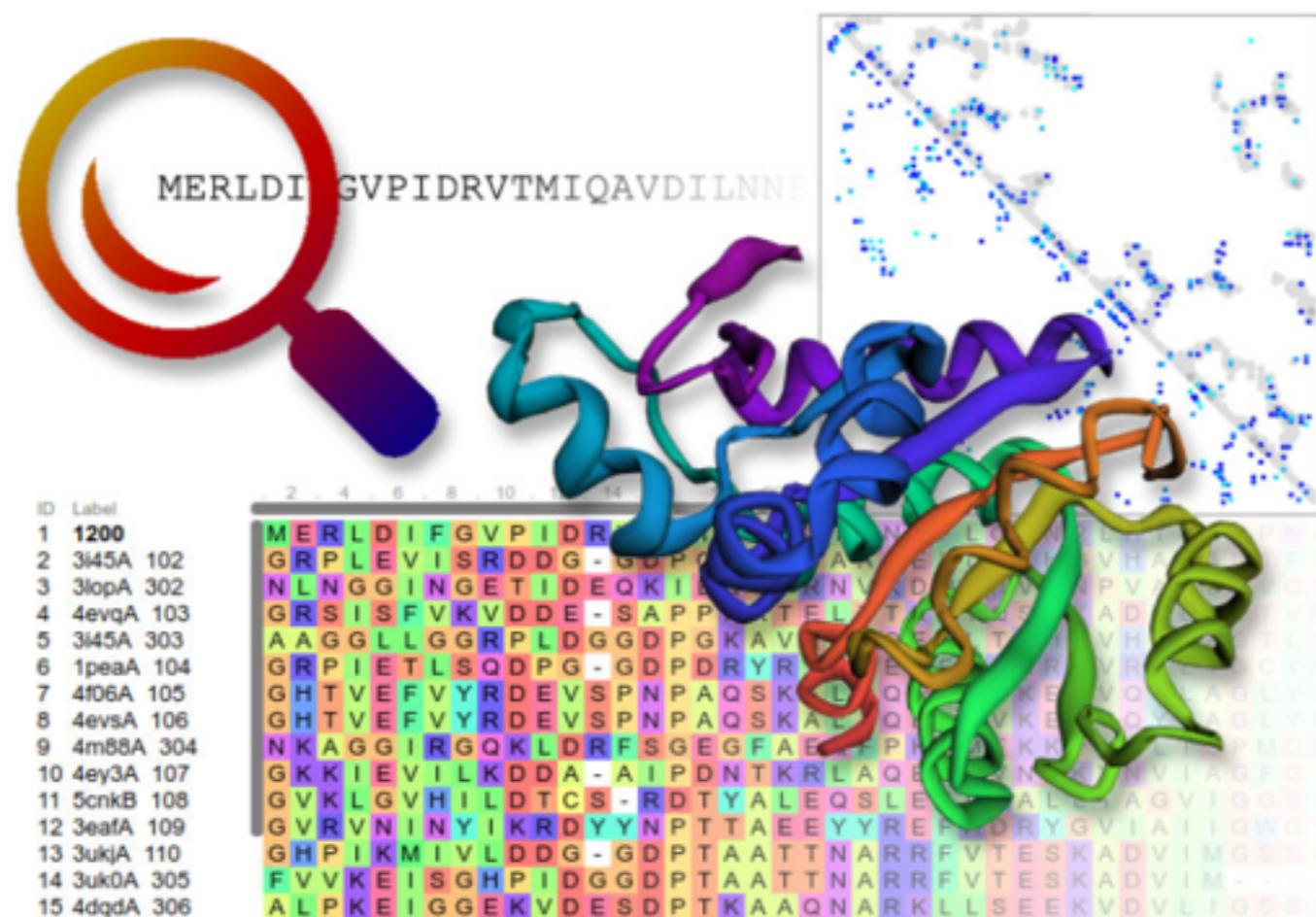


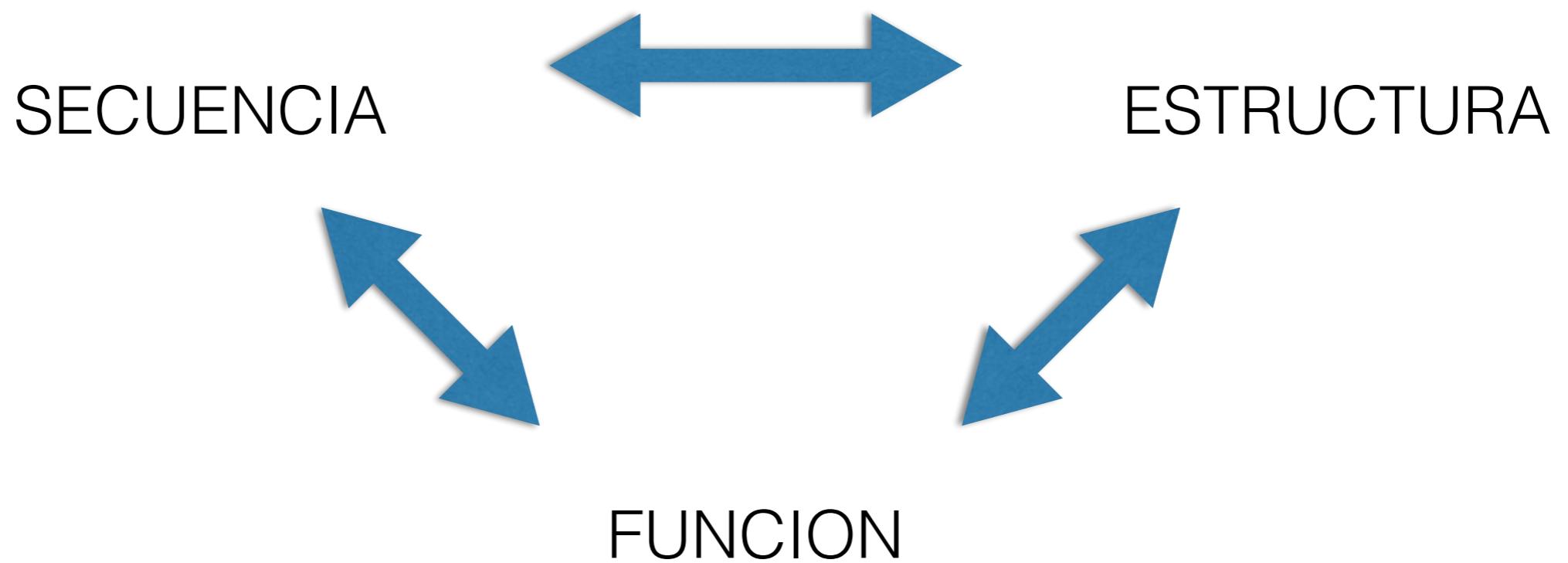
INTRODUCCION A PREDICCION DE ESTRUCTURA DE PROTEINAS: MODELADO POR HOMOLOGIA

Prof. Lucía Chemes

Instituto de Investigaciones Biotecnológicas
Universidad Nacional de San Martín



PROTEINAS



Las secuencias y las estructuras de las proteínas evolucionan y la información contenida en ambas nos dice mucho acerca su función

PROTEINAS

actualmente contamos con un gran número de secuencias y estructuras, que nos permiten elucidar regiones funcionales en las proteínas



Bases de datos de secuencias

~80.000.000



(500.000 curadas)



Bases de datos de estructuras

~160.000

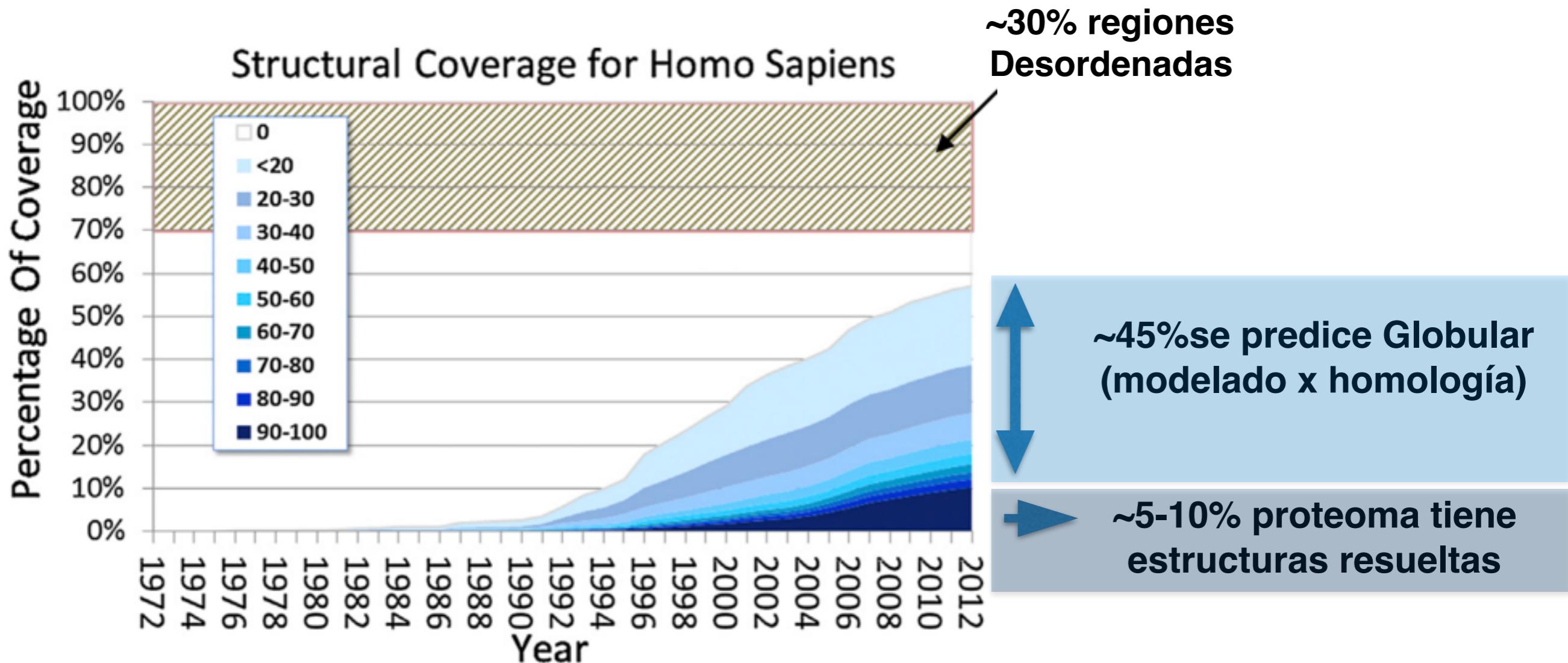


Bases de datos de regiones desordenadas

~3.000

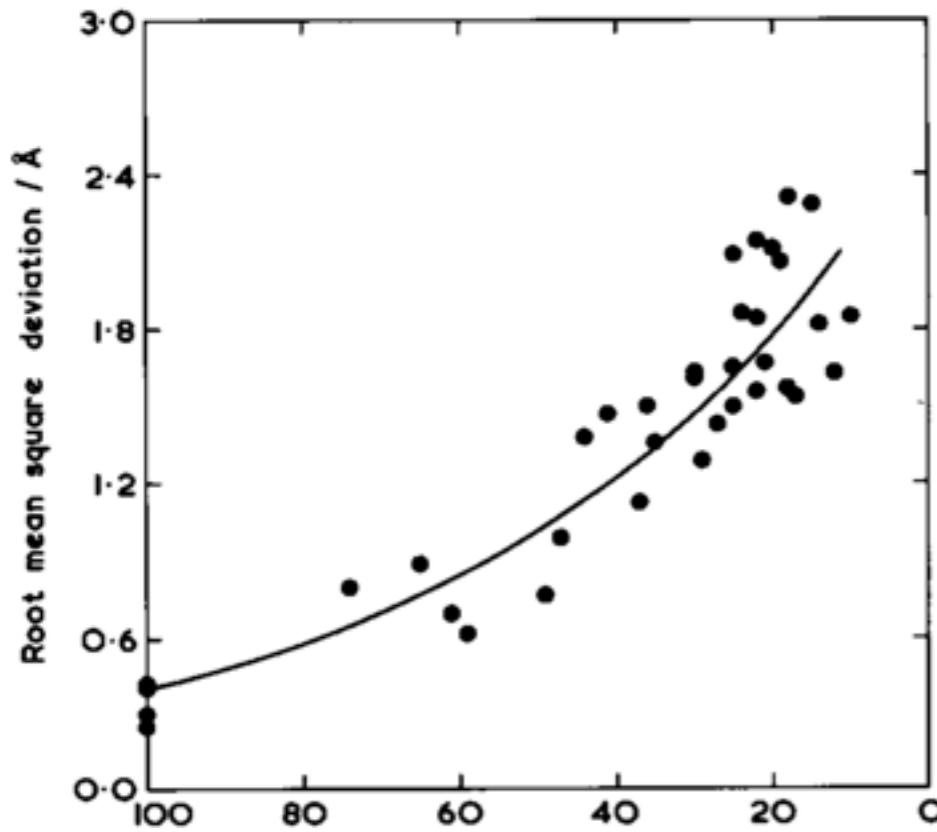


Qué fracción del proteoma humano corresponde a regiones estructuradas versus desordenadas?



En azul se muestra la fracción de aminoácidos que pueden asignarse a estructuras conocidas de la PDB dependiendo del %identidad de secuencia

En qué se basa el éxito del modelado por homología?



A NIVEL TEORICO

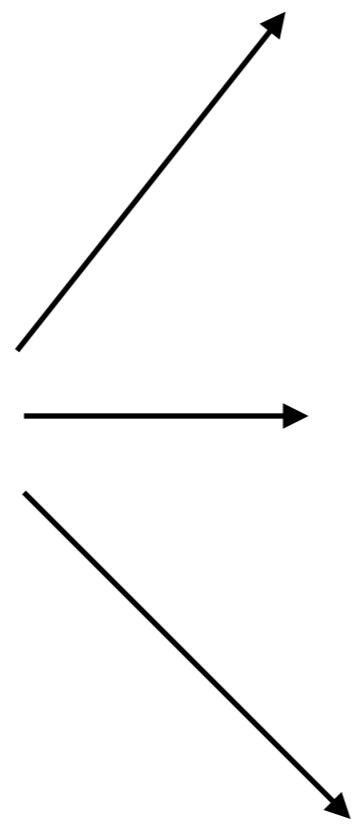
El análisis de la relación entre las secuencias y estructuras de proteínas revela que la estructura se encuentra altamente conservada aún a alta divergencia de secuencia

Chothia & Lesk (1986) *EMBOJ* 5:823-6

A NIVEL ALGORITMICO

- Desarrollo de **herramientas estadísticas para detectar homólogos remotos** extrayendo información evolutiva de alineamientos múltiples (**Profiles - HMMs**)
- Explosión de la secuenciación de genomas (alto número de secuencias a partir de las cuales construir HMMs)
- Desarrollo de **herramientas que permiten el modelado** de una secuencia **query** a partir de su alineamiento a una secuencia **target** de estructura conocida

Posibles métodos

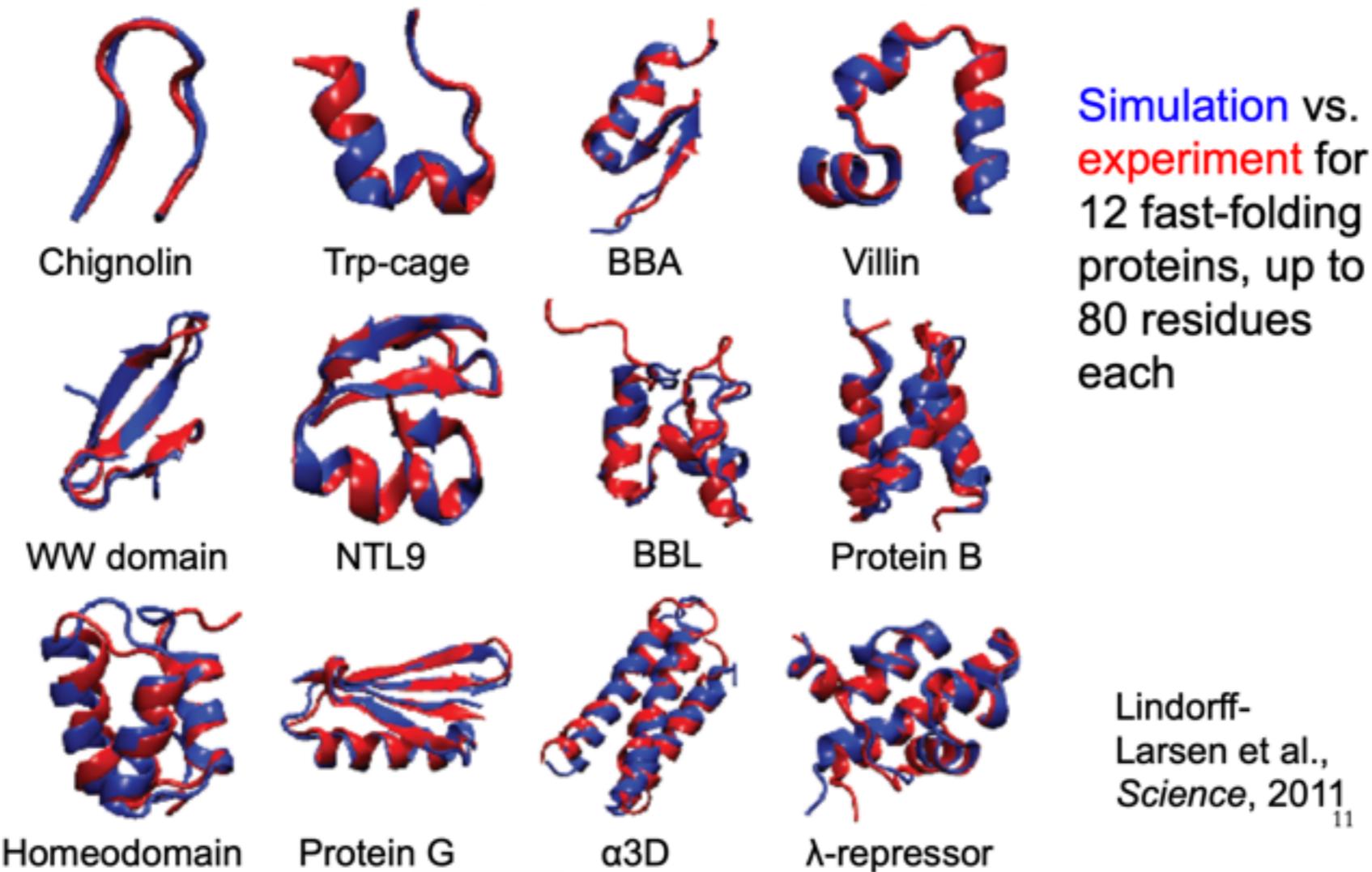


Simulación basada en principios físicos
(dinámica molecular)

Modelado de novo usando
bibliotecas de fragmentos
(Rosetta)

**Modelado por homología
(HHPred/Modeller, I-TASSER)**

Podemos obtener el modelo simulando el plegado de la cadena por dinámica molecular?

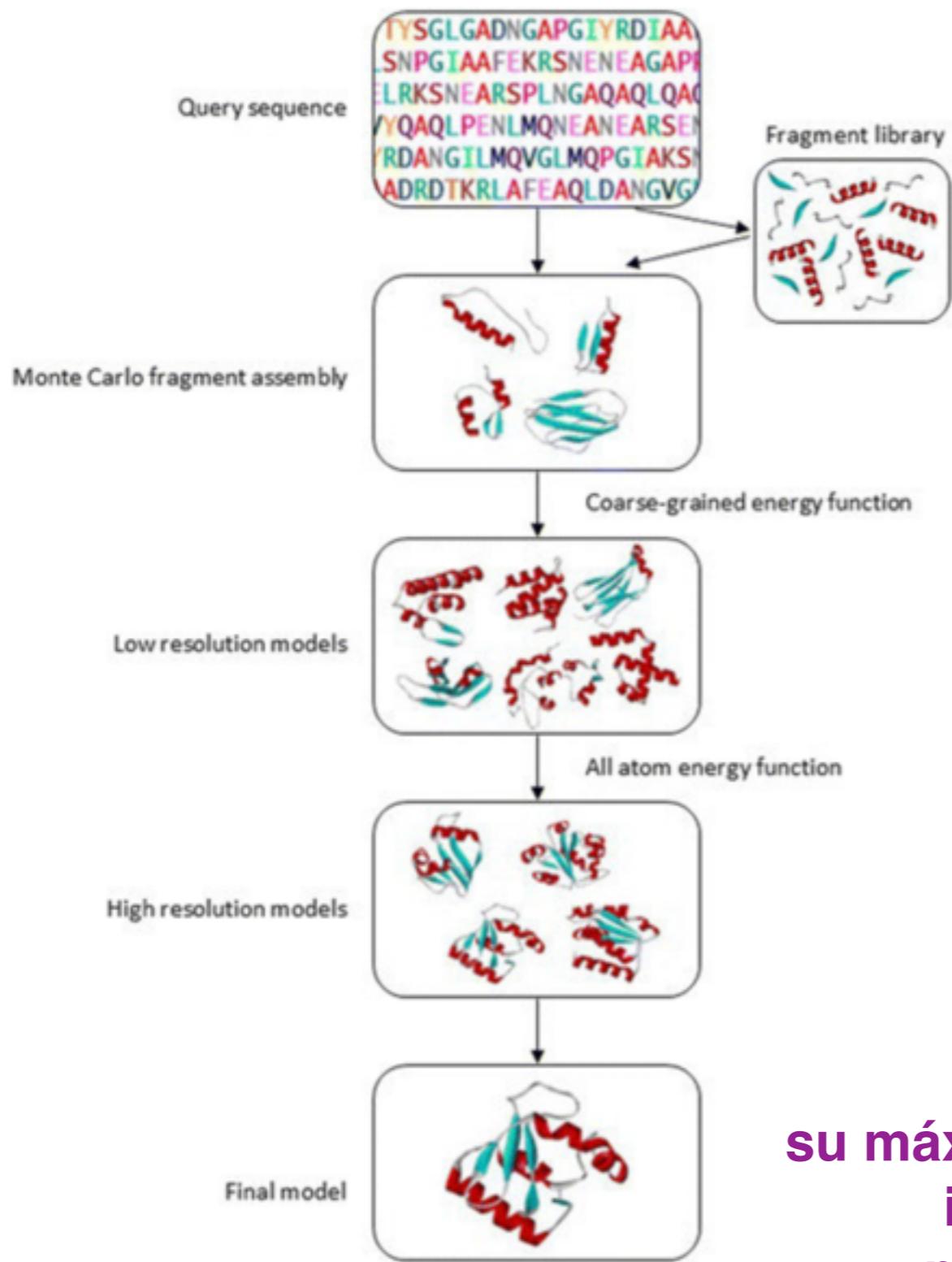


suena bien...pero hay un gran problema: este proceso consume mucho tiempo y puede producir más errores. Es confiable sólo para dominios pequeños <80 res



Esto hace que -en general- el modelado por homología sea mas confiable que la simulación de plegado de novo

Modelado de novo usando bibliotecas de fragmentos



ROSETTA (D. Baker)

RMSD típicos 4-6 Å

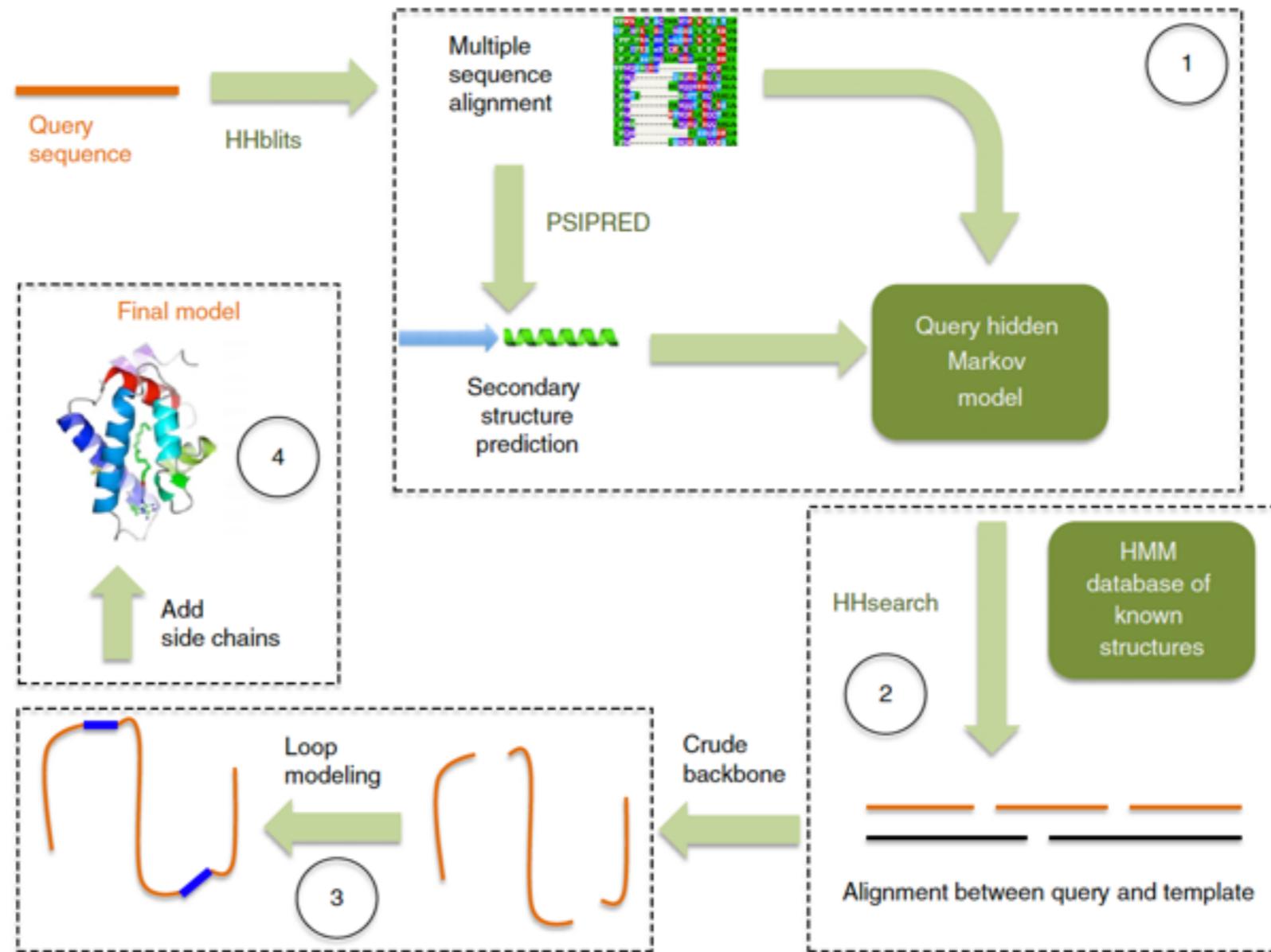
su máxima utilidad es cuando no se puede identificar un buen templado a partir de secuencias homólogas

El modelado por homología se impuso como el mejor método de predicción de estructura 3D

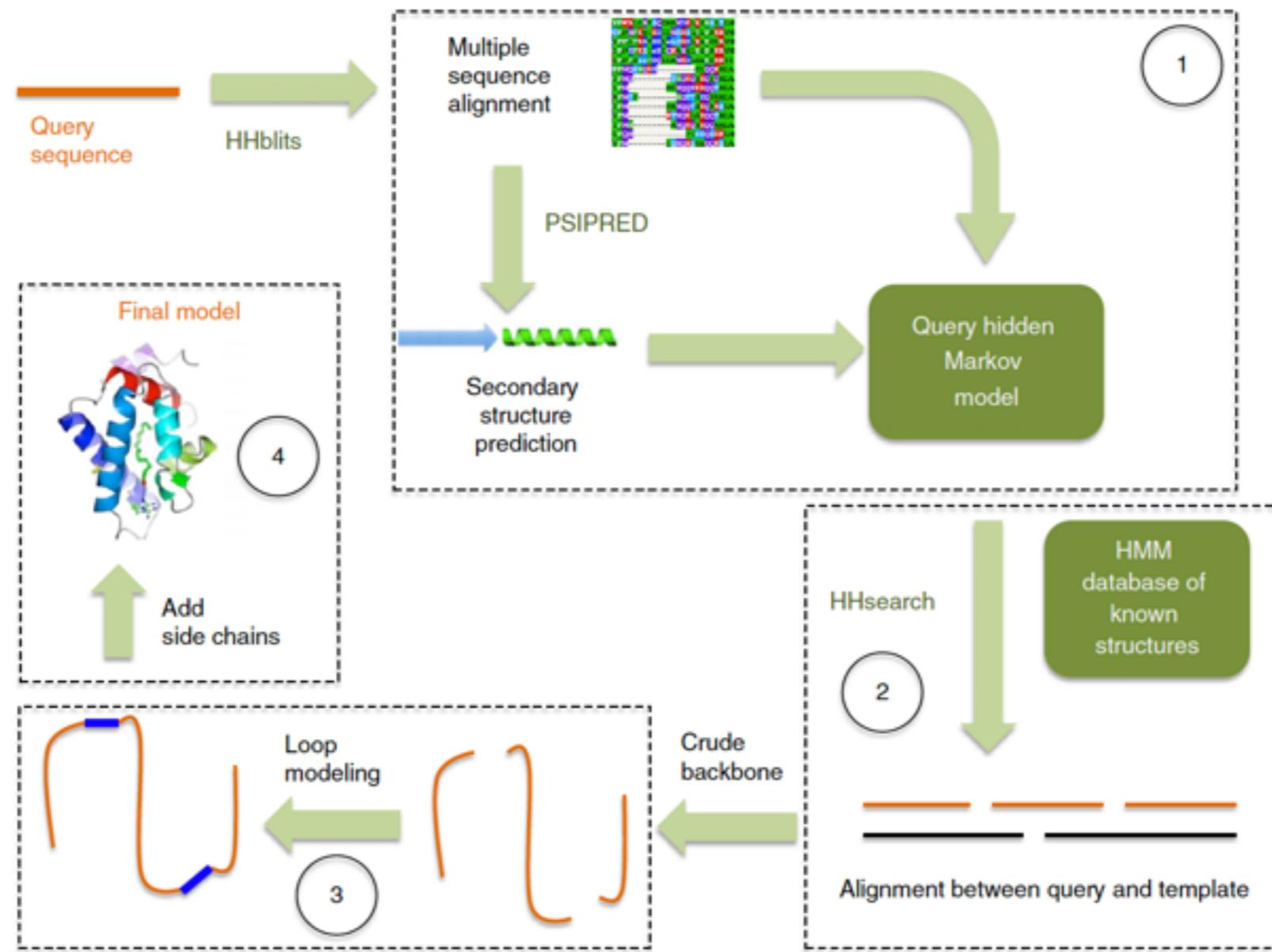
Cuáles son los pasos a seguir?

- A partir de una secuencia **query**, encontrar una secuencia homóloga cuya estructura se encuentre resuelta y resulte el mejor **template de modelado**
- Realizar un alineamiento de la secuencia **query** y la secuencia **template**
- Obtener un **modelo 3D de la secuencia query** en base al alineamiento y la estructura del template
- Evaluar la calidad del modelo obtenido

Esquema del algoritmo para modelado por homología



Esquema del algoritmo para modelado por homología



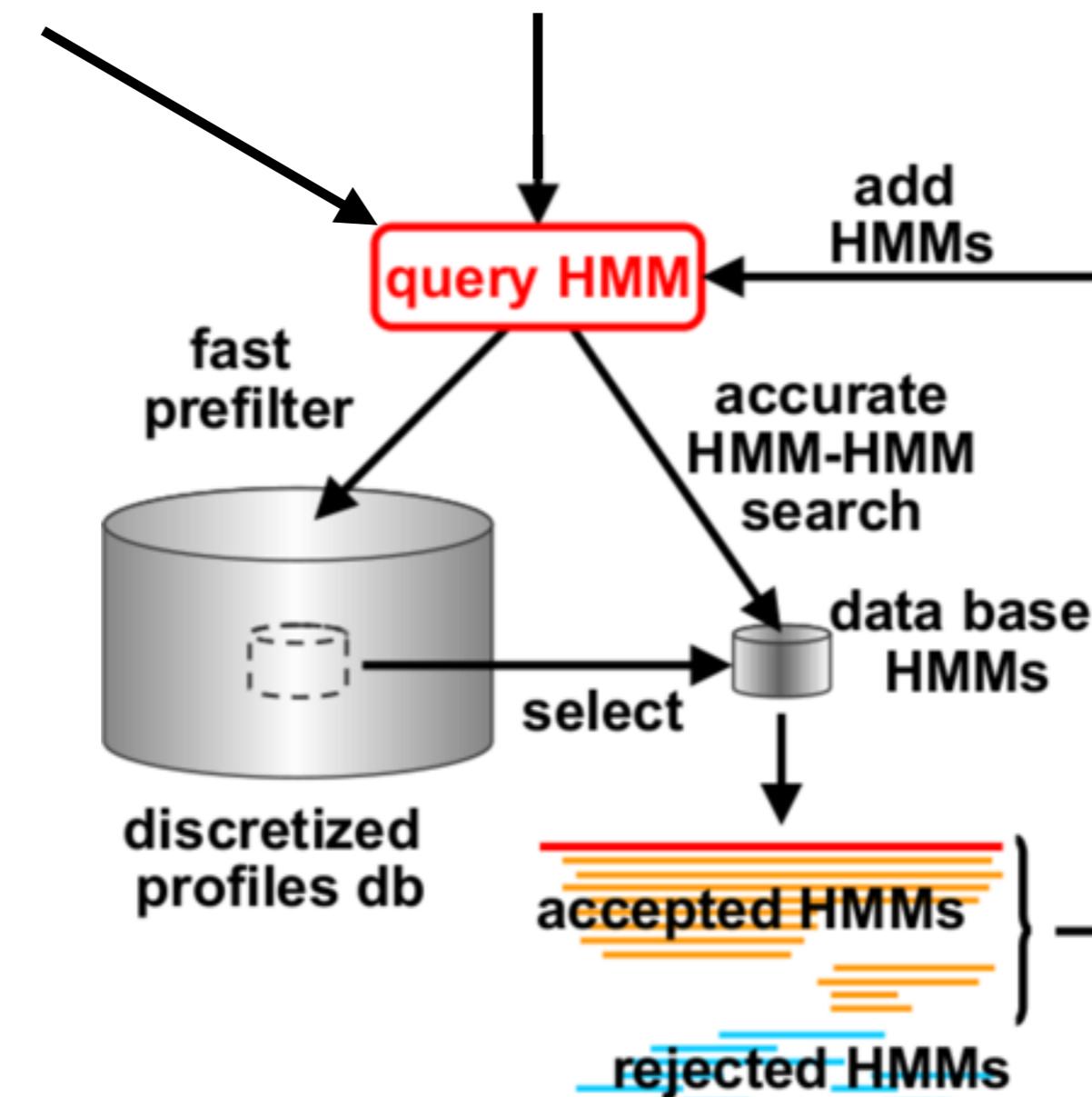
Paso 1: compilación de secuencias homólogas del query y construcción del HMM

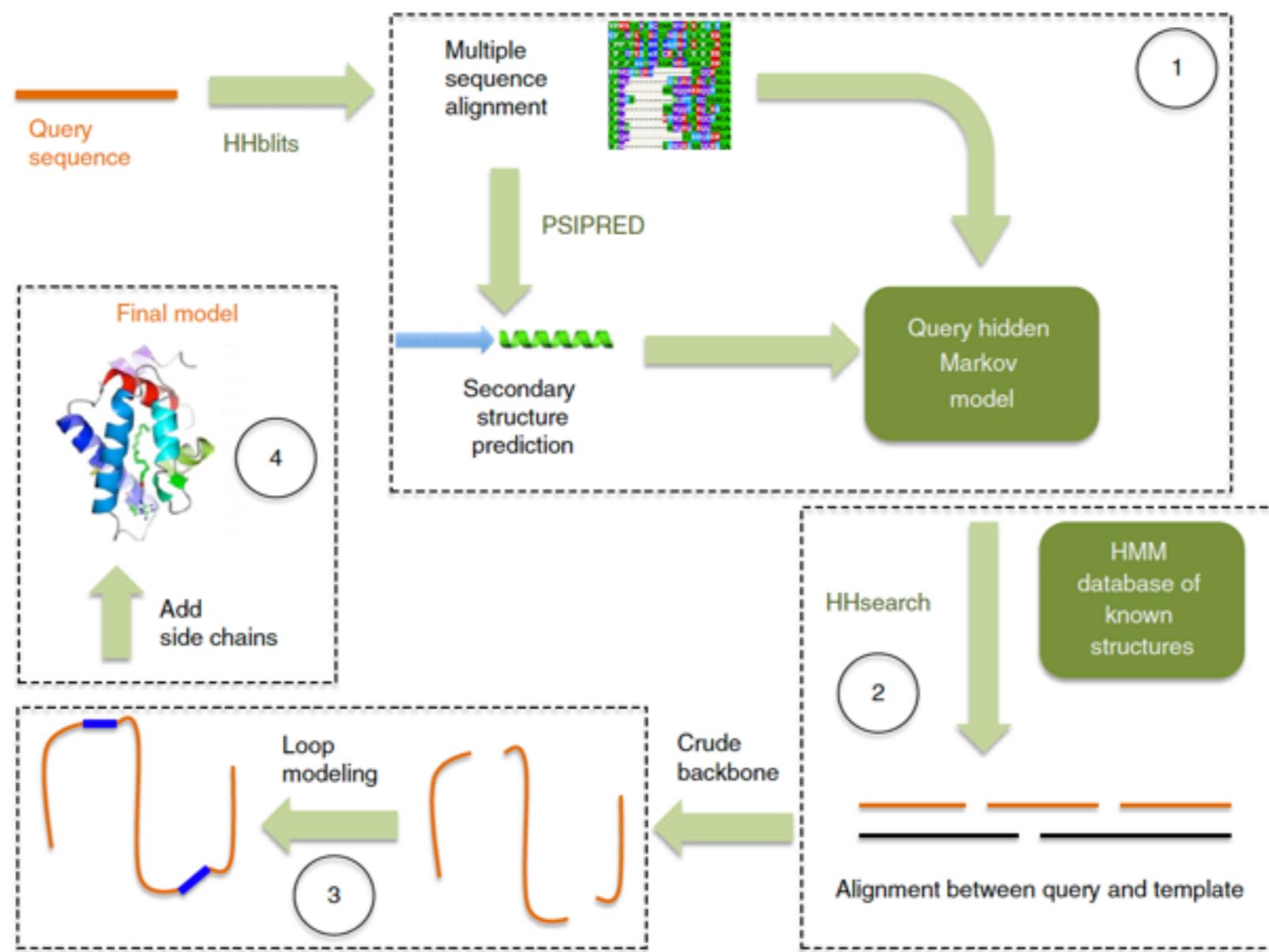
query: se realiza contra la base de datos nr20 (secuencias con <20% ID secuencia) usando HHBlits, y por separado se predice estructura secundaria (PSIPRED). El MSA y la predicción de estructura 2ria se combinan en un HMM de la secuencia **query**

Secuencia query

Predicción
Estr Secundaria

Alineamiento
Base





Paso 2: identificación de un templado por alineamiento del HMM query con una base de datos de HMM de templados: se realiza buscando en una biblioteca de HMMs de una serie de modelos de la PDB. Una vez identificado un HMM templado, se realiza un alineamiento entre **query** y **templado**.

La comparación de HMM-HMM es crítica para la detección de homología (sensibilidad) así como para mejorar la calidad del alineamiento

BLAST

comparación secuencia-secuencia. Matriz de scoring basada en sustituciones observadas en un conjunto de dominios globulares (**sin información posicional**)

PSI-BLAST

comparación secuencia-profile basada en un MSA de una familia de proteínas homólogas. Las probabilidades de ocurrencia de aminoácidos varían en diferentes posiciones del MSA (**contiene información posicional**)

COMPASS
CLUSTAL

comparación profile-profile basada en un MSA de una familia de proteínas homólogas. Las probabilidades de ocurrencia de aminoácidos varían en diferentes posiciones del MSA (**contiene información posicional**)

HHPred

comparación HMM-HMM basada en un MSA de una familia de proteínas homólogas. Las probabilidades de ocurrencia de aminoácidos varían en diferentes posiciones del MSA e incorpora probabilidades sitio-específicas de inserción o delección (**contiene información posicional de aa e INDELS**)

BLAST

PSI-BLAST

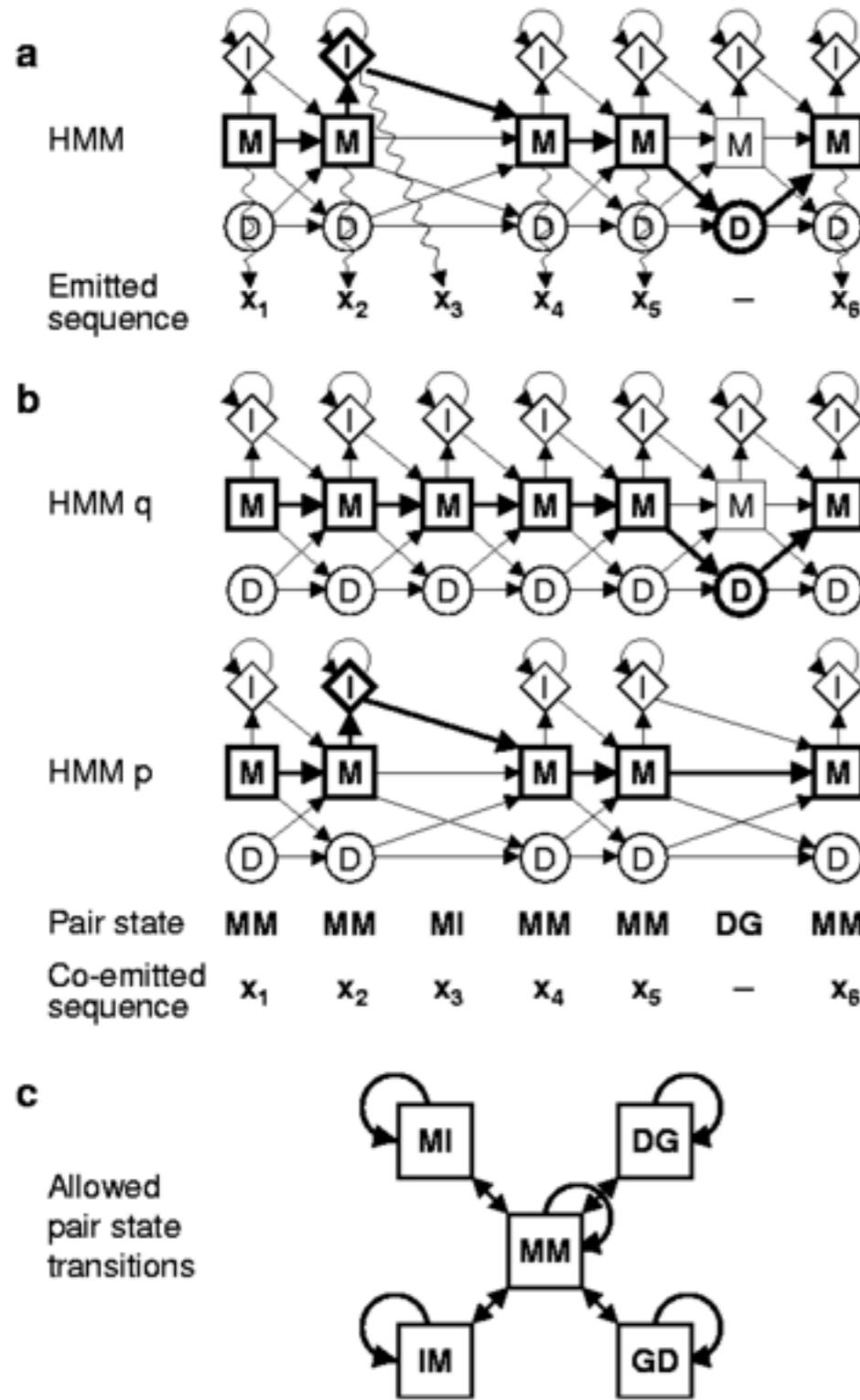
CLUSTAL

HHPred

sensibilidad y calidad de alineamiento

profile = PSSM (Position Specific Scoring Matrix)

HMM = Hidden Markov Model



La detección de homólogos y el alineamiento de query y target son determinantes principales del éxito del modelado por homología

Los mejores métodos hacen uso del alineamiento de un HMM del query con un HMM del templado

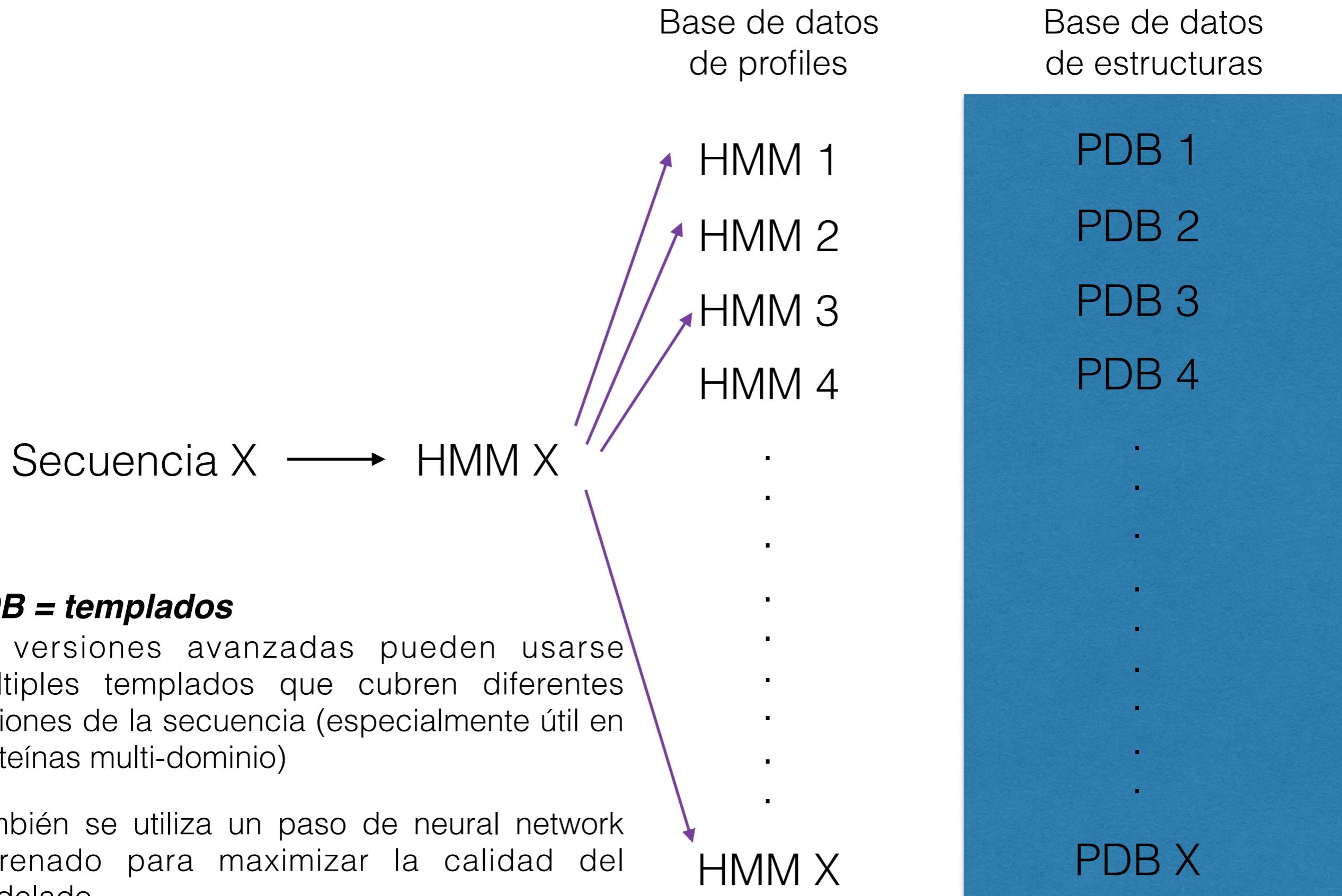
El método incorpora peso de secuencias (sequence weighting) y pseudocuentas

Función de scoring: Log-sum-of-odds score

$$S_{LSO} = \log \sum_{x_1, \dots, x_L} \frac{P(x_1, \dots, x_L | \text{co-emission on path})}{P(x_1, \dots, x_L | \text{Null})}.$$

El algoritmo maximiza la función de scoring

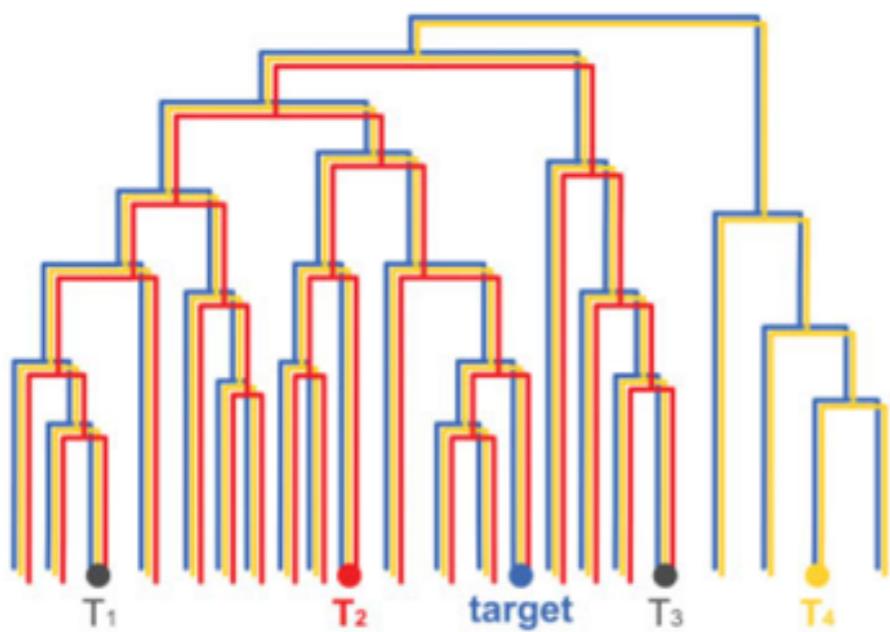
Alineamos una secuencia para la que no conocemos la estructura con una secuencia de estructura conocida



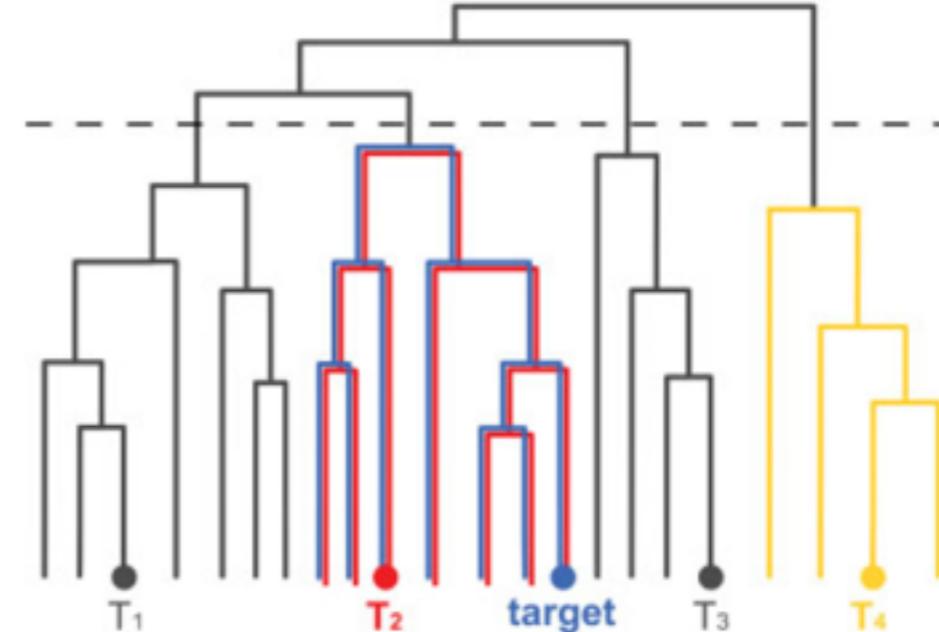
en los pasos finales de la selección de templados, se incluye un paso en el cual se reduce la diversidad de secuencias del HMM para facilitar la detección del homólogo más cercano

Se muestra un target (**query**) y 4 templados T1-T4

Sin filtrado

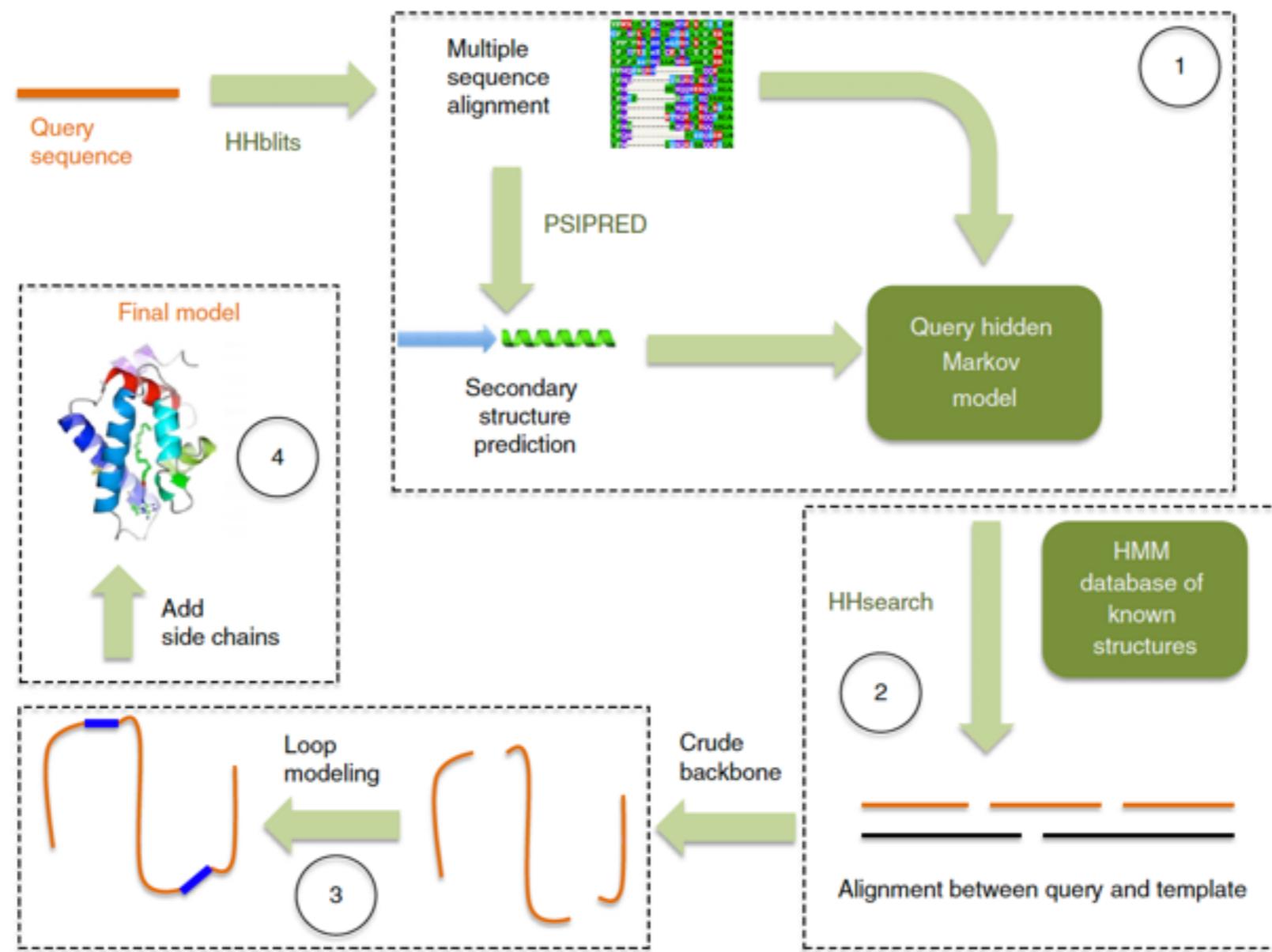


Con filtrado

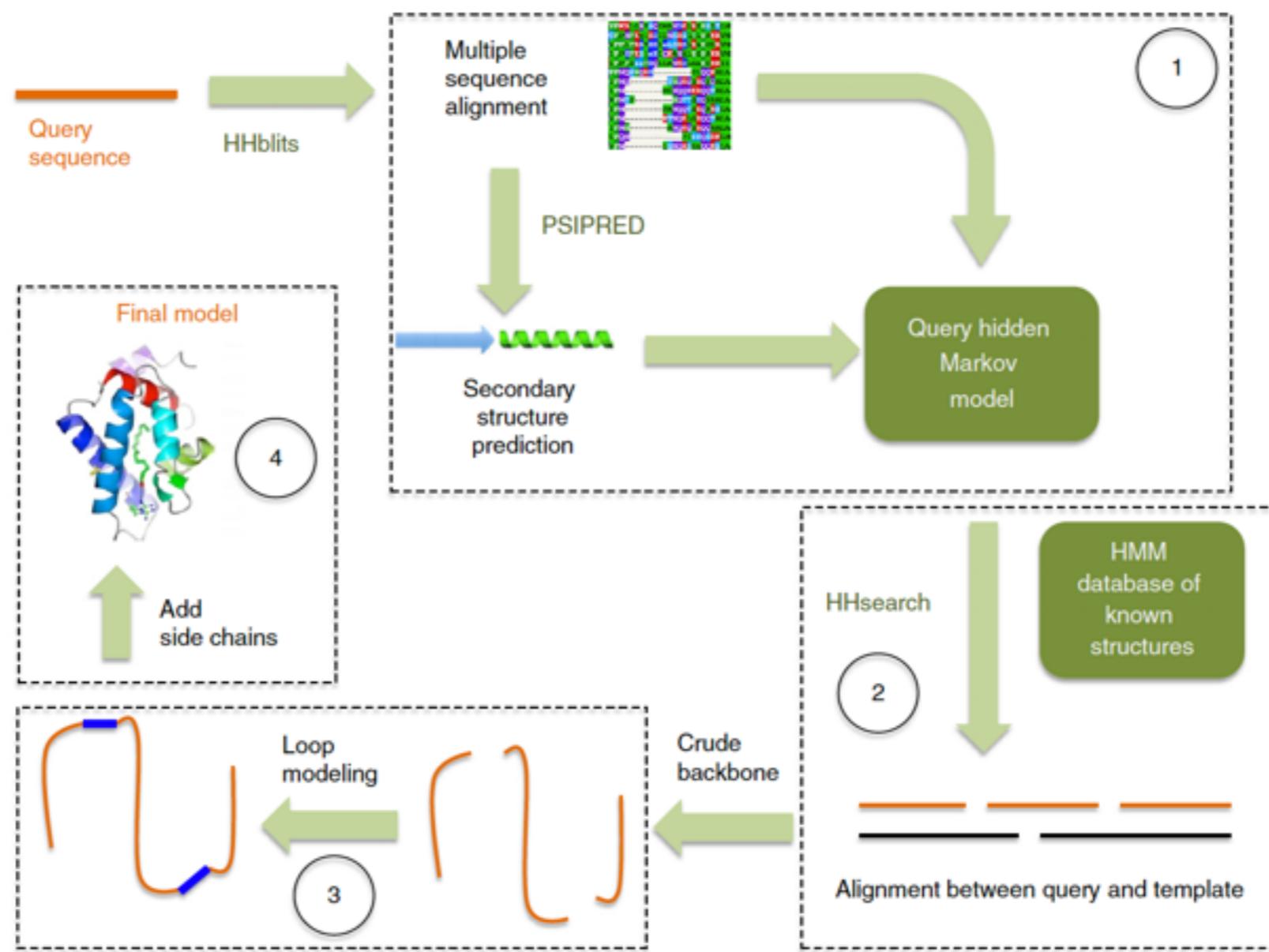


Al aplicar un filtro de diversidad, se detecta que el HMM del templado T2 es el más cercano al HMM del **query**

- Seq en HMM del query
- Seq en HMM de T2
- Seq en HMM de T4

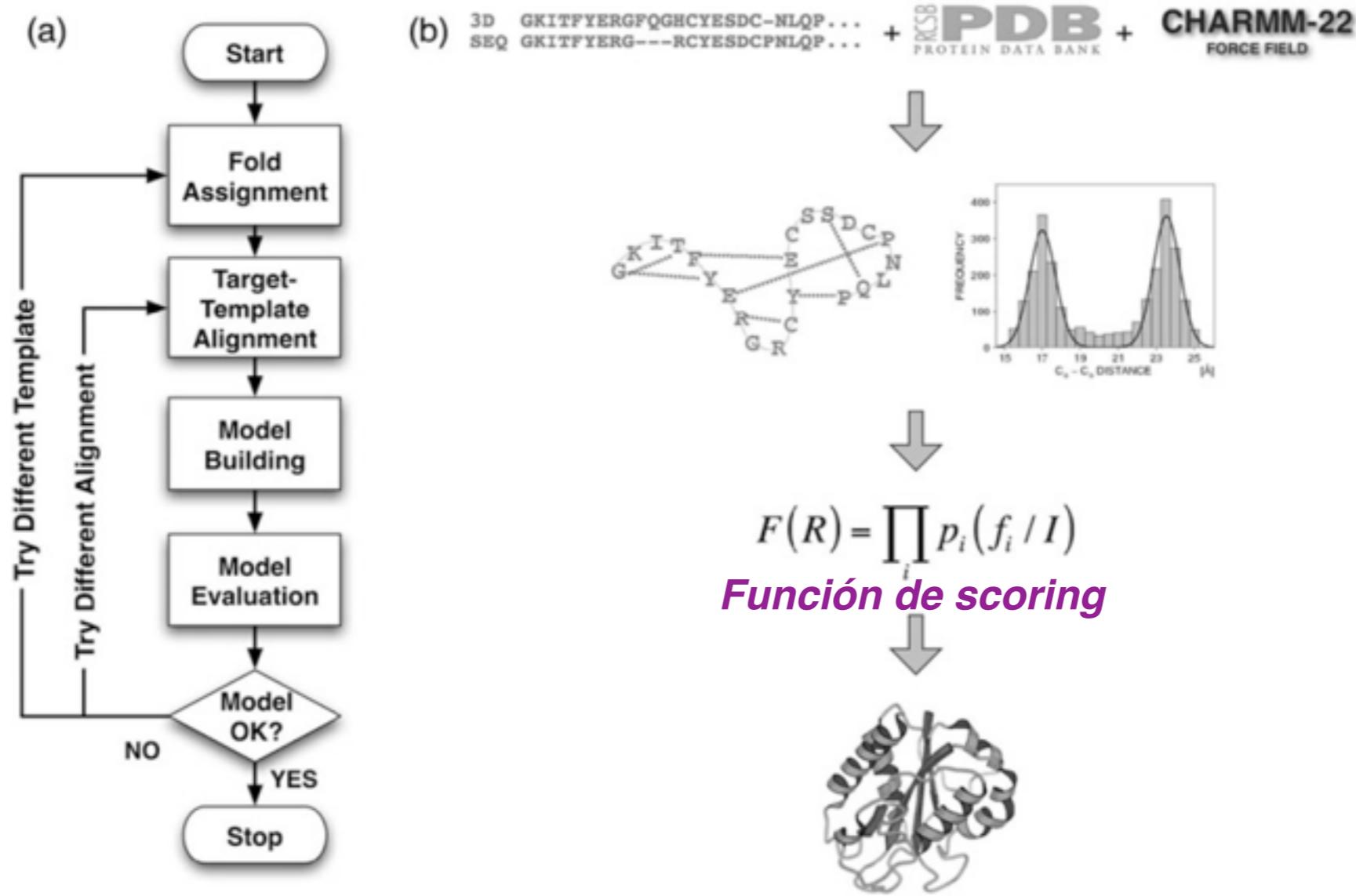


Paso 3: Producción de un modelo primario y modelado de loops: En primer lugar el alineamiento de (2) se utiliza para producir un modelo de la cadena principal (backbone), sin cadenas laterales. En segundo lugar, a esta cadena principal, se adicionan los “loops” por modelado.



Paso 4: Producción de un modelo completo por adición de cadenas laterales: Una vez construida completamente la cadena principal y los loops, se agregan las cadenas laterales, buscando minimizar los choques estéricos, y respetando la estereoquímica de los diferentes residuos.

Pasos: construcción de un modelo 3D MODELLER



- restricciones de distancia: provenientes del alineamiento **query-templado**
- restricciones estadísticas: provenientes de estructuras conocidas (Protein Data Bank)
- restricciones estereoquímicas de un campo de fuerzas físico (CHARMM-22)

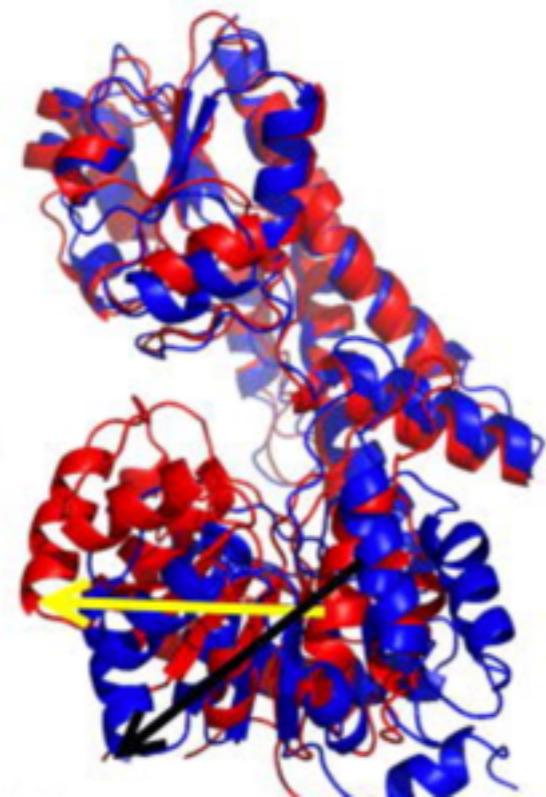
Cómo se comparan dos estructuras?

superposición espacial

Ej: **query** vs **templado**

RMSD

root mean square deviation

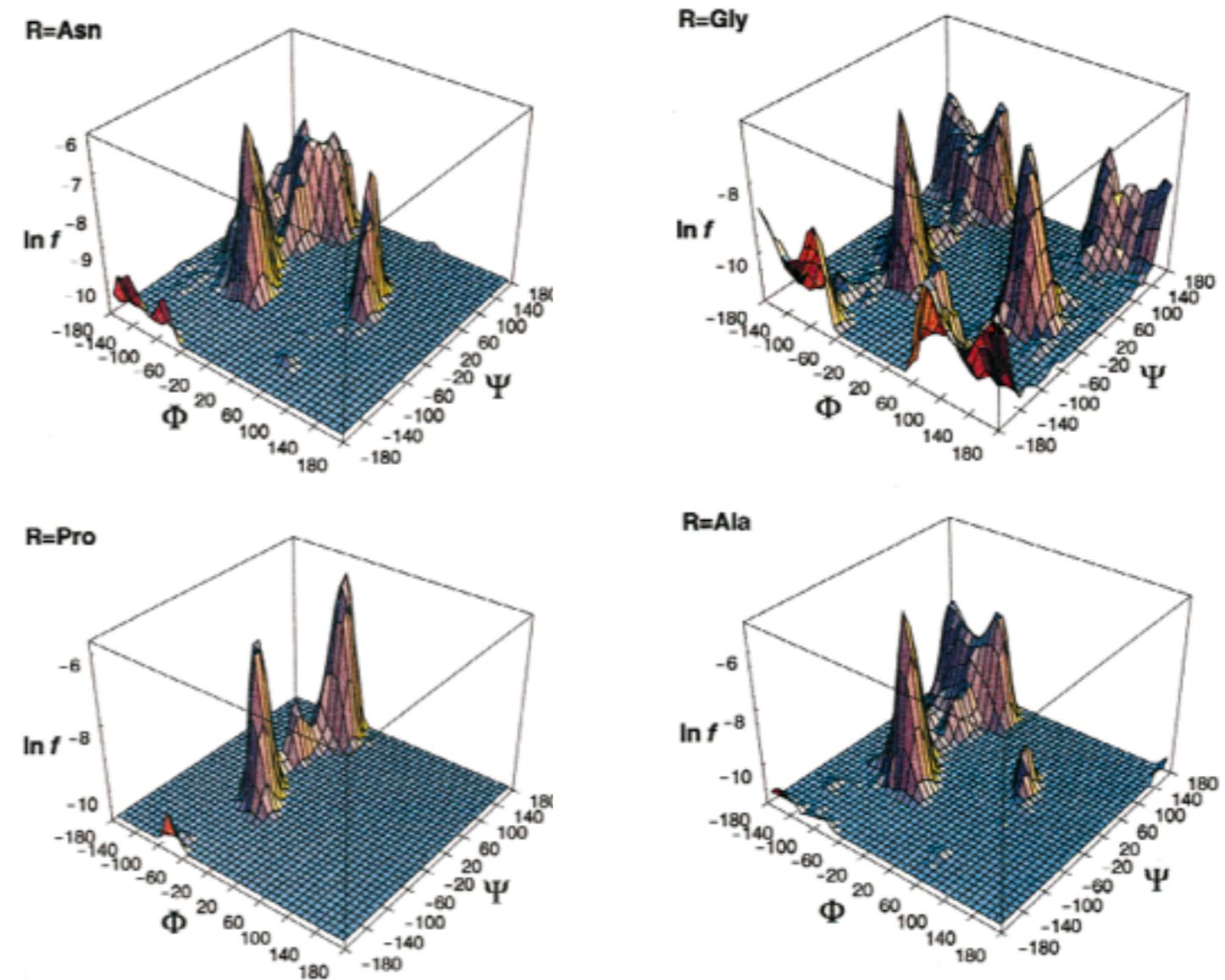


$$\begin{aligned} \text{RMSD}(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)} \end{aligned}$$

RMSD: es una medida de similitud entre dos estructuras. Es una diferencia cuadrática media sobre las coordenadas x, y, z de cada átomo

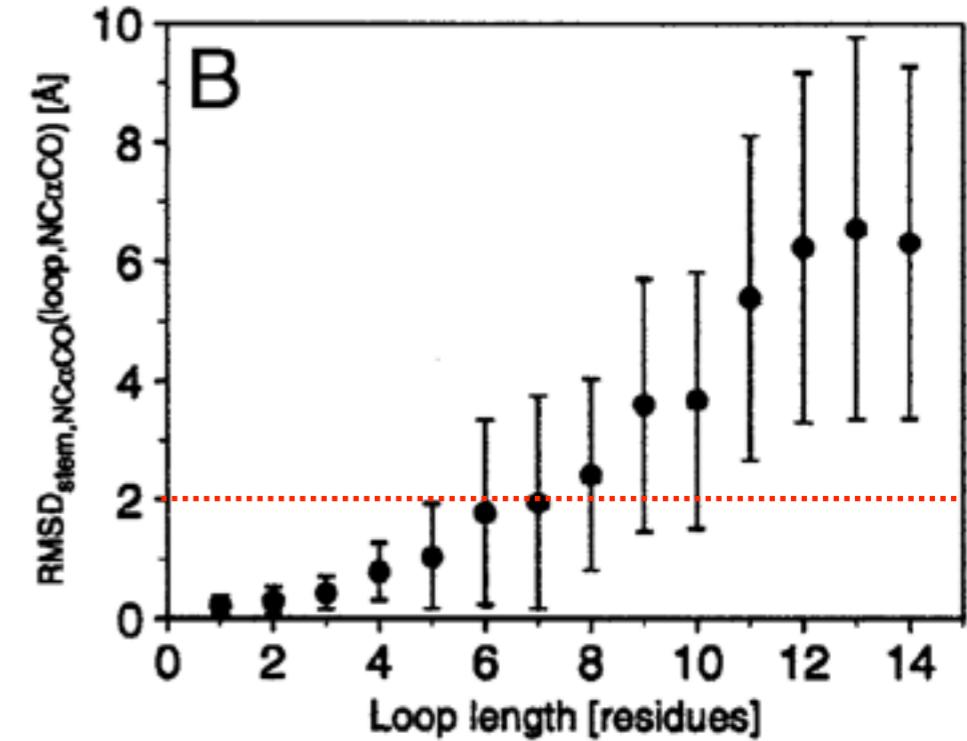
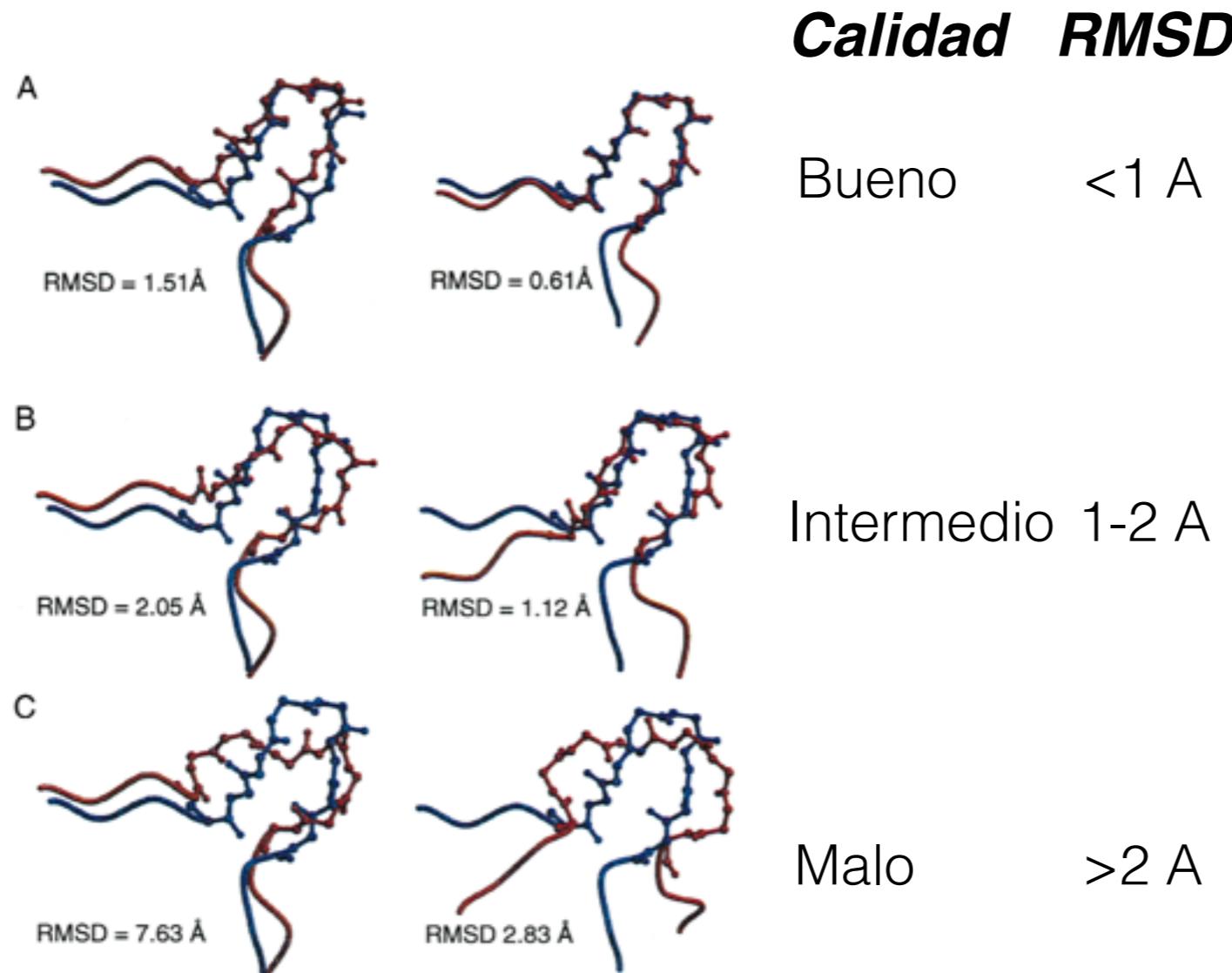
Cadena principal (***backbone***)

Se modela buscando satisfacer el alineamiento con el templado y los ángulos de ramachandran



Modelado de “loops”

Es otro de los pasos más críticos y limitantes del modelado por homología



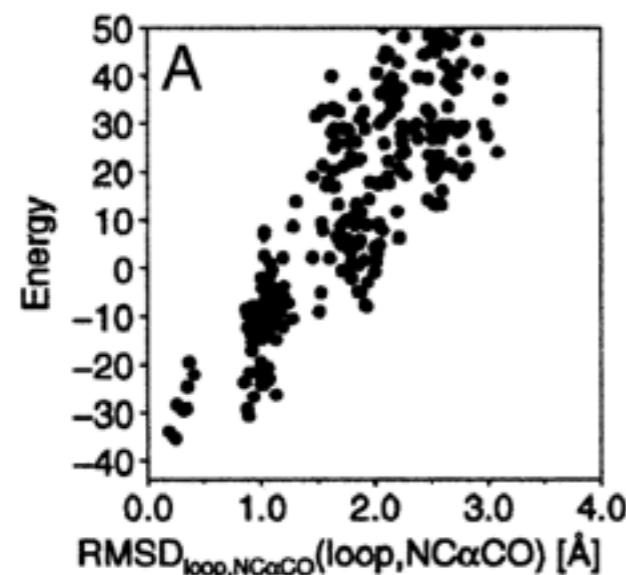
Limitado a loops de 3-15 residuos

Es difícil modelar con buena exactitud loops mayores a 6-8 res

Se utiliza una biblioteca de fragmentos obtenidos por “clustering” de la PDB

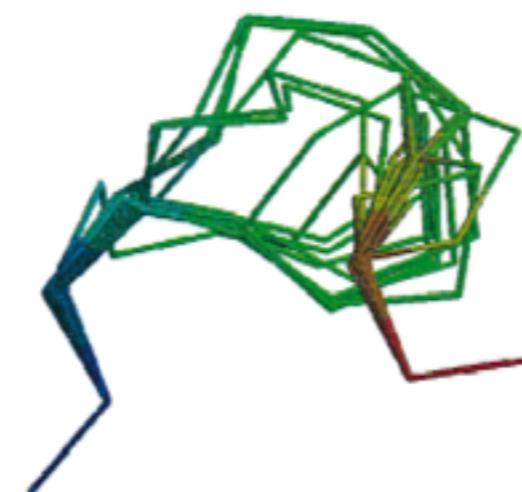
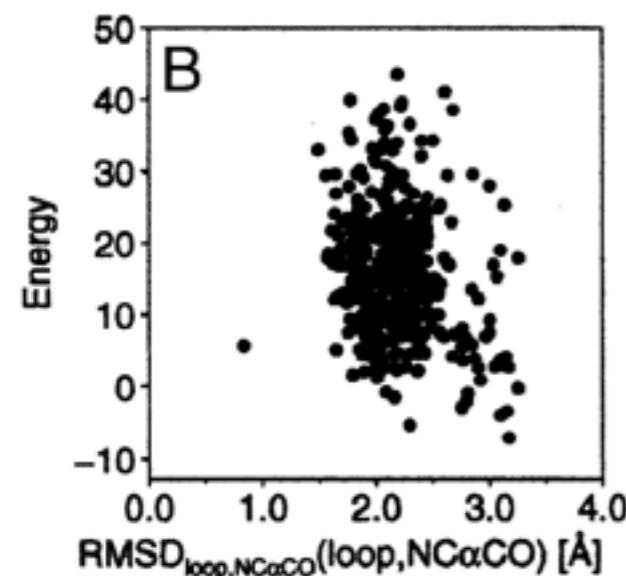
Se toma en cuenta la secuencia de las regiones flanqueantes, y la distancia entre los puntos de inicio y final del loop

Ejemplos: Modelado de “loops”



Modelado exitoso

las estructuras de menor energía coinciden y tienen bajo RMSD



Falla en modelado

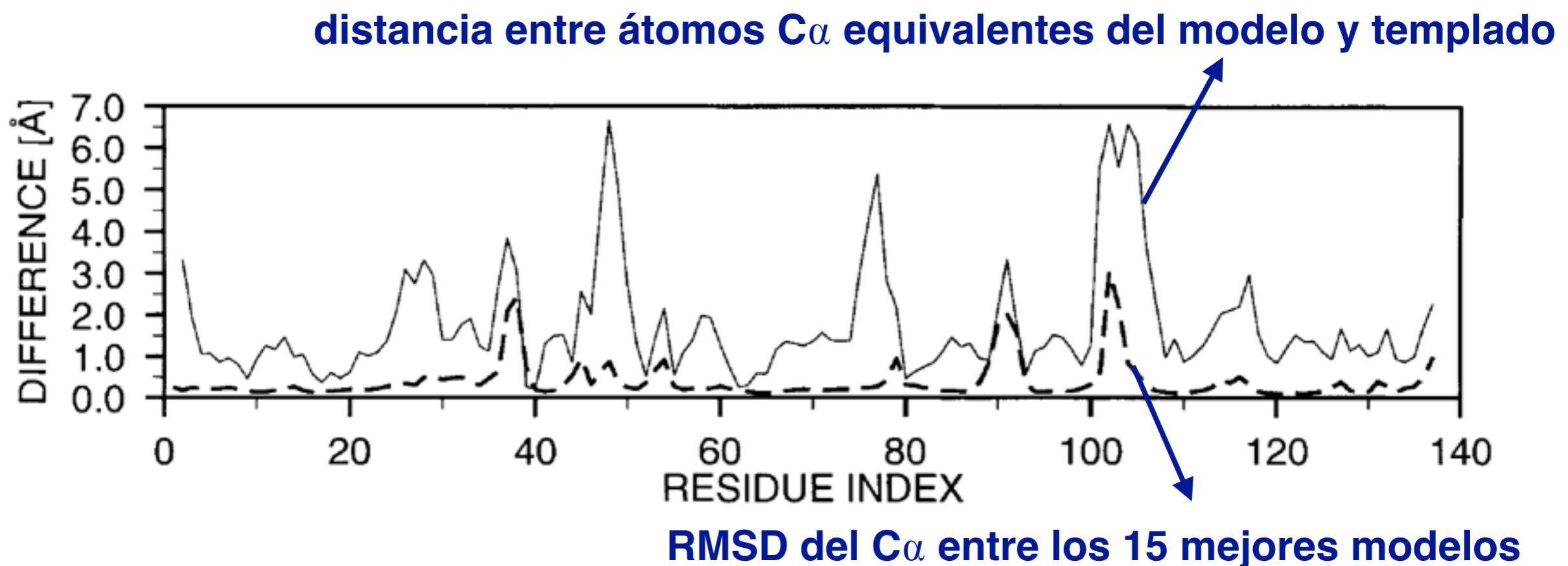
las estructuras de menor energía no coinciden y tienen mayor RMSD

Adición de cadenas laterales

Se hace utilizando una Biblioteca de rotámeros

Se busca evitar choques estéricos

Precisión de ~80% ***si la asignación del backbone es buena***

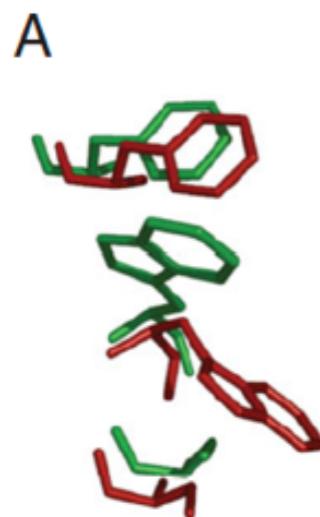


La variabilidad entre los mejores modelos correlaciona con el error de modelado



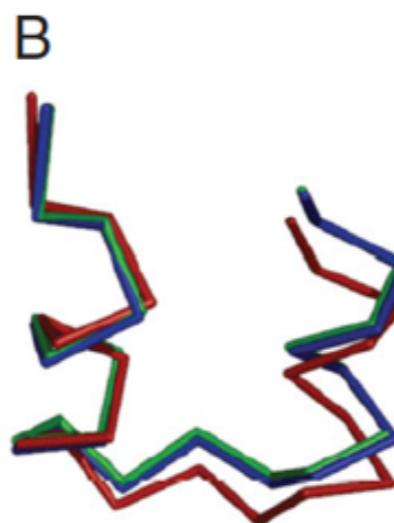
En un buen modelo, los mejores templados coinciden entre sí

ERRORES MAS COMUNES DEL MODELADO POR HOMOLOGIA



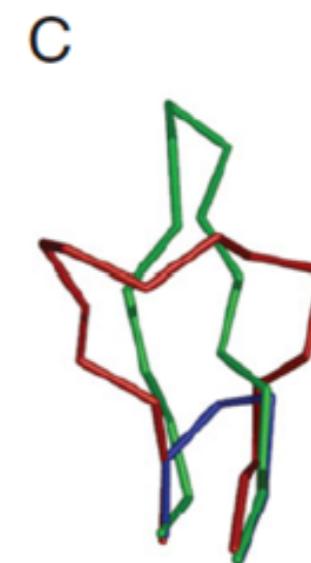
Errores en el posicionamiento de cadenas laterales

W109 en retinoid acid binding protein Xray (rojo) y modelo (verde)



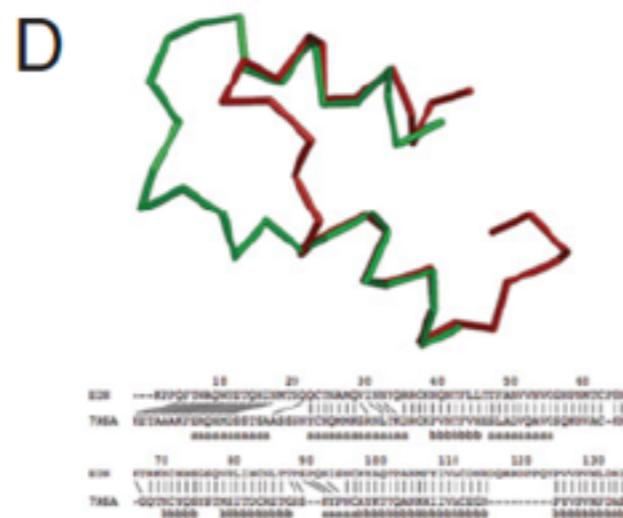
Distorsiones y corrimientos en regiones bien alineadas

retinoid acid binding protein:
estructura (rojo) y modelo (verde)
comparado con templado fatty acid
binding protein (azul)



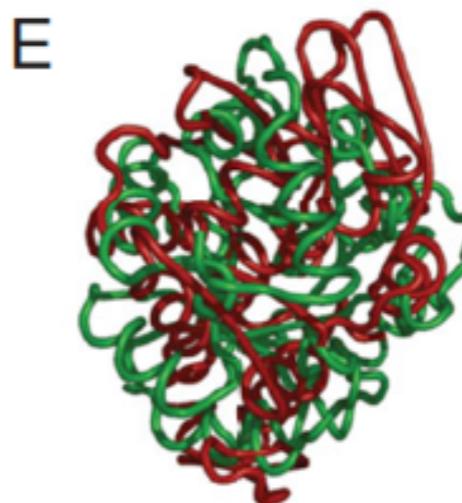
Errores en regiones que carecen de templado

El loop 112-117 de EDN (verde)
modelado en base a la estructura
de Ribonucleasa (azul).
Rojo: estructura correcta



Errores debidos fallas en el alineamiento

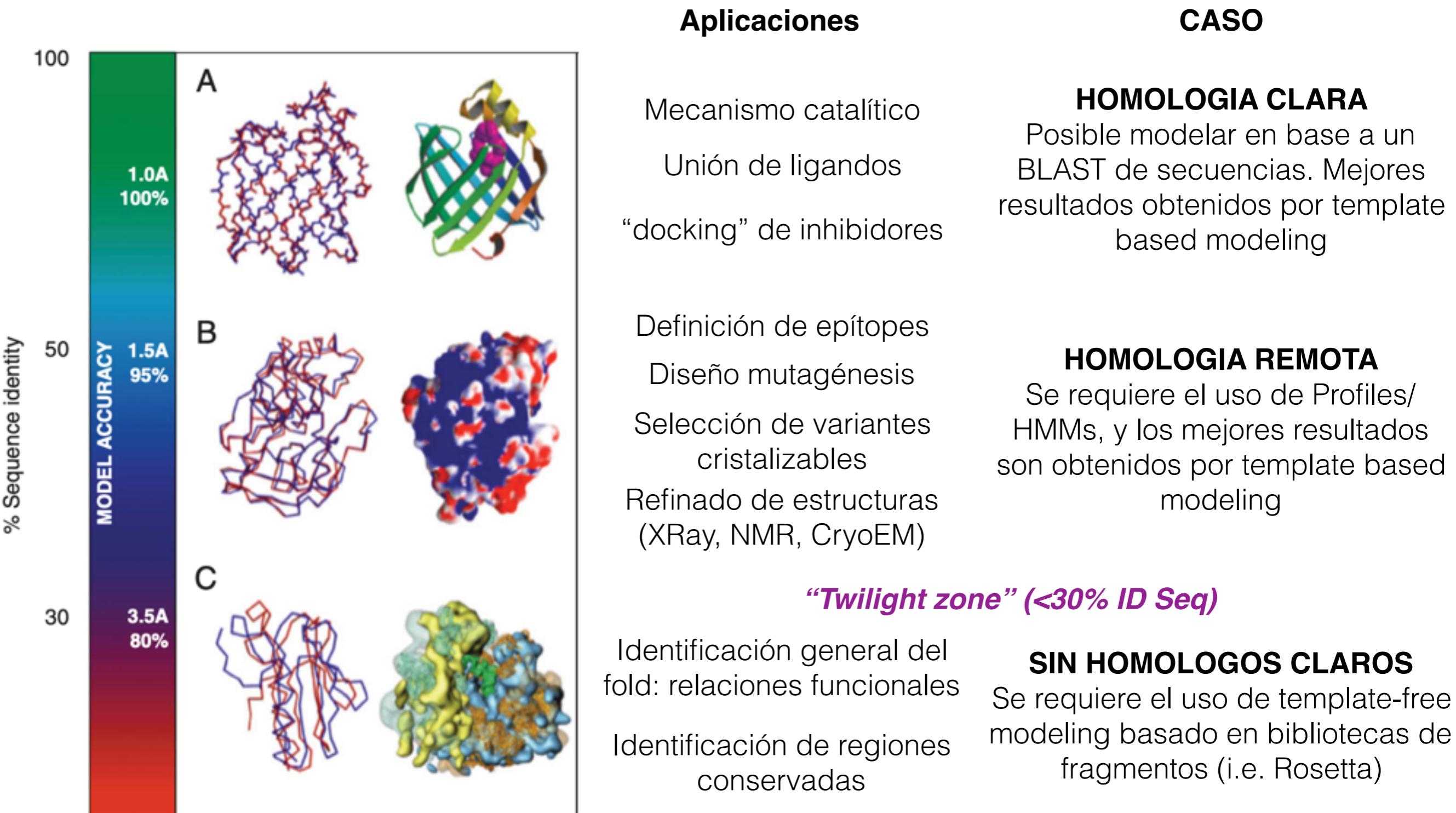
región N-terminal de neurotoxina de eosinófilos (EDN) (rojo) comparada con su modelo (verde)



Errores debidos a un templado incorrecto

α -Tricosanthin modelada usando la estructura de 3-glicerofosfato sintasa

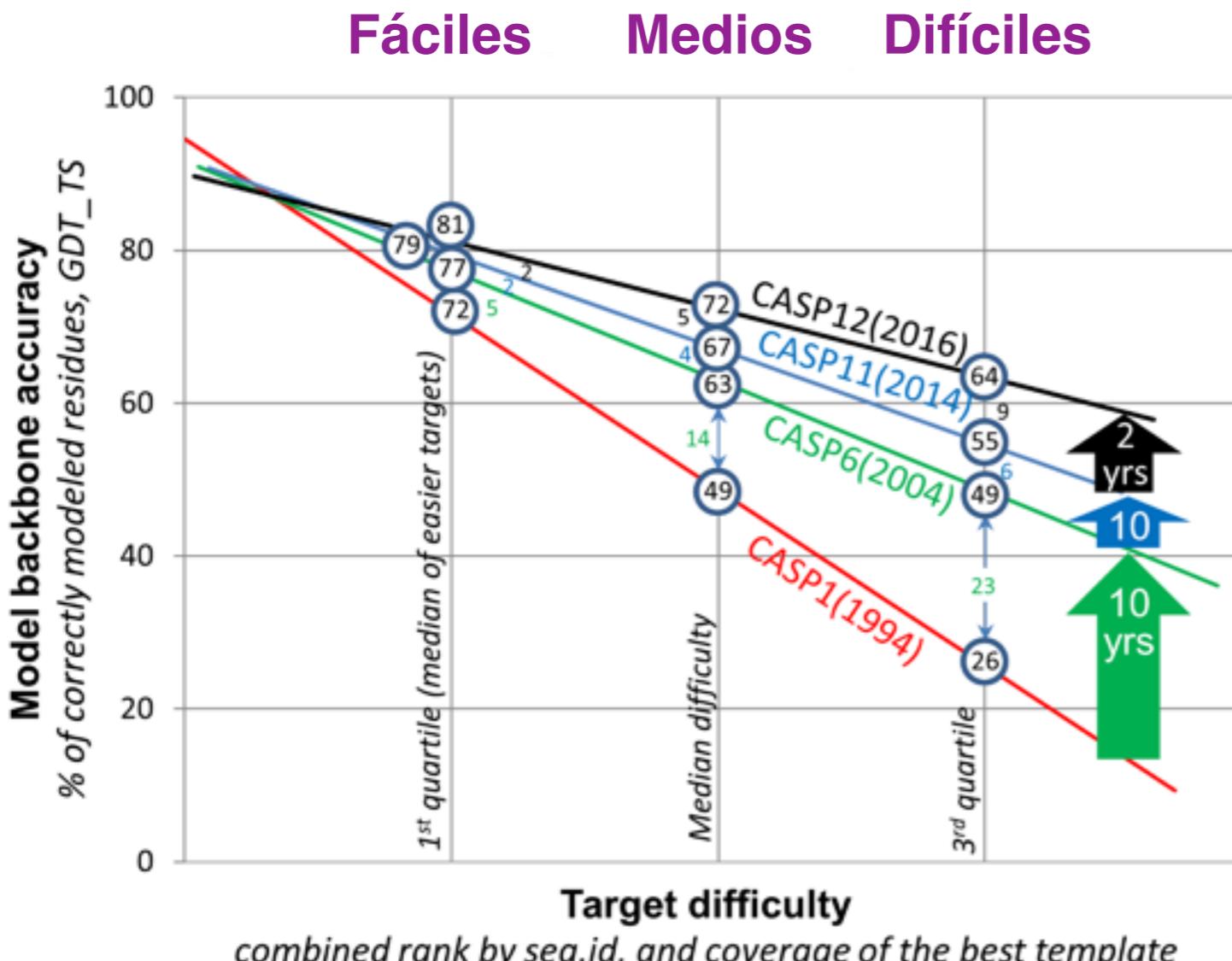
Propiedades generales del modelado estructural



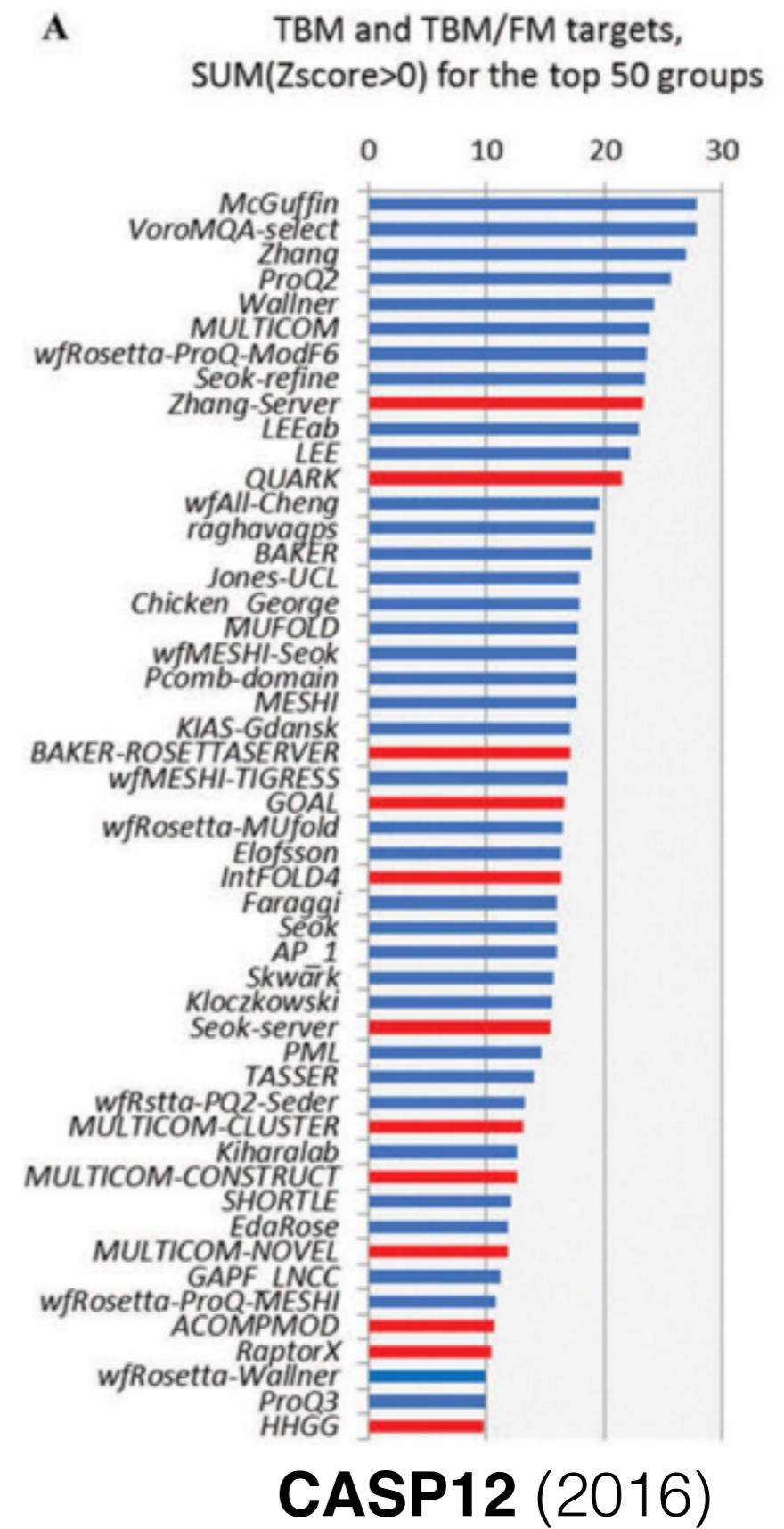
Comparación de métodos de modelado

Competición CASP

critical assessment of structure prediction

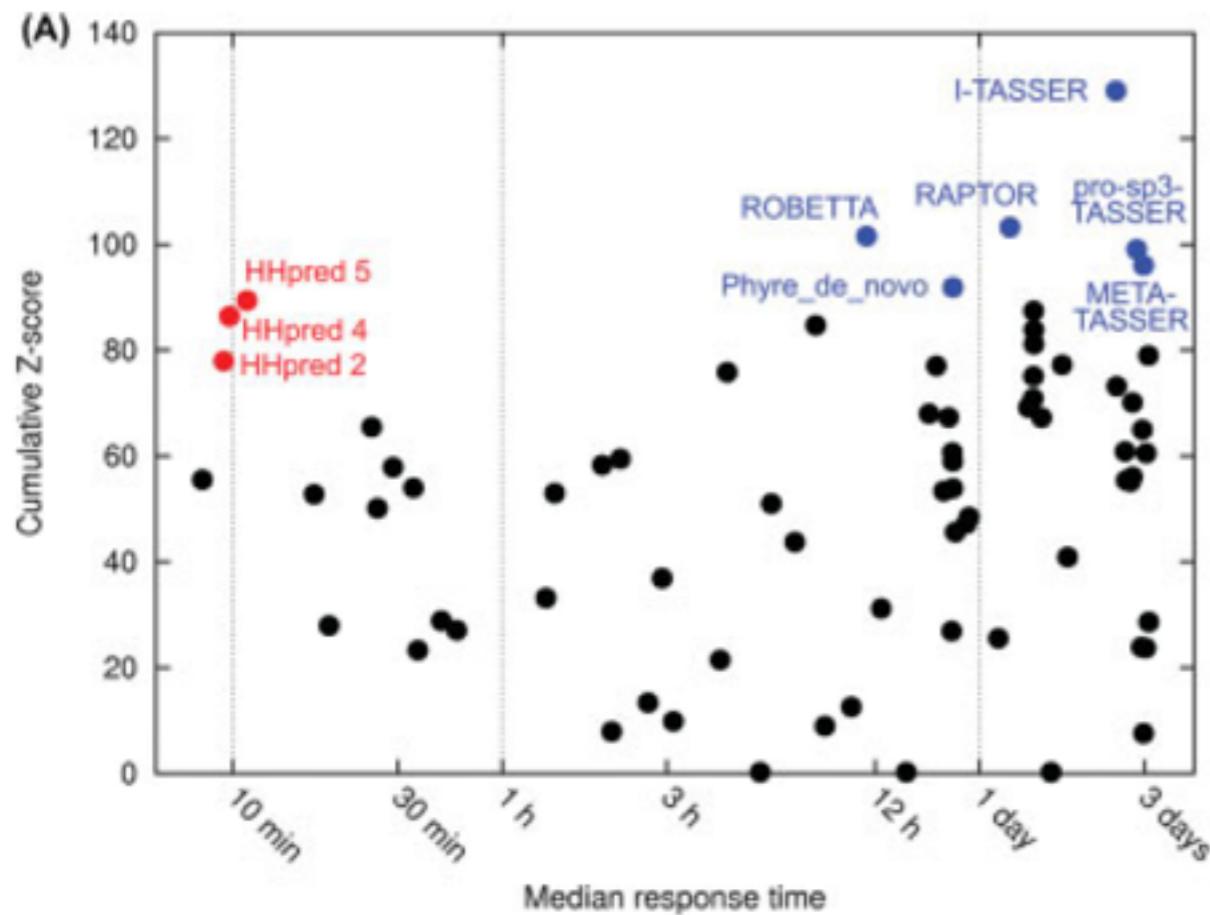


La calidad de predicciones fue aumentando a lo largo de los años

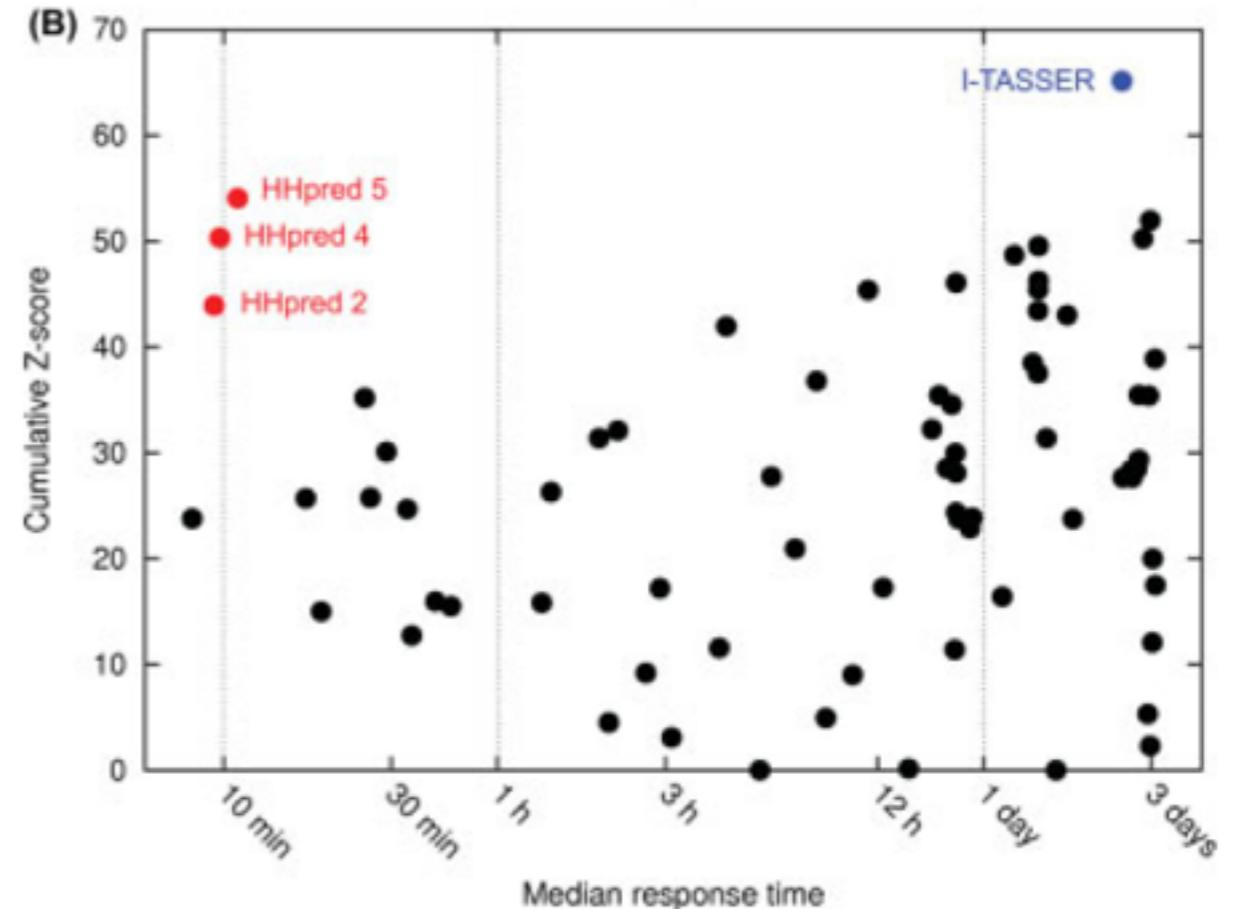


Rapidez versus exactitud de métodos de predicción

164 proteínas target



85 proteínas de un sólo dominio



HHPred modela mucho más rápido que los métodos que le compiten en exactitud

Puede observarse que HHPred funciona mucho mejor sobre dominios individuales

Azul: servers mejores que HHPred

Z-Score: score global para evaluar los servers

resultados CASP8

FIN