Juritz, Ezequiel Iván

# Caracterización estructural de proteínas por métodos evolutivos

# Caracterización estructural de proteínas por métodos evolutivos

*TESIS DOCTORAL*

**Ezequiel Iván Juritz**

ejuritz@unq.edu.ar

**Resumen**

Si bien está bien establecido hace tiempo ya que la dinámica conformacional de una proteína es un factor esencial para entender su función y su evolución secuencial, los métodos y técnicas más difundidas para el estudio y estimación de parámetros a partir de la estructura proteica consideran el estado nativo de la proteína como una entidad rígida y única.

Es sabido que la estructura de una proteína condiciona su evolución, esto es, su divergencia secuencial. Por consiguiente, una comprensión integral del proceso evolutivo a nivel molecular debe estar en relación con un entendimiento profundo de la estructura funcional de la proteína. El presente trabajo presenta una contribución hacia una consideración más realista de la estructura proteica, tomando en cuenta la dinámica estructural asociada a su estado nativo. Consideramos que es necesario un punto de inflexión en lo referente a la información asociada a la estructura nativa de las proteínas. Con tal fin, desarrollamos aquí un estudio de la caracterización de la diversidad conformacional proteica y su impacto en la divergencia evolutiva.

Decidimos independizarnos de métodos teóricos para análisis de dinámica estructural, y utilizar en su lugar únicamente estructuras provenientes de cristalizaciones. Nos basamos en el premisa que estructuras de la misma proteína, cristalizadas en distintas instancias, pueden contener información acerca de su diversidad conformacional.

Como primer paso generamos una base de datos con la colección redundante de estructuras cristalográficas disponibles para cada dominio proteico: Protein Conformational Diversity DataBase (PCDB, (Juritz, Alberti, & Parisi, 2011)), disponible en línea en http://www.pcdb.unq.edu.ar. El estudio de la diversidad conformacional fue efectuado utilizando estructuras cristalográficas de proteínas separadas en dominios, con el fin de sustraerse de los movimientos entre dominios que, al no representar necesariamente un reacomodo significativo entre los residuos que componen los núcleos funcionales proteico, dificultarían el análisis de las deformaciones estructurales intra-dominios. Estas estructuras, al pertenecer a la misma proteína, pueden ser consideradas instancias estructurales representativas del conjunto de confórmeros en equilibrio dinámico que define el estado nativo y sus diferencias estructurales, evaluadas a partir del cálculo de la raíz cuadrada de la desviación cuadrática media (RMSD, del inglés Root Mean Square Deviation) son una medida de la diversidad conformacional de la proteína.

Los análisis indican valores de diversidad conformacional significativos para gran parte de las proteínas estudiadas, que sugieren la importancia de tomar en consideración este factor en estudios que involucren la estructura proteica. Con tal fin se efectuaron estudios para evaluar la influencia de la diversidad conformacional en ciertas prácticas bioinformáticas.

Mediante la aplicación un modelo de evolución molecular que toma en cuenta explícitamente los condicionamientos estructurales de la proteína en estudio, y análisis estadísticos de máxima verosimilitud, evaluamos cuán adecuado condice cada confórmero estructural la divergencia secuencial observada en las familias proteicas. Hemos encontrado que en la mayoría de los casos, el confórmero que une el ligando, no necesariamente el de menor energía, es el que se condice mejor con la diversidad secuencial observada. Adicionalmente, al usar la información codificada en la totalidad de los confórmeros estudiados, es cuando se obtienen los mejores valores de máxima verosimilitud.

Estos resultados indican que las sustituciones aminoacídicas derivan en efectos específicos respecto al confórmero proteico estudiado. Basados en esta afirmación; en la asunción de que la estructura proteica condiciona la diversidad secuencial; y en que las diferencias estructurales entre los confórmeros pueden presentar valores significativos, decidimos estudiar la incidencia de la diversidad conformacional en la estimación de la perturbación estructural ante un evento mutacional. La medida más ampliamente difundida para estudiar este efecto en las proteínas es el valor de $\Delta\Delta G$ [Kcal mol-1]. Muchas líneas de

investigación apuntan al desarrollo y a la mejora de algoritmos de predicción de ΔΔG. Presentamos aquí resultados que indican que tomar en consideración la diversidad conformacional de las proteínas se traduce en un incremento en la precisión de las estimaciones teóricas del valor de ΔΔG de una mutación en la secuencia proteica en todos los métodos predictivos estudiados.

La capacidad de estimación del ΔΔG de una mutación reviste de importancia en numerosos campos, entre los cuales destaca el de la salud humana. Una aplicación concreta es la mejora en la predicción referente a si una mutación puede desencadenar un cuadro patológico o no mediante la desestabilización de la estructura nativa proteica. Con tal fin, procedimos a estudiar la incidencia de la diversidad conformacional en la capacidad predictiva sobre el efecto patógeno o neutral de una mutación. Nuestros resultados indican que, si bien esta capacidad predictiva aún dista de ser precisa, la consideración de la diversidad conformacional deriva en un incremento.

Concluimos que la consideración de la diversidad conformacional es un elemento trascendental para una compresión integral de la estructura, la función y la evolución proteica. A su vez, son necesarios nuevos enfoques y aplicaciones con el fin de lograr un provecho de la información codificada en la diversidad conformacional.

Como continuación del presente trabajo, se encuentra en las últimas etapas de desarrollo, y disponible en línea a partir de Abril del 2012, una base de datos de diversidad conformacional de proteína enteras, derivadas de la base de datos PDB. Esta nueva de base de datos presenta una variedad de información acerca de las proteínas y las estructuras depositadas y permitirá efectuar nuevos análisis y correlaciones de la diversidad conformacional del estado nativo, asimismo como comparaciones entre la diversidad conformacional intra e inter dominios.

Deseo dedicar mi tesis al mundo de la investigación y la docencia que nos da tantas satisfacciones a quienes estamos inmersos en él. A las instituciones que lo componen y a todos los que día a día ponemos nuestro pequeño aporte.

A Gustavo, por enseñarme tanto, por dirigirme, y sobre todo, por volverme al camino cuando me pierdo.

A todos mis colegas y profesores, por hacer de la actividad diaria una alegría.

A Lili por apoyarme, por corregirme, y por hacerme feliz.

A mi familia, por saber que siempre allá en el Sur hay alguien con quien contar.

A Gustavo Pierdominici, por su nobleza, su malhumor y sus consejos.

A Nicolás Palopoli, por todo el camino recorrido juntos.

A Sebastián por su buen humor y su energía.

A mis abuelos por pensar que sigo siendo un chico.

Indice

1. Introducción

Objetivos

2. Desarrollo de la base de datos PCDB

2.1 Introducción

2.2 Implementación de la base de datos PCDB

2.3 Lenguajes de programación y motores de bases de datos

2.4 Servidores, bases de datos y métodos bioinformáticos utilizados

CATH

PDB

MAMMOTH

Catalytic site atlas

UniProt

Enzyme Commission

Gene Ontology

PQS

Procognate

3. Servidor en línea de la base de datos PCDB

3.1 Funcionalidad

3.2 Efectuar una búsqueda en la base de datos PCDB

3.2.1 Limitar por posibles causas de diversidad conformacional.

3.2.2 Limitar por extensión de la diversidad conformacional.

3.2.3 Limitar por código PDB o por código CATH.

3.2.4 Limitar por código de clasificación estructural CATH.

3.2.5 Información de salida.

3.2.6 Enviar la búsqueda.

3.2.7 Resultados de la búsqueda.

3.3 Navegación por la base de datos PCDB.

3.3.1 Clasificación estructural.

3.3.2 Seleccionar Arquitectura

3.3.3 Dominios pertenecientes a la clasificación estructural actual.

4. Diversidad conformacional en la base de datos PCDB.

4.1 Introducción

4.2 Diversidad conformacional bajo condiciones homogéneas y heterogéneas.

4.3 Incidencia de diversos parámetros sobre la diversidad conformacional.

5. Efecto de la diversidad conformacional en la divergencia secuencial

5.1 Resumen

5.2 Introducción

5.3 Set de datos utilizado

5.4 Simulaciones SCPE

5.5 Cómputos de máxima verosimilitud

5.6 Confórmeros que unen ligando y confórmeros de mínima energía

5.7 Resultados

5.7.1 Evolución bajo condicionamientos estructurales

5.7.2 Evolución bajo escasos o nulos condicionamientos estructurales

5.7.3 Evolución y diversidad conformacional

Conclusiones

6. Diversidad conformacional y estimación de la estabilidad proteica

6.1 Resumen

6.2 Introducción

6.3 Metodología

6.4 Resultados y discusión

7. Diversidad conformacional y predicción del efecto de las mutaciones

7.1 Resumen

7.2 Introducción

7.3 Metodología

7.4 Resultados 96

7.4.1 Dversidad conformacional

7.4.2 Diferencias en los valores de ΔΔG obtenidos

7.4.3 Análisis estadísticos

7.5 Conclusiones

Conclusiones

Perspectivas

Referencias

## Capítulo 1

### Introducción

Los procedimientos y análisis computacionales representan un rol cada vez más trascendental en cualquier disciplina científica, y la biología no es una excepción. Particularmente, la gestión, el análisis computacional y el desarrollo de métodos teórico-computacionales y asentados en bases de datos con información de índole biológica, se engloban dentro de lo que conocemos actualmente como bioinformática.

La bioinformática, como todas las ciencias de base teórica, se sustenta mayormente de datos de origen experimental, siendo su principal fuente de información la secuenciación de ADN, la cristalización de proteínas y los estudios de interacciones y regulaciones proteicas. Los continuos progresos tecnológicos así como los avances de la biotecnología generan actualmente un crecimiento exponencial en el volumen de datos biológicos disponibles.

El continuo progreso de las técnicas en el campo del secuenciamiento y cristalización de proteínas genera un aumento exponencial en el volumen de las bases de datos de datos biológicas disponibles. En pos de lograr un aprovechamiento eficiente de esta información, son necesarios nuevos enfoques y perspectivas originales a fin de evaluar, clasificar y vincular esta nueva información estructural y secuencial de manera óptima. Simultáneamente, el grado de avance de las técnicas de biología computacional es notorio, y nuevas herramientas y perfeccionamientos emergen continuamente.

Con el fin de aprovechar todo el potencial de estos datos, y asentándose en el continuo crecimiento de la capacidad de cómputo, es preciso no sólo la mejora y aplicación de técnicas de procesamiento de datos orientadas al estudio de sistemas biológicos, sino también el desarrollo de nuevos modelos teóricos.

Dentro de este contexto, los autores consideran que, paralelamente con el desarrollo y refinamiento de las técnicas de biología computacional y bioinformática, es necesario iniciar un proceso de actualización conceptual de la estructura nativa proteica. El avance experimental en los campos de la cristalografía y el secuenciamiento; los adelantos de la bioinformática y la biología computacional; sumados al aumento de la capacidad de cálculo y almacenamiento computacional componen un escenario fértil para generar un punto de inflexión en el enfoque y el planteo de ciertos problemas bioinformáticos.

Así, en el presente trabajo, pretendemos hacer un aporte hacia la incorporación de una visión más realista de la estructura nativa de una proteína.

Desde las primeras apreciaciones sobre la estructura tridimensional de las proteínas de Hermann Emil Fischer (Fischer, 1894) hasta hoy, el concepto de estado nativo ha sido continuamente, y será, moldeado en base a nuevos modelos propuestos.

Entre finales del siglo XIX y principios del XX, Fischer demostró que las proteínas estaban constituidas por cadenas lineales de aminoácidos. Sus estudios acerca de la función enzimática de las proteínas fueron los primeros indicios acerca de la especificidad de la actividad enzimática. Basado en sus investigaciones, Fischer propuso la hipótesis de la "Llave y cerradura" como modelo relacional entre la estructura y la función proteica.

En la década de 1930, el modelo de la estructura tridimensional proteica era un tema latente en las investigaciones de los principales laboratorios de la época y ya se intuía su relación respecto a las propiedades biológicas de las proteínas. Las bases bioquímicas de la biología estructural de proteínas se iniciaron con manuscritos clásicos como los de Mirsky y Pauling (Mirsky & Pauling, 1936; L. Pauling & Coryell, 1936), orientados al estudio de la importancia de la estructura proteica para la conservación de su función. Estos investigadores, a pesar de no contar con modelos estructurales cristalinos de proteínas, intuyeron y definieron el estado nativo de las proteínas como aquella estructura única ("configuración", según la nomenclatura de Pauling) que confería las propiedades biológicas a las proteínas. Por el contrario, cualquier estado no funcional se denomino "desnaturalizado". Sin embargo, en 1936 Landsteiner notó que los anticuerpos presentaban cierto grado de promiscuidad en su unión a ligandos (Landsteiner, 1936). Landsteiner, en 1940, escribía sobre sus investigaciones "Uno podría deducir que los distintos plegamientos, conjuntamente con los cambios en la orientación de los aminoácidos, podría alcanzar hasta el más amplio espectro de variaciones estructurales" (Landsteiner, 1936).

En 1940, hace ya 70 años, Pauling, en base a estudios de anticuerpos y su promiscuidad, propuso que las proteínas (anticuerpos, particularmente) pueden existir como un conjunto de estructuras con estabilidades energéticas cercanas (Linus Pauling, 1940). Según su propuesta, la promiscuidad enzimática era explicada por la especificidad diferencial de cada uno estos "isómeros". Aún así, este modelo no fue aceptado por muchas décadas.

Las investigaciones acerca de la actividad enzimática y su promiscuidad daban indicios de que el modelo de "Llave y cerradura" encerraba una simplicidad excesiva. En 1958, Daniel Koshland, en base a sus resultados de la actividad de numerosas enzimas, propone un nuevo modelo de la actividad enzimática, conocido como el modelo del

“Ajuste inducido”. En este modelo la estructura proteica es conceptualizada bajo cierto dinamismo, y el sustrato ejerce un efecto modulador sobre ésta en el proceso de unión.

Los posteriores (y más exactos) estudios de Max Perutz (luego de haber trabajado en un proyecto secreto para hacer un portavión de Pykrete y de demostrar que los glaciares fluyen) acerca de las formas T y R de la hemoglobina (Perutz et al., 1960), y de cómo éstas complementaban el modelo propuesto por de Monod (Monod, Wyman, & Changeux, 1965) explicando su cooperatividad y alosterismo, evidencian ya la inexactitud en la que se incurre al representar el estado nativo bajo la forma simplista de una única estructura.

En 1968 Cyrus Levinthal notó que, dado el elevado número de grados de libertad en un polipéptido desplegado, la molécula posee una cantidad astronómica de conformaciones posibles (Levinthal, 1968). En uno de sus trabajos, estimó su número aproximadamente entre 3300 y 10143 (Levinthal, 1969). Por más veloz que una proteína de 100 aminoácidos “ensaye” cada uno de estos confórmeros hasta dar con el adecuado (esto es, el de menor energía), tardaría aproximadamente la edad del universo, cuando en la realidad este proceso ocurre en el rango de milisegundos a microsegundos. Este concepto se conoce bajo el nombre de la “Paradoja de Levinthal” (para una información más detallada ver (Zwanzig, Szabo, & Bagchi, 1992)), y apunta a que el proceso de plegado proteico se encuentra conducido por la formación rápida de interacciones locales que dirigen los subsiguientes pasos de plegado estructural hasta un mínimo, o conjunto de mínimos, de energía.

En la década de 1980, la importancia del modelo de embudo para el proceso del plegado proteico era indiscutible (J. D. Bryngelson & Wolynes, 1989). Adicionalmente, por la década de 1990, el modelo incorporó el concepto de “embudo rugoso” como plausible interpretación de la paradoja de Levinthal (Joseph D Bryngelson, Nelson, Nicholas, & Wolynes, 1995; Nienhaus, Muller, McMahon, & Frauenfelder, 1997; P. G. Wolynes, 1997). Este modelo toma en consideración no solo un mínimo de energía, sino un conjunto de confórmeros estables, situados en mínimos locales y separados por pequeños reordenamientos estructurales, lo que le da su carácter de rugoso al fondo del paisaje termodinámico. Este último concepto es inherente a la diversidad conformacional del estado nativo, y expresa la imposibilidad de escindir la dinámica estructural con el proceso del plegado proteico.

Posteriores estudios han analizado la relación entre la diversidad conformacional y la promiscuidad enzimática (Copley, 2003; O’Brien & Herschlag, 1999), el pH de cristalización (Bocharov et al., 2008), la divergencia secuencial (Friedland, Lakomek, Griesinger, Meiler, & Kortemme, 2009; Juritz, Palopoli, Fornasari, Fernandez-Alberti, &

Parisi, 2012) y la capacidad de evolución (Nobuhiko Tokuriki & Tawfik, 2009), por mencionar algunos ejemplos. Estos estudios han puesto en tela de juicio la capacidad de estudiar ciertos aspectos biológicos y fisicoquímicos de las proteínas siguiendo el concepto de rigidez de la estructura y especificidad de la función proteica.

El concepto de la “nueva visión” de las proteínas (James & Tawfik, 2003) ha surgido durante el transcurso de los últimos años. Este concepto se basa en la premisa de que una secuencia puede adoptar varias estructuras (o confórmeros), otorgándole diversidad estructural, funcional y permitiendo una alta velocidad de evolución. El modelo de la “nueva visión” sugiere que existe una conformación predominante en la estructura nativa proteica, con afinidad a un ligando específico. En menor medida, conformaciones alternativas en el estado nativo, presentan diversas afinidades por otros sustratos secundarios. Esta promiscuidad no se desarrolla de forma notoria debido a que la conformación primaria secuestra la mayor parte de la proteína. La actividad secundaria puede ser incrementada mediante eventos mutacionales que estabilicen las conformaciones con máxima afinidad para dicha actividad secundaria redistribuyendo las poblaciones relativas en el equilibrio, pero sólo en cierto grado, dado que se puede perder la afinidad por el sustrato primario. Ahora bien, luego de un evento de duplicación génica, una copia del gen puede evolucionar incrementando su actividad catalítica hacia el nuevo sustrato, al punto de perder su función original. Asimismo, estas nuevas funciones pueden servir como punto de partido para un nuevo proceso.

Finalmente, los últimos trabajos en el tema han demostrado que una descripción integrada a la dinámica es esencial para interpretar aspectos biológicos funcionales de las proteínas como sus procesos catalíticos (Henzler-Wildman et al., 2007; Wolf-Watz et al., 2004), reconocimiento proteína-proteína (Fuentes, Der, & Lee, 2004; Lange et al., 2008; Yogurtcu, Erdemli, Nussinov, Turkay, & Keskin, 2008), procesos macromoleculares como replicación del ADN y el plegamiento de proteínas por medio de chaperonas (Russel et al., 2009), la promiscuidad enzimática (Khersonsky, Roodveldt, & Tawfik, 2006), los procesos de transducción de señales (Bai et al., 2010; Smock & Gierasch, 2009) y la capacidad de las proteínas para desarrollar nuevas funciones (propiedad conocida como "capacidad de evolución») (Aharoni et al., 2005; Nobuhiko Tokuriki & Tawfik, 2009).

Es así como en el presente trabajo apoyamos el concepto de un estado nativo mejor representado por un conjunto de confórmeros proteicos y caracterizado por un equilibrio dinámico inherente existente entre ellos, y no por el modelo de una sola estructura (B. Ma, Kumar, Tsai, & Nussinov, 1999). El concepto de una estructura tridimensional rígida ha sido muy útil para comprender ciertos factores asociados a la biología y a la

fisicoquímica de la proteína. Pero actualmente sabemos que la proteína debe pasar por muchos cambios energéticos y estructurales para cumplir su función. Aún así, el paradigma actual se encuentra basado en una relación estructura–función esencialmente rígida, y los métodos computacionales teóricos así lo asumen.

Se han efectuado diversos avances en el estudio de los dinamismos proteicos utilizando distintas técnicas. Entre los métodos experimentales, la resonancia magnética nuclear (RMN) se encuentra entre los enfoques más ampliamente utilizados (Lindorff-Larsen, Best, Depristo, Dobson, & Vendruscolo, 2005). Asimismo, se han obtenido buenas correlaciones en los análisis comparativos entre las estructuras obtenidas mediante métodos de dinámica molecular, para simular la dinámica de proteínas y aquellas estructuras resueltas por RMN (Best, Clarke, & Karplus, 2005; Prabhu, Lee, Wand, & Sharp, 2003)

En cuanto a los métodos computacionales, los resultados obtenidos mediante métodos de cálculo como la dinámica molecular de grano grueso ("Coarse-Grained Molecular Dynamics") y el método de Monte Carlo, usados en combinación con análisis de modos normales, han puesto de manifiesto que éstos constituyen herramientas válidas para explorar la diversidad conformacional de proteínas (Bahar & Rader, 2005; Chng & Yang, 2008; M Karplus & Kuriyan, 2005; J. Ma, 2005; Tozzini, 2005).

A pesar de estos avances en el campo, la caracterización del equilibrio del conjunto de confórmeros, consistente en el estudio estructural y las características termodinámicas de cada confórmero individual, representa un gran reto a superar.

Un enfoque completamente diferente para abordar la diversidad conformacional es considerar que las estructuras cristalográficas obtenidas en diferentes instancias representan confórmeros presentes en el estado nativo de la proteína. Chotia y Lesk han explorado la relación entre la divergencia secuencial y la divergencia estructural (Chothia & Lesk, 1986); Matthews ha estudiado la capacidad de adaptación de la estructura proteica a los cambios secuenciales (Matthews, 1996); y Karplus ha investigado acerca de variaciones de las estructuras cristalizadas (Zoete, Michielin, & Karplus, 2002).

Actualmente, con miles de estructuras redundantes depositadas en bases de datos de estructuras (Burra, Zhang, Godzik, & Stec, 2009), la extensión y distribución de la diversidad conformacional puede ahora ser explorada en un gran número de proteínas, no accesible mediante las metodologías citadas previamente.

**Objetivos**

El objetivo global de nuestro trabajo es relacionar la extensión de la diversidad conformacional con parámetros fisicoquímicos y procesos biológicos, para lograr una mayor comprensión de la relación entre diversidad conformacional, estructura, función y evolución proteica.

Los objetivos específicos son:

I. Diseño y desarrollo de una base de datos de diversidad conformacional proteica.

II. Estudiar la influencia de la diversidad conformacional sobre la divergencia secuencial durante la evolución.

III. Estudiar la variación relativa de la estabilidad proteica a las mutaciones en función de los distintos confórmeros disponibles.

Nuestro primer objetivo se basa en el diseño y desarrollo de una base de datos de diversidad conformacional proteica basada en comparaciones entre estructuras de las mismas proteínas cristalizada en distintas instancias. Adicionalmente a la colección redundante de estructuras, nos proponemos vincular las entradas de la PCDB con una gran variedad de información fisicoquímica y biológica a fin de estudiar correlaciones entre la extensión de la diversidad conformacional y diversos parámetros siendo, hasta nuestro conocimiento, la primera base de datos de estas características. En sucesivos capítulos describiremos el diseño y la funcionalidad a través de red de la base de datos PCDB, así como algunos estudios efectuados basados en la misma, para finalmente enumerar posibles futuras aplicaciones y mejoras. Esta base de datos abre nuevas posibilidades para el estudio y correlación de la diversidad conformacional proteica con diversos y variados parámetros.

Un segundo objetivo es estudiar la incidencia diferencial que cada confórmero estructural ejerce sobre la evolución en una familia de proteínas. Está bien establecido que la estructura de la proteína posee un efecto modulador sobre la diversidad secuencial. Con el concepto de un estado nativo representado por un conjunto de confórmeros en equilibrio, resulta interesante estudiar cómo la estructura de cada uno de ellos influye sobre la evolución observada en la naturaleza, y relacionar esta información con características termodinámicas (mayor o menor estabilidad) y biológicas (unión al ligando) de cada confórmero en particular. Para tal fin aplicaremos el SCPE (Gustavo Parisi & Echave, 2001), un modelo de evolución molecular que toma en consideración la estructura de la proteína y los condicionamientos estructurales que

ésta impone a cada sitio. Del SCPE se derivan matrices de sustitución sitio-específicas (M. S. Fornasari, Parisi, & Echave, 2002) que poseen información acerca de la posibilidad de ocurrencia de los 20 aminoácidos en cada sitio de la proteína. Mediante las técnicas estadísticas de máxima verosimilitud, nuestro objetivo es analizar cuán bien condice cada una de estas matrices, derivadas de los distintos confórmeros, la divergencia secuencial observada en la familia de homólogas.

Asimismo, la posibilidad de disponer de los confórmeros estructurales para una misma proteína permite efectuar análisis teóricos para un conjunto de estructuras y comparar los múltiples resultados. Creemos que este enfoque es un primer paso en tomar en consideración la información de la diversidad conformacional asociada al estado nativo. En ese sentido, estudiaremos la variación de la perturbación estructural ante un evento mutacional en los diferentes confórmeros. Los confórmeros estructurales que representan el estado nativo poseen características termodinámicas diferentes entre sí, por lo que estudios termodinámicos presentarán diferencias en base a qué confórmero se utilice.

**Capítulo 2**

**Desarrollo de la base de datos PCDB**

## 2.1 Introducción

Como mencionáramos anteriormente, uno de los primeros objetivos del presente trabajo es el desarrollo de una base de datos con de proteínas con distintos grados de diversidad conformacional. La base de datos PCDB (Protein Conformational Database) (Juritz et al., 2011), disponible en línea en http://www.pcdb.unq.edu.ar/, representa un paso en el camino hacia considerar la estructura nativa de una proteína como una entidad dinámica e integrada por un conjunto de estructuras en equilibrio.

Este servidor recopila casos de diversidad conformacional de dominios. Para cada dominio representado en la base de datos, PCDB contiene la recopilación redundante de todas las estructuras cristalográficas disponibles hasta el momento. La decisión de analizar las estructuras proteicas descompuestas en dominios se basa en que nuestro interés radica en la diversidad conformacional del dominio como un todo. Aquellos movimientos que involucren un desplazamiento o traslación entre dominios, pueden dar lugar a valores RMSD significativos sin reflejar un cambio apreciable en la disposición de los residuos que componen la unidad de los dominios funcionales.

A su vez, PCDB ofrece una completa colección de información de tipo biológica (e.g. función, localización subcelular, longitud, organismo al que pertenece) acerca de todas las proteínas representadas. De cada una de las estructuras cristalográficas disponibles se reúne información adicional concerniente a las condiciones de cristalización (pH de cristalización, presencia o ausencia sustrato, estado oligomérico, etc.). La información recopilada en la PCDB proviene de un gran número de bases de datos y servidores bioinformáticos especializados.

El objetivo de este servidor es reunir información y permitir consultas referentes a todas las estructuras reales disponibles para un mismo dominio. De esta manera es posible efectuar comparaciones entre estas estructuras, contrastar las diferencias y similitudes entre ellas, distinguir las partes más móviles de las menos móviles, etc. Y, adicionalmente, efectuar estudios para correlacionar estos aspectos estructurales con aspectos biológicos o fisicoquímicos de la proteína.

Dos características fundamentales diferencian la base de datos PCDB de otras bases de datos ya disponibles conteniendo información acerca de diversidad conformacional (Gerstein & Krebs, 1998; Kamp et al., 2010). En primera medida, PCDB usa exclusivamente estructuras resueltas experimentalmente. Es decir, no extrae estructuras teóricas adicionales derivadas a partir de una estructura experimental

mediante la aplicación a ésta de métodos computacionales como lo pueden ser modos normales, dinámica molecular, diversas minimizaciones energéticas, etc. Adicionalmente, la base de datos de estructuras, agrupa una nutrida y variada colección de datos (fisicoquímicos, biológicos, de condiciones de cristalización) relacionados a cada una de las estructuras y de los dominios representados. De esta manera es posible, mediante búsquedas o consultas definidas en el servidor en línea, buscar casos de diversidad conformacional que coincidan con ciertos parámetros de interés del usuario, como lo pueden ser la amplitud de la diversidad conformacional, la clasificación estructural de la proteína, presencia o ausencia de ligando, taxonomía, etc. La información contenida en PCDB posibilita un amplio espectro de estudios y de análisis de correlaciones entre diversidad conformacional y diversos parámetros contenidos en la base de datos, como lo son función proteica, localización subcelular, parámetros GO asociados, información sobre el sitio activo, estado oligomérico y longitud proteica, entre otros.

Una de las aplicaciones de mayor importancia de la base de datos PCDB es la posibilidad de reclutar estructuras pertenecientes a una misma proteína, con el objetivo de aplicar métodos bioinformáticos a distintas de sus estructuras. Consideramos que las diferencias que se aprecian en los resultados obtenidos son indicio de que es necesario considerar la diversidad conformacional en métodos teóricos basados en estructuras proteicas.

## 2.2 Implementación de la base de datos PCDB

A fin de estudiar la diversidad conformacional, se buscaron aquellas proteínas que han sido cristalizadas dos o más veces. Para ello se derivaron de la base de datos CATH (Greene et al., 2007; C. A. Orengo et al., 1997; Pearl et al., 2003) todos los dominios proteicos que tenían 2 o más estructuras cristalográficas depositadas. Las estructuras cristalográficas resueltas pueden ser consideradas como instancias de la diversidad conformacional del estado nativo (Best, Lindorff-Larsen, DePristo, & Vendruscolo, 2006). Se encuentran diferencias estructurales entre cristales de la misma proteína, ya sean bajo idénticas o diversas condiciones. Es necesario destacar que la completitud de la descripción del estado nativo depende en modo directo de la disponibilidad de estructuras del dominio en estudio.

Los movimientos de traslación o rotación de dominios enteros, o de secciones de proteínas sin una estructura definida que no forman parte de dominios, suelen conducir a altos valores de RMSD entre estructuras de proteínas completas. Para evitar considerar aquellos ejemplos en donde la diversidad conformacional se debe únicamente a este tipo de movimientos, y no a deformaciones de la estructura interna de los dominios funcionales, se decidió utilizar estructuras de dominios y no estructuras de proteínas enteras.

Las estructuras depositadas en la PCDB fueron extraídas de la base de datos CATH[1], guardando su clasificación estructural. A su vez, la base de datos CATH deriva de la base de datos "Protein Data Bank"[2] (PDB; (Berman et al., 2000).

Se efectuaron en una primera instancia comparaciones estructurales entre los confórmeros aplicando los algoritmos de los siguientes programas: Mammoth (Ortiz & Strauss, 2002), TM-Align (Yang Zhang & Skolnick, 2005), LGA (Zemla, 2003) y Profit[3]. Varios "scores" de similitud estructurales fueron obtenidos a partir de éstos: MaxSub, RMSD, TM-Score y GDT. Los valores obtenidos mediante análisis estadísticos de correlación entre todos los "scores" resultados de estos métodos fueron muy altos, por lo que decidió utilizar solamente un "score". El seleccionado fue el RMSD, dado que es más apropiado para la evaluación de diferencias estructurales, siendo los otros "scores" esencialmente diseñados para evaluar similitudes estructurales.

Como medida de la diversidad conformacional de cada dominio representado, se computaron los valores de Root-Mean-Square Deviation (RMSD) en Å entre todos los

[1] http://www.cathdb.info/
[2] http://www.pdb.org/pdb/home/home.do
[3] http://www.bioinf.org.uk/software/profit/

pares posibles de estructuras pertenecientes al mismo dominio. En base a estos datos, se registraron los valores máximos, mínimos y los promedios de RMSD para cada una de las entradas de la base de datos. En el servidor de la base de datos se encuentran disponibles todos los parámetros estadísticos citados registrados entre todos los pares de estructura para cada dominio. Consideramos el RMSD máximo y el RMSD promedio como buenos parámetros estadísticos descriptivos de la diversidad conformacional proteica.

En la base de datos PCDB hay un total de 7.989 dominios proteicos representados. La cantidad de estructuras contenidas en la base de datos es de 36.581, dando un promedio de 4,6 estructuras por dominio representado. La mínima cantidad de estructuras registradas por dominio es de 2, y el máximo de estructuras reclutadas para un dominio determinado es de 39 estructuras. El 21% (1.641) de los dominios representados poseen más de 5 estructuras. Esta información se muestra en la

Tabla 1.

Las diferencias entre las estructuras pertenecientes al mismo dominio reflejan de modo directo los movimientos a los que está sujeta el mismo. El hecho de contar únicamente con estructuras provenientes de cristalizaciones y no derivadas por técnicas computacionales excluye la posibilidad de arribar a conclusiones atribuibles a artefactos o inexactitudes propios de los métodos predictivos teóricos.

| | |
|---|---|
| Mínima cantidad de estructuras por dominio | 2 |
| Promedio de cantidad de estructuras por dominio | 4,6 |
| Máxima cantidad de estructuras por dominio | 39 |
| Cantidad de dominios proteicos representados | 7.989 |
| Número de estructuras comprendidas | 36.581 |

Tabla 1. Estadísticas acerca de las estructuras que componen la base de datos PCDB.

Al pertenecer todas las estructuras comparadas a la misma proteína, y al mismo dominio proteico, es esperable que la identidad secuencial entre estas sea del 100%. Aún así, existen casos en donde se detectan leves diferencias en las secuencias de las distintas estructuras de un mismo dominio, causadas bien por gaps o por disimilitudes en los cortes para obtener los dominios. El origen de estos gaps se debe mayormente a la existencia de residuos no detectados en los métodos de rayos X. Aún así, la influencia de estos gaps es sumamente reducida, puesto que el 70% de los dominios poseen un 100% de identidad secuencial entre todas sus estructuras, un 87% posee

una identidad secuencial mayor a al 95%, y solo un 5% presenta un porcentaje de identidad secuencial menor al 90%. Se decidió incluir aquellos dominios con un porcentaje de identidad menor al 100%, con la posibilidad de filtrar por este criterio en la página de búsqueda del servidor. La identidad secuencial se estimó por medio del programa Profit.

En la Figura 1 se esquematizan los pasos efectuados para el desarrollo de la base de datos. Asimismo, se detalla toda la información disponible para cada uno de los dominios, y para cada una de las estructuras que representan a cada dominio, y la fuente de donde se extrajo esa información. Más adelante se describen los servidores y las bases de datos vinculadas a la PCDB.

Si bien la diversidad conformacional es un factor inherente a la estructura proteica y esencial para explicar la catálisis enzimática, existen ciertos factores que pueden modular su extensión. Al cristalizar una proteína en las mismas condiciones, es esperable encontrar ciertas diferencias estructurales entre ellas. Adicionalmente, cuando las condiciones de cristalización varían, estas diferencias se acentúan, debido a desplazamientos del equilibrio dinámico entre los confórmeros. Esto es debido a la existencia de ciertos factores que modulan la extensión de la diversidad conformacional, como el pH (Bocharov et al., 2008), la variación del estado oligomérico de la proteína (Bernardes et al., 2012), mutaciones (Laine, Chauvot de Beauchêne, Perahia, Auclair, & Tchertanov, 2011) y la presencia de ligandos (Ostermann, Waschipky, Parak, & Nienhaus, 2000), entre otros. Estos cuatro factores fueron considerados como moduladores de la diversidad conformacional en la PCDB, vinculando cada estructura representada en la base de datos con esta información concerniente a sus condiciones de cristalización. Es posible de esta manera, con la herramienta de búsqueda, seleccionar solo aquellos dominios que presenten estructuras que hayan sido cristalizadas en condiciones heterogéneas para cualquier combinación de los parámetros citados. Asimismo, es posible seleccionar solo aquellos dominios cuyas estructuras hayan sido cristalizadas en las mismas condiciones respecto a estos parámetros.

En futuras actualizaciones de la PCDB se incluirán más factores que puedan estar asociados a la modulación de la diversidad conformacional, con lo que se espera una mayor diferenciación entre las distribuciones de RMSD de la población de dominios con estructuras cristalizadas en iguales y en diversas condiciones.

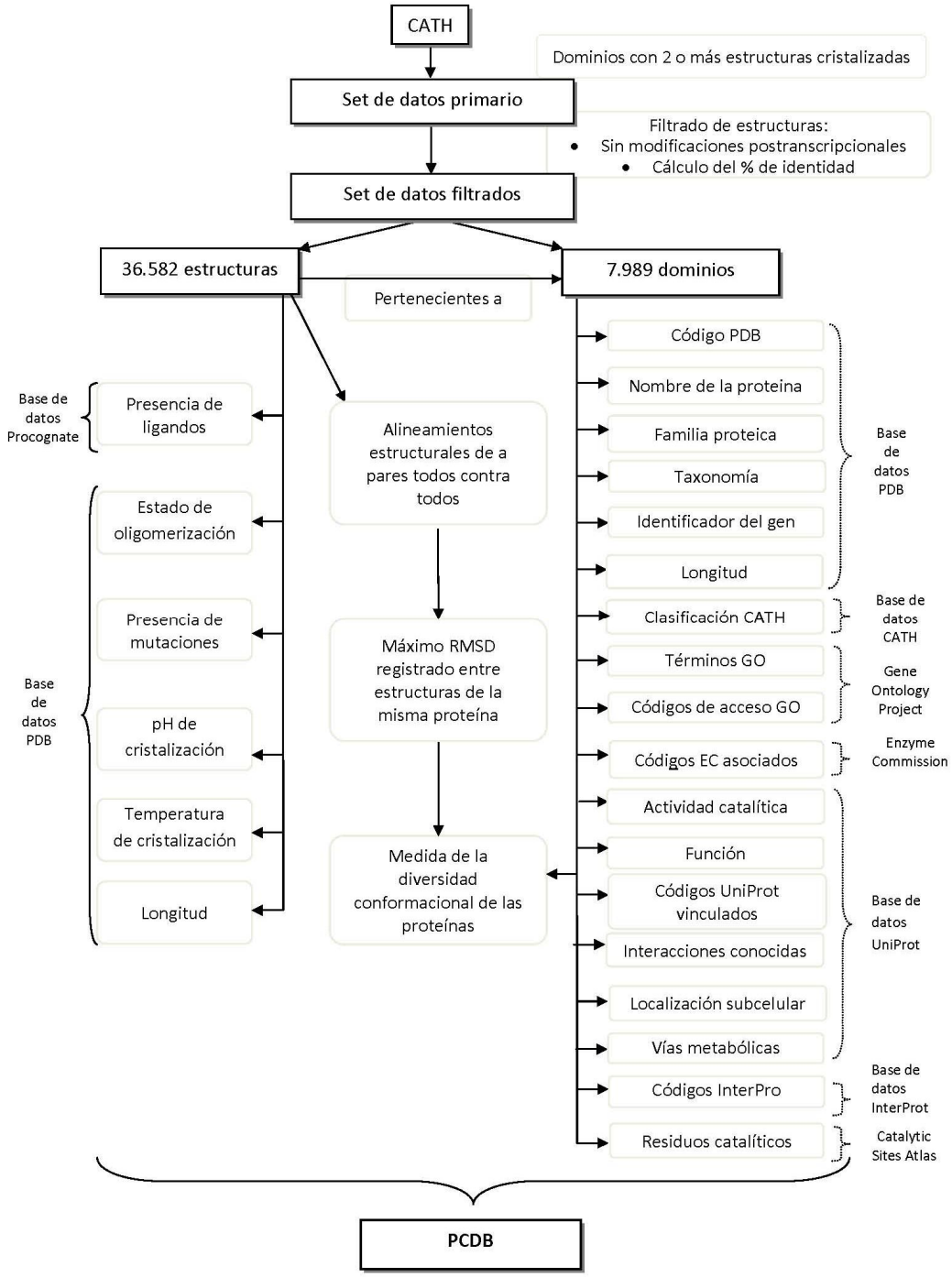



**Figura 1. Esquema representativo del desarrollo de la base de datos PCDB**

### 2.3 Lenguajes de programación y motores de bases de datos

El sitio de internet de la base de datos ha sido diseñado e implementado mediante los lenguajes HTML y PHP versión 5.3.1.

La base de datos está basada en el lenguaje SQL, y ha sido desarrollada mediante el motor de bases de datos MySQL 5.0.76

El gestor y administrador de MySQL utilizado fue el SQLyog Community – MySQL GUI v8.224.

Las consultas que se envían a la base de datos a través del sitio de internet se generan dinámicamente mediante lenguaje PHP del lado del servidor en el sitio web, siguiendo las pautas definidas por el usuario. Las consultas enviadas se efectúan basadas en el lenguaje SQL, y han sido implementadas mediante el motor de base de datos MySQL.

La base de datos y la página de internet a la que acceden sus usuarios se encuentran alojadas en los servidores de la Universidad Nacional de Quilmes, y son mantenidos por el equipo técnico establecido para tal fin por dicha institución.

[4] http://www.webyog.com/

## 2.4 Servidores, bases de datos y métodos bioinformáticos utilizados

Las estructuras comprendidas en la PCDB han sido vinculadas con otras bases de datos y servidores, y han sido expuestas a diversos métodos bioinformáticos en pos de enriquecer la información disponible al usuario, referente tanto a las proteínas como a las estructuras representadas. Esto permite efectuar estudios de correlaciones entre un amplio espectro de factores y la diversidad conformacional. De esta manera es posible correlacionar la extensión de la diversidad conformacional con parámetros tales como función proteica, clasificación estructural, taxonomía, etc, con el fin de comprender ciertos aspectos de la dinámica proteica.

Se listan a continuación las bases de datos y los servidores que han sido utilizados en el presente trabajo.

**CATH**

Todas las estructuras presentes en la PCDB han sido derivadas de la base de datos de dominios proteicos CATH (Alex D Michie, Orengo, & Thornton, 1996).

Se expone bajo estas líneas fundamentalmente la motivación del autor a favorecer CATH sobre otras bases de datos estructurales como punto de partida para la implementación de una base de datos de diversidad conformacional.

El servidor CATH5 es una base de datos de estructuras de dominios proteicos. La base de datos se basa en una clasificación estructural altamente fiable debido a su combinación de criterios tanto estructurales como funcionales, en adición a sus controles y refinaciones efectuadas manualmente. Cada proteína presente en esta base de datos ha sido dividida en dominios, para luego ser asignados a superfamilias de homólogos (esto es, dominios que se encuentran relacionados por la evolución mediante la existencia de un ancestro pasado común a todos ellos). Este procedimiento de clasificación usa una combinación de técnicas automatizadas y manuales que incluyen algoritmos computacionales, empíricos y evidencia estadística, revisión de bibliografía y análisis por parte de especialistas.

La clasificación estructural de los dominios proteicos en CATH se basa en un código de 8 dígitos, en donde cada uno representa uno de los 8 niveles de organización vertical, basados en criterios estructurales, secuenciales y funcionales. Los 8 niveles CATH son detallados en la Tabla 2. La base de datos CATH se caracteriza por su alto grado de depuración, utilizando para tal fin técnicas automatizadas así como también refinaciones manuales basadas en bibliografía específica.

[5] http://www.cathdb.info/

| Clase | Nivel | Criterio |
|---|---|---|
| C | Clase | Determinada por la estructura secundaria y el empaquetamiento dentro de la estructura proteica. Existen tres clases principales: 1) “Mayormente hélices α”; 2) “Mayormente láminas β”; y 3) “Hélices α - láminas β combinadas”. Esta última clase incluye combinaciones de hélices/láminas y de hélices+láminas, como fueron descriptas por Levitt en 1978 (Levitt, 1978). Una cuarta clase contiene una diversidad de estructuras de bajo contenido de estructura secundaria. |
| A | Arquitectura | Describe la forma global de la estructura proteica, determinada por las orientaciones de la estructura secundaria, ignorando las conexiones entre sí. Se asigna manualmente, usando una descripción simple del arreglo de estructura secundaria observado (barril, sándwich de 3 capas, entre otros arreglos). |
| T | Topología (Familia de plegamientos) | Las estructuras son agrupadas en función de si comparten una topología o un plegamiento en el núcleo del dominio. Esto es, si comparten las estructuras secundarias y la conectividad entre ellas dentro de su núcleo estructural. Dominios dentro del mismo nivel pueden presentar diversas decoraciones al núcleo común. Algunas topologías son particularmente pobladas (C. A. Orengo et al., 1997) como las arquitecturas “Mayormente lámina β, sándwich de doble capa” y la “Mayormente hélice α, sándwich de triple capa”. |
| H | Superfamilia de homólogos | Este nivel agrupa dominios proteicos que se suponen descendientes de un ancestro común y pueden por ende ser descriptos como dominios homólogos. Las similitudes son identificadas por una alta similitud secuencial o estructural mediante SSAP. Las estructuras son clasificadas en la misma superfamilia de homólogos si satisfacen uno de los siguientes criterios: valoración<br>1) Identidad secuencial >= 35%, superposición >= 60%.<br>2) Valor SSAP >= 80.0, identidad secuencial >= 20%.<br>3) Valor SSAP >= 70.0, igual función, según base de datos Pfam (Bateman et al., 2004; Finn et al., 2008).<br>4) Similitud significativa mediante búsquedas por HMM-secuencia y comparaciones tipo HMM-HMM usando SAM (Hughey & Krogh, 1996), HMMER (Finn, Clements, & Eddy, 2011) o PRC (Madera, 2008). |
| S | >35% de identidad secuencial | Los dominios dentro de cada Superfamilia de homólogos son clasificados de acuerdo a criterios excluyentes de identidad secuencial. |
| O | >60% de identidad secuencial | |
| L | >95% de identidad secuencial | |
| I | >100% de identidad secuencial | |

Tabla 2. Clases jerárquicas de la base de datos CATH, junto con la descripción del criterio para cada una. Las primeras cuatro clases incluyen criterios estructurales y secuenciales, mientras que las últimas cuatro se basan en criterios exclusivamente secuenciales.

Solo estructuras resueltas con una resolución mejor que 4,0 Å son incluidas en la base de datos CATH, y, consecuentemente, también en la base de datos PCDB. Aquellas estructuras no proteicas, modelos y estructuras con un porcentaje de átomos Cα mayor al 30% en referencia al número total de átomos son excluidas también. Este filtrado se lleva a cabo en CATH mediante el protocolo SIFT (Ng & Henikoff, 2003).

La versión de CATH utilizada para la generación de la base de datos PCDB es la v3.3, la última al momento de su implementación.

**PDB**

De la base de datos PDB6 (Protein Data Bank, H. M. Berman et al., 2000) se ha extraído numerosa información con el fin de completar datos referidos a los dominios presentes en la PCDB.

A saber, la base de datos PDB es el repositorio a nivel mundial de estructuras tridimensionales de moléculas biológicas, incluidas proteínas y ácidos nucleicos. Fue creado en 1971 en el Laboratorio Nacional de Brookhaven, con un total de 7 estructuras depositadas. Actualmente hay alrededor de 74.000 estructuras depositada y es actualizado semanalmente. Cuenta con la creación de la organización Worldwide Protein Data Bank7 (wwPDB) (H. Berman et al., 2003) que trabaja en el procesado y como centro de distribución de la información de la PDB.

La PDB posee información de las estructuras de moléculas biológicas cristalizadas hasta el momento, incluyendo su conformación tridimensional, su secuencia, funciones biológicas asociadas, presencia de ligandos, mutaciones, estado oligomérico, parámetros fisicoquímicos de cristalización, entre otros. Mucha de esta información fue aprovechada para nutrir la base de datos PCDB.

[6] http://www.pdb.org/
[7] http://www.wwpdb.org/

**MAMMOTH**

El programa MAMMOTH (MAtching Molecular Models Obtained from Theory; (Ortiz & Strauss, 2002) es un programa de comparación y alineamiento de estructuras proteicas.

Se basa en un método heurístico para encontrar, en un modo independiente de la secuencia, el subconjunto máximo entre dos proteínas con el mismo esqueleto de carbono y conformación tridimensional. El programa reporta el alineamiento estructural optimizado en formato PDB, conjuntamente con un valor del significado estadístico del alineamiento devuelto, basado en la probabilidad de obtener dicha superposición estructural por azar, cuando dos plegamientos de esa longitud son comparados.

El proceso está altamente optimizado y ha sido utilizado para efectuar el gran número de cómputos (163.040 cálculos) que han debido llevarse a cabo para estudiar la diversidad conformacional entre todas las estructuras representadas en la base de datos PCDB.

En la página de la base de datos se encuentra disponible el alineamiento estructural en formato PDB del par de estructuras que presentan el máximo valor de RMSD entre todos los pares posibles.

También se proporciona el valor máximo, promedio y mínimo de RMSD registrado para cada dominio representado en la base de datos.

**Catalytic site atlas**

La base de datos Catalytic Site Atlas8 (CSA, Porter et al., 2004) es una base de datos con documentación de sitios activos y residuos catalíticos de enzimas. En esta base de datos se encuentra definida una clasificación de residuos catalíticos incluyendo únicamente aquellos involucrados directamente en aspectos de una reacción catalizada por la enzima.

Todas las estructuras presentes en la PCDB se vincularon con la base de datos CSA. Aquellas estructuras que cuentan con información de sitios catalíticos en CSA presentan estos datos en el servidor en línea de la base de datos PCDB. De esta manera, es posible estudiar la diversidad conformacional de los sitios catalíticamente activos.

La información de CSA se basa en dos tipos de datos. En primer lugar, ingresos anotados manualmente, derivados de literatura. Las referencias para cada una de estas entradas son suministradas por la base de datos CSA. En segundo lugar, ingresos homólogos a estos primeros, obtenidos mediante búsquedas tipo PSI-BLAST (Altschul et al., 1997). Una secuencia se considera homóloga cuando su alineamiento secuencial presenta un valor de e-value mejor que 5 x 10-5 respecto a un ingreso previo. De esta manera, los residuos equivalentes, que alinean secuencialmente con los residuos catalíticos en la secuencia original, son registrados como sitios catalíticos.

Es posible acceder a los datos de la base de datos CSA mediante un código PDB, Swiss-Prot9, o un número EC10. Se listan de cada ingreso los sitios catalíticos, usando la numeración de residuos según PDB. Cada sitio es, adicionalmente, etiquetado según “Referencia de literatura” o bien “PSI-BLAST”. En este último caso, es posible a su vez seguir el vínculo hasta la entrada original, basada en literatura.

La versión utilizada de CSA fue la 2.2.12, la última al momento de la implementación de la PCDB, y la última también al momento de la generación del presente documento.

[8] http://www.ebi.ac.uk/thornton-srv/databases/CSA/
[9] http://www.uniprot.org/
[10] http://www.chem.qmul.ac.uk/iubmb/enzyme/

**UniProt**

La base de datos UniProt11 (Jain et al., 2009) es el centro medular de la colección de información funcional de proteínas. Sus datos son conocidos por su precisión, consistencia y riqueza. Se considera en esta base de datos información concerniente a la secuencia aminoacídica, nombre de la proteína, descripción, taxonomía, términos de ontología biológica, indicaciones de la calidad de la anotación, entre otros.

La base de datos PCDB presenta para cada dominio representado el código UniProt (bajo el nombre de Accession number). De esta manera el usuario puede vincular los dominios recuperados en una búsqueda o mediante la navegación de la base de datos con UniProtKB12 para extraer información adicional. Adicionalmente, se vinculó mediante este código información concerniente a localización subcelular, función, a

Esta base de datos consiste en dos secciones, una sección conteniendo ingresos anotados manualmente con información basada en literatura y refinada mediante análisis computacionales (UniProtKB), y una sección con información basada puramente en métodos computacionales (Swiss-Prot).

En pos de reducir la redundancia y maximizar la fidelidad de las secuencias, todas las proteínas codificadas por el mismo gen son agrupadas en una misma entrada UniProt. Diferencias entre diversos secuenciamientos son reportadas y analizadas.

[11] http://www.uniprot.org/
[12] http://www.uniprot.org/help/uniprotkb/

**Enzyme Commission**

Para aquellos dominios asociados a enzimas, el usuario puede extraer de la base de datos PCDB la clasificación de la Enzyme Commission13 (A. Bairoch, 2000). La clasificación EC representa una categorización vertical de reacciones enzimáticas, y son utilizadas comúnmente como identificadores de enzimas o genes de enzimas en el análisis de genomas completos.

El método de asignación del código EC se basa en un esquema de clasificación de reacciones enzimáticas, conocido bajo el nombre de RC (de sus siglas en inglés, Reaction Classification). Cada reacción en la base de datos EC primeramente se descompone en pares reactivos. Cada par es luego alineado estructuralmente para identificar el centro de reacción, las regiones coincidentes y las regiones no coincidentes. El número RC representa los patrones de conversión de los átomos en estas tres regiones. La precisión de este método es del 90%, resultando altamente confiable en la clasificación final.

[13] http://www.chem.qmul.ac.uk/iubmb/enzyme/

**Gene Ontology**

El proyecto Gene Ontology14 (GO) (Ashburner et al., 2000; Consortium, 2008) es una colaboración multidisciplinaria que tiene como objetivo lograr una descripción consistente de los productos génicos a través de numerosas bases de datos.

La base de datos PCDB provee a los usuarios la lista de términos GO asociados a cada uno de los dominios representados, así también como el GO ID (rotulado como GO Accession en la PCDB) asociado al mismo. Esto permite efectuar búsquedas subsecuentes en la base de datos GO en caso de requerir más información acerca de un dominio particular encontrado en PCDB.

Se han desarrollado en el marco de esta colaboración tres vocabularios estructurados que describen los productos génicos en términos de sus procesos biológicos asociados, componentes celulares y funciones de manera dependiente de la taxonomía.

El proyecto comenzó en el año 1998 como una colaboración entre tres bases de datos de organismos (FlyBase15 [de Drosophila]; The Saccharomyce Genome Database16 y Mouse Genome Database17). El número de bases de datos contenidas fue aumentando y, actualmente, las principales bases de datos secuenciales de animales, plantas y microorganismos forman parte en el proyecto GO.

El uso de términos GO por las bases de datos facilita la uniformidad de búsquedas entre ellas. Los vocabularios controlados son estructurados de tal manera que pueden ser consultados a diversos niveles, lo que convierte a los términos GO en un parámetro clave a la hora de búsqueda de información respecto a una proteína de interés, motivo por el cual se incorporó la base de datos GO a la PCDB.

[14] http://www.geneontology.org/
[15] http://flybase.org/
[16] http://www.yeastgenome.org/
[17] http://www.informatics.jax.org/

**PQS**

El servidor Protein Quaternary Structure18 (Henrick & Thornton, 1998), mantenido por el grupo Macromolecular Structure Database (MSD) del European Bioinformatics Group19 (EBI), provee al usuario de una herramienta de análisis de la unidad biológica para una entreada PDB. En el presente trabajo se vincularon las estructuras de la PCDB con el servidor PQS a fin de obtener información acerca del estado oligomérico de las mismas.

El estudio del estado cuaternario es importante en la comprensión de la función biológica de una proteína. El servidor PQS es un recurso disponible en línea que presenta las coordenadas de los estados cuaternarios para las estructuras contenidas en el Brookhaven Protein Data Bank (PDB).

De la cristalografía de una macromolécula se obtiene un conjunto de coordenadas que no son independientes de la simetría cristalográfica. Las coordenadas depositadas describen los átomos necesarios para el refinamiento de los datos observados experimentalmente. Sin embargo, estas coordenadas no necesariamente describen la molécula completa en estudio, o pueden incluir múltiples copias de la molécula. En una entrada PDB las coordenadas depositadas generalmente consisten en el contenido de la unidad asimétrica (ASU), es decir, la fracción de la celda de la unidad cristalográfica que no tiene simetría cristalográfica. El servidor PQS dispone de un procedimiento automatizado que ha sido diseñado para reconocer la existencia de varias copias en el cristal, o aquellos casos en donde la simetría es necesaria para generar una lista de puntos que describen la totalidad de la macromolécula en estudio. Hay varias posibilidades relación entre las coordenadas cristalográficas depositados únicos y la macromolécula en estudio:

(I) El contenido de la ASU define una sola copia de la macromolécula.

(Ii) El contenido de la ASU consiste en más de una copia de la macromolécula.

(Iii) Los contenidos de la ASU requieren operaciones cristalográficas de simetría que deben aplicarse a la generación de la macromolécula completa.

(Iv) Una combinación de los anteriores, incluidas las copias múltiples y las transformaciones de simetría.

Para cada estructura PDB, todos los contactos atómicos entre residuos a una distancia menor a 3,7 Å entre todas las cadenas, incluyendo cada uno de las 27 posibles traducciones de la celda unidad por operación de simetría, son calculados

[18] http://www.ebi.ac.uk/pdbe/pqs/
[19] http://www.ebi.ac.uk/

utilizando los algoritmos CCP4 (Collaborative_Computational_Project, 1994) y WHATIF (Vriend, 1990). Un conjunto cuaternario potencial es construido en base a la adición progresiva de las cadenas monoméricas que se consideran para contribuir al ensamblado. La selección de la cadena se basa en el número de contactos entre las cadenas que se encuentran y el número de residuos en cada cadena. Este algoritmo es recursivo, permitiendo la detección de las estructuras cuaternarias donde los contenidos de la ASU no están en contacto con todos los elementos de simetría relacionados en el montaje final. Todos los conjuntos de potenciales están disponibles en el servidor con la anotación asignada automáticamente.

Posteriormente, en una segunda fase, se determina si los contactos proteicos encontrados son específicos (un oligómero macromolecular verdadero) o no específicos (cristal packing). Las caracterizaciones estructurales de las interfaces proteína-proteína han estudiadas ampliamente estudiadas y se han relacionado con muchas propiedades: hidrofobia, análisis estructurales, preferencias de los residuos, número de enlaces de hidrógeno, etc. Para discriminar entre contactos no específicos (cristal packing) y una interacción proteína-proteína funcional, el valor del área superficial accesible al solvente (ASA) es ampliamente utilizado. El valor de Δ(ASA) fue estimado para cada complejo potencial cuaternario, y se calculó la diferencia media sobre el número de cadenas que se asocian para formar el conjunto completo. Para aquellas entradas PDB de las que se tiene información acerca del estado cuaternario, la diferencia de superficie accesible por cadena oscila entre los 370 y los 4750 Å2 para homo-dímeros y entre los 640 y los 3230 Å2 para los hetero-dímeros (S. Jones & Thornton, 1995). El servidor aquí descripto utiliza un punto de corte de Δ(ASA) de 400 Å2 por cadena para clasificar a los interacciones proteína-proteína. El proceso completo de discriminación utiliza una puntuación empírica ponderada con aportes del valor de Δ(ASA), el número de residuos enterrados en la interfaz, la energía de solvatación, el número de puentes salinos presentes en la interfase y la presencia de puentes de tipo di-sulfuro.

**Procognate**

La base de datos Procognate20 (Bashton, Nobeli, & Thornton, 2008) es una base de datos de ligandos asociados a los dominios estructurales de enzimas representadas en CATH (Greene et al., 2007), SCOP21 (Structural Classificarion Of Proteins) (Lo Conte, Brenner, Hubbard, Chothia, & Murzin, 2002) y Pfam22 (Protein Family) (Finn et al., 2008). Procognate presenta un alto grado de depuración de información, mediante técnicas tanto automáticas como manuales, y fue elegida para integrar la PCDB debido a la confiabilidad y exactitud de sus datos.

Los ligandos aquí presentados han sido identificados mediante información extraída de las bases de datos ENZYME23 (A. Bairoch, 2000) y KEGG24 (M Kanehisa & Goto, 2000; Minoru Kanehisa, Goto, Kawashima, & Nakaya, 2002) y comparados con los ligandos extraídos de la base de datos PDB mediante coincidencias de grafo para evaluar su similitud química. Aquellos ligandos involucrados en reacciones conocidas en ENZYME y KEGG para una enzima en particular son asignados a estructuras ya referidas a un código EC.

Se han vinculado las estructuras presentes en la PCDB con la base de datos Procognate a fin de obtener información acerca del tipo de ligando presente en cada entrada de la PCDB.

[20] http://www.ebi.ac.uk/thornton-srv/databases/procognate/
[21] http://scop.mrc-lmb.cam.ac.uk/scop/
[22] http://pfam.sanger.ac.uk/
[23] http://expasy.org/enzyme/
[24] http://www.genome.jp/kegg/

**Capítulo 3**

**Servidor en línea de la base de datos PCDB**

### 3.1 Funcionalidad

El servidor de la base de datos PCDB25 fue diseñado para reclutar ejemplos de dominios proteicos que exhiban diversidad conformacional y relacionar la misma con propiedades fisicoquímicas y/o biológicas.

Es posible restringir los resultados de la búsqueda en función de una variedad de parámetros: extensión de la diversidad conformacional, diferencias en las condiciones de cristalización de las distintas estructuras representantes del dominio proteico (pH, ligandos), etc. La búsqueda puede ser también limitada mediante un código PDB o por una clasificación de dominios CATH, para reclutar solo las entradas de la PCDB que coincidan con la clasificación ingresada. El campo para completar las restricciones se encuentra en la parte superior de la solapa “Search”.

Asimismo, es posible efectuar una búsqueda sin restricciones, cuyo resultado será la base de datos completa. El usuario puede descargar la información presentada en los resultados de las búsquedas en formato .csv para un posterior tratamiento de los datos

La página de inicio del servidor de la base de datos se exhibe en la .Figura 2.

[25] http://www.pcdb.unq.edu.ar/

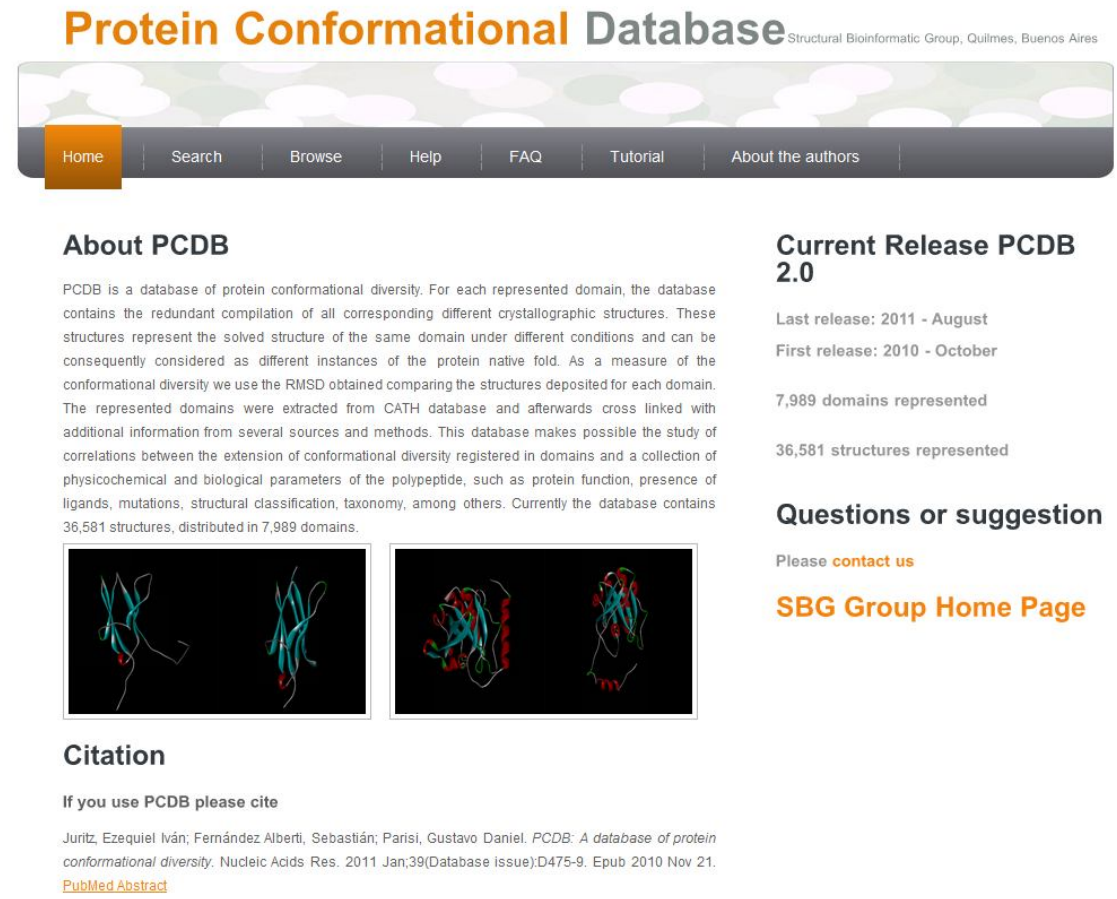



**Figura 2. Página de inicio del servidor de la base de datos PCDB.**

Adicionalmente a las búsquedas, otra modalidad de utilizar la base de datos consiste en la navegación a través de la misma. Para esto se sigue la clasificación estructural CATH (Whitney, 1997) de los dominio pasando por los 8 niveles de jerarquía. A medida que el usuario seleccione un determinado nivel (C, A, T, H, S35, O60, L95, I100) se mostrarán los posibles niveles inmediatamente inferiores. Cada uno de los niveles inferiores se presentará como un hipervínculo que el usuario puede utilizar para ingresar al mismo. A su vez, se listarán también todos los dominios pertenecientes a la clasificación actual, junto con una serie de datos relevantes, como familia proteica a la que pertenece el dominio, el RMSD máximo registrado entre sus pares de estructuras, y la clasificación estructural CATH completa.

Se detallan a continuación los diversos métodos y opciones para restringir las búsquedas.

### 3.2 Efectuar una búsqueda en la base de datos PCDB

Una de las maneras de utilizar el servidor en línea de la base de datos PCDB consiste en efectuar una búsqueda de dominios representados en la misma, que cumplan con ciertos requisitos definidos por el usuario. Es posible delimitar las búsquedas por medio de combinaciones de un amplio espectro disponible de parámetros y criterios. La página de búsqueda de la base de datos tal cual como se encuentra en línea de muestra en la Figura 3. Uno de los posibles criterios para limitar las búsquedas efectuadas, por ejemplo, es por la extensión de la diversidad conformacional exhibida por el dominio. Esto es, el valor de RMSD presentado por los pares de estructuras acumuladas de tal dominio. Esta función puede ser de utilidad, por ejemplo, para el usuario que se encuentre interesado en estudiar casos extremos de diversidad conformacional, que superen determinado valor de RMSD. De esta manera puede efectuar una búsqueda estableciendo como valor mínimo de diversidad conformacional 7 Å, por ejemplo. Una búsqueda delimitada por un mínimo de 7 Å resultará en 99 dominios cuya diversidad conformacional supera dicho valor. De manera similar, es posible limitar los resultados a aquellos dominios que expresen una diversidad conformacional menor a cierto margen. Se pueden limitar los dominios arrojados también por un código PDB [1abc o 1abcA] o CATH [1abcA00, compuesto por el código PDB, y en donde el quinto carácter representa la cadena y el número final el número de dominio] que sea de interés por el usuario, o por una clasificación estructural CATH deseada. La clasificación estructural CATH se compone de 8 dígitos, cada uno asociado a cierta jerarquía vertical, y es tratada con mayor profundidad más adelante.

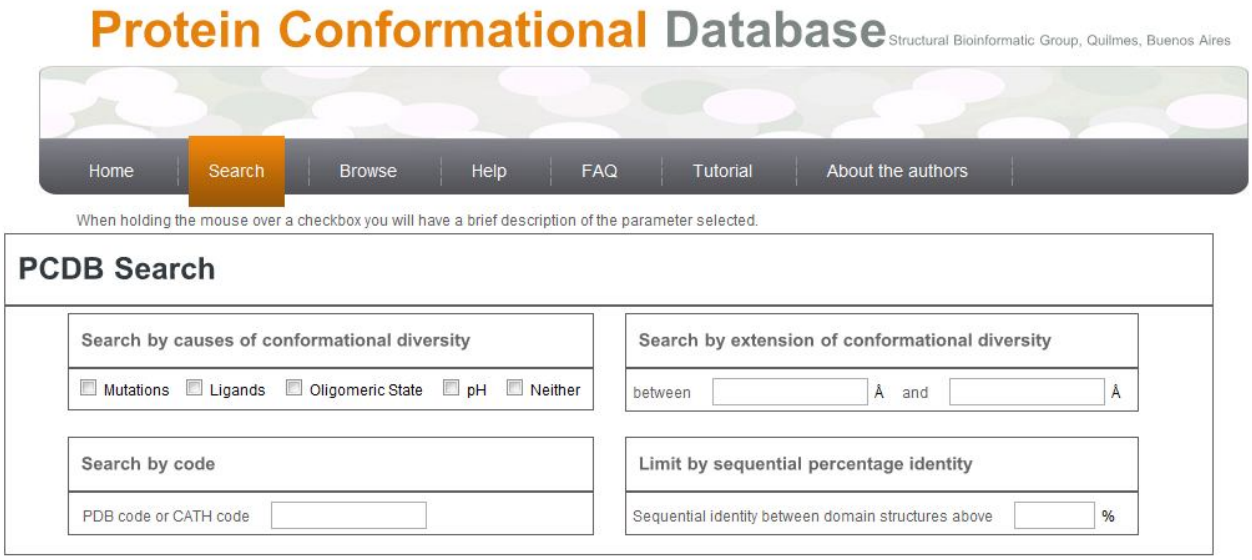



**Figura 3. Página de búsqueda del servidor de PCDB. Se exhiben aquí las diversas formas en que el usuario puede limitar una búsqueda sobre la base de datos.**

### 3.2.1 Limitar por posibles causas de diversidad conformacional.

Una de los métodos de filtrar los resultados de una búsqueda se basa en las condiciones de cristalización de las diversas estructuras representantes de los dominios proteicos. Las condiciones comprendidas actualmente por PCDB comprenden presencia o ausencia de mutaciones, presencia o ausencia de ligandos, diferencias en el estado oligomérico de la proteína cristalizada y diferencias en el pH de la cristalización. Esta opción está orientada a recuperar aquellos dominios que han sido cristalizados varias veces en condiciones homogéneas o bien heterogéneas para estos parámetros citados.

A modo de ejemplo, si uno está interesado en estudiar la diversidad conformacional proteica causada por cambios en el estado oligomérico, debe activar el cuadro “Oligomeric State” dentro del campo “Search by causes of conformational diversity” en la solapa “Search”. De esta manera, PCDB desplegará en el resultado de la búsqueda sólo aquellos dominios en donde existen estructuras cristalizadas en distintos estados oligoméricos. Las otros cuadros disponibles en el campo “Search by causes of conformational diversity” son “Mutations”, “Ligands”, “pH” y “Neither”. El cuadro “Mutations”, en caso de estar activado, limitará el resultado a aquellos dominios proteicos en donde haya al menos una estructura cristalizada con alguna mutación y otra sin ninguna mutación o con una mutación. El cuadro “Ligands” en caso de estar

activado, limitará el resultado a aquellos dominios en donde haya al menos una estructura que haya sido cristalizada con ligando y una estructura que haya sido cristalizada sin ligando. El cuadro “pH”, al activarse, limitará el resultado a aquellos dominios representados en PCDB en donde haya al menos dos estructuras cristalizadas con una diferencia de más de 1 (una) unidad de pH. El cuadro “Neither”, por otro lado, al activarse, limita el resultado sólo a aquellos dominios en donde las estructuras han sido cristalizadas bajo idénticas condiciones de los parámetros citados.

La descripción de cada uno de estos campos se resume en la Tabla 3.

| Cuadro | Efecto |
|---|---|
| “Mutations” | Limita a aquellos dominios proteicos cristalizados tanto con como sin mutaciones |
| “Ligands” | Limita a aquellos dominios proteicos cristalizados tanto con como sin ligando |
| “Oligomeric State” | Limita a aquellos dominios proteicos cristalizados en diversos estados oligoméricos |
| “pH” | Limita a aquellos dominios proteicos cristalizados en diferentes condiciones de pH, presentando una diferencia igual o mayor a 1 unidad de pH |
| “Neither” | Limita a aquellos dominios proteicos cristalizados en idénticas condiciones de los parámetros citados |

Tabla 3. Factores influyentes en la diversidad conformacional que son considerados en la base de datos PCDB.

### 3.2.2 Limitar por extensión de la diversidad conformacional.

Es posible restringir los dominios reclutados por la PCDB en una búsqueda en función de la amplitud de la diversidad conformacional registrada en el mismo. La amplitud de la diversidad conformacional de una proteína está definida, como ya se ha explicado, mediante los valores de RMSD entre las distintas estructuras pertenecientes a dicha proteína. Específicamente, la diversidad conformacional de una proteína en la PCDB está determinada por el máximo valor de RMSD registrado entre pares de sus estructuras.

En el campo “Search by extention of conformational diversity” de la solapa “Search” es posible definir el rango de extensión de diversidad conformacional fijando el valor mínimo y el valor máximo de RMSD, en unidades de Å. En caso de dejar algún campo vacío (mínimo o máximo), esta variable no será restringido a ningún valor. Estos valores se aplican al RMSD máximo registrado entre las estructuras de los correspondientes a los dominios representados.

### 3.2.3 Limitar por código PDB o por código CATH.

Dentro del campo “Search by code” es posible especificar un código PDB (1abc) o un código CATH (1abcA00) para recuperar aquellos dominios de la PCDB en donde alguna estructura coincide con el código buscado.

### 3.2.4 Limitar por código de clasificación estructural CATH.

En el campo “Limit search by CATH structural classification” es posible limitar los dominios expuestos en los resultados de la búsqueda de acuerdo a cualquier combinación de los 8 niveles de clasificación estructural de dominios de la base de datos CATH. Se encuentran en este campo los 8 cuadros correspondientes a los 8 niveles de clasificación estructural CATH (C; A; T; H; S35; O60; L95; I100). Estos campos pueden ser completados con la identificación de un nivel estructural CATH para limitar los resultados a los dominios que corresponden ese nivel estructural. Estos campos pueden completarse en cualquier combinación, aunque lo más frecuente es limitar solo entre el primer nivel (C) y el cuarto nivel (H).

### 3.2.5 Información de salida.

En el campo “Format Output Information”, dentro de la solapa “Search”, se enumera la información que contiene la PCDB sobre cada uno de los dominios representados en la base de datos. Esta información se ha recopilado mediante vinculaciones con numerosas bases de datos y métodos de biología computacional, enumerados y descriptos en el capítulo 2. Activando y desactivando cada cuadro, es posible personalizar el conjunto de datos que cada usuario se encuentra interesado en extraer de los dominios que resultarán de la búsqueda a efectuar.

El campo está dividido temáticamente a fin de facilitar la navegación visual del usuario, como se aprecia en la Figura 4.

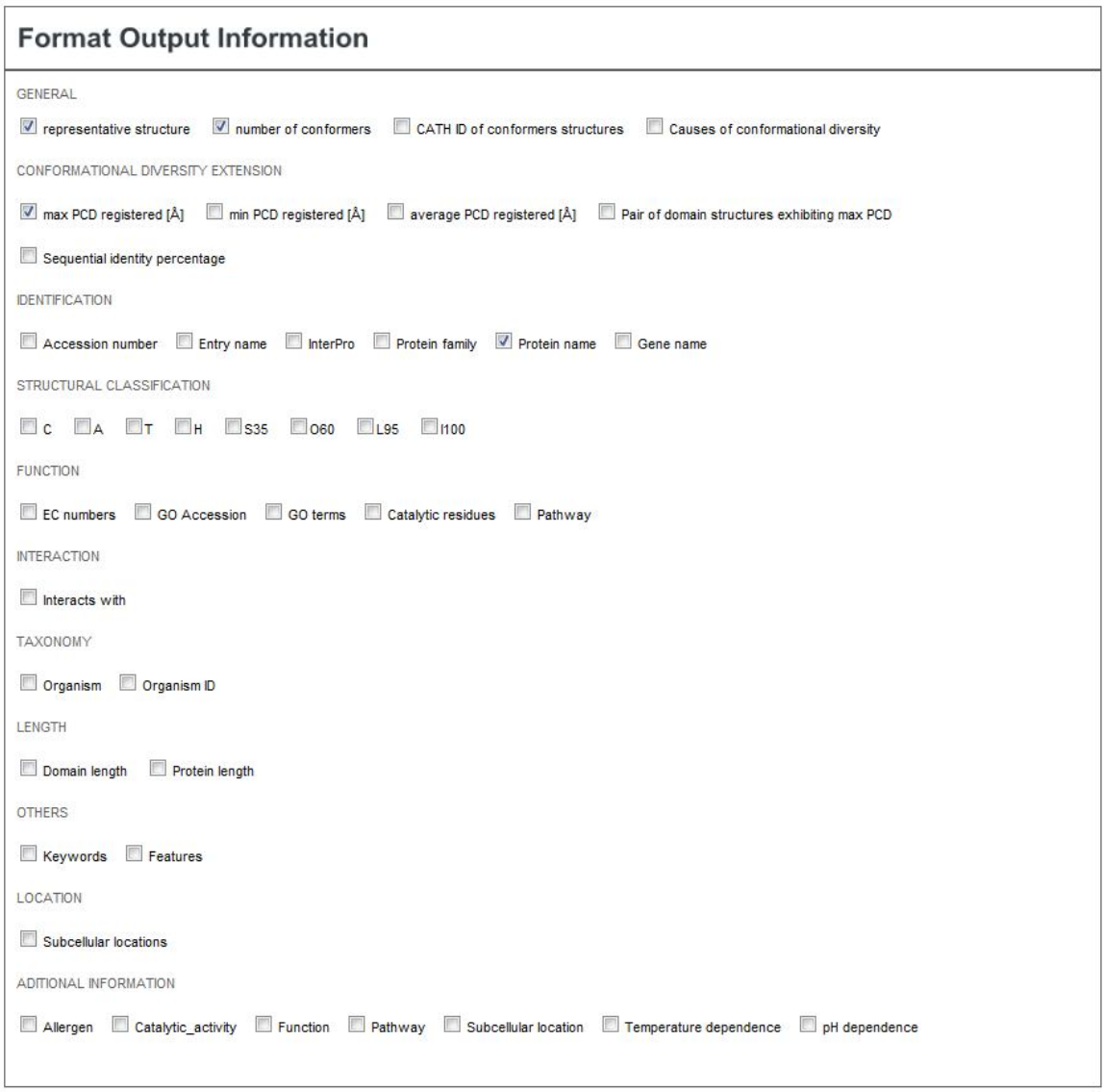



Figura 4. Página en donde se da formato a la salida de una búsqueda. Se listan aquí los datos disponibles al usuario para los dominios que resulten de una búsqueda. Adicionalmente, muchos de estos datos son identificaciones de otras bases de datos, con lo que es posible vincularlas a fin de profundizar sobre un dominio en particular.

### 3.2.6 Enviar la búsqueda.

Una vez establecidos los parámetros de la búsqueda a realizar, tanto las restricciones de los parámetros en los que el usuario está interesado y la información de salida que desea compilar, al presionar el botón de “Search” de la pestaña del mismo nombre a fin de dar inicio a la búsqueda. El botón se encuentra en la parte inferior de la pestaña, y se encuentra destacado con un círculo de color rojo en la Figura 5.

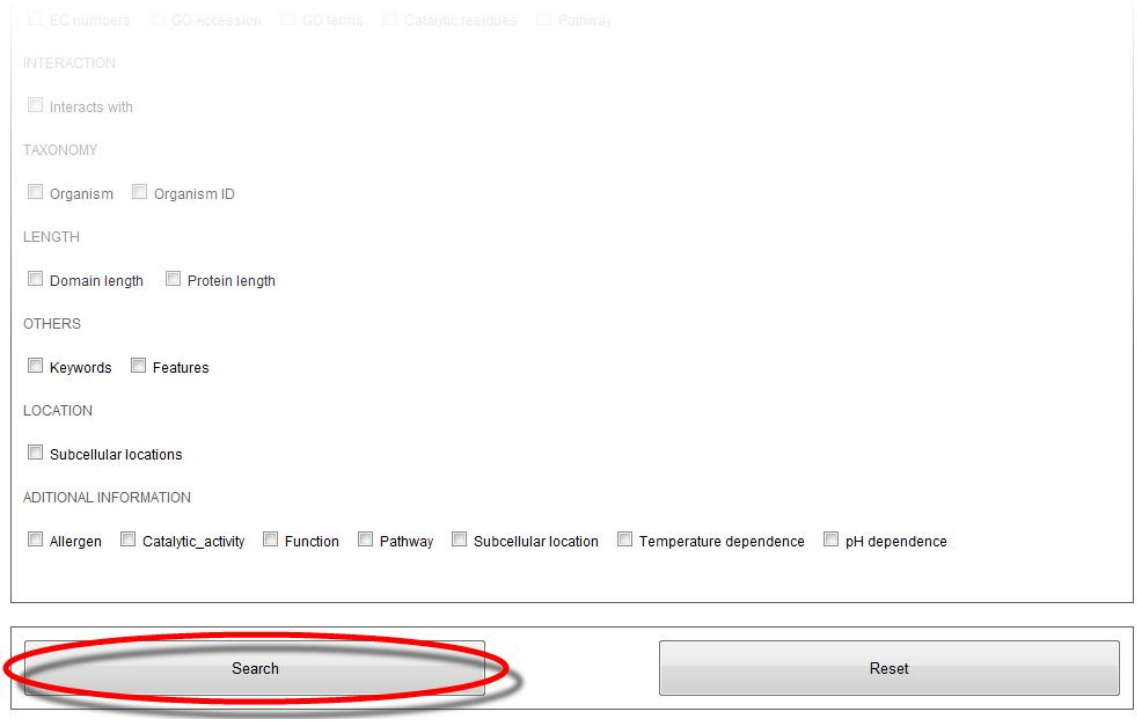



Figura 5. Se destaca con un óvalo de color rojo la ubicación en la página de “Format output” del botón “Search”, cuya función es enviar la búsqueda a la base de datos, definida con los parámetros previamente fijados.

### 3.2.7 Resultados de la búsqueda.

En la página siguiente se presentará, en formato de tabla de doble entrada, la información requerida de aquellos dominios que coinciden con los parámetros de búsqueda previamente definidos. En la Figura 6 se presenta a modo de ejemplo el resultado de una búsqueda sin restricciones y con la información de salida predeterminada.

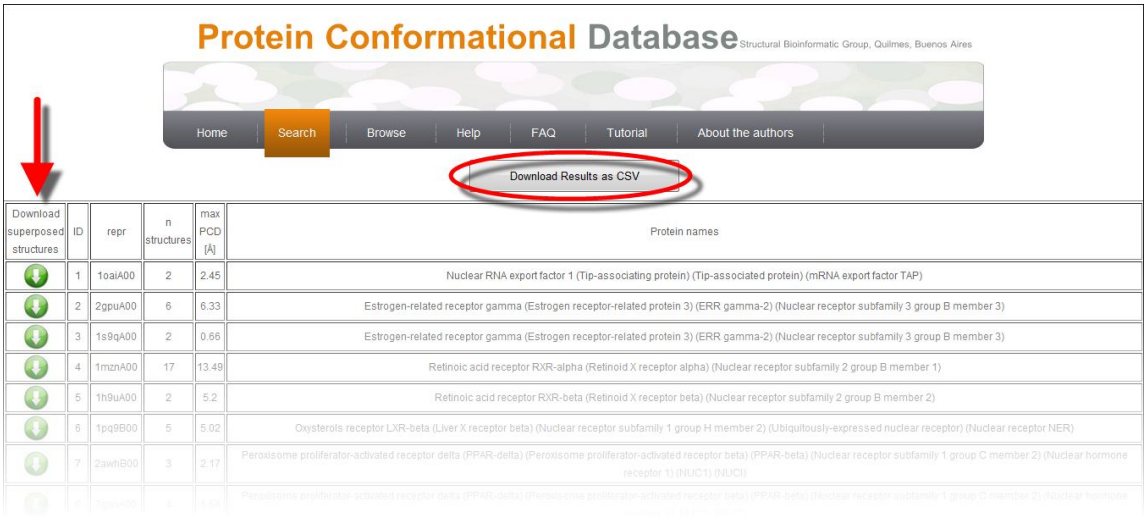



Figura 6. Ejemplo de los resultados de una búsqueda en la PCDB. En este caso, los resultados aquí mostrados se obtuvieron mediante una búsqueda sin restricciones, y los datos de salida solicitados son aquellos que vienen seleccionados por defecto en la página. Se destaca con un óvalo de color rojo el botón para descargar los resultados como una tabla de doble entrada en formato CSV (comma separated value). Con una flecha de color rojo se indica la columna con los vínculos que permiten descargar el alineamiento entre el par de estructuras que presentan el mayor valor de RMS. El alineamiento de descarga en formato PDB.

Dos posibilidades adicionales presenta la página correspondiente a la presentación de los resultados. La primera consiste en descargar bajo formato .csv (“comma separated values”) un archivo con la información solicitada de los resultados expuestos. Esto se efectúa mediante un botón situado por encima de la tabla exhibida, con la leyenda “Download Results as CSV”. Dicho botón se destaca en la Figura 6 mediante un círculo de color rojo.

La segunda, consiste en que el usuario puede descargar a un archivo el alineamiento estructural del par de estructuras que presentan el máximo RMSD entre todas las estructuras dadas de cualquier dominio. El formato del archivo a descargar es PDB, y el alineamiento estructural es derivado del programa Mammoth. Esto se efectúa

mediante un botón de color verde disponible para cada dominio presentado al usuario, situado en la primera columna hacia la izquierda de la tabla presentada, con el encabezado de “Download superposed structures”. Dicha columna se encuentra destacada con una flecha de color rojo en la Figura 6.

### 3.3 Navegación por la base de datos PCDB.

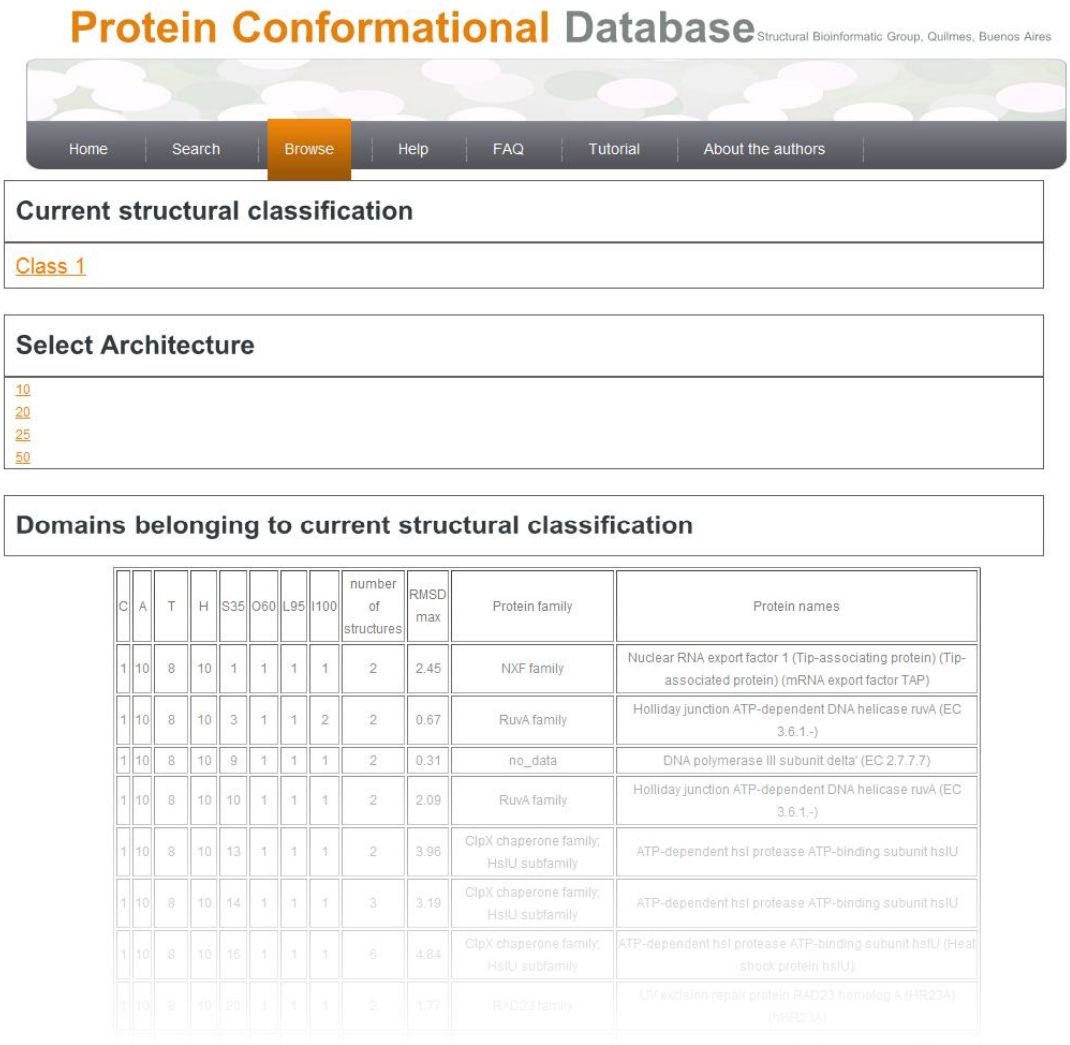
Existen dos maneras de utilizar la base de datos PCDB. Una es efectuando una búsqueda de acuerdo a algún criterio o combinación de criterios, como ya se ha explicado en la sección anterior. La segunda forma consiste en navegar por la base de datos, siguiendo la estructura de 8 niveles de la clasificación CATH. La pestaña “Browse” de la base de datos PCDB permite esta funcionalidad. Una vez seleccionada esta pestaña, el primer paso consiste en elegir una de las 4 Clases de la clasificación CATH (nivel C). Las 4 Clases de CATH son:

1. Mayormente Hélices α.
2. Mayormente Láminas β.
3. Hélices α - Láminas β Combinados.
4. Estructuras secundarias restantes.

Una vez elegida alguna de las 4 Clases, se expondrán 3 campos: “Current structural classification” (“Clasificación estructural actual”), “Select Architecture” (“Seleccionar Arquitectura”) y “Domains belonging to current structural classification” (“Dominios pertenecientes a la clasificación estructural actual”). En la Figura 7 se presenta un recorte de pantalla de la pestaña “Browse” luego de seleccionar la Clase 1. La información que provee cada uno de los campos presentados se detalla a continuación.

### 3.3.1 Clasificación estructural.

Este campo indica la ubicación actual del usuario dentro de la base de datos PCDB. En la medida que el usuario navega por la base de datos PCDB, este campo se actualizará indicando el nivel estructural CATH que se está exponiendo en pantalla. En el ejemplo de la Figura 7, la clasificación en donde el usuario se encuentra es “Clase 1”.




| C | A | T | H | S35 | O60 | L95 | I100 | number of structures | RMSD max | Protein family | Protein names |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 8 | 10 | 1 | 1 | 1 | 1 | 2 | 2.45 | NXF family | Nuclear RNA export factor 1 (Tip-associating protein) (Tip-associated protein) (mRNA export factor TAP) |
| 1 | 10 | 8 | 10 | 3 | 1 | 1 | 2 | 2 | 0.67 | RuvA family | Holliday junction ATP-dependent DNA helicase ruvA (EC 3.6.1.-) |
| 1 | 10 | 8 | 10 | 9 | 1 | 1 | 1 | 2 | 0.31 | no_data | DNA polymerase III subunit delta' (EC 2.7.7.7) |
| 1 | 10 | 8 | 10 | 10 | 1 | 1 | 1 | 2 | 2.09 | RuvA family | Holliday junction ATP-dependent DNA helicase ruvA (EC 3.6.1.-) |
| 1 | 10 | 8 | 10 | 13 | 1 | 1 | 1 | 2 | 3.96 | ClpX chaperone family; HslU subfamily | ATP-dependent hsl protease ATP-binding subunit hslU |
| 1 | 10 | 8 | 10 | 14 | 1 | 1 | 1 | 3 | 3.19 | ClpX chaperone family; HslU subfamily | ATP-dependent hsl protease ATP-binding subunit hslU |
| 1 | 10 | 8 | 10 | 16 | 1 | 1 | 1 | 6 | 4.84 | ClpX chaperone family; HslU subfamily | ATP-dependent hsl protease ATP-binding subunit hslU (Heat shock protein hslU) |
| 1 | 10 | 8 | 10 | 20 | 1 | 1 | 1 | 2 | 1.77 | RAD23 family | UV excision repair protein RAD23 homolog A (HR23A) (hHR23A) |

Figura 7. Página de navegación de la PCDB, cuando el usuario ha seleccionado la Clase 1 (Mayormente Hélices α).

### 3.3.2 Seleccionar Arquitectura.

En este campo se listan los niveles inmediatamente inferiores disponibles en la base de datos PCDB. El usuario puede elegir uno de los niveles aquí expuestos para continuar con la navegación sobre la base de datos.

En el caso del presente ejemplo la página listará las Arquitecturas (Nivel A) para la Clase 1 (Nivel C) elegida por el usuario. Una vez elegida una de las Arquitecturas listadas, el rótulo cambiará a “Select Topology”, dado que la base de datos estará listando las Topologías (Nivel T) disponibles dentro de la Arquitectura elegida por el usuario.

### 3.3.3 Dominios pertenecientes a la clasificación estructural actual.

Dentro de este campo se enumeran todos los dominios comprendidos bajo la clasificación actual. De los dominios listados se detalla la siguiente información: los 8 niveles de clasificación estructural CATH a la que pertenece el dominio, el número de estructuras disponibles en la base de datos PCDB para el dominio listado, el valor máximo de RMSD computado entre los pares de estructuras, la familia de proteínas a la que pertenece el dominio y los nombres de las proteínas asociadas. Se describen brevemente estos campos en la

Tabla 4.

| Campo | Información suministrada |
|---|---|
| C | Clase a la que pertenece el dominio listado |
| A | Arquitectura a la que pertenece el dominio listado |
| T | Topología a la que pertenece el dominio listado |
| H | Superfamilia de homólogos a la que pertenece el dominio listado |
| S35 | Nivel S35 al que pertenece el dominio listado |
| O60 | Nivel O60 al que pertenece el dominio listado |
| L95 | Nivel L95 al que pertenece el dominio listado |
| I100 | Nivel I100 al que pertenece el dominio listado |
| number of structures | Número de estructuras disponibles en PCDB del dominio listado |
| RMSD max | RMSD máximo registrado entre par de estructuras del dominio listado |
| Protein family | Familia proteica a la que pertenece el dominio listado |
| Protein names | Nombres de proteínas asociadas al dominio listado |

Tabla 4. Información presentada al usuario de todos los dominios que pertenecen a la clasificación actual cuando se ejecuta la opción de navegación de la PCDB.

## Capítulo 4

## Diversidad conformacional en la base de datos PCDB.

### 4.1 Introducción

A fin de evaluar la extensión de la diversidad conformacional de los dominios proteicos representados en la base de datos PCDB, se procedió a calcular los valores de Root-Mean-Square Deviation (RMSD) entre cada par de estructuras representantes del mismo dominio proteico. Para tal fin se utilizó el programa Mammoth (MAtching Molecular Models Obtained from Theory; (Ortiz & Strauss, 2002), diseñado para encontrar la máxima superposición entre dos estructuras proteicas. Este programa calcula el valor de RMSD mínimo entre dos estructuras y guarda el alineamiento estructural correspondiente en formato PDB. El MAMMOTH fue utilizado aplicando los parámetros definidos por defecto, y es descripto con mayor profundidad en otro capítulo. Se aplicó el algoritmo del MAMMOTH a todos los pares posibles de estructuras relacionadas al mismo dominio. La base de datos consta de 7.989 proteínas, con un total de 36.581 estructuras y el número de cómputos de RMSD llevados a cabo asciende a un total de 163.040 cálculos. Estos cálculos fueron realizados en el cluster de cálculos del grupo de trabajo Structural Bioinformatic Group de la Universidad Nacional de Quilmes. Todos los archivos generados por el programa, incluyendo la superposición del par de estructuras, fueron almacenados. Aquellos alineamientos que presentan la máxima divergencia estructural por dominio se encuentran disponibles en el servidor en línea para ser descargados por los usuarios de la base de datos en formato PDB.

El mínimo valor de RMSD registrado en la PCDB es de 0,0 Å, y el máximo presenta un valor de 26,7 Å. El promedio global de RMSD entre los pares de estructuras pertenecientes a la misma proteína es 1,3 Å. Estos valores se resumen en la Tabla 5.

| | |
|---|---|
| Mínimo RMSD registrado | 0,0 Å |
| Promedio RMSD registrado | 1,3 Å |
| Máximo RMSD registrado | 26,7 Å |

Tabla 5. Estadísticas de los valores de RMSD registrados entre pares de estructuras del mismo dominio.

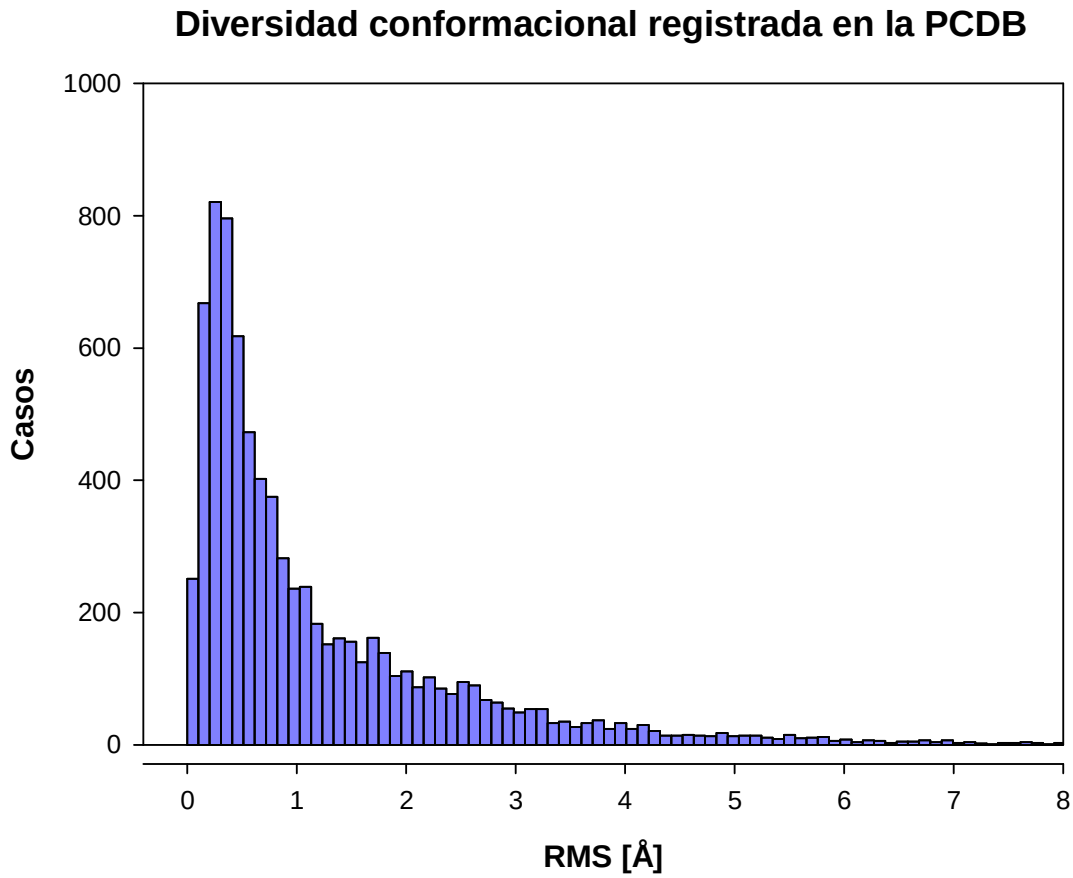



Figura 8. Distribución de RMSD entre pares de estructuras de los dominios proteicos representados en la base de datos PCDB.

La distribución de valores de RMSD entre los pares de estructuras que componen los dominios representados en la PCDB se expone en la Figura 8. Se aprecia una tendencia hacia valores relativamente bajos de RMSD, algo esperable tratándose de comparaciones de la misma secuencia. Aún así, se distinguen una gran proporción de casos que exhiben de mediana a alta diversidad conformacional: un 40% de los dominios estudiados presentan un RMSD mayor a 1 Å, y un 20% presenta un valor mayor a 2 Å. Más de un 5% presenta un RMSD mayor a 4 Å. Basándonos en estudios previos (Burra et al., 2009), consideramos que dos estructuras presentan diversidad conformacional cuando su RMSD supera el valor de 1 Å.

### 4.2 Diversidad conformacional bajo condiciones homogéneas y heterogéneas.

Con el fin de explicar la diversidad conformacional observada en la base de datos PCDB, se efectuaron análisis de correlación entre diversos parámetros para estudiar la incidencia de éstos en la extensión de la diversidad conformacional.

De esta manera, las estructuras contenidas en la base de datos PCDB se encuentran vinculadas con ciertos datos de la proteína y parámetros de las condiciones de cristalización: presencia o ausencia de mutaciones; presencia o ausencia de ligando/s; estado oligomérico y pH de la cristalización. Dentro de las mejoras que estamos realizando actualmente, se incluye incrementar el número de estos datos a considerar para ampliar las posibilidades de estudio. Adicionalmente, cada dominio representado está asociado a un valor que indica si las estructuras que lo representan presentan valores idénticos o diferentes para cada uno de los datos mencionados. En caso de que las estructuras representantes de un domino hayan sido cristalizadas todas en el mismo estado oligomérico, por ejemplo, el dominio será homogéneo para ese parámetro. Un dominio puede obviamente presentar valores homogéneos para ciertos datos y heterogéneos para otro. Si un dominio presenta valores homogéneos para todos los datos, significa que todas las estructuras presentes en la base de datos han sido cristalizadas bajo las mismas condiciones con respecto a los parámetros aquí considerados. Es posible efectuar búsquedas en la base de datos limitando para heterogeneidad en cualquier combinación de estos datos, como se explica en el capítulo 3.

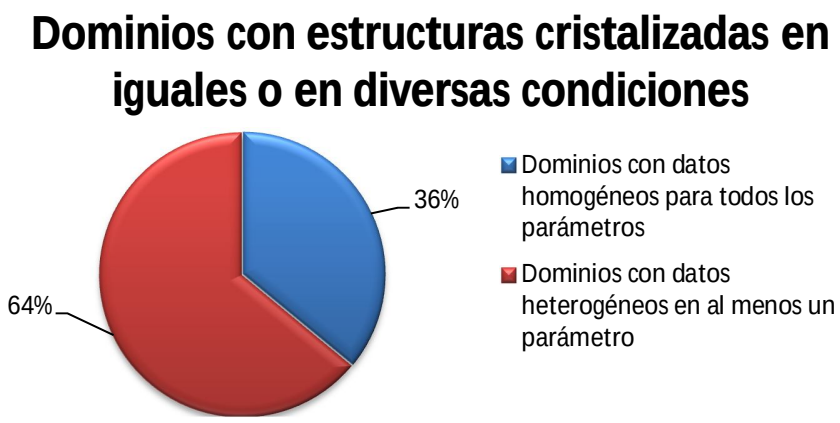



**Figura 9. Porcentaje de dominios representados en la PCDB, compuestos por estructuras cristalizadas en varias instancias bajo las mismas condiciones (en color azul oscuro) o bajo condiciones diversas (en color rojo claro).**

Al discriminar los dominios representados en la PCDB en función de la homogeneidad o heterogeneidad de los parámetros estudiados, se aprecia una divergencia entre aquellos dominios cuyas estructuras derivan de instancias de cristalizaciones en condiciones diversas, con respecto a aquellos dominios cuyas estructuras provienen de cristalizaciones en condiciones idénticas. Coincidentemente con lo esperado, un mayor grado de diversidad conformacional se expone en aquellos dominios con valores heterogéneos para algún parámetro. Este dato es de gran trascendencia, y representa una de las relaciones fundamentales a ser estudiadas mediante la PCDB. La diversidad conformacional se exhibe en la estructura nativa de una proteína, y está definida por fluctuaciones dinámicas de su estructura. A su vez, al efectuar cambios en ciertas condiciones (pH, presencia de ligandos, estado oligomérico, entre otros) la estructura nativa de la proteína se reacomoda para adaptarse a las nuevas condiciones. Estos reacomodos se traducen en alteraciones de las concentraciones relativas entre confórmeros presentes en el estado nativo, así también como la aparición de nuevos confórmeros y la desaparición de otros (Bernardes et al., 2012; Bocharov et al., 2008; Laine et al., 2011).

Como se expone en la Figura 9, de los 7.989 dominios representados en la base de datos PCDB, el 36% (2.875 dominios) poseen datos homogéneos en todos los parámetros evaluados (mutaciones, presencia de ligandos, estado oligomérico y pH de cristalización). Esto es, todas las estructuras de estos dominios han sido cristalizadas en idénticas condiciones para estos parámetros. El restante 64% (5.114 dominios) poseen

estructuras que han sido cristalizadas en diversas condiciones para al menos uno de estos parámetros.

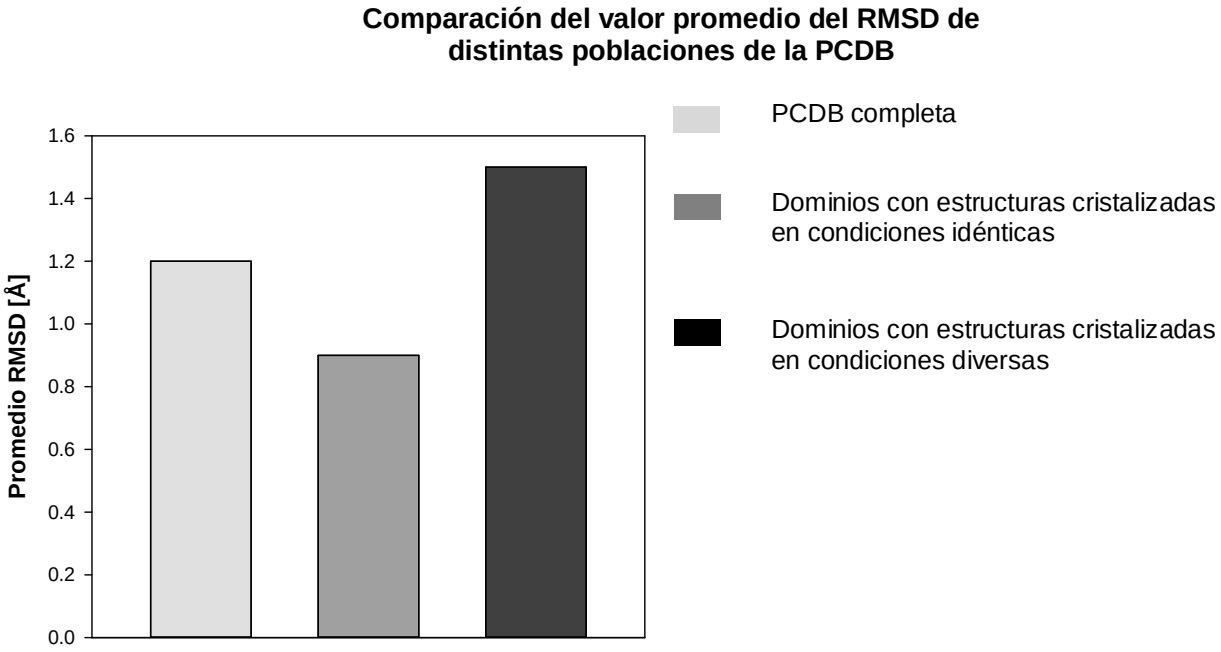
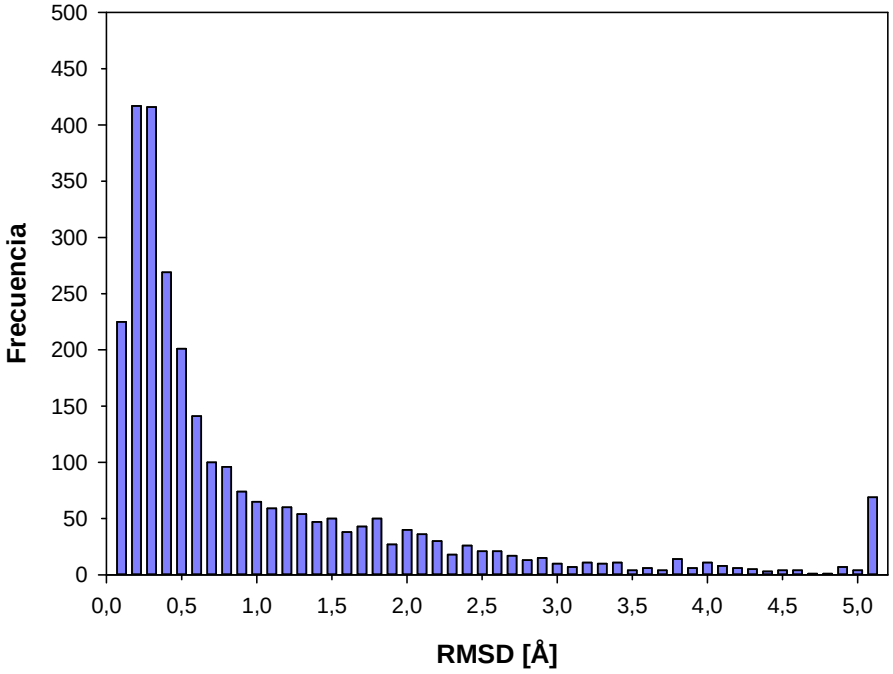



**Figura 10. Comparación del valor promedio de RMSD de diversas poblaciones de dominios de la PCDB. En gris claro se grafica el promedio de RMSD entre las estructuras de todos los dominios de la base de datos. En negro, el promedio de RMSD entre las estructuras de dominios que presentan instancias de cristalización con condiciones diversas. En color gris oscuro, el promedio de RMSD entre las estructuras de dominios que presentan instancias de cristalización con condiciones idénticas.**

Las dos poblaciones de dominios (homogéneos y heterogéneos en condiciones de cristalización), como se aprecia en la Figura 10, difieren en la extensión de su diversidad conformacional. Sus distribuciones fueron sometidas al test estadístico Mann–Whitney, y la discrepancia entre la distribución de las dos poblaciones es estadísticamente significativa, presentando un valor de Z-score de 21,8. En la Figura 11.a se detalla la distribución de los valores de RMSD para aquellos dominios con estructuras provenientes de condiciones homogéneas, y en la Figura 11.b se exhibe la distribución de valores de RMSD para aquellos dominios con estructuras cristalizadas en instancias diversas bajo condiciones diferentes. Se nota una tendencia hacia valores de RMSD más grandes por parte de la población de dominios con estructuras cristalizadas bajo diversas condiciones.

Es interesante resaltar que en la distribución de la Figura 11.a los dominios cristalizados en condiciones homogéneas presentan casos con RMSD altos, indicando la presencia de diversidad conformacional (aproximadamente 200 casos con RMSD mayor a 5). Como se mencionó anteriormente es muy probable que existan parámetros adicionales (por ejemplo tipo de solvente en la cristalización) causales de la diversidad conformacional no considerados en este trabajo. La inclusión de nuevos parámetros a la información de la PCDB es una de las líneas en las que los autores se encuentran trabajando.

**a. Histograma de valores RMSD en población de dominios con parámetros homogéneos**




**b. Histograma de valores RMSD en población de dominios con parámetros heterogéneos**

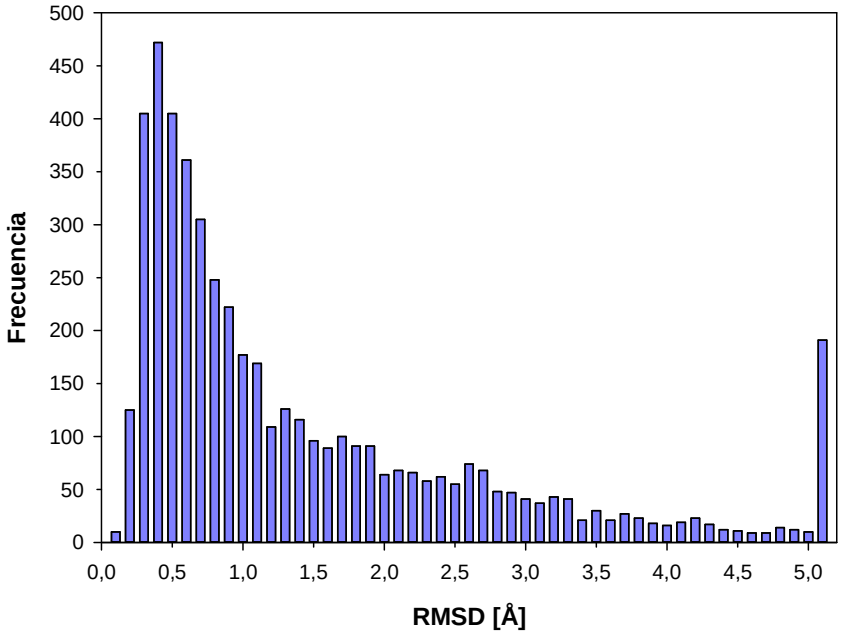



**Figura 11. Se grafica la distribución de los valores de RMSD [Å] entre los pares de estructuras de aquellos dominios cuyas estructuras han sido obtenidas siempre en iguales condiciones (a) o en condiciones diversas (b). Se denota una mayor ocurrencia de valores de RMSD altos en la segunda población.**

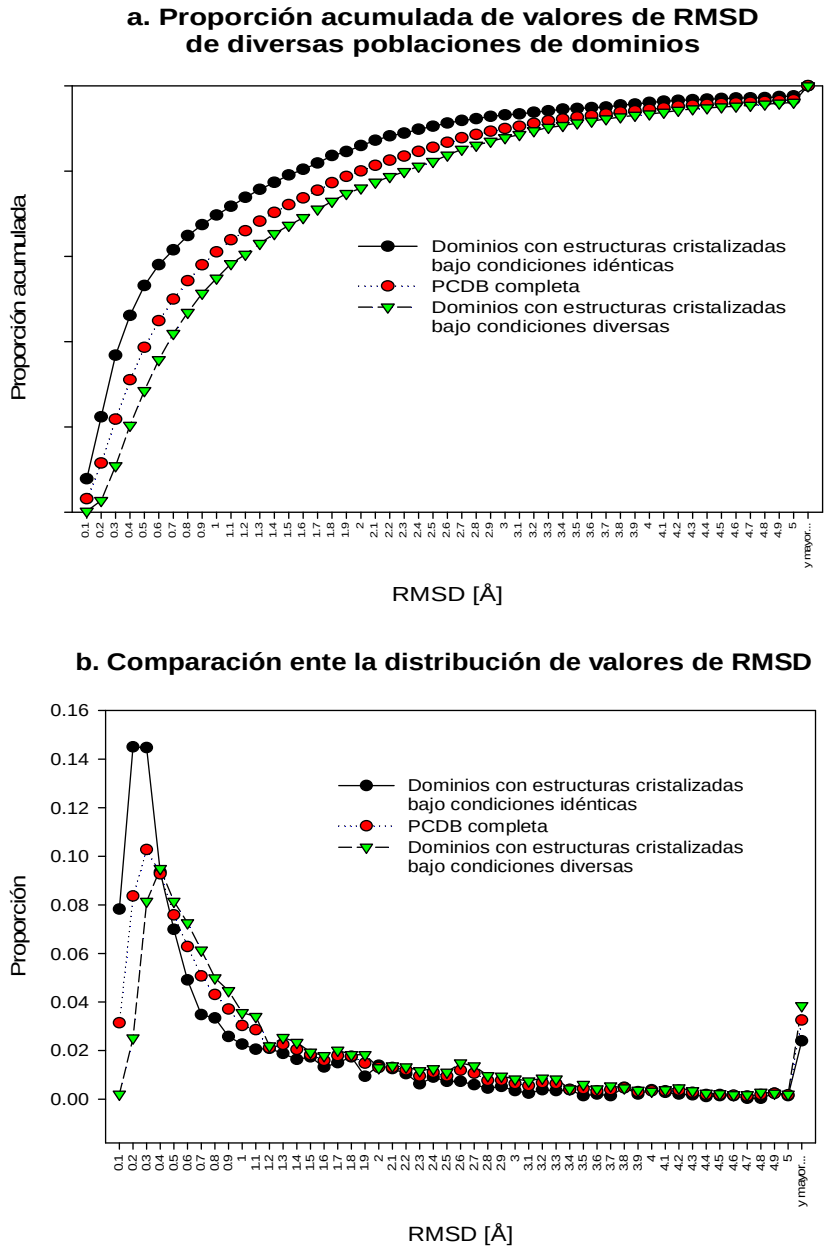



Figura 12. (a) Se grafica el porcentaje acumulado de valores de RMSD totales en función del RMSD de las población total de dominios de la PCDB, de la población con estructuras cristalizadas en iguales condiciones y de la población de dominios con estructuras cristalizadas en condiciones diversas. Se aprecia que la población de dominios con condiciones heterogéneas se condice con valores mayores de RMSD. (b) Comparación entre la distribución de valores de RMSD entre las tres poblaciones de dominios (total, homogénea y heterogénea). Se distingue la tendencia de la población de dominios homogénea hacia valores menores de RMSD.

En la Figura 12.a se comparan las proporciones acumuladas de la población de dominios con estructuras con datos homogéneos para los parámetros estudiados, con la población de dominios con estructuras con datos heterogéneos, y con toda la base de

datos PCDB. En la Figura 12.b se detallan las distribuciones de los valores de RMSD presentados por cada una de estas poblaciones. En ambas figuras se aprecia la mayor diversidad conformacional expuesta por la población de dominios con estructuras cristalizadas bajo diversas condiciones en referencia a la población de dominios compuestos por estructuras cristalizadas bajo idénticas condiciones.

### 4.3 Incidencia de diversos parámetros sobre la diversidad conformacional.

Con el fin de tasar la incidencia y el peso relativo de cada uno de los parámetros estudiados en la extensión de la diversidad conformacional proteica, estudiamos las distribuciones de RMSD para cada parámetro en forma individual. Se comparó la diversidad conformacional observada por poblaciones compuestas por dominios cuyas estructuras han sido cristalizadas en condiciones diversas respecto a determinado parámetro.

Las poblaciones comparadas fueron cinco. Aquellos dominios cuyas estructuras fueron cristalizadas bajo iguales condiciones componen la población “Homogénea”. Aquellos dominios cuyas estructuras fueron cristalizadas en condiciones diferentes exclusivamente respecto a las mutaciones, componen la población “Heterogénea respecto a mutaciones”. Análogamente ocurre con las poblaciones Heterogénea respecto a ligandos, Heterogénea respecto a estado oligomérico y Heterogénea respecto a pH. Aquellos dominios en donde se encuentran estructuras cristalizadas en condiciones diversas para todos los parámetros componen la población “Heterogénea respecto a todos”. Los resultados obtenidos a partir de este estudio se exhiben en la Tabla 6.

| Población | RMSD máximo (media) | RMSD promedio (media) |
|---|---|---|
| Heterogénea respecto a todos | 2,5 | 1,2 |
| Homogénea | 0,9 | 0,7 |
| Heterogénea respecto a pH | 1,5 | 1,1 |
| Heterogénea respecto a estado oligomérico | 1,1 | 0,8 |
| Heterogénea respecto a ligandos | 0,9 | 0,7 |
| Heterogénea respecto a mutaciones | 0,9 | 0,7 |

Tabla 6. Valores de RMSD de poblaciones de dominios con estructuras cristalizadas bajo condiciones diversas para distintos parámetros considerados.

La primera conclusión que se extrae de los resultados aquí expuestos, es la marcada diferencia de diversidad conformacional exhibida entre la población Homogénea y la población Heterogénea con respecto a todos. De aquí se desprende una reafirmación del hecho que los parámetros contemplados por la base de datos PCDB efectivamente provocan cambios conformacionales en la estructura proteica. Este dato confirma el enfoque basal del presente trabajo, e insta a ampliar las líneas de investigación en este sentido. Tanto el valor mínimo como el valor promedio del RMSD calculado entre las estructuras de los dominios que conforman estas dos poblaciones presentan valores significativamente alejados: una media de 2,5 Å para el valor máximo de RMSD en la población Heterogénea contra una media de 0,9 Å en la población Homogénea. En cuanto a los valores promedios, la primera población presenta una media de 1,2 Å mientras que la segunda presenta una media de 0,7 Å.

De la comparación entre las poblaciones heterogéneas solo para un parámetro en particular, se observa que aquellos dominios con estructuras cristalizadas en condiciones diversas respecto pH son los que presentan una mayor diversidad conformacional, tanto si consideramos el máximo RMSD como si consideramos el RMSD promedio. Las poblaciones Heterogénea respecto a ligandos y Heterogénea respecto a mutaciones presentan una diversidad conformacional similar.

La separación entre los valores correspondientes a la población con estructuras cristalizadas en condiciones homogéneas, la población con heterogeneidad en ligandos

y la población con heterogeneidad en mutaciones es muy reducida. La presencia de mutaciones es sabido que puede o no inducir un cambio conformacional, en dependencia de la ubicación y la sustitución en cuestión. Aún así, consideramos que existen numerosos factores no tenidos en cuenta en el presente trabajo que pueden ejercer una modulación sobre la diversidad conformacional..

## Capítulo 5

## Efecto de la diversidad conformacional en la divergencia secuencial

### 5.1 Resumen

Está bien establecido que la conservación de la estructura a través de la evolución se manifiesta en la divergencia secuencial observada en una familia de homólogos. La conservación de determinados entornos fisicoquímicos en pos de mantener el plegamiento proteico, y por ende la función biológica, origina un patrón de sustitución sitio-específico distinguible en alineamientos de secuencias homólogas. Adicionalmente, y como ya hemos mencionado, es importante destacar el hecho que el estado nativo de una proteína no se representa de forma adecuada mediante una única estructura, sino que es mejor descripto por un conjunto de confórmeros asociados a un equilibrio dinámico entre sí. A continuación se presenta nuestro estudio acerca de la influencia de la diversidad conformacional en la divergencia secuencial y en la evolución proteica.

Como punto de partida, derivamos de la base de datos PCDB una serie de 900 proteínas con diferentes grados de diversidad conformacional. Con el propósito de desarrollar un modelo de evolución molecular estructuralmente condicionado y que contemple la diversidad conformacional, exploramos la influencia de diversas conformaciones dentro de la diversidad secuencial.

Hemos encontrado que la presencia de diversidad conformacional ejerce una fuerte modulación dentro del patrón de sustitución secuencial. A pesar de que los confórmeros comparten muchos de los sitios condicionados estructuralmente, en promedio 30% de estos sitios son específicos de un determinado confórmero. Hemos encontrado que en el 76% de las proteínas estudiadas, un confórmero es significativamente mejor que los restantes en la predicción de la diversidad secuencial. Es particularmente interesante notar que en la mayoría de los casos este confórmero es el asociado al ligando, participando en la función biológica de la proteína.

Adicionalmente, mostraremos que la mejor descripción de la diversidad secuencial observada en proteínas homólogas se obtiene utilizando la totalidad del conjunto de confórmeros disponibles en nuestra base de datos.

La existencia de un patrón de sustitución sitio-específico diferencial para cada confórmero indica que la diversidad conformacional puede jugar un papel central en una mejor comprensión de la evolución proteica.

## 5.2 Introducción

La secuencia proteica es una fuente de información utilizada en una gran diversidad de campos, tales como evolución molecular, bioinformática estructural y proteómica entre otros. Esta información puede ser derivada de patrones de sustitución sitio-específicos codificados en alineamientos, dada por la probabilidad diferencial de encontrar la ocurrencia de un aminoácido en diferentes posiciones de la cadena polipeptídica. Un mayor entendimiento de los orígenes de los patrones de sustitución puede representar un avance en el estudio de los mecanismos subyacentes a los procesos evolutivos. Este es uno de los principales ejes para desarrollar y perfeccionar herramientas bioinformáticas que se basan en información secuencial. El descubrimiento de que la conservación de la estructura proteica está íntimamente relacionada con la diversidad secuencial es la base fundamental de este campo de investigación (Lesk & Chothia, 1980; Chothia & Lesk, 1986). A medida que las proteínas evolucionan, son blanco de presiones selectivas para conservar sus estructuras condicionando su divergencia secuencial. Esto indica la presencia de información codificada en sus patrones de sustitución secuenciales observados en la naturaleza. Esta información es usada para búsquedas de relaciones evolutivas usando exclusivamente información secuencial.

Una de las primeras herramientas que implícitamente captura la importancia de esta información fue el uso de profiles para búsquedas de proteínas con homología distante (Gribskov et al, 1987), que inmediatamente fue seguida por el uso de Modelos Ocultos de Markov (HMM, de sus siglas en inglés Hidden Markov Models) (R Hughey & Krogh, 1996; K. Karplus et al., 1997). Los profiles provienen de alineamientos múltiples secuenciales, y son matrices de dimensiones (longitud de secuencia) x (20) que representan la frecuencia relativa de cada uno de los 20 aminoácidos por cada una de las posiciones de la cadena polipeptídica. Cuando las firmas estructurales fueron consideradas explícitamente, ambas metodologías aquí mencionadas aumentaron su desempeño en búsquedas de homólogos remotos, reconocimiento de plegado, y asignación de modelos estructurales (Gribskov et al. 1988; Luthy, McLachlan, and Eisenberg 1991; Eisenberg et al. 1992; Luthy, Bowie, and Eisenberg 1992). Esta visión apoya la idea que diferentes ambientes proteicos modulan diferencialmente el patrón de sustitución aminoacídico (Overington et al. 1990; Overington 1992), y que la inclusión de la información estructural en pos de mejorar la descripción de los patrones de sustitución se condice con el conocido mecanismo que aumenta la ocurrencia de ciertos aminoácidos en determinados elementos de estructura secundaria (Guzzo 1965; Levitt

1978). Como en el caso de los profiles y los HMMs, los modelos evolutivos incrementaron su capacidad al considerar explícitamente en el modelado de los patrones de sustitución secuencial información derivada de la estructura proteica. Muchos de estos modelos han sido desarrollados considerando una variedad de propiedades estructurales referidas al plegado proteico y su estabilidad termodinámica, a potenciales derivados de información evolutiva, propiedades de ambientes fisicoquímicos, estructura cuaternaria y perturbaciones originadas en mutaciones no sinónimas (Koshi and Goldstein 1995; Bastolla, Roman, and Vendruscolo 1999; Dokholyan and Shakhnovich 2001; Parisi and Echave 2001; Fornasari, Parisi, and Echave 2007; Kleinman et al. 2010).

A pesar de que las relaciones entre los patrones de sustituciones y la estructura proteica están bien establecidas, y han sido ampliamente utilizadas en el campo de la bioinformática, la mayor parte de las conclusiones fueron obtenidas describiendo el estado nativo proteico aceptando la visión de una única estructura. Como mencionáramos anteriormente (ver Introducción), la evidencia experimental indica que esta visión debe ser reformulada para introducir la noción de un grupo de confórmeros en equilibrio dentro del concepto del estado nativo (Volkman et al. 2001; James and Tawfik 2003; Henzler-Wildman et al. 2007; Lange et al. 2008). Esta noción ha sido previamente usada para explicar la heterogeneidad en las propiedades de unión de la seroalbúmina bovina (Karush, 1950). Una descripción más formal fue incluida en el modelo Monod-Wyman-Changeux de regulación alostérica propuesto en 1965 (Monod et al., 1965). La idea fue extendida posteriormente en el concepto de embudos de plegado (folding funnels) (J. D. Bryngelson & Wolynes, 1989; Joseph D Bryngelson et al., 1995), explicando las vías de plegado proteico, con un fondo rugoso representando una colección de isómeros conformacionales. De acuerdo a esta visión, las poblaciones de estos confórmeros siguen distribuciones estadísticas y las barreras energéticas que los separan definen el equilibrio conformacional (B. Ma et al., 1999; Nienhaus et al., 1997). La extensión de la diversidad conformacional está entonces relacionada con el grado de rugosidad en la base del embudo, incluyendo la distribución y las alturas de las barreras entre los confórmeros. Las proteínas que exhiben una mayor rigidez en su estructura (como lo pueden ser proteínas con relativamente baja diferencias entre sus confórmeros), estarán asociadas a paisajes más suaves e uniformes, mientras que aquellas proteínas más flexibles, presentarán paisajes rugosos y más complejos. Adicionalmente, se ha demostrado que el equilibrio relativo entre la población de confórmeros está relacionado con el tipo de plegamiento proteico (O Keskin, Jernigan, & Bahar, 2000), con la presencia de ciertas mutaciones (Sinha & Nussinov, 2001), como

así también con la historia evolutiva de las proteínas (Maguid, Fernández-Alberti, Parisi, & Echave, 2006). La idea de embudos de plegado “estáticos” fue extendida posteriormente con la noción de embudos dinámicos al incluir el efecto del ambiente en su descripción (S Kumar, Ma, Tsai, Sinha, & Nussinov, 2000). Los paisajes dinámicos apoyan las hipótesis de selección conformacional de unión, según la cual el ligando escoge la conformación con mayor afinidad, así como el caso de la selección entre anticuerpos y antígenos (Footet & Milstein, 1994). Más aún, a pesar de que el confórmero de mayor afinidad con el ligando puede corresponderse con una conformación con una energía relativamente alta, los confórmeros que pertenecen a una población escasamente poblada en el estado nativo pueden también unir al ligando y desplazar el equilibrio hacia confórmeros con mayor afinidad a éste (S Kumar et al., 2000). La descripción de los paisajes dinámicos ofrece una visión clave para comprender las relaciones entre estructura, función y dinámica (James & Tawfik, 2003; Nobuhiko Tokuriki & Tawfik, 2009), y se condice con los últimos datos experimentales que describen el comportamiento proteico (Boehr, McElheny, Dyson, & Wright, 2006; Hilser, 2010). Estos conceptos nos proporcionan claras elucidaciones de datos experimentales y han reemplazado el bien conocido y casi establecido modelo de llave-cerradura (Fischer, 1894) así también como el modelo inducido para describir las interacciones entre las proteínas y sus ligandos (Koshland, Ray, & Erwin, 1958).

Considerando el estado nativo proteico como un conjunto de confórmeros, así como también los condicionamientos que la conservación de la estructura impone a la diversidad secuencial, nos conlleva a la subsecuente existencia de un patrón de sustitución característico para cada confórmero. Siguiendo con esta idea, la información secuencial contenida en un alineamiento de proteínas homólogas puede representar una combinación compleja de diversos condicionamientos estructurales impuestos por las conformaciones que pueblan el estado nativo de las proteínas.

Por todo lo expuesto anteriormente, en esta parte del presente trabajo nos propusimos estudiar el efecto que puede tener la diversidad conformacional en el patrón de sustitución secuencial originado a partir de los condicionamientos estructurales. Para tal fin, estimamos la diversidad conformacional de un grupo de proteínas extraídas de la base de datos PCDB desarrollada por el autor (Juritz et al., 2011). Como mencionáramos en el capítulo 2, el procedimiento empleado está validado por trabajos previos que han demostrado la correspondencia entre deformaciones estructurales detectadas bajo condiciones de cristalización diferentes y cambios conformacionales relacionados a la flexibilidad propia del esqueleto proteico, expresado por las diferencias

estructurales entre cristalizaciones bajo las mismas condiciones (Best et al., 2006; Zoete et al., 2002).

Para cada proteína con más de dos estructuras, estimamos el patrón de sustitución por medio del SCPE (Structurally Constrained Protein Evolution), un modelo de evolución molecular desarrollado por nuestro grupo de trabajo (Gustavo Parisi & Echave, 2001). El modelo calcula una matriz de sustitución sitio-específica para cada una de las posiciones de la cadena polipeptídica blanco (M. S. Fornasari et al., 2002). Mediante técnicas de máxima verosimilitud estadística, se estimó la influencia de cada confórmero en el patrón de sustitución encontrado en alineamientos de proteínas homólogas.

Nuestros resultados indican que la diversidad conformacional modula fuertemente el patrón de sustitución derivado de los condicionamientos estructurales. La mejor descripción del patrón de sustitución observado, por otro lado, se obtiene mediante el uso de todos los confórmeros para la proteína.

A pesar de que cada confórmero tiene sus propios condicionamientos estructurales, encontramos que en el 76% de las proteínas estudiadas, un único confórmero se encuentra asociado al mejor valor de máxima verosimilitud obtenido, y supera significativamente al resto de los confórmeros. Interesantemente, en el 62% de los casos, este confórmero es el involucrado en la unión del ligando, y sólo en un 25% se corresponde con el confórmero de menor energía del conjunto.

### 5.3 Set de datos utilizado

Dado el alto costo computacional asociado a los cálculos de máxima verosimilitud utilizados en esta parte del trabajo, se derivaron de la PCDB 900 proteínas monodominio elegidas al azar, con diferentes grados de diversidad conformacional. El promedio de confórmeros disponibles para cada una de estas proteínas es de 4,3.

Se tomó el RMSD máximo (RMSDmax) entre los Cα de diferentes confórmeros de la misma proteínas como medida de la diversidad conformacional. Los valores de RMSDmax registrados en las proteínas del set de datos utilizado se encuentran entre un mínimo de 0 Å y un máximo de 22,65 Å. El valor promedio de RMSDmax entre pares de confórmeros para cada proteína es de 7,5 Å. Más del 40% de las proteínas en el conjunto de datos muestran un valor de RMSDmax por encima de 0,4 Å, que es el valor comúnmente observado entre diferentes cristales de una proteína obtenida bajo las mismas condiciones de cristalización.

El ASA (área accesible al solvente) para cada residuo en cada estructura se obtuvo usando el programa NACCESS26, desarrollado por el Departamento de Bioquímica y Biología Molecular de la Universidad College de Londres.

### 5.4 Simulaciones SCPE

Las matrices de sustitución sitio-específicas se obtuvieron utilizando el SCPE (Structurally Constrained Protein Evolution Model) (Gustavo Parisi & Echave, 2001), un modelo de evolución molecular que toma en cuenta explícitamente los condicionamientos estructurales de la estructura proteica. El SCPE simula la evolución de proteínas mediante la introducción de mutaciones al azar en una secuencia proteica de estructura conocida. Las mutaciones son aceptadas o no, de acuerdo a una evaluación del nivel de perturbación estructural generado, utilizando para tal fin un puntaje basado en la diferencia energética introducida por la mutación. La puntuación de cada mutación se compara con un parámetro ($\lambda$) que es una medida representativa de la presión selectiva para la aceptación de las sustituciones no sinónimas. El principal resultado de una simulación SCPE es un conjunto de matrices de sustitución aminoacídicas de carácter sitio-específico, obtenidas por conteo de las mutaciones aceptadas (M. S. Fornasari et al., 2002). Para el presente trabajo se ejecutaron 8000 simulaciones SCPE mediante cálculos independientes para cada estructura, con un valor predeterminado de $\lambda$ de 0,25 y un tiempo de divergencia media de 20 sustituciones no sinónimas por sitio. Estos valores paramétricos fueron evaluados previamente como el mejor conjunto de valores para diferentes proteínas no relacionadas.

Para el estudio particular de los sitios mutacionales, fue utilizado un valor de $\lambda$ de 0,0001; que implica una completa relajación de la presión selectiva respecto a condicionamientos estructurales. De esta manera, la sustitución de las matrices obtenidas está dominada por el proceso que subyace a las mutaciones, definido en el SCPE mediante un modelo empírico de sustitución de codones (Schneider, Cannarozzi, & Gonnet, 2005).

[26] http://www.bioinf.manchester.ac.uk/naccess/

### 5.5 Cómputos de máxima verosimilitud

Los cálculos del examen estadístico de máxima verosimilitud (ML, de sus siglas en inglés Maximum Likelihood) requieren como datos de entrada un modelo de evolución, un alineamiento múltiple y el árbol filogenético correspondiente a las secuencias allí representadas. El programa HYPHY27 (Pond, Frost, & Muse, 2005) fue utilizado con secuencias de comandos personalizadas para permitir la inclusión de las matrices de sustitución sitio específicas generadas a partir del modelo SCPE. Un alineamiento de secuencias homólogas para cada proteína en el conjunto de datos se obtuvo de la base de datos HSSP28 (Sander & Schneider, 1993). Cada alineamiento contiene como mínimo 20 secuencias proteicas con un porcentaje de identidad secuencial mayor al 35%, y es referido a la secuencia con estructura cristalográfica conocida.

Una inferencia filogenética se realizó para cada alineación mediante el enfoque de máxima parsimonia según PROTPARS29 (J Felsenstein, 1989). El modelo JTT (D.T. Jones, Taylor, & Thornton, 1992) fue utilizado como un modelo de referencia sin restricciones estructurales por ser un modelo de evolución molecular ampliamente utilizado y aceptado. Para evitar conflictos por residuos faltantes en estructuras cristalográficas, los cálculos de máxima verosimilitud fueron computados únicamente sobre aquellos residuos compartidos por todos los confórmeros de cada proteína en el conjunto de datos. Se considera que un modelo es superior al otro cuando su valor de máxima verosimilitud es mayor, y esta diferencia es estadísticamente significativa. Las comparaciones entre los modelos se analizaron mediante el coeficiente AIC (de sus siglas en inglés Akaike information criterion) (Akaike, 1974) y una clasificación de los modelos estimados se calcularon utilizando ΔAIC (Burnham & Anderson, 2002). Un valor de ΔAIC <2 se tomó como medida superior del umbral de tolerancia del modelo.

Para cada proteína estudiada, se identificaron los sitios condicionados estructuralmente (SCS, del inglés Structurally Constrained Sites), comparando los resultados de los modelos SCPE y JTT para cada sitio individual. Si el modelo SCPE en un sitio representa mejor el patrón de sustitución que JTT, el sitio es considerado SCS. En caso contrario, se define como un sitio sin restricciones estructurales (UCS, de las siglas en inglés Unconstrained Sites). Los sitios en donde la performance de los dos modelos es indistinguible se denominan sitios mutacionales (MS, de sus siglas en inglés Mutational sites). Los sitios SCS se estimaron para cada proteína con una sola

[27] http://www.hyphy.org
[28] http://swift.cmbi.kun.nl/swift/hssp/
[29] http://cmgm.stanford.edu/phylip/protpars.html;

estructura o estructuras múltiples. En este último caso, se registraron aquellos sitios SCS comunes a todas las estructuras, así también como aquellos sitios SCS específicos de determinada conformación.

### 5.6 Confórmeros que unen ligando y confórmeros de mínima energía

Para estudiar las relaciones entre la presencia de ligando, la diferencia de energía relativa entre confórmeros, el número de contactos interatómicos y los valores de máxima verosimilitud obtenidos para las diferentes estructuras, se ha utilizado un conjunto de 55 proteínas (y 320 estructuras correspondientes), extraídos del set de datos original. Estas proteínas cumplen con las condiciones necesarias para su análisis. Por ejemplo, la energía relativa entre los confórmeros para cada proteína se calcula utilizando el potencial obtenido por Ferrada y Melo (Ferrada & Melo, 2009), que requiere longitudes superiores a 90 residuos. También se ha asignado la ocurrencia de ligandos afines al conjunto resultante, utilizando para tal fin la base de datos Procognate (Bashton et al., 2008). Los ligandos afines depositados en esta base de datos son los que participan en la función biológica de las proteínas. Cada proteína presente en el conjunto de datos contiene al menos un confórmero con y un confórmero sin ligando (datos extraídos de la base de datos PCDB). El número de contactos en cada estructura se determinó utilizando la definición de contacto desarrollado por el grupo de trabajo de Berrera (Berrera, Molinari, & Fogolari, 2003), que también es utilizado por el programa SCPE (Gustavo Parisi & Echave, 2001).

### 5.7 Resultados

#### 5.7.1 Evolución bajo condicionamientos estructurales

Nuestro interés radica en el análisis del impacto de la diversidad conformacional de las proteínas en el patrón de sustitución aminoacídico observado en la naturaleza. Para dilucidar esta relación, es importante saber en qué medida la información estructural opera en el proceso evolutivo de las proteínas. Para tal fin se computaron cálculos de máxima verosimilitud (Joseph Felsenstein, 1981) entre los alineamientos derivados de la base de datos HSSP y sus estimaciones filogenéticas correspondientes. En primer lugar, se efectuaron comparaciones entre las estimaciones de máxima verosimilitud obtenidos utilizando dos modelos de evolución de proteínas, el modelo SCPE (Gustavo Parisi & Echave, 2001) y el modelo JTT (D.T. Jones et al., 1992). El SCPE es un modelo de evolución que considera explícitamente la conservación de la estructura de la

proteína para simular la divergencia secuencial. Por esta razón, el SCPE es clasificado dentro de los modelos “constrained” o condicionados por la estructura. El principal resultado de una simulación SCPE es un conjunto de matrices de sustitución sitio-específicas. El modelo demostró ser adecuado para la detección de firmas derivadas de las limitaciones estructurales de las familias de proteínas (Gustavo Parisi & Echave, 2004). Por otro lado, JTT es un modelo bien establecido y utilizado ampliamente en el campo de la evolución molecular, que utiliza una matriz de sustitución única para todos los sitios estudiados. Esta matriz se deriva de la acumulación de sustituciones de aminoácidos a partir de un gran número de proteínas. En consecuencia, JTT se considera como un modelo “unconstrained” o sin restricciones estructurales, en el sentido que no toma en cuenta de forma explícita las limitaciones estructurales específicas. Los cálculos de máxima verosimilitud nos permiten comparar la capacidad de ambos modelos para describir el patrón de sustitución real observado en el alineamiento. Como se mencionó anteriormente, la comparación de modelos se realizó utilizando los criterios de Akaike de Información (AIC, (Akaike, 1974)) y su significación estadística se evaluó posteriormente con el método de Burnham y Anderson (Burnham & Anderson, 2002). En los casos en donde el SCPE supera al modelo JTT, se espera que las restricciones estructurales sean un componente importante en el alineamiento de secuencias. Al contrario, cuando el modelo JTT supera al SCPE, puede ocurrir debido a dos situaciones diferentes. Por un lado, es posible que el alineamiento no logre comprender la suficiente información estructural. Por otro lado, una tendencia fisicoquímica u otro determinante del modelo de sustitución podría ser mejor descrito por el modelo JTT. Adicionalmente, es posible también que ambos modelos expliquen igualmente bien el patrón de sustitución encontrado.

El set de datos aquí utilizado está compuesto por un conjunto redundante de 3.896 dominios de la base de datos CATH (Greene et al., 2007), que pertenecen a 900 proteínas de diferentes familias estructurales extraídos de la base de datos PCDB (Juritz et al., 2011). Mediante cálculos de máxima verosimilitud para cada dominio, se encontró que el modelo SCPE supera al modelo JTT en el 89% de los casos. Este resultado sugiere que las proteínas presentan posiciones condicionadas por parámetros estructurales que ejercen un efecto modulador sobre la diversidad secuencial. Esta hipótesis fue explorada con más detalle, realizando estimaciones de máxima verosimilitud y evaluaciones estadísticas por posición para calcular el porcentaje de los sitios donde el modelo SCPE supera a JTT. Se encontró un promedio de 37% de estos sitios por proteína estudiada, identificados como sitios SCS (Figura 13 y Figura 14). Por otro lado, el modelo JTT supera al SCPE en promedio en sólo el 12% de los sitios,

llamados UCS. Los sitios en donde los dos modelos de evolución explican igual de bien la divergencia secuencial, son llamados sitios mutacionales (MS).

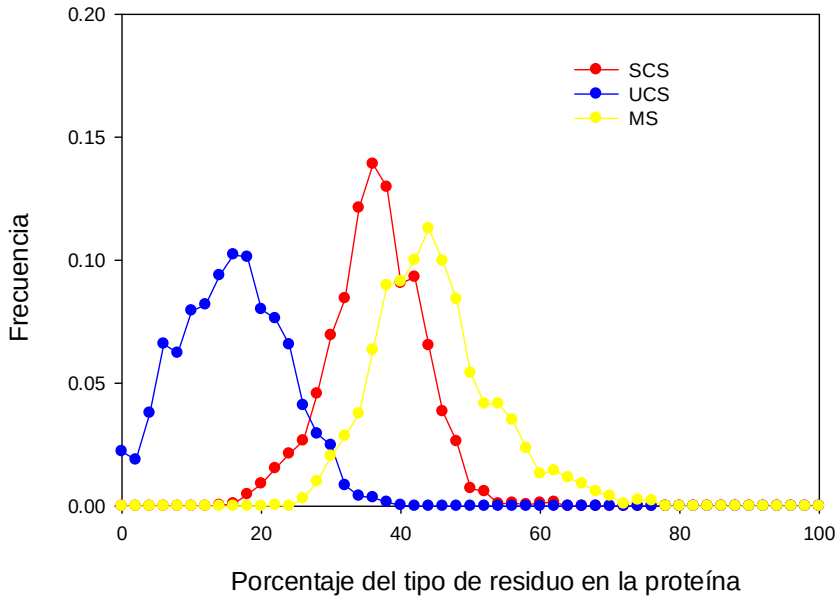



**Figura 13. Comparación entre la distribución de ocurrencia de diferentes clases de sitio en las estructuras estudiadas en nuestra base de datos. La clasificación de los sitios se deriva de la comparación entre el valor de máxima verosimilitud obtenido mediante un modelo evolutivo que contempla información estructural (SCPE) y otro modelo que no considera esta información (JTT). Los sitios donde el modelo SCPE supera a JTT se denominan sitios estructuralmente condicionados (SCS, de sus siglas en inglés Structurally Constrained Sites) y aquellos en los que el modelo JTT supera al SCPE, son sitios no restringidos (UCS, de sus siglas en inglés Unconstrained Sites). Los sitios en donde los resultados de ambos modelos son indistinguibles estadísticamente se denominan sitios mutacionales (MS, de sus siglas en inglés Mutational Sites), ya que no están sometidos a una presión selectiva de carácter estructural o fisicoquímica.**

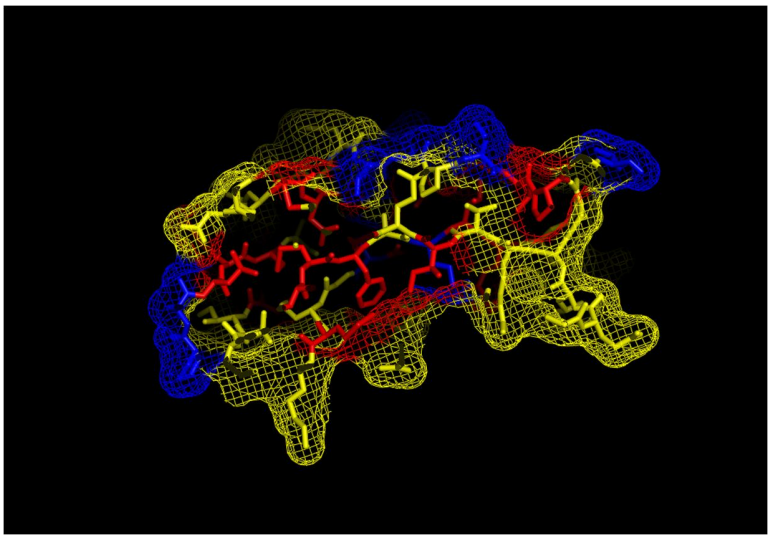


Figura 14. Ejemplo, usando la representación de mallas y varas, de la distribución de diferentes clases de sitio en una proteína que denota una evolución bajo limitaciones estructurales (proteína receptora morfogenética ósea de tipo IA, código PDB 1goo, código CATH 1gooB00). La figura muestra la distribución relativa de los sitios estructuralmente condicionados (SCS; en color rojo), los sitios no restringidos estructuralmente (UCS; en color azul) y los sitios mutacionales (MS; en color amarillo). Las limitaciones estructurales en el proceso de sustitución a lo largo de la evolución explican por qué los sitios tipo SCS suelen corresponderse con los residuos enterrados, mientras que los sitios tipo UCS son, en su mayoría, expuestos a solventes.

Las posiciones clasificadas como SCS son las posiciones proteicas que evolucionan bajo limitaciones estructurales de tipo terciario. Es evidente que limitaciones estructurales adicionales podrían influir en la divergencia de proteínas, como las interacciones proteína-proteína (M. S. Fornasari et al., 2002), plegamiento (Drummond & Wilke, 2008) o la agregación de proteínas (Monsellier & Chiti, 2007). Las limitaciones estructurales en el SCPE surgen en función del número de contactos entre los átomos de los residuos. En la Figura 15 se muestra la distribución de sitios SCS y UCS en función del número de contactos. La mayoría de los sitios SCS presentan entre cuatro y seis contactos por residuo, mientras que los sitios UCS poseen una media de uno. Adicionalmente, hemos notado que los sitios SCS suelen corresponder a los residuos más ocultos al solvente, mientras que los sitios UCS son en su mayoría expuestos al solvente Figura 14..

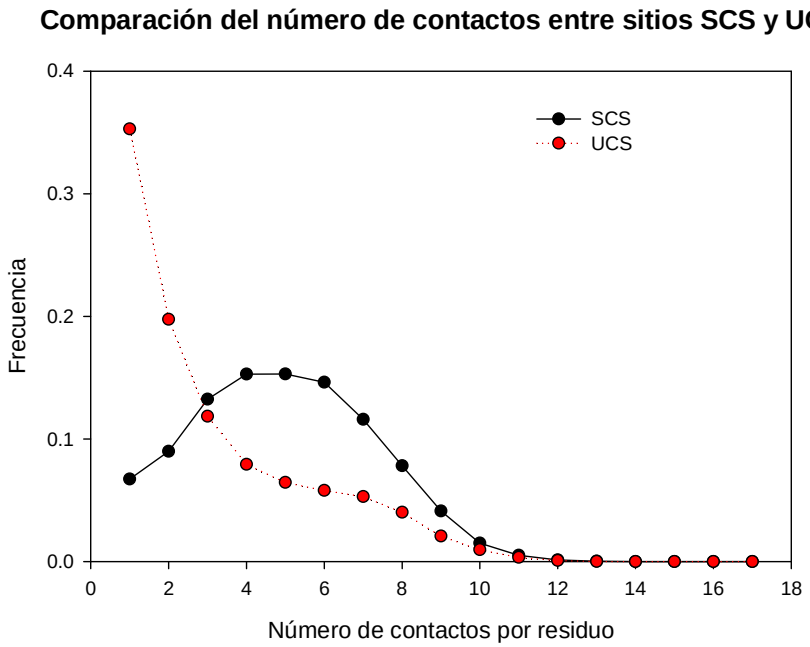



**Figura 15. Distribución de los sitios estructuralmente condicionados (SCS) y los sitios no restringidos (UCS) en función del número de contactos entre átomos de residuos. La mayoría de los sitios de tipo SCS tienen en promedio 5 contactos, mientras que los sitios restringidos presentan sólo un contacto como media.**

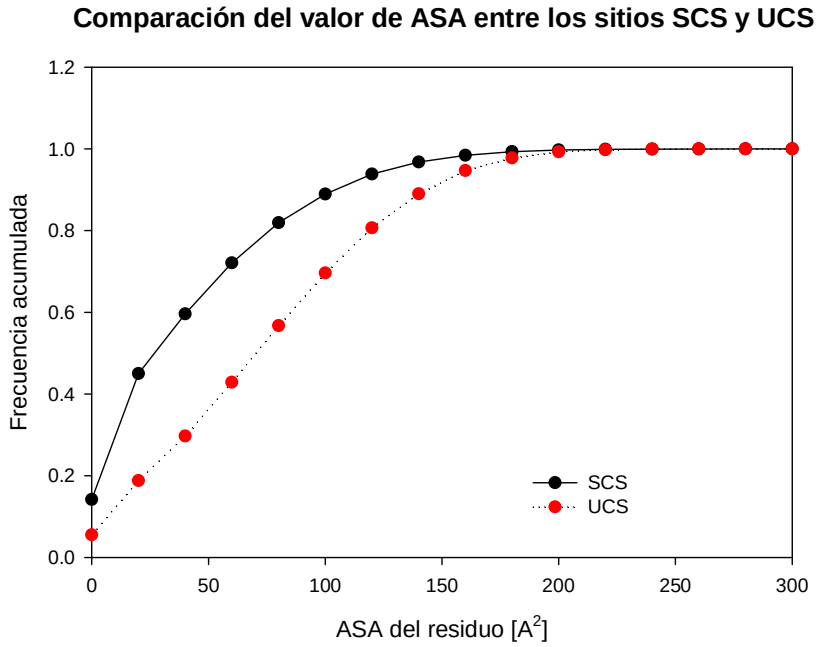



**Figura 16. Comparación entre la distribución de la exposición al solvente de los sitios estructuralmente condicionados (SCS) y los sitios no restringidos (UCS). La mayoría de los sitios tipo SCS están más enterrados en el núcleo proteico, mientras que UCS son en su mayoría solvente expuestos, encontrándose en el contorno exterior proteico.**

Es interesante notar que la mayoría de los sitios (45% en promedio) no puede ser clasificado como sitios SCS ni como sitio UCS (Figura 14, en color amarillo), pues se encuentran igual de bien descritos por ambos modelos. Para el estudio de sus propiedades se llevó a cabo un conjunto independiente de simulaciones SCPE, pero sin imponer restricciones estructurales. En estas condiciones, los patrones de sustitución resultantes corresponden al modelo de sustitución de codones empírico utilizado por SCPE en el proceso de mutación (A. Schneider et al., 2005). Esto significa que las matrices de sustitución obtenidas por estas simulaciones no contienen información acerca de las limitaciones estructurales o fisicoquímicas. Se encontró de esta manera que el 89% de los sitios que son explicados igualmente bien por los dos modelos (SCPE y JTT), también se explican igualmente bien por estas matrices de sustitución mutacionales. En consecuencia, estos sitios se denominan sitios mutacionales (MS), y de acuerdo a nuestros resultados, evolucionan bajo ninguna otra restricción más allá del proceso netamente mutacional.

En este punto, es interesante analizar cómo los sitios de SCS, siendo en promedio sólo el 37% del total, pueden lograr una interpretación más exacta que el modelo JTT en el 89% de las 3.896 estructuras estudiadas. La razón de dicha diferencia se puede encontrar en el cálculo de la diferencia del valor de máxima verosimilitud promedio por sitio ($ML_{SCPE} - ML_{JTT}$). Hemos encontrado un valor de 2,47 para los sitios SCS, valor que representa una mejora considerable en la descripción de la evolución de estos sitios de acuerdo con el criterio AIC. Por lo tanto, estas mejoras obtenidas en los SCS son lo suficientemente significativas como para conducir a una mejora generalizada en toda la estructura. Sin embargo, la mayoría de los sitios (45% en promedio) son igualmente bien descriptos por los dos modelos, lo que significa que ni las consideraciones estructurales ni los aspectos fisicoquímicos son necesarios para explicar la evolución observada de estos sitios. Veremos más adelante que la consideración de la diversidad de conformación va a alterar esta opinión.

### 5.7.2 Evolución bajo escasos o nulos condicionamientos estructurales

Como se ha mencionado anteriormente, encontramos que el modelo JTT supera al SCPE en el 11% de los dominios CATH de nuestro set de datos. Para estos casos, las limitaciones estructurales pueden estar ausentes, o bien el SCPE no describe de modo acertado su modelo de sustitución. El análisis individual de estas estructuras ha demostrado dos causas principales que explican el desempeño observado de los modelos. Por un lado, hemos encontrado que en ciertos casos estas estructuras están asociadas con grandes ligandos, alojados en cavidades vastas, enterradas en la proteína. Este tipo de interacción (proteína-ligando) no se considera en el modelo SCPE, que sólo considera interacciones entre aminoácidos. Por lo tanto, la evolución de los residuos dentro de las limitaciones estructurales relacionados con la unión de los ligandos grandes no sería así simulada correctamente en el SCPE. Por otra parte, hemos constatado que la mayor parte de las estructuras que pertenecen a este grupo corresponden a proteínas clasificadas como “estructuras sueltas” o “loosy”, indicando la ausencia de estructura secundaria o de limitaciones estructurales. La Figura 17 muestra ejemplos de proteínas. Algunas de estas proteínas tienen muy pocos elementos estructurales secundarios, grandes bucles, o pliegues con baja densidad entre los residuos o los contactos de largo alcance. En estos casos, la distribución de sitios SCS y UCS se invierte en relación con las proteínas estructuradas, previamente analizadas, mientras que la distribución de los sitios MS sigue siendo esencialmente la misma (Figura 18). Por lo tanto, la mayor parte de los residuos de estas proteínas son sitios no condicionados estructuralmente o sitios mutacionales. El modelo JTT deriva sus probabilidades de sustitución de proteínas diferentes, y subsecuentemente de una diversidad de entornos estructurales y funcionales. Esta información generalizada explica mejor en promedio el modelo de sustitución encontrado en sitios sin limitaciones estructurales (en promedio un 30%), pero aún así muestra un sesgo de sustitución con algún tipo de patrón fisicoquímico. Como los sitios mutacionales son explicados con desempeño comparable por los tres modelos (SCPE, JTT y mutacionales), es la fracción de los sitios UCS la que explica la evolución de proteínas con bajo contenido estructural.

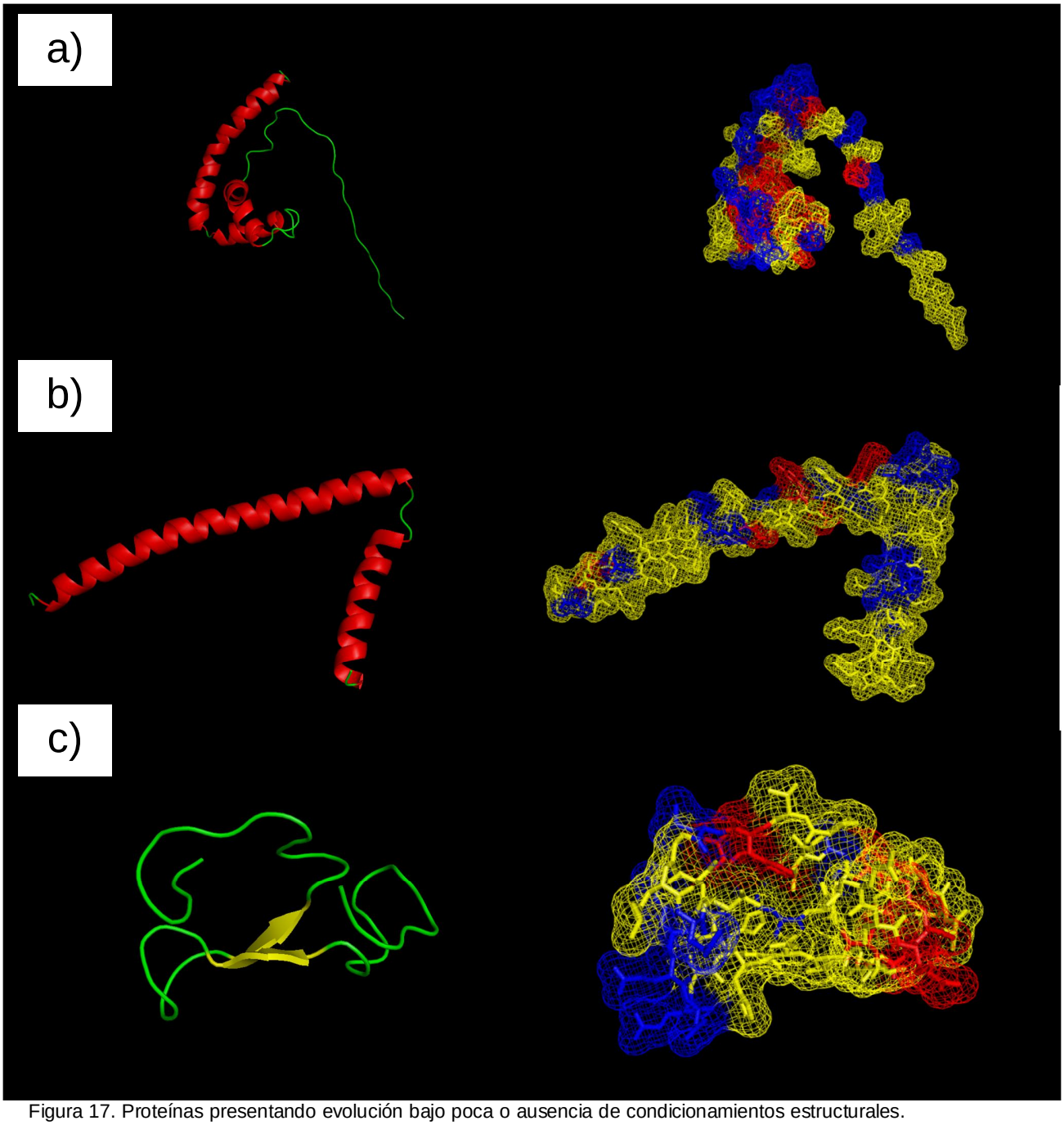



Figura 17. Proteínas presentando evolución bajo poca o ausencia de condicionamientos estructurales. a) proteína NHP6A (PDB 1CG7; CATH 1cg7A00). b) Troponina T (PDB 1j1d; CATH 1j1dB00). c) Dihydroorotato Deshidrogenasa B (subunidad PYRK) (PDB 1ep2; CATH 1ep2B03). En todos los casos la representación esquemática de la izquierda y la representación en mallas y varas de la derecha se obtuvieron mediante PyMOL. En color rojo se representan los sitios tipo SCS, en amarillo los sitios tipo MS amarillo y en azul los sitios tipo UCS.

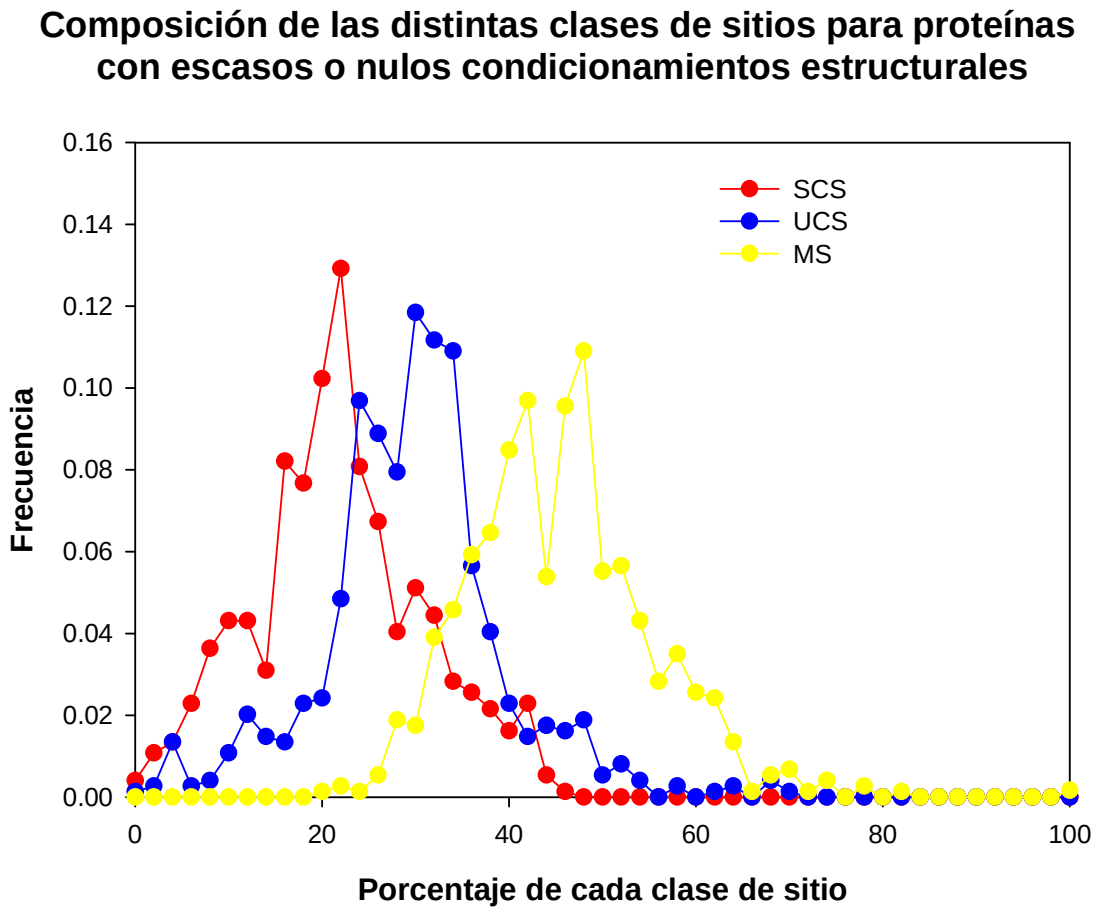



Figura 18. Distribución de la ocurrencia en las distintas clasificaciones de residuos en proteínas que exhiben una evolución bajo escasos o nulos condicionamientos estructurales.

### 5.7.3 Evolución y diversidad conformacional

Con el fin de estudiar el efecto de la diversidad conformacional en la divergencia secuencial de proteínas homólogas, se utilizó la misma colección de estructuras redundantes previamente descripta para describir el patrón de sustitución de 900 proteínas. En este sentido, disponemos en promedio de 4,3 estructuras por proteína para describir la influencia del conjunto de confórmeros en el proceso de sustitución a evaluarse con técnicas de máxima verosimilitud. Como mencionáramos en el capítulo 2, como una medida de la diversidad de conformación se utilizó el máximo valor de RMSD (RMSDmax) derivado de una comparación de todos contra todos entre todas las estructuras recolectadas para cada proteína desde la base de datos PCDB. La

distribución de estos valores se muestra en la Figura 19, y coincide con la distribución de estudios previos en un conjunto de datos más grandes (Burra et al., 2009).

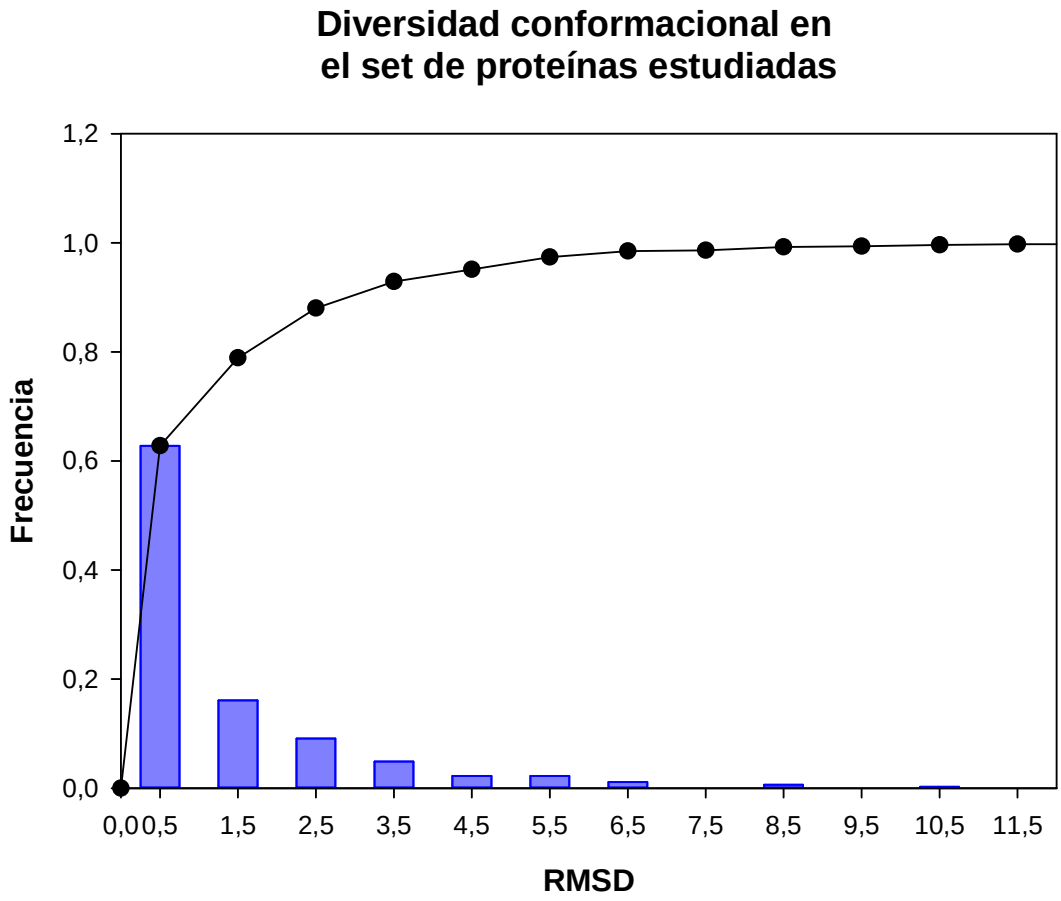



Figura 19. Distribución del RMSD [Å] máximo calculado entre las estructuras de las 900 proteínas estudiadas.

Utilizando el mismo método ya explicado anteriormente para clasificar los sitios, fue calculado el número de sitios SCS, UCS, y MS para cada proteína como la suma de la cantidad de sitios diferentes en todas las diferentes conformaciones. La Figura 21 muestra la distribución obtenida de estas tres clases. Comparando con las distribuciones presentadas en la Figura 13, se aprecia que el promedio del número total de SCS se incrementa del 37% al 48% cuando todos los confórmeros del conjunto son considerados. Al comparar las dos distribuciones está claro que la principal consecuencia de considerar la diversidad conformacional es un incremento en el número de sitios SCS, que representan ahora la mayoría de los sitios de las proteínas. Por otra parte, los sitios SCS pueden ser clasificados como: (a) sitios SCS compartidos

por todas las conformaciones (un 73% del total de sitios SCS), es decir, sitios que pueden ser clasificados como sitios SCS en todas las conformaciones de la proteína, y (b) sitios SCS distintivos de una única conformación (27% del total de sitios SCS), es decir, sitios que pueden ser clasificados como SCS en sólo una de las conformaciones del conjunto. Para estudiar la relación entre el porcentaje de estos sitios y la extensión de la diversidad conformacional, se calculó el coeficiente de correlación de Spearman con una distribución log-log entre RMSDmax y el porcentaje de sitios SCS totales. Un valor de 0,36 se obtuvo utilizando los sitios SCS recogidos en todas las conformaciones. Sin embargo, una mejor correlación (0,48) se obtiene al considera el porcentaje de SCS específicos de las diferentes conformaciones individuales. Este último valor, evidentemente, refleja mejor la relación entre la movilidad del estado nativo proteica y las limitaciones estructurales en la evolución.

Aunque la mayoría de los confórmeros comparten la mayoría de los sitios SCS, nuestros resultados indican que el modelo de sustitución de un conjunto de secuencias homólogas contiene información sobre la dinámica de las proteínas. Por esta razón, la consideración explícita de la diversidad conformacional de proteínas mejora la descripción del modelo de sustitución de aminoácidos observado. De hecho, la mejor descripción del proceso evolutivo, en términos de los cálculos de máxima verosimilitud, puede ser obtenida usando el mejor valor de máxima verosimilitud por posición entre todas las conformaciones de cada proteína. La consideración del mejor valor obtenido por máxima verosimilitud por posición entre todas las conformaciones de cada proteína resulta en un aumento de hasta un 25% por cada sitio (en promedio 1,7% por cada sitio, Figura 20) en comparación con los mejores valores obtenidos mediante un única conformación del conjunto disponible para cada proteína. Adicionalmente, la correlación tipo log-log entre el RMSDmax y la diferencia entre el mejor valor de máxima verosimilitud y el promedio obtenido por todas las conformaciones tiene un coeficiente de correlación de Spearman de 0,498, lo que indica una vez más la relación entre el dinamismo de las proteínas y las limitaciones de la secuencia evolutiva.

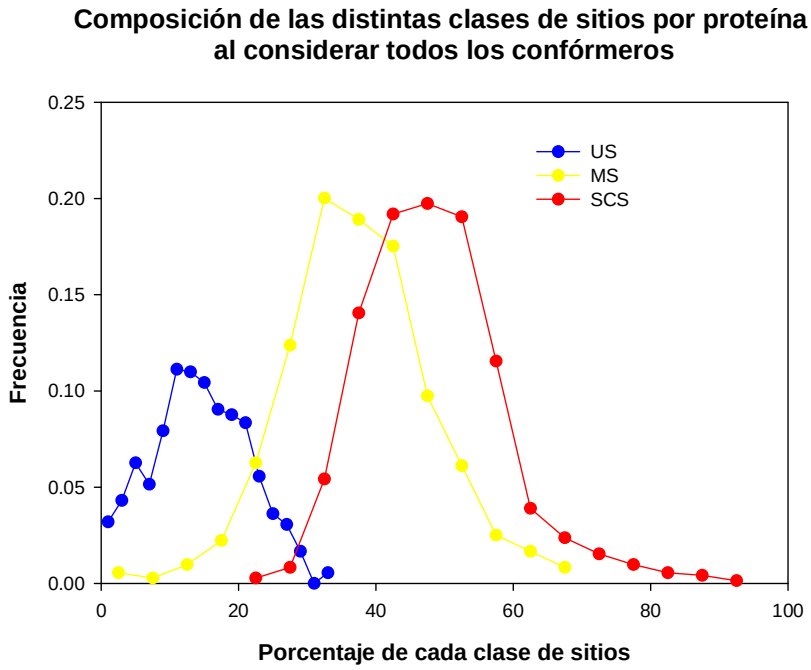



**Figura 21. Distribución de la ocurrencia de las distintas clases de sitios cuando son considerados todos los confórmeros de las proteínas estudiadas.**

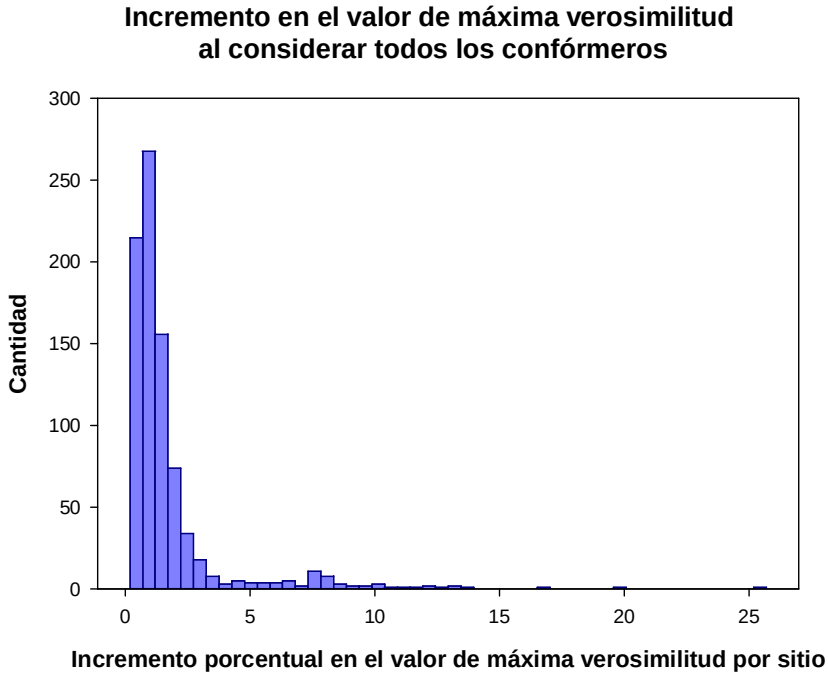



**Figura 20. Incremento registrado en el valor de máxima verosimilitud por sitio. Se grafica como la diferencia porcentual entre el valor calculado considerando una sola conformación, y considerando la mejor conformación en términos de máxima verosimilitud de todas las estructuras consideradas.**

A.

Aunque la mejor descripción del patrón de sustitución se obtiene utilizando todas las conformaciones de cada proteína, es importante señalar que en el 76% de las proteínas se encontró que una sola conformación supera estadísticamente al resto en términos de la prueba de máxima verosimilitud y de AIC. La importancia de la existencia de esta "mejor conformación" tiene que ver con su influencia para explicar la divergencia secuencial. Esta característica podría estar asociada con una mayor presión selectiva en su evolución, por lo tanto, estos confórmeros podría ser importante desde el punto de vista biológico. Con el fin de obtener una comprensión más profunda de este hallazgo, se estudió la relación entre la presencia de una mejor conformación, las energías relativas entre los distintos confórmeros y la presencia de los ligandos. Se utilizó un potencial (Ferrada & Melo, 2009) para calcular las energías relativas entre los confórmeros y la base de datos Procognate (Bashton et al., 2008) para asignar la presencia de los ligandos para cada conformación del conjunto. El potencial estadístico ha sido utilizado con éxito en diferentes contextos como predicción de estructura de proteínas, evaluación de la calidad estructural e identificación de errores, el reconocimiento de plegamientos, y el diseño de proteínas.

Hemos encontrado que la mejor conformación podría corresponder a la conformación con el mínimo de energía en sólo el 28% de los casos; pero, por otro lado, en el 62% se corresponde con la conformación con el ligando unido. Estos resultados podrían indicar que confórmeros con ligandos unidos, a pesar de ser menos representados en el conjunto de confórmeros en equilibrio, son responsables de una fuerte presión selectiva que influye en el patrón de sustitución secuencial durante la evolución.

Los datos aquí presentados sugieren que nuevos modelos de evolución molecular y herramientas y aplicaciones bioinformáticas deberían ser desarrolladas considerando explícitamente estas tendencias

## Conclusiones

En el estudio del comportamiento de los sitios en las distintas conformaciones de una proteína, hemos detectado que ciertos sitios son estructuralmente condicionados en su divergencia secuencial en todas o algunas de las conformaciones (sitios SCS, del inglés Structurally Constrained Sites), mientras que algunos no lo son en ninguna (UCS, del inglés Unconstrained Sites). A pesar de que los confórmeros comparten la mayoría de sus sitios tipo SCS, casi el 30% de ellos son específicos para cada conformación. Esto significa que un alineamiento de secuencias de proteínas homólogas contiene información conformacional acerca de las proteínas, una observación que está de acuerdo con los hallazgos previos que describen que el dinamismo es una característica conservada en la evolución de proteínas (Maguid, Fernandez-alberti, & Echave, 2008; Maguid et al., 2006). Hemos encontrado que la mejor descripción del patrón de sustitución observado en la naturaleza para una proteína se obtiene cuando todas sus conformaciones estructurales son consideradas. Se obtiene una mejora significativa en los valores de máxima verosimilitud por posición cuando se considera el mejor valor por cada sitio entre todos los confórmeros respecto a cuando se considera solo el mejor confórmero. El promedio en el incremento del valor de máxima verosimilitud es del 1,7%, y el aumento alcanza el 25% en algunos casos, testimonio de la mejora en la sustitución de las matrices sitio específicas mediante la incorporación de información adicional proveniente del conjunto de confórmeros. Sin embargo, las diferentes conformaciones del conjunto no tienen el mismo impacto en el patrón de sustitución. Es interesante notar que entre todas las conformaciones, en el 76% de las proteínas una única conformación supera al resto en términos de cálculos de máxima verosimilitud. En un estudio más profundo de este resultado, hemos relacionado la distribución de los confórmeros a dos de los principales factores (energía relativa y la presencia de ligandos) que influyen en el equilibrio entre los confórmeros de acuerdo a la hipótesis de pre-equilibrio (C.-jung Tsai, Kumar, Ma, & Nussinov, 1999). Hemos encontrado que la "mejor conformación" es asociada mayormente con la conformación de unión del ligando más que con la conformación de la energía relativa más baja. En consecuencia, la conformación de la energía más baja, que normalmente se asocia con una baja actividad de unión (Kantrowitz & Lipscomb, 1990; Velyvis, Yang, Schachman, & Kay, 2007), tiene menor influencia en la divergencia secuencial que otros confórmeros, menos poblados en el equilibrio pre-existente, apoyando la función biológica debido a la unión de ligandos afines. Es posible que la mejor conformación sea una conformación compacta, con un mayor nivel de contactos entre los residuos, favoreciendo la aparición

de sitios SCS. En general se establece que la llamada "conformación cerrada" se asocia con la capacidad de unión y la actividad biológica (Gutteridge & Thornton, 2005). Sin embargo, en nuestra base de datos la mejor conformación se asocia con la conformación con el número relativo más alto de contactos solo en el 25% de los casos. De esta forma, el sesgo reportado en los patrones de sustitución, específicos para cada conformación, podría reflejar el resultado de la diferencia de presiones selectivas durante la evolución, pudiendo no presentar ninguna relación con el nivel de compactación de los confórmeros

Nuestros resultados están de acuerdo con las ideas previas que explican las tendencias generales en la evolución de las proteínas y el plegamiento de proteínas, en particular, la teoría de la "frustración mínima" (minimum frustration) (Joseph D Bryngelson et al., 1995). De acuerdo con este punto de vista, las sustituciones aminoacídicas son seleccionadas para la ocurrencia de interacciones de las cadenas laterales, de tal manera que se favorezca el estado plegado de la proteína. Sin embargo, una pequeña fracción de residuos energéticamente "frustrados" pueden permanecer, y han sido asociados con los residuos de unión (Ferreiro, Hegler, Komives, & Wolynes, 2007). De manera similar, y aunque los confórmeros comparten la mayoría de los sitios SCS, hemos encontrado sitios SCS específicos de confórmeros menos poblados (de mayor energía), y, por lo tanto, energéticamente frustrados. Las simulaciones SCPE no incluyen las funciones selectivas para discriminar los residuos funcionales, utilizando este término en la forma clásica para indicar residuos relacionados con la catálisis y la unión. Sin embargo, como la función de la proteína se basa en la diversidad conformacional (M Karplus & Kuriyan, 2005), el modelo de sustitución confórmero específico desarrollado en este trabajo podría indicar la existencia de una clase más amplia de "residuos funcionales", relacionados con el dinamismo necesario para mantener la actividad biológica. Por otra parte, en una determinada familia de proteínas con residuos altamente conservados que participen en la catálisis y/o en la unión de ligandos, se espera cierta diversificación funcional que puede estar relacionada con la presencia de promiscuidad de sustrato, así como con la especificidad múltiple (N Tokuriki & Tawfik, 2009). Esta divergencia funcional puede ser explicada en términos de muestreo conformacional de los residuos catalíticos o de unión debido a rearreglos locales (flexibilidad de los lazos y cadenas laterales) y globales (movimientos de dominio o transiciones de plegamiento) entre el conjunto de confórmeros. Diferentes mutaciones que ocurren durante la evolución podrían afectar a la población relativa de los confórmeros, originando de esta manera la divergencia funcional (Buyong Ma, Shatsky, Wolfson, & Nussinov, 2002). La información asociada a

los sitios SCS confórmero específicos podría ayudar en la estimación del espacio secuencial definido por un conjunto dado de confórmeros, y podría representar una herramienta clave en la caracterización de grupos funcionales o subfamilias (Abhiman & Sonnhammer, 2005).

Otro hallazgo interesante es que los sitios estructuralmente condicionados representan hasta un 37% de los sitios en promedio teniendo en cuenta una única estructura, y un máximo de 48% cuando la diversidad conformacional se tiene en cuenta. Adicionalmente, se observó que la distribución de los sitios estructurales es muy amplia (abarca desde el 20% al 90%). A pesar de que está fuera del alcance del presente trabajo, la cantidad de limitaciones estructurales ha sido recientemente relacionada con la tasa de evolución de las proteínas. Por el momento, una de las correlaciones más fuertes y consistentes entre los datos genómicos y las tasas de cambio evolutivo es el nivel de expresión de los genes. Estimaciones anteriores han establecido que las limitaciones estructurales podrían explicar hasta un 10% de la variación de la tasa de evolución . Sin embargo, estudios más recientes indican que características particulares de la estructura y de la función, así como también la tasa de traducción pueden contribuir comparablemente para explicar las tasas de evolución. La observación de que cada conformación modula diferencialmente el patrón de sustitución, y que la diversidad conformacional implica en promedio el 48% de los residuos de una proteína, es indicio de que las restricciones estructurales son pilares fundamentales en la divergencia de las proteínas, y que desempeña a su vez un papel crucial en la comprensión de las tasas de evolución.

**Capítulo 6**

**Diversidad conformacional y estimación de la estabilidad proteica**

## 6.1 Resumen

La estimación teórica de los efectos de una mutación en la estabilidad de la estructura de proteínas es un tema clave en la biología computacional. En este trabajo se describe cómo tomar en cuenta la información codificada en la diversidad conformacional proteica aumenta la precisión de los cálculos teóricos de ΔΔG. Para tal fin, hemos utilizado un conjunto de datos compuesto por los valores de ΔΔG experimentales calculados para 1.993 mutaciones puntuales de 89 proteínas distintas. Hemos calculado teóricamente el efecto de estas mutaciones en la estabilidad de la estructura de proteínas con las aplicaciones bioinformáticas más difundidas para el cálculo de ΔG. A continuación, se relacionó cada proteína presente en el conjunto de datos con la Base de Datos de Diversidad Conformacional PCDB, con el fin de obtener la compilación redundante de todas las estructuras cristalográficas disponibles para cada una de estas proteínas. Se computaron los mismos cálculos teóricos pero teniendo en cuenta no sólo una estructura, sino todos los confórmeros disponibles derivados de la PCDB.

La comparación de los resultados obtenidos demuestra que considerar la diversidad conformacional del estado nativo incrementa la correlación entre los valores teóricos y experimentales de ΔΔG.

## 6.2 Introducción

Una mutación puntual es un cambio de una base por otra en la secuencia de ADN. Debido a la naturaleza redundante del código de codones, algunas mutaciones a nivel de ADN no se traducirán en un cambio de la secuencia de la proteína (mutación sinónima). Por otro lado, muchas mutaciones en la secuencia de ADN codificante se traducen en un cambio aminoacídico en la secuencia proteica.

La predicción del efecto que una mutación puntual tendrá sobre una proteína es un campo de estudio de gran importancia en la actualidad. Es por esto que muchos esfuerzos se están haciendo en pos de aumentar nuestra capacidad de predecir los efectos de mutaciones puntuales (Capriotti, Fariselli, & Casadio, 2004; Capriotti, Fariselli, Calabrese, & Casadio, 2005; Cheng, Randall, & Baldi, 2006; Frenz, 2005). Han sido desarrollado varios métodos computacionales para estimar los cambios de

estabilidad causados por la sustitución de las cadenas laterales de las proteínas (ΔΔG = ΔGwild type - ΔGmutated). La mayoría de ellos se basan en el análisis de la perturbación energética y/o estructural inducido a la estructura nativa de la proteína por las mutaciones estudiadas. A pesar de ser computacionalmente muy demandantes, los primeros métodos diseñados utilizaban un modelo en donde se incluían todos los átomos de la molécula en la estimación del ΔΔG (Bash, Singh, Langridge, & Kollman, 1987). Luego se desarrollaron potenciales simplificados, acoplados con búsquedas conformacionales limitadas (C. Lee, 1994; C. Lee & Levitt, 1991) y el uso de distintos tipos de potenciales, como aquellos basados en interacciones hidrofóbicas (Koehl & Delarue, 1994), en los tipos de estructura secundaria (Munoz & Serrano, 1994), en los contactos establecidos entre los residuos (Miyazawa & Jernigan, 1994) y aquellos basados en el conocimiento (Sippl, 1995). Estos métodos permiten el estudio del efecto termodinámico de diferentes mutaciones en las proteínas en un tiempo computacional razonable. Recientemente, los enfoques basados en un aprendizaje automático han sido implementados para la predicción de ΔΔG en proteínas. Algunos de estos métodos pueden utilizar como entrada bien la estructura de la proteína o su secuencia (Khan & Vihinen, 2010).

La estabilidad de una proteína se puede describir como su capacidad para perder su estado estructural nativo. La forma tradicional de cuantificar la estabilidad de una proteína es mediante el cálculo de la variación de la Energía libre de Gibbs en el proceso de plegado (ΔGpleg), como se ilustra en la Figura 22.

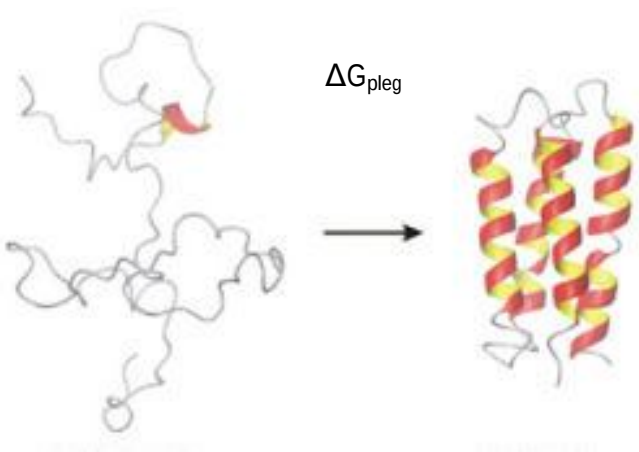
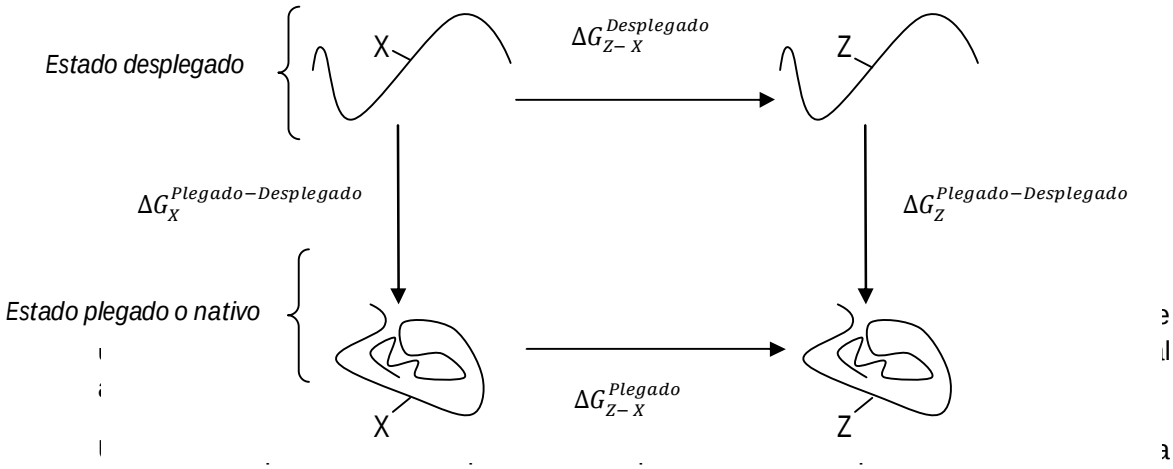


Figura 22. El ΔG de plegado de una proteína representa el nivel de estabilidad del estado plegado respecto al estado desplegado. Se calcula mediante la diferencia entre la energía libre de Gibbs entre el estado plegado y el estado deplegado.

Asimismo, una forma de estudiar la medida en la que afecta una mutación a la estabilidad de la proteína es mediante el cálculo la variación del valor de ΔGpleg entre la proteína wild type y la proteína mutada. Este valor se simboliza ΔΔG, y de acuerdo a las leyendas de la Figura 23, se define como se describe en la Ecuación 1.

$$\Delta\Delta G = \Delta G_{WT}^{Plegado-Desplegado} - \Delta G_{mut}^{Plegado-Desplegado} = \Delta G_{WT-mut}^{Plegado} - \Delta G_{WT-mut}^{Desplegado}$$

Ecuación 1. Cálculo de la variación de la Energía Libre de Gibbs del proceso de plegado de una proteína. El subíndice "WT" hace referencia a la proteína no mutada ("Wild Type"), mientras que el subíndice "Mut" hace referencia a la proteína mutada.




proteica reviste gran importancia en numerosas aplicaciones biotecnológicas y en aéreas de la salud.

Existen numerosos programas orientados a calcular el valor del ΔΔG para una proteína respecto a una mutación, utilizando una diversidad de enfoques, como funciones de energía física efectiva (Gilis & Rooman, 1997; Saraboji, Gromiha, & Ponnuswamy, 2006; Topham, Srinivasan, & Blundell, 1997), potenciales empíricos de energía (Bava, Gromiha, Uedaira, Kitajima, & Sarai, 2004; M Michael Gromiha et al., 1999), métodos de aprendizaje automático (Guerois, Nielsen, & Serrano, 2002) y potenciales estadísticos de energía (Capriotti et al., 2004, 2005).

Si bien estos métodos toman como parámetro la estructura tridimensional de una proteína, ninguno toma en cuenta la diversidad conformacional en ninguna medida. Es decir, se asume rígida la estructura del estado nativo.

Está bien establecido que la divergencia secuencial de una proteína se encuentra condicionada por la diversidad conformacional de su estructura nativa. A su vez, como mostráramos en el capítulo anterior, a pesar de que los distintos confórmeros de una proteína comparten muchos de sus sitios estructuralmente restringidos, cada uno presenta un patrón de clasificación de sitios SCS, UCS y MS propio. La presencia de diferencias estructurales entre los confórmeros genera la presencia de un patrón de sustitución secuencial confórmero-específico. Consecuentemente, la robustez de una proteína a las mutaciones puede depender del confórmero considerado para su estimación. Un análisis del efecto de una mutación sobre un estado nativo considerado como una única estructura dista mucho de ser la visión más realista con la que podemos trabajar. El estado nativo debe ser contemplado con una dinámica asociada, para tener en cuenta los efectos de la mutación en todos los confórmeros de la proteína estudiada.

La hipótesis central en este trabajo es que la diversidad conformacional puede tener una incidencia respecto a la estimación del valor de ΔΔG. Nuestro aporte aquí consiste en tomar en consideración la diversidad conformacional de las proteínas en la estimación de la perturbación estructural inducida por una sustitución.

### 6.3 Metodología

Se vincularon las entradas de proteínas de un solo domino de la base de datos PCDB con la base de datos termodinámica Protherm30 (H. Zhou & Zhou, 2002). Protherm posee valores de ΔΔG de un gran número de mutaciones puntuales sobre diversas proteínas. Un set de datos de 1993 mutaciones puntuales no sinónimas con valores de ΔΔG determinados experimentalmente, cada una asociada a varias estructuras derivadas de la PCDB, fue reclutado de esta manera. Estas 1993 mutaciones pertenecen a 91 proteínas distintas, con 619 confórmeros estructurales en total. La cantidad de estructuras reclutadas por cada proteína oscila entre 2 y 37.

Se descartaron las proteínas con 2 o más dominios con el objetivo de evitar incongruencias entre un valor de ΔΔG experimental calculado en base a una proteína y un valor de ΔΔG teórico estimado en base a uno solo de sus dominios.

Por convención, el ΔΔG se calcula efectuando (ΔGmut - ΔGwild-type), como se señala en la Ecuación 1. Por ende, un valor negativo en el ΔΔG de una mutación es indicador de que ésta genera un efecto desestabilizante en la estructura proteica.

[30] http://www.protherm.eu/

Se efectuaron los cálculos de ΔΔG para las mutaciones pertenecientes al set de datos previamente descripto utilizando 3 (tres) programas ampliamente difundidos para tal fin: FoldX (Lewontin, 1974; Risch & Merikangas, 1996; Stitziel, Binkowski, Tseng, Kasif, & Liang, 2004; Suh & Vijg, 2005), I-Mutant (Sunyaev, Ramensky, & Bork, 2000) y D-Mutant(Stenson et al., 2003).

En una primera fase, se estimaron los valores de ΔΔG utilizando para eso una sola estructura (seleccionada al azar entre todos los confórmeros disponibles) para cada sustitución evaluada. Para obtener resultados estadísticamente significativos, fueron tomadas 10 sub-muestras. Cada sub-muestra está compuesta por una estructura tomada al azar entre todos los confórmeros disponibles de un polipéptido. Posteriormente, para evaluar el valor de ΔΔG de una mutación utilizando una sola estructura, se procedió a promediar los 10 valores de ΔΔG observados. Se evaluó la correlación de los valores teóricos computados por cada uno de los programas por los valores experimentales

En una segunda fase, se estimó el valor de ΔΔG para la mutación correspondiente a cada una de sus estructuras vinculadas, a fin de considerar no sólo una estructura sino un conjunto de confórmeros para describir el estado nativo proteico.

Se calcularon valores estadísticos para cada una de las mutaciones: el valor mínimo de ΔΔG obtenido para entre todas las estructuras, el valor promedio, el valor máximo, y el valor más cercano al ΔΔG experimental.

Finalmente, se compararon las correlaciones obtenidas entre los datos experimentales y los teóricos, ya sea considerando solo una estructura o algún parámetro estadístico derivado de las múltiples estructuras.

Al estimar la correlación entre los valores experimentales y teóricos considerando sólo una estructura, se generaron series de poblaciones de estructuras aleatorias para cada mutación, se calculó la correlación para cada una de ellas y se utilizó el promedio de las correlaciones.

### 6.4 Resultados y discusión

Los valores de ΔΔG de las 1993 mutaciones estudiadas presentan valores comprendidos entre -6.8 kcal/mol y 5.0 kcal/mol, con un promedio de 1.0 kcal/mol. La distribución de los valores de ΔΔG experimentales se representa en la Figura 24.

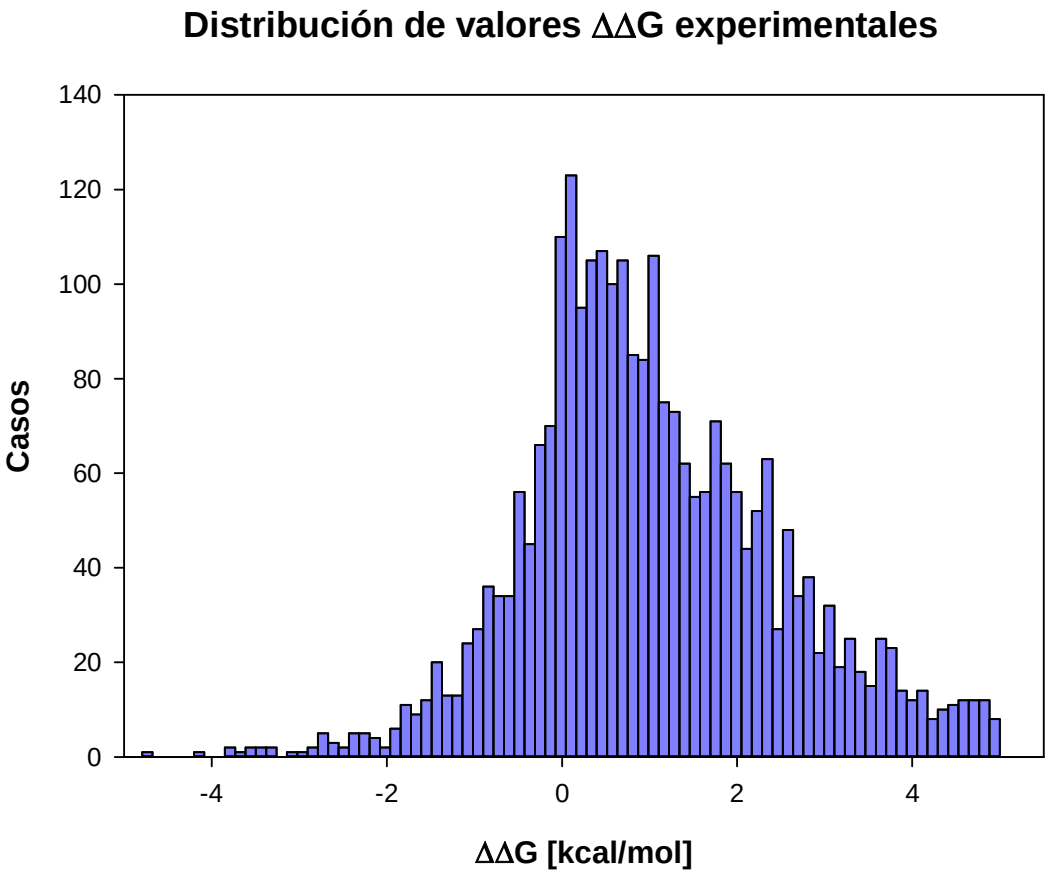



**Figura 24. Distribución de los valores de ΔΔG experimentales de las 1993 mutaciones pertenecientes al set de datos derivados de la base de datos *Protherm.***

El primer resultado notable es la diferencia entre los valores de ΔΔG obtenidos para distintos confórmeros de una proteína, cuando se evalúa una mutación en particular. La distribución de estas diferencias se esquematiza en la Figura 25. Se han detectado diferencias mayores 5 kcal/mol, lo que es un claro indicio de la importancia del confórmero utilizado al efectuar estimaciones teóricas.

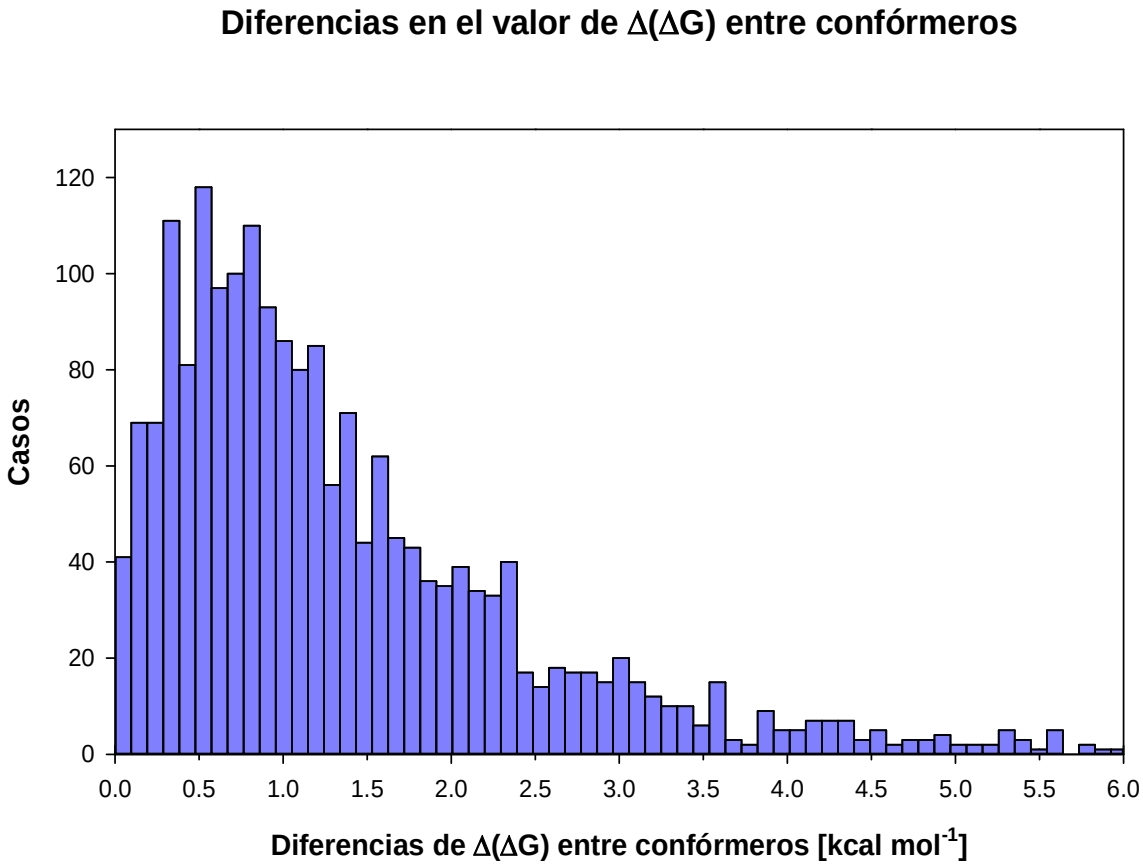



**Figura 25. Distribución de las diferencias entre los valores de ΔΔG de las mutaciones estudiadas obtenidos a partir de distintos confórmeros de la proteína.**

El ASA relativo de un residuo representa un criterio útil para discriminar en residuos ubicados en la superficie de la proteína. Un residuo que presenta un valor de ASA relativo mayor a 25, es considerado como un residuo de superficie, mientras que un residuo que presenta un valor de ASA mayor a 25, se considera un residuo perteneciente al núcleo estructural de la proteína. Al comparar la distribución de los ΔΔG de estos dos tipos de residuos, se observan dos comportamientos distintos, como se aprecia en la Figura 26.

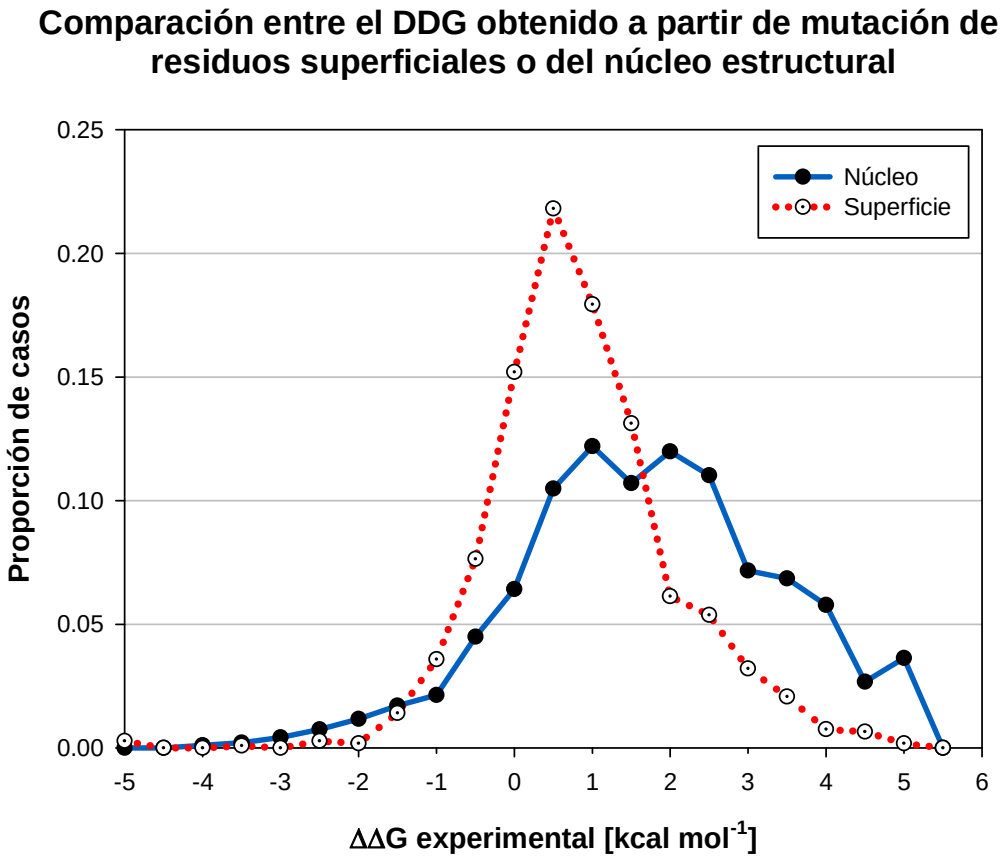



**Figura 26. Distribución diferencial de los valores de ΔΔG obtenidos a partir de mutaciones de residuos de la superficie y de residuos del núcleo estructural de la proteína.**

Se evaluó la existencia de relaciones entre el grado de exposición al solvente del residuo mutado y el nivel de desestabilización que genera en la estructura proteica. Para tal fin se efectuaron cálculos del Area Accesible al Solvente (ASA, del inglés "Accessible Solvent Area") de los residuos mutados mediante el programa Naccess, desarrollado por el Departamento de Bioquímica y Biología Molecular de la Universidad College de Londres.. Se utilizó el valor relativo del ASA. En la Figura 27 se aprecia que, a menores valores de ASA (esto es, cuando el residuo mutado se encuentra dentro del núcleo estructural de la proteína) se corresponden mayores valores de ΔΔG. Cuando el ASA del residuo es grande, es decir, cuando el residuo está altamente expuesto al solvente (y sus contactos con el resto de los residuos son pocos), el valor de ΔΔG se acerca a 0.

Este resultado es esperable, puesto que las mutaciones en el núcleo estructural de la proteína suelen tener un efecto mayor que aquellas mutaciones sobre residuos de su superficie (M. M. Gromiha, Oobatake, Kono, Uedaira, & Sarai, 2000), debido a su mayor número de contactos y su consecuente menor tolerancia a las sustituciones.

Las correlaciones ente los valores experimentales de ΔΔG y los calculados teóricamente por los programas I-Mutant, Fold-X y D-Mutant utilizando una sola

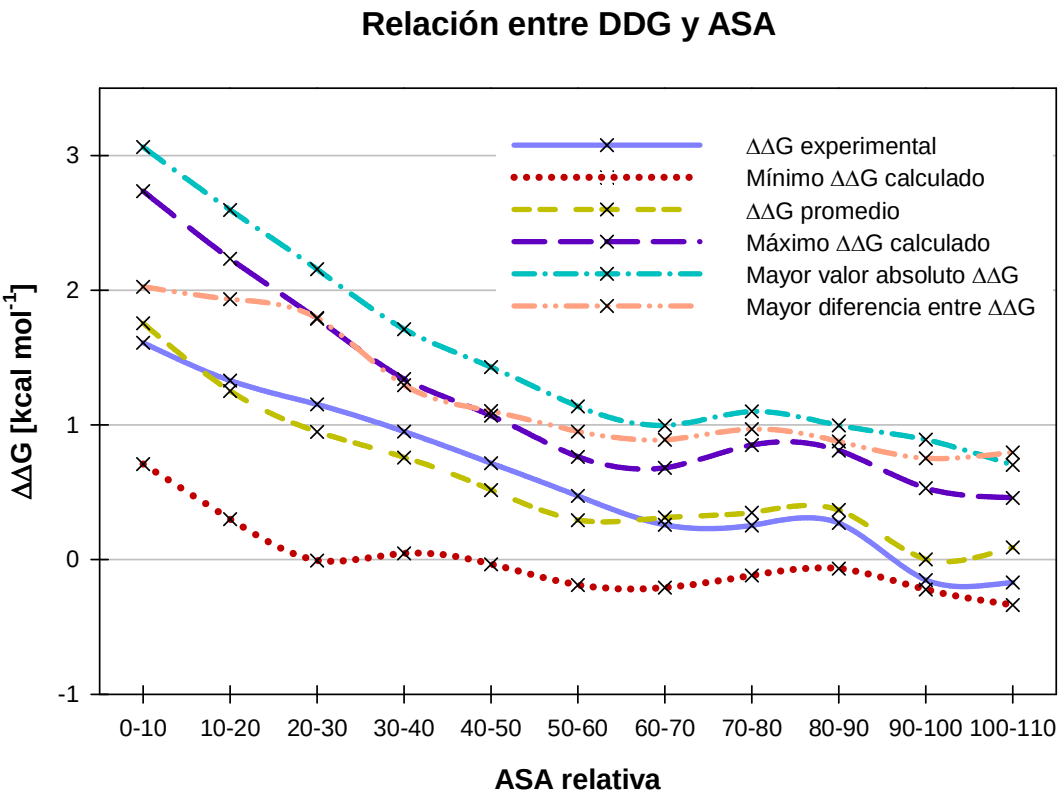



**Figura 27. Relación entre ASA relativa, derivada de cálculos con el programa Naccess, y el valor promedio de ΔΔG de todos los confórmeros considerados para evaluar una mutación.**

estructura por mutación son de 0,53; 0,57 y 0,52 respectivamente. Estas correlaciones de presentan en la Figura 28.

Las correlaciones ente los valores experimentales de ΔΔG y los calculados teóricamente por los programas I-Mutant, Fold-X y D-Mutant utilizando el promedio de todos los valores de ΔΔG calculados para cada confórmero mejoran notablemente, obteniéndose los valores de 0,59; 0,64 y 0,60 respectivamente. Se grafican estos valores en la Figura 29.

Al comparar las correlaciones presentadas en la Figura 28 y las presentadas en laFigura 29, se aprecia la mejora obtenida al considerar la información codificada en la diversidad conformacional, sobre todo cuando se utiliza el programa Fold-X.

Las correlaciones entre los mínimos y los máximos valores de ΔΔG obtenidos entre todos los confórmeros para una determinada proteína y los valores experimentales, no presentan una mejora significativa respecto a la correlación al utilizar una sola estructura.

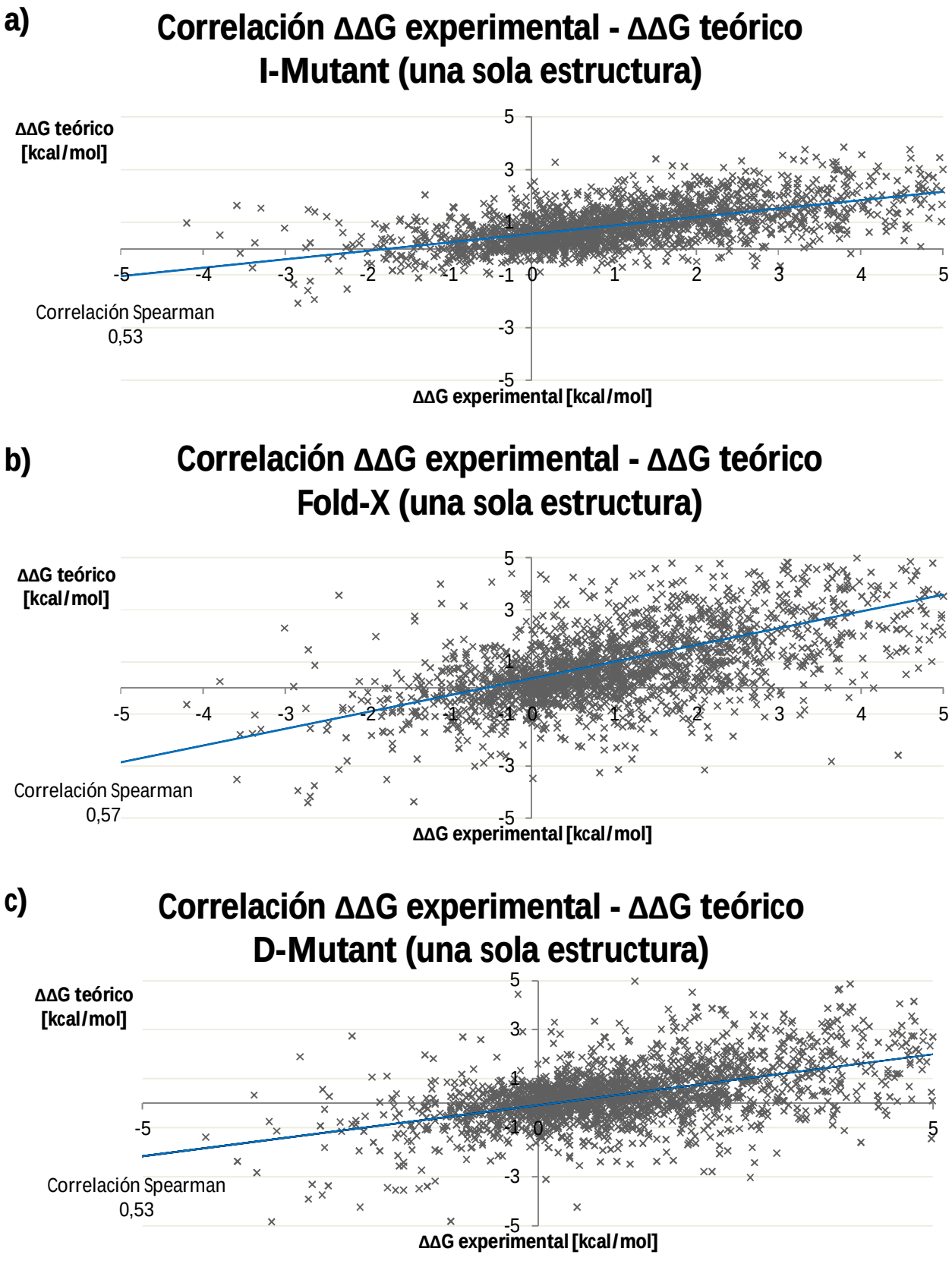



Figura 28. Correlación entre los valores de ΔΔG experimentales y ΔΔG teóricos según los métodos (a) I-Mutant, (b) Fold-X y (c) D-Mutant, utilizando sólo una estructura por mutación.

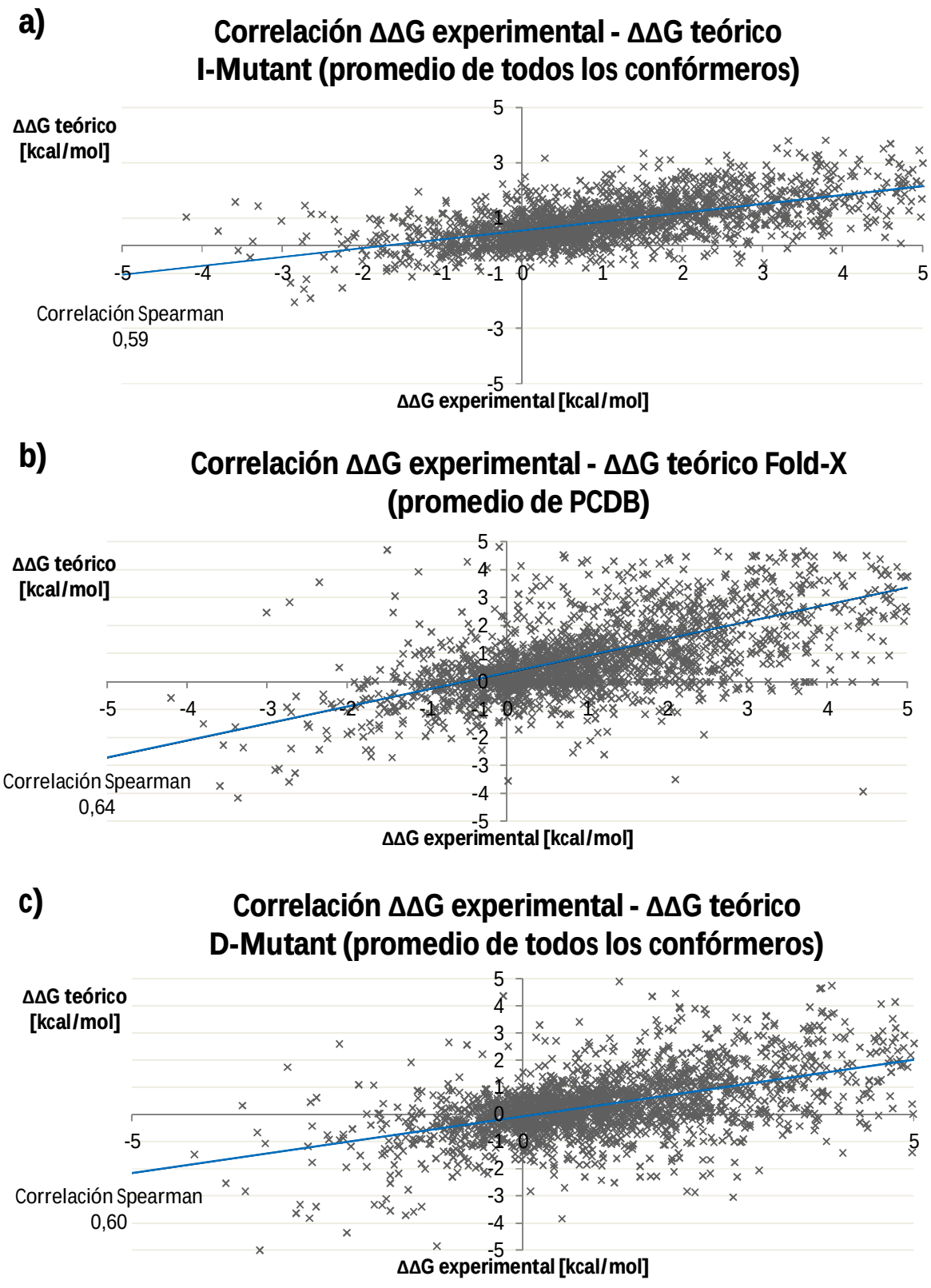



Figura 29. Correlaciones obtenidas por los métodos (a) I-Mutant, (b) Fold-X y (c) D-Mutant, utilizando todos los confórmeros disponibles derivados de la PCDB.

| Método de predicción de ΔΔG | Estadístico de ΔΔG teórico utilizado | Correlación Spearman<br>ΔΔG teórico<br>ΔΔG experimental |
|---|---|---|
| I-Mutant | una sola estructura | 0,53 |
| | promedio de todas las estructuras disponibles en PCDB | 0.59 |
| | confórmero con valor de ΔΔG más cercano al experimental | 0,70 |
| Fold-X | una sola estructura | 0,57 |
| | promedio de todas las estructuras disponibles en PCDB | 0,64 |
| | confórmero con valor de ΔΔG más cercano al experimental | 0,86 |
| D-Mutant | una sola estructura | 0,52 |
| | promedio de todas las estructuras disponibles en PCDB | 0,60 |
| | confórmero con valor de ΔΔG más cercano al experimental | 0,72 |

Tabla 7. Comparación entre las correlaciones de Spearman obtenidas entre los valores experimentales de ΔΔG según Protherm y los valores de ΔΔG teóricos estimados según diferentes métodos, considerando y no considerando la diversidad conformacional derivada de la base de datos PCDB.

Adicionalmente, hemos encontrado que, al utilizar los valores de ΔΔG del confórmero estructural que más se aproxima al valor experimental, la correlación mejora notablemente, como se aprecia en la Tabla 7. Se logra alcanzar un coeficiente de correlación de Spearman de 0,86 con el método Fold-X. Los métodos I-Mutant y D-Mutant también mejoran significativamente su correlación entre los valores teóricos y los experimentales al seleccionar el confórmero con el valor de ΔΔG más cercano al valor experimental.

Aunque este resultado puede parecer trivial, es necesario enfatizar dos aspectos. Por un lado, si bien esperable cierto aumento, los autores destacan el alto incremento en el valor de coeficiente de correlación obtenido.

Por otro lado, este resultado deja abierta la pregunta acerca de cuáles son las características comunes (si es que las hay) de los confórmeros que otorgan el mejor valor de ΔΔG. Muchos esfuerzos apuntan a mejoras en los algoritmos y parámetros utilizados por los métodos de predicción teórica de ΔΔG. Otra perspectiva para abordar la capacidad de estimación del ΔΔG puede ser una correcta selección del confórmero o los confórmeros a utilizar.

## Capítulo 7

## Diversidad conformacional y predicción del efecto de las mutaciones

### 7.1 Resumen

Un mejor entendimiento de los factores que conllevan a una mutación a manifestar una enfermedad constituye un pilar para los avances a futuro en el área de la salud humana (Wang & Moult, 2001). Así como la función de la proteína es mejor explicada mediante la existencia de conjunto de confórmeros en un equilibrio dinámico, creemos que el estudio del efecto de una sustitución aminoacídica sobre una la estructura de una proteína puede ser mejor comprendido mediante su análisis todos los confórmeros estructurales que pueblan su estado nativo.

El objetivo del presente trabajo es analizar si la discriminación entre aquellas sustituciones aminoacídicas relacionadas a una enfermedad y aquellas no relacionadas (variaciones polimórficas o neutras) presenta una mejora al tomar en consideración todas las estructuras disponibles de una proteína. Para tal fin, se utilizó una colección de 803 sustituciones aminoacídicas (482 relacionadas a enfermedades y 323 variaciones neutrales) en 119 proteínas distintas, que muestran diferentes grados de diversidad conformacional. Estas proteínas fueron tomadas de la base de datos de diversidad conformacional de proteínas PCDB. Se estimó el valor del ΔΔG, tratado en el capítulo anterior, para cada sustitución aminoacídica en todas las estructuras disponibles en la PCDB para cada una de las proteínas. Se encontró que el valor estimado para el ΔΔG de una sustitución aminoacídica depende en gran medida la conformación utilizada en la estimación. Considerando como desestabilizante un valor umbral de ΔΔG de ± 2Kcal/mol (esto es, un valor de ΔΔG < -2kcal/mol o bien ΔΔG > 2kcal/mol), hemos encontrado que el 35% de las sustituciones aminoacídicas estudiadas pueden dar lugar a resultados tanto neutros como desestabilizantes, en dependencia del confórmero utilizado. Al considerar el valor umbral de ΔΔG desestabilizante en ± 1 Kcal/mol, el más difundido en la bibliografía, el porcentaje de resultados ambiguos asciende a 58%.

Asimismo, los resultados indican que la consideración de la diversidad conformacional aumenta el rendimiento de la predicción de la enfermedad relacionada con sustituciones aminoacídicas sobre la base de análisis de ΔΔG. Hemos encontrado que los distintos confórmeros correlacionan de diferentes maneras con el fenotipo asociado a la sustitución (enfermedad o neutra) y que, en la mayoría de los casos, existe un confórmero por proteína que presenta una correlación perfecta con el fenotipo correspondiente. Nuestros resultados indican que el uso de la diversidad conformacional

reviste de gran importancia para comprender la relación entre las sustituciones aminoacídicas, sus efectos en la estabilidad de las proteínas y su correspondencia con fenotipos relacionados a enfermedades.

### 7.2 Introducción

Los polimorfismos no sinónimos de nucleótido simple (nsSNPs, de sus siglas en inglés non synonymous Single Nucleotide Polymorphism) representan el tipo más frecuente de variación genética en humanos. La base de datos Humsavar31 clasifica aproximadamente 65.000 nsSNPs del genoma humano como a) variantes relacionados a una enfermedad o b) variantes neutras (esto es, no relacionados a ninguna enfermedad), según estudios y reportes de literatura especializada. 40% de los nsSNPs registrados en esta base de datos están relacionados con algún tipo de enfermedad. El resto de las mutaciones puntuales registradas no causan ningún efecto manifiesto sobre la proteína en cuestión. Por otro lado, algunas mutaciones puntuales tienen un efecto sobre una variedad de aspectos de la proteína (Lofgren & Banerjee, 2011; Orosz, Olah, & Ovadi, 2009; Shah, Bonapace, Hu, Strisciuglio, & Sly, 2004), como lo pueden ser su estructura, su función, su capacidad de unión a ligando, etc. Particularmente, ciertas sustituciones pueden conducir a una enfermedad, debido a que la proteína, por distintas razones, pierde su actividad (Almeida-Souza et al., 2010; Orosz et al., 2009).

Está bien documentado que los polimorfismos de nucleótido pueden causar pérdida de función en diversas maneras. A pesar ser una de las menos frecuente (Wang & Moult, 2001; Yue, Li, & Moult, 2005), la causa más evidente de enfermedad a raíz de un nsSNPs es la sustitución de residuos involucrados directamente en la función proteica, como lo pueden ser residuos pertenecientes al sitio activo, a sitios de unión a sustrato y/o a reguladores alostéricos (Lofgren & Banerjee, 2011). Cuando la unidad funcional está formada por un complejo de proteínas, un nsSNP ubicado en residuos de la interfase de unión interproteica puede alterar la función y provocar una enfermedad (Almeida-Souza et al., 2010; Orosz et al., 2009). Otro posible mecanismo está relacionado con la perturbación de la estabilidad proteica. La sustitución de residuos puede sostener un efecto de desestabilización de la estructura nativa (Wang & Moult, 2001; Yue et al., 2005). A su vez, numerosos nsSNPs que presentan un efecto estabilizante de la estructura proteica nativa han sido asociados a enfermedades (Ling et al., 2010). Por otra parte, el origen de la patogenia causada por nsSNPs se relaciona

[31] http://www.uniprot.org/docs/humsavar

también con modificaciones anómalas post-traduccionales (Alonso, Mederlyova, Novak, Grundke-Iqbal, & Iqbal, 2004) y agregación proteica (C M Dobson, 2001).

Aún así, estudios previos demuestran que en el 80% de los casos en que una mutación simple conlleva a una enfermedad, es debido a la pérdida de estabilidad en la estructura proteica (R Casadio, Vassura, Tiwari, Fariselli, & Luigi Martelli, 2011). Aproximadamente sólo un 10% de los casos restantes se compone de mutaciones que afectan de manera directa ciertos aspectos de la función molecular, y el restante 10% de mutaciones que producen efectos no identificados en la proteína, como lo pueden ser aquellas mutaciones que favorecen el agregado de proteínas.

Es así que la predicción del efecto termodinámico en la estabilidad proteica producido por sustituciones aminoacídicas es un campo de estudio de gran complejidad y que involucra una amplia variedad de causas y efectos. En este trabajo, nuestro objetivo consiste en aportar una mejora a los métodos de predicción de enfermedades causadas por la pérdida de estabilidad de la proteína debido a mutaciones puntuales.

La estabilidad de la estructura proteica se puede estimar mediante una medición de la variación de energía libre de Gibbs (ΔG) entre el plegado y desplegado estado de la proteína, como se ha explicado en detalle en el capítulo anterior.

Es posible establecer un criterio teórico para discriminar, con cierto grado de tolerancia, aquellas sustituciones relacionadas a enfermedades y aquellas sustituciones neutras, analizando el valor del ΔΔG inducido por el cambio de residuos en la proteína (Reumers, Schymkowitz, & Rousseau, 2009). Esto se basa en la noción de que las sustituciones aminoacídicas más perjudiciales para las proteínas están asociadas con perturbaciones termodinámicas de la estructura proteica por encima de un cierto umbral de ΔΔG (aproximadamente de ± 1 kcal/mol).

La precisión y la exactitud de estos métodos ha demostrado ser en promedio moderada. En la mayoría de los métodos, la predicción del efecto termodinámico de la sustitución puntual sobre la proteína se estima utilizando una única estructura de la proteína en estudio. Este enfoque aparentemente subestima el concepto bien establecido de que el estado nativo de una proteína está mejor representada por un conjunto de confórmeros (James & Tawfik, 2003). Desde un punto de vista práctico, la diversidad conformacional podría ser descrita utilizando varias estructuras experimentalmente disponibles de la misma proteína derivadas de la base de datos PCDB. Estas estructuras de la misma proteína pueden ser consideradas como instancias de la dinámica proteica del estado nativo, y representan confórmeros estructurales de la proteína (Best et al., 2006). De esta manera, estimamos que la

descripción del estado nativo de la proteína será más o menos completa dependiendo de las estructuras experimentales disponibles

En el presente trabajo, se investigó cómo la presencia y extensión de la diversidad conformacional afecta a la estimación del ΔΔG en proteínas, tanto con sustituciones neutrales cono en sustituciones relacionadas a enfermedades. Tomando en consideración las estructuras cristalográficas disponibles para una cierta proteína, se estimó en qué medida el efecto de una misma sustitución aminoacídica puede divergir en diferentes confórmeros.

### 7.3 Metodología

Se partió de un set de datos de 3.678 mutaciones en proteínas monodominio, provenientes de la base de datos Humsavar32. 2.281 de estas mutaciones están relacionadas con un fenotipo asociado a una enfermedad, mientras que las restantes 1.395 son mutaciones neutras que no se encuentran asociadas a ninguna enfermedad.

Estas 3.678 mutaciones del set de datos inicial se encuentran distribuidas en 226 proteínas distintas. Con el objetivo de tomar en consideración la diversidad conformacional de estas proteínas, se vincularon estas 226 proteínas con la base de datos PCDB, a fin de reclutar todas las estructuras disponibles resueltas al momento para cada proteína..

Se vinculó cada una de las proteínas presentes en el set de datos con la base de datos PCDB, con el fin de extraer todas las estructuras cristalográficas disponibles para cada proteína del set de datos. La vinculación se efectúa mediante búsquedas limitadas por el código PDB de la proteína en estudio. Luego de este proceso, se reclutaron un total de 1.773 estructuras a partir de las 226 proteínas; resultando en un promedio de 7,8 estructuras reclutadas por proteína. Fue registrado un máximo de 70 estructuras por proteína, y un mínimo de 2.

Se procedió a calcular el RMSD entre los pares de estructuras de proteínas obtenidas de la base de datos PCDB mediante el software Mammoth.

Se utilizaron programas de cálculo teórico para computar los valores de ΔΔG de las mutaciones referidas en el set de datos inicial. Se emplearon cinco diversos programas para calcular el efecto de las mutaciones, ampliamente difundidos: Fold-X (Schymkowitz et al., 2005), I-Mutant (Capriotti et al., 2005), D-Mutant (Hongyi Zhou & Zhou, 2002), Pop-Music (Gilis & Rooman, 2000) y Automute (Masso & Vaisman, 2010). Las

[32] http://www.uniprot.org/docs/humsavar

configuraciones de los programas utilizados, en todos los casos, fueron establecidas con los valores predeterminados por defecto.

En una primera fase, se utilizó una sola estructura de la proteína para estudiar una mutación en particular. Esto es, el procedimiento tradicional para el cálculo de ΔΔG de una proteína ante una mutación.

En una segunda fase, se procedió a computar el ΔΔG de la mutación en todas las estructuras reclutadas de la vinculación con la PCDB. De esta manera, la información codificada en este conjunto de valores de ΔΔG, representa una medida de la perturbación estructural de la mutación en todo el conjunto de confórmeros estructurales del estado nativo. Por criterios propios de cada programa (e.g. longitud de la cadena, anomalías en las numeraciones de los átomos), muchas estructuras no son admitidas como entrada para el cálculo de ΔΔG. En la Tabla 8 se resumen la cantidad de mutaciones estudiadas y el número de cálculos de ΔΔG efectuados por cada programa utilizado.

Se estableció un valor umbral de ΔΔG, a fin de disgregar aquellas mutaciones que desestabilizan en tal grado a la estructura de la proteína que ocasionan la pérdida de su función, y aquellas mutaciones en donde la consecuencia en la estabilidad proteica no es lo suficientemente grande como para hacer esta inferencia. Los valores umbrales aplicados para el presente trabajo dependen del programa utilizado, y fueron establecidos en base a bibliografía específica de los autores de cada programa. Los valores umbrales para cada programa se encuentran especificados en la Tabla 8. Una mutación que introduce un ΔΔG por fuera de dicho rango umbral (ya sea superior o inferior), es considerada como potencialmente asociada a un cuadro patológico. Aquellas mutaciones cuyo ΔΔG computado se encuentra por debajo de él, son clasificadas como mutaciones con efecto neutro.

| Método | Rango de ΔΔG [Kcal/mol] considerado neutral | Número de mutaciones estudiadas | Número de cálculos de ΔΔG considerando la diversidad conformacional (vinculación a PCDB) | Número de cálculos de ΔΔG sin consideración de la diversidad conformacional (sin vinculación a PCDB) |
|---|---|---|---|---|
| **Fold-X** | (-2,1 \| 2,1) | 1.742 mutaciones (496 neutras; 1.246 relacionadas a enfermedad) | 26.802 | 1.742 |
| **I-Mutant** | (-0,8 \| 0,8) | 1.720 mutaciones (483 neutras; 1.237 relacionadas a enfermedad) | 26.327 | 1.720 |
| **D-Mutant** | (-1,4 \| 1,4) | 1.743 mutaciones (497 neutras; 1.246 relacionadas a enfermedad) | 26.986 | 1.743 |
| **POPMUSIC** | (-0,6 \| 0,6) | 1.001 mutaciones (212 neutras; 789 relacionadas a enfermedad) | 15.531 | 1.001 |
| **AUTOMUTE** | Clasificación propia del programa | 1,117 mutaciones (308 neutras; 809 relacionadas a enfermedad) | 16.701 | 1.117 |

**Tabla 8. Número de mutaciones estudiadas y cálculos de ΔΔG realizados por los distintos programas utilizado. Se detallan asimismo los rangos umbrales de ΔΔG clasificados como neutrales para cada programa utilizado.**

En base a los valores obtenidos de ΔΔG para una mutación en una dada estructura proteica, y con lo expuesto anteriormente, es posible inferir si esa mutación puede estar asociada a una enfermedad, debido a la disminución de la función biológica de la misma por pérdida de estabilidad, razón primaria de las enfermedades ocasionadas por mutaciones puntuales. Vale mencionar que el valor calculado de ΔΔG puede ser tanto desestabilizante (en los casos en donde el ΔG de la proteína no mutada es mayor que dicho valor en la proteína mutada), así también como estabilizantes (cuando el valor de ΔG de la proteína mutada es mayor que dicho valor en la proteína no mutada). En el presente trabajo, sostenemos que una mutación que manifiesta un valor de ΔΔG suficientemente alejado del 0 (esto es, altamente estabilizante o desestabilizante), debe ser considerada como desestabilizante y, por lo tanto, posiblemente asociada a un fenotipo patogénico. Esto se sustenta, en parte, por el hecho de que una sustitución aminoacídica que estabilice demasiado un confórmero en particular puede perturbar el equilibrio dinámico del estado nativo, inhibiendo la función biológica.

En la segunda fase, al disponer de un cierto número de estructuras sobre las cuales computar el cálculo de ΔΔG para cada mutación en una cierta proteína, se tomaron varios valores estadísticos: El valor promedio de todos los ΔΔG calculados a partir de todas las estructuras de dicha proteína respecto a la sustitución estudiada, el valor máximo, el valor mínimo, y el mayor valor absoluto.

Una vez calculados los valores de ΔΔG para cada mutación en la primera fase, se clasificaron las mutaciones en neutras o asociadas a enfermedad de acuerdo a si caían dentro o fuera del valor umbral establecido para cada método usado. Se determinó el Matthews Correlation Coefficient (MCC), la precisión, la especificidad y la sensibilidad para comparar cuán bien se ajustan la clasificación teórica de cada uno de los programas utilizados con los datos derivados de la base de datos Humsavar, considerando y no considerando la diversidad conformacional.

Se detallan en la Ecuación 2 las fórmulas de los estadísticos utilizados para efectuar las comparaciones.

$$MCC = \frac{(pv)(nv) - (nf)(pf)}{\sqrt{(pv + nf)(pv + pf)(nv + nf)(nv + pf)}}$$

$$Precisión = \frac{pv + nv}{pv + nv + pf + nf}$$

$$Especificidad = \frac{nv}{nv + pf}$$

$$Sensibilidad = \frac{pv}{pv + nf}$$

Ecuación 2. Fórmulas de los estadísticos utilizados. pv = Positivos verdaderos. pf = Positivos falsos. nv = Negativos verdaderos. nf = Negativos falsos.

### 7.4 Resultados

#### 7.4.1 Dversidad conformacional

En la Figura 30 se expone la distribución de los valores de RMSD máximos obtenidos por el programa Mammoth entre un los pares de confórmeros estructurales para cada una de las proteínas utilizadas en el presente trabajo. El promedio de RMSD entre todos los pares posibles de estructuras de la misma proteína es 2,5 Å; presentando un mínimo de 0,0 Å y un máximo de 21,6 Å. Se aprecia un máximo de

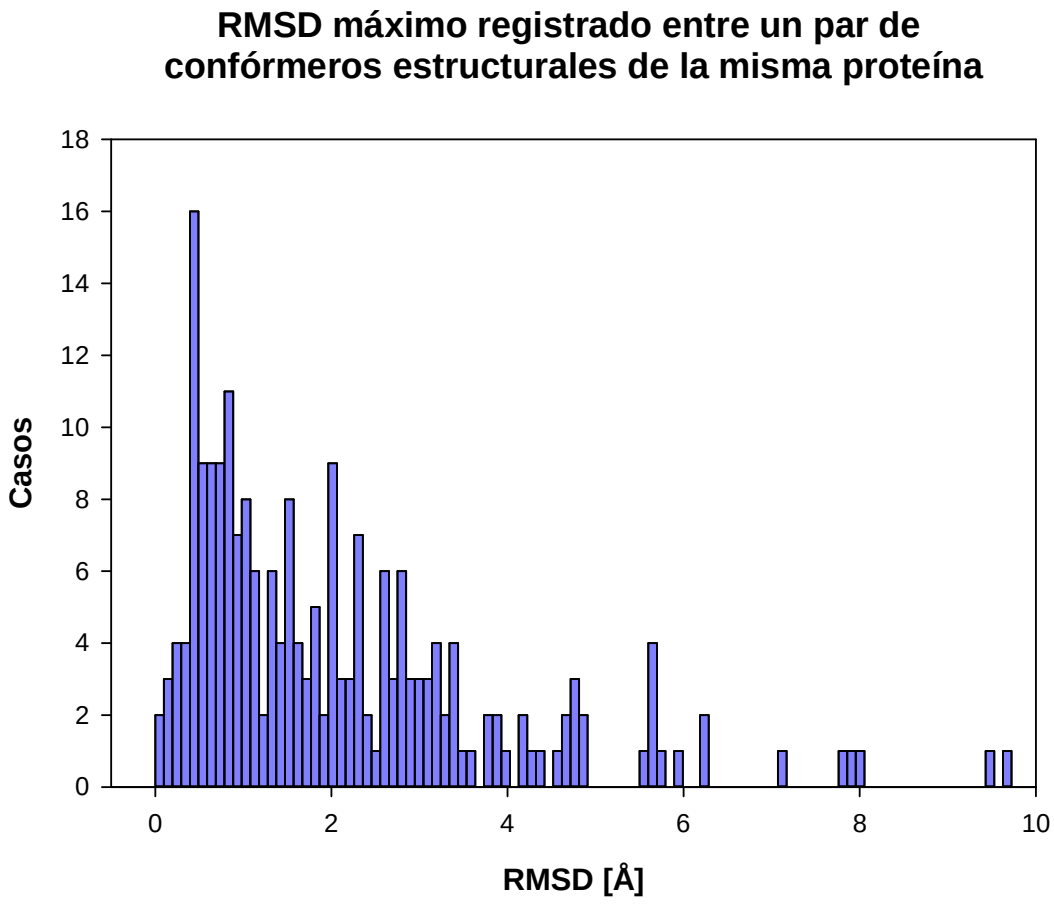



Figura 30. Distribución del RMSD máximo registrado (representativo de la diversidad conformacional proteica) entre pares de confórmeros estructurales de la misma proteína.

casos en 0,6 Å, así también como la presencia de casos con diversidad conformacional mayor a 4 Å. Teniendo en cuenta que el promedio registrado de RMSD entre dos estructuras resueltas de la misma proteína cristalizada bajo las mismas condiciones oscila entre 0,1 Å y 0,4 Å (Burra et al., 2009), a partir de la distribución de la Figura 30, llegamos a la conclusión de que el set de datos utilizado contiene proteínas con diversidad conformacional entre moderada y extrema.

### 7.4.2 Diferencias en los valores de ΔΔG obtenidos

Nuestros resultados indican que las variaciones de los ΔΔG obtenidos a partir de distintos confórmeros, para una misma mutación de una misma proteínas, son altamente significativos. En la Figura 31 se exponen los porcentajes acumulados de las

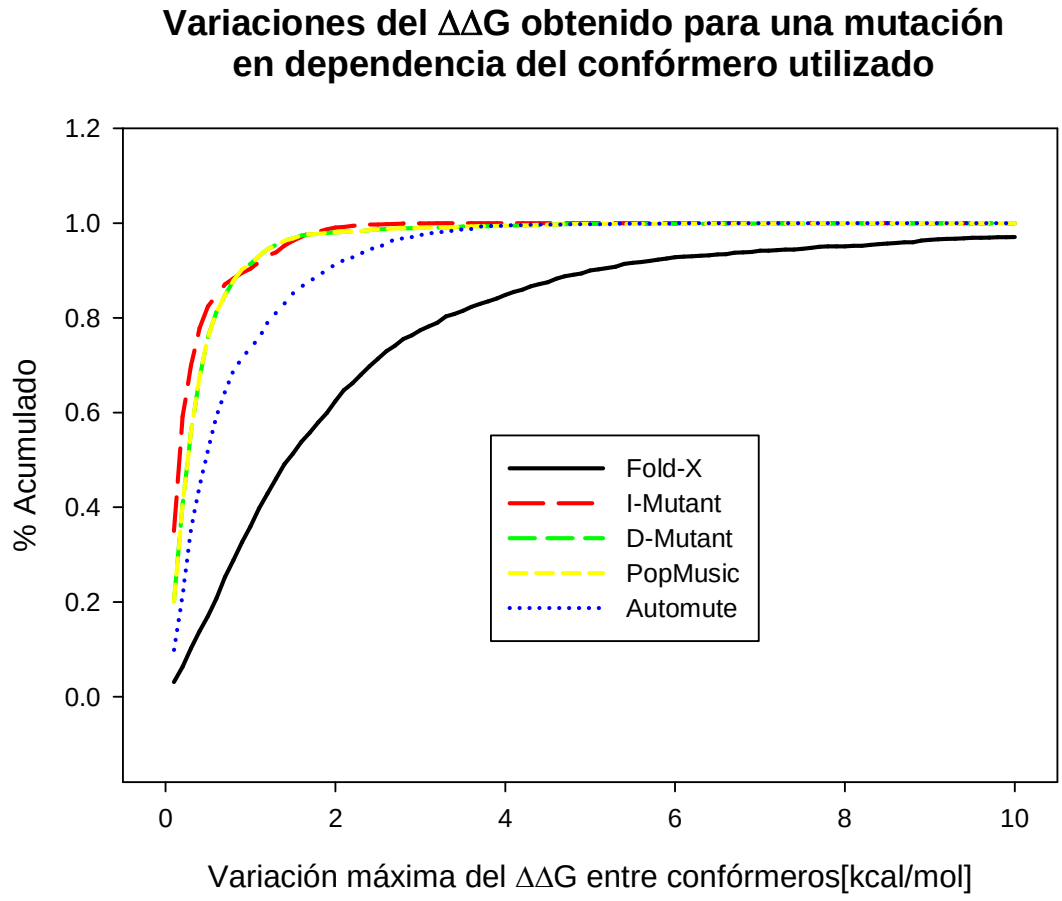



**Figura 31. Distribución acumulada de los valores de las variaciones de ΔΔG obtenidos mediante distintos métodos entre distintos confórmeros de una misma proteína.**

variaciones de los ΔΔG obtenidos mediante distintos métodos en dependencia del confórmero utilizado. Se aprecia que ciertos métodos, particularmente el Fold-X, son más susceptibles a la estructura utilizada que otros métodos. Este hecho se debe al algoritmo utilizado y a los parámetros que cada método contempla a la hora de efectuar el cálculo.

Nuestros resultados indican que muchas mutaciones son consideradas como altamente desestabilizantes para ciertos confórmeros de la proteína, mientras que en otros confórmeros, esa misma mutación no produce alteración significativa de la estabilidad proteica.

En la Figura 32 se compara el número de mutaciones en donde todos los confórmeros de la proteína mutada dan lugar a valores de ΔΔG desestabilizantes, el número de mutaciones en donde todos los confórmeros de la proteína mutada dan lugar a valores de ΔΔG desestabilizantes, y el número de mutaciones en donde hay confórmeros que son desestabilizados y otros que no lo son, para los programas Fold-X, I-Mutant y PopMusic, los 3 métodos computacionales que mejor se correlacionan con los valores experimentales según la Tabla 9 (ver más adelante). Cabe aclarar que estos métodos no fueron aplicados exactamente al mismo conjunto de mutaciones, puesto que cada método efectúa sus propios filtros sobre los datos de entrada. Aún así, el 90% de las proteínas utilizadas son compartidos entre las poblaciones utilizadas para todos los métodos.

Es interesante notar la alta proporción de resultados que dan una y otra predicción, en función del confórmero proteico utilizado. Este hecho indica la importancia de tomar en cuenta la información de la diversidad conformacional en estimaciones estructurales.

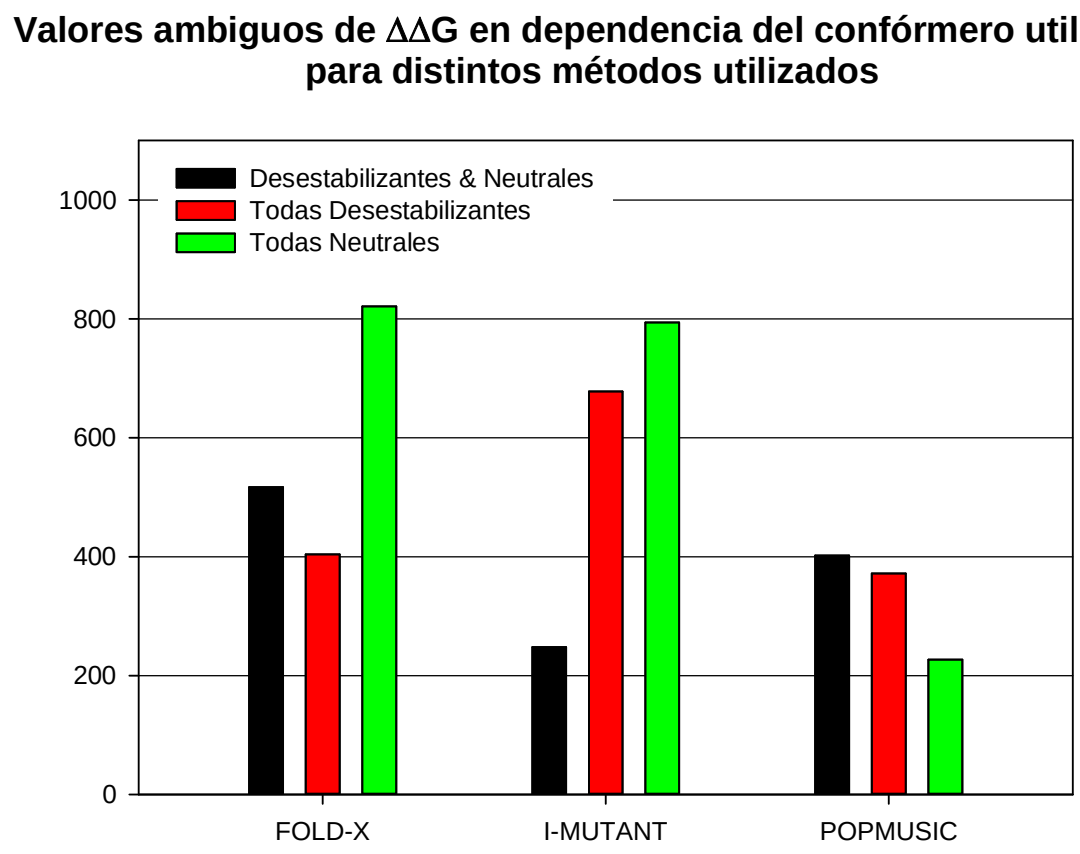



Figura 32. Comparación del número de mutaciones clasificadas como desestabilizantes para todos sus confórmeros, como neutrales para todos sus confórmeros y aquellas en donde hay confórmeros desestabilizantes y otros neutrales (mutaciones con resultados ambiguos de ΔΔG).

1hrzA00
1hryA00

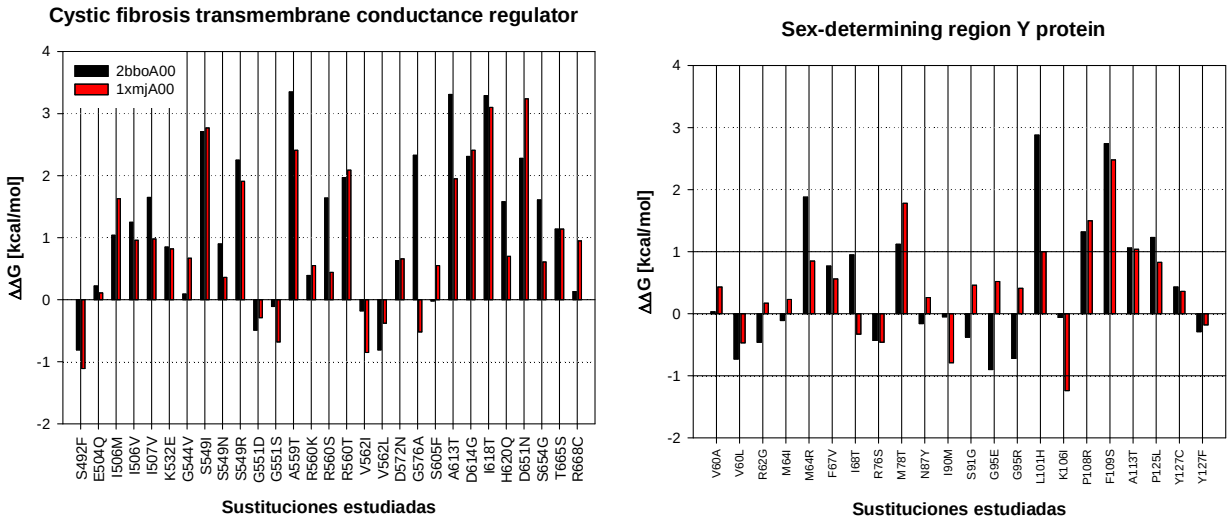



**Figura 33. Comparaciones de los valores de ΔΔG obtenidos para distintas mutaciones en dos proteínas (*Regulador de la conductancia transmembrana de la fibrosis quística* y la *Proteína de la región Y determinante del sexo*) utilizando para su estimación dos confórmeros distintos de cada proteína. Las mutaciones estudiadas se representan sobre el eje horizontal. Se grafica en barras negras los valores obtenidos en base a un confórmero, y en barras rojas los valores obtenidos en base al otro confórmero utilizado. Se aprecian significantes diferencias en ciertas mutaciones estudiadas.**

En la Figura 33 se muestran dos casos proteínas, con dos confórmeros disponibles cada una. Se aprecia en este gráfico, por ejemplo en las mutaciones S91G, G95E y G95R cómo una misma mutación puede ejercer distintos efectos sobre confórmeros distintos. Cabe destacar que no se asume que las estructuras disponibles completan todos los confórmeros presentes en el estado nativo.

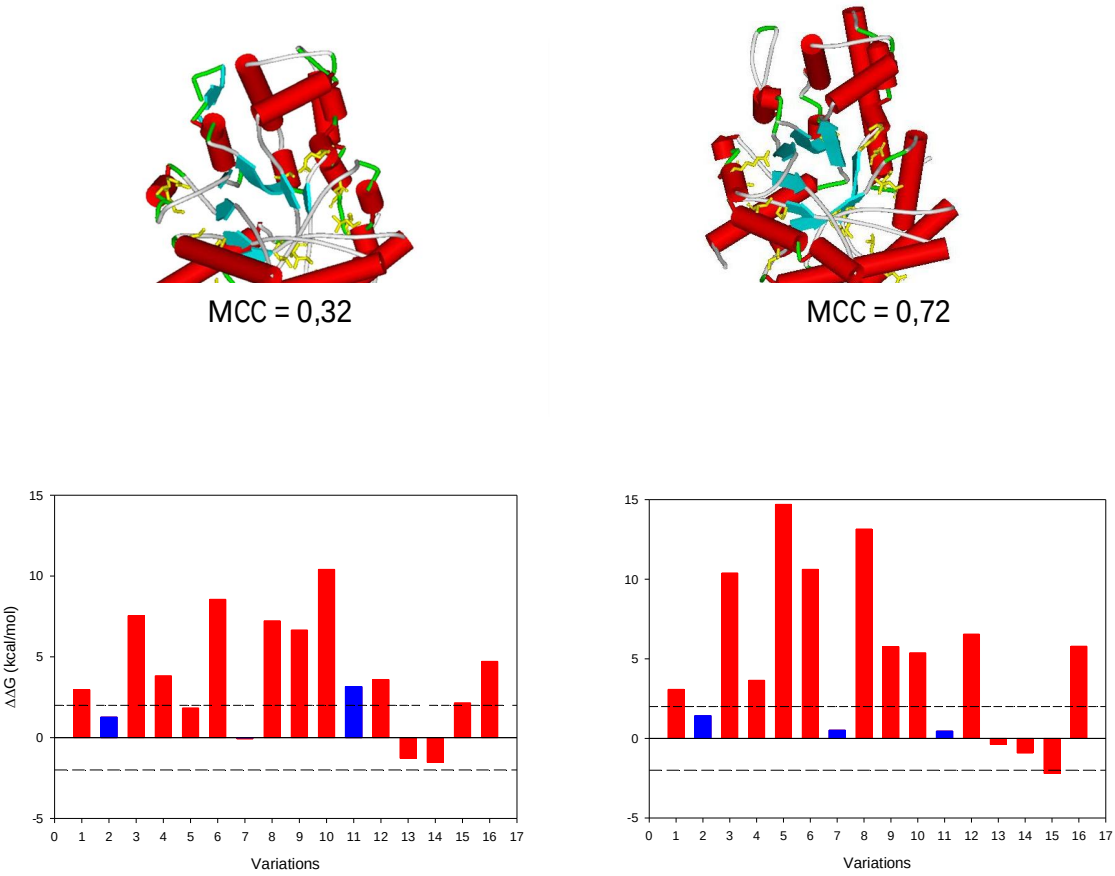



**Figura 34. Ejemplo que muestra cómo la diversidad conformacional afecta a la estimación del valor de ΔΔG y la MCC. La proteína de fructosa-bifosfato aldolasa B (P05062). Los dos confórmeros pertenecientes a esta proteína presentan un RMSD de 1,3 Å. En la parte superior de la figura se muestra una representación gráfica de los dos confórmeros con los ácidos mutados amino en color amarillo. Los gráficos de barra representan los valores ΔΔG (en rojo para las mutaciones asociadas a enfermedades y en azul para las mutaciones neutras). Se señala el valor correspondiente de MCC para cada conformación. Las flechas negras indican las predicciones equivocadas basadas en el intervalo de referencia de ± 2kcal/mol. Como los cambios estructurales producen variaciones en ΔΔG, los confórmeros presentan valores de MCC diversos.**

| | Fold-X (-2,1\|2,1) | | I-Mutant (-0,8 \| 0,8) | | D-Mutant (-1,4 \| 1,4) | | POP-Music (-0,6 \| 0,6) | | AUTOMUTE | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *sin PCDB* | *con PCDB* | *sin PCDB* | *con PCDB* | *sin PCDB* | *con PCDB* | *sin PCDB* | *con PCDB* | *sin PCDB* | *con PCDB* |
| MCC | 0,24 | **0,31** | 0,18 | 0,19 | 0,19 | 0,21 | 0,25 | 0,32 | 0,09 | 0,16 |
| Precisión | 0,54 | 0,67 | 0,60 | 0,63 | 0,47 | 0,51 | 0,68 | 0,80 | 0,41 | 0,71 |
| Sensibilidad para enfermedad | 0,46 | 0,65 | 0,59 | 0,67 | 0,31 | 0,38 | 0,7 | 0,85 | 0,27 | 0,87 |
| Sensibilidad para neutras | 0,80 | 0,72 | 0,60 | 0,53 | 0,87 | 0,84 | 0,6 | 0,48 | 0,82 | 0,27 |
| Especificidad para enfermedad | 0,88 | 0,88 | 0,79 | 0,78 | 0,86 | 0,86 | 0,86 | 0,86 | 0,80 | 0,77 |
| Especificidad para neutras | 0,32 | 0,38 | 0,37 | 0,39 | 0,33 | 0,35 | 0,35 | 0,45 | 0,28 | 0,42 |

**Tabla 9. Se exhibe aquí las mejoras en diversos parámetros que se obtienen al considerar más de una estructura en la predicción de estabilidad de una mutación en una proteína. Se detalla bajo cada método de predicción de DDG el valor umbral utilizado para decidir si una mutación es desestabilizante o no.**

Fueron efectuadas una serie de comparaciones entre los valores estadísticos obtenidos en la primera y en la segunda fase, con el fin de evaluar si la inclusión de la diversidad conformacional se traduce en un incremento en la correlación entre los valores obtenidos teóricamente y aquellos experimentales.

Se presenta en la Tabla 9 la comparación entre dichos resultados. Estos mismos valores comparados mediante un gráfico de barras se presenta en la Figura 35. Se aprecia una mejora significativa en la correlación obtenida al incluir los confórmeros estructurales. Este hecho que sugiere que la información codificada en la diversidad conformacional puede abrir una línea de investigación para incrementar la capacidad de predicción del efecto de mutaciones sobre una proteína.

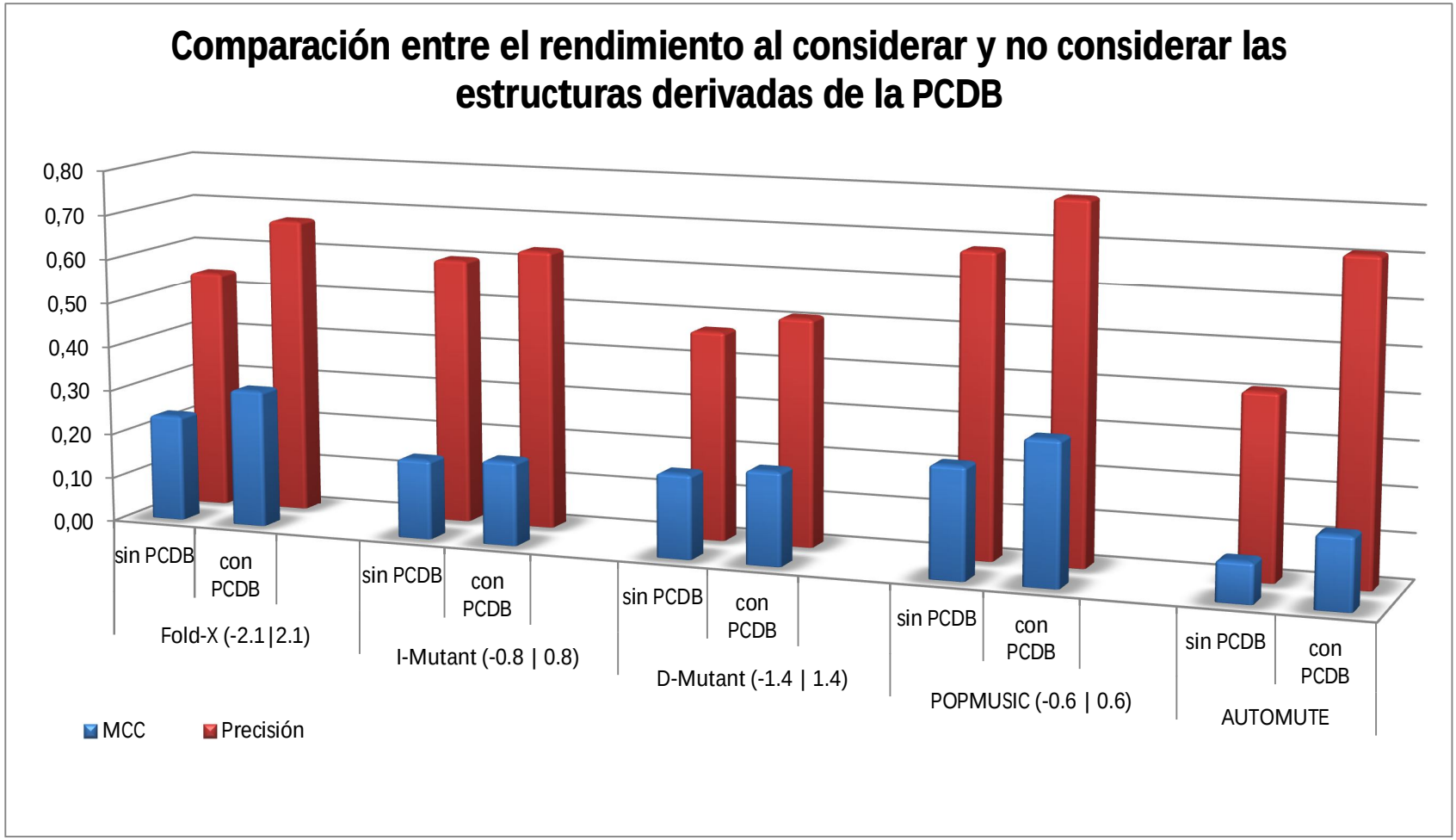



Figura 35. Comparación entre los valores de MCC (Matthews Correlation Coefficient) y la precisión para los métodos de predicción de enfermedades, considerando o no la diversidad conformacional contenida en la base de datos PCDB. Se denota siempre un mejor desempeño en todos los métodos al incluir explícitamente la información codificada en la diversidad conformacional.

### 7.5 Conclusiones

En primer término es necesario destacar que las diferencias entre los valores de ΔΔG en dependencia del confórmero utilizado para evaluar una sustitución dada en una cierta proteína no son despreciables. Este resultado pone en tela de juicio el nivel de confiabilidad con el que debe evaluarse un valor de ΔΔG obtenido a partir de una sola estructura y pone de manifiesto la importancia de considerar más de una estructura en la aplicación de métodos de predicción teóricos.

En todos los métodos se ha logrado una mejora sobre todos los parámetros estadísticos estudiados al considerar la diversidad conformacional proteica en la predicción de si una mutación conllevará a una enfermedad o no. En ciertos métodos de predicción de enfermedades estas mejoras son más significativas que en otros. Esta diferencia es atribuible a las diferencias entre los algoritmos de cada método de predicción, en donde se le confiere importancia a diversos parámetros.

Si bien incluso es posible que en algunos métodos el incremento en el costo computacional debido a la vinculación con la base de datos PCDB y a los cálculos adicionales no se justifiquen, los resultados no dejan de ser alentadores, instando a seguir esta línea de investigación a fin de perfeccionar la inclusión de la información codificada en la diversidad conformacional para efectuar diversas predicciones por métodos computacionales.

A su vez, las estructuras disponibles en la actualidad, utilizadas en el presente trabajo, pueden no ser una completa descripción del estado nativo. Estimamos que el incremento en el volumen de estructuras y/o el avance de los métodos teóricos de dinámica molecular pueden lograr una descripción más completa en el futuro.

Consideramos que esta mejora se puede incrementar mediante técnicas más complejas de introducir información codificada en la diversidad conformacional.

## Conclusiones

En este estudio hemos encontrado que el uso de la información codificada en la diversidad conformacional estructural mejora el desempeño de métodos de biología computacional y modula la divergencia de secuencias de proteínas.

Nuestros resultados demuestran la relevancia de considerar la diversidad conformacional al estudiar el mecanismo de evolución de las proteínas, y, por lo tanto, la posibilidad de desarrollar mejores modelos de evolución molecular.

Adicionalmente, el estudio del patrón de sustitución, así como la comprensión de la conservación de la heterogeneidad en las secuencias son temas centrales en diversas técnicas de bioinformática. Sugerimos en este contexto que la consideración de la diversidad conformacional y el sesgo que exhiben en el patrón de sustitución aminoacídico constituye un pilar fundamental de la próxima generación de herramientas de biología computacional.

El vínculo entre el patrón de sustitución de residuos y la extensión de la diversidad conformacional es un campo prometedor para aumentar nuestro entendimiento sobre la evolución de proteínas y, sin embargo, es un campo no explorado aún. Adicionalmente, y siguiendo la línea de investigaciones previas, otro objetivo consiste en enriquecer el volumen de estructuras en la base de datos PCDB considerando proteínas homólogas cercanas, con el fin de aumentar la representación de los distintos confórmeros del dominio representado.

Adicionalmente, la consideración de distintos confórmeros para la predicción de la perturbación estructural que una mutación puntual ejerce sobre una proteína mejora la correlación con los datos experimentales. Esto puede ser debido a que los cálculos de ΔΔG experimentales son un promedio de todos los confórmeros estructurales presentes en el estado nativo de la estructura proteica en estudio. A pesar de los avances en las técnicas teóricas de cálculo de ΔΔG, es necesario avanzar en la consideración de la diversidad conformacional para lograr una aproximación más realista.

La predicción de si una mutación puede conllevar a un fenotipo patogénico es una aplicación del cálculo teórico del ΔΔG. La importancia y la aplicación de estas investigaciones en el campo de la salud son trascendentales, y muchos estudios abordan el incremento de esta capacidad predictiva. Nuestros resultados, basados en una variedad de predictores de valores de ΔΔG, indican que esta capacidad puede ser aumentada al tomar en consideración la diversidad conformacional de las proteínas, derivada de la base de datos PCDB.

Consideramos que este hecho puede representar una nueva línea de investigación bioinformática para incrementar la capacidad predictiva de los métodos teórico-computacionales en general.

**Perspectivas**

Nuestro interés radica en incrementar la disponibilidad en la cantidad y diversidad de datos biológicos y estructurales disponibles para cada proteína representada en la base de datos, en pos de mejorar las posibles correlaciones encontradas entre diversidad conformacional y un amplio espectro de parámetros fisicoquímicos. Una de nuestras metas a corto plazo es introducir la secuencia de alineamientos para cada proteína depositada, para obtener información evolutiva, como la conservación relativa en diferentes posiciones y las tasas de evolución.

Un estudio a futuro consiste en explorar mediante métodos evolutivos la relación entre la robustez a las mutaciones y el tipo de plegamiento, para estudiar las relaciones entre determinadas patologías y ciertos tipos de plegamientos, que posiblemente presenten una robustez disminuida a las mutaciones, y por ende sean más propensos al desarrollo de enfermedades.

Este tipo de análisis puede terminar en el desarrollo de herramientas computacionales para predecir el efecto de las mutaciones sobre la función biológica considerando explícitamente su diversidad conformacional y estudiar su correlación con el establecimiento de enfermedades.

Mi tesis de posgrado estará fundamentada en estos estudios preliminares, y tendrá como objetivo a largo plazo el desarrollo de un método para predicción de enfermedades por mutaciones puntuales en proteínas a partir del estudio del efecto de éstas sobre la estabilidad de las estructuras de sus confórmeros.

Adicionalmente, este método a desarrollar incluirá información concerniente a los sitios catalíticamente activos de cada proteína, datos ya incluidos para tal fin en la base de datos PCDB. La consideración de la diversidad conformacional y de los sitios activos supondrá una mejora significativa en el desempeño respecto a los métodos actuales de predicción de enfermedades.

**Referencias**


Abhiman, S., & Sonnhammer, E. L. L. (2005). FunShift: a database of function shift analysis on protein subfamilies. Nucleic acids research, 33(Database issue), D197-200. doi:10.1093/nar/gki067

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., & Tawfik, D. S. (2005). The “evolvability” of promiscuous protein functions. Nature genetics, 37(1), 73-6. doi:10.1038/ng1482

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Almeida-Souza, L., Goethals, S., de Winter, V., Dierick, I., Gallardo, R., Van Durme, J., Irobi, J., et al. (2010). Increased monomerization of mutant HSPB1 leads to protein hyperactivity in Charcot-Marie-Tooth neuropathy. J Biol Chem, 285(17), 12778-12786.

Alonso, A. del C., Mederlyova, A., Novak, M., Grundke-Iqbal, I., & Iqbal, K. (2004). Promotion of hyperphosphorylation by frontotemporal dementia tau mutations. J Biol Chem, 279(33), 34873-34881.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., et al. (2000). Gene Ontology : tool for the unification of biology. Nature genetics, 25(1), 25-29. doi:10.1038/75556.Gene

Bahar, Ivet, & Rader, a J. (2005). Coarse-grained normal mode analysis in structural biology. Current opinion in structural biology, 15(5), 586-92. doi:10.1016/j.sbi.2005.08.007

Bai, F., Branch, R. W., Nicolau, D. V., Pilizota, T., Steel, B. C., Maini, P. K., & Berry, R. M. (2010). Conformational spread as a mechanism for cooperativity in the bacterial flagellar switch. Science (New York, N.Y.), 327(5966), 685-9. doi:10.1126/science.1182105

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic acids research, 28(1), 304-5. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102465&tool=pmcentrez&rendertype=abstract

Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of Catalytic Residues in Enzyme Active Sites. Journal of Molecular Biology, 324(1), 105-121. doi:10.1016/S0022-2836(02)01036-7

Bash, P. a, Singh, U. C., Langridge, R., & Kollman, P. a. (1987). Free energy calculations by computer simulation. Science (New York, N.Y.), 236(4801), 564-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3576184

Bashton, M., Nobeli, I., & Thornton, J. M. (2008). PROCOGNATE: a cognate ligand domain mapping for enzymes. Nucleic acids research, 36(Database issue), D618-22. doi:10.1093/nar/gkm611

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., et al. (2004). The Pfam protein families database. Nucleic acids research, 32(Database issue), D138-41. doi:10.1093/nar/gkh121

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic acids research, 32(Database issue), D120-1. doi:10.1093/nar/gkh082

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., et al. (2000). The Protein Data Bank. Nucleic acids research, 28(1), 235-42. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract

Bernardes, A., Batista, F. a H., de Oliveira Neto, M., Figueira, A. C. M., Webb, P., Saidemberg, D., Palma, M. S., et al. (2012). Low-Resolution Molecular Models Reveal the Oligomeric State of the PPAR and the Conformational Organization of Its Domains in Solution. PloS one, 7(2), e31852. doi:10.1371/journal.pone.0031852

Berrera, M., Molinari, H., & Fogolari, F. (2003). Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics, 4(8).

Best, R. B., Clarke, J., & Karplus, M. (2005). What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. Journal of molecular biology, 349(1), 185-203. doi:10.1016/j.jmb.2005.03.001

Best, R. B., Lindorff-Larsen, K., DePristo, M. A., & Vendruscolo, M. (2006). Relation between native ensembles and experimental structures of proteins. Proceedings of the National Academy of Sciences, 103(29), 10901–10906. NATL ACAD SCIENCES. doi:10.1073/pna.0511156103

Bocharov, E. V., Mayzel, M. L., Volynsky, P. E., Goncharuk, M. V., Ermolyuk, Y. S., Schulga, A. a, Artemenko, E. O., et al. (2008). Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. The Journal of biological chemistry, 283(43), 29385-95. doi:10.1074/jbc.M803089200

Boehr, D. D., McElheny, D., Dyson, H. J., & Wright, P. E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. Science (New York, N.Y.), 313(5793), 1638-42. doi:10.1126/science.1130258

Bryngelson, J. D., & Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model with applications to protein folding. J Phys Chem, 93, 6902– 6915.

Bryngelson, Joseph D, Nelson, J., Nicholas, O., & Wolynes, P. G. (1995). Funnels , Pathways , and the Energy Landscape of Protein Folding : A Synthesis. PROTEINS: Structure, Function, and Genetic, 21(3), 167-195. Retrieved from http://wolynes.ucsd.edu/Wolynes_Papers/Funnels Pathways 135.pdf

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. New York: Springer.

Burra, P. V., Zhang, Y., Godzik, A., & Stec, B. (2009). Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. PNAS, 106(26), 10505–10510.

Capriotti, E., Fariselli, P., & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics (Oxford, England), 20 Suppl 1, i63-i68. doi:10.1093/bioinformatics/bth928

Capriotti, E., Fariselli, P., Calabrese, R., & Casadio, R. (2005). Predicting protein stability changes from sequences using support vector machines. Bioinformatics (Oxford, England), 21 Suppl 2, ii54-8. doi:10.1093/bioinformatics/bti1109

Casadio, R, Vassura, M., Tiwari, S., Fariselli, P., & Luigi Martelli, P. L. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. Hum Mutat, 32(10), 1161-1170.

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins, 62(4), 1125-32. doi:10.1002/prot.20810

Chng, C.-P., & Yang, L.-W. (2008). Coarse-grained models reveal functional dynamics--II. Molecular dynamics simulation at the coarse-grained level--theories and biological applications. Bioinformatics and biology insights, 2, 171-85. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2735960&tool=pmcentrez&rendertype=abstract

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO Journal, 5(4), 823-826. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1166865/pdf/emboj00167-0187.pdf

Collaborative_Computational_Project. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr, 50(Pt5), 760-763.

Consortium, T. G. O. (2008). The Gene Ontology project in 2008. Nucleic acids research, 36(Database issue), D440-4. doi:10.1093/nar/gkm883

Copley, S. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr. Opin. Chem. Biol., 7, 1–8.

Dobson, C M. (2001). The structural basis of protein folding and its links with human disease. Philos Trans R Soc Lond B Biol Sci, 356(1406), 133-145.

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell, 134(2), 341-52. doi:10.1016/j.cell.2008.05.042

Felsenstein, J. (1989). Mathematics vs. Evolution: Mathematical Evolutionary Theory. Science, 246(4932), 941-942.

Felsenstein, Joseph. (1981). Evolutionary Trees from DNA Sequences : A Maximum Likelihood Approach. Journal of Molecular Evolution, 17(6), 368-376.

Ferrada, E., & Melo, F. (2009). Effective knowledge-based potentials. Protein science : a publication of the Protein Society, 18(7), 1469-85. doi:10.1002/pro.166

Ferreiro, D. U., Hegler, J. a, Komives, E. a, & Wolynes, P. G. (2007). Localizing frustration in native proteins and protein assemblies. Proceedings of the National Academy of Sciences of the United States of America, 104(50), 19819-24. doi:10.1073/pnas.0709915104

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39 Suppl 2(May), W29-37. doi:10.1093/nar/gkr367

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., et al. (2008). The Pfam protein families database. Nucleic acids research, 36(Database issue), D281-8. doi:10.1093/nar/gkm960

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft, 27(3), 2985–2993. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/cber.18940270364/abstract

Footet, J., & Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. Proc. Natl. Acad. Sci. USA, 91(October), 10370-10374.

Fornasari, M. S., Parisi, G., & Echave, J. (2002). Site-Specific Amino Acid Replacement Matrices from Structurally Constrained Protein Evolution Simulations. Molecular biology and evolution, 19(3), 352-356. Retrieved from http://mbe.oxfordjournals.org/content/19/3/352.full.pdf+html

Frenz, C. M. (2005). Neural network-based prediction of mutation-induced protein stability changes in Staphylococcal nuclease at 20 residue positions. Proteins: Structure, Function, and Bioinformatics, 59(2), 147-151. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/prot.20400/abstract

Friedland, G. D., Lakomek, N.-A., Griesinger, C., Meiler, J., & Kortemme, T. (2009). A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. PLoS computational biology, 5(5), e1000393. doi:10.1371/journal.pcbi.1000393

Fuentes, E. J., Der, C. J., & Lee, A. L. (2004). Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain. Journal of Molecular Biology, 335(4), 1105-1115. doi:10.1016/j.jmb.2003.11.010

Gerstein, M., & Krebs, W. (1998). A database of macromolecular motions. Nucleic acids research, 26(18), 4280-90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12211036

Gilis, D., & Rooman, M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. Journal of molecular biology, 272(2), 276-90. doi:10.1006/jmbi.1997.1237

Gilis, D., & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein engineering, 13(12), 849-56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11239084

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Research, 35(Database issue), D291-7. doi:10.1093/nar/gkl959

Gromiha, M Michael, An, J., Kono, H., Oobatake, M., Uedaira, H., & Sarai, A. (1999). ProTherm : Thermodynamic Database for Proteins and Mutants. Nucleic acids research, 27(1), 286-288.

Gromiha, M. M., Oobatake, M., Kono, H., Uedaira, H., & Sarai, A. (2000). Importance of surrounding residues for protein stability of partially buried mutations. Dynamics, Journal of Biomolecular Structure and Dynamics, 18(2), 281-295.

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology, 320(2), 369-87. doi:10.1016/S0022-2836(02)00442-4

Gutteridge, A., & Thornton, J. M. (2005). Conformational changes observed in enzyme crystal structures upon substrate binding. Journal of molecular biology, 346(1), 21-8. doi:10.1016/j.jmb.2004.11.013

Henrick, K., & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. Trends Biochem Sci, 23(9), 358-361.

Henzler-Wildman, K. a, Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. Nature, 450(7171), 838-44. doi:10.1038/nature06410

Hilser, V. J. (2010). An ensemble view of allostery. Science (New York, N.Y.), 327(5966), 653-4. doi:10.1126/science.1186121

Hughey, R, & Krogh, a. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. Computer applications in the biosciences : CABIOS, 12(2), 95-107. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8744772

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., et al. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. BMC bioinformatics, 10, 136. doi:10.1186/1471-2105-10-136

James, L. C., & Tawfik, D. S. (2003). Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends in Biochemical Sciences, 28(7), 361-368. doi:10.1016/S0968-0004(03)00135-X

Jones, D.T., Taylor, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. Nature, 358(6381), 86-89.

Jones, S., & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol, 63(1), 31-65.

Juritz, E. I., Alberti, S. F., & Parisi, G. D. (2011). PCDB: a database of protein conformational diversity. Nucleic acids research, 39(Database issue), D475-9. doi:10.1093/nar/gkq1181

Juritz, E. I., Palopoli, N., Fornasari, M., Fernandez-Alberti, S., & Parisi, G. (2012). Protein conformational diversity modulates protein divergence. Mol Biol Evol. doi:10.1093/molbev/mss080

Kamp, M. W. V. D., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., Benson, N. C., et al. (2010). Dynameomics: A Comprehensive Database of Protein Dynamics. Structure, 18(4), 423-435. Elsevier Ltd. doi:10.1016/j.str.2010.01.012

Kantrowitz, E. R., & Lipscomb, W. N. (1990). Escherichia coli aspartate transcarbamoylase: the molecular basis for a concerted allosteric transition. Trends in Biochemical Sciences, 15(2), 53-59.

Karplus, K., Sjo, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., et al. (1997). Predicting protein structure using hidden Markov models. PROTEINS: Structure, Function, and Genetics, Suppl., 29(Supplement 1), 134-139.

Karplus, M, & Kuriyan, J. (2005). Molecular dynamics and protein function. Proceedings of the National Academy of Sciences of the United States of America, 102(19), 6679-85. doi:10.1073/pnas.0408930102

Karush, F. (1950). HETEROGENEITY OF BINDING SITES OF BOVINE SERUM ALBUMIN. Journal of the American Chemical Society, 72, 2705-2713.

Keskin, O, Jernigan, R. L., & Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophysical journal, 78(4), 2093-106. doi:10.1016/S0006-3495(00)76756-7

Khan, S., & Vihinen, M. (2010). Performance of protein stability predictors. Hum Mutat, 31(6), 675-684.

Khersonsky, O., Roodveldt, C., & Tawfik, D. S. (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. Current opinion in chemical biology, 10(5), 498-508. doi:10.1016/j.cbpa.2006.08.011

Koehl, P., & Delarue, M. (1994). Polar and nonpolar atomic environments in the protein core: implications for folding and binding. Proteins, 20(3), 264-278.

Koshland, D. E. J., Ray, W. J. J., & Erwin, M. J. (1958). Protein structure and enzyme action. Federation proceedings, 17(4), 1145-1150.

Kumar, S, Ma, B., Tsai, C. J., Sinha, N., & Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. Protein science : a publication of the Protein Society, 9(1), 10-9. doi:10.1110/ps.9.1.10

Laine, E., Chauvot de Beauchêne, I., Perahia, D., Auclair, C., & Tchertanov, L. (2011). Mutation D816V alters the internal structure and dynamics of c-KIT receptor cytoplasmic region: implications for dimerization and activation mechanisms. PLoS computational biology, 7(6), e1002068. doi:10.1371/journal.pcbi.1002068

Landsteiner, K. (1936). The Specificity of Serological Reactions. Dover Publications (reprinted 1962).

Lange, O. F., Lakomek, N.-A., Farès, C., Schröder, G. F., Walter, K. F. a, Becker, S., Meiler, J., et al. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science (New York, N.Y.), 320(5882), 1471-5. doi:10.1126/science.1157092

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. J Mol Biol, 236(3), 918-939.

Lee, C., & Levitt, M. (1991). Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature, 352(6334), 448-451.

Levinthal, C. (1968). Are there pathways for protein folding? Journal de Chimie Physique et de Physico-Chimie Biologique, 65, 44–45.

Levinthal, C. (1969). How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 22–24.

Levitt, Michael. (1978). Conformational preferences of amino acids in globular proteins. Biochemistry, 17, 4277-4285.

Lewontin, R. C. (1974). The Genetic Basis of Evolutionary Change. London: Columbia University Press. Retrieved from http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/papers/lewontin1974/lewontin_gboec.html

Lindorff-Larsen, K., Best, R. B., Depristo, M. a, Dobson, C. M., & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. Nature, 433(7022), 128-32. doi:10.1038/nature03199

Ling, S. C., Albuquerque, C. P., Han, J. S., Lagier-Tourenne, C., Tokunaga, S., Zhou, H., & Cleveland, D. W. (2010). ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. Proc Natl Acad Sci, 107(30), 13318-13323.

Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2002). SCOP database in 2002 : refinements accommodate structural genomics. Nucleic acids research, 30(1), 264-267.

Lofgren, M., & Banerjee, R. (2011). Loss of allostery and coenzyme B12 delivery by a pathogenic mutation in adenosyltransferase. Biochemistry, 50(25), 5790-5798.

Ma, B., Kumar, S., Tsai, C. J., & Nussinov, R. (1999). Folding funnels and binding mechanisms. Protein engineering, 12(9), 713-20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10506280

Ma, Buyong, Shatsky, M., Wolfson, H. J., & Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site : A matter of pre-existing populations. Protein Science, 11(2), 184-197. doi:10.1110/ps.21302.or

Ma, J. (2005). Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure, 13(3), 373-380. doi:10.1016/j.str.2005.02.002

Madera, M. (2008). Profile Comparer : a program for scoring and aligning profile hidden Markov models. Bioinformatics, 24(22), 2630-2631. doi:10.1093/nar/gkn065

Maguid, S., Fernandez-alberti, S., & Echave, J. (2008). Evolutionary conservation of protein vibrational dynamics. Gene, 422, 7-13. doi:10.1016/j.gene.2008.06.002

Maguid, S., Fernández-Alberti, S., Parisi, G., & Echave, J. (2006). Evolutionary Conservation of Protein Backbone Flexibility. Progress in Biophysics and Molecular Biology, 63(4), 448-457. doi:10.1007/s00239-005-0209-x

Masso, M., & Vaisman, I. I. (2010). AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel., 23, 683-687.

Matthews, B. W. (1996). Structural and genetic analysis of the folding and function of T4 lysozyme. FASEB J, 10, 35–41.

Michie, Alex D, Orengo, C. A., & Thornton, J. M. (1996). Analysis of Domain Structural Class Using an Automated Class Assignment Protocol. Journal of molecular biology, 262(20), 168-185.

Mirsky, A. E., & Pauling, L. (1936). On the Structure of Native, Denatured, and Coagulated Proteins. Proc. Natl. Acad. Sci. USA, 22(7), 439-447.

Miyazawa, S., & Jernigan, R. L. (1994). Protein stability for single substitution mutants and the extent of local compactness in the denatured state. Protein Eng, 7(10), 1209-1220.

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the Nature of Allosteric Transitions : A Plausible Model. Journal of molecular biology, 12(December), 88-118.

Monsellier, E., & Chiti, F. (2007). Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep., 8((8)), 737–742. doi:10.1038/sj.embor.7401034

Munoz, V., & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins, 20(4), 301-311.

Ng, P. C., & Henikoff, S. (2003). SIFT : predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814. doi:10.1093/nar/gkg509

Nienhaus, G. U., Muller, J. D., McMahon, B. H., & Frauenfelder, H. (1997). Exploring the conformational energy landscape of proteins. Physica D, 107(2-4), 297-311.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. Structure (London, England : 1993), 5(8), 1093-108. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9309224

Orosz, F., Olah, J., & Ovadi, J. (2009). Triosephosphate isomerase deficiency: new insights into an enigmatic disease. Biochim Biophys Acta, 1792(12), 1168-1174.

Ortiz, A. R., & Strauss, C. E. M. (2002). MAMMOTH ( Matching molecular models obtained from theory ): An automated method for model comparison. Protein Science, 11(11), 2606-2621. doi:10.1110/ps.0215902.next

Ostermann, A., Waschipky, R., Parak, F. G., & Nienhaus, G. U. (2000). Ligand binding and conformational motions in myoglobin. Nature, 404(6774), 205-208.

O'Brien, P. J., & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. Chem. Biol., 6, R91–R105.

Parisi, Gustavo, & Echave, J. (2004). The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. BMC evolutionary biology, 4(41). doi:10.1186/1471-2148-4-41

Parisi, Gustavo, & Echave, J. J. (2001). Structural constraints and emergence of sequence patterns in protein evolution. Molecular biology and evolution, 18(5), 750-6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11319259

Pauling, L., & Coryell, C. D. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. Proc Natl Acad Sci U.S.A., 22(4), 210-216.

Pauling, Linus. (1940). A Theory of the Structure and Process of Formation of Antibodies. Journal of the American Chemical Society, 62, 2643-2657.

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. Oxford Univ Press. Retrieved from http://nar.oxfordjournals.org/cgi/content/abstract/31/1/452

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Murihead, H., Will, G., & North, A. C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature, 185(4711), 416-422.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy : hypothesis testing using phylogenies. Bioinformatics, 21(5), 676-679. doi:10.1093/bioinformatics/bti079

Porter, C. T., Bartlett, G. J., & Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic acids research, 32(Database issue), D129-33. doi:10.1093/nar/gkh028

Prabhu, N. V., Lee, A. L., Wand, A. J., & Sharp, K. A. (2003). Dynamics and Entropy of a Calmodulin-Peptide Complex Studied by NMR and Molecular Dynamics. Biochemistry, 42(2), 562-570.

Reumers, J., Schymkowitz, J., & Rousseau, F. (2009). Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. BMC Bioinformatics, 10(8), S9.

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. Science, 273(5281), 1516-1517.

Russel, D., Lasker, K., Phillips, J., Schneidman-duhovny, D., Velazquez-Muriel, J. A., & Sali, A. (2009). The structural dynamics of macromolecular processes. Current Opinion in Cell Biology, 21, 97-108. doi:10.1016/j.ceb.2009.01.022

Sander, C., & Schneider, R. (1993). The HSSP data alignments base of protein structure-sequence. Nucleic Acids Research, 21(13), 3105-3109.

Saraboji, K., Gromiha, M. M., & Ponnuswamy, M. N. (2006). Average assignment method for predicting the stability of protein mutants. Biopolymers, 82(1), 80-92.

Schneider, A., Cannarozzi, G. M., & Gonnet, G. H. (2005). Empirical codon substitution matrix. BMC Bioinformatics, 6(134). doi:10.1186/1471-2105-6-134

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res Web Server issue, 33, W382--388.

Shah, G. N., Bonapace, G., Hu, P. Y., Strisciuglio, P., & Sly, W. S. (2004). Carbonic anhydrase II deficiency syndrome (osteopetrosis with renal tubular acidosis and brain calcification): novel mutations in CA2 identified by direct sequencing expand the opportunity for genotype-phenotype correlation. Hum Mutat, 24(3), 272.

Sinha, Neeti, & Nussinov, R. (2001). Point mutations and sequence variability in proteins : Redistributions of preexisting populations. Proc. Natl. Acad. Sci. USA, 98(6), 3139-3144.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. Curr Opin Struct Biol, 5(2), 229-235.

Smock, R. G., & Gierasch, L. M. (2009). Sending Signals Dynamically. Science, 324, 198-203.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., et al. (2003). Human Gene Mutation Database (HGMD®):2003 update. Hum Mutat, 21(6), 577-581.

Stitziel, N. O., Binkowski, T. A., Tseng, Y. Y., Kasif, S., & Liang, J. (2004). topoSNP: A topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Research, Database issue, 32, D520-D522. doi:10.1093/nar/gkh104

Suh, Y., & Vijg, J. (2005). SNP discovery in associating genetic variation with human disease phenotypes. Mutation Research, 573, 41-53. doi:10.1016/j.mrfmmm.2005.01.005

Sunyaev, S., Ramensky, V., & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends in genetics : TIG, 16(5), 198-200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10782110

Tokuriki, N, & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. Curr Opin Struct Biol, 19(5), 596-604.

Tokuriki, Nobuhiko, & Tawfik, D. S. (2009). Protein Dynamism and Evolvability. Science, 324(5924), 203-207.

Topham, C. M., Srinivasan, N., & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Engineering, 10(1), 7-21.

Torrance, J. W., Bartlett, G. J., Porter, C. T., & Thornton, J. M. (2005). Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. Journal of molecular biology, 347(3), 565-81. doi:10.1016/j.jmb.2005.01.044

Tozzini, V. (2005). Coarse-grained models for proteins. Current Opinion in Structural Biology, 15, 144-150. doi:10.1016/j.sbi.2005.02.005

Tsai, C.-jung, Kumar, S., Ma, B., & Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. Protein Science, 8, 1181-1190.

Velyvis, A., Yang, Y. R., Schachman, H. K., & Kay, L. E. (2007). A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. Proceedings of the National Academy of Sciences, 104(21), 8815–8820.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J Mol Graph., 8(1), 52-56.

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. Human mutation, 17(4), 263-270. doi:10.1002/humu.22

Whitney, J. (1997). Testing for differences with the nonparametric Mann-Whitney U test. J Wound Ostomy Continence Nurs, 24(1).

Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z., & Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nature structural & molecular biology, 11(10), 945-9. doi:10.1038/nsmb821

Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci U S A., 94(12), 6170-6175.

Yogurtcu, O. N., Erdemli, S. B., Nussinov, R., Turkay, M., & Keskin, O. (2008). Restricted Mobility of Conserved Residues in Protein-Protein Interfaces in Molecular Simulations. Biophysical Journal, 94(May), 3475–3485. doi:10.1529/biophysj.107.114835

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. Journal of molecular biology, 353(2), 459-73. doi:10.1016/j.jmb.2005.08.020

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research, 31(13), 3370-3374. doi:10.1093/nar/gkg571

Zhang, Yang, & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research, 33(7), 2302-9. doi:10.1093/nar/gki524

Zhou, Hongyi, & Zhou, Y. (2002). Distance-scaled , finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science, 11, 2714-2726. doi:10.1110/ps.0217002.2002

Zoete, V., Michielin, O., & Karplus, M. (2002). Relation between Sequence and Structure of HIV-1 Protease Inhibitor Complexes: A Model System for the Analysis of Protein Flexibility. Journal of molecular biology, 315(1), 21-52. doi:10.1006/jmbi.2001.5173

Zwanzig, R., Szabo, A., & Bagchi, B. (1992). Levinthal's paradox. Proc Natl Acad Sci USA, 89(1), 20–22. doi:doi:10.1073/pnas.89.1.20