UNIVERSIDAD ANDRES BELLO

Predicción de estructura de proteínas

Bioinformática

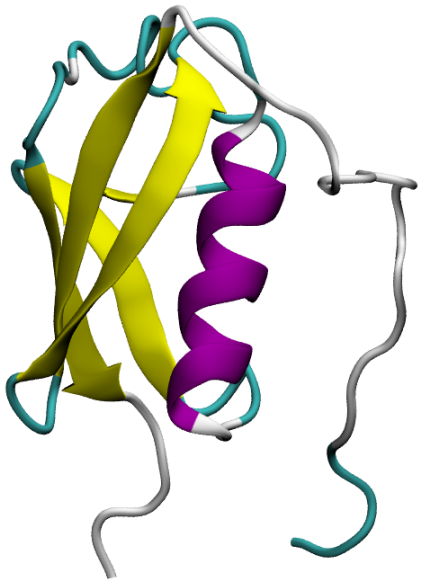
5 de Mayo 2017

Predicción de Estructura Secundaria de Proteínas

Se estima que casi el 50% de los residuos de una proteína se empaquetan en forma de *α-hélices o sábana-β*

Predicción de estructura secundaria se refiere a la predicción del **estado conformacional** de cada residuo de la secuencia de una proteína como uno de tres posibles estados: hélice (H), sábana (E) o coil (C).

Las predicciones se basan en el hecho de que los elementos de estructura secundaria tienen un arreglo regular de amino ácidos, estabilizados por puentes de hidrógeno.



Predicción de Estructura Secundaria de Proteínas: Aplicaciones

Durante la evolución, las estructuras terciarias, y por tanto secundarias son mucho más conservadas que las secuencias. Como resultado, la identificación de **elementos de estructura secundaria ayuda a guiar los alineamientos de secuencia.**

Es útil para la clasificación de proteínas y para la identificación de dominios y motivos.

La predicción de estructura secundaria es un paso intermediario en la predicción de estructura terciaria.

Predicción de Estructura Secundaria de Proteínas Globulares

La predicción de estructura secundaria es dependiente del contexto:

- **α-hélices** determinada por interacciones de corto alcance
- **sábanas -β** determinada por interacciones de largo alcance (más difícil)

La exactitud de las predicciones actualmente llega al 75%.

Los métodos de predicción pueden ser:

- **ab initio**: Predicen la estructura secundaria empleando información estadística calculada a partir de una sola secuencia.
- **basados en homología**: usan información de alineamientos múltiples, ya que no confian en estadisticas de una sola secuencia, ya que patrones estructurales son conservados entre multiples secuencias homologas.

Métodos Ab-Initio

- Fueron inicialmente desarrollado en los años 70.
- Miden la **propensidad** de cada aminoácido de pertenecer a cierto elemento de estructura secundaria, basado en datos estructurales de proteínas conocidas.

Propensidad = P = (x /m) / (y / n)

n = número total de residuos
m = número de residuo que aparecen en cierto tipo de estructura secundaria (e.g. α-hélice)
y = número total de un residuo (e.g. alanina)
x = número total de un residuo en un tipo de estructura secundaria (e.g. alaninas en α-hélices)

Interpretación:

Propensidad > 1: el residuo aparece preferentemente en este tipo de estructura secundaria
Propensidad < 1: el residuo no aparece preferentemente en este tipo de estructura secundaria

Métodos Ab-Initio: Chou-Fasmany, GOR,

Métodos Ab-Initio: Chou-Fasman

TABLE 14.1. Relative Amino Acid Propensity Values for Secondary Structure Elements Used in the Chou–Fasman Method

Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

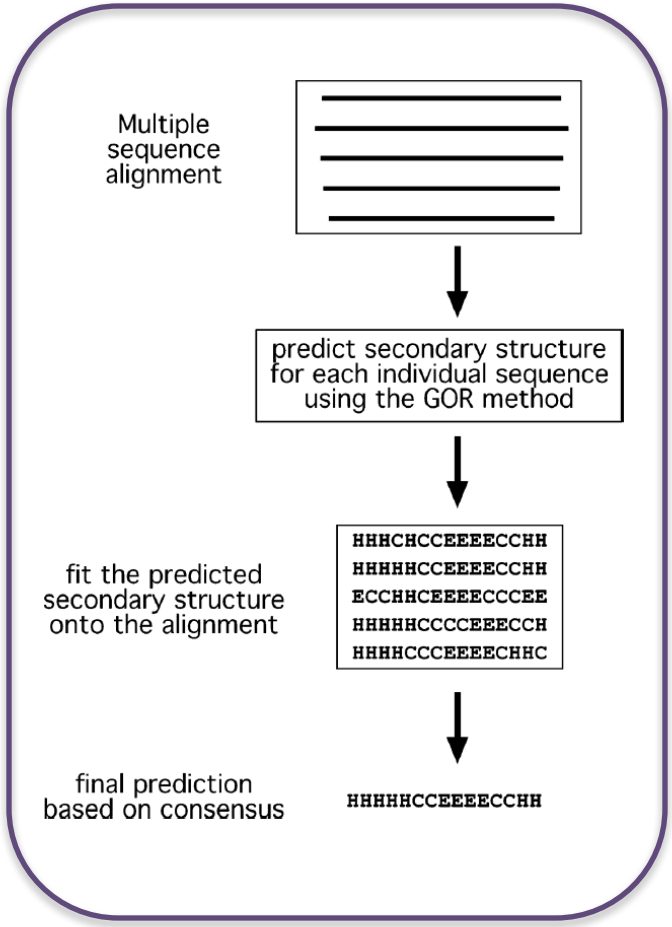
	Común	**Poco común**
α-hélices	alanina Ácido glutámico Metionina	Glicina Prolina
sábana-β	Isoleucina Tirosina Valina	Ácido glutámico Ácido aspártico Prolina
coil	Asparragina Glicina Prolina	Isoleucina Leucina Valina

Métodos Ab-Initio: Limitaciones

- Solo consideran interacciones de corto alcance.
- Exactitud de las predicciones sólo llega al 50%. (al azar uno esperaría: 30% α-hélice, 20% sábana-β, 50% coil)
- Tienden a no encontrar sábanas-*β* y a acortar tanto las sábanas como las hélices.

Métodos Basados en Homología

- Fueron desarrollados en los 90.
- Combinan predicción ab-initio de secuencias individuales e información de alineamientos con múltiples secuencias homólogas, debido a que homólogos cercanos adoptan la misma estructura terciaria, y por tanto secundaria.
- Residuos alineados entre las múltiples secuencias debieran adoptar la misma estructura secundaria, por lo que se usa una regla de "**voto mayoritario**" para corregir inconsistencias.
- Exactitud de este método es de alrededor de 70%



Métodos según su precisión

TABLE 14.2. Comparison of Accuracy of Some of the State-of-the-Art Secondary Structure Prediction Tools

Methods	Q_3 (%)
Porter	79.0
SSPro2	78.0
PROF	77.0
PSIPRED	76.6
Pred2ary	75.9
Jpred2	75.2
PHDpsi	75.1
Predator	74.8
HMMSTR	74.3

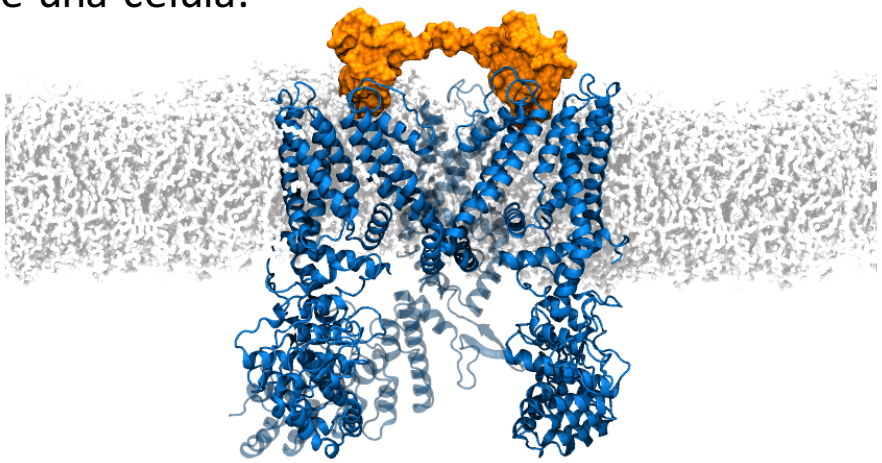
Note: The Q_3 score is the three-state prediction accuracy for helix, strand, and coil.

Puntaje Q3: Se usa un set de proteínas de estructuras conocidas y se determinar el porcentaje de residuos que están correctamente predichos para 3 tipos de estructuras secundarias: *α-hélice, sábana-β, coil*.

Predicción de Estructura Secundaria de Proteínas Transmembrana

- Constituyen hasta el **30%** de las proteínas de una célula.

- Funciones de las proteínas de transporte:
  - Traducción de señales
  - Transporte
  - Conversión energética
  - Target de drogas.



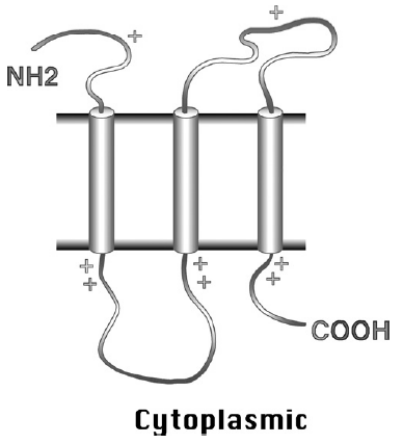
- Su estructura es difícil de resolver por métodos experimentales.

- Los métodos de predicción de estructura de proteínas solubles no funcionan bien para estas proteínas ya que son mucho más hidrofóbicas.

Predicción de Proteínas con α-hélices

- Las **hélices trans-membrana son de aproximadamente 17 a 25 residuos de largo** cada una, y los residuos son predominantemente **hidrofóbicos**.

- Las **hélices están conectadas por loops hidrofílicos** que comúnmente tienen menos de 60 residuos.



- **Regla del interior positivo**: Los residuos que bordean las hélices transmembrana hacia el citoplasma son cargados positivamente.

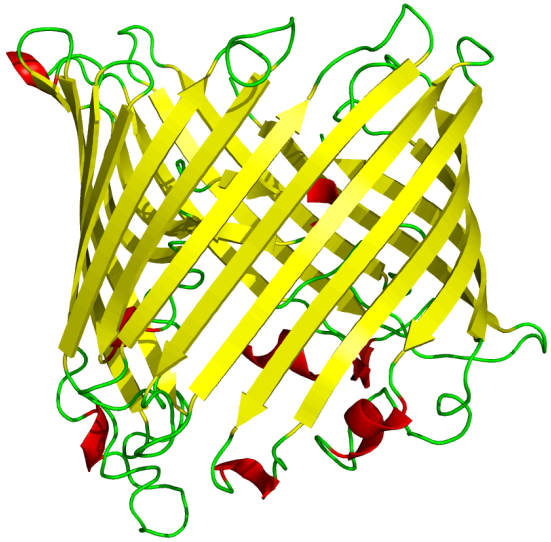
- Muchas proteínas transmembrana tienen **péptidos señal** que los localizan a la membrana, y muchos algoritmos sufren el problema que predicen este péptido como una hélice. La solución es remover peptidos señal antes de hacer la predicción.

Predicción de Proteínas con barriles β

- Existe sólo un número limitado de estas estructuras, por lo que el poder predictivo no es muy alto. Ocurren sólo en organismos procariontes.

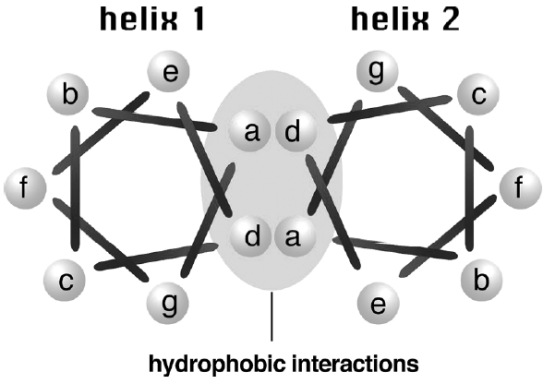
- Las sábanas-β que forman los barriles β transmembrana son de carácter anfipático: un residuo hidrofóbico hacia las bicapas lipídicas, seguido de uno más hidrofílico hacia el poro del barril. Contienen entre 10 a 22 residuos.

- **TBBpred** es un algoritmo usado para este propósito. Implementa dos tipos de métodos de aprendizaje automático: "redes neuronales" y "máquinas de vectores de soporte".



Predicción de Coiled Coils

- Involucran dos o más α-hélices de la misma o diferentes proteínas. Facilitan interacciones entre proteínas, principalmente en la regulación de la transcripción (cierres de leucina) y en la mantención de la integridad del citoesqueleto.



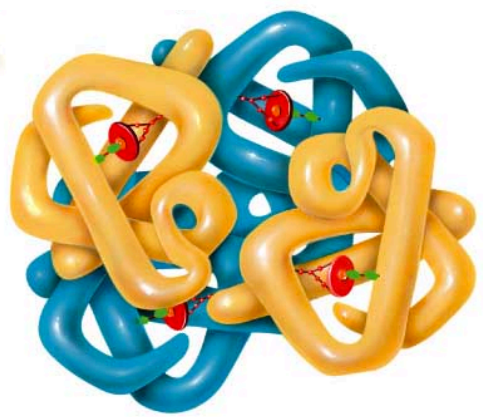
- **Cada siete residuos**: El primero y el cuarto son hidrofóbicos y miran hacia la interface, y el resto son hidrofílicos y se exponen al solvente.

Predicción de Estructura Terciaria

- Al día de hoy hay reportadas ~56.5 millones de secuencias de proteínas diferentes, mientras que **sólo se conocen ~100 mil estructuras tridimensionales**.

- Para entender la función de una proteína necesitamos conocer su estructura.

- Un **modelo 3D** de una proteína sirve para generar hipótesis que pueden ser validadas experimentalmente, tales como mutación sitio dirigida, estabilidad de la proteína, y análisis funcionales.



Métodos

Existen tres tipos de aproximaciones computacionales para predecir estructuras 3D de proteínas:

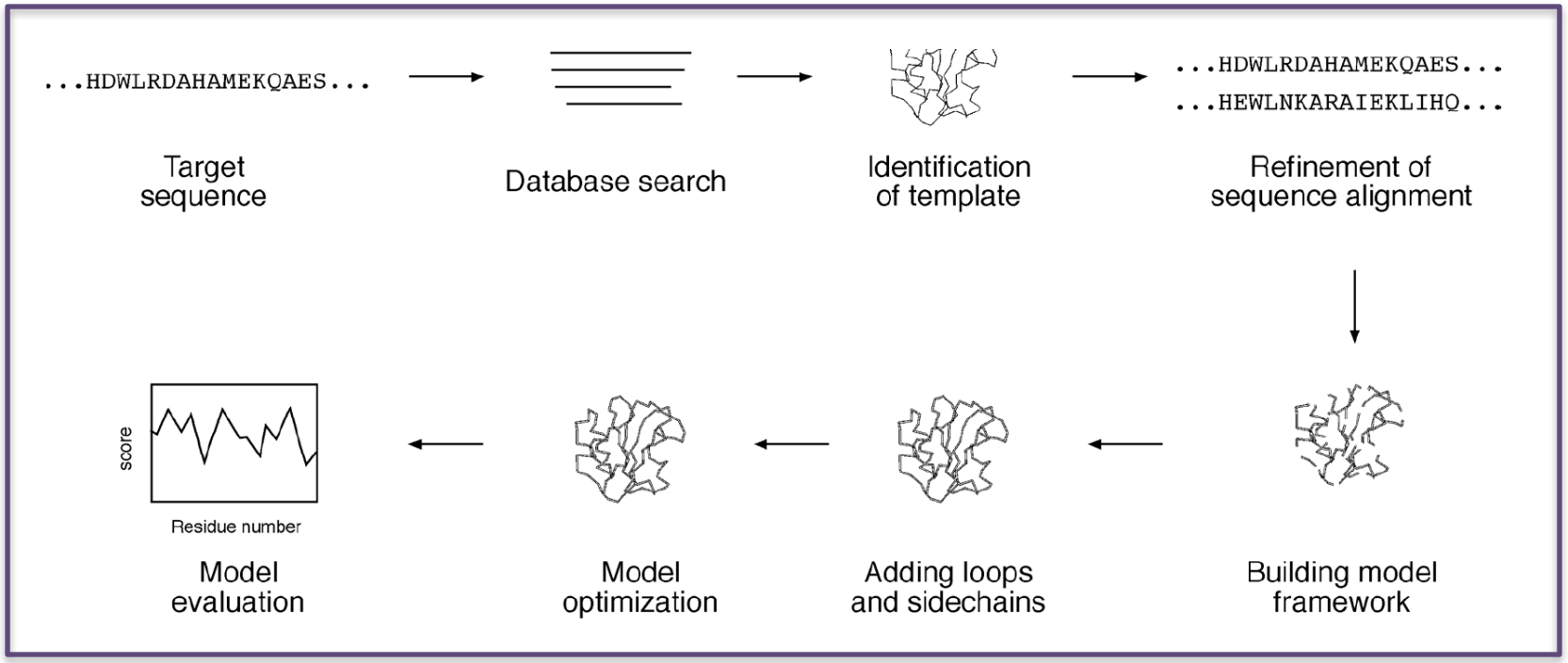
1. **Modelamiento por homología:** Basado en proteína de estructura conocida con alta identidad de secuencia.

1. **Threading:** Identifica proteínas que son estructuralmente similares independiente de la similitud de secuencia.

1. **Predicción Ab-Initio:** Basado en principios fisicoquímicos que gobiernan el empaquetamiento de proteínas, sin la necesidad de identificar un templado inicial.

1. Modelamiento por Homología

- También conocido como *Modelamiento Comparativo.*
- Basado en el principio que sí dos proteínas tienen **alta similitud de secuencia**, muy probablemente tienen estructuras 3D similares.
- El producto final es un modelo atómico de la proteína, basado en alineamiento con una proteína **templado**.

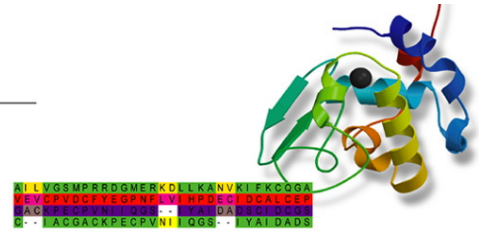


1. Método por Homología: Servidores Web y Bases de Datos

Modeller: Servidor web donde el usuario entrega un alineamiento de la secuencia problema con uno o más templados. Distingue regiones conservadas de las no tan conservadas. Estas últimas, incluyendo loops, se dejan bien libres de cambiar, mientras que las regiones conservadas quedan casi fijas. Utiliza minimización de energía y dinámica molecular para optimizar el modelo.

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



ModBase. Contiene modelos generados para casi 5 millones de secuencias, a partir del programa Modeller. Por ejemplo, contiene modelos para 58631 de las proteínas del proteoma humano. Son muy útiles, por ejemplo, para estudiar interacciones entre proteínas y drogas.

ModBase: Database of Comparative Protein Structure Models

2. Threading

- **La estructura de proteínas es mucho mas conservada que la secuencia**. Por tanto, proteínas pueden tener empaquetamientos similares en ausencia de similaridad de secuencia.

- Threading predice el empaquetamiento estructural de una secuencia de estructura desconocida comparando su **predicción de estructura secundaria, accesibilidad al solvente y polaridad** con la de empaquetamientos conocidos.

- En adición a la información estructural anterior, **para cada familia también se determina un perfil de PSI-BLAST**.

- De esta manera, threading **encuentra relaciones evolutivas** no dependiendo solamente de identidades de secuencia.

3. Predicción de Estructura Ab-Initio

- Produce un modelo 3D de una proteína basado solamente en información de secuencia, sin necesidad de identificar homólogos de secuencia o de estructura.

- La ventaja es que puede predecir estructuras con nuevos empaquetamientos.

- La desventaja es que **nuestro conocimiento de los procesos que gobiernan el empaquetamiento de proteínas es limitado**, y por tanto los resultados son bastante inexactos.

- Se basan en la idea de que las estructuras de proteínas buscan el mínimo de energía, y por tanto **utilizan el procedimiento de minimización energética**.

- **Rosetta es un método híbrido**. Divide una secuencia a modelar en fragmentos, y utiliza threading para predecir la estructura de cada uno de ellos. Luego los fragmentos se unen al azar, y usando el principio de minimización energética, elige el mejor modelo.