencia pueda ser artefactual, considerando que previamente se mostró que existen patrones conservados (los que fueron identificados mediante alineamientos estructurales). Un error de esta naturaleza puede impactar de manera negativa la construcción de alineamientos de secuencia para buscar nuevos miembros o para identificar características de conservación en ella.

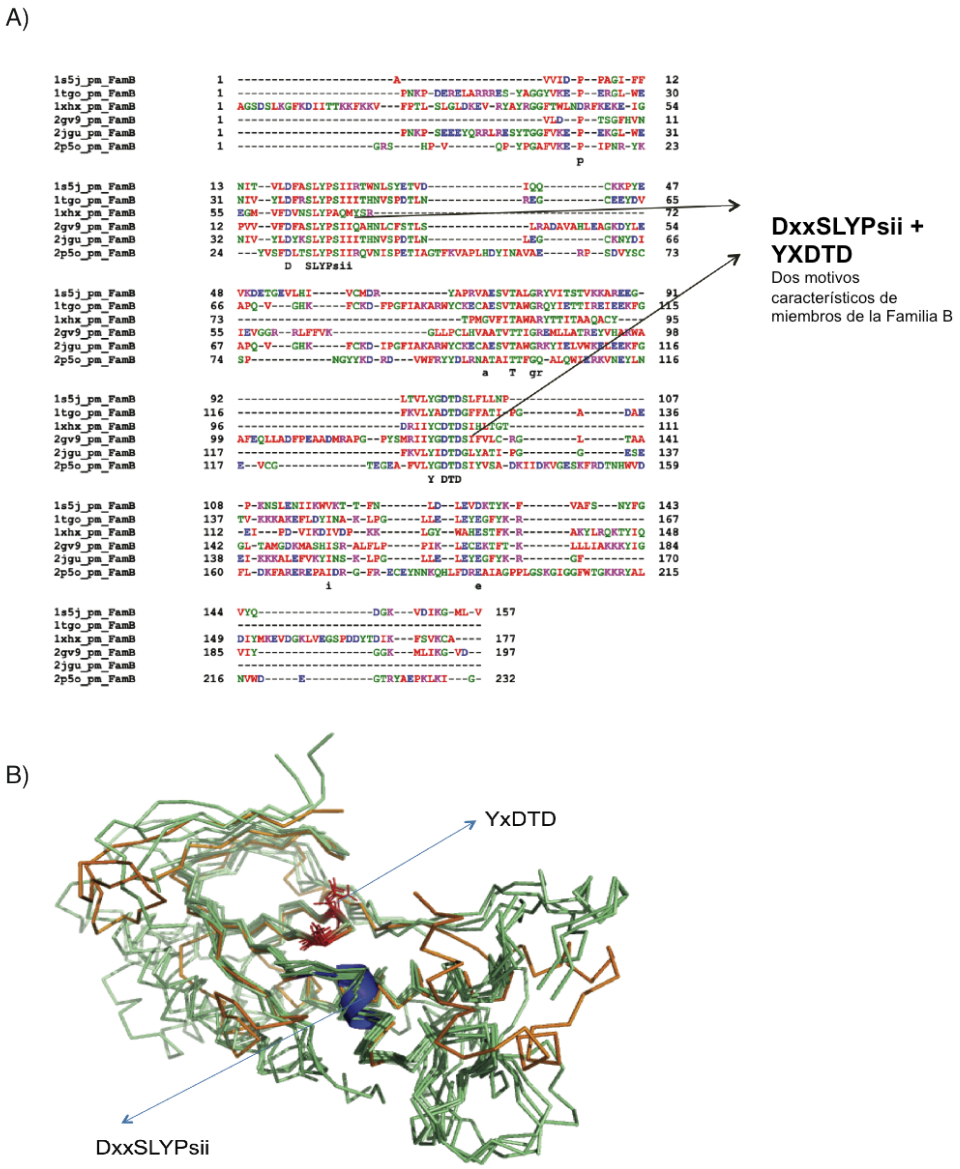


Figura 21. Conservación estructural y de secuencia en las palmas de polimerasas de la familia B. En la figura se muestra un alineamiento de múltiples estructuras de polimerasas de la familia B. A) Representación en forma de secuencias del MStA de palmas de la familia B. Se destacan mediante flechas dos elementos altamente conservados a nivel de la secuencia y que son característicos de esta familia de polimerasas. El elemento **YxDTD** es parte de la díada catalítica de esta familia. B) Superposición de palmas de polimerasas de la familia B. Las proteínas están representadas en su esqueleto de la cadena principal. La estructura de la polimerasa del fagoPhi29 (código PDB 1xhx) se encuentra destacada en color naranjo, mientras que el resto de las polimerasas de la familia B se encuentran en color verde. Los motivos de secuencia altamente conservados se han destacado en la superposición óptima. El motivo **YxDTD** se destacó en color rojo, en tanto que el motivo **DxxSLYPsii** se destacó en color azul.

Al comparar alineamientos derivados de la información de secuencia únicamente (aproximación clásica), con alineamientos que utilizan información estructural (MStAs), es posible observar dos hechos importantes. El primero es que los alineamientos de secuencias contienen errores, y el segundo es que los alineamientos estructurales corrigen los errores de los primeros. En la Figura 22 se muestra un ejemplo de esta situación. El MSA construido presenta un error en el alineamiento del motivo `DxxSLYPsii`, que con base en la información recopilada de los alineamientos de múltiples estructuras, es posible afirmar que el motivo de secuencia `DxxSLYPsii` es altamente conservado en polimerasas de la familia B, incluso para aquellas que se alejan del prototipo de esta familia (como es el caso de la polimerasa del fago Phi29). Por otra parte, la utilización de un MStA complementado con las mismas secuencias presentes en el MSA corrige este error, permitiendo así el correcto alineamiento de estas secuencias. De esta manera, el MStA incrementa la relación señal/ruido en los alineamientos de secuencias proteicas de familias altamente divergentes.

Si bien la variación a nivel de secuencia y estructura observada para la polimerasa Phi29 es significativa, aún existen elementos conservados de la secuencia que la ligan a la familia B, como son los patrones presentes en el dominio de la palma. Por otra parte, su situación es alejada en relación a los otros miembros de la familia B, ya que entre ellos tienen una variación mucho más contenida.

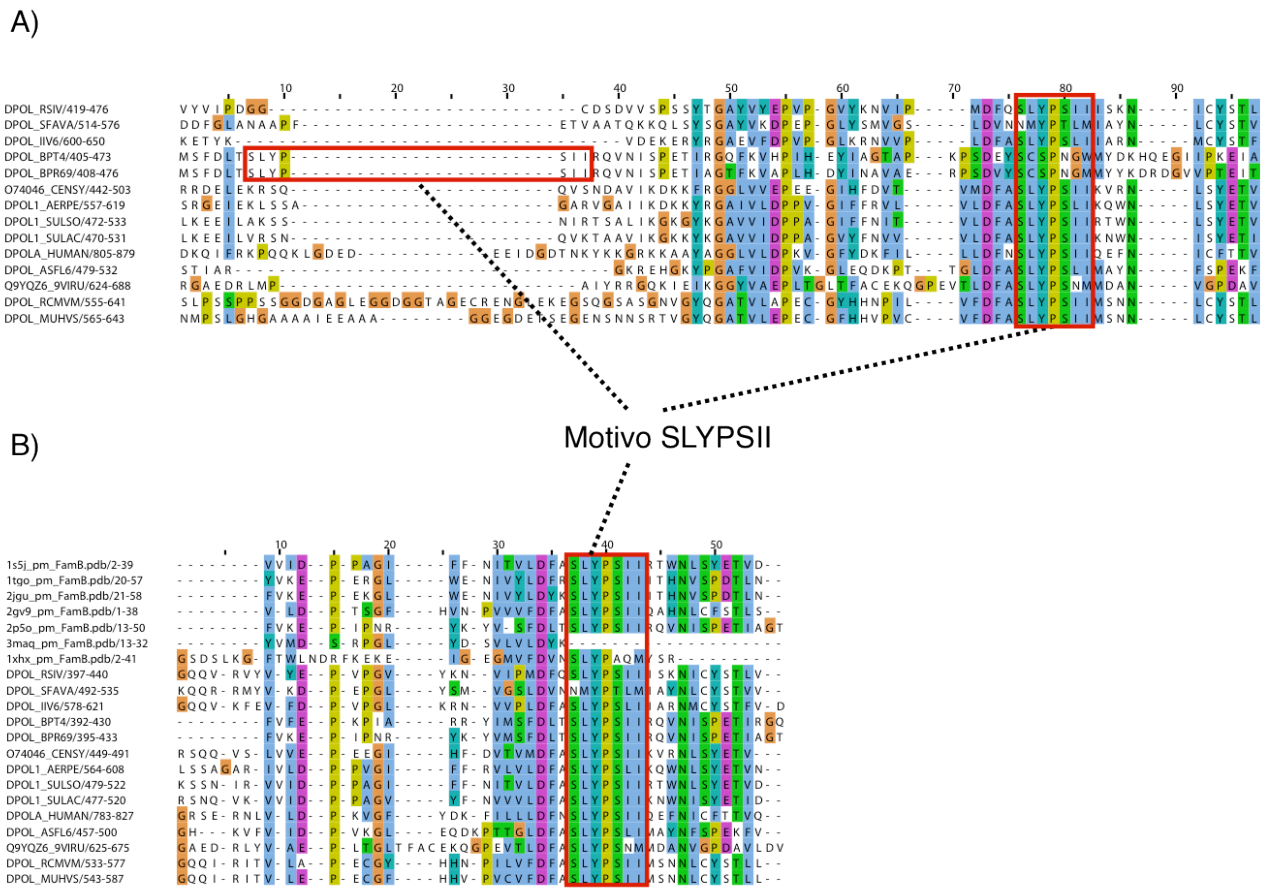


Figura 22. Comparación entre alineamientos de secuencia y alineamiento estructural. En la figura se muestran dos alineamientos múltiples, uno derivado con información de secuencia únicamente (A) y otro utilizando un MStA de base que fue complementado con las mismas secuencias presentes en el MSA (B). El alineamiento de múltiples secuencias presenta errores pues, el motivo de secuencia **SLYPSii,** altamente conservado en proteínas de la familia B se encuentra mal alineado en dos secuencias (DPOL_BPT4 y DPOL_BPR69). Los errores se encuentran destacados con cajas rojas. En B), el MStA corrige dichos errores, dado que ahora todas las secuencias contienen el motivo **SLYPSii** correctamente alineado; además se generan alineamientos más compactos y con menor proporción de gaps. También se encuentra destacado el motivo de secuencias con una caja roja en el MStA.

La familia C presenta un grado contenido de variación estructural y de secuencia. Las estructuras analizadas comparten una similitud relativa del 61%, mientras que su similitud de secuencias es del 33% en promedio. De acuerdo a lo reportado en la literatura hasta la actualidad, estas polimerasas sólo se han encontrado en el dominio Bacteria. Mediante el uso de perfiles construidos a partir de alineamientos estructurales y complementados con secuencias de bases de datos (según lo descrito en la sección de métodos), se realizó una búsqueda en la base de datos de secuencias de proteínas no redundantes de NCBI. Esta búsqueda entregó resultados negativos tanto para Arqueas, Eucariontes y Virus, recuperándose únicamente secuencias pertenecientes al dominio Bacteria. Adicionalmente, se realizó una segunda búsqueda exhaustiva en genomas virales disponibles en NCBI, sin embargo los resultados para polimerasas de la familia C también fueron negativos.

La familia X también presenta un grado contenido de variación estructural, con una similitud relativa del 54%. De todas formas, cuando se estudian en detalle los miembros de esta familia es posible encontrar algunas variaciones que no necesariamente se ven reflejadas en el porcentaje mencionado anteriormente. Por ejemplo, en esta familia la polimerasa ASFV se caracteriza por ser la polimerasa más pequeña cristalizada hasta el momento. Además, la estructura de su dominio catalítico carece del dominio de los dedos. A nivel de secuencia la variación también es importante, en este caso a nivel de cadenas completas es de un 26%. Otra característica de esta familia es que posee dominios catalíticos muy pequeños en relación a lo observado en las familias A, B y C. En promedio tienen una longitud de 225 aminoácidos, comparados con los 430 aminoácidos de las familias A, B y C.

A diferencia de lo revisado en las familias anteriores, las palmas de las polimerasas de la familia X poseen motivos de secuencia altamente conservados. Un ejemplo es el caso de un motivo de 21 aminoácidos donde se encuentran dos de los ácidos aspárticos que forman parte de la tríada catalítica de la familia X. Estructuralmente, estos aminoácidos forman un *loop* que conecta a las hebras 1 y 2 de la palma. Otro segmento altamente conservado posee 10 aminoácidos y estructuralmente corresponde a la parte terminal de la hebra 5 de la palma más un segmento sin estructura secundaria definida, que conecta con el dominio de los dedos que se encuentra altamente conservado. Sobre la hebra 5 es posible encontrar el tercer ácido aspártico que forma parte de la tríada catalítica. (Figura 23)

La familia Y también presenta un alto grado de conservación estructural; la similitud relativa promedio observada para las estructuras de esta familia es de un 72%. A nivel de secuencia la conservación en cadenas completas tiene un promedio de 29% de identidad de secuencia a partir de los alineamientos estructurales. Al igual que la familia X, los dominios catalíticos son pequeños teniendo en promedio una longitud de 268 aminoácidos. Desde el punto de vista estructural, el dominio catalítico de esta familia es altamente conservado y no se observan miembros dentro de la familia que se diferencien de manera considerable del resto. Los análisis independientes a nivel de cadena completa, dominio catalítico y dominios estructurales son consistentes en este sentido. La principal forma de variación observada en esta familia tiene que ver con la inserción de pequeños segmentos que actúan como conectores de los dominios estructurales de estas polimerasas

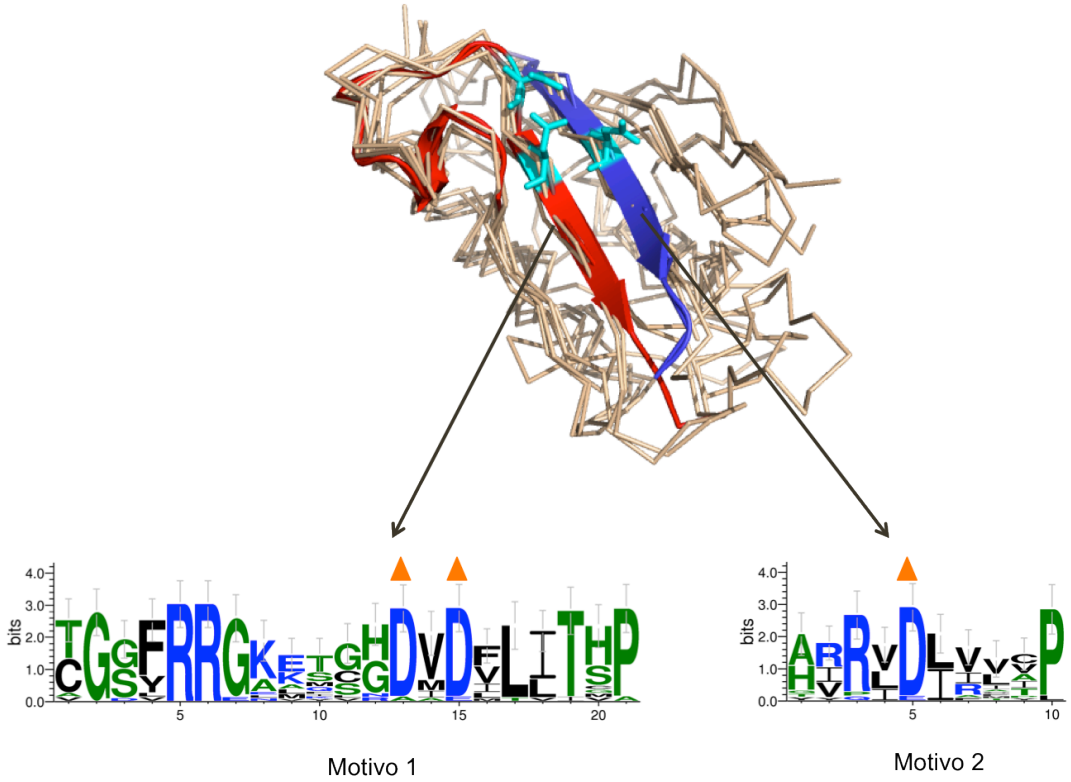


Figura 23 Motivos conservados en palmas de polimerasas de la familia X. En la parte superior de la figura se muestra el alineamiento de múltiples estructuras de dominios palma de la familia X. Las proteínas están representadas con su cadena principal y en color café claro. En la superposición se han destacado dos segmentos (en rojo y azul) que presentan mayor conservación de secuencia y utilizando representación de *cartoons*. La hebra en color rojo corresponde a la hebra número 2 de la palma en la familia X, mientras que la hebra en color azul corresponde a la hebra 5 de este dominio. Las flechas apuntan a los logos de secuencia obtenidos para cada uno de los motivos destacados. En cada uno de ellos se han marcado con triángulos de color naranja los residuos de ácido aspártico que forman parte de la tríada catalítica.

5.7.2. Organización de los dominios en la estructura primaria.

De acuerdo a los datos mostrados en secciones anteriores de este escrito, el dominio de la palma es el más conservado dentro del dominio catalítico de polimerasas de ADN. Sin embargo, cuando se estudia la organización de los dominios en la estructura primaria es posible también encontrar algunos patrones que relacionan algunas familias.

Antes de describir más en detalle la organización se utilizarán en adelante las siguientes convenciones: P para referirse al dominio palma (*Palm)*, F para referirse al dominio dedos (*Fingers) y* Th para el dominio pulgar (*Thumb)*. Estos tres dominios estructurales forman el dominio catalítico funcional. El detalle de la descomposición en dominios estructurales para las polimerasas de ADN de la base de datos se muestra en la Tabla VI.

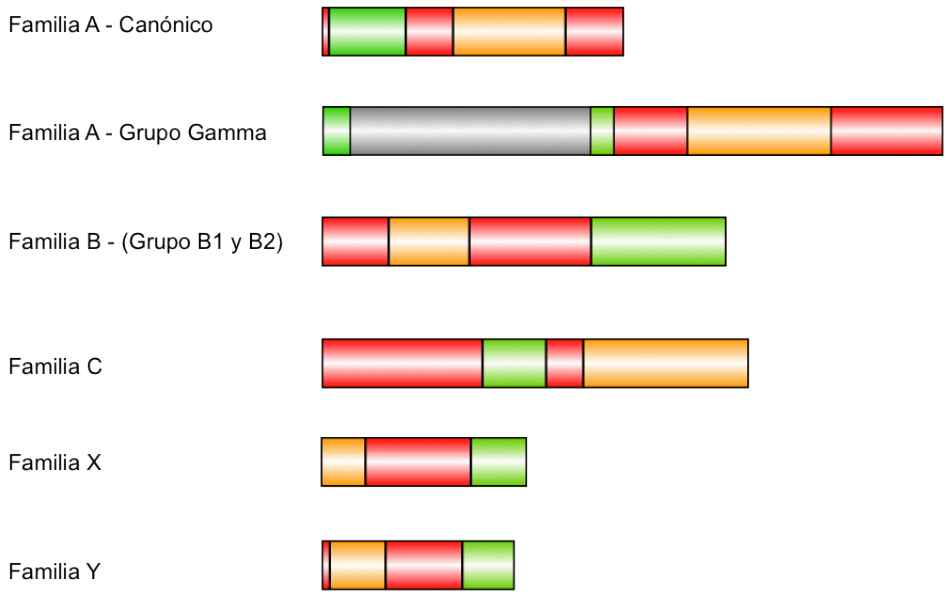
Tabla VI. Descomposición del dominio catalítico de polimerasas de ADN en sus dominios estructurales.

		Límites de los dominios				
Código PDB	**ID de Cadena**	**Palma (P)**	**Dedos (F)**	**Pulgares (Th)**	**Nombre Común**	**Familia**
1taq	A	444-452:552-612:758-832	613-757	453-551	Taq Polymerase	A
1lv5	A	485-495:595-655:801-876	656-800	496-594	Bacillus Fragment	A
2ajq	B	212-235:411-476:617-704	477-617	236-264:332-410	T7 Pol	A
2kfn	A	519-548:657-707:853-928	708-852	546-656	Klenow	A
3ikm	A	815-910:1095-1239	910-1095	440-475:785-815	Pol Gamma	A
2p5o	A	383-468:573-729	469-572	730-903	RB69	B
3maq	A	391-421:520-634	422-519	635-783	Pol II	B
1tgo	A	369-449:500-585	452-492	586-773	Pol TGO	B
2gv9	B	701-766:826-956	767-825	957-1197	HSV-1	B
2jgu	A	369-450:501-588	451-500	589-775	Pfu	B
1s5j	A	493-559:621-718	560-620	719-882	Pol B1 Ss	B
1xhx	_	190-261:427-531	359-395	531-575	Phi29	B
3iay	A	576-660:715-835	661-714	836-985	Delta	B
2hnh	C	271-432:511-560	561-911	433-510	Pol III	C
3f2b	A	829-1002:1050-1103	1103-1294	1002-1050	Pol III	C
2hpi	C	286-492:575-622	623-835	493-574	Pol III	C
1huo	A	152-262	88-151	263-335	Beta	X
1xsl	A	386-494	328-385	495-575	Lambda	X
2ihm	A	289-424	232-288	425-496	Mu	X
1jaj	_	1-105	N/A	106-174	ASFV Pol X	X
2w9m	A	161-241	91-161	241-306	Pol X	X
2asd	B	1-10:78-166	11-77	167-233	Dpo4	Y
2dpi	A	26-37:99-224	38-98	225-288	Iota	Y
2wtf	A	1-32:132-294:380-389	33-131	295-379	Eta	Y
2oh2	A	98-109:170-337	110-169	338-411	Kappa	Y
1k1s	A	1-19:78-171	20-35:39-77	172-236	DinB	Y
2aq4	A	356-365:438-536	366-437	537-603	yRev1	Y
3gqc	A	417-426:542-632	427-445:505-545	632-697	hRev1	Y

Las polimerasas de la familia A tienen una estructura del tipo *P-Th-P-F-P*. Los dominios palma están formado por tres segmentos a excepción de la polimerasa Gamma donde es de dos segmentos. Los dominios dedos son monosegmento en todas las polimerasas de esta familia y siempre se encuentran flanqueados por segmentos del dominio palma. El caso de los dominios pulgares también presenta variaciones al interior de la familia. Existen dos casos donde el dominio de los pulgares está compuesto por dos segmentos (polimerasa T7, código PDB 2ajq y polimerasa Gamma, código PDB 3ikm). Entre estos segmentos es posible encontrar un segmento espaciador. Las polimerasas de la familia B tienen una organización del tipo *P-F-P-Th*. En general los dominios de la palma están conformados en todos los casos por dos segmentos. Los dominios de los dedos son siempre monosegmento y se encuentran flanqueados por los segmentos de la palma. Finalmente el dominio del pulgar es siempre monosegmento en esta familia. La única excepción a lo revisado anteriormente aparece en la polimerasa del fago Phi29 (código PDB 1xhx), en la cual los dedos no se encuentran flanqueados directamente por la palma, sino por otros dominios. De todas maneras se respeta la organización P-F-P. La familia C presenta una organización del tipo *P-Th-P-F*. La palma es un dominio formado por dos segmentos y entre éstos se encuentra el dominio pulgar que es monosegmento en los casos estudiados. Finalmente, hacia el C-terminal en la secuencia, se encuentra el dominio de los dedos que es monosegmento en las estructuras estudiadas. La familia X tiene una organización del tipo *F-P-Th*. Todos los dominios de esta familia son del tipo monosegmento. Por último, la familia Y presenta una organización del tipo *P-F-P-Th*. Los dominios de la palma contienen dos segmentos y se encuentran flanqueando al dominio de los dedos. El dominio del pulgar de esta familia es siempre monosegmento. En resumen, prácticamente en todas

las polimerasas de ADN los tres dominios estructurales que forman el dominio catalítico forman una unidad continua en la estructura primaria. La excepción a esta observación se encuentra en el grupo de las polimerasas Gamma que actualmente pertenecen a la familia A (Figura 24).

Recapitulando los resultados expuestos en esta sección podemos decir que las estructuras analizadas en promedio comparten un 28% de identidad de secuencia a nivel de cadena completa (derivada a partir de datos estructurales). Se identificaron algunas variaciones en miembros de las familias A y B, donde algunos de ellos se distancian de manera importante de los grupos principales que existen dentro de cada una de esas familias. Estas diferencias ocurren tanto a nivel de secuencia como de estructura y sugieren que la actual clasificación en familias es incompleta, pues no da cuenta de la real diversidad que existe en esta clase de enzimas.



	Longitud Promedio		
Familia	**Palma**	**Dedos**	**Pulgares**
A	171	151	95
B	173	60	178
C	230	251	69
X	107	61	69
Y	138	68	68

Figura 24. Organización de dominios estructurales de polimerasas de ADN en la estructura primaria. Cada rectángulo muestra la organización de los dominios estructurales de la palma, dedos y pulgares en las diferentes familias de polimerasas de ADN. El tamaño del rectángulo es proporcional al número de aminoácidos que constituyen a cada dominio. Los dominios palma se muestran en color rojo, los dedos en color naranja y los pulgares en color verde. Cualquier otro dominio distinto a los anteriores es asignado a una única categoría llamada “otros”, que aparece de color gris en la figura. En la esquina superior derecha de la figura se muestra una tabla con la longitud promedio observada para cada dominio estructural en las diferentes familias de polimerasas de ADN. Finalmente en la esquina inferior derecha, se muestra la leyenda de colores utilizados para cada dominio estructural.

5.8. Relaciones de similitud estructural entre dedos y pulgares de polimerasas de ADN.

Con la finalidad de establecer si existen relaciones cruzadas entre algunos dominios estructurales de polimerasas de ADN, se exploraron comparaciones de los dominios de pulgares y dedos. Para estos efectos se generaron listas de archivos que contenían las estructuras aisladas de cada subdominio y se realizó la comparación estructural de todos contra todos siguiendo la metodología descrita en el punto 4.4. De acuerdo a esto, un total de 55 dominios fueron comparados dando un total de 1,485 comparaciones posibles.

Al analizar el agrupamiento jerárquico obtenido de estas comparaciones es posible apreciar algunas relaciones cruzadas entre dominios de polimerasas de ADN. La primera de ellas corresponde a la relación entre los pulgares de polimerasas de la familia A y dedos de polimerasas de familia B. El agrupamiento jerárquico muestra que estos dominios estructurales descienden desde un nodo común. Dentro de él es posible apreciar dos subgrupos que contienen a los pulgares de la familia A y los dedos de la familia B. La segunda relación cruzada que se observa en el dendrograma es la que ocurre entre los dedos de la familia X y los pulgares de la familia Y. De acuerdo al agrupamiento jerárquico, también comparten un nodo común y también es posible distinguir dos subgrupos independientes para cada dominio estructural por familia. Los agrupamientos fueron generados a partir de un criterio de corte del 50% sobre la similitud estructural (Figura 25). En primer lugar se describirá la relación que ocurre entre dominios de la familia A y B.

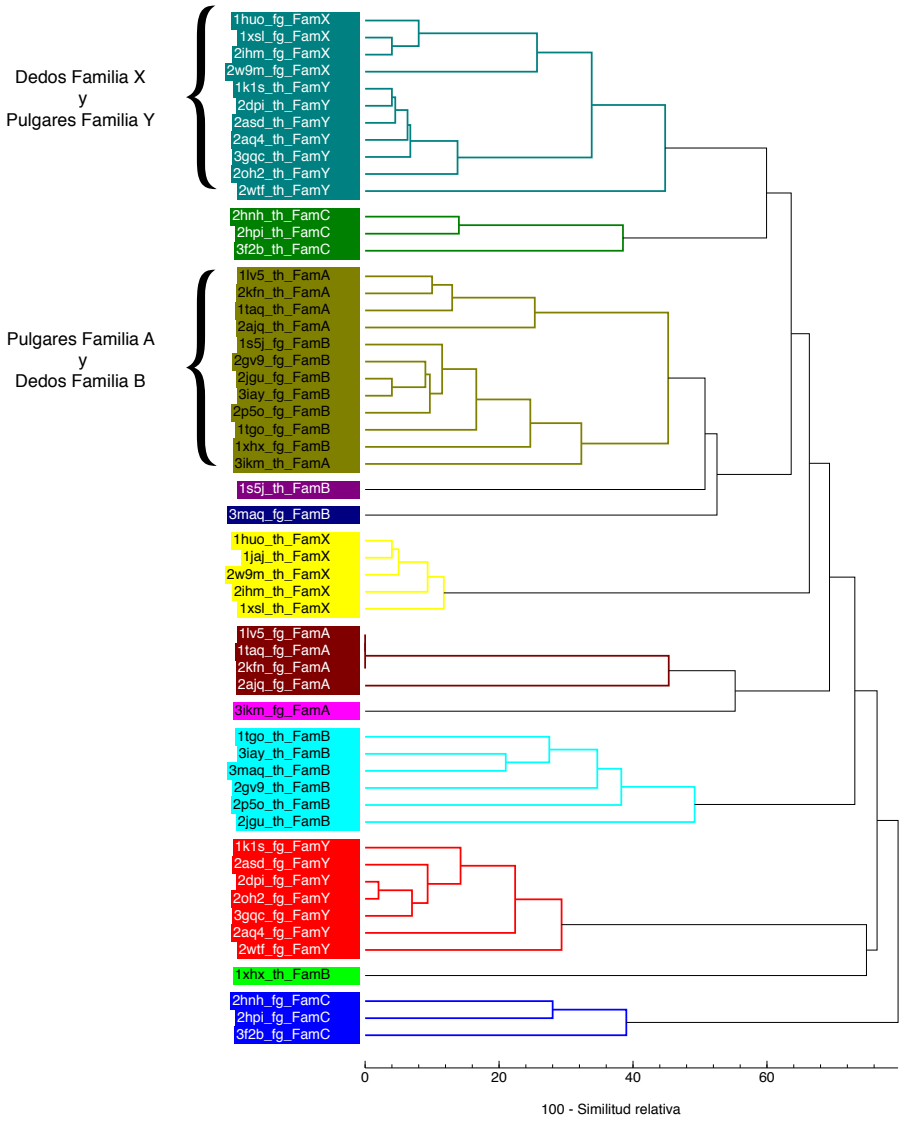


Figura 25 Dendrograma de similitud estructural de dedos y pulgares de polimerasas de ADN. En el dendrograma se muestran las relaciones de similitud estructural emtre dedos y pulgares de polimerasas de ADN. Se destacan dos grupos de pulgares y dedos relacionados. El primero de ellos corresponde al grupo de dedos de la familia X y pulgares de la familia Y destacados en color verde claro. El segundo corresponde al grupo de pulgares de la familia A y dedos de la familia B destacados en color verde en el árbol. Las relaciones fueron determinadas estableciendo un criterio de corte de 50% de similitud.

Los pulgares de la familia A son dominios de segmento continuo que están constituidos por cuatro hélices alfa. Dos de ellas, con longitudes entre 23 y 28 aminoácidos y disposición antiparalela, constituyen la base del pulgar. En la punta del pulgar se observan otras dos hélices alfa con tamaños entre 10 y 12 aminoácidos conectadas por un número variable de aminoácidos sin una estructura secundaria definida. Esta punta del pulgar se encuentra orientada hacia la cavidad que recibe al ADN en la polimerasa. La estructura completa tiene una forma análoga a la de un gancho.

Esta relación de similitud fue explorada de manera más amplia empleando la herramienta *TopSearch*. Cada uno de los dominios (dedos y pulgares de la familia A y B) fueron utilizados como sonda para realizar búsquedas sobre cadenas de proteínas. En ambos casos se recuperan alineamientos estructurales tanto del dominio de la misma familia como aquel con el que se tiene relación cruzada. A parte de estos, se obtienen una serie alineamientos entre el dominio sonda y proteínas no relacionadas con la función de polimerasa de ADN. Las longitudes de alineamiento observadas en estos casos, son similares a aquellos valores obtenidos entre pulgares de la familia A y dedos de la familia B. Lo anterior sugiere que este arreglo estructural puede que no esté relacionado directamente con la función de polimerización, sino con un rol en la estabilidad del plegamiento en proteínas.

Si bien la similitud estructural de estos segmentos es alta (un promedio de 48% de similitud relativa), a nivel de secuencia la señal es muy baja, pues en promedio los alineamientos estructurales entregan un promedio de 20% de identidad de secuencia. Pese a

esto, no se puede descartar la existencia de algún patrón de conservación dado que ambos tipos de dominio presentan interacciones con ADN según lo reportado en las estructuras cristalográficas. Por lo anterior, se exploraron en detalle los alineamientos de múltiples estructuras para las superposiciones de este segmento y posteriormente se representaron en la forma de logos de secuencia. En este análisis se observó que no existen patrones de conservación entre los pulgares de la familia A y dedos de la familia B (Figura 26). En conjunto, la evidencia recolectada a nivel estructural y de secuencia, sugiere que estos segmentos corresponden a análogos estructurales.

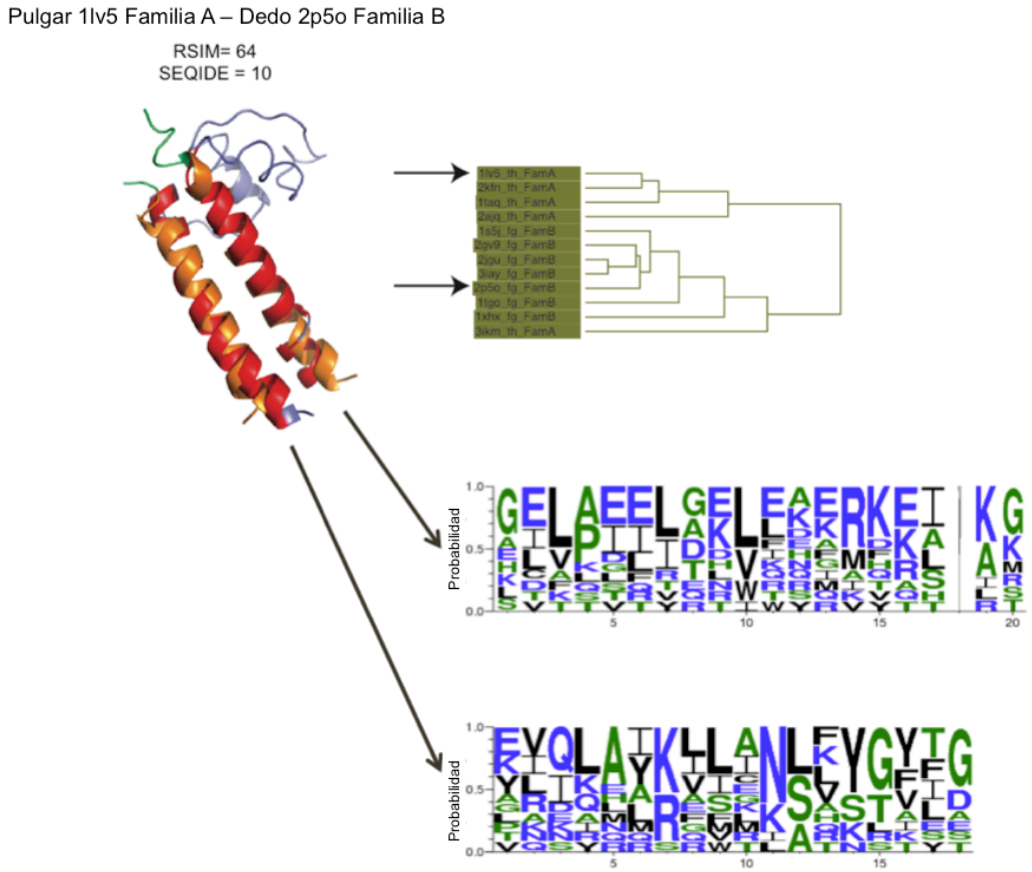


Figura 26. Ejemplo de las relaciones de similitud estructural entre dedos y pulgares de polimerasas de ADN en las familias A y B. Arriba y al costado izquierdo de la figura se muestra una superposición del pulgar de la polimerasa de *Geobacillus sthearothermophilus* (estructura de color azul, carbonos equivalentes en rojo) y el dedo de la polimerasa del fago RB69 (estructura de color verde, carbonos equivalentes en color naranjo). Los valores de similitud estructural e identidad de secuencia del alineamiento estructural según el cálculo mediante STOVCA se encuentran reportados sobre la superposición óptima. En la derecha, se presenta un extracto del dendrograma de la Figura 25. Se indican con flechas negras la posición de las estructuras comparadas en esta figura. Finalmente abajo y a la derecha se encuentran los logos de secuencia derivados a partir del alineamiento estructural de los pulgares de la familia A y dedos de la familia B. Se generó un logo de secuencia para cada una de las hélices que se superponen (en el caso de la figura, pares de hélices). Las flechas indican a que segmento estructural corresponde cada logo de secuencia.

La segunda relación cruzada encontrada en el análisis de la Figura 25, , corresponde a la similitud estructural entre dedos de la familia X y pulgares de la familia Y. Los dedos de la familia X están formados por una estructura compuesta fundamentalmente por cuatro hélices alfa organizadas en dos pares antiparalelos, que se disponen con una rotación de 90º, formando en conjunto una estructura de apariencia cuboidal. Cada hélice está compuesta por 10 a 13 aminoácidos. Todos estos segmentos se encuentran conectados por *loops* pequeños. Los pulgares de la familia Y también poseen cuatro hélices, sin embargo poseen algunas variaciones, por ejemplo la hélices 1 y 3 (numeradas del extremo N- a C-terminal del subdominio) son más pequeñas (5 a 6 aminoácidos). Por otra parte la conexión entre la hélice 2 y 3 ocurre a través de un segmento que tiene una pequeña hélice alfa también (de 4 a 5 aminoácidos).

El agrupamiento establecido en el dendrograma sugiere la existencia de una relación de similitud estructural entre ambos tipos de dominios de las diferentes familias. En promedio se observan similitudes relativas superiores al 60. Las identidades de secuencia calculadas para estas superposiciones óptimas pueden superar el 20%. A diferencia de lo que ocurría en la comparación de los dedos de familia A y pulgar de familia B, en esta comparación cruzada de dominios y familias, fue posible identificar algunos residuos conservados a partir de la información de los alineamientos estructurales. La región N-terminal de ambos dominios se superpone óptimamente en una longitud de 31 aminoácidos. En primer lugar, existen dos glicinas altamente conservadas que conectan las dos primeras hélices de estos dominios. Entre ellas se encuentra preferentemente isoleucina, aunque no es tan conservada. De todas maneras el patrón de sustitución de este aminoácido es siempre por

otro residuo de tipo no polar. Posteriormente residuos de lisina, ácido aspártico y un residuo de carga positiva se encuentran conservados en las tercera hélice de ambos dominios (Figura 27).

De acuerdo a datos recopilados desde la PDIdb (*Protein DNA Interface database*), muestran que tanto en los dedos de la familia X como en los pulgares de la familia Y, los residuos de glicina se encuentran realizando contactos con el esqueleto fosfato del ADN, en un modo de unión similar.

Comparados desde la perspectiva de su organización en la estructura primaria, ambos dominios están compuestos por segmentos continuos en la secuencia aminoacídica. Aprovechando este tipo se similitud, se construyó para cada uno de ellos alineamientos de múltiples estructuras, que posteriormente fueron complementados con secuencias provenientes de bases de datos de acuerdo a la metodología descrita para esta tesis. Finalmente se transformaron a modelos de Markov (según lo descrito en el punto 4.6). También se construyeron alineamientos de múltiples secuencias de acuerdo a lo descrito en el punto 4.7.

Cuando se realizó la búsqueda en la base de datos Swissprot empleando el perfil *Fingers-FamX-MSA*, se recuperaron solamente dedos de polimerasas de la familia X. La misma situación ocurrió cuando se empleó el perfil *Thumbs-FamY-MSA*, el cual sólo recuperó a polimerasas de la familia Y. Una situación diferente se produjo al emplear perfiles derivados de alineamientos estructurales. El perfil *Fingers-FamX-MStA*, recuperó dominios pulgares de polimerasas de la familia Y. A modo de ejemplo el de la polimerasa

IV de *Saccharomyces cerevisiae* fueron recuperados con un *e-value* de $8.3e^{-5}$, lo que es considerado como un resultado significativo de acuerdo a los criterios de corte establecidos por defecto en el programa HMMer. La búsqueda realizada con el perfil *Thumbs-FamY-MStA*, logró recuperar dominios dedos de polimerasas de la familia X. Como ejemplo los dedos de la polimerasa beta de *Rattus norvegicus* fueron recuperados con un e-value de $9.4e^{-5}$, por lo que, al igual que en el caso anterior, corresponde a un resultado significativo. Esto actúa como evidencia adicional para relacionar estos dos dominios (además de la evidencia estructural presentada).

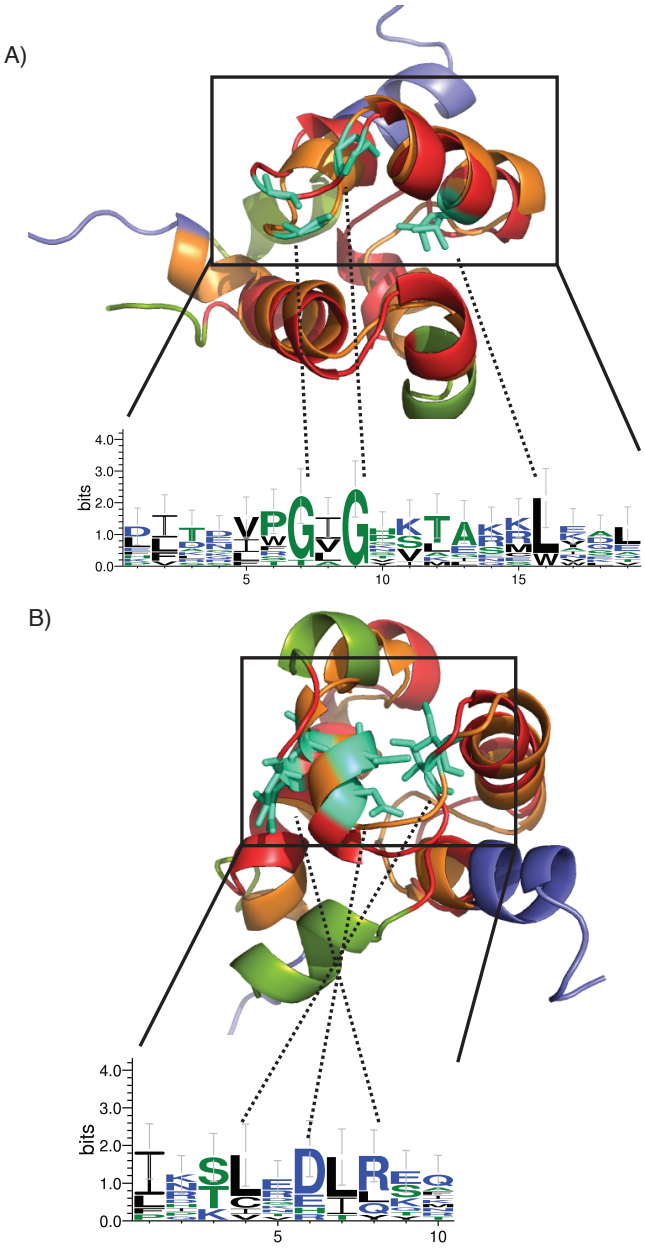


Figura 27. Patrones de conservación de secuencia en alineamientos estructurales de dedos de familia X y pulgares de familia Y. En la figura se muestran patrones de conservación derivados de alineamientos estructurales de dominios dedos de familia X y pulgares de familia Y. Las estructuras superpuestas corresponden a representantes de cada tipo de subdominio y familia (dedos familia X: código PDB 1huo en color azul, pulgar familia Y: código PDB 1k1s, color verde). Los aminoácidos estructuralmente equivalentes se muestran en rojo para la estructura azul y en naranjo para la estructura verde. En cada superposición se destaca una zona con un cuadro negro, de la cual se construyeron logos de secuencia (bajo cada superposición). Los aminoácidos conservados se destacan con sus cadenas laterales y color cyan en cada superposición y están ligados al logo de secuencia mediante líneas punteadas.

5.9. Mapa de relaciones.

Las polimerasas de ADN además de tener un dominio catalítico que es el fundamental para establecer la función que desempeñan, poseen dominios llamados accesorios en la literatura. Estos dominios accesorios tienen funciones diversas. Adicionalmente se ha demostrado en otras publicaciones que la eliminación de estos dominios accesorios no afecta la actividad de polimerización de la enzima (H Klenow, 1970). No obstante, para comprender de manera más sistémica las relaciones entre las distintas familias de polimerasas de ADN, se utilizó una poderosa herramienta de análisis estructural llamada TopSearch. Esta herramienta única a nivel mundial permite explorar el espacio estructural de manera simple y rápida. Las búsquedas en TopSearch y la construcción de relaciones se llevaron utilizando las cadenas completas de polimerasas de ADN presentes en la base de datos creada en esta tesis. Los resultados de las relaciones se muestran en detalle en esta sección.

Una relación compartida entre la familia A y B ocurre con los dominios 3'-Exonucleasa. Esta relación también es compartida con polimerasas de la familia C (Figura 28). Ambas polimerasas poseen este dominio accesorio, el cual se encuentra altamente conservado. Por otra parte las polimerasas de las familias A e Y se relacionan con las ciclasas a través del subdominio palma. Las polimerasas de la familia A, B e Y se relacionan entre ellas a través del dominio palma. Es importante destacar, que las polimerasas de la familia B no recuperan relaciones estructurales con las ciclasas. Por último cabe mencionar una relación que es exclusiva para miembros de la familia A con proteínas del tipo 5'-Endo-Exo-Ribonucleasas 5' (Figura 29).

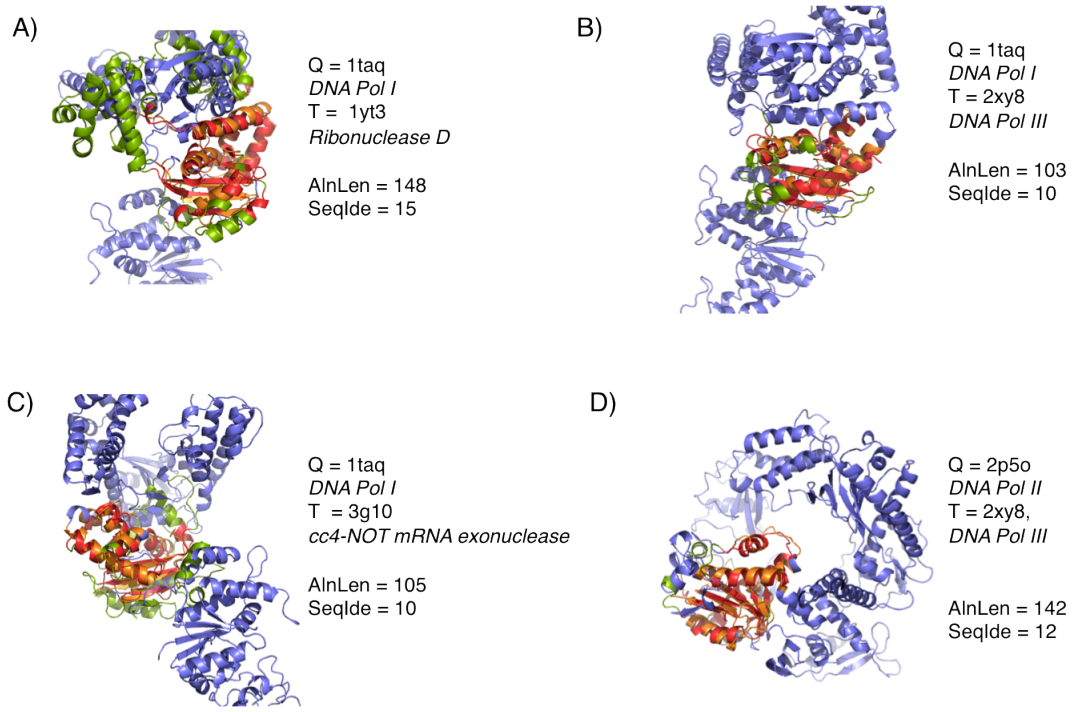


Figura 28. Relaciones de similitud estructural de polimerasas con dominios 3'-5' Exonucleasa. En la figura se muestran ejemplos de relaciones entre cadenas de polimerasas de ADN a través de sus dominios 3'-5' exonucleasa con otras proteínas que contienen este dominio. El esquema de colores utilizados en las superposiciones óptimas es el mismo que se ha empleado en otras figuras de esta tesis. A) Superposición óptima del dominio 3'-5' exonucleasa de la ADN polimerasa I de *Thermus aquaticus* con la Ribonucleasa D de *Escherichia coli*. B) Superposición óptima del dominio 3'-5' exonucleasa de la ADN polimerasa I de *Thermus aquaticus* con el dominio 3'-5'exonucleasa de la ADN polimerasa III de *Escherichia coli*. C) Superposición óptima del dominio 3'-5' exonucleasa de la ADN polimerasa I de *Thermus aquaticus* con la exonucleasa de ARN cc4-NOT de *Saccharomyces cerevisiae*. D) Superposición óptima del dominio 3'-5' exonucleasa de la ADN polimerasa II del bacteriófago RB69 con el dominio 3'-5'exonucleasa de la ADN polimerasa III de *Escherichia coli*

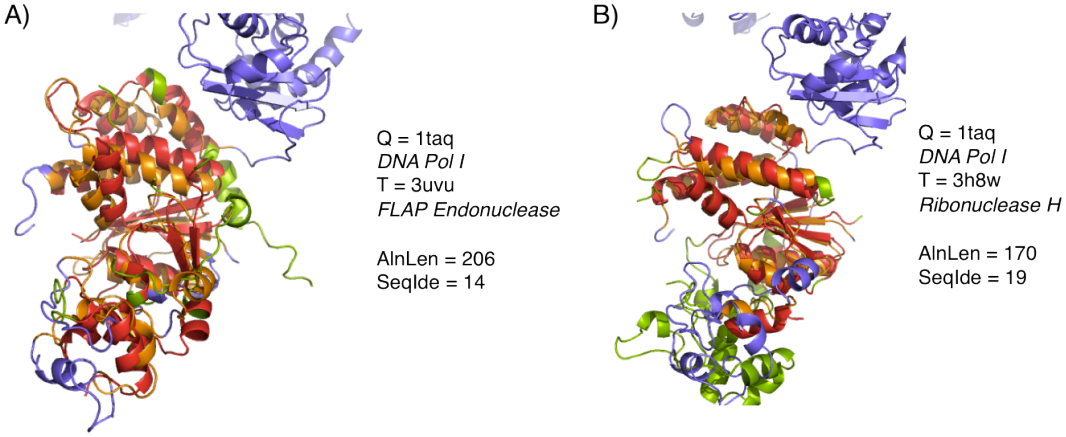


Figura 29. Relaciones de similitud estructural entre polimerasas de ADN y dominios 5'-3' Exo-Endo-Ribonucleasa. En la figura se muestran ejemplos de las relaciones estrucrurales que se dan entre los dominios 5'-3' exonucleasa de polimerasas de ADN y otras proteínas que contienen dominios similares. A) Superposición óptima del dominio 5'-3' exonucleasa de la ADN polimerasa I de *Thermus aquaticus* con la Endonucleada FLAP de *Homo sapiens*. B) Superposición óptima entre el dominio 5'-3' exonucleasa de la ADN polimerasa I de *Thermus aquaticus* con la Ribonucleasa H del bacteriófago T4.

Las polimerasas de la familias C y X se conectan a través de su relación con los dominios PHP, los cuales están presentes de manera exclusiva en estas dos familias.

En el caso particular de las polimerasas de la familia X se relacionan con un número importante de nucleotidiltransferasas, a partir del subdominio de la palma (Figura 30). Cuando se utilizan estas nucleotidiltransferasas recuperadas como *query* en el programa TopSearch se recuperan únicamente palmas de polimerasas de la familia X, y no de otras familias, haciendo exclusiva esta relación.

Un aspecto estructural particular de las polimerasas de la familia Y, es la presencia de un dominio accesorio llamado PAD, (polymerase accesory domain). Las búsquedas estructurales con el dominio PAD como *query* en *TopSearch*, no recuperan ninguna otra polimerasa que no sea de la familia Y.

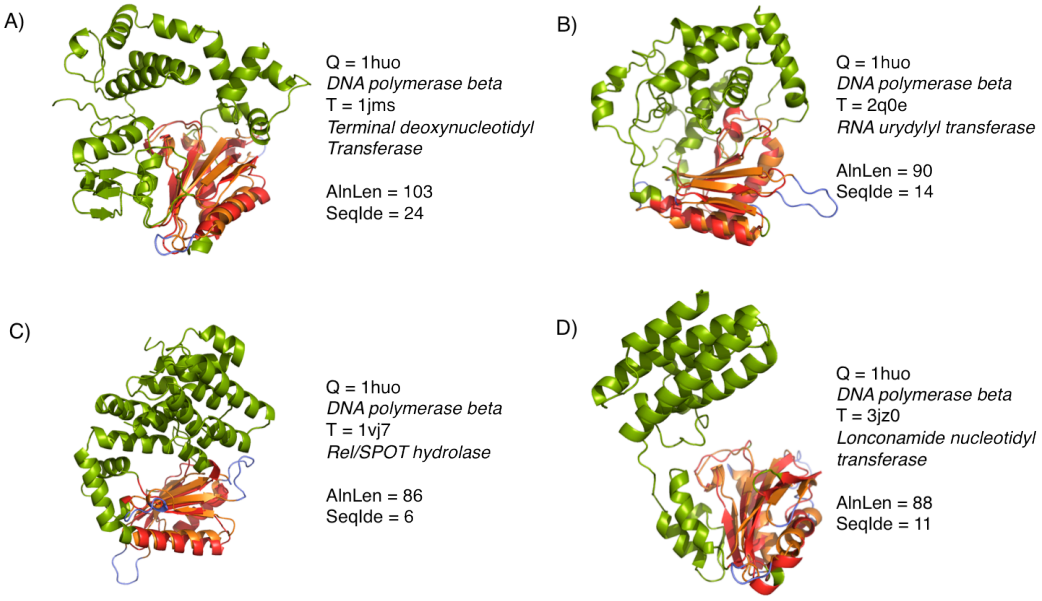


Figura 30. Relaciones de similitud de la palma de polimerasas de la familia X. En la figura se muestran ejemplos de las relaciones de similitud recuperadas con palmas de polimerasas de la familia X. Cada superposición está construida superponiendo la palma de la polimerasa beta de *Rattus norvegicus* contra alguno de los hits obtenidos mediante *TopSearch*. En cada caso se presentan los valores de la longitud del alineamiento y la identidad de secuencia que se obtiene en el alineamiento estructural.

Finalmente a modo de síntesis de las relaciones identificadas por diferentes métodos en este trabajo, se creó un mapa de relaciones para las polimerasas de ADN. En este mapa se muestran conexiones que existen entre familias de polimerasas de ADN a través de la similitud de dominios estructurales que son parte del dominio catalítico, así como por la presencia y ausencia de dominios accesorios y la similitud estructural presente en ellos. El mapa es consistente con algunas relaciones presentadas previamente a lo largo del análisis de esta tesis. En primer lugar, se distingue al grupo ABY, que se encuentra formado por polimerasas de ADN de las familias A, B e Y. Estas conforman un grupo muy relacionado, fundamentalmente a partir de la conservación identificada en el dominio de las palmas. Esta relación fue definida como de carácter fuerte, pues la palma es un dominio central en la función de la polimerización al contener el aparato catalítico responsable de la polimerización de ADN. Dentro de este grupo es posible observar otras relaciones que conectan a las polimerasas de este grupo ABY. Las polimerasas de la familia A e Y poseen un dominio palma que comparte alta similitud con proteínas del tipo ciclasas que también tienen una función en la modificación de nucleótidos. Esta relación no se encuentra de manera directa para las polimerasas de la familia B. Las polimerasas de la familia A y B se relacionan entre ellas por la presencia en ambas de un dominio 3'-5' exonucleasa, que también se observa en polimerasas de la familia C. Al analizar la similitud estructural de este dominio, se aprecia que altamente conservado, de manera tal que cuando se comparan dominio 3'-5' de polimerasas de las familias A, B y C, la similitud estructural promedio es de un 70%, sin embargo la identidad de secuencia es tan baja como un 15% en promedio. En este grupo también se observan conexiones que son exclusivas de una familia. Por ejemplo en el caso de la familia A, se observa la presencia de un dominio 5'-3' exonucleasa

que tiene alta similitud estrctural con la superfamilia de dominios 5'-endo-exo-ribonucleasa. Por otra parte, en la familia Y se observa la presencia de un dominio PAD (*polymerase associated domain*). Las búsquedas estructurales mediante *TopSearch*, no entrega resultados de similitud relevante con algún otro dominio presente en el PDB. Finalmente una relación de carácter débil que existe entre la familia A y B, es la que se observa entre los pulgares de la familia A y los dedos de la familia B. Si bien se mide una similitud estructural importante entre estos dominios, se considera que esta relación es débil a partir de la evidencia mostrada en puntos anteriores.

Las familias X y C muestran un número mucho menor de conexiones, y aparecen como grupos separados e independientes del grupo ABY. Las familias C y X aparecen relacionadas por la presencia de un dominio PHP de alta similitud estructural entre ambas familias, y con otros dominios PHP presentes en otras familias. A partir de las búsquedas mediante la herramienta TopSearch, se detectó una relación muy débil entre las palmas de la familia X e Y. Si se utiliza como dominio query una palma de la familia X, es posible recuperar palmas de la familia Y. Si se realiza el ejercicio inverso, también se recuperan palmas de la otra familia. No obstante este resultado, se califica como débil, pues el valor de similitud estructural reportado en estos alineamientos es menor que el que se observa entre palmas del grupo ABY (70% promedio comparado con un 45%). Las palmas de la familia X presentan una alta similitud con aquellas proteínas clasificadas en la superfamilia de beta-nucleotidiltransferasas, donde se encuentran proteínas como Poli-A-ARN polimerasas y nucleotidiltransferasas de ARN. Las palmas de la familia C no presentan relación con ninguna otra proteína relacionada a funciones que involucren modificación o

interacción con nucleótidos. Finalmente existe una relación calificada como débil entre dedos de la familia X y pulgares de la familia Y que fue presentada en puntos anteriores. Las relaciones descritas anteriormente se resumen en la Figura 31.

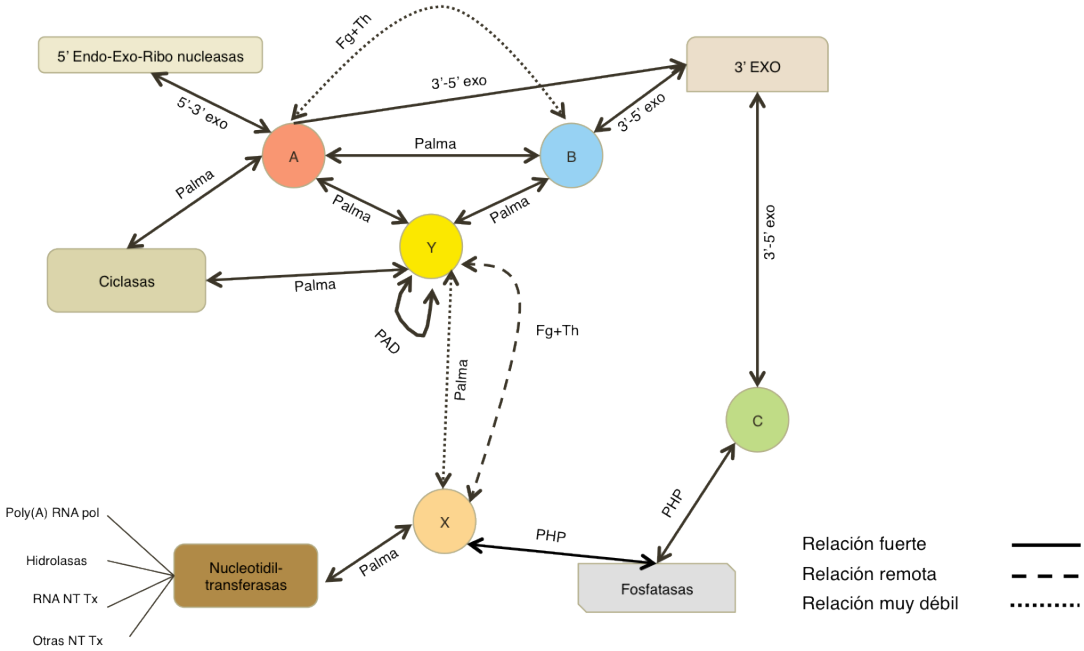


Figura 31. Mapa de relaciones de polimerasas de ADN. La siguiente figura presenta un esquema de las relaciones que se observan entre polimerasas de ADN. Con líneas gruesas se conectan elementos que tienen un soporte estructural fuerte (*Relative similarity > 50)*. Una segunda categoría de relaciones cuando denominadas remotas que ocurren con similitudes relativas entre 40 y 50. Finalmente relaciones de tipo muy débil donde la similitud relativa se encuentre entre 30 y 40. Sobre la línea de conexión se indica el dominio que genera la relación.

6. Discusión

6.1. Conjunto de datos para la validación de STOVCA.

El conjunto de datos obtenido para la validación de STOVCA si bien no es un conjunto plenamente representativo de la diversidad estructural existente en el PDB, representa de buena manera a una variedad importante de pliegues y arquitecturas conocidas. Por lo demás, el propósito de este filtro de datos fue generar un conjunto de alta diversidad estructural y de secuencia para la validación de diferentes programas de alineamiento estructural. En la literatura los artículos publicados en el área de la comparación de estructuras de proteínas, usualmente toman a la base de datos HOMSTRAD como un estándar de comparación, sin embargo no consideran la posibilidad de redundancia de secuencia y estructura contenida en ella. En este caso el filtro de exigencia aplicado permitió eliminar este sesgo, de manera tal de indagar en las reales diferencias que se pueden encontrar entre los programas de alineamiento estructural comparados. Este conjunto de datos finalmente ha quedado disponible a la comunidad científica, para que desarrolladores de programas de alineamiento estructural puedan ocuparlo en posteriores análisis del rendimiento de sus programas.

6.2. Estandarización de alineamientos estructurales con STOVCA.

Tal como se demostró previamente, la construcción de alineamientos estructurales tiene dos partes independientes: transformaciones geométricas para superposiciones óptimas y la

lectura del alineamiento asociado a la secuencia a partir de la superposición óptima. La transformación captura el desempeño de un programa determinado de alineamiento estructural, mientras que la derivación del alineamiento puede ser estandarizada. De esta forma es posible mapear la transformación geométrica a un número específico, llamado longitud del alineamiento, que puede ser utilizado para determinar el desempeño de varios métodos de alineamiento estructural sobre un par de estructuras proteicas. A partir de esta aproximación, se mostró que varios programas de alineamiento estructural tienen desempeños diferenciales.

Los programas utilizados se pueden dividir en dos grandes grupos, donde uno supera al otro en desempeño. Pese a esta división, se observa que ningún programa se desempeña de manera consistente como el mejor. Por lo tanto, dependiendo de la aplicación, puede ser ventajoso utilizar varios programas de alineamiento estructural en paralelo para seleccionar la mejor solución a partir del conjunto de alternativas reportadas y estandarizadas.

El programa STOVCA, se utiliza únicamente para calcular alineamientos de secuencia de máxima longitud derivados de información de superposiciones de estructuras previamente calculadas con otros programas. Los alineamientos de STOVCA son óptimos y dependen de parámetros que pueden ser modificados por el usuario. Finalmente, dado ese conjunto de parámetros fijados por el usuario, los alineamientos que se obtengan serán estandarizados y comparables entre ellos.

La comparación de programas fue realizada con dominios simples de tamaño similar,

esto corresponde al escenario más sencillo que se encuentra en la comparación de estructuras de proteínas. En otros casos, la situación es más compleja, puesto que las proteínas consisten frecuentemente de varios dominios y repeticiones. Es posible que se requieran permutaciones en la secuencia y además varios alineamientos alternativos para describir de manera completa la similitud estructural entre dos proteínas. Este tipo de situaciones no fueron tratadas con STOVCA ya que sólo unos pocos programas en la actualidad tienen la capacidad de realizar estas tareas (Sippl & Wiederstein, 2012)

En esta evaluación, tampoco fueron incluidos programas de alineamiento estructural conocidos como alineadores flexibles. Estos programas entregan como resultado una serie de alineamientos cortos y de bajo RMSD, por lo que no sería adecuado evaluarlos con el criterio adoptado acá que busca maximizar la longitud del alineamiento. También es necesario mencionar que no todos los programas mencionados acá necesariamente maximizan la longitud del alineamiento como función objetivo al momento de realizar un alineamiento estructural, lo que podría explicar en parte las diferencias de desempeño registradas mediante la evaluación con STOVCA. Claramente el propósito de STOVCA no es comparar y ordenar por un criterio del mejor al peor software, sino que enfatizar que casos complejos como los incluidos en la evaluación pueden dar resultados que difieren de manera sustancial. Una manera razonable de aproximarse a esto último es generar todos los alineamientos con diferentes programas y compararlos de manera estandarizada.

6.3. Base de datos de polimerasas de ADN.

La base de datos de estructuras de polimerasas de ADN construida para esta tesis constituye una recopilación de las estructuras disponibles en el PDB, y cuenta con un alto grado de curación manual de sus entradas, por lo tanto constituye un importante recurso para el estudio de esta clase de enzimas. Esta base de datos se caracteriza por representar de manera adecuada la diversidad de polimerasas de ADN en cuanto a su estructura, organismo de procedencia, función y familia. Durante el desarrollo de esta tesis, la base de datos PDB ha tenido notables avances en sus sistemas de consulta de datos. En la actualidad es posible crear consultas personalizadas que se mantienen en ejecución permanente y notifican a los usuarios de actualizaciones relevantes cuando corresponda. En el caso particular de esta tesis, la consulta ha sido actualizada mensualmente, y en el transcurso no se ha depositado ninguna estructura nueva que deba ser incluida como nuevo representante en esta base de datos. Por ello, los análisis presentados en este escrito siguen siendo vigentes.

6.4. Alineamientos estructurales de cadenas completas de polimerasas de ADN.

Las relaciones de similitud estructural exploradas a nivel de cadenas completas de polimerasas de ADN, mostraron que la división actual de estas polimerasas en las familias propuestas es válida. Sin embargo, quedó de manifiesto que en general existe un importante grado de diversidad estructural dentro de cada familia. En algunos casos se observan ejemplos extremos como lo que ocurre en las Familias A y B. En éstas es posible encontrar

miembros que se alejan considerablemente del promedio observado en la familia (polimerasa Gamma código 3ikm y polimerasa Phi29 código 1xhx). La interpretación de estos resultados fue hecha de acuerdo al esquema de clasificación de proteínas propuesto por Sippl en su base de datos COPS (Suhrer et al., 2009). En el artículo de Sippl se definen tres umbrales de clasificación de proteínas que permiten establecer jerarquías de similitud tal como existen en otras bases de datos de clasificación de proteínas (por ejemplo FSSP y CATH). La novedad de este criterio está en que fue definido a partir de la variación estructural observada en proteínas. De acuerdo a esto, todas aquellas ramificaciones que ocurren bajo el umbral definido como *Distante,* dan origen a grupos de proteínas que se relacionan estructuralmente. En adelante se definen dos umbrales adicionales que dan cuenta de la importancia de esta relación. Los casos presentados de las familias A y B, son casos límite, pues a nivel de cadena completa están casi en el umbral *Distante*, por tanto se consideran como miembros altamente divergentes. Una explicación para los resultados obtenidos puede encontrarse en el hecho que las polimerasas de ADN presentan una serie de dominios accesorios en la cadena completa, por lo que la alta divergencia podría explicarse por la variabilidad de estos dominios y no por la variabilidad del dominio catalítico. Esta posibilidad será discutida más adelante conforme se profundice en el estudio a nivel de los dominios catalíticos. Una segunda posibilidad es considerar la variación producto de la desviación estructural a consecuencia de movimientos propios de la estructura como resultado de las condiciones de cristalización. En efecto algunas estructuras fueron cristalizadas en presencia o ausencia de ADN, lo que genera movimiento de algunos segmentos. Sin embargo este efecto está corregido por la forma en que el algoritmo de STOVCA calcula la similitud estructural, pues se intenta maximizar la

cantidad de residuos estructuralmente equivalentes a un valor de desviación estructural constante. De esa manera es el usuario quien decide qué grado de variación estructural se desea tolerar. La comparación de cadenas completas, permitió la identificación de estas variaciones a nivel global, que finalmente son útiles para profundizar en la comprensión de las relaciones estructurales de este tipo de proteínas.

6.5. Alineamientos a nivel del dominio catalítico.

El análisis de la similitud del dominio catalítico (dominio funcional formado por palma, pulgar y dedos), también mostró consistencia con la clasificación actualmente aceptada. En este caso se esperaba que el dominio catalítico presentase un alto grado de diversidad estructural, sin embargo se observa un comportamiento similar al comparar a nivel de cadenas completas. Lo anterior demuestra que el dominio catalítico también contiene un grado importante de variación estructural, por lo que las diferentes familias de polimerasas de ADN no constituyen un grupo de alta conservación estructural.

En el caso de las familias A y B en términos generales se repite el patrón de divergencia obtenido a nivel de cadenas completas. De manera más particular es posible observar que la variación del dominio catalítico de la familia A es muy acentuada, ya que el dominio catalítico de la polimerasa Gamma se distancia mucho de los otros miembros de la familia. En el caso de la familia A se identificaron elementos estructurales concretos que dan origen a esta diversidad, como es una inserción en el dominio del pulgar. Este segmento estructural no pudo ser asociado a ninguna otra estructura conocida en el PDB, lo que

sugiere que se trata de un nuevo pliegue. La familia B, también es una familia muy diversa. Como caso especial tenemos la estructura del fago Phi29, el que se caracteriza por tener un pulgar con una estructura diferente al pulgar típicamente observado en otros miembros de la familia B. La estructura de este pulgar tiene similitud con los *clamps* de ADN que actúan como factores de procesividad en la polimerización. Sin embargo esta similitud debe analizarse con cautela, pues si se toma este segmento específico y se buscan similitudes estructurales en otras proteínas del PDB es posible encontrar una gran cantidad de proteínas que tienen arreglos estructurales similares.

6.6. Relaciones estructurales de los dominios palma, pulgares y dedos.

El análisis del dominio de las palmas mostró que existe una relación estructural entre esta clase dominios en las familias A, B e Y. De acuerdo a lo analizado, forman un grupo común bajo el umbral de corte *Remoto*. Este resultado sugiere que estos dominios pueden ser considerados homólogos estructurales. Lo anterior se apoya en varias fuentes de evidencias. Por una parte, cuando se exploran alineamientos estructurales entre palmas de las familias A, B e Y, se observan longitudes de alineamiento que en promedio son de 70 aminoácidos. Cuando se expresa como un valor de superposición estructural es superior al 50%. Además del valor numérico, se aprecia una conservación de la arquitectura y topología de estos dominios, ya que en las tres familias se encuentran constituidos por cuatro hebras betas y dos hélices alfa. En la superposición de estos segmentos también coincide la posición de los aminoácidos del aparato catalítico. Por último, la relación entre las palmas de estas tres familias no es plenamente simétrica, siendo más similares las

palmas de las familias A e Y. Lo anterior se desprende no sólo del valor de similitud estructural que es mayor, sino también de algunas firmas de secuencias en torno al aparato catalítico.

Las evidencias encontradas anteriormente sugieren que las palmas de las polimerasas de las familias A, B e Y podrían haber evolucionado de un ancestro común con alguna clase de actividad nucleotidiltransferasa. La hipótesis se basa en el hecho que si se toman como sondas a representantes de las palmas de las familias A, B e Y para realizar búsquedas mediante la herramienta *TopSearch*, se obtiene un número importante de nucleotidiltransferasas de diverso tipo, siendo las más destacadas las ciclasas. Sin embargo es importante recalcar que esta similitud se encuentra contenida en lo que se puede definir como un núcleo de la palma que en total conforma una estructura de entre 60 a 80 aminoácidos con dos elementos de arquitectura altamente conservados: un techo constituido por 4 hebras beta antiparalelas y una base de dos hélices alfa. Este patrón también se observó en otras proteínas que no guardan relación funcional con polimerasas de ADN o transformación química de nucleótidos (ej: carboxipeptidasas), aunque la similitud estructural presentó valores más bajos que los registrados con proteínas que interactúan con algún tipo de nucleótidos. Lo anterior sugiere que este arreglo arquitectónico podría haber sido seleccionado en la evolución debido a su estabilidad termodinámica. En el caso de polimerasas de ADN una característica es la presencia de la tríada catalítica constituida por aminoácidos del tipo ácido aspártico que participan en la coordinación de un ión bivalente como el magnesio para el proceso de polimerización.

Los datos analizados mostraron la imposibilidad de establecer una conexión entre las palmas del grupo ABY con las de las familias C o X. Si bien es posible superponer parte de estas palmas, los segmentos superpuestos no tienen ninguna relación funcional conservada, lo que apoya la idea que las palmas del grupo ABY no son homólogas con las de la familias C o X. En efecto, las polimerasas de la familia X se relacionan con un grupo diferente de proteínas del que se relacionan las familias A, B e Y. Además, el hecho que la arquitectura y topología sean diferentes también sugiere que no se encuentran relacionadas. Lo mismo es válido para la familia C, para la que además existe un argumento que se apoya en su distribución filogenética, pues se encuentran únicamente en el dominio de las bacterias.

Por otra parte, el análisis del dominio de los dedos, mostró que la clasificación actual para familias de polimerasas de ADN no se cumple. Acá se identificaron casos importantes de variación estructural como ocurrió con la familia A, donde se observaron diferencias en la arquitectura y organización del dominio de los dedos. Para el caso particular de la polimerasa Gamma, los dedos guardan similitud en una porción constituida por dos hélices alfa antiparalelas que forman la estructura principal del dedo. Sin embargo carece de un segmento característico que es reemplazado por la inserción de un segmento de 310 aminoácidos que de acuerdo a las búsquedas realizadas no presenta similitud significativa con otras estructuras disponibles en el PDB, sugiriendo que se trataría de un pliegue nuevo. Alternativamente, también es posible pensar que éste se trate de un segmento que en las condiciones de cristalización de la proteína no haya adquirido su estructura más estable, pues al analizar los datos obtenidos a partir del cristal, se observa poca estructuración secundaria. De acuerdo a esto, se requeriría observar cristales en otras condiciones para

verificar si este segmento permanece estructurado de la misma manera. Por último, es importante destacar que a nivel de secuencia las búsquedas tradicionales arrojan sólo polimerasas Gamma como resultado, lo que apoya la primera explicación.

Al igual que en el caso del dominio de los dedos, el dominio del pulgar no respeta la clasificación actualmente propuesta para polimerasas de ADN. Este dominio es el más variable a nivel estructural de los analizados. En este caso la variación tiene varias fuentes de origen. Una de ellas es el cambio de la composición de elementos de estructura secundaria, tal como se observó en el pulgar de la polimerasa Phi29 en la familia B. En la familia A, también se observó un grado importante de variación estructural explicada por pequeñas inserciones que ocurren fundamentalmente en regiones de los pulgares que quedan en la superficie de la proteína.

6.7. Las polimerasas de ADN constituyen una familia diversa a nivel estructural y de secuencia.

De acuerdo a lo expuesto en los resultados de esta sección, las familias A y B son las que presentan una mayor diversidad estructural y de secuencia. En el caso particular de la familia B, al contar con mayor número de miembros con estructura cristalizada, se hace más robusto el análisis estructural y las inferencias que se puedan obtener de éste. Uno de los aspectos más interesantes es lo referente a la situación de la polimerasa del fago Phi29, que presenta una variación estructural importante; el dominio pulgar es completamente distinto a los observados en los otros miembros de esta familia.

Adicionalmente desde el punto de vista de la secuencia, presenta una muy baja identidad (del orden del 19%), con lo que conforman un grupo lejano en relación a las otras polimerasas de esta familia. Estos resultados sugieren que las polimerasas virales similares a las del fago Phi29 podrían ser clasificadas en una nueva familia. Sin embargo, un análisis más fino y robusto como el presentado a partir de los alineamientos de múltiples estructuras, revela la conservación de algunos motivos de secuencia, que si bien son cortos, son característica propia de todas las polimerasas que pertenecen a la familia B. Si se toman en cuenta estas evidencias, se pone de manifiesto la complejidad del problema de la clasificación de proteínas. Conforme a los resultados obtenidos, los sistemas de clasificación jerárquicos, y que dividen a las proteínas en grupos discretos llamados familias, no es suficiente para describir de manera precisa la variación que se observa tanto a nivel de secuencia, estructura y función en proteínas. Por lo mismo, una posibilidad sería generar niveles adicionales de discriminación más finos en la clasificación. En este caso concreto la familia B podría ser dividida en dos clases, una de ellas llamada B_1 que contiene a todas las polimerasas canónicas de la familia B, representadas por la histórica polimerasa II de *E. coli*, y una segunda clase B_2 representada por la polimerasa del fago Phi29. Este tipo de situaciones puede ser relevante al momento de comprender como estas las proteínas han ido acumulando variación y generando nuevas estructuras que pueden ser en términos globales conservadas, pero que en algunos puntos locales presentan variaciones importantes, incluso en segmentos que habitualmente se presumen conservados pues están en directa relación con la función que ésta cumple. Esto ocurre en este caso concreto, pues en efecto el subdominio de los dedos es un dominio que tiene una participación en la

función de la polimerización acomodando en un ángulo adecuado el polímero de ADN que se produce.

En el caso de la familia A, la estructura de la polimerasa Gamma es un miembro altamente divergente dentro de su familia. A nivel de organización de sus dominios en la estructura primaria, se observó una inserción en el dominio del pulgar a la que no se le identificó similitud estructural con algún segmento de otra proteína presente en el PDB. Las búsquedas basadas en secuencia tampoco tuvieron éxito en buscar elementos similares en otras proteínas, por tanto es una característica propia de las polimerasas Gamma. A nivel de secuencia, también aparece como miembro distante dentro de la familia A. Esta polimerasa también presenta algunas inserciones de menor tamaño en el segmento de la palma, pero fundamentalmente en segmentos que se encuentran en la superficie de la proteína, por lo que no se les puede asociar un rol directo en modificar la función de ésta. Al igual que como sucede con la familia B, las clasificaciones jerárquicas no dan cuenta de manera precisa de estas diferencias. Los datos recopilados sugieren que en este caso también podría dividirse a la familia A en dos clases diferentes, la clase Pol I que contiene a todas aquellas polimerasas que son altamente similares a la polimerasa I de *E. coli* y el grupo Gamma que contiene a todos los ejemplares similares a la polimerasa Gamma. Este último grupo además comparte la característica funcional de ser responsables de la polimerización de los genomas mitocondriales.

Casos como los discutidos anteriormente, llevan a la idea que la clasificación actual

en familias para las polimerasas de ADN puede ser incompleta. Lo anterior se sustenta en al menos dos casos estudiados en detalle en la familias A y B. En ambos casos existe una distancia importante a nivel de secuencia y estructura. Además de los casos presentados anteriormente, que ocurren en familias de alta diversidad, también es posible encontrar otros en familias menos diversas como es el caso de la familia X. Si bien la variación estructural es más baja cuando se observan los valores de similitud estructural, existen también casos donde aparecen miembros lejanos al grupo canónico, como es el caso de la polimerasa ASFV que carece del dominio dedos. Aquí, la variación viene dada por la ausencia de parte del dominio catalítico.

Finalmente, en relación a la familia C, es posible mencionar que se trata de la familia que menores similitudes guarda con las otras familias de polimerasas de ADN. Las búsquedas de polimerasas de la familia C en otros arqueas, eucariontes y virus arrojó resultados negativos empleando perfiles construidos a partir de alineamientos estructurales. La búsqueda mediante perfiles de secuencia ha demostrado ser más sensible para la identificación de relaciones remotas entre secuencias. En este caso al ser construido a partir de alineamientos estructurales, se esperaba detectar alguna polimerasa de esta familia o fragmentos de ella en algún dominio diferente a Bacteria, considerando que el alineamiento estructural mejora la relación señal/ruido y corrige errores en alineamientos tal como se mostró en la sección de resultados. Una explicación para este resultado proviene de la naturaleza misma de los datos. Al complementar los alineamientos estructurales con nuevas secuencias, la diversidad resultante es baja. Esto limita la posibilidad de encontrar nuevas secuencias que son comparadas con un patrón

conservado. Todas las nuevas polimerasas que han sido anotadas como familia C (en bases de datos como *SwissProt*) comparten una elevada similitud con alguna de las tres estructuras que pertenecen a la base de datos. Adicionalmente si se realiza el complemento utilizando secuencias anotadas como miembros de la familia C, pero esta vez provenientes de la base de datos NR, se obtiene un resultado similar. Lo anterior entonces permite concluir que al menos en el dominio de las Bacterias la familia C es un grupo altamente conservado. Esto contrasta de manera importante con lo observado para otras familias. Si se toman grupos filogenéticos grandes y se analiza su diversidad, se observa una mayor variabilidad. En este sentido la familia C es un grupo particular de polimerasas. Otro aspecto que llama la atención, es que hasta la fecha tampoco se hayan descrito polimerasas de la familia C en virus, situación que ocurre para el resto de las familias con estructura disponible (recordar que la familia D, fue excluida de este estudio porque no existen estructuras depositadas en el PDB). Desde el punto de vista estructural estas polimerasas tienen una similitud remota con las polimerasas de la familia X a nivel del dominio de la palma. Sin embargo esta similitud es sólo al nivel de arquitectura, pues ambas se encuentran constituidas por 5 hebras beta que forman el techo de la palma (y esto las diferencia del grupo A-B-Y que tiene una arquitectura de 4 hebras beta en el techo). Sin embargo, en un análisis más fino de las superposiciones de ambas palmas, se ve que la posición de los ácidos aspárticos que forman la tríada catalítica tienen una posición diferente, luego la geometría del sitio catalítico es diferente. Lo anterior sugiere que el aparato catalítico de las polimerasas de las familias X y C no están relacionadas. Estas evidencias sugieren que la familia C tendría un origen independiente en la historia de la evolución.

6.8. Mapa de relaciones de polimerasas de ADN.

Tal como se demostró en los puntos referentes al análisis de la variación estructural y de secuencias, las polimerasas de ADN constituyen una familia de proteínas altamente diversa tanto a nivel de estructura, como a nivel de secuencias. Con una metodología de clasificación basada en algoritmos de agrupamiento del tipo jerárquico, se logró reproducir la clasificación actualmente aceptada para esta clase de familias a nivel de cadenas y dominios catalíticos, pero no en dominios estructurales como los dedos y pulgares.

Del análisis de los resultados obtenidos mediante agrupamientos jerárquicos, se desprende que por sí solos constituyen una herramienta útil para comprender las relaciones estructurales en proteínas, sin embargo poseen ciertas limitaciones. Estos sistemas asumen que el flujo de la información genética es siempre en línea vertical, con acumulación de variación a lo largo del tiempo, y no consideran la posibilidad de relaciones cruzadas o transferencias horizontales de genes, hecho que se ha demostrado que existe.

De lo anterior, surge la necesidad de expresar de una manera diferente la integración de este tipo de información, lo que se consiguió mediante la creación de un mapa de relaciones de polimerasas de ADN, donde cualquier clase de similitud estructural actúa como conector entre diferentes familias. De esta manera, se pueden cubrir todas las posibles conexiones que existan entre las diferentes familias.

De acuerdo a los datos recabados en esta investigación desde la perspectiva estructural es posible identificar 3 grandes grupos de polimerasas de ADN. Uno de ellos

relaciona fuertemente a las polimerasas de las familias A, B e Y a través del dominio palma. Los datos de similitud estructural sugieren una conexión evolutiva de estas tres familias. Estas relaciones de similitud no se observan cuando se realizan análisis a nivel de secuencia. Por ejemplo, al buscar mediante perfiles construidos únicamente con información proveniente de la secuencia, no se obtienen señales que los relacionen de manera concreta. Búsquedas con métodos más avanzados, tales como los alineamientos de perfiles, tampoco revelan conexión entre ambos. Si se repiten ambos ejercicios pero con perfiles derivados de alineamientos estructurales tampoco se obtienen resultados positivos. No obstante lo anterior, si es posible identificar ciertos aspectos específicos de la secuencia que analizados en el contexto estructural permiten validar las relaciones. Como ejemplo, encontramos a la posición de los aminoácidos del aparato catalítico, que se encuentran correctamente alineados cuando se comparan palmas de las familias A, B e Y. Además, en el caso particular de comparaciones entre palmas de la familias A e Y, se aprecian firmas de secuencia que comparten similitud entre sus secuencias. Un segundo argumento estructural para justificar la relación proviene del uso de la herramienta *TopSearch*. Al utilizar como sonda cadenas completas de polimerasas de la familias A, B e Y, se obtienen proteínas de las dos familias restantes. En estos casos, el punto que concentró la similitud corresponde al dominio de la palma. Un análisis más fino, utilizando sólo el dominio de la palma como sonda, reveló para el caso de las familias A e Y, un importante grado de similitud con proteínas del tipo ciclasas, situación que no se observa con palmas de la familia B. Por lo anterior, en estricto rigor la relación entre las familias A e Y es más fuerte. Adicionalmente, es importante considerar que la palma tiene un rol funcional fundamental para las polimerasas de ADN, dado su rol en la catálisis. En el grupo de polimerasas ABY,

también se pueden encontrar otra clase de relaciones fuera del dominio catalítico. Por ejemplo, las polimerasas de la familia A poseen un dominio 5'-3' exonucleasa que es característico de esta familia. Este dominio es estructuralmente conservado cuando es comparado con otras 5'-endo-exo-ribonucleasas. Por otra parte, las polimerasas de las familias A y B se relacionan por la presencia de un dominio 3'-5' exonucleasa que permite la corrección de errores durante la polimerización, incrementando la fidelidad de copiado. Adicionalmente, la familia C, también posee dominios 3'-5' exonucleasa, que aparecen relacionados a los presentes en las familias A y B, por lo que este es un punto de relación entre las tres familias. El dominio exonucleasa es altamente conservado a nivel estructural, cuando se comparan dominios 3'-'5 exonucleasa de las familias A, B y C, la similitud estructural es en promedio de un 75%, aunque sus identidades de secuencia tienen un 25% en promedio. Esta alta conservación no es consistente con la variación observada a nivel de los dominios catalíticos de polimerasas de ADN. En algunos casos la distancia estructural es muy alta (sobrepasa el nivel *Distante*, lo que indica que no están relacionados estructuralmente), de hecho al comparar algunas familias se encuentran arquitecturas diferentes. De manera específica, en las familias A y B, el dominio de las palmas se encuentra constituido por un techo de 4 hebras beta y un piso de 2 hélices alfa, mientras que en la familia C está constituido por un techo de 5 hebras beta y un piso de 2 hélices alfa. Pese a que es posible superponer arquitecturas de este tipo, la topología presentada es diferente por tanto no es factible encontrar una superposición razonable entre estos dos tipos de dominios palma. Si se toma en cuenta esta inconsistencia, es posible plantear como hipótesis que los dominios exonucleasa han evolucionado de manera independiente a las polimerasas de ADN. Sin embargo, es interesante la relación que se origina a nivel

funcional, pues pese a que los dominios catalíticos y los dominios exonucleasa puedan haber evolucionado independientemente, las polimerasas de las familias A, B y C comparten un rol funcional. Estas tres familias se encuentran involucradas en la replicación de los genomas de los tres dominios de la vida y a nivel viral.

Las familias A y B también se encuentran conectadas, aunque de manera muy débil, a través de la similitud encontrada entre los dedos de la familia A y los pulgares de la familia B. En la literatura se planteó la posibilidad que estos dominios correspondiesen a análogos estructurales (Hübscher, Maga, & Spadari, 2002). A partir de los agrupamientos jerárquicos, se logró reproducir esta relación cruzada entre estos dominios. No obstante, al utilizar como sonda cualquiera de estos segmentos en una búsqueda mediante *TopSearch*, revela que existe un número importante de estructuras donde es posible encontrar este arreglo estructural con valores de similitud similar a los reportados cuando se comparan los dedos de la familia A con pulgares de la familia B. Estos resultados sugieren que más que una relación funcional, es posible que este arreglo tenga una relación con la estabilidad termodinámica y plegamiento de la proteína. Estos segmentos en general presentan pocas interacciones con el ADN, y la gran mayoría de ellas son inespecíficas. En efecto, de los estudios realizados, la principal fuente de variación en las estructuras de polimerasas de ADN se encuentra a nivel del dominio de los dedos y pulgares. Los resultados obtenidos apoyan más la idea que ésta es no es una verdadera relación de homología o analogía estructural.

Las familias C y X forman grupos independientes en el mapa de las polimerasas de ADN. Sin embargo, en el mapa de relaciones estas familias aparecen conectadas porque ambas contienen dominios del tipo PHP, los cuales están involucrados en la degradación de los fosfatos liberados durante la incorporación de nuevos nucleótidos al ADN. Tal como ocurre con los dominios 3'-5'-exonucleasa, los dominios PHP son altamente conservados a nivel estructural y también aparecen en otras clases de proteínas no relacionadas a polimerización de ADN. Las polimerasas de las familias C y X también comparten algunos rasgos de similitud en la arquitectura de sus dominios palma. En ambas familias, se encuentran constituidos por un techo de 5 hebras beta y un piso de 2 hélices alfa. En particular para la familia X se identificó relación estructural con una serie de nucleotidiltransferasas tales como la Poly(A)-polimerasa, nucleotidiltransferasas de ARN e hidrolasas.

Las familias X e Y presentan una relación cruzada que ocurre entre los dedos de la familia X y los pulgares de la familia Y. Esta relación fue descubierta inicialmente a nivel estructural, donde ambos segmentos tienen similitudes estructurales de un promedio de 65%. Los segmentos alineados tienen una longitud de 110 aminoácidos en promedio. Dado que estos dominios son pequeños, cabe la posibilidad que la similitud encontrada no represente una relación estructural real, sino que se trate de una arquitectura común en el PDB. Lo anterior fue descartado mediante búsquedas utilizando como sonda uno u otro dominio. Como resultado siempre se obtienen ejemplares de la familia contraria, indicando que esta relación de similitud estructural es simétrica. Adicionalmente, otra clase de proteínas no relacionadas a polimerasas de ADN se recupera en la búsqueda con estas

sondas, sin embargo la similitud estructural obtenida es del orden del 40%, lo que sugiere que efectivamente existe una relación estructural entre estos dominios. Una segunda línea de evidencia corresponde a la encontrada a nivel de secuencia, puesto que luego de analizar los alineamientos estructurales se identificaron aminoácidos conservados en ambas estructuras. Al utilizar perfiles de secuencia basados en alineamientos estructurales se logró recuperar la relación obtenida a nivel estructural, mientras que utilizando perfiles basados únicamente en la información de secuencia, no se logró recuperar dicha relación. Lo anterior demuestra que utilizar alineamientos estructurales mejora la relación ruido/señal para la detección de relaciones remotas entre proteínas. De acuerdo a los datos obtenidos, estos segmentos se encuentran relacionados estructuralmente. Una posibilidad que surge al estudiar esta relación es que efectivamente estos dominios estén relacionados no sólo a nivel estructural, sino a nivel funcional, con lo que se espera que deberían tener el mismo rol en el dominio catalítico, y por tanto interacciones similares con el ADN. De ser así, esto cambiaría la perspectiva que se tiene acerca de la analogía propuesta por Steitz sobre el dominio catalítico de las polimerasas de ADN (Steitz, 1994). Este modelo, compara el dominio catalítico con la estructura de una mano derecha. Para evaluar esta posibilidad, se estudiaron en detalle los contactos entre estos dominios relacionados y el ADN. Este estudio pudo ser realizado en algunas de las estructuras de cada familia, ya que no todos los ejemplares fueron cristalizados con ADN. El análisis reveló que si bien los tipos de interacciones que tiene cada dominio con el ADN son similares, éstos interactúan con partes diferentes de esta molécula. El caso de los dedos tiene interacciones fundamentalmente a través del surco mayor del ADN, mientras que el pulgar lo hace por el surco menor. En ambos casos, los tipos de interacciones entre la proteína y el ADN son de

tipo inespecíficos entre la cadena principal de la proteína y el fosfato del esqueleto de la proteína. Por lo tanto, del análisis se concluye que estos dominios son homólogos estructurales, pero no homólogos funcionales.

En esta tesis se han estudiado las polimerasas de ADN tomando como principal fuente de información la comparación de sus estructuras. El uso de información combinada de secuencia y estructura permitió establecer relaciones de similitud entre polimerasas de ADN que no se pueden detectar mediante el uso de métodos tradicionales de comparación de secuencias. Lo anterior demuestra que esta aproximación incrementa la relación señal/ruido en comparaciones de familias de proteínas altamente divergentes como es el caso específico de las estudiadas en esta tesis. Las relaciones estudiadas fueron graficadas inicialmente mediante agrupamientos jerárquicos, sin embargo conforme se fueron explorando más relaciones, se estableció que este tipo de metodología tiene limitaciones que impiden explorar otra clase de relaciones más complejas entre proteínas. Si se quiere comprender de mejor manera el origen y evolución de familias complejas estructuralmente como es el caso de las polimerasas de ADN se requieren metodologías de agrupamiento más complejas como mapas o grafos en lugar de agrupamientos jerárquicos.

Por otra parte, los algoritmos y tecnologías de comparación de estructuras de proteínas se han diversificado de manera importante en los últimos años. A Marzo del 2013 existen publicados más de 100 programas de alineamiento estructural (datos no mostrados), sin embargo ninguno de ellos entrega una solución definitiva al problema, pues por definición la comparación estructural involucra dos parámetros sobre los que no se puede encontrar un

óptimo de manera simultánea. Cada programa ha intentado ofrecer la mejor solución posible en diferentes contextos de comparación y entregando la información del alineamiento estructural de diferentes maneras no necesariamente comparables. Por lo mismo, se requiere un sistema de comparación estandarizada de alineamientos estructurales. Esta comparación estandarizada además permite seleccionar de acuerdo a parámetros definidos por el usuario, escoger el mejor alineamiento estructural de un conjunto de alineamientos previamente calculados por diferentes programas. De esta manera también se puede maximizar la calidad de la información al escoger, por ejemplo, alineamientos estructurales con mayor cantidad de residuos alineados.

Los datos obtenidos de la comparación estandarizada fueron empleados para la construcción del mapa permitió comprender las relaciones estructurales de polimerasas de ADN desde una perspectiva más global. Si bien la cantidad de estructuras disponibles es poca comparada con la cantidad de secuencias de polimerasas disponibles, de todas formas es posible inferir algunas relaciones, pues de acuerdo al principio que relaciona la conservación de la estructura y secuencia en proteínas, estas estructuras representan un espacio de secuencias amplio, por lo que las inferencias son válidas para un número importante de secuencias de proteínas que no tienen una estructura cristalizada disponible en el PDB.

Los resultados obtenidos en esta tesis también establecen interrogantes acerca del origen y evolución de esta clase de proteínas. En la literatura actual, el problema de comprender la historia evolutiva de los sistemas que se encargan de copiar y mantener las moléculas de

ADN, ha sido abordada a partir de la comparación de secuencias únicamente. Una hipótesis acerca de este origen, es plantear que todas las polimerasas provengan de un ancestro común. Según esto, debe haber existido en la historia una polimerasa de ADN ancestral presente en el llamado LUCA (*last common universal ancestor*). De manera alternativa, se puede plantear como hipótesis que se hubiesen originado de manera independiente. La respuesta a esta interrogante es compleja, sin embargo el mapa de relaciones estructurales construido en esta tesis, permite recolectar información novedosa para resolver esta pregunta. Si la primera hipótesis fuese cierta, los datos estructurales deberían mostrar relaciones en algunos segmentos del dominio catalítico que provengan de diferentes familias de polimerasas de ADN. Lo anterior significa de manera concreta que en la comparación estructural de los dominios catalíticos, deberían obtenerse valores de distancia entre las estructuras menores al umbral *Distant*. En tales condiciones, las polimerasas de ADN formarían un único grupo relacionado estructuralmente. Se espera que lo anterior se cumpla de manera especial en el dominio de la palma, pues contiene los aminoácidos responsables de la catálisis. Los resultados obtenidos de la comparación a nivel de dominios catalíticos y del dominio de la palma sugieren que existen tres grupos principales de polimerasas de ADN: el grupo ABY (formado por polimerasas de las familias A, B e Y), el grupo de la Familia C y finalmente el grupo de la Familia X. De acuerdo a la comparación, estos grupos se diferencian en la arquitectura y topología de los dominios de la palma, lo que apoya la hipótesis que no están relacionados evolutivamente, y por tanto se habrían originado de manera independiente en la evolución. Además de lo anterior, el hecho que los ejemplares de la Familia C se encuentren únicamente en Bacterias, apoya desde el punto de vista filogenético a la hipótesis anterior. En la construcción del mapa

también se generaron relaciones entre estas familias no relacionadas, a partir de los dominios accesorios de polimerasas de ADN. Sin embargo, el patrón de conservación y la distribución filogenética sugiere que la historia evolutiva de estos dominios es independiente de la del dominio catalítico de las polimerasas de ADN. Finalmente, la familia de las polimerasas de ADN corresponde a una familia de proteínas altamente divergente a nivel de secuencias y divergente a nivel de estructuras. Al parecer en lo que se refiere a copiar ADN, se generaron diferentes estrategias en los sistemas biológicos en tiempos y grupos de organismos diferentes. Sin embargo, es importante considerar que al tratarse de un gen tan ancestral es difícil seguir la pista de los eventos que pueden haber ocurrido. Por ejemplo, aún a nivel estructural es complejo trazar eventos de transferencia horizontal de genes, en especial en relación con el rol que podrían haber tenido los virus en la evolución de esta familia de proteínas.

Finalmente la metodología empleada en el desarrollo de esta tesis puede ser ampliada al estudio de cualquier familia de proteínas que posean estructuras cristalizadas en la base de datos del PDB. Este tipo de estudio puede ayudar a comprender relaciones más complejas, tales como las descritas en esta tesis, dentro de familias de proteínas, así como entre familias de proteínas. Actualmente, la visión jerárquica de la clasificación de proteínas limita la comprensión de estas relaciones, por lo que una estrategia basada en mapas de relaciones complejas puede ayudar a comprenderlas de manera más acabada.

7. Conclusiones

i. Ningún programa de alineamiento estructural tiene un desempeño mejor que otros programas de alineamiento estructural de manera consistente.

ii. Se creó un algoritmo que permite estandarizar alineamientos estructurales de proteínas provenientes de diferentes programas que proveen superposiciones de estructuras.

iii. El análisis estandarizado de alineamientos estructurales provenientes de diferentes algoritmos permite la detección de errores finos y de gran importancia en alineamientos pareados de estructuras de proteínas.

iv. La utilización de información combinada de secuencia y estructura permitió establecer relaciones entre diferentes familias de polimerasas de ADN.

v. Los patrones de similitud estructural y de secuencia sugieren que la familia A puede dividirse en dos clases: A1 representada por la polimerasa I de *E.coli* y A2 representada por la polimerasa Gamma de *H.sapiens*

vi. Los patrones de similitud estructural y de secuencia sugieren la división de la familia B en dos clases: la clase B1 formada por polimerasas de ADN similares a la polimerasa II de *E.coli* y la clase B2 formada por polimerasas similares a la polimerasa del fago Phi29.

vii. Los dedos de la familia A y pulgares de la familia B no comparten una relación de homología, y corresponden a análogos de tipo estructural.

viii. Los dedos de la familia X y pulgares de la familia Y son homólogos estructurales pero no funcionales.

ix. Los perfiles de secuencia derivados de alineamientos estructurales mejoran la

sensibilidad en la detección de relaciones remotas en dominios estructurales de polimerasas de ADN.

x. Las polimerasas de ADN pueden ser clasificadas en 5 familias de acuerdo a la similitud estructural y de secuencia de sus cadenas proteicas.

xi. Se reconocen tres clases estructurales de dominios palma en polimerasas de ADN: los tipos ABY que agrupan a las palmas de familias A, B e Y, el tipo C que agrupa a las palmas de la familia C, y finalmente el tipo X que agrupa a las palmas de la familia X.

xii. De acuerdo al análisis de la relación secuencia/estructura del dominio catalítico y dominios estructurales de polimerasas de ADN, éstas se habrían originado en dos o más eventos independientes en la historia de la evolución.

xiii. Los dominios accesorios de polimerasas de ADN comparten una historia evolutiva independiente a la de los dominios catalíticos de esta clase de enzimas.

xiv. La clasificación jerárquica de estructuras de proteínas limita la posibilidad de descubrir y representar relaciones complejas en dominios catalíticos de polimerasas de ADN.

8. Referencias

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402.

Altschul, S. F., & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in biochemical sciences*, *23*(11), 444–447.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410.

Arndt, J. W., Gong, W., Zhong, X., Showalter, A. K., Liu, J., Dunlap, C. A., Lin, Z., et al. (2001). Insight into the catalytic mechanism of DNA polymerase beta: structures of intermediate complexes. *Biochemistry*, *40*(18), 5368–5375.

Bailey, S., Wing, R. A., & Steitz, T. A. (2006). The structure of T. aquaticus DNA polymerase III is distinct from eukaryotic replicative DNA polymerases. *Cell*, *126*(5), 893–904.

Bairoch, A. (2004). The Universal Protein Resource (UniProt). *Nucleic acids research*, *33*(Database issue), D154–D159.

Barker, W. C., Garavelli, J. S., McGarvey, P. B., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S. L., et al. (1999). The PIR-International Protein Sequence Database. *Nucleic acids research*, *27*(1), 39–43.

Berglund, H., Olerenshaw, D., Sankar, A., Federwisch, M., McDonald, N. Q., & Driscoll, P. C. (2000). The three-dimensional solution structure and dynamic properties of the human FADD death domain. *Journal of molecular biology*, *302*(1), 171–188.

Birzele, F., Gewehr, J. E., Csaba, G., & Zimmer, R. (2007). Vorolign--fast structural alignment using Voronoi contacts. *Bioinformatics (Oxford, England)*, *23*(2), e205–11.

Boeckmann, B. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, *31*(1), 365–370.

Brautigam, C. A., Sun, S., Piccirilli, J. A., & Steitz, T. A. (1999). Structures of Normal Single-Stranded DNA and Deoxyribo-3‘- S-phosphorothiolates Bound to the 3‘-5‘ Exonucleolytic Active Site of DNA Polymerase I from Escherichia coli†,‡. *Biochemistry*, *38*(2), 696–704.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, *5*(4), 823–826.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics (Oxford, England)*, *20*(3), 426–427.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–1423.

Csaba, G., Birzele, F., & Zimmer, R. (2008). Protein structure alignment considering phenotypic plasticity. *Bioinformatics (Oxford, England)*, *24*(16), i98–104.

Ellenberger, T., Doublié, S., Tabor, S., Long, A. M., & Richardson, C. C. (1998). Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature*, *391*(6664), 251–258.

Enoiu, M., Jiricny, J., & Schärer, O. D. (2012). Repair of cisplatin-induced DNA interstrand crosslinks by a replication-independent pathway involving transcription-coupled repair and translesion synthesis. *Nucleic acids research*, *40*(18), 8953–8964.

Evans, R. J., Davies, D. R., Bullard, J. M., Christensen, J., Green, L. S., Guiles, J. W., Pata, J. D., et al. (2008). Structure of PolC reveals unique DNA binding and fidelity determinants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(52), 20695–20700.

Fiala, K. A., Brown, J. A., Ling, H., Kshetry, A. K., Zhang, J., Taylor, J.-S., Yang, W., et al. (2007). Mechanism of template-independent nucleotide incorporation catalyzed by a template-dependent DNA polymerase. *Journal of molecular biology*, *365*(3), 590–602.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, *39*(Web Server issue), W29–37.

Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z., & Godzik, A. (2007). Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, *23*(2), e219–24.

Garcia-Diaz, M., Bebenek, K., Krahn, J. M., Kunkel, T. A., & Pedersen, L. C. (2005). A closed conformation for the Pol lambda catalytic cycle. *Nature structural & molecular biology*, *12*(1), 97–98.

Gibrat, J.-F., Madej, T., & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current opinion in structural biology*, *6*(3), 377–385.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(13), 4355–4358.

Guerler, A., & Knapp, E.-W. (2008). Novel protein folds and their nonsequential structural analogs. *Protein science : a publication of the Protein Society*, *17*(8), 1374–1382.

H Klenow, I. H. (1970). Selective Elimination of the Exonuclease Activity of the Deoxyribonucleic Acid Polymerase from Escherichia coli B by Limited Proteolysis. *Proceedings of the National Academy of Sciences of the United States of America*, *65*(1), 168. National Academy of Sciences.

Hasegawa, H., & Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current opinion in structural biology*, *19*(3), 341–348.

Hendrickson, W. A. (1979). Transformations to optimize the superposition of similar structures. *Acta Crystallographica Section A*, *35*(1), 158–163. International Union of Crystallography.

Hogg, M., Wallace, S. S., & Doublié, S. (2004). Crystallographic snapshots of a replicative DNA polymerase encountering an abasic site. *The EMBO journal*, *23*(7), 1483–1493.

Holm, L., & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics (Oxford, England)*, *16*(6), 566–567.

Hopfner, K. P., Eichinger, A., Engh, R. A., Laue, F., Ankenbauer, W., Huber, R., & Angerer, B. (1999). Crystal structure of a thermostable type B DNA polymerase from Thermococcus gorgonarius. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(7), 3600–3605.

Hübscher, U., Maga, G., & Spadari, S. (2002). Eukaryotic DNA polymerases. *Annual review of biochemistry*, *71*(1), 133–163.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*.

Jiang, Y., Nock, S., Nesper, M., Sprinzl, M., & Sigler, P. B. (1996). Structure and importance of the dimerization domain in elongation factor Ts from Thermus thermophilus. *Biochemistry*, *35*(32), 10269–10278.

Johnson, S. J., Taylor, J. S., & Beese, L. S. (2003). Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(7), 3895–3900.

Joyce, C. M., & Benkovic, S. J. (2004). DNA polymerase fidelity: kinetics, structure, and checkpoints. *Biochemistry*.

Joyce, C. M., & Steitz, T. A. (1994). Function and structure relationships in DNA polymerases. *Annual review of biochemistry*, *63*, 777–822.

Kamtekar, S., Berman, A. J., Wang, J., Lázaro, J. M., de Vega, M., Blanco, L., Salas, M., et al. (2004). Insights into Strand Displacement and Processivity from the Crystal Structure of the Protein-Primed DNA Polymerase of Bacteriophage ϕ29. *Molecular cell*, *16*(4), 609–618.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), 3059–3066.

Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., & Leberman, R. (1996). The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 A resolution. *Nature*, *379*(6565), 511–518.

Kiefer, J. R., Mao, C., Hansen, C. J., Basehore, S. L., Hogrefe, H. H., Braman, J. C., & Beese, L. S. (1997). Crystal structure of a thermostable Bacillus DNA polymerase I large fragment at 2.1 A resolution. *Structure*, *5*(1), 95–108.

Kim, S. W., Kim, D.-U., Kim, J. K., Kang, L.-W., & Cho, H.-S. (2008). Crystal structure of Pfu, the high fidelity DNA polymerase from Pyrococcus furiosus. *International journal of biological macromolecules*, *42*(4), 356–361.

Kim, Y., Eom, S. H., Wang, J., Lee, D.-S., Suh, S. W., & Steitz, T. A. (1995). Crystal structure of Thermus aquaticus DNA polymerase. *Nature*, *376*(6541), 612–616.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: A multiple structural alignment algorithm. *Proteins*, *64*(3), 559–574.

Kunkel, T. A., & Bebenek, K. (2000). DNA replication fidelity. *Annual review of biochemistry*, *69*, 497–529.

Lamers, M. H., Georgescu, R. E., Lee, S.-G., O'Donnell, M., & Kuriyan, J. (2006). Crystal structure of the catalytic alpha subunit of E. coli replicative DNA polymerase III. *Cell*, *126*(5), 881–892.

Lee, Y.-S., Kennedy, W. D., & Yin, Y. W. (2009). Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations. *Cell*, *139*(2), 312–324.

Leslin, C. M., Abyzov, A., & Ilyin, V. A. (2007). TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT method. *Nucleic acids research*, *35*(Database issue), D317–21.

Letunic, I., & Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128.

Leulliot, N., Cladière, L., Lecointe, F., Durand, D., Hübscher, U., & van Tilbeurgh, H. (2009). The family X DNA polymerase from Deinococcus radiodurans adopts a non-standard extended conformation. *The Journal of biological chemistry*, *284*(18), 11992–11999.

Liu, S., Knafels, J. D., Chang, J. S., Waszak, G. A., Baldwin, E. T., Deibel, M. R., Thomsen, D. R., et al. (2006). Crystal structure of the herpes simplex virus 1 DNA

polymerase. *The Journal of biological chemistry*, *281*(26), 18193–18200.
Maciejewski, M. W., Shin, R., Pan, B., Marintchev, A., Denninger, A., Mullen, M. A., Chen, K., et al. (2001). Solution structure of a viral DNA repair polymerase. *Nature structural biology*, *8*(11), 936–941.
Madhusudhan, M. S., Webb, B. M., Marti-Renom, M. A., Eswar, N., & Sali, A. (2009). Alignment of multiple protein structures based on sequence and structure features. *Protein engineering, design & selection : PEDS*, *22*(9), 569–574.
Martínez, L., Andreani, R., & Martínez, J. M. (2007). Convergent algorithms for protein structural alignment. *BMC bioinformatics*, *8*, 306.
Mizuguchi, K., Deane, C. M., Blundell, T. L., & Overington, J. P. (1998). HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, *7*(11), 2469–2471.
Moon, A. F., Garcia-Diaz, M., Bebenek, K., Davis, B. J., Zhong, X., Ramsden, D. A., Kunkel, T. A., et al. (2006). Structural insight into the substrate specificity of DNA Polymerase μ. *Nature structural & molecular biology*, *14*(1), 45–53. Nature Publishing Group.
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. *Journal of molecular biology*, *247*(4), 536–540.
Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S., & Aggarwal, A. K. (2005). Rev1 employs a novel mechanism of DNA synthesis using a protein template. *Science*, *309*(5744), 2219–2222.
Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S., & Aggarwal, A. K. (2006). Hoogsteen base pair formation promotes synthesis opposite the 1,N6-ethenodeoxyadenosine lesion by human DNA polymerase ι. *Nature structural & molecular biology*, *13*(7), 619–625.
Ortiz, A. R., Strauss, C., & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science*.
Panjkovich, A., Melo, F., & Marti-Renom, M. A. (2008). Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs. *Genome biology*, *9*(4), R68.
Pearl, F. M. G., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., et al. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic acids research*, *31*(1), 452–455.
Pei, J. (2008). Multiple protein sequence alignment. *Current opinion in structural biology*, *18*(3), 382–386.
Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *35*(Database), D61–D65.
Rechkoblit, O., Malinina, L., Cheng, Y., Kuryavyi, V., Broyde, S., Geacintov, N. E., & Patel, D. J. (2006). Stepwise translocation of Dpo4 polymerase during error-free bypass of an oxoG lesion. *PLoS biology*, *4*(1), e11.
Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., et al. (2010). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, *39*(Database), D392–D401.
Sacan, A., Toroslu, I. H., & Ferhatosmanoglu, H. (2008). Integrated search and alignment of protein structures. *Bioinformatics*, *24*(24), 2872–2879.

Savino, C., Federici, L., Johnson, K. A., Vallone, B., Nastopoulos, V., Rossi, M., Pisani, F. M., et al. (2004). Insights into DNA replication: the crystal structure of DNA polymerase B1 from the archaeon Sulfolobus solfataricus. *Structure*, *12*(11), 2001–2008.

Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering*, 11(9) 739-747.

Silvian, L. F., Toth, E. A., Pham, P., Goodman, M. F., & Ellenberger, T. (2001). Crystal structure of a DinB family error-prone DNA polymerase from Sulfolobus solfataricus. *Nature structural biology*, *8*(11), 984–989.

Sippl, M. J., & Wiederstein, M. (2012). Detection of Spatial Correlations in Protein Structures and Molecular Complexes. … *structure*, *20*(4), 718–728.

Sippl, M. J., Suhrer, S. J., Gruber, M., & Wiederstein, M. (2008). A discrete view on fold space. *Bioinformatics*, *24*(6), 870–871.

Slater, A. W., Castellanos, J. I., Sippl, M. J., & Melo, F. (2013). Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*, *29*(1), 47–53.

Steitz, T. A. (1999). DNA polymerases: structural diversity and common mechanisms. *Journal of Biological Chemistry*.

Steitz, T. A., & Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(14), 6498–6502.

Suhrer, S. J., Wiederstein, M., Gruber, M., & Sippl, M. J. (2009). COPS--a novel workbench for explorations in fold space. *Nucleic acids research*, *37*(Web Server), W539–W544.

Swan, M. K., Johnson, R. E., Prakash, L., Prakash, S., & Aggarwal, A. K. (2009a). Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase delta. *Nature structural & molecular biology*, *16*(9), 979–986.

Swan, M. K., Johnson, R. E., Prakash, L., Prakash, S., & Aggarwal, A. K. (2009b). Structure of the human Rev1-DNA-dNTP ternary complex. *Journal of molecular biology*, *390*(4), 699–709.

Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *Journal of molecular biology*, *208*(1), 1–22.

Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2002). *Multiple Sequence Alignment Using ClustalW and ClustalX*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Uljon, S. N., Johnson, R. E., Edwards, T. A., Prakash, S., Prakash, L., & Aggarwal, A. K. (2004). Crystal Structure of the Catalytic Core of Human DNA Polymerase Kappa. … *structure*, *12*(8), 1395–1404.

Wang, F., & Yang, W. (2009). Structural insight into translesion synthesis by DNA Pol II. *Cell*, *139*(7), 1279–1289.

Wang, S., & Zheng, W.-M. (2008). CLePAPS: fast pair alignment of protein structures based on conformational letters. *Journal of bioinformatics and computational biology*, *6*(2), 347–366.

Xiao, T., Towb, P., Wasserman, S. A., & Sprang, S. R. (1999). Three-dimensional structure of a complex between the death domains of Pelle and Tube. *Cell*, *99*(5), 545–555.

Zhang, Y. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, *33*(7), 2302–2309.

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, *57*(4), 702–710.