

## Capítulo 1

# Software Toolkit para análisis de trayectorias de plegamiento de proteínas

Este capítulo describe un toolkit que fue desarrollado para evaluar 16 propiedades estructurales y energéticas de proteínas seleccionadas de la literatura y otras de desarrollo propio. Las propiedades serán usadas en los próximos capítulos para lograr una representación de una proteína en términos de *características de plegamiento*. Empezamos con la presentación de una descripción de las propiedades mostrando su definición y su importancia asociada con características de plegamiento; y luego presentamos el toolkit que desarrollamos con los algoritmos y herramientas usadas para calcular sus valores.

### 1.1. Propiedades de proteínas

Las propiedades descritas en la tabla 1.1 han sido usadas en otros estudios para caracterizar diferentes aspectos del proceso de plegamiento de proteínas, y han sido asociadas con algunas características de plegamiento no observables directamente a partir de la estructura de la proteína.

## Capítulo 1. Software Toolkit para análisis de trayectorias de plegamiento de proteínas

Nombre	Id	Descripción	Asociación
Contactos Nativos	NC	Contactos formados por una conformación de proteína que también están presentes en la estructura nativa.	Grado de cercanía con la estructura nativa. Cantidad de estructura nativa.
Orden de Contactos	CO	Separación promedio de secuencias entre residuos en contacto en la estructura nativa.	Propiedades cinéticas del plegamiento de proteínas.
Radio de Giración	RG	Cuánto se esparce la proteína de su centro.	Nivel de compactación de la estructura.
Enlaces de Hidrógeno	HB	Número de enlaces de hidrógeno formados por la estructura.	Formación y estabilización.
Área de Superficie Accesible.	AS	Superficie accesible para el solvente.	Compactación y estabilidad conformacional.
Vacios	VD	Espacios sin llenar dentro de una proteína que no son accesibles desde el solvente.	Empaquetado y estabilidad de la proteína.
RMSD	RM	Distancia promedio en posiciones entre los átomos de dos proteínas.	Similitud de las proteínas.
RMSD Local	LR	Promedio del RMSD entre los elementos de la estructura secundaria (desarrollado por Garreta).	Topología de la proteína y similitud estructural.
Energía Potencial	PE	Energía asociada con las fuerzas de atracción y repulsión entre los átomos de la proteína.	Estabilidad termodinámica.
Momento Dipolar	DM	El producto de la magnitud de las cargas y su distancia de separación en una proteína.	Forma de la proteína. Propiedades eléctricas.
Residuos en Estructuras Secundarias Correctas	RC	Residuos formando la misma estructura secundaria que la conformación nativa (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Residuos en cualquier estructura secundaria	RA	Residuos en la proteína objetivo formando cualquier elemento de estructura secundaria (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Puntaje Estructural	SS	Residuos que forman “correctamente” motivos locales cortos o bloques de proteínas (desarrollado por Garreta).	Similitud de estructuras entre proteínas.
Clusters Rígidos	CL	Grupos de átomos acoplados uno al otro a través de enlaces rígidos.	Estabilidad mecánica de una proteína.
Grados de libertad	DF	Relacionada con el número de restricciones necesarias para hacer rígida la estructura de la proteína.	Estabilidad mecánica de una proteína.
Regiones estresadas	SR	Regiones con sobrecarga de restricciones que reducen el grado de libertad de una proteína.	Estabilidad mecánica de una proteína.

**Cuadro 1.1: Resumen de las propiedades seleccionadas para describir el plegamiento.** La tabla muestra el nombre de la propiedad, el id que será usado como nombre corto a lo largo de este documento, una descripción corta y su asociación con el plegamiento de proteínas.

### 1.1.1. Contactos Nativos

El número de contactos nativos de una conformación proteica puede ser considerado como una medida global de la "cantidad" de estructura que contiene respecto a su estructura nativa. [?]. ...Dos residuos de proteína distintos forman un contacto cuando la distancia máxima entre residuos de un par de átomos pesados (C, O, S, or N) está dentro de un valor umbral (ej. 6 o 7 Å). Otras restricciones pueden ser una distancia euclidiana mínima entre los residuos (e.g. 3 Å); una separación secuencial de al menos 3 residuos entre sí; o una distancia entre átomos pesados específicos como C- $\alpha$  or C- $\beta$ .

El número de contactos nativos se ha utilizado para evaluar la "semejanza nativa" de las conformaciones (cuánta estructura tiene una conformación) en varios estudios relacionados con el plegamiento de proteínas como la construcción de rutas utilizando métodos probabilísticos de hoja de ruta.[?, ?]; el estudio de las primeras fases del plegado mediante simulaciones de dinámica molecular [?]; y en la construcción del camino usando eventos de desplegado [?].

Se calcula contando los contactos de la conformación que también están en la estructura nativa. Como se ha definido anteriormente, dos átomos de dos aminoácidos diferentes  $a_i$  y  $a_j$  están en contacto, si su distancia entre residuos  $\delta(a_i, b_j) \leq \delta^{max}$ , donde

$$\delta(a_i, b_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

es la distancia euclidiana entre dos átomos(ej. carbono C- $\alpha$  o C- $\beta$ ) de los dos residuos diferentes con coordenadas 3D  $x_i, y_i, z_i$  y  $x_j, y_j, z_j$ ; y  $\delta^{max}$  es algún umbral de distancia máxima permitida (ej. 6 o 7 Å) [?].

### 1.1.2. Orden de Contacto

La complejidad topológica de una proteína está relacionada con el tipo de interacciones, locales o no locales, que impulsan su plegamiento.[?]. El orden de contacto (CO) se define como una medida del número relativo de interacciones locales en comparación con las interacciones no locales [?]. Las estructuras dominadas por interacciones locales como las proteínas altamente  $\alpha$  helicoidales corresponden a bajo CO y se pliegan más rápidamente que las estructuras dominadas por interacciones de largo alcance -necesarias para su estabilidad—como  $\alpha/\beta$  y  $\beta$  proteínas con CO más alto.

Esta medida ayuda a cuantificar la topología y la estabilidad de las proteínas en sus estados nativos.[?]. Se ha correlacionado bien con las propiedades

de la cinética de plegado de proteínas como las tasas de plegado y la colocación del estado de transición de proteínas. [?]. Además, se ha utilizado en programas de estructura de proteínas *ab initio* para ayudar a seleccionar los mejores candidatos para una etapa de refinamiento posterior. [?]. En proteínas con cinética simple -plegamiento en dos estados- ha descrito el tipo de interacciones proteicas [?, ?].

Se calcula a partir de las coordenadas de la estructura de una proteína de la siguiente manera[?]:

$$CO = \frac{1}{LN} \sum_{i,j}^N \Delta S_{i,j} |i - j|$$

Donde  $L$  es el número total de residuos en la proteína,  $N$  es el número total de contactos, y  $\Delta S_{i,j}$  es la separación secuencial de residuos entre  $i$  y  $j$ .

### 1.1.3. Radio de giro

El radio de giro (RG) es una medida de la forma o dimensiones de la conformación de una proteína que describe su tamaño y compacidad midiendo cuánto se extiende la estructura de la proteína desde su centro. En el análisis de proteínas, la RG ayuda a inferir cuán plegada o desplegada está una proteína, ya que es indicativa del nivel de compactación de la estructura[?]. Ayuda a trazar los cambios de conformación de una proteína durante su trayectoria de plegado trazando su RG en función del tiempo. [?]. Además, se ha demostrado que las arquitecturas de las proteínas ( $\alpha, \beta, \alpha/\beta$ , y  $\alpha + \beta$ ) tienen un RG característico [?].

El radio de giro de una conformación se calcula como la distancia entre su centro de gravedad y sus extremos.[?]:

$$R_g = \sqrt{\left( \frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i} \right)}$$

donde  $m_i$  es la masa del átomo  $i$  y  $r_i$  la posición del átomo  $i$  con respecto al centro de masa de la molécula.

### 1.1.4. Enlaces de Hidrógeno

Los enlaces de hidrógeno (HB) juegan un papel central en el plegamiento de proteínas, influyendo en la forma final en 3D de una proteína al contribuir en la formación de sus estructuras secundarias y terciarias. Un enlace de

hidrógeno es un tipo de fuerza atractiva que existe entre un hidrógeno unido a un átomo electronegativo de una molécula y un átomo electronegativo de la misma molécula o de una molécula diferente. Los enlaces de hidrógeno participan tanto en la estabilización de hélices  $\alpha$  como en la interacción entre hebras  $\beta$  para formar hojas  $\beta$ , así que son parcialmente responsables de la estabilidad de la estructura final de la proteína plegada [?]. Varios estudios han utilizado el HB para evaluar la estabilidad [?] y para indicar cierta organización en proteínas [?]. Generalmente, este tipo de enlace es más fuerte que otras fuerzas intermoleculares como el van der Waals pero más débil que los enlaces iónicos y covalentes. [?].

Para determinar el número de enlaces de hidrógeno formados en una estructura proteica, se cuentan los pares de átomos -donantes y aceptantes- que están dentro de un valor umbral [?]. El umbral puede variar de 2.6 a 3.6 Å, que es la distancia entre átomos pesados. [?].

#### 1.1.5. Área de superficie Accesible

La superficie accesible (ASA) de una proteína contribuye a determinar su tamaño y forma. El ASA es un concepto que corresponde a la superficie de una proteína accesible al solvente. [?]. Desempeña un papel importante en varias propiedades de las proteínas como la compactidad, el plegamiento de proteínas y la estabilidad conformacional. [?]. El ASA se ha utilizado para caracterizar la compactidad de una estructura proteica definiendo una relación entre su ASA y la superficie de la esfera ideal del mismo volumen. [?].

La idea general para calcular el ASA es utilizar una "sonda", correspondiente a una esfera del disolvente, que se hace rodar sobre la superficie de la proteína explorando el área accesible por la sonda. [?]. El radio de la sonda corresponde al radio de una molécula del disolvente, por ejemplo 1.4 Å para una molécula de agua.

#### 1.1.6. Vacíos

Los vacíos son espacios vacíos o cavidades vacías, dentro de una proteína, que no son accesibles al disolvente. Permanecen vacías aunque pueden ser lo suficientemente grandes como para ser llenadas por otros átomos o moléculas como el agua. Se han relacionado con la estabilidad de las proteínas, ya que al rellenar las cavidades -en la ingeniería de proteínas- en algunos tetrámeros se puede estabilizar su estructura en general [?]. Además, los vacíos están relacionados con la evaluación de la calidad de empaquetamiento de las pro-

teínas. [?]. Menos cavidades de vacíos dentro de una proteína implican que está más empaquetada, y las altas densidades de empaquetado en los núcleos de proteínas a menudo se consideran más como sólidos que como líquidos. [?].

Los vacíos se calculan utilizando un método de búsqueda de vacíos basado en cuadrículas que mapea una proteína en una cuadrícula 3D y asigna cada punto de cuadrícula a la proteína, al solvente o a la cavidad de acuerdo con su ubicación en la estructura de la proteína.[?]. Dos tipos de sondas con un radio pequeño (ej. 0.0 Å) y un radio grande (ej. 1.4 Å) se usan para detectar vacíos y delimitar las regiones accesibles con solventes, respectivamente. La sonda más grande delimita las regiones accesibles al disolvente, mientras que la más pequeña identifica los vacíos sin fusionarlos con el disolvente en bruto.

### 1.1.7. Error cuadrático Medio

El error cuadrático medio(*Root-Mean Square Deviation* en inglés, y más conocido como RMSD)refleja la distancia media en posiciones entre los átomos de dos proteínas superpuestas. Generalmente, estructuras similares tienen valores bajos de RMSD, en el orden de 1-3 Å, mientras que los valores RMSD más grandes corresponden a valores más disímiles [?].

El RMSD ha sido ampliamente utilizado como una medida del grado de similitud entre dos estructuras de proteínas. El experimento CASP ha utilizado este método -además de otras medidas- para evaluar los protocolos de predicción utilizados por los expertos para predecir la estructura 3D entre proteínas.[?].

Se calcula a partir de las coordenadas del objetivo  $i$  and una referencia  $j$  como sigue:

$$RMSD(i, j) = \sqrt{\frac{1}{N} \sum_{k=1}^N \left| r_k^{(i)} - r_k^{(j)} \right|^2}$$

donde  $i$  y  $j$  son dos estructuras de proteínas óptimamente superpuestas de igual tamaño .  $N$  es el número de átomos en cada estructura,  $k$  es un índice sobre estos átomos, y  $r_k^{(i)}, r_k^{(j)}$  son las coordenadas cartesianas del átomo  $k$  en conformaciones  $i, j$ . El valor de RMSD resulta mínimo porque está determinado por la superposición de las estructuras que minimiza su distancia. En nuestro proyecto, el RMSD se utiliza siempre para comparar una determinada conformación proteica con la nativa.

### 1.1.8. Error cuadrático Medio Local (Local RMSD)

El citado RMSD es una medida global incapaz de evaluar las similitudes en las estructuras locales, y considerada como una "medida imperfecta" para comparar estructuras como conformaciones intermedias donde los elementos estructurales pueden estar presentes pero no todavía organizados espacialmente.[?, ?].

En su proyecto, Garreta desarrolló el *local root-mean square deviation* (LRMSD). Compara una conformación de referencia  $i$  de una proteína con una conformación objetivo  $j$ , utilizando el RMSD sólo en cada elemento de estructura secundaria (EES) de la estructura de referencia y promediando los valores obtenidos. Se relaciona con la topología de una proteína, y contribuye a medir la similitud estructural entre dos conformaciones de proteínas comparando localmente sus elementos estructurales secundarios.

Se calcula extrayendo las posiciones de los aminoácidos de los SSEs desde la primera conformación (la referencia), usando el programa DSSP[?]. Las mismas posiciones se extraen de la segunda conformación (objetivo). Se instala cada par de EES y se calcula su valor RMSD, finalmente se determina el promedio de todas las SSEs en la estructura de referencia:

$$LRMSD(i, j) = \frac{1}{S} \sum_{s=1}^S RMSD(i_s, j_s)$$

donde  $i$ , es la conformación de referencia y  $j$  la conformación objetivo.  $S$  es el número de EESs en la estructura de referencia.  $RMSD(i_s, j_s)$  es el valor RMSD entre el número EES  $s$  en la conformación de referencia  $i$  y los átomos correspondientes en la conformación del objetivo.

En este trabajo, la conformación de referencia es siempre la nativa.

### 1.1.9. La energía potencial

Si una proteína se toma como una molécula estática con muchos átomos interactuando, su energía potencial puede ser definida como la energía dada por su estructura tridimensional.

La energía potencial está relacionada con la estabilidad termodinámica, considerando que una proteína se pliega al estado de menor energía al disminuir su energía potencial, y alcanza la estabilidad termodinámica en su estado nativo.[?].

Si las interacciones entre los átomos de la proteína son consideradas como fuerzas elásticas, entonces pueden ser escritas en términos de ecuaciones conocidas como funciones de energía potencial (o campos de fuerza) en las

que la energía total es dada por la suma de términos enlazados y no enlazados que describen átomos enlazados por enlaces covalentes e interacciones de largo alcance, respectivamente. A continuación se presenta una función general de energía potencial; para funciones más específicas, véase[?]:

$$E_{total} = E_{enlaces} + E_{ángulos} + E_{torsiones} + E_{van-der-waals} + E_{electrostatica}$$

donde las primeras tres energías corresponden a términos enlazados que describen enlaces, ángulos y rotaciones de enlaces en la proteína. Y, los dos últimos términos describen interacciones entre átomos no ligados o átomos separados por 3 o más enlaces covalentes, la energía de interacción de van der Waals y la energía electrostática.

#### 1.1.10. El momento dipolar

El momento dipolar puede definirse como el producto de la magnitud de la carga y la distancia de separación entre las cargas de una molécula.

El momento dipolar está relacionado con las estructuras geométricas y eléctricas de una molécula; y su magnitud puede afectar la forma de una proteína, su interacción intermolecular, la solución biomolecular y sus actividades bioquímicas.[?]. Su magnitud también puede utilizarse para distinguir entre moléculas polares y no polares. En las proteínas, el momento dipolar se calcula como la suma vectorial del momento dipolar de tres componentes: cargas fijas en la superficie, grupos polares en las cadenas principales y laterales y las fluctuaciones móviles de iones. [?].

#### 1.1.11. Residuos en estructuras secundarias correctas o en cualquiera

Hélices alfa, hebras beta formando hojas beta, y otros elementos de estructura secundaria (EESs) son contribuyentes esenciales para la estabilización del pliegue proteínico general. Estas estructuras son segmentos de la cadena de proteínas en los que los ángulos de torsión del backbone  $\phi$  y  $\psi$  se repiten en patrones regulares. Hemos creado dos medidas para cuantificar la proporción de EESs en una conformación de proteína. La primera, *residuos en estructura secundaria correcta*, evalúa la proporción de residuos de una conformación objetivo que coinciden con en su EES con el de la nativa. La segunda, *residuos en cualquier estructura secundaria* evalúa la proporción de residuos formando cualquier EESs en la conformación objetivo.

Ambas medidas se calculan obteniendo y comparando los EESs asignados por el programa DSSP[?] para las conformaciones de la proteína: objetivo



y nativo para la medida correcta de SSEs, y sólo objetivo para cualquier medida de EESs.

#### 1.1.12. Puntaje estructural

Esta medida fue desarrollada por Garreta y está relacionada con una tesis doctoral de nuestro grupo de investigación. Se basa en la similitud estructural entre proteínas, pero las comparaciones se realizan utilizando motivos locales cortos o bloques de proteínas (BP) [?] en lugar de EESs. La representación de las proteínas con BP proporciona más información estructural local que los EESs porque los BP resaltan las variaciones sutiles y las conservaciones estructurales de la estructura. [?, ?]. La estructura 3D de la proteína fue representada como una cadena unidimensional de BPs a partir de una biblioteca de 40 que fueron construidos utilizando fragmentos de 5 residuos que consisten en las coordenadas 3D de sus átomos C- $\alpha$ . Para su cálculo, se codifican ambas estructuras de proteínas: objetivo y nativa, transformando su representación 3D en una secuencia 1D de BPs. Luego, la proporción de BPs del objetivo que se calcula que también están en el nativo.

#### 1.1.13. Clusters rígidos, regiones estresadas y grados de libertad

Estas medidas se toman del análisis de rigidez que se ocupa de analizar regiones rígidas y flexibles de una proteína. [?, ?]. *Clusters rígidos* son las partes de la proteína que se comportan como cuerpos rígidos (es decir, las distancias entre todos los átomos permanecen fijas). [?]. *Grados de libertad* está relacionado con el número de restricciones necesarias para rigidizar la estructura de la proteína [?]. Y *regiones estresadas* son regiones que contienen restricciones excesivas [?], como restricciones de distancia entre los átomos impuestas por las fuerzas de enlace, que reducen el número total de grados de libertad disponibles para la proteína. [?]. La rigidez puede estar relacionada con la estabilidad mecánica de una proteína [?], que se explica como la resistencia de la proteína a desplegarse en respuesta a una fuerza externa [?].

Estas medidas se calculan utilizando el software FIRST que implementa el algoritmo del juego de los guijarros [?]. Éste genera un gráfico dirigido para modelar las fuerzas físicas de las restricciones debidas a la unión presente dentro de la proteína. Luego FIRST comprueba cada enlace para determinar si forma parte de dos tipos de elementos de rigidez: un clúster rígido o una región sometida a tensión que forma una conexión flexible.

## 1.2. Toolkit para cálculo de métricas

Se trata de un conjunto de herramientas de software para el análisis de proteínas que ofrece un conjunto de funciones listas para usar para calcular las propiedades que se describen a continuación. Además, como estos cálculos se ejecutarán en rutas de plegado completas o trayectorias con cientos o miles de conformaciones de proteínas, por lo que usamos un marco de trabajo distribuido propio para evaluarlas de forma distribuida en ordenadores de escritorio sencillos.

### 1.2.1. Funciones

Las principales funciones del conjunto de herramientas se presentaron en la tabla 1.1. Aquí, presentamos en la Tabla 1.2.2 las propiedades con el nombre de la función implementada y un resumen de las diferentes herramientas y parámetros utilizados para calcular cada propiedad. El kit de herramientas ofrece un conjunto de funciones a través de programas y bibliotecas independientes de python que pueden utilizarse directamente desde una interfaz de línea de comandos o llamarse desde programas externos, respectivamente.

### 1.2.2. Implementación

El kit de herramientas implementa diferentes algoritmos y sirve como *wrapper* de varias funcionalidades incluidas en otras herramientas de software para la manipulación de proteínas desarrolladas en los últimos años como programas, librerías, paquetes y frameworks. La figura 1.2.2 presenta una vista de las funciones implementadas en el kit de herramientas. En la capa externa están nuestras funciones utilizadas directamente para evaluar las propiedades de una estructura proteica. En el medio hay una capa de programas externos como bibliotecas y paquetes que están 'envueltos' por nuestras funciones para implementar su funcionalidad. Y en la capa interna están los diferentes lenguajes en los que se implementan algunas de nuestras funciones o programas externos.

## 1.3. Instalación y despliegue

El sistema consiste en un único archivo ('vm-eval.ova') que corresponde a un dispositivo virtual (también conocido como máquina virtual). Es un completo paquete de software listo para usar (sistema operativo con aplicaciones) que no necesita configuración o instalación especial. El usuario

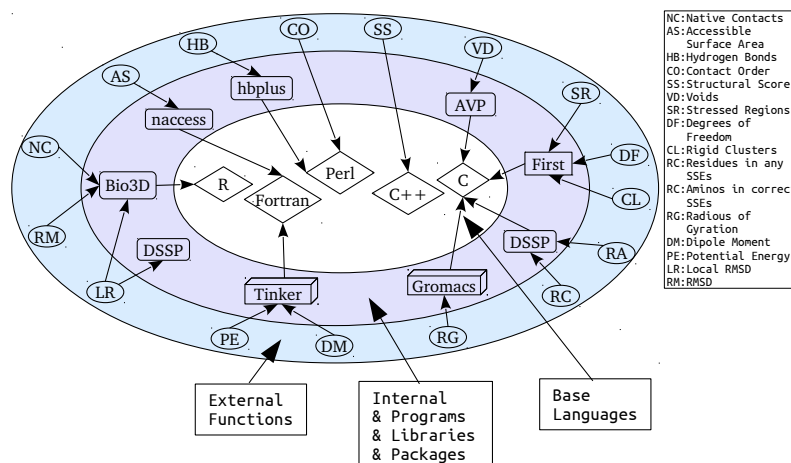
### 1.3. Instalación y despliegue

Función	ID	Descripción
native_contacts	NC	El NC utiliza mapas de contacto creados con la función <code>cmap</code> del paquete <code>bio3d</code> del sistema R [?]. Define un umbral de contacto de 7 Å y un umbral de vecindad de 3 átomos.
contact_order	OC	OC uses the code developed by Baker and co-workers [?] con un umbral de 6 Å.
radius_gyration	RG	Utiliza el <code>g_gyrate</code> del paquete <code>Gromacs</code> para la simulación de dinámica molecular [?].
hydrogen_bonds	HB	Usa el programa <i>hbplus</i> [?], identificando enlaces de hidrógeno con una distancia máxima de 2.6 y 3.9 Å para el donante-aceptor y el receptor de hidrógeno, respectivamente.
asa	AS	Usa el programa <i>naccess</i> [?], con una sonda de 1.4 Å para calcular el área total polar y no polar.
voids	VD	Usa el programa <i>AVP</i> [?] para calcular el volumen de los vacíos.
rmsd	RM	Usa la función <i>rmsd</i> del paquete <i>bio3d</i> .
local_rmsd	LR	Usa el programa <i>DSSP</i> [?] para obtener los elementos de estructura secundaria de las proteínas.
potential_energy	PE	Usa el programa <i>analyse</i> del paquete <i>tinker</i> para simulación de dinámica molecular [?] con un campo de fuerza <i>amber96</i> .
dipole_moment	DM	También usa el programa <i>analyse</i> del paquete <i>tinker</i> para simulación de dinámica molecular.
correct_sse	RC	Usa <i>DSSP</i> para obtener los elementos de estructura secundaria asignados para cada residuo.
any_sse	RA	Como RC, usa <i>DSSP</i> para obtener los elementos de estructura secundaria asignados para cada residuo.
structural_score	SS	Codifica la estructura objetivo y la nativa como una secuencia 1D de bloques de proteína usando una versión modificada del método PB-ALIGN [?].
rigid_clusters	CL	Usa la herramienta <i>FIRST</i> para análisis de rigidez [?].
degrees_freedom	DF	Usa la herramienta <i>FIRST</i> para análisis de rigidez [?].
stressed_regions	SR	Usa la herramienta <i>FIRST</i> para análisis de rigidez [?].

Cuadro 1.2: Principales funciones del toolkit de análisis de proteínas.

sólo necesita descargar el dispositivo (<http://bioinformatica.univalle.edu.co/software/vm-eval>) e importarlo a un virtualizador preinstalado (un software de virtualización como VirtualBox, VMWare, Xen).

En este proyecto hemos utilizado el virtualizador VirtualBox(<http://www.virtualbox.com>) para construir, importar y ejecutar el artefacto. Se



#### Estructura

del Toolkit para análisis de proteínas. Tres capas. En la primera capa, las funciones para evaluar las propiedades de las proteínas. En la segunda los programas externos 'envueltos' por nuestras funciones. En la tercera, los lenguajes de computación usados por las funciones y programas externos.

preconfiguró con requisitos básicos de hardware que los usuarios pueden cambiar de acuerdo con la capacidad del software del virtualizador y de la máquina, que en VirtualBox son la cantidad de RAM asignada al dispositivo, el número de núcleos de CPU virtual que debe ver el dispositivo y el porcentaje de uso de la CPU (por defecto es del 100 %).

Actualmente, el dispositivo se ejecuta en un sistema operativo Linux reducido (servidor Ubuntu 12.04) con una interfaz de línea de comandos. El usuario predefinido que ejecuta la aplicación es el usuario *fm* (*/home/fm*) en cuyo directorio hay dos subdirectorios, uno para ejecutar el procesamiento distribuido (*/home/fm/framework*) y otro para la aplicación del usuario (*/home/fm/application*). El repositorio de nube creado para el servicio de almacenamiento en nube es el */home/fm/Dropbox*). Otros directorios son el repositorio de los resultados */home/fm/results*, y un sistema de archivos RAM para una lectura y escritura rápida de archivos temporales */home/fm/ramdisk*.

#### 1.3.1. Resultados

El *framework* se utilizó para evaluar distributivamente todas las vías y trayectorias, y para realizar los análisis necesarios para culminar el presente trabajo. El framework fue operado en los PCs de escritorio del Laboratorio de Bioinformática en Universidad del Valle <http://eisc.univalle.edu.co/index.php/grupos-investigacion/bioinformatica>. Específicamente, realizamos la evaluación de 16 propiedades de las proteínas en tres conjuntos de datos diferentes: (1) en las 16000 conformaciones de una trayectoria de plegado de proteínas del subdominio de la villin headpiece (HP-35NleNle) from the folding@home project generated by molecular dynamics (see chapter 5); (2) las 37 vías obtenidas del aprendizaje basado en PRM tomado del Servidor Parasol, que comprende alrededor de 6000 conformaciones (ver capítulo 6); y (3) en las conformaciones 630000 de una trayectoria de plegado de proteínas del subdominio de villina. (HP-35NleNle) obtenidas por el grupo de investigación de David E. Shaw [?] usando dinámica molecular.

El framework funcionó sin problemas; y los tiempos de ejecución disminuyeron considerablemente en comparación con nuestros intentos antes de desarrollar el framework.